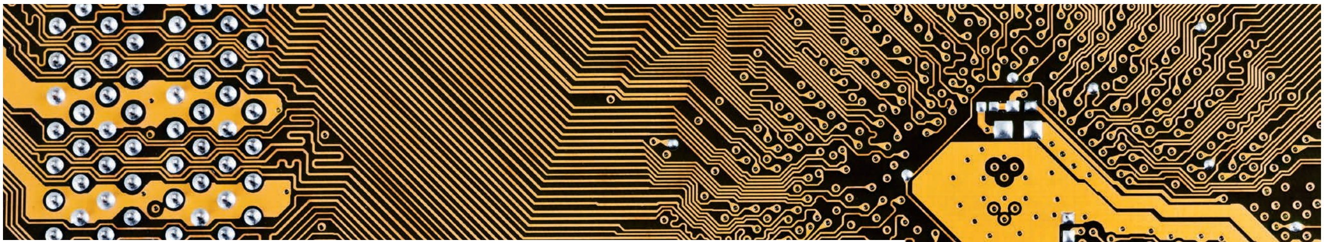
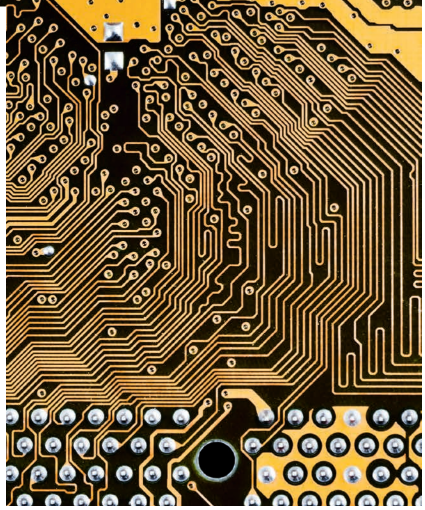

IMAGE LICENSED BY INGRAM PUBLISHING

Challenging Trends in Energy of Computing for Data Centers

Sadasivan Shankar, SLAC National Laboratory and Stanford University

The unsustainable energy requirements of computing data centers are driven by the growth of artificial intelligence, crypto coin mining, and other emerging applications underscoring the need for energy efficiency across all layers of computing.

Worldwide digitalization has led to exponential increases in the processing of digital data across all aspects of the economy. Specifically, this correlates very strongly with the use of computing in large applications, including artificial intelligence (AI), crypto coin mining, large-scale scientific computing, and sensing. The computing needs for these intensive applications are primarily supported by large data centers. Computers, which form the workhorse of these data centers, require significant amounts of energy to operate as well as resources for the subsequent cooling and thermal management of all the systems. The total energy intensity in data centers results in what we term the *3E effect*, in which a single unit of compute may need three or more units of energy for computing, cooling, and advanced electronics for energy and resource management. This problem can be traced to energy use across layers of computing. The unsustainable trends discussed in this article underscore the need for energy-efficient computing in all aspects of data processing, where energy efficiency should be a design variable bridging atoms to algorithms.

INTRODUCTION

Emerging applications such as AI, crypto coin mining, digital sensors, and advanced model-based applications such as digital twins (DTs) have been leading to exponential growth in computing usage. In turn, these trends have magnified the energy of computing as one of the grand challenges as we move into a digitalization era.[1,2,3,4,5] The increase in energy is more evident in large data centers where data are stored,

Digital Object Identifier 10.1109/MC.2024.3468568
Date of current version: 15 November 2024

EDITOR CONRAD JAMES
Sandia National Laboratories, USA;
cdjame@sandia.gov

accessed, and computed, making them very resource and energy intensive.[6] This article addresses the challenges due to the exponentially increasing resource requirements for data storing and processing. In the next section, we use specific examples to illustrate applications that are driving large needs in computing and estimate energy in a subset of these key applications. In the "Energy and Resource Requirements in Data Centers" section, we discuss the relationships of energy in computing with the trends in data centers. We conclude with specific pointers to reduce the energy requirements in computing. In conclusion, we summarize the key problem and specific pointers for enabling energy efficiency across all layers of computing so that, in effect, the data centers become energy efficient.

ENERGY OF COMPUTATION

In this section, we shall focus on a few specific applications resulting in significant growth in computing and the specific trends that have led to the exponential growth of data centers. Our estimates are based on published literature using power and the peak number of instructions per second for integer four-bit (INT4), integer eight-bit (INT8), floating-point (FP) 16-bit (FP16), FP 32-bit (FP32), and FP 64-bit (FP64). For Top500 supercomputers, we use the peak power and standard benchmarks.[7] In all of our analyses, we observe that the energy needed per simulation can be several tens of orders of magnitude or higher than the minimum energy required for switching a single bit or the thermodynamic limit of information processing.[8]

Applications driving the growth of data centers

Consistent with the previous analysis, the *application* is defined to quantify the operations associated with the entire simulation by the process of computation. Correspondingly, the energy for the application, defined as *energy per application* (*EPA*), is a metric that includes algorithms and software. Analogously, energy/instructions represent the architecture implemented at the hardware system level and energy/bit switching at the transistor level of a single switch.[7] The emerging applications of AI/machine learning (ML) methods have been driving significant growth in compute operations over the last decade.[7,9,10] As the AI/ML methods stabilize, these methods have led to several emerging applications for computing, including autonomous cars with Level 4 or higher automation, DTs for enabling the accelerated discovery of chemicals and materials, intelligent sensing, and industrial instrumentation, further driving new innovations like smart grids requiring more computing. We will briefly illustrate these applications in the following sections.

