INTELLIGENCE UPDATE

Hardware for AI: options and directions

Max Smolaks

7 Mar 2025

Developments in AI research continue to raise questions about the design and features of the data centers required to deliver AI-based services at scale.

AI is not a uniform workload, and the infrastructure required to deliver AI will be far from uniform. At least three approaches to server hardware for AI have emerged to support different model types at various stages of their life cycle — with often contrasting data center requirements.

The first approach, focused on high-density GPU deployments, places a much higher-than-average strain on power and cooling equipment, requiring either purpose-built data centers or a complete redesign of legacy sites.

The second approach distributes AI accelerator resources across existing data halls, increasing average rack densities while often fitting into existing data center specifications.

The third approach prioritizes efficient inference, using specialist chips that consume less power than CPUs or GPUs and servers that are in line with the average rack densities of enterprise IT.

All three types of infrastructure will be essential as businesses attempt to balance the cost and revenue opportunities emerging from generative AI.

Division of labor

When equipping data centers for enterprise IT, organizations can choose from dozens of CPU models and hundreds of server configurations across on-premises, colocation, cloud and edge environments. The modern IT stack has evolved to support all budgets and levels of quality of service. A mature AI hardware market will work in the same way.

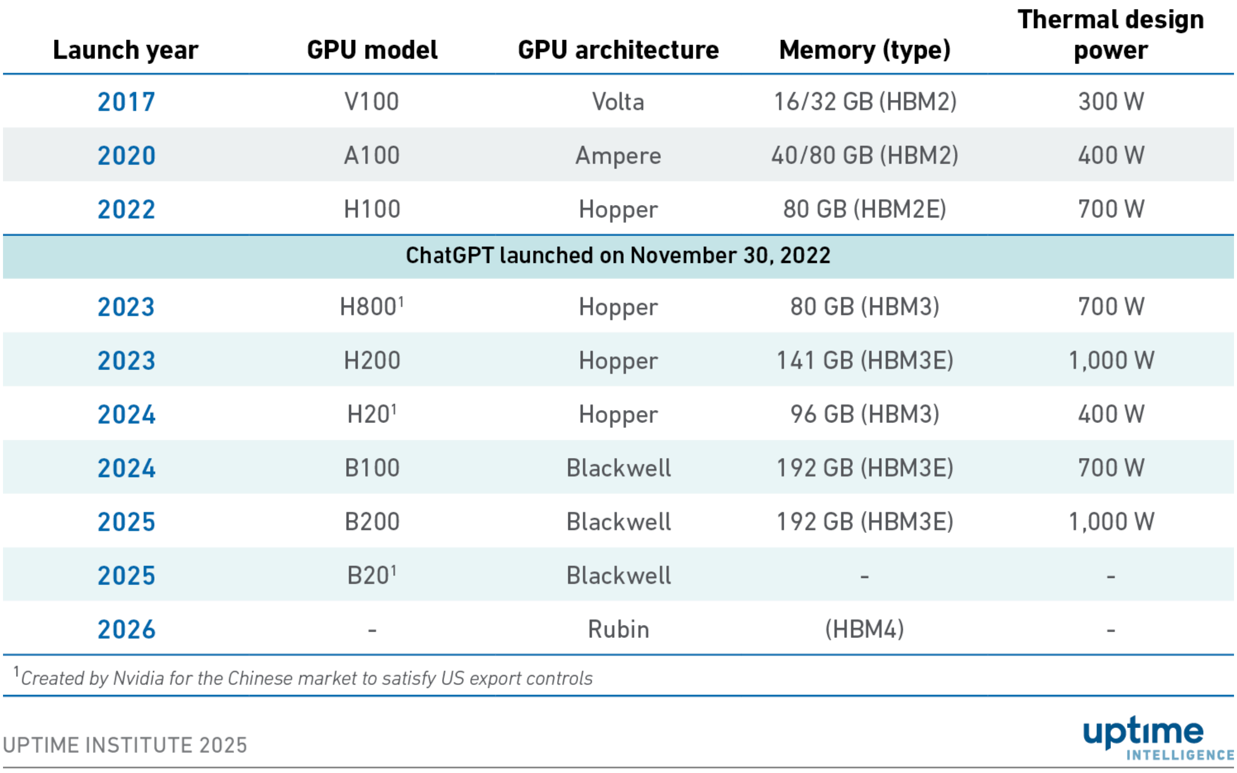
Workload diversity is built into every machine learning (ML) model's life cycle. Training the model — running algorithms on a dataset and configuring them to identify patterns — typically requires large clusters of specialized servers to maximize compute, network and storage performance.

