WHITE PAPER

The Coming Inference Divide

Why where you run AI will matter as much as which AI you run

May 2026 | Confidential

COLOVORE

Table of Contents

Executive Summary

Artificial intelligence infrastructure is entering a structurally different phase than the one that defined the last decade of cloud computing. Training and inference, often discussed together, are diverging economically, operationally, and strategically. Training is consolidating into a small number of hyperscalers and frontier labs. Inference is moving in the opposite direction.

The risk is not about predicting exactly when fragmentation arrives at full scale. It is about what happens to enterprises that make the wrong infrastructure commitment at a moment of genuine architectural uncertainty. Infrastructure decisions made today carry five to seven-year economic lives. The probability that AI inference hardware and deployment models change materially within that window is not a theoretical concern. It is high enough to make lock-in risk a board-level issue.

Enterprises that commit deeply to a single hyperscaler inference stack, or to a purpose-built facility optimized for a single chip architecture, are making the same category of mistake that defined each prior technology transition, from mainframe to client-server, from client-server to virtualization, from on-premises to cloud. Each transition left enterprises with stranded assets, competitive disadvantage, and expensive migrations. The AI inference transition is structurally similar. The window to avoid that trap is now.

Colovore's strategy is built around this reality. Rather than betting on a single winning silicon architecture, Colovore positions itself as the neutral, high-density, low-latency execution layer where multiple inference platforms can coexist. By designing facilities capable of supporting extreme power density, advanced cooling, and diverse accelerator form factors in critical metro locations, Colovore absorbs silicon innovation rather than chasing it. The more fragmented and specialized inference hardware becomes, the more valuable this neutral execution layer becomes.

The question is not whether enterprises should wait for fragmentation to fully materialize before acting.

The question is whether the cost of being wrong about a five-year infrastructure bet exceeds the cost of building flexibility in from the start.

We believe it does, by a wide margin.

1.

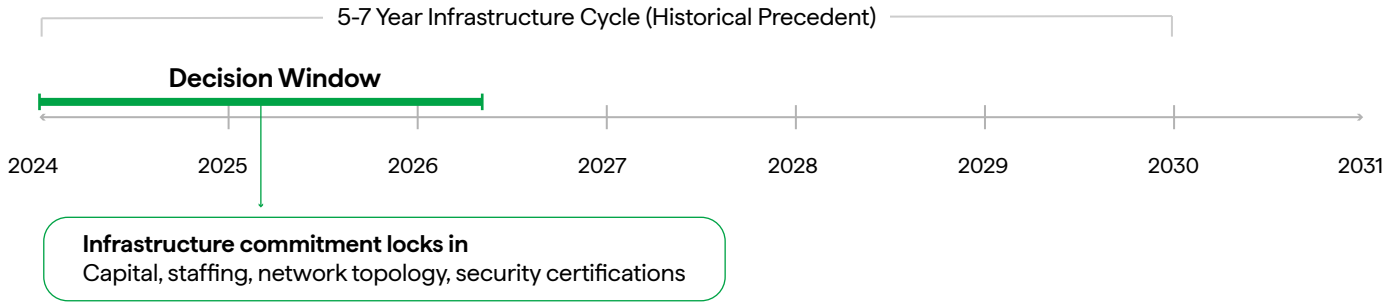
The Infrastructure Decision Enterprises Are Making Right Now

Most large enterprises are currently somewhere in the early stages of committing to an AI inference infrastructure strategy. For many, that means negotiating enterprise agreements with hyperscalers, evaluating GPU cloud options, or in some cases considering purpose-built on-premises or colo deployments optimized for a specific chip architecture.

Each of these decisions has a multi-year economic life. Enterprise infrastructure commitments are not quarterly experiments. They involve capital expenditure, depreciation schedules, staffing, network topology, security certifications, and operational processes that do not change quickly. A decision made today about where and how to run inference will shape enterprise AI capability through 2030 and beyond.

The risk embedded in that decision is asymmetric. If an enterprise commits to a hyperscale inference stack and inference hardware and economics remain stable, the cost of that commitment is predictable. If inference hardware fragments, if specialized silicon displaces general-purpose GPUs for specific workloads, if latency requirements tighten, if data sovereignty regulations intensify, or if the economics of owned infrastructure shift relative to cloud token pricing, the cost of unwinding a deep hyperscale commitment becomes very large.

