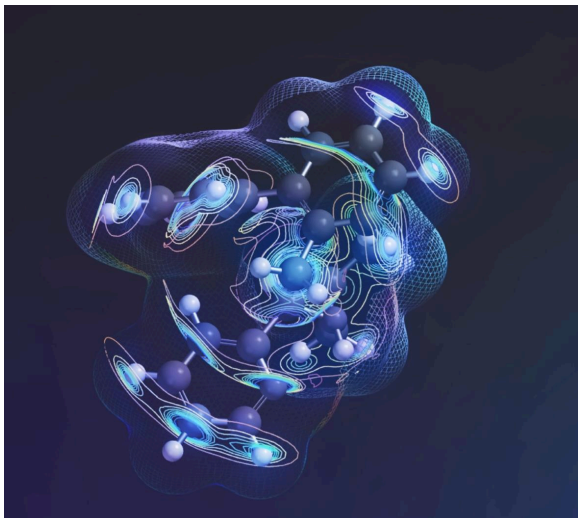
CASC | Coalition for Academic Scientific Computation

Energy Efficiency in Research Data Centers



Background

Research computing and data facilities today face unprecedented opportunities alongside complicated challenges. Opportunities stem from the explosion of high performance computing applications leveraging artificial intelligence (AI), machine learning (ML) and deep learning, promising groundbreaking discoveries and new ways of doing business. But AI and the large language models (LLMs) now making their way into research computing come with a cost. AI applications need to run fast and work best on graphic processing units (GPUs), which are built for multitasking. GPUs have thousands of compute cores [1] and are designed to run complex workloads in parallel.

The enormous growth of AI applications is due to the compute power of GPUs and the availability of very large data sets. Both CPU and GPU chips continue to pack more transistors into them and require more power. Future CPUs will require more than 500W of power, while GPUs will exceed a kilowatt, and both will require direct to chip liquid cooling to manage the heat load. Moreover, training AI models takes time and effort; often several tries are needed to create a good model. Meta recently reported that it required nearly 32 million GPU hours across more than 16,000 GPUs

Contact Us kelley@casc.org | (202) 670-5798 | casc.org

to train Llama 3.1, during which they experienced 466 hardware failures that interrupted the training. This effort consumed more than 22 million kilowatt-hours, or enough electricity to power more than 2,300 homes in the US for a full year [2].

To complicate matters even more, AI and the use of GPUs comes as the world struggles to mitigate the impacts of climate change. Data IT and data centers are responsible for 4% of the nation's energy use (3% worldwide) and that figure is expected to rise to 6% by 2026[3]. This means data centers must balance the need for more capable AI-enabled infrastructure with the need to manage the environmental impact of research computing. Innovative new architectures have been developed to address the problem, but technologists and researchers need access and education to use them effectively. ***This paper highlights the main avenues to energy savings in information technology and computing infrastructure and suggests actions that can be taken to create research data centers capable of handling the most demanding problems, while meeting energy efficiency requirements at all stages of computing workflows and reducing their carbon footprint.***

Challenges

Cooling: HPC data centers generate significant heat, necessitating extensive cooling from air conditioning and water and leading to substantial energy costs at a time when energy prices are rising. AI and ML applications exacerbate these requirements, as superfast GPUs consume more energy and generate more heat than CPUs, making it challenging for data centers to manage costs and achieve carbon neutrality. Liquid cooling, where a liquid absorbs and dissipates heat from data center components, allows for more precise temperature control and closer component packing, which can mean greater compute capacity within the same carbon footprint. However, liquid cooling means a considerable investment in the system itself (pipes, pumps, liquid handling units, etc.) and in new expertise for monitoring leaks and performing maintenance.

Hardware: Hardware breakthroughs have helped data centers become more efficient as the number of research data centers grows and interest in AI accelerates. Among the innovations that address energy use are high bandwidth memory (HBM) chips, which reduce the power needed to transfer data between memory and processor, extend battery life and reduce power consumption [4, 5]; wafer-scale processors that greatly speed up the process of

developing ML/AI models [6]; and data processing units (DPUs), which can be used to offload key functions and reduce server power consumption [7]. However, if these innovations are to have an impact, data centers operators must implement them and software developers must adopt them. That requires investments not only in hardware but in human resources through technical training and ongoing user support.

