



VMware Memory Tiering on Lenovo ThinkAgile VX V4 Solution Brief

Enterprise workloads, from virtual desktops and databases, keep demanding more memory, yet DRAM remains one of the most expensive resources in the data center. As core counts climb with each processor generation, memory capacity and cost, not compute, increasingly cap how many virtual machines a host can support. Market pricing makes the tradeoff acute: server DRAM prices rose more than 90 percent quarter over quarter in early 2026, constrained DDR5 supply pushed per-node prices up roughly 1.7 times, and lead times for memory-dense servers stretched to between 4 and 16 weeks.

VMware Cloud Foundation 9.1 (VCF 9.1) addresses this directly with a production ready Memory Tiering capability. By combining DRAM with NVMe flash under a software defined memory scheduler, VCF 9.1 expands addressable memory without new DIMMs. Validated on Lenovo ThinkAgile VX V4 nodes powered by Intel Xeon 6 processors, the solution turns dense, all NVMe hyperconverged hardware into a platform that doubles virtual machine density while holding performance close to an all DRAM baseline.



Figure 1. VMware Cloud Foundation 9.1 is validated on the ThinkAgile VX650 V4 (top) and VX630 V4 (bottom)

Business Challenge

The problem is not simply that applications need more memory; it is that the memory they need has become dramatically more expensive. DRAM is costly, and infrastructure teams are under constant pressure to do more with the same budget. Traditional memory overcommitment techniques introduce problems of their own:

- **Unpredictable performance:** ballooning, compression, and swapping reclaim memory at the cost of latency spikes that are difficult to plan around.
- **Ongoing administrative effort:** overcommitment requires continual tuning and monitoring to avoid degrading production workloads.
- **Overprovisioned capacity:** to stay safe, teams size DRAM to each virtual machine's peak, leaving memory underused during off peak periods while paying for capacity that is rarely needed.
- **Balancing Memory Expansion with CPU Utilization Efficiency:** CPU resources in existing servers often remain underutilized due to memory constraints, but DRAM upgrades are typically available

only in large increments 512 GB/1TB. This creates a trade-off where adding memory to unlock CPU capacity can itself result in excess, underutilized memory.

The resulting problem is a familiar one: expensive, underutilized hardware and imbalanced expansion to keep pace with workload growth. It is a safe approach, but an inefficient and costly one that limits how far each server refresh can stretch.

Solution

VMware Memory Tiering in VCF 9.1 introduces enhancement in software defined memory scheduler that runs transparently beneath the virtual machine layer. It continuously classifies memory pages as hot or cold and places each on the appropriate tier:

- **Tier 0 (DRAM):** high speed system memory that holds frequently accessed, active pages for low latency access.
- **Tier 1 (NVMe SSD):** high-performance flash that holds infrequently accessed pages at a fraction of the cost of DRAM.

At the default 1:1 ratio, a node with 1 TB of DRAM can present roughly 2 TB of addressable memory to virtual machines, effectively doubling capacity without adding a single DIMM. Lenovo ThinkAgile VX is purpose built for this architecture, offering up to 8 TB of DDR5 across 32 DIMM slots and 24 onboard PCIe 5.0 NVMe ports with 1:1 lane mapping and no oversubscription. The Micron 7600 MAX PCIe Gen 5 SSD combines 3 DWPD endurance with low latency and exceptional mixed random read/write performance, making it ideal for memory-tier and cache-acceleration deployments. Memory Tiering is transparent to the guest operating system and applications, is tunable per cluster, host, or virtual machine, and remains DRS and vMotion aware, with Intel VMD and PCIe Gen 5 NVMe providing the low latency secondary tier. Lenovo enables memory tiering on both ThinkAgile VX V4 servers powered by Intel Xeon 6 processors and ThinkAgile VX V3 servers powered by AMD EPYC processors.

With VCF 9.1, you get:

- **Better performance:** VDI and database workloads show linear scaling for use cases with active memory below 50% of DRAM.
- **Software NVMe mirroring:** native mirroring built into vSphere removes the need for a RAID controller, a natural fit for the direct NVMe ports on the VX V4.
- **Simplified configuration:** a single vSphere Configuration Profiles workflow with no host reboot cuts setup time by more than half.
- **Enhanced observability:** new vCenter summary cards and a VCF Operations dashboard with What-If analysis model savings before tiering is enabled.
- **Broader compatibility:** security, low latency, and fault tolerant virtual machines, plus nested virtualization, are all supported.

Key benefit of NVMe tiering: The NVMe tiering solution helps to improve server efficiency and maximize the value of existing infrastructure investments by balancing CPU and memory usage optimally without impacting application performance.

