THE HIDDEN COST OF THE AI

POWER RISK

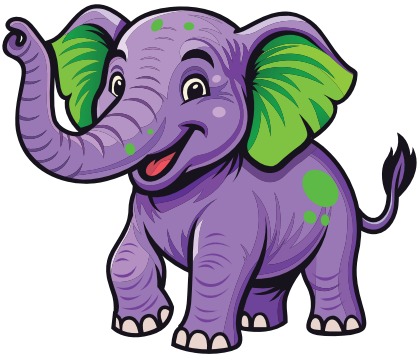
exactmarket

ELEPHANT IN THE ROOM EBOOK SERIES

TABLE OF CONTENTS

WELCOME TO THE ELEPHANT IN THE ROOM!

You have in your hands a guide designed to call out the elephant in the room—a topic that's too important to be ignored but maybe isn't getting the attention it deserves.



INTRODUCING THE ELEPHANT

Turning Power Constraints into a Vendor Advantage

Nobody wants to admit this yet: AI's next barrier isn't innovation. It's electricity.

Hyperscale data centers are consuming grid capacity faster than it can be expanded. And vendors embedding AI into their products are about to face a significant obstacle, not because they lack features, but because the grid can't keep up.

This book aims to highlight that uncomfortable truth and to give vendors an advantage amid these challenges.

INTRODUCTION

THE ENERGY DILEMMA:
Supply and Demand

AI features are being integrated into products faster than any significant shift since mobile technology, but this speed masks a core truth that most haven't fully understood:

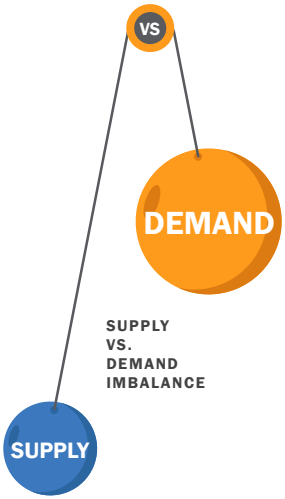
01 AI scale is tied to the laws of physics concerning power, meaning your solutions could be similarly constrained.

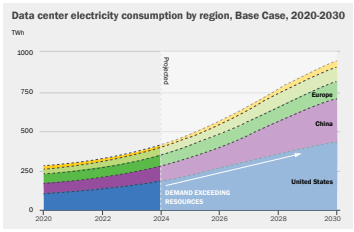
02 Model training requires immense energy, approximately three times more power per square foot of data center space.[1]

03 According to Statista, data volumes are soaring from 149 zettabytes of data in 2024, to an expected 394 zettabytes by 2028; this accounts for 5% of data center energy consumption.[2]

04 Inference continually consumes energy.

05 Low-latency networking for inference tasks and hybrid communications accounts for 5% of data center energy consumption.[3]





Source: IEA, Energy demand from AI, Dec 2024

To meet AI data center requirements, the demand is expected to double from 415 TWh in 2024 to 945 TWh in 2030.[4] The only feasible solution to meet the demand is to build small modular nuclear reactors (SMRs), which take between two and four years on average,[5] meaning **there will be a considerable shortage** in the coming years, as many are still in contractual stages and should be being built now to meet pending demands.

This means the old assumption "cloud = unlimited capacity" is no longer neutral; instead, AI adoption and consumption within your product will relate to increased exposure to energy-driven cost volatility.

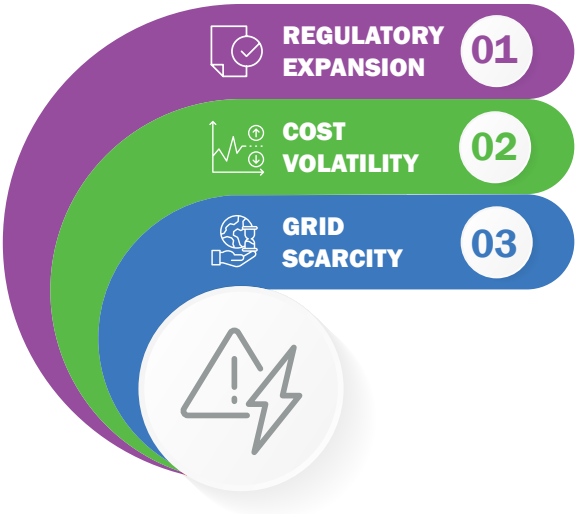
For vendors, this matters because:

- Gross margin on AI features now depends on someone else's energy procurement, essentially out of your control.
- Your cost to serve at scale may not grow in a predictable, linear manner and may become more expensive if energy resources, such as SMRs, are delayed. Power could become a premium in data centers.
- Your buyers (CFO, CIO, regulators) will increasingly view AI through a risk perspective rather than just as a feature.

CHALLENGE

THE POWER CEILING

Three Forces Collide



EU/APJ increasingly require locality + provenance.

Cloud AI cost → now shadow-priced by energy.

Hyperscalers are securing early access to power.

Grid scarcity, in terms of power required and connection to the grid, could impact hyperscale elasticity for essential inference tasks of solutions, resulting in supply vs. demand contention.

