Microsoft File Sharing Protocols (SMB 3) Update

Tom Talpey

Architect, File Server team, Microsoft

Outline

- SMB 3.1.1
- Windows timeline and Document timeline(s)
- SMB3 Family News
- More speed! Less latency!

SMB 3.1.1 Feature Review

Since SDC (September 2014)

0. Dialect now 3.1.1

1. Extensible Negotiation
2. Preauthentication Integrity
3. Encryption Improvements
4. Cluster Dialect Fencing
5. Cluster Client Failover (CCF) v2
6. SMB3.x behaviors

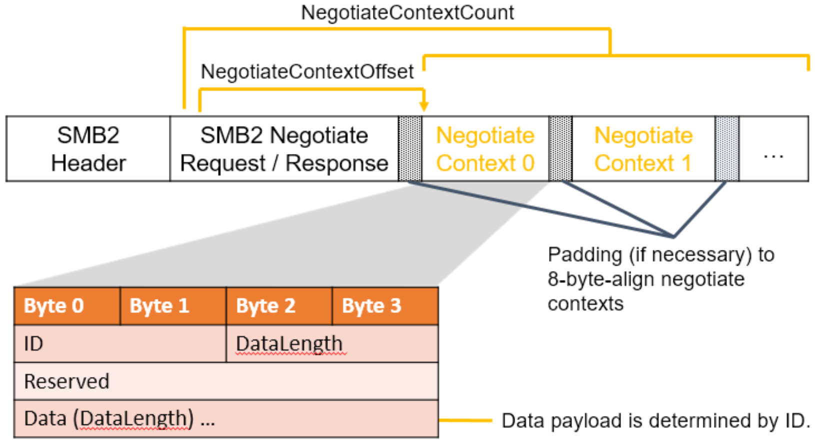
Dialect

- The Windows 10 SMB dialect will be **3.1.1**
    - At SDC time (Technical Preview 1) it was 3.1
    - 3.1 is now unsupported, and unimplemented – will be rejected
    - We expect, but cannot promise, that Windows Server 2016 will also be 3.1.1
- Minor number bump (3.**1**) is by design
    - We will likely continue this on future major cycles of Windows Server
- Sub-minor bump (3.1.**1**) needed due to Preview protocol updates
    - The dialect disambiguates packet formats and behavior changes from 3.1
- Dialects in documents and PowerShell now “Major.Minor.Sub”
    - Or simply Major.Minor if Sub==0
    - E.g. “2.002” is now “2.0.2”, “3.0” remains “3.0”, etc.
    - Simplified documentation and scripting

Extensible Negotiation (review)

SDC 2014 Review

- How to negotiate complex connection capabilities?
    - Very few unused bits left in the negotiate messages
- SMB 3.1.1 Extensible Negotiation
    - Exchange additional negotiate information via negotiate contexts (same idea as create contexts)
    - Repurpose unused fields in negotiate request / response as *NegotiateContextOffset* and *NegotiateContextCount* fields
    - Add list of negotiate contexts to end of existing negotiate messages
- UNCHANGED


NegotiateContextCount
NegotiateContextOffset
SMB2 Header
SMB2 Negotiate Request / Response
Negotiate Context 0
Negotiate Context 1
…
Padding (if necessary) to 8-byte-align negotiate contexts
Data payload is determined by ID.

Byte 0	Byte 1	Byte 2	Byte 3
ID		DataLength	
Reserved			
Data (DataLength) …			

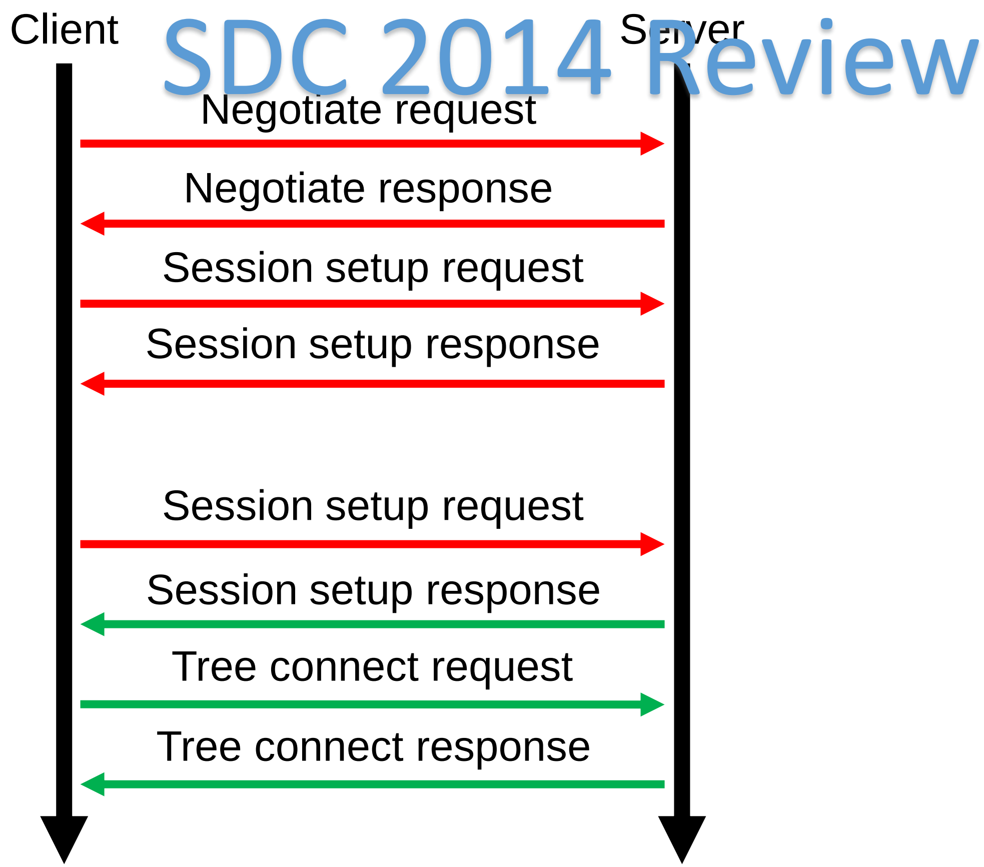
Key Points

SDC 2014 Review

- Client sends negotiate contexts if it supports the 3.1.1 dialect.
- Server sends negotiate contexts if it selects 3.1.1 as the connection’s dialect.
- Receiver must ignore unknown negotiate contexts.
- SMB 2/3 server implementations must be willing to accept negotiate requests that are larger than the SMB2_HEADER + SMB2_REQ_NEGOTIATE + Dialects array.
  - A client does not know whether a server supports SMB 3.1.1 before it negotiates, so must assume that it does and send negotiate contexts.
  - Windows accepts negotiate requests as large as 128 KiB

