

Advancements in Hyper-V Storage

Todd Harris, Senthil Rajaram
Microsoft

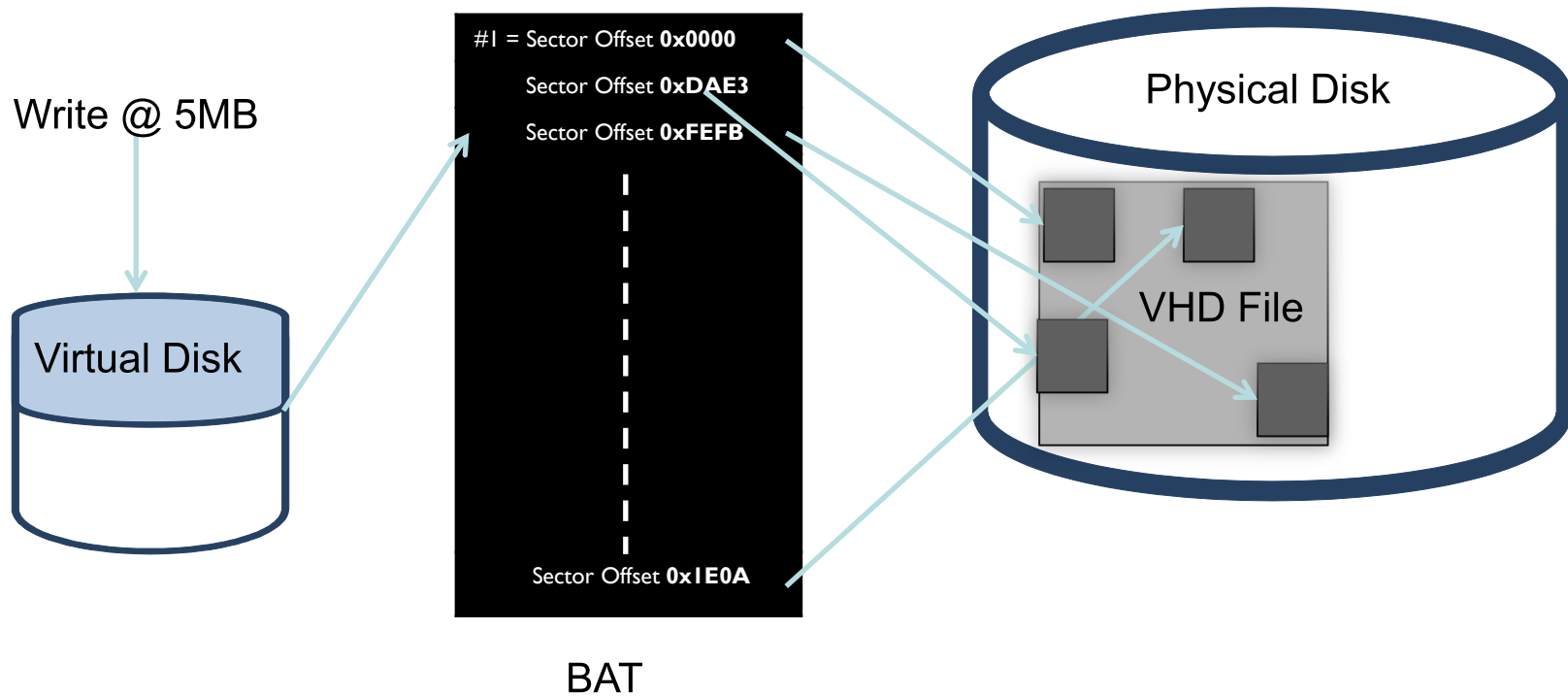
- Hyper-V storage stack changes for upcoming “Windows 8” release
 - VHDX
 - Online meta-operations
 - Hyper-V over SMB2.2
 - Virtual Fibre Channel support
 - Changes related to industry innovations
 - Offloaded Data Transfer(ODX) integration
 - Trim/Unmap integration
 - Large sector disk support

- ❑ Container format - encapsulate disks as files
 - ❑ VM
 - ❑ Files on the host OS
 - ❑ If attached to VM, appear as a disk to guest OS
 - ❑ Native mount
 - ❑ Files on the OS
 - ❑ If mounted, appear as a disk to the OS as well

- ❑ Three types – fixed, dynamic, differencing

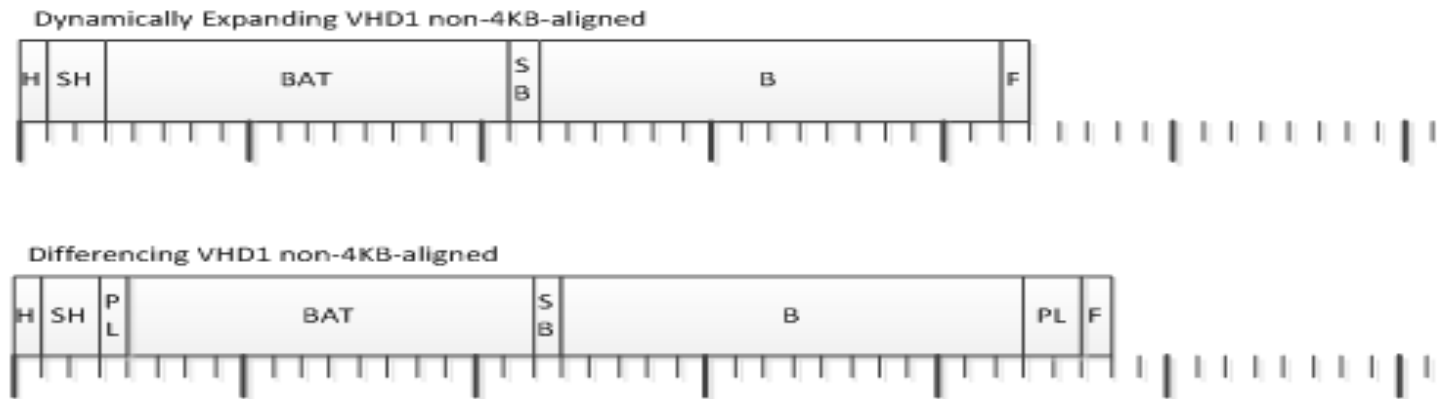
VHDX: Why?

