Benefits of RDMA In Accelerating Ethernet Storage Connectivity

Mike Jochimsen, David Fair, Brian Hausauer,
Jim Pinkerton, Tom Talpey

March 4, 2015

SNIA™
Ethernet Storage Forum

Today's Presenters

David L. Fair
SNIA-ESF Chair
Unified Networking Mktg Mgr, Intel

Mike Jochimsen, SNIA–ESF Business Development Director of Alliances, Emulex

Brian Hausauer
RDMA Hardware Architect, Intel

Jim Pinkerton
Windows Server Architect, Microsoft

Tom Talpey
Architect, Microsoft

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

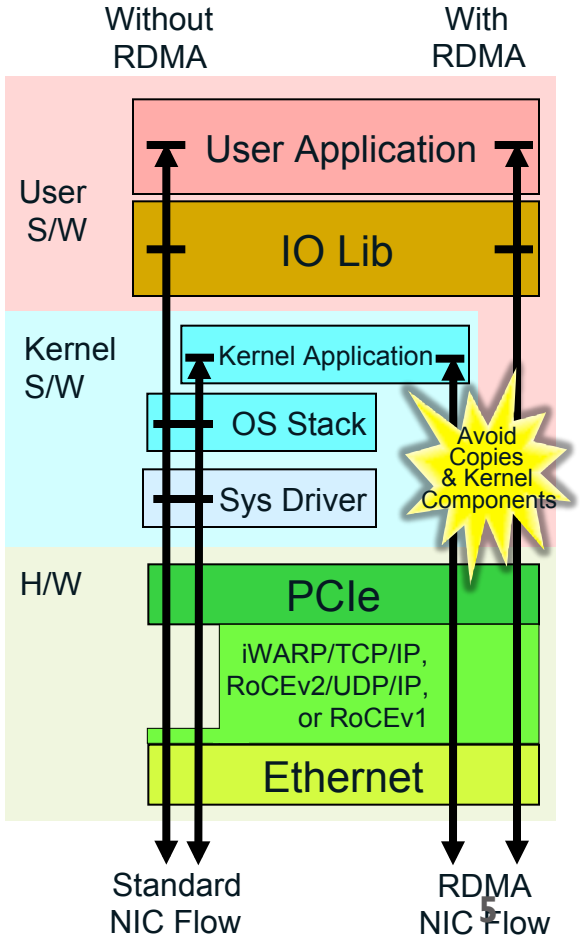
Agenda

SNIA
Ethernet Storage Forum

- What is RDMA?
- Overview of key RDMA protocols for storage
- Where is RDMA used in storage?
- Deep Dive on SMB Direct

What Is Remote Direct Memory Access (RDMA)?

- RDMA increases bandwidth while lowering latency, jitter, and CPU utilization – great for networked storage!
- RDMA bypasses system software stack components that processes network traffic
  - For user applications, RDMA bypasses the kernel altogether
  - Kernel applications that support RDMA get the same benefits as user applications, e.g. SMB Direct in Windows® Server 2012



What Is Remote Direct Memory Access (RDMA)?

SNIA Ethernet Storage Forum

- RDMA can deliver direct data placement of data from one machine (real or virtual) to another machine -- without copies
  - Applications can write directly into the memory space of another application or to a storage target
  - A storage system can write to an application's memory space or to a storage target
- RDMA enables header/data separation by preserving message boundaries

What Are RDMA Verbs?

SNIA Ethernet Storage Forum

- RDMA defines an asynchronous “verbs” programming interface (traditional sockets are synchronous)
- For applications to take advantage of RDMA, they have to be written to verbs
- However, verbs can be hidden “under the hood,” i.e., below the application APIs, to avoid the requirement of rewriting applications
  - HPC MPI libraries have been re-written to verbs so applications written to MPI can take advantage of RDMA without having to be rewritten
- The Open Fabrics Alliance maintains verbs and libraries of code written to them

OPENFABRICS ALLIANCE

Where Can I Find RDMA?