Software: As HPC has become more heterogeneous, many application frameworks have become more portable. Yet many – including modeling systems that utilize AI – are far from being mature and universally available. A key challenge to creating energy efficient, sustainable computing environments is helping end users and computational scientists author applications that aim for energy efficiency and consider the resources needed for a job in addition to speed and performance, which has been the traditional focus of developers. In turn, data center operators must consider how those applications use resources and their energy demands if they are to make the most informed decisions about energy use in their data centers.

Data Security: The rapid growth of generative AI means large datasets are needed for training – often containing sensitive details – and that poses more challenges for data centers. Privacy breaches can be devastating and governments are strengthening data protection protocols as AI proliferates. Although outsourcing data-intensive computing to the cloud or a colocated data center can save costs and reduce an institution-owned-and-operated data center's energy footprint, it can also mean less control of security procedures.

Actions: Creating an energy-aware data center ecosystem

How can data centers adapt to a new age in computing and computing technologies?

The hybrid approach. Some research computing and data (RCD) centers have adopted hybrid solutions, where technologies such as *virtualization, cloud* and *software-defined networking* mean greater flexibility with the same or less hardware and thus lower energy use. *Optimized use of equipment* also reduces the manufacturing carbon footprint per delivered computing service and translates into lower costs for operation per service. Using the cloud for AI workloads, for example, saves users of the AI applications

the effort of building, deploying and maintaining their own GPU clusters, which for many workloads could lower costs. However, cloud-based computing comes with its own drawbacks. While cloud computing can mean more flexibility and scalability without more local energy use, it also means less control of the overall infrastructure and less ability to customize for specific uses. Many IT service providers take the hybrid approach of offering users cloud-based services while maintaining their own compute clusters. The ability to send complex workloads to the cloud gives service providers the flexibility to handle complex AI and big data computing problems without having to worry about extra heat and cooling.

Hardware efficiency. On the hardware side, systems that run on GPUs are more energy efficient than CPU-based systems, but the size of the problems they handle, including training AI models using massive amounts of data, mean more heat released and more water used for cooling. MIT researchers have shown that modifying hardware using *dual-rail logic circuit design*, a type of digital logic circuit that uses two complementary signals, and recycling the extra bits from each signal can cut energy use and heat loss in half, although some of that savings is lost because of the need for more wires and transistors [8]. Power tradeoffs among the components of an HPC system can also achieve energy efficiency. These include *dynamic power management* (DPM), *dynamic voltage and frequency scaling* (DVFS), and *power capping mechanisms* [9].

Software efficiency. As hardware becomes more energy efficient, the impact of software on overall energy consumption becomes significant. To achieve greener software, "we need to consider energy efficiency and sustainability of software as important parameters, in addition to functionality, security, scalability and accessibility," writes the Green Software Foundation, an industry collaboration aimed at building a more efficient software ecosystem [10]. Moreover, says the group, software development must aim for *reuse, extended longevity of use*, and the most minimal computational and memory resource requirements possible. *Energy profiling* of applications at all phases of the software development lifecycle, using energy adaptive APIs, and careful resource management within the software stack to minimize energy consumption are all techniques that software engineers can use to become more aware of energy consumption and develop energy aware software [11].

Training and education. Any move toward greener data centers requires awareness – on the part of data center operators, software developers and domain users of HPC systems – and an emphasis on

training and education. Numerous organizations, including the U.S. Department of Energy, universities, and corporations including Intel and Nvidia offer a variety of classes and certificate programs on data center and computing infrastructure energy efficiency. But these trainings vary based on who presents them and their ideas of best practices. ***As AI/ML/DL continues to proliferate in data centers, CASC can help develop best practices and use cases for sustainable, energy aware, data center operations and energy efficient/aware software development practices.***

Key Takeaways

Innovation is key. As AI and GPU processors increase the heat generated by data centers and the need for cooling, research computing facilities must be flexible and ready to change. Both hardware innovations and new software development practices must be used to reach the goal of the carbon neutral data center.

