Realtime
publishers

The Shortcut Guide[tm] To

Architecting iSCSI Storage for Microsoft Hyper-V

sponsored by

hp®

Greg Shields

Copyright Statement

If you have any questions about these terms, or if you would like information about licensing materials from Realtime Publishers, please contact us via e-mail at info@realtimepublishers.com.

Chapter 3: Critical Storage Capabilities for Highly-Available Hyper-V

Chapter 2 highlighted the fact that *high availability is fundamentally critical to a successful Hyper-V infrastructure*. This is the case because uncompensated hardware failures in any Hyper-V infrastructure have the potential to be much more painful than what you're used to seeing in traditional physical environments.

A strong statement, but think for a minute about this increased potential for loss: In any virtual environment, your goal is to optimize the use of physical equipment by running multiple virtual workloads atop smaller numbers of physical hosts. Doing so gives you fantastic flexibility in managing your computing environment. But doing so, at the same time, increases your level of risk and impact to operations. When ten workloads, for example, are running atop a single piece of hardware, the loss of that hardware can affect ten times the infrastructure and create ten times the pain for your users.

Due to this increased level of risk and impact, you must plan appropriately to compensate for the range of failures that can potentially occur. The issue here is that no single technology solution compensates for every possible failure. Needed are a set of solutions that work in concert to protect the virtual environment against the full range of possibilities.

Depicted in Figure 3.1 is an extended representation of the previous chapter's fully-redundant Hyper-V environment. There, each Hyper-V server connects via multiple connections to a networking infrastructure. That networking infrastructure in turn connects via multiple paths to the centralized iSCSI storage infrastructure. Consider for a minute which failures are compensated for through this architecture:

- Storage and Production Network traffic can survive the loss of a single NIC due to the incorporation of 802.3ad network teaming and/or MPIO/MCS.
- Storage and Production Network traffic can also survive the loss of an entire network switch due to the incorporation of 802.3ad network teaming and/or MPIO/MCS that has been spread across multiple switches.
- Virtual machines can survive the planned outage of a Hyper-V host through Live Migration as a function of Windows Failover Clustering.
- Virtual machines can also be quickly returned to service after the unplanned outage of a Hyper-V host as a function of Windows Failover Clustering.
- Network oversubscription and the potential for virtual machine denial of service are inhibited through the segregation of network traffic across Storage, Production, Management, and Heartbeat connections.

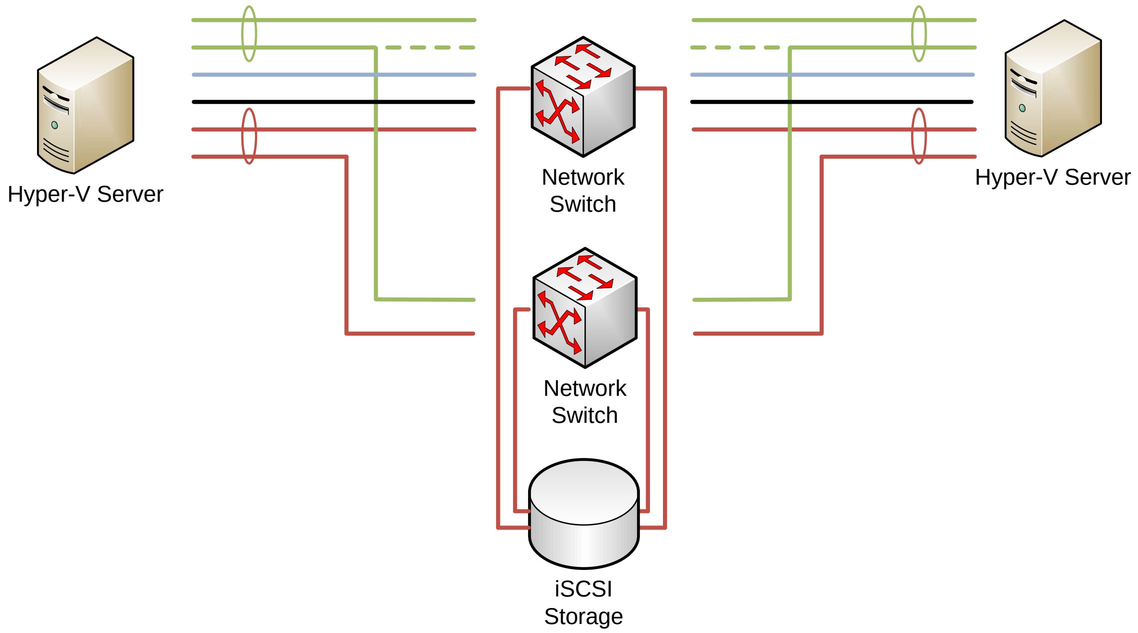

Figure 3.1: Hyper-V environments require a set of solutions to protect against all of the possible failures.

