Data center operators enter their industrial engineering era

Companies that can master bottlenecks by detecting, resolving, and anticipating constraints from site selection to refresh will gain competitive advantage.

kpmg.com

Executive summary

AI supremacy has shifted from a software sprint to an industrial-scale race defined by infrastructure, resilience, and governance.

Leaders are deploying industrial-grade operating models, using cross-functional control towers, and integrating capital and energy planning strategies by design.

Winning strategies expand the economic lens beyond cost, factoring in return on resilience, speed to market, and the value of optionality.

AI has fundamentally changed the scale and complexity of digital infrastructure build-outs. For years, leaders treated AI as a software puzzle to be solved with better models, more data, and faster iteration. That still matters, but the race is no longer won in code. Instead, tomorrow's winners will be defined by their ability to build and operate industrial-scale systems for intelligence.

Today, AI is a capital-intensive, infrastructure-dependent industrial system. Not only is this build-out fueling a historic surge in capital expenditure,[1] but also domestic AI infrastructure spending was estimated to account for roughly 1.1 percent to 1.2 percent of GDP growth in the first half of 2025.[2]

The organizations pulling ahead are not simply algorithm innovators; they are industrial strategists. They understand that AI performance is shaped by more than the language models they buy, build, or borrow. Instead, beyond the technology itself are energy access, facility design, hardware availability, permitting timelines, and operating discipline, all of which influence capacity delivery and the ability to conduct research.

That shift changes the definition of leadership. In this new era, the limiting factor is rarely just compute demand or capital. It is the ability to design, build, and operate hyperscale ecosystems that can balance extreme density, continuous uptime, and accelerated growth with environmental scrutiny, community expectations, and geopolitical uncertainty.

These constraints are dynamic. Yesterday's bottleneck may have been GPU availability. Today's may be substation capacity or local permitting. Tomorrow's may be attrition among power engineers or site reliability teams. AI infrastructure leadership is therefore becoming a continuous cycle of detection, prioritization, and resolution. The strategic challenge is not merely to solve each constraint in isolation, but to orchestrate the system so that throughput, resilience, and growth remain intact as bottlenecks move.

The companies that adopt an agile industrial model will redefine competitive advantage. Those that do not will find that their AI ambitions are constrained not by algorithms, but by electricity, cooling, talent, regulation, and supply-chain reality.

[1] Matt Day and Annie Bang, "Big Tech to Spend $650 Billion This Year as AI Race Intensifies," Bloomberg.com, February 5, 2026.

[2] "Competition in artificial intelligence infrastructure," OECD, November 14, 2025.

The license to operate is a design requirement

The ability to scale AI infrastructure increasingly depends on earning and maintaining a license to operate. Carbon intensity, water and land use, renewable sourcing, and community impact now influence approval timelines, financing conditions, and brand trust. That means sustainability and compliance cannot be treated as downstream reporting exercises. They must be embedded into the design of the infrastructure itself, from siting and facility configuration to operating models and supplier choices.

> "Sustainability and compliance cannot be treated as downstream reporting exercises. They must be embedded into the design of the infrastructure itself, from siting and facility configuration to operating models and supplier choices."

Organizations that move fastest include regulatory requirements, sustainability, and stakeholder engagement as core design parameters. Doing so shortens approval cycles, reduces execution risk, and helps preserve strategic flexibility as scrutiny intensifies.

Strengthening delivery discipline across a global data center portfolio

A hyperscaler operates a multibillion-dollar global portfolio of data center capital projects across diverse regulations, vendors, delivery models, and maturities. Rapid scaling strained consistent governance, policy compliance, and execution discipline while meeting aggressive timelines.

The company launched a large-scale audit program to improve control, transparency, and performance across projects, addressing contract compliance, insurance, scheduling, payment applications, and change management, with deep dives on high-risk areas in collaboration with leadership.

The program identified more than $25 million in potential recoverable costs, including over $5 million in noncompliant spend, and supported the recovery of more than $5 million to date. Audits surfaced nearly 300 findings on process gaps and execution inconsistencies, creating a foundation for continuous improvement.

With robust testing, communication, and issue tracking, the audit program strengthened scalable capital delivery across regions, improving governance, enhancing financial discipline, and boosting confidence in executing complex, schedule-critical infrastructure at global scale.

