INTELLIGENCE UPDATE

AI’s growth calls for useful IT efficiency metrics

Jacqueline Davis

10 Oct 2025

The digital infrastructure industry is under pressure to measure and improve the energy efficiency of the computing work that underpins digital services. Enterprises seek to maximize returns on cost outlay and operating expenses for IT hardware, and regulators and local communities need reassurance that the energy devoted to data centers is used efficiently. These objectives call for a productivity metric to measure the amount of work that IT hardware performs per unit of energy.

With generative AI projected to boost data center power demand substantially, the stakes have arguably never been higher. Fortunately, organizations monitoring the performance and efficiency of their AI applications can benefit from experiences in the field of supercomputing.

In September 2025, Uptime Intelligence participated in a panel discussion about AI energy efficiency at the Yotta 2025 conference in Las Vegas (Nevada, US). The panelists drew on their extensive experience in supercomputing to weigh in on discussions around AI training efficiency. They discussed the need for a productivity metric to measure it, as well as a key caveat organizations need to consider.

Organizations such as Uptime Intelligence and The Green Grid have published guidance on calculating work capacity for various types of IT. Software applications and their supporting IT hardware vary significantly, so consensus on a single metric to compare energy performance remains out of reach for the foreseeable future. However, tracking energy performance in a given facility over time is important, and is achievable practically for many organizations today.

Defining AI computing work

The work capacity of IT equipment is needed to calculate its utilization and energy performance when running an application. The Green Grid white paper IT work capacity metric V1 — a methodology provides a methodology for calculating a work capacity value for CPU-based servers. Uptime Intelligence has proposed methodologies to extend this to accelerator-based servers for AI and other applications (see *Calculating work capacity for server and storage products*).

Floating point operations per second (FLOPS) is a common and readily available unit of work capacity for CPU- or accelerator-based servers. In 2025, an AI server's capacity usually ranks in the trillions of FLOPS, or teraFLOPS (TFLOPS).

Not all FLOPS are the same

Even though large-scale AI training is radically reshaping many commercial data centers, the underlying software and hardware are not fundamentally new. AI training is essentially one of many applications of supercomputing. Supercomputing software, along with the IT selection and configuration, varies in many ways — and one of the most relevant variables when monitoring energy performance is floating point precision. This precision (measured in bits) is analogous to the number of decimal places used in inputs and outputs.

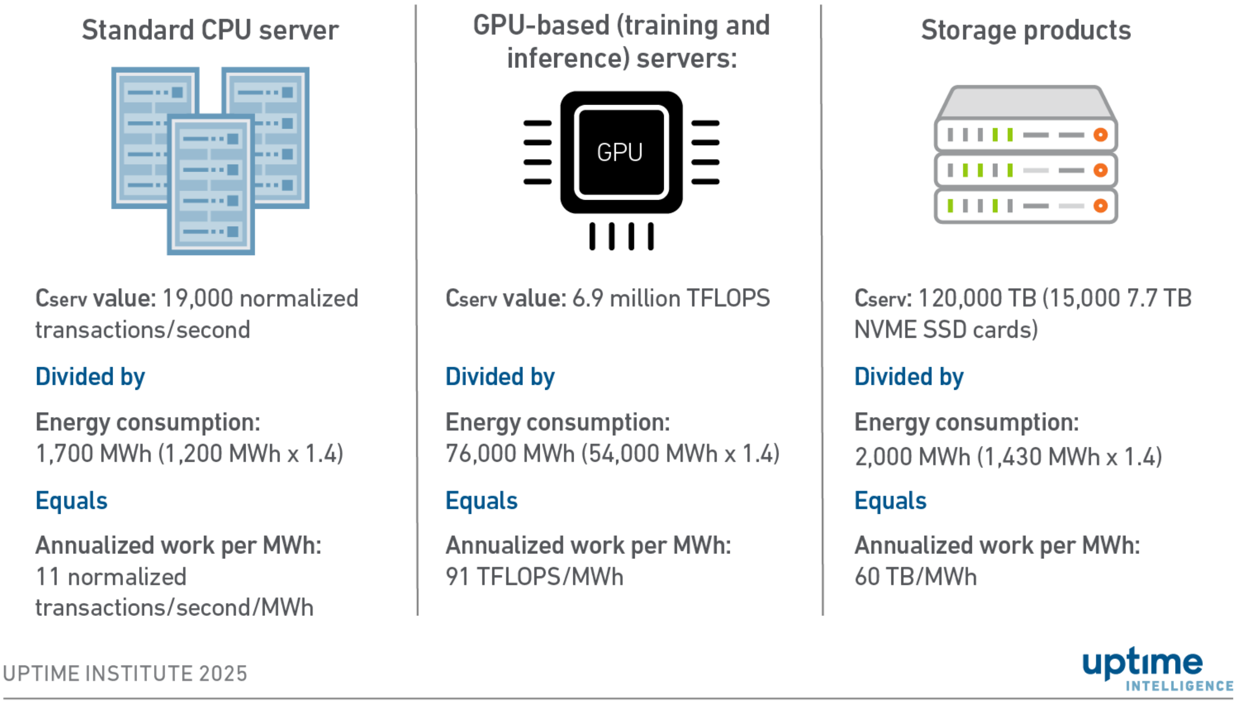
GPUs and other accelerators can perform 64-, 32-, 16-, 8- and 4-bit calculations, and some can use mixed precision. While an high-performance computing (HPC) workload such as computational fluid dynamics might use 64-bit ("double precision") floating point calculations for high accuracy, other applications do not have such exacting requirements. Lower precision consumes less memory per calculation — and, crucially, less energy. The panel discussion at Yotta raised an important distinction: unlike most engineering and research applications, today's AI training and inference calculations typically use 4-bit precision.

