Microsoft | Services

Building High performance fault tolerance storage using Windows Server 2012 R2

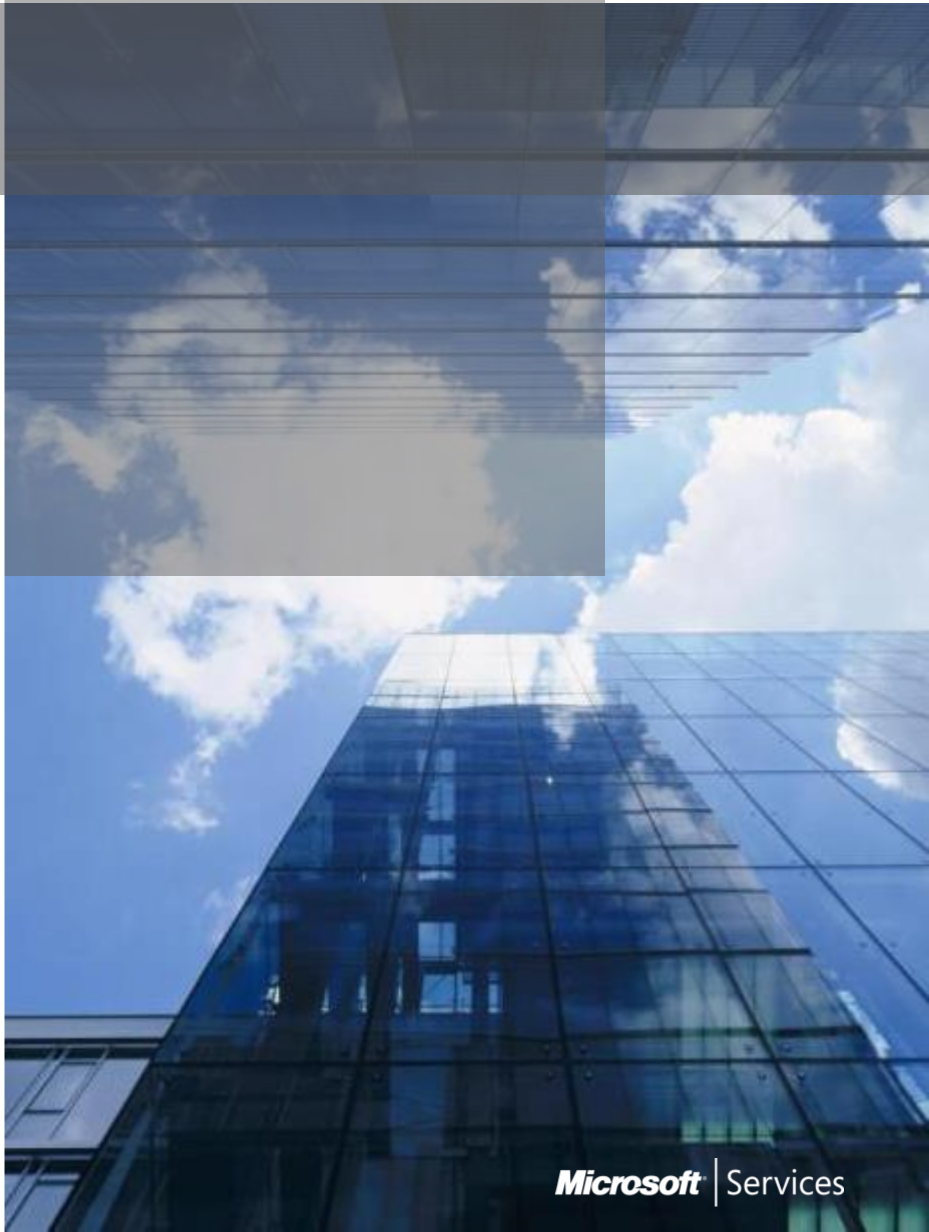
STRATEGY ■ CONSULTING ■ SUPPORT

Moti Bani, Premier Field Engineer

1/5/2014

Table of Contents

- **Storage new model**
- SAN
- JBODs
- SMB 3
- Wrapping all together



Microsoft | Services

Storage Changes

- **Azure**
  - 1,000,000 physical machines (48 cores, 96 GB RAM)
  - Azure is actually Windows vNext
  - Major challenge: How to scale storing VM, make it cheaper without compromise performance or features?
- **Traditional Fiber-channel solutions**
  - Slow (only 8GB)
  - Expensive
  - Complex

Storage Approach

- Strategic decision: Lower the storage costs

  - **Storage spaces (**management layer)

