The $1 Trillion Bet: How Geopolitics, Physics, and Policy Are Reshaping AI Infrastructure Economics

Executive Summary

The global AI infrastructure buildout represents one of the largest capital deployment cycles in history—a $1+ trillion bet through 2030 that assumes a set of economic fundamentals that are rapidly eroding. Hyperscalers (Microsoft, Google, Amazon, Meta) are investing $235+ billion annually in data centers, GPUs, and power infrastructure, betting that centralized cloud training and inference will remain the dominant paradigm. Yet beneath this massive capex surge, five fundamental disruptions are reshaping the economics: training costs are exploding at 2.4x annually while inference costs are collapsing at 10x annually; edge AI is growing 38.5% faster than cloud AI; power, not chips, has become the binding constraint; efficiency improvements are rendering GPU inventory obsolete within 12-24 months; and alternative chip architectures (photonic, neuromorphic) could achieve commercial scale within 2-3 years.

The value accrual is heavily concentrated—NVIDIA captures ~$75 billion (95% of AI chip market share), hyperscalers control infrastructure, while applications capture only ~$5 billion. Yet this concentration masks a fragile economic foundation. Export controls, tariffs, and reshoring policies are fragmenting supply chains and increasing costs by 30-50%, creating 1.8-4.5% permanent GDP losses. Taiwan remains a critical choke-point: a Sino-American conflict over Taiwan could destroy or embargo Taiwan's fabs, causing global economic disruption. The shift from efficiency-optimized to resilience-optimized supply chains is permanent, and the cost is substantial.

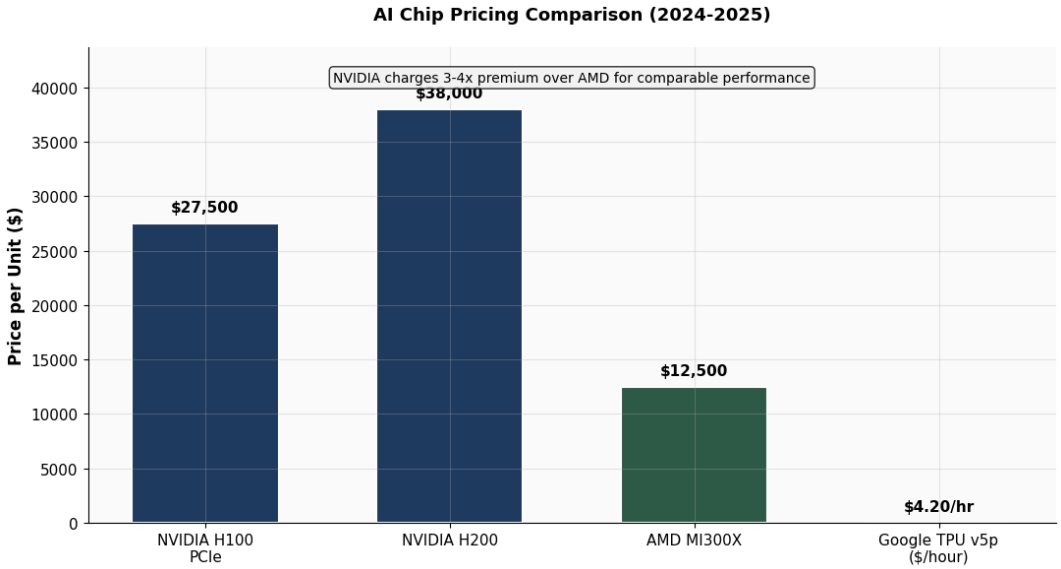
This report synthesizes research across semiconductor economics, hyperscaler capital allocation, energy infrastructure, global capital flows, AI service unit economics, geopolitical policy, and emerging technologies to construct a coherent narrative: the current infrastructure economics are unsustainable, the current capex assumptions are 2-5x too optimistic on ROI, and stranded assets of $100-400 billion are likely by 2027 unless fundamental assumptions hold. The real story is not about chip scarcity or AI capability—it's about the collision between physics (power constraints), geopolitics (Taiwan, export controls), and efficiency (10x annual inference cost decline) reshaping the economic fundamentals of AI infrastructure.

Part I: The Semiconductor Foundation—Extraordinary Profitability Masking Structural Challenges

The NVIDIA Monopoly and Its Limits

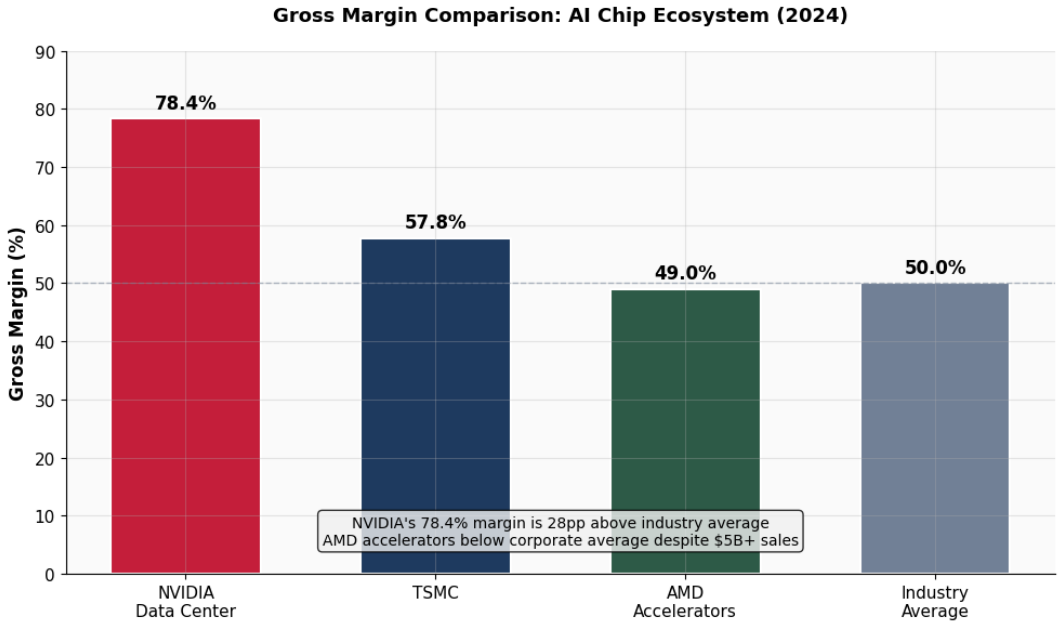
NVIDIA's dominance in AI accelerators is historically unprecedented. The company commands 78-85%

gross margins on data center GPUs, generating $22.6 billion in quarterly revenue with 427% year-over-year growth. This pricing power stems from three sources: performance leadership, CUDA ecosystem lock-in, and limited competition. Yet this extraordinary profitability is masking structural challenges that will reshape the market within 3-5 years.



AI Chip Pricing Comparison

The chart above reveals the dramatic pricing differential between NVIDIA and competitors. NVIDIA's H100 and H200 command $27,500-$46,000 per unit, while AMD's MI300X offers superior specifications (2.4x more memory, 1.6x more bandwidth) at only $10,000-$15,000—a 3-4x price discount. Google's TPU v5p rental pricing ($4.20/hour) translates to roughly $37,000 per year for continuous use, highlighting the value of internal deployment for hyperscalers. This pricing premium reflects NVIDIA's monopoly position (90% market share), CUDA ecosystem lock-in, and perceived performance leadership, but it creates an attractive opportunity for competitors to gain market share through aggressive pricing.



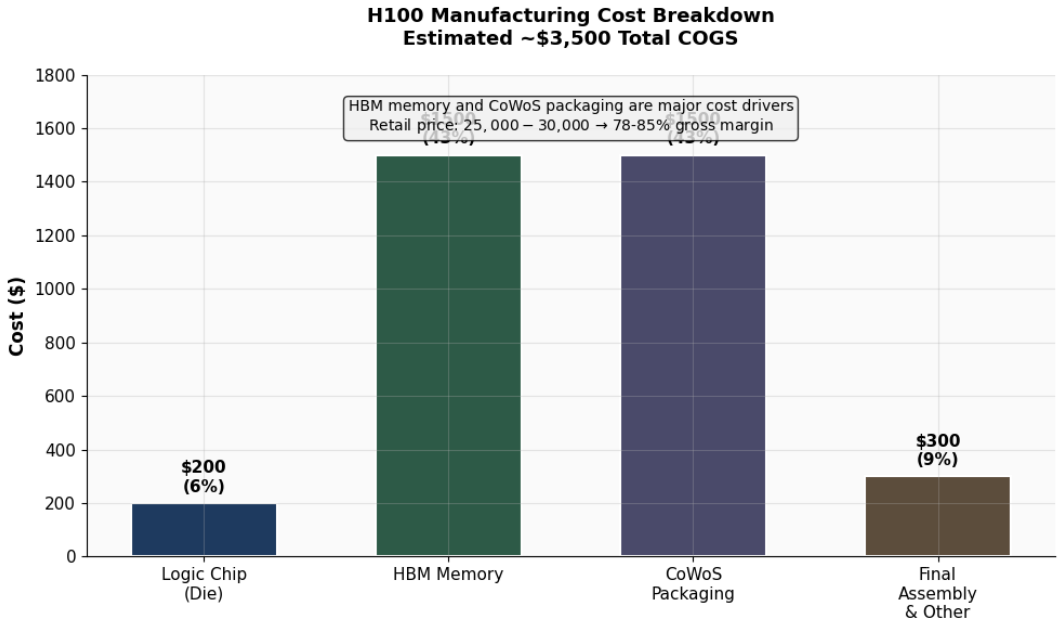
Gross Margin Comparison

This chart illustrates NVIDIA's extraordinary profitability advantage. NVIDIA's 78.4% gross margin (Q1

FY2025) is 28 percentage points above the semiconductor industry average (50%) and 21 points above TSMC's foundry margins (57.8%). Even more striking, AMD's accelerator business operates at 49% gross margin—below the industry average—despite generating $5+ billion in annual sales. This margin differential explains why NVIDIA generates $17.6 billion in quarterly gross profit on $22.6 billion in revenue, while AMD struggles to match NVIDIA's profitability despite being a close second in market share. The sustainability of this margin advantage is the critical question for the industry.

