PRODUCT BRIEF

SUPERMICRO, INTEL, AND MICROSOFT DEMONSTRATE HOW TO ACHIEVE 35% PERFORMANCE GAINS ON MICROSOFT SQL SERVER 2025 OLAP WORKLOADS ON INTEL® XEON® 6 PROCESSORS

New Intel® Flat Memory Mode Simplifies Memory Expansion

SSG-222B-NE3X24R

TABLE OF CONTENTS

Executive Summary 1
The Challenge of Scaling SQL Server Analytics Workloads 2
Introducing Intel Flat Memory Mode 3
A New Approach to Expanding Memory Capacity 3
Supermicro and Intel Bring a New Approach to Expanding Memory Capacity 4
Test Parameters and Workload Overview: Microsoft SQL Server Data Analytics Workload 4
Ideal Use Cases 7
Total Cost of Ownership Considerations. 7
Conclusion 8
For More Information 8

Executive Summary

As organizations increasingly rely on data-driven decision-making, SQL Server Online Analytical Processing (OLAP) workloads continue to grow in size and complexity. These workloads are read-intensive complex queries intended to perform advanced analytical calculations and trend analysis without disrupting transactional systems. Performance of these workloads is typically limited by memory capacity rather than compute power.

Intel and Microsoft collaborated on Microsoft SQL Server to optimize the database engine for the Intel Xeon 6 processors in FMM configurations. In addition, Intel and Supermicro collaborated to leverage Supermicro's X14-generation servers powered by Intel® Xeon® 6 processors. The unique combination

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

of Intel Xeon 6 processor performance, Supermicro's advanced design, cooling innovations, and rapid delivery model for its X14 servers created an opportunity for these long-standing partners to tackle this challenge and deliver a better customer solution.

To improve performance and easily increase memory capacity, Intel Flat Memory Mode (FMM), available on Intel® Xeon® 6 Processors, was leveraged as a new approach to memory scaling by aggregating DRAM and CXL-attached memory into a single, hardware-managed memory pool.

Intel and Supermicro evaluated the performance impact of using Flat Memory Mode with a Microsoft SQL Server OLAP workload, derived from the TPC-H benchmark on a Supermicro X14 enterprise server platform. (Note: This workload is derived from the TPC-H Benchmark and is not comparable to published TPC-H Benchmark results, as the implementation does not comply with all requirements of the TPC-H Benchmark). Expanding memory capacity and using FMM delivered substantial performance gains[1]:

- Up to 30% improvement in single query stream performance
- Up to 40% improvement in parallel query stream performance
- Up to 35% improvement in overall query performance

The Challenge of Scaling SQL Server Analytics Workloads

Memory-intensive workloads create unique performance challenges. OLAP workloads are uniquely characterized by:

- Large data sets that must be scanned and joined across complex queries
- Increasing numbers of concurrent users
- The need for predictable, low-latency response times

To run these workloads, maximizing the amount of data resident in memory is critical. However, DRAM capacity is often limited by cost and platform constraints. When memory is insufficient, systems rely more heavily on storage, leading to significantly higher I/O latency, slower query execution, and reduced throughput.

To expand capacity and share system memory, organizations can now leverage CXL memory, which allows devices to access it as if it were part of the system RAM. While software-managed memory-tiering solutions can be implemented to deliver this capability, they often require operating-system support, application awareness, or page-based data movement, which increases overhead and complexity.

Enterprises need a solution that can expand memory capacity without disrupting existing applications or operational models.

[1] See Appendix for workloads and configurations. Results may vary.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Introducing Intel Flat Memory Mode (FMM)

Introduced with Intel® Xeon® 6 processors featuring P-cores, Intel® Flat Memory Mode (FMM) provides a new, unique approach to memory scaling by aggregating DRAM and CXL-attached memory into a single, hardware-managed memory pool.

Prior generations of HW-managed tiering were legacy 2LM, a memory setup in which Near Memory (NM) functions as a cache for Far Memory (FM), but Near Memory capacity is not directly accessible or visible to applications.

