

Futurum

Futurum Research 2026

Key Issues & Predictions

Last updated on February 10, 2026

© 2026 The Futurum Group. All rights reserved.

Introduction

2026: From AI Experimentation to Operational Excellence

As we enter 2026, the honeymoon phase of AI experimentation has officially ended. We've moved past the "science projects" and are now facing the cold, hard reality of operationalizing these technologies at scale. At Futurum Research, we've been closely tracking this shift. The theme for this year is clear: **execution over hype**.

Five critical pivots are defining the 2026 agenda:

The Power & Cooling Bottleneck: Energy constraints have officially surpassed silicon availability as the primary hurdle for AI expansion. If you can't power it or cool it, you can't scale it.

- **The Rise of Agentic Commerce:** We are moving from "Read-Only AI" (chatbots) to "Read-Write AI"—autonomous agents capable of negotiating and executing transactions across cloud marketplaces.
- **The Shift to Outcome-Based Value:** As AI breaks the link between effort and value, we're seeing a surge in outcome-based pricing models. In many cases, enterprises are no longer willing to pay for "steps"; they are paying for results.
- **Decentralized Buyer Shifts:** Technology purchasing is decentralizing as CMOs, CROs, and operations leaders adopt AI to drive growth, transforming the CIO from primary buyer to enterprise orchestrator.
- **The Foundation of Trust:** Governance, data hygiene, and security are now the primary constraints for expansion. Strong data integrity and security are no longer "back-office" concerns—they are the non-negotiable prerequisites for the applicability of AI and organizational resilience.

This isn't just about technical upgrades; it's a structural rebalancing of cloud strategy, data governance, and talent. This year's winners won't just have the smartest models—they'll have the most resilient, cost-efficient, and reliable architectures to run them.

Here's to shaping what's next, together.



Tiffani Bova

Chief Strategy and Research Officer
The Futurum Group



AI Platforms: Power and Cooling Constraints Become the Primary Scaling Bottleneck

Prediction: By the end of 2026, energy and cooling constraints will surpass silicon availability as the primary bottleneck for AI expansion. We will see several planned AI data center deployments experience delays of six months or more due to power or cooling limitations, while the emergence of carbon-aware AI scheduling will begin shifting non-critical workloads across time zones and energy grids to balance sustainability goals with operational demands.

Why This is Trending:

- **AI Heat Density Outstrips Traditional Cooling:** Modern AI accelerators generate heat levels that make air cooling increasingly impractical at scale. The mandatory shift to liquid cooling is forcing fundamental data center redesigns - a transition requiring significant capital investment, specialized expertise, and lead times measured in years rather than months.
- **Power Demand Exceeds Grid Capacity:** The electricity requirements of advanced AI facilities are reaching scales that can overwhelm local grid capacity. Some planned deployments are being blocked entirely by utilities unable to guarantee supply, while others face multi-year waits for new transmission infrastructure. In response, data center operators and hyperscalers are increasingly moving to behind-the-meter generation - dedicated natural gas plants, on-site solar installations, or direct nuclear Power Purchase Agreements (PPAs) - that bypass grid constraints entirely. This shift from grid dependency to captive power supply is widening the gap between providers that can guarantee capacity and those that cannot.
- **Regional Strategies Diverge Sharply:** The urgency of AI infrastructure buildout is exposing fundamentally different approaches to the energy challenge. China is leveraging its coal-dominated grid - coal still provides about 70% of electricity in eastern data center hubs, according to the IEA's Energy and AI report (2025) - prioritizing speed over sustainability. The United States faces a fragmented grid driving pragmatic embrace of every available source, natural gas plants contracted specifically for AI clusters, aging nuclear facilities being reconsidered, and hyperscalers signing unprecedented long-term PPAs. The Middle East is positioning itself as a global AI infrastructure destination, combining sovereign wealth, abundant energy, and electricity tariffs roughly half of US rates - the Stargate UAE project alone envisions a 5-gigawatt campus in Abu Dhabi. Europe finds itself constrained by climate commitments and fragmented grids; the Nordic countries are emerging as the continent's AI energy oasis, while broader European deployment faces seven-year grid connection timelines in Germany and tension between growth ambitions and EU decarbonization targets.
- **Sustainability Mandates Meet Operational Reality:** Corporate sustainability commitments are colliding with the carbon intensity of AI workloads, prompting new approaches: carbon-aware scheduling that shifts non-urgent processing to renewable-rich windows, geographic load balancing across time zones, and renewable PPAs to offset fossil fuel dependency.

Use Cases:

- **Cooling Infrastructure Requires Upfront Investment:** The shift to liquid cooling is becoming mandatory for enterprises deploying current-generation accelerators on-premises. Organizations that did not factor cooling upgrades into initial budgets are discovering retrofit costs can add 20-30% to total cost of ownership. Power and cooling assessments are becoming prerequisites for any significant AI hardware procurement.
- **Power Availability Becomes a Site Selection Prerequisite:** For enterprises planning private AI infrastructure, power availability has overtaken real estate cost, connectivity, and proximity to headquarters as the primary site selection criterion. In constrained markets, utilities are quoting multi-year lead times, forcing organizations to accept delays, pay premiums for sites with existing capacity, or adopt hybrid strategies across multiple locations.
- **Cloud AI Capacity and Pricing Become Less Predictable:** Enterprises relying on cloud AI services will increasingly encounter downstream effects of provider power constraints, including capacity limits in preferred regions, longer provisioning times, and pricing volatility as providers pass through energy costs. Organizations are responding by negotiating reserved capacity, diversifying across providers and regions, and building contingency plans for capacity-constrained scenarios.
- **Carbon-aware Scheduling Becomes Operationally Viable:** Enterprises with significant batch AI processing now have tooling to shift non-time-sensitive workloads to lower-carbon and lower-cost grid windows, thus reducing electricity spend and reported emissions without impacting latency-sensitive inference.
- **Edge Deployment Offers a Path Around Centralized Power Constraints:** For enterprises facing power limitations at primary data centers, distributing inference to edge locations offers an alternative to waiting years for grid upgrades. The trade-off is increased operational complexity in managing a fleet of smaller deployments rather than a single centralized facility.