SNIA
Ethernet Storage Forum

- RDMA is in production today in InfiniBand* fabrics and in two Ethernet fabrics, iWARP and RoCE
  - An application should not care what wire, transport, or network is delivering its traffic
- RDMA is supported today in Linux and Microsoft Windows® Server

iWARP and RoCE Are Different "RDMA over Ethernet" Technologies

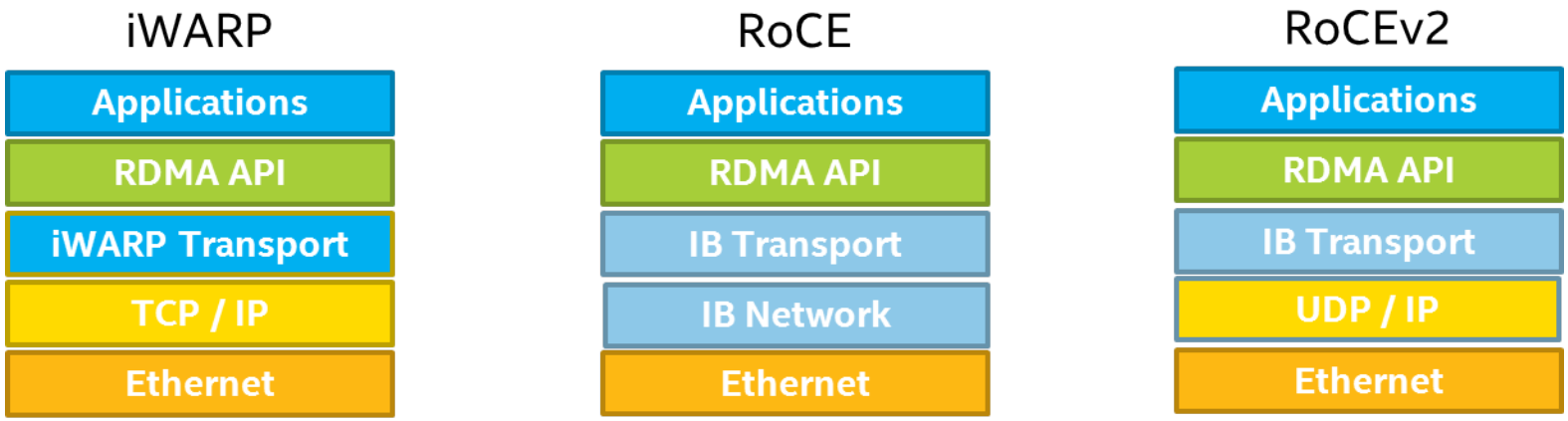
SNIA Ethernet Storage Forum

- Both can deliver RDMA storage over Ethernet
- Applications shouldn't care between them
- iWARP
  - 2007 Internet Engineering Task Force (IETF) standard
  - Runs on top of TCP/IP
- RDMA over Converged Ethernet (RoCE)
  - 2010 InfiniBand Trade Association (IBTA) standard
    - Uses an InfiniBand transport on an Ethernet wire
    - Requires Data Center Bridging to establish a lossless L2 fabric
  - 2014 IBTA update for routable "RoCEv2"
    - Adds a UDP/IP ecapsulation in support of routability

Under The Hood: iWARP and RoCE

SNIA Ethernet Storage Forum

- iWARP and RoCE cannot talk RDMA to each other because of L3/L4 differences
  - iWARP adapters can talk RDMA only to iWARP adapters
  - RoCE adapters can talk RDMA only to RoCE adapters



Agenda

SNIA
Ethernet Storage Forum

- What is RDMA?
- Overview of key RDMA protocols for storage
- Where is RDMA used in storage?
- Deep Dive on SMB Direct

Overview of RDMA storage protocols

SNIA Ethernet Storage Forum

- SMB 3 with SMB Direct
  - RDMA-enabled network protocol for file sharing
  - In production. Broad ecosystem including Windows Server 2012 and Windows Server 2012 R2
  - Deep dive by Microsoft later in this presentation
- NFS/RDMA
  - RDMA-enabled Remote Procedure Call (RPC) layer for NFS
  - Defined in RFC 5532, RFC 5666, RFC 5667 (2009-2010)
  - In open source distros for many years, but not widely used
  - Server OS support
    - Linux kernel NFS/RDMA client and server (recent industry effort to reanimate)
    - Oracle Solaris 11 NFS/RDMA client and server, for InfiniBand only