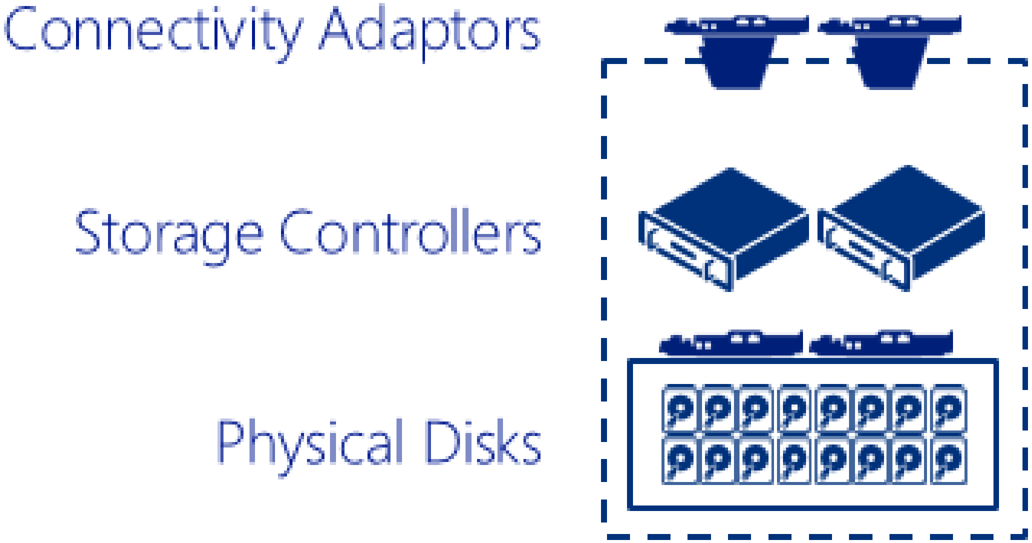
    Cost effective business storage

  - **File based storage** (infrastructure layer)

    High performance SMB for application workload

What is a SAN really

- **Enterprise-grade Physical Disks**
- **Storage Controllers**
- **Connectivity Adapters**

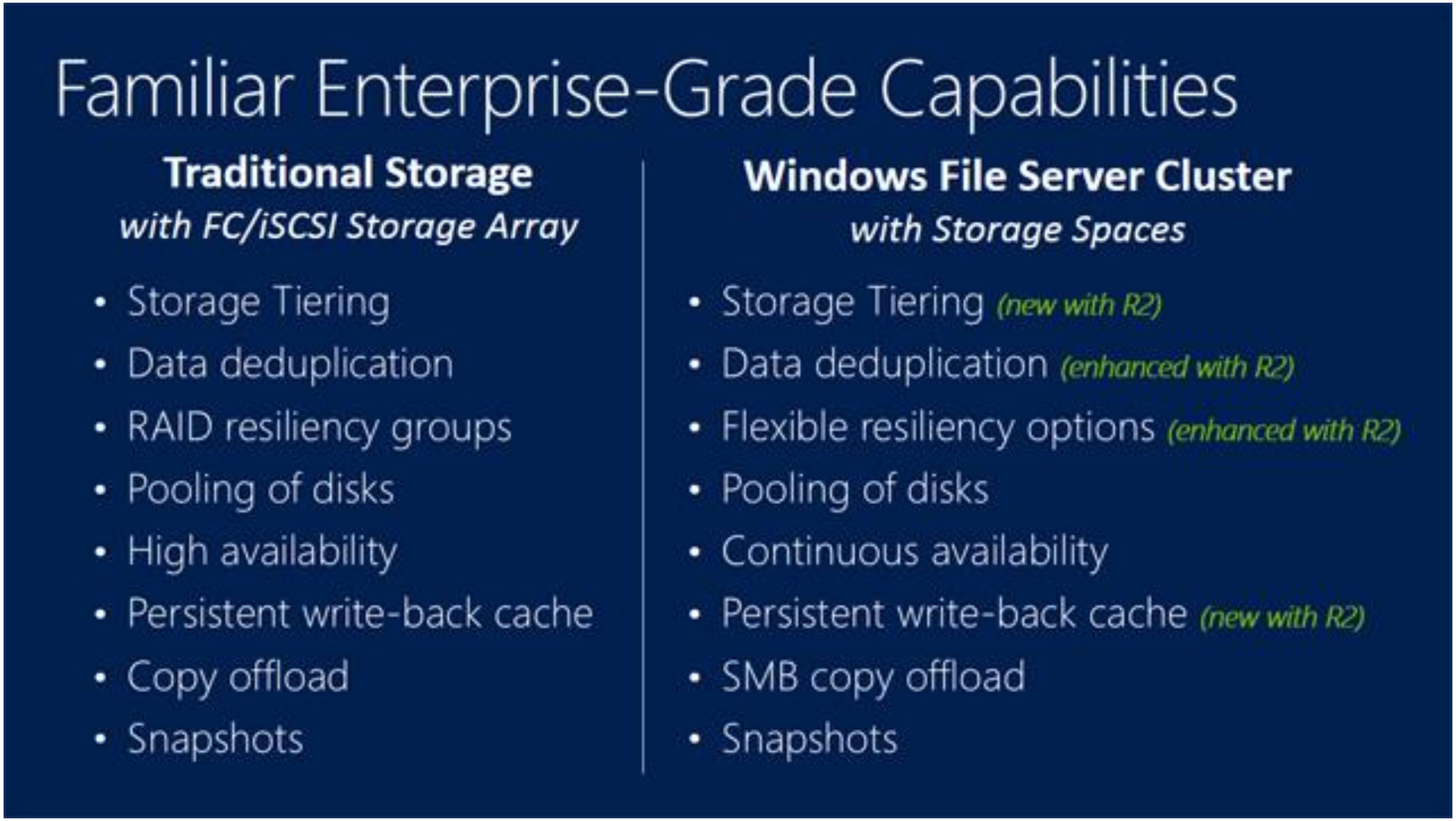


Connectivity Adaptors

Storage Controllers

Physical Disks

Enterprise-grade capabilities



Familiar Enterprise-Grade Capabilities

Traditional Storage
with FC/iSCSI Storage Array

- Storage Tiering
- Data deduplication
- RAID resiliency groups
- Pooling of disks
- High availability
- Persistent write-back cache
- Copy offload
- Snapshots