"For enterprise IT leaders, power and cooling have moved from infrastructure afterthoughts to strategic constraints that shape every AI deployment decision. In 2026, we're seeing organizations factor cooling retrofits into TCO calculations, negotiate cloud capacity commitments months in advance, and rethink site selection criteria entirely. For technology vendors - whether hyperscalers, infrastructure providers, or other types of AI platform companies - the ability to guarantee power availability, deliver efficient cooling at density, and offer predictable capacity is becoming as important as model performance or feature differentiation. The enterprises that recognized this shift early are deploying while their competitors wait; the vendors that solved for power and cooling first are winning deals that others cannot even bid on."



Nick Patience

Vice President & Practice Lead,
AI Platforms

[Request Analyst Session](#)



Cybersecurity & Resilience: The Blurring Lines of Data Security & Recovery

Prediction: As we move deeper into 2026, the rigid distinctions between Data Security Posture Management (DSPM), Data Loss Prevention (DLP), and Backup/Recovery are beginning to dissolve. While our previous focus was on the agents themselves, the conversation has shifted toward the data they rely on. We expect to see a growing trend of "Data Resilience" convergence, where organizations stop treating data security and data recovery as separate disciplines and start managing them as a single, continuous lifecycle—driven largely by the need to prepare data for broader AI applicability.

Why This is Trending:

- **Interest in AI applicability Drives Data Hygiene:** Organizations are eager to apply AI technology to their processes, but this requires a level of data visibility and cleanliness that siloed tools cannot provide. As businesses explore how to "activate" their data for AI, they are finding that the lines between discovering data (DSPM), protecting it (DLP), and retaining it (Backup) must blur to create a usable, secure data foundation.
- **The Complexity of Fragmented Data Policies:** Managing data risk has become a complex undertaking as data volume explodes. Security leaders are realizing that maintaining separate policy engines, such as one for "don't leak this" (DLP) and another for "save this forever" (Backup), is inefficient. The growing interest in unifying these controls stems from the need to simplify governance across increasingly complex hybrid environments.
- **Storage Technology is Being Applied to Security Outcomes:** On the flip side, traditional storage and backup technologies are increasingly being viewed as well-suited for active security use cases. We are seeing a trend where backup repositories are no longer just passive archives but can support activities such as threat hunting, classification, and compliance scanning, effectively bringing "backup" tools into the security operations center.

Use Cases:

- **Unified Discovery and Classification:** The convergence of these tools allows for a "scan once, use everywhere" approach. Instead of running separate vendors for backup, DLP, and DSPM, organizations can use a single technology stack to identify sensitive assets. This is beneficial for organizations trying to understand their data layout to determine which datasets are safe and ready for AI experimentation.
- **Integrated Risk Remediation:** Blurring the lines allows for more cohesive responses to data risks. For example, if a DSPM tool identifies a misconfigured database containing sensitive PII, it can not only alert the security team but also automatically trigger an immutable backup snapshot to ensure resilience before remediation occurs. This is evidenced by the increasing cross-pollination of features in platforms from multiple vendors.
- **Forensic Preparedness via Backup:** Security teams are increasingly leveraging backup infrastructure for investigations. In scenarios involving ransomware or data corruption, the ability to mount a "clean" historical copy of the environment instantly allows analysts to understand the scope of an incident without disrupting production. This moves backup from a disaster recovery task to a proactive security operations capability.

"Data security has evolved from a technical control to a strategic enabler, serving as one of the key bridges between technology investments and tangible business outcomes. As organizations look to unlock the value of their data for AI and innovation, ensuring that data is available, trustworthy, and recoverable is paramount. By unifying these disciplines, security leaders can move beyond simple risk mitigation and provide the confidence the business needs to move fast."



Fernando Montenegro
Vice President and Practice Lead,
Cybersecurity & Resilience

[Request Analyst Session](#)



Data Intelligence, Analytics, and Infrastructure: The Shift to Composable Trust and Industrialized AI

Prediction: By the end of 2026, the "do-it-all" data platform will fracture into a Composable Intelligence Stack, where the Semantic Layer and Universal Catalog become primary control points for Enterprise AI. As the focus shifts from experimentation to production scale, organizations will prioritize "Data Contracts" and "Intelligent Caching" to enforce reliability and control the exploding costs of generative AI.

Why This is Trending:

This trend is driven by the harsh reality of moving AI from "potential to production." The low-hanging fruit has been picked, and three distinct pressures are forcing a new architectural approach.