Preauth Integrity (review)

- How to protect messages from tampering prior to authentication?
  - No protection prior to SMB 3.0
  - SMB 3.0.x Negotiate Validation doesn't protect negotiate contexts or session setup messages
- SMB 3.1 Preauthentication Integrity
  - Provides end-to-end protection of preauthentication messages
  - Session's secret keys derived from hash of the preauthentication messages
  - Signature validation/decryption of subsequent authenticated messages will fail in case of preauthentication message tampering
- UNCHANGED



SDC 2014 Review

🟥 Unsigned/Unencrypted, now protected by hash

🟩 Signed/Encrypted

Key Points

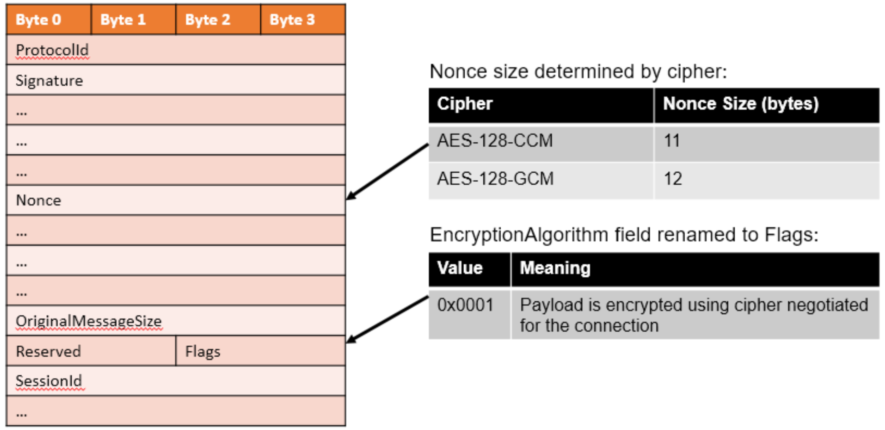
SDC 2014 Review

- Preauthentication Integrity is mandatory for SMB 3.1.1
- Session setup hashes are only calculated for master and binding session setup exchanges, not reauthentication
- Preauthentication Integrity supersedes SMB 3.0.x Negotiate Validation for SMB 3.1.1 connections
- Expect additional hardening based on security reviews over time
- Document significantly updated for clarity

Encryption Improvements (review)

SDC 2014 Review

- SMB 3.0.x mandates the AES-128-CCM cipher
  - What if a different cipher is required for performance, regulatory requirements, etc?
- SMB 3.1.1 Encryption Improvements
  - Ciphers are negotiated per-connection
  - Adding support for AES-128-GCM
  - Clients can mandate that sessions be encrypted even if the server does not require encryption.
- **Mostly** UNCHANGED



Byte 0	Byte 1	Byte 2	Byte 3
ProtocolId			
Signature			
...			
...			
...			
Nonce			
...			
...			
...			
OriginalMessageSize			
Reserved		Flags	
SessionId			
...			

Nonce size determined by cipher:

Cipher	Nonce Size (bytes)
AES-128-CCM	11
AES-128-GCM	12

EncryptionAlgorithm field renamed to Flags:

Value	Meaning
0x0001	Payload is encrypted using cipher negotiated for the connection

Client-mandated Encryption (change)

SDC 2014 Review

- No longer present:
  - **Client** mandates session encryption by setting the SMB2_SESSION_FLAG_ENCRYPT_DATA flag in its session setup request.
- Removed in 3.1.1
  - Not a complete solution to mandating encryption
  - Client can readily detect and reject server behavior, and decline to continue
  - Therefore, removed from protocol

Key Points

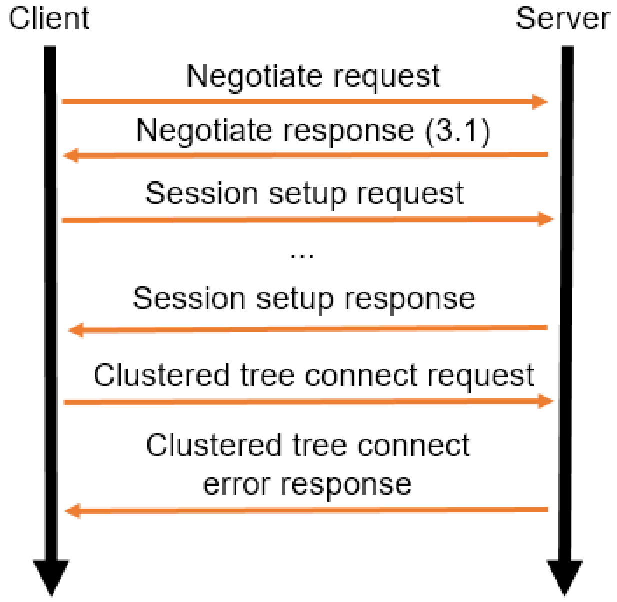
SDC 2014 Review

- AES-CCM required for SMB 3.0x compatibility.
- AES-GCM provides significant performance gains and should be supported.
- Session binding (multichannel) requires all of a session's channels to negotiate the same cipher as the session's original connection.
- ~~Client-mandated encryption depends on SMB 3.1 and Preauthentication Integrity to guarantee security.~~
  - ~~Not sufficient for client to simply send encrypted requests and verify encrypted responses.~~

