HyperFRAME RESEARCH

VULTR

STRATEGIC WHITE PAPER

Enterprise AI Use Cases on the Vultr + NVIDIA Open Stack

How to Evaluate Use Cases to Achieve AI Results

Authors:

Stephen Sopko
Analyst-in-Residence
Semiconductors & Deep Tech

Ron Westfall
VP and Practice Leader for
Infrastructure and Networking

MARCH 2026

The Outcome Gap

Enterprises Have the Infrastructure. The Question Is What to Build.

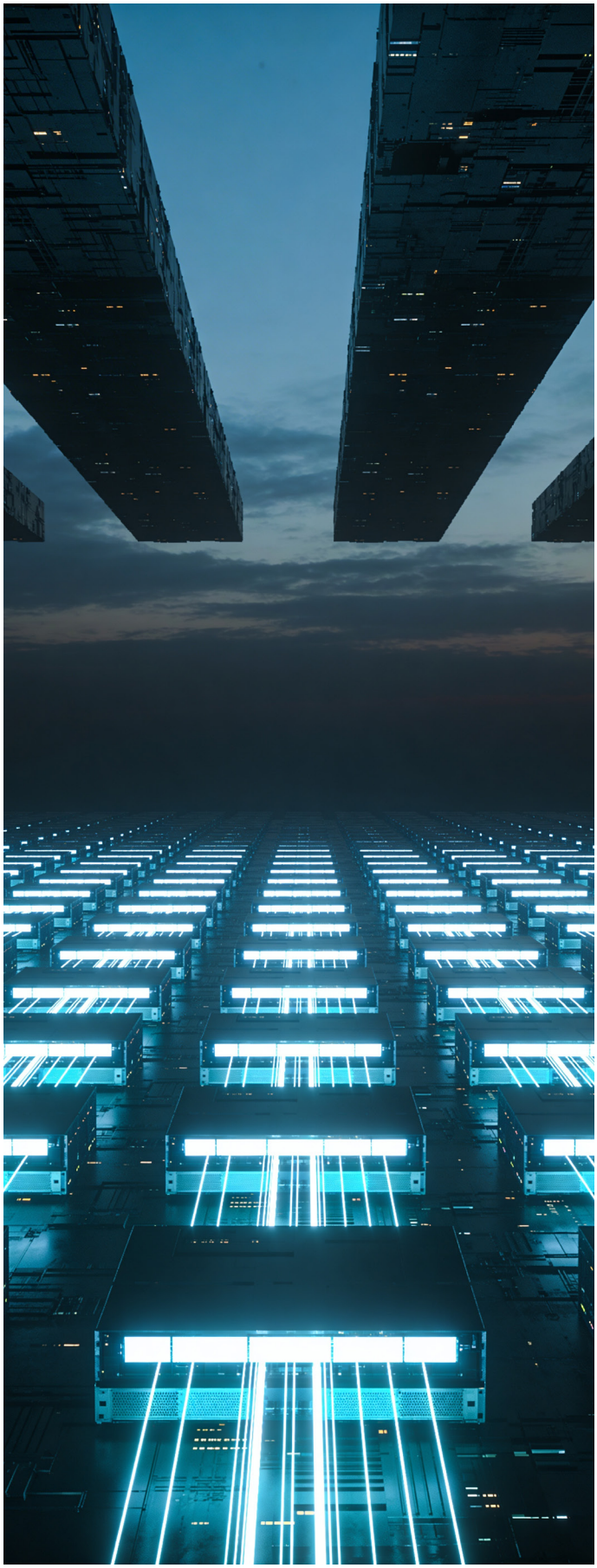
The enterprise AI conversation shifted in 2026. The preceding 3 years were defined by infrastructure acquisition: procurement cycles, GPU commitments, cloud agreements, and extensive experimentation. Most organizations with meaningful AI budgets can now access compute. Fewer organizations possess a clear answer to a far more fundamental and consequential decision "What business outcome to deliver with AI, and when?"

