Futurum®

Lenovo

AI Inference: Enterprise Infrastructure and Strategic Imperatives

JANUARY 2026

1. Executive Summary

Artificial intelligence has entered its production phase. While the development and training of foundational models captures media attention, the real economic value is generated through inferencing - the deployment of trained models to make predictions from trained models and drive business decisions. Enterprises across every major industry are now planning or executing substantial infrastructure investments to support inference workloads at scale, making the technical and architectural choices around inferencing infrastructure decisions of immediate strategic consequence.

Futurum estimates the global AI inference infrastructure market growing from $5.0 billion in 2024 to $48.8 billion by 2030, representing a six-year CAGR of 46.3%. This Base Case projection reflects balanced adoption and pricing normalization across enterprises. The Bull Case - assuming hyper-adoption, rapid cost/performance improvements and minimal regulatory friction - projects the market reaching approximately $137 billion by 2030 (2.8x Base). The Bear Case - reflecting slower enterprise uptake and margin compression from model commoditization - projects approximately $26 billion (0.53x Base).[1]

Critically, the deployment model mix is shifting dramatically: hybrid and edge inference deployments are growing at 65% CAGR compared to 46% for public cloud, signaling a structural transition toward distributed inference architectures.

However, deploying inference at scale presents substantial challenges. The computational patterns, memory access requirements, and power density constraints of AI inference differ materially from training workloads. Success requires specialized infrastructure and deep technical expertise applied strategically across cloud, edge, and on-premises deployment models.

This paper examines why AI inferencing has become a critical strategic priority, the technical and operational bottlenecks enterprises encounter, and the framework through which organizations should evaluate technology choices. The central thesis is straightforward: the right infrastructure and operational choices for inferencing will determine competitive advantage; the wrong choices will constrain growth, inflate costs, and render AI investments unproductive.

1 2H 2025 AI Platforms Market Sizing & Five-Year Forecast, Futurum Research, December 2025.

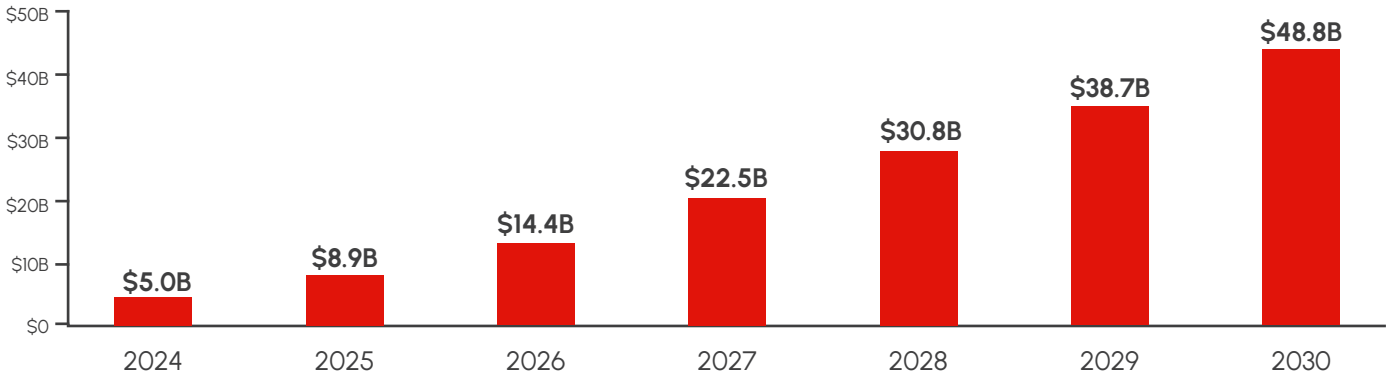
2. AI Inferencing Market Outlook

The AI inference market is experiencing accelerated growth driven by three convergent forces: the maturation of generative AI, the emergence of agentic AI systems requiring real-time decision-making and the proliferation of edge and physical AI deployments.

Market Scale and Trajectory

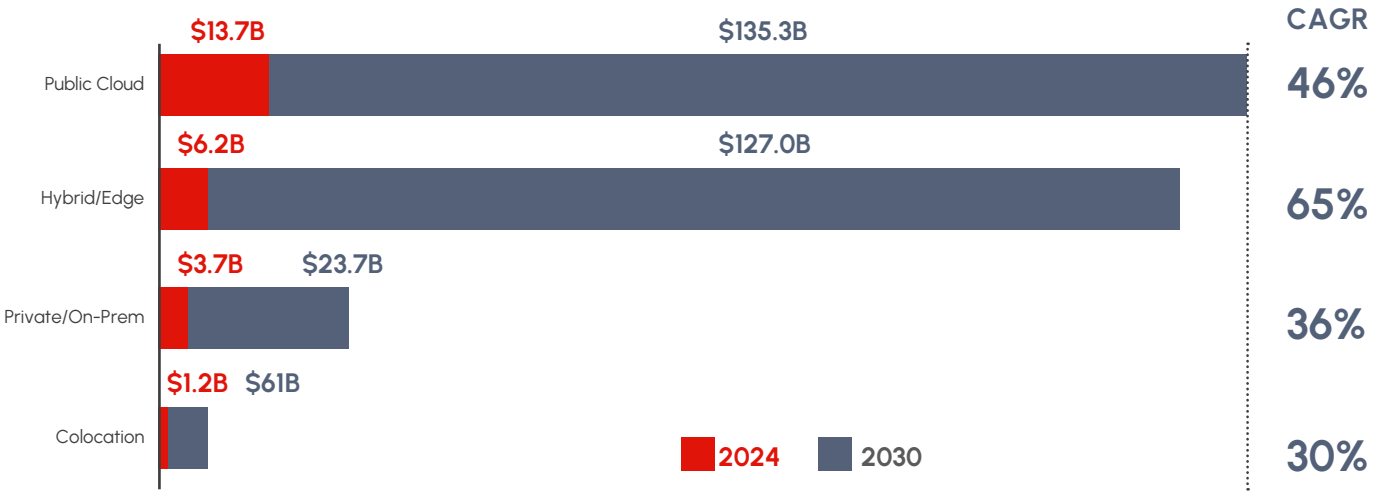
The growth trajectory of AI inference infrastructure reflects both the proliferation of generative and physical AI applications and a fundamental reorientation of enterprise compute spending toward production inference rather than model development. Futurum's base forecast projects the following market evolution:

Figure 1: AI Inference Market Size 2024-2030



Source: 2H 2025 AI Platforms Market Sizing & Five-Year Forecast, Futurum Research, December 2025.

