

APTLY TECHNOLOGY

AI INFRASTRUCTURE INTELLIGENCE

Enterprise AI Infrastructure Benchmark Report 2026

How Ready Are Enterprises for Production-Scale AI?

A global readiness framework for production-scale AI infrastructure

Published July 2026

A research report from Aptly Technology

Table of Contents

TL;DR: The Report in Ninety Seconds 3

Executive Summary 4

01 The State of Enterprise AI Infrastructure 6

02 The Enterprise AI Infrastructure Stack 9

03 The Enterprise AI Infrastructure Benchmark 12

04 Key Enterprise Challenges 16

05 What's Next 19

06 Executive Recommendations 21

Frequently Asked Questions 23

Appendix: Research Methodology & Sources 25

TL;DR: The Report in Ninety Seconds

You do not need to read twenty pages to get the point. Here is the whole argument, compressed.

- Enterprise AI spending is no longer a rounding error. Gartner puts worldwide AI spending at $2.59 trillion in 2026, up 47% year over year, with infrastructure absorbing more than 45% of that total.
- Infrastructure spending is growing even faster than software spending. IDC tracked $318 billion in worldwide AI infrastructure spend for 2025 and forecasts $487 billion in 2026, a 53% jump.
- Spending is not the same as readiness. Popular surveys find that 88% of organizations now use AI in at least one business function, yet only 39% report any EBIT impact at the enterprise level, and no single business function has scaled AI agents past roughly 10% adoption.
- Infrastructure has replaced budget as the top blockage. Flexential's 2026 survey of 350+ IT leaders finds infrastructure constraints now outweigh budget concerns for AI expansion, and the share of leaders expecting measurable AI returns within a year dropped from 51% to 36% in a single year.
- Power, not chips, sets the ceiling. IEA data shows AI-focused data center electricity demand grew 50% in 2025 alone, and the capital expenditure of five large technology companies is set to rise a further 75% in 2026.
- The network fight is settled, mostly. Ethernet now carries roughly two-thirds of AI back-end network switch sales, reversing a market InfiniBand held at an 80% share only three years ago, according to Dell'Oro Group.
- This report introduces the Aptly Enterprise AI Infrastructure Readiness Framework, a seven-pillar model scoring your organization from Beginner to AI Ready across compute, networking, storage, data centers, AI operations, security, and scalability.
- Aptly builds and runs the infrastructure layer this report benchmarks, as the only Microsoft-trusted supplier authorized to build and support third-party hyperscale datacenters worldwide, with direct NVIDIA and Supermicro partnerships and 24x7 Global Operations Centers.
- See also: GPU Datacenter Buildout Services and AI Infrastructure Managed Services.

Executive Summary

Why This Report Matters

AI infrastructure sat in the IT budget for two years. This sits on the board agenda now. Worldwide AI spending will reach $2.59 trillion in 2026, a 47% increase over 2025, and Gartner attributes more than 45% of this spending to infrastructure alone: AI-optimized servers, network fabric, processing silicon, and the facilities housing them.

Behind the spending sits a harder question. Is your infrastructure ready to run AI at production scale, or are you funding pilots stuck in the lab? Popular surveys on Industry leaders report 88% of organizations now use AI somewhere in the business, yet only 39% see any EBIT impact at the enterprise level, and no single business function has scaled AI agents past roughly 10% adoption.

The gap between adoption and impact is an infrastructure story as much as a strategy story. Deloitte's Tech Trends 2026 research finds enterprises moving from proof of concept to production find their existing compute, network, and data center footprint was built for a different era of workload. Inference costs have dropped 280-fold over two years, yet total AI spending keeps climbing, because usage is outpacing the cost curve. Some enterprises now see monthly AI compute bills in the tens of millions of dollars, driven largely by the always-on, continuous inference agentic AI systems demand.

This report exists to answer one question directly: how ready is your organization, really? We built the Aptly Enterprise AI Infrastructure Readiness Framework to give CIOs, CTOs, Chief AI Officers, and infrastructure leaders a common language for that answer. Chapter 3 walks through all seven pillars in full and gives you a maturity model you can apply to your own environment this quarter, not next year.

Executive Highlights

Top infrastructure trends

- Inference, not training, now dominates AI compute. Deloitte predicts inference will account for roughly two-thirds of all AI compute in 2026, most of this running on power-intensive data center chips rather than cheaper edge silicon.
- Token consumption is set to double in two years. Deloitte's enterprise AI infrastructure survey finds 30% of organizations already consuming more than 10 billion tokens a month, a share expected to reach 61% by 2028.
- Rack power density has broken past what air cooling handles. Current-generation AI training racks now exceed 40 kW, 2025-class systems pass 100 kW, and vendor roadmaps point toward hundreds of kilowatts per rack within a few years, according to Uptime Institute.

Biggest investment priorities

- Compute capacity: GPU clusters, high-bandwidth memory, and AI-optimized servers, the single largest AI spending category tracked by Gartner.

- High-speed networking and AI fabrics, as Ethernet-based scale-out architectures pull ahead of InfiniBand in cost and vendor diversity.
- Data center modernization: power provisioning, liquid cooling, and rack density upgrades to support next-generation accelerators.
- AI operations and governance: observability, capacity planning, and model risk controls, the layer most enterprises underinvest in relative to compute.

Most common enterprise challenges

- Skills, not silicon, is the top barrier most leaders name. Deloitte's State of AI in the Enterprise finds the AI skills gap is the single biggest reported barrier to integrating AI into existing workflows.
- Confidence in readiness has fallen even as strategy has matured. Forty-two percent of leaders now call their AI strategy highly prepared, yet the same organizations report feeling less prepared on infrastructure, data, risk, and talent than they did a year earlier.
- GPU and memory supply remain tight. High-bandwidth memory capacity is sold out through 2026 at major suppliers, pushing memory toward roughly 30% of hyperscaler AI spending, up from about 8% in 2023, per SemiAnalysis research cited by Data Center Knowledge.

Aptly's key recommendations

- Short term: run a formal infrastructure readiness assessment before your next AI budget cycle, and stop treating GPU procurement as the whole strategy.
- Mid term: modernize data center power and cooling ahead of your next hardware refresh, and build observability into AI operations from day one.
- Long term: design for hybrid and multi-cloud portability now, so sovereign AI, edge inference, and next-generation accelerators do not force a rebuild later.

01: The State of Enterprise AI Infrastructure

Every dollar enterprise spend on AI models eventually turns into a demand for compute, network, and power. This chapter sets the baseline: how big the market is, what is driving demand, and where the money goes.

AI Infrastructure Market Overview

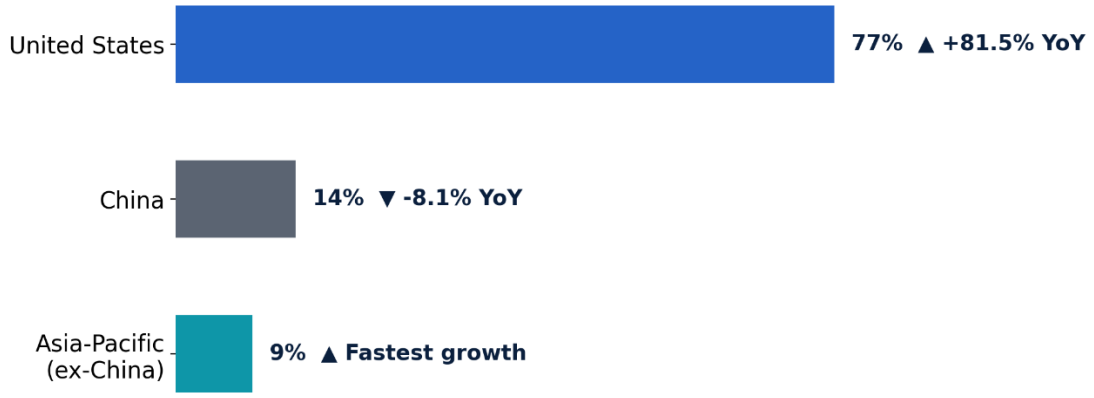
Global AI infrastructure spending reached $318 billion in 2025, more than double the $153 billion recorded in 2024, according to IDC's Worldwide Quarterly AI Infrastructure Tracker. IDC forecasts the market will grow another 53% in 2026, to $487 billion, and expects it to pass $1 trillion before the end of the decade. Growth did not slow through the year. Q4 2025 alone brought in $89.9 billion, a 62% jump from the same quarter a year earlier.



Zoom out to the full AI stack, hardware, software, and services, and the number gets much larger. Gartner forecasts total worldwide AI spending of $2.59 trillion in 2026, a 47% increase year over year, inside a broader IT budget of $6.31 trillion, itself growing 13.5% in the same year. AI now accounts for roughly 41% of total IT spending, up from about 32% a year earlier.

The United States leads by a wide margin. IDC found the US responsible for 77% of global AI infrastructure spending in Q4 2025, growing 81.5% year over year, while China's spending fell 8.1% amid continued export

controls on advanced semiconductors. Mordor Intelligence projects Asia-Pacific as the fastest-growing region through 2031, driven partly by sovereign AI programs across the Gulf states and Southeast Asia.



Hardware still dominates where the money lands. Compute, memory, and networking hardware account for the large majority of AI infrastructure spend today, though software, particularly inference optimization and MLOps tooling, is the fastest-growing line item as enterprises shift focus from raw training capacity to efficient, sustained inference.

What's Driving Demand?

Five forces are pulling enterprise infrastructure spend upward at once, and they compound rather than substitute for each other.

Generative AI and agentic AI

Generative AI moved the industry from experimentation to infrastructure planning. Agentic AI is now moving it from infrastructure planning to infrastructure strain. Capgemini's Rise of agentic AI report finds that while 61% of organizations are still only exploring AI agents and 23% have launched pilots, just 14% have deployed agents at least partially at scale and only 2% have reached full-scale deployment across the enterprise. Agents behave differently from chatbots on infrastructure: instead of a single prompt and response, an agent runs multi-step, tool-using workflows generating continuous, compounding inference load. Deloitte points to agentic AI as the single biggest driver behind enterprise monthly AI bills reaching the tens of millions of dollars.

AI factories

Hyperscalers and a new class of specialized providers are building what the industry now calls AI factories, purpose-built facilities designed around accelerated compute rather than adapted from general-purpose data center designs. The IEA reports these facilities have more than tripled in capacity over the past eighteen months using satellite-based tracking, and the capital expenditure behind them, more than $400 billion across five large technology companies in 2025, is set to rise a further 75% in 2026.

