U.S. DEPARTMENT OF
ENERGY

Office of
ENERGY EFFICIENCY &
RENEWABLE ENERGY

ADVANCED MATERIALS &
MANUFACTURING
TECHNOLOGIES OFFICE

DRAFT REPORT

Energy Efficiency Scaling for Two Decades Research and Development Roadmap

Version 1.0

August 2024

Disclaimer

Authors

The initial draft of this report was prepared by Energetics, Inc., for the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy, Advanced Materials and Manufacturing Technologies Office (AMMTO) based on inputs from a year-long roadmapping effort involving the participating organizations below documented at **https://live-slac-microelectronics-d9.pantheonsite.io/sustainable-computing/energy-efficient-and-sustainable-computing**.

The primary AMMTO author and Energy Efficiency Scaling for Two Decades (EES2) lead is Tina Kaarsberg, AMMTO.

The primary EES2 analysis lead and author is Sadasivan Shankar, SLAC National Accelerator Laboratory.

The primary Energetics, Inc., authors of this report are I-Hsi Daniel Lu, Nick R. Johnson, Kenta Shimizu, Emmanuel Taylor, Russ Jones, Harper Alerion, and Ken Weaver.

The authors acknowledge Matt Roney, Dwight Tanner, and Jessica Blackburn from AMMTO's technical writing team who provided invaluable editorial support.

They also acknowledge Thomas Finamore, the Energetics, Inc., producer who provided graphic editorial support.

Participating Organizations[1]

Government

U.S. Department of Energy AMMTO

U.S. Department of Commerce National Institute of Standards and Technology

National Laboratories

Argonne National Laboratory

Brookhaven National Laboratory

Fermi National Laboratory

Lawrence Berkeley National Laboratory

Lawrence Livermore National Laboratory

National Renewable Energy Laboratory

Oak Ridge National Laboratory

Pacific Northwest National Laboratory

Sandia National Laboratories

SLAC National Accelerator Laboratory

Industry

3D Epitaxial Technologies

Advanced Micro Devices

Aligned Carbon

America's Frontier Fund

America's Frontier Fund

Applied Materials

ARM

BRDG Bridge to Connect

Cadence

Carbice

Carbon Technology, Inc.

Dedalo AI

Dexmat

Energetics

Finwave Semiconductor

General Electric Vernova

GoogleGreat Lake Crystal Technologies

Hyperion

IBM

IEEE- USA

Infineon Technologies

Intel

Iris Light

Liqid

Metis Microsystems

Micron Technology

Microsoft

Multi3D

NanoSonic Inc.

Nantero

Nhanced Semiconductors, Inc

Paragraf

PseudolithIC

Quantum Silicon

Radiation Monitoring Devices, Inc.

SAP

SEMI

Semiconductor Research Corporation

Siemens

[1] Except for the National Institute of Science and Technology, all participants' organizations have formally signed the EES2 pledge.

Sixline Semiconductor
Synopsys
TechSearch International
Tetramer
Tiptek
VIEE
Wolley Tech
Zyvex Labs

Universities

Duke University
Florida Semiconductor Institute
Hasso Plattner Institute
Los Angeles Trade-Technical College
Stony Brook University
University of Colorado at Boulder
University of Nebraska at Lincoln
University of Texas at Dallas
University of Texas at San Antonio

Working Group Co-Chairs

Algorithms and Software

Brian Hirano, Micron Technology

Tapan Shah, General Electric Vernova

Advanced Packaging and Heterogeneous Integration

Moinuddin Ahmed, Argonne National Laboratory

Na Li, Carbice

Circuits and Architectures

Azeez Bhavnagarwala, Metis Microsystems

Emre Salman, Stony Brook University

Education and Workforce Development

Russell Harrison, IEEE-USA

David Shahoulian, Intel

Manufacturing Efficiency and Environmental Sustainability

Josh Ballard, Zyvex Labs

Prashant Nagapurkar, Oak Ridge National Laboratory

Steve Putna, Texas A&M University

Materials and Devices

Jayasimha Atulasimha, Virginia Commonwealth University

John Baniecki, SLAC National Accelerator Laboratory

Paul Fischer, Intel

Shashank Misra, Sandia National Laboratories

Metrology and Benchmarking

Craig Green, Carbice

James Booth, National Institute of Science and Technology

Power and Control Electronics

Paul Sharps, Sandia National Laboratories

Tim McDonald, Infineon Technologies

Acknowledgments

The authors would like to acknowledge the additional significant contributions of the following working group members and Energetics staff members:

Aaron Fisher, Lawrence Livermore National Laboratory

Alexander Paramonov, Argonne National Laboratory

Amanda Petford-Long, Argonne National Laboratory

Amir Ziabari, Oak Ridge National Laboratory

Angel Yanguas-Gil, Argonne National Laboratory

Anil Mane, Argonne National Laboratory

Antonino Tumeo, Pacific Northwest National Laboratory

Ashfia Huq, Sandia National Laboratories

Ben Tang, Infineon Technologies

Bill Gervasi, Nantero

Brian Rowden, Oak Ridge National Laboratory

Caecilia Gotama, BRDG

Can Bayram, University of Illinois

Carlos Gutierrez, Sandia National Laboratories

Chad Husko, Iris Light

Conrad James, Sandia National Laboratories

Dan Green, PseudolithIC

Daniel Gopman, NIST

Danny Clavette, Infinenon Technologies

David Gilmer, Nantero

David Gotthold, Pacific Northwest National Laboratory

David Voss, Energetics

Dhanushkodidurai Mariappan, GE Vernova

Dhireesha Kudithipudi, University of Texas - San Antonio

Elizabeth Neville Reyes, Applied Materials

Eungsan Cho, Infineon Technologies

Francesco Musso, Dedalo AI

Godwin Maben, Synopsys

Hal Finkel, DOE/ASCR

Harish Bhandari, RMDI

Hsien-Hsin Sean Lee, Intel

Hyunim Chung, Sandia National Laboratory

Jack Kotovsky, Lawrence Livermore National Lab

James Provine, Aligned Carbon

Jan Vardaman, TechSearch International

Jigesh Patel, Synopsys

Jo Luo, Nantero

Joe Kline, NIST

Joel Varley, Lawrence Livermore National Laboratory

John Leung, Intel

Josh Fryman, Intel

Julia Deitz, Sandia National Laboratories

Keita Teranishi, Oak Ridge National Laboratory

Keith Lanier, Synopsys

Maria DiGiulian, Infineon Technologies

Martin Frank, IBM

Matthew Weimer, Forge Nano

May Gokhale, Lawrence Livermore National Lab

Muralidharan Govindarajan, Oak Ridge National Lab

Nick Lalena, DOE

Pamela Klabbers, Fermilab

Percy Kawas, Infineon Technologies

Peter Dowben, University of Nebraska - Lincoln

Petr Sushko, Pacific Northwest National Laboratory

Radislav Potyrailo, GE Vernova

Rahul Sen, Nantero

Ricardo Ruiz, Berkeley Lab

Richard Kainradl, SAP

Rick Ridgley, Nantero

Sam Salama, Hyperion

Sameh Khalil, Infineon Technologies

Sean Shaheen, University of Colorado - Boulder

Shari Liss, SEMI Foundation

Srilatha Manne, AMD

Steffen McKernan, Carbon Technology

Steve Buffat, Nantero

Steve Pawlowski, Micron

Sushanta Pal, Nantero

Timothy Wei, University of Nebraska-Lincoln

Tommy Finamore, Energetics

Victor Zhirnov, Semiconductor Research Corporation

Vishal Saxena, University of Delaware

Volker Sorger, University of Florida

Walt Zalis, Energetics

Wayne Johnson, ICF

Wiley Kirk, 3DET

Yiran Chen, Duke University

Yuepeng Zhang, Argonne National Laboratory.

List of Abbreviations

3DHI	3-dimensional heterogeneous integration
AC	Alternating current
ADK	Assembly design kit
AI	Artificial intelligence
ALU	Arithmetic logic unit
AMD	Advanced Micro Devices
ANL	Argonne National Laboratory
ANN	Artificial neural network
AP	Advanced packaging
APHI	Advanced packaging/heterogeneous integration
ASIC	Application-specific integrated circuit
AST	Abstract syntax tree
BEOL	Back-end-of-line
CAD	Computer-aided design
CDO	Carbon-doped silicon oxide
CDR	Clock and data recovery
CD-SAXS	Critical-dimension small-angle X-ray scattering
CD-SEM	Critical-dimension scanning electron microscope
CIM	Compute-in-memory
CISC	Complex instruction set computer
CMOS	Complementary metal–oxide–semiconductor
CMP	Chemical mechanical polishing
CNT	Carbon nanotube
CNTFET	Carbon nanotube field-effect transistor
CPO	Co-packaged optics
CPU	Central processing unit
Cu	Copper
CVD	Chemical vapor deposition
CXL	Compute express link
D2W	Die-to-wafer
DBI	Direct bond interconnect

DC	Direct current
DDR	Double data rate
DFT	Density functional theory
DNN	Deep neural network
DoD	Department of Defense
DOE	U.S. Department of Energy
DRAM	Dynamic random-access memory
DSA	Domain-specific architecture
DSL	Domain-specific language
DTCO	Design technology co-optimization
DTM	Dynamic thermal management
EAM	Electro-absorption modulator
EDA	Electronic design automation
EES2	Energy Efficiency Scaling for Two Decades
EMI	Electro-migration issues
EMIB	Embedded multi-die interconnect bridge
EOM	Electro-optic modulator
ESD	Electrostatic discharge
FeFET	Ferroelectric field-effect transistor
FEOL	Front end of line
FET	Field-effect transistor
FinFET	Fin field-effect transistor
FMEA	Failure mode effect analysis
FP16	Floating point 16 (16-bit representation)
FP32	Floating point 32 (32-bit representation)
FPGA	Field programmable gate array
FSG	Fluorosilicate glass
FTJ	Ferroelectric tunnel junction
GDDR	Graphics double data rate
GPU	Graphics processing unit
GWP	Gross World Product
HBM	High bandwidth memory

HCl	Hydrochloric acid
HDI	High-density interconnects
HDI	Hydrogen Deficiency Index
HF	Hydrofluoric acid
HI	Heterogeneous integration
HIST	Heterogeneous interconnect stitching technology
HMC	Hybrid memory cube
HNO_3	Nitric acid
HPC	High-performance computing
HVAC	Heating, ventilation, and air conditioning
HVM	High-volume manufacturing
I/O	Input/output
IC	Integrated circuit
IC	Interconnect
ICT	Information and communication technology
IEEE	Institute of Electrical and Electronics Engineers
ILD	Interlayer dielectric
IMEC	Interuniversity Microelectronics Centre
InP	Indium phosphide
Int32	Integer 32 (32-bit precision)
Int8	Integer 8 (8-bit precision)
IoT	Internet of Things
IP	Intellectual property
Ir	Iridium
IRDS	International Roadmap for Devices and Systems
ISA	Instruction set architecture
JVM	Java Virtual Machine
KOH	Potassium hydroxide
LED	Light-emitting diode
LMP	Liquid metal paste
LPDDR	Low-power double data rate
MAC	Multiply-accumulate

MAPT	Microelectronics and Advanced Packaging Technologies
MBE	Molecular-beam epitaxy
MIEC	Mixed ion-electric conductor
MIV	Monolithic inter-tier via
ML	Machine learning
MLIR	Multi-level intermediate representation
MOCVD	Metalorganic chemical vapor deposition
MOSFET	Metal–oxide–semiconductor field-effect transistor
MWNT	Multi-walled carbon nanotube
MZM	Mach-Zehnder modulator
NAND	Not AND memory
NaOH	Sodium hydroxide
NASA	National Aeronautics and Space Agency
NDE	Non-destructive evaluation
NIST	National Institute of Standards and Technology
NRAM	Nanotube random-access memory
NRAM	Non-volatile random-access memory
NVM	Non-volatile memory
OEM	Original equipment manufacturer
OOK	On/off keying
ORNL	Oak Ridge National Laboratory
OSAT	Outsourced semiconductor assembly and test companies
OSC	Organic semiconductor
PACE	Power and control electronics
PAM4	Pulse amplitude modulation 4-level
PCB	Printed circuit board
PCIe	Peripheral Component Interconnect Express
PCRAM	Phase change random-access memory
PDK	Process design kit
PDU	Power distribution unit
PFC	Power factor correction
PIM	Process-in-memory

PINN	Physics-informed neural network
pJ	Picojoule
PUE	Power use effectiveness
QD	Quantum dot
QW	Quantum well
RC	Resistive-capacitive
R&D	Research and development
RD&D	Research, development, and demonstration
RDL	Redistribution layer
ReRAM	Resistive random-access memory
RF	Radio frequency
Rh	Rhodium
RISC-V	educed instruction set computer 5
RM	Ring modulator
Ru	Ruthenium
SerDes	Serializer/deserializer
Sandia	Sandia National Laboratories
SIA	Semiconductor Industry Association
SiC	Silicon carbide
SiCN	Silicon carbon nitride
SiGAA	Silicon gate-all-around
SiGe	Silicon germanium
SiO_2	Silicon dioxide
SiP	System in package
Si_xN_y	Silicon nitride
SLAC	Stanford Linear Accelerator Center
SNN	Spike neural network
SNR	Signal-to-noise ratio
SOA	Semiconductor optical amplifier
SoC	System on Chip
SOIC	Small outline integrated circuit
SOI-TFET	Silicon-on-insulator tunnel field-effect transistor

SOP	Small outline package
SRAM	Static random-access memory
SRC	Semiconductor Research Corporation
SSD	Solid state drive
STCO	System technology co-optimization
STEM	Science, technology, engineering, and mathematics
STTRAM	Spin transfer torque random-access memory
SWNT	Single-walled carbon nanotube
TCB	Thermocompression bonding
TCO	Total cost of ownership
TEM	Transmission electron microscopy
TFET	Tunnel field-effect transistor
TIM	Thermal interface material
TMDC	Transition-metal dichalcogenide
TPU	Tensor processing unit
TRL	Technology readiness level
TSMC	Taiwan Semiconductor Manufacturing Company
TSV	Through silicon via
UCIe	Universal Chiplet Interconnect Express
UMC	United Microelectronics Corporation
UPS	Uninterruptible power supply
VCache	Vertical cache memory
W2W	Wafer-to-wafer
WDM	Wavelength division multiplexing

Executive Summary

This R&D roadmap is part of the U.S. Department of Energy's (DOE's) pledge to increase microelectronics' energy efficiency 1,000-fold in two decades. With rapidly emerging challenges such as the increase in electricity use of data centers, innovations that exponentially increase energy efficiency are urgently needed to put microelectronics' electricity use on a more sustainable path (see Figure ES-1). Just as President Kennedy did with his moonshot goal 60 years ago, DOE pledged to achieve this energy efficiency goal not because it is easy but because it is hard.

Background

Since the invention of the integrated circuit or "chip" 65 years ago, semiconductor-based electronics, or microelectronics, have enabled growth of information technology (IT)—computing, communication, and other electronics applications. Chip manufacturers now layer billions of semiconductor-based switches (i.e., transistors, the foundational unit of electronic devices) onto silicon to make the microelectronics that are essential for modern life. IT growth in the last century was propelled forward by the biennial doubling of transistor density on chips, which led to greater performance and lower cost per function. This tradition of exponential performance improvements is why much of the semiconductor industry already sets exponential technical goals. For example, the initial pledge signer, Advanced Micro Devices (AMD), had already set the goal to increase the efficiency of its chips 30 times by 2025; since signing DOE's pledge, AMD has increased its goal to 100 times by 2027.

As transistors were miniaturized, chip power density initially remained constant (Dennard scaling), leading to more than doubling energy efficiency biennially. By 2005, however, this biennial efficiency doubling began to slow markedly as it reached certain physical limits. The slowing of efficiency doubling coupled with the rapid rise in energy and computation-intensive IT applications, has led to sharp increases in global IT energy consumption. According to the Semiconductor Research Corporation (SRC), by 2010, global computing energy use began to double every 3 years.

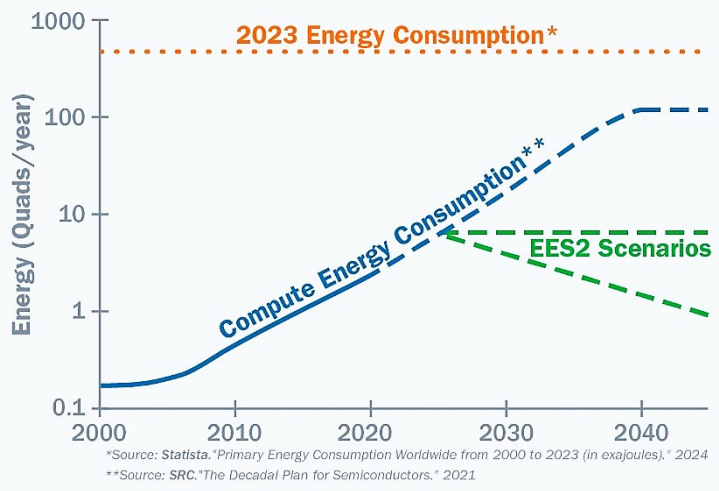


Figure ES 1. SRC energy consumption forecast and the EES2 efficiency goal in two energy consumption scenarios

Energy Efficiency Scaling for Two Decades

The U.S. Department of Energy's Advanced Materials and Manufacturing Technologies Office (AMMTO) launched a national initiative with industry partners, national labs, and academia, called Energy Efficiency Scaling for Two Decades (EES2) in 2022. This ambitious program aims to double the energy efficiency of microelectronics biennially, targeting a 1,000-fold

improvement over two decades. A key component of EES2 is this version 1.0 roadmap, the first in a series of research, demonstration, and demonstration (RD&D) roadmaps. This document is a product of extensive literature review and energy analysis, collaboration meetings between nine working groups, and expert input during the writing process. The working groups met monthly, with the organizing committee engaging in literature review and analysis to prepare for the following meeting.

The EES2 roadmap focuses on the largest and fastest growing IT energy user, the "compute stack" (see Figure ES-2), which comprises everything from devices to software. The stack shown is from the seminal DOE report *Basic Research Needs for Microelectronics* (DOE Office of Science, 2018), which extended the notion of co-design from simply designing hardware and software together to specifically co-designing adjacent layers of the hardware with adjacent layers of the software. This roadmap examines innovative technologies co-designed by experts on different parts of the stack that can exponentially increase computing energy efficiency. This roadmap is a first step in a multi-year research effort to develop and deploy portfolios of cutting-edge microelectronics technologies that are 10-, 100-, and even 1,000-times more energy efficient than the technologies they replace. Alone, none of the technologies will achieve the industry-wide biennial efficiency doubling leading to the 21 1,000-times goal.

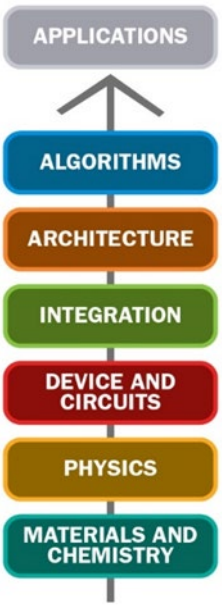


Figure ES-1. Compute Stack. *Source: DOE Office of Science, 2018*

DOE's Undersecretary for Energy and Science and the now sixty-five other external industry-based EES2 pledgers were inspired to work toward this goal and join the roadmap effort on the strength of DOE's 2021–2022 "Semiconductor R&D for Energy Efficiency" virtual workshop series and its 2022 sponsored assessment of computing energy use (Shankar and Reuther 2022), which contributed additional insight on how the stack could be co-designed with rigorous analysis of computing performance using the metrics of energy per bit, per instruction, and per application.

Strategies for Efficiency First

To directly support the goals of EES2, co-design strategies are prioritized to optimize for efficiency first. Simply put, this means that where multiple properties are desired for a given technology solution, energy efficiency should be the first property for optimization in the co-design. In addition, three sub strategies emerged related to the three different energy metrics for the near-, mid-, and long-term, as shown in Figure ES-2.

Near-Term: Optimization of Energy per Instruction/Operation

EES2-sponsored analysis by Stanford Linear Accelerator Center (SLAC) (Shankar and Reuther 2022) showed large variation in energy per instruction or operation for different types of computational tasks. This suggests that a chip design strategy that ensures instruction complexity for a given task is as low as possible is the correct "efficiency first" hardware co-design strategy. Graphic processing units (GPUs) for gaming and artificial intelligence (AI) are a successful example of an approach of this type.

Mid-Term: Device-Level Innovations To Minimize Energy Use per Bit

Because they are so foundational, innovations at the device level, especially with transistors, are critical. In the near- and mid-term, the EES2 roadmap highlights innovations that sharpen the subthreshold swing slope and lower switching voltage, such as tunnel field-effect transistors (TFETs). In the mid- to long-term, device level innovations from quantum and nature-inspired computing will be critical for widespread advances from 100x to 1,000x energy efficiency.

Long-Term: Full-Stack Software-Driven Co-design To Minimize Energy per Application

The goal of full-stack co-design has yet to be implemented. This strategy is accelerating toward this goal by focusing on a subset of full-stack co-design that is software-driven, requiring that hardware developers understand what the software needs to do, and software designers understand the needs of hardware. Full implementation would require a major change in pedagogy and curriculum since software and hardware engineers have become more and more specialized in recent decades. But in the meantime, steps in this direction include specifying in algorithms that do not require high precision to save energy. Figure ES-2 illustrates the interaction between different layers of the compute stack, the timeline for the innovation, and the correlated energy metric.

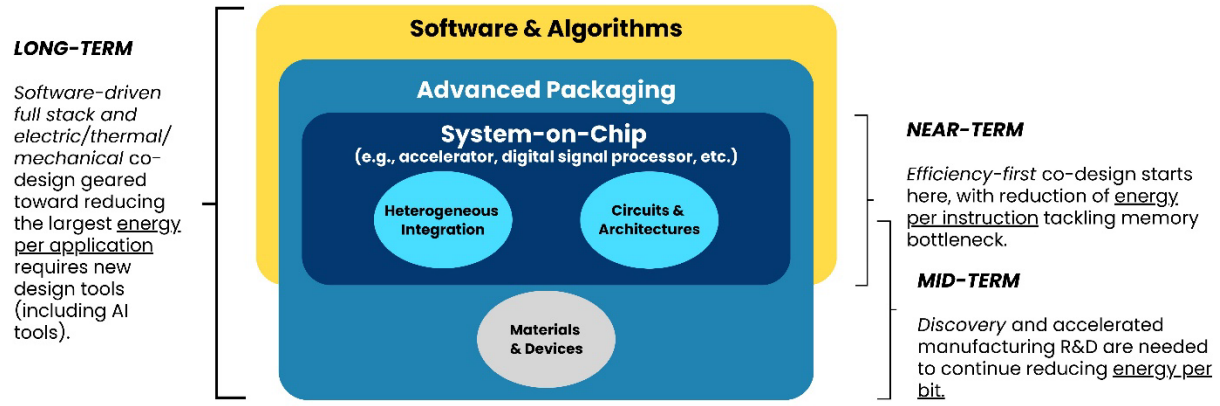


Figure ES-2 Relationship of compute stack elements to achieving energy efficiency goals with different time frames

Expanding Co-Design for the Compute Stack→ Stack Working Groups

When the concept of co-design was first applied to microelectronics, it simply meant the integration of hardware and software design in computing. Compute is the largest microelectronics system energy user—hence the focus of this roadmap. As the complexity of the compute stack grew (see Figure ES-3 left side), numerous subcategories of hardware and software were developed. In order to achieve the benefits of co-design envisioned by DOE in its seminal *Basic Research Needs for Microelectronics* (2018) report, co-design for energy efficiency must ensure that adjacent elements of the stack work together. AMMTO's DOE partners in SC defined co-design in this 2018 report as "where each of the technical abstraction layers in modern computer system design (the compute stack), from fundamental materials research through applications, inform and engage other abstraction layers." Furthermore, "co-design activities largely occur between adjacent technology abstraction layers (e.g., between materials and devices or computer architects and software designers)." Therefore, the initial four

EES2 working groups (WGs) were assembled, as shown in Figure ES-3, to ensure co-design among adjacent layers.

Pledger Experience Led to Inclusion of Enabling Layers of Co-Design

EES2 industry and laboratory pledging partners with experience in the rapidly growing data center sector also urged the inclusion of power in the WGs' co-design approach. Thus, a Power and Control Electronics WG was added. In addition, since the National Institute of Standards and Technology (NIST) had been involved with pre-EES2 efficiency efforts—and the EES2 team knew the importance of metrology to keep track of efficiency goals—a Metrology and Benchmarking WG was included from the beginning. Finally, the analysis EES2 was conducting in parallel with the WGs showed that manufacturing energy use, complexity, and chemical intensity also had begun to rise rapidly in recent years, so a Manufacturing Energy Efficiency and Sustainability WG was included.

Early in the process, the working groups realized that past efforts at co-design had not generally involved software for hardware, such as the proprietary electronic design automation (EDA) software used to design circuits. To rectify this issue, the Circuits and Architectures WG began to meet with other WGs. In 2018, the co-design was mainly between adjacent layers, but by 2023, it had become clear that every aspect of the compute stack, plus every aspect of the enablers, needed to be aligned to minimize energy use.

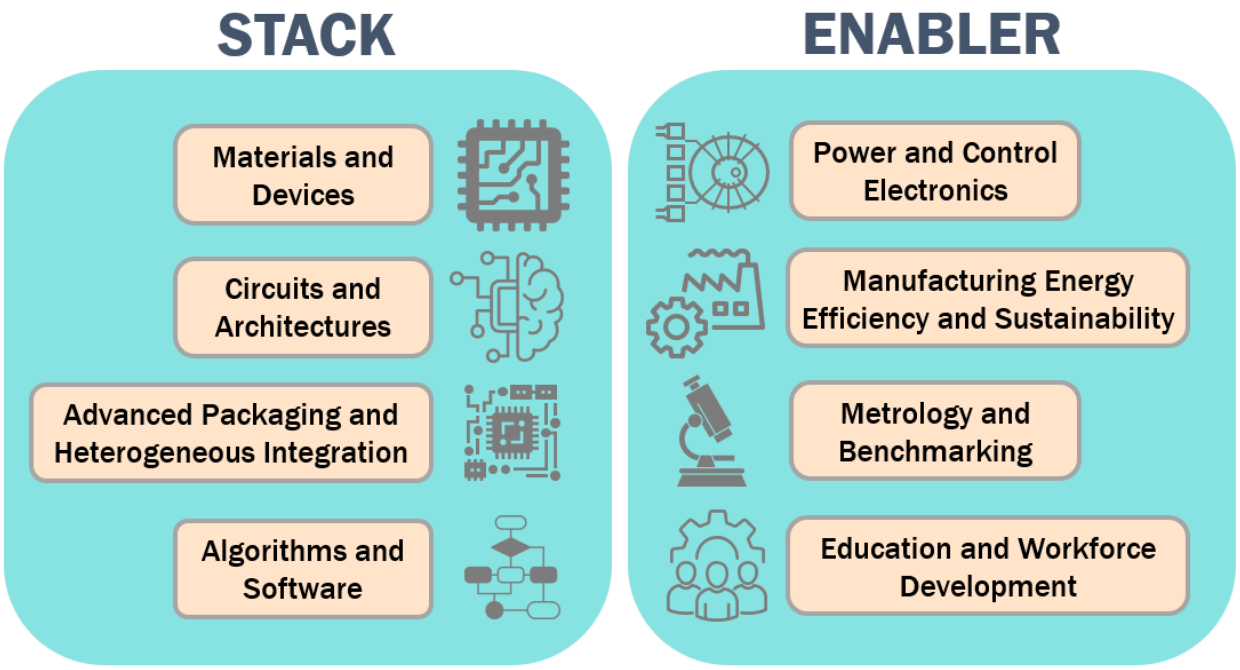


Figure ES-3. Organization of the EES2 working groups

With the help of cross cutting pledgers such as SRC and IEEE (the world's largest professional society), an Education and Workforce Development WG was formed. A co-design focus between working group areas is essential to make the rapid efficiency progress needed for biennial efficiency doubling and to ensure that the effort is both technically feasible and commercially viable.

The general scope of each WG is described in the following section.

Compute Stack Co-Design Working Groups

The **Materials and Devices** group tackled energy efficiency through materials and devices, such as carbon nanotubes and spintronic memory. This included scalability, thermal management, and interface issues in current materials.

The **Circuits and Architectures** group worked to overcome the challenges of slowing planar geometric scaling of transistors and memory. This group pioneered alternative, energy-efficient designs in processors and memory systems, including compute-in-memory technologies.

The **Advanced Packaging and Heterogeneous Integration** group at the next level up in the compute stack, worked on advanced thermal management techniques, and optimizing data movement strategies such as optical interconnects.

The **Algorithms and Software** group emphasized software-driven co-design and were inspired by natural systems such as dragonflies and human brains to benchmark neuromorphic algorithms matched directly with accelerator hardware.

Table ES-1. Condensed Focus Areas for Energy Efficiency and their Manufacturing Challenges & Solutions[a]

Focus Areas for Energy Efficiency	Manufacturing Challenges & Solutions
Materials and Devices (Mid-Term)	
Innovate in materials such as 2D materials, carbon nanotubes (CNTs), and ferroelectric materials for future CMOS alternatives.	Address production and integration challenges by investing in scalable high-quality material manufacturing and creating industry-wide standards and protocols.
Circuits and Architectures (Near-Term)	
Enhance energy efficiency in compute architectures and memory technologies.	Prioritize advanced electronic design automation (EDA) and new architectures integrated with algorithms to optimize power distribution and increase energy efficiency, backed by continued investment in novel device technologies.
Advanced Packaging and Heterogenous Integration (Near-Term)	
Develop vertically integrated, energy-efficient 3D technology stacking.	Pair novel technologies with state-of-the-art processors/memories to show durability and enhance intra-chip energy efficiency, improving EDA for system-level cooling and interconnect scaling.
Algorithms and Software (All Time Scales but Especially Long-Term)	

Innovate in machine learning algorithms and software that efficiently support diverse computing architectures.	Develop machine learning optimization through meta-learning and exploit massively parallel computing systems more effectively, using advanced parallelization of code.

[a] For complete list, refer to Table 85.

Crosscutting Co-Design Working Groups (Also Known as Enablers)

The **Power and Control Electronics** group focused on enhancement and innovation of power delivery systems on chip as well as in energy intensive applications such as data centers.

The **Manufacturing Energy Efficiency and Sustainability** group looked at the correlation between less efficient products and less efficient manufacturing processes to make them. The group also explored other energy-related environmental impacts of manufacturing.

The **Metrology and Benchmarking** group defined measurement and benchmarking standards necessary to evaluate emerging microelectronic technologies.

The **Education and Workforce Development** group took advantage of the compelling EES2 benefits to the planet for efforts to convince policy makers and potential new industry employees.

Table ES-2. Condensed Focus Areas for Energy Efficiency and Their Grand Challenges and Solutions[a]

Focus Areas for Energy Efficiency	Manufacturing Challenges and Solutions
Power and Control Electronics (Very Near-Term)	
Enhance power delivery and control across microelectronics to data centers by migrating loads to higher-efficiency regions and utilizing renewable resources.	Develop resource-aware scheduling strategies and implement advanced co-design tools to optimize power provisioning and thermal management, reducing overall energy consumption.
Manufacturing Efficiency and Environmental Sustainability (Near-Term)	
Improve manufacturing processes to lower greenhouse gas emissions and energy consumption.	Introduce alternative gases and processes with lower environmental impacts and invest in alternate lithography technologies like nanoimprint lithography (NIL) for energy-efficient manufacturing.
Metrology and Benchmarking (All Time Scales)	
Advance metrology by integrating AI/ML in nondestructive, high-resolution techniques to evaluate complex structures and materials accurately.	Establish comprehensive benchmarking standards and develop advanced metrology tools for real-time analysis, bridging the gap between traditional methods and the needs of emerging technologies.
Education and Workforce Development (All Time Scales but Especially Long-Term)	

Cultivate a skilled workforce attuned to the demands of energy-efficient microelectronics and sustainability.	Align educational outcomes with industry needs, implement targeted training programs, and promote inclusivity to build a diverse workforce capable of driving global innovation.

[a] For complete list, refer to Table 86.

Key Technologies Identified

Figure ES-4 graphically illustrates some promising technology options identified in this roadmap by adjacent co-design approaches to the compute stack, sorted by working group.

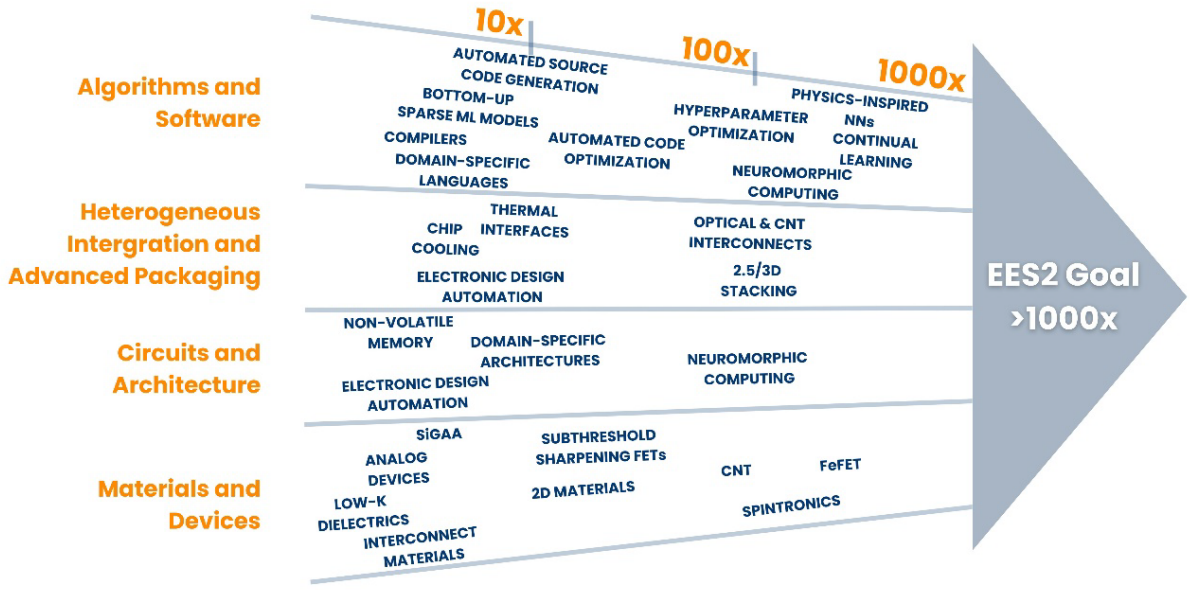


Figure ES-4. Key energy-efficient technologies for industry researchers to beat in each epoch

The figure illustrates one of two major criteria used by the working groups in evaluating candidate EES2 technologies: factor of efficiency improvement based on energy metrics (e.g., energy per bit, energy per switch, memory access) compared to state-of-the-art technologies. Note the semi-log tick marks where the efficiency factor increases logarithmically to the right.

Note that these technologies do not in any way represent a government plan for energy efficiency. Rather these technologies and the technology areas from which they spring are technologies with “energy efficiency to beat.” Rather than being a plan or even a forecast, the roadmap seeks to provide benchmarks that will inspire technology developers to apply the recommended efficiency first design principles and possibly prove wrong dire predictions for future computing energy use.

Next Steps

This version 1.0 of the EES2 roadmap is the end of the beginning of a two-decade effort to take energy efficiency scaling from historical fact to future reality. The demand for computing and the critical need to curb emissions require an acceleration and expansion of current initiatives.

In addition to the roadmap, DOE intends for the EES2 partners to begin a cycle of challenging one another on increasingly ambitious energy efficiency goals. For example, AMD has already begun to challenge the industry on AI chip efficiency—and a response from other AI chip makers is hopefully forthcoming soon. Other non-roadmap EES2 activities planned for the pledgers include the establishment of a testing facility to measure the relative efficiency of highly energy-intensive software (e.g., AI training, Transformer) due to the emergence of many different AI chip architectures. Such a testing facility would also verify the orders of magnitude efficiency improvements of AMMTO and other government funded hardware, such as TFET and neuromorphic chips.

Finally, AMMTO hopes that the EES2 partners will continue to document and learn from microelectronics' past and forecasted future ability to enable all sectors of the economy to become more energy efficient and sustainable. EES2 partners will also continue to identify and publicize the problems solved and the opportunities offered by the roadmap 1.0 and the analysis performed for EES2. A surge in energy use forecasted for data centers is the first of many challenges to which the EES2 community will forcefully respond. Future potential energy-use surges related to communications (such as those that may accompany 6G+) will also be identified, documented, learned from, and publicized by the EES2 community as it evolves from a government-led organization to one that is privately led.

While this report documents myriad potential efficiency improvements across 55 technologies, achieving their full benefits requires an integrated approach that emphasizes software-driven co-design across the entire technology stack. Ultimately, EES2 hopes to reboot the energy efficiency doubling pace of Dennard scaling doubling efficiency every two years—with the goal of reaching 1,000 times more in the next 20 years.

Plans for roadmap 2.0 are already underway. As DOE and its partners recruit more industrial, academic, and national laboratory members for the EES2 innovation ecosystem, the initiative will not only have more policy impact, but it will also boast even broader technical expertise among the WGs. Now that the first roadmap is published, EES2 will actively turn to broaden its recruiting into new microelectronics application sectors, such as communications. In addition, while EES2 started with electronics and electrons, it will also broaden to promising new information carriers, such the photons used in optoelectronics and photonics. EES2 already includes pledgers whose research includes long-term transformational technology areas such as quantum computing as well as the latest advances in nature-inspired architectures. EES2 will work with these pledgers to help recruit additional pledgers from their respective sectors and to attract more volunteers for the version 2.0 WGs.

Although much can change before the start of version 2.0 of the roadmap in spring 2025, future WGs will continue to build upon a solid base of peer-reviewed research while continuing to work with EES2 pledgers to lower barriers toward immediate deployment of technologies for biennial microelectronics energy efficiency doubling. This dual R&D and deployment strategy ensures flexibility and responsiveness to emerging technologies and market shifts, thereby fostering a sustainable evolution of the microelectronics sector.

As the EES2 Initiative continues to grow and build momentum for massive improvements in computing energy efficiency, the EES2 team will work further with stakeholders in microelectronics and related applications to develop the technology base and to assess progress toward the goal every 2 years.

This roadmap is not intended to serve as a forecast or to pick winners and losers among technologies. Rather, it is the opening salvo in a new energy efficiency "space race," where instead of outer space, the EES2 team explores the fascinating realm of increasingly tiny and ultra energy efficient information systems. The roadmap sets a high bar to challenge and motivate technology developers and to counteract grim forecasts that humanity cannot achieve the clean energy transition due to rising computing energy use trends. The semiconductor industry's inspiring past successes in improving energy efficiency indicate that ambitious EES2 efficiency goals can be met as well. Let's do it now.

Table of Contents

List of Figures

List of Tables

SECTION

1

Introduction

1 Introduction

The semiconductor manufacturing industry makes the integrated circuits, or chips, that drive innovation and productivity throughout the global economy. The U.S. Department of Energy (DOE) has significant expertise and experience with the semiconductor industry and has led related research on everything from specialized chips to artificial intelligence. Both DOE's semiconductor expertise and semiconductor-based technologies themselves are critical for its missions in national security, scientific research, and clean energy technologies.

1.1 Background

The U.S. economy was damaged when southeast-Asia-dominated semiconductor supply chains for chips were severely disrupted during the pandemic. For example, the Federal Reserve estimated that, due to the chip supply shortage, the slowdown in U.S. automobile manufacturing alone cost the U.S. economy nearly $240 billion, or more than 1% of the U.S. gross domestic product (GDP). In response, the White House commissioned a Supply Chain Report that included a report chapter on semiconductors led by DOE (Mann and Putsche 2022), which showed that the U.S. semiconductor manufacturing industry had shrunk to only 10% worldwide manufacturing. The chapter also showed that the last time the U.S. civilian government had been involved in the semiconductor industry—in the 1990s, when it provided billions for the SEMATECH consortium—the U.S. accounted for more than 37% of semiconductor manufacturing. In August 2022, the Administration signed the CHIPS and Science Act into law.

As public support grew for government support of the domestic semiconductor industry, so did discussions among federal agencies about investment in semiconductor research and development (R&D). AMMTO's predecessor office--the Advanced Manufacturing Office—sponsored a series of virtual workshops on Semiconductor R&D for Energy Efficiency. In September 2022, DOE launched the Energy-Efficiency Scaling for 2 Decades (EES2) initiative for the semiconductor industry and its major energy using applications. DOE's intent in developing EES2 was to have a simple goal to drive research progress. In November 2022, DOE's Advanced Materials and Manufacturing Technologies Office (AMMTO) launched the EES2 R&D Roadmap effort that resulted in this report.

This version 1.0 roadmap focuses on the largest and fastest growing semiconductor application—computing. As detailed in the next section, the historical semiconductor scaling that inspired energy-efficiency scaling applied only to chip energy use. With energy-efficiency scaling, DOE hopes to drive innovation across the entire compute stack from transistors to software. Subsequent roadmaps will apply this broad principle to other fast-growing microelectronics-based applications, such as communications.

1.1.1 Moore's Law

The semiconductor manufacturing industry is unique compared to traditional manufacturing industries since its key product's performance improvements—including the size, cost, density, and speed of components over the past half century—increase exponentially. Hence, the exponential trends of the semiconductor industry must be plotted on a semi-log-scale plot where time in years on the horizontal axis is on a linear scale and the vertical axis is on a logarithmic scale with each notch representing an order of magnitude increase. The classic example of such a trend is Moore's Law, where the number of transistors on an integrated circuit (IC) or

chip doubles about every two years. While originally Moore's Law was simply an observation by Intel founder Gordon Moore of a trend based on his experience in manufacturing—not a law of physics—it has served the semiconductor industry well as a unifying benchmark. However, Moore's Law needed to evolve (e.g., from two-dimensional [2D] to three-dimensional [3D]) to continue. Predictions beginning decades ago that Moore's law was ending proved premature. For example, in late 2023, chips were launched with more than 100 billion transistors, continuing the biennial efficiency doubling trend.

1.1.2 Dennard Scaling and the Initiative

The scaling relationship that ended about two decades ago was Dennard scaling. When transistors were planar (i.e., 2D), Robert H. Dennard showed that as the number of transistors on a chip doubled, their power use remained constant. This doubling of transistors on a chip of the same size does not increase the power it uses because power use stays in proportion with area, while both voltage and current scale downward with length. As a result, energy efficiency of chips following Moore's Law doubled every 2 years until Dennard scaling ended.

Most experts say that Dennard scaling ended between 2005 and 2006. As the voltage needed to switch the transistor steadily declined (as transistors used less power), it neared the limit where random thermal noise could also cause unintentional switching (i.e., classical "leakage" of current). Additionally, as transistor dimensions dropped to the nanoscale, quantum tunnelling (i.e., quantum leakage) began to occur. Both types of leakage also cause the chip to heat up, which further decreased its energy efficiency due to the additional energy needed to keep the chip cool. The end of Dennard scaling was one of several factors that contributed to the beginning of exponential growth in computing electricity use that became noticeable by 2010.

1.1.3 Why 20 Years? Why 1,000 Times?

During the three decades or so of Dennard scaling, semiconductor chips had biennial efficiency doubling. Even after this, continuing innovation driven by Moore's law maintained efficiency doubling, although at a slower pace. The EES2 goal is based on the notion that a simple goal such as Moore's Law can be a driver of progress and a unifying theme for the industry. The 20-year duration of the EES2 goal is a result of the desire to be 1,000 times more energy efficient using the same efficiency doubling that worked previously for the industry. As shown in Table 1, 1,000 times (actually 1,024 times) is simply the mathematical result of doubling something 10 times: 2^{10}=1,024 over 20 years.

Table 1. 20 Years of Biennial Energy Efficiency Doubling

Year	n=number doublings	Energy Efficiency: 2^n
2	**1**	**2**
4	**2**	**4**
6	**3**	**8**
7	**~10 times**	
8	**4**	**16**
10	**5**	**32**
12	**6**	**64**
13	**~100 times**	
14	**7**	**128**

16	**8**	**256**
18	**9**	**512**
20	**10**	**1,024 (~1,000 times)**

The 1,000 times rationale is similarly straightforward based on the question: where do we want to be in 2043? Figure 1 shows an upper curve efficiency scenario for EES2 deployment that has energy use flattening onto a trajectory parallel to the global energy consumption. A more optimistic EES2 efficiency deployment scenario is depicted as a lower curve returns to mid-2010's energy use over time. The industry can achieve 1,000 times because it has done so before. The continuation of Moore's Law is the result of numerous innovations in thin film, lithography, and various other microscale manufacturing processes. The EES2 initiative seeks to turn this innovation engine toward efficiency. The EES2 Analysis, the EES2 predecessor workshops, and this roadmap show that this can be done.

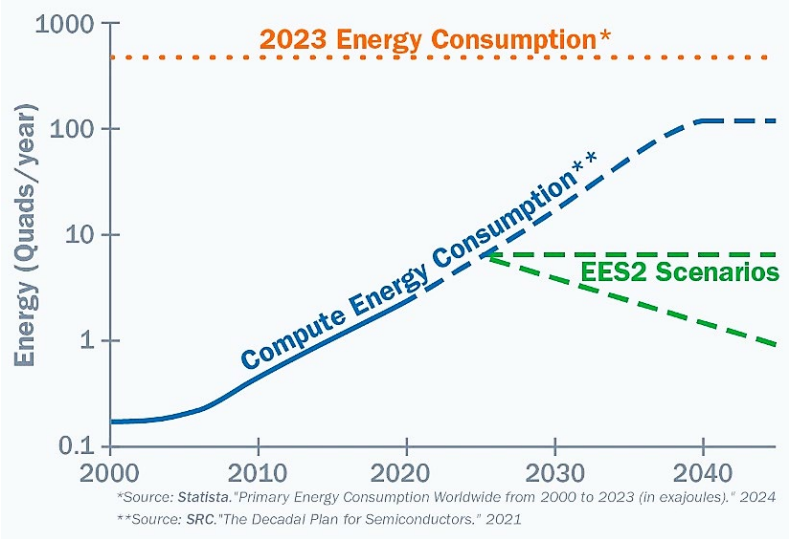


Figure 1. The 2023 version of the SRC computing energy use forecast

1.1.4 Linear Versus Exponential Growth

Understanding the distinction between linear and exponential growth is crucial for understanding why computing and communication electricity use—currently still just a few percent of electricity use—could very rapidly become difficult to sustain economically or environmentally. While linear growth is intuitive and manageable, exponential growth is not. For example, if the doublings shown in the rightmost column of Table 1 were instead for electricity growth, what seemed like a nonproblem in years 1–10 would begin to become an issue in years 11–15 and become a serious problem verging on emergency as the unit of growth doubled from 256 to 512 and increased again to 1024 in years 16–20. Since exponential growth is so nonintuitive, key graphs depicting exponential growth in this report (e.g., Figure 1 for Moore's law and Figure 1 showing the Semiconductor Research Corporation (SRC) analysis and the EES2 goals) show all the zeros of the actual number rather than using shorthand exponential notation (e.g., 1,000,000,000 rather than 10^9). Space considerations prevent the avoidance of exponential notation in many other key plots of the report, but readers are advised to keep Table 1 in mind when interpreting them.

1.1.5 History of the Pledge and Pledgers

EES2 counters unsustainable exponential growth in electricity demand with exponential growth in efficiency. The need for this concerted joint industry-government was first articulated on Jan. 12, 2022, when DOE announced its goal to increase microelectronics (and its applications')

energy efficiency by 1,000 times in 20 years *or less*. “Or less” is an important part of the goal since preliminary EES2 analysis was already showing that the doubling time for AI computing electricity use was beginning to shorten. From the outset, the EES2 team knew that the EES2 efficiency doubling time might need to shorten if demand accelerated. As a result, another key part of the EES2 initiative involves partnering with data collection agencies (e.g., U.S. Energy Information Administration [EIA]) on more comprehensive collection of microelectronic applications’ (e.g., computing and communication) energy use. DOE also announced its intent to work with industry on joint R&D road mapping based on DOE’s new concept of energy efficiency scaling. DOE then developed the concept of the EES2 pledge to organize this effort.

By September 2022, DOE, together with an initial group of 20 other organizations, pledged to cooperate on identifying solutions to drive energy efficiency scaling by developing the first EES2 roadmap and by the end of 2023. These partners also pledged to cooperate on updates needed to the pledge and the roadmap and to catalyze deployment of the technology solutions identified in the roadmap(s). The number of signatories of the EES2 cooperation pledge more than tripled since then to 65 organizations at the time of this writing, with the current EES2 Pledgers listed in the acknowledgements of this report.

The EES2 cooperation pledge reads as follows.

We the undersigned agree to cooperate:

- *To document and learn from the extraordinary record of microelectronics’, including power electronics’, energy efficiency, such as increases greater than 1,000,000 times in energy efficiency since the invention of the transistor nearly 75 years ago.*
- *To document and learn from microelectronics’ past and forecasted future ability to enable all sectors of the economy to become more energy efficience and sustainable.*
- *To identify and publicize problems solved and opportunities offered by microelectronics’ Energy Efficiency Scaling over 2 Decades (EES2).*
- *To publicize and identify sources to fund version 1.0 (2022–2023) of the EES2 RD&D roadmap.*
- *To participate in version 2.0 (2024–2025) of the AMMTO-led EES2 RD&D roadmap.*
- *To explore formation of a partnership, perhaps “EES2 Allies,” that enable the EES2 1,000 times efficiency goal by leading EES2 RD&D roadmapping after 2025 and by catalyzing the deployment of cost-effective technologies, including power electroncs,*

needed to stay on the EES2 path of doubling microelectronics' energy efficiency every 2 years.

We do this because:

- *Microelectronics' life cycle energy use is rapidly becoming unsustainable as microelectronics demand begins to outpace continuing efficiency improvements due to burgeoning computing, communication, and electrification demands.*
- *EES2 is a key organizing principle that aims to help meet new energy demands.*
- *The EES2 is a technology leadership path that provides economic and other public benefits.*

To achieve the EES2 goal, this version 1.0 roadmap identifies numerous candidate technologies to beat that were identified by working groups comprising paired elements of the compute stack.

It's important to note that the EES2 version 1.0 WG volunteers may not have had (or have been able to share) all the technology insights developed by their respective organizations, and that not every single member of the semiconductor innovation ecosystem was represented in our working groups. Nevertheless, the "technologies to beat" are meant to represent an aggressive challenge to the entire computing innovation ecosystem—especially amongst its highly competitive industry members—to foster rapid change and a refocus on efficiency; to boldly outdo each other, and even themselves, in technological innovation.

1.1.6 Analysis Metrics: Energy per Bit, Instruction, and Application

In the EES2 analysis of the headroom for efficiency innovation (see Figure 2) we used three metrics: energy per bit, energy per instruction, and energy per application to identify opportunities, we also used them whenever possible in benchmarking technology candidates. The first metric, energy to flip a bit, is the lowest energy and has historically been the driver of pre-2005 microelectronics efficiency gains.

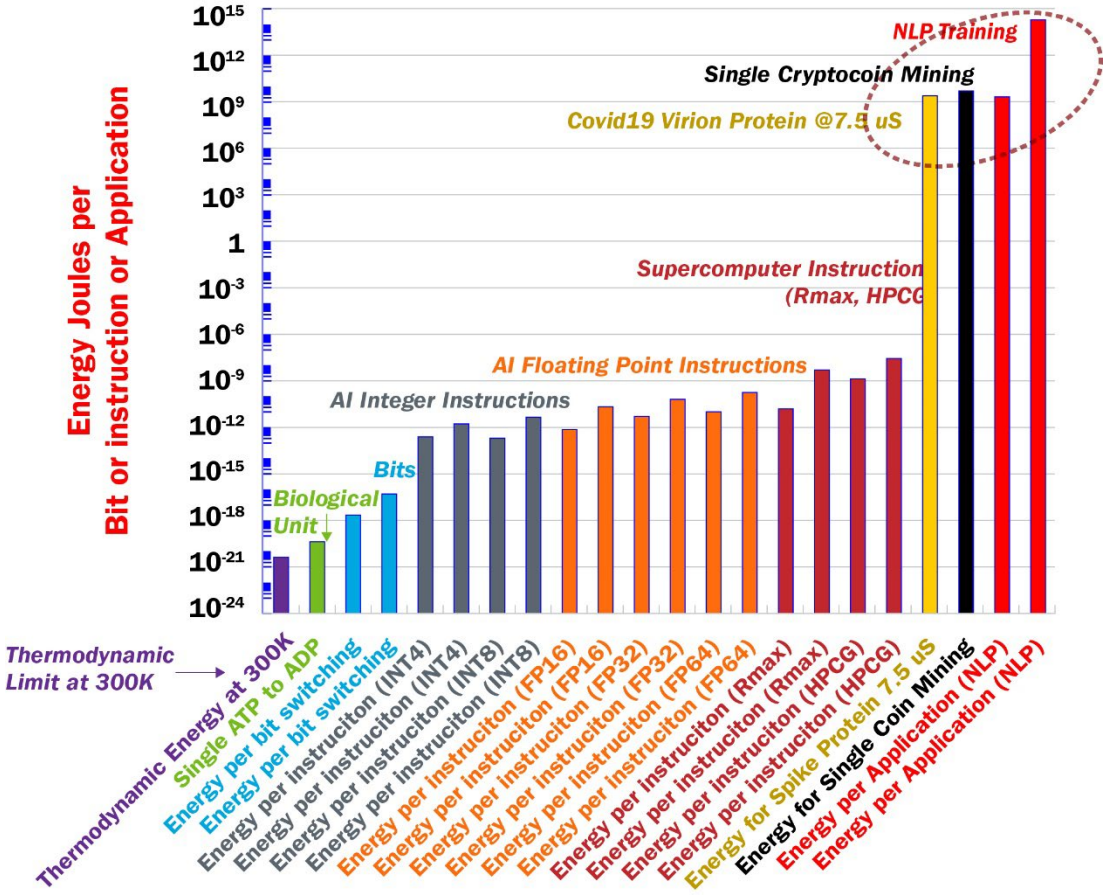


Figure 2. Analysis of the opportunity space for energy efficiency (from bits to bitcoin) according to the bits, instructions and application metrics. *Source: Shankar 2022*

Figure 4 and the analysis that accompanied it show that instructions are the new “low-hanging fruit” of potential microelectronics efficiency gains.

Instructions

This category of innovation potential focuses on reducing the 1,000,000-times difference between the highest and lowest energy per instruction—and then using in algorithms and software the lowest energy operation possible. For example, the energy hungry inference part of an AI calculation can often use far less precise instruction types. The “efficiency first” strategy is to ensure each instruction maximizes its contribution to overall system performance while minimizing energy consumption. A very high leverage approach identified by the WGs would be to provide electronic design automation (EDA) firms the tools (e.g., SLAC’s CompJoule tool) to optimize for efficiency first. Given the path-dependence of some designs, such a tool could rapidly accelerate the deployment of more energy efficient EDA into the innovation ecosystem.

Bits

These are the fundamental units of data within electronic systems. The focus is on new materials and devices that increase efficiency of how bits are manipulated and transferred through transistors. Co-design enables the development of transistors that are precisely tuned to software requirements, reducing unnecessary energy expenditure. It also facilitates the

creation of data pathways that are optimized for specific data processing tasks, reducing latency and energy per bit. Additionally, co-design supports the integration of cutting-edge transistor technologies like fin field-effect transistors (FinFETs) and gate-all-around transistors, which offer superior control over electricity flow and significantly minimize leakage currents at smaller scales.

Applications

System energy use is captured by the energy per application metric to perform a particular task. Software-driven co-design of complex applications such as those involving AI or eventually quantum computing is the major opportunity. Note that NLP—now known as large language models—exceed the next closest application by more than 1,000,000 times.

1.2 Scope of the Problem

1.2.1 Scaling Problems and Innovations to Overcome Them

As illustrated by historical data in Figure 3 (Rupp 2022), microelectronics exponential performance improvements or "scaling" of various performance indicators have been fairly flat for decades. Although most began to plateau in the mid-2000—for example, clock speeds at 3 gigahertz—single-thread performance (blue dots) continued to improve although much more slowly (10 times over two decades). Power per die (red dots at 100 watts) seems to have plateaued even earlier in the late 1990s. The fast innovating semiconductor community responded to these varied trends with new innovations such as architectures involving exponentially increasing numbers of logical cores (black dotted trend of Figure 3). Multiple cores and other innovations allowed CPU performance to continue improve at an exponential pace because Moore's Law still holds (orange dots).

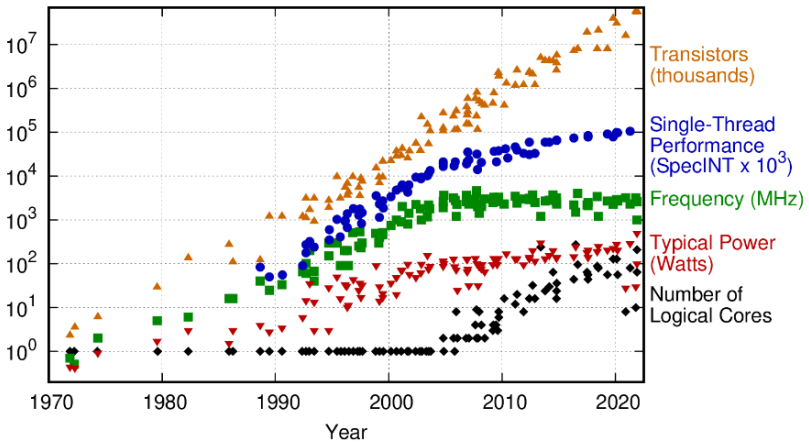


Figure 3. 50 years of microprocessor trend data. *Original data up to the year 2010 collected and plotted by Horowitz et al; new plot and data collected for 2010–2021 by Rupp 2022*

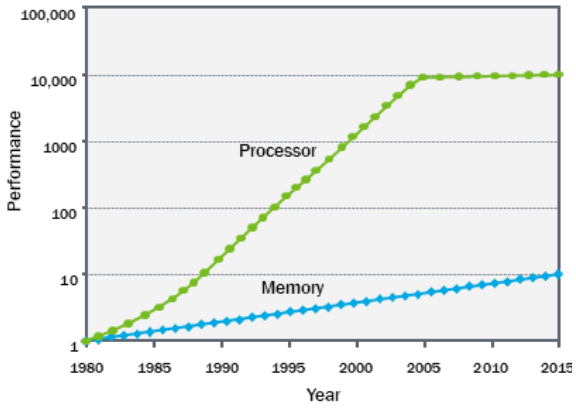


Figure 4. The gap between processor performance and DRAM latency. *Latency is the time between processor memory requests and data return. Source: Hennessey and Patterson 2019*

These developments have effectively ended the era of consistent single-CPU performance enhancement with each new generation of processors, at least under current technological paradigms. The continuation of Moore's Law since allowed the exponential increase in the number of cores per die, spreading computational load (and heat) around the die with active thermal management and clock throttling to keep total dissipated power within thermal design limits.

The Memory Wall Problem

The scaling problems began to cause divergence in the progress curves of memory versus logic. While processor logic speed increased rapidly pre-2005, progress in memory speed (bandwidth and especially latency) lagged, as illustrated in Figure 4. (Hennessey and Patterson 2019). Since 2005, multicore processors have given rise to new complexity in coordinating the data movements of simultaneously executing threads across multiple cores while maintaining memory coherency. This divergence has given rise to the "memory wall" problem, where the speed of data transfer between memory and logic components has become a significant bottleneck, limiting overall system performance. As a result, these disparities in scaling relationships are not only presenting new challenges and driving innovative approaches in the design and operation of microelectronic devices, but they are also exacerbating the problem of energy efficiency in the semiconductor industry.

1.2.2 The S-Curve

The "memory wall" and other limitations within the current scaling paradigms make clear that the next wave of microelectronic innovations will demand interdisciplinary expertise. The rapid advancements that have defined the semiconductor industry are reaching an inflection point, reminiscent of the stages described by the S-curve model in technology adoption (see Figure 5). This model not only reflects the developmental stages of technology, but also signals the evolving demands on the workforce that support it. Initially, industry emphasis is on innovation and high energy consumption, but as technologies progress along the S-curve, the industry must recalibrate to focus on performance optimization, efficiency, and sustainable practices. This transition carries significant implications for workforce development, calling for a comprehensive revamp of traditional educational programs to face the upcoming challenges and opportunities in the industry.

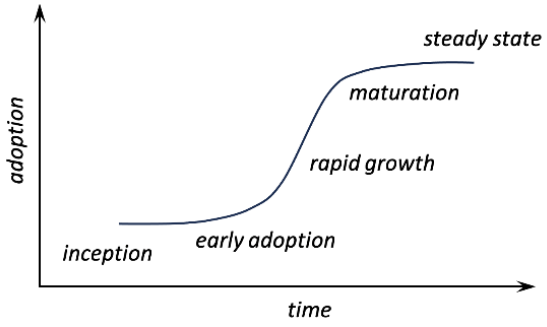


Figure 5. S-curve model

If EES2 is to succeed multiple areas of semiconductor design and manufacture for

EES2 technologies will be on the rapid growth part of S-curve. Educators must work closely with innovators to rapidly reshape, update, and make full-stack codesign a key part of in shaping a curriculum that also integrates multiple disciples—for example, the principles of power efficiency, heat management, and parallel processing architectures, thereby enabling students to become highly innovative workers and innovators in the semiconductor manufacturing and design industries.

1.2.3 Major Sources of Computing Energy Use

Various hardware advancements and their associated energy demands have significantly contributed to the overall energy use in computing. As advanced machine learning AI technologies become increasingly complex, they require not just extensive computational resources but also more specialized hardware. These components, while designed for efficiency in certain tasks, still contribute to the overall energy footprint due to their need for high power to perform trillions of operations per second.

Cryptocurrency mining exemplifies this trend, where the specialized application-specific integrated circuit (ASICs) consume vast amounts of electricity to sustain continuous, intensive computation. These devices, with billions of transistors packed into a single chip, are pushing the limits of energy efficiency in semiconductor technologies.

The hardware underpinning cloud computing infrastructure also plays a significant role in energy usage. Data centers, now equipped with servers featuring high-density chips and advanced 3D heterogeneous integration to manage the massive data processing requirements, have seen an escalation in energy consumption. Innovations such as System on a Chip (SoC) and advanced memory technologies have mitigated some of this increase, but the sheer volume of processing offsets these improvements.

Moreover, the proliferation of Internet of Things (IoT) devices and the rollout of 5G networks add to energy use. While each individual sensor, actuator, or communication module in IoT solutions might consume little energy, the aggregate energy required to support billions of these devices globally is substantial. Additionally, the infrastructure supporting 5G networks, despite being more energy-efficient on a per-bit basis, is expected to increase overall energy consumption due to the sheer increase in data rates and network density.

1.2.4 Estimation of Inefficiencies

A wide-ranging study of energy efficiency in computing and losses compared to fundamental limits has been conducted by Shankar (Shankar and Reuther 2022; Shankar 2023). He surveyed the energy intensity per instruction for the top 500 supercomputers for some widely reported benchmarks as well as for some of the largest-scale applications, including cryptocurrency mining and natural language processing machine learning applications. Shankar then compared those application-level energy measurements to the energy used by lower-level individual machine instructions, biological systems (brains), and the fundamental thermodynamic limit. As shown in Figure 6, energy use varies massively.

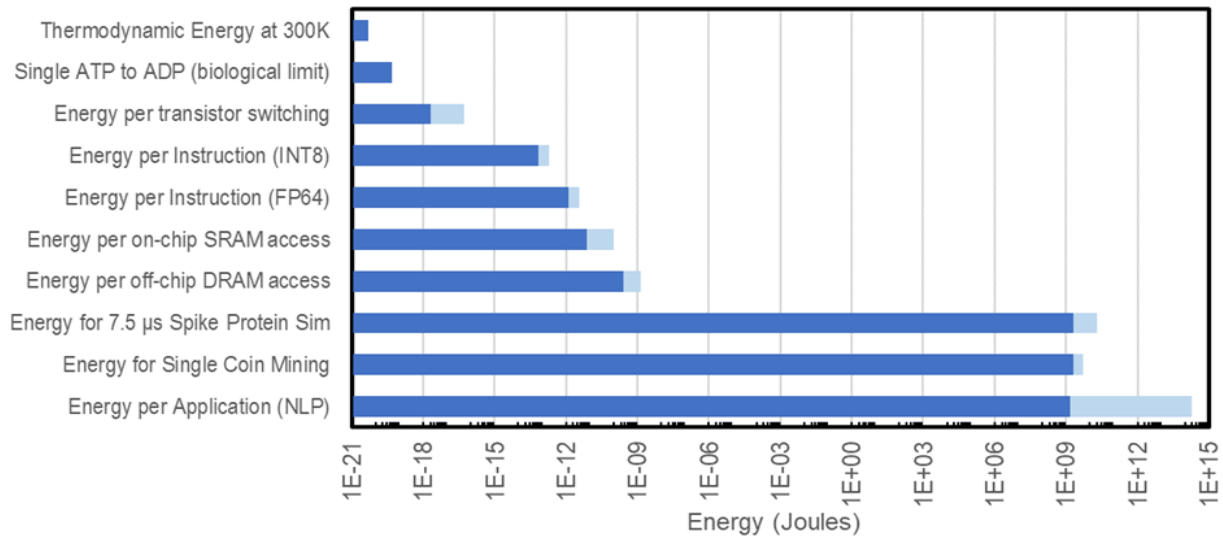


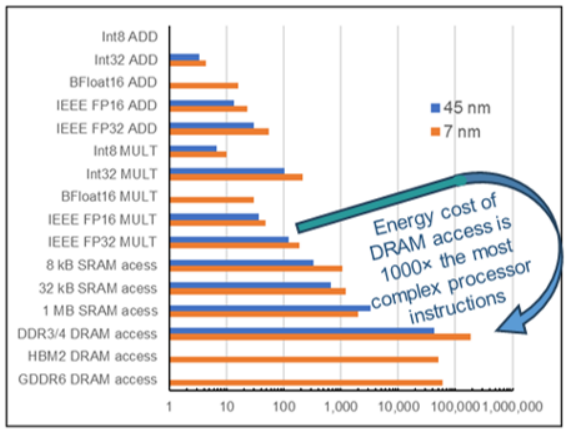
Figure 6. Scale of energy use from bits to applications. *Source: Jouppi et al. 2021 for INT8, FP64, SRAM and DRAM access; Shankar 2023 for the remaining values*

The high energy cost of memory access has been shown by measurements by Horowitz (Han et al. 2016) for a 45-nm process and later updated with a comparison to a 7-nm process (Jouppi et al. 2021) with results as shown in Figure 7.

Figure 7(a) compares the energy cost of the 45-nm and 7-nm processes and shows that for every processor instruction, the energy is reduced for the smaller geometry process. The cost of external DRAM access, however, remains the same. When compared to the energy cost of on-chip instructions in Figure 7(b), the off-chip DRAM access is 185,000 times more energy than the least costly INT8 ADD instruction and about 1,000 times more costly than the most complex compute instructions. Note that in Figure 7(b), the energy cost of operations for each process node is normalized with the energy of an Int8 operation (0.03 pJ for the 45nm node and 0.007 pJ for the 7nm node).

Operation	Energy per operation (pJ) 45 nm	7 nm	45 mm/7 nm Ratio
Int8 ADD	0.03	0.007	4.3
Int32 ADD	0.1	0.03	3.3
BFloat16 ADD	—	0.11	—
IEEE FP16 ADD	0.4	0.16	2.5
IEEE FP32 ADD	0.9	0.38	2.4
Int8 MULT	0.2	0.07	2.9
Int32 MULT	3.1	1.48	2.1
BFloat16 MULT	—	0.21	—
IEEE FP16 MULT	1.1	0.34	3.2
IEEE FP32 MULT	3.7	1.31	2.8
8 kB SRAM acess	10	7.5	1.3
32 kB SRAM acess	20	8.5	2.4
1 MB SRAM acess	100	14	7.1
DDR3/4 DRAM access	1300	1300	1
HBM2 DRAM access	—	250-450	—
GDDR6 DRAM access	—	350-480	—

(a) Energy per operation (pJ)



(b) Energy per operation relative to energy for an INT8 ADD

Figure 7. Energy cost for various operations. *Source: Jouppi et.al, 2021.*

Computation inevitably involves data movement into and out of DRAM and longer-term storage. Data movement is generally much more energetically expensive than the computations, so approaches to reduce the energy cost of data movement—as well as to avoid data movement when possible—are key pathways for energy savings.

At its core, the reason computers use energy is straightforward. Digital circuits are like light switches, turning on and off to represent the 1-second and zero-second in computer language. The energy needed to flip these switches depends on a few things: the electrical pressure (voltage), how much electrical storage capacity there is (capacitance), and how fast the switches are flipping (frequency). Conductors in circuits have inherent “parasitic” capacitance simply due to the presence of charge in neighboring conductors, and this capacitance is proportional to the conductor length. The 8 kilobits (kB), 32 kB, and 100 kB SRAM listed in Figure 7 correspond to the first, second, and third level on-chip cache memories, organized at progressively further distances from the core and therefore with progressively higher latency and energy costs.

This roadmap seeks to comprehensively identify opportunities for reduced energy intensity in all aspects of microelectronics. The twin issues of the high energy cost of memory access and the latency of access are recurring underlying motivations for many topics throughout this roadmap, with numerous approaches for improvement. Fortuitously, solutions to reduce energy consumed by memory access are also solutions to address the key bottleneck in speed, and thus also improve overall compute performance.

1.2.5 Efficiency First

Powerful computational tools and decades of manufacturing knowledge now enable approaches that compare “efficiency first” designs (e.g., using advanced technologies such as those in this roadmap) to conventional designs and achieve EES2 goals without compromising performance. During the course of this roadmap, it became clear that efficiency optimization along another axes (i.e., thermal and mechanical considerations in addition to electronic) was needed to prevent the performance demands of ever tinier next generation microelectronics from

contributing to electricity demand problems. While thermal and mechanical dimensions had previously been considered, including them at the very outset—for example, combining electronic, thermal, and mechanical modelling—still needs to be done.

1.2.6 Energy Use and Sustainability in Semiconductor Manufacturing

Semiconductor manufacturing has become more energy intensive in recent years due to the growing demand for advanced, high-performance chips and the complexities inherent in their production. The manufacturing process, characterized by steps like material deposition, lithography, etching, and polishing, has become more demanding with the introduction of advanced processing technology such as extreme ultraviolet lithography (EUV). This intensification in manufacturing complexity not only escalates energy consumption but also heightens the sustainability concerns associated with semiconductor production.

Key sustainability issues in semiconductor manufacturing include the use of potent greenhouse gases like SF_6, NF_3, and perfluorocarbons in etching processes, and the presence of PFAS materials in standard processing equipment. Additionally, the high demand for ultrapure water, exacerbated by a fivefold increase in water usage over the past decade (Crawford, King, and Wu 2023), poses significant environmental challenges. With many new fabrication facilities located in water-insecure regions like Arizona and northern Taiwan, the industry's water usage is a growing concern.

1.3 Key Concepts for Microelectronic Energy Efficiency

In the EES2 roadmap, co-design emerges as a pivotal R&D strategy essential for catalyzing significant advancements in energy efficiency within the microelectronics sector. This comprehensive approach synthesizes hardware and software design processes from the initial stages, ensuring every component is optimized for minimal energy use while upholding high performance. This strategic integration results in systems that are efficiently tailored to the evolving demands of contemporary technology applications.

1.3.1 Co-Design Process

Co-design is not just a design technique, it is foundational to our strategic approach. Here is an example of a co-design process:

- **Requirement Analysis:** Stakeholders collaboratively define and align on system requirements, establishing clear objectives for performance and efficiency.
- **Concurrent Design:** Teams from two adjacent parts of the stack develop their designs in parallel, enabling real-time adjustments and optimization based on mutual feedback, which ensures that both aspects evolve together seamlessly.
- **Prototyping and Testing:** Early and iterative testing of integrated prototypes allows for quick identification and correction of inefficiencies, ensuring that the final product functions as intended in real-world conditions.
- **Optimization and Refinement:** Continuous refinement based on testing feedback allows for the enhancement of system efficiency and functionality, ensuring that the designs meet the standards set by the roadmap.

This structured approach to co-design directly supports a roadmap's goals by promoting rapid innovation and implementation of energy-efficient technologies, also ensuring that

developments are not only technically feasible but also commercially viable and ready to meet the challenges posed by an energy-intensive technological landscape.

1.4 Organization of the Work

The “compute stack” describes the hierarchy of layers responsible for the development of computational systems, as shown in Figure 8. In order to achieve the benefits of co-design envisioned by DOE in its seminal Basic Research Needs for Microelectronics (DOE Office of Science, 2018) report, co-design for energy efficiency must ensure that adjacent elements of the stack work together. AMMTO’s DOE partners in the Office of Science defined “co-design” in this 2018 report as “where each of the technical abstraction layers in modern computer system design (the compute stack), from fundamental materials research through applications, inform and engage other abstraction layers.” Furthermore, “co-design activities largely occur between adjacent technology abstraction layers (e.g., between materials and devices or computer architects and software designers).” Interdisciplinary co-design is an efficiency imperative. This report reinforces and extends the SC recommendations by providing an order in which co-design needs to be implemented (e.g. efficiency first and for energy intensive applications, major waste heat reduction first). For instance, the development of IBM's NorthPole chip required a holistic approach, integrating breakthroughs across circuits, architecture, and algorithms (Modha et al. 2023). This approach creates electronics that are not only cutting-edge but also sustainable in their energy usage.

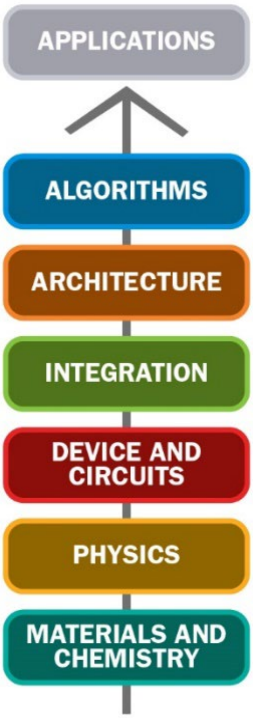


Figure 8. The compute stack.
Source: DOE Office of Science 2018

The EES2 program has effectively divided its scope into eight specialized working groups to enable a comprehensive and collaborative approach to achieve its goal. For a detailed exploration of the microelectronic domain, the initiative has been further partitioned into two categories as depicted in Figure 8, namely: Compute Stack and Microelectronic Enablers. Each working group within the “stack” addresses a layer of the computing stack, with a central focus on energy efficiency. Concurrently, the Enabler category concentrates on enabling technologies, approaches, and workforce, paying specific attention to their manufacturing processes and the energy consumption involved in their computation and operation, including aspects like data center function and energy transport among others.

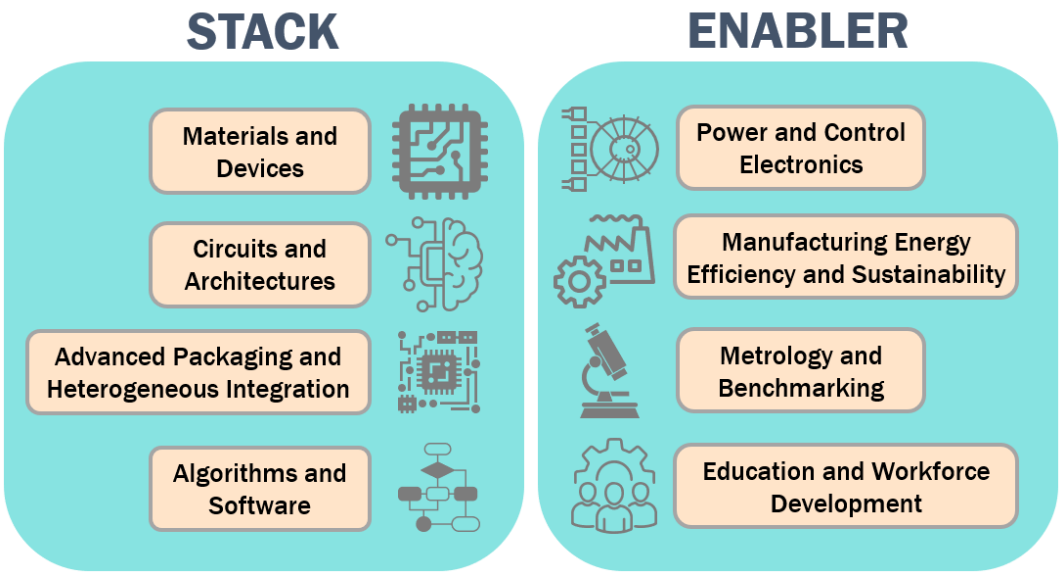


Figure 9. 2022–2023 organization of the EES2 working groups.

The general scope of each WG is described in the following section.

1.4.1 Compute Stack

The **Materials and Devices** group focuses on contemporary challenges in device technology, evaluating innovative materials like carbon nanotubes, and pioneering devices such as spintronic memory. They also address existing issues encompassing current materials, including scalability, contact resistance, and thermal attributes. Given that materials and devices are fundamental to all semiconductor products, key areas of examination include interfaces, interconnects, CMOS compatibility, and novel devices, particularly those leveraging unique switching mechanisms. As Moore's Law decelerates, the discovery of new materials, switching mechanisms, and devices is crucial to meet the efficiency targets of EES2.

The **Circuits and Architectures** group seeks energy efficiency gains in the fundamental building block circuits (transistors, memory cells, etc.) as well as in their organization into an architecture (processors, domain-specific accelerators, high bandwidth memory, etc.). Energy efficiency improvements have primarily come through geometric scaling of the transistor and memory cell. However, since scaling is slowing, this group is focusing on systemic issues, and energy-efficient parallel technologies of processors and memory as well as compute-in-memory technologies for enhancing energy efficiency and performance.

The **Advanced Packaging and Heterogeneous Integration** group emphasizes energy-efficient strategies in integrated circuits and packaging. This is achieved through heterogeneous integration, optical interconnects, and thermal mitigation, utilizing cutting-edge materials and novel packaging techniques. Given that data movement is an energy-intensive operation, this focus provides tangible energy efficiency solutions for semiconductor products. Key areas of consideration include industry standards, optimized thermal management, chiplet-based

integration, reduction of critical dimensions to silicon fabrication sizes, and innovation in interconnects and input/output systems.

The **Algorithms and Software** group focuses on optimizing energy efficiency of microelectronics through strategic utilization of algorithmic design and software development. Examples include bio-inspired/neuromorphic algorithms and algorithmic improvements coupled to accelerator hardware. The group's aim is to champion the energy efficiency goal within software without compromising computational operations.

1.4.2 Microelectronic Enablers

The **Power and Control Electronics** group focuses on enhancement and innovation of power delivery systems spanning from microelectronics to large data centers. Their concentration lies in exploring economically viable and efficient design solutions, which range from the implementation of wide-bandgap devices in switching power supplies to formulating strategies for optimizing renewable energy use and reducing carbon in energy supplies to data centers to improved thermal management strategies for lower overall energy consumption of the data center(s). Understanding that efficient power delivery and control are critical for information-communication technologies, the group acknowledges that managing where, when, and how power is delivered to devices can minimize energy consumption and is integral to handling large-scale renewable resources and electric transport. Key areas of interest include power management, thermal mitigation technologies, pioneering devices, power leakage, and power conditioning circuits and components.

The **Manufacturing Energy Efficiency and Sustainability** group focuses on optimization of energy efficiency and promotion of sustainability in the manufacturing process, especially in response to the rising wave of manufacturing facilities spurred by the CHIPS and Science Act. The Act's core objective is to repatriate manufacturing of microelectronics. As discussed above, manufacturing-related energy usage escalates with each new iteration of advanced semiconductor technology. In response to this, key areas of interest include alternatives for energy-intensive extreme ultraviolet lithography, lower greenhouse gas-emitting dry etch gases, and the implementation of sustainable manufacturing practices.

The **Metrology and Benchmarking** group identifies measurement, characterization, and benchmarking needs for the technologies discussed in other working groups. Recognizing the intricacy of burgeoning integrated circuits and microelectronic systems, this focus is essential for the identification of pioneering metrology technologies and strategies for future systems. As semiconductor products increase in complexity, the importance of metrology in understanding process variability, device function, and troubleshooting amplifies. Benchmarking becomes crucial in defining appropriate metrics for energy efficiency and in comparing extant and emerging technologies. Areas of interest include innovative metrology methods and tools, suitable metrics for each level of the microelectronic stack, and standards stipulated by NIST, inclusive of those provided for energy efficiency by DOE.

The **Education and Workforce Development** group develops strategies to ensure a well-qualified workforce is available not only to support the growth of the domestic microelectronics

industry, but also to lead global innovation in microelectronics technology, especially as it pertains to energy efficiency in microelectronics and computational systems.

1.4.3 Cross-Collaboration

Although the work was organized by the groups listed above, many of the ideas documented in the roadmap are inherently interdependent and require input from multiple working groups. For example, the most impactful implementation of carbon nanotube field-effect transistors (CNTFETs) is in monolithic 3D integration with emerging memory elements and circuit architecture. This requires direct innovations from the Materials and Devices, Circuits and Architectures, Advanced Packaging and Heterogeneous Integration, Metrology and Benchmarking, Manufacturing, and Algorithms and Software working groups. In this way, demonstration and implementation of the ideas from each working group may require innovations from multiple working groups simultaneously. To address this interdependency, team members had opportunities throughout the roadmap development process to discuss cross-cutting opportunities, seek input from other working groups, and coordinate results in each technologies action plan.

1.5 Methodology

This roadmap is a product of extensive literature review and energy analysis, nine working group collaboration meetings, and expert input during the writing process. The working groups met monthly, with the organizing committee engaging in literature review and analysis to prepare for the following meeting. Table 2 and the description below provide a general overview of the roadmapping process. Some working groups may have differed through the series of meetings based on progress and the nature of the topics within each group.

The roadmap formally launched in September 2022 with a pair of meetings—the first to introduce the EES2 pledge and inaugurate its first 20 signers, and the second to discuss energy efficiency considerations of microelectronic devices and identify key technological advancements needed to achieve the EES2 goal. The input gathered during these events along with post-meeting literature review established the working group topics, which were formalized in a meeting held in November 2022. Those that participated in the meeting identified the working groups to which they were interested in contributing.

Starting in January 2023, each working group met monthly to explore various aspects related to the working group topic. January 2023 featured the announcement of a detailed roadmap schedule and strategy, and participant discussions served to determine the general scope of each working group. Working group members were nominated by the pledging organizations and self-selected the working groups supported according to expertise and interest. Some more active members were invited by DOE or volunteered to act as co-chairs of the working groups.

In February 2023, the third meeting built upon the previous by centering around specific energy efficiency technologies. In many cases, the working group proposed more technologies than the working group could effectively discuss and characterize throughout the rest of the working group meetings, so the group prioritized what they thought were the most promising technologies. The number of technologies that were deprioritized depended on the size of the working group. Working group members were also asked to estimate projected energy efficiency contribution and timeline for achievement of that contribution. These estimates were

only meant to gauge an approximate number on the potential efficiency improvement and in most cases, highlighted the need for deeper analysis.

The fourth meeting functioned as an intensive session for the working group to continue the discussions from the previous month. Working group members broke up into small groups to research and analyze existing literature and data to refine initial energy efficiency estimates for the down-selected technologies. The need for an additional working group focused on workforce development was also discussed.

The fifth meeting, held in April 2023, was dedicated to identifying challenges that may arise in the development and implementation of the proposed technologies solutions. A secondary purpose of the fifth meeting was the official establishment of the newly founded workforce development working group.

In May 2023, the sixth meeting focused on a discussion of R&D solution pathways for the challenges that were identified in the previous meeting, and the seventh meeting in June reviewed and discussed the input collected from the previous meetings and made any adjustments necessary. If time allowed, working groups also started on action plans.

Held in July 2023, the eighth meeting was dedicated to developing an action plan for each technology or to address key challenges. Working group members, once again, broke up into small groups to flesh these out. Groups continued to collaborate offline after the eighth meeting to continue to make progress prior to the ninth and final meeting, which was held in August. This meeting, the small groups finished up their action plans and presented them for feedback from the broader group.

Writing of the roadmap began in September 2023, with working group facilitators drafting sections pertaining to their respective groups. Drafts were distributed to working group chairs and participants for comment, and support staff were tasked with drafting introductory and overview sections of the roadmap.

Table 2. Workshop Series Used to Establish the Targeted Technologies and Associated Solution Pathways and Action Plans for thie Roadmap

WG Meeting	Timing	Topic(s) of Discussion
1	November 2022	Working group topics and membership
2	January 2023	Working group charters and processes
3	February 2023	Key energy-efficient technologies, prioritization, and efficiency estimates
4	March 2023	State of the art, baseline energy consumption, and future projections
5	April 2023	Efficiency improvement challenges
6	May 2023	Pathways for advancement
7	June 2023	Review and refinement
8	July 2023	Action planning
9	August 2023	Action planning, review, and refinement

The technologies in this roadmap are assessed against two metrics, timeline to maturity and impact. Timeline to maturity corresponds to the time required to achieve a technology readiness level (TRL) of 6. Tailored definitions of TRLs for the microelectronics industry are detailed in Figure 10. Technologies already at TRL 6 are included for their potential energy efficiency improvements, despite not being incumbent technologies. Impact is measured by comparing future performance in an energy metric against current technology (e.g., energy per bit, energy per switching, memory access, etc.). While true impact from the technologies contained in the roadmap will be dependent on commercialization and deployment, non-technical and market forces play an outsize role in determining this timeline. Therefore, the roadmap does not attempt to estimate when this will occur, nor which technologies should be addressed next. Instead, the roadmap compiles promising energy-efficient technologies and approaches and highlights the technical challenges and potential solution pathways to achieve technical readiness for commercialization, if so desired by industry.

Generic TRL Guidelines		Microelectronics-Specific TRL Guidelines
IDEA Unproven concept, no testing hos been performed	0	IDEA Unproven concept, no testing has been performed
BASIC RESEARCH You can now describe the need(s) but have no evidence	1	BASIC RESEARCH Studied theoretically or experimentally with general ideas for use
TECHNOLOGY FORMULATION Concept and application hove been formulated	2	TECHNOLOGY FORMULATION Concept/application defined, limited experimental confirmation
NEEDS You have an initial 'offering, stakeholders Clke your slideware	3	PROMISING TECHNOLOGY CANDIDATE Promising but not yet demonstrated in a functional system
SMALL SCALE PROTOTYPE Built in a laboratory environment	4	TECHNOLOGY CANDIDATE FOR FUTURE NODES Demonstrated in research labs but too immature for next nodes
LARGE SCALE PROTOTYPE Tested in intended environment	5	TECHNOLOGY CANDIDATE FOR NEXT NODES Included in research fab line for the upcoming node
PROTOTYPE SYSTEM Tested in intended environment close to expected performance	6	THE NEXT NODE BEYOND CURRENT PRODUCTION PLAN Meets performance requirements for the next production node
DEMONSTRATION SYSTEM Operating in operational environment at precommercial scole	7	THE UPCOMING STATE OF THE ART PRODUCTION NODE Vetted in operational environment, ramping towards HVM
FIRST OF A KIND COMMERCIAL SYSTEM All technical processes to support commercial activity in place	8	THE CURRENT STATE OF THE ART PRODUCTION NODE The latest node in HVM, supplying customers
FULL COMMERCIAL APPLICATION Technology 'general availability' for all customers	9	PREVIOUS PRODUCTION NODES The previous node and all preceding nodes, highly vetted

Figure 10. Definitions for technology readiness levels for the microelectronics industry as used in this report

1.6 Related Work

Roadmapping has a long tradition in the semiconductor industry, with industry-led groups coordinating on shared efforts to pursue early-stage R&D, standardize equipment, and reduce capital expenditures while propelling the technology forward in keeping with Moore's Law. There are several important roadmapping activities being undertaken today to facilitate technological progress in the post-Dennard scaling era. The following subsections list and describe the scope of the most prominent roadmaps.

1.6.1 International Roadmap for Devices and Systems

The *International Roadmap for Devices and Systems* (IRDS) (IRDS 2022), the most long-standing roadmap in the industry, evolved from the predecessor International Technology Roadmap for Semiconductors (ITRS). Even earlier, a 1965 paper by Gordon Moore laid out the observation known as Moore's Law (Moore 1965). Moore's paper established a tempo of technology advancement for the semiconductor industry, but it was not until 1991 that a formal roadmap document (the ITRS) was developed by the U.S. semiconductor industry community. From this beginning, in keeping with the expansion of the industry globally, the roadmap grew into an international effort. The ITRS was updated annually through 2015 but was then

supplanted by the IRDS, which had a broader scope that encompassed electronic devices and systems. The intent of the IRDS roadmap is to provide a basis to facilitate cooperation by academic, manufacturing, supply, and research organizations, specifically:

- To identify key trends related to devices, systems, and all related technologies by generating a roadmap with a 15-year horizon.
- To determine generic devices' and systems' needs, challenges, potential solutions, and opportunities for innovation.
- To encourage related activities worldwide through collaborative events, such as related IEEE conferences and roadmap workshops.

1.6.2 Heterogeneous Integration Roadmap

The *Heterogeneous Integration Roadmap* (HIR) (HIR 2022) is a collaborative effort between several IEEE technical societies—the IEEE Electronics Packaging Society (EPS), the IEEE Electron Devices Society (EDS), and the IEEE Photonics Society—as well as the industry group SEMI and the ASME Electronic and Photonic Packaging Division (EPPD). Like the IRDS (from which it is an outgrowth), the HIR provides guidance for the global electronics industry regarding projected technology capabilities, needs, and opportunities. The HIR provides:

- A forecast of industry requirements to maintain the pace of progress for the industry and user community over a 15-to-25-year horizon.
- Identification of difficult challenges that must be addressed to meet these industry requirements, with identified research needs and potential solutions.

1.6.3 2030 Decadal Plan for Semiconductors

Published in January 2021, the *2030 Decadal Plan for Semiconductors* by the Semiconductor Research Corporation (SRC 2021) was instrumental in motivating the work that has resulted in this EES2 roadmap. The decadal plan outlined key research priorities for the semiconductor and computer industries. It followed a June 2020 report by the Semiconductor Industry Association (SIA) calling for a 3-fold increase in federal investment in semiconductor R&D to stimulate U.S. economic growth and job creation, complementing it with specific goals and quantitative targets (SIA 2020). The decadal plan identified five seismic shifts that will influence the industry:

- Analog hardware will enable machine intelligence systems.
- Growing demand for memory will outstrip global supply, creating opportunities for new memory and storage solutions.
- Growing demand for communication capacity to keep up with data generation rates will drive communication technology development.
- Emerging security challenges in highly interconnected systems and in AI systems will drive security technology development.
- Ever-rising energy demands for computing will necessitate new computing paradigms with dramatically improved energy efficiency.

1.6.4 Microelectronic and Advanced Packaging Technologies Roadmap

The ongoing microelectronic and advanced packaging technologies (MAPT) roadmap (MAPT 2023) effort is led by SRC as an expansion of the 2030 decadal plan. MAPT is a multidisciplinary strategy addressing advanced packaging, 3D integration, EDA, nanoscale manufacturing, new materials, and energy-efficient computing, with the aim of assuring future design, development, and manufacturing of heterogeneously integrated chips in the U.S. and like-minded nations. The MAPT Roadmap outlines research priorities and challenges that must be addressed to ensure sustainable growth and innovation, and focuses explicitly on energy sustainability, environmental sustainability, and workforce sustainability.

1.6.5 National Strategy on Microelectronics Research

The White House Office of Science and Technology Policy (OSTP) and its National Science and Technology Council (NSTC) created—as required by the first CHIPS authorization in 2021—a Subcommittee on Microelectronics Leadership (SML) that was tasked with providing a National Strategy on Microelectronics Research (NSTC 2024). The Office of Science represented DOE in the strategy development effort. The National Strategy identified the following goals to guide agency efforts in microelectronics research:

- Enable and accelerate research advances for future generations of microelectronics.
- Support, build, and bridge microelectronics Infrastructure from research to manufacturing.
- Grow and sustain the technical workforce for the microelectronics R&D to manufacturing ecosystem.
- Create a vibrant microelectronics innovation ecosystem to accelerate the transition of R&D to the U.S. industry.

In the national strategy, improving energy efficiency was mentioned as being “increasingly essential for sustainability” and as an important research focus in numerous areas. AMMTO’s Semiconductor R&D for Energy Efficiency workshop series was also referenced in the national strategy.

1.6.6 How This Complements Prior Roadmaps and Strategies

The EES2 roadmap complements the roadmaps and strategies listed above. Whereas the prior reports encourage energy efficiency qualitatively, the EES2 roadmap has quantitative goals for energy efficiency. It also uses a common factor—the energy efficiency improvement factor—to compare technologies, as well as the three specific energy metrics: energy per bit, instruction, and application.

With several EES2 pledging institutions and individual WG participants also involved in other roadmap efforts, cross-pollination is another complementary area. For example, SRC, the first EES2 pledger, provided early and valuable input to the EES2 roadmap process, and the EES2 team likewise has provided input to the MAPT effort. Most of the WG members are also active in at least one other collaborative microelectronics innovation ecosystem effort, such as standards-setting committees, technical societies and councils, and community-organized conferences, thus bringing a wider perspective and greater industry connectivity to the team.

1.7 Introduction References

Crawford, Alan, Ian King, and Debby Wu. 2023. “The Chip Industry has a Problem with Its Giant Carbon Footprint.” Bloomberg. Published April 8, 2023. https://www.bloomberg.com/news/articles/2021-04-08/the-chip-industry-has-a-problem-with-its-giant-carbon-footprint.

Han, Song, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A. Horowitz, and William J. Dally. 2016. “EIE: efficient inference engine on compressed deep neural network.” *ACM SIGARCH Computer Architecture News*. Vol. 44 (Issue 3): pg 243–254. https://doi.org/10.1145/3007787.3001163.

Hennessey, John, and David Patterson. 2019. *Computer Architecture: A Quantitative Approach*. Burlington, MA: Morgan Kaufmann Publishers.

HIR. 2022. *Heterogeneous Integration Roadmap 2021 Edition*. Institute of Electrical and Electronics Engineers (IEEE). https://eps.ieee.org/technology/heterogeneous-integration-roadmap.html.

IRDS. 2022. *International Roadmap for Devices and Systems*. IEEE. https://irds.ieee.org/.

Jouppi, Norman P., et al. 2021. “Ten Lessons from Three Generations Shaped Google’s TPUv4i: Industrial Product.” Presented at ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA). Valencia, Spain. https://doi.org/10.1109/ISCA52012.2021.00010.

Kamiya, George, and Oskar Kvarnström. 2019. “Data centres and energy – from global headlines to local headaches?” International Energy Agency (IEA). Published December 20, 2019. https://www.iea.org/commentaries/data-centres-and-energy-from-global-headlines-to-local-headaches.

Kurzweil, Ray. 2005. The Singularity Is Near: When Humans Transcend Biology. New York: Penguin Books.

Mann, Margaret, and Vicky Putsche. 2022. “Semiconductor: Supply Chain Deep Dive Assessment.” U.S. Department of Energy Response to Executive Order 14017, “America’s Supply Chains.” Published February 24, 2022. https://doi.org/10.2172/1871585.

MAPT. 2023. *Microelectronics and Advanced Packaging Technologies (MAPT) Roadmap*. https://srcmapt.org/.

Masanet, E., A. Shehabi, N. Lei, S. Smith, and J. Koomey. 2020. “Recalibrating global data center energy-use estimates.” *Science*. Vol. 367 (Issue 6481): pg 984–986. https://doi.org/10.1126/science.aba3758.

McKie, Robin, and James Tapper. 2022. “Chaos after heat crashes computers at leading London hospitals.” *The Guardian*. Published August 7, 2022. https://www.theguardian.com/environment/2022/aug/07/chaos-after-heat-crashes-computers-at-leading-london-hospitals.

Modha, Dharmendra S., et al. 2023. “Neural inference at the frontier of energy, space, and time.” *Science*. Vol. 382 (Issue 6668): pg 329–335. https://doi.org/10.1126/science.adh1174.

Moore, Gordon. 1965. “Cramming more components onto integrated circuits.” *Electronics*. Vol. 38 (Number 8). Published April 19, 1965.

National Science and Technology Council. 2024. “National Strategy on Microelectronics Research.” Executive Office of the President of the United States. Published March 2024. https://www.whitehouse.gov/wp-content/uploads/2024/03/National-Strategy-on-Microelectronics-Research-March-2024.pdf.

Rupp, Karl. 2022. “Microprocessor Trend Data.” https://github.com/karlrupp/microprocessor-trend-data.

Shankar, Sadasivan, and Albert Reuther. 2022. “Trends in Energy Estimates for Computing in AI/Machine Learning Accelerators, Supercomputers, and Compute-Intensive Applications.” Presented at 2022 IEEE High Performance Extreme Computing Conference (HPEC). Waltham, MA. https://doi.org/10.1109/HPEC55821.2022.9926296.

Shankar, Sadasivan. 2023. “Energy Estimates Across Layers of Computing: From Devices to Large-Scale Applications in Machine Learning for Natural Language Processing, Scientific Computing, and Cryptocurrency Mining.” Presented at 2023 IEEE High Performance Extreme Computing Conference (HPEC). http://dx.doi.org/10.1109/HPEC58863.2023.10363573.

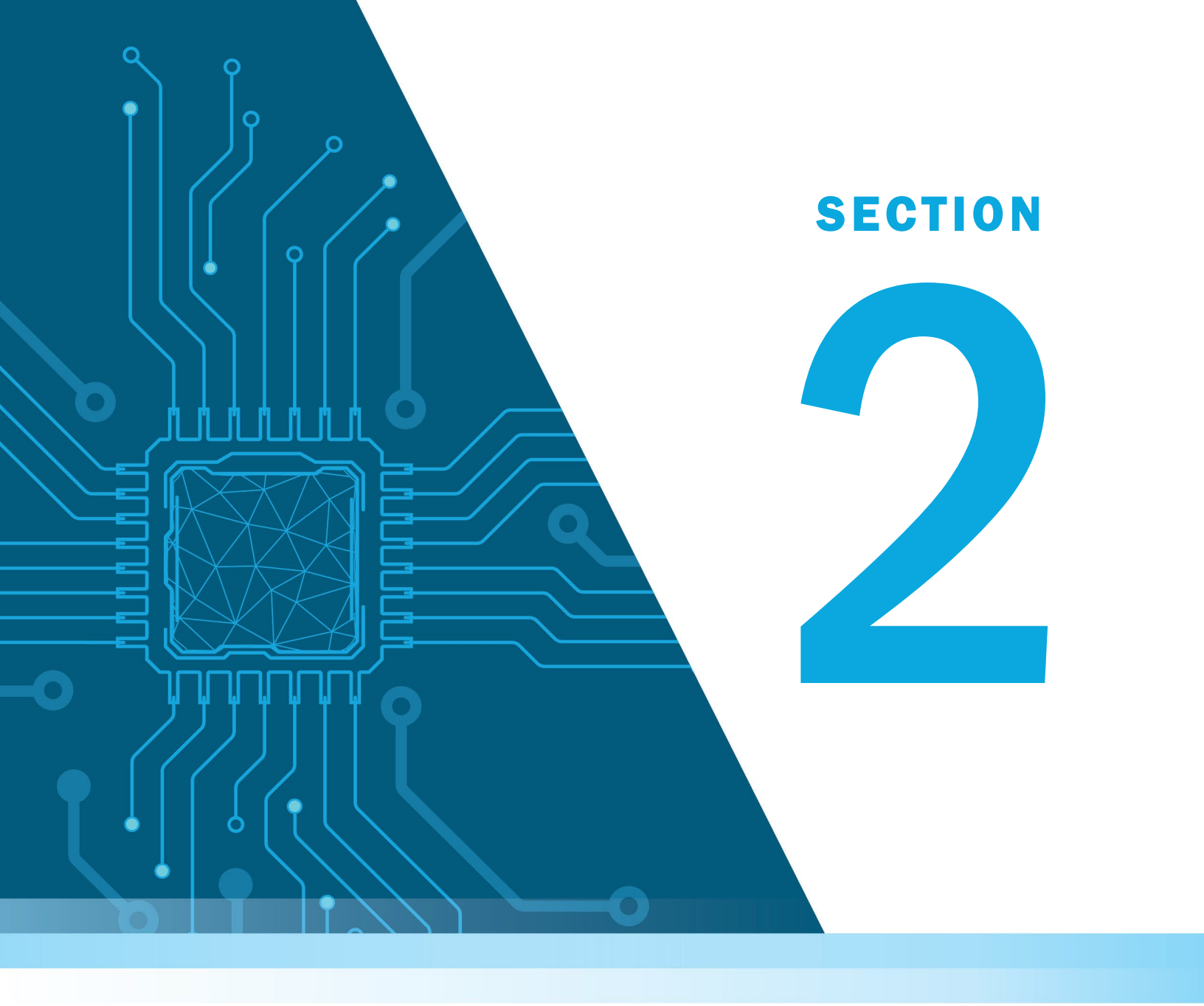
SIA. 2020. “State of the U.S. Semiconductor Industry.” Semiconductor Industry Association (SIA). https://www.semiconductors.org/wp-content/uploads/2020/07/2020-SIA-State-of-the-Industry-Report-FINAL-1.pdf.

SRC. 2021. *2030 Decadal Plan for Semiconductors*. Semiconductor Research Corporation (SRC). https://www.src.org/about/decadal-plan/.

Statista. "Primary Energy Consumption Worldwide from 2000 to 2023 (in exajoules)." Accessed July 24, 2024. https://www.statista.com/statistics/265598/consumption-of-primary-energy-worldwide/.

The White House. 2022. “FACT SHEET: CHIPS and Science Act Will Lower Costs, Create Jobs, Strengthen Supply Chains, and Counter China.” Published on August 9, 2022. https://www.whitehouse.gov/briefing-room/statements-releases/2022/08/09/fact-sheet-chips-and-science-act-will-lower-costs-create-jobs-strengthen-supply-chains-and-counter-china/.

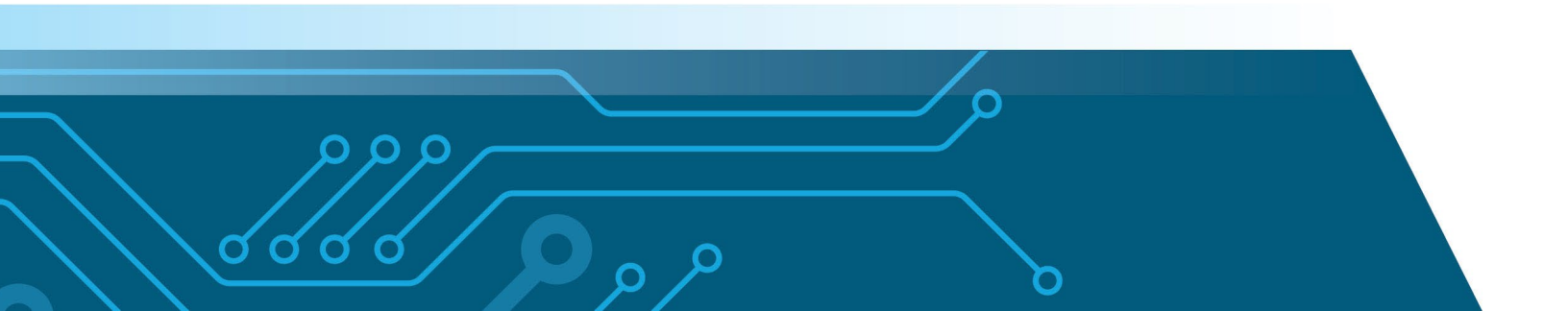
York, Richard, and Julius Alexander McGee. 2016. “Understanding the Jevons Paradox.” *Environmental Sociology*. Vol. 2 (Issue 1): pg 77–87. https://doi.org/10.1080/23251042.2015.1106060.



SECTION

2

Technologies for the Compute Stack



2 Technologies for the Compute Stack

This chapter discusses the hierarchical layers of the microelectronics compute stack and highlights the energy efficiency potential for the technologies within each. The sub-sections are a summary of the input gathered from working group deliberations over the roadmapping period. Each sub section is first framed within the context of microelectronics and makes connections to the broader EES2 goal. This sets the stage for an in-depth analysis of each technological area.

Key technological domains are identified, with their functionality thoroughly explored alongside potential improvement strategies. Precise metrics and projected timelines are provided for each technology's path towards maturity and deployment with emphasis on the future initiatives needed to achieve these outlined objectives.

2.1 Materials and Devices

In 2024, the metal-oxide-semiconductor field-effect transistor (MOSFET) is the foundation for logic and memory devices, serving as the backbone of traditional computing technologies. The complementary metal-oxide-semiconductor (CMOS) process, which pairs complementary and symmetrical MOSFETs, has been the standard implementation of MOSFETs for decades. As CMOS technologies progressed below 10-nm, short channel effects, which result in high standby power consumption and low drive current, became the dominant factor impeding continued scaling (Lee et al. 2015). These short channel effects led to intense interest in alternative device technologies in parallel with CMOS scaling. Examples of alternative device technologies include novel channel materials, device architectures, and switching mechanisms.

To overcome the limitations of Si CMOS scaling, there are generally two approaches: CMOS-extension and Beyond-CMOS (or CMOS-replacement) technologies (Hiramoto 2009; see Figure 11). CMOS-extension technologies utilize thermionic emission to switch charge states using either innovative materials or device architectures. In contrast, Beyond-CMOS strategies employ unconventional transport mechanisms, such as tunneling, or rely on effective carriers distinct from charge, such as spin.

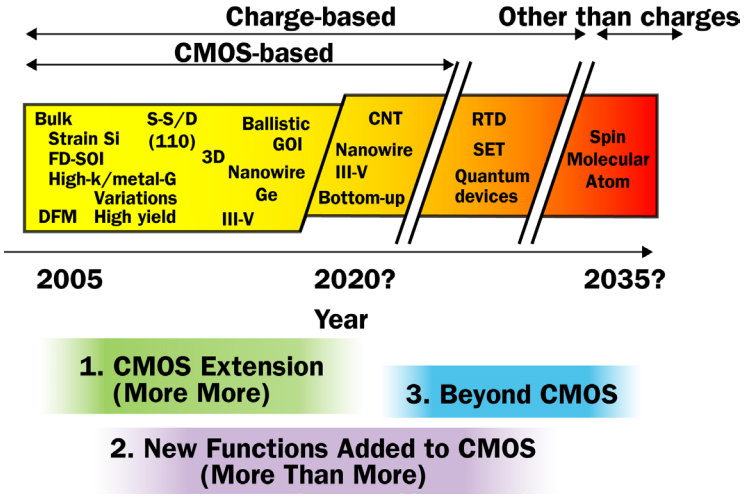


Figure 11. Technology options for new information processing technologies. *Source: Hiramoto 2009*

Innovative materials technologies—such as 2D materials, carbon nanotubes, spintronics, and ferroelectrics—primarily impact the bit level. These materials serve as the foundation for developing advanced transistors, including traditional FETs, TFETs, and Si-GAA transistors, which are crucial for enhancing energy efficiency and performance at the bit level. By fundamentally improving the properties of transistors, these materials play a pivotal

role in the efficient manipulation and storage of bits, laying the groundwork for more energy-efficient and high-performance computing architectures.

Working group methodology

The working group focused on addressing the contemporary challenges in device technology (Si scaling) and evaluating innovative materials and emerging devices (see Table 3). Challenges associated with these alternatives—including scalability, contact resistance, interfaces, and CMOS compatibility—were also addressed.

Table 3. Promising Energy-Efficient Materials and Device Technologies.

Technology
2D materials
CNTFET
CNT memory
Spintronic/magnetoelectric logic
Spintronic memory
Ferroelectric memory
Tunnel field effect transistor
Silicon gate-all-around
Analog devices for neuromorphic
Novel interconnects and contacts
Novel interlayer dielectrics

Figure 12 summarizes the potential energy efficiency improvement factor and timeline for demonstration of the prioritized technologies. The y-axis represents the potential energy efficiency improvement factor, which is quantified based on the energy savings achieved when transitioning from the incumbent technology to the alternative in logarithmic scale. The x-axis, on the other hand, denotes the years it takes for this specific technology to reach TRL 6. For more information on TRL6, refer to section 1.5. The references for each technology are included in the detailed write-ups that can be found in the following sections.

A systematic benchmarking effort is needed to objectively compare the technologies proposed in this chapter. Furthermore, as the lowest rung on the compute stack, system-level efficiency impacts from innovations stemming from this working group must be carefully evaluated.

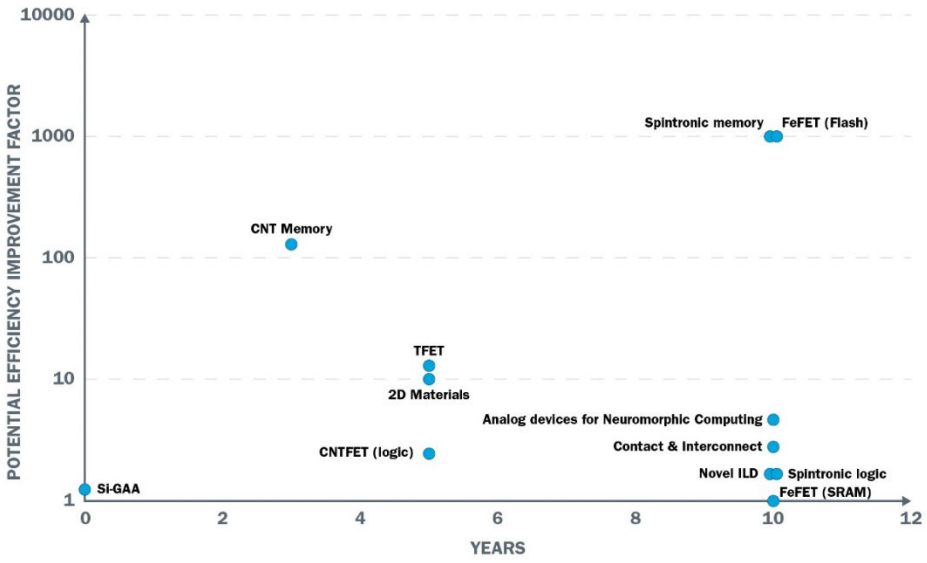
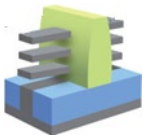
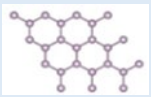
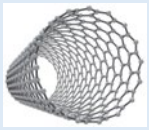
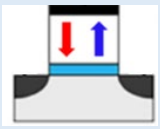
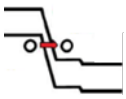
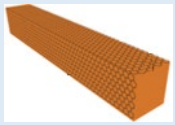


Figure 12. Potential efficiency improvement factor and timeline for selected technologies proposed by the Materials and Devices working group

Key Takeaways

Table 4 summarizes the most significant identified energy efficiency opportunities that can be achieved through advances in materials and devices.

Table 4. Key Takeaways for Energy Efficiency Opportunities in Materials and Devices

Technology Group	Key Opportunities for Energy Efficiency
CMOS-Extension	 • **Si-GAA transistors** offer energy efficiency gains along with other performance improvements, including faster switching speeds and reduced channel leakage. It is currently on the path to being fully realized and integrated into systems, replacing FinFETs, by 2025.
Beyond-CMOS: Conventional Carrier and Transport	   • **2D Materials**, such as TMDCs, are atom-thick layers with unique properties like high electron mobility and thermal conductivity. 2DFETs can offer energy efficiency improvement over traditional silicon-based FETs by reducing total capacitance and operating voltage, as well as reducing device density due to their thin-layered structure. • **CNTs** have unique material properties such as high electron mobility and ultrathin 1D structure. CNT-based devices, especially CNT memory, also exhibit outstanding current density and minimal parasitic effects. Their high carrier mobility and near-ballistic carrier transport allow CNTFETs to mitigate short channel effects present in silicon MOSFETs, significantly enhancing computational energy efficiency. • **Ferroelectric FET (FeFET)** enable non-volatile memory with lower write energy. Hafnia-based FeFETs stand out for their nondestructive read, fast switching, scalability, and potential for multibit operation, offering significant advantages over traditional memory technologies.
Beyond-CMOS: Alternative Carriers and Transport	   • **Spintronic devices** utilize the electron's charge and spin to enable low power consumption and high endurance electronic circuits with the added advantage of non-volatility, offering competitive performance for both logic and memory applications. • **Tunnel FETs (TFET)** leverage quantum tunneling for carrier transport, enabling steeper subthreshold slopes and lower operating voltages. • **Emerging devices and materials for analog computing**, such as memristors, have the potential to transform computational methods. These devices leverage a variety of materials, including organic materials that mimic neuroplasticity and mixed ion-electron conductors that facilitate brain-like processing.
Device Integration Materials	  • Integration of novel materials for interconnect and ILD can reduce resistive loss and capacitive delay, leading to improvements in energy efficiency, performance enhancement, and higher device density. • **ILDs** with lower κ-values are essential to minimize crosstalk and delay time. Innovative materials with structural, thermal, and chemical integrity, along with mechanical hardness and minimal leakage, are needed. • There is a concerted push towards **interconnect and contact** metals, such as ruthenium, which have lower mean free paths and are less affected by boundary scattering. Research is also looking into barrierless alternatives and novel contact materials to overcome metal-induced gap states and reduce contact resistivity.

Grand challenges

The major challenges for achieving energy efficiency gains in microelectronics materials and devices include:

- Achieving consistent manufacture of high-purity and high-quality materials—such as 2D materials, carbon nanotubes (CNTs), and magnetic materials—for energy-efficient microelectronic devices.
- Identifying and developing processing methods to enable integration of new materials.
- Benchmarking emerging devices and material technologies against a consistent set of metrics and test protocols.
- Establishing R&D testbeds or prototyping facilities to demonstrate emerging device concepts and materials.
- Evaluating fundamental and interfacial properties (thermal stability, conductivity, contact resistance, etc.) of emerging materials and heterostructures and understanding their implications on device behavior.
- Bridging the knowledge gap between material science and device engineering through cross-disciplinary collaboration among material scientists, device engineers, and system architects to foster holistic understanding and create highly efficient, scalable, and reliable devices.
- Developing and leveraging high-fidelity device modeling and system simulation platforms that consider nuanced behavior of materials to accelerate R&D.

2.1.1 2D Semiconductor Materials

Two-dimensional (2D) semiconductor materials generally consist of one-to-three atom-thick layers, forming covalently bonded lattice structures. Many 3D semiconductor materials include surfaces with dangling bonds, which are unbonded atoms at the surface that can create reactive sites. In contrast, 2D semiconductor materials generally have saturated bonding configurations with minimal dangling bonds. This characteristic can contribute to their unique electronic, mechanical, and optical properties (Allain et al. 2015). The absence of dangling bonds in 2D semiconductor materials ensures that their surfaces are smooth and uniform, which minimizes electron scattering and enhances electrical conductivity. By reducing reactive sites that could otherwise trap electrons or degrade the material's properties over time, the absence of dangling bonds in 2D semiconductor materials is critical for the development of high-performance energy-efficient devices as it allows for faster electron transport and improved device reliability. Furthermore, 2D materials are adept at operating at lower voltage levels without compromising on speed, offering a promising avenue for reducing energy use in computing and electronic

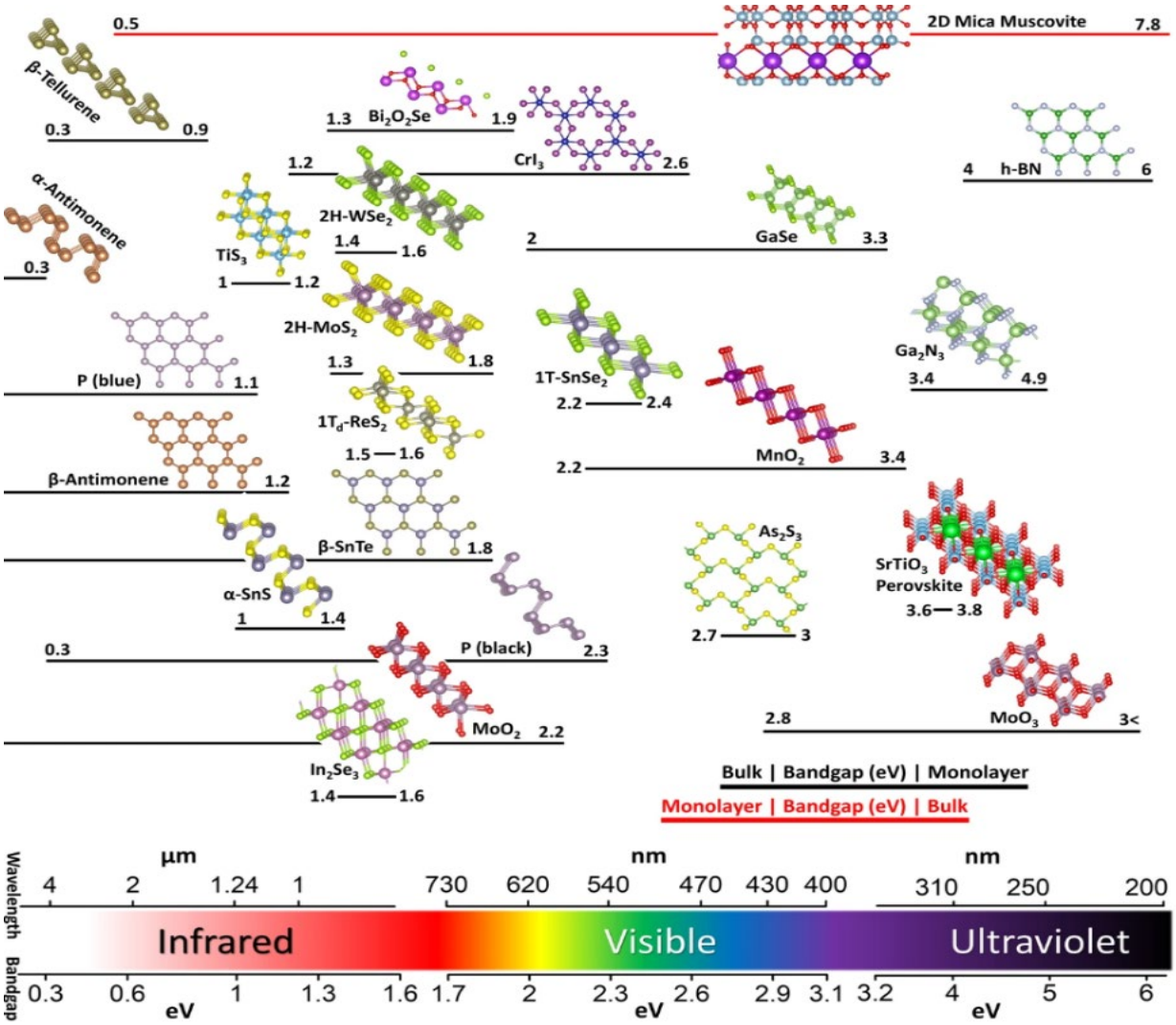


Figure 13. Selected 2D materials and their bandgap. *Source: Chaves et al. 2020*

devices. For example, graphene, a single layer of carbon atoms arranged in a hexagonal lattice, exhibits extraordinary properties, including exceptional electron mobility, mechanical strength, and thermal conductivity. The discovery of the benefits of graphene spurred interest in a whole class of monolayer 2D materials, each with its own unique properties and potential applications.

