SAMSUNG

White Paper

Samsung CMM-D Utilization in IMDB Applications

Application Engineering team (Memory)

Author:
Byeonghun Hwang, Youjin Jang, Jesin Kim, Sungjin Hwang, Kyumin Park, Seungpyo Cho

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 disclosing, 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.

Abstract

Memory Limitations and the Emergence of CXL

Over the past decades, advances in CPU processing power have consistently pushed system performance forward. Yet, despite this progress, memory has remained a fundamental bottleneck. Conventional DRAM is constrained by physical slot counts, channel limits, and cost, making it difficult to scale in line with ever-growing data demands. This imbalance between compute and memory resources is often described as the “memory wall,” and it continues to limit the full utilization of modern processors.

Samsung CMM-D, a CXL-based memory module, is designed to break through these limitations. By connecting via PCIe and supporting the CXL 2.0 standard, CMM-D expands memory capacity and bandwidth beyond the CPU’s native channels. It provides a flexible and scalable approach that enables dynamic memory expansion and pooling, helping data centers use memory resources more efficiently across multiple servers. This not only alleviates the pressure on traditional DRAM but also creates a more balanced and cost-effective infrastructure for data-intensive applications.

The value of Samsung CMM-D is particularly evident in the in-memory database (IMDB) market. IMDB platforms rely on keeping entire datasets resident in memory to deliver real-time analytics and query performance. As dataset sizes grow larger and more complex, conventional DRAM alone struggles to meet these requirements, CMM-D directly addresses this challenge by offering the elasticity and scalability needed to sustain IMDB growth while improving total cost of ownership through reduced over-provisioning and more efficient infrastructure utilization.

In this way, Samsung CMM-D not only overcomes the long-standing limitations of memory but also emerges as a key enabler for rapidly expanding markets like IMDB and other memory-intensive workloads, supporting both performance innovation and operational efficiency.

Introduction

Samsung CMM-D (CXL Memory Module-DRAM)

Flexible and adaptable memory

Samsung’s CMM-D leverages the PCIe 5.0 physical layer and fully supports the CXL 2.0 protocol, ensuring both high-speed data transfer and compatibility with the latest CPU and system architectures. By using the CXL Type-3 device standard, CMM-D can seamlessly integrate into server platforms as memory expanders, providing cache-coherent across with reduced latency compared to traditional PCIe-attached storage. This makes it well-suited for AI/ML workloads, large-scale databases, and memory-intensive cloud applications.

Expansion of Bandwidth and Capacity

A major advantage of CMM-D is its ability to scale system memory for beyond the physical limitations of DDR DIMM slots. While conventional DRAM channels on CPUs are constrained by the number of DIMM slots and supported capacities, CMM-D connected via CXL provides additional memory bandwidth and memory wall, addressing both the growing dataset sizes of AI training/inference and the high-bandwidth requirements of system where CPU computational throughput can be fully utilized without being bottlenecked by memory resources.

Total Cost of Ownership (TCO) Benefits

CMM-D delivers clear advantages in total cost of ownership (TCO). By reducing the need for DRAM over-provisioning, CMM-D minimizes stranded memory capacity and enables more efficient infrastructure utilization. Its ability to expand memory capacity through PCIe/CXL slots reduces the need for costly CPU or server upgrades, lowering capital expenses.

Furthermore, memory pooling allows dynamic allocation across multiple nodes, improving workload flexibility and consolidation. Combined with reduced power and cooling requirements, these features translate into lower operational costs and a longer lifecycle for existing infrastructure, making CMM-D a cost-efficient solution for large-scale and memory-intensive deployments.

Flexible System Operation with Memory Pooling

CMM-D also unlocks a new paradigm in system resource management through memory pooling. By disaggregating memory from the CPU, CMM-D allows multiple servers to allocate dynamically memory capacity from a shared pool. This reduces standard memory resources, improves total cost of ownership, and supports highly variable workloads by providing memory "on demand." For cloud operators and data centers, this means greater efficiency, better workload consolidation, and the ability to meet diverse application requirements without over-provisioning physical DRAM in each node

SAP HANA

SAP HANA (High-performance ANalytic Appliance) is an in-memory database management system that enables real-time data processing and analytics. By leveraging in-memory computing and advanced data modeling, it supports complex queries and high-speed transactions, making it ideal for enterprise applications. By storing data in column-based tables in main memory and bringing online analytical processing (OLAP) and online transactional processing (OLTP) together, SAP HANA is unique - and significantly faster than other database management systems (DBMS) on the market today.

