RADIANT

WHITEPAPER

Tenancy Determines Trust

The Architecture of Secure, Programmable AI Clouds

A whitepaper for operators, sovereign buyers and AI infrastructure leaders.

radiant.co

Contents

1. The problem tenancy must solve

AI is scaling faster than the infrastructure that supports it. Capacity is scarce and demand is growing with a force that breaks traditional planning cycles. Moreover, AI workloads now span training, tuning, inference and retrieval with wildly different performance and governance requirements.

For the better part of a decade, cloud providers have masked complexity behind multitenant architectures optimized for their own efficiency. That approach translates poorly for AI.

Four hard truths define the landscape:

1. **Performance is unpredictable when tenants collide**: AI workloads are bursty. A single misbehaving job can introduce jitter across an entire slice of the cluster. Hyperscalers solve this with over-provisioning, not isolation.
2. **Compliance boundaries cannot be abstracted away**: Sovereign programs, regulated industries and national AI factories require proof of isolation at the hardware level. Logical boundaries are not enough. You cannot achieve sovereignty by placing US-controlled cloud hardware inside your borders. Unless the control plane, data paths and operational authority are severed from the parent jurisdiction, the environment remains exposed.
3. **GPU supply is constrained**: GPU supply is constrained in every market segment. Lead times stretch into quarters, not weeks, and operators cannot buy their way out of scarcity. Every idle GPU hour is a direct economic loss because the fixed cost of the cluster continues regardless of utilization. The economics of AI infrastructure leave no room for underuse. Sharing becomes a requirement at scale, but it only works when isolation holds and performance remains predictable.
4. **Cloud control planes were not designed for AI**: Traditional cloud control planes and the NVIDIA software stack were not built to natively support fine-grained, multi-tenant AI environments. When a single large GPU cluster must be safely and dynamically partitioned into multiple smaller clusters, operational gaps quickly emerge. Achieving secure isolation, predictable performance, and fair resource allocation requires significant automation across compute, storage, and networking—well beyond what existing control planes provide out of the box.

AI infrastructure providers need a new primitive. Not a repackaged VPC and not a thin wrapper around Kubernetes. The primitive itself is tenancy—designed for AI from first principles. It must support multiple tenancy modes, from strict isolation to softer, utilization-optimized sharing, while still delivering predictable performance, enforcing strong boundaries, raising utilization, and proving sovereignty on demand.

2. The Tenancy Spectrum

Tenancy is not a binary choice between shared and dedicated. It is a spectrum of isolation, governance and performance guarantees. The Radiant AI Fabric defines this as three modes calibrated for real workloads.

Soft Tenancy

Shared GPUs with namespace separation. Ideal when utilization is the priority and governance risk is low.

Benefits

- Maximum fleet utilization
- MIG partitions for hardware-level QoS
- Rapid experimentation across teams

Use cases

Internal AI teams, inference-heavy pipelines, RAG workloads, startups optimizing cost.

Strict Tenancy

Dedicated GPU nodes with isolated data planes, governed by a shared control plane. Provides hard performance and security boundaries while retaining centralized automation, policy, and lifecycle management.

Benefits

- Predictable, non-contended performance
- Isolated compute, storage paths, and network planes
- Centralized identity, policy enforcement, and observability
- No noisy-neighbor impact on workloads

Use Cases

Business-critical enterprise workloads, financial models, telco AI services, model training with strict SLOs.

Private Tenancy

A dedicated control plane plus dedicated GPUs. The gold standard for sovereignty, privacy and operational separation.

Benefits

- Complete isolation
- Independent lifecycle management
- Deterministic governance with auditable boundaries

Use Cases

Sovereign clouds, national AI programs, regulated industries, air-gapped deployments.

Why These Three Modes Matter

Every AI buyer is trapped between two extremes: the cost of sharing compute and the risk of sharing compute. The answer is not more shared infrastructure or more dedicated clusters. The answer is a programmable tenancy spectrum that adapts to each workload, each customer and each regulatory environment.

Hyperscalers do not offer this. Most GPU clouds cannot. Systems integrators bolt on strict tenancy off-platform without automation. Only a cloud-native control plane with hardware-aware segmentation can move fluidly between Soft, Strict and Private.

