THE AGI ADJACENCY PROBLEM

Compute, Energy, and Geopolitical Friction as Strategic Constraints

Thorsten Meyer

ThorstenMeyerAI.com

February 2026

Executive Summary

Hyperscalers will spend **$602 billion** on infrastructure in 2026, a 36% increase over 2025. NVIDIA's Blackwell GPUs are sold out through mid-2026, with a backlog of **3.6 million units**. Global datacenter electricity demand is projected to hit **945 TWh by 2030** — nearly 3% of global consumption. Export controls, rare-earth retaliations, and packaging bottlenecks are fragmenting the supply chain that makes all of it possible.

AGI discourse fixates on model capabilities. Executives and policymakers should fixate on what actually determines who deploys frontier AI at scale: chips, energy, packaging capacity, grid reliability, water, and geopolitical access. This is the **AGI adjacency problem** — the infrastructure layer that governs whether intelligence breakthroughs translate into operational advantage or remain trapped in research labs.

The implication is stark: AI strategy is now inseparable from infrastructure strategy. Organizations that treat AI as a software procurement problem will discover, too late, that they've made a hardware dependency bet they didn't understand.

Metric	Value
Hyperscaler capex 2026	**$602B (36% YoY increase)**
NVIDIA Blackwell backlog	**3.6M units, sold out through mid-2026**
Global datacenter power 2026	**~500 TWh (260 TWh US alone)**
TSMC CoWoS capacity end-2026	**120,000–130,000 wafers/month**
NVIDIA share of TSMC CoWoS	**>60% of 2025–2026 capacity**
H100 cloud rental price decline	**64–75% from peak**
Sovereign cloud market projection	**$154B → $823B by 2030**

This article maps the constraint landscape across compute, energy, geopolitics, and enterprise exposure — and provides a framework for strategic planning under infrastructure-governed conditions.

1. The Capability Race vs. the Infrastructure Reality

Public discourse frames AI as an intelligence race. Model benchmarks dominate headlines. Every quarter brings a new frontier model that edges past the previous leader on reasoning, coding, or multimodal tasks.

Operationally, none of it matters without a physical chain:

- **Advanced chips** — designed by NVIDIA, AMD, or custom ASICs; fabricated almost exclusively at TSMC's leading-edge nodes
- **Advanced packaging** — CoWoS and other 2.5D/3D packaging that integrates HBM memory with logic dies
- **Datacenter build-out** — physical facilities with power, cooling, networking, and physical security
- **Power contracts** — long-term electricity agreements sufficient for multi-megawatt deployments
- **Cooling and water systems** — liquid cooling infrastructure now essential for AI-density racks
- **Transmission infrastructure** — grid connections capable of delivering sustained high loads

Any break in this chain delays or reprices AI programs, regardless of model quality. A company with access to GPT-5 but insufficient inference capacity is no better positioned than a company running GPT-4 with abundant compute.

The Chain Has Already Broken — Multiple Times

The bottleneck has shifted repeatedly since 2023. First, HBM memory was the constraint. Then TSMC CoWoS packaging. Then GPU supply. Now, increasingly, the binding constraint is power — the most fundamental physical resource in the chain.

Bottleneck	Peak Period	Current Status
HBM3e memory	2023–2024	Easing; SK Hynix/Samsung expanding
TSMC CoWoS packaging	2024–2025	Still tight; fully booked through 2026
GPU supply (Blackwell)	2025–2026	Sold out through mid-2026
Datacenter power	2025–ongoing	Binding constraint in key regions
Grid transmission	2025–ongoing	Multi-year permitting timelines
Water/cooling capacity	2025–ongoing	Emerging constraint; regulatory pressure

"The AI race is not an intelligence race. It's a kilowatt race, a packaging race, and a permitting race — and no foundation model can solve any of them."

