SUSTAINABLE DATA CENTERS ROADMAP

CHAPTER 2.1

Information Technology (IT) Equipment

Alp Kucukelbir

October 2025

ICEF
Innovation for Cool Earth Forum

2.1 Information Technology (IT) Equipment

Alp Kucukelbir

Data centers are complex facilities that include information technology (IT) and infrastructure equipment, often distributed across several buildings. IT equipment typically comprises electronics housed in server racks, networked to one another and powered by electrical infrastructure. Each rack contains general purpose and/or accelerated processors, connected to memory chips and short- and long-term storage. Electrical infrastructure converts electricity to the appropriate levels for IT equipment operations.

A data center running artificial intelligence (AI) workloads operates like a massive, high-speed postal sorting facility designed to handle an enormous volume of specialized packages. The servers act as sorting stations where incoming data packets (like letters and parcels) are processed by workers (processors) who must quickly analyze, categorize and route information. Memory serves as temporary holding areas where packages wait to be processed, while storage systems function as vast warehouses storing reference materials needed for sorting decisions. The networking infrastructure resembles conveyor belts that move packages between stations.

As described in Chapter 1, AI workloads are particularly energy intensive; processing AI packages requires cross-referencing millions of previous deliveries simultaneously—each “package” (data point) must be compared against vast databases of patterns and examples before determining where it should go next.

A. Defining and Measuring Energy Efficiency

Computing energy efficiency has improved by an extraordinary factor of 10 billion over 5 decades, from ENIAC's (Electronic Numerical Integrator and Computer) 150-kilowatt power consumption in 1946 to modern processors achieving billions of operations per watt-hour. This transformation followed Koomey's Law, which demonstrated that computations per watt-hour doubled every 1.6 years during the "golden age" from 1946 to 2000. The improvement was driven by transistors shrinking while maintaining nearly constant power density, creating exponential performance gains with stable or decreasing power consumption.[1]