AI inference cost volatility is driven by power demand and supply, despite the energy line item not being visible. When electrical grid stress rises, for example, heatwaves, transmission constraints, seasonal demand, the real cost under AI features can change, even if you never change your code.

Regulatory expansion, such as the European Union Artificial Intelligence Act (EU AI Act) and Network and Information Systems Directive (NIS2), among others, aims to tie down AI deployments to specific physical jurisdictions, affecting not only where inference is executed but also how it is executed. This results in more pressure in energy-constrained markets.

RESULT:

AI PRODUCT STRATEGY = **ENERGY** + **GEOGRAPHIC AVAILABILITY**

Hyperscale Hits the POWER CEILING

Vendors have been conditioned to assume that hyperscale equates to “infinite backend capacity.”

This assumption is no longer valid.

AWS, Azure, and Google Cloud are now projecting their physical limits and incurring unprecedented capital expenditures to sustain growth. To maintain AI infrastructure growth, they must:

- Acquire more land
- Construct additional data centers
- Secure a more stable silicon supply
- Secure more grid capacity contracts
- Underwrite more power purchase agreements (PPAs)

They have committed to approximately **$240 billion a year in data center, energy, and silicon capital expenditures** to sustain the AI growth curve.[5] Where is that huge investment coming from? **YOU.**

This shift introduces a new reality for any vendor whose AI inference relies solely on hyperscale infrastructure: **UNCERTAINTY**

01

MARGIN UNPREDICTABILITY
Fluctuating power becomes an opaque factor that can influence cost paths.

02

VOLATILITY IN THE COST TO SERVE
A grid shock (like a heatwave, outage, or capacity limit) in one region can instantly alter energy costs, silently affecting your AI feature’s cost to serve.

03

LATENCY FLUCTUATIONS
GPUs are being centralized into fewer, larger, and more power-dense data centers.

04

PRICING PRESSURES
These are arising from the inability to shield buyers from energy inflation.

This is not a technical issue; instead, it highlights the fragility of the business model. When the platform itself is constrained by power, AI feature costs are also constrained by the same power limitations.

When grid conditions change, the cost to serve could drift even if the code remains unchanged.

The Land Grab

AWS, Azure, and Google are not sitting still. They are now spending extraordinary capital to secure one thing above all else: **future power availability**.

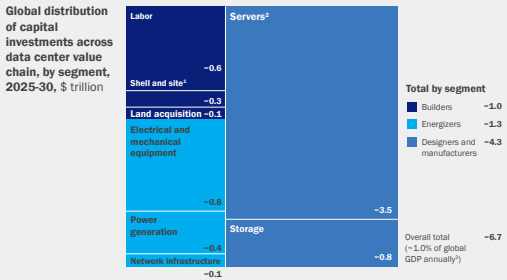
This is why you are seeing:

- **Hyperscalers pre-buying land before power is even provisioned**
- **Multi-billion PPA agreements signed years in advance**
- **Entire new data center campuses being built around power access, not just comms**

PPAs are long-term contracts between an electricity generation company and a buyer, primarily for renewable energy projects. They set a fixed price for a duration, giving stability and reducing costs through long-term contractual discounts.

In 2025, hyperscale data centers were the largest consumers of PPAs. In 2025, Amazon signed a 1,920-MW nuclear energy PPA through 2042, the largest ever signed by a private company, and Google signed a 200-MW nuclear PPA, not set to be ready for production until 2030. Essentially, they are buying up the resources before they are even available.

$6.7 trillion of capital expenditure will be cumulatively deployed for data center infrastructure through 2030.



[1] Includes mechanical, electrical, and plumbing.
[2] Including graphics processing units and central processing units.
[3] Global GDP in 2023: $106 trillion.
Source: Goldman Sachs; S&P Capital IQ; McKinsey analysis

McKinsey & Company

Analysis shows that by 2030, about $7 trillion in CapEx will be spent on data center infrastructure globally.[7]

More than $4 trillion of it will be allocated toward computing-hardware investments, with the remainder allocated to areas such as real estate and power infrastructure.

More than 40% of this spend will be in the United States.[8] This isn't about greed; it's simply the price we pay for challenging the laws of physics.

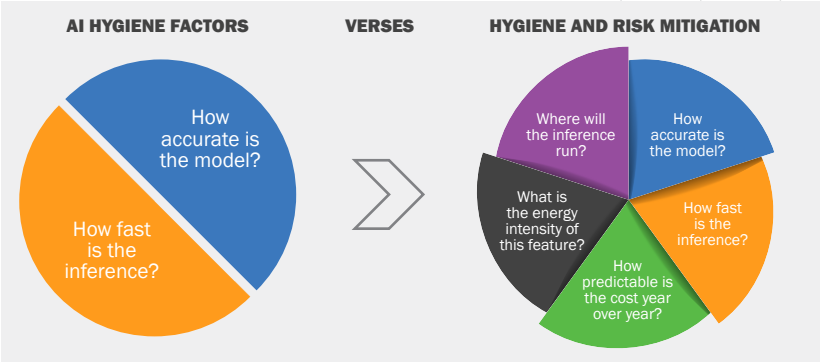
IMPLICATION FOR VENDORS

Even though inference feels virtual and abstract in product roadmaps, the economic substrate underlying it will become scarce, physical, and price-sensitive. This is why "just run inference in hyperscale" is no longer a neutral architectural default. **It is a dependency on someone else's power future**.

