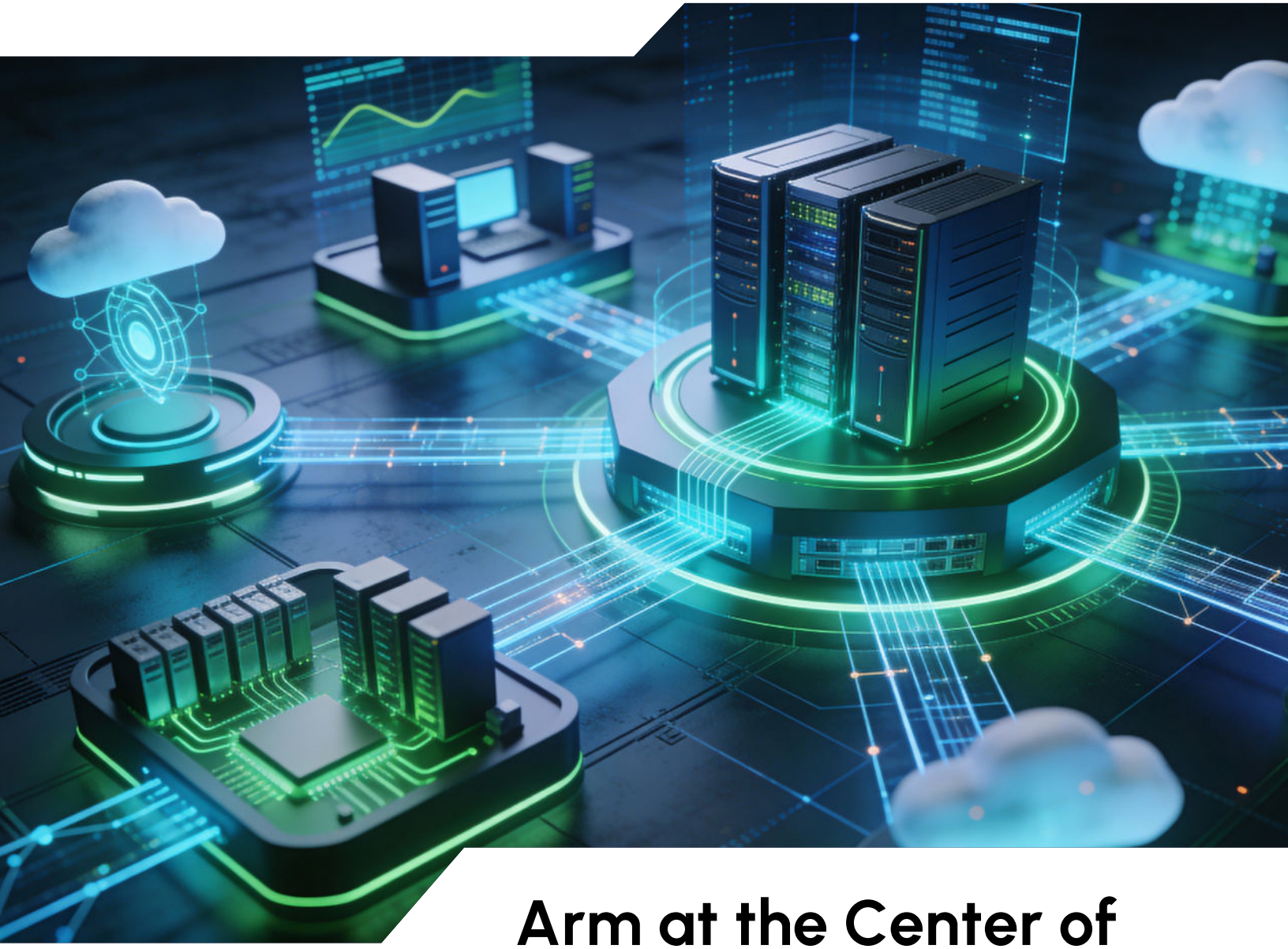
Futurum®

arm



Arm at the Center of the AI & Data Center Revolution

FEBRUARY 2026

As artificial intelligence (AI) features proliferate in enterprise applications, so does the need for the flexibility to build an effective platform for these applications. Arm is uniquely positioned to enable this proliferation, whether the applications run in the public cloud, on-premises data centers, edge locations, or end-user devices. The library of Arm-based processors and associated intellectual property (IP) can be licensed to suit specific needs, from megawatt racks in datacenters to milliwatt IoT devices. The combination of proven CPU and GPU designs with an open partnership model allows system vendors and ODMs to build optimized solutions with a standard set of tools and supporting software. The result is that you'll find Arm technology everywhere AI runs and wherever data is generated or moved to support AI capabilities.

