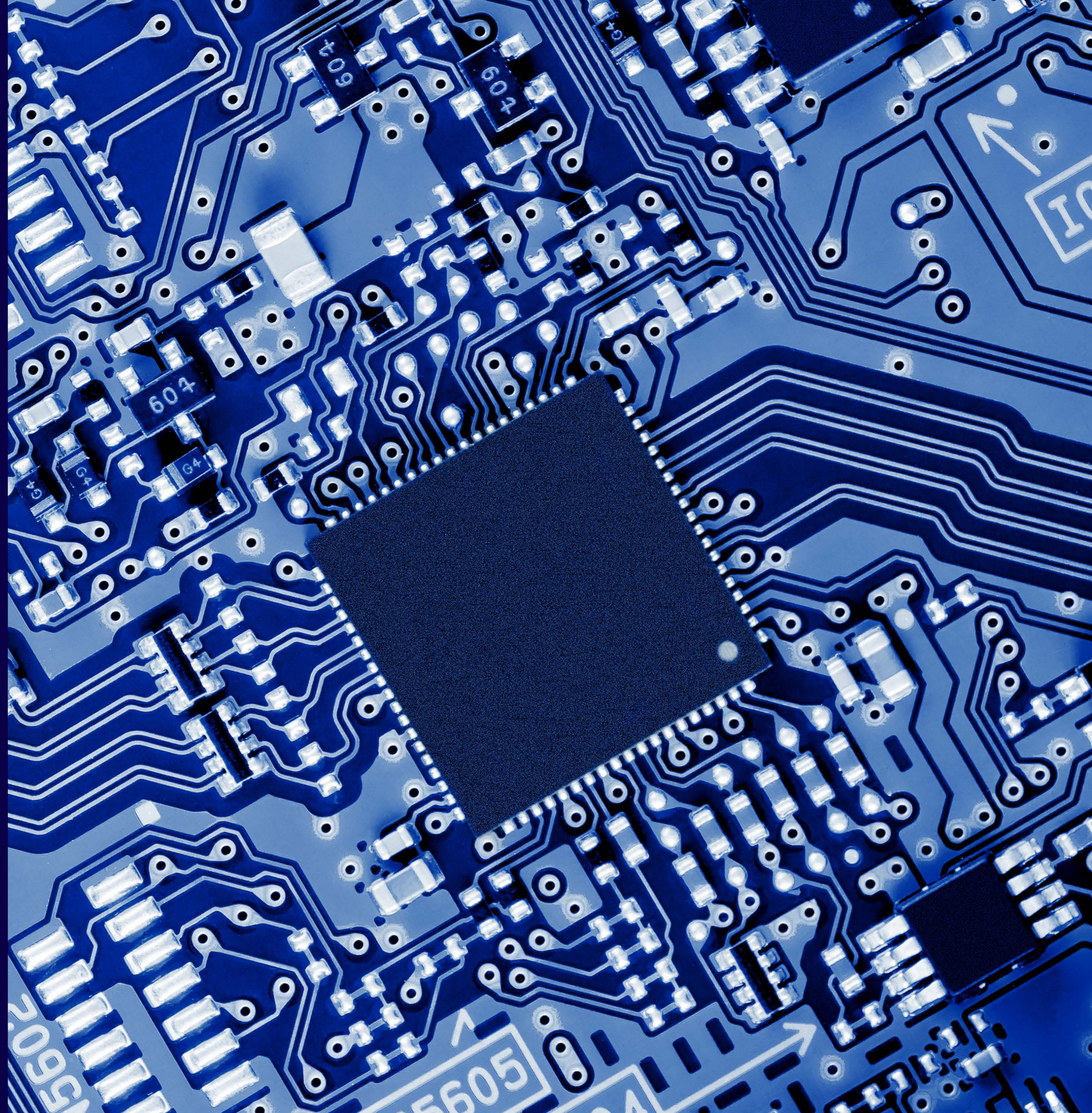
ERM
SUSTAINABILITY
INSTITUTE

Turning Bottlenecks into Advantage: **Future-Proofing the AI Infrastructure Value Chain**

June 2026



Contents

Executive summary 3

Introduction 6

AI infrastructure as a catalyst for transformation 8

Why AI infrastructure requires a new operating model 10

The AI infrastructure value chain: When one node fails, the system breaks 14

What companies get wrong in the AI infrastructure value chain 24

How to future-proof the AI infrastructure value chain: Seven moves leaders should make now 27

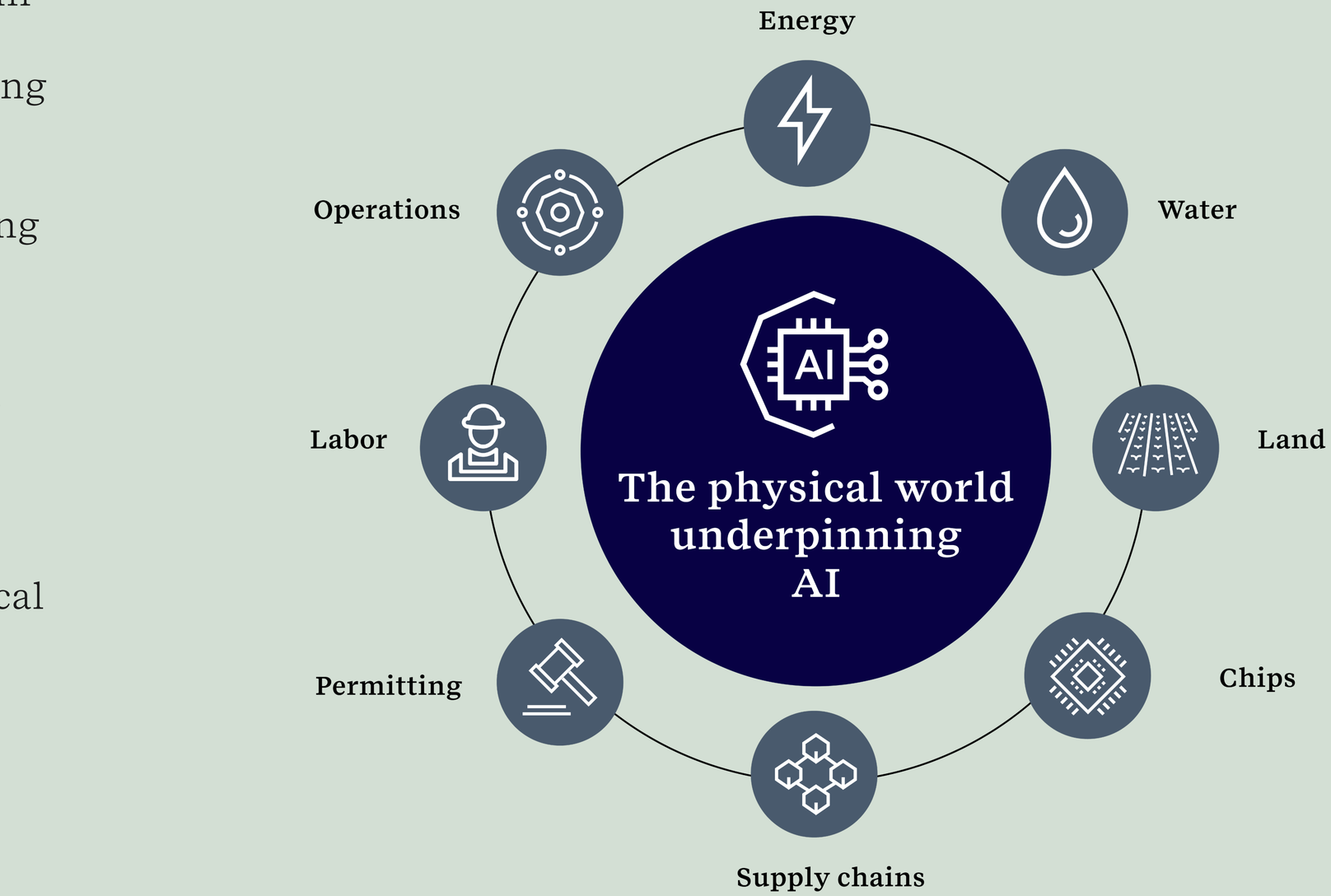
Conclusion: Scaling AI with confidence in a constrained world 31

