pract.ai + Wallaroo.ai
Joint Research Publication

INFRASTRUCTURE SOVEREIGNTY IN THE AGE OF AI

The Enterprise Playbook for Escaping Hardware Lock-In

How unified orchestration enables strategic flexibility in an era of chip constraints, cloud concentration, and data residency mandates

Enterprise White Paper

Contents

Executive Summary

90%

of AI projects fail to deliver ROI - most stall at the 'last mile' between prototype and production

The enterprise AI landscape has reached an inflection point. After years of rapid adoption, organizations are waking up to understand that the infrastructure powering their most strategic investments is controlled by a remarkably small number of vendors. This concentration creates fragility, limits negotiating power, and exposes enterprises to risks that few boards have not yet fully contemplated.

This white paper examines the emerging imperative of infrastructure sovereignty, which is the strategic capability to deploy, migrate, and operate AI workloads across heterogeneous environments without vendor dependency. We argue that sovereignty is not merely a technical preference but a business-critical competency that will separate resilient enterprises from vulnerable ones.

> **The Core Challenge**
>
> *The AI supply chain has a single point of failure, and most enterprises have no contingency plan. When 80% of AI training and inference runs on one vendor's hardware, concentration risk becomes existential risk.*

The path forward requires a fundamental rethinking of AI infrastructure strategy. Organizations must move from point-solution thinking to platform thinking - from hardware-specific deployments to orchestration-layer abstractions that preserve optionality while delivering performance.

This paper provides a practical framework for assessing sovereignty maturity, understanding the dimensions of AI infrastructure independence, and executing a 90-day roadmap to strategic flexibility. The goal is not to eliminate vendor relationships but to ensure that no single vendor relationship becomes an enterprise vulnerability.

1. The Concentration Problem

The modern AI stack exhibits a degree of vendor concentration which is unlike what preceded it in enterprise technology. Understanding this concentration is the first step toward addressing it.

The Hardware Bottleneck

Nvidia currently commands approximately 80-90% of the data center GPU market for AI workloads. This dominance extends beyond raw market share, the CUDA ecosystem, cuDNN libraries, and TensorRT optimization tools create deep technical dependencies that make migration costly and complex.

The practical implications are significant. Lead times for high-end GPUs routinely exceed 6-12 months. Pricing power rests entirely with the supplier. Technical roadmaps are dictated by a single vendor's priorities rather than enterprise needs.

> **Market Reality**
>
> *In 2024, data center sales for AI inference hardware exceeded $60 billion. The projected fourfold increase by 2028 will intensify competition for limited supply and deepen dependency for those without alternatives.*

The Cloud Concentration

Hyperscaler dominance compounds the hardware bottleneck. Three providers (AWS, Azure, and Google Cloud) control approximately 65% of the global cloud infrastructure market. For AI-specific services, concentration is even higher, as these providers offer the most comprehensive managed AI platforms.

This creates a dual dependency. First, enterprises are locked into both the hardware vendor (through CUDA dependencies) and the cloud provider (through proprietary APIs, data gravity, and egress costs). Second, migration becomes not just technically challenging but economically prohibitive.

The Last Mile Problem

Beyond concentration, enterprises face an execution gap. The journey from AI prototype to production (the 'last mile') is where most initiatives stall. Data scientists build compelling models that never reach deployment because the infrastructure required to operationalize them at scale is fragmented, complex, and vendor-specific.

This last mile problem compounds concentration risk: enterprises become dependent not just on specific hardware but on the specific toolchains, APIs, and deployment patterns of their chosen

vendors. Breaking free requires addressing both the infrastructure lock-in and the operational complexity simultaneously.

The Emerging Alternative Landscape

The market is responding to concentration concerns. AMD's MI300 series offers competitive performance for many workloads. Arm-based architectures from Ampere, AWS Graviton, and others provide compelling economics for inference. Specialized AI accelerators from companies such as Cerebras, Groq, and SambaNova address specific use cases with superior efficiency.

Yet adoption remains limited, not because alternatives lack capability, but because enterprises lack the infrastructure abstraction to leverage them. The missing piece is not better hardware; it is the orchestration layer that makes hardware choice a configuration decision rather than an architectural commitment.