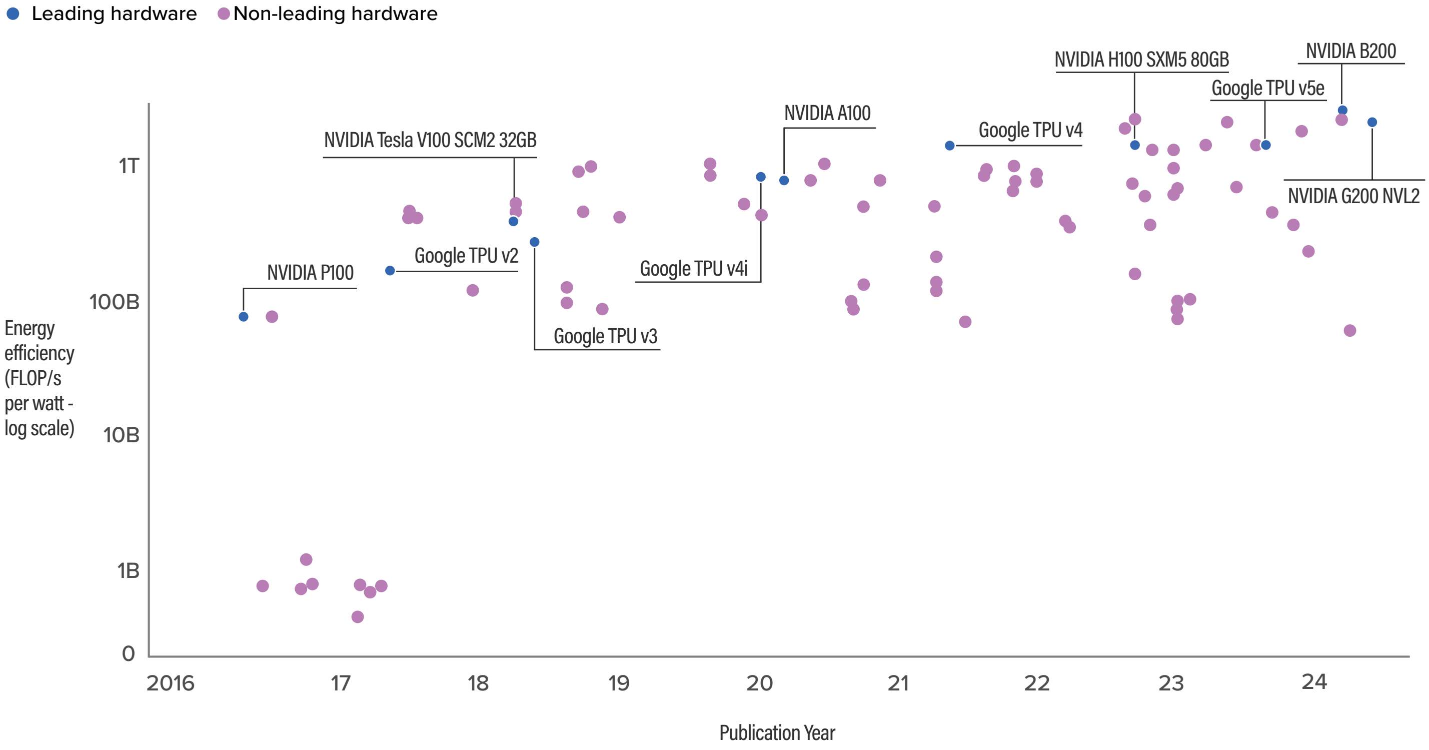
Figure 2.1-1. *Image of a data center.*

Around 2005, this golden age ended as Dennard scaling broke down due to quantum effects at atomic scales. The industry pivoted to architectural innovations, with recent advances showing continued but slower progress—efficiency now doubles every 2.3 years instead of every 1.6 years.[2] The last decade has seen exceptional technological progress. Processors advanced through smaller manufacturing nodes (22 nm to 3 nm), multi-core architectures and dynamic power scaling. Memory, storage and networking designs evolved. Critically, ARM processors and specialized AI chips challenged x86 dominance with superior performance-per-watt.

Improvements in the energy efficiency of IT equipment components have slowed but not stopped. These improvements are expected to continue in the years ahead. (See the second part of this chapter.) Yet when these component-level improvements are viewed at the system scale, a different picture emerges. Chip-level efficiency gains are increasingly offset by exponential demand growth from AI and other workloads. Despite these energy efficiency improvements, carbon emissions at data centers continue to rise.[3]

Companies, researchers and engineering consortia have yet to align on a comprehensive way to measure IT equipment energy efficiency in AI workloads. Epoch.AI, a research institute, tracks hardware efficiency trends using FLOPS per watt—a traditional metric applied to scientific computing involving many "floating point operations per second" (FLOPS). While this metric does not provide a complete picture due to IT equipment diversity and the nature of AI workloads, it serves as a good approximation. The MLCommons Power Work Group recently released a benchmark intended to include applications from data centers to mobile devices, from training to inference.[4] Stanford's Human-Centered AI 2025 Index Report, using Epoch.AI data, calculates a 40% reduction in energy consumption in AI-specific hardware per FLOP over the past three years, indicating that progress may be slowing down even further (Figure 2.1-2).[5]

