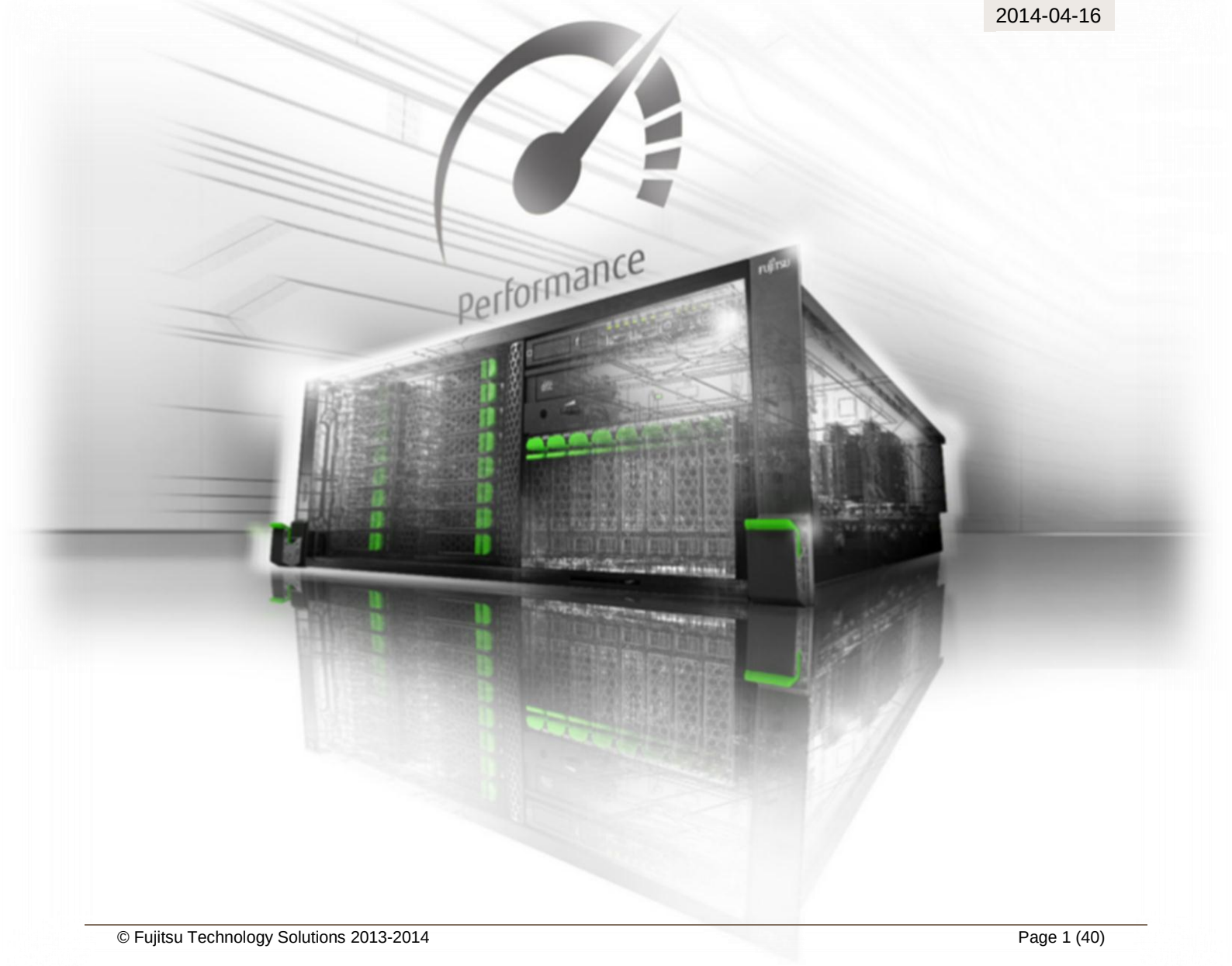
WHITE PAPER
FUJITSU Server PRIMERGY
WINDOWS STORAGE SPACES PERFORMANCE

Microsoft is introducing the Storage Spaces functionality with Microsoft Windows Server 2012 and Microsoft Windows 8. This document will give you an overview of this function and its performance. In addition to the fundamental features and configuration approaches, also with regard to a cluster environment, basis performance values as well as comparative values for the HW RAID environment are mentioned.

Version
1.0a
2014-04-16



Contents

Document history 2
Introduction 3
Measurement method 4
Measurement environment 5
Functionality 6
  Features 6
  Configurations 12
  Storage Spaces in the cluster 14
Performance 19
  Storage Spaces 19
    “Simple” 19
    “Mirror 2-way” 21
    “Parity” 23
    “Mirror 3-way” 24
  Storage Spaces vs. HW RAID (default stripe size) 25
    “Simple” vs. HW RAID 0 25
    “Mirror” vs. HW RAID 1/10 27
    “Parity” vs. HW RAID 5 29
  Storage Spaces vs. HW RAID (256 kB stripe size) 31
    “Simple Mirror Parity” vs. HW RAID 0 / 10 / 5 31
  Storage Spaces and Shared Volume 34
Conclusion 39
Literature 40
Contact 40

Document history

Version 1.0

First version

Version 1.0a

Revised diagrams

Introduction

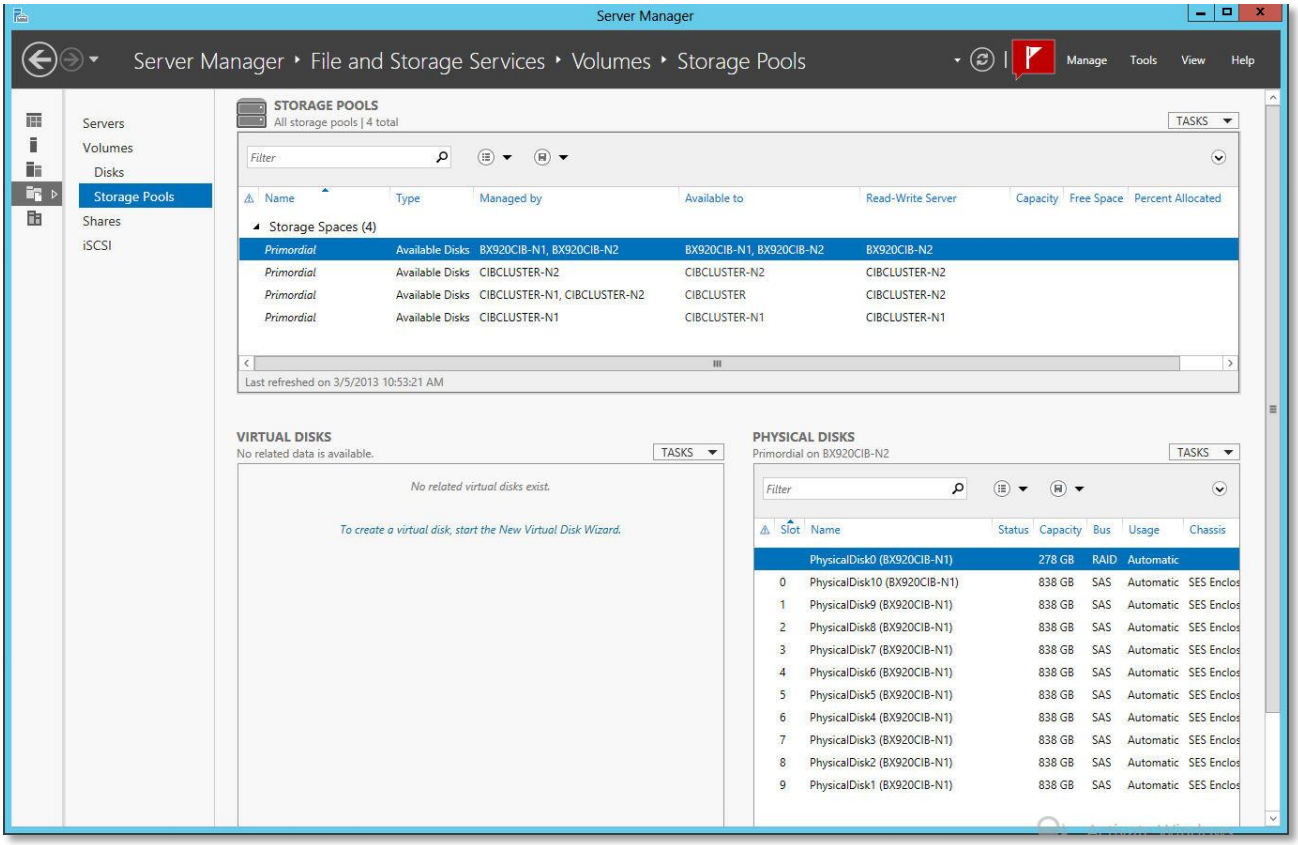
As part of the handling of file and storage subsystems Microsoft is introducing the functionality of storage pools and Storage Spaces with Microsoft Windows Server 2012 and Microsoft Windows 8. Basically, we are dealing with the following functions here:

- Storage virtualization
- RAID functionality
- Thin provisioning (storage reservation)
- Script-based management via PowerShell
- Efficient file system repair
- Integration with cluster shared volumes (CSV)

Storage pools support various connection interfaces, but it is also possible to operate different hard disk types and sizes in a pool. This functionality, combined with a simple and flexible storage configuration, can make sense for minor requirements, e.g. where substantial investments in SAN (Storage Area Network) or NAS (Network Attached Storage) hardware are not appropriate. In contrast to previous data medium management, an entire abstraction as well as required scaling of the physical data media are possible.

An interface (GUI) for the management of the Storage Spaces is provided in the Server Manager of Microsoft Windows 2012 under File and Storage Services. The complete functionalities, plus a number of special functions, are also available as PowerShell commands.

The picture below shows the appropriate Server Manager interface.



Measurement method

As standard, performance measurements of disk subsystems in PRIMERGY servers are carried out with a defined measurement method, which models the hard disk accesses of real application scenarios on the basis of specifications.

The essential specifications are:

- Share of random accesses / sequential accesses
- Share of read / write access types
- Block size (kB)
- Number of parallel accesses (# of outstanding I/Os)

A given value combination of these specifications is known as “load profile”. The following five standard load profiles can be allocated to typical application scenarios:

Standard load profile	Access	Type of access		Block size [KB]	Application
		read	write		
File copy	random	50%	50%	64	Copying of files
File server	random	67%	33%	64	File server
Database	random	67%	33%	8	Database (data transfer) Mail server
Streaming	sequential	100%	0%	64	Database (log file), Data backup; Video streaming (partial)
Restore	sequential	0%	100%	64	Restoring of files

In order to model applications that access in parallel with a different load intensity, the “# of Outstanding I/Os” is increased, starting with 1, 3, 8 and going up to 512 (from 8 onwards in increments to the power of two).

The measurements of this document are based on these standard load profiles.

The main results of a measurement are:

- Throughput [MB/s] Throughput in megabytes per second
- Transactions [IO/s] Transaction rate in I/O operations per second
- Latency [ms] Average response time in ms

The data throughput has established itself as the normal measurement variable for sequential load profiles, whereas the measurement variable “transaction rate” is mostly used for random load profiles with their small block sizes. Data throughput and transaction rate are directly proportional to each other and can be transferred to each other according to the formula

Data throughput [MB/s]	*= Transaction rate [IO/s] × Block size [MB]*
Transaction rate [IO/s]	*= Data throughput [MB/s] / Block size [MB]*

All the details of the measurement method and the basics of disk I/O performance are described in the white paper “Basics of Disk I/O Performance”.

Measurement environment

All the measurement results discussed in this document were determined using the hardware and software components listed below:

System Under Test (SUT)	
Hardware	
Models	1 × PRIMERGY RX200 S7 “Storage Spaces” 1 × PRIMERGY RX300 S7 “HW RAID” 2 × PRIMERGY BX920 S3 + 1 × PRIMERGY SX980 S2 “Cluster”
Processor	PRIMERGY RX200 S7: 2 × Xeon E5-2643 (3.3 GHz) PRIMERGY RX300 S7: 2 × Xeon E5-2643 (3.3 GHz) PRIMERGY BX920 S3: 2 × Xeon E5-2430 (2.2 GHz)
Data medium	PRIMERGY RX200 S7: 8 × MK1401GRRB PRIMERGY RX300 S7: 8 × MK1401GRRB PRIMERGY SX980 S2: 10 × ST9900805SS
Software	
BIOS	PRIMERGY RX200 S7: V4.6.5.3 R2.4.0 PRIMERGY RX300 S7: 1.18.0 PRIMERGY BX920 S3: v0625
BIOS settings	PRIMERGY RX200 S7/RX300 S7: For measurements of type “Performance”: Execute disable bit = Disabled; Frequency Floor Override = Enabled; Power Technology = Custom; Energy Performance = Performance; CPU C6 Report = Disabled; Package C State limit = C0 PRIMERGY BX920 S3: Performance/Watt=Performance; NX Memory Protection=Disabled
Operating system	PRIMERGY RX300 S7: Microsoft Windows Server 2008 R2 Enterprise PRIMERGY RX200 S7/RX300 S7/BX920 S3: Microsoft Windows Server 2012 Standard:
Controller	PRIMERGY RX300 S7: RAID Ctrl SAS 6G 5/6 1GB (D3116) Driver name: megasas2.sys, Driver version: 6.505.05.00 Firmware package: 23.2.1-0049, Firmware version: 3.152.75-1658 Controller cache: 1 GB PRIMERGY RX200 S7: PSAS CP200i Driver name: lsi_sas2.sys, Driver version: 2.055.84 Firmware version: 14.00.00.00-IT PRIMERGY BX920 S3: PY SAS HBA Mezz Card 6Gb Driver name: lsi_sas2.sys, Driver version: 2.055.84 Firmware version: 10.00
Cache settings	PRIMERGY RX300 S7: RAID Ctrl SAS 6G 5/6 1GB (D3116) Controller cache: Read-ahead, Write-back, Read Direct Disk cache: enabled
Initialization of RAID arrays	Stripe size: Storage Spaces: 256 kB (default) HW RAID: 64 kB (default), 256 kB
File system	NTFS
Measuring tool	Iometer 2006.07.27
Measurement data	Measurement file of 32 GB

Note: Some components may not be available in all countries or sales regions.

Functionality

The fundamental features of storage pools as well as the essential configuration processes are described below. Integration in a failover clustering functionality, e.g. via shared volumes, is only dealt with here to some extent.

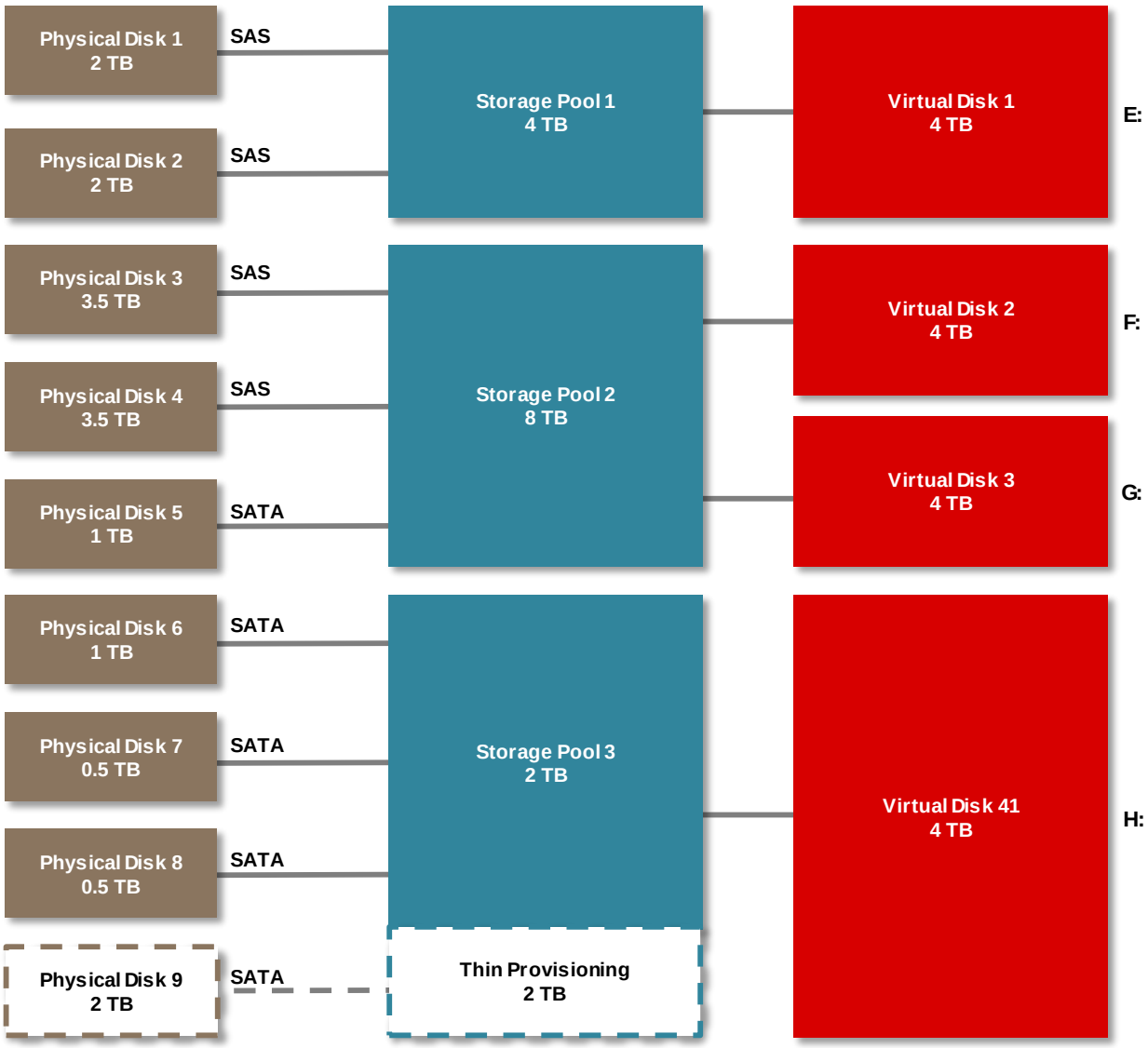
Features

Storage pools can use USB, SATA and SAS technology to group hard disks (now also of different size) into physical hard disk clusters (but only SAS in failover clustering). These pooled hard disks are then no longer available to classic data media management (Disk Manager).

