The SC11 Conference
11/12-18/2011, Seattle

SMB 2.2 Over RDMA

Dan Lovinger
Principal Architect
Microsoft Corporation

Agenda

WHO WILL BENEFIT FROM THIS TALK

- System Integrators
- App Developers

TOPICS

- Overview of the "Windows 8" File Server
- Features
  - SMB2 Transparent Failover
  - SMB2 Scale Out
  - SMB2 RDMA

WHAT YOU'LL LEAVE WITH

- Better understanding of how "Windows 8" File Server features light up new platform capabilities

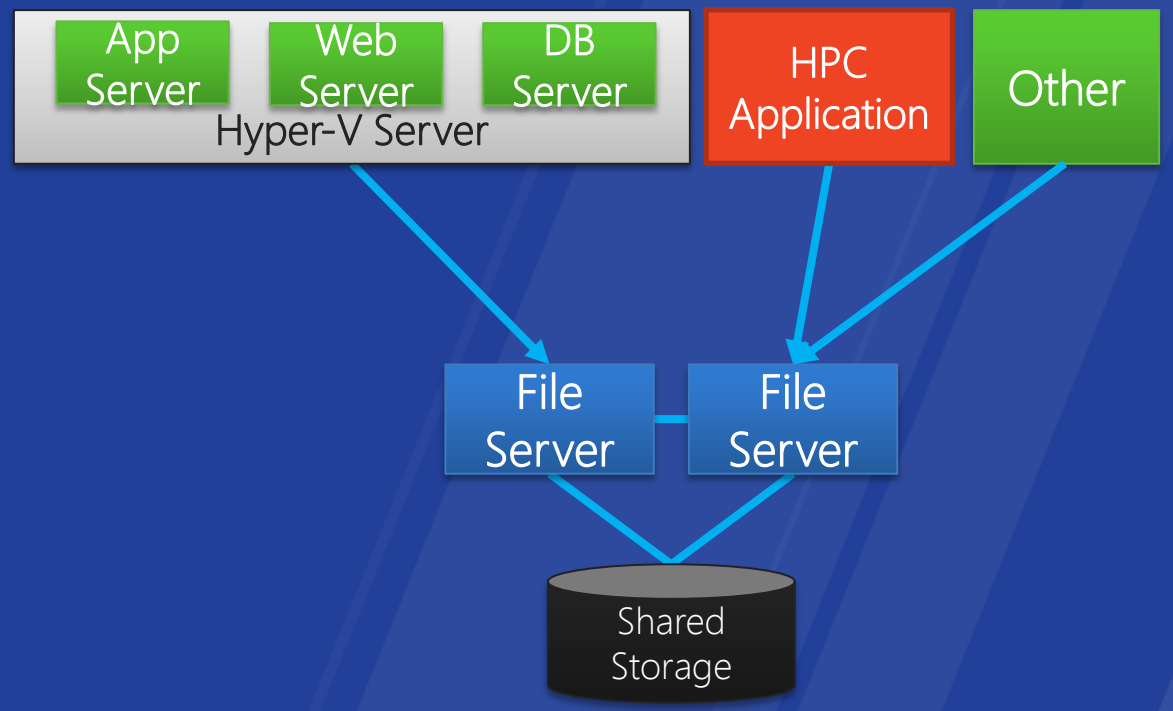
Remote File Storage for Server Applications

Remote File Storage for Server Applications

What is it and why?

- What is it?
  - Server applications storing their data files on SMB2 file shares (UNC paths)
  - Examples:
    - Hyper-V: Virtual Hard Disks (VHD),configuration files, snapshots etc.
    - HPC: application state, configuration, archive, etc.
- What is the value?
  - Easier provisioning – shares instead of LUNs
  - Easier management – shares instead of LUNs
  - Flexibility – dynamic server relocation
  - Leverage network investments – no need for specialized storage networking infrastructure or knowledge
  - Lower cost – Acquisition and Operation cost