This article addresses the challenges due to the exponentially increasing resource requirements for data storing and processing.

AI and ML. Nature language processing (NLP) based on large language models has become the application of choice for ML from language to scientific analysis.[11] Among these approaches, the Transformer-based models, which can be parallelized, enable significant scale-up to larger modes.[12] The high number of computations required by these models is mostly driven by two of their characteristics: a neural network-based model (Transformer-based) with a large number of parameters (hundreds of billions or more) and the significant amounts of data needed to train these models. The data for training consists of words, phrases, part-of-speech requirements, existing collections of text from academic journals, books, social network websites, Wikipedia, and Common Crawl (that is, crawling the web).

The number of FP operations ranges from 6×10^{18} to 3×10^{24} for each model training.[7] Based on estimating the energy per instruction (EPI) to range from 1×10^{-12} J/FPI (FP16) (lower bound) to 1×10^{-11} J/FPI (FP64) (upper bound), the energy required varies from 6 million joules to 30 trillion joules. Inference in AI/ML methods is also energy intensive given the duration of their use in applications. Typically, the time during which inference is used is longer (estimated here annually, given its regular estimated usage) compared to that for training (in this example, estimated to be about 55 days, once per year as indicated in Shankar[8]). For inference, the energy estimates are based on the usage patterns of large language models as reported in the literature.[13,14] The number of FP operations to compute an inference for a model with 175 billion parameters varies between 1 million FPs to 175 billion FPs per word. The details of the estimates of the bounds are provided in a previous paper assuming 10 million queries/day, each query limited to 500 words.[8] The energies estimated range from 41 million joules (1 million FPs per word) to 798 billion joules (175 billion FPs) per year. At the upper limit, the energy used in inference is significantly higher than per capita electricity usage in the United States and even higher than the total energy expended by a human in a lifetime of 73 years.

Scientific computations. Similar to the AI/ML applications, we illustrate the energy for scientific simulation of the dynamics of a virion.[15] Although the overall simulations included AI models in addition to molecular dynamics, we use this case to illustrate the application in scientific computing. Simulations were run for 8.77 days on 80 P100 GPUs for ~7.5 ms. Based on the simulation time and the total performance and assuming the energy per FP operation to be bounded by 1×10^{-12} to 1×10^{-11} J as before, the energy for the Spike ACE Complex and SARS COVID virion simulation is estimated to be about 24.9 billion joules as estimated before.[8] Using the estimates from the previous calculations,[7,8] the average energies are estimated for a single operation, where the operation could be a transistor switching [energy per bit (EPB)], an instruction (from four bits to 64 bits as EPI), or the application covering the entire simulation (for example, EPA for training a single model using a large language model or the simulation of a specific dynamic problem for a given duration), and are plotted in Figure 1. Two specific observations can be highlighted from this analysis. As the computing proceeds from that of a single bit switching to a single simulation, as in the application mentioned previously, the energy goes from 1×10^{-17} J to between 1×10^{9}-3×10^{12} J, more than 25 orders of magnitude. The second observation is that compared to the thermodynamic limit (~4.2×10^{-21} J at 300 K), digital computations are generally energy intensive, increasing as we move from *bits* to *algorithms*.

The U.S. Department of Energy and its laboratories are leaders in the use of scientific computing for the analysis, discovery, and design of problems covering biological, chemical, physical, and materials problems. In addition, measurement data in these laboratories from several user facilities and characterization labs need significant computing power to measure, extract, and analyze. For example, in 2023, 30% of the top 10 and 20% of the top 20 supercomputers in the world were installed and operated by the U.S. Department of Energy labs. Analysis of 500 supercomputers maintained by Top500, the most powerful computer systems, with statistics on performance on the high-performance LINPACK (HPL) benchmark quantified the energy used by these computers.[16,17] The energy estimates for high-performance computing have two components: *hardware/system-level architecture and algorithms/software*, which are strongly codependent. As the integrated systems in large-scale supercomputers have many components, the energy needs to be estimated using benchmarks and also the time for the successful completion of simulations. The estimates have been discussed in previous papers.[7,8]

