October 2025 | White Paper, sponsored by Intel

# EDGE AI:
# Paving the Way for Intelligent, Resilient Deployments with a Systems Mindset

Nina Turner
Research Director,
Semiconductor Technology Supply Chain Intelligence;
Enabling Technologies: Edge AI Semiconductors, IDC

THIS PDF USES HYPERLINKS

# Table of Contents

**CLICK ANY HEADING TO NAVIGATE DIRECTLY TO THAT PAGE.**

# IDC Opinion

**As AI adoption expands at the edge, the market for AI processors and accelerators supporting these workloads is forecast to reach $64 billion, reflecting a five-year CAGR of 16.1%.**

As AI continues to rapidly evolve, AI compute at the edge is emerging as a foundation for running a diverse mix of AI and legacy applications. As AI adoption expands at the edge, the market for AI processors and accelerators supporting these workloads is forecast to reach $64 billion, reflecting a five-year CAGR of 16.1%. The edge is dominated by distributed systems, and companies will also need to distribute their AI compute and operations to be effective. Taking a systems-oriented approach reflects the real-world needs of industries where distributed processes, cost-effectiveness, operational efficiency, reliability, and scalability are paramount. Avoiding siloed AI systems and integrating AI into existing workflows can enable companies to combine traditional compute with AI to enhance and improve operations.

AI at the edge can create entirely new workflows, or fit within existing workflows, by either replacing certain functions or adding additional capability. For example, in industrial and healthcare settings, AI-driven computer vision may increase productivity and accuracy by assisting with detection or identification and make decision suggestions, while final decision-making and control functions remain governed by conventional rules-based software or human oversight.

Organizations must evaluate the capabilities and strengths of the AI being implemented and leverage AI while maintaining continuity with established operational processes.

Open standards and open software can reduce vendor lock-in and provide flexibility to choose best-of-breed solutions from multiple suppliers. Open platforms allow enterprises to build upon existing software investments, leveraging open source foundational models (e.g., Llama from Meta) and software frameworks and enabling organizations with limited in-house expertise to leverage work from the larger AI expert community.

Total cost of ownership (TCO) remains a central consideration, and systems with the ability to run multiple AI-enabled workloads alongside legacy software will enable organizations to balance operational functionality, performance, energy efficiency, and system cost. The edge AI market growth drivers include the need for real-time on-premises data processing, reduced latency, and enhanced privacy. On-premises systems may have limited available power and connectivity, or the enterprise may wish to limit cloud or datacenter compute costs.

The future of advanced edge AI computing lies in system-level solutions that seamlessly blend AI and legacy workloads, are built on open standards, and are designed for operational efficiency and adaptability. This approach will be essential for unlocking the next wave of innovation and value at the edge.

## IN THIS WHITE PAPER

**This white paper explores the evolving landscape of edge AI, examining the drivers of adoption and the shifting mix of AI workloads. It highlights the importance of a systems-level approach to integration, the role of open software and standards in supporting vendor flexibility, and total cost of ownership considerations that can impact deployment decisions. The discussion addresses how energy efficiency, interoperability, and the rise of open source foundational models are accelerating the democratization of AI for the edge. The paper provides a future outlook for edge AI adoption, outlining key challenges and offering recommendations for successful edge AI deployments.**

# Situation Overview

## Edge AI Adoption and Workload Trends

### As AI frameworks and algorithms continue to develop in the cloud, new techniques and innovations are enabling and evolving edge AI capabilities.

Optimization techniques such as quantization, pruning, structured sparsity, and knowledge distillation reduce model size and improve compute efficiency, helping bring AI models from the datacenter to the edge. In parallel, small language models (SLMs), today generally under 10 billion parameters, have emerged to deliver conversational functionality on embedded systems with lower memory footprints. Large vision models (LVMs) are also beginning to move from the cloud to the edge, with multimodal versions that combine image, video, or other sensor inputs with language models to deliver richer context and more actionable insights in real time. Emerging vision-language-action (VLA) models that integrate natural language instructions with visual AI perception are trained on “physics of the real world” and can deliver an action plan such as “There is a hazard in the path, move to the right to go around, and create alarm for maintenance,” enabling physical edge AI system capabilities. Edge systems attach the physical world to AI, driving intelligence and increased capabilities into robotics, autonomous vehicles, drones, and many other systems that interface with the real world.