2. Compute Is Strategic, Not Merely Technical

Compute Access as Industrial Capacity

Compute access increasingly resembles strategic industrial capacity — more akin to steelmaking or semiconductor fabrication than to software licensing. Firms with privileged access to high-performance infrastructure gain compounding advantages:

- **Faster iteration cycles.** Training runs that take weeks at scale take months without scale. The difference compounds through more experiments, faster feedback, and earlier product launches.
- **Lower marginal inference costs.** H100 cloud rental prices have fallen 64–75% from their 2023 peaks, but only for those with reserved capacity or owned infrastructure. Spot pricing remains volatile.
- **Stronger product defensibility.** Companies that can serve inference at lower cost can offer capabilities competitors can't match economically.

The Dependency Trap

Risk Category	Manifestation	Enterprise Impact
Procurement delays	Blackwell sold out through mid-2026	12–18 month gaps in capacity
Pricing volatility	H100 spot prices swing 30–40% quarterly	Unpredictable AI unit economics
Provider concentration	3 hyperscalers control ~65% of cloud GPU	Single-provider lock-in risk
Capacity allocation	Priority goes to largest customers	SMEs and public sector rationed
Model availability	Providers can restrict/deprecate models	Sudden capability regression

If your organization's most important AI workflows run on a single cloud provider, you don't have an AI strategy. You have a vendor dependency.

3. Energy as the Hidden Governor

The Power Wall

As AI workloads scale, energy moves from a background variable to the binding constraint on AI deployment. The numbers are now impossible to ignore.

IEA projects US datacenter power demand at **260 TWh in 2026**. Gartner forecasts global datacenter electricity rising from **448 TWh in 2025 to 980 TWh by 2030**. AI-optimized servers will represent **44% of total datacenter power** by 2030, up from 21% in 2025.

The GB200, NVIDIA's latest Blackwell platform, draws over **1,200W per chip at peak**. Average rack density is expected to grow from **36 kW in 2023 to 50 kW by 2027**. These aren't marginal increases — they represent a fundamental change in what datacenters require from the grid.

Energy Economics Shape AI Economics

Energy Variable	AI Impact	Strategic Implication
Power price per MWh	Sets inference cost floor	Location arbitrage = competitive weapon
Grid reliability	Latency workloads disrupted	Redundancy increases cost
Permitting timelines	New capacity delayed 2–5 years	Supply-demand mismatch persists
Decarbonization targets	Constrains where to build	RE procurement = advantage
Transmission capacity	Limits power delivery	Grid bottleneck ≠ generation

"The constraint on frontier AI is no longer intelligence. It's electricity. And unlike intelligence, you can't download more of it."

Water: The Other Resource Constraint

Water is AI infrastructure's second physical dependency. Water absorbs heat **3,000 times more effectively** than air, making liquid cooling essential for AI-density deployments. The global liquid cooling market will surge from **$2.8 billion in 2025 to over $21 billion by 2032**, growing at 30%+ annually.

The tension is real: **45% of datacenter facilities globally** may face high water stress by the 2050s. Among facilities currently under construction, **30% are in regions where water scarcity is expected to intensify**. The EU expects to roll out regulation in 2026 requiring datacenter operators to meet minimum water-usage standards.

4. Geopolitical Controls and Supply Chain Fragmentation

The New Geography of Compute

Export controls, technology restrictions, and industrial policy are reshaping who gets access to frontier AI hardware. This is no longer a trade policy footnote — it's a structural feature of AI deployment.

The US has tightened semiconductor export controls repeatedly since October 2022. The latest rounds extend restrictions to foreign affiliates, with congressional pressure (the AI OVERWATCH Act, advanced January 2026) pushing to prohibit Blackwell chip sales to foreign entities of concern for two years. China has retaliated with licensing requirements on rare-earth oxides, metals, and magnet products — materials essential to the semiconductor supply chain.