Cluster Dialect Fencing (review)

SDC 2014 Review

- How to support clustered file servers whose nodes have different maximum SMB dialects (for example 3.0.2 vs. 3.1.1)?
  - Currently, all cluster nodes must support the same maximum SMB dialect to allow a client to transparently failover between cluster nodes.
- SMB 3.1.1 Cluster Dialect Fencing
  - Define a maximum SMB cluster dialect that all nodes support.
  - Fence access to cluster shares based on the maximum SMB cluster dialect.
  - Fenced clients instructed to reconnect at a cluster-supported dialect.
- UNCHANGED



Key Points

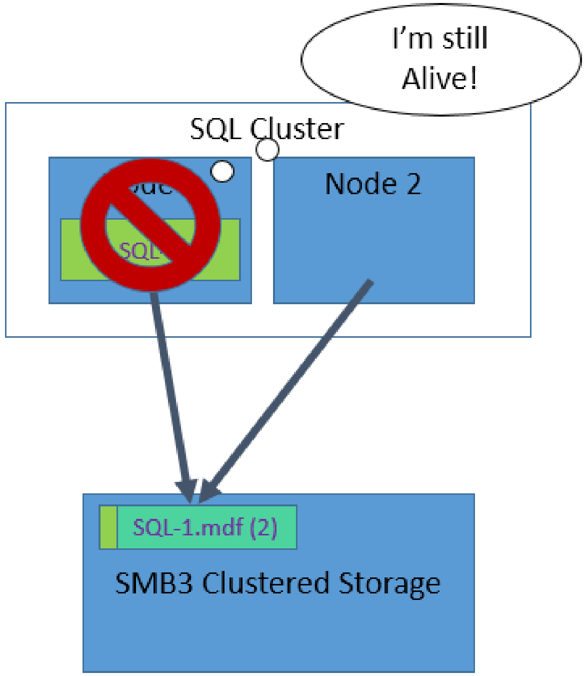
SDC 2014 Review

- Dialect fencing only affects clustered share access.
  - Clients can still access non-clustered shares using dialect X even if the maximum SMB cluster dialect is < X.
  - Can’t mix clustered and non-clustered access on same connection.
- Client implementation should protect against infinite loop of tree connect failure, disconnect, reconnect, tree connect failure, …

Cluster Client Failover v2 (review)

SDC 2014 Review

- Introduced with SMB 3 for clustered applications using SMB 3 storage
- Permits clustered application to tag an open with *ApplicationInstance* identifier
- An open issued by a different client with the same *ApplicationInstance* indicates workload has transitioned to the new node, so old opens are closed
- UNCHANGED



Key Points

SDC 2014 Review

- The CCF2 extension permits a client machine to keep your clustered application running during failover situations, but release it when the workload has been formally moved.
- Extending your storage cluster to support CCF2 is simple

Changes coming (review)

SDC 2014 Review

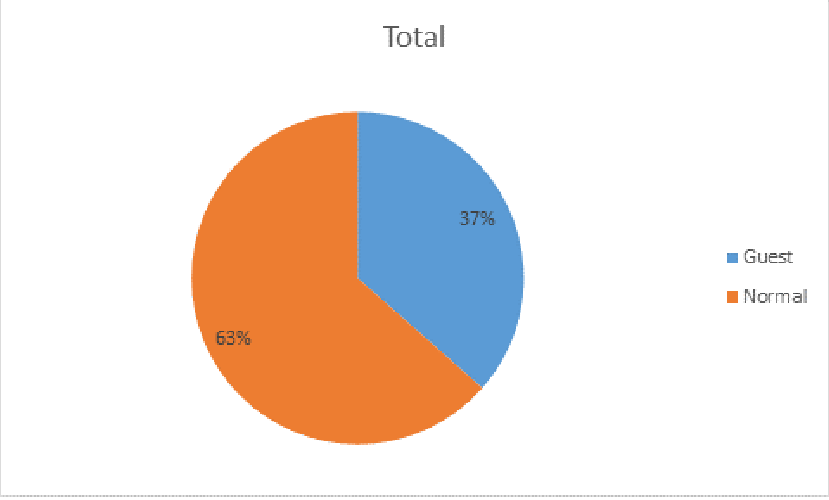
- ~~Removing RejectUnencryptedAccess setting~~
    - ~~Always reject clients that don’t support encryption when connecting to a server/share that requires encryption.~~
    - CHANGED RejectUnencryptedAccess – see the updated document
        - When a server receives an unencrypted request, and the server is configured to require encryption, RejectUnecryptedAccess is checked
        - Default of TRUE means all such client requests are rejected (legacy-free, most secure)
        - Setting to FALSE rejects encryption-capable SMB3.x but allows others to connect (legacy compatibility)
            - Defers enforcement until all legacy clients upgraded
- Removing RequireSecureNegotiate setting
    - Always perform negotiate validation if the connection’s dialect is 2.x or 3.0.x.
    - UNCHANGED

Changes #2 (review)