AI leaders use agile industrial models

As AI facilities become factory-like in scale and complexity, leaders are managing them with the discipline of advanced manufacturers. Throughput depends on how effectively the entire system is orchestrated across procurement, commissioning, operations, maintenance, security, and refresh cycles, rather than theoretical capacity. The strongest operators are establishing cross-functional control towers that bring together finance, supply chain, engineering, risk, sustainability, and technology teams into a single view of capacity, cost, performance, and constraint exposure. In this model, infrastructure management is more than a back-office priority. Asset lifecycle management—from sourcing and installation to optimization, maintenance, and recovery—becomes a strategic lever for scaling AI predictably under changing conditions.

AIOps and automation are accelerating this shift. By improving monitoring, governance, orchestration, and cost transparency across large GPU fleets, automation reduces human error and speeds response when new bottlenecks emerge.

“The strongest operators are establishing cross-functional control towers that bring together finance, supply chain, engineering, risk, sustainability, and technology teams into a single view of capacity, cost, performance, and constraint exposure.”

Scaling high-performance infrastructure for advanced research workloads

A specialized research group within a technology company relied on fragmented, unreliable compute environments to handle increasingly complex workloads. As demand surged, it required a scalable infrastructure and a more mature systems engineering operating model.

The team moved critical workloads from scattered compute rooms to a high-performance data center suite with next-generation technologies, including liquid-cooled networks, GPUs, and advanced storage. They built a structured demand pipeline; executed a more than $100 million procurement using standardized rack configurations; and coordinated construction, operations, and engineering in parallel. Legacy clusters were rationalized into a shared resource pool, supported by lifecycle management for operations, enhancements, and decommissioning.

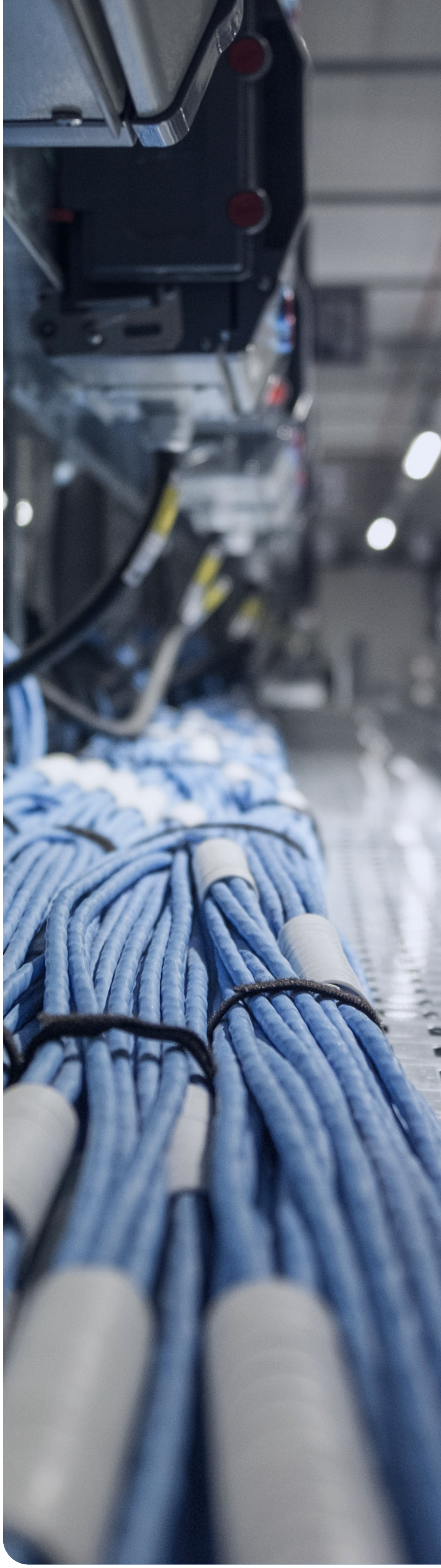
Agile delivery, formalized capacity planning, and strategic roadmapping strengthened the operating model and aligned to priorities. Customer engagement shifted to centralized intake, priority-based SLAs, and regular alignment with research teams.

The program delivered consistent high-performance compute, greater reliability, and improved alignment between engineering and end users—reducing risk, increasing scalability, and enabling next-generation research.

AI infrastructure economics move beyond total cost of ownership

AI leaders need an economic lens that goes beyond traditional total cost of ownership. Investment decisions increasingly need to account for return on resilience, speed-to-market advantages, and the value of optionality when bottlenecks move or technology changes. This is particularly important as hardware refresh cycles compress and organizations look for new ways to recover value through refurbishment, redeployment, remanufacturing, and materials recovery. In this environment, circular asset strategies become resilience levers as they help manage unpredictable costs, downtime, and sustainability concerns.

> "Investment decisions increasingly need to account for return on resilience, speed-to-market advantages, and the value of optionality when bottlenecks move or technology changes."



