Solution brief

Edge AI inference: powering the next wave of innovation

Secure, scalable AI at the edge with HPE ProLiant Compute

Business value of AI inferencing at the edge

In today's data-driven world, the ability to make real-time decisions—right where data is created—has become a defining advantage. Edge AI inferencing empowers organizations to act instantly on insights from sensors, cameras, and connected devices, delivering new levels of operational efficiency, customer experience, and resilience. By processing data locally, businesses avoid the cost and delay of sending high-volume streams to the cloud, while also strengthening privacy and compliance.

The momentum behind edge AI is accelerating:

— The global edge AI market is projected to reach $25–26 billion in 2025, with growth rates exceeding 20% annually through 2030.

— More than 90% of CIOs now view edge AI as essential for driving innovation and ensuring business continuity.

— Regulatory frameworks, such as the EU AI Act, increasingly require on-premises, auditable inferencing for sensitive workloads—making edge solutions not just strategic, but necessary.

Edge AI inferencing is more than a technology trend—it's a catalyst for growth, enabling organizations to compete, adapt, and lead in a rapidly evolving landscape.

Vertical	Top edge use cases and value
Manufacturing	**Vision QC:** Real-time defect detection on production lines **Predictive maintenance:** Early equipment failure alerts **Autonomous robotics:** AI-driven process optimization and safety
Defense	**Threat detection:** Real-time surveillance and anomaly detection **Situational awareness:** AI-powered data fusion from multiple sensors **Mission planning:** Automated resource allocation and logistics optimization
Telecommunications	**Network optimization:** AI-managed traffic and resource allocation **Fraud detection:** Real-time call/data anomaly detection **Service assurance:** Predictive outage prevention
Retail	**Loss prevention:** Instant video analytics for theft detection **Shelf analytics:** Automated inventory tracking **Autonomous checkout:** Seamless customer experience with AI
Healthcare	**Imaging analysis:** On-site diagnostics for faster treatment **Remote monitoring:** Real-time patient alerts **Asset tracking:** Efficient supply chain and equipment management
Public sector	**Traffic management:** Smart city video analytics for flow and safety **Emergency response:** Rapid event detection and dispatch **Crowd analytics:** Real-time monitoring for public safety
Energy/utilities	**Grid monitoring:** Detect anomalies and optimize grid performance **Asset health:** Predictive maintenance for infrastructure **Demand forecasting:** AI-driven load balancing and planning
Transport/logistics	**Fleet management:** Predictive maintenance and route optimization **Cargo tracking:** Real-time location and condition monitoring **Autonomous vehicles:** AI-powered navigation and safety
Hospitality	**Guest personalization:** AI-driven recommendations and services **Security/access control:** Real-time video inference for safety **Inventory optimization:** Automated supply chain management
Education	**Campus safety:** Video analytics for secure access **Adaptive learning:** Real-time feedback for personalized education **Facility management:** AI-driven resource and space optimization

Technical requirements for inferencing at the edge

Effective edge AI inferencing relies on a set of core technical requirements that ensure performance, reliability, security, and compliance across distributed environments.

1. Performance and acceleration

Edge inferencing demands high-throughput, low-latency processing to support real-time analytics and decision-making. Systems should support hardware acceleration, such as GPUs or specialized AI processors, and enable efficient computation for workloads like computer vision and natural language processing. Support for optimized data types (e.g., INT8, FP16) and robust cooling is essential for sustained performance.

2. Scalability and deployment flexibility

Edge environments vary widely, from small retail sites to large industrial facilities and telecommunications networks. Infrastructure should offer modular and scalable options to accommodate different deployment sizes and mission-critical scenarios, supporting both compact and ruggedized configurations.

3. Security and data integrity

Hardware-embedded security features, such as silicon root of trust, secure boot, and runtime firmware validation help safeguard systems from tampering and unauthorized access. Physical security measures may also be required to prevent theft or intrusion.

4. Centralized management and automation

Managing distributed edge nodes efficiently requires unified control platforms that enable centralized lifecycle management, remote monitoring, automated updates, and integration with hybrid cloud environments. These capabilities reduce operational overhead and facilitate rapid troubleshooting.

5. Compliance and governance

Edge AI deployments must adhere to evolving regulatory requirements, including data privacy and auditability standards. Solutions should support policy-based orchestration, maintain audit trails, and enable responsible AI governance across all locations.