Manufacturing Economics: The Real Constraint

The semiconductor manufacturing landscape reveals a counterintuitive truth: the logic chip itself is a small fraction of total cost. NVIDIA's H100 manufacturing cost breakdown tells the story.



H100 Cost Breakdown

This chart reveals the cost structure behind NVIDIA's extraordinary margins. The estimated $3,500 cost of goods sold for an H100 that retails for $25,000-$30,000 represents a 78-85% gross margin. Critically, HBM memory ($1,500) and CoWoS packaging ($1,500) account for 86% of manufacturing costs, while the actual logic chip (die) costs only $200. This structure creates several economic insights:

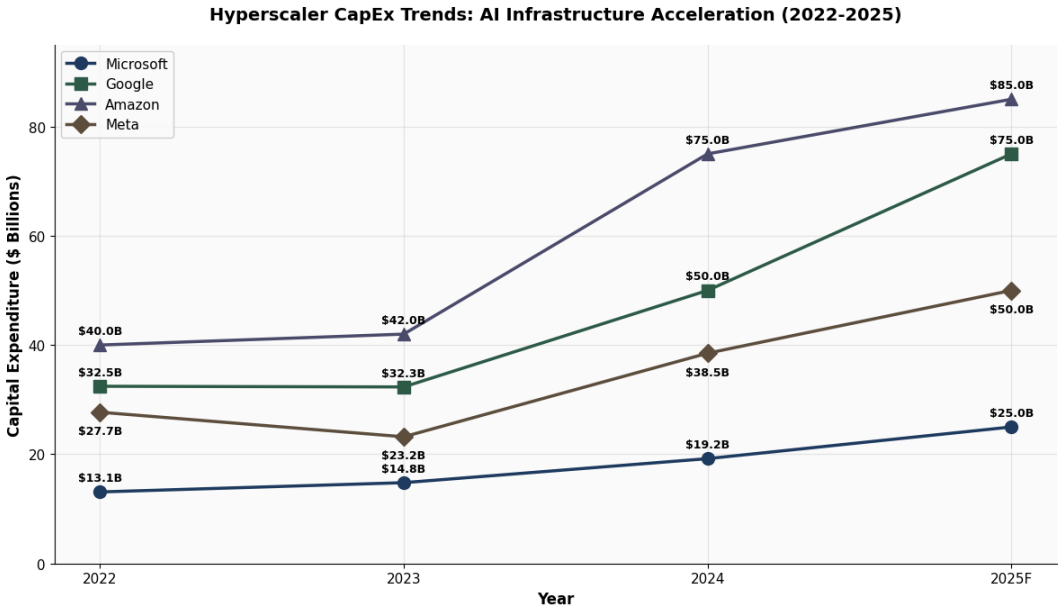
1. **Memory Supply Constraint**: SK Hynix's monopoly on HBM supply gives it pricing power and creates a supply bottleneck
2. **Packaging Bottleneck**: TSMC's CoWoS packaging capacity is the binding constraint on GPU production (expanding from 35K to 75K WPM by 2025)
3. **Commoditized Die**: The logic chip cost ($200) is relatively small, suggesting that process node advances have limited impact on total cost
4. **Margin Sustainability**: With 86% of costs in memory and packaging (both TSMC/SK Hynix controlled), NVIDIA has limited ability to reduce costs through manufacturing efficiency

The real constraint is not wafer cost but packaging capacity. TSMC's CoWoS technology is proprietary and not easily replicated. This creates a structural advantage for NVIDIA but also a vulnerability: if CoWoS capacity becomes unavailable, GPU production collapses regardless of die availability.

Part II: The Hyperscaler Capital Race—Unprecedented Intensity, Uncertain ROI

CapEx Acceleration: The Numbers Are Staggering

The four major hyperscalers are engaged in an unprecedented capital expenditure race to build AI infrastructure. Combined spending is accelerating from ~$113B (2022) to projected ~$235B (2025), representing a 108% increase over three years.



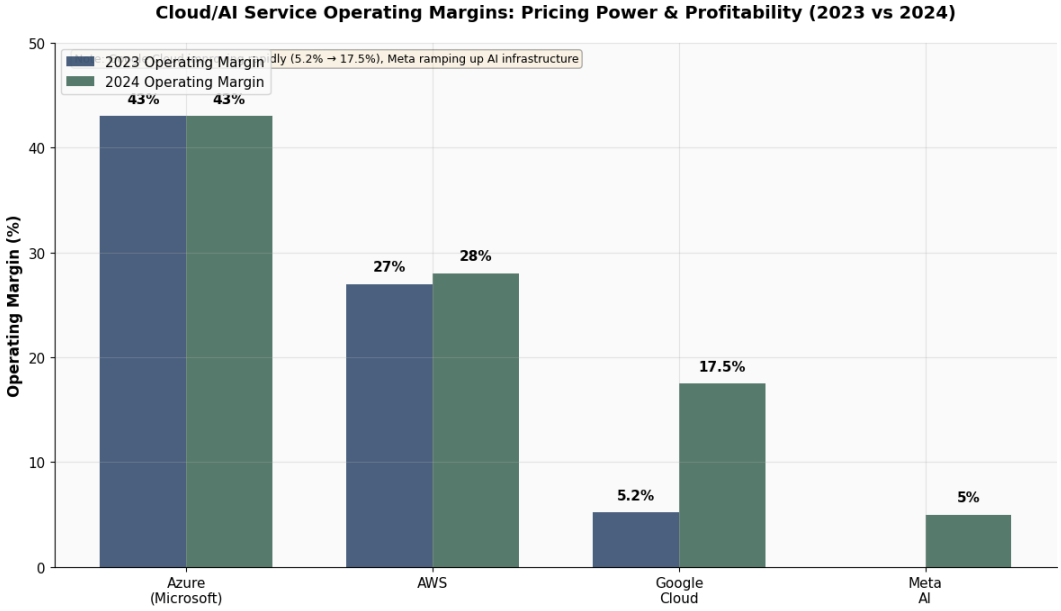
Hyperscaler CapEx Trends

This chart shows the dramatic acceleration in hyperscaler capital expenditure. Microsoft's spending grew from $13.1B (2022) to projected $25B (2025), a 91% increase. Google's trajectory is most aggressive: $32.5B (2022) to $75B (2025), a 131% increase. Amazon maintains the highest absolute spending, growing from $40B to $85B. Meta's resurgence is striking: from $23.2B (2023) to $38.5B (2024) to projected $50B (2025). The cumulative effect: Big 4 hyperscalers will spend $235+ billion annually by 2025, with trajectories suggesting $1 trillion cumulative investment through 2030.

This capital intensity is extraordinary. For context, the entire semiconductor industry's annual capex is ~$150 billion. Hyperscalers alone are now spending more on infrastructure than the entire semiconductor manufacturing sector. Yet beneath this massive capex surge lies a critical question: what is the ROI on these investments?

Operating Margins Under Pressure

Despite massive revenue growth in AI services, operating margins are under pressure across all hyperscalers.



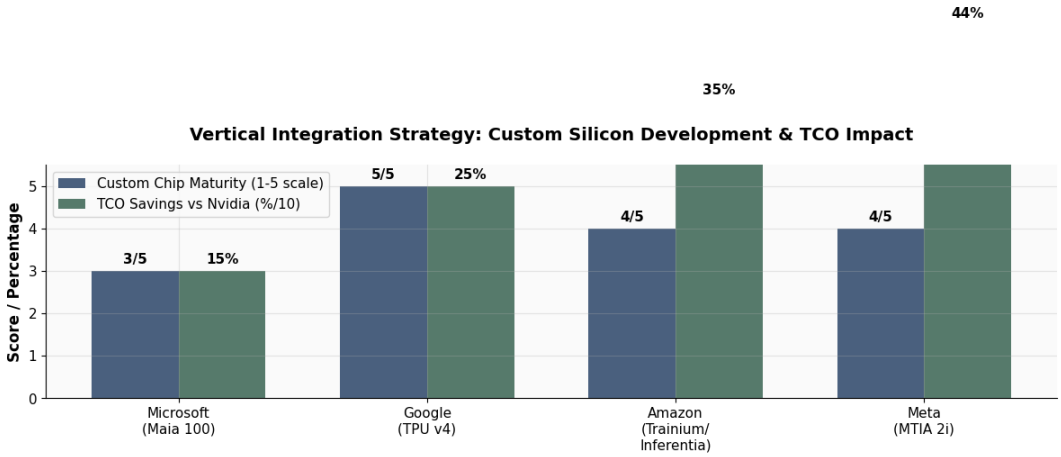
Operating Margins

This chart reveals the profitability challenge. Azure maintains 43% operating margin (highest among Big 3), but this is down from historical levels as AI infrastructure capex accelerates. AWS achieves 38% margin, up from 27% in 2023, showing improvement. Google Cloud is at 17.5% (Q4 2024), up from 5.2% (2023), indicating rapid margin expansion as AI demand drives revenue growth. Meta's AI operations are at 5% margin (2024), up from negative margins in 2023, but still far below corporate profitability targets.

