INTELLIGENCE UPDATE

AI super-densification: how far will it really go?

Daniel Bizo

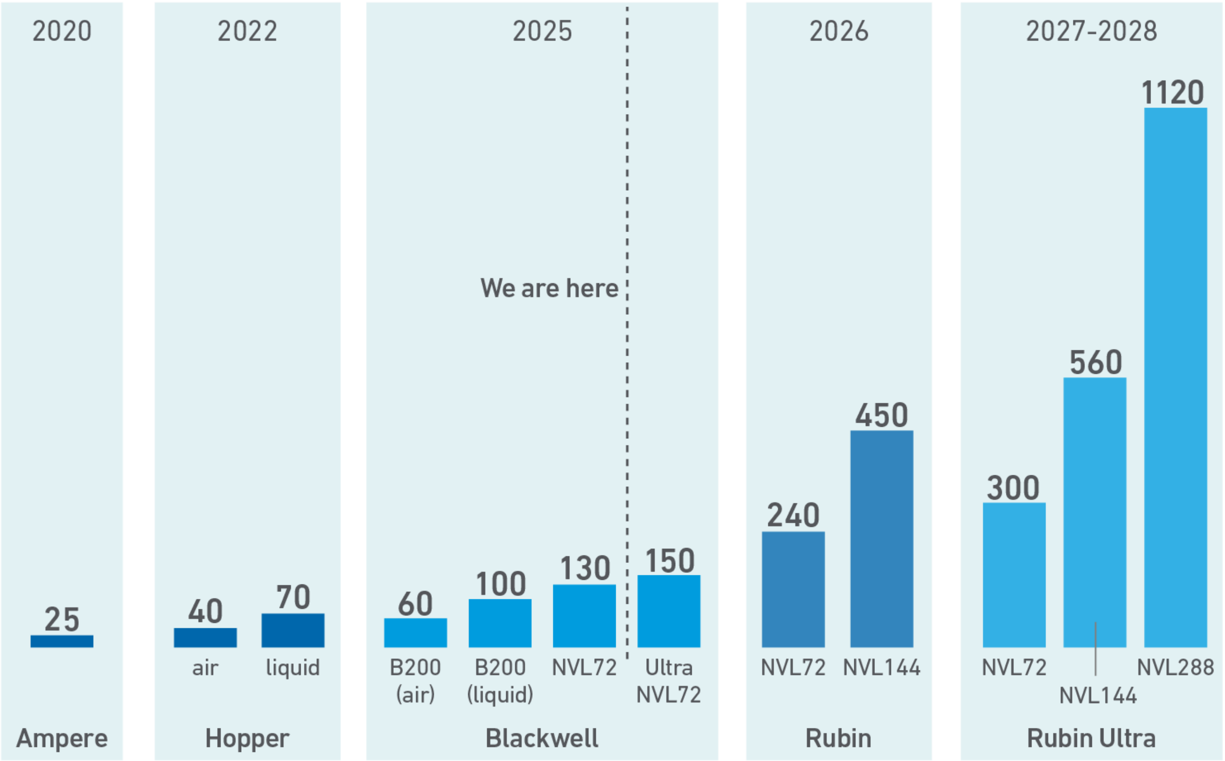
28 Jul 2025

Uptime Intelligence first observed that server silicon power was going to follow a steep upward trajectory in 2021. We anticipated the emergence of today’s processors with 500 watt thermal power and GPUs with ratings exceeding 1 kilowatt (kW) — and even more powerful products for the future. This was not prophecy, but the extrapolation of where silicon power is headed, pushed on by the combined forces of semiconductor physics and the pursuit of absolute performance.

Uptime Intelligence has long anticipated that realized rack densities would begin to climb, largely in line with escalating server power levels. When a new mainstream server is capable of reaching 1 kW at full load, rack power can quickly exceed 10 kW at partial load, even if not fully populated. This shift is now visible across the industry. An Uptime Intelligence report published in 2022 projected that 40 kW racks will become much more common in densified infrastructure, characterizing this as “ultra-high density“ (see *Silicon heatwave: the looming change in data center climates*).

