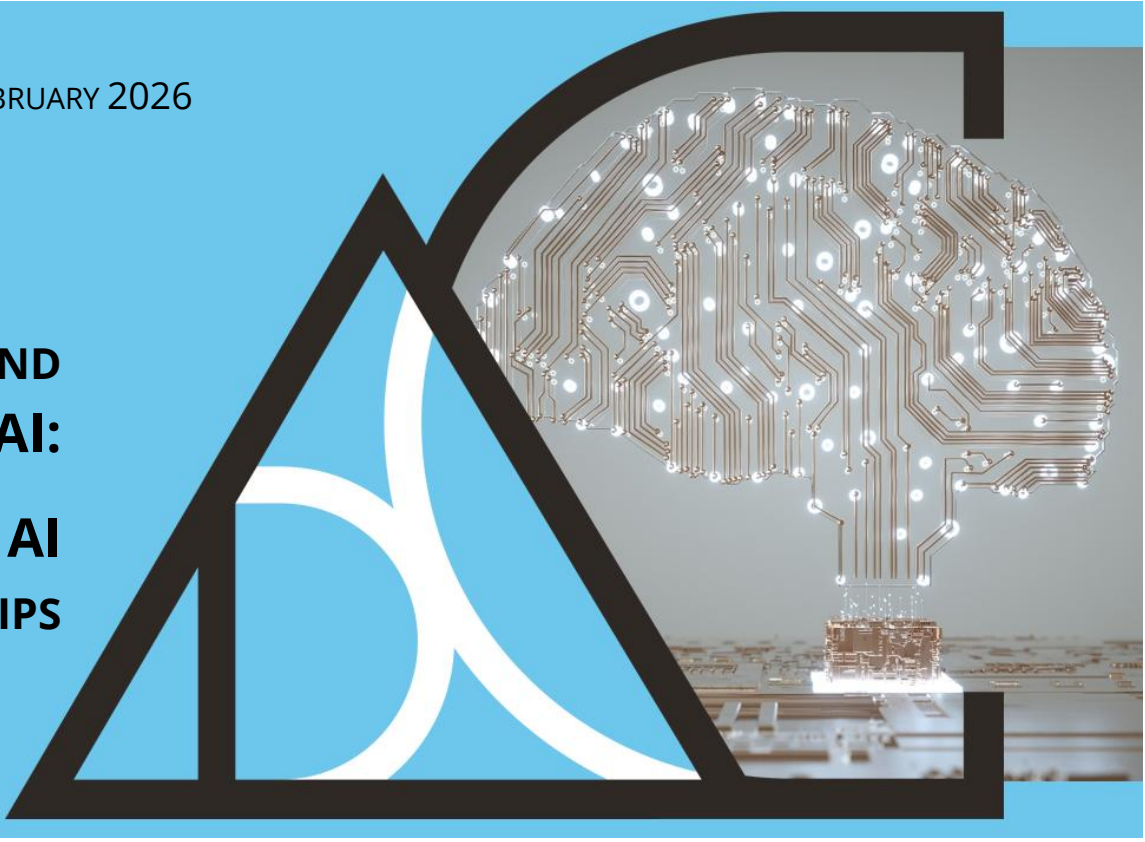
AdC Short Papers I February 2026



Competition and Generative AI:

Access to AI Chips

The Portuguese Competition Authority (AdC) has been following the generative artificial intelligence (AI) sector since late-2022. The AdC has published an issues paper on AI in November 2023[1] and initiated a short paper series on AI in 2024[2].

This is the fourth short paper of the series and examines competition issues related to access to chips for training and running AI models.

I. Introduction

Compute[3] and memory are critical inputs for generative AI, both during training and inference. The leading AI models compete on performance, encompassing the quality of AI outputs and the time it takes to produce them. To increase output quality, developers have sought to scale up their models along several dimensions. This includes increasing model size, larger volumes of training data, increasing context windows, or extending training or inference time[4]. AI developers have also focused on improving the efficiency of AI systems, namely

[1] Available here.

[2] So far, three short papers have been published. The first short paper, from September 2024, focused on the access and use of data in generative AI and its implications for competition (here). The second paper, from December 2024, covered issues related to access to AI models by downstream third-party AI developers and the role of openness of AI models in promoting competition and innovation (here). The third paper, from July 2025, addresses competition issues related to access to talent in the generative AI sector (here).

[3] Compute refers to the computational processing capacity required to execute AI workloads, including the mathematical operations (such as large-scale matrix multiplications and related functions) that underpin model training and inference. It encompasses both the hardware resources (e.g. GPUs, ASICs) and the effective availability of those resources over time.

[4] As the returns of scaling pre-training have seemingly diminished (see more here and here), AI developers have recently shifted to scaling inference-time compute. An example are the so-called reasoning models. Rather than producing the final answer in one go, these models generate intermediate reasoning steps as part of the output. This uses more compute during inference but often results in better responses. This kind of processing also requires greater memory capacity, which may open the door to alternatives to existing semiconductors, such as NVIDIA's chips, whose strengths lie in handling very large volumes of similar calculations (Financial Times," How 'inference' is driving competition to Nvidia's AI chip dominance", March 2025, here).

through architectural innovations[5]. Scaling up on these dimensions, combined with a sharp rise in demand for AI services over the last few years, have significantly increased the need for high-performance compute and memory resources.

Access to such compute depends on specialised hardware, namely AI accelerator chips (AI chips). These are special-purpose chips optimised to handle the core operations of AI models[6], for which they are significantly more efficient than traditional computer processors.

The most widely used AI chips are GPUs[7]. These are specialised semiconductor devices that are optimised for parallel processing and are at the cornerstone of the leading models. AI developers are reportedly using hundreds of thousands of high-performance GPUs, mostly from NVIDIA, to train and run their models (see Box 1) and plan on acquiring more[8]. AI developers may also rely on ASICs[9], custom-designed chips for specific tasks and applications, in particular for inference tasks. These chips are more efficient at those tasks (e.g., faster, cheaper, lower energy consumption), at the cost of flexibility[10]. Google's Tensor Processing Units (TPUs), for example, are ASIC chips that are used to train and run models like those behind Gemini. It has been reported that GPUs accounted for approximately 72% of the total AI chips market in 2023, while ASICs represented around 22%[11].

Compute capacity is also highly concentrated. According to a dataset by Epoch AI, the firms with the most compute capacity are xAI, Meta AI, Google, Microsoft, Tesla and NVIDIA, accounting approximately for 90% of the existing known capacity[12]. In addition, approximately 75% of the

[5] An example is the Mixture of Experts technique, which works by dynamically activating only the parts of the model that are most useful to process a particular input, reducing overall computation and cost.

[6] The core operations of modern AI workloads consist of massively parallel linear algebra (primarily matrix multiplication) and supporting elementwise nonlinear functions.

[7] Graphics Processing Units.