Virtual disks (spaces) can now be created in the defined pool by means of Storage Spaces. These virtual disks should then be used like normal drives, e.g. through the allocation of a drive letter.

It is possible to increase and use reserved areas of both a storage pool (through adding a further hard disk) and a virtual disk (through reserved storage). However, the reserved storage feature (thin provisioning) is not supported in failover cluster mode.

The diagram below shows the structure of storage pools as an example - using the various physical hard disks and the virtual disks created from it and ultimately the allocated drives. These *virtual disks* should not be confused with the *virtual hard disks* (VHD, or VHDX) from the virtualization environment.



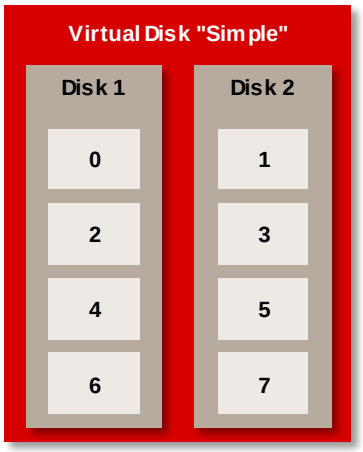
The area marked “Thin Provisioning” means in an exemplary way in this diagram that without the existence of physical hard disk 9 (2 TB) storage pool 3 is initially created with 2 TB. However, virtual disk 4 is already configured with 4 TB and the provisioning type “Thin” is set (instead of “Fixed”). If more than 2 TB of storage space is subsequently required for drive H:, it is possible to utilize the full capacity on drive H: by simply adding hard disk 9 and integrating it into storage pool 3.

The Storage Spaces (virtual disk) can be created with various storage layouts. You can select from “Simple”, “Mirror” and “Parity”.

The following table shows the meaning and general conditions of the possible settings.

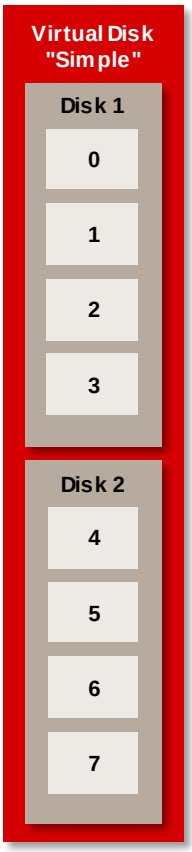
Layout	RAID comparison	RAID description	General condition	Fail-safety
Simple	RAID 0	Striping without redundancy	None	None
Mirror 2-way	RAID 1 RAID 10	Striping with simple redundancy	= 2 physical disks = n × 2 disks (n>1)	1 drive failure
Mirror 3-way	Not comparable	Striping with double redundancy	>= 5 physical disks	2 drive failures
Parity	RAID 5	Striping with parity redundancy	>= 3 physical disks	1 drive failure

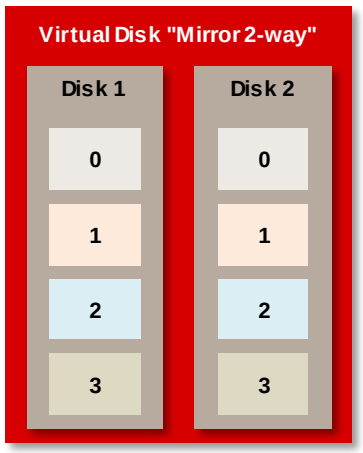
The various layouts can be clearly illustrated as follows:



Simple: The data blocks (here 0 - 7) are distributed over all the hard disks for accesses via so-called columns (picture left). The more columns (with at least one hard disk) there are, the more parallel accesses are possible. The best performance is achieved with one hard disk per column.

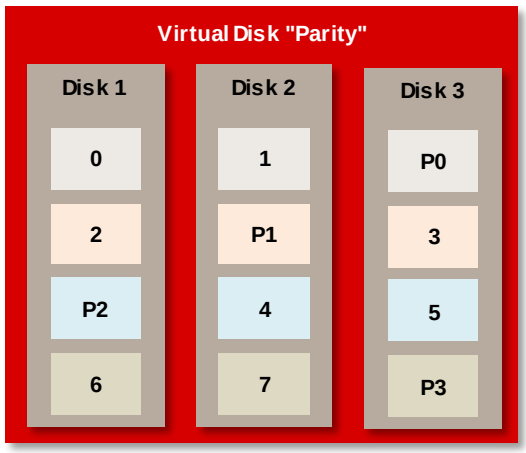
The “same” configuration with only one column would entail losses as regards performance (picture right). The columns can be configured (only PowerShell) and will among other things be considered to some extent a little later.





Mirror 2-way: The data blocks (here 0 - 3) are duplicated (picture left), thus of course also reducing the capacity by half.

Mirror 3-way: Here the usable capacity is only one third of the available hard disk capacity, but with two data copies, which means that even two drives can fail.



Parity: The data blocks (here 0 - 7) with parity information (here P1 - P3) are distributed over all the hard disks (picture left). This option offers a good compromise between speed and fail-safety.

PowerShell

As regards the overview, here are some details about the standard values in relation to the storage layout. Some parameters can only be changed from the default via the PowerShell and cannot be modified with the GUI of the Server Manager.

The PowerShell command

Get-VirtualDisk

can be used to show various settings. The command

New-VirtualDisk

can be used e.g. to create and parameterize new virtual disks.

Columns

The standard number of the above mentioned columns is one per hard disk for the “Simple” layout. This applies for up to eight hard disks, above that the value remains fixed or would have to be modified via the PowerShell.

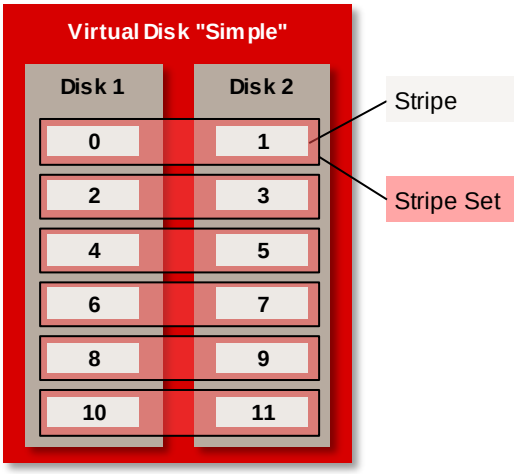
In the case of the “Parity” layout three columns are configured as a default setting and the “Mirror” default setting follows from the number of data copies. In other words, Mirror 2-way: *number of disks divided by 2*.

Stripe size

The default stripe size (interleave) is 256 kB for all storage layouts.

Stripes (chunks) are sub-units of a stripe set and stripe size refers to the size of a data block that is saved on a disk. A stripe set is composed of one stripe for each data medium of a RAID array.

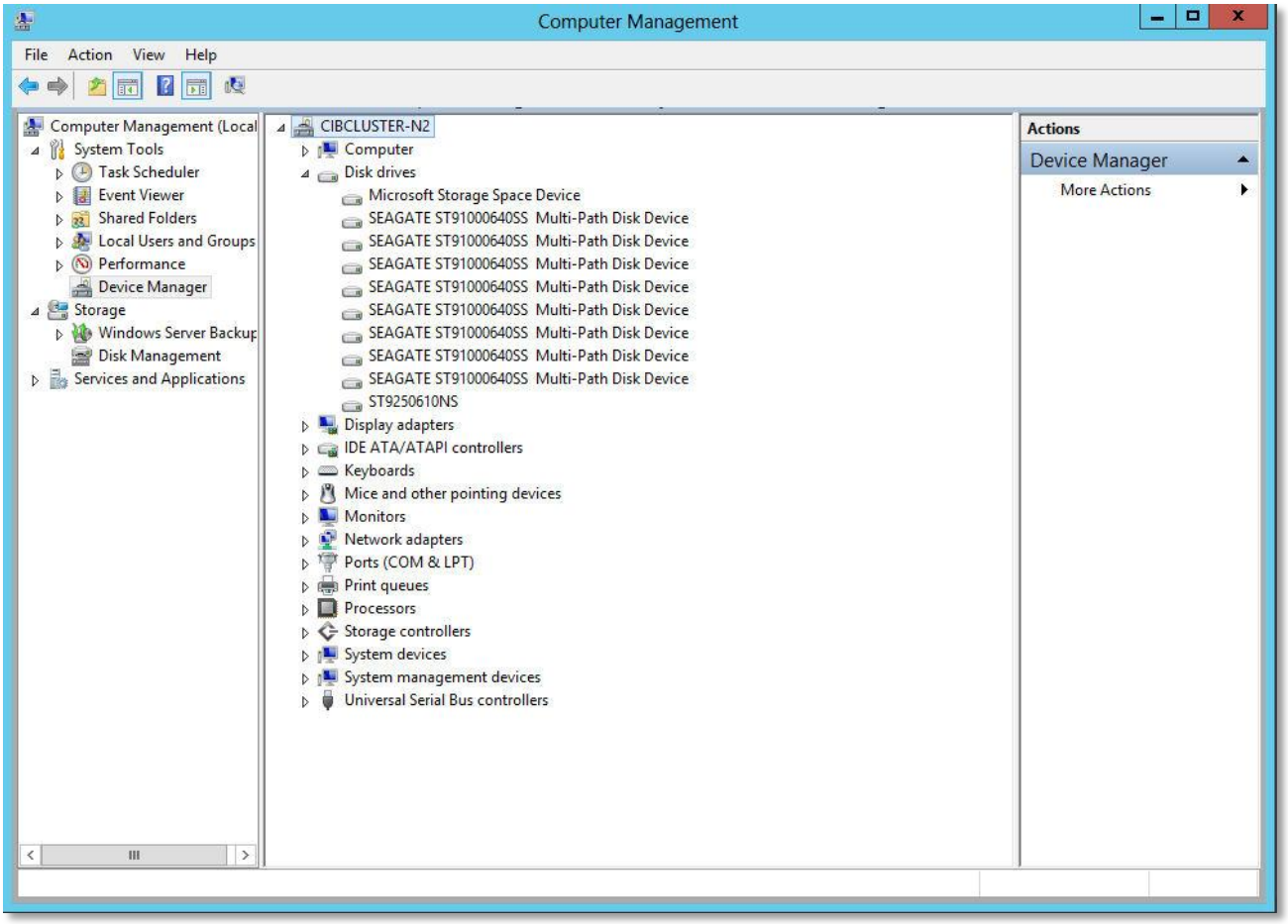
The diagram shows a RAID 0 array or a “Simple” Storage Spaces layout with two hard disks (disk 1 and disk 2) and the corresponding logical distribution of the data blocks.



Cache settings

There are also a series of cache settings, which are shown here. In addition to the disk cache, which is relevant to the physical hard disk, there are caches for the virtual disk, which is mapped via the Microsoft Storage Space Device.

The picture shows how the disk drives are presented in the Device Manager:



The settings can now be made via <Properties> and <Policies>.

The impact of the individual caches is described in more detail in various test scenarios in the later sections on the topic of performance.

Settings for the Microsoft Storage Space Device ("Enable write caching on the device" cannot be disabled here).

Setting for the single hard disk.

Summary:

The main difference between classic data media management (Disk Manager) and the storage pool/spaces is the significant flexibility and scaling option of the new features as regards the physical data media, combined with optimized storage space management and use via completely abstracted storage mapping.

(see:

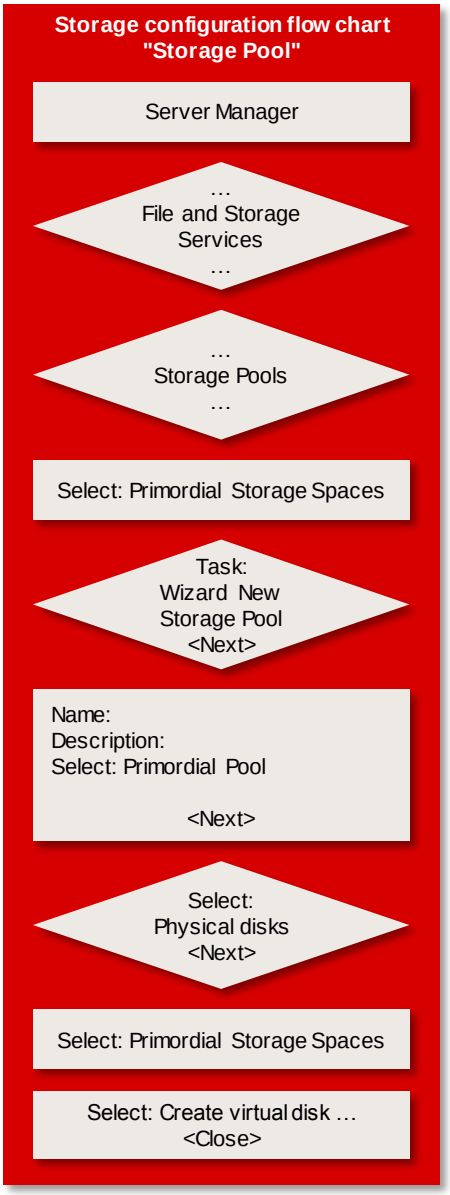
http://www.windowsnetworking.com/articles-tutorials/windows-server-2012/Getting-Used-Using-Windows-Storage-Spaces-Part1.html)

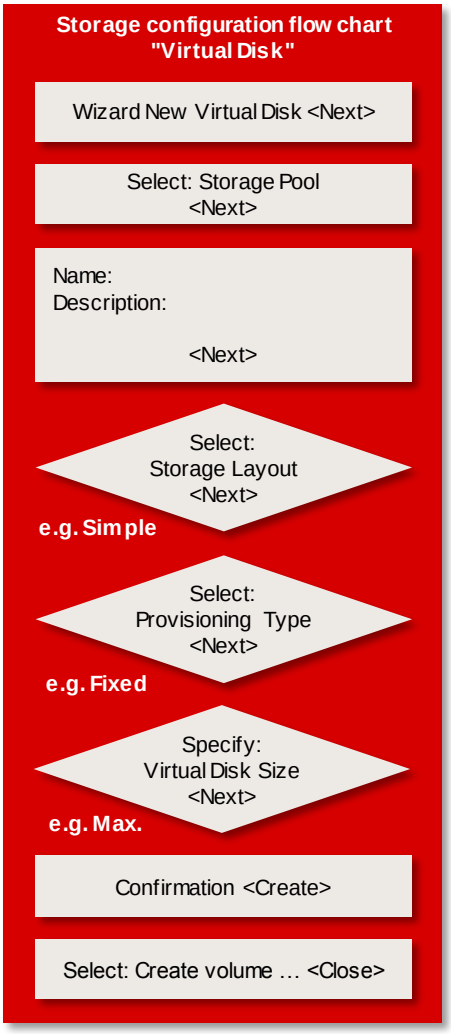
Configurations

By selecting File and Storage Services under Windows Server Manager you can easily configure the hard disk storage.

The previously unallocated hard disks are displayed in the so-called Primordial Storage Space in the Storage Pool section. The wizards New Storage Pool, New Virtual Disk and New Volume can be used to configure the data media, data area and storage layouts.

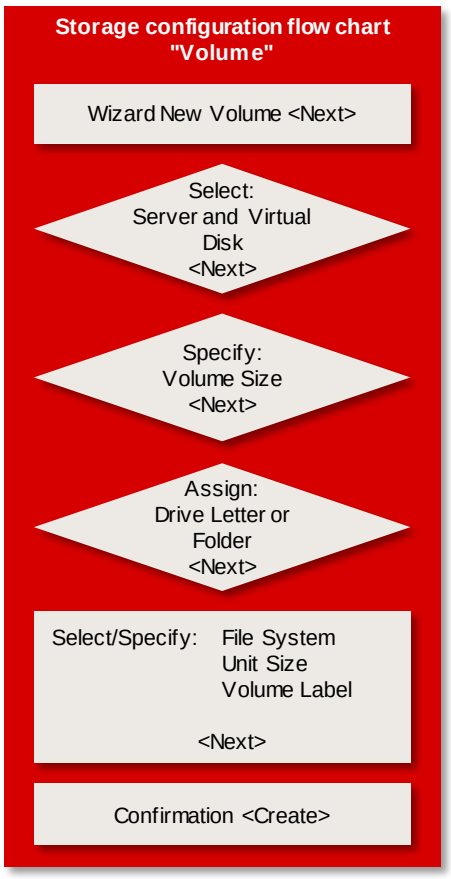
The following sequence schematically depicts the fundamental procedure:





According to this process a configured virtual disk is shown in the Server Manager. We can also see this disk in the Disk Manager, but as the corresponding unallocated basic disk, i.e. not yet operational.