Executive Summary



ERM
SUSTAINABILITY
INSTITUTE

The buildout of Artificial Intelligence (AI) infrastructure will be one of the most consequential industrial expansions of the coming decade. However, the factors that determine AI scale are no longer limited to compute—the processing power needed to train and run AI models—but the physical world that underpins deployment: energy, water, land, chips, supply chains, permitting, labor, and operations. This marks a shift in the AI infrastructure operating model where orchestrating physical systems is paramount.



As demand intensifies, constraints are emerging simultaneously across the AI infrastructure value chain. Material shortages can delay infrastructure buildout, power and water limitations can force suboptimal siting decisions, and accelerated hardware cycles can reinforce upstream supply pressure.

At the same time, the scale of AI infrastructure investment presents a significant opportunity to accelerate speed to scale, secure speed to power, preserve optionality, strengthen deployment resilience, maintain capital efficiency, and build infrastructure that can adapt as conditions evolve. However, these benefits will not emerge from siloed approaches that optimize individual components without accounting for system-level interactions.

Future-proofing the AI value chain requires orchestrating an integrated physical system.

ERM helps organizations connect decisions across the system—turning physical constraints into practical pathways for more sustainable, resilient, and scalable AI growth.

Seven moves to future-proof the AI infrastructure value chain

1

Plan AI infrastructure as a connected physical system:

A system-level approach to AI infrastructure enables companies to anticipate and manage interconnected constraints across the value chain—avoiding bottlenecks and delays while improving speed and utilization.

2

Use portfolio-level planning to preserve optionality:

Portfolio-level planning enables companies to coordinate resource, regulatory, and stakeholder constraints across assets—improving capital allocation, adaptability, and overall system performance.

3

Treat power, water, and compute as one infrastructure planning model:

Integrated power, water, and compute planning aligns infrastructure with physical constraints—avoiding inefficiencies and bottlenecks while improving performance and speed to scale.

4

Move sourcing from lowest cost to risk reduction:

Considering sourcing constraints alongside cost considerations helps companies anticipate chokepoints—improving supply resilience and operational continuity.

5

Design for refresh, reuse, recovery, and redeployment from day one:

Designing AI hardware for refresh and circularity extends asset life—reducing obsolescence, lowering lifecycle costs, and easing reliance on constrained inputs.

6

Scenario-test binding constraints before committing capital:

Scenario planning helps companies anticipate which constraints will become binding under different conditions—reducing risk and enabling more adaptive, resilient execution.

7

Build deeper supplier and ecosystem partnerships:

Supplier partnerships improve visibility into upstream constraints—reducing supply risks and enabling more proactive, coordinated scaling.