The history of enterprise technology suggests that major infrastructure transitions happen on roughly five to seven year cycles. We are currently in the early years of AI inference deployment at enterprise scale. The probability that the hardware landscape, deployment models, and economics look materially different by 2030 is not speculative. It is the base case.



Enterprises are not being asked to predict the future precisely.

They are being asked to avoid building infrastructure that cannot absorb it.

2.

Training and Inference Are Diverging, Not Converging

Training and inference are frequently described as two sides of the same AI infrastructure problem. In practice, they behave very differently and are driven by different economic incentives.

Training is episodic, extremely capital intensive, highly parallelizable, and largely insensitive to latency. What matters most is aggregate throughput and utilization over time. These characteristics strongly favor large, homogeneous clusters operated at massive scale. As a result, training naturally consolidates into a small number of hyperscalers and frontier labs with access to cheap capital, cheap power, and the operational discipline to run very large fleets.

Inference is fundamentally different. Inference is the continuous execution of trained models to produce outputs that drive real-world actions. It runs all the time. It is frequently latency sensitive, sometimes extremely so. It is tightly coupled to enterprise data, workflows, and regulatory obligations. It is increasingly embedded in systems where tail latency, jitter, and determinism matter more than peak throughput.

As inference becomes core intellectual property rather than an experimental capability, decision-making authority shifts away from research teams and toward CIOs, CTOs, CFOs, and risk and compliance leaders. These stakeholders prioritize predictability, control, and long-term economics. This shift is the root cause of inference fragmentation.

Infrastructure options being evaluated now:

Hyperscaler
Enterprise agreements

- Capex via contracts
- Deep stack integration
- Multi-year pricing
- Shared depreciation

GPU cloud
Spot / reserved capacity

- Shorter commitments
- Hardware-agnostic
- Scaling flexibility
- Vendor portability

On-prem / colo
Purpose-built silicon

- Owned hardware
- Chip-architecture bet
- Sovereignty control
- Full depreciation sched.

Commitment Depth & Unwind Cost

Lower — Higher

3.

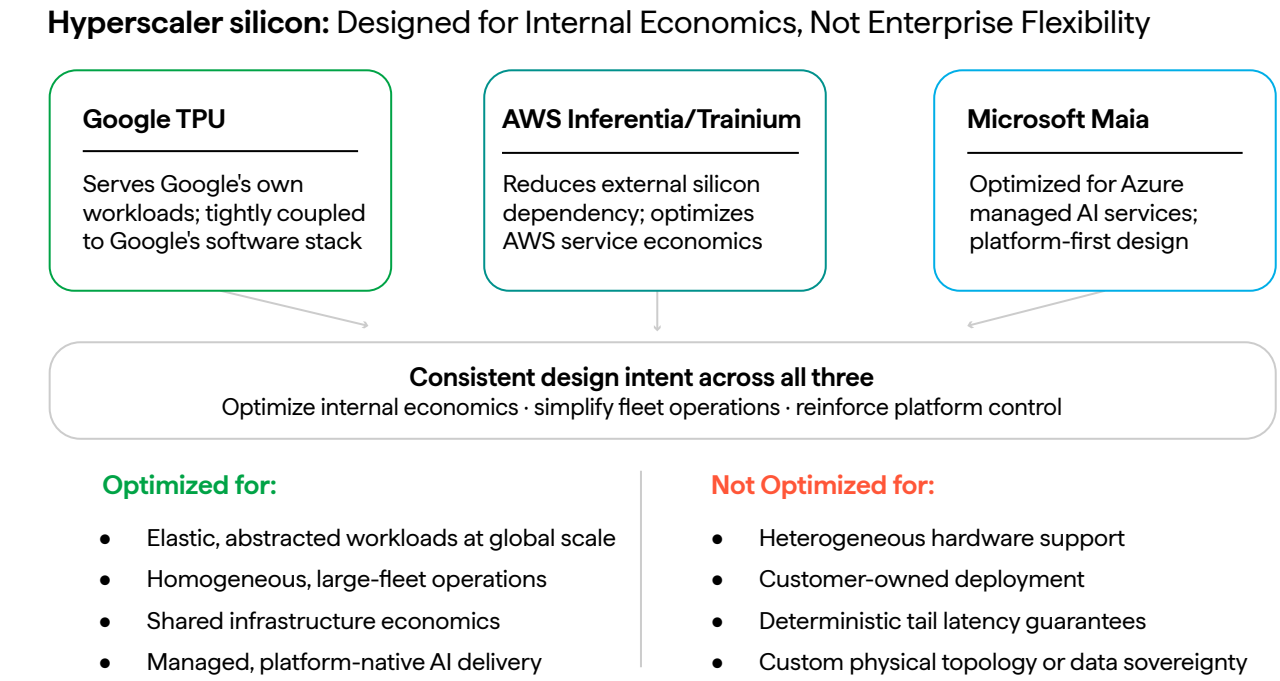
Why Hyperscaler Inference Has Structural Limits