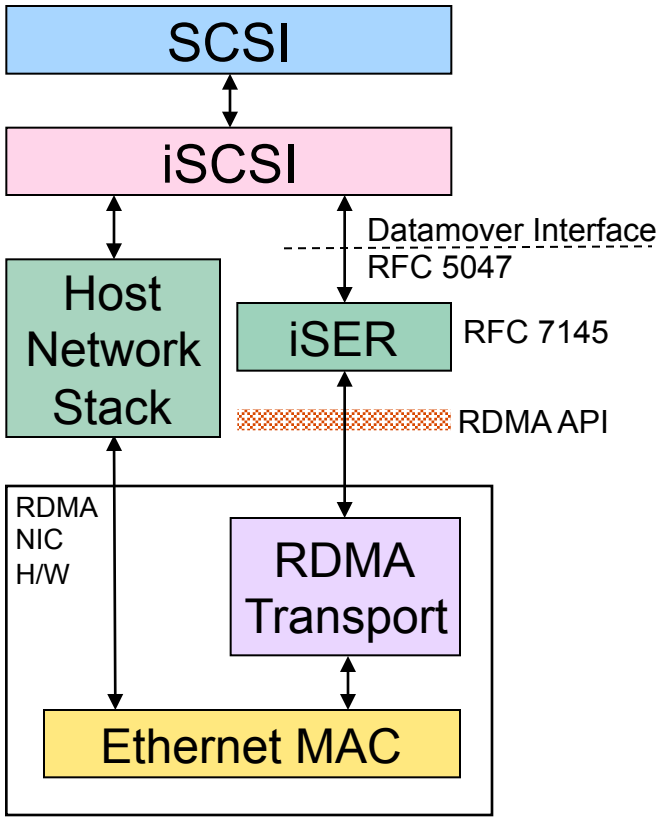
Overview of RDMA storage protocols

SNIA Ethernet Storage Forum

- iSCSI Extensions for RDMA (iSER)
  - Extends iSCSI with an RDMA-enabled Datamover service
  - Defined in RFC 5047 (2007), and RFC 7145 (2014)
  - In production. Growing interest by users and storage vendors
  - Server OS support
    - Linux kernel iSER initiator (kernel-integrated)
    - Linux kernel targets supporting iSER: LIO (kernel-integrated), SCST
    - Linux userspace target supporting iSER: TGT
    - Oracle Solaris 11 iSER initiator/target, for InfiniBand only

iSER Architecture

- iSER provides iSCSI with a Datamover service, which offloads data transfers (Data-In, Data-Out, etc.) to an RDMA NIC
- iSER maintains the existing iSCSI management infrastructure including MIB definition, bootstrapping, negotiation, naming and discovery, and security
- If iSER is not supported by initiator or target, then traditional iSCSI data transfer mode is used
- iSER enables a general-purpose RDMA NIC to handle iSCSI with the efficiency and performance of an iSCSI host bus adapter



Drawing adopted from Chadalapaka 2004, slide 8:
www.ietf.org/proceedings/60/slides/ips-1/ips-1.ppt

Overview of RDMA storage protocols

SNIA Ethernet Storage Forum

- SCSI RDMA Protocol (SRP)
  - Maps SCSI initiator to target using RDMA
  - Defined in T10 (ANSI INCITS 365-2002)
  - Requires out-of-band management for e.g. target discovery
  - Only used with InfiniBand. On Ethernet, iSER is preferred.
- NVMe over Fabrics
  - Proposed open standard for remote block access to NVMe SSDs across an RDMA-enabled fabric
  - Proposed standard currently in definition in the NVM Express Work Group
  - Detailed in another SNIA-ESF webcast available on demand: *The Performance Impact of NVMe and NVMe over Fabrics*

