Datacenter power and energy management: past, present, and future

Ricardo Bianchini, Microsoft Azure

Christian Belady, Microsoft Cloud Operations and Innovation

Anand Sivasubramaniam, Penn State University

Abstract

This paper overviews some of the key past developments in cloud datacenter power and energy management, where we are today, and what the future could be. This topic is gaining enormous, renewed interest in the context of the conflicting needs of the AI revolution and the climate crisis.

Keywords: Datacenters, power, energy, sustainability.

Introduction

Datacenters underly nearly all aspects of modern society as our lives become ever more reliant on online services, like banking, e-commerce, social networks, search, gaming, cloud computing, and Artificial Intelligence (AI). In this context, datacenter power and energy management has never been more important to service providers, like Amazon, Microsoft, Meta, and Google, and society at large due to the combination of two major challenges: the climate crisis and the recent unprecedented expansion in datacenter demand due to AI. Both power and energy are critical: (peak) power draw drives datacenter designs, construction costs, and embedded carbon emissions, whereas energy (or average power over time) translates into electricity costs and electricity generation emissions. The management of datacenter power and energy involves actively modulating power draw, eliminating inefficiencies, or introducing optimizations in software, hardware, and physical infrastructure.

This paper provides a retrospective of some of the key advances in this space over its more than 2 decades, from its infancy in the early 2000s until today. It also looks forward, highlighting key advances that we believe will be required for the AI era that started recently. Instead of attempting to be comprehensive, *we focus on the advances in software, hardware, and physical infrastructure (especially power delivery and cooling) that had the most impact on production cloud datacenters*. Academic research has produced equally important efforts in software for server and cluster-wide power and energy management, sub-thresholding transistors, new power states for hardware components, hardware heterogeneity for power-performance tradeoffs, battery-based power re-shaping, locating datacenters based on power/cooling/carbon considerations, and geographic load distribution across datacenters, to name a few; but many of these ideas have not yet been deployed at scale. Ultimately, our goal is to entice more researchers and practitioners to join the effort to address the massive challenges that industry faces in this space.

Early days: learning to build cloud datacenters (2000s)

Up until 2000, all the published work on power and energy management focused on laptops and other battery-operated devices; cloud datacenters had not yet started to proliferate and face scaling and operational challenges, whereas servers were "connected to the wall" so were not viewed as a problem. In 2001, two academic papers [1, 2] introduced the notion of managing datacenter energy in software based on the diurnal pattern of the offered load. Because servers consumed a large fraction of their maximum power while idle, the papers proposed to concentrate the load on a smaller number of servers during periods of light load, and shut the idle ones down entirely. To this day, shutting servers down is only used in production during emergencies or other rare scenarios for multiple reasons, including reliability concerns and the overhead of bringing complex services back up in an orderly way.

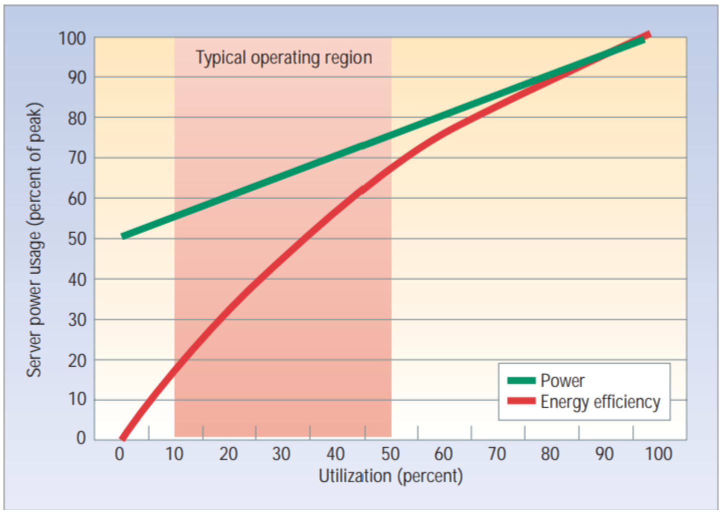

Figure 1: Server with poor power proportionality. The server's power draw is a large percentage of its peak power even at 0% utilization, which is highly inefficient especially when servers run at low utilization. Energy efficiency was computed by dividing utilization by power. Reproduced from [3].