2. The Three Dimensions of AI Sovereignty

Infrastructure sovereignty is not a single capability but a composite of three distinct dimensions. Enterprises should develop competency across all three to achieve genuine strategic flexibility.

DIMENSION 1	DIMENSION 2	DIMENSION 3
Hardware Sovereignty	*Data Sovereignty*	*Operational Sovereignty*
Chip vendor optionality	Residency compliance	Deployment flexibility
Architecture portability	Encryption at rest/transit	Migration capability
Performance parity	Access control	Vendor independence
Cost optimization	Audit capability	Air-gap readiness

Hardware Sovereignty

Hardware sovereignty means the ability to deploy AI workloads across different chip architectures (x86, Arm, various GPU vendors) without rewriting applications or accepting significant performance degradation. This requires abstraction at the inference layer and tooling that handles optimization automatically.

The goal is not to abandon any particular vendor but to ensure that vendor relationships are choices rather than constraints. An enterprise with hardware sovereignty can respond to supply disruptions, leverage competitive pricing, and adopt emerging architectures without lengthy migration projects.

Data Sovereignty

Data sovereignty encompasses the ability to control where data resides, how it moves, and who can access it. For AI systems, this extends beyond traditional data governance to include model weights, training data provenance, inference inputs, and output logging.

Regulatory requirements increasingly mandate data residency within national borders. Sector-specific regulations add complexity i.e. healthcare data cannot be processed in the same environments as general commercial workloads. Data sovereignty requires infrastructure that can enforce these boundaries while maintaining operational efficiency.

Operational Sovereignty

Operational sovereignty is the capability to deploy, manage, and scale AI workloads independently of any single provider's management plane. This includes the ability to operate in air-gapped environments, migrate workloads between environments without downtime, and maintain consistent observability regardless of deployment location.

> **Strategic Imperative**
>
> *Data sovereignty is table stakes. Infrastructure sovereignty is the next mandate. Organizations that achieve operational sovereignty gain not just compliance but competitive advantage through deployment flexibility.*

The Edge and Air-Gapped Imperative

Sovereignty takes its most concrete form at the edge and in air-gapped environments. These deployment contexts (factory floors, oil derricks, defense installations, healthcare facilities) represent the frontier of AI infrastructure independence.

Manufacturing: Computer vision models for quality assurance must run on factory floor servers with sub-millisecond latency. Cloud dependencies introduce unacceptable delays and connectivity risks.

Energy & Utilities: Predictive maintenance models on oil platforms and pipelines operate in environments with no IP connectivity. Air-gapped deployment is not optional, it is the only option.

Defense & Government: National security applications require complete network isolation. Models must be deployable via physical media (USB drives) to disconnected secure facilities.

Healthcare: HIPAA compliance and patient privacy requirements often mandate on-premises inference. Diagnostic AI cannot send patient data to external cloud services.

Telecommunications: 5G network optimization requires AI at the edge (in cell towers and local data centers) where decisions must happen in real-time without round-trip latency to centralized clouds.

> **Edge Reality**
>
> *For many enterprises, sovereignty is not an abstract strategic goal, it is a deployment requirement. Air-gapped capability separates vendors who understand enterprise reality from those who assume perpetual cloud connectivity.*

3. The Hidden Cost of Lock-In

Infrastructure lock-in carries costs that extend far beyond licensing fees and compute charges. Understanding these hidden costs is essential for building the business case for sovereignty investments.

1

Technical Debt Accumulation

Every deployment optimized for a specific hardware platform creates switching costs. CUDA-specific code, vendor-specific APIs, and hardware-tuned configurations become liabilities when alternatives emerge or existing vendors raise prices.

2

Negotiating Leverage Erosion

Vendors understand lock-in dynamics. As dependency deepens, pricing discussions become increasingly one-sided. Enterprises without credible alternatives accept terms they would otherwise reject.

3

Compliance Exposure

Regulatory requirements evolve faster than locked-in infrastructure can adapt. Data residency mandates, industry-specific requirements, and emerging AI regulations create compliance gaps that proprietary platforms cannot easily address.