[8] See, for example, the State of AI Report Compute Index which estimates that the largest AI developers may be currently using hundreds of thousands of high-performance GPUs, mostly by NVIDIA (here). See also reports claiming that xAI may currently use 200,000 Nvidia GPUs, or that OpenAI plans to reach 1 million GPUs by the end of 2025 (here). In addition, data on GPU clusters from Epoch AI shows that firms like Meta AI, Oracle, DataVolt, OpenAI/Microsoft, Sesterce, xAI and Reliable Industries likely plan to have more than 1 million H100-equivalent GPUs in the next 5 years. This data includes existing and planned, confirmed and likely, non-duplicate, private GPU clusters as per Epoch AI. See more in Pilz, Rahman, Sanders & Heim (2026), "*Data on GPU Clusters*" (here). Accessed 12 February 2026.

[9] Application-Specific Integrated Circuits.

[10] There is a trade-off between generalisation and specialisation in chip design. On one end of the spectrum, general-purpose processors like CPUs offer flexibility and can handle a wide range of tasks but are less efficient at specific tasks. On the other hand, ASICs deliver the best performance and efficiency for a narrow set of tasks but lack flexibility. GPUs strike a balance in this trade-off, given they have very high performance for parallel computation, specific operations required for AI, but they are still rather flexible.

[11] According to a report by The Information Network, a semiconductor industry analyst, from May 2025, as cited by eeNews Europe. Projections for 2025 expect ASICs to increase their market share, primarily at the expense of GPUs.

[12] Data based on the dataset from Epoch AI on GPU clusters. This data includes existing, confirmed, non-duplicate, private GPU clusters as per Epoch AI. If one were to include confirmed planned GPU clusters as well, the largest firms in terms of compute capacity would be xAI, Meta AI, Google, Amazon, Microsoft and NVIDIA. Compute capacity based on maximum theoretical performance in 16-bit FLOP/s. See more in Pilz et al. (2026).

currently deployed capacity is owned by private firms[13], and public entities have seen their share of total compute capacity decline in recent years[14].

AI model performance also depends heavily on memory. During training and especially inference, models must process large volumes of data quickly. Delays in memory access can degrade model speed and impair real-time applications[15]. To support this, AI chips integrate high-bandwidth memory (HBM) directly onto the same package. HBM is an advanced type of memory that enables significantly faster data access and transfer speeds compared to conventional memory used in consumer devices.

AI developers often access compute and memory via cloud providers, which have taken a central role in the AI stack. The cost of building a computing infrastructure suitable for AI is prohibitively high for most AI developers. Cloud providers offer scalable, on-demand access to advanced computing and memory resources, allowing AI developers to train and deploy models without incurring the substantial capital costs associated with owning and maintaining dedicated hardware.

In parallel with hardware, software plays a critical role in determining how efficiently AI chips run. Software stacks – including drivers, compilers, runtimes, and development frameworks – bridge the gap between AI models and the underlying hardware.

Achieving state-of-the-art performance requires more than just computational capacity; it also relies on the interaction between compute, memory and communication resources[16]. Sustained access to these inputs is crucial for performance and is shaped by concentrated markets for advanced chip technologies. This raises questions regarding market dynamics, structural dependencies, and potential risks to competition across the AI supply chain and the AI stack. These issues are addressed in the sections that follow.

II. Bottlenecks across AI chip supply chain

The production of AI chips relies on a complex global value chain encompassing multiple interconnected stages, from design to manufacturing. Box 1 (next page) presents a visual overview of this supply chain, along with a description of each stage and its key components and stakeholders.

---

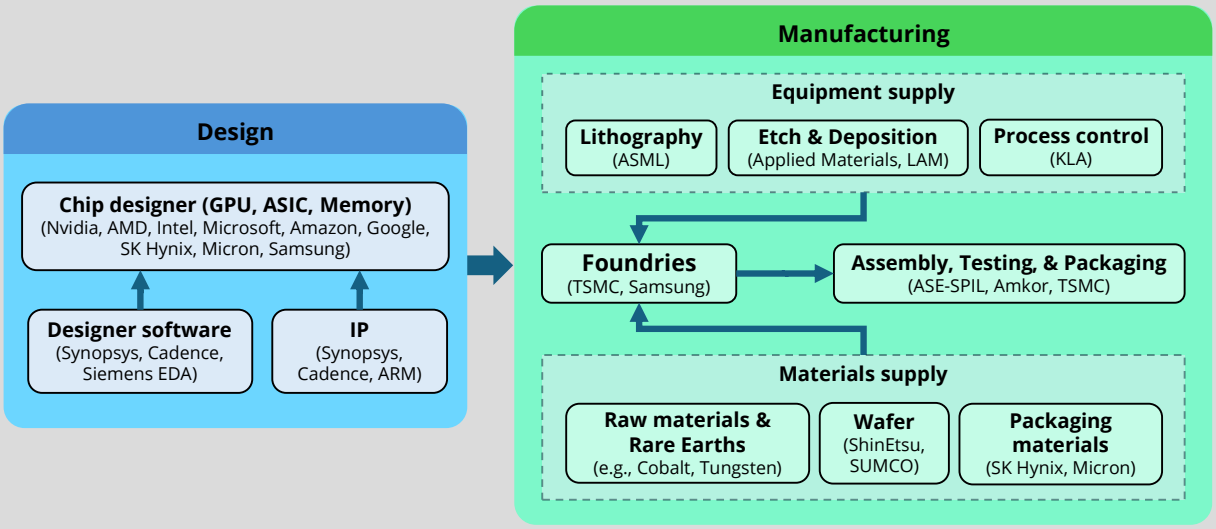
[13] According to the dataset from Epoch AI on GPU clusters, approximately 75% of the compute capacity is owned by private firms, 12% by public entities, and 13% by joint public–private partnerships. If one includes planned GPU clusters, however, 57% of the of compute capacity is owned by private firms, 43% by joint public-private partnerships and less than 1% by public entities. This data includes existing and planned, confirmed and likely, non-duplicate GPU clusters as per Epoch AI. Compute capacity based on maximum theoretical performance in 16-bit FLOP/s. See more in Pilz et al. (2026).

[14] See Pilz, Sanders, Rahman & Heim (2025). Trends in AI Supercomputers. Available here.

[15] Memory performance has struggled to keep pace with advances in compute. According to an AI-focused industry publication, since 2023, AI compute power has increased by 750% and processing speeds have tripled, while memory bandwidth has grown by only 1.6X.

[16] These utilisation gaps are largely attributed to memory bottlenecks, synchronisation overheads and software inefficiencies, rather than just insufficient raw compute (here and here).

Box 1 – AI chip value chain and key players



Chip Design

Chip design creates the blueprint of the chip, which defines its parts and how they operate. This process is highly complex and intensive in R&D and aims at ensuring chips are performant, power efficient, reliable and cost effective.

