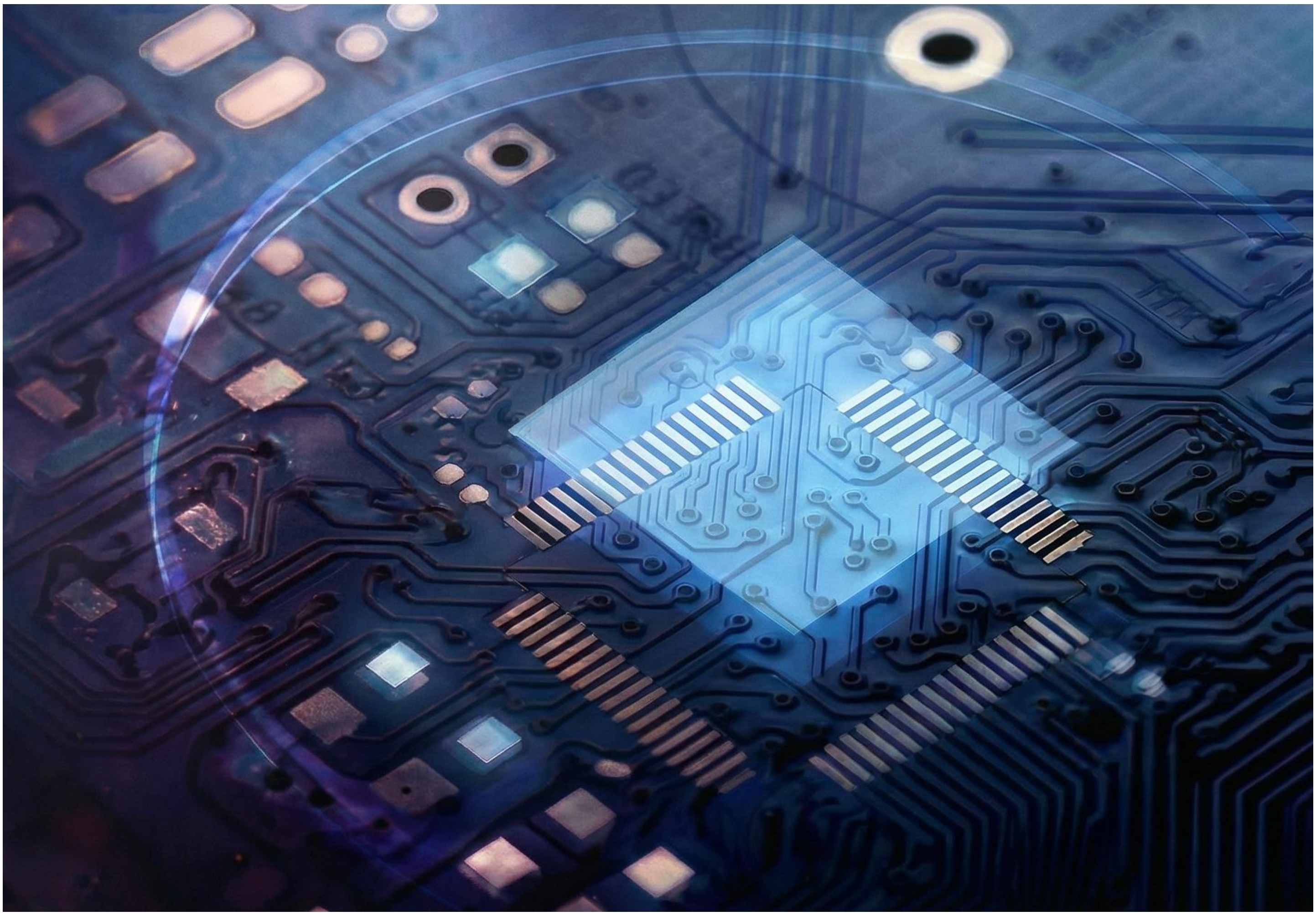
White Paper

Usage-aware Memory Auto Tiering

SLA optimization by Samsung Cognos for memory-bound workloads

Memory Solutions Lab

Authors:
Vasanthi Jagatha, Sr. Manager, Data Fabric Solutions
Mandar Sawant, Principal Engineer, Data Fabric Solutions
Mayank Saxena, Sr. Director, Data Fabric Solutions



March 2026

Legal Disclaimer

This document has been prepared by Samsung Electronics Co., Ltd. ("Samsung"). The contents of this document are the property of Samsung, protected by applicable laws and non-disclosure agreements between Samsung and you or your employer, as applicable. You are strictly prohibited from, including, but not limited to disclose-ng, copying, reproducing, distributing, transmitting, modifying, rebroadcasting, re-encoding, re-presenting, exploiting, or creating derivative works of this document or any parts of this document without the prior written permission from Samsung.