Intel® Flat Memory Mode fundamentally differs from legacy 2LM. Instead of cache-based near/far semantics, FMM exposes the entire aggregated DRAM and CXL memory capacity directly to the operating system and applications as a unified address space. Memory placement and tiering are handled transparently in hardware, eliminating the need for software-defined tiering, OS policies, or application changes. This architecture enables near-DRAM performance characteristics, improved performance predictability, and reduced operational complexity when scaling memory capacity using CXL.

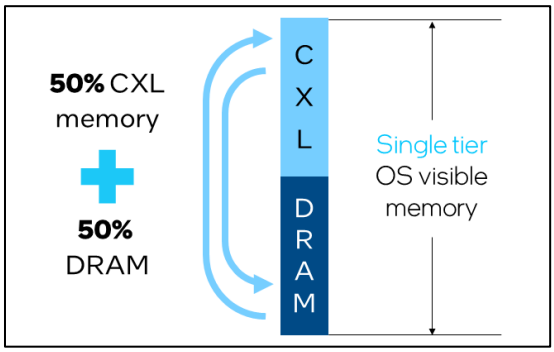

Figure 1 - How CXL Flat Memory Mode Works

A key differentiator of Intel Flat Memory Mode is that memory management occurs entirely in hardware within the processor. The processor dynamically manages the placement of hot and cold data between near memory (DRAM) and far memory (CXL), operating at cache-line granularity rather than at the page level, as in traditional software-managed tiered memory systems.

Since the operating system treats memory as a single unified pool, applications and databases such as SQL Server can immediately benefit from expanded memory capacity without requiring software awareness or modification. In fact, it can be enabled through a single step in the BIOS. This approach eliminates the overhead of software-based tiering solutions and enables systems to scale memory while maintaining performance close to native DRAM performance.

A New Approach to Memory Expansion

Unique characteristics of Intel Flat Memory Mode include:

- Unified Memory Pool: DRAM and CXL memory are aggregated and exposed to the operating system as a single memory address space.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

- Hardware Managed Data Movement: The processor manages the movement of hot and cold data between near memory (DRAM) and far memory (CXL).

Cache Line Granularity: Data movement occurs at cache line granularity rather than page granularity, reducing latency compared to software-based tiering.

- Transparency: Applications and the operating system see a single NUMA node and do not need to be modified or revalidated.
- No OS programming model dependency: Unlike software tiering solutions, FMM does not depend on specific operating system versions.

Supermicro and Intel Bring a New Approach to Expanding Memory Capacity

Supermicro's X14 servers are high-density, all-flash storage servers designed for extreme-performance workloads. It is specifically designed for scale-out all-flash NVMe storage environments, massive storage capacity, and low latency, all in a compact 2U form factor. These systems excel at petabyte-scale flash storage workloads, all while delivering ultra-fast access via PCIe 5.0 NVMe.

Test Parameters and Workload Overview: Microsoft SQL Server Data Analytics Workload

Analytical performance was assessed using a workload derived from TPC Benchmark™ H (TPC-H), a decision-support benchmark that measures how well a system handles complex business queries over large datasets. Designed to challenge systems with complex, ad hoc SQL queries that scan and join large volumes of data to answer business-critical questions. The benchmark workload consists of 22 parameterized queries that represent a supply chain-style data model, which reflects overall query-processing efficiency by accounting for database size.

The workload was generated using the MSTPCH benchmark tool (v. 3.0.1) with a scale factor of 10TB database, and measures both:

- Single-stream query performance
- Multi-stream performance with concurrent users

The overall query performance reflects both the speed of single-query execution and the concurrent query throughput. The workload is based on, but not fully aligned with, the official TPC-H benchmark specification. Consequently, the results should not be directly compared with any published TPC-H benchmark disclosures.

The Supermicro system tested is detailed in the Appendix.

Impressive Performance Improvement Results

Expanding memory capacity using Intel Flat Memory Mode delivered substantial performance gains[2].

[2] See Appendix for workloads and configurations. Results may vary.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

- Up to 30% improvement in single-query stream performance
- Up to 40% improvement in parallel-query stream performance
- Up to 35% improvement in overall query performance

Figure 2 illustrates the results.

