How Microsoft.com Moved to a Virtualized Infrastructure

Technical White Paper

Published: May 2010

Contents

[Executive Summary 3](#_Toc262792465)

[Situation 4](#_Toc262792466)

[Business Drivers 4](#_Toc262792467)

[Start of the Solution 5](#_Toc262792468)

[Dynamic Compute Infrastructure 6](#_Toc262792469)

[Logical Overview 8](#_Toc262792470)

[Operational Considerations 11](#_Toc262792471)

[System Health 15](#_Toc262792472)

[Results 18](#_Toc262792473)

[System Stability 18](#_Toc262792474)

[Virtual Machine Deployment Agility 18](#_Toc262792475)

[Resource Optimization and Power Usage 19](#_Toc262792476)

[Virtual Machine Migration 20](#_Toc262792477)

[Plan for High Availability 20](#_Toc262792478)

[Optimize Use of Capacity 21](#_Toc262792479)

[Consider and Implement Automation 23](#_Toc262792480)

[Find New Team Efficiencies 23](#_Toc262792481)

Executive Summary

Situation

Microsoft.com Operations, a team within the Microsoft IT division, needed to put existing hardware to better use and find ways to control capital and operational expenses.

Solution

Virtualization advancements in the latest releases of Microsoft technologies, combined with SAN-based clustered storage, solved numerous ongoing challenges. Implementing a dynamic compute infrastructure enabled Microsoft.com Operations to increase architectural flexibility, reduce power usage, and optimize use of resources.

Benefits

* Increased architectural flexibility
* Optimization of resource utilization
* Significant power savings
* New operational efficiencies

Products & Technologies

* Windows Server 2008 R2
* Windows clustering
* Cluster Shared Volumes
* SAN-based storage
* Hyper-V
* Live migration
* Quick migration
* Microsoft System Center Virtual Machine Manager 2008 R2
* Maintenance mode
* Quick storage migration
* Virtual machine templates
* Windows PowerShell cmdlets

Microsoft.com Operations (MSCOM Ops), a team within the Microsoft Information Technology (Microsoft IT) division that manages the Microsoft.com Web site, faced the same budgetary pressures that affect many businesses today: It needed to constrain capital spending, optimize the use of existing hardware, and reduce operational costs. Using the traditional IT model of “Decommission the old and buy newer, more-powerful servers” continued to result in a lot of unused storage, network, and compute capacity. The number of physical servers provisioned to support business needs was growing at a rate of approximately 20 percent per year during a time when budget continued to shrink. In spite of earlier and significant investments in physical server hardware, Microsoft.com was using less than 10 percent of the available processing power and 30 percent of the storage space.

MSCOM Ops saw a great opportunity to address these challenges while fulfilling the Microsoft IT mission of being an early adopter of the Windows Server® 2008 R2 operating system, Hyper-V™, and Microsoft® System Center Virtual Machine Manager 2008 R2. This paper introduces the logical design of the dynamic compute infrastructure (DCI), a virtualized hosting environment that combines these new technologies with clustering and storage area network (SAN)–based storage to provide a flexible, cost-effective hosting environment.

This new virtualized hosting infrastructure enables the team to:

* Provide capacity on demand for customers. Allocating and deploying virtual machines takes only days instead of weeks.
* Optimize the use of resources. CPU utilization has more than doubled on average, and storage utilization has reached 50 percent on the longest-running DCI cluster.
* Actualize significant power savings. MSCOM has seen an overall 67 percent power savings with virtual machines on two-socket host clusters when compared to the power that equivalent physical servers use.
* Develop new operational efficiencies. No clusters have gone offline since MSCOM Ops deployed the first DCI instance, in spite of monthly maintenance that included hardware and software upgrades.

This paper also describes some of the automation that the team has developed to help manage Microsoft.com, and it includes operational considerations and best practices aimed at IT pros who may be considering a similar virtualization strategy. However, this paper is based on the experience and recommendations of MSCOM Ops as an early adopter. It is not intended to serve as a procedural guide. Each enterprise environment has unique circumstances; therefore, each organization should adapt the plans and lessons learned described in this paper to meet its specific needs.

Note: For security reasons, the sample names of internal resources and organizations used in this paper do not represent real resource names used within Microsoft and are for illustration purposes only.

Situation

Reliance on physical servers to provide processing power and storage capacity for hosted applications presented numerous challenges for the MSCOM Ops team, including:

* A lack of architectural flexibility. It could take weeks or months to adapt the environment to changing business needs, such as:
* Accommodating temporary spikes in traffic.
* Providing incubation capacity for new or updated properties.
* More efficiently validating pre-release and release candidates of Microsoft technologies in a production environment.
* Higher-than-necessary power costs to keep underutilized servers running.
* Suboptimal use of:
* Data-center space to house the servers.
* CPU capacity—CPU utilization remained under 10 percent of the total amount available.
* Disk storage space—approximately 30 percent of the available disk space was used.

Managing the traditional operational tasks of acquiring, maintaining, and decommissioning hardware for keeping the Microsoft.com infrastructure running remained complex and expensive in terms of capital and operational expenses. Controlling costs has always been a focus of the team, but business and economic conditions necessitated cost-control measures that were even more aggressive.

MSCOM Ops clearly needed to find ways to contain capital and operating expenses while continuing to deliver high performance and highly available solutions. The team needed a solution that would offer architectural flexibility as well. The classic IT model for requisitioning, deploying, and supporting physical servers was slow and cumbersome in comparison to the capabilities of new server tools and technologies.

Business Drivers

Providing better value from capital and operational expenses emerged as the key business driver for creating a virtualized architecture for Microsoft.com properties. The MSCOM Ops team sought to identify ways to maximize architectural flexibility while still providing a resilient and highly available infrastructure for service delivery. The expected results from designing and building a virtualized infrastructure included:

* Application teams would have access to incubation capacity that could be increased or decreased as needed during development cycles.
* Deployment time for servers and applications would shift from weeks or months to hours or days.
* The self-healing capabilities gained from combining these technologies into a virtualized architecture would result in faster dynamic recoveries from failures, increasing the number of failure scenarios handled without human interaction.
* MSCOM Ops would be able to adapt the production environment more quickly to maintain availability of applications during short-term increases in load or unexpected load-driving events on systems.
* Gains in architectural flexibility would make it easier to deploy servers for evaluation of pre-release and release candidate versions of Microsoft server products.
* Operations engineers would be able to optimize capacity while continuing to deliver infrastructure that facilitates 99.9 percent or better availability for the hundreds of individual applications hosted in the MSCOM Ops environment.

