Energy of Computing as A Key Design Aspect for Sustainability

Sadasivan Shankar
SLAC National Laboratory, Menlo Park, CA; Materials Science and Engineering
Stanford University, CA, USA
{sshankar@slac.stanford.edu; sadas.shankar@stanford.edu}

Tina Kaarsberg
Advanced Materials and Manufacturing Technology Office, U.S. Department of Energy

Challenge: *What is the Problem?*

Based on the exponentially increasing energy demands for computing and its applications and the ubiquitous use of digitalization, the trajectory of computing appears unsustainable both from energy and materials perspectives.[1],[2],[3] This problem is further compounded by the slowing of Moore's law and the increasing use of Artificial Intelligence Methods in all aspects of the modern economy in the 21st century. The energy needed per simulation can be *several tens of orders of magnitude* or higher than the thermodynamic energy limit at room temperatures[4]. This disparity is due to several factors [5]. Some of these factors include slowing down of the traditional geometrical scaling semiconductor devices (Moore's law), widespread use of smart devices which are connected to internet, wider adoption of use of data centers for processing information at lower cost to consumers, and limitations of traditional digital computer architectures[6] in processing large amounts of data. In addition, demands on demands on data and accuracy can be computationally intensive requiring large numbers of higher precision operations. In combination, these effects, further accelerated by the digitization of the economy and ubiquitous

---

[1] https://www.imf.org/en/Blogs/Articles/2023/03/21/how-pandemic-accelerated-digital-transformation-in-advanced-economies
U.S. Department of Energy, "Semiconductor Supply Chain Deep Dive Assessment," February 24, 2022, https://www.energy.gov/sites/default/files/2022-02/Semiconductor%20Supply%20Chain%20Report%20-%20Final.pdf

[2] Shankar, S.; Reuther, A. Trends in Energy Estimates for Computing in AI/Machine Learning Accelerators, Supercomputers, and Compute-Intensive Applications. In Proceedings of the 2022 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 19–23 September 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–8.

[3] Semiconductor Research Corporation. (2020, October). The Decadal Plan for Semiconductors: A Pivotal Roadmap Outlining Research Priorities. SRC. https://www.src.org/about/decadal-plan/
https://srcmapt.org/chapter2/

[4] S. Shankar, "Energy Estimates Across Layers of Computing: From Devices to Large-Scale Applications in Natural Language Processing, Scientific Computing, and Cryptocurrency Mining", submitted to IEEE HPEC, 2023.

[5] S. Shankar, "Lessons from Nature for Computing: Looking beyond Moore's Law with Special Purpose Computing and Co-design," 2021 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 2021, pp. 1-8, doi: 10.1109/HPEC49654.2021.9622865.

[6] Hennessy, J.L. and Patterson, D.A., 2019. A new golden age for computer architecture. Communications of the ACM, 62(2), pp.48-60.

availability of data analytic tools, computing devices, and increasing use of sensors, have all together led to an increase in energy requirements across all layers of computing systems as illustrated below.

Very large-scale integration in microelectronics from materials to systems using advanced manufacturing technologies has enabled a revolution in computing to become a critical component of all aspects of civilization. This revolutionary era known by the moniker "Moore's law" has been leading this charge since 1965.[7] In turn, this era has led to the application of computing accelerating innovations in all areas from computer-aided design to drug discovery, weather predictions, and autonomous vehicles. Moore's law, which is an empirical prediction, has proved to be accurate for over five decades accelerating the economy of the developed and developing worlds. By early 2010s[8], the internet powered by the revolutions in computing and microelectronics has been estimated to be responsible for a more than a fifth of the world economic growth in developed countries alone and that between 11.74 and 18.63 percent of productivity growth during 1960 to 2019 can be attributed to physical changes in the size of electronic components[9]. Although geometric scaling of circuit elements in microelectronics has led to improved energy efficiency over the past five decades, the difficulty in sustaining Moore's law cadence coupled with the ever-increasing need for computing threatens to undercut any improvements in energy efficiency moving forward for the surging digital economy.

The Cause: W*hy is computing needing so much energy?*

Estimates of energy requirements of computing in different systems from chips to racks indicate the following[1]: energy efficiency of computing systems decreases as transistors and chips are integrated together. The integration of these components into complex systems are required by Machine Learning and Artificial Intelligence[10], Crypto Coin mining[11] and other large scale scientific and business simulation applications. It follows that realizing these applications has been associated with an overall increase in energy requirements of computing systems. Moreover, based on the current trends, the shift to clean and sustainable energy sources is alone not sufficient to meet this increasing energy use for computing.

Energy used by computing leads to a non-linear multiplicative effect (we are terming the effect of energy for computing as a "3E" or more to indicate that the use of energy has additional collateral needs for more energy) as illustrated below. During computing, the highly efficient form of electronic energy is converted to heat, the least efficient form of energy[4] ("1E"). Consequently, all the heat generated needs to be removed for enabling safe operation of computing devices ("2E"). Next, to install complex data centers and large computer systems, refrigeration and automation requires additional electronic components for operation[12] ("3E"). Further, with the digitalization of all aspects of the economy, incorporation of renewable energy sources, and widespread use of artificial intelligence, the electrical grid system with its own computing intelligence requires more energy for its computing needs (beyond "3E").

---

[7] Schaller, R.R.: Moore's law: past, present and future. IEEE spectrum 34(6) (1997) 52-59

[8] Manyika, J., & Roxburgh, C. (2011, October). The great transformer: The impact of the Internet on economic growth and prosperity; https://www.mckinsey.com/~/media/McKinsey/Industries/Technology%20Media%20and%20Telecommunications/High%20Tech/Our%20Insights/The%20great%20transformer/MGI_Impact_of_Internet_on_economic_growth.pdf

[9] https://www.newyorkfed.org/research/staff_reports/sr970

[10] Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M. and Villalobos, P., 2022. Compute trends across three eras of machine learning. arXiv preprint arXiv:2202.05924.

[11] New York Times (2023), The Real-World Costs of the Digital Race for Bitcoin, Published April 9, 2023, Updated April 11, 2023

[12] Sverdlik, Y. (2018, August 2). *Google is Switching to a Self-Driving Data Center Management System*. Data Center Knowledge. https://www.datacenterknowledge.com/google-alphabet/google-switching-self-driving-data-center-management-system

Opportunity: *How do we solve the problem of energy of computing for a sustainable world?*

The pathway to sustainable computing is through *innovations* in which energy itself is a key design variable, i.e., use energy as an additional basis during the design. The semiconductor and computing industry over its five-decade plus history has been the leader in using innovations in materials, devices, systems, architectures, hardware systems, algorithms, software, processing, and manufacturing to provide solutions to overcome problems facing the world. Computing systems are complex and multi-layered, spanning nanometer sized devices to large city block-sized data centers processing complex signals. Hence, considerations of energy efficiency are needed at all these levels from the bit switching to system operations and algorithms and software used for simulations. Moreover, losses in energy efficiency become amplified when these components (e.g., software-hardware-devices-systems-manufacturing) are integrated together. As these layers have historically been designed independently, innovations at these intersections provide an important opportunity for addressing the computing energy usage challenge.

It is important that design strategies spanning the long road from atoms to algorithms be adopted for addressing unsustainable trends in energy for computing. One example of identifying innovation pathways includes recent Department of Energy efforts on Energy Efficient Computing in which the US Department of Energy has initiated a collaborative ecosystem with the national laboratories, industrial partners, academia, and international agencies to work for a bi-decadal plan to reduce energy use in computing.[13] As part of this effort, initial estimates[4] reveal that there is enough headroom for significant energy reduction (over a million times) during computing from atoms to algorithms. While it is unlikely that any single solution can address the increasing energy demand for computing challenge, new approaches undoubtedly will need to critically address the energy efficiency of computing at all levels – chips, devices, software, systems, manufacturing, including optimization of energy and material distributions and the corresponding supply chains.

The pathways to sustainability are dependent on quantification and qualification of information, which needs development of specific metrics at all levels from components to systems. As illustrated with examples above, metrics help with monitoring indicators of outcomes thereby enabling measuring progress and setting new directions. This is especially important for estimating the degree of success for sustainability, without which it will be hard to assess practicality and progress. We think that that the path to sustainability for all aspects of computing is achievable to the industry known for its relentless innovations. In this regard, the efforts for being undertaken in scoping the problem, formulating the metrics, and setting up a long roadmap by IEEE[14] for a planet positive 2030 world, the US Department of Energy EES2 Roadmap for Energy Efficient Computing (EES2)[15] for a 1000X or more reduction, and other organizations can enable a new and exciting era where energy efficient computing using lessons from nature is realizable, leading to potential cascade effects for a sustainable future!

---

[13] https://www.energy.gov/eere/articles/department-energy-announces-pledges-21-organizations-increase-energy-efficiency

[14] https://globalpolicy.ieee.org/promoting-strong-sustainability-by-design/#:~:text=Planet%20Positive%202030%20is%20an,future%20for%202030%20and%20beyond.

[15] https://live-slac-microelectronics-d9.pantheonsite.io/events/us-doe-ees2-meetings/doe-ees2-pledge-signing-event