Microsoft SQL Server 2019 on VMware vSphere 7

Performance Study - September 1, 2021

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

Table of Contents

Executive Summary

Microsoft SQL Server® databases can achieve great performance and fully utilize current-generation server and storage technology using VMware vSphere®. With the introduction of vSphere 7.0 Update 2, recently released servers with 3rd Generation Intel® Xeon® "Ice Lake" processors are supported with up to 40 cores per socket. In this paper, a variety of vCPU and virtual machine combinations were tested on current-generation Intel hardware to show that vSphere 7 can handle tens of thousands of online transaction processing (OLTP) database operations per minute. This represents an increase in performance over previous generation servers.

Introduction

VMware vSphere provides an ideal platform on which customers can virtualize their business-critical applications, including databases, ERP systems, and emerging technologies. A full discussion of the benefits is included in the whitepaper "Virtualizing Business-Critical Applications on vSphere" [1].

A business-critical application that is often run on vSphere is Microsoft SQL Server. Consolidating these deployments onto modern multi-socket, multi-core, multi-threaded server hardware is an effective solution for administrators and their organizations. Achieving optimal SQL Server performance on vSphere has been an ongoing focus for VMware. Previous series of performance tests with SQL Server on vSphere 5.5 [2] using a four-socket Intel Xeon E7-4800 and vSphere 6 using a four socket E7-4800 v2–based server [3] were published. Those whitepapers showed excellent performance up to the maximum size virtual machine supported at the time: 64 vCPUs for vSphere 5.5 and 128 vCPUs for vSphere 6.

This new study uses a 2-socket–based server, but still approaches the testing with similar methodology as the previous tests. Using 3rd Generation Intel Xeon Scalable "Ice Lake" processors, a 2-socket server can host one large SQL Server virtual machine **or** a significant number of smaller SQL Server VMs. This is because of processor improvements [4] including increased number of cores per socket, more PCI Express lanes, higher memory bandwidth, and higher memory capacity. The maximum size of VM supported with vSphere 7.0 Update 2 is much larger now at 768 vCPUs; however, this performance study is limited to 76 vCPUs to match the total core count in the 2-socket host used.

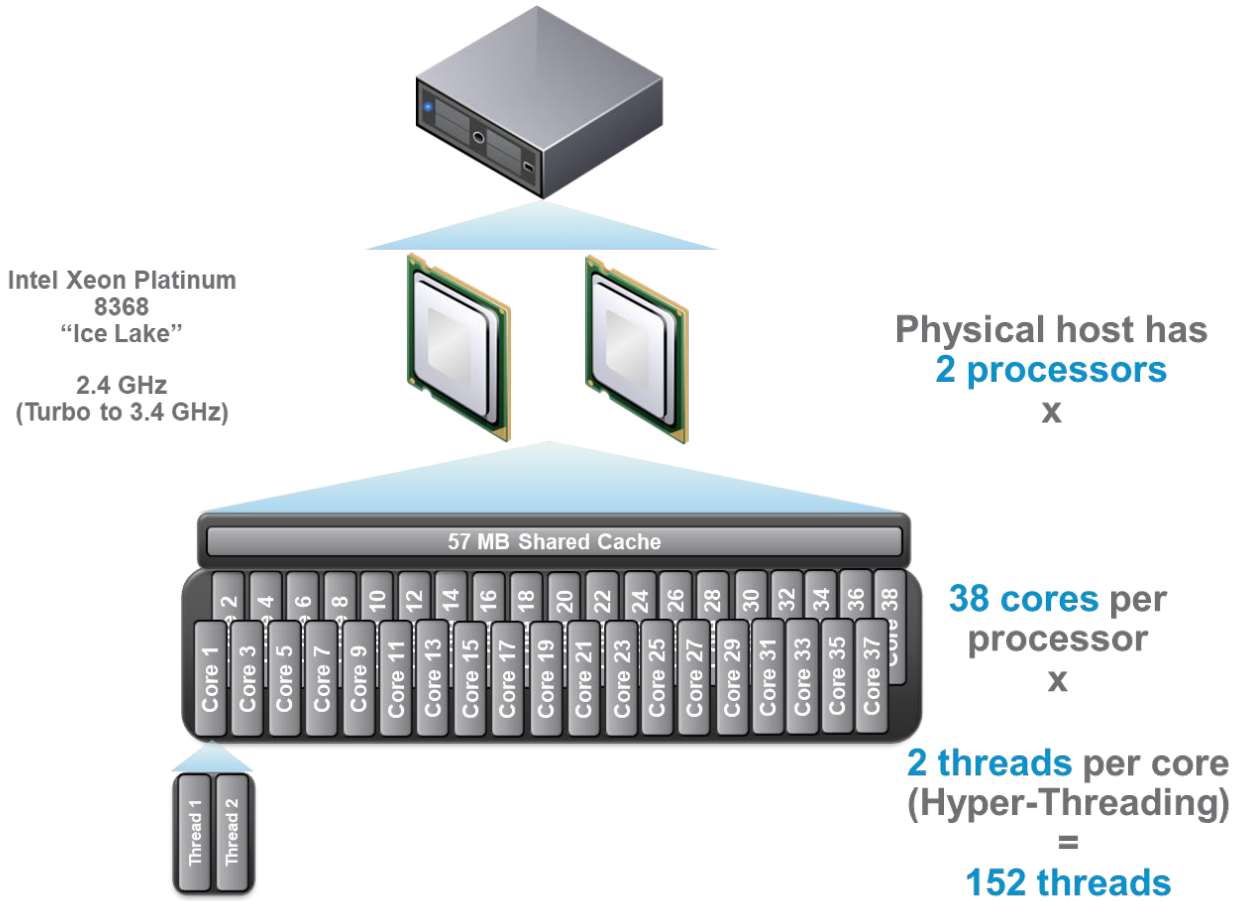
The new tests show large SQL Server database instances continue to run extremely efficiently, achieving great performance in a variety of virtual machine configurations on vSphere 7.0 Update 2. Only minor tunings to the SQL Server application and the vSphere 7 ESXi hypervisor were needed to fully optimize these large virtual machines.

