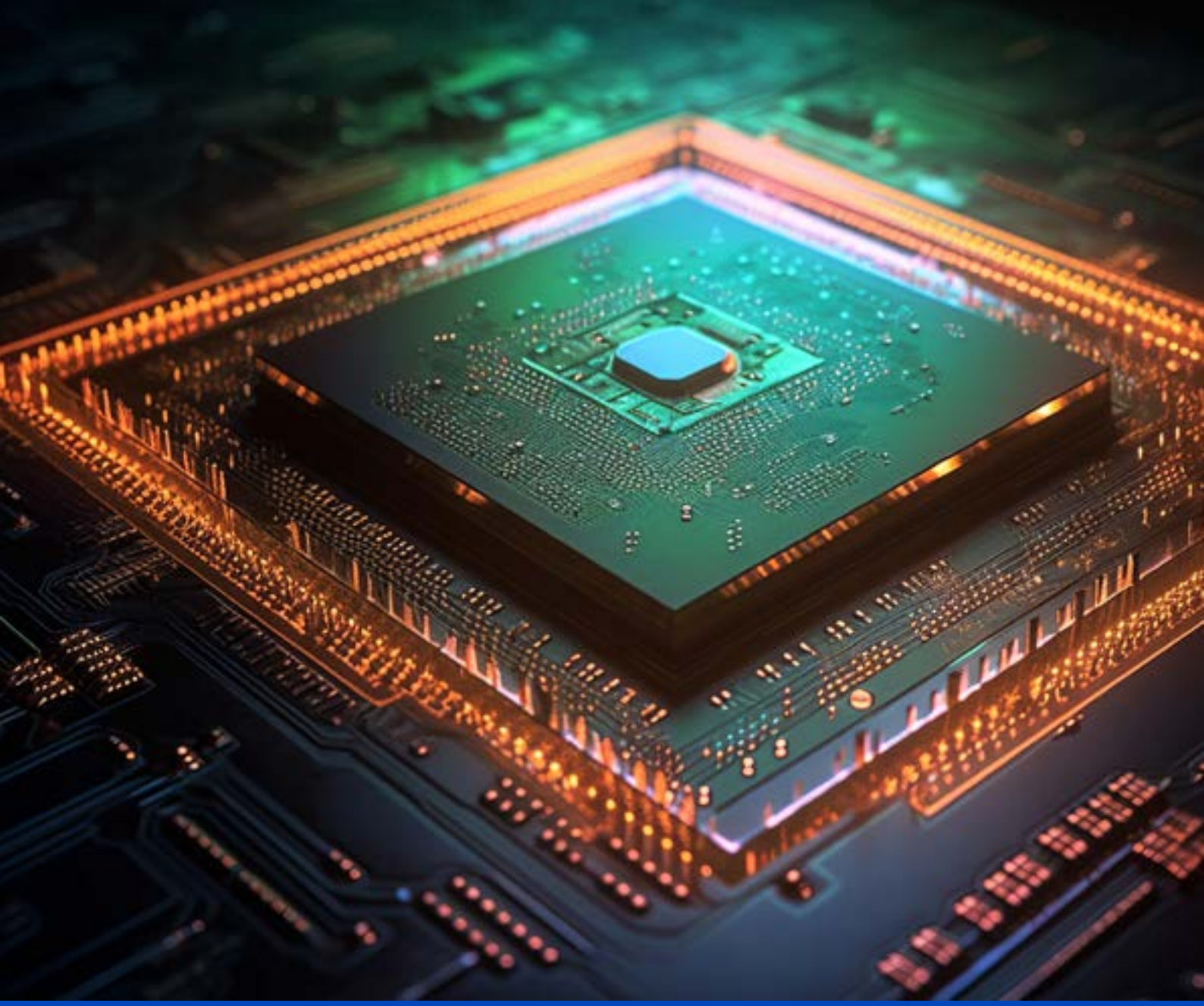


The case for AI Infrastructure

WisdomTree®

WisdomTree.eu
+44 (0) 207 448 4330

The case for AI Infrastructure: the physical foundation for a digital future

Artificial intelligence (AI) infrastructure is becoming a foundational layer of the global economy. As model intelligence improves and AI agents begin to take their share of digital workloads, demand for compute is rising at a rapid pace.

