2026 Edge Computing Pivot: Privacy, Control, and Latency

Key Trends Driving the Future of Enterprise Software

Author: Jim Lundy

Video Producer: Adam Pease

Topic: Cloud and Edge Computing

Issue: How will enterprises leverage AI applications in cloud and on-premise (edge) environments?

May 4, 2026 | Research Note 2026-10

SUMMARY

For the last decade, the enterprise mantra has been "Cloud-First." However, in 2026, the cracks in centralized SaaS models have become impossible to ignore. Rising AI costs, increasing hyperscaler outages, and the constant threat of IP leakage have triggered a massive shift toward Edge Computing.

This Research Note explores the emergence of the AI Data Center at the Edge: ruggedized, small-form-factor infrastructure that brings high-density compute power directly to the source of data. While SaaS remains a staple for administrative tasks, mission-critical AI applications are moving to the edge to achieve 10,000x efficiency gains and true data sovereignty.

RESEARCH NOTE

Number: 2026-10
May 4, 2026

TABLE OF CONTENTS

Introduction

As the SaaS Era winds down, Edge Computing is emerging to transform the way enterprises run and manage their AI applications. This shift is made possible by the Neural Processing Units and Small Language Models that can now be specifically optimized for the power and memory constraints of local hardware. Enterprise needs are driving this shift, too. Proliferating cloud costs for responses needed in milliseconds, as well as growing concerns surrounding governance of critical data, are driving enterprises to turn away from the cloud. With Edge Computing, AI investments will go further thanks to improved latency, lower costs, and increased data sovereignty and privacy.

While the cloud will still be useful for smaller workloads and tasks thanks to its processing power, the bottom line is that corporate computing is being restructured from the ground up. The centralized control model of yesterday is no longer; Edge Computing now offers a model of distributed autonomy. To remain competitive, enterprises must migrate their mission-critical AI applications to the edge. This research note provides an overview of this transformative shift as well as guidance on how enterprises can get started with Edge Computing.

The SaaS Fatigue and the Edge Awakening

Over the last few years, especially, enterprises have raced to adopt AI, which led to explosive growth, but also, simultaneously, explosive SaaS constraints. Cloud latency is only getting slower due to the millions of applications running and limited hyperscaler options, not to mention outages. Enterprises have become disenchanted with the increasing data egress costs needed to run their applications, only to get unsatisfactory results. The status quo has become unsustainable.

Moving AI processes to localized devices enables real-time, offline inferencing. With Edge Computing, PCs turn into high-performance edge devices capable of running specialized AI models locally. The result is a significant reduction in costs, a decrease in response time, and increased data sovereignty and privacy. Let's dive into more of these benefits in depth by exploring the trade-offs of SaaS.

The Trade-Offs of the SaaS Era: Privacy, Control, and Latency

The SaaS model of managing AI applications has been popular due to its sheer processing power, as well as the consistency, ease, and expertise that come with using a hyperscaler. Enterprises don't have to worry about running or managing a data center on premises when they can rely on SaaS; it's convenient. However, with that comes the trade-off of putting your total and complete trust in the vendor, and sometimes, this does not meet stringent regulatory requirements. Increasingly, this no longer meets requirements by an enterprise's government, especially if they are using a foreign hyperscaler.

Today, hyperscalers are at more risk than ever of cybersecurity attacks. And, even with their robust security protocols and dedicated teams in place, the centralized model of SaaS inherently

allows more opportunities for sensitive data to be leaked by the very nature of how it is processed and transmitted.

The same goes for latency. The physical distance between a user and the data center hosting an application directly impacts latency, or delay, which in turn affects the user experience. Enterprises have found that tasks needing responses in milliseconds are too costly to run in the cloud. Today, applications are competing with each other for speed, and data costs are quickly spiraling out of control. Enterprises are throwing money at the cloud only to get lackluster results.

The Rise of the AI Data Center at the Edge

Enterprises are seeking secure, sovereign AI capacity to ensure economic and technological independence, especially as governments grow stricter about data. Many will increasingly require information to remain within specific jurisdictions or borders. An enterprise's regional presence is also necessary for delivering responsive applications. This is where a new kind of AI data center will come into play.

