Peeking At the Future with Giant Monster Virtual Machines

Project Capstone: A Performance Study of Running Many Large Virtual Machines in Parallel

TECHNICAL WHITE PAPER

Hewlett Packard Enterprise

IBM

vmware

Table of Contents

Executive Summary

This technical white paper examines the extraordinary possibilities available when leading edge servers and storage push the boundaries of current technology in terms of capacity and performance. Tests were run with many different configurations of extremely large virtual machines (known as monster virtual machines) and results show that VMware vSphere® 6 successfully managed to run all of the virtual machines in a high performing and efficient manner. vSphere 6 is ready to run the largest systems and workloads of today with great performance. vSphere 6 is also ready for the future to take on the high performant systems and workloads that will become more common in datacenters.

Introduction

The rate of increases in capacity and performance of computing is dramatic. Starting in 1975, Moore's law observed that the number of transistors in an integrated circuit doubles every two years. This doubling of transistors has translated into chip performance also doubling every two years. VMware vSphere has also rapidly increased its capacity to support larger virtual machines and hosts to keep up with this compute capacity that increases over time.

Two- and four-socket x86-based servers are commonly used today. While the number of cores per socket in these servers does not exactly follow Moore's law (because each core itself is more powerful with each generation of processors) it can be used as a rough proxy. The current generation Intel Xeon chips have a maximum of 18 cores per socket and 36 logical processes with hyper-threading enabled. This is almost a doubling of the 10 cores per socket in Xeon chips from two generations and about four years before.

Many four-socket servers that use the current generation Intel Xeon processors have 72 cores, but the HPE Superdome X has 16 sockets with 240 cores. By using this cutting edge server, it is possible to have the type of compute capacity in a single server that, by following Moore's Law, won't be available in a four-socket server for many years. It is a peek into the future.

Project Capstone

Project Capstone brings together VMware, HPE, and IBM in a unique opportunity to combine each of these industry leading companies and their respective leading-edge technologies to build a test environment that shows the upper bounds of what is currently possible with such giant compute power. Running numerous heavy workloads on monster virtual machines on a vSphere 6, HPE Superdome X, and IBM FlashSystem configuration in parallel exemplifies the present capabilities of these combined technologies.

Project Capstone became a centerpiece of the 2015 VMware conference season as it occupied center stage at VMworld US in San Francisco as the subject of a highly anticipated Spotlight Session that included individual presentations from senior management of VMware, HP, and IBM. VMworld 2015 Europe in Barcelona included a Capstone-themed breakout session as well. But perhaps most significantly, the VMworld floor presence at Oracle Open World in San Francisco in October brandished a complete demo version of the Capstone Stack to include the Superdome X as well as the IBM FlashSystem.

VMware vSphere 6.0

VMware vSphere 6.0 includes new scalability features that enable it to host extremely large and performance-intensive applications. The capabilities of individual virtual machines has increased significantly from the previous versions of vSphere. A single virtual machine can now have up to 128 vCPUs and 4TB of memory. While these levels of resources are not commonly required, there are some large applications that do require and make use of resources at this scale. These are usually the last applications to be considered for virtualization due to their size, but it is now possible to move this last tier of applications into virtual machines.

HPE Superdome X

HPE Integrity Superdome X sets new, high standards for x86 availability, scalability, and performance; it is an ideal platform for critical business processing and decision support workloads. Superdome X blends x86 efficiencies with proven HPE mission-critical innovations for a superior uptime experience with RAS (reliability, availability, and serviceability) features not found in other x86 platforms, allowing this machine to achieve five nines (99.999%) of availability. Breakthrough scalability of up to 16 sockets can handle the largest scaled-up x86 workloads. Through the unique **nPars** technology, HPE Superdome X increases reliability and flexibility by allowing for electrically isolated environments to be built within a single enclosure [1]. It is a well-balanced architecture with powerful Xeon processors working in concert with high I/O and a large memory footprint that enables the virtualization of large and critical applications at an unprecedented scale.

Whether you want to maximize application uptime, standardize, or consolidate, HPE Superdome X helps virtualize mission-critical environments in ways never before imagined.

The HPE Superdome X is the ideal system for Project Capstone because it is uniquely suited to act as the physical platform for such a massive virtualization effort. The ability of vSphere 6 to scale up to 128 virtual CPUs can be easily realized on the HPE Superdome X because it allows for massive, individual virtual machines to be encapsulated on a single system while huge aggregate processing is parallelized.