Regionally, North America currently leads adoption with approximately 48% of global inference market share in 2024, driven by early enterprise maturity and substantial hyperscaler infrastructure investments. However, Asia-Pacific is experiencing the fastest regional growth at approximately 58% CAGR, propelled by government-backed sovereign AI initiatives, semiconductor ecosystem investments, and accelerating edge and physical AI deployments across manufacturing and consumer electronics. Europe, Middle East & Africa represents approximately 30% of market share, with growth accelerating from 2026 as EU AI Act compliance requirements drive infrastructure investment.[2]



Source: 2H 2025 AI Platforms Market Sizing & Five-Year Forecast, Futurum Research, December 2025.

2 2H 2025 AI Platforms Market Sizing & Five-Year Forecast, Futurum Research, December 2025.

The Hybrid Shift

The most strategically significant trend is the rise of hybrid and edge deployment. By 2030, hybrid and edge inference ($127.0B combined) will nearly equal public cloud inference ($135.3B). This structural shift is driven by latency requirements, data sovereignty mandates, and the proliferation of small language models (SLMs) enabling localized inference. Enterprises can no longer assume a cloud-first default for all inference workloads; instead, workload-optimized placement is becoming the standard with architectural and operational choices.

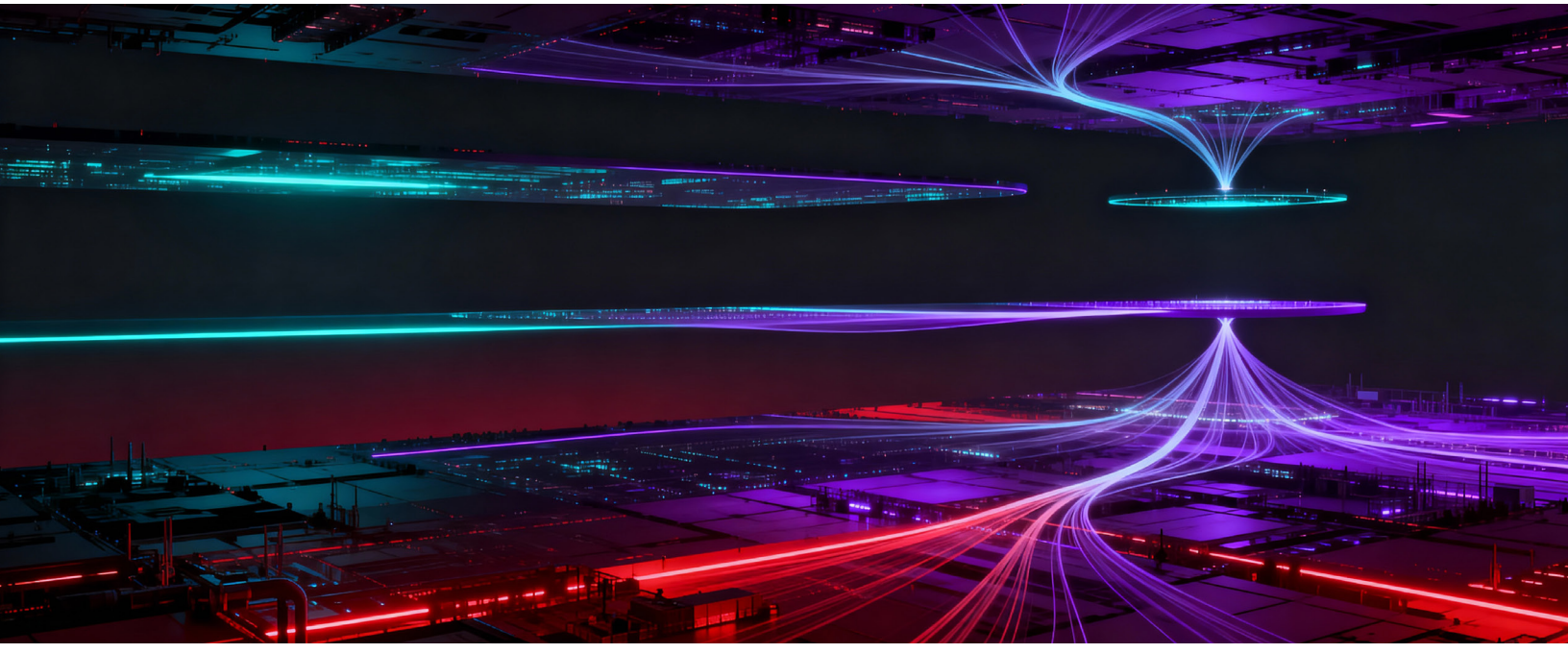
Key Adoption Drivers

Generative AI & SLMs: While LLMs drive initial volume, SLMs are democratizing inference, allowing powerful reasoning to run on cost-effective edge hardware and workstations.

Agentic AI: Unlike static chatbots, agentic systems shift inference from request-based interactions to persistent, multi-step reasoning. This continuous chain-of-thought processing drives sustained GPU utilization, increased memory pressure, and heightened latency sensitivity, challenging standard cloud architectures.

Edge & IoT: Distributed intelligence is pushing inference to the data source - whether autonomous vehicles, factory controllers, robotics, smart devices, or employee PCs - to minimize bandwidth costs and latency.

Regulatory Compliance: Frameworks such as the EU AI Act and GDPR are forcing data residency decisions that favor on-premises and local inference over global cloud processing.



3. Definition and Scope of AI Inferencing

AI inferencing is the production deployment of trained models to generate predictions based on new data. While model training is an episodic, centralized, and batch-oriented process, inference is continuous, real-time, and distributed.