Although graphene exhibits excellent properties suitable for transistors, it lacks an innate bandgap. To introduce a bandgap, graphene must be fashioned into nanoribbons. However, this modification can lead to complications stemming from edge scattering effects and a significant decrease in carrier mobility, eliminating graphene as a candidate for FET and limiting it to use in interconnects and applications where switching is not the primary concern, such as thermal management (Lin et al. 2010). Thus, there has been a shift in focus to other 2D semiconductor materials, predominantly transition metal dichalcogenides (TMDCs) and hexagonal boron nitride (h-BN) with innate bandgap ranging from 0.3 to 6 eV (see Figure 13), which is suitable for conventional CMOS applications such as logic and memory (Chaves et al. 2020). With a layer-dependent tunable bandgap and strong light-matter interaction, 2D materials are also suitable for diverse optical devices such as photodetectors, modulators, lasers, and light-emitting diodes.

Logic and Memory

The evolution of the electronics industry has largely been propelled by scaling of the contacted gate pitch (CGP) and metal pitch (MP). This scaling has consistently enabled platforms with superior performance and optimized power consumption. However, achieving further area reductions isincreasingly challenging due to processing limitations and intrinsic device constraints. One major component of CGP scaling in silicon-based technology, the gate length, appears to plateau beyond the 3-nm node, as depicted in Figure 14 (Ahmed et al. 2020). As the gate length diminishes, a thinner channel becomes essential to keep short channel effects under control. 2D-FETs have the potential to prolong geometric scaling by overcoming traditional scaling challenges associated with these short channel effects due to their innate channels, enhanced electrostatic control, and superior theoretical mobilities. To demonstrate the advantages of 2D-FETs, Interuniversity Microelectronics Centre (IMEC) has demonstrated a circuit-level power-performance-area evaluation at 2 nanometers between stacked 2D-nanosheets and Si-based counterparts. The results shown in Figure 14 indicate an ~18% reduction in total capacitance (less energy required to switch a transistor on/off) and ~22% improvement in drain current, which indicates lower operating voltage (Ahmed et al. 2020). From these results, we can assume roughly 1.2 times the energy efficiency improvement for 2D materials.

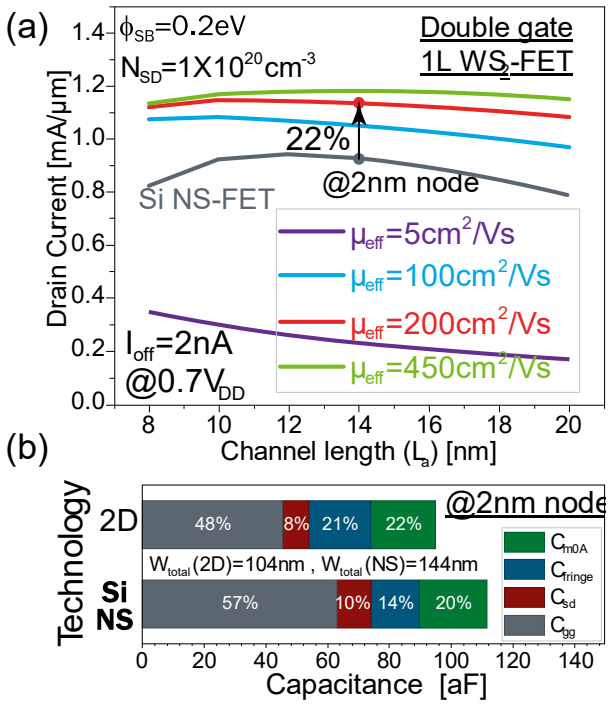


Figure 14. Si NS-FET versus 2D-FET. *(a) Drain current improvement and (b) capacitance improvement. Source: Ahmed et al. 2020*

Several pressing challenges remain before this 2D-FET technology can be commercialized. Most prominent among those challenges are material growth and transfer, which are detailed below. Ultimately, to realize 2D-FETs, 2D materials are needed that are stable to the environment, not adversely affected by edge scattering, and also CMOS- and HVM-compatible. The knowledge gained from previous efforts with graphene and TMDCs may be helpful in investigating other 2D materials.

Applications in Neuromorphic Computing

2D materials, notably transition metal dichalcogenides (TMDCs), have a tunable bandgap, which allows for dynamic control over charge transport, enabling multi-level storage in analog memory devices. 2D materials have imperfections that serve as charge trapping sites, and with external voltage, these sites can hold charge, modifying the device's conductance for analog storage (Cao et al. 2020). The nature of these traps, however, can be tailored by adjusting defect characteristics. 2D materials (as well as other materials) also display resistive switching, where resistance changes with applied voltage (M. Wang et al. 2018). This behavior, influenced by factors like metal ion movement or charge dynamics at defects, can result in varied resistance states for broader analog data representation. Furthermore, 2D materials show promising retention times (>10 years), which are crucial for the longevity of stored states in analog memories (Rehman et al. 2020).

Challenges and Solution Pathways for 2D Semiconductor Materials

Material Growth

Monolayer 2D devices have garnered attention through significant lab-scale demonstrations, predominantly involving single-TMDC flakes. However, for these 2D materials to be broadly adopted, the development of wafer-scale growth of 2D films is needed. In Figure 15, three main techniques for producing high-quality monolayer TMDCs are shown: powder-based chemical vapor deposition (CVD), metal-organic CVD (MOCVD), and molecular beam epitaxy (MBE) (Briggs et al. 2019).

Powder-based CVD is preferable for research applications because of its low manufacturing cost in synthesizing high-quality, defect-free TMDCs (Liu et al. 2015). However, source concentrations cannot be independently adjusted, which restricts the potential of power-based CVD for producing large-area TMDCs.

For the commercialization of 2D TMDCs, MOCVD and MBE are the preferred pathways. MOCVD offers tight control over domain sizes and density, thanks to its precursor switching and pulsing techniques. On the other hand, MBE heats ultra-pure sources in Knudsen effusion cells, directing beams of atoms or molecules onto a heated substrate. Due to the exceptionally pure materials and ultra-high vacuum conditions, MBE excels at producing high-quality, expansive 2D TMDCs (Yue et al. 2017). However, MBE for TMDCs faces challenges due to the high vapor pressure of sulfur, which limits its application mainly to selenides and tellurides, barring a few exceptions.

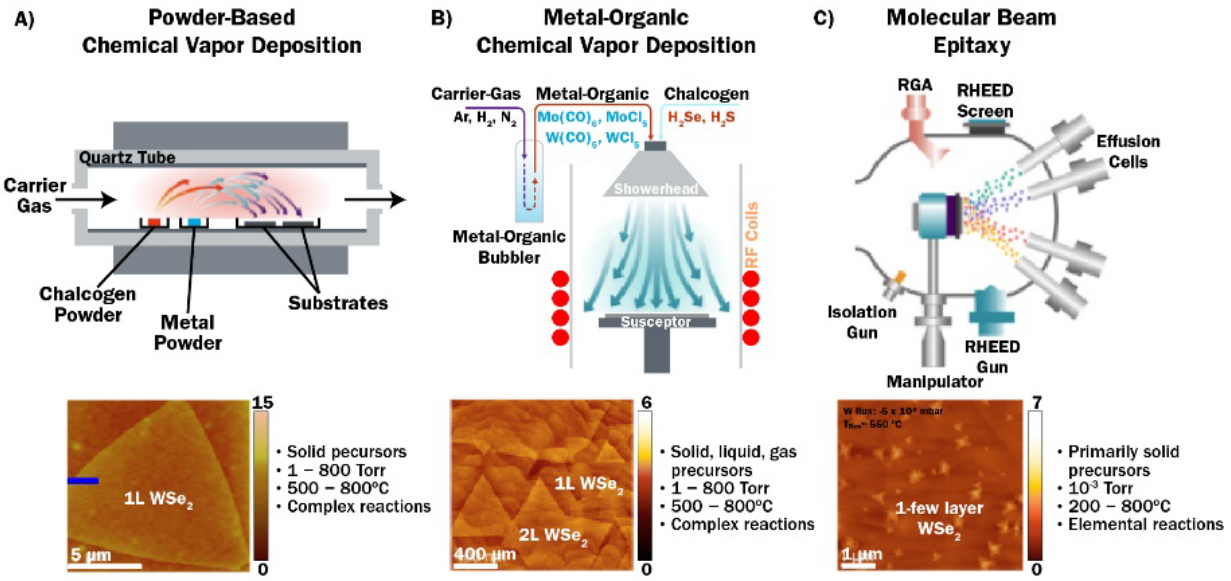


Figure 15. Overview of synthesis techniques of single to few layer TMDC flakes. *Source: Briggs et al. 2019*

While MOCVD and MBE hold promise for synthesizing large-area 2D TMDCs, several challenges must be addressed before they can be widely adopted for these materials. With their high-vacuum and ultra-high vacuum environments, these techniques come at a higher cost compared to other methods. Furthermore, their inherent complexity—coupled with concerns over scalability, growth rate, and consistent production of defect-free large areas--poses significant barriers. For precise parameter control, robust modeling and simulation practices are still in development.

Material Transfer

Once high-quality, defect-free monolayer 2D materials are produced, an efficient, high-quality, and repeatable transfer technique is needed for device fabrication. Presently, the transfer of 2D materials often involves the use of hazardous chemicals, including hydrofluoric acid (HF), hydrochloric acid (HCl), and nitric acid (HNO_3) (Elías et al. 2013). Not only are these substances environmentally unfriendly, but they also compromise the quality of the film during the process. Strong bases like potassium hydroxide (KOH) and sodium hydroxide (NaOH) have been explored as alternatives, but these etchants can similarly degrade the film quality and impair device performance due to their high corrosiveness and inadvertent doping effects (Wang et al. 2014). Research on a suitable transfer methodology is still ongoing. For example, the ultrasonic bubbling transfer method utilizes microbubbles produced by ultrasonication to lift the film from the substrate (Ma et al. 2015). However, this technique can introduce cracks and wrinkles to the transferred film. Regardless of approach, the transfer process must be CMOS-compatible to leverage existing processes and tools for wafer fabrication.

Action Plan for 2D Semiconductor Materials

Table 5. Action Plan for 2D Semiconductor Materials

Scope			
Technology for Energy Efficiency:	2D material-based devices, primarily TMDCs and h-BN		
Technology of Interest:	Logic		
Challenges		**Solution Pathways**	
• Develop process for scalable synthesis and post-growth transfer (if needed) of 2D materials and post. • Maintain consistent thickness of 2D materials. • Minimize defects in 2D materials. • Alleviate thermal stability and contact resistance issues at 2D materials interfaces.		• Transition to other 2D materials (away from graphene) with innate bandgaps suitable for various applications, including Stanford University's Nano-Engineered Computing Systems Technology (N3XT). • Leverage 2D-FETs for further device miniaturization and enhanced energy efficiency. • Investigate key TMDC growth techniques: powder-based CVD, MOCVD, and MBE. • Address challenges in transferring high-quality 2D materials using innovative methods.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline (years)**
Discovery and characterization	Electrical conductivity, bandgap measurement, mechanical strength	Establish standardized characterization protocols	1–2
Bandgap engineering	Bandgap values, uniformity across samples	Achieve tunable bandgap ranges suitable for semiconducting applications	2–3
Research on production techniques	Rate of synthesis, purity, defect density	Develop methods for large scale production with less than 1% defect density	3–5
Transfer Technique Development	Quality retention post-transfer, throughput	Refine transfer methods to maintain >95% quality retention	1–2
Commercialization roadmap	Cost per area, scalability metrics	Reduce production cost by 50%, create a scalable commercial process	4–6
Integration with current technologies	Compatibility, performance benchmarks	Integrate with existing CMOS processes with demonstrated performance benefits	3–4
End-to-end testing	Device failure rates, performance metrics	Achieve less than 5% device failure over standard testing procedures	2–3
Stakeholders and Potential Roles in Project			
Stakeholder	**Role**		
Product Manufacturers/Suppliers	• Synthesize high-quality, scalable 2D materials. • Develop transfer methods preserving material integrity. • Innovate in device integration and performance testing.		
End Users/Original Equipment Manufacturers (OEMs)	• Specify performance and integration requirements. • Pilot test new 2D material-based components. • Provide feedback for material and device optimization.		
Academia	• Conduct fundamental research on 2D materials. • Explore novel properties and potential applications. • Collaborate on translating lab-scale successes to industry-scale processes.		
Required Resources		**Cross Collaboration Needs of Working Groups**	

• Advanced material synthesis equipment. • High-precision characterization tools. • Funding for interdisciplinary research and development projects.	• Circuits and Architectures: Optimize the interplay between algorithmic design, software, and hardware for enhanced energy efficiency. • APHI: Ensure novel materials are integrated into next-generation packaging solutions. • MEES: Integrate 2D materials into current infrastructure.

2.1.2 Carbon Nanotube-Based Devices

Single-wall carbon nanotubes (CNTs) have unique characteristics—including ultrathin (1D) body, high and symmetric electron and hole mobility (Franklin et al. 2012a), outstanding current density, and very low parasitic—that make them promising materials for transistors that are more energy-efficient than conventional Si CMOS.

CNTs come in two electrical types, semiconducting and metallic, determined by their chirality. As shown in Figure 16, CNT chirality is defined by the rolling angle of a graphene sheet into a CNT. Uncontrolled growth of CNT results in approximately two-thirds of grown CNTs being semiconducting and one-third metallic. CNT diameters vary from roughly 0.7–3 nanometers, and the bandgap of semiconducting CNTs approximately follows 0.9 eV/diameter. Therefore, variations in diameter can significantly influence the characteristics of CNT-based devices, particularly affecting threshold voltage and off-current due to thermally excited charge carriers.

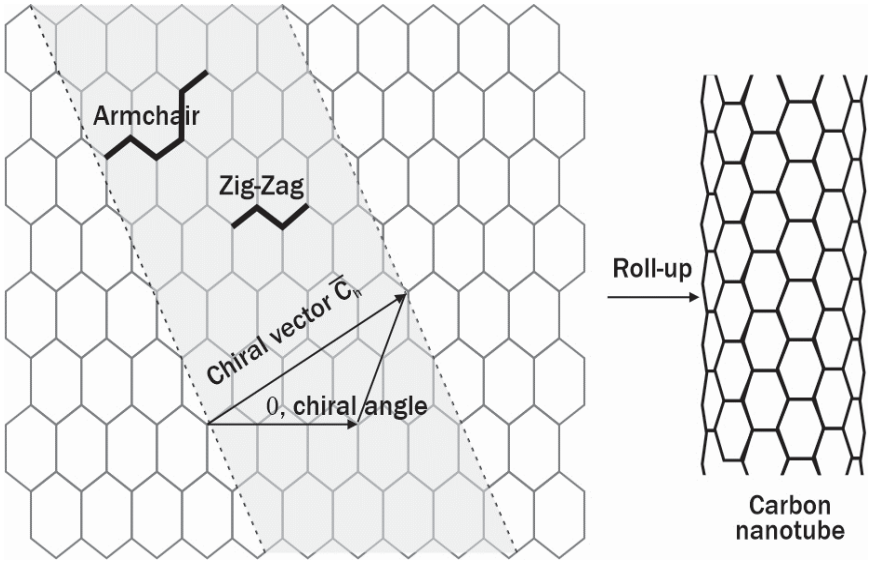


Figure 16. Graphene lattice with chiral angle and vector for CNT. *Two dominant CNT configurations (armchair and zigzag) are shown. Source: Ávila and Lacerda 2008*

Conductivity in CNT is sensitive to impurities and defects that cause scattering. With their single-atom thickness, external adsorbents and contamination directly influence carriers through scattering but also by gating semiconducting CNTs. This makes CNTs candidates for single-molecule sensors but also raises contamination as a key issue. Contaminants introduce significant hysteresis through charge traps that fill slowly, leading to low yield, high device variability, and performance degradation. While significant progress has been made in controlling the chirality and diameter of CNT material while keeping it clean, challenges remain. Requirements for material purity (including chirality) and quality (e.g., lack of defects and consistent length) will vary for different classes of devices (logic and memory-for-compute), as detailed in the proceeding sections.

While the benefits of CNT-based devices are clear, most research has been in laboratory settings. Commercialization requires more research, testing, and progress before justifying new facilities or risking contamination of current, expensive, and entrenched semiconductor processes and equipment.

2.1.2.1 Carbon Nanotubes Field-Effect Transistors

In CNT field-effect transistors (CNTFETs), CNTs are used as the channel material where an applied gate field lowers the barrier for carrier injection from metal contacts. Due to their ultra-thin 1D geometry and high carrier mobility, CNTs overcome short channel effects that hamper Si MOSFETs below the 10nm node. CNTs also demonstrate near-ballistic carrier transport, further amplifying their value for energy-efficient computation. To date, several groups have demonstrated the potential for CNTFETs in various devices, including complementary devices (Ding et al. 2012; Han et al. 2013) and devices with channel lengths scaled below 9nm (Franklin et al. 2012b) and operating voltage below 0.5 V (Wei et al. 2013). However, these demonstrations typically rely on a single transistor or a few CNT transistors, and CNT-material production is the critical bottleneck to commercialization. The next major step is to demonstrate these characteristics with CNT-dense devices that are comparable or better than Si.

To be competitive with 2nm silicon technology, CNTFETs require >99.9999% semiconducting CNTs, a diameter of < 1.2nm for low off-current, and a density of >125 CNTs/μm^2 (IRDS 2022). The CNTs must also be sufficiently long and free of contaminants. These requirements are far more stringent than those needed for CNT memory, RF CNTFETs, and CNT-based sensors. Atomic-level control of CNT production is paramount to the demonstration of CNTFETs for logic.

First-generation CNT devices will use metal interconnects like Si CMOS, where most energy is lost in chips. Thus, the energy improvement over current silicon technology (Si FinFET) will be modest (~3 times). However, the ability of CNTFETs to be stacked and monolithically 3D-integrated is where they truly provide efficiency benefits. Aly et al. proposed a novel computing approach (Nano-Engineering Computing Systems Technology [N3XT]) that monolithically integrates logic and memory, leveraging CNTFETs as the logic component, and shows a 1,000x improvement in energy-delay product compared with conventional Si technology (Aly et al. 2015).

Table 6. Energy Impact and Timeline Estimates for Carbon Nanotube Field-Effect Transistors

Technology	Expected Performance	Commercial Benchmark Product	Commercial Benchmark	Energy Impact Factor	Timeline for TRL 6
CNTFET (logic)	0.2 pJ/cycle	SiGe FinFET or GAA-FET	0.6 pJ/cycle	3	3–5 years

Studies like these (and many more) convey the potential impact of CNTFETs. However, significant challenges remain.

Challenges and Solution Pathways for Carbon Nanotube Field-Effect Transistors

Carbon Nanotube Material Production

Currently, there are three dominant methods of producing CNTs for electronics, each method has its own challenges.

CVD Growth With a Catalyst

CNT growth via CVD produces pristine/contamination-free and high-performance CNTs but has not yet shown adequate chirality and diameter control for digital applications. Early CVD growth using iron catalyst and other materials yielded CNTs with varying diameters and a mix of semiconducting and metallic tubes (Molckovsky et al. 2019). Current efforts have used controlled catalyst size to selectively grow CNTs in a narrow diameter distribution and used CVD conditions to drive higher semiconducting CNT yield. While progress has been made, more work is needed to fine-tune the process and increase selectivity and CNT density. More broadly, CNT CVD growth is not well understood, and fundamental experimental and modeling efforts are needed to accelerate progress.

Purification Through Polymer-Conjugation Sorting

CNT purification with polymer conjugation was pioneered by Mark Hersam and Mike Arnold at Northwestern University nearly 20 years ago. The polymers differentially bind to CNTs by diameter and chirality, allowing them to be separated in solution via one or more cycles of ultracentrifugation. Enriched semiconducting purity CNTs can then be distributed from solution on wafer, either without order or aligned using varying methods including floating evaporative self-assembly (Brady et al. 2016). Selectivity of >99.99% semiconducting purity and sufficient density for digital applications has been shown, but the processing results in damage and short tubes. While significant progress has been made using this method—including the first 100 GHz (Rutherglen et al. 2019) and THz CNTFET (Z. Zhang et al. 2023) demonstrations—the complete removal of the polymer wrapping remains a key challenge that continues to hinder device performance.

Purification Using Nonpolymer-Conjugation Sorting

Post-growth on-wafer purification involves removing the metallic CNTs from pristine CVD-grown tubes directly on a wafer. This method utilizes the electrical or optical response of the CNTs themselves to identify and remove the metallic CNTs. The electrical response is used in VLSI-compatible metallic CNT removal (VMR), which removes metallic CNTs in formed CNTFETs by flowing current through them with the semiconducting CNTs turned off by a sufficient gate voltage. The current in the metallic CNTs can be either sufficient to destroy the CNTs like a fuse or sufficient to heat and thereby pattern a masking layer, leaving the metallic CNTs exposed to etch chemistry (Shulaker et al. 2015). The primary challenge with this approach has been degrading the remaining semiconducting CNTs, thereby degrading CNTFET performance. Rapid progress is now being made using the optical response of CNTs. Here, like in electrical heating post-CNTFET formation, selective heating of metallic CNTs through electromagnetic energy absorption in the RF (Xie et al. 2014) or visible/IR (Du et al. 2014) ranges have been demonstrated.

Regardless of approach, continued and sustained R&D of CNT material production to fabricate CNTFETs that meet digital logic requirements is needed.

Contact Resistance

In addition to CNT material production, contact resistance is another fundamental materials challenge holding CNTFETs from their theoretical/projected potential. Because ballistic transport with extremely low resistance can be achieved for CNT lengths below 40nm, the contact resistance ultimately determines CNTFET performance for scaled devices (Franklin et al. 2014). The best CNTFET contacts allow for much lower bias voltages and thereby significant energy savings over Si. As with Si, the choice of metals and heat treatments can vary these resistances across a few orders of magnitude. Broadly, the interaction and transport phenomena at the CNT-metal interface are not well understood. Previous studies (Franklin et al. 2014; Pitner et al. 2019) investigating side contacts, varying overlap, and carbide formation with various metals have generated useful insights. Most studies were carried out on single-CNT devices, and devices with sufficient, dense CNTs will likely have a distribution of overlaps and resistances. A more fundamental understanding of interface and transport behavior may ultimately accelerate the time to a feasible solution. As such, *ab initio* modeling of the CNT-metal interface and experimental validation on high-density CNTFETs is a key pathway. Other solutions include experimental study of new metals, including carbide-based materials and work function matching.

Dielectric Materials

Consistent with silicon-based FETs, high-k gate dielectrics are needed for sufficient gate capacitance and current control in CNTFETs. Identifying the appropriate gate dielectric material and processing steps remains an open challenge. Atomic layer deposition (ALD) is the dominant method and aluminum, hafnium, and zirconium-based dielectrics have been explored. Owing to the geometry and chemistry of (subject), conformal, uniform growth on CNTs has been difficult. An additional complexity for CNTFETs is the potential for uncontrolled, secondary reactions of the ALD precursor and the CNTs (Simmons et al. 2006), resulting in unexpected and degraded CNTFET performance. Various strategies, such as self-assembling monolayers prior to deposition, have been employed to mitigate this problem with varying results.

Dielectrics have also been found to dope CNTs, reflecting the CNT-dielectric interface properties. Typically, CNTFETs are p-type devices at ambient conditions, but recent studies have shown n-type behavior using hafnium dioxide (HfO_2), thought to be due to the positive fixed charges at the CNT-HfO_2 interface (Moriyama et al. 2010). This finding provides a pathway to enable complementary CNT devices through dielectric doping in the spacer region through the right material combination. Regardless of gate or spacer dielectric, further research is necessary to refine deposition techniques for CNTFETs. This includes exploring requisite pre-treatments and post-processing steps aimed at minimizing CNT damage or contamination. Additionally, an in-depth examination of material chemistry is essential to develop CNTFETs that feature high-density CNT arrays.

Device Performance, Modeling, and Simulation

A consistent device model is needed to allow system simulations for CNTFETs. While several compact models for CNTFETs have been created, consistent device designs and data are needed for parameter extraction, most of which come from university labs at present. A consistent fabrication capability and large datasets are also necessary, which is only possible in industrial facilities with process control. In fact, because there are no clear winners in any of the core CNTFET components (e.g., CNT production, dielectrics, contacts, device architecture), it is

a challenge to generate a compact model that holistically considers all combinations. At present, each model consists of the components that the modeler finds appropriate or most promising. Until these pieces are firmed up, a consistent model may be difficult to formulate.

Foundry/Process Integration

The way in which CNTFETs will be integrated with existing foundry and CMOS processes is still an open question. Contamination stemming from CNTs and their likely contact metals is a serious concern for industry. While the general processing steps (e.g., lithography, deposition, or etch strip) for a CNTFET are not significantly different from CMOS, changes will be required to accommodate the materials used in these processes and may require a parallel set of processing tools. As a first step in facilitating this transition, a collaboration between CNTFET developers and commercial R&D facilities—like Skywater, SUNY NanoTech, and IMEC—with the goal of developing a process design kit (PDK) for CNT-based devices, would accelerate progress. It should be noted that CNT-CMOS co-fabrication was done at Skywater through the DARPA ERI program. While the project was ultimately unsuccessful due to insufficient system-level performance resulting from poor CNT material, it nevertheless provides a blueprint for how CNTs can be integrated into a fab.

Action Plan for Carbon Nanotube Field-Efficient Transistors

Table 7. Action Plan for Carbon Nanotube Field-Efficient Transistors

<table>
<tr><th colspan="4">Scope</th></tr>
<tr><td colspan="2">Technology for Energy Efficiency:</td><td colspan="2">CNTFETs</td></tr>
<tr><td colspan="2">Technology of interest</td><td colspan="2">Logic</td></tr>
<tr><th colspan="2">Challenges</th><th colspan="2">Solution Pathways</th></tr>
<tr><td colspan="2">• Achieve consistent CNT quality for improved device energy efficiency.
• Develop dielectrics for "CNT doping" to boost device performance.
• Understand CNT-metal interfaces for charge transfer.
• Ensure device performance for real-world applications.
• Bridge the gap between lab innovations and mass production.</td><td colspan="2">• Continue to develop primary CNT manufacturing pathways (listed above).
• Leverage advanced dielectrics and spacers to tune and enhance device performance and energy requirements.
• Complete a comprehensive and fundamental study of contact metals and CNT.
• Develop compact models for accurate simulation and design.
• Scale CNT technologies for integration with current manufacturing processes.</td></tr>
<tr><th>Major Tasks/Milestones</th><th>Metrics</th><th>Targets</th><th>Timeline (years)</th></tr>
<tr><td rowspan="3">Improving CNT material production</td><td>Semiconducting to metallic CNT ratio</td><td>>99.99%
>99.9999%</td><td>1–3
3–5</td></tr>
<tr><td>Current per CNT</td><td>>15 uA/CNT</td><td>1–3</td></tr>
<tr><td>CNT density</td><td>100 CNTs/um</td><td>3–5</td></tr>
<tr><td rowspan="4">Explore dielectric materials and deposition techniques</td><td>Breakdown field (gate)</td><td>6 MV/cm</td><td>1–3</td></tr>
<tr><td>Effective oxide thickness (gate)</td><td>3nm</td><td>1–3</td></tr>
<tr><td>Dielectric constant (spacer)</td><td>SiOx or better</td><td>1–3</td></tr>
<tr><td>Doping (spacer)</td><td>Effective field of ±1V normalized by an EOT similar to the gate</td><td>1–3</td></tr>
<tr><td>Improving contact resistance</td><td>Resistance</td><td><30 kOhm/CNT across devices</td><td>1–3</td></tr>
</table>

CNTFET performance	Energy delay product	10x improvement over Si-GAA	3–5
	Gate resistance	<5 kOhm/CNT contact	3–5 for e-beam, many more for EUV
Introduction of CNT material into multi-user commercial R&D foundry	Throughput	1,000 wafers/month without degradation in material properties	3–5

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	R&D companies • Develop CNT material and devices. As of this writing, the authors are aware of only 4 U.S. startups doing active CNT development (versus dozens 10+ years ago): Aligned Carbon, Carbon Technology, Nantero, and SixLine. Commercial R&D foundries • Fabricate devices, develop PDKs, and complete electrical characterization (commercial R&D foundries).
Academia	• Engage in fundamental materials and device research, modeling, and simulation. • Develop and conduct advanced metrology.

Required Resources	Cross Collaboration Needs of Working Groups
• Material simulation: ab initio calculation for CNT-metal contact interface. • Circuit modeling and simulation: compact models and system models. • Facilities, including access to nanofabrication foundries, advanced nanoscale metrology, and electrical testing capabilities.	• Circuits and Architectures: Evaluate benefits of monolithic 3D integration of CNTFETs; modeling and simulation of system level performance from device characteristics. • Metrology and Benchmarking: Evaluate and understand CNT contamination (atomic scale) and its impacts on device performance, benchmark CNTFET performance.

Carbon Nanotube Memory

Unlike logic, memory applications have significantly less stringent requirements for CNT-based devices. CNT memory (e.g., NRAM from Nantero) utilizes a mat of CNTs that deflect and open a gap between electrodes when a voltage is applied. This gap creates a sufficient change in resistance between the two electrodes to register a 0 or 1. The CNT mat is produced via commercial fab standard spin coating and annealing. Because CNTs in this configuration are multi-layered and are not channel materials, requirements for uniformity in chirality and selectivity (semiconducting vs. metallic) are relaxed.

The current instantiation of CNT memory (NRAM) is a single-layer, 2-gigabit capacity memory on a 22nm die. NRAM's performance is comparable to existing DRAM products, but its non-volatile nature significantly enhances efficiency, reducing power consumption by an average of 33% in the DDR4 performance in active. Research indicates that approximately 50% of DRAM in data centers is idle, consuming 30% of the power used during active periods. Substituting DRAM with NRAM could lead to an average power saving of 15% in idle due to NRAM's lower idle power requirements (Zhang et al. 2014). Additionally, NRAM's non-synchronous bank-level operations for reading and writing, coupled with a GHz clock only present in interface circuitry, further reduce active power consumption. Nvidia's research suggests that smaller data accesses (512B or 256B) are more efficient in GPUs, occurring 80% of the time. NRAM's adaptability with flexible page sizes (down to 256 bits) and the ability to handle multiple pages

simultaneously can lead to a 75% power reduction during these frequent operations (Chatterjee et al. 2017).

Table 8 estimates energy impact and timeline for comparing NRAM and DDR4 DRAM in 2Kbit/page mode.

Table 8. Energy Impact and Timeline Estimates [a] for Carbon Nanotube Memory

Metric	NRAM	DDR4 DRAM	Impact Factor	Timeline for TRL 6
Energy per bit	5 fJ/bit	7 fJ/bit	1.4x	3 years
Non-volatile	Yes	No	-	
Latency	5 ns	15 ns	3x	
Frequency	64 GB/s	64 GB/s	1x	
Active power	260 mW	408.3 mW	1.6x	
Idle power	0.8 mW	85.5 mW	106x	

[a] *Source: Micron Technology 2017*

Challenges and Solution Pathways for Carbon Nanotube Memory

Foundry/process integration and contamination

At present, NRAM processes and materials have been defined and fab integration remains the dominant challenge in bringing NRAM to market. Unsurprisingly, concerns over contamination have barred access to fabs. Contamination tests have been completed and previous NRAM runs in R&D fabs have had no issues. Ultimately, fab runs at leading edge nodes are needed to identify areas of further development to continue moving this technology forward.

Action Plan for Carbon Nanotube Memory

Table 9. Action Plan for Carbon Nanotube Memory

Scope			
Technology for Energy Efficiency:	CNT memory		
Technologies of Interest:	Memory		
Challenges		**Solution Pathways**	
• Secure access to manufacturing facilities for CNT technology to demonstrate technology at leading edge nodes. • Scale technology to relevant memory capacity/density.		• Alleviate CNT contamination concerns through BEOL short loops and post-process contamination characterization.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline (years)**
Complete specification for non-volatile DRAM replacement	JEDEC approval of the DRAM specs CXL specs UCIe specs	DDR5/DDR6 SDRAM HBM4/HBM5 CXL UCIe	2–4
Fab integration	CNT fabrication process in current infrastructure	Deployment and adoption of CNT fabrication process in current infrastructure from industry	1

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Integrate CNT process and materials into fab operations. • Integrate CNT memory into product.
End Users/OEMs	• Define requirements for various applications.
Academia	• Define standards for certain applications (e.g., data security). • Develop standards that comprehend CNT memory semantics.

Required Resources	Cross Collaboration Needs of Working Groups
• Production of NVM technology in commercial fabs. • Integration of NVM technology with existing systems.	• Circuits and Architectures: Develop comprehension and application of CNT memory to memory protocol. • APHI: Provide chiplet support and integration for NVM devices.

2.1.3 Spintronic Devices

Spintronic materials, the building blocks for spin transport-based electronic devices, rely on an electron's charge, as well as its magnetic spin to perform computations and store data. These materials offer the potential for circuits that can achieve low power consumption and high endurance, with competitive read and write performance. While ideal spintronic devices promise negligible standby power dissipation, practical implementations have yet to achieve this due to perfect spin polarization or detection efficiencies. Current state-of-the-art spintronic devices, as demonstrated by Ikeda et al., achieve on-off ratios of 7:1 at room temperature. While not approaching the on-off ratios of experimental CMOS technologies, these 7:1 at room temperature ratios still mark a significant advance owing to their non-volatility (Ikeda et al. 2007). This inherent non-volatility in spintronic devices is a distinct advantage over CMOS with respect to energy efficiency because it enables data retention without power. Recent advances in 300 mm processing and manufacturing tools have led to the availability of spintronics manufacturing capacity within back-end-of-line facilities at leading semiconductor foundries (Lee et al. 2018).

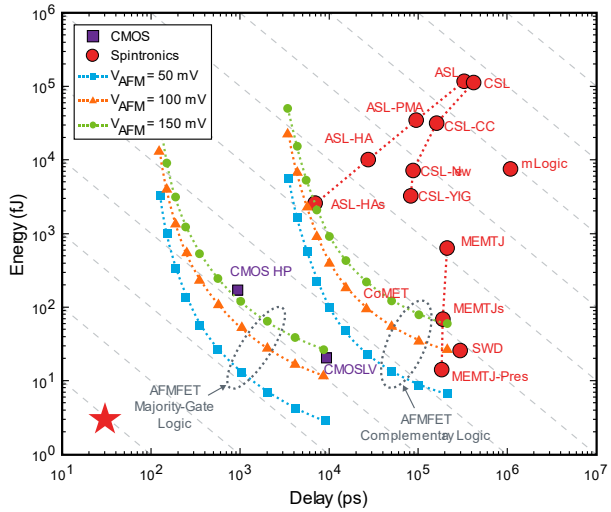


Figure 17. Comparison of energy and delay of a 32-bit adder among various charge- and spin-based devices. *Benchmarks show performance and switching energy versus delay time. Source: Pan and Naeemi 2018*

A key focus for researchers in recent years has been to address the challenge of high switching error rates in spintronics (Sun et al. 2022). Both memory and logic devices in this field rely on precise control over electron spins to perform dependable data read and write operations. While factors like thermal noise and spin-orbit interactions can introduce variability in these spin states, innovative solutions are being developed to minimize errors during state transitions (Tan et al. 2021). Despite these challenges, the technology's substantial energy efficiency and scaling benefits remain promising, and spintronics continue to offer opportunities for advancing energy-efficient microelectronics.

2.1.3.1 Spintronic Logic

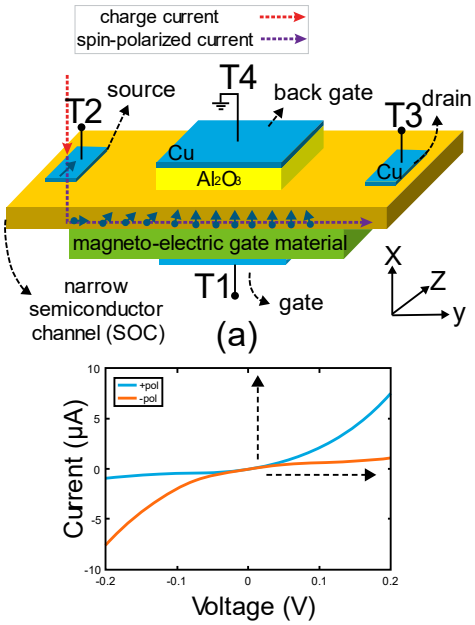


Figure 18. The magnetoelectric FET with performance shown for a channel with a spin-orbit splitting of only 100 meV. *With a channel of 0.5 eV spin-orbit splitting the on/off ratio might approach 10^4 (Dowben et al. 2018)*

Spintronic logic uses an electron's intrinsic spin to encode data in nanoscale magnets, including single domain magnets or, alternatively, non-uniform magnetic textures like magnetic domain walls, spin waves, and magnetic skyrmions. These data elements can be written using spin-polarized currents, as well as various electrostatic gating-induced effects, such as voltage-control of magnetic anisotropy, voltage-control of interlayer exchange coupling, and magnetoelectric reversal of exchange bias. The data state can be read out either by integrating the data element into a magnetic tunnel junction or by the spin Hall voltage generated in an adjacent spin-orbit coupling layer.

Generally, there are two types of spintronic logic devices: current-controlled and voltage-controlled. Voltage-controlled devices are orders of magnitude more energy-efficient than current-controlled and are the preferred type. Voltage-controlled magnetoelectric spintronic logic devices (e.g., MESO) have the potential for 30x (or more) improvement over CMOS. However, many of the devices are reliant on ferromagnetic materials, which are slower than CMOS since moving spin textures (e.g., domain walls) or switching the magnetization of a single-domain element is slower than charging the gate capacitors. Thus, more recent work focuses on magnetoelectric transistors without ferromagnetism.

Structurally, magnetoelectric transistors are quite straightforward, as shown in Figure 18. Magneto-electric FETs can streamline a full adder from 28 transistors to merely eight device elements, which could lead to a considerable reduction in energy costs and an increase in operational speed (Sharma et al. 2020). However, more complex structures pose integration challenges with current CMOS technologies (Mahmood et al. 2021). Despite these hurdles, recent strides have been made in developing magnetoelectric or multiferroic non-volatile technologies, which have the potential to transform the landscape of electronics (Manipatruni et al. 2019; Vaz et al. 2021; Kosub et al. 2015; Kosub et al. 2017; Mahmood et al. 2021; He et al. 2022).

Table 10. Energy Impact and Timeline Estimates[a] for Spintronic Logic

Technology	Expected Performance	Commercial Benchmark Product	Commercial Benchmark	Energy Impact Factor	Timeline for TRL 6
Spintronic (Logic)	50 fJ/switch	CMOS HP	100 fJ/switch	2	10 years

[a] Source: Puebla et al. 2020

Challenges and Solution Pathways for Spintronic Logic

Switching Error

In Boolean logic, write error rates of approximately 10^{-15} are required for reliable device function (Manipatruni et al. 2019). As a standalone, spintronic devices fall short with typical error rates around 10^{-5} or worse (Manipatruni et al. 2018). Furthermore, if integrating with CMOS, the discrepancy in on/off will cause significant switching errors.

One key strategy to address these issues involves the exhaustive exploration of alternative materials and switching mechanisms with the potential for a thousandfold energy reduction. A rigorous approach would entail computational modeling to identify materials most likely to achieve the requisite low error rate, followed by experimental validation. This process would include initial film measurements and subsequent prototyping to empirically validate switching performance. Additionally, a comprehensive understanding of materials, defects, inhomogeneities, and process issues is essential for identifying the root causes of switching errors.

Energy To Switch Magnetization

Improving the energy efficiency of switching mechanisms in spintronic devices requires both material and architectural innovation. Various magnetoelectric materials are being studied for their potential to minimize leakage and reduce coercive voltage. Component magnetic materials must be identified that can be heterogeneously integrated into a layered structure for optimal performance, while also considering scalability.

Currently, MESO and other logic devices could benefit from four classes of materials: (1) spin-orbit coupling materials for spin-to-charge conversion, (2) magnetoelectric materials for charge-to-spin conversion, (3) interconnects scalable to nanoscale widths, and (4) nanomagnets (Manipatruni et al. 2019). For magneto-electric FET, the key challenges are finding a spin-orbit coupling material for the semiconductor channel with large spin orbit coupling and a demonstration that the magneto-electric can be scaled to small volumes and low coercive voltage while still retaining fidelity to 400 K. Material selection should account for attributes like

coupling strength, temperature stability, scalability, chemical resilience, and non-volatility. Considerable R&D efforts are needed to achieve this goal.

Fabrication

Fabrication of spintronic logic devices presents challenges, especially when targeting integration with established CMOS technology. Material contamination is of note. During the ion milling stages of fabrication, contamination can lead to the shorting of the oxide tunnel barrier, which is crucial for the spintronic memory/logic read-out circuit. This problem becomes pronounced when forming both memory and logic devices on CMOS substrates. The introduction of new materials often necessitates specialized buffer layers or substrates. Moreover, these materials might require high-temperature processing, rendering them incompatible with the amorphous or polycrystalline texture of existing surface materials, such as TEOS or metallized via stub.

One possible pathway to advance spintronic devices within the CMOS framework is the development of CMOS-based test vehicles. An exemplary initiative is the Daffodil chip at NIST. This chip delivers a flexible design intended to enable research and development into two-terminal resistive memory and selector devices and can make assessments of write-energy, write-delay, and switching error metrics across diverse prototypes, thereby evaluating integration-level performance (Hoskins et al. 2021). This chip can integrate spintronic device arrays in the BEOL on CMOS reticles. To ensure the viability of this integration, it is vital to pinpoint small-batch tape-out opportunities and leverage the appropriate circuit topologies to efficiently integrate BEOL spintronic devices with CMOS. Opportunities like the Google-NIST partnership, Nanotechnology Accelerator Program, will deliver even more test vehicles that could help researchers transition new spintronic materials and spintronic logic devices using an industrially relevant platform (Google Open Source Blog 2022).

Action Plan for Spintronic Logic

Table 11. Action Plan for Spintronic Logic

Scope	
Technology for Energy Efficiency:	Spintronic logic device
Technology of Interest:	Logic

Challenges	Solution Pathways
• Reduce switching error rates, error rates around 10^{-15} for reliable operation are required. • Integrate spintronics with CMOS technology, addressing poor on/off ratios of standalone spintronic devices. • Identify and utilize materials and mechanisms that can significantly reduce the energy required to switch magnetization in spintronic devices. • Address nanofabrication intricacies, including contamination issues during ion milling and the incompatibility of new materials with established CMOS processes.	• Explore alternative materials and switching mechanisms through computational modeling and empirical validation to achieve low switching error rates. • Develop magnetoelectric materials to improve energy efficiency, focusing on leakage minimization and reduction of coercive voltage. • Advance CMOS-compatible test vehicles like the Daffodil chip at NIST for research and development of two-terminal resistive memory and selector devices. • Leverage small-batch tape-out opportunities and appropriate circuit topologies for efficient BEOL integration of spintronic devices with CMOS.

Major Tasks/Milestones	Metrics	Targets	Timeline (years)
Material selection for switching error	Switching Error Rate	1 in 10^{15}	5
Energy efficiency of spintronic devices	Energy to Switch Magnetization	< 100 aJ/switch	5–10
Integration with CMOS technology	Successful Demonstration	BEOL integration with CMOS	5–10

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Fund academia and start-ups (e.g., Intel through SRC funded materials and benchmarking work.
End Users/OEMs	• Develop specialized synthesis or patterning tools to accommodate diverse materials. • Develop new patterning IP due to cross-junction type stacks.
Academia	• Simulate and demonstrate materials and devices. • Test materials and devices (includes materials growth, device fabrication, and testing over millions of cycles).

Required Resources	Cross Collaboration Needs of Working Groups
• Testing capabilities, develop metrology and infrastructure. • Access to small sample prototypes. Small number of runs on their equipment for demonstration on their platforms. • Supplements to academia to upgrade current infrastructures.	• Circuits and Architectures: Develop novel approach to integrate with CMOS. • Metrology and Benchmarking: Measure materials.

Spintronic Memory

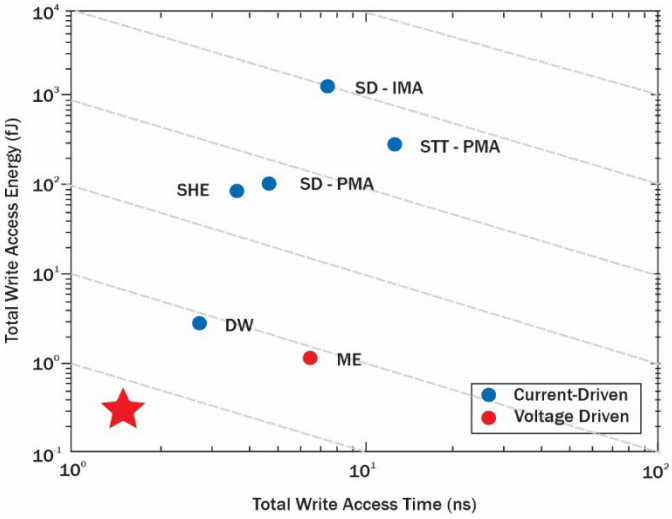


Figure 19. Write energy vs. write delay for various types of spintronic memory cells. *Star shows the desired target (Pan and Naeemi 2017)*

At its core, spintronic memory relies on precise manipulation of electron spins for data storage. In spintronic memory, the orientation of the spins can be aligned or anti-aligned. These alignments represent the binary states (0 and 1) of digital data. Because the magnetic states are inherently stable without the need for a continuous power supply, the data remains "non-volatile" or persistent even when the device is powered off. Non-volatility is a key parameter for energy-efficient computing and AI hardware. One technology that offers non-volatility and endurance (Bhatti et al. 2017) is spintronic memory realized with magnetic tunnel junctions (MTJs). In MTJs, the magnetization orientation of a soft magnetic layer is switched utilizing spin-transfer-torque (STT) by a current that is polarized by the reference (or hard) magnetic layer, whose magnetization orientation is fixed (Slonczewski 1996). The problem, however, is that such spin-transfer-torque magnetic random-access memory (STT-RAM) devices require ~100 fJ/bit to switch (Nowak et al. 2016), which is *1,000 times more than the ~100 aJ energy required to switch CMOS devices* (Datta, Diep, and Behin-Aein 2015). Furthermore, the need for large currents to switch MTJs necessitates the use of CMOS devices with larger node sizes, thus impeding the scaling of this technology. But if the switching current can be reduced by approximately one-half, these devices can be integrated with smaller CMOS devices, which would make MTJ MRAM competitive or superior to embedded DRAM and SRAM at the last level cache (Worledge 2022). This possibility motivates a search for more energy-efficient and low current switching mechanisms to reverse the magnetization while simultaneously keeping the switching error low.

Table 12. Energy Impact and Timeline Estimates for Spintronic Memory

Technology	Expected Performance	Commercial Benchmark Product	Commercial Benchmark	Energy Impact Factor	Timeline for TRL 6
Spintronic (Memory)	100 aJ/bit	STT-MRAM	100 fJ/bit	1,000	10 years

Some of the key methods studied as an alternative to conventional STT-MRAM include spin orbit torque (SOT) (Liu et al. 2012) and voltage control methods, which include direct voltage control of magnetic anisotropy (Kanai et al. 2012) and strain-mediated voltage control (Atulasimha and Bandyopadhyay 2010). Additionally, there are other emerging techniques being explored to enhance the functionality and efficiency of spintronic devices.

Applications in Neuromorphic Computing

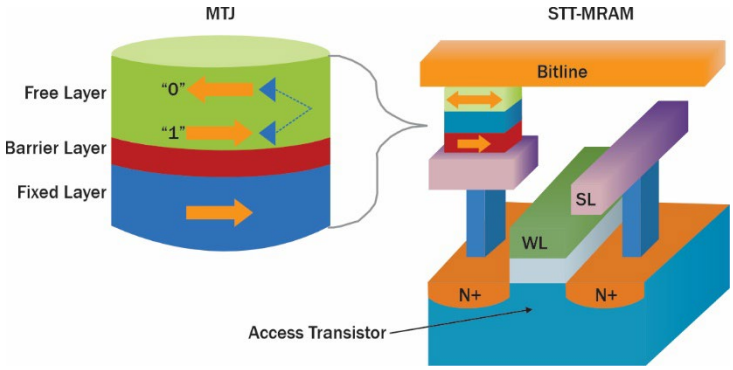


Figure 20. STT-MRAM device structure. *Source: MRAM-info 2023*

Nonvolatility and nonlinear magnetization dynamics in spintronic materials, such as those found in STT-MRAMs and MTJs, are quintessential for the development of neuromorphic analog devices because they enable the emulation of complex synaptic and neuronal functionalities, akin to biological counterparts (Vincent et al. 2015; Borders et al. 2017). Non-volatility ensures that these devices retain information without a constant power supply, mirroring the human brain's energy-efficient information retention, which is vital for instant-on capabilities and reducing power-intensive operations. The non-linear response of these spintronic materials is analogous to biological synapses, whose strength is modulated by the timing and frequency of neural signals, thereby enabling synaptic plasticity, which is central to learning and memory. Furthermore, the potential for three-dimensional stacking of spintronic devices echoes the dense neural networks of the brain, allowing for a compact yet complex network that facilitates vertical communication, which optimizes both space and functionality for advanced neuromorphic computing architectures (Grollier et al. 2020).

The primary challenge to realize MTJ crossbar arrays is the relatively low resistance ratio between the on/off states, making it difficult to read the junction state. Addressing this challenge calls for both material advancements to increase resistance variations and the design of efficient low-power circuits for state reading (Parkin et al. 2004). Given that spintronic behavior can be predictively described based on physical phenomena, implementation for neural networks is achievable (LeCun, Bengio, and Hinton 2015).

Challenges and Solution Pathways for Spintronic Memory

While spintronic memory is more mature than spintronic logic, the development of spintronic memory still faces its own set of challenges.

Switching Error

While memory inherently has a more forgiving threshold for switching error compared to logic—owing to its primary role in data storage and retrieval versus the real-time computational demands of logic devices—this tolerance narrows considerably as device dimensions diminish and industry pushes for enhanced memory density and energy efficiency. Stochastic processes and thermal fluctuations can interfere with the accurate switching of electron spins, failing to change the magnetic state as intended. Such switching errors compromise the reliability of the memory. The leading commercialized STT-MRAM technology currently exhibits a switching error rate of 10^{-6}. The newer magnetoelectric (ME) spintronic devices are comparable to conventional SRAM at 10^{-14} with the additional advantage of being non-volatile (Manipatruni et al. 2017). Despite being a relatively recent area of study, ME devices harness electric fields to adjust the topologically protected spin current in the semiconductor channel. This mechanism

affords ME devices advantages in terms of reduced power consumption and enhanced switching speeds compared to conventional STT-MRAM that depends on spin-polarized currents. To achieve better switching error, R&D is focused on identifying optimal magnetic materials with minimized variability and exceptional thermal stability. Efforts are also directed towards refining the geometry of MTJs for improved stability, implementing error correction codes for post-event error mitigation, and utilizing sophisticated simulation tools, allowing researchers to gain deeper insights into the root behaviors contributing to switching errors.

Energy and Current to Switch Magnetization

Alternative magnetic order, such as ferromagnetic and antiferromagnetic configurations, can offer several advantages over existing spintronic memories, including those based on ferromagnetic materials and MTJs. Antiferromagnetic materials, when integrated with SOT mechanisms or voltage-switched schemes, demonstrate potential for ultra-fast switching with minimal energy costs. Furthermore, recent works suggest that there are suitable room-temperature readout mechanisms for antiferromagnet-based non-volatile memory (Xiong et al. 2022).

Key areas to explore include identification of materials and processing to fabricate double-barrier magnetic tunnel junctions (Hu et al. 2015; Khanai et al. 2021), the use of voltage-modulated perpendicular magnetic anisotropy (Bi et al. 2017), and voltage-modulated exchange coupling (Zhang et al. 2022). Other worthwhile areas to explore include spin-transfer torque for switching, as well as spin-orbit torque combined with spin-transfer torque switching (Grimaldi et al. 2020).

Action Plan for Spintronic Memory

Table 13. Action Plan for Spintronic Memory

Scope			
Technology for Energy Efficiency:	Spintronic memory, specifically MTJs for non-volatile memory		
Technology Interest:	Memory		
Challenges		**Solution Pathways**	
• Reduce energy and current requirements for switching MTJs in Spin-Transfer-Torque Random Access Memory (STT-MRAM). • Reduce switching error rates. • Address scaling challenges for spintronic memories in reducing switching energy and current.		• Explore alternative magnetic orders and materials, like antiferromagnetic and ferrimagnetic, to reduce switching energy and enhance stability. • Develop and utilize spin-orbit torque (SOT) and voltage-switched mechanisms for efficient and fast switching. • Identify optimal magnetic materials with minimized variability for better switching accuracy. • Explore double-barrier magnetic tunnel junctions and voltage-modulated anisotropy or exchange coupling for efficient switching. • Investigate system-level changes, including stochastic MTJ operation for energy efficiency, and MTJ use in compute-in-memory systems.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline (years)**
Switching error reduction	Switching error	1 in 10^6	5–10
Energy efficiency improvement	Energy to switch magnetization	100 aJ/switch	5–10

Switching current reduction	Current to switch magnetization	100 μA or less	5–10

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Fund academia and start-ups.
End Users/OEMs	• Develop capabilities for manufacturing scaled devices.
Academia	• Simulate and demonstrate materials and devices. • Test materials and devices (includes materials growth, device fabrication, and testing over millions of cycles).

Required Resources	Cross Collaboration Needs of Working Groups
• Facilitation of international collaborations while protecting US IP or sharing IP equitably. • Fabrication, characterization, and testing capabilities. • Amortizing design and R&D, including industry scale fabrication equipment, in critical areas. • Supplements to academia to upgrade current infrastructures. • Test bed development.	• Circuits and Architectures: Develop novel approach to integrate with CMOS. • Metrology and Benchmarking: Measure materials.

2.1.4 Ferroelectric Memory/Ferroelectric Field-Effect Transistors

Ferroelectric materials are nonvolatile, exhibiting spontaneous polarization of discrete, stable, or metastable states without an applied electric field. In ferroelectric materials, it is possible to switch between polarization states using an electric field, forming the basis of ferroelectric memory devices (Rabe et al. 2007). The field where the polarization switches to the opposite state is known as the coercive field. Over the past couple of decades, ferroelectric memories have been intensely studied as a replacement or supplement to existing memory technologies. Compared with other non-volatile alternative memory technologies, such as phase change memory and resistive memory, ferroelectric memories require lower write energy, making them more attractive as a more energy efficient competitive memory technology (Hwang and Mikolajick 2019).

Ferroelectric field-effect transistors (FeFETs), which incorporate a ferroelectric oxide or organic ferroelectric between the channel and gate electrode, utilize the permanent polarization of the ferroelectric material to enable memory capabilities. Compared with perovskite-based FRAM, hafnia FeFETs provide numerous advantages, including nondestructive read, fast switching, scalability, high coercive field, and CMOS compatibility. Hafnia FeFETs have also achieved the smallest physical gate lengths of reported FeFETs (Mueller, Slesazeck, and Mikolajick 2019). This advancement in FeFET technology is pivotal as it contributes significantly to the development of more energy-efficient computing systems.

Ferroelectricity has also been observed or predicted in other materials as possible alternatives to hafnia-based and organic ferroelectrics. These include films made of wurtzite-structured materials and 2D van der Waals materials (Liu et al. 2021; Blinov et al. 2000; H. Wang et al. 2018; Si et al. 2018; Guan et al. 2020). Table 14 provides a comparison of hafnia-based FeFETs with incumbent technology. While state-of the-art (22-nm node), hafnia-based FeFETs do not provide improvement in the write energy per bit, compared to embedded SRAM at the 7-nanometer node, SRAM is volatile, lacks multibit operation, and has high standby power. This leads to performance degradation in energy consumption at the system level when standby is

frequent, such as in applications at the edge. By contrast, hafnia-based FeFETs have low standby power, multibit operation potential, and much smaller cell size (10–30 F^2) enabling monolithic 3D integration of FeFETs in the back-end-of-line (BEOL) for compute-in-memory architecture, with significant area, energy, and latency benefits. With this configuration, Dutta demonstrated 3 times' improvement in energy efficiency (TOPS/W) of a 3D monolithically integrated 22 nm BEOL FeFET (2 bit per cell) compared against 7-nm SRAM (Dutta et al. 2020), with potential up to 10 times' improvement with a 4-bit cell. Moreover, for storage applications, FeFETs can provide up to 1,000 times benefits in energy performance compared with Flash (SONOS).

Table 14. Energy Impact and Timeline Estimates[a] for FeFETs

Technology	Expected Performance	Commercial Benchmark Product	Commercial Benchmark	Energy Impact Factor	Timeline for TRL 6
FeFET	1 fJ/bit	eSRAM	1 fJ/bit	1	10 years
		eFlash (SONOS)	1,000 fJ/bit	1,000	10 years

[a] *Source: Khan, Keshavarzi, and Datta 2020*

More recently, ferroelectric tunnel junctions (FTJ) have been studied as a more energy efficient emerging memory technology. As seen in Figure 21, FTJ is an ultra-thin ferroelectric film sandwiched between asymmetric electrodes or interfaces. Polarization states are determined by non-volatile modulation of the barrier height. The ferroelectric dipole orientation ultimately determines the high or low resistance state and can be read non-destructively. To date, research activity has largely been only in academic settings, on perovskite-based ferroelectrics, and on single devices. More work is needed to better understand the ferroelectric/metal interfacial properties, deviations between experimental data and modeled behavior, scalability, CMOS compatibility and the potential for hafnia-oxide-based ferroelectrics for the tunnel junction (IRDS 2021; Garcia and Bibes 2014). The potential for organic ferroelectrics gating a narrow channel transistor also requires further study (Xia and Hu 2022; Kang et al. 2019; Zheng et al. 2009).

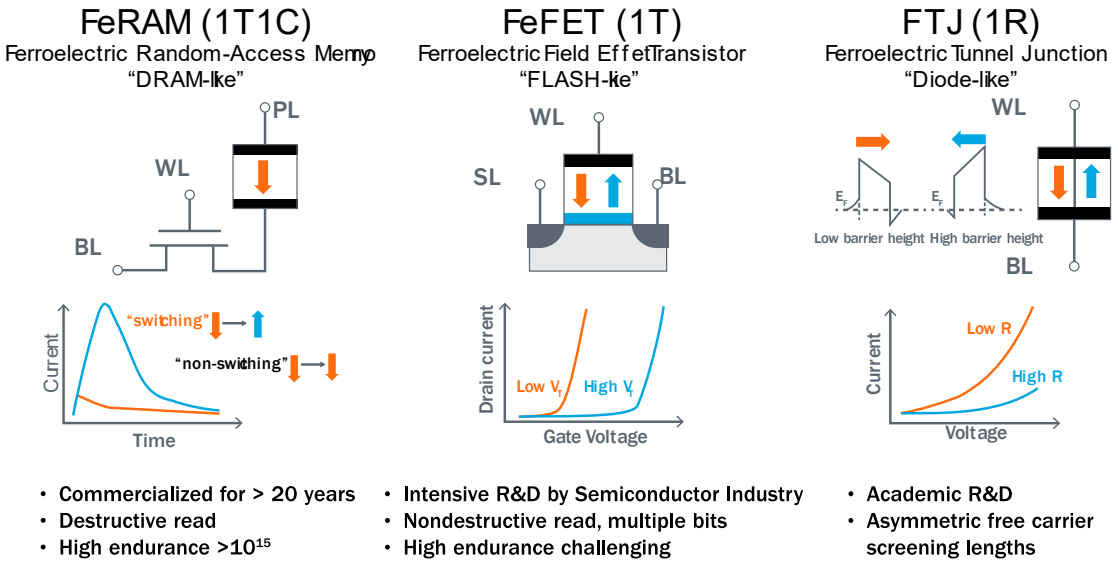


Figure 21. Operating principles of ferroelectric memory. *Source: Mikolajick et al. 2021*

The working group deliberations primarily focused on FeFETs as a promising energy-efficient memory technology; the sections below, therefore, only pertain to the challenges, solution pathways, and action planning for FeFETs.

Applications in Neuromorphic Computing

The switchable electrical polarization in ferroelectric materials allows them to mimic the synaptic weights inherent to brain function. This polarization arises from the formation of stable electric dipoles within their non-centrosymmetric crystal structures. The capacity for controlled and incremental switching in ferroelectric materials makes them prime candidates for neuromorphic architectures, potentially enabling energy-efficient and high-density computational paradigms.

Key challenges include: achieving uniform polarization behavior at the nanoscale, especially in devices less than 100nm in size; extending lifetime, particularly in silicon-based FaFET devices given their limited cycling endurance (Christensen et al. 2022); and optimizing the current density and reading speed in ferroelectric tunneling junctions (FTJ), which is complex and often influenced by the thickness of the ferroelectric layer and the intricacies of multi-layer stacks (Slesazeck and Mikolajick 2019).

Addressing these challenges requires a combination of material science innovations and a deeper understanding of their intrinsic properties. Potential solutions include the stabilization of specific ferroelectric phases in crystallized thin films, introducing dopants, and epitaxial growth of monocrystalline layers. In the context of FTJs, understanding domain wall motion might lead to more refined and analog switching behaviors. Additionally, the exploration of newer materials, with combined ferroelectric and piezoelectric properties, can further expand the horizon for neuromorphic applications.

Challenges and Solution Pathways for Ferroelectic Memory/FeFETs

Device Characteristics: Endurance, Retention, and Write Voltage

Compared with SRAM and DRAM, FeFETs have the advantage of being non-volatile, smaller in cell size, and more energy efficient in standby power. Compared to Flash, FeFETs have superior cycling and the potential for deeper scaling. For example, Hafnia-based FeFETs have been demonstrated at the 22nm node for silicon-on-insulator (SOI) (Khan et al. 2020; Dunkel et al. 2017). However, reliability challenges, related to endurance and retention, are the major barriers yet to be addressed.

FeFETs are becoming an advantageous alternative to existing memory technologies due to their compact size and energy efficiency. Recent studies have demonstrated Si-channel, hafnia-based FeFETs with 10^5-10^9 cycles for deterministic switching, which (while better than Flash) does not compare to SRAM (>10^{16}). Reduced cycling can be attributed to the degradation of the ferroelectric and interfacial layers (notably, the interfacial layer (IL) formed between the ferroelectric (FE) and the Si channel). In hafnia-based FeFETs with Si channels, the mechanisms of charge trapping and trap generation at interfaces (FE-IL and IL-Si) are particularly important. Charge trapping and de-trapping play a crucial role in defining the read speed, particularly influencing the read-after-write latency. This inferior switching speed to conventional SRAM is primarily due to the slow kinetics involved in neutralizing charged interfacial states. These states act as a screen, effectively masking the polarization inherent to the ferroelectric material, thereby impacting the overall speed and efficiency of the reading process (Wang et al. 2021). Compounding this issue, retention and endurance can be inversely

related, depending on the operation mechanism, such that strategies to improve one can degrade the other.

Write voltage is another FeFET characteristic that must be addressed to be a truly competitive technology. Although oxide channel FeFETs have achieved 1.6V (Dutta et al. 2022), at present, typical FeFET write voltage is <4V, compared with <1V for both SRAM and DRAM. At these voltages, FeFET is not compatible for logic. Write voltage is ultimately dependent on coercive voltage (the voltage to switch a polarization state) of the ferroelectric layer. Thus, endurance, retention, and write voltage are intrinsically tied through the materials and material combinations of the channel and gate stack.

High mobility and disorder-tolerant oxide semiconductor channel materials integrated with ultra-thin (e.g., sub-5nm) ferroelectric layers may improve endurance and lower write voltage. Previous studies have shown promise for n-type tungsten oxide and indium tin oxide (Dutta et al. 2022), but p-type oxide channel is severely lacking. R&D solutions are needed for p-type oxide channel materials (for CMOS) that exhibit good stability and electrical performance. Furthermore, defect-enhanced leakage current and/or threshold voltage instability are major challenges that also need to be addressed through R&D. Gating a two-dimensional electron gas with a ferroelectric may also be worthwhile due to higher channel mobilities and increased on/off ratios.

Atomic layer deposition (ALD) research--both modeling and experimental approaches—was proposed as a possible pathway towards fabrication of FeFETs with ultra-thin hafnia-based ferroelectric layers for CMOS-compatible logic voltages with good endurance and retention properties. Film growth by ALD is also highly desired for process integration in advanced CMOS nodes. Density functional theory (DFT) can be leveraged to identify chemical pathways and growth mechanisms to better understand the structure and characteristics of films and interfaces and design better stacks and processes. Experiments could then test and validate these approaches. Plasma-enhanced ALD was noted as a potential solution for BEOL integration of hafnia-based FeFETs because it lowers post-deposition annealing requirements, changes the phase transformation sequence of the ferroelectric, and controls heterogeneity in dopant and defect concentrations, among other factors (Yu et al. 2022).

Working group members also proposed adopting a more practical, results-oriented strategy: a complete full understanding of the underlying principles governing device function might not be essential, but exploring pathways that show improved and promising results could still drive ongoing R&D. Ultimately, to overcome issues related to the endurance-retention trade-off and the write voltage, it is crucial to develop optimized stack designs and materials. Whether these advancements come from theoretical or empirical research, potential optimized solutions might involve tailored polarization hysteresis, reduced trap density, or innovative stack architectures.

Materials-Related Challenges

Throughout working group deliberations, several materials-related challenges emerged, listed below:

- Control/selection of the desired phase of a ferroelectric material during device processing and field cycling.
- BEOL-compatible transistors to support monolithic 3D integration.

- Contact between channel material and ferroelectric (FeFET).
- Contact between metal electrode and ferroelectric material.
- CMOS-compatible ferroelectric materials.
- Leakage at ferroelectric domain walls.

Challenges associated with controlling and selecting the desired phase of the ferroelectric material are related to those of BEOL-compatible transistors. To enable monolithic 3D integration, FeFET process flow temperatures must be kept below 400°C (for those using conventional rapid thermal annealing), mitigating any deleterious effects on FEOL transistors and structures. Above this temperature, electro-migration, damage to underlying dielectric materials, and changes in dopant profiles become a concern. While the deposition temperature of the ferroelectric material commonly deposited by ALD for hafnia-oxide based ferroelectrics is below this limit, subsequent annealing to achieve the desired phase for ferroelectricity via crystallization—typically through wafer-scale rapid thermal annealing— is above 400°C. Any further processing after this phase has been established must be below the annealing temperature so the ferroelectric material retains its characteristics and does not revert to a more stable phase.

New approaches (e.g., new materials, localized annealing, and/or dopants) are needed to address these challenges. A Stanford-SLAC project is developing a holistic BEOL-compatible, ML-guided process integration approach to control the HfO_2-ZrO_2 (HZO) ferroelectric phase at temperatures compatible with CMOS-BEOL integration. This approach includes a novel, non-equilibrium flash annealer; electrical characterization (e.g., endurance and fatigue) and structural characterization (e.g., using XRD and TEM); real-time x-ray synchrotron measurements of behavior; and ML-assisted process exploration (Karigerasi et al. 2022; Biswas et al. 2021). Significant attention is being given to organic ferroelectric memory as well (Asadi 2010) since it is considered to be very scalable (Blinov et al. 2000).

Contact between channel and ferroelectric, and between metal and ferroelectric, may be addressed through fundamental interfacial studies and experimental approaches. Modeling and characterization can provide fundamental understanding of interfacial phenomena and structural information to inform and guide experimental approaches. Experimental approaches can include co-optimizing gate electrode material and ALD growth conditions, as well as novel post-processing techniques (e.g., localized annealing) to reach the desired characteristics.

Before the discovery of hafnia-based and organic ferroelectrics (Zheng et al. 2009), ferroelectric materials (primarily PZT-based perovskites) were not CMOS-compatible. As noted in the introduction to this chapter, given the existing CMOS infrastructure, any viable future solution for ferroelectric-based devices must be CMOS-compatible.

Leakage at domain walls, caused by structural defects in the material, may lead to performance and efficiency degradation in oxide ferroelectric devices. For example, it is thought that these play a role in wake-up phenomenon, fatigue, and delay. However, because of the complexity of the ferroelectric material systems, including meta-stable phases and the transformation between these phases under external field, this is an area of intense research with many open questions as to its true effects on ferroelectric device performance (Saini et al. 2023; Stolichnov et al. 2018; S. Zhang et al. 2023; Lee et al. 2020).

Action Plan for Ferroelectric Memory/FeFETs

Table 15. Action Plan for Ferroelectric Memory/FeFETs.

Scope			
Technology for Energy Efficiency:	FeFET		
Technology of Interest:	Memory		
Challenges		**Solution Pathways**	
• Overcome volatility and multi-state storage limitations in 3D integrated non-volatile memory systems. • Achieve high density and low switching energy while ensuring CMOS logic voltage compatibility. • Enhance endurance and power efficiency for deeply scaled memory technologies. • Address retention and rad-hard issues of 3D monolithic integration.		• Develop high-mobility, disorder-tolerant oxide semiconductors with thin ferroelectric layers for improved endurance. • Innovate ALD deposition techniques and materials, like HfO2-based ferroelectrics, for better scalability and stability. • Optimize thermal processing through advanced crystallization techniques. • Focus on R&D for process-property relationships to fine-tune ferroelectric capacitors and FeFETs for deep scaling.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline (years)**
Enhance the interface between ferroelectric materials and channel for durable cycle life.	Electric field cycling	10^{10} to 10^{12} cycles	5 years for initial improvements, 5–10 years for advanced targets
Innovate materials and design for lower write voltages and improved channel material stability.	Required voltage for reliable switching	1.2V and 0.7V	5–10 years, with current technology at 1.5V
Develop solutions for reducing write energy in ferroelectric devices.	Energy consumed per bit during write operations	Less than 1 fJ/bit	5–10
Advance deposition methods and optimize thermal processing for FeFET integration.	Structural and electrical properties of ferroelectric transistors		5–10
Improve the electrical characteristics of ferroelectric devices for efficient operation.	Electrical performance parameters like transconductance and Hall measurements	--Subthreshold swing of 65–70 mV/decade --Carrier mobility to n-type 50 cm^2/Vs at Vt stress + 1.2V --Reduce hysteresis to below 20 mV	2–5 for subthreshold swing 5 for n-type 10 for p-type 2–5 for reduction hysteresis
Stakeholders and Potential Roles in Project			
Stakeholder	**Role**		
Product Manufacturers/Suppliers	• Provide advanced materials and devices structures for testing. • Utilize extensive contact with industry partners.		
End Users/OEMs	• Innovate metrology tools specific for ferroelectric memory. • Define performance requirements and validation protocols. • Collaborate on testbeds for ferroelectric memory applications.		
Academia	• Develop new materials. • Research into new characterization methods for ferroelectric materials. • Develop AI/ML techniques for predictive metrology.		

Required Resources	Cross Collaboration Needs of Working Groups
• Access to advanced metrology equipment and facilities. • Funding for long-term research and development. • Collaboration platforms between industry, academia, and national labs.	• APHI: Focus on 3D device structure to integrate ferroelectric memory. • Circuits and Architectures: Develop new computing paradigms to ensure compatibility with next-gen computer near memory architectures and neurocomputing chips. • Algorithms and Software: Facilitate configurable hardware for efficient algorithm mapping, maximizing the energy efficiency gains in computational applications. • MEES: Adopt low-temperature synthesis techniques like flash lamp annealing to enhance fabrication process.

2.1.5 Tunnel Field-Effect Transistors

Subthreshold Slope Sharpening Transistor Technologies

A major impetus for exploring Beyond-CMOS technologies such as alternative switching methods (e.g., quantum tunnelling) is their potential to achieve an order of magnitude (~10x) increase in energy efficiency by steepening the subthreshold slope that defines the transition between the transistor's off and on states. In a MOSFET, the subthreshold swing (SS), the inverse of the subthreshold slope, is usually limited to 60 mV/decade at room temperature (see Figure 22). This limitation is often referred to as the Boltzmann tyranny (Pananakakis et al. 2023).

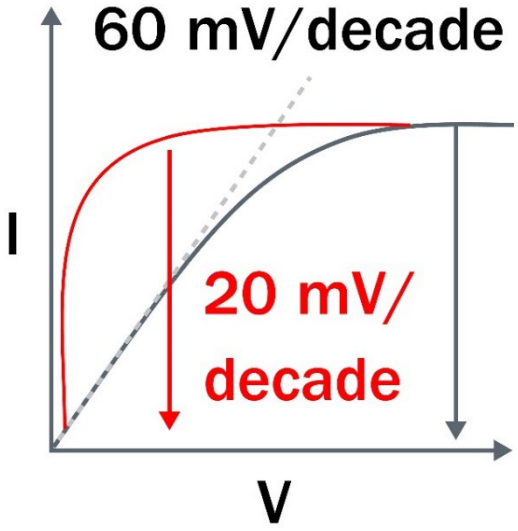


Figure 22. Subthreshold Slope of I/V on/off curve for typical FET and subthreshold sharpening tech (e.g. TFET). *Source: Cristoloveanu et al. 2016*

At the same current, steepening the slope of I-V curve for on/off reduces the switching voltage. Additionally, because the power used by the transistor is proportional to the square of the voltage, reducing voltage is a powerful energy efficiency lever (e.g., reducing voltage by a factor of 3 reduces the power by 9 times). However, during the roadmap process, it became clear that Beyond-CMOS switching technologies were not the only way to sharpen the subthreshold slope. For example, since this limit is related to thermal excitation of electrons in the MOSFET at room temperature, the 60 mV/decade limit also can be overcome in conventional CMOS by running the transistor at cryogenic temperatures (Södergren et al. 2023). Some academic research further shows that ultrathin dielectrics and ultrathin devices in general may enable better control of mobile charge and current, thus also steepening the slope (Cristoloveanu and Ghibaudo 2022).

TFET

Tunnel field-effect transistors (TFETs) are a promising alternative to traditional MOSFETs for continuing to decrease the voltage of operation, thus improving energy efficiency (Seabaugh

and Zhang 2010). In recent years, TFET has advanced rapidly as scientists focus on optimizing their performance and overcoming existing limitations. Most notably, the breakthrough to achieve steeper subthreshold slopes and lower operating voltages came from successful integration of novel materials, such as 2D materials and heterojunctions (Kanungo et al. 2022).

Compared with MOSFETs, TFETs rely on a fundamentally different mechanism for carrier transport: MOSFETs switch by modulating thermionic emission over a barrier, where thermal excitation of carriers limits the steepness of the turn-on current. TFETs, on the other hand, switch by modulating quantum tunneling *through* a barrier. Tunneling is enabled by the overlap of electron-like and hole-like wavefunctions through an energy barrier. The successful transmission through the energy barrier is dependent on the mass of the particle, the thickness of the barrier, and relative energy levels. Notably, this process is independent of thermal excitation, allowing TFETs to operate with much steeper turn-on, which, in turn, enables lower operating voltages than MOSFETs and significant energy savings.

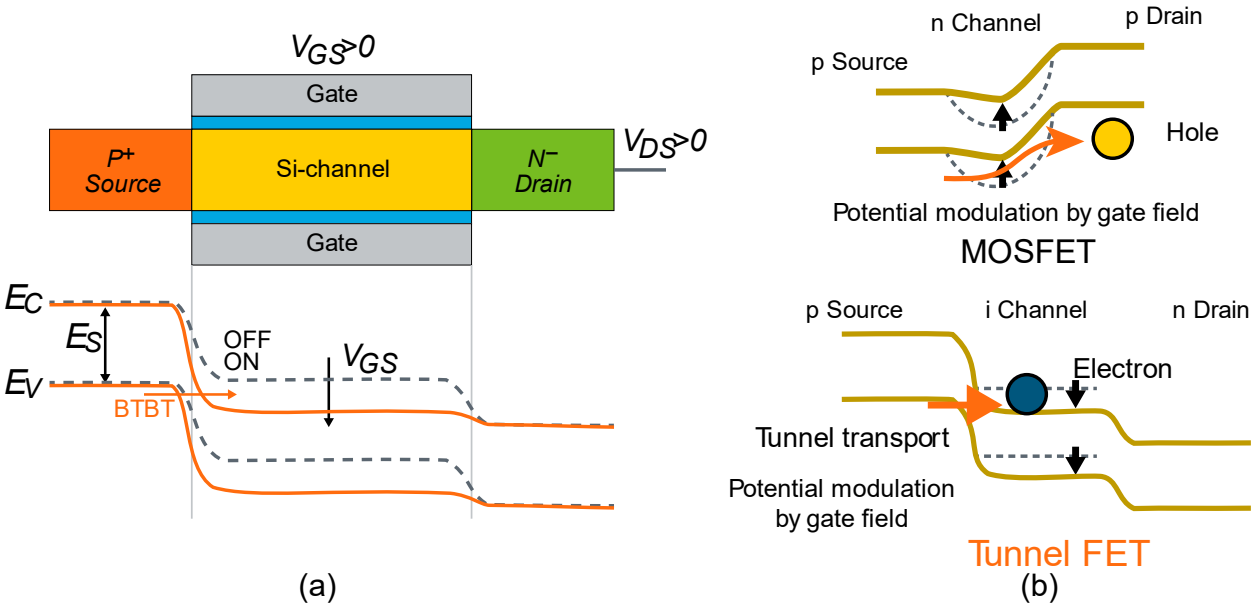


Figure 23. Device operation of a tunnel field-effect transistor (TFET). *(a) the schematic of a TFET cross section and band diagram along the channel in on and off states and (b) comparison between TEFET and MOSFET operation principle. Source: Agha et al. 2021*

For conventional silicon FETs, subthreshold slope is fundamentally limited by thermal energy fluctuations and is fixed at 60 mV/decade. Conversely, because carrier transport in TFETs is ultimately dependent on its wavefunction and not constrained by thermal energy, TFETs can theoretically achieve significant improvements, with variables like band mass and defect density limiting the subthreshold slope (Lu and Seabaugh 2014). Taking a modestly aggressive

subthreshold slope of 20 mV/decade, this translates to roughly 10 times improvement in power consumption and energy efficiency compared with conventional FETs (IRDS 2020).

Table 16. Energy Impact and Timeline Estimates[a] for TFETs

Technology	Expected Performance	Commercial Benchmark Product	Commercial Benchmark	Energy Impact Factor	Timeline for TRL 6
TFET	0.5 fJ/bit	5nm Si FinFET	7.02 fJ/bit	14	5 years

[a] Source: Huang et al. 2017

However, several challenges must be addressed for TFETs to achieve their theoretical performance and be a viable alternative to incumbent technology. These challenges can be grouped into two categories: device performance and manufacturability. The same band-to-band tunneling (BTBT) mechanism that enables TFETs to have 10 times improvement in energy efficiency introduces low ON-state current. And the fabrication of high-performance devices has, thus far, relied on unconventional means, with conventional manufacturing methods having yielded only poor-performance devices, which highlights the challenges of developing a process suitable for conventional CMOS processes and equipment.

To date, there have been limited successful demonstrations of TFETs that meet targets for on/off ratio, on-state current, and threshold voltage. Innovations in materials (Nazir, Rehman, and Park 2020), device structure, and manufacturing approaches are needed for TFETs to truly be a viable alternative.

Challenges and Solution Pathways for Tunnel Field-Effect Transistors

Low On-State Current

Since transistor speed is determined by current density, TFETs must have comparable current to MOSFETs while continuing to maintain a steep subthreshold slope to be a competitive alternative technology. To enhance the on-state current in TFETs, various techniques have been proposed, including gate and spacer engineering, band engineering, and innovative TFET structures. Examples of innovative TFET structures include vertical TFET, stacked gate junction-less TFET, and SOI-TFET with interface trap charges (Choi and Lee 2010; Eyvazi and Karami 2020; Rahi, Asthana, and Gupta 2017; Mitra and Bhowmick 2019; Kao et al. 2012).

Traditional approaches to enhancing the on-state current have focused on improved electrostatic design and band engineering in a traditional transistor geometry. Gate engineering approaches enhance on-state current by leveraging a multi-metal gate to improve electrostatic control over the tunneling interface. This approach effectively changes the work function along the channel length and modulates the distance between the conduction band of the channel and the valence band of the source (Nigam, Kondekar, and Sharma 2016; Kumar et al. 2020). By comparison, spacer engineering entails separation of the gate terminal from the drain and source regions using spacers. Coupled with high-K value, spacers reduce channel resistance, thereby improving on-state current. Band engineering, through heterostructure and material design, can be used from the energetics to the carrier masses to manipulate the quantum mechanical variables underlying band-to-band tunneling.

While these approaches can improve on-state current for traditional horizontal TFETs, they will always have less current than MOSFETs of the same geometry. Conventional MOSFETs have an uninterrupted channel where charge carries flow between source and drain. TFETs, on the other hand, rely on tunneling through a barrier, across an *interrupted* channel and, therefore, will always sustain less current given other operating constraints (e.g., off-state current).

Some recent research focuses on vertical geometry that enables many parallel tunneling channels across the gate length, increasing the total current of the device. These device structures universally rely on tunneling between a buried layer of charge and a gated surface layer, so the tunneling region is not only the width of the channel but also the length of the gate, effectively doubling the relevant dimensionality (Revelant et al. 2014). This geometry introduces requirements for even stronger electrostatic control by leveraging quantum confinement to define the buried layer, such as with a stack of 2D materials (Kanungo et al. 2022). An alternative approach is being explored in a Sandia National Laboratories project that couples atomically precise manufacturing techniques with a vertical TFET design to simultaneously improve subthreshold slope and device current. An atomically abrupt 2D layer of dopants is created at the surface of silicon to define the source contract using a process called atomically precise advanced manufacturing (APAM). The source is then buried in intrinsic silicon to define the channel and a gated drain layer is created over it.

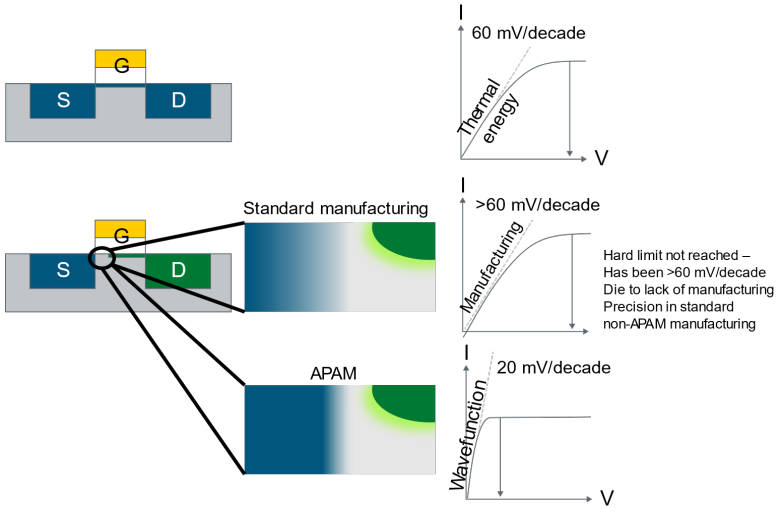


Figure 24. Enhanced ON current and subthreshold slope with atomically precise advanced manufacturing (APAM). *Source: Kaarsberg, Misra, and Shimizu 2023*

Subthreshold Slope

Though TFETs can theoretically achieve subthreshold slopes that exceed Si MOSFETs, this has been hard to achieve in practice (Avci, Morris, and Young 2015), primarily due to limitations in standard manufacturing. Using conventional manufacturing techniques, the doping profile of the source is diffused, due to diffusion during activation. This diffuse dopant profile essentially produces an overlapping sequence of turn-on currents corresponding to different dopant densities. The first of these may have a very steep subthreshold slope but supports very little current. Those later in the sequence may support significant current but have a subthreshold slope that is masked by the earlier curves. These sequences combine to produce a poor subthreshold slope that can be much worse than the thermal limit for MOSFETs.

Alternatively, processes like band-to-trap tunneling compete with band-to-band tunneling and can also produce overlapping turn-on curves that deteriorate the subthreshold slope or produce unacceptably high off-state current. Understanding the role of defects in TFETs is a significant area of need, starting with the metrology required to monitor their creation and evolution in fabrication.

APAM, as previously noted, can fabricate abrupt doping profiles to mitigate this issue. Layered heterostructures of 2D materials take an alternative route to providing abrupt charge profiles. However, other solution pathways are needed and may again rely on numerous engineering approaches to the gate, source, and dielectric regions.

Manufacturability

Manufacturability is also becoming an issue central to TFET development, such as wafer-scale uniformity of tunnel junction formation. Device-to-device variations across the wafer will result in variations in electrical characteristics, including subthreshold slope and ON-state current, resulting in poor yield, unexpected device performance, and challenges with circuit design. Modeling and experimental validation can mitigate some of these issues and will be especially important for transitioning to high-volume manufacturing. Extensive measurements from actual devices can help develop, feed into, and refine process and device models. A key area of need is the development of tunnel junction-specific metrology so that processes can be refined and monitored.

In other situations, a needed tool or process may not yet exist. Taking APAM as an example, transitioning this process to high-volume manufacturing will require the development of a tool that can accomplish the surface cleaning, doping, and silicon capping at high throughput and wafer scale. Development of a tool to support APAM TFET with unknown commercialization prospects is a hard sell. However, identifying more near-term applications can jumpstart engagement with tool developers.

While incumbent manufacturing approaches have yielded poor-performance devices, manufacturing/fabrication innovations, like APAM and other previously discussed engineering approaches seek to overcome these limitations. However, the best performing devices still require the incorporation of the many innovations made in high-volume manufacturing with these new techniques. Thus, a significant milestone is that these new techniques be manufacturable at the wafer scale and compatible with existing infrastructure, or they will ultimately become dead ends.

To date, there is no clear winner in the TFET process flow (e.g., materials, structures), so the process requirements to establish a fab process are not yet known. Efforts up to now have primarily focused on improving device performance at the lab scale, but practical factors like manufacturability, lab-to-fab transition, and process development/integration now need to be considered.

Action Plan for TFETs

Table 17. Action Plan for TFETs

Scope			
Technology for Energy Efficiency:	TFET		
Technology of Interest:	Logic		
Challenges		**Solution Pathways**	
• Increase ON-state current. • Achieve competitive subthreshold slope (<60 mV/dec). • Develop processes that are compatible with existing CMOS manufacturing processes.		• Explore engineering approaches (e.g., gate and spacer engineering), and structural modifications (e.g., vertical TFET), to improve electrostatic control and ON-state current. • Leverage new techniques (e.g., atomically precise advanced manufacturing) for abrupt dopant profiles to reduce subthreshold slope. • Achieve wafer-scale uniformity in tunnel junction formation for consistent device performance.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline (years)**
Validation of device model for TFETs	Material parameters	validation of both modeling and experimental	2–3
Control junction abruptness and doping densities limits	eV/nm	>0.1	2–3
Establish good DC device metrics	High ON current: μA/μm	>100	4
	Low OFF current: nA/μm	<1	4
	Low SS slope: mV/decade	<20	4
Establish good high speed circuit metrics	Operating voltage & speed: V, ps	<0.3, <100	6
Assess scalability of TFETs	Area	<50 nm^2	6
Feasibility of TFET-based memory	Equal state retention to CMOS SRAM	retention rate matching CMOS SRAM	6
Stakeholders and Potential Roles in Project			
Stakeholder	**Role**		
Product Manufacturers/Suppliers	• Develop TFET-specific fabrication tools and processes.		
End Users/OEMs	• Integrate TFETs into low-power devices and systems.		
Academia	• Engage in fundamental research on TFET materials, interfaces, and device physics.		
Required Resources		**Cross Collaboration Needs of Working Groups**	
• Device models validated by experiments. • Combine novel techniques with cutting edge fab tools. • Engagement between academia/national labs and industry.		• Materials and Devices: Develop novel TFET materials. • Circuits and Architectures: Integrate TFETs into existing systems. • Metrology and Benchmarking: Tailor measurement techniques to TFETs.	

2.1.6 Silicon Gate-All-Around Transistors

Gate-all-around (GAA) transistors build on the successes of classical, two-dimensional planar transistors, as well as the more recently dominant fin field-effect transistors (FinFETs). Whereas planar designs had the transistor's gate positioned along one side of the channel it modulates,

and FinFETs improved upon this approach by wrapping around a raised fin-like channel on three sides, GAA transistors exhibit gate designs that fully wrap around the device's channels. A general schematic of all three transistor types is presented in Figure 25 to show the different relationships between their respective gates and channels.

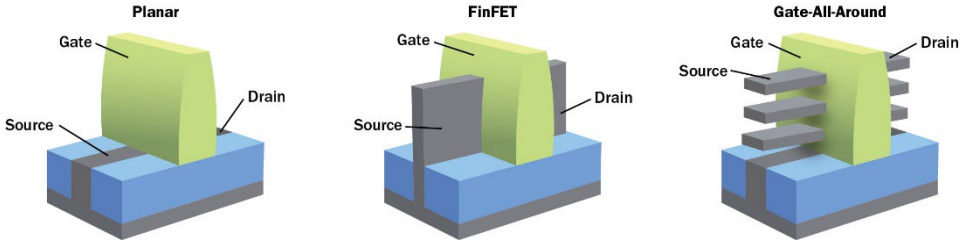


Figure 25. Typical source, drain, and gate arrangements for planar, FinFET, and GAA transistors. *Source: Semiconductor Engineering 2023*

While specific design details of GAA transistors can differ somewhat by manufacturer, designs generally involve silicon nanosheets or nanowires stacked vertically and passing through a high-k metal gate, such that the gate surrounds these small channels on all sides. These stacked nanosheets/wires are formed through alternating layers of epitaxially grown silicon (Si) and SiGe, with the latter layers containing only small concentrations of germanium. These interspersed SiGe layers are then selectively etched away later in the fabrication process and replaced by the transistor's high-k dielectric metal gate (Mukesh and Zhang 2022). The thicknesses and widths of these layers of Si and SiGe can be carefully controlled, allowing manufacturers to vary their designs to optimize for properties such as higher drive current (using wider nanosheets) or more energy-efficient power consumption (using narrower nanosheets) (Hofman 2022). Multiple stacked nanosheets/wires must be implemented within a GAA transistor's design to gain performance advantages over current FinFET designs, and some of the challenges that result from trying to create these very precise stacks are discussed further below.

However, the all-around design of these gates in GAA transistors can afford significant benefits and scaling advantages relative to current FinFET designs. GAAs are in the process of superseding FinFETs as the dominant technology for high-performance logic, offering benefits such as modest improvement in switching energy (Huang et al. 2017; Barraud et al. 2017) (Table 18), faster switching speeds, lower current/power usage, greater transistor density, and reduced channel leakage (Alcorn 2023). The channel thicknesses created in FinFETs are defined (and, in turn, limited) by lithographic resolution, while the gate structure of GAA designs affords greater control over the channel and more opportunities for channel length scaling—and thus greater potential transistor density (Singh 2021), allowing for continued dimensional improvements. GAA transistors can also be manufactured at an acceptable price point for chipmakers, such that they are expected to see wide-ranging application in—among other applications--AI systems, gaming, graphics, medical and automotive technologies, and advanced 5G networks.

Table 18. Energy Impact and Timeline Estimates[a] for Si-GAA

Technology	Expected Performance	Commercial Benchmark Product	Commercial Benchmark	Energy Impact Factor	Timeline for TRL 6
5nm Si-GAA	6.94 fJ/bit	5nm Si FinFET	7.02 fJ/bit	1.1	Current

[a] *Source: Huang et al. 2017*

While FinFETs have played a leading role in high-performance logic through the 2010s up to the present day, there are limits as to how tall the fins can be and how many can be placed next to one another without negative electrical effects (Hofman 2022). For example, 3nm FinFETs have been fabricated, but there are prohibitive issues with current leakage and short-channel effects as FinFET devices get progressively smaller while attempting to continue the dimensional scaling of Moore's law (Singh 2021).

Moving from FinFETs' fin-based design to stacked, fully surrounded nanosheets has been done to help mitigate these electrical effects, and GAA technologies have been in development for decades. Toshiba demonstrated the first GAA transistor back in 1988, called the Surrounding Gate Transistor, and IBM has been working on their GAA devices and the accompanying nanosheet technology for over a decade (Singh 2021). But the first significant performance benchmarking of GAA transistors has come out in the past five years (Mukesh and Zhang 2022).

The latest *International Roadmap for Devices and Systems* (IRDS™) confirmed IEEE's earlier prediction that FinFETs would gradually be usurped by GAA devices in high-performance logic, with a transition beginning around 2022 and expected to be fully realized in 2025 (IRDS 2022). Samsung began 3nm chip production using their Multi-Bridge-Channel FET (MBCFET™) GAA technology around mid-2022 (Samsung Semiconductor 2022). A recent IMEC roadmap for transistors projects a transition timeline like that of IEEE's IRDS, expecting higher-volume GAA production from Samsung and Intel in 2024, followed by TSMC's GAA production in 2025 (Alcorn 2023).

Both the IRDS and IMEC roadmaps expect GAA transistors to be a crucial component of at least the next few generations of logic. Other technologies with potential to play a role further down the line include complementary FET (CFET) transistors (Alcorn 2023), vertical-transport FETs, and stacked transistors (Mukesh and Zhang 2022). And though the default channel material in GAA devices is silicon, there are other semiconductive materials under review that could eventually play a role within GAA transistors and/or other transistor technologies. Examples of other semiconductive materials under consideration include InGAAs and other III-V semiconductors (Semiconductor Engineering 2023), molybdenum disulfide, graphene, and indium oxide (Mukesh and Zhang 2022).

Challenges and Solution Pathways for Si-GAA

Since Si-GAA was identified as a promising energy-efficient technology, the discussion below is primarily based on a review of existing work due to the high motivation and intensive R&D already occurring in industry.

Many of the pathways for development of GAA transistors have been relatively well-established from preceding FinFET designs. Common components between GAAs and FinFETs include the shallow trench isolation, high-k metal gate, source/drain epitaxial elements (Mukesh and Zhang 2022), and pillar patterning. Initial fabrication of GAA's alternating Si and SiGe nanosheets/nanowires is considered generally straightforward. For the SiGe layers, decreasing concentrations of germanium helps to minimize defects such as lattice distortion, but increasing concentrations of germanium makes it easier to etch these layers away later in the process and limit erosion of the purely silicon nanosheets (Semiconductor Engineering 2023). Most current challenges with GAA technologies stem from this tradeoff in the latter steps in the production process, such as etching away the SiGe layers from between the Si channels and depositing the gate's high-k metal and dielectric materials.

"A Review of the Gate-All-Around Nanosheet FET Process Opportunities" by Sagarika Mukesh and Jingyun Zhang at IBM Research provides a thorough picture of GAA transistors' remaining technological challenges, which are summarized below.

The "fat-fin" effect of GAA nanosheets (known as sub-fin leakage for FinFETs) results from an increased capacitance in the area below the stacked nanosheets. This effect is generally mitigated by adding a SiGe layer with a higher concentration of germanium at the bottom of the stack and then selectively etching it away and replacing it with a full bottom dielectric isolation layer to minimize channel leakage. The "narrow sheet effect," in contrast, results from the thinness of the silicon nanosheets and involves a decreased mobility of electrons/holes due to combinations of phonon scattering and surface roughness. These narrow sheet effects can typically be offset by increasing the nanosheet's width. Similarly, a third challenge—accommodating multiple threshold voltages, as generally required by industry—results from the minimal space between these stacked nanosheets that is available to deposit work function metals. Proposed solutions include alternative etching methods and/or increasing the nanosheets' spacing.

Overall, GAA architecture includes various "unique design knobs" that allow for manufacturers to negotiate these various performance tradeoffs. Processing challenges for these layered nanosheets were categorized by researchers as either mechanical stability, device variability, thermal intermixing, or self-heating effects. For self-heating, research into these effects is ongoing (and includes novel substrates like diamond on silicon), but solutions more appropriate to high-volume production are still under investigation as the technology proceeds into smaller and smaller dimensions.

Action Plan for Si-GAA

Table 19. Action Plan for Si-GAA

Scope		
Technology for Energy Efficiency:	Si-GAA	
Technology of Interest:	Logic	
Challenges		**Solution Pathways**

• Manage "fat-fin" effects to minimize channel leakage. • Counter narrow sheet effects impacting electron/hole mobility. • Accommodate multiple threshold voltages within tight nanosheet spacings. • Address mechanical stability, device variability, thermal intermixing, and self-heating effects during processing.	• Utilize higher germanium concentration SiGe layers for ease of selective etching. • Offset narrow sheet effects by adjusting nanosheet width. • Explore alternative etching methods and increasing nanosheet spacing. • Continue research into novel substrates and high-volume production solutions for thermal management.

Major Tasks/Milestones	Metrics	Targets	Timeline (years)
Transition to GAA transistors for high-performance logic	X Factor	1.6x improvement over FinFET	Immediate to 5
Addressing "fat-fin" and "narrow sheet" effects	Leakage	Minimize channel leakage	Immediate to 5
Achieving multiple threshold voltages with tight nanosheet spacings	Voltage control	Multiple threshold voltages within GAA design	Immediate to 5
Mitigating self-heating and thermal management challenges	Thermal management	Effective heat dissipation	Immediate to 5
Integration of GAA transistors into industry applications	Integration success	Widespread application in AI, medical, and automotive technologies	Immediate to 5

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Develop and optimize GAA transistor manufacturing.
End Users/OEMs	• Implement GAA transistors in high-performance logic applications.
Academia	• Research options for overcoming physical challenges and device variability.

Required Resources	Cross Collaboration Needs of Working Groups
• Advanced material synthesis facilities. • High-precision etching and patterning equipment. • Novel substrate materials for thermal management studies.	• APHI, Circuits and Architectures, and Materials and Devices: Address integration and performance optimization.

2.1.7 Emerging Devices and Materials for Analog Computing

To bridge the gap between conventional computing and neuromorphic computing, analog devices emerge as the prime technology choice. Analog devices differ fundamentally from their digital counterparts in that they process continuous signal values. Unlike digital devices that encode information into discrete states, typically represented by 0s and 1s, analog devices can handle an infinite range of values and provide a more natural and efficient way of simulating biological neural networks. They also have the potential to be integrated seamlessly with current CMOS technology, adding the capability for brain-like computation and storage within a single unit.

While there are numerous types of analog devices, this discussion will focus on several key types that have been the subject of extensive deliberation among the working group. These devices—memristors, organic semiconductors (OSCs), and mixed ion-electron conductors (MIECs)—are integral to advancing the capabilities of analog circuits within neuromorphic computing. Memristors, with their ability to remember previous states of electrical resistance, serve as the cornerstone for creating artificial synapses, thereby enabling the emulation of synaptic plasticity critical for learning and memory in neuromorphic systems. OSCs contribute to

the analog device landscape with their flexibility, low cost, and biocompatibility, which are advantageous for organic neuromorphic circuits that require low-power, flexible computing substrates. Meanwhile, MIECs offer a dual conduction mechanism that is pivotal for memristive devices due to their ability to emulate the ionic motion akin to biological synapses, thus adding another layer of biomimicry to analog neuromorphic devices. These components are integral to the analog paradigm, each addressing different aspects of the neuromorphic challenge and collectively moving the field closer to achieving brain-like computational efficiency within a silicon-based technology framework.

The integration of analog devices in neuromorphic computing is not just a matter of transplanting existing technology into new applications. It requires a fundamental rethinking of device architecture and operation to harness the full potential of analog computation. As we explore these new horizons, the Circuits and Architecture section of the roadmap provides a more detailed examination of how neuromorphic devices can contribute to system-level energy savings and the broader implications for higher-level systems and architectures.

Memristor

Silicon-based devices struggle to replicate the complex dynamics of biological processes efficiently. Two of these challenges are of note: 1) the number of connections between individual switching elements: a neuron has thousands of connections while a transistor only has two, and 2) storage of data: conventional silicon devices store data separately from logic operations, whereas neurons perform both computation and storage functions (U.S. Department of Energy, Office of Energy Efficiency & Renewable Energy 2021). In response to these limitations of silicon-based devices, the development of memristive device represents a pivotal shift towards emulating the more intricate and energy-efficient functionalities of the human brain, bridging the gap between traditional computing architectures and the dynamic capabilities of biological neural networks.

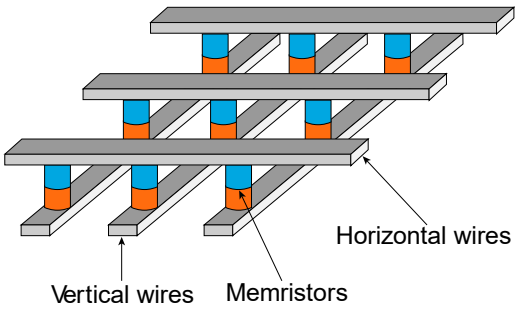


Figure 26. Basic schematic of memristors in crossbar arrays. *Source: Yadav et al. 2023*

A memristive device (or memristor) is a two-terminal electronic component that regulates the flow of electric current in a circuit and remembers the amount of change that has previously flowed through it (Yang, Strukov, and Stewart 2013). The key property of a memristor is its ability to retain its resistance state even when power is turned off (non-volatility), which makes it an attractive device for neuromorphic computing applications (Xiao et al. 2023). In biological systems, synaptic weights between neurons adjust over time based on activity, a process that underlies learning and memory. The resistance states can be adjusted to mimic synaptic weights, and the ability of memristors to change and remember these states can be used to simulate synaptic plasticity, the strengthening or weakening of synapses based on activity. For these reasons, memristors, typically in crossbar arrays (see Figure 26), are mainly studied as an analog device for neuromorphic computing.

Expected performance for memristor devices leveraging novel materials are typically on the order of 1 fJ/switch or spike (Zhu et al. 2020). Typical silicon-based neuromorphic systems

utilize FinFET technology in GPUs (see Table 20). While a device-level comparison indicates only modest energy impact, the true energy-efficiency impact is derived from the neuromorphic computing architecture that these devices enable. A more complete discussion of neuromorphic computing is provided in the Circuits and Architectures section.

Table 20. Device-Level Energy Impact and Timeline Estimates[a] for Analog Devices for Neuromorphic Computing.

Technology	Expected Performance	Commercial Benchmark Product	Commercial Benchmark	Energy Impact Factor	Timeline for TRL 6
Analog for Neuromorphic	1 fJ/bit	5nm Si FinFET	7.02 fJ/bit	7	10 years

[a] Source: Huang et al. 2017

A broad range of materials with different maturity levels have been explored, some of which are summarized below.

Organic materials

Owing to their ample free volume, organic semiconductors (OSCs) are characterized by their low switching energies, remarkable tunability, and efficient ion migration. Central to their appeal for neuromorphic computing is their ability to emulate neuroplasticity at a single unit level, with a wide range of synaptic switching mechanisms. These mechanisms range from two-terminal devices employing filament formation and charge trapping to advanced three-terminal systems such as ion-gated electrochemical transistors (van de Burgt et al. 2017).

However, OSCs for neuromorphic computing face several challenges. Speed optimization remains a top priority. The intrinsic rate limitations of OSCs, stemming from their slow charge carrier mobility, result in a longer response time compared to their inorganic counterparts. Endurance is also a concern due to repetitive conduction. Issues also arise in enhancing device density, especially given the incompatibilities between OSCs and certain solvents used in photolithography (Zakhidov et al. 2011). Integrating these organic devices with traditional binary digital systems presents further hurdles, mainly due to the low degradation temperature (>150°C) of OSCs, while the traditional nanofabrication process for annealing Cu interconnects requires ~400°C (Christensen et al. 2022). Environmental factors—such as exposure to moisture or oxygen—alongside intrinsic electronic stability issues, amplify these challenges (Keene et al. 2019).

By identifying and fine-tuning rate limitation of organic materials, the speed of these devices can be improved. To improve device density, novel fabrication processes have been proposed that are capable of accurately fabricating OSCs in vertical architectures (Lenz et al. 2019). Moreover, refining the crystallinity of OSCs, in tandem with advanced encapsulation techniques (see Figure 27), presents a compelling approach to mitigate stability issues (Keene et al. 2019; Go et al. 2020).

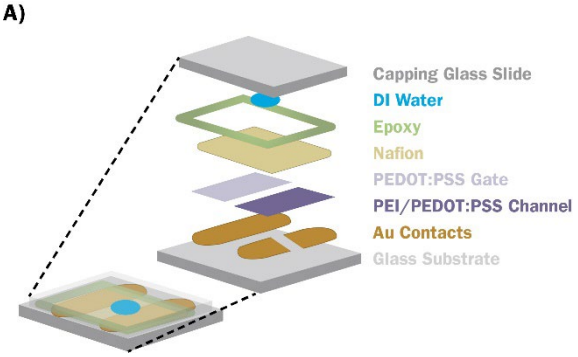


Figure 27. Novel encapsulation strategy.*Source: Keene et al. 2019*

Mixed ion-electron conductors (MIECs)

Mixed ion-electron conductors (MIECs)—most commonly oxides such as cerium oxide, amorphous gallium oxide, and lanthanum nickel oxide—are unique materials capable of simultaneously conducting ions and electrons. This dual conduction mechanism is particularly useful in memristors. By utilizing the ionic motion within MIECs, memristors can emulate the gradual strengthening and weakening of synaptic connections (Shenoy et al. 2014).

Challenges include device variability, arising from the inherent inhomogeneities and defects in MIEC materials, which leads to inconsistent device behavior (Narayanan et al. 2015). Additionally, the long-term stability and endurance of MIEC-based devices can be compromised due to repetitive ion movement, which may induce degradation or drift in the device's performance over time (Burr et al. 2013). Finally, the speed of ionic movement, in comparison to electron motion, can also limit the device's switching speed, potentially slowing down computations in neuromorphic circuits.

To address these challenges, dopants or novel MIEC compounds can be used to enhance ion mobility and reduce undesired defects (Liu and Wang 2020). Device architecture can also be optimized to mitigate degradation, for instance, by implementing protective barrier layers that minimize detrimental ion migration (Yoon, Oh, and Park 2022). Additionally, hybrid device designs, which combine the benefits of MIECs with other materials or mechanisms, provide pathways to harness the advantages of ionic conduction while offsetting its limitations (Maas et al. 2020).

Action Plan for Emerging Devices and Materials for Analog Computing

Table 21. Action Plan for Emerging Devices and Materials for Analog Computing

<table>
<tr><th colspan="4">Scope</th></tr>
<tr><td>Technology for Energy Efficiency:</td><td colspan="3">Emerging Devices and Materials for Analog Computing</td></tr>
<tr><td>Technology of Interest:</td><td colspan="3">Neuromorphic</td></tr>
<tr><th colspan="2">Challenges</th><th colspan="2">Solution Pathways</th></tr>
<tr><td colspan="2">• Identify and synthesize materials for analog devices compatible with CMOS processes.
• Reduce power usage to match biological systems' efficiency.
• Integrate analog devices for neuromorphic architectures.</td><td colspan="2">• Leverage foundry expertise in memory technologies like resistive random-access memory (RERAM) for neuromorphic device development.
• Create PDKs and leveraging multi-material accelerators for diverse computing applications.
• Advance spiking neural network implementations for real-time learning and adaptability.
• Enhance computational models to closely mirror the physics of biological systems, such as using carbon nanotube networks for processing.</td></tr>
<tr><th>Major Tasks/Milestones</th><th>Metrics</th><th>Targets</th><th>Timeline (years)</th></tr>
<tr><td>R&D for neuromorphic materials</td><td>Neuromorphic functionality such as multistate memory and nonlinear activation.</td><td>Identify memristive materials with switching energy < 100 aJ/bit and spiking network frequency > GHz</td><td>3–5</td></tr>
<tr><td>Commercialization Feasibility Analysis</td><td>Endurance, lifecycle, cost-effectiveness</td><td>Achieve >10^{10} cycles and reduce the cost to approximately $10/synaptic cycle</td><td>3–7</td></tr>
<tr><td>Large-Scale Neuromorphic System Integration</td><td>Scalability to simulate a large number of neurons</td><td>Implement systems with ~10^{11} synthetic neurons</td><td>5–10</td></tr>
</table>

Hybrid Integration with Existing Technologies	Integration efficacy with current tech	Develop high-performance devices for edge computing under 100 mW	5–10
High-Performance Computing (HPC) and Data Center Adoption	Neuromorphic computing at a massive scale	Integration in systems requiring >10^{15} synthetic neurons for advanced scientific computation	10–15

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Build up new infrastructure for fabrication and define requirements.
End Users/OEMs	• Implement neuromorphic computing solutions, provide feedback. • Develop new deposition, lithography, and metrology equipment.
Academia	• Increase R&D efforts on emerging materials.

Required Resources	Cross Collaboration Needs of Working Groups
• Collaboration between academia and industry. • Seed funding for startups. • EDA tools. • Open-source design tools.	• Circuits and Architectures: Define material properties/metrics requirements. • Algorithms and Software: Define how close to brain inspired computing is required vs more general distributed computing architecture. • APHI: Define metrics in thermal heat transfer and develop material integration methodology. • Metrology and Benchmarking: define metrics and develop metrology methods for this technology.

2.1.8 Novel Materials for Silicon Scaling

As contemporary CMOS technology continues to scale beyond 3nm, the parasitic delay and dissipation from conventional interconnect materials increasingly dominate overall transistor performance. At the same time, as feature sizes (e.g., trenches, vias) approach the limits of existing fabrication equipment, issues like electromigration and crosstalk become more problematic. Thus, there is intense research, and high industry motivation, into identifying novel materials and process schemes for the integration of contacts in the middle-of-line (MOL) processes and interconnects and intermetal dielectrics in the backend of line (BEOL) processes. Not only does this integration offer the prospect of performance enhancement and allow for smaller device dimensions, but it also offers potential energy efficiency improvements by reducing resistive loss and capacitive delay.

One of the most pressing challenges is determining which novel materials ensure CMOS compatibility and seamless integration. Given the extensive array of potential material options, down-selection is a daunting task. It's a complex problem, primarily because there are no clear "winners" in material choices; the suitability and tradeoffs often vary on a case-by-case basis depending on specific applications and technological frameworks. The subsequent sections summarize the challenges and prospective solutions for interlayer dielectrics (ILD), interconnects, and contacts.

2.1.8.1 Interlayer Dielectric (ILD)

The interlayer dielectric (ILD) is an insulating layer used between interconnect layers in the BEOL. The primary function of the ILD is to electrically isolate the layers to minimize crosstalk

and ensure accurate circuit behavior. Additionally, the ILD offers thermal management and structural support for further processing.

In conventional CMOS devices, ILD primarily consists of silicon dioxide (SiO_2) and its derivatives, such as fluorosilicate glass (FSG) or carbon-doped silicon oxide (CDO). These materials work well due to their reliable dielectric properties and CMOS compatibility. However, as device nodes advance and dimensions shrink, these materials are increasingly prone to breakdown under electric fields. Moreover, these materials' higher parasitic capacitance contributes signifiantly to the overall delay time and switching energy. Thus, there is a need for novel ILD materials with lower dielectric constants (κ-values).

Table 22. Important Properties for Materials in Low-κ Applications[a]

Structural	Electrical	Mechanical	Chemical
• Small, closed pores • Thickness uniformity • No channel continuity	• Low κ • Low leakage current • Low charge trapping • Low dielectric loss • High breakdown resistance	• High Young's modulus • High hardness • Low residual stress • High thickness threshold • High adhesion strength	• Low moisture absorption • No metal corrosion • No fluorine/chlorine loss • Etch selectivity • Good chemical/thermal stability

[a] Source: Hatton et al. 2006

Novel ILD Materials

Ideally, ILD materials have very low κ-value while also exhibiting the following charateristics: structural, thermal, and chemical integrity; sufficient hardness; a large band gap for minimal leakage; and compatibility with existing manufacturing processes. (Ryan et al. 2003). In reality, the industry transitioned from SiO_2 to various other materials (Table 23) over the past couple of decades that were sufficient, but not ideal, including FSG for the 180nm node and SiCOH for the 120nm and 90nm nodes. At the advanced nodes, the industry is focused on porous organo-silicon ILDs. Despite its favorable dielectric constant, process integration, as detailed below, is a major challenge. Further R&D to identify alternative ILD options and process integration pathways are needed.

Table 23. Dielectric Constants of Various Contemporary Low-κ Materials[a]

Classification	Material	Fabrication	Dielectric Constant (κ)
Silicon dioxide	$Si0_2$	CVD	3.9–4.5
Silsesquioxane-based	Hydrogen-Silsesquioxane (HSSQ)	Spin-on	2.9–3.2
	Methyl—Silsesquioxane (MSSQ)	Spin-on	2.6–2.8
Silica-based	FSG	CVD	3.2–4.0
	SiCOH	CVD	2.7–3.3
Porous	Porous HSSQ	Spin-on	1.7–2.2
	Porous MSSQ	Spin-on	1.8–2.2
	Porous SiCOH	Spin-on/CVD	1.5–2.5
Air gaps	Air	—	1.0

[a] Source: Sekhar 2012

Integration With CMOS

The transition from silicon dioxide to alternative ILD materials has added complexity to their integration process. Despite their low κ-values, ILDs are mechanically weak, lack thermal stability, and have diminished adhesive strength, making them susceptible to trapping chemicals and delaminating. This adhesive weakness often stems from a high carbon concentration during the PECVD process. One mitigation involves depositing an initial oxide layer to enhance the film's adhesion. However, the PECVD process often induces plasma-related damage (PID) that weakens the film both mechanically and thermally, making it more hydrophilic. The copper-integrating dual-damascene process is particularly vulnerable to introducing PID at multiple stages. To mitigate PID, alternative precursor and deposition techniques are being explored, with the pore stuffing method—employing materials like PMMA to shield the surface—emerging as a promising solution (see Figure 28) (Zhang et al. 2015).

Action Plan for Interlayer Dielectrics

Figure 28. Schematic of pore stuffing method. *Source: Zhang et al. 2015*

Table 24. Action Plan for Interlayer Dielectrics

Scope			
Technology for Energy Efficiency:	Interlayer dielectrics		
Technologies of Interest:	Novel materials for silicon-based logic devices		
Challenges		**Solution Pathways**	
• Achieve ultra-low κ-values while ensuring mechanical integrity and thermal stability. • Balance dielectric properties with structural and mechanical robustness. • Address integration challenges with CMOS processes, including thermal and plasma-induced damages. • Down-scale pore size without compromising dielectric properties.		• Develop novel materials and deposition methods to create stable, low-κ porous ILDs. • Optimize pore size and distribution to ensure structural integrity and low dielectric constants. • Innovate integration techniques to mitigate plasma-induced damage and improve film adhesion. • Explore the use of protective materials during fabrication to prevent damage to porous structures.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline (years)**
Discover novel oxide with air gaps	Experimental validation	Meet metal compatibility requirements	2+
Lab demonstration	Dielectric constant	κ < 2.5	5
Mechanical testing	Mechanical strength	> 4 Gpa	5
Dielectric breakdown analysis in capacitors	Electric field vs. thickness	High breakdown resistance	5
BEOL processing compatibility	Materials compatibility	Compatible with sub-400°C processes	5
Develop novel deposition methods	Deposition techniques	Suitable precursors for low-κ ILDs	5
Identify etching processes for ILD	Etching efficiency	Minimize defect and maintain low-κ	5
Test material robustness	Accelerated lifetime	Comparable to industry standards	5

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Develop and supply novel low-κ ILD materials. • Collaborate on the integration of these materials into existing fabrication lines.
End Users/OEMs	• Provide specification for device performance that drive the requirements for ILD material properties and feedback on the integration impact.
Academia	• Increase R&D on new ILD materials, explore innovative integration techniques, and contribute to understanding the science behind material behavior and process development.
Required Resources	**Cross Collaboration Needs of Working Groups**
• Collaboration between academia and industry. • National lab with EWD effort to bring in more experts. • Access to CMP resources.	• Circuits and Architectures: Define material properties and metrics requirements. • Metrology and Benchmark: Develop methods to research lower dimension materials.