Later papers [3, 4] articulated the poor proportionality of server hardware power draw as a function of utilization (Figure 1, reproduced from [3]), which led to many efforts to improve it. These works were part of the impetus for the Energy Star program in the US to include energy consumption requirements for servers. Server CPU vendors introduced new *idle* low-power states that lowered idle energy consumption substantially. Unfortunately, other server components (e.g., memories, storage, network interfaces) have not seen similar improvements. Recognizing the persistent low resource utilization with short idle times in production datacenters, a different approach advocated for *active* low-power modes that allowed components to continue to operate (while at low utilization) but at a lower performance and power draw [5, 6, 7]. Again, this idea has been adopted for server CPUs (dynamic voltage and frequency scaling) but has not yet been as successful for other components. We encourage hardware vendors to revisit providing such modes and bring GPUs and other AI accelerators on-par with the power management capabilities of CPUs.

At the same time as these works on datacenter energy management, it became clear that power management was an even more important factor at scale [8, 9], especially given the highly skewed distribution of power draw in production datacenters where near-peak draws happen rarely if ever (Figure 2, reproduced from [9]). Any power that is rarely used can be oversubscribed, i.e., allocated for hosting more servers in the datacenters, as long as power can be quickly capped during those rare times. Hosting more servers via power oversubscription can reduce costs and emissions tremendously, as the provider better utilizes its existing datacenters instead of building new ones. The findings from this decade have persisted. Today, oversubscribing power allocations via hierarchical power capping has become a staple in production cloud datacenters [10, 11, 12]. Interestingly, since servers are still more expensive than power, an even better approach is to increase server utilization, leaving less room for power oversubscription. Today, resource oversubscription, resource harvesting, and workload co-location are used pervasively for this purpose.

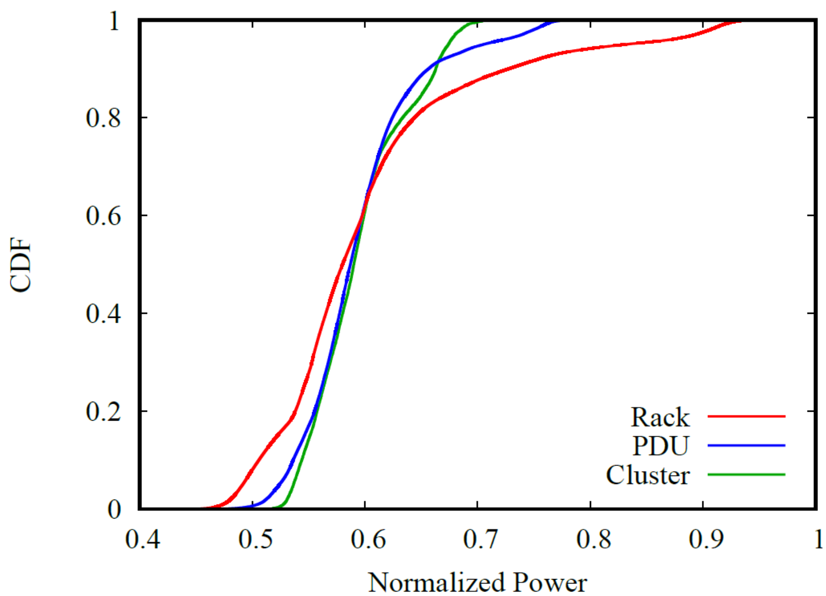

Figure 2: Cumulative Distribution Function of power draw of a cloud datacenter aggregated at the rack, power delivery unit, and cluster. The coarser the aggregation, the more power is available for oversubscription. Reproduced from [9].