The Fragmentation Map

Control Layer	Mechanism	Strategic Effect
US chip export controls	Performance thresholds, end-use	Regional capability divergence
China rare-earth controls	Export licensing for materials	Supply chain vulnerability
EU sovereignty requirements	CADA, €180M cloud tender	Mandatory local processing
Data localization mandates	National data residency	Multi-jurisdiction complexity
ITAR/EAR compliance	Defense/dual-use restrictions	Cross-border collaboration limits

Taiwan: The Single Point of Failure

TSMC fabricates virtually all frontier AI chips. NVIDIA has booked **over 60% of TSMC's CoWoS advanced packaging capacity** for 2025–2026. TSMC's monthly CoWoS capacity stands at ~80,000 wafers per month, projected to reach 120,000–130,000 by end of 2026 — but demand continues to outpace expansion.

Taiwan's centrality to the AI supply chain is the single largest geopolitical risk factor in AI infrastructure. TSMC is building facilities in Arizona and Japan, but these won't reach the scale or advanced process capability of Taiwan fabs for years. Diversification is real but slow.

Every AI strategy that assumes uninterrupted access to frontier chips is, implicitly, a bet on Taiwan Strait stability. Make that bet explicit in your risk register.

"Geopolitics doesn't pause for product roadmaps. The organizations that plan for fragmentation — not just efficiency — will be the ones still operating when supply chains shift."

5. Enterprise Exposure: Where the Risks Sit

The Cloud Abstraction Illusion

Many enterprises underestimate their infrastructure exposure because they consume AI as cloud APIs. The abstraction is real — you don't see the GPU, the cooling system, or the power contract. But the risks propagate through.

Exposure Channel	Risk Vector	Warning Signal
Pricing changes	Cloud GPU costs passed through	Quarterly cost increases >15%
Service-level variability	Inference latency spikes	SLA violations in critical periods
Model deprecation	Provider restricts models	Capability regression without alternative
Regional compliance	New sovereignty mandates	Workloads forced to migrate
Capacity allocation	Provider prioritizes large clients	Degraded throughput
Supply chain disruption	Chip shortage → cloud capacity	Extended waitlists

The Financial Exposure

AI cloud spending is now a capital planning issue, not just an IT budget line. Hyperscalers' capital intensity has reached staggering levels — Microsoft dedicating **45% of revenue** to capex, Oracle reaching **57%**. These costs will eventually flow through to customers.

The inference cost trajectory offers both promise and risk. Median LLM inference prices have fallen **200x per year** since 2024, but this masks significant variance. Frontier models remain expensive. The gap between frontier and commodity pricing creates a bifurcated market where capability access depends on budget scale.

Enterprise boards should scrutinize AI business cases for:

- **Realistic capacity assumptions.** Don't plan for compute availability you haven't contracted.
- **Downside scenarios.** Model a 50% cost increase and 30-day capacity disruption simultaneously.
- **Single-supplier concentration.** If >70% of AI workloads run on one provider, you're exposed.
- **Architecture flexibility.** Can you migrate workloads across providers within 90 days?

6. Public Sector and National Compute Capacity

The Sovereignty Spectrum

Governments face a strategic choice that most haven't confronted honestly: rely on global hyperscaler ecosystems or cultivate domestic capability. Full-stack sovereignty — owning everything from chip design to cloud platform — is unrealistic for all but two or three nations. The question is where on the spectrum to invest.

The EU has moved from strategy documents to physical infrastructure. The European Commission's **€180 million sovereign cloud tender** (expected to award by early 2026) demands quantifiable metrics for technological sovereignty. The **AI Factories initiative** pools EuroHPC supercomputing resources — LUMI in Finland, Leonardo in Italy — to let startups and researchers access training compute at reduced cost. France and Germany convened a **Summit on European Digital Sovereignty** in November 2025, launching a joint task force to report in 2026.

