What’s New in Performance for VMware vSphere 7?

Performance Study - October 5, 2021

Updated for vSphere 7.0 U1, U2, and U3

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Table of Contents

Introduction

Underlying each release of VMware vSphere® are many performance and scalability improvements. The vSphere 7.x platform continues to provide industry-leading performance and features to ensure the successful virtualization and management of your entire software-defined datacenter.

This document will be regularly updated with highlights of new features and performance information with each subsequent release of the vSphere 7.x platform.

vSphere 7.0 (released Apr 2020)

Virtual Machine Scalability

Virtual Hardware Version	Virtual CPUs	Virtual Memory
17	256	6TB

Selective Latency Sensitivity

When moving physical network infrastructure to network function virtualization (NFV), customers expect that the data plane applications within the guest operating system will remain high performing with minimal jitter. To meet this expectation, the VMware performance team advises customers to deploy NFV VMs with `latency-sensitivity = HIGH` and full CPU reservation. By doing this, all vCPUs of the VM can achieve low-latency responsiveness by consuming dedicated physical cores from the host, regardless of the applications running inside the guest.

However, a key issue here is that such a configuration is wasteful for vCPUs running non-critical tasks. The dedicated physical cores assigned to non-critical vCPUs can be better utilized to fit other data path applications from different guests to drive up the consolidation ratio and improve virtualization return on investment for a variety of workloads, including Telco NFV. Because of this, ESXi needs to be aware of different application requirements within a VM when provisioning vCPUs with the latency setting.

This new feature allows the VM to pin a subset of its vCPUs to individual cores for greater performance. Note that this is an enhancement to the existing VM latency-sensitivity setting, which pins all vCPUs in the VM to cores. Pinning a specified vCPU subset reduces the resource requirements while maintaining needed performance.

We can now demonstrate workloads running on pinned vCPUs that show similar or better performance whether the latency-sensitive setting is used for individual CPUs or when it’s used for the whole VM.

Deprecation of vmklinux

All previously supported devices that are not supported by native drivers will not function and will not be recognized when installing or upgrading vSphere 7.0.

This will ensure the continued efficiency, performance, and availability of new features for the vSphere platform.