NVIDIA is the main player in the design of high-end GPUs[17] used in AI, accounting for over 90% of market share of discrete GPUs for use in datacentres[18]. Other competitors in this stage include AMD and Intel. The AI-specific ASIC segment includes start-ups such as Cerebras Systems and Groq, as well as cloud providers like Google (TPUs) and Amazon (Inferentia and Trainium).

Most AI chips are not designed entirely from scratch[19]. Instead, chip designers integrate licensed pre-designed modules known as IP cores[20], developed by firms such as Synopsys, Cadence, and Arm, into custom architectures to reduce complexity and accelerate time-to-market. These same players, together with Siemens EDA, also supply electronic design automation (EDA) software, which is essential for verifying circuit layouts, optimising physical design, and simulating performance prior to physical fabrication.

Memory chips, particularly HBM, are also critical to scaling and accelerating training and inference in AI accelerators. The HBM stacks are physically co-packaged with the accelerator chip (GPUs, TPUs, etc), creating a single integrated unit where the compute die provides the processing power and the HBM delivers the required memory bandwidth. The global supply of HBM is concentrated, with SK Hynix, Samsung, and Micron as the major suppliers[21].

Chip Manufacturing

Chip design blueprints are sent to foundries which are contracted to manufacture the physical

[17] Key NVIDIA AI chips include the A100 and H100, as well as the most recent Blackwell-generation chips. NVIDIA is also the developer of the advanced Hopper (H100/H200) and recently unveiled Blackwell (B200/B300) chips (here).
[18] Market share used in the Commission decision Case M.11766 – NVIDIA / RUN:AI (see decision here). Data on GPU clusters from Epoch AI shows a similar figure. NVIDIA is the primary GPU supplier for approximately 88% of the clusters whose suppliers are known, while AMD is second with 7% and Google is third with 4%. This data includes only existing, confirmed, non-duplicate, private GPU clusters as per Epoch AI. See more in Pilz et al. (2026).
[19] Embedded, "Silicon Integrated Circuit Intellectual Properties Design", October 2024 (here).
[20] IP cores are standardised, reusable circuit blocks providing essential functionalities like processors, memory controllers, or connectivity interfaces.
[21] The Economist, "Memory chips could be the next bottleneck for AI", October 2024 (here).

chips as per the designer's specifications. This stage is also concentrated, with TSMC being the key player in manufacturing high-end AI chips, particularly at process nodes of 5 nanometer and below[22].

Upstream, the manufacturing process relies on specialized materials, including silicon wafers, rare-earth elements, and critical chemicals, sourced through a complex global supply network.

Once fabricated, chips are sent to ATP (Assembly, Testing, and Packing), where they are encapsulated, tested, and connected to memory and interconnect systems. The global ATP market is led by ASE and Amkor, while TSMC also provides advanced packaging services in-house[23].

Both the design and the manufacturing stages of the AI chip supply chain are concentrated, namely due to structural factors. Both stages exhibit significant economies of scale. On the design side, developing leading-edge chips entails intensive R&D investments, extensive verification cycles, and relies on sophisticated tools. On the manufacturing side, producing high-end chips requires multibillion-dollar investments in fabrication facilities[24] and access to highly specialised equipment, such as extreme ultraviolet lithography systems.

The structural characteristics of the AI chip supply chain may also contribute to supply rigidity. High fixed costs, long lead times and time-to-market, and limited supplier capacity constrain the ability to scale production quickly. As demand for more powerful chips accelerates[25], these constraints intensify across the supply chain, contributing to the persistent chip scarcity affecting developers and cloud service providers.

Indeed, demand for high-end GPUs continues to exceed available supply[26]. A similar situation is unfolding in the memory segment, with reports indicating that HBM chips are largely sold out for 2025 and much of 2026[27]. In the manufacturing stage, access to advanced foundry services remains limited[28], with reported prioritisation of well-capitalised clients[29]. Constraints on packaging capacity also cause delays, particularly

[22] CNN Business, "Why this key chip technology is crucial to the AI race between the US and China", June 2025 (here).
[23] See reports here and here.
[24] Estimates indicate that new facilities can cost between $15 billion and $20 billion or more (here), with construction and equipment timelines ranging from 18 months to over four years (here, here, and here).
[25] Comparing planned and existing GPU clusters in the dataset from Epoch AI, the total compute capacity of planned systems in the next ~5 years is approximately 37 times larger than the current existing capacity. This data includes confirmed and likely, existing and planned, non-duplicate GPU clusters as per Epoch AI. Compute capacity based on maximum theoretical performance in 16-bit FLOP/s. See more in Pilz et al. (2026). See more in Pilz et al. (2026).
[26] Despite expected production increases of Blackwell-generation chips, NVIDIA's supply has not yet caught up with demand, according to the Q3 2025 NVIDIA Corp Earnings Call.
[27] Yahoo Finance, "Micron Sells Out 2025 HBM Supply: Can It Meet Soaring Demand in 2026?", June 2025 (here).
SAM Mobile, "SK Hynix will soon sell out 2026 HBM chips while Samsung struggles", March 2025 (here).
[28] Over 50% of generative AI companies report GPU shortages as a major obstacle to scaling their operations. Vertu, "Unveiling Trends in the Global AI Chip Market", May 2025 (here).
[29] In mid-2024, TSMC revealed that its advanced packaging capacity for 2024 and 2025 was fully booked by just two clients – NVIDIA and AMD – racing to secure capacity for AI processors.

as advanced packaging facilities are capital-intensive and cannot be scaled rapidly[30].

The lifecycle of AI chips is also shortening, driven by rising computational demands from AI developers and the rapid pace of innovation in chip technology[31]. This dynamic further intensifies barriers to entry and expansion in the AI supply chain, as firms must adjust to faster turnover cycles while still operating within capacity and investment constraints[32].

Shorter chip lifecycles translate into faster depreciation of AI hardware, increasing financial risk for firms that rely on owning and renting computational capacity. This is particularly relevant for smaller cloud intermediaries, whose business model depend on maintaining high utilisation rates over longer periods to amortise investments in GPUs. By contrast, major cloud providers benefit from access to cheaper capital and sustained internal demand, which allows them to absorb faster depreciation, smooth utilisation over time, and renew hardware more frequently.

Structural and demand-side pressures, and the resulting scarcity, have given market power to key players in the AI supply chain and contributed to sustained increases in the price of AI chips[33].

This environment raises barriers to entry and expansion for AI developers, due to the high cost of accessing the compute needed to train and deploy advanced models. Most AI developers, particularly start-ups, cannot build their own computing infrastructure due to its high upfront cost. While relying on cloud computing may partially solve this scale effect, the high marginal cost of accessing compute may constrain their ability to enter the market, innovate, and compete across the AI ecosystem[34].