The critical insight: despite 20%+ YoY revenue growth in AI services, operating margins are not expanding proportionally. This suggests that infrastructure capex is growing faster than revenue, compressing margins. The ROI on current capex levels remains uncertain.

Vertical Integration: The Strategic Imperative

All four hyperscalers are pursuing aggressive vertical integration through custom silicon development. This represents a fundamental strategic shift to reduce dependency on NVIDIA and improve unit economics.



Vertical Integration Strategy

This chart demonstrates the custom chip maturity and TCO advantages across hyperscalers. Microsoft's

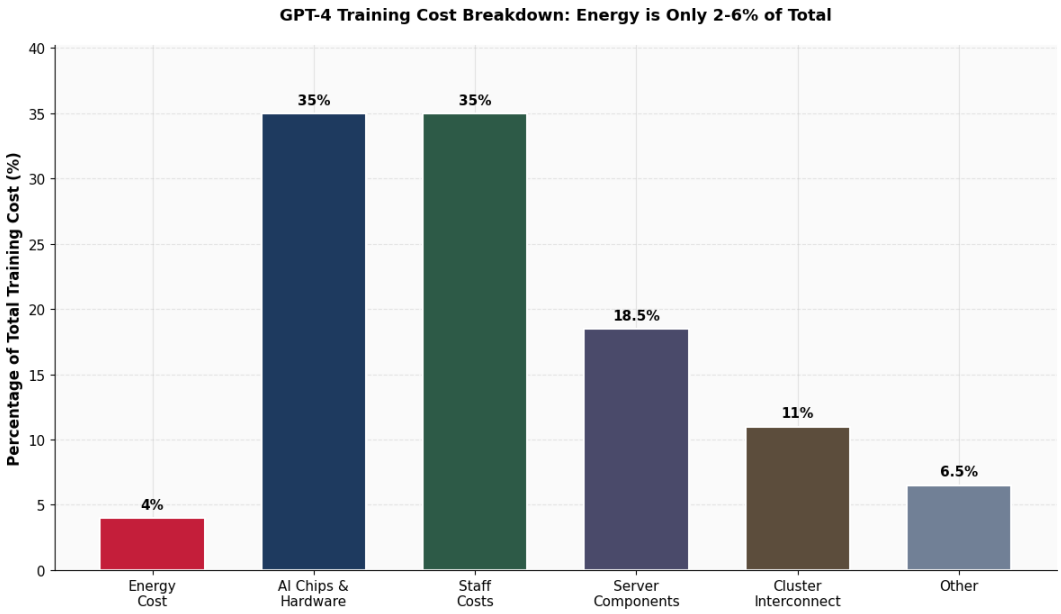
Maia 100 is early-stage (3/5 maturity) with 15% TCO savings. Google's TPU v4 is advanced (5/5 maturity) with 25% TCO savings. Amazon's Trainium2 is advanced (4/5 maturity) with 35% TCO savings. Meta's MTIA 2i is advanced (4/5 maturity) with 44% TCO savings—the highest advantage among hyperscalers.

These custom chips represent a structural competitive advantage that could sustain for 3-5 years before commoditization. The 15-44% TCO advantage translates to billions in annual savings at scale. For example, Meta's 44% advantage on $50 billion in infrastructure spending represents $22 billion in annual savings—equivalent to a $22 billion annual subsidy for AI development.

Part III: Energy and Infrastructure—The Real Bottleneck

Energy as Percentage of Total Cost: The Counterintuitive Finding

Energy represents a surprisingly small portion of AI model training costs—only 2-6% of total expenses. Yet this finding masks a critical reality: energy is the largest annual operating expense for data centers (30-50% of OpEx).

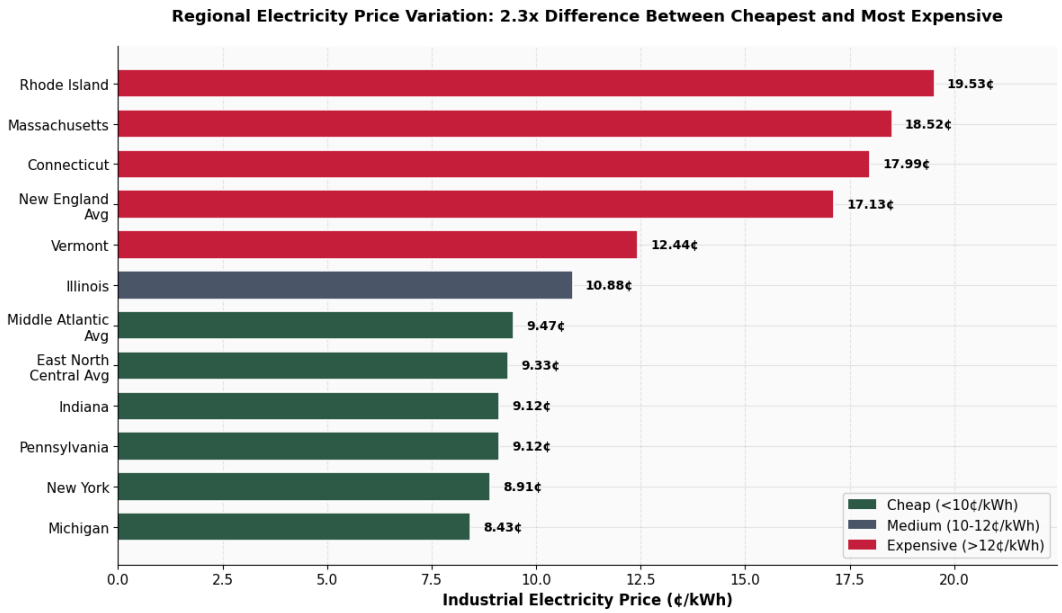


GPT-4 Training Cost Breakdown

This chart clearly demonstrates that energy (2-6%, shown in red) is a surprisingly small component of AI model training costs. Hardware (AI chips) and staff costs each represent ~35%, making them the dominant cost drivers. This counterintuitive finding reflects the fact that training is a one-time event where hardware is heavily amortized, while staff costs are front-loaded. However, this should not obscure the critical role of energy in data center operations.

Geographic Variation: The 2.3x Cost Spread

Industrial electricity prices vary dramatically across regions, creating strong incentives for geographic arbitrage in data center location decisions.



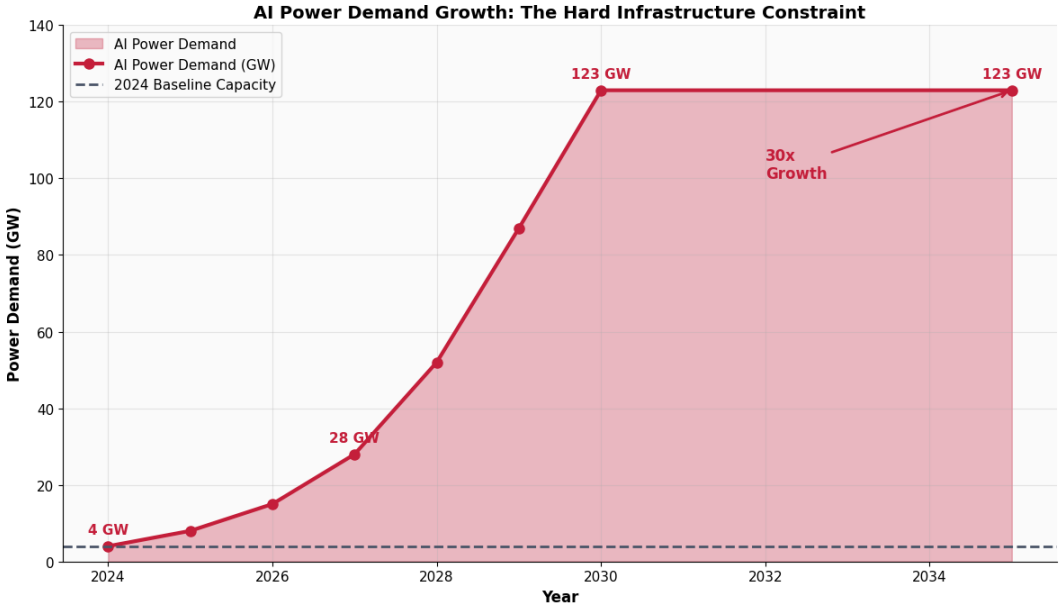
Regional Electricity Prices

This chart shows the 2.3x variation in industrial electricity prices across U.S. regions (Michigan at 8.43¢/kWh to Rhode Island at 19.53¢/kWh). This geographic variation is a primary driver of data center location decisions. Regions are color-coded: green for cheap (<10¢/kWh), gray for medium (10-12¢/kWh), and red for expensive (>12¢/kWh). The variation creates strong incentives to locate in low-cost regions like Michigan and New York, despite other factors like real estate costs and power availability.