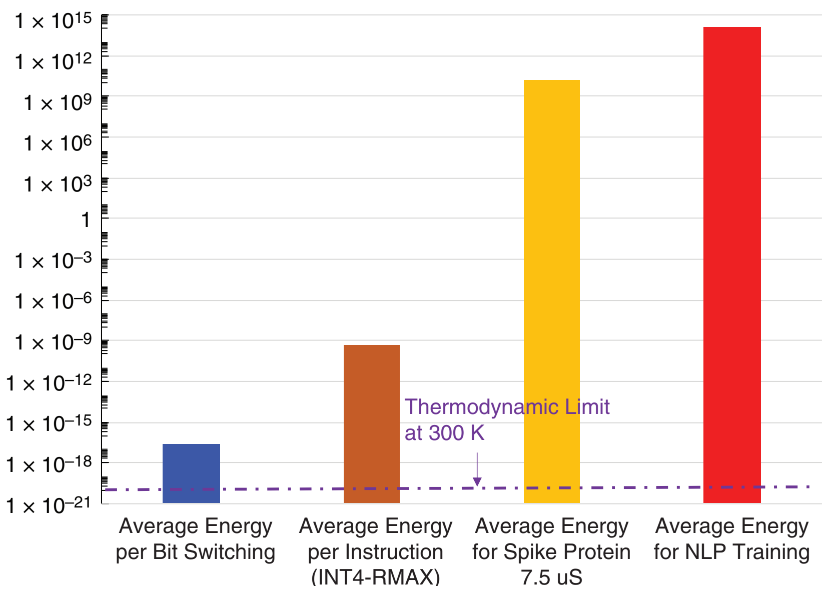

FIGURE 1. Average energy in joules per switching, per operation, or per application, adapted from S. Shankar[8]. RMAX is the maximum performance achieved on the High Performance LINPACK benchmark.

Crypto coin mining. The significant computing needed for cryptocurrency mining is the result of computational operations of a virtual online transaction included in a ledger, which needs to be validated by several connected online computer nodes. The design of this operation is compute intensive both due to the nature of the one-way operation termed as *hashing* that maps digital inputs into a fixed length of output digits and due to the requirements for verification.[18,19] Computing power for crypto coins is realized with the help of farms, represented as the plurality of a large number of video cards connected to computers and a number of application-specified integrated circuit (ASIC) computing modules using ASIC.[20,21]

A cryptocurrency network's effectiveness can be measured by its hash rate, which is the number of computer users (termed as *miners*) working to verify transactions and the rate of hash generation. The users run computations on complex mathematical puzzles based on transaction data and have to generate a significant number of guesses (more than six orders of magnitude from million to trillion) per second. These are hashes—alphanumeric codes that are randomized to identify a single unique piece of data. The application in this case computes to reach a block of transaction data validated as the "correct solution" based on the amount of used computation.

The validated block is added to the existing chain, and the user receives a newly minted cryptocurrency. By the very nature of the workflow setup for the application, significant computational power is needed for advancement.[18] Energy estimates of computation for crypto coin applications are compared with others in a previous publication.[8]

Electricity demand associated with cryptocurrency mining operations in the United States has grown very rapidly over the last several years. Estimates indicate that annual electricity use from cryptocurrency mining probably represents from 0.6% to 2.3% of U.S. electricity consumption.[19] As indicated in other publications,[22] 34 large-scale operations termed *bitcoin mining* operate in the United States, putting tighter pressure on the local power grids. As illustrated later, the energy usage in computer-based mining of crypto coins is turning out to be a significant fraction of the electricity used in computing, including data centers and AI-based applications. In Figure 2, the lower energy estimates (electrical energy associated with crypto computing) from 2016 to 2024 are compared with annual electricity production in different states (for example, Arizona and California), countries (for example, Australia, the Netherlands, and the United Kingdom), and annual energy generation by Hoover Dam hydroelectric project.[23]

Intelligent sensing and smart grids. Intelligent sensors (sensors closely connected with the information processing of the data) range from large scientific instrumentation to distributed devices or automated systems (for example, autonomous/driverless cars, advanced manufacturing and assembly, and sensing of the weather and ambient conditions), which are collectively named the *Internet of Things*. Sensors from highly specialized and complex measurement equipment to widely distributed devices of more than 30 billion connected devices generate significant amounts of data (~100 zettabytes/year), which are growing exponentially annually, driven by classical digital, analog, and quantum instrumentations, all of which require large amounts of computing power.

In addition, with the electrification of transportation, incorporation of renewable energy sources, and widespread use of AI, the electrical grid system for a sustainable world with heterogeneous energy sources is going to be driven by a combination of digital technologies for optimal energy generation, distribution, and utilization. The current grid system in the United States and in the world consists of a patchwork for an electrical generating system connected by sectors into thousands of miles of transmission lines. Throughout the world, the patchworked grids cover 80 million km. This system results in several challenges to overcome:

1. nonoptimal current usage
2. lack of security
3. lack of flexibility in balancing different heterogeneous energy sources
4. energy inefficiency in storage and transmission with asynchronous matching of sources to loads