3. Why Tenancy Determines Trust

Trust is now the gating resource for AI adoption. Even in a capacity-constrained market trust remains the gating resource for adoption because organizations will not place mission-critical workloads into environments they cannot verify.

Trust has three dimensions:

1. **Isolation**: The customer must know that data, traffic and execution paths are not shared unless explicitly allowed. Hardware segmentation, encrypted volumes and network partitioning must enforce this, not policy documents.
2. **Predictability**: AI pipelines break under jitter. Training costs inflate when jobs slow down. Trust requires performance that is not influenced by unknown neighbors or regional congestion.
3. **Sovereignty**: Governments and regulated institutions must prove that their workloads remain within boundaries they control. Residency and compliance are not contractual features. They are architectural outcomes.

Tenancy is the mechanism that connects architecture to trust. Without a robust tenancy model, no amount of hardware scale or marketing narrative can convince a sovereign buyer or enterprise operator that the platform is safe for their data or stable for their workloads.

4. Why NVIDIA Cares

NVIDIA's Cloud Partner program is defining the reference architecture for the next wave of AI clouds. These clouds must:

- Serve multiple customers on the same fleet
- Provide sovereign or private environments programmatically
- Guarantee performance isolation at scale
- Deliver automation that aligns with NVIDIA's GPU roadmap

What is missing in the ecosystem is a control plane that provides strict and private tenancy without duplicating infrastructure or introducing heavy operational overhead. NVIDIA needs partners that can create on-demand private environments for regulated, sovereign and enterprise buyers.

Radiant delivers this through a control plane that can carve and re-carve clusters in minutes. No one else in the market provides strict and private tenancy under a single platform.

5. The Architecture of Secure Segmentation

Tenancy is only real when enforced across all three planes of the stack.

Compute Segmentation

- MIG for hardware-isolated GPU partitions
- SR-IOV for virtual GPU instances
- Node-level fencing for strict and private environments
- GPU-aware scheduler for consistent performance

This ensures that one job cannot degrade another regardless of priority or workload type.

Network Segmentation

- EVPN and VXLAN overlays for tenant-specific fabrics
- InfiniBand partitioning for high-performance isolation
- Virtual NICs for strict separation of data paths
- Policy-driven routing

Boundaries hold only when the network is part of the segmentation model.
AI workloads are network-heavy. Isolation must extend to the fabric.

Storage Segmentation

- Encrypted volumes per tenant
- Namespace isolation
- Policy-based access controls
- Auditability from the control plane

AI pipelines depend on predictable access to datasets, checkpoints and embeddings.
Storage isolation is as important as compute isolation.

Control Plane Governance

- Centralized policy enforcement
- Dedicated identity per tenant
- RBAC, Quotas, audit logs and cost attribution
- API-driven provisioning of Soft, Strict and Private environments

The control plane is the arbiter of boundaries. Without it, segmentation becomes a collection of low-level settings without consistency.

6. Economic Implications

Tenancy is not only a security feature. It is an economic multiplier.

The economics of AI infrastructure are defined by one variable above all others: utilization. GPUs are scarce, capital-intensive and power-hungry. Every hour they sit idle erodes the business case for the entire platform. A tenancy model that cannot raise utilization without degrading performance will fail on the balance sheet long before it fails in production.

A programmable tenancy system changes this equation.

Soft tenancy increases fleet yield by reclaiming idle GPU cycles while maintaining hardware-backed boundaries. Provisioning resources happens in a matter of seconds. Instead of stranded capacity scattered across clusters, operators convert partial loads into productive compute. This shift reduces effective cost per model iteration and accelerates the payback period of deployed hardware.

Strict and private tenancy unlock higher-value workloads without requiring duplicate infrastructure. In traditional architectures, satisfying isolation demands from regulated or sovereign buyers requires operators to stand up separate fleets. This process is largely manual or only partially automated, taking days to weeks to build, provision, and validate—and remains error-prone and costly to maintain over time.