Figure 2.1-2. *Energy efficiency of leading machine learning hardware, 2016-24.*

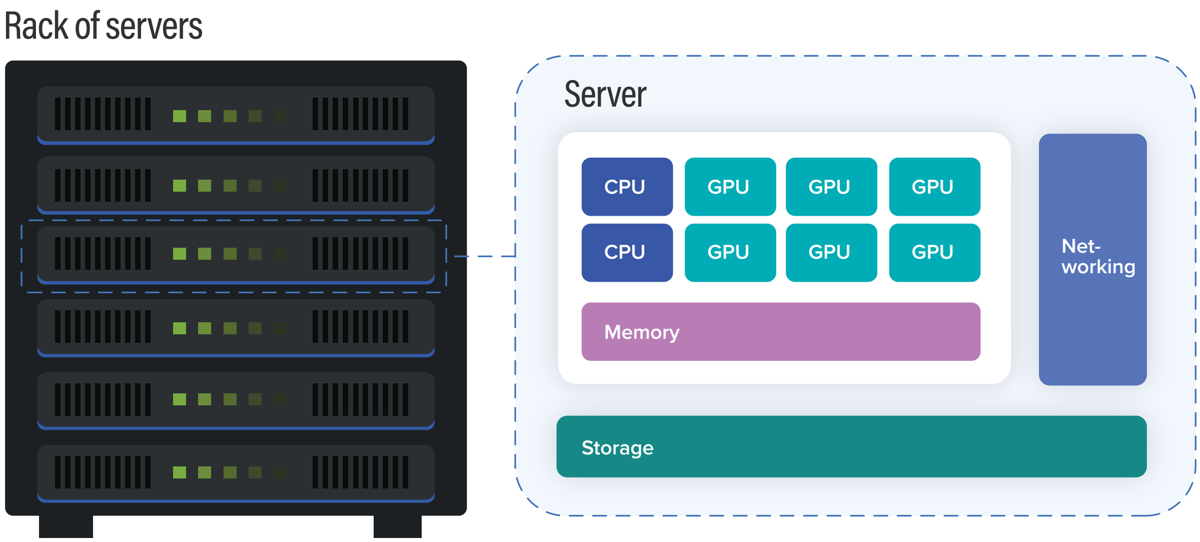


B. Components

Chapter 1 discussed energy use trends at the level of entire servers, showing how average power draw has increased over time. Below we discuss the components within those servers and how each contributes to overall energy use. Figure 2.1-3 presents a sketch of typical components found in a server. Each component consumes different amounts of energy in different ways.

Central processing units (CPUs) handle general-purpose computing tasks with moderate power draw (50-200 W), while accelerated processors, like graphics processing units (GPUs) and custom chips like tensor processing units (TPUs), excel at parallel processing but consume much more electricity (400-1000+ W). Memory provides fast temporary storage for active data, consuming negligible power that scales with capacity. Storage devices, such as solid-state drives (SSDs), provide persistent data storage, with SSDs using less power than older mechanical drives. Networking components move data between systems, with power consumption varying dramatically based on speed and distance requirements.

Figure 2.1-3. Components of a server.



i. Computation: processors and memory

Chapter 1 shows that servers consume the majority of power in data centers. Processors and memory are a primary source of energy usage within servers. Historically, data centers relied mainly on CPUs, which ran at roughly 50-200 W per chip. GPUs for AI have been drawing 400-700 W, with newer (2024+) chips expected to run at over 1000 watts.[6] GPUs contain thousands of smaller cores compared to CPUs' dozens of larger cores, meaning far more transistors are actively switching and consuming power simultaneously during parallel computations.

Major chip designers are pursuing different energy efficiency strategies. NVIDIA claims its most recent processor architecture uses up to 25x less energy relative to its own previous generation chips during AI inference.[7] Advanced Micro Devices (AMD) is focusing on using fewer GPUs to achieve similar outputs, claiming its recent offerings have a 30% advantage over competitors.[8] Intel targets energy savings with open system standards, claiming a 40% average power reduction relative to competitors.[9] Hyperscalers are targeting energy efficiency by specializing for AI applications. Google claims its most recent accelerated processor TPUs are 67% more energy efficient than previous versions.[10] Amazon Web Services similarly claims a 40% improvement in energy efficiency for its own accelerated processors.[11] Microsoft, Meta, IBM and other hyperscalers appear to be aiming to achieve similar outcomes.