Deployment and lifecycle management

Much of the Lenovo value comes from how simple Memory Tiering is to adopt. On Lenovo ThinkAgile VX V4, tiering is enabled through a single vSphere Configuration Profiles workflow that provisions the NVMe tier automatically and needs only maintenance mode, not a host reboot, so a cluster reaches a tiered configuration in minutes. The Lenovo VX Deployer guides cluster setup with validated firmware, drivers, and a tested bill of materials, and customers can size and order the exact configuration through the

Lenovo Data Center Solution Configurator (DCSC), and the Lenovo Top Choice Express program ships in-demand validated configurations in two weeks or less.

Getting started is low risk because assessment comes first. Before any hardware change, the Memory Tiering Assessment Tool (MTAT) and the VCF Operations What-If analysis let teams model whether NVMe tiering fits a given workload and estimate the DRAM-to-NVMe ratio and savings, so customers can confirm the approach is right for them before they deploy.

Once tiering is enabled, VCF Operations provides a dedicated Memory Tiering dashboard with per-VM tier bandwidth, latency, and consumed memory for day-to-day monitoring, while Distributed Resource Scheduler, high availability, and vMotion keep workloads balanced. Lenovo lifecycle services then cover deployment, support, and ongoing updates under a single point of contact, so teams gain the density and cost benefits without added operational overhead.

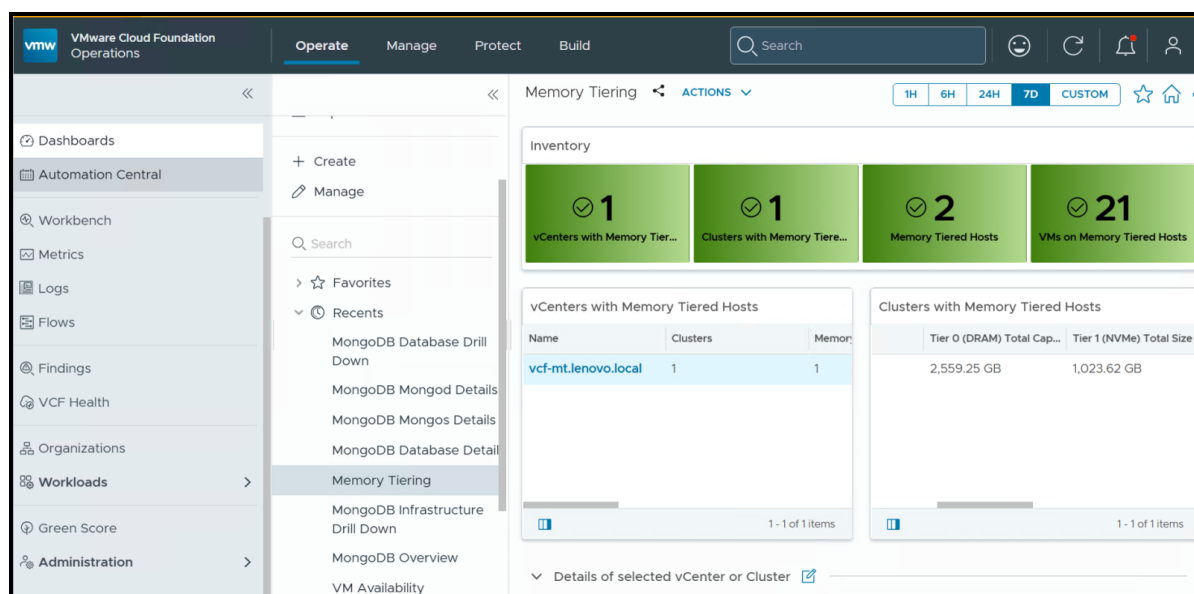


Figure 2. VCF Operations Memory Tiering dashboard

Use Cases

Memory Tiering delivers the most value for workloads that reserve far more memory than they actively use at any given moment. The scenarios below are the highest value deployment targets validated on VCF 9.1 and Lenovo ThinkAgile VX V4, spanning virtual desktops and databases.

- **Virtual Desktop Knowledge worker**

Knowledge worker virtual desktops typically require only 4 to 8 GB of memory and remain idle for extended periods between bursts of user activity. With Memory Tiering, VX V4 nodes keep active desktop memory pages in DRAM while transparently moving inactive pages to NVMe storage. This optimization enables up to **2x higher desktop density per node** without significantly impacting user experience. Login Enterprise testing at double the desktop density demonstrated only a **10% impact on end-user experience scores**, which remains well within acceptable thresholds for VDI deployments, delivering improved infrastructure efficiency and a better return on investment.