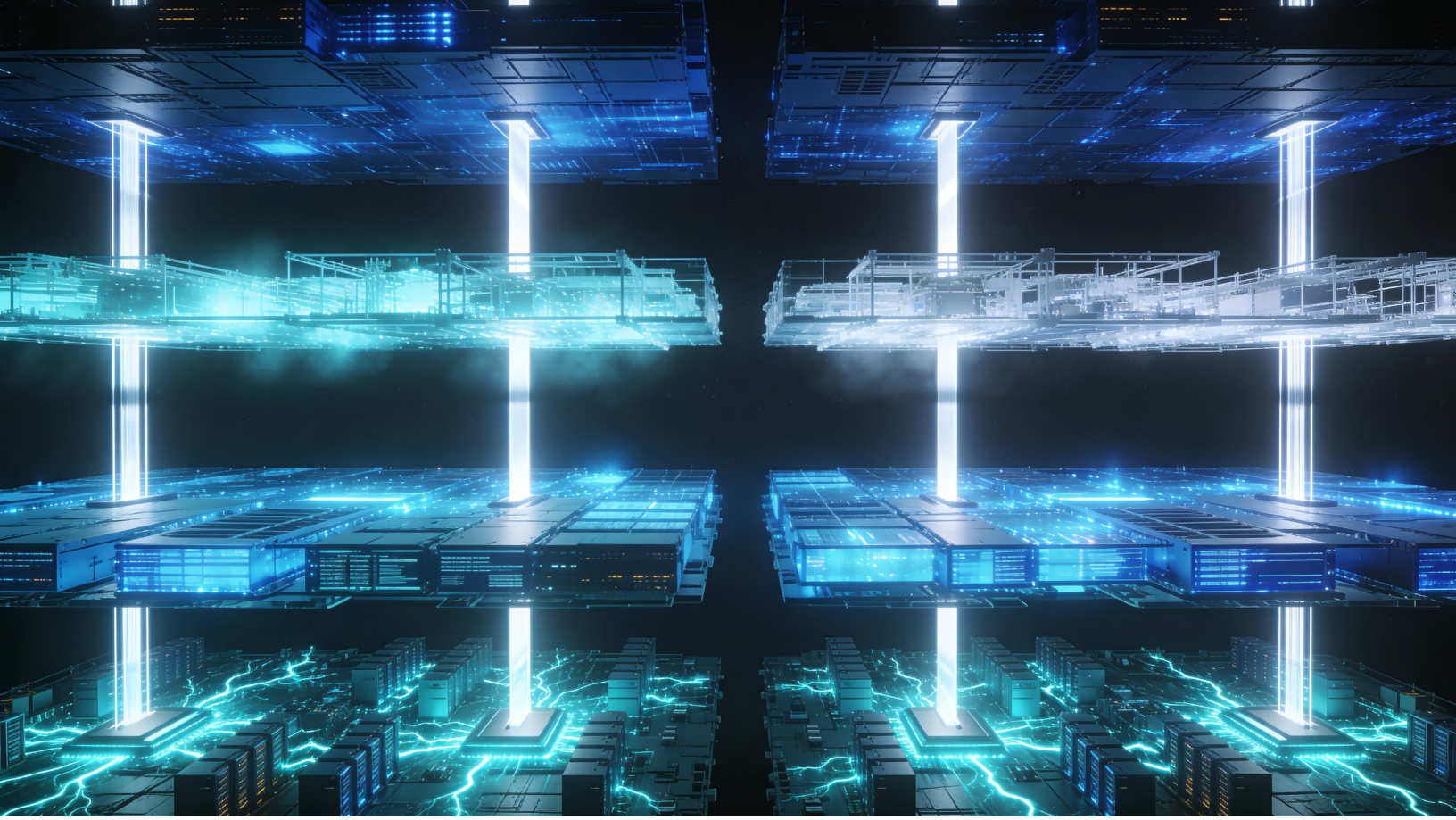
Infrastructure-first thinking dominated the opening years of the AI-age; while necessary, it is now increasingly outmoded. An overdue return to business-outcome-first thinking now changes the conversation around evaluating AI use cases and the infrastructure needed to deliver them. The relevant questions now go beyond GPU specifications or cloud provider footprints; AI-empowered organizations are increasingly addressing which AI-driven outcomes are achievable, which require next-generation hardware to realize at scale, and what the path from prototype to production actually looks by specific industry use case.

If infrastructure is no longer the critical bottleneck for most enterprises (HyperFRAME Research Lens data confirms that it's now 21% with poor data quality (27%) and cost (24%) now key), ***clear decisions about what to build on that infrastructure move back to the forefront.***

This paper addresses those decisions. It examines use cases across four industries, all designed to run on the Vultr, NVIDIA, NetApp, and DDN platform described in our first research brief, organized against a three-part framework: **grow revenue, improve operational efficiency, and reduce risk.** We assess what the case studies indicate that the NVIDIA stack delivers for each outcome, where the use cases are production-ready today, and where they represent emerging capabilities that warrant a prototype-first approach.

The goal is not to present a catalog. It is to help the IT leader and the line-of-business executive have a more productive conversation about where AI infrastructure investment converts to measurable business outcomes.





The Stack in Brief

Platform Recap for Business Decision-Makers

A full treatment of the platform architecture is available in our first research brief. For this paper, the relevant summary is as follows.

Vultr, NVIDIA, and NetApp have co-engineered a pre-integrated inference stack designed to reduce the time between infrastructure provisioning and running production AI workloads. The hardware layer centers on NVIDIA GPUs, with Vera Rubin-based instances expected from Vultr in the second half of 2026 and current Blackwell infrastructure available now. The NVIDIA software layer includes: Dynamo for inference orchestration, the Nemotron family of open-source models and weights for domain-specific customization, and NeMo for fine-tuning. This is expanded by NetApp ONTAP and DDN, providing the storage and data management layer that production inference workloads require.

Practical implications for deploying this type of use case are straightforward: you are not building a platform, you are achieving a desirable business goal. The infrastructure layer is already assembled. Engineering effort focused on the application layer, model customization, and integration with existing business systems. That is a meaningfully different starting point than what most enterprises are working from in hyperscaler environments, where the NVIDIA open-source stack is available but not the primary integration path. By surfacing the entire NVIDIA hardware+software ecosystem, leaving out any external layers, Vultr enables direct deployment and portability between NVIDIA environments.

What follows is an assessment of use cases organized around three outcome categories.

Category 1: Grow Revenue

Where AI Inference Has the Most Immediate Revenue Line

Growing revenue is comparatively easy to measure; make changes, and revenue either increases or doesn't. In 2026, this is now table stakes for AI deployment.

Gaming: Low-Ping Cloud Rendering via AI-Driven Asset Optimization

Business objective: Deliver consistent low-latency streaming experiences globally, without requiring high-end local hardware, thus expanding the addressable player base and reducing churn.