Concurrently with these advances in software and hardware, providers made tremendous strides towards improving the efficiency of the physical infrastructure in terms of the power delivery hierarchy and cooling systems. A key driver for these improvements was the creation of the Power Usage Efficiency (PUE) metric. PUE quantifies the power overhead beyond what the IT hardware requires by dividing the full datacenter power by the IT power; the closer PUE is to 1, the more efficient the datacenter. As a result of the focus on PUE, providers were able to reduce power delivery losses to a few percent (e.g., by reducing the number of power conversions under normal operation), while creating more efficient cooling systems (e.g., by bringing outside air into the datacenter for cooling in addition to the earlier hot-aisle/cold-aisle layout for server racks). Over this period, cloud datacenter PUEs improved from around ~1.25 to ~1.1 on average. PUE awareness also made locating datacenters based on cooling costs, along with power and latency goals, equally important – whether it be close to water sources or in climatic conditions where using outside air for cooling would be effective. It has also led smaller datacenter operators to consolidate their IT resources and smaller machine rooms into larger facilities where their scale offered better efficiency. Though PUE as a metric has its limitations (e.g., it does not account for server fans as an overhead), it is still used today and will likely continue to be useful going forward.

Maturity: the cloud takes off (~2010s)

During the 2010-2020 decade, cloud computing exploded in popularity, driven mostly by Amazon Web Services, Microsoft Azure, and Google Cloud Platform. The migration of workloads from on-premise datacenters to run virtualized in the cloud continues to have positive power and energy implications as cloud datacenters are more efficient than on-premise ones. However, concerns about the climate crisis and datacenter sustainability became more widespread and, towards the end of the decade, online service providers started making pledges to limit their carbon footprint and leverage more renewable energy. The cloud's growth and sustainability concerns fueled further innovations in power and energy management, which at this stage was reaching its maturity phase.

In the software space, the notion of software-defined hardware and infrastructure was introduced. Software-defined power and energy led software to (1) take control of power and energy management functions that used to be done by the hardware or the physical infrastructure, and (2) manage the power and energy usage across the three tiers in a more holistic and workload-aware manner [11, 12]. For example, Microsoft built software systems that maximize the allocation of power in a datacenter, by leveraging all its redundant power during normal operation and limiting power draw when a power emergency or a planned maintenance occurs [13]. To lower its carbon footprint, Google used workload knowledge to shift load to periods of high renewable energy [14]. Along similar lines, providers introduced AI systems responsible for enabling efficiency improvements and/or energy management. For example, Google announced its use of AI to control the cooling knobs in its datacenters [15], whereas Microsoft introduced AI for resource management into the Azure control plane [16].

On the hardware side, the decade saw the impact of the end of Dennard scaling (which happened around the middle of the prior decade) and the slowdown of Moore's law (which started in this decade and we are still witnessing). These factors meant that more servers would be needed to service the same load, and the increases in transistors in the upcoming server generations would not benefit from commensurate reductions in power draw. As a result, hardware accelerators became the best way to improve efficiency (Figure 3, reproduced from [17]). For example, Microsoft introduced its first accelerator (Catapult [18], for efficient Web search) in the first half of the decade, later Google introduced its Tensor Processing Units [19] for efficient AI acceleration, and Amazon introduced its Nitro host virtualization accelerator [20]. Today, accelerators for many software functions continue to be built, including compression, encryption, and video transcoding.