Value is no longer captured solely by GPUs[1] or at the model layer, but in the computing ecosystem that surrounds and enables them. AI is evolving into a physical system that depends on energy, hardware and increasingly complex systems and supply chains. This is driving sustained expansion across:

Semiconductors and manufacturing

Networking

Data centres and power

1 GPU: Graphics Processing Unit. Originally designed for rendering graphics, GPUs are now the primary processors used for AI training and inference due to their parallel processing capabilities.

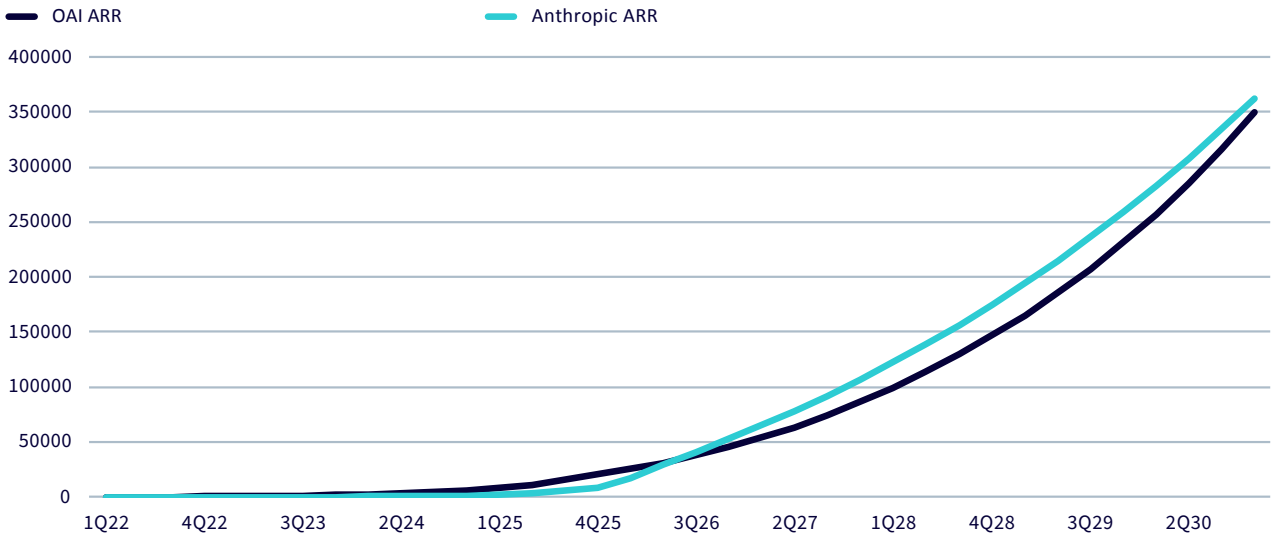
Inference demand and compute scaling

The defining feature of this cycle is the nature of demand. Demand for AI compute is now visible and persistent across enterprise and consumer use cases. Training a model requires immense compute for a fixed period, while inference is continuous.

At the most basic level, this demand is measured in tokens, the units of text, code, or data processed by AI models. Every query, workflow and agent-driven task generates tokens, each requiring compute to process. As usage grows, token generation scales directly with it, creating structural demand for compute.

This is reflected in the growth of the leading AI model labs, the 'producers of AI tokens'.

Figure 1: Annualised revenue of leading LLM[2] providers



Source: SemiAnalysis, April 2026. **Forecasts are not an indicator of future performance and any investments are subject to risks and uncertainties. Historical performance is not an indication of future performance and any investments may go down in value.**

Annualised recurring revenues (ARR)[3] of AI model providers, such as OpenAI and Anthropic, have moved from zero to over $30 billion within just a few years. The pace of adoption exceeds prior technology cycles[4] and follows an exponential growth curve, with SemiAnalysis' forecasts reaching over $300 billion in ARR by 2030.

Hyperscalers[5] and neocloud[6] providers supply much of the compute these AI labs rely on, while

2 LLM: Large Language Model. A type of AI model trained on large datasets of text to generate, understand and reason over language. Examples include OpenAI's GPT series and Anthropic's Claude.

3 ARR: Annualised Recurring Revenue. A normalised measure of predictable, recurring revenue on an annual basis, commonly used to track the growth of subscription-based technology businesses.