The gaming industry's shift toward cloud-rendered, cross-device streaming is not some future requirement – it has been a constantly evolving competitive pressure for years. Players expect instant access, cinematic rendering quality, and consistent multiplayer performance regardless of device. Platforms that cannot deliver consistent streaming quality lose players at acquisition and at retention.

The technical challenge is not rendering quality in isolation. It is maintaining that quality at scale across geographically distributed player bases, through variable network conditions, while managing the GPU cost implications of demand spikes during launches and live events. Those three problems compound each other in ways that static infrastructure planning cannot address.

The Vultr, NetApp, and NVIDIA architecture addresses them as a system. Workloads are rendered closer to players through globally distributed Vultr Cloud GPU and Bare Metal infrastructure - thus reducing input lag for latency-sensitive competitive and multiplayer titles. In the same infrastructure model, NetApp ONTAP with FlexCache distributes frequently accessed game assets across regions, reducing cross-region data pulls and simultaneously reducing load-time spikes and delays during scene transitions. Real-time session telemetry, streamed through Vultr Managed Kafka®, feeds NVIDIA NIM microservices running Nemotron 3 Nano models that dynamically adjust adaptive bitrate, recommend GPU scaling actions, and trigger predictive asset preloading. Finally, NVIDIA Dynamo handles the inference orchestration layer keeping optimization services stable under heavy concurrency and reducing the likelihood of bottlenecks at peak demand.

The business outcome is comparatively easy to measure: player acquisition rates improve thanks to lower local hardware barriers, retention rates stabilize as consistent performance drives them, and cost discipline at the GPU level improves through telemetry-driven scaling rather than capacity buffers.

Hospitality: Dynamic Pricing and Revenue Optimization

Business objective: Increase Revenue Per Available Room (RevPAR) by using AI-accelerated demand forecasting that allows dynamic pricing reflecting live booking signals instead of legacy static seasonal models.

Hotel revenue optimization is a precision problem more than a volume problem. While occupancy rates are increasingly stabilized across much of the market, in 2026, margin opportunity will be driven by pricing that responds to real-time demand signals and scenario modeling, anticipating shifts in booking patterns before they affect revenue windows.

Most hotel organizations have the data required to support AI-driven pricing. The problem is that it lives across booking, revenue management, loyalty, and operational systems, often across properties with different regional compliance requirements. Without a governed, unified data foundation, AI models produce forecasts that revenue managers cannot trust, so they ignore them. The technology problem is half organizational, half infrastructure.

The NetApp, NVIDIA, and Vultr stack addresses both. Hotel data is ingested and consolidated into a governed storage layer built on NetApp ONTAP within Vultr's cloud environment, with encryption, versioning, and regional compliance management as structural features rather than afterthoughts. That unified, auditable foundation feeds NVIDIA GPU-accelerated demand forecasting, cancellation prediction, and dynamic pricing optimization, deployed through Vultr Cloud GPU with NVIDIA Dynamo managing inference at operational speed.

The result is a system that allows revenue managers to act on explainable pricing recommendations during active booking windows. While simulation engines test pricing and demand scenarios pre-execution, they enable real-time inference to update recommendations as booking patterns shift. The business case sharpens thanks to a compounding effect across multi-property portfolios: small improvements in forecast accuracy, applied consistently across many properties, accumulate into material gains in RevPAR and gross operating profit per available room.

Category 2: Improve Efficiency

Operational AI That Reduces Cost and Throughput Friction

Business outcomes in the Improve Efficiency area are more difficult to measure because the underlying baseline is often poorly understood. In this case, the observable outcome is a trend moving in the right direction, rather than additional revenue dollars.

Hospitality: Labor Optimization and Operational Cost Management

Business objective: Reduce labor cost per occupied room by aligning staffing models to AI-generated demand forecasts rather than historical patterns.

The revenue optimization framing above addresses one dimension of hotel profitability. The operational **efficiency** dimension is equally material and runs on the same infrastructure, which is the structural argument for treating the hospitality use case across both sections rather than forcing it into one.

Rising labor costs are among the most significant margin pressures in hotel operations today. The American Hotel and Lodging Association has documented consistent year-over-year increases in labor cost per occupied room across most hotel categories. The traditional response, staffing to historical patterns with manual adjustments during peak periods, leaves significant efficiency on the table because it cannot respond quickly enough to the booking pattern volatility that modern demand environments produce.

The same AI-accelerated forecasting infrastructure that drives dynamic pricing in the revenue optimization use case also powers labor modeling. Demand forecasts update in near real time as booking signals shift. Labor scheduling recommendations are generated and updated continuously rather than weekly. Operations leadership can run staffing scenarios before committing to scheduling decisions. The difference from conventional revenue management tools is not just speed; it is that the recommendations are connected to the same governed data layer and the same real-time inference engine that produces pricing decisions, creating consistency across revenue and operations functions that siloed systems cannot achieve.

For hotel groups operating across multiple independently owned franchise properties, the ability to deploy this capability

consistently without requiring system redesign at each site is the deployment-level differentiator. Vultr Kubernetes Engine handles workload portability across properties; NetApp's data governance framework manages the compliance requirements that vary by region.

E-Commerce Fulfillment: Autonomous Shop-to-Ship Robotics

Business objective: Reduce fulfillment time and error rates by connecting digital order triggers directly to autonomous warehouse execution, eliminating manual handoff steps.