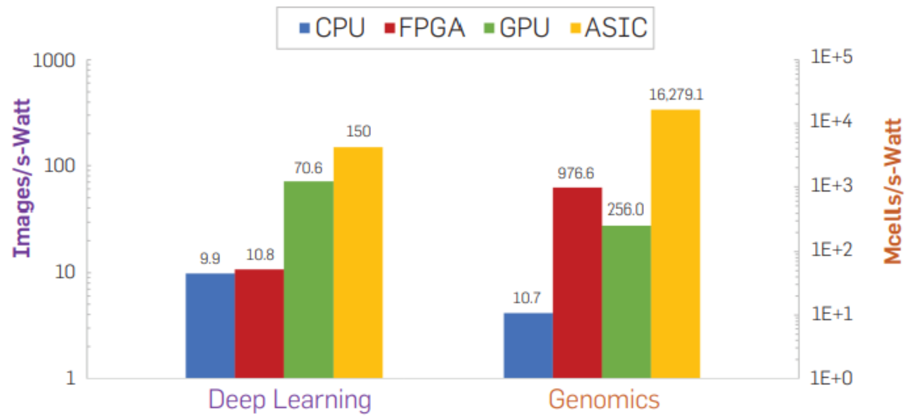

Figure 3: Computation efficiency in tasks/s-Watt for CPU, FPGA, GPU, and ASIC for deep learning (Resnet-50 inference) and genomics (Banded Smith-Waterman algorithm). Accelerators are substantially more efficient than CPUs. Reproduced from [17].

There were also advances in the physical infrastructure from more efficient power delivery topologies to liquid cooling. For example, Google created a large medium-voltage power plane (tens of MW) with a generator farm to enable greater power oversubscription and lower costs [11], whereas Microsoft went from block-redundant to distributed-redundant power topology [13]. In maximizing the allocation of this redundant power, Microsoft leveraged earlier ideas on leveraging energy storage to shave transient peaks in power draw [21]. In the cooling space, high power densities started to challenge the limits of air cooling, leading Google to use cold plate-based liquid cooling for its TPUs starting in 2018. Though liquid cooling may incur higher capital costs, it can also reduce energy consumption because (1) water is a better heat transfer medium than air and (2) many server fans can be eliminated. Today's GPUs and other AI accelerators are exacerbating the power density problem, so liquid cooling will likely become widespread across providers.

Future: the AI era (~2020s+)

The beginning of the 2020 decade saw trends and technologies from the previous decade accelerate. However, two additional major disruptions are heavily influencing power and energy management: (1) the massive expansion in computational and power demand that AI systems are prompting; and (2) the increasing constraints that local communities and existing electrical grids are imposing on new datacenter development. The computational needs of AI systems exploded when large language models like ChatGPT and Bard gained enormous adoption right after being made available to the public. This caused OpenAI/Microsoft, Google, and AWS to start a race to train increasingly large and more capable models, leading to a corresponding massive computational demand for model inference. Compounding this trend, both training and inference rely on GPUs and other AI accelerators that can draw more power than CPUs by multiple fold. In many cases, the race is demanding more power faster than electrical grids can accommodate. As a rough approximation, deploying new power capacity might lag demand by 5-10 years, the time it takes to get permits, build large power plants, and possibly lay out new transmission lines. Not to mention the enormous capital demands of such a large and accelerated expansion. Large cloud providers will need to help provision the extra power supply, while reducing the environmental impact of AI [22]. They will also be asked to interact more closely with power utilities given their large amount of dispatchable power and energy storage "behind the meter".

At the same time, we now see environmentally-conscious localities imposing limits on new datacenter construction and on the use of generators powered by fossil fuels. Electrical grids have also been challenged by the enormous power needs of AI-capable datacenters, the power usage characteristics of large AI training workloads, and the greater penetration of variable-availability renewable power (e.g., solar and wind). In fact, as cloud providers pledged to become carbon-negative until the end of the decade, efforts to procure and integrate enough renewable power into the electrical grid (for both running AI workloads and reducing the carbon intensity of the providers' supply chains) are accelerating as well.

The combination of these factors is leading to a flurry of innovation in physical infrastructure, hardware, and software. In terms of infrastructure, we will see liquid cooling become pervasive even in cloud providers; first, by cooling GPUs and other AI accelerators (as had already started happening in the prior decade), and later cooling general-purpose compute servers, as they become increasingly power-hungry and air cooling becomes less economical and efficient. In this decade, two-phase immersion cooling showed promise to greatly reduce cooling power and energy (as well as engineering costs) but started facing regulatory challenges, so facility water piped to cold plates will likely be the main liquid cooling technology in the short and medium terms. We will also see electrical grids and datacenter infrastructure address the large power swings involved in AI training. To disperse the need for AI inference power and energy and reduce latencies, we might see inference move to the edge of the network (for large models) or consumer devices (for small ones). Along similar lines, power generation might become more distributed, reducing the need for massive plants and more transmission capacity. Further in the future, we would love to see datacenters that are more "ecologically-integrated", i.e., better integrated into and contributing to the environment around them. For example, think of a datacenter providing a massive green space in its roof for a park, using renewable sources for its power and rain for its cooling, providing heat for community spaces, etc.

