White Paper
Data Centers | Tiered Memory
June 2023

Tiered Memory in VMware's Production Tanzu Environment

VMware and Intel collaboration demonstrates how tiered memory with Intel® Optane™ persistent memory (PMem) enables up to a 66% reduction in the number of servers while increasing memory per server from 384 GB to 4 TB[1]

Author
John Hubbard
Solutions Architect

Table of Contents

Executive Summary

VMware Tanzu is a popular container platform—and VMware itself runs its own containerized applications on Tanzu. Like its customers that use Tanzu, VMware Platform Services (VPS) faces data center challenges—tight IT budgets, memory-hungry modern workloads, and outdated hardware. VMware recently collaborated with Intel to determine the viability of upgrading hardware to consolidate servers and use tiered memory to provide Tanzu's containerized workloads with more memory than the existing legacy hardware could support.

Following best practices developed by Intel for right-sizing tiered memory systems, VMware data center architects monitored real-world, production, containerized workloads running on Tanzu to understand average memory and CPU utilization. The memory metrics gathered for the legacy server environment indicated that VMware's Tanzu deployment was a good fit for a tiered memory system with Intel® Optane™ persistent memory (PMem). In a tiered memory configuration, Intel Optane PMem serves as main capacity memory and a small amount of 3,200 MT/s DRAM serves as a cache. Tiered memory enabled VMware to replace 27 legacy blade servers with nine newer 1U servers equipped with 3rd Gen Intel® Xeon® Scalable processors. As a result, per-server memory capacity increased from 384 GB to 4 TB, while lowering memory costs by up to 33%.[1,2]

In summary, VMware's deployment of tiered memory for their production Tanzu environment proves that Intel Optane PMem enables massive server consolidation—reducing the number of servers by as much as 66%—and provides vast amounts of memory for VMware Tanzu containers at an affordable $/GB.[2]

At a Glance

Using Intel® Optane™ persistent memory (PMem) with upgraded Intel® hardware running VMware Tanzu provides the following benefits:

- **Server consolidation.** IT departments can run their containers on fewer servers, decreasing data center footprint, reducing infrastructure complexity, and increasing efficiency.[3]
- **Vast increases in system memory.** Memory capacity can be increased far beyond what is physically or budgetarily possible with DRAM.[3]
- **More compute power per square foot.** Upgrading to Intel® Xeon® Scalable processors with more memory channels and enhanced per-core performance (compared to previous-generation processors) can support compute-hungry containers.[3]
- **Lower CapEx.** Tiered memory systems can reduce $/GB by up to 33%.[4]

Business Challenge: Scale Memory while Increasing Data Center Efficiency

Many data center architects face a perfect storm of challenges. Legacy hardware cannot keep up with the compute and memory demands of modern workloads. But a sluggish economy is causing IT budgets to remain flat or even shrink, making it difficult to convince stakeholders to purchase additional servers. The situation is complicated even further by the need to support not only traditional virtual machine (VM)-based workloads, but also containerized workloads.

The container ecosystem offers a number of choices. One choice is to use open-source Kubernetes, with or without VMware's suite of products such as VMware vSphere and vCenter. An alternative is to use VMware vSphere with Tanzu, which offers a VM Service functionality that enables DevOps engineers to deploy and run VMs, in addition to containers, in a common, shared Kubernetes environment. But even with Tanzu, data centers still need to scale memory while increasing data center efficiency by consolidating servers. Is it possible to scale memory while actually decreasing the number of servers needed in the data center? Intel has the answer: tiered memory using Intel® Optane™ persistent memory (PMem). See the sidebar, "A Closer Look at Tiered Memory" for a high-level explanation.