Introduction

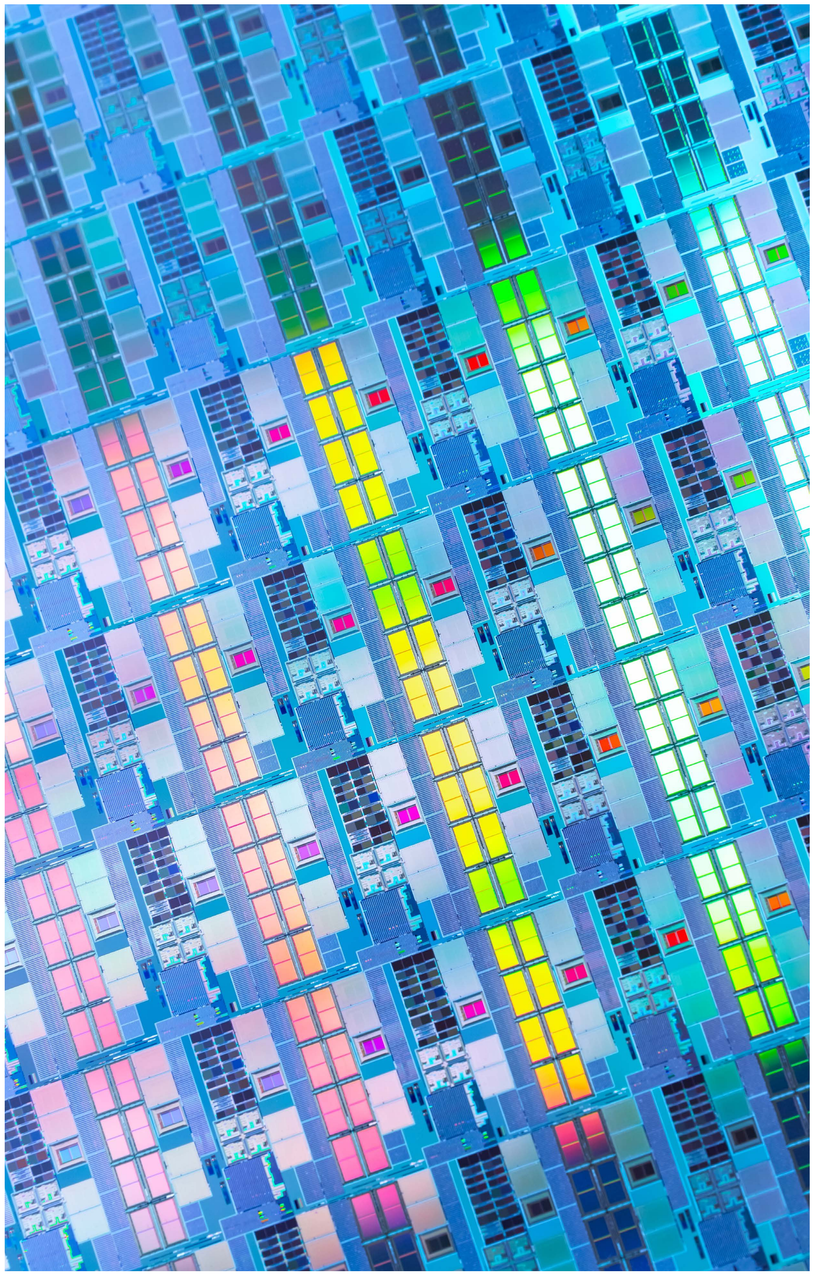


ERM
SUSTAINABILITY
INSTITUTE

Artificial Intelligence (AI) is increasingly shaped by physical systems, anchored in tangible inputs and infrastructure. In this context, AI is no longer constrained by compute, cloud, and software abstractions, but by the infrastructure value chain—from silicon and servers to site development and operations—that makes development and delivery possible.

For companies already managing value chain risk, this shift introduces new complexity: constraints are no longer isolated risks to be mitigated individually. They are interdependent pressures that evolve dynamically and propagate across the value chain. Disruptions in one layer can rapidly cascade, amplifying risk exposure and delaying timelines.

Traditional risk management approaches are insufficient in this environment. Competitive advantage will depend on treating the AI infrastructure value chain as an integrated system and coordinating decisions across upstream, midstream, and downstream layers. Companies that adopt integrated approaches can reduce risk, unlock faster scaling, and build more resilient and sustainable growth pathways.



AI infrastructure as a catalyst for transformation



ERM
SUSTAINABILITY
INSTITUTE

AI infrastructure is not only a source of new physical constraints to be overcome; it is also a catalyst for system-level transformation. The scale of AI demand is pushing companies, utilities, suppliers, developers, and communities to rethink how energy, water, land, materials, and compute are planned together. Done well, this buildout can accelerate grid investment, create demand for lower-carbon materials, improve water stewardship, strengthen supplier transparency, and drive more circular approaches to hardware and components.

The opportunity is not simply to build more data centers. It is to build a more adaptive infrastructure system around AI—one that aligns growth with resource availability, community expectations, regulatory realities, and long-term asset performance. Companies that treat physical constraints as early design inputs, rather than late-stage project risks, will be better positioned to move faster, protect margins, reduce disruption, and earn trust with stakeholders.



ERM LENS:

ERM sees AI infrastructure as an integrated physical system, not a sequence of separate projects. Our work across energy, water, permitting, materials, supply chains, EHS, sustainability, and operations gives us visibility into the intersections where AI infrastructure risks and opportunities emerge.

Why AI infrastructure requires a new operating model



ERM
SUSTAINABILITY
INSTITUTE

AI infrastructure growth will be shaped by the availability, accessibility, reliability, and social acceptability of inputs like energy, water, materials, land, labor, and capital that determine where AI infrastructure can be built, how quickly it can scale, and whether it can operate reliably over time.

As demand for these inputs intensifies, a set of system-wide constraints is emerging through emissions pressures, supply chain concentration, policy interventions, community responses, and investor pressures, shaping AI deployment. Together, these constraints highlight how traditional AI infrastructure operating models focused solely on compute are no longer sufficient.



Carbon emissions growth runs into on-the-ground realities

KEY CONSTRAINTS

The rapid expansion of AI and the data centers that power them is driving a significant increase in carbon emissions from both energy and upstream inputs.

150% The average increase in operational emissions from four leading AI firms between 2020 and 2023 (largely attributed to growing data center electricity use and dependence on grid energy mixes beyond company control).

33 to 80 million tons of CO2 AI systems' estimated global carbon footprint in 2025 (New York City's carbon footprint was 52.2 million tons in 2023).

24-44 million metrics tons of CO2 Estimated carbon footprint of U.S. AI systems in 2030 based on current growth rates (equivalent to adding 5 to 10 million cars to U.S. roads).

IMPLICATIONS FOR AI DEPLOYMENT

- **Policy pressures:** Growing AI-related carbon emissions are likely to face policy pressures like renewable energy mandates, emissions limits for on-site power generation, and requirements not to pass on new electricity generation costs to consumers, which could slow speed to scale.
- **Finance access:** Investors condition financing on decarbonization plans that provide credible pathways to reduce carbon emissions.
- **Hardware limits:** Reducing the embodied carbon of input materials may run into hardware realities where emissions cannot be reduced further without compromising AI performance.

Supply chain concentration imposes material constraints

KEY CONSTRAINTS

AI systems depend on a narrow set of materials, components, and manufacturing nodes that are highly concentrated by geography and supplier.

99% and 74% China's share of global refining capacity for AI-relevant critical minerals gallium and germanium, respectively.

70% Share of the global foundry market controlled by Taiwan Semiconductor Manufacturing Company (TSMC).

36 to 52 weeks Current lead times for data center Graphic Processing Units (GPUs), reflecting persistent supply shortages across advanced AI hardware.

IMPLICATIONS FOR AI DEPLOYMENT

- **Bottleneck risks:** Concentrated supply chains mean that disruptions at individual upstream nodes can delay AI capacity expansion globally.
- **Cost pressures:** Scarcity in materials, chips, and hardware increases price volatility, raising the marginal cost of AI training.
- **Emerging deficits:** Emerging deficits of some essential materials due to upstream pressures and competing needs raises risk of future shortages.

Policy scrutiny increases permitting and compliance risk

KEY CONSTRAINTS

Growing concerns over data centers are heightening policy scrutiny of projects.

200+ The number of bills introduced across all 50 U.S. states in 2025 to regulate data centers.

27 The number of U.S. states advancing legislation requiring data center developers to cover energy costs and report electricity usage as of April 2026.