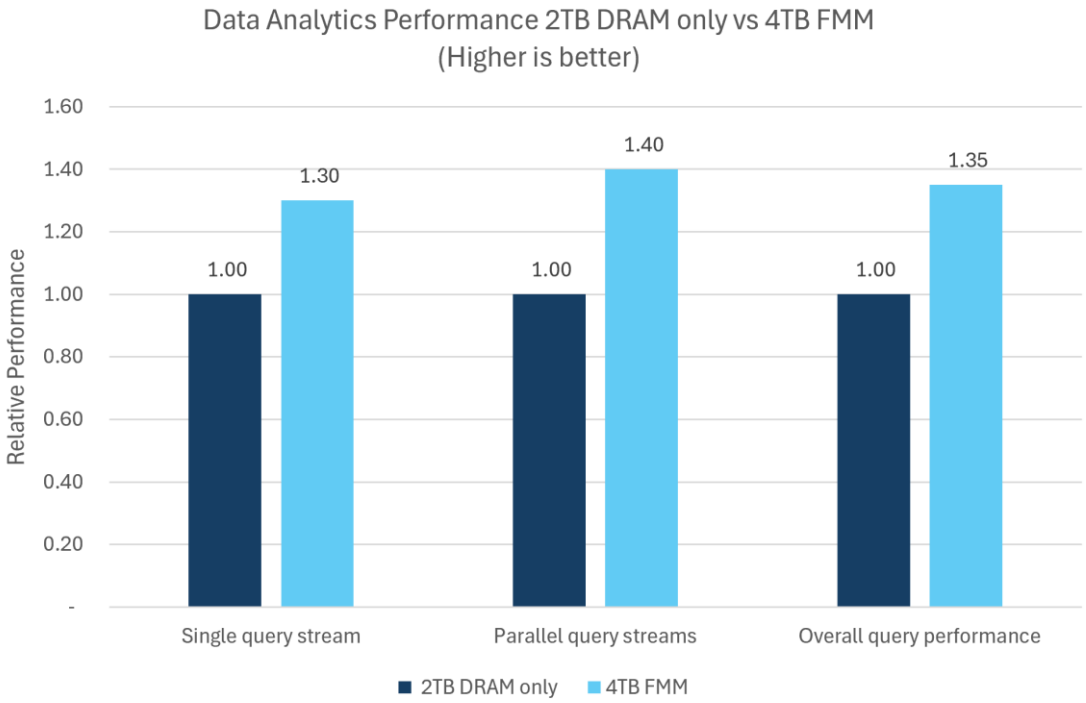

Figure 2 – Data Analytics workload with 10TB database size, measuring 2TB DRAM-only vs. 2TB DRAM and 2TB CXL with Intel FMM enabled. Higher is better.

In addition, Intel Flat Memory Mode demonstrated near-DRAM performance with minimal overhead for this workload.[3] as shown in Figure 3 below.

- 6% performance impact in single-query stream performance
- 2% performance impact in parallel-query streams performance
- 4% impact in overall query performance

[3] See Appendix for workloads and configurations. Results may vary.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