SAP HANA uses a compressed, columnar storage layout for both fast read accesses and a low memory footprint. The columnar data is stored in the read-optimized 'main storage' and maintains a separate 'delta storage' for optimized writes. The main storage contains table data and retrievals of these table data consume higher memory bandwidth, especially for OLAP workloads. Fortunately, accesses to the table data are mostly sequential, which enables efficient prefetching hiding the access latency. The delta storage is write-optimized columnar storage to save newly inserted or modified data. It is periodically merged with the main storage. In addition, a designated portion of memory is allocated for operational data and intermediate results during query processing by HANA Execution Engine (HEX). This allocated space is referred to as 'HEX heap memory'.

OLTP (Online Transaction Processing)

Online Transaction Processing (OLTP) is a type of data processing that manages and processes a high volume of real-time, short transactions concurrently, such as insertions, updates, and deletions, in a database. OLTP systems are optimized for rapid data entry and retrieval, ensuring quick response times and maintaining data integrity in environments like retail, banking, and e-commerce.

OLAP (Online Analytical Processing)

Online Analytical Processing (OLAP) is a software technology used to perform high-speed, complex, multidimensional analysis on large volumes of business data. OLAP systems are designed to support complex queries and enable users to perform data analysis and reporting with speed and efficiency. These systems are typically used for decision support, allowing businesses to perform operations like trend analysis, forecasting, and financial reporting across large volumes of data. OLAP provides the capability to view data from different perspectives, often using data cubes for dimensional analysis.

TPC (Transaction Processing Performance Council)

The Transaction Processing Performance Council (TPC), founded in 1988, defines benchmarks for transaction processing and databases, and to publish objective, verifiable TPC performance data to the industry. TPC benchmarks are used in evaluating the performance of computer systems, and TPC publishes the results. TPC-C is one of representative On-Line Transaction Processing benchmarks. TPC-DS is a decision support benchmark that models several generally applicable aspects of a decision support system.