IMPLICATIONS FOR AI DEPLOYMENT

- **Permitting challenges:** More data center-specific policies are likely to translate to more onerous permitting processes that could extend project timelines.
- **Compliance requirements:** Increased data center compliance requirements are likely to raise project costs.

Community issues challenge social license to operate

KEY CONSTRAINTS

Local opposition to the perceived impacts of data centers on local communities is increasingly slowing and, in some cases, halting AI infrastructure projects.

$64 billion: Value of U.S. data center investments blocked or delayed by local opposition groups between March 2024 to March 2025.

38% and 39%: Americans who view data centers as having a negative effect on the environment and home energy costs, respectively.

37%: Americans who would support a new data center built in their communities.

IMPLICATIONS FOR AI DEPLOYMENT

- **Geographic limitations:** Resistance in established data-center hubs may push developers toward less optimal locations for power, water, or network connections, raising costs and reducing performance efficiency.
- **Reputational exposure:** As AI's physical footprint becomes more visible, negative local sentiment toward data centers can spill over into broader perceptions of AI companies, raising brand and trust risks.

Capital intensity raises investor scrutiny and prioritization pressure

KEY CONSTRAINTS

Scaling AI infrastructure requires enormous capital, but for leading technology companies, the issue is not access to capital, it is whether investments can deliver credible returns and reliable capacity.

$6.7 trillion Estimated investment needed by 2030 for data centers to match AI and traditional compute demands.

$750 billion The projected Capex spend of the 14 largest publicly owned data center operators globally in 2026.

57% The rise in global data center Capex spend in 2025 compared to 2024.

IMPLICATIONS FOR AI DEPLOYMENT

- **Disciplined capital allocation:** Investor interest in disciplined capital allocation will likely lead to companies favoring projects with demonstrable use cases and returns
- **Resilient siting demands:** Project sites with access to power, water, and other resilience-boosting factors are likely to become more attractive as companies seek to de-risk capital deployment and secure investor confidence.
- **Demonstrated business value:** Companies will likely prioritize projects with credible business value pathways to justify capital intensity to investors.

ERM LENS:

For AI infrastructure leaders, the challenge is not simply identifying constraints. It is understanding how decisions in areas such as energy sourcing, water strategy, materials selection, or site design shift risk and value elsewhere in the system. ERM helps organizations evaluate those trade-offs early, before they become project delays, stranded investments, or stakeholder conflicts.

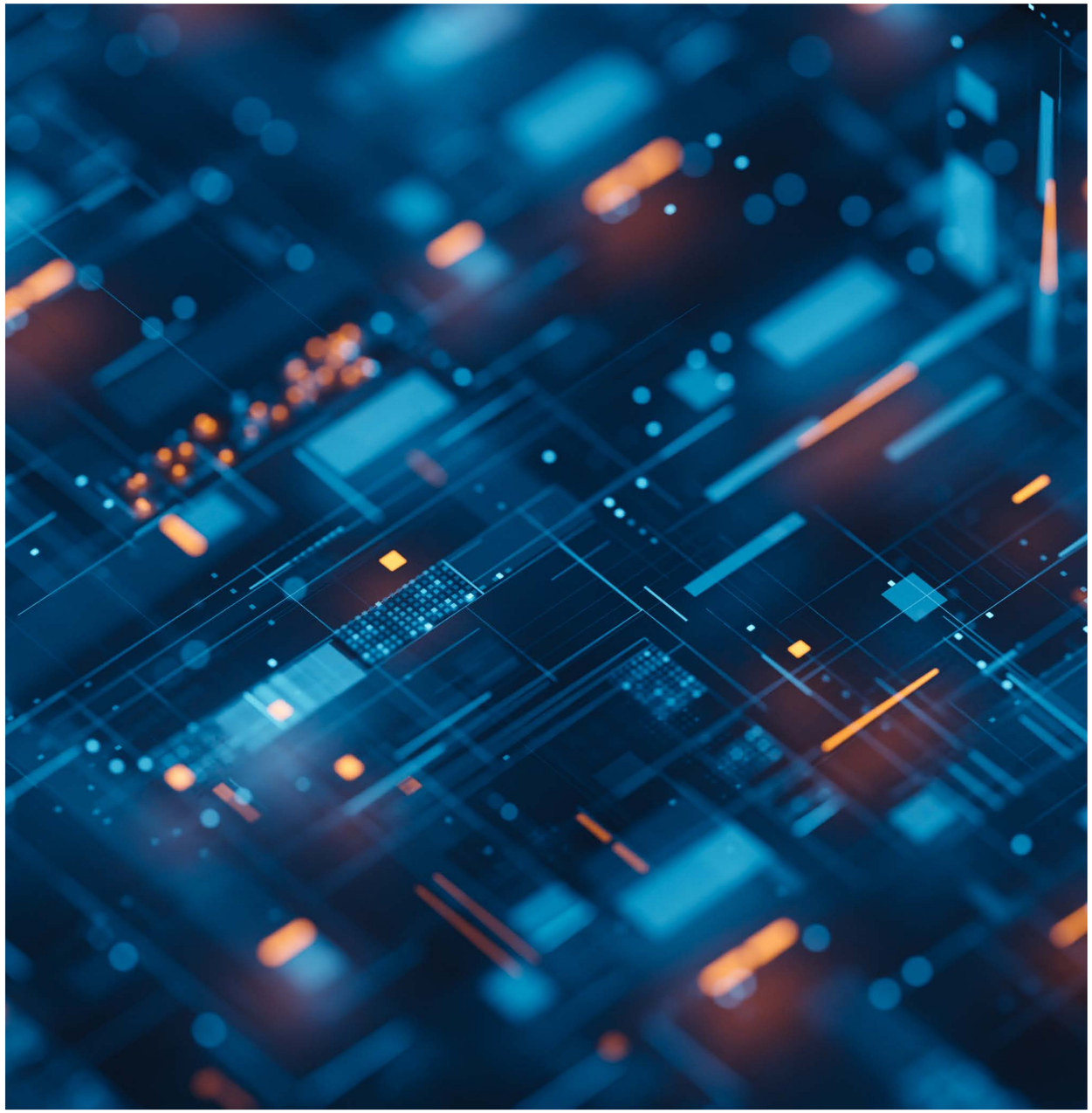
The AI infrastructure value chain: When one node fails, the system breaks



ERM
SUSTAINABILITY
INSTITUTE

While system-wide constraints pose challenges, the AI infrastructure value chain is not a single coherent mass; rather, it is a set of three interconnected layers—upstream, midstream, and downstream—that form an interconnected system reflecting the global scale and complexity of the technology.