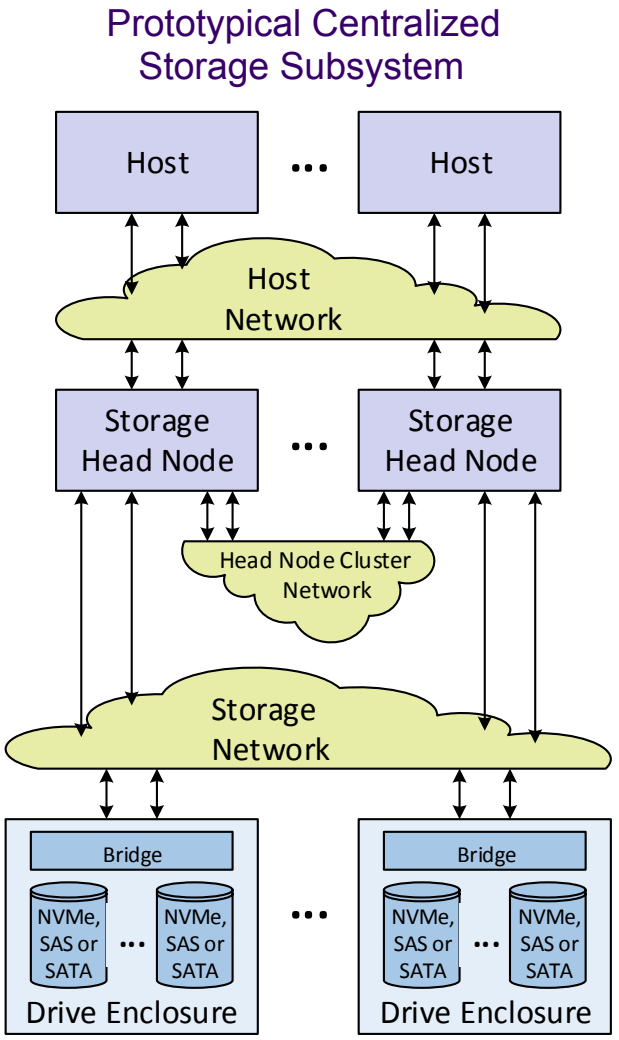
Overview of RDMA storage protocols

- Future topics, out-of-scope for this presentation
  - RDMA-enabled distributed filesystems
    (e.g. Lustre, GPFS, HDFS, Ceph, etc.)
  - RDMA-enablement of scale-out distributed SAN or caching
    (e.g. EMC ScaleIO, VMware Virtual SAN, Dell Fluid Cache, etc.)

SMB 3 with SMB Direct, NFS/RDMA, iSER, and NVMe over Fabrics are designed to work with either iWARP or RoCE

Where is RDMA Used In Storage?

- Prototypical Centralized Storage Subsystem
  - Hosts are storage clients
  - Storage Head Nodes serve files, blocks, objects, or combinations
  - Drive enclosure could be
    - A typical JBOD
    - A server with direct-attached storage
    - A single network-connected NVMe, SAS, or SATA device



Where is RDMA Used In Storage?

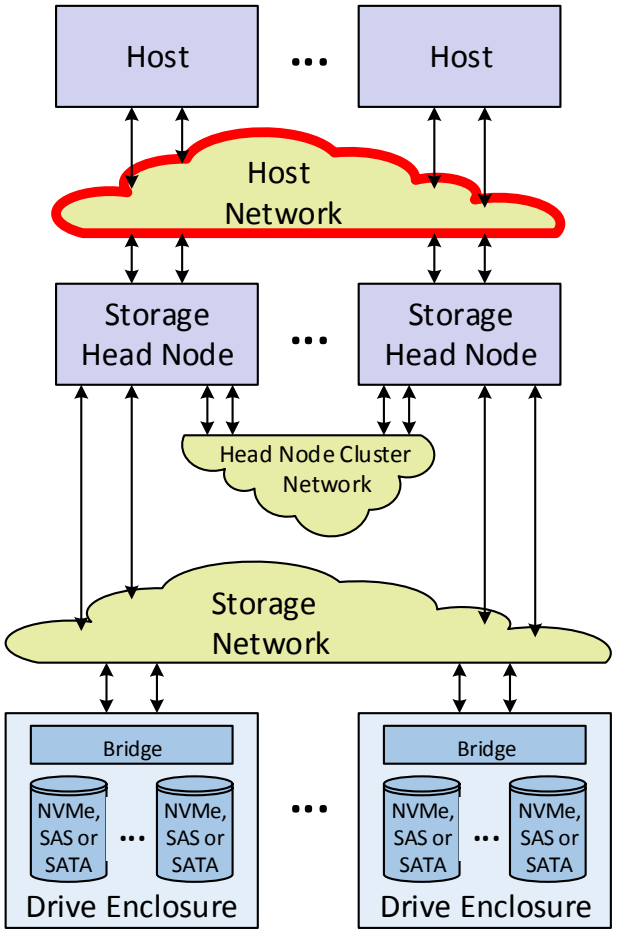
SNIA Ethernet Storage Forum

Host Network

As RDMA-enabled Hosts become common, Host Network protocols will shift to take advantage :

- File Storage protocols will shift from SMB or NFS on Ethernet towards *SMB Direct* or NFS/RDMA.
- Block Storage will shift from Fibre Channel, iSCSI, or FCoE towards iSER or *NVMe over Fabrics* for high-performance all-flash storage.