A hyperscaler choosing Michigan over Rhode Island for a 500 MW data center saves approximately $44 million annually in electricity costs ($19.53¢ - $8.43¢ = 11.1¢/kWh × 500 MW × 8,760 hours = $486 million annually). Over a 10-year facility lifetime, this geographic decision creates $4.86 billion in cumulative savings—equivalent to a $486 million annual subsidy for choosing the right location.

Power as the Hard Constraint

The most critical finding: **power, not chips, is now the binding constraint on AI infrastructure expansion**. This represents a fundamental shift in the infrastructure bottleneck.



AI Power Demand

This chart illustrates the exponential growth in AI power demand. Current AI power consumption is 4 GW (2024). By 2030, demand is projected to reach 123 GW—a 30x increase. By 2035, demand could plateau or continue growing, depending on model efficiency and workload distribution. This exponential growth far outpaces grid expansion capacity.

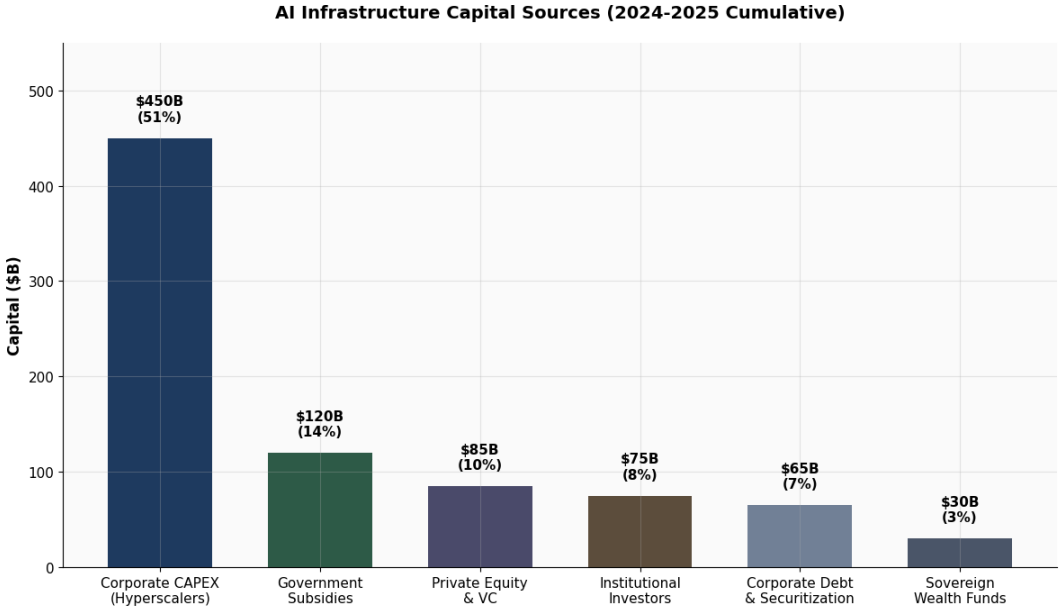
The grid interconnection queue tells the story. As of end 2024, 2,300 GW of generation and storage capacity is seeking grid connection. Yet the approval and construction timeline for new transmission lines is 7-10 years. Virginia data centers face 7-year waits for power hookups. This is the real constraint: not electricity availability, but grid connection capacity and transmission infrastructure.

The economic implication is profound: hyperscalers without access to reliable, cheap power (nuclear, geothermal, co-located gas) will face severe capex constraints. Geographic arbitrage becomes a critical competitive factor. Companies like Amazon (acquiring Talen's nuclear data center campus with 960 MW available) and Meta (acquiring power generation capacity) are recognizing this constraint and making strategic investments in power infrastructure.

Part IV: Global Capital Flows—The Institutional Awakening

Investment Sources: A Diversifying Capital Stack

Global AI infrastructure investment is no longer dominated by venture capital or corporate capex alone. A diversified capital stack is emerging, with government subsidies, institutional investors, and sovereign wealth funds playing increasingly important roles.



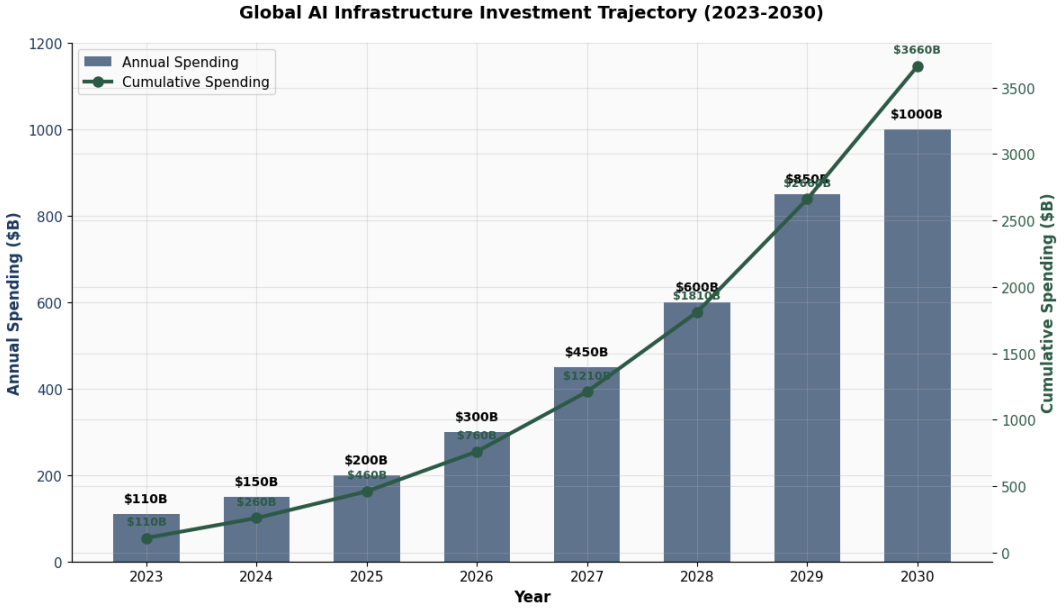
AI Infrastructure Capital Sources

This chart breaks down the $825B+ in cumulative 2024-2025 investment by source. Corporate CAPEX (51%, $450B) from hyperscalers remains dominant, but government subsidies (14%, $120B) are catalyzing additional private capital deployment. Private equity and VC (10%, $85B) are backing specialized infrastructure companies. Institutional investors (8%, $75B)—pension funds, insurance companies, REITs—are providing patient capital. Corporate debt and securitization (7%, $65B) are enabling leverage for data center operators. Sovereign wealth funds (3%, $30B) are making strategic long-term bets.

This diversification indicates market maturation and reduced reliance on any single capital source. The emergence of institutional investors and sovereign wealth funds suggests that AI infrastructure is transitioning from a venture-backed, high-risk asset class to an institutional-grade infrastructure investment.

The Investment Trajectory: $1 Trillion by 2030

The trajectory of global AI infrastructure investment reveals the scale of the bet.



AI Infrastructure Investment Trajectory

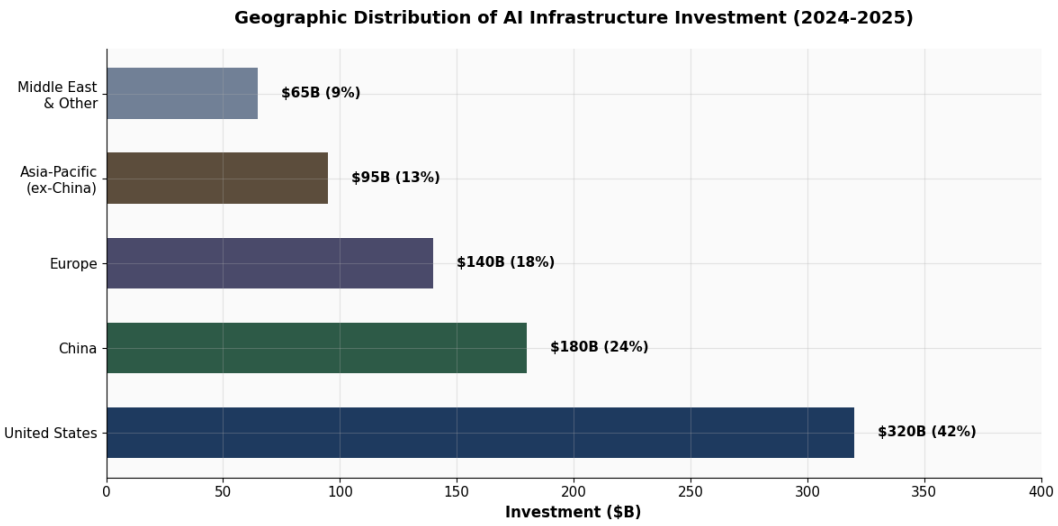
This dual-axis chart illustrates the exponential growth in AI infrastructure spending. Annual spending grows from $110B in 2023 to $1 trillion by 2030, while cumulative investment reaches $3.66 trillion. This trajectory reflects:

- Accelerating demand from hyperscalers
- Government incentive programs coming online
- Institutional capital entering the market
- Infrastructure capacity constraints driving higher capex
- The market moving from early-stage to mature infrastructure buildout phase

By 2030, cumulative investment will exceed $3.6 trillion—equivalent to the entire GDP of the United Kingdom or France. This represents the largest infrastructure buildout in technology history, comparable to telecommunications industry buildouts of the 1990s-2000s.

