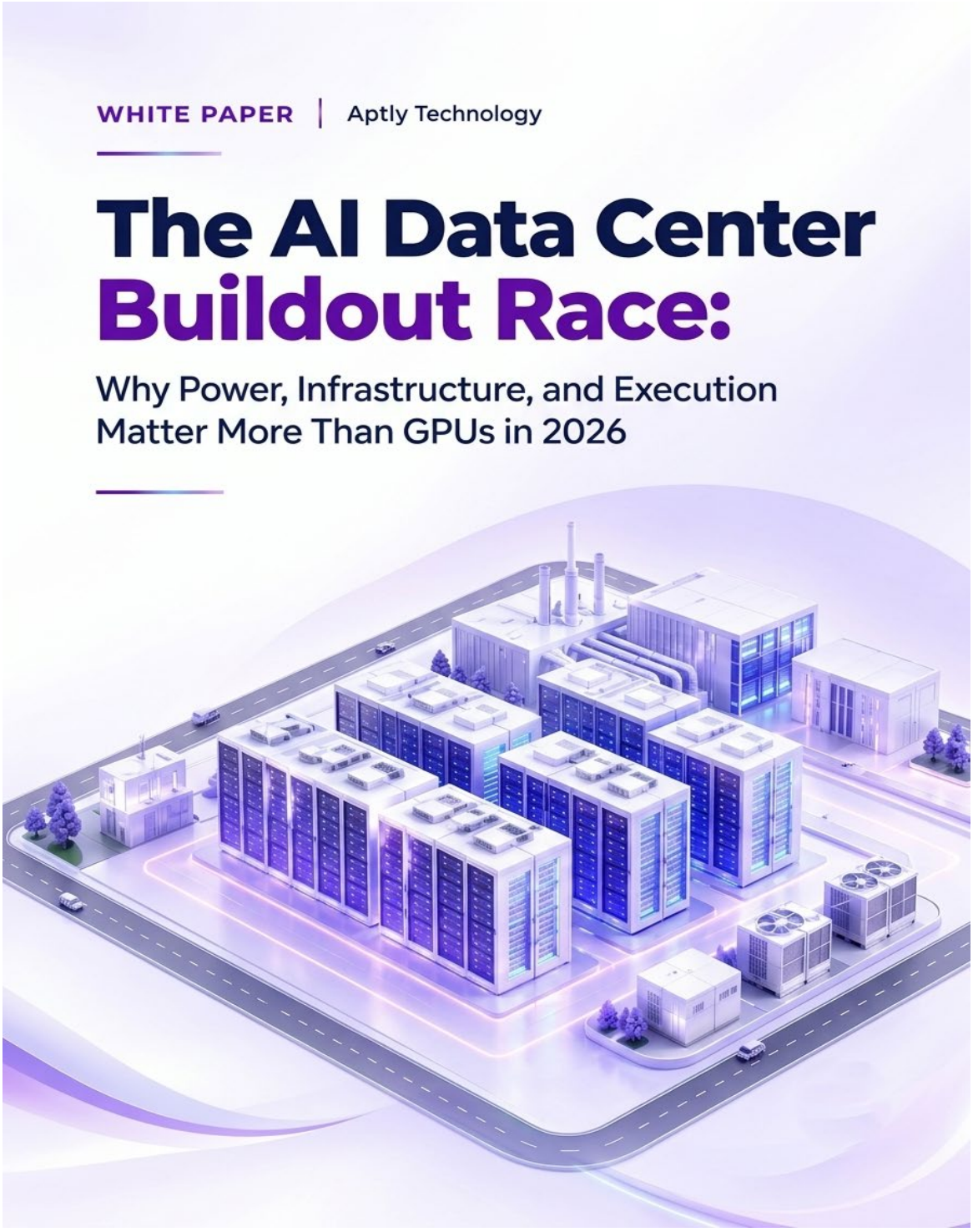


White Paper | Aptly Technology | June 2026

How enterprises can prepare for the next phase of AI growth when power, cooling, and deployment speed become the true competitive differentiators.