Note: For more information about how the team maintains, monitors, and measures availability for Microsoft.com, see “Maintaining High Availability for Microsoft.com” at <http://technet.microsoft.com/en-us/library/ff467943.aspx>.

* New tools would enable the team to make significant improvements in automation and standardization during the transition.

In short, MSCOM Ops expected many improvements in business agility while containing costs through use of virtualization and the associated tools for managing the operating environment.

Start of the Solution

Advances in Microsoft server and management technologies pointed to virtualization as a potential solution to various challenges. The team's goal was to take the stability successes previously achieved with Hyper-V into a scalable SAN-based environment.

Note: To read about the team's early findings on using Hyper-V, download the paper "Microsoft.com Operations Virtualizes MSDN and TechNet on Hyper-V" at <http://download.microsoft.com/download/6/C/5/6C559B56-8556-4097-8C81-2D4E762CD48E/MSCOM_Virtualizes_MSDN_TechNet_on_Hyper-V.docx>.

Members of MSCOM Ops began developing new plans for a DCI after they learned more about the following features available in Hyper-V in Windows Server 2008 R2 and in System Center Virtual Machine Manager R2:

* Hyper-V in Windows Server 2008 R2:
* Live migration for moving virtual machines with no downtime
* Hundreds of virtual machines per logical unit number (LUN) through Cluster Shared Volumes (CSVs)
* Performance improvements in dynamic virtual hard disks (VHDs)
* System Center Virtual Machine Manager 2008 R2:
* Single-console live migration support
* Template-based rapid provisioning
* Virtual network configuration for clusters
* Maintenance mode to automate the evacuation of virtual machines off host machines

The team designed a resilient, self-healing, and flexible architecture based on these technologies. In June 2009, the team deployed the first DCI instance in one of the data centers used to deliver Microsoft.com services.

Dynamic Compute Infrastructure

The DCI design takes advantage of improvements in Hyper-V, especially live migration, giving MSCOM Ops the ability to move virtual machines from one host to another while the virtual machines are still running. This flexibility stems from running clustered hosts attached to CSVs in a SAN. Having this capability enables the team to optimize capacity as needed and to make the process of performing routine maintenance more efficient.

Table 1 provides an overview of the key Microsoft technologies used to build and manage the MSCOM Ops DCI.

Table 1. Technologies of the MSCOM Ops DCI

|  |  |  |
| --- | --- | --- |
| Products and features | Details | Benefits |
| Windows Server 2008 R2  Cluster service | MSCOM Ops considers a cluster the lowest common component for redundancy for an application hosted in the DCI. | Enables failover from one host within a cluster to another host in order to maintain high availability. |
| Hyper-V  Live migration | Hosts and clusters can be managed without affecting the operations of the hosted virtual machines. | Streamlines management of infrastructure and enables operators to rebalance use of capacity by moving virtual machines as needed. |
| Quick migration | In conjunction with clustering, this feature enables moving multiple virtual machines simultaneously by copying the memory to a disk in storage. | Enables operational efficiencies in cases where using live migration is not possible. |
| CSVs | CSVs are designed and used to store virtual machines and their associated VHDs, not for file or content sharing. They give every host access to every dynamic VHD in the cluster. | Provides easier storage management, greater resiliency against some failures, and the ability to store many virtual machines on a single LUN and have them fail over individually.  Provides the infrastructure to improve support for live migration of virtual machines running Hyper-V. |
| Dynamic VHDs | Using dynamic VHDs allows better use of storage size because it provisions only the space needed while still allowing for growth up to a pre-declared upper limit. | Provides flexibility for greater agility in managing storage capacity. |
| Dynamic Host Configuration Protocol (DHCP) | DHCP facilitates automation of management processes for IP and media access control (MAC) addresses. | Eliminates complexities in assigning and re-using IP and MAC addresses.  Enables the same tooling to be applied to both physical and virtual servers.  DHCP server acts as the repository of, and provides the tracking mechanism for, IP mappings. |

Note: For more information about Hyper-V and the benefits of using clustering and CSVs to enable live migration, visit the Virtualization with Hyper-V: Overview page at <http://www.microsoft.com/windowsserver2008/en/us/hyperv-overview.aspx>.

System Center Virtual Machine Manager enables operators to manage the rapid provisioning of new virtual machines and to provide unified management of physical and virtual infrastructure through one console. MSCOM Ops deploys one instance of System Center Virtual Machine Manager per data center. Table 2 provides an overview of how the team uses System Center Virtual Machine Manager to optimize operations management for the DCI.

Table 2. Use of Virtual Machine Manager to Optimize Operations Management

|  |  |  |
| --- | --- | --- |
| Feature | Description | Benefit |
| Maintenance mode | Use to designate a server to not be available for virtual machine deployments. | Maintenance mode enables operators to perform live migrations in a prescribed fashion rather than allowing virtual machines to be sent to more than one host within the cluster.  Also, if a node fails, this feature enables faster recovery time because the virtual machines from the failed node quickly migrate to the maintenance-mode node. |
| Quick Storage migration | Use to specify storage locations for each virtual hard disk (.vhd) file. | Migration of a running virtual machine’s files to a different storage location on the same host entails minimal or no service outage. |
| Self-service portal | Use to remotely start, stop, pause, or restart a virtual machine, even if remote desktop services are not available. | The self-service portal makes it easier for owners of virtual machines to interact with virtual machines without needing to be physically in front of the host. |
| Templates | Use to enable rapid provisioning for virtual machines by enabling the team to create and maintain a standard set of server configurations. | Using templates provides better control of ongoing capacity forecasting and optimization efforts. It also makes it easier to track remaining not-yet-provisioned capacity in the DCI and simplifies creation of virtual machines. |
| Metadata | Add custom metadata or edit default metadata about virtual machines to enable more detailed reporting. | By using custom data, MSCOM Ops can track who requested the virtual machines, on which ticket the request came in, and to which department the requestor belongs. The team collects custom details in System Center Virtual Machine Manager reports. |
| Windows PowerShell™ cmdlets | Use to facilitate automation and reporting functions. | Using Windows PowerShell cmdlets enables MSCOM Ops to automate routine administrative tasks, including report generation.  Custom reporting provides information about how full a cluster is, gives system health information that aids the process of determining when to migrate virtual machines to another cluster, and shows which hosts have the needed capacity when such a change is needed. |

