White Paper

MICROSOFT HYPER-V SCALABILITY WITH EMC SYMMETRIX VMAX

Abstract

This white paper highlights EMC's Hyper-V scalability test in which one of the largest Hyper-V environments in the world was created. This testing was performed to demonstrate how well such an environment scales when deployed on an EMC® Symmetrix VMAX™ enterprise storage platform. Descriptions of technologies that can be applied, such as storage tiering, provisioning, and replication, show how Symmetrix® features and EMC software products are used to provide scalable, reliable, and highly available virtualization solutions to assist customers on their journey to the cloud.

December 2010

EMC believes the information in this publication is accurate of its publication date. The information is subject to change without notice.

The information in this publication is provided “as is”. EMC Corporation makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com.

All other trademarks used herein are the property of their respective owners.

Part Number h8130

Table of Contents

Executive summary

For many customers, there has been a growing need to provide more and more physical server deployments to service increasing business needs. This has subsequently led to a number of inefficiencies in operational areas:

- Servers and storage have typically been overprovisioned in terms of CPU and memory resources.
- Each additional server requires additional data center floor space and increases both power and cooling costs

This server sprawl can quickly become unmanageable, resulting in increased cost, lower resource utilization, management complexity, and impact on ROI.

With the release of the Windows Server 2008 operating system, Microsoft introduced the latest in its series of server virtualization solutions. This hypervisor, known as Hyper-V, is available on Windows Server 2008 or Microsoft Windows Server 2008 R2. Additionally, the hypervisor functionality is available as a free download as Windows Hyper-V Server 2008 and Windows Hyper-V Server 2008 R2. The Hyper-V Server product is a bare metal install option. Hyper-V products are only available for the 64-bit (x64) release of Microsoft Windows Server 2008, and require that the server hardware platform supports hardware assisted virtualization (Intel VT or AMD-V). As of Windows Server 2008 R2, Windows Server products are only available for x64 hardware. Support for the x86 platform ends with the Windows Server 2008 release.

Hyper-V provides customers an ideal platform for key virtualization scenarios, such as production server consolidation, business continuity management, software test and development, and development of a dynamic data center. Scalability and high performance can be achieved by supporting features like guest multi-processing support and 64-bit guest and host support, while features like quick migration of virtual machines from one physical host to another, and integration with System Center Virtual Machine Manager, provide users with flexibility and ease of use.

Hyper-V allows customers to achieve significant space, power, and cooling savings while maintaining availability and performance targets. EMC® Symmetrix® storage systems are able to provide additional value to customers by providing the ability to consolidate storage resources, implement advanced high-availability solutions, and provide seamless multi-site protection of customer data assets.

As customers seek to consolidate data center operations, Microsoft's Hyper-V hypervisor provides a scalable solution for virtualization on the Windows Server platform. To further facilitate cost savings, large-scale consolidation efforts can benefit by optimizing and consolidating storage resources to a single storage repository. Additionally, many of the advanced features of the Hyper-V environment are either facilitated by, or enhanced with, the implementation of a scalable storage array.

Beyond providing protection and performance requirements through core system performance and RAID protection, Symmetrix arrays provide complementary technologies for Hyper-V environments that improve dynamic placement capabilities for Hyper-V landscapes. High availability and multi-site disaster protection can be transparently integrated to produce comprehensive solutions for customer deployments, providing significant value-added solutions for consolidated Hyper-V deployments on Symmetrix storage systems.

EMC Symmetrix VMAX™ storage arrays easily scale to meet the demands of large-scale consolidation efforts. With support of thousands of connected hosts, presentation of tens of thousands of logical units, and advanced internal mechanisms such as snapshot and clone operations, and with multi-site replication solutions to provide disaster restart/recovery solutions, Symmetrix systems are a central part of Windows Server consolidation efforts.

Introduction

This white paper presents methods of scaling a large Microsoft Hyper-V cluster configuration with Symmetrix VMAX. It shows how a 16-node Microsoft Failover Cluster running Hyper-V is scaled to support 64 virtual machines per cluster node, and how to use various methods to quickly duplicate and deploy them. It shows I/O performance on the array correlates to the Hyper-V environment as each node is brought online and I/O is executed from each virtual machine. It concludes by describing approaches to optimize performance in such an environment.

Audience

This white paper is intended for Microsoft Windows Server 2008 administrators, storage architects, customers, and EMC field personnel who want to understand the implementation of Hyper-V solutions on EMC Symmetrix storage platforms.

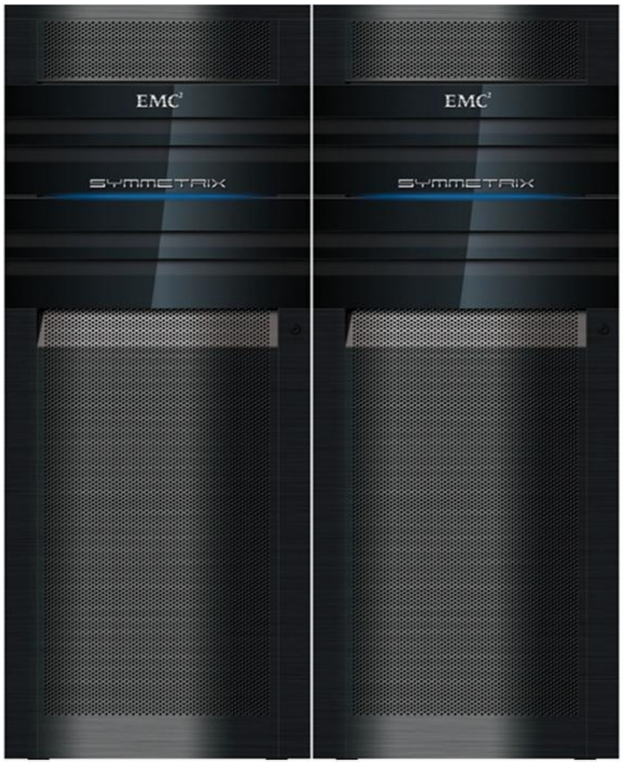
Technology overview

Symmetrix overview

The EMC Symmetrix VMAX Series with Enginuity™ is the latest generation of the Symmetrix product line. Built on the strategy of simple, intelligent, modular storage, it incorporates a new scalable fabric interconnect design that allows the storage array to seamlessly grow from an entry-level configuration into the world's largest storage system. Symmetrix VMAX arrays provide improved performance and scalability for demanding enterprise storage environments such as those found in large virtualization environments, while maintaining support for EMC's broad portfolio of platform software offerings.