The AI Inflection and Arm's Central Role

AI has created the largest compute inflection point in decades, fundamentally redefining how the world designs, deploys, and scales compute infrastructure. From model training that pushes the boundaries of high-performance computing to inference workloads, AI has become the vital computing workload of our time. Yet AI is not a single workload with a single ideal infrastructure. Instead, AI is a set of workloads that require a cohesive strategy to accommodate diverse requirements cost-effectively and at high performance. AI increases the demand for performance, power efficiency, and scalability across every tier of compute, from cloud to edge, and from data center to device.

As Generative AI transitions from news bites about ChatGPT to production business applications, the need for infrastructure that supports AI wherever it runs has become clear. AI is being built into business applications, whether on-premises or in the Public Cloud, and increasingly into distributed locations such as retail premises, autonomous vehicles, and manufacturing plants. The most visible AI is in the hands of every consumer, on their smartphones and laptops. One common thread running through all these locations is the range of Arm CPU and AI accelerator designs permeating every location where AI runs. Arm has steadily become the indispensable foundation of the modern computing era—whether it's powering the hyperscale data centers, training frontier models, enabling on-device intelligence in smartphones, or accelerating the AI-defined evolution of automotive, robotics, and industrial systems. The Arm architecture forms the connective tissue across the AI continuum. Its ecosystem of silicon partners, system integrators, and software developers has become the bedrock upon which AI experiences are built and delivered at scale.

The AI pipeline is inherently distributed and differentiated. Foundation model training occurs in the cloud, where models ingest massive datasets and demand exascale compute. Fine-tuning and inference, however, are executed not only in cloud clusters but also increasingly in on-premises data centers, edge servers, personal computers, vehicles, and mobile

devices. The computing requirements across different applications require a consistent, power-efficient foundation that scales seamlessly across environments. The Arm architecture uniquely enables this continuum. It delivers high performance and efficiency, workload flexibility, architectural consistency, and an extensive software ecosystem across form factors.

In hyperscale data centers, Arm-based CPUs deliver performance and efficiency at scale. Their superior performance-per-watt allows cloud providers to maximize throughput while minimizing total cost of ownership (TCO), a vital metric as AI models grow in size and training consumes more energy with increasing performance requirements. Meanwhile, the rapidly expanding ecosystem of accelerators and heterogeneous compute designs built to work with Arm CPUs ensures that workloads run optimally, from transformer models to vector inference. SmartNICs and NPUs contain Arm cores to offload functions from the main CPU. Specialized accelerators such as NVIDIA GPUs and Google TPUs are often paired with Arm CPUs to handle general control and data management duties without incurring high cost or power overhead.

On the other end of the spectrum, Arm's low-power heritage has become the driving force of mobile AI and edge intelligence across smartphones, PCs, wearables, XR, and embedded systems. Smartphones powered by the Arm compute platform now deliver generative AI experiences once thought possible only in the cloud: multimodal assistants, real-time translation, and local image generation, all at PC-class performance levels and with unmatched battery efficiency. Arm Mali GPUs deliver AI with power and cost options to meet application requirements, with more than 10 billion shipped to date.

In physical AI applications such as autonomous machines, vehicles and robotics, the Arm platform delivers real-time AI that can sense, decide, and act under the uncompromising power, safety, and reliability constraints these systems demand. As AI scales across physical systems, Arm's power efficiency and a broad, deeply established software ecosystem are decisive advantages. For example in automotive, Arm's designs underpin the compute platforms that enable advanced driver-assistance systems (ADAS), autonomous navigation, and connected in-vehicle AI, all of which require consistent performance and safety.

Arm's ability to power the whole AI spectrum makes it uniquely positioned in this AI era, from hyperscale to handheld. As models grow more capable and ubiquitous, compute capability and efficiency will define the winners of the next decade. This strongly reinforces the advantages of Arm's scalable architecture, power-efficient design philosophy, and comprehensive ecosystem, all underpinned by its full-stack presence from cloud to edge.

AI Compute Grows Up, Everywhere

AI is catalyzing the most profound architectural transformation in modern computing. In recent years, discussions around AI infrastructure have focused primarily on building foundation models and AI accelerators (i.e., GPUs and ASICs). Yet as AI extends from training massive foundation models in the cloud to running inference across billions of connected devices, it is increasingly clear that the future of AI compute will not be defined by accelerators alone, but will require a coherent, flexible architecture to enable AI everywhere.