Windows File Server Cluster
with Storage Spaces

- Storage Tiering *(new with R2)*
- Data deduplication *(enhanced with R2)*
- Flexible resiliency options *(enhanced with R2)*
- Pooling of disks
- Continuous availability
- Persistent write-back cache *(new with R2)*
- SMB copy offload
- Snapshots

How it works

- The new Storage Spaces feature in Windows Server 2012 is a software virtualization and management layer

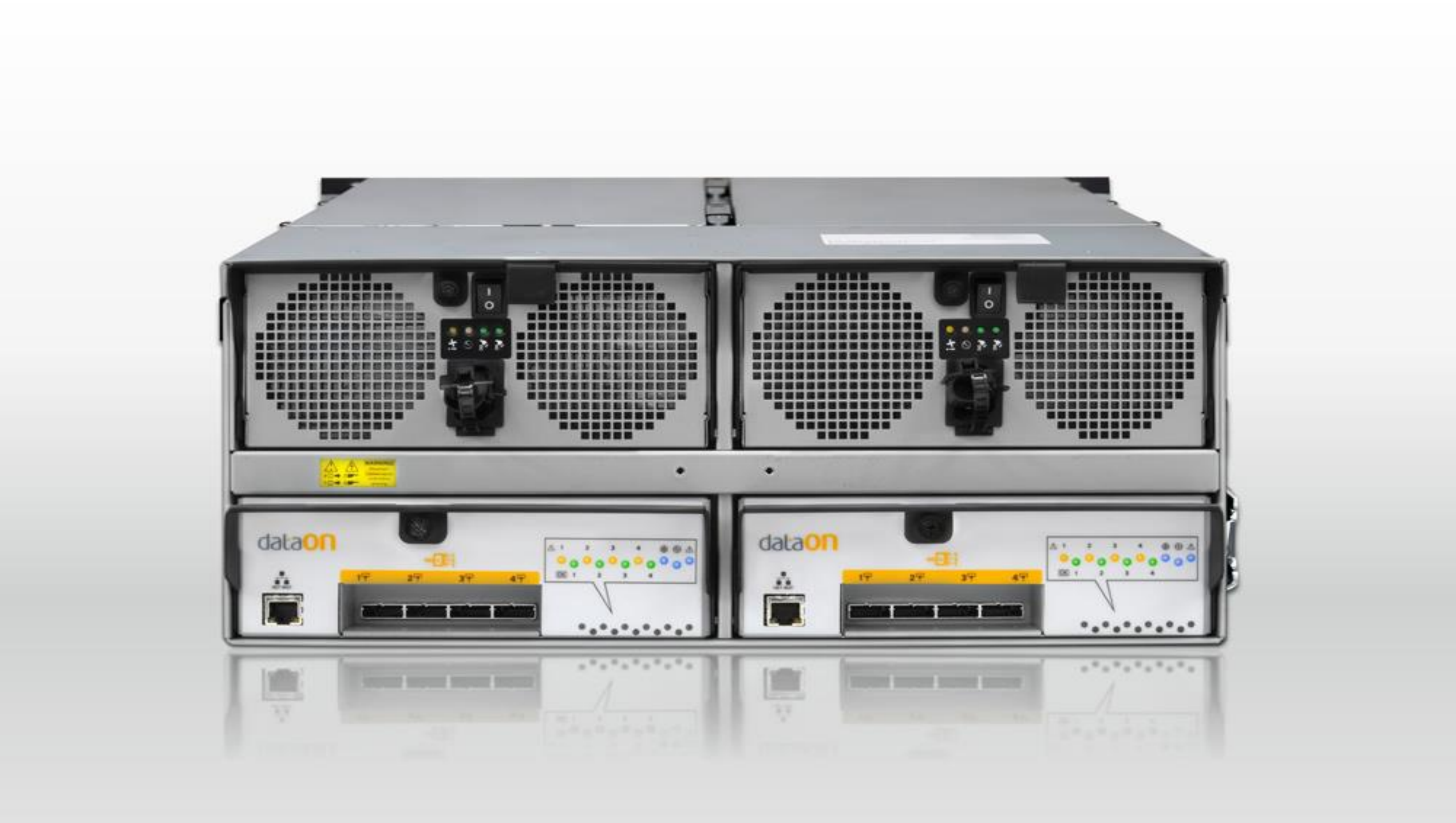
- Storage administrators can group inexpensive disks into Storage Pools
    - Enable storage aggregation, elastic expansion, and delegated administration

- From this pool you create Virtual Disks (LUN)
    - Each Virtual Disk can be simple, mirrored or parity