□ VHD issue: 2TB limit



VHDX: Why?

- ❑ VHD issue: Sub-optimal alignment
 - ❑ Dynamic and Diff VHD format not 4K aligned



- ❑ Perf degradation on 4K disks

VHDX: Goals

- ❑ Keep format simple
 - ❑ Encourage wide adoption - similar to VHD
- ❑ Usable in all VHD scenarios
- ❑ Solve existing VHD format issues
 - ❑ Increase maximum virtual size beyond 2 TB
 - ❑ Better alignment of large structures (payload)
- ❑ Support emerging storage technologies
 - ❑ Advanced format disks, trim/unmap
- ❑ Performance parity, or better

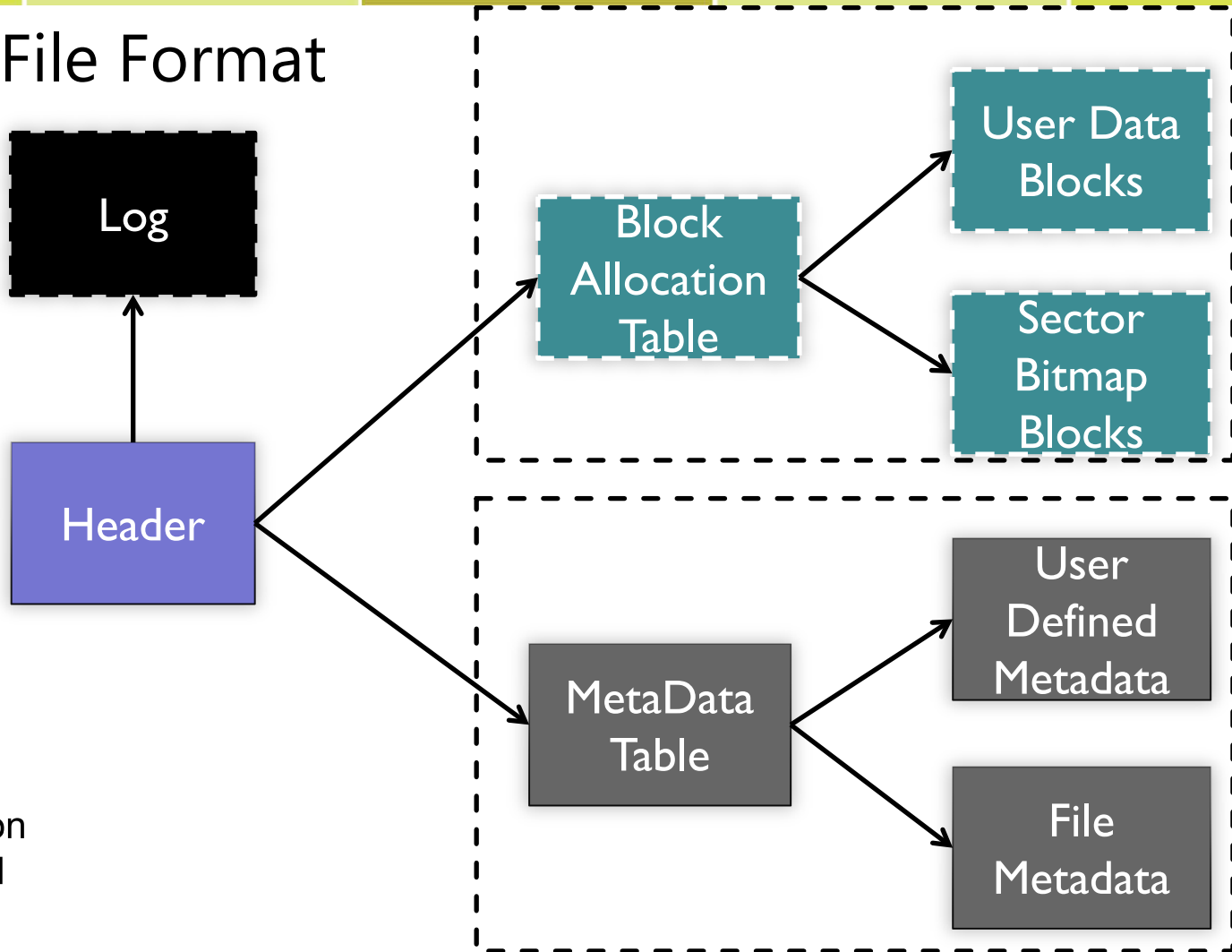
VHDX: Goals

- ❑ Enable new features/scenarios
 - ❑ Better resiliency to power failure events
 - ❑ Larger block sizes
 - ❑ Allow user to embed metadata in VHDX file

- ❑ Metadata updates are logged, enabling resiliency
 - ❑ Block allocations, block state changes, etc.
 - ❑ Payload data is not logged!
- ❑ Block sizes up to 256 MB
- ❑ User metadata supported
 - ❑ Key/value pairs
 - ❑ Up to 1024 entries, 1 MB per entry
- ❑ **All** internal I/O is 4k-aligned
- ❑ logical/physical sector size can be specified

