PRODUCT BRIEF

SUPERMICRO AND AMD DEVELOP LARGE-SCALE VIRTUALIZATION ENVIRONMENTS FOR WORKLOAD CONSOLIDATION

SQL Server and Virtualization on a Single High-Performance Platform

Supermicro Server AS -2125HS-TNR

Supermicro Server AS -2126HS-TN

TABLE OF CONTENTS

Executive Summary 1
Challenges 2
Solution Overview 2
SQL Server Benchmarking Summary-High-concurrency OLTP 2
Testing Methodology 4
Results 5
Key Highlights 6
More Information 8

Executive Summary

Modern enterprise infrastructure is under increasing pressure as organizations scale mission-critical workloads such as transactional databases and large-scale virtualization environments. Traditional approaches that rely on simply adding more hardware are no longer sufficient. Instead, enterprises require systems that deliver predictable performance, efficient scaling, and optimized cost structures.

Supermicro, in collaboration with AMD and key ecosystem partners, provides a validated platform designed to address these challenges. Built on AMD EPYC™ 9005 high-frequency processors and Supermicro's H14 and Hyper server architectures, this solution enables organizations to consolidate workloads, improve performance and efficiency, and accelerate time-to-deployment. The platform is optimized for both SQL Server OLTP workloads and virtualization environments, making it a versatile foundation for modern data centers.

Challenges

Enterprise workloads are becoming increasingly complex and demanding. In database environments, organizations face growing concurrency requirements, rising licensing costs tied to core counts, and the need for consistent low-latency performance. At the same time, virtualization and cloud deployments must balance virtual machine density with predictable quality of service, all while managing operational overhead and infrastructure sprawl.

Physical infrastructure constraints, including limitations on power, cooling, and rack space, further compound these challenges. As systems scale, inefficiencies emerge, particularly at higher core counts where contention and diminishing returns become more pronounced. Enterprises are therefore shifting their focus from simply scaling up hardware to achieving efficient, predictable scaling.

Solution Overview

The Supermicro platform powered by AMD addresses these challenges through a unified architecture that supports both database and virtualization workloads. At its core, the solution leverages dual-socket AMD EPYC 9005 processors, particularly the high-frequency SKUs, which are designed to deliver strong per-core performance and low-latency response for transactional workloads.

In SQL Server OLTP environments, the platform demonstrates strong, predictable scaling as core counts increase. Testing shows that throughput can scale to approximately 2.5x on the same system architecture using EPYC processors with different core counts and high frequencies, compared to a baseline EPYC 16-core-powered system, with the most efficient scaling typically achieved before reaching the highest core counts.[1] This reflects real-world database behavior, where resource contention and system overhead introduce non-linear scaling at higher densities.

For virtualization workloads, validated VMmark results highlight the platform's ability to support dense virtual machine environments while maintaining consistent quality of service. Infrastructure operations such as live migration and workload balancing are executed reliably, demonstrating the system's readiness for enterprise and cloud-scale deployments.

Together, these results confirm that the platform is not only capable of delivering high performance but also maintaining stability and predictability under mixed and dynamic workloads.

SQL Server Benchmarking Summary - High-concurrency OLTP Transaction

This evaluation demonstrates that SQL Server 2022 OLTP throughput scales strongly across AMD EPYC 9005 High Frequency options on the Supermicro AS -2125HS-TNR platform under a consistent test environment.

Scaling Summary

To contextualize scaling, the table below compares measured throughput to ideal linear scaling (based strictly on total core count). Scaling efficiency declines with higher core counts, as expected in real-world OLTP systems, due to contention for shared resources and database engine overhead.[1]

The table below summarizes OLTP scalability results for Microsoft SQL Server 2022 running a TPC-E derivative workload on a Supermicro dual-socket platform using AMD EPYC 9005 Series High Frequency processors. Across four CPU options (8 to 32 cores per socket), the system demonstrates increasing throughput with core count, while scaling efficiency declines relative to ideal linear scaling. [1]

