KONSERT
STRATEGY & IP

The race to own the AI data center infrastructure stack

How changing technology and commercial dynamics are redrawing competitive boundaries

At a glance

AI is reshaping data center infrastructure in both technical and commercial dimensions.

On the technical side, the data center infrastructure technology stack is becoming more tightly integrated. Power density requirements are increasing, forcing new electrical designs. Cooling is moving from air to liquid and becoming part of system architecture. Racks are evolving from passive enclosures into integrated units that combine compute, power, and thermal management. Data centers are being redesigned as AI-specific environments rather than general-purpose facilities.

On the commercial side, the market is concentrating. Demand sits with a small group of hyperscalers and frontier labs whose capex plans set the pace. Supply is dominated by a short list of specialized vendors whose bargaining position has only strengthened. A new class of dedicated AI infrastructure providers has also emerged that is absorbing capex risk.

Future advantages will be with companies that can hold and defend positions in this shifting market through a combination of technical leadership, operational execution, capital, and intellectual property.

AUTHORS

Aniket Rao
Manager

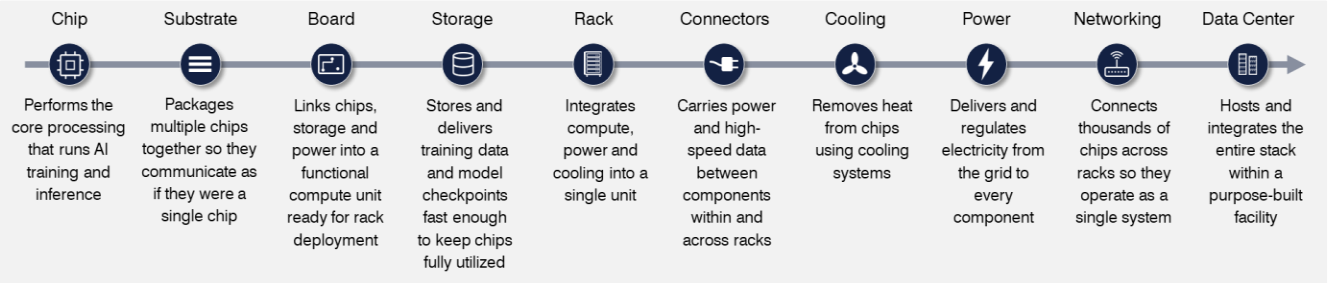
The AI buildout is driving technical integration and commercial consolidation in equal measure

Historically, most of the data center value chain could be optimized relatively independently. Compute, networking, power, cooling, and facilities were loosely coupled. Data centers were designed for a wide range of workloads such as enterprise SaaS, web hosting, storage - and demand came from a long tail of enterprise buyers. Traditional data centers typically run at 6 to 15 kW per rack, rely on air cooling, and follow general-purpose layouts.

However, AI training and high-density inference workloads are demanding something far more specialized and integrated: rack densities of 50 to 130 kW, liquid cooling, and facilities designed from the ground up around GPU clusters and high-bandwidth interconnects. In these environments, improvements in one layer cascade through all the others. A higher-power chip generation outstrips existing power distribution and forces new electrical designs. The heat it generates requires a shift to liquid cooling. Faster compute creates new interconnect bottlenecks, reshaping cluster topology and how floor space is allocated.

AI data center infrastructure is therefore being designed as a system rather than as a collection of independent components.

Illustration 1: AI data center infrastructure stack



This integration is also happening at a commercial level. The buyer and supplier side is consolidating into a small number of companies and frontier labs whose capex commitments are now driving almost every meaningful commercial contract. Additionally, a new class has emerged: dedicated AI infrastructure providers that are taking on the capex burden of GPU clusters and selling capacity back to the same companies and labs.