Hyperscalers are rational actors. Their inference strategies are optimized for a specific operating model built around massive, homogeneous fleets. These platforms excel at elastic, abstracted workloads delivered at global scale. Understanding why they excel also explains what they cannot do.

Google TPUs are designed primarily to serve Google's own workloads and are tightly coupled to Google's software stack. AWS Inferentia and Trainium are designed to reduce dependency on external silicon vendors and optimize the economics of AWS services. Microsoft Maia is optimized for workloads that fit Azure's managed AI services. Across all three, a consistent pattern emerges: hyperscaler chips are designed to optimize internal economics, simplify fleet operations, and reinforce platform control.

What they do not optimize for are heterogeneous hardware support, customer-owned deployment, deterministic tail latency, or custom physical topology. These omissions are not failures. They are the inevitable result of designing silicon for fleet-scale abstraction.

For enterprises whose inference requirements align with elastic, abstracted delivery at global scale, hyperscaler inference is a reasonable choice. The problem is that a large and growing class of enterprise inference workloads does not align with that model, and the enterprises running those workloads are currently making infrastructure commitments that will constrain them for years.



4.

The Use Cases That Break in Abstracted Environments

The following categories of enterprise inference workloads share a common characteristic: they degrade materially when placed behind multi-tenant schedulers, geographic distance, or platform-managed upgrade cycles. These are not edge cases. They represent a large and growing share of enterprise AI deployment

Financial Services and Real-Time Decisioning

- Real-time transaction fraud scoring at sub-10ms latency with proprietary feature stores, model versioning tied to regulatory audit requirements, and behavior that cannot change without formal change control. The model cannot phone home to a shared environment and the latency requirement cannot tolerate geographic round trips.
- High-frequency and algorithmic trading systems where inference is embedded directly in order routing and risk decisions. These systems already run co-located infrastructure for microsecond advantages. Adding AI inference to those decision loops only intensifies the physical proximity requirement.
- Real-time insurance underwriting and claims decisioning where a customer is in a digital funnel and the model output determines pricing or approval in under a second. Regulatory audit trail requirements mean platform-managed upgrades are not acceptable.
- Relationship manager decision support in banking, where inference runs against a customer's full transaction history, current market conditions, and proprietary risk models simultaneously during a live customer interaction. The data cannot leave the perimeter and the latency must be invisible to the human conversation.

Healthcare and Life Sciences

- Clinical decision support at the point of care, where HIPAA constraints, model explainability requirements, and the need to run inference against a patient's longitudinal record in real time all combine to make hyperscale cloud delivery problematic.
- Pharmaceutical and biotech drug discovery pipelines where proprietary compound libraries and genomic datasets represent core intellectual property that cannot transit a shared network or reside in a multi-tenant environment during inference.
- Adaptive clinical trial management where reinforcement learning models are adjusting treatment protocols based on patient response data in near real time, and reproducibility requirements mean complete control over the execution environment is mandatory.

Manufacturing and Industrial Control

- Robotic process automation in manufacturing where the inference model is controlling physical equipment and the feedback loop between sensor input, inference output, and actuator command must complete in milliseconds. A cloud round trip introduces physics-based latency that no software optimization can overcome.
- Autonomous vehicle and advanced driver assistance system validation pipelines where inference runs

continuously against sensor data and the feedback loop between inference output and system behavior is extremely tight.