- Measured throughput increases from 1.0× (EPYC 9015, 16 total cores) up to ~2.55x (EPYC 9375F, 64 total cores) on the same server design, corresponding to a scaling efficiency of 63.7% relative to linear scaling. [1]
- The results follow the three-phase scaling pattern typically seen in database systems.
- Phase 1 — Linear scaling (16 → 32 cores), minimal contention.
- Phase 2 — Soft saturation (32 → 48 cores), shared resources start contending.
- Phase 3 —Contention (48 → 64 cores)

Processor	Cores/Socket	Total Cores	Measured (Normalized)	Efficiency
EPYC 9015 (2S)	8	16	1.00x	100.0%
EPYC 9175F (2S)	16	32	1.71x	85.6%
EPYC 9275F (2S)	24	48	2.29x	76.5%
EPYC 9375F (2S)	32	64	2.55x	63.7%

Throughput Behavior

The platform shows strong throughput increases as the number of cores increases, with the highest-throughput configuration delivering up to ~2.55x the baseline performance. [1]

Measured performance does not scale linearly with increasing core counts. This is typical for OLTP workloads where contention (locks/latches), log and checkpoint activity, and shared memory/IO paths introduce non-linear overheads.
The largest throughput gain occurs when moving from 16 to 48 total cores (~2.29x), after which incremental increases to 64 total cores yield diminishing returns (~2.55x). In practice, the 48-core or 64-core options may be selected based on workload priorities, such as maximizing throughput, or on other factors like SQL Server per-core licensing, which were outside the scope of this performance evaluation.

Why High Frequency Matters

SQL Server workloads frequently include short, latency-sensitive transactions that benefit from high sustained clock speeds, fast cache access, and efficient inter-core communication. High-Frequency EPYC 9005 processors are designed to deliver higher per-core performance, improving responsiveness under heavy concurrency.

Practical Deployment Guidance

- If the primary goal is maximum throughput on a single node, the 64-core (2S) EPYC 9375F configuration delivered the highest normalized throughput in this study.
- If the goal is a balanced point between throughput and scaling efficiency, the 48-core (2S) EPYC 9275F configuration achieved most of the scaling benefit before diminishing returns increased.
- For deployment, maintain consistent BIOS and SQL Server configuration practices (Soft-NUMA and NUMA awareness) to preserve predictable scaling.

Test Architecture Overview

The evaluation used a consistent test environment across all runs. The only component changed between test points was the processor OPN.

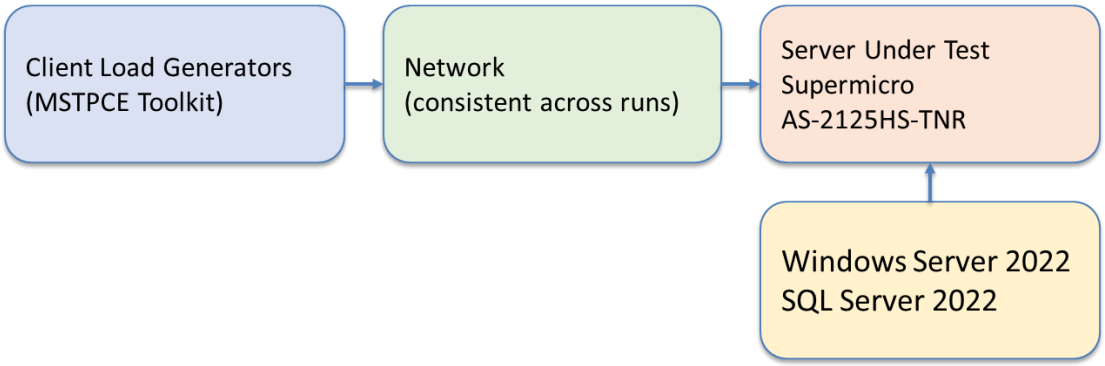

Figure 1 - Test environment overview (consistent storage, memory, and network across runs).