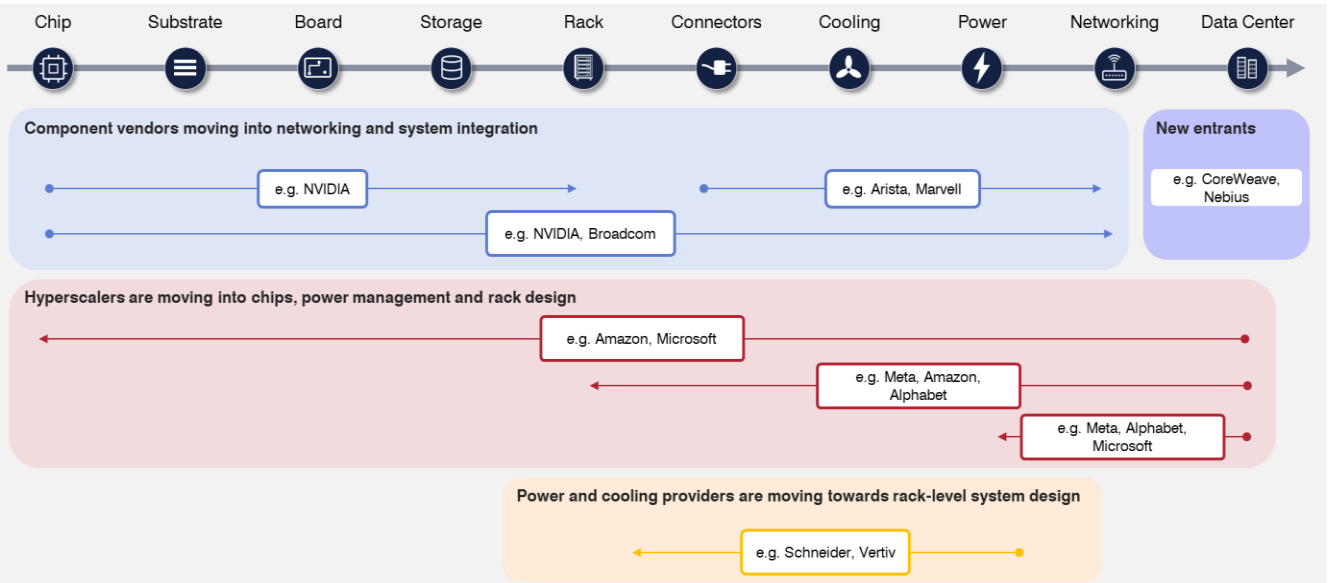
However, not all positions in this technical and commercial integration are equal. Some shape system performance, deepen customer dependency, or reinforce ecosystem lock-in: Chips and networking are the most obvious examples. Others act as bottlenecks where deployment is difficult, capacity is constrained, or substitution is limited: Power, cooling, and data center design increasingly fit that description. A third set derives strategic value from only their commercial position: Deploying capital ahead of demand, securing long-term customer commitments, or standing between buyers and suppliers whose own balance sheets cannot fund the buildout alone.

The strategic question is therefore not only which technical position to occupy, but also which commercial position to take in a market with new technical considerations, concentrated demand, concentrated supply, and a new set of intermediaries.

Companies are moving beyond their traditional positions

We are already seeing companies move beyond their traditional positions to take new, broader positions where technical performance, economics, and deployment become harder to separate.

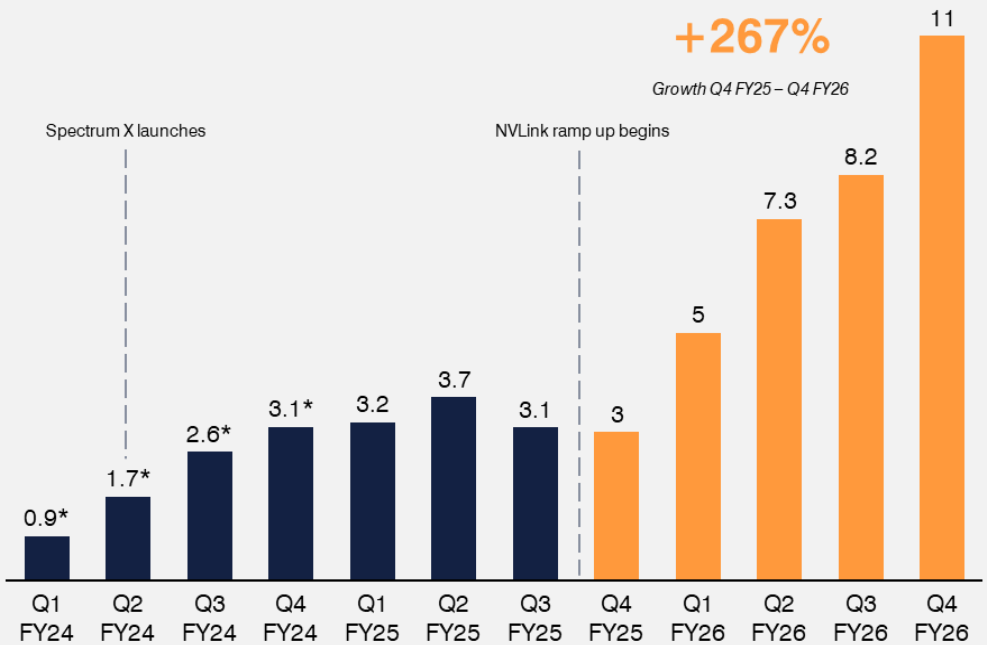
Illustration 2: Company movements within the AI data center infrastructure stack