The contents of this document are provided for informational purposes only. No representation or warranty (whether express or implied) is made by Samsung or any of its officers, advisers, agents, or employees as to the accuracy, reasonableness or completeness of the information, statements, opinions, or matters contained in this document, and they are provided on an "AS-IS" basis. Samsung will not be responsible for any damages arising out of the use of, or otherwise relating to, this document. Nothing in this document grants you any license or rights in or to information, materials, or contents provided in this document, or any other intellectual property.

The contents of this document may also include forward-looking statements. These forward-looking statements include all matters that are not historical facts, as well as statements regarding Samsung's intentions, beliefs and current expectations concerning, among other things, market prospects, growth, strategies, and the industry in which Samsung operates. By definition, forward-looking statements involve risks and uncertainties because they relate to events and depend on circumstances that may or may not occur in the future. Samsung hereby reminds you that forward-looking statements are not guarantees of future performance and that the actual developments of Samsung, the market, or the industry in which Samsung operates may differ materially from those made or suggested by the forward-looking statements contained in this document or in the accompanying oral statements. In addition, even if the information contained herein or the accompanying oral statements are shown to be accurate, those developments statements may not be indicative of future developments.

All contents in this document may be subject to change without notice. Without limiting the generality of the foregoing,
1. All design, features and specifications represented herein may change without notice;
2. Images shown here have been adjusted for demonstration purposes and may appear differently on the actual products;
3. All data on products herein, including their performances, are based on internal testing using standard Samsung benchmarks under laboratory conditions. Test results do not guarantee future performance under such test conditions, and the actual throughput or performance that any user will experience may vary depending upon many factors; and
4. All images on screen are simulated, except where otherwise noted.

BY CONTINUING TO ACCESS THIS DOCUMENT, YOU ARE DEEMED TO HAVE READ, UNDERSTOOD, AND AGREED WITH THE FOREGOING TERMS AND CONDITIONS.

Definitions

CXL	Compute Express Link
TM	Tiered Memory
Cognos	Samsung distributed memory orchestration and management platform
ATM	Auto Tiered Memory (plugin for memory tiering and other performance features)
RAS	Reliability, Availability, Serviceability
SLA	Service Level Agreement
Promote	Data moved from a slower tier to faster tier with respect to latency
CHMU	CXL Hot-range Monitoring Unit

Abstract

As data-intensive workloads continue to grow in scale and complexity, the traditional reliance on expensive, high-speed DRAM creates a critical imbalance between memory capacity, cost, and performance. This issue leads to underutilized CPUs and high total cost of ownership (TCO) in modern data centers. A tiered memory architecture addresses this challenge by using DRAM as a small, fast caching layer on top of a large cost-effective, highly dense CXL or NVMe memory.

We introduce Cognos — a memory software orchestrator that abstracts the memory tiering based on the user application SLAs.

Our analysis, supported by test case evaluations on bare metal and virtualized platforms, shows that adding Cognos Auto Tier to the existing technology stack can deliver substantial TCO savings and increase workload density without sacrificing performance. We discuss the benefits of Cognos Auto Tier for various demanding workloads, including in-memory databases and machine learning applications. The findings underscore how SLA tracking is an essential strategy for balancing performance, capacity, and cost in the next generation of data centers. This is achieved by monitoring application SLAs and RAS (device failure prediction) events and automatically tiering the data to faster or slower memory devices.

Introduction

Many enterprises manage and maintain their own (on-prem) data centers. They are increasingly running AI and HPC workloads that are constrained by memory capacity and/or memory performance. HBM and DRAM are the current main memory solutions, but they are fixed per core. So, to get more memory, enterprises will have to add more compute. The quest for more memory results in:

- Purchasing more or larger DRAM devices per server
- Adding more servers to get more memory thereby leading to underutilization of compute

Both of these solutions drive up the TCO, which is a critical metric for many enterprises.

