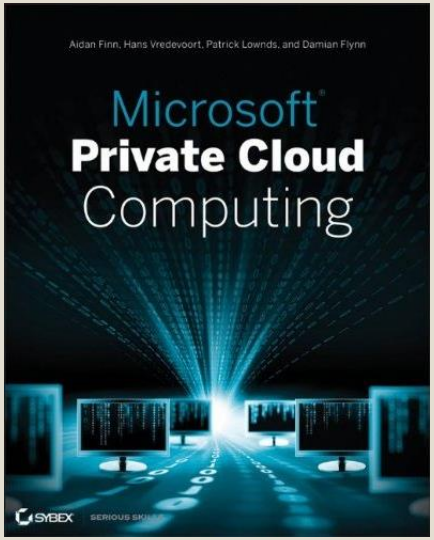
Storage Spaces

Aidan Finn

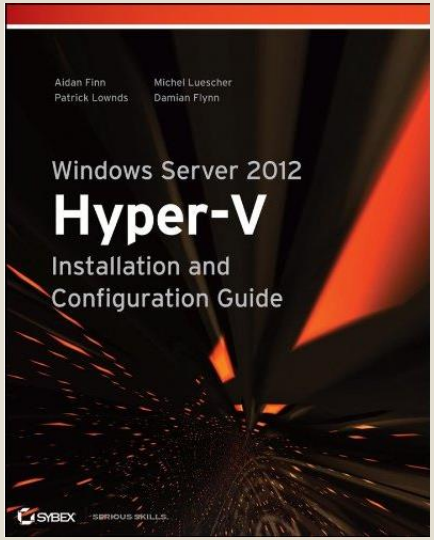
About Aidan Finn

- Technical Sales Lead at MicroWarehouse (Dublin)
- Working in IT since 1996
- MVP (Virtual Machine)
- Experienced with Windows Server/Desktop, System Center, virtualisation, and IT infrastructure
- @joe_elway
- http://www.aidanfinn.com
- http://www.petri.co.il/author/aidan-finn
- Published author/contributor of several books