Despite the work and research on new edge AI frameworks and optimizations, adoption of production level AI is largely focused on machine vision AI with many companies in the early stages of understanding AI and many still in proof-of-concept evaluation. The implementation of AI into a company's processes goes beyond the AI algorithm itself and into the complexities of operational processes, integration with existing infrastructure, and deployment and management of the AI algorithms across many edge systems. Once companies better understand and learn from pilot programs and operational best practices with AI, the barriers to implementation will decrease and the number and complexity of edge AI deployments will accelerate.

Once companies better understand and learn from pilot programs and operational best practices with AI, the barriers to implementation will decrease and the number and complexity of edge AI deployments will accelerate.

## The drivers that are accelerating the interest and acceptance of edge AI are:

- **Real-time responsiveness** is required for safety-critical tasks across many industries, such as autonomous material handling, energy grid-fault detection, and autonomous driving where critical millisecond decisions are necessary. In industrial settings, real-time responsiveness enables systems to react swiftly to identify defects and out-of-specification parameters and adjust operations in milliseconds, preventing costly errors and downtime. In healthcare, real-time AI can monitor vital measurements and create alerts.

- **On-premises data processing** reduces the need to transmit large volumes of data to the cloud, which lowers data transfer costs and ensures compliance to regulations that restrict sensitive data from crossing borders or cloud boundaries, such as the European Union's (EU's) GDPR and Brazil's LGPD. By keeping processing local, on-premise edge AI also improves operational efficiency and ensures compliance with industry-specific requirements in sectors such as finance, healthcare, government, and other critical infrastructure that mandate control of sensitive data.

- **Energy efficiency and power constraints** are key drivers of edge AI. Minimizing data transmission to centralized datacenters and leveraging optimized application-specific edge AI algorithms reduce the resources required to run AI. Edge AI systems are often designed with a limited power budget targeted for the specific application, avoiding the energy overhead of cloud infrastructure. Edge AI enables companies to spend less on cloud compute, network bandwidth, and energy while staying within the power constraints of operational environments.

- **Security and privacy** are central drivers of edge AI adoption, as local processing on edge devices reduces the potential attack surface, lowers the risk of sensitive data exposure, and supports compliance with evolving data protection regulations. Taking a multilayered security approach safeguards personal and proprietary data while building user and consumer trust.

# Systems Integration and a Systems Mindset

Edge AI implementations that utilize standalone, "bolt-on" AI systems are unable to truly unlock the full potential of AI, both financially and operationally. Isolation from business processes results in fragmented workflows, duplicative effort, and limited impact. Siloed systems may deliver single benefits but rarely can deliver the business transformation that AI promises. Real efficiencies and operational value emerge when AI is integrated into existing operations and workloads, enabling seamless data flow, automated decision-making with the ability to improve and optimize processes. Integrating edge AI with other systems across the enterprise allows for more sophisticated and impactful decision-making.

Edge AI system form factors such as edge gateways, embedded servers, and industrial PCs are increasingly focused on energy efficiency and compact designs, especially in industrial environments or in applications where environmental factors favor ruggedized, fanless enclosures to minimize failures due to dust or debris. Wide temperature operating ranges are common in edge deployments with limited active cooling. Beyond the physical factors, edge systems must deliver more than just AI inference. They need to integrate AI and non-AI functions, such as real-time edge AI process control with legacy business logic. Edge AI must work together with long-established business functions and rules while providing new ways to deliver insights and recommended actions. For example, in manufacturing, it can improve production line efficiency with early defect detection and predictive maintenance recommendations. In retail, it can optimize inventory management, and in healthcare, it can support faster diagnostics at point of care. Edge AI augments and enhances operations while allowing companies to continue operating without disruption. Companies are looking to upgrade to system-level edge solutions where multiple functions are consolidated into a platform rather than siloed, standalone deployments. These unified system-level solutions simplify deployment and management while supporting diverse workloads. This unified system-level solution reduces IT/OT system management and networking, simplifies maintenance, and helps ensure consistent system security and governance. Edge AI can be deployed alongside existing applications without the need for complex reengineering or parallel infrastructure. System-level solutions enable companies to adapt to evolving demands with less complexity and lower total cost of ownership.