Table of Contents

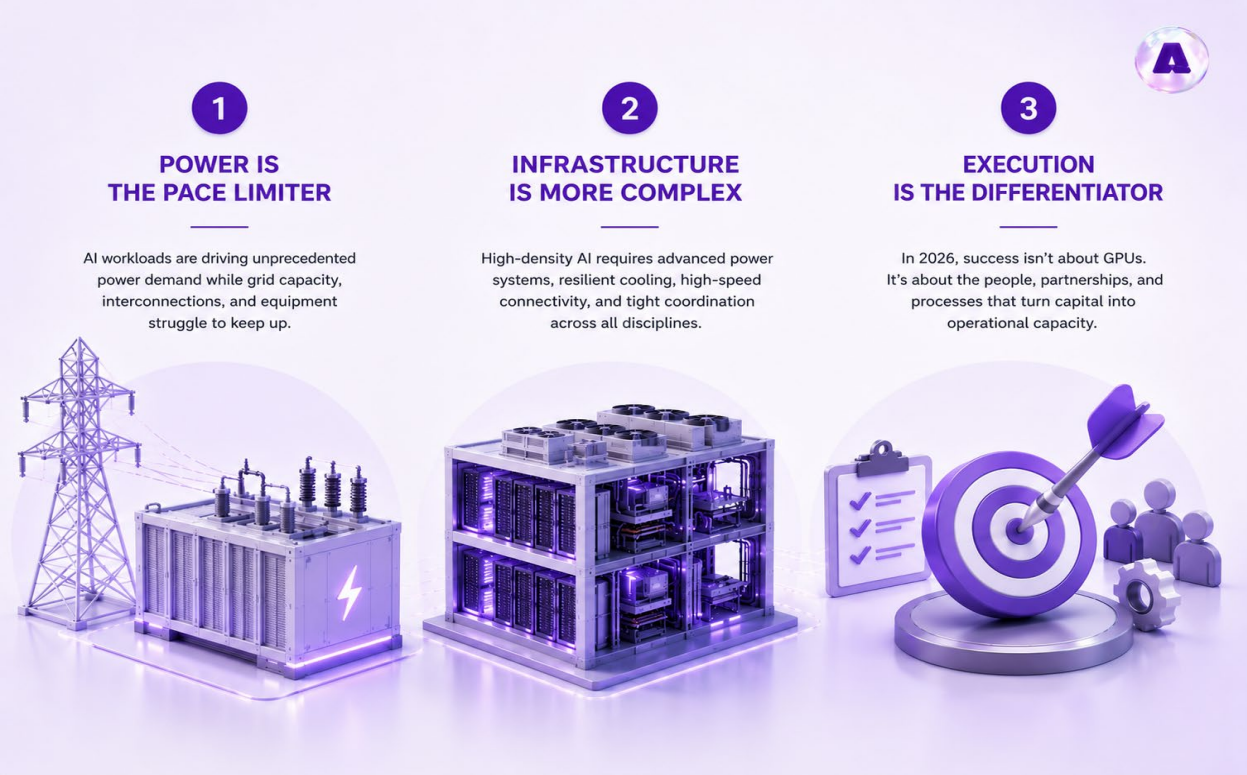
Executive Summary....3
1. The Shift from Compute Scarcity to Capacity Scarcity....5
1.1 Power Has Become the Primary Constraint....7
1.2. Infrastructure Bottlenecks Extend Beyond the Grid....8
1.3. Execution Is the New Differentiator....10
2. GPUs Still Matter, but They No Longer Decide the Race Alone....12
For Operators....14
For Investors....14
For Policymakers....15
4. Geography, Market Structure, and the New Logic of AI Capacity....15
5. An Executive Action Agenda for 2026....16
6. Building AI Infrastructure for Scale and Speed....16
Secure Power Early and Treat It as a Strategic Asset....17
Design Infrastructure Around AI Workloads....17
High-Density Electrical Architectures....17
Advanced Cooling Strategies....17
Flexible Infrastructure Designs....18
Build Network Architectures Designed for AI....18
Establish Strategic Vendor Partnerships....18
What Leading Organizations Are Doing:....19
Standardize and Accelerate Execution....19
Standardized Reference Architectures....19
Integrated Program Governance....19
Automation-Driven Deployment....20
Build for Continuous Scalability Rather Than Immediate Demand....20
Maximizing Returns on AI Infrastructure Investments....20

Building Scalable Foundations for Future Growth 22
Strengthening Competitive Position Through Execution Excellence 22
Reducing Risk Through Strategic Infrastructure Planning 23
Economic Influence Beyond Technology 23
Aptly's Core Capabilities 25
Why This Matters 26

Executive Summary

In the early years of the generative AI boom, GPUs were the industry's biggest concern. Organizations competed aggressively to secure AI accelerators, making compute availability the primary factor driving AI deployment.

By 2026, the situation has changed. GPUs remain important, but they are no longer the only constraint on AI growth. The **three forces driving the AI data center buildout race** are:



As a result, the race is about delivering AI-ready capacity that is powered, connected, cooled, and operational at scale.



The conclusion of this white paper is straightforward: ***In 2026, success is determined by the ability to deliver operational megawatts, live racks, and revenue-generating AI services faster than competitors.*** GPUs still matter enormously, but they no longer decide the race alone. That is why the center of gravity in AI infrastructure strategy has moved from chips alone to the broader system of energy, land, cooling, networking, construction, and execution.

1. The Shift from Compute Scarcity to Capacity Scarcity



The first phase of the generative AI boom was defined by a scramble for accelerators. Enterprises, cloud providers, and model developers competed for scarce GPU supply, and access to chips was widely viewed as the primary source of competitive advantage. By 2026, however, that assumption has changed.

Organizations increasingly recognize that owning or reserving compute does not automatically translate into deployable AI capacity. A GPU that cannot be installed into a rack with sufficient power, cooling, networking, and operational support has limited economic value. As a result, the bottleneck has shifted from compute procurement to infrastructure deployment and readiness.

This distinction matters because AI infrastructure is now a systems problem. The strategic question is no longer simply, ***"Who can buy the most chips?"*** It is, ***"Who can convert capital into operating megawatts, usable racks, and production-grade clusters at scale?"*** That question moves the conversation beyond semiconductor supply into utility relationships, substation design, cooling architecture, network engineering, workforce availability, and program management. In this environment, time-to-power and time-to-revenue matter more than procurement commitments alone.

The shift is particularly evident as investment continues to accelerate. According to IDC, worldwide AI infrastructure spending reached $318 billion in 2025—more than double 2024 levels.



This highlights that capital is increasingly available even as physical infrastructure remains constrained. As a result, infrastructure is now valued less by announced capacity and more by energized, commissionable, and monetizable capacity.

At the same time, new opportunities are emerging for hyperscalers and operators with access to power-rich sites, specialized cooling technologies, faster development pathways, or regional delivery expertise. In 2026, AI capacity is produced through coordinated execution across an increasingly interconnected infrastructure ecosystem.