Symmetrix VMAX systems now deliver new software capabilities that improve capacity utilization, ease of use, business continuity, and security. These features provide significant advantage to customer deployments in a virtualized environment, where ease of management and protection of virtual machine assets and data assets are required.

Symmetrix VMAX arrays extend the scalability of previous generations of Symmetrix DMX™ technology, by providing a superior level of scalability, and support for a broad new range of drive technologies as detailed in Figure 1. Additionally, Symmetrix VMAX offers the ultimate in flexibility, including the ability to incrementally increase back-end performance by adding VMAX Engines and storage bays. Each high-availability VMAX Engine controls eight redundant Fibre Channel loops that support up to either 240 or 360 drives depending upon configuration. Subsequently, each high-availability VMAX Engine provides front-end as well as back-end connectivity, providing enhanced scalability.



- 2 to 16 Director boards
- Up to 2.1 PB usable capacity
- Up to 128 FC front-end ports
- Up to 64 FICON front-end ports
- Up to 64 Gig-E / iSCSI front-end ports
- Up to 472 GB global memory (Mirrored)
- 48 to 2400 disk drives
- Enterprise Flash Drives: 200/400 GB
- Fibre Channel drives
  - 146/300/450 GB 15,000 rpm
  - 400 GB 10,000 rpm
- SATA II drives 1 TB 7,200 rpm

Figure 1. Symmetrix VMAX hardware scalability

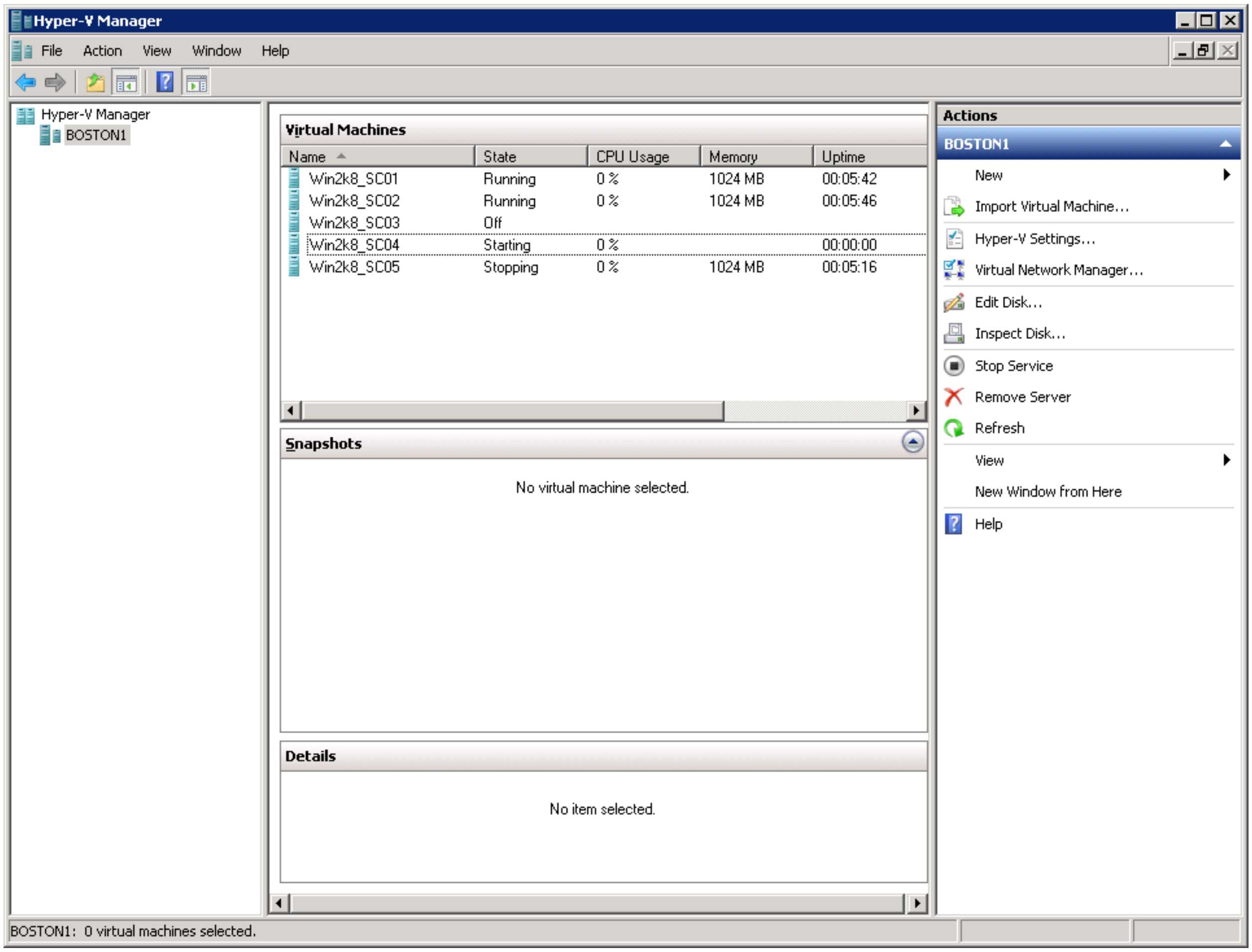
The Symmetrix VMAX storage systems also meet customer expectations for high-end storage in terms of availability. High-end availability is more than just redundancy; it means nondisruptive operations and upgrades, and being “always online.” Beyond previous Symmetrix generations, Symmetrix VMAX arrays provide:

- Nondisruptive expansion of capacity and performance at a lower price point
- Sophisticated migration for multiple storage tiers within the array
- The power to maintain service levels and functionality as consolidation grows
- Simplified control for provisioning in complex environments

Microsoft Hyper-V hypervisor

Microsoft Windows Server 2008 provides the Hyper-V server role on the applicable versions of Windows Server. In the initial release of Windows Server 2008, separate product releases included the Hyper-V role, and were required to be ordered explicitly. When a Windows Server instance has the Hyper-V role installed, the original operating system instance is referred to as the "parent partition."