AI Training vs. Inference

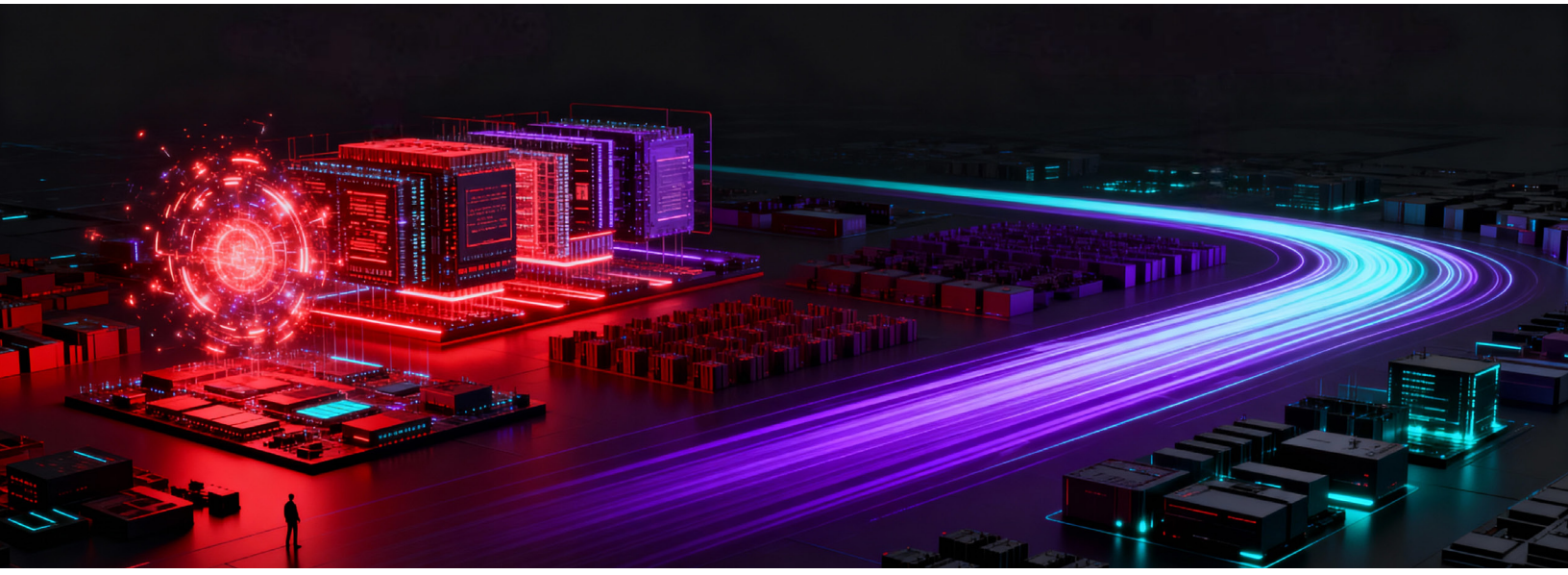
The technical requirements for these two phases differ significantly:

- **Optimization Goals:** Training optimizes for massive throughput over long periods, whereas inference prioritizes milliseconds of latency per user request.
- **Infrastructure Suitability:** Standard training rigs often make poor inference engines because they lack the specific agility and memory bandwidth required for efficient token generation.
- **Performance Benchmarks:** In production, **Time to First Token (TTFT)**—the elapsed time from a user prompt submission to the first token appearing in the response—has emerged as a critical benchmark.

Deployment Environments

The choice of environment directly impacts the speed and efficiency of the inference cycle:

- **Cloud:** Best for variable or bursty workloads where the need for elasticity outweighs higher marginal costs.
- **Edge:** Essential for applications requiring sub-100ms latency, such as industrial automation or retail analytics, and for operations that must function without constant connectivity.
- **Data Center/On-Premises:** Provides the lowest cost-per-inference for stable, high-volume workloads while maintaining strict control over data sovereignty.
- **AI Workstations/PCs:** A critical tool for developer productivity and privacy, allowing the local execution of AI apps and quantized models without data leaving the device.



4. Enterprise Adoption Challenges

Despite substantial interest and rapid market growth, enterprises encounter significant obstacles in deploying AI inference at scale. These challenges span business, technical, and operational dimensions.

Business and Operational Challenges

Cost management represents the most immediate concern for enterprise decision-makers. Cost management manifests differently in inferencing than traditional IT. Cloud-based inference costs scale linearly with token generation. As successful pilots transition to production, monthly inference bills can grow rapidly, a phenomenon organizations often call 'bill shock'. This creates a fundamental problem: many enterprises lack metrics to understand whether inference is profitable.

Scalability from pilot to production compounds cost challenges. Pilots typically run on modest hardware with minimal operational discipline. Scaling to thousands of concurrent users requires operational infrastructure - monitoring, incident response, capacity planning - that many organizations have not built. Optimization strategies that work at small scale often fail at large scale. A partitioning strategy effective for single-digit GPU concurrency becomes a bottleneck at hundreds of concurrent users.

ROI expectations create additional pressure. Many enterprises expect to realize returns within 12 months, a timeline unrealistic for generative AI initiatives requiring longer development and adoption cycles. This pressure often results in infrastructure investments misaligned with genuine strategic priorities.

Talent gaps pose a fourth constraint. Deploying inference at scale requires cross-functional expertise in data science, ML engineering, systems engineering, and infrastructure operations. In addition to scarce skills in quantization, distributed inference, and MLOps, many organizations lack expertise in planning for and managing the token economics of inferencing—an increasingly critical capability for demonstrating returns on infrastructure investments and sustained business value. While organizations retain ultimate authority over technology decisions, these skills gaps exacerbate existing Hybrid IT challenges, often requiring internal upskilling and/or engagement with external partners to effectively operationalize inference at scale.

Technical Bottlenecks

Memory bandwidth saturation is the primary technical bottleneck in LLM inference. Large models must load billions of parameters efficiently, but the bandwidth between GPU compute cores and memory often becomes the limiting factor in inference throughput. When the rate at which data moves from memory to compute cores is insufficient, GPUs sit idle - resulting in poor utilization despite expensive infrastructure investment. This phenomenon directly degrades Tokens per Second (TPS) performance and inflates Tokens per Watt, meaning infrastructure consumes power without generating value.

