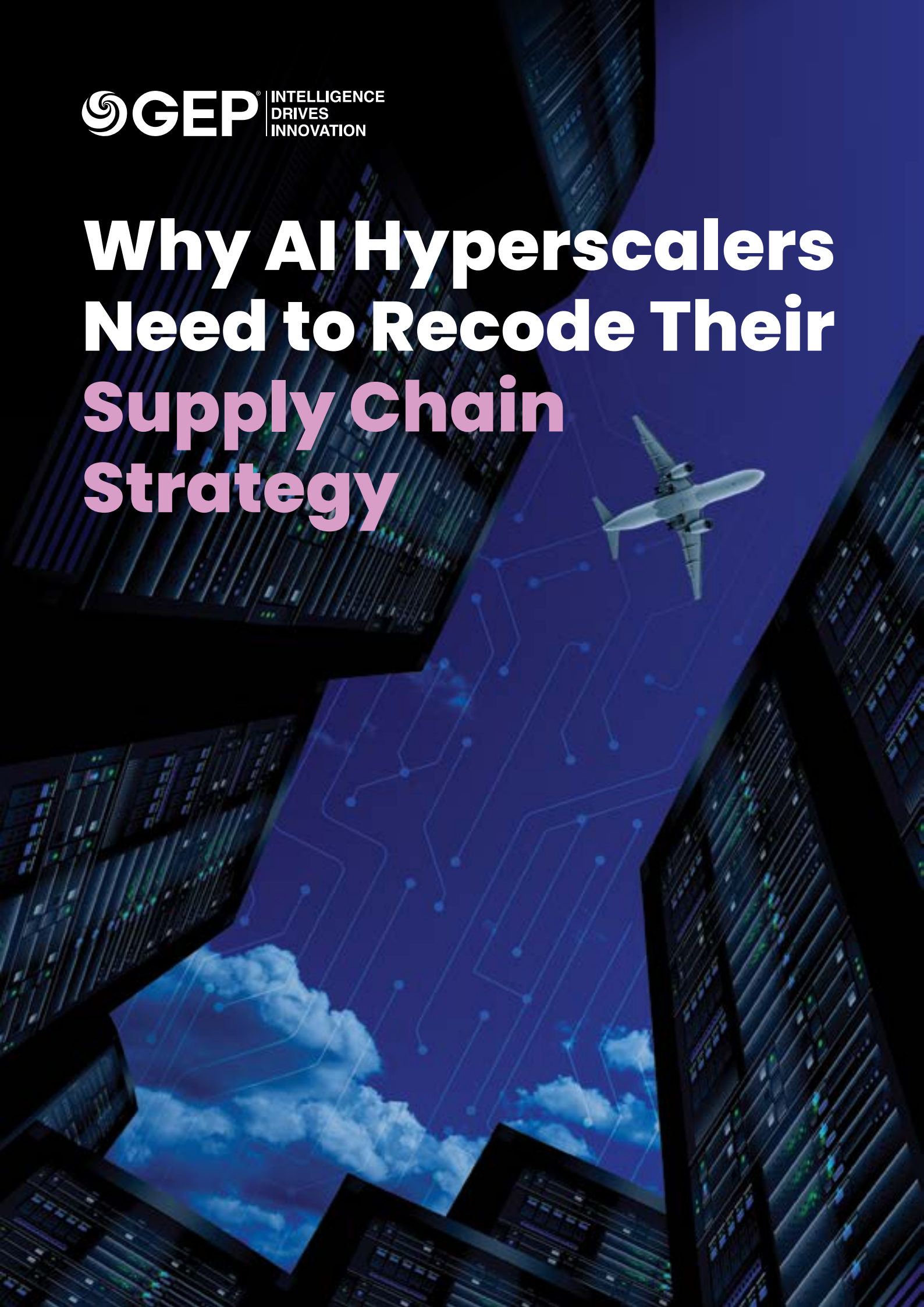
GEP INTELLIGENCE DRIVES INNOVATION

Why AI Hyperscalers Need to Recode Their Supply Chain Strategy



The AI industry has spent the past few years engaged in a high-stakes infrastructure arms race—one defined by billion-dollar bets, rapid technological leaps, and an unrelenting demand for computational power. The companies that can build the fastest, most powerful AI systems have long believed they will emerge as the industry's dominant players. And until recently, that belief seemed indisputable.

Venture capital poured in. Tech giants doubled down. Governments got involved. The $500 billion Stargate Project in the US, an unprecedented investment in AI infrastructure, made it clear that the race to scale AI isn't just about private competition—it was about national strategy. The fundamental equation seemed simple:

Maximize raw performance—the sheer computational muscle needed to train ever-larger AI models.

Maximize deployment speed—the ability to scale new models at an industrial pace.

To sustain this, an immense, highly specialized, and rigid physical supply chain emerged—what insiders now call the "**Supply Chain of AI.**" This infrastructure isn't just about silicon chips or cloud storage; it's a sprawling ecosystem of data centers, power generation, fiber optic networks, and high-performance cooling systems, all working in tandem to sustain AI's insatiable appetite for compute power.

But as 2025 unfolds, the assumptions underpinning this strategy are beginning to crack. Two new imperatives are reshaping the industry's approach to AI infrastructure:

Flexibility—the ability to pivot instantly in response to new breakthroughs, regulatory shifts, or unexpected constraints.

Cost efficiency—ensuring that AI inference (the execution of AI models at scale) can remain economically viable, rather than a black hole of operational costs.