Chip companies are moving into networking and rack design

NVIDIA has expanded beyond accelerators into networking, systems, software, and full-rack architectures. The rapid rise in networking revenue, driven primarily by NVLink compute fabric for its Blackwell systems and increasingly by Spectrum-X in the Ethernet for AI market, highlights how NVIDIA is now building the entire AI data center fabric. NVLink is NVIDIA's proprietary high-bandwidth interconnect that binds GPU clusters together at rack scale, while Spectrum-X (its AI-optimized Ethernet platform combining switches, DPUs, and software) is designed to ensure GPUs can communicate efficiently across clusters, together turning networking into a critical performance layer rather than a commodity. Broadcom's position is now spanning custom silicon, connectivity, and AI data center networking infrastructure. These moves reflect a broader reality: in AI, chip performance alone is not enough. Increasingly, the value lies in controlling how compute is connected, integrated, and deployed. The bottleneck is therefore the efficiency of the system, and particularly the network that binds it together.

Figure 1: NVIDIA quarterly networking revenue (in USD billions, Q1 FY2024 – Q4 FY2026)



NVLink and Spectrum-X have extended NVIDIA's reach beyond GPUs into the full networking fabric

Source: NVIDIA quarterly commentaries.

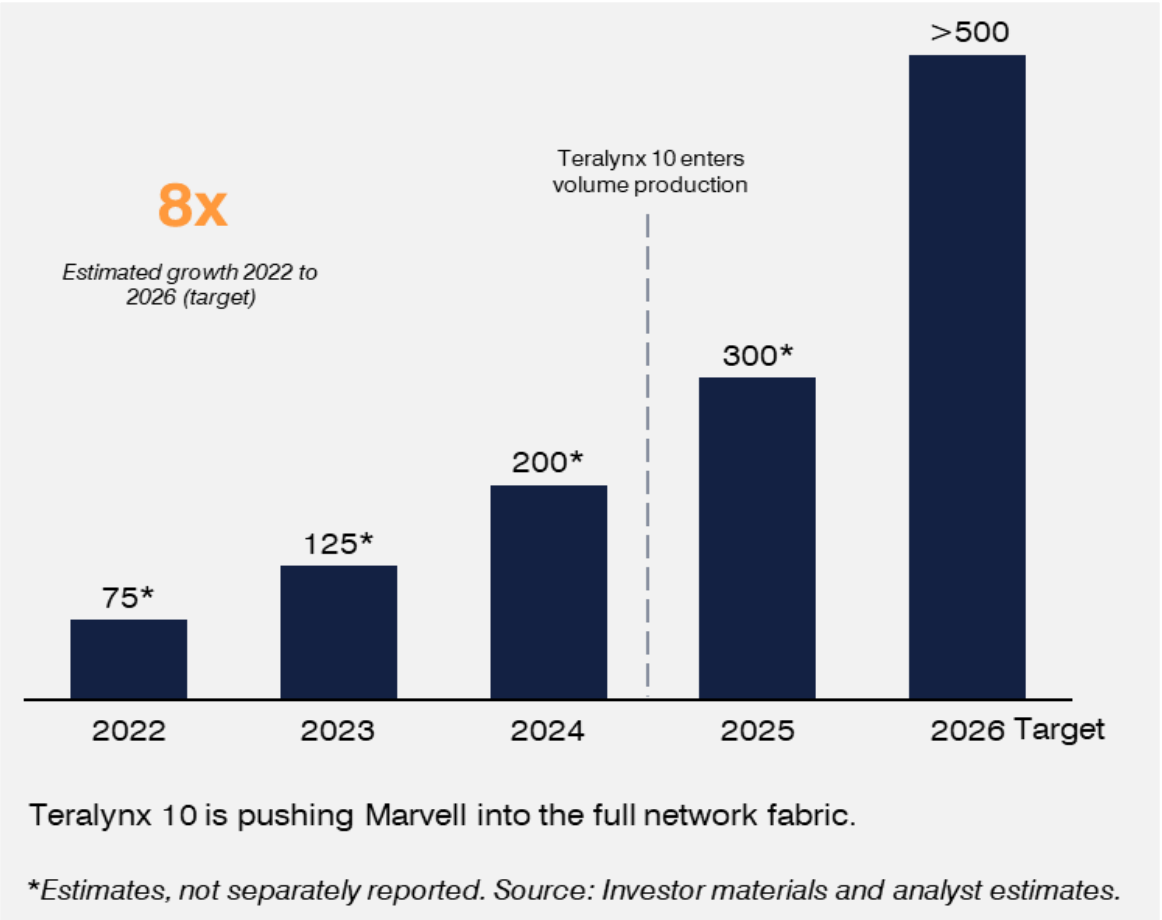
** Estimated. Discrete networking revenue not separately reported by NVIDIA prior to Q1 FY25.*