Startup companies are developing novel processor architectures with promising energy efficiency performance. Untether AI recently demonstrated a 3-6x improvement across multiple machine learning (ML) Commons benchmark categories.[12] Groq claims to achieve 10x energy efficiency improvement through a novel chip design.[13]

The software that trains AI models and runs inference requires large amounts of memory. These chips consume energy as data are written to and read from memory. AI systems leverage higher memory bandwidth (the rate at which data are transferred) to accelerate AI training and inference. But higher bandwidth typically correlates with higher power consumption. Researchers and industry are working toward novel architectures to increase bandwidth while minimizing energy usage.[14]

ii. Storage

Chapter 1 distinguishes energy usage between servers and storage. Below we discuss storage technology within the context of servers (for short- to medium-term information needs during processing).

SSDs provide the high-speed data access required for training large AI models. During training, AI systems must process massive data sets containing billions of examples, which requires continuous, rapid data transfer from storage to processing units. For AI inference operations, SSDs enable rapid access to AI models and any reference data needed for generating responses.

AI workloads create unique energy demands for SSDs due to their sustained high-throughput patterns. During training, SSDs experience bursty read and write patterns as data sets are streamed repeatedly, keeping the drives in active states rather than allowing them to enter lower-power idle modes. In contrast, memory usage is read-heavy, continuous and sustained during inference.

Current SSD technology is almost twice as fast as the previous generation but consumes 50-100% more energy during peak operations.[15] Memory chip manufacturers, such as Samsung and Micron, are developing new designs that maintain performance while reducing energy consumption by half.[16,17]

iii. Networking

Data center networking controls how information flows throughout the system. Switches direct data to their destinations—some locally within server racks, others regionally within building sections. Load balancers ensure that no single server becomes overwhelmed. Together, these devices aim to provide fast and reliable information flow.

Similar to memory and storage requirements, AI workloads require high bandwidth networks. These expand beyond traditional enterprise networking standards. AI data centers' bandwidth requirements can range from several gigabits per second to terabits per second.[18]

Traditional enterprise networking equipment consumes dramatically less power per port than specialized AI networking hardware, with power differences 3-10x higher for AI equipment. Despite this increase, networking remains under 5% of total data center power consumption due to the massive scale of AI compute requirements. Networking equipment manufacturers are adopting technologies, such as co-packaged optics, to increase bandwidth while reducing power consumption, with projections of up to an 80% reduction in energy usage.[19,20]

C. Innovations and Forecasted Efficiency Gains

Data centers present significant opportunities for improving energy efficiency through both established best practices and emerging innovations. While conventional strategies benefit all data center operations (see Table 2.1-1 below), AI workloads demand specialized approaches, including high-bandwidth networking, sustained storage performance and power oversubscription tailored to inference patterns. Hardware advances in electronics, photonics and power distribution can improve energy efficiency, while tailoring server designs and data center operations to specific AI workloads—training versus inference—offers further energy savings. Although the rate of improvement in energy efficiency appears to be slowing, the innovations below may lead to new upticks in progress.

Table 2.1-1. Conventional data centers typically operate with legacy equipment at low utilization rates, basic power distribution systems, minimal virtualization, reactive operational practices and outdated hardware refresh cycles.

Category	Best Practice	Baseline
Electronics Manufacturing	Select ENERGY STAR certified servers and equipment	Standard servers without efficiency certifications
	Deploy solid-state drives (SSDs) over mechanical drives	Traditional spinning disk drives
	Use efficient memory technologies (e.g., DDR5, low-voltage)	Older DDR3/DDR4 standard voltage memory
Processors	Enable processor power management (e.g., C-states, P-states)	Processors running at constant maximum frequency
	Implement dynamic voltage/frequency scaling	Fixed voltage and frequency operation
Storage	Storage virtualization and deduplication	Direct-attached storage with redundant data
	Tiered storage with automated data placement	Single-tier storage systems
	Implement storage power management	Always-on storage arrays
Networking	Deploy energy-efficient network switches	Legacy switches without power scaling
	Network virtualization and consolidation	Physical dedicated network infrastructure
	Use adaptive link rate and port power management	Fixed-rate network interfaces
Power Conversion	High-efficiency power supplies (>90% efficiency)	Standard 80% efficient power supplies
	High-efficiency uninterruptible power supply (UPS) systems (>95%)	Traditional UPS systems at 85–90% efficiency
	Power factor correction (>0.95)	Uncorrected power factor (0.7–0.8)
Operations	Real-time power monitoring and management	Manual monitoring with monthly readings
	Automated workload scheduling and migration	Static workload placement

i. Electronics manufacturing and packaging

Chip packaging is the process of enclosing semiconductors in protective materials and connecting them to external circuits through pins. It provides physical protection, electrical connections and thermal management for delicate silicon chips.