The risk associated with each of these potential failures has been mitigated through the implementation of multiple layers of redundancy. However, this design hasn't necessarily taken into account its largest potential source of risk and impact. Take another look at Figure 3.1. In that figure, one element remains that in and of itself can become a significant single point of failure for your Hyper-V infrastructure. *That element is the iSCSI storage device itself.*

Each and every virtual machine in your Hyper-V environment requires storage for its disk files. This means that any uncompensated failure in that iSCSI storage has the potential to take down each and every virtual machine all at once, and with it goes your business' entire computing infrastructure. As such, there's a lot riding on the success of your storage infrastructure. This critical recognition should drive some important decisions about how you plan for your Hyper-V storage needs. It is also the theme behind this guide's third chapter.

Virtual Success Is Highly Dependent on Storage

In the end, storage really is little more than just a bunch of disks. You must have enough disk space to store your virtual machines. You must also have enough disk space for all the other storage accoutrements that a business computing environment requires: ISO files, user home folders, space for business databases, and so on. Yet while raw disk space itself is important, the architecture and management of that disk space is exceptionally critical to virtualization success in ways that might not be immediately obvious.

Chapter 2 introduced the suggestion that the goal for SAN availability is “no nines,” or what amounts to 100% availability. Although this requirement might seem an impossibility at first blush, it is in fact a necessity. The operational risk of a SAN failure is made even more painful by the level of impact such an event will have on your environment. As a result, your goal in selecting, architecting, and implementing your SAN is to ensure that its design contains no single points of failure.

Today’s iSCSI SAN equipment accomplishes this lofty goal through the concurrent implementation of a set of capabilities that layer on top of each other. This layered approach to eliminating points of failure ensures that surviving hardware always has the resources and data copies it needs to continue serving the environment without interruption.

> **“Non-Interruptive” Is Important**
>
> This concept of non-interruptive assurance during failure conditions is also critical to your SAN selection and architecture. Your selected SAN must be able to maintain its operations without interruption as failures occur. Although non-interruptive in this definition might mean an imperceptibly slight delay as the SAN re-converges after a failure, that delay must be less than the tolerance of the servers to which it is connected. As you’ll discover later in this chapter, non-interruptive is important not only during failure operations but also during maintenance and management operations.

The easiest way to understand how this approach brings value is through an iterative look at each compensating layer. The next few sections will discuss how today’s best in class iSCSI SAN hardware has eliminated the SAN as a potential single point of failure.

Modular Node Architecture

Easily the most fundamental new approach in eliminating the single point of failure is in eliminating the “single point” approach to SAN hardware. Modern iSCSI SAN hardware accomplishes this by compressing SAN hardware into individual and independent modules or “nodes.” These nodes can be used independently if needed for light or low-priority uses. Or, they can be logically connected through a storage network to create an array of nodes.

Figure 3.2 shows a logical representation of how this architecture might look. Here, four independent storage nodes have been logically connected using their built-in management software and a dedicated storage network. Within each node are 12 disks for data storage as well as all the other necessary components such as processors, power supplies, NICs, and so on. The result of connecting these four devices is a single logical iSCSI storage device. That device has the capacity to present the summation of each device’s available storage to users and servers.