Networking is shifting from commodity infrastructure to a core determinant of system performance

Companies such as Arista Networks and Marvell Technology are becoming more central as AI workloads scale across clusters and data centers. Marvell's data center switching revenue is projected to grow significantly through 2026, driven in part by its Teralynx 10 platform entering volume production. Teralynx 10 is Marvell's latest high-performance switching silicon, designed for AI data center networks, enabling high bandwidth, low latency, and efficient traffic management across large GPU clusters. It represents a shift beyond traditional interconnect components towards owning a larger share of the switching and network fabric layer.

In AI systems, interconnect performance directly shapes cluster efficiency, utilization, and overall workload economics, as thousands of accelerators must operate in tightly synchronized environments. This makes networking a core part of the system architecture, where performance bottlenecks can shift from compute to the fabric that connects it.

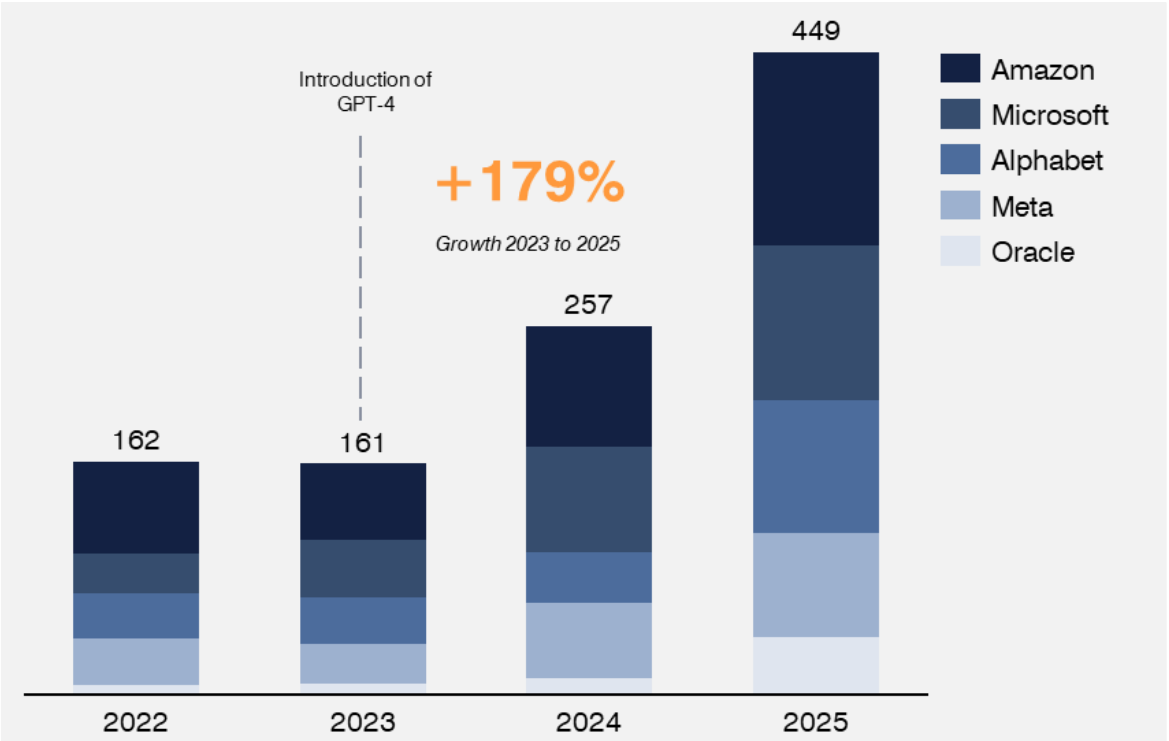
Figure 2: Marvell data center switching revenue (in USD millions, 2022-2026)



Hyperscalers are moving from data center operators to data center infrastructure owners

Amazon, Microsoft, Alphabet, Meta, and Oracle are now committing capital at a level that suggests something more fundamental than capacity expansion. They are increasingly shaping the technology stack through custom silicon, AI-optimized networking, rack-scale system design, and AI-ready data center deployment. This means that they are looking at influencing more of the stack that matters most for AI performance, economics, and deployment speed.