When the Hyper-V server role is installed, the Windows Hyper-V virtualization hypervisor is installed for the parent partition. Utilizing the functionality implemented by the hypervisor, and managed through the Hyper-V Manager Management Console (MMC) shown in Figure 2, it is possible to define virtual machine instances.



Name	State	CPU Usage	Memory	Uptime
Win2k8_SC01	Running	0 %	1024 MB	00:05:42
Win2k8_SC02	Running	0 %	1024 MB	00:05:46
Win2k8_SC03	Off			
Win2k8_SC04	Starting	0 %		00:00:00
Win2k8_SC05	Stopping	0 %	1024 MB	00:05:16

No virtual machine selected.

No item selected.

BOSTON1: 0 virtual machines selected.

Figure 2. Hyper-V Manager Management Console

In more complicated Hyper-V deployments that may be comprised of a large number of physical servers, and a multitude of virtual machine instances, customers can use Microsoft System Center Virtual Machine Manager 2008 (SCVMM). The SCVMM solution provides a comprehensive management framework with centralized command and control features. SCVMM also includes additional functionality in the form of the Performance and Resource Optimization (PRO) subsystem. The PRO functionality has a dependency on Microsoft System Center Operations Manager

(SCOM), allowing customers to build automatic and dynamic management capabilities into a Hyper-V landscape. Such configurations may allow for dynamic placement of virtual machine resources based on changing characteristics of the data center. More information regarding SCVMM and its integrated functionality can be found at http://www.microsoft.com/systemcenter.

Further detailed information about Microsoft Hyper-V with EMC Symmetrix can be found in the white paper EMC Symmetrix with Microsoft Hyper-V Virtualization on EMC.com and Powerlink®.

Goal of testing

The goal of EMC's Hyper-V Scalability test was to build and test one of the largest Hyper-V environments in the world to demonstrate how well such an environment scales when deployed on a Symmetrix VMAX enterprise storage platform. The desire was to use a combination of EMC and Microsoft technologies to show the many options available to customers who are looking at virtualization as a way to reduce the cost of doing business.

Environment description

In order to create a 16-node cluster capable of generating 64 virtual machines per node, and to drive high I/O rates to the Symmetrix array, EMC chose new servers with multi-core, hyper-threaded x64 CPUs with virtualization capability enabled. Each server had 96 GB of RAM and a single, dual-port, 8 GB Fibre Channel HBA to connect to a Fibre Channel fabric to be zoned to the VMAX array (Figure 3). The servers were also configured to boot from the array in order to take advantage of the array's data-protection and high-availability functions as well as simulate a low-cost customer environment, saving money on internal server hard disks.

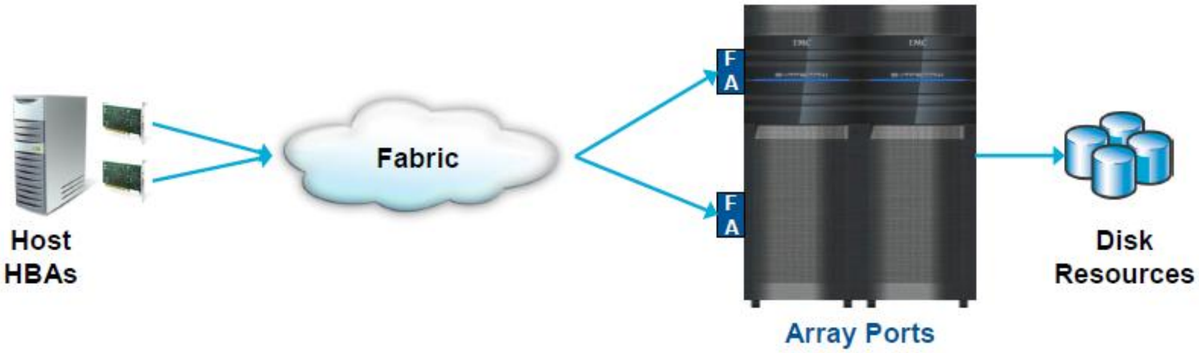
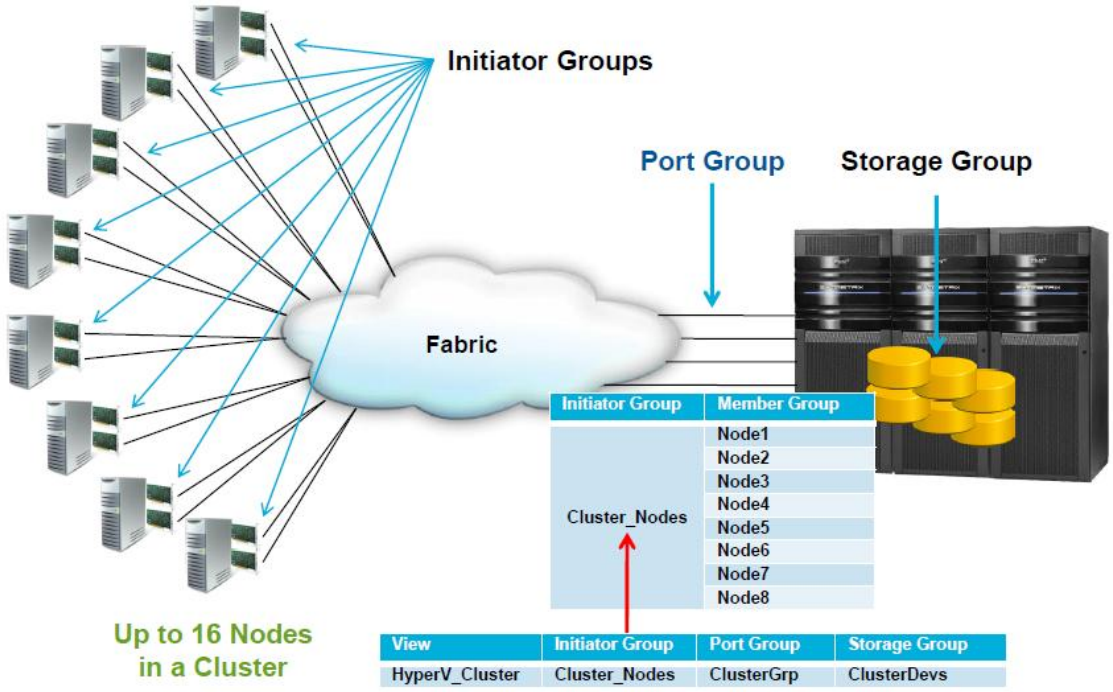