Testing Methodology

TPC-E DISCLOSURE

This benchmark is derived from the TPC-E™ benchmark specification. Results are not comparable to published TPC-E results. The benchmark was not audited and does not comply with all requirements of the TPC-E specification. The Transaction Processing Performance Council (TPC) has not reviewed or approved these results. The benchmark was configured to maximize CPU utilization. A pre-built database backup containing 4.8 million customers was restored for each run.
For each processor option, the workload was executed three times; results below represent the average of the three runs and are normalized to the dual-socket EPYC 9015 result (set to 1.0).

Single-run procedure:

1. Install the target processor OPN in the server.
2. Verify BIOS settings are optimized for performance.
3. Configure SQL Server Soft-NUMA based on the CPU's L3 cache/compute layout.
4. Restore the database from backup.
5. Verify memory and database settings are optimal for the workload.
6. From the client, start connection/user processes.
7. From the client, start regular checkpoints (every 7.5 minutes).
8. Start the OLTP workload and monitoring (BenchCraft).

9. Run for 90 minutes steady-state.
10. Stop workload and generate reports.

Tested System Configuration

- Server Model: Supermicro AS -2125HS-TNR
- CPUs: 2x AMD EPYC 9015, 9175F, 9275F and 9375F
- Memory Configuration: 1.5 TB total system memory; 24x 64 GB DDR5 6000 MT/s Micron DIMMs
- Storage:
    - O/S: 2x Micron 440 GB M.2 NVMe
    - SQL Server Logs: 1x Samsung 3.84 TB NVMe PCIe Gen5 drives.
    - SQL Server Data: 7x Samsung 3.84 TB NVMe PCIe Gen5 drives.
    - Network: 1x Dual Port Mellanox ConnectX-6 100 Gbps NIC

Load Levels by Processor Option

User load was increased with core count to keep the server CPU-bound. The table below summarizes the client-side configuration used to drive load.

	2×9015	2×9175F	2×9275F	2×9375F
Total Cores (2S)	16	32	48	64
#CU and ME Servers on Client	4	32	16	16
#Users per CU Server	80	20	80	80
Total User Count	320	640	1280	1280

Results

The chart below shows normalized OLTP throughput for each CPU option. Normalization baseline is the dual-socket EPYC 9015 configuration (16 total cores).

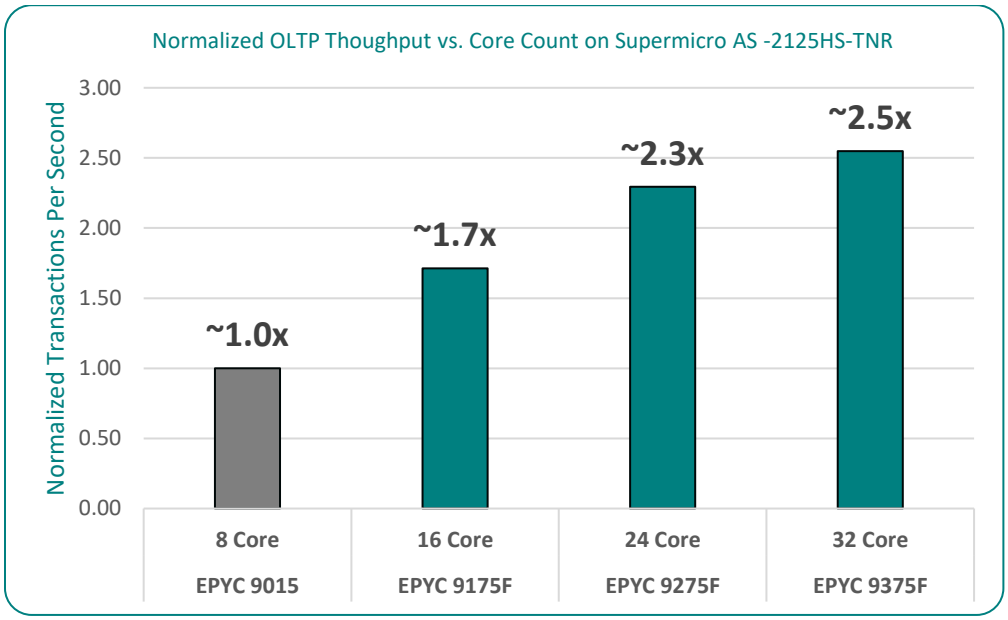

Figure 2 - Normalized OLTP throughput vs. core count.