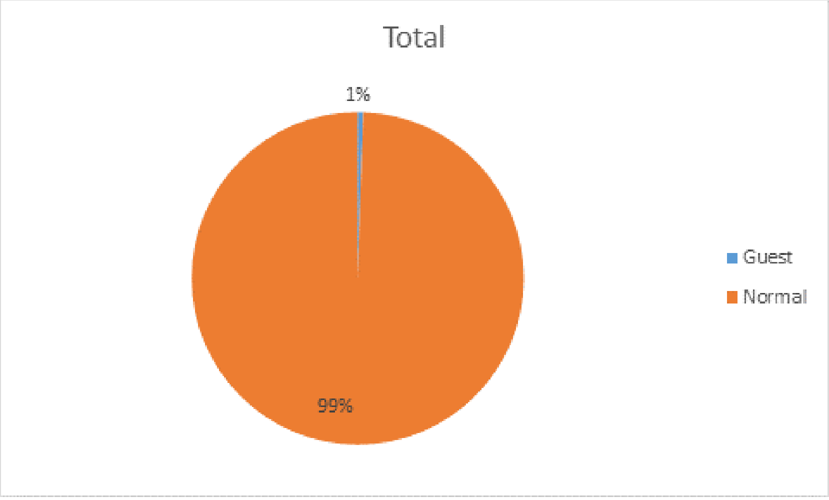
SDC 2014 Review

- ~~Restricting use of guest sessions~~
  - Indistinguishable from man-in-the-middle attack.
  - ~~Let us know ASAP if you have a scenario that requires guest logons using SMB 2.x or 3.x.~~
  - Interop experience indicates many third-party servers return SESSION_IS_GUEST, and most Samba-based (and many other) NAS server connections fail when clients reject guest with this
- Windows 10 Preview telemetry showed a high percentage of guest usage
  - Somewhat limited sample, used here for illumination
  - Guest usage still quite prevalent in Home/SMB1 settings (sigh) but less common in SMB2 (good!)
  - ☞ Added a setting (and an organization-wide group policy), but default to "off" for Windows 10

SMB1 Guest (Browser sessions removed)



SMB2/3 Guest (Loopback sessions removed)

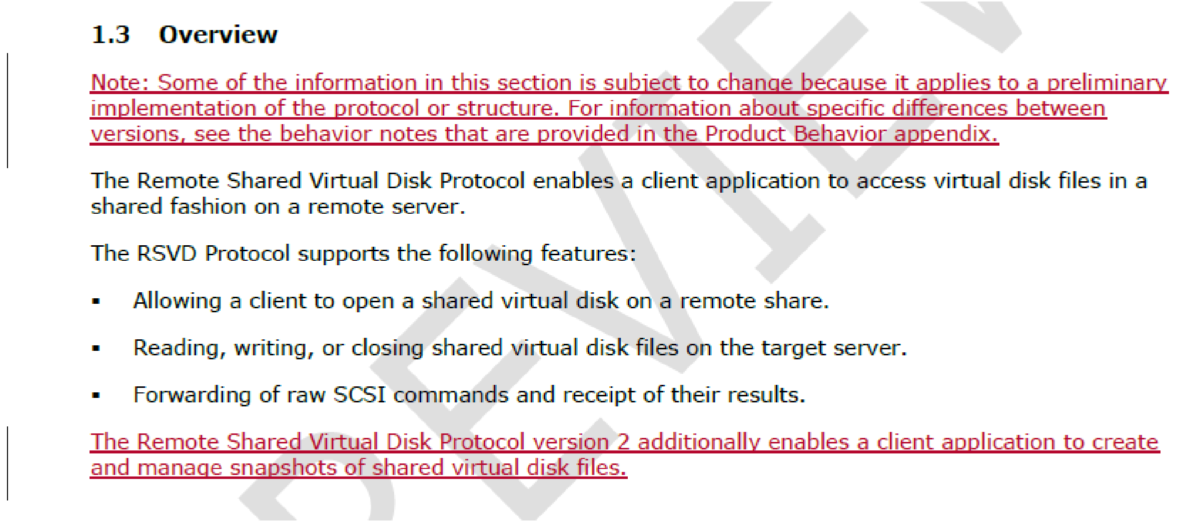


SMB2_TREE_CONNECT Change

- SMB 3.1.1 now requires a non-anonymous, non-guest TREE Connect to be **signed**
- Provides additional hardening of the NEGOTIATE/SESSION_SETUP/TREE_CONNECT initial exchange
- Non-strongly authenticated sessions still subject to previous checks
  - E.g. limited pipe access, etc

MS-RSVD Preview

- New MS-RSVD Version 2
  - Supports VHDX snapshots, VHD Sets
- New updated Preview
  - http://download.microsoft.com/download/C/6/C/C6C3C6F1-E84A-44EF-82A9-49BD3AAD8F58/Windows/[MS-RSVD-Diff].pdf



1.3 Overview

Note: Some of the information in this section is subject to change because it applies to a preliminary implementation of the protocol or structure. For information about specific differences between versions, see the behavior notes that are provided in the Product Behavior appendix.

The Remote Shared Virtual Disk Protocol enables a client application to access virtual disk files in a shared fashion on a remote server.

The RSVD Protocol supports the following features:

- Allowing a client to open a shared virtual disk on a remote share.
- Reading, writing, or closing shared virtual disk files on the target server.
- Forwarding of raw SCSI commands and receipt of their results.

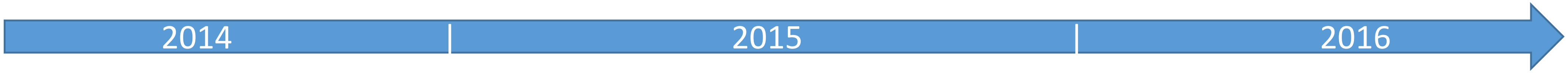
The Remote Shared Virtual Disk Protocol version 2 additionally enables a client application to create and manage snapshots of shared virtual disk files.

