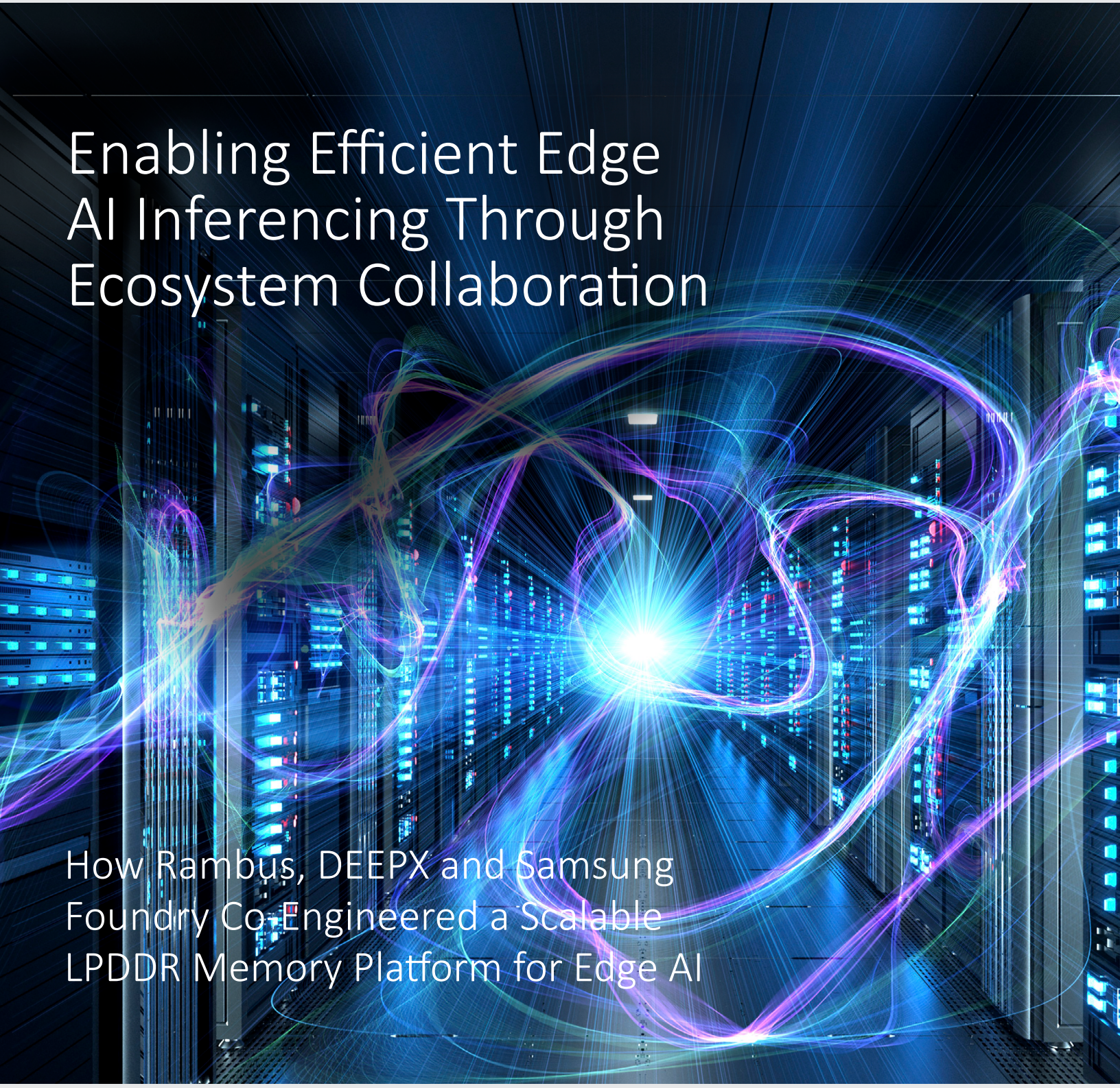
Enabling Efficient Edge AI Inferencing Through Ecosystem Collaboration



How Rambus, DEEPX and Samsung Foundry Co-Engineered a Scalable LPDDR Memory Platform for Edge AI

Table of Contents

Executive Perspective

Artificial intelligence is no longer defined by centralized compute alone. As inference workloads shift from hyperscale data centers into cameras, robots, industrial systems, edge servers, and intelligent infrastructure, the definition of system performance is changing. Latency, power efficiency, data locality, and cost increasingly outweigh peak throughput as primary design constraints. At the center of this transformation is the memory subsystem.