Note: For more information about System Center Virtual Machine Manager, visit the Virtual Machine Manager page at <http://www.microsoft.com/systemcenter/en/us/virtual-machine-manager.aspx>.

This section provides an overview of the underlying physical components and network elements and includes a description of the logical design of the DCI.

Note: Some of the specific network and hardware details given in this paper reflect the realities of building this infrastructure within the larger Microsoft corporate network and may not be worth pursuing in other organizations. As with designing any architecture, an organization should evaluate the business requirements and any existing technical constraints to determine which design elements to apply.

Logical Overview

The DCI design uses host clusters as the element from which to build redundancy for hosted applications. To take the best advantage of live migration through use of CSVs, MSCOM Ops placed a high priority on the ability to balance storage performance. This led to a decision to use one or more CSVs in order to provide the opportunity for redundancy and load balancing.

Basing the host clusters on Windows Server 2008 R2 enables up to 16 nodes that use CSVs for storage. As depicted in Figure 1, the logical design of the DCI consists of:

* 16-node host clusters with one node in maintenance mode.
* Clustered hosts attached to a SAN with CSVs from 6 terabytes to 12 terabytes, depending on the number of nodes in the cluster and the storage requirements of the applications being hosted.

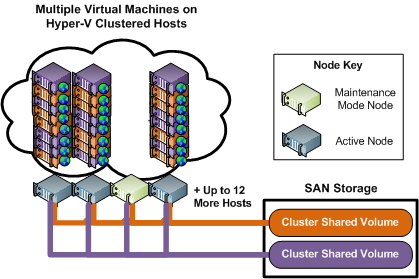


Figure 1. Logical design overview of the DCI

Note: For more information about using CSVs and Hyper-V to enable failover clustering, see “Hyper-V: Using Hyper-V and Failover Clustering” at [htt://technet.microsoft.com/en-us/library/cc732181(WS.10).aspx](http://technet.microsoft.com/en-us/library/cc732181(WS.10).aspx).

## Cluster Communications

In the MSCOM Ops solution, host servers utilize multiple networks. Each network has a specific purpose:

* The Cluster network is the primary (active) network for supporting cluster communication and live migrations within the DCI.
* The Administration network is used for remote management of the host, of virtual machines, and for other administrative functions. It is also the secondary (failover) network for cluster communications and live migrations.
* The virtual machines that reside on Web server host clusters use two production networks, Production 1 and Production 2, to deliver content and functionality to Microsoft.com users. These networks are not exposed for the hosts to use.

Figure 2 illustrates the connections and network communication flow for the DCI.

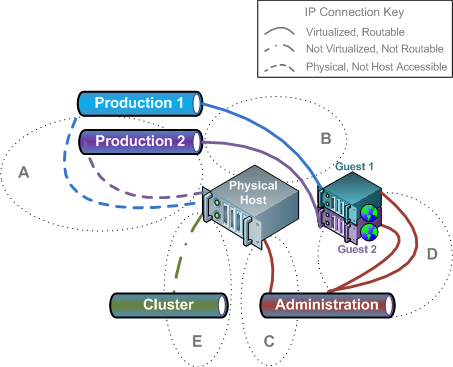


Figure 2. DCI network connections

Servers communicate over these networks through the network interface cards (NICs), as described in Table 3. The letters included Figure 2 align with rows in the table according to the “Figure reference” column.

Table 3. DCI Network Details

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Figure reference | Network | NIC | Routable? | Public or private? | Host accessible? |
| A | Production 1 and Production 2 | Physical on host | Yes | Public | No |
| B | Production 1 and Production 2 | Virtualized on guest | Yes | Public | Not applicable |
| C | Administration | Physical on host | Yes | Private | Yes |
| D | Administration | Virtualized on guest | Yes | Private | Not applicable |
| E | Cluster | Physical on host | No | Private | Yes |

## Cluster Configuration

MSCOM Ops used many defaults when setting up the clusters for the DCI. However, the team encountered some issues, including unplanned virtual machine migrations and CSV redirection, when initially evaluating the performance of its DCI design. Because of these issues, the team identified a number of design-specific cluster configuration settings. This section introduces those settings and describes the specific changes that the team made to help ensure that the clusters communicate across the networks as intended for the DCI.

To help ensure that communications on networks work optimally for the design, the team uses the following settings:

* Administration Network. Configured to allow cluster communications and clients to connect through this network.
* Cluster Network. Configured to allow cluster communications. Clients are not allowed to connect through this network.

MSCOM Ops configures clusters to use a disk witness by selecting Node and Disk Majority as the quorum setting for determining how many nodes can fail before the cluster goes offline.

Note: For more information about quorum options, see “Understanding Quorum Configurations in a Failover Cluster” at <http://technet.microsoft.com/en-us/library/cc731739.aspx>.

The network preference for live migration can be changed only after a virtual machine has been deployed to a node. This is a global cluster setting available in Failover Cluster Manager. After it is changed, it affects all nodes in the cluster. For the steps on how to change the preference for networks for live migration, see “Configure Cluster Networks for Live Migration” in “Hyper-V: Using Live Migration with Cluster Shared Volumes in Windows Server 2008 R2” at <http://technet.microsoft.com/en-us/library/dd446679(WS.10).aspx>.