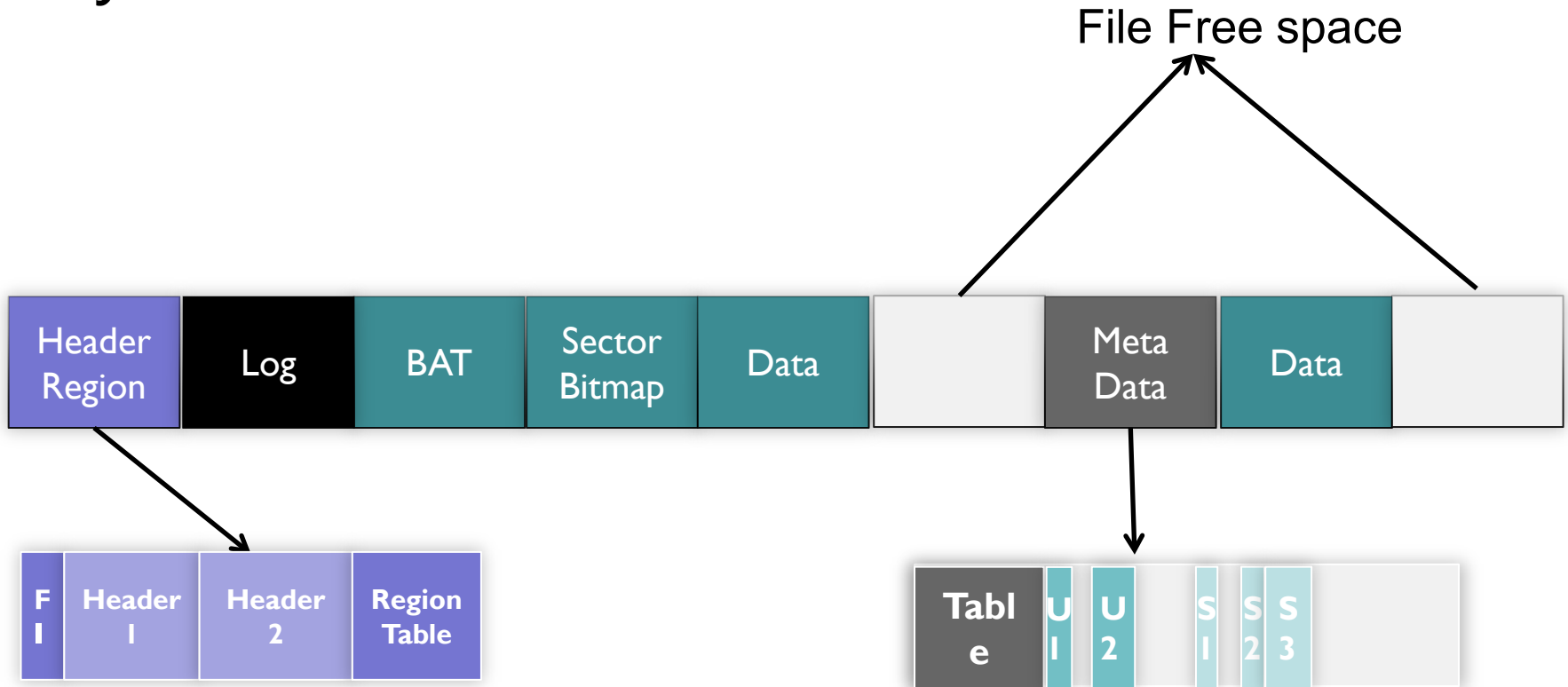
VHDX: Layout

Logical File Format



Large Allocation
& 1MB Aligned

Physical File Format



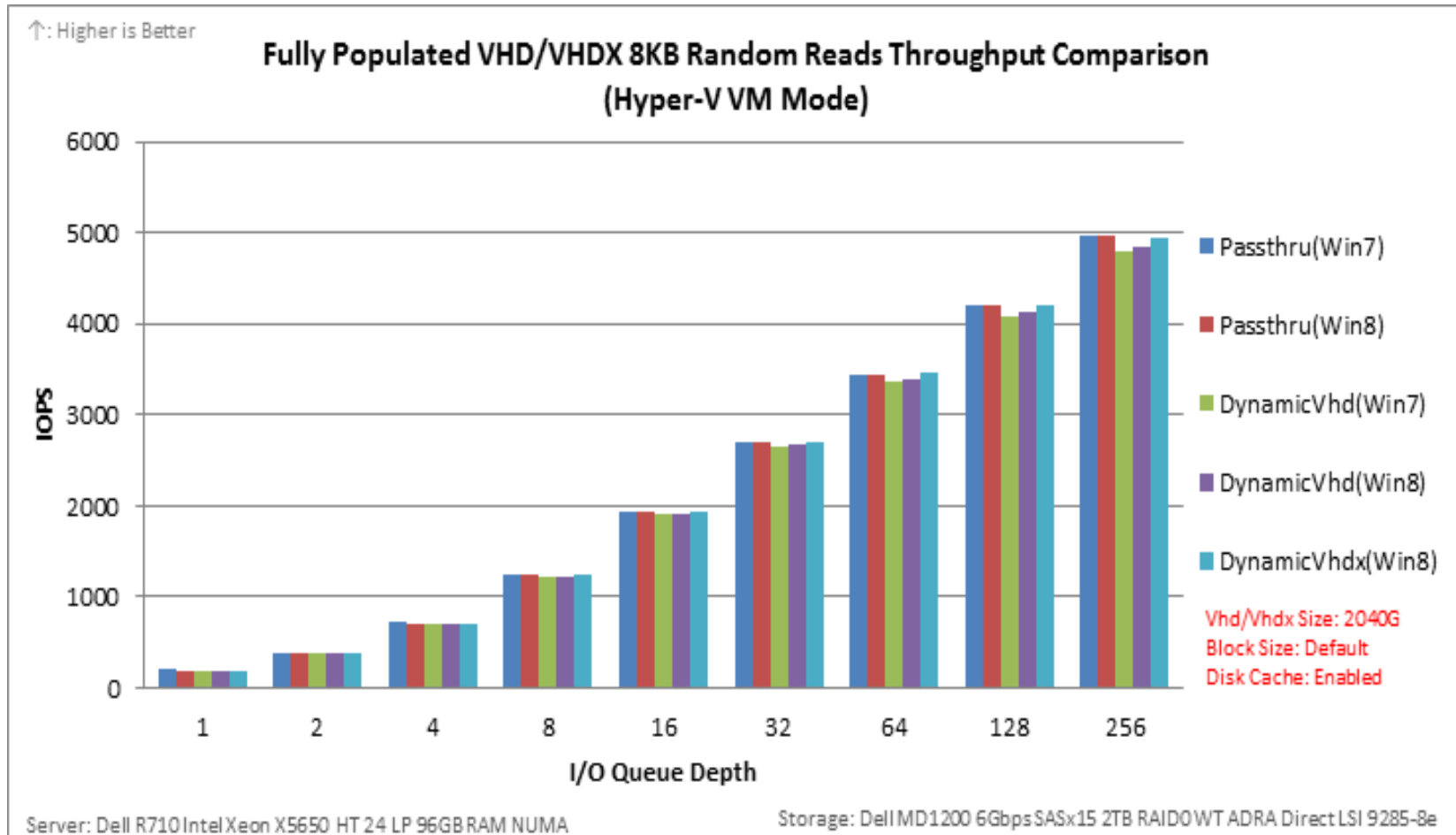
VHDX: Parameters

Parameter	Min	Max	Defaults	Notes
Virtual Size	3 MB	16 TB*	None	16 EB ~ (2 ⁶⁴)
Block size	1 MB	256 MB*	32MB* dyn 2MB* diff	
Logical sector size	512 B	4 KB	512 B	
Physical sector size	512 B	4 KB	4 KB	