2.1.8.2 Interconnects and Contacts

Contacts and interconnects form the backbone of multi-layered microelectronic chips, ensuring a coherent flow of data and power. Interconnects are the horizontal and vertical conductive pathways that link the various components on a chip, ensuring smooth communication and power distribution. Contacts connect the interconnects to the transistor switch itself. For different physical reasons, as the dimension of interconnects and contacts decreases, the dissipation associated with them increases. This dissipation was much less than that from the transistors for decades but has grown to be comparable to the dissipation in the transistor itself. The integration of novel materials into interconnects and contacts offers a promising avenue to not only address dissipation but also improve the overall performance, reliability, and longevity of next-generation microelectronic devices.

Novel Interconnects

The ohmic dissipation in metal interconnects necessitates the use of metals with lower resistivity and prompted the move from aluminum to copper decades ago. As the dimension of these interconnects shrink, grain boundaries and boundary scattering plays as much a role as the bulk resistivity and motivates the search for metals without grain problems and with a low mean free path. This issue has led to significant research into metals like ruthenium, which have slightly worse bulk resistivity than copper but with short mean free paths that limit the effect of boundary scattering. Ideal materials also need to be compatible with the dual damascene process, where a barrier protects the ILD, and the interconnect metal fills both the vertical vias and the horizontal trenches that form the wiring layer. The tightest geometry comes from the vertical part of the interconnect: the interlayer via, which is discussed below.

As silicon continues to scale, the liner/barrier's thickness in the interlayer via for Cu interconnects becomes the bottleneck for further miniaturization. Liners that are applied post-barrier enhance adhesion between the metal and the barrier, act as a precursor for subsequent metal deposition, and support electromigration resistance. Ta/TaN is the incumbent liner/barrier for the Cu dual damascene process, and TaN's barrier thickness of 0.8nm has been demonstrated without compromising its efficacy (Witt et al. 2018). Switching to Co or Ru from the Ta further improves TaN barrier integrity (Witt et al. 2018).

While a Ru/TaN combination can achieve linewidths down to 2nm, the drive for further miniaturization sparked interest in barrierless alternatives (Wu et al. 2018), including Co, Ru, Ir, Rh, Mo, and W. These metals enable both hybrid metallization and semi-damascene processes. Co and Ru have garnered substantial experimental validation. Because these processes enable higher aspect ratio lines, it reduces resistance, and in turn, improves energy efficiency. Figure 29 shows Cu's resistance becoming higher than Ru and Co at smaller dimensions (van der Veen et al. 2018).

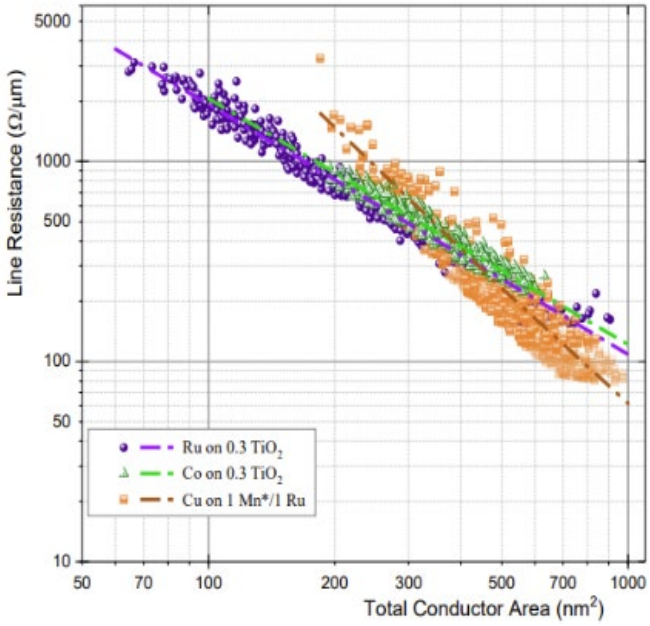


Figure 29. Logarithmic comparison of the damascene line resistance vs. the total conductor cross-sectional area of Ru, Co, and Cu nanowires. *Source: van der Veen et al. 2018*

Recent innovations have explored the potential of 2D materials, such as graphene and MoS_2, as alternatives to traditional barriers (Nogami et al. 2021; Lo et al. 2018). These materials may hold promise for enabling a new generation of metals that could surpass the performance of copper.

Action Plan for Novel Interconnects

Table 25. Action Plan for Novel Interconnects.

Scope			
Technology for Energy Efficiency:	Novel interconnects and vias		
Technologies of Interest:	Novel materials for silicon-based logic devices		
Challenges		**Solution Pathways**	
• Reduce energy consumption of traditional interconnects and improve signal transmission. • Address issues with the increased resistivity and reduced fill in copper vias as CMOS technology scales down to 3nm and beyond. • Integrate novel materials for interconnects and vias with existing production technologies.		• Explore new low-resistivity, high-thermal conductivity metals (e.g., ruthenium, molybdenum) and compounds for interconnects. • Investigate materials like graphene and TMDCs as alternatives to traditional barriers in vias. • Implement semi-damascene processes for via filling to achieve higher aspect ratio lines and reduce resistance.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline (years)**
Develop models to find alloys with optimal resistivity and thermal conductivity	Resistivity and Thermal Conductivity	Lower than known single-element metals	2
Differentiate grain, edge, and bulk resistance in nanowires	Grain, Edge, and Bulk Resistance	Techniques that provide clear resistance differentiation	2

Perform lab tests on nanowire resistance with potential low rho-lambda metals	Nanowire Resistance and Contact Resistance	< 200 Ohm-microns for contact resistance	5
New methods for depositing low-resistivity materials	Deposition Technique Efficiency	Successful integration with CVD/ALD	5
BEOL process compatibility with new interconnect materials	Compatibility with BEOL Oxides & Thermal Properties	High compatibility ratings	5
Test accelerated lifetime of new interconnect materials for robustness	Accelerated Lifetime Testing Results	Performance on par with current standards	5

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Develop new materials and processes for advanced interconnects and vias. • Supply the semiconductor industry with innovative materials and solutions that meet advanced performance specifications.
End Users/OEMs	• Provide specifications and performance requirements for new devices. • Integrate and test new interconnect and via technologies in finished products.
Academia	• Increase R&D effort on the properties of new materials. • Collaborate on developing new methodologies for material synthesis and integration and contribute to workforce education and training.

Required Resources	Cross Collaboration Needs of Working Groups
• Collaboration between academia and industry. • National lab with EWD effort to bring in more experts. • Access to innovative materials. • Investment in laboratory facilities, testing equipment, and simulation tools for material development and device integration.	• Circuits and Architectures: Define material properties/metrics requirements. • Metrology and Benchmark: Develop methods to research lower dimension materials.

Novel Contacts

Contacts refer to the regions where an interconnect makes a direct electrical connection to an active device region, such as the source, drain, and gate of a transistor. Contacts allow for the transfer of electrical signals and power between the transistor (or other active device) and the interconnecting metal layers. The ohmic dissipation from contacts depends on the area of the contact and becomes larger as transistor dimensions shrink. This dissipation is significant since contacts are the largest feature of transistors in leading technologies. Integration of novel contact materials has the potential to drastically reduce contact resistivity, enhance electron transport, and minimize leakage currents. Such advancements can lead to significant improvements in energy efficiency and overall device performance.

According to IRDS 2022, in some current technologies, the series resistance can degrade the saturation current by 40% (IRDS 2022). As the gate pitch continues to scale, the repercussions on the drive current due to external resistance are expected to intensify. This scaling, combined with the anticipated rise in interconnect resistance, requires a drastic decrease in device contact resistance. PMOS devices, which use holes as carriers, require metal contacts with a high work function to reduce the Schottky barrier for holes, whereas NMOS devices, which employ electrons as carriers, need metals with a lower work function. In practice, the Schottky barrier height (SBH) is set by the metal-induced gap states (MIGS) and not the metal work function.

A fundamental understanding of the variables underlying MIGS is needed to identify or engineer contact materials to overcome MIGS. Another pathway is to explore other contact materials—like semimetals and metal-insulator-semiconductor (MIS) contacts—that may be able to inhibit or mitigate MIGS altogether. Once promising materials or approaches are identified, practical challenges like process integration and compatibility must be considered.

Action Plan for Novel Contacts

Table 26. Action Plan for Novel Contacts.

Scope			
Technology for Energy Efficiency:	Novel contacts		
Technologies of Interest:	Novel materials for silicon-based logic devices		
Challenges		**Solution Pathways**	
• Minimize contact resistivity as device dimensions shrink. • Align metal work functions with semiconductor energy levels to mitigate Schottky barriers. • Manage metal-induced gap states (MIGS) that affect Fermi-level pinning and contact performance. • Integrate new contact materials with current manufacturing processes.		• Employ MIS contacts with ultra-thin dielectrics to reduce SBH. • Research and deploy semimetals or other innovative materials that can potentially inhibit MIGS. • Complete a systematic study of MIGS to understand and engineer their influence on SBH and contact resistance. • Develop compatible fabrication techniques for novel contact materials within existing semiconductor manufacturing workflows.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline (years)**
Identify novel materials with potential to decrease contact resistivity and compatibility with PMOS and NMOS applications	Contact resistivity and work function compatibility	Contact resistivity < 10 μΩ·cm Compatible work function for PMOS and NMOS	3–5
Design and test MIS contacts to reduce SBH, as well as engineering solutions for MIGS	SBH reduction and MIGS control	SBH < 0.3 eV Effective mitigation of MIGS effects	3–5
Develop prototypes with novel contact materials and MIS structures	Electrical performance and interface quality of prototypes	Prototype contacts meeting or exceeding current industry standards (10–100 μΩ·cm)	5–7
Integration with advanced FET technologies	Compatibility and performance in FinFETs and GAA-FETs	Demonstration of integration without performance loss	2–4
Commercialization and manufacturability testing	Process integration success and yield rate	Demonstration of scalability and reliability in manufacturing	3–5
Performance and reliability analysis	Long-term stability and failure rates	Failure rates below industry-standard thresholds; extended device lifetimes	2–3
Stakeholders and Potential Roles in Project			
Stakeholder	**Role**		
Product Manufacturers/Suppliers	• Develop and supply innovative materials that meet the specific resistivity and work function requirements for advanced contacts. • Engage in R&D for scalable production methods of new contact materials and integration technologies.		
End Users/OEMs	• Provide performance specifications and reliability requirements for contacts in various applications. • Test and validate the new contact technologies in real-world scenarios to provide feedback for further development.		

Academia	• Conduct fundamental research on material properties, contact interface physics, and novel contact architectures. • Collaborate with industry partners for knowledge transfer and to guide research towards commercially viable solutions.

Required Resources	Cross Collaboration Needs of Working Groups
• Advanced material synthesis facilities. • High-resolution characterization tools for material analysis. • Computational resources for modeling and simulation. • Fabrication facilities for prototype development.	• Algorithms and Software: Develop predictive models for contact performance and to simulate the effects of new contact materials on overall device efficiency. • APHI: Integrate new contact materials into multi-chip modules. • Circuits and Architectures: Evaluate how new contact materials affect the performance of circuits and overall system architecture, including their impact on signal integrity and speed.

2.1.9 Conclusion for Materials and Devices

The advancements discussed in the Materials and Devices chapter are fundamental to the EES2 roadmap's mission to significantly reduce energy consumption across various sectors, from consumer electronics to large-scale data centers. The integration of novel materials such as 2D materials, CNTs, and ferroelectric materials, alongside advancements in transistor technologies from traditional device structure to Si-GAA, is vital for energy efficiency improvement.

Addressing the challenges of thermal stability, conductivity, and contact resistance is essential and requires robust collaboration between materials science and device engineering, fostering a co-design approach across working groups. New materials must be seamlessly integrated into systems to create more energy-efficient next-generation devices. To achieve this, Metrology and Benchmarking group is crucial to develop standardized testing protocols and precise measurement techniques, ensuring material innovations are rigorously evaluated and reliably transitioned from benchtop discoveries to industry-standard solutions.

To achieve our EES2 energy efficiency goals, strategic investment in the Materials and Devices is crucial for mid-term success. By advancing technologies such as tunnel field-effect transistors (TFETs) and leveraging innovative materials like 2D materials and carbon nanotubes (CNTs), we can significantly reduce energy consumption at the bit level. These advancements in MnD are essential for driving the next wave of energy efficiency improvements, laying the hardware foundation for a more sustainable future in microelectronics.

With the urgent need to deploy advanced energy-efficient devices due to escalating environmental concerns, EES2 has set TRL 6 as the baseline for the deployment of these advanced technologies. This target highlights the necessity for accelerated research and development, the establishment of industry-wide benchmarks, and the development of dedicated testbeds to validate and expedite the market adoption of emerging technologies. Engaging all stakeholders—from policymakers to industry leaders—is crucial to ensure that the pace of innovation matches the pressing timelines for achieving energy sustainability and environmental preservation.

2.1.10 Materials and Devices References

Agha, Firas Natheer Abdul-Kadir, Yasir Hashim, and Waheb Abduljabbar Shaif Abdullah. 2021. "Temperature Characteristics of Gate All Around Nanowire Channel Si-TFET." Presented at the 5th International Conference on Electronic Design (ICED). Perlis, Malaysia. https://doi.org/10.1088/1742-6596/1755/1/012045.

Ahmed, Z., A. Afzalian, T. Schram, D. Jang, D. Verreck, Q. Smets, P. Schuddinck, et al. 2020. "Introducing 2D-FETs in Device Scaling Roadmap using DTCO." Presented at the 2020 IEEE International Electron Devices Meeting (IEDM). San Francisco. https://doi.org/10.1109/IEDM13553.2020.9371906.

Alcorn, Paul. 2023. "Imec Reveals Sub-1nm Transistor Roadmap, 3D-Stacked CMOS 2.0 Plans." Tom's Hardware. Published May 26, 2023. https://www.tomshardware.com/news/imec-reveals-sub-1nm-transistor-roadmap-3d-stacked-cmos-20-plans.

Allain, A., J. Kang, K. Banerjee, and A. Kis. 2015. "Electrical contacts to two-dimensional semiconductors." *Nature Materials*. Vol. 14 (Issue 12): pg 1195–1205. https://doi.org/10.1038/nmat4452.

Aly, M.M.S., M. Gao, G. Hills, C.-S. Lee, G. Pitner, M.M. Shulaker, T.F. Wu, et al. 2015. "Energy-Efficient Abundant-Data Computing: The N3XT 1,000x." *Computer*. Vol. 48 (Issue 12): pg 24–33. https://doi.org/10.1109/MC.2015.376.

Asadi, Kamal. 2010. "Organic non-volatile ferroelectric memories and opto-electronics." https://www.researchgate.net/publication/42788130_Organic_non-volatile_ferroelectric_memories_and_opto-electronics.

Atulasimha, J., and S. Bandyopadhyay. 2010. "Bennett Clocking of Nanomagnetic Logic Using Multiferroic Single-Domain Nanomagnets." *Applied Physics Letters*. Vol. 97 (Issue 17): 173105. https://doi.org/10.1063/1.3506690.

Avci, U.E., D.H. Morris, and I.A. Young. 2015. "Tunnel Field-Effect Transistors: Prospects and Challenges." *IEEE Journal of the Electron Devices Society*. Vol. 3 (Issue 3): pg 88–95. https://doi.org/10.1109/JEDS.2015.2390591.

Ávila, Antonio Ferreira, and Guilherme Silveira Rachid Lacerda. 2008. "Molecular Mechanics Applied to Single-Walled Carbon Nanotubes." *Materials Research*. Vol. 11 (Issue 3): pg 325–333. http://dx.doi.org/10.1590/S1516-14392008000300016.

Barraud, S., V. Lapras, B. Previtali, M.P. Samson, J. Lacord, S. Martinie, M.-A. Jaud, et al. 2017. "Performance and Design Considerations for Gate-All-Around Stacked-NanoWires FETs." Presented at the 2017 IEEE International Electron Devices Meeting (IEDM). San Francisco. https://doi.org/10.1109/IEDM.2017.8268473.

Bhatti, Sabpreet, Rachid Sbiaa, Atsufumi Hirohata, Hideo Ohno, Shunsuke Fukami, and S.N. Piramanayagam. 2017. "Spintronics based random access memory: a review." *Materials Today*. Vol. 20 (Issue 9): pg 530–548. https://doi.org/10.1016/j.mattod.2017.07.007.

Bi, Chong, Congli Sun, Meng Xu, Ty Newhouse-Illige, Paul M. Voyles, and Weigang Wang. 2017. "Electrical Control of Metallic Heavy-Metal—Ferromagnet Interfacial States." *Phys. Rev. Appl*. Vol. 8 (Issue 3): 034003. https://doi.org/10.1103/PhysRevApplied.8.034003.

Biswas, Arpan, Anna N. Morozovska, Maxim Ziatdinov, Eugene A. Eliseev, and Sergei V. Kalinin. 2021. “Multi-objective Bayesian optimization of ferroelectric materials with artificial control for memory and energy storage applications.” *Journal of Applied Physics*. Vol. 130 (Issue 20): 204102. https://doi.org/10.1063/5.0068903.

Blinov, L.M., Vladimir M. Fridkin, Sergei P. Palto, A.V. Bune, P.A. Dowben, and Stephen Ducharme. 2000. “Two-dimensional ferroelectrics.” *Physics-Uspekhi*. Vol. 43 (Issue 3): pg 243. http://dx.doi.org/10.1070/PU2000v043n03ABEH000639.

Borders, William A., Hisanao Akima, Shunsuke Fukami, and Satoshi Moriya. 2017. “Analogue Spin–Orbit Torque Device for Artificial-Neural-Network-Based Associative Memory Operation.” *Applied Physics Express*. Vol. 10 (Issue 1): 013007. http://dx.doi.org/10.7567/APEX.10.013007.

Brady, Gerald J., Austin J. Way, Nathaniel S. Safron, Harold T. Evensen, Padma Gopalan, and Michael S. Arnold. 2016. “Quasi-ballistic carbon nanotube array transistors with current density exceeding Si and GaAs.” *Science Advances*. Vol. 2 (Issue 9). https://doi.org/10.1126/sciadv.1601240.

Briggs, Natalie, Shruti Subramanian, Zhong Lin, Xufan Li, Xiaotian Zhang, Kehao Zhang, Kai Xiao, et al. 2019. “A roadmap for electronic grade 2D materials.” *2D Materials*. Vol 6 (Issue 2): 022001. http://dx.doi.org/10.1088/2053-1583/aaf836.

Burr, G.W., K. Virwani, R.S. Shenoy, G. Fraczak, C.T. Rettner, A. Padilla, R.S. King, et al. 2013. “Recovery Dynamics and Fast (sub-50ns) Read Operation with Access Devices for 3D Crosspoint Memory Based on Mixed-Ionic-Electronic-Conduction (MIEC).” Presented at the 2013 Symposium on VLSI Technology. Kyoto, Japan. https://ieeexplore.ieee.org/document/6576688.

Cao, Guiming, Peng Meng, Jiangang Chen, Haishi Liu, Renji Bian, Chao Zhu, Fucai Liu, and Zheng Liu. 2020. “2D Material Based Synaptic Devices for Neuromorphic Computing.” *Advanced Functional Materials*. Vol. 31 (Issue 4): 2005443. https://doi.org/10.1002/adfm.202005443.

Chatterjee, N., M. O’Connor, D. Lee, D.R. Johnson, S.W. Keckler, M. Rhu, and W.J. Dally. 2017. “Architecting an Energy-Efficient DRAM System for GPUs.” Presented at the 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA). Austin, TX. https://doi.org/10.1109/HPCA.2017.58.

Chaves, A., J.G. Azadani, H. Alsalman, D.R. da Costa, R. Frisenda, A.J. Chaves, S.H. Song, et al. 2020. “Bandgap engineering of two-dimensional semiconductor materials.” *npj 2D Materials and Applications*. Vol. 4 (Article no. 29). http://dx.doi.org/10.1038/s41699-020-00162-4.

Choi, W. Y., and W. Lee. 2010. “Hetero-Gate-Dielectric Tunneling Field-Effect Transistors.” *IEEE Transactions on Electron Devices*. Vol. 57 (Issue 9): pg 2317–2319. https://doi.org/10.1109/TED.2010.2052167.

Christensen, Dennis V., Regina Dittmann, Bernabe Linares-Barranco, Abu Sebastian, Manuel Le Gallo, Andrea Redaelli, Stefan Slesazeck, et al. 2022. “2022 roadmap on neuromorphic computing and engineering.” *Neuromorphic Computing and Engineering*. Vol. 2 (Issue 2): 022501. http://dx.doi.org/10.1088/2634-4386/ac4a83.

Cristoloveanu, S., J. Wan and A. Zaslavsky. 2016. "A Review of Sharp-Switching Devices for Ultra-Low Power Applications," in IEEE Journal of the Electron Devices Society, vol. 4, no. 5, pp. 215-226, Sept. 2016, http://doi.org/10.1109/JEDS.2016.2545978

Cristoloveanu, S., Gérard Ghibaudo. 2022. "Breaking the subthreshold slope limit in MOSFETs." *Science Direct.* Vol. 198: 108465. https://doi.org/10.1016/j.sse.2022.108465

Datta, S., V.Q. Diep, and B. Behin-Aein. 2015. "What constitutes a nanoswitch? A perspective." In *Emerging Nanoelectronic Devices*, Chapter 2, edited by A. Chen, J. Hutchby, V. Zhirnov, and G. Bourianoff. New York: Wiley. https://doi.org/10.48550/arXiv.1404.2254.

Ding, Li, Shibo Liang, Tian Pei, Zhiyong Zhang, Sheng Wang, Weiwei Zhou, Jie Liu, and Lian-Mao Peng. 2012. "Carbon nanotube based ultra-low voltage integrated circuits: Scaling down to 0.4 V." *Applied Physics Letters*. Vol. 100 (Issue 26): 233116. https://doi.org/10.1063/1.4731776.

Dowben, P. A., C. Binek, K. Zhang, L. Wang, W.-N. Mei, J.P. Bird, U. Singisetti, et al. 2018. "Towards a Strong Spin–Orbit Coupling Magnetoelectric Transistor." *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*. Vol. 4 (Issue 1): pg 1–9. https://doi.org/10.1109/JXCDC.2018.2809640.

Du, Frank, Jonathan R. Felts, Xu Xie, Jizhou Song, Yuhang Li, Matthew R. Rosenberger, Ahmad E. Islam, et al. 2014. "Laser-induced nanoscale thermocapillary flow for purification of aligned arrays of single-walled carbon nanotubes." *ACS Nano*. Vol. 8 (Issue 12): pg 12641–12649. https://doi.org/10.1021/nn505566r.

Dunkel, Stefan, Martin Trentzsch, Ralf Richter, Peter Moll, Christine Fuchs, Oliver Gehring, M. Majer, et al. 2017. "A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond." Presented at the 2017 IEEE International Electron Devices Meeting (IEDM). San Francisco. https://doi.org/10.1109/IEDM.2017.8268425.

Dutta, S., H. Ye, W. Chakraborty, Y.-C. Luo, M. San Jose, B. Grisafe, A. Khanna, et al. 2020. "Monolithic 3D Integration of High Endurance Multi-Bit Ferroelectric FET for Accelerating Compute-In-Memory." Presented at the 2020 IEEE International Electron Devices Meeting (IEDM). San Francisco. https://doi.org/10.1109/IEDM13553.2020.9371974.

Dutta, S., H. Ye, A.A. Khandker, S.G. Kirtania, A. Khanna, K. Ni, and S. Datta. 2022. "Logic Compatible High-Performance Ferroelectric Transistor Memory." *IEEE Electron Device Letters*. Vol. 43 (Issue 3): pg 382–385. https://doi.org/10.1109/LED.2022.3148669.

Elías, Ana Laura, Néstor Perea-López, Andrés Castro-Beltrán, Ayse Berkdemir, Ruitao Lv, Simin Feng, Aaron D. Long, et al. 2013. "Controlled Synthesis and Transfer of Large-Area WS_2 Sheets: From Single Layer to Few Layers." *ACS Nano*. Vol. 7 (Issue 6): pg 5235–5242. https://doi.org/10.1021/nn400971k.

Eyvazi, K., and M. A. Karami. 2020. "A New Junction-Less Tunnel Field-Effect Transistor with a SiO_2/HfO_2 Stacked Gate Oxide for DC Performance Improvement." Presented at the 28th Iranian Conference on Electrical Engineering (ICEE). Tabriz, Iran. https://doi.org/10.1109/ICEE50131.2020.9260621.

Franklin, Aaron D., George S. Tulevski, Shu-Jen Han, Davood Shahrjerdi, Qing Cao, Hong-Yu Chen, H.-S. Philip Wong, and Wilfried Haensch. 2012a. "Variability in Carbon Nanotube

Transistors: Improving Device-to-Device Consistency." *ACS Nano.* Vol. 6 (Issue 2): pg 1109–1115. https://doi.org/10.1021/nn203516z.

Franklin, Aaron D., Mathieu Luisier, Shu-Jen Han, George Tulevski, Chris M. Breslin, Lynne Gignac, Mark S. Lundstrom, and Wilfried Haensch. 2012b. "Sub-10 nm Carbon Nanotube Transistor." *Nano Letters*. Vol. 12 (Issue 2): pg 758–762. https://doi.org/10.1021/nl203701g.

Franklin, Aaron D., Damon B. Farmer, and Wilfried Haensch. 2014. "Defining and Overcoming the Contact Resistance Challenge in Scaled Carbon Nanotube Transistors." *ACS Nano*. Vol. 8 (Issue 7): pg 7333–7339. https://doi.org/10.1021/nn5024363.

Garcia, V., and Manuel Bibes. 2014. "Ferroelectric tunnel junctions for information storage and processing." *Nature Communications*. Vol. 5 (Article no. 4289). https://doi.org/10.1038/ncomms5289.

Go, Gyeong-Tak, Yeongjun Lee, Dae-Gyo Seo, Mingyuan Pei, Wanhee Lee, Hoichang Yang, and Tae-Woo Lee. 2020. "Achieving Microstructure-Controlled Synaptic Plasticity and Long-Term Retention in Ion-Gel-Gated Organic Synaptic Transistors." *Advanced Intelligent Systems*. Vol. 2 (Issue 11): 2000012. https://doi.org/10.1002/aisy.202000012.

Google Open Source Blog. 2022. "Google and NIST Partner on Nanotechnology Development Platform." Published September 13, 2022. https://opensource.googleblog.com/2022/09/google-and-nist-partner-on-nanotechnology-development-platform.html.

Grimaldi, E., V. Krizakova, G. Sala, et al. 2020. "Single-shot dynamics of spin–orbit torque and spin transfer torque switching in three-terminal magnetic tunnel junctions." *Nature Nanotechnology*. Vol. 15: pg 111–117. https://www.nature.com/articles/s41565-019-0607-7.

Grollier, J., D. Querlioz, K.Y. Camsari, et al. 2020. "Neuromorphic Spintronics." *Nature Electronics*. Vol. 3: pg 360–370. http://dx.doi.org/10.1038/s41928-019-0360-9.

Guan, Zhao, He Hu, Xinwei Shen, Pinghua Xiang, Ni Zhong, Junhao Chu, and Chungang Duan. 2020. "Recent Progress in Two-Dimensional Ferroelectric Materials." *Advanced Electronic Materials*. Vol. 6 (Issue 1): 1900818. https://doi.org/10.1002/aelm.201900818.

Han, S.-J., H. Oida, H. Park, J.B. Hannon, G.S. Tulevski, and W. Haensch. 2013. "Carbon nanotube complementary logic based on Erbium contacts and self-assembled high purity solution tubes." Presented at the 2013 IEEE International Electron Devices Meeting. Washington, DC. http://dx.doi.org/10.1109/IEDM.2013.6724664.

Hatton, Benjamin D., Kai Landskron, William J. Hunks, Mark R. Bennett, Donna Shukaris, Douglas D. Perovic, and Geoffrey A. Ozin. 2006. "Materials Chemistry for Low-k Materials." *Materials Today*. Vol. 9 (Issue 3): pg 22–31. https://doi.org/10.1016/s1369-7021(06)71387-6.

He, Keke, Bilal Barut, Shenchu Yin, Michael D. Randle, Ripudaman Dixit, Nargess Arabchigavkani, Jubin Nathawat, et al. 2022. "Graphene on Chromia: A System for Beyond-Room-Temperature Spintronics." *Advanced Materials*. Vol. 34 (Issue 12): 2105023. http://dx.doi.org/10.1002/adma.202105023.

Hiramoto, T. 2009. "Transistor Evolution for CMOS Extension and Future Information Processing Technologies." Presented at the 2009 International Workshop on Junction Technology. Kyoto, Japan. https://doi.org/10.1109/IWJT.2009.5166205.

Hofman, Sander. 2022. “What is a gate-all-around transistor?” ASML. Published October 3, 2022. https://www.asml.com/en/news/stories/2022/what-is-a-gate-all-around-transistor.

Hoskins, Brian, Wen Ma, Mitchell Fream, Osama Yousuf, Mathew Daniels, Jonathan Goodwill, Advait Madhavan, et al. 2021. “A System for Validating Resistive Neural Network Prototypes.” Presented at ICONS 2021: International Conference on Neuromorphic Systems 2021 (Article no. 26). Published October 13, 2021. https://doi.org/10.1145/3477145.3477260.

Hu, G., J.H. Lee, J.J. Nowak, J.Z. Sun, J. Harms, A. Annunziata, S. Brown, et al. 2015. “STT-MRAM with Double Magnetic Tunnel Junctions.” Presented at the 2015 IEEE International Electron Devices Meeting (IEDM). Washington, DC. https://doi.org/10.1109/IEDM.2015.7409772.

Huang, Y.-C., M.-H. Chiang, S.-J. Wang, and J.G. Fossum. 2017. “GAAFET Versus Pragmatic FinFET at the 5nm Si-Based CMOS Technology Node.” *IEEE Journal of the Electron Devices Society*. Vol. 5 (Issue 3): pg 164–169. https://doi.org/10.1109/JEDS.2017.2689738.

Hwang, Cheol Seong, and Thomas Mikolajick. 2019. “Ferroelectric Memories.” In *Advances in Non-Volatile Memory and Storage Technology, Second Edition*, edited by Blanka Magyari-Köpe and Yoshio Nishi, 393–441. Woodhead Publishing. ISBN: 9780081025840. https://doi.org/10.1016/B978-0-08-102584-0.00012-7.

Ikeda, S., et al. 2007. “Magnetic Tunnel Junctions for Spintronic Memories and Beyond.” *IEEE Transactions on Electron Devices*. Vol. 54 (Issue 5): pg 991–1002. https://doi.org/10.1109/TED.2007.894617.

IRDS. 2020. *International Roadmap for Devices and Systems (IRDS™) 2020 Edition.* Institute of Electrical and Electronic Engineers (IEEE). https://irds.ieee.org/editions/2020.

IRDS. 2021. *International Roadmap for Devices and Systems: 2021 Update Beyond CMOS*. IEEE. https://irds.ieee.org/images/files/pdf/2021/2021IRDS_BC.pdf.

IRDS. 2022. *International Roadmap for Devices and Systems: 2022 Edition*. IEEE. https://irds.ieee.org/images/files/pdf/2022/2022IRDS_BC.pdf.

Kaarsberg, Tina, Shashank Misra, and Kenta Shimizu. 2023. “Manufacturing an Extremely Efficient Transistor for Decarbonization.” U.S. Department of Energy (DOE), Advanced Manufacturing Office (AMO). Sandia National Laboratories/Energetics Incorporated. Accessed December 2023. https://aceee.org/sites/default/files/pdfs/sis21/panel-2/Kaarsberg.pdf.

Kanai, S., M. Yamanouchi, S. Ikeda, Y. Nakatani, F. Matsukura, and H. Ohno. 2012. “Electric Field-Induced Magnetization Reversal in a Perpendicular-Anisotropy CoFeB-MgO Magnetic Tunnel Junction.” *Applied Physics Letters*. Vol. 101 (Issue 12): 122403. https://doi.org/10.1063/1.4753816.

Kang, Minji, Sang-A Lee, Sukjae Jang, Sunbin Hwang, Seoung-Ki Lee, Sukang Bae, Jae-Min Hong, et al. 2019. “Low-Voltage Organic Transistor Memory Fiber with a Nanograined Organic Ferroelectric Film.” *ACS Appl. Mater. Interfaces*. Vol. 11 (Issue 25): pg 22575–22582. https://doi.org/10.1021/acsami.9b03564.

Kanungo, S., G. Ahmad, P. Sahatiya, et al. 2022. “2D Materials-based Nanoscale Tunneling Field Effect Transistors: Current Developments and Future Prospects.” *npj 2D Materials and Applications*. Vol. 6 (Article no. 83). https://doi.org/10.1038/s41699-022-00352-2.

Kao, K.-H., et al. 2012. "Optimization of Gate-on-Source-Only Tunnel FETs With Counter-Doped Pockets." *IEEE Transactions on Electron Devices*. Vol. 59 (Issue 8): pg 2070–2077. https://doi.org/10.1109/TED.2012.2200489.

Karigerasi, Manohar, et al. 2022. "Simulation-Guided Thermal Process Discovery for Flash Lamp Annealing Crystallization of On-Chip HfO_2-ZrO_2 Ferroelectric Memories." Presented at the 2022 MRS (Materials Research Society) Spring Meeting. Honolulu, HI. https://www.mrs.org/meetings-events/presentation/2022_mrs_spring_meeting/2022_mrs_spring_meeting-3664701.

Keene, Scott Tom, Armantas Melianas, Yoeri van de Burgt, and Alberto Salleo. 2019. "Mechanisms for Enhanced State Retention and Stability in Redox-Gated Organic Neuromorphic Devices." *Advanced Electronic Materials*. Vol. 5 (Issue 2): 1800686. https://doi.org/10.1002/aelm.201800686.

Khan, A.I., A. Keshavarzi, and S. Datta. 2020. "The future of ferroelectric field-effect transistor technology." *Nat Electron*. Vol. 3: pg 588–597. https://doi.org/10.1038/s41928-020-00492-7.

Khanai, Pravin, Bowei Zhou, Magda Andrade, Yanliu Dang, Albert Davydov, Ali Habiboglu, Jonah Saidian, et al. 2021. "Perpendicular Magnetic Tunnel Junctions with Multi-interface Free Layer." *Applied Physics Letters*. Vol. 119 (Issue 24): 242404. https://doi.org/10.1063/5.0066782.

Kosub, Tobias, Martin Kopte, Florin Radu, Oliver G. Schmidt, and Denys Makarov. 2015. "All-Electric Access to the Magnetic-Field-Invariant Magnetization of Antiferromagnets." *Phys. Rev. Lett.* Vol. 115 (Issue 9): 097201. https://doi.org/10.1103/PHYSREVLETT.115.097201.

Kosub, T., M. Kopte, R. Hühne, et al. 2017. "Purely antiferromagnetic magnetoelectric random access memory." *Nat Commun*. Vol. 8: 13985. http://dx.doi.org/10.1038/ncomms13985.

Kumar, S., Y. Singh, B. Singh, and P.K. Tiwari. 2020. "Simulation Study of Dielectric Modulated Dual Channel Trench Gate TFET-Based Biosensor." *IEEE Sensors Journal*. Vol. 20 (Issue 21): pg 12565–12573. https://doi.org/10.1109/JSEN.2020.3001300.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature*. Vol. 521: pg 436–444. https://doi.org/10.1038/nature14539.

Lee, C.-S., E. Pop, A.D. Franklin, W. Haensch, and H.-S.P. Wong. 2015. "A Compact Virtual-Source Model for Carbon Nanotube FETs in the Sub-10-nm Regime—Part I: Intrinsic Elements." *IEEE Transactions on Electron Devices*. Vol. 62 (Issue 9): pg 3061–3069. https://doi.org/10.1109/TED.2015.2457453.

Lee, K., et al. 2018. "22-nm FD-SOI Embedded MRAM Technology for Low-Power Automotive-Grade MCU Applications." Presented at the 2018 IEEE International Electron Devices Meeting (IEDM). San Francisco. https://doi.org/10.1109/IEDM.2018.8614566.

Lee, Hyun-Jae, Minseong Lee, Kyoungjun Lee, Jinhyeong Jo, Hyemi Yang, Yunseok Kim, Seungchul Cha, Umesh Waghmare, and Jun Hee Lee. 2020. "Scale-free ferroelectricity induced by flat phonon bands in HfO_2." *Science*. Vol. 369 (Issue 6509): pg 1343–1347. https://doi.org/10.1126/science.aba0067.

Lenz, J., F. del Giudice, F.R. Geisenhof, et al. 2019. "Vertical, electrolyte-gated organic transistors show continuous operation in the MA cm^−2 regime and artificial synaptic behavior." *Nat. Nanotechnol*. Vol. 14: pg 579–585. https://www.nature.com/articles/s41565-019-0407-0.

Lin, Y-M., C. Dimitrakopoulos, K.A. Jenkins, D.B. Farmer, H.-Y. Chiu, A. Grill, and Ph. Avouris. 2010. "100-GHz transistors from wafer-scale epitaxial graphene." *Science*. Vol. 327 (Issue 5966): pg 662.

Liu, Bilu, Mohammad Fathi, Liang Chen, Ahmad Abbas, Yuqiang Ma, and Chongwu Zhou. 2015. "Chemical Vapor Deposition Growth of Monolayer WSe_2 with Tunable Device Characteristics and Growth Mechanism Study." *ACS Nano*. Vol. 9 (Issue 6): pg 6119–6127. https://doi.org/10.1021/acsnano.5b01301.

Liu, Luqiao, Chi-Feng Pai, Y. Li, H. W. Tseng, D. C. Ralph, and R. A. Buhrman. 2012. "Spin-Torque Switching with the Giant Spin Hall Effect of Tantalum." *Science.* Vol. 336 (Issue 6081): pg 555–558. https://doi.org/10.1126/science.1218197.

Liu, Z.-C., and L. Wang. 2020. "Applications of Phase Change Materials in Electrical Regime From Conventional Storage Memory to Novel Neuromorphic Computing." *IEEE Access*. Vol. 8: pg 76471–76499. https://doi.org/10.1109/ACCESS.2020.2990536.

Liu, Xiwen, D. Wang, K.-H. Kim, K. Katti, J. Zheng, P. Musavigharavi, J. Miao, E.A. Stach, R.H. Olsson III, and D. Jariwala. 2021. "Post-CMOS Compatible Aluminum Scandium Nitride/2D Channel Ferroelectric Field-Effect-Transistor Memory." *Nano Letters*. Vol. 21 (Issue 9): pg 3753–3761. https://doi.org/10.1021/acs.nanolett.0c05051.

Lo, C.-L., K. Zhang, R.S. Smith, K. Shah, J.A. Robinson, and Z. Chen. 2018. "Large-Area, Single-Layer Molybdenum Disulfide Synthesized at BEOL Compatible Temperature as Cu Diffusion Barrier." *IEEE Electron Device Letters.* Vol. 39 (Issue 6): pg 873–876. https://doi.org/10.1109/LED.2018.2827061.

Lu, H., and A. Seabaugh. 2014. "Tunnel Field-Effect Transistors: State-of-the-Art." *IEEE Journal of the Electron Devices Society*. Vol. 2 (Issue 4): pg 44–49. https://doi.org/10.1109/JEDS.2014.2326622.

Ma, D., J. Shi, Q. Ji, et al. 2015. "A universal etching-free transfer of MoS_2 films for applications in photodetectors." *Nano Research*. Vol. 8: pg 3662–3672. http://dx.doi.org/10.1007/s12274-015-0866-z.

Maas, Klaasjan, Edouard Villepreux, David Cooper, Carmen Jiménez, Hervé Roussel, Laetitia Rapenne, Xavier Mescot, Quentin Rafhay, Michel Boudard, and Mónica Burriel. 2020. "Using a Mixed Ionic Electronic Conductor to Build an Analog Memristive Device with Neuromorphic Programming Capabilities." *Journal of Materials Chemistry C.* Issue 2. http://dx.doi.org/10.1039/C9TC03972D.

Mahmood, Ather, Will Echtenkamp, Mike Street, Jun-Lei Wang, Shi Cao, et al. 2021. "Voltage Controlled Néel Vector Rotation in Zero Magnetic Field at CMOS-Compatible Temperatures" (Version 1). Research Square. Published March 15, 2021. https://doi.org/10.21203/rs.3.rs-38435/v1.

Manipatruni, Sasikanth, Dmitri E. Nikonov, Ramamoorthy Ramesh, Huichu Li, and Ian A. Young. 2017. "Spin-Orbit Logic with Magnetoelectric Nodes: A Scalable Charge Mediated Nonvolatile Spintronic Logic." arXiv. Last modified March 5, 2017. https://doi.org/10.48550/arXiv.1512.05428.

Manipatruni, S., D.E. Nikonov, and I.A. Young. 2018. “Beyond CMOS computing with spin and polarization.” *Nature Physics.* Vol. 14: pg 338–343. http://dx.doi.org/10.1038/s41567-018-0101-4.

Manipatruni, S., D.E. Nikonov, C.C. Lin, et al. 2019. “Scalable energy-efficient magnetoelectric spin–orbit logic.” *Nature*. Vol. 565: pg 35–42. https://www.nature.com/articles/s41586-018-0770-2.

Micron Technology, Inc. 2017. “TN-40-07: Calculating Memory Power for DDR4 SDRAM.” Technical Note. Accessed January 26, 2024. https://www.micron.com/-/media/client/global/documents/products/technical-note/dram/tn4007_ddr4_power_calculation.pdf.

Mikolajick, T., S. Slesazeck, H. Mulaosmanovic, M.H. Park, S. Fichtner, P.D. Lomenzo, and M. Hoffmann. 2021. “Next Generation Ferroelectric Materials for Semiconductor Process Integration and Their Applications.” *Journal of Applied Physics*. Vol. 129 (Issue 10): 100901. https://doi.org/10.1063/5.0037617.

Mitra, Suman Kr., and Brinda Bhowmick. 2019. “Impact of Interface Traps on Performance of Gate-on-Source/Channel SOI TFET.” *Microelectronics Reliability.* Vol. 94: pg 1–12. https://doi.org/10.1016/j.microrel.2019.01.004.

Molckovsky, A., et al. 2019. “Bridging the Gap between Performance and Energy-Efficiency in Emerging Applications.” *ACM Transactions on Design Automation of Electronic Systems*. Vol. 24 (Issue 5, Article 54). doi:10.1145/3350663.

Moriyama, N., Y. Ohno, T. Kitamura, S. Kishimoto, and T. Mizutani. 2010. “Change in Carrier Type in High-k Gate Carbon Nanotube Field-Effect Transistors by Interface Fixed Charges.” *Nanotechnology*. Vol. 21 (Issue 16): 165201. https://doi.org/10.1088/0957-4484/21/16/165201.

MRAM-info. 2023. “STT-MRAM.” Accessed December 2023. https://mram-info.com/stt-mram.

Mueller, Johannes, Stefan Slesazeck, and Thomas Mikolajick. 2019. “Ferroelectric Field Effect Transistor.” In *Ferroelectricity in Doped Hafnium Oxide: Materials, Properties and Devices*, edited by Uwe Schroeder, Cheol Seong Hwang, and Hiroshi Funakubo, pg 451–471. Woodhead Publishing Series in Electronic and Optical Materials. Woodhead Publishing. ISBN: 9780081024300. https://doi.org/10.1016/B978-0-08-102430-0.00022-X.

Mukesh, Sagarika, and Jingyun Zhang. 2022. “A Review of the Gate-All-Around Nanosheet FET Process Opportunities.” *Electronics*. Vol. 11 (Issue 21): 3589. https://doi.org/10.3390/electronics11213589.

Narayanan, P., et al. 2015. “Exploring the Design Space for Crossbar Arrays Built With Mixed-Ionic-Electronic-Conduction (MIEC) Access Devices.” *IEEE Journal of the Electron Devices Society*. Vol. 3 (Issue 5): pg 423–434. https://doi.org/10.1109/JEDS.2015.2442242.

Nazir, Ghazanfar, Adeela Rehman, and Soo-Jin Park. 2020. “Energy-Efficient Tunneling Field-Effect Transistors for Low-Power Device Applications: Challenges and Opportunities.” *ACS Applied Materials & Interfaces.* Vol. 12 (Issue 42): pg 47127–47163. https://doi.org/10.1021/acsami.0c10213.

Nigam, Kaushal, Pravin Kondekar, and Dheeraj Sharma. 2016. “Approach for Ambipolar Behaviour Suppression in Tunnel FET by Workfunction Engineering.” *Micro & Nano Letters*. Vol. 11 (Issue 8): pg 460–464. https://doi.org/10.1049/mnl.2016.0178.

Nogami, T., et al. 2021. “Electromigration and Line Resistance of Graphene Capped Cu Dual Damascene Interconnect.” Presented at the 2021 IEEE International Electron Devices Meeting (IEDM). San Francisco. https://doi.org/10.1109/IEDM19574.2021.9720525.

Nowak, J.J., et al. 2016. “Dependence of Voltage and Size on Write Error Rates in Spin-Transfer Torque Magnetic Random-Access Memory.” *IEEE Magnetics Letters.* Vol. 7 (Article no. 3102604): pg 1–4. https://doi.org/10.1109/LMAG.2016.2539256.

Pananakakis, G., Gérard Ghibaudo, and Sorin Cristoloveanu. 2023. “Detailed comparison of threshold voltage extraction methods in FD-SOI MOSFETs.” *ScienceDirect.* Vol. 209: 108764. https://doi.org/10.1016/j.sse.2023.108764.

Pan, C., and A. Naeemi. 2017. “Nonvolatile Spintronic Memory Array Performance Benchmarking Based on Three-Terminal Memory Cell.” *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*. Vol. 3: pg 10–17. https://doi.org/10.1109/JXCDC.2017.2669213.

Pan, C., and A. Naeemi. 2018. “Complementary Logic Implementation for Antiferromagnet Field-Effect Transistors.” *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*. Vol. 4 (Issue 2): pg 69–75. https://doi.org/10.1109/JXCDC.2018.2878635.

Parkin, S., C. Kaiser, A. Panchula, et al. 2004. “Giant Tunnelling Magnetoresistance at Room Temperature with MgO (100) Tunnel Barriers.” *Nature Materials*. Vol. 3: pg 862–867. http://dx.doi.org/10.1038/nmat1256.

Pitner, Gregory, Gage Hills, Juan Pablo Llinas, Karl-Magnus Persson, Rebecca Park, Jeffrey Bokor, Subhasish Mitra, and H.-S. Philip Wong. 2019. “Low-Temperature Side Contact to Carbon Nanotube Transistors: Resistance Distributions Down to 10 nm.” *Nano Letters*. Vol. 19 (Issue 2): pg 1083–1089. https://doi.org/10.1021/acs.nanolett.8b04370.

Puebla, J., J. Kim, K. Kondou, et al. 2020. “Spintronic Devices for Energy-Efficient Data Storage and Energy Harvesting.” *Communications Materials*. Vol. 1 (Issue 24). http://dx.doi.org/10.1038/s43246-020-0022-5.

Rabe, Karin M., Matthew Dawber, Céline Lichtensteiger, Charles H. Ahn, and Jean-Marc Triscone. 2007. “Modern Physics of Ferroelectrics: Essential Background.” In *Physics of Ferroelectrics: A Modern Perspective*, edited by Karin M. Rabe, Charles H. Ahn, and Jean-Marc Triscone, pg 1–30. New York: Springer. http://dx.doi.org/10.1007/978-3-540-34591-6_1.

Rahi, S.B., P. Asthana, and S. Gupta. 2017. “Heterogate Junctionless Tunnel Field-Effect Transistor: Future of Low-Power Devices.” *Journal of Computational Electronics*. Vol. 16: pg 30–38. https://doi.org/10.1007/s10825-016-0936-9.

Rehman, Muhammad Muqeet, Hafiz Mohammad Mutee Ur Rehman, Jahan Zeb Gul, Woo Young Kim, Khasan S. Karimov, and Nisar Ahmed. 2020. “Decade of 2D-materials-based RRAM devices: a review.” *Science and Technology of Advanced Materials*. Vol. 21 (Issue 1): pg 147–186. https://doi.org/10.1080/14686996.2020.1730236.

Revelant, A., et al. 2014. “Electron-Hole Bilayer TFET: Experiments and Comments.” *IEEE Transactions on Electron Devices*. Vol. 61 (Issue 8): pg 2674–2681. https://doi.org/10.1109/TED.2014.2329551.

Rutherglen, C., A.A. Kane, P.F. Marsh, et al. 2019. “Wafer-scalable, aligned carbon nanotube transistors operating at frequencies of over 100 GHz.” *Nature Electronics*. Vol. 2: pg 530–539. https://www.nature.com/articles/s41928-019-0326-y.

Ryan, E.T., A.J. McKerrow, J. Leu, and P.S. Ho. 2003. “Materials Issues and Characterization of Low-k Dielectric Materials.” In *Low Dielectric Constant Materials for IC Applications*, edited by P.S. Ho, J.J. Leu, and W.W. Lee, Springer Series in Advanced Microelectronics, Vol. 9. Berlin and Heidelberg, Germany: Springer. http://dx.doi.org/10.1007/978-3-642-55908-2_2.

Saini, Balreen, Fei Huang, Yoon-Young Choi, Zhouchangwan Yu, Vivek Thampy, John D. Baniecki, Wilman Tsai, and Paul C. McIntyre. 2023. “Field-Induced Ferroelectric Phase Evolution During Polarization ‘Wake-Up’ in $Hf_{0.5}Zr_{0.5}O_2$ Thin Film Capacitors.” *Advanced Electronic Materials.* Vol. 9 (Issue 6): 2300016. https://doi.org/10.1002/aelm.202300016.

Samsung Semiconductor. 2022. “Samsung Begins Chip Production Using 3nm Process Technology With GAA Architecture.” Published June 30, 2022. https://news.samsung.com/global/samsung-begins-chip-production-using-3nm-process-technology-with-gaa-architecture.

Seabaugh, Alan C., and Qin Zhang. 2010. “Low-Voltage Tunnel Transistors for Beyond CMOS Logic.” *Proceedings of the IEEE.* Vol. 98 (Issue 12): pg 2095–2110. https://doi.org/10.1109/JPROC.2010.2070470.

Sekhar, Vasarla Nagendra. 2012. “Mechanical Characterization of Black Diamond (Low-k) Structures for 3D Integrated Circuit and Packaging Applications.” In *Nanoindentation in Materials* Science, edited by Jiri Nemecek. IntechOpen. https://doi.org/10.5772/53198.

Semiconductor Engineering. 2023. “Gate-All-Around FET (GAA FET).” Accessed December 4, 2023. https://semiengineering.com/knowledge_centers/integrated-circuit/transistors/3d/gate-all-around-fet/.

Sharma, N., J.P. Bird, Ch. Binek, P.A. Dowben, D. Nikonov, and A. Marshall. 2020. “Evolving Magneto-electric Device Technologies.” *Semiconductor Science and Technology*. Vol. 35 (Issue 7): 037001. http://dx.doi.org/10.1088/1361-6641/ab8438.

Shenoy, Rohit S., Geoffrey W. Burr, Kumar Virwani, Bryan Jackson, Alvaro Padilla, Pritish Narayanan, Charles T. Rettner, Robert M. Shelby, Donald S. Bethune, and Karthik V. Raman. 2014. “MIEC (mixed-ionic-electronic-conduction)-based access devices for non-volatile crossbar memory arrays.” *Semiconductor Science and Technology*. Vol. 29 (Issue 10): 104005. http://dx.doi.org/10.1088/0268-1242/29/10/104005.

Shulaker, Max M., Gage Hills, Tony F. Wu, Zhenan Bao, H.-S. Philip Wong, and Subhasish Mitra. 2015. “Efficient metallic carbon nanotube removal for highly-scaled technologies.” Presented at the 2015 IEEE International Electron Devices Meeting (IEDM). Washington, DC. https://doi.org/10.1109/IEDM.2015.7409815.

Si, Mengwei, Pai-Ying Liao, Gang Qiu, Yuqin Duan, and Peide D. Ye. 2018. “Ferroelectric Field-Effect Transistors Based on MoS_2 and $CuInP_2S_6$ Two-Dimensional van der Waals Heterostructure.” *ACS Nano*. Vol. 12 (Issue 7): pg 6700–6705. https://doi.org/10.1021/acsnano.8b01810.

Simmons, J.M., B.M. Nichols, S.E. Baker, Matthew S. Marcus, O.M. Castellini, C.-S. Lee, R.J. Hamers, and M.A. Eriksson. 2006. “Effect of Ozone Oxidation on Single-Walled Carbon Nanotubes.” *The Journal of Physical Chemistry B.* Vol. 110 (Issue 14): pg 7113–7118. https://doi.org/10.1021/jp0548422.

Singh, Kamaldeep. 2021. “Gate-All-Around (GAA) FET – Going Beyond The 3 Nanometer Mark.” Copperpod Intellectual Property. Published September 15, 2021. https://www.copperpodip.com/post/gate-all-around-gaa-fet-going-beyond-the-3-nanometer-mark.

Slesazeck, Stefan, and Thomas Mikolajick. 2019. “Nanoscale Resistive Switching Memory Devices: A Review.” *Nanotechnology*. Vol. 30 (Issue 35, Article no. 352003). http://dx.doi.org/10.1088/1361-6528/ab2084.

Slonczewski, J.C. 1996. “Current-driven excitation of magnetic multilayers.” *Journal of Magnetism and Magnetic Materials*. Vol. 159 (Issues 1–2): pg L1–L7. https://doi.org/10.1016/0304-8853(96)00062-5.

Södergren, L., P. Olausson and E. Lind, "Cryogenic Characteristics of InGaAs MOSFET," in IEEE Transactions on Electron Devices, vol. 70, no. 3, pp. 1226-1230, March 2023, https://doi.org/10.1109/TED.2023.3238382

Stolichnov, Igor, Matteo Cavalieri, Enrico Colla, Tony Schenk, Terence Mittmann, Thomas Mikolajick, Uwe Schroeder, and Adrian M. Ionescu. 2018. “Genuinely Ferroelectric Sub-1-Volt-Switchable Nanodomains in $Hf_xZr_{(1-x)}O_2$ Ultrathin Capacitors.” *ACS Applied Materials & Interfaces*. Vol. 10 (Issue 36): pg 30514–30521. https://doi.org/10.1021/acsami.8b07988.

Sun, Jonathan Z., and Christopher Safranski. 2022. “Metrology and Metrics for Spin-Transfer-Torque Switched Magnetic Tunnel Junctions in Memory Applications.” *Journal of Magnetism and Magnetic Materials*. Vol. 563: 169878. https://doi.org/10.1016/j.jmmm.2022.169878.

Tan, J., J.H. Lim, J.H. Kwon, V.B. Naik, N. Raghavan, and K.L. Pey. 2021. “Role of temperature, MTJ size and pulse-width on STT-MRAM bit-error rate and backhopping.” *Solid-State Electronics*. Vol. 183: 108032. https://doi.org/10.1016/j.sse.2021.108032.

U.S. Department of Energy, Office of Energy Efficiency & Renewable Energy. 2021. “Advanced Manufacturing Office Workshop on Manufacturing and Integration Challenges for Analog and Neuromorphic Computing.” Workshop Report, August 11–13, 2021. https://energy.gov/sites/default/files/2022-08/AMO%20Semiconductors%20Workshop%20Report_2022.pdf.

van de Burgt, Y., E. Lubberman, E. Fuller, et al. 2017. “A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing.” *Nature Mater.* Vol. 16: pg 414–418. http://dx.doi.org/10.1038/NMAT4856.

van der Veen, M.H., et al. 2018. “Damascene Benchmark of Ru, Co and Cu in Scaled Dimensions.” Presented at the 2018 IEEE International Interconnect Technology Conference (IITC). Santa Clara, CA. https://doi.org/10.1109/IITC.2018.8430407.

Vaz, D.C., et al. 2021. “Functional Demonstration of a Fully Integrated Magneto-Electric Spin-Orbit Device.” Presented at the 2021 IEEE International Electron Devices Meeting (IEDM). San Francisco. https://doi.org/10.1109/IEDM19574.2021.9720677.

Vincent, A.F., et al. 2015. "Spin-Transfer Torque Magnetic Memory as a Stochastic Memristive Synapse for Neuromorphic Systems." *IEEE Transactions on Biomedical Circuits and Systems*. Vol. 9 (Issue 2): pg 166–174. https://doi.org/10.1109/TBCAS.2015.2414423.

Wang, Xingli, Yongji Gong, Gang Shi, Wai Leong Chow, Kunttal Keyshar, Gonglan Ye, Robert Vajtai, et al. 2014. "Chemical Vapor Deposition Growth of Crystalline Monolayer $MoSe_2$." *ACS Nano*. Vol. 8 (Issue 5): pg 5125–5131. http://dx.doi.org/10.1021/nn501175k.

Wang, Hanlin, Qiang Zhao, Zhenjie Ni, Qingyuan Li, Hongtao Liu, Yunchang Yang, Lifeng Wang, et al. 2018. "A Ferroelectric/Electrochemical Modulated Organic Synapse for Ultraflexible, Artificial Visual-Perception System." *Advanced Materials*. Vol. 30 (Issue 46): 1803961. https://doi.org/10.1002/adma.201803961.

Wang, M., S. Cai, C. Pan, et al. 2018. "Robust memristors based on layered two-dimensional materials." *Nature Electronics*. Vol. 1: pg 130–136. https://www.nature.com/articles/s41928-018-0021-4.

Wang, Z., et al. 2021. "Standby Bias Improvement of Read After Write Delay in Ferroelectric Field Effect Transistors." Presented at the 2021 IEEE International Electron Devices Meeting (IEDM). San Francisco. https://doi.org/10.1109/IEDM19574.2021.9720502.

Wei, H., H. Shulaker, H.-S. P. Wong, and S. Mitra. 2013. "Monolithic three-dimensional integration of carbon nanotube FET complementary logic circuits." Presented at the 2013 IEEE International Electron Devices Meeting. Washington, DC. https://doi.org/10.1109/IEDM.2013.6724663.

Witt, C., et al. 2018. "Testing The Limits of TaN Barrier Scaling." Presented at the 2018 IEEE International Interconnect Technology Conference (IITC). Santa Clara, CA. https://doi.org/10.1109/IITC.2018.8430289.

Worledge, D.C. 2022. "Spin-Transfer-Torque MRAM: the Next Revolution in Memory." Presented at the 2022 IEEE International Memory Workshop (IMW). Dresden, Germany. https://doi.org/10.1109/IMW52921.2022.9779288.

Wu, Z., et al. 2018. "PVD- Treated ALD TaN for Cu Interconnect Extension to 5nm Node and Beyond." Presented at the 2018 IEEE International Interconnect Technology Conference (IITC). Santa Clara, CA. https://doi.org/10.1109/IITC.2018.8430433.

Xia, J., and Y. Hu. 2022. "Organic ferroelectric non-volatile memory transistors." Presented at the 2022 IEEE International Flexible Electronics Technology Conference (IFETC). Qingdao, China. https://doi.org/10.1109/IFETC53656.2022.9948506.

Xiao, Yongyue, Bei Jiang, Zihao Zhang, Shanwu Ke, Yaoyao Jin, Xin Wen, and Cong Ye. 2023. "A review of memristor: material and structure design, device performance, applications and prospects." *Science and Technology of Advanced Materials*. Vol. 24 (Issue 1). https://doi.org/10.1080/14686996.2022.2162323.

Xie, X., S. Jin, M. Wahab, et al. 2014. "Microwave purification of large-area horizontally aligned arrays of single-walled carbon nanotubes." *Nature Communications.* Vol. 5: 5332. http://dx.doi.org/10.1038/ncomms6332.

Xiong, Danrong, Yuhao Jiang, Kewen Shi, Ao Du, Yuxuan Yao, Zongxia Guo, Daoqian Zhu, et al. 2022. “Antiferromagnetic Spintronics: An Overview.” *Fundamental Research.* Vol. 2 (Issue 4): pg 522–534. https://doi.org/10.1016/j.fmre.2022.03.016.

Yadav, Dev Narayan, Phrangboklang Lyngton Thangkhiew, Sandip Chakraborty, and Indranil Sengupta. 2023. “Efficient Grouping Approach for Fault Tolerant Weight Mapping in Memristive Crossbar Array.” *Memories - Materials, Devices, Circuits and Systems*. Vol. 4: 100045. http://dx.doi.org/10.1016/j.memori.2023.100045.

Yang, J., D. Strukov, and D. Stewart. 2013. “Memristive devices for computing.” *Nature Nanotechnology*. Vol. 8: pg 13–24. http://dx.doi.org/10.1038/nnano.2012.240.

Yoon, Chansoo, Gwangtaek Oh, and Bae Ho Park. 2022. “Ion-Movement-Based Synaptic Device for Brain-Inspired Computing.” *Nanomaterials*. Vol. 12 (Issue 10): 1728. http://dx.doi.org/10.3390/nano12101728.

Yu, Zhouchangwan, et al. 2022. “Nanocrystallite Seeding of Metastable Ferroelectric Phase Formation in Atomic Layer-Deposited Hafnia-Zirconia Alloys.” *ACS Applied Materials & Interfaces*. Vol. 14 (Issue 47): pg 53057–53064. http://dx.doi.org/10.1021/acsami.2c15047.

Yue, Ruoyu, Yifan Nie, Lee A. Walsh, Rafik Addou, Chaoping Liang, Ning Lu, Adam T. Barton, et al. 2017. “Nucleation and growth of WSe_2: enabling large grain transition metal dichalcogenides.” *2D Materials*. Vol. 4 (Issue 4): 045019. http://dx.doi.org/10.1088/2053-1583/aa8ab5.

Zakhidov, Alexander A., Jin-Kyun Lee, John A. DeFranco, Hon Hang Fong, Priscilla G. Taylor, Margarita Chatzichristidi, Christopher K. Ober, and George G. Malliaras. 2011. “Orthogonal Processing: A New Strategy for Organic Electronics.” *Chemical Science*. Issue 6. http://dx.doi.org/10.1039/C0SC00612B.

Zhang, Dongli, Moussa Ehsan, Michael Ferdman, and Radu Sion. 2014. “DIMMer: A Case for Turning Off DIMMs in Clouds.” *SOCC ’14: Proceedings of the ACM Symposium on Cloud Computing*. New York: Association for Computing Machinery. http://dx.doi.org/10.1145/2670979.2670990.

Zhang, Liping, Jean-Francois de Marneffe, Markus H. Heyne, Sergej Naumov, Yiting Sun, Alexey Zotovich, Ziad el Otell, Felim Vajda, Stefan De Gendt, and Mikhail R. Baklanov. 2015. “Improved Plasma Resistance for Porous Low-k Dielectrics by Pore Stuffing Approach.” *ECS Journal of Solid State Science and Technology*. Vol. 4 (Issue 1): N3098. http://dx.doi.org/10.1149/2.0121501jss.

Zhang, Delin, Mukund Bapna, Wei Jiang, Duarte Sousa, Yu-Ching Liao, Zhengyang Zhao, Yang Lv, et al. 2022. “Bipolar Electric-Field Switching of Perpendicular Magnetic Tunnel Junctions through Voltage-Controlled Exchange Coupling.” *Nano Letters*. Vol. 22 (Issue 2): pg 622–629. http://dx.doi.org/10.1021/acs.nanolett.1c03395.

Zhang, Sirui, Qinghua Zhang, Fangqi Meng, Ting Lin, Binjian Zeng, Lin Gu, Min Liao, and Yichun Zhou. 2023. “Domain Wall Evolution in $Hf_{0.5}Zr_{0.5}O_2$ Ferroelectrics under Field-Cycling Behavior.” *Research*. Vol. 6 (Article ID 0093). https://doi.org/10.34133/research.0093.

Zhang, Zhiyong, Jianshuo Zhou, Li Ding, Lin Xu, Xiaohan Cheng, et al. 2023. “Terahertz Metal-Oxide-Semiconductor Transistors Based on Aligned Carbon Nanotube Arrays.” Preprint, submitted March 4, 2023. https://doi.org/10.21203/rs.3.rs-2526224/v1.

Zheng, Yi, Guang-Xin Ni, Ming-Gang Zeng, Shu-Ting Chen, Kui Yao, and Barbaros Özyilmaz. 2009. “Gate-controlled nonvolatile graphene-ferroelectric memory.” *Applied Physics Letters*. Vol. 94 (Issue 16): 163505. https://doi.org/10.1063/1.3119215.

Zhu, Jiadi, Teng Zhang, Yuchao Yang, and Ru Huang. 2020. “A Comprehensive Review on Emerging Artificial Neuromorphic Devices.” *Applied Physics Reviews*. Vol. 7 (Issue 1): 011312. https://doi.org/10.1063/1.5118217.

2.2 Circuits and Architectures

In the pursuit of energy-efficient computing, the design of new circuits and architectures plays a pivotal role. As new circuits and architectures are designed to address emerging computing needs while also adapting to evolving CMOS/IC, memory, and interconnect technologies, energy efficiency must be a core consideration.

There is a stark contrast in energy consumption between logic operations and memory access (as shown in Figure 7 in the Introduction). Compared to an Int8 ADD operation, accessing on-chip SRAM—which is closest to the processor and the most energy-efficient form of memory—can be up to 2,000 times more energy-intensive, while accessing off-chip DRAM can be up to 190,000 times more energy-intensive (Jouppi et al. 2021). The primary energy cost arises from the capacitive charging and discharging associated with data transfer between compute elements and memory, highlighting data movement as not only a performance bottleneck but also a major energy sink.

This chapter synthesizes the collective insights from the Circuits and Architectures working group, highlighting technologies that have significant potential for energy savings while also considering economic viability. While the technologies discussed represent a selection of the myriad options for improving energy efficiency within circuits and architectures, they exemplify the type of innovation required to meet the dual demands of performance and efficiency. The energy impact factors for each proposed technology, as compared with these technologies' current counterparts, are specific to their applications and critical to understanding their potential benefits.

Circuits and architectures bridge the gap between bits, instructions, and applications where technologies can apply to one or more of these defined hierarchical levels. Nonvolatile memory, for example, offers significant energy reductions at the bit level that also extend to the instruction level. Technologies like compute-in-memory (CIM) decrease energy consumption per instruction and enhance application performance through architectural innovations. Similarly, technologies such as the compute express link (CXL) enhance instructional energy efficiency and application performance by optimizing resource allocation. Finally, ASICs and domain-specific architectures (DSAs) elevate efficiency at both the instruction and application levels by tailoring hardware to specific computational tasks.

Working group methodology

The working group focused on high-impact technologies to improve energy efficiency and performance related to memory access, domain-specific and application-specific architectures, digital and analog compute-in-memory technologies, novel non-volatile memories, and EDA. To quantify the potential energy efficiency gains of these technologies, the working group conducted benchmarking through a literature search and compared the results to incumbent technologies.

Table 27 shows the proposed technologies organized by group. Specific energy contributions can be found in each of the following sections where applicable. Some technologies, such as EDA or instruction set architecture (ISA), do not contribute directly to energy consumption. However, proposed energy savings through utilizing these technologies are mentioned in their respective sections. While compute-near-memory is discussed, this was considered more of an integration scheme rather than a new architecture.

Table 27. Technology Groups Addressed by the Circuits and Architectures Working Group.

Technology Group	Specified Technology
Memory Access	CXL Fabric
	UCIe
	Instruction Set Architecture
Non-Volatile Memory (NVM)	NRAM
	ReRAM
	STTRAM
	PC-RAM
SRAM	Metis SRAM bit line variation reduction, energy reuse
Compute-near-Memory	Vcache
	MIV stacked ReRAM
	DRAM Cache
Compute-in-Memory (Digital)	SRAM CIM HBM PIM
Compute-in-Memory (Analog)	Neuromorphic
Domain-Specific Architectures (DSAs)	GPU
	TPU
	FPGAs
	Anton-3
EDA	Energy per bit simulations
	Advanced PDKs
	DTCO

Figure 30 shows the technologies of interest with their potential energy efficiency improvement factors and timelines to TRL 6, as determined by the working group. For more information on TRL6, refer to section 1.5.

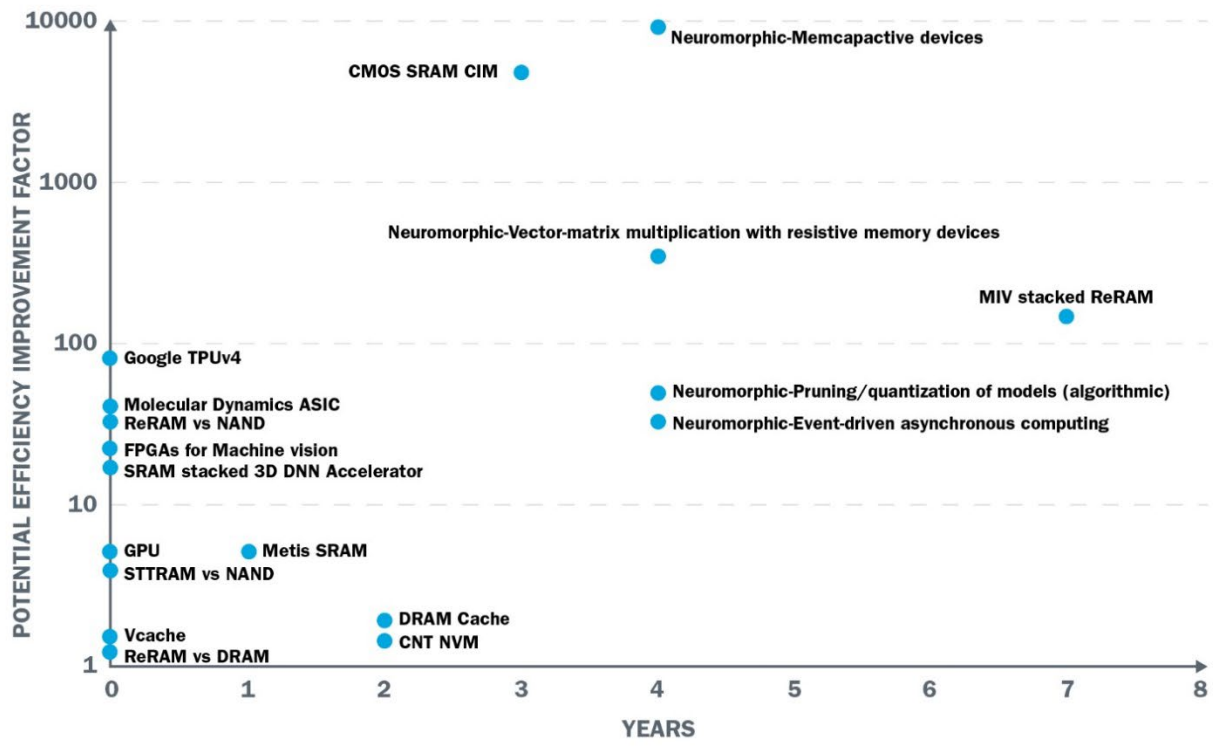
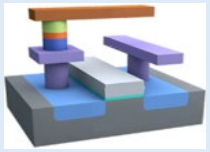
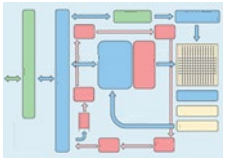


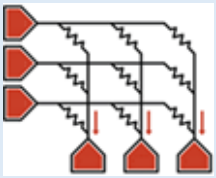
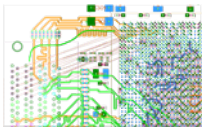
Figure 30. Potential efficiency improvement factor and timeline for selected technologies of the Circuits and Architectures working group.

Key takeaways

Table 28 summarizes the most significant identified energy efficiency opportunities that can be achieved through advances in circuits and architectures.

Table 28. Key Takeaways for Energy Efficiency Opportunities in Circuits and Architectures.

Technology Group		Key Opportunities for Energy Efficiency
Alternative NVM		• Utilize low-power NVM for AI weight applications with infrequent read requirements. • Improve energy per bit of read and write as well as density to be comparable with DRAM and NAND memories,
Domain-Specific Architecture		• Investigate large application spaces that would result in significant energy and performance savings. • Leverage the appropriate architecture for the specified use case.

Compute-in-Memory (Digital and Analog)		• Reduce the extreme barrier to entry by creating new algorithms and software, along with security protocols and advanced hardware, for both digital and analog CIM technologies. • Improve ADC to DAC overheads for neuromorphic computing to improve computational efficiency. • Evaluate IoT/Edge applications with alternative NVM to improve energy performance for power-limited operations.
EDA		• Create and use open-source EDA for advanced architecture and circuit development to reduce hidden overheads. • Co-design software and architecture to improve energy efficiency and reduce circuit overheads.

Grand challenges

Major challenges that must be overcome by circuits and architectures to achieve significant energy savings include:

- Enabling compute-in-memory, whether digital or analog, through the creation of new architectures, new security protocols, new EDA software for co-design, and significant development of instruction set architectures and new software/algorithms.
- Demonstrating non-volatile memory (NVM) technologies that are comparable in cost and density to DRAM or NAND, particularly when implemented in monolithic integration.
- Leveraging advanced EDA simulation software and co-design to create novel architectures and simulation of device function, whether through optimization of current structures or the use of alternative approaches such as 3D integration.
- Establishing R&D testing facilities to enable integration of novel materials and architectures with current state-of-the-art technologies, while also testing and evaluating energy efficiency and performance improvements.
- Minimizing energy overheads and enhancing overall performance by effectively educating the current workforce, particularly from academic institutions, about the importance of selecting the appropriate architecture for the right applications (e.g., opting for GPUs over CPUs for intensive image processing tasks).
- Reducing the total cost of ownership of new interconnect fabrics that can reduce overheads and optimize memory access.

2.2.1 Memory Access

Classical computers use a von Neumann architecture (see Figure 31), where a large fraction of CPU instructions involve moving data between CPU registers and memory. However, the classical architecture has limitations, given the pace at which processor speed has outstripped memory speed. To mitigate this, architects first incorporated fast cache memories located very close to the processor, and then multi-core designs offering parallel processing, with individual caches for each core to lower the memory access time. More recently, the industry moved to system on chip (SoC) and system in package (SiP) architectures to improve the time required to access memory beyond the on-chip caches.

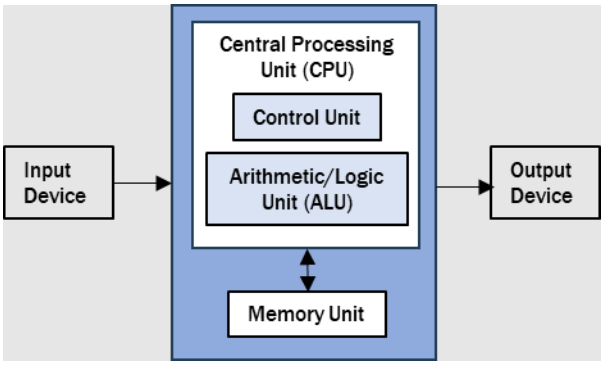


Figure 31. Classical von Neumann computer

However, cache memory, while fast and energy efficient, is not sufficient for all applications because its storage is small relative to its large 2D footprint. To further improve the memory read or write time, a hierarchy was created by computer architects based on three factors: access rate (in terms of clock cycles, or how many times transistors actuate per unit of time), storage size, and cost. Cache memory (SRAM) is built near the processor and is structured into levels (L1, L2, L3) based on storage capacity and access time. DRAM is off chip from the processor and is structured as the main non-immediate memory storage, slower than SRAM but with a much higher memory density. Long-term storage of information is in NAND or disk memory, which, unlike DRAM, retains stored information when power is removed. A pictorial representation of the memory hierarchy in modern von Neumann machines is shown in Figure 32.

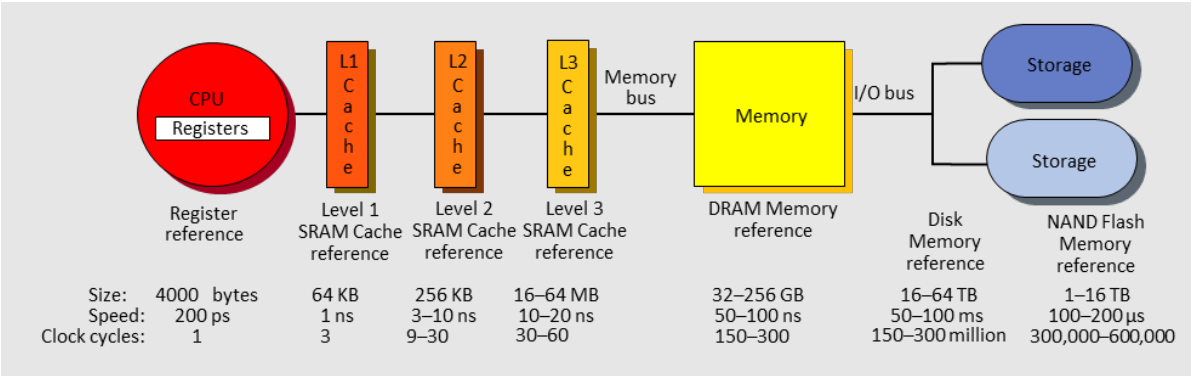


Figure 32. Typical memory hierarchy sizes and access times (c. 2019). *Variations on this hierarchy exist for other structures such as mobile, laptop, etc. Source: Hennessy and Patterson 2019; clock cycles assume a 3.33 GHz clock.*

Recognizing that the transfer of data between memory and compute is the largest energy consumer, exploring the following areas can yield efficiency improvements in memory access:

- Page size

- Interconnect fabrics
- Instruction set architecture
- Universal Chiplet Interconnect Express (UCIe)

As system architectures grow increasingly complex, featuring multiple memory domains and a variety of processors and accelerators, there is a pressing need for a cohesive framework that facilitates communication, prioritization, and resource allocation among these components. Traditionally, this coordination has been managed through bus systems, which are pathways that transmit data among the various components of a computer. However, the task of resource allocation—such as accessing open memory banks or SRAM when necessary—is complex. Moreover, managing the placement and transfer of data between memory and processors adds further complications. Interconnect fabrics like CXL offer a robust communication protocol that ensures coherency across processors, accelerators, and memory units (Bowman 2023). This integration not only accelerates communication but also enhances overall system performance and reduces costs. By enabling disparate components to function cohesively, such technologies transform a collection of individual parts into a seamlessly operating high-performance chip.

The adoption of advanced technologies in memory access significantly enhances performance, as demonstrated in Table 29. CXL technology, for instance, supports full duplex operations, offering lower latency and improved energy efficiency compared to traditional LPDDR DRAM, with potential energy savings up to a factor of 1.7. Furthermore, sophisticated interconnect structures can markedly decrease the energy consumed during memory access. Although not the primary focus of the Circuits and Architectures group, emerging interconnect standards such as UCIe are poised to dramatically reduce energy use per bit by 20–40x and significantly diminish latency due to shorter interconnects.

Table 29. CXL and UCIe Energy Impact Factor Comparison and Timeline for Improvements to Memory Access.

Sources: Sharma et al. 2022; Gervasi 2023; Gervasi and Chang 2023

Specified Technology	Baseline Energy Performance	Commercial Benchmark Product	Commercial Benchmark Energy Performance	Energy Impact factor	Timeline for Lab Scale Demonstration (TRL 6, years)	Projection on Data Centers
CXL optimized DDR5	5.8 GB/W for 64 GB	Standard DDR5	3.8 GB/W for 64 GB	1.5	0	6% power reduction, 2% cooling savings
CXL optimized DDR5	6.4 GB/W for 128 GB	Standard DDR5	4.9 GB/W for 128 GB	1.3	0	
CXL Native DRAM 8-lane PCIe Gen5	1.5–2.0 pJ/bit	2x LPDDR-6400 x 16	2.0–2.5 pJ/bit	1–1.7	0	N/A
UCIe (chiplet interconnect architecture)	0.5–0.25 pJ/bit	PCIe	10 pJ/bit	20–40x	1–3	N/A

There is also potential to achieve lower losses and higher transfer speeds through various printed circuit board (PCB) materials and connectors, but these choices also affect cost. The cost-performance trade-off must be considered in terms of total cost of ownership.

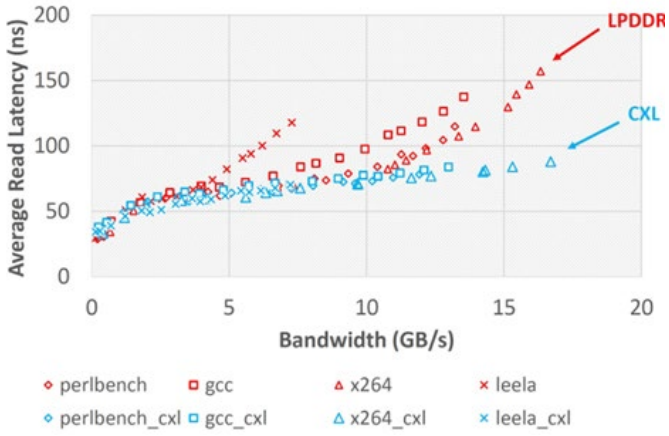


Figure 33 CXL Native DRAM 8-lane PCIe Gen5 vs. LPDDR 2x LPDDR-6400 latency vs. bandwidth comparison. *Source: Gervasi and Chang 2023*

Challenges and solution pathways for memory access

Reducing unnecessary bit overhead through page size adjustments

The page size is the lowest number of bits/cells of a memory architecture that can be accessed. A memory page of 8,192 bits requires all 8,192 bits to be accessed, when only 256 or 512 bits are needed. Utilizing a memory mechanism with a more granular page size enabled with a smart memory controller could potentially eliminate the overhead of unused bits. Another possible implementation is the memory "buffets" concept (Pellauer et al. 2019), which provides explicit, composable data transfers between a processor and the external memory, and access requests decoupled from the request receiver, thereby reducing or eliminating the need for on-chip buffering. The design of buffets has been publicly released in register transfer language (RTL) code and is flexible enough to fulfill the needs for memory access architecture in a variety of use cases.

Improving memory access granularity

Significant energy wastage occurs in program execution since only a small fraction of the page size is utilized, yet the entire page is pre-charged. This leads to all bit lines being energized, all sense amplifiers being employed to detect signals, and the complete transfer of all bits of information back to the memory bus. Possible solutions are to create memory controllers that have optimized closed page memories or more SRAM-like addressing. Implementing the DRAM address scheme to perform the Row Address Select (RAS) and Column Address Select (CAS) cycles without delay between them could help reduce activation and recharging required for memory access. In addition, eliminating open page mode for applications with low hit rates would reduce overhead by eliminating constant power to word lines of DRAM cells. Lastly, multi-page-sized memory could be a possible solution for reducing wasted energy through smart memory address buses; however, this may impact chip size and performance.

Optimizing power in system fabrics

As systems became more complex, a standard interface method to connect all the devices to the CPU and establish communication protocols was needed. The Peripheral Component Interconnect Express (PCIe) was created to help support this through interconnect standards of the PCBs and the component connections. What made PCIe popular was the ability for backwards compatibility between older devices and improved data transfer speed through a parallel bus architecture. With the emerging emphasis on energy efficiency and speed,

interconnect fabrics and components must be adapted. Possible solutions include exploiting new and upcoming interconnect standards, such as UCIe (Sharma 2022) and CXL. They are designed for reduced latency and employ memory pooling, which can significantly reduce data center memory requirements per device (Bowman 2023). These advanced interconnect technologies can be combined with improved PCB materials targeting reduction of dielectric losses at high frequencies and improving impedance matching to minimize signal reflections.

Reflecting true total cost of ownership (TCO)

Re-architecting memory for power optimization could lead to an increase in die size and, consequently, higher costs. Similarly, enhancements in PCB technology, such as improved connectors and board materials, might elevate initial expenses in favor of energy savings. These higher upfront costs must be evaluated in conjunction with the long-term savings derived from decreased energy consumption. Companies should promote the energy efficiency of these devices and articulate their total cost of ownership (TCO), which integrates reduced energy and cooling expenses, particularly in data centers. This comprehensive TCO perspective should also extend to consumer electronics, potentially through an initiative like the ENERGY STAR program (Energy Star 2024).

Integration of multiple IP stacks with reduced latency

The relentless miniaturization of chip technologies necessitates innovative approaches to enhance latency, yield, and energy efficiency. While striving for increased on-chip functionality to boost performance, it is crucial to acknowledge the challenges associated with larger chips, which often exhibit lower yields due to constraints such as die reticle size limitations. Additionally, transitioning to newer manufacturing nodes introduces complexities related to cost, time to market, and supply chain management (Sharma 2022). Chiplet integration offers a compelling solution to these issues. This design strategy allows for the incorporation of cutting-edge technologies alongside established ones, reducing time to market and enhancing energy efficiency through shorter interconnects. Chiplet architectures also enable the integration of diverse process technologies—such as different cores, memories, inputs and outputs, photonics, and mixed-signal components—into a single package, optimizing energy usage. Furthermore, UCIe's versatile interconnect standards facilitate compact device designs and closer component placement, enhancing communication speeds and overall system performance. Establishing UCIe-based chiplet technology standards will significantly influence both performance and energy efficiency.

Action plan for memory access

Table 30. Action Plan for Memory Access.

Scope	
Technology for Energy Efficiency	Memory Access
Technologies of Interest:	• All semiconductors utilizing memory outside of SRAM cache • System architectures such as data centers • Smart fabrics • Signal quality improvement of PCB materials
Challenges Addressed	**Solution Pathways**

• Interconnect fabrics optimization of resources. • PCB signal quality. • Memory overheads, such as page size.	• Reduce open page mode for low hit rates. • Implment arbitrary page sizes (e.g., multi-page-size DRAM with smart memory buses). • Optimize memory movement such as memory pooling. • Improve signal transmission, interconnect, and sockets on PCB to reduce resistance, capacitance, inductance, etc. • Utilize system fabrics with power optimization for performance vs. energy efficiency. • Find alternatives to copper for signal and power. distribution in PCBs such as CNTs.

Major Tasks/Milestones	Metrics	Targets	Timeline
Protocol changes for memory access optimization	Protocol approval	JEDEC, CPU/GPU/APU makers	3–5 years
Increase effectivity of memory semantic storage	Memory models for all devices adopted	Operating systems, hypervisors, applications	4–6 years
Improved PCB materials development	PCBs with improved signal transmission characteristics	All system buses	4–6 years
Improved socket development	Sockets with improved signal transmission characteristics	All system bus add-in strategies	4–6 years
Memory architectures that exploit SRAM-like command interface to reduce wasted access	Power-efficient memory	Memory suppliers	5–7 years
System architectures that optimize memory resources and minimize data movement	New system architectures	Operating systems, hypervisors, smart fabric	6–8 years
Increase use of energy-efficient data movement	Increased use of remote direct memory access (RDMA), data left in place when power states are available, smart fabrics that reduce traffic	System architects, fabric device suppliers, system software stack, hypervisors	6–8 years
Software and applications that are power-aware	Applications tuned for power conservation	Operating systems, hypervisors, applications	8–10 years

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• DRAM suppliers to define power-efficient core design options. • Processor/memory controller suppliers to support protocol for new memory types. • Standards organizations to consider power efficiency as a memory design requirement. • PCB suppliers to improve quality of materials.
End Users/OEMs	• End users evaluate TCO requirements in the context of power vs. performance. • End users to evaluate improved PCB materials that may require additional upfront capital but have backend energy savings.
Academia	• Architectural studies in memory architecture for power reduction. • Development of power-efficient data buses. • Development of power-efficient support circuits (phase-locked loop [PLL], etc.). • Exploration of new materials for PCBs, signal transmission.

Required Resources	Cross Collaboration Needs of Working Groups

• Innovation needed in on-chip power and signal distribution. • Improved PCBs' quality may require new materials. • System architects look for improvements to card edge connectivity. • Accurate power simulation models for memory architectures. • Device power requirements documented early in design phase. • Metrics for power utilization in TCO calculations. • Research reducing data movement and redundancy, especially exploiting new technologies such as NVMs. • Research in smart fabrics that use energy efficiency as a design target. • Develop alternatives to Cu for signal distribution to reduce power expended to achieve high signal quality.	• Circuits and Architectures: Include resource fabrics as key design goal for memory efficiency; remove open page memories with low hit rates. • Materials and Devices: Improve PCB material for improved signal loss and alternatives to Cu for signal and power distribution. • Algorithms and Software: Help with memory protocols to provide greater granularity of access, which will possibly require cooperation between the architecture, compiler, and operating system.

2.2.1.1 Interconnect Fabrics

The working group created an additional action plan for interconnect fabrics because of their importance to memory access. CXL is proposed as a replacement to PCIe given its open-source nature and its ability to target data centers, which collectively use 1–2% of global energy (Masanet et al. 2020). CXL leverages memory pooling, which is the ability to have multiple devices store data in the same memory bank, removing the need for excessive memory storage and, as a result, simplifying the software/algorithms (Woo 2021). It also includes new coherency protocols for accessing cache and device memory. Finally, CXL utilizes low latency connectivity, which is advantageous given the increase in memory needed for AI and other large memory workloads.

For further energy reduction, we propose utilizing UCIe with CXL, as well as replacing or supplementing PCIe where feasible. UCIe has been shown to reduce the energy per bit to 0.25 pJ/bit, a 40x savings over PCIe at 10 pJ/bit (Sharma et al. 2022). UCIe also allows for the designer to create SoCs and SiPs of chiplets from different sources or nodes for a complex design and function. In addition, UCIe could allow for integration of 2D and 2.5D integration schemes for improved energy per bit.

Action plan for interconnect fabrics

Table 31. Action Plan for Interconnect Fabrics.

Scope			
Technology for Energy Efficiency	Interconnect Fabrics		
Technologies of Interest:	Communication between data and processors in computing systems		
Challenges Addressed		**Solution Pathways**	
• Expand computing systems architectures with unified connectivity between resources. • Provide a standard electrical, protocol, and command structure usable by all resources.		• Utilize CXL (Compute Express Link), which provides a system-level interconnect standard with a fixed electrical and protocol solution as well as a limited number of physical modules and sockets to support it. • Adopt UCIe (Universal Chiplet Interconnect Express) to expand the CXL concept to chiplets. UCIe solutions likely use CXL interconnects for system interaction.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline**
Standard resource module: Storage	NVMe (NVM Express) moving to CXL, evolving with memory semantics	Data centers	Sampling now. Standards in 3 years.
UCIe adoption	Chiplet assemblies shipping, integrating devices from multiple suppliers	System-on-chip (SoC)	Now: CPU/GPU + HBM. 6 years: CXL-like chiplets. 20 years: order chiplets on Digikey.
CXL adoption	Systems shipping with CXL modules	Data centers	3 years
Standard resource module: Memory	Memory modules with CXL interfaces	Data centers	3 years: custom solutions. 6 years: standard commodities.
Memory semantic data access mechanisms	Continued development of Direct Access (DAX), OS support for memory pooling and sharing, applications using these modes	Data centers	5–10 years
Stakeholders and Potential Roles in Project			
Stakeholder	**Role**		
Product Manufacturers/Suppliers	• Provide compatible modules for fabric-based systems. • Provide compatible chips for modules. • Provide compatible chiplets for chiplet assemblies.		
Academia	• Migrate from filesystems to memory mapped data.		
Standards Organizations	• Define standards for CXL, UCIe, and chiplets.		
Required Resources		**Cross Collaboration Needs of Working Groups**	
• Power requirements and TCO analysis communicated to suppliers, end users, and standards bodies. • Operating systems, drivers, and education. • Research in improved PCB materials, replacements for Cu for signal and power distribution. • Software development of memory semantic access such as DAX must continue and not get stalled despite discontinuation of Optane.		• Algorithms and Software: Develop software frameworks that support the latest interconnect technologies to ensure compatibility and performance optimization across different platforms. • Metrology and Benchmarking: Ensure standard specifications are aligned to enable seamless interoperation among different interconnect technologies.	

2.2.2 Digital Compute-in-Memory

In traditional computing architectures, logic gates are fundamental components that perform basic operations on binary data—essentially, the ones and zeros that are the backbone of digital systems. These gates process signals and generate outputs based on inputs, enabling the execution of complex computational tasks. For example, a simple operation like addition or the comparison of two numbers is carried out by a series of logic gates interpreting and processing binary data.

The energy and latency cost to access memory in conventional computing architectures, paired with the increase in AI and ML (Mehonic and Kenyon 2022) and their immense off-chip memory access requirements (Biswas and Chandrakasan 2019), contribute significantly to the overall energy consumption and latency issues in microelectronics. Hence, new architectures such as compute-in-memory (CIM) are required to not only significantly lower the unsustainable energy usage of AI/ML technologies, but also improve performance.

In CIM architectures, data is both processed and stored within memory cells, eliminating the need to constantly move data back and forth between the memory and the processor—a process that typically consumes large amounts of energy. This method is particularly advantageous for operations that involve repetitive calculations such as vector multiplication, which is a basic mathematical process often used in computing (Agrawal et al. 2018; Kim et al. 2021; Lin et al. 2022).

CIM technology can utilize various types of memory, such as SRAM (static random-access memory), DRAM PIM (dynamic random-access memory processing-in-memory), STTRAM (spin-transfer torque random-access memory), and FeRAM (ferroelectric random-access memory). These memory cells need to be organized in a way that allows them to perform calculations effectively, similar to how logic gates in traditional computing are arranged to perform operations.

However, standard SRAM designs, which typically use a structure with six transistors per memory cell, are not suitable for complex operations like matrix multiplication, due to potential disruptions when reading and writing data. Thus, new SRAM architectures that can handle these computations without such disturbances must be developed. This need for new designs applies not just to SRAM but to all memory technologies if they are to support CIM effectively.

Utilizing CIM technologies is projected to yield an estimated energy efficiency improvement of 21,000x, as shown in Table 32. These projections, highlighted in Marvin Chang's presentation at the 2022 IEEE VLSI symposium (Chang 2022), suggest that adopting the Metis Microsystems approach could improve the balance between energy consumption and processing speed (energy-delay product) by approximately 10x (Bhavnagarwala 2021). The Metis Microsystems approach strategically integrates advanced computational and memory components to increase energy efficiency and minimize the delay in data processing tasks. This approach particularly emphasizes SRAM due to its superior performance benefits compared to other non-volatile memory (NVM) technologies. While SRAM-based CIM offers considerable advantages in speed and energy efficiency, making it highly suitable for applications in accelerators where memory such as register files and cache are predominantly SRAM-based, it is also essential to explore alternative CIM architectures (Gao et al. 2017; Sze et al. 2017). This exploration is crucial for adapting to diverse application needs such as IoT devices, which may

face space constraints, or systems managing extremely large datasets where the properties of SRAM might not yield the most energy-efficient outcomes. Pursuing development in other memory technologies like DRAM, NAND, or other non-volatile memories will ensure that energy efficiency gains are maximized across a wider range of computing scenarios.

Table 32. Comparison of SRAM-Based CIM at 1-Bit Precision. *Current (Chang 2021) and projected energy impact of utilizing efficiency improvements on SRAM technologies (Bhavnagarwala 2023), as compared to an average of 1-bit precision of currently available accelerators (Shankar and Reuther 2022).*

Specified Technology	Tops/W 1-Bit Precision	Baseline Energy Performance	Commercial Benchmark Product	Commercial Benchmark Energy Performance	Energy Impact Factor	TRL 6 Timeline (years)
advanced CMOS SRAM CIM + analog multiply-accumulate (MAC)	9,600 (current), 96,000 (mature)	0.1 fJ/op (current), 0.01 fJ/op (mature)	current commercial AI/ML accelerators	63 fJ/op (1-bit precision)	480 (current), 4,800 (mature)	1–3
advanced CMOS SRAM CIM + digital MAC	32,000 (current), 320,000 (mature)	0.03 fJ/op (current), 0.003 fJ/op (mature)	current commercial AI/ML accelerators	63 fJ/op (1-bit precision)	2,100 (current), 21,000 (mature)	1–3
NVM	1,440	0.7 fJ/op	current commercial AI/ML accelerators	63 fJ/op (1-bit precision)	90	3–5
DRAM	1,120	0.9 fJ/op	current commercial AI/ML accelerators	63 fJ/op (1-bit precision)	70	3–5

Challenges and solution pathways for digital CIM

Active energy and latency overhead from bitcell transistor variability

Active energy and latency are significantly impacted by the variability in memory bitcells, which are the fundamental storage units in memory arrays. Each bitcell's performance varies, affecting the speed and accuracy of reading data. Slower bitcells may consume more of the initial energy supply (precharge) from other cells through leakage mechanisms—unintended electrical flow that depletes charge. This issue often necessitates increased voltage to accurately write data to these less responsive, or "worse," cells.

Because memory plays a crucial role in computing speed, addressing these variations is essential. About 70% of an ASIC's energy is consumed by SRAM buffers and register file arrays, which manage data temporarily for quick access (Gao et al. 2017). Reducing the energy lost to inefficient bitcells and leakage can therefore lead to significant improvements in overall energy efficiency.

The Metis Microsystems approach mentioned earlier includes innovations like harvesting energy from data movements and employing self-regulating circuits. These strategies aim to significantly reduce energy wastage. This approach not only addresses energy consumption but also optimizes the operational latency of memory-intensive tasks.

Power delivery

Power delivery in SRAM-based CIM architectures involves challenges that need to be addressed for efficient operation. SRAM CIM is typically structured around an advanced base cell design, such as an 8-transistor setup, which fundamentally alters how power is supplied and managed within the memory chip.

Firstly, SRAM CIM operations frequently activate numerous word lines (WL), the control lines that select memory cells for reading or writing. This requires robust power delivery systems for the WL drivers, which are circuits that activate these lines. However, noise—fluctuations in electrical signals—generated by WL activations can interfere with the computational processes, introducing errors. To mitigate this, implementing a denser power grid has been proposed. This approach distributes power more uniformly across the chip, helping to stabilize voltage levels and reduce computational noise.

Secondly, activating many bit cells simultaneously for static and dynamic computing tasks generates considerable noise in power delivery. Such noise issues are also prevalent in alternative CIM structures that employ NVM technologies. These challenges highlight the need for innovative power delivery solutions to ensure reliable and accurate memory operations in advanced CIM architectures (Verma et al. 2019).

Architectural challenges related to the specific use case

CIM technologies excel at performing matrix vector operations rapidly, which is essential for tasks like image processing and machine learning. However, these operations represent just a fraction of the computational needs. CIM systems often require integration with other technologies to handle other types of computational tasks. This integration demands careful management of non-idealities that arise from analog signals, especially those affected by temperature and voltage variations. For CIM to function effectively alongside different accelerators and ASICs, it is crucial to ensure that these components can communicate seamlessly, adapting to the unique signal requirements of each device. This compatibility and efficient communication between various components are critical for the successful deployment of CIM technologies.

Disruption to 50 years of software

For the past 50 years, von Neumann architecture has been the choice for all computer architectures, and all software has been built with compute and memory separated. While the hardware has specific challenges, including power delivery, CIM array structure, ADC to DAC limitations, and lack of new EDA software for CIM layout and simulation, the largest bottleneck is software. Software must map its automatic code generators in the compiler stack to the hardware for optimal functionality. This will require new compilers and likely new instruction set extensions as well. Additionally, even with CIM technologies, conventional storage—whether DRAM/NAND or others—is still needed, and the transfer of information to and from CIM architectures must be programmed.

CIM size-related challenges

While CIM using SRAM significantly reduces energy consumption by eliminating the need for data transfers between the multiply-accumulate (MAC) unit and cache memory, it necessitates the use of larger MACs within the accelerator. The MAC, responsible for performing arithmetic operations essential for processing tasks, becomes considerably larger due to the multiple transistor configurations used in SRAM-based CIM. No single architecture has yet emerged as dominant, leading to increased sizes of MACs. Additionally, SRAM transistor scaling is not progressing as rapidly as the latest node transistors are due to technological and manufacturing challenges, which limits the miniaturization and efficiency improvements typically seen in newer semiconductor technologies (Heyman 2023). As SRAM continues to scale down, leakage currents increase, which requires more standby power. This becomes a challenge for edge or IoT devices that have limited space and power resources. A potential solution lies in adopting smaller non-volatile memory technologies such as STTRAM, FRAM (ferroelectric random-access memory), spintronics, or ReRAM (resistive random-access memory), although these alternatives also come with their own set of challenges. For more details on non-volatile technologies, refer to the Materials and Devices chapter.

Action plan for digital CIM

Table 33. Action Plan for Digital CIM.

Scope	
Technology for Energy Efficiency	Compute-in-Memory
Technologies of Interest:	• Architecture for software to target CIM platforms for efficient utilization of CIM arrays. • Energy and latency cost of moving data from local memory along wire paths for each computation. • Bitcell transistor and read current variability limitations on performance, energy efficiency, and accuracy of CIM arrays.

Challenges Addressed	Solution Pathways
• Reduction of memory data access and movement for MAC operations. • ADC overheads and bitcell variability limiting TOPS/W and accuracy.	• Using CIM architectures would significantly reduce the data transfer between processor and memory. Two-thirds of memory energy overhead could be eliminated. • Developing new architectures that do not require ADCs and eliminating overheads of bitcell transistor variability could enable CIM arrays to reach much higher efficiencies (3,000–5,000 TOPS/W). • Harvesting electrostatic energy with self-regulating circuit action can reduce the energy consumed by a read access in a CIM array by >80% while doubling performance.

Major Tasks/Milestones	Metrics	Targets	Timeline
Demonstrator of circuits to minimize the overheads of CMOS memories from bitcell transistor and read current variability	Bitpath energy use/read access; latency of bitpath from word line select to capture of data at array output	10x improvement in the energy-delay product metric	12 months
Demonstrator of circuits to minimize costs of moving data along local and global bitpaths in large CMOS memories from harvesting evaluation energy	Bitpath energy use/read access; latency of bitpath from word line select to capture of data at array output	10x improvement in the energy-delay product metric	12 months
Demonstrator of array peripheral circuits using digital in-memory arithmetic operation eliminating ADC overheads incurred from analog approaches	TOPS/W (maximum possible # of OPS/Joule that can be accomplished with the CIM array)	5,000 TOPS/W	12 months

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• EDA tool vendor (Synopsys, Cadence, Siemens, etc.) and foundry providing PDKs relevant to multi-project wafer (MPW) of test chip.
End Users/OEMs	• Industry members (fabless, foundry, and integrated device manufacturers [IDMs]) and government labs designing their own chips.
Academia	• Interdisciplinary: Johns Hopkins University (JHU) Applied Physics Laboratory (APL) could enable CMOS memory solutions to be more competitive in a traditionally NVM leadership domain of 'always-on' availability for fast cognitive wake-up function.

Required Resources	Cross Collaboration Needs of Working Groups
• EDA tools for designing SRAM/RF/CIM arrays and custom arithmetic components, including access to a CMOS platform PDK, servers equipped with these tools, and MPW test chips for creating demonstrator chips. • Detailed specifications for new circuits and architectures, with a comparative analysis against standard industry designs. • JHU APL's potential contribution of thermoelectric energy generator (TEG) integration into chip packages to enhance CMOS memory retention energy, offering a more competitive solution than NVM.	• Education and Workforce Development: Principal investigators (PIs) should seek partnerships across industry and academia. Workforce development is enhanced when interdisciplinary programs offer targeted internships for university students. National laboratories, with their extensive lab and computing resources, are ideal for developing prototypes for these programs.

2.2.3 Analog Compute-in-Memory/Neuromorphic Computing

While SRAM-based CIM uses digital memory techniques, or discrete values, alternative CIM architectures are viable using analog technologies with continuous (non-discrete) values. Analog CIM uses architectures with non-volatile analog memories as synaptic weights (Z. Wan et al. 2022) or signal combinations at each node to determine the output variable. These

architectures can be built using various NVMs or specialized circuitry tailored for different mathematical functions, effectively creating an analog computer for dedicated applications.

Of particular interest is the application of neuromorphic computing, or brain-inspired computing. This technique uses neural networks that model the structure of the brain, creating artificial neurons such as memristors (Chu 2020; Kumar et al. 2022), spintronics (Grollier et al. 2020), phase change memory (Sebastian et al. 2017), SRAM (Jhang et al. 2021), and more. It should be noted that neuromorphic structures do not need to be purely digital or analog. They can be a combination of both, such as SRAM with an analog MAC. Neuromorphic computation can occur through spike encoding or spike compute through spiking neural networks (SNN). SNNs for AI and ML utilize biological models of neurons to carry out computations or pattern recognition in a more energy-efficient manner compared to conventional deep neural networks (Yamazaki et al. 2022) by having compute and memory at the same location.

Table 34 illustrates that neuromorphic computation offers substantial energy savings. Simulations predict that highly parallel memcapacitive devices could achieve up to 9,000 times the energy savings per operation over traditional accelerators. For analog neuromorphic systems, projections show a potential 350-fold energy savings. In the digital realm, SRAM-based CIM adapted for neuromorphic architectures could provide up to 2,100 times, and potentially even greater, energy savings compared to existing accelerators in the near future. Notably, digital CIM technologies offer higher precision than their analog counterparts (Mehonic and Kenyon 2022).

Table 34. Neuromorphic CIM Technologies Compared to Current Commercial AI Accelerators at 1-Bit Precision.

Sources: Demasius, Kirschen, and Parkin 2021; Zimmer et al. 2020; W. Wan et al. 2022; Krishnan et al. 2022; T. Xiao et al. 2022; Chang 2022; current and projected efficiency improvements on SRAM technologies from Bhavnagarwala 2021, 2023.

Technology Group	Specified Technology	Baseline Energy Performance (1-bit precision)	Commercial Benchmark Product	Commercial Benchmark Energy Performance	Energy Impact Factor (X Factor)	Timeline for Lab Scale Demonstration (TRL 6, years)
Compute-in-Memory Architectures (Neuromorphic)	Memcapacitor devices enabling parallel MAC operations	0.007 fJ/op (simulation)	current commercial AI/ML accelerators	63 fJ/op (1-bit precision)	9,000	3–5
	Neuromorphic-Vector-matrix multiplication with resistive memory devices	5.8 fJ/op (current), 0.18 fJ/op (mature)	current commercial AI/ML accelerators	63 fJ/op (1-bit precision)	11 (current), 350 (mature)	3–5
	Neuromorphic-Pruning/quantization of models (algorithmic)	5.8 fJ/op (current), 1.3 fJ/op (mature)	current commercial AI/ML accelerators	63 fJ/op (1-bit precision)	11 (current), 48 (mature)	3–5
	Neuromorphic-Event-driven asynchronous computing (clockless) for deep learning	5.8 fJ/op (current), 1.9 fJ/op (mature)	current commercial AI/ML accelerators	63 fJ/op (1-bit precision)	11 (current), 33 (mature)	3–5

Compute-in-Memory Architectures (Digital)	advanced CMOS SRAM CIM + Digital MAC	0.03 fJ/op (current), 0.003 fJ/op (mature)	current commercial AI/ML accelerators	63 fJ/op (1-bit precision)	2,100 (current), 21,000 (mature)	0–3

Challenges and solution pathways for analog CIM/neuromorphic computing

Spike encoding and computation

In contrast to other neural networks, spike encoding neural networks (SNNs) use signal timing to convey information. This more closely mimics the brain's synaptic responses through time and its ability to transfer information between neurons. In addition to emulating the brain more closely, SNNs are more powerful than traditional artificial neural networks (Zhang et al. 2022). SNNs can allow for immense energy savings of between 100,000–300,000x over continuous value networks and three orders of magnitude over CPUs, depending on the task (Zhang et al. 2022).

However, significant challenges remain for SNNs. For example, analog devices are not yet robust enough for long-term use (cycling) compared to CMOS transistors (Merolla et al. 2014), necessitating further investigation into how to better develop and enable such devices. (For more information on neuromorphic devices, see Section 2.1.7.) Programming of images or speech for SNNs will require new programming methodologies/languages or translation from existing ANNs. Lastly, because SNNs are still relatively new, investigation into viable applications is ongoing.

ADC and DAC overheads limiting TOPS/W

Any compute-in-memory structure must communicate with a microprocessor using digital inputs and outputs. Therefore, continuous analog signals from SRAM CIM signals must be converted to digital values. Current challenges with analog-to-digital converters (ADC) and digital-to-analog converters (DAC) include:

- Increased IC area from DAC inputs and outputs from the analog memory arrays (Xiao, Jiang, and Chee 2022).
- Multiple devices per input/output circuit, requiring heavy power use (Amirsoleimani et al. 2020), with some ADC accounting for up to 92% of total power consumption of the circuit (Yao et al. 2020).
- Difficulty in achieving more than four-bit accuracy (Danial, Sharma, and Kvatinsky 2020).

To move past size, speed, accuracy, and power concerns for digital CIM with analog MACs for best energy performance, the EES2 working group and others (Shafiee et al. 2016; Zhang, Huang, and Shen 2020) suggest using digital technologies for immediate implementation to avoid overhead in the near term. Some techniques are already being employed, such as improved conversion algorithms and circuit design innovation, to improve ADC to DAC overhead by ~7.5x (Danial, Sharma, and Kvatinsky 2020). Research in this area should continue, especially for neuromorphic computing.

Electronic Design Automation Tool Development

Currently, there is no open-source EDA software suitable for neuromorphic design, nor are there standards set by the community for which device structures and architectures are the

most viable for neural networks. An open-source or academic-licensed EDA software with up-to-date process design kits specific to advanced neural network structures would allow for advancement of neural networks through simulation for speed and energy efficiency, as well as application space-testing prior to hardware creation. In addition, co-designing algorithms with devices and architectures can address the ADC to DAC power and area overheads before device creation (Christensen et al. 2022).

Table 35. Action Plan for Analog CM/Neuromorphic Computing

Scope			
Technology for Energy Efficiency	Neuromorphic and analog computing		
Technologies of Interest:	Potential areas of interest include edge applications, SoC, scientific computing (e.g., climate forecasting, detectors, etc.), sensing, IoT, AI hardware (inference, on-chip learning), AI/ML, wearables/embedded devices, healthcare (e.g., smart/body-integrated sensors), and smart homes.		
Challenges Addressed		**Solution Pathways**	
• Hardware intrinsically distributed through neuronal processing (spike encoding and spiking compute). • Plasticity at neuronal circuit level and architecture level (removes bottleneck of memory accesses). • Harnesses multiple neuronal/sub-neuronal units that operate at low precision. • Opportunity to bring computing closer to physics domain, which provides more feasibility to heterogeneous architectures (low power, so can be closer to other compute devices), data bandwidth advantage (data preprocessing to reduce processing requirements), and takes maximum advantage of 3D architecture.		• Focus on developing robust neuromorphic architectures (devices/hardware needed for real-time learning) and then scale up architectures/systems for complex or large compute applications (scalable learning rules). • Engage users by developing appropriate benchmarks/applications of analog computing. • Create tools for programming scalable neuromorphic systems. Software system is lacking. Better connections with ML community (large LLM); neuromorphic should not evolve separately. Open-source tutorials and repositories. • Develop EDA tools optimized for neuromorphic and related approaches.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline**
Benchmark current neuromorphic systems	Learning and inferencing standard datasets vs. existing hardware	• Increasing access to neuromorphic systems for testing • Benchmarking against alternative ASICs • Implementing a specific number of neurons and degrees of free parameters, with the timeline depending on when scale definitions are established.	• 3–5 years • 1–2 years (depends on benchmarks) • 2–3 years (commercially viable chips/systems)
Develop benchmarks that stress different aspects of neuromorphic system	Metrics and targets will depend on the type/application of neuromorphic system	• Edge accelerators: < 1 milliwatt (mW), 4 bit precision, > 1 Giga operation per second (GOPS/s) • General compute: 4–8 bit precision, up to 10 GOPS/s • HPC: Often, high bit precision is not required; typically, adaptive precision uses 4–8 bits but can increase as needed, facilitating a new computing paradigm that enhances speed and energy efficiency beyond exascale.	• 5 years (edge) • 5–10 years (general compute) • 10–20 years (HPC) • (Also possible that all three categories could evolve on similar timescales)
Develop EDA tools the community can agree upon and use together. Open-source or academic licensing.	System/network scale	Toward millions or tens of millions of synthetic neurons	2–3 years
Availability of high-level programming language	High-level language development	More intuitive programming; high-level abstraction	3–5 years

Extreme increase in power efficiency	Energy per benchmark application	• 10x • 1,000x • Much larger; approach or exceed human brain efficiency	• 2–3 years • 5–10 years • 10–20 years
Removing energy barrier for neuromorphic devices	Joules/operation, total energy per program/algorithm, average power (single W and lower, ~100 mW)	• Reducing the ADC to DAC penalty ensures compatibility with Boolean hardware. • Staying in the analog domain before output (removing analog conversion). • Transitioning to neuromorphic code (image recognition, time series classification). • Requiring adaptive precision for training and inference due to precision issues.	• Tesla has claimed success with end-to-end neuromorphic. • 2–3 years in some cases (80%–90% chip running in neuromorphic) • 5–10 years (90%+ chip running in neuromorphic) • 10+ years (all running 100% neuromorphic)
Unconventional computing (avoiding digital)	Joules/operation, total energy per program/algorithm, average power (single W and lower, ~100 mW)	Exploring alternative methods of computation such as reservoir computing, chaotic computing, coupled oscillator computing, and cellular state machine computing. These approaches overlap with some aspects of CIM and offer potential for significant improvements in energy efficiency.	Most of these technologies are in the early stages of research and development, with a commercial viability horizon of 10+ years.

Stakeholders and Potential Roles in Project

Stakeholder	Role
Product Manufacturers/Suppliers	• Develop cost-effective neuromorphic devices. It is important to collaborate as needed so that a software stack is also available for ease of use.
End Users/OEMs	• Act as system integrator to incorporate neuromorphic devices into large systems for wide use; again, the software stack is available for heterogeneous system.
Academia	• Provide methods and techniques to facilitate ease of use of neuromorphic devices/architectures. • Develop applications that efficiently utilize neuromorphic architectures.
National Laboratories	• Provide methods and techniques to facilitate ease of use of neuromorphic devices/architectures. • Develop applications aligned with DOE mission that efficiently utilize neuromorphic architectures. • Act as both consumer and producer of neuromorphic technology. • Act as testbed host and provider.
Government	• Provide the support needed for a neuromorphic ecosystem: hardware, software, applications, and workforce.

Required Resources	Cross Collaboration Needs of Working Groups

• EDA tools and open-source design tools, startup and project seed funding, fablets and P-line facilities, and employee training. • Access to evaluation systems and testbed hardware. Well-integrated solutions: chips, board, rack, and software. Access to example designs/use cases. • Resources to hire new faculty and create new courses, certifications, and degree programs. • Resources for hiring well-trained students and postdocs. Funding for larger projects and initiatives (center level). • Access to latest information to help in outreach to public and policymakers. Access to well-trained grad students and postdocs for program development and management.	• Materials and Devices: Develop robust, energy-efficient switching mechanisms; explore new memory and memristive materials; integrate neuronal behaviors such as non-linear response and spike plasticity with CMOS technology. • Algorithms and Software: Establish a more mature programming environment for neuromorphic systems, addressing the absence of robust software frameworks and the potential need for new programming languages. • Metrology and Benchmarking: Standardize methods to meaningfully compare different neuromorphic hardware. • Manufacturing: Optimize deposition and fabrication processes for neuromorphic devices, ensuring CMOS compatibility and integration of diverse materials. • Education and Workforce Development: Develop educational programs and tools at various levels from K-12 to college to foster expertise in neuromorphic computing; promote interdisciplinary training and create new open-source curricula to enhance recruitment and awareness in hardware-related fields.

2.2.4 Nonvolatile Memory

The general memory architectures have evolved slowly over the last 40 years, e.g., with larger and multi-level on-chip caches, and NAND memory gradually replacing traditional magnetic disks. SRAM serves as a crucial component in this evolution, primarily used as cache memory due to its higher speed compared to DRAM, despite its higher cost per bit and larger cell size which limits its density. DRAM remains the primary volatile memory, given its combination of speed, ease of manufacturing, cost per bit, and continued planar scaling. NAND, the primary nonvolatile storage, has achieved an extremely high density, and its cost per bit is lower than other memories. While DRAM and NAND are unlikely to be replaced with new upcoming technologies, it is important to understand their strengths and shortcomings, to anticipate how emerging memory technologies might be advantageously incorporated into the memory hierarchy, depending on the application.

DRAM continues to scale via shrinking of the cell pitch size, but fundamental scaling issues such as leakage current are forcing DRAM to move to 3D structure (Pires 2023). Additionally, DRAM requires constant refreshing of its bitcells to maintain data, which consumes 30% of its energy. Continued scaling of DRAM is only increasing the refresh rate and associated energy use. DDR4 and DDR5 are different generations of DRAM; DDR4, the 4th generation, has a refresh rate of 64 milliseconds (ms), which means it must renew its stored data every 64 ms to maintain its integrity. DDR5, the 5th generation, has an improved refresh rate at 32 ms (Vogelsang et al. 2022), allowing for more frequent data renewal. This increase in refresh rate attempts to mitigate energy inefficiency, especially from standby power, yet the faster refresh rate also presents challenges in maintaining the energy efficiency gains achieved through advancements in scaling and design (Vogelsang 2010).

NAND energy efficiency improvements are made through geometric shrinking of the memory cell and with multi-tier stacking of more memory cells. The primary issue for NAND is the access energy cost, which is nearly 10 times that of DRAM at 100 pJ/bit (Pawlowski 2023). This value does not include the cost of access through the interconnects, which adds significant overhead.

Although DRAM and NAND will likely remain the dominant forms of memory, supplementary memories can help reduce access and standby power consumption. Memories such as MRAM (STTRAM), FeFET, ReRAM, and NRAM are viable next-generation technologies. Most of them offer significant reduction in energy per bit and improved speed compared to NAND memory. Compared to DRAM, these technologies offer similar speeds with similar read and write costs and can be placed closer to the processing unit. Most importantly, they do not suffer from memory volatility and can store data without power supplied to the cell, saving nearly 30% of energy costs (or more if not accessed within 32 ms). This makes them particularly useful for neural network applications (Veksler et al. 2020; Mukherjee et al. 2021; Chang 2021), high-radiation environments (Nantero 2023; Marinella 2021), and IoT devices where space and power are limited (Saito et al. 2021).