Figure 3. Server connectivity

EMC Auto-provisioning Groups were used to provision storage to each Hyper-V parent node. This method reduces the complexity of storage connectivity by putting storage devices, host initiators, and array ports into a defined "view" that links the devices, initiators, and arrays ports together. This allows simple control over how the Hyper-V nodes accessed the storage devices assigned to them (Figure 4).



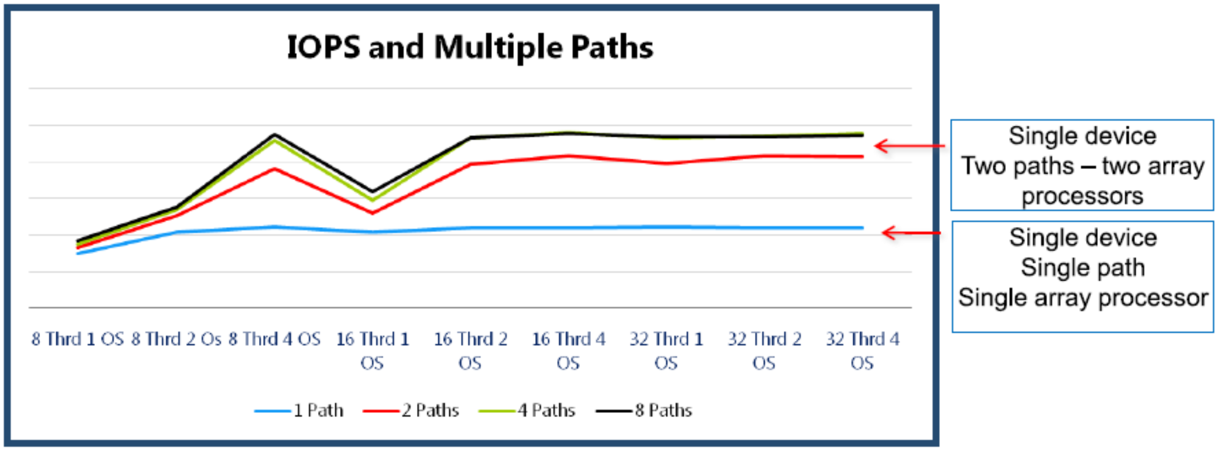
Initiator Group	Member Group
Cluster_Nodes	Node1
	Node2
	Node3
	Node4
	Node5
	Node6
	Node7
	Node8

View	Initiator Group	Port Group	Storage Group
HyperV_Cluster	Cluster_Nodes	ClusterGrp	ClusterDevs

Figure 4. Cascading initiator groups into a single view

When configuring the VMAX array for this large-scale environment, it was important to determine how the workload from the virtual machines would be consolidated on the array's devices. Could the LUNs support the I/O load? Could the LUN be scaled to accommodate the virtual machines that would be accessing it simultaneously? These questions were considered along with how the I/O stack from the server HBA down to the disks within the array themselves would be affected by the overall workload.

To improve performance of the parent nodes, EMC PowerPath® multipathing software was used. Figure 5 shows how higher performance is achieved through the use of multiple paths to separate array ports and processors.



Single device
Two paths – two array processors

Single device
Single path
Single array processor

Figure 5. Single path vs. multipath performance

Building the environment

The servers were configured with Microsoft Windows Server 2008 R2 and failover clustering and Hyper-V installed. For simplicity, a single server was installed and configured, and then the boot image was duplicated across all boot LUNs within the array. This allowed all 16 nodes in the cluster to be configured identically. Manual intervention was required to give each server node a unique name and network characteristics.

HBA zoning and storage provisioning were done in a way that spread the load over the array's front-end director ports and attempted to limit connectivity to a single processor on each director SLIC to increase performance from each CPU on the array.

Data LUNs for the cluster configuration were configured as cluster shared volumes (CSVs). CSVs are a feature of Windows 2008 R2 failover clustering and allow a single namespace for LUNs connected to the cluster. All LUNs are online and directly accessible by each cluster node. One node acts as a coordinator for locking, and locking is limited to individual files on create and open functions. This means that virtual machines are not tied to the LUN on which they reside, and there is no failover requirement because the LUN is always online and available.

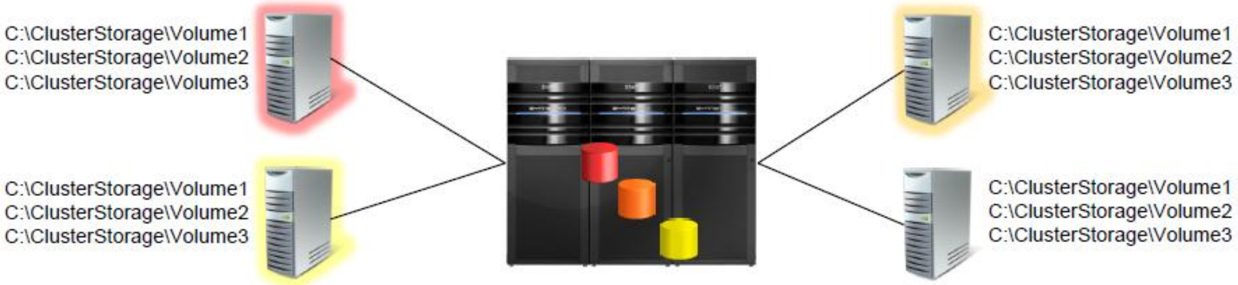

C:\ClusterStorage\Volume1
C:\ClusterStorage\Volume2
C:\ClusterStorage\Volume3
C:\ClusterStorage\Volume1
C:\ClusterStorage\Volume2
C:\ClusterStorage\Volume3
C:\ClusterStorage\Volume1
C:\ClusterStorage\Volume2
C:\ClusterStorage\Volume3
C:\ClusterStorage\Volume1
C:\ClusterStorage\Volume2
C:\ClusterStorage\Volume3