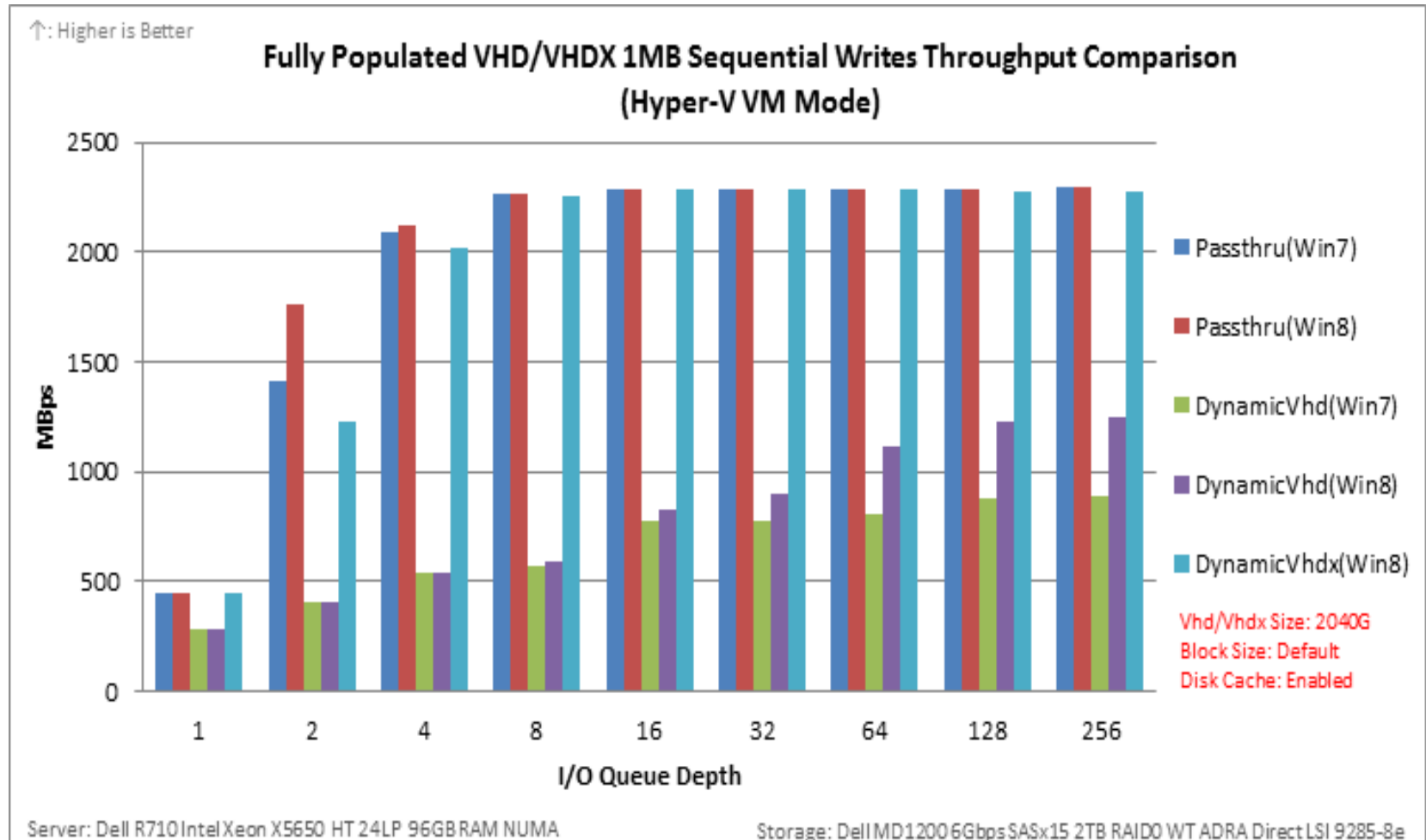
* Being tuned; not a file format limitation

- ❑ VHDX Not Supported On Versions < “Windows 8”
- ❑ Inbox tools for conversion between VHD & VHDX
 - ❑ API, UI & script support
- ❑ Mixed format differencing disks chain not allowed
 - ❑ i.e no child VHDX with parent VHD
- ❑ Format specification will be published