Running critical data and real-time processes on localized devices at the edge of the network offers a few significant advantages to enterprises. By moving to the edge, enterprises decrease the risk of sensitive data being leaked or intercepted. Keeping applications on-premise also allows enterprises to directly govern their data, ensuring they meet local, industry, and/or governmental compliance and regulatory standards. And, by processing data at the point of origin, these systems enable real-time inferencing and decision-making without the need for constant internet connectivity, allowing for reliable speed. Thanks to advances in AI, the cost to run and scale AI operations will also decrease. We will dive into this more specifically on page five of the report, in the section, *The Economic Reality: Scaling AI Without Breaking the Bank.*

Security in the Crosshairs: Why Hyperscaler Vulnerability Is Your Risk

Hyperscalers are more vulnerable than ever to breaches and cybersecurity threats, in part due to how dependent enterprises are on the cloud. This vulnerability is also exacerbated by the weaponization of autonomous, agentic AI. Massive amounts of data are being transmitted 24/7, and the processes they are tied to are growing increasingly complex. Even with the most robust security protocols in place and real-time patching, there is ample opportunity to intercept sensitive information without detection.

Outages are also on the rise, which can render an enterprise's entire infrastructure inaccessible until the problem is rectified. Running critical AI processes in the cloud means an enterprise is entirely dependent on a hyperscaler and is subject to the bottlenecks, outages, and security breaches that affect the entire platform. Moreover, relying on a foreign hyperscaler for AI infrastructure means an enterprise is also ceding control of its most valuable asset: its sensitive data, which is then subject to infrastructure governed by foreign laws. This level of risk means strategic information could potentially be exposed to foreign government surveillance and legal orders, fundamentally undermining intellectual property, national security, and digital autonomy.

As more enterprises seek to adopt new AI applications into their everyday processes, such as meetings, these problems will only worsen. We've already reached the tipping point. Enterprises

are frustrated by the constraints and vulnerabilities that exist when using the cloud. Using Edge Computing offers a way to mitigate many of these concerns and enables enterprises to take charge of their security by using localized intelligence, rather than acting as passive participants.

The Economic Reality: Scaling AI Without Breaking the Bank

As enterprises turn to Edge Computing, they might be wondering how to scale their AI operations while avoiding driving up costs, especially as data center bills increase. The good news is that the cost to run AI is decreasing overall, thanks to advances in technology. Large Language Models and Small Language Models will run on smaller hardware footprints thanks to AISCs, or AI-Specific Integrated Circuits. AISCs are a specialized class of semiconductors, such as Application-Specific Integrated Circuits (ASICs) and Tensor Processing Units (TPUs), designed exclusively to execute artificial intelligence workloads with maximum efficiency.

AISCs are hard-coded to perform specific calculations, and this narrowed focus eliminates the hardware overhead of non-essential functions and will also increase energy efficiency, both of which will drive costs down. GPUs will still be used for experimental research and flexible training, but this shift to ASICs for production will help enterprises scale AI without being burdened with the massive power draw and cooling requirements associated with legacy GPU clusters.

For intensive applications that demand efficiency and adaptability at the Edge, such as deep learning applications, Field Programmable Gate Arrays (FPGAs) will become an integral part of an enterprise's hardware strategy. Unlike GPUs, which are powerful but consume significant energy, FPGAs are known for their power efficiency and low latency. Their key advantage, however, is their inherent re-programmability. An FPGA is a silicon chip that can be reconfigured for new applications or optimized for specific tasks after it has been manufactured. This capability means FPGAs offer significant long-term cost advantages and adaptability for the rapidly evolving field of AI.

Defining the Edge-Native Infrastructure: Key Elements

Enterprises should keep the following key elements in mind when building out their Edge Computing infrastructure to support their real-time processing needs:

1. **Advanced Hardware:** Using FPGAs for deep learning applications at the Edge will drive costs down. Their ability to be reconfigured for specific deep learning tasks makes them a versatile and cost-effective solution for localized AI. When designing custom chips, enterprises should focus on performance-per-watt, which will lower costs while also maintaining efficiency.
2. **Efficient Cooling Solutions:** Leveraging liquid cooling vs. traditional air conditioning will reduce energy costs while also maintaining ideal temperatures for applications and devices at the Edge to operate efficiently.
3. **Renewable Energy:** Enterprises must consider integrating renewable energy as part of a sustainable AI strategy. By powering a data center with sources like solar, wind, and geothermal, organizations can significantly reduce their carbon footprint and even meet ESG mandates.