We can now use the New Volume wizard to complete the configuration according to the following process:



The drive is now listed in the Server Manager, in the Disk Manager and also in the File Explorer as an operational hard disk drive.

Note:

In addition to the number of hard disks for a pool, the decisive steps of the selection process are the specifications for the storage layout (Simple, Mirror, Parity) and for the provisioning (Thin or Fixed).

Storage Spaces in the cluster

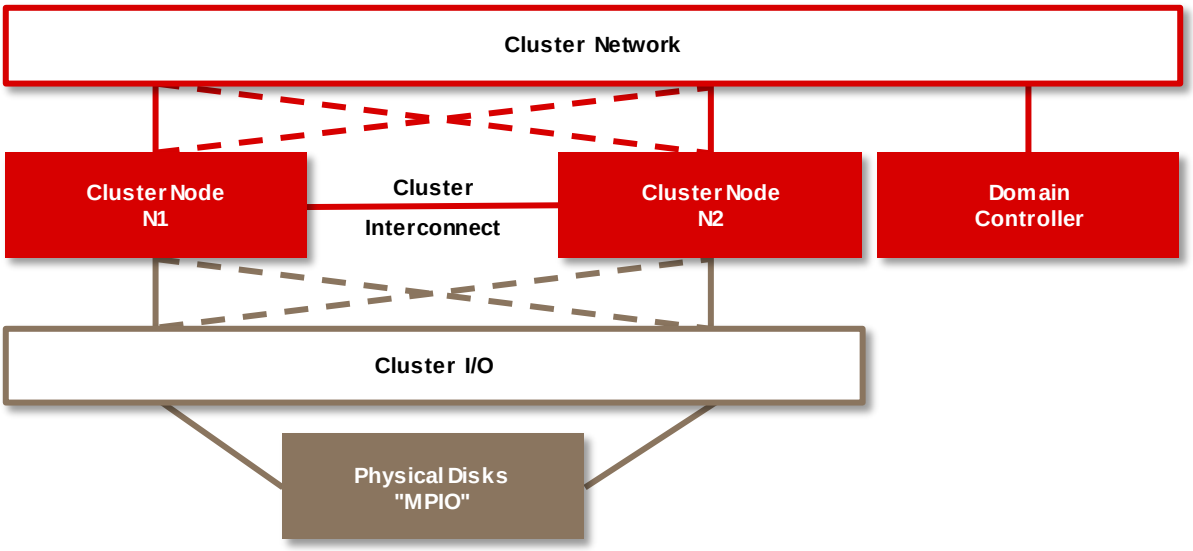
Clusters play a central role in the provision of error-tolerant systems. Further developed cluster management is provided with Windows Server 2012. The following features are always covered with this cluster:

- High availability and scaling
- Use of cluster shared volumes (CSVs)
- Cluster servers (nodes) on physical servers or virtual machines
- Failover-secured services and roles
- Service and maintenance optimization
- Controlling and monitoring
- GUI and PowerShell-based functionality

Suitable hardware and software configurations, which ensure fail-safety, e.g. redundant communication / access components (LAN, I/O, etc.) and dual-ported storage media, also known as multi-path IO (MPIO), are prerequisite for the cluster. These components are subject to a consistency and functional check in the cluster management of Microsoft Windows Server 2012.

Note: In addition to the cluster with two nodes as described here, a failover strategy with a 1-n node cluster can entail other hardware requirements.

The following diagram shows the fundamental structure of a failover cluster:

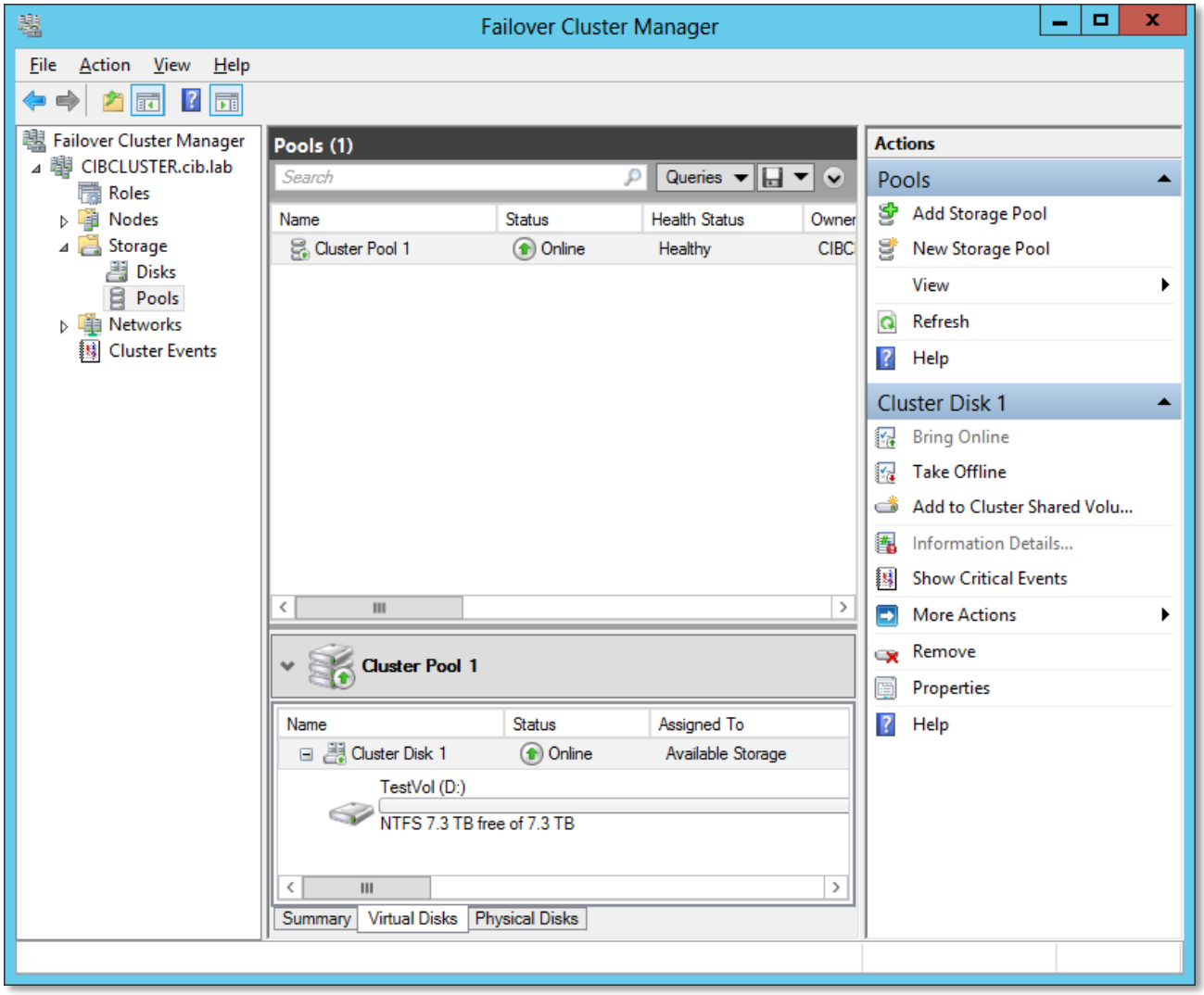


The use of cluster shared volumes is now a partial aspect in the cluster. The Failover Cluster Manager, as the central management unit, is based for storage mapping on storage pools and Storage Spaces and also offers additional functionalities.

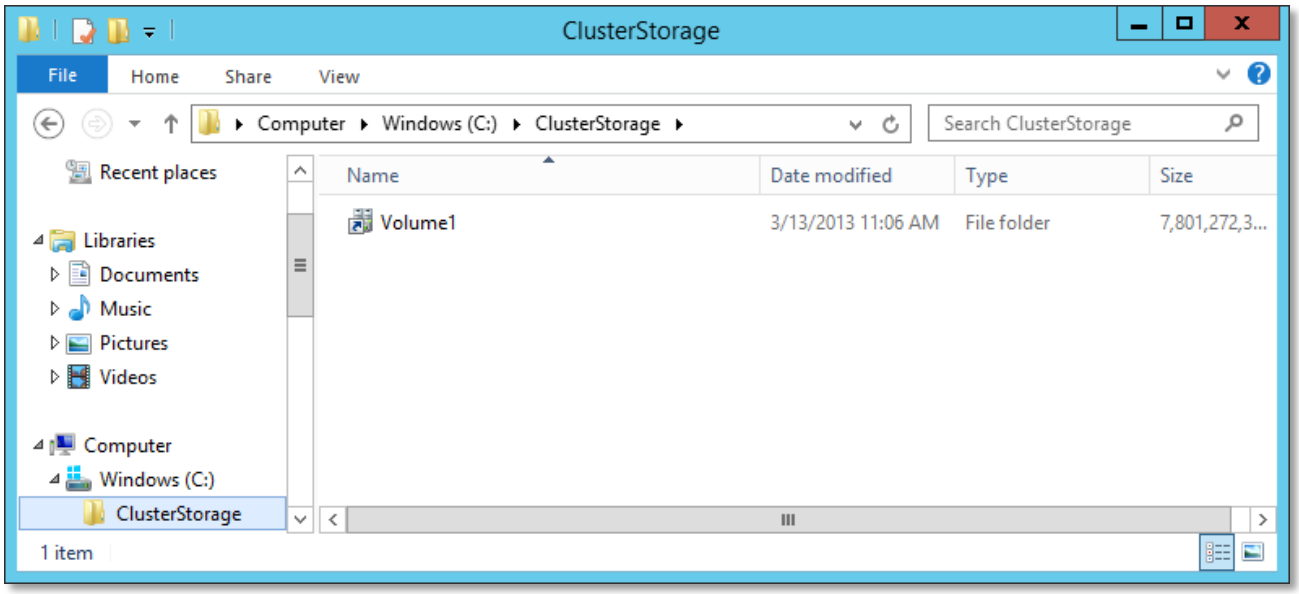
Cluster shared volumes allow several nodes to simultaneously access the same storage areas (but only NTFS). Thus in the event of an error, defined roles in the cluster can be assumed very quickly and without any further interaction between the nodes. CSV also supports simple management of a large number of storage mappings in the cluster.

The configuration corresponds to the process as described before, but now via the Failover Cluster Manager, which can also configure the CSVs. After a drive has been configured, it is initially available to one of the nodes in turn in the Server Manager, in the Disk Manager and also in the File Explorer as a listed, operational hard disk drive.

It is of course also depicted in the Failover Cluster Manager as “Available Storage”, see the following picture.



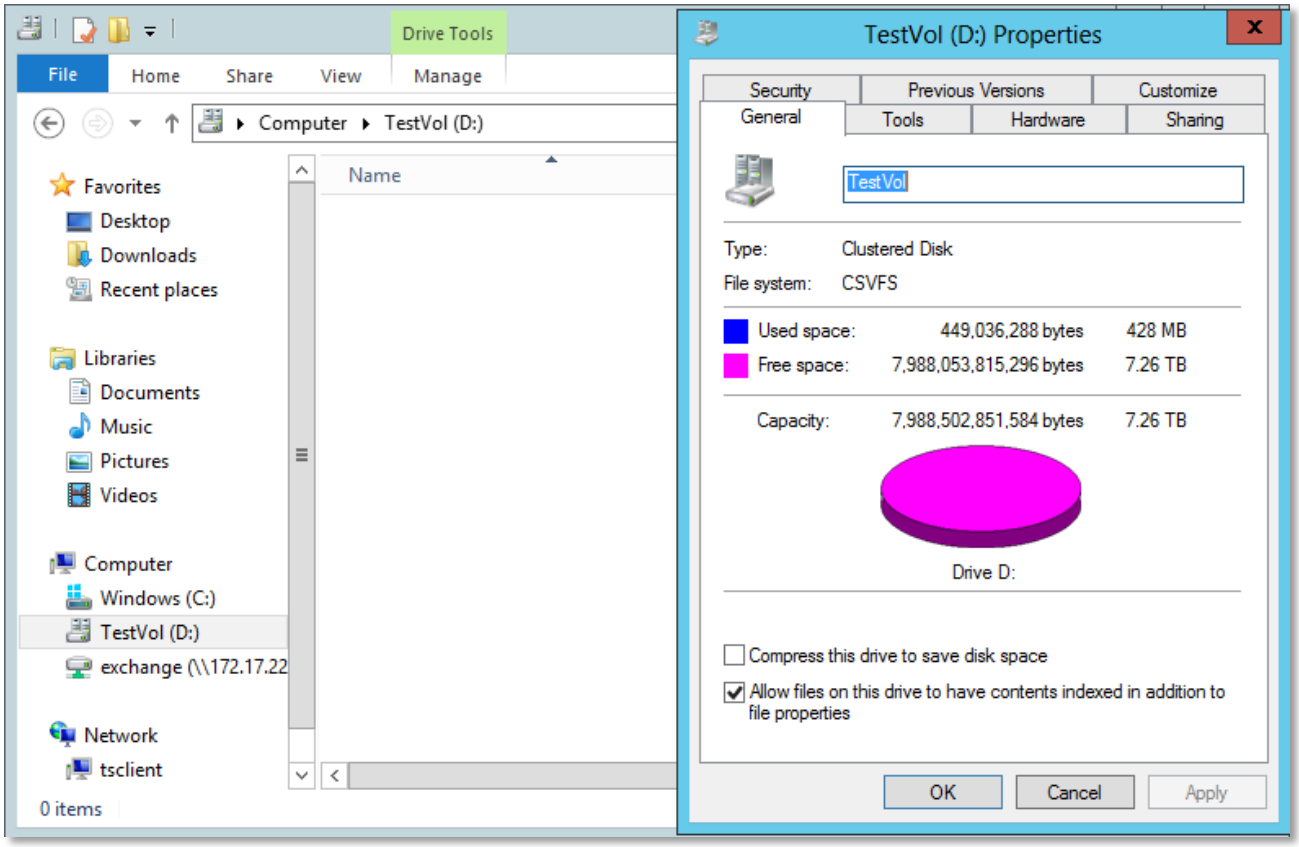
The created drive, which to begin with should only be operated on the assigned node, can be put into the "ClusterStorage". The action "Add to Cluster Shared Volumes" is used to mount the drive under **Windows(C:)\ClusterStorage** as a volume.



This volume is now available in the cluster to the nodes. The volume can be edited in the File Explorer like a directory and appears in the Disk Manager and in the Server Manager as a reserved device.

Depending on requirements, it is now also possible to assign drive letters on the node, to which the CSV is allocated. The volume then reappears as a separate drive, e.g. in the File Explorer. However, the allocation of a drive letter is only possible in the Disk Manager (direct drive).

In order to assign drive letters on further nodes it is necessary for the CSV to be moved to the appropriate node and then assigned again via the Disk Manager. This procedure is e.g. needed in order to have the test program Iometer (as an example for appropriate applications) simultaneously measure the same storage area (in this case drives) from several nodes.



Further options of accessing a number of disks from various nodes at the same time can be achieved by means of the file server role or with VHDs (virtual hard disks) from the virtualization environment (Hyper-V).

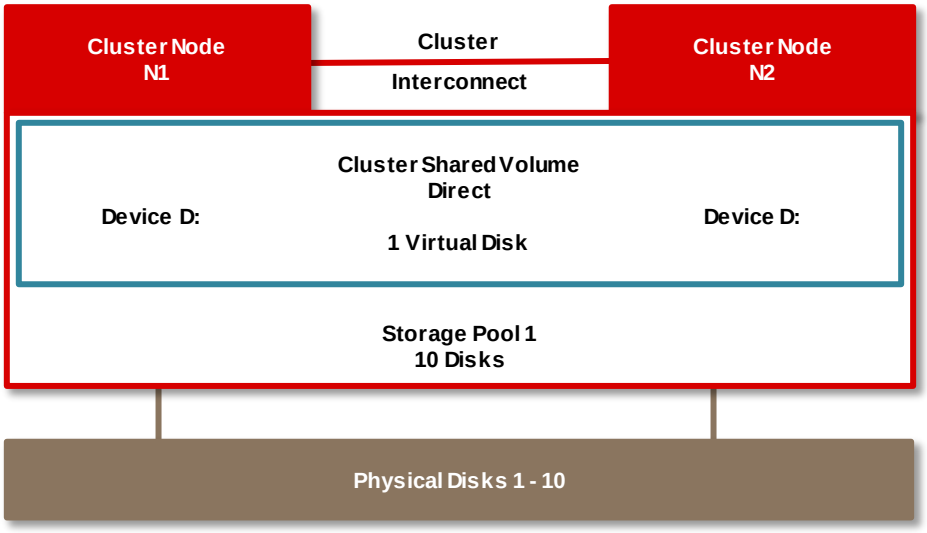
In the first case, e.g. two virtual disks (as initially described) are created via a disk pool and identified with drive letters. A storage area is now assigned via the file server role from one node in each case. Both nodes thus access the same disks, but different storage areas. On the other hand, it is in the case of shared volume exactly the same storage area in the disk pool used.