Each of these layers is composed of individual nodes that together enable AI delivery when functioning in concert. However, as at the system level, these nodes also face constraints. Managing them effectively requires orchestration across layers rather than optimization within them.



Three interconnected value chain layers

Upstream: Build the machine

Value chain node	Why it is essential to AI growth	Critical constraints
 **AI hardware**	Without key hardware components, AI data centers would be little more than empty buildings. *Examples of key hardware components for AI data centers:* **Servers** Servers host and execute AI trainings and workloads. **Storage systems** Hold the datasets (training data, model artifacts, etc.) that make AI operations possible. **Network infrastructure** Enables communication between servers, storage systems, external systems, and model users.	**Hardware lifecycle carbon footprint:** AI hardware components carry significant carbon footprints. The estimated lifecycle emissions of a typical AI server, for example, is approximately 2,748 $kgCO_2e$. **Memory shortages and price pressures:** Memory technologies like High Bandwidth Memory (HBM) are critical to AI, providing the necessary bandwidth for large workloads. With demand rising and shortages emerging, memory prices are surging, increasing by as much as 90% in Q1 2026 compared to Q4 2025.
 **Chips**	AI chips perform the math operations, memory access, and data movement that train and operate AI models. *Examples of AI chips:* **Application-specific Integrated Circuits (ASICs)** Run AI inference. **Graphic Processing Units (GPUs)** Train AI systems and run AI inference. **Field-programmable Gate Arrays (FPGAs)** Run AI inference.	**Chip bottlenecks:** Advanced AI chips are produced in limited volume by a few companies, with demand outpacing manufacturing capacity. Taiwan Semiconductor Manufacturing Company's (TSMC) dominance in chip production also poses risks, given Taiwan's geographic vulnerabilities.

Value chain node	Why it is essential to AI growth	Critical constraints
 **Concrete**	AI data centers depend on significant amounts of concrete to house servers, cooling technologies, and other equipment, making the input a key factor influencing how quickly new AI capacity can come online. U.S. data centers are forecasted to consume approximately 860,000 metric tons of concrete between 2025-2027, equal to 25.7% of all office-related construction consumption.	**Concrete embodied carbon:** Concrete production is carbon-intensive, locking in substantial embodied carbon. New U.S. Al data centers could generate approximately 2 million metric tons of CO2 through 2030, equivalent to the annual emissions of 400,000 cars.
 **Critical minerals and metals**	Critical minerals and metals form the foundation for many of the components and systems that make AI possible, not to mention their criticality to AI supporting activities such as power generation. *Examples of critical mineral and metal inputs for various AI data center components:* **Serverboards and circuitry** • Copper • Gold • Silver **Heat sinks** • Aluminum • Copper **Semiconductors** • Gallium • Germanium • Silicon	**Competing needs for materials:** Critical minerals and metals are needed for many other sectors beyond AI, including aerospace, defense, power, and transportation. Global copper demand, for example, is projected to grow by 24% by 2035. **Growing materials demand, inelastic supply:** Demand for AI-essential critical minerals and metals is rising, with each MW of data center capacity requiring between 60-75 to metric tons of the materials. However, bringing online new mining and refining capacity can take decades, with new mines taking on average 29 years to commence operations in the U.S. **Materials supply chain concentration:** The mining and processing of many AI-essential critical minerals is concentrated in China. For example, China controls 98% of global gallium production and 60% of germanium refining.
 **Labor**	From mining and hardware manufacturing to site development and delivery, AI data centers rely on a wide variety of labor. For data center construction alone, a typical 250,000 square foot facility can require up to 1,500 on-site workers. *Examples of key labor roles in the AI infrastructure value chain:* **Mining** • Geologists • Miners • Refinery operators **Site development** • Heavy equipment operators • Concrete workers and finishers • Welders **Data center delivery** • Electricians • HVAC technicians • Fiber optic technicians	**Supply chain labor shortages:** The skilled labor needed across the AI infrastructure value chain is in increasingly short supply. For example, 71% of mining leaders report talent shortages affecting production target achievement. **Data center labor shortages:** Data centers face their own shortages of specialized engineering and construction talent, with 52% of U.S. data center construction sites reporting staffing gaps in 2025.

Value chain node	Why it is essential to AI growth	Critical constraints
 **Land**	AI data centers need large, secure parcels of land with ready access to power, water, and fiber connections to operate. Hyperscale data center campuses with more than 1 GW of capacity can now require upwards of 800 acres of land.	**Power and water-related land scarcity:** While land itself is not a constraint, land with sufficient power and water availability to operate AI data centers, coupled with local community and political support, is scarce. Many potential sites may have plenty of power and water, but lack community support, while those with community support may lack water. **Local policy and regulation restrictions:** Local policy and regulation are increasingly restricting where data centers can be built, as local governments respond to constituent concerns. Loudoun County, VA, for example, eliminated by-right zoning for data centers in 2025.
 **Steel**	Steel provides the structural backbones of AI data centers, supporting the heavy loads of data center hardware. A hyperscale data center campus requires up to 20,000 tons of steel per facility, much more than typical commercial buildings.	**Steel scope 3 emissions:** Major data center operators have seen their Scope 3 emissions surge because of embodied emissions from steel and other data center inputs. Globally, the average ton of virgin steel carries a carbon footprint of 2.3-2.9 tCO2. **Niche steel product shortages:** Niche products like grain-oriented electrical steel (GOES) used in power transformers that step down grid power for data centers is in short supply. There is effectively one supplier of GOES in the U.S. and production is not keeping pace with demand, leading to four plus year wait times for new transformers.

Midstream: Power the machine

Value chain node	Why it is essential to AI growth	Critical constraints
Power	AI data centers are energy intensive, requiring abundant and reliable power to run the models that handle AI training and workloads. The power capacity of a hyperscale facility typically exceeds 100 MW, equal to the annual electricity demand of 80,000 U.S. households.	**Power generation capacity:** Global power networks are struggling to keep up with growing demand. In 2024 and 2025, global electricity demand grew by 4.4% and 3%, respectively, the fastest in more than a decade. Add in AI data center power demands (U.S. AI data centers could require 14 GW of new capacity by 2030), and the generation capacity picture becomes even more complicated. **Gas turbine delays:** Natural gas will play a significant role in meeting AI power demands in the near-term. However, sourcing natural gas turbines is currently a challenge. Since 2021, average gas turbine lead times went from 2 years to 5 years due to constrained manufacturing capacity. **Power interconnection lags:** Significant power capacity is held up in interconnection queues, slowing AI data centers connections to new generation. In the U.S., the typical 2024 generation project spent 55 months from interconnection request to commercial operations, compared to 22 months in 2008.
Water	AI data centers require significant water to run at peak efficiency and to produce components and their raw materials inputs. A 100 MW hyperscale facility in the U.S. can consume more than 500,000 gallons of water (both directly for cooling purposes and indirectly via embodied water) per day, equivalent to 6,500 households.	**Data center indirect water consumption:** Over 60% of data centers' water consumption is indirect (e.g., from electricity generation and hardware). As AI growth continues, this footprint will amplify exposure to water intensity risks across the value chain. **Local water availability:** AI data centers are increasingly built in areas with limited water availability, placing them in conflict with other local demands. Since 2022, approximately two-thirds of data centers in the U.S. were built in areas of high-water stress.