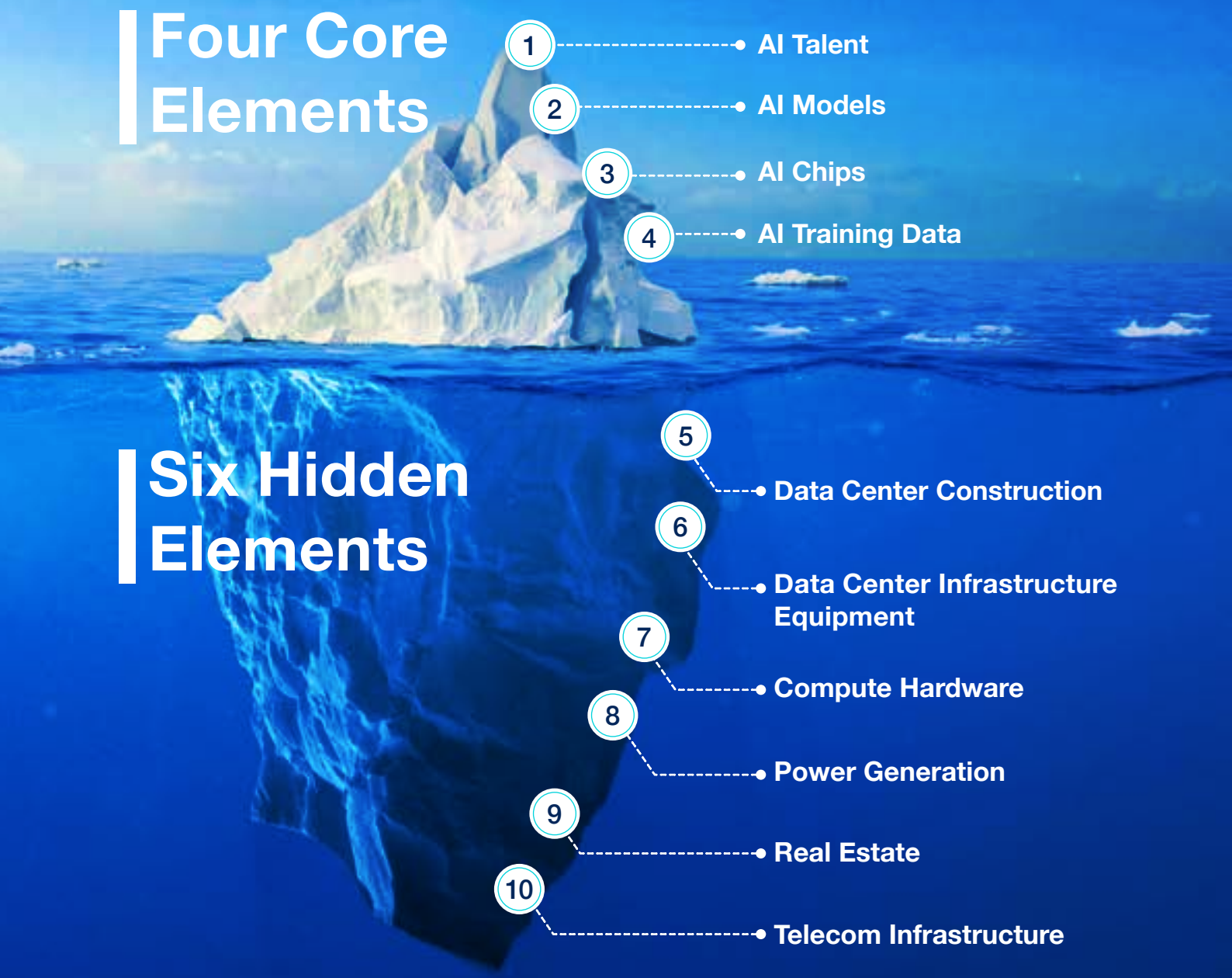
The problem? The infrastructure built to maximize performance and speed was never designed to be flexible or cost-efficient. Vertical integration—where a company owns and controls everything from chips to data centers—has long been seen as the optimal approach. But now, as the financial realities of AI inference come into focus, the industry is being forced to rethink whether a more externalized, modular supply chain might offer a better path forward.

Achieving that balance isn't a simple choice between owning infrastructure or outsourcing it. The trade-offs will vary across different components of the supply chain, from compute hardware to energy sourcing. A rigid, one-size-fits-all approach is no longer feasible—what's needed instead is a portfolio strategy, one that optimizes for speed and performance while maintaining enough flexibility and cost control to sustain AI's long-term growth.

The Supply Chain of AI is often framed around four core elements—Talent, Models, Data, and Chips—the battlegrounds where companies compete for dominance. But beneath these high-profile components lies another set of six hidden elements that are just as critical and, in many cases, the biggest bottlenecks slowing AI deployment.

For a more detailed look into these elements and the roadblocks shaping AI's future, see our full analysis here:

These infrastructure layers—data center construction, power generation, compute hardware, infrastructure equipment, real estate, and telecom networks—are facing mounting challenges. U.S. data centers are projected to consume up to 9.1% of the nation's electricity by 2030,[1] Microsoft is reportedly reassessing its data center plans,[2] and grid connection times for new data centers in Northern Virginia now exceed seven years.[3]



Evolving Goalposts of AI Deployment

For the past five years, the race to deploy artificial intelligence has been driven by two fundamental metrics:

Raw performance—Who has the most advanced models or chips?

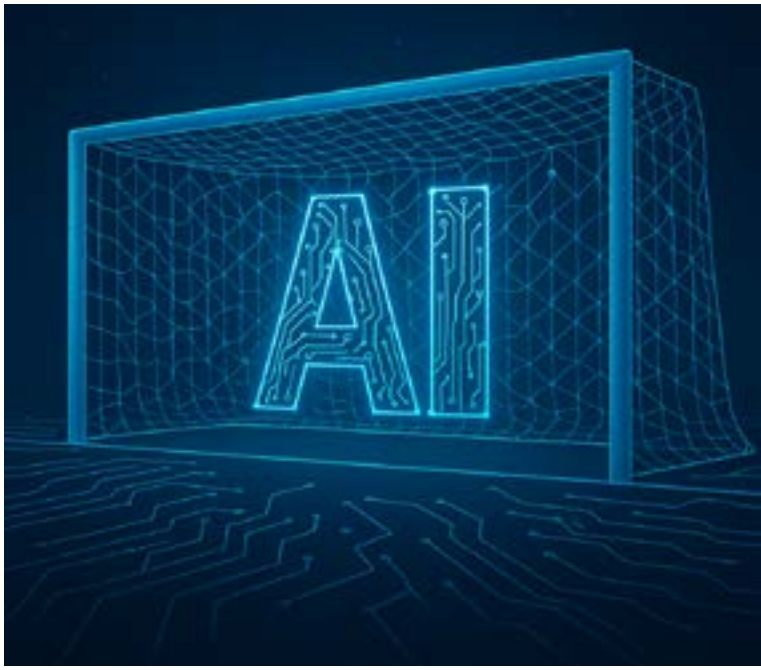
Deployment speed—How fast can these technologies be integrated into massive data centers?

This relentless focus on power and speed fueled a wave of vertical integration, with companies developing proprietary chips, stockpiling electrical infrastructure equipment, and locking in long-term control over their AI supply chains. The assumption was clear: owning everything from silicon to server farms would ensure dominance.