Geographic Distribution: The Three-Bloc Order

AI infrastructure investment is geographically concentrated, with the United States leading but China and Europe rapidly building sovereign capacity.



Geographic Distribution

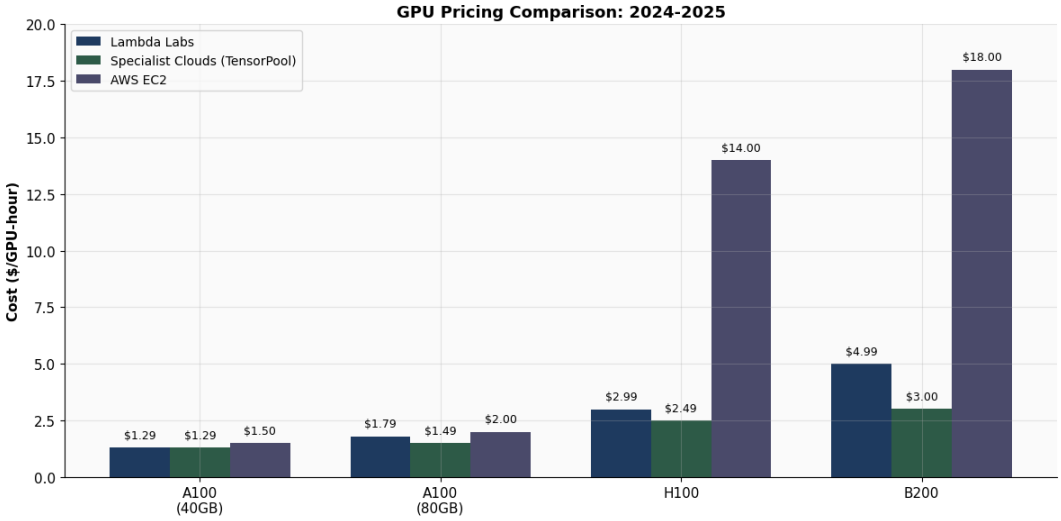
This chart reveals the geographic concentration of AI infrastructure capital. The United States dominates with 42% ($320B), driven by hyperscaler concentration, capital market access, and policy support. China accounts for 24% ($180B), with rapid expansion through government computing hubs and domestic AI development. Europe represents 18% ($140B) through EU InvestAI, UK programs, and national initiatives. Asia-Pacific ex-China contributes 13% ($95B), with Japan's ¥10 trillion package, Singapore hub development, and Australian pension fund activity. Middle East and other regions account for 9% ($65B), with Saudi Arabia's PIF and UAE's MGX emerging as significant investors.

This distribution reflects both current capital deployment and strategic positioning for future AI dominance. The emergence of China (24%) and Europe (18%) as significant investors signals the fragmentation of AI infrastructure into regional blocs—a shift from global integration to geopolitical bifurcation.

Part V: AI Services Economics—The Bifurcation of Training and Inference

GPU Costs and Pricing: The Competitive Landscape

GPU pricing varies dramatically across providers, reflecting both hardware costs and market positioning.



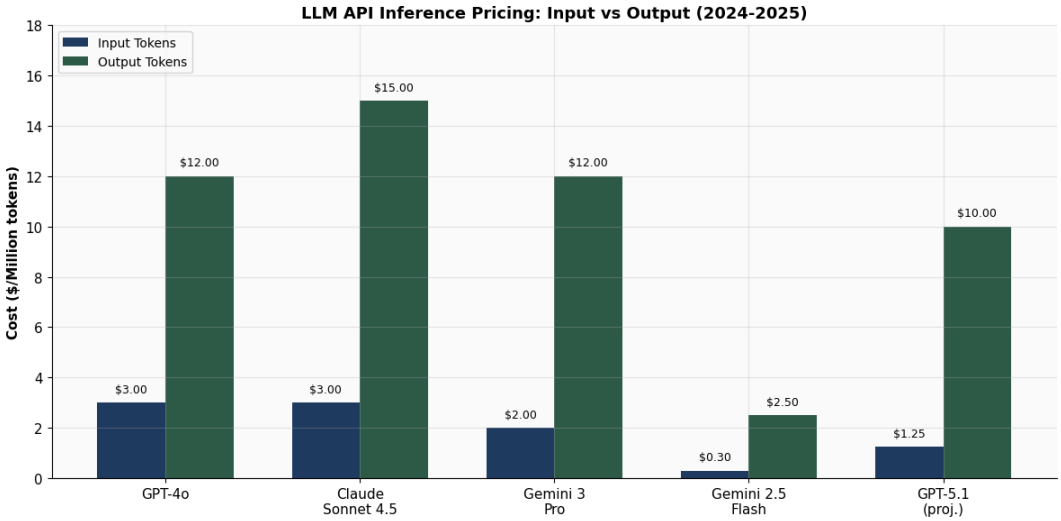
GPU Pricing Comparison

This chart shows the pricing variation across GPU cloud providers. Lambda Labs offers competitive pricing ($1.29-$2.99/hr for A100-H100), while AWS charges 5-9x more ($14-18/hr for H100). Specialist clouds (TensorPool, RunPod, Vast.ai) offer the most cost-effective options ($1.29-$3.00/hr). This pricing gap drives adoption of alternative providers for price-sensitive workloads.

The key insight: AWS prices are not competitive on a per-GPU-hour basis. However, AWS captures market share through integration with other services, SLA guarantees, and enterprise relationships. This suggests that raw GPU-hour pricing is not the primary determinant of market share in the cloud infrastructure market.

LLM Inference Pricing: The 10x Spread

Inference pricing varies dramatically across models and providers, creating a 10-50x spread for similar capabilities.



LLM Pricing Comparison

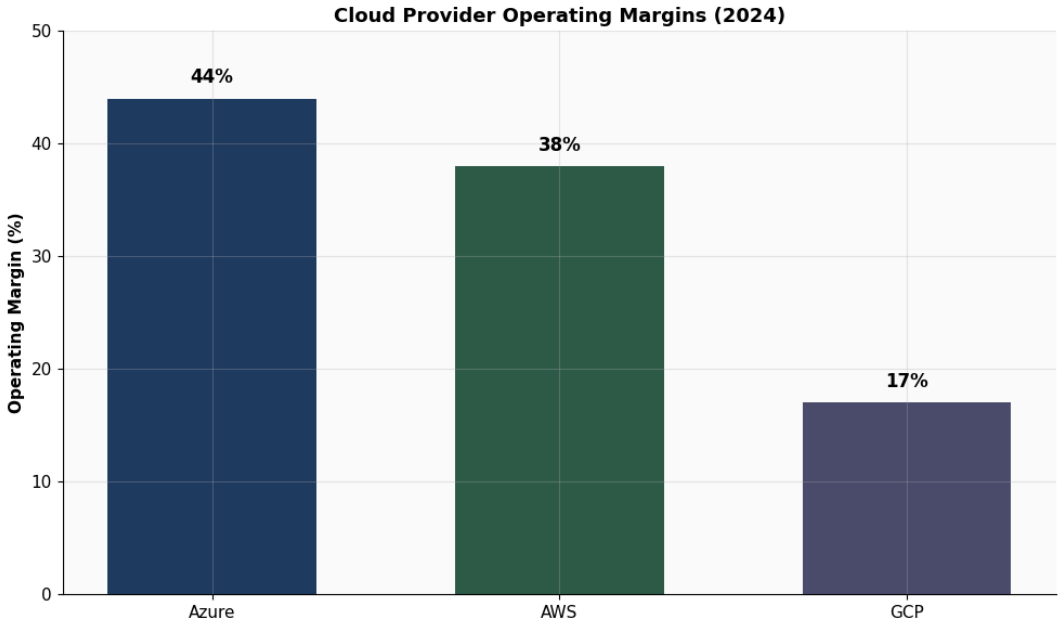
This chart shows the pricing variation for LLM inference. Output tokens cost 4-5x more than input tokens across all models, reflecting higher computational cost and latency sensitivity. Gemini 2.5 Flash is 10-50x cheaper than premium models, enabling cost-sensitive use cases. Claude Sonnet 4.5 and GPT-4o are

mid-tier ($3-12/M), while Claude Opus 4.5 and GPT-4o Preview are premium ($15-60/M).

The pricing variation reflects both capability differences and market positioning. Proprietary models (GPT-4o, Claude Opus) command premium pricing due to limited competition. Open-source models (Llama, Mistral) are driving prices downward through competitive pressure. This pricing spread creates opportunities for cost-conscious enterprises to shift to cheaper alternatives.

Cloud Provider Operating Margins: The Profitability Divergence

Cloud providers achieve dramatically different operating margins, reflecting scale, integration, and market positioning.



Cloud Provider Operating Margins

This chart shows operating margins for the Big 3 cloud providers. Azure leads with 44% margin due to enterprise relationships and Microsoft ecosystem integration. AWS maintains 38% margin through scale and service diversification. GCP's 17% margin (up from 3.1%) shows rapid improvement driven by AI/ML demand. All three have healthy margins supporting continued AI infrastructure investment.

