uptime INTELLIGENCE

INTELLIGENCE UPDATE

# GPU power management is a work in progress

Max Smolaks

4 Jun 2025

Specialized IT hardware designed to perform generative AI training and inference is much more efficient at the job than generic IT systems. However, it is becoming clear that GPU-heavy servers are subject to many of the same problems that negatively impact the efficiency of traditional enterprise IT: low infrastructure utilization, high idle power and wasteful peak performance states.

The data center industry is at the precipice of an escalating power problem. Infrastructure for generative AI consumes a substantial amount of power and accounts for a significant portion of new data center capacity. Training of large language models (LLMs) currently dominates AI infrastructure needs, with the objective of squeezing the most performance out of hardware to shorten training times. Even if total energy is often a secondary consideration for AI training runs, system utilization levels are high, which means relatively high (if suboptimal) power efficiency.

But the efficiency problem becomes greater with inference workloads, which will be responsible for a growing — and ultimately much larger — share of data center power consumption. In inference, AI-based services respond to users' queries and workload behavior is less predictable. The priority is cost-efficient token generation, which is often tied to revenue or business functions. Businesses will want to optimize inference hardware for power efficiency (while meeting quality of service) to supress costs and avoid waste of energy — on a colossal scale globally.

Better GPU power management mechanisms can significantly reduce energy spent per token, capital costs and carbon emissions in AI data centers. However, the control mechanisms are not currently well developed or understood.

## New hardware, old problems

Generic server processor (CPU) power management is a well-established discipline in traditional IT. The theoretical goal is simple: to ensure the server only uses the necessary amount of energy for the job. The underlying mechanisms are complex, involving software components in the operating system and hardware features in the silicon, such as voltage-frequency controls. A

© COPYRIGHT 2025 UPTIME INSTITUTE. ALL RIGHTS RESERVED.

key aspect of power management is addressing idle power: CPU power can be cut considerably during periods of low workload activity — even reduced to near-zero for large parts of the chip.

This technology is mature, well understood and clearly explained by vendors. It is known to reduce IT energy consumption effectively at certain points of load (see *Understanding how server power management works*). Still, it is not always used since it involves a trade-off between processing power and energy consumption, introducing latency into chip responses.

The process is more complex for GPUs. A current-generation high-performance data center GPU is not a single chip but a complex multi-chip package. It combines several compute and memory dies, which means at least two, but likely more clock domains whose operating frequencies impact power draw. Default power management tools offer only the most basic functionality: Nvidia's System Management Interface (nvidia-smi) supports power capping and offers three performance modes that are likely limiting clock frequencies. Little detail is available about how these affect overall efficiency.

## GPU power management: what we know

- Tokens (collections of characters that are the fundamental units of data processed by AI models) are used to measure both the input and output of LLMs, and to calculate the cost of queries. In AI training, the focus is on maximizing overall hardware performance to shorten the model development time. In inference, the primary focus shifts to maximizing the number of tokens generated per watt; this metric allows model owners to compare and manage the efficiency of inference. It can be improved by using power management features.
- Similarly to CPUs, GPU performance does not scale linearly with power consumption. The sweet spot for most efficient token generation will not necessarily be the same as the point where a GPU-equipped server delivers the most tokens per second. Counter-intuitively, achieving the best token per watt output will often require GPUs to run at reduced capacity. With modern CPUs, the best server efficiency is typically achieved in the 60% to 80% capacity utilization range (see *The strong case for power management*).
- Power management features can enable GPUs to deliver more inference within the same power envelope — by keeping devices in power and performance states close to their power efficiency optimum. The drawback of this approach is the need to overprovision the number of GPUs, which can become expensive.
- Default GPU power management features are dictated by drivers and firmware, which is not nearly as developed as the firmware that governs CPUs. Conversations with experts suggest that current generation Nvidia GPUs idle at 20% of nameplate power and will not idle at all if they have data stored in on-board memory. For comparison, modern CPUs from Intel and AMD typically idle at 1% to 10%.

