McKinsey & Company

Technology, Media & Telecommunications Practice

The next big shifts in AI workloads and hyperscaler strategies

AI workloads are reshaping data center economics, power planning, and leasing decisions. Their influence is already evident across most new builds and will only intensify in the coming years.

This article is a collaborative effort by Chhavi Arora, Marc Sorel, and Pankaj Sachdeva, with Arjita Bhan, Jess He, Nicholas Shaw, Riya Garg, and Shriya Ravishankar, representing views from McKinsey's Technology, Media, and Telecommunications Practice.

December 2025

AI is now the primary growth engine for data centers in the United States and is projected to be one of several drivers that will grow in supply and increase power capacity from about 30 or more gigawatts (GW) in 2025 to 90 or more GW or more by 2030, a CAGR of approximately 22 percent. That capacity is larger than the entire power demand of California today[1]—and it's completely reshaping the nation's data center infrastructure.

Hyperscalers are expected to capture about 70 percent of the forecast capacity in the US market through owned or leased options,[2] and their infrastructure decisions will define how the broader data center ecosystem evolves. AI compute today is primarily split between two workload types: training and inferencing. Both workloads are rapidly shaping hyperscaler strategies and are driving a paradigm shift in site selection, power strategy, and architectural design across hyperscalers' portfolios.

Training workloads are driving the need for large-scale, high-density campuses with advanced mechanical, electrical, and plumbing (MEP) systems and specialized hardware integration patterns. Meanwhile, AI inference workloads are accelerating site build-outs in greater metro and surrounding areas that are optimized for low round-trip time, high network interconnectivity, and energy efficiency. What's more, our research finds that inferencing workloads are projected to make up a little more than half of AI workloads by 2030, causing hyperscalers to reconsider their design approach and the location of data center builds. And energy constraints are changing the way hyperscalers think about the market and ways to build faster.

This article explores how these trends are reshaping hyperscaler strategies, outlining five key shifts in how they are choosing to scale, meet surging AI demand, and optimize capacity. It also examines how hyperscalers are balancing greenfield expansion and brownfield retrofits, as well as how capital is being redeployed across the data center value chain to sustain this rapid growth.

AI compute is tilting toward a high-availability, inference-heavy future

Training workloads focus on developing and refining large language models and other AI systems. These workloads require high power densities of 100 to 200 or more kilowatts (kW) per rack, specialized low-loss interconnects, and advanced liquid-cooling systems to sustain compute-intensive jobs.[3] Training workloads are insensitive to latency and can tolerate delays of up to 100 milliseconds between adjacent regions,[4] allowing hyperscalers to site them in remote, power-rich areas where grid capacity, land, and water are more available.

Inference workloads deploy and operate these trained models. They power real-time applications such as search, chatbots, and recommendation engines and require 30 to 150 kW per rack.[5] Older hardware can be repurposed to accommodate simultaneous atomized workloads. While training costs are capital-intensive and often hard to link directly to commercial impact, inference costs are usually recurring and directly tied to revenue generation.

[1] "Current and forecasted demand," California ISO, accessed December 2, 2025.
[2] "AI power: Expanding data center capacity to meet growing demand," McKinsey, October 29, 2024.
[3] "Power update: A surging data center tide lifts the power sector," S&P Global, October 7, 2025.
[4] *AI Network Blog*, "The anatomy of a training job: Where your billion-dollar GPUs actually spend their time," blog entry by Art Fewell, May 26, 2025.
[5] "Power update: A surging data center tide lifts the power sector," S&P Global, October 7, 2025.