4

Innovation Constraints

Lock-in limits the ability to experiment with emerging approaches. New model architectures, specialized hardware, and novel deployment patterns become inaccessible when infrastructure lacks flexibility.

5

Talent Dependency

Specialized skills for proprietary platforms create organizational fragility. Teams optimized for one vendor's ecosystem struggle to leverage alternatives, even when those alternatives offer clear advantages.

The Compounding Effect

Lock-in costs compound over time. The longer an enterprise operates within a constrained infrastructure, the more expensive escape becomes and the more attractive continued dependency appears. This is not a technical phenomenon but a strategic trap.

4. The Orchestration Layer

Architecture for Optionality

The solution to infrastructure lock-in is not to avoid vendors but to insert an abstraction layer between AI workloads and underlying infrastructure. This orchestration layer serves the same strategic function that hypervisors served for compute virtualization, it transforms hardware choice from an architectural decision into a configuration parameter.

> **The Strategic Insight**
>
> *The orchestration layer is to AI what the hypervisor was to compute, the unlock for optionality. Just as VMware enabled enterprises to escape hardware vendor lock-in for servers, AI orchestration platforms enable escape from chip and cloud concentration.*

Key Capabilities of the Orchestration Layer

A mature AI orchestration platform provides several critical capabilities that enable sovereignty:

Unified Deployment: The ability to deploy identical workloads across cloud, on-premises, and edge environments using consistent tooling. Models packaged once run anywhere without reconfiguration.

Hardware Abstraction: Automatic optimization for different chip architectures. The same model performs efficiently on Nvidia GPUs, AMD GPUs, Arm CPUs, and specialized accelerators without manual tuning.

Auto-Packaging: Automatic conversion of models from any framework (PyTorch, TensorFlow, ONNX, Hugging Face) into optimized, portable formats. This eliminates manual deployment engineering and ensures consistency across environments.

Workload Migration: The capability to move running workloads between environments without downtime. This enables response to supply constraints, cost optimization, and compliance requirements.

Centralized Observability: Consistent monitoring, logging, drift detection, and performance tracking across all deployment environments, from cloud to factory floor to air-gapped facility.

Resource Orchestration: Dynamic allocation and scaling of compute resources based on demand, with automatic routing to the most cost-effective available infrastructure and intelligent batching for throughput optimization.

The Performance Question: Abstraction Without Penalty

Critics argue that abstraction layers add overhead and reduce performance. This concern misunderstands modern orchestration technology. Well-designed platforms, particularly those built with performance-first languages such as Rust, add minimal latency while enabling optimizations that would be impractical to implement manually.

High-performance inference engines built in systems languages deliver C-like execution speeds with memory safety guarantees. This approach (drawn from high-frequency trading and real-time systems) enables sub-millisecond latency even in resource-constrained edge environments.

Smart batching and caching maximize hardware utilization automatically. Dynamic batching adjusts to workload patterns, and KV-cache optimization accelerates LLM inference without manual tuning.

Hardware-aware optimization routes different pipeline stages to appropriate compute resources (CPU for preprocessing, GPU for inference, specialized accelerators for specific operations) without requiring developers to manage this complexity.

Up to 80%

infrastructure cost reduction may be achieved through optimized resource utilization and hardware-appropriate workload routing

The Economics of Abstraction

The economic benefit compounds over time. Initial investment in orchestration infrastructure pays dividends through reduced vendor lock-in, improved negotiating position, and the ability to leverage emerging hardware without migration costs. Organizations report deployment time reductions from months to days, infrastructure utilization improvements of 50-80%, and the operational capacity to scale from dozens to thousands of model deployments.

Perhaps most importantly, the orchestration layer transforms AI infrastructure from a specialized capability requiring dedicated engineering teams into a self-service platform that data scientists can leverage directly, freeing up to 40% of team capacity previously consumed by deployment mechanics.

5. The Sovereignty Maturity Model

Enterprises exist along a spectrum of infrastructure sovereignty. This maturity model provides a framework for assessment and a roadmap for advancement.