# Open Software, Standards, and Vendor Flexibility

As companies implement their edge AI projects, the advantages of open software and frameworks and open standards allow companies to have more vendor flexibility, reducing costs, and avoiding vendor lock-in. Proprietary, closed ecosystems add a layer of risk that can lead to increased costs, limited innovation, and barriers to adopting new solutions. Open standards and software can provide a base for interoperability, allowing edge AI systems to work within a diverse ecosystem of hardware and software. The flexibility enables customers to select best-of-breed systems and components, migrate workloads as the business needs evolve, and negotiate with a wider base of suppliers to lower costs and operational expenses. In addition to providing a competitive supplier base, open source software can help prevent fragmentation and integration challenges with siloed systems and simplify deployment and management across heterogeneous edge system environments. Companies can benefit from the open source community to accelerate return on investment and create a path to scale edge AI implementations.

**Companies can benefit from the open source community to accelerate return on investment and create a path to scale edge AI implementations.**

Leveraging open source foundational models can greatly reduce the time and cost of model training, enabling developers to quickly prototype and deploy edge AI. Open source frameworks and runtimes such as PyTorch, ONNX Runtime, and Intel's OpenVINO provide community-supported solutions for model optimization and deployment across diverse hardware and software. In addition, newer tools designed for large language model deployment such as Ollama, LlamaIndex, and LangChain are lowering the barrier to entry for edge AI deployments, making it easier for companies to fine-tune and customize AI to meet specific edge application requirements.

# Total Cost of Ownership

As companies evaluate AI initiatives, edge AI can deliver long-term cost advantages due to its focus on local data processing, which reduces the cost of constant cloud connectivity, consumption-based cloud AI processing fees, and cloud data storage fees. Companies recognize the need to carefully evaluate the total cost of ownership against the desired AI performance. TCO includes not only the up-front capital expenditures for the hardware but also ongoing operational costs that will be used over the system lifespan like power, software licensing, management, and maintenance. For edge AI, these operational costs can be significant and exceed the up-front capital costs. All these factors must be considered as models scale and grow in complexity and require more computational resources. Over a multiyear period, on-premises edge deployments can offer significant savings compared with cloud-based alternatives, especially for stable, high-throughput workloads where infrastructure is sized appropriately and energy costs are managed efficiently.

## AI and the Drive for Energy Efficiency

Worldwide, AI is being recognized as a significant driver of energy consumption, leading to initiatives in many countries and regions. In Europe, energy efficiency is a critical factor where it is a regulatory and strategic priority. The European Union has enacted initiatives such as the EU Energy Efficiency Directive and Green Deal that set aggressive targets for reducing energy consumption and carbon emissions. Under the EU Data Center sustainability rating regulation, datacenter operators with an installed IT power demand of at least 500kW must report a set of 24 key performance indicators including energy use, renewable energy sourcing, water usage, cooling efficiency, and waste heat reuse. This reporting is designed to increase transparency and drive innovation in energy and resource efficiency.

While Europe is at the forefront of energy policy, other regions are developing energy policy as the growth of AI surges. China's national strategy is to integrate AI development with energy planning, coordinating the expansion of datacenters with renewable energy infrastructure. Japan is prioritizing the expansion of renewables as it recognizes the increased energy demand from AI and datacenters. South Korea has launched a number of funds to support AI innovation and the development of next-generation semiconductors and is promoting smart grid technologies. The United States policy shifts make the previous grid modernization and clean energy initiatives unclear in the short term, but leading datacenters operators recognize the importance of energy efficiency as AI compute becomes energy limited rather than compute or memory bound.

The challenges of high energy use of centralized datacenters are another key factor in the attractiveness of edge AI as an alternative. The local data processing of edge AI systems minimizes the energy-intensive transfer of large volumes of data. According to the World Economic Forum, processing AI tasks locally at the edge rather than via cloud datacenters can reduce energy consumption by 100x to 1,000x per task. Offloading AI workloads that benefit from the real-time availability of edge AI reduces the need for low-latency, energy-intensive cloud infrastructure. Edge AI systems often are optimized for lower power consumption. As the number of connected devices grows, distributing the AI workloads across the edge reduces the need for centralized datacenters. Companies deploying edge AI solutions can control operational costs by reducing datacenter dependency, better meet sustainability goals, and operate in areas with limited connectivity or power infrastructure.