IBM FlashSystem

The IBM FlashSystem® family of all-flash storage platforms includes IBM FlashSystem 900 and IBM FlashSystem V9000 arrays. Powered by IBM FlashCore™ technology, the FlashSystem 900 delivers the extreme performance, enterprise reliability, and operational efficiencies required to gain competitive advantage in today's dynamic marketplace.

Adding to these capabilities, FlashSystem V9000 offers the advantages of software-defined storage at the speed of flash. These all-flash storage systems deliver the full capabilities of FlashCore technology's hardware accelerated architecture, MicroLatency™ modules, and advanced flash management coupled with a rich set of features found in only the most advanced enterprise storage solutions, including IBM Real-time Compression™, virtualization, dynamic tiering, thin provisioning, snapshots, cloning, replication, data copy services, and high-availability configurations.

While virtualization lifts the physical restraints on the server room, the overall performance of multi-workload server environments enabled by virtualization are held back by traditional storage because disk-based systems struggle with the challenges posed by the resulting I/O. consolidated. As virtualization has enabled the consolidation of multiple workloads run on a fewer physical servers, disks simply can't keep up, and this limits the value enterprises gain from virtualization.

IBM FlashSystem V9000 solves the storage challenges left unanswered by traditional storage solutions. It handles random I/O patterns with ease, and it offers the capability to virtualize all existing data storage resources and bring them together under one point of control. FlashSystem V9000 provides a comprehensive storage solution that seamlessly and automatically allocates storage resources to address every application demand. It moves data to the most efficient, cost-effective storage medium—from flash, to disk, and even to tape—without disrupting application performance or data availability, and more capacity can be added without application downtime or a lengthy update process. IBM FlashSystem V9000 helps enterprises realize the full value of VMware vSphere 6.

Test Environment

The test environment was designed to allow testing for extremely large monster virtual machines. vSphere 6 provides the capability to host virtual machines of up to 128 vCPUs. This is the foundation for running larger monster virtual machines than in the past. The HPE Superdome X and IBM FlashSystem storage array provided the hardware server and storage platforms respectively. The Superdome X used in this project had 240 cores and

480 logical threads with hyper-threading enabled. This was coupled with 20TB of extremely low latency, all-flash storage within the IBM FlashSystem array. A four-socket server was used as a client load driver system for the testbed. The diagram below shows the testbed setup.

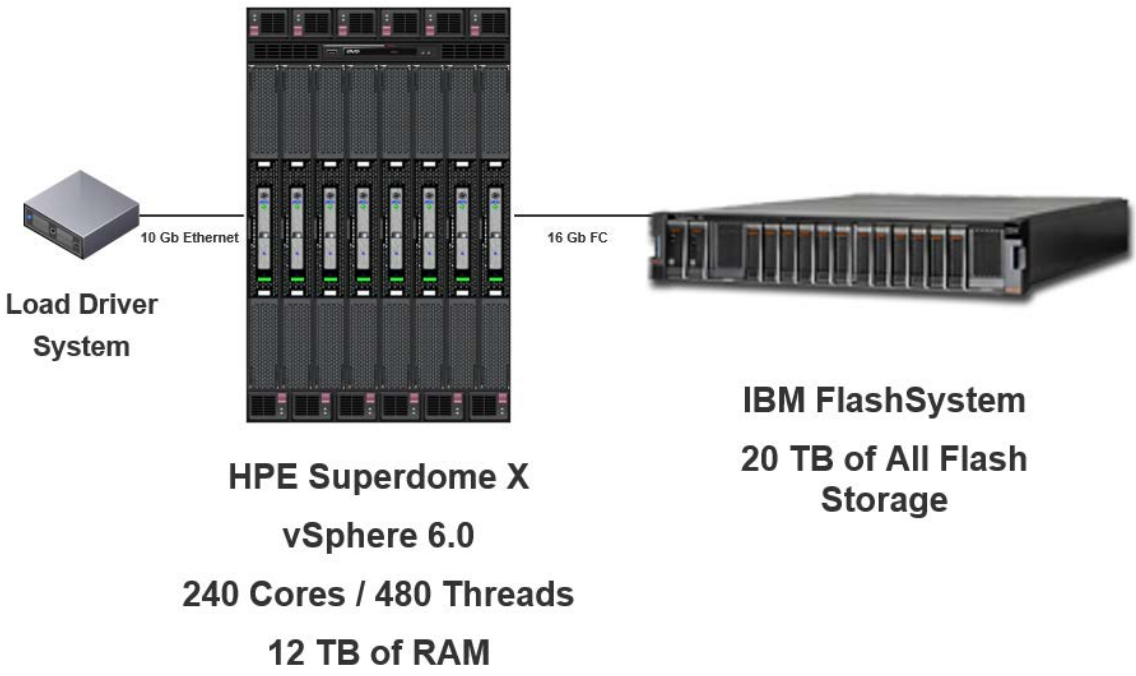


Figure 1. Testbed hardware

Test Configuration Details

HPE Superdome X Server:

- vSphere 6.0
- 16 Intel Xeon E7-2890 v2 2.8GHz CPUs (15 cores per CPU)
- 240 cores / 480 threads (hyper-threading enabled)
- 12TB of RAM
- 16Gb Fibre Channel
- 10Gb Ethernet

IBM FlashSystem 900:

- 20TB capacity
- All-flash memory
- 16Gb Fibre Channel

Client load driver server:

- 4 x Intel Xeon E7-4870 2.4 GHz
- 512GB of RAM
- 10Gb Ethernet

Virtual Machine Configuration

The configuration of the virtual machine was kept constant in all tests except for the number of virtual CPUs and related virtual NUMA. In all tests, the total number of vCPUs across all virtual machines under test was equal to the number of cores or hyper-threads on the server. In the maximum size virtual machine test case, there were four virtual machines each with 120 vCPUs for a total of 480 vCPUs assigned on the server. This matches the 480 hyper-threads available on the server. Table 1 shows the number of virtual machines with their vCPU configurations that were tested.

Number of VMs	vCPUs per VM	Virtual Sockets per VM	Total vCPUs Assigned on Server	Total Physical Threads On Server With HT Enabled
4	120	4	480	480
8	60	2	480	480
16	30	1	480	480

Table 1. Virtual machine configuration

The configuration parameter PreferHT was used for these tests to optimize the use of the system's hyper-threads in this high CPU utilization benchmark. By default, vSphere 6 schedules each vCPU on a core where another vCPU is not scheduled. In other words, vSphere will not use the second thread that is created on each core with hyper-threading enabled until there is a vCPU already scheduled on all of the physical cores on the system. Using the PreferHT parameter changes this and instructs the scheduler for a virtual machine to prefer to use hyper-threads instead of physical cores.

The best performance for two vCPUs would be to use a thread from two physical cores, and this is the default scheduling behavior. Using two threads of the same core results in lower performance because hyper-threads share most of the resources of the physical core.

However, in the case of high overall system utilization, all threads on all cores are in use at the same time. PreferHT provides a performance advantage because each virtual machine is spread across fewer NUMA nodes and this results in increased NUMA memory locality. By using PreferHT, a highly utilized system becomes more efficient because the virtual machines all have more NUMA locality while still using all the logical threads on the server.

Standard best practices for database virtual machines were used for the configuration. Each virtual machine was configured with 256GB of RAM, two pvSCSI controllers, and a single vmxnet3 virtual network adapter.

The virtual machines were installed with Red Hat Enterprise Linux 6.5 as the guest operating system. Oracle 12c was installed following the installation guide from Oracle.

Test Workload

The open source database workload DVD Store 3 was used for these tests [2]. DVD Store simulates an online store that allows customers to log in, browse products, read and submit product reviews, and purchase products. It uses many database features to run the database including primary keys, foreign keys, full text indexing and searching, transactions, rollbacks, stored procedures, triggers, and simple and complex multi-join queries. It is

designed to be CPU intensive, but also requires low latency storage in order to achieve good throughput. DVD Store includes a driver program that simulates user activity on the database. Each simulated user steps through the full process of an order: log in, browse the DVD catalog, browse product reviews, and purchase DVDs. Performance is measured in orders per minute (OPM).

DVD Store 3, which was recently updated from version 2, adds product reviews and a few other features that are designed to make the workload include the typical product reviews commonly found today on many Web sites, and version 3 is also more CPU intensive. The increased CPU usage makes it possible for a DVD Store 3 instance to fully saturate larger systems more easily than what was possible with the previous version of DVD Store.

For these tests, a 40GB DVD Store 3 database instance was created on each virtual machine. The direct database driver was used on the client load system to stress the database without running a middle tier because the focus of these tests was on the large database virtual machines. The database buffer cache was set to same size as the database to optimize performance. The number of driver threads running against each monster virtual machine was increased until the maximum OPM began to decrease. At the point of maximum OPM the CPU usage and other performance metrics were checked to verify that the system had reached saturation.