Testbed and Methodology

To test the performance of SQL Server database virtual machines, a benchmark testbed was assembled. This infrastructure was tested with many small virtual machines (scale-out) as well as a single large VM (scale-up). Multiple databases were created within these virtual machines, and the open-source DVD Store workload was used to measure the aggregate throughput.

Testbed Configuration

The test server was equipped with two 3rd Generation Intel Xeon Scalable "Ice Lake" Platinum 8368 processors with 38 cores per socket (76 cores total), 2 TB of physical memory, and two 10GbE NICs. Since the goal of this study was to achieve maximum application performance, the BIOS system profile was set to **Performance**. Other BIOS settings, such as Intel Hyper-Threading and Turbo Mode, were left at their defaults (enabled) because they provided the highest throughput. ESXi 7.0 Update 2 was booted from an image on the network, while all the virtual machines' disks resided on the local NVMe storage devices.



SQL Server and OS Performance Tuning

Microsoft Windows® Server 2019 was used as the guest operating system for the virtual machines. The number of virtual CPUs and virtual RAM varied depending upon the test, which is detailed later. The operating system, SQL data files, and SQL logs were stored on separate virtual disks. The latest version of VMware Tools was installed, and each virtual hard disk was connected to a separate VMware Paravirtual SCSI (PVSCSI) adapter to ensure optimal throughput and lower CPU utilization. SQL Server 2019 Enterprise Edition was installed for each database instance with the following features: Database Engine Services, Full-Text Search, and Client Tools Connectivity. The Database Engine was configured with Mixed Mode Authentication.

For tuning SQL Server, here are two great resources:

1. Architecting Microsoft SQL Server on VMware vSphere Best Practices Guide [5] has a wealth of recommendations and guidance for running SQL Server in a vSphere-based environment.

2. Performance Tuning Guidelines for Windows Server 2019 [6] is a Microsoft document with a specific set of tuning recommendations for OLTP workloads. This document includes tuning guidelines that are applicable to virtualized configurations. The OLTP tuning section [7] is located under **Additional Tuning Resources → Performance Tuning for Online Transaction Processing** (OLTP).

The tunings were found to have a noticeable benefit for the DVD Store workload and are summarized below:

- Set the power plan to **High Performance**.
- Enable SQL Server to use large pages by enabling the Lock pages in memory user right for the account that will run the SQL Server in Group Policy.
- Enable the TCP/IP protocol to allow communication with client systems: In Server Configuration Manager, navigate to **SQL Server Network Configuration** > **Protocols for MSSQL Server**, right-click **TCP/IP**, and click **Enable**.
- SQL Server Startup Parameter -T834 Use Microsoft Windows large-page allocations for the buffer pool.
- Max degree of parallelism (MAXDOP) [7] is an advanced configuration option that controls the number of processors used to execute a query in a parallel plan. Setting this value to 1 disables parallel plans altogether, which is not recommended in a large virtual machine because it could result in unused virtual CPUs and wasted resources. Microsoft has guidelines for setting this value, depending on whether the host has NUMA and (or) Hyper-Threading, and there are numerous recommendations online, but many predate virtualization and modern server hardware. The default value of 0 is recommended because it allows SQL Server to utilize all virtual CPUs available, and this was the value used for all testing.
- Cost threshold for parallelism [8] is an option that specifies the threshold at which parallel plans are used for queries. The value is specified in seconds, and the default is 5, which means a

parallel plan for queries is used if SQL Server estimates it would take longer than 5 seconds if run serially. This value started to show very high response times with large virtual machines with 8 or more virtual CPUs, particularly for the DVD Store **login** query. Increasing this value to 50 resolved this performance bottleneck and this value was used for all testing.

- SQL Server licensing can affect performance by potentially limiting scaling. Product features vary depending upon the edition (the two main editions are Standard and Enterprise), and SQL Server 2019 licensing has two different licensing models (Core-Based or Server+CAL). During previous testing, the type of license was determined to be limiting the performance of virtual machines with more than 20 vCPUs. The SQL instance's ERRORLOG showed this message when the VM had more than 20 vCPUs: SQL Server detected 32 sockets with 1 core per socket and 1 logical processor per socket for 32 total logical processors, and SQL Server licensing detected 20 logical processors. This was due to a 20 logical processor limitation imposed on the Server+CAL licenses we had used initially. Fortunately, using a Core-Based licensing model can allow customers to license by VMs/vCPUs instead of the entire physical host. See "Licensing SQL Server 2019 in a Virtualized Environment" [9] in the *Microsoft SQL Server 2019 Licensing Guide* for all definitive statements about SQL Server 2019 licensing in virtualized environments.
- To scale up load on a SQL Server virtual machine, large tables can be partitioned, multiple databases can be hosted on one SQL instance, or multiple "named instances" of SQL Server can be installed.

DVD Store

Performance tests were conducted using the open source DVD Store 3.5 benchmark [10]. DVD Store simulates an online store that allows customers to log in, browse products, read reviews, leave new reviews, join as a premium member, rate reviews, and purchase products. Simple and complex database queries are included in the workload in addition to its use of many database features such as indexes, foreign keys, and stored procedures.

A new feature of DVD Store 3.5 allows for multiple "stores" to be created within a single instance. Essentially, it models the case where there are a number of stores that each have their own set of tables representing their own customers, products, reviews, and purchases. This allows for the database to be scaled to larger sizes with data spread across more tables and indexes instead of a single set of tables as in previous versions of DVD Store.

For these tests, the general guideline for creating and running DVD Store instances was to have a SQL Server instance for each 8 vCPUs of the VM. An 8-vCPU VM had 1 SQL Server instance, while a 32-vCPU VM had 4 SQL Server instances. Each SQL Server instance was loaded with DVD Store 3.5 with 5 stores totaling about 50 GB of space on disk.

Scale-Up Results

In this set of tests, a single VM was increased in size in terms of both vCPUs and memory to measure how well its performance scales. Testing started with an 8-vCPU VM and increased up to a maximum of 76 vCPUs. The results are shown below.

