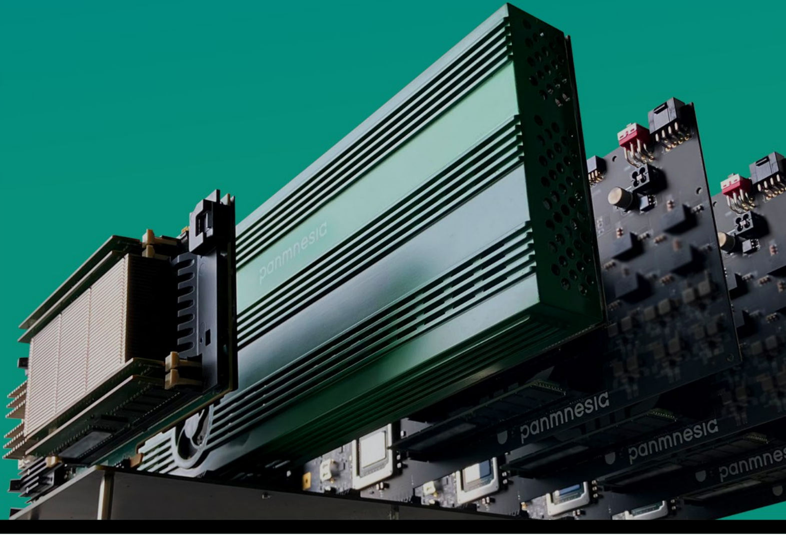
Panmnesia: AI × CXL

Hearst: Bringing Limitless Memory for Large-Scale AI-Powered Applications with CXL Memory Pool



Contact: contact@panmnesia.com

This technology is safeguarded by one or more patents.
May 11, 2023

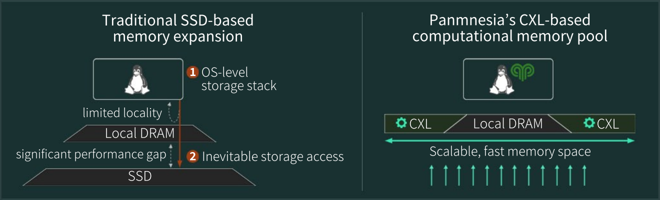
AI-driven services, including recommendation systems, search engines, and graph neural networks, have become increasingly popular due to their unparalleled capabilities and precision. As these services directly impact user experience, major tech companies must enhance their offerings to maintain their market position amidst the competitive AI landscape. To achieve this, they are rapidly increasing the amount of data their AI-driven services handle to attain greater accuracy. Consequently, there is a growing need to expand the available memory in their computing systems to accommodate such vast amounts of data.

To tackle this issue, Panmnesia offers a new type of **CXL-based disaggregated memory pool**, called *Hearst*. This CXL memory pool boasts a composable architecture that supplies extensive memory space, catering to each application's needs. Panmnesia has also investigated new architectural concepts like near-data processing in CXL, resulting in significant success for mainstream data center applications. Hearst paves the way for large-scale applications with memory disaggregation, fostering growth in the CXL ecosystem.

The Challenges

Reduced Performance, Excessive Data Duplication, Software Interventions

To handle the vast amounts of data, existing technologies have turned to high-capacity SSDs to increase available memory. However, traditional SSDs have a latency that is several orders of magnitude slower than DRAM, prompting the use of local DRAM as a cache to minimize storage access time. Despite these efforts, performance remains limited due to two primary factors: First, caching effectiveness is highly dependent on input data locality, which can vary significantly based on application behavior. Second, the OS-level storage stack, including the file system and block layer, generates unnecessary data duplication and software interference, further increasing application latency. These drawbacks make SSD-based approaches a substantial trade-off in performance for larger capacities.


Traditional SSD-based memory expansion
1 OS-level storage stack
limited locality
Local DRAM
significant performance gap
2 Inevitable storage access
SSD
Panmnesia's CXL-based computational memory pool
CXL
Local DRAM
CXL
Scalable, fast memory space

The Solution

CXL-based Computational Memory Pool

Panmnesia has developed Hearst, a cutting-edge system designed for large-scale AI-driven applications to overcome these challenges. Hearst leverages a CXL-based disaggregated memory pool, providing boundless memory to applications without sacrificing performance. The system features a scalable architecture that connects multiple memory expanders to the CPU, offering extensive memory capacity as needed. Although these connections introduce additional latency when transferring data between memory expanders and the CPU, our solution minimizes this overhead by utilizing near-data processing to reduce data movement. Furthermore, CXL enables the CPU to access the internal memory of the memory expander seamlessly, eliminating the need for software interference.

Panmnesia’s Hearst Prototype

Panmnesia's Prototype includes fully-realized all-in-one components implemented on actual hardware:

CXL-enabled CPU

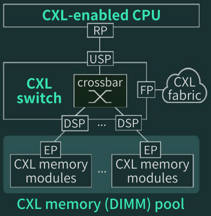
A custom 4-core processor with CXL RC capability

CXL switch

Capable of connecting **500+ memory resources** using port-based routing (PBR)

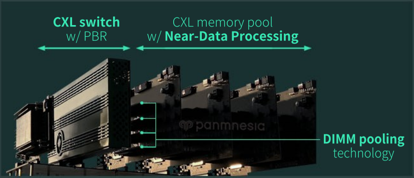
CXL memory (DIMM) pool

CXL endpoint controllers with replaceable memory modules (1~2TB per device)



What’s New?