When power ceilings constrain the platform, the economics of AI features are also constrained.

Power Is Now a Procurement Variable

AI is no longer just a cloud consumption item. It now introduces a new category of enterprise risk, energy-driven cost instability. Buyers should start considering evolving from evaluating hygiene factors to minimizing exposure.



Customer evaluation of AI products should shift toward:

How predictable is my AI OpEx spend?
Procurement teams dislike unpredictable variable costs.

What is the energy usage per unit (token/frames/predictions) of inference?
Although baked in today, energy could become a benchmark metric, similar to $/GB in cloud storage.

What are the locality constraints and jurisdictional triggers?
Regulators increasingly focus on where data and compute physically occur.

What is the vendor's plan for mitigating regional grid stress events?
Grid stress equals energy volatility, which could lead to cost volatility.

Emerging FinOps practices should start to connect AI expenses directly to energy risks.

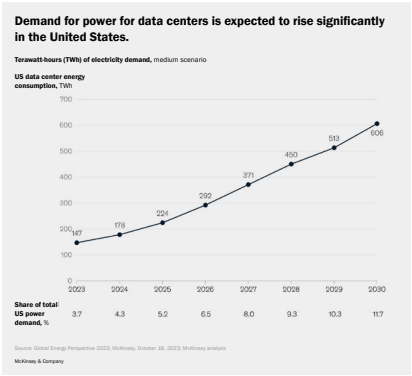
The new enterprise evaluation model should be:
AI = energy exposure + regulatory exposure (jurisdiction)

Vendors who can cut that exposure for the buyer—through product design and placement control—become less risky for customers to procure, easier to pass security and risk reviews, and simpler to forecast in multi-year budgets. That is the premium position in the market and the voice that should be amplified into vendor marketing. Accuracy is basic; predictability is a way to stand out.

In 2026 and beyond, de-risking might be a more compelling sales point than "smarter AI."

Cost Becomes Energy-Aware

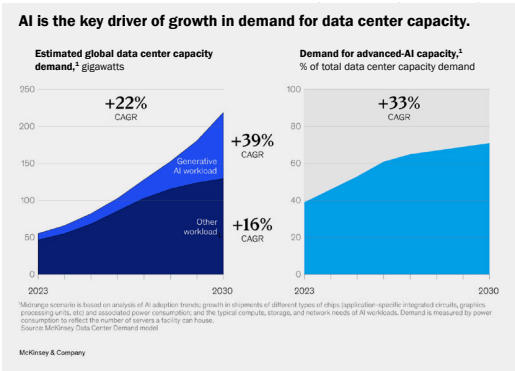
When it comes to the landscape of identities, accounts, and other credentials, organizations struggle with volume, velocity, and variety.



In 2025, wholesale electricity was up to 267% more expensive than in 2020 in areas near US data hubs.[9]

Data center energy consumption is expected to increase in tandem with growing AI demand. Greater deployment of high-performance infrastructure increases power density, now **5-20%**, and is expected to reach **35-50% by 2030** due to AI and HPC.[10]

AI economics traditionally relied on cost-per-call, per-month, or per-token models. When energy becomes the key factor, the benchmark could shift to: How much power does your model consume? How efficiently can customers operate it? What deployment flexibility do you provide?



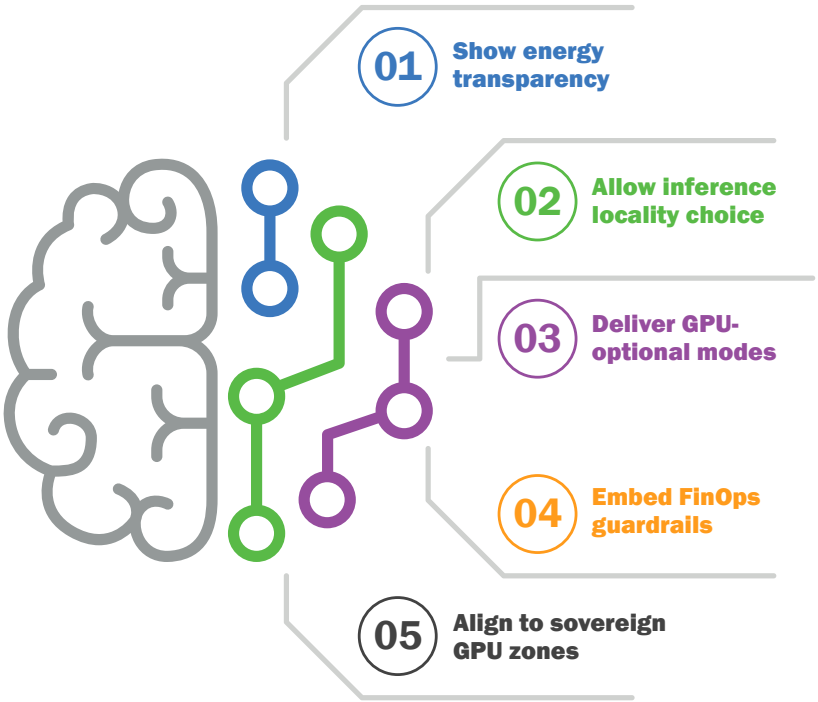
In the future, the financial model is likely to evolve to include a cost per kWh needed to run AI features. The winners will be vendors who can maintain smaller model footprints, implement GPU-optional CPU inference modes, and offer customers control over jurisdiction and locality. This represents the new premium positioning.