On the software side, we will likely see an expansion of the software-defined concept to performance and liquid cooling. Software-defined performance takes control of frequency increases from hardware techniques like Intel's Turbo Boost, providing additional performance only to the most critical workloads (e.g., from customers paying a premium for the extra performance) that can meaningfully benefit (e.g., they are compute-bound). This approach is more efficient as it only consumes additional energy when it can be highly effective for the provider and its customers. Software-defined liquid cooling could leverage software control (e.g., load balancing, load prediction, management of leaks) for reducing peak and average PUEs, and increasing availability. With the coming increase in power density, especially in AI training datacenters where GPUs must communicate frequently and synchronously, cooling failures will need to be tackled quickly before the equipment overheats. We are also already seeing software-driven power oversubscription for AI-capable datacenters and their workloads [23]. Power- and energy-aware scheduling and batching of AI inference will likely come next, whereas delay-insensitive AI training could enable matching of computation to renewable energy availability. At a coarser grain, we will likely see software (and physical infrastructure) help electrical grids with high renewable penetration stabilize and regulate themselves, by increasing or decreasing datacenter power draw on demand. On all these fronts, we expect that AI will continue to play a major role in helping manage power and energy.

In terms of hardware, the trend towards increasingly efficient accelerators will continue, especially in the AI space. We are also likely to see GPUs and other AI accelerators start to integrate similar power management mechanisms as CPUs, like fine-grained frequency control, power capping, and idle power states (all under software control), which will enable greater efficiency and sustainability. Interestingly, custom accelerators might also have a negative impact on sustainability, as they might increase embodied carbon emissions if they become obsolete and are replaced too quickly. On the cooling front, we might see new liquid cooling technologies that can extract a greater amount of heat per cm^2, for example, by bringing the liquid closer to silicon [24]. The future might also be bright in cloud datacenters for technologies with a strong impact on power and energy, like 3D die stacking and silicon photonics.

Finally, we hope that this will be the decade when cross-layer power and energy management finally becomes a reality at scale, with physical infrastructure (datacenter, electrical grid, and possibly power plant), hardware, and software all working together to achieve maximum efficiency and sustainability. Coordinated optimization across these layers is a great opportunity for innovation.

Conclusion

Over the last 2.5 decades, cloud datacenter power and energy management has come a long way: a large amount of excellent research and practice introduced many ideas and innovations in physical infrastructure, hardware, and software. Many of them have had a significant impact in production, increasing efficiency and sustainability, and reducing costs. Lately, we have been faced with enormous disruptions and challenges brought about mainly by the conflicting needs of the AI revolution and the climate crisis. In light of them, we believe that datacenter power and energy management has never been as critical and as ripe for innovation; but we need all hands on deck from researchers to practitioners, from academia to industry. Industry should continue to host graduate and undergraduate interns in datacenter power, energy, and sustainability, and publish traces (e.g., https://github.com/Azure/AzurePublicDataset) and papers around the main challenges it faces. For their part, academics could leverage the industry data and hosting of interns to renew their focus on these areas, produce more graduates with expertise in them, and pursue ambitious ideas that might be too risky for industry at first. Let us work together!

This paper is dedicated to the memory of our friend and colleague, Luiz André Barroso.

References

[1] Pinheiro, Eduardo, Ricardo Bianchini, Enrique V. Carrera, and Taliver Heath. "Load balancing and unbalancing for power and performance in cluster-based systems." In *Proceedings of the Workshop on Compilers and Operating Systems for Low Power*, vol. 180, pp. 182-195. 2001.

