uptime NETWORK

INTELLIGENCE UPDATE

The problem with energy per token

Dr. Owen Rogers 19 May 2026

As adoption of generative AI grows, the energy demand of this technology is becoming an increasingly important consideration for data center operators, end users and policymakers. However, infrastructure power consumption alone reveals little about the useful work being performed: the same AI cluster may deliver very different levels of output depending on how it is configured and used.

Energy-per-token metrics aim to address this problem by linking energy consumption directly to the AI-generated output. Tokens are fragments of words that represent the basic units of data manipulated by most large language models. By expressing efficiency as joules per token (or its derivatives such as tokens per watt), data center operators can compare AI models, hardware platforms, and inference architectures using a common unit tied to application activity. The metric is also attractive because it can be translated into economic and sustainability measures, including cost per token and carbon emissions per token.

In practice, however, calculating this metric for a given deployment is difficult, and determining a fair benchmark is even harder.

Why token energy is hard to calculate

A crucial factor in calculating energy per token is the model's inference throughput, usually expressed as tokens per second. This throughput varies significantly depending on the lengths of the input queries and output responses, concurrency levels, the model being used, and the hardware. Some of these factors, such as the model and hardware, are under the operator's control. Others, such as prompt length and the responses given, depend on interactions between the model and end users. As a result, predicting inference throughput is extremely challenging.

Benchmark data from Nvidia shows significant variability in achievable throughput, even when using the same model and hardware. This data shows optimized performance of Nvidia hardware when utilizing its containerized inference software, NIM.

Assume an enterprise owns a pair of Nvidia H100s running the AI model Llama 3.3 with 70 billion parameters on NIM.

Benchmark data shows that if 250 concurrent end users submit simple prompts and receive simple responses of around 200 tokens each, the hardware can process around 5,067 tokens per second.

However, if those same users are permitted to include images or documents in their prompts and receive comprehensive responses — involving 20,000 input tokens and 2,000 output tokens — the number of tokens delivered per second drops to 427.

The difference in energy use between these scenarios is slight. These benchmarks are performed under conditions designed to achieve maximum throughput utilizing the hardware as effectively as possible.

Consequently, energy per token is lower in the high-throughput use case than in the lower-throughput use case. The hardware is the same, and the power is likely to be similar, but changes in user behavior or model configuration have now substantially increased the energy per token.

Capacity planning

To a large degree, throughput is determined by capacity planning. Enterprises aim to provision sufficient GPU (or other AI accelerator) capacity to respond effectively to end-user requests during peak demand, while avoiding excessive idle infrastructure during periods of lower demand.

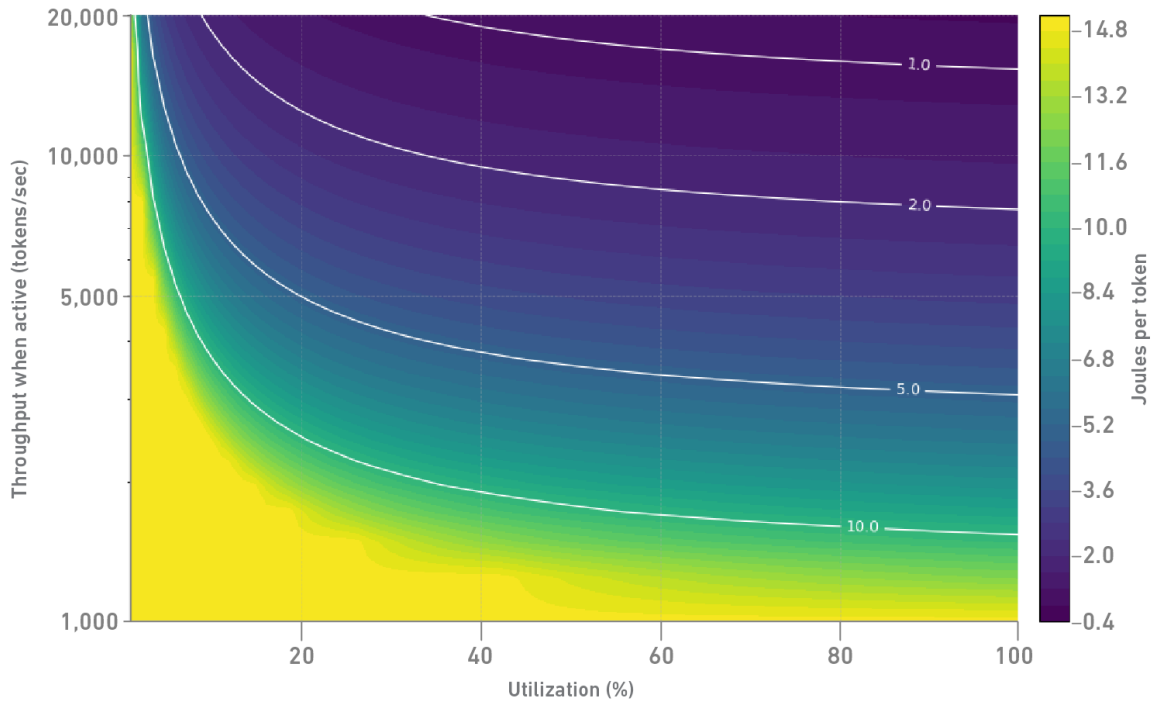
Another important factor in capacity planning is lifetime utilization. Throughput reflects how efficiently a model generates tokens during active inference periods, while utilization reflects how effectively the underlying hardware is utilized over its operational lifetime.