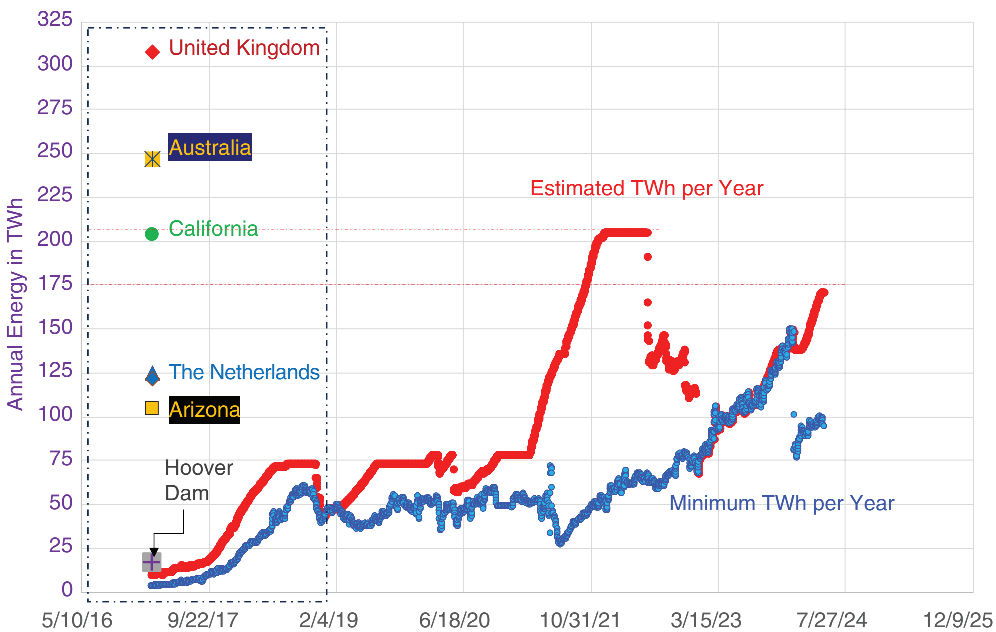

FIGURE 2. Energy estimates of crypto coin mining. Comparisons to the electricity produced in two states in the United States, three countries, and that generated by the Hoover Dam project. Adapted and expanded from "Bitcoin energy consumption index"[23] with the U.S. Energy Information Administration (EIA) and IEA data.

5. inability to anticipate changing weather events.

According to the International Energy Agency (IEA), although investments in electricity grids increased by 6% in 2021, with advanced economies accelerating investment to support and enable the electrification of buildings, industry, and transport, they need to double to reach net zero in a new sustainable economy by 2030. A smart grid is a dynamic network that will maximize the use of a carbon-neutral energy network for the minimization of loss and variability, enabling a two-way flow of electricity and data. The complexity of analog physical systems with digital communications technologies is necessary to sense, modulate, and respond to the dynamics and variations in energy generation, storage, and usage. This system can be termed the *Internet of Energy* (*IoE*).[24] The smart IoE will consist of the integration of sensors, digital controls, computers, automation, machine intelligence, and new technologies and equipment working together to optimize energy distribution and utilization. As can be seen, the IoE is at the nexus of information and energy processing requiring access to local and significant computing resources, which may require additional data centers. The IEA and the governments of the United States, China, European nations, and other countries are trying to evaluate the design of a new grid system and are seeing this transition as a fundamental step.

Digital Twins. A DT is defined as a virtual model that accurately represents a physical object embedded in a corresponding digital version of its environment using either computational models or experimental data.[26,27,28] Smart R&D and manufacturing, referred to as *Industry 4.0*, are expected to encompass the optimal use of sensors, data communication/processing, and AI/ML methods to connect real manufacturing systems with their virtual counterparts. As the specific name indicates, a DT is a computer-based virtual model, digital in terms of its informational basis, and uses the same inputs as the physical entity that it represents. It can be a system, a process, or a combination of elements that represents a real physical entity. Although many of the real systems can be nondigital, including information processing in analog, quantum, or hybrid systems, *digital* here refers to a virtual computer representation of a real-world system. For example, the data from the physical sensors are inputted into the DT, and its output is a computer model that has all the critical characteristics of the original entity.

The effective use of DTs from R&D to material characterization and to automation will require three components that make this application data and compute intensive: measurement-based experimental data, large-scale computer simulations, and computer-based exploration of what-if counterfactuals. This is illustrated in Figure 3 as applied to accelerated material discovery or autonomous manufacturing integrated systems such as semiconductor processing.[29] Research using DTs should have access to high-performance computing with appropriate infrastructure ranging from handheld mobile devices to large-scale computer centers. Collecting and processing data need access to the data centers and communication with the measurement facilities where the experiments are being run.

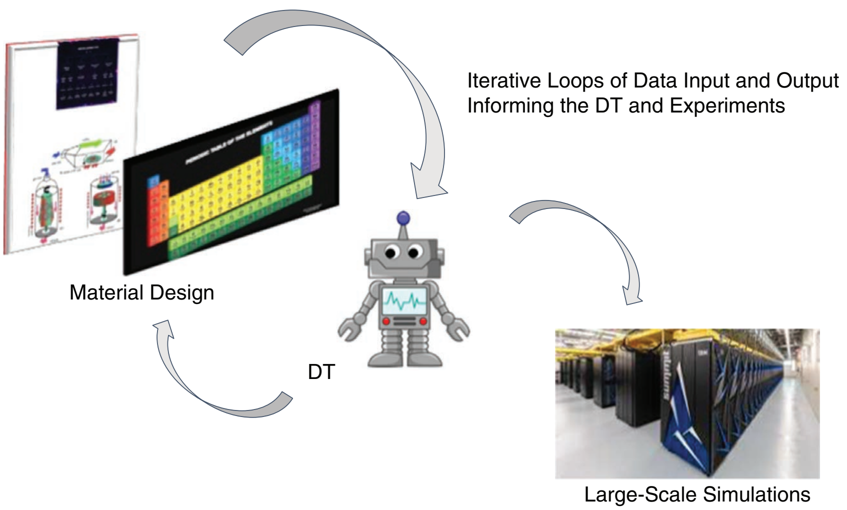

FIGURE 3. DTs. Components that make these application data and compute intensive are experimental data, large-scale simulations to analyze and generate data, and computations to explore what-if counterfactuals.