CMM-D Performance in SAP HANA
: RDIMM vs CXL Memory Configurations

System Configuration

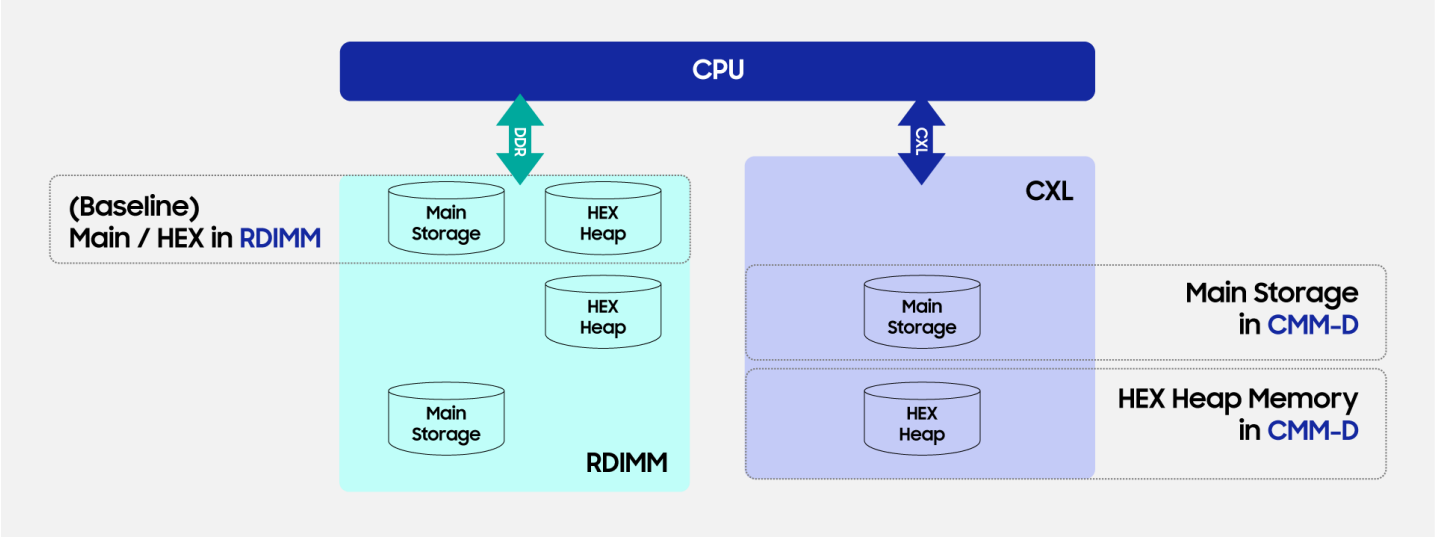


Figure 1. SAP HANA System Configuration Driven by Main Storage and HEX Heap Memory Deployment

The evaluation is conducted using three experimental cases to assess the operational behavior and performance of main storage and HEX heap memory.

(1) Baseline : Main storage and HEX heap memory in RDIMM

(2) Main CXL : HEX heap memory in RDIMM and Main storage in CMM-D

(3) HEX CXL : Main storage in RDIMM and HEX heap memory in CMM-D

HW	Host		Intel Emerald Rapids A2 stepping 2-sockets
	DIMM		DDR5 4800Mbps 64GB x16pcs / 1 host
	CMM-D	HW	CS Sample 256GB x 4pcs / 1 host
		FW	14.40.1.060f.46
	SSD		3.84TB SSD x 2pcs (Data/Log) / 1 host
SW	OS		SUSE Enterprise Server15 SP3 for SAP
	Kernel		V6.7.0-57
	BIOS		EGSPCRB1.SYS.0107.D52.231107228
	Application		SAP HANA TPC-C / TPC-DS (SF100)

Table 1. System Configuration for Experimental Evaluation

In each experimental case, the memory configuration is evaluated across six setups including the baseline: Remote DRAM, Local CXL, Local CXL x2, Remote CXL, and Remote CXL x2.

(a) Baseline : Use 8-channel RDIMM in the local NUMA node

(b) Remote DRAM: Use 8-channel RDIMM in the remote NUMA node

(c) Local CXL: Use a single CMM-D device on the local NUMA node

(d) Local CXL x2: Use two CMM-D devices on the local NUMA node

(e) Remote CXL: Use a single CMM-D device on the remote NUMA node

(f) Remote CXL x2: Use two CMM-D devices on the remote NUMA node

Synthetic Memory Bandwidth

Before conducting the main evaluation, we first measured the baseline performance of the memory devices used in our test environment. Specifically, we assessed the synthetic bandwidth of eight RDIMMs and two CMM-D devices, measuring each in both local and remote configurations.

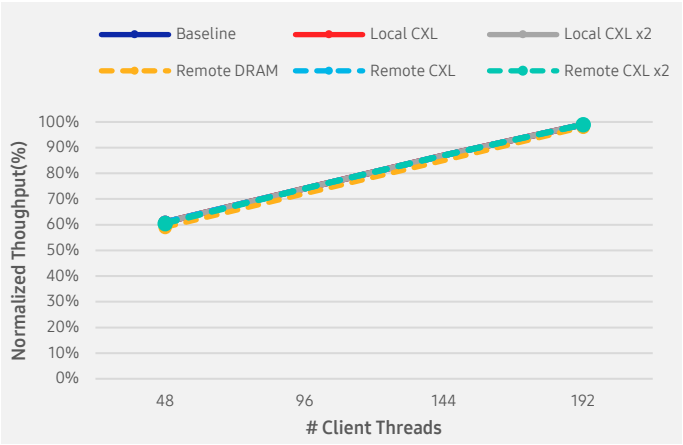
Intel MLC (v 3.10)		RDIMM (x8ch)		CMM-D (x2pcs)	
		Local	Remote	Local	Remote
Idle latency (ns)		144	239	254	353
Bandwidth (GB/s)	All Read	278.4	155.8	52.4	52.4
	All Write	236.1	152.7	47.2	47.2
	3R 1W	247.5	206.1	61.3	61.4
	2R 1W	243.9	230.0	61.1	61.3
	1R 1W	237.2	236.2	59.7	59.9
	2R1W (Stream Triad Like)	246.1	196.2	62.2	62.3