The next era of AI will hinge on system-level harmony, a balanced architecture that unites performance, efficiency, and scalability across the entire compute stack. The conversation is shifting from "how much raw compute can we deploy" to "how intelligently can we orchestrate compute across diverse requirements and environments at a system level." AI requires more than just compute power; it demands integration. This includes AI accelerators, CPUs, memories, and software that must function as one cohesive platform, scaling efficiently from hyperscale data centers to handheld devices. In this transition, Arm's design philosophy of openness, scalability, and power efficiency has emerged as the common architectural denominator across global AI infrastructure.

1. Cloud Dominance

One great example of Arm's success in this AI revolution is its collaboration with hyperscalers, who are vital to the buildout of AI infrastructure. These world-leading hyperscalers are now rethinking how data centers are built and optimized for AI. This is not a transient experiment but a structural realignment toward architectures that can deliver exponential compute growth without exponential energy consumption. Arm has become central to this realignment.

Public cloud providers are, at their core, the operators of the world's most efficient and scalable computing environments. All of the major cloud providers have embraced Arm as central to their current and future efficiency. AWS Graviton, Azure Cobalt, and Google Axion are all Arm instance types for the best price-performance on public cloud. Yet some of the most significant gains for cloud providers are in using Arm processors in their custom network equipment and hardware acceleration devices. Having licensed IP for Arm CPU designs that can be integrated into their own products is critical to building the exact infrastructure required for each element of a cloud infrastructure.

These developments are reshaping the priorities of hyperscale infrastructure. As AI model complexity, memory demands, and power requirements multiply, hyperscalers are prioritizing architectures that can sustain performance while reducing energy and capital intensity. Arm's modular, efficient design enables this shift by allowing partners to customize silicon for specific workloads while preserving a consistent architectural base across fleets. According to Arm's data, nearly half of the compute shipped to top hyperscalers at the end of 2025 was Arm-based, meeting massive enterprise demand for AI training and inference.

2. Data Center Shifts

At the very beginning of the AI revolution, the goal was to tick a technology checkbox (e.g., an expensive GPU), and there were a few options for deploying AI applications in a data center. As companies realized that AI is a long-term strategic component of their application estate, a more measured, deliberate, and systematic approach has prevailed. A systems approach to infrastructure design leads to integration across technologies—power-efficient CPUs for tasks they can accommodate; higher-performance CPUs and accelerators for cases where they deliver more value. Various Arm CPUs are used where their specific characteristics suit the task: powerful Neoverse CPUs for performance, Cortex CPUs as supporting components, or Ethos NPUs for lightweight AI processing. The catalog of Arm designs covers a wide swath of these diverse system requirements.

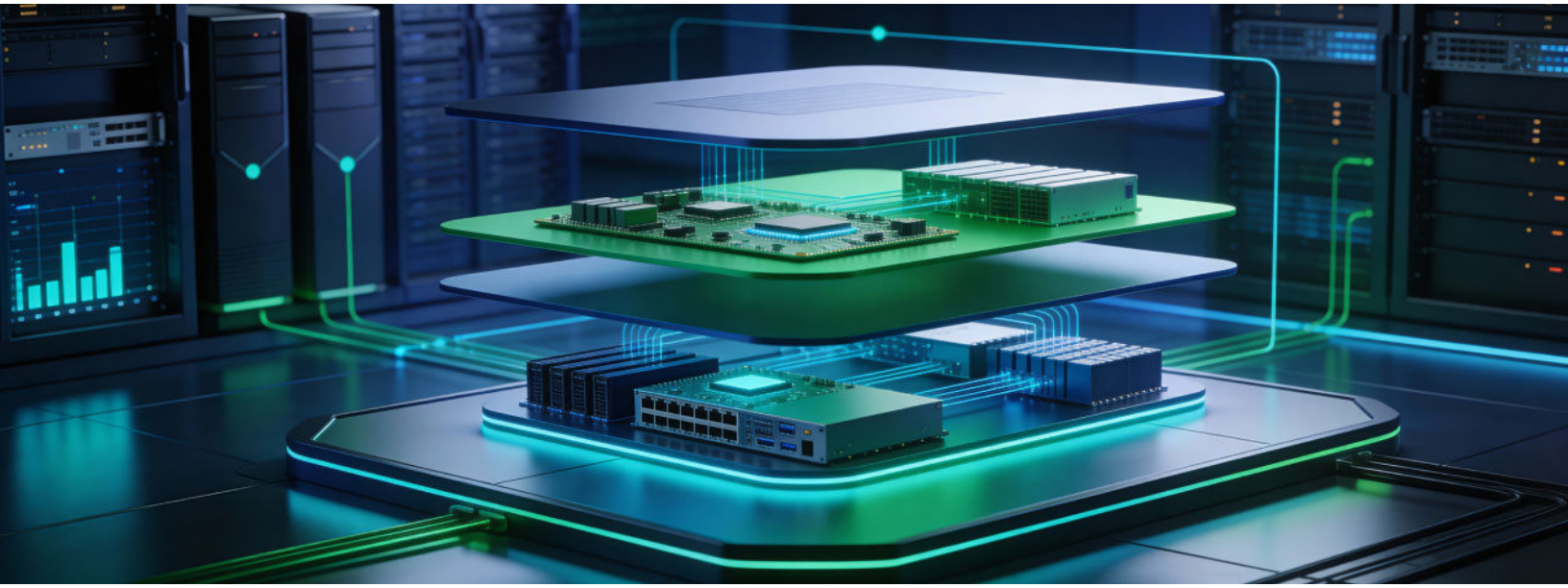
Meanwhile, the world's most crucial AI compute provider, NVIDIA, has been building its offerings on Arm. Its Grace and Vera CPUs, designed to pair with NVIDIA's Rubin and Blackwell GPUs or to operate their Bluefield-4 DPU, are built on the Arm platform, and illustrate how Arm architectures can be seamlessly integrated into heterogeneous systems. This forms the foundation for balanced compute nodes that deliver exceptional efficiency across both AI training and inference workloads.

