INTELLIGENCE UPDATE

GPU utilization is a confusing metric


Max Smolaks

1 May 2025

The specialized IT equipment required to perform AI training and inference is relatively new. These devices are expensive and need to be used effectively, especially GPUs for AI. Yet research literature, disclosures by AI cluster operators and model benchmarks suggest that — similarly to other types of IT infrastructure — GPU resources are often wasted. Many AI teams are unaware of their actual GPU utilization, often assuming higher levels than those achieved in practice.

On average, GPU servers engaged in training are only operational 80% of the time. When these servers are running, even well-optimized models only reach 35% to 45% of compute performance that the silicon can deliver. The numbers are likely worse for inference, where the workload size is dynamic and less predictable, fluctuating with the number and complexity of end-user requests.

Many factors can limit the performance and efficiency of GPUs:

- **Availability of work.** Queuing up a training workload that might take weeks or even months to complete is a complex process with many dependencies. Delays between projects are to be expected.
- **Network latency and throughput.** Training generative AI models requires GPUs to exchange significant amounts of data as quickly as possible. Insufficient network throughput or latency can lead to GPUs waiting for input.
- **GPU failure rates.** Hardware failures can involve GPU chip or on-board memory defects, communication errors and power management issues. The larger the cluster, the higher the probability that a failure will disrupt the training process. In an example shared by Meta in 2024, the training run involving a cluster of 16,384 GPUs encountered 419 component failures over 54 days, or one every three hours.
- **Checkpointing frequency.** To prevent loss of work in the event of a hardware failure, developers save the intermediate model training steps called checkpoints at regular intervals (typically every few minutes). Creating a checkpoint takes time because data needs to be aggregated and moved from GPU memory to server memory, and from server memory to storage. GPUs are not engaged in computations during this time, with a corresponding drop in power demand across all nodes.
- **Storage performance.** Higher throughput of storage systems can accelerate access to training data, the checkpoint creation process and recovery from hardware failures. Less performant storage will extend the duration of training and the amount of time

GPUs are not engaged in computation.
- **Model architecture.** Most of all, training workload efficiency is impacted by the software architecture of the specific model. This includes batch size, data buffering and memory configuration, floating-point precision and other settings that can be used to improve model performance on specific hardware.

These are just some of the factors that can affect the amount of compute delivered by GPUs, the overall power consumption of the cluster and its cooling requirements. Even in computationally intensive workloads, node-level power demand rarely approaches manufacturer-rated maximums.

Having a simple utilization metric for GPUs would be a boon for the industry; unfortunately, GPUs are unlike other server components and require new ways of accounting for performance. Potential metrics are complex but useful to understand as operators prepare for the arrival of GPUs in their data centers.

What are we measuring and why?

The most basic approach to defining and tracking GPU utilization looks at average server operational time. This is useful since it accounts for the AI accelerators and other server components such as CPUs, memory and storage devices.

Estimates by the Lawrence Berkeley National Laboratory (LBNL) state that GPU servers are, on average, engaged in useful work between 75% and 85% of the time, spending the rest consuming idle power — at around 20% of nameplate power. This metric is of limited use to data center operators; while it does affect the overall power consumption of a cluster over time, it does not describe the level of sustained power required to support it.

The second approach tracks individual GPU load via the tools and functionality provided by hardware designers such as Nvidia and AMD (e.g., running “nvidia-smi” command line utility). This represents the most common definition of GPU utilization.

This utilization data can be easily accessed and is used by many observability tools but is not always the best metric for understanding GPU efficiency. What it measures is the share of the discrete GPU processing elements — called streaming multiprocessors by Nvidia and compute units by AMD — that are executing at a given time. It does not distinguish between the work done by the compute cores, and the work done by moving data in and out of memory; in fact, 100% GPU utilization can be achieved while doing no computation. Therefore, it is not a suitable metric to establish whether a workload takes full advantage of GPU capabilities.

In addition, while essential to training and inference, memory operations have a much lower power consumption profile than compute operations. A 100% utilized GPU moving data consumes a fraction of the power that would be consumed by a 100% utilized GPU running matrix multiplication calculations (the foundation of generative AI workloads). This discrepancy makes most GPU utilization data unsuitable for power consumption estimates — and many other

applications.

The third method of accounting for GPU performance and efficiency — and likely the most objective — is model FLOPS (floating point operations per second) utilization (MFU). Initially introduced by Google Research in 2022, this metric tracks the ratio of the observed performance of the model (measured in tokens per second) to the theoretical maximum of the underlying hardware operating at peak throughput (as reported by manufacturer, with no memory or communication overhead). A higher MFU indicates better efficiency, which means cheaper and shorter training runs.

While this metric sounds convenient, it is notoriously difficult to calculate and the results are often surprising: even well-optimized models only reach between 35% and 45% MFU. Why are these numbers so low? Manufacturer-specified maximum throughput ignores implementation details, but the observed performance is impacted by the factors described earlier in this report, such as network and storage throughputs, and the patchwork of software products, custom code and mathematics that make a model out of training data.

MFU has quickly gained prominence as the target metric for model developers. Due to the physical limits in chip-to-chip communications, it will never reach 100%; results above 50% were considered state-of-the-art in early 2025.

The main drawback of MFU is that the formula involves the core elements of architecture that can be radically different from one model to the next. It is a good indicator of whether specific features make a particular model more or less efficient, but not a perfect basis for comparing models.

Even so, MFU should interest data center operators, because it has a more-or-less direct relationship with power consumption. Higher MFU generally means that more of the GPU resources are engaged in performing work and therefore consuming more power. However, it is not a perfect linear correlation.

What about power?

For data center operators, hardware utilization metrics are a useful indicator of the demand for operational power and cooling. LBNL predicts that, between 2024 and 2028, GPU-equipped servers will average between 60% and 80% of their nameplate power.

Making these predictions is difficult. At present, much of the GPU performance data is inconsistent. Performance maximums are either theoretical (provided by GPU designers) or obtained via benchmarks where systems are optimized to run specific software. Few operators know what “good” levels of utilization of AI infrastructure in production actually look like.

More information is required on how GPUs perform in real-world settings, and the exact scale of the effects that the various bottlenecks have on AI hardware cluster power consumption. Many organizations treat this information as proprietary. What makes matters worse is the people that

run AI workloads — and have knowledge of how they behave over time — operate in small teams that are far removed from both facilities and IT operations teams. Obtaining this information for comparison and meaningful analysis will be difficult and is likely to take some time.

In the interim, operators can run their own experiments, metering the total power delivered to a training cluster or group of inference servers. Combining this data with utilization measurements will demonstrate how different AI workloads affect power consumption.

Unfortunately, as they try to understand GPUs, operators face an unusual obstacle: GPU vendors. Nvidia has repeatedly claimed that its systems “operate at or near peak utilization continuously when running AI workloads” in its data center design reference literature. This is, at best, a half-truth.

The Uptime Intelligence View

GPU resources must be used effectively but the industry lacks the data to know what levels of performance can be deemed “effective.” The metric commonly called GPU utilization is not particularly useful for understanding efficiency, but MFU shows promise. More data about real-world deployments needs to be collected to establish what “good” looks like for an efficient AI cluster.

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

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.