Table 2. Synthetic Benchmark Results for DRAM and CMM-D Memory Bandwidth

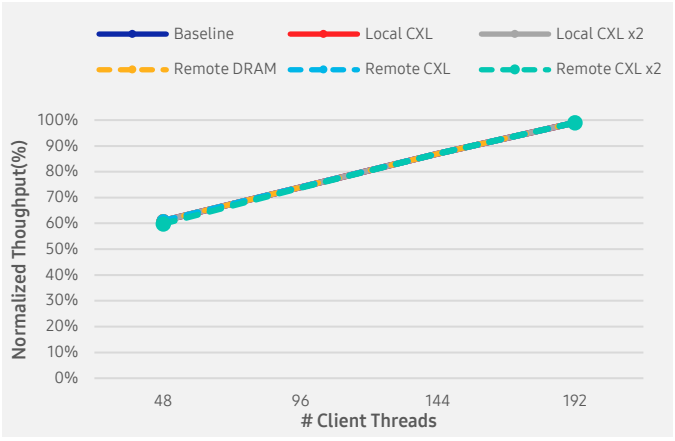
As shown in the results of Table 2, for latency, DRAM showed 144 ns locally and 239 ns remotely, while CMM-D recorded 254 ns locally and 353 ns remotely—representing approximately 40-65% higher latency compared to DRAM.
From a bandwidth perspective, RDIMM achieves up to 278 GB/s under an all-read workloads, whereas CMM-D delivers 52.4 GB/s for the same workloads and approximately 60 GB/s under a mixed read/write workloads. In addition, unlike RDIMM, CMM-D shows the same bandwidth when accessed through a remote NUMA node. Because CMM-D operates over a direct CPU-to-device CXL/PCIe link, rather than through the CPU's memory controller, so NUMA locality affects latency but has little impact on sustained bandwidth.