Monster Virtual Machine Tests

A series of tests were run with different sizes of virtual machines. Each test is briefly described with the results and analysis. The first tests discussed are all similar in that each configuration is a set of virtual machines that fully consume all the CPU threads on the host. The configurations are four 120-vCPU VMs, eight 60-vCPU VMs, and 16 30-vCPU VMs. In each case, the total number of vCPUs running across all the virtual machines is 480, which equals the number of CPU threads on the host.

In addition to these tests with maximum configurations, some tests were run with a virtual machine configuration that under-provisions the server, and a test comparing CPU affinity (pinning) vs. PreferHT configurations.

Storage Performance

For these tests, the goal was to use all the CPUs on the server. In order to accomplish this, the amount of disk I/O was minimized by specifying a database buffer cache that was approximately the same size as the database on disk. This meant that after the initial warmup phase of running the test, most database queries could be satisfied without a disk I/O operation because most of the database was cached in memory.

In order for all CPUs to be kept busy, the disk I/O operations that occur must be as low latency as possible. The IBM FlashSystem array was able to keep average disk latency below 0.3 milliseconds in all tests and was a highlight of system performance. IOPS peaked at approximately 50,000 during some of the test runs, which was well within the capabilities of the storage array. The array provided extremely low latency storage in all test scenarios. The capabilities of the IBM FlashSystem array in terms of IOPS were never pushed, but the tests did benefit greatly from the consistently low response times.

Four 120-vCPU VMs

The maximum size virtual machine in vSphere 6 is 128 vCPUs. So with the limit of 480 total CPU threads on the host, running four 120-vCPU VMs is the maximum size possible while keeping all virtual machines the same size and staying under the vSphere maximum. While not many environments today have a single virtual machine running at this size, this test ran four of them on a single host under high load.

To measure the scalability of the solution at full capacity, tests were run first with just a single monster virtual machine. In additional tests, all four virtual machines were run at the same time. Maximum performance was found for each test case by increasing the number of threads in the client drivers to find the point at which the most orders per minute (OPM) were achieved. This point of maximum throughput was also found to be at near CPU saturation, indicating that performance had peaked.

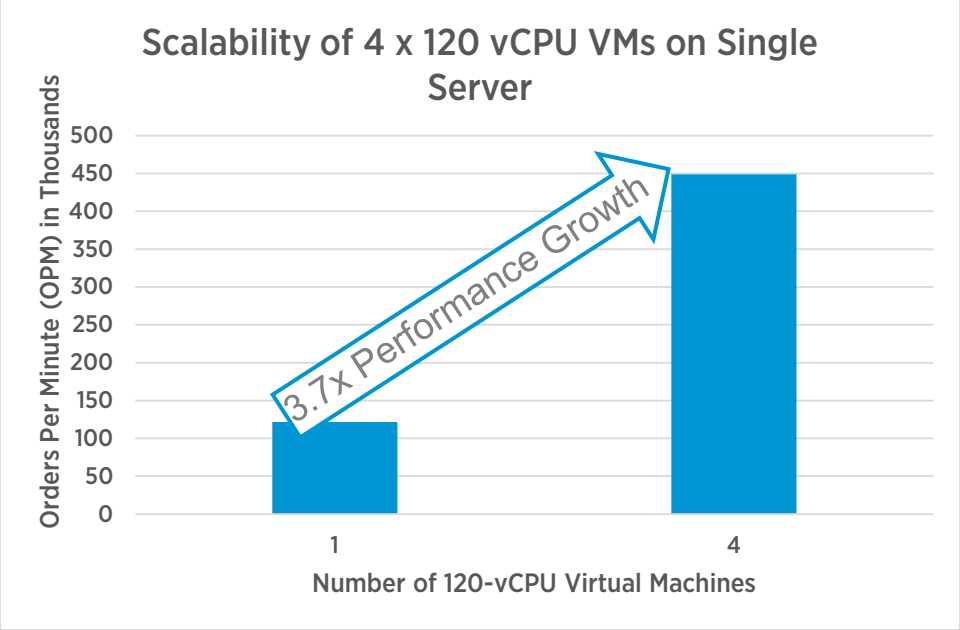


Figure 2. Almost linear scalability of 4 x 120 vCPU VMs on a single server

In this type of test, the ideal is linear scalability. This would be a 4 times performance growth going from a single virtual machine to four virtual machines. As Figure 2 shows, the four 120-vCPU virtual machines achieved 3.7 times the throughput of the single 120-vCPU virtual machine, which is 92% of perfect linear scalability. Storage performed at a very high level maintaining 0.3 milliseconds latency and 20,000 IOPS during the test.