## Importance of Integration and Interoperability

The promise of edge AI is centered not just in faster, real-time analytics but in tighter integration with legacy systems and with broader operational workflows. The real value of edge AI is in driving enhanced automation and will only be fully realized when AI not only works within the existing data and system architectures but also finds insights that can feed directly into an organization's operations. Examples include the detection of defects on a production line or predictive maintenance insight that feeds directly into ERP and factory management applications that can autonomously adjust process settings, schedule maintenance, or trigger an alarm.

AI will usually fit within and enhance existing workflows and processes rather than operate in isolation. Most enterprises and organizations have long-established software stacks and are looking to avoid disruptive changes. Platforms that can both run AI and perform non-AI tasks mean IT and OT managers do not need to deploy AI-specific hardware and training or perform extensive rearchitecting of the operational infrastructure. This avoids disruptive changes to accommodate the use of AI.

Edge AI that offers interoperability with existing IT and OT infrastructure and utilizes standardized protocols and open middleware enables simplified deployments for both greenfield and brownfield. Edge AI running on systems that implements open software ensures simplified implementation, testing, and reduced implementation costs, making edge AI an upgrade rather than an upheaval.

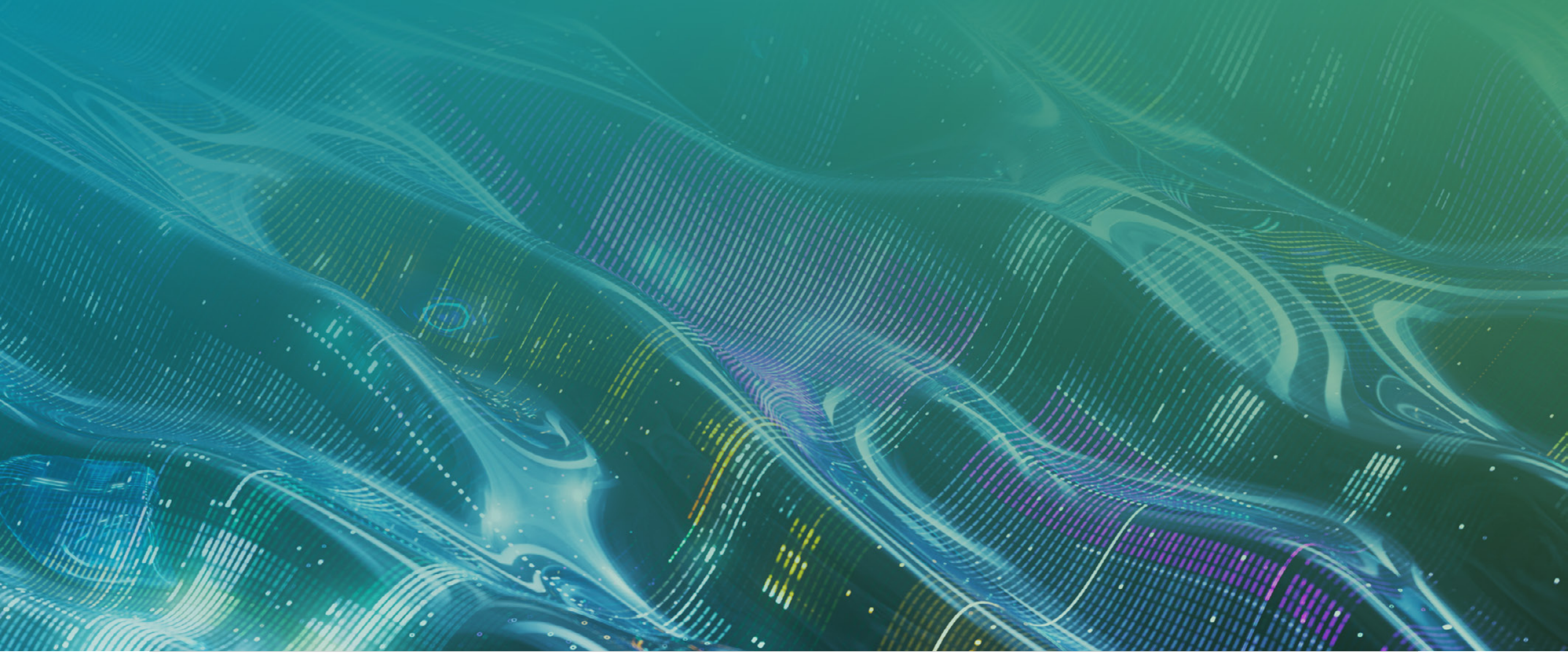
**Edge AI running on systems that implements open software ensures simplified implementation, testing, and reduced implementation costs, making edge AI an upgrade rather than an upheaval.**