As described, a virtual disk is moved to the CSV with the “VHD method”. The Disk Manager is now used to create a VHD in the volume from each node (e.g. Windows(C:)\ClusterStorage\Volume\DiskN1.vhd). Using the Disk Manager this “hard disk” can now also be configured in each node as a drive.

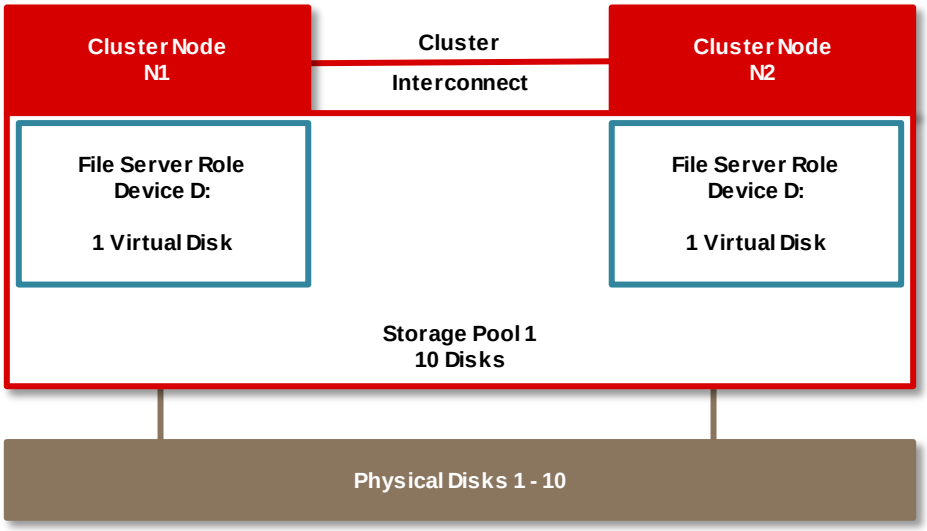
The different configuration options will be dealt with again during the course of the performance considerations.

Here is a graphic overview to illustrate the various drive assignments:

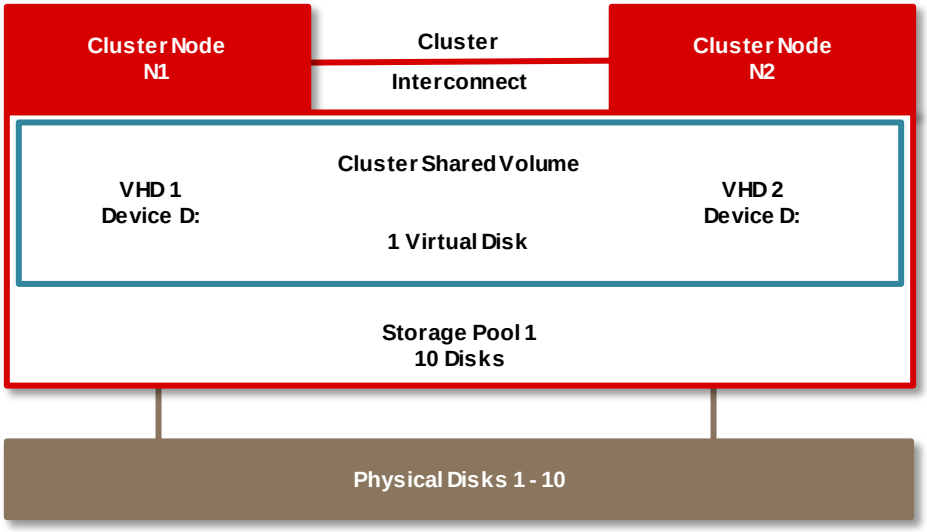
In the first case, a virtual disk is moved to the “ClusterStorage” and addressed with a direct drive assignment from both nodes. The nodes thus work simultaneously on one storage area for all the hard disks of the storage pool.



In the second case, two virtual disks are created and accessed via the configuration of a file server role and the appropriate drive assignment. The cluster shared volume is not used here. Thus the same storage pool, but different storage areas are used.



In the third and last case, a virtual disk is again first moved to the cluster storage. However, drive assignment is now achieved by creating in each case one VHD in the cluster storage from every node. The same storage area in the storage pool is also used in this case.



Performance

Various performance considerations have been made in connection with Storage Spaces. In the Storage Spaces and Shared Volume section the listed configuration options for CSV are also compared under performance features. The deployed measuring tool Iometer with the standard measurement method and the measurement environment have been used as described in the introduction.

Storage Spaces

First of all, various measurements are performed for the described storage layouts of the Storage Spaces. In particular the scaling when the number of hard disks is doubled, and also the use of the write cache receive special consideration.

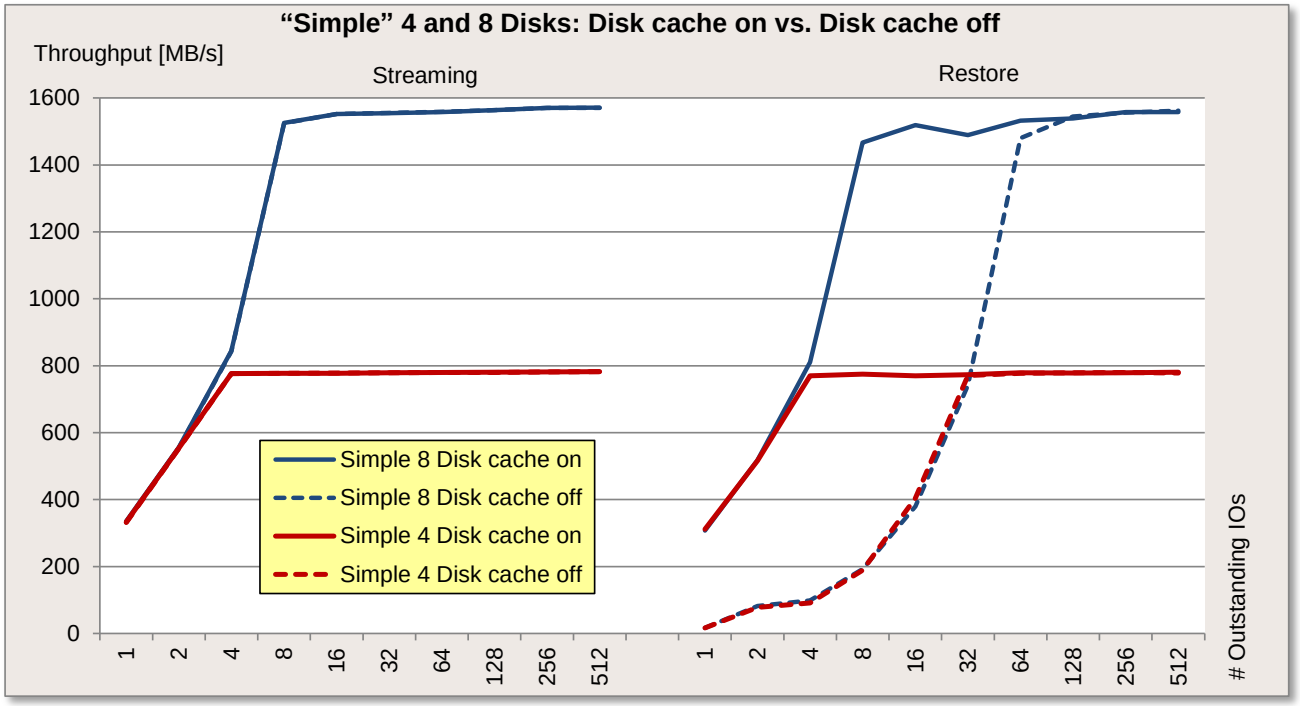
The access/response time (latency [ms]) in this environment is not considered further, reference is only made to the appropriate document under Literature ("Basics of Disk I/O Performance"). Summed up in one sentence:

Latency depends on the transaction rate (I/O operations per second) and on parallelism when carrying out the transactions.

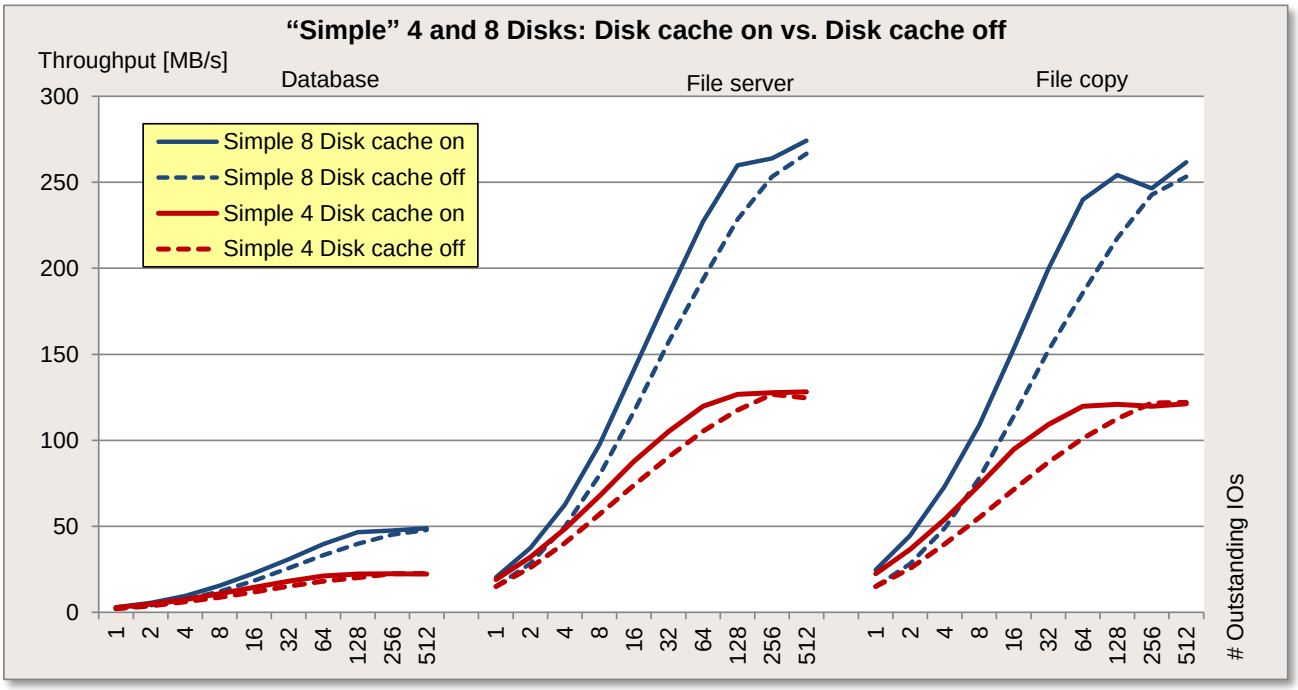
"Simple"

The first measuring configuration describes the storage layout "Simple" with four and eight hard disks. The write cache of the hard disks is enabled (Disk cache on), as is the write cache of the Microsoft Storage Space Device (Storage cache on), which is permanently enabled for all measurements. In the case of a comparative measurement the write cache of the hard disks was then disabled (Disk cache off).

The first diagram initially shows the measurement results of the standard load profiles "Streaming" (sequential access, 100% read, 64 kB block size) and "Restore" (sequential access, 100% write, 64 kB block size).



The next diagram shows the course of the standard load profiles “Database” (random access, 67% read, 8 kB block size), “File server” (random access, 67% read, 64 kB block size) and “File copy” (random access, 50% read, 64 kB block size).



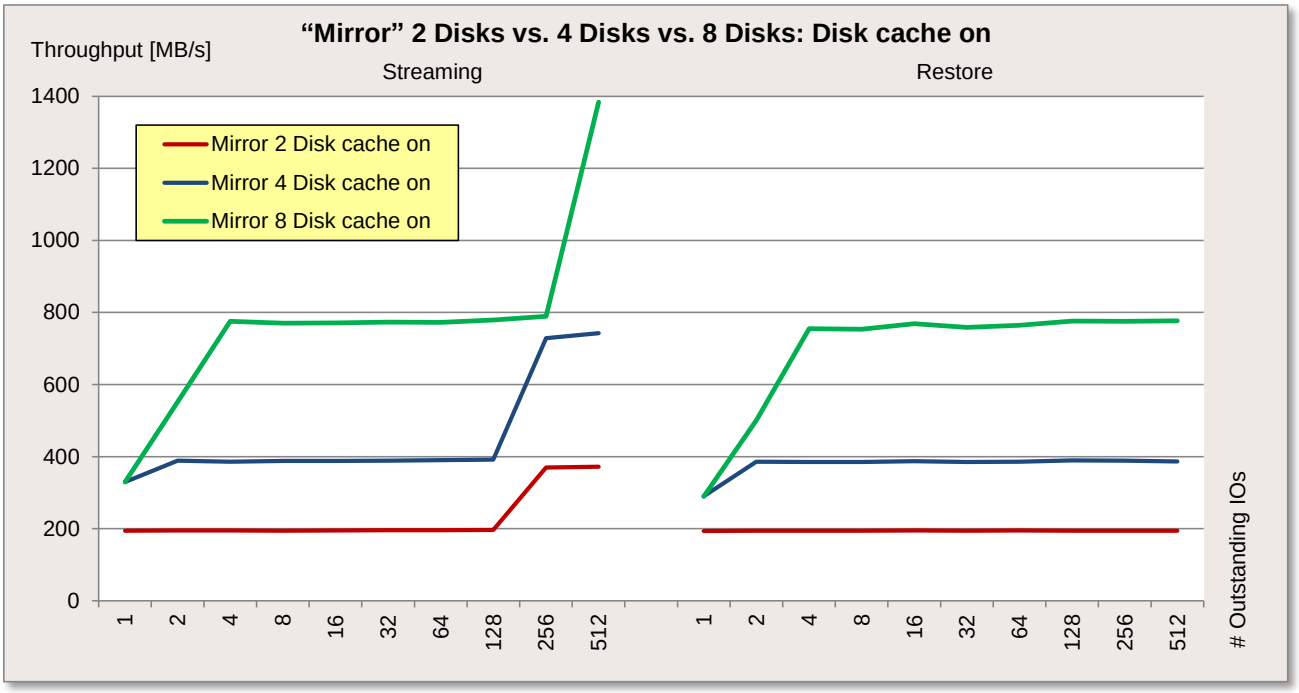
It becomes clear in both diagrams that related to the write share the write cache has a positive impact, particularly in the case of medium load intensity. Furthermore, throughput scales by factor 2 when the number of hard disks is doubled.

However, it should not be forgotten that the use of the write cache must be accompanied with precautions against data loss in the event of a power failure, e.g. by using an uninterruptible power supply (UPS). At this juncture it should be added that if you enable the “Turn off Windows write cache buffer flushing on the device” setting of the hard disks (see cache settings), the result is a further slight gain in throughput.

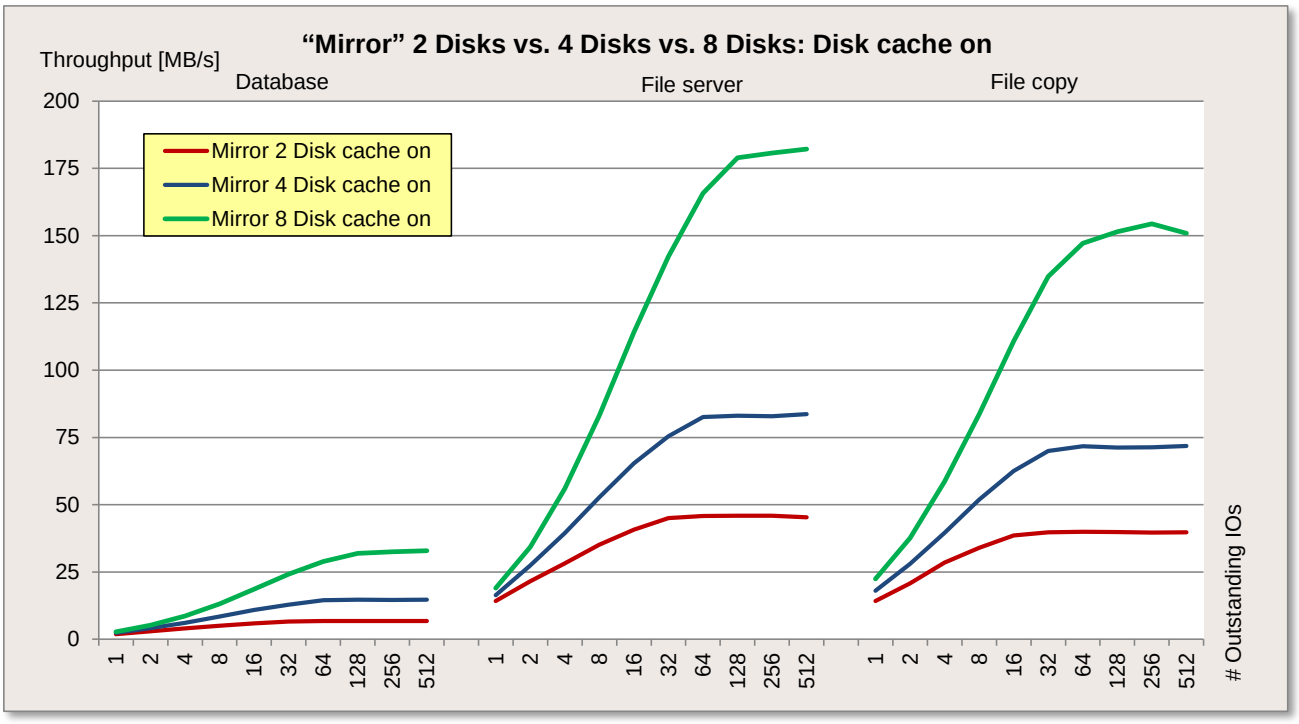
"Mirror 2-way"

The second measuring configuration describes the storage layout "Mirror 2-way" with two, four and eight hard disks. The write cache of the hard disks is enabled (Disk cache on), as is the write cache of the Microsoft Storage Space Device (Storage cache on).