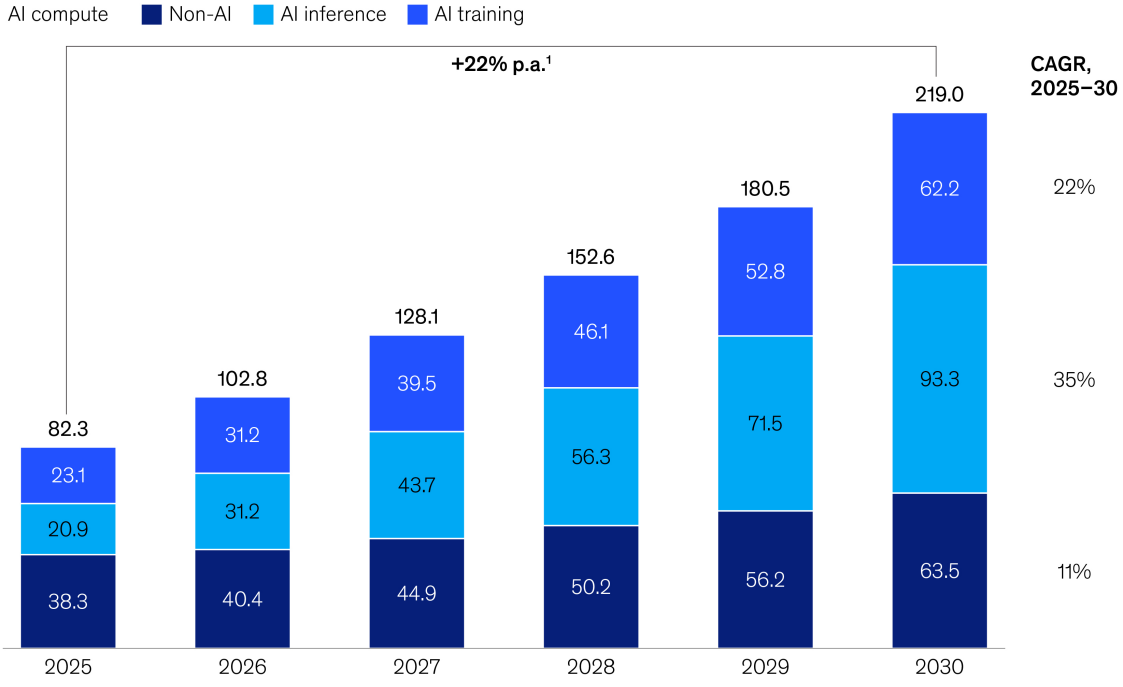
As both training and inference footprints expand, hyperscalers are demanding higher levels of infrastructure resiliency to ensure uptime and service continuity. Many new AI-ready data centers are being designed to meet full 2N redundancy standards[6] to minimize the risk of downtime from component or utility failures. This shift toward fault-tolerant design reflects both the criticality of continuous AI operations and the growing economic exposure tied to inference-driven applications.

By 2030, inference will surpass training to become the dominant workload in AI data centers, representing more than half of all AI compute and roughly 30 to 40 percent of total data center demand (Exhibit 1). This move from one-time model training to sustained inference activity will increasingly shape hyperscalers' site strategy, network design, and power provisioning.

Exhibit 1

Inference workloads could make up more than 40 percent of data center demand in 2030, growing 35 percent CAGR until 2030.

Global data center demand by workload, 2025–30, gigawatts



Note: Includes all provider types.
[1]Per annum.
Source: McKinsey Data Center Demand Model

McKinsey & Company

[6] These data centers have two completely independent power and cooling systems that can handle full loads.

That said, the precise growth trajectory of inference workloads remains uncertain. While query volumes and reasoning-heavy workloads continue to rise, several efficiency curves are improving just as quickly. Hardware advances are lowering the amount of energy needed per compute (tokens per watt). Software optimizations, including shifts toward smaller, fine-tuned models, are further reducing runtime requirements. And improvements in precision formats and model specialization are advancing just as rapidly. These trends could moderate growth toward CAGRs of 4 to 7 percent, particularly because permitting timelines, regulatory limits, and regional grid constraints cap how quickly new capacity can come online.

Designing for dual frontiers: Power-hungry training versus round-trip, latency-critical inference

AI infrastructure needs are split between the two workloads. Training workloads will demand up to one megawatt (MW) per rack in some frontier systems,[7] requiring the use of ultradense stacks for graphics processing units (GPUs) and tensor processing units as well as liquid cooling. By contrast, inference workloads, though still far above legacy compute in terms of kW draw, operate at 30 to 150 kW per rack,[8] aligning more with enhanced cloud-compute infrastructure than full high-performance computing infrastructure. Part of this is because inference workloads are highly atomizable—in other words, individual tasks can be handled independently—unlike training, which relies on large-scale, tightly synchronized GPU clusters. The result is two radically different build archetypes, each determining where and how hyperscalers build for AI.