# The Democratization of Edge AI

The rapid development and increased availability of open source foundational models toolkits are lowering the barriers to AI innovation.

As of mid-2025, Hugging Face, an open source AI community, hosts approximately 1.86 million models, reflecting an active open source community. Improved performance between the top open weight and closed models shows improving parity in capability. However, people and skills continue to be critical for large-scale rollouts that are integrated into the enterprise. The AI ecosystem from semiconductor vendors to AI companies recognizes the criticality of investing in training, education, and workforce development. Vendors are launching "academy" programs and credentialing training programs to encourage the education and development of more AI proficient professionals. Universities are adjusting curricula to enable skills in edge AI development, from time-sensitive networking and scheduling to security and protocol communication. These efforts will gradually lead to a broader community capable of developing and supporting edge AI solutions.

The rapid development and increased availability of open source foundational models toolkits are lowering the barriers to AI innovation.



# Future Outlook

## Over the next five years, edge AI will evolve from isolated "AI add-on" applications to being fully integrated within an organization's systems, becoming valued enhancements and delivering critical new functionality.

Early edge AI deployments are largely being built around computer vision, providing real learnings as organizations become more comfortable with edge AI implementations and the corresponding TCO. As companies better understand the benefits of edge AI and how to optimize algorithms, new innovations in small language models, domain-specific language models, multimodal AI, large vision models, and vision language action models will be increasingly adopted at the edge. Virtual machines will be utilized to run multiple programs on a single system — from the edge AI algorithm to other functions such as control, scheduling, or orchestration.

By 2030, IDC forecasts edge systems to make up 89% of the systems sold worldwide. This rapid proliferation of edge systems will require increasingly diverse compute capabilities tailored to support the wide range of AI workloads expected at the edge. To provide system developers with the flexibility to adapt to edge AI workloads, semiconductor vendors are integrating AI accelerators on die or in-package and also developing discrete accelerators. The ubiquity of edge AI acceleration will shift the market from raw performance toward ease of integration, security, and the maturity of supporting software toolchains. As IT or OT operations grow more familiar with edge AI, the variety and scope of algorithms will expand from single sensor siloed systems to multimodal inputs capable of aligning with operational requirements. Semiconductor vendors are also working to enable complex VLA models to run within the limited processing, memory and power envelopes of edge systems.

The power demands of datacenter AI encourage offloading some functionality toward the more power efficient edge as enterprises evaluate both the up-front capital costs and the ongoing operational costs of generated tokens/watt. This will be especially true in regions with high energy costs, energy grid constraints, or where concerns about the sustainability of AI exist. Edge AI adoption may be accelerated in those regions, and learnings from those best practices will gradually migrate to other regions. As open source model libraries expand, user-friendly model optimization tools mature and targeted training programs reach operations and IT teams; the talent bottleneck that once constrained edge AI projects will loosen. A larger, better-equipped practitioner community will keep refining the edge stack, enabling more integrated workloads and pushing edge AI beyond proof-of-concept pilots into full-scale, day-to-day production. Shared best practices will shorten deployment cycles from months to weeks, while standardized certification paths will give executives confidence in the skills of their teams. This cycle of collaboration, tooling, and education will accelerate innovation at the edge, making AI-driven insights a routine part of operational decision-making.


Shared best practices will shorten deployment cycles from months to weeks, while standardized certification paths will give executives confidence in the skills of their teams.