VHDX – Performance



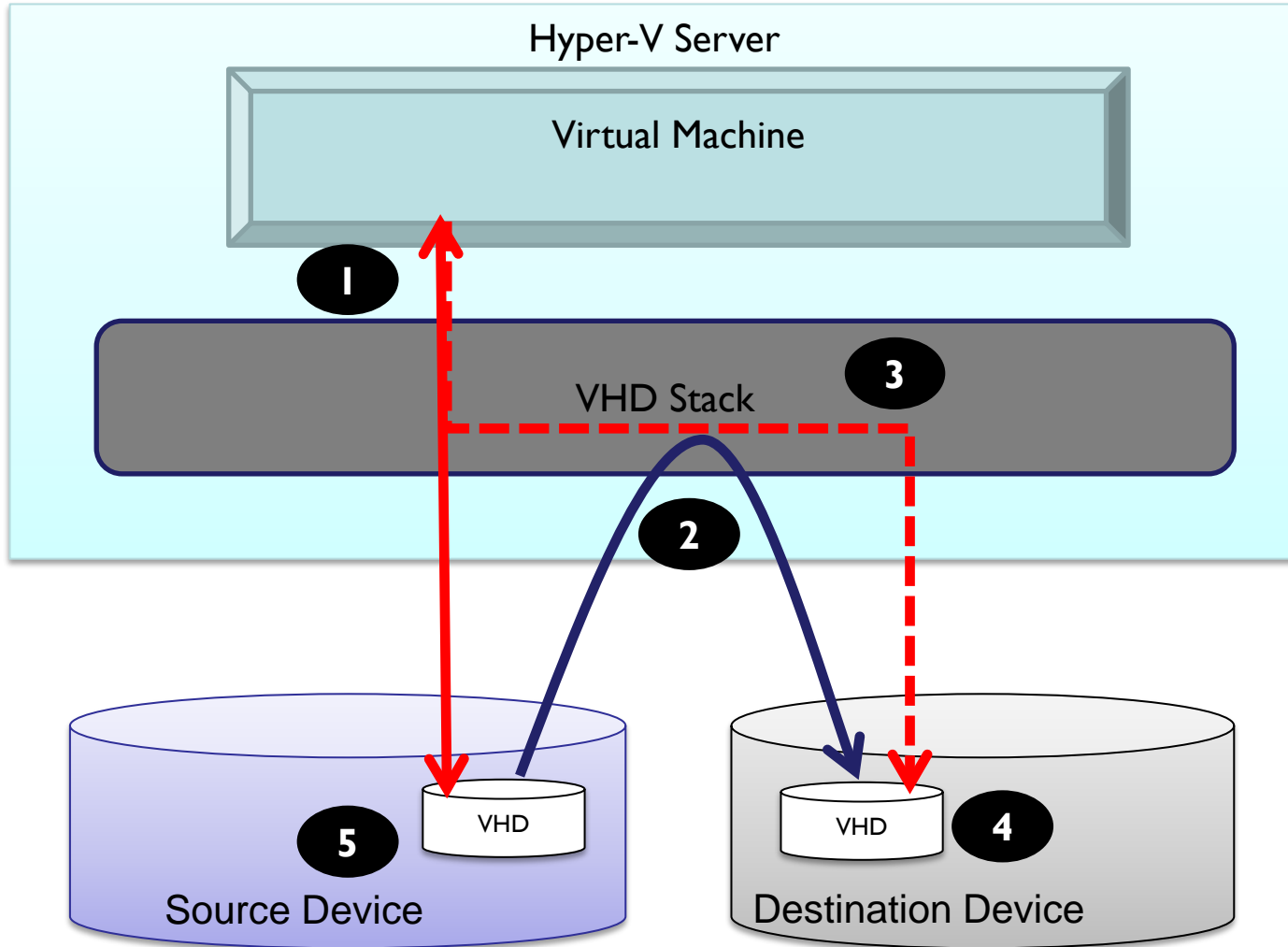
VHDX – Performance



Online Operations – Motivation

- ❑ VM storage maintenance currently means downtime
 - ❑ Moving VHDs
 - ❑ Evacuate failing storage, reorganize storage
 - ❑ Merging VHD chain
 - ❑ Reclaiming space from deleted snapshots
- ❑ Solution: online meta-operations
 - ❑ Mirror
 - ❑ Re-Parent
 - ❑ Merge

Online Operations – Mirror



- ❑ New scenarios enabled
 - ❑ Live storage migration
 - ❑ Quick provisioning
 - ❑ Start a VM from local diff
 - ❑ Copy master in the background
 - ❑ Re-Parent to the local master
 - ❑ Standalone server and cross-cluster live migration
 - ❑ Move VM storage to a network fileshare
 - ❑ Detach VM from source
 - ❑ Migrate & attach VM to destination

Hyper-V Network Storage Support

- ❑ Store all VM files on network fileshares
 - ❑ All Hyper-V operations possible - Snapshots, Save/Resume, SLP, Live migration etc..
- ❑ Customer value:
 - ❑ Leverage customers file storage investments
 - ❑ Enable new scenarios
 - ❑ Standalone server to Server VM Live Migration
 - ❑ Cross cluster live migration

Hyper-V & SMB2.2

- ❑ Supported configurations will need SMB 2.2
 - ❑ No blocks for < SMB2.2 versions, but
UNSUPPORTED

Network Failure Resiliency

- ❑ Resilient to network failures (P0)
 - ❑ Intermittent – transparently recoverable
 - ❑ SMB2.2 Resiliency
 - ❑ Permanent - transparently failover to another path
 - ❑ SMB2.2 Multichannel

File Server Failure Resiliency

- ❑ Resilient to file server failures (P0)
 - ❑ No disconnection of open handles & in-flight operations
 - ❑ SMB2.2 Continuous Availability

Hyper-V Node Failure Resiliency

- ❑ Seamless Clustered Hyper-V Failover(P0)
 - ❑ Storage available to failover node within blackout time (TCP timeout)
 - ❑ Satisfied by CSV – single namespace
 - ❑ Satisfied by SMB – common UNC path
 - ❑ No additional wait time for opening files on failover node
 - ❑ SMB2.2 Cluster Client Failover

- Host based VM backup support (P0)
 - VSS provider for remote file shares

- Performance - Similar to local storage
 - SMB2.2 Direct (RDMA)
 - SMB2.2 Multichannel

- ❑ Management - Hyper-V & SMB PowerShell
- ❑ Full permissions on SMB share and NTFS folders for
 - ❑ Hyper-V Host machine-accounts
 - ❑ VM administrators
- ❑ Remote Management of VMs: Double Hop Issue
 - ❑ Constrained Delegation or
 - ❑ Perform Operation from Hyper-V node