Figure 6. Single namespace for clustered LUNs with CSVs

For Hyper-V configuration, it was decided to use Virtual Hard Disks (VHDs) for virtual machine storage. Each virtual machine would exist on a VHD residing on one of the large data LUNs presented to the cluster. This would allow a virtual machine instance to be migrated to any node in the cluster using Microsoft's live migration feature. Figure 7 shows how multiple VHDs can reside on a physical volume on a parent node. In our configuration we presented 16 LUNs, with each 1 TB in size to the cluster. These large LUNs were used to deploy the 64 virtual machines per node, with each LUN housing 64 VHDs for a total of 1,024 virtual machines in the environment.

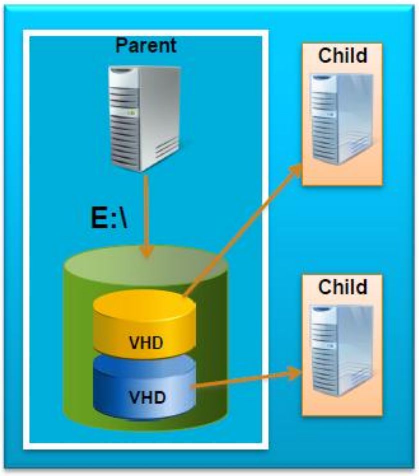


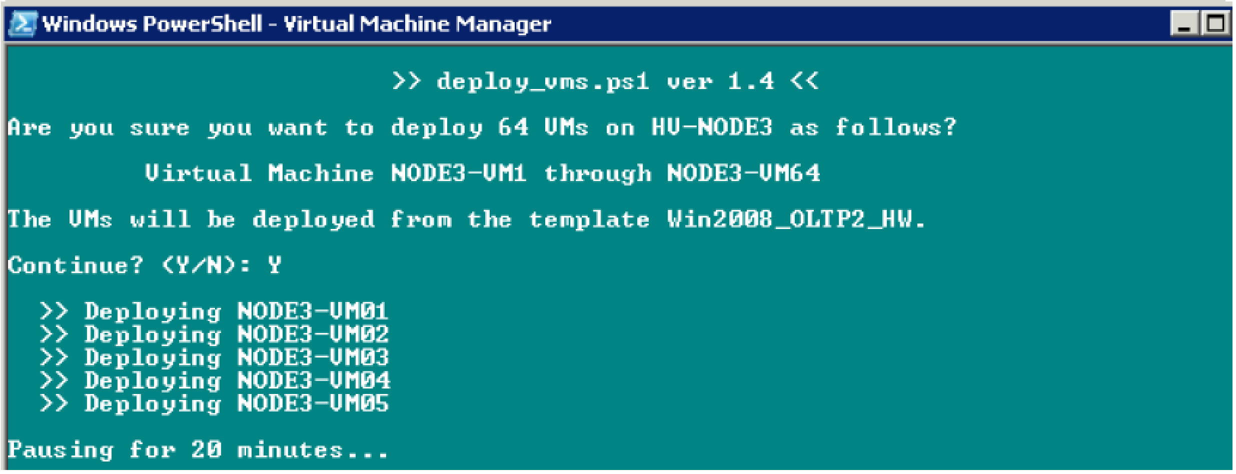
Figure 7. Multiple VHDs for virtual machines

Deploying the virtual machines

Once the cluster was in place, the question of how to quickly and accurately deploy such a large number of virtual machines was raised. A couple of methods were attempted to determine how well the criteria of speed and integrity could be accomplished to scale the configuration.

Microsoft System Center Virtual Machine Manager (SCVMM) was used as our initial deployment method. SCVMM provides centralized management of physical and virtual infrastructures, and centrally creates and manages virtual machines across entire datacenters. The software also consolidates multiple physical servers onto virtual hosts, and rapidly provisions and optimizes new and existing virtual machines, among other features. For EMC's purposes, we wanted to use SCVMM's virtual machine deployment capabilities combined with Windows PowerShell scripting to quickly create and deploy the large number of VMs we needed to create.

SCVMM employs a template model when deploying multiple VMs in an environment. Similar to creating the master image of our parent server node's boot images, we were required to create a running virtual machine that was configured how we wanted other virtual machines in the environment to behave. Once we had this "gold" virtual machine, an SCVMM template was created from it. Using SCVMM PowerShell modules, a deployment script was developed that took the gold image template and created multiple virtual machines from it and deployed them on a specified node (Figure 8).



```
                         >> deploy_vms.ps1 ver 1.4 <<

Are you sure you want to deploy 64 VMs on HV-NODE3 as follows?

         Virtual Machine NODE3-VM1 through NODE3-VM64

The VMs will be deployed from the template Win2008_OLTP2_HW.

Continue? (Y/N): Y

  >> Deploying NODE3-VM01
  >> Deploying NODE3-VM02
  >> Deploying NODE3-VM03
  >> Deploying NODE3-VM04
  >> Deploying NODE3-VM05

Pausing for 20 minutes...
```

Figure 8. Deploying VMs from an SCVMM template via PowerShell script

Because SCVMM uses the local area network to deploy VMs, it was discovered that scaling a Hyper-V configuration such as ours could take days due to the speed of the network and the nature of SCVMM's deployment method. It was decided that EMC TimeFinder® may provide a deployment solution superior to SCVMM's for a large-scale configuration.