As the only supplier trusted by Microsoft to build and support third-party hyperscale datacenters, Aptly operates directly in this AI factory ecosystem, bringing hyperscale-grade GPU datacenter buildout, AI

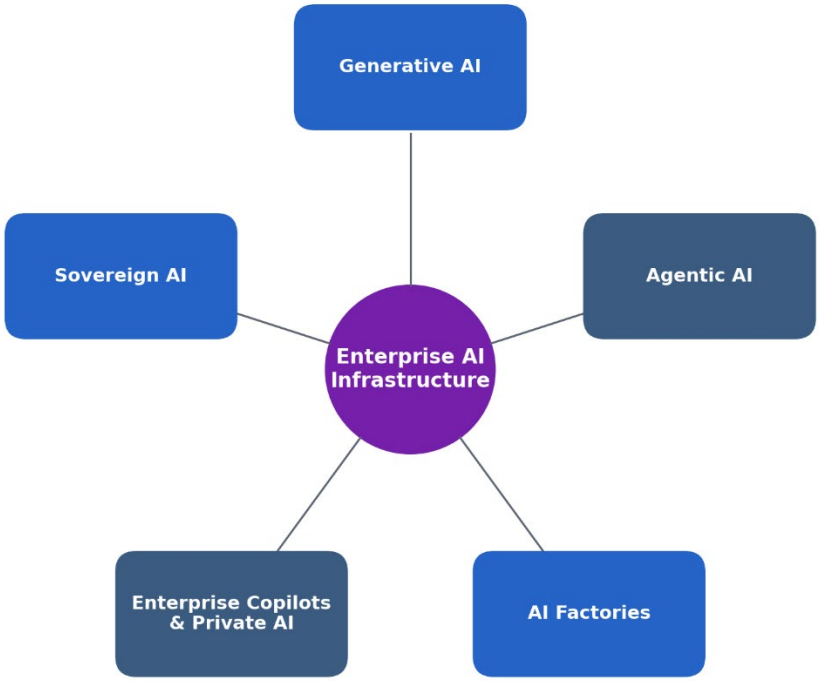
workload optimization, and enterprise AI application development expertise to organizations that need the same accelerated infrastructure patterns inside their own environments.

Enterprise copilots and private AI

Inside the enterprise, demand shows up as copilots embedded in existing software and as private, domain-specific models trained or fine-tuned on proprietary data. Both push infrastructure decisions toward control: enterprises want inference to run close to their data, on infrastructure they audit directly, rather than through a shared, opaque API.

Sovereign AI

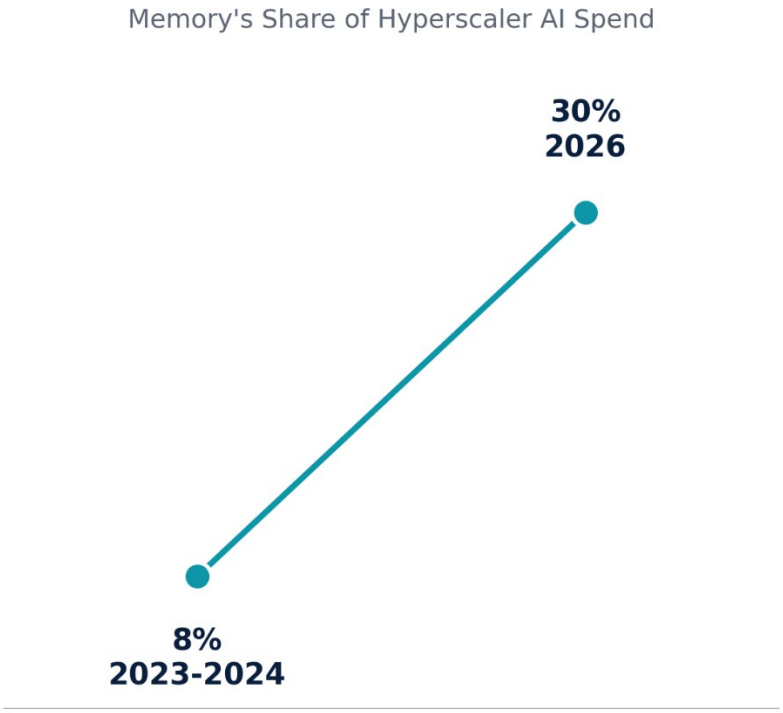
National governments are now infrastructure buyers in their own right. Sovereign AI programs, where a country builds and governs AI compute under its own laws and data residency rules, have become a material demand driver rather than a policy footnote. Deloitte's enterprise AI survey frames sovereign AI as a matter of strategic independence, not only ownership, and this is reshaping where enterprises with multinational operations choose to locate inference workloads.



Market Trends

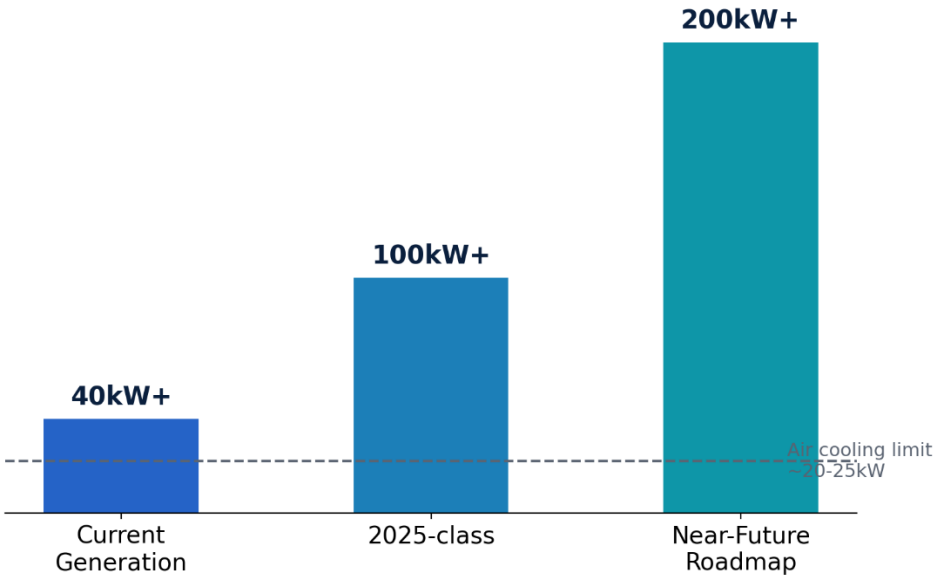
GPU demand and the memory bottleneck

GPU demand still outpaces supply, but the binding constraint has shifted from the processor itself to the memory around it. High-bandwidth memory production for 2026 was sold out at major suppliers before the year began, and producing a single bit of HBM consumes roughly three to four times the wafer capacity of standard DRAM. Memory is on track to represent close to 30% of hyperscaler AI spending in 2026, up from about 8% in 2023 and 2024, according to SemiAnalysis research reported by Data Center Knowledge. If you are planning hardware procurement, budget for memory lead times, not only GPU lead times.



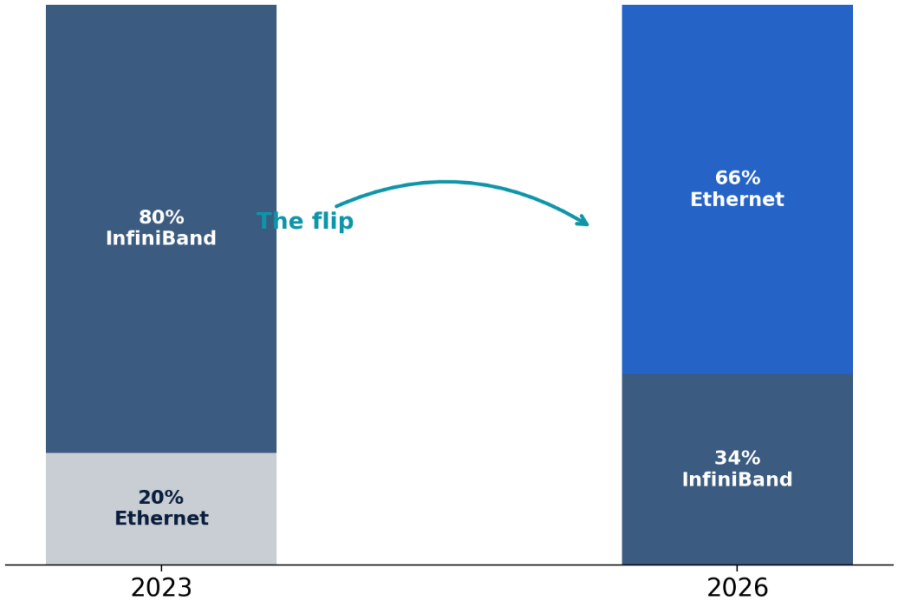
AI-ready data centers

Rack power density has moved from a design detail to a site-selection constraint. Current-generation AI training racks exceed 40 kW, 2025-class deployments pass 100 kW, and product roadmaps point toward hundreds of kilowatts per rack within a few years, according to Uptime Institute. Air cooling stops working economically above roughly 20 to 25 kW per rack, which is why direct liquid cooling has gone from a specialist option to a default assumption for any new AI cluster.



High-speed networking

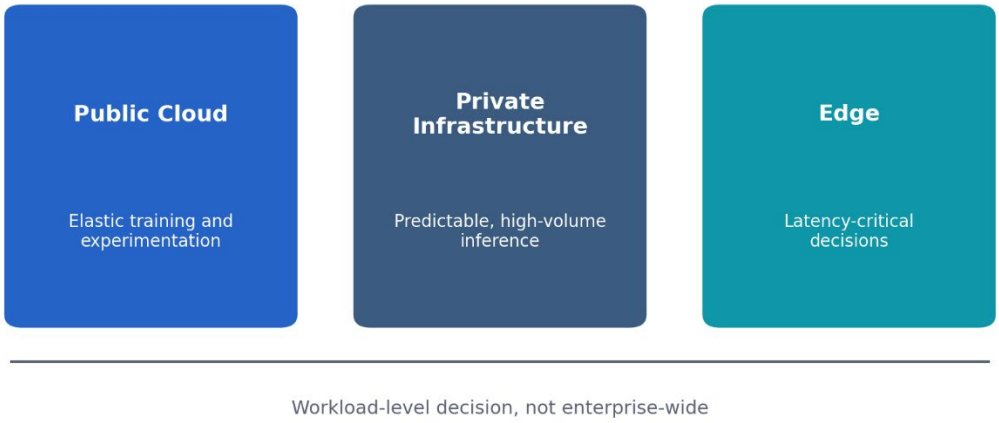
The network architecture debate has largely settled. Ethernet now carries roughly two-thirds of data center switch sales in AI back-end networks, a full reversal from 2023, when InfiniBand held about 80% of this same market, according to Dell'Oro Group.