Selective Sovereignty: What's Feasible

Sovereignty Layer	Feasibility	Current Leaders
Secure processing for sensitive workloads	High	EU (CADA), UK, Japan
Regional compliance zones	High	EU, India, Brazil
Domestic chip design	Medium	US (CHIPS Act), EU, China
Advanced chip fabrication	Low	US, Taiwan, South Korea only
Public-interest compute programs	Medium	EU (AI Factories), UK (ARIA)
Domestic talent pipelines	Medium	All jurisdictions (varying)

Uncertainty label: The economic viability of broad public compute initiatives remains uncertain and highly jurisdiction-dependent. Most national programs will supplement, not substitute, commercial cloud.

"Sovereignty isn't about building everything yourself. It's about ensuring that no single external decision — a chip export ban, a price increase, a service policy change — can shut down your critical AI operations."

7. Environmental and Social Externalities

Infrastructure Friction Is Rising

AI infrastructure expansion doesn't happen in a vacuum. It happens in communities with power grids, water systems, and neighbors. Organizations that ignore these externalities are already paying the price.

Externality	Scale	Policy Response
Power competition	DCs competing with residential/industrial	Utility commission pushback
Water consumption	45% of facilities may face high stress	EU minimum standards (2026)
Land and permitting	Multi-year approval timelines	Expedited in some jurisdictions
Carbon emissions	Rising despite RE commitments	Scope 2/3 reporting mandates
Community opposition	NIMBYism vs. DC concentration	Social license requirements

Northern Virginia's Loudoun County — home to the world's densest cluster of datacenters — has become the canonical example. Local power utility Dominion Energy faces strain from demand that has doubled in a decade. Community opposition to new facilities is intensifying. The pattern is replicating in Dublin, Singapore, Amsterdam, and emerging hubs in the US Southeast.

The lesson: social license is now a constraint on AI infrastructure as real as kilowatts and cooling capacity. Organizations that engage local stakeholders early, share economic benefits, and mitigate environmental impacts will secure capacity faster than those that don't.

8. Strategic Planning Under Constraint

The Constraint-Governed Portfolio

Traditional AI planning starts with models and use cases. Infrastructure-governed planning starts with constraints and works backward to what's deployable. Executives should classify AI workloads by their constraint profile:

Workload Class	Primary Constraint	Architecture Requirement	Planning Horizon
Compute-critical	Hardware, inference cost	Long-horizon capacity contracts	18–36 months
Latency-critical	Location, network	Edge/regional deployment	6–12 months
Regulation-critical	Data governance, audit	Jurisdiction-aware architecture	Ongoing

Resilience-critical	Uptime, continuity	Multi-provider; model fallback	Immediate
Cost-critical	Budget, unit economics	Commodity models; on-premise	Quarterly

This shifts AI planning from model-centric roadmaps to infrastructure-governed portfolios. The question isn't "What's the best model?" — it's "What's the best model we can reliably operate, at acceptable cost, under the constraints we face?"

Financial Discipline for AI Infrastructure

• **Long-term infrastructure commitments.** Reserved capacity contracts of 1–3 years are becoming standard for serious AI deployments.

• **Variable inference cost risk.** Budget for ±50% variance in inference costs over a 12-month horizon.

• **Productivity realization lag.** Goldman Sachs projects $7 trillion in AI-driven GDP impact over a decade, but enterprise-level returns may take 3–5 years to materialize.

• **Capex vs. opex trade-off.** On-premise infrastructure offers up to **18x cost advantage** per million tokens over API access — but requires capital, talent, and operational maturity.

9. Practical Implications and Actions

For Enterprise Leaders

1. Add infrastructure dependency mapping to AI governance. Track chip, cloud, energy, and network dependencies for every critical AI workflow. If you can't map the dependency chain from model to power grid, you can't manage the risk.

2. Negotiate capacity and price protections. For mission-critical AI use cases, reserved capacity with price caps isn't optional — it's risk management. Treat GPU capacity like energy futures.

3. Diversify deployment architecture. Multi-region is table stakes. Multi-provider is the next requirement. Target no more than 70% of AI workloads with any single provider.