- First, the "Science Project" era is over; the "FinOps" era has begun. As AI workloads scale, the cost of redundant LLM queries is becoming unsustainable. We are seeing a pivot toward Intelligent, AI-driven caching and vector/semantic optimization. Infrastructure decisions are no longer just about speed; they are directly linked to profit margins. If you can't lower the "Token Bill" via semantic caching, your AI project dies on the finance desk.
- Second, raw data is dangerous without context. We've learned that pointing a raw LLM at a data warehouse is a recipe for hallucination – even with a model context protocol (MCP) server as an intermediary. This has elevated the Semantic Layer from a nice-to-have to a non-negotiable prerequisite. It serves as a translator, teaching AI what business metrics (like "churn" or "revenue") actually mean, ensuring that an AI agent and a CEO's dashboard speak the same language.
- Third, the monolithic platform is giving way to open standards. No single vendor can be best-in-class at everything. The market is settling on a composable architecture, held together by open standards such as Apache Iceberg and Project Nessie. The battle has moved from the storage format to the Universal Catalog. This serves as the "Git for Data" that manages governance across a fragmented, hybrid landscape.

Use Cases:

- **The FinOps-Aware Semantic Cache:** Instead of hitting an expensive model like GPT-4 for every user query, an intelligent caching layer intercepts the request and serves it from cache. It recognizes the meaning (semantics) of the question (e.g., realizing "What are sales in Brazil?" is semantically identical to a previous query "Brazil revenue stats") and serves the answer from a local vector store. This dramatically reduces latency and slashes inference costs.
- **Data Contracts as CI/CD Gatekeepers:** Moving beyond reactive data quality, engineering teams are implementing Data Contracts that "shift left." A software engineer attempting to merge a code change that alters a table schema will be blocked by the CI/CD pipeline if that change violates the contract defined by downstream data consumers. The idea is to stop "bad data" at the source before it breaks executive dashboards or AI agents.
- **The Data Analyst as AI Shepherd:** The workflow for data professionals is moving rapidly from manual SQL coding to auditing AI outputs. In this use case, an analyst uses a natural language interface to request complex analysis. Their primary task is not writing the code, but reviewing the "Diff View" (i.e., comparing their human intent against the AI-generated logic) to spot hallucinations and validate the narrative before it reaches decision-makers.

"In 2026, we are done with the 'magic' of AI and are now facing the cold, hard mechanics of production. It's no longer enough to have AI-ready data. You need a Semantic Layer to explain your business to the model, and a Universal Catalog to govern it. The winners this year won't be the ones with the smartest models, but the ones with the most reliable, impactful, and cost-efficient data architecture to run them."



Brad Shimmin

Vice President & Practice Lead, Data Intelligence, Analytics, and Infrastructure

[Request Analyst Session](#)



Digital Leadership, CIOs, and Tech Buyer: AI's Operational Reckoning Forces CIOs and Enterprise Tech Buyers to Reset Cloud, Governance, and Buying Models

Prediction: Enterprise technology strategy in 2026 is entering an operational reckoning as CIOs and business technology buyers move AI from experimentation into core execution. Futurum's Q3 2025 CIO Survey shows that 89% of CIOs are focused on AI-driven strategic improvement and 80% prioritize AI as central to business transformation. At the same time, 71% of CIOs are reevaluating where cloud workloads should run, reflecting mounting pressure from AI cost structures, data gravity, security exposure, and decentralized business-led technology buying. As AI expands across functions, enterprises are being forced to reinvent cloud placement, governance frameworks, and operating models to sustain scale.

Why This Is Trending:

- **The Enterprise AI Scaling Reality Check:** AI pilots are widespread, but enterprise-scale execution remains elusive. While 63% of CIOs already apply AI to accelerate processes and customer service, production deployments are exposing weaknesses in data quality, orchestration, integration, and lifecycle governance. Buyers increasingly view 2026 as the inflection point for moving from fragmented experimentation to standardized operating models.
- **Business-Led AI Buying Accelerates Beyond IT:** Technology purchasing continues to decentralize as CMOs, CROs, and operations leaders adopt AI to drive personalization, forecasting, productivity, and growth. This shift is transforming the CIO role from primary buyer to enterprise orchestrator, responsible for enabling speed while preventing platform sprawl, data fragmentation, and security gaps.
- **Data Security and Governance Emerge as the Primary Constraint:** With 80% of CIOs citing data security, privacy, and information leakage as top concerns, governance — not ambition — is now the limiting factor for AI expansion. CISOs and risk leaders are exerting greater influence over buying decisions, slowing deployments that lack explainability, control, and accountability.
- **Cloud Strategy Enters a Structural Rebalancing Phase:** AI economics are reshaping infrastructure decisions. Survey results show 66% of CIOs prioritizing storage and application migration and 51% focusing on cloud modernization. Enterprises are increasingly optimizing workload placement across public, private, and hybrid environments to balance performance, cost predictability, sovereignty, and inference efficiency.

- **Talent Scarcity Becomes a Dominant Buying Filter:** Despite aggressive AI and cloud adoption goals, 90% of CIOs report continued challenges acquiring and retaining skilled IT talent. As a result, CIOs and senior business technology buyers are actively prioritizing solutions that minimize dependence on scarce specialists. Platforms that are easier to deploy, operate, govern, and integrate — requiring fewer highly trained engineers or data scientists — are increasingly favored over technically elegant but operationally complex alternatives.