1.1 Power Has Become the Primary Constraint

Power availability has become the defining variable in the 2026 AI data center race.



Many industry reports about the same challenge: companies have the funding, land, and demand needed to build new AI data centers, but projects are delayed because power infrastructure cannot be delivered quickly enough. Utility connections, substations, transmission upgrades, and critical electrical equipment frequently take years to complete. As a result, the ability to secure and activate power has become more important than simply announcing new capacity.

Today, hyperscalers and data center operators are requesting massive amounts of power within very short timeframes. EPRI estimates that data center electricity demand projections for 2030 are now **60% higher than its 2024 forecasts**, driven by the accelerated pace of AI infrastructure development. In many regions, utilities are struggling to process the growing number of large-load requests, creating long interconnection queues.

Even when power can eventually be provided, transmission upgrades, substations, and distribution infrastructure require extensive planning, permitting, and construction. In some cases, the wait for power can exceed the lifespan of a typical GPU generation.

These constraints are changing where data centers are built. While traditional data center hubs remain important, they no longer automatically attract new investment if power cannot be delivered on schedule. Instead, hyperscalers are prioritizing regions that offer faster utility coordination, available grid capacity, supportive permitting processes, and reliable power sources. Site selection decisions are increasingly driven by a simple question: ***where can large amounts of power be delivered within the next two to four years?***

Many organizations are also exploring hybrid and behind-the-meter power strategies to improve reliability and reduce deployment risk. If power is the primary constraint on AI growth, then securing reliable access to power becomes a key competitive advantage.



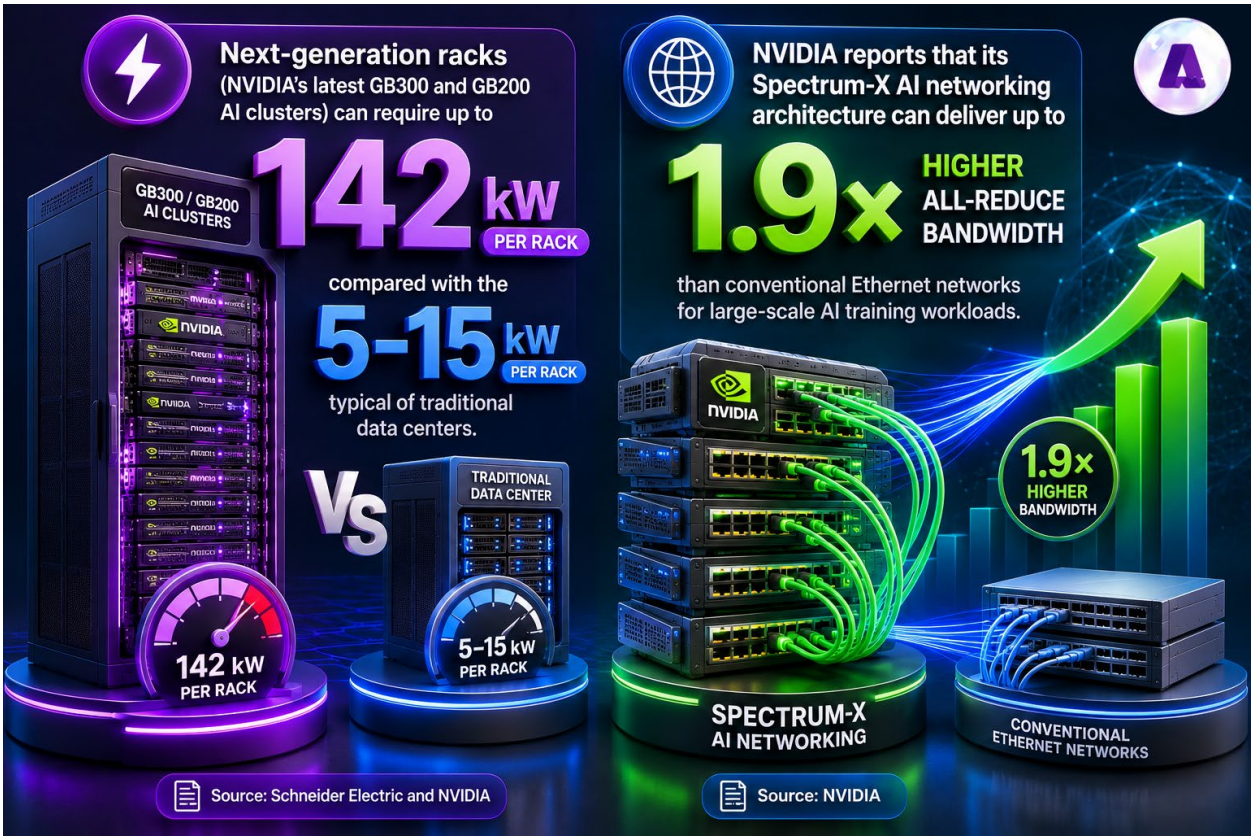
1.2. Infrastructure Bottlenecks Extend Beyond the Grid

Even when sufficient power is available, AI-ready infrastructure introduces a new set of challenges. AI training and inference clusters are driving rack densities far beyond what traditional data centers were designed to support. As a result, operators must rethink power distribution, cabling, floor layouts, backup systems, and cooling infrastructure.

The industry is increasingly recognizing that not all data center capacity is equal. Capacity designed for high-density AI workloads is fundamentally different from capacity built for traditional enterprises.