Operational Considerations

Moving to a DCI presents unique opportunities to make infrastructure management easier through use of the Microsoft System Center tools and strategic use of automation. Many routine maintenance activities for the DCI hosts occur without affecting service delivery. MSCOM Ops now focuses more of its efforts on managing infrastructure as a service. This section introduces how the team determines when to virtualize a particular application workload, what automation the team has implemented, and what tools it uses to manage the DCI.

## Virtual Machine Deployment

This section provides background about how MSCOM Ops determines whether to host an application in the DCI, including a description of the evaluation process for existing and new properties. Whether the team is deploying virtual machines as part of a server consolidation effort or in the process of deploying a new application to the DCI, the team deploys them across clusters by design rather than deploying on only one cluster.

Determining how to manage IP addresses for a virtualized environment is crucial to the success of building and operating such an infrastructure. MSCOM Ops uses DHCP combined with automation to make managing IP and MAC addressing for virtual machines more efficient.

Server Consolidation Analysis

Deciding whether an application or service is a candidate for moving to the dynamic infrastructure is the first step toward virtualization in the MSCOM Ops environment. Not every server workload is a candidate for virtualization because of its processing, memory, or storage needs.

Each attempt to consolidate servers through virtualization begins with analyzing utilization of existing physical servers. When performing this analysis, MSCOM Ops system engineers review a number of factors against certain guidelines for capacity optimization. One constraint involves the number of processors available in a guest virtual machine. Though some guests running in the DCI are on older versions of Windows Server, this discussion focuses on virtual machines running Windows Server 2008 R2.

Note: For more information about operating systems that can run on guest virtual machines, see “Virtualization with Hyper-V: Supported Guest Operating Systems” at <http://www.microsoft.com/windowsserver2008/en/us/hyperv-supported-guest-os.aspx>.

To determine whether a physical server is a good candidate for virtualization, MSCOM Ops tracks certain utilization factors for a minimum of 30 days:

* CPU utilization
* Memory utilization
* Storage utilization
* Current server input/output (I/O) workload (optional)

For more guidance on conducting a server consolidation analysis, download the “Infrastructure Planning and Design: Windows Server Virtualization” guide at <http://go.microsoft.com/fwlink/?LinkID=147617>.

New Server Deployment

When MSCOM Ops considers whether to provision the servers to host a new application to the DCI, it considers the same factors mentioned previously in “Server Consolidation Analysis.” The difference with this scenario is that there is no performance history to review before the team makes a determination. The team reviews the technical specifications of the application and any available test reports, including details on the anticipated storage needs and expectations of how I/O intensive the application may be, to make an initial determination.

If the decision is to provision virtual machines in the DCI for hosting the application, the next phase is to establish a performance baseline. Conducting a baseline performance analysis helps the team ensure that the virtual machines and the hosts have sufficient memory for the anticipated production loads. This early performance assessment enables the team to make changes based on reviewing whether the initially provisioned capacity was:

* Sufficient for production loads. No changes to the virtual machines are necessary.
* More than what the application typically needs. In this case, the team can quickly reclaim unused or underutilized virtual machine resources.
* Not enough capacity. Depending on how much more capacity may be needed, the team may opt to add to the virtual machine’s resources or deploy more virtual machines on which to host the application. Or, the team may decide that the application is no longer a good candidate for operating in the DCI.

Automation for IP and MAC Address Management

Managing and tracking IP address assignments and utilization in a large-scale hosting environment is an operational necessity. Use of a DHCP server can remove much of the complexity of this routine task. For MSCOM Ops, introducing the use of dynamic IP or dynamic MAC addresses presented potential problems because of the use of hardware load balancers to facilitate high availability of the Microsoft.com properties. The team recognized that having greater flexibility in provisioning, de-provisioning, and rebuilding virtual machines could result in increased operational costs if it did not proactively plan for managing how IP addresses are assigned.

MSCOM Ops evaluated the available scenarios, taking into consideration the uptime requirements for Microsoft.com Web properties and using a DHCP server for tracking IP address assignments. The team decided to use static IP and static MAC addresses for hosts and virtual machines in the DCI. Choosing to identify NICs in this way avoided the following potential problems:

* Dynamic IP addresses:
* Provisioning and rebuilding machines introduces the risk of some instability because each time this type of change happens, a new IP address is created and applied. Waiting for propagation of such changes throughout the environment results in increases to operational expenses because of the need for manual intervention to resolve issues.
* If the DHCP server becomes unavailable and the lease time for an IP expires, the client becomes unavailable.
* Dynamic MAC addresses:
* There is a risk of encountering problems in this scenario when using live migration to move virtual machines from one host to another because a new IP is always retrieved when the MAC address changes. Even if an organization is using static IP addresses, the virtual machine retains its IP connection only until it needs to be rebuilt, at which point the DHCP reservation is no longer valid because the machine has been given a new MAC address. The effect in this case is that the machine will receive a new IP address, causing the same types of issues noted previously with rebuilding servers that use dynamic IP addresses.
* Users more often experience slow-loading sites as a result of waiting for dynamically assigned MAC addresses to propagate to servers and switches.
* Another issue inherent with waiting for dynamic MAC addresses to propagate is that packets are more likely to be dropped. Pages or parts of pages may then appear to be missing to site users.

Whether an operator chooses to use either dynamic MAC or dynamic IP addresses for virtual machines, such a scenario makes use of hardware load balancers more complicated. For example, using dynamic MAC and static IP addresses for virtual machines requires operators to monitor more systems for connection issues or dropped data packets. These issues typically are the result of delays in MAC address propagation throughout the system and are avoidable if static MAC addresses are used. Likewise, though the root issues are different, choosing to use static IP addresses simplified the process of using hardware load balancing for MSCOM Ops because it eliminated the need to update IP mappings on these devices every time a dynamic IP change occurs.

Another important reason to choose to use static IP and MAC addresses was that this choice enabled MSCOM Ops to design automation that brought additional operational efficiencies to deploying physical and virtual servers for the DCI. This static IP/static MAC addressing scenario, when combined with a new approach to using the DHCP service, enabled the team to use the same IP management automation while provisioning new servers and rebuilding existing servers, whether physical or virtual.