4. Integrate energy planning with AI expansion. Include grid reliability stress tests and decarbonization constraints in datacenter site selection. Your sustainability team and your AI team need to talk.

5. Develop continuity playbooks. Predefined degradation modes when advanced model access is constrained. What happens if your frontier model provider goes down for 72 hours?

6. Model the financial downside. Every AI business case should include a scenario where compute costs increase 50% and capacity is constrained for 30 days.

For Policymakers and Public Sector Leaders

7. Establish geopolitical risk review for AI sourcing. Quarterly review of export controls, sanctions exposure, and compliance obligations. The regulatory landscape is moving faster than annual planning cycles.

8. Invest in selective sovereignty. Full-stack sovereignty is a fantasy for most nations. Target sovereign processing capacity for the 10–15% of workloads that genuinely require it.

9. Engage local stakeholders early. Communities that feel heard during planning approve projects faster than communities that feel bulldozed. Water, power, and economic benefit sharing should be part of every infrastructure proposal.

10. Monitor the "public-interest compute" trajectory. The EU AI Factories model, UK ARIA, and similar programs are experiments worth watching. If they prove viable, they may become templates for how democracies ensure broad access to AI infrastructure.

What to Watch Next

- New export-control rounds affecting advanced compute access
- Power market stress in key AI datacenter regions
- Emergence of "public-interest compute" policy programs

• Enterprise shift from single-model dependence to resilience-oriented AI stacks

The Bottom Line

The AGI adjacency problem is not a futurist concern. It's a 2026 planning constraint. Every organization deploying AI at scale is making bets on chip supply, energy availability, grid reliability, and geopolitical stability — whether they recognize those bets or not.

The capability race is real, but capabilities without infrastructure are academic. The winners of the next five years won't be the organizations with the best models. They'll be the organizations that secured the chips, contracted the power, diversified their suppliers, and built the governance to operate under constraint.

The AGI adjacency layer — compute, energy, and geopolitics — is where AI strategy meets physical reality. Plan for the infrastructure you can secure, not the intelligence you wish you had.

AI strategy without infrastructure strategy is just a presentation deck.

Thorsten Meyer is an AI strategy advisor who reads utility commission filings and TSMC earnings calls with equal enthusiasm — because the future of AI runs on kilowatts and wafer starts, not just parameters. More at ThorstenMeyerAI.com.

Sources

1. Gartner — Electricity Demand for Data Centers to Grow 16% in 2025 and Double by 2030 (November 2025)
2. IEA — Energy Demand from AI (2025)
3. Goldman Sachs — Why AI Companies May Invest More than $500 Billion in 2026 (2025)
4. IEEE ComSoc — Hyperscaler Capex > $600 Bn in 2026 (December 2025)
5. SM Daily Press — NVIDIA Blackwell B200/GB200 Sold Out Through Mid-2026 (December 2025)
6. DigiTimes — TSMC CoWoS Capacity and NVIDIA Booking (December 2025)
7. TrendForce — TSMC CoWoS-L/S Fully Booked (December 2025)
8. Epoch AI — LLM Inference Price Trends (2025)
9. S&P; Global — Beneath the Surface: Water Stress in Data Centers (2025)
10. MSCI — When AI Meets Water Scarcity (2025)
11. Congressional Research Service — US Export Controls and China: Advanced Semiconductors (2025)
12. Mayer Brown — Administration Policies on Advanced AI Chips Codified (January 2026)
13. European Commission — €180M Sovereign Cloud Tender (October 2025)

14. McKinsey — Accelerating Europe's AI Adoption: Sovereign AI Capabilities (2025)
15. Introl — Hyperscaler CapEx Hits $600B in 2026 (January 2026)
16. Goldman Sachs — AI to Drive 165% Increase in Data Center Power Demand by 2030 (2025)
17. Fusion Worldwide — Inside the AI Bottleneck: CoWoS, HBM, and Capacity Constraints (2025)
18. CSIS — Sovereign Cloud–Sovereign AI Conundrum (2025)