Vendors won't succeed by simply being cheaper. They will succeed by being energy-efficient and by offering intelligent placement.

SOLUTION

Five Plays That Build Competitive Advantage

AI features may no longer be distinguished by model intelligence. Instead, they differentiate by how controllable, predictable, and energy-efficient they are in real customer settings. These five plays don't require "bigger models." They need choice architecture, the ability to let the customer decide where and how the AI runs.



Vendors who deliver choice will be seen as lower risk, more future-proof partners.

PLAY 1

Energy Transparency

Vendors that showcase how their AI features use energy will earn the trust of customers in the upcoming cycle. Why is that? Because the main cost factor of AI is moving from just “usage” to “power.”

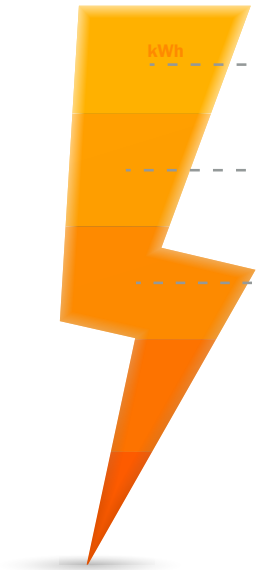
Energy transparency will be how vendors:

- Protect margins
- Protect buyers’ budgets
- Build trust at CFO + CRO + regulator levels

This isn’t a sustainability pitch. This is about financial predictability and risk reduction.

What should vendors expose?

Useful transparency metrics include:



kWh per 1,000 inferences
(as a benchmark for comparison)

Latency and energy multipliers
across regions like US-East, Asia Pacific and Japan (APJ) and the European Union (EU)

Energy consumption differences
between GPU and CPU (highlighting cost variations)

Why this matters to enterprises

If vendors can demonstrate how costs behave under different modes and locations, then the customer can model budget, regulatory, and performance requirements. FinOps personas, CFOs, and CIOs all seek key insights linked to these behaviors, **which can help predict operational costs over time**.

Incorporating energy transparency into your product shifts the narrative from “AI is costly” to “Here’s how we ensure AI’s financial sustainability.” This strategy offers a competitive advantage because vendors who can clearly explain their cost are more likely to attract buyers and loyalty.

PLAY 2

Inference Placement Setting

Factors such as distance to the server, network infrastructure, data routing, and server load influence traditional inference placement. Placement control could also be considered an economic lever, because where inference runs determines the **energy price, legal exposure, latency path, and data residency**.

Placement is no longer optional; it has become a part of the economic agreement between vendors and customers. In practice, this means offering the customer a clear, structured choice.

PLACEMENT OPTION	WHEN IT'S BEST
Hyperscale (AWS/Azure/GCP)	Burst capacity, experimentation, fast scaling
Private cloud/ On premises	Stable, long-lived inference with sensitive data
Sovereign GPU zones	Regulated industries + regulated regions

The placement settings can be set at the global (platform level), per workspace/tenant, or even per feature.

Why this matters to enterprises

Data protection authorities now regulate not just data at rest, but also **where and how** AI inference occurs.

Medium- and large-sized enterprises are establishing AI risk teams to conduct **model risk ratings and approve models**. Without knowing inference locations, they can't assess risks, leading to deployment blocks or vendor pushback.

CFO and FinOps teams are responsible for **ensuring cost predictability and are affected when inference costs rise unexpectedly, especially if linked to grid volatility, which makes** forecasts unreliable.

When they choose a locality, they are, in effect, choosing power cost exposure, legal exposure, and performance characteristics. They can select the exposure profile they prefer.

Strategic implication for vendors

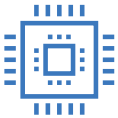
Vendors should consider integrating placement selection into the onboarding process, offering default placements based on the use case, and allowing overrides.

Placement choice becomes a confidence feature. Vendors turning placement into a knob turn risk into control. This helps you become the "safer vendor" without compromising your AI capabilities.

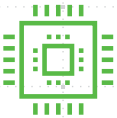
PLAY 3

GPU-Optional Modes

Enterprise AI often focuses on inference rather than training. AI used to rely solely on GPUs for the necessary processing. Still, modern inference runtimes, compression techniques for smaller models, and low-volume or latency-sensitive workloads now enable AI workloads to utilize more abundant and affordable CPUs, especially on lightweight edge devices.



VS.