Eight 60-vCPU VMs

For the next set of tests, the virtual machines were readjusted down to 60 vCPUs and cloned so that there was a total of eight 60-vCPU VMs. In this test, each virtual machine had the same number of vCPUs as each Superdome X compute blade. It is by no means a requirement to match virtual machine size to the underlying hardware so specifically, but this can allow for optimized results in some environments.

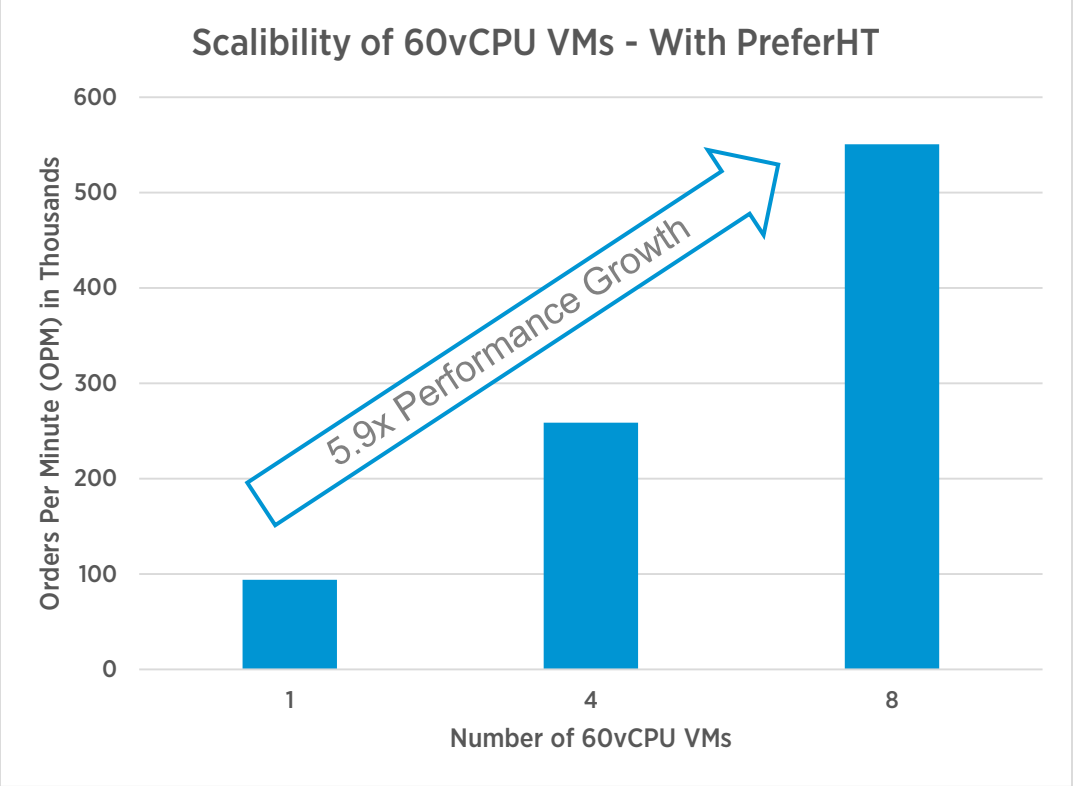
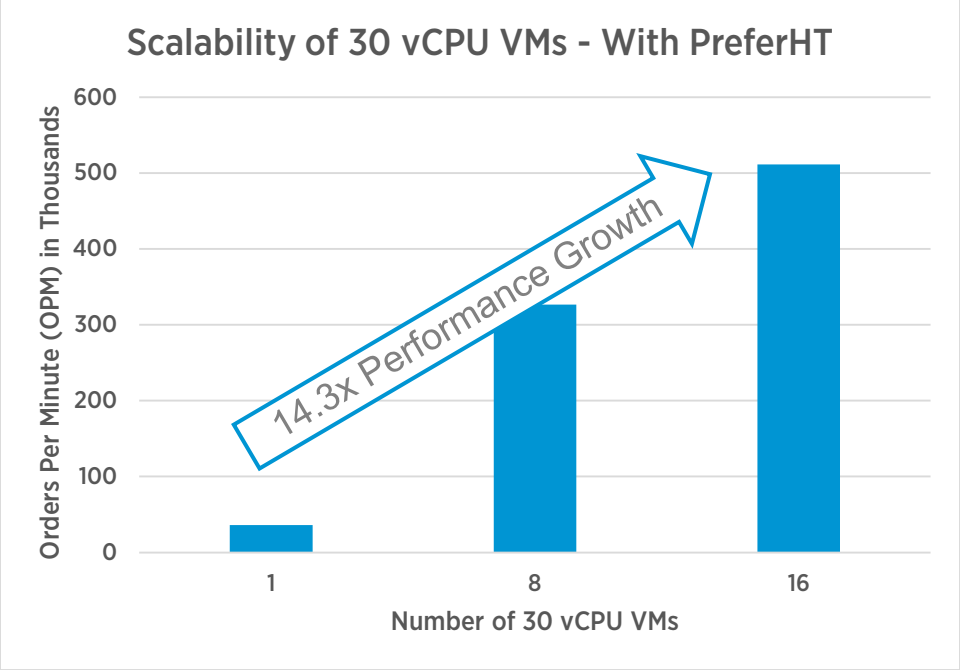