Public high-performance computing (HPC) infrastructure may partially mitigate this bottleneck by broadening access to large-scale computing capacity[35]. These

[30] Though NVIDIA's CEO Jensen Huang highlighted that the industry had "*probably four times*" more advanced packaging capacity in early 2025 than two years prior, the demand was so high that it remained the limiting factor (here). By late 2024, NVIDIA was reportedly selling its newest "Blackwell" AI chips as fast as TSMC could assemble them, with packaging remaining "*a bottleneck due to capacity constraints*", according to TSMC Chairman Mark Liu.

[31] A report finds that the median time between the release of leading AI chips and the publication of the final frontier model trained on them is approximately four years. This trend is also recognised by cloud providers. CoreWeave disclosed in its IPO filing that the launch of Blackwell immediately devalued its Hopper-based systems (previous NVIDIA's chips), requiring accelerated depreciation. AWS and others have similarly shortened server depreciation timelines (here).

[32] Microsoft, for instance, is facing delays in the development and deployment of its AI chips, struggling to match Nvidia's performance and timelines (here).

[33] One of the consequences is the cost of training AI models, which has surged over 300% in just a few years, driven by rising GPU prices. Vertu, "Unveiling Trends in the Global AI Chip Market", May 2025 (here). According to Cloudzero, on Google Cloud, a single A100 GPU instance can cost over 15X more than a standard CPU instance.

[34] A survey conducted by Civo, a cloud services provider, found that 85% of organisations reported delays in their AI/ML projects due to limited GPU availability, with more than one-third citing delays of three to six months. See here article reporting on Civo's research.

[35] See OECD Roundtables on Competition Policy Papers, No. 330, "Competition in artificial intelligence infrastructure", November 2025 (here).

supercomputers may support compute-intensive AI workloads by aggregating thousands of interconnected advanced chips.

Although their current scale remains limited relative to overall demand for AI compute[36], **more public HPC infrastructure is being built to support industrial and commercial AI projects**[37]. At the European level, the recent launch of AI Factories[38] and the planned AI Gigafactories[39] aims to scale up publicly accessible computing infrastructure to support the development and deployment of AI solutions.

The pro-competitive potential of public HPC infrastructures depends on their governance and access conditions. Given the scarcity of compute and memory for AI developers, the way access to these supercomputers is allocated may influence competitive conditions in downstream AI markets. For this reason, access criteria to this infrastructure should remain transparent, objective and non-discriminatory.

Structural bottlenecks in the AI chip supply chain may undermine market access

Persistent supply constraints, high concentration across the value chain, and rising demand for cutting-edge hardware have created bottlenecks in the production and delivery of AI chips, limiting smaller developers' ability to access the compute needed to compete in advanced AI development.

Public HPC initiatives may expand access to AI compute, although their scale remains limited

Public HPC infrastructures may broaden access to large-scale computing capacity and partially alleviate supply constraints. While still limited in scale, recent European initiatives seek to expand their role in supporting AI development. Their impact on competition depends on transparent and non-discriminatory access criteria.

[36] According to the dataset from Epoch AI on GPU clusters, the total existing compute capacity under the EuroHPC programme is comparable to that of private firms like Meta AI (~1.59 times that of EuroHPC) or Google (~0.93 times). This data includes confirmed, existing, non-duplicate GPU clusters as per Epoch AI. Compute capacity based on maximum theoretical performance in 16-bit FLOP/s. See more in Pilz et al. (2026). See more in Pilz et al. (2026).

[37] For example, some recent programmes in Portugal such as InovIA (a voucher programme for innovation that provides easy access to supercomputers) have aimed at providing SMEs and start-ups with HPC capabilities for R&D and experimentation purposes.

[38] Under the EuroHPC Joint Undertaking, the EU intends to establish at least 19 AI Factories: HPC ecosystems optimized for AI, offering computing resources and support services to the European industry and scientific users for the development of large AI models (see more here and here).

[39] The Commission announced AI Gigafactories will be capable of training the most complex AI models and provide at least four times more capacity than the AI Factories. Investments will be made through a public-private partnership, where the Commission will commit €20 billion from the InvestAI initiative (here).

III. Cloud providers expanding across the AI stack

Major cloud providers have become a cornerstone for AI development. These – often referred to as hyperscalers – include Amazon Web Services (AWS), Google Cloud, and Microsoft Azure. Even leading developers, with state-of-the-art models and a very large number of users, such as OpenAI, rely on cloud computing to train, do inference, and deploy their AI models.

To reduce dependency on chip suppliers, hyperscalers have increasingly invested in the design of proprietary AI chips – such as Amazon's Trainium and Inferentia[40], or Google's TPUs[41].

Although these chips were initially designed for internal use, most cloud providers are now integrating them into their own distribution channels[42]. These chips remain integrated with the providers' own cloud infrastructure and are generally not available for purchase on the open market (i.e., developers can only access them through the corresponding cloud platforms). Within these environments, cloud providers may offer access to a mix of proprietary chips and third-party GPUs, such as those from NVIDIA or AMD.

Some cloud providers may begin to supply their own chips to AI developers. For example, there are reports that Google's TPUs may be used in the future by Meta[43].

The entry of hyperscalers into chip design can have pro-competitive effects. These developments may help diversify supply and reduce dependency on any single set of players, which may help ease existing bottlenecks.

However, it may also raise competition risks, particularly the risk of exclusionary strategies. Vertical integration may provide hyperscalers the ability and, in certain circumstances, the incentive to leverage their dual role as infrastructure providers and chip designers to reinforce their competitive position across the AI stack. On the one hand, they control the cloud environments that constitute the main channel through which AI developers access compute resources. On the other hand, they design and deploy their own proprietary accelerators, which are integrated within the same cloud environments. In such scenarios, hyperscalers may have the incentive to favour their own chips through differentiated pricing structures, discount eligibility, allocation of capacity, or closer integration with proprietary software stacks[44].

As cloud providers expand into chip design, chip firms are moving beyond hardware supply into cloud-based AI compute services. Examples of this shift includes recent initiatives