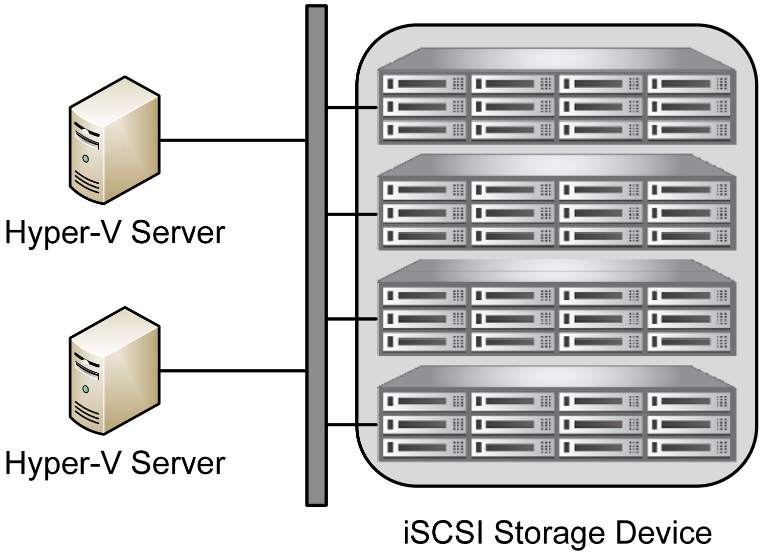

Figure 3.2: Multiple storage nodes aggregate to create a single logical device.

Important to recognize here is that each device can be an independent entity or aggregated with others to modularly increase the capacity of the SAN. This modular approach can be added to or subtracted from as the data needs of its owner changes over time. This presents a useful benefit to the ownership of such a SAN over more traditional monolithic approaches: *Its capacity can be expanded or otherwise modified as necessary without the need for wholesale hardware replacements.*

Consider as an alternative the more traditional monolithic SAN. These devices rely on the population of a storage "frame" with disks, storage processors, and switch fabric devices. In this type of SAN, there is a physical limit to the amount of storage that can be added into such a frame. Once that frame is full to capacity, either additional frames must be purchased or existing disks or frames must be swapped out for others that have greater capacity. The result can be a massive capital expenditure when specific threshold limits are exceeded.

Using the modular approach, new modules can be added to existing ones at any point. Management software within each module is used to complete the logical connection through the dedicated storage network. That same software can be configured to automatically accomplish post-augmentation tasks such as volume restriping and re-optimization on behalf of the administrator. This chapter will talk more about these management functions shortly.

Redundant Storage Processors Per Node

Modularization alone does nothing to enhance storage availability. It also does nothing to enhance the resiliency of the individual node and its data. However, it does provide the framework in which much of the aforementioned advanced availability features lie.

Every storage device requires some sort of processor in order to accomplish its stated mission. Although some processors leverage entirely proprietary code, many processors today rest atop highly-tailored distributions of existing operating systems (OSs) such as Linux or Windows Storage Server. No matter which OS is at its core, one architectural element that is critical to ensuring node resiliency is the use of redundant storage processors within each individual node.

Figure 3.3 shows how this might look in a storage device that is comprised of four nodes. Here, each individual node includes two storage processors that are clustered for the purposes of redundancy. With this architecture in place, the loss of a storage processor will not impact the functionality of the individual node.

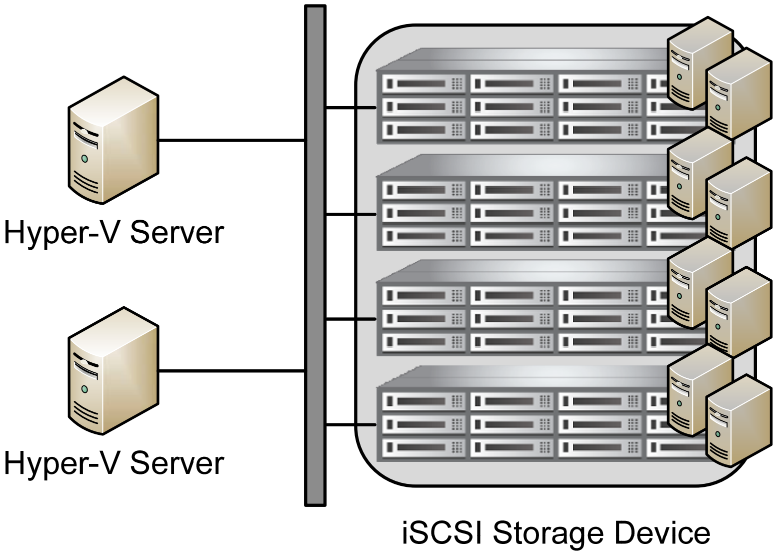

Figure 3.3: Multiple storage processors per node ensure individual node resiliency.

This architecture comes in particularly handy when nodes are used independently. In this configuration, a single node can survive the loss of a storage processor without experiencing an interruption of service.