InfiniBand has not disappeared, and remains the choice for the largest frontier training runs, but Ethernet's cost structure, vendor diversity, and the Ultra Ethernet Consortium's 1.0 specification have made Ethernet the default for enterprise-scale deployments.

Cloud AI and hybrid deployment

No enterprise runs AI on a single environment anymore. Deloitte's compute strategy research describes a three-tier model taking hold across the enterprises studied: public cloud for elastic training and experimentation, private infrastructure for predictable, high-volume inference, and edge compute for time-critical decisions in manufacturing and autonomous systems. The choice is workload by workload, not enterprise-wide.



Sustainable infrastructure

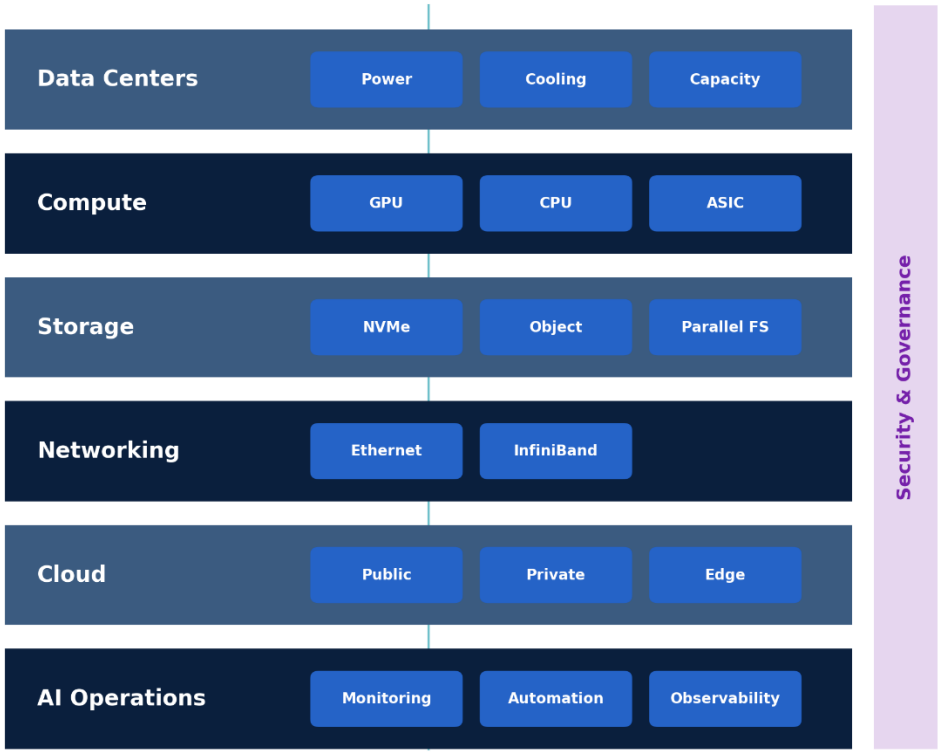
Power has become the hard ceiling on AI ambition. Global data center electricity demand grew 17% in 2025, and demand from AI-focused facilities alone grew 50% in the same year, according to the IEA. The agency projects data center electricity consumption will double by 2030, with AI-focused facilities roughly tripling their power draw in the same window. Help Net Security's coverage of NTT Data's 2026 outlook states the point plainly: access to electricity, not chip availability, is becoming the deciding factor in where new AI capacity gets built, since large power transformers and grid connections carry lead times measured in years.

WHY IT MATTERS

Enterprises treating AI infrastructure as a procurement exercise, buy GPUs, sign a cloud contract, declare victory, consistently underperform organizations treating infrastructure as a systems design problem spanning compute, network, power, and operations together. Chapter 3 gives you a structured way to tell which camp you are in.

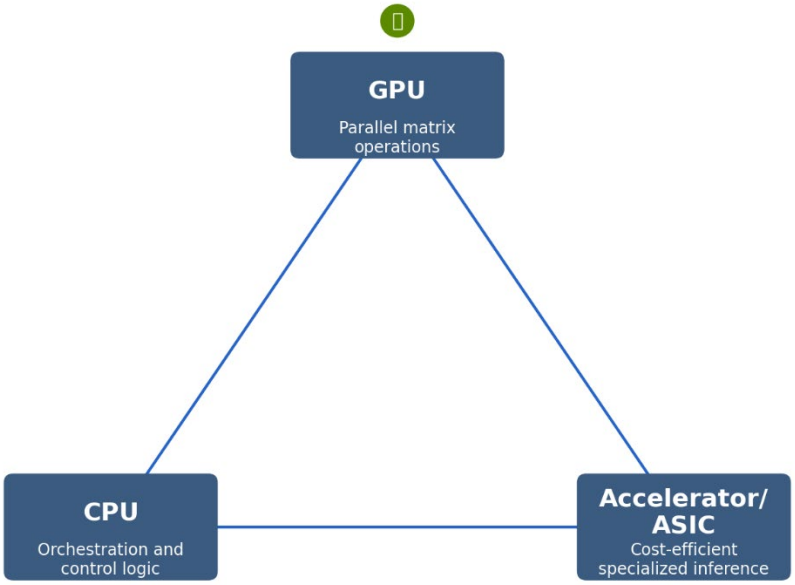
02: The Enterprise AI Infrastructure Stack

Production AI infrastructure has seven layers. Each one becomes the bottleneck stalling a deployment, regardless of how strong the other six are. This chapter walks through them in the order most enterprises encounter the constraints.



Compute

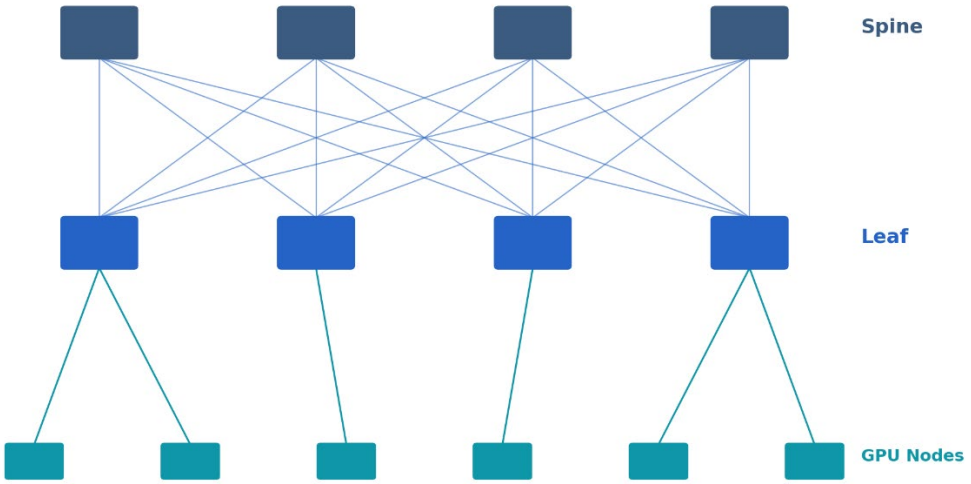
Compute is the layer everyone budgets for first, and the one most enterprises still get wrong. GPUs handle the parallel matrix operations behind model training and inference. CPUs orchestrate everything around them: data loading, preprocessing, and the control logic agentic workflows depend on. Purpose-built accelerators, ASICs and FPGAs, are gaining ground for specific inference workloads where cost per token matters more than raw flexibility.



NVIDIA supplies an estimated 80% or more of AI training GPUs in production today, though AMD's MI350 series and custom silicon from hyperscalers are gaining share as enterprises look to diversify supply and reduce total cost of ownership. The practical question for you is rarely which chip. The real question is utilization. Idle GPU capacity is one of the most common, and most expensive, findings in any infrastructure assessment.

Networking

Training and inference clusters live or die on the network fabric connecting nodes, not on the chips themselves. Ethernet has overtaken InfiniBand as the majority technology for AI back-end networks, carrying roughly two-thirds of switch sales as of early 2026, per Dell'Oro Group, while InfiniBand still leads in the largest frontier-scale training runs where sub-two-microsecond latency justifies the cost. Topology choices, leaf-spine, dragonfly, fat-tree, determine how well a cluster scales past a few hundred nodes without new bottlenecks appearing at the spine.



Storage

AI workloads read and write data in patterns conventional enterprise storage was never built for: massive parallel reads during training, high-throughput checkpointing, and low-latency retrieval for inference and retrieval-augmented generation. Three storage types now sit side by side in most AI environments: NVMe all-flash for hot training data, object storage for model artifacts and data lakes, and parallel file systems for distributed training jobs needing every GPU fed at once.

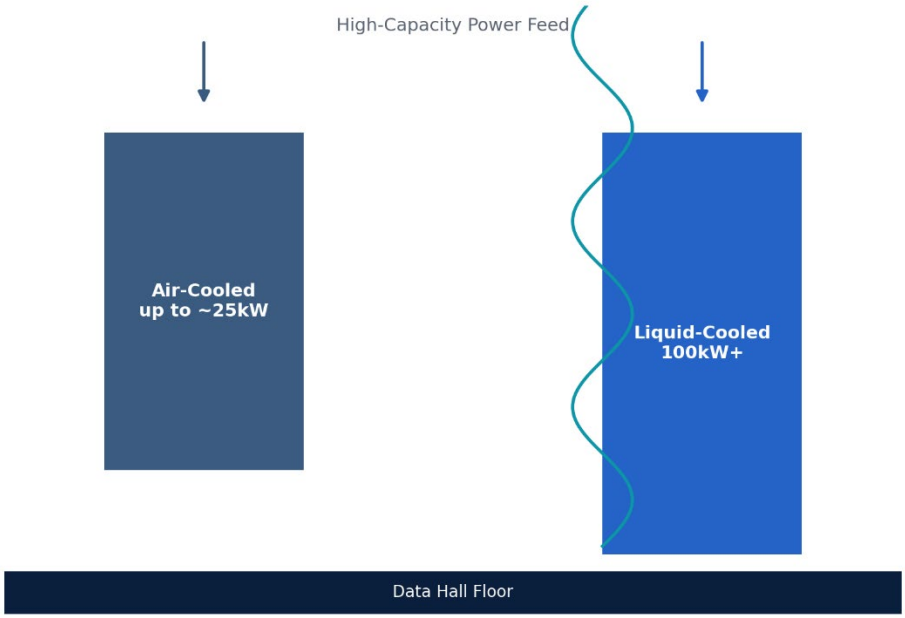


Throughput and IOPS, not raw capacity, decide whether storage becomes the constraint leaving expensive GPUs waiting.