Determining the "Right Fit" for Tiered Memory with Intel® Optane™ Persistent Memory

Intel has developed a set of best practices for sizing a tiered memory system that can meet a data center's current and future scaling needs (such as over the next five years):

1. Use metrics, such as vSphere Memory Monitoring and Remediation (vMMR), to determine how memory is being used in the legacy (DRAM-only) environment. These measurements will reveal the active memory footprint as a percentage of consumed memory.
   a. Note: As is the case with statistics, larger datasets offer the most insight. Ideally, measure and record the active and consumed memory metrics for a week or more. It should be easy to identify the usage "highs and lows" throughout the week. Often, traffic patterns and behaviors can be hidden by averages; thus, ensure that the sample rate or granularity is set to no less than one minute.
   b. Note: Live Optics can also measure active memory over time, but the consumed memory metric is not available outside of vSphere. Instead, use the total physical memory to calculate the active memory footprint relative to total system memory. Consumed memory provides a measurement that more accurately represents the workload(s) at that moment. Therefore, when calculating the footprint with physical memory, do so with only 80% of the total physical memory to account for headroom.

$$\frac{\textbf{Active Memory}}{\textbf{80\% of Physical Memory}}$$

2. Active memory footprints less than 25% of consumed memory are a great fit for tiered memory.

$$\frac{\textbf{Active Memory}}{\textbf{Consumed Memory}}$$

3. Obtain CPU usage information as well, which can provide insight into trends and opportunities for consolidation.

Read Intel's Best Practices Guide for a deeper dive into these steps, such as an explanation of active versus consumed memory.

Tiered Memory in Action: VMware's Own Real-World Tanzu Environment

Working closely with Intel engineers, VMware data center architects investigated how Intel Optane PMem could benefit VMware's existing Tanzu production environment. VMware's legacy blade servers had limited memory capacity and bandwidth. With a maximum of 24 slots, 8 memory channels (4 channels per socket) and support for only 2,400 MT/s DRAM, VMware found it necessary to scale out servers to balance CPU, memory, and container workload growth. By moving to newer 1U servers with tiered memory and 8 memory channels per socket, VMware could achieve a more efficient balance.

vMMR measurements indicated that the legacy environment, with only 384 GB of DRAM per node, was characterized by the following (see Figure 1):

- About 300 GB of consumed memory (78% of memory was in use or not free for other processes and VMs)
- About 30 GB of active memory

That means that 10% of consumed memory is active, or less than 8% of the physical memory—making VMware's Tanzu environment a great fit for tiered memory.

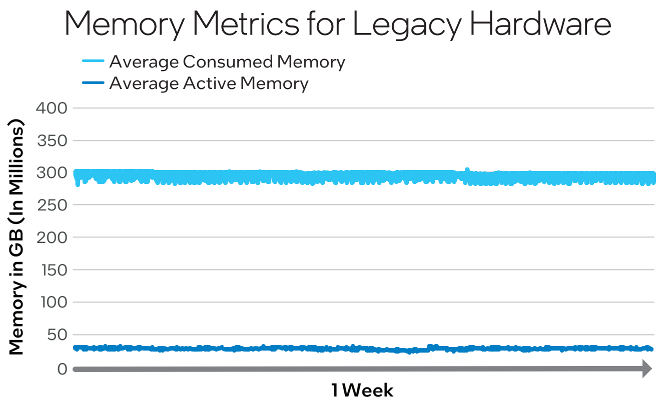

Figure 1. Memory metrics in VMware's DRAM-only Tanzu production environment show that active memory is only about 10% of consumed memory, making it an excellent candidate for tiered memory with Intel® Optane™ PMem.

By moving to newer 1U servers with tiered memory and 3rd Gen Intel® Xeon® Scalable processors, VMware achieved the following notable benefits (see Figure 2):[5]

- Replaced 27 older blade servers with just nine newer 1U servers—a 66% consolidation.
- Grew memory capacity per server from 384 GB to 4 TB—more than 10x greater capacity for memory-hungry containerized workloads.
- Reduced memory costs by up to 33% with tiered memory compared to DRAM-only systems.[6]
- Reduced infrastructure complexity by eliminating the storage area network and moving to VMware vSAN.