Inference, the process of running new data through a trained model to calculate an output, typically involves a single chip — either a CPU, GPU or another type of AI accelerator. Here, the focus of the infrastructure shifts to serving requests economically and efficiently at the targeted token generation latency — a complex balancing act.

This means there are at least two distinct hardware markets for AI: training and inference. Further adding to workload diversity are non-generative AI models that predate ChatGPT. These include models for computer vision, speech recognition, recommendation systems, older chatbots and many other applications that have modest infrastructure requirements and can be developed or deployed with just a few servers.

Many AI accelerator products — including all GPUs — can be used for both training and inference. These are listed in **Tables 1** and **2**. A rapidly growing category of accelerators is designed exclusively for inference (see **Table 3**).

Table 1 Flagship general-purpose GPUs launched by Nvidia since 2017 (SXM form factor)

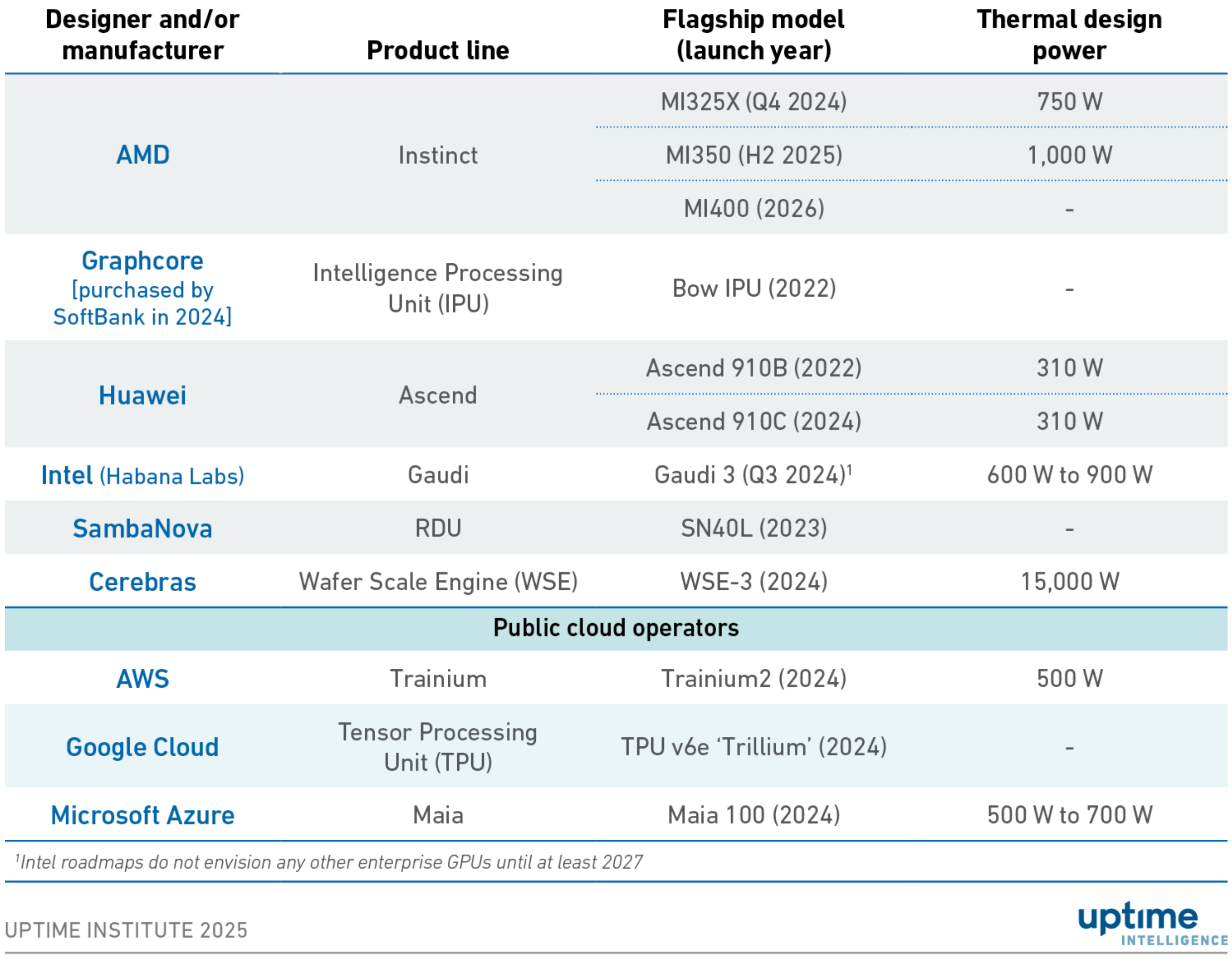


Launch year	GPU model	GPU architecture	Memory (type)	Thermal design power
2017	V100	Volta	16/32 GB (HBM2)	300 W
2020	A100	Ampere	40/80 GB (HBM2)	400 W
2022	H100	Hopper	80 GB (HBM2E)	700 W
ChatGPT launched on November 30, 2022				
2023	H800[1]	Hopper	80 GB (HBM3)	700 W
2023	H200	Hopper	141 GB (HBM3E)	1,000 W
2024	H20[1]	Hopper	96 GB (HBM3)	400 W
2024	B100	Blackwell	192 GB (HBM3E)	700 W
2025	B200	Blackwell	192 GB (HBM3E)	1,000 W
2025	B20[1]	Blackwell	-	-
2026	-	Rubin	(HBM4)	-

[1] *Created by Nvidia for the Chinese market to satisfy US export controls*

UPTIME INSTITUTE 2025

Table 2 Alternative platforms for AI model training and inference



Designer and/or manufacturer	Product line	Flagship model (launch year)	Thermal design power
AMD	Instinct	MI325X (Q4 2024)	750 W
		MI350 (H2 2025)	1,000 W
		MI400 (2026)	-
Graphcore [purchased by SoftBank in 2024]	Intelligence Processing Unit (IPU)	Bow IPU (2022)	-
Huawei	Ascend	Ascend 910B (2022)	310 W
		Ascend 910C (2024)	310 W
Intel (Habana Labs)	Gaudi	Gaudi 3 (Q3 2024)[1]	600 W to 900 W
SambaNova	RDU	SN40L (2023)	-