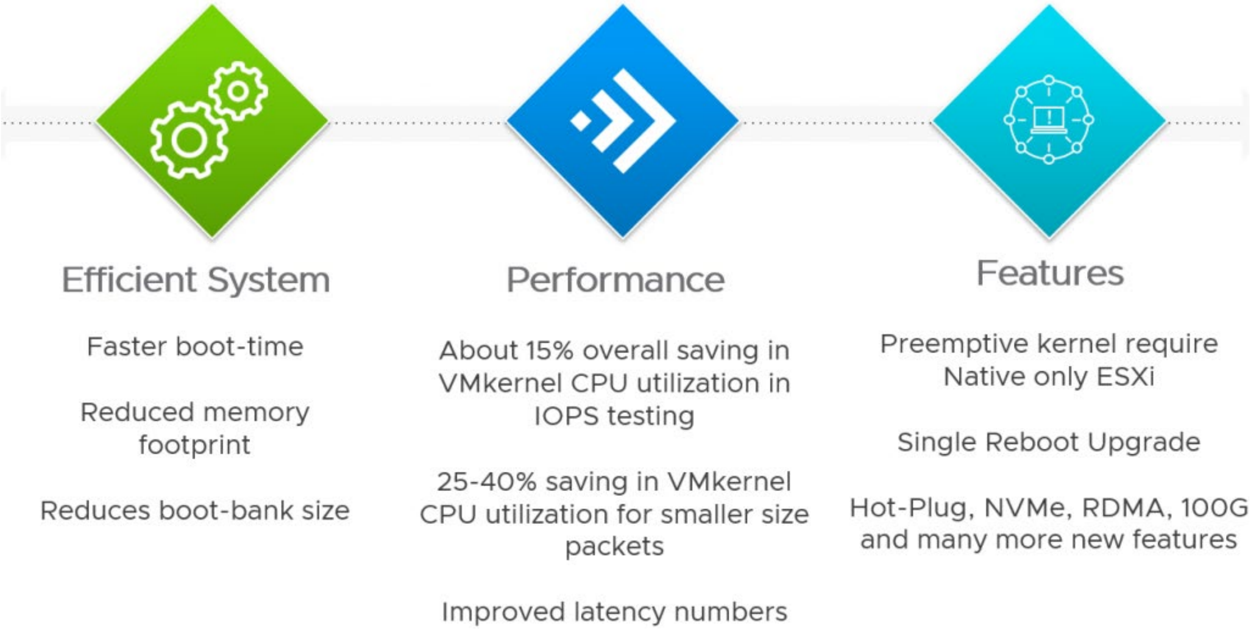

Efficient System
Faster boot-time
Reduced memory footprint
Reduces boot-bank size
Performance
About 15% overall saving in VMkernel CPU utilization in IOPS testing
25-40% saving in VMkernel CPU utilization for smaller size packets
Improved latency numbers
Features
Preemptive kernel require Native only ESXi
Single Reboot Upgrade
Hot-Plug, NVMe, RDMA, 100G and many more new features

Figure 1. Improved performance for latency-sensitive workloads

References:

- Devices deprecated and unsupported in ESXi 7.0
  https://kb.vmware.com/s/article/77304
- What is the Impact of the VMKlinux Driver Stack Deprecation?
  https://blogs.vmware.com/vsphere/2019/04/what-is-the-impact-of-the-vmklinux-driver-stack-deprecation.html

Assignable Hardware

In vSphere versions prior to vSphere 7.0, a VM specified a PCIe passthrough device by using its hardware address. This was an identifier that pointed to a specific physical device at a specific bus location on the VM's ESXi host. This restricted that VM to that particular host. The VM could not easily be migrated to another ESXi host with an identical PCIe device. This impacted the availability of the application using the PCIe device in the event of a host outage.

Assignable hardware in vSphere 7.0 provides a flexible mechanism to assign hardware accelerators to workloads. This mechanism identifies the hardware accelerator by attributes of the device rather than by its hardware address, allowing for a level of abstraction of the PCIe device. The assignable hardware feature implements compatibility checks to verify that ESXi hosts have assignable devices available to meet the needs of the VM.

Virtual NVMe Controller

Virtual NVMe devices have reduced guest I/O processing overhead. The virtual NVMe controller is the default disk controller for the following guest operating systems when using hardware version 15 or later:

- Windows 10
- Windows Server 2016
- Windows Server 2019

vMotion Enhancements

vSphere 7.0 introduces several new vMotion enhancements, including memory pre-copy optimizations, loose page trace installs, improved page table granularity, and switch-over phase enhancements. These features are designed to support the vMotion of “monster” VMs (those with a large number of vCPUs) but benefit all sizes of vMotions as well.

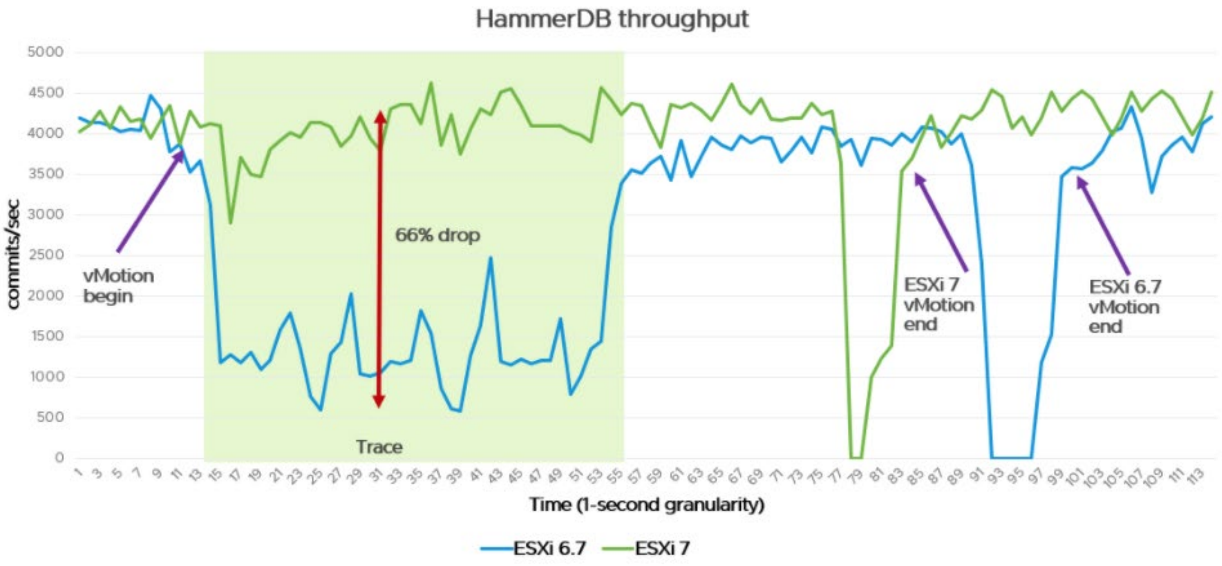


Figure 2. vMotion enhancements in vSphere 7 over vSphere 6.7

As a result, vMotion no longer exhibits a performance impact during the page trace phase, the stun time remains within 1 second instead of taking multiple seconds, and the overall live-migration time is almost 20 seconds shorter.