The CarphaCom platform, developed during the AI Meets Robotics challenge in collaboration with Vultr and LabLab, demonstrates a use case worth tracking for retail and manufacturing supply chain leaders. The core architecture connects digital commerce directly to robotic warehouse execution: a B2B order triggers autonomous picking, packing, and dispatch workflows without manual handoff between systems.

The prototype's performance metrics are informative; order-to-robot trigger latency under 300 milliseconds, system state synchronization under 500 milliseconds, simulated pick accuracy of approximately 99 percent, and end-to-end fulfillment time is 35-60% faster than manual testing workflows. The architecture is built on Vultr Cloud GPU infrastructure, with NVIDIA Isaac Sim powering the digital twin simulation environment and Gemini models handling command processing and operational decision support.

An honest maturity assessment is warranted. CarphaCom was built in an 8-day hackathon sprint, and, according to the development team's characterization, is approximately 75 percent production-ready at the commerce layer. Enterprise deployment of this case will require additional hardening and scalability testing. We include it because, for IT leaders evaluating warehouse automation, the relevant observation is not the current implementation maturity. The relevance is the demonstrated architectural pattern: digital twin simulation for validation, event-driven microservices for integration, and GPU-hosted AI for decision orchestration. Organizations with existing investments in warehouse automation should evaluate how this pattern has the potential to extend their current capabilities.

Category 3: Reduce Risk

Governance and Reliability Systems That Protect Operations and Assets

Risk is the toughest area to measure, and often ends up in a philosophical rather than an accounting-based discussion. The question of “if a risk is avoided, how do we value it versus true cost impacts?” The reality is that risks created by AI use in autonomous scenarios must be addressed, regardless of assigned ROI. That said, even here, hard performance and cost benefits can be tracked.

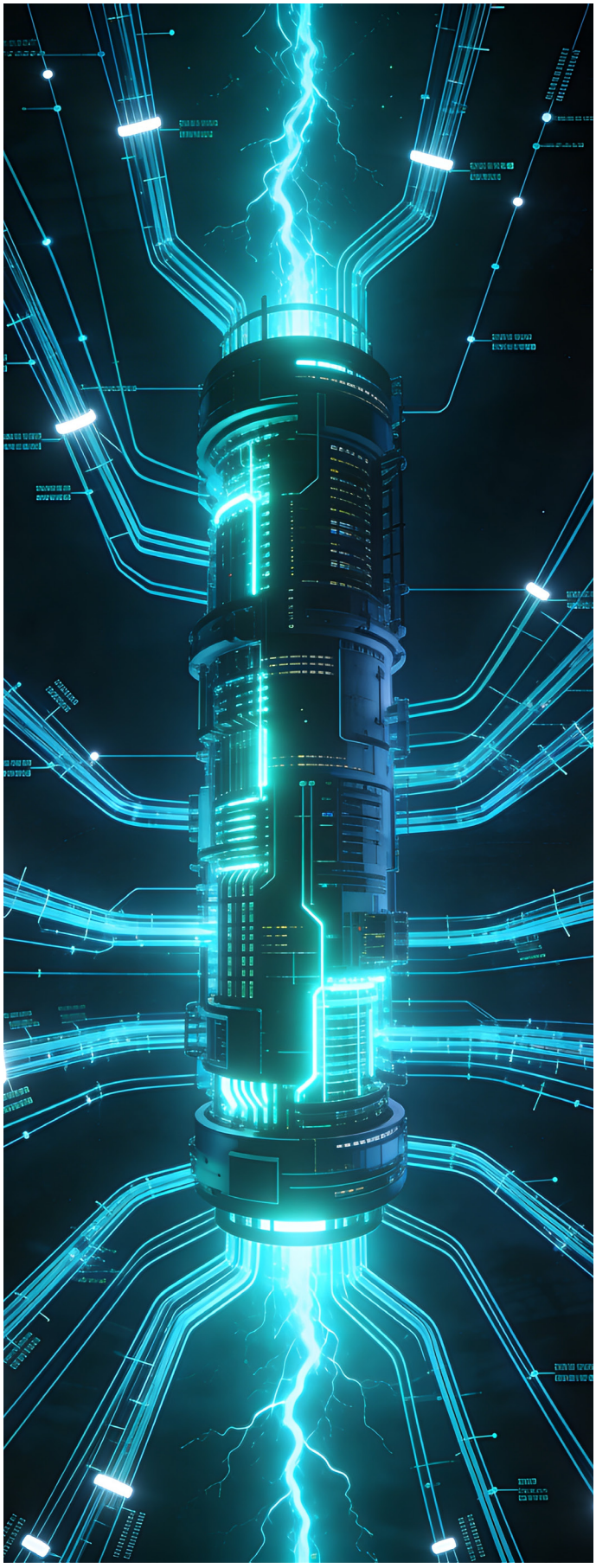
Synthetic Data for Computer Vision: Reducing Model Risk in High-Stakes Deployments

Business objective: Eliminate the data quality and coverage gaps that cause computer vision models to fail in production by replacing fragile, manually collected datasets with physics-accurate synthetic training data.

Synetic.ai represents one of the more analytically interesting use cases in this paper because it addresses a risk that most enterprises have not yet quantified: the cost of deploying computer vision models trained on incomplete or unrepresentative real-world data. In industries where those models control physical systems, including manufacturing quality inspection, agricultural monitoring, defense applications, and industrial robotics, model failure is not a software bug. It is an operational incident.