Little did we know that, only three years later, handling racks exceeding 100 kW would become a real-world technical requirement for dozens of new facilities; let alone that megawatt-racks (although with footprints significantly larger than the standard 19”) would be considered a plausible option. **Figure 1** shows Uptime Intelligence’s estimates (compiled from various sources) of the power rating for dense GPU racks, as envisioned by Nvidia. Power density figures for future rack systems (Blackwell Ultra, Rubin, Rubin Ultra generations) are unofficial estimates and are subject to change as product development progresses.

Figure 1 Estimates of high-density rack configurations across Nvidia GPU generations (unofficial estimates)



(Note: Names Ampere, Hopper, Blackwell and Rubin are Nvidia development codenames refer to generations of GPUs, not a specific product. The numbers in the NVL rack-scale system designations are kept consistent across generations to signify the number of GPU modules in the rack, deviating from Nvidia's official nomenclature.)

UPTIME INSTITUTE 2025 (VARIOUS INDUSTRY SOURCES)

Enter the era of hyperscale AI supercomputers that will use 1 megawatt (MW) racks before the end of the decade. If current plans hold, these AI training systems will be even more powerful and denser than the current cutting-edge exascale systems (at 20 MW to 30 MW of IT load each, and at up to 400 kW per custom rack). These systems are funded by the US government to run advanced simulations in nuclear physics and scientific research. Is this level of hyperscale supercomputing really imminent? Could we see racks that are 100 times denser than the current industry average within a few years?

GPU-interconnection is all that is needed

Numerous technical and business-related issues are involved in addressing the above questions. Data center design needs a major overhaul to support hundreds of kilowatts per rack at scale — let alone a megawatt. Effective planning starts with understanding the layout of the site, the size of the electrical infrastructure relative to the IT space and the need for more extensive use of medium voltage in site power distribution. The IT floor will change dramatically too, as systems with 1 MW racks will require an external power supply cabinet feeding high-voltage direct current to compute workloads. These systems will also need a coolant distribution unit for every one or two compute racks to remove all the generated heat.

Purely technically, it is all possible. The fundamental question is not how, but why? Why would 1 MW racks be so compelling that they justify the considerable costs of overcoming the related facility design difficulties?

It comes down to sharing data at scale. In the case of training large generative pre-trained transformer (hence the acronym GPT) models, sharing data at high speed across hundreds (or thousands) of GPUs is essential to performance. This need for inter-GPU communication stems from the fundamental architecture of generative AI models: learning the complex patterns of language (in the case of large language models) across vast contexts, where the presence of every word influences the meaning and relationship between words.

This feature, called global attention mechanism, makes modern models outperform previous neural network architectures in natural language processing and many other tasks. In compute terms, this mechanism has introduced massive parallelism into the training process. Ever-larger generative models are being trained on ever-larger compute clusters because it is now possible to improve model performance through scale effectively.

The successful processing of the training steps performed by any GPU depends on data being held across potentially thousands of other GPUs. This processing requires extreme network bandwidth and ultra-low latency. Every additional inch of copper wire or optical cable results in more stalling for the GPUs and more power wasted. When GPUs are spread across multiple racks, it can penalize performance and efficiency substantially. This is what drives the engineering choices behind the close coupling of GPUs (or any other types of processors), which in turn pushes rack power density toward 1 MW and beyond.

Finance will rule again

The key consideration that will counteract unrestrained densification is economics. Engineering extreme hardware systems is extremely expensive — only Nvidia knows how expensive — but the annual investment required in the development of high-end supercomputing architectures is likely to be hundreds of millions of dollars. Today, the total market for large AI training compute workloads (and other types of supercomputing workloads) is worth tens of billions annually — large enough to absorb enormous non-recurring engineering costs at Nvidia, and other silicon and system developers.