---

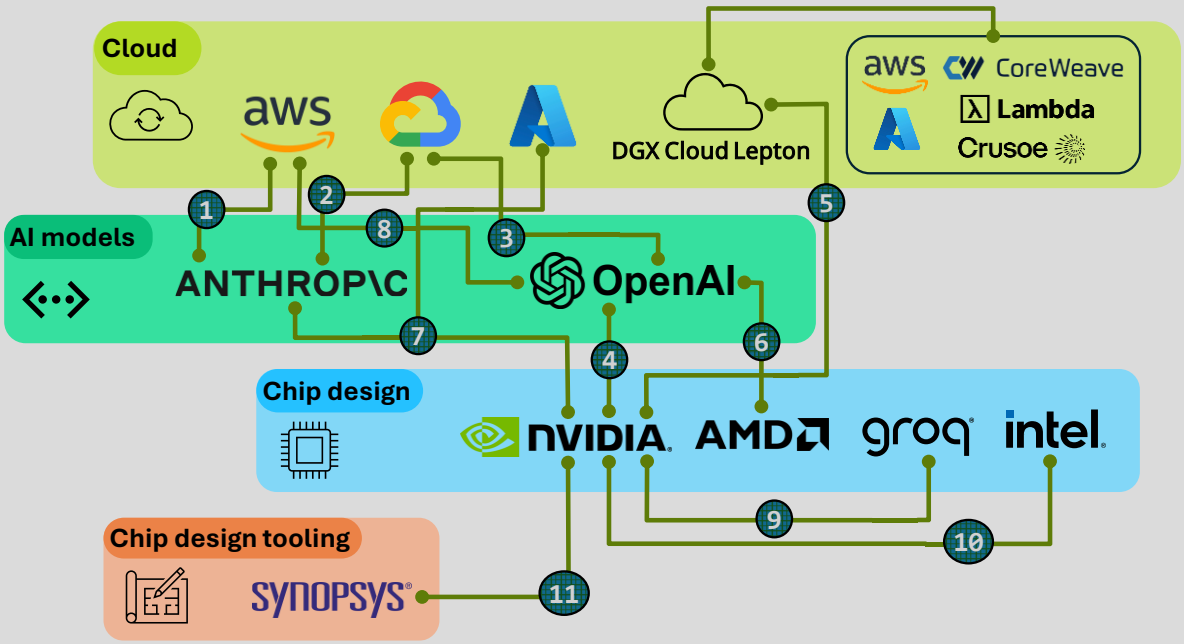
[40] AWS Tranium (here); AWS Inferentia (here).
[41] Google's TPUs are specifically optimised for Google's own machine learning workloads, demonstrating efficiency in both inference and training processes (see more here and here). Over half of the AI workloads on Google Cloud now rely on TPU chips (here).
[42] OpenAI, for example, is now renting Google TPUs to diversify its AI computing power from NVIDIA chips and Azure cloud services (see more here). AWS also offers proprietary AI chips (see here, here, and here).
[43] The Wall Street Journal, "Meta Is in Talks to Use Google's Chips in Challenge to Nvidia", November 2025 (here).
[44] For instance, Google Cloud excludes several NVIDIA GPU families – including A100, H100, H200, B200 (Blackwell) – from automatic price reductions under Sustained Use Discounts (here), and also excludes A4 instances (based on Blackwell) from Committed-Use Discounts (here). TPU instances, by contrast, remain eligible.

such as Intel's Tiber AI Cloud[45], AMD's Developer Cloud[46], and NVIDIA's DGX Cloud[47].

To date, these platforms seem to be used primarily to showcase the capabilities of the underlying hardware and ecosystems to companies seeking to invest in AI. Over time, however, they may take on a more strategic role by reducing dependence on hyperscalers, while ensuring that proprietary products are available through hosted environments adapted to their performance characteristics, and reinforcing control over the hardware–software interface.

Partnerships play a prominent role in operationalising these strategies across the AI stack. Hyperscalers have established several partnerships with AI developers, particularly involving access to cloud infrastructure and simultaneously promoting the use of their proprietary accelerators. Chip designers, in turn, have established partnerships with smaller cloud providers that are seeking to expand in the AI market. Box 2 outlines a selection of publicly disclosed partnerships.

Box 2 – AI partnerships across the stack



[45] Intel's Tiber AI Cloud was launched in October 2024 as an evolution of the earlier Intel Tiber Developer Cloud, aiming to support production-scale AI workloads rather than just development trials (here).
[46] AMD's Developer Cloud launched in June 2025 as a developer platform aimed at increasing access to AMD's Instinct GPUs, with preconfigured ROCm software stacks (here).
[47] In June 2025, NVIDIA extended its DGX Cloud with DGX Cloud Lepton, a global GPU aggregation layer that pools spare capacity across multiple cloud partners (see Box 2). This enables developers to dynamically access unused GPU resources without being locked into a single provider, thereby easing transitions between different cloud platforms (see more here and here).

Partnerships have become an important feature of the AI ecosystem, shaping how compute resources are developed and distributed. They often reflect the dual trend of hyperscalers moving into chip design and chip designers expanding into cloud services, while both seek to attract AI developers into their ecosystems. The following examples illustrate these developments, though the list is not exhaustive.

1. **AWS and Anthropic**: Shortly after the partnership with Google, Anthropic started its relationship with AWS, establishing AWS as its primary cloud and training partner. This followed a substantial Amazon investment, with an initial $1.25 billion in September 2023, reaching $8 billion of total investment in 2024. This collaboration explicitly includes working on the development of Trainium, AWS's proprietary chip designed for training AI models[48].
2. **Google Cloud and Anthropic**: Anthropic initially collaborated with Google Cloud, announcing their partnership in February 2023. Anthropic accessed and used Google's GPU and TPU clusters for its AI model development[49].
3. **Google Cloud and OpenAI**: OpenAI partnered with Google Cloud starting in early 2025 to gain access to further compute resources, including Google's custom-built TPUs[50].
4. **NVIDIA and OpenAI**: In September 2025, NVIDIA and OpenAI signed a partnership to deploy up to 10 gigawatts of NVIDIA systems for OpenAI's next-generation AI datacentres. Under the arrangement, NVIDIA will invest up to $100 billion, in $10 billion tranches per gigawatt of capacity, with the first 1 gigawatt scheduled for the second half of 2026. For each $10 billion invested by NVIDIA, OpenAI is reportedly purchasing around $35 billion worth of NVIDIA's GPUs under the deal[51]. Recent reports, however, indicate that negotiations over the agreement have stalled and that no definitive contract has yet been finalised[52].
5. **NVIDIA's DGX Cloud Lepton**: as a GPU marketplace where NVIDIA lists capacity from multiple clouds under one catalogue, Lepton aggregates supply from partner clouds. Confirmed participants in this platform include hyperscalers, such as AWS and Microsoft Azure, and other cloud providers like CoreWeave, Crusoe, Nebius, Lambda Labs, among others[53]. Some of the commercial terms became public in September 2025. For instance, NVIDIA reportedly entered into a $1.5 billion, four-year lease to rent back approximately 18,000 GPUs it had previously sold to Lambda[54]. In addition, NVIDIA is said to have signed a $6.3 billion agreement with CoreWeave, under which NVIDIA committed to purchase any unused cloud capacity not taken up by customers[55]. In January 2026, NVIDIA also reportedly invested $2 billion in CoreWeave, increasing

[48] Anthropic, "Powering the next generation of AI development with AWS", November 2024 (here).
[49] Anthropic, "Anthropic Partners with Google Cloud", February 2023 (here).
[50] Revolgy, "Google Cloud signs deal with OpenAI, ending Microsoft's exclusive role", June 2025 (here).
[51] The Economist, "Nvidia's $100bn bet on OpenAI raises plenty of questions", September 2025 (here).
[52] CNBC, "Nvidia, OpenAI appear stalled on their mega deal. But the AI giants still need each other", February 2026 (here).
[53] Announced by NVIDIA in June 2025. Google Cloud is not listed among Lepton participants (here).
[54] Tom's Hardware, "Nvidia signs $1.5 billion deal with cloud start-up Lambda to rent back its own AI chips", September 2025 (here).
[55] Reuters, "CoreWeave, Nvidia sign $6.3 billion cloud computing capacity order", September 2025 (here).

its equity stake[56].