Performance Analysis

TPC-C



<Main CXL>



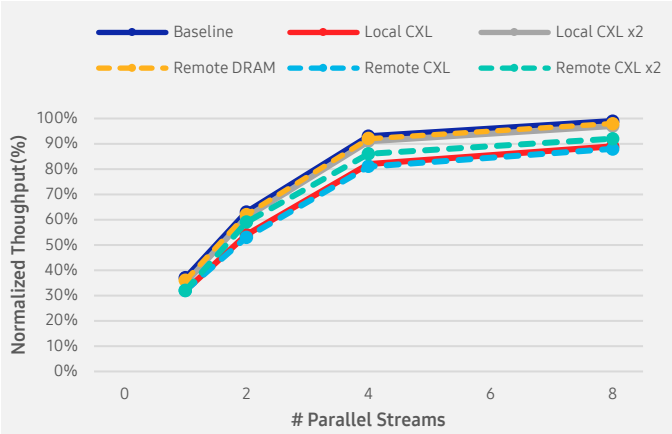
<HEX CXL>

Figure 2. TPC-C Workload Throughput Comparison between Main CXL and HEX CXL

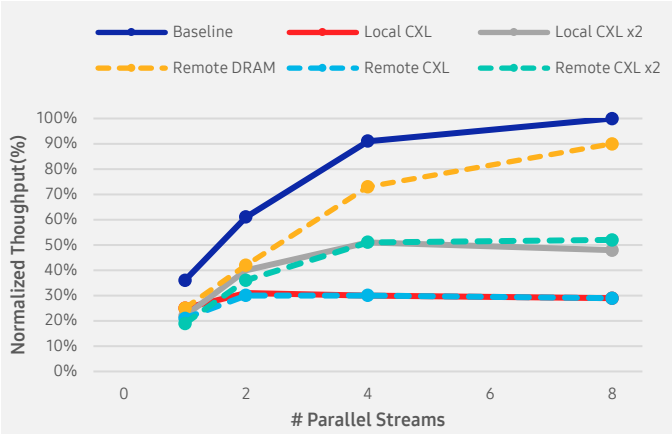
Figure 2 shows the performance comparison of Main CXL and HEX CXL for each system configuration when running the TPC-C workloads. In the TPC-C workloads, both Main CXL and HEX CXL showed no performance differences between RDIMM and CMM-D configurations. This indicates that assigning main storage and HEX heap memory operations to CMM-D does not introduce any performance degradation, especially in OLTP workloads thereby enabling additional capacity expansion through using of CMM-D.

TPC-DS

Using the same methodology applied to the previous TPC-C workload evaluation, we divided the analysis into two cases—Main CXL and HEX CXL.



<Main CXL>



<HEX CXL>

Figure 3. TPC-DS Workload Throughput Comparison between Main CXL and HEX CXL

Figure 3 illustrates the performance of Main CXL and HEX CXL under the TPC-DS workloads across various system environment configurations. For Main CXL, using two local CMM-D devices interleaving resulted in approximately a 4% performance reduction compared to the baseline at 8 streams. Since this performance remains effectively equivalent to the baseline environment, it suggests that adding CMM-D can provide capacity expansion benefits without incurring meaningful performance loss.

In contrast, the HEX CXL evaluation showed a significant performance degradation, with a 48% reduction observed at 8 streams when using two local CMM-D devices interleaving. This indicates that CMM-D is not suitable for serving as HEX heap memory.

	QPM	Bandwidth (GB/s)							
		RDIMM (x8ch)				CMM-D x2			
		Main	HEX		Total	Main	HEX		Total
		Read	Read	Write		Read	Read	Write	
Baseline	100%	12	12	27	51	-	-	-	-
Main CXL	93%	-	7.5	20	27.5	12.3	-	-	12.3
HEX CXL	52%	7.4	-	-	7.4	-	7.2	16.1	23.3

Table 3. Memory Bandwidth Measurements for Each Test Case in the TPC-DS Workload

Table 3 shows the bandwidth results of RDIMM and CMM-D measured under the Main CXL and HEX CXL configurations when using 4 parallel streams. According to the RDIMM performance evaluation, RDIMM provides a peak read-only bandwidth of approximately 278 GB/s, whereas the maximum bandwidth observed during the TPC-DS workload was only 51 GB/s, indicating that the workloads' bandwidth demand is relatively modest.

Although two CMM-D devices are sufficient to sustain 51 GB/s of bandwidth, the application still experiences performance degradation, indicating that further analysis is required to determine the root cause.

Workload Analysis

Access Pattern Analysis

To analyze the behavior of the SAP HANA TPC-DS workloads, we captured runtime packet traces using a LeCroy Summit T516 Protocol Analyzer. The collected data was then examined with our in-house analysis solution, PA-TAP (Protocol Analyzer-Test Agent Platform), enabling an in-depth assessment of workload characteristics and memory access patterns.

Operation	Duration (s)	Access Count			Access Count per second
		Read	Write	Total	
Main storage	2.5 s	2,940,749 (99.7%)	7,639 (0.3%)	2,948,388	1,179,355
HEX heap memory	0.25 s	2,062,931 (54.8%)	1,704,804 (45.2%)	3,767,735	15,070,940

Table 4. Comparison Access Counts Observed in Main Storage and HEX Heap Memory Operations

Table 4 presents the packet analysis results extracted using the Protocol Analyzer during main storage and HEX heap memory operations. For Main Storage, the analysis revealed a trace duration of approximately 2.5 seconds, during which about 2.9 million accesses were recorded. This corresponds to an access rate of roughly 1.18 million accesses per second, indicating a moderately intensive workloads. Additionally, 99.7% of all accesses were reads, confirming that Main Storage exhibits a read-only-dominant workload profile.