4. **Intelligent Workload Management:** Part of optimizing Edge computing comes down to using the right software that maximizes efficiency and productivity and minimizes idle time and wasted energy.

Trends Driving the Shift from Centralized to Distributed AI

Erosion of Hyperscaler Dominance:

The extreme concentration of AI infrastructure among a few cloud giants has created significant resilience and competition risks, prompting enterprises to seek "neocloud" alternatives and on-premises solutions to avoid vendor lock-in. For years, the reliance on a handful of providers was seen as a convenience; however, in 2026, this concentration is viewed as a single point of failure. Furthermore, most hyperscalers have not yet fully migrated their legacy data centers to be AI-native. This gap in infrastructure capability means that a standard public cloud instance may not offer the specialized cooling or power density required for high-intensity generative AI workloads.

By adopting a hybrid cloud strategy, enterprises gain greater agency over where to run their applications–whether in the public cloud, a private cloud, or a locally managed private data center. This move is accelerated by the rise of what Aragon calls "AI Factories" – data centers specifically architected with high-density GPU and NPU clusters designed for the "agentic era." The implication is that enterprises are no longer tethered to the slow upgrade cycles of traditional hyperscalers. Instead, they can deploy their most sensitive AI workloads in localized AI Factories that offer superior performance and predictable cost structures. This shift ensures that organizations can maintain operational continuity and leverage the most advanced hardware without being sidelined by the capacity constraints or "noisy neighbor" issues of centralized global clusters.

The figure below shows the massive shift of data centers toward AI Factories. Enterprises need to understand this shift and how they can leverage Edge AI.

Total Data Center Spending (Traditional vs AI) 2025-2031

Category \ Year	2025	2026	2027	2028	2029	2030	2031
Total Data Center	384,000	426,240	481,651	545,229	618,835	699,284	790,191
Legacy Data Center	332,544	353,779	379,541	402,924	422,664	434,954	441,717
AI Factory	51,456	72,461	102,110	142,305	196,171	264,329	348,474

Table 1: Total Data Center vs AI Factory Spending 2025-2031.

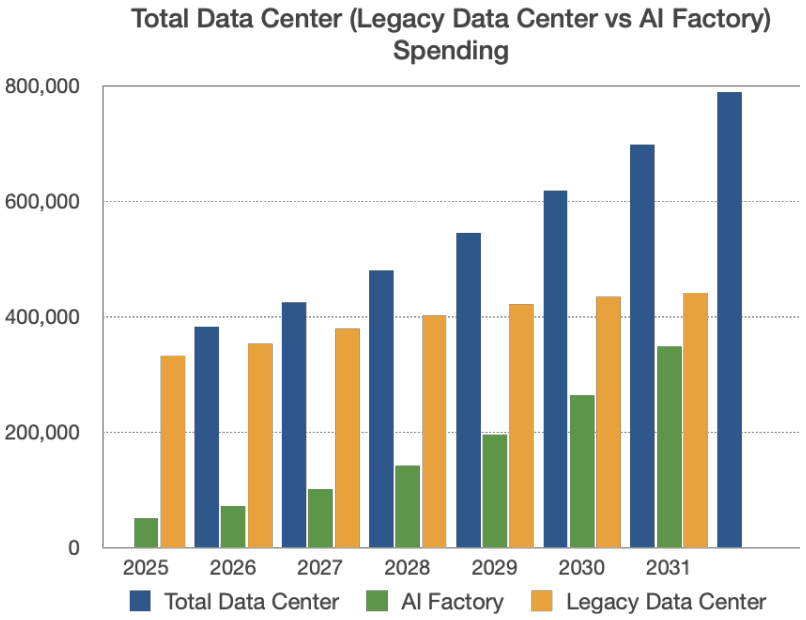

Figure 1: Total Data Center vs AI Factory Spending Globally (2025-2031).

Data Sovereignty and Regulatory Pressures:

Stringent global regulations such as the EU's GDPR and new data localization laws in APAC are mandating that data be stored and processed within specific jurisdictions, making centralized global clusters a compliance liability.