Note: To review the concepts of how DHCP works when running on Windows Server 2008, see “DHCP Architecture” at <http://technet.microsoft.com/en-us/library/dd183602(WS.10).aspx>.

The automation that MSCOM Ops developed takes advantage of the ability to define a scope and to make IP reservations on the DHCP server. Most of the virtual machines in the MSCOM Ops DCI have two NICs, as depicted in Figure 2: one for communication with one of the Production local area networks (LANs) and one for communication with the Administration LAN. Each LAN has a scope defined within the DHCP server that the team uses to identify and reserve an IP address for a particular NIC. How this happens varies depending on whether the server is physical or virtual. Table 4 provides an overview of how the automation assigns IP and MAC addresses by server and deployment types.

Table 4. Overview of IP and MAC Address Automation

|  |  |  |
| --- | --- | --- |
| Server type | Deployment type | Automation notes |
| Physical | New | DHCP dynamically assigns an IP address from those available for the particular scope. The assigned address is stored with the MAC address as a DHCP reservation for future use. The tool changes the configuration on the server to static after the address has been reserved. |
| Physical | Rebuild | Via DHCP, the physical server receives the same IP address that it was using before the rebuild, based on looking up the MAC address. The tool changes the configuration on the server to static. |
| Virtual | New | Before the virtual machine is created, the tool selects an available IP address and generates a MAC address based on the selected IP address. Both are stored as a DHCP reservation and later applied to the virtual machine. |
| Virtual | Rebuild | Via the DHCP reservation, the server receives the same IP address for the Administration network that it was using before the rebuild based on looking up the MAC address. The address information from the DHCP reservation is then applied again to the virtual machine. |

Process Description for IP Automation While Provisioning a New Virtual Machine

Using a new virtual machine as an example, the automation logic for the IP provisioning process is as follows:

1. The operator calls automation and specifies which networks are required.
2. The tool checks the specific network scopes to find available IP addresses. The tool selects the next available IP address for each network specified in step 1.
3. A MAC address is generated based on combining the assigned IP address with unassigned manufacturer MAC identifiers to create an address that is unique in the MSCOM Ops operating environment.
4. The IP and MAC address pairs are reserved in DHCP.
5. The operator creates a virtual machine with the generated MAC addresses. When the virtual machine goes online, it automatically receives the reserved IP address for the Administration NIC from DHCP.
6. Automation determines to which network each NIC is connected. The tool then renames the NIC to specify the network on which it resides.
7. For the Administration NIC, the tool collects the IP details from the virtual machine and applies the settings, including making them static.
8. For the Production NIC, the automation pulls the IP details from the DHCP server and applies the settings, including making them static.
9. The automation applies network routes.

Relying on a DHCP server also entails risks when an organization needs to plan for business continuity and high availability. Because DHCP may be a point of failure in this automation, MSCOM Ops has taken some additional precautions:

* It extended IP lease times so that even if the DHCP server fails, the team has up to two weeks to restore data from backups and to recover service.
* DHCP uses a default destination for its automatic backups. If a manual backup is stored in that location, it will be overwritten the next time an automated backup occurs because only one backup is retained by default. MSCOM Ops uses a scheduled process to move and rename the backup of the DHCP database once every few hours. This renamed backup database is also copied to a secondary storage location.
* MSCOM Ops monitors the number of available IP addresses on a per-network basis. This enables the team to request additional IP addresses if needed before a problem can arise.

The DCI combined with the IP automation and novel use of DHCP described in this section enables IP affinity when a virtual machine is rebuilt.

System Health

A key operational benefit of having a DCI is the ability to rapidly provision and configure hosted systems. If the infrastructure approaches maximum capacity, this benefit may be substantially eroded. To avoid this, the MSCOM Ops team monitors a number of performance counters. This section describes the system health issues that the team currently considers to be most important. This section also includes details about a few known issues that may assist in planning and deploying a virtualized infrastructure for IT pros who are considering moving to a similar design.

Host Memory Utilization

In planning the infrastructure and provisioning virtual machines on the available hosts, MSCOM Ops erred on the side of caution and set aside more than the minimum memory for the parent partition suggested by the Hyper-V system requirements at <http://www.microsoft.com/hyper-v-server/en/us/system-requirements.aspx>. The team started with the guidance in “Correct Memory Sizing for Root Partition” in “Performance Tuning Guidelines for Windows Server 2008” at <http://go.microsoft.com/fwlink/?LinkId=121171>. Early evaluations suggested that the demands of the Microsoft.com properties after virtualization might be even greater than indicated from simply applying the Windows Server tuning formula, and the team made adjustments accordingly.

As the cluster’s memory resources fill, MSCOM Ops pays more attention to placing virtual machines in a way that best utilizes the remaining available memory. MSCOM Ops reviews such reports on a weekly or monthly basis.

CPU Oversubscription Ratio

When determining the oversubscription ratio for a particular host, MSCOM Ops considers the same processor utilization factors that it uses to determine whether to virtualize a server:

* Peak usage times
* Peak usage loads
* When sustained loads are expected and how much CPU is needed to avoid performance degradations

The hardware for the DCI, though standardized within each cluster, is not uniform across clusters. What is possible in terms of CPU oversubscription depends on the capabilities of the underlying components.

Table 5 introduces the objects and the associated performance counters that MSCOM Ops uses for tracking CPU overutilization, including descriptions of what the counters mean in relation to the DCI.