Reference:

- vMotion Enhancements in vSphere 7
  https://core.vmware.com/resource/vmotion-enhancements-vsphere-7#section6

Precision Clock Device

vSphere 7 introduces a new timekeeping option with the ESXi host precision timing protocol (PTP), and a new precision clock device. This is a virtual clock device that provides a VM with access to the system time of the primary ESXi host. This is an important performance feature that maintains accurate timekeeping across the cluster for time-sensitive workloads.

Scalable Shares

The *scalable shares* feature is a new dynamic resource management mechanism within DRS to automatically manage VM shares within and across resource pools to ensure priority is dynamically rebalanced as the population of resource pools change. Enabling scalable shares allows resource pools to have dynamic and relative entitlements. If you've used resource pools, you've likely seen a situation where resource pools configured with higher share levels would not necessarily guarantee more resources to their workloads. This is no longer the case with scalable shares. An improvement like this is also important for vSphere with Kubernetes because the vSphere pod service needs this feature to ensure performance.

References:

- DRS with Scalable Shares in vSphere 7
  https://www.youtube.com/watch?v=jkp25I4R0R8
- vSphere 7 DRS Scalable Shares Deep Dive
  https://frankdenneman.nl/2020/05/27/vsphere-7-drs-scalable-shares-deep-dive/

vSphere 7.0 U1 (released Oct 2020)

Virtual Machine Scalability

Virtual Hardware Version	Virtual CPUs	Virtual Memory
18	768	24TB

Paravirtual RDMA and Native Endpoints

RDMA between VMs is known as paravirtual RDMA (PVRDMA); this was introduced in vSphere 6.5. VMs with a PCIe virtual NIC that supports a standard RDMA API can leverage PVRDMA technology. VMs must be connected to the same distributed virtual switch to leverage PVRDMA.

In vSphere 7.0 U1, PVRDMA-capable VMs can now communicate with native endpoints. Native endpoints are RDMA-capable devices such as storage arrays that do not use the PVRDMA adaptor type (non-PVRDMA endpoints). Use of RDMA technology for communication between nodes in a cluster and storage arrays is quite common because it offers very low latency and high performance, especially for third-tier applications.

With this feature, customers can enhance the performance for applications and clusters that use RDMA to communicate with storage devices and arrays.

Monster VM Enhancements

Performance enhancements in ESXi that support the larger scale of VMs include the widening of the physical address, address space optimizations, better NUMA awareness for guest operating systems, and more scalable synchronization techniques. This allows VMs to now be sized to 768 vCPUs and 24TB RAM. ESXi hosts with AMD processors can support VMs with twice the previous number of vCPUs (256), and up to 8TB of RAM.

vMotion Enhancements

VMware vSphere® vMotion® incorporates several cutting-edge enhancements in vSphere 7.0 U1 that provide dramatic improvements in performance when migrating VMs. These include rearchitecting memory pre-copy with an innovative page-tracing mechanism, among other things, that greatly reduce the performance impact on guest workloads during live migration and significantly reduce vMotion duration. Performance data from Tier-1 workloads show these optimizations can scale to hundreds of vCPUs and terabytes of memory.

Figure 3 compares the impact of vMotion on an Oracle Database server running inside a 72-vCPU/1TB VM on both vSphere 6.7 and vSphere 7.0 U1. Thanks to all the new performance optimizations, we observed great improvement in guest performance during the 7.0 U1 vMotion. First, by optimizing the memory allocation path on the destination, the duration of migration was cut by half. By completely reinventing hypervisor page-tracing mechanisms, the average throughput loss of 50% during 6.7 vMotion was brought to less than 5% during 7.0 U1 vMotion. Finally, all the enhancements added to vMotion improved Oracle Database resume time during the switchover phase by an order of magnitude.