Data Centers

Power and cooling now define what a data center runs, more than floor space or rack count. Current AI training racks exceed 40 kW, with 2025-generation deployments passing 100 kW, according to Uptime Institute. Air cooling remains viable up to roughly 20 to 25 kW per rack. Past this threshold, direct liquid cooling, cold plates

or immersion, becomes a requirement rather than an upgrade. As of early 2024, only 22% of data center operators had deployed direct liquid cooling in any form, and most of those ran it on a small share of total racks, according to Uptime Institute survey data reported by Network World.



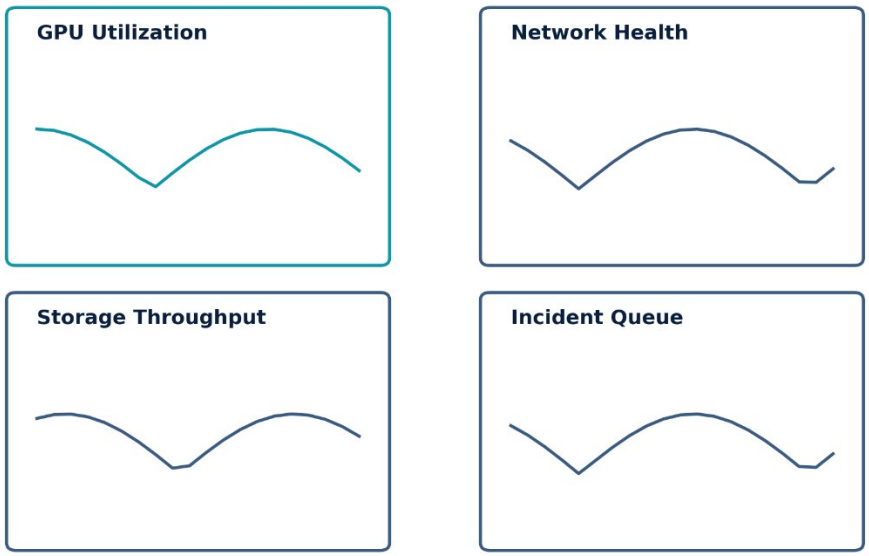
If your facility was built before 2023, assume a retrofit plan is needed.

Cloud

Public cloud, private infrastructure, and edge compute now function as three distinct tools rather than three competing strategies. Deloitte's research describes the pattern taking hold across enterprises: public cloud absorbs elastic training and experimentation, private infrastructure runs predictable, high-volume inference at controlled cost, and edge handles workloads where latency measured in milliseconds decides whether a use case works at all, manufacturing quality control and autonomous systems among them. The decision belongs at the workload level, not the enterprise level.

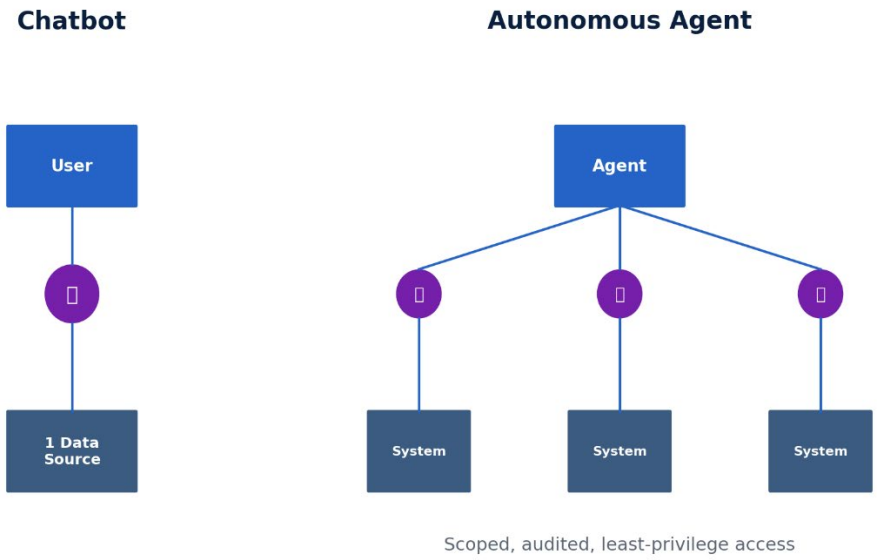
AI Operations

This is the layer most enterprises underbuild. Monitoring GPU utilization, automating capacity planning, and maintaining observability across a heterogeneous compute estate take dedicated tooling and dedicated people, not an extension of existing IT operations. Nutanix's 2026 Enterprise Cloud Index found 79% of IT executives already encounter AI applications or agents deployed by non-IT functions without central visibility, and 87% expect containerization to expand over the next three years specifically to regain control over AI workload portability and consistency.



Security

AI infrastructure inherits every conventional security requirement and adds new ones: model risk, data lineage, and agent permissioning chief among them. Zero Trust architecture, strong identity controls, and clear AI governance policy are no longer separable from infrastructure design, because an AI agent with broad system access is a materially different risk than a chatbot answering questions in a browser tab.



Only one in five enterprises reports a mature governance model for autonomous AI agents today, according to Deloitte, even as agentic deployment accelerates.

The stack at a glance

Layer	What It Does	Benchmark Signal to Watch	Aptly's Role
Compute	Runs training and inference workloads across GPUs, CPUs, and accelerators	GPU utilization rate, not only GPU count	GPU cluster design, deployment, and NVIDIA-partnered procurement
Networking	Connects nodes with low-latency, high-bandwidth fabric	Fabric bottlenecks at scale, not peak port speed	InfiniBand and Ethernet AI fabric design and buildout
Storage	Feeds and checkpoints training and inference pipelines	Sustained throughput under concurrent load	High-performance storage architecture for AI pipelines
Data Centers	Provides power, cooling, and physical capacity	kW per rack versus site cooling capacity	Hyperscale datacenter buildout and liquid cooling retrofits
Cloud	Distributes workloads across public, private, and edge	Workload-to-environment fit, not cloud spend alone	Hybrid and multi-cloud AI infrastructure strategy
AI Operations	Monitors, automates, and plans capacity	Observability coverage across the full AI estate	24x7 Global Operations Center monitoring and support
Security	Governs identity, access, and model risk	Maturity of agentic AI governance controls	Zero Trust implementation and compliance alignment

03: The Enterprise AI Infrastructure Benchmark

This is the signature chapter of the report. This chapter introduces the Aptly Enterprise AI Infrastructure Readiness Framework, a diagnostic model built from the market data in Chapters 1 and 2 plus Aptly's own delivery experience across GPU datacenter buildouts, and gives you a way to score your own organization against the model.

Introducing Aptly's Framework

Most AI infrastructure assessments stop at compute. They count GPUs, check a utilization dashboard, and call the job done. This approach misses the layers determining whether AI reaches production: the network fabric between the GPUs, the storage feeding them, the power and cooling keeping them alive, the operations discipline running them day to day, the governance controlling them, and the architecture letting them scale past a single site.

The Aptly framework scores readiness across seven pillars. Each pillar gets evaluated independently, because a strong compute score cannot compensate for a data center unable to deliver the power a new GPU cluster needs, or a security program with no answer for autonomous agent access. Score each pillar honestly, plot the results, and the gaps holding your AI program back become visible in a single view instead of staying buried inside separate team reports.



Pillar 1: Compute Readiness

Compute readiness asks whether your organization has the right processing capacity, in the right place, used well, not simply whether you own enough GPUs.

- GPU capacity: current fleet size against forecast training and inference demand for the next twelve months
- Scalability: how quickly you add compute capacity when a workload outgrows its current footprint
- Resource utilization: the share of provisioned GPU capacity running workloads rather than sitting idle

Pillar 2: Networking Readiness

Networking readiness measures whether your fabric moves data between nodes fast enough to keep expensive compute from waiting.

- Network speed: current fabric bandwidth against the requirements of your largest active or planned training and inference workloads
- AI fabrics: whether you have purpose-built AI networking (InfiniBand, AI-optimized Ethernet) rather than repurposed general-purpose network infrastructure

- Bottlenecks: known chokepoints at the spine or leaf layer appearing only once a cluster scales past pilot size

Pillar 3: Storage Readiness

Storage readiness measures whether data reaches your compute fast enough, and whether your storage layer scales with your model and dataset growth.

- Throughput: sustained read and write performance under concurrent, multi-job load, not single-stream benchmark numbers
- Scalability: capacity headroom against projected data and checkpoint growth over the next planning cycle
- Data availability: how reliably training and inference pipelines reach the data they need, when they need it

Pillar 4: Data Center Readiness

Data center readiness measures the physical foundation: whether your facilities deliver the power and cooling next-generation accelerators require.

- Rack density: current supported kW per rack against the density of the accelerators you plan to deploy
- Power: available capacity and contracted headroom for AI workload growth, not only current utilization
- Cooling: air cooling limits versus liquid cooling readiness for racks above roughly 20 to 25 kW
- Sustainability: progress on power usage effectiveness and renewable sourcing commitments tied to AI growth

Pillar 5: AI Operations

AI operations readiness measures whether you run AI infrastructure reliably day to day, not only stand it up once.

- Monitoring: real-time visibility into GPU utilization, network performance, and workload health across the full estate
- Automation: how much of capacity planning, scaling, and routine maintenance runs without manual intervention
- Observability: end-to-end tracing across the AI pipeline, from data ingestion through model output
- Day-2 operations: incident response, patching, and lifecycle management maturity once infrastructure is in production

Pillar 6: Security & Governance

Security and governance readiness measures whether your controls have caught up to what AI systems, especially autonomous agents, now do.