Cooling is one of the clearest examples of this shift. As rack power densities increase, conventional air-cooling systems become less effective. This makes liquid cooling a growing requirement for many AI deployments. However, implementing liquid cooling is not as simple as replacing one technology with another. Operators must evaluate different options including direct-to-chip cooling, rear-door heat exchangers, immersion cooling, and hybrid approaches. Local factors such as including water availability, environmental regulations, permitting requirements, and access to specialized talent can also influence which cooling strategy is viable.

Networking is becoming another critical infrastructure requirement. AI factories depend on high-bandwidth, low-latency connections to link thousands of GPUs across racks. As AI clusters scale, optical networking, fiber capacity, switching infrastructure, and interconnection design become essential to overall system performance. A site with adequate power but insufficient networking capability may still struggle to operate effectively as an AI platform.



These evolving requirements are reshaping data center development strategies. While retrofitting existing facilities can be attractive when valuable power and network connections are already in place, the ***technical complexity and execution risk*** increase significantly as density requirements rise.

In 2026, the most successful infrastructure builders are not necessarily those creating the most complex designs. Instead, they are the organizations that standardize wherever possible, reducing deployment risk while focusing innovation on the areas that deliver the greatest competitive advantage.



1.3. Execution Is the New Differentiator

One of the most important lessons of the AI infrastructure race is that capital alone does not guarantee results. Many projects struggle to move from planning to operation because of delays in power delivery, equipment procurement, permitting, construction, and commissioning. As AI campuses become complex, execution capability is becoming a competitive advantage. Industry estimates suggest that **30%–50% of AI data center projects planned for 2026 may be delayed**. This shows the growing gap between investment commitments and actual capacity delivery.

Time-to-market has become a critical business metric. Every month of delay can postpone model training, reduce available inference capacity, delay customer commitments, and lower returns on expensive infrastructure investments. As a result, project controls, procurement visibility, integrated scheduling, and commissioning readiness are becoming executive priorities rather than purely operational concerns.

Without realistic planning, leading operators can no longer assume that signed agreements will automatically translate into energized facilities. Instead, they actively manage critical-path risks, challenge optimistic assumptions, and recognize that many project milestones depend on external stakeholders.

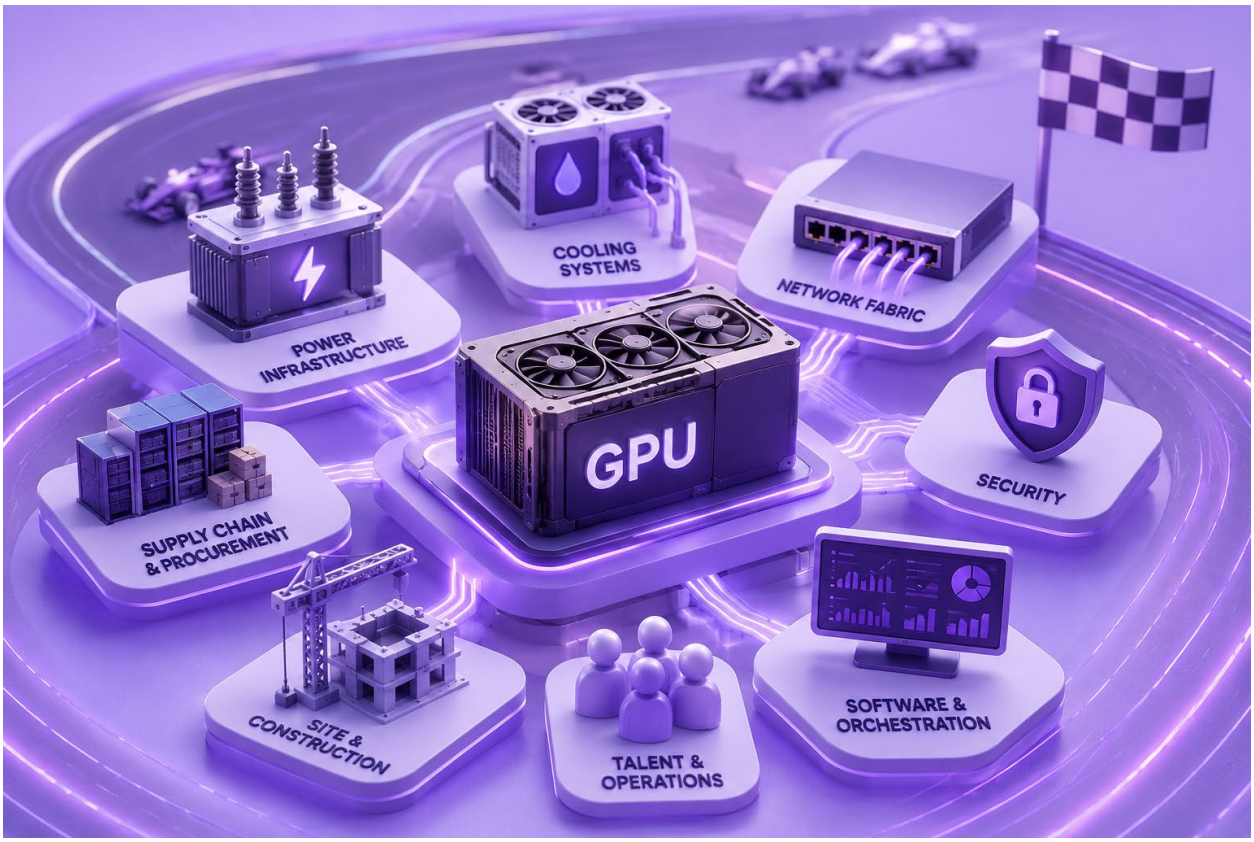
Workforce availability has become another major constraint. The AI infrastructure boom depends on electricians, mechanical contractors, commissioning specialists, controls engineers, and network professionals. ***Supply chains*** also remain a challenge.