The first diagram initially shows the measurement results of the standard load profiles "Streaming" (sequential access, 100% read, 64 kB block size) and "Restore" (sequential access, 100% write, 64 kB block size).



The next diagram shows the course of the standard load profiles “Database” (random access, 67% read, 8 kB block size), “File server” (random access, 67% read, 64 kB block size) and “File copy” (random access, 50% read, 64 kB block size).



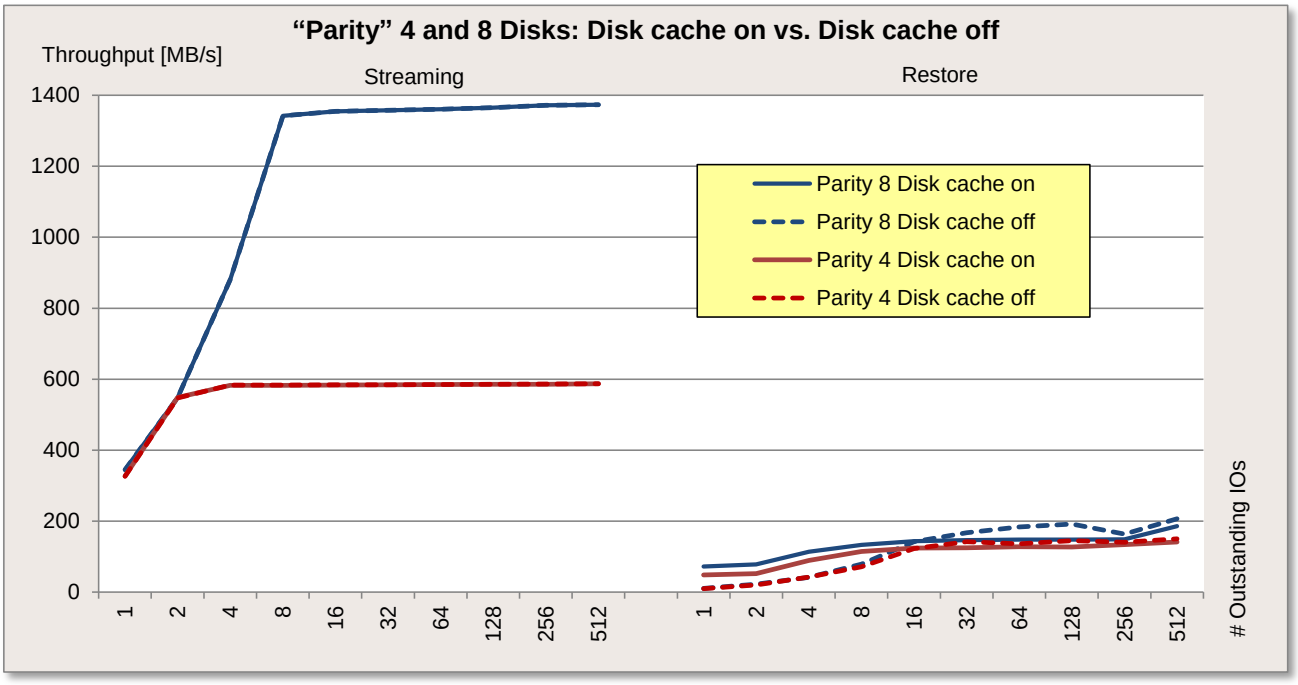
You can really see the scaling due to the doubling of the number of hard disks. In the case of the two sequential load profiles up to factor 2, otherwise with a somewhat different characteristic between factor 1.2 and 2.2.

As with the “Simple” configuration, the result without the write cache of the hard disks (Disk cache off) is an accordingly lower throughput, which for the sake of clarity is not specifically presented here.

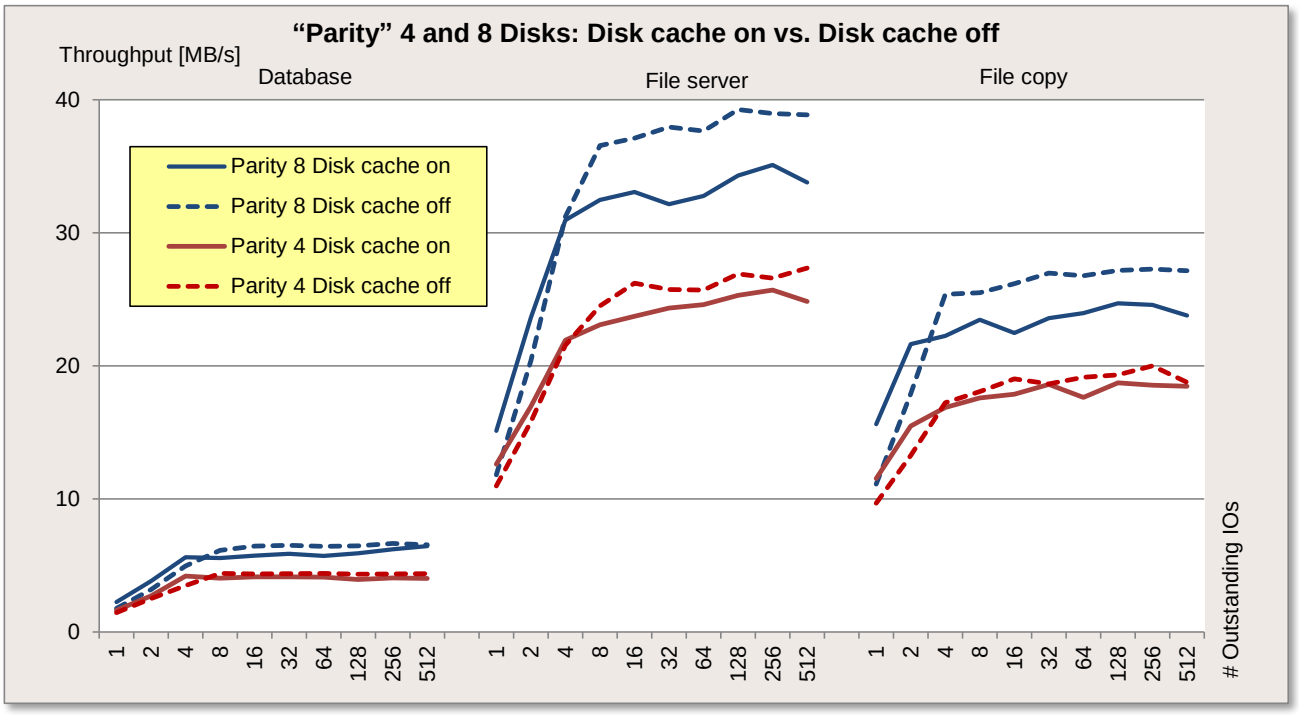
"Parity"

The third measuring configuration describes the storage layout "Parity" with four and eight hard disks. The write cache of the hard disks is enabled (Disk cache on), as is the write cache of the Microsoft Storage Space Device (Storage cache on). In the case of a comparative measurement the write cache of the hard disk was then disabled (Disk cache off).

The first diagram initially shows the measurement results of the standard load profiles "Streaming" (sequential access, 100% read, 64 kB block size) and "Restore" (sequential access, 100% write, 64 kB block size).



The next diagram shows the course of the standard load profiles “Database” (random access, 67% read, 8 kB block size), “File server” (random access, 67% read, 64 kB block size) and finally “File copy” (random access, 50% read, 64 kB block size).



This storage layout does not show any disadvantages with pure read accesses. However, the result for “Parity” is unsatisfactory for write applications; so the value is too low for “Restore” and the random load profiles. The write cache even has a negative impact in part.

The generally lower write performance of the “Parity” variant results from a “journaling procedure”, with a correspondingly additional overhead. This storage layout is therefore only recommended for predominantly read application scenarios.

“Mirror 3-way”

The storage layout “Mirror 3-way” is not considered in this version of the white paper. Here is only an overview of the usable disk capacity with different numbers of hard disks, which were put together as an example for the hard disk type used in this white paper.

Mirror 3-way Layout (1 physical disk = 147 GB)		
#Disks	Storage Pool GB	Max. Virtual Disk GB
5	679	225
6	815	270
7	950	314
8	1060	360

Storage Spaces vs. HW RAID (default stripe size)

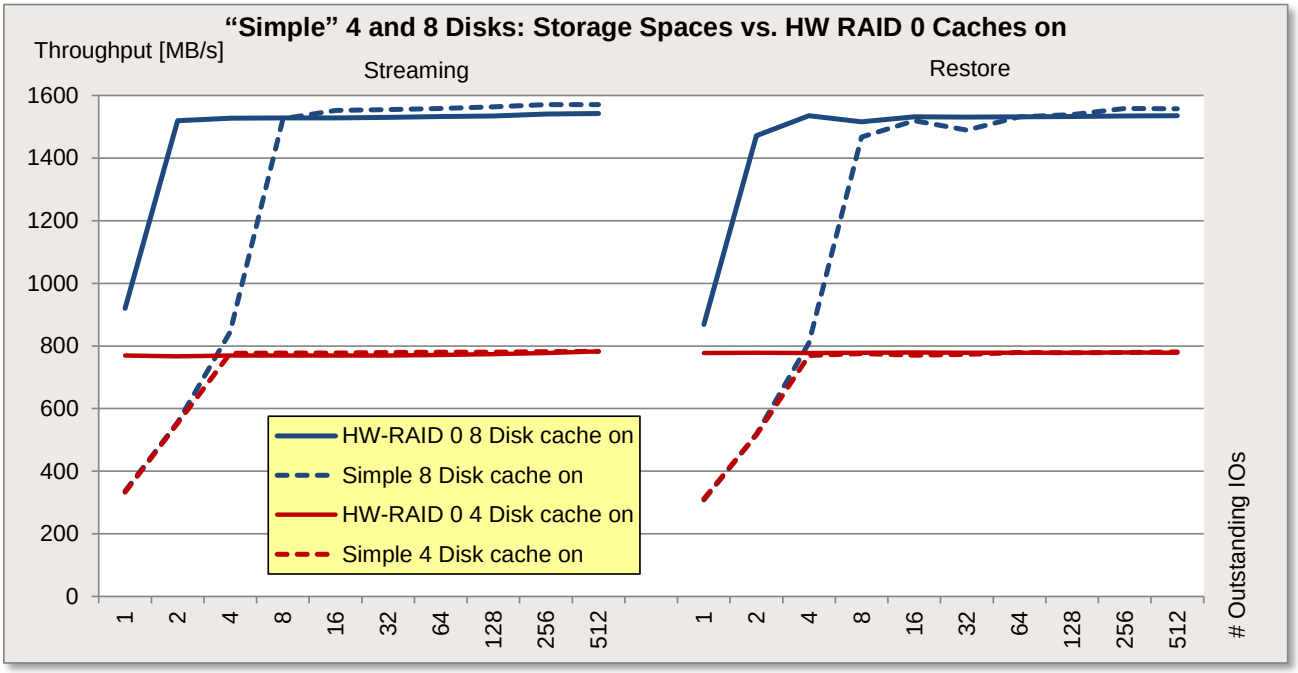
The comparison of a classic hardware RAID with the Storage Spaces function is of particular interest. The storage layouts “Simple/Mirror/Parity” with the SAS controller “PSAS CP200i” are compared with the appropriate HW RAID levels of a “RAID Ctrl SAS 6G 5/6 1GB (D3116)” controller. The behavior of the appropriate write caches is also taken into account here.

Note:
However, the following results should be put into perspective, as the standard stripe size (interleave) was used in each case for the measurements. The Storage Spaces have a value of 256 kB and HW RAID has a value of 64 kB - this is a difference that has an impact on throughput. This is why there is an overview of all the storage layouts compared with the appropriate HW RAID in the next section, where the stripe size 256 kB was consistently used as an example for the measurements.

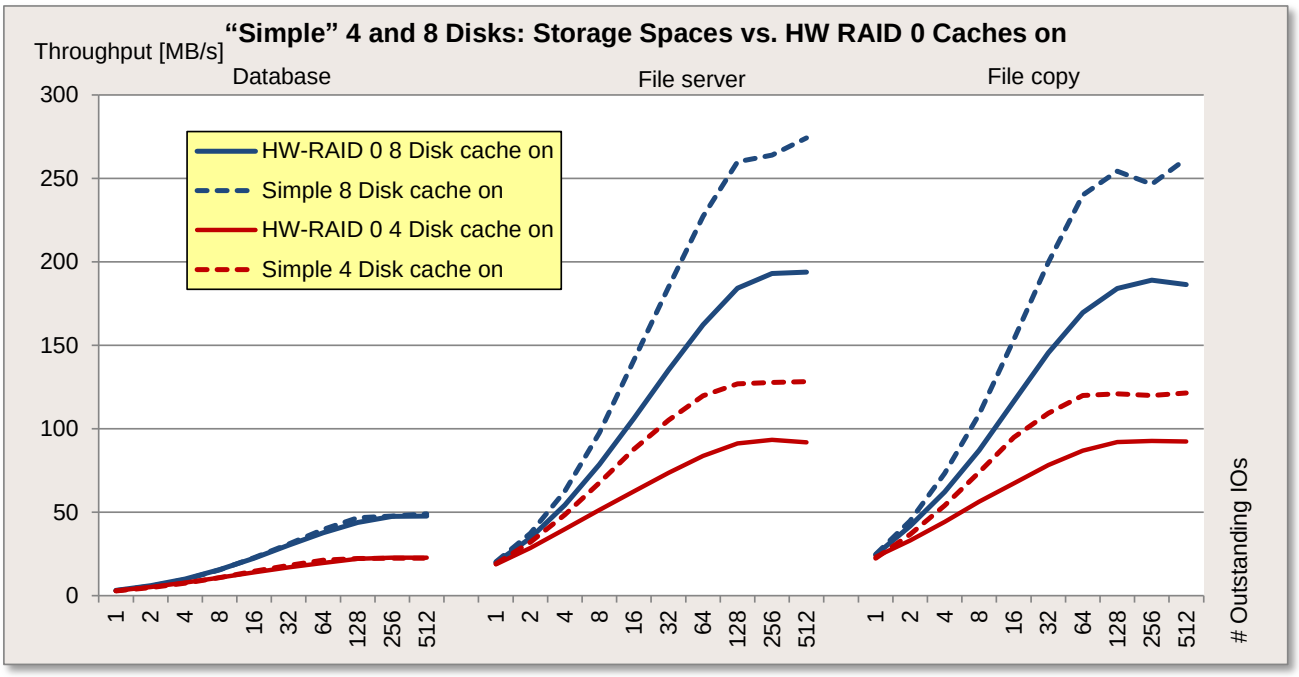
“Simple” vs. HW RAID 0

The first measuring configuration describes the storage layout “Simple” with four and eight hard disks compared with a HW RAID 0 configuration. The write cache of the hard disks is enabled (Disk cache on).

The first diagram initially shows the measurement results of the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).