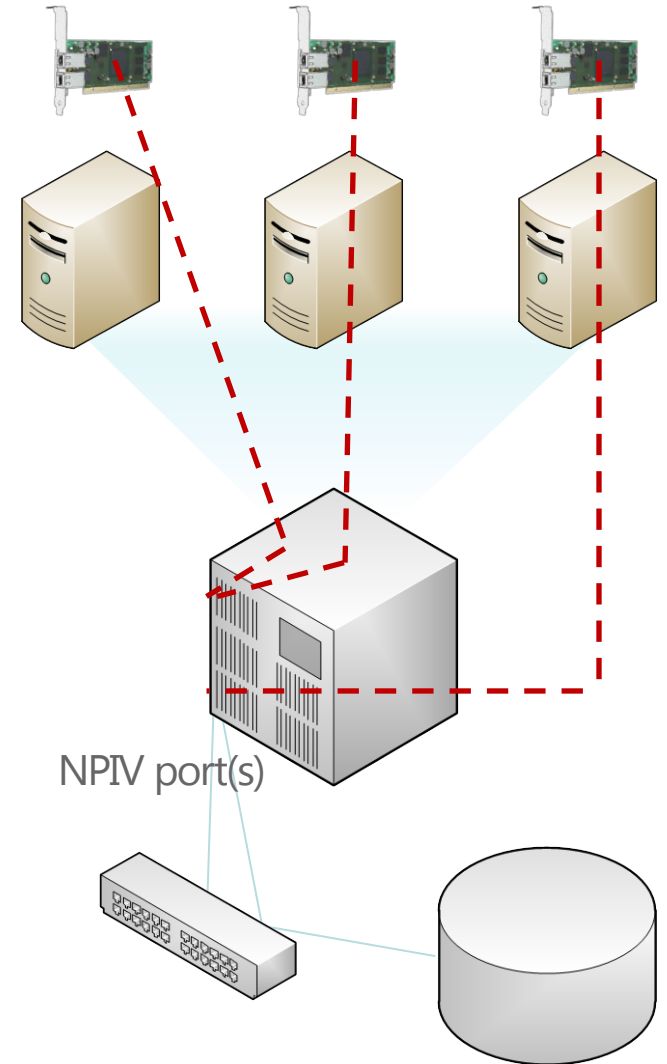
*<http://blogs.technet.com/b/josebda/archive/2008/06/27/using-constrained-delegation-to-remotely-manage-a-server-running-hyper-v-that-uses-cifs-smb-file-shares.aspx>

Fibre Channel – New Virtual Device

- ❑ Presents as a new FC port in the guest
- ❑ Direct VM access to the FC fabric
 - ❑ Use existing infrastructure and management
 - ❑ VM storage can be managed like physical storage
 - ❑ No CDB filtering
- ❑ Enables guest clustering of virtualized applications

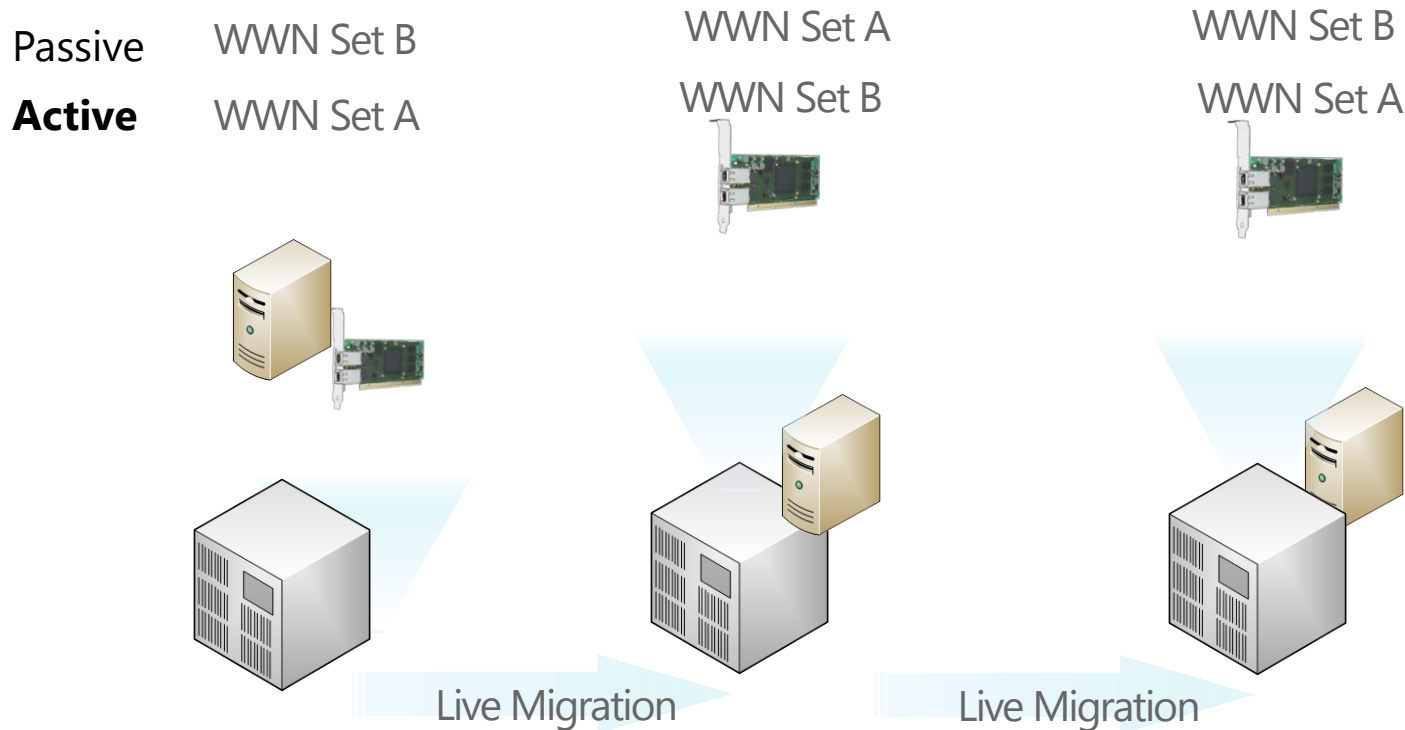
Fibre Channel – Architecture

- ❑ Up to four HBAs assigned to each guest
- ❑ WWNs assigned to each HBA
- ❑ NPIV utilized to surface guest ports on the host
- ❑ Multi-path I/O Support
 - ❑ Both guest & host can use MPIO
 - ❑ Guests might use different DSMs



Fibre Channel – Live VM Migration

- LM of a VM with FC HBA uses two sets of WWNs



Fibre Channel – Support & Limits

- ❑ Guest support for Windows Server 2008 and up
- ❑ Requires “Windows 8” server Hyper-V host
- ❑ Requires an updated NPIV HBA driver
 - ❑ Work in progress with HBA vendors

- ❑ Offloaded Data Transfer (ODX)
 - ❑ Token-based mechanism to offload copy to storage hardware
 - ❑ Reduces CPU and memory required on host
- ❑ Support in virtual storage stack
 - ❑ Decrease time required for VM maintenance
 - ❑ Speed up Merge, Mirror meta-operations
 - ❑ Increasing VM workload performance
 - ❑ Passing down ODX from guest to the host hardware