As a result, organizations that consistently bring capacity online on schedule can build a reputation for reliability, while those that repeatedly miss delivery targets risk losing credibility with customers, investors, and partners.



2. GPUs Still Matter, but They No Longer Decide the Race Alone

None of this suggests that GPUs have become unimportant. Accelerators remain indispensable, and leading-edge compute continues to influence performance, cost, model capability, and competitive positioning. GPU access still matters for training frontier models, serving high-demand inference workloads, and maintaining bargaining power within the AI ecosystem.



But in 2026, GPUs are best understood as one component in a much larger production system. Their strategic value depends on the environment into which they are deployed:

- Adequate power
- Efficient cooling
- Resilient networking
- Skilled operations
- Infrastructure capable of supporting increasingly dense AI clusters without compromising reliability

The scale of these requirements is growing rapidly. According to the International Energy Agency (IEA), global data center electricity consumption is projected to increase from approximately **415**

TWh in 2024 to nearly 945 TWh by 2030, with AI serving as a major driver of that growth. This underscores a fundamental reality: ***acquiring compute is only part of the challenge; supplying and operating that compute at scale is becoming equally important.*** Where these supporting foundations are weak, GPU access alone does not create durable competitive advantage.

This is why market conversation has broadened from chips to full-stack infrastructure. Operators are increasingly judged not only by how much compute they can acquire, but by how efficiently they can translate that compute into revenue-generating service capacity. Infrastructure readiness changes the economics of silicon. If a company procures large numbers of accelerators but cannot deploy them quickly, the carrying cost of that capital rises while commercial returns are delayed.

Conversely, a company that acquires fewer chips but brings them online rapidly within a well-designed environment can often generate stronger effective returns. In this context, power contracts, electrical architecture, liquid-cooling capability, fiber density, and execution maturity are not supporting details; they are key determinants of return on AI capital.

The same logic applies to supply-chain strategy. The companies best positioned in 2026 are not simply those with strong GPU relationships; they are those with coordinated visibility across semiconductors, optics, transformers, switchgear, liquid-cooling components, and construction schedules. A delay in any one of these layers can reduce the realized value of all the others. At the same time, networking infrastructure is becoming increasingly strategic. Industry forecasts estimate that the market for AI-focused optical transceivers will grow from approximately **$16.5 billion in 2025 to $26 billion in 2026**, reflecting the growing importance of high-speed connectivity in large-scale AI deployments.

Competitive advantage therefore comes from integration: the ability to synchronize chip delivery with power availability, facility readiness, network activation, and commissioning schedules. Organizations that recognize this early are more likely to convert infrastructure spending into durable market share rather than stranded or underutilized assets.

This broader framing also changes who is perceived as strategically important. In a GPU-centric narrative, the obvious winners are chip designers and the largest buyers of silicon. In an infrastructure-centric narrative, the winner set expands to include utilities, power-equipment manufacturers, cooling specialists, network providers, project-delivery firms, and operators capable of bringing capacity online faster than their peers. That does not diminish the importance of accelerators. Rather, it reflects a more mature understanding of how value is created across the AI stack. Sustainable advantage increasingly depends on the ability to integrate compute, power, cooling, networking, and execution into a single operational system.

In 2026, compute remains essential, but it is no longer sufficient on its own.

3. Strategic Implications for Operators, Investors, and Policymakers

For Operators

- Shift from compute-centric planning to infrastructure-centric planning.
- Treat grid access, utility relationships, cooling strategy, transmission dependencies, network architecture, and delivery governance as board-level priorities.
- Base capacity planning on realistic assumptions about energization timelines and infrastructure readiness, not solely on demand forecasts and hardware roadmaps.
- Standardize campus designs where possible to improve speed, scalability, and operational consistency.
- Secure critical equipment and supply-chain dependencies early to reduce execution risk.
- Strengthen cross-functional coordination across power, construction, networking, procurement, and commissioning teams.
- Focus on reducing time-to-capacity rather than maximizing announcement-driven growth.

For Investors

- Evaluate not only a company's ability to acquire AI hardware, but also its ability to deploy and utilize capacity efficiently.
- Assess project pipelines with particular attention to energization assumptions, utility relationships, and power procurement strategies.
- Examine whether announced capital expenditures are translating into operational capacity on realistic timelines.
- Monitor utilization rates and infrastructure returns alongside traditional measures of technology investment.
- Identify gaps between spending commitments and delivered capacity, as these can become significant sources of execution risk.
- Recognize that long-term value creation increasingly depends on infrastructure delivery capability, not semiconductor access alone.

For Policymakers

- Accelerate permitting processes for power, transmission, and digital infrastructure projects.
- Invest in grid modernization and transmission expansion to support growing AI-related electricity demand.
- Strengthen domestic manufacturing capacity for critical electrical and infrastructure components.
- Expand workforce development programs for electricians, engineers, commissioning specialists, and skilled construction labor.
- Improve coordination between digital infrastructure policy and energy policy.
- Create regulatory environments that enable infrastructure deployment while maintaining reliability and sustainability goals.
- Recognize that the AI infrastructure race is increasingly a competition among jurisdictions to provide the power and infrastructure conditions required for next-generation digital capacity.