ENERGY AND RESOURCE REQUIREMENTS IN DATA CENTERS

Data centers are buildings that house infrastructures for computer systems and related components, including hardware for computing, storage, and communications. The data centers, which serve as large hubs in which big data are processed and stored, serve as the backdrop of all digital computing. Data centers also need access to continuous electricity for all aspects of computing and subsequent air conditioning for dispersing and cooling the computational and storage systems. Most of the aforementioned applications need large data centers for enabling the simulations that are typically continuously active around the clock depending on the requirements. In 2024, the total number of data centers in the world exceeds more than 10,000, with more than 50% of them residing in the United States alone.[30] Data center distribution across different countries as of March 2024 is shown in Figure 4.[30] Of the 10,593 data centers worldwide, 5,381 are in the United States, followed by 521 in Germany, 514 in the United Kingdom, 449 in China, 336 in Canada, 315 in France, and 307 in Canada, together adding up to more than 75%. As the United States has approximately 50% of the data centers, the energy of operation driven by computing is of bigger significance to the United States for its grid system and sustainability.

According to the IEA,[31] estimated global data center electricity consumption in 2022 was 460 terawatt-hours (TWh) or around 1–1.3% of global final electricity demand, excluding energy used for cryptocurrency mining (estimated to be 110 TWh in 2022) or for data transmission network energy use (estimated to range from 260 to 360 TWh in 2022). Using the value of 350 TWh (40 GW) for the data centers in 2024 (a lower estimate excluding crypto coin mining and data transmission), we have plotted in Figure 5(a) the energy of all the data centers and compared them to a few of the countries. The following observations highlight the trends. With our estimates (excluding several factors as indicated previously), *data centers consume more energy than all but a few countries*. For example, data centers consume more annual electricity than the continent of Australia or the entire United Kingdom. In Figure 5(b), we plot the data center energy estimate versus the annual electricity of the 50 states in the United States. The total energy requirements by the data centers *exceed all of the states except one*. Assuming that the United States hosts 50% of the data centers, the total energy of data centers installed in the United States exceeds the annual electricity production of all but five states.

In addition, in Figure 5(a), we have also included the energy required for running the highest performing 500 supercomputers under three conditions: 1) adding up the estimated total energy of the TOP 500 of the supercomputers (TOP500)[8]; 2) assuming that all the TOP500 have the same computing power as that of the first exaflop machine (20 MW); and 3) assuming that all of the TOP500 supercomputers have the same computing power as the most powerful supercomputer (~2 exaflops and maximum power of 60 MW). In the first case, the estimates were used based on reported specifications, while in the second and third cases, we used the same power for all of the 500 supercomputers. It appears that there are only a few countries that have enough electricity to host TOP500 supercomputers. In reality, these machines are distributed across the developed nations with the United States in the lead (about 25% of the top 10 computers were installed in 2022).

Even as the current significant energy requirements of large-scale installations are evident, the expected trends also need to be examined. According to IEA, the data center growth is even more exponential in China and Europe. By 2026, while the energy spent in the United States on data centers is estimated to increase by 1.5 times, these centers constitute an even larger fraction of the total annual electricity usage in other countries. For example, one-third of the electricity demand is estimated to be for data centers in Ireland, and one-fifth is estimated to be for data centers in Denmark. The growth of data centers is mainly driven by AI and crypto coin mining and could potentially be accelerated due to other new applications discussed previously.

The energy usage problem is expected to be multiplicative and termed the *3E+ effect*.[32] Energy used by computing leads to a multiplicative effect as the use of energy has collateral needs, leading to additional requirements. As computing is driven by electronics, the energy after computing is converted to heat, from one of the most efficient forms (electronic) to the least efficient form (heat). This is 1E. Consequently, the heat dissipated needs to be removed to enable the reliable processing of computing devices, which is 2E. Next, to install complex data centers and large computer systems, refrigeration and automation require additional electronic