IA – Offload Implementation

- ❑ Mirror and merge operations fully utilize offload
 - ❑ Internal operations make offload calls directly
 - ❑ Will fall back to data read/write if offloads fail
 - ❑ Files copied during migration use OS support
- ❑ Offload operations from VMs are passed through
 - ❑ Mapping is non-trivial, due to file layout
 - ❑ Simple support for now—truncate to first extent
- ❑ Fixed VHD and VHDX creation time reduced
 - ❑ Well known token: “zero token”

- ❑ Trim/Unmap
 - ❑ System & Apps can inform storage stack of unused space
- ❑ Efficiencies at the virtual & physical storage layer
 - ❑ Pass down the Unmap from the guest to the physical hardware
 - ❑ As with offload, non-trivial
 - ❑ Allows explicit control of block state
 - ❑ Sub-block-sized unmaps are not tracked

IA – Trim/Unmap Implementation

- ❑ Only supported for VHDX and pass-through
- ❑ Only supported for Windows 8 Guests
- ❑ Storage optimizer sends Unmap on block boundaries
 - ❑ Allows TP disks to skip tracking sub-block unmaps
- ❑ Works for virtual disks on remote storage
 - ❑ Unmap flows over SMB
 - ❑ Virtual Stack uses FSCTL_FILE_LEVEL_TRIM

IA – Large Sector Size - Today

- ❑ Large sector size disks (Advanced format) are here
- ❑ 512e
 - ❑ Performance issues for VHDs due to RMW
- ❑ Native 4K
 - ❑ Hyper-V does not currently support these drives

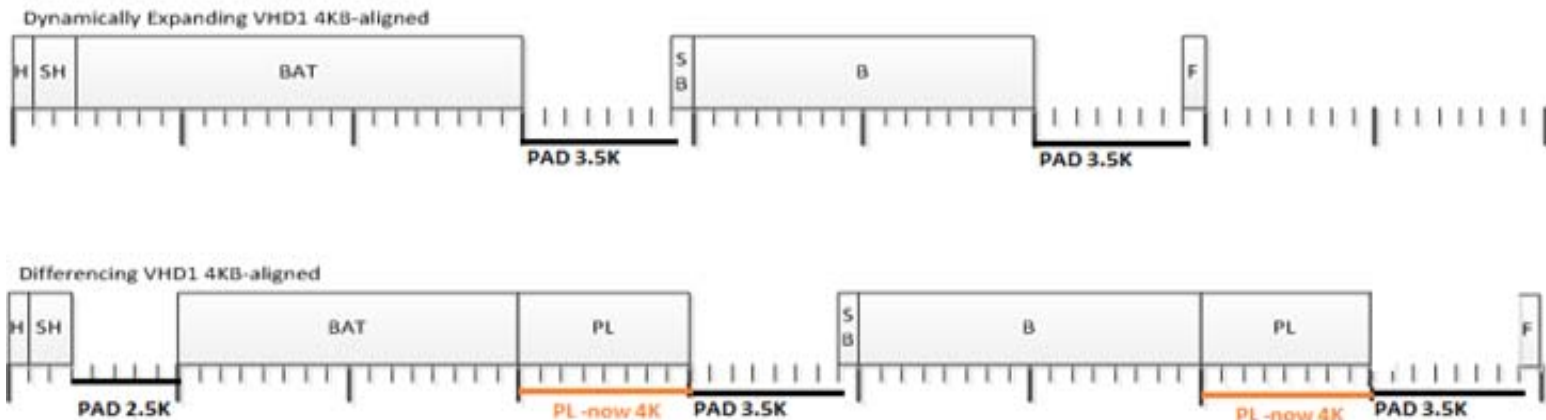
IA – 4K Native Support

- ❑ RMW layer implemented in the virtual storage stack
 - ❑ Allows virtual disks to work on 4K native disks

- ❑ VHDX is completely 4K-Aligned
- ❑ VHDX can present different sector size information based on metadata
 - ❑ 512 physical/logical
 - ❑ 4k physical / 512 logical (512e)
 - ❑ 4k physical/logical