A comparison of energy and performance metrics of conventional memory architectures and next-generation NVM technologies is shown in Table 36. The energy impact factors compared to NAND and DRAM are shown in Table 37. All technologies have the potential to improve energy efficiency, read/write times, and durability compared to NAND; however, the biggest issue they face in replacing NAND is density. STTRAM, NRAM, and RRAM have the potential to improve on the energy cost per bit of DRAM. Not shown in the table are the improvements through eliminating a refresh every 32 or 64 ms, which is at least 30% of the cost of DRAM operation. They can also be monolithically integrated with logic, which potentially produces a significant energy reduction cost over a DRAM-based GPU (An et al. 2022). Lastly, while SRAM is highly energy-efficient, it demands continuous power for data retention and occupies a larger area. In contrast, NVM technologies do not require constant power, allowing for either monolithic integration or achieving densities up to 10 times greater than SRAM. This capability from NVM technologies significantly alleviates the energy bottleneck associated with accessing higher levels of cache or DRAM for additional memory (Gopireddy and Torrellas 2019; Hankin et al. 2019).

Table 36. Comparison of Conventional Memory Architectures to Alternative Nonvolatile Memories. *Due to variation among reported data, these values should be taken as estimations. Sources: Marinella 2021; Pawlowski 2023; Bhavnargawala 2023; Yu 2016; Zhang et al. 2021; Vogelsang et al. 2022; Chatterjee et al. 2017; Sivan et al. 2019; Pan and Naeemi 2017.*

Parameter	SRAM	DRAM	NAND	STTRAM	ReRAM	NRAM (NVM)	NRAM (AI NVM)	PCRAM
Cell Area (F^2)	>100	<6	<4	6–20	<4 if 3D	4	8	~4
Voltage (V)	<1	<1	>10	<1	1–3	<1–3	<1–3	1–3
Read Time (nanoseconds [ns])	<1	10–20	10,000	~10	~10	15	2.5	~10
Write Time (ns)	<1	10–20	10,000	~13	~2–10	1,000	40	~50
Retention	N/A	~32 ms	Years	Years	Years	Years	Years	Years
Read/Write Energy	~18 fJ/bit	5–10 pJ/bit	>100 pJ/bit	299 fJ/bit for write	2–13 pJ/bit (lowest energy states)	0.2 pJ/bit for read, 30 pJ/bit for write	0.4 pJ/bit for read, 60 pJ/bit for write	>100 pJ/bit

Endurance (cycles)	>1E16	>1E16	>1E4	~1E12	~1E12	~1E9	~1E9	~1E7

Table 37. Energy Impact Factors of NVM Technologies Compared to DRAM and NAND.

Technology	Energy Per Bit (Array Level)	Energy Impact Factor Compared to DRAM	Energy Impact Factor Compared to NAND
STTRAM	0.299 pJ/bit (write)	16–32	333
ReRAM	2–13 pJ/bit (lowest energy states)	0.77–5	7.7–50
NRAM (NVM)	0.2 pJ/bit (read), 30 pJ/bit (write)	25–50 (read), 0.17–0.33 (write)	3–50
NRAM (AI NVM)	0.4 pJ/bit (read), 60 pJ/bit (write)	12.5–25 (read), 0.08–0.17 (write)	1.7–25
PCRAM	>100 pJ/bit	N/A	1

Challenges and solution pathways for non-volatile memory

Support from processors through application space

The primary challenges for NVM integration are the infancy of applications that use it, the lower endurance compared to DRAM, and the lower density compared to NAND. However, NVM technologies do offer improvements in energy cost over NAND and approach energy cost parity with DRAM, while avoiding overhead energy from refresh. Understanding which applications NVM could be used for will influence the rate of adoption. Examples of such applications include AI/ML (Chakraborty, Gupta, and Suri 2020; Chang et al. 2021; Mukherjee et al. 2021), 3DICs for compute-near-memory (Hosseini et al. 2022), and IoT (Saito et al. 2021). In addition, NVM provides new use cases in conditions where conventional memory may break down, such as high temperature, high shock, and high radiation environments (Marinella 2021; Strenz 2020). As these use cases become better understood by the community, circuitry and architecture improvements (Mukherjee et al. 2021) and integration steps (such as CXL adoption) can increase NVM adoption and reduce energy consumption of conventional memories through new and hybrid designs.

Electronic Design Automation Tools and Process Design Kits for Application Space and Total Cost of Ownership Analysis

EDA tools and PDKs are primarily tailored to existing technologies and standard architectures (Mifsud and Constandinou 2023). However, as NVM technologies emerge, there is a pressing need for updated EDA tools and PDKs to explore their potential applications more effectively. These tools should accommodate NVM's unique signal behaviors and memory access characteristics, which differ from those of traditional memories. Additionally, it is crucial to develop EDA software that can simulate both the performance and the total cost of ownership (TCO) for systems incorporating NVMs. This will allow designers to assess the feasibility and benefits of integrating NVM into new circuit architectures.

Cost

While some commercial products use NVMs, the market is small due to higher cost and lower density. DRAM and NAND are supported by multiple fabs across the globe, which keeps costs low simply due to economies of scale. NVM scaling is required to keep up with advanced CMOS devices, and this cannot occur without increased utilization to lower costs. One possible solution would be to create an advanced fab that allows for production of 3D-integrated NVMs, where speed and energy efficiency are improved over conventional devices. Another would be to identify strong uses cases to increase production line adoption of NVMs and reduce cost.

Action plan for non-volatile memory

Table 38. Action Plan for Non-Volatile Memory.

Scope			
Technology for Energy Efficiency	• Non-volatile memory		
Technologies of Interest:	• All NVM media types and memory controllers (e.g., CPU) • All NVM media cell and control logic types (ReRAM, NRAM, STTRAM, Spintronics, etc.)		
Challenges Addressed		**Solution Pathways**	
• Enable processor support. • Application space for new NVM. • Interface width becomes arbitrary. Operating frequency can be lower. Flexible error correction schemes. Temperature sensitivity lower compared to DRAM or eFlash. Smaller footprint compared to SRAM. • Integration of NVM processes and materials into fabs. • Data density issue, which depends on specific application; NVM only impacts certain markets but can be used to supplement DRAM/NAND. • ReRAM may be sensitive to temperature for memory window, read-write noise.		• Negotiate NVM-friendly protocols with controller suppliers and establish standards. • Develop processor support to enable NVM for multiple applications. • Design libraries available for design integration (software design to write protocol). • Provide funding for fabs to integrate new processes. • Design NVM circuitry to handle specific variabilities such as flexible error correction schemes. • Target Applications for NVM. Focusing on specific use cases can accelerate the adoption of NVMs. MRAM and STTRAM are positioned to replace parts of DRAM in environments requiring durability and in AI systems to enhance power efficiency, particularly beneficial for edge devices where space and power are constrained. NRAM, offering competitive read speeds and greater density, is a viable replacement for SRAM in AI applications. Strategic circuit and system co-design is essential to address the inherent challenges of these devices.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline**
Software (memory tiering)	DAX (Direct Access) is in Windows/Linux, ongoing development, growing, other emerging memory. Processor direct vs. CXL, must be incorporated in memory-tiering software.	Software enablement for data centers and hyperscalers	4 years
Access protocol: communicating characteristics to key players	Education of AMD, ARM, Microsoft, Intel, IBM	DRAM replacements	4 years
Control suppliers: product demonstration of NVM	Fab availability and costs. Demonstration of reliability.	Comparative reliability, failure rate, cost to DRAM. TCO competitive. Thermal management requirements. Device lifespan.	6–7 years

Co-design of circuitry to enable next-gen NVM	Integration of on and off-chip components including chiplets and 3D ICs. System performance is modeled and benchmarked using SPEC, evaluating against competing technologies.	Design package for NVM adheres to budget constraints and standards. It includes chiplet designs for UCIe, on-chip SRAM replacement, and off-chip DRAM/NAND/NOR as supplemental memory, targeting specific applications.	Approximately 6–10 year timeline for monolithic implementations.
FEOL/BEOL integration: fablet	Initial development phase for technologies that may not reach widespread production	Next-gen NVM	5–10 years

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
End Users/OEMs	• Provide software support for new memory types.
Academia	• Develop modeling and benchmarking methodologies. • Provide workforce.
National Laboratories	• Perform radiation hard tests.
Government	• Provide exploratory fabs where new materials can be introduced and tested.

Required Resources	Cross Collaboration Needs of Working Groups
• Fablet for BEOL processing out of "standard" fabs. • Overall migration of filesystem models, support libraries, and programming constructs. • Education on new materials, process integration, and contamination controls.	• Software: New architectures; power requirements will require new algorithms. • Materials and Devices: Continue development at bit level. • Advanced Packaging and Heterogeneous Integration: New architectures require new cooling and interconnect methods. • Power and Control Electronics: New power delivery requirements (off- and on-chip monolithic/stacked methods).

2.2.5 Domain-Specific Architectures

Domain-specific architectures (DSAs) and ASICs implement alternative architectures designed to reduce energy and speed overhead for some workloads. Although DSAs and ASICs only complete a handful of tasks, they are incredibly fast and energy-efficient compared to the conventional CPU architecture for those tasks. For example, some microprocessors today include domain-specific processing sections dedicated to tasks like audio and video coding and decoding.

Specific examples of DSAs with significant energy savings over CPUs are the graphics processing unit (GPU), tensor processing unit (TPU), and field-programmable gate array (FPGA). The GPU has a fundamentally different architecture than the CPU does, with large memory banks for massive parallel computing capabilities at significant reduction of energy over the CPU. Google created the TPU after noticing that speech searches were increasing and threatened to double their data center computational power use (Jouppi et al. 2018). The TPU was implemented as a coprocessor for speech matrix multiplication, boasting 30–80x energy savings over CPUs for speech searches. FPGAs provide a dynamically reconfigurable accelerator architecture, allowing the hardware to adapt and accelerate various functions like search algorithms, signal processing, matrix multiplication, and machine learning based on changing workload demands (Putnam et al. 2016). Additionally, FPGAs can serve as a development platform for designing custom ASICs.

DSA and ASIC designs generally follow the strategy outlined by David Patterson and John Hennessy in their book *Computer Architecture: A Quantitative Approach* (Hennessy and Patterson 2019) to improve processing speed and significantly reduce energy per application:

- Minimize the distance that data is moved through hardware and compiler design.
- Invest the savings of on-chip real estate from simplified domain-specific microarchitecture to add arithmetic or memory units depending on which is needed most.
- Through *a priori* knowledge of the target application, utilize parallelism that is easiest for the programmer.
- Use the smallest data type and size needed for the problem.
  - Utilize domain-specific programming languages that are already in use on other systems to reduce complexity of programming.

Table 39 shows a comparison of some recent domain-specific architectures utilizing understanding of the application to create optimized architectures (TPU, FPGA, Anton). Significant gains can be made using domain-specific architectures in place of the conventional CPU/GPU. This is not an exhaustive list and should be used only as supporting evidence for DSAs/ASICs as an important strategy for large use applications.

Table 39. Performance Comparison of Some Recent[a] Domain-Specific Architectures

Technology Group	Specified Technology	Baseline Energy Performance	Commercial Benchmark Product	Commercial Benchmark Energy Performance	Energy Savings Multiplier (X Factor)	Timeline for Lab Scale Demonstration (TRL 6, years)
Domain Specific Architectures	Google TPUv1	92 TOPS/W	Haswell CPU	2.6 TOPS/W	35	0
Domain Specific Architectures	FPGA (embedded computer vision)	1.2–30.8 mJ/frame	ARM 57 CPU	4.5–227 mJ/frame	3.8–7.6	0
Domain Specific Architectures	Anton 3	~150–100K Watt-hour/microsecond (Wh/µs) Simulation	NVIDIA A100 GPU	3,300–1,400,000 Wh/µs Simulation	14–22	0

[a] Sources: Jouppi et al. 2018; Qasaimeh et al. 2019; Shaw et al. 2021

Challenges and Solution Pathways for Domain-Specific Architectures

Cycle Design Time and Simulation

Designing a new chip takes significant time and resources. Allowing for all companies to utilize design software and simulation may enable application-specific architectures suited to specific needs and spur innovation. Historical methods have involved understanding the use case with an existing architecture, then creating iterations of the hardware through design, chip creation, and testing to make an improved device. Enhancing companies' ability to assess device performance through simulations instead of multiple physical test iterations will significantly reduce device costs and time to market. Additionally, adoption of UCIe could allow for easier mixing and matching of different IP components for specified use cases.

Use Cases and Total Cost of Ownership

Creating a custom chip with a new design-specific architecture will require new masks, interconnect design, different IP, and various other factors that together may cost tens of millions of dollars. For DSAs to be adopted on a wide scale, more use cases must be identified where companies can realize significant impacts on total cost of ownership (TCO) or device power, thus providing justification for these investments. As noted earlier in this section, Google developed the TPU because speech searching trends indicated a future doubling in data center computation (Jouppi et al. 2018). In addition, image recognition in power-limited applications, such as smartphones and autonomous vehicles, led to the development of an FGPA architecture (Shaw et al. 2021; Shankar 2022). These examples illustrate that identifying applications that could benefit from reduced power consumption and faster processing can continue to push DSAs into the mainstream and significantly reduce energy consumption in specific use cases.

Table 40. Action Plan for Domain-Specific Architectures

Scope			
Technology for Energy Efficiency	DSAs and ASICs		
Technologies of Interest:	• Large-scale technologies addressing a specific task (e.g., Google's Tensor Processing Units [TPUs], Meta's Meta Training Inference Accelerator [MTIA]). • Domain-specific, high-performance, large-scale scientific compute tasks such as climate simulation or molecular dynamics. Can be applied to different products (CPUs, GPUs, ASICs, FPGAs, etc.) or applications (ML, communication networks, edge computing [CPUs, GPUs, Neural Processing Units, Bluetooth]).		
Challenges Addressed		**Solution Pathways**	
• Operational efficiency • Run-time performance • Real-time requirements (latency) • IP protection • Cost • Make sure approaches not only meet an application's functional requirements but also ensure solutions are achieved within a feasible timeframe.		• Identify large use case applications where development, production, and debugging of DSA/ASIC is cost effective for superior performance and energy efficiency. • Ensure proper use of fidelity requirements. • Explore potential solutions, including cheaper mask sets or silicon production sleds, reasonably priced IP available for reuse, compiler frameworks that can be leveraged for new technologies, better CAD tools for silicon design and debugging, and increased output of skilled engineers from academic institutions.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline**
Identify use cases or applications requiring DSAs	Large-scale applications or tasks that do not run well on existing hardware	2–3 applications	1 year
Strawman design for killer app(s)	Power improvement with iso-performance, or improved performance for iso-power	10 times or greater	1 year for one app
EDA tools tailored for quicker workflow	Design-build-test cycle time	2 times or greater reduction in time	~2 years
Stakeholders and Potential Roles in Project			
Stakeholder	**Role**		
Product Manufacturers/Suppliers	• Design silicon and software		
End Users/OEMs	• Implement in specific markets • Define markets and applications		

Academia	• Develop fundamental design concepts • Accelerate ecosystem growth • Foster workforce development
National Laboratories	• Train senior student/postdoctoral • Develop metrology and standards
Government	• Incentivize the ecosystem • Conduct Outreach (communication to public, members of Congress, etc.)

Required Resources	Cross Collaboration Needs of Working Groups
• Government incentives, clear return on investment and total accessible market, and industry momentum. • Demonstrator/prototype chips/products, evaluation testbeds, access/information about new technologies, and conferences/workshops to bring end users together. • Access to the latest information about advantages, etc., to help in disseminating to the public and policymakers. • Access to EDA tools, prototype hardware, and foundries (at scale). • Funding programs. • Access to well-trained graduate students and postdocs for program development and management. • Resources for outreach and talent development. • Engineers with skills for ASIC and software development, hardware/software co-design of systems. Consider degree programs or certificates in ASIC design and co-design.	• Algorithms and Software: Collaborate to address DSS. • Metrology and Benchmarking: Benchmark DSAs against other solutions. • Power and Control Electronics: Examine adaptive power management in a heterogeneous compute environment. • APHI: Develop possible interaction with 3D architectures because DSAs may require this technology. • Manufacturing Energy Efficiency and Sustainability: Collaborate to reduce cost and number of layers required, cost of packaging and heterogenous integration, and fabrication time.

2.2.6 Instruction Set Architecture

An instruction set architecture (ISA) serves as the crucial software interface to a computer's hardware, enabling software to command the physical components. It defines the supported codes or instructions that a processor can execute, bridging the gap between hardware and software. Commercially significant ISAs include Intel's and ARM's proprietary sets and the open-source RISC-V, all vital for CPU operations. Higher-level virtual machine ISAs like the Java Virtual Machine (JVM) and NVIDIA's Parallel Thread Execution (PTX) provide a further layer of abstraction, primarily used in GPU computing.

While creating a new ISA for novel hardware is technically possible, the substantial software ecosystem required makes it increasingly impractical. Instead, enhancing existing ISAs for energy efficiency and integrating them into DSAs is more commercially viable. This approach taps into existing development tools, speeds up time to market, and cuts costs, while aligning with the energy efficiency goals by optimizing data handling and computational tasks more efficiently in DSAs.

Challenges and Solution Pathways for Instruction Set Architecture

Power Management

ISAs face critical challenges in managing power efficiently across varying workloads. Traditional ISAs are not always optimized for power conservation, leading to excessive energy consumption during idle or low-activity periods. As systems become more complex and energy efficiency becomes a greater concern, particularly in mobile and embedded devices, the need for effective power management strategies becomes paramount. Additionally, existing ISAs may

lack the flexibility to dynamically adjust power settings based on real-time processing demands, resulting in suboptimal power usage.

To address these challenges, incorporating explicit power management instructions into ISAs can significantly enhance energy efficiency. Such instructions would enable processors to adjust their power usage dynamically, ensuring that energy consumption is aligned with the workload requirements. For example, lower precision numeric formats could be employed to reduce the memory bandwidth requirements for certain applications like neural network processing, thereby conserving energy. Additionally, designing ISAs with built-in power-saving modes, like those seen in ARM processors, can minimize power consumption when devices are not in full use. Integrating these power management capabilities into the ISA design would help in reducing overall energy expenditure while maintaining performance (Keller et al. 2017).

Compute-in-Memory

The integration of CIM technologies within traditional ISA frameworks is a significant challenge. CIM aims to reduce the energy and latency costs associated with data movement by performing computations directly where data is stored. However, adapting software to fully leverage CIM capabilities can be complex due to the need for significant changes in program architecture and memory management. Additionally, traditional ISAs may not support the operations needed for effective CIM, which limits the potential gains from this technology.

To overcome these obstacles, ISAs could be extended to include specialized instructions that support compute-in-memory operations. This approach would involve developing intermediate representations, such as tensor dataflow graphs, to optimize data layout and computation strategies directly within the memory array (Wang et al 2022). Such innovations would not only facilitate the integration of CIM into existing system architectures but also enhance the efficiency of data processing tasks. Moreover, incorporating CIM transparently within the ISA could shield programmers from complex hardware details, making it easier to develop applications that benefit from in-memory computing. Collaborative efforts between hardware designers, software developers, and standards bodies are crucial to standardize and propagate these advancements across the industry.

Table 41. Action Plan for Instruction Set Architectures.

Scope			
Technology for Energy Efficiency	Instruction Set Architecture		
Technologies of Interest:	x86, RISC-V, FPGA, GPU, and other CPU		
Challenges Addressed		**Solution Pathways**	
• Power management integration • Compute-in-memory integration		• Involve ISA developers in the definition of PIM functions. • Use cache control structures to match PIM granularities. • Avoid wasted power via control mechanisms to enable or disable speculative operations based on hit rates. • Lower fidelity requirements to minimum. Int8, if it can be used, is significantly less energy intensive than FP32. • Require improved understanding of cache line minimums. Possible to increase or decrease size dependent upon need.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline**

Memory PIM functions	Memory specification with PIM	Cache memory (HBM) or main memory (DDR, LPDDR)	Currently underway
Improve compiler efficiency	High-level language compilation closer to hand-coded assembly	All operating systems and applications	Ongoing
Speculative hit rate monitors	On-the-fly ability to enable or disable speculative functions based on success hit rates	CPUs, GPUs, FPGAs, etc.	4 years
Cache line efficiency	Memory accesses allow granularity closer to the application requirement	CPUs, GPUs, FPGAs, etc.	4 years
ISAs that comprehend PIM	PIM instructions removed from xPU if redundant with memory PIM	CPUs, GPUs, FPGAs, etc.	4 years

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Improve energy efficiency of ISAs, caches, and memory accesses.
End Users/OEMs	• Consider TCO in performance analyses.
Academia	• Revise compiler, reinterpreter writers to consider power utilization.
Government	• Implement algorithms to consider power, ISA issues.
Other	• Develop standards for memory access granularity, PIM functions.

Required Resources	Cross Collaboration Needs of Working Groups
• Power analysis tools for ISA definition, compiler writing. • Benchmarks that calculate TCO factors such as power and cooling. • Power models defined before command sets and access granularity are defined.	• Algorithms and Software: Revise compilers, interpreters, etc., need additional work to improve energy efficiency of instruction streams and data accesses. • Education and Workforce Development: Reeducate compiler and interpreter writers are needed to consider energy use as part of the optimizer functions.

2.2.7 Electronic Design Automation for Circuits and Architectures

EDA is critical in the field of microelectronics, where circuit designs and layouts are meticulously planned and executed. As energy-efficient microelectronic devices continue to miniaturize and increase in complexity, the challenges in IC design have escalated, involving intricate design rules, evolving circuit sizes, diverse masks, specialized measurement needs, and continuously developing processes. EDA tools are instrumental in managing these complexities by facilitating the design of circuits, devices, and systems that meet performance standards and manufacturing requirements.

A key component of EDA is design-technology co-optimization (DTCO). This process involves a collaborative effort between designers and process engineers to optimize a circuit or system's performance, power efficiency, and area density, while also aiming to reduce process development time and costs (Yuan 2022; Synopsys 2023a). DTCO enables teams to refine process technologies to achieve ambitious targets such as precise linewidths, specific dopant profiles, high-quality films, and robust electrical benchmarks.

The DTCO process begins with the development of the fundamental transistor or circuit component. Following this, a comprehensive set of design rules—usually geometric in nature (Ferguson 2018)—is established along with the requisite process steps. This ensures that the device's performance is optimal and that it can be manufactured with a high yield. These design rules and process steps are encapsulated in what is known as the Process Design Kit (PDK).

The PDK, used alongside EDA tools, allows for the precise creation of energy-efficient device and interconnection layouts essential for modern circuits.

One critical aspect of EDA and PDKs, aside from their ability to support circuit and architecture design, is their ability to simulate the behavior of the individual circuit components as well as the full device. It is important to know that each circuit component behaves as indicated and each device is performing at the expected speed and power. EDA creates a 3D model of the circuitry that can then be simulated to show device performance. This can enable manufacturers to reduce parasitic early in the design process rather than later when the device is near production, leading to significant cost savings (Synopsys 2023b).

While EDA and PDK tools do not result in direct savings on the semiconductor device itself, the ability to leverage them for circuit, device, and packaging components will allow for energy savings in other ways:

- Enabling rapid prototyping and testing through advanced simulation tools ensures devices function correctly and are manufacturable without extensive physical trial and error. This efficiency not only saves resources but also enhances manufacturing energy efficiency and sustainability.
- Utilizing EDA tools to refine digital twin architectures in AI/ML applications can lead to more efficient processing and energy use.
- Employing EDA tools to conduct preliminary energy metric testing helps set standards for computing energy per application, promoting energy-efficient designs.
- Implementing EDA tools helps reduce unwanted parasitic effects in circuit designs, improving overall energy efficiency and device performance.
- Incorporating alternative materials for interconnects in the PDK, such as graphitic carbon, CNTs, or advanced devices that are naturally more energy-efficient, such as TFET or GAA, will save energy.
- Implementing EDA tools facilitates the creation of 3D and other innovative architectures that inherently improve energy efficiency by optimizing space and reducing interconnect lengths.
- Reducing the effort needed for design and verification tasks through automated and more intelligent EDA tools will lead to faster development cycles and lower energy consumption during testing.

Challenges and solution pathways for Electronic Design Automation for Circuits and Architectures

Architecture Design for the Most Energy-Efficient Layout

EDA tools play a crucial role in identifying and managing the parasitic components of resistance and capacitance (RC) delays in circuit designs, as highlighted by Thiruvengadam and Borges (2022). These tools are instrumental in optimizing circuit structures to enhance signal transmission and overall performance. However, EDA does not necessarily provide information about either energy per bit performance or standby power (Bhavesh et al. 2022), nor is it necessarily designed primarily for energy efficiency. Power optimization is needed for all aspects of the design flow to reduce overhead (Reis 2015). One solution is to have specific

power constraints of the device added into the system with energy per bit and energy per application integrated into the EDA/PDK software. Utilizing ML (Bhavesh et al. 2022; García-Martín et al. 2019) or AI tools that have been already developed for EDA (Hilson 2023) for power performance standards such as energy per bit for read/write of memory, along with energy per application, could improve the overall energy efficiency of the architecture with little or no compromise in performance.

Verification Bottleneck

Device verification represents a significant challenge for manufacturers, with any major issues in the later stages of product development potentially adding substantial costs and extending the time-to-market (Synopsys 2023b). The conventional approach has been to perform verification late in the IC development process. Moving verification earlier in the process (i.e., with simulation-testing) can uncover performance and function issues that otherwise would not arise until later stages (Aboagye, Patel, and Vig 2014; Synopsys 2023c). EDA providers such as Synopsys® and Cadence® already have this capability. Making early verification more widespread will help prevent unexpected costs and reduce the environmental harm from wasted resources.

Process Design Kit With Sufficient Information for Electronic Design Automation

For the EDA software to construct a device—e.g., an architecture such as DRAM or a processor unit such as an ALU—it must rely on the specifications provided by the PDK. Each PDK has design rules, constraints, schematics, circuit models, and more (Worthman 2014). For optimal design and simulation, the PDK must provide the expected modeling data to the designer for their analysis with EDA. Developing a set standard for what information a PDK must provide could enable better simulation and energy consumption analysis. Such a standard would also benefit designers trying to model newer energy-efficient devices that may not have the same market share or popularity as the incumbent technologies.

Circuitry Parasitics

Circuitry parasitics, primarily resulting from interconnects and components within a circuit, are responsible for a significant portion of energy consumption in microelectronics, often accounting for more than 80% of the total energy use. Parasitic capacitance and resistance in these elements can lead to energy losses, especially during the transmission of signals. To combat these inefficiencies, it is crucial to develop PDKs that incorporate novel materials and innovative designs. These might include CNTs for interconnects, energy-efficient devices like MRAM or TFETs, optimized SRAM architectures, or 3D ICs. Such advancements can significantly reduce parasitic losses and, consequently, the overall energy consumption in microelectronic devices.

Open-Source Electronic Design Automation and Process Design Kits

The high cost to purchase and use EDA and PDKs impedes academic research groups and small companies from developing new and next-generation energy efficient technologies (Chen et al. 2021). For these stakeholders to provide innovative and commercially relevant designs, open-source PDKs such as SKY130 (Chen et al. 2021) and open-source EDA platforms such as DoD's OpenROAD (Moore 2018) are needed to lower design costs.

Table 42. Action Plan for Electronic Design Automation Improvements.

Scope	
Technology for Energy Efficiency	Electronic Design Automation for chip development
Technologies of Interest:	• Design and simulation software (i.e., EDA and PDKs) • Reduction of circuitry parasitics

Challenges Addressed	Solution Pathways
• Industry focuses solely on reducing design time when developing EDA and PDKs. Tools are needed to infer the most energy-efficient solutions more effectively. Current approaches usually require a human expert in the loop. • Better algorithms, performance, and energy efficiency are all needed. • EDA is becoming very expensive, though DARPA is working on an open-source approach to reduce cost. • Co-design is needed for PDK and EDA tools to improve power modeling. (EDA is not currently aware of PDK.)	• Work with vendors to develop application-specific EDA to improve efficiency. • Implement simulation will require higher-level computing languages, which can help solve the current verification bottleneck. • Accelerate EDA tools using AI. • Incentivize (using government incentives) private EDA vendors to cooperate with researchers on PDK and even EDA more broadly. • Move from rectilinear to more open (e.g., curvilinear) device shape to help energy bottlenecks and be thought of as 0.5D expansion. • Investigate and develop PDKs for novel less parasitic devices such as a TFET with lower leakage, reducing bitcell variability, etc. • Implement interoperability between tools to leverage the specialties of different vendors without having to use custom scripts or error-correction mechanisms.

Major Tasks/Milestones	Metrics	Targets	Timeline
Ease of use for EDA/PDK for better circuit design (translating from user language to higher-level language)	Lines of code (C vs. Python)	Compilers to translate	Some compilers are already available, but quality needs improvement.
Demonstrator of Design Technology Co-Optimization (DTCO) flow to optimize Sandia TFET device for memory design to improve energy efficiency	Sub-threshold swing, Vmin	20 mV/dec (1/3 of MOSFET), 3x reduction	12 months
EDA simulation (circuitry parasitics reduction): Demonstrator of EDA tools for simulation and performance analysis of memory design energy efficiency	Energy-delay product	5x–10x reduction	1–2 years
EDA and PDK co-development	Power simulation, photonics, advanced packaging	EDA tools to be aware of PDK for utilization	1–2 years

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	EDA tool vendor (Synopsys, Cadence, Siemens, etc.)
End Users/OEMs	Industry members (fabless and IDMs) and government labs designing their own chips
National Laboratories	Collaborative interaction with Sandia to model/simulate TFETs

Required Resources	Cross Collaboration Needs of Working Groups

• EDA tools for 3D integration, semi-custom SRAM array design, and PDK of CMOS platform. • Inputs on specs of the components for which new circuits/architectures are implemented. Review and assessment based on industry's/lab's own designs. • Device descriptions and measurements; equivalent of BSIM (Berkeley Short-channel IGFET Model) models for circuit simulation.	• Materials and Devices: Complete characterization of transistors needs to be in the PDK tool. • Algorithms and Software: Develop lessons regarding high cost and long time needed for experts. EDA tools need to be developed by analogy to processing units: general purpose, to broadest application (e.g., graphics), to application specific. • EWD: Help with curriculum development for developing application-specific experts that emphasize energy efficiency, further parasitics reduction.

2.2.8 Conclusion for Circuits and Architectures

The Circuits and Architectures chapter highlights the vital importance of prioritizing the development of new circuitry designs to bolster energy efficiency. Domain-specific architectures, in-memory computing, and neuromorphic technologies have emerged as promising solutions to lower energy consumption, particularly in computationally intensive workloads.

Meeting these targets requires robust collaboration across different fields to design new circuits and architectures capable of reducing memory access costs and improving power delivery. This collaboration can be enabled by enhancing EDA tools to streamline device integration and performance simulations, strengthening instruction set architectures to support memory pooling, and developing standards to incorporate chiplet-based designs.

To rapidly translate these innovations into real-world impact, EES2 has set TRL 6 as a baseline to accelerate the deployment of new designs and architectures. Achieving this baseline demands significant investment in co-design strategies, advanced EDA software, and standards to ensure seamless integration across the computing stack. Dedicated cross-collaboration among various stakeholders will be essential to ensure these efforts deliver transformative gains in energy efficiency.

2.2.9 Circuits and Architectures References

Aboagye, A., M. Patel, and N. Vig. 2014. “Standing up to the semiconductor verification challenge.” McKinsey. https://www.mckinsey.com/~/media/McKinsey/dotcom/client_service/Semiconductors/Issue%204%20Autumn%202014/PDFs/MoSC2014_Standing_up_to_the_semiconductor_verification_challenge.ashx.

Agrawal, A., A. Jaiswal, C. Lee, and K. Roy. 2018. “X-SRAM: Enabling In-Memory Boolean Computations in CMOS Static Random Access Memories.” *IEEE Transactions on Circuits and Systems I: Regular Papers*. Vol. 65 (Issue 12): pg 4219–4232. https://doi.org/10.1109/TCSI.2018.2848999.

Amirsoleimani, Amirali, et al. 2020. “In-Memory Vector-Matrix Multiplication in Monolithic Complementary Metal–Oxide–Semiconductor-Memristor Integrated Circuits: Design Choices, Challenges, and Perspectives.” *Adv. Intell. Syst.* Vol. 2: 2000115. http://dx.doi.org/10.1002/aisy.202000115.

An, Ran, et al. 2022. “A Hybrid Computing-In-Memory Architecture by Monolithic 3D Integration of BEOL CNT/IGZO-based CFET Logic and Analog RRAM.” Presented at the 2022 International

Electron Devices Meeting (IEDM). San Francisco. https://doi.org/10.1109/IEDM45625.2022.10019473.

Bhavesh, Modi D., et al. 2022. "Power Consumption Prediction of Digital Circuits using Machine Learning." Presented at the 2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP). Vijayawada, India. https://doi.org/10.1109/AISP53593.2022.9760542.

Bhavnagarwala, Azeez. 2021. "Circuits and Methods to harvest energy from transient on-chip data." United States Patent No. 20220321123. https://patents.justia.com/patent/20220321123.

Bhavnagarwala, Azeez. 2023. "Fast, Energy Efficient Cmos 2piriw Register File Array Using Harvested Data." United States Patent Application No. US 2023/0120936 A1. https://patentimages.storage.googleapis.com/5b/cf/8c/1f7d1885fdee5f/US20230120936A1.pdf.

Biswas, A., and A.P. Chandrakasan. 2019. "CONV-SRAM: An Energy-Efficient SRAM With In-Memory Dot-Product Computation for Low-Power Convolutional Neural Networks." *IEEE Journal of Solid-State Circuits*. Vol. 54 (Issue 1): pg 217–230. https://doi.org/10.1109/JSSC.2018.2880918.

Bowman, Kurtis. 2023. "Explaining CXL Memory Pooling and Sharing." CXL Consortium. Published August 2, 2023. https://computeexpresslink.org/blog/explaining-cxl-memory-pooling-and-sharing-1049/.

Chakraborty, Supriya, Abhishek Gupta, and Manan Suri. 2020. "Unified Characterization Platform for Emerging NVM Technology: Neural Network Application Benchmarking using Off-the-Shelf NVM Chips." Presented at the 2020 IEEE International Symposium on Circuits and Systems (ISCAS). Seville, Spain. https://doi.org/10.1109/ISCAS45731.2020.9180590.

Chang, Liang, et al. 2021. "Trend of Emerging Non-Volatile Memory for AI Processor." Presented at the 2021 18th International SoC Design Conference (ISOCC). Jeju Island, South Korea. https://doi.org/10.1109/ISOCC53507.2021.9613905.

Chang, Meng-Fan (Marvin). 2022. "Advance in Embedded Compute-in-Memory Macros." Presented at the 2022 IEEE VLSI Symposium on Technology and Circuits. Honolulu, HI.

Chatterjee, Niladrish, et al. 2017. "Architecting an Energy-Efficient DRAM System for GPUs." Presented at the 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA). Austin, TX. https://doi.org/10.1109/HPCA.2017.58.

Chen, Kaiquan, et al. 2021. "FreePDK15TFET: An Open-Source Process Design Kit for 15nm CMOS and TFET devices." Presented at the 2021 IEEE International Symposium on Circuits and Systems (ISCAS). Daegu, South Korea. https://doi.org/10.1109/ISCAS51556.2021.9401190.

Christensen, Dennis V., et al. 2022. "2022 Roadmap on neuromorphic computing and engineering." *Neuromorph. Comput. Eng.* Vol. 2: 022501. http://dx.doi.org/10.1088/2634-4386/ac4a83.

Chu, Jennifer. 2020. "Engineers put tens of thousands of artificial brain synapses on a single chip." MIT News Office. Published June 8, 2020. https://news.mit.edu/2020/thousands-artificial-brain-synapses-single-chip-0608.

Danial, Loai, Kanishka Sharma, and Shahar Kvatinsky. 2020. “A Pipelined Memristive Neural Network Analog-to-Digital Converter.” Presented at the 2020 IEEE International Symposium on Circuits and Systems (ISCAS). Seville, Spain. https://doi.org/10.1109/ISCAS45731.2020.9181108.

Demasius, Kai-Uwe, Aron Kirschen, and Stuart Parkin. 2021. “Energy-efficient memcapacitor devices for neuromorphic computing.” *Nat Electron.* Vol. 4: pg 748–756. https://www.nature.com/articles/s41928-021-00649-y.

Energy Star. "Home Page." ENERGY STAR. Accessed May 9, 2024. https://www.energystar.gov.

Ferguson, John. 2018. “The process design kit: protecting design know-how.” Siemens. https://static.sw.cdn.siemens.com/siemens-disw-assets/public/82831/en-US/Siemens-SW-The-process-design-kit-WP-81843-C2.pdf.

Gao, Mingyu, Jing Pu, Xuan Yang, Mark Horowitz, and Christos Kozyrakis. 2017. “TETRIS: Scalable and Efficient Neural Network Acceleration with 3D Memory.” *ASPLOS ’17: Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems*. Xi’an, China. http://dx.doi.org/10.1145/3037697.3037702.

García-Martín, Eva, et al. 2019. “Estimation of energy consumption in machine learning.” *Journal of Parallel and Distributed Computing*. Vol. 134: pg 75–88. https://doi.org/10.1016/j.jpdc.2019.07.007.

Gervasi, Bill. 2023. “Introduction to DRAM Technology.” Presented at the 2023 Flash Memory Summit. Santa Clara, CA.

Gervasi, Bill, and San Chang. 2023. “CXL Native Memories.” Presented at the Storage Developer Conference (2023). Freemont, CA.

Gopireddy, Bhargava, and Josep Torrellas. 2019. “Designing vertical processors in monolithic 3D.” *Proceedings of the 46th International Symposium on Computer Architecture (ISCA '19)*, pg 643–656. New York: Association for Computing Machinery. https://doi.org/10.1145/3307650.3322233.

Grollier, Julie, Damien Querlioz, Kerem Y. Camsari, et al. 2020. “Neuromorphic spintronics.” *Nat Electron*. Vol. 3: pg 360–370. https://doi.org/10.1038/s41928-019-0360-9.

Hankin, Alexander, et al. 2019. “Evaluation of Non-Volatile Memory Based Last Level Cache Given Modern Use Case Behavior.” Presented at the 2019 IEEE International Symposium on Workload Characterization (IISWC). Orlando, FL. https://doi.org/10.1109/IISWC47752.2019.9042051.

Hennessy, John L., and David A. Patterson. 2019. *Computer Architecture: A Quantitative Approach* (6th Edition). Burlington, MA: Morgan Kaufmann Publishers.

Heyman, Karen. 2023. “The Uncertain Future of In-Memory Compute.” Semiconductor Engineering. Accessed December 13, 2023. https://semiengineering.com/the-uncertain-future-of-in-memory-compute/.

Hilson, Gary. 2023. “AI Can’t Design Chips Without People.” EETimes. Published July 3, 2023. https://www.eetimes.com/ai-cant-design-chips-without-people/.

Hosseini, Maryam S., et al. 2022. “Near Volatile and Non-Volatile Memory Processing in 3D Systems.” *IEEE Transactions on Emerging Topics in Computing.* Vol. 10 (Issue 3): pg 1657–1664. https://doi.org/10.1109/TETC.2021.3115495.

Jhang, Chuan-Jia, Xue Chen-Xin, Je-Min Hung, Fu-Chun Chang, and Meng-Fan Chang. 2021. “Challenges and Trends of SRAM-Based Computing-In-Memory for AI Edge Devices.” *IEEE Transactions on Circuits and Systems I: Regular Papers.* Vol. 68 (Issue 5): pg 1773–1786. https://doi.org/10.1109/TCSI.2021.3064189.

Jouppi, Norman P., Cliff Young, Nishant Patil, and David Patterson. 2018. “A domain-specific architecture for deep neural networks.” *Communications of the ACM*. Vol. 61 (Issue 9): pg 50–59. https://doi.org/10.1145/3154484.

Jouppi, Norman P., et al. 2021. “Ten Lessons from Three Generations Shaped Google’s TPUv4i : Industrial Product.” Presented at the 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA). Valencia, Spain. https://doi.org/10.1109/ISCA52012.2021.00010.

Keller, Ben, et al. 2017. “A RISC-V Processor SoC With Integrated Power Management at Submicrosecond Timescales in 28 nm FD-SOI.” *IEEE Journal of Solid-State Circuits*. Vol. 52 (Issue 7): pg 1863–1875. https://doi.org/10.1109/JSSC.2017.2690859.

Kim, Youngbae, Shuai Li, Nandakishor Yadav, and Kyuwon Ken Choi. 2021. “A Novel Ultra-Low Power 8T SRAM-Based Compute-in-Memory Design for Binary Neural Networks.” *Electronics*. Vol. 10 (Issue 17): 2181. http://dx.doi.org/10.3390/electronics10172181.

Krishnan, Gokul, et al. 2022. “Exploring Model Stability of Deep Neural Networks for Reliable RRAM-Based In-Memory Acceleration.” *IEEE Transactions on Computers*. Vol. 71 (Issue 11): pg 2740–2752. https://doi.org/10.1109/TC.2022.3174585.

Kumar, Suhas, et al. 2022. “Dynamical memristors for higher-complexity neuromorphic computing.” *Nat Rev Mater.* Vol. 7: pg 575–591. http://dx.doi.org/10.1038/s41578-022-00434-z.

Lin, Zhiting, et al. 2022. “A review on SRAM-based computing in-memory: Circuits, functions, and applications.” *Journal of Semiconductors*. Vol. 43 (Issue 3): 031401. http://dx.doi.org/10.1088/1674-4926/43/3/031401.

Marinella, Matthew J. 2021. “Radiation Effects in Advanced and Emerging Nonvolatile Memories.” *IEEE Transactions on Nuclear Science*. Vol. 68 (Issue 5): pg 546–572. https://doi.org/10.1109/TNS.2021.3074139.

Masanet, E., A. Shehabi, N. Lei, S. Smith, and J. Koomey. 2020. “Recalibrating global data center energy-use estimates.” *Science*. Vol. 367 (Issue 6481): pg 984–986. http://dx.doi.org/10.1126/science.aba3758.

Mehonic, A., and A.J. Kenyon. 2022. “Brain-inspired computing needs a master plan.” *Nature*. Vol. 604: pg 255–260. https://doi.org/10.1038/s41586-021-04362-w.

Merolla, Paul A., et al. 2014. “A million spiking-neuron integrated circuit with a scalable communication network and interface.” *Science*. Vol. 345 (Issue 6197): pg 668–673. http://dx.doi.org/10.1126/science.1254642.

Mifsud, Andrea, and Timothy G. Constandinou. 2023. "Towards a CMOS-Process-Portable ReRAM PDK." Presented at the 2023 21st IEEE Interregional NEWCAS Conference (NEWCAS). Edinburgh, United Kingdom. https://doi.org/10.1109/NEWCAS57931.2023.10198171.

Moore, Samuel K. 2018. "DARPA Plans a Major Remake of U.S. Electronics: The defense department's research wing is pouring $1.5 billion into projects that could radically alter how electronics are made." IEEE Spectrum. Published July 16, 2018. https://spectrum.ieee.org/darpas-planning-a-major-remake-of-us-electronics-pay-attention.

Mukherjee, Avilash, et al. 2021. "A Case for Emerging Memories in DNN Accelerators." Presented at the 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE). Grenoble, France. http://dx.doi.org/10.23919/DATE51398.2021.9474252.

Nantero. 2023. "Explore our Technology Ecosystem." Accessed December 2023. https://www.nantero.com/technology/.

Pan, Chenyun, and Azad Naeemi. 2017. "Nonvolatile Spintronic Memory Array Performance Benchmarking Based on Three-Terminal Memory Cell." *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*. Vol. 3: pg 10–17. https://doi.org/10.1109/JXCDC.2017.2669213.

Pawlowski, Steve. 2023. "The resurgence of shared memory systems." Presented at the EES2 Working Group Meeting, March 2023. https://ees2.slac.stanford.edu/sites/default/files/2023-09/EES2%20Conversation%202023%203%2015-Pawlowski%20Micron.pdf.

Pellauer, Michael, Yakun Sophia Shao, Jason Clemons, et al. 2019. "Buffets: An Efficient and Composable Storage Idiom for Explicit Decoupled Data Orchestration." *ASPLOS '19: Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*. Pg 137–151. https://doi.org/10.1145/3297858.3304025.

Pires, Francisco. 2023. "3D DRAM Proposal Paves the Road for a Density Increase." Tom's Hardware. Published August 29, 2023. https://www.tomshardware.com/news/3d-dram-proposal-paves-the-road-for-a-density-increase.

Putnam, Andrew, et al. 2016. "A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services." *Communications of the ACM.* Vol. 59 (Issue 11): pg 114–122. https://doi.org/10.1145/2996868.

Qasaimeh, Murad, et al. 2019. "Comparing Energy Efficiency of CPU, GPU and FPGA Implementations for Vision Kernels." Presented at the 2019 IEEE International Conference on Embedded Software and Systems (ICESS). Las Vegas, NV. https://doi.org/10.1109/ICESS.2019.8782524.

Reis, Ricardo. 2015. "Trends on EDA for low power." Presented at the 2015 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO). Ottawa, ON, Canada. http://dx.doi.org/10.1109/NEMO.2015.7415104.

Saito, Hitoshi, et al. 2021. "Development of 16 Mb NRAM Aiming for High Reliability, Small Cell Area, and High Switching Speed." Presented at the 2021 IEEE International Memory Workshop (IMW). Dresden, Germany. https://doi.org/10.1109/IMW51353.2021.9439617.

Sebastian, Abu, et al. 2017. “Temporal correlation detection using computational phase-change memory.” *Nat Commun*. Vol. 8: 1115. https://www.nature.com/articles/s41467-017-01481-9.

Shafiee, Ali, et al. 2016. “ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars.” Presented at the 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA). Seoul, South Korea. https://doi.org/10.1109/ISCA.2016.12.

Shankar, Sadas. 2022. “Energy Efficiency in Computing.” Presented at the EES2 Working Group Meeting, September 2022. https://vimeo.com/769660021/518fb74820.

Shankar, S., and A. Reuther. 2022. “Trends in Energy Estimates for Computing in AI/Machine Learning Accelerators, Supercomputers, and Compute-Intensive Applications.” Presented at the 2022 IEEE High Performance Extreme Computing Conference (HPEC). Waltham, MA. https://doi.org/10.1109/HPEC55821.2022.9926296.

Sharma, Debrenda Das. 2022. “Universal Chiplet Interconnect Express (UCIe): Building an open chiplet ecosystem.” White paper. Universal Chiplet Interconnect Express. https://www.uciexpress.org/_files/ugd/0c1418_c5970a68ab214ffc97fab16d11581449.pdf.

Sharma, Debendra, Gerald Pasdast, Zhiquo Qian, and Kemal Aygun. 2022. “Universal Chiplet Interconnect Express (UCIe): An Open Industry Standard for Innovations With Chiplets at Package Level.” *IEEE Transactions on Components, Packaging and Manufacturing Technology*. Vol. 12 (Issue 9): pg 1423–1431. https://doi.org/10.1109/TCPMT.2022.3207195.

Shaw, David E., et al., 2021. “Anton 3: Twenty Microseconds of Molecular Dynamics Simulation Before Lunch.” *SC ‘21: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. Article no. 1: pg 1–11. https://doi.org/10.1145/3458817.3487397.

Sivan, Maheswari, et al. 2019. “All WSe_2 1T1R resistive RAM cell for future monolithic 3D embedded memory integration.” *Nat Commun*. Vol. 10 (Article no. 5201). https://www.nature.com/articles/s41467-019-13176-4.

Strenz, Robert. 2020. “Review and Outlook on Embedded NVM Technologies – From Evolution to Revolution.” Presented at the 2020 IEEE International Memory Workshop (IMW). Dresden, Germany. https://doi.org/10.1109/IMW48823.2020.9108121.

Synopsys. 2023a. “Synopsys TCAD: Unleash the Power of Smart Technology Modeling – from Atoms to Circuits.” Accessed December 2023. https://www.synopsys.com/manufacturing/tcad.html.

Synopsys. 2023b. “What is EDA (Electronic Design Automation)?” Accessed December 2023. https://www.synopsys.com/glossary/what-is-electronic-design-automation.html.

Synopsys. 2023c. “Scalable SOC Verification: Early Software Bring-Up & System Validation.” Accessed December 2023. https://www.synopsys.com/verification.html.

Sze, Vivienne, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. 2017. “Efficient Processing of Deep Neural Networks: A Tutorial and Survey.” *Proceedings of the IEEE*. Vol. 105 (Issue 12): pg 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740.

Thiruvengadam, Anand, and Ricardo Borges. 2022. "Why DTCO is Critical to Modern Memory Design Techniques." Synopsys. Published August 24, 2022. https://www.synopsys.com/blogs/chip-design/why-dtco-is-key-to-memory-design-techniques.html.

Veksler, Dmitry, et al. 2020. "Memory update characteristics of carbon nanotube memristors (NRAM®) under circuitry-relevant operation conditions." Presented at the 2020 IEEE International Reliability Physics Symposium (IRPS). Dallas, TX. https://doi.org/10.1109/IRPS45951.2020.9128335.

Verma, Naveen, et al. 2019. "In-Memory Computing: Advances and Prospects." *IEEE Solid-State Circuits Magazine*. Vol. 11 (Issue 3): pg 43–55. https://doi.org/10.1109/MSSC.2019.2922889.

Vogelsang, Thomas. 2010. "Understanding the Energy Consumption of Dynamic Random Access Memories." *Proceedings of the 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO '43)*. Atlanta, GA. Pg 363–374. http://dx.doi.org/10.1109/MICRO.2010.42.

Vogelsang, Thomas, Brent Haukness, Eric Linstadt, Torsten Partsch, and James Tringali. 2022. "DRAM Refresh with Master Wordline Granularity Control of Refresh Intervals." *Proceedings of the International Symposium on Memory Systems (MEMSYS '21)*. New York: Association for Computing Machinery. Article no. 3: pg 1–6. https://doi.org/10.1145/3488423.3519321.

Wan, Weier, et al. 2022. "A compute-in-memory chip based on resistive random-access memory." *Nature*. Vol. 608: pg 504–512. http://dx.doi.org/10.1038/s41586-022-04992-8.

Wan, Zhe, et al. 2022. "Accuracy and Resiliency of Analog Compute-in-Memory Inference Engines." *J. Emerg. Technol. Comput. Syst*. Vol. 18 (Issue 2, Article no. 33): pg 1–23. https://doi.org/10.1145/3502721.

Wang, Z., C. Liu, and T. Nowatzki. 2022. "Infinity Stream: Enabling Transparent and Automated In-Memory Computing." *IEEE Computer Architecture Letters*. Vol. 21 (Issue 2): pg 85–88. https://doi.org/10.1109/LCA.2022.3203064.

Woo, Steven. 2021. "CXL Ushers in a New Era of Data-Center Architecture." Electronic Design. Published October 13, 2021. https://www.electronicdesign.com/technologies/embedded/article/21176870/rambus-cxl-ushers-in-a-new-era-of-datacenter-architecture.

Worthman, Ernest. 2014. "A Guide To Advanced Process Design Kits." Semiconductor Engineering. Published April 14, 2014. https://semiengineering.com/a-guide-to-advanced-process-design-kits/.

Xiao, Rui, Wenju Jiang, and Piew Yoong Chee. 2022. "An Energy Efficient Time-Multiplexing Computing-in-Memory Architecture for Edge Intelligence." *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*. Vol. 8 (Issue 2): pg 111–118. https://doi.org/10.1109/JXCDC.2022.3206879.

Xiao, T. Patrick, et al. 2022. "On the Accuracy of Analog Neural Network Inference Accelerators." *IEEE Circuits and Systems Magazine*. Vol. 22 (Issue 4): pg 26–48. https://doi.org/10.1109/MCAS.2022.3214409.

Yamazaki, Kashu, Viet-Khoa Vo-Ho, Darshan Bulsara, and Ngan Le. 2022. “Spiking Neural Networks and Their Applications: A Review.” *Brain Sci*. Vol. 12 (Issue 7): 863. http://dx.doi.org/10.3390/brainsci12070863.

Yao, Peng, et al. 2020. “Fully hardware-implemented memristor convolutional neural network.” *Nature*. Vol. 577: pg 641–646. https://www.nature.com/articles/s41586-020-1942-4.

Yu, Shimeng. 2016. “Resistive random access memory (RRAM): from devices to array architectures.” San Rafael, CA: Morgan & Claypool. https://searchworks.stanford.edu/view/11650805.

Yuan, Lipen. 2022. “What is DTCO?: An Introduction to Design-Technology Co-Optimization.” TSMC. Published June 15, 2022. https://www.tsmc.com/english/news-events/blog-article-20220615.

Zhang, Sai, Kejie Huang, and Haibin Shen. 2020. “A Robust 8-Bit Non-Volatile Computing-in-Memory Core for Low-Power Parallel MAC Operations.” *IEEE Transactions on Circuits and Systems I: Regular Papers*. Vol. 67 (Issue 6): pg 1867–1880. https://doi.org/10.1109/TCSI.2020.2971642.

Zhang, Yang, et al. 2021. “EnTiered-ReRAM: An Enhanced Low Latency and Energy Efficient TLC Crossbar ReRAM Architecture.” *IEEE Access*. Vol. 9: pg 167173–167189. https://doi.org/10.1109/ACCESS.2021.3129878.

Zhang, Anguo, Xiumin Li, Yueming Gao, and Yuzhen Niu. 2022. “Event-Driven Intrinsic Plasticity for Spiking Convolutional Neural Networks.” *IEEE Transactions on Neural Networks and Learning Systems*. Vol. 33 (Issue 5): pg 1986–1995. https://doi.org/10.1109/TNNLS.2021.3084955.

Zimmer, Brian, et al. 2020. “A 0.32–128 TOPS, Scalable Multi-Chip-Module-Based Deep Neural Network Inference Accelerator With Ground-Referenced Signaling in 16 nm.” *IEEE Journal of Solid-State Circuits*. Vol. 55 (Issue 4): pg 920–932. https://doi.org/10.1109/JSSC.2019.2960488.

2.3 Advanced Packaging and Heterogeneous Integration

As transistor nodes shrink below 20 nm, the cost benefits diminish, shifting the focus toward advanced packaging (AP) and heterogeneous integration (HI)—collectively referred to as APHI—as essential methods for enhancing energy efficiency and device performance. The focus of APHI is on a variety of different approaches for packaging chips together. Until recently, integration technologies focused on planar chip interconnects, chip-to-chip connections, and air flow and heat sinks for thermal management. In the shift to APHI, new technologies such as 2.5D and 3D geometries, as well as advanced interconnect schemes and new thermal mitigation strategies between stacked chips, are some of the key energy efficiency approaches. Multiple semiconductor organizations, including MAPT, IRDS, IEEE, and the Semiconductor Industry Association (IEEE IRDS 2023; SIA 2022), have stated that HI will be the key technology driver for at least the next decade due to its performance and energy efficiency improvement potential.

Energy efficiency of logic and memory operations have not improved at the same rate. Memory technologies have improved more rapidly than logic operations in terms of energy efficiency, largely due to significant advancements in memory design and integration techniques. Logic operations, involving arithmetic instructions, have improved by approximately 2x to 4x depending on the calculation type (Jouppi et al. 2021). In contrast, memory technologies like HBM2 and GDDR6 have seen a 6-foild increase in efficiency compared to older DDR3/4 standards (Vogelsang 2010; Smith 2022; O'Conner et al. 2017). However, memory operations still exhibit an energy cost nearly 4x higher than that of the most energy intensive logic operation, primarily due to the energy costs associated with data transfer through interconnects. For instance, accessing DDR3/4 memory is about 1,300 times more energy-intensive per bit than logic operations are. Advances in HBM2 and GDDR6 have reduced this disparity to about 250x to 350x through packaging improvements, yet accessing memory remains significantly more energy-costly than performing logic operations (Jouppi et al. 2021).

The Circuits and Architectures chapter emphasized computational strategies to minimize data transfer, whereas this chapter on APHI will explore next-generation interconnect technologies. These technologies aim to enhance data transfer efficiency and incorporate thermal mitigation strategies to lower energy consumption by reducing chip parasitics and secondary energy expenditures.

An instruction in microelectronics refers to a command given to a computer processor to perform a specific operation. At the instruction level, various technologies discussed in the APHI chapter have the potential to significantly impact energy consumption. Innovations such as carbon nanotube interconnects decrease parasitic losses, while approaches like 2.5/3D interconnects and chip stacking technologies shorten the distances between interconnected components. These advancements collectively aim to reduce the overall energy requirements of memory and logic operations, demonstrating a crucial step toward more energy-efficient microelectronic systems.

Working Group Methodology

APHI technologies are at the forefront of performance improvement in microelectronics today and align well with the EES2 goal of reducing the overall energy consumption of microelectronics. The APHI working group proposed nearly 30 technologies, divided into six groups, that tackle both the energy consumed during operation and the secondary costs of

cooling (see Table 43). The proposed technologies' energy efficiency factors (compared to incumbent technologies) are found in their respective sections throughout the chapter.

Certain technologies discussed in the Circuits and Architectures chapter are discussed in further detail here because they also have implications for APHI. Compute-near-memory, for example, is enabled through interconnect and thermal management technologies. Additionally, EDA, while not a physical device, can have significant impacts on the system package through co-design and with initial energy consumption simulations.

Table 43. APHI Technology Groups and Technologies of Interest

Technology Group	Specified Technology
Next-Gen Interconnects, all levels	Graphene, CNT (SWNT, MWNT)
	Ru, Ir, Rh
	Optical
Foundational 2.5/3D Interconnect Technologies	Carbon based
	2.5D-Bridge Chip, EMIB/Foveros, Interposer, Chiplet
	Through silicon via
	Monolithic 3D (Monolithic Inter-tier Vias)
	Hybrid Bonding (Cu-Cu)
	UCIe
Application-driven 3D Integration	Vcache
	MIV stacked ReRAM
	DRAM Cache
Advanced Thermal Interface Materials (TIM)	LMP Metal Solder with polymer (Indium-based)
	Nanostructure engineering to increase thermal surface area contact
	CNT based thermally conductive matrix
	Graphene based conductive matrix
EDA for Systems Design (SOIC, SiP, PCB)	Energy per bit simulations
	Architecture level PDKs
	STCO
	Thermal Co-Design

Figure 34 shows the technologies of interest with their potential energy efficiency improvement factors and timelines to TRL 6, as determined by the working group. For more information on TRL6, refer to section 1.5.

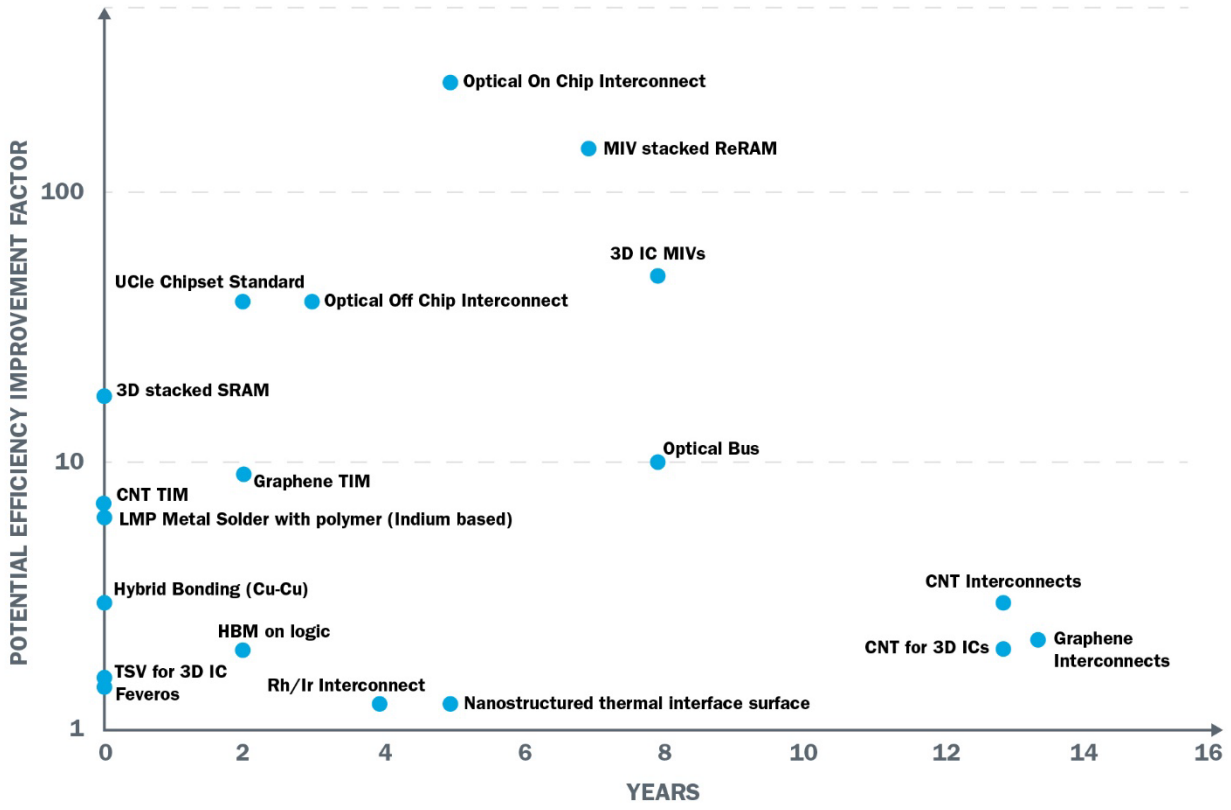
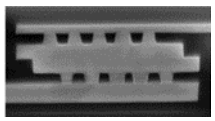
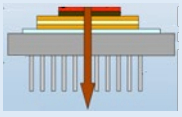


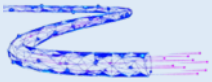
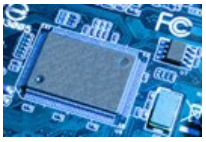
Figure 34. Potential efficiency improvement factor and timeline for selected technologies of the APHI working group

Key Takeaways

Table 44 summarizes the most significant identified energy efficiency opportunities that can be achieved through advances in APHI.

Table 44. Key Takeaways for Energy Efficiency Opportunities in APHI.

Technology Group	Key Opportunities for Energy Efficiency
Interconnects for 2.5/3D stacking	 • Transition to 3D hybrid bonding to surpass the limitations of copper microbumps, enhancing energy efficiency with reduced signal delay and improved bandwidth. • Implement copper-to-copper and dielectric bonding for submicron pitch sizes, resulting in significant energy savings compared to traditional methods. • Utilize advanced packaging techniques for legacy nodes to improve energy efficiency.
Thermal Interface Materials	 • Advance thermal interface materials that offer lower thermal resistance and obviate the need for polymer adhesives, leading to better heat dissipation and energy efficiency.
Advanced system cooling technologies	 • Develop novel cooling technologies capable of managing the heat generated by high-density chip stacks, thus reducing the energy required for cooling operations.

Alternative interconnect materials		• Advance optical interconnects to sidestep parasitic capacitance for increased bandwidth and lower energy per bit through component miniaturization and monolithic integration. • Explore carbon-based interconnects like CNTs to reduce resistive losses and improve thermal conduction over copper, aiming for reduced RC delays and enhanced energy efficiency.
Packaging EDA		• Enhance energy efficiency through EDA that facilitates advanced integration/packaging co-design and initial energy consumption simulations. • Encourage the use of system technology co-optimization (STCO) for predictive modeling and optimization of energy use in packaging designs.

Grand Challenges

Achievement of the identified energy savings opportunities with APHI requires overcoming the following major challenges:

- Creating an R&D fablet open to universities and small businesses for feasibility testing to address challenges associated with APHI technologies and workforce development.
- Implementing UCIe as the standard interconnect to facilitate miniaturization, increase bandwidth, accommodate legacy nodes, and resolve supply challenges.
- Developing and testing novel devices that may not currently be CMOS compatible but can be monolithically integrated.
- Exploring innovative thermal interface materials, like carbon-based matrices and topographically engineered interfaces, to effectively address thermal management challenges.
- Reducing memory access costs and heat generation, primarily for SRAM and DRAM, through innovations in interconnect technologies, including alternative materials and mechanisms.

2.3.1 Carbon Nanotube-Based Interconnects

Interconnect technologies currently use ~80% of all on-chip power (Karkar et al. 2016). While 3D ICs will significantly reduce this high energy consumption, investigations into alternative materials are paramount. Currently, copper is the material of choice for interconnects at all levels, but as devices continue to scale, copper's resistivity, grain boundary effects, and thermal issues will increase. Additionally, copper electromigration is an issue as line widths are decreased (Mittal and Lin 2017). The APHI working group proposes carbon nanotubes (CNTs) as one possible solution for moving beyond copper interconnects.

CNTs were introduced in the Materials and Devices chapter as a possible new transistor technology. As the quality of CNT production improves, along with other carbon allotrope production technologies (such as wet-based chemistry methods, deposition methods, and graphitic sheet transfer technologies), research priorities should include investigation of alternative opportunities for carbon, such as interconnects and interposer technologies, for improvements in energy efficiency and performance.

CNTs and graphene provide significant reduction in resistance and capacitance, reducing RC delays without the need for as many repeater amplifiers; they also provide increased thermal conduction over copper, likely reducing localized hot spots (Alam et al. 2011; Mittal and Lin 2017). One important property of CNTs and graphene is the ability to carry significantly higher current density than copper does at smaller sizes (Soldano, Talapatra, and Kar 2013), opening up CNTs for power distribution, vias, and possibly smaller chips. Table 45 provides a comparison of simulated graphene layers, measuring resistance, capacitance, and correlating impact factors of CNT bundles compared to conventional copper interconnects.

Table 45. Comparison of Simulated Graphene Layers and Resistance, Capacitance, and Correlating Impact Factors of CNT Bundles Compared to Conventional Copper Interconnects

Technology Group	Specified Technology	Baseline Energy Performance	Commercial Benchmark Product	Commercial Benchmark Energy Performance	Impact Factor	Timeline (years)
Carbon-Based Interconnects	Graphene (simulation 20 layers)	Resistance: 600 Ω/μm	Copper	Resistance: 950 Ω/μm	1.6	10–15
		Capacitance: 0.08 femtofarad/ micron (fF/μm)	Copper	Capacitance: 0.17 fF/μm	2.1	
	CNT (simulation)	4 Ω	Copper (simulation)	12 Ω	3.0	
	CNT Local Interconnect (simulation)	Capacitance 100 μm: 12.6 fF 500 μm: 62.8 fF 1,000 μm: 142.3 fF	Copper	Capacitance 100 μm: 14.3 fF 500 μm: 71.4 fF 1,000 μm: 184.4 fF	1.1–1.3	

Although graphene and CNT bundles provide improvement over copper, the implementation of these materials as interconnects is still in its infancy. The working group identified no current device or prototype using CNT interconnects, only simulation or initial rudimentary test structures. For example, graphene interconnects with 7nm technology (Wang et al. 2017) showed an 8% improvement in the energy delay product via EDA simulation, with more room for improvement.

Challenges and Solution Pathways for Carbon Nanotube-Based Interconnects

Contact Resistance

Contact resistance measures the impedance electrons face when transitioning between different media. For carbon-based interconnects, the contact resistance is dominated by metal-carbon distance, adhesion to the metal contact, and the metal work function. Researchers have employed various techniques to reduce this resistance, including using joule heating to eliminate interfacial impurities, forming metallic-carbon interfacial layers, and applying ultrasonic nanowelding. Despite these efforts, the best achievable contact resistivity for CNT to metal remains around 10^{-5} Ω·cm^2. This value is still an order of magnitude higher than the contact

resistance of $5.8*10^{-6}$ Ω·cm^2 for Cu at the 22nm node, highlighting ongoing challenges in achieving comparable efficiency (Todri-Sanial, Dijon, and Maffucci 2017).

Ab initio calculations play a crucial role in tackling these challenges by providing a theoretical foundation to explore and optimize the atomic and electronic structures at the interfaces. These calculations help predict the optimal configurations for reducing contact resistance, focusing on parameters such as the type of metal used, the number and dimensions of multi-walled nanotubes, graphene layer properties, and interface characteristics. The working group suggested avoiding metals altogether in situations where the carbon-based interconnect contact resistance is too high (Wang et al. 2017). CNT-to-graphene contact resistance is poorly referenced in the literature, but a value of 10^{-6} μΩ·cm^2 has been reported (Ramos et al. 2016) and could likely be further optimized. Another pathway could be to create graphitic nanosheets through laser ablation of SiC (Salama et al. 2002). This technique is patented for an interposer technology (Salama 2023) but may be expandable to create vertical and horizontal interconnects with limited contact resistance.

Production of Carbon-Based Interconnects and Process Integration

As detailed in the Materials and Devices chapter's CNTFETs section, the initial step for any new material integration, such as carbon-based interconnects, involves a rigorous industry vetting process to ensure that no contaminants are introduced. Once cleared, integration of these technologies will bring additional challenges. Chemical vapor deposition (CVD) of CNT and graphene generally requires seed layers and may not be BEOL-compatible to produce the needed material properties with current methods. Alternatives like spin coating CNTs for horizontal interconnects or transferring graphene sheets are BEOL-compatible but still require significant process optimization to meet high-volume manufacturing standards and to ensure they are free of contaminants.

If the industry opts for converting silicon carbide (SiC) to graphitic carbon, this would necessitate not only new equipment but also extensive optimization of both the laser systems used and the resulting material properties. Establishing a solid foundation for the integration of carbon interconnects will be crucial. A dedicated fabrication facility, or fablet, that allows for the exploration of new processes and their refinement to ensure CMOS compatibility could greatly accelerate the transition of these technologies to full-scale high-volume manufacturing.

Action Plan for Carbon Nanotube-Based Interconnects

Table 46. Action Plan for Carbon Nanotube-Based Interconnects

Scope		
Technology for Energy Efficiency:	Carbon nanotube-based interconnects	
Technologies of Interest:	• Carbon-based interconnects for chip stacking (carbon through silicon vias, flip-chip pads) • Carbon-based interconnects for PCBs (replacing/complementing Cu, though hole vias are quite large and cannot coat via with Cu). CNT for interconnects into SiC (integrated circuits, substrate fabrications, multiple applications). • Carbon-to-carbon-based vertical and horizontal interconnects; carbon-to-carbon HDI (CNTs, graphene) • CNTFets for compute to memory bus • Flexible electronics	
Challenges Addressed		**Solution Pathways**

• Reduce Ohmic contact at carbon/metal, carbon/carbon interface. • Improve production of global interconnects with correct diameter, length, and chirality for optimal material properties along with improved filtration purification processes. • Research low-temp. deposition processes (under 300°C), seed layer, or conversion technologies for of vertical interconnects. • Integrate processes for CNT interconnects that are CMOS-compatible. • Understand junction contact resistance of carbon based horizontal and vertical interconnects.	• Develop contact promoter between metal interconnect and C (Ag, Ni, Pd, TiN) in addition to *ab initio* calculations to help understand the electron transport between carbon-to-metal and carbon-to-carbon. • Produce vertical CNTs through low-temp. (<400°C) with a cobalt seed. Produce pure metallic SWNTs and MWNTs containing electrical properties like metallic SWNT via wet chemistry processes with filtration/purification processes to remove impurities. • Optimize CMOS-compatible, carbon-based interconnects using spin coating by fine-tuning the solution viscosity and substrate roughness and by enhancing inkjet printing processes. • Investigate SiC to graphitic interposer technology (no EDA, PDK for this technology).

Major Tasks/Milestones	Metrics	Targets	Timeline
CMOS-compatible spin coating for horizontal interconnects	• Ability to demonstrate a metal layer with 90nm technology	• Achieve legacy nodes • Achieve chip-to-chip compatibility	0–1 years
Vertical interconnects	• Vertical via fill for CNTs (inkjet, squeegee); Molarity of solvent layers, interlayer resistance; Comparable to the resistance of copper. • Efficient routing on interlayers; Low interlayer resistance applicable to global interconnects.	• CNT with lower resistance than copper, suitable for lower RC delay at approximately 20nm technology scale • <350°C for BEOL compatibility	0–2 years (ink jet, squeegee) 3–5 years
SiC conversion to graphitic carbon interposer technology	• Advanced packaging. • High power and energy savings applications. • Thermal considerations, high-frequency. performance, and current carrying capacity.	• Used in high-performance computing (HPC), AI interposer technology, and power electronics (especially for thermal management to enable operations at higher temperatures).	2 years (application development)
Leverage knowledge of community to reduce repeat experiments	• On par with the contact resistance of copper. • Knowledge sharing to reduce iterative testing.	• Develop CNT technology with resistance lower than or comparable to copper for both local and chip-to-chip interconnects.	3–4 years (academic demonstration)
Ab initio calculations for interface production	• Develop accurate interface models to determine the electronic and physical structure at the CNT-metal interface. • Assess Fermi level and band structure.	• Enhance electrical structure to optimize performance • Define physical dimensions of SWNTs and MWNTs	3–4 years
Omnidirectional interconnect (*in situ*, *ex situ*)	• On par with resistance of copper	• Multi-die stacking. • Suitable for high-power application	7 years

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Supplying high metallic CNTs, graphitic interposer technology
End Users/OEMs	• Testing and integration
Academia	• Basic/experimental research to enable technologies

National Laboratories	• Basic/experimental research to enable technologies, take to higher TRL
Government	• Funding opportunities • Explore extreme uses (temperature, radiation)

Required Resources	Cross Collaboration Needs of Working Groups
• Develop a fablet capable of integrating CNTs into the BEOL process, serving as a generic institute to lower the barriers to entry. • Enhance the availability and development of simulation tools and libraries, including Density Functional Theory (DFT) and *ab initio* methods. • Amplify focus on thermal characteristics in academic research to improve material performance under varying thermal conditions. • Use high-volume manufacturing laser machines for high throughput of SiC conversion to graphitic carbon. • Address the limited scope of current carbon-based technologies by adjusting architectural, power delivery, and manufacturing processes. Development of dedicated EDA/PDK software tailored to these materials is essential. • Develop curricula at both collegiate and workforce training levels to educate on the unique requirements and applications of carbon-based technologies.	• Materials and Devices: Synthesize metallic SWNT, MWNT. Develop interconnect interface. • Circuits and Architectures: Develop new systems and circuitry for difference in potential and current given ballistic e-transport and lower capacitance and voltages. • Metrology and Benchmarking: Investigate voltages, capacitance, failure mechanisms, interfacial issues, benchmarking new performance, etc. • Algorithms and Software: Update software/algorithms if Circuits and Architectures effort enables different architectures. • Power and Control Electronics: Implement new power paradigm with lower resistance and capacitance; voltage drop of carbon-based interconnects will require power supply changes.

2.3.2 Optical Interconnects

Optical interconnects provide a superior alternative for connectivity between on-chip cores and within multi-chip modules, especially as the performance of electrical links degrades over longer distances like those found in traditional circuit boards. As discussed in the beginning of this chapter, a significant portion of energy in computer systems is consumed by interconnections rather than logic operations, particularly at the board and chip levels. This is due to high signal volume and the associated energy costs of charging and discharging wires with high capacitance. Figure 35 presents benchmark data from around 2018, comparing the bandwidth density and energy efficiency of state-of-the-art electrical and optical interconnects, expressed in terms of bandwidth density multiplied by energy efficiency (specific energy in pJ/bit).

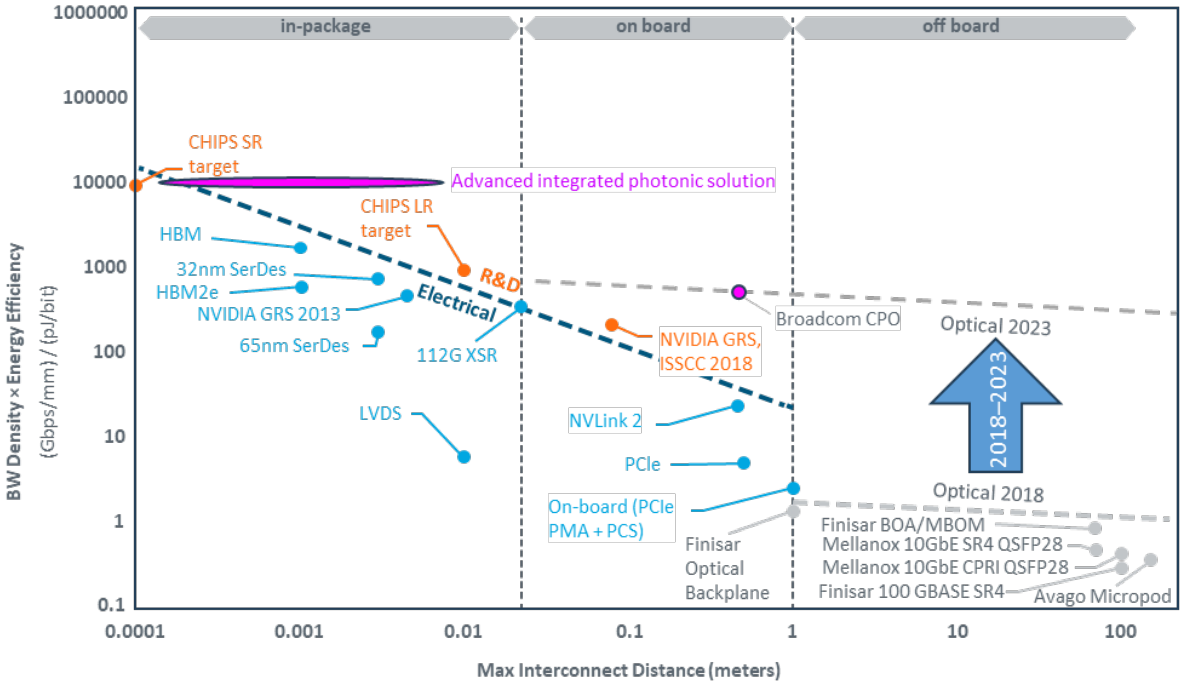


Figure 35. Interconnect figure of merit benchmarks (circa 2018) with 2023 commercial and R&D optical interconnect benchmark references. *Source: Stojanovic 2020; Original source: Gordon Keeler, DARPA MTO, ERI Summit 2019*

Photonic transceivers enable high bandwidth optical signaling but currently exist commercially as board-level pluggable components connected to chips and multi-chip packages via electrical wires whose power dissipation and density limit overall performance. The DARPA Photonics in the Package for Extreme Scalability (PIPES) program (Tauke-Pedretti 2023) has pushed the boundaries of the electrical/photonic interface toward the package and board level through further miniaturization of the photonic components. As a result of work done in the PIPES program and elsewhere, the crossover point between optical and electrical interconnects has moved from about 1 m in 2018 to about 10 cm by 2023 (Sorger 2023), with exact crossover depending on many factors, such as underlying component technology, signal processing, and signal modulation format. A commercial co-packaged optics (CPO) product (Broadcom 2023) and an estimate based on recent R&D for sub-centimeter interconnections have been added to the interconnect benchmark plot (see Figure 35). The latter is based on photonic waveguides with 10 μm pitch, a 10 Gbps on-off keyed data rate (without multiplexing), transceiver-less operations assuming 10 fJ/bit for the electro-optic modulator (EOM), 15% laser wall-plug efficiency, and 30 fJ losses. These assumptions are supported by current R&D on the component devices, as discussed in the following section.

Challenges and Solution Pathways for Optical Interconnects

To further exploit the benefits of optical links, photonics must become more intimately integrated in the microelectronics package. However, the target of 1pJ/bit associated with technologies such as those pursued by PIPES, while achieving impressive progress over the prior state of the art, does not fully address the grand challenge sought after in the EES2 vision of 100–1,000x power reductions for data communication and systems. Integrating optical transceivers

(ideally monolithically) into electronic chips presents an opportunity to move closer to the more ambitious EES2 goal.

An optical link needs components at each end to transmit and receive the signal. The source needs a light source (generally a laser) and an electro-optical modulator, and the receiver needs a photodetector. The potential for drastic reductions in the energy demanded by optical interconnects through device scaling has been discussed in detail by Miller (2017), which suggests a pathway for interconnects from ~ 1 cm to ~ 10 m that have the same energy (~ 10fJ/bit) as local electrical wires on chip or better. Achieving this level of performance will have far-reaching impact on the energy consumption of computer systems, substantially knocking down the “memory wall” both in terms of latency and energy consumption. A more recent review by Mekawey et al. (2022) provides more up-to-date references to the state of the art.

A fundamental quantum mechanical advantage of optical interconnects is known as “quantum impedance conversion,” meaning that the optical signal only needs to charge the capacitance of the photodetector and not the channel itself (Miller 1989). This avoids the major energy use of electrical interconnects but trades that loss for the energy that must be consumed to power the optical transmitter and receiver. To achieve very low-energy optical interconnects, the key challenge is to reduce capacitance of photodetectors, optical sources, and their associated circuitry.

Solutions to improving power consumption of optical links are rooted in (a) component performance improvement, which includes clever designs, emerging materials, and deployment of device optimization algorithms, (b) link-level optimization such as is enabled by multiplexing of signals (Winzer and Neilson 2017), and (c) system synergies enabled by HI of multiple technologies, each optimized for a specific purpose. Here, HI is key since it allows reduction of parasitic capacitances (e.g., between CMOS drivers and optoelectronic components). Regarding the latter, emerging chip manufacturing capabilities are promising, such as Global Foundries 45SPCLO, a 45nm SOI CMOS technology monolithically integrating RF, analog, and silicon photonics capability (GlobalFoundries 2022).

Semiconductor Lasers

Semiconductor lasers act as light sources that are modulated to encode information. For transmission over long distances, most transmitters use externally modulated lasers. However, for short-reach links, lasers can be directly modulated, avoiding the need for a separate modulator, and thereby offering savings in energy consumption and transmitter footprint.

Integrating lasers with photonics has its challenges, though significant advancements have been reported recently (Li et al. 2022). To achieve optical gain on Si, an effective and common solution is to integrate a III–V semiconductor gain medium on a Si photonics platform. III–V lasers have the advantages of high gain, high optical output power, and the ability to operate using electrical pumping. Since III–V materials are not CMOS-compatible today, integration approaches with Si photonics platforms include flip-chip integration, transfer printing, and heterogenous bonding. Direct growth of III-V gain material on silicon substrate may bring the cost down and improve scalability. Quantum dot (QD) lasers integrated on Si through bonding have also been reported recently (Norman et al. 2019; Shang et al. 2022).

Electro-Optic Modulators

Achieving efficient modulation of light in optical interconnects presents significant challenges. An electro-optic modulator (EOM) operates similarly to a transistor, with an optical source and drain, and an electronic gate that modulates the refractive index of the optical medium to modulate light. The most common optical modulators are the lithium-niobate-based Mach-Zehnder modulator (MZM), indium-phosphide-based electro-absorption modulator (EAM), and silicon ring modulator (RM). MZM is currently more widely used (Wooten et al. 2000) than the other types of modulators, especially in long-haul applications, because of better extinction ratio (the ratio between signal energies for the “1” and “0” states), larger modulation bandwidth, and relatively low influence of thermal and polarization variations on the modulator performance. Indium-phosphide EAMs offer advantages in terms of lower drive voltage requirements and smaller form factor (Wu et al. 2017). Silicon RMs are favored for their compact size, low loss, low energy consumption (~6fJ/bits), and compatibility with CMOS technology (Li et al. 2013). However, their low extinction ratio and strong sensitivity to temperature remain obstacles to adoption. Research is underway to address these limitations.

Sorger et al. (2015) laid out an evolutionary path for future EOMs in terms of technological advancements and limitations, as shown in Figure 37. EOMs, crucial for converting electrical data to optical signals, can operate by directly modulating the light source or through other mechanisms. The focus is on reducing the physical size of EOMs to lower capacitance and energy requirements while increasing data transmission rates. Nanoscale EOMs, approximately 1 μm in size, can achieve switching rates over 100 Gbps and switching energy as little as 1 fJ/bit, which reduces the power needed to just 1 μW. This is a 1,000x reduction compared to classical modulators. However, enhancing light-matter interaction to reduce size further without sacrificing performance remains a challenge. Current efforts explore various promising materials and techniques such as free carriers in silicon and indium tin oxide, quantum-confined Stark effects in germanium quantum wells, and permittivity tuning in graphene (Xu et al. 2005; Amin et al. 2018; Srinivasan et al. 2019; Ye et al. 2014). These advancements have led to significant reductions in device size and improvements in modulation efficiency, marking substantial progress toward integrating these technologies into standard manufacturing processes.

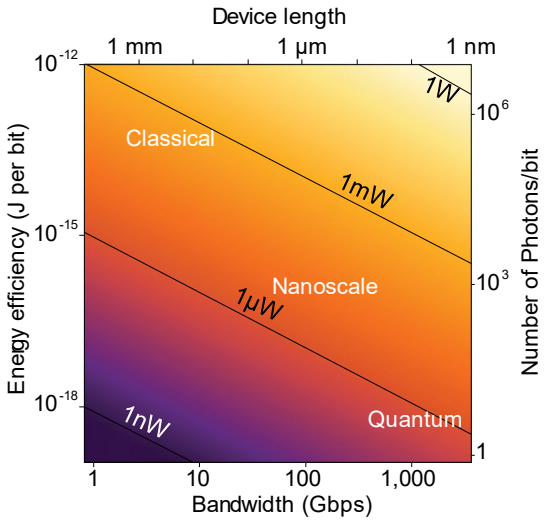


Figure 36. Optoelectronic modulator device scaling laws. *Source: Sorger et al. 2015*

Photodetectors

The photodetector circuits in long distance communication links are generally designed for maximum sensitivity for weak signal recovery in the presence of noise. The amplification

circuitry in these links is, in some cases, the largest energy consumer (Krishnamoorthy et al. 2015). However, for short distance interconnects, the requirements change substantially because the link operates far above the noise floor (Krishnamoorthy and Miller 1996), analogous to short point-to-point electrical wiring. The requirement for short chip-to-chip or intra-chip optical interconnects is to make the optical interconnect behave just like a short electrical wire, without the overhead associated with low-noise amplification, line coding, clock and data recovery (CDR), or serialization/deserialization (SERDES) needed for long-distance interconnects. These simplifications make it easier to conceive of optical interconnects between chips while also enabling those low-energy interconnections to be extended to much greater lengths (meters rather than centimeters) because of the absence of charging capacitance and very low signal attenuation.

If a photodetector can be made small enough and close enough to the input gate of a CMOS transistor, the photodetector can generate a voltage swing large enough to drive the transistor directly without any amplification—a so-called "receiverless" photodetector (Miller 2017) able to achieve ~1 fJ/bit total energy for the receiving system. To achieve the low photodetector capacitance to reach this performance level, the photodetector's volume must be on the order of 1 cubic μm or less (Miller 2017). Adequate absorption can be achieved in the direct-bandgap III-V semiconductors commonly used for commercial fiber optic receivers. Furthermore, emerging concepts of quantum-thin layered materials, e.g., transition-metal dichalcogenides (TMDCs), show a very high absorptivity ($\alpha \approx 10^5$/cm), exceeding that of III-V materials by an order of magnitude or more. Combining such emerging materials with optical-mode compression and impedance-matching techniques can enable receiverless and self-powered photodetection schemes and hence more efficient links (Wang, Sorger, and Dalir 2022; Wang et al. 2023).

The photodetector must also be very close to the input transistor (within about 1 μm), necessitating monolithic integration. Monolithic integration of III-V devices on silicon ICs is a challenge that has attracted significant research attention over time, including epitaxial deposition of III-V quantum dots on Si (Wu, Tang, and Liu 2019). Recent work (Wen et al. 2022) demonstrated monolithic fabrication of InP/$In_{0.5}Ga_{0.5}As$/InP p-i-n heterojunction photodiodes that were also capable of working as LED emitters. These photodiodes could be an alternative to traditional lasers and EOMs as optical transmitters. The key to their effectiveness as transmitters lies in the ability to confine their light emissions to a single mode, which enhances the directionality and intensity of the light. Concentrating structures are used to achieve this confinement, optimizing the photodiode's output for more efficient optical communication systems (Miller 2017).

Interconnection Optics

The optical medium interconnecting transmitters and receivers also poses several challenges to enabling widespread use in chip-to-chip or intra-chip connections. The list of challenges includes integrating a practical number of channels, achieving high bandwidth transfer, minimizing losses of signal power within the medium and at interfaces, and managing costs.

Optical fiber coupling may be accomplished via grating couplers, edge couplers, or evanescent couplers (Mekawey et al. 2022). Challenges for the waveguide include coupling losses at either end, losses in propagation, and bending radius (which can directly impact losses or crosstalk within the waveguide due to reflections). Very compact and massively parallel optical interconnects will require advances in these areas.

In the design of communication links, there is a fundamental trade-off between the number of physical channels, the data rate supported by each channel, and the complexity of circuitry at either end of the link. In long-distance communications, Serializer/Deserializer (SERDES) circuitry is commonly used to merge multiple data channels into a single serial stream. This approach simplifies the physical interconnect but can constrain the flexibility needed for shorter connections on a chip or between boards. For these shorter connections, parallel communication architectures are preferred due to their lower latency and simpler circuitry. Additionally, in the optical domain, signals can be combined using Wavelength-Division Multiplexing (WDM) rather than time-multiplexing. WDM allows for substantial bandwidth on a single optical channel by utilizing the high frequency of optical carriers. For instance, a demonstration by Liu et al. (2019) achieved a transmission rate of 4.1 Tbps using 64 WDM channels, showcasing how optical technologies can efficiently manage the trade-offs between channel count and data rate to achieve high bandwidths with reduced complexity.

Monolithic integration is crucial in the context of intra-chip optical links because it allows the entire system—transmitters, receivers, and waveguides—to be fabricated as a single structure directly on the semiconductor substrate. This integration enhances compatibility with existing semiconductor processes and significantly improves the efficiency and compactness of the communication system. For instance, a 2020 study by Liu et al. offers a practical example of these benefits. The authors developed a monolithic plasmonic waveguide that drastically outperformed traditional electrical connections in terms of signal latency and energy dissipation. Their design achieved signal latencies of approximately 0.18 to 0.19 picoseconds (ps) and energy dissipation rates between approximately 2.5×10^{-3}–3.8×10^{-3} fJ/bit. Additionally, they reported minimal crosstalk with a coupling length of 155 to 125 μm, demonstrating effective isolation between channels over short distances. This example underscores the potential of monolithic integration to significantly enhance the performance of optical interconnects on chips.

Optical signaling also opens the possibility of free-space communication links between chips or boards. The theoretical diffraction-limited density of such connections is enormous; for example, two 1×1 cm surfaces separated by 1 cm could theoretically support up to 100 million channels, or up to 10,000 channels if separated by 1 meter (Miller 2000). The optical interface could consist largely of conventional imaging optics or lenslet arrays. Interconnections of this type can not only transfer tens of Tbps between chips, but also can enable clock signals to propagate reliably over distances of meters to enable much larger machines to operate in strict synchronization than is possible with electrical connections.

Finally, cost is one of the biggest obstacles to the widespread replacement of electrical connections with optical connections. Photonics packaging is far more expensive than conventional electronics packaging (Mekawey et al. 2022), making cost reduction a central concern for the development of optical interconnect technology. Automating the chip packaging ecosystem is expected to significantly reduce costs. Furthermore, achieving high-performance and low-energy photonic links that advance heterogenous technology system-on-chip solutions may drive costs down at the system level.

Action Plan for Optical Interconnects

Table 47. Action Plan for Optical Interconnects.

Scope			
Technology for Energy Efficiency	Using optical interconnects to replace metal for on-chip and off-chip communications to lower energy costs, improve latency and bandwidth, and move past the data deluge		
Technologies of Interest	• Pluggable transceivers (400 Gbps) (100 m to 1 m) (co-package electronics and optics) • Optical engines (close to processors for TB/s) (<1 m in general, depends on use case) • Optical interconnect for data center architectures, rack to rack (100 m to 1 m) • Optical interconnect for chip-to-chip connection (cm scale) • Optical interconnects for intra-chip and 2.5/3D HI connections (mm scale)		
Challenges		**Solution Pathway**	
• Miniaturize key optical components. • Minimize supporting circuitry and associated capacitance. • Integrate electro-optic devices monolithically on silicon CMOS chips. • Improve manufacturability and cost.		• Improve micron-scale electro-optic modulators and light sources. • Achieve monolithic integration of III-V photodetectors on CMOS silicon. • Develop advanced interconnection optics including free-space links.	
Major Tasks / Milestones	**Metrics**	**Targets**	**Timeline (years)**
Advanced electro-optical modulators	50 fJ/bit, linear footprint < 500 µm	ER/IL = 1.0	3–5
	10 fJ/bit, linear footprint < 250 µm	ER/IL = 2.0	6–9
	1 fJ/bit, linear footprint < 100 µm	ER/IL = 5.0	10–15
	0.1 fJ/bit, linear footprint < 10 µm	ER/IL = 10.0	16–20
Advanced photodetectors	GBP = Responsivity x Speed [A/W x b/s]	0.7 x 40 = 28 G	0–3
		1 x 50 G	3–5
		2 x 100 G	6–9
		10 x 200 G	10–15
Optical I/O	Coupling Efficiency x Channel count [% x n]	50% x 1–4	0–3
		70% x 8	4–8
		90% x 64	9–15
		99% x 256	15+
Laser source	Efficiency-Channel-Product [% x N]	10% x 1	0–3
		20% x 8	3–6
		30% x 64	7–12
		50% x 256	12+
Waveguide (passive) platform	Loss x bending-radius [dB/cm x um]	0.2 x 50	0–3
		0.1 x 30	3–6
		0.01 x 15	7–12
		0.001 x 2	12+
Stakeholders and Potential Roles in Project			
Stakeholder	**Role**		
Industry Groups	Free MPW runs, then dedicated runs as partnerships		
End Users/OEMs	Joint demo projects, sharing what end-use/product limitations exist, to help design new photonic ICs and components		
Academia	Long-range and exploratory device and system technology development		
National Laboratories	Metrology, placement for initial demonstrations, initial technology ‘leaps’		
Government	Funding support, convener role through centers of excellence		
Other	Standards development (e.g., IEEE, SPIE)		
Required Resources		**Cross-Collaboration with Other Working Groups**	

• Academia: Photonics 10-year R&D center support for long-term planning (e.g., $20M for 10 years, $2M/year to fuel the technology pipeline) • Government: Funded centers of excellence, 2–3 programs, $50+M each, to accelerate the technology and connect R&D to product demonstrations	• Materials and Devices: Develop more efficient EOMs (e.g., ER/IL) and photodetectors; optical coupling losses • Circuits and Architectures: Implement co-design of driver/control of photonic components • Algorithms and Software: Optimize algorithms to increased bandwidth • Education and Workforce Development: Accelerate photonic IC design education and expertise

2.3.3 3D Hybrid Bonding

As 3D packaging approaches continue to evolve, a critical need is to develop capacity for robust wafer and die stacking with improved interconnect methods. The traditional C4 technology, which involves soldering connections at the corners of stacked chips, has been gradually superseded by copper bumps or microbumps. These create numerous vertical copper-to-copper interconnects between the stacked elements. offering enhanced bandwidth and energy efficiency compared to traditional soldering approaches. However, reducing the pitch—the distance between each connection—to less than 10–15 µm is essential for further bandwidth and efficiency gains. This miniaturization presents significant challenges, but is necessary to meet the demands of high-performance, energy efficient devices (Albright 2022).

3D hybrid bonding, which creates chip interconnects using both metal (copper) and adjacent dielectric elements (SiO_2, SiCN, Si_xN_y), facilitates chip-stacking connections below the 10 µm level. These permanent dielectric-to-dielectric and metal-to-metal bonds can in turn deliver orders-of-magnitude improvements over copper microbumps, reducing signal delay, enhancing bandwidth and memory density, and improving energy efficiency (Hiebert 2023). 3D hybrid bonding is also referred to throughout the industry as a direct bond interconnect (DBI). Comparisons to other bonding methods and an example schematic of 3D hybrid bonding are shown in Figure 37 and Figure 38, respectively.

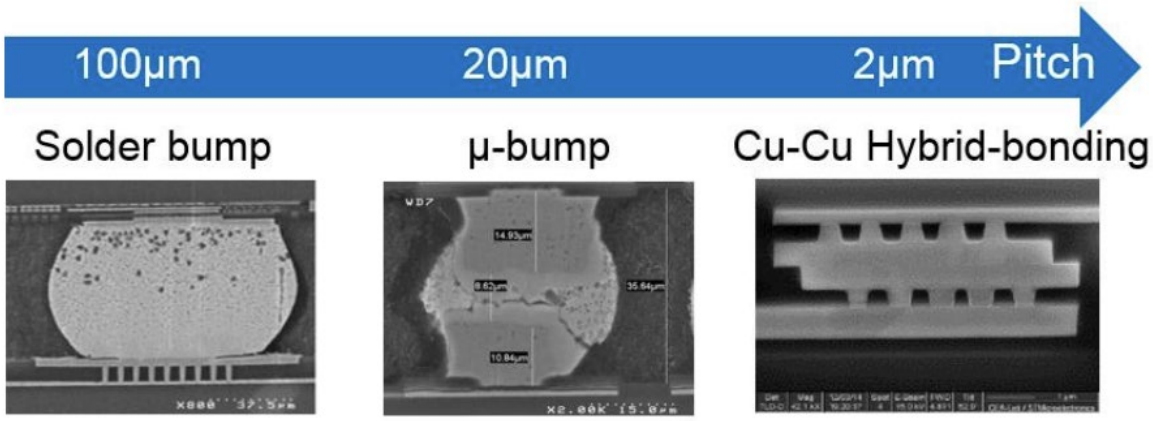


Figure 37. Comparative images and size scales for solder, microbump, and 3D hybrid bonding interconnects. Source: Jani 2019

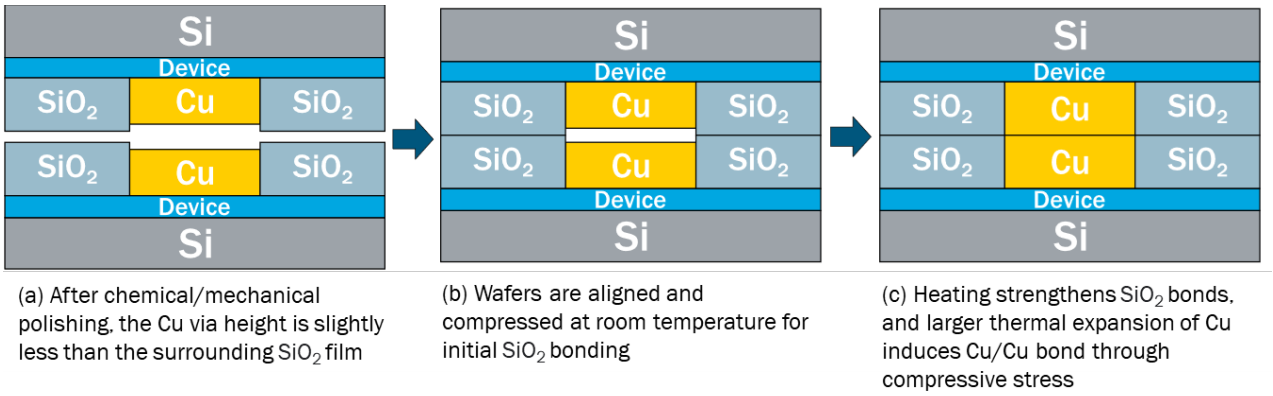


Figure 38. 3D hybrid bonding process with Cu vias and SiO_2 films. *Source: Ong et al. 2022*

Copper-to-copper bonding (μBump) is frequently used for 3D packaging due to its low resistivity, high energy efficiency, and ability to accommodate 15–20 μm pitch sizes. However, its high bonding temperatures and the difficulties in scaling below 10 μm make it impractical within many heterogeneous integration applications or with components (such as certain logic devices) that have a limited thermal budget. 3D hybrid bonding—using copper and a dielectric material—has proven suitable for mass production of CMOS devices and has the potential to achieve the interconnect densities that will be required for the next generation of vertically stacked packaging (e.g., 3D memory on logic integration). The dielectric materials bond well at lower temperatures without external pressure, eliminating many of Cu-to-Cu bonding's thermally induced issues while potentially allowing for submicron copper pad pitch sizes and, most importantly, improving energy efficiency.

Table 48. Impact and Timeline Estimates for 3D Hybrid Bonding

3D Hybrid Bonding		
Energy Efficiency Improvement	Incumbent Technology	Timeline to Demonstration
3x	Cu microbumps	0 years

Reduced pitch sizes and shortened electrical paths allow 3D hybrid bonding to achieve lower power consumption and latency relative to Cu-to-Cu bonds, while also reducing thermal resistance. Polymer adhesives/underfill materials are no longer needed because the dielectric materials themselves serve as the underfills. Wafer-to-wafer hybrid bonding approaches have been utilized in image-sensing applications for the past few years, and the potential benefits of hybrid bonding for enabling heterogeneous integration has led to a significant industry push in this direction (Albright 2022).

Recent examples of commercial applications using 3D hybrid bonding include AMD's 3D V-Cache™ technology. It debuted in 2021 and has been implemented across various gaming and high-performance server processors. AMD described it as the industry's first 3D product for HPC applications and the first demonstration of hybrid bonding used in these applications (AMD 2023). The company has indicated >15x improvement in interconnect density and >3x improvement in interconnect energy efficiency relative to using 3D microbumps. IBM and ASMPT presented a new hybrid bonding method at the 2023 IEEE Electronic Components and

Technology Conference, touting a bond thickness between chiplets of around 0.8 microns (Murphy 2023) (significantly thinner than what is possible with solder). Finally, leading foundry United Microelectronics Corporation (UMC) has been working with Cadence Design Systems, recently certifying their 3D hybrid-bonding technologies within Cadence's design platforms, allowing the foundry's customers to develop 3D systems more easily and accelerate these systems' time to market (UMC 2023) while improving energy efficiency.

Challenges and Solution Pathways for 3D Hybrid Bonding

Hybrid-bonding techniques are being implemented in various advanced 3D packaging applications. However, as the industry pushes toward single digit micron and submicron pitch sizes, key challenges such as device alignment and process controls become more critical and need enhanced solutions. Additionally, reliability issues such as electromigration and copper diffusion become more pronounced at these smaller scales and must be addressed.

Alignment, Metrology, and Process Controls

As pitch sizes shrink, tighter control of bond-pad alignment—with submicron accuracy—to ensure secure connections will be increasingly important. New metrology and process control techniques are needed, both to meet these requirements and to maintain satisfactory process yield. Different bonding structures and increased keep-out zones (i.e., unused die areas) may also be needed in some cases, such as with double-sided bonding. More stringent process control over aspects like chemical mechanical polishing (CMP), wafer dicing, wafer/die cleaning, dielectric thickness and surface topologies, and the dish-like shape of the copper pads must also be considered.

Contamination control, bonding temperature precision (Hiebert 2023), and thermal-budget management—especially for cases involving elaborate 3D architectures—must also be dealt with going forward. Efforts to address these challenges will center around process optimization and continuous improvement.

Die Integrity and Manufacturing Standards

Particularly for die-to-wafer hybrid bonding, ensuring that only good dies are used is vital to the integrity of the final chip's performance. Additional process control steps, along with accurate manufacturer-provided die-quality information, will be essential within such applications (Hiebert 2023). New manufacturing standards will also be necessary to guide hybrid bonding's strength, durability, cost, and emissions considerations going forward. These new standards will likely follow comparable copper-bond standards.

Challenges at Reduced-Length Scales

Potential challenges for bond reliability due to pitch shrinking and higher-density interconnects must also be identified and addressed going forward. These could include local current concentration and electromigration, short circuiting (due to dielectric breakdown and copper diffusion to dielectrics), dielectric reliability deterioration (due to shorter conduction paths), and path breakdowns (due to misalignment of the copper bonding pads). Additional work will also generally be needed in developing and improving 3D EDA tools for both multilayer and heterogenous integration.

Action Plan for 3D Hybrid Bonding

Table 49. Action Plan for 3D Hybrid Bonding

Scope			
Key Technology for Energy Efficiency	High-performance and energy-efficient 3D hybrid bonding		
Technologies of Interest:	• Ru-, Co-, Ir-, and Rh-based metal interconnects • Chip-to-chip: EMIB/Foveros, TSMC interposer, chiplet, Through silicon vias, hybrid bonding • Edge bonding		
Challenges Addressed		**Solution Pathways**	
• Replace solder-based interconnect technologies with smaller, more densely packed interconnects featuring increased input and output pins to enhance the energy efficiency of 3D technologies. • New bonding methods enabling reduced pad size are required and must show thermal and mechanical reliability and electrical properties on par with those of metal-to-metal bonding. • Scaling Cu pitch sizes below 10 µm with high thermal budget. • Thermal budget differences for different components • Removal of CuOx for improved Cu-to-Cu connections.		• Develop low-temperature bonding methods that can prevent Cu oxidation and enable fine-pitch structure with a high density. • Continue to integrate bumpless Cu hybrid bonding technology and surrounding dielectrics (e.g., SiO_2, SiCN, SIN). Shown to be suitable for mass production of CMOS devices while increasing interconnect density below 10 µm, which can enable 3D stacking. • Develop inorganic dielectrics with low bonding temperature, alleviating thermal gradient issues. • Achieve lower power use and latency from reduced pitch size and shortened interconnect length with hybrid bonding. • Investigate optimal surface treatment methods to remove copper oxide, including methods like using cohydroxylated and cohydrophilic copper oxide. Explore selective thermal atomic layer deposition of copper and adjust bonding temperatures to improve the bonding interface.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline**
Understand the effect of bonding process parameters on hybrid bonding quality to allow good process control (e.g., mechanical polishing, wafer/die cleaning, and wafer dicing)	• Improved device-to-device communication (TOPS/W) • Lower power delivery	• Sufficient yield in comparison to solder techniques • Further reduction of thermal budget in hybrid bonding • Precise alignment with submicron accuracy, different bonding structures (for double-sided bonding), and increased keep-out zones to realize successful bonding and high yields	0–2 years (initial deployment) 2–5 years (process improvement)
Require manufacturing standards to guide hybrid bonding strength, cost, and emissions	• Comparable Cu bond standards	• Meet testing standards for durability • Meet or beat emissions standards	0–2 years
Require high processing quality standards to control surface flatness for D2D or D2W hybrid bonding. Debris, burrs, and other particulates generated during die singulation would induce uneven topography of bonding surface and void formation, leading to poor electrical connectivity or open-circuit failure.	• Flatness, low dishing	• Via precisely controlled chemical mechanical polishing (CMP), achieve <1-nanometer surface roughness and <0.1-nanometer leveling of copper dishing for well-controlled thermal expansion. • Carefully determine bonding pattern density, configuration, and topology of adjacent layers to minimize impact to surface topography.	0–2 years

Identify and address bonding reliability issues associated with high-density interconnects.	• Pitch shrinking induced local current concentration. • Electromigration effects. • Short circuit due to dielectric breakdown. • Copper diffusion to dielectrics. • Dielectric reliability deterioration due to shorter conduction paths or breakdown paths caused by misalignment or overlay of bonding pads.	• Improve the reliability of interconnects to sustain pitch shrinking down to submicron sizes. • Enhance the stability and performance under high-density conditions.	Ongoing efforts with targeted improvements expected to be implemented within the next 3-5 years

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Continue to create and advance hybrid bonding, ensure reliability • EDA suppliers must advance current technology and simulation using high-density interconnect (HDI)
End Users/OEMs	• Redesign for the use of this technology to achieve speed and energy efficiency gains. Possibly add redundancy given reduction of size (HDI).
Academia	• EDA for design • Die thinning, co-planarity experimentation
National Laboratories	• EDA for design • Die thinning, co-planarity experimentation • Initial device prototyping • Simulation to understand transport properties with different bonding configurations
Government	• Funding, part stability (NASA, military, etc.)

Required Resources	Cross Collaboration Needs of Working Groups
• Funding and locations for die thinning, application specific HDI equipment for smaller companies • Allocate resources for academic and national laboratory research on integration techniques and simulations, including device functionality, heat management, and power distribution. • Reliability testing • Hardened electronics	• Manufacturing Energy Efficiency and Sustainability: Develop the hybrid bonding and Cu conductive pathway process; ensure alignment while moving to smaller pitch size for stacking. • Circuits and Architectures: Develop circuit designs and architectures that leverage advanced packaging techniques. Focus on scaling down global interconnects and enhancing thermal management to accommodate smaller device footprints and increased density. • Algorithms and Software: Explore potential need for new programming. • Power and Control Electronics: Deliver power through different architectures.

2.3.4 Vertical Integration (2.5D/3D)

Vertical 2.5D and 3D packaging approaches offer significant opportunities for creating more efficient and faster microelectronic devices compared to traditional 2D architectures, such as PCIe 2D interconnects. There has been a surge in advanced-packaging technologies over the past 25 years, building on earlier efforts in wire bonding and flip-chip approaches (see Figure 39) and incorporating a diverse range of vertically oriented solutions that allow more efficient access to memory and other IC components.

2.5D/3D Nonmonolithic Technologies

In advanced packaging, vertical 2.5D approaches generally utilize an interconnect carrier such as a silicon interposer layer, heterogeneous interconnect stitching technology (HIST), or bridge chips to route wires horizontally. 3D non-monolithic packaging involves two or more chips being stacked vertically, typically connected either with a through silicon via (TSV) and microbumps or through bumpless 3D hybrid bonding (Burkacky, Kim, and Yeom 2023). The wire lengths for 2.5D integrations are generally in the 100 micron–5 millimeter range, while vertical 3D layers use TSVs and nanoscale vias with lengths around 100 nanometer–100 microns, affording significant efficiency and latency improvement potential in the transition from 2.5D to fully 3D architectures (Zhang, Zhang, and Bakir 2018). A detailed benchmarking study by Zhang et al. evaluated various 2.5D and 3D integrations—including bridge chips, interposers, HISTs, and 3D monolithic (discussed in the next sub-section) and non-monolithic approaches—based on typical configurations and component sizing. Figure 39 illustrates the schematics and relative capacitance breakdowns for these different signal channels.

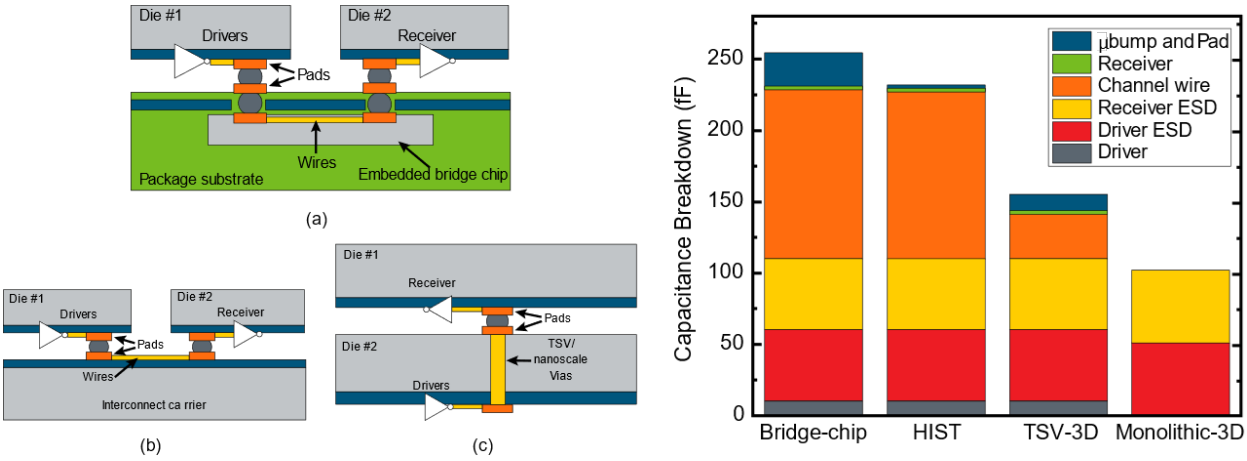


Figure 39. Digital signal channel paths and associated capacitance. *(Left) schematics of digital signal channels: (a) bridge-chip 2.5D, (b) interposer and HIST 2.5D, and (c) 3D integrations; (right) capacitance breakdowns for these signal channels (ESD = electrostatic discharge capacitor). Source: Zhang, Zhang, and Bakir 2018*

Silicon interposers, which have been available for over a decade, incorporate horizontal electrical connections between adjacent die, combined with TSVs that pass vertically through the silicon die/wafer. Taiwan Semiconductor Manufacturing Company (TSMC) is the main industry supplier of a range of interposer technologies (Burkacky, Kim, and Yeom 2023). While interposers generally serve as a 2.5D interconnect technology, TSVs can also be utilized for fully 3D chip-stacking configurations. HIST uses a 2.5D and 3D design similar to an interposer configuration but uses stitch chips with high-density, closely packed wires along with multi-height, compressible micro-interconnects (Jo et al. 2018) to enable chiplet-based designs.

Bridge chips provide high-density interconnects between die and can achieve larger sizing than is often possible for silicon interposers, which can face cost and technical limitations (Bakir 2022). Compared to interposers, bridge chips are a relatively new technology and typically use less silicon (Burkacky, Kim, and Yeom 2023). The following examples help illustrate the diversity of such technologies being explored or brought to market: Intel's embedded multi-die

interconnect bridge (EMIB) silicon bridge technology uses thin silicon pieces embedded inside an organic substrate to connect adjacent die (Keser and Kroehnert 2019). Samsung, IMEC, and others have pursued similar 2.5D silicon-bridge approaches (Lapedus 2018). Intel's Foveros technology is a 3D face-to-face connection between dice/chiplets that complements the EMIB's 2.5D functionality, using small microbumps to accomplish the chip-on-chip bonding and improve overall interconnect density (WikiChip 2023). Another significant development in chip stacking is AMD's Ryzen technology, which employs hybrid bonding—not a bridge chip technology but rather a different approach to achieve a 3x improvement in energy efficiency (AMD 2023).

A major bottleneck in advancing 2.5D and 3D HI technologies is the lack of standardization, notably the absence of universally adopted frameworks like the Universal Chiplet Interconnect Express (UCIe). This issue is compounded by the limited number of third-party vendors in the industry. Most major fabs typically do not incorporate wafers from external sources, limiting the diversity and innovation typically brought by third-party contributions. This restriction is particularly challenging for advanced packaging and fabrication processes like Through-Silicon Vias (TSVs) and die-to-die bonding, which require complex and extensive manufacturing infrastructure. Moreover, 3D monolithic architectures face significant challenges related to yield, cost, availability of suitable materials and devices, and effective thermal management (Zhang, Zhang, and Bakir 2018).

3D Monolithic Technologies

While 2.5D and 3D non-monolithic approaches offer significant energy and signal-delay improvements over traditional 2D packaging, the relatively large size and capacitance of TSVs remains a limiting factor. Current technology nodes are in the single-digit nanometer range, yet TSVs generally have diameters of a few micrometers, along with large pitch (30–50 μm), large keep-out-zones, and, accordingly, large capacitances (Dhananjay et al. 2021). Monolithic 3D integration technologies allow device layers to be sequentially assembled in the vertical direction; thus, multiple layers of transistors can be fabricated above a single substrate (Dhananjay et al. 2021). Monolithic inter-tier vias (MIVs) serve to interconnect these vertical device layers, where the vias' diameters are orders of magnitude smaller than those of both TSVs and mini-TSVs (see Figure 40).

The diagram shown in Figure 40 visually represents the size differences among 14-nm NAND gates, Monolithic Inter-tier Vias (MIVs, 50-nm), Mini TSVs (2μm), and regular TSVs (5-μm) compared to a 28-nm NAND gate. Each shape and size reflects the relative scaling: the small rectangle for the 14-nm NAND gate signifies its base size of 1 times; circles are used for MIVs and TSVs to indicate their cylindrical nature, which makes them appear disproportionately larger due to their area encompassing both the core and insulation; the larger rectangle represents the older technology of a 28-nm NAND gate, emphasizing the significant reduction in scale over time. The energy-saving opportunities from these various 2.5D and 3D vertical integration approaches are estimated in Table 49.

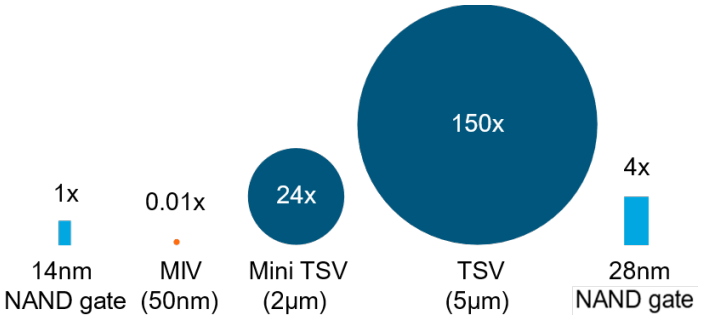


Figure 40. Relative sizes of typical NAND gates, MIVs, and TSVs. Source: Samal et al. 2016

Lowering interconnect capacitance means lower energy per bit and higher bandwidth. For analog-sensing applications, it also means lower noise and power use in the subsequent signal/amplifier stage. While 3D monolithic integration faces several challenges distinct from those of the 2.5D and 3D non-monolithic approaches—including thermal management, cost, modeling, and yield constraints—the energy efficiency and latency benefits of 3D monolithic approaches make them promising for high-performance applications. With the ability to mix different process technologies into the vertical layers (logic + memory, logic + logic, logic + analog, etc.), monolithic 3D opens nearly endless IC design opportunities.

Although there are currently no examples of commercially available 3D monolithic technologies (Dhananjay et al. 2021), these advanced packaging approaches are expected to play an increasingly important role over the next decade. Current limiting factors include overall manufacturing complexity, the need for low-temperature processing of the stacked device layers, and insufficient 3D design infrastructure and metrology.