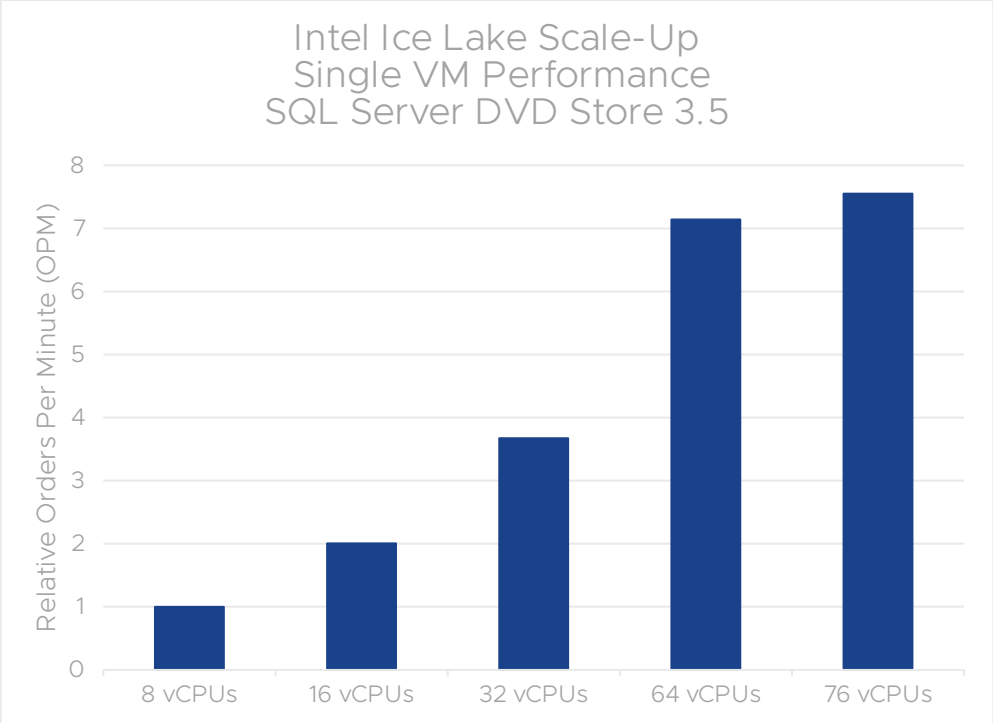


Figure 1. When scaling up a VM with varying amounts of vCPUs, performance improves steadily until reaching a VM sized with 76 vCPUS - the same as the number of physical cores in the server.

The host has 38 cores per socket for a total of 76. Because 38 is not easy to divide evenly, a commonly used size of 8 vCPUs was selected as the base size. That was then doubled to 16, 32, and 64, with a final run at 76 to measure performance of a run that is "right-sized" at one physical core per vCPU.

Performance scales almost linearly from 8 to 16, falls off slightly from 16 to 32, and is then again almost linear from 32 to 64. The final test using the last few vCPUs shows an additional gain, taking advantage of the final cores by moving from 64 to 76.

Scale-Out Results

For this set of tests, multiple 8-vCPU VMs were used to scale out the tests with many VMs, starting with a single 8-vCPU VM and moving up to a total of 10 VMs (80 vCPUs assigned across the 76 physical cores in the system). For these tests, we compared against a previous generation Intel server using "Cascade Lake" processors, which have 24 cores per socket for a total of 48 cores for

the two-socket system (compared with 38 cores per socket and 76 total cores for the Ice Lake–based server).

The current generation Intel server provided better performance that was mostly due to having more cores, but even in the cases with 1 or 2 VMs (where the core count doesn't really come into play), the Ice Lake system still had a small advantage.