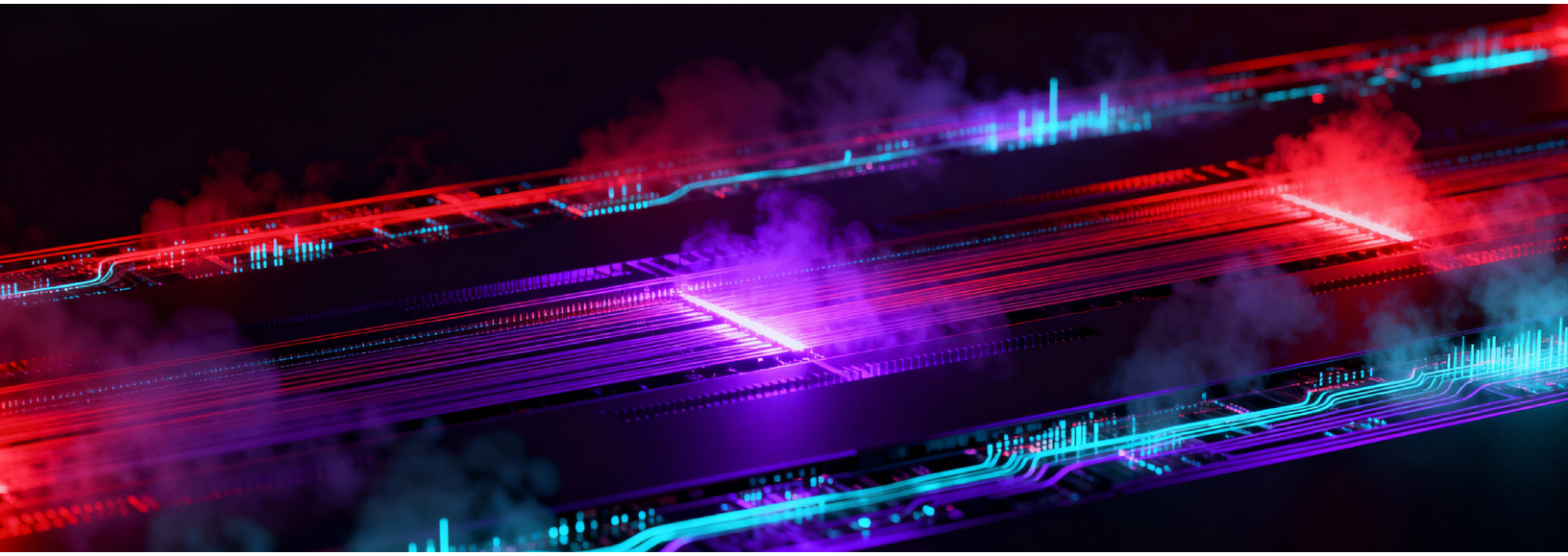
This bottleneck is particularly acute in inference because computational efficiency is inherently lower than in training. Training amortizes memory accesses across large batches; inference processes data with lower compute intensity, making memory bandwidth the limiting factor. Very large models exceeding single-accelerator memory capacity require sophisticated distributed strategies such as tensor parallelism (dividing layers across GPUs) or pipeline parallelism (dividing models sequentially across GPUs). These introduce complexity and coordination overhead that can negate performance benefits if not managed well.

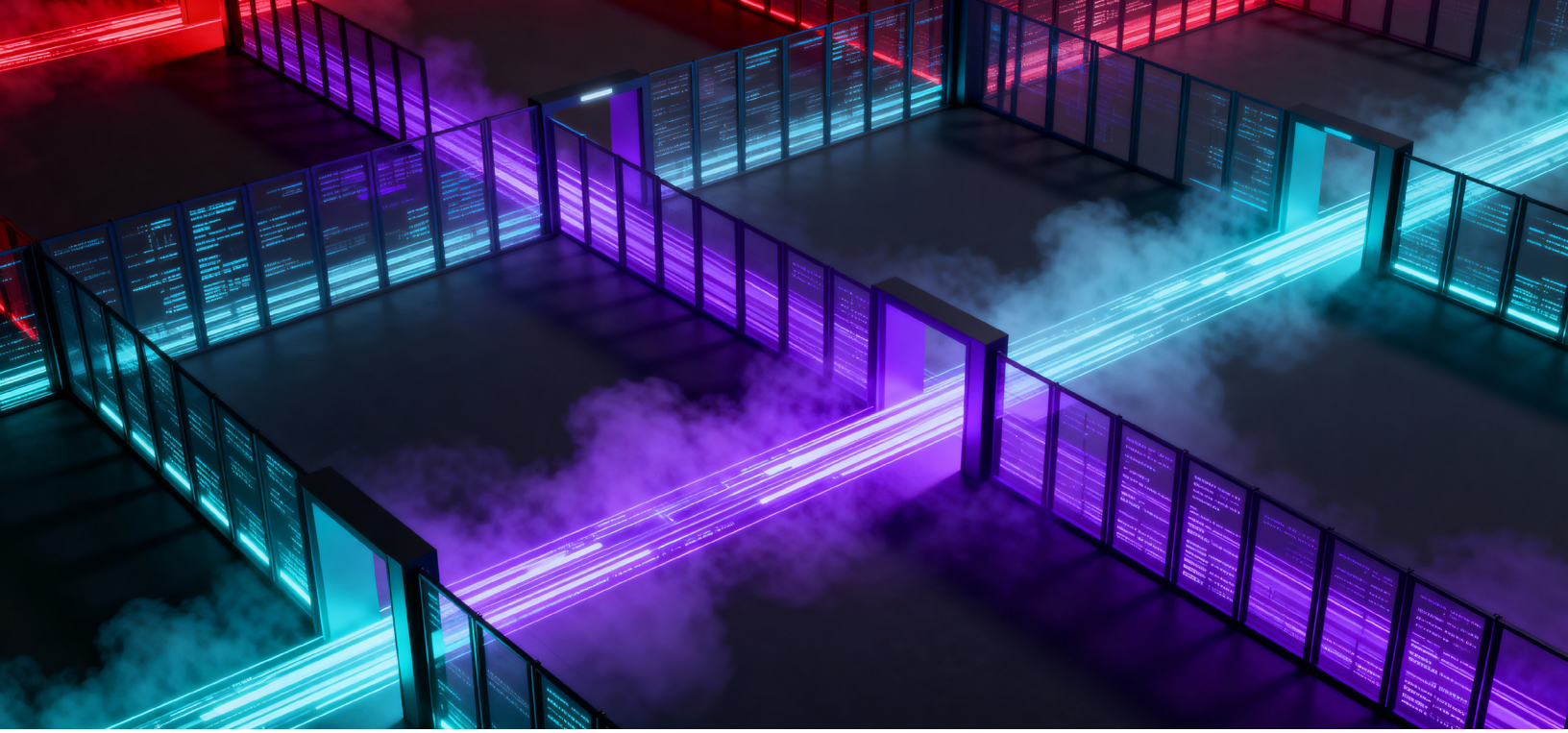
Latency requirements for real-time applications create design constraints. Many applications demand sub-100-millisecond response times, measured as Time to First Token (TTF) - the elapsed time from user prompt submission to the first token appearing in the response. Cloud infrastructure with network round-trips and queuing struggles to meet TTF requirements consistently. Meeting latency constraints reliably often requires on-premises or edge deployment rather than shared cloud resources.