Downstream: Run the machine

Value chain node	Why it is essential to AI growth	Critical constraints
AI hardware lifecycles	AI's rapid advances have been made possible by innovations in chips and other hardware. The Nvidia GB200 NVL72 data center supercomputer, launched in 2025, runs AI inference 30-40 times faster than the company's older chips like the H100, which were launched in 2022.	**Effective obsolescence:** Rapid AI advances mean hardware quickly goes obsolete, straining sourcing and end-of-life management. GPUs running typical AI workloads may last only 1-3 years, depending on utilization.
AI hardware utilization	AI model advances rely on increasing amounts of compute, which depends on GPUs and other expensive and energy intensive hardware, for training. GPU utilization (e.g., compute used in training or inference) in real-world use cases often hovers between 10% and 40%, offering significant opportunity for improvement.	**Underutilization costs:** Underutilized AI hardware creates capacity limits that show up in higher costs and slower deployment. Salesforce Research cut AI training costs by 42% by raising GPU utilization from 60% to close to 100%.
Data center cooling	AI data center operations generate significant heat that must be dissipated by cooling infrastructure. The average power density of data center server racks is expected to increase from 36 kW in 2023 to 50 kW by 2027, raising cooling demands.	**Power consumption:** As power densities rise, larger portions of data center power must be dedicated to cooling rather than AI training and inference, reducing overall capacity and project economics. Cooling accounts for between 7% and 30% of a data center's energy consumption.

Energy, labor, and water as structural inputs and value levers

Energy, labor, and water are structural inputs and value levers critical to future-proofing the AI value chain.

Energy

Energy affects whether data centers can be built and where, how reliably they can operate, and at what cost. Future-proofing requires companies to assess energy availability, reliability, and cost during sourcing and site selection, align infrastructure development with grid capacity and decarbonization pathways, and secure diversified, low-carbon energy sources.

Labor

Labor shapes the speed, cost, and feasibility of scaling data center capacity. Future-proofing requires companies to assess workforce capacity and capability, integrate talent considerations into site selection and supply chain planning, invest in training and reskilling programs, and build partnerships with local communities to ensure a resilient pipeline of skilled labor.

Water

Water influences input sourcing, site viability, energy production, and cooling strategy. Future-proofing requires companies to evaluate water security as part of site selection, energy planning, supplier risk, cooling strategy, and long-term asset performance.



One layer’s challenge is another’s impact

In the AI infrastructure value chain, challenges in one layer reshape feasibility and performance in others. Whether from upstream to midstream or downstream to upstream, the value chain cannot deliver deployment at scale without managing cross-layer interactions in concert.

Upstream to midstream

UPSTREAM CHALLENGE	MIDSTREAM IMPACT
AI hardware shortages	Shortages of AI hardware hinder the scope of data center build-out, limiting AI capacity increases and asset performance.
Limited input availability (e.g., chips, critical minerals, etc.)	Reduced availability of key inputs slows data center build-out, curtailing AI capacity increases.
Turbine and transformer scarcity	Insufficient turbine and transformer availability means data centers cannot get the power they need to commence operations.

Midstream to downstream

MIDSTREAM CHALLENGE	DOWNSTREAM IMPACT
Land permitting delays	More stringent regional regulations delay permitting, extending data center development and operational timelines.
Grid shortfalls and interconnection delays	Inadequate regional grid capacity and interconnections force data centers to operate at reduced capacity or not at all.
Site-level power and water supply insufficiencies	Suboptimal data center siting for power and water needs reduces operational efficiency and increases cost per unit of compute.

Downstream to upstream

DOWNSTREAM CHALLENGE	UPSTREAM IMPACT
Accelerated AI hardware refresh cycles	Increased demand for hardware inputs reinforces upstream bottlenecks, further constraining supply availability.
AI hardware burnout	High and unoptimized demands accelerate hardware degradation, necessitating more frequent replacement that places heightened pressures on already stretched upstream supply chains.
Increased water-cooling demands	The increasing adoption of water-cooling technologies stretches upstream supply chains' ability to supply components and systems.

ERM LENS:

ERM operates at interfaces such as materials and manufacturing, power and permitting, water and siting, operations and sustainability, and circularity and supply resilience. At these points, fragmented decisions create cascading risk while integrated planning can create competitive advantage by accounting for interdependencies.



What companies get wrong in the AI infrastructure value chain



ERM
SUSTAINABILITY
INSTITUTE

Why siloed approaches fail

Most AI infrastructure strategies still reflect a compute-centric operating model, where decisions are optimized in isolation. In a constrained environment, these approaches fail and ultimately reinforce the very dependencies they aim to solve by eroding speed to scale, reducing optionality, weakening resilience, and increasing capital inefficiencies. To unlock AI's full potential, companies must address each value chain layer and their interlinkages simultaneously.

Designing data centers without regard to supply chain realities

WHY IT FAILS:

Not integrating supply chain realities (e.g., critical minerals, chip capacity) into design decisions limits optionality and build feasibility, creating dependencies on constrained inputs while hindering design-for-circularity and end-of-life planning.

Committing to an evaporative cooling strategy before assessing water supply security

WHY IT FAILS:

Cooling decisions made without understanding long-term water availability can limit adaptability and increase exposure to water curtailments, costly retrofits, permitting friction, and operational disruption.

Optimizing Power Usage Effectiveness (PUE) in isolation from future throughput plans

WHY IT FAILS:

Optimizing PUE for current data center loads before accounting for future throughput plans risks driving systems back into inefficient operation as workloads scale or shift.

Sourcing data center components without considering end-of-life

WHY IT FAILS:

Sourcing data center components without considering end-of-life pathways (e.g., reuse, refurbishment, and material recovery) locks companies into linear consumption models that increase their exposure to future supply challenges and drive higher lifecycle costs due to disposal, re-procurement, and unrealized value retention.