Distributed architecture is a resilience strategy

Although hyperscale campuses remain the engines for large-scale training, regional facilities and edge deployments are becoming essential for inference, latency-sensitive workloads, and compliance with local requirements.

This is not a binary choice between centralized and decentralized infrastructure. It is a continuum. The most resilient operators can place workloads across hyperscale hubs, regional nodes, and edge environments based on economics, performance requirements, regulatory constraints, and capacity.

That flexibility matters because infrastructure topology is now a constraint-management tool. Organizations can use distributed architectures to sidestep localized bottlenecks, improve failover options, and accelerate speed to market while maintaining governance and security discipline.

Defending AI at scale requires unified resilience

As AI infrastructure becomes mission-critical, security and resilience move to the center of operating strategy. Hyperscale campuses, regional nodes, and edge deployments are high-value assets that must be protected against cyber threats, physical disruption, operational failures, and regulatory exposure.

The strongest operators integrate cybersecurity and physical security rather than managing them separately. They are combining zero-trust architectures, threat monitoring, incident response, secure facilities, redundancy, cross-region failover, and recovery testing into one operating discipline.

Resilience is no longer a cost center. It is a prerequisite for uninterrupted throughput, stakeholder confidence, and long-term growth.

What this means for decision-makers

The forces reshaping AI infrastructure are not isolated trends. They are redefining every stage of the lifecycle—from strategy and capital allocation to site selection, design, execution, operations, and refresh.

In practical terms, leaders should focus on four priorities:

01 Develop an integrated capital and energy strategy.

Recognize that grid timelines are longer than tech timelines. Instead, consider alternative or supplemental power strategies and plan power procurement, facility design, capacity expansion, and funding as one interdependent system. Wire every dollar of capital to a kilowatt of capacity, sync every expansion to energy availability, and configure every facility for cost, resilience, and environmental profile. The strongest operators turn power access from a constraint into a catalyst, and use procurement to dictate speed, scale, and competitive position.

02 Embed license-to-operate requirements into design.

Treat compliance, sustainability, and community impact as design inputs, not downstream hurdles. By weaving sustainability targets, regulatory compliance, and community engagement into the earliest site and facility discussions, you front-load trust into the project pipeline. This shifts the conversation from compliance to competitiveness and opens doors to capital, talent, and partnerships.

03 Implement an agile industrial model.

The operating model must be designed to move as fast as the bottlenecks shift. Use system-level management disciplines, including a cross-functional control tower, to coordinate assets, workflows, and performance across the lifecycle. Capacity planners are contending with a multiconstraint environment that requires the adaptation of industrial supply chain planning solutions such as MRP-II into a digital infrastructure construct. This includes:

- Building capacity deployment schedules, including demand growth, sovereign requirements, tenancy commitments, and real estate (buy, build, lease)
- Just-in-time thinking that can adapt to technology shifts that occur within procurement cycles
- DC bill of materials that covers energy, cooling, physical structure, networks, and technology components
- Materials and capacity requirement planning that covers the bill of materials as well as trade labor and specialty contractor capacity
- DC floor control, including construction sequencing and technology work packages
- Inventory management that involves material staging, procurement buffers, and spare parts management
- Program and project management and capital project accounting.

A control tower with real-time visibility across assets, workflows, and constraints enables leaders to see emerging choke points before they emerge in performance metrics. An agentic AI application that operates autonomously and optimizes target business outcomes by coordinating procurement, engineering, finance, risk, and operations in one shared system makes capacity growth predictable even when external shocks hit.

04 Build a unified resilience framework.

Resilience only works when it's treated as a single organism. Integrating cyber, physical, and operational defenses into one framework reduces blind spots and accelerates response. This means designing recovery protocols, failover routes, and threat detection that work across regions, scales, and facility types, with governance that treats uninterrupted throughput as a measure of operational health, not a lucky outcome.

Keep your options open

Choice is the antidote to constraint whiplash, so use distributed architectures, circular asset strategies, and proactive scenario planning to stay ahead of shifting constraints. This gives operators the ability to shift workloads, repurpose assets, and redirect growth with minimal disruption. This is how leaders design for optionality and turn uncertainty into a platform for competitive moves.

For executive teams, this agenda should translate into a near-term action plan: identify the current bottleneck, quantify its impact on growth, align capital and operating decisions around the next two to three constraints, and establish governance that allows the enterprise to respond at industrial speed.

Leadership in the industrial era of AI is not a one-time achievement. It is a continuous capability: the ability to identify the current bottleneck, mobilize the enterprise to resolve it, and anticipate the next source of constraint before it limits growth.