Power density and cooling represent practical constraints often underestimated during planning. For example, a single NVIDIA H200 GPU draws ~700 watts; clusters with dozens of these devices generate thermal challenges standard data center infrastructure cannot handle. Many organizations discover too late that facility upgrades are necessary, extending timelines and budgets substantially.

I/O throughput and storage bandwidth can limit performance in applications requiring continuous access to feature stores. Models may need to access millions of features per second, requiring sophisticated caching and workload placement strategies.

Interoperability challenges arise in multi-vendor, heterogeneous environments, which describes virtually every large enterprise. Managing inference across multiple hardware vendors, integrating with legacy systems, and coordinating across cloud and on-premises infrastructure requires careful architectural planning.





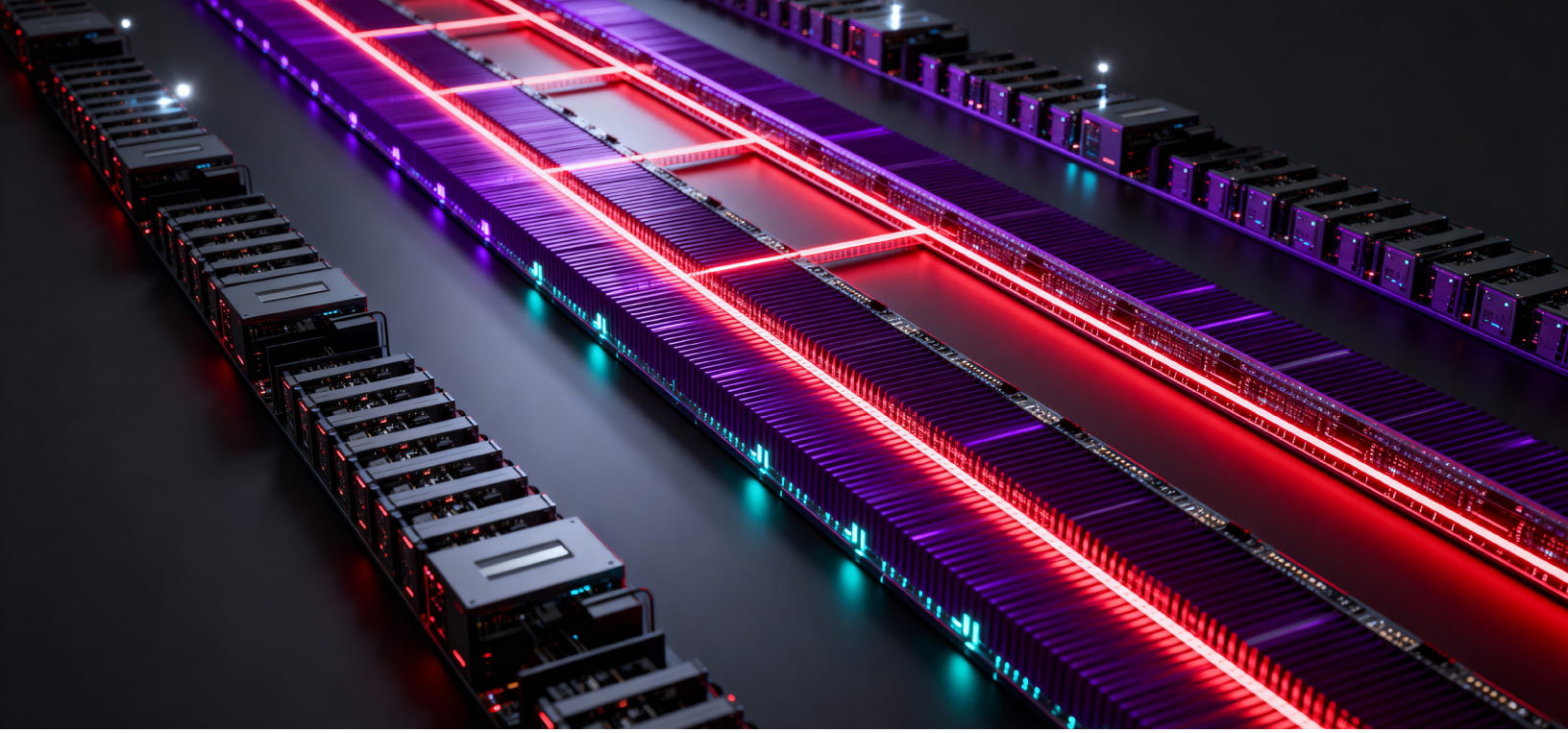
Data and Security Challenges

Data quality issues underlie many unsuccessful projects. Inference models are only as effective as production data. Concept drift - systematic differences between training and production data - degrades accuracy over time. Silent model failures due to malformed inputs require continuous monitoring and governance processes.

Regulatory compliance introduces substantial operational requirements. GDPR, the EU AI Act, and sector-specific regulations (e.g. HIPAA, GLBA) require understanding how personal data is processed. Large language models trained on internet-scale data may encode personal information, raising complex compliance questions. Privacy-preserving inference techniques and compliance audit capabilities must be architected into infrastructure decisions.

Data residency requirements further constrain deployment flexibility. Many jurisdictions mandate that data remain within specific geographic regions. This requirement effectively mandates on-premises or regional deployment for certain workloads rather than centralized cloud systems, adding architectural complexity.

Model security becomes increasingly important as models become valuable intellectual property. Models can be reverse-engineered through adversarial testing. Inference APIs must be designed to reveal minimal information about model structure while remaining functional. Model watermarking and protective techniques remain immature.



5. Requirements for Specialized AI Inferencing Infrastructure

General-purpose infrastructure, such as servers designed for database workloads, web applications, or virtualized environments, does not adequately serve AI inference requirements. Specialized infrastructure addressing the distinct characteristics of inference workloads is essential.