Redundant Network Connections and Paths

Redundancy in processing is a great feature, but even multiple storage processors cannot assist when network connections go down. The risk of network failure is in fact such a common occurrence that the entirety of Chapter 2 was dedicated to highlighting the necessary server-to-SAN connections that are required for Hyper-V.

Yet that discussion in Chapter 2 did not include one critical redundancy element that is shown in Figure 3.4. This redundancy becomes relevant when used in the framework of a modular SAN architecture. There, each individual storage node has also been connected to the storage network using redundant connections.

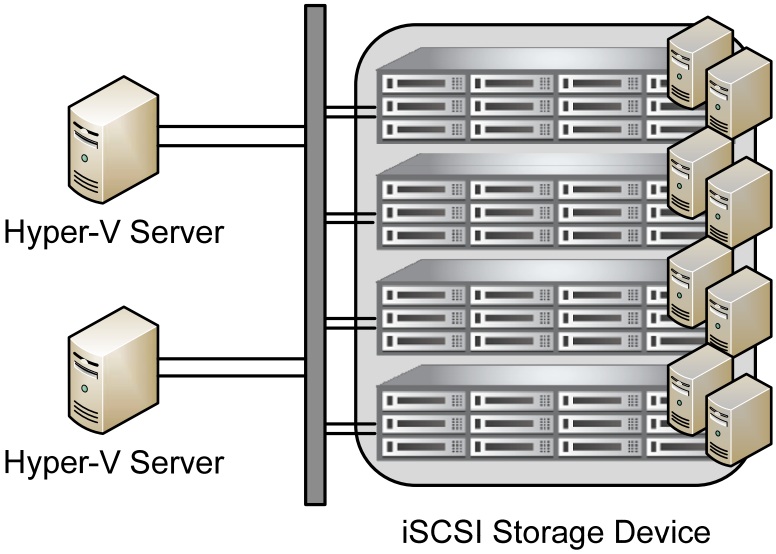

Figure 3.4: Redundant connections and paths relate to inter-node communication as well as server-to-node.

Important to recognize here is that this configuration is necessary not only for resiliency but also for raw throughput. Because each individual storage node is likely connected to by multiple servers, the raw network performance in and out of each node can be more than is possible through a single connection. Although all iSCSI storage nodes have at least two network connections per node, those that are used in support of extremely high throughput may include four or more to support the necessary load.

> **Note**
>
> Measuring that performance is a critical management activity. iSCSI storage nodes tend to come equipped with the same classes of performance counters that you're used to seeing on servers: Processor, network, and memory utilization are three that are common. Connecting these counters into your monitoring infrastructure will ensure that your Hyper-V server needs aren't oversubscribing any part of your SAN infrastructure.

Disk-to-Disk RAID

RAID has been around for a long time. So long, in fact, that it is one of those few acronyms that doesn't need to be written out full when used in guides like this one. Although RAID has indeed had a long history in IT, it's important to recognize that it is another high-availability feature that you should pay attention to as you consider a SAN storage device for Hyper-V.

The reason behind this special consideration has to do with the many types of RAID protection that SANs can deploy over and above those traditionally available within individual servers. These added RAID levels are made possible in many ways due to the sheer number of disks that are available within an individual storage node.

Figure 3.5 shows a graphical representation of how some of these might look. In addition to the usual RAID 1 (mirroring), RAID 5 (striping with parity), and RAID 1+0 (disks are striped, then mirrored) options that are common to servers, SANs can often leverage additional RAID options such as RAID-with-hot-spares, RAID 6 (striping with double parity), and RAID 10 (disks are mirrored, then striped), among others.