Figure 3. Scalability of 60-vCPU VMs

The total throughput achieved with eight 60-vCPU VMs was the highest of any of the tests conducted. The IBM FlashSystem array also continued to achieve impressive performance with latency under 0.3 milliseconds and average 16,000 IOPS.

Sixteen 30-vCPU VMs

This test consisted of running a sixteen 30-vCPU VMs. Each of these 30-vCPU VMs was essentially using all the threads on a server socket because of the use of the preferHT parameter.



This large number of monster VMs running at the same time still resulted in very good total throughput and excellent scalability moving from a single virtual machine running to all sixteen. The throughput of the sixteen virtual machines was 14.3 times of a single VM or 89% of perfect linear scalability. Once again, disk latency remained below 0.3 milliseconds with average IOPS of 13,000.

Under-Provisioning with Four 112-vCPU VMs

In the other tests covered in this paper, the server has been fully committed with a vCPU allocated for every thread on the host. This means that all threads will be used for virtual machine vCPUs. In most environments, this still leaves lots of CPU available because not all virtual machines are running at full CPU utilization. In an environment where all assigned vCPUs are at 100% usage, there isn't anything left over for the ESXi hypervisor to use for its functions. This includes virtual networking and disk I/O handling. The hypervisor then competes directly with the virtual machines for CPU.

In this case, performance of the virtual machines can actually be improved by reducing the number of vCPUs to leave some CPU threads available on the host for the use of ESXi. In this specific configuration, while running the 4 x 120 virtual CPUs with the DVD Store 3 workload, the network traffic is about 900 Megabits per second (Mb/s) transmitted and 200 Mb/s received and an average 30,000 of disk IOPS is also being processed. In order to allow for the host to have some CPU resource available to handle this workload, the number of vCPUs for each of the four virtual machines was reduced from 120 to 112. This leaves one core (two hyper-threads) per socket unassigned to a virtual machine.