4. Geography, Market Structure, and the New Logic of AI Capacity

One of the most important consequences of the power-and-execution shift is that geography matters more than it did in earlier cloud cycles. Traditional data center markets still benefit from ecosystem density, fiber connectivity, workforce familiarity, and proximity to customers. Yet in 2026, those advantages can be offset by grid congestion, community opposition, water constraints, or long utility timelines. As a result, the map of AI capacity is beginning to diversify.

Operators are still drawn to the largest established hubs, but they are also pursuing secondary and emerging markets that offer a more favorable combination of power availability, land, permitting speed, and network access.

This is changing market structure in at least three ways.

1. AI training and large-scale inference are no longer assumed to consolidate in only a handful of flagship regions.
2. Colocation, utility partnerships, and energy-adjacent development models are becoming more strategically significant because they can shorten time-to-capacity.
3. Regional differentiation is increasing: some markets will specialize in very large training campuses with exceptional power envelopes, while others will focus on lower-latency inference, sovereign capacity, or mixed enterprise AI deployments.

Land strategy is therefore changing as well. In prior cycles, proximity to large peering hubs dominated siting decisions. In the AI cycle, operators increasingly value parcels that can support substation development, cooling infrastructure, staged campus expansion, and future transmission access. Over time, this will likely widen the circle of relevant data center geographies and create new winners among jurisdictions that combine energy abundance, permitting realism, and infrastructure-friendly policy.

5. An Executive Action Agenda for 2026

For executive teams, the immediate priority is to align AI ambition with physical delivery reality. That begins by reframing infrastructure planning around a few questions that are often more decisive than GPU allocations:

- Where will the next increments of capacity be energized?
- What equipment and utility dependencies sit on the critical path?
- What cooling and network assumptions are embedded in the current design?
- What contingencies exist if any of those assumptions fail?

Organizations that can answer those questions with precision will make better capital decisions than those that focus only on hardware road maps or market narratives.

A practical executive agenda for 2026 includes five priorities.

1. Treat power procurement and interconnection as strategic workstreams with direct executive oversight.
2. Build integrated visibility across design, procurement, construction, networking, and commissioning so that schedule risk is visible early rather than discovered late.
3. Standardize facility designs and operating playbooks wherever possible to accelerate repeatability.
4. Strengthen partnerships with utilities, engineering firms, equipment suppliers, and regional stakeholders rather than treating them as downstream vendors.
5. Measure success in terms of energized capacity delivered, utilization achieved, and time-to-revenue captured.

6. Building AI Infrastructure for Scale and Speed

The organizations winning the AI infrastructure race are not simply procuring more GPUs. They are adopting a holistic strategy that aligns power, infrastructure, networking, vendor ecosystems, and execution under a unified operating model. Success increasingly depends on eliminating

bottlenecks across the entire infrastructure lifecycle rather than optimizing individual components in isolation.

Secure Power Early and Treat It as a Strategic Asset

Power availability has emerged as one of the most significant determinants of AI project timelines. Global data center capacity demand is projected to increase from **82 GW in 2025 to 219 GW by 2030**, representing a 3.5x increase in demand. AI workloads alone are expected to account for approximately **70% of total capacity demand by 2030**.

As AI demand accelerates, leading organizations are moving power planning to the earliest stages of AI infrastructure development by:

- Conducting power feasibility studies before site selection
- Engaging utility providers during planning rather than construction
- Reserving future capacity for expansion
- Evaluating alternative and distributed power strategies
- Incorporating long-term energy availability into infrastructure roadmaps

Organizations that secure power early gain a significant advantage in deployment speed, scalability, and investment certainty.

Design Infrastructure Around AI Workloads

Many organizations invest heavily in GPUs but fail to achieve expected performance because supporting infrastructure cannot keep pace. Traditional facility designs were never intended to support these density requirements. NVIDIA GB200-based systems may require up to **120 kW per rack**, making power and cooling design critical to infrastructure performance.

AI-ready environments should prioritize the following to improve utilization rates and maximize return on AI investments.:

High-Density Electrical Architectures

- Scalable power distribution systems
- Intelligent power monitoring
- Redundant electrical pathways
- Future-ready capacity planning

Advanced Cooling Strategies

- Direct-to-chip liquid cooling
- Rear-door heat exchangers
- Hybrid cooling environments
- Thermal optimization and monitoring

Flexible Infrastructure Designs

- Modular deployment models
- Expansion-ready rack configurations
- Adaptable mechanical systems

Organizations that continue relying on legacy data center designs risk encountering infrastructure limitations long before they exhaust compute capacity.

Build Network Architectures Designed for AI

In AI environments, network performance directly influences compute performance. Large-scale AI clusters depend on thousands of accelerators communicating continuously across the network fabric. Even minor latency issues can reduce overall cluster efficiency and extend model training timelines.

Leading organizations are investing in:

- High-bandwidth spine-leaf architectures
- Low-latency networking fabrics
- AI-optimized east-west traffic designs
- Scalable network automation
- End-to-end observability and monitoring

As AI deployments scale, networking is becoming a primary determinant of infrastructure utilization and return on investment.

Establish Strategic Vendor Partnerships

The complexity of AI infrastructure has made vendor strategy a critical success factor. AI deployments now involve interconnected ecosystems spanning:

- Power and utility providers
- Data center infrastructure vendors
- Cooling technology suppliers

- Network equipment manufacturers
- Compute and GPU providers
- Systems integrators
- Managed service partners

Organizations that approach procurement as a series of isolated transactions often experience integration challenges, supply chain delays, and operational inefficiencies.