When comparing Panmnesia's AI x CXL (Hearst) to CXL 2.0 expander-based memory pooling systems from a high-level hardware design perspective, there are three distinct differences. Firstly, Hearst utilizes **port-based routing** instead of the hierarchy-based routing (HBR) found in PCIe. Secondly, Hearst boasts an exceptional ability to process data in CXL memory through intelligent **near-data processing**. Finally, users are likely to prefer not to incur the financial cost of near data processing when memory modules need replacement. To address this, Hearst incorporates **DIMM pooling** technology into CXL endpoint device architecture and decouples all CXL controllers and data processing mechanisms from the memory modules as opposed to expander-based memory pooling.



Port-Based Routing CXL Switch

CXL 2.0-based HBR CXL switches face scalability limitations in memory pooling due to the inherent nature of PCIe bus enumeration. Panmnesia addresses this issue by introducing a port-based routing mechanism that allows for 4 petabytes per root complex, effectively overcoming CXL 2.0's scalability constraints. Hearst's CXL switch connects multiple endpoint devices, each supporting numerous DIMM modules, to the CPU via upstream ports (USPs) and downstream ports (DSPs). The USP links the CPU to the switch, while the DSPs connect memory expanders to the switch. The USPs and DSPs are interconnected by configuring the internal crossbar, ultimately forming a CXL network that supplies scalable memory space to the CPU. Additionally, Hearst's network can be connected to the CXL fabric through a distinct port type known as a fabric port, further expanding the available memory space.

Affordable Near-Data Processing through DIMM Pooling

Scalability and low maintenance costs are critical aspects that users prioritize. In AI acceleration and data processing, CXL memory pooling may impede performance, as each memory request to the CXL network involves data transfer through the CXL interconnect, leading to latency equal to or exceeding DRAM access itself. Moreover, the existing expander architecture necessitates the replacement of entire endpoint components for memory module maintenance.

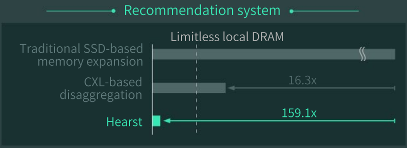
Panmnesia overcomes this challenge by integrating near-data processing within an economical, modular endpoint architecture. Hearst processes data inside the memory endpoint device by merging a domain-specific accelerator (DSA) with its controller, while maintaining separation from the underlying DIMM modules for simplified maintenance. In contrast to CXL memory expander approaches, Hearst allows the system to independently replace memory modules, eliminating the need for other module replacements or generating waste. Hearst's DSA is designed to manage data with high parallelism, returning only a small-sized result to the host rather than the original data, meeting the demands of various AI frameworks and applications. By minimizing data transfer volumes, Hearst reduces data transfer overhead and achieves superior performance.

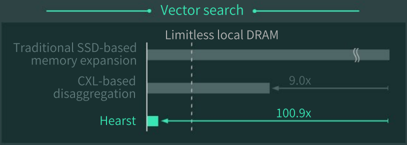
Evaluating the Performance of Panmnesia's Hearst

In a case study, we implemented our solution on two representative workloads commonly found in modern data centers and assessed the performance outcomes. We compared our solution's performance to that of a previous memory expansion approach utilizing SSDs and a baseline CXL system that does not incorporate the proposed near-data processing architecture. Furthermore, we evaluated a hypothetical system with unlimited local DRAM. For the SSD, we employed the Intel Optane 900P.

Workloads

Recommendation system	Vector search
The recommendation system predicts the items that have a high chance of attracting user's interest. It consumes 52.4% of Meta's data center resources for deep learning.	Vector search finds the objects that mostly match the user's intent using AI-generated vectors. It is widely used in various production services, such as Microsoft's Bing search.





Experiment Results

The baseline CXL system demonstrates a 12.7x performance improvement over the traditional SSD-based solution, but its performance is constrained by the overhead of transferring embedding vectors through the CXL interconnect. On the other hand, Hearst exhibits a 10.5x performance enhancement over the baseline CXL system, exceeding the hypothetical system's performance by 3.6x. This performance gain can be attributed to our near-data processing approach, which mitigates data transfer overhead and accelerates vector processing.

Conclusion

Panmnesia's Hearst provides scalable memory disaggregation by connecting multiple memory expanders to the CPU. Moreover, our solution does not sacrifice performance for memory capacity, delivering even better performance than a hypothetical system with unlimited local DRAM resources. We believe our innovative solution sets a new benchmark for harnessing the full potential of CXL technology. Panmnesia's Hearst is safeguarded by one or more patents. For additional information, please visit panmnesia.com or contact contact@panmnesia.com.

panmnesia

Brininging All Types of System Devices
to Life with Perfect Memory

Homepage: panmnesia.com
Linkedin, Youtube: Panmnesia