- Identity and access: how tightly AI agent permissions are scoped and audited against least-privilege principles

- Compliance: alignment with data residency, sector regulation, and emerging AI-specific requirements such as the EU AI Act
- AI governance: model risk management, data classification, and a defined escalation path for autonomous agent decisions

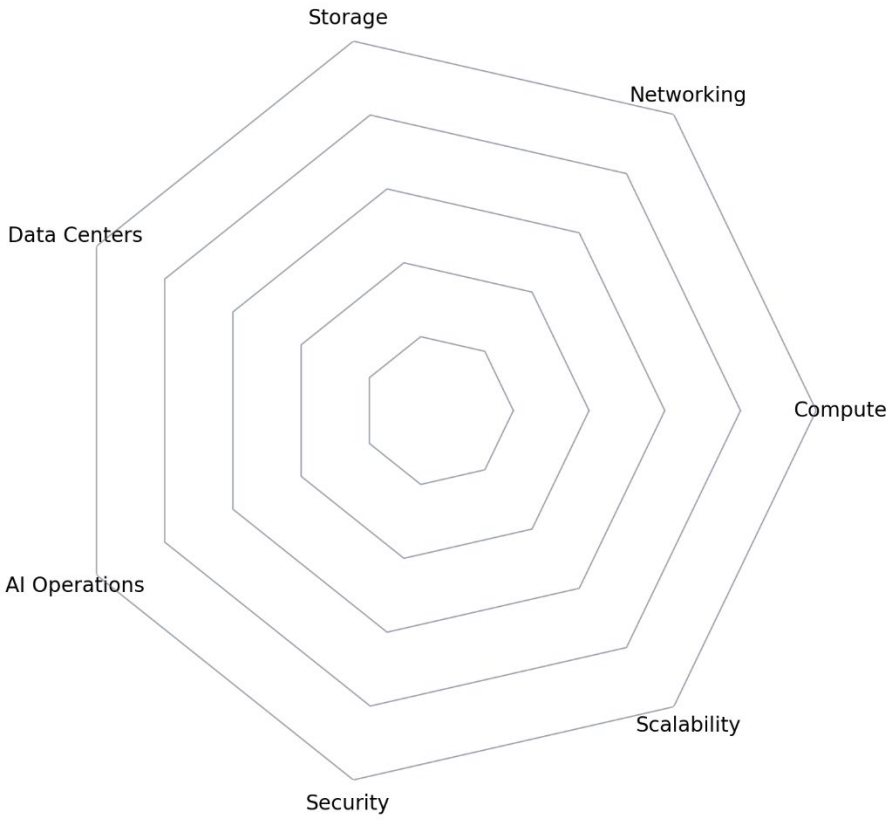
Pillar 7: Scalability

Scalability readiness measures whether your architecture grows from a single pilot to an enterprise-wide AI factory without a rebuild.

- Hybrid deployment: ability to run the same workload across on-premises, cloud, and edge without redesigning it each time
- Multi-cloud: portability across cloud providers, avoiding dependence on a single vendor's AI stack
- Global deployment: readiness to run compliant, low-latency AI workloads across multiple regions and jurisdictions at once

The Readiness Scorecard

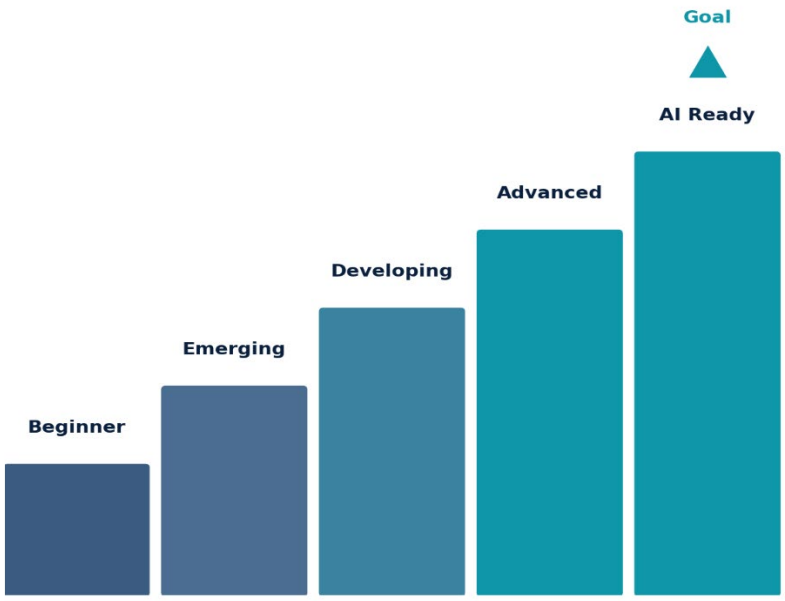
Score each pillar against the five-level maturity model below, then look at the pattern across all seven scores rather than the average.



Plot your score on each axis, then connect the points.

A single weak pillar, most often data center readiness or AI operations, caps what the other six deliver.

Level	What It Looks Like	Typical Infrastructure Profile	Next Step
1. Beginner	AI runs in isolated pilots with no production path	Shared, general-purpose infrastructure repurposed for AI experiments	Run a formal infrastructure readiness assessment before the next AI budget cycle
2. Emerging	A few use cases are live, but each was built one-off	Dedicated GPU capacity exists, but network, storage, and power were not planned around it	Standardize the stack, stop custom-building each new use case
3. Developing	Multiple business functions run AI in production	Purpose-built compute and networking, but operations and governance lag behind	Build unified AI operations and observability across all functions
4. Advanced	AI is embedded across most core workflows	AI-ready data centers, high-speed fabric, and mature day-2 operations	Formalize governance and extend the model to hybrid and multi-cloud
5. AI Ready	Infrastructure scales ahead of AI demand, not behind it	Full-stack readiness across all seven pillars, with sustainability and governance built in	Prepare for next-generation workloads: agentic AI at scale, sovereign compute, edge inference



Two organizations, two starting points

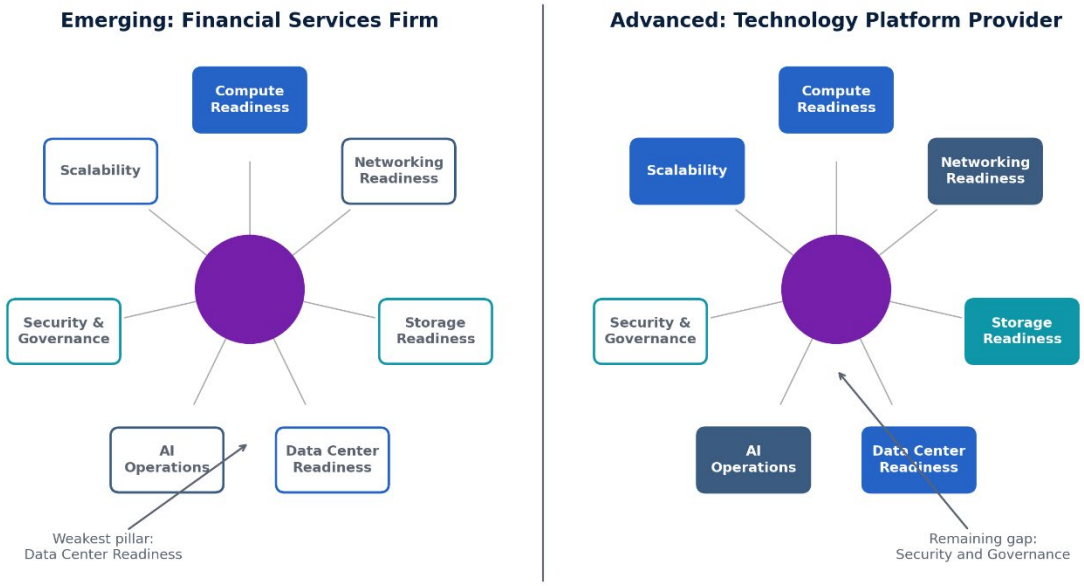
Two illustrative profiles, drawn from patterns Aptly sees repeatedly across enterprise engagements, show how differently the same framework plays out in practice.

PROFILE: EMERGING

A mid-size financial services firm sits at Emerging. It runs three AI copilots in production, each on infrastructure procured separately for the project. Compute readiness scores well: GPUs are modern and well utilized. Data center readiness scores poorly: the facility hosting the newest cluster has no path to the liquid cooling the next hardware refresh will need. The firm's constraint is not ambition. The constraint is a data center built for a workload profile no longer in use.

PROFILE: ADVANCED

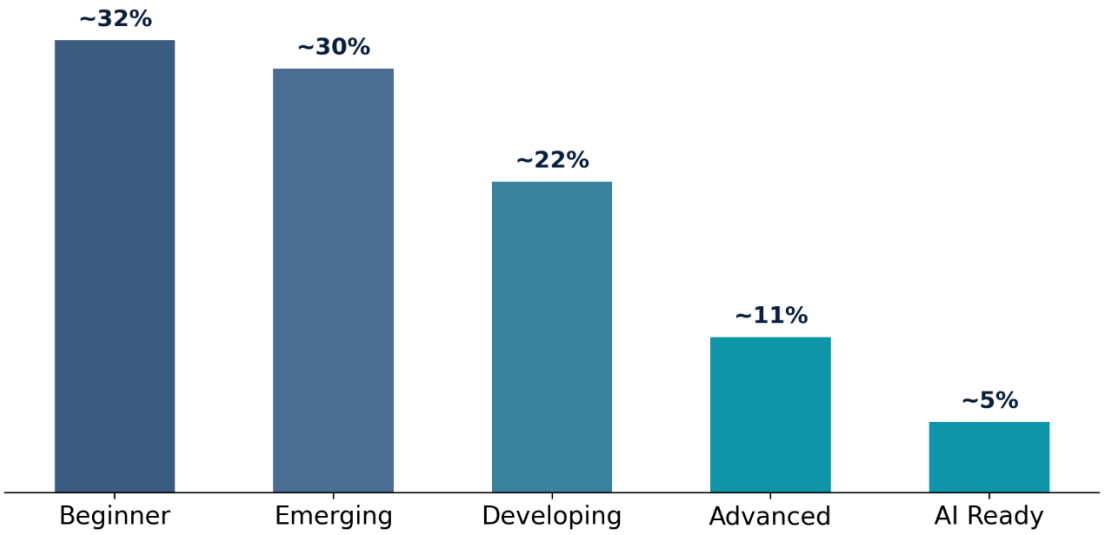
A technology platform provider sits at Advanced. AI runs across product, support, and internal operations, all on a common, purpose-built stack. Compute, networking, and AI operations all score strongly. Security and governance is the one pillar still catching up: agent permissioning was designed for last year's chatbot-style tools, not this year's autonomous, multi-step agents. The provider's next move is governance, not more infrastructure.