Why General-Purpose Architectures Fall Short

General-purpose server architectures balance compute, memory, I/O, and power across a wide range of possible workloads. This balanced design is optimal for a few specific workloads. AI inference exhibits characteristics that make general-purpose architectures inefficient. CPU-to-GPU balance in general-purpose systems is typically sized for workloads requiring significant CPU processing. Inference workloads, by contrast, are accelerator-dominated; the GPU or specialized accelerator performs the vast majority of computation, while CPU activity is limited to orchestration and I/O management. General-purpose architectures overprovision CPU capacity and underprovision memory bandwidth and accelerator interconnection.

Hardware and Platform Requirements

GPU accelerators remain the dominant compute platform for inference workloads due to their combination of performance, programmability, and ecosystem maturity. NVIDIA's data center GPUs (H200, Blackwell B200, GB200) dominate the market, though competition is increasing from AMD (MI300X) and Intel. GPU selection must account for model requirements, memory constraints, and total cost of ownership.

High-bandwidth memory (HBM) is essential for large model inference. HBM, integrated directly on GPU packages, provides 3–5x higher bandwidth than standard memory, substantially improving inference throughput for memory-bandwidth-limited workloads. However, HBM capacity constraints remain a limiting factor for very large models, creating trade-offs between memory capacity and bandwidth.

CPU-to-GPU architecture must be optimized for inference communication patterns. PCIe Generation 5 and higher-speed interconnects (NVLink) reduce data movement latency between CPU and GPU. Multi-GPU systems require high-bandwidth interconnection to coordinate distributed inference efficiently.

Cooling and power delivery infrastructure must account for concentrated heat density. Liquid cooling, optimized airflow design, and direct-to-chip cooling approaches are becoming necessary rather than optional in high-density deployments. Power delivery infrastructure must provide substantial, consistent power supply with redundancy.

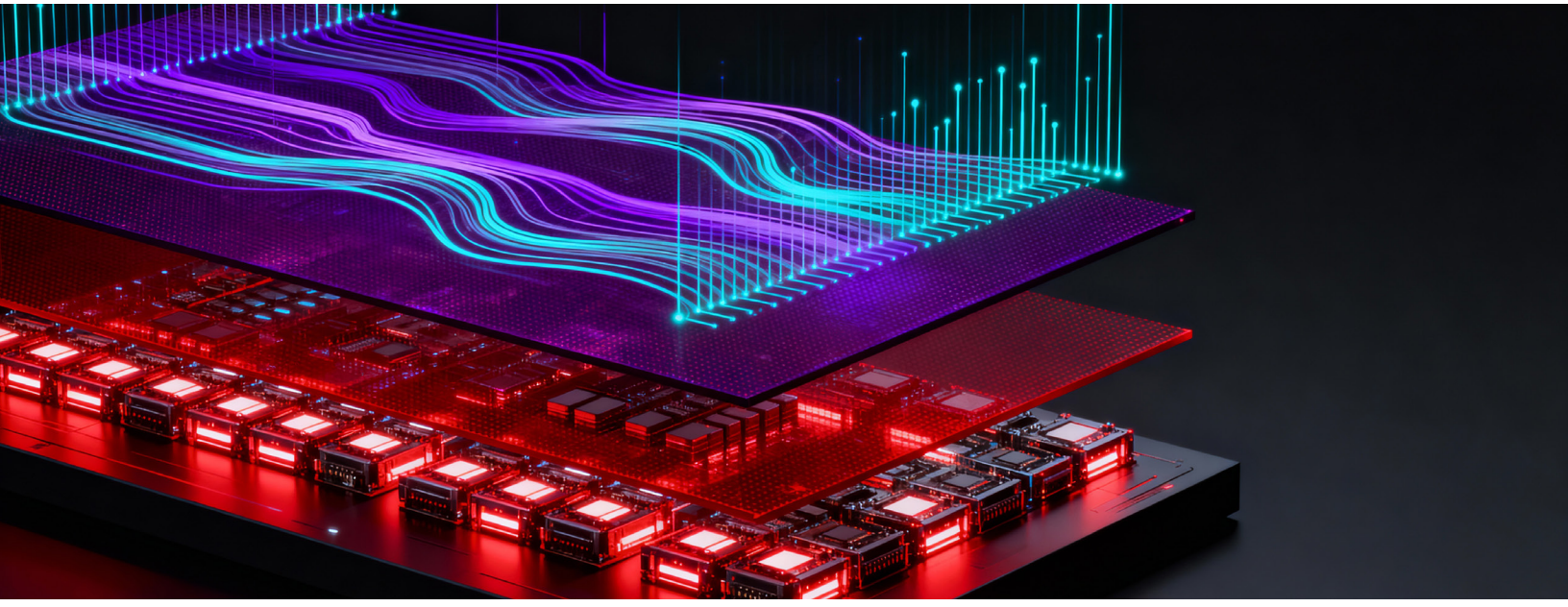
Networking infrastructure must provide low-latency, high-bandwidth connectivity. RDMA capabilities, 100+ Gbps connectivity, and optimized switching reduce inference latency in distributed scenarios. Storage architecture must provide high-speed access to input features. NVMe drives, parallel file systems, and distributed object storage enable rapid data access.

Software and Deployment Stack

Effective inference infrastructure requires a comprehensive software stack addressing model optimization, runtime execution, resource orchestration, and operational management.

Model optimization is foundational. Deploying large models efficiently requires quantization (reducing model precision from float32 to float16 or int8), knowledge distillation (training smaller models to replicate larger ones' behavior), and pruning (removing parameters with minimal contribution to accuracy). These techniques are essential to make models practical for deployment but require specialized expertise. PyTorch and TensorFlow offer quantization tooling, but effectively applying these tools requires deep understanding of trade-offs.

Inference engines - runtime systems optimized for executing trained models - represent the next critical layer. Examples here include NVIDIA's TensorRT, Intel's OpenVINO, and the open source ONNX Runtime.. These runtimes apply graph optimization, operator fusion, kernel selection, and other techniques to maximize throughput. Critical to modern LLM



inference, these engines implement efficient KV cache management - a technique that stores key-value pairs from previously processed tokens to avoid recomputing them during token generation. Efficient KV cache handling directly impacts both latency (improving Time to First Token) and memory usage, often determining whether inference remains practical at scale. Inference engine selection affects model portability, vendor dependency, and long-term flexibility.

Containerization and DevOps processes have become standard. Models, inference code, and dependencies are packaged in containers managed through Kubernetes. This approach enables consistent deployment across cloud, edge, and on-premises environments. However, it requires organizations to develop DevOps competency.