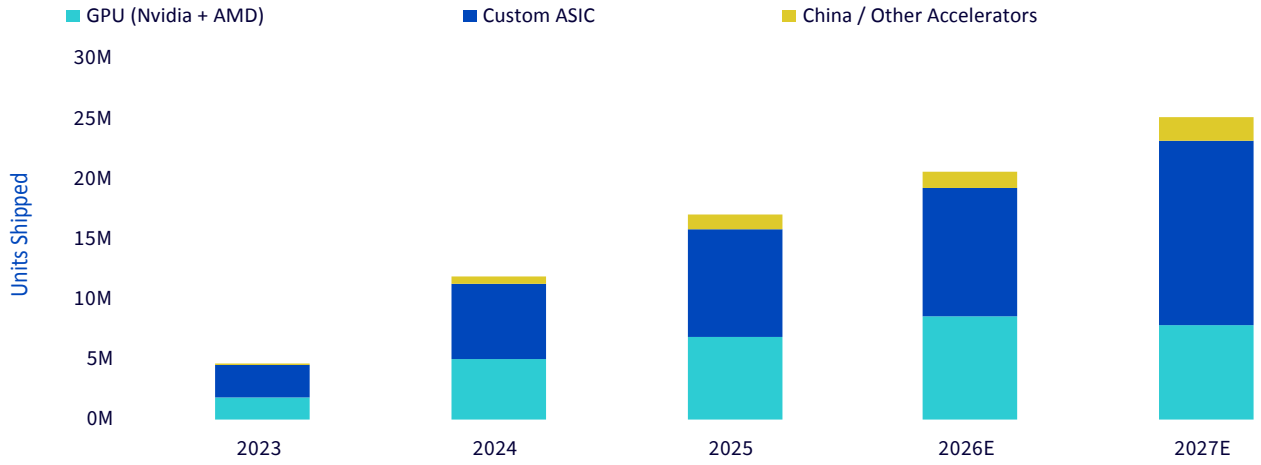
4 ChatGPT was widely recognised as the fastest-growing consumer application in history upon its release, reaching 100 million monthly active users in just two months.

5 Hyperscalers are large cloud service providers that operate enormous, global data centre infrastructures.

6 Neoclouds are specialised, AI-focused cloud service providers offering on-demand, high-performance GPU-as-a-Service for AI training and inference.

also building their own competing models. Both sides are committing hundreds of billions in capital expenditure to expand capacity. These are forward-looking decisions with multi-year horizons, focused on deploying increasingly specialised chips for maximal efficiency. This is resulting in new chip designs and growing orders to manufacturing partners.

Figure 2: XPU[7] unit shipment



Source: SemiAnalysis Accelerator Model, April 2026. **Forecasts are not an indicator of future performance and any investments are subject to risks and uncertainties. Historical performance is not an indication of future performance and any investments may go down in value.**

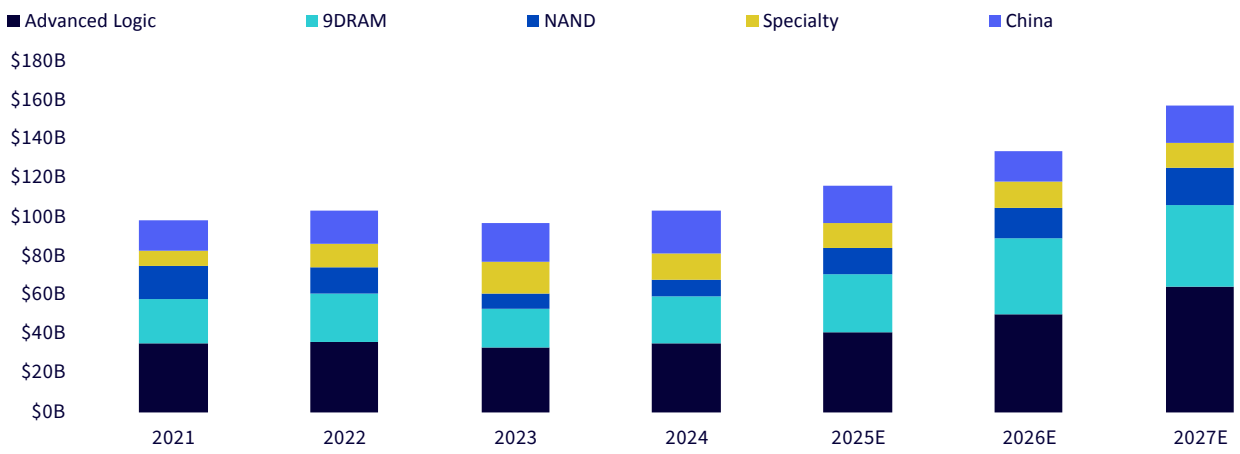
AI accelerator shipments are expected to scale from roughly 5 million units back in 2023 to over 25 million by 2027, increasing 5x over the period. This reflects both the growing need to support AI computing workloads and a broadening ecosystem beyond a single supplier or architecture.