Inference workloads are often co-located with applications and storage to minimize latency. In terms of MEP and hardware requirements, training workloads demand advanced switchgear and resilient, uninterruptible power supply systems that are battery-backed to absorb the sharp step-load events that rapid GPU power fluctuations cause. These fluctuations stem from highly variable computational intensity during AI training cycles. As GPUs move between phases of training events, such as matrix multiplication, memory transfer, and synchronization, instantaneous power draws can surge or decline by 30 to 60 percent within milliseconds.[9] Managing these surges requires oversized power-delivery networks, harmonic filtering, and fast-response uninterruptible power supply systems to prevent voltage dips and protect downstream components.

AI training data centers demand much higher power densities than standard data centers because of the intense computational requirements of training large AI models. Data center demand for training AI is expected to grow at a CAGR of 22 percent over the next five years, reaching more than 60 GW by 2030. But as inference workloads become more dominant—expected to grow at a CAGR of 35 percent over the next five years and reach more than 90 GW by 2030—data centers will adapt to support inferencing at scale, focusing on real-time, low-latency processing. As such, a shift from large, power-intensive facilities to smaller, modular, and distributed data centers will be expected to meet data center capacity at the GW scale.

[7] Erik van Klinken, "How data centers are making the giant leap to 1 megawatt per rack," Techzine, October 22, 2025.
[8] "Power update: A surging data center tide lifts the power sector," S&P Global, October 7, 2025.
[9] Imran Latif et al., "Empirical measurements of AI training power demand on a GPU-accelerated node," arXiv, December 11, 2024.

Cloud campuses are morphing into mixed-use engines for inference and general compute
Already, about 70 percent of new core campuses combine general compute and inference, often separated by building or data halls. To ensure instant responsiveness, next-generation data center designs are co-locating inference clusters within existing cloud-campus footprints rather than isolating them in stand-alone training sites. Inference racks are fixed closer to access points, storage, and networking zones, redrawing the blueprint of traditional core-cloud campuses (Exhibit 2).

Exhibit 2

To avoid latency, inference workloads should be housed with core cloud assets, while training workloads can be separate.

	Strategic considerations	Cloud compute	AI inference	AI training
IT requirements	IT capacity	40–200 megawatts (MW)	50–250 MW[1]	300 MW to 2 gigawatts
	Latency	~15 milliseconds (ms) between adjacent regions	~15 ms between adjacent regions	100 ms or more
	Resiliency	Tier 4 or 4+	Tier 4	Tier 3 or 4
	Power density	1–50 kilowatt (kW)/rack	30–150 kW/rack	100–200 kW/rack, proof-of-concept ~1 MW/rack setups
Infrastructure requirements	Cooling	Air cooling	Liquid cooling preferred[1]	Liquid cooling required
	Network	Standard configuration	Standard configuration with some adjustments (eg, NVLink interconnect, limited)	Specialized networking solutions that are more complex and fragile (need near-zero-loss cabling)
		AI inference workloads are integrated within historically cloud-dominated campuses to ensure low-latency access to data and tools		Training is typically located in dedicated training campuses that can be removed from main cloud sites because latency is not an issue

[1]Data center campuses with capacity of 50 MW or more and liquid cooling are best positioned to handle shifting workloads; sites without these capabilities risk higher material costs and lower leasing yields.

McKinsey & Company

Smaller data centers will be connected by high-speed networks, optimized for real-time inferencing and large-scale training. A significant portion of inferencing will continue shifting to the edge, reducing latency and bandwidth demands. Moreover, these new data center builds will

increase the adoption of power-efficient hardware, such as custom silicon (for example, application-specific integrated circuits and neural processing units and ARM-based[10] architectures.