Prototypical Centralized Storage Subsystem

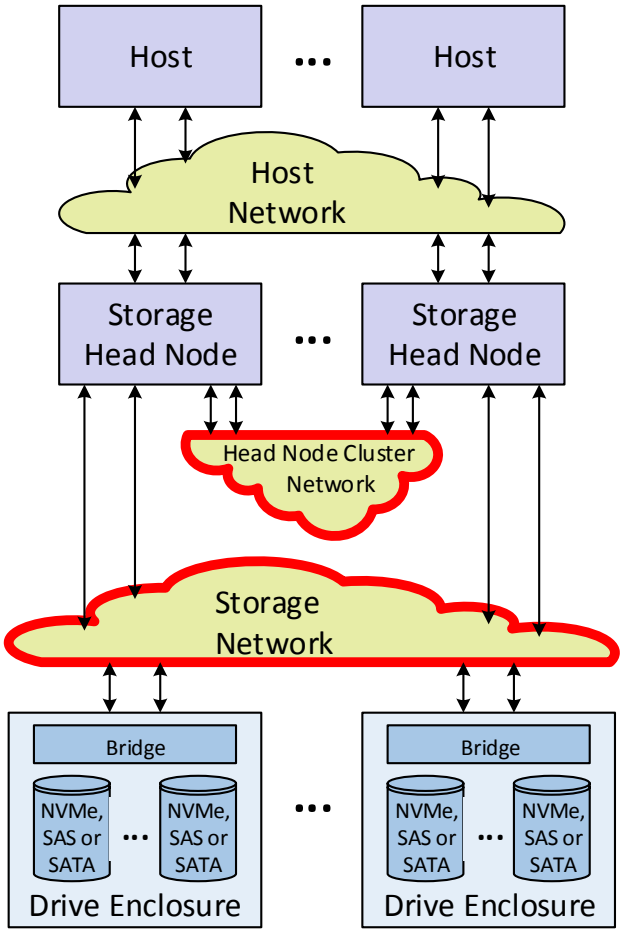


Where is RDMA Used In Storage?

SNIA Ethernet Storage Forum

- Head Node Cluster Network
  - Wide variety of networking technologies in use: Ethernet, Fibre Channel, PCIe non-transparent bridging, RapidIO, InfiniBand, etc.
  - Some use RDMA, but storage protocols on this network are typically vendor-specific.
- Storage Network
  - Today, this is SAS, Fibre Channel, or a PCIe fabric with high-performance all-flash storage
  - Industry momentum to transition the Storage Network to Ethernet
  - RDMA protocols are well-suited to this transition
    - iSER, *NVMe over Fabrics*, or combination
  - As the Drive Enclosure Bridge becomes more intelligent/autonomous, other RDMA protocols may emerge

Prototypical Centralized Storage Subsystem



Agenda

SNIA Ethernet Storage Forum

- What is RDMA?
- Overview of key RDMA protocols for storage
- Where is RDMA used in storage?
- Deep Dive on SMB Direct

The Origins of SMB3

SNIA Ethernet Storage Forum

- Enterprise-Class Storage Features with Much Simpler Management
- File sharing semantics (rather than block semantics)
  - Increased flexibility, easier provisioning and management
  - Easy deployment of encryption & signing
- Enterprise class RAS
  - No application downtime for planned maintenance or unplanned failures
  - Extremely fast failover (<10 sec)
- Excellent Performance
  - Near line rate performance for both small and large IOs with RDMA
  - Very low CPU utilization with RDMA
- Work with existing storage investments
- Work on existing and next generation network infrastructure

SMB3 Features

- Remote File Protocol
  - Full file semantics
  - Read/write/metadata
    - Caching, leasing, locking
    - Access control, sharing
  - Enterprise Application-ready
- Performance
  - Scale-Out (scalable server)
  - SMB Direct (RDMA)
  - Multichannel (trunking for bandwidth)
  - Directory Leasing (metadata caching)
  - BranchCache™ V2
  - Server Copy Offload
- Availability
  - Transparent Failover
  - Witness (fast failover)
  - Multichannel (path redundancy)
- Strong Security
  - Encryption – AES-CCM/GCM
  - Signing - AES-CMAC
- Backup
  - Volume Shadow Copy (VSS) for SMB File Shares
- Management
  - PowerShell™ over WS-Man
  - SMI-S File

SMB3 Multichannel

SNIA Ethernet Storage Forum

Full Throughput

- Bandwidth aggregation with multiple NICs
- Multiple CPUs cores engaged when NIC offers Receive Side Scaling (RSS) or Remote Direct Memory Access (RDMA)