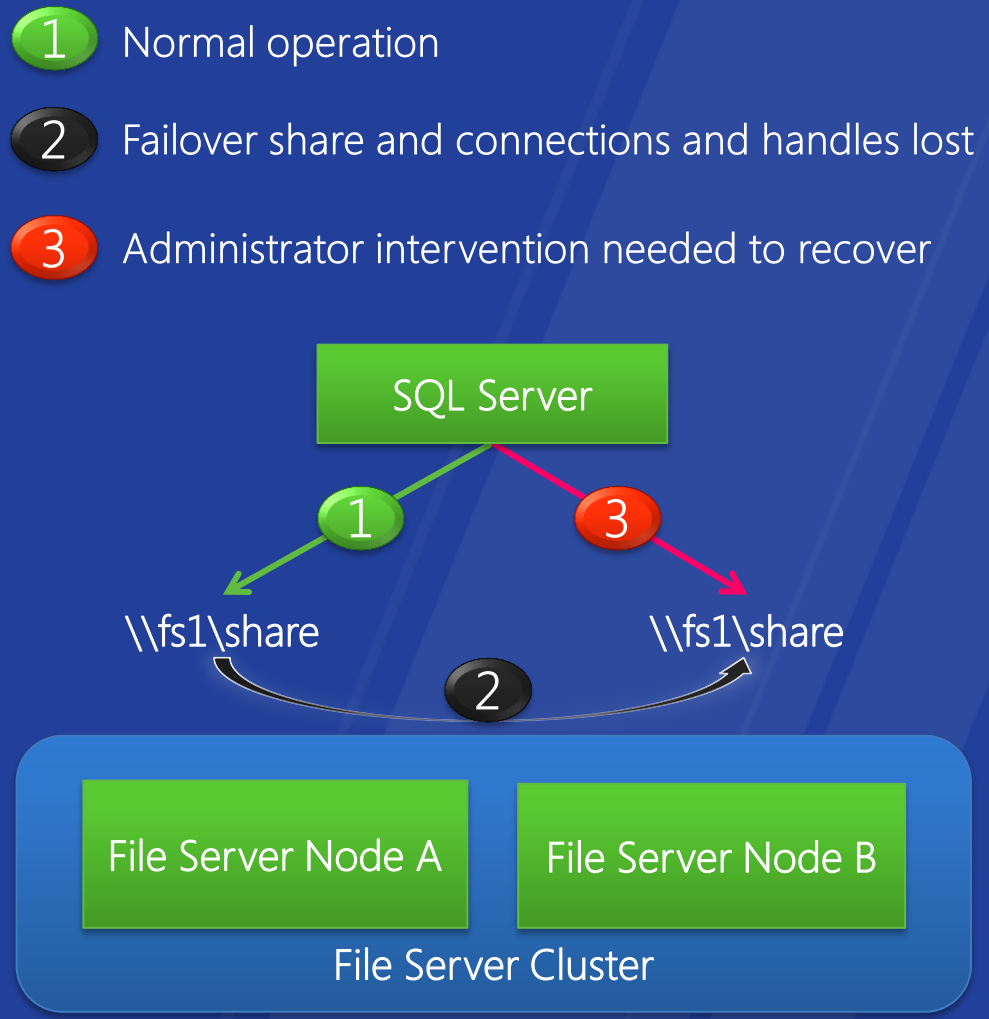
- Example:



Historical - Windows Server 2008 R2

Failovers are not transparent

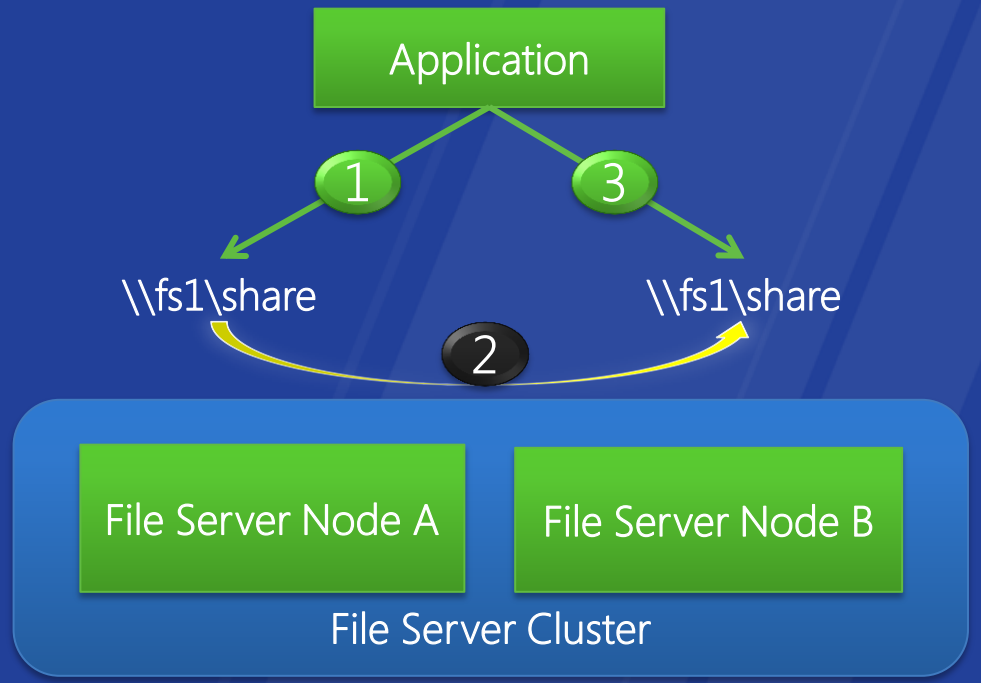
- Works great for traditional file server use scenarios
- Server applications expect storage to be continuously available
- With current solution connection and file handles are lost on share failover ->
  - Application disruption
  - Administrator intervention required to recover