Protocol Documentation releases

- “Official” (full-support) documents are released:
  - Prior to any RTM of Windows
  - New Errata process replaces interim document releases
    - And provides more prompt resolution of Technical Document Issues
- “Preview” documents are released:
  - Prior to major Windows Technical Preview releases
  - Whenever new preview content is completed, significant, or “interesting”
    - Case-by-case basis, but we try to push the SMB family early and often
- However…
  - Windows and Windows Server releases are now becoming decoupled

Timeline

- ROUGH dates ☺
  - Windows 10 has 3 Technical Previews (“beta”), Release expected mid-2015
  - Windows Server 2016 has 2 Technical Previews, and a third expected sometime in 2015, with Release in 2016
- Windows Client and Server are no longer on the same release schedule
- Documentation will need to adapt



Windows 10	TP1	TP2	TP3	RTM	…	…		…
Windows Server 2016	TP1		TP2		TP3?		RTM	…

Preview/RTM “Mix”

- SMB 3.1 and 3.1.1 are Previewed since September 2014
- Windows 10 releasing with SMB 3.1.1
- Windows Server 2016 still in Preview, also implements 3.1.1
- ☞ Published MS-SMB2 will cover **RTM** Windows 10 and **Preview** Server
  - This is new, and “experimental”
  - Considered best way to simplify licensee document experience
  - Potentially may result in some confusing WBNs
- Bear with us, and give feedback if needed
- Still unclear how to manage future release interleave

SMB's Growing Family

- SMB3 (update in preview)
- Use by Cloud Platform System (CPS)
- Use by Storage Replica
- Use by Storage Spaces Direct
- Hyper-V related:
    - RSVD (update in preview doc)
    - Storage QOS (preview doc)
- Azure Files
    - Azure joining SMB family – with an Azure SMB2 server
        - New implementation, derived from MS-SMB2
        - Currently supporting SMB 2.1
        - http://azure.microsoft.com/en-us/services/storage/files/

Cloud Platform System (CPS)
Integrated solution for HW and SW

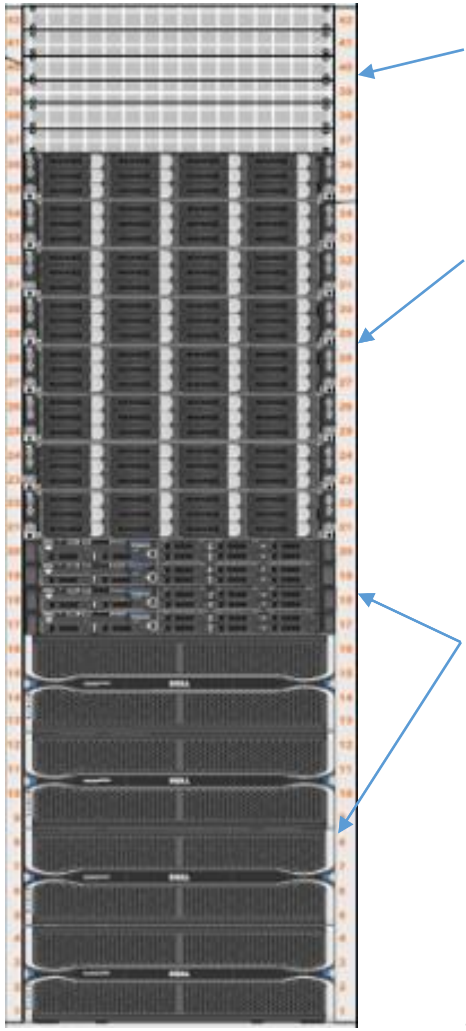
Per rack (1–4 racks)

512 cores
8TB RAM
262 TB usable storage

1360 Gb/s internal to rack
560 Gb/s inter-rack
60 Gb/s external

2322 lbs.
42U
16.6 KW maximum

Today's solution with
Windows Server 2012 R2
and System Center 2012 R2



Networking

5 x Force 10 – S4810P (64 port @ 10GbE – Data)

1 x Force 10 – S55 (48 port @ 1GbE – Management)

Compute scale unit (32 x Hyper-V hosts)

Dell PowerEdge C6220ii – 4 Compute Nodes per 2U

- Dual socket Intel IvyBridge (E5-2650v2 @ 2.6GHz), 256 GB memory
- 2 x 10 GbE Mellanox NIC's (LBFO Team, NVGRE offload) - tenants
- 2 x 10 GbE Chelsio (**iWARP/RDMA**) – storage **(SMB3 / SMB Direct)**
- 1 local SSD @ 200 GB (boot/paging)

Storage scale unit (4 file servers, 4 JBODs)

Dell PowerEdge R620v2 (4 Server for Scale Out File Server)

- Dual socket Intel IvyBridge (E5-2650v2 @ 2.6GHz)
- 2 x LSI 9207-8E SAS Controllers (shared storage)
- 2 x 10 GbE Chelsio T520 (**iWARP/RDMA**) **(SMB3 / SMB Direct)**

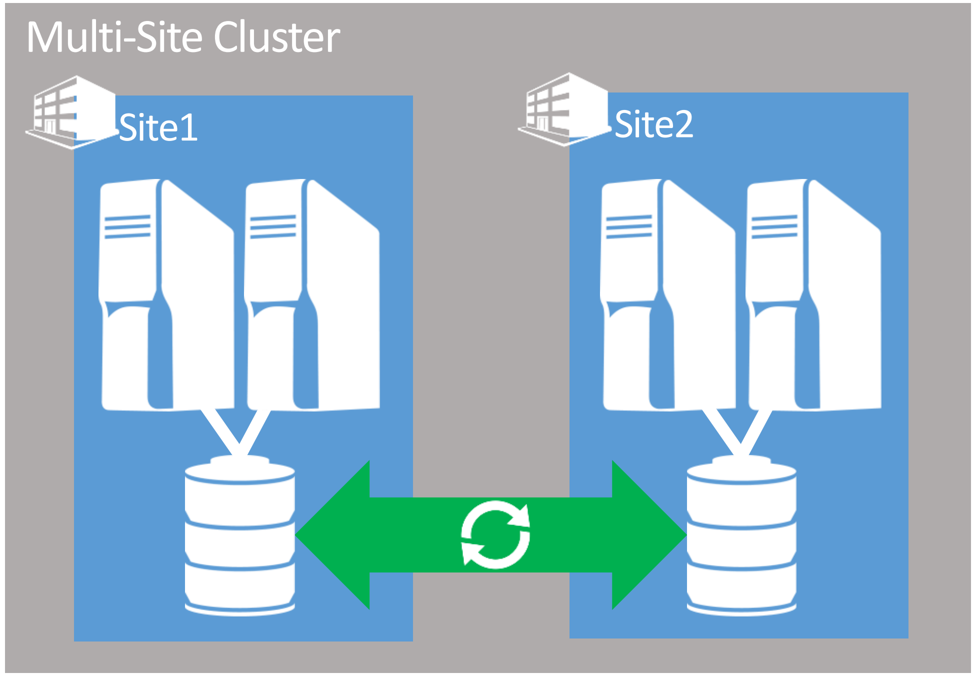
PowerVault MD3060e JBODs (48 SAS HDD, 12 SAS SSD)