Advanced packaging techniques, such as Taiwan Semiconductor Manufacturing Company's (TSMC's) chip-on-wafer-on-substrate (CoWoS), are delivering a 30%

reduction in power consumption compared to traditional packaging through reduced interconnect distances and improved thermal management.[21] The company is targeting a doubling of its production capacity to 75,000 wafers in 2025.[22]

AI-designed chips represent the next frontier of chip designs with better power efficiency.[23] For example, researchers at Oregon State University leveraged AI to design chips from new materials, achieving a 6x improvement in energy efficiency.[24]

ii. Photonics innovation

Silicon photonics innovation is revolutionizing data center networking. This technology uses light (photons) instead of electricity to transmit data through silicon-based optical components like waveguides and modulators. It achieves better power efficiency because photons do not generate resistive heat during transmission and can carry data over longer distances without signal amplification, unlike electrical signals that require constant power to overcome resistance and maintain signal integrity.[25]

Companies are leading research and development (R&D) efforts in photonics. Intel's co-packaged optics demonstrated a 67% power reduction over pluggable optical transceivers.[26] Ayar Labs' TeraPHY chiplets claim 4-8x greater power efficiency than traditional interconnects,[27] while NVIDIA's co-packaged solutions claim 3.5x better power efficiency.[28]

iii. Power conversion improvements

Power electronics are circuits that convert and control electrical power between different voltage levels or frequencies or from AC to DC. Gallium nitride (GaN) and silicon carbide (SiC) semiconductors have wider bandgaps than silicon, allowing them to operate at higher voltages, temperatures and switching frequencies with lower resistance. This enables smaller, lighter power converters with less energy lost as heat during conversion.

GaN and SiC adoption is already driving efficiency gains in data centers.[29] These technologies are showcasing up to 98% energy conversion efficiency[30] versus the 85-90% baselines seen with older power electronics. Both technologies are in active development, with pathways to an increased range of operation and reliability.[31]

iv. Data center operations

AI workflows are creating specific opportunities for operating data center IT equipment. Microsoft reports that the average and peak power usage in AI inference are bounded. This implies that AI data centers used solely for inference offer substantial headroom for power oversubscription, allowing the deployment of 30% more servers per AI inference data center with minimal performance loss.[32]

v. Server design

Server design innovations center on liquid cooling advances that dramatically improve energy efficiency. (Chapter 2.3 explores this topic in detail.) Other strategies include Microsoft Azure's efforts to use existing components to assemble server racks, leading to a net 8% reduction in emissions.[33] Amazon Web Services (AWS) has focused on standardized yet modular equipment to help retrofit existing data centers.[34]

Box 2.1-1

Moving computation to the edge

Some AI use cases offer opportunities to avoid data centers altogether. So-called "edge AI" processes data locally on devices (smartphones, industrial robots, autonomous vehicles) rather than communicating with data centers. Motivated by the ability to work offline, with low latencies and strong data privacy guarantees, running AI workloads "on the edge" can also reduce energy consumption.

Edge AI workflows eliminate data transmission to data centers, thereby reducing their computational load. Specialized processors, such as field-programmable gate arrays (FPGAs) and edge-optimized AI chips, can effectively run AI inference workflows.[35] The tradeoff is that edge AI devices are typically more expensive and less energy efficient than traditional counterparts. Multiple companies, such as DEEPX, Hailo and Axelara, are actively developing energy efficient and affordable edge AI processors.

