Energy Efficiency Scaling for 2 Decades (EES2) Roadmap for Computing

T. Kaarsberg[1], J. Atulasimha[4], J. Baniecki[2], P. Fischer[5], S. Pawlowski[5], S. Misra[6], A. Bhavnagarwala[7], E. Salman[8], M. Ahmed[9], N. Li[10], R. Aggarwal[11], B. Hirano[12], T. Shah[13], C. Green[10], J. Booth[14], P. Sharps[6], T. McDonald[15], J. Ballard[16], Y. Chen[17], P. Nagapurkar[18], W. Huang[15], D. Kudithipudi[19], A. Paramonov[9], F. Musso[20], A. K. Ziabari[18], J. Luo[22], A. K. Petford-Long[9], D. Gopman[14], C. Gotama[21], T. Wei[21], S. Shaheen[22], Y. Zhang[9], I. Lu[23], K. Shimizu[23], E. Taylor[23], N. Johnson[23], R. Jones[23], and S. Shankar[2,3]

[1]U.S. Department of Energy, Advanced Materials and Manufacturing Technology Office, email:tina.kaarsberg@ee.doe.gov [2]SLAC, Menlo Park, CA, USA [3]Stanford University, Stanford, CA, USA [4]Virginia Commonwealth University, Richmond, VA, USA [5]Intel Corporation, Santa Clara, CA, USA [6]Sandia National Laboratory, Albuquerque, NM, USA [7]Metis Microsystems, Hawleyville, CT, USA [8]Stony Brook University, Stony Brook, NY, USA [9]Argonne National Laboratory, Lemont, IL, USA [10]Carbice, Atlanta, GA, USA [11]Meta, Menlo Park, CA, USA [12]Micron, Boise, ID, USA [13]GE Vernova, Cambridge, MA, USA [14]National Institute for Standards and Technology, Gaithersburg, MD, USA [15]Infineon Technologies, El Segundo, CA, USA [16]Zyvex Labs, Richardson, TX, USA [17]Duke University, Durham, NC, USA [18]Oak Ridge National Laboratory, Oak Ridge, TN, USA [19]University of Texas at San Antonio, San Antonio, TX, USA [20]Dedalo AI, Berkeley, CA, USA [21]BRDG, Orange County, CA, USA [22]University of Colorado Boulder, Boulder, CO, USA [23]Energetics, Columbia, MD, USA

***Abstract*—In response to the looming crisis in global energy consumption required for advanced computing applications, the United States Department of Energy (DOE) Advanced Materials and Manufacturing Technology Office (AMMTO) is leading a multi-organizational effort to define a roadmap for energy efficiency scaling for two decades (EES2) with the aim to reduce energy use in all aspects of computation by more than a factor of 1000 in two decades. By July of 2024, over 60 organizations representing industry, academia, and the national laboratories have pledged to work in various aspects of research and development to enable energy efficiency in computing including in the development of the EES2 roadmap, with an initial public release in 2024 as the first phase of an ongoing commitment to energy-efficient and sustainable computation.**

I. Introduction

Semiconductors are instrumental in driving innovation and productivity throughout the global economy, with significant implications in virtually every economic sector and are critical for U.S. global competitiveness, economic stability, national security, and climate resilience. Advances in microelectronics, encompassing scientific computing, machine learning, automation, and more, are pivotal in driving technological innovations across the economy [1]. The semiconductor industry, essential for continuous technological progress, faces challenges in energy efficiency and carbon footprint. According to the Semiconductor Research Corporation (SRC), semiconductor energy use has increased unsustainably since 2010. By 2030, semiconductors could consume nearly 25% of human planetary energy production [2]. This increase is driven by the end of Dennard scaling, increased digitalization, AI advancements, and the proliferation of smart devices, exacerbating the energy and carbon impact of semiconductor production. At a critical juncture, the industry could benefit immensely from targeted R&D investments in new technologies and innovative manufacturing processes, steering towards sustainable energy usage in the areas of microelectronics and computing.

II. Scope of the Problem

The semiconductor industry faces critical challenges in energy efficiency and sustainability due to increasing demands from ubiquitous computing and manufacturing complexities. This has led to significant energy consumption, highlighting the need for innovative solutions beyond geometrical scaling to reduce energy use and improve efficiency.

A. The Memory Wall and Communication Problem

Historically, processor speed increased rapidly pre-2005, while memory bandwidth lagged, leading to the "memory wall" issue [3]. The coordination of data movements across multiple cores and maintaining memory coherence has become increasingly complex, limiting overall system performance. This divergence in scaling relationships has driven innovative approaches in microelectronic device design and operation but has also exacerbated energy efficiency problems. Addressing the memory wall requires reducing the energy cost of data movement and improving memory access speed.

B. Major Sources of Computing Energy Use

Advanced machine learning and AI technologies, cryptocurrency mining, and cloud computing have significantly increased energy demands. Specialized hardware for these technologies, while efficient for specific tasks, contributes to the overall energy footprint due to high-power requirements. Cryptocurrency mining, in particular, involves specialized ASICs that consume vast amounts of electricity [4][8]. The hardware supporting cloud computing infrastructure, including data centers with high-density chips and advanced 3D heterogeneous integration, also contributes to escalating energy consumption [5]. The proliferation of IoT devices and the rollout of 5G networks further add to energy use. Although each IoT device consumes little energy individually, the aggregate energy required for billions of devices globally adds up to significant values. 5G infrastructure, while more energy-efficient on a per-bit basis,

increases overall energy consumption due to higher data rates and network density [6].

C. Energy Inefficiency at multiple levels

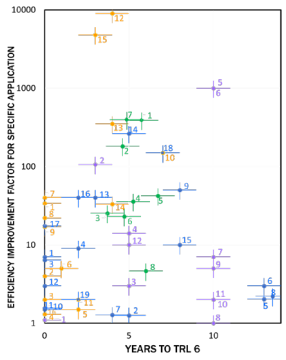
Wide-ranging studies have highlighted significant inefficiencies in computing energy use. Shankar's research [7][8] compared energy intensity per instruction for top supercomputers against fundamental thermodynamical and biological limits. The findings revealed massive energy use disparities, with memory access being particularly energy intensive. For instance, Horowitz's measurements [9] and subsequent updates [10] showed that, despite improvements in processor energy efficiency with smaller geometries, the energy cost of external DRAM access remains unchanged. On-chip instruction energy costs are significantly lower than off-chip DRAM access, underscoring the need to reduce data movement energy, by exploring innovative memory technologies to address the high energy costs of data movement.

III. EES2 Cooperation Pledge

The US DOE's AMMTO is leading a collaborative effort involving other Federal agencies, industry leaders, universities, national laboratories, and international partners to develop an R&D roadmap for energy efficiency scaling over the next two decades. The EES2 goal—achieving a 1000X improvement in microelectronics energy efficiency—is designed to reverse the trend of increasing semiconductor energy use through innovations in research and development. As of July 2024, more than 60 organizations have signed the pledge [11].