The need for data sovereignty is not new – large enterprises have used this strategy for decades. What is new is the need for privacy and security, so the provider community will need to respond.

Global regulatory environments are becoming increasingly fragmented, mandating that data be processed and stored within specific jurisdictions. Stringent frameworks like the EU’s GDPR and emerging data localization laws across the APAC region have transformed centralized global cloud clusters into significant compliance liabilities. While the concept of data sovereignty is a long-standing strategy for large enterprises, the modern imperative is driven by the acute need for advanced privacy and security in the age of generative AI. This shift is forcing a departure from the "borderless" cloud model toward localized edge architectures that ensure sensitive intellectual property and personal data never cross international boundaries.

The implication for the technology market is a fundamental restructuring of how software is delivered. The provider community must respond by shifting away from one-size-fits-all hyperscale deployments toward modular, edge-native solutions that can be hosted within a clients specific geographic or physical perimeter. For the enterprise, this transition means that regulatory compliance is no longer just a legal hurdle but a driver for architectural resilience. By adopting edge-based sovereignty, organizations can mitigate the risk of massive "blast radius" breaches associated with central cloud providers while ensuring they remain ahead of tightening global data protection mandates.

The Transition to Inference Economics:

As AI models migrate from the intensive training phase toward large-scale enterprise deployment, the focus has shifted from raw compute power to inference economics. By 2026, inference is projected to account for two-thirds of all AI computing power, creating a financial and operational imperative for organizations to decentralize their infrastructure. Relying solely on centralized cloud models for high-frequency queries is becoming cost-prohibitive and introduces significant latency. Consequently, enterprises are moving toward hybrid architectures where processing occurs at the edge, closer to the data source. This shift allows for millisecond response times and a drastic reduction in data egress costs, ensuring that AI agents can operate autonomously and sustainably in real-time environments.

This movement is further accelerated by the introduction of specialized hardware, such as Google's brand-new TPU v8 series. The eighth-generation lineup introduces a dual-processor strategy specifically designed to optimize the "agentic era": the TPU 8t for massive-scale training and the TPU 8i for high-efficiency inference. By utilizing the inference-optimized TPU 8i, which features significantly expanded on-chip SRAM to host larger memory caches, organizations can eliminate the common "memory wall" that plagues traditional processors. The implication of this hardware specialization is a 10,000x gain in operational efficiency, enabling enterprises to deploy complex multi-agent systems at a fraction of the cost and energy required by general-purpose GPUs.

Rise of Domain-Specific and "Tiny" Models

The enterprise AI landscape is shifting away from monolithic, general-purpose models in favor of Domain-Specific Language Models (DSLMs) and TinyML. This trend is driven by the need for higher precision, lower operational costs, and the ability to process data on resource-constrained local devices such as smartphones, sensors, and workstations. While massive models offer broad capabilities, they often lack the deep, nuanced understanding required for professional workflows in specialized fields. By 2026, the proliferation of these smaller, fine-tuned models will challenge the necessity of "one-size-fits-all" AI, as organizations realize that highly optimized models can outperform generalists in specific tasks while requiring significantly less compute.

A primary example of this shift is the emergence of models like Google Gemma 4, which are specifically designed for high-performance local execution. Unlike their predecessor models that required massive cloud clusters, Gemma 4 is optimized to run efficiently on standard enterprise hardware, including desktops and Mac Minis. This capability allows organizations to deploy sophisticated AI agents directly onto employee workstations without sacrificing speed or security. The implication of this development is a radical change in the "cost-per-task" equation; by running intelligence locally on existing hardware, enterprises can eliminate recurring API costs and ensure that sensitive intellectual property never leaves the local machine. This democratization of high-performance AI ensures that even the most data-sensitive industries can scale digital labor at the edge with total architectural control.

Escalating Compute and Energy Constraints:

The exponential growth of artificial intelligence is directly linked to a massive and accelerating demand for electrical power. This soaring energy consumption is emerging as a primary bottleneck to scaling AI, presenting significant economic and environmental challenges. Addressing this energy imperative through efficiency and renewable sources is not just a matter of corporate responsibility; it is a prerequisite for building a viable and sustainable AI future.