Vertical integration, sometimes referred to as “More than Moore,” is of great interest to the EES2 community, given its vast landscape of integration technologies with associated performance and energy efficiency improvements. Table 49 compares the various vertical integration approaches in terms of energy per bit; the TSMC interposer represents the commercial benchmark product for 2.5D as a mass-produced state of the art.

Table 49. Energy Per Bit Comparisons of Different Vertical Integration Schemes

Specified Technology	Baseline Performance	Commercial Benchmark Product	Commercial Benchmark Performance	Performance Multiplier	Timeline (years)
Bridge Chip, EMIB/Foveros	150 fJ/bit	TSMC Interposer	263 fJ/bit	1.6	0
TSMC Interposer	263 fJ/bit	TSMC Interposer	263 fJ/bit	1	0
Heterogeneous Interconnect Stitching Technology (HIST)	259.9 fJ/bit	TSMC Interposer	263 fJ/bit	1	0
Through-Silicon Vias (TSVs)	176.2 fJ/bit	TSMC Interposer	263 fJ/bit	1.5	0
Monolithic Inter-Tier Vias (MIVs)*	0.1 fF	TSV	5 fF	50	7–10
Monolithic 3D	135.1 fJ/bit (with ESD) 3.7 fJ/bit	TSMC Interposer	263 fJ/bit	1.9 (with ESD) 71.1	7–10

	(without ESD)			(without ESD)	
UCIe	250 fJ/bit (2.5D)	PCIe	10 pJ/bit (2D)	40	1–3
3D Hybrid Bonding**	<1	3D microbump	3	>3	0

Sources: Mahajan 2016; Angelini 2020; Zhang, Zhang, and Bakir 2018; Salman 2023
ESD for monolithic 3D refers to ESD protection capacitor.
*Energy per bit for MIVs are not found from literature; Using femtofarad to represent capacitance.
*3D hybrid bonding energy metric taken from AMD. No approximation was done given the unknown size of 3D microbumps and how this affects energy per bit.

Challenges and Solution Pathways for Vertical Integration (2.5D/3D)

Development of Standard Interconnect Schemes for 2.5D/3D ICs

There is currently a lack of standardization for 2.5D and 3D heterogenous-integration approaches. The ability to more readily mix and match dice/chiplets and other IC devices from different manufacturers and different technology nodes, for example, would allow for integrators to adjust more easily to supply chain issues or incompatibilities between different generations of technology nodes. An interconnect standard for 3D ICs (similar to UCIe) alongside creation and adoption of fabrication standards for heterogeneous integration—as well as 3D-compatible process design kits (PDKs) from foundries—would all serve to facilitate more rapid development and deployment of HI solutions.

New Electrical Design Automation Tools for 3D IC Co-Design

To reduce design time and risk, development of 3D electronic design automation (EDA) tools for both multilayer and heterogeneous integration will be essential. These resources will need to allow for co-design of different technologies and applications, satisfying key energy and performance constraints while modeling important criteria such as thermal loads, electromagnetic interference, and resource optimization. The collective expertise from the industry and academia ensures the development of comprehensive security and verification methodologies, as discussed in sections 2.2.7 and 2.3.6 of the roadmap.

Prototype Development Issues

If fabs limit or do not allow other intellectual property into their facilities, this will hinder widespread success of APHI. The creation of an integration-minded fab/fablet might be of use in overcoming such obstacles. Academia will need access to integration-minded facilities to develop prototypes and help grow a workforce with the essential working knowledge of 3D ICs. Potential solutions include establishing access to small-volume production and prototyping facilities to foster innovation and workforce development, along with building a domestic supply chain that incorporates integration vendors capable of accepting and combining wafers from various foundries.

Bolstering domestic fabrication capabilities would improve prototyping, allow for PDK development, and ideally establish libraries that encompass chiplets, 3D stacks, and monolithic 3D. Ultimately, the industry must be able to provide small-scale 3D manufacturing/packaging capabilities at a reasonable cost.

Power Delivery

2.5D/3D architectures present unique power delivery challenges. Unless die on different process nodes have separate power sources, stacked chips can compete for power resources, and transistors toward the top of a stack will see greater drops in voltage due to power traveling through multiple TSVs (SRC 2023). New power delivery methods and novel materials for power delivery are both needed. Technical targets identified include maintaining voltage noise within 5% to 10%, power efficiencies greater than 95%, and on-die temperatures less than 80°C. Additional considerations related to power delivery can be found in the Power and Control Electronics section of the roadmap.

2.5D/3D Nonmonolithic Package Assembly

The transition to AP represents a significant shift in how wafer packaging is typically handled today. Currently, back-end packaging is most often outsourced to semiconductor assembly and test companies (OSATs) (Burkacky, Kim, and Yeom 2023). However, some of the APHI technologies mentioned will require processing conditions more typical of front-end fabrication and/or have stringent processing requirements. This may alter the role and importance of OSATs.

Thermal Budget Constraints

Novel packaging schemes that integrate devices with one another will increase thermal density, requiring new methods for heat removal. As thermal densities reach over 100 W/cm^2, conventional air cooling will reach its limits. This requires investigation into improved TIMs, such as nano-etched surfaces with high elastic modulus and thermally conductive materials; improved heat sinks (such as nanodiamond copper); or water-cooled microchannels etched in copper; alongside system-cooling technologies such as immersion cooling or direct cooling through microfluidic channels (IEEE HIR 2023). Additional information on thermal budget challenges and solution pathways is found in section 0, Thermal Management.

Action Plans for Vertically Integrated Devices and 3D Monolithic Integration

Table 51. Action Plan for Vertically Integrated Devices

Scope		
Technology for Energy Efficiency	• Energy-efficient vertically integrated devices	
Technologies of Interest:	• Chiplets for 2.5D integration • Wafer-to-wafer 3D stacking • Monolithic 3D integration with Multi-tier Vias (MIVs)	
Challenges Addressed		**Solution Pathways**

• All functional units implemented with the same technology node, leading to inefficiencies. • High energy consumption for data movement within and between chips. • Significant RC delays and signal issues in large interconnects. • Low yield in monolithic 3D technologies. • Limited availability of EDA tools for effective 3D integration. • Lack of access to small business and university facilities, which hampers innovation.	• Develop and adopt universal standards and PDKs for HI and 3D fabrication. • Standardize interfaces for different functions within chiplets and 3D stacks. • Create and refine 3D EDA tools to support co-design across different technologies. • Enhance security and verification for new EDA tools. • Establish a robust supply chain for heterogeneous integration. • Enable small-volume production and prototyping to encourage innovation. • Apply co-design strategies to manage thermal integration and reduce required energy in 3D and monolithic architectures. • Utilize novel materials and advanced techniques for high-yield monolithic 3D integration. • Integrate memory technologies like NVM in 3D architectures to reduce latency and energy consumption.

Major Tasks/Milestones	Metrics	Targets	Timeline
Development of standards, including pin maps	• Adoption time • Cost	• Quick turnaround time for R&D • Cost reduction • Adoption of standards	3 years
3D native EDA tools	• Complexity/runtime accuracy • Ability to capture thermal issues and thermal optimization	• Reduction in design time and risk, Improvement in runtime accuracy	3–5 years
Power delivery and thermal management	• Voltage noise • Temperature (transient and steady-state) • Power loss	• Maintain voltage noise within 5%–10% • Power efficiencies larger than 95% • On-die temperatures <80°C	3–5 years
Domestic fab/fablet	• Cost • Production capacity	• Prototyping capability • Build PDKs, libraries for chiplets • 3D stacks • Monolithic 3D	5 years
Development of novel materials	• Thermal dissipation capability • Yield for monolithic 3D	• Reduce I/O parasitics • Increase thermal conductivity	5 years

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• E.g., integration/assembly vendors/suppliers • Provide small-scale manufacturing/3D packaging capability at reasonable cost
End Users/OEMs	• Negotiate with stakeholders for providing IP resources
Academia	• Models and frameworks for benchmarking different 3D technologies • Cross-layer optimization methods to satisfy runtime thermal constraints • Provide workforce that is knowledgeable in advanced 3D packaging and monolithic 3D integration
National Laboratories	• Development of novel materials, thermal solutions, simulation capability for temperature, 154ultiphysics • Lead support; act as a support center for academic community • Lead development of standards for ease of adoption of 3D heterogeneous integration (3DHI)

Government	• Financially support the supply chain • Establish ecosystem providing design and prototyping capability

Required Resources	Cross Collaboration Needs of Working Groups
• Manufacturers/Suppliers: Funding and availability of users, potential tax incentive by government • Academia: Test equipment for validation, funding for prototyping and education • National labs: HPC equipment • Government: Funding and resources such as real estate, water, and electricity; tax incentives for using national vendors/suppliers	• Algorithms and Software: Develop new design, new algorithms • Power and Control Electronics: Develop new design, new power paradigm • Metrology and Benchmarking: Understand and mitigate failures • Education and Workforce Development: Train next-generation workforce in chiplet design and test flow, chip stacking, and monolithic 3D integration. As 3D native tools are developed, the workforce should also be familiar with them and 3D-specific design/test methodologies. With these advanced technologies, the boundary between package and die designs is less clear. New educational materials should be developed to consider these characteristics. As domestic fabs for chiplet design and test flow technologies emerge, they will need more workers in diverse fields, including electrical engineering, mechanical engineering, materials science, and chemistry.

Table 50. Action Plan for 3D Monolithic Integration

Scope	
Technology for Energy Efficiency	3D Monolithic Integration
Technologies of Interest:	• Monolithic inter-tier vias (MIVs), interlayer dielectric (ILD) interconnect • Alternative low-temperature devices, processes (CNTFETs, NRAM, ReRAM, rapid annealing [e.g., thermal, laser, other]) • Thermal coupling between devices integrated on one another; may be more challenging (need to design for thermal coupling).

Challenges Addressed	Solution Pathways
• Reduction of RC delay, interconnect length. • Faster access to upper or lower tiered device (memory, logic, other in different tier). • Reduced footprint vs. co-planar solution. • Thermal management for 3D ICs. • TSV RC delays.	• Develop MIVs for improved data transfer, reduced RC delay, and significant improvements in energy savings. • Enable mixing of process technologies (logic + memory, logic + logic, logic + analog, etc.). • Co-design 3D monolithic architecture with thermal performance and software.

Major Tasks/Milestones	Metrics	Targets	Timeline
Standards development	• Adoption rate of new standards • Compatibility with existing fabrication processes • Ease of integration into existing manufacturing environments	• Achieve industry adoption • Ensure compatibility with current leading-edge fabrication processes • Develop guidelines that simplify integration into any standard semiconductor manufacturing environment	Varies
Fabrication of monolithic integration	• Low-temperature processes compatible with FEOL • MIV process development • Thermal mitigation	• CNTFETs (low T), NRAM (SRAM replacement), FeFETS, low-temp. silicon (CEA-List), junctionless transistors (UIUC; III-V materials on BEOL for RF application, carbon nanowires, GaN devices) • Low-temp. annealing methods: Laser, quick thermal anneal without	Certain technologies are available now, NRAM integration, ReRAM integration

		distribution to bottom layers, microwave annealing	
Co-design and tests	• Thermal simulation, device performance, EDA software for design is lacking • Conventional test methods are not applicable (non-idealities of upper tiers)	• See Action Plan for Vertically Integrated Devices • Better probing techniques (finer granularity, deeper), focused ion beams (FIBs) for SEM/TEM for analysis. X-ray/acoustic imaging: look at devices at a given depth, something high throughput.	Varies

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	Focus on FEOL processes by testing chips and wafers; extend to commercial foundries for scale-up.
End Users/OEMs	Drive demand for efficient and low-energy chips through NMI; facilitate efficient logistics for intra-fab shipments.
Academia	Lead in demonstrating technology, developing new materials, optimizing designs, and training the next generation of engineers.
National Laboratories	Serve as early technology adopters and supporters, providing MPW access and fostering the commercial availability of advanced technologies.
Government	Support research and technology transfer between public institutions and private industry, ensure funding and foster inter-sector collaboration.

Required Resources	Cross Collaboration Needs of Working Groups
• Universities and Start-ups: Setting up fablet facility for innovation and practical applications. • Industry: Circuit design, EDA, and fabrication resources to enable the technology, and new PDKs that support the technology (see Action Plan for Vertically Integrated Devices). • Academia: Education and workforce development, such as training students on important challenges (design for thermal integrity, testing). • National Labs: HPC for materials and process development (EDA, simulation at material level of device, computation).	• Circuits and Architectures: Develop new designs. • Power and Control Electronics: Implement new power distribution (ultra-thin dielectric). • Algorithms and Software: Develop new computing paradigm. • MEES: Implement new process development. • Metrology and Benchmarking: Implement new reliability test • See Action Plan for Vertically Integrated Devices for additional considerations.

2.3.5 Thermal Management

The advancement of next-generation, energy efficient APHI technologies necessitates an intensive focus on thermal management strategies at both the die and system levels, the latter of which is detailed in the Power and Control Electronics (PACE) chapter. With the ongoing increase in the density of interconnects, transistors, and memory cells, there is a corresponding rise in power density, which essentially refers to the amount of heat generated per unit area. This increase in power density leads to greater energy consumption due to heightened device and chip parasitics and increases the demand for effective heat dissipation solutions. This problem is further exacerbated by chip stacking, which concentrates the power density and heat of multiple chips into the footprint of just one chip (IEEE HIR 2021).

Fundamentally, power consumed in the circuitry is manifested as heat that must be dissipated, and while lower-resistance interconnects, memory cells, and transistors will help, heat generation remains inevitable due to the inherent switching energy and contact resistance of these components. With the shrinking of transistor devices, DRAM, and other components, the leakage current is increasing, as is the heat produced (Kao, Kuo, and Dai 2016). The performance of all cell types—logic, memory, and even interconnects—worsens with increased heat (Weste and Harris 2011), while energy consumption rises. For example, DRAM operating outside its nominal temperature range can have a performance degradation of 8.6% and increased power consumption (due to leakage) of 16.1% (Zhang, Sarvey, and Bakir 2014). It is paramount to devise secondary methods for heat removal at the device/die and system levels, both to maintain device performance and to avoid unnecessary increases in energy consumption.

Chip stacking and 3D monolithic integration are inevitable, as are the complex heat-removal technologies that must be developed to enable them and boost their energy efficiency. Temperature coupling between device stacks only worsens when they are integrated vertically (Sarvey et al. 2015; Zhang, Sarvey, and Bakir 2014). Additionally, heat removal is more difficult in 3D IC designs, given the increased distance from the heat spreader (Kumar and Naeemi 2017). To help enable the next revolution in microelectronics packaging, the APHI working group focused on interfacial heat removal technologies (called thermal interface materials [TIMs]). The sections below describe the background, technology comparisons, challenges, and solution pathways for TIMs.

Thermal Interface Materials

In conventional packaging, heat generated by the device is transferred through the interconnects and BEOL layers, which may have heat-isolating properties, toward the heat spreader and subsequently to the heat sink. However, when chips are stacked, the path for heat to travel from the inner chips to the heat spreader lengthens, exacerbating thermal management challenges. To mitigate these heat conduction issues, TIMs are placed at the interfaces between chips to help facilitate heat transport toward the heat sink for eventual removal from the system. The use of TIMs is crucial for mitigating localized hot spots, which tend to be less energy efficient and can degrade overall device performance.

TIMs must be thermally conductive to remove the generated heat. They should also be capable of completely filling the gaps between contact surfaces, accommodating the surface roughness of the contacting layer, and remain mechanically stable through numerous thermal cycles as the device powers on and off. TIMs have two distinct levels: TIM 1 which conducts heat from the die to the heat spreader (typically made of copper), and TIM 2, which facilitates heat transfer from the heat spreader to the heat sink (Jensen and Lasky 2020). The choice of TIMs usually depends on their thermal conductivity and elastic modulus to ensure efficient heat management

As shown in Table 51, TIMs can make significant improvements over the state of the art in terms of materials' thermal conductivity and can incorporate engineering features that decrease the temperature of the chip overall. However, significant challenges related to TIMs remain, as discussed below.

Table 51. Performance of Advanced Thermal Interface Materials Compared to Baseline Technologies

Technology Group	Specified Technology	Baseline Energy Performance	Commercial Benchmark Product	Commercial Benchmark Energy Performance	Performance Multiplier	Timeline (years)
Advanced Thermal Interface Materials (TIMs)	Liquid metal paste (LMP) solder with polymer (Indium-based)	70 W/m·K (thermal conductivity)	polymer-based paste, conductive filler particles	10 W/m·K	7	2
	Carbon nanosprings in conductive polymer	100 W/m·K	polymer-based paste, conductive filler particles	10 W/m·K	10	2
	CNT-based thermally conductive matrix	63.7 W/m·K	polymer-based paste, conductive filler particles	10 W/m·K	6.4	2
	Graphene-based conductive matrix	40–90 W/m·K	polymer-based paste, conductive filler particles	10 W/m·K	4–9	2
	Nanostructure engineering to increase surface contact area of TIM with 5.4 W/m·K at 17 CFM air flow	58°C (device temperature)	Indium (~70 W/m·K)	73°C	1.26	5

Challenges and Solution Pathways for Thermal Interface Materials

Poor Understanding of Thermal Interface Resistivity

Although thermal interface resistivity is often poorly understood, it likely defines heat removal characteristics. Advanced TIMs presented in the literature tend to generate excitement because of the high thermal conductivity of these new materials. However, thermal conductivity alone is not an adequate performance indicator. Materials such as CNTs and graphene provide bulk thermal conductivity in the thousands of W/m·K, but there have been no measurements of graphene or CNT TIMs showing greater than 40–90 W/m·K. The primary reasons for differences between theoretical and measured heat transfer are:

- The materials must have adequate contact (lowering thermal interface resistance with the substrates) to promote heat transfer (Jensen and Lasky 2020).
- Material phonon modes need to overlap for adequate heat transfer.
- Heat transfer in certain materials is anisotropic, meaning it varies with direction, which can restrict heat transfer across different axes (Guo et al. 2021; Refai-Ahmed et al. 2018). In the case of anisotropic TIMs, interfacial heat transfer can be limited because the heat at a hot spot cannot be conducted laterally. In contrast, isotropic (direction independent) heat transfer can be approached through alignment technologies such as embedding Carbice™

CNT forest (anisotropic heat transfer) in an aluminum “sandwich” (isotropic heat transfer) (Green, Prinzi, and Cola 2016).

Ensuring the best possible surface contact is important for maximizing heat transfer. For instance, increasing the surface contact area of the flexible TIM between the IC and heat sink significantly reduces temperature compared to normal higher-thermal-conductivity TIMs, as thermal resistance is a function of surface area (Guo et al. 2021). Nanoengineering the IC surface and heat sink to promote increased thermal transfer via improved interfacial conductance will complicate the manufacturing process. However, it is likely needed to help ensure adequate heat removal in 3D ICs.

Because interfacial contact and resistance is the primary indicator of whether a TIM with good thermal conductivity will promote adequate heat removal, the heat transfer process must be better understood, tested, and simulated. Conventional heat measurements for thermal conductivity should be performed, but there must also be adequate device performance demonstration with each TIM integrated. An industry-standard device or device architecture should be implemented to allow standardized testing for publication of new TIM specifications that will be adequate for next-generation 3D ICs. All of the considerations listed above could be added into EDA software via a TIM PDK for improved thermal modeling.

Compatibility With New Advanced System Cooling Technologies

Conventional technologies utilize a heat sink combined with forced air flow for removing heat from the system. While this approach is viable for 2D electronics, it may not be adequate as the industry moves into 3D ICs. Next-generation cooling techniques—such as microfluidic cooling, immersion cooling, and direct liquid cooling—are in development and need the help of advanced TIMs. The TIMs must be compatible with the cooling environment.

Mechanical Durability Through Coefficients of Thermal Expansion Mismatch Affecting Long-Term Thermal Cycling Stability

Effective TIMs must be capable of accommodating the differential coefficients of thermal expansion (CTE) between the two surfaces they connect. The CTE measures how much a material expands when heated and contracts when cooled. This differential in expansion rates can lead to mechanical stress during thermal cycling, which is the process of repeated heating and cooling that occurs in operational environments (Guo et al. 2021). Conventional materials such as Cu, In, and Al, which have great thermal conductivity, lack the malleability to adjust to thermal expansion and contraction when in contact with surfaces. These more rigid materials also will have more trapped air pockets, which reduce thermal transport and increase localized heating. This in turn leads to lower thermal cycling lifetimes. However, materials that do have the ability to survive the changing expansion—such as Carbice CNT forests with polymer, CNT/graphene pastes, and high-elastic-modulus materials—will be better suited for long-term stability (Green, Prinzi, and Cola 2016; Guo et al. 2021).

Action Plan for Thermal Interface Materials

Table 52. Action Plan for Thermal Interface Materials

Scope	
Technology for Energy Efficiency	Thermal interface materials

Technologies of Interest:	Thermal interface materials (TIM1–2)
Challenges Addressed	**Solution Pathways**
• Interfacial adherence and defect impact on heat removal • Testing and simulation standardization • Compatibility with advanced cooling technologies such as immersion cooling	• Design materials whose intrinsic properties can be simulated. Contemporary example: TIM based on elastic CNT on aluminum backbone. • Explore *in situ*, real time interfacial monitoring. • Scale technologies that enable mapping and simulation of thermal resistance distribution at the real-application interface. • Expand focus on TIM for ongoing consortia working on material design. • Develop scalable algorithms that accurately translate technology-enabled interface mappings into simulations that reflect the actual distribution of interface resistance in practical applications. • Propose standardized testing protocols that transition from measuring thermal bulk properties to evaluating thermal interface conductivity, ensuring realistic performance expectations are met.

Major Tasks/Milestones	Metrics	Targets	Timeline (Years)
Gather data to bridge the gap between bulk material properties and system-level performance, enhancing insights at the interface through *in situ* monitoring.	--	Understand failure mechanism (bulk properties vs. system level)	2
Understand and list optimal material properties (at bulk and system interface).	--	Distill a list of optimal material properties for TIM	2
Check chemical compatibility with cooling environment.	Toxicity, durability in advanced cooling environment (e.g., immersion cooling)	Non-toxic, reliable thermal performance under advanced cooling conditions such as immersion cooling (dielectric fluid or refrigerant coolant)	2–5
Communicate findings between standard-making entities (NIST) and industry.	Proper communication between NIST and industry	White paper/guideline published by NIST on TIM design and application	2–5
Develop technology and algorithm for interface mapping (interface thermal resistance).	Accuracy, resolution, throughput, and compatibility with i*n-situ* testing	Develop technology for mapping thermal interfaces that integrates with simulation tools, enabling high-throughput design of thermal management solutions	2–5
Develop standard set of tests to use (inter-agency NIST, DOE, ASHRAE, etc.).	--	Testing including over-time performance to evaluate reliability of TIM	3–6
Overall Goal: Design material with optimal properties. Conduct gap analysis for engineered materials and composites.	Better TIM with improved system-level thermal performance	Improve system-level thermal resistance through the development of technology and simulation tools that enhance interface performance and throughput	5–10

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Encourage to do more research via, e.g., government providing findings or potential tax benefit.
End Users/OEMs	• Provide wish list for DOE or assigned national laboratory.
Academia	• Identify novel thermal materials suitable for coating on chips and packages to serve as heat sinks or thermal redistributors. • Find better and easier equivalent or alternative testing method for HI.

National Laboratories	• Provide examples or standards for testing routine and equipment. • Provide testing capabilities for researchers and suppliers.

Required Resources	Cross Collaboration Needs of Working Groups
• Government funding for material design innovation and manufacturing upscaling. • Access to HPC server application testing facilities. • Reliability, energy efficiency, and ease-of-use (modularity, sustainability) standards. • Funding for fundamental research and tools on understanding/improving interfacial thermal resistance and on low-energy, scalable manufacturing methods. • Measuring and testing equipment and tools for understanding interface performance, reliability, energy efficiency, and ease of use. • Shared findings, best practices, and handbook.	• Power and Architecture: Consider thermal management in designs; have better thermal understanding when running simulations. • Materials and Devices: Define requirements that materials must meet to be selected for use as TIMs at different layer levels. • Education and Workforce Development: Develop updated tools and comprehensive education for industry. Universities should offer cross-disciplinary training in materials science, heat transfer, mechanical engineering, and simulation to better prepare system engineers for industry challenges. Enhancements may include developing cross-disciplinary curricula, establishing endowed chairs in APHI, and creating university/lab centers.

2.3.6 Packaging Electronic Design Automation/Process Design Kits/Assembly Design Kits

Historically, packaging design focused on providing mechanical stability and facilitating power distribution and data transfer through features like ball grid arrays on printed circuit boards and redistribution layers (RDLs). With the advent of more complex devices, however, the shift toward 2.5D components using embedded multi-die interconnect bridge (EMIB), interposer technologies, and 3D integration necessitates advanced EDA tools. These tools are essential for managing increased interconnect density, refining design protocols, and enhancing simulations to keep pace with evolving packaging technologies (Acito 2019; de Geus 2023).

Today, the chip industry navigates between two distinct design paradigms. The first encompasses traditional SoCs connected to PCBs. The second focuses on shrinking interconnects near the IC scale for multi-stacked chips, such as flip chip and hybrid bonding used in High Bandwidth Memory (HBM), and for stacked systems like DRAM Cache, both of which necessitate innovative packaging techniques. The use of diverse IP blocks, defined as reusable units of logic or data such as microprocessors or memory arrays, has led to a proliferation of design capabilities, complicating the design and simulation processes. Additionally, multi-domain components such as analog, digital, RF, and photonics increase design complexity. The scope for simulation now extends to multi-physics problems that involve electrical, mechanical, thermal, optical, and acoustic properties. Consequently, issues such as heat and electrical crosstalk, which have become more prevalent in design layouts, modeling, and device performance simulations, must be thoroughly addressed. EDA software must continually evolve to accommodate these increasing complexities in design and simulation.

Because EDA is a software-based tool, it does not directly affect the overall energy usage of microelectronics. However, it does help reduce microelectronics' energy consumption through advanced design and simulation, which creates effective, energy-efficient devices with improved performance. Beyond this, EDA reduces costs through prototype failure-mode analysis.

Challenges and Solution Pathways for Packaging Electronic Design Automation/Process Design Kits/Assembly Design Kits

System Technology Co-Optimization

System technology co-cptimization (STCO) is an approach that extends design technology co-optimization (DTCO) principles to the system level. It deconstructs the conventional SoC into distinct functional components—such as I/O, cache memory, and HBM—and optimizes their integration on a silicon die, typically using chiplets (Moore 2023; Siemens 2021). This method facilitates collaboration between design and manufacturing teams to refine IP block functionalities, communication protocols, and interconnections. Similar to DTCO, where circuit design and process teams work together to optimize device elements like transistors, STCO leverages chiplet-based designs or standardized interconnect I/O layouts. This approach enables the assembly of advanced circuitry that delivers enhanced performance and reduced power consumption.

There are two main benefits of using the STCO technique. First, it allows for improved design through standardization, and second, it facilitates early analysis to discover related issues in the design phase rather than in prototype hardware. Simplifying the components and standard interfaces (I/O interconnects) should enable development of an early package prototype with enough information for initial performance simulations, especially when combined with EDA software assembly design kits (ADK) that provide design rules for chiplet devices with associated simulation parameters (Siemens 2021; Heinig and Fischbach 2015). With the aid of simulations, the design and manufacturing teams can work to improve all components simultaneously for optimal performance and energy efficiency. As a result, STCO will enable more energy efficient 3D stacked architectures.

Standards for Package Assembly Design Kits With Vendor Support

Process design kits (PDKs) allow for continued improvements in the performance of device architecture through collaboration with foundries and designers. Current 3D design software is limited to specific packaging types requiring high-effort, user-specific scripting to enable advanced layouts (Heinig and Fischbach 2015). Without specific information about the design and manufacturing of components, the package designer cannot adequately simulate the electrical, thermal, or mechanical behavior of the package. To help enable package and device simulations, an ADK is needed.

ADKs contain design rules and simulation information analogous to a PDK. They contain the manufacturing steps, including the assembly technologies used (e.g., copper wires), the materials properties relevant to simulations (such as mechanical, thermal, and electrical); the geometrical information for interfaces such as input/output for die and substrate technologies; and, lastly, design rules such as component clearance, interconnect sizes, and pitches. With all this information, along with EDA and STCO, package designers and manufacturers can complete initial prototype design and simulations for the next generation of packaging technologies. This information should also enable 3D chip stacking and simulate the relevant challenges. Importantly, standards should be set for assembly design kits, such as utilizing chiplet technology for interconnect schemes, that can be bought or shared with vendors. This standardization is particularly important for managing thermal challenges associated with device stacking (Kao, Kuo, and Dai 2016).

Enhancing Simulation Algorithms To Reduce Computational Demands

As the industry adopts a "shift left" approach—moving testing and verification earlier in the development process—through STCO, there is an escalating need for sophisticated simulations that can accurately model the multi-physics aspects of devices and packaging, including heat, power, and energy consumption. While traditional CPU and GPU cores generally suffice for these tasks, there is room for optimization. Recognizing the specific requirements for multi-physics simulations, the development of specialized algorithms and dedicated hardware architectures—such as CIM or DSA—could drastically reduce both simulation times and energy costs. This targeted approach would enable faster development cycles for prototypes, yielding devices with markedly better performance and energy efficiency compared to existing solutions.

Artificial-Intelligence-Driven System-Level Optimization

System-level optimization in packaging design involves multiple teams, including device manufacturers, designers, IC layout teams, and ADK providers. Leveraging AI in this process can streamline the design, enhancing both performance and energy efficiency. AI has already proven effective in EDA for optimizing complex architectures, such as next-generation processors (de Geus 2023). By implementing AI, the design time for intricate projects, like GPUs, has been dramatically reduced—from the traditional months-long process requiring extensive engineering resources to significantly shorter periods (Hilson 2023). Utilizing AI specifically tailored for package design could accelerate prototype development, minimizing both time and resource expenditure, and fostering quicker iterations and enhancements in package solutions.

Action Plan for Packaging Electrical Design Automation/Process Design Kits/Assembly Design Kits

Table 55. Action Plan for Packaging Electrical Design Automation/Process Design Kits/Assembly Design Kits

Scope		
Technology for Energy Efficiency	Electronic Design Automation for Packaging.	
Technologies of Interest:	• Chip-stacking, 2.5D and 3D technology. Efficient handling of system complexity. • Multi-domain floor planning (digital, analog, RF, and photonic ICs) • Co-design/simulation and verification of different domains from different foundries • Mixed-signal (digital/analog/optical) functional verification • Package design: mechanical, PCB	
Challenges Addressed		**Solution Pathways**

• 3DHI challenge: Moving from D(esign)TCO to S(ystem)TCO • Standard pin map options • EDA algorithms for simulation of the full system • AI utilization at the system level	• Optimize STCO to include component model, design full system, and handle interface where packaging meets chips. Designs are first optimized with individual DTCO flows for each 3DHI layer. Chips are integrated with packaging/interconnect layers using DTCO + packaging co-design. System is combined and optimized through the STCO flow to create the final optimized stack. • Develop and adopt standards for package assembly design kits that different vendors can support. Standard pin maps can give DC/AC models for each option. • Optimize simulation algorithms to reduce reliance on CPU and GPU cores, thereby decreasing power and memory usage, and incorporate neuromorphic computing and deep learning techniques to enhance problem-specific processing efficiency. • Implement AI-driven optimization across system levels to enhance cost-efficiency and performance, focusing on analog and digital design optimization, and design verification. Extend optimization efforts to encompass all domains, including packaging, to achieve comprehensive improvements in reliability, yield, thermal management, and signal integrity.

Major Tasks/Milestones	Metrics	Targets	Timeline
Technology definition for standardization	Size, height, technology for PCB/packaging	Library standardization for EDA	1–3 years
Demonstrate how STCO can be applied to a 3DHI phased array antenna to consume less energy	Size, weight, and power (SWAP)	10x improvement	3–5 years
AI-driven, system-level optimization	Cost and Efficiency	2x improvement (cost) and 10x improvement (efficiency)	5–7 years
Improving solution algorithm	Simulation time, memory, compute power, energy, thermal load	100x–1,000x improvement	8–10 years

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• EDA vendors • Define, develop the software
End Users/OEMs	• Help with standardization • Validate software for simulations • Provide feedback on new designs/design flow • Work with foundries
Academia	• Train students in system knowledge (3DHI, APHI, packaging) • Fund PhD research (new algorithms, designs, AI)
National Laboratories	• Develop HPC comparison, new algorithms, and AI development
Government	• Fund academic/national lab research

Required Resources	Cross Collaboration Needs of Working Groups
• Partnerships between EDA/partner manufacturers for continual development (loaning parts, paying for parts, funding for hardware evaluation). HPC resources for improvements in algorithm design for simulations. • Funding for initial development of AI, algorithms, and software. • HPC center for algorithm innovation (national lab). • Dedicated centralized resource at national lab or center for R&D. Ease of access with IP protection (CRADA).	• Algorithms and Software: Collaborate on minimizing energy use for EDA simulation. • Circuits and Architectures: Develop EDA solutions. Circuits and Architectures solutions will influence APHI solutions (e.g., PDK will have an impact on assembly design kit). • Education and Workforce Development: Implement EDA tool training. Partner validation and verification.

2.3.7 Conclusion for Advanced Packaging and Heterogeneous Integration

The Advanced Packaging and Heterogeneous Integration (APHI) chapter of the EES2 roadmap emphasizes the crucial role of advanced packaging technologies in enhancing energy efficiency across the semiconductor industry. Vertically integrated devices and system-level cooling strategies represent key areas where significant advancements can lead to major energy savings. By employing energy-efficient 3D technology and optimizing the thermal interface materials, APHI aims to manage heat more effectively, thus reducing the thermal challenges associated with dense packing of high-performance chips.

Innovation in APHI is targeted towards solving scaling challenges for optical interconnects, enhancing intra-chip signal integrity, and increasing the energy efficiency of memory access. The deployment of these technologies demands rigorous EDA improvements to support new ADKs, facilitating a streamlined pathway from design to simulation and, ultimately, to manufacturing.

To meet energy-efficiency goals, EES2 emphasizes the need for accelerated development and integration of novel packaging solutions, such as the establishment of dedicated R&D facilities that allow for rapid prototyping and early-stage testing of APHI technologies. Such initiatives are vital for overcoming current barriers in thermal management, material integration, and system-level integration, ensuring that advanced packaging can keep pace with the evolving demands of modern computing environments.

Overall, advancing APHI technologies is about not just enhancing individual components, but also ensuring a synergistic integration that maximizes overall system performance and energy efficiency. The roadmap sets a clear directive for industry-wide collaboration, standardized practices, and focused R&D efforts to rapidly bring these critical technologies to market readiness, aligning with the urgent needs for sustainable energy management in the semiconductor sector.

2.3.8 Advanced Packaging and Heterogenous Integration References

Acito, Bill. 2019. “Leveraging the Best of Package and IC Design for System Enablement.” Presented at the 2019 International Wafer Level Packaging Conference (IWLPC). San Jose, CA. https://doi.org/10.23919/IWLPC.2019.8914109.

Alam, Tafseer, Rohit Dhiman, Rajeevan Chandel, and Dhrub Solanki. 2011. “Mixed carbon nanotube bundle: Capacitance analysis and comparison with copper interconnect.” Presented at the 2011 International Conference on Emerging Trends in Electrical and Computer Technology. Nagercoil, India. https://doi.org/10.1109/ICETECT.2011.5760207.

Albright, Jessica. 2022. “Hybrid Bonding Basics: What Is Hybrid Bonding?” Semiconductor Engineering. Accessed September 28, 2023. https://semiengineering.com/hybrid-bonding-basics-what-is-hybrid-bonding/.

AMD. 2023. “AMD 3D V-Cache™ Technology.” Undated. Accessed October 4, 2023. https://www.amd.com/en/technologies/3d-v-cache.

Amin, Rubab, Can Suer, Zhizhen Ma, Ibrahim Sarpkaya, Jacob B. Khurgin, Ritesh Agarwal, and Volker J. Sorger. 2018. “Active material, optical mode and cavity impact on nanoscale electro-optic modulation performance.” *Nanophotonics*. Vol. 7 (Issue 2): pg 455–472. http://dx.doi.org/10.1515/nanoph-2017-0072.

Angelini, Chris. 2020. “Intel Learns Hard on Advanced Chip Packaging Technology in Battle for Computing Supremacy.” Venture Beat. Accessed October 15, 2023. https://venturebeat.com/business/intel-leans-hard-on-advanced-packaging-technologies-in-battle-for-computing-supremacy/.

Bakir, Muhannad S. 2022. “Bridge-chip Interconnect Technologies.” IEEE Electronics Packaging Society. Accessed October 14, 2023. https://eps.ieee.org/publications/enews/march-2022/846-bridge-chip-interconnect-technologies.html.

Broadcom. 2023. “Co-Packaged Optics.” Undated. Accessed November 22, 2023. https://www.broadcom.com/info/optics/cpo#advantages.

Burkacky, O., T. Kim, and I. Yeom. 2023. “Advanced chip packaging: How manufacturers can play to win.” McKinsey & Company. Accessed October 16, 2023. https://www.mckinsey.com/industries/semiconductors/our-insights/advanced-chip-packaging-how-manufacturers-can-play-to-win.

De Geus, Aart. 2023. “How Quickly Will Multi-Die Systems Change Semiconductor Design?” Synopsys. Undated. Accessed November 2023. https://www.synopsys.com/multi-die-system/how-multi-die-systems-will-change-semiconductor-design.html.

Dhananjay, K., P. Shukla, V.F. Pavlidis, A. Coskun, and E. Salman. 2021. “Monolithic 3D Integrated Circuits: Recent Trends and Future Prospects.” *IEEE Transactions on Circuits and Systems II: Express Briefs.* Vol. 68 (Issue 3): pg 837–843. https://doi.org/10.1109/TCSII.2021.3051250.

GlobalFoundries. 2022. “GlobalFoundries Announces Next Generation in Silicon Photonics Solutions and Collaborates with Industry Leaders to Advance a New Era of More in the Data Center.” Published March 7, 2022. https://gf.com/gf-press-release/globalfoundries-announces-next-generation-silicon-photonics-solutions-and/.

Green, Craig, Leonardo Prinzi, and Baratunde A. Cola. 2016. “Design and evaluation of polymer-carbon nanotube composites for reliable, low resistance, static and dynamic thermal interface materials.” Presented at the 2016 15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (Itherm). Las Vegas, NV. https://doi.org/10.1109/ITHERM.2016.7517611.

Guo, Xiaoxiao, Shujian Cheng, Weiwei Cai, Yufeng Zhang, and Xue-ao Zhang. 2021.”A review of carbon-based thermal interface materials: Mechanism, thermal measurements and thermal properties.” *Materials & Design*. Vol. 209: 109936. https://doi.org/10.1016/j.matdes.2021.109936.

Heinig, Andy, and Robert Fischbach. 2015. “Enabling automatic system design optimization through Assembly Design Kits.” Presented at the 2015 International 3D Systems Integration Conference (3DIC). Sendai, Japan. https://doi.org/10.1109/3DIC.2015.7334602.

Hiebert, Stephen. 2023. “Hybrid Bonding takes Heterogeneous Integration to the Next Level.” 3D InCites. Accessed September 28, 2023. https://www.3dincites.com/2023/05/hybrid-bonding-takes-heterogeneous-integration-to-the-next-level/.

Hilson, Gary. 2023. “AI Can’t Design Chips Without People.” EETimes. Published July 3, 2023. https://www.eetimes.com/ai-cant-design-chips-without-people/.

IEEE HIR. 2021. “Heterogeneous Integration Roadmap, Chapter 20: Thermal.” In *Heterogeneous Integration Roadmap: 2021 Edition*. Institute of Electrical and Electronic Engineers (IEEE). https://eps.ieee.org/images/files/HIR_2021/ch20_thermal1.pdf.

IEEE HIR. 2023. “Heterogeneous Integration Roadmap, Chapter 20: Thermal.” In *Heterogeneous Integration Roadmap: 2023 Edition*. IEEE. https://eps.ieee.org/images/files/HIR_2021/ch20_thermal1.pdf.

IEEE IRDS. 2023. “IEEE International Roadmap for Devices and Systems.”IEEE. Undated. Accessed November 11, 2023. https://irds.ieee.org/.

Jani, Imed. 2019. “Test and characterization of 3D high-density interconnects.” Thesis in Micro and nanotechnologies/Microelectronics, Université Grenoble Alpes, 2019. https://theses.hal.science/tel-02634259/.

Jensen, Timothy, and Ronald Lasky. 2020. “The Basics of Metal Thermal Interface Materials (TIMs).” Presented at the 2020 Pan Pacific Microelectronics Symposium. Hawaii. https://doi.org/10.23919/PanPacific48324.2020.9059395.

Jo, P.K., X. Zhang, J.L. Gonzalez, G.S. May, and M.S. Bakir. 2018. “Heterogeneous Multi-Die Stitching Enabled by Fine-Pitch and Multi-Height Compressible Microinterconnects (CMIs).” *IEEE Transactions on Electron Devices.* Vol. 65 (Issue 7): pg 2957–2963. https://doi.org/10.1109/TED.2018.2838529.

Jouppi, Norman P., et al. 2021. “Ten Lessons from Three Generations Shaped Google’s TPUv4i : Industrial Product.” Presented at the 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA). Valencia, Spain. https://doi.org/10.1109/ISCA52012.2021.00010.

Kao, C.T., A.Y. Kuo, and Y. Dai. 2016. “Electrical/thermal co-design and co-simulation, from chip, package, board to system.” Presented at the 2016 International Symposium on VLSI Design, Automation and Test (VLSI-DAT). Hsinchu, Taiwan. https://doi.org/10.1109/VLSI-DAT.2016.7482540.

Karkar, Ammar, Terrence Mak, Kin-Fai Tong, and Alex Yakovlev. 2016. “A Survey of Emerging Interconnects for On-Chip Efficient Multicast and Broadcast in Many-Cores.” *IEEE Circuits and Systems Magazine*. Vol. 16 (Issue 1): pg 58–72. https://doi.org/10.1109/MCAS.2015.2510199.

Keser, B., and S. Kroehnert. 2019. “Embedded Multi-die Interconnect Bridge (EMIB).” In *Advances in Embedded and Fan-Out Wafer Level Packaging Technologies*, 487–499. Wiley-IEEE Press. https://ieeexplore.ieee.org/book/8726249.

Krishnamoorthy, A.V, and D.A.B. Miller. 1996. “Scaling Optoelectronic-VLSI Circuits into the 21st Century: A Technology Roadmap.” *IEEE Journal of Selected Topics in Quantum Electronics*. Vol. 2 (Issue 1): pg 55–76. https://doi.org/10.1109/2944.541875.

Krishnamoorthy, A.V., H. Schwetman, X. Zheng, and R. Ho. 2015. “Energy-Efficient Photonics in Future High-Connectivity Computing Systems.” *Journal of Lightwave Technology*. Vol. 33 (Issue 4): pg 889–900. https://doi.org/10.1109/JLT.2015.2395453.

Kumar, Vachan, and Azad Naeemi. 2017. “An overview of 3D integrated circuits.” Presented at the 2017 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization for RF, Microwave, and Terahertz Applications (NEMO). Seville, Spain. https://doi.org/10.1109/NEMO.2017.7964270.

Lapedus, Mark. 2018. “Bridges Vs. Interposers.” Semiconductor Engineering. Accessed October 14, 2023. https ://semiengineering.com/using-silicon-bridges-in-packages/.

Li, G., et al. 2013. “Ring resonator modulators in silicon for interchip photonic links.” *IEEE Journal of Selected Topics in Quantum Electronics*. Vol. 19 (Issue 6). https://doi.org/10.1109/JSTQE.2013.2278885.

Li, N., et al. 2022. “Integrated lasers on silicon at communication wavelengths: A progress review.” *Advanced Optical Materials*. Vol. 10 (Issue 23). https://doi.org/10.1002/adom.202201008.

Liu, S., X. Wu, D. Jung, J.C. Norman, M.J. Kennedy, H.K. Tsang, A.C. Gossard, and J.E. Bowers. 2019. “High-channel-count 20 GHz passively mode-locked quantum dot laser directly grown on Si with 41 Tbit/s transmission capacity.” *Optica*. Vol. 6 (Issue 2): pg 128–134. https://doi.org/10.1364/OPTICA.6.000128.

Liu, Y., et al. 2020. “The Design of CMOS-Compatible Plasmonic Waveguides for Intra-Chip Communication.” *IEEE Photonics Journal*. Vol. 12 (Issue 5, Article no. 4800810): pg 1–10. https://doi.org/10.1109/JPHOT.2020.3024119.

Mahajan, R., et al. 2016. “Embedded Multi-die Interconnect Bridge (EMIB) – A High Density, High Bandwidth Packaging Interconnect.” Presented at the 2016 IEEE 66th Electronic Components and Technology Conference (ECTC). Las Vegas, NV. https://doi.org/10.1109/ECTC.2016.201.

Mekawey, H., M. Elsayed, Y. Ismail, and M.A. Swillam. 2022. “Optical Interconnects Finally Seeing the Light in Silicon Photonics: Past the Hype.” *Nanomaterials*. Vol. 12 (Issue 3): 485. https://doi.org/10.3390/nano12030485.

Miller, D.A.B. 1989. “Optics for low-energy communication inside digital processors: quantum detectors, sources, and modulators as efficient impedance converters.” *Optics Letters*. Vol. 14 (Issue 2): pg 146–148. https://doi.org/10.1364/OL.14.000146.

Miller, D.A.B. 2000. “Communicating with Waves Between Volumes – Evaluating Orthogonal Spatial Channels and Limits on Coupling Strengths.” *Applied Optics.* Vol. 39: pg 1681–1699. https://doi.org/10.1364/AO.39.001681.

Miller, D.A.B. 2017. “Attojoule Optoelectronics for Low-Energy Information Processing and Communications.” *Journal of Lightwave Technology*. Vol. 35 (Issue 3): pg 346–396. https://doi.org/10.1109/JLT.2017.2647779.

Miscuglio, Mario, Zibo Hu, Shurui Li, Jonathan K. George, Roberto Capanna, Hamed Dalir, Philippe M. Bardet, Puneet Gupta, and Volker J. Sorger. 2020. “Massively parallel amplitude-only Fourier neural network.” *Optica.* Vol. 7: pg 1812–1819. https://doi.org/10.1364/OPTICA.408659.

Mittal, Jagjiwan, and K.L. Lin. 2017. “Carbon nanotube-based interconnections.” *J Mater Sci.* Vol. 52: pg 643–662. https://doi.org/10.1007/s10853-016-0416-4.

Moore, Samuel K. 2023. “Keeping Moore’s Law Going Is Getting Complicated: CMOS 2.0 will require exceptional creativity to rewire and 3D-stack the chip.” IEEE Spectrum. Published May 24, 2023. https://spectrum.ieee.org/stco-system-technology-cooptimization.

Murphy, Mike. 2023. “IBM Research unveils hybrid bonding for packaging chips.” IBM. Accessed June 28, 2023. https://research.ibm.com/blog/hybrid-bonding-chip-packaging-chiplets.

Norman, J.C., et al. 2019. “A review of high-performance quantum dot lasers on silicon.” *IEEE Journal of Quantum Electronics*. Vol. 55 (Issue 2). https://doi.org/10.1109/JQE.2019.2901508.

O’Connor, Mike, Niladrish Chatterjee, Donghyuk Lee, John Wilson, Aditya Agrawal, Stephen W. Keckler, and William J. Dally. 2017. “Fine-grained DRAM: energy-efficient DRAM for extreme bandwidth systems.” *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture* (MICRO-50 ’17). Pg 41–54. New York: Association for Computing Machinery. https://doi.org/10.1145/3123939.3124545.

Ong, Jia-Juen, Wei-Lan Chiu, Ou-Hsiang Lee, Chia-Wen Chiang, Hsiang-Hung Chang, Chin-Hung Wang, Kai-Cheng Shie, et al. 2022. “Low-Temperature Cu/SiO_2 Hybrid Bonding with Low Contact Resistance Using (111)-Oriented Cu Surfaces.” *Materials*. Vol. 15 (Issue 5): 1888. https://doi.org/10.3390/ma15051888.

Ramos, Raphael, A. Fournier, Murielle Fayolle, Jean Dijon, C.P. Murray, and J. McKenna. 2016. “Nanocarbon interconnects combining vertical CNT interconnects and horizontal graphene lines.” Presented at the 2016 IEEE International Interconnect Technology Conference/Advanced Metallization Conference (IITC/AMC). San Jose, CA. https://doi.org/10.1109/IITC-AMC.2016.7507676.

Refai-Ahmed, Gamal, Hao Do, Brian Philofsky, and Jason Strader. 2018. “Extending the performance of high heat flux 2.5D and 3D packaging from component-system interaction.” Presented at the 19th International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE). Toulouse, France. https://doi.org/10.1109/EuroSimE.2018.8369893.

Salama, Islam. 2023. “Interposer and Packaging Device Architecture and Method of Making For Integrated Circuits.” United States Patent Application 20230298964. https://patents.justia.com/patent/20230298964.

Salama, Islam, Nathaniel Quick, Arivinda Kar, and Gilyong Chung. 2002. “Electrical Characterization of Laser-Irradiated 4H-SiC Wafer.” *Materials Research Society Symposium – Proceedings*. Vol. 719: pg 73–78. https://doi.org/10.1557/PROC-719-F3.2.

Salman, Emre. 2023. “Energy Inefficiencies of Data Movement and Benefits of Monolithic Inter-tier Vias for Datacentric Applications.” Presented at the EES2 April 2023 Meeting.

Samal, S.K., D. Nayak, M. Ichihashi, S. Banna, and S.K. Lim. 2016. “Monolithic 3D IC vs. TSV-based 3D IC in 14 nm FinFET Technology.” Presented at the 2016 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S). Burlingame, CA. https://doi.org/10.1109/S3S.2016.7804405.

Sarvey, Thomas, et al. 2015. “Embedded cooling technologies for densely integrated electronic systems.” Presented at the 2015 IEEE Custom Integrated Circuits Conference (CICC). San Jose, CA. https://doi.org/10.1109/CICC.2015.7338365.

Shang, C., et al. 2022. “Electrically pumped quantum-dot lasers grown on 300mm patterned Si photonic wafers.” *Nature*. Vol. 11 (Article no. 299). http://dx.doi.org/10.48550/arXiv.2206.01211.

SIA. 2022. *American Semiconductor Research: Leadership Through Innovation.* Semiconductor Industry Association (SIA). https://www.semiconductors.org/wp-content/uploads/2022/11/American-Semiconductor-Research-Report-FINAL1.pdf.

Siemens. 2021. "Using a System Technology Co-Optimization (STCO) Approach for 2.5/3D Heterogeneous Semiconductor Integration." White paper. https://static.sw.cdn.siemens.com/siemens-disw-assets/public/15hO367mMjEXqQxw385h2O/en-US/Siemens-SW-Using-a-system-technology-co-optimization-(STCO)-approach-for-2.53D-WP-83719-C1.pdf.

Smith, Ryan. 2020. "Micron Spills on GDDR6X: PAM4 Signaling For Higher Rates, Coming to NVIDIA's RTX 3090." AnandTech. Published August 20, 2020. https://www.anandtech.com/show/15978/micron-spills-on-gddr6x-pam4-signaling-for-higher-rates-coming-to-nvidias-rtx-3090.

Soldano, Caterina, Saikat Talapatra, and Swastik Kar. 2013. "Carbon Nanotubes and Graphene Nanoribbons: Potentials for Nanoscale Electrical Interconnects." *Electronics.* Vol. 2 (Issue 3): pg 280–314. https://doi.org/10.3390/electronics2030280.

Sorger, Volker, et al. 2015. "Nano-optics gets practical." *Nature Nanotechnology.* Vol. 10: pg 11–15. https://doi.org/10.1038/nnano.2014.314.

Sorger, Volker. 2023. "Can Photonic & Analog Computing Paradigms enable 1000x Power (Performance?) Savings?" Presented at the EES2 Working Group Meeting, August 2023. https://ees2.slac.stanford.edu/doe-meetings-events/doe-ees2-roadmap-meeting-9.

SRC. 2023. "Advanced Packaging and Heterogeneous Integration." In *Microelectronics and Advanced Packaging Technologies (MAPT) Roadmap – Interim Report*. Semiconductor Research Corporation (SRC). https://srcmapt.org/chapter9/.

Srinivasan, S.A., et al. 2019. "High Absorption Contrast Quantum Confined Stark Effect in Ultra-Thin Ge/SiGe Quantum Well Stacks Grown on Si." *IEEE Journal of Quantum Electronics*. Vol. 56 : 5200207. https://doi.org/10.1109/JQE.2019.2949640.

Stojanovic, Vladimir. 2020. "Super-fast optical interconnects." *Compound Semiconductor Magazine.* Vol. 26 (Issue 8). https://compoundsemiconductor.net/article/112506/Super-fast_Optical_Interconnects/feature.

Tauke-Pedretti, Anna. 2023. "Photonics in the Package for Extreme Scalability (PIPES)." Defense Advanced Research Projects Agency (DARPA). Undated. Accessed November 9, 2023. https://www.darpa.mil/program/photonics-in-the-package-for-extreme-scalability.

Todri-Sanial, Aida, Jean Dijon, and Antonio Maffucci. 2017. *Carbon Nanotubes for Interconnects*. Ebook. Cham, Switzerland: Springer Cham. https://link.springer.com/book/10.1007/978-3-319-29746-0.

UMC. 2023. "UMC and Cadence Collaborate on 3D-IC Hybrid Bonding Reference Flow." Accessed June 6, 2023. https://www.umc.com/en/News/press_release/Content/technology_related/20230201.

Vogelsang, Thomas. 2010. "Understanding the Energy Consumption of Dynamic Random Access Memories." Presented at the *2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture*. Atlanta, GA. https://doi.org/10.1109/MICRO.2010.42.

Wang, H., B.M. Nouri, H. Dalir, and V.J. Sorger. 2023. "30 GHz plasmonic slot $MoTe_2$ photodetector integrated with silicon photonic circuits at telecom wavelength." *Ultrafast Phenomena and Nanophotonics XXVII*. Vol. 12419. https://doi.org/10.1117/12.2651013.

Wang, H., V.J. Sorger, and H. Dalir. 2022. "Plasmonic Slot Waveguide – Integrated $MoTe_2$ Photodetector with 30-GHz Bandwidth at Telecom Wavelength." Presented at the 2022 IEEE Photonics Conference (IPC). Vancouver, BC, Canada. https://doi.org/10.1109/IPC53466.2022.9975520.

Wang, N.C., S. Sinha, B. Cline, C.D. English, G. Yeric, and E. Pop. 2017. "Replacing copper interconnects with graphene at a 7-nm node." Presented at the *2017 IEEE International Interconnect Technology Conference (IITC).* Hsinchu, Taiwan. https ://doi.org/10.1109/IITC-AMC.2017.7968949.

Wen, P., P. Tiwari, S. Mauthe, et al. 2022. "Waveguide coupled III-V photodiodes monolithically integrated on Si." *Nature Communications.* Vol. 13 (Article no. 909). https://doi.org/10.1038/s41467-022-28502-6.

Weste, Neil, and David Harris. 2011. *CMOS VLSI Design: A Circuits and Systems Perspective.* Boston: Pearson Education, Inc.

WikiChip. 2023. "Foveros – Intel." Undated. Accessed October 14, 2023. https://en.wikichip.org/wiki/intel/foveros.

Winzer, P.J., and D.T. Neilson. 2017. "From Scaling Disparities to Integrated Parallelism: A Decathlon for a Decade." *Journal of Lightwave Technology.* Vol. 35 (Issue 5): pg 1099–1115. https://doi.org/10.1109/JLT.2017.2662082.

Wooten, E.L., et al. 2000. "A review of lithium niobate modulators for fiber-optic communications." *IEEE Journal of Selected Topics in Quantum Electronics*. Vol. 6 (Issue 1): pg 69–82. https://doi.org/10.1109/2944.826874.

Wu, Jiang, Mingchu Tang, and Huiyun Liu. 2019. "Chapter 2 – III-V quantum dot lasers epitaxially grown on Si substrates." In *Nanoscale Semiconductor Lasers, Micro and Nano Technologies series*, 17–39. http://dx.doi.org/10.1016/B978-0-12-814162-5.00002-9.

Wu, R., et al. 2017. "Compact modeling and circuit-level simulation of silicon nanophotonic interconnects." Presented at the Design, Automation & Test in Europe (DATE) Conference & Exhibition. Lausanne, Switzerland. https://doi.org/10.23919/DATE.2017.7927057.

Xu, Q., B. Schmidt, S. Pradhan, et al. 2005. "Micrometre-scale silicon electro-optic modulator." *Nature.* Vol. 435: pg 325–327. https://doi.org/10.1038/nature03569.

Ye, C., S. Khan, Z.R. Li, E. Simsek, and V.J. Sorger. 2014. "λ-Size ITO and Graphene-Based Electro-Optic Modulators on SOI." *IEEE Journal of Selected Topics in Quantum Electronics*. Vol. 20 (Issue 4, Article no. 3400310): pg 40–49. https://doi.org/10.1109/JSTQE.2014.2298451.

Zhang, Yang, Thomas E. Sarvey, and Muhannad S. Bakir. 2014. "Thermal challenges for heterogeneous 3D Ics and opportunities for air gap thermal isolation." Presented at the *2014 International 3D Systems Integration Conference (3DIC)*. Kinsdale, Ireland. https://doi.org/10.1109/3DIC.2014.7152174.

Zhang, Y., X. Zhang, and M.S. Bakir. 2018. “Benchmarking Digital Die-to-Die Channels in 2.5-D and 3-D Heterogeneous Integration Platforms.” *IEEE Transactions on Electron Devices.* Vol. 65 (Issue 12). https://doi.org/10.1109/TED.2018.2876688.

2.4 Algorithms and Software

Significant opportunities exist for reducing energy consumption in computing through improved algorithms and software. As shown in Figure 6 in the Introduction, there are about 20 orders of magnitude in computational energy consumption between representative large application programs and the individual instructions being executed (Shankar 2023). This domain bridges architectural designs and the software that maps to them. Energy improvements in software must come from an understanding of what the software does and how, as well as finding means to accomplish software tasks more efficiently, either purely through improved algorithms or through improvements in both the underlying machine architecture and the algorithms implementing problem solutions on that architecture. The large-scale applications benchmarked in Figure 41 are from distinct problem domains, and algorithmic improvements that reduce the energy cost of training of large-language models may have little or no impact on the energy used for spike protein simulation, and vice versa. This chapter is divided into four sections after a brief summary of the key aspects of Algorithms and Software and the working group that contributed to the key aspects: 1) energy efficiency in algorithms, mainly as applicable to machine learning; 2) software for general purpose architectures (e.g., CPU and GPU); 3) software for special purpose architectures (e.g., ASIC); and 4) measurements, tools, and benchmarking to enable energy efficiency.

Working group methodology

As described in Section 1.4, after an initial definition of candidate technologies for inclusion in the roadmap, members of the Algorithms and Software working group performed an initial estimate of the potential energy efficiency improvement factor of the various technologies and the timeline over which the estimated energy efficiency can be achieved. This assessment (with results shown in Figure 41), although subjective, provides general directions for a quick review. Specific points to be considered: 1) it is not possible to accurately quantify potential improvements for algorithms and software not yet implemented, and 2) the expected gains are more a curve than a point in time because real software is continually and incrementally refined over time. The diamonds in Figure 41 represent a collection of technologies that were expanded in later meetings. For ease of organizing this chapter, technologies have been grouped into topics as shown in Table 53 and are discussed in detail in the proceeding subsections.

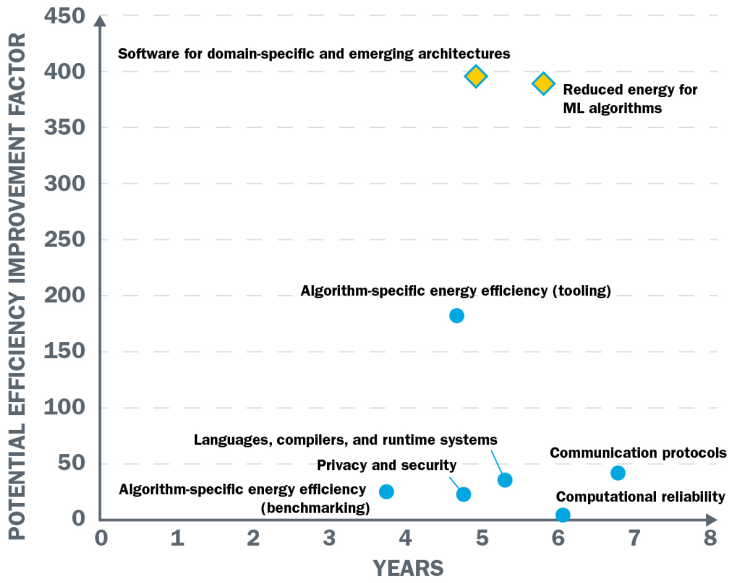


Figure 41. Algorithms and Software working group potential efficiency improvement factor and timeline initial assessment.

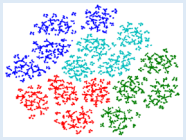
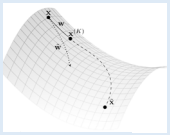
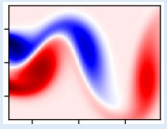
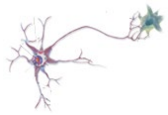
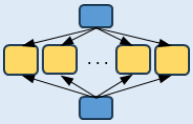
Table 53. Algorithms and Software Technology Grouping

Technology group	Technology
Algorithm-specific energy efficiency	Tooling
	Benchmarking
Algorithms for machine learning	Meta-learning of hyperparameter optimization
	Physics-informed machine learning models for scientific computing
	Continual learning (sequential training without catastrophic forgetting)
	Bottom-up sparse machine learning model development
	Benchmarking hyperparameter optimization methods
	Benchmarking and methodology to quantify training and inference costs
	Energy-efficient alternative training methods
	Approximate/efficient matrix/tensor multiplication
Software for conventional architectures	Languages compilers and run-time systems
	Privacy and security
	Computational reliability
	Communication protocols
	Data compression
	Precision of data types
Software for domain-specific and emerging architectures	Domain-specific languages
	Adoption of existing compute cores in domain-specific architectures
	Reusable memory access control architecture
	Compute-in-memory
	Tightly coupled architecture and software co-design

Table 54 summarizes the most significant identified energy efficiency opportunities that can be achieved through advances in algorithms and software.

Table 54. Key Takeaways for Energy Efficiency Opportunities in Algorithms and Software

Technology Group	Key Opportunities for Energy Efficiency

Improved efficiency in machine learning algorithms	   	• Continual learning, in which ML systems can build on knowledge gained without retraining from the ground up (incremental learning) analogously to biological systems. • Bottom-up sparse ML model development, in which the larger ML model is made up from smaller models that are trained for more narrow tasks and combined as a “mixture of experts” or hierarchical model of knowledge. • Meta-learning, or “learning to learn,” for optimization of model hyperparameters (such as number of nodes, number of layers, learning rate), which can greatly reduce the effort required to develop efficient models. • Physics-inspired ML models, in which the neural network model incorporates physics models, such as differential equations, and can be used to solve problems in applied mathematics with the potential to enhance or displace finite-element solvers, greatly accelerating numerical problem solutions.
Software for domain-specific and emerging architectures	 	• Domain-specific languages for conventional as well as new and emerging architectures, which can express a problem solution in high-level operators that are amenable to intermediate representations that can be better targeted for machine-level optimization. • In co-design with advanced architectures, exploiting compute-in-memory, data compression, and data types for more efficient brain-inspired representation.
Languages, compilers, and runtimes	  	• Languages such as Mojo that aim to replace interpreted Python with a source-code-compatible, incrementally compiled alternative. • Better automatic code optimization to exploit machine parallelism and maximize the speed and energy benefit of cache memory. • Application of machine learning to code writing and code optimization.

Grand challenges are to:

- Optimize energy efficiency of algorithms by improving use of parallel resources and minimizing data movement.
- Improve advanced profiling tools and benchmarking to measure software's energy impact.
- Integrate new hardware architectures into existing systems and their codebases within commercially tolerable compatibility constraints and continue to measure and benchmark energy estimates.
- Reduce the energy consumption of machine learning algorithms with new strategies in training and inference stages.
- Advance fundamental understanding of intelligence and learning to realize the transformative potential of machine intelligence. Machine learning systems are still far from the observed performance of learning in humans and other animals.

2.4.1 Algorithm-Specific Energy Efficiency Tooling and Benchmarks

A comprehensive capability for systematic profiling, enabling the performance and energy impact of software to be accurately measured, is a prerequisite for achieving the aim of benchmarking energy performance for specific algorithms in a wide variety of computing systems and environments.

2.4.1.1 Tooling

For traditional architectures, there are three components to energy use: movement of instructions (program) to the CPU, executing instructions in the execution units, and reading (loading) data from and writing (storing) data back to memory or external devices such as network and storage controllers. If power is measured during the execution of the program, the measurement captures the energy used in all three components. However, if the temperature distribution is measured, it will represent different distributions of energy as the heat distribution is an effect caused by computing. Thus, to understand and analyze software for different workloads, a more detailed and component-level measurement is desired.

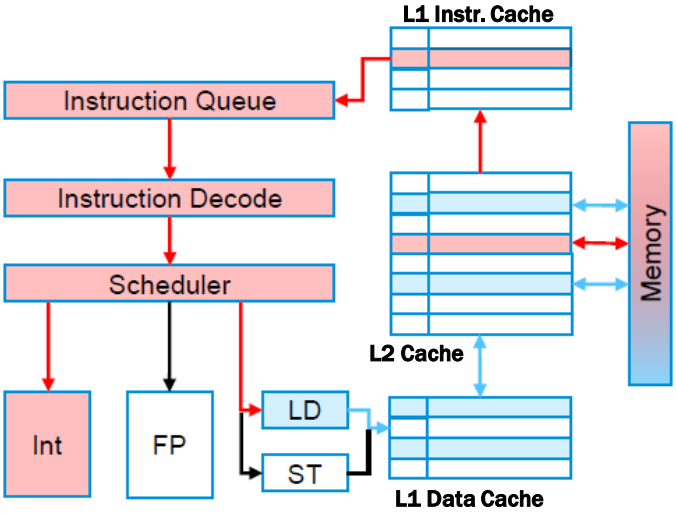
The current state of the art for measuring energy usage of algorithms is inadequate. For example, the Intel Runtime Average Power Level (RAPL) facility is normally used for thermal management of microprocessors but is only able to provide an aggregate estimate of power over the whole chip or a whole core at an interval of about 1 ms. Knowing the energy consumption at a more detailed level is desirable but faces some fundamental challenges. For example, consider the very simple case illustrated in Figure 42. In this example, the function `foo(*a,*b)` simply returns the sum of two variables whose addresses are passed to it. In the first call to `foo`, the local cache is "cold," and the processor must fetch the arguments

```
int foo (unsigned int *a, unsigned int *b)
{
      return *a + *b;
}
```

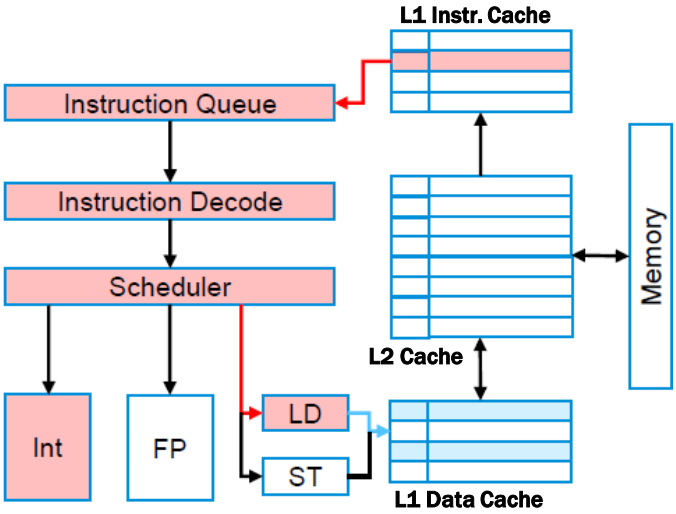
(a) Example C function foo(*a, *b)

```
; r1 - r4 input registers,; r0 return reg
_foo:
ld [%r1], r5         ; *a - > register 5
ld [%r2], r6         ; *b -> register 6
add %r5, %r6, %r0    ; sum in return reg.
ret
```

(b) Compiled assembly code for foo



(c) First call to foo



(d) Subsequent call to foo

Figure 42. Interaction of software and CPU architecture

as well as the function's instructions from the DRAM main memory. On the next call, if called with the same arguments, the processor finds both the instructions and data in the L1 cache, and no DRAM activity is triggered. For a 7nm process, the energy cost of DRAM access is 173x the cost of accessing L1 cache (see Figure 7 in the Introduction), and 43,333x the cost of an INT32 add, and the energy expended to call foo differs by at least 173x from one call to the next. Other complications may arise. For example, other functions executed between calls to `foo`, or other processes running on the same core, could claim all the available L1 cache space and cause foo's instructions and data to be evicted. The energy usage for code at this level of granularity is a complex interplay of the current workloads being executed by the system, the runtime environment, and details of the memory hierarchy implementation.

This example illustrates that software meant to probe the energy expended at a granular level must carefully consider not only the software being probed, but also other workload characteristics spanning the software and hardware present in the system at runtime.

Modern microprocessors provide subsystems that monitor many details of processor performance and events during operation, with the Intel Performance Management Unit (PMU) being a prime example (Intel 2022). These systems do not, however, provide adequate source-level traceability to facilitate the tooling needed for precise energy measurements. A selection of sophisticated architecture-specific profiling tools available include the Intel Vtune Profiler (Intel 2023) and Intel Processor Trace (Yagemann 2023), the AMD Research Instruction Based Sampling Toolkit (Greathouse 2021), IBM POWER9 Performance Monitor Unit (IBM 2018), various NVIDIA tools for GPU performance analysis (NVIDIA 2023), and some Linux-specific tools (Gregg 2023). Generally, these tools provide no facility or very limited capability to perform energy profiling.

There has been progress in addressing the need for energy profiling recently. Variorum (2023), which recently won an R&D 100 award, is an extensible vendor-neutral library for Linux that exposes power and performance monitoring and control of low-level hardware dials. Variorum's application programming interface (API) abstracts the details of the vendor-specific implementations and makes low-level machine performance dials available to both general and advanced users in a portable manner. In its internal implementation, Variorum uses different kernel interfaces, such as model-specific registers (MSRs) on Intel and AMD or NVML for NVIDIA, to expose the available dials on the platform. These dials allow for measurement and control of various physical features on processors and accelerators, such as power, energy, frequency, temperature, and performance counters.

Challenges and Solution Pathways for Algorithm-Specific Energy Efficiency Tooling

There is no efficient way to take memory access traces (e.g., cache misses) from software. Such a capability may be required to perform the detailed energy profiling needed. One tool which may serve as a starting point is the Intel Pin dynamic binary instrumentation framework for the IA-32, x86-64, and MIC instruction-set architectures (Intel 2023a. This framework performs measurements at run time on the compiled binary files and requires no recompiling of source code. At this juncture the granularity needed is not well-understood, and requirements may evolve with the software and architecture over the coming years.

Measurements can be supplemented by energy-aware compute simulation which provides open libraries for estimating energy at different levels from instruction to system-level. Such software

simulates a computing chip or subsystem running a workload, collecting energy use data. An example of this type of software has been developed for GPUs at Purdue University (Kandiah et al. 2021). Expanding on this approach with other hardware and with an open-source approach can enable more widespread, fine-grained understanding of energy uses at every stage of a computation. In addition, a DOE funded effort is developing software for estimating energy usage of different applications on different software-hardware combinations.

2.4.1.2 Benchmarks

Benchmarks have long been used to provide quantitative comparisons of the performance of computer systems, starting with the first widely reported "Whetstone" benchmark (Curnow and Wichmann 1976). The need to incorporate energy efficiency measures in computer equipment benchmarks has been recognized for more than a decade (Fanara, Haines, and Howard 2009).

An industry non-profit group called the Standard Performance Evaluation Corporation (SPEC) was formed to establish, maintain, and endorse standardized benchmarks and tools to evaluate performance and energy efficiency for the newest generation of computing systems. SPEC has developed the SERT benchmark suite to assess energy efficiency of servers. This suite has been incorporated into an international standard (ISO/IEC 21836:2020) for server energy effectiveness metrics (ISO 2020) and has also been adopted as a requirement for the DOE Energy Star rating system for computer servers (U.S. Department of Energy 2018).

Benchmarks generally attempt to provide a useful measure of a computer system's performance by executing a workload that is representative of some important class of real-world applications. There is a proliferation of benchmarks across many application domains. For example, in machine learning, the MLCommons organization—a consortium of AI community researchers and developers from more than 30 organizations—was formed in part to develop and promulgate benchmarks specific to machine learning. MLPerf is an independent, objective benchmark suite published by MLCommons used to evaluate training and inference performance of machine learning systems. The MLPerf Training benchmarking suite measures the time it takes to train machine learning models to a target level of accuracy. MLPerf Inference benchmarks measure how quickly a trained neural network can perform inference tasks on new data.

Challenges and Solution Pathways for Algorithm-Specific Energy Efficiency Benchmarks

To support measurable achievement of the EES2 goal, a suite of benchmarks needs to be established and used to track performance of systems as they evolve over the next decades. There are many benchmarks already in existence covering the range of use cases across domains. A suite of standard benchmarks for assessing energy use, such as the preliminary list shown in Table 55, will be necessary going forward. The term "benchmarking the benchmarks" has been coined to describe this selection process, which may uncover gaps and potentially lead to the development of new more specialized benchmarks for some cases, especially for low-level performance assessment in coordination with the tooling development. There are other use cases among the benchmarks identified.

Table 55. Algorithm-Specific Use Cases and Benchmark Suite Selection

Domain	Software	Benchmark

AI/ML	• Frameworks • ML Compilers • Integrating new AI/ML • accelerators • Data prep techniques	• MLCommons benchmarks • NeuroBench • DataPerf • Domain-specific • Training/inference perf tests Models for Science
Cloud (“Data center tax”)	• Open Source • “Cloud” enterprise apps • REST API services	• Fleet Bench (Google) • Others should be coming out soon
HPC	• (pick some target kernels)	• Rmax • HPL • HPGC • Graph500 benchmarks • (based on kernels picked)
Enterprise	• Enterprise-class Database • In-Memory Databases • Back-Office Applications • Supply Chain • CRM	• TPC Benchmarks (C, E, H, DS) • SpecJBB • SpecVIRT/Vmmark Virtualization

The needs identified for both tooling and benchmarking are best addressed by an industry- or government-sponsored organization that can fill the following roles:

- Perform benchmarking across diverse hardware and software platforms from multiple vendors and provide a centralized repository for reporting results.
- Develop energy profiling tools able to operate on multiple hardware and operating systems platforms and provide and support those tools as open-source solutions, enabling energy measurements to become commonplace rather than the difficult-to-get data, as is currently the case.

- Provide support tools for energy use estimates in computer performance simulations, allowing highly granular assessment of energy performance throughout the simulated system.
- Coordinate with industry benchmarking standards groups to disseminate findings and tools throughout the industry.
- The ability to precisely measure the energy impact of software is crucial for assessing and enhancing energy efficiency throughout the entire computing stack in order to monitor and quantify the industry's overall progress toward the EES2 goals, as illustrated in Figure 43. A combination of benchmarks selected to cover all prominent use cases, as suggested in Table 55, along with the tools to accurately measure and simulate energy efficiency performance will provide clear feedback to stakeholders regarding energy reduction progress and opportunities.

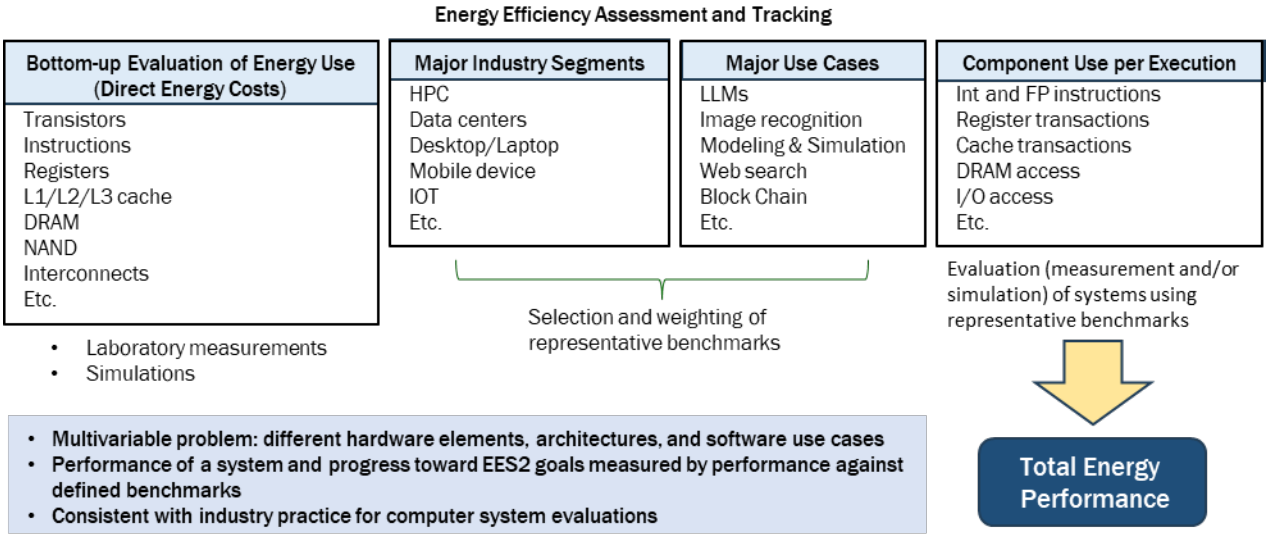


Figure 43. EES2 proposed approach to evaluation of computer system energy performance and progress toward long-term improvement goals

Action Plan for Algorithm-Specific Energy Efficiency Tooling and Benchmarks

Table 56. Action Plan for Algorithm-Specific Energy Efficiency Tooling and Benchmarks.

Scope	
Key Technology for Energy Efficiency	Benchmarks and tooling for algorithm-specific energy efficiency measurement and improvements.
Technologies of Interest	Energy efficiency metrics for conventional general-purpose computing systems, domain-specific accelerators, and emerging compute paradigms.

Challenges	Solution Pathways
• Programmers have inadequate tools for access to information about the energy impact of programs or what programming decisions affect the energy impact. • Modern microprocessors have counter subsystems, but these are generally difficult to access and provide no information about energy associated with events. • Architectural limitations inhibit effective means of making energy measurements in some cases. • Inadequate attention is given to energy efficiency during systems and software development. • The breadth of the EES2 energy efficiency goal requires an industry-wide view of energy performance across a wide range of machine architectures and software use cases.	• Determine and publish energy expended for operations at a granular level, similar to the Horowitz (2014) and Jouppi et al. (2021) papers table of energy costs, for each architecture (CPU, GPU, accelerators). • Provide profiling tools that enable executed programs to be measured in terms of these operations, enabling energy measurements for program sections. • Provide simulation tools for simulating energy performance of systems and software, especially during development. • Define a set of energy vs performance benchmarks covering all prominent use cases in the industry. • Provide a neutral source for collecting and promulgating energy use information, methodology, and tools.

Major Tasks / Milestones	Metrics	Targets	Timeline (years)
Develop ongoing research/metrology capability to replicate the Horowitz (2014)/Jouppi et al. (2021) measurements on any computer system.	Energy per operation (Joules/op)	Measurements on multiple CPU and GPU processor platforms with different memory configurations	1–2
Develop energy models to map counter measurements to energy use and to establish traceability from event counters to processes and instructions	Energy consumption of a program or program fragment (Joules per execution)	Robust profiling tools made available for multiple platforms	2–3
Develop benchmark codes for different algorithm classes to be run on different systems as a rating metric	Energy ratings similar to Energy Star	Open-source benchmark codes that work with the profiling tools to adaptable perform benchmarking on a wide variety of commercial systems	3

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
End Users/OEMs	Run tests (particularly for data centers)
Academia	Perform most of the research work through graduate student projects
National Laboratories	Potentially host the metrology institute
Government	Funder

Required Resources	Cross Collaboration with Other Working Groups
Product manufacturers or suppliers: Make hardware and software available for testing Government: Provide funding and facilitate access to national-level computing facilities	Circuits and Architectures: because algorithm energy use measurement depends on measurements at the level of circuits & architectures as input. Metrology and Benchmarking: these measurements should fit into some standardized benchmarking framework. Education and Workforce Development: - A course similar to MIT 6.016 is needed for “Design for Energy Performance.” - Some of the benchmarking studies needed would be good graduate student research projects.

2.4.2 Reduced Energy for Machine Learning Algorithms

Worldwide, the market for machine learning (ML) and AI applications is growing at an astounding rate. This trend is expected to continue over the remainder of this decade as shown

in Figure 44 rising from $208 billion in 2023 to more than $1.8 trillion projected in 2030 (approximately 9x, equivalent to 36.6% annual growth). Thus the need to find ways to reduce the energy intensity of machine learning applications is urgent.

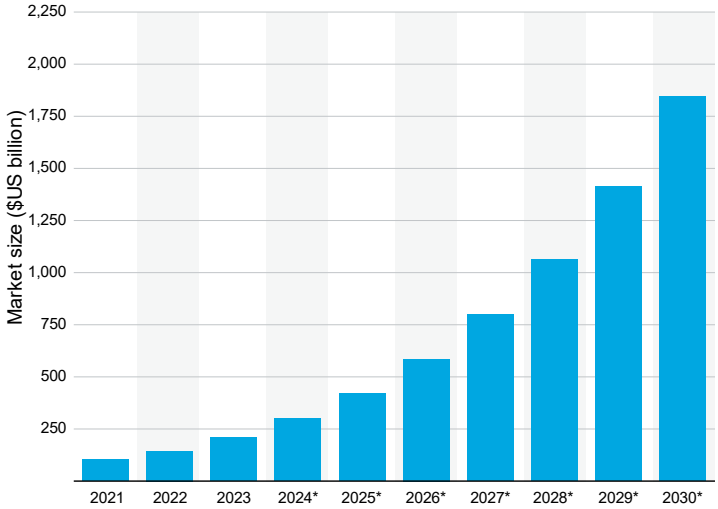


Figure 44. Market growth worldwide for machine learning and artificial intelligence through 2030. *The market is anticipated to grow at a 36.6% compounded annual rate. Source: Statistica 2023*

The term artificial intelligence (AI) broadly means the ability of computers to exhibit independent intelligence (as opposed to executing explicit human-designed algorithms for solving specific problems), but practically speaking, AI systems today are all based on machine learning using neural network algorithms. A simple neural network is depicted in Figure 45, comprising an input layer with three inputs, an output layer with two outputs, and a "hidden" layer with four nodes. The arrows depict connections between these pseudo-neurons (crudely mimicking biological neurons) and each arrow has an associated "weight" coefficient that adjusts the influence of each node's input to its output. Values of these weights are determined by a training process that combines a set of example inputs and outputs (the training set) with the goal of making the network produce the correct outputs for the training set. The trained network may then be used with a wider set of inputs to produce estimated outputs (with a degree of accuracy that depends on the design of the network, the size and quality of the training set, and the difficulty of the problem).

Large-scale deep neural networks (DNNs) (neural networks with more than one hidden layer) have shown impressive performance in many domains, including computer vision and natural language processing (O' Neill 2020). Many of the remarkable gains in machine learning performance have been enabled using increasingly large models, with a growth of about five orders of magnitude in the number of parameters over eight years (see Figure 46). As a result, training and using DNNs require immense amounts of energy and contribute to a large carbon footprint. For example, GPT-3, with more than 175 billion parameters, reportedly consumed 1,287 MWh for training (de Vries 2023).

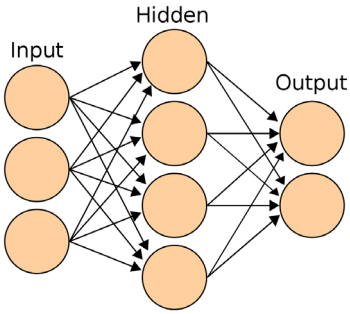


Figure 45. Neural network with one hidden layer. *Source: Wikipedia 2024*

This growth in model size has been spurred by the discovery that model performance can be improved by over-parameterizing, where the number of model parameters greatly exceeds the number of data points in the training set. Overparameterization, combined with the ever-increasing training set sizes needed to reduce statistical error rates, has caused overall computational growth to increase by at least the fourth power of the target performance. In practice,

computational demand has grown even faster (Thompson et al. 2022).

The biggest driver of the increasing energy use of computing is the feedback loop among the demand for model performance, model size, and adoption of ML applications: The desire for continuous improvement in performance drives ever-larger datasets, with a corresponding increase in the number of parameters in the latest models. Successful application of these increasingly large models leads to massive proliferation of ML applications. Each new success drives users to desire even better model performance, and so on.

Challenges and solution pathways for machine learning

ML training is much more energy-intensive than ML inference is. However, because a trained model is typically used for many thousands or even millions of inferences, the total energy cost of inference for a model may be equal to or greater than that of its training. Hence, reducing the energy use for both ML training and inference is important for achieving the EES2 goals.

At a fundamental level, the limit to what quality of training can be achieved for a given dataset and training task for different machine learning algorithms is not known. This theoretical knowledge gap may have a large impact on the energy cost of training once the limits become better understood. Even though current understanding is limited, several important observations have been made that suggest areas ripe for improvement in terms of reducing energy use in ML:

- The energy cost of training is inversely dependent on training data quality (i.e., training cost is higher when the training data quality is lower).
- Mechanisms to transfer model learning or model hyperparameters (tuning) from one model to another are not well-understood; finding effective transfer methods would greatly reduce energy consumption.
- Precision requirements are different for different aspects of training and inference. (Generally, higher precision is needed for training than for inference). Adaptive approaches may be able to optimize precision at different steps of the training and inference processes to reduce energy use with minimal impact on accuracy.
- Animals can learn from noisy data. In addition, natural learning is robust, retaining inference accuracy well even when data distribution shifts. Producing algorithms that come closer to the performance of natural systems (e.g., evolutionary algorithms) will require research and experimentation. Such nature-inspired algorithms may eventually deliver major reductions in the quantity and quality of training required.

- Note that, in addition to the opportunities discussed below, ML applications to compilers, code writing, and runtime systems are covered in section 2.4.5.1. Opportunities for software gains in connection with compute-in-memory and neuromorphic hardware architectures are discussed in section 2.4.6.4.

2.4.2.1 Energy-Efficient Alternative Training Methods

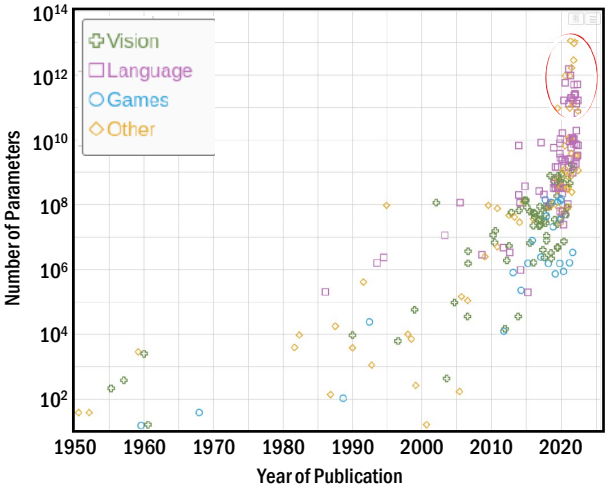


Figure 46. Complexity of machine learning models.*Source: Villalobos et al. 2022*

There are many potential opportunities to improve the energy efficiency of ML training and inference processes. These are applicable for both data center operations and edge devices such as mobile phones but are particularly important for the latter due to their limited energy budget. Some methods for improving training efficiency include (list adapted from Verhelst and Murmann 2020):

- **Software optimizations:** Design algorithms to maximize spatial and temporal locality of data access in a CPU memory hierarchy; optimize data flow in systolic arrays (such as in the Google TPU) or in compute-in-memory architectures; or enable wholly new emerging architectures, such as neuromorphic spintronics (Grollier et al. 2020).
- **ML processing:** Expand ML processing on edge devices rather than in the cloud. This would avoid the time and energy cost of transferring huge amounts of data collected at the edge to data centers.
- **Model compaction:** Manipulate network topology to co-design the computation with available hardware resources, avoiding bottlenecks when operating on constrained embedded processors.
- **Model quantization:** Manipulate the numeric precision of model weights and activations. Limited fixed-point representations of eight, four, or even fewer bits have been shown to be adequate for many inference tasks. Math computations run much faster in reduced precision, especially on GPUs with Tensor Core support for that precision (NVIDIA 2023a), while also reducing required memory bandwidth.
- **Pruning:** Selectively remove near-zero weights after or during training. This can result in a sparse neural network which can be exploited for more efficient storage. It may also result in more efficient computation when supported by sparse neural processor hardware that skips zero-valued multiplication operations.

Gains from application of these strategies may be much greater when used together, but combining them is a research challenge. For example, model compaction may not work well with model quantization in some instances.

Additional energy efficiency gains are possible when algorithms are combined with architectural features during co-design. New neuromorphic architectures are emerging with good systems and spiking models, but to date, there is no corresponding end-to-end software design stack to support model adoption for large scale problems.

Another potential direction for efficiency improvement in training is in data preparation, which often requires manual intervention or bespoke data quality filtering software. Better software tools for automating data and feature preparation could boost efficiency both by consuming less development time and by providing better training data quality, allowing a target level of inference accuracy with less training data.

2.4.2.2 Approximate/Efficient Matrix/Tensor Multiplication

Linear algebra, specifically matrix multiplication, is at the heart of most of the computational algorithms including machine learning and is generally at the core of both training and inference. Therefore, methods to accelerate matrix multiplication have the potential to make ML tasks faster. Extensive efforts have been expended on efficient exploitation of data spatial and temporal locality when subdividing the work to minimize references to slow external memory, with differing requirements and strategies depending on the hardware available. Additionally, lower precision numeric representations have effectively reduced resource demands while maintaining overall fidelity, suggesting that approximate matrix multiplication could also be beneficial in many ML applications.

Many approaches have been tried with varying degrees of success. Recently, Blalock and Guttag (2021) demonstrated an approximate matrix multiplication method called Multiply-ADDitioN-lESS (MADDNESS) that accomplishes an approximate multiplication with bounded error 10x faster than other approximate methods and 100x faster than exact multiplication. The algorithm is especially efficient when one of the two matrices is known beforehand (e.g., representing model weights). In this case, the product matrix is approximated using no multiplications at all, instead using a simple binary tree hash function to compute an index into a pre-computed lookup table of dot products. Blalock and Guttag (2021) provide a brief description of and comparison to several other approximation algorithms.

Other methods include extending approximation methods to GPUs and other existing accelerators, convolutional networks with weight reuse, memory use optimization, and custom accelerator design for edge devices. Efficiency gained through a combination of advanced architecture and algorithms is further discussed in section 2.4.6.4.

2.4.2.3 Meta-learning of Hyperparameter Optimization

In machine learning, meta-learning, or “learning to learn,” is a data-driven approach in which metadata from prior ML is used to assess which approaches have been most effective and adapt learning strategies accordingly to speed up the learning process. This, in turn, suggests the possibility of effective machine learning with drastically reduced computational and energy cost by converging on a trained network with far fewer iterations.

Meta-learning has been explored in a wide variety of applications, including computer vision, robotics, neural architecture search, hyperparameter optimization, language and speech, and others (Hospedales et al. 2021). Ansótegui et al. (2021) employed a meta-learning approach to demonstrate a 50% reduction in the computational cost to optimize arbitrary “black box” functions using results from each iteration and applying machine learning to the selection of

next iteration parameters. Such an approach is desirable when the function to be optimized is itself computationally expensive, for example engineering design optimizations involving complex simulations.

Hyperparameters are variables that determine the configuration of a neural network, distinct from the parameters in the dataset that are used for training and inference. The hyperparameters are set before training a model. Examples of model hyperparameters include the number of hidden layers in a neural network, the number of nodes in each layer, and even the type of neural network model to be used. Algorithm hyperparameters include learning rate and batch size. Hyperparameters cannot be learned directly from the training data but nevertheless can be optimized with measures such as inference accuracy or training time. Commonly used methods to optimize hyperparameter values are:

- Bayesian search, which conditions probability of an outcome on the state of current knowledge.
- Grid search, an exhaustive trial of all possible combinations of the hyperparameters to determine the best one. Grid search is computationally intensive, especially with large numbers of hyperparameters (the “curse of dimensionality”).
- Random search selects combinations of hyperparameters randomly rather than systematically to find a near-optimum combination with far fewer trials compared to grid search. It can be combined with grid search over a smaller search space determined by an initial random search.
- Evolutionary or genetic algorithms and other heuristic approaches use mutation, crossover, and selection from an initially randomized set of hyperparameter values to evolve toward an optimal solution.
- Multi-fidelity searches, especially for very large models, evaluate the hyperparameter selection on a small subset of the dataset using one or a combination of the other methods and infer performance over the full dataset from the results.

Each of these approaches presents opportunities for improved learning performance guided by past experience. Key challenges include application of meta-learning methods to diverse tasks rather than tasks drawn from closely related tasks, and improving the ability to generalize from learning metadata. In some cases, computational cost is a challenge, and a number of solutions have been explored to devise computational shortcuts for the training of meta-learning parameters. Use of meta-learning techniques, in combination with different methods, especially for very large neural networks, can help in optimization depending on the specific application.

Systematic collection of data from previous ML applications can be applied to reduce learning costs, including energy cost. Effective methods of applying past ML training experience to new applications are just beginning to be explored in this extremely active area of research.

2.4.2.4 Continual Learning (Sequential Training without Catastrophic Forgetting)

Many approaches to continual learning rely on the stochastic gradient descent training method and must adopt strategies such as memory buffers or replay to avoid catastrophic forgetting—the tendency of a neural network to abruptly forget previously learned information as a result of new incoming information. Madireddy, Yanguas-Gil, and Balaprakash (2023) developed a biologically inspired ML architecture that incorporates synaptic plasticity mechanisms and

neuromodulation to enable continual learning without stochastic gradient descent. This memory-free architecture achieves continual learning performance superior to that of other memory-constrained learning approaches and matches the performance of memory-intensive replay-based approaches. The high accuracies achieved rely in part on a novel inelastic rule that implements a simple form of memory consolidation for synaptic weights that deviate from the presynaptic weights of each neuron, leading to a stabilization of weights that mitigates catastrophic forgetting.

Harun et al. (2023) performed a comparative assessment of the efficiency of multiple continual learning systems and found that, despite recent methods that have largely solved the catastrophic forgetting problem, many of the methods for incremental learning are highly inefficient in terms of computation, memory, and storage, with some methods requiring more computation than training from scratch does. Ideally, a model should adapt to a growing training dataset without increasing the computation or memory, but most continual learning methods lack this ability.

Biological organisms are able to learn throughout their lifetimes from interactions with their environment. It is desirable for neural network machines to be able to similarly learn on a continual basis, without expending disproportionate amounts of energy. This challenge is known as lifelong learning and largely remains unsolved. Kudithipudi et al. (2022) identified a set of key capabilities that artificial systems will need to achieve lifelong learning and described biological mechanisms that help explain how organisms solve these challenges. Examples include transfer of knowledge for application in new circumstances; exploitation of task similarity by decomposing tasks into more elementary, reusable components; noise tolerance; and hierarchical distributed neural networks for specialized functions that enable both fast response and reduced complexity of higher-level brain functions.

Progress in bio-inspired continual learning has the potential to play a critical role in reducing the energy cost of ML training by focusing on energy-efficiency in both the formalisms of learning and in the implementations of training methods.

2.4.2.5 Physics-Informed ML Models for Scientific Computing

Machine learning is increasingly being used to solve problems in applied mathematics, engineering, and physics, using equations that model the problem to guide the training of the neural network. Physics-informed neural networks (PINNs) are neural networks that incorporate physics in appropriate model equations, such as partial differential equations, as a component of the neural network itself (Cuomo et al. 2022). The framework for such models was first introduced by Raissi et al. (2017), although there were many prior examples of related work before the formulation of a formalized framework. Figure 47 shows a generalized flow diagram of such a neural network model.

Such ML applications are a promising research direction in scientific computing in general, offering the potential to displace or enhance other computationally intensive engineering calculations, such as finite element solvers, for substantial energy savings. PINN are an active area of research, with many examples in power systems (Huang and Wang 2023; Misyris, Venzke, and Chatzivasileiadis 2020), fluid dynamics, quantum mechanics, materials science, optics, electromagnetics, and other fields (Cuomo et al. 2022).

There are still many unresolved challenges, such as convergence and stability, as well as implementation issues with software architectural design, including boundary conditions management, neural network hyperparameter selection, and optimization strategies. As is the case with other fields of ML research, the choice of the best type of neural network (feed-forward, deep learning, convolutional, recurrent, or others) is not well-understood. Integration of PINN into scientific analysis code written in conventional programming languages such as C++ and Python is also a challenge.

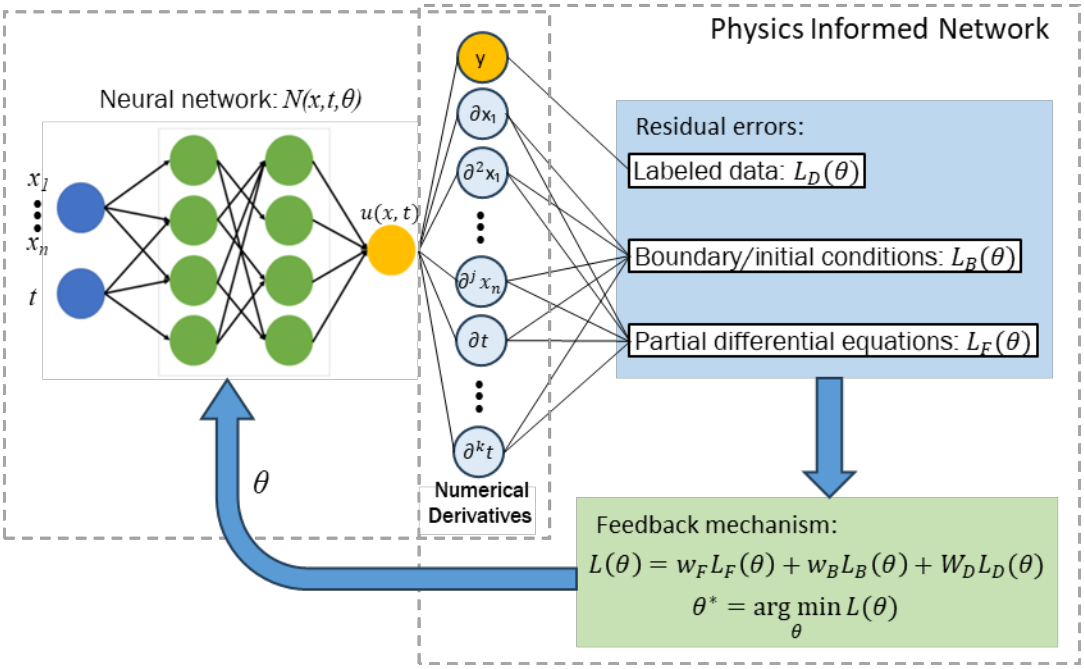


Figure 47. Physics-informed neural network differential equation solver. *The network is defined by θ. Its input variables are transformed into network output field u. Derivatives are calculated from the given equation(s) and u, and the residuals are used as feedback to train the network. Source: Cuomo et al. 2022.*

2.4.2.6 Bottom-up Sparse ML Model Development

One strategy to enable large models with high performance but better scaling is to use sparse connectivity. This approach routes individual inputs to different "experts" in a potentially huge network instead of passing every input to every part of the neural network. This is known as a "mixture of experts" model. For example, a mixed image and text recognition/classification model could have separate specialized expert sub-models for image and text analysis (Mustafa et al. 2022).

The mixture of expert model approach mainly benefits performance during inference, not training. As mentioned earlier, although training takes more computational energy, inference is typically used hundreds or thousands of times for each training instance and thus accounts for most energy use. Recently, Huang et al. (2023) proposed a method for a dimensional reduction technique that can be applied during the training phase and can generally be applied to any network architecture. The authors showed that the reduced version of the neural network maintains high accuracy but takes much less energy, model storage, and computational time in both training and inference for a specific application. This approach thus presents a promising direction for further research.

2.4.2.7 Benchmarking Hyperparameter Optimization Methods

The number of choices in and increasing size of ML designs makes research on effective methods for hyperparameter optimization (HPO) essential. However, the research community lacks realistic, diverse, computationally cheap, and standardized benchmarks that can be used to compare optimization approaches. Eggensperger et al. (2022) has proposed a set of containerized, multi-fidelity benchmarks, allowing them to be reproducibly run for computationally affordable and statistically sound evaluations. The suite of benchmarks is called HPOBench and was tested in a large-scale study evaluating 13 optimizers from 6 optimization tools. This kind of benchmarking capability will not only shed light on the most effective optimization strategies but can also provide the dataset for a meta-learning approach to highly efficient hyperparameter optimization.

2.4.2.8 Benchmarking and Methodology to Quantify Training and Inference Costs

ML training presents three key benchmarking challenges. First, multiple implementation factors (such as processor architecture, memory architecture, and network size) simultaneously affect both training throughput (speed) and the training time to reach a specified quality threshold. Second, the stochastic nature of training causes run-to-run variation in time to solution. And third, the diversity of software and hardware systems makes fair benchmarking difficult. MLPerf, as noted in Section 2.4.1.2, aims to address these challenges. MLCommons has developed the MLPerf benchmark suite to measure how fast systems can train models to a target quality metric (Mattson et al. 2020). An MLCommons Power Working Group has also been established to create power measurement techniques built on industry-standard tools in support of MLPerf benchmarks. Future work should benchmark a more comprehensive set of accelerator configurations and include benchmarks for inference on both servers and edge devices.

Action plan for reduced energy for ML algorithms

Table 57. Action Plan for Reduced Energy for ML Algorithms.

Scope	
Technical Challenge for Energy Efficiency	Reduce the energy cost of software (both training and inference) for machine learning applications.
Technologies of Interest:	• All types of machine learning algorithms • All hardware architectures • Training set data quality • Incremental/progressive training
Challenges	**Solution Pathways**
• Data/dimensional reduction and specificity • Network architecture and optimization • Hyperparameter optimization • Numerical operations • Adaptation of nature-inspired learning in ML methods	• Develop large-scale benchmarks and methodology to quantify training and inference costs • Demonstrate energy-efficient alternative training methods • Develop approximate matrix/tensor multiplication methods • Achieve continual learning without catastrophic forgetting • Optimize physics and data requirements for scientific machine learning • Develop sparse models bottom-up, avoiding the need to build and prune large models • Develop hyperparameter optimization methods for large language models and meta-learning methods for predicting optimal values • Improve meta-learning of hyperparameter optimization for training large models

Major Tasks / Milestones	Metrics	Targets	Timeline (years)
Meta-learning of hyperparameter optimization for training large models	Iterations required to reach optimal hyperparameters for training, energy savings from optimized training for a target performance	Few shot (5 data points) to one shot optimization (1 data point)	1–3
Continual learning: achieve sequential training on multiple tasks without catastrophic forgetting	Accuracy on all tasks compared to that obtained from full retraining, total energy cost for training	95% relative accuracy without having to retrain on prior data. 95% energy savings with respect to a specialist model	5
Establish trade-off between physics and data requirement for SciML	%data/dimension reduction	90% reduction in required data if 90% physics known	5
Develop sparse models bottom-up, i.e., avoiding the need to build and prune large models	% reduction in parameters of language/vision/time series models	95% of state-of-the-art accuracy with 90% reduction in size	5
Develop benchmark hyperparameter optimization methods for large language models and develop meta-learning methods for predicting optimal values	Number of iterations required to achieve optimal hyperparameters	Identify optimal hyperparameter for a new model with <100 iterations	3
Develop large scale benchmark and methodology to quantify training costs	Energy cost per training experiment	One dataset/architecture per main use case as defined by its footprint in energy consumption	1–3
Demonstrate energy-efficient alternative training methods (low precision training, non-gradient methods, neuromorphic computing architectures)	Energy cost per training experiment	90% reduction in training costs	3–5
Approximate/efficient matrix/tensor multiplication	Energy cost per flop	50% reduction in energy with 90% accuracy maintained	5

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Software Developers	Develop and implement improved algorithms
Academia	Develop and implement improved algorithms
National Laboratories	Develop and implement improved algorithms and benchmarks
Government	Provide targeted funding programs for high impact areas

Required Resources	Cross Collaboration with Other Working Groups
	Circuits and Architectures: Breakthroughs in this area will require a combination of algorithms and hardware accelerators, co-designed to work together for maximum efficiency.

2.4.3 Reduced Energy for Algorithms Used in Scientific Computing

The U.S. Department of Energy (DOE) and its affiliated laboratories are at the forefront of employing scientific computing to tackle a wide range of challenges in biology, chemistry, physics, and materials science. These computations require significant processing power, so it should come as no surprise that as of 2023, DOE laboratories operate three of the world's top ten supercomputers. A comprehensive analysis of 500 supercomputers, as reported by the TOP500 list, which ranks the most powerful computer systems worldwide, highlights the intensive energy demands of these systems (TOP500 2022; Barrett et al. 2010). The Top500 analysis, detailed in various studies, spans systems from 2010 to 2022 and includes the first reported exascale computer (Shankar and Reuther 2022).

In these high-performance computing systems, energy is consumed by two main sources: the hardware/system-level architecture and the algorithms/software, which are inherently interdependent. Accurate energy assessments thus require benchmarks that consider both the raw performance metrics and the actual time needed to complete scientific simulations, emphasizing the critical co-dependency of hardware and software in achieving energy efficient scientific computing.

Energy benchmark analysis

Energy benchmark analysis is vital for scientific computing because it directly impacts the efficiency and sustainability of supercomputers engaged in complex simulations. An analysis by Shankar and Reuther (2022) evaluates the performance of the world's most powerful computing systems, as ranked by the TOP500 list, focusing on the High-Performance Linpack (HPL) and High-Performance Conjugate Gradient (HPCG) benchmarks.

The HPL benchmark assesses a supercomputer’s ability to solve a dense linear system of equations using double or higher precision arithmetic (FP64). It employs LU decomposition, where a matrix is factored into a lower and an upper triangular matrix, followed by back substitution to find the solution. This process, crucial for high-performance computing systems, largely consists of complex matrix multiplications that benefit from parallel processing capabilities of modern CPUs, GPUs, and memory subsystems. Rmax and Rpeak denote the actual maximum and theoretical peak performances of these systems, respectively, reported in teraFLOPS or petaFLOPS.

The HPCG benchmark, in contrast, evaluates the efficiency of data access and computations in a conjugate gradient solver, which is foundational for simulating physical systems. It tackles a structured sparse linear system of equations using stencils, a method that relies on FP64 for its accuracy and stability. Due to its focus on sparse data patterns, HPCG typically shows lower performance rates compared to HPL, emphasizing the necessity for both benchmarks to ensure comprehensive evaluation and numerical stability of supercomputing systems (Heroux and Dongarra 2013).

Because not all systems are evaluated using the HPL and HPCG benchmarks, the importance of relevant benchmarks is underscored by Figure 48, which graphs Rmax for the top 500 systems. For instance, the Frontier system, recognized as the most energy-efficient on HPL benchmark among the TOP500, is over 200 times less energy-intensive than the LLNL CTS-1 Quartz, another system operated by the DOE. This stark contrast in energy efficiency, based on analysis from 2022, highlights the need for more comprehensive benchmarks to better understand energy disparities and identify strategies to enhance energy efficiency across all high-performance computing systems.

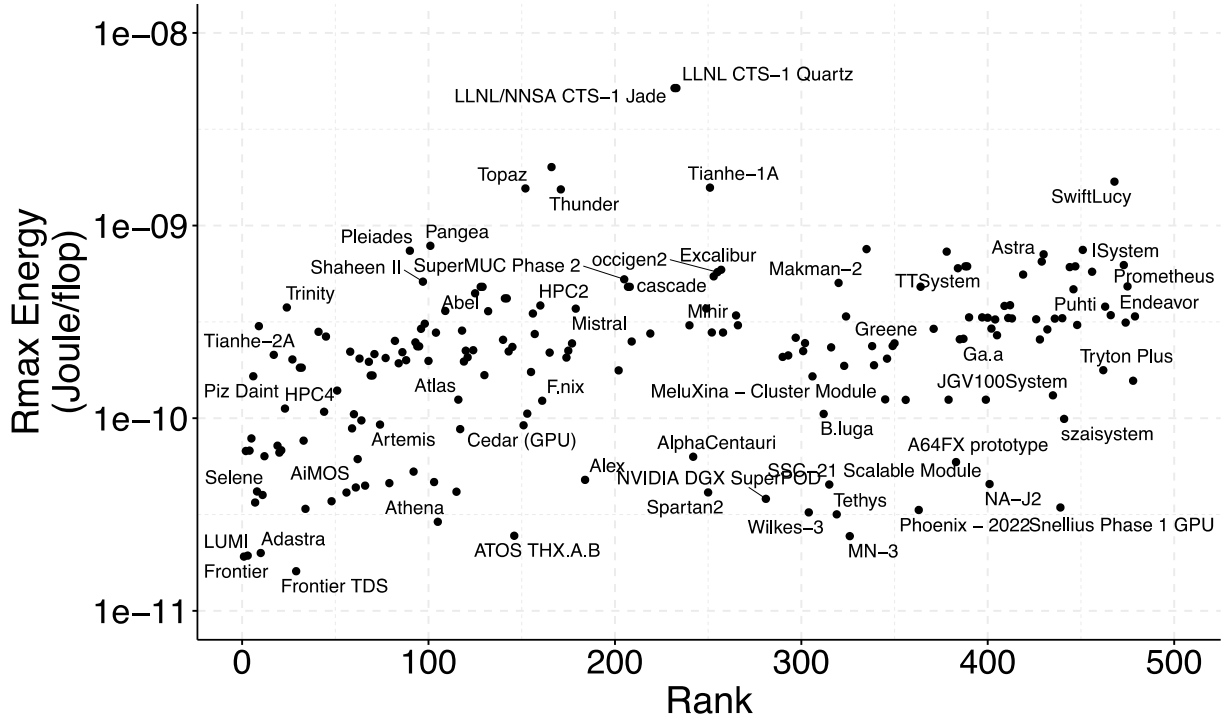


Figure 48: Energy/Instruction based on HPL benchmarks Rmax. *The X-axis consists of the top 500 supercomputers with the fastest on the left and slowest on the right. Source: Shankar and Reuther 2022.*

To illustrate the energy requirements for large-scale scientific computations, we consider the simulation of a SARS-CoV-2 spike protein, a crucial component of the coronavirus's infection mechanism. Conducted on a supercomputer, this simulation analyzed the dynamics of the viral envelope consisting of 305 million atoms. The simulation ran for 8.77 days on 80 P100 GPUs at the San Diego Supercomputer Center, achieving a sampling duration of approximately 7.5 microseconds. Parameters from this study are detailed in Table 58. Assuming energy costs for floating-point operations range between 1 x 10-12 to 1 x 10-11 joules, the total energy expended for this simulation was approximately 24.9 billion joules, as outlined in Table 59 (Casalino et al., 2021). This case exemplifies the significant energy demands of advanced scientific simulations.

Table 58: Simulation parameters for Covid Virion particle simulations.

Source: Shankar 2023

NAMD Simulation	Atoms	Nodes	Sim rate	Performance
Spike-ACE2 complex	8.5M	1024	162 ns/day	229 TFLOP/s
SARS-CoV-2 virion	305M	4096	68 ns/day	3.06 PFLOP/s

Table 59: Energy estimate in Joules and kWh for simulation of a single virion particle.

Source: Shankar 2023

Application	Energy (joules)	Energy (kWh)
Spike ACE Complex	1.74E+09	4.82E+02
SARS Covid Virion	2.32E+10	6.44E+03

Total (Max)	2.49E+10	6.92E+03

The energy consumed during the entire simulation of the SARS-CoV-2 spike protein significantly surpasses that used by a large language model application, despite the shorter simulation duration. This vast difference, spanning over twenty orders of magnitude greater than the energy for a single floating-point operation, is attributable to the immense power requirements and extensive compute cycles demanded by high-performance supercomputers for scientific computations. This scenario underscores the inherent energy intensity of high-precision scientific computing, necessitated by the stringent accuracy requirements of such simulations. This pattern is likely representative of a broad spectrum of large-scale scientific computations, highlighting the critical need for comprehensive system benchmarking. The analysis makes it clear that ongoing efforts to advance and optimize software across different systems and applications are essential to achieve energy efficiency in scientific computing.

2.4.4 Reduced Energy for Cryptocurrency Mining

Electricity demand from U.S. cryptocurrency mining operations has surged dramatically in recent years. Current estimates suggest that annual electricity consumption from cryptocurrency mining accounts for between 0.6% and 2.3% of the nation's total electricity use (EIA 2023). According to *The New York Times*, 34 large-scale Bitcoin mining operations now function in the United States, further straining local power grids (Dance et al 2023). As shown in Figure 49, the energy usage for computer-based cryptocurrency mining, including data centers and AI applications, is becoming a significant share of the electricity used in computing.

Figure 49 presents a comparison of energy estimates (electrical energy associated with computing) from 2016 to 2024 against the annual electricity production of various states (such as Arizona, California), countries (like Australia, the Netherlands), and the annual energy generation of the Hoover Dam hydroelectric project. Additionally, the figure includes lower bounds and estimated energy requirements for cryptocurrency mining (EIA 2024).

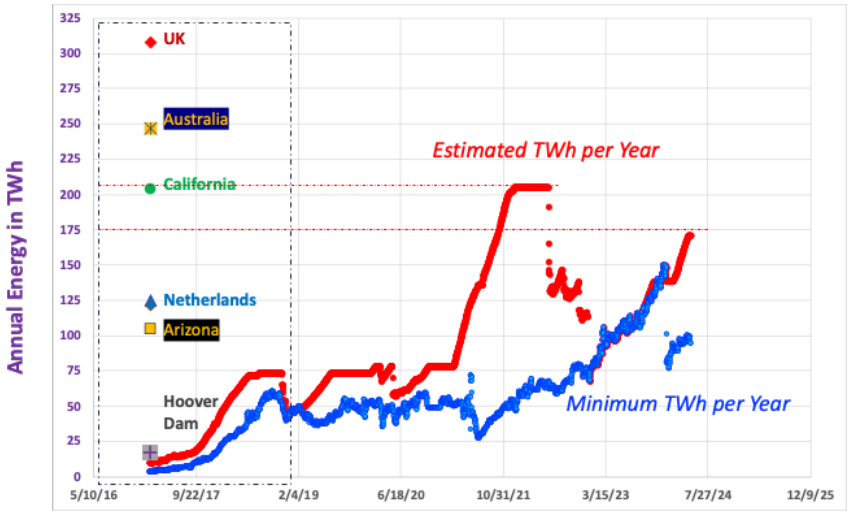


Figure 49. Energy use estimates of cryptocurrency mining. *Energy is compared to the electricity generated in the states of California and Arizona; in Australia, the Netherlands, and the United Kingdom; and by the Hoover Dam project. Source: EIA 2024*

After dipping briefly in 2021, the energy consumption from cryptocurrency mining has seen continued growth since 2023. The cost relative to transaction volume has not changed

significantly since 2010, but the overall energy usage has grown into a significant component of computing (Song and Aste 2020). Current estimates suggest around 170 TWh is consumed, roughly ten times the total output of the Hoover Dam and exceeding the annual electricity production of Arizona and the Netherlands. At its peak between 2021 and 2023, cryptocurrency mining consumed more electricity than the entire state of California did. The U.S. Energy Information Administration estimates that about 38% of global cryptocurrency mining is done in the United States, using approximately 3.5 times the annual electricity output of the Hoover Dam. This energy consumption now rivals the 200 TWh used annually by the world's conventional data centers and constitutes a significant fraction of the 460 TWh consumed by all data centers combined (International Energy Agency 2017).

Strategic responses for the United States

The substantial electricity demand from cryptocurrency mining has prompted specific actions from policymakers and grid planners to mitigate adverse effects on electricity cost, reliability, and related emissions (The White House 2022; de Vries 2018). Challenges in tracking the energy use of cryptocurrency mining arise from the difficulty in partitioning energy use from system-level components to architecture and software. With the trillion-dollar market capitalization associated with cryptocurrency mining, it is expected that both hobbyists and commercial miners will continue engaging in this resource-intensive activity (Thompsett 2024).

Understanding the extent of cryptocurrency mining's profound impact on the U.S. energy landscape is difficult, particularly regarding the consumption of significant computing power and associated energy and material resources. To address potential instability in the electrical grid due to the intense electricity demands of cryptocurrency mining, the U.S. government has issued a report pursuant to *Executive Order 14067,* Ensuring Responsible Development of Digital Assets, raising four critical inquiries (Thompsett 2024):

1. How do digital assets affect energy usage, including grid management and reliability, energy efficiency incentives and standards, and sources of energy supply?
2. What is the scale of climate, energy, and environmental impacts of digital assets relative to other energy uses, and what innovations and policies are necessary for robust comparisons?
3. What are the potential uses of blockchain technology that could support climate monitoring or mitigation technologies?
4. What key policy decisions, critical innovations, research and development, and assessment tools are required to minimize or mitigate the climate, energy, and environmental implications of digital assets?