- Energy grid management and predictive load balancing where inference models are integrated with SCADA systems and the physical consequences of a wrong or delayed output are material. These environments also carry near-air-gapped requirements that are fundamentally incompatible with hyperscale cloud delivery.

Agentic and Reinforcement Learning Systems

- AI agent assist platforms where agents are chaining multiple model calls together in real time, orchestrating workflows across internal systems, and maintaining stateful context across interactions. Latency compounds across hops. Abstraction layers multiply the problem.
- Reinforcement learning feedback loops in production systems where the model is continuously updating based on observed outcomes and the tightness of the feedback loop determines the rate of improvement. These systems require stable, deterministic execution environments that cloud schedulers cannot reliably provide.
- Personalized medicine and precision treatment systems where RL models adjust protocols based on patient response, and regulatory reproducibility requirements make controlled, auditable execution environments non-negotiable.

Regulated, Sovereign, and Sensitive Environments

- Government and defense contractor workloads where FedRAMP, ITAR, or classified data handling requirements create hard constraints on where inference can physically execute. This is a large and growing market that is structurally excluded from standard hyperscale inference offerings.
- Legal and financial document processing where attorney-client privilege or material non-public information constraints create hard boundaries around where models can run and what data can leave the enterprise perimeter.
- Supply chain optimization and dynamic pricing systems running continuous inference against live inventory, logistics, and demand signals across thousands of SKUs simultaneously, deeply integrated with ERP and order management systems. Inference behaves as an application component, not an external API.

What unites all of these use cases is that inference has moved from being a service to being an application component. It sits inside a larger system with defined inputs, explicit dependencies, and measurable consequences when behavior changes or degrades.

That shift changes everything about the infrastructure requirements.

5.

The Economics of the Current Decision Window

The economic argument for rethinking hyperscale inference dependency is not abstract. Enterprises are currently at an inflection point where several forces are converging simultaneously.

First, the fully loaded cost of hyperscale inference is not declining at the same rate as headline token pricing. Data egress fees, network transit costs, latency penalties that require over-provisioning, and the engineering overhead of working around abstraction limits are real costs that do not appear in per-token pricing comparisons. **Beyond compute costs, companies can expect egress, networking, and storage fees amounting to an additional 20-40% of their bill. Standard egress fees alone estimate $0.09 per GB for AWS, Azure, and GCP.** As inference becomes embedded in high-volume production systems, these costs compound.

Second, owned inference hardware is depreciating faster than any prior generation of enterprise compute because the silicon roadmap is moving so quickly. **Enterprise cloud re-platforming costs range from $40K to $600K+ when switching infrastructure stacks. Meanwhile, 30-50% of GPU spend may be wasted on idle, over-provisioned capacity.** This creates a genuine tension: enterprises want long-lived infrastructure assets with predictable economics, but inference silicon is volatile by design.The resolution to this tension is not to avoid ownership. It is to own the infrastructure layer that is stable, meaning the facility, the power, and the cooling, while retaining flexibility on the hardware layer that is volatile.

Third, the neo-cloud model that many emerging inference silicon vendors have adopted is transitional, not strategic. Operating a cloud requires continuous capital investment in facilities, networking, staffing, security, and compliance. **Due to overlooked costs, organizations estimate they've spent 25-35% more than planned in the first 12 months post-migration.** Inference startups cannot amortize these costs the way hyperscalers do. As usage grows, neo-cloud vendors face margin compression as customers compare token pricing directly with hyperscalers. More importantly, selling inference as an API abstracts away the very hardware differentiation that made the silicon valuable. For most inference silicon companies, the sustainable end state is selling hardware into environments where customers own the infrastructure. Neo-cloud is a bridge to credibility, not the destination.

The window to establish flexible, neutral execution infrastructure before lock-in closes is now.

Enterprises that wait until fragmentation is fully visible will find that the cost of transition has already been incurred.

6.