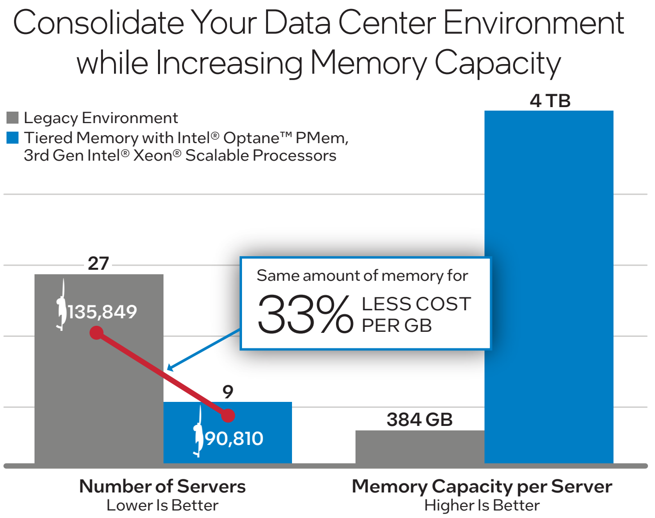

Figure 2. Intel® Optane™ PMem enables significant server consolidation while providing terabytes of memory per server and substantially lower memory costs.[7]

Memory monitoring in the upgraded environment is shown in Figure 3—less than 15% of the consumed memory is active, confirming that this system is still a good fit for tiered memory and is meeting predictions for performance. In other words, the active working dataset size (100 GB) easily fits in the DRAM cache, with 90% free. The consumed memory that is not active (that is, not as hot) is automatically stored on Intel Optane PMem.

Leveraging the vMMR "mem.missrate.latest" metric in VMware vCenter, we can confirm that tiered memory is doing its job to ensure the hottest data resides in cache with less than a 1% miss rate per node.

CPU usage comparison over the course of one week confirmed an increase in data center efficiency: 27 legacy systems with DRAM averaged about 20% CPU usage per node, while nine newer systems with tiered memory averaged about 30% CPU usage per node.

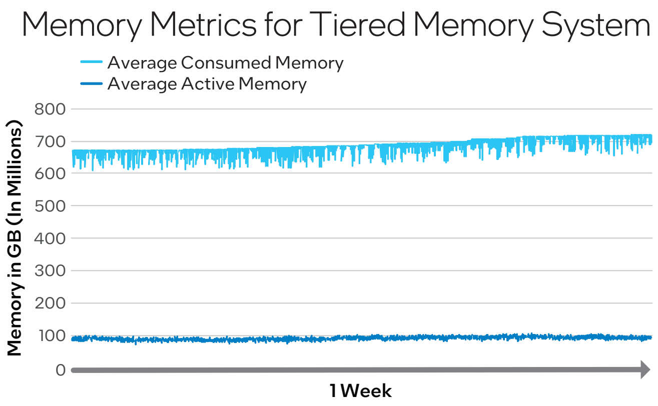

Figure 3. On newer servers with tiered memory, the active memory footprint is still well below the 25% threshold for sizing tiered memory systems.

Solution Architecture: Tiered Memory for Containerized Environments

Figure 4 illustrates the overall tiered memory architecture for the VMware Tanzu environment, augmented by a vSAN storage cluster. As a rule of thumb, Intel recommends a DRAM-to-PMem ratio of 1:4. For example, a 4 TB server would use 1,024 GB of DRAM and 4,096 GB of Intel Optane PMem.

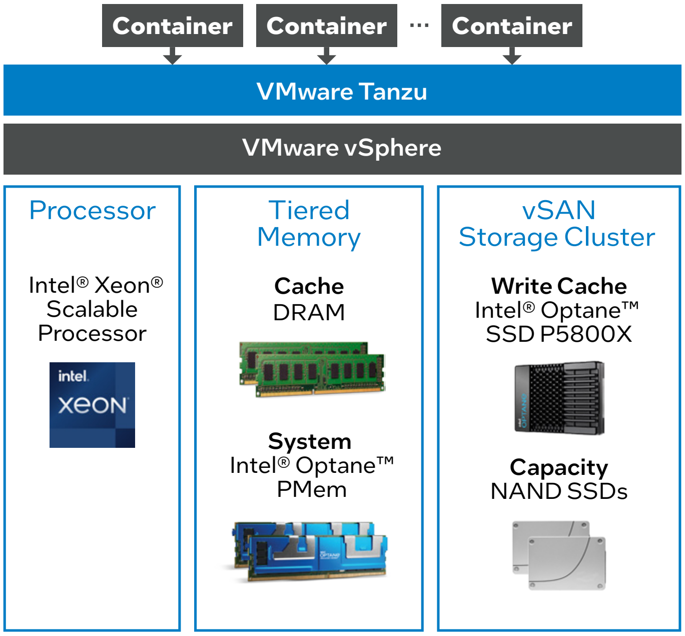

Figure 4. Intel® Optane™ PMem comprises main system memory (essentially a capacity tier), while a small amount of DRAM serves as a memory cache tier.