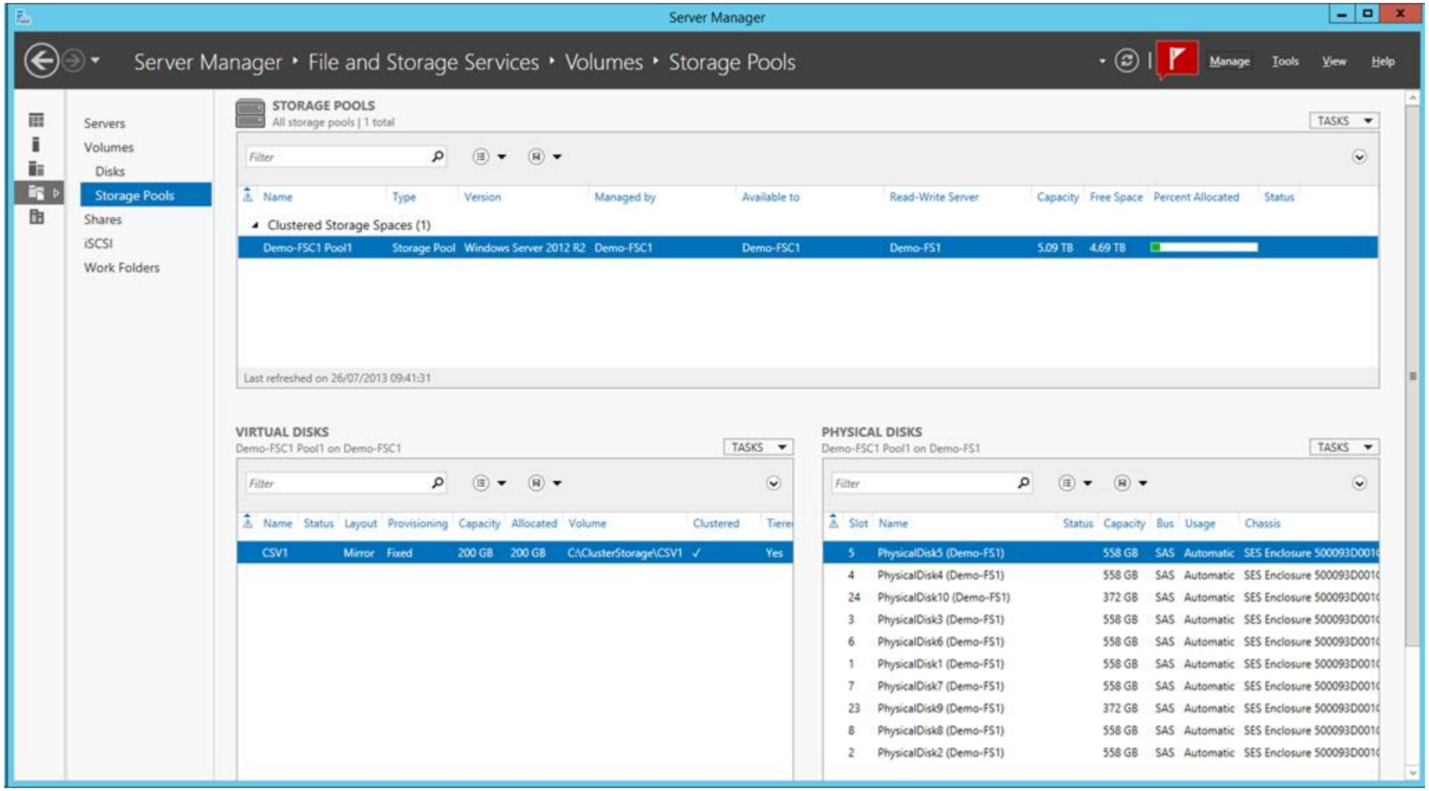
Leveraging the same enterprise-grade SAS SSD and HDD disks

See www.windowsservercatalog.com under the “Storage Spaces” Category

Closer look on the JBOD …



Connecting the JBOD disks



STORAGE POOLS
All storage pools | 1 total

Name	Type	Version	Managed by	Available to	Read-Write Server	Capacity	Free Space	Percent Allocated	Status
Clustered Storage Spaces (1)									
Demo-FSC1 Pool1	Storage Pool	Windows Server 2012 R2	Demo-FSC1	Demo-FSC1	Demo-FS1	5.09 TB	4.69 TB		

Last refreshed on 26/07/2013 09:41:31

VIRTUAL DISKS
Demo-FSC1 Pool1 on Demo-FSC1

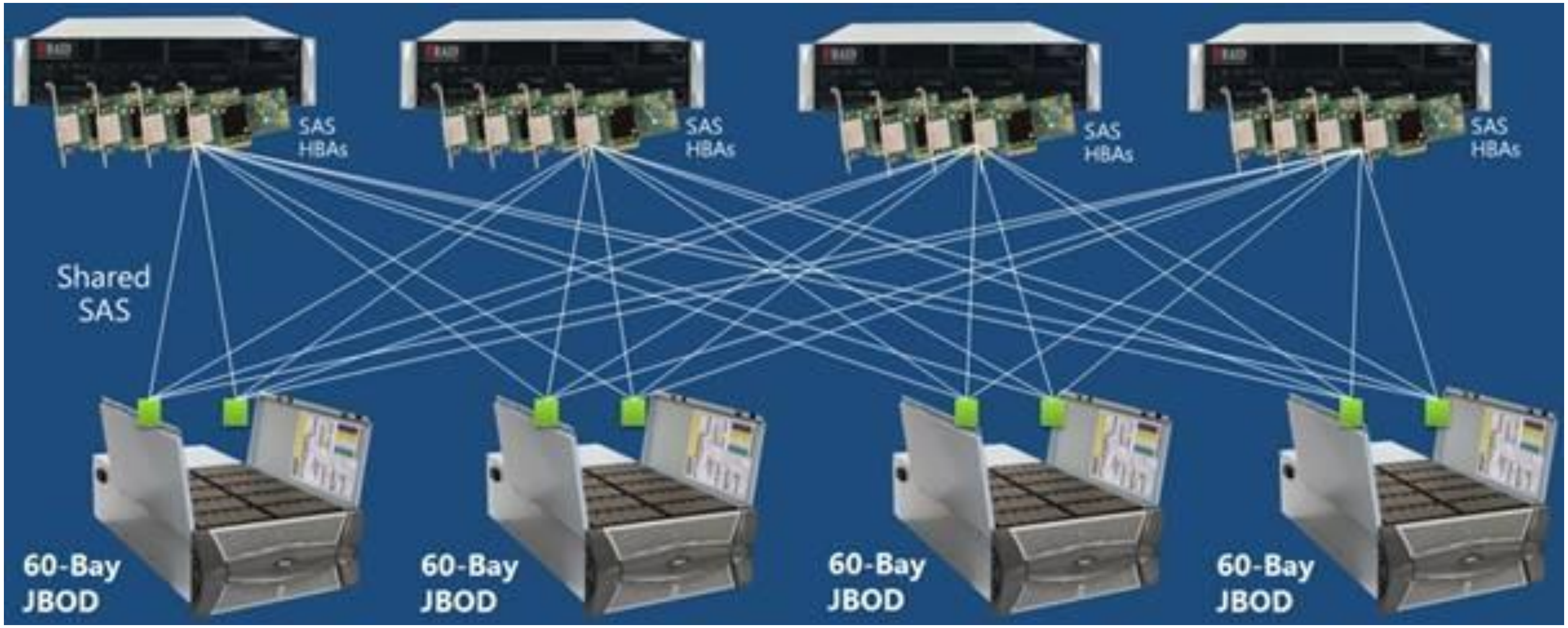
Name	Status	Layout	Provisioning	Capacity	Allocated	Volume	Clustered	Tiere
CSV1		Mirror	Fixed	200 GB	200 GB	C:\ClusterStorage\CSV1	✓	Yes