The Synetic.ai platform generates physics-accurate synthetic data and adaptive simulation environments, replacing manual data collection, labeling, and retraining pipelines. By building high-fidelity simulated environments that capture real-world physics, lighting, and materials, the platform produces photorealistic, perfectly annotated datasets for computer vision training. The results have been externally validated: in collaboration with researchers from the University of South Carolina, Synetic.ai’s synthetic datasets demonstrated a 34 percent improvement in model generalization compared to real-world datasets.

The infrastructure requirements are substantial. Each simulation involves physics calculations at a scale that demands consistent, high-throughput GPU access. Synetic.ai runs NVIDIA HGX B200 Bare Metal instances on Vultr, leveraging the platform’s global footprint to enable distributed simulation pipelines and multicloud redundancy. The compliance posture,



HIPAA, GDPR, and SOC coverage, addresses the certification requirements that healthcare, defense, and manufacturing customers impose on their infrastructure vendors.

> *The risk profile of deploying a computer vision model trained on insufficient real-world data versus one trained on validated synthetic data is not a marginal difference in regulated and high-consequence industries. It is the difference between a system that performs reliably in production and one that fails in the conditions it was never exposed to during training.*

The business impact is documented. **Vultr's infrastructure has enabled Synetic.ai to reduce rendering costs by over 40 percent, triple simulation capacity in six months, and support new enterprise contracts across defense, industrial robotics, and agriculture, contributing to more than $600,000 in ARR.** R&D iteration cycles that previously took weeks are now measured in days. Those compression rates translate directly into faster time-to-market for customers deploying vision-driven systems.

For enterprise IT leaders in manufacturing, agriculture, healthcare, or defense evaluating AI deployment risk, the practical question this use case raises is: how was your computer vision model trained, and does the training data reflect the conditions the model will encounter in production? For organizations that cannot answer that question with confidence, synthetic data platforms running on GPU infrastructure with the reliability characteristics Vultr provides represent a risk mitigation lever that the model selection decision alone cannot.

Sovereign Robotics: AI Governance Layer for Autonomous Industrial Systems

Business objective: Enable safe AI-driven robot autonomy in manufacturing and energy environments by intercepting and evaluating AI-generated actions before execution, maintaining auditability without eliminating autonomy.

The autonomous robotics adoption curve in manufacturing and energy is encountering a governance problem that traditional safety-only approaches do not address. Regulatory and operational pressure to ensure safe execution, enforce policy controls, and maintain auditability is intensifying. The typical response is either to constrain AI autonomy to the point where the economic case weakens or to accept compliance risk and manage it reactively. Neither is stable, as autonomous systems take on more consequential roles.

Sovereign Robotics Ops charts a third path: a real-time governance layer positioned between the AI planning engine and robot actuators. Every AI-generated action passes through a governance API that evaluates human proximity, speed thresholds, and geofencing rules before execution. The system then decides amongst 4 branches: execute the action as planned, modify it to meet safety thresholds, halt it, or trigger replanning. This decision-making function generates a SHA-256 hash chain, thus producing tamper-proof audit trails for compliance reporting. Policy evaluation latency is under 100 milliseconds, maintaining operational throughput without compromising the safety gate.

The same maturity caveat applies here as with CarphaCom. Sovereign Robotics Ops emerged from a hackathon context in collaboration with Vultr and LabLab. The architecture is sound, the performance benchmarks are credible, and the sim-to-real consistency, applying the same policy layer in Gazebo and Isaac Sim environments as in production, de-risks the validation pathway. But this is a reference implementation rather than a production-hardened enterprise product. Organizations evaluating autonomous robotics governance should treat it as a design pattern to inform procurement and custom development decisions.

The governance gap in industrial AI autonomy is real and underserved by current enterprise software vendors. The architectural direction this implementation demonstrates, a platform-level governance layer running on cloud infrastructure rather than embedded in individual robot control systems, is the right approach for organizations that need both autonomy and auditability.

The Pattern: What These Use Cases Share

Why the Same Platform Serves Multiple Outcomes

Across the use cases examined here, a consistent architectural profile emerges regardless of industry or outcome category. All of them are inference-heavy at the decision layer. All of them benefit from models that can be fine-tuned on domain-specific data rather than generic foundation models. All of them require enterprise-grade storage and data-pipeline infrastructure at the scale required by production inference. And all of them must deploy quickly, either because of business-case-driven time-to-value or because competitive pressure is compressing development cycles.

That architectural profile is the genesis of the Vultr, NVIDIA, and NetApp stack. Dynamo handles the inference orchestration complexity that would otherwise require custom engineering. Nemotron model weights provide the domain-adaptable foundation that industry-specific use cases require. NetApp ONTAP provides the governed, auditable data layer that regulated industries and franchise-distributed operations need. Vultr's pre-integrated environment removes the infrastructure assembly work that typically consumes engineering capacity before application development can begin.

The composable approach the stack enables is worth emphasizing. None of the use cases described here require a wholesale infrastructure transformation. The appropriate starting point is a single use case with a clear business outcome, deployed as a defined workload on the pre-integrated stack, measured against the specific metric the business cares about, and expanded to adjacent use cases once the architecture is understood in production.