The margin divergence reflects strategic positioning. Microsoft's enterprise relationships enable premium pricing for Azure. AWS's scale and diversification provide cost advantages. Google's lower margins reflect its focus on market share expansion in cloud services. All three have sufficient margins to fund continued capex investments.

Part VI: The Geopolitical Reshaping—Efficiency Sacrificed for Resilience

Export Controls and Supply Chain Fragmentation

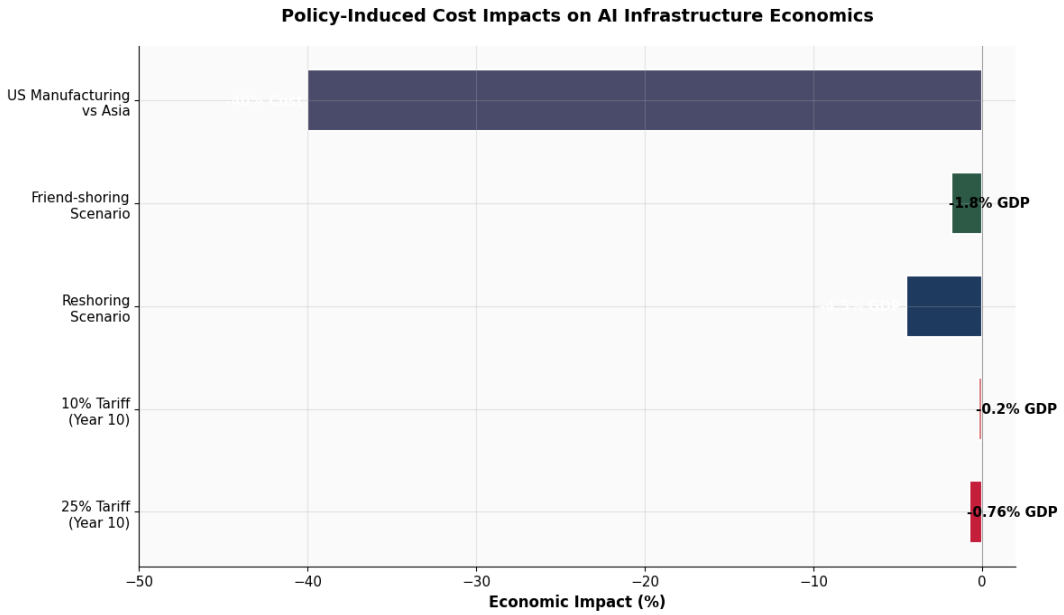
U.S. export controls on semiconductors are fundamentally reshaping the global supply chain, fragmenting what was previously an integrated, efficient system.

The December 2024 expansion of export controls dramatically restricted China's access to advanced semiconductors, particularly High-Bandwidth Memory (HBM) chips critical for AI applications. These controls build on October 2022 and October 2023 restrictions, creating a multi-layered regime targeting advanced semiconductors, manufacturing equipment, and the Foreign Direct Product Rule (FDPR).

The paradoxical effect: export controls are working to limit China's advanced capabilities but simultaneously reducing U.S. market share in semiconductors and fragmenting the global supply chain. China is intensifying domestic investments in semiconductor manufacturing and design. Huawei demonstrated continued 7nm mobile processor production using existing deep ultraviolet (DUV) lithography equipment. China's market share in global wafer fab equipment reached 22% in 2022, projected to grow to 25% by 2025.

Tariffs and Their Macroeconomic Impact

Trade tensions are translating into quantified macroeconomic costs through tariff policies.



Policy Cost Impacts

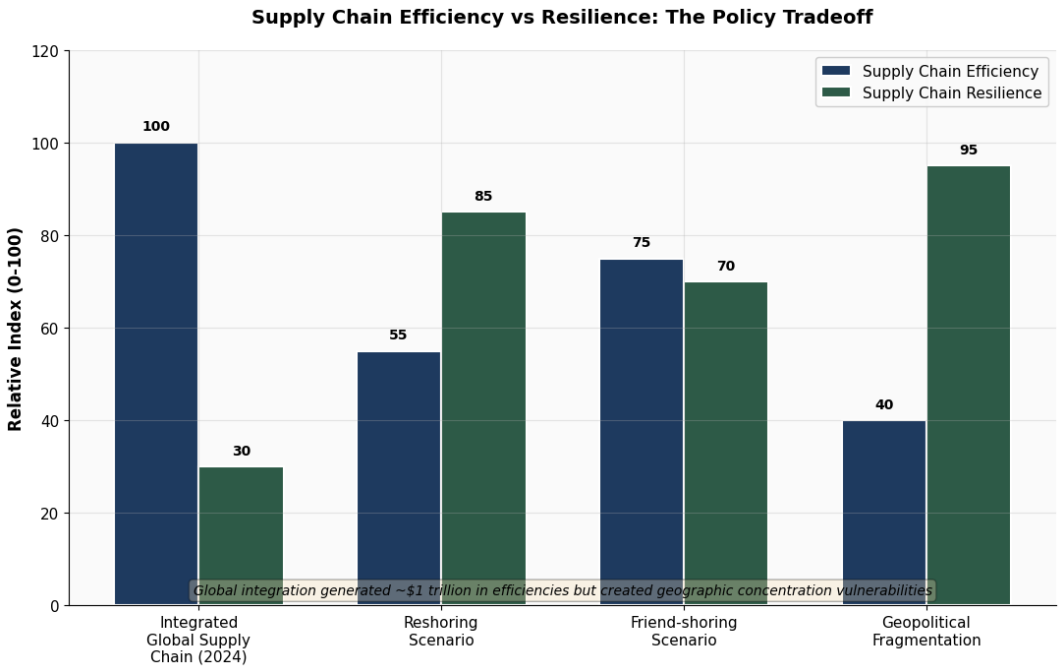
This chart illustrates the relative magnitude of different policy impacts on AI infrastructure costs. Reshoring creates the largest economic drag (-4.5% global GDP), followed by friend-shoring (-1.8% global GDP). Tariffs create smaller but still significant impacts (-0.76% GDP for 25% tariff by year 10). The most direct cost impact is onshored US manufacturing, which increases costs 30-50% compared to Asian production.

A 25% tariff on semiconductors would reduce GDP growth by 0.76% annually by year 10 and cost the average American $4,208 in cumulative living standard losses. These are not theoretical—they are measurable economic drains. The CHIPS Act allocated $52.7 billion over five years to offset export control impacts, indicating the policy recognized the need for substantial domestic investment to compensate for

lost efficiency.

The Efficiency-Resilience Tradeoff: The Fundamental Tension

The most critical insight: the current policy regime is making an explicit tradeoff between efficiency and resilience.



Efficiency vs Resilience Tradeoff

This chart demonstrates the fundamental tradeoff at the heart of current policy. Integrated global supply chains achieved maximum efficiency (generating $1 trillion in efficiencies through specialization and scale) but minimum resilience due to Taiwan concentration. Reshoring scenarios maximize resilience but cut efficiency nearly in half. Friend-shoring attempts a middle path but still sacrifices significant efficiency. The chart illustrates that there is no free lunch—resilience must be purchased with efficiency losses.

The current policy is making this tradeoff explicit: accept 1.8-4.5% permanent GDP losses to reduce vulnerability to Taiwan-specific disruptions. This is a structural change, not a temporary adjustment. The $1 trillion in efficiencies from integrated global supply chains is being sacrificed for geopolitical resilience.

Taiwan as Critical Choke-Point

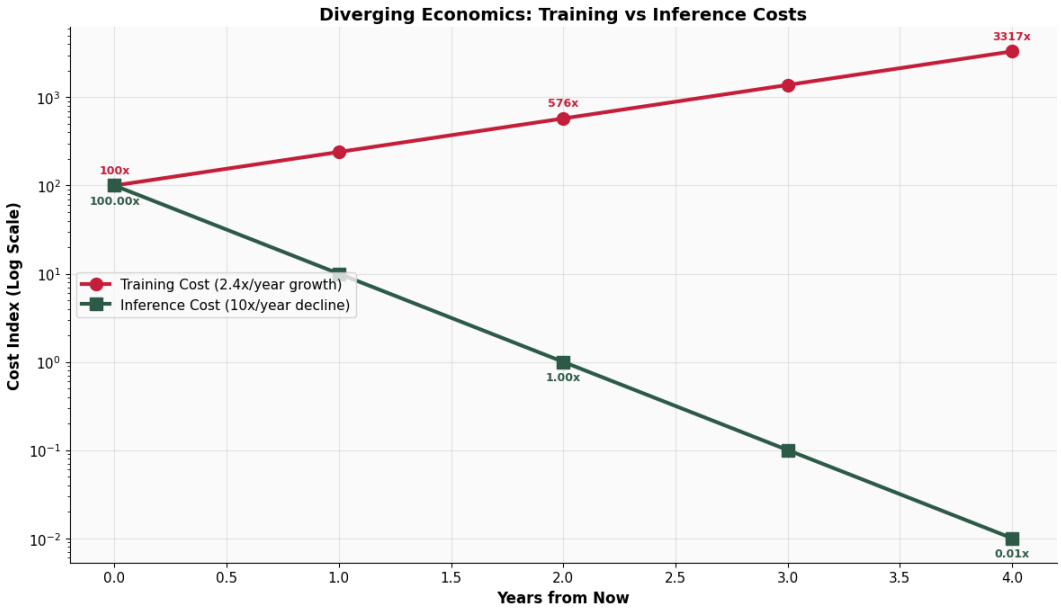
Taiwan remains the critical vulnerability in the global AI infrastructure supply chain. TSMC supplies 60%+ of world's chips including AI applications. A Sino-American conflict over Taiwan could destroy or embargo Taiwan's fabs, causing massive global economic disruption.