PHYSICAL DISKS
Demo-FSC1 Pool1 on Demo-FS1

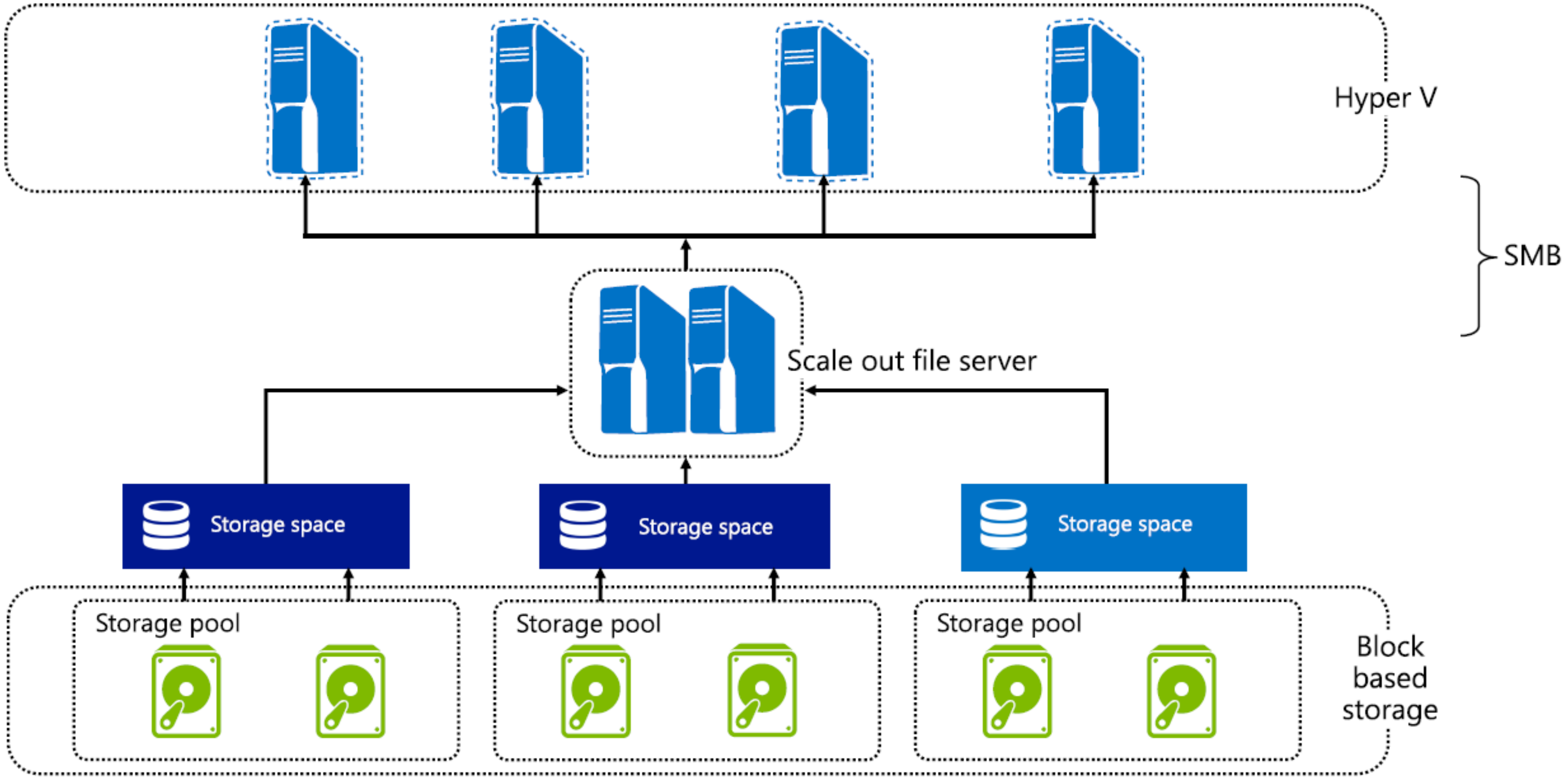
Slot	Name	Status	Capacity	Bus	Usage	Chassis
5	PhysicalDisk5 (Demo-FS1)		558 GB	SAS	Automatic	SES Enclosure 500093D001
4	PhysicalDisk4 (Demo-FS1)		558 GB	SAS	Automatic	SES Enclosure 500093D001
24	PhysicalDisk10 (Demo-FS1)		372 GB	SAS	Automatic	SES Enclosure 500093D001
3	PhysicalDisk3 (Demo-FS1)		558 GB	SAS	Automatic	SES Enclosure 500093D001
6	PhysicalDisk6 (Demo-FS1)		558 GB	SAS	Automatic	SES Enclosure 500093D001
1	PhysicalDisk1 (Demo-FS1)		558 GB	SAS	Automatic	SES Enclosure 500093D001
7	PhysicalDisk7 (Demo-FS1)		558 GB	SAS	Automatic	SES Enclosure 500093D001
23	PhysicalDisk9 (Demo-FS1)		372 GB	SAS	Automatic	SES Enclosure 500093D001
8	PhysicalDisk8 (Demo-FS1)		558 GB	SAS	Automatic	SES Enclosure 500093D001
2	PhysicalDisk2 (Demo-FS1)		558 GB	SAS	Automatic	SES Enclosure 500093D001

```
$pooldisks = Get-PhysicalDisk | ? {$_.CanPool –eq $true }
New-StoragePool -StorageSubSystemFriendlyName *Spaces* -
FriendlyName TieredPool1 -PhysicalDisks $pooldisks
```