"Windows Server 8"
SMB2.2 Transparent Failover

- Failover transparent to server application
  - Zero downtime – small IO delay during failover
- Supports planned and unplanned failovers
  - HW/SW Maintenance
  - HW/SW Failures
  - Load Rebalancing
- Resilient for both file and directory operations
- Requires:
  - Windows Failover Clusters
  - Both server running application and file server cluster must be "Windows Server 8"

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



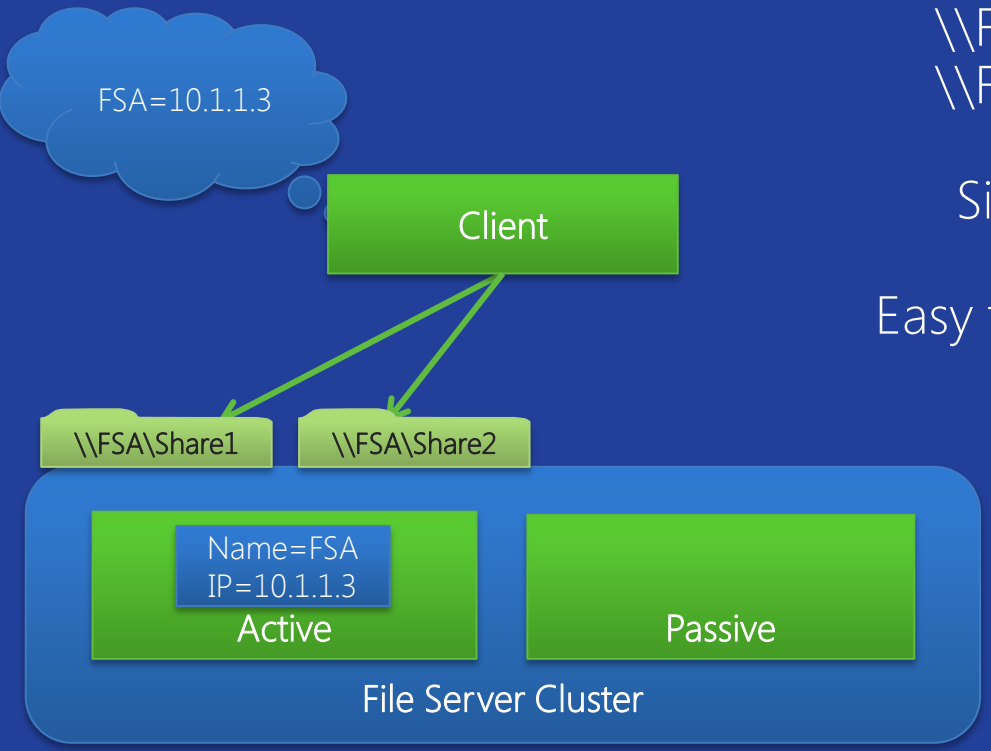
Historical: Windows Server 2008 R2

Active-Passive Single File Server

1 logical file server
1 virtual IP address
Active/Passive

\\FSA\Share1
\\FSA\Share2

Single name
Simple
Easy to manage

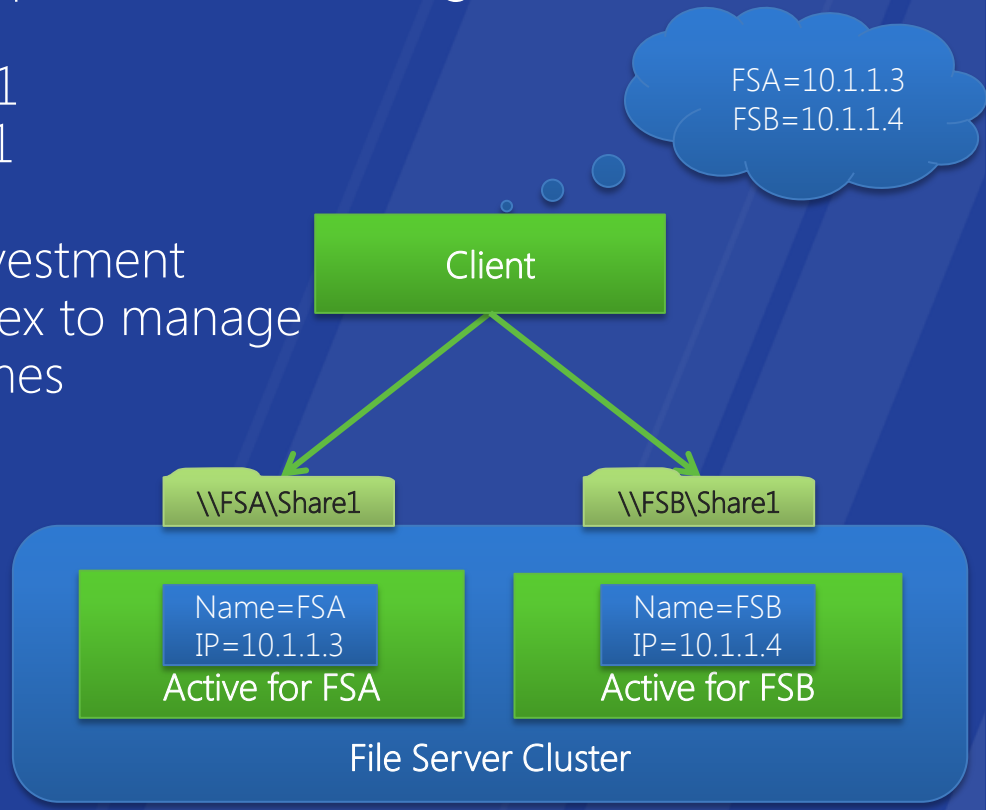


Active-Passive Multiple File Servers

2+ logical file servers
2+ virtual IP addresses
Access to disparate shares through different nodes