Books



System Center 2012 VMM



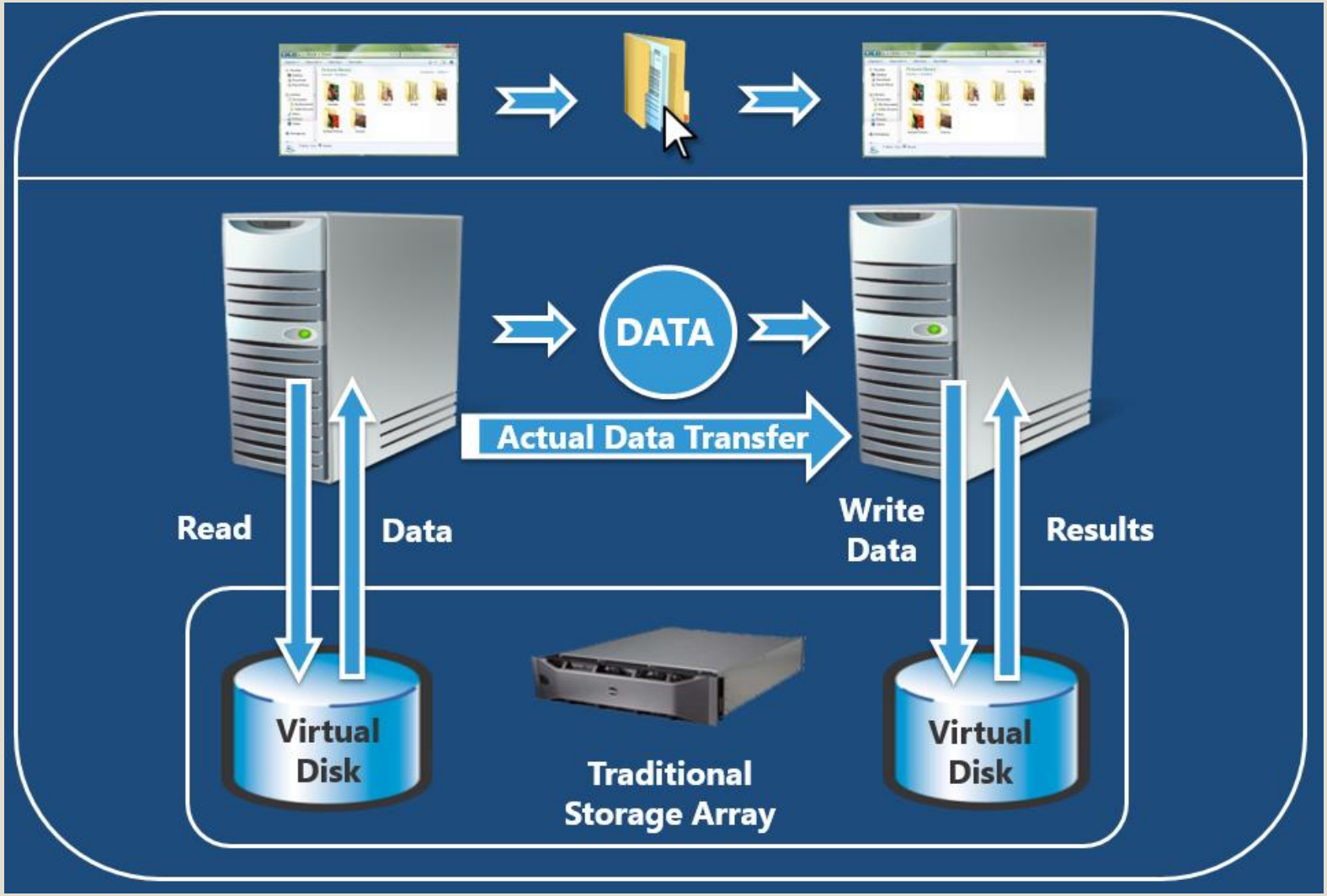
Windows Server 2012 Hyper-V

Agenda

- Item 1
- Item 2
- Item 3

Traditional Block Storage

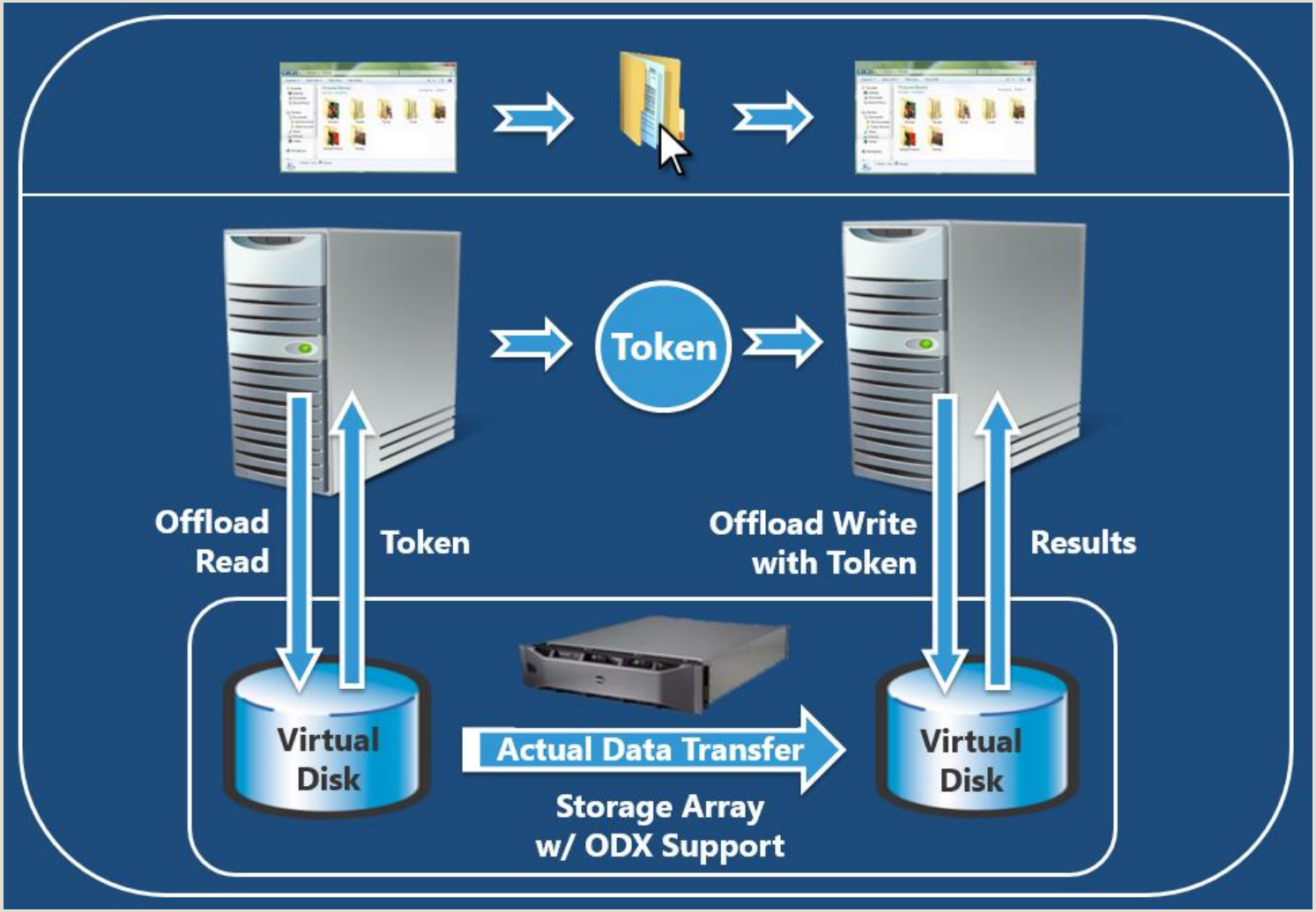
Normal Flow Of Data



Offloaded Data Transfer (ODX)

- Don't waste time/effort involving servers in data operations
- Offload operations to ODX-capable storage
- Results in faster operations:
  - Moving files
  - Creating fixed virtual hard disks
  - Deploying VM templates (requires VMM 2012 R2)
- Some manufacturers doing better than others!
- You might need to disable ODX when not using that storage:
  - On by default and *might* cause issues with incompatible SANs
  - http://technet.microsoft.com/library/jj200627.aspx

Storage With ODX



Highly Available Services

- Host clustering provides HA for hosts:
  - Preventative host maintenance
  - Automatic reactive failover after host failure
- HA hosts are an infrastructure element
- Business doesn't care about infrastructure
  - They care about services
- Actual *service* HA is done at the guest level:
  - By design of the service application
  - Or by using NLB/Guest Clustering