There are further challenges when more servers are added to the data center:

- Increased data movement causing degraded performance and increased power
- Adding extra networking
- More points of failures complicating RAS maintenance.

Our solution works toward a vision of creating highly memory dense systems with multiple tiers of memory that is managed by Cognos. Cognos Auto Tier automatically promotes, demotes, or migrates data between tiers based on the provided application SLA.

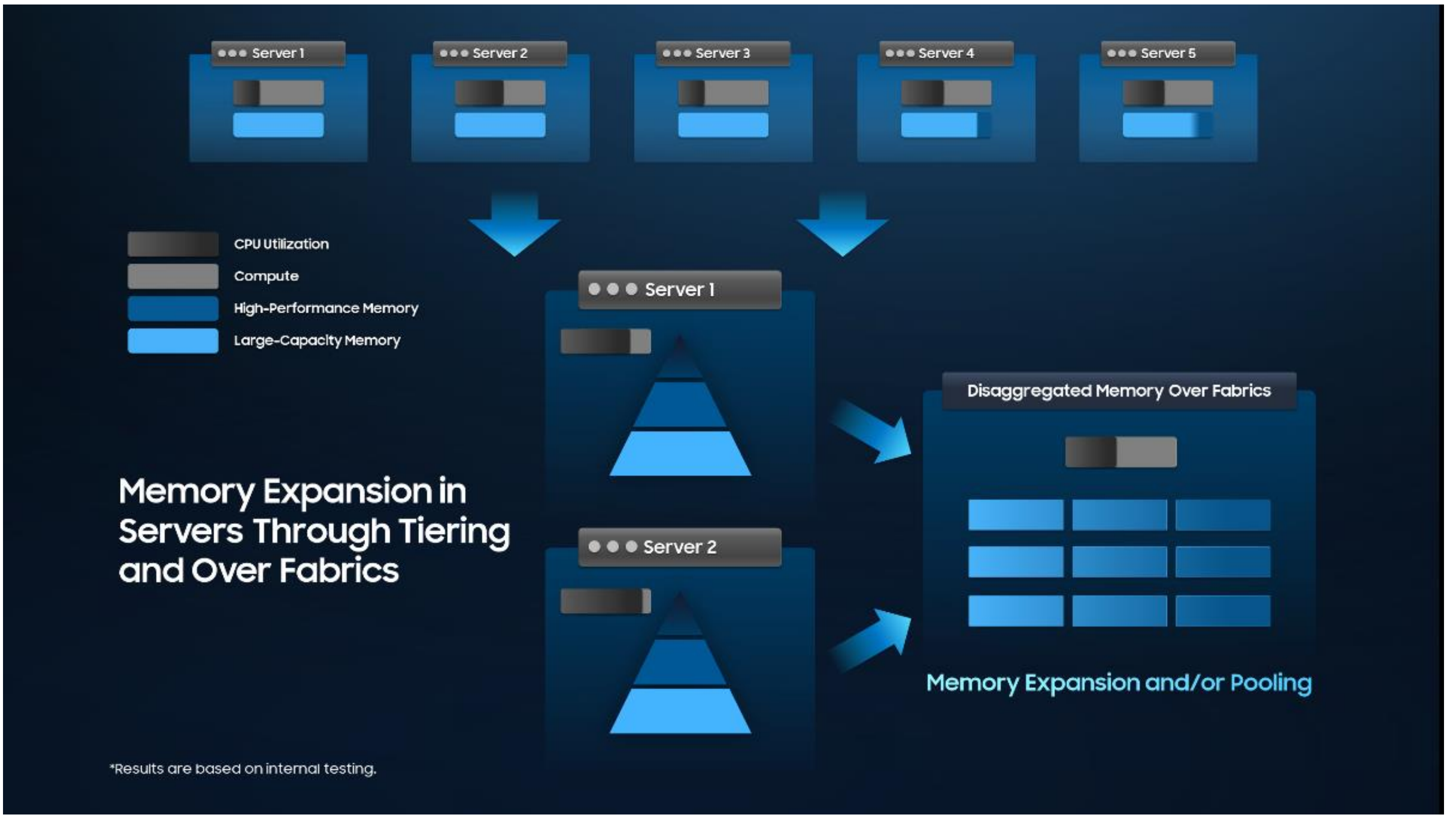

Server 1
Server 2
Server 3
Server 4
Server 5
CPU Utilization
Compute
High-Performance Memory
Large-Capacity Memory
Memory Expansion in Servers Through Tiering and Over Fabrics
Server 1
Server 2
Disaggregated Memory Over Fabrics
Memory Expansion and/or Pooling
*Results are based on internal testing.