The Emergence of Specialized Inference Silicon

Inference fragmentation creates space for specialization. As workloads diverge, new silicon architectures emerge to optimize for specific bottlenecks rather than generality. NVIDIA H100, H200, and Blackwell currently account for the overwhelming majority of production enterprise inference — but the specialized silicon ecosystem is real and growing. Fragmentation is already underway. The pace of adoption will accelerate as these use cases move from pilot to production.

Deterministic, Ultra-Low-Latency Platforms

Groq's architecture is designed around compile-time scheduling rather than dynamic runtime scheduling, delivering extremely predictable token latency. This makes it attractive for interactive agents, real-time decision systems, and workflows where tail latency determines business outcomes. These platforms lose much of their value when placed behind multi-tenant schedulers or abstracted cloud APIs.

Cerebras, while initially known for wafer-scale training systems, is increasingly positioned as a deterministic, high-throughput inference platform for very large models. Its inference deployments favor dedicated, single-tenant environments where consistent performance and memory locality matter more than elastic scaling. Cerebras systems are fully liquid cooled by design and represent some of the highest-density inference-capable platforms available.

The NVIDIA Repositioning

NVIDIA is not displaced by inference fragmentation. It is repositioned. As hyperscaler fleets adopt proprietary accelerators, NVIDIA's most advanced systems increasingly deploy outside those environments in customer-owned and partner-operated infrastructure. The next phase of inference, which increasingly blends real-time inference with reinforcement learning and feedback loops, amplifies the importance of low-latency execution between tightly coupled components. Neutral, high-density metro facilities become a natural home for NVIDIA-based inference, aligning the interests of silicon providers, enterprises, and infrastructure operators rather than placing them in conflict.

Memory-Bound and Long-Context Platforms

SambaNova targets the growing mismatch between compute and memory with a dataflow-oriented architecture that emphasizes efficient data movement. Its systems support long-context inference and concurrent execution of multiple models, increasingly important for enterprise knowledge systems and agentic workflows. d-Matrix approaches the same problem with digital in-memory compute techniques to reduce the energy and latency cost of memory access during inference.

Open-Stack and Enterprise-Sovereignty Platforms

Intel Gaudi is positioned as an enterprise-friendly alternative that integrates into existing data center environments while avoiding deep lock-in to hyperscaler-specific platforms. Tenstorrent emphasizes architectural flexibility and open tooling, appealing to enterprises concerned about long-term sovereignty and vendor risk. AMD Instinct accelerators, with very high memory bandwidth and large on-package memory, are

increasingly adopted for inference workloads that benefit from memory locality, and are sold broadly into the enterprise market without a proprietary control plane.

7.

Why Single-Vendor Inference Strategies Are Structurally Unstable

The most important strategic conclusion is that single-vendor inference strategies are structurally unstable over time. This instability is not driven by vendor incompetence. It is driven by the pace of inference innovation, the diversity of enterprise workloads, and the mismatch between hardware evolution and infrastructure depreciation.

Inference hardware is evolving faster than traditional enterprise infrastructure cycles. New architectures emerge to address specific bottlenecks. At the same time, model architectures, context requirements, and application patterns continue to change. Enterprises cannot rationally commit to a single inference platform for five to seven years when the underlying performance envelope and cost structure are shifting on much shorter timelines.

Single-vendor inference strategies also break down because no single architecture can optimally serve the full spectrum of enterprise inference workloads. Low-latency interactive agents, long-context reasoning systems, high-volume batch inference, vision pipelines, and control systems place conflicting demands on hardware. Any platform that attempts to generalize across all of these inevitably makes tradeoffs that leave value on the table.

The only viable way to reconcile the tension between long-lived infrastructure assets and volatile inference silicon is to decouple hardware choice from the physical execution environment. Neutral infrastructure allows enterprises to change inference platforms without changing facilities, network topology, or operational models.

The enterprises that win the next decade of AI competition will not be those that picked the right chip in 2026.

They will be those that built infrastructure flexible enough to absorb whichever chips prove right across a rapidly evolving landscape.

8.

Why This Capability Does Not Emerge from Generic Colocation

The execution layer described does not naturally emerge from traditional colocation models, even those operated by sophisticated, well-capitalized providers. The requirements imposed by modern enterprise inference workloads are not incremental extensions of existing data center design assumptions. They are directionally different.