Floating point precision is necessary information when evaluating a TFLOPS benchmark. A 64-bit precision calculation TFLOPS value is one-half of a 32-bit TFLOPS value— or one-sixteenth of a 4-bit TFLOPS value. For consistent AI work capacity calculation, Uptime Institute recommends that IT operators use 32-bit TFLOPS values supplied by their AI server providers.

Working it out: work per energy

The maximum work capacity calculation for a server can be aggregated at the level of a rack, a cluster or a data center. Work capacity multiplied by average utilization (as a percentage) produces an estimate of the amount of calculation work (in TFLOPS) that was performed over a given period. Operators can divide this figure by the energy consumption (in MWh) over that same time to yield an estimate of the work's energy efficiency, in TFLOPS/MWh. Separate calculations for CPU-based servers, accelerator-based servers, and other IT (e.g., storage) will provide a more accurate assessment of energy performance (see **Figure 1**).

Figure 1 Examples of IT equipment work-per-energy calculations



Standard CPU server

Cserv value: 19,000 normalized transactions/second

Divided by

Energy consumption: 1,700 MWh (1,200 MWh x 1.4)

Equals

Annualized work per MWh: 11 normalized transactions/second/MWh

GPU-based (training and inference) servers:

Cserv value: 6.9 million TFLOPS

Divided by

Energy consumption: 76,000 MWh (54,000 MWh x 1.4)

Equals

Annualized work per MWh: 91 TFLOPS/MWh

Storage products

Cserv: 120,000 TB (15,000 7.7 TB NVME SSD cards)

Divided by

Energy consumption: 2,000 MWh (1,430 MWh x 1.4)

Equals

Annualized work per MWh: 60 TB/MWh

UPTIME INSTITUTE 2025

Even when TFLOPS figures are normalized to the same precision, it is difficult to use this information to draw meaningful comparisons between the energy performance of significantly different hardware types and configurations. Accelerator power consumption does not scale linearly with utilization levels. Additionally, the details of software design will determine how closely real-world application performance aligns with simplified work capacity benchmarks.

However, many organizations can benefit from calculating this TFLOPS/MWh productivity metric and are already well equipped to do so. This calculation is most useful to quantify efficiency gains over time, e.g., from IT refresh and consolidation, or refinements to operational control. In some jurisdictions, tracking FLOPS/MWh as a productivity metric can satisfy some regulatory requirements. IT efficiency is often overlooked in favor of facility efficiency — but a consistent productivity metric can help to quantify available improvements.

The Uptime Intelligence View

Generative AI training is poised to drive up data center energy consumption, prompting calls for regulation, responsible resource use and return on investment. A productivity metric can help meet these objectives by consistently quantifying the amount of computing work performed per unit of energy. Supercomputing experts agree that operators should track and use this data, but they caution against interpreting it without the necessary context. A simplified, practical work-per-energy metric is most useful for tracking improvement in one facility over time.

The following participants took part in the panel discussion on energy efficiency at Yotta 2025:
- Jacqueline Davis, Research Analyst at Uptime Institute (moderator)
- Dr Peter de Bock, former Program Director, Advanced Research Projects Agency–Energy
- Dr Alfonso Ortega, Professor of Energy Technology, Villanova University
- Dr Jon Summers, Research Lead in Data Centers, Research Institutes of Sweden

Other related reports published by Uptime Institute include:

Calculating work capacity for server and storage products

The following Uptime Institute experts were consulted for this report:

Jay Dietrich, Research Director of Sustainability, Uptime Institute

ABOUT THE AUTHOR

Jacqueline Davis

Jacqueline is a Research Analyst at Uptime Institute covering global trends and technologies that underpin critical digital infrastructure. Her background includes environmental monitoring and data interpretation in the environmental compliance and health and safety fields.

jdavis@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.