VMware Tanzu Use Cases

VMware Tanzu is a modular, cloud-native application platform that helps reduce the complexity of building, delivering, and operating cloud-native apps, whether they are on-premises, in the public cloud, or hosted in a hybrid environment. Here are some of the ways organizations are using VMware Tanzu[9]:

- Modernize existing mission-critical applications by re-platforming or re-architecting the software, while simultaneously enhancing security and resiliency.
- Streamline the developer experience so they can focus on creating new digital experiences.
- Improve the business' security posture through automated security and compliance.
- Monitor and manage the container environment from the data center to the cloud from a central hub.

Reduce Data Center TCO with Tiered Memory[8]

Three imperatives combine to frustrate data center architects: IT budgets are always tight. Workloads—especially artificial intelligence, machine learning and big data analytics—are increasingly memory hungry. And failing to meet service-level agreements due to outdated hardware isn't an option. A tiered memory system using Intel® Optane™ persistent memory (PMem) can help on all three fronts by providing system memory for less $/GB and lower overall system cost.

For example, consider a server equipped with 2x 3rd Gen Intel® Xeon® Scalable processors and 2,048 GB of DRAM per socket (4,096 GB total memory). On average, this memory subsystem would cost $135,849. In contrast, a tiered memory system (same processor) with 2,048 GB of Intel Optane PMem (as system memory) and 512 GB of DRAM (as cache) per socket (4,096 GB total memory) would cost only $90,810. In summary, you have the same amount of memory for 33% less $/GB.[7] Thus, tiered memory systems with Intel Optane PMem free up substantial budget dollars that can be used for other data center priorities.

Capacity Planning and a Familiar Story

VMware and Intel worked together to craft a new environment with capacity planning at the heart of the design. We can use the statistics from VMware's new vMMR environment to predict how much more room there is to grow for the current production workloads.

Today VMware's production environment is operating at 30% CPU utilization, 700 GB of consumed memory and 100 GB of active memory. Assuming the workload's behavior remained the same, how far could it scale in this new environment? Best practices dictate that it is wise to allow 20% headroom of any resource. Put differently, infrastructure planning should expect not to exceed 80% of a resource, whether it be CPU, memory, storage, network, etc.

- Following best practices and limiting CPU utilization to 80%, we can predict 1.8 TB of consumed memory, with 267 GB of active memory.
- Outside of best practices, driving CPU to a maximum of 100%, each node would theoretically consume 2.3 TB of memory with 333 GB of active memory.

But workloads do not always scale linearly, nor are all workloads created equal. More times than not, things change. So, while VMware's current production workloads have an average active memory footprint less than 15% of consumed memory, it is entirely possible that could change. For example, what if new workloads were introduced into the environment? What if VMware discovered a situation where a new application needed a last-minute home? What are the requirements? How much CPU and memory is needed? This wasn't part of the plan; where could this application land? Does this situation sound familiar?

The beauty of a correctly sized tiered memory system, such as the one VMware is now using for Tanzu, is that even at 100% CPU consumption, each node would have 1.7 TB of free memory (42%), with two-thirds of the DRAM cache unused. This means that this memory-abundant environment could easily accommodate more memory-hungry containers and VMs, assuming their CPU requirements were less than the containers operating in production today.

Balancing utilization across resources while ensuring headroom is an ongoing challenge. Fortunately, VMware introduced new algorithms into their Distributed Resource Scheduler (DRS) in vSphere 8.0. Leveraging statistics from vMMR, vSphere 8.0 and vCenter can make migration decisions based on memory utilization. This means that balancing memory usage across the cluster can be completely automated, even if only for initial placement. Assuming our earlier scenarios, VMware could easily collocate new workloads automatically. New DRS detection capabilities include:

- The host is running low on memory capacity.
- A VM is consuming excessive memory bandwidth.
- An increase in host DRAM cache miss rate is above recommended thresholds (10%).
- An increase in host Intel Optane PMem memory bandwidth has occurred (indicates rising DRAM cache misses).