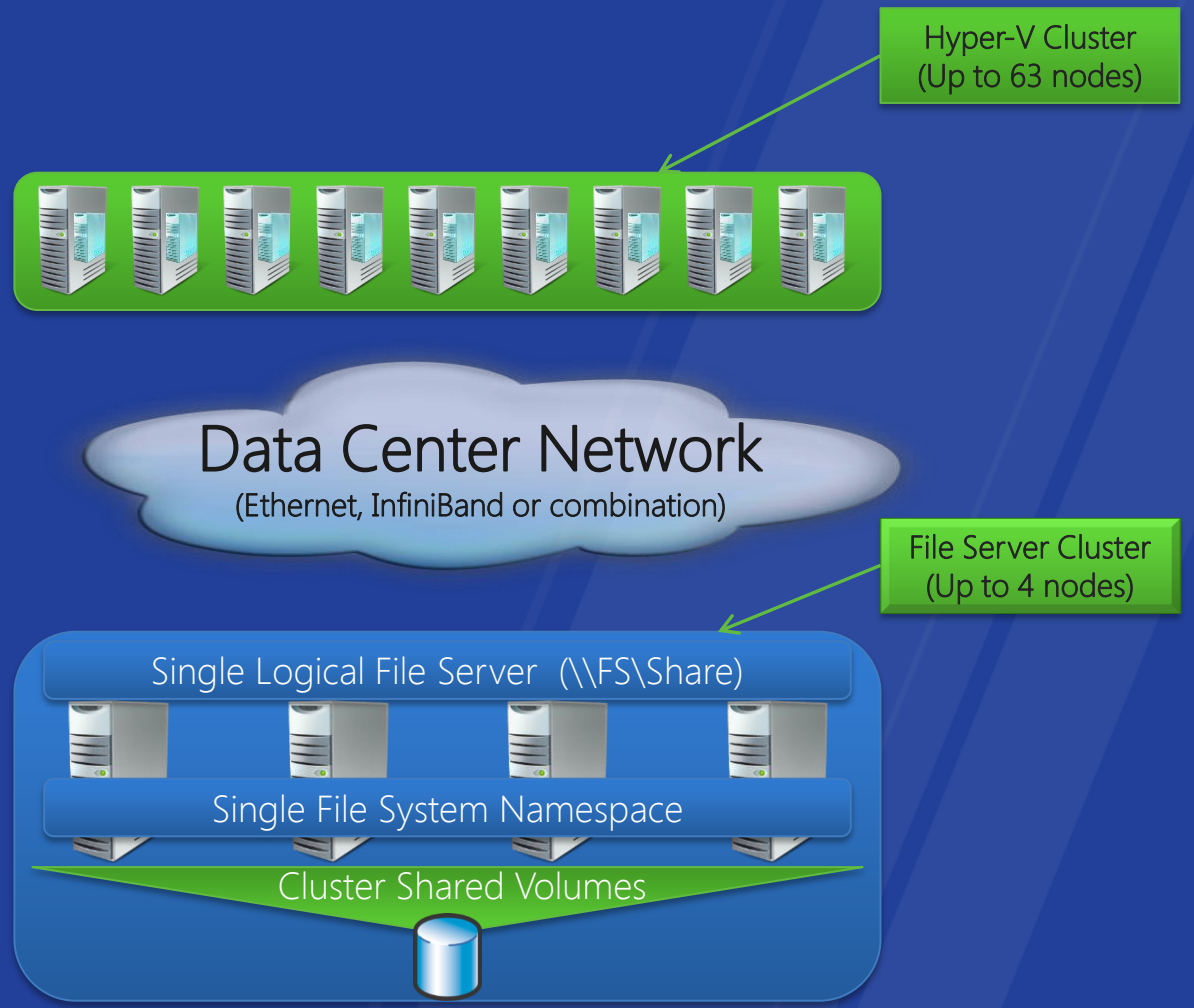
\\FSA\Share1
\\FSB\Share1

Leverage investment
More complex to manage
Multiple names



"Windows Server 8"
SMB2 Scale Out

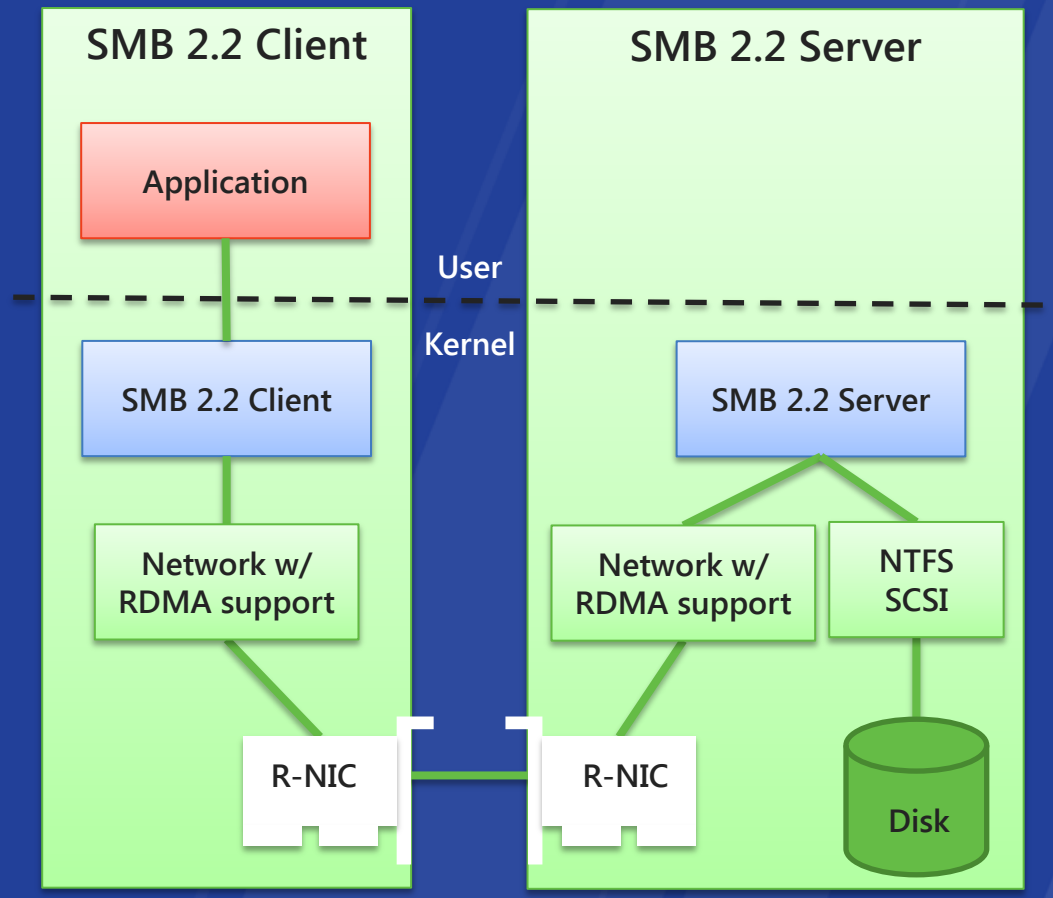
- Active/Active
  - Simultaneous access to a single share through all cluster nodes
  - Multiple shares still useful to isolate traffic
- Targeted for Server Applications
  - Server applications with few metadata operations - Hyper-V and Data-Heavy HPC
  - Bandwidth intensive applications – Increase available bandwidth by adding cluster nodes
- Simplified and easy management
  - Single logical file server – fewer DNS names, IP addresses (no need for virtual IPs)
  - Single file system namespaces – no drive letter limitation, larger file systems
  - No cluster disk resources to manage