However, high-end proprietary architectures are only advantageous in systems with many thousands of processing units, GPUs or other types, where they attain substantial (reaching well into double-digit percentages) performance gains compared with less richly interconnected designs. At more moderate system sizes, performance gains become incremental, making the facility infrastructure and IT system cost premium (from more expensive chip fabrication and packaging to system manufacturing and rack integration) less attractive. Shaving a few minute off a training pass due to denser GPU-interconnectivity will rarely, in most cases, be valuable enough to justify the added infrastructure engineering effort and facility capital expenditure.

Although the total market size for AI clusters will continue to allow Nvidia and others to recoup significant research and development investments, most customers and use cases will not need AI supercomputers with thousands of GPUs. It is also worth noting that future generations of AI supercomputers will deliver several times more performance per GPU. This means the training performance of large AI clusters spanning hundreds of 40 kW compute racks installed in 2023 and 2024 (based on Nvidia’s H-series GPUs) will be matched by a handful of compute cabinets in 2028/2029 at a fraction of the footprint.

The direct financial and overall economic viability of AI supercomputers with 1 MW compute racks largely hinges on the development trajectory of AI models. The size of the market segment for hyperscale AI systems will be defined by the practical value of being able to efficiently handle models that are an order of magnitude larger and more complex than today’s largest compute systems.

There is no consensus on the practical scalability limits of existing generative model architectures — a blurred line beyond which improvements to model performance become marginal. It is not simply a matter of the sheer scale and complexity of the most powerful AI models, but also architectural decisions, such as monolithic versus modular (mixture-of-experts) neural networks, tunings to computational precision (resource use versus response quality) and batch sizes of the training passes, as well as the number of passes on the same training data.

Presently, there is no evidence of a profitable business model for state-of-the-art AI products (costing hundreds of millions or billions of dollars to develop) or that scaling GPT models another 10 times to 100 times would improve the financial viability of these offerings. If anything, the economics of leading-edge models are becoming more difficult due to ballooning compute intensity (and cost) of inference. Consequently, planning and assessing investments in data centers for hyperscale AI supercomputers are built on counterfactual scenarios.

Yet, even without a solid business case, there is no sign of the surge in investments into ever-larger AI models slowing. As a result, the likely scenario is that densification of AI supercomputing will continue. Some in the data center industry consider it as almost a certainty that the coming years will see racks rated at hundreds of kilowatts (exceeding greater than 50 kW power per square meter / 4.6 kW per square foot in the data hall) in volume deployments, although these will remain limited to a relatively few number of specialist AI training facilities. Even if 1 MW racks become more of a technical curiosity than a commercially viable product shipping in volume, dramatic densification seems inevitable — for now.

Future Uptime Intelligence reports will examine the engineering and technical innovations required to accommodate this increasing densification.

The Uptime Intelligence View

Few in the data center industry expected to be discussing the possibility of megawatt-scale racks only a couple of years ago. No matter how improbable this outlook may seem, the drive for denser AI supercomputers and the facilities housing them will persist, fueled by the billions of dollars in speculative investments aimed at developing enormous AI models. Unless these investments drop sharply or the current trajectory in AI hardware development changes, 1 MW racks may prove to be more than ambitious speculation. Either way, a bout of intense research and development activity in both IT and facility equipment in response to this potential outcome will bring benefits to the broader data center industry.

Other related reports published by Uptime Institute include:
AI power fluctuations strain both budgets and hardware
Electrical considerations with large AI compute
Data center AI strategies are mixed in early 2025
GPU power management is a work in progress

ABOUT THE AUTHOR

Daniel Bizo

Over the past 15 years, Daniel has covered the business and technology of enterprise IT and infrastructure in various roles, including industry analyst and advisor. His research includes sustainability, operations, and energy efficiency within the data center, on topics like emerging battery technologies, thermal operation guidelines, and processor chip technology.

dbizo@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.