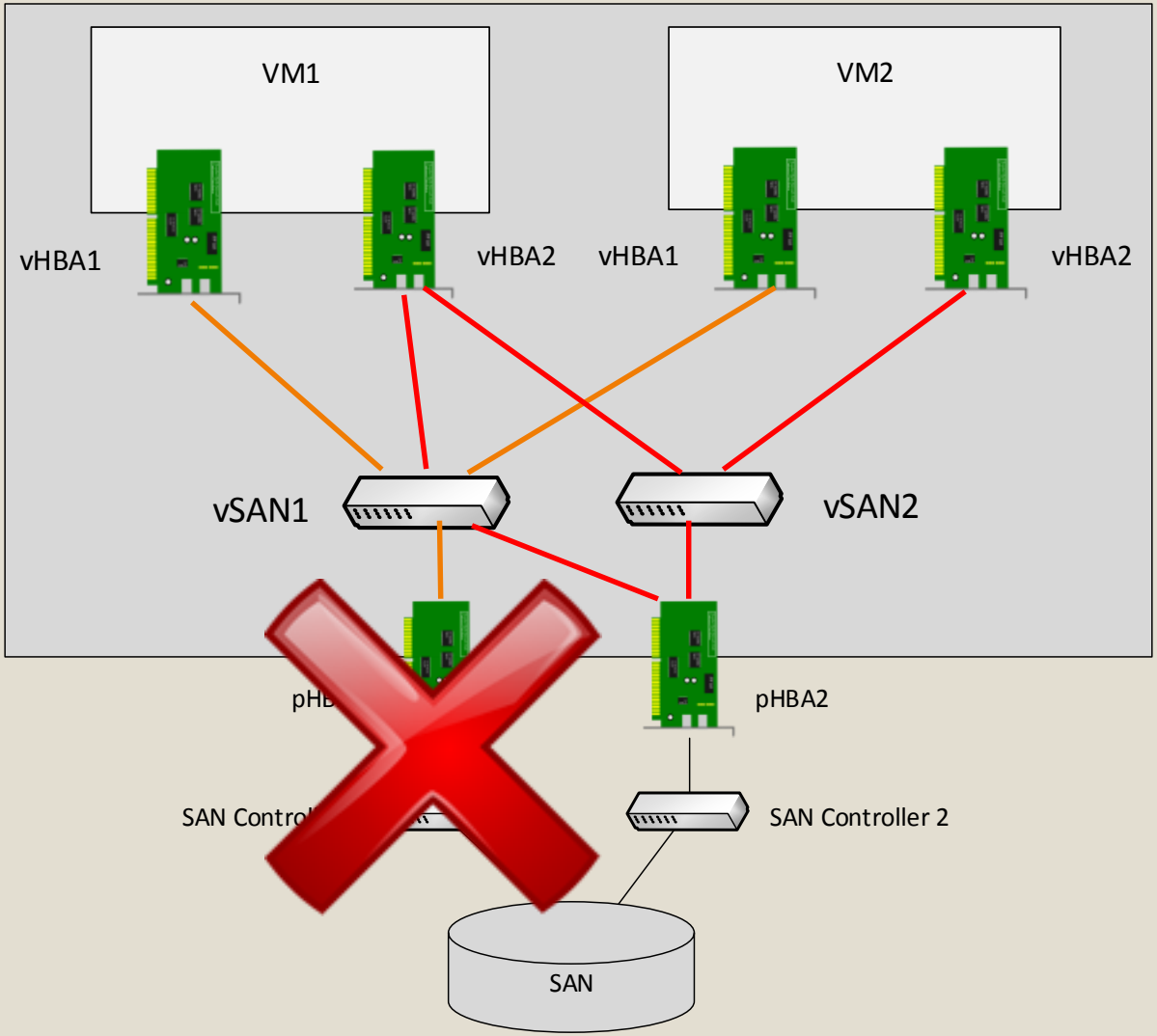
Guest Clustering

- A cluster that is implemented in the guest OS of VMs
- VMs will need shared storage
- Traditionally only possible with iSCSI
- WS2012 added support for virtual fibre channel

Virtual Fibre Channel

- Uses N_Port Virtualization ID (NPIV) feature in the host's HBA
- VMs have 1-4 virtual FC HBAs
- Each FC HBA has two worldwide names (WWNs) for SAN zoning
- VMs can connect to and use LUNs on FC SAN
- MPIO enabled in the VM guest OS
- VMs can live migrate without disconnect to the LUNs

Configuring Virtual Fibre Channel



Virtual Fibre Channel Live Migration

- Each vHBA gets two WWNs: A & B
- SAN administrator must configuring VM zoning for each WWN
- The VM switches between WWNs A and B as it Live Migrates
  - Allows SAN LUN dependency to be connected before VM is started on destination host & removed from source host
- Live migration … when on:
  - Host 1, VM uses WWNA to connect to SAN
  - Host 2, VM uses WWNB to connect to SAN
  - Host 3, VM uses WWNA to connect to SAN

Virtual Fibre Channel Tips

- Update firmware AND drivers throughout the entire stack:
  - SAN
  - HBAs
  - Servers
- Download and install recommended hotfixes for Hyper-V and Failover Cluster (they are not available via Windows Update)
- Go back to steps 1 & 2 and make sure you did them
- Make sure each host is configured identically
- Live Migrate new VMs to get WWNB active on the SAN for zoning

Storage Spaces

Storage Challenges

- Hardware RAID is inflexible and locks in manufacturer
- SAN (SAS/iSCSI/FC/FCoE) is expensive
- Can preclude SMEs from clustering
- Can be a challenge for enterprise/cloud too
  - Matching storage across sites
  - Held captive by storage sales for all data
  - Cost per TB
- So Microsoft decided to offer new storage options

Storage Spaces

- Let's get this out of the way:

  THIS IS NOT WINDOWS RAID OF THE PAST

- Software solution that aggregates and provisions LUNs (aka Virtual Disks) similarly to a SAN
- Aggregate just a bunch of disks (JBOD) into a Storage Pool
  - No hardware RAID
- Create "Virtual Disks" from the Storage Pool
  - Define level of fault tolerance for the virtual disk
  - Virtual disk consumes a bit of space from each disk in the JBOD – how depends on chosen fault tolerance

Use Cases

- Cheap storage on the back of a single server
- Storage Spaces is a support storage type for Failover Clustering
  - You don't need that expensive SAN anymore!

- Note: you can use Storage Spaces on Windows 8/8.1 to aggregate some USB drives
  - I use 2 * 3 TB USB 3.0 drives to store my photos, videos, etc
  - 1 disk mirrors the other