SMB2 Direct (SMB2 over RDMA)

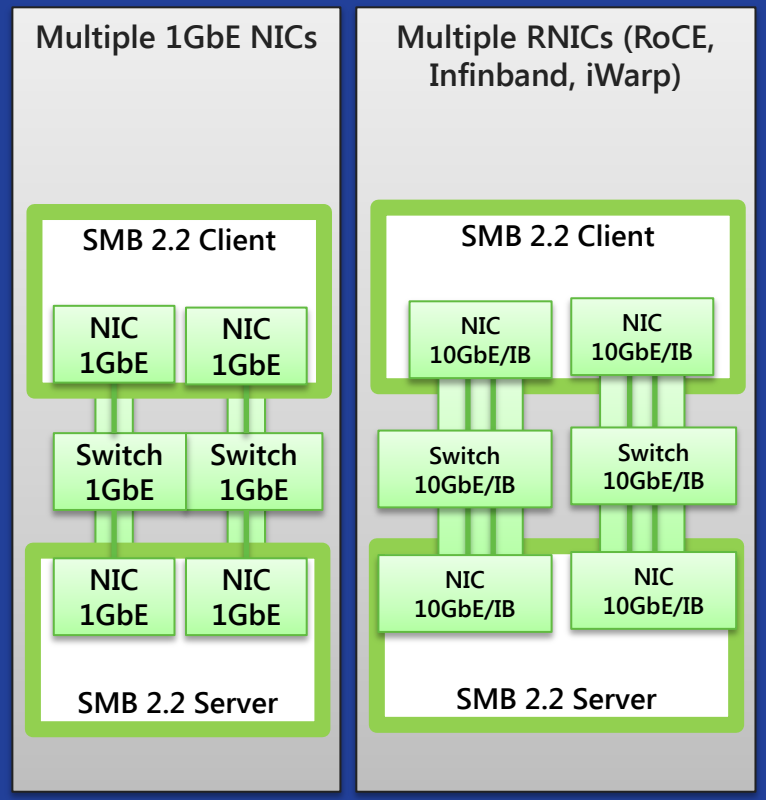
RDMA is "Remote Direct Memory Access" – a secure way to enable a DMA engine to transfer buffers between two machines across the network

- Used by File Server and Clustered Shared Volumes (CSV version 2) for storage communication within a cluster
- Advantages
  - Scalable, fast and efficient storage access
  - Choice of faster fabrics (QDR / FDR InfiniBand)
  - **Minimal CPU utilization** for I/O processing
    - High throughput with low latency
- **Transparent to applications**
- Can aggregate links with SMB2 Multichannel for load balancing and failover



SMB 2.2 Multichannel

Sample Configurations



- **Multiple TCP streams for one SMB 2.2 session**
  - Single NIC – with RSS enables more CPUs to process traffic
  - Multiple NICs, with NIC Teaming – SMB 2.2 can use a single IP address per team
  - Multiple NICs, Without NIC Teaming – each NIC must have a unique IP addresses (Required for RDMA NICs)
- **Failover**
  - SMB 2.2 Multichannel implements end-to-end failure detection
  - Fully leverages NIC teaming failover but does not require it
- **Throughput**
  - Bandwidth aggregation with multiple NICs
  - Multiple CPUs to process network interrupts with single RSS-capable NIC or multiple NICs
- **Automatic Configuration**
  - SMB 2.2 detects and uses multiple network paths
- Preliminary performance:
  - http://go.microsoft.com/fwlink/p/?LinkId=227841