What Leading Organizations Are Doing:

- Developing long-term vendor relationships rather than project-based engagements
- Aligning infrastructure roadmaps with strategic technology partners
- Diversifying suppliers to reduce supply chain risk
- Conducting joint planning sessions with key vendors
- Standardizing technology stacks across deployments

Strategic vendorship like Aptly Technology enables faster deployment cycles, improved interoperability, better pricing leverage, and reduced implementation risk.

In an environment where infrastructure delivery timelines increasingly determine AI competitiveness, strong vendor ecosystems have become a strategic advantage.

Standardize and Accelerate Execution

Infrastructure execution has emerged as one of the most important competitive differentiators in the AI era. Leading organizations are accelerating delivery through:

Standardized Reference Architectures

- Repeatable infrastructure designs
- Pre-validated deployment models
- Consistent operational frameworks

Integrated Program Governance

- Unified project management offices
- Cross-functional stakeholder alignment
- Real-time project visibility
- Proactive risk management

Automation-Driven Deployment

- Infrastructure-as-Code
- Automated provisioning
- Configuration standardization
- Continuous validation processes

The ability to deploy infrastructure repeatedly, consistently, and at scale is rapidly becoming more valuable than isolated technology advantages.

Build for Continuous Scalability Rather Than Immediate Demand

The next generation of AI workloads will require substantially more infrastructure than organizations deploy today. Inference workloads alone are expected to become the dominant driver of AI data center demand by 2030, creating continuous pressure on power, networking, and cooling systems. Leading organizations design around future growth. This requires:

- Expansion-ready facility designs
- Scalable power and cooling architectures
- Modular network fabrics
- Long-term capacity forecasting
- Infrastructure lifecycle planning

The organizations that succeed in the next phase of AI adoption will be those that view infrastructure as a continuously evolving strategic capability.

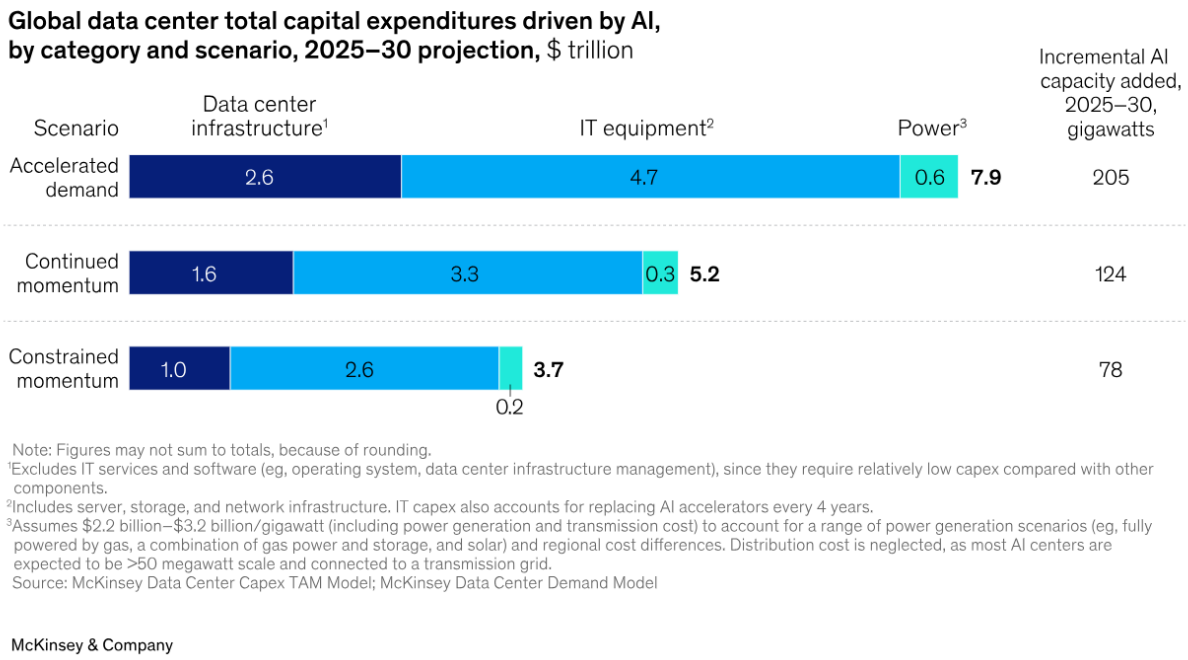
The ROI of Getting AI Infrastructure Right: The Real Business Impact

Organizations that recognize the shift from a GPU-centric mindset to an infrastructure-first strategy are positioning themselves to capture AI value faster, scale more efficiently, and reduce deployment risk. The following analysis is based on IEEE ComSoc Technology.

Maximizing Returns on AI Infrastructure Investments

AI infrastructure spending is reaching unprecedented levels, raising the stakes for infrastructure planning and execution.

Capital investments to support AI-related data center capacity demand could range from about $3 trillion to $8 trillion by 2030.



According to IEEE ComSoc's analysis, major technology companies are expected to invest between **$364 billion and $400 billion annually** in AI data centers, AI hardware, and cloud infrastructure. At the same time, Sequoia Capital estimates that AI infrastructure investments made in 2023 and 2024 alone require approximately **$800 billion in AI-generated revenue** over the useful life of those assets to achieve acceptable investment returns.

Organizations that maximize infrastructure utilization through effective power planning, networking, cooling, and operational governance are better positioned to generate value from these significant investments.