Figure 1. Reducing server utilization through local and remote memory expansion

However, managing heterogenous memory devices with different performance and capacity profiles adds complications in the software stack when moving data between these devices.

Existing data access monitoring and movement solutions, e.g., DAMON, TPP, are more reactive than predictive by nature. Certain tunable parameters, e.g., watermarks, quotas, access parameters (sampling, aggregation intervals, etc.), age, help to label virtual memory regions as hot or cold and accordingly move (promote, demote) data between the tiers. Some support merging of regions, but merging parameters are more based on access frequencies and are not much extendible. Due to lack of control, it is possible that data movements may result in memory thrashing. Because of reactive nature and lack of control, existing solutions don't provide a way to pre-fetch the data from slower to faster tiers. Existing solutions cannot be used as they are to tier with a specific device (a hybrid memory device comprising of a cache and persistent memory), they work only on the exposed NUMA nodes.

The goal of this paper is to:

1. Introduce Cognos auto-tiering architecture and SLA-based data movement capabilities, which significantly reduce TCO while still maintaining application SLAs
2. Present experiments and results from different types of workloads using in-memory databases that showcase the value of auto tiering

System architecture/design overview

The below diagram describes the system architecture of the cluster running the AI/data warehousing application. The underlying system is a 3-node cluster consisting of 4TB DRAM and 8TB of CXL memory. The application uses VMware Tanzu GemFire, an in-memory database over the DRAM+CXL memory tiers across the 3 nodes. Cognos Auto Tier is enabled on this cluster and manages the memory and orchestrates tiering based on the application SLAs.