Table 5. CPU Overutilization Objects, Counters, and Further Investigation

|  |  |  |
| --- | --- | --- |
| Object name | Counter name | Details |
| Hyper-V Hypervisor Logical Processor | % Total Run Time | Provides the process utilization as a total, including all the virtual machines and the host.  This measure is useful for considering how to balance CPU utilization across the DCI as a whole.  If this measure reaches 60% or higher on a particular host, MSCOM Ops begins investigating to determine what is causing this level of utilization: Is it an issue with the application itself? Does the application need more processing power to handle current usage levels? Resolution may involve fixing an application issue or rebalancing processing power to provide more to the application. |
| Hyper-V Hypervisor Logical Processor | % Guest Run Time | Provides the total process utilization of the virtual machines.  If this value suggests an unexpected increase in processor utilization with the virtual machines, MSCOM Ops reviews additional counters. In particular, looking at % Processor Time from the operating system of the virtual machines can help the operator determine a solution. |
| Hyper-V Hypervisor Logical Processor | % Hypervisor Run Time | Provides the process utilization of the host.  MSCOM Ops uses this to consider how to balance utilization across the DCI as a whole. |

CSV Drive Space Utilization

The DCI uses dynamic VHDs in the MSCOM Ops CSVs, a feature that enables the team to overallocate SAN storage space. MSCOM Ops monitors space utilization on the SAN on a daily basis because of a major risk inherent in overallocating the physical storage that contains the dynamic VHDs. The risk is that if a dynamic VHD cannot expand to its fully allocated space because the underlying storage has already been filled, that virtual machine will pause unexpectedly.

Note: Unexpected pausing of a virtual machine is an issue that is associated with choosing to use dynamic VHDs. It will happen whether or not the dynamic VHD is on a CSV and is designed to safeguard the virtual machine from data corruption.

Monitoring of space utilization occurs on a per-CSV basis. MSCOM Ops sets initial monitoring thresholds based on expected business growth and reviews them regularly. Daily reports include calculations of subscribed SAN and used SAN space.

This close monitoring of the actual used SAN space means that the team can take advantage of another Hyper-V feature, the use of multiple CSVs. The ability to dynamically add CSVs or to dynamically grow them means that the team can add storage capacity before virtual machines exceed storage capacity. After the team adds the CSV, it uses quick storage migration to move storage for some virtual machines to that CSV.

## Host Maintenance

MSCOM Ops uses a new host configuration, maintenance mode, in combination with live migration while performing host maintenance. This configuration setting, available through System Center Virtual Machine Manager R2, prevents the creation or addition of virtual machines to the maintenance-mode node within the cluster. In other words, it allows a host server to be set aside for maintenance activities and therefore also becomes that cluster’s target for a passive quick migration if another node fails.

**Note:** Enabling a maintenance-mode node through System Center Virtual Machine Manager does not prevent the addition of virtual machines to the node through other administrative tools like the Failover Cluster Manager snap-in.

When host maintenance is needed, MSCOM makes updates on the maintenance-mode node first within each cluster. The team uses scripting to facilitate the maintenance process as follows:

1. Use live migration to move virtual machines from the active node to the maintenance-mode node and change its configuration to active.
2. Change configuration of what was the active mode to maintenance-mode node.
3. Perform host maintenance activity on the maintenance-mode node.
4. Validate the health of the maintenance-mode node.

This process repeats until the team has performed maintenance on all nodes in a particular cluster.

## Other Considerations

This section contains operational considerations that MSCOM Ops discovered in the course of troubleshooting issues after deploying the first instance of the DCI. These issues occur because of the way a feature was designed or require the application of a fix that is not yet available in a service pack. As such, future versions of Windows Server, Hyper-V, and related management tools may offer other ways to address them.

“Network Dead” Issue

A virtual NIC can be disabled unexpectedly while the virtual machine is taking live traffic, resulting in a “network dead” issue. The hotfix for this issue, “The Network Connection of a Running Hyper-V Virtual Machine Is Lost Under Heavy Outgoing Network Traffic on a Windows Server 2008 R2-Based Computer,” is available at <http://support.microsoft.com/kb/974909>.

Binary File Issue

MSCOM Ops discovered that the hypervisor maintains binary (.bin) files on the storage location for each virtual machine. The size of each binary file is equal to the amount of memory allocated to the guest machine. It is important to account for this space when planning storage for a virtualized infrastructure.

Results

Virtualization enables MSCOM Ops to mix high-scale and low-scale systems on the same physical server resources. By oversubscribing CPUs and overallocating storage space, the team improves overall hardware utilization while also maintaining a reduced physical footprint and requiring less power to run the environment. New features of Hyper-V and Windows Server 2008 R2 provide the architectural flexibility that helps the team optimize the utilization of available resources.

System Stability

MSCOM Ops currently runs multiple 16-node clusters within each of two data centers and is in the process of deploying a third. An individual DCI deployment does not have nodes in multiple data centers. This design is for geographically isolated instances of the solution. The first instance was operational as of June 2009, and the second came online in November 2009. The third instance is expected to be online in June or July 2010. Since the initial deployment, no clusters have gone offline—even during maintenance activities and unexpected failures.

Hyper-V has continued to be exceptionally stable. It is still delivering the same end-to-end availability of MSDN® and TechNet compared to that of the previous physical platform, based on the experience of running the DCI solution for over a year and a half.

Self-Healing During Failures

The DCI instances have continued to deliver services in spite of experiencing some unexpected failures:

* When a physical server failed, the virtual machines were failed over automatically to the maintenance-mode node in the cluster. This self-healing capability resulted from configuring the virtual machines to be highly available when deployed. The virtual machines experienced only a few minutes of downtime, just long enough for them to be failed over. This outage was short enough that by the time the first-tier support team received and began reviewing the first alert, the system had self-healed.
* A similar resolution occurred when a network port failed. In this case, the use of redundancy for network communications enabled the virtual machines to be failed over to the maintenance-mode node.

Uptime During Maintenance

MSCOM Ops has performed maintenance activities on at least a monthly basis since the initial deployment of a DCI instance in June 2009. These maintenance activities have included software updates and hardware upgrades. It is worth repeating here that no clusters have gone offline during this time.

Virtual Machine Deployment Agility

Previously, server deployment took weeks or months to complete because the process entailed acquiring new hardware and locating data-center space before building the system. In the virtualized environment that the DCI provides, virtual machines are often deployed the same week that they are requested. Better still, servers can be deployed in the DCI in less than an hour if business needs require it.

Because opening and modifying the hardware of an already deployed server was prohibitively expensive, MSCOM Ops did not undertake such direct changes with the physical servers. Instead, if business needs dictated increasing storage or memory, the team purchased and deployed a new server—often taking months to address the change request. No such lengthy process exists for reconfiguring virtual machines in the DCI. Making these types of changes takes only hours or days in the DCI environment.