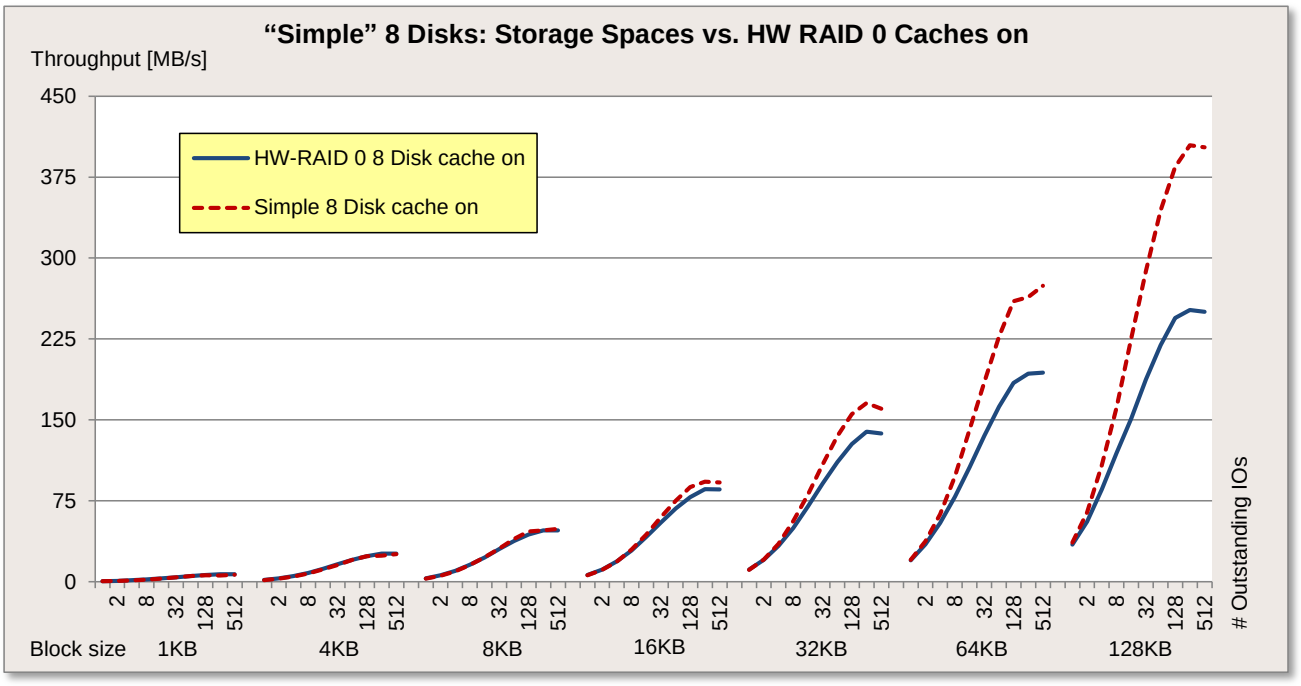
The next diagram shows the course of the standard load profiles “Database” (random access, 67% read, 8 kB block size), “File server” (random access, 67% read, 64 kB block size) and finally “File copy” (random access, 50% read, 64 kB block size).



The two diagrams above show that when comparing “Simple” with HW RAID 0 with the load profile used and with the respective default stripe size the Storage Spaces have slight advantages. When you double the number of hard disks, the scaling behaves in a comparable way in both measurement environments.

The advantages can also be seen in this configuration if you not only compare (as above) the selected standard load profile, but possible further load profiles of server applications.

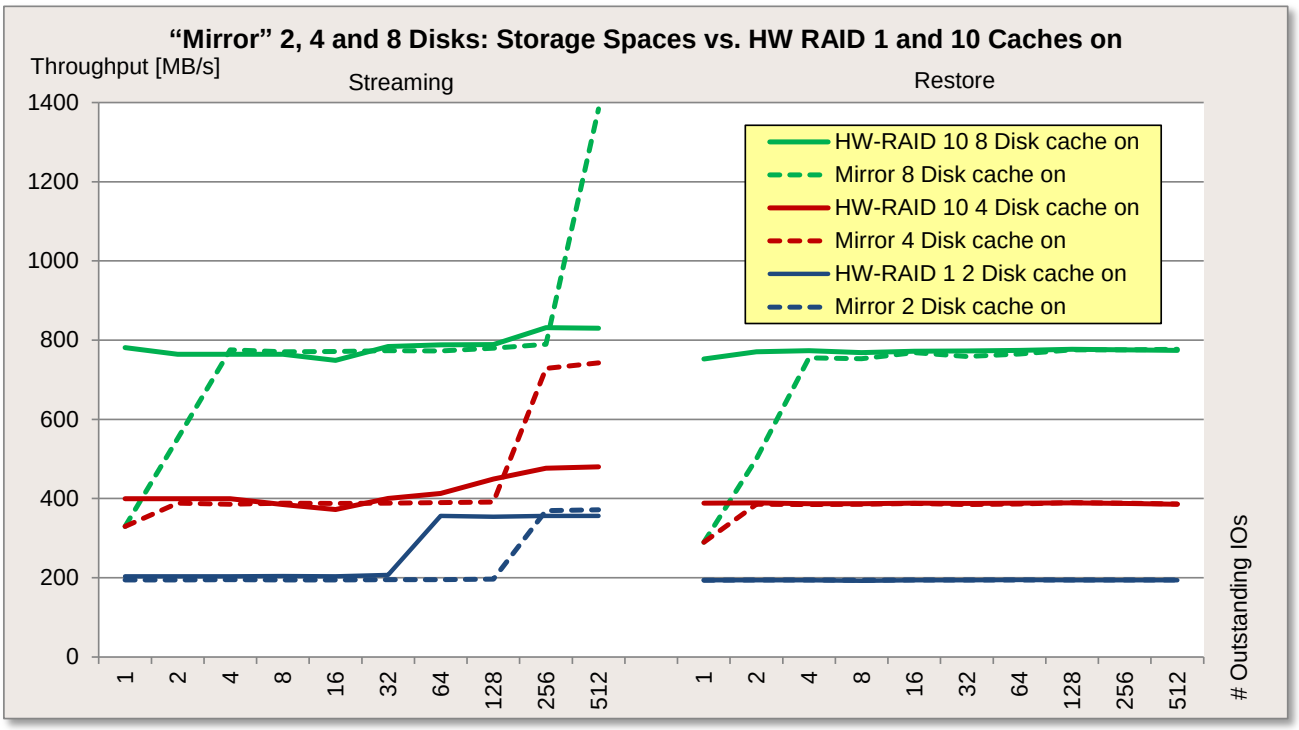
Diagram of a load profile of server applications with random access and 67% read with different block sizes.



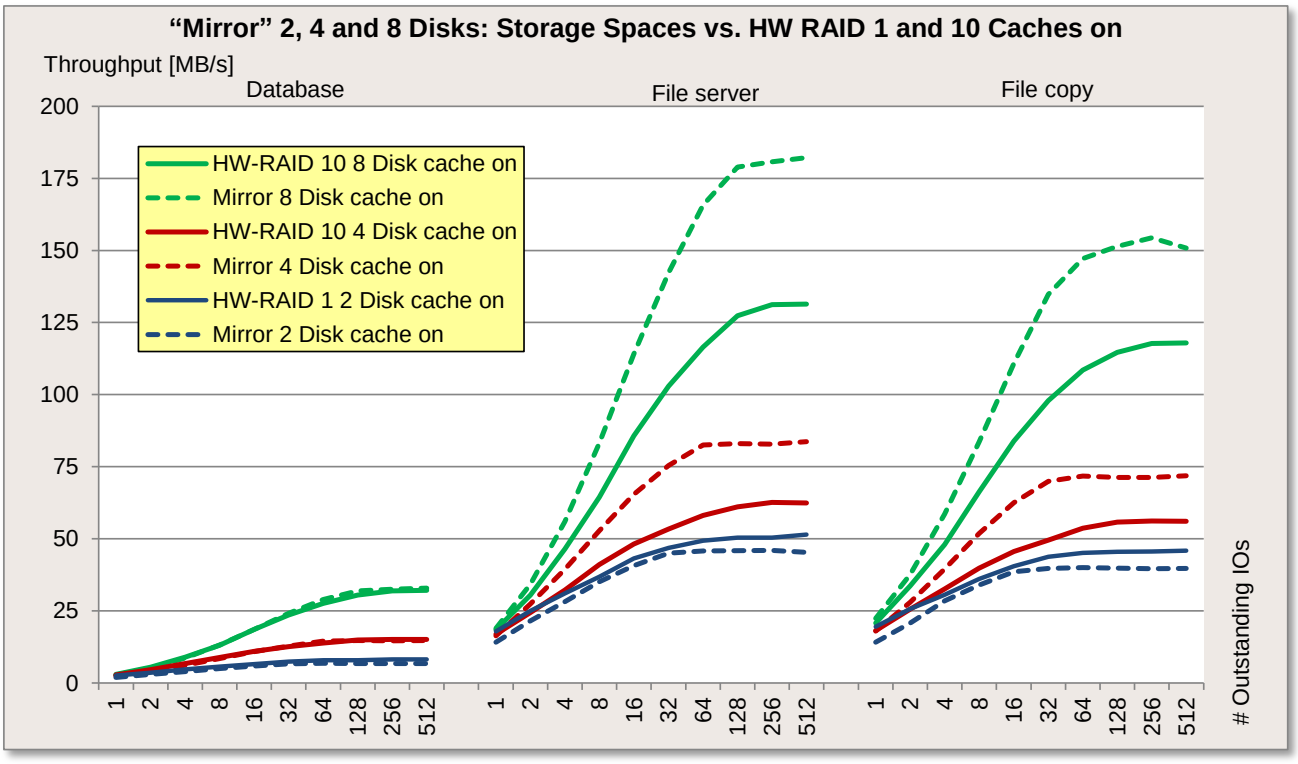
"Mirror" vs. HW RAID 1/10

The second measuring configuration describes the storage layout "Mirror 2-way" with two, four and eight hard disks compared with a HW RAID 1 and HW RAID 10 configuration. The write cache of the hard disks is enabled (Disk cache on).

The first diagram initially shows the measurement results of the standard load profiles "Streaming" (sequential access, 100% read, 64 kB block size) and "Restore" (sequential access, 100% write, 64 kB block size).



The next diagram shows the course of the standard load profiles “Database” (random access, 67% read, 8 kB block size), “File server” (random access, 67% read, 64 kB block size) and finally “File copy” (random access, 50% read, 64 kB block size).

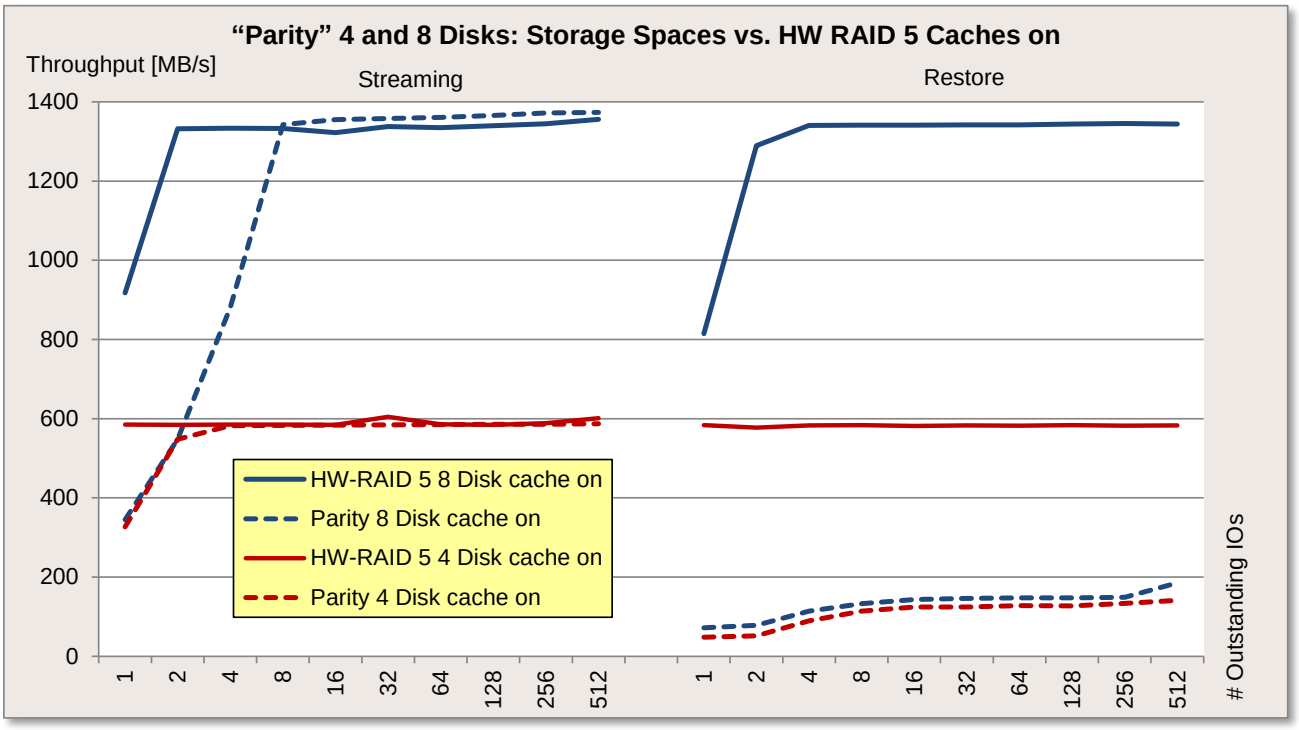


And also when comparing “Mirror” with HW RAID 1 and HW RAID 10, we see slight advantages for the Storage Spaces in the defined measurement environment. The same general conditions apply as with the previous “Simple” – RAID 0 comparison and the scaling for the increasing number of hard disks also meets expectations.

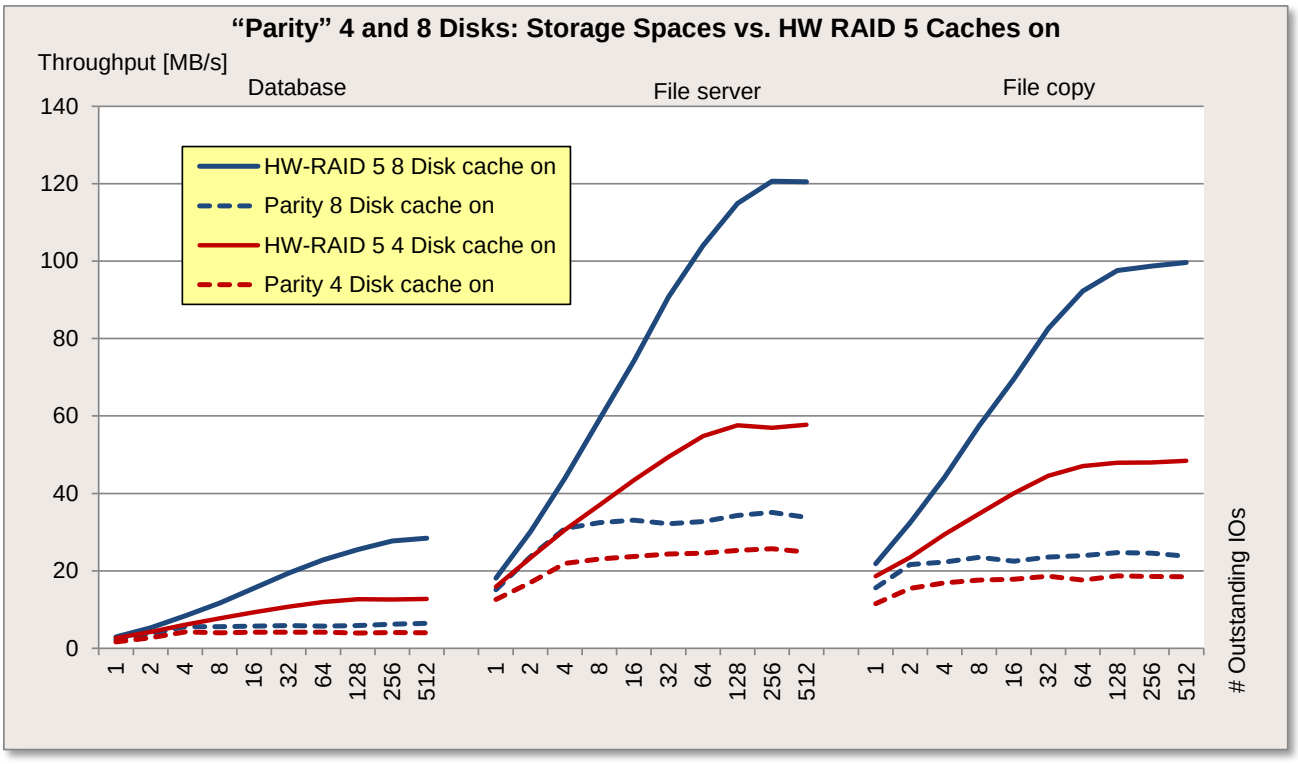
"Parity" vs. HW RAID 5

The last measuring configuration describes the storage layout "Parity" with four and eight hard disks compared with a HW RAID 5 configuration. The write cache of the hard disks is enabled (Disk cache on).

The first diagram initially shows the measurement results of the standard load profiles "Streaming" (sequential access, 100% read, 64 kB block size) and "Restore" (sequential access, 100% write, 64 kB block size).



The next diagram shows the course of the standard load profiles “Database” (random access, 67% read, 8 kB block size), “File server” (random access, 67% read, 64 kB block size) and finally “File copy” (random access, 50% read, 64 kB block size).



In the comparison of “Parity” with HW RAID 5 the Storage Spaces show considerable losses, especially with sequential write and the load profiles with random access. This behavior of the Storage Spaces has already been explained before. A HW RAID benefits particularly in the case of a RAID 5 thanks to the cache of the RAID controller, which is essential for RAID arrays with parity.

However, the “Parity” variant is equivalent for almost exclusively sequential read applications.