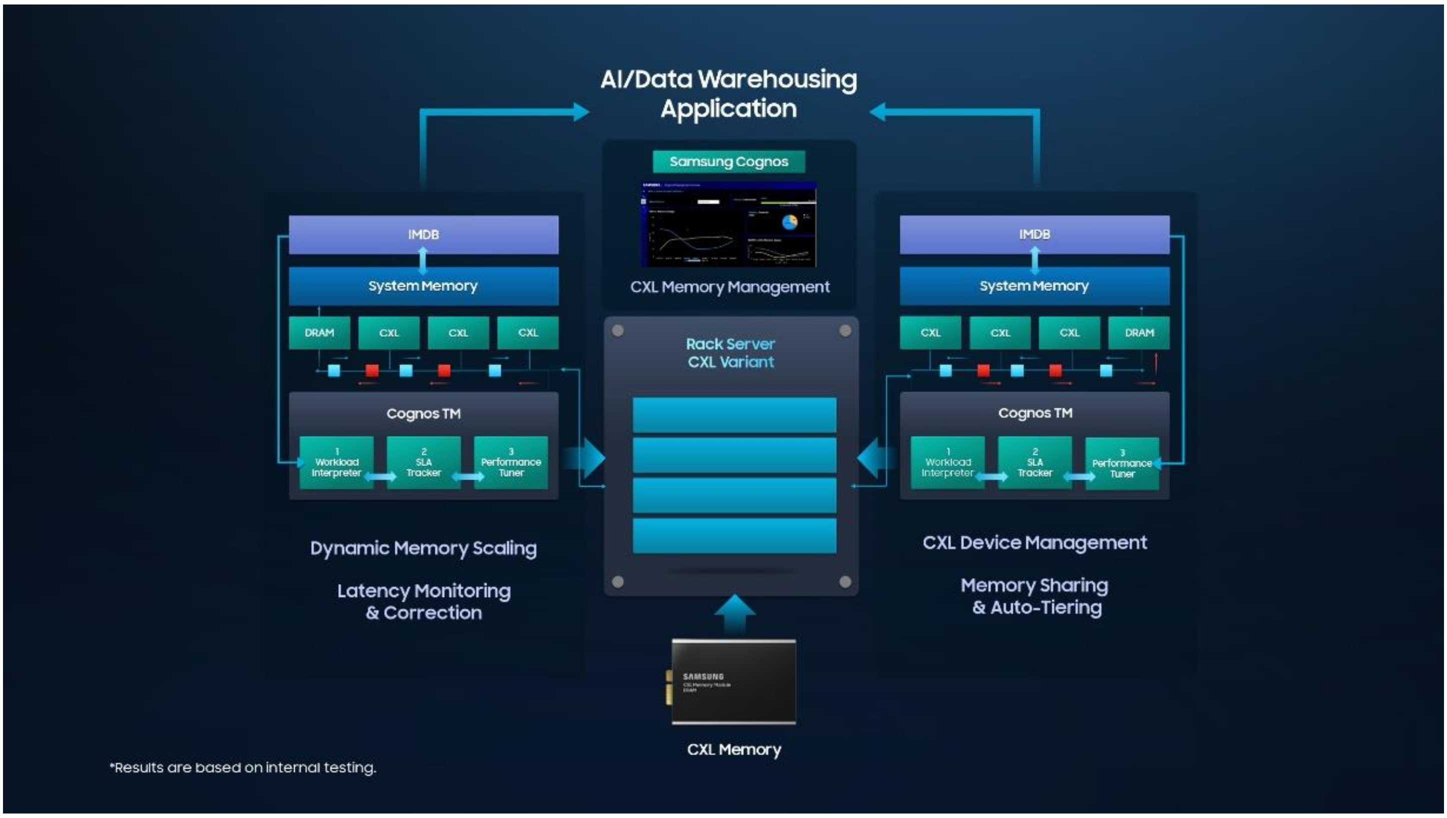


Figure 2. System configuration with CXL memory expansion

Customer-specific performance target

- Application SLA of 100kops and sub 1ms latency
- Achieve 3X TCO

Scalability

- Scalable across multiple nodes and clusters
- Move data for multiple applications and across more than two memory tiers (or devices)

Reliability

- Design must handle various failure domains, e.g., node, service, and device
- Corresponding events must be generated to notify user
- Data movement must avoid application data corruption

Application transparency and integration with existing systems or standards

- No user application changes required
- Although Cognos as a software does not have a restriction and largely depends on feature availability from the devices, Cognos defines a southbound interface to integrate with heterogenous CXL device types, which can support varying CXL protocols, CXL 1.1 (current), CXL 2.0 (current), CXL 3.x (future)
- ATM can tier across a variety of memory devices that can be exposed as a NUMA node or not, including CXL (volatile or persistent), DAX, libnvdimm, SSD, or HBM