Where Most Enterprises Stand Today

Put the market data from Chapter 1 next to this framework, and a pattern emerges. Gartner's AI maturity survey finds that most organizations are still in the early and intermediate stages of AI maturity, with only a

minority sustaining meaningful, enterprise-level impact from AI over multiple years, which maps closely to Beginner and Emerging on this scorecard.

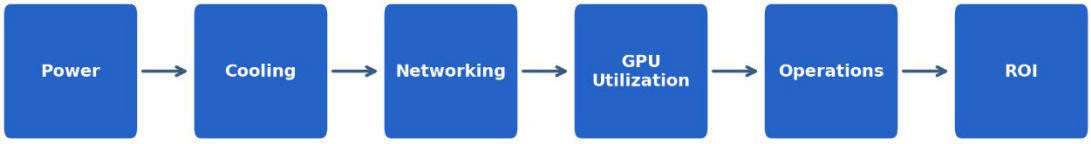


Illustrative distribution, synthesized from McKinsey and Deloitte findings, not a standalone Aptly survey

The roughly 6% of organizations Gartner identifies as AI high performers, seeing more than 5% EBIT impact from AI, look far closer to Advanced or AI Ready: they redesign workflows around AI rather than bolting AI onto existing ones, and Deloitte's enterprise research finds this same group reports feeling less prepared on infrastructure and talent than their strategy alone would suggest, evidence even leading organizations know their infrastructure has not fully caught up to their ambition

04: Key Enterprise Challenges

Ask ten infrastructure leaders what is blocking their AI program and you will hear three categories of answer, in roughly the same order every time: the physical infrastructure is not ready, the operating model cannot run what gets built, and the business case will not hold still long enough to get funded twice.



Infrastructure Challenges

GPU and semiconductor supply

GPU allocation now runs years ahead. Hyperscalers placed multi-billion-dollar forward orders for current-generation accelerators through 2027, absorbing most available supply and leaving mid-market enterprises competing for what remains. The deeper constraint sits one layer down, in memory: high-bandwidth memory capacity was sold out at major suppliers before 2026 began, and meaningful new capacity is not expected before late 2026 at the earliest. If your procurement plan assumes GPU lead times only, revisit it.

Power limitations

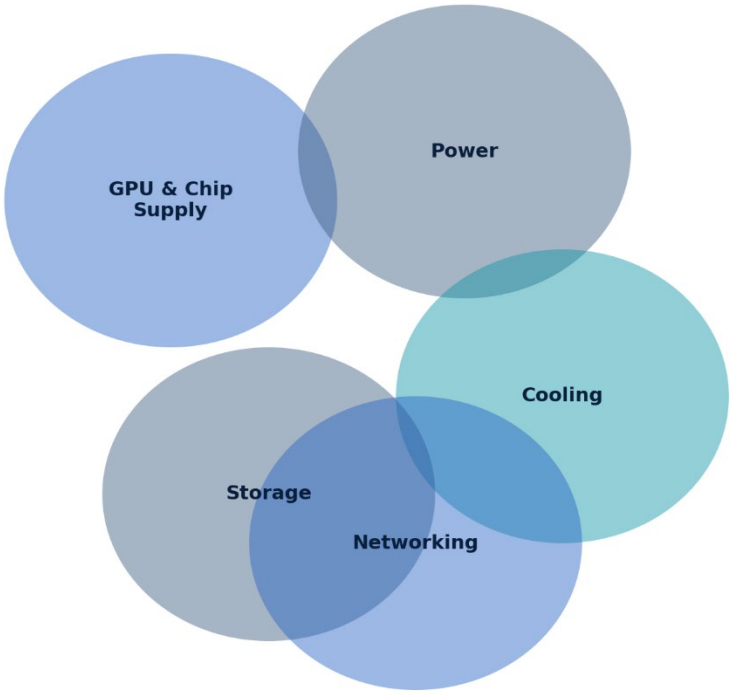
Power has overtaken chip availability as the binding constraint on where and how fast AI capacity grows. Data center electricity demand grew 17% globally in 2025, with AI-focused facilities growing 50%, according to the IEA. Large power transformers and grid interconnections carry lead times measured in years, not months, which means power planning now has to start before the hardware order, not after it.

Cooling constraints

Air cooling tops out around 20 to 25 kW per rack. Current AI training racks already exceed 40 kW, and next-generation systems are heading toward hundreds of kilowatts, according to Uptime Institute. Enterprises retrofitting older facilities for liquid cooling face real engineering constraints: piping, floor loading, and leak containment were rarely part of the original design.

Storage bottlenecks and networking complexity

Storage systems sized for conventional enterprise workloads routinely buckle under the sustained, parallel read patterns of AI training. Networking complexity compounds the problem: as clusters scale past a few hundred nodes, bottlenecks appear at the spine layer never visible in smaller pilot deployments, and diagnosing them takes specialized expertise most enterprise network teams have not yet built.



Operational Challenges

Skills shortages

The single most commonly cited barrier to AI integration is not technology. The barrier is people. Deloitte's State of AI in the Enterprise research identifies the AI skills gap as the biggest reported barrier to integrating AI into existing workflows, ahead of budget, data quality, or technology maturity. AI infrastructure architects, AIOps specialists, and engineers who understand both GPU clusters and enterprise networking remain scarce, and most organizations are competing with hyperscalers for the same small talent pool.

Immature AI operations

Many enterprises stand up an AI cluster. Far fewer run it reliably at 2 a.m. when a node fails mid-training-run. Fragmented tooling, siloed infrastructure and data science teams, and a lack of unified observability across the AI estate turn routine incidents into extended outages. Nutanix found 79% of IT executives already encounter AI applications deployed by non-IT functions with no central visibility, a governance gap becoming an operational one the moment something breaks.

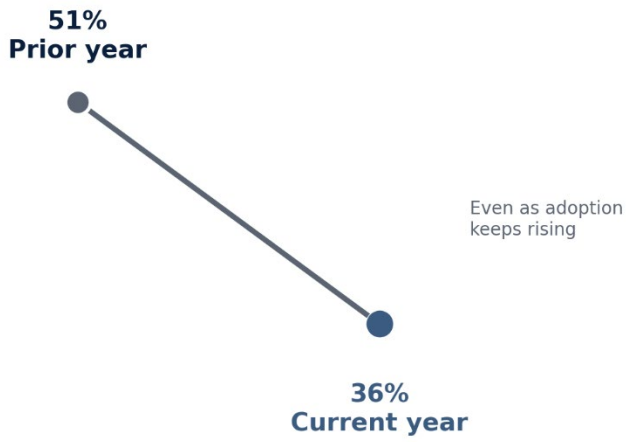
Resource utilization

Idle GPU capacity is one of the most consistent findings across infrastructure assessments. Enterprises overprovision to avoid queueing, then find a large share of expensive, power-hungry compute sits unused for long stretches. Without real-time utilization monitoring, the problem stays invisible until the next capacity planning cycle forces a reckoning.

Business Challenges

ROI uncertainty

Confidence in AI returns is heading the wrong direction. The share of enterprise leaders expecting measurable AI financial returns within one year dropped from 51% to 36% in a single year, according to Flexential's 2026 State of AI Infrastructure Report, even as the share already seeing returns held flat at 20%.



Separate research from MIT's NANDA initiative, reported by Fortune found roughly 5% of generative AI pilots achieve measurable revenue impact, based on interviews with 150 leaders and analysis of 300 public AI deployments. The pattern across both studies points the same direction: infrastructure and integration discipline, not model quality, decide which pilots become production systems.

Solution:

Enterprises should institutionalize an AI investment governance layer that ties every AI initiative to explicit business KPIs, with stage-gates that require measurable impact before scaling beyond pilot. This can include building a standardized AI value scorecard per use case scenario, running controlled A/B experiments, and decommissioning low-yield pilots so infrastructure and talent are redirected to the few workloads that prove repeatable returns.

Cost management

AI costs do not behave like conventional IT costs. Inference is continuous, not periodic, and agentic workloads compound token consumption in ways traditional cloud budgeting was never built to forecast. Some

enterprises now see monthly AI compute bills reach the tens of millions of dollars, and without workload-level cost attribution, finance teams cannot tell which use cases are earning their keep.

Solution:

Implement AI-specific cost observability and FinOps practices that attribute spend down to the model, workload, team, and agent level, so that the finance team can see cost per request, per token, and per business outcome in near real time. Optimize on four core layers – model (smaller/quantized where possible), runtime (batching & efficient servers), infrastructure, and platform (budgets, alerts, reviews, etc.) – to stabilize AI compute bills.

Governance

Only one in five enterprises reports a mature governance model for autonomous AI agents, even as agentic deployment accelerates across the business, according to Deloitte. Governance built for a chatbot answering questions does not transfer cleanly to an agent taking actions inside production systems.

Solution:

Design and adopt an agentic AI governance framework that extends traditional AI and API governance into action-taking agents, covering roles, policies, guardrails, and human-in-the-loop controls for every class of autonomous behavior. This should include explicit scopes for agent-authority, tiered risk classifications, change-controlled tool access, logging and audit trails, and cross-functional oversight.

Vendor lock-in

Enterprises standardizing early on a single cloud provider's AI stack now find portability costs more to build after the fact than it would have cost to design in from the start. Multi-cloud and hybrid architecture is no longer a resilience nice-to-have. This is what keeps a GPU shortage at one provider from becoming your outage.

Solution:

Reduce lock-in risk by deliberately designing a modular, vendor-agnostic AI architecture: separate data, models, orchestration, and applications; favor open standards; and use portable building blocks such as containers, Kubernetes and infrastructure-as-code across clouds.

WHY IT MATTERS

None of these challenges sit in isolation. A skills shortage slows the response to a power constraint. A power constraint delays the hardware refresh needed to fix a utilization problem. Treat infrastructure, operations, and business readiness as one system, since the constraints behave this way in practice.

05: What's Next

The infrastructure decisions you make in the next twelve months will either absorb the next wave of AI workloads or force a rebuild in eighteen months. Here is what the next wave looks like.