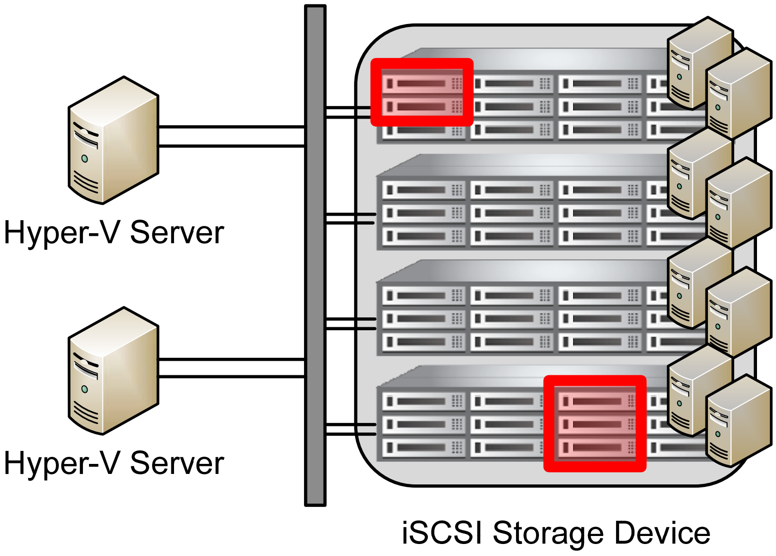

Figure 3.5: Disk-to-disk RAID in iSCSI storage devices is functionally similar to RAID within individual servers.

These alternative options are often necessary as the size of SANs grow due to the potential for multiple disk failures. Although the traditional RAID levels used in servers are designed to protect against a single disk failure, they are ineffective against the situation where more than one disk fails in the same volume. The added level of protection gained through advanced RAID techniques becomes increasingly necessary when large numbers of individual disks are present in each storage node.

Node-to-Node RAID

Another RAID capability that is not common to server disk drives is the capacity to span volume redundancy across multiple nodes. In fact, this feature alone is one of the greatest reasons to consider the implementation of a multiple-node architecture for the storage of Hyper-V virtual machines as well as other critical data.

In Figure 3.6, the red boxes that represent intra-node RAID have been augmented with another set of purple boxes. This second set of boxes highlights how node-to-node RAID configurations can span individual nodes. In this configuration, volumes have been configured in such a way that every piece of data on one node (or its parity information) is always replicated to one or more additional nodes in the logical device.

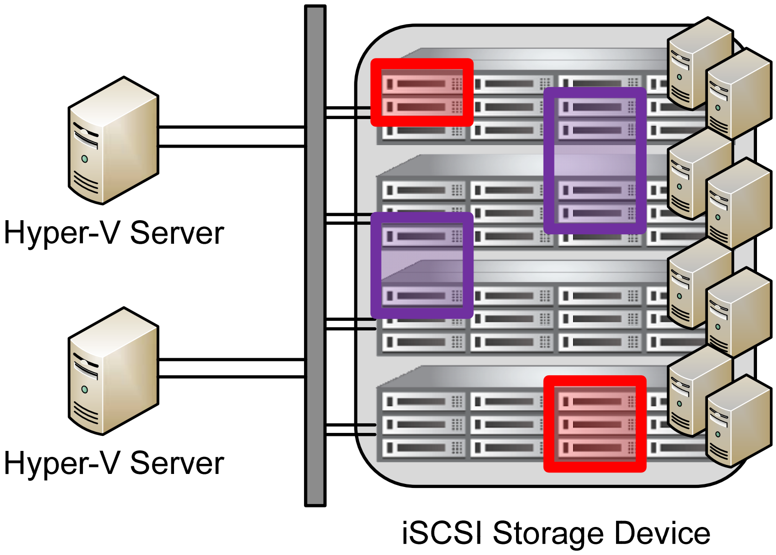

Figure 3.6: Node-to-node RAID ensures that entire nodes can fail with no impact to operations.

Note

Although Figure 3.6 shows an example of a RAID set that has been created across only a few disks in a few nodes, it is more common that RAID sets are created across every disk in the entire logical storage device. By creating a hardware RAID set in this manner, the entire device's storage can then be made available to exposed volumes.

Depending on the storage device selected, multiple levels of node-to-node RAID are possible with each having its own benefits and costs. For example, each block of data can be replicated across two nodes. This configuration ensures that a block of data is always in two places at once. As an alternative that adds redundancy but also adds cost, each block can be replicated across three nodes, ensuring availability even after a double-node failure.

This architecture is critically important for two reasons. First, it extends the logical storage device's availability *to protect against failures of an entire node or even multiple nodes*. The net result is the creation of a storage environment that is functionally free of single points of failure.

As a second reason, such an architecture also increases the capacity of the logical storage device's volumes *to greater than the size of a single node*. Considering the large size of Hyper-V virtual machines, extremely large volume sizes may be necessary, such as those that are larger than can be supported by a single node alone.

Modularization Plus Disk-to-Disk RAID Equals Swap-Ability

Interesting to note here is how the combination of disk-to-disk RAID goes hand-in-hand with modularization. This combination of capabilities enables SAN hardware to be very easily replaced in the case of an entire-node failure, making the individual node itself *a hot-swappable item*.