- **SQL Server OLTP Transactional**

SQL Server databases are designed to utilize available memory efficiently to maximize performance. However, the amount of active memory required for transaction processing varies based on database size, working set, transaction volume, the number of concurrent users, and workload intensity. Memory tiering with NVMe enables approximately **1.5x to 2x higher consolidation** for different OLTP scenarios on standalone deployment while maintaining acceptable throughput and latency tradeoffs. This approach helps maximize ROI by improving resource utilization across both horizontal and vertical scaling of SQL Server instances.

- **MongoDB NoSQL Memory Bound CRUD Operations**

MongoDB performance, when heavily memory-bound, can be optimized by using NVMe memory tiering to extend the effective capacity beyond physical DRAM. This approach keeps hot data in DRAM while tiering cooler data to NVMe, preventing severe, high-latency disk thrashing. The strategy allows for higher VM consolidation density and reduced infrastructure costs, while maintaining acceptable throughput for mixed CRUD workloads. The WiredTiger engine and OS cache utilize this extended memory to sustain performance, making it an effective, cost-efficient solution for memory-constrained environments.

Workloads

Lenovo and VMware validated Memory Tiering across a range of enterprise workloads on the ThinkAgile VX650 V4, a 2U dual socket platform with Intel Xeon 6700P or 6500P processors, up to 8 TB of DDR5, and 24 PCIe 5.0 NVMe bays. All results below are at twice the virtual machine density of an all DRAM baseline:

- **Virtual desktops:** Login Enterprise 5.14, 0 to 8 percent performance impact.
- **Databases:** HammerDB 5.0 on SQL Server 2022 at twice the density, with 0 to 5 percent performance impact; SQL Server and MongoDB are common Memory Tiering consolidation targets.
- **Sizing guidance:** use high endurance NVMe (3 DWPD or higher), keep host CPU utilization below 75 percent, and avoid virtual machines larger than 32 vCPUs or 512 GB in tiered configurations.

Memory Tiering is the strongest fit when host memory, not CPU, is the constraint and the active working set stays at or below roughly 50 percent of DRAM, so hot pages remain in Tier 0 and only latency-tolerant cold pages move to NVMe. It is not recommended for latency-critical workloads where nearly every page is hot, or where the active working set exceeds available DRAM.

Test Results

Lenovo and Intel engineering teams validated VCF 9.1 Memory Tiering on Lenovo ThinkAgile VX V4 across three representative workloads, each tested at double the virtual machine density of a DRAM-only baseline using the default 1:1 DRAM-to-NVMe ratio.

Table 1. Summary of validated Memory Tiering test results

Workload	Benchmark	Density gain	Result with NVMe tiering
Knowledge worker VDI	Login Enterprise	2x desktops per node	EUX stayed above the 7.6 satisfactory threshold (7.9 to 8.7)
SQL Server 2025 OLTP	HammerDB (TPC-C style)	1.5x to 2x consolidation	80 to 90 percent of DRAM-only throughput (TPM and NOPM)
MongoDB CRUD	YCSB	Approximately 2x consolidation	76 to 82 percent of DRAM-only throughput

Across all three workloads, Memory Tiering roughly doubled consolidation while retaining the majority of performance, with host CPU utilization staying within recommended ranges. Knowledge worker VDI showed the smallest impact, since idle desktop memory tiers efficiently to NVMe. Full hardware

configurations, workload parameters, and per-scenario analysis are available in the companion Lenovo reference architecture.

Business Outcome

Deploying VCF 9.1 Memory Tiering on ThinkAgile VX V4 produces measurable results across cost, capacity, and efficiency. Because NVMe costs roughly 4 to 6 times less per GB than DDR5 DRAM, doubling addressable memory in software rather than hardware drives most of the savings:

- **Up to 45 percent** lower infrastructure total cost of acquisition for the same virtual machine capacity.
- **2x** virtual machine density per host with more than 90 percent performance retention.
- **Up to 50 percent** fewer physical servers by consolidating twice the virtual machines per node.
- **More than 50 percent** faster Memory Tiering configuration in VCF 9.1, with no host reboot required.

As a published VMware example, reaching 2 TB of effective memory per node with tiering, using 1 TB of DRAM plus a 1 TB NVMe tier, costs roughly 205,000 dollars compared with about 318,000 dollars for an all-DRAM 2 TB node. That is the same effective capacity for close to 40 percent less, before the additional savings in power, cooling, and rack space from deploying fewer servers.