Under a unified control plane that enforces strict and private boundaries programmatically, the same physical fleet can serve mixed customer types. Isolation, governance, and lifecycle management are automated end-to-end, allowing new environments to be instantiated in minutes and continuously maintained without bespoke operational workflows. The result is compressed capital expenditure, simplified procurement, and materially lower long-term TCO.

Finally, predictable performance becomes an economic lever. Stable, interference-free throughput ensures training cycles complete on schedule and inference pipelines operate within budget. Variability translates into cost drift. Isolation eliminates that drift.

Together these effects turn tenancy into an economic multiplier.

Platforms that master this dynamic outperform on cost, scale more efficiently and offer price points competitors cannot match. In AI infrastructure the return curve is dictated by how well a provider can share without compromising trust. Programmable tenancy is the mechanism that makes that possible.

7. Market Impact

Programmable tenancy does more than secure workloads. It expands the addressable market for AI infrastructure.

Most providers are constrained not by demand but by the types of customers they can serve. Each segment imposes different requirements on isolation, governance and performance. A single-mode tenancy model limits which markets a platform can enter.

A programmable spectrum of Soft, Strict and Private tenancy changes the strategic landscape.

Sovereign AI clouds gain a model that aligns with national requirements. They can offer shared national infrastructure for general workloads yet create fully private, hardware-isolated environments for sensitive missions on demand. This replaces the practice of building multiple sovereign clusters with a single, flexibly partitioned environment.

Enterprise AI factories benefit from the internal version of the same pattern. Large organizations can consolidate scattered AI efforts into shared infrastructure while providing dedicated, isolated environments for business units with strict compliance needs. This increases internal adoption and reduces the cost of AI expansion.

Cloud service providers unlock a tiered service strategy. They can offer cost-efficient shared environments for broad segments while capturing premium value from regulated or performance-sensitive customers who require strict or private tenancy. Few CSPs can execute this model today because their control planes cannot enforce boundaries programmatically.

Telcos emerge as regional AI providers only when they can support mixed customer profiles on the same fleet. Tenancy transforms telco infrastructure into multi-tenant AI platforms capable of serving enterprises, governments and application providers concurrently.

Programmable tenancy does not just secure workloads. It creates markets by enabling infrastructure providers to serve customer groups that were previously inaccessible or uneconomical.

8. Proof Points

The model is no longer theoretical. It is validated in production by institutions that cannot tolerate architectural risk.

Dell has a Radiant SKU because no other system could provide strict multitenancy as a licensable, operator-controlled capability. It is foundational to their AI Factory strategy. This decision reflects a structural gap in the ecosystem, not a feature comparison.

NVIDIA is evaluating the tenancy model for inclusion in the NCP reference architecture because the industry lacks a unified way to generate private, hardware-isolated environments programmatically. NCP clouds must support sovereignty, performance guarantees and shared economics. Only a programmable tenancy system can satisfy all three.

Sovereign and telco buyers have prioritized the model in competitive evaluations because it offers something hyperscalers cannot: strict and private tenancy under a single control plane without duplicating infrastructure or compromising performance.

In each case the same pattern appears.

Platforms that must meet the highest standards of governance, scale and performance have converged on a tenancy model that treats isolation as a programmable primitive. Competitors are fragmenting their designs to approximate this capability. None achieve full-stack segmentation with cloud-native automation.

The proof is not anecdotal. It is structural.

Organizations with the strongest constraints have validated this model first because they have no tolerance for architectural ambiguity.