Tier 2 markets rise as power bottlenecks strain the coasts

Time to power[11] has become the biggest limiting factor for new capacity, often dictating where and how hyperscalers expand. In 2022, the time to market in unconstrained markets was between 12 and 18 months. Now, time to market can be 36 months or longer in regions such as northern Virginia, prompting hyperscalers to value speed and power assurance as much as cost efficiency. As a result, hyperscalers are no longer choosing partners only for capital efficiency but also for their ability to gain access to power faster, secure permits, and deliver the capacity needed for sustained AI growth.

Tier 1 hubs, such as northern Virginia and Santa Clara, together account for roughly 30 percent of US data center capacity. Now, these locations are constrained by grid congestion, multiyear permitting timelines, and land prices exceeding $2 million per acre. As a result, hyperscalers are pivoting toward tier 2 markets—including Des Moines, Iowa; San Antonio, Texas; and Columbus, Ohio—where power can be delivered 12 to 24 months faster and land costs are up to 70 percent lower (Exhibit 3. This rebalancing is pushing hyperscalers to adopt power-first and energy-anchored site selection models by, for example, partnering directly with states' utility providers to build new data centers.

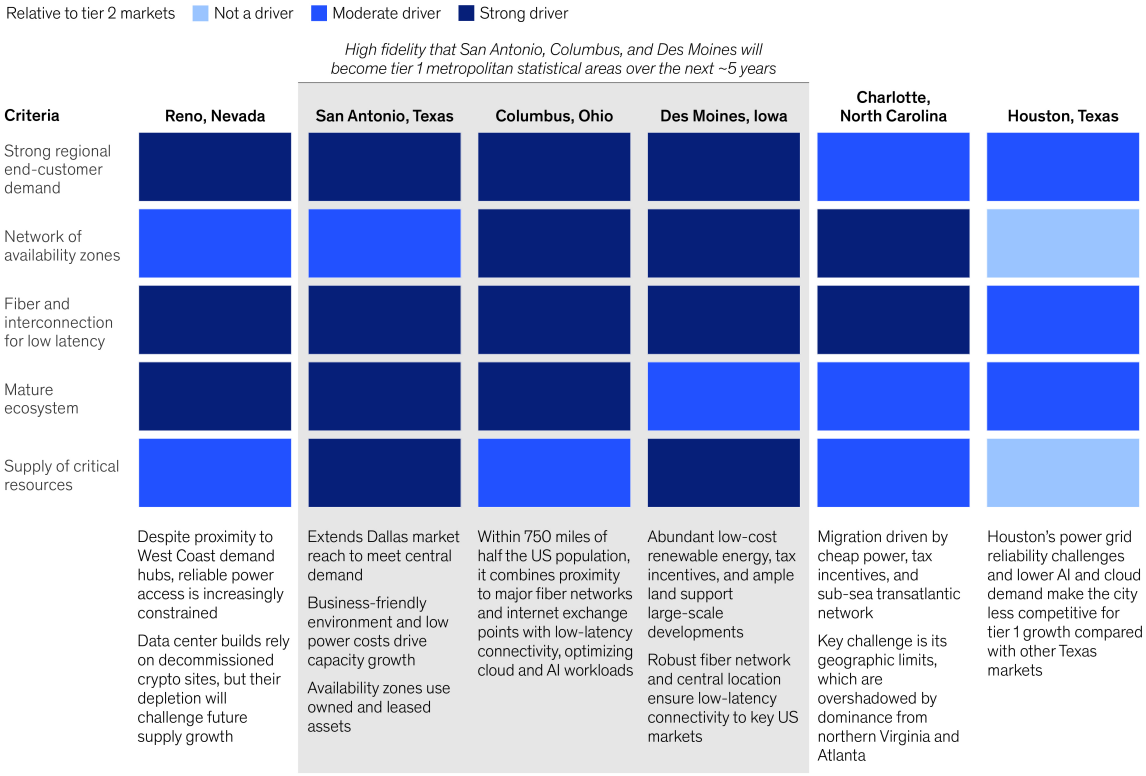
While data centers of less than 500 MW are still frequently self-funded, larger, multi-GW campuses increasingly rely on joint ventures (JVs with infrastructure funds, utilities, or private credit partners. The capital intensity (up to $25 million per MW and speed-to-market pressures require hyperscalers to use various capital deployment strategies, including using developer or fund-backed balance sheets rather than waiting for internal capital cycles. While these structures broaden capital access, they often introduce complexity as more parties get involved. For example, aligning interests, negotiating contracts, allocating risk, and structuring exit strategies can become more difficult, which can slow preconstruction steps, due diligence, and legal negotiations. Moreover, utility partners can require more negotiations for grid upgrades, capacity reservations, and permitting coordination to serve new loads, which can add to timelines.