## GPU power management: what is still unknown

- Hardware utilization ranges for GPUs that would enable them to achieve the optimal levels of efficiency, as expressed in tokens per watt are not available. Optimal utilization ranges for specific CPUs are established in experimental conditions, through benchmarks such as The Green Grid’s publicly available Server Efficiency Rating Tool (SERT) database. No widely recognized benchmarks of this kind exist for

© COPYRIGHT 2025 UPTIME INSTITUTE. ALL RIGHTS RESERVED.

modern GPUs.
- The impact of high (near 100%) utilization on the lifespan of GPUs is uncertain. Silicon and electronic components can deteriorate rapidly when exposed to high current, high voltage and high temperature. Anecdotal evidence suggests GPUs can fail in as little as two to three years of sustained heavy load.
- The impact of specific power management features on the performance of GPUs is unclear. For modern CPUs, power management can reduce server work capacity (measured in SSJ transactions per second) by up to 6.4%, while reducing power consumption by up to 21%.
- The key question concerns the total cost of ownership for inference hardware: would it be more cost-efficient to run GPUs at near 100% utilization, with an associated reduction in tokens per watt and a potentially shorter lifespan, or run more GPUs at lower utilization levels, increasing capital expenditure required but reducing the cost of power?

# Immediate solutions

Today, GPU vendors such as Nvidia and AMD are more interested in improving maximum performance of their products than in promoting efficient AI compute. This has left the door open for smaller AI accelerator vendors, such as Qualcomm, SambaNova, Groq, Blaize, Tenstorrent and others, to specialize in efficient inference. Several cloud vendors have developed their own inference chips, promising power-efficient operation. It is important to note that these companies do not make GPUs — their designs are simpler, cheaper and lack some of the features required to train large models. These hardware platforms might represent a more economic choice for inference workloads, at least in the near term.

On the software front, a wave of new products aimed at fixing GPU power management shortcomings is coming. American startup Neuralwatt is developing GPU power consumption optimization software powered by machine learning. Firmware specialist AMI has improved the DCM software platform (previously developed by Intel) to provide insights into GPU health, performance and power consumption to facilitate improved resource utilization. In time, integration of accurate GPU power monitoring capabilities into IT management tools will enable users to investigate energy-related aspects of AI compute in more detail.

What the industry needs right now is clarity from the GPU vendors: descriptions of the trade-offs between power and efficiency, efficiency benchmarking, and a simple admission that it might not always be desirable to drive these servers at 100% load.

Tools such as the SERT database enable easy comparisons of products from different CPU vendors, helping customers make more informed choices, and to pick the right chip for the job. If GPUs are to find a home in enterprise data centers, they need to offer the same opportunity to compare the efficiency of hardware — not just performance.

## The Uptime Intelligence View

© COPYRIGHT 2025 UPTIME INSTITUTE. ALL RIGHTS RESERVED.

AI is going to be responsible for a rapidly increasing share of data center capacity and much of this capacity will rely on GPUs from Nvidia and AMD. Similarly to other types of IT resources, GPUs need to be managed effectively to achieve the most efficient operation, but there's little information available on the features designed to conserve GPU power.

CPU power management features and CPU benchmarking via tools such as the SERT database, provide a blueprint for what is required of enterprise-grade silicon. GPU vendors will have to follow this blueprint — or risk customers moving to alternative hardware platforms that made efficient inference their sole mission.

Thumbnail photograph courtesy of Fritzchens Fritz / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

© COPYRIGHT 2025 UPTIME INSTITUTE. ALL RIGHTS RESERVED.

## ABOUT THE AUTHOR

### Max Smolaks

Max is a Research Analyst at Uptime Institute Intelligence. Mr Smolaks' expertise spans digital infrastructure management software, power and cooling equipment, and regulations and standards. He has 10 years' experience as a technology journalist, reporting on innovation in IT and data center infrastructure.

**msmolaks@uptimeinstitute.com**

© COPYRIGHT 2025 UPTIME INSTITUTE. ALL RIGHTS RESERVED.

## About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. Its Tier Standard is the IT industry's most trusted and adopted global standard for the proper design, construction, and operation of data centers – the backbone of the digital economy. For over 25 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions.

With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.

© COPYRIGHT 2025 UPTIME INSTITUTE. ALL RIGHTS RESERVED.