But as 2025 unfolds, the equation is changing. The explosive growth of AI applications, rising operational costs, and mounting revenue pressures are forcing hyperscalers to rethink their priorities.

The Inference Bottleneck: ChatGPT may soon surge past 1 billion active users, driving demand for AI inference—processing real-time user requests—far beyond initial projections.[4] Unlike training, which is a one-time CapEx investment, inference is an ongoing OpEx, meaning it must eventually operate at a cost sustainable within the bounds of revenue.

The Revenue Squeeze: The launch of DeepSeek R1 has shattered existing AI pricing models. Chinese AI giants Alibaba, Baidu, Tencent, and ByteDance were forced to slash subscription fees by up to 97% to remain competitive, eroding profit margins overnight.[5] In the U.S., this pressure is manifesting differently: Big Tech is rolling out increasingly resource-intensive AI features for free, betting on future monetization strategies that remain uncertain.



The Cost Explosion: AI compute costs are spiraling out of control, and more waves of change are coming. In particular, agentic consumption—where AI agents autonomously generate more AI workloads, creating an exponential growth in resource demand, is very unpredictable. Microsoft, OpenAI, and Google are already struggling with forecasting these infrastructure costs, as each successive AI model iteration demands exponentially more compute power.[6].

In response to these pressures, two new critical metrics are emerging for AI deployment:

Flexibility—The ability to pivot infrastructure strategies rapidly in response to breakthroughs, market shifts, or economic constraints.

Cost Efficiency—Ensuring AI operations generate sustainable revenue rather than consuming resources at unsustainable rates.

These shifts are forcing hyperscalers to make new trade-offs, weighing long-term infrastructure investments against the increasing need for agility. The dominance of vertically integrated AI stacks is being challenged by new economic realities, reshaping the way AI is built, deployed, and sustained.

GEP INTELLIGENCE DRIVES INNOVATION

Evolution of critical metrics in the Supply Chain of AI

Leading Thoughts

Critical Metrics

2022

"AI compute needs custom hardware and chips"

"The AI race is winner-take-all and will go to a vertically integrated player"

"Core elements (Talent, Models, Data, and Chips) FTW!

Raw Performance

2023

"AGI is the finish line; whoever crosses it first wins it all!"

"Build'em, and they will come"

"Quick buildouts needed to train the next big model"

Deployment Speed

2024

"Rate of societal change, regulatory frameworks, and the hidden supply chain elements are all unpredictable"

"Supply-side investment needs to follow the rate of demand growth"

"The 'Agents spawning agents' scenario throws compute infrastructure forecasts for a toss"

"Governments investing in AI creates a new dimension of uncertainty"

"Will corporations lease AI data centers from the big tech hyperscalers, or build their own?

"Quantum computing can overturn current forecasts again in a few years"

Agility

2025

"DeepSeek R1 ground-breaking not in performance, but in 'performance per token' "

"If fast-follower models can get close in performance, where is the ROI on $100B+ investments?"

"Buildouts need to focus on inference, which has to run on very low (or zero) token prices"

Cost Efficiency

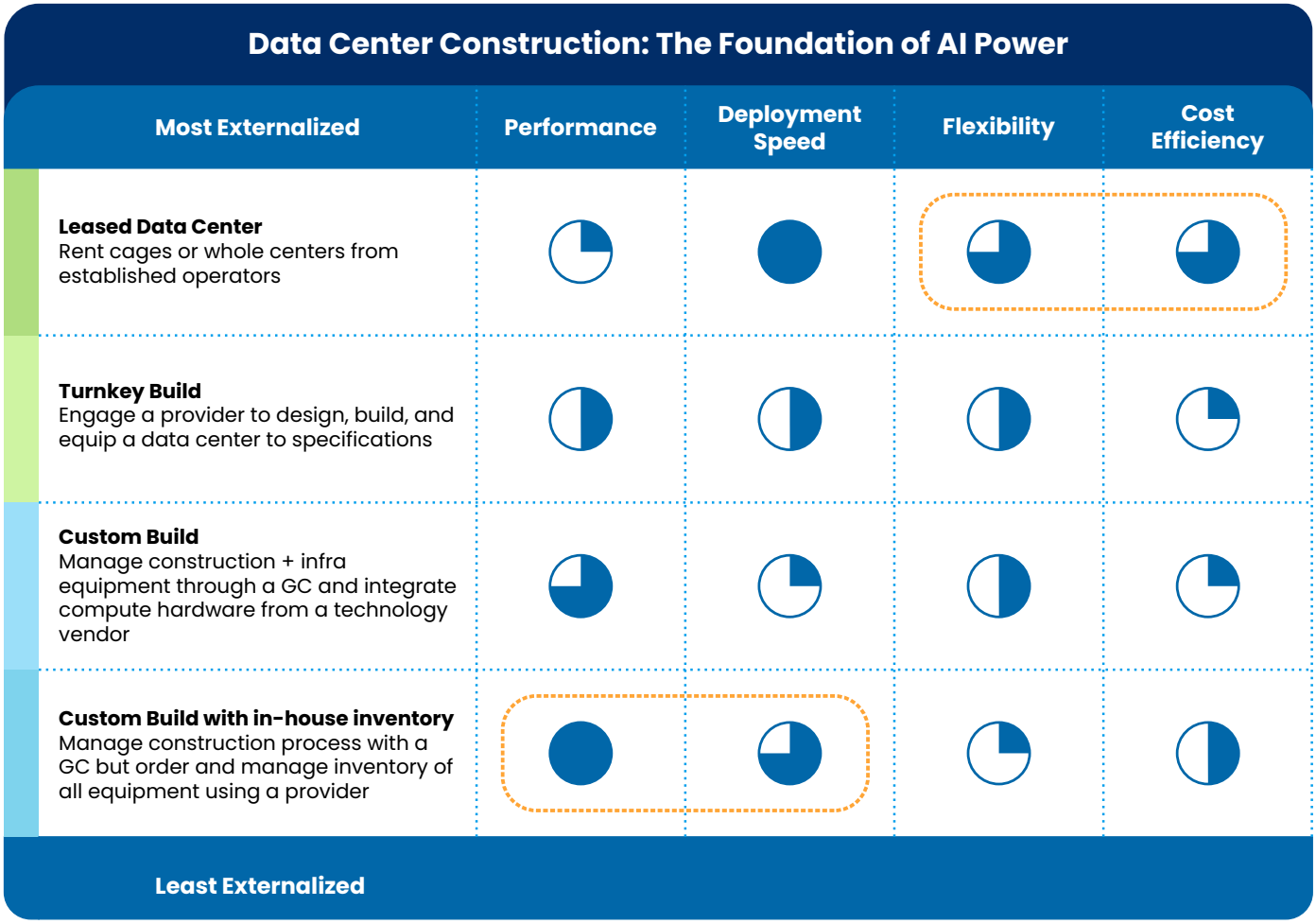
The industry's priorities are shifting from absolute performance to economic and operational sustainability. The initial race for custom hardware and massive training capacity has given way to concerns about long-term viability, unpredictable demand surges, and the fundamental economics of inference. The question is no longer just "How fast and powerful can AI be built?", but rather "How adaptable and cost-effective can it remain?"