9. Conclusion

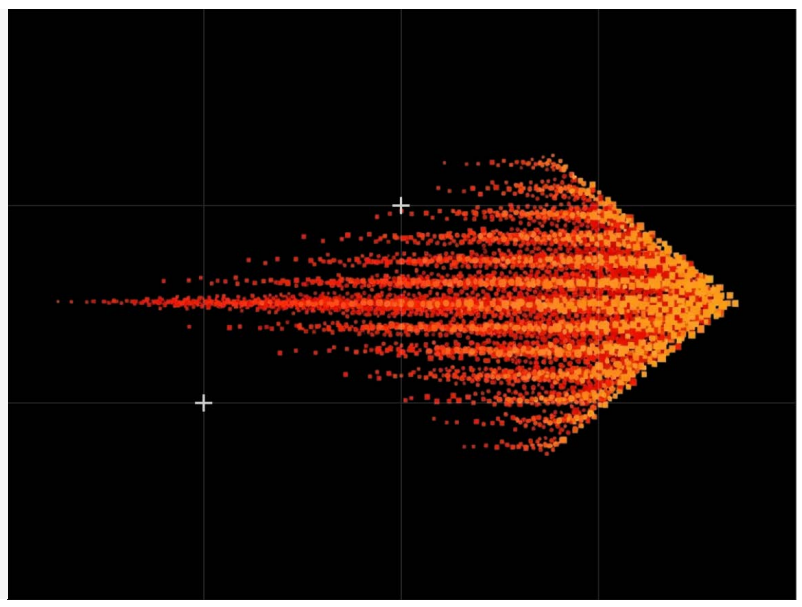
Tenancy determines trust. Trust determines adoption. Adoption determines scale.

AI clouds that cannot provide programmable, hardware-backed tenancy will fail to meet the demands of sovereign buyers, regulated enterprises and AI-native organizations. The market is shifting from capacity scarcity to trust scarcity. The winners will be the operators who deliver isolation and performance as programmable primitives, not static configurations.

Soft tenancy maximizes utilization. Strict tenancy guarantees performance. Private tenancy proves sovereignty.

A single control plane that orchestrates all three is not a feature. It is the foundation of modern AI infrastructure. This is the architecture the industry will standardize on. This is the architecture NVIDIA needs in NCP. This is the architecture operators will demand.

The future of AI infrastructure is not shared or dedicated. It is programmable.

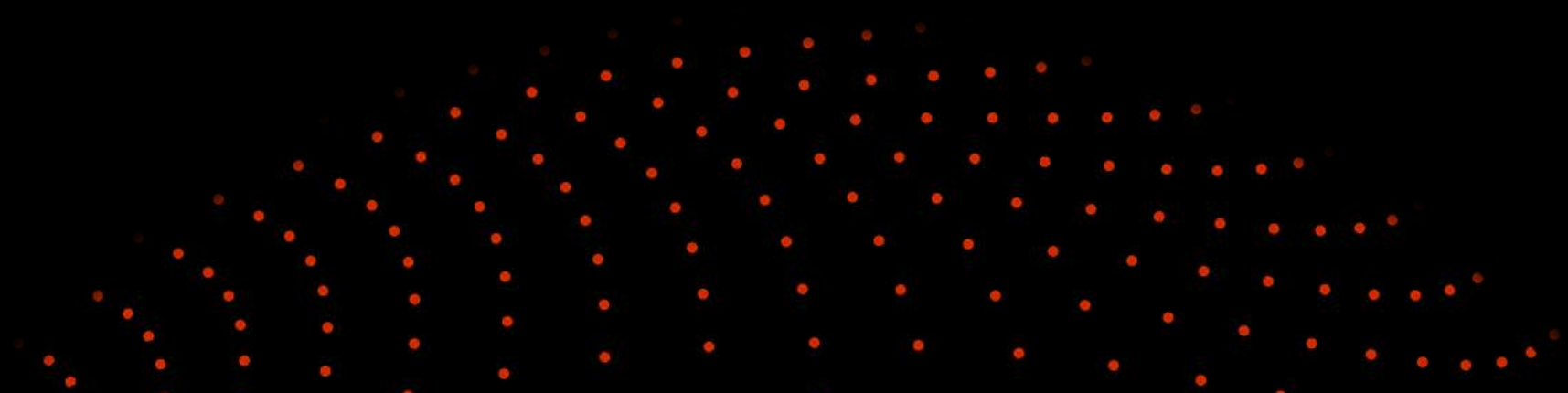


RADIANT

About Radiant

Radiant is building the utility model for compute - ubiquitous, always on and offering superior economics. Through vertical integration of capital, powered land, compute, and software, we deliver sovereign-grade control for nations, turnkey certainty for enterprises, and limitless capacity for builders. This model will democratize AI, equipping every individual with the ability to unlock their creative potential and tackle world-changing problems.

LEARN MORE AT RADIANT.CO



radiant.co