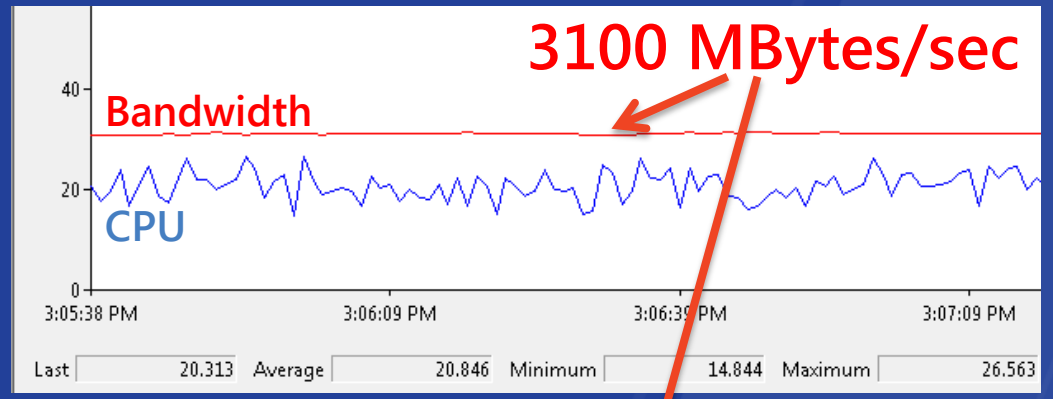
SMB 2.2 Direct Performance

Windows Server 8 supports all standards based RDMA networks

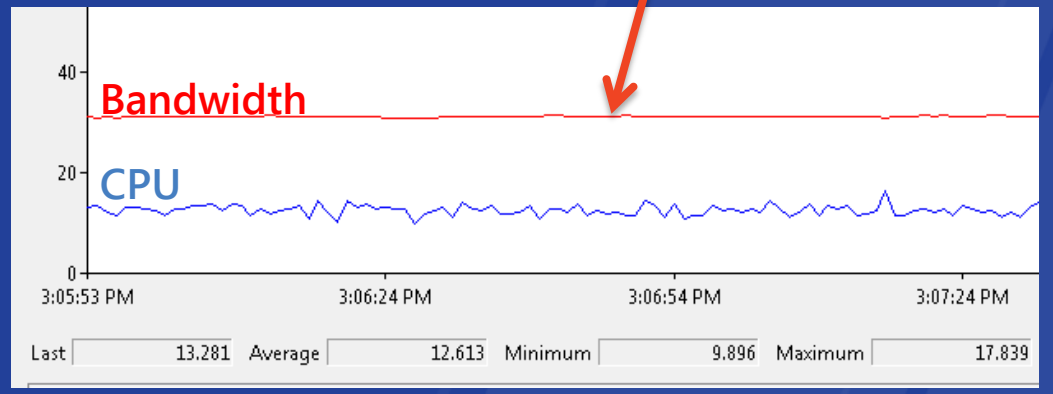
- Ethernet based protocols
  - ROCE (RDMA Over Converged Ethernet) – InfiniBand protocols on Ethernet
    - IBTA standard – http://infinibandta.org
    - Requires switches to support Data Center Bridging (DCB)
  - iWARP – RDMA on top of TCP/IP
    - IETF standard over TCP/IP – http://www.ietf.org
    - Only routable RDMA protocol
- InfiniBand based protocols
  - InfiniBand protocols on an InfiniBand network (32 gbits/sec, 56 gbits/sec…)

InfiniBand Preliminary 512 KB I/O Results:

Client: ~20% of 4 cores, or 80% of one core



Server: ~12% of 12 cores, or 140% of one core



SMB2 Direct Specification

- New document
  - MS-SMBD
- Sits "below" MS-SMB2 in the SMB2 stack
  - As a transport framing layer
  - Peer to Direct TCP
    - Optional

[MS-SMBD-Preview]: SMB2 Remote Direct Memory Access (RDMA) Transport Protocol Specification

Intellectual Property Rights Notice for Open Specification Documentation

- **Technical Documentation.** Microsoft publishes Open Specifications documentation for protocols, file formats, languages, standards as well as overviews of the interaction among each of these technologies.