The team reconfigured or decommissioned the virtual machines that host approximately 25 percent of the new applications deployed in a DCI. Changes included reducing the allocated memory or adding or reducing storage for a virtual machine.

Resource Optimization and Power Usage

MSCOM Ops is still in the process of deploying a new instance of the DCI. The team is decommissioning physical servers that are currently, or are about to be, out of warranty. As much as is possible, the team is shifting those workloads to virtual machines. The team does not expect to reach maximum capacity in any of its DCI deployments anytime soon, and yet is already seeing a more efficient use of DCI resources and power as compared to use of physical servers.

Table 6 shows how moving server workloads from an entirely physical environment to the virtualized environment provided by the DCI results in increased utilization of hardware resources and a significant reduction in power usage. Metrics included in the “DCI” column are based on an analysis of the longest-running DCI instance.

Table 6. Comparison of Storage, CPU, and Power Utilization

|  |  |  |  |
| --- | --- | --- | --- |
| Resource | Description | Physical | DCI |
| Storage | Figures show averages based on used hard disk drive space from what are now decommissioned physical servers versus used CSV space for the same set of hosted applications. | Approximately 30% | Approximately 50% and growing as content for hosted systems increases |
| CPU | Figures show average CPU utilization based on what are now decommissioned physical servers versus CPU utilization from virtual machines deployed on the oldest DCI cluster. | Less than 10% utilized on average | More than 20% utilized on average for 16 hosts, including maintenance-mode node |
| Power | Using three popular DCI virtual machine configurations as the basis for this comparison, these figures show the average power that physical servers draw versus the power that comparable virtual machines draw. These figures take into account the power that the two-socket host and the SAN use. | 187 watts | 45 watts |

MSCOM Ops expects to continue to achieve better utilization of storage and CPU capacity. The team also expects to continue saving power as the number of DCI-hosted applications increases.

Virtual Machine Migration

The DCI enables MSCOM Ops to use live migration, quick migration, and quick storage migration for managing virtual machines. These capabilities provide a much higher degree of architectural flexibility than was previously possible for Microsoft.com properties.

MSCOM Ops uses live migration to rebalance host server utilization (CPU and memory) within the same cluster. The team also moves virtual machines to the maintenance-mode node via live migration while patching software on the host server or while performing maintenance on physical hardware components.

In some cases, quick migration or quick storage migration is the only option for performing an operational task. For example, the team uses quick storage migration to balance storage between CSVs within the same cluster. Some operational scenarios require migrating virtual machines between two clusters, a function that quick migration makes possible.

**Best Practices**

Many of the lessons that MSCOM Ops learned from designing and implementing the DCI may be valuable to other IT pros. This section summarizes generalized architecture and process best practices based on the team’s experience. These best practices can help guide other IT pros who are considering the implementation of a virtualized architecture.

Plan for High Availability

This section addresses several considerations that are useful to IT pros who are interested in designing and building systems with high availability as a key requirement. Not all of these strategies will apply to all scenarios. Thus, wherever possible, this section provides references to additional background information.

Operators may approach planning for uptime in many ways when using Windows Server and other Microsoft technologies as part of the hosting infrastructure. Becoming familiar with the following pages can aid the process of planning to move to a fully or partially virtualized architecture:

* Windows Server Enhanced Capability Editions at <http://www.microsoft.com/windowsserver2008/en/us/high-availability-ent.aspx>
* Virtualization with Hyper-V: Technical Resources at <http://www.microsoft.com/windowsserver2008/en/us/hyperv-technical-resources.aspx>
* Failover Clustering at <http://www.microsoft.com/windowsserver2008/en/us/failover-clustering-multisite.aspx>

Identify Hardware Requirements

From the start, the MSCOM Ops team made careful decisions about what hardware to purchase. It based these decisions on an understanding of current and forecasted needs for processing power, memory, and storage, and on the expected life of the hardware. Expecting resiliency, reliability, and high availability from discount-quality hardware is probably unrealistic. An organization should carefully analyze availability requirements and then buy accordingly.

Design for Automatic Failover

Using Hyper-V as the basis for virtualization and CSV for storage enables operators to include automatic failover for hosts. To do this, an organization should deploy the virtual machines as highly available. Guidance on how to do this is available in “Hyper-V: Using Hyper-V and Failover Clustering” at <http://technet.microsoft.com/en-us/library/cc732181(WS.10).aspx>.

Use Live Migration

After a virtualized infrastructure based on Hyper-V is in place, IT pros can use it in a number of ways to complete tasks more efficiently. Examples based on the experience of MSCOM Ops in running its DCI include:

* Rebalancing host server utilization by moving virtual machines to maximize use of existing CPU, network, and memory capacity within the same cluster.
* Performing security updates or replacing hardware for host servers. Specifically, operators should consider using live migration to move virtual machines to other nodes before completing maintenance tasks.

Note: Virtual machines slow down during the live migration process. Using live migration does not enable high availability for virtual machines. For background information and guidance on how to make a virtual machine highly available, see “Configure a Virtual Machine for High Availability” at <http://technet.microsoft.com/en-us/library/cc742396.aspx>.

Use Maintenance Mode

System Center Virtual Machine Manager enables the configuration of maintenance-mode nodes. The larger a cluster is, the higher the probability is that one host will fail. Keeping one node in maintenance mode enables an operator to fail over to it if such an incident occurs.

In addition to the total number of nodes, an organization should consider the type of cluster being deployed when it is determining the number of maintenances modes to have. For example, if the cluster hosts databases or other systems that maintain state, a minimum of two maintenance-mode nodes provides better protection against unexpected downtime. To learn more about the possible scenarios for combining use of live migration with use of a maintenance-mode node, see “About Maintenance Mode” at <http://technet.microsoft.com/en-us/library/ee236481.aspx>.