6. **AMD and OpenAI**: In October 2025, AMD signed an agreement to supply 6 gigawatts of computing capacity using its forthcoming MI450 AI chips for OpenAI's next-generation AI datacentres, with the first shipments scheduled for the second half of 2026. The deal includes giving OpenAI the option to acquire up to 10% of AMD's shares at a nominal price of $0.01 per share[57].
7. **Microsoft, NVIDIA and Anthropic**: In November 2025, Microsoft and NVIDIA reportedly committed to invest up to $5 billion and $10 billion, respectively, in Anthropic. In parallel, Anthropic committed to purchase approximately $30 billion in compute capacity from Microsoft, with Microsoft's data centres relying on NVIDIA AI chips[58].
8. **AWS and OpenAI**: In November 2025, AWS and OpenAI reportedly agreed on a seven-year arrangement valued at $38 billion, granting OpenAI access to large volumes of NVIDIA GB200 and GB300 (Blackwell) GPUs hosted on AWS infrastructure[59].
9. **NVIDIA acquihire of Groq**: In December 2025, NVIDIA entered into a non-exclusive technology licensing agreement with Groq, a company specialising in inference-focused AI chips. In parallel, NVIDIA hired several key Groq executives and engineers, including its founder, while Groq continues to operate as an independent firm[60].
10. **Intel and NVIDIA**: In December 2025, NVIDIA reportedly acquired shares worth approximately $5 billion (around a 4% stake) in Intel, a competing semiconductor manufacturer[61].
11. **NVIDIA and Synopsys**: In December 2025, NVIDIA invested $2 billion (around 2.6% stake) in semiconductor design software provider Synopsys as part of a strategic partnership to accelerate computing and artificial intelligence engineering solutions[62].

Some of these partnerships and investments raise concerns, as they combine procurement commitments, technology dependencies and transfers of strategic assets in ways that may soften rivalry between actual or potential competitors.

In particular, the growing use of reverse acquihires – as discussed in the AdC's analysis of labour markets in generative AI – **may lead to the concentration of scarce technical talent and know-how**. **A reverse acquihire may also amount to a concentration under EU Merger Regulation and the Portuguese Competition Act**[63].

[56] Reuters, "Nvidia invests $2 billion in CoreWeave to boost data center build-out", January 2026 (here).
[57] Reuters, "AMD signs AI chip-supply deal with OpenAI, gives it option to take a 10% stake", October 2025 (here).
[58] Financial Times, "Microsoft and Nvidia to invest up to $15bn in OpenAI rival Anthropic", November 2025 (here).
[59] OpenAI, "AWS and OpenAI announce multi-year strategic partnership", November 2025 (here).
[60] Groq, "Groq and Nvidia Enter Non-Exclusive Inference Technology Licensing Agreement to Accelerate AI Inference at Global Scale", December 2025 (here).
[61] Reuters, "Nvidia takes $5 billion stake in Intel under September agreement", December 2025 (here).
[62] NVIDIA, "NVIDIA and Synopsys Announce Strategic Partnership to Revolutionize Engineering and Design", December 2025 (here).
[63] See the AdC Short Paper "Competition and Generative AI: Labour Markets", from July 2025, here.

They may also grant the partners involved privileged access to sensitive technical and commercial information that may be unavailable to rivals[64]. Such arrangements could fall within the scope of Article 101 of the TFEU, in particular if the exchange of commercially sensitive information reduces strategic uncertainty or sustains a collusive outcome[65].

In addition, partnerships between firms at different levels of the AI stack may alter incentives across adjacent markets. A firm holding a strong position in one segment (e.g. cloud infrastructure or chip supply) may acquire the ability – and, in some circumstances, the incentive – to influence or restrict access to key AI models or AI chips. This could affect competition both upstream (model development and compute supply) and downstream (distribution of AI models and AI-enabled services via cloud)[66].

Circular investments, whereby large firms invest in AI developers that subsequently rely on the same firms for chips or cloud services, may reinforce incumbents. Even without explicit exclusivity, such arrangements can limit exposure to alternative suppliers, align incentives across the ecosystem and accelerate market tipping in AI infrastructure markets characterised by scale effects and input concentration. This dynamic is particularly relevant in markets where chip designers have extensive software ecosystems, proprietary development frameworks, or strong network effects. In such contexts, technological dependencies and aligned incentives may reinforce incumbent positions, reduce contestability and limit effective choice for AI developers. These risks are further examined in the next section.

Expanding control across the AI stack may increase market concentration

As cloud providers move into chip design and chip providers expand into cloud services, control over hardware and infrastructure may become more concentrated. This could reinforce ecosystem incentives that lock-in developers and raise concerns about interoperability, self-preferencing, and foreclosure risks.

IV. Hardware-software integration reinforcing lock-in effects

AI chips combine hardware and software components to enable complex parallel computing. A key part of this stack is the

[64] U.S. Federal Trade Commission Staff Report, "Partnerships Between Cloud Service Providers and AI Developers", January 2025 (here) and European Commission, Policy brief "Competition in Generative AI and Virtual Worlds", September 2024 (here).

[65] See European Commission, Guidelines on the applicability of Article 101 TFEU to horizontal cooperation agreements (2023 revision), July 2023 (here). In particular, the Guidelines explain that information exchange may be assessed under Article 101(1) if it goes beyond what is objectively necessary or proportionate for the implementation of the agreement. Additional exchanges between competitors may further reduce strategic uncertainty, and a unilateral disclosure accepted by a competitor can constitute a concerted practice where there is a link between the disclosure and subsequent market conduct.

[66] See CMA's decision on Microsoft/OpenAI partnership merger inquiry, March 2025 (here) where the CMA examined whether an increase in Microsoft's influence over OpenAI could create incentives to restrict rivals' access to OpenAI's foundation models in markets, including namely cloud compute and downstream productivity software markets.

software platform that bridges AI models and the underlying hardware within a single chip and across multiple chips operating jointly.

Among the accelerator software stacks available for AI development[67], **CUDA**[68], **together with its associated libraries**, **has become the most widely adopted**, **and it is only compatible with NVIDIA hardware**. According to NVIDIA, this ecosystem is used by over five million developers and more than 3,700 applications[69]. In addition, most modern AI development tools and programming libraries were first written for CUDA and remain primarily optimised for that framework, although multi-backend support is growing[70].