Critical Metric	Key Considerations for Externalization
Raw Performance	• Viability of in-house AI chip design as companies like Meta and OpenAI explore proprietary hardware[7,8] • Strategic partnerships with chipmakers to balance cost, supply chain risks, and performance
Deployment Speed	• Bottlenecks in long lead-time supplies (e.g., switchgear, generators) delaying AI infrastructure rollouts • Leasing generic data center compute to bypass hardware supply chain delays[9]
Flexibility	• Build vs. lease decisions for inference loads, as hyperscalers weigh capital expenditures against agile infrastructure[10] • Achieving positive ROI despite short amortization periods, with modular designs becoming a key strategy
Cost Efficiency	• Sourcing goods and services at speed and scale, with AI chip shortages impacting cost structures • Extending the life of new builds amid rapid technological change through scalable, adaptable infrastructure

Across all four metrics, strategic outsourcing (externalization) is shaping AI's next phase of growth. Companies must now reassess how much infrastructure they control directly versus what is sourced from partners, as the trade-offs between cost, speed, and flexibility become increasingly complex.

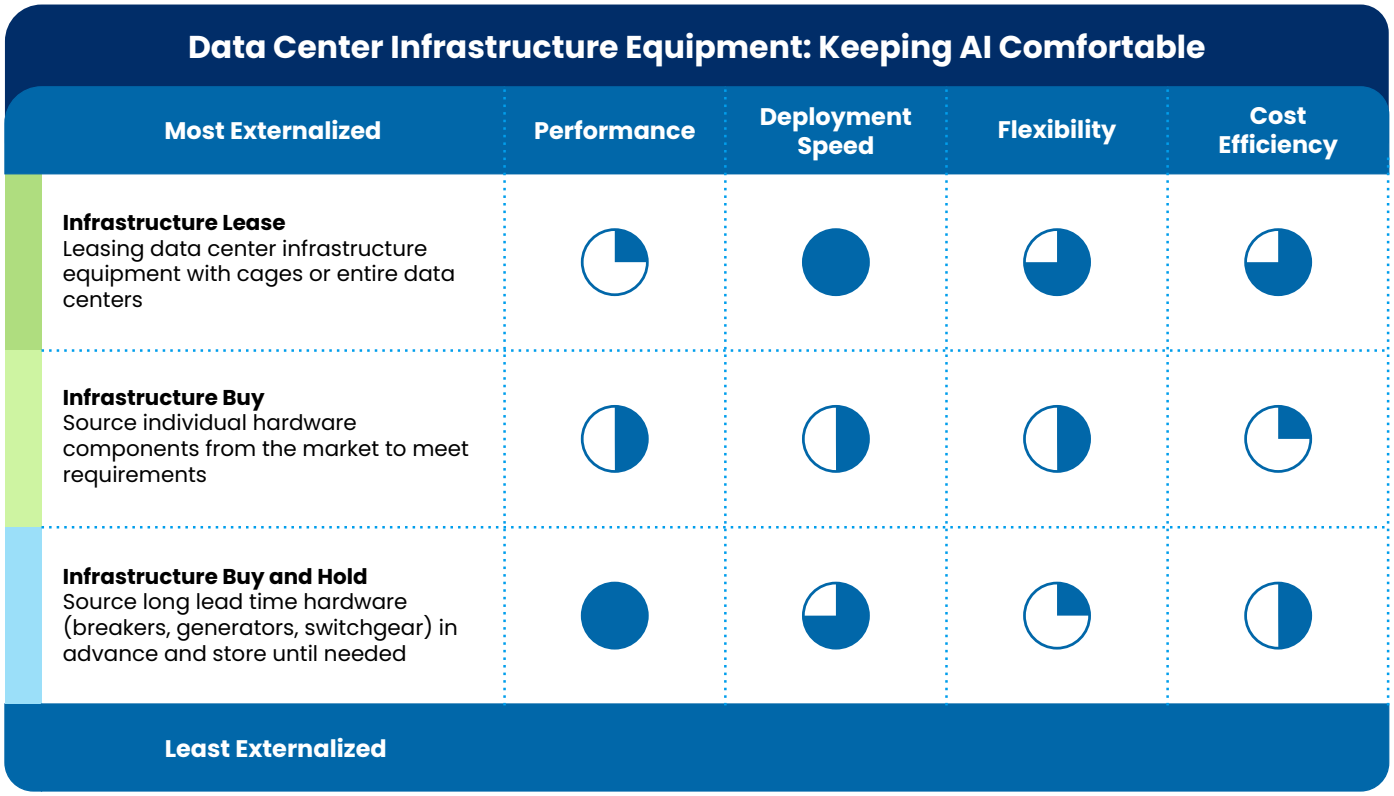
Strategic Outsourcing Options for the Hidden Elements of AI's Supply Chain

As seen in the preceding sections, strategic outsourcing (externalization) is emerging as a key approach to balancing raw performance, deployment speed, flexibility, and cost efficiency. Each of the six hidden elements of AI's supply chain presents multiple options that can reduce capital outlays while increasing adaptability.



For years, technology firms have prioritized custom-built data centers with in-house inventories to maximize performance and deployment speed. Owning infrastructure ensures tailored optimization and direct control, aligning with industry demands for raw performance and rapid deployment.