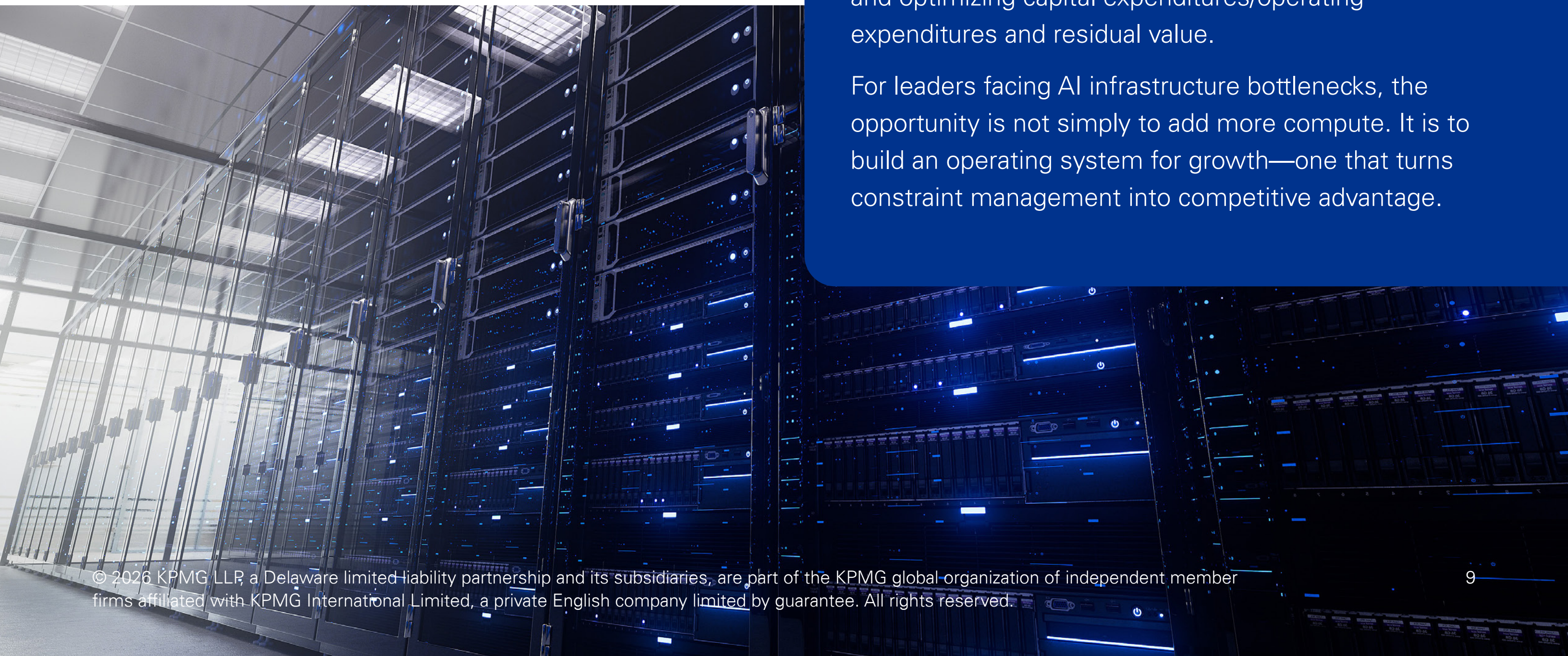
How KPMG can help

The transition to industrial-scale AI operations demands more than technical experience. It requires an adviser that understands the intersection of infrastructure, operations, finance, regulation, risk, and transformation—and can connect strategy to execution across the full asset lifecycle.

KPMG can help organizations bring industrial-grade discipline to AI infrastructure planning and execution—linking strategy to implementation across site selection, capital planning, operating model design, risk and resilience, regulatory navigation, and long-term value realization.

That support can include wide-ranging lifecycle experience, multidisciplinary teams that bridge technology and infrastructure, global-local insight into market constraints, and established methodologies adapted from complex industrial and capital-intensive environments. We also offer asset management across the full lifecycle: establishing asset registers/CMDBs and financial/operational tracking; commissioning, onboarding, governing, and decommissioning refresh and technology upgrades; and optimizing capital expenditures/operating expenditures and residual value.

For leaders facing AI infrastructure bottlenecks, the opportunity is not simply to add more compute. It is to build an operating system for growth—one that turns constraint management into competitive advantage.