Connecting the JBODs



Microsoft vision



Speeds

HBA	Throughput
3Gb SAS x4	~1.1 GB/sec
6Gb SAS x4	~2.2 GB/sec
4Gb FC	~0.4 GB/sec
8Gb FC	~0.8 GB/sec
16Gb FC	~1.5 GB/sec

NIC	Throughput
1Gb Ethernet	~0.1 GB/sec
10Gb Ethernet	~1.1 GB/sec
40Gb Ethernet	~4.5 GB/sec
32Gb InfiniBand (QDR)	~3.8 GB/sec
54Gb InfiniBand (FDR)	~6.5 GB/sec

Bus Slot	Throughput
PCIe Gen2 x4	~1.7 GB/sec
PCIe Gen2 x8	~3.4 GB/sec
PCIe Gen2 x16	~6.8 GB/sec
PCIe Gen3 x4	~3.3 GB/sec
PCIe Gen3 x8	~6.7 GB/sec
PCIe Gen3 x16	~13.5 GB/sec

Numbers are rough approximations.
Actual throughput in real life will be lower than these theoretical maximums

SMB 3

- **SMB Transparent Failover** - Continuous availability
- **SMB Scale-Out** – Active/Active file server clusters
- **SMB Direct (SMB over RDMA)** - Low latency, low CPU use
- **SMB Multichannel** – Network throughput and failover
- **SMB Encryption** - Security
- **VSS for SMB File Shares** - Backup and restore
- **SMB PowerShell and VMM Support**

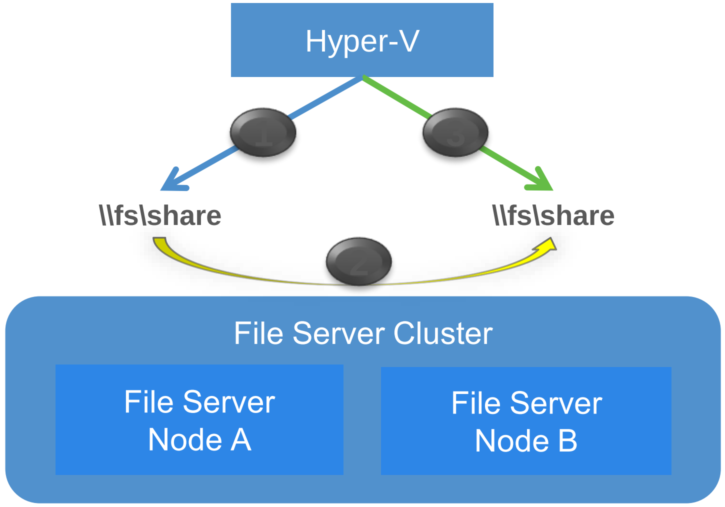
SMB Transparent Failover

- Failover transparent to server application
- Zero downtime – small IO delay during failover
- Supports planned and unplanned failovers

