PENGUIN SOLUTIONS

NEWS RELEASE

Penguin Solutions Unveils Industry's First Photonics-Based HBM-Class Memory Appliance to Break the GPU Memory Wall for Agentic AI

2026-11-16

> MemoryAI™ Photonic Appliance, featuring Marvell® Photonic Fabric™ technology, is designed to enable the transparent sharing of disaggregated high-bandwidth memory across multiple CPU and GPU-based servers

FREMONT, Calif.--(BUSINESS WIRE)-- **Penguin Solutions**, Inc. (Nasdaq: **PENG**), the AI Factory Platform Company, today unveiled the MemoryAI™ Photonic Appliance, the industry's first photonics-based HBM-class memory appliance designed to overcome GPU memory capacity constraints for agentic AI workloads. Using silicon photonics, the appliance is built to enable CPUs and GPUs to transparently share memory at high bandwidth memory (HBM)-class data rates. Penguin is showcasing the appliance at SC26 in Chicago, alongside its underlying MemoryAI™ Photonic Memory Module, the world's first silicon photonics (SiPh) memory module designed to deliver HBM-class performance, built on the Marvell® Photonic Fabric™ technology platform.

HBM provides the high-speed data access GPUs need for token generation that is essential to running agentic workloads. However, localized ultra-fast HBM technology must sit within millimeters of the GPU, so its capacity is limited by package size and silicon I/O constraints. These limitations can create data bottlenecks that leave GPUs underutilized, contributing to the growing AI memory wall.

Penguin Solutions' new MemoryAI™ Photonic Appliance is designed to address these constraints by targeting terabytes-per-second data rates across a multi-rack AI infrastructure, feeding memory-starved GPUs to support more efficient execution of agentic AI workloads. The appliance will use silicon photonics, which consumes less power than traditional copper interconnects, to extend access to memory at HBM-class data rates across a photonic memory fabric. Located meters from GPUs rather than millimeters, the appliance expands available memory capacity using HBM or standard DDR5, complementing rather than replacing existing GPU-attached HBM.

Memory is shared transparently, allowing applications to access appliance memory as if it were directly attached, without specialized access routines or data copying. Memory can also be statically or dynamically allocated. Through CXL 3.0 connectivity and Marvell photonic network interface cards (NICs), the same shared memory pool can be accessed by CPUs, enabling flexible memory utilization across AI infrastructure.

The appliance is part of the Penguin MemoryAI™ Platform Architecture for scalable, tiered, and disaggregated memory that supports production AI inference workloads. It works alongside the MemoryAI™ KV Cache Server, which provides network-attached memory for KV cache offloading, and SMART® CXL add-in cards, which expand CPU memory capacity. Together, these solutions help reduce pressure on GPU-attached HBM by offloading reusable KV cache data, preserving high-speed GPU memory for active token generation, and enabling customers to scale inference workloads more efficiently without a proportional increase in GPU resources.

“Agentic AI continues to put stress on HBM memory and the GPU, and the MemoryAI™ Photonic Appliance is designed to relieve that stress,” said Andy Mills, VP, Advanced Product Development, at Penguin Solutions. “Photonics is the step-function change that makes this possible. Instead of hopping out to a CPU over a slow network, a GPU will have a direct way to access memory at scale, and CPUs and GPUs will be able to transparently share terabyte-class memory as though it were directly attached.”

“For AI inference to scale, we need to deliver tokens with greater economic efficiency,” said Preet Virk, senior vice president and general manager, Photonic Fabric Business Unit at Marvell. “By integrating Marvell Photonic Fabric technology into Penguin Solutions' MemoryAI Photonic Appliance, we are creating a path to high-bandwidth, low-latency and high-capacity memory sharing across multiple racks. This approach is designed to improve accelerator utilization and increase the quality of tokens and throughput.”

Key features of the MemoryAI™ Photonic Appliance will include:

- High-bandwidth performance: Targets 57.6 Tbps of total appliance bandwidth to support demanding AI inference workloads.
- Scalable memory capacity: Designed to support up to 16 TB of DDR5 RDIMM memory and 576 GB of HBM3E memory per appliance.
- Flexible memory access: Offers three memory pool access modes: HBM Direct Mode, HBM-DDR5 Cached Mode, and DDR5 Direct Mode.
- Shared CPU and GPU memory: Enables memory pooling across GPUs and CPUs to improve memory utilization across AI infrastructure.
- Intelligent memory management: Penguin management software enables static or dynamic memory allocation to GPUs and CPUs based on workload requirements, while managing memory capacity, performance, and service levels.

Availability

The MemoryAI™ Photonic Appliance is planned for sampling to select customers in the second half of 2027.

About Penguin Solutions

Penguin Solutions is the AI Factory Platform Company. We design, build, and manage next-generation data centers for enterprises, sovereign AI initiatives, and AI cloud providers.

With deep design expertise at the intersection of data center AI infrastructure and memory solutions, our Full-Stack AI Factory Platform combines differentiated infrastructure software, advanced memory, compute systems, end-to-end services, and industry-leading partner technologies to help customers accelerate deployment, optimize token economics, and maximize the return on their AI investments.

For more information about Penguin Solutions and its services, visit **www.penguinsolutions.com**.

Penguin Solutions and MemoryAI are trademarks or registered trademarks of Penguin Solutions, Inc. or its affiliates. Marvell and Photonic Fabric are trademarks of Marvell or its affiliates. All other trademarks are the property of their respective owners.

Media Relations

Lerin O'Neill
Sr. Director, Corporate Communications & Public Relations
Penguin Solutions
pr@penguinsolutions.com

Source: Penguin Solutions, Inc.