Technology stack overview

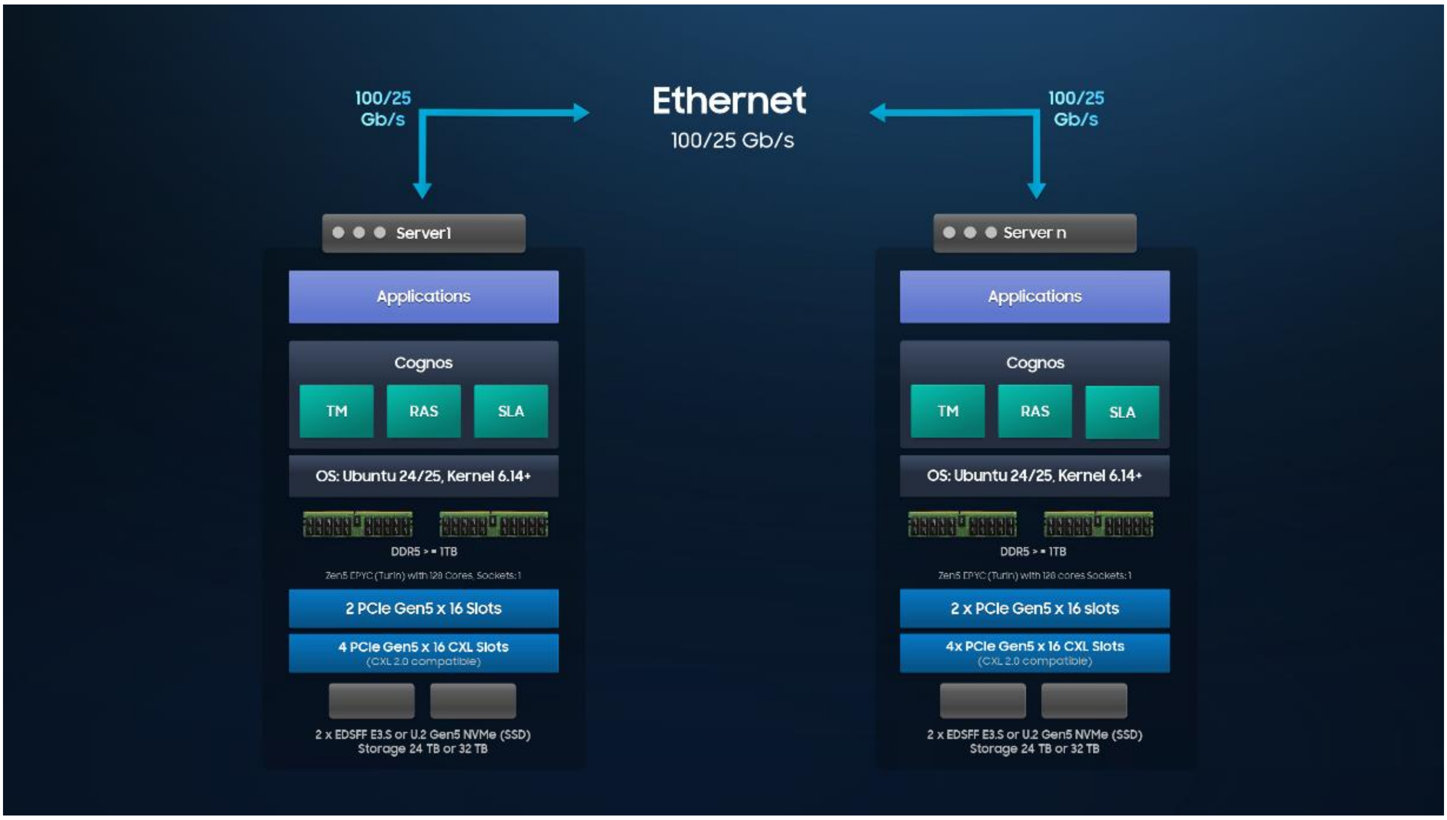

100/25
Gb/s
Ethernet
100/25 Gb/s
100/25
Gb/s
Server1
Applications
Cognos
TM
RAS
SLA
OS: Ubuntu 24/25, Kernel 6.14+
DDR5 > = 1TB
Zen5 EPYC (Turin) with 128 Cores, Sockets: 1
2 PCIe Gen5 x 16 Slots
4 PCIe Gen5 x 16 CXL Slots
(CXL 2.0 compatible)
2 x EDSFF E3.S or U.2 Gen5 NVMe (SSD)
Storage 24 TB or 32 TB
Server n
Applications
Cognos
TM
RAS
SLA
OS: Ubuntu 24/25, Kernel 6.14+
DDR5 > = 1TB
Zen5 EPYC (Turin) with 128 cores Sockets: 1
2 x PCIe Gen5 x 16 slots
4x PCIe Gen5 x 16 CXL Slots
(CXL 2.0 compatible)
2 x EDSFF E3.S or U.2 Gen5 NVMe (SSD)
Storage 24 TB or 32 TB

Figure 3. Cognos application stack

Customer-specific implementation challenges and solutions

Varying nature of application workload and IO sizes

Different workload types and IO sizes impose data profiling challenges. For example, small IO sizes (less than page size) and random workloads do not generate a specific access pattern. PTE-based profiling algorithms (DAMON) can take longer durations to identify data hotness pattern and can thus be less accurate and delay the data movement, affecting the application performance overall. Furthermore, as profiling takes longer, not only is accuracy impacted, but resource consumption is increased as well, e.g., CPU and memory.

As an initial implementation, Cognos uses DAMON to profile and move application data. Cognos configures DAMON based on a given application SLA. DAMON configurations were tested with respect to data promotion and demotion. It becomes extremely difficult and time consuming to tune DAMON for small object sizes. In certain cases, it is observed that aggressive data promotion from slower to faster memory tiers with NUMA balancing and memory interleaving provides >10% of application throughput and latency benefits. This is proved through in-memory database workloads with IO sizes as small as 1KB.

Hardware support for application data profiling

With lack of hardware (memory devices) support to provide data hotness information and counters, it is challenging to devise an accurate and an efficient memory tiering solution.