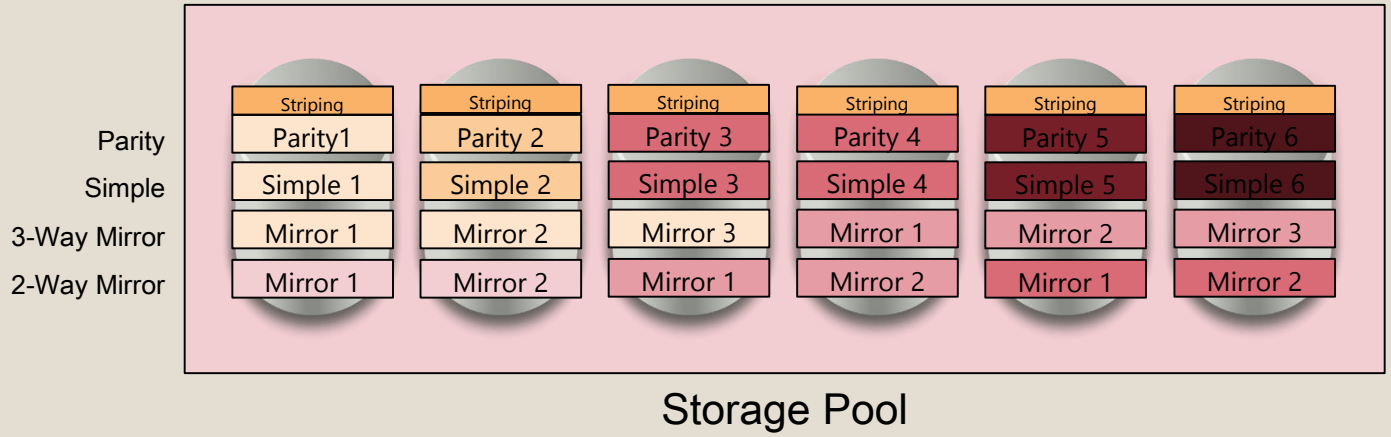
Storage Spaces Hardware

- Pointless implementing Storage Spaces on RAID hardware
- Use JBOD trays
  - Will support SCSI Enclosure Services (SES)
  - Connected via SAS adapter/cables with MPIO for fault tolerance
  - Can be daisy chained (see OEM guidance)
- There is a special HCL category for Storage Spaces supported hardware
  - You won't find big block storage vendors there ☺

Common Hardware Questions

- Can I add disks?
  - Yes, and new larger disks will be fully utilized!
- Is there disk failure detection?
  - Supported JBODs have SES support
  - JBOD can detect failure and inform Windows
  - Windows can detect failure and detect JBOD
  - Disk illumination as usual
- How does disk replacement work?
  - Support for hot spare (WS2012 and later) and parallelized restore (WS2012 R2)
  - Just pop in a new disk and it is seen as hot new capacity

Visualising Storage Spaces

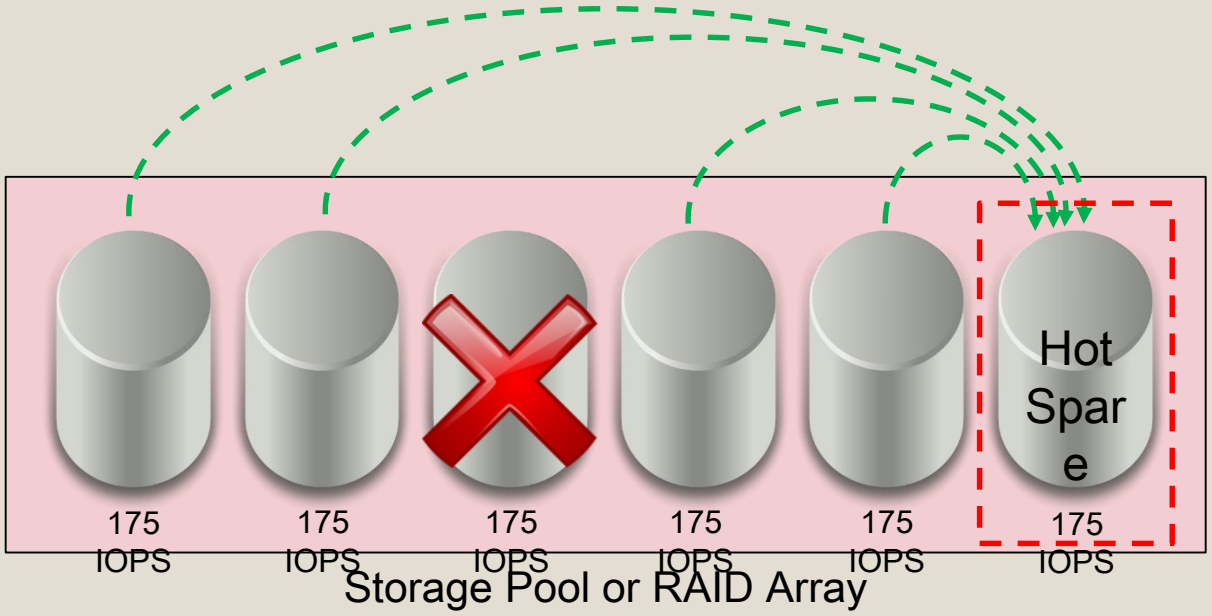


Hot Spares & Parallelized Restore

Hot Spares

- Supported in WS2012 and later Storage Spaces
- Hot Spares are old school … and slow to restore
- What happens when:
  - You are using RAID5 or parity disk
  - Disk fails
  - Hot spare restore kicks in
  - Another disk fails during the restore to the hot spare?
- You pray that backups worked