Training and operating large-scale AI models is an exceptionally energy-intensive process. Training a single large language model can consume gigawatt-hours of electricity, equivalent to the annual energy consumption of thousands of homes. This is because AI Factories must run thousands of power-hungry GPUs at near-maximum capacity for weeks or months at a time.

Furthermore, the inference phase, when the model is actively used to answer queries or generate content, also contributes a significant and continuous energy load. As models grow larger and their use becomes more widespread, this energy footprint is set to expand dramatically.

The Critical Role of Renewable Energy

Even the most efficient facility will still require a vast amount of electricity. Integrating renewable energy is, therefore, a critical component of a sustainable AI strategy. By powering AI Factories with sources like solar, wind, and geothermal, organizations can significantly reduce their carbon footprint and meet ESG (Environmental, Social, and Governance) mandates.

Beyond the environmental benefits, this is also a sound economic decision. Co-locating AI Factories with renewable energy sources can provide stable, predictable, and often lower long-term energy costs, insulating operations from the price volatility of fossil fuels. This strategic approach to power ensures that the AI revolution can continue to scale without being constrained by unsustainable energy costs or environmental impact.

Industry Spotlight: Who Wins at the Edge in 2026?

Organizations across highly regulated and mission-critical sectors stand to benefit most from the pivot to Edge Computing. By moving computing power to the source of data, these industries can bypass the security risks and latency issues inherent in centralized cloud models. This shift is particularly vital for entities that must balance rapid AI-driven decision-making with stringent data sovereignty requirements.

Figure 2: Three Edge Compute examples: Manufacturing, Healthcare, and Smart Cities.

Financial Services: Real-Time Fraud Detection and Regulatory Compliance

The financial services sector has always had a local data center approach to a majority of its applications and workloads. There is a keen interest in maintaining these workloads at the edge to manage escalating data costs and meet tightening global localization laws. Banks and Insurance firms are finding that tasks requiring responses in milliseconds, such as automated fraud detection or high-speed trading analysis, are often too slow to run efficiently in the cloud.

By utilizing specialized AI hardware at the edge, these enterprises can process massive transaction volumes with maximum energy efficiency and lower operational overhead. The move to the edge also mitigates the risk of systemic infrastructure failure caused by massive cloud outages, which can render financial services inaccessible. For the enterprise, this means achieving a higher cost-to-value ratio for AI investments while maintaining the high-intensity regulation requirements that govern the global financial landscape.

Government: Strengthening Data Sovereignty and Public Trust

For government agencies, the move to Edge Computing is driven by the mandate to protect citizen data and maintain service continuity. Centralized cloud models often present a significant risk, as a single vulnerability can expose sensitive personal records or state secrets to unauthorized access.

Hyperscalers have responded to this with ‘Government Cloud’ options. However, these are still shared infrastructure – so they offer the same risks as the commercial Hyperscaler data centers.

By deploying localized edge infrastructure, Government agencies can ensure that data processing remains within specific jurisdictions and adheres to strict local governance laws. This reduces the "blast radius" of potential cyberattacks and minimizes reliance on third-party cloud providers. The implication for government entities is a more resilient digital infrastructure that fosters public trust through enhanced privacy and the ability to deliver essential services even during broader network outages.

Defense: Tactical Intelligence and Field Resiliency

In defense environments, Edge and mobile Computing are a strategic necessity for maintaining operational superiority in disconnected or contested areas. Traditional cloud reliance is often unfeasible for field operations where low latency and high reliability are non-negotiable. By leveraging ruggedized, small-form-factor AI data centers, defense units can process mission-critical intelligence locally and in real time. This enables immediate situational awareness and autonomous decision-making at the tactical edge without requiring a persistent link to a centralized command center.

For the defense and intelligence communities, sovereignty is non-negotiable, as modern warfare is increasingly an information-driven endeavor. Relying on foreign-controlled AI presents an unacceptable risk to national security. A sovereign AI capability is the only way to ensure that mission-critical systems are secure, reliable, and under national command. This is critical for developing and deploying the next generation of defense technologies.

Advanced AI platforms are now central to gaining a strategic advantage. They ingest and analyze vast amounts of data from satellites, drones, and sensor grids to enable the predictive analysis of object and troop movements, giving commanders the ability to anticipate and counter adversary actions, a capability demonstrated by platforms like Palantir Gotham.