- **Copyrights.** This documentation is covered by Microsoft copyrights. Regardless of any other terms that are contained in the terms of use for the Microsoft website that hosts this documentation, you may make copies of it in order to develop implementations of the technologies described in the Open Specifications and may distribute portions of it in your implementations using these technologies or your documentation as necessary to properly document the implementation. You may also distribute in your implementation, with or without modification, any schema, IDL's, or code samples that are included in the documentation. This permission also applies to any documents that are referenced in the Open Specifications.

- **No Trade Secrets.** Microsoft does not claim any trade secret rights in this documentation.

- **Patents.** Microsoft has patents that may cover your implementations of the technologies described in the Open Specifications. Neither this notice nor Microsoft's delivery of the documentation grants any licenses under those or any other Microsoft patents. However, a given Open Specification may be covered by Microsoft Open Specification Promise or the Community Promise. If you would prefer a written license, or if the technologies described in the Open Specifications are not covered by the Open Specifications Promise or Community Promise, as applicable, patent licenses are available by contacting iplg@microsoft.com.

- **Trademarks.** The names of companies and products contained in this documentation may be covered by trademarks or similar intellectual property rights. This notice does not grant any licenses under those rights.

- **Fictitious Names.** The example companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted in this documentation are fictitious. No association with any real company, organization, product, domain name, email address, logo, person, place, or event is intended or should be inferred.

Reservation of Rights. All other rights are reserved, and this notice does not grant any rights other than specifically described above, whether by implication, estoppel, or otherwise.

Tools. The Open Specifications do not require the use of Microsoft programming tools or programming environments in order for you to develop an implementation. If you have access to

- Available *now* at MSDN protodoc preview node:
  - http://msdn.microsoft.com/en-us/library/ee941641.aspx

Use of RDMA

- The SMB2.2 client **directs** all use of RDMA
  - For SMB2 Reads and Writes only
- The SMB2.2 server **performs** all RDMA
  - Improves security, integrity and performance
- Zero-copy, zero-touch
  - Buffer cache use is supported optionally on both peers
- Uses a simple RDMA profile
  - Allows use of any transport type (iWARP, IB, RoCE)
  - Any memory registration type
  - No optional features required
    - E.g. atomics, remote invalidate, etc

In summary

Remote File Storage for Server Applications

- Feature set enables Continuously Available storage from Windows File Server
- Scale and Performance
- Reliable

Application Developers

- Add support for storing data on file servers

System Integrators

- Design systems for Continuous Availability, if needed
- Design for RDMA

“Windows Server 8” File Server Features at a glance
For Server Applications (SMB2.2)

Primary : //build Talk 444

- Server Fault Tolerance – Transparent Failover
  - Server Fault Tolerance with zero application downtime
  - Hardware and software maintenance
- Server Scale Out
  - Active/Active file shares – single share across all nodes
  - Increased bandwidth, optimized for FLASH
- Application Consistent Backups
  - VSS for SMB2 File Shares (Extension to VSS)
  - Shadow copy of Server Application data on File Shares
- Performance for Server Applications
  - Optimizations for server application IO profiles
  - Performance analysis & tuning
- Flexible Storage Options
  - External Storage Arrays provide Offloaded Data Transfer (ODX) and sophisticated management
  - Shared JBOD SAS (Storage Spaces or Clustered PCI RAID)

Other talks that cover SMB2.2 Features

- Network Fault Tolerance & Scale
  - Multichannel (Talk 446 & 451)
    - Network Fault Tolerance with zero application downtime
    - Bandwidth Aggregation across multiple network adapters
  - SMB2 Direct & RDMA (Talk 446 & 451)
    - Support for RDMA enabled network adapters
    - High bandwidth, low latency and CPU consumption
- Scalable Management & Performance Optimization
  - PowerShell (Talk 451)
    - End to end CLI and scripting
  - Performance Counters and Events (Talk 451)
    - Extend local analysis techniques to the file server
  - Appliance deployments (Talk 449)
- Designing Building Blocks for the Cloud (Talk 430)
  - Use of File Server for Hosted Cloud deployments

thank you

- More Resources – Server+Cloud Sessions
  http://www.buildwindows.com

Microsoft®