Use Cases:

- **Enterprise AI Operating Models:** Defining shared governance frameworks that enable business-led AI adoption while maintaining centralized data, security, and integration controls.
- **Cross-Functional AI Deployments:** Expanding AI use cases across marketing, revenue operations, finance, and customer experience with consistent oversight and performance metrics.
- **AI-Aware Cloud Optimization:** Rebalancing training, inference, and data workloads across environments to manage cost, latency, and regulatory exposure.
- **Data Security and Privacy Reinforcement:** Implementing stronger data lineage, access controls, and leakage prevention as AI usage scales.
- **Low-Talent-Dependency Platforms:** Adopting technologies designed to reduce operational burden through automation, embedded governance, and simplified management models.



Dion Hinchcliffe

Vice President & Practice Lead,
Enterprise Technology Buyers

[Request Analyst Session](#)

"2026 marks the moment when enterprise ambition collides with operational reality. AI adoption is no longer the challenge; scaling it safely and economically across dozens of business buyers is. CIOs are becoming orchestrators of platforms and governance, while CMOs, data leaders, and revenue executives drive demand at unprecedented speed. The organizations that win will be those that align decentralized buying with centralized control and choose solutions that reduce talent dependency—turning AI momentum into durable business advantage."



Ecosystems, Channels & Marketplaces: Cloud Marketplaces at the Nexus of Agentic Commerce

Prediction: By the end of 2026, agentic commerce will move from assisted discovery to autonomous procurement, with marketplaces providing the governance layer required for "agent-to-agent" (A2A) deal execution. This transition is underpinned by the rapid growth of marketplace-associated revenue, which surpassed \$21 billion in 2025 and is projected to exceed \$41 billion by 2029.

Why This is Trending:

- **From Conversation to Execution:** The industry is shifting from "Read-Only AI" (chatbots for discovery) to "Read-Write AI" (agents that act). In 2026, commerce is defined by agents that don't just suggest solutions but possess the authority to execute transactions without human clicks.
- **Maturation of Governance Protocols:** The launch of standards such as the Universal Commerce Protocol (UCP) and the Agent Payments Protocol (AP2) provides the "common language" needed for agents to negotiate carts and verify identities securely.
- **Marketplaces as the "Nexus":** As organizations deploy thousands of individual digital tools and agents, cloud marketplaces have become the essential platforms for managing this sprawl. They provide the verified inventory, pricing mechanisms (consumption-based and negotiated), and secure checkout environments that agents require to operate safely. According to a 2025 Futurum Decision Maker survey of 705 technology partners, 40% claimed that their customers frequently purchased solutions from a hyperscaler marketplace.
- **The "Negotiation Game" Goes Digital:** Enterprise buying remains a negotiation game, but it is now being digitized through private offer programs and "agent mode" features that allow buyers to upload RFPs for autonomous matching and deal-making.

Use Cases:

- **Autonomous B2B Procurement:** Organizations like the National Gallery Singapore have implemented "procure-to-pay" systems using Coupa's AI agents to automate the entire lifecycle from RFx builds to invoice matching and payment, claiming to cut cycle times by 50%.
- **Scalable Negotiation:** Startups like nnamu utilize game theory to deploy agents that autonomously define strategies and finalize pricing with hundreds of suppliers simultaneously—a task impossible for human teams.
- **Agentic GTM for Service Providers:** Global System Integrators (GSIs) and ISVs are increasingly using marketplaces as a key route to market for agentic AI, leveraging marketplace billing visibility to support new outcome-based pricing models.

"The road to agentic commerce has truly begun, and cloud marketplaces are the critical nexus of this evolution. By providing the governance and interoperability required for autonomous agents to negotiate and transact at scale, marketplaces are no longer just a procurement option; they are the engine driving the next flywheel of enterprise software usage and service consumption."



Alex Smith

Vice President & Practice Lead
Ecosystems, Channels & Marketplaces

[Request Analyst Session](#)



Enterprise Software & Digital Workflows: Outcome-Based Pricing To Proliferate in Agentic AI Use Cases

Prediction: The use of outcome-based pricing will continue to proliferate throughout 2026, largely for agentic AI use cases in which the resolution or goal is known, but the actual steps and processes used to achieve the goal may vary with each agentic interaction. According to Futurum's IH 2026 Enterprise Software Decision Maker survey, the percentage of respondents indicating they were using an outcome-based pricing model for AI features increased to 22%, up from 18% in 2025. Conversely, respondents indicating they were using a consumption-based model declined to 30% from 36%, according to the previous survey.

Why This is Trending:

- From the customer's perspective, these use cases are a strong fit for outcome-based or resolution-based pricing because they are defined less by how much software is consumed and more by whether a business result is actually achieved. Because these tasks can range from trivial to highly complex, a simple case and a complex escalation may be billed very differently under usage-based models, even though the business value of the resolution may be identical.
- When assessing AI usage, customers often prefer pricing models that incentivize the vendor to resolve issues efficiently, so that the cost of retries, escalations, or system inefficiencies does not flow back to the buyer. An outcome-based pricing approach shifts operational risk from the customer to the provider, which feels fair when vendors claim superior automation, AI, or orchestration capabilities.
- Vendors have sought to demonstrate why their agents are superior to others in the market, and an outcome-based approach provides them with an opportunity to demonstrate their effectiveness. By architecting guardrails and controls around outcome definitions, time frames, lookback periods, and implementing monetization controls (platform fees, overage pricing for out-of-scope scenarios, and renegotiation triggers), vendors can use outcome pricing as a key market differentiator.

Use Cases:

- Tasks whose execution may vary significantly in the number of steps, time to complete, or types or amount of data that must be accessed, are ripe for this pricing approach, including handling customer service or support inquiries, validating service entitlements, deal qualification and pricing, updating internal or external catalogs or databases, or handling order fulfillment exception handling.
- High volume, well-understood workflows with historically predictable resolution patterns, such as first-line customer, employee, or partner inquiries, requests, or workflows.