This geopolitical risk is the primary driver of all onshoring and redundancy policies. The U.S., China, and Japan are all subsidizing chip production elsewhere to reduce Taiwan dependence. Yet this diversification comes at the cost of efficiency and increased capex.

Part VII: Emerging Disruptions—The Five Forces Reshaping Economics

Disruption #1: The Training-Inference Bifurcation

Training costs are rising 2.4x annually while inference costs are falling 10x annually. These opposite trajectories create fundamentally different infrastructure economics.



Training vs Inference Costs

This chart visualizes the fundamental divergence in AI infrastructure economics. Training costs are growing exponentially at 2.4x per year (reaching $1B+ by year 4), while inference costs are collapsing at 10x per year (reaching near-zero by year 4). This creates a bifurcated infrastructure economics where training becomes an elite, hyperscaler-only activity while inference commoditizes.

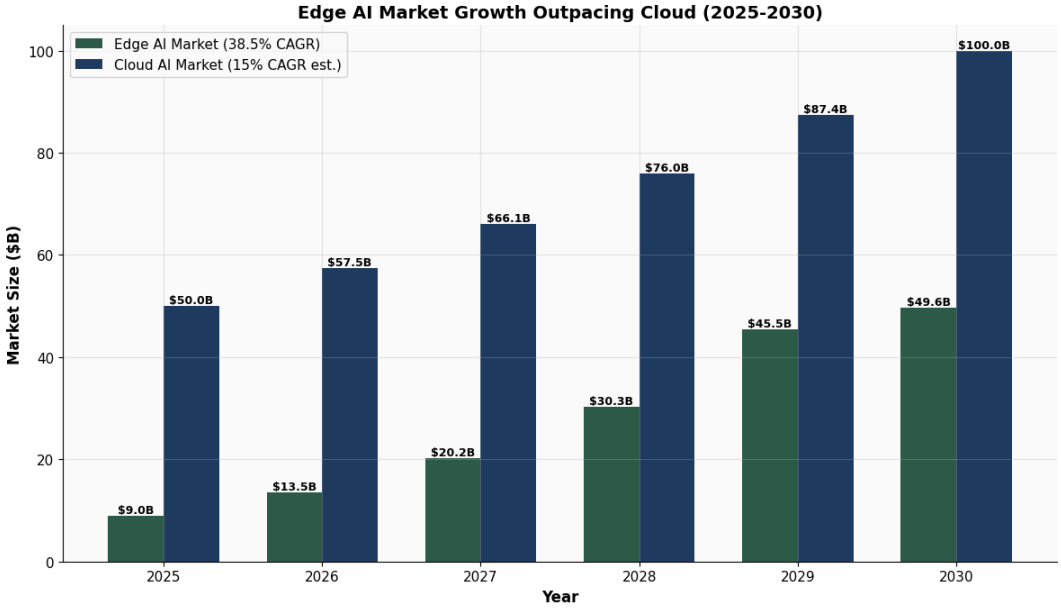
By 2027:

- Training runs: $1B+ each (only hyperscalers can afford)
- Inference: $0.01-0.1 per 1M tokens (commodity pricing)

This divergence is the single most important disruption to current economic assumptions. Hyperscaler ROI depends on training value capture, not inference volume. If inference becomes a commodity service with razor-thin margins, the ROI on massive inference infrastructure capex collapses.

Disruption #2: Edge AI as Genuine Alternative

Edge AI is growing 38.5% CAGR (vs. cloud AI ~15% CAGR) with 10,000x efficiency advantage and 5-10ms latency.



Edge vs Cloud Market

This chart shows edge AI market growth outpacing cloud AI. Edge AI market grows from $9B (2025) to $49.6B (2030), representing 38.5% CAGR. Cloud AI market grows ~15% CAGR (estimated). By 2030, edge AI represents 33% of total AI infrastructure market ($49.6B of $150B total).

The efficiency advantage is staggering: edge AI offers 10,000x efficiency advantage over cloud processing for inference. Edge provides 5-10ms latency vs. 100-500ms for cloud. Combined with small language models (5-29x cheaper) and efficiency improvements (10x/year), edge represents a credible threat to hyperscale cloud inference ROI.

Disruption #3: GPU Obsolescence Acceleration

Economic GPU life is 12-24 months, not 5-6 years. This creates massive stranded asset risk.

Current depreciation assumptions assume 5-6 year GPU economic life. However, actual economic life is 12-24 months due to:

- Rapid model efficiency improvements (10x/year inference cost decline)
- Potential alternative chip architectures (photonic, neuromorphic)
- Token price deflation (98% decline in 3 years)

This creates $100-400B stranded asset risk by 2027. If hyperscalers are depreciating GPUs over 5-6 years, their ROI assumptions are 2-5x too optimistic. When actual GPU life is 12-24 months, the capex investment cannot be amortized over the assumed period, creating massive writedowns.

Disruption #4: Alternative Chip Architectures

Photonic computing and neuromorphic chips could achieve commercial scale within 2-3 years, potentially disrupting GPU-centric economics.

Photonic computing is projected to surpass state-of-the-art electronic hardware (Google TPU v4) by two

orders of magnitude in throughput, power efficiency, and compute density. A neuro-optical computing platform announced in September 2025 achieves 16.2 Petaflops/s (FP8) per 1 cm² module with 1.83 femtojoules per operation (vs. 5+ pJ/op for electronic systems).

Neuromorphic chips offer ultra-low power consumption (15+ TOPS/W vs. 1-3 TOPS/W for GPUs) but remain far from commercial deployment. Intel's Loihi 2 is a research system, not commercially available. The neuromorphic chip market is $56.2M (2023), projected to reach $2.3B by 2034 (40.1% CAGR)—minuscule compared to GPU market ($100B+).

Probability of photonic/neuromorphic breakthrough by 2026-2027: 15-25%. Impact if realized: $200-400B asset writedown.

Disruption #5: Open-Source Models and Efficiency

Open-source models provide 5-29x cost reduction vs. proprietary APIs. This shifts capex from cloud providers to enterprises, reducing hyperscaler revenue while enabling edge/on-premise alternatives.

Meta AI's LLaMA-3 training required 350,000 NVIDIA H100 GPUs (~$9 billion) but models are now freely available for deployment. This shifts capex from cloud providers to enterprises. Enterprises can deploy small language models (SLMs) on-premise for $10-100M (vs. $1B+ hyperscaler capex), achieving comparable performance for specific tasks.

Part VIII: The Stranded Asset Trap—Why Current CapEx Assumptions Are Wrong

Critical Assumption #1: "GPU Inventory Will Amortize Over 5-6 Years"

Current hyperscaler accounting assumes 5-6 year GPU economic life. Actual economic life is 12-24 months.

This creates a 2-5x ROI gap. If hyperscalers invested $100B in GPUs assuming 5-6 year amortization, but actual useful life is 12-24 months, the investment cannot be amortized over the assumed period. This forces early writedowns and creates stranded assets.

Critical Assumption #2: "Cloud Datacenters Will Capture All Inference Workloads"

Current assumption: hyperscalers' massive capex investments will drive down inference costs, making cloud dominant.

Disruption: Edge AI growing 38.5% CAGR (vs. cloud ~15% CAGR); 10,000x efficiency advantage; SLMs 5-29x cheaper. By 2030: Edge AI market $49.6B vs. cloud AI market $100B (but growing faster). Workload shift: 30-40% of inference workloads move to edge by 2030.

Economic reality: Hyperscaler inference revenue growth may plateau or decline despite capex investments.

Critical Assumption #3: "Power Will Not Be a Constraint"

Current assumption: renewable energy and grid expansion will support 30-300 GW AI demand by 2030.

Disruption: Grid expansion is slow; renewable energy not scaling fast enough. Natural gas co-located plants likely primary solution (not green). Power becomes geographic constraint (not technical). Hyperscalers in power-constrained regions face severe capex constraints.

Economic reality: Power costs will rise 30-50%; geographic arbitrage becomes critical; hyperscalers without power access face severe capex constraints.

Quantified Stranded Asset Risk

The potential stranded asset writedown is substantial:

- **Scenario A (Efficiency Wins)**: $100-150B in datacenter/GPU inventory
- **Scenario B (Power Constraint)**: $50-100B in datacenters built in power-constrained regions
- **Scenario C (Photonic Breakthrough)**: $200-400B in GPU inventory + datacenters
- **Scenario D (Bifurcated Infrastructure)**: Minimal (infrastructure specialized for its purpose)

Probability-weighted expected stranded assets: $100-200B by 2027.