3. Edge Proliferation: Billions of Arm Devices Becoming AI-Native

While cloud and data centers are clearly priorities for this AI revolution, those are not the only locations. The most profound change in the next few years may be unfolding at the edge, where intelligence meets the physical world, and where Arm's ubiquity makes it the default platform for distributed AI.

Over 99% of global smartphones are powered by the Arm compute platform; these devices are rapidly evolving into AI-native systems. Tasks such as real-time translation, visual recognition, voice synthesis, and generative assistance that once required the cloud now execute locally, leveraging Arm's power-efficient cores and integrated AI engines. This shift enables immersive, private, and responsive AI experiences, all within tight power and thermal budgets.

The same architectural foundation is extending to wearables, AR/VR devices, and a new generation of Arm-based PCs that combine PC-class AI capabilities with mobile-level efficiency. While Arm-based PCs are still in the early stages, their trajectory is clear: enabling on-device generative workloads and personal copilots that operate locally, continuously, efficiently, and securely.

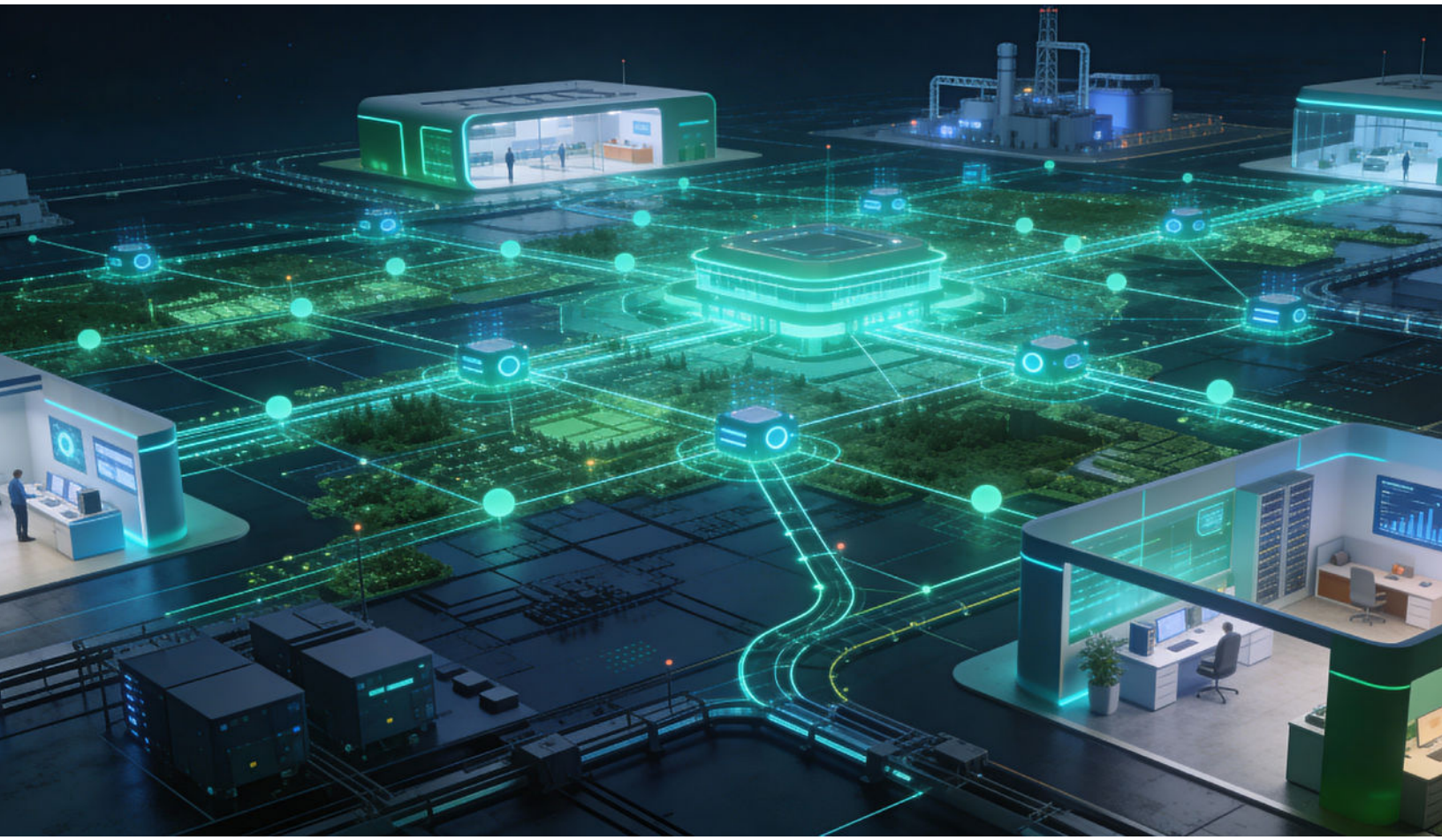


4. Intelligent and Autonomous Systems

Physical AI is a significant part of Arm's ecosystem. As AI reshapes the mobility landscape, nearly 100% of global automakers already deploy Arm technologies across compute domains ranging from microcontrollers and infotainment to centralized ADAS and autonomous driving systems. More than 80% of central compute and ADAS SoCs are Arm-based, and leading companies such as Rivian, Tesla, and XPeng are designing Arm-based custom autonomy stacks that span AI models, software, and silicon. In robotics, a combination of more capable AI models with better and cheaper sensing and compute that can run locally within real-world constraints is driving deployment beyond manufacturing into logistics, healthcare, retail, and service applications. The vast majority of today's robotic systems rely on Arm-based processors for intelligence, control, perception, and system coordination.