For edge AI, memory is no longer a background consideration. It directly shapes achievable model size, sustained inference throughput, and overall system efficiency. Solving this challenge requires more than faster DRAM or higher interface speeds. It demands close coordination between AI compute architecture, memory interface IP, and advanced silicon platforms.

This paper describes how DEEPX, Rambus, and Samsung Foundry worked together to deliver an optimized LPDDR memory solution for edge AI inference. By aligning architecture, IP, and manufacturing early in the design process, the three companies reduced risk, accelerated development, and enabled AI systems that scale across product generations and process technologies. More broadly, this collaboration illustrates how ecosystem level engineering is becoming a requirement for next generation edge AI.

Edge AI is Reshaping System Priorities

The rapid expansion of AI into edge environments is driven by practical necessity. Applications such as smart cities, factory automation, robotics, and autonomous systems require real time responsiveness, predictable latency, and local decision making. Transmitting raw data to the cloud is often impractical due to bandwidth, power, privacy, or connectivity constraints.

At the same time, AI models continue to grow in size and sophistication. Even when training occurs in the data center, models deployed at the edge increasingly require large parameter counts, extended memory footprints, and sustained memory bandwidth. This combination creates intense pressure on memory subsystems in devices that must remain power efficient, compact, and cost effective.

Traditional memory approaches reveal clear tradeoffs. On-chip memory enables speed but does not scale economically. High bandwidth memory delivers exceptional throughput but introduces cost structures that may be poorly suited to most edge deployments as well as packaging complexities in edge deployments. Graphics oriented DRAM offers high data rates but often exceeds acceptable power budgets.

Edge AI systems therefore operate in a difficult middle ground. They require both performance and efficiency within strict physical and economic limits.

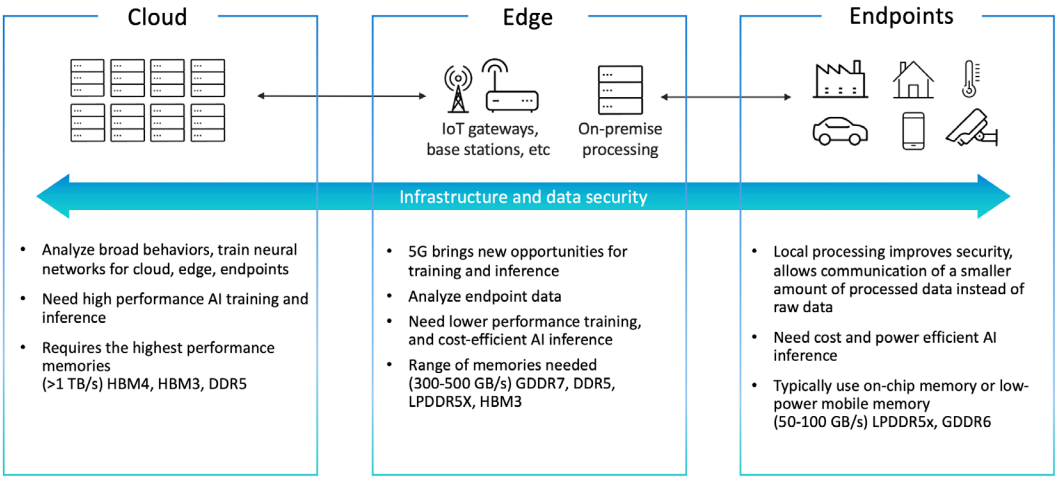


Figure 1: AI Infrastructure Memory Requirements (Source: Rambus)

LPDDR as a Strategic Enabler for Edge Inferencing

LPDDR has become a crucial technology for addressing this middle ground. Although originally developed for mobile applications, LPDDR has evolved significantly. LPDDR5 and LPDDR5X now deliver bandwidth levels that support advanced AI inference workloads while consuming far less power than traditional high performance memory options.

LPDDR also provides architectural flexibility. System designers can tune channel counts, interface widths, and memory densities to match the needs of specific edge applications. This enables a more precise balance of bandwidth, capacity, and power consumption.

However, LPDDR delivers the greatest value when the memory controller and physical interface are designed and validated together as part of a complete system, providing you a complete solution. As data rates increase, late-stage integration risk becomes unacceptable. This reality has driven a shift away from assembling discrete components toward ecosystem level cooperation.