Hot Spare Restore



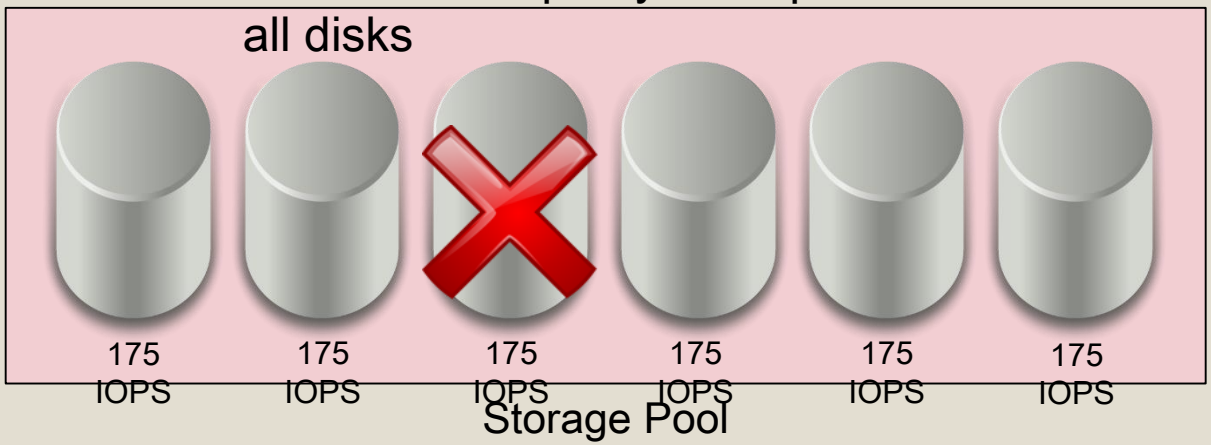
- 4 Disks (175 IOPS each) reading to restore data = 700 read IOPS
- Recovering to one disk with write speed of 175 IOPS

Too much opportunity for second disk failure and data loss

Parallelized Restore (WS2012 R2)

All Disks have copies of data via parity

Interleaves from failed disk restored from parity and spread to all disks



- 5 Disks (175 IOPS each) reading to restore data = 875 read IOPS
- Recovering to all disks = 875 write IOPS