Semiconductor scaling and supply chain bottlenecks

Meeting this demand for chips requires a major expansion in semiconductor manufacturing. Today, foundries do not have enough capacity to produce the volume that customers need. Building that capacity means constructing new facilities and investing in the specialised equipment, known as wafer fabrication equipment (WFE), used at each stage of turning a raw silicon wafer into a finished chip.

7 XPU: A general term for any accelerator processing unit used in AI workloads, including GPUs, custom ASICs and other specialised silicon.

Figure 3: Global WFE[8] spend by application



Source: SemiAnalysis WFE Model, Feb 2026. **Forecasts are not an indicator of future performance and any investments are subject to risks and uncertainties. Historical performance is not an indication of future performance and any investments may go down in value.**

Global Wafer Fabrication Equipment (WFE) spend is expected to approach $160 billion by 2027, reflecting a structural shift toward AI-driven capacity expansion. Despite this increase, key constraints persist, particularly in advanced lithography[9] and cleanroom[10] capacity, which continue to limit throughput.

At the same time, the chips themselves are becoming more complex. Each new generation of AI accelerator needs more memory and greater bandwidth to keep pace with rising compute demands, all within similar space and power constraints.

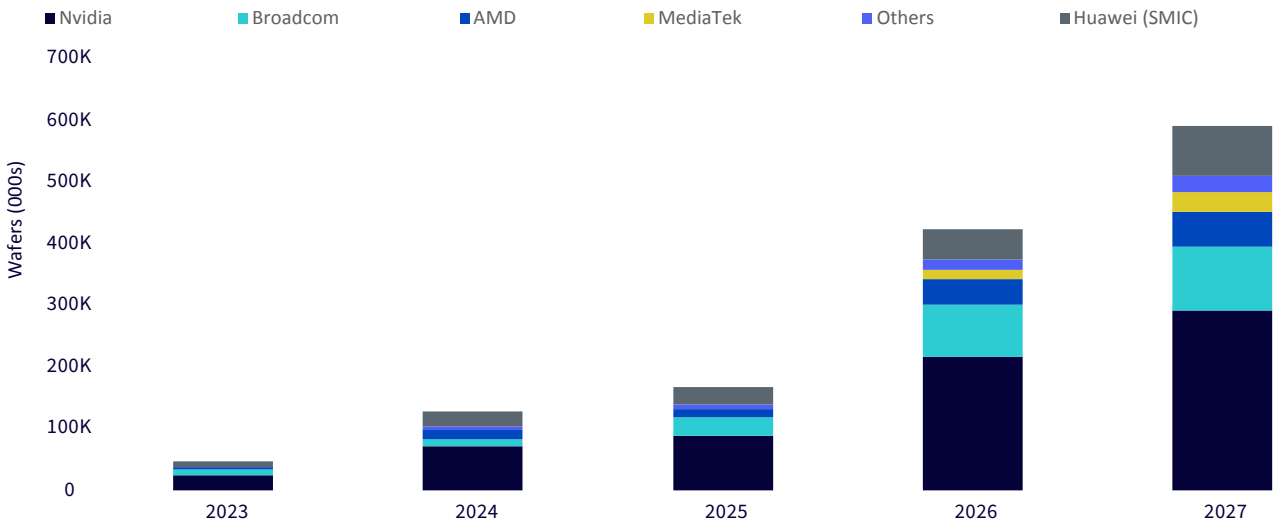
This growing complexity means that individual components can no longer simply be placed side by side on a traditional chip package. Instead, they must be tightly integrated using advanced packaging techniques such as CoWoS (Chip-on-Wafer-on-Substrate), which stack and connect multiple chiplets into a single unit. As a result, advanced packaging has become one of the most immediate bottlenecks in the supply chain.

8 WFE: Wafer Fabrication Equipment. The specialised machinery used to manufacture semiconductor chips, including lithography, deposition, etching and inspection systems.

9 Lithography: The process of transferring circuit patterns onto a silicon wafer using light. In particular, Extreme Ultraviolet (EUV) lithography, is essential for manufacturing the most advanced AI chips, and remains one of the most capital-intensive steps in semiconductor fabrication.