Optimization of cryptocurrency mining operations

The cryptocurrency mining process, inherently compute-intensive, necessitates an ever-increasing amount of computational power (The White House 2022). Mining operations are performed by networks that execute a one-way hashing function to map digital inputs into fixed-length output digits, essential for validating transactions within a blockchain. Each validation involves solving mathematical puzzles that incorporate transaction data, where the miners generate a vast number of guesses—from millions to trillions—per second to identify unique,

alphanumeric hashes. Once a block of transaction data is verified as correct, it is added to the blockchain, and the successful miner is rewarded with newly minted cryptocurrency.

Mining operations are typically conducted in farms, which consist of numerous video cards and ASIC modules connected to computers, collectively enhancing the network's hash rate and its ability to process and verify transactions swiftly (Kim 2021; Bondarev 2020). However, the continuous operation of these energy-intensive farms poses significant challenges to power infrastructure, particularly in countries like the United States where a substantial number of mining centers are located. This sustained high energy demand highlights the urgent need for research into optimized consensus algorithms and system efficiencies (Lei et al 2021).

In a notable advancement, Ethereum drastically reduced its energy consumption by over 99% in 2022 by transitioning its algorithm to a Proof of Stake (PoS) consensus mechanism. This change aligns the potential of algorithmic and software innovations in reducing the energy footprint of blockchain technologies (The White House 2022). Further research in this area could lead to more sustainable practices across the industry, alleviating stress on global energy resources.

2.4.5 Software for Conventional Architectures

The technologies described in this section refer to software “for conventional architectures,” with emphasis on CPUs and GPUs, but applicable to emerging architectures as well. Some general themes that emerged from working group discussions include the following:

- Tooling as discussed in section 2.4.1 is needed to find new ways to optimize (i.e., incrementally improve) high-use software.
- More efficient languages, compilers, and libraries (e.g., math kernels) will allow for more efficient programming by leveraging a given microarchitecture’s capabilities.
- Some common software functions such as encryption, error correction, and communications offer opportunities for energy-saving optimizations. These are necessary functions in computer systems that can be viewed as a sort of “tax” that must be paid for systems to work.
- Newer hardware needs updated compilers and libraries to allow it to be used in old problems (e.g., NVIDIA Rapids, AMD ROCm). This need exists for incrementally improving performance in conventional hardware, as well as for emerging architectures/devices. Emerging hardware or software architectures need to be integrated into existing processes and show they can solve existing problems, and that the benefits of new architectures justify the overheads of integrating them.

2.4.5.1 Languages, Compilers, and Runtime Systems

The proliferation of multi-core processors has spurred a need for software that can harness the full potential of these systems. However, parallelization (restructuring code to enable portions to run simultaneously on multiple processors) is an advanced topic in computer science education and code optimization via explicit parallelization in source code is labor intensive and potentially error prone. Profiling tools (discussed in section 2.4.1), when adapted to provide accurate and detailed energy reporting, can help skilled programmers to more easily and quickly identify

opportunities for optimization across complex computing environments, thus facilitating optimization with less investment in programmer labor.

It is often possible to achieve significant improvements in program speed by optimizing software implementation. Bentley (1984) described a well-known example in one of his famous "Programming Pearls," where an orbital mechanics program was sped up by a factor of 400, reducing run time from about a year to less than a day. Of that 400x factor, 2x was from faster hardware and the rest was from optimizing the code design: 40x from better data representation, 2x from tuning of loop step sizes, and 2.5x from recoding a critical procedure in assembly language. Other examples, such as using optimal vector lengths in do-loops and rewriting logic for specific architectures (e.g., vector processing machines like Cray supercomputers in the 1980s and 1990s), can contribute to co-optimization between hardware-software components that uses the machines to reduce computational effort.

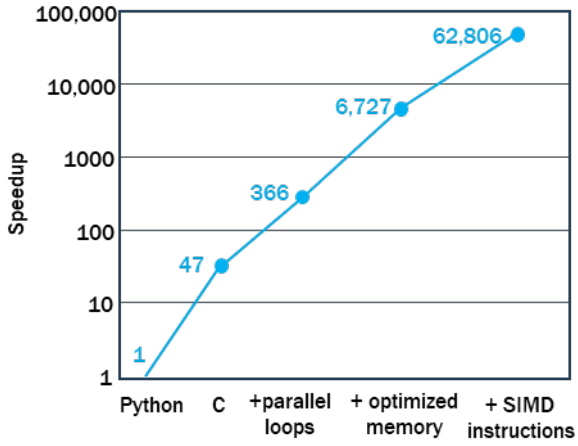


Figure 50. Matrix multiplication speedup over native Python. *Source: Hennessy and Patterson 2019*

A more recent well-known example, shown in Figure 50, comes from Hennessy and Patterson's (2019) Turing Award paper. A matrix multiply operation written in interpreted Python was sped up by almost 63,000x by reimplementing the code in C, writing it explicitly to operate in parallel, optimizing cache memory access, and exploiting vector multiplication hardware.

Faster programs typically use less energy. A recent study (Pereira et al. 2021) compared the performance of many programming languages on a common Linux-based desktop platform. This comparison was part of a project called the Computer Language Benchmarks Game, in which benchmark programs are collected in as many programming languages as possible and are run in a common system operational setup. Figure 51 illustrates the results for one of the benchmarks. (Results for other benchmarks were similar but not identical.) Energy consumed was measured using The Intel Running Average Power Limit (RAPL) tool measured the programs' energy consumption. The negative correlation between speed and energy consumed was strong across all languages.

Although a skilled programmer can make major efficiency gains in some cases by explicit handling of optimization, it is more scalable if the compilers perform optimization automatically. Foundational software components that run continuously justify the expenditure of considerable effort at manual optimization. However, most software development is conducted either in situations where such time investment is not feasible, or by personnel who lack the training (for example, when scientific research code is written by research scientists, not computer scientists).

The Python optimization results shown in Figure 50 may be an extreme example, but it underscores the widespread inefficiency of software due to both language choice and programming practices. Python is an interpreted rather than compiled language, trading

execution efficiency for ease of coding. The prevalence of Python continues to grow (Cass and Goldstein 2023), thanks both to the availability of a rich ecosystem of libraries and packages for every variety of programming problem and to the ease of learning and experimenting afforded by the interpreted language. Although the Python interpreter imposes an inefficiency in program execution, much of the widely used Python code infrastructure (e.g., NVIDIA Rapids and CUDA) is built on fast, compiled, optimized libraries implemented in C/C++ with convenient Python language wrappers for ease of use.

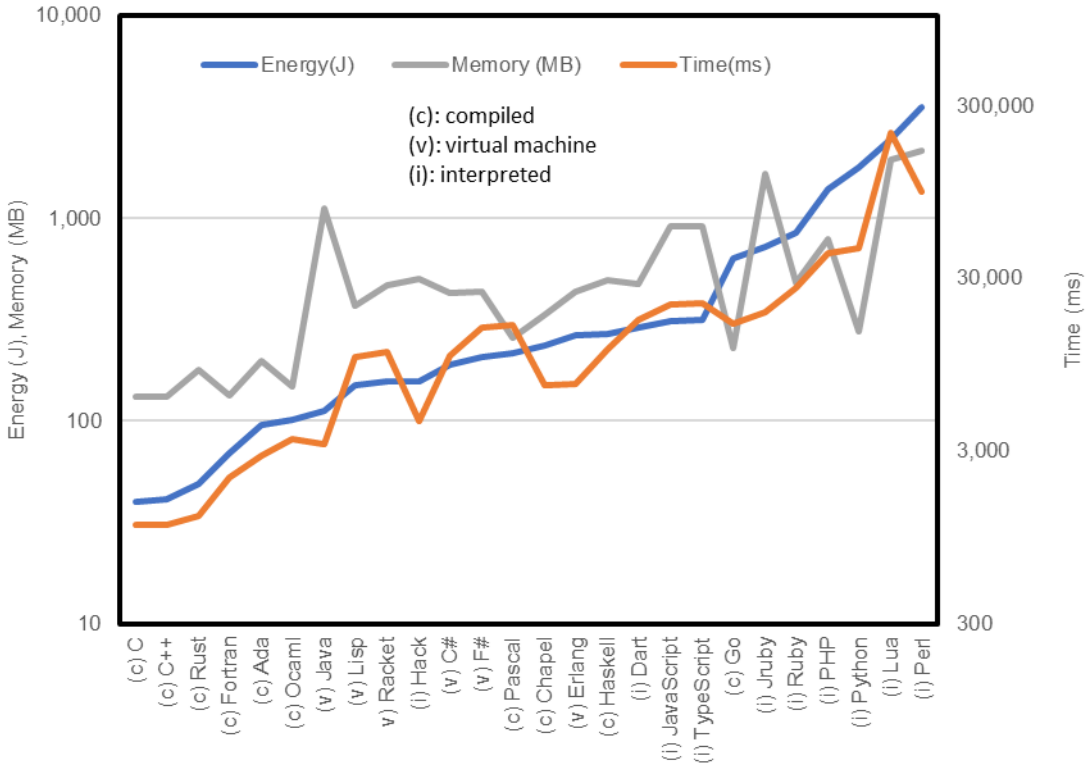


Figure 51. Comparison of the energy, speed, and memory used for various programming languages. *Results are for the binary-trees benchmark. Source: Pereira et al. 2021.*

Some more recently introduced languages such as Julia and Mojo aim to be comparable to Python in terms of user friendliness but in a higher-performance compiled implementation. Julia is a high-level, general-purpose programming language, increasingly being used for numerical analysis and scientific computational problems (Fischer 2022). Testing Julia with codes indicates improved energy efficiency compared to other high-level languages like Python for specific applications (Pereira et al. 2021). Mojo is particularly notable because it aims to be code-compatible with Python 3.x, supporting both ahead-of-time and just-in-time compilation as well as full compatibility with the popular Jupyter notebook style of Python programming. Mojo is still under development and not yet capable of full compatibility. Further maturation of Julia, Mojo, and similar language initiatives will yield large practical benefits in program speed and energy efficiency.

All major commercial and open-source compiler systems (e.g., gcc, Visual Studio, LLVM) now have built-in optional optimizers for auto-parallelization of code to run efficiently on multiple cores, as well as auto-vectorization of code (automatic execution of an arithmetic operation on

multiple elements of an array simultaneously) to use single-instruction-multiple-data (SIMD) vector instructions. Code "optimization" generally cannot be provably optimal; rather, code optimizations exploited by compilers make use of heuristic rules that are painstakingly discovered through intuition and experimentation by compiler writers.

Of particular importance in mitigating the memory performance bottleneck is the need for cache management optimizations. Although the software has no direct control over the cache (which is managed completely by the hardware), software optimization must take into account the processor's cache operations in order to maximize the speed and minimize the energy cost associated with data movement. The Python matrix multiplication example in Figure 50 underlines the importance of cache management optimization: while parallelization provided a speedup factor of 7.8x, cache optimization provided a boost of 18.4x.

Challenges and solution pathways for languages, compilers, and runtime systems

Compiler optimizers are constantly improving through normal software management practices. Further opportunities for improvement in compiler optimization include auto-parallelizing and auto-vectorizing optimizations. Machine learning techniques may be able to improve compiler optimization to improve the speed and thus energy efficiency of executable code. Wang and O'Boyle (2018) provided a comprehensive review of research in applying machine learning in compilers and runtime systems, summarized in Table 60. Some studies have reported success in applying machine learning to discover improved optimization heuristics or fine-tune known heuristics. Others have used machine learning to optimize use of the myriad compiler optimization flags of popular compilers, such as gcc, to achieve the highest speedup factor.

Table 60. Machine Learning Methods in Compiler and Runtime Design. *Source: Wang and O'Boyle 2018.*

Approach	Problem	Application Domains	Models
Supervised learning	Regression	Useful for modeling continuous values, such as estimating execution time, speedup, power consumption, latency, etc.	Linear/non-linear regression, artificial neural networks (ANNs), support vector machines (SVMs)
Supervised learning	Classification	Useful for predicting discrete values, such as choosing compiler flags, #threads, loop unroll factors, algorithmic implementations, etc.	K-nearest neighbor (KNN), decision trees, random forests, logical regression, SVM, Kernel Canonical Correlation Analysis, Bayesian
Unsupervised learning	Clustering	Data analysis, such as grouping profiling traces into clusters of similar behavior	K-means, Fast Newman clustering
Unsupervised learning	Feature engineering	Feature dimension reduction, finding useful feature representations	Principal component analysis (PCA), autoencoders
Online learning	Search and self-learning	Useful for exploring a large optimization space, runtime adaption, dynamic task scheduling where the optimal outcome is achieved through a series of actions	Genetic algorithm (GA), genetic programming (GP), reinforcement learning (RL)

Because the effective use of parallel computing resources depends on the workload at runtime, which may not be accurately estimated at compile-time, schemes to efficiently manage runtime resource allocation are an important element of overall performance optimization.

Another important and promising direction for optimization is the direct generation of source code by machine learning systems. AI assistants for human code-writing are currently experiencing a rapid and widespread uptake in the software developer community. GitHub has introduced an AI tool called GitHub Copilot (GitHub 2024) that provides sophisticated interactive code completion capabilities as well as code generation from textual prompts. As of this writing more than 37,000 businesses have begun using GitHub Copilot in their code development. Other AI-based code generation tools such as ChatGPT are also available.

A recent survey of ML-based code generation approaches (Dehaerne et al. 2022) compared 37 studies from 2016–2022 describing ML-based source code generation: generating original code from a textual description of requirements, generating documentation from code, or translating code between languages. The authors noted that although ML models can generate code, it is often not as optimized or effective as human-written code is. A study comparing human-written to AI-assisted C++ code showed that the human-written code was 15–26% faster on average, and that the difference was greater for more expert programmers, up to 6x faster in some cases (Erhabor et al. 2023).

However, future work promises to address these challenges. Automated source-level rewriting of human-written code for optimization without the use of machine learning has demonstrated some successes. Baziotis, Kang, and Mendes (2023) demonstrated a system called Dias to automatically rewrite code in exploratory Jupyter data analysis notebooks. Dias was able to rewrite individual Jupyter cells to be 57x faster compared to hand-written code calling the Python Pandas library and 1,909x faster compared to the same code calling the Modin library (a drop-in replacement for Pandas that supports parallel processing). Whole Jupyter notebooks were accelerated by up to 3.6x when using the Pandas library and 26.4x using the Modin library. Application of such automated source-level optimization to AI-generated source code is a logical next step. This is an extremely active area of research. Given the boost in programmer productivity afforded by such tools, AI-generated or AI-assisted code development may be used to optimize algorithms for energy efficiency, depending on other system-level constraints.

2.4.5.2 Privacy and Security

Security, at the intersection of software and architecture, is computationally expensive. Of relevance to EES2, from a high-level confidential computing perspective, security is an overhead cost as we seek to improve the overall energy efficiency of computing. Privacy and security solutions must meet needs as efficiently as possible. Current solutions, which put encryption into datapaths to enable a trusted execution environment, have support from AMD, Intel, and NVIDIA. Currently, encrypted computing has limitations related to the management and security of encryption keys and the necessity of multiple encryption/decryption steps for data processing.

To enable privacy and security as a built-in feature of data flow, the internet of the future needs a secure multi-party compute infrastructure that uses homomorphic encryption and private information retrieval (PIR). Homomorphic encryption allows computations to be performed directly on encrypted data—such that computation results are also encrypted, and when they are decrypted, are identical to the results of those computations on the unencrypted data. Homomorphic encryption, therefore, can be used for cloud storage and computation that allows data to be processed in the cloud while remaining encrypted and private (Munjal and Bhatia 2022).

PIR has been a popular research topic since it was first described in 1995 (Hsu et al. 2020). With PIR, when a user makes a query to retrieve information, the request is sent to the service provider using homomorphic encryption. The encrypted response can only be decrypted by the user. There are crypto technologies that can return a desired item without having knowledge of the request content, but the algorithms are extremely computation-intensive (10x to 1,000,000x compared to unencrypted). Nevertheless, because of the demand for privacy, PIR will likely be deployed in at least some internet transactions in the future and will result in a large computational workload for those transactions.

Challenges and solution pathways for privacy and security

There are many trade-offs between computational cost and security/privacy guarantee in implementing homomorphic encryption and PIR with different algorithms. For example, PIR can be implemented in a single server using a distributed point function or in two servers using two-level homomorphic encryption, and the two alternatives will have different costs.

In the design space for homomorphic encryption, there are even more trade-offs, for example, supporting additive homomorphic encryption, using integer or floating-point algorithms, and other design possibilities. The myriad options available must be evaluated to determine practical solutions for different applications. Accompanying these design trade-offs are differences in energy consumption.

A high degree of parallelism in the computations is needed to make PIR practical. Currently, implementations of PIR use conventional hardware (e.g., CPUs, GPUs). Ultimately, for both speed and energy efficiency, domain-specific architectures should be developed, and the hardware and software should be co-designed. As of 2024, at least six companies are testing or commercializing the first chips implementing homomorphic encryption (Moore 2024).

2.4.5.3 Computational Reliability

In general, increased reliability requires more energy, a kind of “tax” on the system. Computational reliability encompasses a broad range of technologies and techniques in modern computer systems. Algorithms and software used to increase reliability include various forms of RAID, N-way replication, two-phase commit, and active-passive configurations. There are also protocols for dealing with hardware failure due to age or external influences like radiation-induced bit flip or array failure.

Computational reliability has been recognized as an important topic for decades. Well-established industry groups study the issues and set standards and requirements for fault management, such as the Open Compute Project (OCP) working groups on fleet-scale memory fault management and silent data corruption errors.

Bit errors may be the result of electromagnetic interference but are most commonly induced by cosmic rays. Error correction coded (ECC) memory is a type of DRAM used in data centers, servers, and generally any application where high reliability is critical. (It is typically not used in personal computers.) ECC memory uses additional non-data “parity” bits to encode the data bits in a Hamming code (Hamming 1950) or a triple modular redundancy code (Shooman 2002) in a way that permits detection of errors and reconstruction of the correct data if an error occurs. ECC memory may be implemented using an extra DRAM on a memory module containing the parity bits, or with the parity checking on-chip. “Chipkill,” an IBM-specific technology, is a more effective version of ECC that also corrects for multiple bit errors, including the loss of an entire

memory chip. This is accomplished by spreading the bits of a Hamming-coded word across multiple chips, so that the word can be reconstructed even if an entire chip fails.

Error checking is also present in CPU cache memories, typically with a single bit error detection capability in the L1 cache backed up by a single bit error correction encoding in the L2 cache. If a 1-bit L1 cache error is detected, it can be refreshed from the L2 cache.

For very large systems subject to the combined error rate of all components (such as data centers and supercomputers), or systems that run critical applications (such as financial transaction processing), further protection from errors is achieved by checkpointing, in which the system or application state is periodically backed up to non-volatile memory. In the event of an uncorrectable upset, the system or application can restart from the most recent checkpoint rather than starting over from the beginning or completely rebooting.

For non-volatile storage, redundant array of independent disks (RAID) technology is conceptually similar to the triple modular redundancy or Chipkill technologies discussed above. In this scheme, multiple drives contain multiple copies of the data or parity calculations from the data, enabling full data recovery in the event of a failure of a drive sector or the whole drive.

Not every transistor in a computer can be protected from state errors. Silent data corruption—when data errors go undetected by the larger system—is a widespread problem for large-scale infrastructure systems. It can propagate through erroneous computations and manifest as application-level problems. It can also result in data loss and can be difficult to debug and resolve. Dixit et al. (2021) described best practices for detecting and remediating silent data corruptions, finding that reducing silent data corruption requires not only hardware resiliency and production detection mechanisms, but also robust fault-tolerant software architectures.

Challenges and solution pathways for computational reliability

Future improvements in computational reliability are likely to come from refinements in both the degree of protection and the methods for handling failures. Some challenges and potential solutions are as follows.

Application checkpointing

While ECC protection is implemented in hardware, application checkpointing (periodic saving of the state of a computation to use as a restart point in case of a failure) is managed by software and is an active area of development. Google recently reported developing a checkpointing scheme used in very large-scale LLM training (a system with more than 50,000 TPUv5e chips organized in UCIe-interconnected pods, with 256 chips per pod) that boosts efficiency 150x by loading checkpoints in a single pod and broadcasting the checkpoint to all other pods, rather than have each pod separately load the checkpoint data. Industry is actively developing more optimizations like this.

Single-event upsets and other sources of random error are stochastic processes, whereas checkpointing algorithms are deterministic or nearly so. An energy cost can be ascribed to the overhead necessary to implement checkpoints and a (stochastic) energy cost can be ascribed to lost work in the event of an error. These two costs can be subjected to formal minimization analysis that can be used to reduce overall energy cost.

Combining reliability and security

For memory, there is a new Open Compute sub-workstream called Fleet-scale Memory Fault Management, a spinoff of the Hardware Management workstream. There are possible options under investigation by this industry group for co-design of both security and reliability features (Aiken et al. 2021) for reduced overhead energy.

2.4.5.4 Communication Protocols

Software can increase energy efficiency by minimizing communication overhead, and in some circumstances software can be used to enable efficient workload data flow. For example, disaggregated resources (e.g., memory) could be used as a more convenient place to store data which is processed incrementally (e.g., weights in a neural network), with the need only to pass pointers instead of passing blocks of data across the interconnect.

Communication protocols such as the NVIDIA Collective Communication Library (NCCL) aim to streamline common communication patterns for AI workloads (NVIDIA 2023b). NCCL implements multi-GPU and multi-node communication primitives optimized to achieve high bandwidth and low latency over PCIe and NVLink high-speed interconnects (Jeaugey 2019). Development frameworks such as PyTorch and TensorFlow have integrated NCCL to accelerate deep learning training on multi-GPU systems.

Microsoft has implemented its own Azure-based platform on top of NCCL known as Microsoft Collective Communication Library (MSCCL), described as “an inter-accelerator communication framework”. It provides programmable communication algorithms for inter-connection among accelerators with different latencies and bandwidths

Challenges and solution pathways for communication protocols

The driving challenge for communication protocols are in estimating quantitatively the trade-offs between speed (performance) and reliability. Standardization and widespread adoption will benefit future system development. The EES2 community can play a role in this effort by ensuring that energy efficiency is a consideration in such standardization efforts.

Action plan for software for conventional architectures

Table 61. Action Plan for Software for Conventional Architectures.

Scope			
Technical Challenge for Energy Efficiency	Software for Conventional Architectures		
Technologies of Interest	• Programming systems, including compilers, languages, runtime libraries • Privacy and security • Communication protocols • Computational reliability		
Challenges		**Solution Pathways**	
• Improvement in compiled code performance without specialized effort by programmers • Reduction of energy cost associated with application checkpointing • Energy-efficient implementation of privacy protocols		• Use machine learning in source code and compiled code optimization • Develop stochastic optimization of checkpointing for energy efficiency • Consider energy efficiency trade-offs in PIR implementation • Use co-design methods to design for efficiency	
Major Tasks / Milestones	**Metrics**	**Targets**	**Timeline**
Compilers and runtimes: Improved profiling tools for optimization	Level of detail and usability of automated code analysis	20% reduction in coding time to implement optimization	2–3 years
Languages and compilers: AI code generation wizards	Coding time; accuracy and efficiency of coded algorithms;	20% improvement in speed of generated code	4–5 years

	suitability for IR code optimization by compilers		
Compilers and runtime libraries: ML approaches to discovering compiler heuristics	Improvement in execution speed of generated code	10% improvement on speed of generated code	2–3 years
Computational reliability: checkpointing energy optimization studies	Overall energy cost of checkpointing	15% improvement	4–5 years
Communication protocols: Promote standardization of NCCL communication framework	Adoption rate for new design projects	>90%	2–3 years
Privacy and Security: Perform energy efficiency trade-offs for candidate PIR implementations	Energy efficiency or energy cost per transaction	TBD	5–7 years

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Compiler Developers	Implement optimization improvements
Hardware Suppliers	Provide improved profiling tools
Data Center Operators	Support modeling and simulation
Academia	Demonstrate prototype software; support modeling and simulation
National Laboratories	Demonstrate prototype software; support modeling and simulation; demonstrate improved checkpoint in HPC centers
Government	Provide targeted funding opportunities to stimulate work

Required Resources	Cross-Collaboration with Other Working Groups
• Research funding is needed for improved compiler systems. • Human capital is needed for participation in standards bodies and working groups, bringing an energy efficiency focus to their work.	Education & Workforce Development: Promote funding for studies at universities; catalyze improved energy efficiency coursework.

2.4.6 Software for Domain-Specific and Emerging Architectures

This section is focused on software challenges and opportunities tied to computer architectures outside of the traditional von Neumann architecture. For the purposes of this discussion, “emerging architectures” refers to those that are not currently in commercial use. Some of the software issues related to emerging architectures have already been discussed in Chapter 2.2. Likewise, many of the solutions outlined in the previous section for “conventional” architectures, such as compiler optimization, data compression, data type precision, and communication protocols, can also be applied to domain-specific and other emerging architectures. This section highlights some additional software related opportunities for energy efficiency gains in emerging architectures.

For decades, the imbalance of compute speed to memory bandwidth, as measured by the number of computations the machine can perform in the time it takes to read a data value from memory (see Figure 52), has made it ever more difficult to ignore communication costs. Beyond just increasing the size of hardware caches, new algorithms must be designed to minimize and hide communication, sometimes at the expense of duplicating memory and computation (Dongarra et al. 2020).

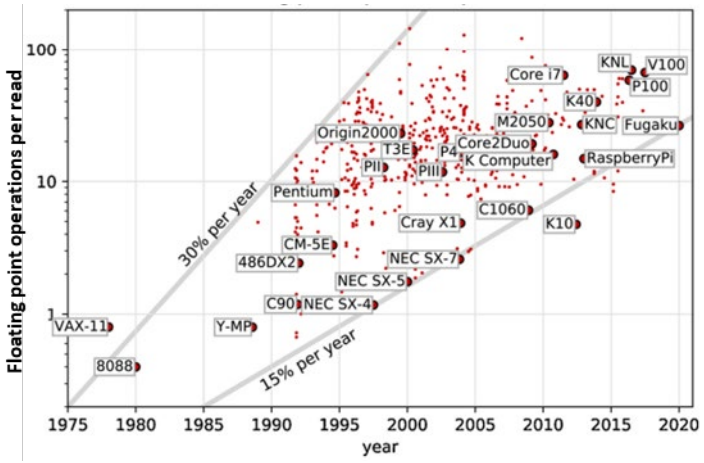


Figure 52. Growing machine imbalance over time.*Source: Dongarra et al. 2020*

Today, architecture and software go hand-in-hand to implement new capabilities in computer systems. Hennessy and Patterson (2019) have called this a new "golden age of computer architecture" in which, at the end of Moore's Law, progress will be made by exploring many new architectural concepts beyond the von Neumann architecture. Domain-specific architectures (DSAs) serve as an illustrative example. By providing combinatorial rather than sequential logic to perform the computation, and by orchestrating efficient data flow with *a priori* knowledge of what the application demands, DSAs enable radically improved performance in the targeted domains compared to what is achievable with a general-purpose CPU. To realize this potential for improved performance, however, the support software system must be able to provide high-level abstracted access to manage the unique aspects of the specific architecture and design. DSAs can partially address machine imbalance by managing data movement more efficiently, but more advanced research is needed.

The machine imbalance problem is exacerbated by the most significant trend in computing today: the exponential increase of machine learning applications that, by their nature, are extremely data intensive. Machine learning algorithms such as generative AI consume a large amount of energy as a direct result of data movement (Sze et al. 2017). Compute In-Memory (with some architecture changes) would be a good candidate to implement low power inferencing for cloud as well as for edge devices, as this could reduce data movement. Cao et al. (2023) and Gozalo-Brizuela and Garrido-Merchán (2023) provide a survey of generative AI as well as use cases.

The following subsections describe some challenges for software developers for emerging architectures and potential solutions for improved energy efficiency.

2.4.6.1 Domain-Specific Languages

Domain-specific languages (DSLs) are a natural fit to work with domain-specific architectures (DSAs), although their application is by no means limited to new architectures. DSLs are appropriate for application domains in which the most used operations can be expressed as high level operators and are then amenable to intermediate representation techniques to target and optimize for the specific hardware that is to be used. A familiar example of a DSL is Structured Query Language (SQL), a language explicitly designed for manipulating relational

databases. Domain-specific languages can make expression of hardware operations and programmer intent more natural and straightforward. At the same time, DSLs can borrow heavily from the syntactic and semantic idioms of popular general-purpose languages to reduce the learning curve for programmers. Nevertheless, the learning curve for any new language can limit its application to a wide variety of problems. The alternative to a DSL is a framework (system of libraries and an associated runtime system) implemented in a general-purpose language.

The trade-offs between these two alternatives, frameworks and DSLs, are illustrated by two popular programming systems for machine language systems widely used today: PyTorch and TensorFlow. PyTorch is an open-source machine learning framework based on the Torch library with strong support for tensor computing and deep neural networks, both of which have significant matrix-vector operations. The tensor (multi-dimensional array) computational workflow can run on NVIDIA GPUs through the NVIDIA CUDA parallel processing compiler. Originally developed by Meta (formerly Facebook), PyTorch is now managed by a non-profit foundation as part of the Linux Foundation.

TensorFlow is also an open-source library (Abadi 2016) but is more properly viewed as a DSL whose computations are expressed as “stateful dataflow graphs,” a data-centric intermediate representation that enables separating program definition from its optimization (Ben-Nun et al. 2019). The TensorFlow language was developed concurrently with the Google tensor processing unit (TPU) in a true example of hardware/software co-design.

Compiler infrastructure has evolved in response to the trend toward specialized DSLs targeting DSAs. The LLVM compiler infrastructure mentioned in section 2.4.5.1 is a suite of libraries enabling multiple language front ends to be represented in a common intermediate representation for optimizations and then targeted to multiple machine-specific back ends. A significant advancement is the introduction of multi-level intermediate representations (MLIR). MLIR simplifies the process of mapping programmatic constructs from DSLs or frameworks directly to DSAs (Lattner and Pienaar 2019). As depicted in Figure 53a, the compiler architecture allows for the integration of multiple programming languages. Initially, source code is processed into an abstract syntax tree (AST), followed by a language-specific intermediate representation (IR) that supports unique features such as novel data types.

In the case of TensorFlow, the front end produces abstract data flow graphs which are translated to the high level operations (HLO) intermediate language and optimized by the accelerated linear algebra (XLA) optimizer. The outcome of these language-specific optimizers is then lowered to the LLVM IR for further optimization and code generation for target hardware (which may yet have additional machine-specific optimizations). The aim of MLIR is to provide a super-extensible system that allows DSLs to lower naturally to MLIR and LLVM IR, accelerating innovations in hardware, compiler algorithms, and high-level abstractions.

For emerging architectures, DSLs allow a path for integration into an existing ecosystem and likely help them target which operators are most important to examine. They also set a bar for measuring the benefit of the emerging architecture, independently of marketing claims. For example, neuromorphic architecture holds enormous promise for building more powerful and efficient machine learning systems, yet it has been difficult to integrate these architectures with runtime systems to enable problem-solving. The use of IR can help decouple the evolution of neuromorphic hardware and software, ultimately increasing the interoperability between

platforms and improving accessibility to neuromorphic technologies as shown in Figure 53b (Pedersen et al. 2023). The potential for DSLs to unlock practical use of these architectures is worthy of extensive study.

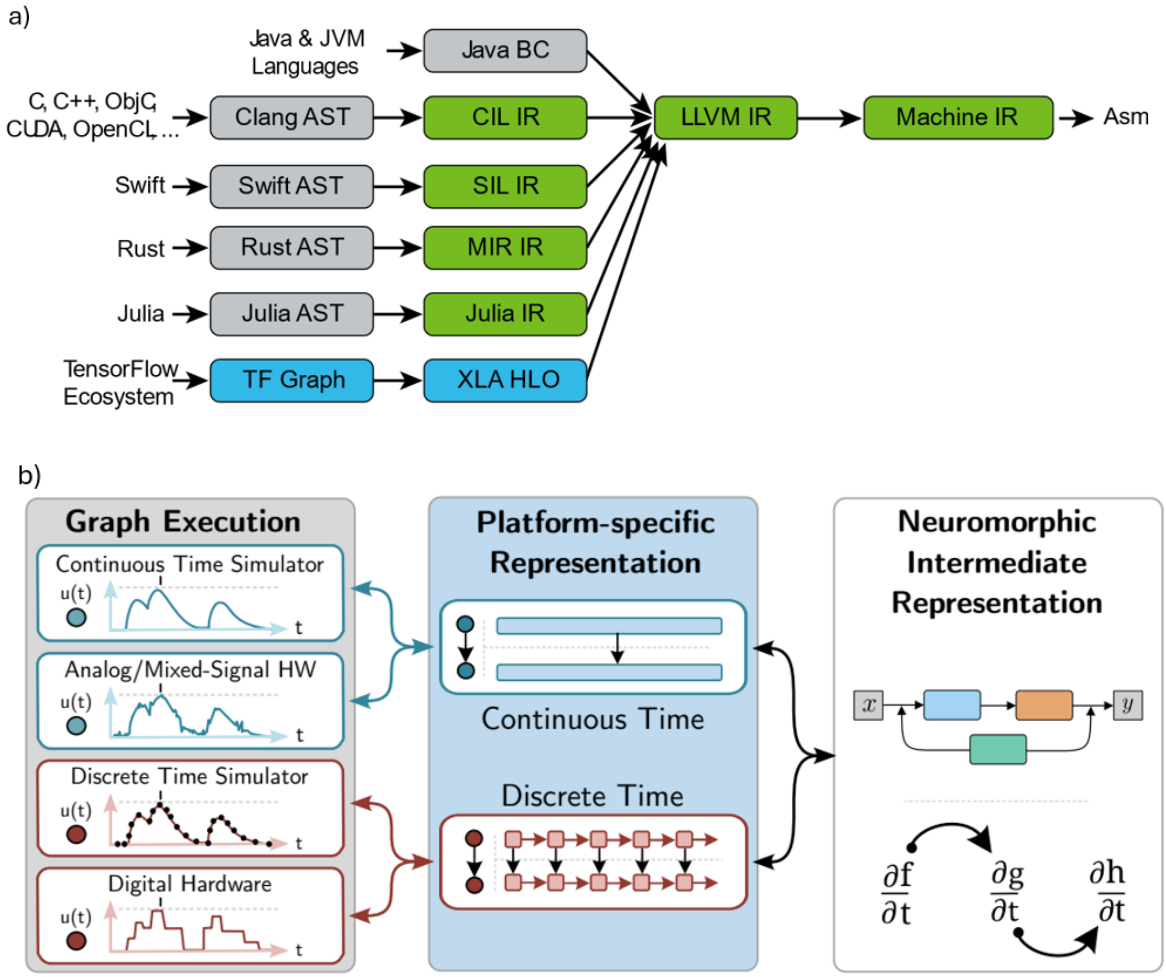


Figure 53. Neuromorphic intermediate representation. *a) MLIR compilation process through language-specific intermediate representations. b) Example of IR that allows for continuous-time representation of nodes that can then be executed on continuous-time hardware or simulators, or discretized for use on discrete-time hardware or simulators. Source: (a) Lattner and Pienaar 2019; (b) Pedersen et al. 2023.*

2.4.6.2 Adoption of Existing Compute Cores in Domain-Specific Architectures

Even though DSAs are tailored for domain-specific workloads, the modern design approach allows DSAs to tap into well-developed software ecosystems, depending on the overlap with existing architectures. Licensable processor cores, such as Arm and x86, and open-source RISC-V can be incorporated into the DSA chip design, thereby gaining access to operating systems, compilers, and high-level applications with comparative ease. The minimal RISC-V core can be implemented in as few as 15,000 gates (Lattner 2021), making its incorporation a very low burden on a custom chip with billions of transistors. Figure 54 shows an example of this design approach for an experimental neuromorphic computer architecture with an embedded RISC-V core.

The pace of software innovation will become a limiting factor unless it can keep up with the pace of architectural innovation. There is room for further maturation and standardization to enable existing software ecosystems to be seamlessly recompiled and executed on new DSA architectures. Development of a fairly standardized design framework for the interface between custom accelerator architectures and the higher-level environment will streamline software development for these accelerators using principles of co-design.

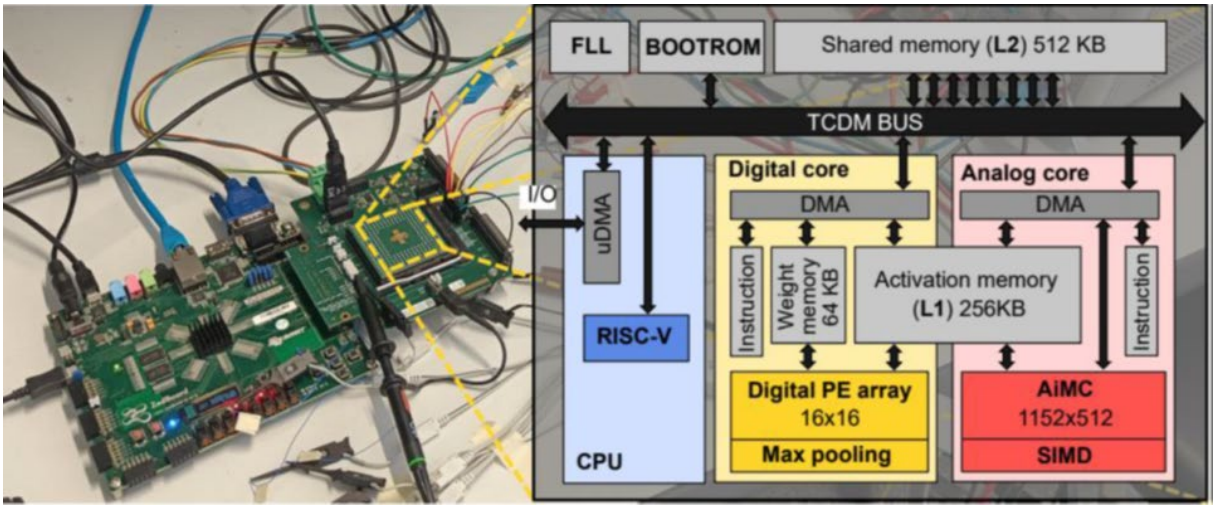


Figure 54. An example neuromorphic computer architecture with embedded RISC-V processor. *Source: MICAS 2023*

2.4.6.3 Reusable Memory Access Control Architecture

One challenge in typical design practice today is for each accelerator to have its own bespoke memory hierarchy and associated circuitry design. This follows quite naturally from one of the main motivations of custom accelerators, which is to take explicit control of data flow as specifically demanded for a specific planned workload. Yet there is potential to accomplish this control while adhering to a reusable memory access control architecture.

The memory "buffets" concept (Pellauer et al. 2019) provides for explicit, composable data transfers between a custom chip and the external memory. Access requests are decoupled from the request receiver, thereby reducing or eliminating the need for on-chip buffering. The design of buffets has been publicly released in RTL code and is flexible enough to fulfill the needs for memory access architecture in a variety of use cases. Such flexibility in efficient memory access could facilitate acceleration of sparse matrix math operations, taking advantage of the data sparsity that is not accommodated well in current accelerators by avoiding time and energy expended transferring mostly zeroes and instead transferring only nonzero data.

2.4.6.4 CIM, Neuromorphic Computing, and Spiking Neural Networks (SNNs)

The successes of large neural network models have spurred innovation in machine architectures aimed explicitly at neural network processing. Many architectural adaptations have been made in GPU design, as well as in FPGA and TPU accelerators (as discussed in Chapter 2.2), to process neural networks more efficiently. To achieve still greater energy efficiency and performance, computer architects have turned to brain-inspired "neuromorphic" computing designs, which mimic observed features of biological brains in silicon. Figure 55 illustrates two such architectural paradigms being actively explored as alternatives to the von Neumann

architecture and their associated challenges. Although the term “neuromorphic computing” most often refers to architectures that include inter-neuron signaling via variable-timed spikes in “spiking neural networks” (SNNs), all neural networks are brain-inspired. The term “neuromorphic” can refer to a wide class of architectures, including all-digital designs and mixed-signal designs with some analog circuitry as part of the network, particularly in the form or programmable resistance elements to represent the model weights. The existing and emerging device technologies for neuromorphic computing were discussed in section 2.1.7, while emerging circuit architectures for both digital and analog compute-in-memory were discussed in sections 2.3.2 and 2.3.3, respectively.

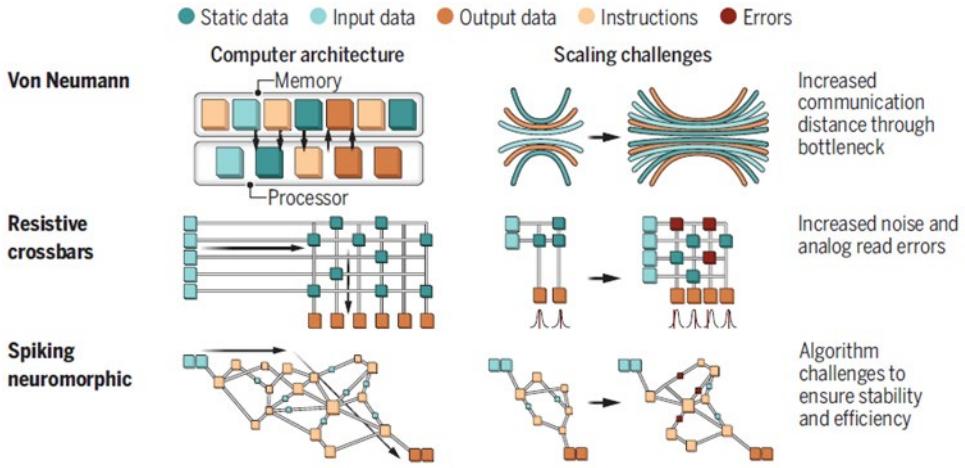
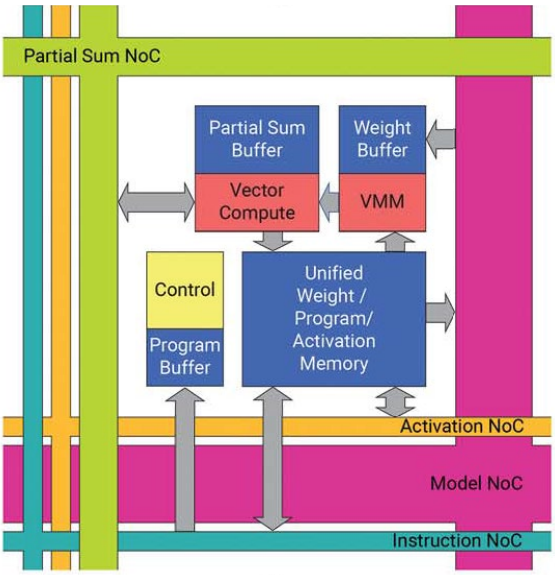


Figure 55. Von Neumann, resistive crossbar, and spiking neuromorphic architecture paradigms and challenges.*Source: Aimone and Agarwal 2024*

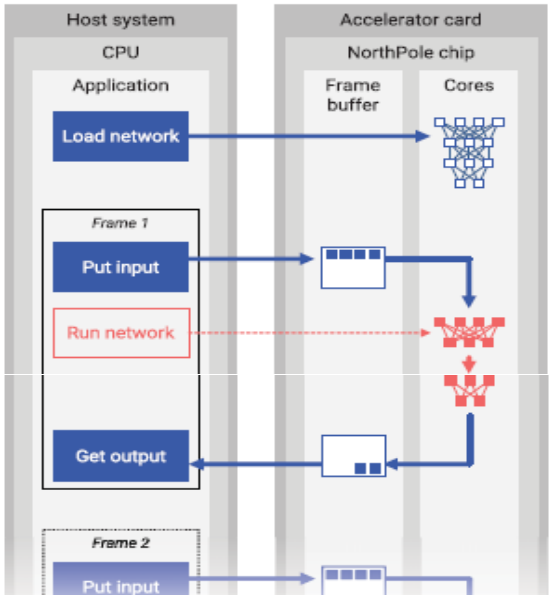
The common feature of brain-inspired or neuromorphic architectures is reorganization of the compute and memory elements to situate them as close as possible to one another to minimize the distance that data must move. Because most ML networks have static hyperparameters (once trained and optimized), compute-in-memory provides an alternative to massive transfers of data by storing the network hyperparameters (weights/kernels) within the memory/compute array where MAC (multiply-accumulate) operations take place. This reduces the neural network model’s memory transfers during runtime. Storing the network parameters once as part of the initialization of the compute array means there is no need for memory transactions during run-time, which results in lower latency as well as lower power consumption per inference.

Digital compute-in-memory (CIM)

CIM is a promising avenue to alleviate data movement bottlenecks but poses challenges for implementation in software. Digital CIM is not an extreme departure in terms of an architecture: it is moving compute closer to or in memory within a digital architecture. Some examples of digital CIM are already commercially available, including UPMEM (UPMEM 2023) and the IBM NorthPole architecture (Modha et al. 2023). In the case of UPMEM, the operators look like hardware instructions or programs with accelerator semantics. Essentially traditional compute done with different tradeoffs, this requires avoidance of certain OS and system architecture rules. In the case of NorthPole (see Figure 56), the chip runs a network model with its own custom-defined instruction set but appears to the host processor as an active memory with just 3 commands (write inputs, run network, read results) and the minimum possible I/O bandwidth. NorthPole is a brain-inspired, all-digital inference engine exploiting CIM specifically optimized for neural network performance. Each of its 256 cores is capable of massively parallel neural network computations (2,048, 4,096, or 8,192 operations per cycle for 8-, 4-, and 2-bit precision, respectively), with memory in each core intimately intertwined with the computing circuits. As a result, it has demonstrated 25x greater energy efficiency (measured in frames per second per watt) for the ResNet50 image classification benchmark, compared to energy efficiency performance with an H100 GPU.



ıematic of an individual core (one of 256)



(b) Host interface

Figure 56. IBM NorthPole digital neuromorphic chip.*Source: Modha et al. 2023*

Analog CIM

Analog crossbar multiplier arrays (see section 2.1.7), an alternative to digital matrix multiplication for linear algebra operations, are the foundation of existing machine learning algorithms. They offer the possibility of dramatic reductions in energy (see section 2.2.3), but have thus far found limited application due to the low precision achievable with analog circuitry. However, as Aimone and Agarwal (2024) have pointed out, precision is an emergent property of digital circuit design because individual transistors have only single-bit precision. Song et al.

(2024) recently demonstrated a method combining architecture and algorithm to achieve arbitrarily high precision with analog crossbar arrays, as depicted in Figure 57. The method dedicates subsequent crossbars to address the residual error (the difference between desired and realized precision) to reach the overall desired precision while maintaining a substantial energy advantage over conventional digital operations. This approach may not only enable more energy-efficient neural network processing but it also may be applied to more conventional numerical analysis tasks that require high-precision matrix multiplication.

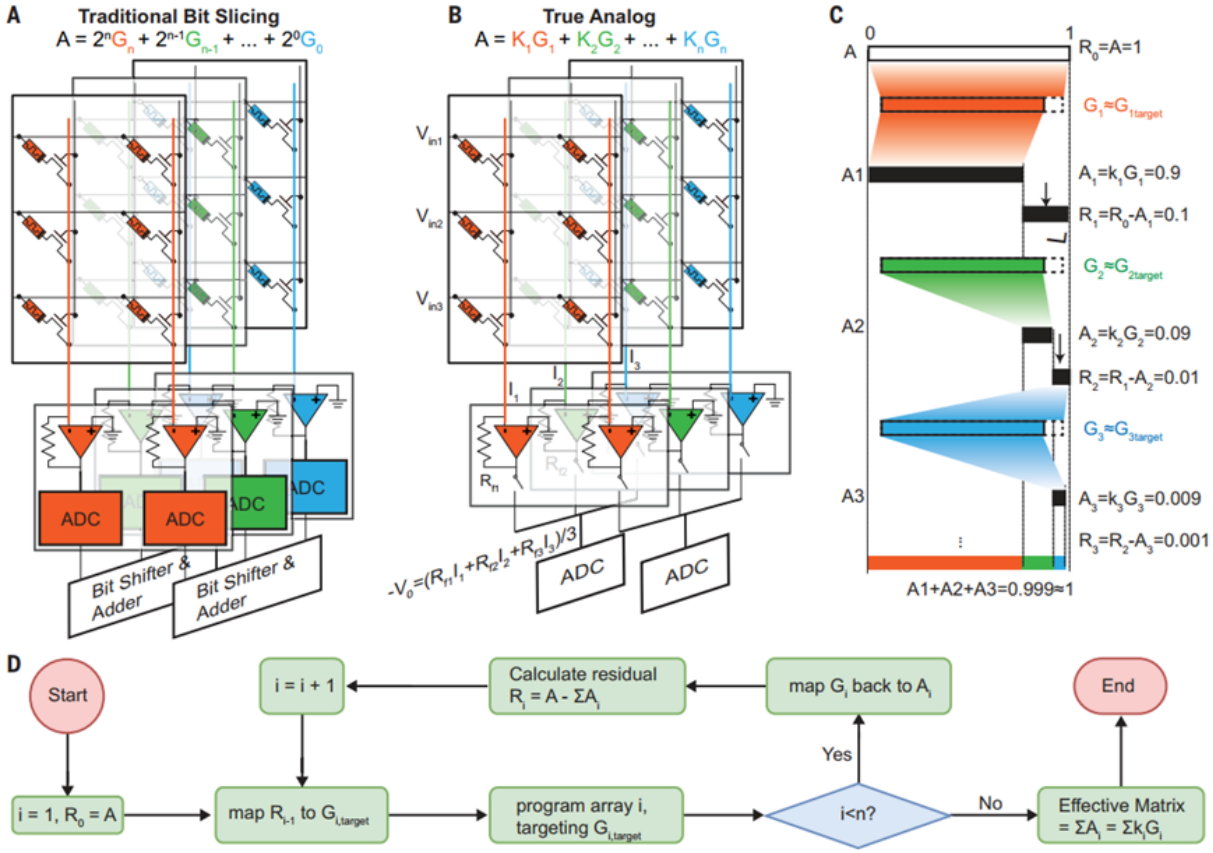


Figure 57. Architecture and algorithm to achieve arbitrarily high precision with analog crossbar multipliers. *(a) Traditional crossbar arrays with ADCs and additional postprocessing circuits; (b) proposed arbitrary precision programming circuit with shared ADCs; (c) example of programming a numerical value A=1 into multiple memristor devices step by step; (d) flowchart of the arbitrarily high-precision programming algorithms. Source: Song et al. 2024.*

Spiking neural networks (SNNs)

Prominent examples of spiking neuromorphic computers include the SpiNNaker computer of the Human Brain Project (Human Brain Project 2023), the Intel Loihi project (Intel 2023b), and the IBM TrueNorth project (Akopyan et al. 2015). Analog neuromorphic computing using ReRAM or silicon photonics, Ising rings, etc., offers huge potential gains in energy efficiency, but the discovery of effective algorithms for training SNNs has proven to be a difficult software challenge. The standard back-propagation method of training other neural networks is incompatible with SNNs. Alternative approaches for training have included strategies to approximate the back-propagation algorithm, to convert networks trained on conventional DNNs to run on SNNs, and to train SNNs directly through evolutionary algorithms (simulating evolution

through survival of the fittest, reproduction with intermixing, and mutation) (Schuman et al. 2022). Despite the difficulties in training them, SNNs have demonstrated some success. For example, the Intel Loihi consumes 5–100x less energy than conventional DNNs for keyword spotting in speech recognition. For bio-inspired odor recognition, Intel Loihi is 3,000x more data-efficient than DNNs (Intel 2020).

Challenges and solution pathways for CIM, neuromorphic computing, and SNNs

At present, digital CIM architectures are much better positioned than analog CIM architectures to gain widespread use, but these systems also present challenges for software. It is desirable to implement CIM relatively transparently while shielding hardware details from programmers in the same way the memory hierarchy does. Issues of automatically parallelizing sequential programs, managing the data layout in memory to implement computations, and incorporating in-memory operations within the memory hierarchy logic remain to be addressed at scale.

Although neuromorphic computing is being evaluated widely, there are currently no real-world applications of neuromorphic computing that have exploited native hardware implementation. Many challenges must be addressed to realize the energy efficiency benefits of CIM, neuromorphic, and SNN architectures (Schuman et al. 2022):

- **Widening algorithmic focus:** The lack of good native training methods for SNNs has meant that much of the reported SNN performance has come from applications where conventional software-based neural network solutions already exist, and SNN implementations were mapped from a DNN to the SNN. Further exploration of neuroscience-inspired approaches may yield higher performance networks. The use of SNNs for exploratory neuroscience is itself an important research direction that may yield better understanding of both SNNs and biological brains.
- **Wider availability of machines and simulators:** There have been several high-profile neuromorphic computer systems, as previously mentioned, that have provided access for diverse groups to experiment with. Wider availability of development systems will enable a much larger community to develop, leading to faster discovery of viable algorithms.
- **Enabling use in heterogeneous compute environments:** Neuromorphic compute engines' reliance on the facilities of a host computer may impose overheads that hinder their performance and prevent commercial viability. It is important to achieve the optimal balance between performance and energy efficiency. Careful design of interfaces between neuromorphic chips and other compute elements—especially for edge computing, where the low power requirements of neuromorphic processors are most attractive—is a must.
- **Better benchmarks:** Because it has been difficult to find problems for which neuromorphic computing is particularly well-suited, current benchmarks tend to rely on problems already effectively solved with conventional networks. Articulate important use cases to define appropriate benchmarks for neuromorphic computers.
- **Better programming abstractions:** The fundamental lack of understanding of computational primitives, abstractions, and representations means that further studies for emerging architectures are essential.

2.4.6.5 Data Compression

Data compression reduces the high costs associated with data movement by increasing the information density of the data, potentially aligning with near-term EES2 goals. It is particularly valuable in contexts with slow file I/O systems, long-distance communication links, and significant associated energy costs. When the energy saved by reducing data volume exceeds the energy spent in compressing and decompressing data, the efficiency gains are substantial. However, the benefits are less pronounced within computer systems, such as in data transfers between main memory and virtual memory or between main memory and cache, where the overhead of compression often outweighs the energy savings, resulting in modest compression ratios.

Nevertheless, there has been work in compression. NVIDIA offers a standard compression library called nvcomp that enables compressed I/O in GPU systems (Sakharnykh, LaSalle, and Karsin 2020). ZeroPoint Technologies has introduced an IP product called Cache-MX that provides data compression of cache lines for L2 and L3 caches (but not for the most time-sensitive L1 cache). Cache-MX is an add-in to the last-level cache controller that performs compression, decompression, and compaction to effectively double the (logical) size of the cache, with an increased latency penalty of 9 cycles for a 1.6 GHz (or potentially faster) clock. This compression delivers higher performance for the same power expenditure, either with larger apparent cache size or with chip real estate freed up for other functionality.

Generally, data compression in a computer system must be lossless. But in connection with approximate computing or analog computing architectures, there may be opportunities for higher compression in lossy compression schemes. For data compression to be beneficial, the full accounting must include energy estimates for access, compression, decompression, transmission, and the relevant computation. Although compression may worsen latency, the overall effect may lead to an efficiency gain because fewer bits are transferred after compression.

Exploiting *a priori* knowledge of the information domain of the data being communicated can yield large improvements in compression (Weissman 2022). For example, the algorithms used widely for audio and video compression exploit knowledge of human perceptual processing to achieve much better compression than would be possible without this knowledge. Tsai and Sanchez (2019) proposed a method for compressing software objects using the logical structure of those objects to achieve better compression ratios. Analogously to domain-specific computer architectures, domain-specific compression algorithms may lead to significant reductions in data movement with consequent improvements in energy efficiency.

2.4.6.6 Precision of Data Types

As shown in Figure 7 of the Introduction, higher numerical precision comes with an energy cost. For example, adds of 32-bit floating point (“FP32”) require 5.4 times more energy than adds of 8-bit integers (“INT8”), while multiplies require 18.7 times more energy for FP32 versus INT8 data precision. These comparisons account for only the arithmetic operation itself and not the energy cost of moving the operands to and from memory, which is of course also proportional to precision. Clearly, data precision has an important energy impact.

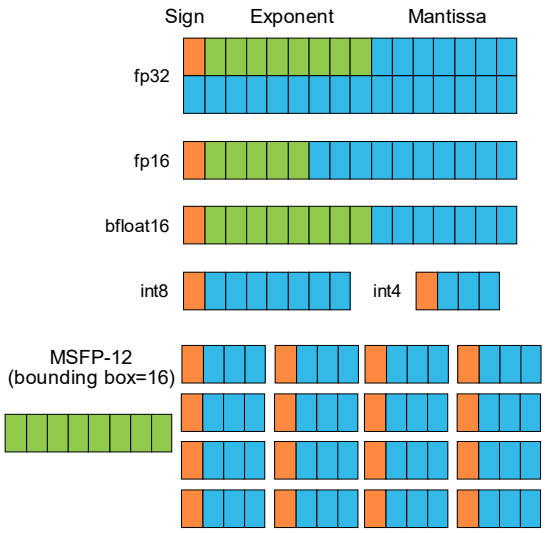


Figure 58. Integer and floating-point numeric representations. *MSFP is a distributed floating-point representation proposed by Microsoft.*

Source: Rouhani et al. 2020

Floating point representation of real numbers in terms of a number of bits of a mantissa, or "significand," with additional bits representing an exponent and a bit representing sign (essentially a binary version of scientific notation), were introduced with the first electromechanical calculators. The IEEE 754 standard (IEEE 2019) for single-precision, double-precision, and quad-precision floating point was issued in 1985, leading to stability and identical results among computers (Athow 2014). There has been continued significant innovation in floating point representation, driven primarily by the massive data throughput required for many ML applications. The "brain float" 16-bit representation that has become widely used in ML applications trades some precision in the mantissa for additional dynamic range in the exponent. A team at the Barcelona Supercomputer Center, together with Intel, has developed a method to achieve higher precision by combining bfloat16 values, thus eliminating the need to implement both bfloat16 and fp32 hardware on a chip (Genkina 2022). Figure 58 illustrates several numeric formats, including a distributed "MSFP" floating point format proposed by Microsoft (Rouhani et al. 2020), in which a single exponent is used in common for a block of mantissa values. This is useful for very hardware-efficient matrix dot-product computations and makes a compromise between the efficiency of integer math, which is subject to underflow or overflow with numerical outliers, and floating point, which has a separate exponent allocated to each value.

Innovation in efficient number representation is ongoing. Opportunities exist for improvement, not only for machine learning but for other applications such as scientific computing, which has typically used higher precision floating point formats. In an effort to increase effective memory bandwidth, a team at Lawrence Livermore National Laboratory is developing floating point compression techniques to discard bits lacking useful information (Hittinger et al. 2019). Further exploration of novel combinations of smaller data types able to realize higher precision, efficient mixed-precision computation, and distributed representation similar to the MSFP format can reduce the memory traffic required to support computations.

Proliferation of numeric type representations also means proliferation of software required for conversion between formats and carries the risk of many different implementations having slight incompatibilities. Therefore, efforts to promote standardization of new, more efficient formats will be beneficial.

2.4.6.7 Tightly Coupled Architecture and Software Co-design

Every novel computing paradigm—quantum computing, neuromorphic computing, biologically inspired computing, optical computing, etc.—must first be reduced to practice sufficiently that a machine architecture can be defined. It is then the job of software to provide appropriate logical abstractions to relate what programmers want to do to the low-level compute paradigm. Energy savings from new architectures are possible only when the new architecture is accompanied by software enabling it to perform its improved functions. Shortening the timescale delivers energy savings sooner, and this can be facilitated by tightly coupled architecture and software co-design.

This is already happening in emerging technologies. For example, IBM already has an assembler and compiler called OpenQASM for their quantum computing hardware (Cross et al. 2022) which has been used to develop and test a robust set of benchmarks (Li et al. 2020). Another more recent example is the NorthPole brain-inspired neural processor developed by IBM (Modha et al. 2023). NorthPole's chip design, aimed at energy-efficient neural inference at the edge, was introduced alongside a full software development suite that includes a compiler, chip simulator, and validator to validate both the compiler and input algorithms. The concurrent availability of these tools should facilitate rapid testing and commercial implementation of solutions using the NorthPole architecture. In general, as illustrated conceptually in Figure 59, software must follow architecture, but the gap between that hardware availability and usable software can be reduced by tightly coupled hardware/software co-design teams. Software may need a longer time to mature in order to adapt the new architecture into the wider computing ecosystem and provide compatibility with existing software interfaces. Co-design early in the process can lead to quicker adoption of new energy-saving architectures and acceleration of overall energy efficiency.

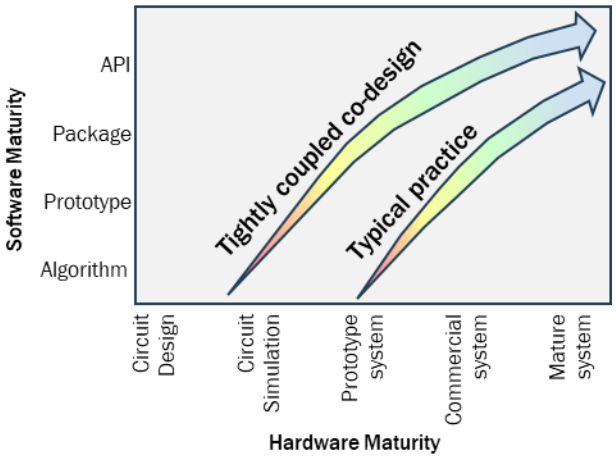


Figure 59. Early engagement between hardware and software designers yields better software sooner.

Action plan for software for domain-specific and emerging architectures

Table 62. Action Plan for Software for Domain-Specific and Emerging Architectures.

Scope	
Technical Challenge for Energy Efficiency	Software for Domain-Specific and Emerging Architectures
Technologies of Interest	• Domain-specific languages • Data buffets • Quantum programming • Neuromorphic programming • Data compression • Hierarchical algorithms for different scales

Challenges	Solution Pathway
• Achieving reduction in energy use through lower-level data compression applications • Ensuring compatibility and efficient interchangeability of emerging numeric data types • Lack of systematic studies on accuracy versus performance	• Support interface standards for domain-specific architectures • Support infrastructure and community to aid development of neuromorphic computing and compute-in-memory software • Develop domain-specific data compression strategies • Promote standardization of data types for information interchange

Major Tasks / Milestones	Metrics	Targets	Timeline
Quantification of energy requirements for larger set of scientific simulations and machine learning algorithms	Identify metrics for different sets of algorithms that span the areas of interest	Measure benchmarks	1–2 years
Support compiler infrastructure for domain-specific languages	Time to implement a working compiler	<3 months	2 years
Promote adoption of data buffet architecture in domain-specific hardware through standards	Number of chips incorporating buffet design	>80% of new designs	3 years
Develop robust software libraries exploiting data buffets	Maturity and functional completeness	Fully implemented libraries in C/C++, Python, Java	3 years
Develop software prototypes for compute-in-memory architecture	Compatibility with language code base	100% compatibility	5–7 years
Discovery of effective training strategies for SNNs	Computational effort to reach target accuracy	No more than other neural network models	7–10 years
Develop high-level programming tools for SNNs	Useful models/use cases supported	Multiple commercially significant models	7–10 years
Proliferate open-source hardware/simulation platforms for SNNs	Availability of development ecosystems	At least one robust development ecosystem	7–10 years
Data compression for domain-specific information	Compression ratio	>50%	3–5 years
Standards for reduced precision/higher efficiency numeric representations	Compatibility of implementations	100% compatibility	5 years

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Industry Groups	formulate and adopt software standards for domain-specific and emerging architecture.
End Users/OEMs	Commercialize new capabilities.
Academia	Explore cutting edge algorithms and architectures.
National Laboratories	Host laboratories and services for the development community.
Government	Provide funding for shared high-performance hardware resources.

Required Resources	Cross-Collaboration with Other Working Groups
• Industry working groups for standardization of interfaces. • Research funding for architectural and software innovation.	Circuits and Architectures: Software opportunities described in this section follow from architectural innovation;close collaboration in software and architecture will yield better and faster results.

2.4.7 Conclusion for Algorithms and Software

Some improvement in energy efficiency can be gained by optimizing common software functions, particularly in making effective parallelization of software more routine. These improvements must remain fully compatible with existing codebases to be acceptable. Although this is a major challenge, the application of new approaches such as machine learning in the

optimization of software may realize major gains. Still greater gains can be achieved in the emerging field of machine learning, where the fundamental limits of algorithmic efficiency are yet to be discovered, as well as in software supporting emerging architectures, where innovative designs continue to be developed for many applications. All software development will benefit from profiling tools that enable programmers to probe energy efficiency of code at a fine-grained level. Those same tools, combined with benchmarking across the major use cases of computing, will enable tracking of industry progress toward EES2 goals.

2.4.8 Algorithms and Software References

Abadi, Martin. 2016. “TensorFlow: Learning Functions at Scale.” In Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming – ICFP 2016, September 1, 2016. http://doi:10.1145/2951913.2976746.

Aiken, A., et al. 2021. “Co-designing from Atoms to Architectures.” In Position Papers for the ASCR Workshop on Reimagining Codesign, edited by James A. Ang, Andrew A. Chien, Si Hammond, Adolfy Hoisie, Ian Karlin, Scott Pakin, John Shalf, and Jeffrey S. Vetter, pg 1–26. March 2021.

Aimone, James B., and Sapan Agarwal. 2024. “Overcoming the noise in neural computing.” *Science*. Vol. 383 (Issue 6685): pg 832–833. https://doi.org/10.1126/science.adn8545.

Akopyan, F., *et al.* 2015. “TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip.” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*. Vol. 34 (Issue 10): pg 1537–1557. https://doi.org/10.1109/TCAD.2015.2474396.

Ansótegui, Carlos, et al. 2021. “Learning to Optimize Black-Box Functions with Extreme Limits on the Number of Function Evaluations.” Presented at Learning and Intelligent Optimization: 15th International Conference, LION 15. Athens, Greece. Springer International Publishing. https://doi.org/10.1007/978-3-030-92121-7_2.

Athow, Desire. 2014. “Pentium FDIV: The processor bug that shook the world.” TechRadar. Published October 30, 2014. https://www.techradar.com/news/computing-components/processors/pentium-fdiv-the-processor-bug-that-shook-the-world-1270773.

Barrett, R.F., T.H.F. Chan, E.F. D’Azevedo, E.F. Jaeger, K. Wong, and R.Y. Wong. 2010. “Complex Version of High Performance Computing LINPACK Benchmark.” *Concurrency and Computation: Practice and Experience*. Vol. 22 (Issue 5): pg 573–587. https://doi.org/10.1002/cpe.1476.

Baziotis, Stefanos, Daniel Kang, and Charith Mendis. 2023. “Dias: Dynamic Rewriting of Pandas Code.” arXiv. Submitted March 28, 2023. https://doi.org/10.48550/arXiv.2303.16146.

Ben-Nun, Tal, Johannes de Fine Licht, Alexandros N. Ziogas, Timo Schneider, and Torsten Hoefler. 2019. “Stateful Dataflow Multigraphs: A Data-Centric Model for Performance Portability on Heterogeneous Architectures.” SC ’19: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Denver, CO. https://doi.org/10.1145/3295500.3356173.

Bentley, Jon. 1984. “Programming Pearls: Perspective on Performance.” *Communications of the ACM*. Vol. 27 (Issue 11): pg 1087–1092. https://doi.org/10.1145/1968.381154.

Blalock, Davis, and John Guttag. 2021.“Multiplying Matrices Without Multiplying.” *Proceedings of the 38th International Conference on Machine Learning*. arXiv. Submitted June 21, 2021. https://doi.org/10.48550/arXiv.2106.10860.

Bondarev, Mikhail. "Energy Consumption of Bitcoin Mining." International Journal of Energy Economics and Policy 10, no. 4 (2020): 524-529. https://doi.org/10.32479/ijeep.9276.

Cao, Yihan, et al. 2023. “A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT.” arXiv. Submitted March 7, 2023. https://arxiv.org/abs/2303.04226.

Cass, S., and H. Goldstein. 2023. “How Python Swallowed the World: Lessons from Compiling Top Programming Languages.” IEEE Spectrum. Vol. 60 (Issue 9): pg 2. Published September 5, 2023. https://spectrum.ieee.org/python.

Cross, Andrew, Ali Javadi-Abhari, Thomas Alexander, Niel de Beaudrap, Lev S. Bishop, Steven Heidel, Colm A. Ryan, et al. 2022. “OpenQASM3: A broader and deeper quantum assembly language.” *ACM Transactions on Quantum Computing*. Vol. 3 (Issue 3). https://doi.org/10.48550/arXiv.2104.14722.

Cuomo, Salvatore, et al. 2022. “Scientific machine learning through physics–informed neural networks: Where we are and what’s next.” *Journal of Scientific Computing.* Vol. 92 (Article no. 88). https://doi.org/10.1007/s10915-022-01939-z.

Curnow, H.J., and B.A. Wichmann. 1976. “A synthetic benchmark.” *The Computer Journal*. Vol. 19 (Issue 1): pg 43–49. http://dx.doi.org/10.1093/comjnl/19.1.43.

Dance, Gabriel J.X., Tim Wallace, and Zach Levitt. "The Real-World Costs of the Digital Race for Bitcoin." The New York Times, April 9, 2023. https://www.nytimes.com/2023/04/09/business/bitcoin-mining-electricity-pollution.html.

de Vries, Alex. "Bitcoin's Growing Energy Problem." Joule 2, no. 5 (May 16, 2018): 801-805. https://doi.org/10.1016/j.joule.2018.04.016.

De Vries, Alex. 2023. “The Growing Energy Footprint of Artificial Intelligence.” *Joule*. Vol. 7 (Issue 10): pg 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004.

Dehaerne, Enrique, Bappaditya Dey, Sandip Halder, Stefan De Gendt, and Wannes Meert. 2022. “Code Generation Using Machine Learning: A Systematic Review.” *IEEE Access*. Vol. 10: pg 82434–82455. https://doi.org/10.1109/ACCESS.2022.3196347.

Dixit, Harish Dattatraya, Sneha Pendharkar, Matt Beadon, Chris Mason, Tejasvi Chakravarthy, Bharath Muthiah, and Sriram Sankar. 2021. “Silent Data Corruptions at Scale.” arXiv. Submitted February 22, 2021. https://doi.org/10.48550/arXiv.2102.11245.

Dongarra, Jack, Mark Gates, Piotr Luszczek, and Stanimire Tomov. 2020. “Translational process: Mathematical software perspective.” *Journal of Computational Science*. Vol. 52: 101216. https://doi.org/10.1016/j.jocs.2020.101216.

Eggensperger, Katharina, Philipp Müller, Neeratyoy Mallik, Matthias Feurer, René Sass, Aaron Klein, Noor Awad, Marius Lindauer, and Frank Hutter. 2022. “HPOBench: A Collection of Reproducible Multi-Fidelity Benchmark Problems for HPO.” Presented at the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. https://doi.org/10.48550/arXiv.2109.06716.

EIA. “State Electricity Profiles 2022.” U.S. Energy Information Administration. Last modified November 2, 2023. Accessed May 9, 2024. https://www.eia.gov/electricity/state/.

EIA "Tracking Electricity Consumption from U.S. Cryptocurrency Mining Operations." U.S. Energy Information Administration. Last modified February 1, 2024. Accessed May 9, 2024. https://www.eia.gov/todayinenergy/detail.php?id=61364.

Erhabor, Daniel, Sreeharsha Udayashankar, Meiyappan Nagappan, and Samer Al-Kiswany. 2023. “Measuring the Runtime Performance of Code Produced with GitHub Copilot.” arXiv. Submitted May 10, 2023. https://doi.org/10.48550/arXiv.2305.06439.

Fanara, A., E. Haines, and A. Howard. 2009. “The State of Energy and Performance Benchmarking for Enterprise Servers.” In *Performance Evaluation and Benchmarking*, edited by Raghunath Nambiar and Meikel Poess. Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-10424-4_5.

Fischer, Keno. “Growing a Compiler – Getting to Machine Learning from a General Purpose Compiler.” JuliaHub, February 19, 2019. https://info.juliahub.com/growing-a-compiler-getting-to-machine-learning-from-a-general-purpose-compiler.

Genkina, Dina. 2022. “Machine Learning’s New Math.” IEEE Spectrum. Published October 18, 2022. https://spectrum.ieee.org/number-representation.

GitHub. 2024. “GitHub Copilot: The world’s most widely adopted AI developer tool.” Undated. Accessed February 27, 2024. https://github.com/features/copilot.

Gozalo-Brizuela, Roberto, and Eduardo C. Garrido-Merchán. 2023. “A survey of Generative AI Applications.” arXiv. Submitted June 5, 2023. https://doi.org/10.48550/arXiv.2306.02781.

Greathouse, Joseph. 2021. “AMD Research Instruction Based Sampling Toolkit.” Last updated April 29, 2021. Accessed November 28, 2023. https://github.com/jlgreathouse/AMD_IBS_Toolkit.

Gregg, Brendan. 2023. “Linux Performance.” Undated. Accessed November 28, 2023. https://www.brendangregg.com/linuxperf.html.

Grollier, J., D. Querlioz, K.Y. Camsari, K. Everschor-Sitte, S. Fukami, and M.D. Stiles. 2020. “Neuromorphic Spintronics.” *Nature Electronics*. Vol. 3: pg 360–370. https://doi.org/10.1038/s41928-019-0360-9.

Hamming, Richard Wesley. 1950. “Error detecting and error correcting codes.” *Bell System Technical Journal*. Vol. 29 (Issue 2): pg 147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x.

Harun, Md Yousuf, et al. 2023. “How Efficient Are Today’s Continual Learning Algorithms?” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.48550/arXiv.2303.18171.

Hennessy, John L., and David A. Patterson. 2019. “A New Golden Age for Computer Architecture.” *Communications of the ACM*. Vol. 62 (Issue 2). https://doi.org/10.1145/3282307.

Heroux, Michael Allen, and Jack Dongarra. 2013. “Toward a New Metric for Ranking High Performance Computing Systems.” Technical Report, Sandia National Laboratories, OSTI ID 1089988. https://doi.org/10.2172/1089988.

Hittinger, J.A., et al. 2019. “Variable Precision Computing.” Livermore, CA: Lawrence Livermore National Laboratory. LLNL-TR-795750. https://www.osti.gov/servlets/purl/1573151.

Horowitz, M. 2014. “1.1 Computing’s energy problem (and what we can do about it).” *2014 IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*. San Francisco. https://doi.org/10.1109/ISSCC.2014.6757323.

Hospedales, Timothy, et al. 2021. “Meta-learning in neural networks: A survey.” *IEEE Transactions on Pattern Analysis and Machine Intelligence.* Vol. 44 (Issue 9): pg 5149–5169. https://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3079209.

Hsu, H.-C., Z.-Y. Liu, R. Tso, and K. Chen. 2020. “Multi-value Private Information Retrieval using Homomorphic Encryption.” Presented at the 2020 15th Asia Joint Conference on Information Security (AsiaJCIS). Taipei, Taiwan. https://doi.org/10.1109/AsiaJCIS50894.2020.00024.

Huang, B., and J. Wang. 2023. “Applications of Physics-Informed Neural Networks in Power Systems – A Review.” *IEEE Transactions on Power Systems.* Vol. 38 (Issue 1): pg 572–588. https://doi.org/10.1109/TPWRS.2022.3162473.

Huang, Hao, Tapan Shah, Scott Evans, and Shinjae Yoo. 2023. “Less-Energy-Usage Network with Batch Power Iteration.” Proceedings of the 40th International Conference on Machine Learning, Workshop Neural Compression: From Information Theory to Applications. Honolulu, HI.

Human Brain Project. 2023. “Neuromorphic computing.” Accessed December 12, 2023. https://www.humanbrainproject.eu/en/science-development/focus-areas/neuromorphic-computing/.

IBM. 2018. “POWER9 Performance Monitor Unit User’s Guide.” OpenPOWER Version 1.2. Published November 28, 2018. https://wiki.raptorcs.com/w/images/6/6b/POWER9_PMU_UG_v12_28NOV2018_pub.pdf.

IEEE. 2019. “IEEE Standard for Floating-Point Arithmetic.” IEEE-754-2019. Published July 22, 2019. https://standards.ieee.org/ieee/754/6210/.

Intel. 2020. “Update on Intel’s Neuromorphic Ecosystem Growth and Progress.” Accessed February 2, 2024. https://www.intel.com/content/www/us/en/newsroom/news/neuromorphic-ecosystem-growth-progress.html#gs.3wrub9.

Intel. 2022. “Intel® Performance Counter Monitor – A Better Way to Measure CPU Utilization.” Last updated November 30, 2022. https://www.intel.com/content/www/us/en/developer/articles/tool/performance-counter-monitor.html.

Intel. 2023. “Intel Vtune Profiler.” Undated. Accessed November 28, 2023. https://www.intel.com/content/www/us/en/developer/tools/oneapi/vtune-profiler.html#gs.1iaxrv.

Intel. 2023a. “Pin – A Dynamic Binary Instrumentation Tool.” Undated. Accessed November 28, 2023. https://www.intel.com/content/www/us/en/developer/articles/tool/pin-a-dynamic-binary-instrumentation-tool.html.

Intel. 2023b. “Loihi 2: A New Generation of Neuromorphic Computing.” Undated. Accessed December 12, 2023. https://www.intel.com/content/www/us/en/research/neuromorphic-computing.html.

International Energy Agency. "Digitalisation and Energy." November 2017. https://www.iea.org/reports/digitalisation-and-energy.

ISO. 2020. “Information technology: Data centres server energy effectiveness metric.” Standard ISO/IEC 21836:2020. Published August 2020. https://www.iso.org/standard/71926.html.

Jeaugey, Sylvain. 2019. “Massively Scale Your Deep Learning Training with NCCL 2.4.” NVIDIA Technical Blog. Published February 4, 2019. https://developer.nvidia.com/blog/massively-scale-deep-learning-training-nccl-2-4/.

Jouppi, Norman P., et al. 2021. “Ten Lessons from Three Generations Shaped Google’s TPUv4i : Industrial Product.” Presented at the 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA). Valencia, Spain. https://doi.org/10.1109/ISCA52012.2021.00010.

Kandiah, Vijay, Scott Peverelle, Mahmoud Khairy, Junrui Pan, Amogh Manjunath, Timothy G. Rogers, Tor M. Aamodt, and Nikos Hardavellas. 2021. “AccelWattch: A Power Modeling Framework for Modern GPUs.” *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*. Virtual event. https://doi.org/10.1145/3466752.3480063.

Kim, Christine. "The Rise of ASICs: A Step-by-Step History of Bitcoin Mining." CoinDesk, April 26, 2020. Updated September 14, 2021. https://www.coindesk.com/tech/2020/04/26/the-rise-of-asics-a-step-by-step-history-of-bitcoin-mining/.

Kudithipudi, D., M. Aguilar-Simon, et al. 2022. “Biological underpinnings for lifelong learning machines.” *Nature Machine Intelligence.* Vol. 4 : pg 196–210. http ://dx.doi.org/10.1038/s42256-022-00452-0.

Langroudi, Hamed F., Zachariah Carmichael, John L. Gustafson, and Dhireesha Kudithipudi. « PositNN Framework : Tapered Precision Deep Learning Inference for the Edge. » In 2019 IEEE Space Computing Conference (SCC), 53-59. IEEE, 2019. Doi :10.1109/SpaceComp.2019.00011.

Lattner, Chris. 2021. “The New Golden Age of Compilers.” Presented at the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2021). Virtual event. https://canvas.eee.uci.edu/courses/43849/files/17973444/download?verifier=TL0VdfIvYmS7WJWe6rnEKL6nFSGPZ0RtcsjF5SOM&download_frd=1.

Lattner, Chris, and Jacques Pienaar. 2019. “MLIR Primer: A Compiler Infrastructure for the End of Moore’s Law.” Presented at the Compilers for Machine Learning Workshop, CGO 2019. Washington, DC. https://research.google/pubs/pub48035/.

Lei, Nuuo, Eric Masanet, and Jonathan Koomey. “Best Practices for Analyzing the Direct Energy Use of Blockchain Technology Systems: Review and Policy Recommendations.” Energy Policy 156 (September 2021): 112422. https://doi.org/10.1016/j.enpol.2021.112422.

Li, Ang, Samuel Stein, Sriram Krishnamoorthy, and James Ang. 2020. “QASMBench: A Low-Level Quantum Benchmark Suite for NISQ Evaluation and Simulation.” *ACM Trans. Quantum Comput.* Vol. 37 (Issue 4, Article no. 111). https://doi.org/10.1145/1122445.1122456.

Madireddy, S., A. Yanguas-Gil, and P. Balaprakash. 2023. “Improving Performance in Continual Learning Tasks using Bio-Inspired Architectures.” arXiv. Submitted August 8, 2023. https://doi.org/10.48550/arXiv.2308.04539.

Mattson, Peter, et al. 2020. “MLPerf Training Benchmark.” arXiv. Last revised March 2, 2020. https://doi.org/10.48550/arXiv.1910.01500.

Mead, C. “Neuromorphic Electronic Systems.” Proceedings of the IEEE 78, no. 10 (October 1990): 1629-1636. http://doi:10.1109/5.58356.

MICAS. 2023. “Hardware-Efficient AI and ML.” KU Lueven. Accessed July 19, 2023. https://micas.esat.kuleuven.be/research/domains/hardware-efficient-ai-and-ml.

Misyris, G.S., A. Venzke, and S. Chatzivasileiadis. 2020. “Physics-Informed Neural Networks for Power Systems.” Presented at the 2020 IEEE Power & Energy Society General Meeting (PESGM). Montreal, QC, Canada. https://doi.org/10.1109/PESGM41954.2020.9282004.

Modha, Dharmendra, et al. 2023. “Neural inference at the frontier of energy, space, and time.” *Science.* Vol. 382 : pg 329–335. http ://dx.doi.org/10.1126/science.adh1174.

Moore, Samuel K. 2024. “Chips to Compute with Encrypted Data Are Coming.” IEEE Spectrum. Accessed January 2024. https://spectrum.ieee.org/homomorphic-encryption.

Munjal, Kundan, and Rekha Bhatia. 2022. “A systematic review of homomorphic encryption and its contributions in healthcare industry.” *Complex & Intelligent Systems*. Vol. 9 (Issue 4): pg 3759–3786. http://dx.doi.org/10.1007/s40747-022-00756-z.

Mustafa, Basil, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. “Multimodal Contrastive Learning with LIMoE: the Language-Image Mixture of Experts.” arXiv. Submitted June 6, 2022. https://doi.org/10.48550/arXiv.2206.02770.

NVIDIA. 2023. “Performance Analysis Tools.” Undated. Accessed November 28, 2023. https://developer.nvidia.com/performance-analysis-tools.

NVIDIA. 2023a. “Train with Mixed Precision.” User’s Guide, NVIDIA Document DA-08617-001_v001. Last updated February 1, 2023. https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html.

NVIDIA. 2023b. “NVIDIA Collective Communications Library (NCCL).” Undated. Accessed November 30, 2023. https://developer.nvidia.com/nccl.

O’ Neill, James. 2020. “An Overview of Neural Network Compression.” arXiv. Submitted June 5, 2020. https://doi.org/10.48550/arXiv.2006.03669.

Pedersen, Jens E., et al. 2023. “Neuromorphic Intermediate Representation: A Unified Instruction Set for Interoperable Brain-Inspired Computing.” arXiv. Submitted November 24, 2023. https://doi.org/10.48550/arXiv.2311.14641.

Pellauer, Michael, Yakun Sophia Shao, Jason Clemons, et al. 2019. “Buffets: An Efficient and Composable Storage Idiom for Explicit Decoupled Data Orchestration.” Presented at ASPLOS ’19: Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems. Providence, RI. https://doi.org/10.1145/3297858.3304025.

Pereira, Rui, Marco Couto, Francisco Ribeiro, Rui Rua, Jácome Cunha, João Paulo Fernandes, and João Saraiva. 2021. “Ranking programming languages by energy efficiency.” *Science of Computer Programming.* Vol. 205: 102609. https://doi.org/10.1016/j.scico.2021.102609.

Raissi, M., P. Perdikaris, and G.E. Karniadakis. 2019.“Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations.” *Journal of Computational Physics.* Vol. 378: pg 686–707. https://doi.org/10.1016/j.jcp.2018.10.045.

Rouhani, Bita, et al. 2020. “Pushing the Limits of Narrow Precision Inferencing at Cloud Scale with Microsoft Floating Point.” Presented at the 34th Conference on Neural Information Processing Systems. Vancouver, BC, Canada. https://proceedings.neurips.cc/paper/2020/file/747e32ab0fea7fbd2ad9ec03daa3f840-Paper.pdf.

Sakharnykh, Nikolay, Dominique LaSalle, and Ben Karsin. 2020. “Optimizing Data Transfer Using Lossless Compression with NVIDIA nvcomp.” NVIDIA Technical Blog. Published December 18, 2020. https://developer.nvidia.com/blog/optimizing-data-transfer-using-lossless-compression-with-nvcomp/.

Schuman, Catherine D., Shruti R. Kulkarni, Maryam Parsa, J. Parker Mitchell, Prasanna Date, and Bill Kay. 2022. “Opportunities for neuromorphic computing algorithms and applications.” *Nature Computational Science*. Vol 2: pg 10–19. https://doi.org/10.1038/s43588-021-00184-y.

Shankar, Sadasivan, and Albert Reuther. “Trends in Energy Estimates for Computing in AI/Machine Learning Accelerators, Supercomputers, and Compute-Intensive Applications.” arXiv preprint arXiv:2210.17331, submitted October 12, 2022. https://arxiv.org/abs/2210.17331.

Shankar, Sadasivan. “Energy Estimates Across Layers of Computing: From Devices to Large-Scale Applications in Machine Learning for Natural Language Processing, Scientific Computing, and Cryptocurrency Mining.” arXiv preprint arXiv:2310.07516, submitted October 11, 2023. https://doi.org/10.48550/arXiv.2310.07516.

Shooman, Martin L. 2002. “*N*-Modular Redundancy.” In *Reliability of Computer Systems and Networks: Fault Tolerance, Analysis and Design*, 145–201. Wiley-Interscience. https://doi.org/10.1002/047122460X.ch4.

Song, Wenhao, et al. 2024. “Programming memristor arrays with arbitrarily high precision for analog computing.” *Science.* Vol. 383 (Issue 6685): pg 903–910. https://doi.org/10.1126/science.adi9405.

Song, Yo-Der, and Tomaso Aste. "The Cost of Bitcoin Mining Has Never Really Increased." Frontiers in Blockchain 3 (October 22, 2020). https://doi.org/10.3389/fbloc.2020.565497.

Statistica. 2023. “Artificial Intelligence (AI) Market Size Worldwide in 2021 with a Forecast until 2030.” Published October 6, 2023. https://www.statista.com/statistics/1365145/artificial-intelligence-market-size/.

Sze, Vivienne, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. 2017. “Efficient Processing of Deep Neural Networks: A Tutorial and Survey.” *Proceedings of the IEEE*. Vol. 105 (Issue 12): pg 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740.

The White House. "FACT SHEET: Climate and Energy Implications of Crypto-Assets in the United States." September 8, 2022. https://www.whitehouse.gov/ostp/news-updates/2022/09/08/fact-sheet-climate-and-energy-implications-of-crypto-assets-in-the-united-states/.

Thompsett, Louis. "Bitcoin Reclaims US$1tn Valuation; the Bull Market is Here." FinTech Magazine, February 20, 2024. https://fintechmagazine.com/articles/bitcoin-reclaims-us-1tn-valuation-the-bull-market-is-here.

Thompson, Neil C., Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. 2022. “The Computational Limits of Deep Learning.” arXiv. Last revised July 27, 2022. https://doi.org/10.48550/arXiv.2007.05558.

TOP500. 2022. “TOP500 List: June 2022.” Published June 2022. Accessed May 6, 2024. https://www.top500.org/lists/top500/2022/06/.

Tsai, Po-An, and Daniel Sanchez. 2019. “Compress Objects Not Cache Lines – an Object-Based Compressed Memory Hierarchy.” Presented at ASPLOS ’19: Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems. Providence, RI. http://dx.doi.org/10.1145/3297858.3304006.

U.S. Department of Energy. 2018. “ENERGY STAR® Program Requirements for Computer Servers.” https://www.energystar.gov/sites/default/files/ENERGY%20STAR%20Version%203.0%20Computer%20Servers%20Program%20Requirements_0.pdf.

UPMEM. 2023. “Best performance and efficiency for big data & AI.” Accessed December 12, 2023. https://www.upmem.com/.

Variorum. 2023. “LLNL/variorum: Vendor-neutral library for exposing power and performance features across diverse architectures.” Lawrence Livermore National Laboratory (LLNL). GitHub repository. Last updated June 13, 2023. https://github.com/LLNL/variorum.

Verhelst, M., and B. Murmann. 2020. “Machine Learning at the Edge.” In *NANO-CHIPS 2030*, edited by B. Murmann and B. Hoefflinger, 293–322. Cham, Switzerland: Springer, Cham. https://doi.org/10.1007/978-3-030-18338-7_18.

Villalobos, Pablo, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, Anson Ho, and Marius Hobbhahn. 2022. “Machine Learning model sizes and the Parameter gap.” arXiv. Submitted July 5, 2022. https://doi.org/10.48550/arXiv.2207.02852.

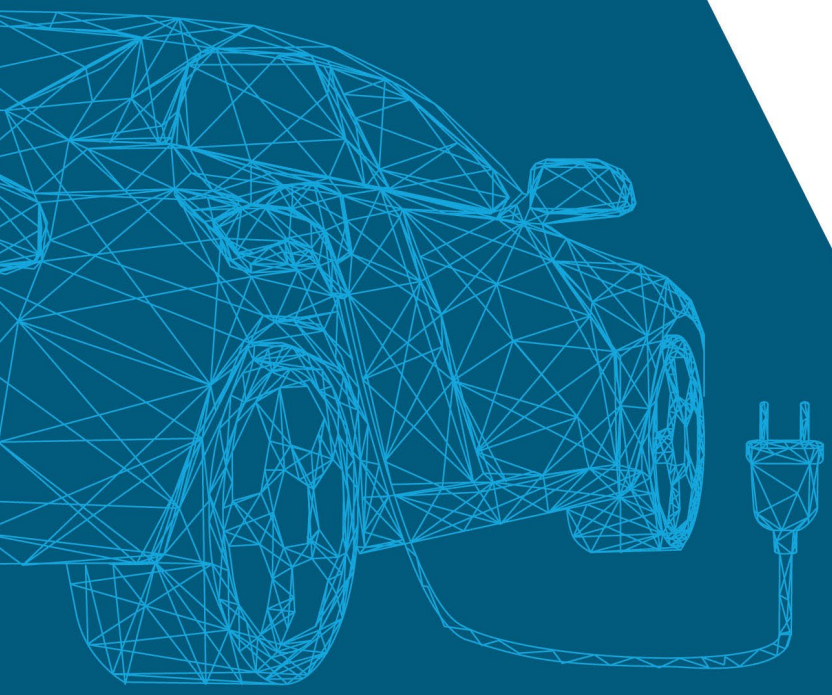
Wang, Z., and M. O’Boyle. 2018. “Machine Learning in Compiler Optimization.” *Proceedings of the IEEE*. Vol. 106 (Issue 11): pg 1879–1901. https://doi.org/10.1109/JPROC.2018.2817118.

Wang, Z., C. Liu, and T. Nowatzki. 2022. “Infinity Stream: Enabling Transparent and Automated In-Memory Computing.” *IEEE Computer Architecture Letters*. Vol. 21 (Issue 2): pg 85–88. https://doi.org/10.1109/LCA.2022.3203064.

Weissman, Tsachy. 2022. “The crucial role of data compression.” The Future of Everything podcast, Stanford University. Uploaded March 11, 2022. https://www.youtube.com/watch?v=qNZGYYZc8Oc.

Wikipedia. 2024. "Neural network (machine learning)." Undated. Accessed May 6, 2024. https://en.wikipedia.org/wiki/Neural_network_(machine_learning).

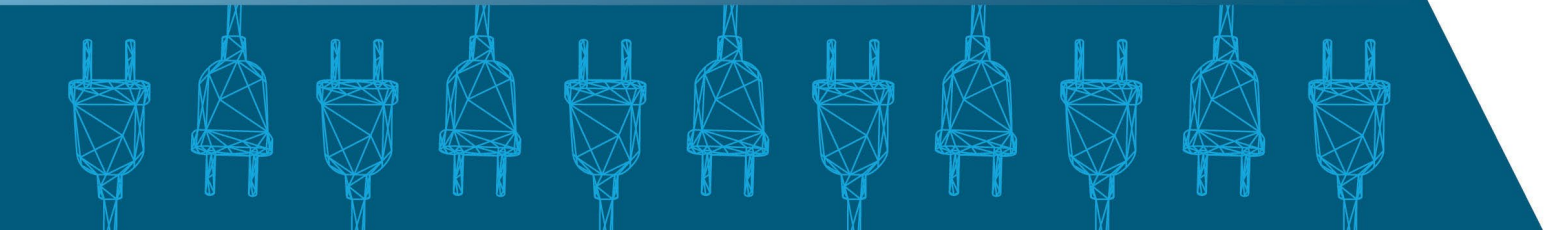
Yagemann, Carter. 2023. “A Practical Beginner’s Guide to Intel Processor Trace.” Published February 24, 2023. https://carteryagemann.com/a-practical-beginners-guide-to-intel-processor-trace.html.



SECTION

3

Enablers



3 Enablers

The four Enabler working groups—Power and Control Electronics, Manufacturing Energy Efficiency and Sustainability, Metrology and Benchmarking, and Education and Workforce Development—cover enabling technologies and approaches in order to address the tools, processes, and technologies needed to support the advances in the compute stack described in the previous chapters.

3.1 Power and Control Electronics (PACE)

Power and control electronics (PACE) refers to an interdisciplinary field with roots in electrical engineering and technology development. This field focuses on the design, development, and application of electronic systems and devices responsible for managing and regulating the use of electricity as an energy source. These systems play a critical role in controlling the generation, conversion, distribution, and utilization of electrical energy. Since they often involve components and circuits designed for the control and automation of various processes, these systems and devices are integral to a wide range of applications, including power supplies, motor drives, renewable energy systems, industrial automation, and more.

PACE and microelectronics are two distinct branches within the broader field of electrical engineering that focus on different aspects of electronic systems. The two branches are generally differentiated by the intended purpose in use, the scale at which they are applied, and the underlying devices and components they utilize. The differences between PACE and microelectronics are further described in Table 63.

Table 63. Power and Control Electronics and Microelectronics Fields.

Field	Intended Purpose	Trend	Devices
Power and Control Electronics	Efficiently manage the generation, distribution, conversion, and control of electrical power.	High voltage and high power over time	Power semiconductors (e.g., thyristors, IGBTs, MOSFETs, diodes), power converters, voltage regulators, motor drives, and control systems
Microelectronics	The miniaturization of electronic components, the fabrication of integrated circuits (Ics), and the development of semiconductor devices, such as microprocessors, memory chips, and other integrated circuits.	Miniaturization over time, currently allowing billions of transistors on a single chip	Transistors, integrated circuits, systems-on-a-chip

In summary, while PACE focuses on managing and controlling the delivery of electric power, microelectronics considers the development and integration of electronic components to create

devices for computation and telecommunication. At their heart, both device classes are built on semiconductor technology, though differences exist in their respective design objectives.

Use of PACE in computing environments

In the computing industry, PACE are essential for ensuring the reliable and efficient operation of electronic devices, from personal computers to data centers. PACE serve several key uses in computing, including:

- **Uninterruptible power supplies (UPS):** UPS systems are used to ensure that critical pieces of equipment never experience a power outage. At their simplest, UPS systems detect when a utility power supply becomes unavailable and use high-speed switches and battery energy storage to provide an alternative electricity supply. In their most advanced form, UPS systems use power electronics to recondition utility electricity supplies in real-time, removing any variations or fluctuations in voltage or frequency, which ensures high power quality for sensitive equipment.
- **Power distribution units (PDU):** PDUs are devices used to distribute electric power to individual server racks. At the simplest level, these devices are analogous to the power strips used in homes and offices to supply electricity to many devices at once. However, PDUs can be much more complicated. In most modern data center facilities, PDUs contain monitoring and control equipment to provide granular insight into energy use and to remotely control power delivery to individual servers or devices. In some data center facilities, PDUs also contain voltage transformers, which are used to reduce power distribution from the UPS output voltage level to a lower voltage that is more suitable for use by the electronic equipment. PDUs may also represent the point at which individual phases branch off from the three-phase utility electrical supply (e.g., step down from 480 V AC, 3-phase to 120 V AC, single phase).
- **Switching power supplies:** These are widely used in computer systems to convert electrical power from the main power source (usually the electrical grid) into the various DC voltages needed by different components within the computer, such as the motherboard, central processing unit (CPU), and graphics processing unit (GPU).
- **Voltage regulators:** Power electronics are used to regulate and stabilize the voltage supplied to sensitive components in computers. Voltage regulators are used to ensure that the voltage supplied to critical devices, like processors, maintains a constant value, as even small fluctuations in voltage can cause damage. This regulation of the voltage is crucial for preventing damage and ensuring the proper functioning of Ics, circuit components, and subsystems.
- **Variable speed fans and pumps for cooling systems:** Power electronics control the speed of fans and pumps in computer systems to optimize air and liquid-based cooling systems. This process is crucial for maintaining the operating temperature of components within acceptable limits. On a larger scale, similar power electronics equipment is often used to control the HVAC systems within data centers.
- **Power factor correction (PFC) systems, which support energy efficiency:** PFC systems, using power electronics, are employed to improve the power factor of computing

equipment and the motor loads that drive data center cooling systems. PFC is a method for reducing energy consumption and improving overall efficiency in alternating current (AC) electrical systems.

In the context of a data center, these power electronic devices are used to provide conditioned electrical power to information and communication technology (ICT) devices, which may include servers and computers; devices such as switches, routers, wireless access points, power-over-ethernet devices, and telecommunication systems; and storage devices and digital signal processing equipment.

Power and control electronics are foundational to the functioning of computing systems, ensuring reliable power delivery, energy efficiency, and the overall performance and longevity of electronic components in a wide range of computing devices and infrastructure.

Relevance to EES2

The PACE working group considered the role that these devices play in supporting the computing infrastructure being explored in other working groups (Materials and Devices, Circuits and Architectures, and Algorithms and Software). PACE supports the proper operation of computing infrastructure and has a direct impact on the energy consumed by microelectronics devices.

All the electrical energy consumed by computing systems passes first through power electronic devices, which are regulated by control systems. If these devices and controls are inefficient, the total energy demand for a computing facility may far exceed the energy input required to run the intended computational equipment. Additionally, energy losses associated with PACE are converted into waste heat that must be removed from computing facilities.

The technical approaches described in the PACE section are organized as follows: Dynamic computing load management techniques are described first. These techniques directly manipulate the power delivered to computing devices to reduce power consumption. Next, advanced and emerging thermal management approaches are described since new approaches will be needed to accommodate emerging circuit architectures with increasing thermal management requirements. Lastly, the PACE section describes the enhanced modeling, analysis, and simulation needs for empowering future co-design efforts in support of the next generation of energy-efficient devices and computing facilities.

Working group methodology

The PACE working group sought to better understand the contribution of power electronics and their control systems to the energy efficiency of computing devices in operation. The working group explored the use of PACE in computing environments, noting standard industry practices and potential areas of innovation. The working group concluded that best-in-class power electronics do not represent significant sources of power consumption or loss in modern and newly constructed data centers. Furthermore, industry-standard practices related to controlling power delivery are effective in eliminating losses and ensuring highly efficient power delivery. However, there are related concepts that warrant further consideration in future iterations of the EES2 roadmap. Though these related concepts may not represent power electronics and control challenges or solutions directly, they do address broader challenges related to the design, implementation, and optimization of power delivery for emerging architectures.

The diagram in Figure 60 presents the estimated energy efficiency factors and related timelines for the technological approaches recommended by the PACE working group. These approaches are described in more detail in the subsequent sections, and activities related to the timelines referenced are noted in the action plans contained in this chapter.

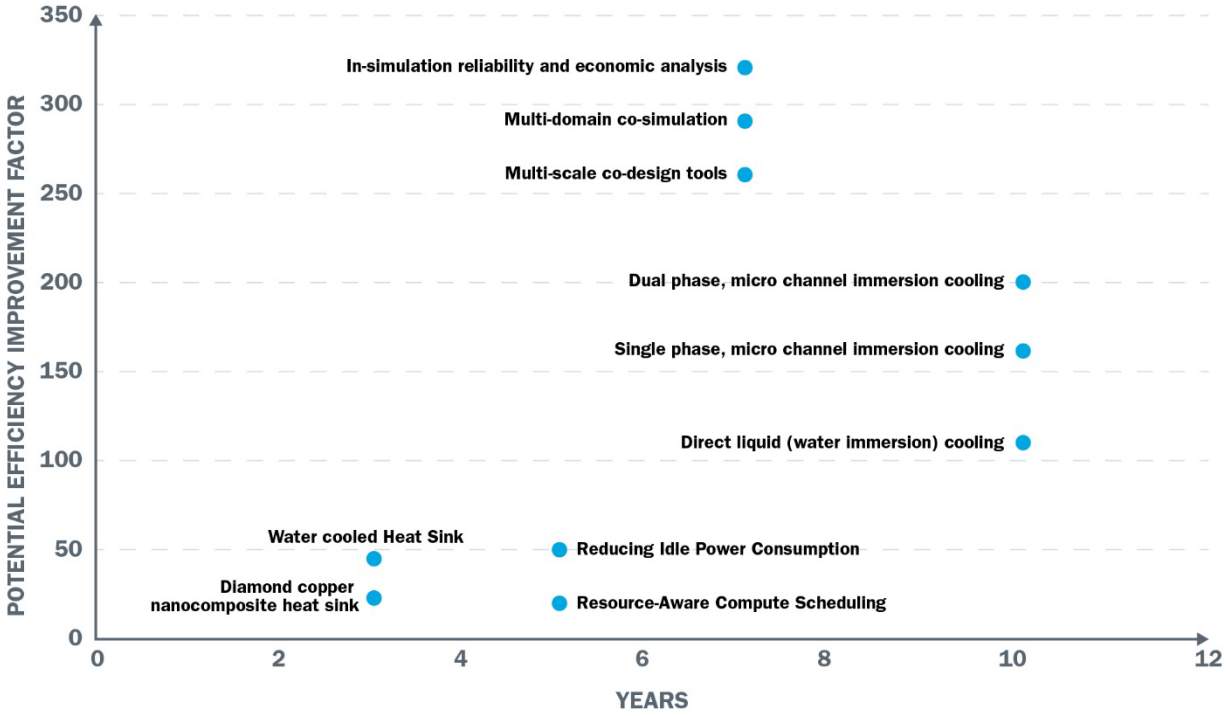
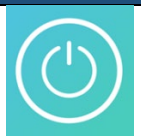
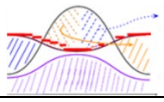


Figure 60. Potential efficiency improvement factor vs. timeline for PACE technologies.

Key takeaways

The following tables present and summarize the PACE-related technologies recommended for further investigation, as well as the major contributions each recommended technology makes to energy efficiency.

Table 64. Key Opportunities for PACE Technology.

Technology Group	Key Opportunities for Energy Efficiency	
Dynamic Computing Load Management	 	• While modern servers have standby modes that consume much less power than in their active states, most servers still consume 80% of their total lifetime electric power while in standby mode. • Turning devices off completely could increase server efficiency by 5X. This represents significant global savings, given the growing installed base of servers worldwide. • Shifting workloads to data centers with more efficient computing resources or available renewable power can effectively reduce global computational energy use.

Advanced Thermal Management Technologies		• Energy densities are increasing in emerging computer architectures, requiring novel strategies for removing heat from circuits while in operation. • Most data centers currently use forced air cooling, but technical limits will reduce the use of this technology in the future. • Techniques for cooling that involve immersion in liquid, coolant distribution via microchannels, and phase-change materials offer new opportunities for managing heat removal in future computing devices.
Enhancing Modeling, Simulation, and Co-Design Capabilities		• Using presently available software packages for microelectronics design, it is difficult to connect energy performance at the device level to overall energy performance at the facility level. • It is nearly impossible to weigh the impact of device-level design changes on high-level system energy consumption. • Extensions are needed to the capabilities of modern design and analysis software programs, to allow co-simulation, co-optimization of system design properties, and validation of design changes. • Enhanced co-design tools will enable the design of future computing systems that are globally optimized to reduce power consumption.

Table 65. PACE Technology Grouping.

Technology Category	Technology
Dynamic Computing Load Management	Reduction of idle power consumption
	Resource-aware compute scheduling
Advanced Thermal Management Technologies	Diamond copper nanocomposite heat sink
	Water cooled heat sink
	Direct liquid cooling
	Immersion cooling, single phase
	Immersion cooling, dual phase
Enhancing Modeling, Simulation, and Co-Design Capabilities	Multi-scale co-design tools
	Multi-domain co-simulation
	In-simulation reliability and economic analysis

Grand challenges

The following challenges must be addressed to realize the potential contributions of PACE technologies toward EES2 goals:

- **Architecture-specific power delivery optimization:** As new device architectures are developed (2.5/3D, neuromorphic, PICs, etc.) power delivery approaches will need to become more specialized to each architecture. The power delivery needs of each emerging architecture will need to be independently assessed and accounted for.
- **Enhanced co-design capabilities:** To fully understand implications for energy efficiency, electricity delivery needs to be co-designed with circuits and architectures.

Improvements in design and simulation tools can help to elucidate design tradeoffs related to energy efficiency.

- **Thermal management:** Despite significant improvements in thermal management at the data center level, emerging architectures will require new approaches to on-chip thermal management due to increases in energy density and dimensionality. Thermal management is an integral part of power delivery optimization and innovations will be required.
- **Applying innovations in non-data-center contexts and legacy computing facilities:** New data centers are being built at massive scales and with impressive innovations included by default. What remains unclear is the percentage of legacy computing equipment, the computational work being performed in non-data-center contexts, and the energy efficiency burden that legacy equipment represents. There may be opportunities to incentivize energy efficiency upgrades for computing equipment housed in non-data-center facilities (e.g., hospitals, research centers, academic institutions, etc.).

3.1.1 State of the Art and Benchmarks

Overview of Data Center Power Distribution

Power distribution systems for data centers are designed to ensure reliability and safety. Any interruptions in the flow of electricity to computational equipment can be costly for data center owners and operators. Therefore, power system design for data centers prioritizes the use of redundant and varied supplies of energy. Since different electricity sources (e.g., an electric utility connection, a diesel generator, an electrochemical battery system, etc.) have different characteristics (voltage fluctuations, disturbances, frequency variations, harmonic distortion levels), methods are used to ensure the same quality of electricity is delivered to sensitive electronic devices. For this reason, double-conversion UPS systems are commonly used in data centers: Each UPS contains a power electronic rectifier, which converts the included AC utility waveform into a DC electrical signal. The UPS then utilizes an inverter (which is another class of power electronics devices) to create a new AC waveform with consistent power quality. These UPS systems also utilize energy storage (typically in the form of electrochemical batteries) to maintain a consistent power supply in the event of a short-duration outage in the primary (utility) electrical supply. For longer-duration outages, onsite diesel backup generators are typically utilized to supply power to data centers until utility power can be restored.

Power distribution architectures can vary between data center facilities. Different facility owners and operators employ a variety of strategies to ensure redundancy, enhance reliability, and achieve high power quality. The diagram in Figure 61 provides an overview of different power delivery architectures for data centers. Though the configuration and interconnection of devices may change, the primary components stay consistent: namely, utility power supplies, distribution transformers, panels and switchgear, backup generators, UPSs, and IT loads.

Figure 62 shows a more streamlined view of power delivery to IT and ICT devices within a data center. This view emphasizes the power electronics devices typically included in the power delivery chain that feed computational devices. As shown, every electron utilized within a computational device must first pass through a long chain of power conversion, conditioning,

and control steps. As energy is converted from one form to another, waste heat is generated, which must be removed from the facility using HVAC equipment.

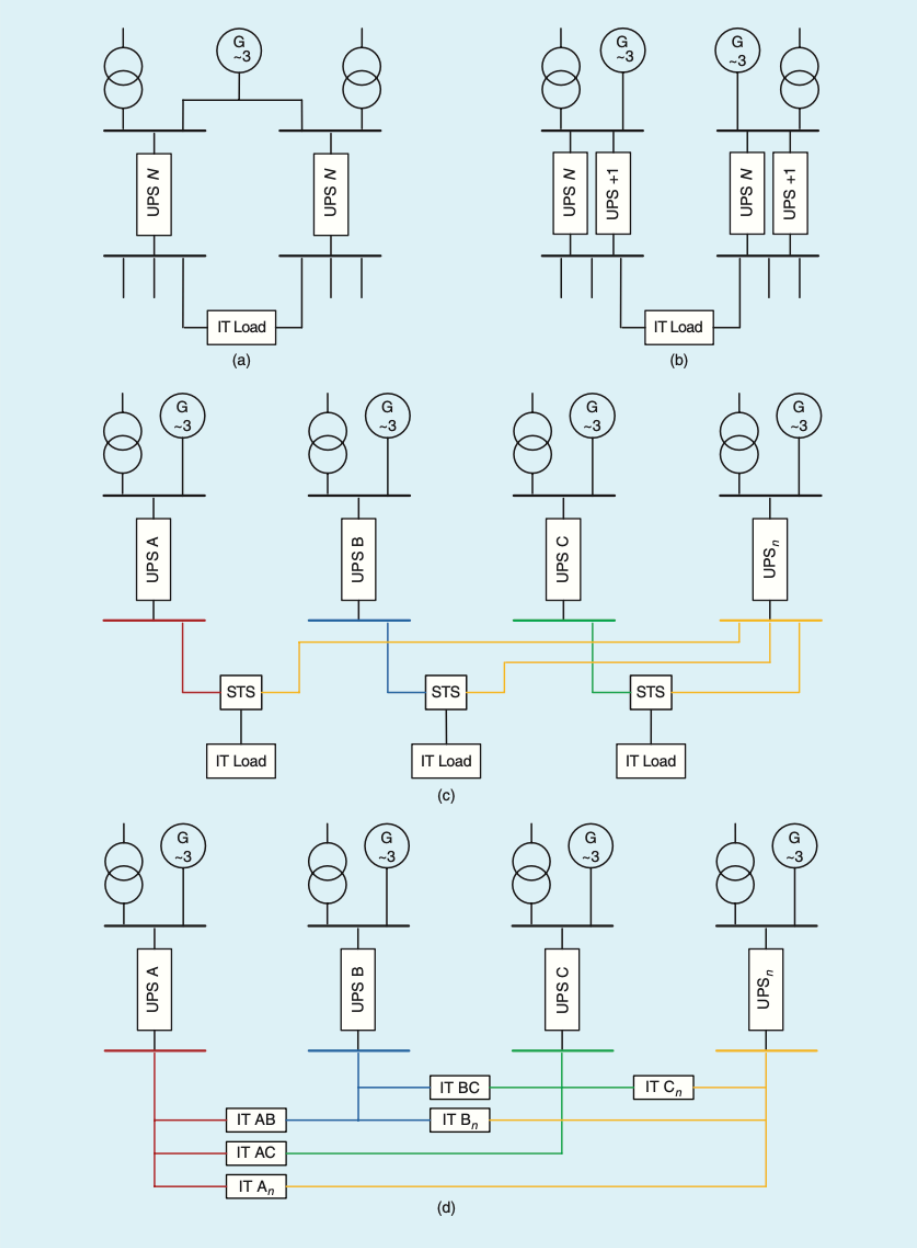


Figure 61. Common power distribution architectures for data centers. *Source: Paananen 2023*

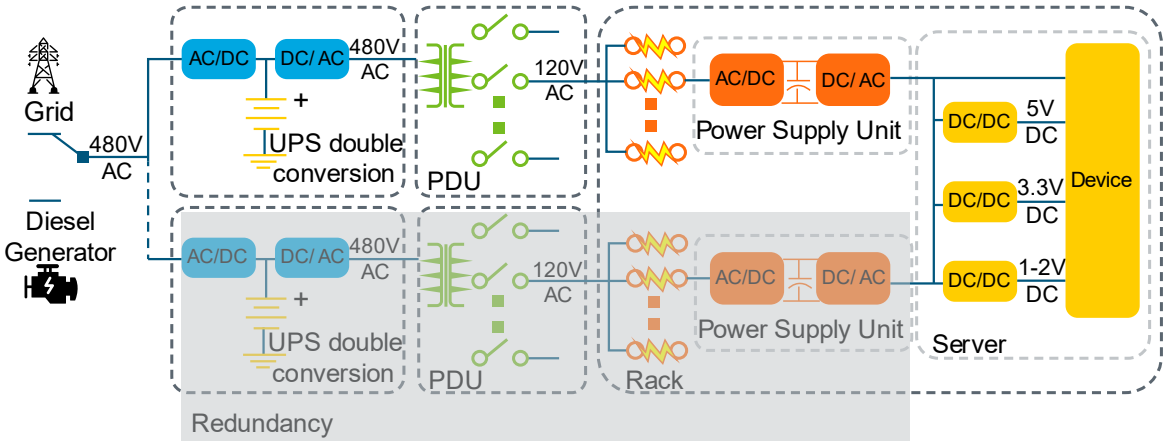


Figure 62. Power electronics in the data center power delivery chain. *Source: Sun et al. 2018*

End-to-End Efficiency in Data Centers

Power use effectiveness (PUE) is a metric that has become popular as a means for communicating the energy efficiency of an operational data center (Gillis and Fontecchio 2022). The PUE metric seeks to capture the extent to which energy consumed within a data center is utilized for computation, the primary purpose of the data center. In practice, data centers consume energy in power conversion equipment, control systems, HVAC systems, and auxiliary building systems. PUE is calculated as the ratio of total energy consumed in a data center to energy consumed specifically by IT equipment. A ratio of 1 would represent 100% of data center facility power being consumed by the intended IT equipment. A PUE efficiency of 3 would mean the facility overall uses three times the amount of power as the IT equipment alone, which is highly inefficient.

In the early 2000s, data centers were reporting an average PUE of 3 or more (de Jong and Vaessen 2007). The industry recognized a need for improvement and resources were dedicated to improvement. In the last two decades, tremendous improvements have been realized. Modern data centers now achieve an average PUE of 1.57 (Bizo et al. 2021), with industry-leading facilities achieving a PUE as low as 1.06. The average PUE for all Google data centers is 1.10 (Google 2023).

Required Cooling Load in Data Centers

To achieve these PUE improvements, data centers have introduced remarkable innovations in the management and control of cooling systems. As computational devices operate within a data center, they produce heat as a natural byproduct. Data center cooling systems work to ventilate this air, replacing it with cooled air, which prevents temperature rises that could damage electronics. Airflow management systems are used to control the flow of air across computing devices. These systems have evolved, incorporating tight seals, plates, and fittings to ensure that exhaust and intake air systems are not able to mix. Precision control of air delivery and removal has greatly improved cooling efficiency for data centers. In some instances, like supercomputing and high-performance computing systems, liquid cooling solutions replace forced air movement.

What could be considered the greatest innovation in cooling efficiency involves the use of “free cooling solutions,” which are made possible by geographically locating data centers in

advantageous locations, such as those with naturally occurring cold water sources or lower ambient air temperatures. In these locations, air and water can be circulated without the use of compressor-based refrigeration systems, which drive energy use in modern cooling and HVAC systems.

With decades of investment, cooling has gone from the highest energy consumer within a data center to a much less significant portion of overall data center power use, which has contributed significantly to the increases seen in data center PUE.

Power Electronics Conversion Efficiency for Data Center Power Delivery

A variety of power electronics devices are required to condition power for computing equipment within a data center. Power electronics converters are devices that are used to convert between alternating current and direct current electricity distribution, or to convert between voltage levels. Table 66 summarizes the common power electronics converters that are found in data centers.

Table 66. Common Power Electronics Converters in Data Centers.

Name	Description	Data Center Use
Rectifier	Converts from AC to DC	Used within UPS systems to eliminate fluctuations in utility or generator supply voltages. Also used in server power supplies to create the DC voltages required to operate electronics.
Inverter	Converts from DC to AC	Used in UPS systems to recreate AC waveforms that are high quality and well-regulated.
Buck Converter	Reduces DC voltage levels	May be used in PDUs or PSUs to further reduce DC voltages after rectification. For instance, rectifiers in data centers often produce voltages between 300 V and 400 V DC. Buck converters are used to produce a regulated 12 V or 48 V supply for on-chip power distribution.

Given the prevalence of power electronics converters within data centers, efforts have been made to increase their efficiency. Inefficient converters produce more waste heat, which building HVAC systems must eliminate. Inefficient power electronics within a data center cause subsequent increases in overall facility energy use. For this reason, power electronics have been the focus of improvements in the last decade. Notably, two approaches have resulted in increased efficiency in power electronics for data centers. Firstly, high-performance, wide-bandgap (WBG) power semiconductors such as gallium nitride (GaN) and silicon carbide (SiC) have replaced traditional silicon devices. These WBG materials allow devices to operate at higher voltages, frequencies, and temperatures with greater efficiency. This means that power electronics utilizing GaN and SiC can switch more quickly and lose less power, resulting in less energy dissipation as heat. Consequently, data centers can reduce cooling requirements, leading to lower energy consumption and operational costs. The robustness of these materials also translates into smaller, lighter, and more reliable devices, making them ideal for the high-density and high-reliability environment of data centers. The ability of GaN and SiC to handle

higher power densities is critical in managing the intense power usage and thermal management challenges inherent in modern data processing centers. By integrating these advanced materials, data centers can significantly enhance their power converters efficiency as high as 98% and 99% (for SiC and GaN devices, respectively) (Horn 2023). The introduction of these advanced materials into the motor drives for cooling systems and Uninterruptible Power Supply (UPS) systems marks a significant milestone, leading to unprecedented operational efficiencies (GlobeNewswire 2023).