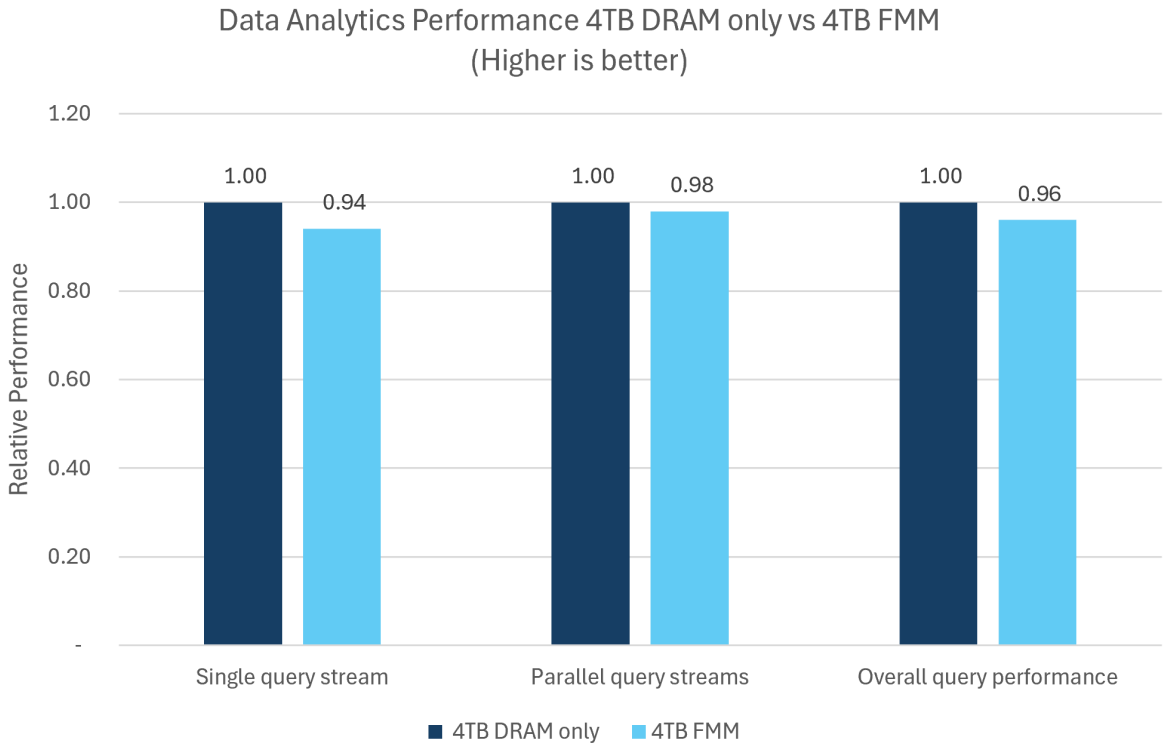

Figure 3 – Data Analytics workload with 10TB database size measuring 4TB DRAM-only vs. 2TB DRAM and 2TB CXL with FMM enabled. Higher is better.

These results demonstrate that increasing effective memory capacity directly improves both query execution speed and system concurrency for SQL Server OLAP workloads, with minimal loss in overall performance.

Why Did We Get These Results?

For memory-intensive workloads like OLAP, the primary driver of performance improvement is the ability to keep a much larger portion of the dataset resident in memory.
Key observations from the performance testing include:

- Over 90% reduction in I/O activity, significantly reducing storage-related stalls
- Most memory accesses are served from DRAM, as evidenced by near-memory cache read miss rates of approximately 10%.
- CXL bandwidth usage of approximately 14 GB/s, with demand read and read for ownership traffic around 6 GB/s, demonstrating a limited exposure to far memory latency

By efficiently managing far memory access and reducing reliance on storage, Intel Flat Memory Mode enables more consistent, higher-performance analytics.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

The Added Benefit of Simple Deployment

One of the key advantages of Intel Flat Memory Mode is that it is simple to deploy.

Once the CXL devices and DIMMs are provisioned, all you need to do is toggle Intel Flat Memory Mode in the BIOS to “enabled” (Socket configuration > Memory configuration > Memory Map > Intel Flat Memory Mode Support). Then the software can use this additional memory capacity just like DRAM.

Other CXL solutions would require software vendors to rewrite, recompile, and revalidate their software stack to use CXL memory.

After FMM is enabled:

- It is exposed transparently to the operating system, so the OS automatically sees the expanded memory
- Microsoft SQL Server uses the additional memory without modification
- There are no application changes, software tiering, recompilation, or revalidation required.

This further emphasizes that this technology is well-suited for production environments where stability and ease of deployment are critical.

Ideal Use Cases

Intel Flat Memory Mode is particularly well-suited for:

- Microsoft SQL Server OLAP and Data Warehouse (DW) workloads
- Decision support systems with large, growing datasets
- Environments constrained by memory capacity rather than CPU performance
- Organizations seeking a low-risk path to adopting CXL memory

Total Cost of Ownership Considerations

Beyond performance improvements, Intel Flat Memory Mode can also potentially help improve the total cost of ownership (TCO) for memory-intensive workloads.

Expanding system memory using CXL could allow organizations to scale memory capacity within a single server rather than deploying additional sockets or additional systems. This can provide several advantages:

- Reduced infrastructure complexity
- Better utilization of existing compute resources
- Lower power and rack footprint
- Simplified system management

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

For database workloads such as SQL Server, expanding memory within a single system may also reduce software licensing costs. Since many enterprise database platforms are licensed per core or per socket, increasing memory capacity within an existing server can allow organizations to process larger datasets without requiring additional licensed compute nodes.
Due to the current volatility in global memory pricing, exact cost comparisons may vary. However, the ability to scale memory capacity using CXL provides a flexible approach that can improve system economics for large analytics deployments.