While AI training will certainly remain at data center levels, forecasting how AI inference demand might shift toward the edge is challenging.[36] Rising data center costs, regulatory pressure for data privacy and real-time applications may push toward edge AI adoption. However, increasingly larger model sizes and edge AI device costs may continue to drive AI inference demand in data centers.

D. Recommendations

1. *Companies, industry standard setters and engineering consortia should **align on common metrics for calculating and reporting the energy efficiency of IT equipment**. Data center operators should support such efforts since energy efficiency directly impacts their operating costs.*

2. *Governments and educational institutions should **develop and distribute resources to assist non-technical audiences in understanding and analyzing the energy requirements of IT equipment.***

3. *Data center operators, utilities and government agencies should **consider the nature of AI computation workloads in designing, provisioning, operating and regulating data centers.** Differences between AI training and inference should be paramount during decision making.*

4. *Governments, utilities and industry consortia should **advance knowledge sharing platforms, case study data, diagnostic tools and training materials related to improving the energy efficiency of IT hardware** from procurement, operations and management perspectives.*

5. *Data center operators should **conduct, support and publish AI inference demand forecasts.** Trends between centralized data center computations and edge applications should be emphasized.*

6. *Data center operators should **redouble efforts to maximize the energy efficiencies of existing IT hardware**, including the adoption of efficient equipment, virtualization, zombie server identification initiatives, and refresh cycles optimized for energy efficiency.*

E. References

1. Jonathan Koomey, Stephen Berard, Marla Sanchez & Henry Wong. Implications of Historical Trends in the Electrical Efficiency of Computing. IEEE Annals of the History of Computing 33, 46-54 (2011). https://doi.org/10.1109/MAHC.2010.28.

2. Alberto Prieto, Beatriz Prieto, Juan José Escobar & Thomas Lampert. Evolution of computing energy efficiency: Koomey's law revisited. Cluster Computing 28, 42 (2024). https://doi.org/10.1007/s10586-024-04767-y.

3. International Energy Agency (IEA). CO_2 emissions associated with electricity generation for data centres by case, 2020-2035; IEA, Paris, France, https://www.iea.org/data-and-statistics/charts/co2-emissions-associated-with-electricity-generation-for-data-centres-by-case-2020-2035 (2025).

4. MLCommons. MLCommons Power Working Group Presents MLPerf Power benchmark at IEEE HPCA Symposium; MLCommons, Dover, Delaware, https://mlcommons.org/2025/03/ml-commons-power-hpca/ (2025).

5. Nestor Maslej, Loredana Fattorin, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika et al. The AI Index 2025 Annual Report; AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, California, https://hai.stanford.edu/ai-index/2025-ai-index-report (2025).

6. Karthik Ramachandran, Duncan Stewart, Kate Hardin, Gillian Crossan & Ariane Bucaille. As generative AI asks for more power, data centers seek more reliable, cleaner energy solutions (Deloitte Insights); Deloitte Center for Technology, Media & Telecommunications, London, United Kingdom, https://www.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2025/genai-power-consumption-creates-need-for-more-sustainable-data-centers.html (2024).

7. Kristin Uchiyama. NVIDIA Blackwell Platform Arrives to Power a New Era of Computing; NVIDIA, Santa Clara, California, https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing (2024).

8. Advanced Micro Devices (AMD). AMD Instinct MI300 Series Accelerators; AMD, Santa Clara, California, https://www.amd.com/en/products/accelerators/instinct/mi300.html (Accessed August 2025).

9. Intel Newsroom. Intel Unleashes Enterprise AI with Gaudi 3, AI Open Systems Strategy and New Customer Wins; Intel, Santa Clara, California, https://newsroom.intel.com/artificial-intelligence/vision-2024-enterprise-ai-gaudi-3-open-systems-strategy (2024).