"Outcome-based pricing is gaining momentum because agentic AI has fundamentally broken the link between effort and value. In many AI-driven use cases, enterprises know exactly what success looks like—a case resolved, an order exception cleared, a deal qualified—but the path to get there is inherently variable. As our latest Futurum data shows, buyers are increasingly unwilling to pay for tokens, steps, or retries when the business value of resolution is the same. Outcome-based pricing shifts that variability and execution risk back to the vendor, where it belongs, and rewards platforms that can consistently deliver results through better orchestration, automation, and intelligence. Over the next year, we expect this model to accelerate as both buyers and vendors recognize it as a more honest, outcome-aligned way to monetize agentic AI."



Keith Kirkpatrick

Vice President and Research Director,
Enterprise Software and Digital
Workflows

[Request Analyst Session](#)



Hybrid Cloud, Infrastructure, and Operations: The AI Platform Provider Boom Will Force Architecture Redesigns, Including On-Premises AI

Prediction: In 2026, the relative costs of the individual components in a complete data center build will change for the first time in years. The boom in AI providers rushing to build massive data centers and contracting years in advance has led to a shortage of both RAM and SSDs, which will keep prices high into 2027. Datacenter, AI infrastructure, and storage system designs that assume SSDs are cost-effective may need to be revisited to favour hybrid solutions with hard disks. An on-premises infrastructure refresh may be too expensive, leading to infrastructure stagnation or increased migration to the public cloud.

Why This is Trending:

- **Large AI Operators have Contracted Years in Advance:** Through the second half of 2025, large AI platform providers announced multi-year, multi-billion-dollar deals to build and operate massive data centers. These announcements often focused on the data center build and power supply contracts. The same providers secured GPU and supporting compute, network, and storage infrastructure to fill those data centers.
- **Building new Chip Fabrication Capacity is Slow and Expensive:** A new semiconductor fabrication plant costs around \$20 billion and takes at least three years to complete. With much of the existing fabrication capacity contracted by the large AI providers, other customers must compete for the limited capacity for the next few years until new capacity comes online.
- **Retail RAM and SSD Prices are Already Climbing:** Even in the first month of 2026, we've seen both RAM and SSD prices more than double. After years of very low prices for both components, the price increase is a shock and will drive changes in purchasing and infrastructure architecture.

Use Cases:

- Vendors with long-term supply contracts will ride out the wait for new fab capacity, relying on those contracts to insulate them from rising RAM and SSD prices. These vendors must ensure

customers know they are insulated from retail market pricing fluctuations. On-premises infrastructure-as-a-service contracts will show their value, as will long-term pricing contracts with public cloud providers.

- Any customer organization with a fixed budget that plans to replace existing server or storage infrastructure is likely to find replacement unaffordable and may choose to extend the life of their existing assets through extended support and maintenance contracts. These companies may find public cloud platforms more cost-effective and migrate more of their on-premises applications to the cloud, freeing up limited on-premises capacity to accommodate expansion of other applications, which would otherwise require purchasing more, now expensive, infrastructure.
- Public cloud providers are not immune to rising costs; customers should expect increases in non-contracted pricing for compute power and storage services over the coming year. Cloud products that rely on excess capacity, such as AWS EC2 spot instances, will likely be less available and more expensive.
- Storage vendors may also be affected by increased component costs, particularly those whose business is based on SSDs, which are 4x the price per GB and 0.1x the price per transaction compared to hard disks. Any vendor without a strong supply contract will need to refocus on hybrid architectures in which SSDs are the performance component and hard drives provide bulk capacity.

"Commodity RAM and SSD prices have risen sharply, and will continue to affect purchasing decisions until the AI wave subsides, or new fab capacity comes online. This year, we will see system architectures in datacenters and cloud adjust to the new balance of component costs."



Alastair Cooke

Research Director, Cloud and Data Center.

[Request Analyst Session](#)



Intelligent Devices: 2026 will see a Significant Acceleration of AI Workloads Expanding to the Edge, from Devices and the IOT to Vehicles and Robots

Prediction: High-performance, power-efficient AI-capable silicon will continue to enable increasingly sophisticated AI use cases at the edge, accelerating the expansion of AI workloads into edge form factors such as devices, vehicles, and robots.

Why This is Trending:

The expansion of AI to the edge is accelerating rapidly, as showcased by CES 2026, which revealed a quickly expanding landscape of new use cases and form factors.

- AI-enabled PCs and mobile handsets are replacing traditional PCs and changing how people work and create on their productivity devices.
- Wearable devices (smart watches, fitness trackers, AI glasses) and hearable devices (smart speakers, earbuds) are integrating context-aware sensing, hands-free assistant features, predictive analysis, and real-time, natural-sounding voice interactions using a mix of on-device and cloud processing.
- Smart appliances like AI-enabled TVs use natural language UIs for search and control, while AI-enabled cameras can identify objects and individuals with precision and increasingly take autonomous action.
- Drones and other semi-autonomous non-passenger vehicles are becoming more context-aware and capable of autonomous action. AI-enabled vehicles are redefining ADAS, performance optimization, maintenance, and in-vehicle experiences through agentic features and contextually aware environmental/entertainment systems that anticipate user needs. The robotics race is increasing investments in physical AI starting in 2026, driven by major technology companies competing in this nascent segment.