Some developers are experimenting with behind-the-meter systems (such as fuel cells, microgrids, and small modular reactors and direct power purchase agreements to reduce their dependence on the grid and accelerate the time it takes to energize a data center. For example, New APR Energy is deploying mobile gas turbines delivering more than 100 MW of behind-the-meter power to a US hyperscaler.[12] Additionally, Active Infrastructure is planning a 362-acre campus in northern Virginia that incorporates hydrogen fuel cells, a microgrid, and battery storage for primary on-site power generation.[13] Access to entitled land and reliable energy has become a strategic differentiator, shaping hyperscaler footprints and their attractiveness as long-term investments.

[10] Advanced reduced instruction set computer machine.
[11] The time required for a location to receive the power it needs to operate.
[12] "New APR Energy deploys 100MW+ of mobile gas turbines for U.S. based AI hyperscaler," APR Energy, February 25, 2025.
[13] Georgia Butler, "Active Infrastructure looks to develop hydrogen-powered data center in Leesburg, Virginia," Data Centre Dynamics, October 28, 2024.

Exhibit 3

Three cities in tier 2 markets have potential to transition into tier 1 markets as hyperscalers pivot.



Source: "North America data center trends H1 2025," CBRE, September 8, 2025; McKinsey analysis

McKinsey & Company

Hyperscalers are adjusting in five major ways

Hyperscaler strategies now set the pace for how digital infrastructure is financed, owned, and operated and how the broader ecosystem responds. Their priorities are shaping where capital flows, revealing opportunities in energy-efficient systems, modular builds, and advanced liquid-cooling technologies. As AI workloads change and scale, hyperscaler strategies are shifting in five major ways.

1. Hyperscalers continue to invest in power to keep scaling AI
As energy becomes a competitive and constrained resource for data centers, hyperscalers are shifting from being passive consumers to active participants in the power ecosystem. The challenge is no longer just about cost; it's about securing reliable, scalable, and ideally clean energy to support the exponential increase in AI demand. For instance, hyperscalers have invested in next-generation energy sources, including small modular reactors and fusion partnerships, to secure clean, scalable power.

Stakeholders are now debating whether to deploy capital into building or cofinancing new infrastructure (such as renewable-energy generation, storage, or next-generation nuclear plants or to move upstream and invest in the developers and suppliers that can guarantee long-term access to clean power. This debate reflects a broader strategic shift: Energy has become a core determinant of capacity growth, and access to it can offer companies a competitive advantage. Regulatory pressures, sustainability commitments, and regional grid bottlenecks are all shaping investment choices in this vein.

Increasingly, hyperscalers are creating special-purpose vehicles and JVs to finance large data center builds to reduce the credit burden and derisk holding large long-term infrastructure assets under their own books, including activities such as direct bond issuance to finance large projects.

2. Hyperscalers are trading ownership for speed-to-power and market access
With time to market averaging 24 to 36 months in power-constrained regions, leasing remains critical to secure near-term compute and power capacity and bridge the gap with customer demands. Yet hyperscalers still seek long-term control of core sites. A McKinsey study finds that, as a result, lease-to-own models now make up 25 to 30 percent of new tier 1 deals, particularly in tier 1–constrained markets such as northern Virginia and Santa Clara, where land and power scarcity make flexibility essential.

Lease renewal rates remain high at about 90 to 95 percent and will likely continue to rise, particularly in tier 1 markets. While hyperscalers pursue AI workloads, they also recognize the strategic importance of retaining sites critical to end-customer demand. To maintain their footprint while also meeting changes in demand, there is growing readiness among operators to divest smaller or non-retrofittable sites and shift their portfolio toward AI-optimized megacampuses and availability zones (AZs).