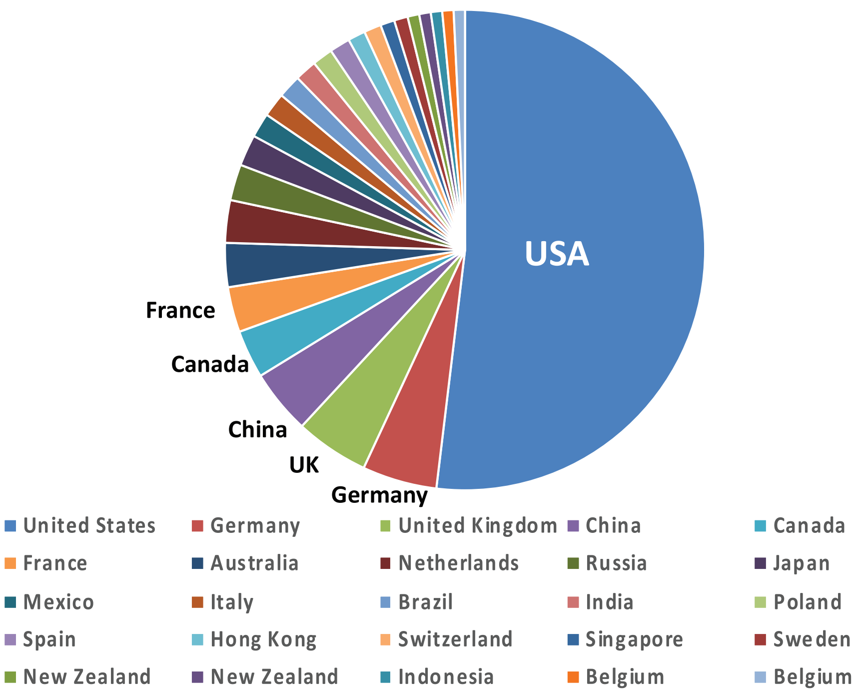

FIGURE 4. Data center numbers in different countries in 2024. (Source: Cloudscene, March 2024.)

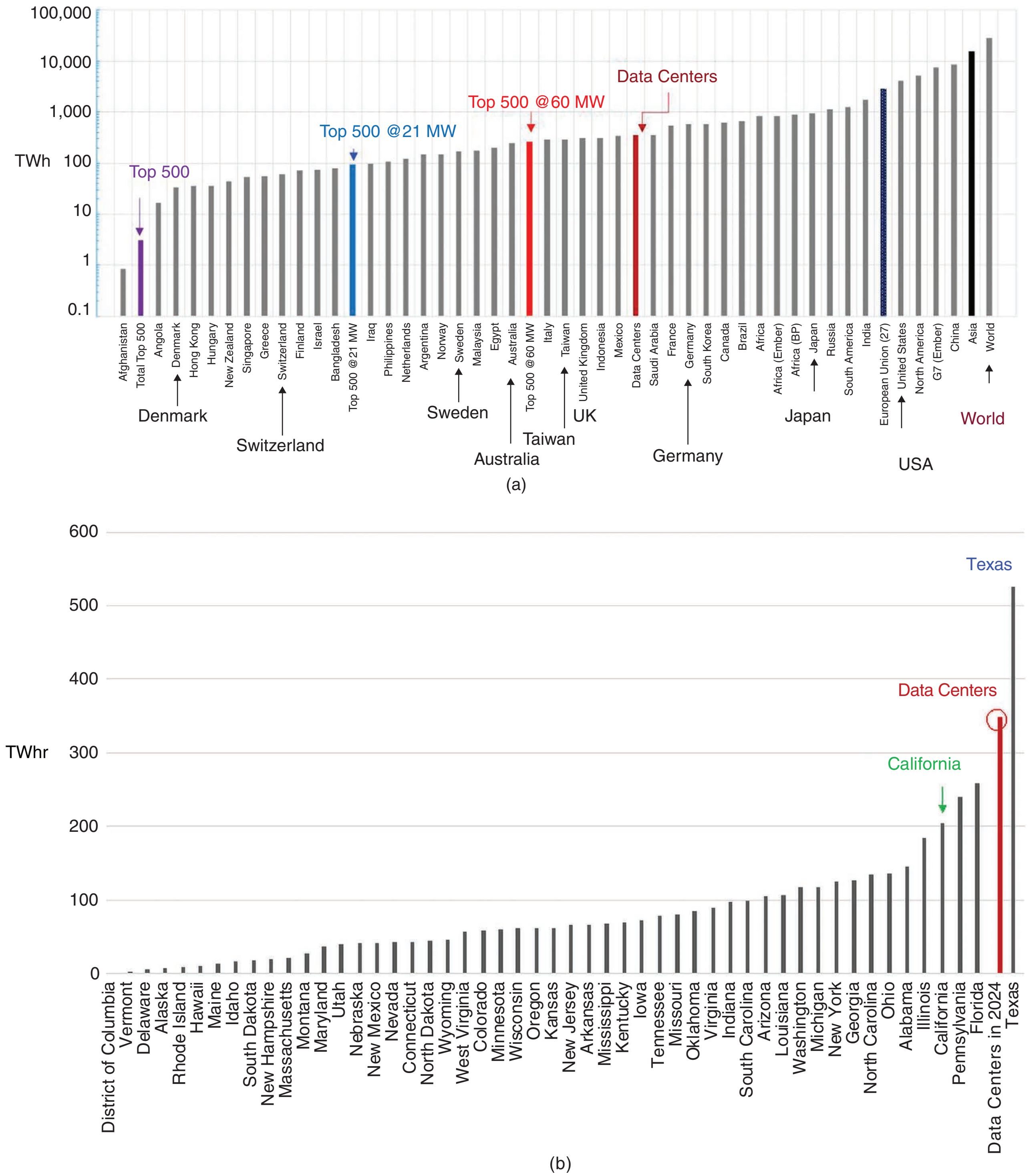

FIGURE 5. (a) The annual electricity use of several countries in the world compared with two scenarios of annual energy needed by Top500 supercomputers and data centers. (b) Annual electricity use in all the data centers compared to the annual electricity production in the fifty states in the United States. More electricity consumed by all states except Texas. Even with 50% of all data centers installed in the United States, total data center energy consumption exceeds all but five states. UK: United Kingdom; USA: United States.

components for operation, leading to 3E. Further, with the digitalization of all aspects of the economy, incorporation of renewable energy sources, and widespread use of AI, the electrical grid system with its own computing intelligence requires more energy for its computing needs, resulting in 3E+.