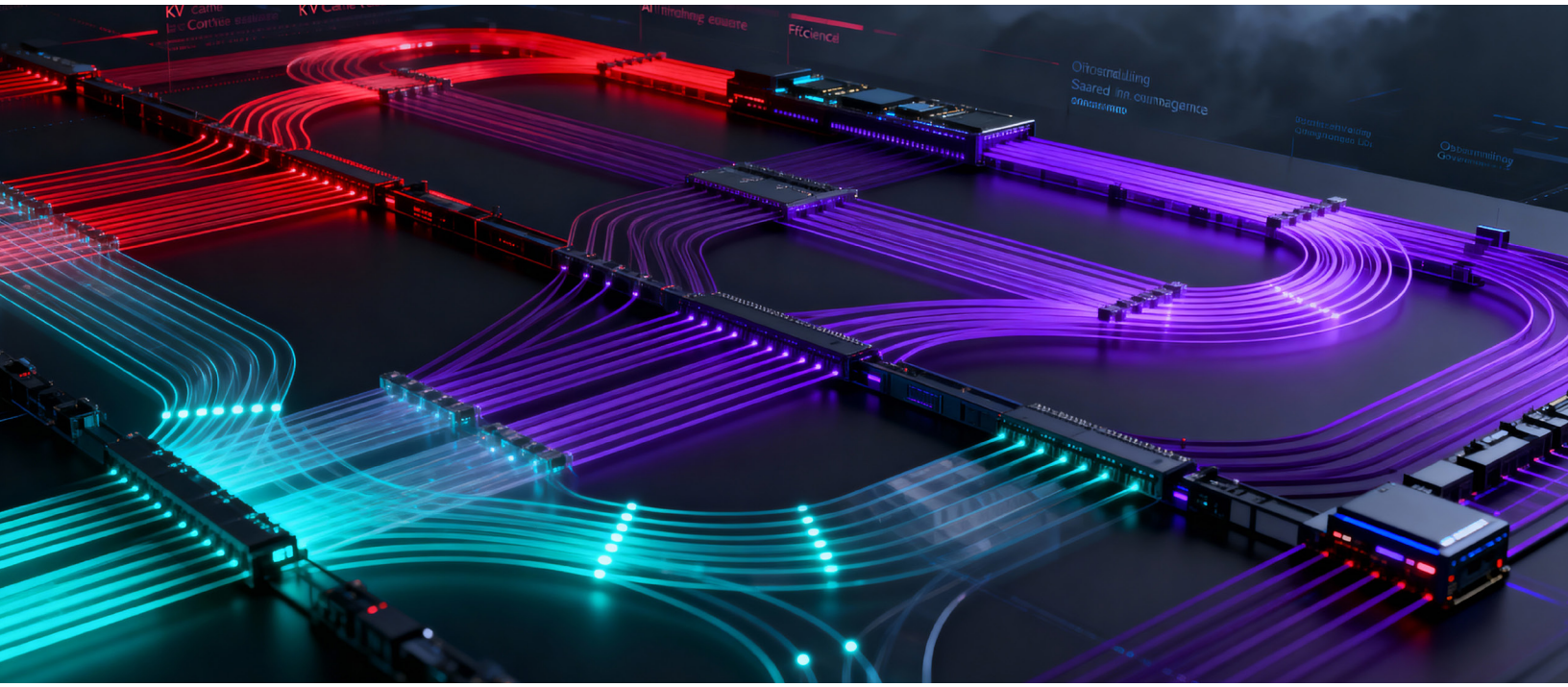
MLOps platforms provide model versioning, A/B testing, performance monitoring, and automated retraining. Organizations deploying at scale find unstructured model management untenable. MLOps platforms provide essential governance but add complexity and overhead.

Resource orchestration and GPU scheduling remain challenging in shared environments. When multiple teams compete for accelerator resources, scheduling mechanisms must ensure fair access, priority management, and efficient utilization. Standard Kubernetes scheduling is often inadequate for GPU workloads.

Observability - visibility into model performance, resource utilization and system behavior - is essential for operational reliability. Inference systems must measure latency, throughput, accuracy, resource utilization, and error rates.

Expertise and Services Requirements

Infrastructure alone is insufficient. Organizations deploying at scale require specialized expertise. Advisory services help organizations assess current state, define architecture, and select appropriate technologies. Deployment and integration services address the integration of new infrastructure with existing systems and validating performance requirements. Optimization and managed services - which provide ongoing tuning of infrastructure, responsive governance, and talent upskilling as workloads mature - maximize performance and value realization from inferencing investments. Cross-functional collaboration facilitation helps organizations align data scientists, ML engineers, infrastructure architects, and security teams.



6. Evaluating AI Inferencing Solutions

The vendor landscape for AI inference infrastructure is complex and fragmented. Organizations evaluating solutions should apply a consistent assessment framework addressing both current capabilities and longer-term alignment.

Comprehensive Capability Assessment

Hardware breadth and depth matter. Vendors should offer a portfolio spanning from AI-optimized workstations for developers through edge-appropriate platforms to hyperscale data center systems. Within each category, support for multiple accelerator options - including NVIDIA GPUs, AMD accelerators, Intel accelerators, and potentially custom silicon - reduces lock-in risk and future flexibility. Hardware roadmaps extending multiple years ahead demonstrate vendor commitment and strategic alignment.

Software integration and optimization tools are often the differentiator between commodity infrastructure and genuinely optimized solutions. Vendors should offer integrated management platforms addressing model deployment, resource orchestration, performance monitoring, and operational support. Integration with enterprise systems, ISV applications and open-source frameworks is essential. Firmware-level optimizations tailored to specific accelerator and model combinations can materially improve inference performance.

Hybrid architecture support - seamless orchestration across cloud, edge, and on-premises infrastructure - is increasingly necessary. Vendors should demonstrate capability to deploy consistent platforms and management approaches across multiple deployment environments rather than requiring separate solutions for each environment.

Expertise depth should extend beyond product support. Vendors offering strategic advisory, engineering support, and ongoing optimization services provide greater value than those offering hardware and software alone. Reference customers deploying at substantial scale provide evidence of vendor capability.

Key Vendor Selection Criteria

Performance validation is essential. Vendors and independent reviewers should publish benchmark results using relevant models and workload patterns. Benchmarks should include inference latency and throughput, Time to First Token (TTF), power efficiency, and cost per inference. Benchmarks on synthetic workloads are far less valuable than results using production models and real application patterns.