Cerebras	Wafer Scale Engine (WSE)	WSE-3 (2024)	15,000 W
Public cloud operators			
AWS	Trainium	Trainium2 (2024)	500 W
Google Cloud	Tensor Processing Unit (TPU)	TPU v6e 'Trillium' (2024)	-
Microsoft Azure	Maia	Maia 100 (2024)	500 W to 700 W

[1] Intel roadmaps do not envision any other enterprise GPUs until at least 2027

UPTIME INSTITUTE 2025

Table 3 Alternative hardware choices for AI inference only

Designer and / or manufacturer (country)	Product line	Flagship model (launch year)	Thermal design power
Groq (US)	LPU/GroqRack	-	-
Enflame (China)	DTU	Yunsui i20 (2021)	-
Qualcomm (US)	Cloud AI	Cloud AI 100 Ultra (2023)	150 W
Untether (Canada)	speedAI/tsunAImi	speedAI240 Slim (2024)	75 W
Blaize (US) [IPO via SPAC in 2025]	Pathfinder / Xplorer	-	-
FuriosaAI (South Korea)	-	RNGD (early 2025)	150 W
Recogni (US)	Pareto	-	-
Etched.ai (US)	Sohu	-	-
Tenstorrent (US)	Grayskull/Wormhole	Grayskull e150 (2024)	200 W
D-Matrix (US)	Corsair	Corsair (2024)	-
Cambricon (China)	MLU	Xuansi 1000 (2021)	-
Public cloud operators			
AWS (US)	Inferentia	Inferentia2 (2023)	-

UPTIME INSTITUTE 2025

Looking at the devices listed above, it is evident that the AI hardware market is developing in at least three directions, depending on customers' objectives.

Objective 1: maximizing compute and bandwidth

The first approach — for developing the largest and most complex models — involves tightly integrated rack-scale systems and appliances to build clusters with thousands of GPUs, all connected by ultra-fast networks.

This infrastructure can require 5 to 20 times more power per rack than typical enterprise IT systems, forcing the adoption of liquid cooling and necessitating more gray space for mechanical and electrical equipment.

The high-density approach is exemplified by the Nvidia NVL72 rack-scale compute system, which consumes up to 132 kW per rack and uses direct liquid cooling to remove the bulk of the heat. The category also includes dense deployments of server systems such as the Nvidia DGX B200 and its derivatives, which pack eight GPUs in a single machine that consumes up to 15 kW in a 10U air-cooled chassis and up to 12 kW in a 4U liquid-cooled version.