10. Amin Vahdat. Announcing Trillium, the sixth generation of Google Cloud TPU; Google Cloud, Mountain View, California, https://cloud.google.com/blog/products/compute/introducing-trillium-6th-gen-tpus (2024).

11. CTOL Editors - Yasmine. AWS Unveils Next-Gen AI Chips with Trainium 3 and Ultra Servers at re:Invent 2024, But Unlikely to Challenge Nvidia's Dominance; CTOL Digital Solutions, Ennetbaden, Switzerland, https://www.ctol.digital/news/aws-trainium-3-chips-vs-nvidia-ai-hardware-reinvent-2024/ (2024).

12. Untether AI. Untether AI Announces speedAI Accelerator Cards As World's Highest Performing, Most Energy Efficient AI Accelerators According to MLPerf Benchmarks; Untether AI, Toronto, Ontario, https://www.untether.ai/untether-ai-announces-speedai-accelerator-cards-as-worlds-highest-performing-most-energy-efficient-ai-accelerators-according-to-mlperf-benchmarks/ (2024).

13. Groq. Products - Groq Is Fast AI Inference; Groq, San Jose, California, https://groq.com/products (Accessed August 2025).

14. Joo-Hyung Chae. High-Bandwidth and Energy-Efficient Memory Interfaces for the Data-Centric Era: Recent Advances, Design Challenges, and Future Prospects. IEEE Open Journal of the Solid-State Circuits Society 4, 252-264 (2024). https://doi.org/10.1109/OJSSCS.2024.3458900.

15. Anshul Rana. SSD Power Consumption: How Much Watts SSDs Consume?; StoredBits.com, https://storedbits.com/ssd-power-consumption/ (2025).

16. Samsung News. Now Available: Samsung 9100 PRO Series SSDs with Breakthrough PCIe® 5.0 Performance; Samsung, Suwon-si, South Korea, https://news.samsung.com/us/samsung-announces-9100-pro-series-ssds-with-breakthrough-pcie-5-0-performance/ (2025).

17. Alvaro Toledo. The world's fastest data center SSD for AI workloads; Micron Technology, Inc., Boise, Idaho, https://www.micron.com/about/blog/storage/ssd/revolutionize-ai-workloads-with-the-worlds-fastest-data-center-ssd. (2024).

18. Alexandra Jonker & Alice Gomstyn. What is an AI data center?; IBM, Armonk, New York, https://www.ibm.com/think/topics/ai-data-center (2025).

19. Laura Peters. Co-Packaged Optics Reaches Power Efficiency Tipping Point; Semiconductor Engineering, https://semiengineering.com/co-packaged-optics-reaches-power-efficiency-tipping-point/ (2025).

20. NVIDIA. NVIDIA Silicon Photonics; NVIDIA, Santa Clara, California, https://www.nvidia.com/en-us/networking/products/silicon-photonics/ (Accessed August 2025).

21. Taiwan Semiconductor Manufacturing Company (TSMC). TSMC Unveils Next-Generation A14 Process at North America Technology Symposium; TSMC, Hsinchu Science Park, Taiwan, https://pr.tsmc.com/english/news/3228 (2025).

22. TrendForce News. TSMC Set to Expand CoWoS Capacity to Record 75,000 Wafers in 2025, Doubling 2024 Output; TrendForce Corporation, Taipei, Taiwan, https://www.trendforce.com/news/2025/01/02/news-tsmc-set-to-expand-cowos-capacity-to-record-75000-wafers-in-2025-doubling-2024-output/ (2025).

23. Jeroen Kusters, Deb Bhattacharjee, Jordan Bish, Jan Thomas Nicholas, Duncan Stewart & Karthik Ramachandran. 2025 global semiconductor industry outlook (Deloitte Insights); Deloitte Center for Technology, Media & Telecommunications, London, United Kingdom, https://www.deloitte.com/us/en/insights/industry/

technology/technology-media-telecom-outlooks/semiconductor-industry-outlook.html (2025).