Because of the interplay between software and hardware, NVIDIA's position in CUDA has strengthened its position in AI chips. The widespread adoption of CUDA creates strong network effects around that framework and NVIDIA's chips[71]. As more AI developers use CUDA, more AI development tools and programming libraries are built on and optimised for that software stack, and NVIDIA chips become more performant and easier to use. This, in turn, attracts more AI developers to CUDA and expands third-party tools and programming libraries for AI development, further reinforcing NVIDIA's position in the AI chips stack.

While the ecosystem of tools around CUDA brings many benefits to AI developers, it may also lock them in on NVIDIA's chips. For AI developers, switching away from NVIDIA's chips often means giving up on CUDA and adopting alternative software stacks. As a result, AI developers lose many of the optimised and state-of-the-art tools and programming libraries they relied on. This switch may require developers to rewrite low-level code, create custom solutions, and retrain their teams, which may only be achievable by the largest AI developers[72].

Although leading AI frameworks have become increasingly accelerator-agnostic at

---

[67] For the purposes of this paper, accelerator software stacks refer to the integrated set of tools, languages, libraries, compilers, and runtimes that enable developers to program and optimise workloads for specific AI chips. This also encompasses language extensions (e.g. CUDA C++, HIP C++) and the associated software development kits (SDKs) provided by chip designers.

[68] CUDA (Compute Unified Device Architecture) is NVIDIA's proprietary programming model and software stack for programming its GPUs. It includes language extensions, compilers, libraries, and tools that allow developers to access the parallel processing power of the chips. By enabling efficient parallel processing, CUDA significantly accelerates complex computing tasks, especially those associated with machine learning, deep learning, scientific simulations, and data-intensive computations.

[69] See here.

[70] While CUDA remains the dominant optimisation target for many production AI workloads, Google, for example, has been improving support for its TPUs in PyTorch (here and here).

[71] See e.g., Hagiu & Wright (2025), available here.

[72] E.g., this has reportedly been a concern regarding Google's TPUs, as developers were historically required to switch to Jax rather than PyTorch to develop their models. These switching costs may limit the attractiveness of Google's TPUs. To address this, Google and Meta have reportedly partnered to reduce these switching costs by enhancing TPU compatibility with PyTorch, making it easier for developers to switch from NVIDIA's chips to Google's (see footnote 70 and here).

the interface level[73]**, this abstraction only partially mitigates switching costs**. Frameworks such as PyTorch or TensorFlow allow applications to run across different hardware platforms without direct interaction with hardware-specific APIs. However, this portability is mainly effective for generic use cases and early development stages and does not fully address performance optimisation, production deployment and large-scale inference, where hardware-specific dependencies remain significant.

As a result, lock-in effects are uneven across different categories of AI developers. Smaller developers with limited performance requirements may tolerate inefficiencies and remain relatively insensitive to hardware choice. Large technology firms, by contrast, often possess the scale and resources to develop bespoke tools and operate across multiple software ecosystems simultaneously. Between these extremes lies a broad segment of medium-sized firms that lack both tolerance for inefficiency and capacity for deep customisation. For this group, switching costs and dependency on dominant platforms may be particularly constraining.

Even when translation tools exist[74]**, they can hardly ensure full compatibility with CUDA's ecosystem** forcing teams to either sacrifice performance or functionality[75].

Alternative accelerator software stacks do exist, but their adoption remains limited. While options such as AMD's ROCm, Intel's oneAPI/SYCL, and cross-platform open standards like OpenCL can in some cases provide competitive performance or enable portability, their ecosystems are generally less mature than CUDA's[76]. This pattern is reflected in the figure below, where CUDA is consistently the most referred platform in GitHub repositories when compared with alternatives, and its share has increased steadily from 2019 to 2025[77]. While based on a keyword-based analysis and therefore only indicative, the evidence points to a continued strengthening of CUDA's relative importance in AI development.

[73] At this level, developers interact with programming libraries specialised on AI, by calling standard functions in their code (e.g., importing and using PyTorch or TensorFlow libraries), rather than directly programming hardware-specific tools such as CUDA.

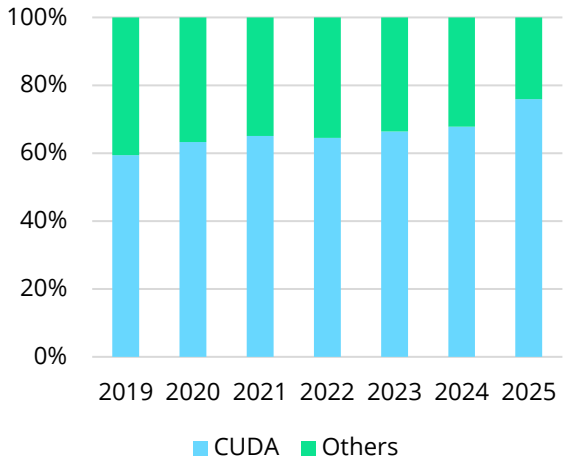
[74] Examples of translation tools include HIPIFY, ZLUDA, SCALE, or Intel's SYCLomatic.

[75] See e.g., Zahid et al. (2024), available here. The authors found that using HIPIFY, a CUDA-to-HIP translator, introduced additional numerical discrepancies between CUDA (on NVIDIA) and HIP (on AMD) code. In addition, according to the AMD documentation, translation of CUDA code to HIP may produce warnings, require manual corrections, and involve iterations of inspection and intervention before the code runs correctly (here).

[76] HPCwire, "Spelunking the HPC and AI GPU Software Stacks", June 2024 (here).

[77] GitHub is the world's largest platform for hosting and sharing software code, widely used by developers to collaborate on open-source and commercial projects. Because it captures a large share of visible developer activity, analysing references in GitHub repositories provides a useful proxy for measuring the relative prevalence of different software stacks in the AI industry.

GitHub repository instances referencing CUDA vs. other accelerator software stacks



Source: Data collected by the AdC in January 2026 based on an automated analysis of public GitHub repositories. The figures report the number of new repositories created each month that reference CUDA compared with those referencing alternative accelerator software stacks (ROCm/HIP, oneAPI/SYCL, OpenCL, Metal, Vulkan, XLA, Neuron).

Differences in ecosystem maturity are also reflected in time lags with which key performance optimisations become available on non-CUDA backends. For example, FlashAttention – a mechanism that reduces both memory footprint and execution time of AI models – became available in CUDA-based stacks in November 2022, whereas comparable support in the ROCm ecosystem emerged only around May 2024, a delay of approximately 18 months. Similarly, bitsandbytes, a library for low-precision inference that reduces memory usage and accelerates large-model workloads, was introduced in CUDA-based stacks in August 2022, with documented ROCm support appearing around December 2024, a lag of roughly 28 months[78].

