AMD

THE NEXT-GENERATION DPU AND NIC OPTIMIZED FOR AI WORKLOADS

SUMMARY

There has been recent debate about whether next-generation AI applications, including generative AI, will truly enable enterprise transformation in the form of improved knowledge-worker productivity and enhanced customer experiences. However, one thing is clear: larger enterprises, public cloud service providers, and hyperscalers are betting big—investing significantly in modern infrastructure to support private AI cloud and other infrastructure-as-a-service offerings designed to unlock new AI monetization opportunities.

AI workloads continue to tax legacy connectivity hardware and software, given the immense data volumes that feed AI models and support the scale-out of larger cluster sizes. At the same time, the high capital and operational expenses associated with new AI infrastructure cannot be overstated. To address these challenges, AI-optimized data processing units and network interface cards can provide valuable performance enhancements and cost-effective offload. An open ecosystem, buoyed by the Ultra Ethernet Consortium, is also making considerable progress to improve the Ethernet standard to support this objective. Consequently, there is an opportunity to leverage all these advancements to deliver cost-effective, highly performant, and highly available connectivity for today's modern AI applications.

Moor Insights & Strategy (MI&S) believes that the third-generation AMD Pensando Salina DPU and the AMD Pensando Pollara 400 NIC can provide what is required to facilitate cutting-edge AI services at scale. AMD Pensando Salina delivers twice the performance of previous DPU generations. Furthermore, AMD Pensando Pollara 400 represents the industry's first Ultra Ethernet-ready AI data interconnect solution, fortified with next-generation RDMA transport capabilities. In pairing these two devices, AMD can help IT infrastructure providers meet the stiff demands of AI applications on-premises, in the cloud, and eventually at the network edge.

AMD PENSANDO SALINA DPU

The AMD Pensando Salina DPU is the company's third-generation system on a chip; it marries high-performance network interconnect capabilities and acceleration engines, providing critical offload to improve AI and ML functions. As with previous generations, Pensando Salina also includes a fully backward-compatible software stack to speed deployment and ensure operational efficiency.

From an architectural standpoint, AMD's newest DPU supports four significant enhancements that will provide needed headroom for next-generation AI application processing at scale.

- 2x400G transceiver support leverages four-level pulse amplitude modulation (PAM4) to facilitate faster and more efficient data transmission.
- 232 P4 match processing units provide the ability to program more I/O functions, significantly improving network traffic steering.
- 2x DDR5 memory dramatically improves bandwidth and supports multi-core processing.
- 16 Arm Neoverse N1 cores facilitate an optimal balance of compute density, scale-out, and lower power consumption for hyperscale workloads.

Combined, these new features allow the Pensando Salina to deliver an impressive twofold improvement in performance, bandwidth, and scale-out over AMD's prior DPU generations. Consequently, MI&S believes that this level of functionality positions AMD's latest DPU as an ideal choice for demanding cloud-native and hyperscaler AI applications. This insight is supported by the company's current sampling of the Pensando Salina DPU with Microsoft Azure, IBM Cloud, and Oracle Cloud Infrastructure, with commercial availability expected in the first half of 2025.

THE ARGUMENT FOR ETHERNET OVER INFINIBAND

The technical interconnect capabilities of InfiniBand are unquestioned, making it ideally suited for supercomputing and other high performance computing workloads. InfiniBand's ultra-low latency makes it an ideal choice for applications that require near real-time data transfer speeds. However, today hyperscalers are effectively leveraging evolved Ethernet for many AI workloads, given its lower cost relative to InfiniBand and its continual performance enhancements. These improvements include broader scale-out and new accelerator-to-accelerator communication efficiencies, as well as support for backend network programmability to meet the evolving needs of AI large language models. For these reasons, AMD has chosen to adopt Ethernet as the interconnect standard for its new network interface card purpose-built for AI.

AMD Pensando Pollara 400

The AMD Pensando Pollara 400 represents a leap forward in the design of network interface cards. As just mentioned, it is purpose-built for AI workloads, with an architecture based on the latest version of RDMA that can directly connect to host memory without CPU intervention. AMD says that this new NIC, which employs unique P4 programmability and supports 400G interconnect bandwidth, can provide up to 6x improvement in performance when compared to legacy solutions using RDMA over Converged Ethernet (RoCE) version 2.

Furthermore, the Pensando Pollara 400 is one of the industry's first Ultra Ethernet-ready AI NICs, meaning that it is supported by an open and diverse ecosystem of partners within the Ultra Ethernet Consortium, including AMD, Arista, Cisco, Dell, HPE, Juniper, and many others. The device is presently sampling with infrastructure providers and is expected to begin commercial shipments in the first half of 2025.

MI&S believes that AMD's decision to invest in Ethernet for its next-generation NIC development effort is wise. Ethernet has a proven record across five decades for supporting affordable and highly evolved connectivity performance. Additionally, plug-and-play support for AI cluster deployments that can be spun up in hours versus days or weeks is a potential game-changer from a deployment perspective.

Call to Action

Next-generation AI applications will continue to tax connectivity infrastructure, and AI-optimized DPUs and evolved Ethernet NICs can address many of these challenges. The AMD Pensando Salina DPU and the UEC-ready AMD Pensando Pollara 400 have the potential to meet the demands of public cloud providers, hyperscalers, and large enterprises. MI&S believes that both of AMD's new solutions provide significant improvements in programmability, bandwidth, and power efficiency for unparalleled data transmission and traffic steering for AI workloads.

Important Information About This Paper

Contributor(s)

Will Townsend, Vice President and Principal Analyst, Networking & Security Practices

Publisher

Patrick Moorhead, CEO, Founder and Chief Analyst at Moor Insights & Strategy

Inquiries

Contact us if you would like to discuss this report, and Moor Insights & Strategy 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 "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

Licensing

Disclosures

AMD commissioned this paper. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.