This distinction matters because energy-per-token metrics are influenced by both operational throughput during active periods and lifetime utilization. Even when inference demand is low, idle infrastructure still consumes power — Nvidia GPUs typically idle at around 10-20% of nameplate power. Systems designed to handle peak demand can generate substantially higher energy-per-token figures if they spend much of their operational lifetime underutilized.

Huge variance

Figure 1 shows the energy per token derived from the architecture and supporting data described in *Where to deploy AI inference: a guide to the economics*, based on a single DGX H100 system used for inference. Energy reflects IT energy for the cluster, with PUE factored in.

Figure 1 Heatmap showing token energy for varying utilization and throughput levels



UPTIME INSTITUTE 2026

Published energy-per-token figures for AI inference are typically derived from benchmarking exercises performed under controlled, ideal conditions by vendors and industry bodies. During these exercises, systems process continuous streams of requests using optimized batching, high concurrency, sustained demand, and tuned software stacks.

Benchmarks of H100 systems show energy intensities in the range of 0.3 to 1.5 joules per token, depending on model size, precision and workload characteristics. In **Figure 1**, these low levels of energy use appear in the top-right corner, around the 1-joule contour. They are achievable only with very high lifetime utilization and sustained token throughput.

Most enterprises will not achieve high levels on either metric. Reductions in lifetime utilization or throughput can substantially increase token energy. At a theoretical 100% utilization and a throughput of 20,000 tokens per second, the energy per token is 0.77 joules, in line with benchmark figures. But at a more realistic 5,000 tokens per second and 50% lifetime utilization, this increases fivefold to 3.5 joules per token.

Regardless of the specific hardware and power assumptions, the overall pattern remains true: energy per token is sensitive to both utilization and throughput.

Plan accordingly

Energy-per-token metrics are useful because they link infrastructure energy consumption directly to AI output. However, they should not be treated as fixed characteristics of hardware. In practice, energy per token is strongly influenced by operational factors including throughput during active inference periods, workload characteristics, use case and lifetime utilization. Published benchmark figures reflect highly optimized, high-utilization conditions that differ substantially from real-world enterprise deployments.

In practice, enterprise inference demand is often uneven and bursty. Infrastructure designed to

support peak throughput may spend significant portions of its operational life lightly utilized, while still consuming substantial power. As a result, real-world token energy intensity can be many times higher than benchmark figures suggest, even when using the same models and accelerators.

The challenge with energy-per-token metrics is that benchmark figures are often interpreted as properties of hardware, when in reality they are heavily shaped by how AI infrastructure is planned, operated and consumed. Enterprises should therefore base capacity decisions on both demand and capability.

The Uptime Intelligence View

As AI inference adoption grows, energy-per-token metrics will likely become more commonplace because they provide a simple way to link infrastructure energy use to the value generated by AI. However, the metric can be misleading if benchmark figures are interpreted as fixed properties of hardware rather than reflections of specific operating conditions. In practice, differences in throughput, use case, workload and utilization are likely to create substantial variation between theoretical and real-world efficiency, particularly as enterprises deploy inference infrastructure to handle uneven and unpredictable production demand.

ABOUT THE AUTHOR

Dr. Owen Rogers

20 May 2026

Dr. Owen Rogers is Uptime Institute's Senior Research Director of Cloud Computing. Dr. Rogers has been analyzing the economics of cloud for over a decade as a chartered engineer, product manager and industry analyst. Rogers covers all areas of cloud, including AI, FinOps, sustainability, hybrid infrastructure and quantum computing.

orogers@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. With over 4,000 awards issued in over 122 countries around the globe, and over 1,100 currently active projects in 80+ countries, Uptime has helped tens of thousands of companies optimize critical IT assets while managing costs, resources, and efficiency. For over 30 years, the company has established industry-leading benchmarks for data center performance, resilience, sustainability, and efficiency, which provide customers assurance that their digital infrastructure can perform across a wide array of operating conditions at a level consistent with their individual business needs. Uptime's Tier Standard is the IT industry's most trusted and adopted global standard for the design, construction, and operation of data centers.

Offerings include the organization's Tier Standard and Certifications, Management & Operations reviews and assessments including SCIRA-FSI financial sector risk assessment, the Sustainability Assessment, and a broad range of additional risk management, performance, availability, and related offerings. Uptime Education training programs have been successfully completed by over 100,000 data center professionals, such as the much-valued ATD (Accredited Tier Designer) and AOS (Accredited Operations Specialist). The Uptime Education curriculum has been expanded by the acquisition of CNet Training Ltd. In 2023.

Uptime Institute is headquartered in New York, NY, with offices in London, Sao Paulo, Dubai, Riyadh, and Singapore, and full-time Uptime professionals based in over thirty-four countries around the world.

For more information, visit www.uptimeinstitute.com