What's emerging is not a fragmented AI landscape but rather a more connected continuum of compute, shaped by Arm's consistent architecture and broad ecosystem participation. From hyperscale data centers to edge and endpoint devices, Arm's designs provide a common foundation for efficient and scalable computing across diverse workloads. The company's open licensing model, mature software ecosystem, and scalable design philosophy give developers the flexibility to innovate across multiple domains while maintaining a unified architecture that supports deployment at virtually any scale.

In the evolving AI era, the industry will be reshaped not only by who builds or owns the most powerful chips but also by who creates the most integrated and efficient systems across the spectrum. Arm's role is central in this regard: providing the architectural foundation on which global, distributed, adaptive, and efficient intelligence can operate cohesively across the cloud-to-edge spectrum.



The Business of Being Arm: Multi-Front Strengths

Despite Arm's deep structural presence across the AI value chain, investor sentiment remains divided, often circling back to a familiar question: "Where is Arm's AI monetization?" Much of the market's attention focuses on the visible layers of AI: GPUs that train foundation models or accelerate inference. Easy headlines from extreme performance numbers or massive implementation costs often drive this focus. Yet Arm operates behind the scenes as the underlying architectural substrate that enables this compute ecosystem to function efficiently and cohesively. Arm is a common factor behind the flashy headlines; wherever a GPU needs vast amounts of data, multiple Arm CPUs are in the pipeline to deliver it. Arm's influence is less about competing at the chip level and more about allowing scale and interoperability of compute across the entire stack.

This architectural neutrality is one of Arm's most critical strategic strengths. Unlike chipmakers engaged in zero-sum competition for unit share, Arm monetizes the expansion of compute itself. As AI workloads proliferate across hyperscale, enterprise, and edge environments, each new domain draws upon Arm's IP, instruction sets, and software ecosystem. This dynamic makes Arm's relevance structural rather than cyclical, tied to the long-term growth of global compute demand rather than to specific product cycles or end markets.