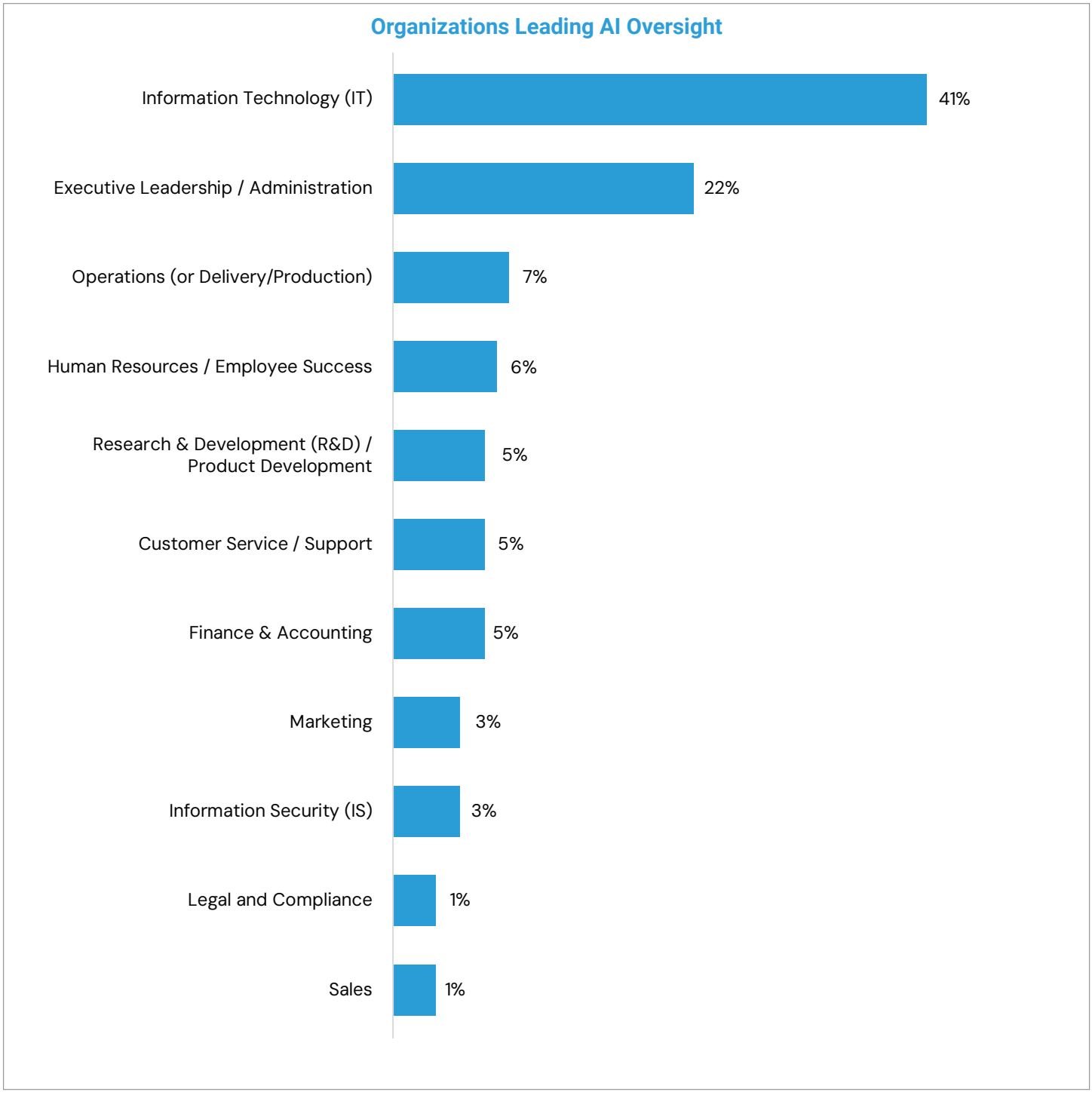
The gaming platform scales to new regions using the same storage and inference infrastructure that powers the first deployment. The hotel revenue management system expands to labor optimization on the same data layer. The synthetic data platform at Synetic.ai tripled simulation capacity on the same Vultr Bare Metal infrastructure it started on. These are not architectural coincidences. They reflect the stack's composable design intent.

None of these use cases are multi-cloud replacement strategies. They layer alongside existing cloud investments, addressing the inference-heavy, open-stack workloads where performance, cost transparency, and model portability matter, while other workloads remain in their existing environments.

How to Shift From Tech to Results-Focused Thinking

A Practical Sequence for Moving from Use Case to Deployment

Pick one use case with a clear business owner on the line-of-business side. Use cases that are more typically subject to failure occur when IT owns the project, and the business owns the outcome, with no shared accountability. Identify the VP of Operations, Chief Revenue Officer, or equivalent executive who will define what success looks like before committing to infrastructure. Current HyperFRAME Research Lens data where IT departments are leading AI oversight in 41% of organizations, executive leadership in 22%, and operations in only 7%. This suggests that a disconnect between business outcomes and technical execution may be organizationally driven.