[2] Chase, Jeffrey S., Darrell C. Anderson, Prachi N. Thakar, Amin M. Vahdat, and Ronald P. Doyle. “Managing energy and server resources in hosting centers.” *ACM SIGOPS Operating Systems Review* 35, no. 5 (2001): 103-116.

[3] Barroso, Luiz André and Urs Hölzle. “The case for energy-proportional computing.” *Computer* 40, no. 12 (2007): 33-37.

[4] Meisner, David, Brian T. Gold, and Thomas F. Wenisch. “Powernap: eliminating server idle power.” *ACM SIGARCH Computer Architecture News* 37, no. 1 (2009): 205-216.

[5] Carrera, Enrique V., Eduardo Pinheiro, and Ricardo Bianchini. "Conserving disk energy in network servers." In *Proceedings of the 17th International Conference on Supercomputing*, pp. 86-97. 2003.

[6] Meisner, David, Christopher M. Sadler, Luiz André Barroso, Wolf-Dietrich Weber, and Thomas F. Wenisch. “Power management of online data-intensive services.” In *Proceedings of the 38th International Symposium on Computer Architecture*, pp. 319-330. 2011.

[7] Deng, Qingyuan, David Meisner, Luiz Ramos, Thomas F. Wenisch, and Ricardo Bianchini. “Memscale: active low-power modes for main memory.” *ACM SIGPLAN Notices* 46, no. 3 (2011): 225-238.

[8] Ranganathan, Parthasarathy, Phil Leech, David Irwin, and Jeffrey Chase. “Ensemble-level power management for dense blade servers.” *ACM SIGARCH Computer Architecture News 34*, no. 2 (2006): 66-77.

[9] Fan, Xiaobo, Wolf-Dietrich Weber, and Luiz André Barroso. “Power provisioning for a warehouse-sized computer.” *ACM SIGARCH Computer Architecture News* 35, no. 2 (2007): 13-23.

[10] Wu, Qiang, Qingyuan Deng, Lakshmi Ganesh, Chang-Hong Hsu, Yun Jin, Sanjeev Kumar, Bin Li, Justin Meza, and Yee Jiun Song. “Dynamo: Facebook’s data center-wide power management system.” *ACM SIGARCH Computer Architecture News* 44, no. 3 (2016): 469-480.

[11] Sakalkar, Varun, Vasileios Kontorinis, David Landhuis, Shaohong Li, Darren De Ronde, Thomas Blooming, Anand Ramesh, James Kennedy, Christopher Malone, Jimmy Clidaras, and Parthasarathy Ranganathan. “Data center power oversubscription with a medium voltage power plane and priority-aware capping.” In *Proceedings of the 25th International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 497-511. 2020.

[12] Kumbhare, Alok Gautam, Reza Azimi, Ioannis Manousakis, Anand Bonde, Felipe Frujeri, Nithish Mahalingam, Pulkit A. Misra, Seyyed Ahmad Javadi, Bianca Schroeder, Marcus Fontoura, and Ricardo Bianchini. “Prediction-based power oversubscription in cloud platforms.” In *Proceedings of the USENIX Annual Technical Conference*, pp. 473-487. 2021.

[13] Zhang, Chaojie, Alok Gautam Kumbhare, Ioannis Manousakis, Deli Zhang, Pulkit A. Misra, Rod Assis, Kyle Woolcock, Nithish Mahalingam, Brijesh Warrier, David Gauthier, Lalu Kunnath, Steve Solomon, Osvaldo Morales, Marcus Fontoura, and Ricardo Bianchini. "Flex: High-availability datacenters with zero reserved power." In Proceedings of the *48th International Symposium on Computer Architecture*, pp. 319-332. 2021.

[14] Radovanovic, Ana. “Our data centers now work harder when the sun shines and wind blows.” Google Data Centers and Infrastructure blog. 2020.

[15] Evans, Richard and Jim Gao. “DeepMind AI reduces Google data centre cooling bill by 40%.” Google DeepMind blog. 2016.