Within an increasingly heterogeneous compute landscape, Arm's role is foundational and indispensable. Its licensing model allows partners to innovate independently while maintaining architectural consistency, ensuring interoperability across devices and platforms. This balance of neutrality and ubiquity has created one of the most durable and scalable business models in the semiconductor industry. Arm is not positioned as a competitor within the ecosystem, but as a foundational enabler that underpins it.

As such, Arm's expanding strategic role across the cloud-to-edge continuum is gaining wider recognition. However, the structural, technological, and economic forces that reinforce Arm's position remain underappreciated and sometimes overlooked. In our view, several key factors are now converging to sustain and accelerate Arm's rise in the AI era.

1. Beyond Performance-per-Watt

For much of its history, Arm's positioning has centered on efficiency, delivering superior performance per watt, optimized TCO, and strong price-to-performance. Often, this was characterized as Arm delivering low performance at low cost. Progressive Arm designs added CPUs that deliver higher performance without compromising power efficiency. In the AI era, Arm's advantage has broadened beyond efficiency alone; the architecture now delivers competitive absolute performance and flexibility while preserving its hallmark of scalable energy efficiency.

Benchmark data from Futurum's Signal65 illustrates this shift. In AI and cloud workloads, AWS Graviton4, built on the Arm Neoverse platform, demonstrated double-digit performance gains over comparable x86 CPUs while maintaining a TCO advantage. These findings highlight that Arm's value proposition is no longer limited to "doing it with less." Arm achieves performance parity or outright leadership, while still consuming substantially less power.

This performance profile has tangible strategic implications for hyperscalers. As AI models scale exponentially, performance per watt has become a defining metric for both cost efficiency and sustainability. By enabling higher throughput without a proportional increase in power or cooling requirements, Arm provides a structural advantage in an industry that is increasingly constrained by the physical and economic limits of energy consumption.

2. Merchant + Custom Silicon: Enabling the AI Ecosystem

Arm's flexible IP licensing model has positioned the company as a unifying foundation for both merchant silicon vendors and custom in-house chip developers. Within the merchant ecosystem, firms such as Qualcomm, MediaTek, and Samsung continue to deploy Arm-based processors across mobile, automotive, and edge AI markets. Each vendor is leveraging the architecture's power efficiency and scalability. In parallel, hyperscalers including AWS, Google, Microsoft, and NVIDIA are designing custom processors with Arm CPUs optimized for workload-specific performance, power efficiency, and infrastructure economics.

This dual-track dynamic of broad merchant scale complemented by deep custom innovation enables Arm to expand its total addressable market without engaging in direct vertical competition with its partners. Every chip designed using Arm IP, whether for mass-market devices or proprietary cloud infrastructure, reinforces the architecture's central role in the global compute ecosystem.

From a financial standpoint, this model provides Arm with a structural revenue advantage. The company's growth is tied not to unit volumes from any single vendor, but to the overall proliferation of compute across the industry. As AI workloads scale in hyperscale data centers, the edge, and endpoint devices, Arm's royalty base expands proportionally. In effect, Arm's monetization model is driven by the broad adoption of AI compute, making its performance a function of industry-wide growth rather than competitive market-share shifts.

3. Cloud-to-Edge Continuity

AI's evolution is increasingly distributed. While model training remains concentrated in large-scale data centers, fine-tuning and inference occur across a diverse range of environments, from hyperscale clusters to edge servers and personal devices. This dispersion of compute introduces significant complexity. Maintaining architectural consistency and efficiency across such heterogeneous systems has become a critical challenge in the semiconductor and infrastructure landscape.

Arm is uniquely positioned to address this challenge. Its architecture spans the full spectrum of compute from high-performance Neoverse platforms in data centers to Cortex, Ethos, and Mali GPUs that power edge and mobile devices. This cross-domain consistency allows developers to optimize once and deploy across multiple environments, reducing software fragmentation and accelerating time-to-market.