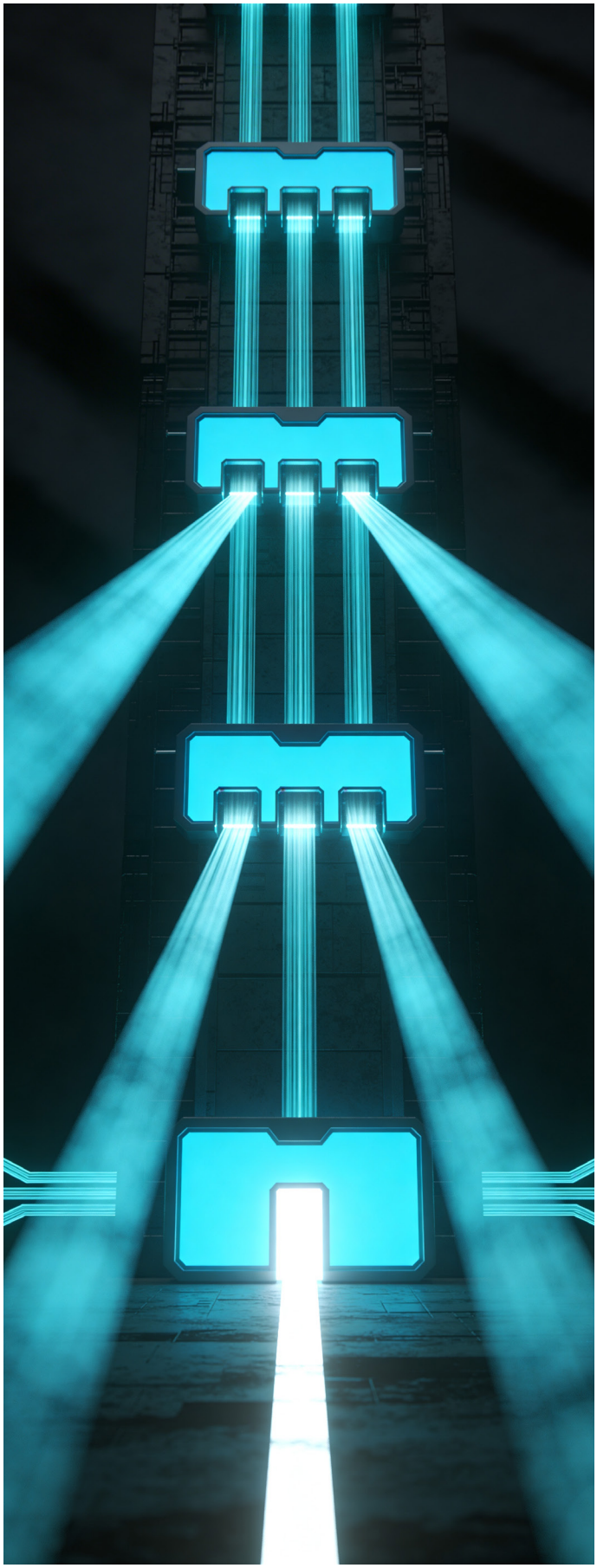
HyperFRAME Research Lens, AI Stack, Oversight of AI, Q1 2026

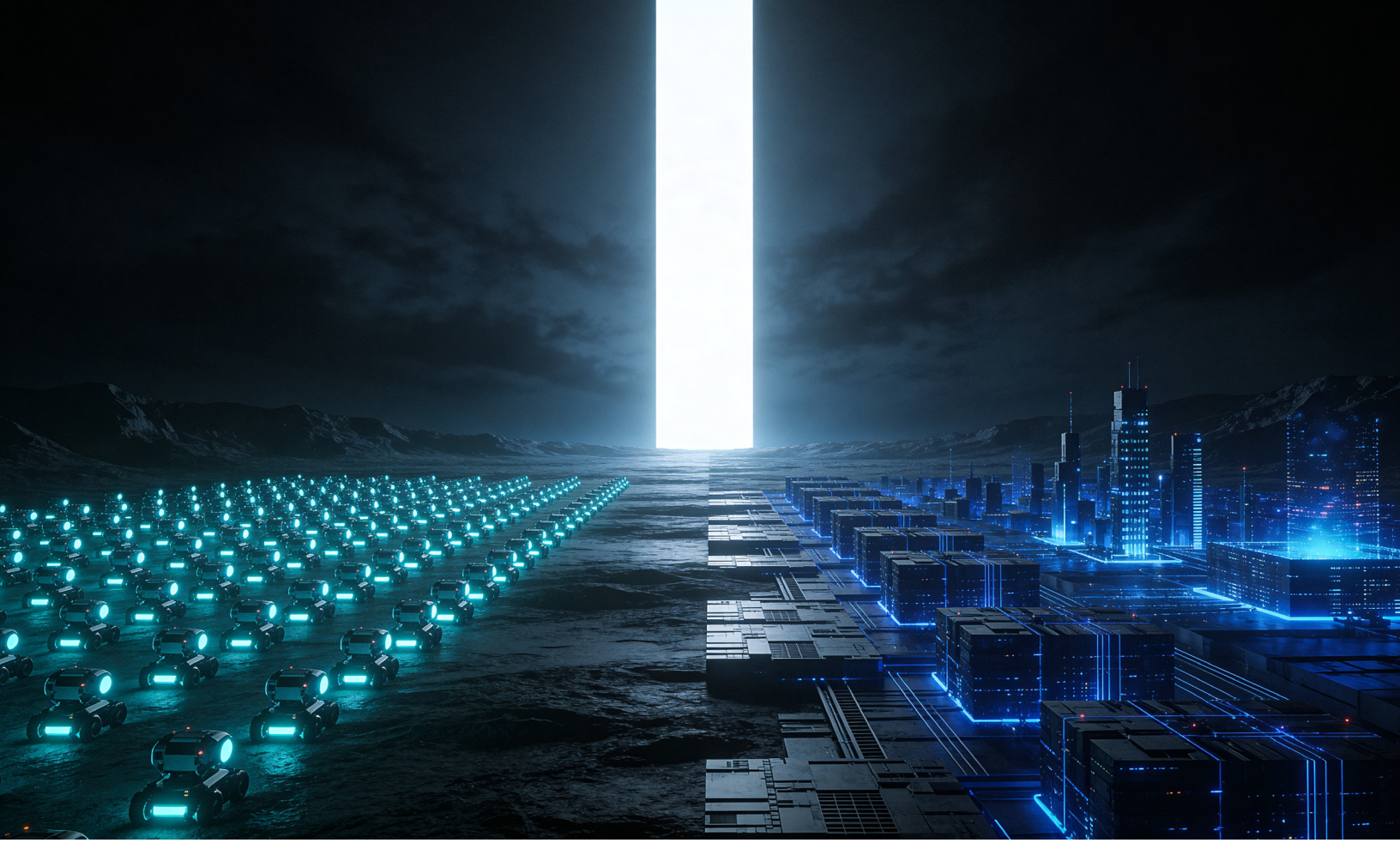
Evaluate production readiness against your timeline. The Gaming, Hospitality, and Synetic.ai use cases have production-grade reference architectures and documented outcomes. The robotics and e-commerce fulfillment cases are at an earlier stage, with meaningful engineering work remaining before enterprise deployment. Match your organization's risk tolerance and delivery timeline to the maturity of the use case, not to the ambition of the outcome.

