Materials opportunities for low-energy computing

Asif Khan*, Hui Jae Yoo, Subhash L. Shinde, Guest Editors, and Prasanna Venkat Ravindran

This issue focuses on the materials opportunities for reducing energy consumption in the different stacks of the compute hardware—devices, memory, interconnects, and microelectronics packaging. The thermodynamics of computing to understand the minimum “compute energy limits” that can be aspired is briefly reviewed. One of the most long-standing limits that remain largely ignored from a technology standpoint, the Landauer limit, is discussed. The relationship between thermodynamics and information, starting with Maxwell’s paradox, is introduced.

Introduction

With the explosion of data and processing required to turn that data into useful information, there is an unprecedented need for energy efficient computation. In the last decade, the rate at which data were generated outpaced improvements in compute efficiency, leading to high energy consumption. For perspective, data centers consume an estimated 200 terawatt hours (TWh) each year.¹ That is close to the national energy consumption of many countries such as Australia and Spain and more than half of the electricity used for transportation worldwide. The information and communication technology (ICT) ecosystem accounts for more than 2% of global CO₂ emissions, which is equivalent to the emissions from fuel in the aviation industry. Data centers and the ICT constitute 1% and 10%, respectively, of the global electricity consumption of 20,000 TWh per year.¹

Nonetheless, the industry has been paying attention to carefully manage energy efficiency in data centers. Over the last couple of years, despite the massive increase in the Internet traffic and data loads, data center electricity demand remained nearly constant, partly because of an industry-wide trend, referred to as the “hyperscale shift.”² Hyperscale data centers are massive information factories, much larger than standard

enterprise data center that provides superior energy efficiency through economies of scale and custom engineering.

Despite that, all indications are that energy usage in computing and its environmental impact will exacerbate in the future. Internet traffic is growing exponentially and had reached more than a zettabyte (10²¹ bytes) in 2017 from 50 exabyte (10¹⁸ bytes) in 2007.¹ Massive proliferation of ICT is inevitable: information services will reach billions of people who do not yet have access to high-speed Internet. Our extensive usage and consumption of data are a precursor to the Fourth Industrial Revolution and underlies the advancement in all areas, ranging from artificial intelligence and machine learning, autonomous transportation and space exploration to health care and drug discovery, public policy, and so on. The path of future technologies is, of course, unpredictable and may prove to be disruptive in terms of energy usage. For example, since the birth of the cryptocurrency Bitcoin in 2008, energy usage in cryptocurrency mining has been a concern. It is estimated that Bitcoin mining uses 20 TWh of electricity per year, which is 0.33% of global electricity use.¹ Altogether, Andrae et al. predicted that by 2030, data centers are likely to constitute 8% and ICT could use 20% of the global electricity demand.³ The recent report by the Semiconductor Research

Asif Khan, School of Electrical and Computer Engineering and Materials Science and Engineering, Georgia Institute of Technology, USA; akhan40@gatech.edu

Hui Jae Yoo, Components Research, Intel Corporation, USA; huijae.yoo.us@ieee.org

Subhash L. Shinde, Center for Sustainable Energy, University of Notre Dame, USA; sshinde@nd.edu

Prasanna Venkat Ravindran, School of Electrical and Computer Engineering, Georgia Institute of Technology, USA; rprasanna22@gatech.edu

*Corresponding author

doi:10.1557/s43577-021-00207-z

Corporation also made similar predictions in their Decadal Plan.⁴

At the very core of the energy challenge in computing lies the fact that the improvement of energy efficiency in compute hardware is waning. Over the last four decades, the number of complementary metal–oxide–semiconductor (CMOS) transistors, the basic building blocks of computing hardware, has been doubling approximately every two years. The number of operations a microprocessor can perform per kilowatt hour of energy has doubled about once every 1.6 years for peak performance and every 2.6 years for average performance. However, the rate of improvement has slowed since early 2000s. This is because the power supply voltage in the microprocessor could no longer be downscaled at the same rate as before due to the limitations imposed by the thermodynamics of the transistor operation. At the same time, transistor technology has nearly reached the ultimate limit of dimensional miniaturization, which may further decrease the generation-to-generation improvement of energy efficiency at the fundamental building block level (**Figure 1**).