Locking in data center sites before determining local grid capacity

WHY IT FAILS:

Site decisions made without understanding grid readiness constrain location flexibility and scalability, forcing retrofits, delays, or stranded capacity when power availability cannot meet projected demand.

Managing Scope 3 emissions as a reporting, not design input

WHY IT FAILS:

Treating Scope 3 as backward-looking reporting rather than a forward-looking design parameter locks in carbon-intensive supply chains, reducing future regulatory agility and decarbonization optionality.

Treating power as a procurement issue instead of a capacity constraint

WHY IT FAILS:

Focusing only on power sourcing rather than system-wide capacity overlooks structural issues, restricting resilience and missing opportunities to diversify through on-site generation or alternative energy models.

Integration wins: What AI companies can learn from other industries

The need for an integrated value chain approach is not unique to the AI industry. Companies have faced the same challenges during previous business transformations. These transformations illustrate the perils of siloed approaches and how integrated operating models create coordination and ecosystem advantages.

The Japanese auto industry vs. The U.S. auto industry

In the 1970s and 80s, Japanese automakers built integrated supplier networks, enabling just-in-time production, reducing inventories, improving quality, and raising productivity. In comparison, U.S. automakers' failed to integrate their suppliers, creating design and sourcing inefficiencies that, alongside factors such as oil shocks, allowed Japanese automakers to pull ahead.

Apple's integrated ecosystem vs. the disconnected PC and Android ecosystems

Apple's tightly integrated ecosystem—spanning chip design, operating systems, and hardware—delivers optimized products through control of critical interfaces. Conversely, the fragmented PC and Android ecosystems, where design is split across firms, make it harder to fine-tune products even when components are advanced.

Walmart and Amazon vs. traditional retailers

Walmart's integrated supply chains reduced inventory and logistics costs, while traditional retailers siloed approaches raised costs and increased inefficiencies. More recently, Amazon's unified operating model for demand data, fulfillment, logistics, and last-mile delivery has enabled faster, more reliable e-commerce that traditional retailers have struggled to match.



Lessons for AI companies

These examples point to a consistent pattern. In periods of rapid technological and market transformation, value chain coordination becomes a primary source of competitive differentiation. Companies that align decision-making and establish ecosystem control outperform those that optimize in silos.

AI infrastructure faces a similar moment, but with a more complex and interdependent physical environment. As a result, the ability to orchestrate this environment across infrastructure value chain decisions will be central to competitive advantage.

How to future-proof the AI infrastructure value chain: Seven moves leaders should make now



ERM
SUSTAINABILITY
INSTITUTE

Future-proofing the AI infrastructure value chain does not mean eliminating physical constraints. It means designing infrastructure, sourcing, siting, operations, and partnerships so the system can adapt as pressures shift. The goal is to reduce fragility, preserve optionality, and make decisions that remain viable under changing conditions.

Leaders who adopt integrated approaches will win

AI infrastructure will be one of the defining buildouts of the next decade. The winners will be companies that can execute across interdependent physical systems, aligning energy, water, materials, land, permitting, and supply chain management under a single, integrated operating model.

Those that do will improve speed to scale, avoid bottlenecks that stall competitors, preserve optionality, and build greater adaptability and resilience, turning constraints into durable advantage.

ERM LENS:

Future-proofing requires more than a supply chain risk platform or a sustainability roadmap. It requires translating physical realities into decision-making—where to build, what to source, how to operate, and when to refresh assets. By embedding these realities into capital allocation, site selection, project design, and operational strategy, decisions remain viable as conditions evolve.

1

Plan AI infrastructure as a connected physical system

WHY THIS MATTERS:

AI infrastructure is a connected physical system, where constraints affect performance across the value chain. Individual project management rather than ecosystem direction will cause sourcing bottlenecks, project delays, and underutilized capacity.

WHY THIS WORKS:

A system-level approach helps companies anticipate and mitigate constraints and their interactions across the value chain, lowering risk and improving deployment speed and adaptability.

2

Use portfolio-level planning to preserve optionality

AI scale is determined at the portfolio level, where factors such as resource availability, permitting risk, community acceptance, and end-of-life must be coordinated across assets. Site-level decision-making alone cannot account for network interactions and shifting macro-trends like regulatory scrutiny and stakeholder pressures.

Portfolio-level planning enables more efficient allocation of capital, resources, and compute across sites, while helping companies more quickly adapt to changing conditions by preserving optionality.

3

Treat power, water, and compute as one infrastructure planning model

WHY THIS MATTERS:

Power, water, and compute must be planned and deployed together. Managing them in isolation risks inefficient site and design choices and misaligned capacity and infrastructure that cannot operate at scale.

WHY THIS WORKS:

One planning model that aligns AI infrastructure with physical realities helps avoid bottlenecks, optimize performance, and accelerate deployment timelines.

4

Move sourcing from lowest cost to risk reduction

WHY THIS MATTERS:

AI scale faces sourcing chokepoints such as critical mineral availability, carbon emissions, supplier capacity, and material recovery. Sourcing decisions made solely on cost minimization can overlook these chokepoints and impede growth.

WHY THIS WORKS:

Integrating chokepoints alongside costs in sourcing decisions reduces companies' exposure to supply shocks and stabilizes operations by ensuring continuity.

5

Design for refresh, reuse, recovery, and redeployment from day one

WHY THIS MATTERS:

AI hardware is resource-intensive and subject to accelerated upgrade cycles that pressure sourcing, operations, and end-of-life management. Without considering refresh, reuse, recovery, and redeployment, companies risk rapid hardware obsolescence, material bottlenecks, and higher lifecycle costs.

WHY THIS WORKS:

Designing for hardware refresh and circularity acts as a strategic hedge for companies, enabling them to extend the useful life of AI infrastructure, recover value at end-of-life, and reduce reliance on constrained inputs.

6

Scenario-test binding constraints before committing capital

WHY THIS MATTERS:

Value chain constraints emerge in different ways, depending on factors such as resource availability, supplier accessibility, demand growth, policy changes, and market shifts. Because of these dependencies, companies must tailor responses to actual impacts rather than expected conditions.

WHY THIS WORKS:

Scenario planning helps companies identify which constraints are likely to become binding first under different conditions and devise mitigative actions, reducing the likelihood of unexpected impacts and keeping projects on track as external situations shift.

7

Build deeper supplier and ecosystem partnerships

WHY THIS MATTERS:

AI scale relies on suppliers like miners, semiconductor manufacturers, utilities, construction companies, and circularity partners who face their own set of value chain challenges.