10 A cleanroom is a room used in semiconductor manufacturing with extremely low levels of airborne particles, temperature fluctuations, humidity variations and other forms of contamination or environmental instability. Small contaminant particles or deviations from ideal conditions can ruin a semiconductor wafer or even an entire batch of chips.

Figure 4: CoWoS[11] wafer shipments by company



Source: SemiAnalysis WFE Model, Feb 2026. **Forecasts are not an indicator of future performance and any investments are subject to risks and uncertainties. Historical performance is not an indication of future performance and any investments may go down in value.**

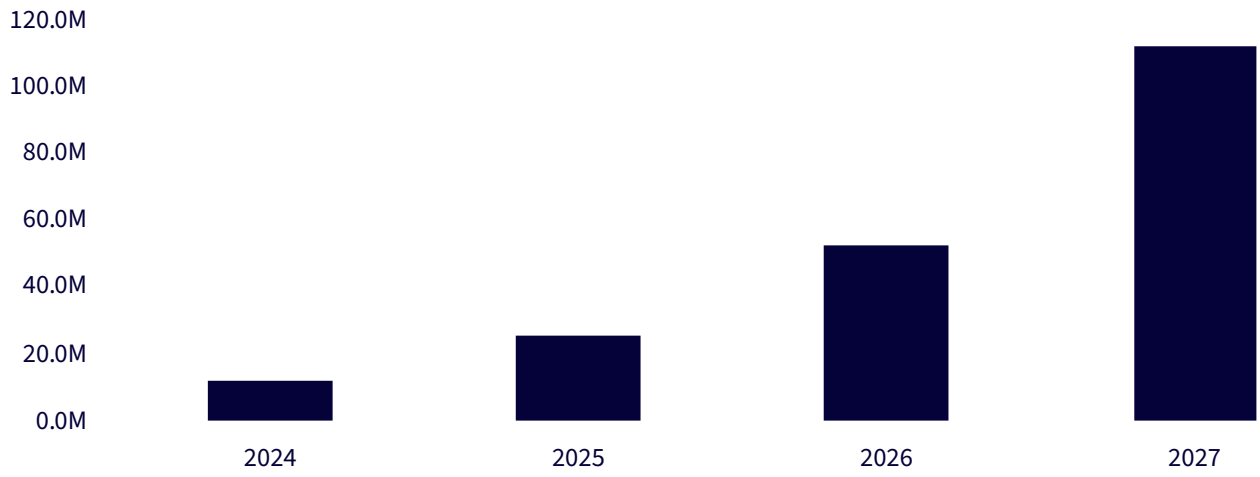
CoWoS wafer shipments are expected to increase more than tenfold from 2023 to 2027, yet demand continues to outpace supply. Chip designers are now securing packaging capacity years in advance. Firms with these specialised capabilities sit at a critical chokepoint, as access to advanced packaging directly determines who can produce the most advanced chips.

Networking and system level constraints

Accelerators are not the only chips that matter. As AI systems scale, thousands of processors must work together, and the networking infrastructure that connects them becomes just as critical. Data must move between chips at enormous speed and volume, making networking chips, optical transceivers and switching equipment essential components of the system. If the network cannot keep pace, even the most powerful AI accelerators sit idle.

11 CoWoS: Chip-on-Wafer-on-Substrate. An advanced semiconductor packaging technology developed by TSMC that enables multiple chips to be integrated onto a single substrate, critical for high-performance AI accelerators.

Figure 5: AI transceiver demand



Source: SemiAnalysis, April 2026. **Forecasts are not an indicator of future performance and any investments are subject to risks and uncertainties. Historical performance is not an indication of future performance and any investments may go down in value.**