The throughput values shown are internally defined, normalized transaction rates and do not represent tpsE or any official TPC metric.

Business Value

- Higher throughput on a consistent platform enables consolidation of multiple smaller database servers into fewer, higher-density nodes.
- Better throughput-per-server can reduce rack footprint and operational overheads (power, cooling, management).
- Optimizing the CPU choice for the required throughput target can help balance performance against per-core licensing and infrastructure cost.

VMmark Benchmarking Summary - Production reality: workloads + operations

The Supermicro AS -2126HS-TN platform, next gen of AS -2125HS-TNR, demonstrates compelling virtualization and cloud-scale performance. Across multiple VMmark runs, the results underscore the system's ability to handle dense virtualized workloads with robust infrastructure operations and consistent quality of service. For organizations running enterprise virtual machines, private or hybrid cloud infrastructures, or large-scale virtualization farms, this platform delivers high performance across varying load levels.

VMmark Benchmarking Summary

VMware validated the platform with several VMmark runs; key results include:

Date	Test Version	System	CPU	Score
2025-03-04	VMmark 4.0.1	AS -2126HS-TN	2 node, 2x AMD EPYC 9755	4.37 @ 5 Tiles VMware

2025-04-29	VMmark 4.0.2	AS -2126HS-TN	2 node, 2x AMD EPYC 9555	3.4 @ 4 Tiles VMware
2024-11-26	VMmark 4.0.1	AS -2126HS-TN	2 node, 2x AMD EPYC 9375F	1.86 @ 2 Tiles VMware

A tile is a repeatable unit of workload consisting of multiple VMs (A mini data center workload bundle):

Each tile includes: a Web server (e.g., OLTP-like workload), a Database VM, an Application server, a File server, a NoSQL workload, and a Social network workload.

VMmark scaling model - Instead of increasing load inside a VM, VMmark adds more tiles (horizontal scaling)

Tiles	Meaning
1 tile	baseline workload
2 tiles	2× workload
5 tiles	large-scale enterprise

Key Highlights

- A total score of 1.86 across two tiles translates to approximately 0.93 performance per tile, indicating strong and consistent scaling efficiency. This shows the platform can expand workload capacity while maintaining stable performance.
- Importantly, the system not only delivers throughput but also maintains quality of service under load, ensuring applications remain responsive even in highly virtualized environments. The virtualization workload breakdown shows strong performance across varied VM types (Auction, DVDStore, NoSQLBench, SocialNetwork), with a geometric mean improvement factor of ~1.03 in one tile. The ability to perform infrastructure operations efficiently under active workloads confirms that the platform is well-suited for cloud and hyper-converged environments, where continuous workload movement and scaling are required. Infrastructure operations (vMotion, Storage VMotion, XvMotion, Deploy) completed successfully, demonstrating infrastructure robustness.
- With its high memory bandwidth, flexible I/O capabilities, and next-generation expansion support, the platform can support a wide range of mixed enterprise workloads, from virtual desktops to business applications and private cloud deployments.
- Finally, the platform is built for enterprise reliability, with redundant power, advanced cooling, and integrated management capabilities that support continuous, 24/7 operation. Combined with flexible storage and networking options, it can serve as a foundation for both traditional virtualization and hyper-converged infrastructure deployments.

Consideration & Best-Practice Insights

- While the VMmark tests are strong indicators, as always, real-world performance will depend on workload mix, storage subsystem, network architecture, and VM sizing.
- Ensure memory configuration and BIOS/NUMA settings are optimized for virtualization workloads to realize full memory bandwidth benefits (EPYC architectures benefit from proper NUMA and memory tuning).

- Cooling and power: with high core counts and dual processors, ensure sufficient cooling and power infrastructure — the spec supports CPUs up to 500W TDP under certain conditions.
- Storage/backplane architecture: Given the high front-hot-swap drive count, use of NVMe or high-speed SAS may maximize I/O performance, and VMmark results assume a robust storage subsystem (see disclosure details).
- Investigate hypervisor version, driver support, and guest VM configuration: The disclosed runs used ESXi 8.0 Update 3 in all three cases. Matching or improving on that stack may yield equivalent results.