A Collaborative Model for Edge AI

The collaboration between DEEPX, Rambus, and Samsung Foundry reflects this shift. Each company contributes a distinct capability, but no participant operates in isolation.

DEEPX develops highly efficient AI inference processors optimized for edge environments where performance per watt and predictable behavior are critical. Rambus brings deep expertise in memory interface architecture, delivering LPDDR controller IP that maximizes bandwidth utilization while minimizing latency and power. Samsung Foundry contributes with leading foundry platforms that deliver scalable, high volume production across multiple process generations—from 5nm to 2nm GAA. Rather than engaging sequentially, the three companies aligned around a common system vision. Memory subsystem requirements were defined alongside AI compute architecture and process technology decisions. This allowed architectural tradeoffs to be evaluated holistically instead of in isolation. As a result, memory transitioned from a late-stage integration challenge into a core system enabler.

A Proven Foundation with DX M1

The collaboration first materialized with DEEPX DX M1, an AI processor manufactured on Samsung Foundry 5nm technology and integrating an LPDDR5 controller IP from Rambus. The DX M1 established a production ready baseline for efficient edge inference and has been deployed across applications including robotics, drones, smart cameras, industrial automation, and edge servers.

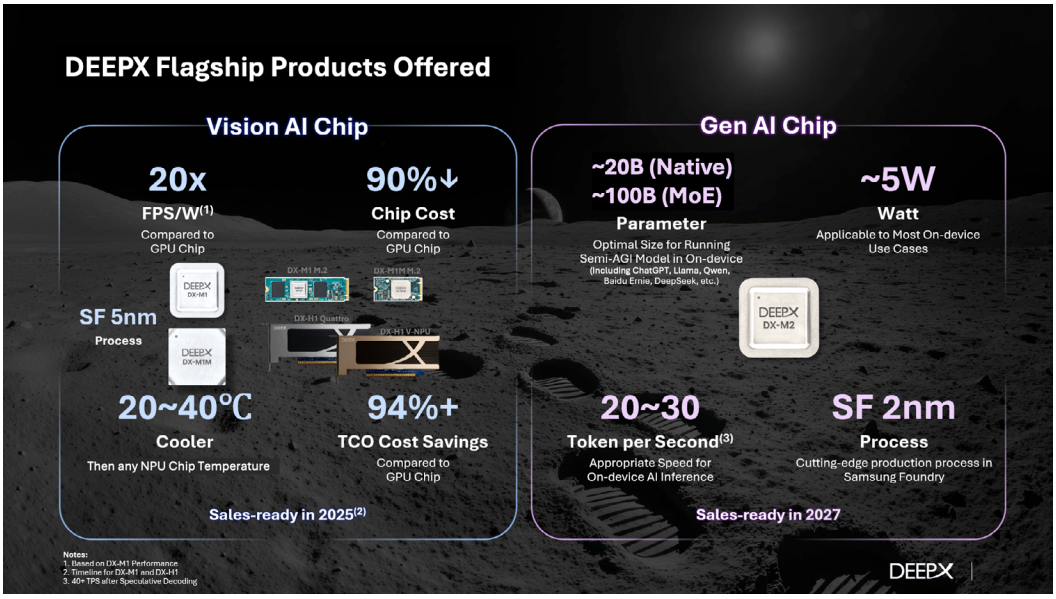


Figure 2: DEEPX Flagship Products (Source: DEEPX)

This first-generation platform validated several key assumptions. LPDDR provides the right balance of bandwidth and power for edge AI. A tightly integrated controller and PHY solution reduces delivery risk. Early ecosystem collaboration shortens the path to deployment.

Just as importantly, DX M1 validated the architectural foundation that would scale forward to the next generation product.

Scaling Forward with DX M2 and Advanced Process Technology

Building on DX M1, DEEPX is developing the next-generation DX M2 processor, targeting ultra-low power generative AI inference at the edge using Samsung Foundry 2nm MBCFET (GAA) process technology. We see the DEEPX M series migration from a 5nm to a 2nm process as well as the evolution from an LPDDR5 to LPDDR5X memory subsystem, leading to an improved memory subsystem and increased performance.