Traditional colocation businesses are designed to maximize stability, predictability, and standardization. Their operating models assume long-lived infrastructure assets, relatively static power envelopes, and gradual technology transitions. This model works well for general-purpose enterprise compute where hardware refresh rates are measured in years. Inference infrastructure behaves differently. Accelerator architectures evolve rapidly, form factors change, and power density increases are nonlinear.

Extreme power density and advanced cooling illustrate this mismatch clearly. While many providers now advertise liquid cooling capability, it is often treated as a special-case accommodation rather than a baseline operating condition. Mixed air- and liquid-cooled deployments across multiple accelerator vendors introduce operational complexity that erodes standardization. As a result, most providers constrain liquid cooling to narrow configurations or bespoke deployments. Inference-driven infrastructure increasingly demands facilities where high-density, mixed-cooling environments are assumed from the outset, not negotiated as exceptions.

Low-latency metro placement introduces an additional structural constraint. Facilities located near dense enterprise networks, financial infrastructure, or major population centers face higher land costs, more complex interconnection requirements, and less flexibility in power sourcing than hyperscale campuses. Traditional colocation economics tend to favor scale and utilization efficiency over physical adjacency and latency discipline.

Colovore›s strategy is intentionally asymmetric relative to these prevailing models. Rather than optimizing for homogeneity, Colovore is designed to tolerate and absorb hardware diversity. Rather than minimizing operational variance, it is built to support rapid silicon evolution. Rather than treating extreme density and advanced cooling as exceptional capabilities, they are treated as foundational design constraints. These differences are structural, not cosmetic.

9.

Inference as an Application Component, Not a Service

For developers and early-stage companies, inference is often treated as a service: a stateless API call that accepts tokens and returns tokens, abstracted from the surrounding system. This mental model aligns well with public cloud delivery and emphasizes convenience, elasticity, and rapid iteration.

Enterprises do not operate this way once inference becomes embedded in core systems. Inference is not a peripheral service but an application component, comparable to a database engine, a transaction processor, a pricing engine, or a risk model. It sits inside a larger system with defined inputs and outputs, explicit dependencies, and measurable consequences when behavior changes or degrades.

Once inference is treated as an application component, enterprises begin asking a different class of questions. Where does inference sit in the transaction flow? What happens when it stalls or times out? How is behavior versioned and pinned over time? How are outputs audited and reproduced? How is blast radius constrained when models are updated? These are not abstract concerns. They are standard questions for any system that participates directly in business logic.

Public cloud inference services abstract away the underlying hardware, scheduling, and topology. That abstraction is valuable for exploratory workloads and non-critical inference. However, it becomes a liability when inference is embedded deeply into application logic. Multi-tenant scheduling introduces variability. Geographic distance introduces latency and jitter. Platform-managed upgrades introduce behavior changes outside the enterprise›s control.

In practice, enterprises increasingly adopt hybrid architectures. Foundation models accessed via API remain valuable for non-latency-critical tasks and general-purpose inference. At the same time, enterprises deploy private inference layers for functions that require determinism, tight integration with internal systems, or proximity to proprietary data. As inference becomes an application component rather than a service, infrastructure decisions move closer to application architecture decisions. Hardware choice, network topology, and physical placement become part of system design. This is the point at which generic, abstracted infrastructure ceases to be sufficient.

10.

The Colovore Execution Model

Colovore›s role is to provide the execution environment for the private inference layer that enterprises increasingly need. By operating neutral, high-density facilities in low-latency metro locations, Colovore allows enterprises to deploy customer-owned inference hardware that integrates tightly with business systems without requiring them to retrofit or rebuild their own data centers.

Colovore facilities are engineered for extreme power density, advanced cooling, and complex interconnection. This allows multiple inference platforms to coexist under identical physical and network conditions. As hardware evolves, vendors and enterprises can pilot, deploy, and migrate without changing facilities or topology. Connectivity is achieved through private, deterministic network paths rather than physical adjacency. From the perspective of the application, inference executes as a local component even though it resides in a specialized external facility.

This model aligns cleanly with enterprise operational boundaries. Enterprises retain ownership of hardware, models, and software, preserving control over cost structures, behavior, and lifecycle management. Colovore assumes responsibility for the physical realities of extreme power density, advanced cooling, and heterogeneous accelerator support.