Future Trends in AI Infrastructure

AI factories become the default design pattern

Purpose-built AI facilities, designed around accelerated compute from the foundation up rather than retrofitted from general-purpose data centers, are becoming the standard for any new large-scale deployment. The IEA already tracks these facilities tripling in capacity over eighteen months. Expect the enterprise version of this pattern, smaller AI-native facilities or purpose-built zones inside existing sites, to spread beyond hyperscalers over the next two to three years.

Agentic AI reshapes infrastructure demand

Agentic AI does not only add workload volume. Agentic AI changes the shape of demand: continuous rather than bursty, multi-step rather than single-request, and increasingly autonomous in what it accesses and acts on. Gartner forecasts AI agent software spending to jump from $206.5 billion in 2026 to $376.3 billion in 2027. Plan infrastructure for sustained, always-on inference load, not the batch-oriented patterns shaping most enterprise data centers.

GPU-as-a-Service matures

As GPU and memory supply stay tight, specialized GPU cloud providers and flexible consumption models are giving enterprises a way to access modern accelerators without committing to multi-year hardware ownership. Expect this model to keep growing fastest among mid-market enterprises priced out of direct hyperscaler allocation.

Sustainable AI infrastructure becomes a procurement requirement

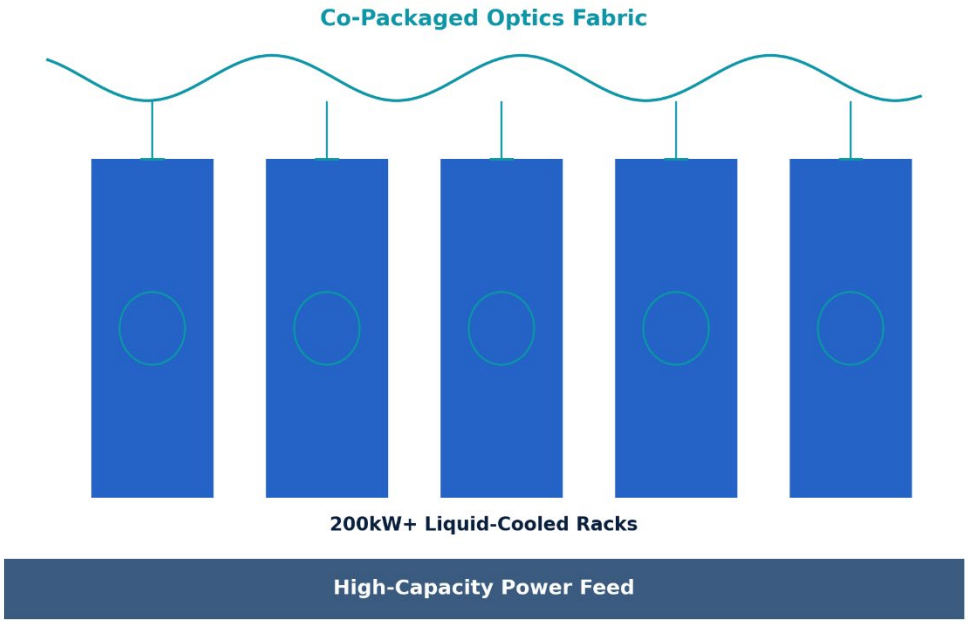
Sustainability disclosure is moving from a corporate responsibility exercise to an enterprise procurement criterion. Renewable power sourcing, water usage, and power usage effectiveness targets increasingly show up in AI infrastructure vendor selection, not only in annual sustainability reports.

Liquid cooling goes from option to default

Vendor roadmaps already point toward 200 kW-plus racks within a few years, according to Uptime Institute. Direct liquid cooling will keep spreading from AI-specific clusters into mainstream enterprise data center design, driven by the same density pressure making air cooling untenable for AI in the first place.

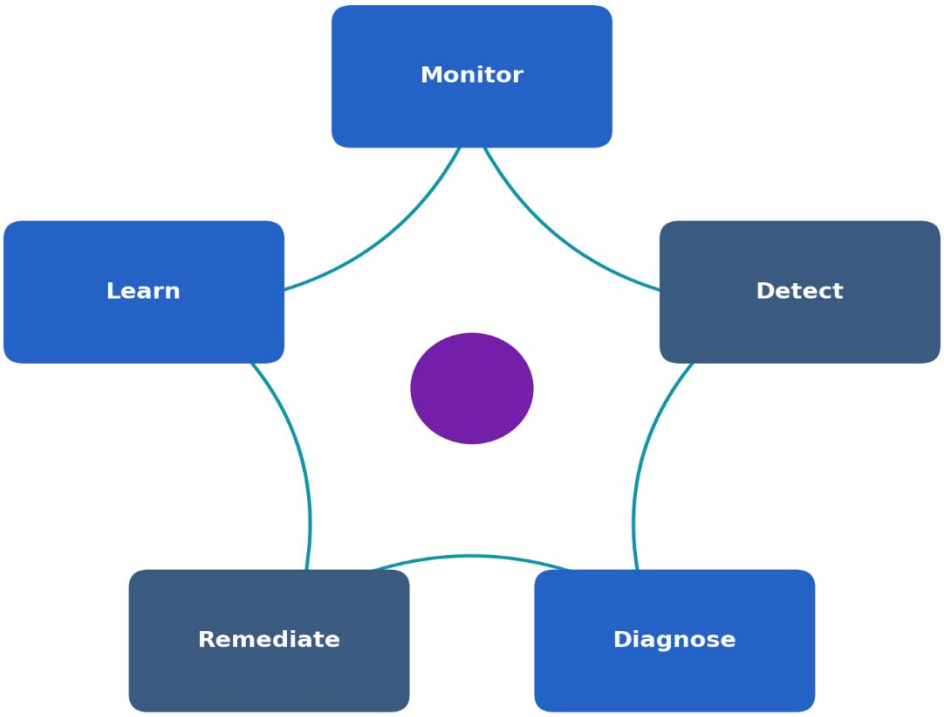
Optical networking enters the mainstream

As switch speeds move from 800 Gbps toward 1.6 Tbps, co-packaged optics, integrating optical interfaces directly with switch silicon, are beginning to ramp in volume across both Ethernet and InfiniBand platforms. Expect this to ease the power and signal integrity limits pluggable optics are starting to hit at the highest bandwidth tiers.



AI infrastructure automation and self-healing operations

AIOps, AI systems monitoring and managing the infrastructure running other AI systems, is moving from a specialist tool to a baseline requirement. Expect automated capacity planning, predictive maintenance, and self-healing failover to become standard features of AI operations platforms rather than premium add-ons, closing part of the skills gap Chapter 4 identifies by reducing how much specialized expertise each incident requires.



Edge AI expands beyond pilots

Manufacturing, autonomous systems, and other latency-sensitive use cases are pushing inference out of centralized data centers and toward the edge, where split-second response times determine whether a use case works at all. Expect edge deployment to grow fastest in industries where a network round trip to a centralized data center is simply too slow to matter.

How These Trends Reshape Architecture Decisions

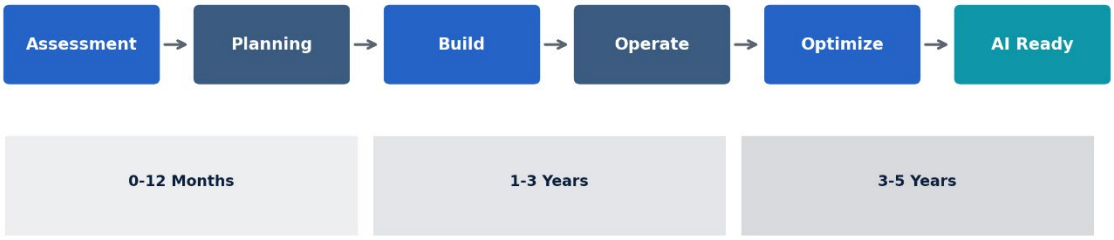
None of these trends point toward a single, fixed architecture. They point toward optionality. Sovereign compute requirements, hybrid workload placement, and rapidly evolving accelerator generations all argue for infrastructure designed to be modular and portable rather than optimized for one moment in time.

WHY IT MATTERS

Do not wait for these trends to fully mature before you plan for them. Liquid cooling, agentic infrastructure demand, and optical networking are already showing up in vendor roadmaps for the next twelve to eighteen months. Enterprises planning a data center refresh around next year's density requirements, not this year's, avoid paying twice.

06: Executive Recommendations

A readiness framework only earns its place in the report if it leads to action. Here is what to do with your score, organized by how soon each step needs to start.



Short Term (0–12 Months)

- Run a formal infrastructure readiness assessment against all seven pillars in Chapter 3 before your next AI budget cycle. Score honestly. The lowest-scoring pillar, not the average, tells you where to start.
- Map your active and planned AI workloads against the infrastructure gaps the assessment surfaces. Most enterprises find the gap is not compute. The gap is data center power, storage throughput, or AI operations coverage.
- Prioritize the AI workloads with the clearest path to measurable value, copilots embedded in existing workflows and domain-specific models tied to a defined business process, and align your first infrastructure upgrades around them. Visible wins fund the next phase.
- Put utilization monitoring in place before you buy more GPU capacity. Idle compute you already own is the fastest infrastructure win available to most enterprises.

Mid Term (1–3 Years)

- Scale compute and networking together, not compute alone. A larger GPU cluster without a matching network fabric upgrade simply moves your bottleneck from the chip to the switch.
- Modernize data centers ahead of your next hardware refresh. Plan power and cooling capacity for the rack densities your next-generation accelerators will need, not the densities your current fleet uses today.
- Build unified observability across the full AI estate, compute, network, storage, and applications, so a single incident does not require four separate teams to diagnose.
- Formalize AI governance for autonomous agents specifically. Chatbot-era access controls do not scope correctly for agents taking actions inside production systems.