TimeFinder provides local storage replication for increased availability and faster data recovery. Using the TimeFinder/Snap feature, up to 128 space-saving copies of live data can be produced within the Symmetrix array. These copies are immediately available for read/write access by host systems. In the case of our Hyper-V virtual machine scaling, a snap of a fully populated data LUN (with 64 VMs) was duplicated 15 times to provide the full complement of VMs to our configuration quickly.

Performance results

With the configuration scaled out and the virtual machines ready, we began starting up the virtual machines and beginning I/O generation from each one. The following figures show the VMAX performance as the virtual machines on groups of four parent nodes were started in parallel every 30 minutes.

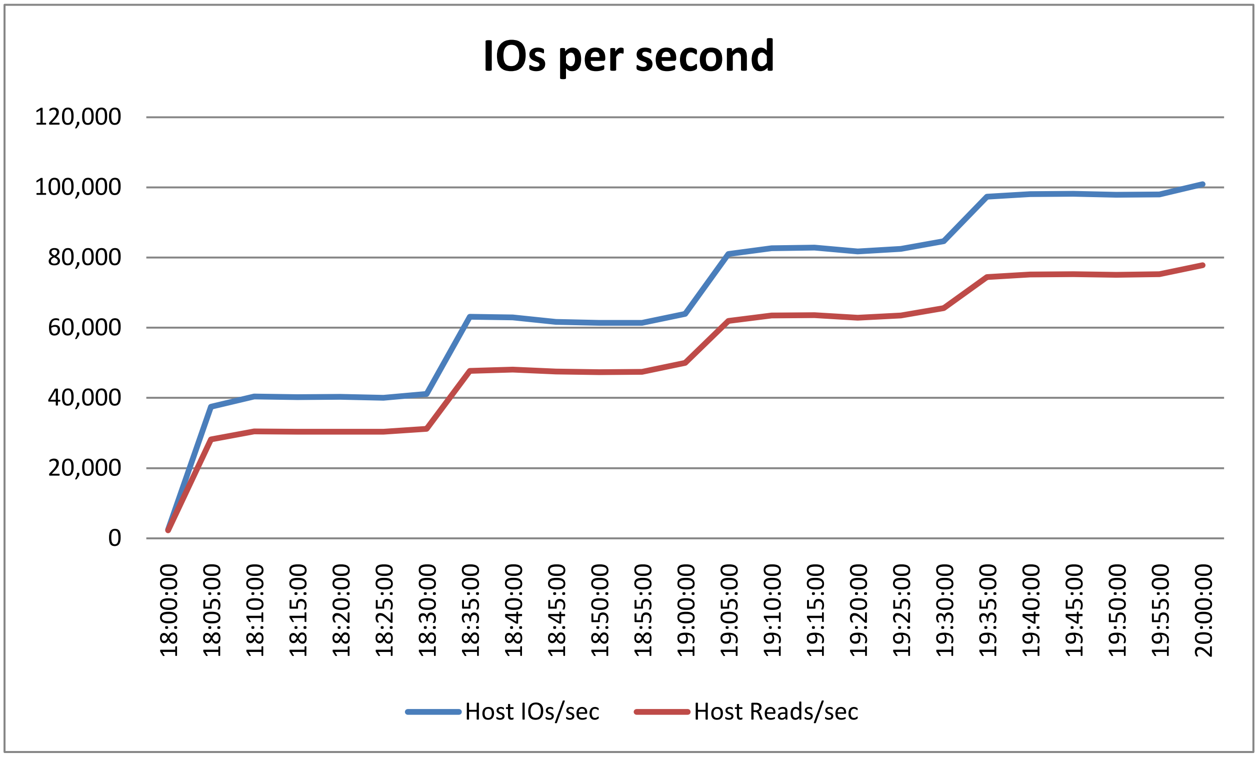


Figure 9. IOPS

Figure 9 shows that I/O increased near linearly as each set of virtual machines was started. It also shows that over 75 percent of the workload being generated was read I/Os.