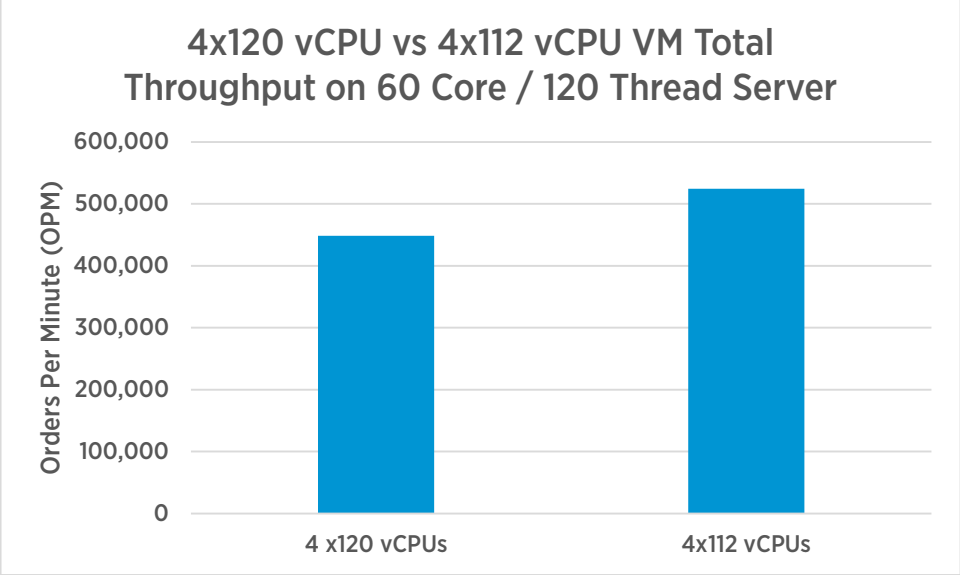


Figure 4. 4x120 vCPU vs. 4x112 vCPU

The results show that overall throughput increased significantly from 448 thousand to 524 thousand OPM. The gain in performance with smaller virtual machines is due to the reduction in contention of resources between the ESXi hypervisor and the virtual machines that is found in this extreme testing scenario when all CPU resources were allocated and fully utilized.

CPU Affinity vs. PreferHT

It is possible to control the CPUs that are used for a virtual machine by using the CPU affinity setting. This allows an administrator to override the ESXi scheduler and only allow a virtual machine to use specific physical cores. The vCPUs used by a virtual machine are pinned to specific physical cores.

In certain benchmarking scenarios, the use of CPU affinity has shown small increases in performance. Even in these relatively uncommon cases, its use is not recommended because of the high administrative effort and the potential for poor performance if the setting is not updated as changes in the environment occur or if the CPU affinity setting is done incorrectly.

Using the Capstone testing environment a test with CPU affinity and PreferHT was conducted to measure which configuration performed better. It was found that PreferHT, which allows the ESXi hypervisor to make all vCPU scheduling decisions, outperformed a CPU affinity configuration by 4%.

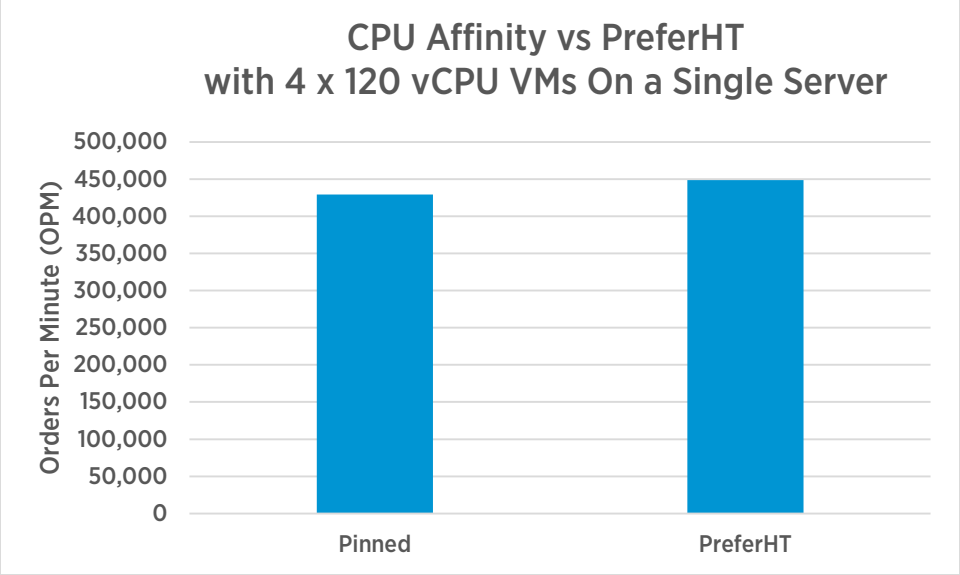


Figure 5. Using PreferHT performed slightly better than setting CPU affinity

Best Practices

Running many very large virtual machines on an even larger server makes it more important to follow monster virtual machine best practices.

Consider the server's NUMA architecture when deciding what size to make the virtual machines. When creating virtual machines, make sure the virtual NUMA sockets match the physical NUMA architecture of the host as closely as possible. For more information, see "Using NUMA Systems with ESXi" [3].

Size and configure storage with enough performance to match the large performance capability of the monster virtual machines. A large server with underpowered storage will be limited by the storage.

Network performance can quickly become an issue if the traffic for the large virtual machines is not correctly spread across multiple NICs. Combining a number of high performance workloads on a single host will also result in high network traffic that will most likely need to use multiple network connections to avoid a bottleneck.

In extremely high CPU utilization scenarios, including benchmark tests, it can be better to leave a few CPU cores unassigned to virtual machines to give the ESXi hypervisor needed resources for its functions.