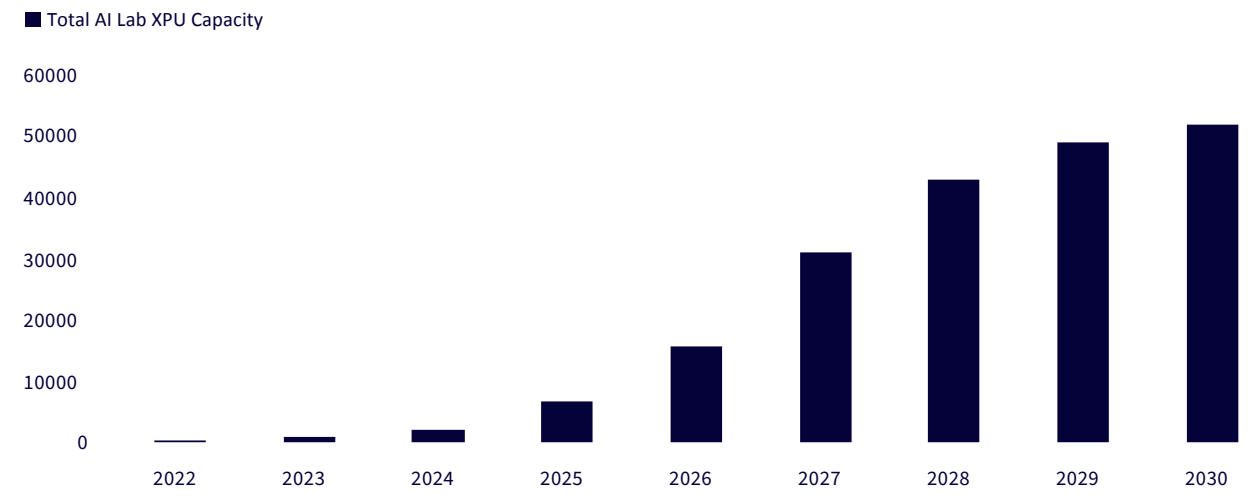
What often connects these processors at high speed within data centres are 800G[12] optical transceivers. Demand for these components is forecast to rise from 25 million units in 2025 to 112 million by 2027, with next-generation 1.6T links also scaling rapidly. This growth reflects the fact that networking is becoming an increasingly significant determinant of overall system performance and cost.

Power demand and data centre expansion

Even if the chips and networking can be delivered, they need somewhere to run. Every accelerator requires power, cooling and physical data centre space to serve tokens to end users. The scale of this demand becomes tangible when viewed through the lens of power consumption, as AI inference workloads drive a sustained increase in energy demand.

12 800G: Refers to 800 Gigabit per second optical transceivers, the current high-speed standard used in AI datacentre networking to move data between servers and switches. Next-generation 1.6T (1,600 Gbps) transceivers are beginning to deploy for higher-bandwidth AI clusters.

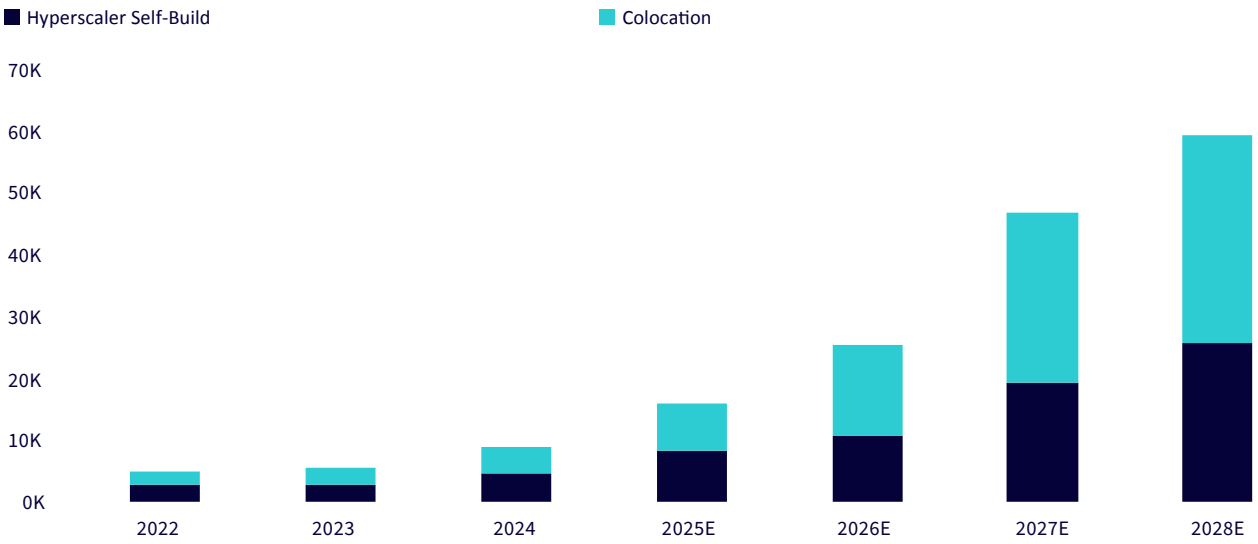
Figure 6: LLM provider XPU capacity (MW[13])