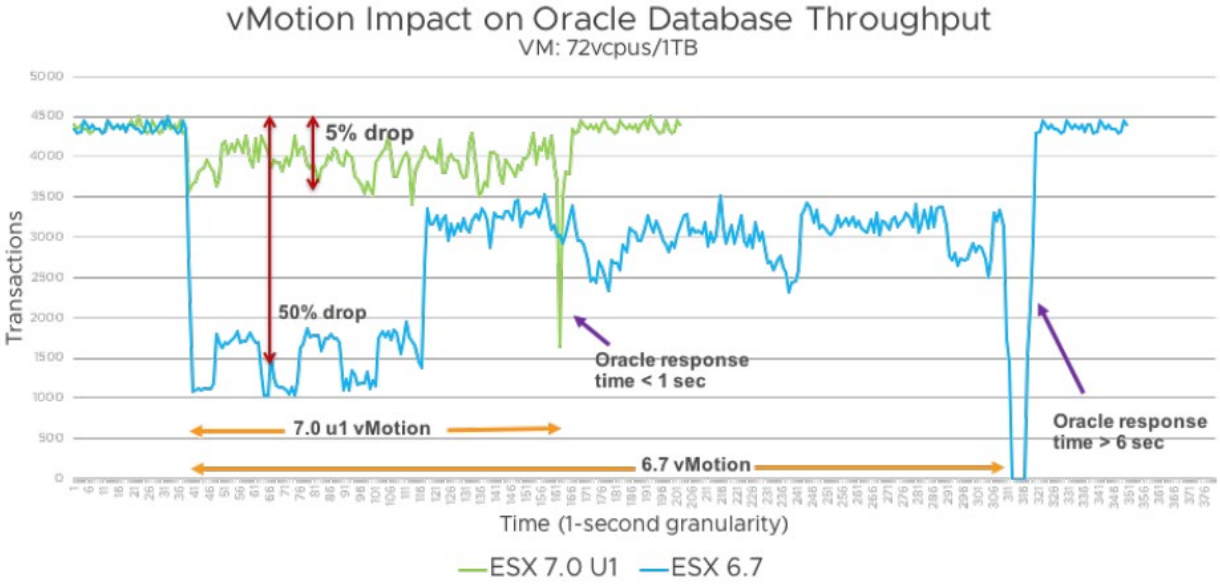


Figure 3. vMotion shows improved performance with an Oracle Database workload

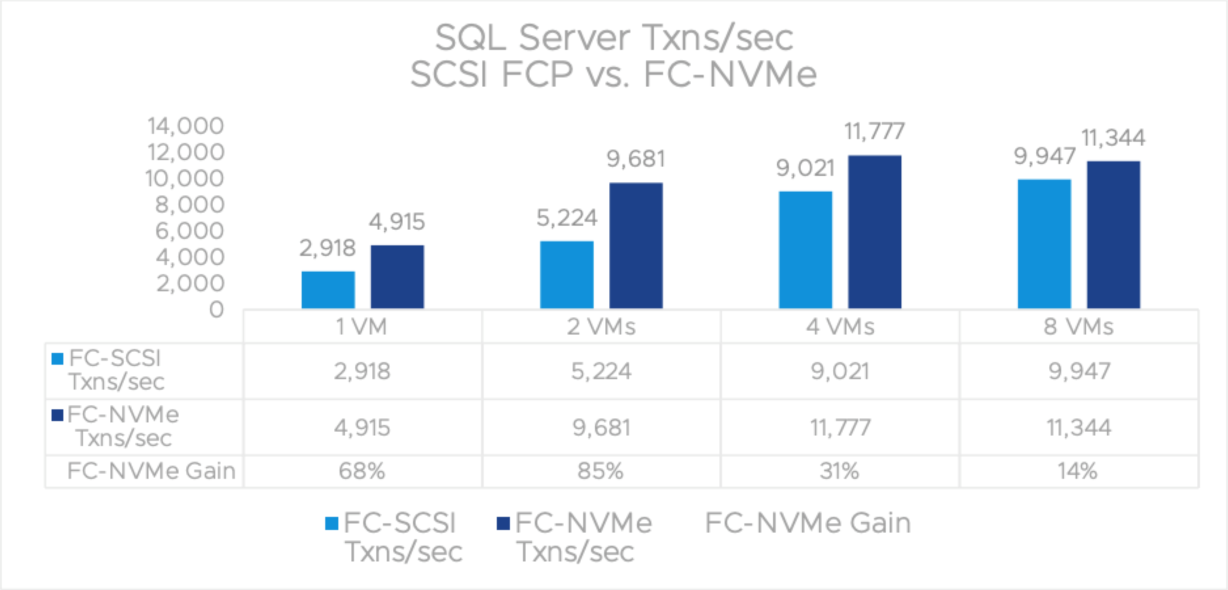
Reference:

- vMotion Innovations in VMware vSphere7.0 U1
  https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/performance/vmotion-7u1-perf.pdf

NVMe-oF

vSphere 7.0 U1 introduces support for NVMe over Fabrics (NVMe-oF), a protocol specification that connects hosts to high-speed flash storage via network fabrics using the NVMe protocol. The fabrics that vSphere 7.0 U1 supports include Fibre Channel (FC-NVMe) and RDMA (RoCE v2). The benchmark results show that FC-NVMe consistently outperforms SCSI FCP in vSphere virtualized environments, providing higher throughput and lower latency.

These tests involved driving a heavy database workload using the Microsoft Cloud Database benchmark using 1, 2, 4, and 8 SQL Server VMs on the ESXi host, both with legacy SCSI FCP and the newer FC-NVMe protocol. The first chart that follows represents the primary CDB metric for measuring database performance: transactions per second (Txns/sec). FC-NVMe is higher in every case and is 85% higher performing with two VMs. Note also that SCSI FCP never quite reached 10,000 Txns/sec, while FC-NVMe was able to surpass 11,000 Txns/sec with both 4 and 8 VMs.



	1 VM	2 VMs	4 VMs	8 VMs
■ FC-SCSI Txns/sec	2,918	5,224	9,021	9,947
■ FC-NVMe Txns/sec	4,915	9,681	11,777	11,344
FC-NVMe Gain	68%	85%	31%	14%

Figure 4. SQL Server transactions per second for FC-SCSI vs FC-NVMe

These results provide a compelling case for customers to upgrade their existing environments to vSphere 7.0 U1 and gain the benefits of NVMe-oF.

Reference:

- Performance Characterization of NVMe-oF in vSphere 7.0 U1 https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/performance/vSphere7u1-NVMe-oF-FC-vs-FC-SCSI.pdf

vSphere 7.0 U2 (released Mar 2021)

Virtual Machine Scalability

Virtual Hardware Version	Virtual CPUs	Virtual Memory
19	768	24TB

Precision Time for Windows

Traditionally, using NTP and Active Directory for time sources is a common practice. While they are mostly accurate, they can sometimes result in too much jitter because of a high stratum value and time synchronization issues. In a highly time-sensitive environment like financial trading, this situation can be quite dangerous. Considering this, VMware has introduced a new precision time architecture that allows VMs to acquire host time over a low overhead and extremely low jitter VM-hypervisor interface.

For vSphere 7.0 U2, VMware introduces a new plugin called vmwTimeProvider as part of VMware Tools. Usually, a driver would be required to access a device. However, the precision clock supports a VMware proprietary paravirtual interface accessible from userspace. Using this paravirtual interface precludes any performance benefits of using memory mapped I/O (MMIO)–based time registers optimized for fast reads.

Reference:

- Achieve Higher Degree of Time Synchronization Accuracy: Precision Time for Windows https://core.vmware.com/blog/achieve-higher-degree-time-synchronization-accuracy-precision-time-windows

Enterprise NVIDIA Infrastructure Support

vSphere 7.0 U2 adds support for the NVIDIA Ampere architecture that enables you to perform high-end AI/ML training and ML inference workloads by using the accelerated capacity of the A100 GPU. vSphere support for A100 GPUs delivers world-class AI performance: up to 20X the performance of previous generation GPUs. As well, you get near bare metal performance and technologies such as GPU Direct communications that enable higher performance for scale-out workloads (adding more VMs to the vSphere system).

In addition, vSphere 7.0 U2 improves GPU sharing beyond time-slice sharing by supporting the Multi-Instance GPU (MIG) technology. vSphere DRS automatically places workloads across AI infrastructure at scale for optimal consumption, and vSphere vMotion provides live migration to simplify infrastructure maintenance such as consolidation, expansion, or upgrades.

With vSphere 7.0 U2, you also see enhanced performance of device-to-device communication, building on the existing NVIDIA GPUDirect functionality, by enabling Address Translation Services (ATS) and Access Control Services (ACS) at the PCIe bus layer in the ESXi kernel.