Storage Spaces vs. HW RAID (256 kB stripe size)

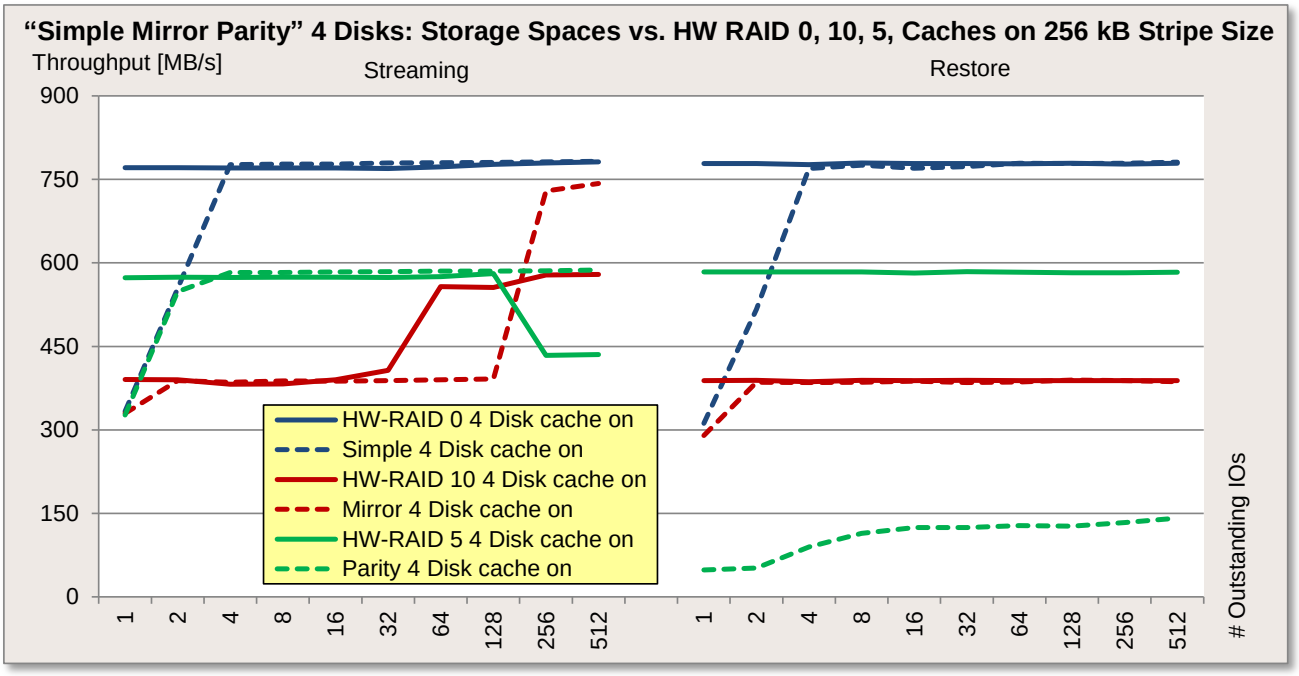
As a comparison there is an overview of all the storage layouts compared with the appropriate HW RAID, where the stripe size 256 kB was consistently used as an example for the measurements.

"Simple Mirror Parity" vs. HW RAID 0 / 10 / 5

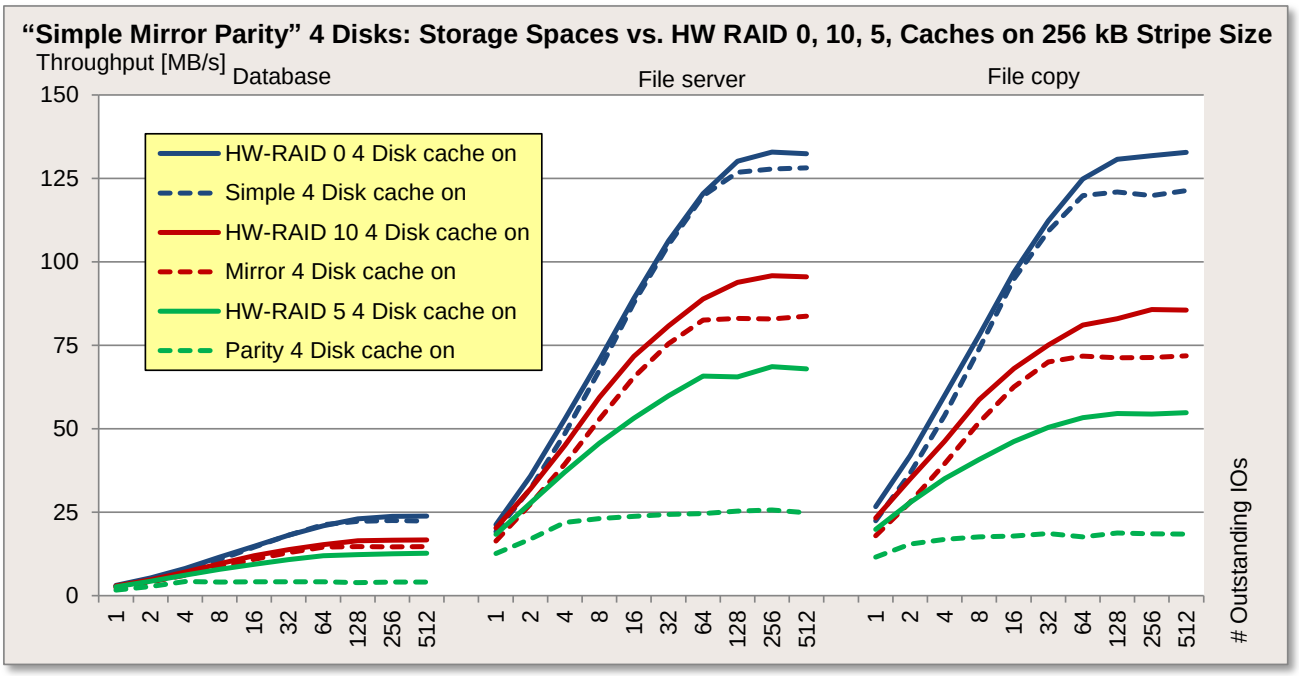
The following measuring configuration describes the storage layouts "Simple", "Mirror" and "Parity" with four hard disks compared with the appropriate HW RAID configuration 0, 10 and 5. The write cache of the hard disks is enabled (Disk cache on).

Important:
The default stripe size of the HW RAID arrays was set from 64 kB to 256 kB here (the default value of the Storage Spaces).

The first diagram initially shows the measurement results of the standard load profiles "Streaming" (sequential access, 100% read, 64 kB block size) and "Restore" (sequential access, 100% write, 64 kB block size).



The next diagram shows the course of the standard load profiles “Database” (random access, 67% read, 8 kB block size), “File server” (random access, 67% read, 64 kB block size) and finally “File copy” (random access, 50% read, 64 kB block size).

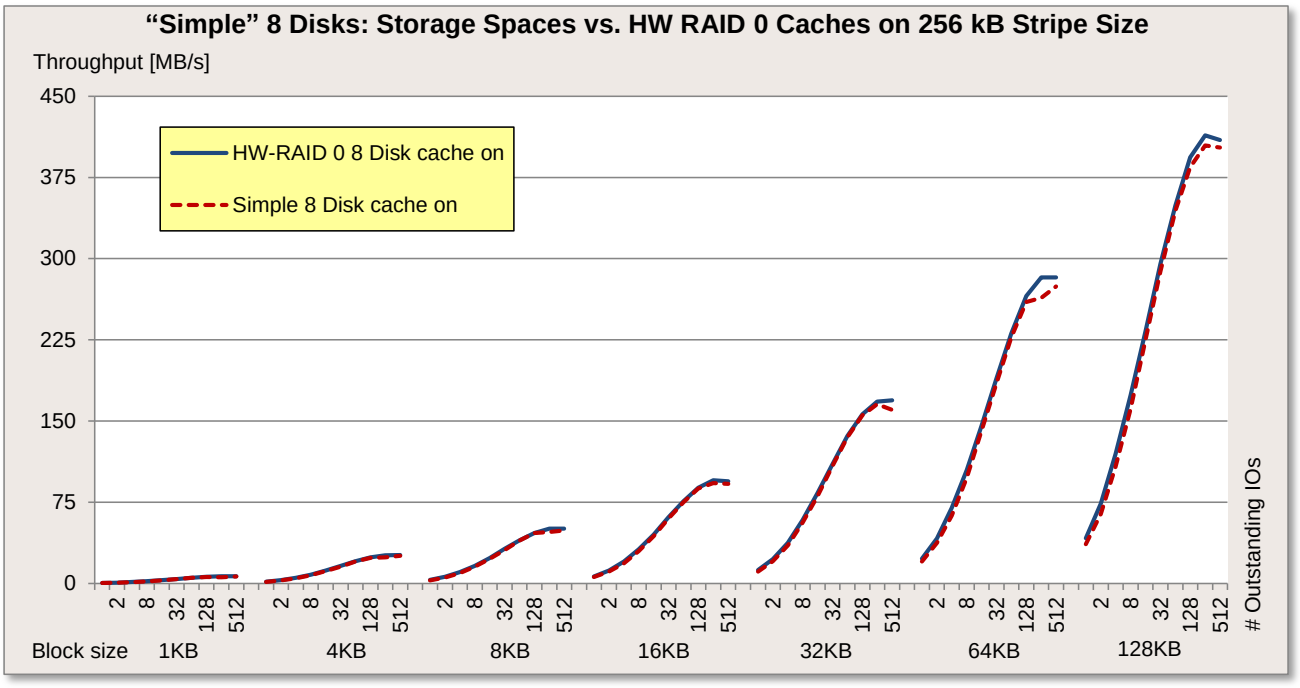


Another picture of the comparison of Storage Spaces with HW RAID configurations can now be seen if you use equally large stripe size values. Using the example with four disks in all storage layouts we can see the equivalence and the slight advantages of the HW RAID. Associated with this a somewhat lower CPU load can be seen for the hardware variant, but which should hardly have any relevance in practice.

It thus becomes clear that the consideration of equivalent configurations for the appropriate use scenario is an important point in the assessing of disk subsystems' performance.

The advantages can also be seen in this configuration if you not only compare (as above) the selected standard load profile, but possible further load profiles of server applications.

Diagram of a load profile of server applications with random access and 67% read with different block sizes.

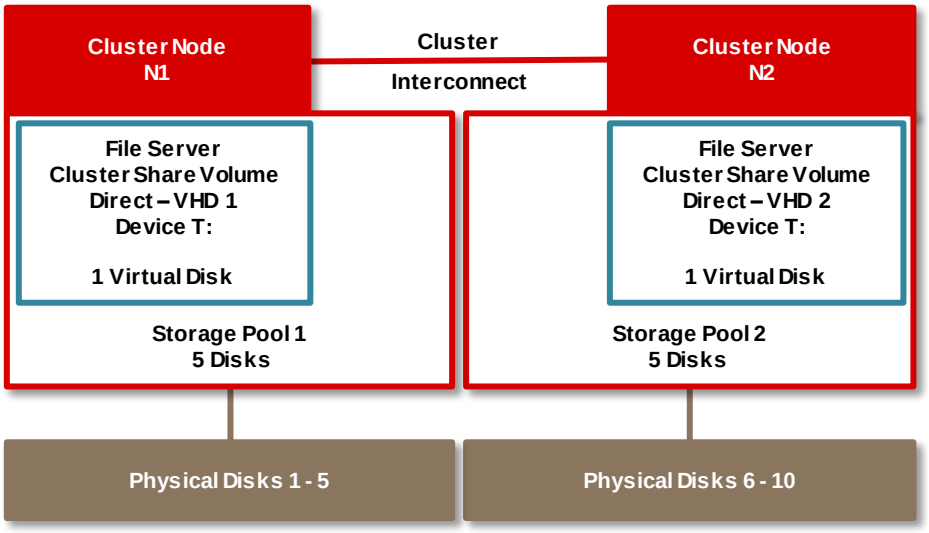


Storage Spaces and Shared Volume

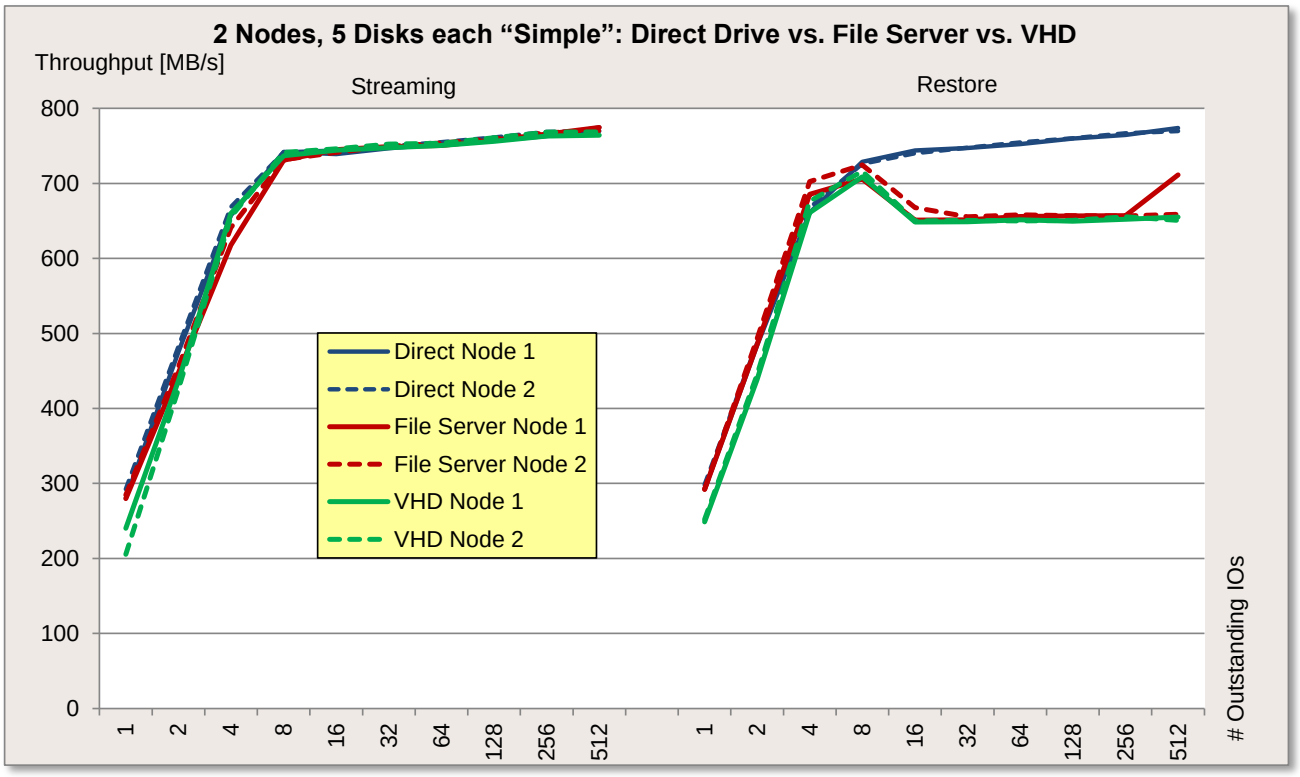
How do Storage Spaces now behave in the cluster environment? The previously described options (direct drives, file server role and VHD) can be used to access the cluster shared volume (CSV). First of all, we will consider these variants in physically separate storage pools. For this purpose, a storage pool of five disks was in the case of two cluster nodes allocated to each node and the configuration was carried out via cluster management. The load was generated on both nodes in parallel, with the different access variants.

The measuring tool Iometer with the standard measurement method and the measurement environment have also been used here, as described in the introduction.

The following picture illustrates the set configuration.