Think for a minute about how this might occur: Every block of data on such a SAN is always replicated to at least one other storage node. Thus, data is always protected when a node fails. When a failure occurs, an administrator needs only to remove the failed node and swap it with a functioning replacement. With minimal configuration, the replacement can automatically reconnect with the others in the logical storage device and synchronize the necessary data. As a result, even an entire node failure becomes as trivial as an individual disk failure.

Integrated Offsite Replication for Disaster Recovery

And yet even these capabilities don't protect against the ultimate failure: the loss of an entire operational site. Whether that loss is due to a natural disaster, one that is man-made, or a misconfiguration that results in massive data destruction, there sometimes comes the need to relocate business operations in their entirety to a backup site.

What's particularly interesting about disaster recovery and its techniques and technologies is that many are newcomers into the IT ecosystem. Although every business has long desired a fully-featured disaster recovery solution, only in the past few years have the technologies caught up to make this dream affordable.

Needed at its core is a mechanism to replicate business data as well as data processing to alternate locations with an assurance of success. Further, that replication needs to occur in such a way that minimizes bandwidth requirements. To be truly useful, it must also be a solution that can be implemented without the need for highly-specialized training and experience. In the case of a disaster, your business shouldn't need specialists to failover your operations to a backup site nor fail them back to the primary site when the disaster is over.

Today's best-in-class iSCSI SANs include the capability to connect a primary-site SAN to a backup-site SAN as Figure 3.7 shows. This being said, such a connection is a bit more than just plug-and-go. There are some careful considerations that are important to being successful, most especially when SAN data consists of Hyper-V virtual machines.

Cross-Reference

Chapter 4 will explore the architectures and requirements for disaster recovery in more detail.

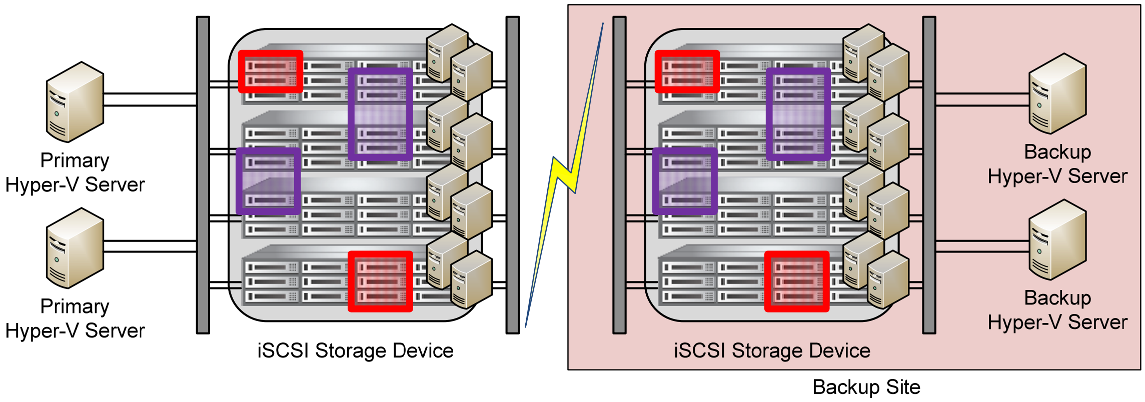

Figure 3.7: Automated replication of changed data to alternate sites protects against entire site loss.

Non-Interruptive Capacity for Administrative Actions

It has already been stated that architecting your storage infrastructure is exceptionally important to be successful with Hyper-V. Yet getting that storage up and operational is only the first step in actually using your Hyper-V virtual environment. It's also the shortest step. Longer in timeframe and arguably more important are the management activities you'll undergo after the installation is complete.

The processes involved with managing Hyper-V storage often get overlooked when the initial architecture and installation is planned. However, these same administrative tasks, when not planned for, can cause complications and unnecessary outages down the road. No matter which action needs to be accomplished, your primary goal should be an ability to invoke those actions with the assurance that *they will not interrupt running virtual machines*.

If these statements sound alarmist, consider the long-running history of storage technologies. In the not-too-distant past, otherwise simple tasks became operational impacts due to their need for volume downtime. These tasks included basic administrative actions such as extending an existing volume to add more disk space, installing a firmware upgrade, or augmenting the environment with additional nodes or frames. In the most egregious of examples, simple tasks such as these sometimes required the presence of on-site assistance from manufacturer storage technicians.