Reference:

- The AI-Ready Enterprise Platform: Unleashing AI for Every Enterprise
  https://blogs.vmware.com/vsphere/2021/03/ai-ready-enterprise-platform-unleashing-ai-every-enterprise.html

Support for Mellanox ConnectX-6 200G NICs

vSphere 7.0 U2 supports the Mellanox Technologies MT28908 Family (ConnectX-6) and the Mellanox Technologies MT2892 Family (ConnectX-6 Dx) 200G NICs offering leading edge network performance.

Reference:

- ConnectX-6 EN Card 200GbE Ethernet Adapter Card
  https://www.mellanox.com/files/doc-2020/pb-connectx-6-en-card.pdf

Performance Improvements for AMD Zen CPUs

vSphere 7.0 U2 includes a CPU scheduler that is architecturally optimized for AMD EPYC. This scheduler is designed to take advantage of the multiple last-level caches (LLCs) per CPU socket offered by the AMD EPYC processors. An extensive performance evaluation using both enterprise benchmarks and microbenchmarks shows that the CPU scheduler in vSphere 7.0 U2 achieves up to 50% better performance on these processors than vSphere 7.0 U1. AMD Zen CPU optimizations allow a higher number of VMs or container deployments with better performance.

We installed a VM with Windows Server 2019 and SQL Server 2019 with 8 vCPUs and 64GB of RAM. We created a DVD Store 3 benchmark database of approximately 50GB for testing. The VM was then cloned 16 times to create the testbed. Results from testing showed performance gains for 7.0 U2 with 8 VMs and 16 VMs of 21% and 13.4%, respectively.