1

Level 1: Captive

Single-vendor dependency with no migration path. Workloads are optimized for specific hardware and deployed through proprietary platforms. Migration would require complete re-architecture. Most enterprises begin here.

2

Level 2: Portable

Workloads use standard formats (ONNX, containerization) and avoid vendor-specific optimizations. Migration is technically possible but operationally complex. The organization has documented alternatives but has not tested them.

3

Level 3: Hybrid

Active deployment across multiple environments with proven migration capability. The organization regularly moves workloads based on cost, performance, and compliance requirements. Orchestration tooling is in place and air-gapped deployment is validated.

4

Level 4: Sovereign

Full infrastructure independence with real-time workload routing. Hardware and cloud choices are configuration parameters. The organization can respond to supply disruptions, price changes, or regulatory requirements within hours rather than months.

> **Maturity Assessment**
>
> *Most enterprises self-assess at Level 2 but operate at Level 1. The test is simple: if your primary GPU vendor doubled prices tomorrow, how long would it take to migrate? If the answer is measured in months, you are Captive regardless of your technical architecture.*

6. 90-Day Sovereignty Roadmap

The journey to infrastructure sovereignty begins with assessment and proceeds through progressive capability building. This 90-day roadmap provides a structured approach to initial sovereignty investments.

PHASE 1
Days 1-30

Audit

- Inventory all AI workloads and their infrastructure dependencies
- Map vendor relationships and contractual constraints
- Identify CUDA-specific code and proprietary API usage
- Assess current maturity level using the sovereignty framework
- Document compliance requirements, data residency constraints, and air-gap needs

PHASE 2
Days 31-60

Abstract

- Evaluate and select orchestration platform with multi-environment support
- Implement abstraction layer for pilot workloads (start with 2-3 models)
- Establish model packaging standards (ONNX, containerization, auto-packaging)
- Deploy unified observability across cloud and edge environments
- Train operations team on multi-environment management and self-service deployment

PHASE 3
Days 61-90

Diversify

- Qualify alternative hardware vendors (AMD, Arm, CPU-based inference)
- Establish secondary cloud, on-premises, or edge capability
- Execute migration test for critical workloads (including air-gapped if required)
- Benchmark performance across environments; validate cost savings
- Negotiate improved terms with current vendors using demonstrated alternatives

Implementation Principle

Portability is not a feature. It is a strategic hedge against market disruption. The 90-day roadmap establishes baseline capability; ongoing investment extends sovereignty to additional workloads and environments.

Conclusion: Board-Level Truths

Infrastructure sovereignty is not a technical preference but a strategic imperative. As AI becomes central to competitive advantage, the infrastructure enabling that AI becomes a critical business asset requiring board-level attention.

Board-Level Truth #1

Concentration risk in AI infrastructure is enterprise risk. When 80% of AI capability depends on a single hardware vendor and three cloud providers, supply chain disruption becomes business continuity threat. Boards must understand and quantify this exposure.

Board-Level Truth #2

The window for achieving sovereignty is narrowing. Every month of continued lock-in deepens dependency and increases switching costs. Organizations that invest in sovereignty now will have strategic options that latecomers cannot easily replicate.

Board-Level Truth #3

Orchestration is the enabler. Just as enterprises invested in virtualization to escape server vendor lock-in, investment in AI orchestration platforms is the path to infrastructure independence. The technology exists; the question is whether organizations will deploy it before lock-in becomes irreversible.

> **The Strategic Question**
>
> *In a world where AI capabilities increasingly define competitive advantage, can your organization afford to have those capabilities controlled by infrastructure you do not own, cannot migrate, and cannot replace? Infrastructure sovereignty is the answer to this question.*

About the Authors

pract.ai can help enterprises navigate AI governance with practical frameworks and implementation support. We believe AI should enable innovation, not obstruct it.

Wallaroo offers a unified AI Orchestration platform built to power the next generation of enterprise AI. The Wallaroo software streamlines the deployment, execution, and operation of AI at scale. It combines state-of-the-art, enterprise-grade inference with portability across diverse cloud to edge infrastructure as well as a unified AI hub to orchestrate and manage the AI.