Scalability from small pilots to production scale is critical. Vendors should demonstrate experience deploying from single-system configurations through multi-thousand GPU installations, with global capabilities to support the adoption and management of inferencing environments. Architecture should scale efficiently without requiring fundamental changes as workload size increases.

Total cost of ownership (TCO) must be transparent and include infrastructure cost, power and cooling, operational overhead and support expenses. Vendors should provide clear cost modeling tools enabling customers to understand long-term cost implications across different deployment scenarios and utilization patterns. Vendor lock-in through proprietary software or incompatible hardware extensions increases long-term cost.

Architecture and deployment flexibility matter substantially. Vendors should support deployment on standard hardware where practical, reducing dependency on vendor-specific systems. Integration with multiple accelerator vendors, container orchestration platforms, and inference frameworks provides flexibility to adapt as technology evolves.

Security and compliance features should be built-in rather than optional. Hardware-level encryption, secure boot, trusted platform modules (TPMs) and integration with enterprise identity and access management systems are essential for compliance-sensitive deployments.

Open ecosystem support is important. Vendors offering support for open-source frameworks, avoiding proprietary lock-in, and contributing to community standards provide greater long-term value. Support for ONNX (Open Neural Network Exchange), Kubernetes, and similar open standards enables organizations to adapt as technology and requirements evolve.

Track record and reference customers demonstrate real-world experience and vendor reliability. Organizations deploying at meaningful scale provide evidence of vendor capability, and willingness to provide references suggests confidence in customer satisfaction.

7. Lenovo for AI Inferencing

Lenovo has established itself as a comprehensive partner for enterprise AI inferencing, combining purpose-built hardware with integrated software platforms and deep technical expertise. Beyond initial deployment, Lenovo positions itself as a long-term inference operations partner, supporting customers across the full lifecycle of AI workloads. By aligning its global manufacturing, services, and extensive partner ecosystem with highly reliable AI infrastructure, Lenovo delivers full-stack solutions that transform AI potential into sustained business value.

Maximize ROI with Efficient Infrastructure

Lenovo's portfolio is engineered to optimize the most critical performance metrics for modern enterprises, specifically focusing on Time to First Token (TTFT). In a production environment, TTFT is the definitive benchmark for measuring how quickly AI compute investments translate into responsive, user-ready applications. By utilizing advanced Neptune liquid cooling, air cooling options and thermally optimized chassis designs, Lenovo systems allow GPUs to run at peak performance without thermal throttling. This sustained performance directly minimizes latency and ensures that enterprises can scale their AI deployments while maintaining the high-speed responsiveness required for real-time decision-making.

Hybrid AI: Personal, Enterprise, and Public

Lenovo's hybrid AI strategy reflects the reality of modern adoption by delivering a consistent experience across personal, enterprise, and public deployment models:

- **Edge and Physical AI:** Lenovo ThinkEdge systems extend inference into physical and edge environments, enabling real-time processing close to data sources while remaining integrated with centralized management and governance.
- **Enterprise Workstations and Data Centers:** Lenovo ThinkStation and ThinkSystem platforms bring high-performance inference to developers and enterprise environments, supporting local experimentation, secure data handling, and scalable production deployments. Lenovo XClarity One provides a unified management plane to simplify operations across distributed environments.
- **AI Gigafactories (Cloud and CSPs):** Lenovo enables AI cloud providers and CSPs to build AI Gigafactories at scale, accelerating time to first token while supporting sustained inference performance as workloads grow.

Advisory & Ecosystem

Infrastructure alone is insufficient to deliver lasting AI value. Lenovo's AI Advisory, implementation, and managed services help customers select and continuously right-size models—such as SLMs versus LLMs—while optimizing cost per million tokens over time. Through ongoing cost, performance, and compliance tuning, Lenovo supports the full AI lifecycle beyond initial deployment. Backed by deep partnerships with industry leaders including NVIDIA, AMD, Intel, and Microsoft, Lenovo provides a validated, risk-reduced path to production that remains adaptable as AI strategies evolve.

Across these tiers, Lenovo enables consistent inference governance from development through production, reducing operational friction and risk as AI environments scale.

8. Conclusion

AI has officially transitioned from an experimental phase into its production phase, where the primary driver of economic value is no longer just model training, but inferencing - the deployment of these models to drive real-time business decisions. As enterprises across every major industry move toward production-scale AI, the infrastructure choices they make today carry immediate and lasting strategic consequences.

For the modern enterprise, investing in specialized inference infrastructure is no longer optional; it is a mission-critical priority. While general-purpose architectures may suffice for initial experiments, they fall short of the computational patterns and power density requirements of AI at scale, often leading to performance bottlenecks and 'bill shock' as costs scale linearly with token generation. Success in this new landscape requires a shift toward hybrid and distributed architectures that place inference at the data source - whether in the cloud, on-premises, or at the edge - to meet the latency demands of real-time applications.

The strategic window for establishing a leadership position is narrow. Organizations that execute a thoughtful, integrated strategy across hardware, software, and services over the next 12 to 18 months will establish a significant competitive advantage as AI becomes central to their operations. Conversely, those that delay investment or make poor technical choices will face escalating costs and constrained capabilities that will be both difficult and expensive to remediate in the future. By prioritizing specialized infrastructure today, enterprises can ensure their AI investments translate into productive, scalable, and profitable business outcomes.

Important Information About This Report

AUTHORS

Nick Patience
Vice President & Practice Lead,
AI Platforms | The Futurum group

PUBLISHER

Futurum Research

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group 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 "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

DISCLOSURES

The Futurum Group 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.

Lenovo

ABOUT LENOVO

Lenovo is a global technology leader that designs and delivers the infrastructure and devices enterprises rely on to run modern, data-intensive workloads. As AI moves from experimentation to production, Lenovo is helping organizations scale AI inferencing across data centers, hybrid environments, and edge locations—where low latency, efficiency, and operational simplicity matter most. Its portfolio spans AI-optimized servers, storage, networking, and workstations, supported by services and deployment expertise that help teams move from pilots to reliable, repeatable operations. With an emphasis on performance, energy efficiency, and manageability, Lenovo enables enterprises to operationalize AI while aligning infrastructure choices to cost, compliance, and business outcomes.

Futurum®

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

Futurum®
CONTACT INFORMATION: The Futurum Group LLC | futurumgroup.com | (833) 722-5337