Furthermore, the battlefield is being transformed by AI assistants for commanders, which process information at machine speed to provide distilled intelligence and recommend courses of action. This leads directly to real-time agentic defense, where AI agents can be authorized to take immediate defensive measures against cyberattacks or drone swarms. The growing role of robotics and autonomous systems for reconnaissance and logistics further underscores the need for a secure, sovereign AI foundation to control these assets.

Healthcare and Localized Patient Privacy

Privacy and data sovereignty remain the primary drivers for Edge Computing in healthcare. With medical data subject to strict regulatory oversight, the centralized processing of patient information often presents unacceptable interception risks.

Localizing AI at the edge allows for the implementation of advanced technologies, such as intelligent staff call recordings and real-time patient monitoring, while ensuring that data remains

within the physical walls of the facility.

By leveraging intelligent call recordings, for example, hospital managers can better train and educate their employees by having real case studies to evaluate and learn from to improve patient outcomes and communication between different departments, without worrying about where their data is going. AI-powered systems can also monitor patients in rural and remote areas, predict health events before they become critical, and provide vital support for local care providers.

This decentralized approach allows providers to leverage predictive AI to identify health events before they become critical, even in rural areas with limited connectivity. Consequently, healthcare administrators can improve patient outcomes and departmental communication without compromising their data sovereignty or inviting the cybersecurity vulnerabilities associated with global cloud platforms.

Three Scenarios Where Edge Outperforms Cloud

While the cloud will still be important for certain tasks, such as long-term analysis and model training, the Edge now outperforms the cloud in multiple areas. Let's examine three of these below:

1. **Organizational Meetings Intelligence:** One thing most enterprises have in common: they have upwards of hundreds of meetings occurring at their organizations every single day. Valuable information is locked inside these meetings, and many enterprises are turning to intelligent transcription applications to automate meeting transcription. The Edge takes it to the next level by eliminating cloud bottlenecks that slow down the transfer of information. Important phrases, timely information, and action items can be pulled from the meeting transcriptions by AI to deliver insights in real-time. By running it at the Edge, sensitive meeting data is kept secure and will meet the high-intensity regulation requirements of many industries. Enterprises will increasingly turn to leveraging AI at the Edge for their meetings. One provider offering comprehensive,

Note 1: AudioCodes Overview

AudioCodes Ltd. is a global leader in enterprise voice and VoiceAI business solutions. The company helps organizations unlock the full value of voice, transforming every conversation, whether human or AI, into a strategic asset that drives better business outcomes. AudioCodes' portfolio spans voice connectivity, unified communications and contact center integration, and next-generation voice AI applications that enhance collaboration, automate workflows and deliver real-time insights. With over 30 years of global experience and trusted by 65 of the Fortune 100, AudioCodes powers the intelligent enterprise, connecting people, platforms and data to move business forward.

CEO: Shabtai Adlersberg

Product Portfolio

AI Offerings: Meeting Insights Cloud Edition and On-Prem, Voice AI Connect/ Live Hub, Voca CIC, Interaction Insights

Other Offerings: SBC enterprise leader, AudioCodes Live Platform, Microsoft Teams Live services

Hardware Offerings: SBC, Gateways, IP Phones, Meeting Rooms, Routers

Availability: Now

Website: https://www.audiocodes.com

intelligent solutions, especially for regulated industries, is AudioCodes (see Notes 1 and 2).

2. **Secure Field Service Operations in Zero-Connectivity Zones:** Service workers are up against numerous challenges when out in the field. In addition to their own expertise and knowledge, they are often reliant on their hand-held device to troubleshoot issues, get approvals for a work order to move forward, and more. Lags in connectivity can severely hinder their work, and the bottom line is that the cloud is just way too slow and costly to meet their needs.

   With Edge Computing, field service operators are supported by up-to-date information on their device, without relying on constant, or any, in some cases, internet connectivity. Using AI Agents at the edge to monitor sensors on field sites can also help workers identify issues faster than ever before. In rural areas or areas impacted by outages, severe weather, and other unforeseen circumstances where service work is often needed, the Edge strongly outperforms the cloud.