Key Insights

1. **NVIDIA's Extraordinary Profitability Is Unsustainable**: 78-85% gross margins represent a 28-29pp advantage over competitors. This pricing power is sustainable in the near term (1-2 years) but faces structural challenges from AMD ($5B+ sales at 49% margin), hyperscaler custom chips (15-44% TCO advantage), and emerging competitors. Expect margin compression to 50-60% by 2027 as competition intensifies.
2. **Manufacturing Economics Are Shifting From Chips to Power**: The logic chip costs only $200 (6% of H100 COGS). HBM memory and CoWoS packaging account for 86% of costs. The real constraint is not wafer cost but packaging capacity and power infrastructure. TSMC's CoWoS bottleneck and electricity supply constraints are now the binding constraints on AI infrastructure expansion.
3. **Hyperscaler CapEx Is Accelerating Into Uncertainty**: $235B annual spending by 2025 ($1T+ cumulative through 2030) assumes ROI that depends on training value capture, inference volume, and power availability. All three assumptions are under pressure. Operating margins are compressing despite revenue growth. The ROI on current capex levels remains uncertain.
4. **Vertical Integration Is Reshaping Competitive Dynamics**: Meta's 44% TCO advantage, Amazon's 35% advantage, and Google's 25% advantage through custom chips represent a structural competitive advantage that could sustain for 3-5 years. This shifts value capture from NVIDIA toward hyperscalers and threatens independent cloud providers.
5. **Power, Not Chips, Is the Real Bottleneck**: AI power demand growing 30x by 2035; grid expansion

slow; renewable energy not scaling fast enough. Natural gas co-located plants likely primary solution. Power becomes geographic constraint, not technical problem. Hyperscalers without power access face severe capex constraints. Geographic arbitrage becomes critical competitive factor.

6. **Global Supply Chain Fragmentation Is Permanent and Costly**: Export controls, tariffs, and reshoring policies are fragmenting supply chains and increasing costs by 30-50%. Global GDP losses from reshoring/friend-shoring are 1.8-4.5%. Taiwan remains a critical choke-point. This is a structural change, not a temporary disruption.
7. **Training and Inference Economics Are Diverging Dramatically**: Training costs rising 2.4x/year; inference costs falling 10x/year. By 2027, training is $1B+ per run (hyperscaler-only); inference is commodity ($0.01-0.1 per 1M tokens). This bifurcation creates fundamentally different infrastructure economics and threatens hyperscaler inference ROI.
8. **Edge AI Is Growing 2.5x Faster Than Cloud AI**: 38.5% CAGR vs. 15% CAGR. Combined with 10,000x efficiency advantage and 5-29x cost reduction from SLMs, edge represents a genuine alternative to centralized cloud inference. By 2030, edge AI represents 33% of total AI infrastructure market.
9. **GPU Obsolescence Is Accelerating**: Economic GPU life is 12-24 months, not 5-6 years. Token price deflation (98% in 3 years) + efficiency improvements (10x/year) + potential alternative chips create $100-400B stranded asset risk by 2027. Current capex assumptions are 2-5x too optimistic on ROI.
10. **The Current Infrastructure Bet Assumes Assumptions That Are Eroding**: The $1+ trillion capex bet through 2030 assumes that training and inference remain hyperscaler-dominated, that power will be available at reasonable costs, that GPU technology remains dominant, and that token prices will stabilize. All four assumptions are under pressure. Stranded assets of $100-400B are likely by 2027 unless fundamental assumptions hold.

Conclusions

The global AI infrastructure buildout represents one of the largest capital deployment cycles in history—a $1+ trillion bet through 2030. Yet this massive capex surge is built on economic assumptions that are rapidly eroding. The real story is not about chip scarcity or AI capability—it's about the collision between physics (power constraints), geopolitics (Taiwan, export controls), and efficiency (10x annual inference cost decline) reshaping the economic fundamentals of AI infrastructure.

Five fundamental disruptions are reshaping the economics: training costs exploding at 2.4x annually while inference costs collapse at 10x annually; edge AI growing 38.5% faster than cloud AI; power becoming the binding constraint (not chips); efficiency improvements rendering GPU inventory obsolete within 12-24 months; and alternative chip architectures potentially achieving commercial scale within 2-3 years.

The value accrual is heavily concentrated—NVIDIA captures ~$75 billion (95% of AI chip market share), hyperscalers control infrastructure, while applications capture only ~$5 billion. Yet this concentration masks a fragile economic foundation. Export controls, tariffs, and reshoring policies are fragmenting supply chains and increasing costs by 30-50%, creating 1.8-4.5% permanent GDP losses. Taiwan remains a critical choke-point: a Sino-American conflict could destroy or embargo Taiwan's fabs, causing global economic disruption.

The shift from efficiency-optimized to resilience-optimized supply chains is permanent, and the cost is substantial. The $1 trillion in efficiencies from integrated global supply chains is being sacrificed for

geopolitical resilience. This is the new baseline for AI infrastructure economics.

The most critical finding: current hyperscaler capex assumptions are 2-5x too optimistic on ROI. Stranded assets of $100-400 billion are likely by 2027 unless fundamental assumptions hold. The current infrastructure economics are unsustainable. The real question is not whether disruption will occur, but which disruption will materialize first—and which hyperscalers will be positioned to survive the transition.

Sources

Semiconductor Economics

- NVIDIA Q1 FY2025 10-Q Filing (April 28, 2024)
- AMD Q3 2024 Earnings Report
- TSMC Q3 2024 Earnings Report
- SemiAnalysis: "H100 Manufacturing Cost Breakdown and GB200 Architecture"
- TechPowerUp: NVIDIA H100/H200 Pricing Guide (2025)
- DIGITIMES Research: TSMC CoWoS Capacity Analysis (2024-2026)

Hyperscaler Capital Expenditure

- Microsoft 2024 Annual Report & 10-K filings
- Google/Alphabet 2023-2024 Annual Reports & 10-K filings
- Amazon 2023-2024 Annual Reports & 10-K filings
- Meta 2024 Quarterly & Annual Reports
- DataGravity: "$150B+ of Annual CAPEX: The trends in Capital Expenditures by Hyperscaler Tech Giants"
- Bloomberg: "Big Tech Capex on track to surpass $1 trillion by 2029 amid AI race"
- McKinsey: "AI power: Expanding data center capacity to meet growing demand"

Energy and Infrastructure

- Lawrence Berkeley National Lab: "Queued Up: 2024 Edition" (interconnection queue analysis)
- U.S. Energy Information Administration (EIA): "Electricity Monthly Update" (2025)
- Turner & Townsend: "Data Centre Cost Index 2024"
- Goldman Sachs: "AI Data Centers' Global Power Surge" (2024)
- Deloitte: "Can US infrastructure keep up with the AI economy?" (2025)

Global Capital Flows

- Stanford HAI AI Index Report 2024

- KPMG Venture Pulse Reports (Q1-Q3 2024, Q3 2025)
- CB Insights State of AI Reports
- BlackRock/Microsoft Global AI Infrastructure Investment Partnership Announcement (September 2024)
- NIST CHIPS Program Office Fact Sheet (April 2024)
- European Commission InvestAI Initiative (January 2025)

AI Services Unit Economics

- Cottier, B., et al. (2024). "The rising costs of training frontier AI models." arXiv:2405.21015
- Sardana, N., et al. (2024). "Accounting for Inference in Language Model Scaling Laws." arXiv:2401.00448
- AWS SageMaker Pricing: https://aws.amazon.com/sagemaker/pricing/
- Azure Machine Learning: https://azure.microsoft.com/en-us/pricing/details/machine-learning
- Google Cloud Vertex AI: https://cloud.google.com/vertex-ai/pricing

Geopolitical and Policy Dimensions

- Bown, Chad P. and Dan Wang. "Semiconductors and Modern Industrial Policy." Peterson Institute for International Economics Working Paper 24-3 (August 2024)
- Cerdeiro, Diego A., et al. "The Price of De-Risking: Reshoring, Friend-Shoring, and Quality Downgrading." IMF Working Paper WP/24/122 (June 2024)
- Ezell, Stephen, et al. "Short-Circuited: How Semiconductor Tariffs Would Harm the U.S. Economy and Digital Industry Leadership." Information Technology and Innovation Foundation (May 2025)
- Sreedharan, Vibhaa, et al. "The Geopolitics of the AI-Relevant Semiconductor Supply Chain." Graduate Institute Geneva (June 2024)
- U.S. Government Accountability Office: "Export Controls: Commerce Implemented Advanced Semiconductor Rules" GAO-25-107386 (January 2025)

Emerging Trends and Future Dynamics

- Ahmed, Sufi R., et al. (2025). "Universal photonic artificial intelligence acceleration." Nature, 640(8058)
- Irugalbandara, Chandra, et al. (2024). "Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's LLM with Open Source SLMs." arXiv:2312.14972
- Marpu, Rhea Pritham. (2024). "The AI Shadow War: SaaS vs. Edge Computing Architectures." arXiv:2507.11545
- Sakib, Tanjil Hasan. (2025). "Small Language Models (SLMs) Can Still Pack a Punch: A survey." arXiv:2501.05465
- Sarkis, Nicholas, et al. (2024). "Compact Language Models via Pruning and Knowledge Distillation." arXiv:2407.14679
- Zhu, Yiqun, et al. (2024). "Q-Sparse: All Large Language Models can be Fully Sparsely-Activated." arXiv:2407.10969
- IDC: "Worldwide Edge Enterprise Infrastructure Workloads Forecast, 2024–2028"
- Transparency Market Research: "Neuromorphic Chip Market Analysis"