However, as flexibility and cost efficiency become equally critical, leasing data center facilities is gaining traction. Outsourcing data center design and construction allows companies to scale rapidly without the capital burden of full ownership while tapping into best-in-class expertise. Microsoft's recent reevaluation of its leasing strategy—canceling planned AI data center deals to avoid potential oversupply—reflects this shift toward a more adaptive approach.[11]



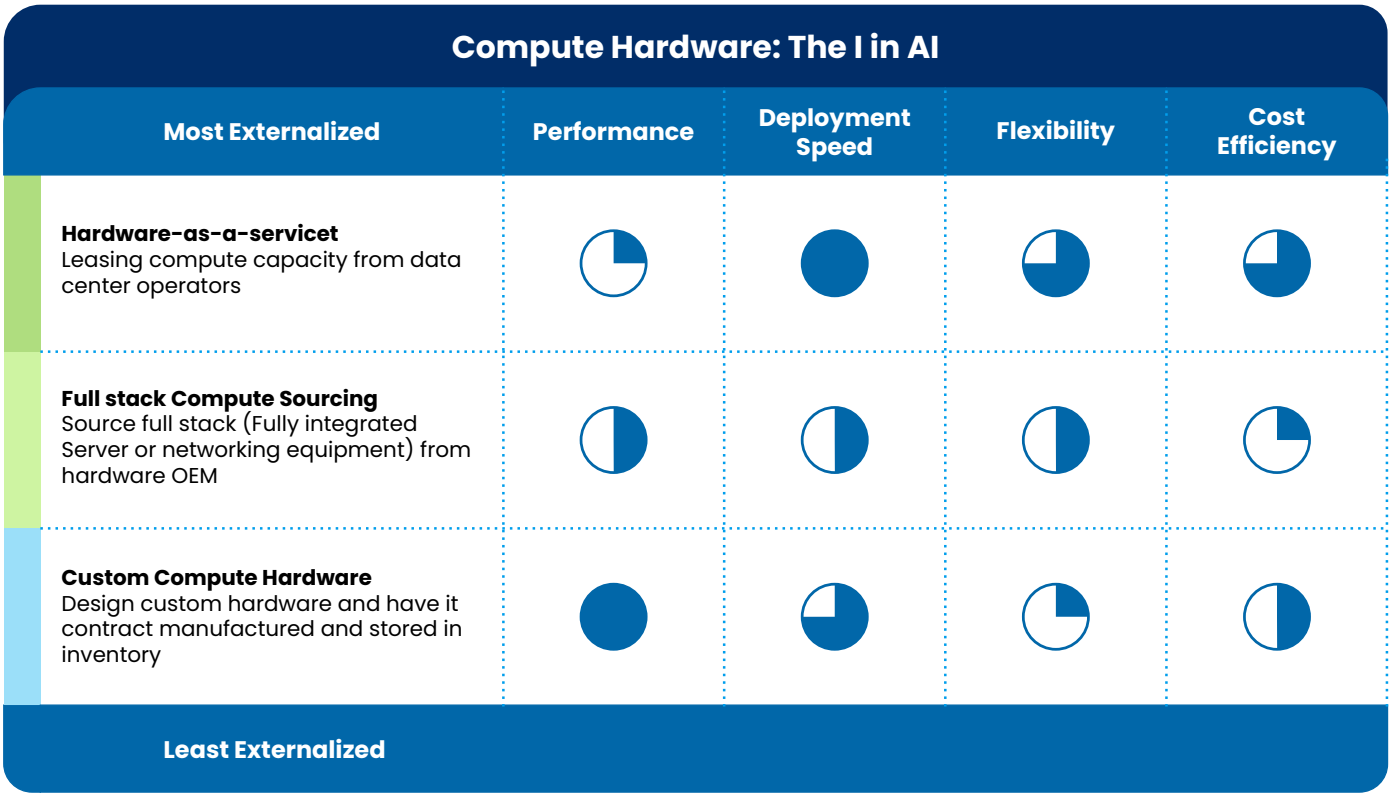
The Trade-Off Between Standardization and Customization in AI Infrastructure

Cooling systems, networking equipment, backup power generators, and racking infrastructure form the quiet backbone of AI's expansion—ubiquitous, often overlooked, yet indispensable. These systems are designed for industrial-scale operations, built by a competitive field of suppliers who specialize in high-reliability, standardized designs that benefit from economies of scale.

But AI is not just another enterprise workload. The sheer intensity of compute, the density of racks, and the need for near-zero downtime push these systems beyond their original design parameters. Standard equipment can do the job, but customization offers an edge—increasing performance, extending uptime, and reducing failure rates under extreme AI loads. And therein lies the dilemma: embrace the efficiency of mass-produced systems or squeeze out the last bit of performance with tailor-made solutions?

Customization comes at a cost—not just in price, but in supply chain fragility. Cooling systems, generators, and networking gear already face long lead times, and the more specialized the configuration, the harder it is to source replacements quickly.

Reports indicate that lead times for critical infrastructure components such as switchgear, generators, and transformers now average around 11 months, a significant factor in AI deployment timelines.[12] Some firms have taken a different approach altogether: buying and holding inventory, securing stockpiles of mission-critical equipment to avoid disruptions.



In the race to scale AI, companies are increasingly outsourcing compute capacity to external cloud

and colocation providers. The rationale is clear: AI hardware is expensive, supply chains are tight, and capital expenditures for in-house data centers can be staggering. Leasing compute from third parties sidesteps long hardware procurement timelines and keeps operations more flexible, allowing companies to scale workloads without committing to multi-billion-dollar infrastructure projects.

It's a model that's already reshaping the industry. OpenAI's recent $12 billion deal with CoreWeave—diversifying its compute power beyond Microsoft—illustrates just how valuable third-party GPU infrastructure has become.13 The appeal is simple: fast access to high-performance compute without the supply chain bottlenecks of custom in-house systems.

But there's a cost to this flexibility. Outsourcing compute means relying on standardized, commodity hardware, where custom AI chips and specialized configurations are off the table. In-house data centers are designed to push performance to the limit, integrating custom accelerators, optimized networking, and hardware fine-tuned for specific AI workloads. Colocation providers, by contrast, offer generic infrastructure, stripping away many of these performance advantages.

The trade-off is unavoidable: lower capital outlay, greater scalability, but at the expense of fine-tuned AI efficiency. As AI's hardware arms race accelerates, companies must decide—is it better to own and optimize, or lease and adapt?

Power Generation: Sustaining the AI Engine

Most Externalized	Performance	Deployment Speed	Flexibility	Cost Efficiency
Retail Power Purchase Relying on existing grid infrastructure for power supply				
PPA/VPPA((Virtual) Power Purchase Agreements) Contracting with (renewable) energy providers for long-term power supply.				
Self-Generation Building on-site power generation facilities such as rooftop solar.				
Co-investing with utility providers to build power stations (e.g., Brookfield Asset Mgmt., Intersect Power) Partnering with utilities to develop new power infrastructure.				
Investing in new technologies for clean power (e.g., BlackRock GIP, Kairos Power, 3 Mile Island) Supporting the development of innovative and sustainable energy solutions.				
Least Externalized				

As AI infrastructure scales at an unprecedented pace, renewable energy partnerships and power purchase agreements (PPAs) have become critical tools for stabilizing electricity costs and advancing sustainability goals. Google's $20 billion investment in renewable-powered data centers[14] and Microsoft's push for on-site nuclear reactors reflect the urgency with which companies are securing long-term energy sources.