Authors

Vijay Subramanyam

Principal, Advisory Technology, Media, and Telecommunications (TMT)
Consulting Sector Leader, KPMG US
vijaysubramanyam@kpmg.com

In this role, Vijay leads the firm's consulting strategy and delivery for clients across the TMT sector, helping them navigate disruption, accelerate digital transformation, and unlock sustainable growth. With more than two decades of experience in management consulting, Vijay brings extensive knowledge in enterprise transformation, business model transitions, technology strategy and modernization, risk & compliance and cyber security. He works closely with C-suite leaders to drive innovation through AI-powered solutions, data-driven decision-making, and agile operating models tailored to the unique challenges of the TMT industry.

Reid Tucker

Principal, Capital Projects Advisory,
KPMG US
reidtucker@kpmg.com

Reid is currently a principal in the Major Projects Advisory practice with more than 25 years of project management, construction management, project accounting and contract administration experience. Reid has performed all aspects of the construction process from performing project due diligence, planning, risk management, design, cost analysis, environmental analysis, contract negotiations, construction management, and construction administration. Reid is currently leading projects in San Francisco, Los Angeles, and Sacramento, including developing project controls, systems, and performing construction audits and risk assessments for major construction programs. Reid has extensive knowledge in alternative delivery methods including alliances, integrated project delivery, lean, and P3s.

Vivek Sudhakar

Principal, Advisory
Platform, Oracle, KPMG US
viveksudhakar@kpmg.com

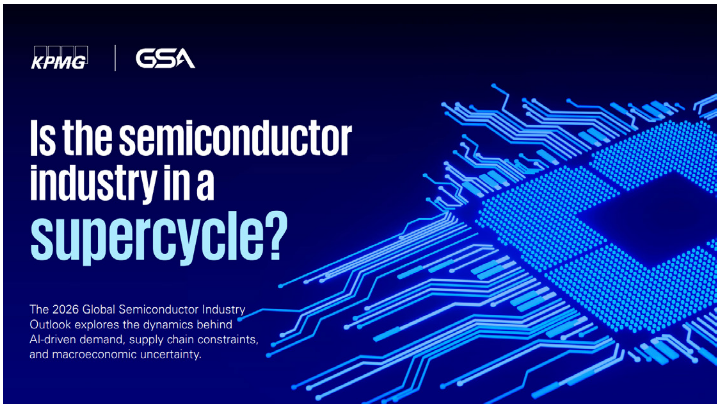
Vivek brings more than two decades of experience in IT and consulting industries and more than 14 years of experience in IT leadership. Vivek fosters long-term advisory relationships with C-level executives by providing guidance, coaching, and support. His focus includes enterprise applications (ERP, CRM, and SCM), business intelligence and data science, Internet of Things (IoT), digital transformation, executive coaching and guidance, technology and business leadership, program management, and Agile/DevOps adoption. As an IT executive, he has an excellent track record helping clients achieve their strategic business goals. Vivek is a research-driven critical thinker with broad and deep knowledge across a wide array of technologies and has industry specialization in industrial manufacturing, automotive, CPG, and healthcare.

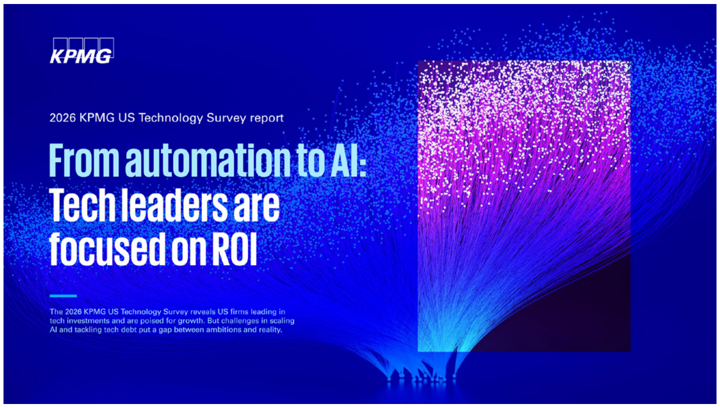
Iacopo Gambino

Managing Director, TMT, Tech & Data Strategy, KPMG US
iacopogambino@kpmg.com

Iacopo is a professional with 20 years of experience supporting a broad spectrum of multinational clients in the US and in highly regulated European markets, leading large scale, multi-year technology and infrastructure transformations. He has hands on experience in supporting the whole datacenter lifecycle, from Strategy to Capacity refresh and decommissioning. His strategic perspective is informed by his deep understanding of AI, modern technologies, and infrastructure.

Related thought leadership





Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

Please visit us: in | **kpmg.com** | **Subscribe**

DASD-2026-20103