Hyperscalers are also promoting proprietary software frameworks tied to their own hardware[79], which, by design, do not rely on CUDA. However, these initiatives remain relatively isolated, and they are feasible mainly because hyperscalers have expanded into chip design, as discussed in the preceding section. In this case, **there is a risk that AI developers may not effectively escape technological lock-in but rather exchange one closed ecosystem for another**.

This tight integration between proprietary software frameworks and underlying hardware may contribute to reinforcing barriers to entry and expansion in chip design. Competing firms must, at the same time, match hardware performance, invest in developing a comparable software environment, and overcome network effects by convincing third-party developers to build and optimise tools for their frameworks.

[78] These optimisation techniques were selected based on a Hugging Face guide ("Optimizing LLMs for Speed and Memory", available here), which identifies FlashAttention and low-precision inference (via libraries such as bitsandbytes) as effective methods for efficient LLM deployment. Release dates were obtained from PyPI, using the first publicly available package versions corresponding to CUDA support and the first versions documenting or enabling comparable non-CUDA (e.g. ROCm) support.

[79] Google's XLA/JAX for TPUs and Amazon's Neuron SDK for Trainium and Inferentia chips.

Switching costs in CUDA's ecosystem may reinforce lock-in and barriers to entry

CUDA's position as the prevailing platform for AI development, combined with its lack of compatibility with alternative hardware, may entail high switching costs for developers and raise barriers for competing chip providers seeking to enter or expand in the AI hardware market.

V. Expanding from chip design into networking

As AI workloads scale beyond the capacity of individual accelerators, performance increasingly depends on the efficiency of communication between chips. Training and inference at scale require frequent data exchange across multiple accelerators, making latency, bandwidth and synchronisation critical determinants of system-level performance. In this context, networking technologies become a core component of effective AI compute.

The high-bandwidth interconnects that connect multiple chips within a server are generally proprietary to each chip designer's ecosystem. Examples include NVLink for NVIDIA and Infinity Fabric for AMD. Because these technologies only function within their own ecosystems, they limit the possibility of combining different AI chips[80].

Despite some initiatives to broaden access to these technologies, these arrangements are still dependent on proprietary architecture and technical conditions[81]. In parallel, industry groups are developing open standards like UALink, intended to create a vendor-neutral layer for interoperability across accelerators[82].

When AI workloads scale beyond the compute and memory capacity of a single server, they must be distributed across clusters of interconnected machines. These clusters rely on high-speed networks to handle that communication efficiently.

Contrary to intra-server scale, interconnects between servers rely more on open standards. Ethernet is the prevailing technology[83], widely used in datacentres[84] and open to all vendors.

An alternative to Ethernet is InfiniBand, a proprietary interconnect technology owned and developed by NVIDIA, and particularly used in high-performance computing[85]. InfiniBand delivers lower latency and higher throughput

[80] In general, mixing chips from different chip designers requires the communication to fall back to PCIe technology, the open standard technology, which has lower bandwidth and higher latency.
[81] An example is NVIDIA's NVLink Fusion, announced in May 2025 as a licensing program that allows selected partners to integrate NVLink ports into their own CPUs or custom accelerators. The initiative broadens access beyond NVIDIA's own GPUs, but participation is conditional, as devices must still connect through NVIDIA's products, like NVSwitch, and rely on NVIDIA's software to manage the communication across the devices (see more here).
[82] See UALink consortium's website here.
[83] Lightwave, "Ethernet maintains a lead over InfiniBand in the AI race", September 2025 (here).
[84] As it is the example of Azure, Oracle, or Meta (here) in recent projects.
[85] The TOP500 list (here), which ranks the world's 500 most powerful supercomputers based on standardised performance benchmarks, indicates that InfiniBand interconnects are used in 57% of all listed systems as of November 2025. The prevalence of this NVIDIA's technology is even higher among recent installations, rising to 65% for systems deployed within the last three years.

than Ethernet, but typically at a higher cost[86]. NVIDIA positions InfiniBand as the benchmark for frontier-scale training, while also promoting its own Ethernet solutions[87].

As chip designers expand into the networking layer, there is a risk that market power at the chip level could be leveraged into adjacent segments of the AI compute infrastructure. Such effects could arise not only through technical integration but also through contractual arrangements, such as exclusivity clauses, bundling practices[88], or loyalty rebates, that might reduce incentives for customers to adopt alternative suppliers[89].

Control over interconnect software may extend market power across the AI stack

As chip designers expand into the networking layer, they may reinforce lock-in effects through technical integration or contractual practices, increasing the risk of leveraging their chip-level position into adjacent segments of the AI stack.

[86] See more here and here.
[87] Fierce Electronics, "Nvidia envisions 'one gigantic GPU' by linking data centers", August 2025 (here).
[88] The US Department of Justice has reportedly launched an investigation into NVIDIA concerning allegations that the company may have abused its market dominance in the supply of AI chips, including potential preferential supply and pricing practices favouring customers purchasing its complete systems (from hardware to networking components). See The Information, "Nvidia Faces DOJ Antitrust Probe Over Complaints From Rivals", August 2024 (here).
[89] In this respect, similar concerns have been identified by other competition authorities and international organisations, including: (i) OECD Roundtables on Competition Policy Papers, No. 330, "Competition in artificial intelligence infrastructure", November 2025 (here); (ii) Autorité de la Concurrence, "Opinion 24-A-05 on the competitive functioning of the generative artificial intelligence sector", June 2025 (here); and (iii) European Commission, Policy brief "Competition in Generative AI and Virtual Worlds", September 2024 (here).

COMPETITION AND GENERATIVE AI: ACCESS TO AI CHIPS – KEY HIGHLIGHTS

Structural bottlenecks in the AI chip supply chain may undermine market access

Persistent supply constraints, high concentration across the value chain, and rising demand for cutting-edge hardware have created bottlenecks in the production and delivery of AI chips, limiting smaller developers' ability to access the compute needed to compete in advanced AI development.

Public HPC initiatives may expand access to AI compute, although their scale remains limited

Public HPC infrastructures may broaden access to large-scale computing capacity and partially alleviate supply constraints. While still limited in scale, recent European initiatives seek to expand their role in supporting AI development. Their impact on competition depends on transparent and non-discriminatory access criteria.

Expanding control across the AI stack may increase market concentration

As cloud providers move into chip design and chip providers expand into cloud services, control over hardware and infrastructure may become more concentrated. This could reinforce ecosystem incentives that lock-in developers and raise concerns about interoperability, self-preferencing, and foreclosure risks.

Switching costs in CUDA's ecosystem may reinforce lock– in and barriers to entry

CUDA's position as the prevailing platform for AI development, combined with its lack of compatibility with alternative hardware, may entail high switching costs for developers and raise barriers for competing chip providers seeking to enter or expand in the AI hardware market.

Control over interconnect software may extend market power across the AI stack

As chip designers expand into the networking layer, they may reinforce lock-in effects through technical integration or contractual practices, increasing the risk of leveraging their chip-level position into adjacent segments of the AI stack.