[16] Cortez, Eli, Anand Bonde, Alexandre Muzio, Mark Russinovich, Marcus Fontoura, and Ricardo Bianchini. “Resource Central: Understanding and predicting workloads for improved resource management in large cloud platforms.” In *Proceedings of the 26th Symposium on Operating Systems Principles*, pp. 153-167. 2017.

[17] Dally, William J., Yatish Turakhia, and Song Han. “Domain-specific hardware accelerators.” Communications of the ACM, vol. 63, no. 7 (2020): 48-57.

[18] Putnam, Andrew, Adrian M. Caulfield, Eric S. Chung, Derek Chiou, Kypros Constantinides, John Demme, Hadi Esmaeilzadeh, Jeremy Fowers, Gopi Prashanth Gopal, Jan Gray, Michael Haselman, Scott Hauck, Stephen Heil, Amir Hormati, Joo-Young Kim, Sitaram Lanka, James Larus, Eric Peterson, Simon Pope, Aaron Smith, Jason Thong, Phillip Yi Xiao, and Doug Burger. "A reconfigurable fabric for accelerating large-scale datacenter services." *ACM SIGARCH Computer Architecture News* 42, no. 3 (2014): 13-24.

[19] Jouppi, Norman P., Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon. "In-datacenter performance analysis of a tensor processing unit." In *Proceedings of the 44th International Symposium on Computer Architecture*, pp. 1-12. 2017.

[20] AWS. “AWS Nitro system”. Re:Invent. 2018.

[21] Govindan, Sriram, Anand Sivasubramaniam, and Bhuvan Urgaonkar. "Benefits and limitations of tapping into stored energy for datacenters." In *Proceedings of the 38th International Symposium on Computer Architecture*, pp. 341-352. 2011.

[22] Wu, Carole-Jean, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. "Sustainable AI: Environmental implications, challenges and opportunities." *Proceedings of Machine Learning and Systems* 4 (2022): 795-813.

[23] Patel, Pratyush, Esha Choukse, Chaojie Zhang, Íñigo Goiri, Brijesh Warrier, Nithish Mahalingam, and Ricardo Bianchini. "Characterizing power management opportunities for LLMs in the cloud." In *Proceedings of the 25th International Conference on Architectural Support for Programming Languages and Operating Systems*. 2024.

[24] Rajan, Sreejith Kochupurackal, Bharath Ramakrishnan, Husam Alissa, Washington Kim, Christian Belady, and Muhannad S. Bakir. "Integrated silicon microfluidic cooling of a high-power overclocked CPU for efficient thermal management." *IEEE Access* 10 (2022): 59259-59269.

Short bios

Ricardo Bianchini is a technical fellow and corporate vice president at Microsoft Azure, Redmond, WA, 98052, USA. His research interests include data center power and energy management, data center sustainability, and cloud computing. Bianchini received his Ph.D. degree in computer science from the University of Rochester. He is a fellow of the Association for Computing Machinery and a Fellow of IEEE. Contact him at ricardob@microsoft.com.

Christian Belady recently retired as vice president of advanced data center technologies in Microsoft's Cloud Operations and Innovation Division, Redmond, WA, 98052, USA. He is now a strategic foresight advisor and board member for Digital Bridge and infrastructure-related companies focusing on data center technologies and sustainability. His research interests include all aspects of data centers. Belady received his M.Eng. degree in mechanical engineering from Rensselaer Polytechnic Institute. He a fellow of the American Society of Mechanical Engineers and the International Microelectronics Assembly and Packaging Society. He was recently elected to the National Academy of Engineering. Contact him at cbelady@hotmail.com.

Anand Sivasubramaniam is a distinguished professor of computer science and engineering at Pennsylvania State University, University Park, PA, 16802, USA. His research interests include data center power and energy management, sustainable computing, and computer architecture. Sivasubramaniam received his Ph.D. degree in computer science from Georgia Tech. He is a fellow of the Association for Computing Machinery and a Fellow of IEEE. Contact him at axs53@psu.edu.