In contrast, HEX heap memory showed significantly heavier activity. Over a duration of 0.25 seconds, approximately 3.8 million accesses were observed—equivalent to 15 million accesses per second, which is more than 12x higher than the access density of main storage. Furthermore, the access distribution consisted of 55% reads and 45% writes, indicating a workload pattern closely resembling a 1:1 read/write mix.

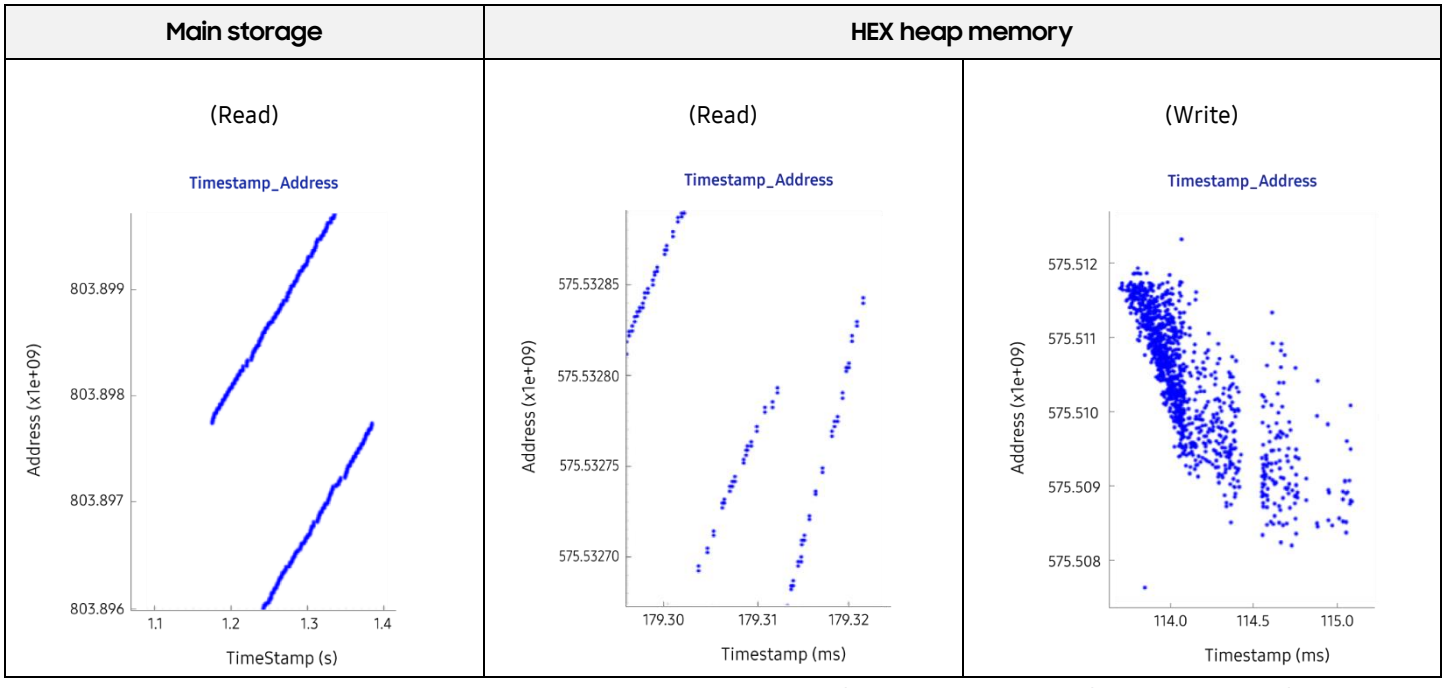


Figure 4. Access Pattern Comparison between Main Storage and HEX Heap Memory during Operation

Figure 4 illustrates the accessed addresses over time for both main storage and HEX heap memory operations. In main storage, read operations exhibit a relatively consistent and predictable address progression, whereas HEX heap memory shows a far more scattered and irregular access pattern.