Device and Semiconductor Vendors' Commitment to Edge AI Expansion:

Every major semiconductor and device vendor is fully committed to AI's expansion to the edge. Aggressive competition among silicon vendors (Qualcomm, AMD, Intel, Apple, NVIDIA, and MediaTek) is

growing to include NXP, Broadcom, Texas Instruments, Amazon, and Google, especially in IOT, automotive, and robotics.

Partnerships across semiconductor, component, AI platform, and system integrator vendors will be critical in 2026 to consolidate ecosystems around interoperability, performance, and scale.

Use Cases:

- **Moving AI Processing from the Cloud to Devices:** While AI training for LLMs (large language models) remains primarily cloud-based, smaller models can be trained on devices for specialized use cases. The biggest opportunity is edge processing for AI inference, driven by the need for real-time, natural-language interactions and ensuring productivity and UX continuity despite network challenges.
- **Agentic AI at the Edge:** As agentic AI transforms user interaction with apps, AI-capable devices are uniquely positioned to deliver agentic-AI-forward experiences. These experiences promise to save users time, significantly improve productivity, and boost the utility of all AI-enabled devices.
- **Awareness, Context, and Agency:** Devices and robots with the capacity to understand surroundings and context will be increasingly equipped to make better decisions. Examples include semi-autonomous vehicles avoiding collisions, smart cameras granting access, AI glasses translating speech in real time, smart wearables providing hyper-personalized recommendations, and humanoid robots autonomously stopping actions if they might result in an accident.

"While the concept of physical AI tends to be associated with robotics, it's important to think of it more broadly: Physical AI encompasses every category of AI-enabled device that increasingly surrounds us today – our phones, our AI-PCs, our smart speakers, our fitness trackers, AI glasses, AI-enabled cars – increasingly connects us to a digital assistant, an AI agent, an AI-optimized camera or audio experience, or a semi-autonomous feature that enables AI at the edge. This rich interconnected ecosystem of form factors, which is also beginning to incorporate robots, weaves the physical fabric that, as it grows, will enable AI to scale at the edge."



Olivier Blanchard

Research Director & Practice Lead,
Intelligent Devices

[Request Analyst Session](#)



Networking: East-West Traffic Dominates Data Centers

Prediction: AI usage has changed traffic patterns in the enterprise data center. Traditional user-focused flows to servers (north-south) have given way to server-to-server traffic (east-west). By the end of 2026, east-west traffic will account for 90% of all data center traffic flows.

Why This is Trending:

Three primary factors have caused the shift:

- **The AI Multiplier:** AI training workloads, such as LLMs or generative media, transform traffic patterns. Training these models requires GPUs to use all-reduce gradients constantly. This creates traffic patterns that are 24 to 32 times more intensive than cloud applications. When combined with the AI Factory model being championed by providers, the majority of traffic is server-to-server.
- **Microservices and Talkative Apps:** Current software design utilizes microservices much more than previous iterations. A user request triggers hundreds of east-west server calls behind the scenes between databases, authentication servers, and other systems. Every year brings more of these applications, making a wider range of calls, which creates more east-west traffic for every AI agent request.
- **Disaggregated Storage:** AI clusters have minimal storage on the servers themselves. High-speed storage lives in its own space connected via fabric. Server DPUs handle I/O requests, making off-system storage usage transparent to the CPU and GPU. These requests still consume network bandwidth, turning every disk read into east-west traffic.

Use Cases:

The massive traffic increase has forced organizations to rethink their architecture and develop new use cases that optimize hardware to best serve the systems that require priority.

- **No More Three-Tier:** The historical three-tier model (Access -> Distribution -> Core) will only serve legacy clients and enterprises. The traffic route through the core to reach other servers cannot meet the needs of AI clusters. Oversubscription of distribution links creates expensive bottlenecks that can't be eliminated with old thinking.
- **Standardizing on Leaf-Spine:** The hyperscale leaf-spine (Clos) architecture provides the most predictable east-west traffic path in the data center. The spine switches' connectivity means servers are only two hops away from their destinations, making load-balancing traffic much easier. Newer switch designs can scale to 128 ports at 800 GbE to support more GPUs connected to a network fabric while still delivering high-speed throughput.
- **The Rise of Rail Optimized Solutions:** 2026 will see increased use of rail-optimized topologies, specifically designed to address GPUs that use all-reduce gradients. The GPUs in a server are each connected to a different switch in the leaf, which means each GPU has a dedicated path to its partner device. This path design eliminates port congestion and ensures that GPUs are never idle because of the network.

"Networks evolve because of traffic. Cloud computing did not change how traffic flows from user to server. AI is fundamentally different because of East-West communication. Practitioners need to understand how to deploy new designs to utilize hardware efficiently and why old-school thinking will only lead to pain down the road."