3. Modular and prefabricated constructions are serving as accelerants for hyperscalers
Rapid AI demand is driving adoption of prefabricated and standardized builds, which can cut delivery timelines by up to 50 percent compared with stick builds.[14] Preapproved powered shells reduce time to market by 50 to 70 percent. Importantly, modular builds are now liquid-cooling ready, enabling hyperscalers to handle next-generation rack densities at scale.

More modular construction will be tailored to hyperscaler preferences, which will drive a higher percentage of prefabricated solutions in data centers. Construction and manufacturing companies that cater to distinct styles can help hyperscalers come online faster.

[14] Jose Luis Blanco, Dave Dauphinais, Garo Hovnanian, and Rob Palter, "Making modular construction fit," McKinsey, May 10, 2023.

4. Hyperscalers are transitioning from scattered sites to unified multifacility campuses
Hyperscalers are consolidating workloads into large, AI-ready campuses and retrofitting portions of their existing fleets to support liquid cooling and higher rack densities, where power, cooling, and structural upgrades can be scaled efficiently.

Hyperscalers are evolving their AZ models, shifting from stand-alone facilities to multifacility clusters. These campuses are projected to account for about 70 percent of deployments by 2030. This trend underscores the interest of companies to cluster their data centers for operational and intra-AZ failover purposes, as compared with allowing stand-alone data centers scattered in various parts of the country to act as their own AZs. For example, should an energy source or piece of software fail, multifacility clusters can replicate and failover data and processes easier than if centers were housed in multiple locations.

From an operational perspective, this approach is advantageous. Housing campuses closer together and consolidating these centers allows operators to maintain control over a sizable footprint without having to staff and oversee several facilities in spread-out locations.

Where retrofits aren't feasible, operators are prioritizing non-AI workloads for legacy sites and migrating AI workloads to new, purpose-built campuses.

5. Hyperscalers are investing in retrofitting existing sites to get them AI ready
As AI reshapes infrastructure needs, hyperscalers are investing heavily to upgrade legacy data centers rather than replace them. Traditional facilities, designed for lower-density cloud workloads, now require major enhancements to support GPU-intensive AI systems that consume up to ten times more power per rack. Retrofitting efforts focus on integrating liquid- and immersion-cooling technologies, reinforcing structures for heavier AI racks, and expanding power distribution and substation capacity to sustain higher loads.

These projects can come with their challenges. First, they are capital-intensive, typically costing $4 million to $7 million per MW for co-locators and $20 million to $30 million per MW for hyperscalers. Second, a retrofitting project may disrupt operations in an existing facility. And third, these projects are only possible in locations with sufficient available power supply. Nonetheless, retrofitting projects are faster and less risky than building new campuses. By upgrading rather than relocating, hyperscalers preserve access to tier 1 network hubs such as northern Virginia and Santa Clara.

Operators are designating facilities where retrofits are technically unfeasible for non-AI workloads and are concentrating high-density compute in new, AI-optimized campuses. Retrofitting has thus become a catalyst for scaling AI further, allowing hyperscalers to extend the life of assets, expand AI capacity quickly, and strengthen resilience across their most valuable data center locations.

Find more content like this on the
McKinsey Insights App
Scan • Download • Personalize

AI has become the gravitational center of digital infrastructure. The current shift in hyperscalers' strategies is rare and important—it is marking the start of a fundamental change in the market, and stakeholders along the data center value chain must recognize the shift and understand how to adapt their position to capture new opportunities that these shifts are creating. The next stage in this evolution will blur the line between data center and power plant as hyperscalers evolve into utility providers, co-developers, and financiers, redefining the pace and geography of AI growth. These changes will further evolve the market as hyperscalers and the data center ecosystem adapt to growing compute demand.

Chhavi Arora is a partner in McKinsey's Seattle office; **Marc Sorel** is a partner in the Boston office, where **Arjita Bhan** is a senior knowledge expert and **Jess He** and **Nicholas Shaw** are consultants; **Pankaj Sachdeva** is a senior partner in the Philadelphia office; **Riya Garg** is a consultant in the Detroit office; and **Shriya Ravishankar** is a consultant in the Bay Area office.

Copyright © 2025 McKinsey & Company. All rights reserved.