Add new disk as blank capacity

Stacking JBOD Trays

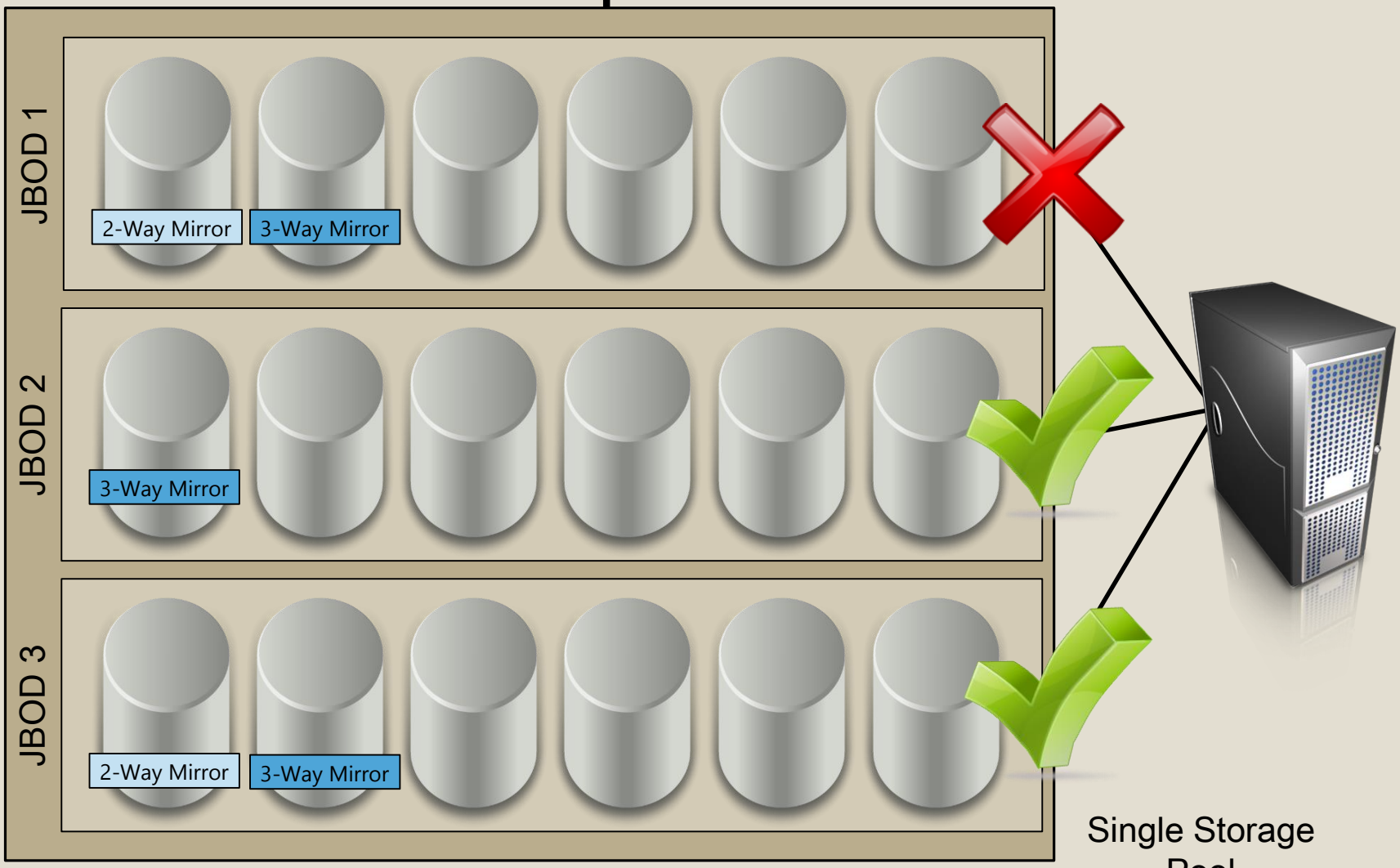
SAS Daisy Chaining

- You can daisy chain JBOD trays
- Add additional capacity
  - Start with one disk tray with up to 12, 24 or even 60 disk slots
  - Scale that out to 4 trays
  - 4 trays * 60 slots * 4 TB drives = .98 Petabyte of raw storage
- There is another benefit of having at least 3 stacked JBOD trays

Fault Tolerant JBODs

- People are worried about:
  - JBOD tray failing and taking all the data with it
  - Scaling beyond the capacity of a single JBOD tray
- You can connect multiple trays to a server
  - Adds storage capacity
  - Fault tolerant virtual disks are interleaved across multiple JBODs instead of just one
- To achieve JBOD fault tolerance
  - You require a minimum of 3 JBODs for cluster quorum