Importantly, this hybrid architecture does not displace the use of foundation models or hyperscaler services. It complements them. Public cloud inference continues to serve appropriate workloads, while private inference layers handle functions that require determinism, proximity, and deep integration. Over time, as inference becomes embedded in more revenue-generating and risk-sensitive systems, the private layer tends to grow faster than the public-facing layer.

Colovore is not competing against hyperscalers, enterprises, or silicon vendors. It is enabling all of them to operate where their models break down.

11.

Competitive Outcomes and Strategic Positioning

The evolution of inference infrastructure is not zero-sum. Hyperscalers will continue to dominate elastic, abstracted workloads where convenience and scale outweigh determinism and control. General-purpose accelerators will remain central to many deployments. Specialized inference platforms will succeed where their advantages justify focused adoption.

Colovore wins by enabling all of these outcomes simultaneously. As inference hardware diversifies and enterprise ownership increases, the value of a neutral, high-density, low-latency execution layer compounds. Each new inference-driven application increases demand for execution capacity that generic infrastructure cannot support. As hardware specialization accelerates and power requirements rise, the gap between enterprise intent and enterprise facility capability widens. Colovore fills that gap by design.

For enterprises evaluating their infrastructure strategy, the question is not whether to use Colovore instead of the cloud. The question is whether the architecture they are building today can absorb the changes that are coming within the economic life of the decisions they are making. The answer, for a growing number of workloads and a growing number of enterprises, is that it cannot without a purpose-built neutral execution layer.

Final Summary

Structurally, AI inference is diverging permanently from AI training, driven by fundamentally different constraints. Inference fragments because it is embedded in business systems constrained by latency, determinism, governance, and physical topology.

This fragmentation is not a transitional phase. It is the direct consequence of inference becoming part of enterprise application architecture rather than an abstract service. The enterprises most at risk are not those that fail to adopt AI. They are those that adopt AI inference deeply and then discover that the infrastructure commitment they made in 2025 or 2026 cannot support the workloads, hardware platforms, or economics of 2029 or 2030.

Hyperscalers are not failing in this transition. They are optimizing correctly for their operating model. Their inference chip strategies excel at elastic, abstracted workloads. At the same time, those design choices make hyperscaler platforms structurally unsuitable for large classes of enterprise inference workloads that require deterministic behavior, proximity to proprietary data, and tight integration with internal systems.

Incumbent silicon platforms such as NVIDIA are not displaced by this shift. They are repositioned into environments where enterprises demand ownership, specialization, and performance fidelity.

Enterprises, however, are not willing to retrofit their existing data centers to support extreme power density, advanced cooling, and rapidly evolving accelerator form factors. This creates a durable structural gap between enterprise intent and enterprise capability.

Colovore fills that gap by design — neutral, high-density, and built specifically for the execution requirements that generic infrastructure cannot meet. By operating high-density, low-latency metro facilities designed explicitly to support heterogeneous, customer-owned inference platforms, Colovore allows enterprises and silicon vendors to deploy advanced inference systems without sacrificing control or performance.

As inference hardware continues to diversify, power density increases, cooling complexity rises, and enterprise integration deepens, the value of a neutral execution layer compounds. This is not a bet on which silicon architecture wins. It is a bet that inference becomes too important to outsource blindly, and that the physical realities of latency, determinism, and density create enduring strategic advantage for infrastructure built to absorb complexity rather than avoid it.

The right time to build flexible infrastructure is before the transition forces your hand.
That time is now.

COLOVORE

colovore.com | info@colovore.com

About Colovore

Colovore is the Bay Area's leading provider of high-density colocation solutions. Located in Santa Clara, CA, Colovore's modern data center features wall-to-wall power densities of 20 kW per rack, leading to significant reductions in TCO for our customers while also ensuring long-term IT scalability. Our facility features 9 MW of power, is LEED-Platinum certified, and operates with a PUE below 1.3. In addition to colocation space, we also feature a range of managed services to make the daily life of an IT professional better, including Ethernet connectivity, managed firewalls and VPNs, rack-and-stack, and cable management services.