Use the pre-integrated environment to answer the application question first. The value of a pre-integrated inference stack is that the first question you answer is whether the use case works for your specific data and business logic, not whether the infrastructure is configured correctly. That sequencing matters. Infrastructure commitment decisions should follow successful application validation, not precede it.

Engage GSI partners early, particularly for the physical integration layer. WWT and comparable systems integrators bring the professional services, compliance frameworks, and enterprise procurement pathways that move a successful prototype into a deployed program. For the robotics and fulfillment use cases specifically, the physical integration work sits outside the cloud infrastructure stack and requires a delivery partner with domain expertise in that layer.

Plan the Vera Rubin transition now, even if you are starting on Blackwell. Use cases prototyped on current-generation hardware can migrate to Vera Rubin on the same Vultr infrastructure with minimal architectural rework, thanks to the platform's extensible design. While Rubin's advancements, such as the new NVIDIA Inference Context Memory Storage (ICMS) Platform powered by BlueField-4, introduce optimized storage tiers for scaling agentic AI, partners like NVIDIA, DDN, and NetApp are committed to assisting with any necessary updates to ensure smooth integration. Organizations that build operational experience on today's stack are not locked into current-generation hardware. They are acquiring the deployment knowledge that should enable them to scale effectively when the next-generation platform arrives in 2H 2026.





Looking Ahead

Two dynamics from these use cases are worth monitoring through the remainder of 2026.

The first is the maturation rate of AI-driven autonomous operations. The Sovereign Robotics and CarphaCom use cases represent early demonstrations of a pattern, centralized cloud AI orchestrating physical systems in real time, that will become significant at enterprise scale within the next two to three years. The governance and compliance frameworks that organizations build now for these systems will determine how quickly they can move from reference architecture to production deployment when the technology matures. Infrastructure portability and auditability are not afterthoughts in that transition. They are prerequisites.

The second is whether synthetic data platforms like Synetic.ai's represent a generalizable risk mitigation approach across industries, or whether the use case remains specific to the computer vision and simulation domains where it has been validated. My analysis suggests the pattern is broader than its current applications, particularly for AI systems deployed in environments where training data coverage is structurally limited by the cost or danger of collecting real-world data. The organizations evaluating this now are ahead of a problem that their competitors have yet to name.

> *The competitive advantage in enterprise AI is shifting from who has the most GPUs to who has the most production inference running against the right data. Both halves of that sentence matter equally.*

The platform is assembled. The use cases are defined and differentiated by production maturity. The sequencing decision, which outcome to pursue first, which use case to prototype, which business executive to put in the room, is the one that will separate the organizations that convert infrastructure investment into business outcomes from those still running experiments in 2027.

ABOUT HYPERFRAME RESEARCH:

HyperFRAME Research delivers in-depth research and insights across the global technology landscape, spanning everything from hyperscale public cloud to the mainframe and everything in between. We offer strategic advisory services, custom research reports, tailored consulting engagements, digital events, go to market planning, message testing, and lead generation programs.

Our industry analysts specialize in rigorous qualitative and quantitative assessments of technology solutions, business challenges, market forces, and end user demands across industry sectors. HyperFRAME Research collaborates closely with your Analyst Relations, Product, and Marketing teams to build and amplify your thought leadership, positioning your expertise to enhance brand and product recognition. Through content that engages readers, viewers, and listeners alike, we ensure your voice resonates across channels.

CONTACT HYPERFRAME RESEARCH:

Steven Dickens
CEO & Principal Analyst | HyperFRAME Research

Email Address:

steven.dickens@hyperframeresearch.com

Telephone Number:

+1 845 505 1678

X: @StevenDickens3

LinkedIn: Steven Dickens

BlueSky: Steven Dickens

CONTRIBUTORS

Stephen Sopko
Analyst-in-Residence
Semiconductors & Deep Tech

Ron Westfall
VP and Practice Leader for
Infrastructure and Networking

INQUIRIES

Contact us if you would like to discuss this report and HyperFRAME Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "HyperFRAME Research." Non-press and non-analysts must receive prior written permission by HyperFRAME Research for any citations.

LICENSING

DISCLOSURES

HyperFRAME Research provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.


HyperFRAME
RESEARCH