4 TB HDDs and 800 GB SSDs

Storage Replica

Cross site High Availability Disaster Recovery: Stretch clusters across sites with synchronous volume replication

Integrated management	End-to-end Windows Server disaster recovery solution Failover Cluster Manager UI and PowerShell
Flexible	Works with any Windows volume, **uses SMB3 as transport** Hardware agnostic - works with Storage Spaces or any SAN volume
Scalable	Block-level synchronous volume replication Automatic cluster failover for low Recovery Time Objective (RTO)



Storage Spaces Direct

Software defined storage for private cloud using industry standard servers with local storage

Cloud design points and management

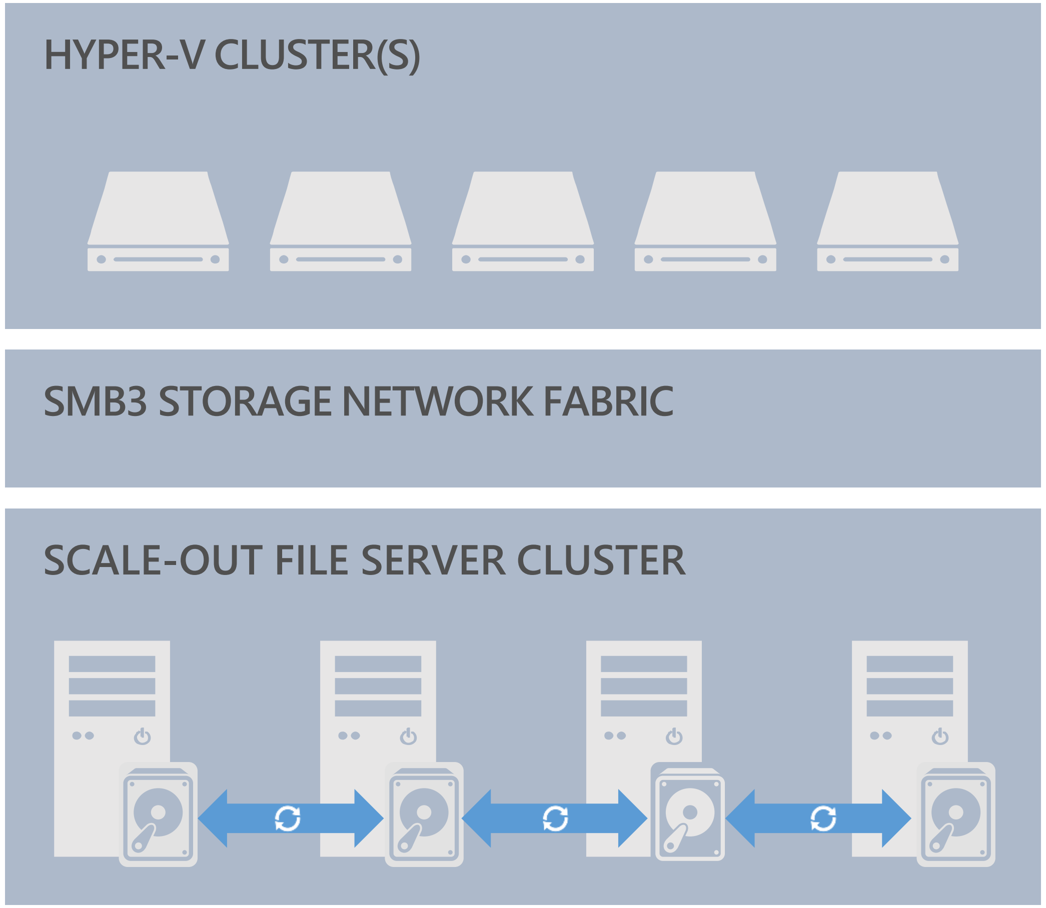
- Standard servers with local storage
- New device types such as SATA and NVMe SSD
- Prescriptive hardware configurations
- Deploy/manage/monitor with SCVMM, SCOM & PowerShell

Reliability, scalability, flexibility

- Fault tolerance to disk, enclosure, node failures
- Scale pools to large number of drives
- Simple and fine grained expansion
- Fast VM creation and efficient VM snapshots

Use cases

- Hyper-V IaaS storage
- Storage for backup and replication targets
- Hyper-converged (compute and storage together)
- Converged (compute and storage separate)



HYPER-V CLUSTER(S)

SMB3 STORAGE NETWORK FABRIC

SCALE-OUT FILE SERVER CLUSTER

Storage Quality of Service (QoS)