Operation	Access Pattern	Access Count	
		Read	Write
Main storage	Sequential Access	2,450,434 (83.3%)	1,928 (25.2%)
	Random Access	490,315 (16.7%)	5,711 (74.8%)
HEX heap memory	Sequential Access	1,177,954 (57.1%)	84,733 (5.0%)
	Random Access	884,977 (42.9%)	1,620,071 (95.0%)

Table 5. Comparison of Sequential and Random Access Characteristics between Main Storage and HEX Heap Memory

To analyze this behavior more precisely, we examined the ratio of sequential and random accesses based on the accessed address patterns. Table 5 summarizes the proportion of sequential and random accesses observed during Main Storage and HEX Heap Memory operations. In this analysis, consecutively generated accesses targeting the same 4KB page were classified as Sequential Access, while all other cases were categorized as Random Access.

For Main Storage, more than 83% of read operations demonstrated sequential characteristics. In contrast, HEX Heap Memory showed markedly different behavior: only 57% of read accesses were sequential, and write accesses exhibited just 5% sequentially. These results indicate that HEX Heap Memory has a significantly stronger random access tendency compared to Main Storage.

Outstanding Analysis

The following section presents an analysis of outstanding behavior during main storage and HEX heap memory operations, along with its impact on latency and overall application performance. Outstanding refers to the number of in-flight requests a device must process after receiving commands from the host. In other words, outstanding increases by one whenever the host issues a request, and decreases by one when the device returns a response.

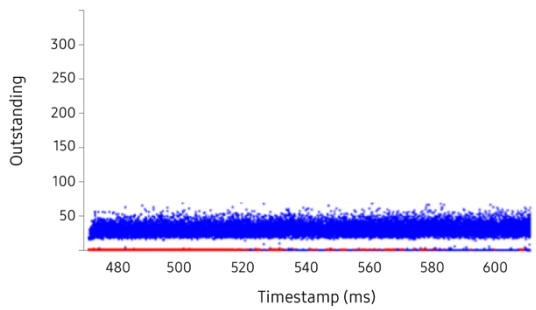
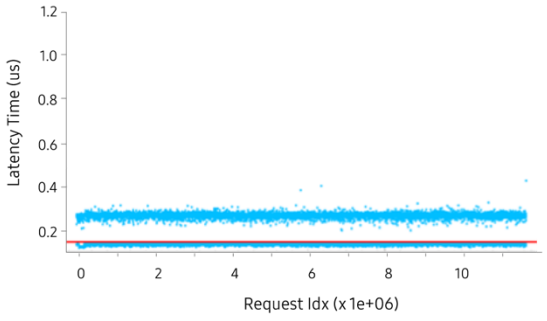
	Outstanding Request Distribution	Latency Distribution
Main storage		
	Avg. 10 Max. 82	Avg. 167 ns 99.999th 330 ns

Figure 5. Correlation Analysis Based on the Distribution of Outstanding Requests and Latency during Main Storage Operation

On the left side of Figure 5, the outstanding level (Y-axis) is plotted over time (X-axis), illustrating the device's processing load and the degree of pressure it experiences during operation. On the right side, latency (Y-axis) is shown as a function of host requests (X-axis), representing the response time observed from the moment a request is issued to when the response is returned. Together, these graphs reveal the correlation between outstanding levels and latency behavior.

During the main storage operations, the outstanding level averaged 10, with a maximum of 82. In terms of latency, analysis of the time from request issuance to response completion showed an average latency of 167ns, while the 99.999th-percentile tail latency reached 330ns.

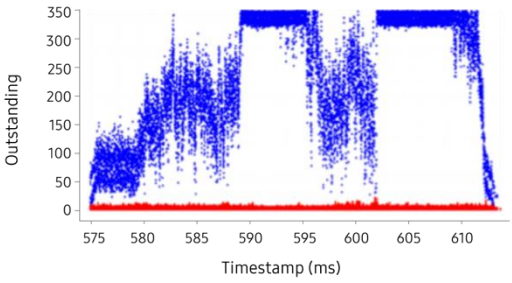
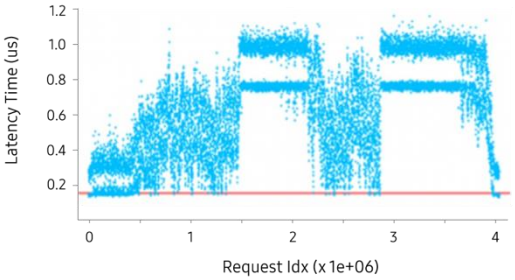
	Outstanding Request Distribution	Latency Distribution
HEX heap memory	 Avg. 198 Max. 349	 Avg. 280 ns 99.999th 1150 ns