3. **Mission-Critical Industrial Automation:** Enterprises are integrating AI into their machinery to improve the safety and efficacy of their industrial processes, and also to address global labor shortages. Unlike traditional industrial machines that are constrained to pre-programmed, repetitive movements, machinery built with AI utilizes world models to predict plausible outcomes of their actions in unstructured environments. Specialized vertical AI for tasks like welding and assembly is also becoming standard, delivering pre-trained intelligence that can be deployed with minimal custom programming. By using AI at the edge, the sensors and monitors that are part of these machines have no delay in flagging, sharing, and updating data in real-time, which is critical in identifying and rectifying situations where safety could be compromised.

The Phased Approach to Migrating AI to the Edge

Enterprises must recognize what to use the cloud for and what to migrate to the Edge. For long-term trend analysis and model training, the cloud is the winner thanks to its sheer processing power. But for tasks and major workloads that require millisecond responses, migrating to the Edge is a necessity due to its speed and reliability, as well as the security of localized intelligence.

Migrating these tasks and workloads to the Edge requires careful planning, but the time to start is now. Security considerations should be a priority. Enterprises should upgrade their edge fleet with AI-capable hardware and invest in "zero-trust" security architectures to minimize entry points for cyber threats. The trade-off with the Edge is that security is organization-dependent, which enterprises can use to their advantage if they prepare appropriately.

A second major consideration is location. Data should work smarter, not harder. Place inference clusters in close proximity to where enterprise data already lives to minimize latency and egress costs.

How to Get Started: Auditing Your AI Latency and Data Sovereignty

Business leaders should evaluate their AI roadmaps to identify stable, repetitive workloads—such as customer service agents and meeting transcriptions— that are prime candidates for migration. They should begin by auditing their "power-to-token" effectiveness (in other words, power in versus output) to ensure their current AI investments are delivering sustainable and scalable value.

Enterprises should also audit their legacy data centers, which are often incapable of supporting the high-density power and cooling required for modern AI. AI-specific data centers equipped with modern, proficient hardware and liquid cooling will be key to keeping costs down.

As data sovereignty increasingly becomes important, enterprises should also consider investing in infrastructure that supports heterogeneous compute

Note 2: Featured AudioCodes Products

Meeting Insights On-Prem: AudioCodes AI-powered Meeting Insights On-Prem solution is a leading edge computing appliance built for organizations that demand the highest levels of security, privacy, and compliance. It leverages real-time transcription and summarization, AI-driven insights, and automated task management, as well as utilizes and adapts to an organization's custom terminology to deliver meaningful transcriptions.

VoiceAI Connect: AudioCodes' VoiceAI Connect enables the integration of any cognitive voice service and bot framework with any voice or telephony channel, helping organizations build powerful voice bots and live-agent assistants. By enabling natural, intelligent voice journeys across contact centers and business applications, VoiceAI Connect delivers full voice functionality in a secure, scalable, and production-ready architecture. The solution is available as a managed service or through the AudioCodes Live Hub self-service portal, making it easy to deploy, scale, and optimize conversational voice experiences.

architectures to avoid vendor lock-in and ensure long-term scalability.

Aragon Advisory

- **Audit Your AI "Cost-to-Value" Ratio**: Enterprises should immediately evaluate the long-term costs of cloud-based AI inference. Shifting high-frequency workloads to the edge can reduce operational costs by orders of magnitude compared to standard API-driven SaaS models.
- **Prioritize Data Sovereignty Over Convenience**: As hyperscalers face more sophisticated nation-state attacks, enterprises must move their "crown jewel" IP to on-premises or private edge environments to minimize the blast radius of a central cloud breach.
- **Invest in Edge-Native Talent and Hardware**: Shift procurement focus toward specialized AI accelerators and Neural Processing Units (NPUs) that can handle local Small Language Models (SLMs). The competitive advantage in 2026 will belong to those who can process data in milliseconds, not hundreds of milliseconds.

Bottom Line

The era of blind reliance on the cloud is over. While SaaS offers ease of entry, it creates significant long-term risks regarding data security and cost volatility. In 2026, Edge Computing is no longer a niche for IoT; it is the primary engine for secure, cost-effective, and real-time enterprise AI. Organizations that fail to decentralize their AI infrastructure will find themselves trapped by rising vendor fees and vulnerable to the systemic risks of a centralized digital economy.