All told, there is a lot of room for improvement of energy efficiency, especially considering the opportunities for innovations at the materials level. For perspective, the minimum energy lost when erasing one bit, is given by the Landauer limit and is 2.87×10^{-21} J at 300 K or $kT \ln 2$, with k and T being the Boltzmann constant and the temperature, respectively. In comparison, today's ultra-scaled transistors dissipate 10^4 – 10^5 times more energy than this thermodynamic limit. Similar

scope for improvement exists in all layers of the computing hierarchy: from devices and circuits levels to architectures and systems.

This special issue focuses on the materials opportunities for reducing energy consumptions in the different stacks of the compute hardware (devices, memory, interconnects, and microelectronics packaging). In the following, we introduce the six articles that comprise the *MRS Bulletin* topical content.

In this Introductory article, we provide a short review of the thermodynamics of computing to layout the groundwork for understanding the minimum “compute energy limits” that we can aspire for. We particularly focus on one of the most long-standing limits that remains largely ignored from a technology standpoint—the *Landauer limit*, a limit arising from the relationship between information and the second law of thermodynamics. In this article, we review the relationship between thermodynamics and information, starting with Maxwell's paradox.

The next article related to “charge-based devices” covers one specific area of materials research in ferroelectric gate oxides for negative capacitance transistors.⁵ In this article, Hoffmann, Cheema, and Salahuddin describe how the inherent native capacitance in ferroelectric materials arising from their microscopic polarization instability makes them very good candidates for high- k metal gate stacks. They also review recent discoveries of ferroelectricity and negative capacitance in HfO₂- and ZrO₂-based thin films of fluorite structure that will lead to improvements in both the energy efficiency and performance of future transistors.

The third article by Heron and Chiang related to “non-charge based devices” focuses on lowering switching energy by employing electronic spin—not charge—for computation.⁶ Magnetic devices offer non-volatility and high endurance but have historically required high charge currents for switching, namely spin currents for their “write” cycle, and longer switching times due to their dependence on magnetic precession. Multiferroic materials, that is, those having more than

one ferroic order—ferroelectric polarization/ferromagnetic magnetization—are studied for their energy-delay reducing properties in devices and heterostructures.

Ye and Hu et al. summarize the memory hierarchy and describe approaches to overcome the “memory wall” as workload demands in modern applications are quickly exceeding the available energy-bandwidth of compute systems.⁷ The authors discuss emerging materials and device technologies for next-generation non-volatile memory technologies, including phase-change memory (PCM), resistive random access memories (ReRAMs), ferroelectric field-effect transistors (FeFETs) and magnetic random access memories (MRAMs). A comparison between different technologies in the back-drop of underlying device architectures (such as back-end-of-the-line vs. front-end, 3-D cross point, 3-D NAND), application space (embedded, last level cache, in-memory, and neuromorphic) and readiness for wide-scale adoption is presented as well.

The fifth article led by Gall’s group covers the important area of research in materials for interconnects.⁸ In their article, they first review how the decrease in interconnects size while increasing their overall length (due to design complexity, etc.) increases energy consumption. They then review future research directions, including (1) new materials with a small electron mean free path to substantially reduce electron scattering at surfaces and grain boundaries become negligible, (2) 2D materials as liner/barrier layers to increase the overall conductor cross-sectional area and enhance specular surface scattering, and (3) the innovations in topological metals that show increasing conductivity with decreasing line width.

The sixth article by Swaminathan’s team describes recent research in microelectronics packaging being carried out in heterogeneous integration (HI) as a “critical enabling technology” for supporting AI applications.⁹ The AI integrated systems are enabled by different materials solutions to support electrical performance, mechanical integrity, and thermal dissipation. They cover the materials research challenges in HI for AI, including selection of materials, their processibility, and compatibility with low cost manufacturing. They also review the past and present work related to requirements for materials for HI.

The seventh article on “materials challenges in brain-inspired computing” is our salutation to this long-pursued research endeavor which is growing very fast.¹⁰ In their article, Zhao, Kang, and Ielmini provide an overview of materials opportunities in emerging neuromorphic devices for brain-inspired computing. They will summarize the biological functions of neuromorphic elements in the nervous system, discuss the requirements for their materials counterparts, review the recent progress in new materials for artificial neurons and synapses, and illustrate some cognitive computing primitives in hardware networks. Finally, they will discuss the current status and upcoming material challenges for brain-inspired computing.

We now briefly review concepts related to the Thermodynamics of Computing in the sections below.

Finally, an article by Banerjee’s group, which will appear in the December 2021 issue of MRS Bulletin, will survey the latest advances in two-dimensional materials, ranging from insulators to superconductors, which can potentially increase the energy efficiency of computation.¹¹ A system incorporating sub-kT switching beyond the Boltzmann limit, resistive memories with high nonvolatility, all interconnected by high-conductivity wiring with minimal IR loss will be envisioned. The need for architectural changes to fully utilize the energy efficiency of these new materials such as in-memory computing will also be recognized.

The thermodynamics of computing

The second law of thermodynamics, as formulated by Clausius, gives the relation between heat transfer and work, and is stated as “Heat can never pass from a colder to a warmer body without some other change, connected therewith, occurring at the same time.” This implies that for a cyclic process $\oint \frac{\delta Q}{T} \leq 0$, where δQ is the energy flowing into the system and T , the temperature of the surrounding. Corollary to this, for a cyclic process, the entropy of the system should increase for any irreversible process and remain constant for reversible ones, that is, $\Delta S \geq 0$.

In 1867, while studying the limitations of the second law, Maxwell postulated a demon, called temperature demon, with sharp-sorting skills who could track individual molecules and measure its velocity.¹² This demon was described to control a tiny valve in the wall in the middle of a box containing gas in thermal equilibrium. The wall separated the box into two sections. The demon controlled the valve to allow only molecules with velocity higher than the average velocity to pass from A to B and those with velocity lower than the average value to pass from B to A. After the first transfer, there is an infinitesimal temperature difference between the two sections. When continued, such a sorting process would eventually lead to an increase in temperature in section B and a decrease in section A, resulting in the transfer of heat from a colder section A to a hotter section B.

Over the years, several interpretations and explanations have been discussed to resolve this paradox. The initial attempts worked towards replacing the demon and its intelligence with a physical mechanism but failed as the demonic action required a perpetual motion. Smoluchowski proposed a spring-loaded trap door that would open only when particles from one side were incident on them, emulating a pressure demon. However, the trapdoor was heated by the incident molecules leading to thermal agitations that beyond a point left the hole open long enough to restore thermal equilibrium. Later, Feynman provided a more sophisticated physical interpretation of the Maxwell’s demon using a ratchet and pawl and an electrical rectifier. Both these interpretations fail to emulate the automatic operation of the demon as they heat up and behave randomly, allowing thermal equilibration and conservation of the second law. Smoluchowski and Feynman ignored the possibility of a perpetual motion machine, and hence, left the paradox open. A major development was made by Szilard when he for the first time considered the

“intelligence” of the demon as information and pointed to its significance in the paradox.

In 1929, Szilard presented a one-molecule experiment intended to find the role of “measurement” in the thermodynamic paradox.^{13,14} He considered a hollow cylinder containing a gas with one-molecule occupying the total volume V and set in a constant temperature bath. A massless barrier is inserted in the middle of the cylinder, dividing it into two equal halves. The demon “measures” whether the molecule is in section A or B and attaches a mass to the side of the barrier with the molecule. The gas expands isothermally to occupy the initial volume V and performs useful work on the mass by extracting the same amount of heat from the heat bath. The amount of work extracted by the isothermal expansion from $V/2$ to V is $kT\ln 2$. This is exactly the Maxwell’s paradox, a violation of the second law as the entropy of the heat bath has reduced by $k\ln 2$. Szilard called this quantity as the fundamental amount and postulated that the measurement accounted for the equal or greater magnitude of entropy increase. Twenty years later, Shannon would describe information theoretic entropy as the information content of a random variable.

The ingenuity of Szilard’s work is highlighted by his identification of measurement and memory as the contributors of the entropy decrease, which even later went on to form the crucial components of information theory and its relationship to thermodynamics. That being said, he did not provide any details on the relationship between memory and the second law.

The paradox was put to rest by independent works by Landauer, Penrose, and Bennett. In 1961, Landauer linked the paradox to the then emerging area of computing and described logical irreversibility and erasing of information as the source of heat generation and entropy increase.¹⁵ Penrose provided a mathematical and abstract version of the same result as Landauer in 1970.¹⁶

The advent of computing had given rise to binary arithmetic using “0” and “1” to store and process the information. Computing systems are physical systems that process information based on a mapping between the logical and physical states. Landauer considers a bistable potential well (physical system) resembling the energy landscapes of materials like ferrites, ferroelectrics, and ferromagnetic films as the storage unit of a bit of information. Each well (physical state) corresponds to “0” and “1” (logical state). The barrier height is assumed to be much higher than the thermal fluctuations to remove the possibility of a probabilistic manipulation of the bit.

If the initial state of the bit is fixed, according to the statistical definition of entropy, $S = k\ln Z$, the entropy of the bit is zero. Z is the number of possible microstates in the system for a given condition. Considering an M bit system (with 2^M possible microstates) and allowing for a random initial state, the entropy of the system is $k\ln 2$ per bit. An erasure of information is emulated here by setting all the bits to “1” irrespective of the initial state of the bit. This process of erasing reduces the entropy to zero, a decrease in entropy of $k\ln 2$ per bit. For a closed system, the entropy of the system cannot reduce, leading to the loss of

Figure 2. Reversible measurements in a system with a double-well energy landscape. Reproduced from Reference 18.

an equivalent amount of energy, given by $kT\ln 2$ per erased bit. This is called the erasure principle or Landauer’s principle.

This erasure maps both “0” and “1” to “1,” that is, a many-to-one mapping and does not have a unique inverse, making it a logically irreversible process, a concept introduced by Landauer and perceived as the more generic origin of the heat dissipation. This would entail that logic gates like NAND and NOR are many-to-one maps and irreversible and would result in heat dissipation due to the erased bits of information.

Bennett extended Landauer’s ideas and proposed a reversible measurement method which involved no entropy change. This measurement method uses a standard state, the state of the memory when no measurement is done. This state can be achieved in a double-well potential by lowering the central barrier and making both the wells undiscernible. In case of ferromagnets, it can be achieved by applying a transverse electric field as shown in **Figure 2**.¹⁷

Landauer’s erasure principle and the reversible measurements proposed by Bennett put the Maxwell paradox to an end after about hundred years and several generations of physicists. At room temperature (~ 300 K), the energy lost when erasing one bit, as given by the Landauer limit is 2.87×10^{-21} J.¹⁹

Conclusions

Energy dissipation remains a major, unsolved challenge in computing, which left unchecked, could lead to adverse environmental impact on a global scale. To this end, there have been collective efforts among academia, industry, and government to explore multi-faceted approaches for advancing low-energy computing. There are plenty of opportunities for materials innovations in computing hardware that can address these challenges. This issue focuses on the materials opportunities for reducing energy consumption in the different stacks of the computing hardware (devices, memory, interconnects, and microelectronics packaging).

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. N. Jones, *Nature* **561**, 7722 (2018)
2. G. Kamiya, O. Kvamström, *Data Centres and Energy—From Global Headlines to Local Headaches?* (International Energy Agency, Paris, France, 2019). <https://www.iea.org/commentaries/data-centres-and-energy-from-global-headlines-to-local-headaches>. Accessed 6 Sep 2021
3. A.S.G. Andrae, T. Edler, *Challenges* (2015). <https://doi.org/10.3390/challe6010117>
4. *The Decadal Plan for Semiconductors* (Semiconductor Research Corp., Durham, NC, 2021). <https://www.src.org/about/decadal-plan>. Accessed 6 Sep 2021
5. M. Hoffmann, S. Salahuddin, *MRS Bull.* **46**, 10 (2021)
6. J.T. Heron, T. Chiang, *MRS Bull.* **46**, 10 (2021)
7. M. Si, H.-Y. Cheng, T. Ando, G. Hu, P.D. Ye, *MRS Bull.* **46**, 10 (2021)
8. D. Gall, J.J. Cha, Z. Chen, H.-J. Han, C. Hinkle, J.A. Robinson, R. Sundararaman, R. Torsi, *MRS Bull.* **46**, 10 (2021)
9. M. Swaminathan, M. Kathaperumal, K.-s. Moon, H. Sharma, P. Murali, S. Ravichandran, *MRS Bull.* **46**, 10 (2021)
10. Y.D. Zhao, J.F. Kang, D. Ielmini, *MRS Bull.* **46**, 10 (2021)
11. A. Pal, K. Agashiwala, J. Jiang, D. Zhang, T. Chavan, A. Kumar, C.H. Yeh, W. Cao, K. Banerjee, *MRS Bull.* **46**, 12 (2021)
12. J.C. Maxwell, in *Life and Scientific Work of Peter Guthrie Tait*, C.G. Knott, Ed. (Cambridge University Press, Cambridge, UK, 1911), p. 213
13. L. Szilard, *Behav. Sci.* **9**(4), 301 (1964)
14. L. Szilard, *Z. Phys.* **53**(11), 840 (1929)
15. R. Landauer, *IBM J. Res. Dev.* **5**(3), 183 (1961)
16. O. Penrose, *Foundations of Statistical Mechanics: A Deductive Treatment*. Dover Publications, Mineola, NY, 2005
17. C. Bennett, *Int. J. Theor. Phys.* **21**(12), 905 (1982)
18. K. Maruyama, F. Nori, V. Vedral, *Rev. Mod. Phys.* **81**(1), 1 (2009)
19. S. Lloyd, *Nature* **406**(6799), 1047 (2000)

Asif Khan is an assistant professor in the School of Electrical and Computer Engineering and of Materials Science and Engineering at the Georgia Institute of Technology. His group focuses on all aspects of ferroelectricity ranging from materials physics, growth, and electron microscopy to device fabrication and ferroelectric circuits for artificial intelligence and machine learning applications. His early career work led to the experimental proof-of-concept demonstration of ferroelectric negative capacitance that can reduce power dissipation in transistors. Khan received a National Science Foundation CAREER Award (2021), a Defense Advanced Research Projects Agency Young

Faculty Award (2021), and an Intel Rising Star Faculty Award (2020). His research portfolio includes two book chapters and 70+ journal and conference publications, including those in *IEEE Electron Device Letters*, *Nature Electronics*, *Nature Materials*, and *IEEE International Electron Devices Meeting*. Khan can be reached by email at akhn40@gatech.edu.

Hui Jae Yoo is a semiconductor process technologist working on integration of novel materials for power/performance/scaling of next-generation logic/memory devices. He has been involved in multiple generations of technology transfer from early research into pathfinding. He is with Components Research at Intel Corporation, and obtained his degrees in electrical engineering from the California Institute of Technology. Yoo can be reached by email at huijae.yoo.us@ieee.org.

Subhash L. Shinde is an associate director of the Center for Sustainable Energy at the University of Notre Dame (ND Energy), leads initiatives in energy research, global development, and corporate programs. He obtained his PhD degree in materials science and engineering from Stanford University. Prior to his position at ND Energy, he was a member of the Research Staff and an R&D manager at Sandia National Laboratories. Previously, he was at IBM's T.J. Watson Research Center and Microelectronics Division. Shinde has carried out research in three-dimensional integration of microsystems (Sandia), advanced thermal management (Sandia, IBM), and materials compatibility issues for future microsystems. He is a Fellow of IEEE and the National Academy of Inventors (NAI). Shinde can be reached by email at sshinde@nd.edu.

Prasanna Venkat Ravindran is a third-year PhD student in Asif Khan's Lab in the School of Electrical and Computer Engineering at Georgia Institute of Technology. He received his BTech degree in electrical engineering from the Indian Institute of Technology Palakkad in 2019, having achieved the Department Gold Medal. He is a recipient of the IEEE Electron Devices Society Masters Fellowship (2020) and Center for Research into Novel Computing Hierarchies PhD Fellowship (2021). His current research interests include ferroelectric materials with emphasis on domain dynamics, polycrystallinity, and quantum paraelectricity, and their applications in novel computing technologies. Ravindran can be reached by email at rprasanna22@gatech.edu.