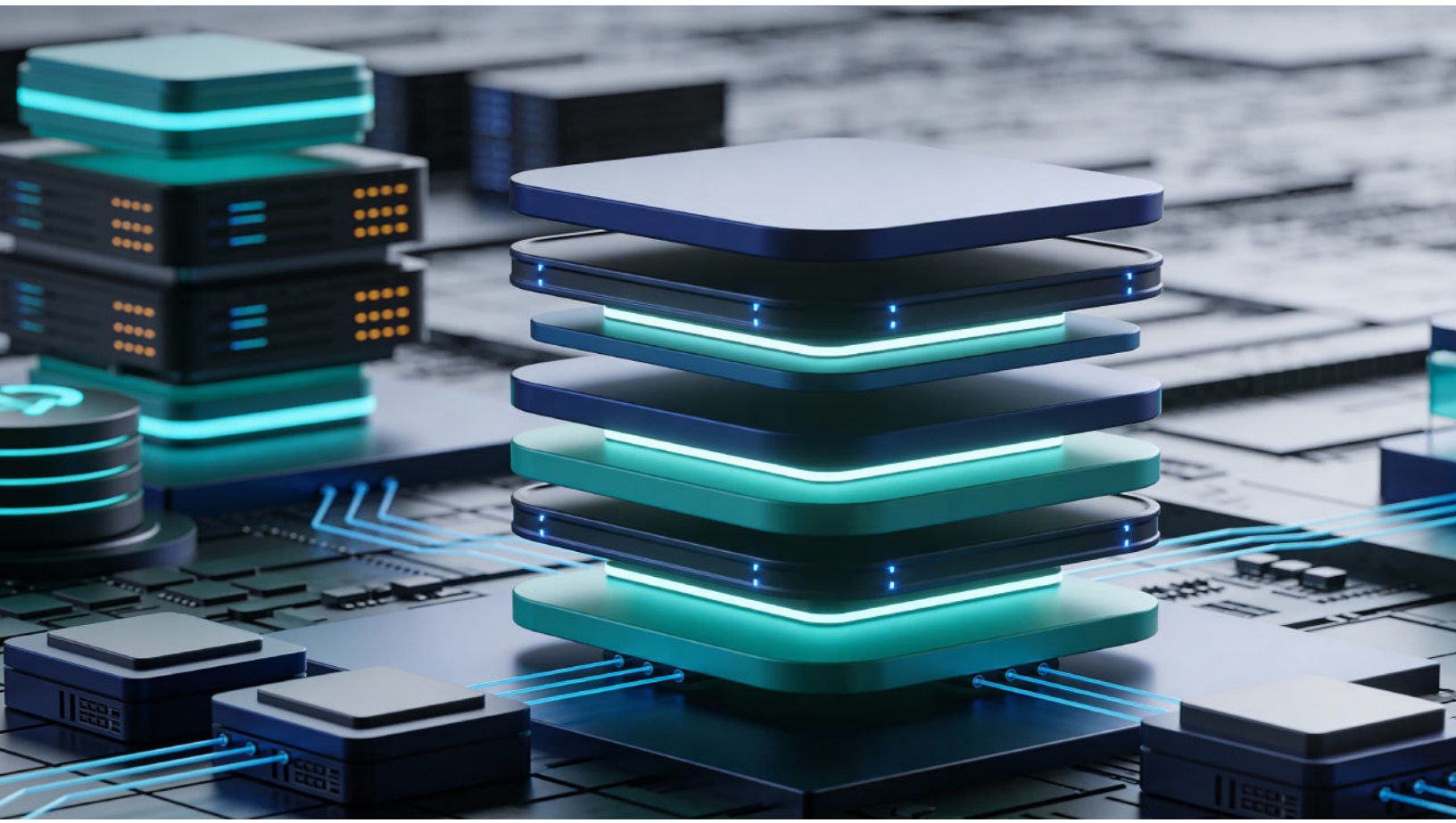
The outcome is a more unified AI continuum with a single architectural framework that underpins compute across all tiers. For hyperscalers, this enables seamless workload portability between cloud and edge environments as customers ask for more distributed AI services. For OEMs and system designers, it provides long-term platform stability, ecosystem continuity, and a consistent developer experience across product generations.

4. System IP, Software Enablement, and Real Switching Costs

In the AI era, hardware alone does not define competitive advantage; software maturity and ecosystem depth are equally decisive. On this front, Arm maintains a structural lead. Its System IP portfolio, spanning interconnects, memory controllers, and coherence fabrics, enables partners to design complete, performance-optimized systems for AI workloads rapidly. Complementing this, Arm's software enablement layer, including compilers, toolchains, and AI frameworks, is deeply integrated across major operating systems such as Linux, Android, and Windows on Arm, ensuring broad developer accessibility and workload portability.

This software and ecosystem maturity has created significant exit costs, protecting long-term business. Arm supports a global developer base of more than 22 million, the largest open ecosystem in computing. Through consistent code portability, optimized runtimes, and cloud-native toolchains, Arm has evolved beyond a CPU architecture into a platform for innovation. Once developers optimize for Arm, migrating to alternative architectures becomes both economically and technically challenging, forming a compounding competitive moat that strengthens over time.