Figure 6. Correlation Analysis Based on the Distribution of Outstanding Requests and Latency during the HEX Heap Memory Operations

Figure 6 presents the outstanding and latency characteristics observed during the HEX heap memory operations. The outstanding averaged 198, with a maximum of 349, representing approximately 20x higher average outstanding and 4.3x higher peak outstanding compared to Main Storage. The average latency increased to 280 ns, and the 99.999th-percentile latency rose to 1,150 ns, which is about 3.5x higher than that of Main Storage.

As shown in the figures, the distributions of outstanding and latency exhibit closely aligned patterns, indicating that increases in outstanding directly contribute to latency inflation. This suggests that high outstanding levels can degrade overall application performance. Therefore, to achieve optimal performance, it is essential to understand the workload characteristics and effectively manage outstanding levels to minimize latency growth.

Conclusion

Through this study, we evaluated multiple experimental scenarios to assess the applicability of CMM-D in IMDB environments. The results demonstrate that for the OLTP workloads (TPC-C), CMM-D can deliver sufficient performance without any degradation compared to RDIMM, confirming its viability as a high-capacity memory alternative. In the OLAP workloads (TPC-DS), CMM-D also proved effective for applications with sequential access patterns and relatively modest traffic demands, further validating its potential as a replacement for RDIMM in such scenarios. These findings highlight that CMM-D can be reliably leveraged in various IMDB workloads, provided that the workload characteristics align with its performance profile.

Based on these experimental insights, it becomes clear why memory scalability remains a critical challenge in modern IMDB environments.

As IMDB-based applications continue to scale, they increasingly encounter memory bottlenecks driven by rapid data growth and expanding in-memory processing requirements. These limitations not only restrict system scalability but can also lead to performance degradation and increased latency.

The introduction of Samsung CMM-D effectively addresses these challenges. By delivering both bandwidth expansion and capacity expansion, Samsung CMM-D alleviates the inherent constraints of traditional DRAM-centric architectures. As a result, IMDB workloads can maintain higher concurrency and achieve more stable latency characteristics, allowing systems to scale past traditional memory boundaries.

The benefits become even more compelling from a TCO perspective. Expanding memory capacity by adding additional server nodes incurs significant costs—including hardware investment, software licensing, and ongoing power and cooling expenses. In contrast, CXL memory enables large-scale capacity expansion without increasing server count, substantially improving cost efficiency compared to provisioning new systems. This allows organizations to eliminate scaling barriers caused by memory limitations while significantly reducing overall infrastructure TCO.

In conclusion, Samsung CMM-D offers a practical and highly effective solution for overcoming memory bottlenecks in modern IMDB environments, empowering the deployment of more stable, scalable, and high-performance next-generation in-memory database infrastructures.

Technical specifications

	Samsung CMM-D	
Part Number	MXFAH2560MB0-CCP	MXFAH1280MB0-CCP
Capacity	256GB	128GB
Form Factor	E3.S 2T	
Interface	PCIe 5.0 x8	
Protocol	CXL 2.0	
DDR	Samsung DDR5	
Speed	6400 Mbps	
Component Composition	(32G x 4) x 80	(32G x 4) x 40
Rank x Organization	4R x 4	2R x 4
Performance (2R 1W)	36 GB/s	

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

Fio is a registered trademark of Fio Corporation. PCI Express and PCIe are registered trademarks of PCI-SIG. Toggle is a registered trademark of Toggle, Inc.

* NVMe is a registered trademark of NVM Express. PCI Express and PCIe are registered trademarks of PCI-SIG.

* CXL® and Compute Express Link® are registered trademarks of the Compute Express Link Consortium.

* 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.