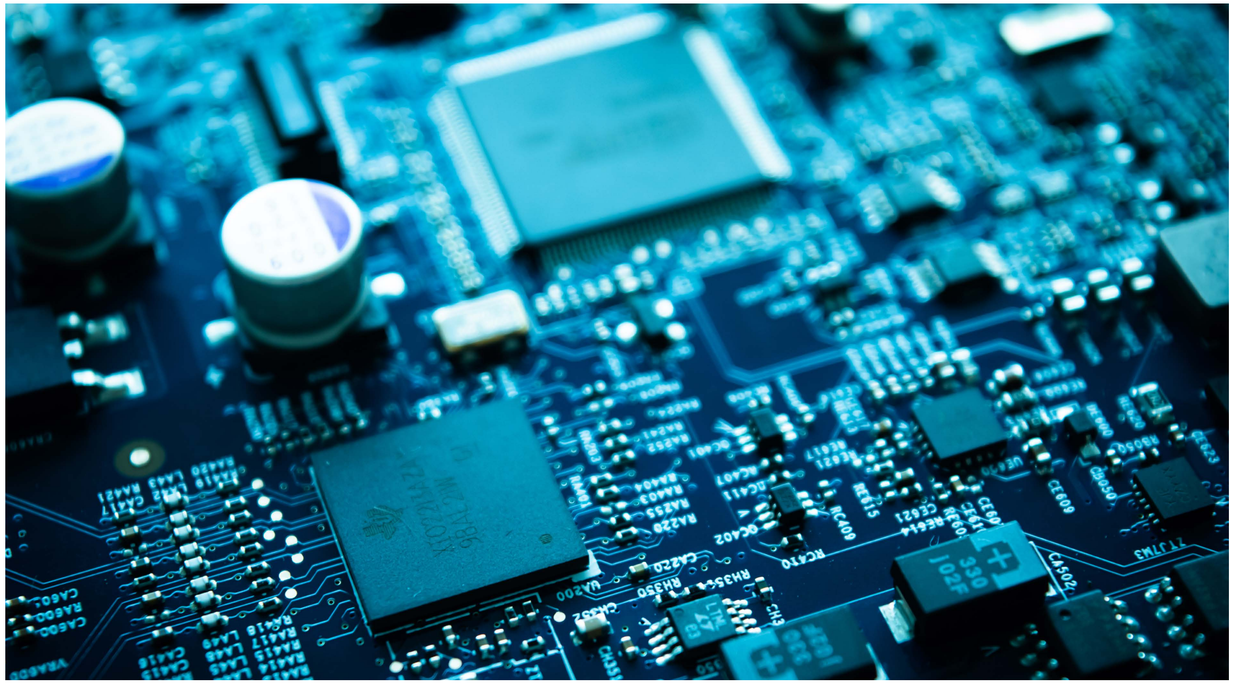
WHY THIS WORKS:

Deep partnerships with key suppliers ensure better visibility into their capacity, delivery timelines, and performance needs, reducing the risk of supply-driven project impediments and helping companies proactively address issues further upstream.

Future-proofing should not create new bottlenecks

Without careful planning, integrated value chain strategies can reduce exposure to one constraint while increasing pressure elsewhere. For example, behind-the-meter power, new transmission, alternative cooling systems, regional sourcing, circularity programs, and supplier diversification can all improve resilience—but they may also introduce new permitting, regulatory, land, interconnection, community, or logistics requirements.

The goal is not to avoid these strategies, but to sequence them earlier and plan them together. Future-proofing requires companies to identify second-order effects before decisions are locked in, so that resilience measures do not become new sources of delay.



Strategic control points: How AI infrastructure leaders are reducing dependency risk

Not every AI infrastructure leader will vertically integrate, and few will control every layer of the value chain. But many are increasing control over the inputs and partnerships that determine speed, reliability, and cost, whether that be by:

- Designing proprietary chips or servers while outsourcing manufacturing
- Securing long-term power arrangements
- Using behind-the-meter generation
- Partnering with utilities to operate onsite generation
- Investing in nuclear small modular reactors or other firm power options
- Pre-qualifying critical suppliers
- Standardizing modular designs
- Building circularity pathways for hardware recovery and reuse

The aim is not ownership for its own sake, but selective control over the decisions and dependencies most likely to affect deployment timelines, asset performance, stakeholder acceptance, and long-term resilience.

Conclusion: Scaling AI with confidence in a constrained world



ERM SUSTAINABILITY INSTITUTE

The next generation of AI leaders will be defined by their ability to orchestrate the physical systems that make the technology possible. Companies that recognize this shift early will be better positioned to move faster, allocate capital more effectively, reduce disruption, and build infrastructure that can adapt as conditions change.

Future-proofing the AI value chain is therefore not a defensive exercise, but a growth strategy. By integrating infrastructure planning with resource strategy, supplier resilience, circularity, permitting, and operational performance, companies can convert constraints into sources of advantage. In practice, success will come down to a small set of differentiators: speed to scale, access to power, capital efficiency, and flexibility. This will require not an incremental evolution of infrastructure strategy, but a transition to a new operating model defined by orchestration, coordination, and resilience.

ERM helps organizations make that shift by connecting decisions across the AI infrastructure system—from materials and manufacturing to power, water, siting, design, permitting, operations, and end-of-life—embedding sustainability and resilience at every point. In a constrained world, the winners will be those that build not just for today's AI demands, but for the physical realities that will define tomorrow's scale.

Integrated approach delivers



Speed to scale



Access to power



Capital efficiency



Flexibility

ERM
SUSTAINABILITY
INSTITUTE

Authors

Alison Drury
Andrew Angle
Olaf Schatteman
Ruzual Boparai

Contributors

Ben Ludtke
Conor Grieve
Jordan Snyder
Mackay Miller
Toby Whincup

Design

Aishwarya Grover
Bruce Stoddart

The ERM Sustainability Institute

The ERM Sustainability Institute is ERM's primary platform for thought leadership on sustainability. The purpose of the Institute is to define, accelerate, and scale sustainability performance by developing actionable insight for business. We provide an independent and authoritative voice to decode complexities. The Institute identifies innovative solutions to global sustainability challenges built on ERM's experience, expertise, and commitment to transformational change.

LinkedIn: linkedin.com/company/erm
Website: erm.com/insights

© Copyright 2026 by the ERM International Group Limited and/or its affiliates ('ERM'). All rights reserved.
No part of this work may be reproduced or transmitted in any form or by any means, without prior written permission of ERM.