6. Reliability and operational resilience

Continuous operation is essential, even during network disruptions or adverse conditions. Infrastructure should feature ruggedized designs, redundant power options, and remote management capabilities to minimize downtime and support mission-critical workloads.

Top considerations for CIOs and AI practitioners

Align edge AI with business outcomes
To maximize the impact of edge AI, organizations should prioritize deployments that directly support their most important business goals, such as operational resilience, customer experience, and regulatory compliance. Focus on use cases where real-time decision-making at the edge delivers measurable results—examples include predictive maintenance, quality control, customer analytics, and network optimization.

Build for scalability and flexibility
It is essential to select edge infrastructure that can scale across a variety of environments, including factories, retail sites, remote offices, and telco networks. Your solution should be adaptable to evolving business needs, new AI models, and changing regulatory requirements, ensuring long-term relevance and agility.

Simplify management and governance
Investing in platforms that offer centralized management, robust security, and compliance across all edge locations is key. Adopting strong governance frameworks will help ensure responsible AI use, protect data privacy, and maintain auditability, especially as regulations like the EU AI Act continue to evolve.

Empower teams with automation and insights
Leverage automation and AI-driven observability to reduce manual intervention, accelerate issue resolution, and optimize performance. By minimizing routine maintenance and enabling remote operations, IT teams can focus on strategic initiatives that drive business growth.

Future-proof your edge investments
Choose partners and solutions with a proven track record in edge innovation, security, and lifecycle support. Consider flexible consumption models and advisory services to maximize return on investment and ensure your infrastructure can adapt to future growth and technological advancements.

Leadership takeaway

Edge AI inferencing is not just a technology upgrade—it's a strategic enabler for business transformation. By focusing on scalable infrastructure, centralized management, robust governance, and continuous innovation, CIOs and AI leaders can realize new value, drive efficiency, and position their organizations for long-term success at the edge.

HPE ProLiant Compute advantage

Multi-layer security

— Silicon root of trust from HPE iLO built into every server protects from manufacturing to end-of-life and provides compliance readiness for future cyber and physical attacks

— Secure boot and runtime firmware validation prevents tampering and unauthorized code

— Optional physical security features (chassis intrusion kit, bezel lock, Kensington lock) protect devices from theft or intrusion

Industry-leading performance and efficiency

— Efficient processing power for low-latency workloads, reducing reliance on cloud infrastructure

— AI-ready with support for a wide range of NVIDIA® GPUs (L4, L40S, H100) to power inference at the edge for manufacturing, public sector, and defense

— Compact, ruggedized form factors for better cooling, airflow, and quiet acoustics; capable of operating in extreme temperatures and conditions

Operational productivity and manageability

— HPE Compute Ops Management enables robust cloud operations management from a single console, with secure connection to servers and enterprise-grade distributed management

— Seamless multi-vendor server monitoring and workflow approvals reduce downtime and keep edge infrastructure running optimally

— Priority response and access to technical assistance 24x7x365

Compliance and trusted supply chain

— TAA compliance and trusted supply chain with secure facilities from manufacturing to delivery

— HPE delivers industry leading server security from chip-to-cloud, with advanced silicon security and custom-built ASICs and only server security from chip-to-cloud, with advanced silicon security and custom-built ASICs

Quantifiable results

— Lower risk of physical and digital tampering; better compliance readiness and security

— Meet business demands and power any workload anywhere, efficiently, for today and the future

— Simplified management and control over distributed server environments, with visibility, insights, and automated capabilities to take instant action

Shape your next mile of AI

Edge AI inferencing is reshaping how organizations operate, compete, and innovate—delivering real-time intelligence, operational resilience, and new business value at the point of data creation. With the HPE ProLiant Compute purpose-built edge portfolio, industry-leading security, and proven manageability, you're equipped to lead in this new era.

What to do next

- Assess your edge opportunities: identify high-impact use cases in your organization where real-time AI can drive measurable outcomes
- Engage your stakeholders: align IT, business, and operations leaders around a shared vision for edge-enabled transformation
- Partner for success: leverage HPE's expertise, services, and flexible consumption models to accelerate deployment and maximize ROI
- Future-proof your investments: choose scalable, secure, and adaptable infrastructure to stay ahead of evolving business and regulatory demands

Visit HPE.com

Ready to put edge AI to work?

Explore the HPE ProLiant Compute edge computing portfolio, access solution resources, and connect with experts.

Learn more at

HPE.com/ProLiant/edge-computing

Chat now

a00156356ENW

HEWLETT PACKARD ENTERPRISE

hpe.com