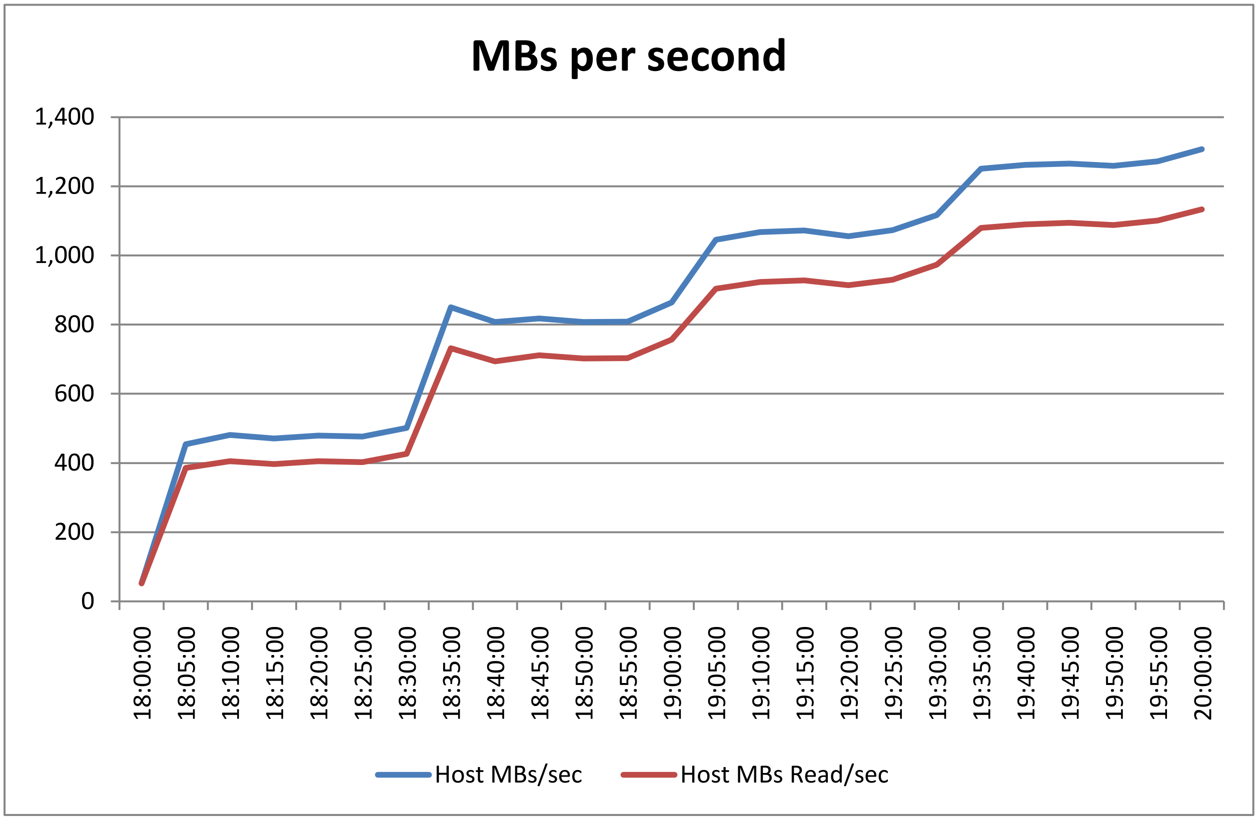


Figure 10. MB/s

Figure 10 shows that the virtual machines were generating over a gigabyte of data at the end of the test.

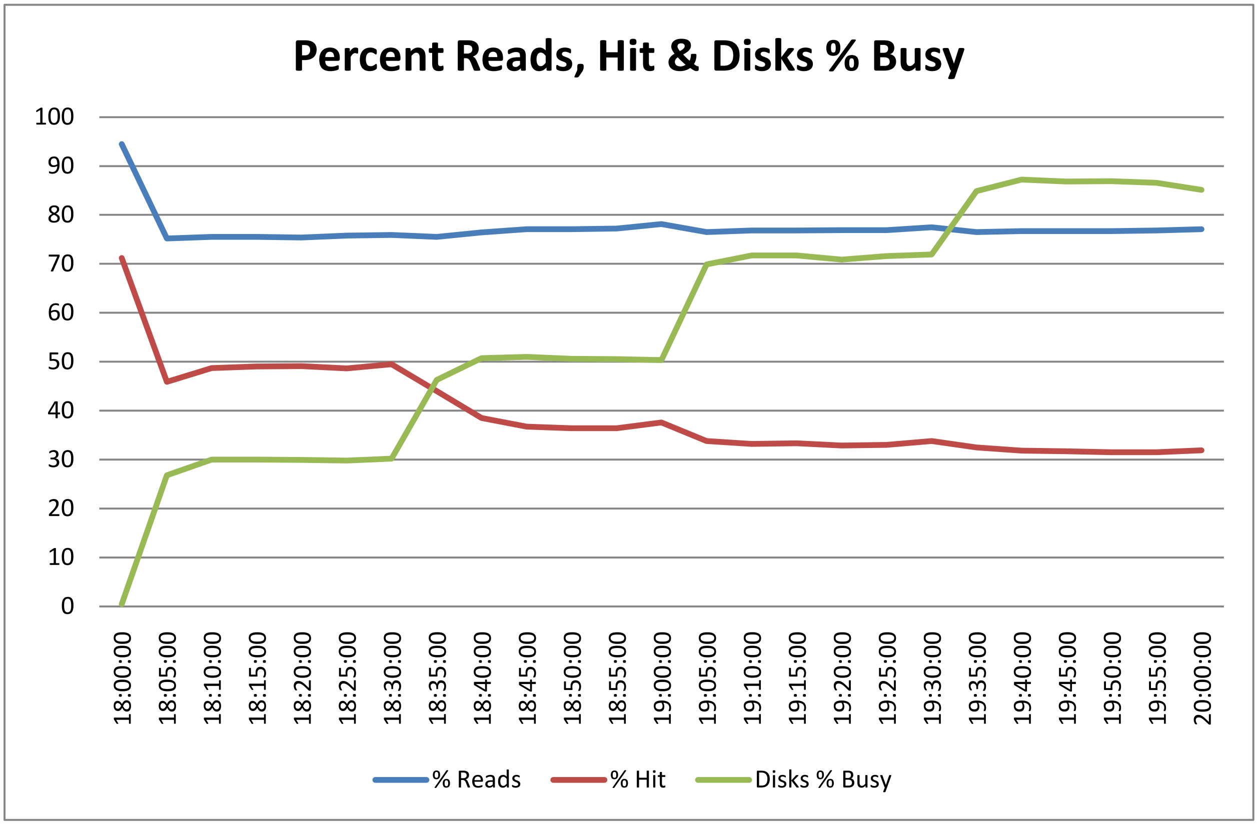


Figure 11. Percentage of reads, hits, and disks percentage busy

Figure 11 shows that as each group of virtual machines was started, the additional I/O workload reduced the cache hit ratio, resulting in more disk activity. By the end of the test, the disks were over 85 percent utilized.

These graphs show that as the I/O from the virtual machines increased, the array activity increased to accommodate the requests. IOPS, after just 5 minutes, reached nearly 40,000. By the end of the test, the disks reached over 85 percent busy. The VMAX, however, was able to sustain this performance.

Approaches to optimizing performance

Creating an optimal configuration of virtual machines is no easy task. It takes careful planning. With the right approach, however, a large-scale virtual environment is not only possible but can be achieved with low overhead and high performance. Keeping the following ideas in mind, and using tested and proven hardware and software products, similar configurations can be easily deployed.

1. Utilize multiple HBAs.

   From the host server perspective, adding more available I/O paths prevents saturation of a path and target port on the storage array. I/O queues on the initiators and targets are less likely to fill, causing a bottleneck.

2. Utilize multiple front-end controllers.

Having more target ports available on the storage array also prevents hitting any queue-full situations that would cause a bottleneck of I/Os from the host server.

3. Use PowerPath to optimize load balancing across the front-end ports.

   PowerPath's path optimization algorithm goes beyond MPIO's traditional round-robin I/O distribution. These advanced algorithms make path saturation less likely to occur.

4. "Quick" format NTFS volumes for thin space saving.