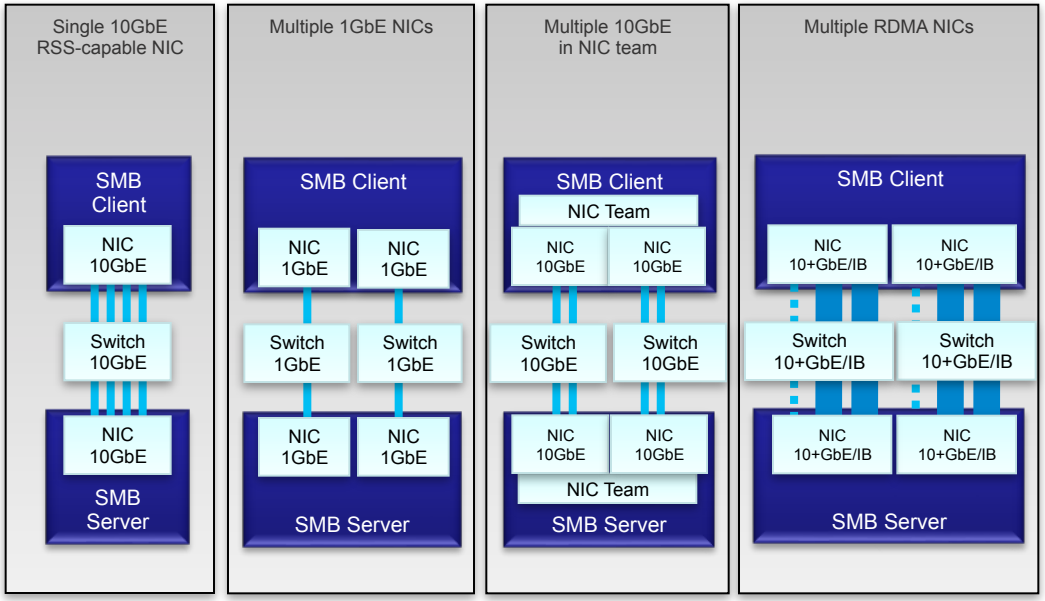
Automatic Failover

- SMB Multichannel implements end-to-end failure detection
- Leverages NIC teaming if present, but does not require it

Automatic Configuration

- SMB detects and uses multiple paths
- Discovers and “steps up” to RDMA

Sample Configurations



SMB Direct

Advantages

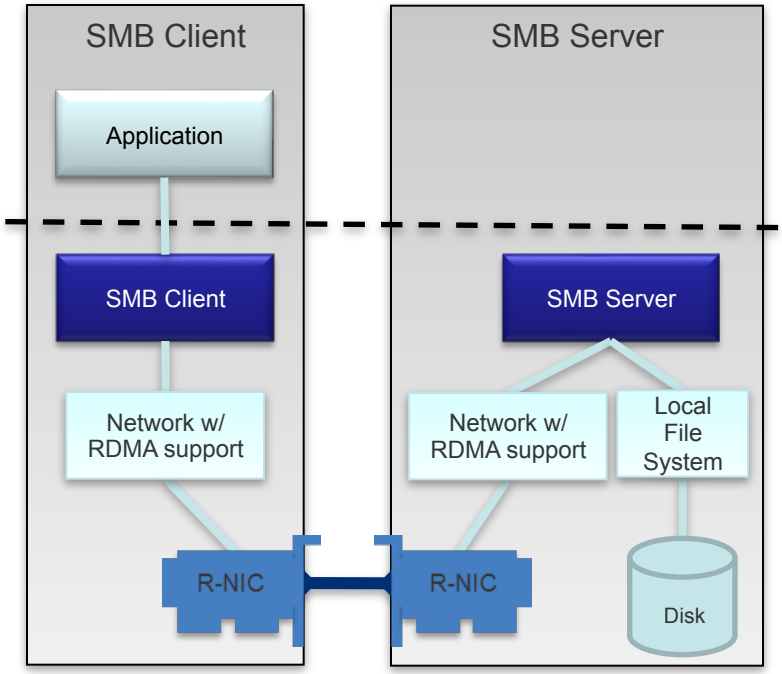
- Scalable, fast and efficient storage access
- High throughput with low latency
- Minimal CPU utilization for I/O processing
- Load balancing, automatic failover and bandwidth aggregation via SMB Multichannel