1 **Normal operation**

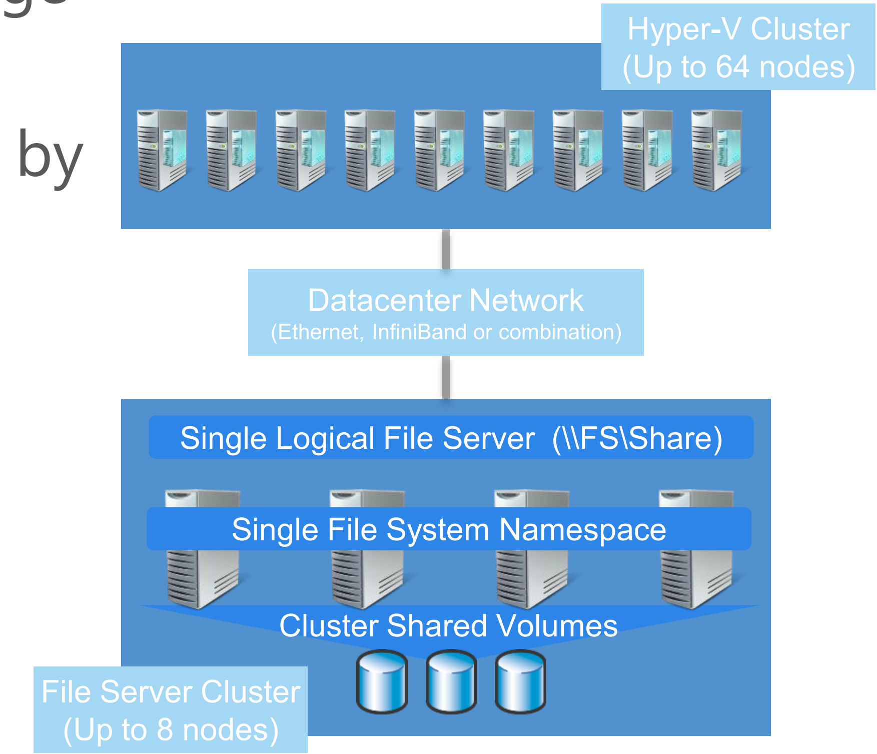
2 **Failover share - connections and handles lost, temporary stall of IO**

3 **Connections and handles auto-recovered Application IO continues with no errors**

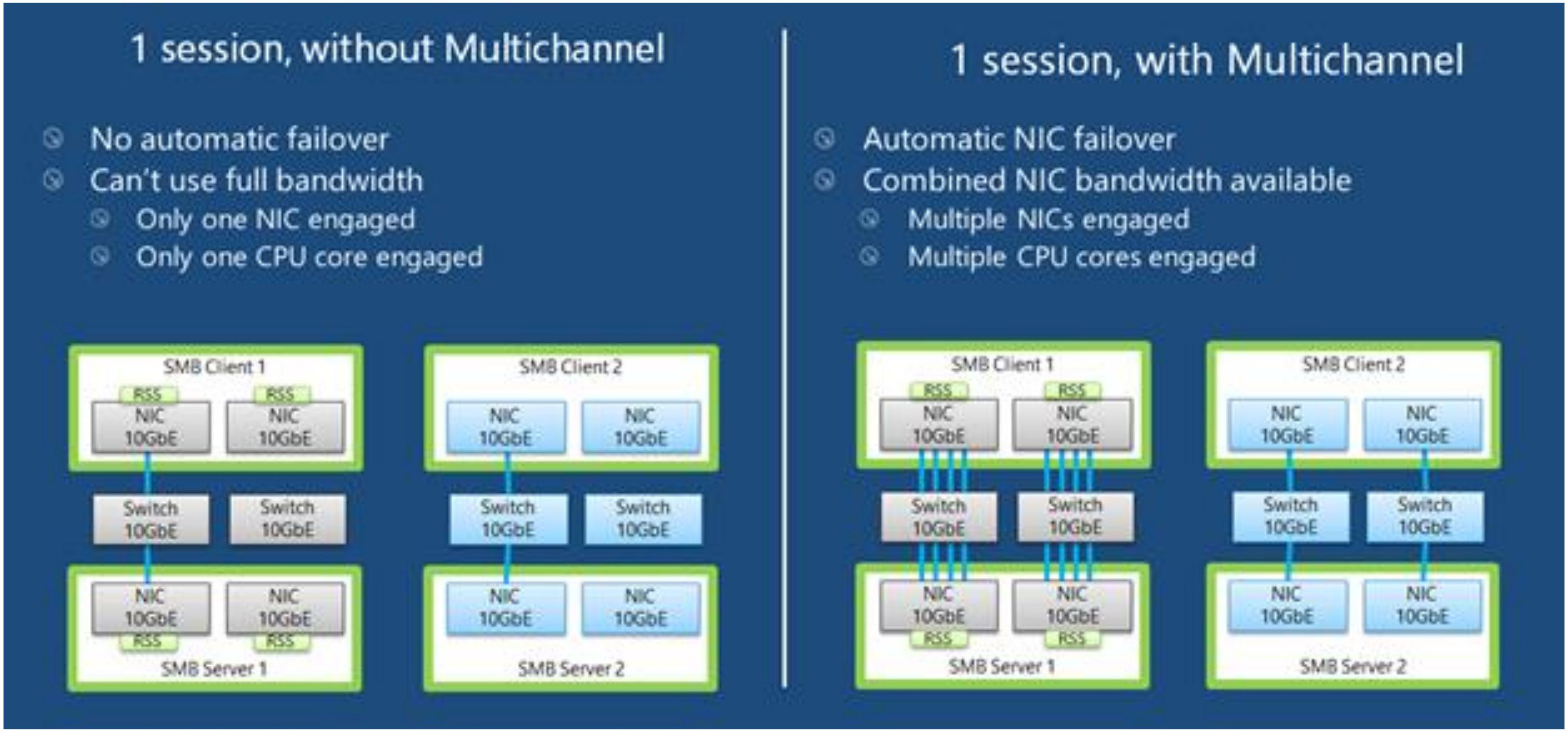


SMB Scale-Out

- Targeted for server app storage
  - Example: Hyper-V and SQL Server
- Increase available bandwidth by adding nodes
  - Leverages Cluster Shared Volumes (CSV)
- Key capabilities
  - Active/Active file shares
  - Fault tolerance with zero downtime
  - Fast failure recovery
  - CHKDSK with zero downtime
  - Support for app consistent snapshots
  - Support for RDMA enabled networks
  - Simple management