CPU	GPU
Optimized for low-latency, sequential tasks	Parallel architecture features thousands of small cores for simultaneous compute
Works best for small models, I/O-bound workloads, predictable latency, and edge applications	Ideal for large language models, high throughput, and low latency at scale.
Less built-in memory cache, and generally, low-cost and readily available	Higher memory bandwidth and large vRAM capacity
	High cost and constrained sourcing

While GPUs remain the dominant choice for high performance in generative AI (GenAI) and parallel computing workloads, not everyone needs this level of high throughput and low latency at scale. On the other hand, CPU capacity is plentiful, unaffected by sourcing and delivery delays, easier to power, much cheaper per cycle, and not limited by the same power constraints that GPUs face. Utilizing CPUs for inference provides flexibility, allowing projects that might have been delayed due to GPU sourcing issues to proceed.

Inference can often operate on CPUs in private clouds, edge locations, and on-premises regulated environments.

By designing your product to scale core inference on CPUs when needed, you give your customers commercial flexibility that hyperscale setups often lack.

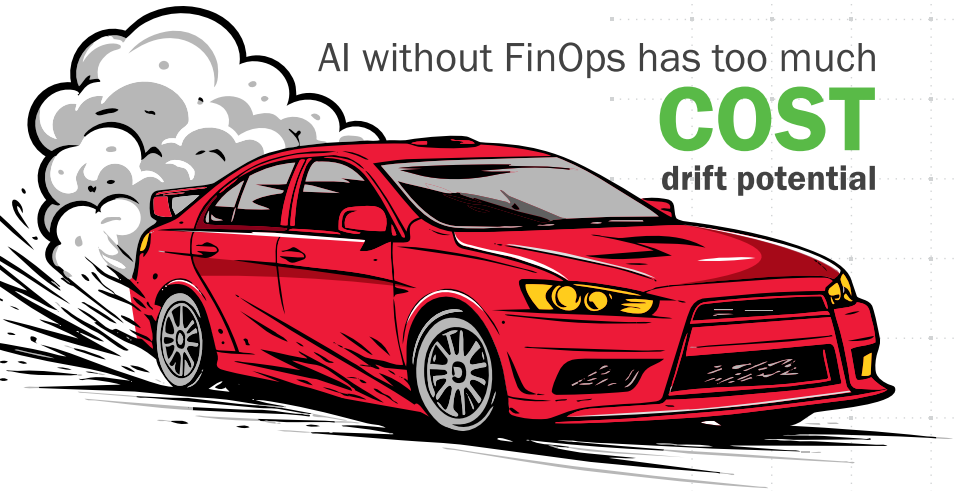
Why this matters for enterprise buyers

CPU-capable inference enables enterprises to scale AI without being hindered by GPU scarcity, cost spikes, or regulatory barriers.

- Lowers cost volatility (cheaper cycles)
- Deploys in more places (private cloud, edge, on-prem)
- Avoids GPU procurement delays
- Simplifies regulatory approval (easier to certify)
- Accelerates time to production AI adoption

PLAY 4

FinOps Built-In



The next advantage for vendors is to integrate FinOps metrics directly into the AI product experience, where decisions are made.

What this means in practice

Instead of the AI feature being a black box, with a price tag attached, the vendor shows:

- Forecasted spend range for a given deployment configuration
- Expected cost-per-unit (per 1,000 tokens/per 1,000 predictions/per 1,000 frames)
- Mode cost analysis (CPU mode vs. GPU mode cost curves)
- Regional pricing overlays (how region A behaves vs. region B)

This creates a new kind of product value: cost intelligence at the point of design.

Strategic implications for vendors

Make AI cost explorable, comparable, and predictable inside your UI. Don't make buyers hunt for the economics. **Show it**. That can help make you, the vendor, feel like a partner, not a future liability.

Why this matters for enterprise buyers

Integrating FinOps into the product gives buyers cost clarity at the exact moment they make deployment decisions.

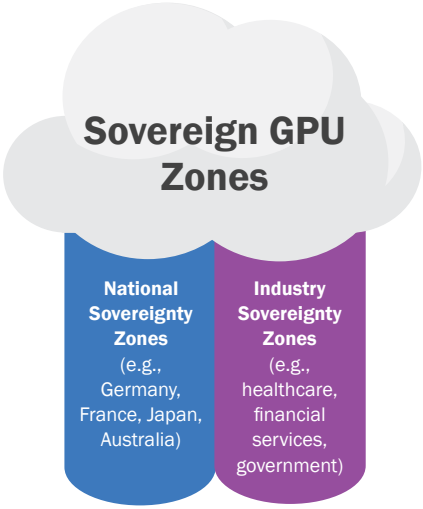
- Show forecasted spend ranges
- Expose cost-per-unit (tokens/predictions/frames)
- Compare CPU vs. GPU cost curves
- Map cost differences by region
- Turn AI pricing into an explorable UI and KPI that is transparent

PLAY 5

Sovereign GPU Zones

AI is no longer “just cloud.” It is now becoming a sovereign capability; examples include the United States taking stakes in quantum technology companies in 2025 and the UK’s AI Opportunities Action Plan investing £25 billion in AI data centers.

Governments, regulators, and national industries are beginning to determine where AI operates, and under what laws, not based on feature value, but on control, locality, and power provenance. This creates a new category of premium compute real estate: **Sovereign GPU Zones**.



These zones are not “niche.” They are the next high-margin, high-trust segment of AI infrastructure.