Figure 3: Capital expenditure of major hyperscalers (in USD billions, 2022-2025)



Capex growth is being driven primarily by AI data center buildouts.

Source: Company filings and investor communications

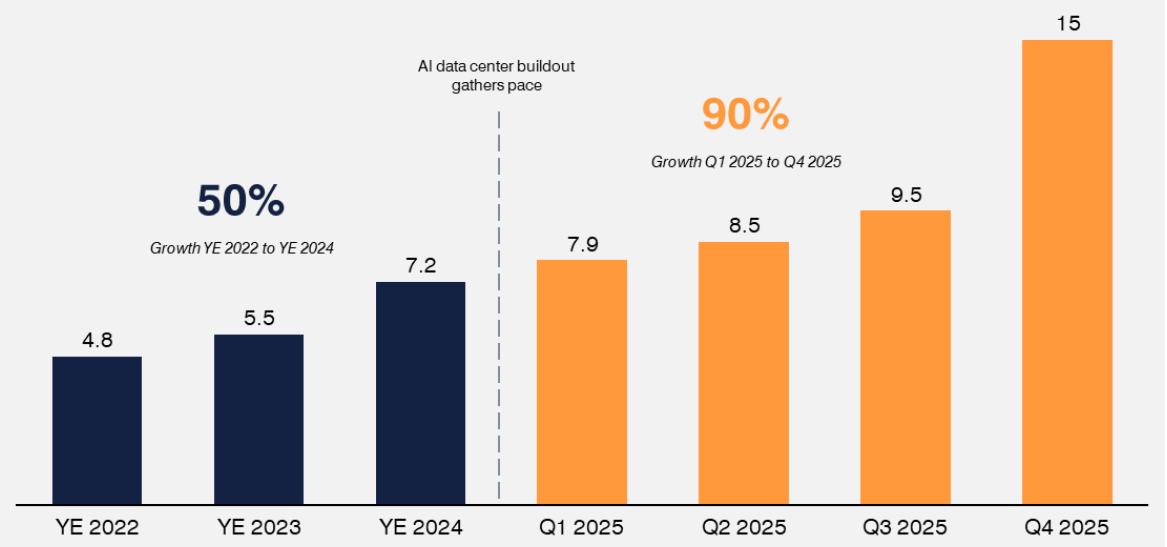
Power and cooling are moving from support functions to strategic positions

Power and cooling are becoming binding constraints on AI deployment. As power density increases and workloads scale, the ability to deliver electricity, manage heat, and prepare sites is increasingly determining how quickly and efficiently AI capacity can be deployed. Companies such as Vertiv, Schneider Electric, and Eaton are benefiting directly from this shift.

Vertiv's order backlog, for example, grew 50% between 2022 and 2024 as data center demand built steadily, then accelerated 90% in 2025 alone as power and cooling capacity became the binding constraint on AI deployment, reaching $15 billion by year-end. Schneider Electric is seeing sustained growth in its data center and networks segment, which is expected to account for around 30% of its revenue mix, driven by demand for power and cooling systems. Eaton is expanding into integrated power and cooling solutions, including acquisitions to strengthen its position in thermal management for AI data centers.

What was once treated as infrastructure overhead is now a determining factor in deployment timelines, operational reliability, and total cost of ownership.

Figure 4: Vertiv's reported order backlog (in USD billions, YE 2022 – Q4 2025)



Vertiv's backlog grew steadily for two years, then nearly doubled in 2025 alone as power and cooling capacity became the binding constraint on AI deployment.