Recommendations

For organizations seeking a virtualization- and cloud-ready platform, the Supermicro AS -2126HS-TN, equipped with dual AMD EPYC processors and validated via VMmark, presents a compelling proposition. It is especially suitable for:

- Private cloud / multi-tenant virtualization environments
- VDI (virtual desktop infrastructure) at scale
- Software-defined storage or hyper-converged infrastructure (HCI) nodes
- Mixed workload consolidation (business apps + containerized services)
- Future-oriented compute nodes with expansion for GPU, CXL, and high-speed networking

To maximize return on investment:

1. Deploy with memory and storage configurations that match your workload's I/O and data locality needs.
2. Leverage the high core-count and threads with appropriately sized virtual machines (avoid under-utilizing the platform).
3. Take advantage of the PCIe 5.0 expansion and AIOM slot for networking or accelerator cards to future-proof your infrastructure.
4. Validate with your workload mix and perform a pilot to confirm quality of service under real operational workloads.

System Under Test

Server Platform: AS -2126HS-TN

- Dual-socket (Socket SP5) supporting AMD EPYC 9005/9004 series processors
- 24 DIMM slots (1DPC) supporting up to 6 TB DDR5 memory at up to 6400 MT/s (for EPYC 9005 series) or up to 4800 MT/s for EPYC 9004 series.
- 24 hot-swap 2.5" NVMe/SATA/SAS drive bays front-accessible, enabling high I/O and storage flexibility.
- Flexible expansion: up to four PCIe 5.0 x16 slots (or eight PCIe 5.0 x8), plus an AIOM (OCP 3.0) networking slot, and CXL 2.0 device support.
- Redundant power supply options (Titanium-level efficiency) and enterprise cooling design, making it suitable for high-density datacenter applications.
- AMD EPYC processors (e.g., the EPYC 9375F in one result) with high core counts, thread counts, and modern features (DDR5, PCIe 5.0, robust memory bandwidth).
    - Example: EPYC 9755F – 128 cores / 256 threads, base 2.7 GHz, boost up to 4.1 GHz, 12-channel DDR5 support, up to 6400 MT/s memory speed, TDP 320 W.

For More Information

Supermicro AMD Servers: https://www.supermicro.com/en/products/aplus

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

AMD

For more than 50 years AMD has driven innovation in high-performance computing, graphics and visualization technologies. Billions of people, leading Fortune 500 businesses and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work and play. AMD employees are focused on building leadership high-performance and adaptive products that push the boundaries of what is possible.

Learn more at www.amd.com

End Notes

[1] 9xx5-288: AMD Internal Testing as of 12/10/2025 based on production systems running MSSQL TPC-E Derivative Workload. The MSSQL TPC-E workload is derived from TPC-Benchmark™ Standard, and as such is not comparable to published TPC-E™ results.

MSTPCE.1.14.0-1048:
2P AMD EPYC 9015 production system, 16 total cores, 1535.57 GiB DDR5-6400, Microsoft Windows Server 2022 Standard 10.0.20348 Build 20348, BIOS 3.6
2P AMD EPYC 9175F production system, 32 total cores, 1535.54 GiB DDR5-6400, Microsoft Windows Server 2022 Standard 10.0.20348 Build 20348, BIOS 3.6
2P AMD EPYC 9275F production system, 48 total cores, 1535.56 GiB DDR5-6400, Microsoft Windows Server 2022 Standard 10.0.20348 Build 20348, BIOS 3.6
2P AMD EPYC 9375F production system, 64 total cores, 1535.56 GiB DDR5-6400, Microsoft Windows Server 2022 Standard 10.0.20348 Build 20348, BIOS 3.6

CPU EPYC 9015 EPYC 9175F EPYC 9275F EPYC 9375F
Metric 3257.22 5565.38 7475.68 8307.54
Normal 1 1.713 2.294 2.548

TPC, TPC Benchmark, and TPC-E are trademarks of the Transaction Processing Performance Council.