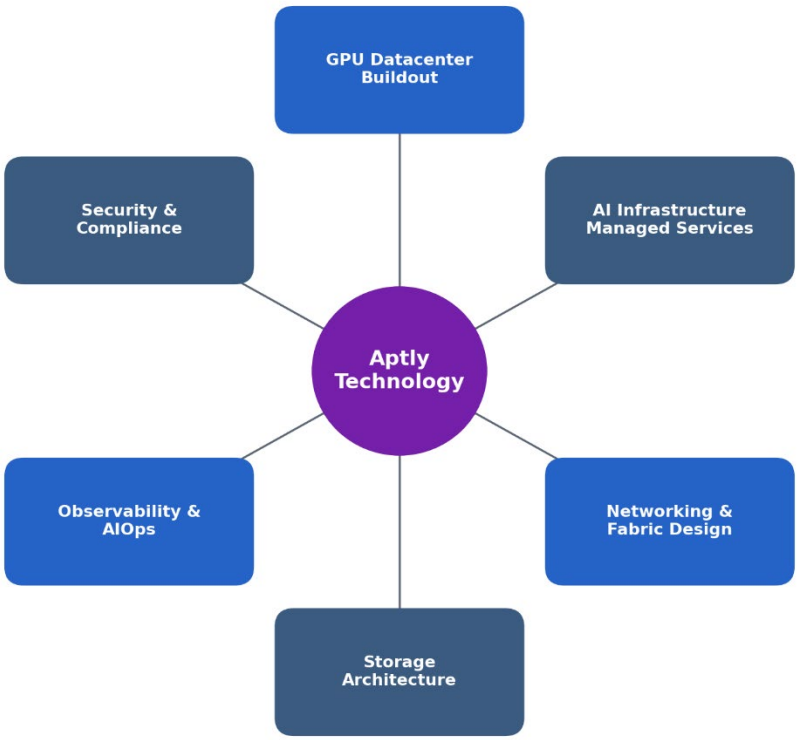
Long Term (3–5 Years)

- Design infrastructure to be AI-ready by default, not AI-ready by exception. New data center capacity, new networking purchases, and new storage platforms should assume AI workloads from the start rather than retrofitting for them later.
- Invest in automation and AIOps to reduce how much specialized expertise each operational incident requires, directly addressing the skills shortage Chapter 4 identifies as the most common barrier enterprises report.
- Embed sustainability into infrastructure design decisions, not only sustainability reporting, as power availability becomes as important a site-selection factor as cost or latency.
- Build modular, portable architecture now, so agentic AI at scale, sovereign compute requirements, and next-generation accelerators do not force a rebuild when they arrive. The enterprises at AI Ready on this report's scorecard built for change. Yours will too.

07: Where Aptly Fits

Aptly Technology designs, builds, and operates the infrastructure layer this report benchmarks, from GPU clusters and AI fabrics to the data centers and operations teams that keep them online. Aptly is the only@ Microsoft-trusted supplier authorized to build and support third-party hyperscale datacenters worldwide, and pairs that role with direct partnerships across the AI hardware ecosystem, including NVIDIA and Supermicro, to deliver AI-optimized compute at scale. The company has deployed and operated tens of thousands of GPU nodes, InfiniBand and high-performance Ethernet fabrics, and AI-ready clusters across 24×7 white-glove Global Operations Centers in North America, Europe, and Asia, underpinned by ISO/IEC 42001:2023 certification from Schellman Compliance, LLC for AI management and governance.

If your readiness assessment surfaces gaps in compute, networking, or data center capacity, Aptly's GPU Datacenter Buildout Services build the physical foundation where production AI runs. These services cover end-to-end design validation against your AI roadmap, racking and cabling at hyperscale, power-up and burn-in, and high-performance networking integration (InfiniBand and AI-optimized Ethernet) so you move from "we need more GPUs" to a fully operational, benchmarked cluster that matches your workload profile. Aptly's experience as a Microsoft-trusted hyperscale partner means power, cooling, rack density, and fabric topology are engineered with the same rigor that underpins public cloud buildouts, but tuned to your enterprise context.



If the gap sits in AI operations, monitoring, or day-to-day reliability, Aptly's AI Infrastructure Managed Services run the estate with uptime SLAs up to 99.9%, closing the operational side of the readiness framework. Aptly's Global Operations Centers provide 24×7 monitoring, incident response, capacity planning, and performance tuning across GPU clusters and AI pipelines, combining SRE-grade practices with AI-specific workload optimization to eliminate idle GPUs, stabilize inference latency, and keep critical AI services within business SLOs. Automation, observability, and governance are built into the managed service so your internal teams can focus on models and applications while Aptly carries the infrastructure reliability burden.

Either way, Aptly positions itself as a single partner for the full AI infrastructure lifecycle—from metal and fabric design through workload deployment, optimization, and ongoing operations. A typical engagement starts with an assessment aligned to the readiness scorecard, maps gaps across compute, networking, storage, data centers, operations, and governance, and then follows with a phased plan that combines buildout services with managed operations where needed. If your assessment surfaces material gaps in any pillar, get in touch with Aptly to benchmark where your organization sits on the scorecard and to design a path from "Developing" or "Advanced" to truly "AI Ready," backed by hyperscale-grade engineering and global delivery capabilities.

Frequently Asked Questions

Straight answers to the questions we hear most often when enterprises start benchmarking their AI infrastructure.

- ***What is enterprise AI infrastructure readiness?***
    - Enterprise AI infrastructure readiness measures whether your compute, networking, storage, data center, operations, security, and scalability foundations support AI workloads at production scale, not only in a pilot. A ready organization moves a working AI use case into production without a redesign of the infrastructure underneath. Most enterprises are not there yet.
- ***What are the seven pillars of the Aptly Enterprise AI Infrastructure Readiness Framework?***
    - The framework scores compute readiness, networking readiness, storage readiness, data center readiness, AI operations, security and governance, and scalability. Each pillar gets scored independently, because a strong score in one, compute is the most common example, cannot compensate for a weak score in another, most often data center power or AI operations maturity.
- ***How is AI infrastructure different from traditional IT infrastructure?***
    - AI infrastructure runs sustained, parallel workloads at power densities traditional IT was never designed for. Current AI training racks already exceed 40 kW, well past what air cooling handles economically, and AI workloads generate continuous inference traffic rather than the bursty, predictable patterns conventional enterprise applications produce. This difference cascades through networking, storage, and cooling design, not only compute.
- ***What is the biggest barrier to enterprise AI adoption in 2026?***
    - Skills, not silicon. Deloitte's State of AI in the Enterprise research names the AI skills gap as the top reported barrier to integrating AI into existing workflows, ahead of budget or technology maturity. Infrastructure constraints now outweigh budget concerns for AI expansion as well, according to Flexential's 2026 survey of enterprise IT leaders, which is why this report treats infrastructure and operations as one connected problem rather than two separate budget lines.
- ***How often should enterprises reassess AI infrastructure readiness?***
    - Reassess at least once a year, and always before a major AI budget or hardware refresh cycle. AI infrastructure requirements move faster than typical IT planning cycles: rack density benchmarks, network architecture standards, and accelerator generations have all shifted meaningfully within a single year recently. An annual reassessment against the seven-pillar framework catches gaps before they become the reason a production rollout stalls.

- ***Should enterprises build AI infrastructure on-premises or in the cloud?***
    - Neither, exclusively. Deloitte's research on enterprise compute strategy describes a three-tier pattern taking hold across the enterprises it studies: public cloud for elastic training and experimentation, private infrastructure for predictable, high-volume inference, and edge compute for latency-critical decisions. Choose the environment workload by workload, based on cost, latency, and data sovereignty requirements, rather than committing the whole AI program to one environment.
- ***What role does Aptly play in enterprise AI infrastructure?***
    - Aptly Technology designs, builds, and operates the physical and operational infrastructure this report benchmarks. As the only Microsoft-trusted supplier authorized to build and support third-party hyperscale datacenters worldwide, with direct NVIDIA and Supermicro partnerships, Aptly has delivered and operated tens of thousands of GPU nodes, InfiniBand fabrics, and AI-optimized clusters, backed by 24x7 Global Operations Centers and uptime SLAs up to 99.9%. If your readiness score points to a gap, Aptly is built to close it.

Appendix: Research Methodology & Sources

Research Methodology

This report combines three sources of evidence. First, published market research and analyst forecasts from Gartner, IDC, McKinsey, Deloitte, the International Energy Agency, Uptime Institute, Dell'Oro Group, and MIT, covering AI infrastructure spending, enterprise adoption, power and cooling trends, and networking architecture through mid-2026. Second, public data from hardware and cloud providers, including NVIDIA, Microsoft, and AMD, on accelerator supply, deployment scale, and platform trends. Third, Aptly Technology's own delivery experience across GPU datacenter buildouts and managed AI infrastructure engagements, which informs the Aptly Enterprise AI Infrastructure Readiness Framework introduced in Chapter 3.



The readiness framework itself is Aptly's proprietary diagnostic model. It synthesizes patterns from the market research above with recurring themes Aptly observes across enterprise infrastructure engagements. Enterprises should use it as a structured starting point for their own assessment, not as a substitute for a detailed technical audit of their specific environment.

All figures in this report are sourced and hyperlinked at the point they appear. Where analyst firms report different figures for related but distinct metrics, for example AI infrastructure hardware spending versus total AI spending across the full technology stack, this report names the specific metric and its source rather than treating the numbers as interchangeable.

Primary Sources Referenced



Source	What This Report Draws From It
Gartner	Worldwide AI and IT spending forecasts, AI agent software spending projections
IDC	Worldwide Quarterly AI Infrastructure Tracker, global and regional spending data
Deloitte	Tech Trends 2026, State of AI in the Enterprise, enterprise AI infrastructure survey
International Energy Agency (IEA)	Data center electricity demand and AI power consumption forecasts
Uptime Institute	Rack density, liquid cooling adoption, and data center survey data
Dell'Oro Group	AI back-end network switch market share, Ethernet and InfiniBand trends

Source	What This Report Draws From It
MIT (NANDA initiative)	Enterprise generative AI pilot outcomes and ROI research
NVIDIA	Enterprise AI adoption survey data and accelerator platform trends
Flexential	2026 State of AI Infrastructure Report, enterprise IT leader survey data
Mordor Intelligence	AI infrastructure market segmentation and regional forecasts
Nutanix	2026 Enterprise Cloud Index, containerization and AI workload governance data

About Aptly Technology

Aptly Technology designs, builds, and operates enterprise AI infrastructure, from GPU datacenter buildouts to 24x7 managed AI operations. Aptly is the only Microsoft-trusted supplier authorized to build and support third-party hyperscale datacenters worldwide, with direct NVIDIA and Supermicro partnerships and Global Operations Centers across North America, Europe, and Asia, certified to ISO/IEC 42001:2023 by Schellman Compliance, LLC. To discuss where your organization sits on the readiness scorecard in this report, contact Aptly.