That historical limitation added substantial complexity and cost to SAN ownership. Today, such limitations are wholly unacceptable when considered with the availability requirements needed by a virtual infrastructure. *Your business simply can't bring down every virtual machine when you need to make a small administrative change to your storage.*

With this in mind, consider the following set of administrative activities that are common to all storage environments. Your SAN hardware should be able to accomplish each of them without interruption to virtual machine processing or other concurrent data access. Further, they also represent actions that a sufficiently-experienced administrator should be able to accomplish with the right hardware and minimal tool-specific instruction.

Note

With these activities, iSCSI isn't alone. Many of the features explained in the following sections should be available in other types of SAN equipment such as those that leverage fibre channel connections. Often, however, these features are only available at extra cost. This is an important consideration when purchasing a new storage infrastructure. Look carefully to the capabilities that are offered by your SAN vendor to ensure that the right set of management activities is available for your needs. For some vendors, you may need to purchase the rights to use certain management functions. As an alternative, look to an all-inclusive SAN vendor that does not price out advanced functionality at extra cost.

Volume Activities

Early monolithic SAN infrastructures required complex configuration file changes when volumes needed reconfiguration. For some vendors, this configuration file change was an exceptionally painful operation, often requiring the on-site presence of trained professionals to ensure its successful implementation.

Today, volume changes are relatively commonplace activities. Administrators recognize that provisioning too much storage to a particular volume takes away disk space from other volumes that might need it down the road. It is for this reason that today's best practices in volume size assignment are to maintain a small but constant percentage of free space. This sliding window of available space can require administrators to constantly monitor and adjust sizes as needed. Some SANs have the capability to automatically scale the size of volumes per preconfigured thresholds. No matter which method you use, this activity on today's iSCSI SANs should not require downtime to either the volume or connected users and servers.

Advanced SANs provide the capability to accomplish other volume-based tasks without interruption as well. These tasks can relate to changing how a volume is provisioned, such as thin-provisioned versus pre-allocated, or configured RAID settings. For example, volumes that start their operational life cycle as a low priority resource may later grow in criticality and require additional RAID protection. That reconfiguration should occur without interruption to operations.

Storage Node Activities

Activities associated with the storage node itself should also be accomplished without impact to data access. For example, adding, removing, or replacing storage nodes from a logical storage device are tasks that can and should be possible without interruption. Important to recognize here are the non-interruptive internal activities that must occur in the background after such a dramatic change to the storage environment:

- Adding a node automatically restripes existing volumes across the new node, balancing storage across the now-larger logical storage device.
- Removing a node automatically relocates data off the node prior to the actual removal activity, ensuring that data remains available even after the node has been removed from the logical storage device.
- Replacing a node automatically rebuilds volumes from surviving data on the remaining nodes.

Another useful cross-node activity is the use of automated volume restriping to reduce spindle contention. This problem of spindle contention was first introduced in Chapter 1 and can have a larger-than-normal impact on storage that is part of a virtualization infrastructure. In essence, when the disk use of virtual machines becomes greater than expected, virtual machines whose disk files share the same disk spindles in the SAN infrastructure will experience a bottleneck. Collocated virtual machines in this situation experience a collective reduction in performance as each vies for attention by the storage device.

To alleviate this situation, some storage devices have the ability to watch for spindle contention and transparently relocate data files to alternate locations on disk. The result is a more optimized distribution of storage hardware resources across the entire logical device as well as better overall performance for virtual machines.

Data Activities

Storage arrays commonly include the ability to snapshot volumes as well as replicate them to other locations within and outside the logical device. Snapshotting activities are critical to reducing backup windows. They also provide the ability to quickly create point-in-time copies of virtual machines for testing or other purposes.

Replication is often necessary when virtual machines or other data must be offloaded to alternate volumes or logical storage devices—this can be due to a forklift upgrade of the logical storage device or because it is necessary to create copies of volumes for device-to-device replication. As with the other activities, completing these data-related activities should be a non-interruptive process.

Firmware Activities

Last, is the not-uncommon activity associated with updating the firmware on individual storage nodes. All storage devices require the occasional update of firmware code in order to add features, eliminate bugs, and update code to prevent known attacks.