Tom Hollingsworth
Research Director, Networking

[Request Analyst Session](#)



Observability: Observability-Native Becomes the Foundation for Agent-Driven Systems

Prediction: By the end of 2026, observability will evolve into an observability-native model that becomes a foundational requirement for operating AI- and agent-driven software systems at scale. Organizations that extend observability beyond infrastructure and application telemetry into agent intent, decision paths, policy evaluation, and action impact will be able to run autonomous systems with trust and accountability. Those that rely solely on traditional observability models will struggle to explain, govern, or safely scale agent-driven work

Why This is Trending:

- **Trust is the Gating Factor for Production AI:** Enterprises will not deploy autonomous systems at scale unless they can prove security, compliance, and intent alignment. Transparency through observability turns AI from experimental projects into production software by making AI behavior visible, secure, auditable, and governable.
- **AI Behavior Must be Observable, not Inferred:** Enterprises can no longer rely on metrics and logs to guess why systems behave the way they do. Agent intent, decision sequences, policy evaluation, and permission use are becoming first-class operational signals that must be generated directly from AI workflows.
- **Observability is Expanding Beyond Operations into Governance:** As agents span development, deployment, and runtime execution, observability is becoming an SDLC-wide system. This shift allows organizations to scale automation with accountability instead of choosing between speed and control.
- **Autonomous Execution Breaks Post-hoc Observability Models:** Agent-driven systems plan, decide, and act continuously across the SDLC. Traditional observability, built for after-the-fact human investigation, cannot explain or govern autonomous behavior operating at machine speed.

Use Cases:

- **Operate Autonomous Agents with Accountability:** As organizations move from AI pilots to production-scale agent ecosystems, accountability becomes the gating factor. Observability-native systems provide traceability across agent intent, decision sequences, and execution outcomes. This allows teams to audit behavior, explain failures, and continuously refine policies without slowing automation or reverting to manual controls.

- **Enable Governed Automation Instead of Opaque Autonomy:** From an enterprise perspective, observability-native platforms allow organizations to align autonomy with risk. Low-impact agents can operate with minimal oversight, while high-impact agents are continuously monitored for policy compliance, permission use, and downstream effects. This enables incremental expansion of autonomy rather than brittle all-or-nothing deployments.
- **Extend Existing Observability Investments into the Agent Era:** Observability-native does not replace existing observability platforms. It extends them. Vendors that integrate agent telemetry, decision context, and policy signals into existing observability stacks allow customers to preserve tools, workflows, and expertise while gaining visibility into autonomous systems. This reduces friction and accelerates adoption compared to standalone AI monitoring tools.
- **Observability Becomes a Competitive Control Layer:** From a vendor perspective, observability-native capabilities are becoming a strategic differentiator. Platform vendors, cloud providers, and observability vendors are converging on the idea that visibility into autonomous execution is inseparable from trust. Vendors that treat AI behavior as opaque or external to observability will force customers to operate blind at precisely the moment when autonomy increases risk. Those that embed observability into agent workflows position themselves as foundational infrastructure for AI-driven software systems.
- **The Shift Sustains Trust at Scale:** As agent-driven development becomes the norm, enterprises will standardize on platforms that provide explainable, governed visibility across autonomous execution. Observability-native systems become the enabling layer that allows observability to remain relevant as software becomes increasingly autonomous. Platforms that cannot participate in this model will be sidelined, regardless of their strength in traditional monitoring.

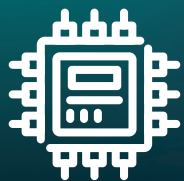
"In 2026, observability becomes the foundation for trusting AI systems that act on their own. Agent-driven systems demand continuous visibility into intent, decisions, permissions, and outcomes as work shifts from human execution to autonomous operation. Observability-native platforms make AI behavior visible, explainable, and governable across the entire lifecycle. This puts observability vendors in a uniquely strong position to become the trust layer for agent-driven software, extending their platforms from operational insight into the control fabric that makes autonomous systems safe to run in production. Without that visibility, autonomous software cannot scale safely."



Mitch Ashley

VP and Practice Lead, Software Lifecycle Engineering, Futurum Research

[Request Analyst Session](#)



Semiconductors, Supply Chain, and Emerging Tech: Token Output Efficiency Becomes the Defining Metric of AI Systems in 2026

Prediction: In 2026, the semiconductor industry will shift decisively from raw compute scaling to efficiency-led system design, with tokens per dollar per watt emerging as the dominant metric shaping AI infrastructure investment. Improvements in memory architecture, rack-scale interconnects, heterogeneous silicon, and software-hardware co-design will materially reduce effective memory pressure and idle compute, enabling long-running, stateful agentic workloads to achieve 10-20x gains in tokens per watt. By the end of 2026, memory shortages will no longer be the primary limiter for AI deployment, and energy constraints will increasingly be addressed through a combination of utilization gains, architectural change, and targeted power expansion.

Why This is Trending:

- **Memory Access Defines AI Efficiency:** As AI shifts from short-form inference to long-running agentic workloads, memory access has become the dominant system constraint. Agentic systems require a persistent state. This state lives primarily in the key-value (KV) cache, stored in GPU high-bandwidth memory (HBM). As reasoning depth and context length increase, KV cache growth accelerates, driving sustained, latency-sensitive memory traffic. In prior generations, GPUs were compute-rich but memory-constrained for inference-heavy workloads, leading to severe underutilization. AI engineers have proven that inference scales with memory capacity and bandwidth first, and compute second.
- **Utilization is the New Performance Frontier:** A critical inefficiency in current AI systems is idle time. During AI inference, GPUs can sit idle for more than 50% of total runtime. Eliminating idle cycles through better orchestration, networking, and silicon design will unlock step-function efficiency gains without proportional increases in silicon area or power. In 2026, the industry will move from monolithic compute density to system-wide scale-out optimization, where keeping accelerators continuously fed with data becomes the primary determinant of performance and cost.
- **CPUs and DPUs as Efficiency Engines:** AI infrastructure is moving away from the GPU doing everything. Data Processing Units (DPUs) increasingly handle networking, storage, security, and protocol processing, eliminating GPU idle time caused by I/O waits and congestion events. At the same time, modern CPUs are evolving from orchestration glue into high-bandwidth participants in reinforcement learning and agentic pipelines, handling environment simulation, scheduling, and control loops that would otherwise starve accelerators. By offloading non-token work and accelerating coordination tasks, CPUs and DPUs materially improve system-level tokens per watt, even without changes to the accelerator itself.