High-density AI server systems are preferred by well-funded "frontier" AI labs, such as OpenAI and Anthropic, as well as hyperscale operators that prioritize speed in developing new AI-based capabilities over infrastructure costs. While large AI training workloads exhibit some unique characteristics, facilities equipped to host them can also support traditional high-performance computing systems used for various technical development and academic research workloads, such as computational chemistry, fluid dynamics, 3D rendering and weather prediction.

Objective 2: making AI easy to deploy and use

The second approach to delivering AI involves less dense server clusters, with AI accelerators distributed across the data hall. These clusters tend to provide less compute capacity — hence less pressure to densify — but are easier to accommodate. They remain predominantly air-cooled and can be handled with standard power delivery and networking equipment.

Many, if not most, GPUs available from public cloud vendors are deployed in this manner. This is likely to be the preferred deployment model for most enterprise AI training workloads, which do not require supercomputing levels of performance. Furthermore, such infrastructure will be just as capable in inferencing workloads as denser, more expensive systems.

The key to this deployment model is flexibility — by using individual accelerator-equipped servers rather than full racks of GPU appliances or rack-scale systems, customers can deploy AI at any level of power density. This approach is a good fit for delivering specific generative AI capabilities such as translation, transcription or summarization, and smaller language models.

Such AI clusters can be built using GPUs from either AMD or Nvidia, or with non-GPU accelerators from challengers such as Graphcore, SambaNova and Cerebras (see **Table 2**).

Intel’s Gaudi is another option, but the cancellation of its next iteration (codenamed Falcon Shores) means Intel will not release another AI accelerator until at least 2027 — effectively ceding the market to faster-moving competitors.

Cloud vendors’ silicon designs for AI, such as Trainium from AWS or Maia from Microsoft, also fit into this category — competing with GPUs on price-to-performance ratio, rather than absolute performance.

Objective 3: efficiency and affordability

Inference is arguably the most important stage of the AI model life cycle — while training only incurs costs for the model owner, inference generates revenue or some other business value. To maximize returns, businesses will want to minimize the cost of delivering inference.

Training workloads benefit from high-density server systems, while inference performance is primarily defined by the size of the model and how it runs on individual chips; increased hardware density brings capacity benefits but does not improve performance.

Inference accelerators are much cheaper than the devices used for training. They rely on simplified silicon architectures and slower types of memory. They also have lower power consumption and are often manufactured using previous generation (i.e., mature, cost-effective) chip fabrication processes.

Some inference accelerators are optimized for specific ML features (e.g., transformers) or workloads, such as computer vision. Many are sold as PCIe cards, allowing them to be added to almost any server, unlike the SXM form factor used by flagship GPUs, which requires purpose-designed AI servers.

Much of the inference hardware development is driven by vendors that have emerged in the past five years. While many are startups, some have grown into public companies.

Starting in 2023, many Intel Xeon and AMD Epic CPUs have been enhanced with silicon and firmware features to improve inference performance. While GPUs can be thought of as throughput machines, designed to process large amounts of data in parallel, CPUs are latency machines, optimized for quick responses. This makes CPUs suitable for high throughput, low complexity AI applications.

Today, most data center operators focus on training, but their attention will shift to inference as organizations adopt AI-based products and services. It is likely that in time, every large data center will run at least some inference workloads. For many small to midsized AI models in low-throughput applications (e.g., in-house enterprise uses), inference will likely gravitate toward standard server processors, which are readily available, affordable and the most attractive option for software developers.

A matter of balance

Delivering AI at scale will require a variety of hardware platforms and facilities, and that is without major disruptions to mainstream ML architectures (see *The DeepSeek paradox: more efficiency, more infrastructure?*).

Organizations' choices of hardware and venue for training will depend on the model's size and complexity, while choices of inference infrastructure will depend primarily on the speed and latency requirements of the AI-based services being delivered.

Underlying both will be concerns about the return on investment in AI, prompting enterprises to consider alternatives to flagship GPUs — platforms that offer lower performance levels at a lower cost.

Operators seeking to future-proof their facility designs will have to consider all three deployment models and avoid being distracted by case studies of extreme performance deployments.

The Uptime Intelligence View

AI will require at least three distinct data center categories to become a mainstream workload.

This means it will need the participation of operators at all levels of scale and sophistication.

Large, purpose-designed campuses will continue to drive AI innovation, but data centers currently serving enterprises will be key to widespread adoption. These facilities will need to accommodate AI workloads with the resources they have.

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.