Why this matters to vendors

If your AI features can run inside sovereign zones, you unlock:

- Access to regulated budgets
- Faster compliance acceptance
- Less resistance in procurement cycles
- More predictable long-term consumption

De-risk customers from hyperscale grid constraints.

Why this matters for enterprises

Sovereign GPU Zones are becoming the highest-trust, highest-margin AI markets, because they combine to negate legal, economic, and infrastructure certainty.

- Unlock regulated budgets
- Accelerate compliance approvals
- Remove procurement friction
- Stabilize long-term consumption patterns
- Protect customers from hyperscale power volatility

Strategic implications for vendors

Partner in Sovereign GPU Zones now. Giving you:

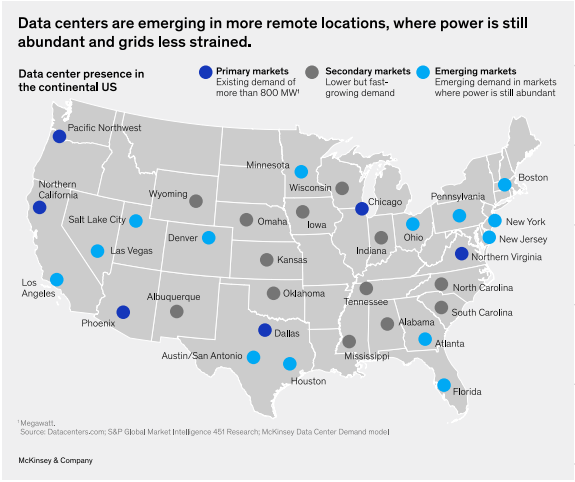
- Legal resilience (jurisdiction clarity)
- Economic resilience (power stability)
- Go-to-market resilience (regulated verticals buy you faster)

The future premium for vendors is not performance alone; it is placement legitimacy.

UNITED STATES LENS

Private Power + Private Equity

Data center power needs in the United States are expected to increase by approximately 460 TWh of demand from 2023 to 2030, which is **three times the current level of consumption**. Similarly, water utilities used for data center cooling are becoming strained, with a predicted **170% increase in demand** expected by 2030.[11]



The United States remains the focal point of the AI race, with hyperscale companies investing around billions annually to stay ahead. This includes expenses for land, new data centers, GPU acquisitions, and power agreements.

Source McKinsey: The Data Center Balance, August 2025

The US regulatory environment is still relatively permissive compared to the EU, but the constraint that will slow down data center buildouts is economic physics: state-by-state grid fragmentation, private equity competitively bidding for megawatts, and hyperscalers overbuying capacity to future-proof AI scale. This means in the US market, **electrical power is becoming a competitive asset class**.

What this means for vendors selling in US markets

Buyers will favor vendors that can place inference in the cheapest and most stable power zones, can offer CPU modes (more affordable to run everywhere), and can demonstrate the energy intensity of their features.

Because the question in every US CIO/CFO is shifting from “Is this good AI?” to “Can this scale without becoming a cost bomb?”

PLACEMENT + TRANSPARENCY WINS HERE.
Not model features or complexity.

EU LENS

Sovereignty + Energy Provenance

Europe will not scale AI through indifference to hyperscale expansion.

More organizations seek reassurance that their **AI data stays within EU borders and complies with EU laws**.

The shift from non-jurisdictional vendors, mainly US hyperscale firms, is due to concerns over privacy, GDPR, and stability. EU countries view these firms as vulnerable to geopolitical shifts and often non-compliant. For EU nations, local data storage is crucial for growth and security.

The EU has acknowledged the data center energy crisis, perhaps more so than the United States, especially since much of the EU's energy previously came from Russia before the war in Ukraine. Data center energy consumption in the EU is projected to double by 2030, primarily driven by AI.[12] Competing against the buildout of new data centers is an already strained grid, which will likely delay the rollout, as existing sites are heavily congested and new sites require grid connectivity.

The EU already regulates data center energy consumption through the Energy Efficiency Directive (EED), Corporate Sustainability Reporting Directive (CSRD), NIS2, and upcoming laws, such as the Cloud and AI Development Act. However, all regulators agree on a key principle:

RISK = **where the data is processed, under which law, and powered by which infrastructure.**

This push promotes moving toward **localized or dedicated private cloud inference deployments** within the EU, ensuring jurisdictional placement with legal provenance, instead of relying on shared, global infrastructure that could route data outside the EU.

Inference locality is not a "preference" here; it is a compliance boundary.

EU buyers (primarily regulated) are increasingly asking:

"Can you guarantee inference stays inside the EU legal perimeter?"

"Can you attest that the power used is under EU regulatory reach?"

"Can I choose a zone where GDPR and sector law are enforceable?"

What this means for vendors

If you can't prove locality and lack transparency in energy, you create risk; if you can, you eliminate it. When vendors are equal, the tiebreaker in EU enterprise buying is: **Who guarantees compliance in energy and placement?** EU companies typically use energy accounting, and transparent solutions become more popular as compliance can be automated. In Europe, the key advantage isn't the largest model but the clearest, **most compliant jurisdiction with transparent energy accounting**.