Scenario

- High performance remote file access for application servers like Virtualization and Databases

Required hardware

- RDMA-capable network interface (R-NIC)
- Three types: iWARP, RoCE and InfiniBand



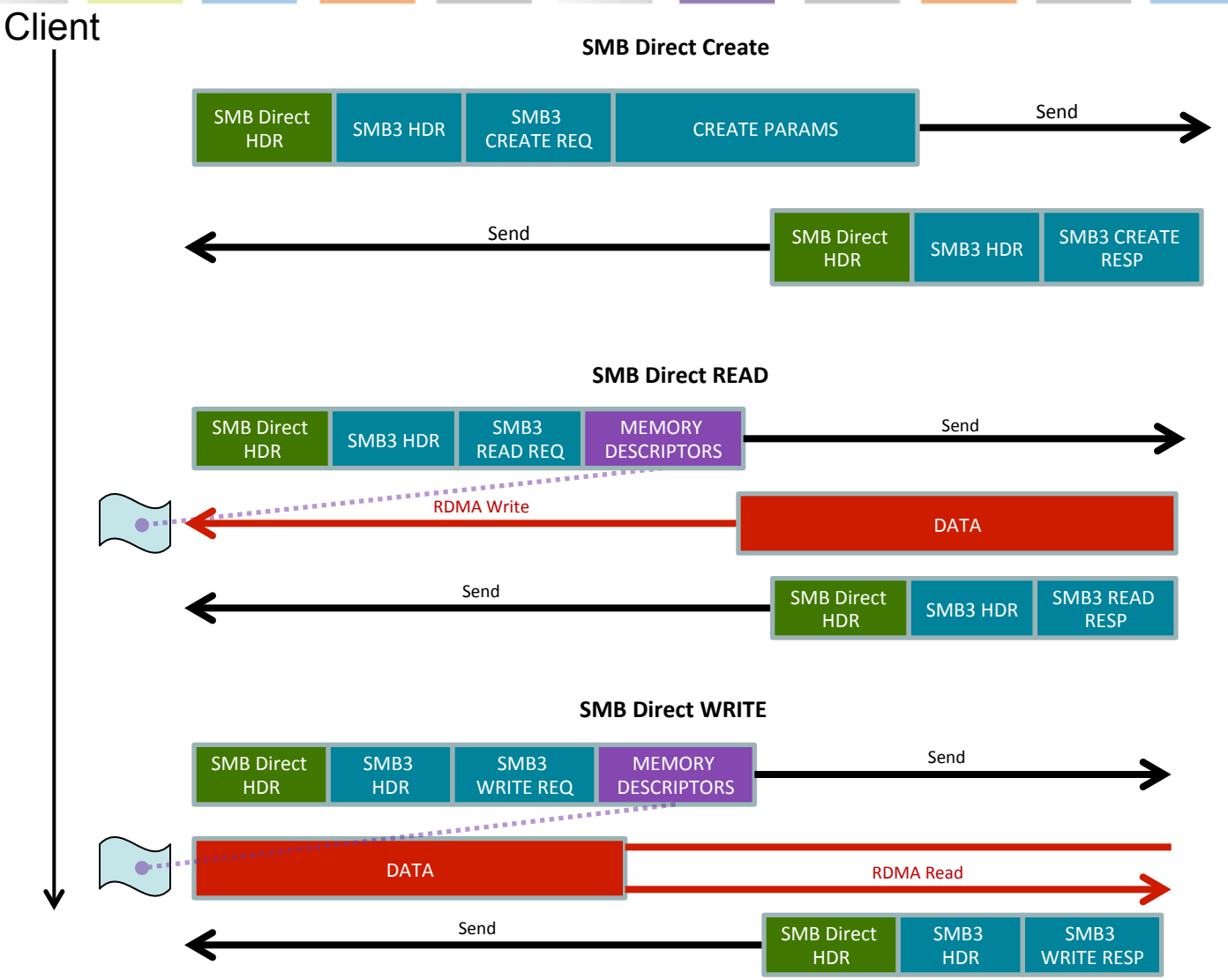
SMB Direct Protocol Basics

SNIA
Ethernet Storage Forum

- SMB Direct is a *transport framing*
  - Only 3 message types
- 2-way full duplex transport which supports:
  - Datagram-type send/receive exchange
    - With fragmentation/reassembly for “large” messages (~2KB and up)
  - Direct RDMA Read/Write
    - Bulk data transfer from 4KB to 8MB per operation
- Secure RDMA transfer model:
  - Client buffer advertisement for READ and WRITE
  - Server RDMA buffer access (push/pull)