Secondly, data centers have increased the voltage at which electric power is distributed. This reduces the current needed to transmit electric power, reducing the power losses associated with electrical conduction (Edmonds 2022). Many data centers have increased DC distribution voltages from a historical norm of 12 V DC to 48 V DC, thereby reducing conduction losses by 16x, as losses are proportional to the square of the current (Maxim Integrated 2023). In some instances, data center owners have eliminated AC power distribution in much of their facilities, choosing instead to distribute 380 V DC from the UPS systems directly to server power supplies (Emerge Alliance 2023). This direct distribution not only increases system efficiency, but also eliminates the need for load balancing and power factor correction, issues that derive explicitly from the use of three-phase alternative current electricity distribution (O'shea 2016).

3.1.2 PACE Approaches for Reducing Computing Energy Use

The EES2 PACE working group brainstormed technology solutions and approaches wherein PACE innovations could be used to enhance the overall energy efficiency of computing infrastructure. The following technology categories emerged as a part of the discussion:

- Electricity supply innovations
- Data center power use improvements
- On-chip / On-package power management
- Architecture-specific power delivery innovations
- Dynamic computing load management
- Advanced thermal management techniques
- Enhanced modeling, simulation, and co-design capabilities

A subset of these technologies was explored by the PACE working group and resulted in the development of various action plans. These action plans detail high-level strategies that can be pursued to solve the challenges identified. The rest of the approaches were investigated by the working group but were not recommended for further consideration as a part of the EES2 roadmap. These approaches are described in a separate section within this chapter, but do not include action plans.

3.1.2.1 Dynamic Computing Load Management

Two techniques for reducing computing facility energy use by manipulating characteristics of the computing load in a facility were discussed: 1) dynamically reducing power supplied to equipment not in use, and 2) shifting demand from one data center location to a different data center location where renewable energy resources were more readily available.

Dynamically turning off power to unused equipment

As demand changes in data centers, equipment utilization varies. Reducing energy delivery in response to demand changes represents an approach for reducing overall energy use. Modern data centers employ demand reduction approaches and can shift workloads over diverse locations, optimizing power use and hardware utilization. Improvements have been made regarding idle power consumption for servers, and power management techniques in data centers empower operators to place limits on server power consumption during periods of low utilization (Matthews and Maclean 2023). Studies have shown that power management resources can reduce servers' idle power consumption by up to 11%. Despite these benefits, the same studies have shown that idle power consumption can still account for 50–90% of overall power consumption for some servers. Turning servers off completely when not in use could reduce server power consumption by an additional 30% beyond the improvements possible through power management systems (IEA 2021), and the potential exists for more broadly utilizing strategies that turn equipment off completely when not in use.

Challenges and solution pathways for dynamic computing load management

Efforts are needed to better understand the operating constraints and limitations associated with cutting power to idle equipment, such as estimating impacts on equipment availability and system flexibility. Modeling and experimentation are also needed to validate potential benefits and tradeoffs. The EES2 community can play a leadership role in exploring the associated risks and opportunities. The federal government can promote investigation through stakeholder engagement and RDD&D investment.

Resource-Aware Distributed Computing

The term resource-aware computing can refer to the efficient scheduling and allocation of workload in a single CPU across threads and cores in a multicore computing environment, across servers in a data center, or across data centers in a regional, national, or global network.

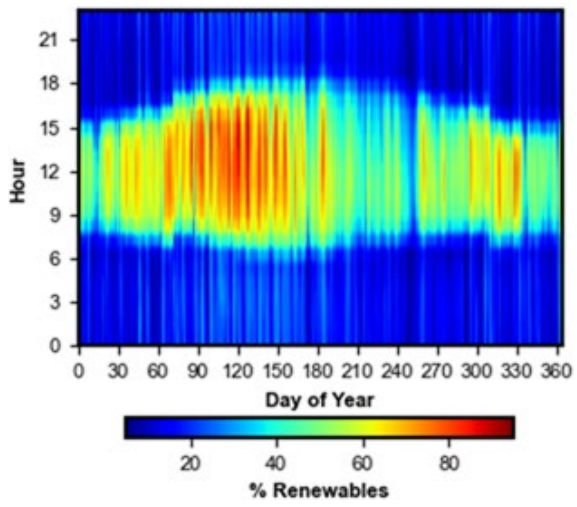


Figure 63. Daily and hourly fraction of renewable energy in the California grid for 2022.*Data source: California ISO 2024*

Scheduling work on a machine to share the computation resources (CPU, memory, I/O) most effectively among the active tasks has been practiced since the advent of the first multitasking computer systems in the 1960s. With the emergence of multicore architectures, the job of the scheduler expanded to allocate tasks across multiple compute cores (see, for example, Tillenius et al. 2015), and this scope quite naturally expanded to scheduling resources across clusters in whole data centers (Vasile et al. 2015). With widespread optical fiber data networks, the scheduling of workload can be expanded to geographically separated data centers as well. Light travels 245 km/ms in optical fiber, so in the 5–10 ms required for a magnetic disk reference, fiber

optic communication can carry data 1,225–2,450 km (roughly the distance from Washington DC to Chicago IL or Austin TX).

Recently, large data center operators have begun to consider the use of scheduling to reduce their carbon footprint. Figure 63 illustrates how large that opportunity could be in the case of electricity supply in California. The statewide supply of renewable energy shows a daily variation from less than 20% at night to more than 50% virtually every day and exceeding 80% regularly during the spring and autumn when longer days and cool weather combine to produce a high output from the state's large solar fleet and low heating and cooling loads. This temporal distribution is typical in regions with high solar energy penetration, but the pattern can be different, with more renewable energy availability at night in the windy regions of the midwestern states.

Recent work by Google (Radovanovic et al. 2023) has implemented a Carbon-Intelligent Compute Management system that is able to selectively delay the execution of temporally flexible workloads to "greener" times (when the local electricity mix is less carbon-intensive). The system, illustrated conceptually in Figure 64, monitors the forecasted carbon intensity of the utility energy supply to the facility. The system then determines heuristically or explicitly which workloads are not time-critical and which can be shifted to times with lower energy carbon content. However, the actual measurements from Google data center clusters demonstrated a power consumption drop of only 1–2% during times with the highest carbon intensity.

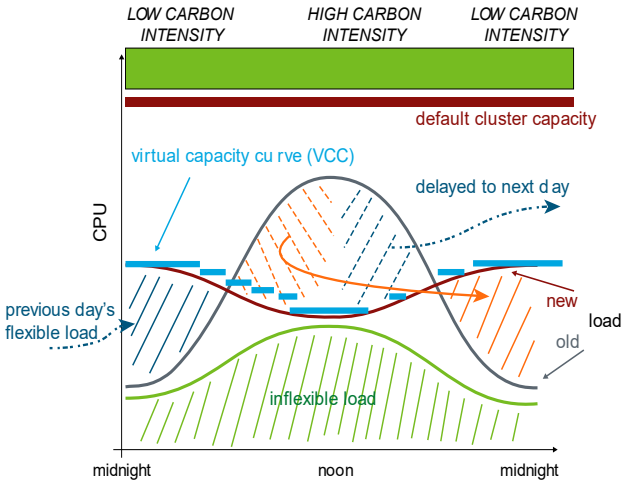


Figure 64. Google carbon-intelligent compute management data center scheduling system concept. *Source: Radovanovic et al. 2023*

A team of engineers at Microsoft and Carnegie Mellon University (Agarwal et al. 2021) introduced the concept of a virtual battery, illustrated conceptually in Figure 65. In a virtual battery model, multiple data centers are located near (perhaps collocated with) renewable energy plants and joined together via a wide-area network (WAN). This approach is a paradigm shift: instead of using techniques to adapt the availability of power to the computation demand, computational demand is

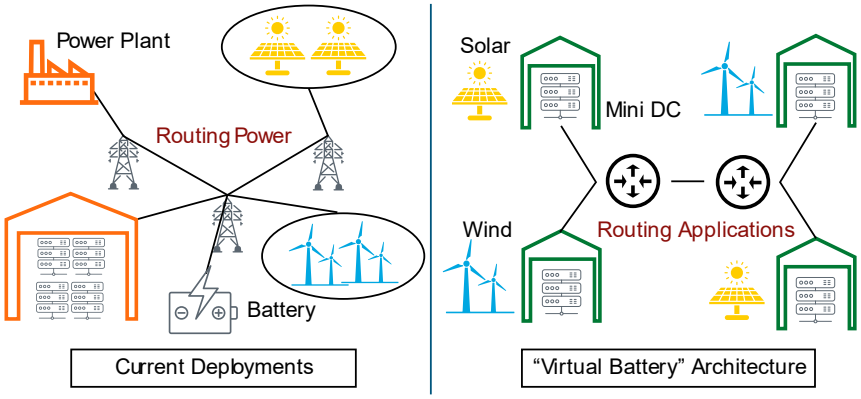


Figure 65. "Virtual battery" shifts workload between data centers in response to renewable power availability. *Source: Agarwal et al. 2021*

shifted to adapt to the availability of clean power. Virtual batteries shift demand by requiring applications to either be flexible and delay-tolerant or proactively migrating to where power is (going to be) available. The engineers noted that about half of the cost of utility-supplied electricity is due to transmission and distribution, and they further claim that the energy cost of shifting workload between data centers is negligible compared to the energy costs of transmission and distribution. Thus, a strategy of relying mostly on self-generated renewable power coupled with opportunistic geographic workload shifting can be very cost-effective.

Challenges and solution pathways for resource-aware distributed computing

This work is still in early stages, and efforts are needed to combine temporal and geographic workload shifting to achieve maximum carbon footprint abatement. But with data centers now consuming more than 1% of total electricity usage (Masanet et al. 2020), these efforts represent immediate opportunities for significant economic benefits.

The migration of workload between data centers can be accompanied by power-saving strategies such as frequency scaling and preferentially powering down components that are older and less energy-efficient.

There is a need for more comprehensive modeling to examine the comparative cost of moving the data versus moving the power and to improve real-time reporting of renewable energy fraction in power generation that can be an input to application routing.

The benefits can be amplified through the management of this process in cooperation with utility companies and grid operators. The transition to a carbon-free grid will come through massive deployment of renewable energy generation and storage systems, but also, importantly, through exploitation of load flexibility as a means of matching supply and demand. Data centers can become a significant source of load flexibility that is needed for a carbon-free grid. A lot of computation, such as database maintenance, is done on a scheduled basis; increasingly we will be able to add periodic retraining updates for ML applications to the suite of flexible computation loads. The EES2 community can play a leadership role in exploration of the intersections of smart grid modernization and data center power management for maximum benefit. Furthermore, the government can act as a convener to coordinate efforts between grid operators and data center operators to engineer the systems for the maximum benefit to both carbon abatement and grid stability.

Action plan for dynamic computing load management

Table 67. Action Plan for Dynamic Computing Load Management.

<table>
<tr><th colspan="4">Scope</th></tr>
<tr><td>Technical Challenge for Energy Efficiency</td><td colspan="3">Reducing energy use in computing facilities through manipulation of electrical load.</td></tr>
<tr><td>Technologies of Interest:</td><td colspan="3">• Reduced idle power consumption
• Resource aware compute scheduling</td></tr>
<tr><th colspan="2">Challenges</th><th colspan="2">Solution Pathway</th></tr>
<tr><td colspan="2">• Optimizing both power consumption and performance in HPC compute center job scheduling; converged cloud/HPC workloads in HPC centers.
• Reducing the power consumption of idle servers in computing facilities.</td><td colspan="2">• Extension of cloud scheduler algorithms and heuristics to optimization across multiple data centers with renewable resource availability as a constraint.
• Investigation of benefits and concerns associated with temporarily eliminating power to idle equipment.</td></tr>
<tr><th>Major Tasks / Milestones</th><th>Metrics</th><th>Targets</th><th>Timeline</th></tr>
<tr><td>Stochastic scheduling</td><td>Resource utilization</td><td>Optimal scheduling under uncertainty</td><td>3 years</td></tr>
</table>

Multi-objective optimization	Time to service/power utilization	Right pareto profile for selecting the optimal scheduling strategy	3 years
Availability of short-term/medium-term and long-term energy forecasts	Accuracy	99%	1–2 years
short/long/medium term power/load forecasting	Forecast accuracy	99% accuracy	3 years
Energy-Reliability tradeoffs in scheduling for on-prem HPC and Cloud	Energy efficiency (e.g., cooling)	Optimality, energy efficiency	3 years
Development of coordinated regional workload scheduling	Total load scheduling flexibility	>90% of data centers participating Daily shiftable load in MW	5 years
Impact assessment for shutting down idle equipment	Power reduction effectiveness	Identify operating constraints, quantify limitations for use, estimate energy impacts, verify potential reductions in equipment availability, assess impact on system flexibility	5 years

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Hardware Suppliers	Provide relevant equipment specifications and perform testing.
Data Center Operators	Evaluate operational limits and tradeoffs and perform studies.
Academia	Perform research and publish findings.
National Laboratories	Develop protocols and support development efforts.
Government	Provide targeted funding opportunities to stimulate work and act as a convener to promote cooperative scheduling and load management between grid operators and data center operators.
Other	Optimize scheduling for regional grid operators and utilities coordinate with data center operators.

Required Resources	Cross-Collaboration with Other Working Groups
Power system renewable energy/carbon intensity data.	Education & Workforce Development: Fund studies at universities; catalyze improved energy efficiency coursework.
Operational data for workload flexibility characteristics	

3.1.2.2 Advanced Thermal Management Technologies

Despite the significant advances made in cooling data centers, computing also takes place in non-data-center environments. Where improvements in cooling do not apply at the facility level, more modular, compact, or chip-level cooling strategies may be needed. In particular, the development of 2.5D and 3D chips will require direct on-chip cooling strategies that exceed the capacity of traditional, forced air cooling methods that are commonly applied today.

As the power density of electronic devices increases, conventional cooling methods become insufficient. For instance, while larger computer systems typically use heatsinks with forced air, and mobile and IoT devices rely on passive heatsinks, these approaches fall short for 3D IC devices. HPC devices are expected to exhibit power densities up to 1,000 W/cm², with stacked logic or memory tiers at 100 W/cm², and IoT devices at 10 W/cm² (Li and Goyal 2017). Moreover, hot spot densities can reach 2–4x the average power density, significantly increasing the risk of performance degradation and device failure due to overheating (IEEE HIR 2021). Current forced air system, even when combined with vapor chambers, is limited as it can only cool up to 85 W/cm², which is inadequate for the cooling demands of 3D ICs. Therefore, novel cooling methods are needed for the next generation of 3D IC cooling.

Planar 2D systems have utilized a software technique to help mitigate extreme temperatures. Dynamic thermal management (DTM) transitions a task to a cooler core when a critical temperature on an existing core is reached. While this technique is important to prevent

overheating, it is unlikely to work well in a 3D IC with a smaller footprint and stacked technologies, which have higher power density (Li and Goyal 2017). This is also a challenge on mobile devices or other applications with passive cooling. For smartphones, the processor case temperature can reach near 43°C, or 18°C above the ambient temperature near the processor (IEEE HIR 2021).

Savings potential for advanced thermal management techniques

Literature sources report investigating methods utilizing liquid or mixed-phase cooling as a potential advanced thermal management strategy. Compared to air cooling (50–85 W/cm²), these methods can range from achieving 100 W/cm² for dielectric immersion cooling, 562 W/cm² for water immersion cooling, and to 1,020 W/cm² for two-phase microchannels. Alternative cool plates can also be used for alleviating heat fluxes of 250 W/cm². These technologies are shown in

Table 68 with their performance compared to conventional air technologies and their impact factor over air cooling. Exact energy impacts are difficult to place as only water-cooled heat sinks were found to project data-center energy savings by 20x. A comprehensive study of the energy impact of these technologies is recommended.

Table 68. Various Device and Package-Level Cooling Technologies, and Their Impact over Conventional Technologies.

All technologies except for direct liquid immersion utilize thermal interface materials (TIMs), although it is not a requirement to forego them. Included is also the timeline to reach TRL 6.

Technology Group	Specified Technology	Baseline Energy Performance	Commercial Benchmark Product	Commercial Benchmark Energy Performance	Impact Factor	Timeline (years)
Device and Package Level Cooling Technologies	Diamond copper nanocomposite heat sink	900 W/m·K	Cu Heat Sink	389 W/m·K	2.3	1–3
	Water cooled Heat Sink	170–250 W/cm²	Forced Air Cooling	50 W/cm²	3.4–5	1–3
		0.23% Chip Power for Cooling		50% of Chip power for cooling	217	
	Direct liquid cooling (water immersion cooling, electrically isolated by dielectric, no TIMs)	562 W/cm²	Forced Air Cooling	50 W/cm²	11.24	5–10
	Immersion Cooling Single Phase (micro channels in the device)	790 W/cm²	Forced Air Cooling	50 W/cm²	15.8	5–10
	Immersion Cooling Dual Phase (micro channels in the device)	1,020 W/cm²	Forced Air Cooling	50 W/cm²	20.4	5–10

In addition to the technologies characterized, many additional technologies exist that have the potential to be used in microelectronics circuit cooling applications, though further

characterization is needed to quantify their potential. These technologies include thermoelectric devices, heat pipes, and magnetocaloric cooling. It is recommended that version 2.0 of the EES2 roadmap further investigate the suitability of different thermal management approaches at different computing scales and in different contexts, including distributed and scalable solutions for non-data-center facilities.

Challenges and solution pathways for advanced thermal management technologies

Infrastructure adoption and standardization

The first paper using microchannels in a device for cooling purposes was published in 1981, but the process was never commercialized due to existing infrastructure limitations at the time (Tuckerman and Pease 1981; Refai-Ahmed et al. 2020). Forced air cooling is reaching its limits and the next-generation data centers employing 3D ICs are likely to leverage water-cooled heat sinks or microchannels. This shift will require new pumps, rack layouts, liquid heat exchangers, and other components. In addition, new standards will be needed for the liquid cooling systems, such as flow rates, pressure drops, pump sizes, line lengths, fin thickness, and channel width for microfluidics, as is currently done with room air conditioners for data centers (AHRI 1360, OCP, ALSI 127, etc.).

Compatibility with chip power and interconnects

Microfluidic channels on the back side of the Si die will have significant integration challenges with chip designs that primarily use backside power. Utilizing microchannels as a TIM will require precise placement of the electrical vias between the fluidic channels (Li and Goyal 2017; Kandlikar 2014). Future designs must balance cooling and power distribution, especially for 3D circuit configurations. Packaging EDA CAD tools must be adapted to evaluate designs that configure these new cooling technologies alongside power distribution and interconnect layouts.

Reliability concerns and serviceability

As these advanced thermal management technologies have not yet been implemented broadly, the potential exists for multiple unknown failure modes and longevity concerns. Issues may include leakage of water or other liquids and the resulting impacts on devices, dielectric coating durability, and boiling of liquid coolants in contact with the devices. Failure mode and effects analysis (FMEA) should be conducted on these approaches to create mitigation plans. In addition, new components should be serviceable. If there is an equipment failure, or preventative maintenance is required, procedures must be established for the removal water or other coolants as needed to safely remove or replace components.

Distributed and scalable solutions for cooling computer equipment in non-data-center facilities

High-efficiency cooling solutions can be applied to various facilities that commonly contain significant computing infrastructure, though computation is not the primary facility purpose (e.g., healthcare facilities, universities, scientific computing centers, and logistics, shipping, and tracking companies). Since computing is often subordinate to other business functions in these environments, investments in computing energy efficiency may be lacking compared to industry-leading data centers. In these contexts, government incentives can promote the adoption of high-efficiency computing and cooling equipment, but the potential effects of this approach have yet to be quantified.

Improving technical and commercial maturity for emerging cooling technologies

Many emerging cooling solutions require further investment in RDD&D before becoming suitable for commercial use in computing facilities. Thermoelectric cooling approaches have the potential to be integrated on-package using deposition processes similar to CMOS microelectronics. However, manufacturing challenges still exist for superlattice and tunneling thermoelectric devices, which may prove to be useful approaches for achieving higher thermal-to-electrical conversion efficiencies. Similarly, heat pipes may be useful for heat removal applications in CNT-based computing architectures. However, the technological development of CNT-based transistors and heat pipes will need to move in parallel, and co-design will be needed, to allow for convergence.

Action plan for advanced thermal management technologies

Table 69 Action Plan for Advanced Thermal Management Technologies.

Scope			
Technology for Energy Efficiency	Effect System level Cooling/Full Chip cooling		
Technologies of Interest:	• Air cooling • Liquid/immersion cooling (including single-/two-phase direct liquid cooling and single-/two-phase immersion cooling) • Microfluidic cooling (single- and two-phase) • Interposer cooling technologies • Heat exchangers (for liquids cooling back down) • Thermoelectric and magnetocaloric cooling • Heat pipes		
Challenges Addressed		**Solution Pathways**	
• Thermal distribution in PCBs. • Server thermal design power (TDP). • Rack density (increasing server density). • Thermal hot spots . • Non-uniform surfaces across multi dies for cooling. • Data center sustainability by increasing facility water temperature (deal with hot air and liquid temperature). • Greater reliability of function during heat waves (e.g., not needing to decrease data use due to heat). • High-efficiency cooling solutions for all compute contexts, including non-data-center environments. • Lower TRL cooling technologies require R&D investment. More mature cooling technologies may still require de-risking. Technologies that have been demonstrated at commercial scale within data centers may still require adaptation for use in other compute contexts. Reliability analysis is required when incorporating novel cooling strategies into device packaging.		• Integrate thermal distribution material in PCB layers for normalized temperature across assembly. These materials can be “free” (e.g., copper floods) or additive (e.g., carbon layers). • Transition from air to liquid by education, economics, and adaptation of the current infrastructure. • Further decrease system thermal resistance through advanced liquid cooling techniques, through standardization and scaling up systems and reliability. • Decrease thermal resistance for removing backside heat and alleviating thermal bottlenecks (i.e., isolation of thermal crosstalk). • Remove heat from the liquid to ensure further cooling of computing components (possibly could use waste heat recovery). • Direct targeted research funding toward required R&D efforts. • Utilize grants, prize competitions, and SBIR programs to drive demonstration and deployment of cooling solutions that need to be de-risked. • Consider use of lending programs, tax incentives, and rebates to encourage the adoption of mature cooling programs in non-data-center contexts. • Measure and publish the results of each effort, to help inform the industry.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline**
Address thermal management for PCBs and assemblies	Thermal distribution across assembly	PCB design, materials and assembly	1 year

Develop R&D efforts for the various low-TRL cooling strategies.	Number of funded projects, research projects awarded	50% industry/government cost share	1–5 years
Promote adoption for higher TRL solutions.	Technology adoption rate	Increase from baseline; utilization of SBIRs and other relevant government programs	1–3 years
Achieve Standardization	Defined by a few standards communities (e.g., Asherie, ALSI 127, ANSI 1360, OCP)	Flow rates, pressure drops, mounting and unit configurations, pump sizes, line lengths, etc.	3–7 years
Investigate reliability concerns/serviceability	Leakage concern, exposure to liquid, MTBF, operations and maintenance costs	Similar device longevity failure rates to conventional air-cooled technologies, reduced O&M costs	3–7 years
Promote infrastructure adoption	Implementation of next-gen cooling techniques by infrastructure upgrades for liquid-based coolants	Microfluidic heat sinks, Immersion cooling, Single- and dual-phase device microchannels	8–12 years
Drive adoption through OpEx (CapEx high for new liquid-cooled data center)	TCO reduction through significant energy savings	Data center builders and planners, data center providers, chip producers	Ongoing

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	Develop cooling technologies and commercialize solutions.
End Users/OEMs	Energy-efficient system integration, adoption, and implementation; Define cooling requirements, verify solutions, participate in demonstrations.
Academia	Design techniques to enable thermal management improvements, such as innovations in new cooling solutions, interface design, coolants, materials (e.g., interposer, heat exchanger), floods, etc.
National Laboratories	Conduct R&D; Develop solutions; Provide testing capabilities and partnerships to enable technology validation and adoption.
Government	Set standards and provide resources and incentives for industry's transition toward higher energy efficiency; government research agencies (like the National Science Foundation and DOE's Office of Science) may be involved in funding R&D for lower TRL cooling solutions.

Required Resources	Cross Collaboration Needs of Working Groups
• Various opportunities for funding, including for manufacturing up-scaling. • Access to HPC system testing facilities. Potentially use DOE or USG-owned facilities as demonstration sites for the technologies developed. • Energy efficiency, reliability, and sustainability standards. • Facilities for testing and validation of solutions' energy efficiency, reliability, and sustainability. • Manufacturing skillset related to scale up, ensuring workforce's ability to manufacture these technologies at scale. Those who service data centers need to be educated on air-cooled data centers and liquid cooling. Possibly a need for thermal engineers to create effective cooling designs. • Education is required for end users (moving from air to liquid): why it makes sense for the business, OpEx savings, etc.	• Materials and Devices: Explore PCB construction issues for thermal distribution. Compatibility with liquid cooling. Chip design materials compatibility with exposure to liquid. • Sustainability: Determine metrics met and water/greenhouse gas footprints. • Metrology and Benchmarking: Develop system-level models to quantify impact. Requirements: Can you accommodate 25 °C water, 45 °C water? Thermal requirements on heat rejection side. • Circuits and Architectures: Design for thermal crosstalk (3D stacking, specifically), where the heat is put in and removed. Microfluidics impact architecture design. Yield at the wafer.

3.1.2.3 Enhancing Modeling, Simulation, and Co-design Capabilities

Modeling and simulation tools will need to evolve to enable the next generation of energy-efficient computing devices and facilities. Current tools do not allow for end-to-end modeling of energy use in data centers and computing facilities. This capability would allow for an improved

understanding of the relationship between device-level operations and overall facility energy use. Additionally, integrated, multi-physics co-design tools are needed to evaluate and optimize design tradeoffs involving thermal management and power delivery. Lastly, modeling and simulation tools will need extended features that enable analyses related to system reliability and economics.

Challenges and solution pathways for enhancing modeling, simulation, and co-design capabilities

End-to-end energy modeling

Simulation environments exist for modeling energy use within different scales (microelectronics level versus data center facility level), but existing tools do not provide end-to-end modeling capabilities across scales. Furthermore, due to a lack of end-to-end simulation capability, it is difficult to predict the way design tradeoffs at the circuit and architecture level will impact aggregate energy use at the facility level.

Due to the complexity and scale of modern computing infrastructure, end-to-end energy use modeling efforts suffer from long simulation times unless significant computational resources are dedicated to the task. These resources, however, are often unavailable to design teams. One potential pathway for resolving this modeling challenge would involve the use of high-performance computing (HPC) resources to generate reduced order models for energy performance. Complex, high-fidelity simulations can be created using HPC, providing realistic energy impact estimates over significant scales. Generating reduced order models would allow the insights gained using supercomputers to be accessible later when using lower capability, personal computer workstations. Using this approach, targeted R&D projects could result in tools made more widely available for design purposes.

Integrated, multi-physics co-design tools

Optimizing the design of power delivery to on-chip devices and ensuring proper removal of heat generated will require co-design in emerging circuit architectures like chip stacking and 2.5D/3D. This co-design process requires detailed simulations conducted in multiple domains (mechanical, thermal, electrical, magnetic, fluid, etc.). Tradeoffs must be evaluated across these domains. At present, software packages have been designed for analysis in each domain, and limited cross-domain analysis capabilities exist. In the future, methods may be needed for tightly coupling simulation packages from different vendors, allowing them to time-synchronize, or pass information between solvers (co-simulation). Alternatively, new simulation tools with expanded features may need to be developed. A dedicated effort may be needed to evaluate the capabilities and shortcomings of current simulation tools with respect to design for emerging architectures. Further recommendations for new features and capabilities can be made, based on that assessment.

Extending simulation tools to analyze reliability and economics

Multi-scale multi-domain co-design tools could also be used to support techno-economic analysis and reliability assessment for complex microelectronic systems. Often, a full bill of materials (BOM) is needed to derive reliability and lifetime information for a computing system. However, due to complexity, a full BOM is not often modeled in software. Large-scale simulations that include a full detailing of components and subsystems could be used to

evaluate failure modes and mitigations and to analyze the cost impacts of design decisions. These analyses are now performed indirectly by industry experts, but in the future, these insights could be made available as extensions of simulation capabilities.

Action plan for enhancing modeling, simulation, and co-design capabilities

Table 70. Action Plan for Enhancing Modeling, Simulation, and Co-Design Capabilities.

Scope			
Technology for Energy Efficiency	Develop multi-scale, single-framework, co-design tools for optimizing circuit design, power delivery, thermal performance, reliability, and economics.		
Technologies of Interest:	• Advanced modeling and simulation • High performance computing and reduced order modeling • Multi-domain physics-based modeling • Reliability analysis, cost optimization, and energy consumption modeling		
Challenges Addressed		**Solution Pathways**	
• Immense fragmentation in engineering analysis software development. • End-to-end (from device to facility) visibility regarding energy consumption within data centers. • Lack of clarity regarding limitations in sharing data between commonly used commercial software packages.		• Improving integration between software packages through design, testing, and standardization. • Utilizing HPC and ROMs to analyze energy consumption across scales. • Extending the features of existing software programs through additional code development.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline**
Baselining current use of design software and limitations regarding data sharing and collaboration	Representation, accuracy	All market-leading software packages characterized; limitations accurately depicted.	1–2 years
Development of tools and methods for enhancing collaboration in design	Usefulness, reception by industry	Broad industry acceptance and high utilization	3–5 years
Testing and validation of co-simulation approaches	Model accuracy, simulation time, insightfulness	Cross-domain optimization, global energy use reductions	5–7 years
R&D projects leveraging HPC and ROMs for end-to-end energy modeling	Model availability and utility	Development of ROMs that can be reused for analyzing different facilities with different hardware configurations	2–5 years
Development projects focused on reliability and economic modeling	Model integration, utility	Development of models that provide useful insights for informing design choices	1–4 years

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Software Vendors	Cooperate in the development of tools for extending capabilities, and in creating co-simulation capabilities.
End Users/OEMs	Define high-priority technical and economic metrics to guide development efforts.
Academia	Conduct R&D to find relevant solutions for the challenges identified.
National Laboratories	Work with industry to scale and customize solutions.
Government	Convene stakeholders, fund R&D, drive progress.

Required Resources	Cross Collaboration Needs of Working Groups
• HPC resources at DOE national labs for developing high speed, highly granular analysis tools and reduced order models. • A pre-competitive consortium for developing software extensions that benefit all vendors.	• Metrology and Benchmarking: Evaluate baseline end-to-end energy performance, assess software capabilities. • Algorithms and Software: Support in optimizing the software tools developed as a result of this effort.

3.1.3 PACE Honorable Mentions

The following energy reduction approaches were explored by the PACE working group but did not result in the development of action plans.

3.1.3.1 Electricity Supply Innovations

This category focuses on methods to reduce the carbon intensity and emissions associated with power data centers. Technologies options and strategies include:

- **Shifting power demand to align with the availability of low-carbon power supplies:** Shifting power demand would involve changing the times at which computational loads are executed (job scheduling) in response to signals provided by the local electric utility. This can be achieved by incorporating emissions implications into the optimization routines used in job scheduling algorithms for data centers. However, data center workloads are often driven by customer demands, and there are practical limits concerning the extent to which loads can be temporally shifted.
- **Utilizing energy storage to optimize low-carbon power delivery to data centers:** During times at which demand cannot be shifted, energy storage may be utilized to minimize the environmental impact of power delivery. This energy-optimization process works by storing energy from low-carbon and carbon-free sources when available (e.g., storing solar energy during peak production periods), and then dispatching the stored energy during times of high demand that cannot be time-shifted. While this energy arbitrage approach has been explored in many industries, data centers are particularly suited since they already have onsite energy storage assets. Ensuring that battery capacities remain sufficiently reserved is critical to assure that data center reliability does not suffer in exchange for decarbonization. The process of utilizing energy storage is a facility-specific optimization, design, and control challenge.

- **Dynamically switching to low-carbon, on-site fuel sources:** Hydrogen fuel represents a high potential energy source for data centers due to its ability to replace both batteries and diesel generators. While diesel generators are typically kept onsite at data center facilities for use in the event of utility power outages, hydrogen fuel cells represent a carbon-free alternative. Hydrogen production can also be used to capture excess on-peak renewable energy, which can be dispatched (instead of battery energy storage) during periods of hard-to-shift computing demand. Additionally, this on-peak renewable energy can also be dispatched during times when local electric power utilities are unable to supply low-carbon electricity. Due to the many potential value propositions offered by hydrogen, developmental efforts are underway by industry and research institutions.
- **Producing on-site renewable generation:** The use of solar PV has been explored for data centers to smooth intermittencies and to reduce power quality impacts on data centers. However, data center power densities continue to increase to the point where onsite (rooftop) PV generation has the potential to provide only a small fraction of the energy needed in an enterprise data center. Countries like Ireland are experiencing unprecedented growth in data center developments, in a region with significant plans for expansion of offshore wind generation. Similarly, data centers have been built in proximity to hydroelectric generation assets to take advantage of lower energy prices and minimize environmental impact. Market drivers have incentivized the exploration of onsite renewables for data centers, and new approaches continue to be explored by industry.

3.1.3.2 Data Center Power Use Improvements

Power distribution architecture changes within data centers

As previously mentioned, the migration to higher voltage levels and the use of DC distribution have largely addressed the efficiency gains possible in these areas.

Data center power delivery equipment efficiency improvements

The introduction of wide bandgap semiconductor devices has created higher-efficiency power delivery equipment for data centers. The introduction of these devices has been coupled with advanced monitoring and control systems that have helped to maximize energy use in modern data centers.

Reducing auxiliary data center power use

Approaches include the use of optimized cooling strategies for data centers and high-efficiency cooling equipment. These approaches have been well-integrated into modern data centers, as evidenced by the significant improvements in PUE over the last two decades.

3.1.3.3 On-Chip/On-Package Power Management

Like power supply reduction techniques used at the server level, approaches exist for reducing on-chip power consumption during times when processors are not actively in use. Dynamic voltage and frequency scaling are two techniques commonly employed by chip developers for reducing idle power consumption from processors. These techniques are routinely employed in modern chip design.

Additional approaches discussed by the PACE working group include ultra-low voltage power delivery and sub/near threshold voltage delivery.

3.1.3.4 Architecture-Specific Power Delivery Innovations

Moving forward, novel device architectures will require integrated design of circuits, cooling, and power delivery mechanisms. Optimizing power delivery approaches for each architecture represents the best path forward. Unique approaches are needed to address power delivery for 3D/chip-stacked/chiplet architectures, photonic integrated circuits, and CNT-based solutions.

As this is essentially an architecture topic, it is recommended that future EES2 roadmap development efforts include power delivery and thermal management as topics within broader architecture discussions.

3.1.4 Conclusion for Power and Control Electronics

Power and Control Electronics (PACE) is one of the critical enablers for efficient compute stacks across varied applications. This chapter emphasizes the necessity of advancing power electronics strategies to handle the increasing power demands and heat densities that accompany the next-generation computing architectures.

Key areas such as eliminating low-power modes in idle equipment and shifting compute loads to more energy-efficient or renewable-powered data centers are highlighted as immediate strategies to reduce power usage significantly. The chapter also stresses the importance of leveraging emerging thermal management technologies that allow for higher power densities in advanced packaging such as 3D integrated circuits.

The roadmap also points to the need for development and standardization of advanced tools and methodologies to assess and quantify the energy impacts at various scales—from device-level to data-center scale. These tools are essential for enabling resource-aware compute scheduling and optimizing thermal management strategies within data centers.

Overall, to align with rapid advancements in computing technology and the escalating pace of environmental concerns, it is imperative to accelerate the deployment of these PACE technologies in high energy impact areas like data centers. This includes investing in R&D to advance cooling technologies, enhancing the functionality of power delivery systems, and developing robust frameworks for continuous performance assessment and improvement.

References

Agarwal, Anup, Jinghan Sun, Shadi Noghabi, Srinivasan Iyengar, Anirudh Badam, Ranveer Chandra, Srinivasan Seshan, and Shivkumar Kalyanaraman. 2021. “Redesigning Data Centers for Renewable Energy.” Presented at the 20th ACM Workshop on Hot Topics in Networks (HotNets ’21), November 10–12, 2021. Virtual Event. https://doi.org/10.1145/3484266.3487394.

Bizo, Daniel, Rhonda Ascierto, Andy Lawrence, and Jaqueline Davis. 2021. “Uptime Institute Global Data Center Survey 2021.” Uptime Institute, Report UII-51 V1.0P. https://uptimeinstitute.com/resources/asset/2021-data-center-industry-survey.

California ISO. 2024. “Managing Oversupply.” Updated March 7, 2024. http://www.caiso.com/informed/Pages/ManagingOversupply.aspx.

De Jong, E.C.W., and P.T.M. Vaessen. 2007. “DC power distribution for server farms.” KEMA Consulting. https://www.directpowertech.com/docs/LEONARDO%20ENERGY.pdf.

Edmonds, Gary. 2022. “Using Distributed 48 V Instead of 12 V in Datacenters.” Published March 1, 2022. https://www.powersystemsdesign.com/articles/using-distributed-48-v-instead-of-12-v-in-datacenters/140/18714.

Emerge Alliance. 2023. “Standard FAQs.” https://www.emergealliance.org/standards/data-telecom/standard-faqs/.

Gillis, Alexander S., and Mark Fontecchio. 2022. “Power usage effectiveness (PUE).” TechTarget. Last updated April 2022. https://www.techtarget.com/searchdatacenter/definition/power-usage-effectiveness-PUE.

GlobeNewswire. 2023. “Wide Band Gap (WBG) Power Devices Market Worth US$ 7,717.69 Million By 2030, Says Consegic Business Intelligence.” Consegic Business Intelligence Private Limited. Published June 08, 2023. https://www.globenewswire.com/news-release/2023/06/08/2684803/0/en/Wide-Band-Gap-WBG-Power-Devices-Market-Worth-US-7-717-69-Million-By-2030-Says-Consegic-Business-Intelligence.html.

Google. 2023. “Efficiency.” Undated. Accessed December 10, 2023. https://www.google.com/about/datacenters/efficiency/.

Horn, Rolf. 2023. “Wide Bandgap Semiconductors Drive Efficiency in Datacenters.” Published April 4, 2023. https://www.digikey.com/en/articles/wide-bandgap-semiconductors-drive-efficiency-in-datacenters.

IEA. 2021. “The Idle Coefficients: KPIs to assess energy wasted in servers and data centres.” International Energy Agency (IEA) | Technology Collaboration Programme on Energy-Efficient End-Use Equipment (4E TCP). https://www.iea-4e.org/wp-content/uploads/2021/10/Server-Idle-Coefficients-FINAL-1.pdf.

IEEE HIR. 2021. “Heterogeneous Integration Roadmap, Chapter 20: Thermal.” In *Heterogeneous Integration Roadmap: 2021 Edition*. Institute of Electrical and Electronic Engineers (IEEE). https://eps.ieee.org/images/files/HIR_2021/ch20_thermal1.pdf.

Kandlikar, Satish. 2014. “Review and Projections of Integrated Cooling Systems for Three-Dimensional Integrated Circuits.” *J. Electron. Packag*. Vol. 136 (Issue 2): 024001. https://doi.org/10.1115/1.4027175.

Li, Yan, and Deepak Goyal. 2017. *3D Microelectronic Packaging: From Fundamentals to Applications.* Cham, Switzerland: Springer International Publishing AG. https://link.springer.com/book/10.1007/978-3-319-44586-1.

Masanet, E., A. Shehabi, N. Lei, S. Smith, and J. Koomey. 2020. “Recalibrating global data center energy-use estimates.” *Science*. Vol. 367 (Issue 6481): pg 984–986. https://doi.org/10.1126/science.aba3758.

Matthews, Lori, and Mark Maclean. 2023. “Reduce Server Power Usage and Save Money with Power Manager.” Dell Technologies. Published January 16, 2023. https://infohub.delltechnologies.com/p/reduce-server-power-usage-and-save-money-with-power-manager/.

Maxim Integrated. 2023. “New 48V Rack Power Architecture for Hyperscale Data Centers.” Accessed December 10, 2023. https://www.maximintegrated.com/content/dam/files/products/power/switching-regulators/48v-rack-power-architecture-for-hyperscale-data-centers.pdf.

O'shea, Paul. 2016. "48 V: The new standard for high-density, power efficient data centers." Power Electronics News. Published August 7, 2016. https://www.powerelectronicsnews.com/48-v-the-new-standard-for-high-density-power-efficient-data-centers/.

Paananen, Janne. 2023. "Grid-interactive data centers enabling energy transition." *IEEE Electrification Magazine.* Vol. 11 (Issue 3): pg 26–34. https://doi.org/10.1109/MELE.2023.3291195.

Radovanovic, Ana, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, et al. 2023. "Carbon-Aware Computing for Datacenters." *IEEE Transactions on Power Systems*. Vol. 38 (Issue 2): pg 1270–1280. http://dx.doi.org/10.1109/TPWRS.2022.3173250.

Refai-Ahmed, Gamal, et al. 2020. "Establishing the Single-Phase Cooling Limit for Liquid-Cooled High Performance Electronic Devices." Presented at the 2020 IEEE 22nd Electronics Packaging Technology Conference (EPTC). Singapore, Singapore. https://doi.org/10.1109/EPTC50525.2020.9315014.

Sun, Mengshu, Yuankun Xue, Paul Bogdan, Jian Tang, Yanzhi Wang, and Xue Lin. 2018. "Hierarchical and hybrid energy storage devices in data centers: Architecture, control and provisioning." *PloS ONE*. Vol. 13 (Issue 1): e0191450. https://doi.org/10.1371/journal.pone.0191450.

Tillenius, Martin, Elisabeth Larsson, Rosa M. Badia, and Xavier Martorell. 2015. "Resource-Aware Task Scheduling." *ACM Transactions on Embedded Computing Systems*. Vol. 14 (Issue 1): pg 1–25. https://doi.org/10.1145/2638554.

Tuckerman, David, and Fabian W. Pease. 1981. "High-performance heat sinking for VLSI." *IEEE Electron Device Letters*. Vol. 2 (Issue 5): pg 126–129. https://doi.org/10.1109/EDL.1981.25367.

Vasile, Mihaela-Andreea, Florin Pop, Radu-Ioan Tutueanu, Valentin Cristea, and Joanna Kołodziej. 2015. "Resource-aware hybrid scheduling algorithm in heterogeneous distributed computing." *Future Generation Computer Systems*. Vol. 51: pg 61–71. https://doi.org/10.1016/j.future.2014.11.019.

3.2 Manufacturing Energy Efficiency and Sustainability (MEES)

The growth and further development of artificial intelligence/machine learning (AI/ML), Industry 4.0, and the Internet of Things (IoT) will increase data analysis, communication, and semiconductor component production (Schume 2020; McKinsey & Company 2022b). With the increased use of microelectronics and their associated energy costs being covered by the compute stacks chapter, this section will now focus on the energy efficiency, resource intensity, and climate impacts of manufacturing.

There is an expectation that a new, faster, and more efficient electronic device will be released every one to two years. New devices require new silicon, new process design kits (PDKs), and—most importantly to this roadmap's scope—more steps per technology node. This resulting increase in steps ultimately results in significantly more energy costs per node.

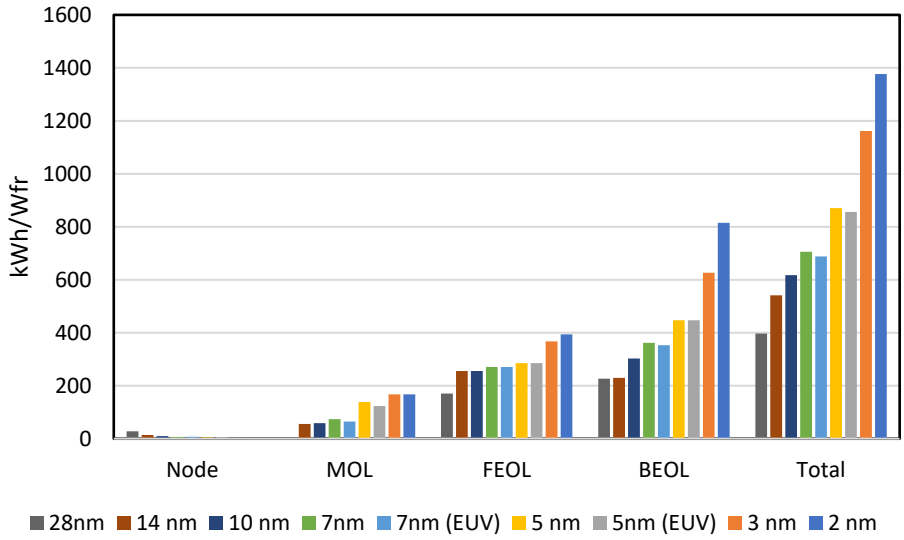


Figure 66. Manufacturing energy costs per wafer for different technology nodes. *MOL = middle of line; FEOL = front end of line; BEOL = back end of line; EUV = extreme ultraviolet. Source: Bardon and Parvais 2023*

Figure 66 illustrates the increase in manufacturing energy required per wafer, with significant increases in back end of line (BEOL) energy costs (Bardon and Parvais 2023). Transitioning from the 3nm to the 2nm node—while yielding a 15% performance improvement and a 30% reduction in power consumption at equivalent transistor counts—incurs a significant production energy increase of approximately 200kWh per wafer. For example, TSMC's nanosheet-based 2nm node, despite its advancements, only enhances chip density by about 1.1X compared to the 3nm node. The substantial manufacturing energy demands potentially outweigh the benefits of performance and power efficiency improvements in semiconductor technology advancements (Shilov 2022). Energy consumption for leading semiconductor fabs is also increasing. TSMC now consumes around 22,000 gigawatt-hours per year, a 2x energy increase over 5 years from 2017 to 2022 (Statista 2023a); Intel consumes 10.9 gigawatt-hours per year, a 2x increase over 7 years from 2015 to 2022 (Statista 2023b). Current production volumes are already taxing many grids, and projected increases in energy consumed during chip production are larger than the energy currently used by most countries, including the U.S. (Knauss 2023). While a

multitude of companies are working on increasing on-site renewable energy production or promising to buy from renewable sources, this may not be an option at all locations.

Other resource and energy consequences emerge from the increase in microelectronics production. For example, ss designs become more complex, the manufacturing process has also become more resource intensive. Each new node requires an increase in the number of lithography, etch, chemical mechanical polishing or planarization (CMP), and deposition steps. These new steps similarly increase the number of wet process steps needed to clean the wafers from residual chemistry on the surfaces and create better interfaces for the next step. Not only does this increase the quantity of water used, but these wet process steps also increase the need for intensive water recycling to remove waste impurities. While fabs recycle a lot of water for reuse (Bassler 2022), this may not work at all locations or in water-stricken areas. In fact, during a 2021 drought in Taiwan, fabs were forced to truck in water to maintain operations, even with an 85%+ recycling rate (Mott 2021).

Along with increasing water use, these processes require more materials and produce more waste gases. To meet the requirements of new technology nodes, an increased number of chemical vapor deposition (CVD), lithography, and dry etch steps produce a variety of fluorinated, high-GWP gases, such as NF_3 and SF_6. When generated, these fluoride-based compounds exhibit a high vapor pressure, enabling them to be readily evacuated toward the abatement system. However, even abatement systems with 95%+ efficiency still release significant amounts of greenhouse gass into the atmosphere. The energy equivalent of the greenhouse gass emitted due to the operational energy use of fabrication facilities will increase unless offset by renewable energy sources or the implementation of alternative non-greenhouse gas process gas.

Advanced technologies will continue to require more and more resources such as electricity and water while producing more waste and greenhouse gas emissions. The Manufacturing Energy Efficiency and Sustainability (MEES) working group focused on technologies and approaches that can mitigate these aspects. Not all technologies that could potentially help manufacturing sustainability are described here, and the technologies discussed were chosen based upon their potential impacts as well as the expertise of the working group members.

Working group methodology

Sustainable production of microelectronics has clear alignment with the EES2 goals of energy reduction and minimizing environmental impacts. Understanding that the next generation of microelectronics will require more resources and produce more waste and greenhouse gass, the MEES working group identified 26 technical areas across 4 different technology groups as technologies worthy of investigation. Table 71 is a list of these technology areas along with specific technologies in these areas that were discussed by the working group. Given the available bandwidth and expertise of the working group members, only the bolded technologies were chosen to be investigated. For the next iteration of the roadmap, additional technologies will be explored.

Table 71. MEES Technology Groups and Specified Technologies.

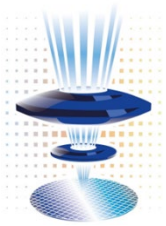
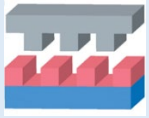
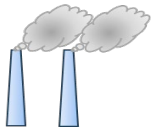
Technical Group	Specified Technology

Alternative and optimized processes for energy and waste	• **Reduce energy consumption of lithography of EUV, and through adoption of nanoimprint lithography** • Identification of high energy bottlenecks, device designs, and process improvements • Selective deposition and etch processes • Bottom-up self-assembly for FEOL • Reduce solvent usage • **Minimize high-GWP (global warming potential) gases (e.g., SF_6, NF_3) through new deposition/etch processes** • Low-PFAS (per- and polyfluoroalkyl substances) materials
Facilities considerations	• Preliminary ideas: wall power, green equipment, water use • Allow for purified compressed air instead of resource-intense gases such as pure N_2 or He when applicable • Energy recovery of waste heat or other • Reduce air filtration to level needed • Facility location optimization primarily in resource- and renewable-energy-rich areas • Additive manufacturing for improved equipment to reduce variations
Sustainable manufacturing practices	• Recycling of key waste streams, e.g., high-value metals from slurries • Water-optimized processes and recycling efforts; target net-zero or net-positive use • Green energy procurement and/or on-site generation • More efficient heating and cooling processes (facilities, tools) • **Improvement of abatement technologies to reduce GWP of byproduct gases or capture for reuse** • Development of life cycle inventory (LCI) identifying energy and materials footprint of advanced integrated circuits (Ics)
E-waste avoidance (including recycling)	• Design for reuse (labeling of components) • Incentives for original equipment manufacturers (OEMs) to recycle • Design hardware for forward compatibility to avoid waste (e.g., chiplet technologies) • Build for disassembly

Key takeaways

summarizes the most significant identified energy efficiency opportunities that can be achieved through advances in Manufacturing Energy Efficiency and Sustainability (MEES).

Table 72. Key Opportunities for Energy Efficiency and Sustainability in MEES.

Technology Group	Key Opportunities for Energy Efficiency
Light-based Lithography	 • Streamline energy usage by refining laser sources for **DUV lithography** and implementing process gas recycling. Enhance photoresist formulations to improve light sensitivity, reducing exposure time. • Optimize plasma generation chamber designs and mirror technology to maximize light utilization for **EUV lithography**. Improve the efficiency of high-powered laser systems and their maintenance processes through innovative system design.
Imprint-based Lithography	 • Eliminate the need for complex light sources and reduce the overall energy footprint. For example, **nanoimprint lithography** has high energy efficiency potential by leveraging its direct mechanical patterning approach.
Process Gas Abatement	 • Develop compact **abatement** solutions with superior destruction and removal efficiency to significantly lower greenhouse gas emissions in semiconductor manufacturing. Foster the adoption of alternative process gases for cleaning, deposition, and etching that have lower global warming potential (GWP).

Grand challenges

The main challenges for improving manufacturing energy efficiency and sustainability are:

- Improving energy efficiency of EUV lithography by seeking breakthroughs in equipment design and process efficiencies.
- Optimizing nanoimprint lithography to compete with traditional photolithography methods, with a focus on minimizing defects, scaling down in size, enhancing alignment precision, and improving stamp lifetime.
- Adopting and creating compact, high-efficiency abatement systems that fit within the spatial constraints of fabs, with higher destruction and removal efficiency and minimal floor space requirements.
- Identifying and implementing lower-GWP gases for etching and cleaning that maintain tool performance while reducing environmental impact.
- Establishing comprehensive energy metrics for each lithographic process, from DUV to EUV, to ensure accurate assessment and benchmarking of energy consumption.
- Developing efficient recycling methods for process gases to curtail emissions and operational costs in photolithography.

3.2.1 Lithography

Lithography, a central process in semiconductor manufacturing, transfers intricate circuit patterns onto a silicon wafer. Since the inception of chip fabrication, lithography has undergone significant transformations to accommodate the persistent demand for smaller, faster, and more energy-efficient microelectronics.

Photolithography, which uses ultraviolet (UV) light to etch designs onto silicon wafers, is the dominant pattern transfer technique. Wafers are first coated with a light-sensitive chemical layer, known as a photoresist. Once applied, the wafer is exposed to UV light that has passed

through a mask (a stencil containing the desired layer design). This UV light exposure causes chemical changes in the photoresist, making the exposed area soluble or insoluble depending on the type of resist used. Following exposure, the wafer undergoes a development process where the soluble regions are washed away to create a precise replica of the mask pattern. Historically, two wavelengths were generally employed: 248nm and 193nm. The 248nm wavelength, used in deep ultraviolet (DUV) lithography, was used early on in photolithographic technology and utilized KrF lasers to produce the desired UV light. As chip technology advanced and there was a pressing need to pattern smaller features, the 193nm wavelength—using ArF lasers—was adopted. Unfortunately, the 193nm lithography process typically consumes more energy than the 248nm process due to the increased complexities in light-source generation and the ancillary equipment required to manage and optimize the shorter wavelength. In fact, energy consumption in the 193nm process can range from several to tens of kWh per wafer depending on the specific photolithographic process, the equipment used, and the design intricacy.

The relentless pursuit of miniaturization in the semiconductor industry has further driven the transition toward extreme ultraviolet lithography. EUV lithography, utilizing a much shorter wavelength of approximately 13.5nm, enables smaller, more precise pattern replication. But generating EUV light requires high-power laser systems and specialized equipment, leading to a substantial increase in energy consumption relative to traditional DUV processes.

As a result, the industry is actively researching energy-efficient solutions within EUV. Additionally, alternative lithography methods—such as nanoimprint lithography, which physically stamps patterns onto surfaces—are being explored as potential successors or complements to EUV since they offer both precision and potentially improved energy profiles.

In this chapter, various lithography techniques are discussed alongside strategies to reduce the energy footprints without compromising technology progression.

DUV lithography

DUV is the primary lithography technique for legacy nodes (10nm and above). As with all lithography, generating the precise UV light requires careful control, stable environments, and intricate machinery, all of which come at a substantial energy cost. Over the years, tool suppliers have been optimizing the light production process. For example, by advancing photoresist sensitivity and refining optics to reduce light scattering, exposure time has decreased, which not only conserves energy per wafer but also accelerates the overall process. Innovations in machinery, such as advanced cooling systems, have further reduced the energy footprint of these lasers. Cymer, a leading manufacturer of DUV laser sources, was able to develop a master oscillator (MO) chamber that helped reduce power consumption by ~15%. Additionally, the shift to neon gas, which offers cost and supply advantages compared to helium, was complemented by a new system adept at capturing, recycling, and supplying over 90% of the neon gas needed by ArF sources (Roman et al. 2017).

DUV lithography has also reached a resolution limit for its most advanced technique, ArF immersion lithography, at around the 40nm to 20nm nodes. Beyond this limit, the resolution and pattern fidelity deteriorate rapidly, making it challenging to produce reliable semiconductor devices. The resolution is proportional to wavelength/numerical aperture, thus the fundamental

limit is governed by the wavelength. To push beyond the 10nm limitation, EUV lithography with a wavelength of 13.5nm has been developed and is currently being commercialized.

EUV lithography

Extreme ultraviolet lithography operates at a wavelength of approximately 13.5nm and can pattern features down to 7nm and below. To generate light at this wavelength, plasma from tiny droplets of molten tin is excited, which emits light within the EUV spectrum. However, conventional optics cannot be used to manipulate EUV light due to its unique absorption characteristics. Instead, EUV light is directed onto the silicon wafer using a series of specialized, multi-layer coated mirrors. The same characteristic that makes EUV able to pattern tiny features (i.e., wavelength) also makes it energy inefficient. Current top-down estimates of power outputs and efficiencies for 13.5nm EUV technology compared to traditional immersion lithography are provided in Table 73.

Table 73. Energy Consumption of EUV vs. DUV Lithography. *Source: Kim 2009*

Metric	200W output EUV	90W output ArF immersion double patterning
Electrical power (kW)	532	49
Efficiency (%)	0.04%	0.18%
Ratio of input power/output	2,660	544.44

Even relative to double-patterning immersion lithography, the lower efficiency and resulting higher input power (~5x) of EUV are evident. In addition to this top-down analysis, a bottom-up evaluation based on bond energies further underscores the heightened energy usage of EUV. Specifically, energy metrics for deposition, lithographic, and etch processes highlight that EUV can require higher energy per bond compared to its 193nm DUV counterpart (Shankar 2023). Table 74 shows that the deposition/growth and etch energetics are bound at the low end (2.1 eV) and high end (8.42 eV), which correspond to copper-metallic bond energy and copper-tantalum bond energy, respectively. The total energy per bond varies between approximately 3x and 5x for EUV compared to double patterning in DUV technology.

Table 74. Total Energy Per Bond for DUV vs. EUV Lithography. *Source: Shankar 2023*

Wavelength	Process Flow		Deposition/ Growth	Litho	Etch	Litho	Etch	TOTAL Energy per Bond (eV)	% for Lithography
193 nm	Double Patterning	Min	2.1	6.41	2.1	6.41	2.1	19.12	67.05%
193 nm	Double Patterning	Max	8.42	6.41	8.42	6.41	8.42	38.08	33.67%
13.5 nm	Single Patterning	Min	2.1	91.67	2.1			95.87	95.62%
13.5 nm	Single Patterning	Max	8.42	91.67	8.42			108.51	84.48%

Addressing the significant energy requirements of EUV lithography is paramount, especially as semiconductor processes inch toward even smaller dimensions. The process of generating EUV light is inherently energy intensive and the energy footprint spans from high-powered lasers to the subsequent cooling and maintenance of the machinery. To address these concerns, EUV

lithography manufacturers like ASML are continuously innovating chamber designs to optimize plasma generation and advancing mirror technology to reduce light losses. See Figure 67 for additional details regarding ASML's roadmap for upcoming EUV technology.

Industry leaders and associated entities are actively pinpointing processes with significant carbon footprints and strategizing how to diminish them (Ragnarsson et al. 2022). Such approaches to curb energy expenditures in lithography/pattern transfer include, enhancing the design of processing equipment for better efficiency, exploring alternative bottom-up processes like directed self-assembly, and refining patterning techniques at nanoscale dimensions. Given the higher energy requirements of EUV relative to traditional immersion lithography, it is important that these energy efficiency efforts be extended to the design of the lithographic processes and equipment.

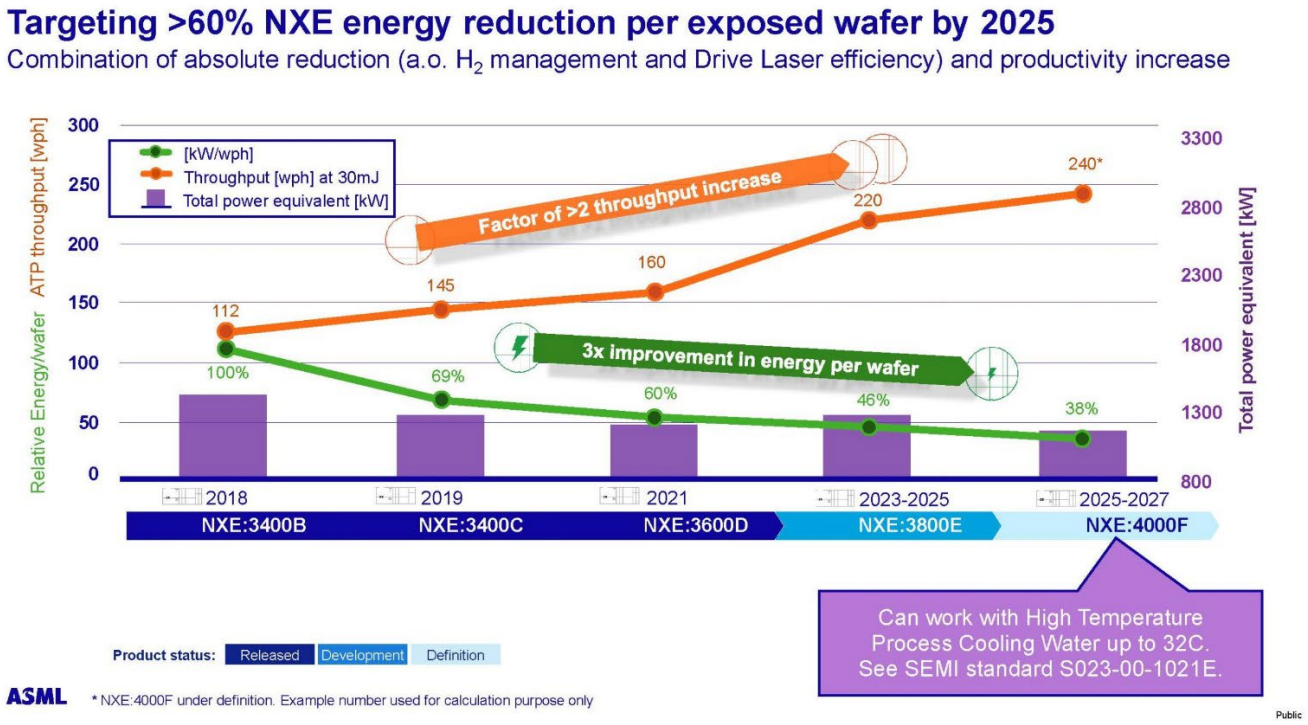


Figure 67. Roadmap of EUV lithography tool developed by ASML. *Source: Jones 2022*

Nanoimprint lithography

Nanoimprint lithography (NIL) is gaining momentum as an alternative to EUV lithography. At its core, nanoimprint lithography operates much like a stamping process (see Figure 68): A mold with the desired patterns is pressed into a resist layer placed on the substrate. This imprinting process physically deforms the resist, replicating the mold's patterns onto the substrate. Once the imprint is made, residual layers are typically removed, followed by etching to transfer the pattern into the substrate. NIL has demonstrated its ability to achieve features well below the 10nm scale and is currently being commercialized by tool providers such as Canon, positioning it as a primary alternative for next-generation semiconductor devices.

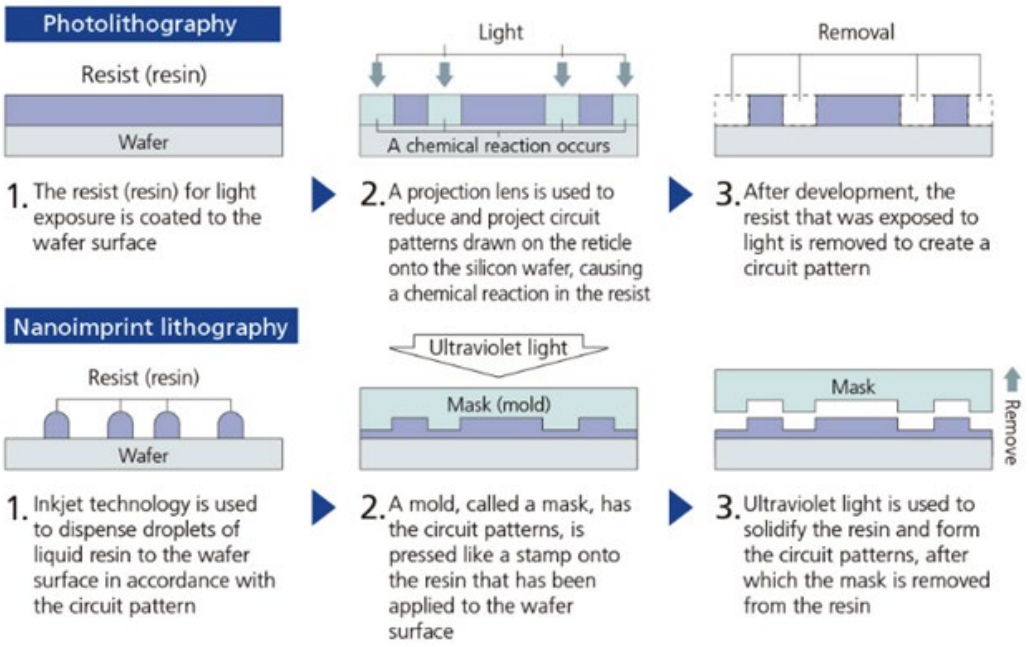


Figure 68. Photolithography vs. nanoimprint lithography processes. *Source: Canon 2019*

From an energy standpoint, NIL is significantly more energy efficient than EUV ilthography (see Table 75). NIL sidesteps the need for complex light sources, such as those used in DUV and EUV lithographies, as well as the associated energy-intensive processes for generating specific wavelengths of light. Instead, NIL's direct mechanical patterning approach substantially reduces its energy footprint, making it a more eco-friendly option.

Table 75. Power Consumption of EUV vs. NIL Processes. *Source: DNP 2023*

Process	Power Consumption of Lithography Process
EUV Lithography	9.7 kWh/wafer
NIL	1.1 kWh/wafer

However, like all innovative technologies, NIL faces its own set of challenges, including size scaling, defect control, and imprint alignment. Among these, defect control remains the biggest challenge. For example, the direct contact between the mold and resist can lead to imperfections or damage that could affect device performance. NIL is currently being used in memory, but Canon has recently announced that they are moving NIL into logic in order to compete with ASML when it comes to precision (Mann 2023). If this comes to fruition, an orders-of-magnitude decrease in energy consumption is possible. But size scaling remains another issue. Currently, NIL masks and EUV masks are made through similar processes. However, there is a nascent process that allows direct duplication of patterns down to 1nm–2nm using NIL—but that template-making process is slow, possibly limiting NIL's near-term prospects for manufacturing critical components (Hua et al. 2004).

Given these considerations, the semiconductor industry is deliberating the best path forward for nanoimprint lithography. While NIL may not completely replace optical lithography due to its

unique challenges, it could very well be used in tandem with other methods. For example, employing NIL for specific layers or processes where its benefits are most pronounced, while also relying on traditional lithographic methods for others, which might offer an optimal blend of precision, energy efficiency, and throughput. As the demand for smaller, more efficient devices grows, integrating techniques like NIL alongside established processes can be key to reducing the energy consumption of chip manufacturing.

3.2.2 Process Gas Abatement Systems

The growth of the microelectronics industry and the increased complexity and number of production steps have resulted in an overall increase in greenhouse gas emissions. Fluorinated gases, which can escape into the atmosphere, are commonly used for etch and chamber cleaning processes. Table 76 presents common gases used in semiconductor processing and their GWPs.

Table 76. GWPs and Atmospheric Lifetimes of Key Waste Gases. *Source: Beu et al. 2019*

Chemical	Global Warming Potential	Atmospheric Lifetime (Years)
CO_2	1	20
CH_3F	150	3
N_2O	310	120
CF_4	6,500	50,000
C_2F_6	9,200	10,000
CHF_3	11,700	160
NF_3	17,200	500
SF_6	23,500	3,200

IMEC recently published a report looking at the sustainability of next-generation chip manufacturing. Their findings showed that there are still significant emissions resulting from N_2O, CHF_3, SF_6, NF_3, and CF_4, with the latter three representing 93% of wafer emissions. The report also showed that emissions per wafer have increased by 2.7x from the 28nm node to the 3nm node. Investigation into abatement systems is needed to effectively remove contaminants from the production line since they can save ~40% of total emissions (excluding onsite power generation) (McKinsey & Company 2022a).

The destruction or removal efficiency values reported in the Intergovernmental Panel on Climate Change's (IPCC) 2019 *Electronics Industry Emissions* report (Beu et al. 2019), alongside IMEC and state-of-the-art values for multiple different gases, are shown in Table 77 (Ko et al. 2014; Hur et al. 2016; Applied Materials 2023; Lee and Chen 2017). For the three gases that contribute 93% of the process waste gases, there is an average of 93% removal according to the IPCC. While this may seem very good, more can still be done to improve abatement (Bardon et al. 2020). While these technologies can be improved significantly with the state of the art, challenges remain to abatement technologies and process gases, which are discussed below.

Table 77. Destruction and Removal Efficiency Values for Key Process Gases. *Source: Beu et al. 2019*

Process Gas	U.S. Destruction and Removal Efficiency (DRE)	IMEC DRE	State of the Art DRE	Improvement Factor
CH_3F	98%	90%	99%	101%
N_2O	60%	90%	98.7	165%
C_2F_6	95%	90%	100%	105%
CF_4	89%	90%	90%–95%	105%
CHF_3	98%	90%	99%	101%
NF_3	95%	95%	99.1%–100%	105%
SF_6	95%	90%	99%	104%

Challenges and Solution Pathways for Process Gas Abatement Systems

Compact Designs With Higher Destruction and Removal Efficiency

Fabs are designed with a subfloor to accommodate additional components to the tool set, which can include pumps, power sources, gas generators, and abatement tools. Since there is limited space available—especially for facilities producing older nodes where increases in floor space are not possible—creation of small-footprint and little-to-no-floorspace abatement systems in line before the pumps is needed. Additionally, the abatement systems need to have a higher destruction and removal efficiency while maintaining a smaller footprint for effective greenhouse gas removal. A one-size-fits-all approach may be difficult to address these challenges without lifetime issues in the tools and parts. While technologies do exist that require no floorspace and are smaller and more efficient than the IPCC-reported values, it may not be cost effective for all smaller producers to upgrade. Providing financial assistance and/or loans for abatement systems to reduce GHG impacts may support wider adoption.

Alternative process gases for cleaning, deposition, and etching

The use of NF_3, SF_6, and perfluorocarbons CH_xF_y occurs primarily in the etching of materials—whether it be for structural or 3D components of films on the wafer, or removal of films from chamber walls to ensure that the tool is performing within its specifications. Despite high removal efficiency, significant gas emissions still occur. Solutions include using alternative chemistries for cleaning atomic layer deposition (ALD) and chemical vapor deposition (CVD) chambers or etching with lower global warming potential (GWP) gases like F_2 plasmas (Hwang et al. 2007; Riva et al. 2009). Additionally, thermal reactivity processes involving metals, metal oxides, and nitrides with agents like ozone, sulfuryl chloride, HF, and transfer ligands could be adapted for chamber cleaning if sufficiently rapid (Partridge et al. 2023a; Partridge et al. 2023b; Johnson et al. 2016). However, challenges remain: NF_3 and SF_6's inertness is hard to match with more reactive chemicals like F_2, HF, or $SOCl_2$, requiring stringent handling precautions. Moreover, many of these alternative methods are still in the research phase, posing significant hurdles before industrial application is feasible.

3.2.3 Conclusion for Manufacturing Energy Efficiency and Sustainability

Manufacturing Energy Efficiency & Sustainability (MEES) plays a vital role in enabling the EES2 roadmap by ensuring that semiconductor manufacturing processes align with modern energy efficiency demands. As an enabler, MEES aims to advance techniques, standards, and methodologies that will shape the next generation of sustainable manufacturing in the semiconductor industry.

Core strategies highlighted include the adoption and efficiency improvement of light- and imprint-based lithography techniques, which offer improved efficiency in the production of advanced microelectronic components. Process gas abatement systems are essential to mitigate emissions and environmental impacts during semiconductor fabrication, or ideally, replace greenhouse gas-emitting process gases altogether with more environmentally friendly alternatives.

MEES acts as a cornerstone for EES2's objectives by improving manufacturing throughput, reducing environmental impact, and enabling the rapid scaling of energy-efficient semiconductor technologies. Collaboration between industry leaders and researchers will ensure these initiatives are accelerated, securing a more sustainable future for semiconductor manufacturing.

References

Applied Materials. 2023. “Aeris®-G Plasma Abatement System.” Accessed December 2023. https://www.appliedmaterials.com/us/en/product-library/aeris-g-plasma-abatement-system.html.

Bardon, M.G., P. Wuytens, L.-Å. Ragnarsson, G. Mirabelli, D. Jang, G. Willems, A. Mallik, A. Spessot, J. Ryckaert, and B. Parvais. 2020. “DTCO including Sustainability: Power-Performance-Area-Cost-Environmental score (PPACE) Analysis for Logic Technologies.” Published in Proceedings of the 2020 IEEE International Electron Devices Meeting (IEDM). San Francisco. https://doi.org/10.1109/IEDM13553.2020.9372004.

Bardon, Marie Garcia, and Bertrand Parvais. 2023. “The Environmental Footprint of Logic CMOS Technologies.” imec. Accessed December 2023. https://www.imec-int.com/en/articles/environmental-footprint-logic-cmos-technologies.

Bassler, Hunter. 2022. “Yes, Semiconductor Plants Use a Lot of Water, but the Vast Majority is Recycled and Returned.” 12 News. Last modified December 8, 2022. https://www.12news.com/article/news/local/water-wars/how-much-water-do-semiconductor-chipmaking-plants-use-arizona-tsmc-fabs/75-bddc3623-b247-408f-a618-1905455009d.

Beu, L., S. Raoux, Y.C. Chang, M.R. Czerniak, F. Illuzi, T. Kitagawa, D. Ottinger, and N. Parasyuk. 2019. “Chapter 6: Electronics Industry Emissions.” In Volume 3: Industrial Processes and Product Use, *2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories*, 6.1–6.81. https://www.ipcc-nggip.iges.or.jp/public/2019rf/pdf/3_Volume3/19R_V3_Ch06_Electronics.pdf.

Canon. 2019. “Nanoimprint Lithography.” Accessed December 2023. https://global.canon/en/technology/frontier07.html.

DNP. 2023. “Nano-Imprint Lithography (NIL).” Accessed December 2023. https://www.global.dnp/biz/solution/products/detail/10162425_4130.html.

Hua, F., Y. Sun, A. Gaur, M.A. Meitl, L. Bilhaut, L. Rotkina, J. Wang, et al. 2004. "Polymer Imprint Lithography with Molecular-Scale Resolution." *Nano Letters.* Vol. 4 (Issue 12): pg 2467–2471. https://doi.org/10.1021/nl048355u.

Hur, M., J.O. Lee, J.Y. Lee, W.S. Kang, and Y-H Song. 2016. "Abatement Characteristics of N_2O in Low-Pressure Plasma Reactor." *Plasma Sources Science and Technology*. Vol. 25 (Issue 1): 015008. https://iopscience.iop.org/article/10.1088/0963-0252/25/1/015008.

Hwang, J.Y., D.J. Kim, N.-E. Lee, Y.C. Jang, and G.H. Bae. 2007. "Chemical Dry Etching of Silicon Nitride in F_2/Ar Remote Plasmas." *Surface and Coatings Technology.* Vol. 201 (Issues 9–11): pg 4922–4925. https://doi.org/10.1016/j.surfcoat.2006.07.081.

Johnson, Nicholas R., Huaxing Sun, Kashish Sharma, and Steven M. George. 2016. "Thermal Atomic Layer Etching of Crystalline Aluminum Nitride Using Sequential, Self-Limiting Hydrogen Fluoride and $Sn(acac)_2$ Reactions and Enhancement by H_2 and Ar Plasmas." *Journal of Vacuum Science & Technology A.* Vol. 34 (Issue 5): 050603. https://doi.org/10.1116/1.4959779.

Jones, Scott. 2022. "ASML EUV Update at SPIE." SemiWiki, TechInsights section, digital image. Accessed February 28, 2024. https://semiwiki.com/semiconductor-services/techinsights/314387-asml-euv-update-at-spie/.

Kim, Hyeong Soo. 2009. "Future of Memory Devices and EUV Lithography." Presented at the EUV Symposium, October 2009, Hynix Semiconductor Inc. PDF presentation. Prague, Czech Republic. http://www.sematech.org/meetings/archives/litho/8653/pres/Keynote3_Kim_Hynix.pdf.

Knauss, Tim. 2023. "How would Micron's electricity-hogging plant here live with NY's war on fossil fuels?" Syracuse.com. Last modified February 28, 2023. https://www.syracuse.com/news/2023/02/how-would-microns-electricity-hogging-plant-here-live-with-nys-war-on-fossil-fuels.html.

Ko, D.G., S.J. Ko, E.K. Choi, S.G. Min, S.H. Oh, J. Jung, B.M. Kim, and I.-T. Im. 2014. "Perfluorocarbon Destruction and Removal Efficiency: Considering the Byproducts and Energy Consumption of an Abatement System for Microelectronics Manufacturing." *IEEE Transactions on Semiconductor Manufacturing.* Vol. 27 (Issue 4): pg 456–461. https://doi.org/10.1109/TSM.2014.2362942.

Lee, How Ming, and Shiaw-Huei Chen. 2017. "Thermal Abatement of Perfluorocompounds with Plasma Torches." *Energy Procedia*. Vol. 142: pg 3637–3643. https://doi.org/10.1016/j.egypro.2017.12.256.

Mann, Tobias. 2023. "Canon Claims Its Nanoimprint Litho Machines Capable of 5nm Chip Production." The Register. Published October 13, 2023. https://www.theregister.com/2023/10/13/canon_nanoimprint_litho/.

McKinsey & Company. 2022a. "Sustainability in Semiconductor Operations: Toward Net-Zero Production." Published May 17, 2022. https://www.mckinsey.com/industries/semiconductors/our-insights/sustainability-in-semiconductor-operations-toward-net-zero-production.

McKinsey & Company. 2022b. "What is the Internet of Things?" Published August 17, 2022. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-the-internet-of-things.

Mott, Nathaniel. 2021. "Report: TSMC and UMC are Trucking in Water to Prevent Further Chip Shortages." Tom's Hardware. Published February 24, 2021. https://www.tomshardware.com/news/tsmc-umc-hit-by-drought-trucking-in-water.

Partridge, J.L., A.I. Abdulagatov, V. Sharma, J.A. Murdzek, A. Cavanagh, and S.M. George. 2023a. "Thermal Atomic Layer Etching of CoO using Acetylacetone and Ozone: Evidence for Changes in Oxidation State and Crystal Structure during Sequential Exposures." *Applied Surface Science.* Vol. 638: 157923. https://doi.org/10.1016/j.apsusc.2023.157923.

Partridge, J.L., J.A. Murdzek, V.L. Johnson, A.S. Cavanagh, A. Fischer, T. Lill, S. Sharma, and S.M. George. 2023b. “Thermal Atomic Layer Etching of CoO, ZnO, Fe_2O_3, and NiO by Chlorination and Ligand Addition Using SO_2Cl_2 and Tetramethylethylenediamine.” *Chemistry of Materials.* Vol. 35 (Issue 5): pg 2058–2068. https://doi.org/10.1021/acs.chemmater.2c03616.

Ragnarsson, Lars-Åke, Cedric Rolin, Sheron Shamuilia, and Els Parton. 2022. “The Green Transition of the IC Industry.” imec. https://imec-int.com/sites/default/files/2022-07/Whitepaper_SSTS_FINAL.pdf.

Riva, M., M. Pittroff, T. Schwarze, R. Wieland, and J. Oshinowo. 2009. “Superior Etch Performance of Ar/N2/F2 for PECVD Chamber Clean.” In Proceedings of the 2009 IEEE/SEMI Advanced Semiconductor Manufacturing Conference, Berlin, Germany. Pg 128–132. http://dx.doi.org/10.1109/ASMC.2009.5155971.

Roman, Y., D. Kanawade, W. Gillespie, S. Luo, M. Thever, T. Duffey, K. O’Brien, et al. 2017. “Advances in DUV Light Source Sustainability.” Published in Proceedings of SPIE 10147, Optical Microlithography XXX, 101471Y (June 20, 2017). https://doi.org/10.1117/12.2260307.

Schume, Philipp. 2020. “Improve Product Quality and Yield with Intelligent, Secure, and Adaptable Manufacturing Operations.” IBM. Published April 17, 2020. https://www.ibm.com/blog/iot-manufacturing-ready/.

Shankar, Sadas. 2023. “DOE EES2 Roadmap Meeting #2.” U.S. Department of Energy | Energy Efficient Computing, Roadmap Meeting 2, Semiconductor Industry Energy Efficiency Scaling (January 12, 2023). https://ees2.slac.stanford.edu/doe-meetings-events/doe-ees2-roadmap-meeting-2.

Shilov, Anton. 2022. “TSMC Reveals 2nm Node: 30% More Performance by 2025.” Tom’s Hardware. Published June 16, 2022. https://www.tomshardware.com/news/tsmc-reveals-2nm-fabrication-process.

Statista. 2023a. “Annual Energy Consumption Taiwan Semiconductor Manufacturing Company (TSMC) from 2016 to 2022.” Published August 2023. https://www.statista.com/statistics/1312965/tsmc-energy-consumption-by-source/.

Statista. 2023b. “Intel’s Energy Use Worldwide from 2015 to 2022, by Type.” Published June 2023. https://www.statista.com/statistics/1200893/intel-energy-use-by-type-worldwide/.

3.3 Metrology and Benchmarking

Metrology plays a central role in the R&D and manufacturing of microelectronic devices and comprised of a diverse array of tools, techniques, and analysis methods. In this roadmap, "metrology" includes both inline, high-throughput measurements as well as off-line advanced characterization techniques and everything in-between. As geometric scaling progressed, the process tolerances and specifications became more stringent and the number and importance of metrology steps increased, as did the measurement requirements (e.g., resolution and measurement time). For instance, comparing the manufacturing processes of two generations of microchips, one can observe a significant increase in the complexity and number of metrology steps required. In manufacturing a chip with 14nm features—smaller and more advanced than one with 65nm features—the number of metrology and inspection steps is four times greater.

More recent advancements, such as 3D stacking and heterogeneous integration, push into the vertical direction. While these innovations exhibit improved performance, energy efficiency, and multifunctionality, they introduce a whole new set of metrology challenges and requirements: minor discrepancies in alignment or defects can cascade through the layers, degrading the system's overall performance and efficiency; novel materials and/or device architectures may require novel characterization techniques; and critical structures and interfaces may no longer be accessible. While traditional approaches continue to provide value, they will not meet the metrology needs for emerging energy efficient devices and systems. To enable the technologies found within this roadmap, advancements in existing metrology techniques or development of new ones—to meet these technologies' specific requirements—are needed.

In conjunction, standardized energy performance measurements (i.e., benchmarking) are needed to objectively evaluate the array of emerging technologies, including those proposed in the roadmap. Benchmarks provide a foundational reference, allowing for the systematic assessment of innovations against a consistent criterion. While each technology may have attributes other than energy that make it suitable for a specific computing application, data from benchmarking will nonetheless help prioritize and possibly downselect the technology options.

Working group methodology

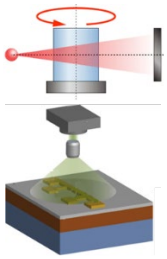
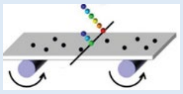
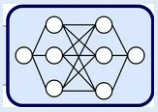
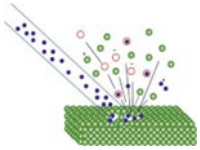
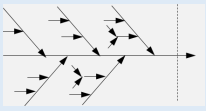
The discussions in this working group were structured differently from the other working groups. Given that innovations in metrology and benchmarking will not confer direct improvements in energy efficiency, the group did not quantify energy efficiency impacts. Instead, the group identified critical metrology and benchmarking approaches and discussed the desired future state for each, summarized in Table 78. These approaches set the context for future discussions of challenges and solution pathways for metrology and benchmarking group.

Key takeaways

Table 78 summarizes the most significant identified energy efficiency opportunities that can be achieved through advances in metrology and benchmarking.

Table 78. Key Opportunities for Energy Efficiency in Metrology and Benchmarking.

Technology Group	Key Opportunities for Energy Efficiency

3D Metrology		• Development of non-destructive **interface metrology techniques** that are strongly connected to characterization and measurement data models. • **X-ray tomography** to evaluate buried structures and interfaces, with 100nm resolution, large field of view, and with scan times of seconds to minutes. • **Thermal metrology** to characterize interfacial thermal resistance and gradients, as well as detect hot spots, with 0.1 milli-Kelvin resolution.
In-situ and In-operando Characterization for Fabrication		• Development of more widely available techniques with fast feedback times, including those for composition, thickness, conformity, etc. • Process-specific measurements, such as those for **RF plasma**, to enable strict process control during fabrication. • The integration of deposition tools with **X-ray measurements** to ensure consistent depth and uniformity across the wafer.
Metrology for High-Aspect-Ratio structures		• Development of **advanced metrology approaches** that are inline, high-speed, and non-destructive to evaluate parameters of interest, e.g., wafer-scale etch and deposition uniformity. • **Multi-modal, multi-scale measurements** to enable seamless integration of measurements and models for holistic characterization.
AI/ML Assisted Metrology and Virtual Metrology		• Integration of **AI/ML** with physics-based modeling for inline high-volume manufacturing metrology. • While there are some instances of **virtual metrology** in high-volume manufacturing, drive wider adoption across industry.
High-Throughput Metrology		• Coupling **high-resolution characterization** with **high-throughput metrology** to improve overall speed, measurement capability, and output of inline metrology techniques. For example, secondary ion mass spectroscopy, which is being adopted as a primary pathway for emerging technologies, can be coupled with an inline technique to reduce offline characterization time.
Failure Analysis		• Establishing advanced, real-time **failure analysis platforms** that integrate multi-modal characterization techniques to swiftly identify and understand failure mechanisms as they occur. • Focusing on the correlation of stressing mechanisms with observed failures through the use of automated, synchronized multi-modal analysis to enable proactive improvements in device design and reliability.
Benchmarking		• Establishment of a sustained **benchmarking program** for emerging devices and systems to objectively evaluate energy efficiency and performance. • Development of an end-to-end **system-level model** to holistically evaluate how energy efficiency improvements at a single level affect system-level efficiency.

Grand challenges

The following represent grand challenges, major resource needs, and key solution pathways distilled from working group discussions:

- **Discrepancy between expected and actual system performance:** During chip manufacturing, the use of simplified metrology structures in non-active areas often fails to

accurately represent the more complex active regions where energy efficiency is critical. This mismatch can lead to unforeseen performance outcomes in devices intended to optimize energy use. Approaches to bridge the gap between expected and actual performance are needed. Solutions may include developing more realistic or representative test structures and providing more samples (without intellectual property [IP]) to metrologists to develop better measurement and modeling capabilities.

- **Buried structures/interfaces (3D metrology)**: The shifts toward 2.5D and 3D devices for energy efficiency introduce layers and interfaces that cannot be easily evaluated. Multi-chamber, multi-process tools are becoming more prevalent to ensure pristine interfaces, as cleanliness is a major contributor to yield losses, resulting in less visibility to underlying structures and interfaces. Traditional methods of evaluating buried structures involve offline and/or destructive characterization, such as focused ion beam (FIB) milling, or slow imaging techniques, such as X-ray tomography. 3D metrology/characterization techniques that are inline or near-inline and non-destructive to minimize defect formation during the production of these devices are needed.
- **Thermal measurements**: The complex structures of emerging energy efficient chip architectures dramatically increase the thermal resistance between various layers and heat sinks, especially for heterogeneously integrated devices. While thermal management is an active area of research, developments in thermal metrology, especially for 3D microelectronics, are lagging. While techniques exist—such as Raman scattering, frequency domain thermo-reflectance (FDTR), and synchrotron photon analysis—improvements in resolution, measurement depth, and the ability to measure heterogeneous materials and interfaces are needed. In addition, as 3D technologies move into high-volume manufacturing, thermal metrology (currently lab-scale) will need to move inline or near-inline.
- **Bringing advanced characterization techniques closer to the "line":** Advanced characterization techniques, such as synchrotron X-ray and scanning tunneling electron microscopy, are typically destructive, slow, and confined to the lab. However, these techniques are invaluable because they provide high-resolution measurement data needed for R&D that conventional metrology techniques cannot provide. As devices and systems get more complex, measurement needs are bordering on those that can only be provided through advanced characterization. Therefore, there is great interest in moving these techniques closer to the "line"—making them non-destructive, faster, and automated.
- The following subsections were distilled from the synthesis of the proposed solution pathways from the working group, each contributes to the advancement of energy efficient devices, and is further discussed in detail:
- **Enhanced metrology:** Improving existing techniques or developing new techniques to meet evolving measurement needs.
- **AI/ML in metrology:** Utilizing artificial intelligence/machine learning algorithms to refine and further enhance metrology, making them more adaptive to ever-evolving measurement needs.

- **Failure analysis (FA):** Developing a holistic FA framework to identify and analyze potential weak links and areas of concern within the device or system, paving the way for improvements in design and functionality.
- **More samples:** Increasing the availability of samples for metrologists and tool developers, while maintaining IP constraints.
- **Benchmarking:** Establishing standards at diverse levels, ranging from narrative frameworks to system-level models. This process ensures a consistent trajectory toward achieving optimal energy efficiency across the board.
-

3.3.1 Enhanced Metrology

Enhanced metrology refers to advanced metrology and characterization techniques that must be developed to meet the measurement needs for emerging energy efficient devices and systems. For instance, the N3XT computing concept (Aly et al. 2015), monolithically integrates carbon nanotube field-effect transistors (CNTFETs), silicon field-effect transistors (FETs), 2D materials for thermal management, and resistive random-access memory (RRAM) along with other components on an energy efficient 3D chip. The complex metrology needs for such an advanced configuration exceed what conventional metrology tools can accommodate, even at the most advanced nodes.

As device become more advanced and smaller, traditional optical microscopy encounters limitations, especially below the 2nm mark. At these scales, more invasive and destructive techniques, such as transmission electron microscopy (TEM) or atom probe microscopy, are often employed. The industry needs to develop non-destructive techniques that offer similar spatial resolution and may benefit from leveraging AI to correlate data from destructive methods.

Summarized below are key techniques and considerations for enhanced metrology to enable the technologies in this roadmap.

Transmission electron microscopy

TEM is an offline characterization technique in which a beam of electrons is transmitted through an ultra-thin sample, interacting with the specimen as it passes through. This interaction produces a magnified image with atomic-scale resolution. It is particularly useful in characterizing device interfaces and structures, crystal structures, and film thicknesses. It is also used as a mechanism to validate measurements of inline tools, such as critical-dimension scanning electron microscopes (CD-SEMs) and critical-dimension small-angle X-ray scattering (CD-SAXS) (Vladar et al. 2014), and, in some instances, develop International System of Units (SI) traceable standards (Orji et al. 2016).

Advancing TEM for specialized measurements demands not only ultra-high resolution but also precise sample preparation coupled with state-of-the-art electron detectors. Presently, TEM is a very slow and destructive process, where the entire TEM process takes approximately 24 hours with engineer oversight. The current pace of R&D related to this roadmap requires faster turnaround with improved resolution.

AI and ML can be leveraged to significantly reduce processing time. By incorporating these technologies, full automation of the TEM process is possible, covering everything from data acquisition to its nuanced interpretation. The introduction of faster detectors and sources with smaller electron dose are anticipated to further shorten processing time. Concurrently, innovations in monochromators and optimizing energy dispersion of the electron beam can improve resolution.

X-ray tomography for 3D stacking and heterogenous integration

X-ray tomography provides a non-destructive method of evaluating hidden interfaces and structures. It is commonly used in failure analysis of packaged devices. It works by generating cross-sectional slices of a 3D structure and reconstructing them to form a 3D image. The push toward advanced packaging and heterogeneous integration (See Chapter 2.3)makes this technique uniquely positioned to support non-destructive evaluation of hidden interfaces and structures. Moreover, with the trend toward multi-step/multi-chamber tools to complete entire process modules, this technique can be further adapted to wafer fabrication to meet metrology needs.

On top of the advancement of X-ray tomography method, high-resolution X-ray tomography for device research is typically confined to large, accelerator-based synchrotron X-ray sources, which makes it not easily accessible to researchers. The development of compact, affordable X-ray sources is needed to reduce the acquisition time and increase accessibility and the number of high-resolution X-ray sources. While X-ray tomography of package- and board-level devices have lower resolution requirements, the availability of compact X-ray sources will still provide significant benefits. Given the complexity of leading-edge heterogeneous integration, integrating board-level electrical testing with 3D, non-destructive mapping techniques becomes essential.

Thermal transport characterization and mapping

Thermal transport characterization and mapping have become increasingly important as 3D heterogeneous devices grapple with heat management challenges. As these devices integrate materials and layers with different functionalities, they often experience uneven thermal gradients, leading to potential performance degradation or even device failure. The development of innovative techniques and metrology tools for thermal characterization can provide insight into these thermal behaviors, facilitating better design and management strategies. Emerging techniques include nitrogen-vacancy (NV) magnetometers and Raman spectroscopy. NV magnetometers, which leverage the quantum properties of defect centers in diamonds, offer high-resolution thermal mapping by sensing temperature-dependent shifts in their luminescence (Kuwahata et al. 2020). Alternately, Raman spectroscopy, which measures the vibrational modes of molecules, can be employed to determine temperature changes based on shifts in peak frequencies (Yue, Zhang, and Wang 2011). Both these techniques not only provide a non-invasive approach to thermal characterization but also allow for real-time monitoring. Techniques such as these can enable more resilient designs or integration schemes by providing a deeper understanding of thermal flows in 3D structures.

Enhanced metrology is essential given the rising complexity of energy efficient devices. The increasing prevalence of heterogeneously integrated devices demand automated measurement that can offer superior characterization and faster rates. Modeling is key to driving these developments forward, ensuring precision and repeatability in measurement as well as analysis. Underpinning these innovations is the availability of samples. Representative test samples and

structures are needed to develop and validate enhanced metrology techniques (further discussion is provided in a separate section below).

Action plan for enhanced metrology technologies

Table 79. Action Plan for Enhanced Metrology Technologies.

Scope			
Metrology and Benchmarking Approach	Enhanced metrology		
Technologies of Interest:	Technologies with complex 3D structures, integration schemes, inaccessible features, and non-standard materials. Technologies requiring in-situ, dynamic, automated measurements with high spatial resolution and multiple measurement modalities.		
Metrology Challenges Addressed		**Proposed Solution Pathways**	
• Inability to evaluate properties at inaccessible points within 3D structures. • Inaccuracy of material and interface properties for non-standard items. • Lack of real-time or near-real-time understanding of interface changes. • Avoid excessive measurements, understanding process dependency to capture only what's necessary.		• Implement non-destructive, deep-penetration metrology with compositional contrast. • Enhance high-spatial-resolution measurements, which may be destructive. • Increase usage of high-speed X-ray tomography at low intensities. • Integrate coupled measurement modalities to provide a holistic view.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline (years)**
Transmission electron microscopy	Critical dimension (CD) and composition measurements	Near-term: 2 hour; mid-term: 1 hour; long-term: 1 hour (fully automated)	3–10
Develop test samples and standards	Identify suitable samples and structures	Identify samples and structures that don't reveal IP but help advance novel metrology	2
Tomography and X-ray scattering	CD measurements and departures	Near-term: correlated magnitudes with baseline CDs; mid-term: 20% error with respect to baseline offline	1–3
Development of compact X-ray sources	Brightness/energy	100x reduction in acquisition time for SAXS and tomography compared to current synchrotron benchmarks	5–15
Integrated board-level electrical testing with 3D mapping	Spatial resolution, throughput	Demonstration with standard electrical testing tools	3
Thermal transport characterization and mapping	Improved spatial resolution	Timeline when needed: soon	Within 1
TEM or atom probe microscopy (destructive). (Optical microscopy is limited below 2nm.)	Non-destructive method with spatial resolution	Far end (if doable at all)	Long-term goal, if achievable
Stakeholders and Potential Roles in Project			
Stakeholder	**Role**		
Product Manufacturers/Suppliers	• Supply critical samples and offer comprehensive details on key product specifications.		
End Users/OEMs	• Spearhead the design, production, and optimization of hardware components and systems.		
Academia	• Innovate and advance research methodologies, harness AI capabilities, and foster collaborative integrations with industry stakeholders.		
National Laboratories	• Undertake specialized research and development projects, often overlapping with academia, to push the boundaries of current knowledge and technology.		
Government	• Facilitate research and industry growth by allocating strategic funding, providing programmatic oversight, and fostering public–private partnerships.		
Required Resources		**Cross Collaboration Needs of Working Groups**	

• High-performance computing (HPC) access for training large-scale models • More beamlines and synchrotrons • Time, funding, and experts	• All working groups are needed to develop requirements related to each technology to ensure that the solutions are practically deployable and the models accurately characterize those parameters.

3.3.2 Artificial Intelligence/Machine Learning in Metrology

Artificial intelligence (AI) and machine learning (ML), with their predictive capabilities and data-driven approach, offer potentially transformative opportunities to metrology. Traditional metrology primarily relies on explicit rule-based systems and manual data analysis and intervention, often rendering the processes time-consuming and occasionally susceptible to operator errors. On the other hand, AI/ML uses data to recognize patterns and detect anomalies, and, in some cases, automatically correct processes. The advantage lies in AI/ML's ability to process and analyze large datasets rapidly, enabling real-time feedback and adjustments, aiding the development of energy efficient devices. However, output from AI/ML is only as good as its input. If these models are working off incomplete or low-quality datasets, its results will be unreliable.

In practice, the introduction of AI/ML into fabs has been gradual. Its current uses are primarily confined to enhancing wafer inspection and defect detection. Virtual metrology, which uses AI to predict process variability on wafers that have not been physically measured, is also slowly being rolled out. Summarized below are additional use cases or applications in which AI/ML can enhance metrology, as well as challenges and potential solutions for each.

Inaccessible points in 3D structures

Conventional metrology tools have limited capability to evaluate or measure points within 3D structures where the probes or tools lack access. By combining inputs from physics-based models, part geometry, and destructive testing, AI/ML may be able to predict, with high accuracy and repeatability, measurements at these inaccessible points—somewhat akin to virtual metrology. However, significant testing, validation, and sensitivity analysis is needed prior to deployment.

Management of large datasets for 3D measurements

Handling vast amounts of data becomes cumbersome, especially when dealing with intricate 3D measurements that produce multidimensional datasets. AI/ML can simplify this challenge by implementing dimensionality reduction techniques. These techniques reduce the size of the dataset without significant information loss, making data management and subsequent analysis more efficient and scalable.

Addressing complex failure analysis

With the intricacy of modern microelectronics, predicting potential device failures becomes a more complex task, especially at the nanoscale. AI/ML-based models can mitigate these challenges by predicting the failure rates for these devices. By training an ML model with historical failure data under different conditions and parameters, the model can forecast potential failures, allowing for preemptive corrective measures.

Scientific data annotation

With the sheer volume of data collected, data annotation quickly becomes the bottleneck for this data being used for AI/ML models. Traditional data annotation is manual and requires domain

expertise to ensure systematic, consistent, and correct annotation. Ironically, the very challenge that must be overcome to make AI/ML function effectively can be addressed by AI/ML. Generative AI and self-supervised techniques, like the Segment Anything Method (developed by META), can complete initial segmentation, annotation, and anomaly detection. While it is still a good idea to keep a human in the loop, especially at the outset, it can reduce much of the upfront manual work. These annotated datasets can then be used to train further models or used in downstream tasks, making the process for development of energy efficient device much more efficient.

Improving Nondestructive Evaluation Techniques

Traditional nondestructive evaluation techniques (NDE) techniques might not offer the granularity required for modern microelectronic devices, often missing out on micro- or nano-scale defects that can compromise device performance. By integrating AI/ML with techniques like X-ray computed tomography (CT), a non-invasive imaging method that captures cross-sectional images using X-rays, a more in-depth and comprehensive assessment of internal structures can be achieved. These algorithms, once trained on annotated data, can improve detection, leading to improved device reliability and longevity.

Complexity in Multimodal, Multiscale Data

Metrology often encompasses diverse datasets, ranging from optical and electron microscopy readings to 3D X-ray CT scans. Additionally, multi-modal data, which may include thermal, mechanical, and electrical measurements, need to be integrated. Navigating and making sense of this multi-modal, multi-scale data is challenging and demands sophisticated computational techniques. By leveraging state-of-the-art AI techniques like multi-task learning, knowledge distillation, and transfer learning, one can simultaneously interpret and correlate data from various sources and scales. Designing an AI model that can effectively amalgamate insights from different data types allows for a holistic understanding, thus enhancing the accuracy and efficiency of metrology. This holistic approach ensures that the unique advantages of each data source are utilized, leading to a more comprehensive assessment.

Action Plan for Artificial Intelligence/Machine Learning in Metrology

Table 80. Action Plan for AI/ML in Metrology

Scope		
Metrology and Benchmarking Approach	Utilization of AI/ML in metrology	
Technologies of Interest:	Inline metrology, non-destructive evaluation, data annotation, 3D metrology, etc.	
Metrology Challenges Addressed		**Proposed Solution Pathways**

• Ensure the reliability of AI/ML predictions, given their dependency on data quality. • Overcome the limited capability of conventional metrology tools to evaluate inaccessible 3D structure points. • Manage large datasets generated from 3D measurements and complex failure analysis. • Automate scientific data annotation to address the bottleneck of manual data processing. • Enhance the granularity of NDE techniques for microelectronic devices. • Address the complexity of integrating multi-modal, multi-scale datasets in metrology.	• Validate AI/ML models using comprehensive testing and sensitivity analysis for accurate virtual metrology. • Implement AI/ML dimensionality reduction to efficiently handle and analyze extensive metrology data. • Train AI/ML models on historical device failure data to predict and prevent potential failures. • Employ generative AI and self-supervised learning for efficient initial data segmentation and anomaly detection. • Integrate AI/ML with advanced NDE methods like X-ray CT to detect internal microscale flaws. • Develop AI algorithms capable of interpreting and correlating multi-modal, multi-scale data for a comprehensive metrology assessment.

Major Tasks/Milestones	Metrics	Targets	Timeline (years)
Inaccessible Points in 3D Structures Curate synthetic and real data, test AI/ML models, understand inaccessible 3D points, design and execute experiments to measure such points, and refine AI models.	Evaluate curated data quantity and quality; assess data completeness; measure accuracy, precision, recall, and F1 score; track inaccessible point detection rates; and monitor model accuracy enhancements.	Overcome challenges in measuring inaccessible 3D structure points.	2–3, varies with the metrology technique used
Management of Large Datasets for 3D Measurements Collect and organize vast 3D datasets, implement dimensionality reduction techniques, evaluate synthetic data quality, incorporate reduced synthetic data into AI training, and track dimensionality reduction and synthesis process enhancements.	Assess curated data quality, measure dataset size reduction and retained information, compare model performances with different dataset sizes, evaluate synthetic data usability, and monitor model performance improvements with synthetic data.	Address challenges in analyzing extensive 3D datasets.	1–1.5
Addressing Complex Failure Analysis Curate relevant parameters, preprocess data, develop and validate AI/ML models for failure prediction, refine the models, and deploy for real-time failure monitoring.	Evaluate parameter collection, assess preprocessing readiness, measure model metrics (accuracy, precision, recall, F1 score, AUC-ROC [area under the receiver operating characteristic curve]), compare model performance across datasets, assess model overfitting, track model enhancements, gauge real-world performance, and implement device improvements post-analysis.	Address intricate device failure analysis challenges.	2–3
Scientific Data Annotation Collect and preprocess relevant data, implement segmentation and anomaly detection, refine labels through active learning, create physics-based simulations, generate and assess synthetic annotated data, and integrate this data into subsequent training or downstream tasks.	Assess data quantity and labeling quality, monitor label enhancement, compare simulation quality to real-world data, evaluate AI/ML model performance, and track downstream task enhancements post-integration.	Address limited annotated scientific data issues.	1–2
Improving Non-Destructive Evaluation Techniques Identify NDE challenges, develop AI/ML solutions, test and refine the algorithms, and incorporate them into the NDE process.	Quantify identified challenges and potential gains, review academic contributions, measure AI improvement in data quality and acquisition time, and track algorithms.	Improve NDE techniques.	3–5, varying with the NDE system type

Complexity in Multi-Modal, Multi-Scale Data Recognize various data types or scales; develop AI for integration and interpretation; and train, validate, and test the models.	Measure identified data type diversity, assess AI performance on different datasets, and examine feature correlations and device performance metrics.	Understand and leverage multi-modal, multi-scale data.	2–3

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Supply AI-enabled devices and tools for model development and testing.
End Users/OEMs	• Share insights on AI/ML application challenges, provide datasets, and offer feedback on model outputs.
Academia	• Conduct foundational research in AI/ML, develop novel algorithms, and collaborate on prototype projects.
National Laboratories	• Offer computational resources and expertise, as well as large-scale testing environments for AI/ML models.
Government	• Fund AI/ML research initiatives, develop policies for ethical AI application, and foster collaboration between various stakeholders.

Required Resources	Cross Collaboration Needs of Working Groups
• High-quality, diverse, and labeled datasets for training and validation • Computing infrastructure with high processing capabilities • Secure data storage and management solutions • Platforms and tools for model deployment and monitoring • Domain-specific expertise for specialized AI/ML applications	• Other working groups in the stack (Materials and Devices, Power and Control Electronics, Algorithms and Software, and Advanced Packaging and Heterogeneous Integration) need to provide requirements from technologies within these groups to better develop platform.

3.3.3 Failure Analysis

Failure Analysis (FA) is a set of techniques aimed at identifying and understanding the root causes of failures in electronic devices and components. With the introduction of new materials, architectures, and integration techniques, traditional FA techniques may no longer be sufficient. The complexity of emerging energy efficient devices means that pinpointing defects and degradation mechanisms using conventional FA can be akin to finding a needle in a haystack. Moreover, the pace of the microelectronics industry demands faster validation of devices and components at the prototype stage, adding another layer of urgency to FA, which is typically slow and methodical.

Recognizing these evolving challenges, FA techniques must undergo a transformation. A holistic, multi-modal in-situ failure analysis platform was proposed as a solution pathway. Such an approach facilitates real-time monitoring of device degradation, combining various characterization modes simultaneously. An automated application can be leveraged to synchronize these modes seamlessly, ensuring a comprehensive understanding of failure mechanisms. In essence, the focus must shift from a post-mortem analysis to a real-time observation, aiding in rapid and precise identification of failure modes. Furthermore, an increased emphasis is needed on correlating specific stressing mechanisms directly with observed failures, allowing for more accurate preemptive measures in device design.

Action Plan for Failure Analysis

Table 81. Action Plan for Failure Analysis.

Scope			
Metrology and Benchmarking Approach	Improved failure analysis		
Technologies of Interest:	• Emerging R&D-stage devices and systems; in particular, those found within the roadmap. • Also relevant to existing integrated circuit (IC) products.		
Metrology Challenges Addressed		**Proposed Solution Pathways**	
• Difficulty of failure analysis as devices get more complex. • Faster validation of R&D-stage devices and components. • Bridge the gap between idealized system metrology and actual system performance.		Multi-modal, in-situ failure analysis platform. This platform will include the following components: • In-situ device stressing and degradation monitoring; • Electrical, thermal, optical, and other modes of characterization during stressing; • An application to synchronize different modes; and • Modeling and failure visualization. These components will work in concert to pinpoint sources of failure and help correlate failure with stressing mechanism.	
Major Tasks/Milestones	**Metrics**	**Targets**	**Timeline (years)**
Integration of electrical, thermal, optical, and other modes of characterization on test stand	Electrical, thermal, optical, and other parameters	Parity with results from single modality testing	1.5
Software/application development for automated synchronization	Application error rate	Minimize	1
Development of testing protocol	Electrical, thermal, optical, and other parameters	Parity with results from single modality testing	1
Development of analysis framework and visualization	Time/spatial resolution	Understanding of where/when/why failure happens	2
Testing and sensitivity analysis	Sensitivity and specificity	Optimal detection of failures with minimal false positives	2.5
Stakeholders and Potential Roles in Project			
Stakeholder	**Role**		
Product Manufacturers/Suppliers	• Provide new devices for characterization.		
End Users/OEMs	• Provide specifications/technical details on problems. Provide characterization tools or components of tools needed to develop platform.		
Academia	• Develop new device designs and first-stage device development.		
National Laboratories	• Provide characterization expertise.		
Government	• Provide support and develop roadmap, coordination of efforts.		
Required Resources		**Cross Collaboration Needs of Working Groups**	
• Test samples that are indicative of emerging devices • End-user requirements to better develop platform		• Other working groups in the stack (Materials and Devices, Power and Control Electronics, and Advanced Packaging and Heterogeneous Integration) need to provide requirements from technologies within these groups to better develop platform.	

3.3.4 More Samples

To meet the ever-stringent process requirements for the next node, as well as the growing catalogue of parameters of interest for R&D-stage devices, metrologists are commonly asked to

develop techniques that are cheaper, better, and faster, or that measure something not measured before. However, these developments are highly dependent on the availability and quality of samples. Through extensive working group discussions and previous workshops held by AMO (Office of Energy Efficiency & Renewable Energy 2021), one thing is clear: increased access to diverse and representative samples is needed to substantially enhance and accelerate the development of novel metrology techniques. Comprehensive and varied samples become paramount to understand, test, and improve metrology, and this is especially the case as 3D devices and systems become more prevalent.

Recognizing that complete elimination of IP restrictions is infeasible and unrealistic, creative solutions to navigate this challenge are necessary. Exploring avenues such as establishing less restrictive non-disclosure agreements (NDAs) or other protective mechanisms could serve as a viable means to bridge this gap. Another idea is to establish a central body (e.g., government-led or industry consortia, similar to SEMATECH) that develops and administers common test structures (without IP) that are available to everyone. By establishing a framework that respects industry sensitivities while ensuring researchers have access to samples, a more collaborative, dynamic, and robust research environment can be fostered. This balanced approach will undoubtedly propel the field of metrology forward, ensuring it continues to address the ever-evolving needs of industry towards a more energy efficient future.

Action Plan for More Samples

Table 82. Action Plan for More Samples

Scope			
Metrology and Benchmarking Approach	Greater availability of samples		
Technologies of Interest:	All IC technologies—emerging and existing		
Metrology Challenges Addressed		**Proposed Solution Pathways**	
• Bridge the gap between idealized system metrology and actual system performance. • Explore mechanisms to overcome IP constraints from device manufacturers, leading to challenges with developing new metrology tools and capabilities. • Inability to evaluate properties of interest at inaccessible points within 3D structures. Inability to measure and verify where devices match designs. • Inaccuracy of material and interface properties that are used in computational models of 3D structures (inclusive of challenges associated with inhomogeneous, anisotropic, and nonlinear materials). • Measure material properties at relevant length scales. Current properties based on bulk materials, which are different than micro/nano scale. • Measure chemical and interfacial properties on side walls and all-around structures.		• Encourage manufacturers or a centralized source to offer realistic/indicative test structures and/or samples rather than model systems (no IP included) for early research. • Fund companies to come up with test structures that are well suited to evaluate a specific metrology. • Consider adopting a Defense Advanced Research Projects Agency (DARPA)-like model with a government group ensuring that devices or common test structures be available for everyone, potentially facilitating metrology standards. • Create standard datasets through testing of a wide variety of samples.	
Major Tasks/Milestones	**Requirements**	**Targets**	**Timeline (years)**
Identify material system and technology	Detailed specifications and types of materials to be studied.	Establish a comprehensive database of materials and technologies.	0.5–1

Identify metrology modality of interest (i.e., electro-thermal, transport, interface, X-ray, thermo-mechanical, dimensional, etc.)	Specific modality devices and equipment relevant for the task.	Develop expertise in multiple modalities, ensuring wide-ranging capabilities.	1–1.5
Develop device/structure design	Standardized designs and schematics based on industry standards.	Achieve efficient and optimal design structures for varied applications.	1–2
Fabricate test devices, including important process variations	Detailed process flow charts and specifications for each device type.	Ensure reliable and repeatable fabrication processes across all devices.	2–3
Evaluate necessary dimensional/material properties	Standardized measurement tools and techniques for diverse material properties.	Acquire accurate and comprehensive material property data for modeling.	2–3
Develop models based on dimensions/material properties	High-fidelity computational tools and software.	Develop predictive models that accurately reflect real-world performance.	3–4
Apply metrology of interest to test devices	Metrology equipment calibrated for the specific test devices.	Ensure accurate and reliable measurements across all test scenarios.	4–5
Document measurement results	Structured database systems and error-estimation algorithms.	Ensure data integrity and reliability for future analysis and application.	4–5

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Develop and fabricate test samples with IP sensitivities in mind • Collaborate with metrologists to address specific issues they face
Tool Vendors	• Create a metrology system (commercial tool) from a metrology technique—bridge valley of death
Academia	• Modeling and simulation • Workforce development • Support circuit design activities
National Laboratories	• Provide characterization expertise and capabilities (e.g., beamlines) • Support circuit design activities
Government	• Investment in fab runs to develop test structures
National Institute of Standards and Technology (NIST)	• Provide standard measurement data and standard material properties (i.e., standardized inputs and outputs)

Required Resources	Cross Collaboration Needs of Working Groups
• Consortium of industry, academia, and national laboratories to foster collaboration and channels of communication	• Other working groups engaged in components and system development (Materials and Devices, Circuits and Architectures, Advanced Packaging and Heterogeneous Integration, and Manufacturing Energy Efficiency and Sustainability) need to provide requirements and samples to the Metrology group to meet these needs.

3.3.5 Benchmarking

Benchmarking enables a consistent comparison of technologies through standardized test methodologies and may help prioritize R&D to support the most promising energy-efficient technologies. It can also establish standards against which performance and efficiency can be measured. Currently, there is a gap in benchmarking the latest advanced technologies and approaches emerging today, including those in packaging, circuits, devices, and

software/algorithms. A sustained and comprehensive benchmarking effort is needed, which may include updating Nikonov and Young's work and expanding it to include higher levels of the stack (Nikonov and Young 2013). In addition, a standardized, system-level model is needed to holistically understand each component's contribution to overall system efficiency.

The primary objective of system-level models is to assess the impact and effectiveness of technological innovations. These models distinguish between innovations that offer real advantages and those that fall short when integrated into larger systems. These models are essential in today's complex technology environments, particularly in data center management and edge devices, and may help illuminate what technologies or technology combinations may provide the largest energy efficiency benefits at the system level. For example, they can analyze the impact of changes in device nodes on data center energy efficiency.

Described below are the key considerations when developing system-level models.

System Complexity

Understanding each component's impact on the overall system is critical for optimizing system efficiency. This requires a model that elucidates the interactions within the system, highlighting the ripple effects of changes in one area on the entire system. Recognizing the interdependencies of components, while complex, is crucial for the model's accuracy and integral for transitioning from a focus on isolated components to a comprehensive system perspective.

Model Development and Simulation

After identifying system components and their interdependencies, these must then be integrated into a modeling framework. Different system layers, from devices to data centers, require distinct models and simulation tools. Ensuring seamless data transfer and interoperability between model levels are key challenges to ensure reliable outputs. Collaboration across various teams such as design engineers, simulation experts, metrologists, and other relevant stakeholders is necessary to achieve a holistic view and effective energy efficient solutions.

Continuous Improvement

To meet evolving energy efficient needs and challenges, continuous improvement must be integral to the refinement and maintenance of these models. Implementing a feedback loop, where insights from system-level measurements consistently inform design and manufacturing processes, enables ongoing refinement and enhancement of the model. This approach not only addresses current challenges but also anticipates future ones, ensuring the model remains robust, efficient, and adaptable in a dynamic technological environment.

Action plan for Benchmarking

Table 83. Action Plan for Benchmarking.

Scope		
Metrology and Benchmarking Approach	Development of energy-focused benchmarking	
Technologies of Interest:	All energy-efficient technologies contained within this roadmap, as well as those emerging in academic and industry R&D settings.	
Metrology Challenges Addressed		**Proposed Solution Pathways**

• Collect benchmarking data and develop models to distinguish between feasible and impractical ideas. • Define energy-efficiency metrics for specific workflows, with a focus on ensuring their relevance over the roadmap scope. • Develop a unified benchmarking standard that assesses the energy efficiency of emerging technologies across all levels of the compute stack. • Quantify the impact of component-level efficiency gains on overall system performance due to complex interdependencies.	• Deploy system-level models to quantify energy consumption trade-offs. • Develop a framework to capture energy consumption trade-offs. • Utilize system-level models to address data center efficiency concerns. • Identify subsystem-level models to understand energy consumption dynamics.

Major Tasks/Milestones	Metrics	Targets	Timeline
System-Level Benchmarking Align metrics with technological advancements and industry needs	Establish standard metrics for system-level modeling	Develop a set of industry-accepted metrics	8 months
Tackle System Complexity Ensure all system components are evaluated holistically, reducing discrepancies	Uniform methods and practices	Streamline system measurements	6 months
Model Development and Simulation Use simulation tools to foresee system behavior and interactions	Integration of advanced simulation tools	Efficient predictive outcomes	10 months
Continuous Improvement Implement a feedback loop to ensure systems stay robust, efficient, and future-ready	Feedback mechanism for insights	Iterative system improvements	Ongoing
Standardize Practices Address complications that arise due to non-standardized practices across sectors	Creation of universal standards	Adoption across all sectors	1 year

Stakeholders and Potential Roles in Project	
Stakeholder	**Role**
Product Manufacturers/Suppliers	• Provide hardware that meets software requirements.
End Users/OEMs	• Update infrastructure to meet hardware needs and improve efficiency. Purchase and install hardware. Collaborate with data center operators for specific requirements.
Academia	• Innovate transformational approaches, such as new materials and computing architectures.
National Laboratories	• Lead in technological development and mature academic innovations.
Government	• Provide funding for new technological approaches and set requirements for efficiency. Fund research and set standards for system design.