Use Cases:

- **Inference XPU:** Inference XPUs will become the economic engine of AI in 2026 as agentic workloads dominate production usage. Unlike training accelerators optimized for peak FLOPS, inference XPUs are architected around memory capacity, sustained bandwidth, and low-latency access to KV cache. Long-running agents with deep context windows stress memory systems far more than tensor throughput, making HBM-rich designs the primary driver of real-world efficiency.

Higher memory capacity per chip will reduce cache thrashing, increase utilization, and improve tokens per watt. In practice, this enables multi-hour inference sessions for research agents, coding agents, and scientific workflows at costs comparable to today's short-form chatbot interactions. Inference XPUs thus represent the clearest path to monetizing AI at scale without proportional increases in power consumption.
- **RL-optimized Servers:** AI racks will increasingly pair accelerators with high-bandwidth CPUs and DPUs to maximize utilization. CPUs handle environment steps, scheduling, and control logic, while DPUs manage data movement, networking, and storage coordination. This division of labor reduces idle GPU time and allows accelerators to focus exclusively on policy evaluation and learning updates. The result makes large-scale RL practical for agent training, robotics simulation, and autonomous decision systems.
- **Physical AI Accelerators:** Physical AI accelerators increasingly rely on heterogeneous integration, combining compute, memory, I/O, and domain-specific accelerators within tightly coupled packages. High-bandwidth memory remains critical for perception and reasoning workloads, while specialized accelerators handle vision, planning, and control without waking unnecessary portions of the model.

"Exploding token efficiency will turn AI from a demo into a digital workforce. When an agent can think for hours for the current cost of a short chat session and not degrade at scale, AI economics shift in favor of autonomous workloads."



Brendan Burke

Research Director, Semiconductors, Supply Chain, and Emerging Tech

[Request Analyst Session](#)



Software Lifecycle Engineering: Agent Control Planes Power Production Scale AI

Prediction: By the end of 2026, agent control planes will determine whether AI-centered software engineering can move from experimentation into sustained, production-scale execution. Organizations that establish unified control planes for agent identity, permissions, lifecycle, policy enforcement, and execution oversight will be able to deploy agent-driven workflows at scale, while those that do not will remain constrained to isolated or low-trust use cases.

Why This Is Trending:

- **Agent Execution Requires Systemic Coordination:** AI agents are moving beyond assistive roles into continuous execution across planning, build, test, deploy, and operate loops. This execution is parallel, asynchronous, long-running, and interdependent, exceeding the coordination capabilities of traditional CI/CD pipelines and human-paced workflow tools.
- **Fragmented Authority Cannot Support Autonomous Execution:** Existing development and DevOps tooling distribute authority across individual tools, pipelines, and environments. As agents operate across stages and systems, this fragmented model breaks down. Teams either slow execution with manual controls or accept unmanaged automation that erodes trust.
- **Vendors are Racing to Define the Execution Authority for Agent-driven Work:** Vendors are responding by racing to define the control layer for agent execution. Cloud providers, platform vendors, IDE vendors, and startups are all expanding beyond point AI features toward platforms that centrally manage agent identity, permissions, orchestration, and policy. This competition is less about who has the best model and more about who becomes the trusted execution authority for AI-driven work.

Use Cases:

- **AI agents at Scale:** As organizations expand agent use beyond pilots, they reach a point where technical capability is no longer the constraint. Individual agents may perform well in isolation, but production deployment exposes gaps in governance, coordination, and lifecycle control. Agent control planes address this by providing the operational foundation for safely and consistently running AI-driven development.

- **Expand AI Autonomy Without Loss of Control:** From an enterprise perspective, control planes allow organizations to classify and manage agents based on risk and responsibility. Low-risk agents may operate autonomously, while higher-impact agents require approvals, tighter permissions, or continuous monitoring. This enables incremental expansion of AI autonomy rather than all-or-nothing adoption decisions.
- **Capture Customers Without Lock-in:** From a vendor and market perspective, agent control planes are becoming the new battleground for platforms. Vendors that previously competed on tooling depth or developer ergonomics are now repositioning themselves as execution platforms. IDE vendors are extending upward into orchestration and workflow authority. Cloud providers are embedding control planes into their AI and platform services. DevOps and automation vendors are refactoring pipelines to operate at agent speed. The market is converging on the idea that whoever owns the control plane shapes how work is executed, governed, and trusted.
- **The Shift Accelerates Adoption:** As agent-driven development scales, enterprises will standardize on platforms with a single, governed control plane rather than adopt incompatible solutions across tools or vendors. Interoperable platforms that can coordinate agents across environments and integrate with security, compliance, and observability will see faster adoption, while tools that cannot participate in governed execution are sidelined, regardless of technical quality.

"In 2026, vendors are competing on who can be trusted with AI execution. Agent control planes are the layer that turns AI from performing limited tasks with limited trust into a production system of autonomous agents. The vendors that establish themselves offering this authority will define the AI era of software platforms."



Mitch Ashley

VP and Practice Lead, Software Lifecycle Engineering, Futurum Research

[Request Analyst Session](#)

Futurum

Become a Client