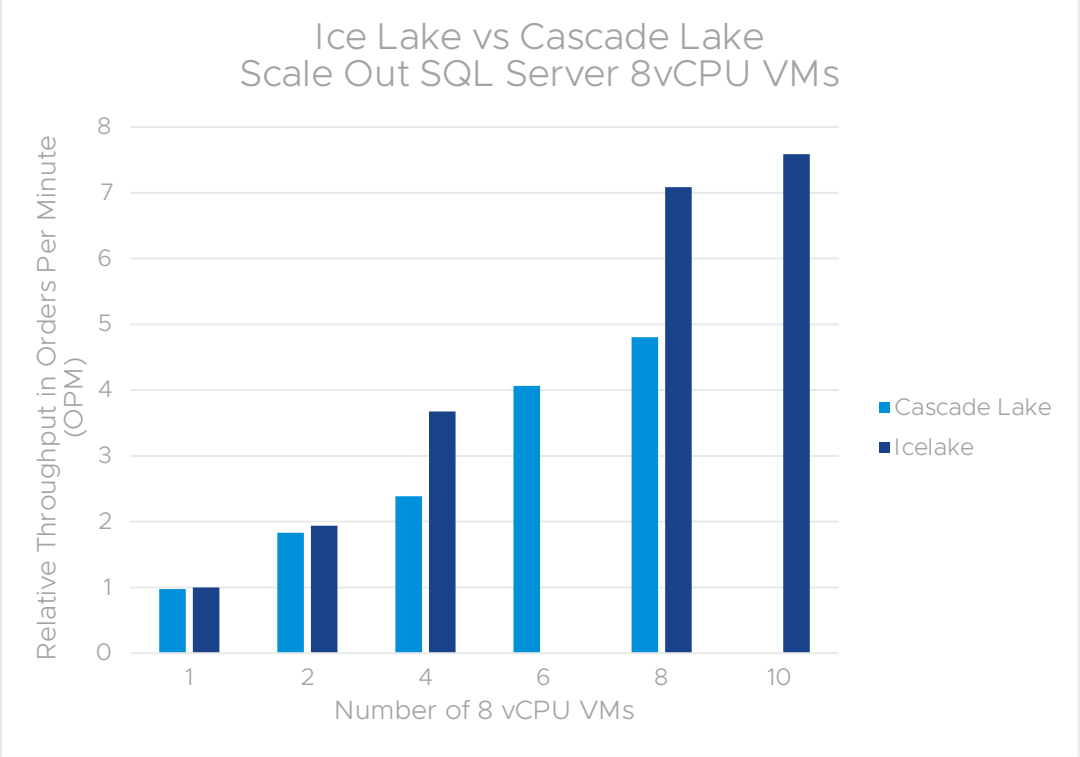


Figure 2. When scaling out the number of VMs, performance improves steadily for both the older and newer processor generations, with Ice Lake seeing a larger improvement and the ability to host more VMs.

The graph above appears to have two "missing" data points. There isn't any data for an Ice Lake test with 6 VMs or for Cascade Lake at 10 VMs. These tests start at 1 VM and then double to 2, 4, 8, and so on, until the maximum number of VMs needed for the test is reached. When the number of cores in a processor is not a multiple of two, then the max number of VMs will break the pattern of doubling. The Cascade Lake system has 48 cores total, which means that 6 VMs with 8 vCPUs is a good data point for the system with 1 vCPU per physical core. The Ice Lake system has 76 cores, so 10 VMs with 8 vCPUs each was close, with only four extra vCPUs assigned. The datapoint with them both running 8 VMs with 8 vCPUs shows a good comparison of the two hosts, and the 6 VMs for Cascade Lake and 10 VMs for Ice Lake provide additional context for that result.

Conclusion

SQL Server has always performed very well on VMware vSphere, as shown in previous performance studies. This new testing with vSphere 7.0 Update 2, SQL Server 2019, and the latest 3rd Generation Intel Xeon Scalable Processors continues to showcase this performance scalability. As new hardware becomes available with faster processors with more cores and more memory, VMware vSphere will continue to support these increased capabilities to provide a great platform for using that additional compute power for SQL Server databases.

References

[1] VMware. (2012) Virtualizing Business-Critical Applications on vSphere. https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/whitepaper/solutions/vmware-virtualizing-business-critical-apps-on-vmware_en-white-paper.pdf

[2] David Morse. (2014) Performance and Scalability of Microsoft SQL Server on VMware vSphere 5.5. https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/vmware-sql-server-vsphere55-performance-white-paper.pdf

[3] David Morse. (2015) Performance Characterization of Microsoft SQL Server on VMware vSphere 6. https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/whitepaper/solutions/vmware-sql-server-vsphere6-performance-white-paper.pdf

[4] Intel. (2021, March) 3rd Gen Intel Xeon Scalable Processors. https://www.intel.com/content/www/us/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html

[5] VMware. (2019, April) Architechting Microsoft SQL Server on VMware vSphere Best Practices Guide. https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/solutions/sql-server-on-vmware-best-practices-guide.pdf

[6] Microsoft. (2017, October) Performance Tuning Guidelines for Windows Server 2019. https://docs.microsoft.com/en-us/windows-server/administration/performance-tuning/

[7] Microsoft. (2016, March) Performance Tuning for Online Transaction Processing (OLTP). https://docs.microsoft.com/en-us/previous-versions/dn567642(v=vs.85)

[8] Microsoft. (2021, March) Configure the max degree of parallelism Server Configuration Option. https://docs.microsoft.com/en-us/sql/database-engine/configure-windows/configure-the-max-degree-of-parallelism-server-configuration-option?view=sql-server-ver15

[9] Microsoft. (2017, March) Configure the cost threshold for parallelism Server Configuration Option. https://docs.microsoft.com/en-us/sql/database-engine/configure-windows/configure-the-cost-threshold-for-parallelism-server-configuration-option?view=sql-server-ver15

[10] Microsoft. (2019) Microsoft SQL Server Licensing Guide. https://download.microsoft.com/download/6/6/0/66078040-86d8-4f6e-b0c5-e9919bbcb537/SQL%20Server%202019%20Licensing%20guide.pdf

[11] Todd Muirhead. (2021, June) dvdstore/ds35. https://www.github.com/dvdstore/ds35

About the Author

Todd Muirhead is a member of the VMware Performance Engineering team where he works on database, storage, and processor performance. He is also one of the co-creators and maintainers of the DVD Store open-source benchmark used in this study.

Acknowledgements

Special thanks to David Morse, James Zubb, and Dave Jaffe.