### Consider Multiple Instances for System Center Virtual Machine Manager

Although managing multiple data centers from one instance of System Center Virtual Machine Manager is possible according to “Supported Configurations for VMM” at <http://technet.microsoft.com/en-us/library/cc764231.aspx>, an organization should consider deploying one instance of System Center Virtual Machine Manager for each data center that uses Hyper-V. In general, deciding how many instances are needed will follow network and environment boundaries. For additional guidance, see “Planning for the VMM Server” at <http://technet.microsoft.com/en-us/library/cc764331.aspx>.

Optimize Use of Capacity

Designing a new architecture based on virtualization can take many forms. This section summarizes best practices specific to creating an initial plan for virtualization with the goal of server consolidation in mind. For additional virtualization scenarios and guidance for beginning to virtualize, visit the Microsoft Virtualization site at <http://www.microsoft.com/virtualization>.

Convert Existing Physical Servers to Virtual

An organization should always strive to virtualize appropriately. Conducting an analysis based on existing performance and utilization data is essential for determining whether a server or an application is an appropriate candidate for virtualization.

Each IT organization has unique business and technical requirements. To learn more about the types of virtualization possible, read the white papers and case studies on the Virtualization Products and Technologies page at <http://www.microsoft.com/virtualization/en/us/products-server.aspx>. In addition, the Microsoft Assessment and Planning (MAP) Toolkit—a free download available at <http://technet.microsoft.com/en-us/library/bb977556.aspx>—can assist in the analysis and planning for a Hyper-V environment.

Conduct Performance Testing for New Applications

An organization should not assume that all server roles and workloads can be virtualized. With new applications, the organization should use as much information as is available to formulate an understanding of all the performance characteristics of the workload. It should keep in mind that virtual machine resources can be added or reclaimed and establish appropriate analysis guidelines and team processes.

Balance Network Use

The ability to monitor and balance network connections is important to providing acceptable levels of service delivery from a virtualized infrastructure because of sharing physical resources between multiple virtual machines. If one virtual machine is putting a heavy load on the existing network connections, it will affect the performance of the other virtual machines.

When planning and building a virtualized infrastructure, an organization should identify which counters are critical to monitor. What these counters are and how best to use the available instrumentation for monitoring largely depend on the hardware components and the design of network communications.

Overallocate SAN

If using a SAN for storage, operators may opt to use dynamic VHDs and to overallocate the storage space. This approach allows for better overall utilization of storage space. As noted previously, operators who implement this strategy must monitor space thresholds closely to avoid having VHDs pause unexpectedly. They should set initial monitoring thresholds based on expected business growth and review them regularly.

Use Multiple CSVs

When an organization is using CSVs on a SAN for storage, LUN growth is possible. For details, see “Add Storage to a Failover Cluster” at <http://technet.microsoft.com/en-us/library/cc733046.aspx>. For more background and access to technical white papers on failover clustering in Windows Server 2008 R2, visit the Failover Clustering page at <http://www.microsoft.com/windowsserver2008/en/us/failover-clustering-technical.aspx>.

Consider and Implement Automation

Novel solutions that combine new technologies with old standards like DHCP servers can simplify routine tasks. An organization should simplify and standardize the host computer and virtual machine configuration as much as possible. These standardizations make it easier to use existing snap-ins, programs that run in the context of Microsoft Management Console, to create new scripts for automating routine tasks and to take advantage of features like virtual machine templates. Operators can create Windows PowerShell cmdlets to automate many tasks, including creation of regular reports used to monitor system health.

Find New Team Efficiencies

Virtualization that results in the abstraction of hardware and platform software from hosted applications enables teams to perform issue resolution and maintenance more efficiently. In addition to having the ability to update operating systems without interrupting service delivery, operators can make hardware changes to the host without disrupting customers. Self-healing capabilities significantly accelerate the process of resolving infrastructure issues. The ability to rapidly provision virtual machines as business needs dictate makes new operational efficiencies possible. With this new agility, operators may need to update team processes to ensure that the right levels of monitoring and analysis are in place for managing the virtualized infrastructure.

**Conclusion**

MSCOM Ops is already reaping the benefits of having a virtualized architecture based on its DCI design. By building this hosting infrastructure that uses SAN-based CSVs and live migration based on Hyper-V and Windows Server 2008 R2, the team has significantly increased architectural flexibility.

The following benefits show that MSCOM Ops has achieved the business goals that it established when it designed a virtualized environment for hosting Microsoft.com, MSDN, and TechNet:

* Customers no longer need to wait weeks or months to access additional capacity for hosted applications.
* When more memory or storage is deployed than is ultimately needed for new applications, the team can reclaim the underutilized capacity in a matter of hours or days.
* Maintenance for the host servers, whether driven by the need to change software or hardware, occurs without interrupting the hosted virtual machines.
* Better still, if hardware fails, quick migration occurs, limiting the downtime to mere minutes instead of hours or days.
* With a power savings of 67 percent for the two-socket host clusters when compared to the power that equivalent physical servers use, the team is well on its way to containing the costs associated with running one of the world's largest and most in-demand corporate Web sites.

The DCI offers a resiliency of systems that simply was not possible with the previous infrastructure. The fact that service delivery has continued in spite of hardware failures, thanks to the self-healing capabilities of this solution, has in itself shown the value of using CSVs for SAN-based storage.

**For More Information**

For more information about Microsoft products or services, call the Microsoft Sales Information Center at (800) 426-9400. In Canada, call the Microsoft Canada Order Centre at (800) 933-4750. Outside the 50 United States and Canada, please contact your local Microsoft subsidiary. To access information via the World Wide Web, go to:

<http://www.microsoft.com>

<http://www.microsoft.com/technet/itshowcase>

The information contained in this document represents the current view of Microsoft Corporation on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.

This White Paper is for informational purposes only. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED, OR STATUTORY, AS TO THE INFORMATION IN THIS DOCUMENT.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Microsoft Corporation.

Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Microsoft, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

Unless otherwise noted, the example companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted herein are fictitious, and no association with any real company, organization, product, domain name, e-mail address, logo, person, place, or event is intended or should be inferred.

 2010 Microsoft Corporation. All rights reserved.

Microsoft, Hyper-V, MSDN, Windows PowerShell, and Windows Serverare either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.

All other trademarks are property of their respective owners.