In sustained testing, active memory stayed below 50 percent of DRAM capacity and NVMe read latency held between 100 and 200 microseconds, within VMware's recommended range. Published figures are based on Lenovo and VMware benchmark testing; actual results vary by workload and configuration.

Conclusion

VCF 9.1 Memory Tiering on Lenovo ThinkAgile VX V4 with Intel Xeon 6 lets organizations double virtual machine density and cut infrastructure cost by up to 45 percent while keeping performance close to an all-DRAM baseline. With software NVMe mirroring, single workflow configuration, and deeper observability, Memory Tiering is ready for production across virtual desktops, databases, and enterprise applications. Lenovo delivers factory integrated VX V4 systems preconfigured for VCF 9.1, so teams can adopt tiering with validated firmware, drivers, and support from day one.

For More Information

To learn more about VMware Memory Tiering on Lenovo ThinkAgile VX V4, contact your Lenovo representative or Lenovo Business Partner, or visit the resources below.

References:

- Lenovo ThinkAgile VX Series:
<https://www.lenovo.com/us/en/servers-storage/hyperconverged/thinkagile-vx/>
- Lenovo Press: ThinkAgile VX Series product guides:
<https://lenovopress.lenovo.com/servers/thinkagile/vx>
- VMware Cloud Foundation:
<https://www.vmware.com/products/cloud-infrastructure/vmware-cloud-foundation>
- Micron 7600 Data Center SSD:
<https://www.micron.com/products/storage/ssd/data-center-ssd/7600-ssd>
- NAND Research, Memory and NAND Flash Crisis (May 2026):
<https://nand-research.com/memory-nand-flash-crisis-may-2026-update/>
- Server Space, Why Server Hardware Is Getting More Expensive in 2026 (August 2026):
<https://serverspace.us/about/blog/why-server-hardware-is-getting-more-expensive-in-2026-and-what-you-can-do-about-it/>

Authors

Markesha Parker is the WW Technical Leader for ThinkAgile VX Hyper-Converged solutions at Lenovo. In this role, she works to deliver quality solutions integrating Lenovo and third party hardware and software. She defines the technical requirements to enable manufacturing, support, and technical sales. Markesha has worked in the IT industry for over 17 years and is currently based in Morrisville, NC.

Chandrakandh Mouleeswaran is a Solution Architect with 18+ years of experience in software development, performance testing and engineering, having worked on designing and architecting many scalable enterprise applications. He has spent a decade in technical enablement and partner solution development for VMware, Nutanix, Oracle and other ISVs across industries and technologies. He specializes in architecting infrastructure solutions for virtualization, VDI, database, cloud, data science, AI/ML solutions and various enterprise workloads.

Cristian Ghetau is an Advisory Engineer for Lenovo in Romania and has experience in Cloud Infrastructure technologies. He has had more than 13 years of experience working with virtual environments from VMware, Microsoft, Oracle, and Linux.

Catherine Maina is a Senior Product Manager for ThinkAgile VX at Lenovo, where she leads portfolio strategy and market direction for VMware based hyperconverged and software defined infrastructure platforms. She works across engineering, sales, marketing, and ecosystem partners to shape co engineered solutions that support modern hybrid cloud architectures, balancing scalability and operational simplicity for enterprise customers.

Chris Honoré is a Solutions Product Manager at Lenovo with deep expertise in datacenter products and solution offerings. He has a strong background in consulting and solution development, helping customers design and support on-premises and hybrid environments. Chris has spent the past 15 years with IBM and Lenovo, specializing in x86 server and data center solutions. Prior to that, he built two decades of experience in the telecommunications industry, serving in both technical and business leadership roles.

Related product families

Product families related to this document are the following:

- [ThinkAgile VX Series for VMware](#)
- [VMware Alliance](#)

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

© Copyright Lenovo 2026. All rights reserved.

This document, LP2518, was created or updated on August 28, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
<https://lenovopress.lenovo.com/LP2518>
- Send your comments in an e-mail to:
comments@lenovopress.com

This document is available online at <https://lenovopress.lenovo.com/LP2518>.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at <https://www.lenovo.com/us/en/legal/copytrade/>.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Lenovo®

ThinkAgile®

The following terms are trademarks of other companies:

AMD and AMD EPYC™ are trademarks of Advanced Micro Devices, Inc.

Intel®, the Intel logo and Xeon® are trademarks of Intel Corporation or its subsidiaries.

SQL Server® is a trademark of Microsoft Corporation in the United States, other countries, or both.

TPC® and TPC-C® are trademarks of Transaction Processing Performance Council.

Other company, product, or service names may be trademarks or service marks of others.