IV. EES2 Working Groups

The EES2 initiative has formed eight working groups to enable a comprehensive and collaborative approach to achieve its goal. The eight groups are further divided into two main categories: the compute stack and enablers. Working groups within the compute stack include *materials and devices*, *circuits and architecture*, *advanced packaging/heterogeneous integration*, and *algorithms and software*, emphasizing the need for co-design from bits and instructions to algorithms and applications, to achieve energy efficiency. Enablers focus on efforts supporting the compute stack, including *power and control electronics, manufacturing energy efficiency and sustainability, metrology and benchmarking*, and *education and workforce development*. This categorization ensures a comprehensive approach to optimizing every aspect of computing technology for minimal energy use. Figure. 1 graphically illustrates the key technology options discussed in version 1.0 of the EES2 roadmap. The technologies are assessed against two metrics, timeline to maturity and energy efficiency improvement factor. Timeline to maturity corresponds to the time required to achieve a technology readiness level (TRL) of 6. Technologies already at TRL 6 are included for their potential energy efficiency improvements, despite not being incumbent technologies. Energy efficiency improvement factor is measured by comparing future performance in an energy metric against current technology (e.g., energy per bit, or switching, energy for memory access, energy per instruction etc.). For each identified technology, the roadmap discusses challenges to commercial realization and solution pathways and action plans to address those challenges. The intent is to illustrate that there are many scientific and engineering options available for energy efficiency in computing given the challenges and complexities in sub-10 nanometer technologies. The current best estimates will serve as a baseline roadmap that will be updated regularly as viability and scalability of the options are identified.



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

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

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. 1. Key technologies for energy efficiency improvements across the compute stack identified by the roadmap.

V. Conclusions and Next Steps

The EES2 goal builds on Moore's Law, guiding industry-wide R&D efforts to maintain the pace of doubling transistor density for over five decades. The US DOE's AMMTO is leading the EES2 effort by fostering collaborations and funding targeted research to transform traditional computing and microelectronic device manufacturing. The ensuing research investments highlight a unified vision to achieve substantial improvements in energy efficiency for computing including AI systems.

The initial EES2 roadmap, created through extensive collaboration and research, identifies key technology options and pathways for energy efficiency. Going forward, EES2 will focus on measuring computing energy use and evaluating new technologies, such as nature-inspired computing and quantum information processing. This strategy aims to stay flexible and responsive to technological advances, promoting a sustainable evolution of the microelectronics sector.

ACKNOWLEDGMENTS

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE).

The work is supported by Argonne National Laboratory, a U. S. Department of Energy (DoE), Office of Science, Office of Basic Energy Sciences, under DoE contract number DE-AC02-06CH11357.

SLAC National Accelerator Laboratory is managed and operated by Stanford University under DOE contract DE-AC02-76SF00515.

Sandia National Laboratories is managed and operated by NTESS under DOE NNSA contract DE-NA0003525.

Nantero is the inventor and manufacturer of the Carbon-Nano-tube memory (NRAM)

BRDG bridge to connect is a nonprofit organization serving first-in-family STEM college students.

REFERENCES

[1] Manpreet Singh, John F. Sargent Jr., and Karen M. Sutter, "Semiconductors and the Semiconductor Industry", Congressional Research Service, R47508, April 19, 2023.

[2] The Semiconductor Research Corporation (SRC) and the Semiconductor Industry Association (SIA), "The Decadal Plan for Semiconductors. Chapter 5: New Compute Trajectories for Energy-Efficient Computing." January 25, 2021. https://www.src.org/about/decadal-plan/.

[3] Hennessey, John, and David Patterson. 2019. Computer Architecture: A Quantitative Approach. 6th Edition, Burlington, MA: Morgan Kaufmann Publishers.

[4] Gallersdörfer, Ulrich, Lena Klaaßen, and Christian Stoll. "Energy Consumption of Cryptocurrencies Beyond Bitcoin." Joule 4, no. 9 (2020): 1843-1846. https://doi.org/10.1016/j.joule.2020.07.013.

[5] Koot, Martijn, and Fons Wijnhoven. "Usage Impact on Data Center Electricity Needs: A System Dynamic Forecasting Model." Applied Energy 291 (2021): 116798. https://doi.org/10.1016/j.apenergy.2021.116798.

[6] Williams, Laurence, Benjamin K. Sovacool, and Timothy J. Foxon. "The Energy Use Implications of 5G: Reviewing Whole Network Operational Energy, Embodied Energy, and Indirect Effects." Renewable and Sustainable Energy Reviews 157 (2022): 112033. https://doi.org/10.1016/j.rser.2021.112033.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] EES2 Meeting website https://ees2.slac.stanford.edu/doe-meetings-events