Control and monitor storage performance

Simple out of box behavior

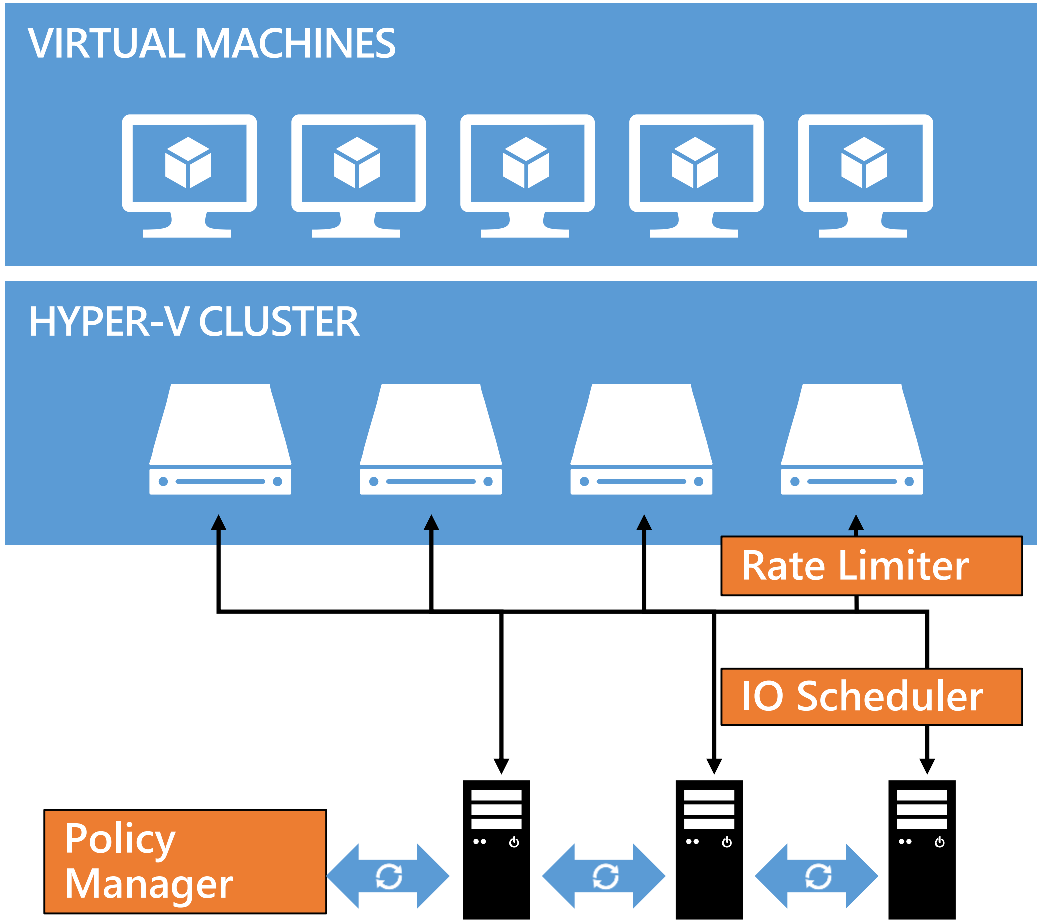
- Enabled by default for Scale Out File Server
- Automatic metrics per VHD, VM, Host, Volume
- Includes normalized IOPs and latency

Flexible and customizable policies

- Policy per VHD, VM, service, or tenant
- Define minimum and maximum IOPs
- Fair distribution within policy

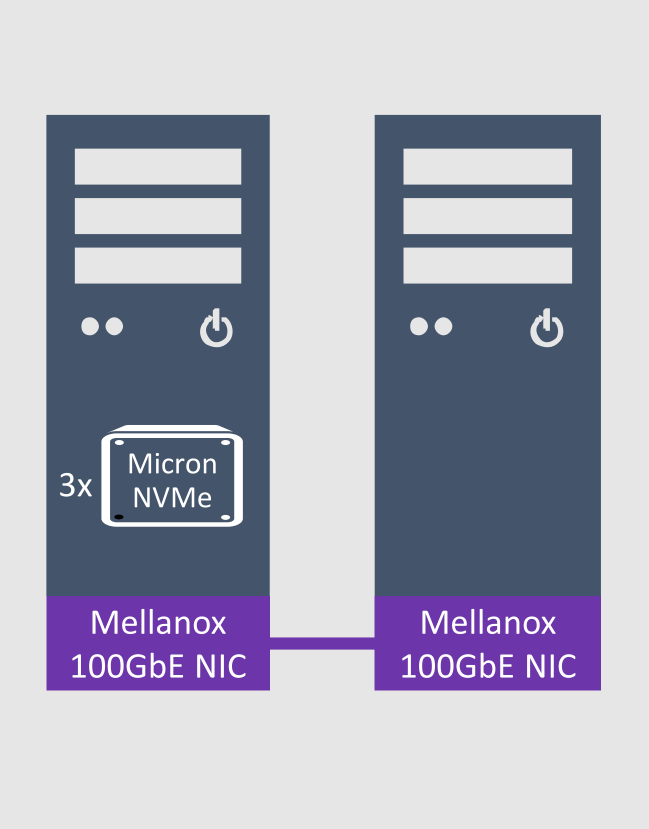
Management

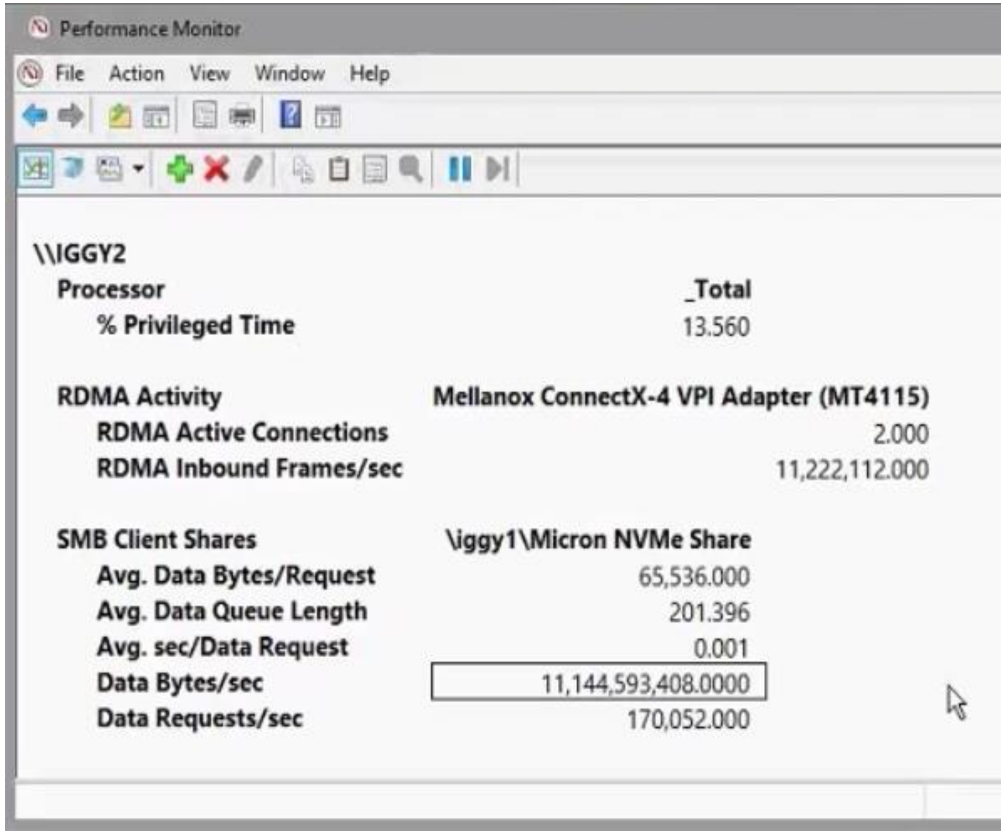
- System Center VMM and Ops Manager
- PowerShell built-in for Hyper-V and SoFS