Yet, the timeline for clean energy deployment lags behind AI's rapid expansion. While companies sign long-term PPAs, the reality is that new energy infrastructure can take years—sometimes decades—to develop. AI demand, on the other hand, is accelerating at a rate that grid operators struggle to forecast. Texas grid operators, for example, are already facing concerns over the surge in AI-driven power consumption[15]

Another challenge is redundancy. AI data centers have grown so large that traditional backup generators are no longer a sufficient safeguard. Facilities now require fully redundant power sources, placing even greater strain on existing grids.

Real Estate: The Actual Foundation

Most Externalized	Performance	Deployment Speed	Flexibility	Cost Efficiency
Lease real estate along with the data center from Providers				
Buy the real estate for the core data center, but lease the real estate for all the ancillary setups such as inventory storage				
Buy and build on owned real estate, with external scouting and Market research				
Buy and build on owned real estate, with internal scouting and Market research				
Least Externalized				

For years, tech firms followed a simple rule: build and own your data centers, control your infrastructure, and optimize for performance. But as AI reshapes the economics of compute, that assumption is now up for debate. The question isn't just about capacity or cost—it's about whether tech companies should be in the real estate business at all.

Leasing AI-specific infrastructure offers a compelling alternative. It eliminates massive capital expenditures, allows for geographic flexibility, and scales up or down based on demand. Unlike traditional enterprise workloads, AI models consume power at unprecedented levels, requiring sites that can handle extreme energy loads, redundancy, and cooling. These variables make data center real estate decisions more strategic than ever.

While physical location doesn't directly affect AI performance, it determines everything else that does:

Connectivity: 70% of the world's internet traffic flows through Northern Virginia, making it an unrivaled hub for low-latency data transfers.[16]

T**ax Incentives:** Virginia's $135.9 million in annual tax breaks has more than doubled since 2017, making it an economic hotspot for data center expansion.[17]

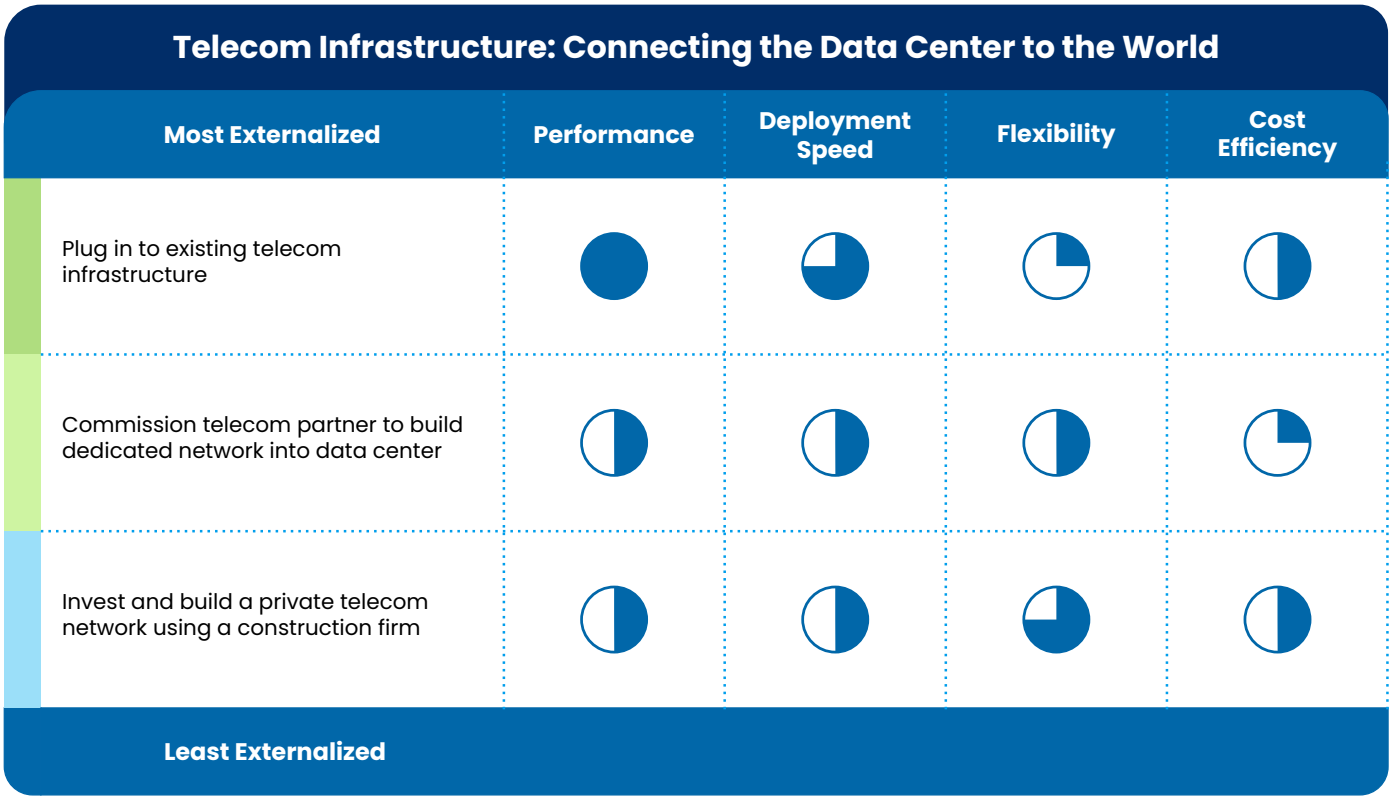
Energy Costs: With the 5th cheapest commercial electricity rate in the U.S., Virginia remains attractive to AI hyperscalers looking to minimize operational costs.

Skilled Workforce: A specialized talent pool in data center construction and maintenance reduces deployment risks and delays.

Regulatory Environment: Virginia's data center-friendly policies give operators confidence that legislative changes won't derail investments.

The challenge, then, is not just about whether to build or lease—it's about who should be making these decisions in the first place.