24. Steve Lundeberg. New computer chips show promise for reducing energy footprint of artificial intelligence; Oregon State University Newsroom, Corvallis, Oregon, https://news.oregonstate.edu/news/new-computer-chips-show-promise-reducing-energy-footprint-artificial-intelligence (2024).

25. Shupeng Ning, Hanqing Zhu, Chenghao Feng, Jiaqi Gu, Zhixing Jiang, Zhoufeng Ying, Jason Midkiff, Sourabh Jain, May H. Hlaing, David Z. Pan & Ray T. Chen. Photonic-Electronic Integrated Circuits for High-Performance Computing and AI Accelerators. Journal of Lightwave Technology 42, 7834-7859 (2024). https://doi.org/10.48550/arXiv.2403.14806.

26. Intel Newsroom. Intel Demonstrates First Fully Integrated Optical I/O Chiplet; Intel, Santa Clara, California, https://newsroom.intel.com/artificial-intelligence/intel-unveils-first-integrated-optical-io-chiple (2024).

27. Ayar Labs. Optical I/O Technology is Essential to Eliminate Bottlenecks; Ayar Labs, Inc., San Jose, California, https://ayarlabs.com/optical-io-products/ (Accessed August 2025).

28. NVIDIA Co-Packaged Silicon Photonics Networking Switches." NVIDIA, https://www.nvidia.com/en-us/networking/products/silicon-photonics/.

29. Gary Elinoff. GaN and SiC Transform AI Data Centers Efficiency; Electropages Media Ltd, Dorset, United Kingdom, https://www.electropages.com/blog/2024/07/gan-and-sic-transform-ai-data-centers-efficiency (2024).

30. Justin Chou. Scaling AI Data Center Power Delivery with Si, SiC, and GaN; Electronic Design, Nashville, Tennessee, https://www.electronicdesign.com/technologies/power/power-supply/article/55289191/scaling-ai-data-center-power-delivery-with-si-sic-and-gan (2025).

31. Buffolo, M., et al. "Review and Outlook on GaN and SiC Power Devices: Industrial State-of-the-Art, Applications, and Perspectives." IEEE Transactions on Electron Devices, vol. 71, no. 3, March 2024, pp. 1344-55. IEEE Xplore, https://doi.org/10.1109/TED.2023.3346369.

32. Pratyush Patel, Esha Choukse, Chaojie Zhang, Íñigo Goiri, Brijesh Warrier, Nithish Mahalingam & Ricardo Bianchini. "Characterizing Power Management Opportunities for LLMs in the Cloud" in Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, La Jolla, CA, USA. 207-222, https://doi.org/10.1145/3620666.3651329 (2024).

33. Jaylen Wang, Daniel S. Berger, Fiodar Kazhamiaka, Celine Irvene, Chaojie Zhang, Esha Choukse, Kali Frost, Rodrigo Fonseca, Brijesh Warrier, Chetan Bansal, Jonathan Stern, Ricardo Bianchini & Akshitha Sriraman. Enabling Sustainable Cloud Computing with Low-Carbon Server Design. IEEE Micro, 1-9 (2025). https://doi.org/10.1109/MM.2025.3572955.

34. Amazon Press Center. AWS Announces New Data Center Components to Support AI Innovation and Further Improve Energy Efficiency; Amazon.com, Inc., Las Vegas, Nevada, https://press.aboutamazon.com/2024/12/aws-announces-new-data-center-components-to-support-ai-innovation-and-further-improve-energy-efficiency (2024).

35. Péter Szántó, Tamás Kiss & Károly János Sipos. “Energy-efficient AI at the Edge” in 2022 11th Mediterranean Conference on Embedded Computing (MECO), 1-6, https://doi.org/10.1109/MECO55406.2022.9797178,(2022).

36. Steven Carlini. The Current And Future Path To AI Inference Data Center Optimization; Forbes Media LLC, Jersey City, New Jersey, https://www.forbes.com/councils/forbestechcouncil/2025/01/28/the-current-and-future-path-to-ai-inference-data-center-optimization/ (2025).