Hardware must adapt to the needs of AI. As AI applications become more common in research computing data centers,, those centers must adapt. Modifications, better power management, liquid cooling, and hybrid solutions that offload work to the cloud must be explored. RCD center technical staff as well as vendors must become knowledgeable on building, deploying and operating energy efficient centers.

Software development practices must be updated. Energy usage and energy efficiency must be considered on par with performance in every stage of the software development lifecycle. Moreover, metrics must be established to measure the cost effectiveness and energy efficiency of software at all stages in the development process.

Use cases and best practices must be developed and shared. CASC can use its network of more than 105 members to educate RCD center staff and students on best practices, and inform administrators and government officials on the best path forward to carbon neutral data centers that are able to handle the workloads of AI-enabled high-performance computing. CASC members can implement innovative approaches as examples of what can be done and develop use cases from those innovations. In turn, faculty can teach students, who are the next generation of RCD center leaders, to implement innovative new approaches.

Authors: *Karen Green; Carolyn Casler, CASC; Eric Coulter, Georgia Institute of Technology; Erik Deumens, University of Florida; Ed Hanna, Pittsburgh Supercomputing Center; Aaron Jezghani, Georgia Institute of Technology; Robert Kalescky, Southern Methodist University; Kathryn Kelley, CASC; Glen MacLachlan, George Washington University; Scott Michael, Indiana University; Dan Stanzione, Texas Advanced Computing Center*

About CASC

The Coalition for Academic Scientific Computation is an educational nonprofit 501(c)(3) organization with 105+ member institutions representing many of the nation's most forward-thinking universities and computing centers. CASC is dedicated to advocating for the use of the most advanced computing technology to accelerate scientific discovery for national competitiveness, global security, and economic success, as well as develop a diverse and well-prepared 21st century workforce. Learn more at http://casc.org.

References

[1] AceCloud: NVIDIA CUDA Cores Explained: How Are They Different? https://acecloud.ai/resources/blog/nvidia-cuda-cores-explained/#:~:text=While%20a%20CPU%20has%20a,40%25%20over%20the%20previous%20generation.

[2] Introducing Llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/

[3] Zahn, M. (April 20, 2024). Data centers fuel AI and crypto but could threaten climate, experts say. ABC News. https://abcnews.go.com/Business/data-centers-fuel-ai-crypto-threaten-climate-experts/story?id=109342525.

[4] Intel Corporation.(2022). Technical Overview Of The Intel® Xeon® Scalable processor Max Series. https://www.intel.com/content/www/us/en/developer/articles/technical/xeon-scalable-processor-max-series.html.

DOI: 10.13140/RG.2.2.25923.41767

---

[5] Simms. (February 5, 2024). What is HBM (High Bandwidth Memory? *https://www.simms.co.uk/tech-talk/what-is-hbm-high-bandwidth-memory/#:~:text=Lower%20power%20consumption%3A%20Its%20simply,benefits%20for%20this%20exact%20reasoning.*

[6] Lavely A. (2022). Powering Extreme-Scale HPC with Cerebras Wafer-Scale Accelerators. *https://8968533.fs1.hubspotusercontent-na1.net/hubfs/8968533/Powering-Extreme-Scale-HPC-with-Cerebras.pdf*

[7] NVIDIA. (2022). DPU Power Efficiency. *https://resources.nvidia.com/en-us-accelerated-networking-resource-library/nvidia-dpu-power-efficiency-white-paper.*

[8] Stauffer, N. (2013). Energy Efficient Computing: Enabling smaller, lighter, faster computers. MIT Energy Initiative. *https://energy.mit.edu/news/energy-efficient-computing/.*

[9] Kocot, B., Czarnul, P., and Proficz, J. (2023). Energy-Aware Scheduling for High-Performance Computing Systems: A Survey. *https://doi.org/10.3390/en16020890.*

[10] Green Software Foundation (2024). 10 Recommendations for Green Software Development. 2024. *https://greensoftware.foundation/articles/10-recommendations-for-green-software-development*

[11] Pinto, G., and Castor, F. (2017). Energy Efficiency: A New Concern for Application Software Developers. CACM2017.https://gustavopinto.org/lost+found/cacm2017.pdf