Given the complexity of site selection, energy procurement, and local policy navigation, leaving real estate market research to specialized external firms is not just an efficiency play—it's a necessity. Owning infrastructure doesn't guarantee an advantage in AI, but choosing the wrong location guarantees a disadvantage.

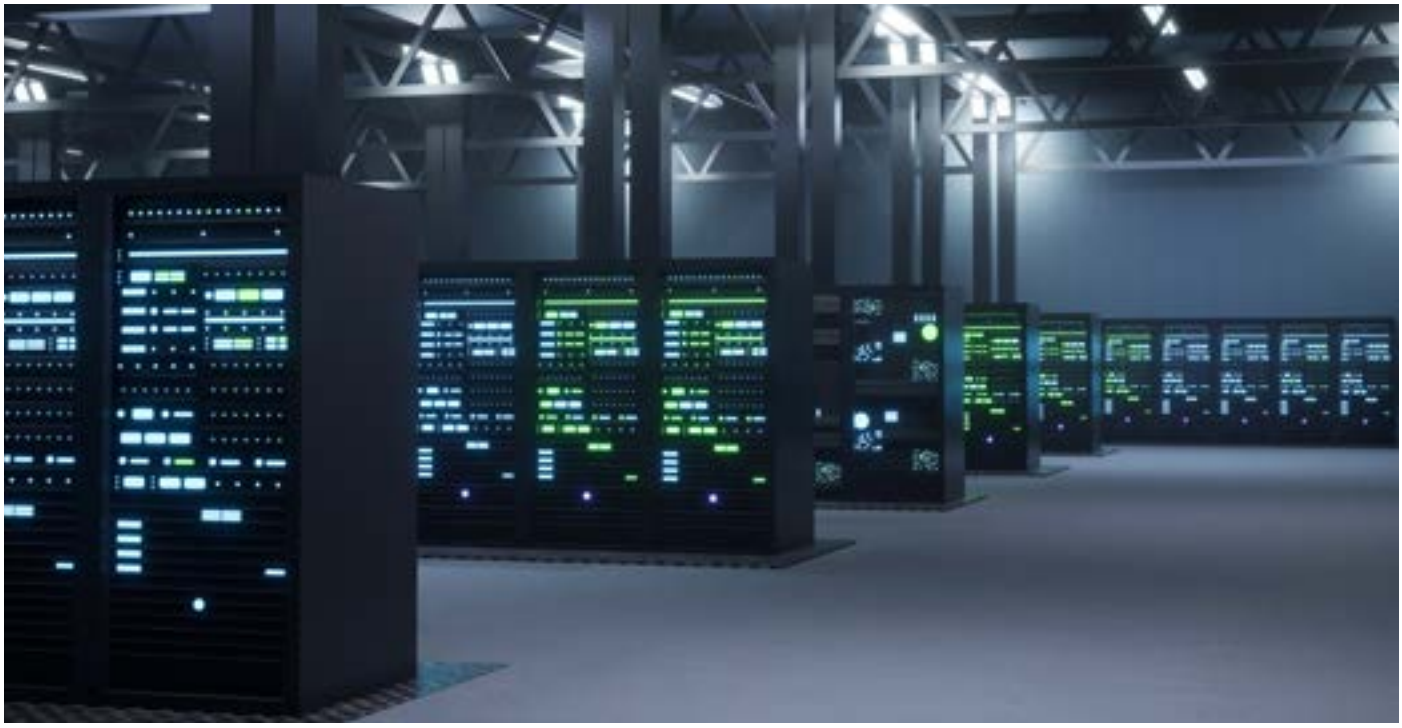


AI is no longer confined to single data centers. The next generation of AI workloads is distributed—trained in one location, processed in another, and executed across a global network. That shift has made high-bandwidth, low-latency data transmission more critical than ever, forcing companies to rethink whether telecom infrastructure should be an asset they own—or a service they lease.

For most, outsourcing telecom is now the only viable option. Laying private fiber networks, securing spectrum rights, and building redundancy for AI-scale traffic requires investments that few companies can justify. The alternative? Leasing from hyperscale telecom providers, who already control the global infrastructure AI depends on.

The industry is moving fast. Google, Microsoft, and Amazon are racing to expand their private fiber networks to support AI workloads. Meanwhile, telecom giants like Lumen and AT&T are rolling out high-speed, AI-optimized networking services, offering direct cloud interconnects with ultra-low latency.

The trade-off is clear: Outsourcing telecom removes the burden of infrastructure ownership, but it also means AI companies are increasingly reliant on a few powerful telecom players. As AI models scale beyond single-location deployments, control over bandwidth and routing will become just as critical as access to compute itself.



AI Infrastructure Can't Be One-Size-Fits-All

The AI race was once about speed and raw performance—who could build the biggest, most powerful data centers. But priorities have shifted. Flexibility and cost efficiency now matter just as much. With inference costs rising and AI workloads becoming more distributed, the challenge is no longer just about building faster, but about building smarter.

No Single Infrastructure Model Solves Everything

Every approach comes with trade-offs:

- Custom-built AI data centers maximize performance and efficiency but lock companies into high-cost, long-term infrastructure that can't easily adapt to shifting demand.
- Leased data centers offer agility and scalability, allowing companies to scale compute capacity up or down without long-term commitments, but limit customization for high-intensity AI workloads.
- Energy sourcing is just as fragmented—some data centers will need on-premise power generation, while others will rely on the wholesale grid, exposing them to potential cost spikes and supply constraints.

At the same time, the pace of AI development is relentless. What works today may be inefficient—or obsolete—tomorrow. Infrastructure decisions can't be static; they need to adjust as AI compute demands, energy costs, and regulatory landscapes evolve.