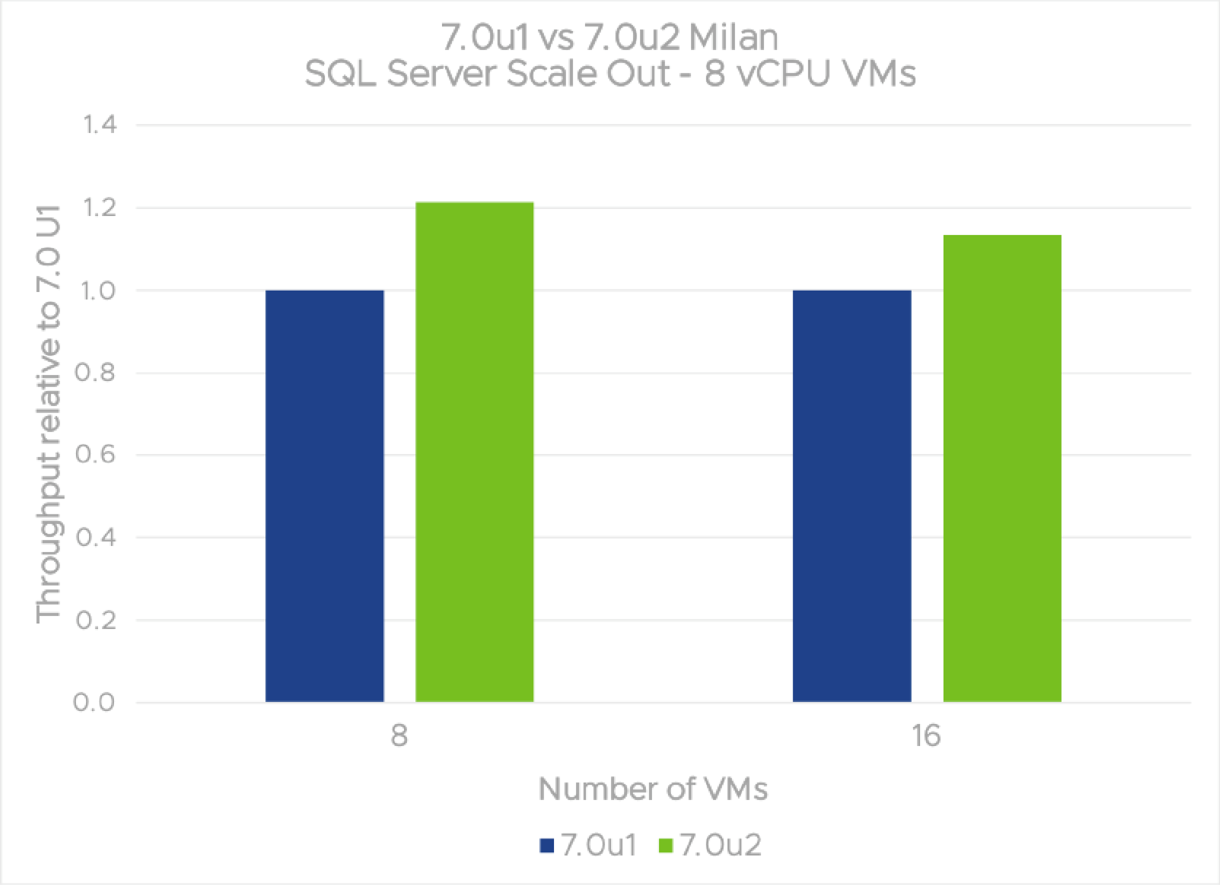


Figure 5. vSphere 7.0 U2 VMs show a performance increase compared to vSphere 7.0 U1 when running on a host configured with AMD EPYC processors

Reference:

- Performance Optimizations in VMware vSphere 7.0 U2 CPU Scheduler for AMD EPYC Processors https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/performance/vsphere70u2-cpu-sched-amd-epyc.pdf

Latency-sensitive Workload Optimizations

Latency-sensitive workloads, such as those in financial and telecom applications, can see a significant performance benefit from I/O latency and jitter optimizations in ESXi 7.0 U2. The optimizations reduce interference and jitter sources to provide a consistent runtime environment. With ESXi 7.0 U2, you can also see a higher speed in interrupt delivery for passthrough devices. This was accomplished through reducing, suspending, or removing system maintenance operations, timer optimizations, interrupt placement, and the use of posted interrupts in the virtual machine monitor. These typically reduced the Cyclictest max latency tails from over 100 microseconds to less than 15 microseconds.

ESXi Suspend-to-Memory

ESXi suspend-to-memory enables faster host upgrades and makes maintenance operations even less disruptive than before. Suspend-to-memory is available in combination with ESXi quick boot. Instead of moving VMs prior to host updates, VMs can be suspended to host memory while only the hypervisor kernel restarts.

vMotion Auto Scaling

vSphere vMotion, in versions prior to vSphere 7.0 U2, does not saturate high bandwidth NICs (such as 25, 40, and 100 GbE) without additional configuration. It required tuning at multiple levels such as vMotion streams, the TCP/IP VMkernel interfaces, and potentially even the NIC driver. That's because, by default, vMotion would use a single stream for handling the vMotion process.

With vSphere 7.0 U2, the vMotion process automatically spins up the number of streams according to the bandwidth of the physical NICs used for the vMotion networks. This is now an out-of-the-box setting. The usable bandwidth for vMotion is determined by querying the underlying NICs.

This means that all VMkernel interfaces enabled for vMotion are checked, along with the underlying physical NIC bandwidths. Depending on the outcome, a number of streams are instantiated. The baseline is 1 vMotion stream per 15 GbE of bandwidth. This results in the following number of streams per VMkernel interface:

- 25 GbE = 2 vMotion streams
- 40 GbE = 3 vMotion streams
- 100 GbE = 7 vMotion streams