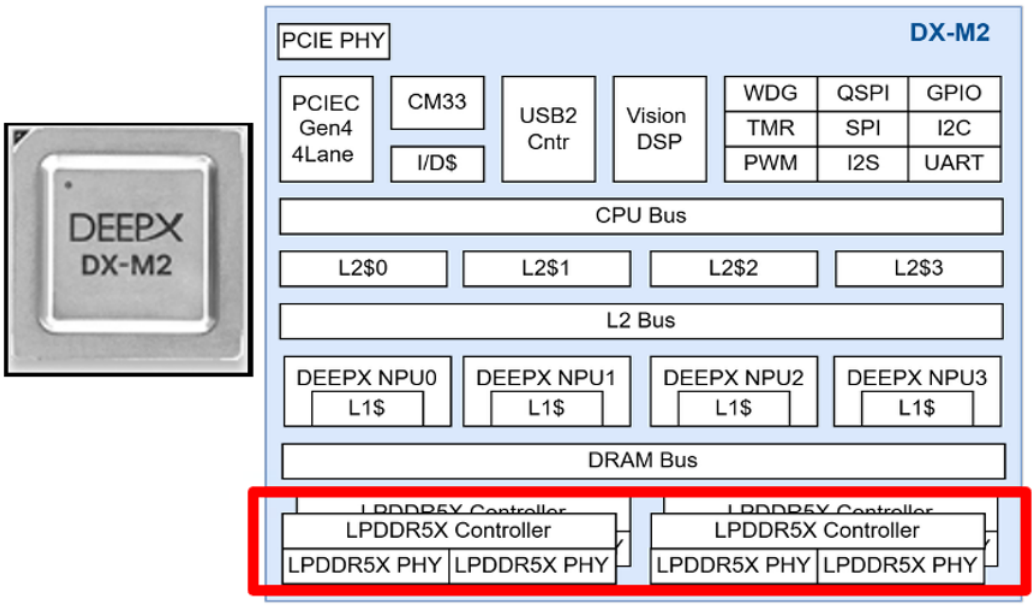


Figure 3: Integrating the Rambus Controller (Source: DEEPX)

As AI workloads become increasingly complex, advanced process nodes are essential not only for compute efficiency but also for overall system power, thermal management, and scalability. Migrating to Samsung Foundry's 2nm technology delivers substantial performance improvements for the DEEPX DX M2 processor while preserving the energy efficiency required for edge deployments. The 2nm node maximizes performance per watt through low resistance redistribution layers (RDL) and the Hyper Cell architecture, providing a performance optimized solution for power sensitive applications.

Through the Samsung Advanced Foundry Ecosystem, Rambus works closely with Samsung Foundry to optimize its memory controller IP across all process technologies, most recently on 2nm. This alignment ensures proven IP scales efficiently, lowering integration risk and accelerating production readiness for customers such as DEEPX.

The LPDDR Memory Subsystem as a Platform

At the heart of both DX M1 and DX M2 is a tightly integrated LPDDR memory subsystem.