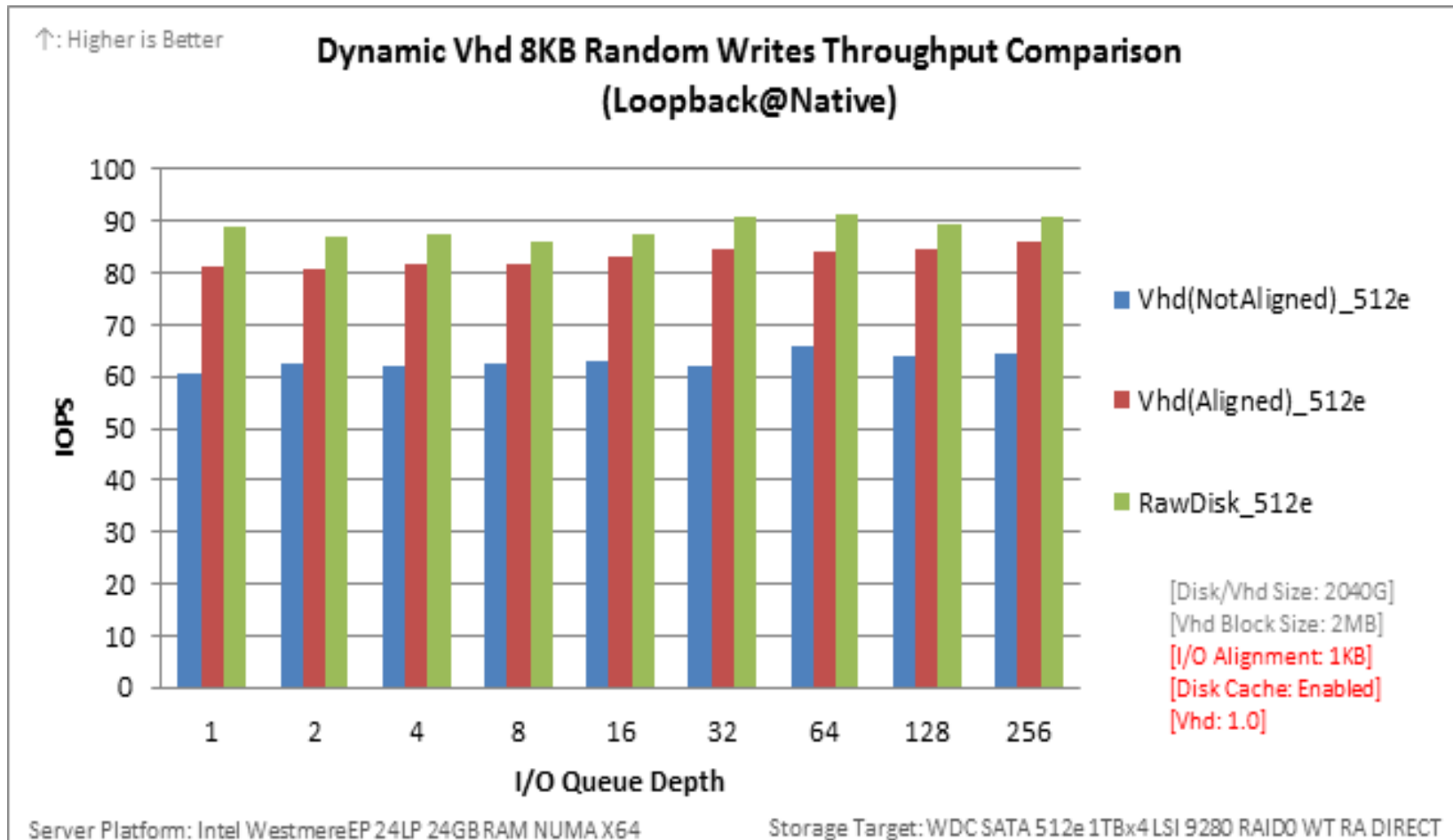
IA – 5 | 2e Support

- Non-fixed VHD is **nominally** not 4K Aligned
 - RMW Performance problem on 5 | 2e
- **New** non-fixed VHDs structures 4K Aligned
 - Within the spirit of the existing format

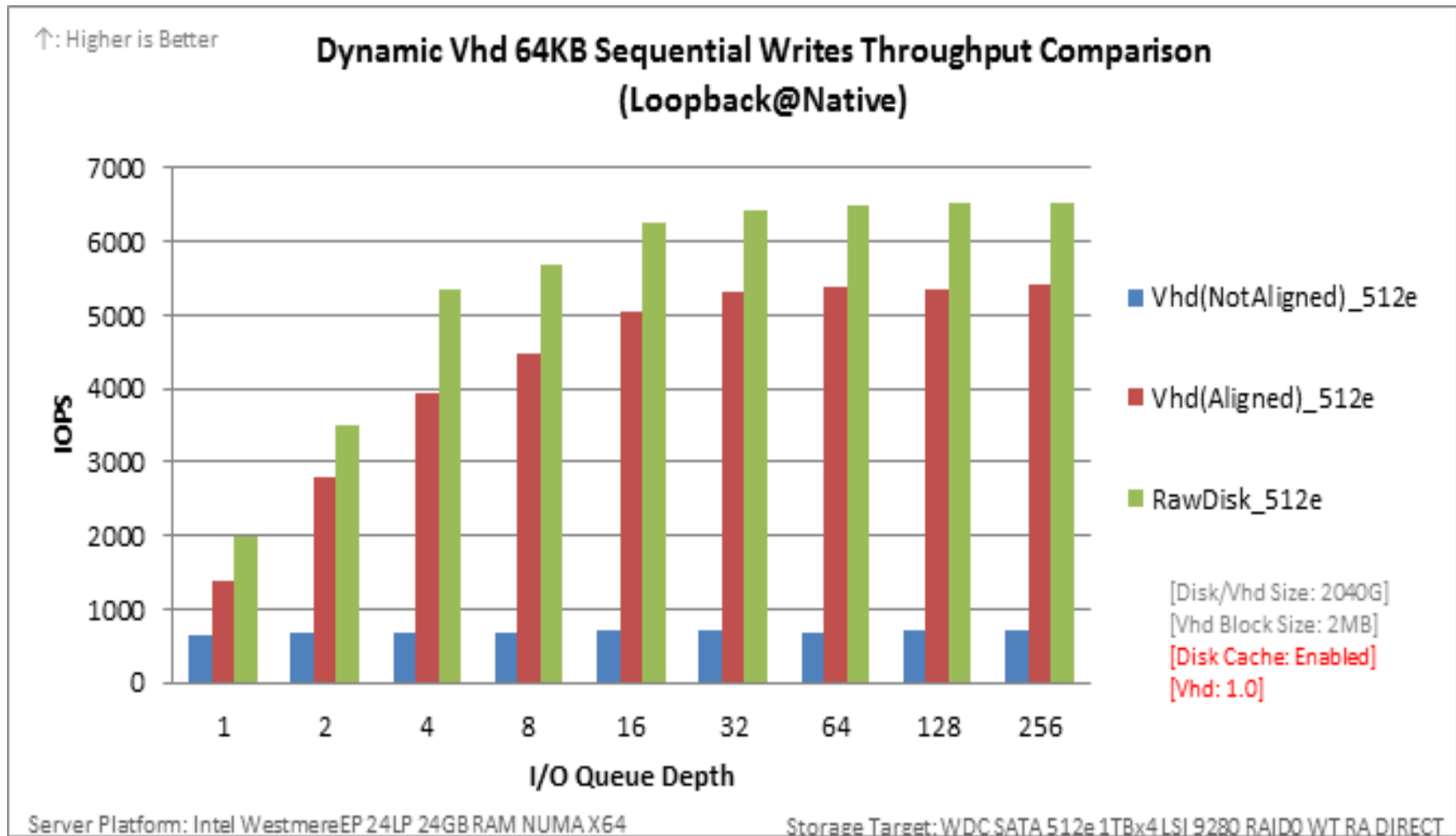


- ❑ Interop
 - ❑ Existing parsers should recognize new VHDs
 - ❑ Unaligned for new allocations
 - ❑ No auto-align for old VHDs on new parsers
- ❑ Only supported for 2 MB block size
 - ❑ Extra space for padding for 512KB block size takes it beyond 2TB file size
- ❑ VHDs report 512 logical/physical by default

IA – Improved perf on 512e



IA – Improved perf on 512e



Q&A