Cognos tackles this problem in multiple dimensions. With flexible architecture, it implements independent profilers based on DAMON stats (current) and performance counters (in-progress) from Intel PEBS or AMD IBS and potentially using CHMU (provided underlying device supports CXL 3.2). Additionally, Cognos also implements independent data movers based on DAMON promote/demote and bulk data movers. Movers extract information from profilers to identify a particular data region to be moved.

Data movers are classified into different categories: aggressive and non-aggressive. This helps in controlling the system resource consumption efficiently. Secondly, ATM also classifies mover configurations into promote heavy, demote heavy, or balanced. In case of small object sizes, along with an optimal interleaving ratio (e.g., 1:2), a promote heavy data mover performance is observed to be better than balanced.

Dependency on application data access hotness percentage

Application data accesses can be random or generate certain hotspots in the data. The larger the hotspots, the better the accuracy of the profiling algorithms. But realistically, this cannot be controlled.

In order to tackle this problem, once again, Cognos' multi-algorithm architecture comes to the rescue. Although application generates minimal hotspots, data accumulated through multiple profilers helps movers to connect the access temporally and spatially (future work). It is observed in some cases that random bulk data movement with adjustable region sizes can help as well as a heuristic algorithm.

Benchmark setup (hardware specs, kernel version, workload, metrics)

- Dual Socket Single node (SMC Intel SRF server) with 4TB Samsung DDR5 + 8 TB (4 x 2TB) CXL Memory
- 4 VMware Tanzu GemFire (IMDB) server instances per node
- Samsung DDR5 : CXL = 1:1 for each GemFire (IMDB) server instances
- YCSB client running on a different machine
- 10 YCSB client processes with 5 threads each totaling 50 connections to IMDB server instances