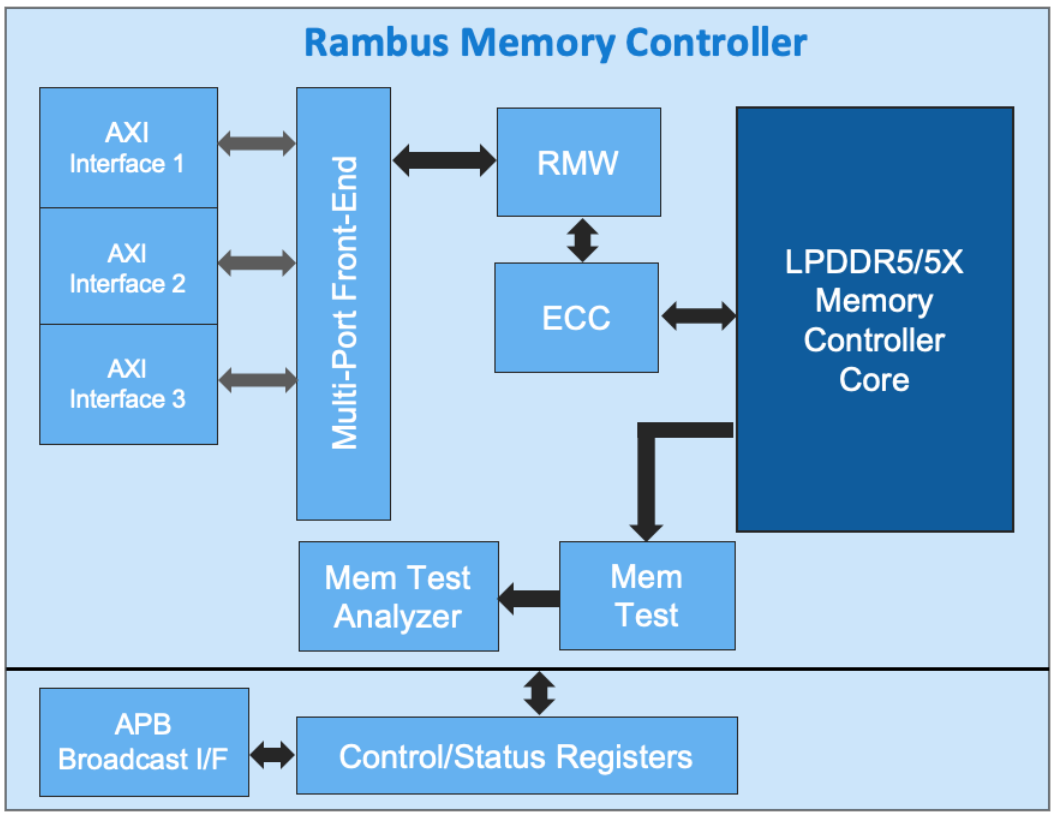


Figure 4: Rambus LPDDR5X Memory Controller (Source: Rambus)

The Rambus controller supports very high data rates and incorporates:

- Advanced scheduling
- Bank management
- Command lookahead to maximize throughput under AI class memory traffic

Delivered as soft RTL, the controller enables architectural reuse across multiple products while supporting features such as:

- Multi-port access
- Inline error correction
- Deep system visibility

Enabling Large Scale Inference at the Edge

For DEEPX, this collaborative memory platform enables capabilities that are uncommon in edge class devices. Wide multi-channel LPDDR interfaces allow high sustained bandwidth, while large aggregate memory capacity supports increasingly complex models without frequent data movement.

This capability is especially important as generative and multimodal AI workflows move closer to the edge. Efficient inference in these domains depends not only on compute throughput but also on the ability to keep large working sets resident in memory.

By avoiding the cost and power penalties of stacked memory while delivering far greater scalability than on chip approaches, the LPDDR-based solution enabled DEEPX to address a wider set of applications without compromising efficiency.

Memory, Process and Architecture as a Unified System

One of the most important lessons from this collaboration is that memory, process selection, and system architecture can achieve the optimal solution when done in collaboration, thus having a strong ecosystem enables this. As AI workloads diversify and edge constraints tighten, marginal gains at the component level are insufficient.

Value increasingly comes from a system working as an integrated whole. Advanced process nodes amplify the benefits of well architected memory subsystems, while mature validated IP reduces

risk when scaling aggressively. Ecosystem collaboration ensures these elements reinforce one another.

The DEEPX, Rambus, and Samsung Foundry collaboration demonstrates how this system driven approach leads directly to faster development, improved power efficiency, and scalable product roadmaps.

Broader Implications for Edge AI

While this work is grounded in specific products, its implications extend across the edge computing landscape. LPDDR is increasingly used in heterogeneous systems that combine accelerators, CPUs, and GPUs, as well as platforms that blend multiple memory technologies to balance capacity and bandwidth.

In each case, success depends less on raw interface speed and more on disciplined integration, validation, and architectural foresight. Ecosystem collaboration is no longer optional. It is becoming a competitive necessity.

Conclusion

Edge AI is redefining the boundaries of system design. Delivering real time intelligence outside the data center demands memory solutions that balance performance, efficiency, scalability, and predictability within strict constraints.

The collaboration between DEEPX, Rambus, and Samsung Foundry shows how these challenges can be addressed through early system level alignment across AI compute, memory interface IP, and leading-edge manufacturing. By co-engineering a scalable LPDDR memory platform and validating it across multiple product generations and process technologies, the three companies have established a blueprint for efficient edge inferencing offering a scalable memory subsystem solution.

As AI continues to expand beyond centralized infrastructure, this model of ecosystem collaboration will be essential not only for better hardware, but also for faster innovation across the edge AI landscape.

For more information, visit
rambus.com/interface-ip/lpddr