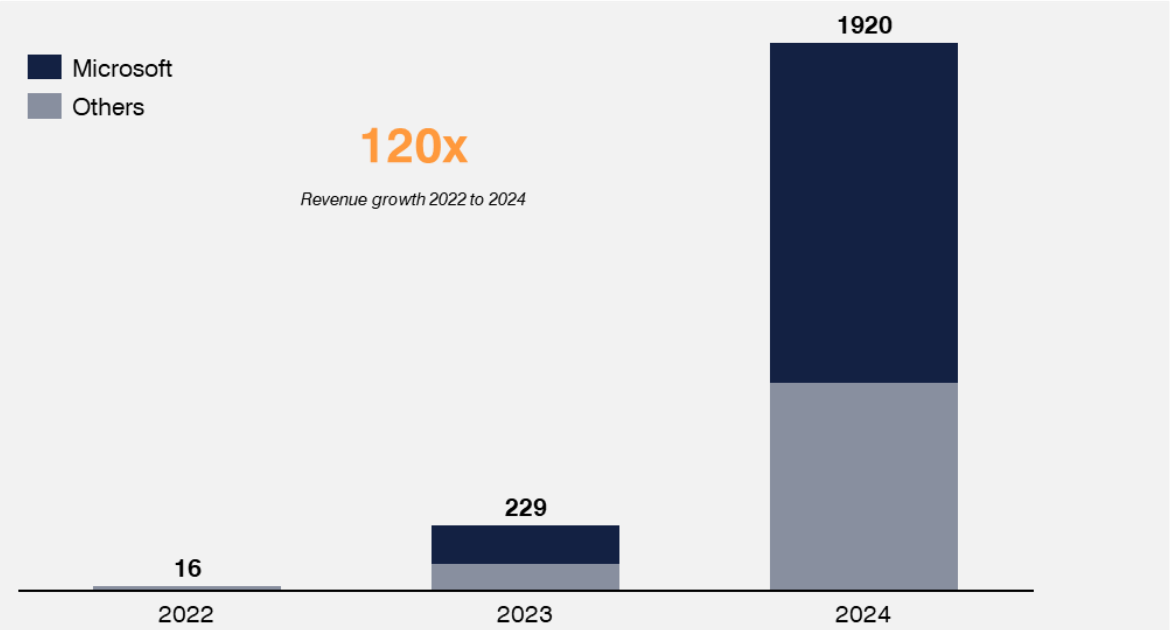
Source: Vertiv's quarterly earnings releases and recent filings

A new class of AI infrastructure providers is absorbing the capex risk that hyperscalers and frontier labs cannot take on alone

Alongside hyperscalers, a parallel set of dedicated AI infrastructure providers have grown rapidly. CoreWeave and Nebius take on the capex burden of building and operating GPU clusters at scale and lease capacity back to hyperscalers, frontier labs, and enterprises whose own deployment timelines or balance sheets cannot meet demand alone. Their economics are distinct from both colocation providers and hyperscalers. CoreWeave, for instance, derives most of its revenue from a single customer (Microsoft). Nebius, NVIDIA-backed, has pivoted into a similar position in Europe.

These providers are an important part of the market because they absorb deployment risk that no one else is currently positioned to take. They allow hyperscalers to use capacity without committing all the capex up front and allow frontier labs to scale training faster than they could on their own infrastructure.

Figure 5: CoreWeave revenue and customer concentration (in USD millions, 2022 – 2024)



Rapid growth and concentrated demand define this layer; Nebius is on a similar trajectory in Europe

Source: CoreWeave public filings.

Positioning in a moving market, and where IP fits in

The AI data center infrastructure market is still taking shape. Who will ultimately control value, and where, remain open questions. The driving technology forces of the AI revolution continue to evolve, and breakthroughs that upend the value chain may still be on the horizon. In the core infrastructure layers, however, the development landscape is maturing, value chain roles are becoming clearer, and competitors are either dropping out or narrowing the gap to the leaders. The positions that matter are still forming, but they will not remain accessible indefinitely.

Competitive advantage in AI infrastructure rests on technical leadership, operational execution, capital, and intellectual property. Of these, IP is one of the parameters that can still alter outcomes where the market otherwise looks path-dependent. For movers, IP amplifies value chain plays and ensures they capture the margins their technology and capital deployment make possible. For incumbents in layers that are commoditising, IP is one of the few remaining moats: a way to hinder or steer well-financed movers and new entrants who would otherwise displace them.

IP pressure is also building from outside the value chain. The complexity of the data center supply chain, with components passing through multiple vendors before reaching the facility, makes it difficult to identify who owes what and to whom. Some of the most sophisticated patent aggregators and non-practising entities have recognised the opportunity this creates and are beginning to act, with cooling technology already the subject of active litigation. In a market where regulatory approval and financing stability are prerequisites for breaking ground, IP exposure will shape who can build, at what cost, and on whose terms.

Companies that want to be part of the next generation of winners need to choose where to compete, move into the layers that matter, and secure those positions through IP before market dynamics lock in.

www.konsert.com