Conclusion

Intel Flat Memory Mode with Intel Xeon 6 processors, deployed on Supermicro X14 enterprise servers, provides a practical and effective solution for scaling SQL Server analytics. By transparently expanding memory capacity using CXL, organizations can achieve significant performance improvements without modifying applications or software stacks.

The benchmark results demonstrate that memory expansion with Intel Flat Memory Mode combined with Supermicro X14 servers delivers faster query execution, higher concurrency, and reduced I/O, enabling enterprises to extract more value from their data while preparing for the next generation of memory-centric computing.

For More Information:

Supermicro SSG-222B-NE3X24R: https://www.supermicro.com/en/products/system/storage/2u/ssg-222b-ne3x24r

Intel Flat Memory Mode: https://www.intel.com/content/www/us/en/support/articles/000102525/processors/intel-xeon-processors.html

Appendix: Test System Configuration Details

Server: Supermicro SSG- 222B- NE3X24R

- Processors: Dual-socket Intel® Xeon® 6760P
- Cores: 64 cores per socket
- Total Threads: 256
- Last Level Cache: 320 MB

Memory Configurations Tested

- Baseline: 2 TB DDR5 DRAM, 4 TB DDR5 DRAM
- Expanded: 2 TB DDR5 DRAM + 2 TB CXL memory using Intel Flat Memory Mode

Software Stack

- Operating System: Windows Server 2025
- Database: SQL Server 2025 CU1
- SQL Server Configuration: Maximum server memory set to 90% of system memory

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

All other system parameters were held constant to isolate the impact of memory expansion.

Configurations:

Configurations		
Workload		SQL Server Data Analytics Workload
Database size		10TB
Processor		Intel Xeon 6760P
TDP (w)		330
Sockets		2
Cores per socket		64
Total LPs		256
LLC Cache (MB)		320
Core Frequency (all-core turbo) (GHz)		3.4
Uncore Frequency (GHz)		2.2
Platform		Supermicro SSG-222B-NE3X24R
BIOS Version		American Megatrends International, LLC. 1.5, 1/8/2026
Memory Capacity		2TB (64GB DDR5 6400 mt/s x 32) / 4TB (128GB DDR5 6400 mt/s x 32)
CXL Capacity		2TB (8x CXL devices x 256GB per CXL)
SQL max server memory		90% of system memory (sp_configure)
DIMM speed		6400 MT/s (throttles down to 5200MT/s at 2DPC)
Operating System		Windows Server 2025 10.0.26100*
SQL Server		SQL Server 2025 CU1
Storage	DB	8x 15TB NVMe (backup, tempdb, data)
	Tempdb	
	Backup	

**CXL support in the Windows Server Operating System is currently in Preview.*

Testing completed by Intel on Jan 26, 2026.

Intel technologies may require enabled hardware, software, or service activation.
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available security updates. No product or component can be absolutely secure.
Your costs and results may vary.
Performance varies by use, configuration, and other factors. Learn more at intel.com/performanceindex.
© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others."

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements. Visit www.supermicro.com

INTEL

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, visit

Visit www.intel.com

MICROSOFT

Microsoft (Nasdaq: MSFT) is a global technology leader focused on empowering people and organizations to achieve more. Guided by its mission to "empower every person and every organization on the planet to achieve more," Microsoft builds platforms, products, and services that help customers innovate, collaborate, and create trusted solutions across the cloud and beyond. To learn more about Microsoft, visit www.microsoft.com

SQL SERVER

Microsoft SQL Server is Microsoft's enterprise database platform—built to help organizations run mission-critical workloads with strong security, performance, and availability, from on-premises to the cloud. SQL Server is positioned as an "AI-ready enterprise database from ground to cloud," supporting modern application needs while enabling scalable, dependable data management across environments. To learn more about SQL Server, visit www.microsoft.com/sql-server

© 2026 Copyright Super Micro Computer, Inc. All rights reserved