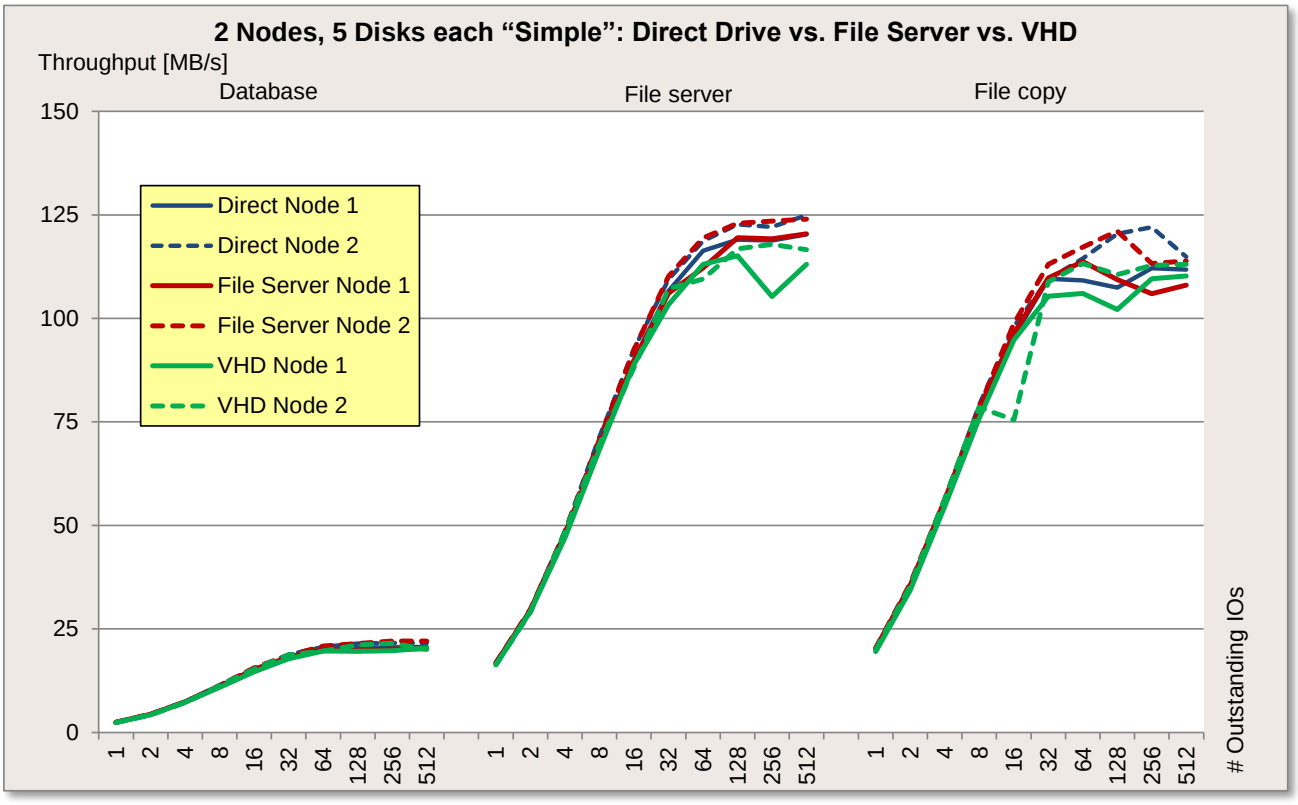


The first diagram initially shows the measurement results of the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).



In the case of sequential read there is no difference between the access variants. And in the case of sequential write direct drive access has slight advantages. The management layer of file server role and VHD organization may incite minor deductions here.

The next diagram shows the course of the standard load profiles “Database” (random access, 67% read, 8 kB block size), “File server” (random access, 67% read, 64 kB block size) and finally “File copy” (random access, 50% read, 64 kB block size).

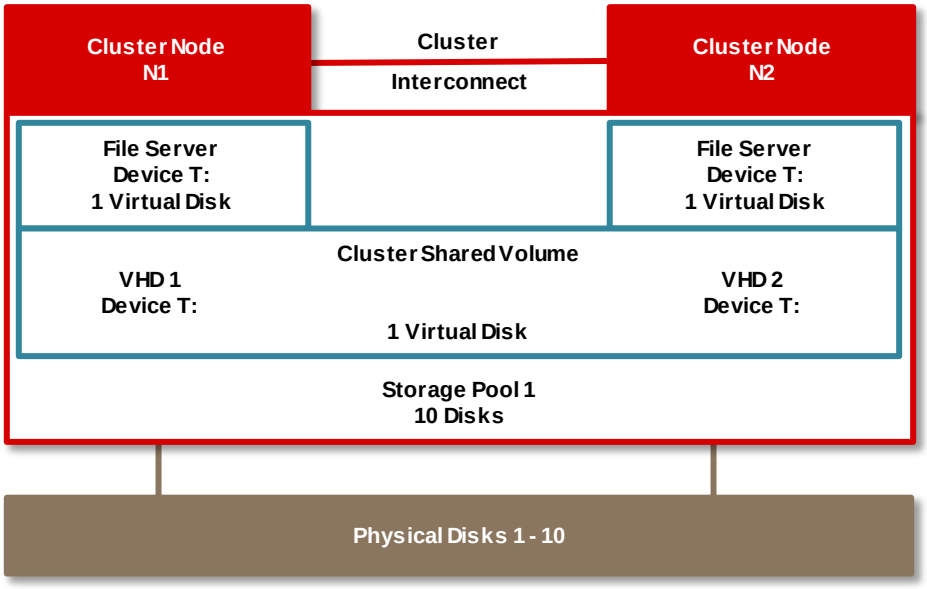


In the case of random accesses there is a somewhat inconsistent picture with the larger number of “Outstanding IOs”, but on the whole we cannot see any major deviations as far as the different access variants are concerned.

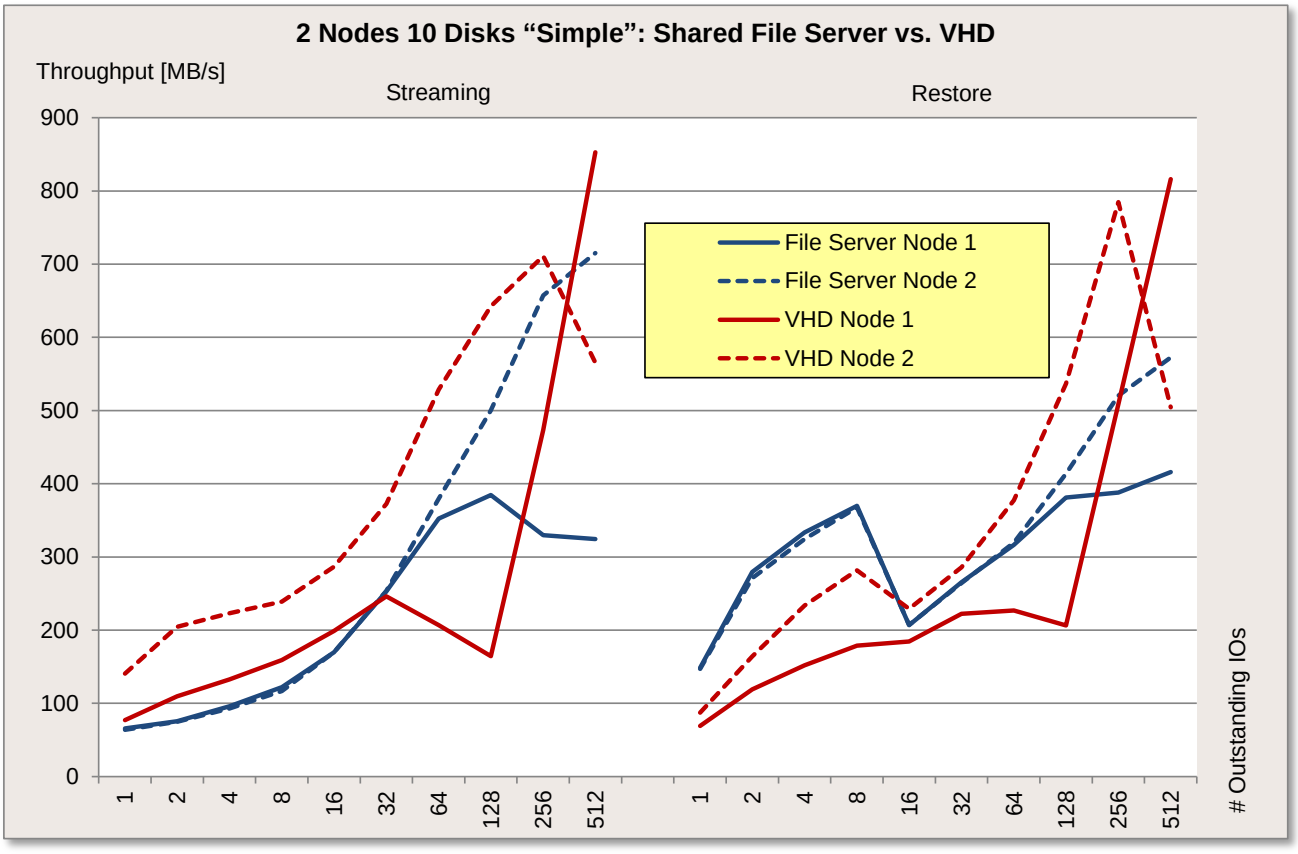
We see an almost identical result for an access configuration without cluster management (not shown here), i.e. with two nodes which simultaneously access the respective disks only via a storage space configuration. So we can see that the logic of drive connection does not play a role in the case of “undisturbed” exclusive access to a node.

A visible management overhead now ensues with the shared volume variants, where various nodes have to synchronize simultaneous accesses. To this end, ten hard disks were made available to both nodes in a storage pool. Both nodes simultaneously accessed all disks - once with the file server role, and the other time as a VHD drive.

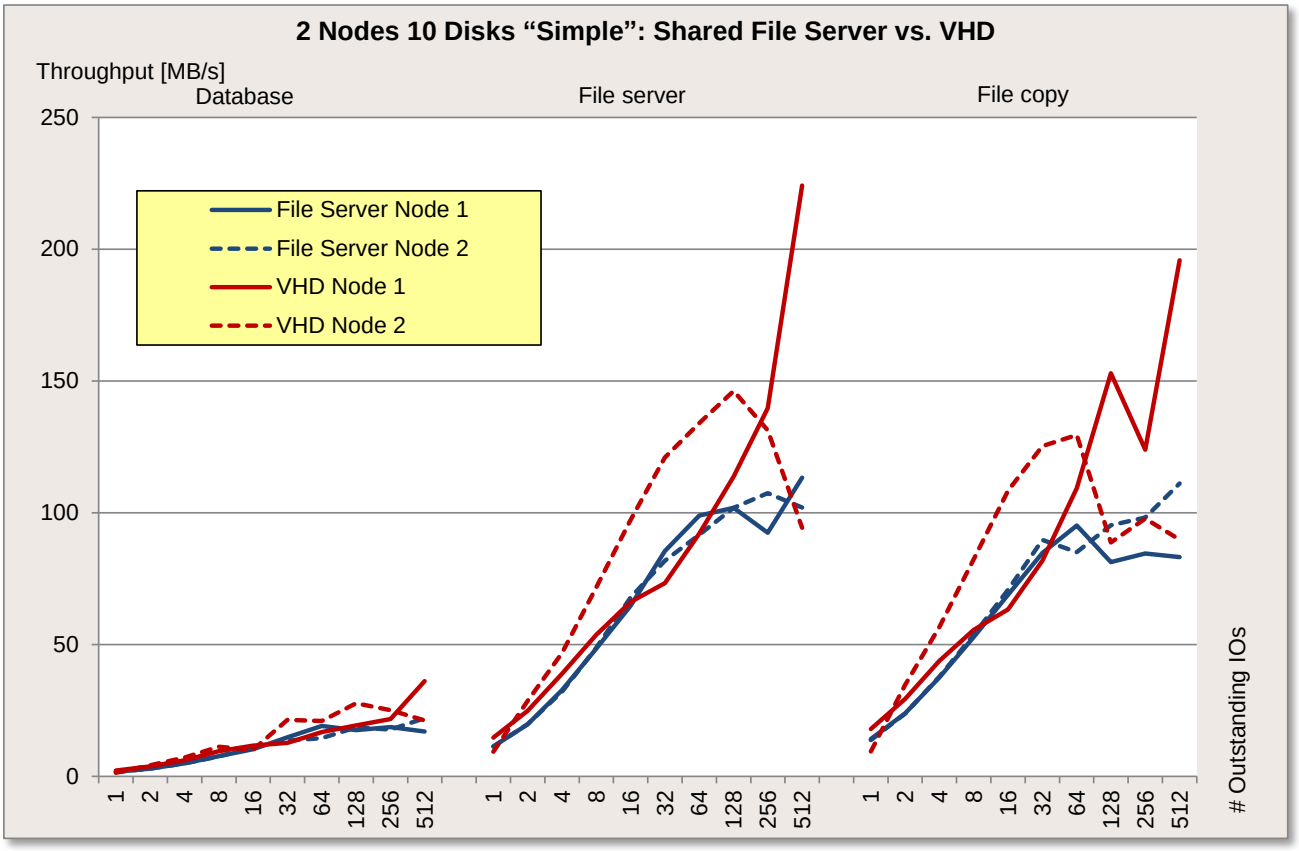
The following picture now also illustrates the corresponding configuration.



The first diagram initially shows the measurement results of the standard load profiles “Streaming” (sequential access, 100% read, 64 kB block size) and “Restore” (sequential access, 100% write, 64 kB block size).



The next diagram shows the course of the standard load profiles “Database” (random access, 67% read, 8 kB block size), “File server” (random access, 67% read, 64 kB block size) and finally “File copy” (random access, 50% read, 64 kB block size).



The effects of the necessary access synchronization can now be seen when both nodes simultaneously access the shared volume. The two nodes now do not work in parallel on their dedicated five hard disks, instead both nodes work in parallel on a pool of 10 hard disks, i.e. on the shared data areas simultaneously. It is possible to see a different load for the individual nodes now, combined with a change in the maximum throughput from one node to the other. The overall result is a lower and less constant throughput. As already determined above, the management layers of cluster management do not remain without losses.

It seems to make more sense here to consider the average performance via the sum of the nodes. In other words, with a shared volume configuration the focus is on the overall cluster performance (e.g. within the framework of a failover strategy).

Conclusion

The handling of files and storage with storage pools and Storage Spaces takes the form in Windows 2012 of a simple and effective option of organizing appropriate hard disk configurations, depending on requirements. The configuration options can be clearly shown and comprehensively automated with the graphic interface of the Server Manager, but also via the PowerShell commands. The flexibility of the connection interfaces and the variable configuration options facilitate the planning of the hard disk systems that are to be connected.

The measurements show that according to load profiles there are typical, different throughputs. This is related to the server applications with very different access requirements, but also the hard disks for which we can determine clear scaling.

The stripe size is also a decisive parameter, which receives a special mention here, because it permits different assessments in a default setting compared with an optimized setting. If 64 kB stripe size for the HW RAID configuration is to be compared with 256 kB stripe size of the Storage Spaces, the Storage Spaces throughputs are almost without exception (e.g. not “Parity”) above the HW RAID results.

In the case of an optimized identical stripe size (in this case 256 kB) the HW RAID throughputs are always slightly higher. Particularly in the case of specific storage layouts and application profiles. Write accesses for the storage layout “Parity” are always clearly behind the appropriate HW RAID 5 configuration. This storage layout is therefore only recommended for predominantly read application scenarios.

In summary, it can be seen that in the environment considered Storage Spaces are an adequate alternative to a comparable HW RAID variant. This also applies for use in the cluster environment. The use of hard disk caches, the dimension of the stripe size and also the briefly mentioned number of columns for the individual storage layouts are decisive for the optimized throughput.

However, it should also be said that RAID 5 should always be covered by the HW RAID solution. Time-sensitive applications with high throughput requirements should also refer to the (optimized) HW RAID solution. Security-relevant applications benefit in any case from the controller cache of some HW RAID solutions, which can be accordingly safeguarded against power failures.

Literature

PRIMERGY Systems

http://primergy.com/

PRIMERGY Performance

http://www.fujitsu.com/fts/products/computing/servers/primergy/benchmarks/

Basics of Disk I/O Performance

http://docs.ts.fujitsu.com/dl.aspx?id=65781a00-556f-4a98-90a7-7022feacc602

Single Disk Performance

http://docs.ts.fujitsu.com/dl.aspx?id=0e30cb69-44db-4cd5-92a7-d38bacec6a99

RAID Controller Performance

http://docs.ts.fujitsu.com/dl.aspx?id=e2489893-cab7-44f6-bff2-7aeea97c5aef

Datasheet: Windows Server 2012

http://docs.ts.fujitsu.com/dl.aspx?id=3c7d9f54-b897-415e-b825-6cb117f5e2cd

Datasheet Fujitsu RAID Controller SAS 6Gbit/s 1 GB

http://docs.ts.fujitsu.com/dl.aspx?id=547c7c85-2530-49af-9370-2aae2a153672

Data sheet: Fujitsu PSAS CP200i SAS Controller

http://docs.ts.fujitsu.com/dl.aspx?id=17fffe0d-36e6-49e4-8f62-47d9ea2c9800

Fujitsu Cluster-in-a-box

Datasheet: Fujitsu Cluster-in-a-box based on PRIMERGY BX400-L
http://docs.ts.fujitsu.com/dl.aspx?id=f37f88b3-8d45-45fa-9776-f69355fbd674

Datasheet: Fujitsu Cluster-in-a-box based on PRIMERGY BX400-M
http://docs.ts.fujitsu.com/dl.aspx?id=1cb3d6c9-973e-4531-ad2d-50288bca21e3

Datasheet: Fujitsu Cluster-in-a-box based on PRIMERGY BX400-S
http://docs.ts.fujitsu.com/dl.aspx?id=f9f8e66f-bcc5-430e-8a5c-3ae366439c9d

Datasheet: Fujitsu Cluster-in-a-box based on PRIMERGY CX420-M
http://docs.ts.fujitsu.com/dl.aspx?id=e0a3a140-39f7-4237-9657-865f957b46ad

Datasheet: Fujitsu Cluster-in-a-box based on PRIMERGY CX420-S
http://docs.ts.fujitsu.com/dl.aspx?id=4e5234e5-fbde-41b1-a157-0d8eb44f5b04

Information about Iometer

http://www.iometer.org

Contact

FUJITSU

Website: http://www.fujitsu.com/

PRIMERGY Product Marketing

mailto:Primergy-PM@ts.fujitsu.com

PRIMERGY Performance and Benchmarks

mailto:primergy.benchmark@ts.fujitsu.com