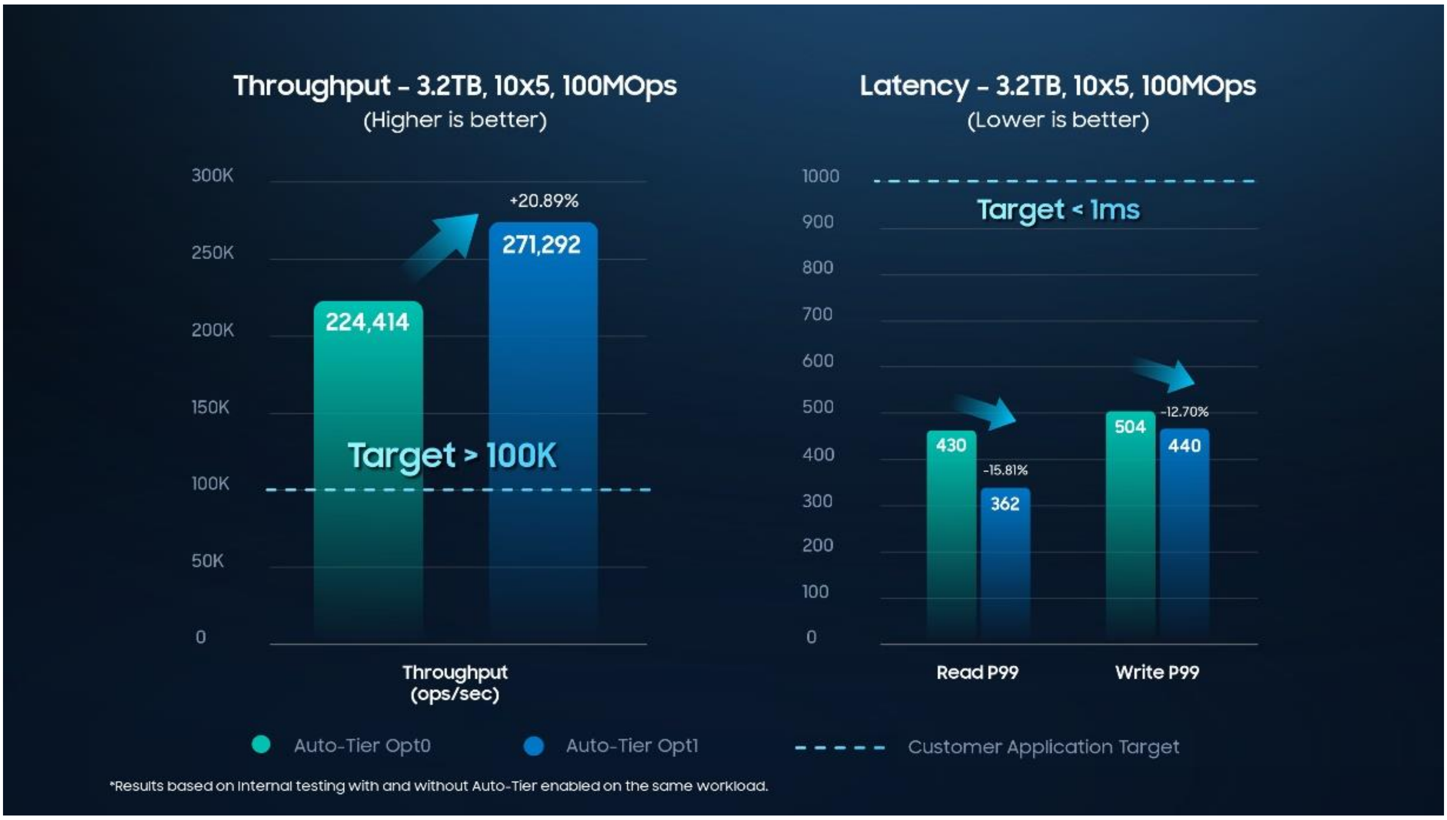


Figure 4. Auto-tiering benchmark results

Discussion: Interpretation of results

Overall, enabling Cognos auto-tiering shows ~20% improvement in throughput and 15% read and 12% write latency improvement. Cognos along with the overall Samsung memory solution resulted in 4X TCO while still maintaining customer application SLAs (100kops throughput, <1ms).

Trade-offs and limitations

- Data is synthetic with varying hotspots through a uniform workload, real-life workload results can vary
- Application software itself can have unknown bottlenecks as cluster expands
- Larger data size and memory profiling must be kept in check with resource consumption

Future work

We are planning to add the following enhancements in the upcoming releases:

- AI-powered predictive data movement to further enhance SLA tracking
- Support Samsung's next generation local and remote attached memory devices as tiers

Conclusion

Overall, Cognos SLA tracking helps to improve adoption of a variety of memory device profiles in the same memory ecosystem, improving the overall TCO benefit of the solution. Various Cognos auto-tiering experiments using heterogenous memory devices helped uncover several system limitations and enhance further collaboration across software and hardware teams.

As data centers modernize and get more memory dense, it is imperative for solutions like Cognos to intelligently and automatically tier the data based on the application SLAs. With upcoming releases, we plan to support more memory tiers and continuously improve our throughput and latency baselines.

References

(1) VMware Tanzu GemFire, https://www.vmware.com/products/app-platform/tanzu-data- intelligence/gemfire

(2) Linux Kernel Documentation, https://docs.kernel.org/

(3) Compute Express Link (CXL) Specification 3.0

(4) Intel Memory Latency Checker (MLC)

Disclaimers

(1) TCO in this paper is defined as $/capacity

(2) Published results (SLA and TCO) are based on the customer-defined workload and hardware configuration only

(3) Cognos is a Research PoC

Contact Us

For those interested in full-stack solutions and wishing to collaborate with Samsung to provide more value to their customers, please reach out to us at rdmsldfscore@ssi.samsung.com or visit our webpage to learn more: https://semiconductor.samsung.com/about-us/locations/us-rnd-labs/memory-labs/data-fabric-solutions/.

About Samsung Semiconductor
Samsung Semiconductor is a global technology leader in advanced memory, logic, and foundry solutions designed for next-generation computing. Our semiconductors power the future of AI, intelligent edge devices, and embedded platforms—delivering high-performance and energy-efficient solutions. Through close collaboration with customers, we help optimize system architectures, accelerate time to market, and enable innovation across various applications, including servers, PCs, mobile and automotive. For more information, please visit semiconductor.samsung.com.

Samsung Electronics Co., Ltd.
1-1, Samsungjeonja-ro, Hwaseong-si, Gyeonggi-do 18448, Korea www.samsung.com 1995-2025

* All product specifications and performance data included in this article reflect internal test results and are subject to variations by user's system configurations. Actual performance may vary depending on use conditions and environment.

* All images shown are provided for illustrative purposes only and may not be an exact representation of the products.