Source: SemiAnalysis AI Datacentre Industry Model, April 2026. **Forecasts are not an indicator of future performance and any investments are subject to risks and uncertainties. Historical performance is not an indication of future performance and any investments may go down in value.**

Inference power demand is expected to rise to well over 50,000 MW by 2030. At that scale, AI is no longer a marginal load on traditional power infrastructure. It becomes a system-level driver of power demand and, as a result, these systems behave more like industrial processes. This is reflected in the term 'AI factories', used by NVIDIA CEO Jensen Huang to describe modern AI data centres. Rising compute and power requirements are forcing a rapid expansion of data centre capacity. The only way to meet this demand is to get more AI factories up and running.

Figure 7: Global data centre MW additions (ex-China)



Source: SemiAnalysis AI Datacentre Industry Model, April 2026. **Forecasts are not an indicator of future performance and any investments are subject to risks and uncertainties. Historical performance is not an indication of future performance and any investments may go down in value.**

13 MW = megawatt.

Global data centre capacity additions are expected to scale from roughly 16,000 MW in 2025 to nearly 60,000 MW by 2028. Hyperscalers are building their own facilities and partnering with neoclouds and other data centre operators to bring capacity online as quickly as possible.

However, power and grid infrastructure operate on multi-year timelines, and are emerging as the binding constraint on how fast new capacity can be deployed. To work around this, operators are increasingly turning to onsite generation, including gas turbines, engine-based systems and fuel cells, prioritising speed over traditional grid dependence. Ultimately, the ability to deliver AI at scale is directly tied to the availability of physical infrastructure, and data centres and power are likely to be a defining bottleneck over the coming decade.

The AI infrastructure investment cycle is underway

AI is evolving into a physical system, and the infrastructure required to support it spans semiconductors and manufacturing supply chains, networking, and data centre and power systems. Demand is accelerating as AI moves into production across the globe, while constraints at each layer of the stack are becoming more visible.

AI is driving one of the largest capital investment cycles in the history of the technology sector. Demand is compounding, but the rate of deployment is governed not by demand alone, but by the pace at which physical infrastructure can be built. This extends the duration of the cycle and positions AI infrastructure as a defining investment theme for the coming decade.

Important Information

Marketing communications issued in the European Economic Area ("EEA"): This document has been issued and approved by WisdomTree Ireland Limited, which is authorised and regulated by the Central Bank of Ireland.

Marketing communications issued in jurisdictions outside of the EEA: This document has been issued and approved by WisdomTree UK Limited, which is authorised and regulated by the United Kingdom Financial Conduct Authority.

WisdomTree Ireland Limited and WisdomTree UK Limited are each referred to as "WisdomTree" (as applicable). Our Conflicts of Interest Policy and Inventory are available on request.

The information contained in this document is for your general information only and is neither an offer for sale nor a solicitation of an offer to buy securities or shares. This document should not be used as the basis for any investment decision. Investments may go up or down in value and you may lose some or all of the amount invested. Past performance is not necessarily a guide to future performance. Any decision to invest should be based on the information contained in the appropriate prospectus and after seeking independent investment, tax and legal advice.

The application of regulations and tax laws can often lead to a number of different interpretations. Any views or opinions expressed in this communication represent the views of WisdomTree and should not be construed as regulatory, tax or legal advice. WisdomTree makes no warranty or representation as to the accuracy of any of the views or opinions expressed in this communication. Any decision to invest should be based on the information contained in the appropriate prospectus and after seeking independent investment, tax and legal advice.

This document is not, and under no circumstances is to be construed as, an advertisement or any other step in furtherance of a public offering of shares or securities in the United States or any province or territory thereof. Neither this document nor any copy hereof should be taken, transmitted or distributed (directly or indirectly) into the United States.

Although WisdomTree endeavours to ensure the accuracy of the content in this document, WisdomTree does not warrant or guarantee its accuracy or correctness. Where WisdomTree has expressed its own opinions related to product or market activity, these views may change. Neither WisdomTree, nor any affiliate, nor any of their respective officers, directors, partners, or employees accepts any liability whatsoever for any direct or consequential loss arising from any use of this document or its contents.

WisdomTree.eu
+44 (0) 207 448 4330