Multiple JBODs

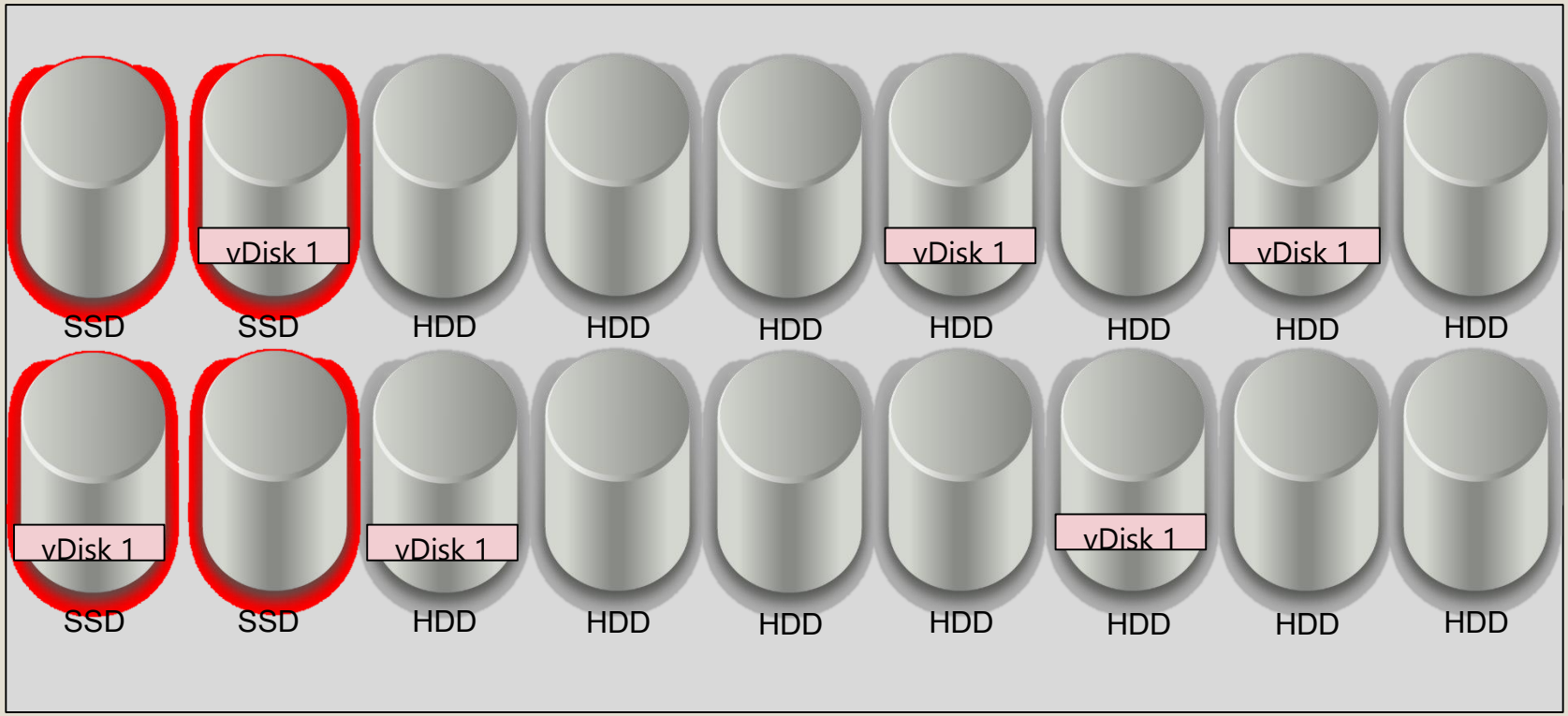


Tiered Storage

Tiered Storage Spaces

- The biggest question on Storage Spaces in WS2012:
  - "Can I tier SSD and HDD?"
  - In WS2012 Storage Spaces: No
  - In WS2012 R2 Storage Spaces: Yes
- Mix 1 tier of SSD with 1 tier of HDD
  - Get the speed of SSD
  - Get the capacity of HDD
- Heap map tracks 1 MB block usage
  - Default schedule optimizes block placement at 1am
  - Configurable (Task Scheduler) and on demand (PowerShell)
- Can pin entire files to the hot SSD tier

Tiered Storage Illustration



When creating Storage Pool:

- Choose which HDD & SSD disks to add

When creating Virtual Disks:

- Is it tiered or not?
- How much SSD/HDD space to use?

Tiered Storage Space Sizing

- Use Storage Tiering *only if you need it*
  - Server grade SSDs make PC SSDs look like penny sweets
- Storage Spaces has concept of “Columns”
  - How many parallel writes can be done to a virtual disk at once
- Example:
  - 12 * HDDs in a storage pool
  - 2-way mirror Virtual Disk has 6 columns
  - 3-way mirror Virtual Disk has 4 columns
- 2 SSDs + 20 HDDs
  - 2-way mirror Virtual Disk has 2 columns
  - Only 2 SSDs, and that’s not all that much write IOPS compared to lots of economic HDDs
- *Suggested* guesstimation technique:
  - Use 4-8 SSDs minimum
  - 10% of disks would be SSD

Write-Back Cache

Write-Back Cache

- Some applications (Active Directory) do “write-through”
  - Tell the storage system to disable write caching to avoid data corruption
  - Hyper-V does this – no hardware write caching
    - *Ever – even with battery backup – no exceptions*
- If you have Tiered Storage Spaces:
  - You have fast SSD
  - Leverage that SSD tier to absorb spikes in write activity
    - Not caching because there is an actual write-to-disk operation
  - Cold data will eventually be demoted to HDD tier

How Write-Back Cache Works

- By default, if you have tiered Storage Pool
  - 1 GB of SSD will be allocated to each new virtual disk for WBC
- You can:
  - Create a virtual disk without WBC cache
  - Create a virtual disk with a larger WBC
    - We hear that there is no point in doing this
- When write activity spikes, the SSD tier absorbs the increased writes
- The cached data is committed so no risk of data loss
- Data is demoted to cold tier

Summarising Storage Spaces

- Scalable storage
  - 4U JBOD with 60 * 4TB drives
  - Daisy chain 4 JBODs
  - 960 TB of raw storage
- Economic storage
  - No manufacturer lock-in for disk
  - 1 TB 7200 HDD Seagate via Amazon: $209.99
  - 1 TB 7200 HDD from “Server Company”: $619.00
- Performance with tiered storage & WBC