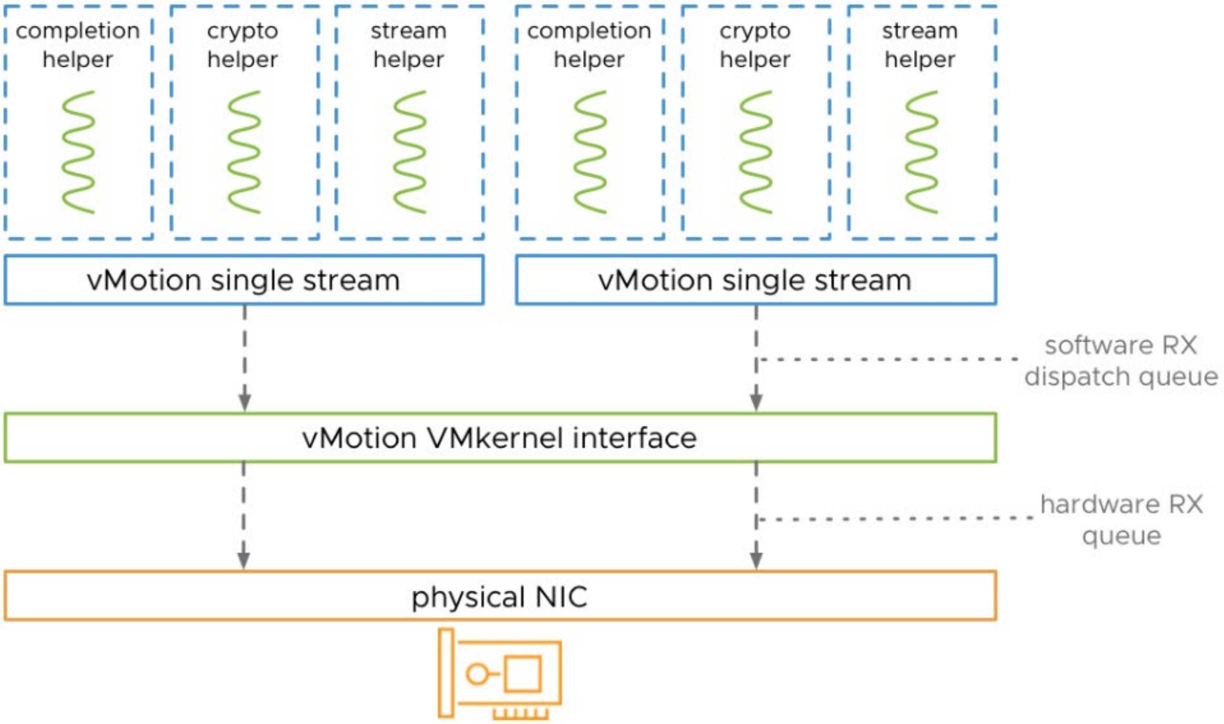


Figure 6. vMotion auto-scaling architecture

Reference:

- Faster vMotion Makes Balancing Workloads Invisible
  https://core.vmware.com/blog/faster-vmotion-makes-balancing-workloads-invisible

vSphere 7.0 U3 (released Sept 2021)

Virtual Machine Scalability

Virtual Hardware Version	Virtual CPUs	Virtual Memory
20	768	24TB

Latency-sensitive Workload Optimizations

ESXi has been further optimized to allow ultra-low-latency applications to perform better with reduced jitter and interference, specifically edge-based, real-time applications. As a part of this feature, a lot of periodic timers have been removed from ESXi when running in low-latency mode and device interrupts have been moved away from CPUs reserved for low-latency applications. Enabling low-latency mode and BIOS optimizations are required to take full advantage of this feature.

vSphere Memory Monitoring and Remediation (vMMR)

The size of DRAM contributes roughly to 50-60% of the server cost. And it's not linear—a 1TB DRAM contributes roughly to 75% of the server cost. So, there's a huge need to reduce the DRAM cost. One solution is Intel Optane Persistent Memory Mode, in which the hardware hides the DRAM as cache and exposes PMem as the memory of the system. PMem is much cheaper, but it has higher latency.

vMMR is a set of telemetry and alerting technology that allows you to use Intel PMem Memory Mode and be alerted when the ESXi host for virtual machine active memory exceeds the host available DRAM, which may start to cause performance degradation.

NVMe over TCP/IP

vSphere 7.0 U3 adds support for NVMe over TCP, which allows ubiquitous TCP/IP networking infrastructure to be used for storage traffic that is better optimized for flash and SSD. With this advancement, organizations can achieve higher performance and lower latency at a reduced cost.

References:

- Announcing vSphere 7 Update 3
https://blogs.vmware.com/vsphere/2021/09/announcing-vsphere-7-update-3.html

- Announcing NVMe/TCP Support with VMware vSphere 7 Update 3
https://blogs.vmware.com/vsphere/2021/09/announcing-nvme-tcp-support-vmware-vsphere-7-update-3.html

Conclusion

Based on these performance, scalability, and feature improvements in vSphere 7.x, VMware continues to demonstrate industry-leading performance.

About the Authors

Mark Achtemichuk currently works as a staff 2 engineer within VMware's Performance Engineering team; he focuses on education, benchmarking, collateral, and performance architecture. @vmMarkA is recognized as an industry expert and holds a VMware Certified Design Expert (VCDX#50) certification. He has worked on engagements with Fortune 50 companies, served as technical editor for many books and publications, and is a sought-after speaker at numerous industry events.

Julie Brodeur is a senior technical writer in the Performance Engineering team at VMware, where she supports the writing activities of a broad organization. She focuses on technical papers, knowledgebase articles, and blogs about the performance of the vSphere suite of products and related virtualization topics.

Acknowledgments

The authors thank the product managers and performance engineers who contributed to and reviewed this paper.