Demo Summary: 100GbE and NVMe

A technology demonstration from May’s Microsoft Ignite





```
\\IGGY2
  Processor                                   _Total
     % Privileged Time                        13.560

  RDMA Activity               Mellanox ConnectX-4 VPI Adapter (MT4115)
     RDMA Active Connections                                    2.000
     RDMA Inbound Frames/sec                           11,222,112.000

  SMB Client Shares           \iggy1\Micron NVMe Share
     Avg. Data Bytes/Request                  65,536.000
     Avg. Data Queue Length                      201.396
     Avg. sec/Data Request                         0.001
     Data Bytes/sec                 11,144,593,408.0000
     Data Requests/sec                       170,052.000
```

Demo highlights:

- Storage Spaces using NVMe SSDs
- SMB3 using 100Gbps RDMA
- **Over 11Gbytes/sec** from one NIC port
- **1ms latency** with SMB3 storage
- Less than **15% CPU** utilization

Jose Barreto’s YouTube Channel or https://www.youtube.com/channel/UCLlf1kxGhvV3b15Q_V7jQJw

Emerging Ultra-Low Latency Storage Technologies

- Traditional block devices
  - HDD – latencies ~1’s msec
  - SSD – latencies ~100’s µsec
  - SMB3 well-matched to these
- NVMe – New high-performance storage interface
  - 2.5”, M.2, PCIe card, etc form factor
  - Block device semantics
  - Latencies ~**10’s µsec** (perhaps even <10 µsec)
  - SMB3 via SMB Direct still in the game
- PM - New class of “Byte-Addressable Storage”
  - Persistent Memory – DIMM form factor
  - Memory semantics, latencies: **<1 µsec**
  - Argues for new paradigm in SMB3 and RDMA

Technology	Latency (high)	Latency (low)	IOPS
HDD	10 msec	1 msec	100
SSD	1 msec	100 µsec	100K
NVMe	100 µsec	10 µsec (or better)	500K+
PM	< 1 µsec	(~ memory speed)	BW/size (>>1M/DIMM)

Note orders of magnitude decrease

SNIA NVM Programming

- SNIA Technical Working Group (TWG) for NVM programming
- Recently published white paper
  - NVM Programming Model v1.1
    - http://www.snia.org/tech_activities/standards/curr_standards/npm
- Ongoing work on Remote Access to PM
  - Numerous tracks at upcoming September SNIA SDC
- Watch this space for SMB3 and SMB Direct
  - And other protocols!

Resources

- http://www.microsoft.com/openspecifications
- http://smb3.info
- YouTube SMB demo videos: https://www.youtube.com/channel/UCLlf1kxGhvV3b15Q_V7jQJw
- MS-SMB2-Diff and MS-RSVD-Diff: http://msdn.microsoft.com/en-us/library/ee941641.aspx
- The Rosetta Stone: http://blogs.technet.com/b/josebda/archive/2015/04/30/smb3-networking-links-for-windows-server-2012-r2.aspx
- SMB3 News
  - http://blogs.technet.com/b/josebda/archive/2015/05/05/what-s-new-in-smb-3-1-1-in-the-windows-server-technical-preview-2.aspx
  - http://blogs.technet.com/b/josebda/archive/2015/04/21/the-deprecation-of-smb1-you-should-be-planning-to-get-rid-of-this-old-smb-dialect.aspx
- Storage QOS
  - http://channel9.msdn.com/Events/Ignite/2015/BRK3504
  - http://blogs.technet.com/b/josebda/archive/2015/05/06/windows-server-2016-technical-preview-2-tp2-and-storage-quality-of-service-qos.aspx
- SNIA SMB presentation (updated)
  - http://www.snia.org/sites/default/files2/DSI2015/presentations/FileSystems/JoseBarreto_SMB3_remote%20file%20protocol.pdf

Resources – Microsoft Ignite

- May 2015, presentations, demos, etc. relevant to topics above
- http://ignite.microsoft.com/Sessions

Code	Session title	Presenters
BRK3496	Deploying Private Cloud Storage with Dell Servers and Windows Server vNext	Claus Joergensen, Shai Ofek, Syama Poluri
BRK3474	Enabling New On-premises Scale-Out File Server with Direct-Attached Storage	Claus Joergensen, Michael Gray
BRK3489	Exploring Storage Replica in Windows Server vNext	Ned Pyle
BRK3504	Hyper-V Storage Performance with Storage Quality of Service	Jose Barreto, Senthil Rajaram
BRK2458	Overview of Microsoft Azure Storage and Key Usage Scenarios	Vamshidhar Kommineni
BRK2472	Overview of the Microsoft Cloud Platform System	Vijay Tewari, Wassim Fayed
BRK2485	Platform Vision & Strategy (4 of 7): Storage Overview	Jose Barreto, Siddhartha Roy
BRK3487	Stretching Failover Clusters and Using Storage Replica in Windows Server vNext	Elden Christensen, Ned Pyle

Questions

ttalpey@microsoft.com