Business Impact

- Higher infrastructure utilization
- Improved ROI on AI investments
- Better asset productivity
- Reduced operational waste

Building Scalable Foundations for Future Growth

The scale of AI infrastructure expansion reflects the expectation that AI will become a foundational business capability across industries.

The four largest hyperscale providers alone are projected to spend nearly **$400 billion in capital investments annually**, exceeding the inflation-adjusted cost of the Apollo space program. Additionally, OpenAI's Stargate initiative and related ecosystem investments represent commitments measured in the hundreds of billions of dollars.

Organizations that build scalable infrastructure today can support future AI growth without repeated redesigns or disruptive modernization projects.

Business Impact

- Future-ready infrastructure
- Simplified expansion planning
- Lower long-term capital costs
- Greater business flexibility

Strengthening Competitive Position Through Execution Excellence

History shows that transformative technologies create winners not only through innovation but through execution.

The IEEE analysis draws parallels between today's AI infrastructure buildout and the fiber-optic expansion of the late 1990s. During that period, telecom providers invested more than **$100 billion** in fiber infrastructure based on anticipated internet demand. While many companies failed, the infrastructure ultimately became the foundation for future digital services such as streaming, cloud computing, and social media.

The lesson for today's leaders is clear: infrastructure readiness can create lasting competitive advantages, even as technologies and business models evolve.

Business Impact

- Faster innovation cycles
- Improved market responsiveness

- Stronger customer experiences
- Sustainable competitive differentiation

Reducing Risk Through Strategic Infrastructure Planning

One of the most significant risks facing AI initiatives is the gap between infrastructure investment and business adoption.

Some of the research highlights that:

- **95% of custom enterprise AI projects fail to generate measurable financial impact or reach full-scale production,** reported by MIT Media Lab.
- Gartner reports that **85% of AI initiatives fail to deliver their expected value.**
- RAND research indicates that **AI projects experience failure rates exceeding 80%**, roughly double that of traditional technology initiatives.

Organizations that combine infrastructure readiness with strong governance, business alignment, and operational planning significantly improve their ability to convert AI investments into measurable outcomes.

Business Impact

- Reduced project failure rates
- Better alignment between technology and business objectives
- Improved operational resilience
- Greater confidence in AI investments

Economic Influence Beyond Technology

AI infrastructure investment is increasingly influencing broader economic growth. According to the IEEE ComSoc analysis, some economists estimate that AI-related infrastructure spending could contribute nearly **half of U.S. GDP growth in 2025**, underscoring the scale and strategic importance of data center construction, power deployment, and supporting infrastructure.

However, analysts also note that the long-term value will ultimately depend on whether organizations can translate infrastructure capacity into sustainable productivity gains and business outcomes.

Business Impact

- Enhanced innovation capacity
- Long-term productivity gains
- Greater organizational resilience
- Strategic positioning for the AI economy

From Strategy to Deployment: Aptly's AI Infrastructure Expertise

The AI infrastructure race is no longer determined solely by who can acquire the most GPUs. It is increasingly defined by who can secure power, design scalable infrastructure, and execute complex deployments efficiently.

Aptly helps organizations navigate this transition by delivering expertise across the entire AI infrastructure lifecycle—from AI-ready data center planning and network architecture to deployment, operations, and optimization.



As the only supplier trusted by Microsoft to build and support third-party hyperscale data centers globally, Aptly brings hyperscale experience to enterprises accelerating their AI initiatives.

Aptly's Core Capabilities

1. **Hyperscale GPU Data Center Deployment:** Aptly has supported the deployment and operation of large-scale GPU infrastructure environments consisting of tens of thousands of GPU nodes and AI-optimized networking fabrics. These environments required end-to-end coordination across power infrastructure, rack deployment, network integration, burn-in testing, and ongoing operational support.
2. **Enterprise AI Infrastructure Modernization:** Organizations increasingly require AI-ready environments capable of supporting model training, inference, and data-intensive workloads. Aptly has helped enterprises design and deploy scalable AI infrastructure environments that combine:
    - GPU compute resources
    - AI-optimized storage architectures
    - High-bandwidth networking
    - Security and compliance frameworks
    - Hybrid cloud integration
3. **AI Networking and Cluster Optimization:** Network architecture has become a critical factor in AI infrastructure performance. Aptly's AI networking expertise includes designing environments that support:
    - InfiniBand fabrics
    - RoCE architectures
    - High-bandwidth east-west traffic
    - GPU-to-GPU communication optimization
    - Large-scale AI cluster scalability
4. **Accelerating AI Infrastructure Deployment:** One of the biggest challenges facing organizations today is execution speed. Aptly leverages standardized deployment methodologies, automation-driven infrastructure practices, and pre-validated implementation playbooks developed through years of hyperscale data center operations.

This approach helps organizations reduce deployment complexity, accelerate time-to-production, and improve infrastructure consistency across multiple environments.

Why This Matters

Industry forecasts indicate that global data center demand may nearly triple by 2030, while AI workloads account for the majority of future growth. At the same time, power constraints, cooling requirements, networking complexity, and execution challenges are emerging as the primary barriers to AI adoption.

Aptly helps organizations overcome these challenges by combining strategic planning with proven infrastructure execution. As organizations compete to operationalize AI at scale, Aptly serves as a trusted partner for designing, building, and optimizing the infrastructure foundations required to support next-generation AI workloads.

The future of AI will not be determined solely by access to compute. It will be determined by the ability to secure power, deploy infrastructure, integrate complex technology ecosystems, and execute at scale. Aptly helps organizations transform AI ambitions into operational reality by delivering the infrastructure foundation on which modern AI depends.