# Challenges/Opportunities

## Challenges for Edge AI Adoption

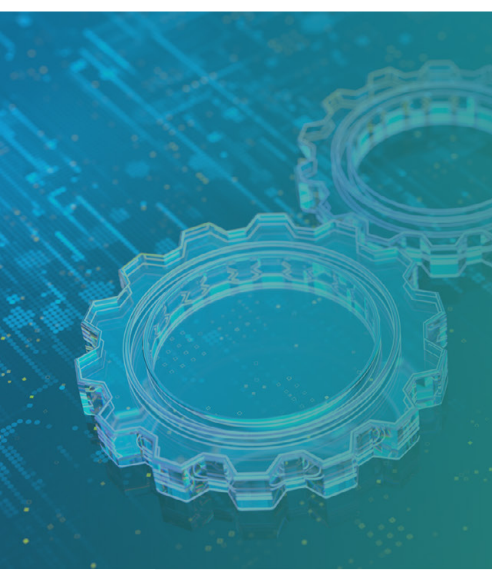
Building intelligent edge systems into an existing operation is less about installing the latest GPU or neural-processing chip and more about mastering the tough, everyday constraints that come with running real workloads outside the datacenter comfort zone.

### Five key challenges that have emerged are:

- **Systems integration complexity:** Every company or organization already runs a complex landscape of control loops, databases, and dashboards that keep the operation running. Inserting AI into the environment without disruptions to production schedules or quality, or data integration issues, requires a system-level mindset. The edge system must juggle multiple algorithms from vision or AI inference, process control, and routine business logic on the same system, under the same real-time conditions. That requires alignment of drivers, kernels, middleware, and algorithms, so that, for example, a sensor data burst does not impact other critical operations.

- **Resource constraints:** Unlike over-provisioned cloud servers with active cooling and kilowatts of headroom, edge devices are often space and energy constrained and may have to operate under harsh environmental conditions. This means processors may have a limited number of cores, have a modest accelerator, and have limited memory. Meeting AI latency targets in that envelope can often call for nontrivial optimizations such as quantized models, smart batching and pruning, and local orchestration policies that throttle noncritical tasks when necessary. Choice of silicon must be carefully considered as a key determiner of performance under these constraints.

- **Operational management at scale:** A proof-of-concept might involve a few homogenous nodes, while production could mean hundreds or thousands of nodes scattered across geographies and time zones and running on heterogenous compute. An enterprise's operations must manage collecting, processing, and analyzing data across those distributed systems, manage model introspection and drift, and perform general hardware and software life-cycle management.
- **Security and privacy:** Every new edge node is another connected endpoint and therefore another potential intrusion point for attackers. Best security practices begin with secure boot and hardware roots of trust, extend through single sign-on and data encryption, and end with continuous monitoring that flags drift from approved software baselines. In addition, edge deployments require a fallback layer of control to ensure recovery and uptime even under attack or failure. Out-of-band manageability provides a secure backup path to monitor and restore edge nodes.
- **Expertise gaps:** Today, few organizations keep a large team of engineers fluent in all the AI frameworks, real-time process control, and industrial networking. Open standards and community supported software tools are critical for enabling the edge AI market. Using widely adopted open source programs and standard data communication protocols lets teams utilize external training, hire from a broader talent pool, and avoid being limited by a proprietary stack.



Addressing these challenges early — ensuring security in automating operations, optimizing for constrained resources, and insisting on system-level unification — moves edge AI from an early pilot program into a resilient, scalable, and integrated part of everyday operations.

# Opportunities and Recommendations for Successful Edge AI Programs

- **Adopt a system-level mindset.** Enterprises must shift from bolt on AI applications to looking at how AI intelligence can improve the entire workflow in a system (this applies to edge AI and cloud AI). Choose edge solutions capable of running all the workloads needed at the edge such as vision inference, process control, data logging, and even those business applications that are required, on the same system, sharing one system's management pipeline and a unified data architecture. When AI flags a defect or out-of-specification parameter and the very next instruction automatically adjusts the operational parameters or updates the quality log, operations can seamlessly continue without significant disruption. This is how edge AI enhances automation.