The estimates illustrated previously are likely to be conservative estimates (as we have not included communication and different specialized applications). The data centers also need other material resources, including metallic racks, cables, and weatherproof infrastructures. As an illustration, the Google data center water usage can vary from 0.4 million liters to 3.4 billion liters/year.[33] It is reported that up to 50% of the water can be recycled, indicating that significant resources are needed to replenish water lost during the cooling process.

As indicated previously, data centers serve as hubs for ubiquitous computing, in turn requiring significant amounts of energy and material resources. Although many companies continue to explore sustainable paths to energy and material usage, energy-efficient computing is an attractive option to help address the utilization itself without impacting computing usage, in contrast to thermal management, which exclusively addresses the effects. Given the 3E+ effect of computing, one unit of computing needing three or more units of energy for energy delivery, management, and heat removal can also have a multiplicative effect on data centers. In addition to the organizational efforts undertaken by several industrial organizations, the U.S. Department of Energy has initiated a road map for energy-efficient computing, encouraging efforts to double energy efficiency biennially, leading to a potential 1,000× increase in energy efficiency.[34] As has been discussed elsewhere,[7] reasons for these energy demands are driven by several factors:

1. Energy efficiency due to geometrical scaling is slowing down.
2. The energy reduction in transistors that is being facilitated by scaling is not getting translated to the energy used at the system level or at the simulation level.
3. The prodigious use of AI/ML applications and emerging new applications of computing across the economy is leading to significant energy demands.

Our analysis has indicated that the energy of computing can be roughly classified across three levels. The first is at the smallest level (for example, materials and transistors) as EPB; the second is the energy at the instruction level (for example, four-bit integer to 64-bit FPs or others depending upon the more advanced benchmarks) as EPI; and the third is the EPA (for example, energy to train one ML model or energy to simulate a single virion or energy to mine a single coin). It is clear that the energy of computing across all these layers needs to be made efficient for overall gains as the current trends appear unsustainable, especially in the context of data centers.

ACKNOWLEDGMENTS

This work was partially supported by the U.S. Department of Energy's Office of Science Contract DE-AC02-76SF00515 with SLAC through an Annual Operating Plan agreement WBS 2.1.0.86 from the Office of Energy Efficiency and Renewable Energy's Advanced Manufacturing and Materials Technology Office. The institutional support from the SLAC National Laboratory and the U.S. DOE's EES2 Working Groups are also acknowledged.

REFERENCES

1. S. K. Moore, "AMD CEO: The next challenge is energy efficiency a 500-megawatt supercomputer is 'probably too much'," *IEEE Spectr.*, Feb. 2023. [Online]. Available: https://spectrum.ieee.org/amd-eyes-supercomputer-efficiency-gains
2. "OpenAI CEO Altman says at Davos future AI depends on energy breakthrough," *Reuters*, Jan. 2024. [Online]. Available: https://www.reuters.com/technology/openai-ceo-altman-says-davos-future-ai-depends-energy-breakthrough-2024-01-16/
3. *The Wall Street J.* [Online]. Available: https://www.wsj.com/tech/ai/artificial-intelligences-insatiable-energy-needs-not-sustainable-arm-ceo-says-a11218c9
4. "Semiconductor: Supply chain deep dive assessment," U.S. Dept. Energy, Washington, DC, USA, Feb. 2022. [Online]. Available: https://www.energy.gov/sites/default/files/2022-02/Semiconductor%20Supply%20Chain%20Report%20-%20Final.pdf
5. "Energy of computing as a key design aspect for sustainability," IEEE SA, Piscataway, NJ, USA, Sep. 2023. [Online]. Available: https://sagroups.ieee.org/planetpositive2030/2023/09/09/energy-of-computing-as-a-key-design-aspect-for-sustainability/
6. N. Jones, "The information factories," *Nature*, vol. 561, no. 7722, pp. 163–166, 2018, doi: 10.1038/d41586-018-06610-y.
7. S. Shankar and A. Reuther, "Trends in energy estimates for computing in AI/machine learning accelerators, supercomputers, and compute-intensive applications," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Waltham, MA, USA, Piscataway, NJ, USA: IEEE Press, Sep. 2022, pp. 1–8, doi: 10.1109/HPEC55821.2022.9926296.
8. S. Shankar, "Energy estimates across layers of computing: From devices to large-scale applications in natural language processing, scientific computing, and cryptocurrency mining," submitted for publication.

9. J. Sevilla, "Parameter counts in machine learning: Public dataset and analysis of the evolution of parameter counts in machine learning," *Towards Data Sci.*, Jul. 2021. [Online]. Available: https://towardsdatascience.com/parameter-counts-in-machine-learning-a312dc4753d0
10. J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos, "Compute trends across three eras of machine learning," 2022, *arXiv:2202.05924*.
11. J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015, doi: 10.1126/science.aaa8685.
12. A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
13. OpenAI, "GPT-4 technical report," 2023, *arXiv:2303.08774*.
14. T. L. Scao et al., "Bloom: A 176 billion-parameter open-access multilingual language model," 2022, *arXiv:2211.05100*.
15. L. Casalino et al., "AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics," *Int. J. High Perform. Comput. Appl.*, vol. 35, no. 5, pp. 432–451, 2021, doi: 10.1177/10943420211006452.
16. "59th edition of the TOP500," TOP500, Sinsheim, Germany, Jun. 2022. [Online]. Available: https://www.top500.org/lists/top500/2022/06/
17. R. F. Barrett, T. H. F. Chan, E. F. D'Azevedo, E. F. Jaeger, K. Wong, and R. Y. Wong, "Complex version of high performance computing LINPACK benchmark (HPL)," *Concurrency Comput., Pract. Exp.*, vol. 22, no. 5, pp. 573–587, 2010, doi: 10.1002/cpe.1476.
18. "Climate and energy implications of crypto-assets in the United States," White House, 2022. Accessed: Oct. 21, 2022. [Online]. Available: https://www.whitehouse.gov/wp-content/uploads/2022/09/09-2022-Crypto-Assets-and-Climate-Report.pdf
19. "Documentation of the National Energy Modeling System (NEMS) modules," U.S. Dept. Energy, U.S. Energy Inf. Admin., Washington, DC, USA. [Online]. Available: https://www.eia.gov/outlooks/aeo/nems/documentation/
20. C. Kim, "The rise of ASICs: A step-by-step history of bitcoin mining," *CoinDesk*, Apr. 26, 2020. [Online]. Available: https://www.coindesk.com/tech/2020/04/26/the-rise-of-asics-a-step-by-step-history-of-bitcoin-mining/
21. M. Bondarev, "Energy consumption of bitcoin mining," *Int. J. Energy Econ. Policy*, vol. 10, no. 4, pp. 525–529, 2020, doi: 10.32479/ijeep.9276.
22. "The real-world costs of the digital race for bitcoin," *New York Times*, Apr. 9, 2023. [Online]. Available: https://www.nytimes.com/2023/04/09/business/bitcoin-mining-electricity-pollution.html#
23. "Bitcoin energy consumption index," *Digiconomist*, Nov. 2023. [Online]. Available: https://digiconomist.net/bitcoin-energy-consumption
24. M. Jaradat, M. Jarrah, A. Bousselham, Y. Jararweh, and M. Al-Ayyoub, "The Internet of Energy: Smart sensor networks and big data management for smart grid," *Procedia Comput. Sci.*, vol. 56, pp. 592–597, 2015, doi: 10.1016/j.procs.2015.07.250.
25. F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital Twin in Industry: State-of-the-art," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2405–2415, Apr. 2019. doi: 10.1109/TII.2018.2873186.
26. "What is a digital twin?" IBM. Accessed: Sep. 23, 2024. [Online]. Available: https://www.ibm.com/topics/what-is-a-digital-twin
27. "What is digital-twin technology?" *McKinsey & Company*, Aug. 26, 2024, pp. 1–6. [Online]. Available: https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-digital-twin-technology.
28. "Digital twin." Wikipedia. Accessed: Dec. 4, 2022. [Online]. Available: https://en.wikipedia.org/wiki/Digital_twin
29. "Revolutionizing manufacturing through mathematics: Illustrating the impact of the mathematical sciences," Nat. Acad., Washington, DC, USA, 2020. [Online]. Available: https://www.nationalacademies.org/event/03-03-2020/revolutionizing-manufacturing-through-mathematics-illustrating-the-impact-of-the-mathematical-sciences
30. "Source of data: Cloudscene." Statista. [Online]. Available: https://www.statista.com/statistics/1228433/data-centers-worldwide-by-country/
31. "Data centres and data transmission networks," IEA, Paris, France, Jan. 2024, p. 8. [Online]. Available: https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks
32. S. Shankar and T. Kaarsberg, "Energy of computing as a key design aspect for sustainability," IEEE SA, Piscataway, NJ, USA, Sep. 2023. [Online]. Available: https://sagroups.ieee.org/planetpositive2030/2023/09/09/energy-of-computing-as-a-key-design-aspect-for-sustainability/
33. M. Zhang. "Data center water usage: A comprehensive guide." Dgtl Infra. Accessed: Sep. 28, 2024. [Online]. Available: https://dgtlinfra.com/data-center-water-usage/
34. "Energy efficiency scaling for two decades research and development roadmap (Draft)," U.S. Dept. Energy, Washington, DC, USA, Aug. 14, 2024. [Online]. Available: https://www.energy.gov/media/330487

SADASIVAN (SADAS) SHANKAR is the research technology manager at SLAC National Laboratory, Menlo Park, CA USA, an adjunct professor at Stanford Materials Science and Engineering, Stanford, CA USA, and a lecturer at the Stanford Graduate School of Business, Stanford, CA USA. Contact him at sshankar@slac.stanford.edu or sadas.shankar@stanford.edu.