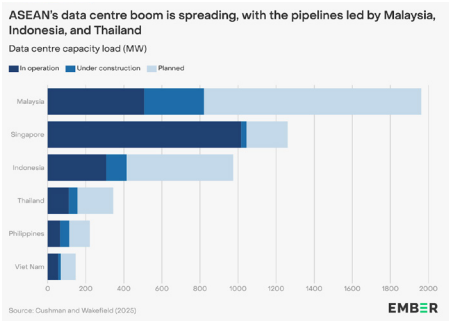
APJ LENS

Speed + National Industrials

Data center energy use in Asia Pacific and Japan (APJ) is rapidly increasing. As the **fastest-growing data market**, APJ faces challenges supporting a data-driven economy, with some regions expecting data centers to use 30% of total energy by 2030, **causing a sevenfold power increase**.[13] Yet, the focus across APJ isn't solely on "performance" but on "**capability independence**"—the ability for nations or sectors to run AI aligned with domestic policy. Three regions in APJ are developing different sovereign models.

REGION	DRIVER	WHAT THIS MEANS FOR VENDORS
Japan/Korea	Industrial AI as national competitiveness	They want domestic capability over single-vendor hyperscale reliance; CPU inference and national cloud alignment fit this need.
Singapore	Highly regulated hub and very tight land/power options	Land restrictions prevent building more DCs. Vendors efficient within capacity may gain privileged access to Singapore's finance and government workloads.
Australia	Following in EU-style laws and regulations	The Australian government is adopting EU-like jurisdictional control, turning placement guarantees into procurement blockers or accelerators. Energy transparency may become a key public sector differentiator.

Energy transparency is limited across Asia, mainly in advanced economies like Japan, Korea, Singapore, and Australia, where it offers a competitive edge. In markets such as India, Thailand, Indonesia, and Malaysia, data center growth is strong, but energy transparency isn't a key factor. Instead, decisions are driven by placement control within sovereign jurisdictions and AI CPU options, due to supply delays in GPUs.



Source: Ember, ASEAN's data centres electricity demand keeps growing, May 2025

For vendors, consider:

- CPU modes = better deployability (more infra options)
- Placement control = lower regulatory friction
- Energy transparency = faster comfort from public sector buyers

APJ will buy vendors that enable national capability, not those necessarily that entrench them deeper into US hyperscale economics.

The 4A Vendor Decision Framework

This is the new strategic sequence for vendors designing AI features in a **power-bounded world**.

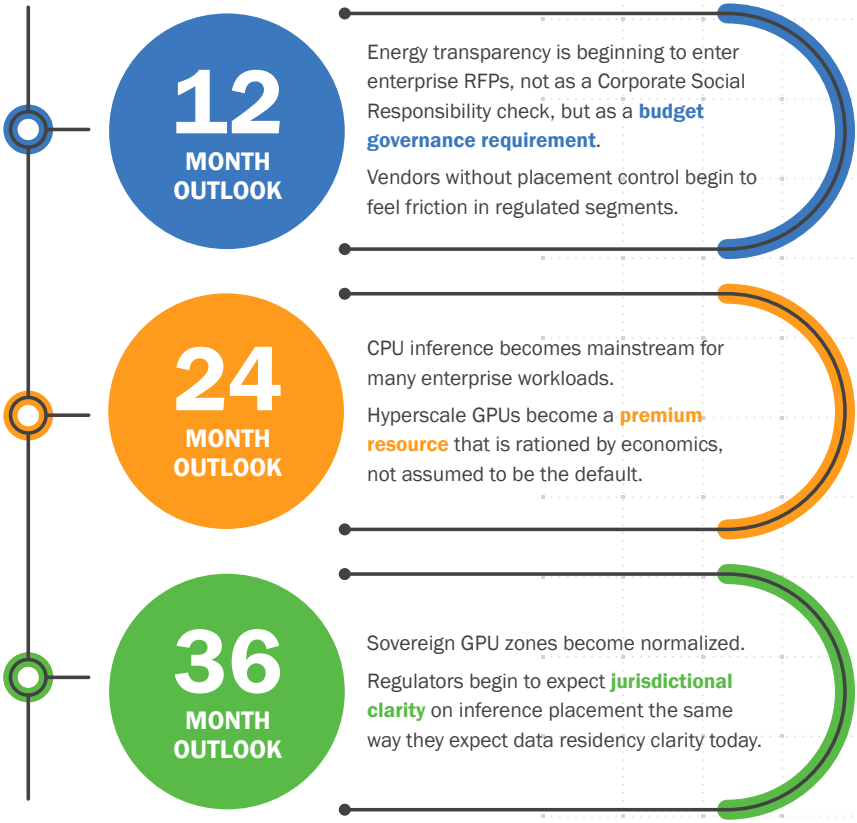
	ASSESS	ALIGN	ACT	ADAPT
SUCCESS	Know your current exposure, both in terms of power and jurisdiction.	Prioritize the roadmap where friction and payoff are highest.	Product de-risks AI consumption at the point of use.	An AI product remains stable even as the market changes beneath it.
VENDOR FOCUS	Map inference locations by region, platform, and energy profile. Identify GPU-dependent vs. CPU-capable workloads.	Identify customer segments that value cost predictability and locality control, such as FSI, Healthcare, Public Sector, and Industrial.	Integrate build placement choice, CPU modes, and energy transparency into the product experience—not just sales verbiage.	Review quarterly: power market signals, GPU pricing trends, sovereign zone availability: update defaults and recommendations.