Conclusion

Intel worked with VMware directly to demonstrate an effective example of sizing a tiered memory configuration for a production environment. The team of VMware and Intel IT experts followed best practices at every step (such as gathering memory monitoring metrics for both the old and new environments). By doing so, VMware now has a modernized and consolidated Tanzu cluster that supports more workloads for less capital outlay, compared to scaling DRAM alone. This robust, efficient environment has more compute power than the previous cluster. Fewer servers with more memory results in more budget and space for additional scaling as VMware's business grows. VMware's customers can gain similar benefits by exploring the memory and CPU usage of their Tanzu environments and putting tiered memory to work in their data centers.

A Closer Look at Tiered Memory

Intel® Optane™ persistent memory (PMem) is unique because it has characteristics of both memory and storage. In the default Memory Mode configuration, Intel Optane PMem does not require any software application changes and is transparent to end users. With tiered memory, a small amount of DRAM serves as a cache for the hottest data and is not seen by the containers or the OS as part of system memory. Main system memory consists of cost-effective Intel Optane PMem modules, which have the same form factor as a DRAM DIMM and can be installed into the same physical DIMM connectors on the memory bus. For most systems, on each memory channel, the DRAM DIMMs plug into the first DIMM slot (slot 0), and the Intel Optane PMem DIMMs into the second slot (slot 1). The Intel Optane PMem Best Practices Guide offers additional details.

Learn More

You may also find the following resources useful:

- Tiered Memory Can Boost Virtual Machine Memory Capacity and Lower TCO brief
- Intel® Xeon® Scalable processors product page
- Boost VMware vSphere Efficiency with Intel® Optane™ Persistent Memory best practices guide

For more details, contact your Intel representative or visit Intel® Optane™ technology.

[1] **Legacy cluster:** 27x Dell PowerEdge M630 Blade Servers, each with 2x Intel® Xeon® processor E5-2680v4 (32 cores, 2.4/3.3 GHz), 384 GB memory (12x 32 GB DRAM) – 4 channels/socket @ 2400 MT/s, SAN storage = QLogic QME2662 16 Gbps Fibre Channel, network = Intel® Ethernet Server Adapter X520 Dual Port 10 GbE.

Modernized cluster: 9x Dell PowerEdge R650 (rack 1U), each with 2x Intel Xeon Gold 6338 processor (32 cores, 2.0/3.2 GHz), 4 TB memory (16x 64 GB DRAM + 16x 256 GB Intel® Optane® persistent memory) – 8 channels/socket @ 3200 MT/s, vSAN storage = 1x Intel Optane SSD P5800X 800 GB + 4x Solidigm P5500 3.84 TB, network = Intel® Ethernet Network Adapter E810-XXVDA2 Dual Port 10/25 GbE.

[2] Pricing information obtained from Intel's Intel® Optane™ PMem TCO Calculator as of February 3, 2023:

DRAM-based solution: DDR4 DRAM cost estimate based on list pricing for DDR4 DIMMs integrated in OEM systems captured on February 3, 2023. DRAM-only memory subsystem cost: $135,849.

Default option uses an average price across 3 OEMs: HPE list price, Lenovo list price, Dell list price.

Intel® Optane™ Persistent Memory and CPU Pricing are based on Intel RCP available on www.ark.intel.com. Tiered memory cost: $90,810.

Intel® Optane™ PMem pricing shown is provided for guidance and planning purposes only and does not constitute a final offer. Pricing guidance is subject to change and may revise up or down based on market dynamics. Please contact your OEM/Distributor for actual pricing.

[3] See endnote 1.

[4] See endnote 2.

[5] See endnote 1.

[6] See endnote 1 for consolidation and memory capacity information. See endnote 2 for pricing information.

[7] See endnote 1 for consolidation and memory capacity information. See endnote 2 for pricing information.

[8] See endnote 2.

[9] VMware, "VMware Tanzu Solution Overview."

Performance varies by use, configuration, and other factors. Learn more at intel.com/PerformanceIndex. Performance results are based on testing by VMware as of October 28, 2022 and may not reflect all publicly available security updates. See configuration disclosures for details. No product or component can be absolutely secure. Your costs and results may vary. Intel technologies may require enabled hardware, software, or service activation. 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.