Using A Portfolio Approach to Optimize the Data Center Fleet

Rather than choosing between custom-built vs. externalized data centers, hyperscalers need to evaluate their data center fleet as a portfolio. Just like diversifying a financial portfolio reduces risk, a data center fleet blending in-house and third-party solutions will remain adaptable to shifting AI demands:

- **Training workloads for next-gen models will continue to demand dedicated, AI-optimized facilities** designed for maximum performance and low-latency processing
- **Inference workloads will run more efficiently in a mix of owned and leased standard data centers,** where capacity can be expanded or reduced based on demand
- **The energy mix will need to combine captive and retail sources,** ensuring resilience against market volatility and regulatory shifts

There is no single best way to scale AI infrastructure. **The companies that stay adaptable—blending ownership with external partnerships—will be the ones best positioned to scale AI efficiently without being trapped by rigid infrastructure models.**

1. "Analyzing Artificial Intelligence and Data Center Energy Consumption," EPRI, 28 May 2024 | https://www.epri.com/research/ products/000000003002020502

2. Britney Nguyen, "Microsoft is pulling back on data centers as it reassesses its AI plans," Quartz, April 3, 2025 | https://qz.com/microsoft-pull- back-data-centers-ai-demand-spending-1851774525

3. International Energy Agency, "Energy and AI," IEA, March 2024 | https://www.iea.org/reports/energy-and-ai

4. Aaron Mok, "Sam Altman Says ChatGPT Might Replace Google Search — and Apple Should Be Nervous," Business Insider, March 2025 | https://www.businessinsider.com/chatgpt-sam-altman-openai-apple-google-search-2025-3

5. Meghan Bobrowsky, "U.S. Businesses Already Love DeepSeek," The Wall Street Journal, March 2025 | https://www.wsj.com/articles/u-s- businesses-already-love-deepseek-1679cde2

6. Camilla Hodgson, "AI Start-Ups Flock to CoreWeave as It Becomes an 'Nvidia-Grade Cloud,'" Financial Times, March 27, 2024 | https:// www.ft.com/content/d5c638ad-8d34-4884-a08c-a551588a9a28

7. Stephen Nellis, "Meta Begins Testing Its First In-House AI Training Chip," Reuters, March 11, 2025 | https://www.reuters.com/technology/ artificial-intelligence/meta-begins-testing-its-first-in-house-ai-training-chip-2025-03-11/

8. Brett Young, "OpenAI Scales Back Foundry Plans, Focuses on In-House AI Chip Development with Broadcom and TSMC," Weights & Biases, March 2025 | https://wandb.ai/byyoung3/ml-news/reports/OpenAI-Scales-Back-Foundry-Plans-Focuses-on-In-House-AI-Chip-Development-with- Broadcom-and-TSMC--VmlIdzo5OTQ2ODk0

9. CoreSite, "Data Centers: Build vs. Lease," CoreSite, 2021 | https://www.coresite.com/blog/data-centers-build-vs-lease

10. DataGarda, "Building vs. Leasing: Strategic Decisions for Data Center Investments," DataGarda, January 24, 2025 | https://datagarda.com/ building-vs-leasing-strategic-decisions-for-data-center-investments/

11. Stephen Nellis, "Microsoft Shelves AI Data Center Deals in Sign of Potential Oversupply, Analyst Says," Reuters, February 24, 2025 | https:// www.reuters.com/technology/microsoft-shelves-ai-data-center-deals-sign-potential-oversupply-analyst-says-2025-02-24/

12. Inside AI News, "Data Center Report: Record-Low Vacancy Pushing Hyperscalers into Untapped Markets," Inside AI News, March 10, 2025 | https://insideainews.com/2025/03/10/data-center-report-record-low-vacancy-pushing-hyperscalers-into-untapped-markets

13. Camilla Hodgson, "The AI Industry's Growth Is Being Slowed by a Surprising Bottleneck: Switchgear," Financial Times, March 13, 2024 | https:// www.ft.com/content/4b52fdbb-ca8e-4208-bb99-f1e7f9313863

14. Reuters, "AI Boom Spurs Big Tech to Build Clean Power on Site," Reuters, February 5, 2025 | https://www.reuters.com/business/energy/ai- boom-spurs-big-tech-build-clean-power-site-2025-02-05/

15. Wikipedia, "Dulles Technology Corridor," Wikipedia, 2024 | https://en.wikipedia.org/wiki/Dulles_Technology_Corridor

16. Dan Swinhoe, "Virginia Granted Data Centers $135.9M in Tax Breaks Last Year," Data Center Dynamics, January 4, 2023 | https:// www.datacenterdynamics.com/en/news/virginia-granted-data-centers-1359m-in-tax-breaks-last-year/"

GEP® delivers AI-powered procurement and supply chain solutions that help global enterprises become more agile and resilient, operate more efficiently and effectively, gain competitive advantage, boost profitability and increase shareholder value.

Fresh thinking, innovative products, unrivaled domain expertise, smart, passionate people — this is how GEP SOFTWARE™, GEP STRATEGY™ and GEP MANAGED SERVICES™ together deliver procurement and supply chain solutions of unprecedented scale, power and effectiveness. Our customers are the world's best companies, including more than 1000 Fortune 500 and Global 2000 industry leaders who rely on GEP to meet ambitious strategic, financial and operational goals.

A leader in multiple Gartner Magic Quadrants, GEP's cloud-native software and digital business platforms consistently win awards and recognition from industry analysts, research firms and media outlets, including Gartner, Forrester, IDC, ISG, and Spend Matters.

GEP is also regularly ranked a top procurement and supply chain consulting and strategy firm, and a leading managed services provider by ALM, Everest Group, NelsonHall, IDC, ISG and HFS, among others. Headquartered in Clark, New Jersey, GEP has offices and operations centers across Europe, Asia, Africa and the Americas. To learn more, visit www.gep.com.

Authors

Pranav Padgaonkar

Alakh Krishnani

Connect With GEP

GEP SMART

GEP SMART is an AI-powered, cloud-native software for direct and indirect procurement that offers comprehensive source-to-pay functionality in one user-friendly platform, inclusive of spend analysis, sourcing, contract management, supplier management, procure-to-pay, savings project management and savings tracking, invoicing and other related functionalities.

GEP NEXXE

GEP NEXXE is a unified and comprehensive supply chain platform that provides end-to-end planning, visibility, execution and collaboration capabilities for today's complex, global supply chains.

Built on a foundation of big data, artificial intelligence and machine learning, GEP NEXXE is next-generation software that helps enterprises make supply chain a competitive advantage.

100 Walnut Avenue, Clark, NJ 07066 I P 732.382.6565 I info@gep.com I www.gep.com

Clark, NJ I Austin I Chicago I Atlanta I Toronto I Mexico City I San Jose I São Paulo I Dublin I London I Amsterdam I Dortmund, Germany I Frankfurt I Oslo I Prague I Stockholm I Toruń, Poland I Cluj-Napoca, Romania I Tampere, Finland I Helsinki/Espoo I Riga, Latvia I Pretoria I Abu Dhabi I Mumbai I Coimbatore I Hyderabad I Kuala Lumpur I Singapore I Surabaya, Indonesia I Shanghai I Dalian I Tokyo I Sydney


GEP
INTELLIGENCE
DRIVES
INNOVATION