Equally important is Arm's architectural flexibility. Its design framework accommodates both open-source collaboration and proprietary extensions, allowing partners to tailor solutions without sacrificing ecosystem coherence. This combination of flexibility, ecosystem scale, and software maturity positions Arm as the platform for AI-ready compute, bridging the worlds of merchant silicon, custom infrastructure, and edge intelligence.



Laser Focus on Collaboration

While some observers express concern that Arm's expanding role in AI compute could position it in competition with its own customers, this view overlooks the foundation of Arm's long-standing success: collaboration. The company's business model is inherently partner-centric, built around enabling innovation rather than displacing it. Arm works closely with ecosystem leaders, including NVIDIA, Qualcomm, AWS, Google, Microsoft, and a wide range of OEMs and ODMs, to co-develop and optimize solutions across markets. Rather than competing for chip sockets, Arm's strategy is to empower partners to create differentiated value on top of its IP, reinforcing the ecosystem's collective growth and ensuring its own relevance across every layer of compute.

Even the newest high-performance, scale-out storage solutions for AI use Arm CPUs in SmartNICs to connect SSDs to high-speed networks. As AI model sizes grow and larger models appear, Arm's IP is everywhere in the high-speed path. Fast storage, fast networks, and high-performance accelerators such as TPUs and GPUs all require more Arm CPUs to fulfill their roles in the AI pipeline. End customers may not know that they are buying devices with Arm IP, but the vendors and ODMs are all choosing Arm for the openness, adaptability, and efficiency to build differentiated products.

RISC-V, often cited as a potential alternative to Arm, is limited in scope and maturity. Its ecosystem today is fragmented, concentrated mainly in niche or verticalized segments such as microcontrollers, Bluetooth controllers, and other cost-sensitive embedded applications where integration and bill-of-materials efficiency matter more than software or ecosystem scale. Outside of China, commercial adoption remains modest, constrained by immature software stacks, limited development tools, and a relatively small base of active developers.

Notably, many RISC-V vendors have adopted business models similar to Arm's by combining IP licensing with royalty structures. This underscores that the business model's scalability, not the instruction set itself, drives sustainable value creation in the semiconductor ecosystem. However, without a unified software foundation or broad developer engagement, RISC-V currently lacks the horizontal leverage and cross-market interoperability that make Arm structurally indispensable across industries.

As of today, Arm maintains a clear advantage over RISC-V in the AI compute revolution. Its mature ecosystem, extensive software support, and deep integration across cloud, edge, and device tiers deliver performance, scale, and reliability that RISC-V has yet to match. That said, leadership in semiconductors is never permanent. The pace of innovation across architectures, tooling, and open-source collaboration continues to accelerate, and RISC-V's progress, particularly in specialized or cost-sensitive domains, warrants close attention.

Central to AI Infrastructure Buildout

Arm's AI trajectory continues to accelerate across multiple dimensions. Strategic collaborations, such as SoftBank's Project Stargate initiative with NVIDIA, Oracle, and OpenAI, place Arm at the center of the next wave of AI infrastructure build-out. These partnerships reinforce Arm's long-standing architectural neutrality, enabling innovation across the ecosystem rather than competing within it.

Momentum among hyperscalers is also gaining strength. Arm-based CPUs are increasingly serving as the orchestration layer in AI-converged data centers at AWS, Microsoft, Google, and Oracle. Concurrently, the edge represents a parallel growth frontier. As generative AI, inference, and autonomous applications extend from the cloud to billions of devices, spanning smartphones, PCs, vehicles, robotics, and embedded systems, Arm's architecture provides the standard fabric that links all stages of the AI lifecycle.

The AI era will be diverse and heterogeneous, defined by specialized accelerators, hybrid clouds, and distributed intelligence. Yet through this complexity, one constant remains: a unified architectural foundation that connects the cloud to the edge. Arm's combination of performance, efficiency, scalability, and universality positions it to underpin the next decade of global compute expansion and quietly power the AI revolution at the global scale we are witnessing.

Important Information About This Report

AUTHORS

Daniel Newman
CEO & Chief Analyst | The Futurum Group

Alastair Cooke
CTO Advisor | The Futurum Group

PUBLISHER

Futurum Research

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

arm

ABOUT ARM

Arm is the industry's highest-performing and most power-efficient compute platform with unmatched scale that touches 100 percent of the connected global population. To meet the insatiable demand for compute, Arm is delivering advanced solutions that allow the world's leading technology companies to unleash the unprecedented experiences and capabilities of AI. Together with the world's largest computing ecosystem and 22 million software developers, we are building the future of AI on Arm.

Futurum®

ABOUT THE FUTURUM GROUP

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

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