Do not use CPU affinity, sometimes referred to as CPU pinning, because it usually does not result in a big increase in performance. In some extreme high utilization scenarios, use the PreferHT setting to get more total performance from a system, but note that using this setting could reduce individual virtual machine performance.

Conclusion

Project Capstone has shown that vSphere 6 is capable of running multiple giant monster virtual machines today on some of the world's most capable servers and storage. The HPE Superdome X and super low latency IBM FlashSystem storage were chosen because of their tremendous performance capabilities, their ease of configuration and use, and their overall complimentary stature to vSphere 6. The unique properties of this stack allowed the testing team to push the limits of virtualized infrastructure to never before seen levels. As stated in the media collateral "Project Capstone, Driving Oracle to Soar Beyond the Clouds," (see Appendix) this example infrastructure stack is possible today and shows that as higher core counts and all-flash storage arrays become more common in the future, a VMware vSphere–based approach will provide the needed scalability and capacity.

This collaboration of VMware, HPE, and IBM shows that applications of the largest sizes can run on a vSphere virtual infrastructure. The limiting factor in most datacenters today is the hardware, but when using the latest technology available, it is possible to lift these limits and bring the flexibility and capabilities of virtualized infrastructure to all corners of the datacenter.

This collaborative achievement between three of the world's most recognized computing companies has solidified the proposition of comprehensive virtualization that VMware has held for a number of years. Very simply put, all applications and databases—regardless of their processing, memory, networking, or throughput demands—are candidates for a virtualized infrastructure. VMware, HPE, and IBM built Project Capstone with leading edge components used as a foundation to prove that 100% virtualization is a reality in even the largest compute environment.

Appendix

An initial blog for Project Capstone was previously published [4].
http://blogs.vmware.com/vsphere/2015/08/vmworld-us-2015-spotlight-session-project-capstone-a-collaboration-between-vmw-hp-ibm-no-application-left-behind.html

A short video on Project Capstone that gives some highlights from the project is available online [5].
https://www.youtube.com/watch?v=X4SRxl04uQ0

Project Capstone was presented at VMworld 2015 in San Francisco and with executives from all three companies participating. A video of this presentation is available online [6].
https://www.youtube.com/watch?v=O3BTvP46i4c

References

[1] Hewlett-Packard Development Company, L.P. (2010) HP nPartitions (nPars), for Integrity and HP 9000 midrange. http://www8.hp.com/h20195/v2/GetPDF.aspx/c04123352.pdf

[2] Todd Muirhead and Dave Jaffe. (2015, July) DVD Store 3. http://www.github.com/dvdstore/ds3

[3] VMware, Inc. (2015) Using NUMA Systems with ESXi. http://pubs.vmware.com/vsphere-60/index.jsp#com.vmware.vsphere.resmgmt.doc/GUID-7E0C6311-5B27-408E-8F51-E4F1FC997283.html

[4] Don Sullivan. (2015, August) VMworld US 2015 Spotlight Session: Project Capstone, a Collaboration between VMW, HP & IBM. http://blogs.vmware.com/vsphere/2015/08/vmworld-us-2015-spotlight-session-project-capstone-a-collaboration-between-vmw-hp-ibm-no-application-left-behind.html

[5] IBM Systems ISVs. (2015, November) Project Capstone - Pushing the performance limits of virtualization. https://www.youtube.com/watch?v=X4SRxl04uQ0

[6] VMworld. (2015, November) VMworld 2015: VAPP6952-S - VMware Project Capstone, a Collaboration of VMware, HP, and IBM. https://www.youtube.com/watch?v=O3BTvP46i4c

[7] VMware, Inc. (2015) Configuration Maximums vSphere 6.0. https://www.vmware.com/pdf/vsphere6/r60/vsphere-60-configuration-maximums.pdf

About the Authors

Leo Demers, Mission Critical Product Manager, HPE

Kristy Ortega, EcoSystem Offering Manager, IBM

Rawley Burbridge, FlashSystem Corporate Solution Architect, IBM

Todd Muirhead, Staff Performance Engineer, VMware

Don Sullivan, Product Line Marketing Manager for Business Critical Applications, VMware

Acknowledgements

The authors thank Mark Lohmeyer, Michael Kuhn, Randy Meyer, Drew Sher, Rawley Burbridge, Bruce Herndon, Jim Britton, Reza Taheri, Juan Garcia-Rovetta, Michelle Tidwell, and Joseph Dieckhans.

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