This updating of SAN firmware must be an operation that does not require downtime. Downtime prevention may occur as a function of multiple storage processors or in using an OS that can implement updates without requiring a reboot.

Storage Virtualization

The concepts that embody storage virtualization share little with those that are associated with traditional server virtualization. However, they do share the same high-level meaning in that storage virtualization is also about *abstraction*. In the case of storage virtualization, the abstraction exists between logical storage (RAID sets, volumes, and so on) and the actual physical storage where that data resides.

You've already been exposed in this chapter to many of the capabilities that fall under the banner of storage virtualization: The ability to snapshot a drive abstracts the snapshot from the bits in its initial volume. Restriping a volume across multiple nodes requires a layer of abstraction as well. Accomplishing this task requires a meta-layer atop the volume that that maps the logical storage to physical locations.

In the context of virtualization atop platforms such as Hyper-V, storage virtualization brings some important management flexibility. It accomplishes this through the introduction of new features that improve the management of Hyper-V virtualization. Let's look at a few of these features in the following sections.

Snapshotting and Cloning

Creating snapshots of volumes enables administrators to work with segregated copies of data but without the need to create entirely duplicate copies of that data. For example, consider the situation where you need to test the implementation of an update to a set of virtual machines on a volume. Using your SAN's snapshotting technology, it is possible to create a duplicate copy of that entire volume. Because the volume has been created as a snapshot rather than a full copy, the time to complete the snapshot is significantly reduced. The level of consumed space is also only a fraction of the overall volume size.

Once created, actions like the aforementioned update installation can be completed on the snapshot volume. If the results are a success, the snapshot can be merged into the original volume or discarded.

Backup and Restore with VSS Integration

Snapshots are useful for other reasons as well. Backup operations are made much easier through the use of snapshots. Integrating those snapshots with Microsoft's Volume Shadow Copy (VSS) ensures that backups successfully capture the state of the virtual machine along with its installed applications. Without VSS integration, installed applications and their data may not be correctly backed up. When seeking a SAN to be used in a virtualized environment, it is important to look for those that support VSS integration to ensure backups of these types of applications.

Volume Rollback

A key advanced feature is the ability for volumes to be rolled backwards in time. This need can occur after a significant data loss or data corruption event. Combining snapshot technology with the capacity to store multiple snapshot iterations gives the administrator a series of time-based volume snapshots. Rolling a volume to a previous snapshot quickly returns the volume to a state before the deletion or corruption occurred. Further, volume rollback can more quickly return a corrupted volume to operations than traditional restore techniques.

Thin Provisioning

Lastly is the capability for volume thin provisioning. It has already been discussed in this chapter that today's best practices suggest that volumes should be configured to maintain only a small level of free space. This small level ensures that available disk space can always be assigned to the volumes that need them.

One problem with this approach relates to how an OS will make use of an assigned volume. Unlike storage devices, OSs tend to create statically-sized volumes for their configured disk drives. Thus, every storage device volume extension must be followed by a manual volume extension within the OS.

A method to get around this limitation is the use of thin provisioning. Here, a volume is presented to the OS for its anticipated size needs. On the storage device, however, the true size of the volume is only as large as the actual data being consumed by the OS. The storage device's volume automatically grows in the background as necessary to provide free space for the OS. The result is that the OS's volume does not need expansion while the storage device's volume only uses space as necessary. This process significantly improves the overall utilization of free space across the storage device.

> **Caution**
>
> Caution must be used in leveraging thin provisioning to ensure that the real allocation of disk space doesn't go above true level of available disk space. Proper monitoring and alerting of storage space is critical to prevent this catastrophic event from occurring.

Storage Architecture and Management Is Key to Hyper-V

You've seen the comment presented over and over that the task of installing the very basics of Hyper-V is excessively simplistic; the real skill comes in creating a Hyper-V infrastructure that can survive the many possible failures that can and will occur in a production computing environment. Preventing those failures happens with the right combination of a good architecture and the capability to accomplish needed management activities without service interruption. You've learned about these needs in this chapter.

But this chapter's discussion on storage capabilities has left one element remaining. You now understand how your iSCSI storage should be architected to ensure the highest levels of availability. But you haven't really come to understand the special needs that arrive when an entire site goes down. Disaster recovery is the theme in the fourth and final chapter. Coming up, you'll learn about the technologies and techniques you'll need to consider when you expand your operations to a full disaster recovery site.

Download Additional eBooks from Realtime Nexus!