This model makes vendors more resilient. It shifts product posture from "Use our AI" to "**Use AI on terms that keep your cost and compliance under control**." Operating this way will help scale safely, sell faster, and retain a higher margin over time.

FUTURE VIEW

12–36 Months

AI capability will continue to accelerate, but the real constraint is economic stability. Enterprises will focus on AI strategies that preserve margins, reduce compliance risks, and ensure transparency of costs. The key shift isn't about AI potential, but about trustworthy AI architectures at scale. Over the next three years, this will shape the industry.



THE WINDOW IS BRIEF, BUT THE BENEFITS ARE SIGNIFICANT.

Vendors who act now will:

- **Become the preferred low-risk option for enterprises**
- **Maintain margin stability amid fluctuating energy prices**
- **Expand AI capabilities without relying on hyperscale power futures**

Vendors who delay may face a defensive stance, not because their AI is ineffective, but because of high physical costs.

The successful players in the next AI cycle will be vendors who view placement as a core product and prioritize energy as a key design factor.

GLOSSARY

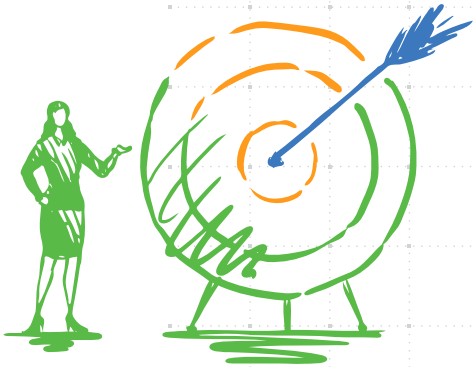
TERM	DEFINITION
AI inference	Running a model to generate output (answers/predictions) is part of the operational phase of AI-
cost-per-unit of inference	Cost model calculated per energy used for each unit of AI output (such as 1,000 tokens, predictions, or frames)
CPU-based Inference	Running inference on CPUs rather than GPUs is more cost-effective, simpler to power, and easier to deploy
energy exposure	Risk that the cost to serve AI increases due to energy price shifts
energy provenance	Proof of the source and jurisdiction of authority used to operate data centers and workloads
energy transparency	Showing model energy performance (e.g., kWh per 1,000 inferences) as a key metric
FinOps	Cloud financial operations discipline, overseeing cloud costs as a managed business function
GPU-optional modes	The product enables inference to run on the CPU when necessary, eliminating the need for a mandatory GPU dependency
grid stress	When the demand for power temporarily surpasses the supply capacity, leading to price spikes
placement control/ inference locality	Allowing customers to choose where inference runs (region/infrastructure/sovereign zone) to control cost and ensure compliance
Power Purchase Agreements (PPAs)	Long-term contracts in which a cloud provider secures future electricity supply (often years in advance) to ensure they have sufficient power for AI data centers
sovereign GPU zones	Compute functions are governed by national or sector-specific jurisdiction rules involving regulated, high-trust AI environments

SOURCES

1. McKinsey, AI power: Expanding data center capacity to meet growing demand, Oct 29, 2024
2. Statista, Data generation volume worldwide 2010-2029, Nov 2025
3. International Energy Agency, Energy and AI World Energy Outlook Special Report, 2025
4. Ibid.
5. IDTechEx, How Long Until Small Modular Reactors Make an Impact on Energy Grids?, Jun 30, 2023
6. theCUBE AI, 286 | Breaking Analysis | Cloud Quarterly – Azure's AI Pop, AWS' Supply Pinch and Google's Execution, Aug 9, 2025
7. McKinsey, The data center dividend, Oct 7, 2025
8. Bloomberg, Eye-Popping Power Price Show AI's Cost to Consumers, Sep 2025
9. Ibid.
10. Kamiya, G. & Coroamă, V.C., Data Centre Energy Use: Critical Review of Models and Results, 2025
11. McKinsey, The data center balance: How US states can navigate the opportunities and challenges, Aug 2025
12. S&P Global, European data center power demand to double by 2030, straining grids, Jul 30, 2025
13. Ember, Solar and wind could power up to a third of ASEANs data centers in 2030, without needing batteries, May 27, 2025

ABOUT

exactmarket™



Exact Market
is here to help your organization understand and capitalize on technology disruptions.

We bring together market insights, content strategies, and digital execution to help enterprises, technology providers, and independent software vendors share complex transformation launches and stories clearly and compellingly.

Our Elephant in the Room series captures the meaningful conversations leaders are already having today—about cloud, cybersecurity, AI, and infrastructure—and transforms them into practical frameworks for making informed decisions.

Whether you are refining your cloud strategies, launching new products, or preparing for the next wave of innovation, Exact Market's specialized marketing resources can help align your business, technology, and brand.

WE DON'T JUST EXPLAIN TRANSFORMATION
WE GUIDE YOU IN SHAPING IT, OWNING IT, AND MOVING IT FORWARD.

Find out more @ www.exactmarket.com

exactmarket℠

ELEPHANT IN THE ROOM EBOOK SERIES