Required Resources	Cross Collaboration Needs of Working Groups
• Consortium or organized body to enable various stakeholders' continuous communication • Demonstration platform such as data center or test bed • High-fidelity modeling, ultimately moving toward reduced-order modeling	• All working groups need requirements related to infrastructure and thermal management, ensuring that the solutions are practically deployable and that the models accurately characterize those parameters.

3.3.6 Conclusion for Metrology and Benchmarking

In the Metrology and Benchmarking chapter, the focus has been on enhancing measurement techniques to keep pace with the rapid advancement of semiconductor technologies. Precision

in metrology is crucial for validating the energy efficiency and performance of emerging devices, especially those that are 3D or heterogeneous in nature.

Given the increasing complex structures and materials in semiconductor manufacturing, traditional metrology techniques often fall short. As a response, there is a push to develop advanced, non-destructive metrology methods that can provide detailed insights without damaging the structures of cutting-edge devices. In addition, the integration of AI and ML into metrology processes not only improves the precision and adaptability of these measurements but also ensures that the evaluations are deeply aligned with actual device performance.

Furthermore, establishing continuous and adaptable benchmarking standards is imperative for accurately assessing the energy efficiency of new technologies. These standards must be robust enough to guide industry-wide R&D efforts, helping to streamline the validation and deployment of innovative materials and architectures.

The chapter stresses the importance of a cohesive approach that bridges the gap between metrology and actual device performance. By fostering the developments in AI-enhanced metrology and advocating for the broad accessibility of diverse test samples, the roadmap aims to support the semiconductor industry's move towards more sustainable and energy-efficient solutions. This strategic focus on advanced metrology and benchmarking is essential for accelerating the deployment of technologies that meet the demands of modern energy efficiency standards.

3.3.7 Metrology and Benchmarking References

Aly, M.M.S., M. Gao, G. Hills, C.-S. Lee, G. Pitner, M.M. Shulaker, T.F. Wu, et al. 2015. “Energy-Efficient Abundant-Data Computing: The N3XT 1,000x.” *Computer*. Vol. 48 (Issue 12): pg 24–33. https://doi.org/10.1109/MC.2015.376.

Kuwahata, A., T. Kitaizumi, K. Saichi, T. Sato, R. Igarashi, T. Ohshima, Y. Masuyama, et al. 2020. “Magnetometer with Nitrogen-Vacancy Center in a Bulk Diamond for Detecting Magnetic Nanoparticles in Biomedical Applications.” *Scientific Reports*. Vol. 10: pg 2483. https://doi.org/10.1038/s41598-020-59064-6.

Nikonov, D.E., and I.A. Young. 2013. “Overview of Beyond-CMOS Devices and a Uniform Methodology for Their Benchmarking.” *Proceedings of the IEEE.* Vol. 101 (Issue 12): pg 2498–2533. https://doi.org/10.1109/JPROC.2013.2252317.

Office of Energy Efficiency & Renewable Energy. 2021. “Semiconductor R&D for Energy Efficiency Workshop 2: Ultra Precision Control for Ultra Efficient Devices.” U.S. Department of Energy. Accessed December 4, 2023. https://energy.gov/eere/amo/events/semiconductor-rd-energy-efficiency-workshop-2-ultra-precision-control-ultra.

Orji, N.G., R.G. Dixson, D.I. Garcia-Gutierrez, B.D. Bunday, M. Bishop, M.W. Cresswell, R.A. Allen, and J.A. Allgair. 2016. “Transmission Electron Microscope Calibration Methods for Critical Dimension Standards.” *Journal of Micro/Nanolithography, MEMS, and MOEMS*. Vol. 15 (Issue 4). https://doi.org/10.1117/1.JMM.15.4.044002.

Vladar, A., J.S. Villarrubia, B. Ming, R.J. Kline, J. Chawla, S. List, and M.T. Postek. 2014. “10 nm Three-Dimensional CD-SEM Metrology.” Presented at SPIE Advanced Lithography. San Jose, CA. Accessed December 3, 2023. http://dx.doi.org/10.1117/12.2045977.

Yue, Y., J. Zhang, and X. Wang. 2011. “Micro/Nanoscale Spatial Resolution Temperature Probing for the Interfacial Thermal Characterization of Epitaxial Graphene on 4H-SiC.” *Small.* Vol. 7 (Issue 23): pg 3324–3333. https://doi.org/10.1002/smll.201101598.

3.4 Education and Workforce Development (EWD)

To address the significant challenges posed by the increasing energy consumption of microelectronics, the EES2 roadmap targets crucial energy efficiency and reduction objectives. This ambition necessitates a workforce that is both expanding and rapidly evolving and is equipped to research, manufacture, and deploy the innovations recommended. Recent strategic frameworks, particularly the National Microelectronics Strategy released on March 8, 2024, have laid the groundwork for education and workforce development (EWD) in this sector (National Science and Technology Council 2024). The EES2 roadmap draws inspiration from this strategy, proposing initiatives that not only enhance technical skills but also foster a sustainability-conscious mindset among both current and future professionals in the field.

The pressing nature of energy efficiency and climate issues demands immediate action, beyond waiting for the next generation of engineers, scientists, and technicians. A report by *The New York Times* on March 14, 2024 highlights the rapid pace of data center construction worldwide, emphasizing the need for swift educational reform to include all learners, particularly the current workforce responsible for deploying these centers (Plumer and Popovich 2024). Our educational recommendations are designed to facilitate the rapid incorporation of workers from adjacent fields into the microelectronics sector, kickstarting the journey towards doubling energy efficiency as early as 2024.

The EES2 roadmap outlines four critical EWD goals to be pursued alongside technological advancements:

- Raise public awareness on the crucial role energy-efficient semiconductors play in global sustainability.
- Engage students and workforce in EES2-driven microelectronics research.
- Empower a future-ready microelectronics workforce through multidisciplinary education, training, and continuous support for educators and learners.
- Navigate demographic shifts and engage diverse talent.

These goals underscore the need for a skilled workforce that not only excels in technical areas but also prioritizes and understands the critical importance of energy efficiency. Discussions from the April 2023 workforce-focused roadmap meeting at SLAC National Accelerator Laboratory further underscore the necessity of rethinking our approach to motivating and training the workforce responsible for microelectronics manufacturing and deployment (SLAC National Accelerator Laboratory 2023).

Achieving and maintaining U.S. leadership in energy-efficient microelectronics hinges on our ability to develop a workforce proficient in every aspect of the field. The substantial role of the semiconductor industry in the U.S. economy, supporting 1.85 million jobs as of 2020 with direct employment numbers rising sharply by 2023, reflects the industry's growth and the attractive nature of its job market (Semiconductor Industry Association 2021, 2023; U.S. Bureau of Labor Statistics 2024).

Women in Electronics and Computing

Despite representing a slim majority of the population and the dominant segment of those pursuing higher education, women remain significantly underrepresented in the fields of physical sciences, engineering, and specifically in microelectronics and power electronics manufacturing and deployment. The National Strategy underscores the importance of starting early, advocating for initiatives beginning in elementary school to bridge this gap and foster sustainable change. We refer readers to both the National Strategy and our own Section 4 for guidance on initiating this crucial shift.

As such, the urgency of today's environmental challenges compels us to accelerate these efforts. The year 2024 marks a critical point in confronting the threats of climate change and the looming energy crisis. It is imperative that women are empowered to rapidly transition into careers within sustainable electronics. Clear communication about the necessity of this shift, coupled with robust support for career changes, can catalyze immediate action.

Historically, women have demonstrated remarkable adaptability and capacity for rapid career shifts in times of need, as exemplified by the iconic Rosie the Riveter during World War II. This historical precedent illustrates the potential for significant workforce transformation. While challenges vary among different groups, many women could transition more readily if given appropriate economic, cultural, and mentoring support.

In response to the pressing challenges of our time, it is crucial to fully engage the potential of women in the electronics and computing sectors. This commitment not only addresses gender disparities but also cultivates a resilient, innovative workforce capable of driving our society towards a more sustainable future.

Working Group Methodology

Recognizing the rapidly evolving landscape of the semiconductor industry, the working group sought to address the gap between existing educational programs and the industry's imminent needs. Emphasis was placed on developing curriculum frameworks that incorporate advanced technical knowledge with an acute awareness of sustainability and energy efficiency. The group also tackled challenges related to diversity in science, technology, engineering, and mathematics (STEM) fields, early childhood education, and the direct linkage between educational pathways and fulfilling career opportunities in the microelectronics sector. Through their deliberations, the working group aimed to lay the groundwork for educational reform that not only meets the immediate technical demands of the EES2 initiative but also ensures the long-term sustainability of the microelectronics industry through a well-informed, skilled, and diverse workforce.

Key Takeaways

The key challenges and opportunities for education and workforce development in support of the EES2 initiative are summarized in Table 84. With a spotlight on the nuanced workforce challenges tied to the roadmap, the DOE's recommendations target the development of industry professionals with the requisite education and training to advance EES2-related technologies both now and in the future.

Table 84. Education and Workforce Development Key Needs and Opportunities

Area	Key Needs and Opportunities

Goal 1: Public Awareness		• Elevate public engagement by promoting the critical importance of energy-efficient semiconductors for global sustainability through interactive and educational initiatives.
Goal 2: Student and Workforce Engagement		• Foster hands-on student and workforce engagement in microelectronics research through dynamic academia-industry collaborations aligned with EES2-driven projects.
Goal 3: Future-Ready Workforce Empowerment		• Craft multidisciplinary and agile educational programs that prioritize energy efficiency and provide continual educator support to empower a future-ready workforce.
Goal 4: Diversity & Demographics		• Harness demographic diversity to invigorate the microelectronics industry with innovative and inclusive strategies for talent development and engagement.

Grand Challenges

The following represent grand challenges, major resource needs, and key solution pathways distilled from working group discussions:

- Evolving education curricula to pace with microelectronics innovation, emphasizing energy efficiency and sustainability from foundational learning.
- Cultivating a technically proficient workforce that integrates environmental considerations into its work, fostering a culture of sustainability within the industry.
- Expanding the diversity and inclusivity of the STEM workforce to incorporate a broader range of perspectives and innovative solutions for energy-efficient microelectronics.
- Creating educational pathways that meld practical experiences with theoretical knowledge, highlighting the importance of co-design in hardware, software, and architecture for enhancing energy efficiency.
- Encouraging continuous professional development and lifelong learning to align with rapid technological advancements and the evolving landscape of energy efficiency.
- Promoting cross-disciplinary collaboration between academia, industry, and government to ensure educational programs meet the microelectronics industry's real-world demands.

We acknowledge the limitations of this report, recognizing it does not encapsulate the entire spectrum of microelectronics and related ICT education and workforce needs for the U.S. and international EES2 participants. Insights from the 2023 MAPT Roadmap and contributions from EES2 members, including those from SRC, SEMI, and IEEE, provide a more comprehensive review and additional depth on education and workforce programs in the semiconductor

industry. These contributions shape our understanding and guide our strategies for fostering a dynamic and capable workforce for the future.

3.4.1 Raise Public Awareness on the Crucial Role Energy-Efficient Semiconductors Play in Global Sustainability

Raising public awareness about the critical role of semiconductors in driving sustainable energy solutions is essential for fostering a broader understanding and support for energy efficiency within the industry. As the backbone of modern technology, semiconductors have the unique potential to significantly reduce global energy consumption through innovative, eco-friendly applications. Educating the public on the importance of semiconductors in achieving a sustainable future not only highlights the industry's commitment to environmental stewardship but also inspires collective action towards a net-zero emissions goal. By engaging communities through informative and interactive initiatives, we can catalyze a shift towards more sustainable practices across industries and encourage the next generation of innovators to prioritize energy efficiency in their creations.

To bolster public engagement and drive sustainability in the semiconductor sector, the following strategies are tailored to emphasize energy efficiency and sustainable practices:

- Develop and promote museum exhibits and public activities that provide insightful, actionable information on the role of semiconductors in achieving energy sustainability, enhancing public understanding of their critical importance in green technology.
- Forge connections with networks of science centers to disseminate region-specific educational content that emphasizes local contributions to sustainable semiconductor practices and energy efficiency.
- Craft and distribute educational kits focused on sustainable microelectronics and designed for use in events celebrating advancements in energy-efficient technologies.
- Maximize the use of multimedia and social media platforms to spread awareness about the environmental impact of semiconductors and the industry's efforts toward sustainability.
- Engage communities through competitions and challenges that highlight the importance of energy efficiency and sustainable innovation in the semiconductor industry, encouraging a new generation to contribute to eco-friendly advancements.

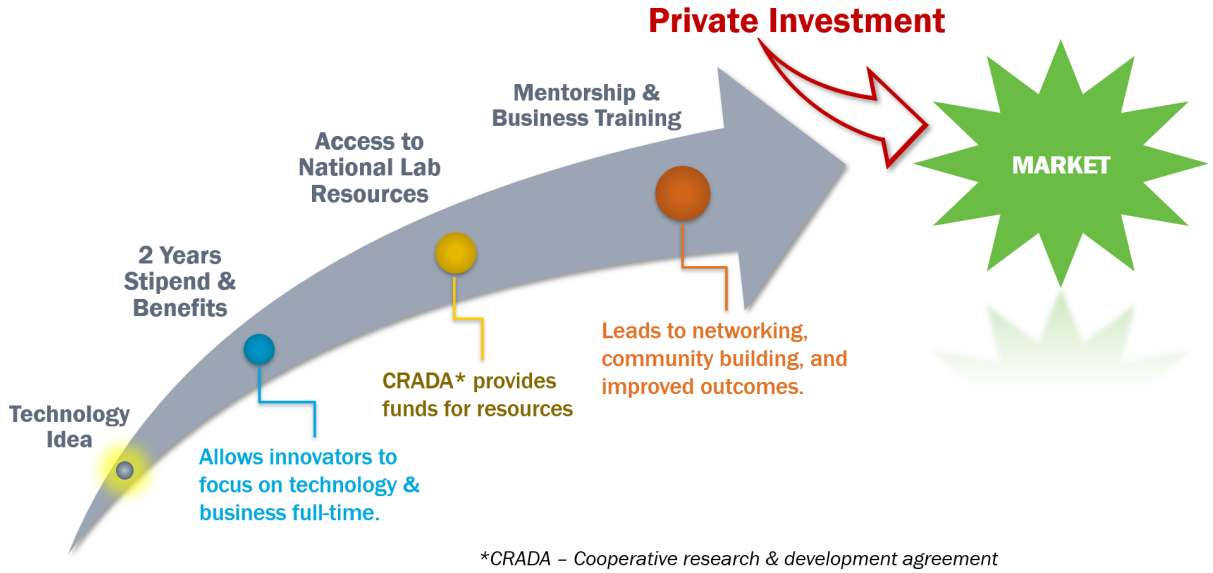
3.4.2 Engage Students and Workforce in Microelectronics Research

To better prepare the next generation for challenges and opportunities listed in this roadmap, leveraging existing educational programs plays a crucial role. These programs, ranging from early childhood development through college, exemplify innovative approaches to integrating STEM principles into various stages of learning. They not only provide valuable resources for educators but also introduce students to the wonders of engineering and technology from a young age. By building on these foundations, we can create a continuum of learning that progressively equips students with the knowledge and skills required for success in the rapidly evolving tech landscape. There are exemplary case studies of existing programs that have made significant strides in STEM education:

- **Project Learning Tree:** This project offers teachers activity guides to enrich classroom learning, with similar initiatives spearheaded by CM partners at CSU and IACMI (https://www.plt.org/).
- **Engineering is Elementary (EIE):** Developed by the Boston Museum of Science, this program integrates STEM content into K–5 reading curricula, providing new material without overburdening the existing curriculum (https://www.eie.org/).
- **Project Lead the Way (PLTW):** PLTW is a curriculum focused on bringing engineering and technology to high schools, introducing students to engineering concepts to prepare them for further education in engineering fields (https://www.pltw.org/).
- **Engineering for Us All (e4usa):** This initiative is aimed at providing a foundational engineering curriculum for high school students, potentially offering college credit upon completion (https://e4usa.org/).
- **TeachEngineering:** A repository of curricular tools and aids for engineering education, this website is a platform where the EES2 initiative can be shared with a broad audience of educators (https://www.teachengineering.org/).
- **Engineering Ambassadors Network:** This network comprises a consortium of around 40 universities training engineering students to deliver compelling, age-appropriate presentations in K–12 schools and after-school programs (https://www.engineeringambassadorsnetwork.org/).

These programs exemplify the diverse approaches to incorporating STEM education across different educational stages in K-12, highlighting the importance of early engagement and continuous learning pathways in building a future-ready workforce.

From the office of AMMTO, the Lab-Embedded Entrepreneurship Program (LEEP) presents a groundbreaking opportunity to engage researchers and the workforce in cutting-edge microelectronics research (EERE 2024). LEEP equips budding entrepreneurs and researchers with the tools, mentorship, and resources to convert their innovative ideas into marketable solutions, focusing on clean energy and technology.



*CRADA – Cooperative research & development agreement

Figure 69. LEEP brings technology idea to market-ready solutions. *Source: EERE 2024*

The LEEP program not only equips innovators with financial support and mentoring but also fosters connections with national lab resources and facilitates industry collaborations. Through initiatives like the annual Demo Day, LEEP encourages the confluence of investors and industry experts, catalyzing the transition from academic research to market-ready solutions. By embedding entrepreneurship in students' educational journey, LEEP ensures that students not only conceive innovative ideas but also possess the tools and knowledge to translate them into viable products that support the nation's energy efficiency and sustainability goals. This engagement is crucial for fostering a competitive edge and positioning the United States as a leader in semiconductor technology and energy-efficient innovation.

As the EES2 roadmap suggests, enriching this educational matrix with programs that nurture a research-driven mindset is paramount. Providing hands-on access to state-of-the-art research facilities and fostering an environment that prioritizes innovation and entrepreneurship ensures that the U.S. continues to pave the way in semiconductor technology and energy efficiency. The concerted effort to bridge education with actionable industry experience will prepare the U.S. workforce to not only face future technological challenges but also to lead the charge in sustainable advancement.

3.4.3 Empower a Future-Ready Microelectronics Workforce Through Multidisciplinary Education, Training, and Continuous Support for Educators and Learners

The burgeoning complexity of semiconductor technologies, such as advancements in neuromorphic and quantum computing, underscores the urgency for interdisciplinary problem-solving skills. This necessitates the development of highly skilled candidates, emphasizing the importance of advanced degrees and highlighting the intense competition for talent, particularly as foreign-born scientists and engineers significantly contribute to this sector (American Immigration Council 2022).

To fulfill the evolving needs of the semiconductor workforce, comprehensive strategies must be developed to empower educators and stimulate students across all educational tiers. Beyond the foundational fields of electrical engineering and computer science, the growing complexity in semiconductor innovations makes interdisciplinary expertise—encompassing chemistry, industrial and environmental engineering, and materials science—increasingly vital for achieving energy efficiency. Enhancing K–12 education through state-aligned, quality resources and hands-on projects is essential for sparking interest in microelectronics careers at an earlier age.

Programs like the NSF's Research Experiences for Teachers and the Robert Noyce Teacher Scholarship Program are instrumental in strengthening STEM education, which is vital for nurturing a future-ready technical workforce. To cultivate talent for skilled technical roles, regionally tailored training programs offering credentials like certificates and diplomas, are often more suitable than traditional degree paths. Such localized training initiatives, especially in burgeoning semiconductor hubs, benefit significantly from partnerships between the industry, educational institutions, and regional training programs.

In higher education, adaptable curricula that keep pace with the swift advancements in semiconductors are necessary to prevent a divergence between academic preparation and industry demands. Proactive collaboration among industry leaders, educators, and labor

representatives is key to developing cutting-edge curricula and programs that cater to imminent industry requirements and encourage interdisciplinary solutions for the multifaceted challenges in semiconductor R&D.

Non-degree programs like Penn State's Microelectronics and Nanomanufacturing Certificate Program exemplify effective industry-academia collaboration. By providing hands-on training and certifications in nanotechnology, these programs forge direct pathways into the semiconductor workforce. Expanding opportunities for mentorship, apprenticeships, and on-the-job training that reflect the current pace of technological innovation in microelectronics is crucial. Creating equitable access to these learning avenues, particularly in underserved regions, and potentially allowing such professional experience to count towards college credit, will streamline career progression in the microelectronics industry.

The following are key action plans to cultivate a future-ready microelectronics workforce:

- Develop comprehensive, interdisciplinary curricula integrating a co-design of hardware and software to address semiconductor complexities and energy efficiency goals.
- Enhance K–12 engagement through quality, state-aligned educational resources and stimulating hands-on projects to spark early interest in microelectronics.
- Bolster STEM teacher training through programs like NSF's Research Experiences for Teachers and the Robert Noyce Teacher Scholarship Program, aiming to build a technically proficient workforce.
- Promote region-specific technical training and credentialing programs, leveraging industry-education partnerships to address local workforce needs in semiconductor hubs.
- Ensure higher education curricula adapt to rapid semiconductor advancements, fostering industry-academia collaboration for innovative, relevant program development.
- Expand non-degree pathways such as certifications and apprenticeships, providing hands-on, industry-aligned training to streamline entry into the semiconductor workforce.
- Create opportunities for mentorship, internships, and professional training that reflect the pace of microelectronics innovation, offering equitable access across diverse communities.
- Allow professional experiences to count towards academic credit, facilitating smoother transitions from education to careers in microelectronics.

3.4.3.1 Curriculum Development

In the subsequent sections, distinct curriculum needs across three critical domains are discussed: Bits, which focus on hardware; Systems, which encompass the co-design of hardware, software, and architecture; and Applications, which deal with algorithms and software. This division is designed to provide a structured approach to curriculum development, enabling targeted educational strategies that cater to the specific skill sets and knowledge areas essential for each aspect of the microelectronics field. By addressing these domains, we lay out a comprehensive educational pathway that supports the EES2 initiative's vision for a sustainable and technologically adept future.

The following coursework and ideas are recommendations derived from this roadmap, designed to address the evolving needs of the microelectronics industry.

For bits (hardware only)

In response to the technological advancements highlighted by the EES2 initiative, the curriculum for hardware engineering and material science must be rigorously updated to equip students with the knowledge and skills necessary to innovate in the field of energy-efficient microelectronics. It should include:

- **Introduction to advanced material science for microelectronics.** Courses should cover the basics of emerging 2D materials, carbon nanotubes (CNTs), ferroelectrics, spintronics, and other novel materials. Focus on their role in enhancing the energy efficiency of interconnects, contacts, and interlayer dielectrics, as well as thermal interface materials.
- **Fundamentals of energy-efficient device physics.** Educate students on the principles of transistor and device-level engineering, including memory and logic devices, analog devices, and the implications of novel transistor structures such as Si-GAA and TFETs for energy savings.
- **Practical applications of novel materials.** Through lab work and projects, provide hands-on experience with fabricating and testing devices made from advanced materials. Emphasize the energy efficiency aspects and performance improvements over traditional silicon-based technologies.
- **Design and simulation of energy-efficient devices.** Integrate courses on CAD and simulation tools specific to devices incorporating novel materials. Teach students to predict device performance, focusing on energy efficiency and power consumption metrics.
- **Capstone projects in energy-efficient hardware design.** Encourage students to undertake comprehensive projects that require them to design, fabricate, and test energy-efficient microelectronic devices, applying their knowledge of advanced materials and device architectures.

By focusing on these critical areas, the “bits” curriculum will prepare students to contribute significantly to the development of next-generation microelectronics, aligning with the EES2 initiative’s goals for a more energy-efficient and sustainable future in computing technology. Collaborative learning experiences, such as team projects and industry internships, will further enhance students’ ability to apply theoretical knowledge to real-world challenges in energy-efficient hardware design.

For Systems (Co-Design of Hardware, Circuits, and Architecture)

The rapid evolution in microelectronics necessitates a curriculum that equips students with advanced knowledge in circuit design and architectural innovations, focusing on energy efficiency and performance optimization.

This includes:

- **Secure and private computing.** Introduce the principles of homomorphic encryption and private information retrieval (PIR) technologies, emphasizing their role in enabling

secure cloud computing and data privacy without compromising on computational efficiency.

- **Computational reliability.** Offer courses on error correction code (ECC) memory technologies, highlighting their importance in enhancing the reliability of data storage and processing in high-stakes environments like data centers and critical servers.
- **Efficient communication protocols.** Educate students on optimizing data movement through efficient communication protocols, including the design and implementation of libraries like NVIDIA's Collective Communication Library (NCCL) for AI workloads.
- **Foundations of neuromorphic computing.** Provide a comprehensive overview of neuromorphic computing, covering the basics of neural computation, including neurons, synapses, dendrites, and cortex operations. Dive into spike encoding mechanisms, spiking and non-spiking brain-inspired networks, and learning rules for spiking neural networks (SNNs).
- **Design and simulation of neuromorphic systems.** Advanced courses on designing digital neural networks and neuromorphic accelerators, including weight quantization, spike design, and learning constraints. Study existing neuromorphic hardware like Intel Loihi, IBM TrueNorth, and others to understand the practical applications and challenges.
- **Quantum computing introduction.** Offer an introductory course on quantum computing, covering the basics of quantum mechanics as applied to computing, qubits, entanglement, and quantum algorithms. Explore the potential energy efficiency benefits and challenges of quantum systems.
- **Advanced architectural design for energy efficiency.** Focus on the design of energy-efficient computing architectures, including the use of computational co-design strategies that integrate hardware and software considerations from the ground up.
- **Practical applications and capstone projects.** Engage students in hands-on projects that involve designing, simulating, and optimizing circuits and architectures for energy efficiency. Encourage projects that incorporate secure computing, neuromorphic systems, and quantum computing concepts.

By addressing these key areas, the "systems" curriculum will prepare students to navigate the complexities of modern circuit design and architecture, with a strong emphasis on energy efficiency and the adoption of next-generation computing paradigms. Through a combination of theoretical knowledge and practical experience, students will be well-equipped to contribute to the advancement of energy-efficient microelectronics, aligning with the goals of the EES2 initiative.

For Applications (Algorithms and Software)

In an era marked by energy-conscious technological innovation, the curriculum for software and applications must evolve to incorporate principles of energy efficiency from the ground up. Students should be trained in the design and implementation of algorithms and software that optimize energy use without sacrificing performance.

- **Foundational programming and code optimizations:** Introduce programming with an emphasis on writing energy-efficient code. Courses must cover runtime optimizations, efficient memory management, and the use of energy-aware compilers. The "Performance Engineering of Software Systems" course currently offered at MIT is an excellent model for addressing this need.
- **Advanced architectures and system integration:** Educate students on the software implications of emerging energy-efficient computational architectures, such as quantum computing and neuromorphic systems.
- **AI and ML for energy efficiency:** Train students to create and optimize AI algorithms that minimize energy consumption, incorporating techniques such as sparse computing and low-power neural networks.
- **Embedded and system-level programming:** Focus on embedded system design with an energy-first approach, including real-time operating systems, microcontroller programming, and IoT applications.
- **Domain-specific software development:** Teach the creation of software for domain-specific architectures, including the use of domain-specific languages that allow high-level problem descriptions to map efficiently to low-power hardware.
- **Application development for energy efficiency:** Offer courses in mobile and web development should emphasize strategies for reducing energy use, from sensor data processing to network communications.
- **AI-enhanced CAD tools:** Include AI methodologies for optimizing chip design in CAD tools, enabling students to contribute to the creation of energy-efficient hardware.

By integrating these components into the software curriculum, students will be prepared to contribute to the EES2 initiative's vision of a sustainable computing future. Collaborations with industry partners for internships and co-op programs can provide practical experience, ensuring that graduates not only understand the theory behind energy-efficient computing but can also apply it in real-world settings.

Cross-Cutting Topics

In addition to the foundational domains of "Bits," "Systems," and "Applications," our curriculum encompasses cross-cutting topics that underscore the importance of co-design in achieving energy-efficient microelectronics. These areas bridge the gaps between hardware engineering, circuit design, architectural innovations, and software development -- ensuring that students grasp the multidisciplinary nature of creating comprehensive solutions for energy-efficient computing.

The recommendations below are derived from this roadmap, aimed at instilling a co-design philosophy in students, preparing them for the collaborative, interdisciplinary challenges of the microelectronics industry as outlined in the EES2 initiative.

This includes:

- **Co-design for energy efficiency.** Introduce students to the principles of co-design, where hardware, software, and system architecture are developed in tandem to optimize

energy efficiency. This includes understanding the trade-offs and synergies between different components of microelectronics systems.

- **Integrated projects in advanced packaging and heterogeneous integration.** Incorporate projects that require students to design and evaluate advanced packaging solutions, such as 2.5D and 3D stacking, focusing on how these technologies impact system performance and energy efficiency. Emphasize the role of heterogeneous integration in enabling high-performance, energy-efficient systems.
- **EDA tools for co-design.** Offer courses that delve into the use of EDA tools in the co-design process, highlighting how these tools facilitate the integrated development of electronic systems, circuits, and components. Special attention should be paid to process design kits (PDKs) and their role in supporting co-design efforts.
- **AI/ML applications in co-design.** Explore how AI and ML algorithms can assist in the co-design process, from optimizing microelectronic device layouts for energy efficiency to predicting the performance of integrated systems. Discuss the use of AI in enhancing metrology tools for better manufacturing precision.
- **Capstone projects on interdisciplinary design.** Engage students in capstone projects that require them to apply knowledge from across the curriculum to design, simulate, and possibly prototype an energy-efficient microelectronic device or system. This could involve integrating advanced packaging techniques, leveraging EDA software for design optimization, and applying AI/ML for performance enhancement.
- **Industry collaborations for practical experience.** Foster partnerships with companies and research institutions involved in advanced packaging, EDA software development, and AI/ML applications in microelectronics. These collaborations can provide students with internships, co-op programs, and access to cutting-edge technologies and methodologies, ensuring their education is directly relevant to industry needs.

By emphasizing co-design in these cross-cutting topics, students will gain a comprehensive understanding of the complexities and interdisciplinary nature of modern microelectronics design and manufacturing. This holistic view is crucial for innovating in the realm of energy-efficient computing and aligns with the ambitious goals of the EES2 initiative.

3.4.4 Navigate Demographic Shifts and Engage Diverse Talent

The talent competition within the STEM sector is increasingly fierce, particularly in hardware engineering and computer software development, highlighting the crucial role of advanced degrees. The past decade has seen Ph.D. hires in the industry double, with foreign-born scientists and engineers constituting 41% of high-skilled technical workers in the semiconductor sector (Hunt and Zwetsloot 2020; National Center for Science and Engineering Statistics 2021). Furthermore, foreign-born individuals represent 30% of all science and engineering workers and hold over half of the doctorates in pivotal fields such as engineering, computer science, and mathematics (Khan, Robbins, and Okrent 2020). However, outdated immigration policies have contributed to a talent drain, diminishing the pool of international talent and simultaneously deterring U.S. students from entering the microelectronics field, thereby jeopardizing the industry's advancements in energy efficiency (Congressional Research Service 2022).

Addressing the talent shortfall necessitates formulating strategies to attract, develop, and retain a diverse talent pool, leveraging both domestic and international expertise. Initiatives such as the Growing Apprenticeships in Nanotechnology and Semiconductors (GAINS) program and the National Talent Hub underscore the critical role of public and private collaboration in aligning workforce development with the goals of the EES2 roadmap, setting a course for a more diverse, innovative, and energy-efficient future in microelectronics.

The semiconductor industry stands at the forefront of addressing global energy challenges, with the potential to significantly impact energy efficiency and sustainability across communities and society. In this vein, it is imperative to prepare an inclusive workforce that is not only well-versed in the current and future landscapes of microelectronics but is also increasingly focused on the benefits of energy efficiency. This preparation extends across educational spectrums, notably within smaller and rural schools, community colleges, Historically Black Colleges and Universities (HBCUs), Tribally Controlled Colleges and Universities (TCCUs), and other minority-serving institutions (MSIs).

By broadening microelectronics education and training across these diverse educational institutions, we unlock opportunities for underrepresented talent in the semiconductor industry. This inclusive approach not only fosters innovation but also cultivates a professional environment that is welcoming and positive, attracting a wider pool of talent. Implementing bridging programs and providing comprehensive support services, such as childcare, further ensures that these opportunities are accessible to all, thus addressing gaps and ensuring a robust pipeline of talent into the industry.

The semiconductor industry is experiencing a pivotal demographic transformation that mirrors the broader shift toward greater diversity in society. This change presents unique challenges and opportunities. Key to this demographic shift is the recognition of the “enthused unfocused” groups who show an interest in semiconductors but perceive the field as daunting or inaccessible. These groups, often inclusive of women and minorities, represent a significant untapped potential (Institution of Mechanical Engineers 2014). To effectively engage them, the industry needs to extend educational outreach efforts that simplify the conceptual presentation of semiconductor technology and make the sector more approachable. Initiatives like mentorship programs, specialized internships, and interactive workshops are essential in providing the necessary insight and encouragement to pursue careers in this field.

3.4.4.1 Underserved Communities

The traditional composition of the STEM workforce—predominantly white, non-Hispanic, and male, particularly at the post-secondary level—is undergoing a transformation. This change is expected to become even more pronounced over the next two decades.

A key insight into this evolving landscape can be gleaned from a comprehensive study by Finegold in 2014 (Institution of Mechanical Engineers 2014). He identifies five distinct groups, or ‘tribes,’ among 11- to 19-year-olds in the United Kingdom, with implications for understanding similar trends in the United States. The study highlights that the majority of youth inclined towards engineering careers belong to the ‘STEM devotees’ group, primarily comprising white males with close ties to adults in STEM professions. However, the study also reveals that other groups are not inherently averse to engineering; they simply need a different approach to

engagement. For instance, the 'social artist' group shows interest in engineering when it intersects with art and design.

A particularly relevant finding for the U.S. context includes a significant number of immigrants and minorities. This group shows an interest in engineering but perceives it as an inaccessible career path. This perception points to a latent potential within these communities that, if properly nurtured, could significantly contribute to the diversity and strength of the STEM workforce, as shown in Figure 70.

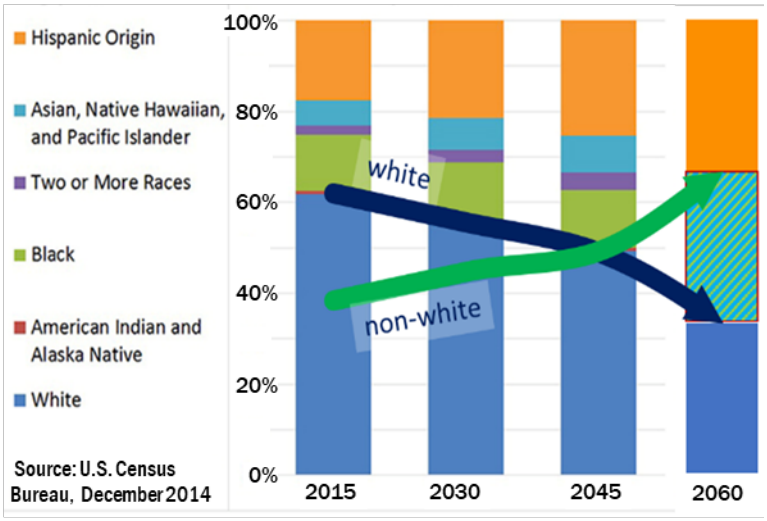


Figure 70. STEM workforce diversity projection. *Source: BRDG program*

Accessible pathways for diverse talents to enter and thrive in the semiconductor industry are needed. This can be achieved through scholarship programs, targeted recruitment initiatives, and collaborations with organizations dedicated to diversity in STEM fields. These pathways should aim to lower the barriers to entry and provide tangible opportunities for these groups to contribute significantly to the semiconductor sector.

3.4.4.2 Women in Science, Technology, Engineering, and Mathematics (STEM)

The semiconductor industry's efforts to engage more women necessitate a nuanced, multifaceted approach, beginning with an understanding of the inequities present from the early educational years. Statistics reveal a significant gender imbalance in STEM, rooted in cultural and educational practices that diverge as early as elementary school. For instance, Lubienski find that girls may begin to doubt their mathematical abilities by the 3rd grade, a stark contrast to boys who may develop an overconfidence in their skills around the same age (Lubienski et al. 2013). This early divergence contributes to a significant underrepresentation of women in the STEM workforce, despite women earning a majority of bachelor's degrees. Specifically, in the 2020–2021 academic year, only 6.7% of women earned degrees in core STEM fields compared to 26.2% for men, highlighting a critical gap at the end of the STEM education pipeline (Statista 2024).

To address this disparity, it is imperative to reshape early childhood messages around STEM, making them inclusive and appealing to all demographics, especially girls and the 'enthused unfocused.' The National Academy of Engineering's "Changing the Conversation" report suggests recasting engineering messaging to resonate with currently disenfranchised demographics. Furthermore, promoting environmental sustainability within STEM disciplines resonates strongly with many women, who often seek careers contributing to societal well-being. Highlighting the role of women in solving environmental challenges through STEM can inspire a new generation to pursue these fields, breaking down stereotypes and broadening the spectrum of opportunities.

Incorporating these insights requires reevaluating legacy educational practices and embracing an educational paradigm that anticipates the rapidly evolving future. As the EES2 initiative tackles pressing sustainability challenges, it also presents an opportunity to pioneer an education and workforce development model that is dynamic and inclusive. Showcasing successful female professionals and emphasizing the industry's commitment to environmental sustainability can inspire and guide aspiring female professionals, fostering a diverse and vibrant workforce ready to address the complex energy challenges of tomorrow.

3.4.4.3 Early-Stage Developments

This demographic evolution shown in Figure 70 coincides with a critical period in STEM education and career decision-making in the United States, where interest often wanes in middle school, particularly among girls and certain minority groups. To counter this trend, it is essential to introduce STEM initiatives at the K–12 level that are specifically designed to engage a diverse student population. Improving STEM instruction, providing experiential learning opportunities, and ensuring access to technology in schools, as well as in afterschool and summer programs, are vital steps in this direction.

For EES2, recognizing and engaging with this rapidly growing but underutilized 'enthused unfocused' group is crucial. Their engagement represents an opportunity to diversify the STEM workforce and challenge existing social paradigms. By creating pathways that make engineering and technology fields more accessible and relatable, EES2 can empower these individuals to become future leaders in technology. This approach is not just about filling workforce gaps; it's about cultivating a rich, diverse pool of talent capable of driving innovation and addressing the complex challenges of our time.

3.4.5 Conclusion for Education and Workforce Development

The Education and Workforce Development chapter has highlighted the critical role of cultivating a technically skilled and diverse workforce to meet the EES2 energy efficiency goals. The future of semiconductor and computing innovation hinges on a workforce capable of understanding, developing, and implementing advanced technologies in a rapidly evolving landscape.

A comprehensive educational framework is needed, ranging from curriculum development for emerging technologies to interdisciplinary training programs that emphasize sustainability. Such programs should align educational outcomes with the specific needs of the semiconductor industry, ensuring that talent pipelines are built to address next-generation challenges.

Moreover, outreach efforts must prioritize diversity and inclusivity to fully leverage the potential of all demographics. This will help secure a workforce that is representative of society and capable of driving innovation forward. Educational pathways should extend beyond traditional academic structures to include targeted training, certification programs, and industry-aligned apprenticeships.

By fostering collaboration across industry, academia, and government, and creating educational programs aligned with industry roadmaps, the EES2 roadmap will help establish a workforce that is ready to tackle the complexities of energy-efficient microelectronics. Ultimately, these initiatives will ensure that the industry remains resilient and innovative in its pursuit of a sustainable and energy-efficient future.

3.4.6 Education and Workforce Development References

American Immigration Council. 2022. “Foreign-born STEM Workers in the United States.” Last modified June 13, 2022. Accessed May 3, 2024. https://americanimmigrationcouncil.org/research/foreign-born-stem-workers-united-states.

Congressional Research Service. 2022. “U.S. Employment-Based Immigration Policy.” Updated July 21, 2022. https://crsreports.congress.gov/product/pdf/R/R47164.

EERE. "Lab-Embedded Entrepreneurship Program," U.S. Department of Energy, Office of Energy Efficiency & Renewable Energy, accessed April 12, 2024. https://www.energy.gov/eere/ammto/lab-embedded-entrepreneurship-program

Hunt, Will, and Remco Zwetsloot. 2020. “The Chipmakers: U.S. Strengths and Priorities for the High-End Semiconductor Workforce.” Center for Security and Emerging Technology (CSET) Issue Brief. Published September 2020. https://cset.georgetown.edu/wp-content/uploads/CSET-The-Chipmakers.pdf.

Institution of Mechanical Engineers. 2014. “Five Tribes: Personalising Engineering Education.” Published October 29, 2014. Accessed February 2024. https://www.imeche.org/policy-and-press/reports/detail/five-tribes-personalising-engineering-education.

Khan, Beethika, Carol Robbins, and Abigail Okrent. 2020. “The State of U.S. Science and Engineering 2020.” U.S. National Science Foundation and National Science Board. Published January 2020. Accessed March 25, 2024. https://ncses.nsf.gov/pubs/nsb20201.

Lubienski, Sarah T., Joseph P. Robinson, Corinna C. Crane, and Colleen M. Ganley. 2013. “Girls’ and Boys’ Mathematics Achievement, Affect, and Experiences: Findings From ECLS-K.” *Journal of Research in Mathematics Education*. Vol. 44 (Issue 4): pg 634–645. https://doi.org/10.5951/jresematheduc.44.4.0634.

National Center for Science and Engineering Statistics. 2021. “National Survey of College Graduates (NSCG) | 2021.” U.S. National Science Foundation. https://ncses.nsf.gov/surveys/national-survey-college-graduates/2021.

National Science and Technology Council. 2024. “National Strategy on Microelectronics Research.” Executive Office of the President of the United States. Published March 2024. https://www.whitehouse.gov/wp-content/uploads/2024/03/National-Strategy-on-Microelectronics-Research-March-2024.pdf.

Plumer, Brad, and Nadja Popovich. 2024. “A New Surge in Power Use Is Threatening U.S. Climate Goals.” *The New York Times*. Published March 14, 2024. https://www.nytimes.com/interactive/2024/03/13/climate/electric-power-climate-change.html.

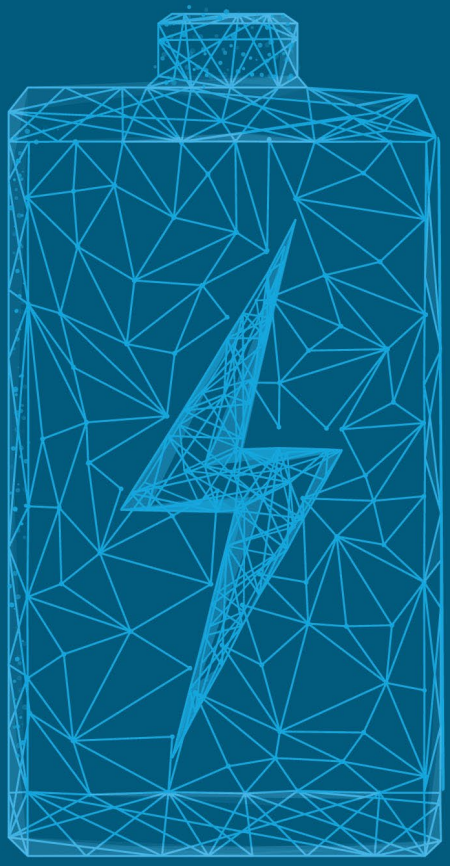
Semiconductor Industry Association. 2021. “Chipping In: The Positive Impact of the Semiconductor Industry on the American Workforce and How Federal Industry Incentives Will Increase Domestic Jobs.” Published May 2021. https://www.semiconductors.org/wp-content/uploads/2021/05/SIA-Impact_May2021-FINAL-May-19-2021_2.pdf.

Semiconductor Industry Association. 2023. “Chipping Away: Assessing and Addressing the Labor Market Gap Facing the U.S. Semiconductor Industry.” Published July 2023. https://www.semiconductors.org/wp-content/uploads/2023/07/SIA_July2023_ChippingAway_website.pdf.

SLAC National Accelerator Laboratory. 2023. “DOE EES2 Roadmap Meeting #5.” Accessed February 2024. https://ees2.slac.stanford.edu/doe-meetings-events/doe-ees2-roadmap-meeting-5.

Statista. 2024. “Number of Bachelor’s Degrees Awarded in the United States During the Academic Year of 2020 to 2021, by Gender and Subject.” Accessed February 2024. https://www.statista.com/statistics/967826/number-bachelors-degrees-awarded-gender-subject-us/.

U.S. Bureau of Labor Statistics. 2024. “Employment and Earnings Table B-3a: Average Hourly and Weekly Earnings of All Employees on Private Nonfarm Payrolls by Industry Sector, Seasonally Adjusted.” Accessed March 25, 2024. https://www.bls.gov/web/empsit/ceseeb3a.htm.



SECTION

4

Conclusion

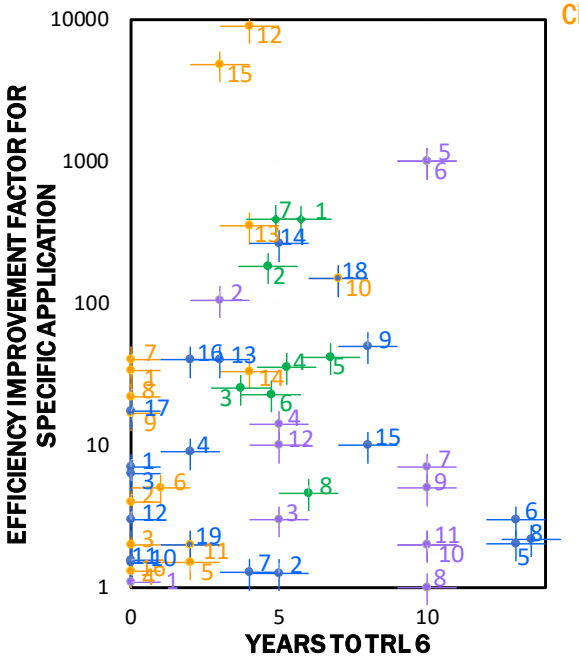
4 Conclusion

The semiconductor industry faces daunting energy challenges. Already confronted with the end of Dennard Scaling and its biennial efficiency improvements, the industry must now contend with explosive data center energy use due to the rise of AI, especially from the growth of natural language programming (NLP)/large language models (LLMs), which are forecasted to drive increases in energy use that could double bi-monthly instead of biennially. In less than two years, LLMs such as ChatGPT have progressed from being a novelty to a commonplace technology that is a standard iPhone feature. The rapid escalation in energy use of just one microelectronics computing application, coming on top of increasing crypto mining electricity use, underscores the urgency of accelerating more energy efficient technologies into the market.

4.1 A New Moonshot and Space Race

Much like JFK's famous Moonshot quote about doing difficult things, the ambitious EES2 goal also serves to organize and measure the best of our energies and skills while similarly providing many public benefits. By setting a straightforward and familiar goal for the industry (biennial efficiency improvements), DOE's EES2 Initiative aims to catalyze an energy efficiency "space race." As version 1.0 and subsequent roadmaps are published, EES2 hopes the industry will compete to better and deploy their own versions of near-term "technologies to beat," as shown in Figure 71. Just setting the goal seems to have already spurred beneficial competition, as evidenced by AMD's announcement of its own goal of 100X efficiency improvement by 2027. EES2 also aims to bolster competition among researchers—especially government funded researchers—to beat these technologies for the mid- and long-term. At the same time, working groups (WGs) in the next versions of the roadmap will race to identify still more ways to co-design energy efficiency into compute stack technology pathways.

EES2 recognizes that the next two decades require a great diversity of technologies and *people* who understand how to design and make them. The further development of curriculum and pedagogy to train and develop the skilled people who can counter the rapidly increasing energy consumption in computing has only just begun.



Circuits and Architectures
1 ReRAM vs NAND
2 STTRAM vs NAND
3 NRAM vs DRAM
4 ReRAM vs DRAM
5 CNT NVM
6 Metis SRAM
7 Molecular dynamics ASIC
8 FPGAs for machine vision
9 SRAM stacked 3D DNN accelerator
10 MIV stacked ReRAM
11 HBM Cache
12 Neuromorphic memcapacitive devices
13 Neuromorphic memristor matrix multiplier
14 Neuromorphic asynchronous computing
15 CMOS SRAM CIM
16 CXL optimized DDR5

Algorithms and Software
1 Reduced energy for ML algorithms
2 Algorithm-specific energy (tooling)
3 Algorithm-specific energy (benchmarking)
4 Languages, compilers, and runtime systems
5 Communication protocols
6 Homomorphic encryption
7 Software for emerging architectures
8 Computational reliability

Materials and Devices
1 Si-GAA
2 CNT Memory
3 CNTFET (Logic)
4 TFET
5 Spintronic memory beyond current STTRAM
6 FeFET (Flash)
7 Analog devices for neuromorphic computing
8 FeFET (SRAM)
9 Contact & interconnect
10 Novel ILD
11 Spintronic logic
12 2D materials

Advanced Packaging & Heterogeneous Integration
1 LMP solder with polymer
2 Nanostructured thermal interface surface
3 CNT TIM
4 Graphene TIM
5 Graphene interconnects
6 CNT interconnects
7 Rh/Ir interconnect
8 CNT for 3D ICs
9 3D IC MIVs
10 Feveros
11 TSV for 3D IC
12 Hybrid bonding (Cu-Cu)
13 Optical off-chip interconnect
14 Optical on-chip interconnect
15 Optical bus
16 UCIe chiplet standard
17 3D stacked SRAM
18 MIV stacked ReRAM
19 HBM on logic

Figure 71. Top energy efficient technologies.

These 55 technologies were identified at the end of 2023 by the roadmap compute stack adjacent layer co-design working groups as those most likely to meet the EES2 1,000x goal. Cost-effectiveness and market projections were considered qualitatively.

4.2 EES2: Putting People and Their Organizations First

This initial release of the EES2 roadmap is the culmination of more than a year of effort by the EES2 pledging organizations and their personnel who participated in the WGs to frame the issues, identify candidate technology options, and formulate the solution pathways and action plans. The EES2 team thoroughly researched the recommended areas and compiled an extensive bibliography. Though not comprehensive, the technologies put forward in this initial roadmap cover the spectrum of the compute stack and enabling technologies with clear potential to achieve EES2 aims. In future versions, as more of the semiconductor innovation ecosystem joins the EES2 Initiative, even more comprehensive roadmaps will be produced.

As of publication, more than 65 organizations have committed to the ambitious pledge (see Figure 72), with many actively participating in the roadmap 1.0 WGs. Pledging and roadmap participation show robust support for the EES2 goals and RD&D agenda from industry leaders, national laboratory leaders, and other educational, workforce, and outreach institutions. This broad base of commitment underscores the potential for strong industry-wide participation in RD&D solicitations aimed at achieving the ambitious objective of enhancing the life-cycle energy efficiency of semiconductor products by at least 10x in 7 years, 100x in 14 years, and 1,000x in two decades. Widespread industry backing further suggests that these efforts will benefit from significant private sector investment, collaboration in education and workforce development, and cost-share participation, all geared towards realizing the transformative energy efficiency targets set forth by EES2.

Figure 72. Pledge signers for EES2 from September 2022– April 2024.

4.3 Technology Results and Co-Design for Efficiency First

The EES2 roadmap highlights the significant factors of energy efficiency improvement possible across the microelectronics compute stack. Figure 71 graphically illustrates the more

promising technology options identified in this roadmap by the compute stack co-design WGs in conjunction with the enabling WGs. The figure presents two key evaluation criteria used by the WGs: time to maturity and efficiency improvement. Time to maturity refers to the time required for a technology to achieve a TRL of 6. Efficiency improvement measures the energy metrics (e.g., energy per bit, energy per switch, memory access) relative to the current state-of-the-art technologies. Additional information is needed to prioritize the potential impact of these technologies and to determine how to allocate resources. EES2 is actively researching some of this information, such as past and current U.S. and global energy use of SOTA technologies, to be included in future versions. Most non-technical factors—apart from the EES2 hypothesis that an ambitious industry-wide biennial goal will drive competition among companies and researchers—are beyond the scope of an R&D roadmap.

Multiple concerted efforts across the full compute stack are necessary over the next two decades. Starting at the bottom of the compute stack with materials and devices, there is an urgent near-term need to consider new materials and device geometries that simultaneously minimize thermal and mechanical forces as well as improve electrical/electronic performance. In the mid-term, continued foundational and manufacturing R&D on materials such as carbon allotropes (graphene and CNTS0) and new switching methods for devices must accelerate. In the long term, device research should include exploration of quantum and nature-inspired approaches and how they can be co-designed across the hardware and full compute stack. In the short term, both in circuits and architecture as well as advanced packaging and heterogeneous integration, industry will take the lead—with strong support from NIST, DOD, NSF, and the CHIPS program—in research on co-design innovations within the circuit/processor and beyond in the hardware stack.

EES2 will further increase competition among researchers in the near and mid-term by supporting the measurement of energy efficiency performance of computing and other microelectronics products. EES2 also will provide benchmarking tools for this efficiency first approach. Finally, at the top of the stack with algorithms and software, the roadmap shows that software-driven full stack co-design is most likely needed to reach 1,000x. Such algorithm and software driven innovations will occur at all time scales, but are especially needed in the long term. Table 85 and Table 86 below expand upon the short-, mid-, and long-term time scales for the EES2 compute stack codesign WGs and enabling WGs, respectively.

Table 85. Key Takeaways for the Compute Stack

Key Areas for Energy Efficiency	Manufacturing Challenges	Solution Pathways
Materials and Devices (mid-term)		

• Novel materials such as 2D materials, CNT, and ferroelectric materials • Current CMOS architecture such as Si-GAA • Future CMOS alternatives with transistors using alternatives for switching such as tunneling, spintronic • Analog devices for neuromorphic	• Streamline production of high-quality novel materials. • Innovate methods for novel materials integration. • Standardize metrics and protocols for emerging technologies. • Assess thermal stability, conductivity, and contact resistance of novel materials. • Connect material science with device engineering. • Leverage detailed device models and simulations.	• Invest in R&D efforts related to scalability of high-purity and -quality novel materials. • Create industry-wide benchmarks and testing protocols to evaluate novel devices and materials. • Fund dedicated testbeds and prototyping labs to demonstrate and refine emerging technologies.
Circuits and Architectures (near term)		
• Compute in and near memory • Domain-specific architectures • New lower-energy non-volatile memory technologies • Neuromorphic computing. • Memory access costs	• Prioritize advanced EDA for improved device architectures. • Develop new algorithmic, power distribution, and additional circuitry changes to bolster new architectures. • Strengthen new device-level technologies to be on par with CMOS. • Increase memory density and reduce cost of new NVM compared to DRAM/NAND. • Eliminate unnecessary overhead power consumption and computational redundancies in architecture systems.	• Improve EDA to enable higher levels of simulation and discover issues before physical device production, increase availability of PDKs for new devices or compute schemes to enable new device/architecture integration. • Design new architectures along with software to enable performance improvements with increased energy efficiency and delve into larger use cases to enable more cost benefits to custom architectures. • Continue funding novel device technologies and concurrent architecture with focus on cost reduction, density increase, and signal variability reduction. • Improve instruction set architectures or instruction level languages and utilize advanced interconnect fabrics such as CXL to enable memory pooling.
Advanced Packaging and Heterogenous Integration (near term)		
• Vertically Integrated devices • Thermal interface materials • Advanced interconnect for Cu replacement • System-level cooling technologies • Interconnect scaling	• Implement STCO for advanced packaging with EDA software. • Remove excessive heat for Energy-efficient 3D technology stacking. • Pair novel technologies with state-of-the-art processors/memory to show proof of durability and energy efficiency. • Address scaling challenges for optical interconnects to enable their use for intra-package and intra-chip signals. • Increase the energy efficiency of memory access.	• Improve EDA to enable ADKs for expanded packaging design and simulation for energy efficiency optimization and FMEA. • Create a fablet allowing for R&D development of advanced packaging and heterogeneous integration technologies, which can alleviate foundry concerns and enable new technology acceleration and proof of concept. • Invest in improved thermal interface materials, heat sinks, and system-level cooling to enable energy-efficient 3D technologies. • Prioritize miniaturization, monolithic integration, and cost reduction of electro-optical light sources, modulators, and detectors. • Enable direct stacking of DRAM or SRAM on processors to help reduce energy costs of the most significant bottleneck of computing.
Algorithms and Software (all time scales, but especially long term)		

• Algorithms that perform tasks more efficiently • Algorithms that avoid data movement • Software that supports new efficient architectures	• Discover and implement machine intelligence algorithms that achieve the abilities of natural systems. • Discover and implement new solutions for scientific computing using machine learning. • Exploit the resources of massively parallel computing systems more effectively.	• Achieve continual, incremental learning in machine learning systems to avoid retraining. • Achieve efficient machine learning through hierarchical models. • Enable fast machine learning design optimization through meta-learning. • Implement fast compiled alternatives for Python. • Improve automatic parallelization of code to exploit available machine resources. • Develop domain-specific languages and frameworks to support emerging architectures.

Table 86. Key Takeaways for Microelectronics Enablers

Key Areas for Energy Efficiency	Grand Challenges	Solution Pathways
Power and Control Electronics (very near term)		
• Migrate computing loads to data centers with available higher-efficiency equipment or onsite renewable energy resources. • Instead of using low-power modes for idle equipment, cut power provisions entirely. • Utilize emerging thermal management strategies to enable higher power densities in stacked die and 3D architectures. • Develop advanced co-design tools for optimizing power delivery along with other key design factors.	• Bridge the gap between metrology and actual device performance. • Challenges with IP constraints and integration of advanced characterization techniques. • Future power delivery approaches will need to be custom fit for circuit architectures. • Co-design tools require improvements to evaluate tradeoffs in the design of complex systems. • Increasing energy density and dimensionality at the chip level necessitate improvements in thermal management. • Computing takes place in non-data center contexts. Scalable solutions are needed to address power management in these locations.	• Develop strategies for resource-aware compute scheduling. • Quantify the impact and challenges associated with idle power reduction strategies. • Pursue RDD&D projects to increase the commercial readiness of emerging cooling technologies. • Extend the functionality of existing software tools to enable co-simulation, reliability investigations, and techno-economic analysis. • Utilize high-performance computing infrastructure to assess the impact of changes in device-level energy use on data center-scale facilities.
Manufacturing Efficiency and Environmental Sustainability (near term)		

• Lower greenhouse gas emitted processes • Abatement systems • EUV efficiency improvements • Alternative lithography development such as NIL	• Multiple processes use high impact greenhouse gas with low removal efficiencies. • Replacement gases that can be used in place of SF_6, NF_3, CF_x are highly caustic. • Viability of replacement processes needs intense scrutiny to test cost effectiveness and yield impacts on devices. • Enable NIL for devices requiring less defect density. • EUV light source is too energy intensive.	• Evaluate novel processes such as thermal ALE and organic vapor plasma etching, which can help reduce greenhouse gas emissions for dry etch but require initial evaluation on 300 mm wafers. • Create alternative processes replacing gases such as SF_6, NF_3, and ClF_3 with F_2, SF_4, or others; this will require better handling because the replacement gases are no longer inert. • Replace abatement systems with improved higher removal efficiencies, which requires only ordering new parts that do not require subfloor space. • Design innovative EUV light source to optimize plasma generation to reduce energy consumption. • Channel R&D efforts toward defect density reduction methods such as stamp material optimization.
Metrology and Benchmarking (all time scales)		
• Advance 3D metrology by developing non-destructive, high-resolution techniques for complex structures and interfaces. • Innovate metrology for precise thermal property measurements of heterogenous materials. • Apply AI/ML to improve precision and efficiency in metrology processes. • Establish continuous and adaptable benchmarking standards for evaluating energy efficiency of new technologies.	• Complexity in metrology due to 3D stacking and heterogeneous integration. • Traditional methods are inadequate for emerging novel devices. • Need for non-destructive techniques and integration with AI/ML. • Bridging the gap between metrology and actual device performance. • Challenges with IP constraints and integration of advanced characterization techniques.	• Develop and adopt advanced, non-destructive metrology methods tailored for complex structures. • Establish comprehensive benchmarking standards for consistent technology evaluation. • Utilize AI/ML algorithms to refine metrology tools for adaptability and precision. • Innovate in metrology to align test structures with actual device performance. • Provide broader access to diverse test samples while respecting IP concerns.
Education and Workforce Development (all time scales but especially long term)		
• Reach people's hearts and minds on the importance of energy efficiency. • Curriculum development for emerging technologies. • Educational pathways for advanced microelectronics. • Interdisciplinary training for sustainability in tech.	• Align educational outcomes with semiconductor industry needs. • Develop talent for next-generation technology roles. • Ensure diversity and inclusivity in STEM fields.	• Create educational programs that align with industry roadmaps. • Implement targeted training for specialized microelectronics roles. • Develop outreach programs to attract a diverse workforce.

4.4 The Future

Version 1.0 of the EES2 roadmap is the beginning of a two-decade effort to take energy efficiency scaling from historical fact to future reality. The exponential demand for computing and the critical need to curb emissions urgently necessitate an acceleration and expansion of these initiatives. While this report documents myriad potential efficiency improvements across fifty-five technologies, achieving their full benefits requires an integrated approach that emphasizes software-driven co-design across the entire technology stack. Ultimately, EES2 hopes to reboot the energy efficiency doubling pace of Dennard scaling doubling efficiency every two years—with the goal of reaching 1,000x more in the next twenty years.

Plans for roadmap 2.0 are already underway. As EES2 recruits more industrial, academic, and national laboratory members of the innovation ecosystem, the initiative will not only have more policy impact, it will also boast even broader technical expertise among the WGs. Now that the first roadmap is published, EES2 will actively turn to broaden its recruiting into new microelectronics application sectors, such as communications. In addition, while EES2 started with electronics and electrons, it will also broaden to promising new information carriers, such the photons used in optoelectronics/ photonics. EES2 already includes pledgers whose research includes long-term transformational technology areas such as quantum computing as well as the latest advances in nature-inspired architectures. EES2 will work with these pledgers to help recruit more from their respective sectors and to attract more volunteers for the version 2.0 WGs.

While much can change before the start of version 2.0 of the roadmap in spring 2025, future WGs will continue to build upon a solid base of peer-reviewed research while continuing to work with EES2 pledgers to lower barriers toward immediate deployment of technologies for biennial microelectronics energy efficiency doubling. This dual R&D and deployment strategy ensures flexibility and responsiveness to emerging technologies and market shifts, thereby fostering a sustainable evolution of the microelectronics sector.

As the EES2 Initiative continues to grow and build momentum for massive improvements in computing energy efficiency, the EES2 team will further work with stakeholders in microelectronics and related applications to develop the technology base and to assess progress toward the goal every 2 years.

This roadmap is not intended to serve as a forecast or to pick winners and losers among technologies. Rather, it is the opening salvo in a new energy efficiency “space race,” where instead of outer space, the EES2 team explores the fascinating realm of increasingly tiny and ultra energy efficient information systems. The roadmap sets a high bar to challenge and motivate technology developers and to counteract grim forecasts that humanity cannot achieve the clean energy transition due to rising computing energy use trends. The semiconductor industry’s inspiring past successes in improving energy efficiency indicate that ambitious EES2 efficiency goals can be met as well. Let’s do it now.

5 Bibliography

This section highlights resources not included in earlier Reference sections of the roadmap.

High-Level Views

Bohr, Mark. 2007. "A 30 Year Retrospective on Dennard's MOSFET Scaling Paper." *IEEE Solid-State Circuits Society Newsletter*. Vol. 12 (Issue 1): pg 11–13. https://doi.org/10.1109/N-SSC.2007.4785534.

Carter, Jonathan, John Feddema, Doug Kothe, Rob Neely, Jason Pruet, Rick Stevens, Prasanna Balaprakash, et al. 2023. "Advanced Research Directions on AI for Science, Energy, and Security." Argonne National Laboratory (ANL) Report ANL-22/91. https://doi.org/10.2172/1986455.

Ding, Li Ping, Ben McLean, Ziwei Xu, Xiao Kong, Daniel Hedman, Lu Qiu, Alister J. Page, and Feng Ding. 2022. "Why Carbon Nanotubes Grow." *Journal of the American Chemical Society.* Vol. 144 (Issue 12): pg 5606−5613. https://doi.org/10.1021/jacs.2c00879.

Geng, Dechao, Xiaoxu Zhao, Ke Zhou, Wei Fu, Zhiping Xu, Stephen J. Pennycook, Lay Kee Ang, and Hui Ying Yang. 2019. "From Self-Assembly Hierarchical h-BN Patterns to Centimeter-Scale Uniform Monolayer h-BN Film." *Advanced Materials Interfaces*. Vol. 6 (Issue 1). https://doi.org/10.1002/admi.201801493.

Kogge, P., K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, et al. 2008. "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems." Defense Advanced Research Projects Agency (DARPA). https://people.eecs.berkeley.edu/~yelick/papers/Exascale_final_report.pdf.

Leiserson, Charles E., Neil C. Thompson, Joel S. Emer, Bradley C. Kuszmaul, Butler W. Lampson, Daniel Sanchez, and Tao B. Schardl. 2020. "There's plenty of room at the Top: What will drive computer performance after Moore's Law." *Science*. Vol. 368 (Issue 6495). https://doi.org/10.1126/science.aam9744.

Net Zero Tracker. 2023. "Net zero targets among world's largest companies double, but credibility gaps undermine progress." Accessed December 22, 2023. https://zerotracker.net/insights/net-zero-targets-among-worlds-largest-companies-double-but-credibility-gaps-undermine-progress.

Patterson, David. 2019. "A New Golden Age for Computer Architecture." Turing Lecture, Association for Computing Machinery. Video presentation, accessed March 1, 2024. https://www.youtube.com/watch?v=aA5pqklkkvI.

Semiconductor Research Corporation (SRC). 2021. *2030 Decadal Plan for Semiconductors: Abridged Report*. https://www.src.org/about/decadal-plan/decadal-plan-abridged-report.pdf.

Shankar, Sadasivan, Victor Zhirnov, and Ralph Cavin. 2009. "Computation from Devices to System Level Thermodynamics." *ECS Transactions*. Vol. 25 (Number 7). https://doi.org/10.1149/1.3203979.

Computing Energy Use Cases

Batmunkh, Altanshagai. 2022. “Carbon Footprint of The Most Popular Social Media Platforms.” *Sustainability.* Vol. 14 (Issue 4): 2195. https://doi.org/10.3390/su14042195

Dance, Gabriel. 2023. “The Real-World Costs of the Digital Race for Bitcoin.” *The New York Times*. Accessed March 1, 2024. https://www.nytimes.com/2023/04/09/business/bitcoin-mining-electricity-pollution.html.

Kanev, S., J.P. Darago, K. Hazelwood, P. Ranganathan, T. Moseley, G.-Y. Wei, and D. Brooks. 2015. “Profiling a warehouse-scale computer.” *ISCA 2015: Proceedings of the 42nd Annual International Symposium on Computer Architecture*. Pg 158–169. https://doi.org/10.1145/2749469.2750392.

Pramanik, P.K.D., N. Sinhababu, B. Mukherjee, S. Padmanaban, A. Maity, B.K. Upadhyaya, J.B. Holm-Nielsen, and P. Choudhury. 2019. “Power Consumption Analysis, Measurement, Management, and Issues: A State-of-the-Art Review of Smartphone Battery and Energy Usage.” *IEEE Access*. Vol. 7: pg 182113–182172. https://doi.org/10.1109/ACCESS.2019.2958684.

Sehgal, Priya, Vasily Tarasov, and Erez Zadok. 2010. “Evaluating Performance and Energy in File System Server Workloads.” *FAST’10: Proceedings of the 8th USENIX Conference on File and Storage Technologies*. https://dl.acm.org/doi/10.5555/1855511.1855530.

Sevilla, J., L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos. 2022. “Compute Trends Across Three Eras of Machine Learning.” Presented at the 2022 International Joint Conference on Neural Networks (IJCNN). Padua, Italy. https://doi.org/10.1109/IJCNN55064.2022.9891914.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. “Energy and Policy Considerations for Deep Learning in NLP.” Presented at the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. Video presentation, accessed March 1, 2024. https://vimeo.com/384787604.

Sudhakar, S., V. Sze, and S. Karaman. 2023. “Data Centers on Wheels: Emissions from Computing Onboard Autonomous Vehicles.” *IEEE Micro*. Vol. 43 (Issue 1): pg 29–39. https://doi.org/10.1109/MM.2022.3219803.

Van Heddeghem, Ward, Sofie Lambert, Bart Lannoo, Didier Colle, Mario Pickavet, and Piet Demeester. 2014. “Trends in worldwide ICT electricity consumption from 2007 to 2012.” *Computer Communications*. Vol. 50 (Issue 1): pg 64–76. http://dx.doi.org/10.1016/j.comcom.2014.02.008.

Materials and Devices

Alexoudi, Theoni, George Theodore Kanellos, and Nikos Pleros. 2020. “Optical RAM and integrated optical memories: a survey.” *Light: Science & Applications.* Vol. 9. https://doi.org/10.1038/s41377-020-0325-9.

Böscke, T.S., J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger. 2011. “Ferroelectricity in Hafnium Oxide Thin Films.” *Applied Physics Letters*. Vol. 99 (Issue 10): 102903. https://doi.org/10.1063/1.3634052.

Cheng, Yi-Lung, and Chih-Yen Lee. 2018. “Porous Low-Dielectric-Constant Material for Semiconductor Microelectronics.” In *Nanofluid Flow in Porous Media,* edited by M.S. Kandelousi, S. Ameen, M.S. Akhtar, and H.-S. Shin. London: IntechOpen. https://doi.org/10.5772/INTECHOPEN.81577.

Cheng, Yi-Lung, Chih-Yen Lee, and Chiao-Wei Haung. 2018. “Plasma Damage on Low-k Dielectric Materials.” In *Plasma Science and Technology – Basic Fundamentals and Modern Applications,* edited by Haikel Jelassi and Djamel Benredjem. London: IntechOpen. https://doi.org/10.5772/intechopen.79494.

Eichfeld, Sarah M., Víctor Oliveros Colon, Yifan Nie, Kyeongjae Cho, and Joshua A. Robinson. 2016. “Controlling nucleation of monolayer WSe_2 during metal-organic chemical vapor deposition growth.” *2D Materials*. Vol. 3 (Issue 2): 025015. http://dx.doi.org/10.1088/2053-1583/3/2/025015.

Empante, T.A., A. Martinez, M. Wurch, Y. Zhu, A.K. Geremew, K. Yamaguchi, et al. 2019. “Low Resistivity and High Breakdown Current Density of 10 nm Diameter van der Waals $TaSe_3$ Nanowires by Chemical Vapor Deposition.” *Nano Letters*. Vol. 19 (Issue 7): pg 4355−4361. https://doi.org/10.1021/acs.nanolett.9b00958.

Favennec, L., V. Jousseaume, V. Rouessac, F. Fusalba, J. Durand, and G. Passemard. 2004. “Porous Extreme Low κ (ELκ) Dielectrics Using a PECVD Porogen Approach.” *Materials Science in Semiconductor Processing*. Vol. 7 (Issues 4–6): pg 277–282. https://doi.org/10.1016/j.mssp.2004.09.084.

Franklin, A.D., M.C. Hersam, and H.-S.P. Wong. 2022. “Carbon nanotube transistors: Making electronics from molecules.” *Science.* Vol. 378 (Issue 6621): pg 726–732. https://doi.org/10.1126/science.abp8278.

Genkina, Dina. 2022. “Computing with Chemicals Makes Faster, Leaner AI.” IEEE Spectrum. Accessed March 1, 2024. https://spectrum.ieee.org/analog-ai-ecram-artificial-synapse.

Geivandov, A.R., S.G. Yudin, and V.M. Fridkin. 2005. “Manifestation of a ferroelectric phase transition in ultrathin films of polyvinylidene fluoride.” *Phys. Solid State*. Vol. 47 (Issue 8): pg 1590–1594. https://doi.org/10.1134/1.2014523.

Gilbert, Simeon J., and Peter A. Dowben. 2020. “Direct measurements of proximity induced spin polarization in 2D systems.” *Journal of Physics D: Applied Physics.* Vol. 53 (Issue 34): 343001. http://dx.doi.org/10.1088/1361-6463/ab8b05.

Gilbert, Matthew J. 2021. “Topological electronics.” *Communication Physics.* Vol. 4 (Article no. 70). https://doi.org/10.1038/s42005-021-00569-5.

Gupta, Anshul, Jürgen Bömmels, Yves Saad, Ivan Cio, and Christopher J. Wilson. 2018. “Integration scheme and 3D RC extractions of three-level supervia at 16nm half-pitch.” *Microelectronic Engineering*. Vol. 191: pg 20–24. https://doi.org/10.1016/j.mee.2018.01.013.

Hanyu, T., T. Endoh, D. Suzuki, H. Koike, Y. Ma, N. Onizawa, M. Natsui, et al. 2016. “Standby-Power-Free Integrated Circuits Using MTJ-Based VLSI Computing.” *Proceedings of the IEEE*. Vol. 104 (Issue 10): pg 1844–1863. https://doi.org/10.1109/JPROC.2016.2574939.

Hennessy, John, and David Patterson. 2018. “A new golden age for computer architecture: Domain-specific hardware/software co-design, enhanced security, open instruction sets, and

agile chip development." Presented at the 2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA). Los Angeles. https://doi.org/10.1109/ISCA.2018.00011.

Ihven, John. 2004. "Precursor Chemistries for the Electronics, Opto-electronics and Optical Industries." *Journal of Materials Chemistry*. Vol. 14 (Issue 21): pg 3071–3080. https://doi.org/10.1039/B405703A.

Intel Corporation. 2017. "Intel® Optane™ Memory Series (16GB, M.2 80mm PCIe 3.0, 20nm, 3D Xpoint™)." Accessed February 2024. https://www.intel.com/content/www/us/en/products/sku/97544/intel-optane-memory-series-16gb-m2-80mm-pcie-3-020nm-3d-xpoint/specifications.html.

Kim, Bum Jun, et al. 2019. "Thickness-Dependence Electrical Characterization of the One-Dimensional van der Waals $TaSe_3$ Crystal." *Materials.* Vol. 12 (Issue 15): 2462. https://doi.org/10.3390/ma12152462.

Kim, Hyung-Woo. 2022. "Recent Trends in Copper Metallization." *Electronics*. Vol. 11 (Issue 18): 2914. https://doi.org/10.3390/electronics11182914.

Kuhn, Kelin. 2018. "CMOS and Beyond CMOS: Scaling Challenges." In *High Mobility Materials for CMOS Applications*, pg 1–44. Woodhead Publishing Series in Electronic and Optical Materials. Elsevier. https://doi.org/10.1016/B978-0-08-102061-6.00001-X.

Kramer, David. 2023. "A computing hardware approach aspires to emulate the brain." *Physics Today*. Vol. 76 (Issue 1): pg 23–26. https://doi.org/10.1063/PT.3.5155.

Lapedus, Mark. 2021. "Breaking The 2nm Barrier." Semiconductor Engineering. Published February 18, 2021. https://semiengineering.com/breaking-the-2nm-barrier/.

Lee, H., A. Lee, F. Ebrahimi, P. Khalili Amiri, and K.L. Wang. 2017. "Analog to Stochastic Bit Stream Converter Utilizing Voltage-Assisted Spin Hall Effect." *IEEE Electron Device Letters*. Vol. 38 (Issue 9): pg 1343–1346. https://doi.org/10.1109/LED.2017.2730844.

Liao, Y.C., C. Pan, and A. Naeemi. 2020. "Benchmarking and Optimization of Spintronic Memory Arrays." *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*. Vol. 6 (Issue 1): pg 9–17. https://doi.org/10.1109/JXCDC.2020.2999270.

Lyu, D., J.E. Shoup, D. Huang, J. García-Barriocanal, Q. Jia, W. Echtenkamp, G.A. Rojas, et al. 2023. "Sputtered $L1_0$-FePd and Its Synthetic Antiferromagnet on Si/SiO_2 Wafers for Scalable Spintronics." *Advanced Functional Materials*. Vol. 33 (Issue 18): 2214201. https://doi.org/10.1002/adfm.202214201.

Moriyama, Naoki, Yutaka Ohno, Kosuke Suzuki, Shigeru Kishimoto, and Takashi Mizutani. 2010. "High-Performance Top-Gate Carbon Nanotube Field-Effect Transistors and Complementary Metal–Oxide–Semiconductor Inverters Realized by Controlling Interface Charges." *Applied Physics Express*. Vol. 3 (Issue 10): 105102. http://dx.doi.org/10.1143/APEX.3.105102.

Moroz, Victor. 2022. "SSDM 2022 Short Course on PPAC of 2DIC & 3DIC." Presented at the 2022 International Conference on Solid State Devices and Materials (SSDM). Virtual event. https://www.researchgate.net/publication/364947453_SSDM_2022_Short_Course_on_PPAC_of_2DIC_3DIC.

Murdoch, G., et al. 2022. “First Demonstration of Two Metal Level Semi-Damascene Interconnects with Fully Self-Aligned Vias at 18MP.” Presented at the 2022 IEEE Symposium on VLSI Technology and Circuits. Honolulu, HI. https://doi.org/10.1109/VLSITechnologyandCir46769.2022.9830150.

O’Brien, K.P., et al. 2021. “Advancing 2D Monolayer CMOS Through Contact, Channel and Interface Engineering.” Presented at the 2021 IEEE International Electron Devices Meeting (IEDM). San Francisco. https://doi.org/10.1109/IEDM19574.2021.9720651.

Onen, Murat, et al. 2022. “Nanosecond protonic programmable resistors for analog deep learning.” *Science.* Vol. 377 (Issue 6605): pg 539–543. https://doi.org/10.1126/science.abp8064.

Pan, Chenyun, and Azad Naeemi. 2017. “Beyond-CMOS Device Benchmarking for Boolean and Non-Boolean Logic Applications.” *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*. arXiv. Submitted November 12, 2017. https://doi.org/10.48550/arXiv.1711.04295.

Peters, Laura. 2022. “Extending Copper Interconnects To 2nm.” Semiconductor Engineering. Published March 17, 2022. https://semiengineering.com/extending-copper-interconnects-to-2nm/.

Pirie, Harris, et al. 2022. “Topological Phononic Logic.” *Physical Review Letters*. Vol. 128: 015501. https://doi.org/10.1103/PhysRevLett.128.015501.

Rehm, L., C.C.M. Capriata, S. Misra, J.D. Smith, M. Pinarbasi, B.G. Malm, and A.D. Kent. 2023. “Stochastic Magnetic Actuated Random Transducer Devices Based on Perpendicular Magnetic Tunnel Junctions.” *Phys. Rev. Appl.* Vol. 19 (Issue 2): 024035. https://doi.org/10.1103/PhysRevApplied.19.024035.

Salahuddin, S., N.K. Nik, and S. Datta. 2018. “The era of hyper-scaling in electronics.” *Nat Electron*. Vol. 1: pg 442–450. https://doi.org/10.1038/s41928-018-0117-x.

Seok, Sang Il, Chang Hoon Ahn, Moon Young Jin, Chang Jin Lee, and Yongku Kang. 2004. “Effect of molecular weight on the mechanical properties of MSSQ films.” *Materials Chemistry and Physics*. Vol. 84 (Issues 2–3): pg 259–262. https://doi.org/10.1016/S0254-0584(03)00326-2.

Shalf, John. 2020. “The future of computing beyond Moore’s Law.” *Philosophical Transactions of the Royal Society A: Mathematics, Physical and Engineering Sciences*. Vol. 378 (Issue 2166). http://doi.org/10.1098/rsta.2019.0061.

Shamiryan, D., T. Abell, F. Iacopi, and K. Maex. 2004. “Low-k Dielectric Materials.” *Materials Today*. Vol. 7 (Issue 1): pg 34–39. https://doi.org/10.1016/S1369-7021(04)00053-7.

Shapiro, M.J., S.V. Nguyen, T. Matsuda, and D. Dobuzinsky. 1995. “CVD of fluorosilicate glass for ULSI applications.” *Thin Solid Films.* Vol. 270 (Issues 1–2): pg 503–507. https://doi.org/10.1016/0040-6090(95)06896-1.

Srivastava, Ashok, Yao Xu, and Ashwani Sharma. 2011. “Carbon nanotubes for next-generation interconnects.” SPIE News. Published January 17, 2011. Accessed March 2024. https://spie.org/news/3220-carbon-nanotubes-for-next-generation-interconnects?SSO=1.

Srivastava, N., and K. Banerjee. 2005. “Performance analysis of carbon nanotube interconnects for VLSI applications.” IEEE/ACM International Conference on Computer-Aided Design, pg 383–390. San Jose, CA. https://doi.org/10.1109/ICCAD.2005.1560098.

Stolyarov, Maxim A., et al. 2016. “Breakdown current density in *h*-BN-capped quasi-1D $TaSe_3$ metallic nanowires: prospects of interconnect applications.” *Nanoscale*. Vol. 8 (Issue 34): pg 15774–15782. https://doi.org/10.1039/c6nr03469a.

Sugibuchi, Kiyoshi, Yukinori Kurogi, and Nobuhiro Endo. 1975. “Ferroelectric field-effect memory device using $Bi_4Ti_3O_{12}$ film.” *Journal of Applied Physics*. Vol. 46: pg 2877–2881. https://doi.org/10.1063/1.322014.

Talin, A. Alec, Yiyang Li, Donald A. Robinson, Elliot J. Fuller, and Suhas Kumar. 2022. “ECRAM Materials, Devices, Circuits and Architectures: A Perspective.” *Advanced Materials*. Vol. 35 (Issue 37): 2204771. https://doi.org/10.1002/adma.202204771.

Toprasertpong, Kasidit, Mitsuru Takenaka, and Shinichi Takagi. 2022. “Memory Window in Ferroelectric Field-Effect Transistors: Analytical Approach.” *IEEE Transactions on Electron Devices*. Vol. 69 (Issue 12): pg 7113–7119. https://doi.org/10.1109/TED.2022.3215667.

Tuchman, Yaakov, Tyler J. Quill, Garrett LeCroy, and Alberto Salleo. 2022. “A Stacked Hybrid Organic/Inorganic Electrochemical Random-Access Memory for Scalable Implementation.” *Advanced Electronic Materials*. Vol. 8 (Issue 8): 2100426. https://doi.org/10.1002/aelm.202100426.

Veloso, A., E. Altamirano-Sánchez, S. Brus, B.T. Chan, M. Cupak, et al. 2016. “Vertical Nanowire FET Integration and Device Aspects.” *ECS Transactions*. Vol. 72 (Issue 4). http://dx.doi.org/10.1149/07204.0031ecst.

Waldrop, M. Mitchell. 2016. “The Chips Are Down for Moore’s Law.” *Nature*. Vol. 530 (Issue 7589). http://dx.doi.org/10.1038/530144a.

Wu, Zhicheng, Ming Zhou, Erfan Khoram, Boyuan Liu, and Zongfu Yu. 2020. “Neuromorphic metasurface.” *Photonics Research*. Vol. 8 (Issue 1): pg 46–50. https://doi.org/10.1364/PRJ.8.000046.

Advanced Packaging and Heterogeneous Integration

Beyene, Wendemagegnehu T. 2022. “Chiplet Technology and Heterogeneous Integration.” IEEE Electronics Packaging Society eNews. Accessed March 2024. https://eps.ieee.org/publications/enews/april-2022/866-chiplet-technology-and-heterogeneous-integration-2.html.

Bhavnagarwala, Azeez J., Blanca L. Austin, Keith A. Bowman, and James D. Meindl. 2000. “A Minimum Total Power Methodology for Projecting Limits on CMOS GSI.” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*. Vol. 8 (Issue 3): pg 235–251. https://doi.org/1063-8210(00)04353-5.

Capozzoli, Alfonso, and Giulio Primiceri. 2015. “Cooling Systems in Data Centers: State of Art and Emerging Technologies.” *Energy Procedia*. Volume 83: pg 484–493. https://doi.org/10.1016/j.egypro.2015.12.168.

Chung, D.D.L. 2022. “Performance of Thermal Interface Materials.” *Small*. Vol. 18 (Issue 16): 2200693. https://doi.org/10.1002/smll.202200693.

Gall, D., J.J. Cha, Z. Chen, et al. 2021. “Materials for interconnects.” *MRS Bulletin*. Vol. 46: pg 959–966. https://doi.org/10.1557/s43577-021-00192-3.

Lau, J.H. 2022. "Recent Advances and Trends in Advanced Packaging." *IEEE Transactions on Components, Packaging and Manufacturing Technology*. Vol. 12 (Issue 2): pg 228–252. https://doi.org/10.1109/TCPMT.2022.3144461.

Peters, Laura. 2023. "The Path to Known Good Interconnects." Semiconductor Engineering. Published January 19, 2023. Accessed March 2024. https://semiengineering.com/the-path-to-known-good-interconnects/.

Razeeb, Kafil M., Eric Dalton, Graham Lawerence, William Cross, and Anthony James Robinson. 2018. "Present and future thermal interface materials for electronic devices." *International Materials Reviews*. Vol. 63 (Issue 1): pg 1–21. https://doi.org/10.1080/09506608.2017.1296605.

Salehi, Soheil, and Ronald F. DeMara. 2015. "Energy and Area Analysis of a Floating-Point Unit in 15nm CMOS Process Technology." Presented at SoutheastCon 2015. Fort Lauderdale, FL. https://doi.org/10.1109/SECON.2015.7132972.

Sebastian, Abu, Manuel Le Gallo, Riduan Khaddam-Aljameh, and Evangelos Eleftheriou. 2020. "Memory devices and applications for in-memory computing." *Nature Nanotechnology*. Vol. 15: pg 529–544. https://doi.org/10.1038/s41565-020-0655-z.

University of Southern California (USC). 2021. "Spiral 2-7: Capacitance, Delay, and Sizing." Presentation at USC Viterbi School of Engineering. https://ee.usc.edu/~redekopp/ee209/slides/EE209Spiral2-7.pdf.

Xing, W., Y. Xu, C. Song, and T. Deng. 2022. "Recent Advances in Thermal Interface Materials for Thermal Management of High-Power Electronics." *Nanomaterials*. Vol. 12 (Issue 19): 3365. https://doi.org/10.3390/nano12193365.

Solid-State Cooling and Thermal Management

Adams, M.J., M. Verosky, M. Zebarjadi, and J.P. Heremans. 2019. "Active Peltier Coolers Based on Correlated and Magnon-Drag Metals." *Physical Review Applied.* Vol. 11: 054008. https://doi.org/10.1103/PhysRevApplied.11.054008.

Aprea, C., A. Greco, A. Maiorino, and C. Masselli. 2017. "Electrocaloric Refrigeration." *Journal of Physics: Conference Series*. Vol. 796: 012019. https://doi.org/10.1088/1742-6596/796/1/012019.

Bailey, Brian. 2022. "Cryogenic CMOS Becomes Cool." Semiconductor Engineering. Published August 11, 2022. https://semiengineering.com/cryogenic-cmos-becomes-cool/.

Heremans, Joseph P. 2016. "Introduction to cryogenic solid state cooling." *Proceedings of SPIE 9821, Tri-Technology Device Refrigeration (TTDR).* https://doi.org/10.1117/12.2228756.

Heremans, Joseph. 2017. "Cryogenic Peltier Cooling." Air Force Research Laboratory Final Report No. AFRL-AFOSR-VA-TR-2017-0084. Published April 6, 2017. https://apps.dtic.mil/sti/pdfs/AD1032532.pdf.

Meng, Yuan, Junhong Pu, and Qibing Pei. 2021. "Electrocaloric cooling over high device temperature span." *Joule*. Vol. 5 (Issue 4): pg 780–793. https://doi.org/10.1016/j.joule.2020.12.018.

Popp, Matthias A., André Erpenbeck, and Heiko B. Weber. 2021. "Thermoelectricity of near-resonant tunnel junctions and their relation to Carnot efficiency." *Nature Research Scientific Reports*. Vol. 11 (Article no. 2031). https://doi.org/10.1038/s41598-021-81466-3.

Shakouri, Ali. 2006. “Nanoscale Thermal Transport and Microrefrigerators in a Chip.” *Proceedings of the IEEE*. Vol. 94 (Issue 8): pg 1613–1638. https://doi.org/10.1109/JPROC.2006.879787.

Sharp, Jeff, Jim Bierschenk, and Hylan B. Lyon, Jr. 2006. “Overview of Solid-State Thermoelectric Refrigerators and Possible Applications to On-Chip Thermal Management.” *Proceedings of the IEEE*. Vol. 94 (Issue 8): pg 1602–1612. https://doi.org/10.1109/JPROC.2006.879795.

Shi, Junye, Donglin Han, Zichao Li, Lu Yang, Sheng-Guo Lu, Zhifeng Zhong, Jiangping Chen, Q.M. Zhang, and Xiaoshi Qian. 2019. “Electrocaloric Cooling Materials and Devices for Zero Global Warming Potential, High-Efficiency Cooling.” *Joule*. Vol. 3: pg 1200–1225. https://doi.org/10.1016/j.joule.2019.03.021.

Spann, Bryan T., et al. 2023. “Semiconductor Thermal and Electrical Properties Decoupled by Localized Phonon Resonances.” *Advanced Materials*. Vol. 35 (Issue 26). https://doi.org/10.1002/adma.202209779.

Tanielian, M.H., R.B. Gregor, J.A. Nielsen, and C.G. Parazzoli. 2011. “Fabrication of nanometer scale gaps for thermo-tunneling devices.” *Applied Physics Letters*. Vol. 99 (Issue 12). https://doi.org/10.1063/1.3641897.

Photonics

Agarwal, Diwakar, Gordon A. Keeler, Christof Debaes, Bianca E. Nelson, Noah C. Helman, and David A. B. Miller. 2003. “Latency Reduction in Optical Interconnects Using Short Optical Pulses.” *IEEE Journal of Selected Topics in Quantum Electronics*. Vol. 9 (Issue 2): pg 410–418. https://doi.org/10.1109/JSTQE.2003.813309.

Cole, Chris. 2021. “Optical and electrical programmable computing energy use comparison.” *Optics Express*. Vol. 29 (Issue 9): pg 13153–13170. https://doi.org/10.1364/OE.420027.

Miller, David A.B. 2000. “Rationale and Challenges for Optical Interconnects to Electronic Chips.” *Proceedings of the IEEE*. Vol. 88 (Issue 6): pg 728–749. https://doi.org/10.1109/5.867687.

Miller, David A.B. 2019. “Waves, modes, communications, and optics.” *Advances in Optics and Photonics*. Vol. 11 (Issue 3): pg 679–825. https://doi.org/10.1364/AOP.11.000679.

Miller, David. 2021. “Optical interconnects to chips – why and how.” Presented at Stanford University. Stanford, CA. https://web.stanford.edu/group/dabmgroup/cgi-bin/dabm/wp-content/uploads/2021/10/IS261.pdf.

Saraswat, Krishna, Hoyeol Cho, Pawan Kapur, and Kyung-Hoae Koo. 2008. “Performance Comparison between Copper, Carbon Nanotube, and Optical Interconnects.” Presented at the 2008 IEEE International Symposium on Circuits and Systems (ISCAS). Seattle, WA. https://doi.org/10.1109/ISCAS.2008.4542034.

Sorger, V.J., R. Amin, J.B. Khurgin, Z. Ma, H. Dalir, and S. Khan. 2018. “Scaling Vectors of Attojoule per Bit Modulators.” *Journal of Optics*. Vol. 20 (Issue 1). https://doi.org/10.1088/2040-8986/aa9e11.

Stojanović, Vladimir, Rajeev J. Ram, Milos Popović, Sen Lin, Sajjad Moazeni, Mark Wade, Chen Sun, et al. 2018. “Monolithic silicon-photonic platforms in state-of-the-art CMOS SOI

processes." *Optics Express*. Vol. 26 (Issue 10): pg 13106–13121. https://doi.org/10.1364/OE.26.013106.

Xu, Jiang. 2016. "Modelling and Analysis of Off-Chip Optical and Electrical Interconnect and Interface." OPTICS Lab. Presentation. https://eexu.home.ece.ust.hk/OPTICS2016/OEIL%20Jiang%20Xu.pdf.

Circuits and Architectures

Anderson, Michael, Benny Chen, Stephen Chen, et al. 2021. "First-Generation Inference Accelerator Deployment at Facebook." arXiv. Submitted July 8, 2021. https://doi.org/10.48550/arXiv.2107.04140.

Arm Developer. 2023. "Arm Custom Instructions." Accessed December 2023. https://developer.arm.com/Architectures/Arm%20Custom%20Instructions.

Arm Developer. 2023. "ARMv8-A Power Management." Accessed December 2023. https://developer.arm.com/documentation/100960/0100/Assembly-language-power-instructions?lang=en.

Athas, W.C., and L.J. Svensson. 1994. "Reversible logic issues in adiabatic CMOS." *Proceedings Workshop on Physics and Computation (PhysComp '94)*. Dallas, TX. https://doi.org/10.1109/PHYCMP.1994.363692.

Bai, S., S. Ranjan, G. Zhou, et al. 2021. "Multistate resistive switching behaviors for neuromorphic computing in memristor." *Materials Today Advances*. Vol. 9. http://dx.doi.org/10.1016/j.mtadv.2020.100125.

French, Robert. 2022. "Catastrophic forgetting in connectionist networks." *Trends in Cognitive Sciences*. Vol. 3: pg 128–135. https://doi.org/10.1016/S1364-6613(99)01294-2.

Gervasi, Bill. 2023. "The Great Convergence; How CXL and UCIe Challenge the Memory Wall." Presented at Flash Memory Summit. Santa Clara, CA.

Hardware Secrets. 2022. "Everything You Need to Know About the CPU Power Management." Published May 15, 2022. Accessed December 2023. https://hardwaresecrets.com/everything-you-need-to-know-about-the-cpu-c-states-power-saving-modes/.

Intel Corporation. 2022. "Technology Guide: Intel® Advanced Vector Extensions 512 - FP16 Instruction Set for Intel® Xeon® Processor Based Products." Accessed March 2024. https://networkbuilders.intel.com/docs/networkbuilders/intel-avx-512-fp16-instruction-set-for-intel-xeon-processor-based-products-technology-guide-1651874188.pdf.

Intel Corporation. 2023. "X86-S External Architectural Specification." White Paper, Rev. 1.0, April 2023, Document Number 351407-001. Accessed March 2024. https://www.intel.com/content/www/us/en/developer/articles/technical/envisioning-future-simplified-architecture.html.

Intel Corporation. 2023. "Nios® V Processor for Intel® FPGA." Accessed December 2023. https://www.intel.com/content/www/us/en/products/details/fpga/nios-processor/v.html.

Jayaraman, Meghana. 2020. "Applications of Neuromorphic Computing." Purdue University. Accessed March 2024. https://docs.lib.purdue.edu/ideas/one/issue3/3/.

Laborieux, Axel, Maxence Ernoult, Tifenn Hirtzlin, and Damien Querlioz. 2021. “Synaptic metaplasticity in binarized neural networks.” *Nat Commun.* Vol. 12: 2549. https://doi.org/10.48550/arXiv.2101.07592.

Patton, Robert, Prasanna Date, Shruti Kulkarni, Chathika Gunaratne, Seung-Hwan Lim, Guojing Cong, Steven Young, et al. 2022. “Neuromorphic Computing for Scientific Applications.” Presented at the 2022 IEEE/ACM Redefining Scalability for Diversely Heterogeneous Architectures Workshop (RSDHA). Dallas, TX. https://doi.org/10.1109/RSDHA56811.2022.00008.

Schmidhuber, Jürgen. 2015. “Deep learning in neural networks: An overview.” *Neural Networks*. Vol. 61: pg 85–117. http://dx.doi.org/10.1016/j.neunet.2014.09.003.

Seshadri, Vivek, Yoongu Kim, Chris Fallin, et al. 2013. “RowClone: Fast and Energy-Efficient In-DRAM Bulk Data Copy and Initialization.” *46th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. Davis, CA. Pg 185–197. https://ieeexplore.ieee.org/document/7847625.

Seshadri, Vivek, Donghyuk Lee, Thomas Mullins, et al. 2017. “Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology.” *50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. Boston. Pg 273–287. https://ieeexplore.ieee.org/document/8686556.

Sze, Vivienne, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. 2017. “Efficient Processing of Deep Neural Networks: A Tutorial and Survey.” *Proceedings of the IEEE*. Vol. 105 (Issue 12): pg 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740.

Teichmann, Ph., J. Fischer, F. Chouard, and D. Schmitt-Landsiedel. 2007. “Design issues of arithmetic structures in adiabatic logic.” *Advances in Radio Science*. Vol. 5: pg 291–295. www.adv-radio-sci.net/5/291/2007/.

van de Burgt, Yoeri, Armantas Melianas, Scott Tom Keene, George Malliaras, and Alberto Salleo. 2018. “Organic electronics for neuromorphic computing.” *Nature Electronics*. Vol. 1: pg 386–397. http://dx.doi.org/10.1038/s41928-018-0103-3.

Vlasov, Danila, et al. 2022. “Reinforcement learning in a spiking neural network with memristive plasticity.” Presented at the 6th Scientific School Dynamics of Complex Networks and their Applications (DCNA). Kaliningrad, Russian Federation. https://doi.org/10.1109/DCNA56428.2022.9923314.

Xia, Z., J. Chen, S. He, and S. Li. 2020. “Neural Synaptic Plasticity-Like Computing: An Ultra-Low Cost Approach for Artificial Neural Networks Implementation.” Presented at the IEEE International Symposium on Circuits and Systems (ISCAS). Seville, Spain. https://doi.org/10.1109/ISCAS45731.2020.9180904.

Yamauchi, H., H. Akamatsu, and T. Fujita. 1995. “An asymptotically zero power charge-recycling bus architecture for battery-operated ultrahigh data rate ULSI’s.” *IEEE Journal of Solid-State Circuits*. Vol. 30 (Issue 4): pg 423–431. https://doi.org/10.1109/4.375962.

Reconfigurable Computing

Boutros, Andrew, Eriko Nurvitadhi, and Vaughn Betz. 2022. “Architecture and Application Co-Design for Beyond-FPGA Reconfigurable Acceleration Devices.” *IEEE Access*. Vol. 10: pg 95067–95082. https://doi.org/10.1109/ACCESS.2022.3204664.

Eckert, Marcel, Dominik Meyer, Jan Haase, and Bernd Klauer. 2016. "Operating System Concepts for Reconfigurable Computing: Review and Survey." *International Journal of Reconfigurable Computing*. Vol. 2016 (Article ID 2478907). http://dx.doi.org/10.1155/2016/2478907.

Gan, Lin, Ming Yuan, Jinzhe Yang, Wenlai Zhao, Wayne Luk, and Guangwen Yang. 2020. "High performance reconfigurable computing for numerical simulation and deep learning." *CCF Transactions on High Performance Computing*. Vol. 2: pg 196–208. https://doi.org/10.1007/s42514-020-00032-x.

Holzinger, P., and M. Reichenbach. 2021. "The HERA Methodology: Reconfigurable Logic in General-Purpose Computing." *IEEE Access*. Vol. 9: pg 147212–147236. https://doi.org/10.1109/ACCESS.2021.3123874.

Putnam, A., A.M. Caulfield, E.S. Chung, D. Chiou, K. Constantinides, J. Demme, H. Esmaeilzadeh, et al. 2016. "A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services." *Communications of the ACM.* Vol. 59 (Issue 11): pg 114–122. https://doi.org/10.1145/2996868.

Shankar, Sadasivan. 2016. "Co-Design 3.0 – Configurable Extreme Computing, Leveraging Moore's Law for Real Applications." SC16 Invited Talk. YouTube video. https://www.youtube.com/watch?v=sijxrGLmzRo.

Tessier, Russell, Kenneth Pocek, and André DeHon. 2015. "Reconfigurable Computing Architectures." *Proceedings of the IEEE*. Vol. 103 (Issue 3): pg 332–354. https://doi.org/10.1109/JPROC.2014.2386883.

Memory Devices, Architectures, and Management

Barla, P., V.K. Joshi, and S. Bhat. 2021. "Spintronic devices: a promising alternative to CMOS devices." *Journal of Computational Electronics*. Vol. 20 (Issue 2): pg 805–837. https://dl.acm.org/doi/abs/10.1007/s10825-020-01648-6.

Bhati, I., M.-T. Chang, Z. Chishti, S.-L. Lu, and B. Jacob. 2015. "DRAM Refresh Mechanisms, Penalties, and Trade-Offs." *IEEE Transactions on Computers*. Vol. 65 (Issue 1): pg 108–121. https://doi.org/10.1109/TC.2015.2417540.

Ghose, S., A.G. Yağlıkçı, R. Gupta, D. Lee, K. Kudrolli, W.X. Liu, H. Hassan, et al. 2018. "What Your DRAM Power Models Are Not Telling You: Lessons from a Detailed Experimental Study." ArXiv. Submitted July 13, 2018. https://doi.org/10.48550/arXiv.1807.05102.

Han, J., R. Cheng, L. Liu, et al. 2023. "Coherent antiferromagnetic spintronics." *Nature Materials*. Vol. 22: pg 684–695. https://doi.org/10.1038/s41563-023-01492-6.

Kennedy, Patrick. 2017. "DDR4 DIMMs and System Power Consumption – We Tested." ServeTheHome. Published January 30, 2017. Accessed March 2024. https://www.servethehome.com/ddr4-dimms-system-power-consumption-tested/.

Li, Shang, Dhiraj Reddy, and Bruce Jacob. 2018. "A Performance & Power Comparison of Modern High-Speed DRAM Architectures." *MEMSYS '18: Proceedings of the International Symposium on Memory Systems*. Alexandria, VA. Pg 341–353. https://doi.org/10.1145/3240302.3240315.

Micron Technology, Inc. 2001. “Calculating DDR Memory System Power.” Micron Technical Note TN-46-03. https://media-www.micron.com/-/media/client/global/documents/products/technical-note/dram/tn4603.pdf.

Micron Technology, Inc. 2019. “Mobile DRAM Power-Saving Features/Calculations.” Micron Technical Note TN-46-12. https://media-www.micron.com/-/media/client/global/documents/products/technical-note/dram/tn4612.pdf.

Raquibuzzaman, Md, Aleksandar Milenkovic, and Biswajit Ray. 2022. “EXPRESS: Exploiting Energy-Accuracy Tradeoffs in 3D NAND Flash Memory for Energy-Efficient Storage.” *Electronics*. Vol. 11 (Issue 3): 424. https://doi.org/10.3390/electronics11030424.

Architectures

Aamodt, Tor M., Wilson Wai Lun Fung, and Timothy G. Rogers. 2018. *General-Purpose Graphics Processor Architectures*. San Rafael, CA: Morgan & Claypool Publishers. https://dl.acm.org/doi/10.5555/3236002.

Hasler, Jennifer, and Bo Marr. 2013. “Finding a roadmap to achieve large neuromorphic hardware systems.” *Frontiers in Neuroscience*. Vol. 7. https://doi.org/10.3389/fnins.2013.00118.

Jouppi, N.P., G. Kurian, S. Li, P. Ma, R. Nagarajan, L. Nai, N. Patil, et al. 2023. “TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings.” Presented at the 50th Annual International Symposium on Computer Architecture. Orlando, FL. arXiv. Submitted April 4, 2023. https://doi.org/10.48550/arXiv.2304.01433.

Open Domain-Specific Architecture (ODSA) Workgroup. 2018. The Open Domain-Specific Architecture: A Chiplet-Based Open Architecture.

Perez, Alberto, Joseph A. Morrone, Carlos Simmerling, and Ken A. Dill. 2016. “Advances in free-energy-based simulations of protein folding and ligand binding.” *Current Opinion in Structural Biology*. Vol. 36: pg 25–31. https://doi.org/10.1016/j.sbi.2015.12.002.

Smagulova, K., M.E. Fouda, F. Kurdahi, K.N. Salama, and A. Eltawil. 2023. “Resistive Neural Hardware Accelerators.” *Proceedings of the IEEE*. Vol. 111 (Issue 5): pg 500–527. https://doi.org/10.1109/JPROC.2023.3268092.

Sun, Yiqiu, Haichao Yang, Wentao Zhang, and Yufeng Gu. 2021. “ASIC Design for Bitcoin Mining.” Association for Computing Machinery. https://zwtaoumich.github.io/paper/EECS570_Final_Report.pdf.

Swirydowicz, Kasia, Eric Darve, Wesley Jones, Jonathan Maack, Shaked Regev, Michael A. Saunders, Stephen J. Thomas, and Slaven Peles. 2021. “Linear solvers for power grid optimization problems: A review of GPU-accelerated linear solvers.” *Parallel Computing*. Vol. 111. https://doi.org/10.1016/j.parco.2021.102870.

Taylor, Michael Bedford. 2017. “The evolution of bitcoin hardware.” *Computer*. Vol. 50 (Issue 9): pg 58–66. https://doi.org/10.1109/MC.2017.3571056.

Algorithms and Software

Anantharaman, Rajesh. 2023. "Google Cloud demonstrates the world's largest distributed training job for large language models across 50000+ TPU v5e chips." Google Cloud blog. Published November 8, 2023. https://cloud.google.com/blog/products/compute/the-worlds-largest-distributed-llm-training-job-on-tpu-v5e.

Anderson, T.E., H.M. Levy, B.N. Bershad, and E.D. Lazowska. 1991. "The Interaction of Architecture and Operating System Design." *ACM SIGOPS Operating Systems Review*. Vol. 25 (Special Issue): pg 108–120. https://doi.org/10.1145/106972.106985.

Arm Developer. 2022. "Armv8.5-A: Memory Tagging Extension." White paper. https://developer.arm.com/documentation/102925/latest/.

Austin, Brian, et al. 2020. "NERSC-10 Workload Analysis (Data from 2018)." Presentation. April 1, 2020. https://portal.nersc.gov/project/m888/nersc10/workload/N10_Workload_Analysis.latest.pdf.

Barroso, Luiz André, Kourosh Gharachorloo, and Edouard Bugnion. 1998. "Memory System Characterization of Commercial Workloads." *ISCA '98: Proceedings of the 25th Annual International Symposium on Computer Architecture*. Pg 3–14. https://doi.org/10.1145/279358.279363.

Barroso, L., M. Marty, D. Patterson, and P. Ranganathan. 2017. "Attack of the Killer Microseconds." *Communications of the ACM*. Vol. 60 (Issue 4): pg 48–54. https://doi.org/10.1145/3015146.

Brown, Tom B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, et al. 2020. "Language Models are Few-Shot Learners." arXiv. Submitted May 28, 2020. https://doi.org/10.48550/arXiv.2005.14165.

Choi, J.W., D. Bedard, R. Fowler, and R. Vuduc. 2013. "A Roofline Model of Energy." Presented at the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing. Cambridge, MA. https://doi.org/10.1109/IPDPS.2013.77.

Clapp, R., M. Dimitrov, K. Kumar, V. Viswanathan, and T. Willhalm. 2015. "Quantifying the Performance Impact of Memory Latency and Bandwidth for Big Data Workloads." Presented at the 2015 IEEE International Symposium on Workload Characterization. Atlanta, GA. https://doi.org/10.1109/IISWC.2015.32.

Cowan, Meghan, Saeed Maleki, Madanlal Musuvathi, Olli Saarikivi, and Yifan Xiong. 2022. "GC3: An Optimizing Compiler for GPU Collective Communication." arXiv. Submitted January 27, 2022. https://doi.org/10.48550/arXiv.2201.11840.

Dai, Wei, and Daniel Berleant. 2019. "Benchmarking Contemporary Deep Learning Hardware and Frameworks: a Survey of Qualitative Metrics." Presented at the 2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI). Los Angeles. https://doi.org/10.1109/CogMI48466.2019.00029.

Dell, Timothy J. 1997. "A White Paper on the Benefits of Chipkill-Correct ECC for PC Server Main Memory."
IBM. https://web.archive.org/web/20150923233043/http:/www.ece.umd.edu/courses/enee759h.S2003/references/ibm_chipkill.pdf.

Jain, Achin, et al. 2023. "A meta-learning approach to predicting performance and data requirements." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition. Vancouver, BC, Canada. https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00353.

Lee, V.W., C. Kim, J. Chhugani, M. Deisher, D. Kim, et al. 2010. "Debunking the 100X GPU vs. CPU Myth: An Evaluation of Throughput Computing on CPU and GPU." *ISCA '10: Proceedings of the 37th Annual International Symposium on Computer Architecture*. Pg. 451–460. https://doi.org/10.1145/1815961.1816021.

Lu, H.J. 2021. "Enable Intel LAM in Linux." Presentation at Linux Plumbers Conference. August 2021. Uploaded to YouTube November 13, 2021. https://www.youtube.com/watch?v=Gh1uuzr1toM.

Marinella, Matthew J., et al. 2022. "Achieving Accurate In-Memory Neural Network Inference with Highly Overlapping Nonvolatile Memory State Distributions." Presented at the 6th IEEE Electron Devices Technology & Manufacturing Conference (EDTM). Oita, Japan. https://doi.org/10.1109/EDTM53872.2022.9797919.

McSherry, Frank, Michael Isard, and Derek G. Murray. 2015. "Scalability! But at what COST?" Presented at the USENIX 15th Workshop on Hot Topics in Operating Systems (HotOS XV). Kartause Ittingen, Switzerland. https://www.usenix.org/conference/hotos15/workshop-program/presentation/mcsherry.

Mustafa, Basil. 2022. "LIMoE: Learning Multiple Modalities with One Sparse Mixture-of-Experts Model." Google Research Brain Team blog post. Published June 9, 2022. https://blog.research.google/2022/06/limoe-learning-multiple-modalities-with.html?m=1.

Nyberg, Chris, Tom Barclay, Zarka Cvetanovic, Jim Gray, and Dave Lomet. 1995. "AlphaSort: A Cache-Sensitive Parallel External Sort." *The VLDB Journal*. Vol. 4: pg 603–627. https://doi.org/10.1007/BF01354877.

Popescu, Diana Andreea. 2019. "Latency-driven performance in data centres." University of Cambridge Computer Laboratory Technical Report No. 937. https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-937.pdf.

Serebryany, K., E. Stepanov, A. Shlyapnikov, et al. 2018. "Memory Tagging and How It Improves C/C++ Memory Safety." arXiv. Submitted February 26, 2018. https://doi.org/10.48550/arXiv.1802.09517.

Shipman, G.M., S. Swaminarayan, G. Grider, J. Lujan, and R.J. Zerr. 2022. "Early Performance Results on 4th Gen Intel® Xeon® Scalable Processors with DDR and Intel® Xeon® processors, codenamed Sapphire Rapids with HBM." arXiv. Submitted November 10, 2022. https://doi.org/10.48550/arXiv.2211.05712.

Stephenson, Mark, Saman Amarasinghe, Martin Martin, and Una-May O'Reilly. 2003. "Meta Optimization: Improving Compiler Heuristics with Machine Learning." *ACM SIGPLAN Notices*. Vol. 38 (Issue 5): pg 77–90. https://doi.org/10.1145/780822.781141.

Vanschoren, J. 2019. "Meta-learning: A Survey." In *Automated Machine Learning: Methods, Systems, Challenges*, 35–61. Cham, Switzerland: Springer, Cham. https://doi.org/10.1007/978-3-030-05318-5_2.

Wang, Yu (Emma), Gu-Yeon Wei, and David Brooks. 2019. "Benchmarking TPU, GPU, and CPU Platforms for Deep Learning." arXiv. Submitted July 24, 2019. https://doi.org/10.48550/arXiv.1907.10701.

Wolfram, Stephen. 2002. *A New Kind of Science*. Wolfram Media, Inc. https://www.wolframscience.com/nks/.

Communication-Avoiding Algorithms

Ballard, G., E. Carson, J. Demmel, M. Hoemmen, N. Knight, and O. Schwartz. 2014. "Communication lower bounds and optimal algorithms for numerical linear algebra." *Acta Numerica*. Vol. 23: pg 1–155. https://doi.org/10.1017/S0962492914000038.

Chen, A., J. Demmel, G. Dinh, M. Haberle, and O. Holtz. 2022. "Communication bounds for convolutional neural networks." *PASC '22: Proceedings of the Platform for Advanced Scientific Computing Conference.* Article no. 1: pg 1–10. https://doi.org/10.1145/3539781.3539784.

Demmel, James. 2022. "Communication-Avoiding Algorithms for Linear Algebra, Machine Learning and Beyond; Integration into Compilers." Oxford Seminars on Tensor Computation. Uploaded to YouTube March 5, 2022. https://www.youtube.com/watch?v=sY3bgirw--4.

Metrology and Benchmarking

Alajlouni, S., K. Maize, and A. Shakouri. 2022. "Full-field pump-probe thermoreflectance imaging for characterization of thin films and 3D integrated circuits." *2022 21st IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm).* San Diego, CA. Pg 1–10. https://doi.org/10.1109/iTherm54085.2022.9899513.

Fisher, S.L., et al. 2019. "Laminography in the lab: imaging planar objects using a conventional x-ray CT scanner." *Meas. Sci. Technol*. Vol. 30 (Issue 3): 035401. https://doi.org/10.1088/1361-6501/aafcae.

Liao, P.-Y., et al. 2020. "Transient Thermal and Electrical Co-Optimization of BEOL Top-Gated ALD In_2O_3 FETs on Various Thermally Conductive Substrates Including Diamond." *2022 International Electron Devices Meeting (IEDM).* San Francisco. Pg 12.4.1–12.4.4. https://doi.org/10.1109/IEDM45625.2022.10019438.

Maize, K., Y. Mi, M. Cakmak, and A. Shakouri. 2023. "Real-Time Metrology for Roll-To-Roll and Advanced Inline Manufacturing: A Review." *Adv. Mater. Technol*. Vol. 8 (Issue 2). https://doi.org/10.1002/admt.202200173.

Nicolai, Lars, Klaus Biermann, and Achim Trampert. 2021. "Application of electron tomography for comprehensive determination of III-V interface properties." *Ultramicroscopy*. Vol. 224: 113261. https://doi.org/10.1016/j.ultramic.2021.113261.

Villarraga-Gómez, H., D. Sirny, M. Terada, M.N. Rad, and A. Gu. 2023. "Workflows for assessing electronic devices with 3D X-ray microscopy and nanoscale computed tomography." *12th Conference on Industrial Computed Tomography (iCT) 2023*. Fürth, Germany. Vol. 28 (Issue 3). https://doi.org/10.58286/27761.

Ziabari, Amirkoushyar, et al. 2018. "Full-field thermal imaging of quasiballistic crosstalk reduction in nanoscale devices." *Nature Communications*. Vol. 9 (Article no. 255). https://doi.org/10.1038/s41467-017-02652-4.

Ziabari, Amirkoushyar, et al. 2020. "Far-field thermal imaging below diffraction limit." *Optics Express*. Vol. 28 (Issue 5): pg 7036–7050. https://doi.org/10.1364/OE.380866.

Manufacturing Energy Efficiency and Sustainability

Chang, Hannah, and Liang-rong Chen. 2020. “Does Taiwan Have Enough Power for TSMC?" CommonWealth Magazine. Vol 703. Published July 28, 2020. https://english.cw.com.tw/article/article.action?id=2766.

Göke, Sebastian, Mena Issler, Demi Liu, Mark Patel, and Peter Spiller. 2022. “Keeping the Semiconductor Industry on the Path to Net Zero.” McKinsey & Company. Published November 4, 2022. https://www.mckinsey.com/industries/semiconductors/our-insights/keeping-the-semiconductor-industry-on-the-path-to-net-zero.

Gupta, U., et al. 2022. “Chasing Carbon: The Elusive Environmental Footprint of Computing.” *IEEE Micro*. Vol. 42 (Issue 4): pg 37–47. https://doi.org/10.1109/MM.2022.3163226.

McKinsey & Company. 2019. “McKinsey on Semiconductors, Issue 7.” Published October 2019. https://www.mckinsey.com/industries/semiconductors/our-insights/mckinsey-on-semiconductors-number-7.

Smeets, C., N. Benders, F. Bornebroek, J. Carbone, R. van Es, A. Minnaert, G. Salmaso, and S. Young. 2023. “0.33 NA EUV Systems for High Volume Manufacturing.” Published in Proceedings of SPIE 12494, Optical and EUV Nanolithography XXXVI, 1249406 (April 28, 2023). https://doi.org/10.1117/12.2658046.

Tasoff, Harrison. 2020. “Moving Bits, Not Watts.” The Current. University of California at Santa Barbara. Published August 25, 2020. https://news.ucsb.edu/2020/019997/moving-bits-not-watts.

Varas, Antonio, Raj Varadarajan, Jimmy Goodrich, and Falan Yinug. 2020. “Government Incentives and US Competitiveness in Semiconductor Manufacturing.” Boston Consulting Group and Semiconductor Industry Association. Published September 2020. https://www.bcg.com/publications/2020/incentives-and-competitiveness-in-semiconductor-manufacturing.

Education and Workforce Development

Committee on STEM Education, National Science & Technology Council. 2018. “Charting a Course for Success: America’s Strategy for STEM Education.” U.S. Department of Energy. https://www.energy.gov/sites/default/files/2019/05/f62/STEM-Education-Strategic-Plan-2018.pdf.

National Institute for Innovation and Technology. N.d. “Registered Apprenticeship Programs: We Identify Roadblocks to Innovation in Strategic Industry Sectors and Ensure They Are Eliminated.” Accessed March 25, 2024. https://www.niit.org.

SEMI. 2024. “SEMI University: Semiconductor Training & Courses.” Accessed March 25, 2024. https://www.semi.org/en/semi-university.

U.S. Department of Labor. 2024. “The Good Jobs Initiative.” Accessed March 25, 2024. https://www.dol.gov/general/good-jobs/principles.

Office of
ENERGY EFFICIENCY &
RENEWABLE ENERGY

ADVANCED MATERIALS &
MANUFACTURING
TECHNOLOGIES OFFICE

For more information, visit: energy.gov/eere/ammto

DOE/EE-XXXX • July 2024