   Use up disk space on an as-needed basis, rather than pre-allocate space that may not be used right away.

5. Avoid Dynamic VHDs for heavy workloads.

   A fixed VHD always performs better than a Dynamic VHD in most scenarios by roughly 10 percent to 15 percent with the exception of 4k writes, where fixed VHDs perform significantly better.

6. Use appropriate drive counts when using Virtual Provisioning™.

   Virtual Provisioning will spread I/O across the available resources in an array. However, I/O must still be written to, and read from, physical disks. Therefore, the data devices in a thin pool must be spread over enough physical disks to keep up with the aggregate I/O requirements for all of the thin devices bound to the pool.

7. Watch the accumulated I/O load for CSVs.

   CSVs are a shared resource, so I/O can come from any number of parent nodes in a cluster in parallel.

8. Size snap pools with sufficient drives for change rate and workload.

   The save devices in a snap pool must be spread across a sufficient number of drives to support the copy on first write activity coming from the source devices, as well as the reads and writes directed at the virtual devices themselves.

9. Metadevices with larger numbers of smaller hypers are better than metas with a lower number of large hypers.

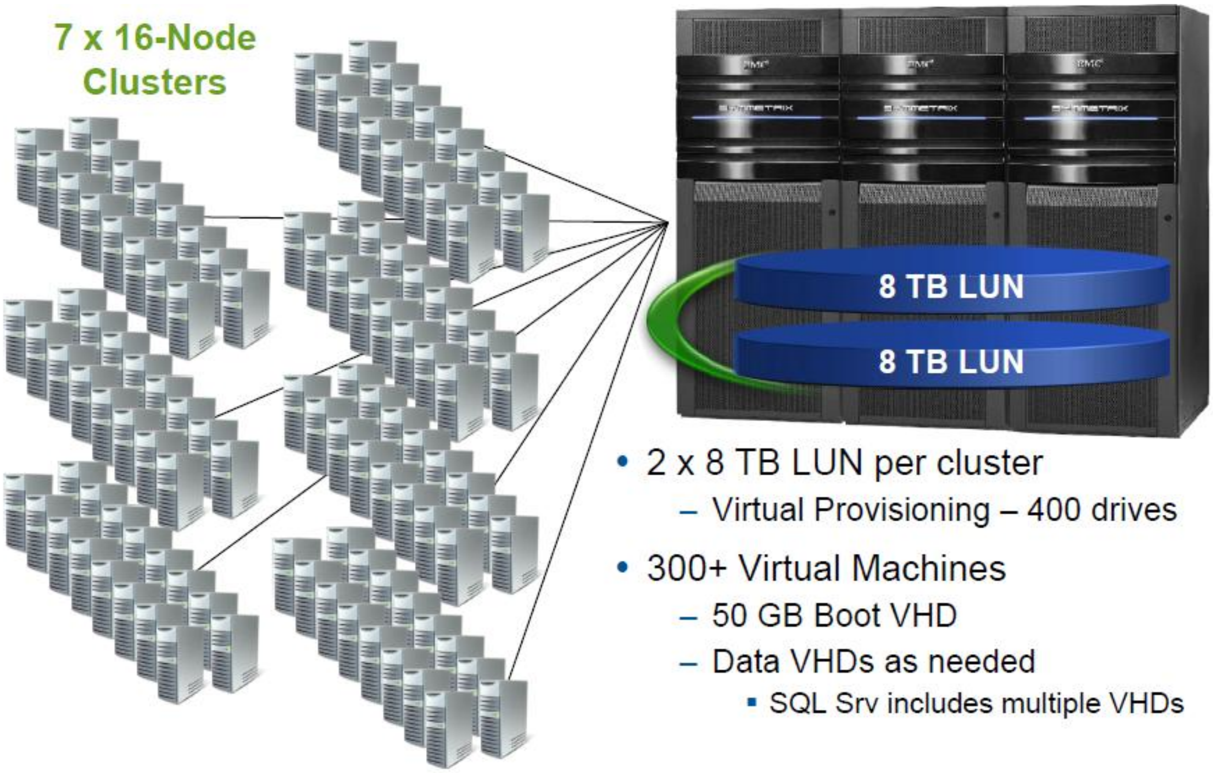
   - A device can have, at most, eight read misses per FA slice queued at a time. By using metadevices with a large number of metamembers, and addressing them down more front-end ports, a device can queue more I/O.

   - For write-intensive workloads, as the metamember count increases, the device has access to more cache and higher write pending limits.

   - Stripe for larger write workloads; concatenate for online growth capability.

Going larger

We've shown how a 16-node cluster running 1,024 VMs performs against Symmetrix VMAX. Because VMAX is a scalable storage architecture, it is possible to support multiple large-scale clusters on a single storage array. With proper provisioning and planning, customers could potentially attach multiple 16-node clusters, with hundreds more virtual machines, to a single Symmetrix VMAX array. Figure 12 shows one possible example.



- 2 x 8 TB LUN per cluster
  - Virtual Provisioning – 400 drives
- 300+ Virtual Machines
  - 50 GB Boot VHD
  - Data VHDs as needed
    - SQL Srv includes multiple VHDs

Figure 12. Configuration of seven 16-node clusters

Conclusion

In this paper, we have discussed methods of scaling a large Microsoft Hyper-V cluster configuration with Symmetrix VMAX. We showed how EMC engineers created a 16-node Microsoft Failover Cluster running Hyper-V and scaled it to support 64 virtual machines per cluster node. We showed how to employ different methods for quickly duplicating and deploying Hyper-V virtual machines using Microsoft and EMC technologies. Finally, we discussed I/O performance on the array and how it correlates to the Hyper-V environment as each node is brought online and I/O is executed from each virtual machine.

As customers move toward implementing the private cloud in their environments, configurations like the one described may be more common. It is hoped that through the combined efforts of EMC and Microsoft, that customers will be able to easily move to a virtualization model that will complement their business by using tested and proven tools and methods designed by each company’s engineering teams. Providing customers with easy-to-use, reliable, and efficient methods to deploy their virtual environments and protect their data is our main focus.