SMB Multi Channel



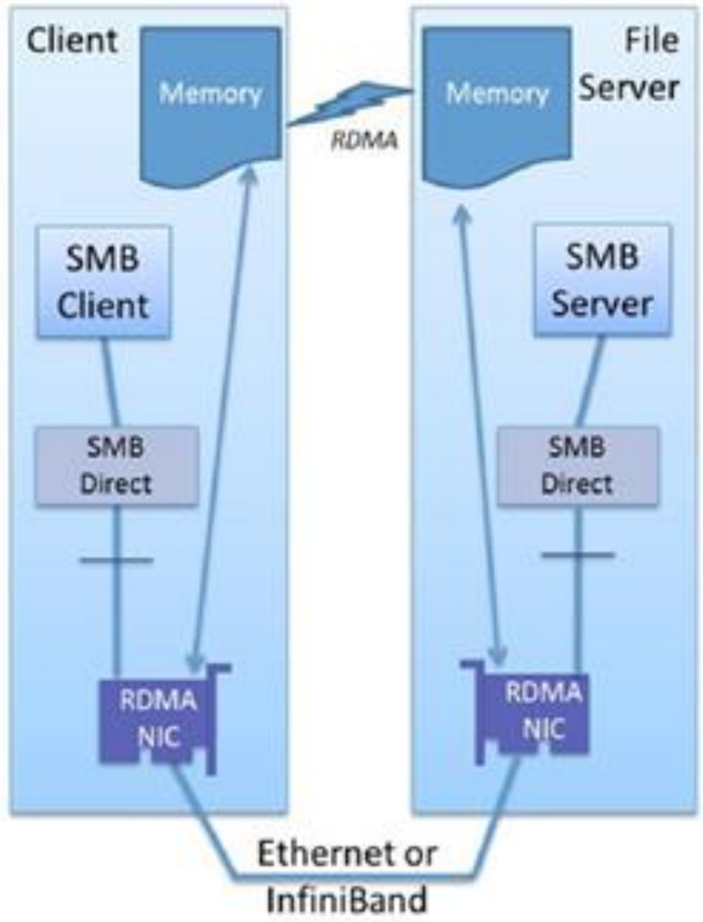
1 session, without Multichannel

- No automatic failover
- Can't use full bandwidth
  - Only one NIC engaged
  - Only one CPU core engaged

1 session, with Multichannel

- Automatic NIC failover
- Combined NIC bandwidth available
  - Multiple NICs engaged
  - Multiple CPU cores engaged

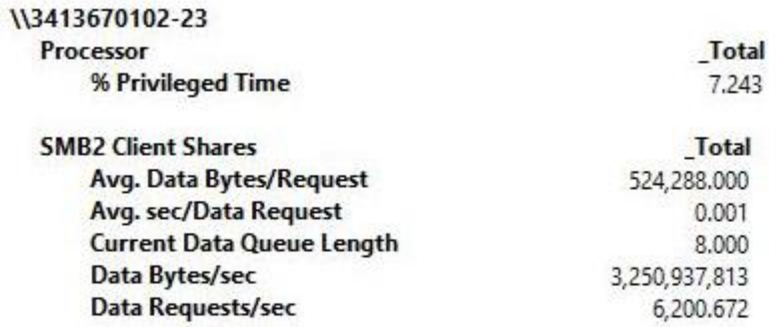
SMB Performance – Direct



TCP/IP to RDMA:

Scenario		Cycles Consumed		Storage bandwidth		Cycles/byte
A	Local baseline	0.9 Gcycles		3.74 GB/s		0.24
B	SMB with 10GbE	1.8 Gcycles	÷	1.19 GB/s	=	1.49
C	SMB Direct with 32GbIB	0.9 Gcycles		**3.28 GB/s**		0.28

Comparing B and C: 3 times faster, half the CPU, 5 times fewer cycles/byte



```
\\3413670102-23
   Processor                                      _Total
        % Privileged Time                          7.243

   SMB2 Client Shares                            _Total
        Avg. Data Bytes/Request              524,288.000
        Avg. sec/Data Request                      0.001
        Current Data Queue Length                  8.000
        Data Bytes/sec                     3,250,937,813
        Data Requests/sec                      6,200.672
```

Network Configuration Options

	Throughput	Fault Tolerance for SMB	Fault Tolerance for non-SMB	Reduced CPU utilization
MC + Single 1GbE				
MC + Multiple 1GbE	▲	▲		
MC + Multiple 1GbE + LBFO	▲	▲▲	▲	
MC + Single 10GbE	▲▲			
MC + Multiple 10GbE	▲▲▲	▲		
MC + Multiple 10GbE + LBFO	▲▲▲	▲▲	▲	
MC + Single RDMA	▲▲▲▲			▲
MC + Multiple RDMA	▲▲▲▲▲	▲		▲

Multichannel is on by default for SMB.
LBFO is helpful for faster failover.
LBFO is helpful for non-SMB traffic (mixed workloads, management).
LBFO is not compatible with RDMA.

Questions?

Microsoft | Services