- **Leverage open, flexible platforms.** Choose open, flexible platforms for edge AI projects to avoid one-off isolated applications. Because these platforms rely on widely used standards instead of proprietary code or formats, you can utilize a variety of hardware, change suppliers when prices or lead times change or for a second source, and add on new sensors or AI features with reduced effort. Open ecosystems also benefit from the global open source community that finds bugs, adds improvements, and shares best practices leveraging the efforts to benefit everyone. Most importantly, open platforms help deliver flexibility when a new use case appears or new capabilities appear, from real-time language translation on a production floor to multimodal AI workload adoption. The same platform can handle it without new capital spending on new isolated systems or painful retraining of AI models. Open platforms protect today's investments while enabling a smoother path for future enhancements.

- **Invest in automation and management.** A thousand edge nodes is only manageable if software deployment, AI model health checks, and remote updates can happen with minimal human intervention. Favor platforms that treat over-the-air updates the same way that computers receive patches: staged rollouts, automatic verification, and a one-click revert if something looks wrong. Automated edge AI performance monitoring with clear, business-level alerts rather than raw data dumps lets small teams oversee large, distributed edge infrastructure without fire drills.

- **Prioritize security and reliability.** Security and reliability should be at the center of every edge AI plan because these systems often run in unattended locations and make split-second decisions that impact safety, quality, and customer trust. Build around devices that start up only with verified hardware and software, encrypt data in motion and at rest, and can automatically revert to a good state if an update fails. Have a clear monitoring and alerting foundation so that small issues are spotted and alerted before they become major issues. When teams treat security features and uptime safeguards as “must-haves,” they protect sensitive data, meet regulatory requirements, and keep operations running smoothly even when networks are unreliable or threats evolve. In the long run, this peace of mind becomes an accelerator for organizations to scale their edge AI without fear of hidden risks or unexpected downtime.
- **Bridge skills gaps with ecosystem support.** Utilize the growing library of open source models, step-by-step tutorials, and low-code tools that let domain experts in maintenance, quality, or other operations to use and contribute directly. Provide internal teams with vendor academies and university educational partnerships that teach edge AI deployment in practical workshops. Spreading know-how across the organization and leaning on community resources reduce dependence on costly specialists and build a learning culture that can shift quickly as needs change.

Together, these practices will accelerate edge AI from a proof of concept into a resilient, scalable capability that will grow with the company or organization, adapt to new requirements, and deliver measurable long-term returns.

# Conclusion

## A system-level approach is the key to unlocking the full value of edge AI — delivering operational efficiency, flexibility, and resilience in real-world deployments.

Edge AI delivers its greatest returns when it is merged into everyday operations, rather than being deployed as an isolated edge AI function. A system-level approach that treats every node as part of a unified, multifunction platform focuses on concrete business outcomes. By running AI side by side with real-time control, monitoring, and business programs, companies eliminate the complexity and cost of data transfers to and from the cloud and can benefit from the focus on real-time reaction, which can, for example, extend the use of existing assets without a significant disruption on operations.

Open source models, software tooling, and standardized data formats simplify hardware changes, the addition of new sensors, or upgrades to the next generation of models. Automation and deployment management keeps large quantities of nodes patched, monitored, and compliant, while end-to-end security provides protection for both intellectual property and customer data.

As energy prices rise and sustainability mandates tighten, the gains from edge AI become strategic. Enterprises that adopt edge AI where appropriate with strategic planning focused on open, system-level architectures will be ready for the new workloads emerging for the edge, from multimodal perception to low-power language assistants and physical AI, without another round of capital equipment spending and upgrades. Treating the edge as a cohesive system to integrate into the larger organization will accelerate the use of edge AI and the promise of increased productivity and functionality.

# About the IDC Analyst

**Nina Turner**
**Research Director,**
**Semiconductor Technology Supply Chain Intelligence;**
**Enabling Technologies: Edge AI Semiconductors, IDC**

Nina Turner is a research director for Edge AI Processor Architectures at IDC. Her research focuses on semiconductor requirements for edge AI and infrastructure workloads, leading forecast models and market analysis across automotive, energy, transportation, smart city, and IoT. She also oversees IDC's broader semiconductor market models and forecasts, delivering market intelligence and supply chain insights to clients.

More about Nina Turner

## IDC Custom Solutions



IDC

IDC Research, Inc.
140 Kendrick Street, Building B, Needham, MA 02494, USA
T +1 508 872 8200

idc.com    @idc    @idc

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives.