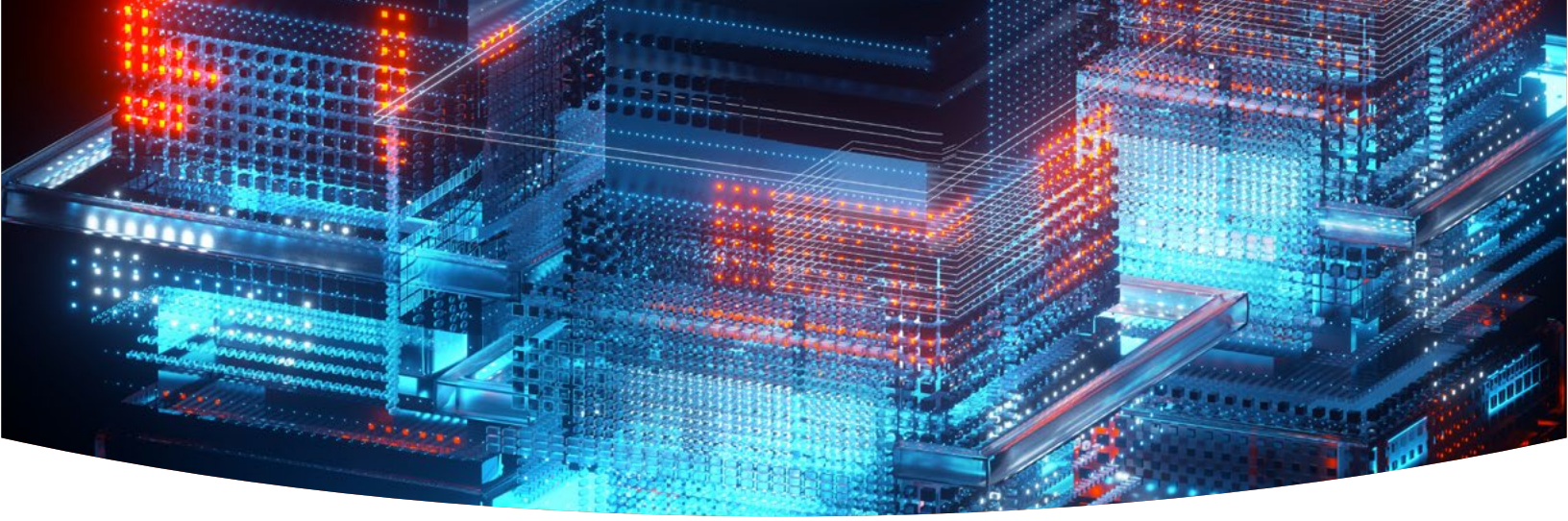


5 strategic decisions for building a scalable compute platform for now and the future

arm

FROM CLOUD TO EDGE: A BLUEPRINT FOR ENTERPRISE AI

“Scalable AI is now a business imperative. And while any company can build out an AI model, fielding it at peak performance across diverse environments is the true design challenge.”

— Mohamad Awad, SVP & GM, Arm Infrastructure Line of Business

Artificial intelligence (AI) is no longer a “nice-to-have” technology—it’s a central driver of competitive advantage and business innovation. Across industries, enterprises are moving beyond experimentation and embedding AI into all their products, workflows, and customer experiences. But as organizations scale, many are discovering a stark reality: their compute infrastructure was not designed for this era of intelligence.

Legacy systems struggle to handle the demands of AI. Latency bottlenecks, due to its architecture, erode real-time insights and experiences. As energy consumption soars and computational complexity increases across data centers and edge environments, stitching together fragmented hardware and software systems is no longer a viable way to meet the growing AI compute demands of modern organizations.

Scaling AI requires more than raw performance; it demands a smarter, future-ready compute stack that can balance efficiency, flexibility, and scalability across cloud, edge, and endpoint deployments.

This whitepaper presents five strategic decisions that enterprise leaders can make to help ensure their infrastructure meets today's AI demands, while remaining adaptable for the future.

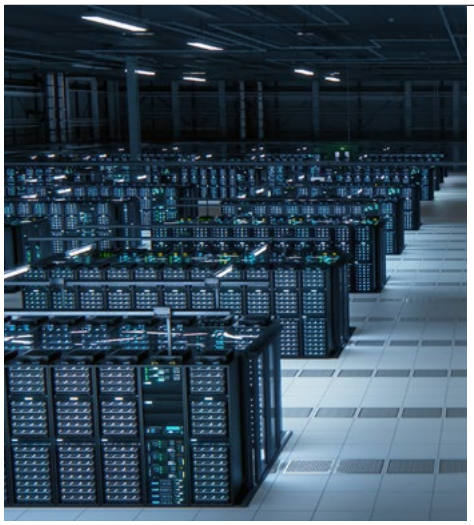
LEGACY INFRASTRUCTURE MEETS ITS LIMITS

The overarching challenge enterprises face is that their infrastructure was never built for AI. Traditional IT environments excel at processing structured data, running transactional applications, and supporting analytics at scale. But AI workloads, from large language models (LLMs) to multimodal systems, operate very differently.

- **Latency becomes critical**: Real-time inference, whether in conversational assistants or predictive maintenance systems, requires response times that legacy stacks cannot deliver.
- **Compute and memory are strained**: Large models with billions of parameters exceed the capacity of conventional systems.
- **Power and thermal constraints bite**: At the edge, power budgets make deploying advanced AI models impractical.
- **Complexity rises sharply**: Disparate environments, tools, and accelerators create silos that increase engineering burden and slow deployment.

"We're seeing organizations across industries hit the same wall," says Mohamad Awad, SVP and GM of the Arm Infrastructure Line of Business. "They've invested heavily in AI models, but the infrastructure to support them isn't keeping up."

The answer is not incremental upgrades; it's fundamentally rethinking how compute should be designed, deployed, and managed.



STRATEGIC DECISION 2: EMBRACE HETEROGENEOUS COMPUTE

As AI workloads diversify, from edge inference to cloud training, performance, efficiency, and responsiveness increasingly depend on the right mix of compute resources:

- **CPUs** orchestrate workloads and provide general-purpose flexibility.
- **GPUs and NPUs** accelerate specialized tasks like inference, computer vision, and training.
- **Domain-specific accelerators** handle emerging workloads in speech, video, or generative AI.

Enterprises must weigh trade-offs carefully:

- **Throughput vs. latency**: High-volume workloads may prioritize throughput, while customer-facing applications demand responsiveness.
- **Performance vs. efficiency**: The most powerful option isn't always the most cost-effective or sustainable.
- **Flexibility vs. simplicity**: Heterogeneity introduces complexity, but it is essential for long-term scalability.

The decision is to embrace heterogeneity from the start, ensuring each workload can run on the right processor at the right time.

STRATEGIC DECISION 3: UNIFY THE SOFTWARE LAYER

Even the most advanced hardware delivers limited value when the software ecosystem is fragmented. Today, many enterprises waste cycles re-optimizing models for each target environment, slowing deployment and increasing costs.

A smarter path is to simplify through unification:

A smarter path is to simplify through unification:

- **Cross-platform libraries** abstract hardware complexity, allowing models to run efficiently on CPUs, NPUs, and GPUs without rewriting.
- **Model optimization tools** enable quantization and pruning, so large models can fit within constrained environments without losing accuracy.
- **Framework integrations** ensure compatibility with popular AI frameworks, eliminating manual translation overhead.

This approach reduces friction for developers, accelerates time to market, and ensures that performance improvements scale across the enterprise.

STRATEGIC DECISION 4: BUILD FOR FUTUREPROOFING

AI evolves at an unprecedented pace. Models, frameworks, and hardware all change rapidly, making rigid or closed systems a recipe for obsolescence. Leaders must prioritize futureproofing their infrastructure.

- **Openness**: Supporting widely adopted standards such as ONNX, MLIR, and PyTorch to maintain interoperability.
- **Ecosystem extensibility**: Engaging with a broad community of silicon providers, hyperscalers, and software vendors to ensure adaptability.
- **Portability**: Enabling workloads to shift seamlessly between cloud, edge, and device as business needs change.

Futureproofing is not a luxury. It is the only way to ensure that today's infrastructure decisions will continue to generate value tomorrow.

STRATEGIC DECISION 5: ALIGN COMPUTE WITH BUSINESS STRATEGY

Compute is no longer just a technical concern; it's a strategic enabler of business outcomes. To maximize impact, enterprises must align infrastructure decisions with broader goals around revenue, customer experience, and innovation.

Key considerations for decision-makers include:

- Which AI workloads benefit from edge deployment to improve responsiveness and reduce latency?
- How do we measure cost-performance over time to avoid runaway inference costs?
- Are our teams equipped with developer tools that support cross-platform AI deployment?
- Do our compute platforms support seamless heterogeneous processing across CPUs, GPUs, NPUs and other AI accelerators?
- Are we working with partners who offer ecosystem-wide support, not just point solutions?
- Do we have a roadmap to unify the compute strategy across edge, data center, and cloud?

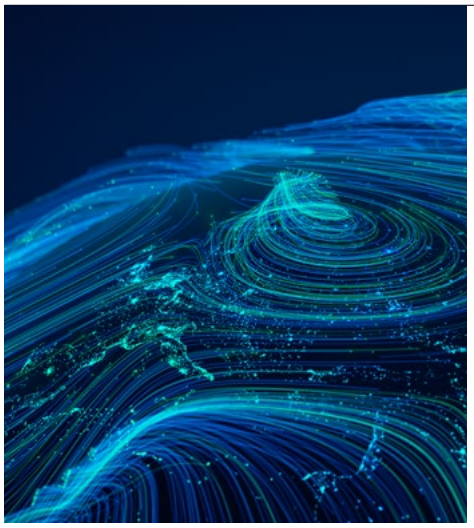
This strategic alignment ensures that compute investments are future-proofed and scalable, supporting both current demands and long-term innovation.

LESSONS FROM REAL-WORLD DEPLOYMENTS

The principles outlined above are already reshaping industries:

- **Mobile AI assistants** are evolving rapidly. Read AI's mobile meeting assistant shows how Arm enables efficient, real-time AI performance at scale. By running its iOS copilot on AWS Graviton–based EC2, built on 64-bit Arm Neoverse cores, Read migrated 89% of its CPU workloads (41% of total compute) to Graviton, including ML inference tasks. This shift delivered a ~20% reduction in infrastructure costs while improving performance and responsiveness, resulting in a faster, more cost-efficient mobile AI experience that Read can update quickly and scale globally.
- **Generative AI models**, such as text-to-audio, are going from minutes-long cloud jobs to real time on phones. Eclipsa Audio, an IAMF-based format unveiled at CES 2025 by Google, Samsung, Arm, and the Alliance for Open Media, shows how Arm made the leap possible. Upstreamed, NEON-accelerated optimizations to Opus and libiamf delivered over 160% speedups on Arm Cortex-A55 (Pixel 7) and decoding in under 16% of real time on AArch64 (<23% on Arm32). That headroom reduces cloud costs, enables on-device personalization, and brings cinematic 28-channel/binaural experiences to mainstream mobile and TV hardware.
- **Cloud-native platforms** at scale require a balance cost and availability. Sprinklr, a unified customer experience management platform, reduced compute costs by 30% while maintaining 99.99% application resiliency by migrating to AWS Graviton-based instances powered by Arm Neoverse cores. This transition allowed Sprinklr to scale global operations with greater efficiency, optimizing infrastructure spend without compromising on performance or reliability—key for delivering real-time customer insights across more than 100 countries.

These examples underscore the larger point: strategic decisions around infrastructure design are unlocking capabilities that were once considered out of reach.



CONCLUSION

The next wave of enterprise AI will not be defined by who builds the biggest models, but by who designs the smartest infrastructure.

By rethinking infrastructure design, embracing heterogeneous compute, unifying the software layer, building for futureproofing, and aligning compute with business strategy, leaders can create a compute platform that is not only scalable but also adaptable to whatever comes next.

The organizations that act on these principles will reduce operational complexity, improve efficiency, and unlock entirely new business models. Most importantly, they will be prepared for the unpredictable future of AI, where workloads will continue to evolve, but the need for scalable, efficient infrastructure will remain constant.

Learn how Arm is enabling scalable, efficient AI, from cloud to edge.