SMB Direct Transfers

RDMA in RED



SMB3 and SMB Direct Workloads

SMB3 Workloads

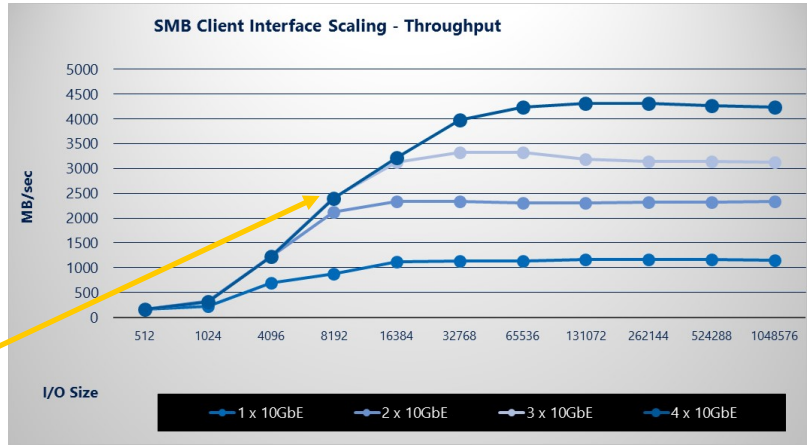
- Storage for Virtualization, database, HPC
- Storage for desktops/laptops/slates (LAN/ WAN)
- Virtual Machine Live Migration between hosts

The Design Point is Private Cloud

- Almost all IOs are small (<64 KB)
  - TCP throughput is significantly **less** due to **CPU saturation**
- SMB Direct (SMB3 support for RDMA)
  - **RDMA enables near line rate** with small IOs

SMB3 Multi-Channel Enables Linear Scaling

- Linear 10GbE scaling with TCP/IP (no RDMA)
- 4300 Mbps with 4x10GbE



Source: Microsoft Sept 2012 white paper provides full details
See http://go.microsoft.com/fwlink/p/?LinkId=227841

SMB3 as Virtual Machine Storage

SNIA Ethernet Storage Forum

- File-based containers
  - Natively support high availability, replication, snapshot/backup
  - Remote file protocols increasingly supported by hypervisors
  - SMB3 Continuous Availability features fully supported by Microsoft Hyper-V
- Workload:
  - 4KB-8KB Random, broad read/write mix
  - Very high **aggregate** IOPS
    - Example rack scaling: 1000 IOPS/VM x 32 VM/U x 32U =~ 1M IOPS
- Low overhead of SMB Direct
  - Less overhead to serve I/O -> more CPU for compute, higher IOPS -> more VM's per rack

SMB3 Storage with RDMA on 40GbE

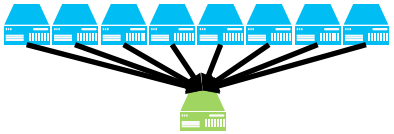
PRELIMINARY

SNIA Ethernet Storage Forum

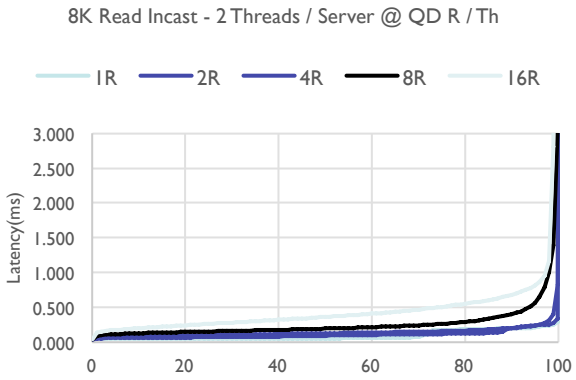
8 KB IOPs* and Latency (ms)
Single client, 8 file servers

QD *R* / Thread	Read IOPs	Latency 50th Read	Latency 99th Read
1	163,800	0.056	0.249
2	291,000	0.094	0.270
4	440,900	0.129	0.397
8	492,400	0.195	1.391
16	510,200	0.363	2.810

Incast Performance



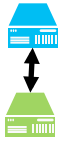
Near line-rate with small IOs
- 33.4 Gbps at 8 KB IO
- Excellent latency variance



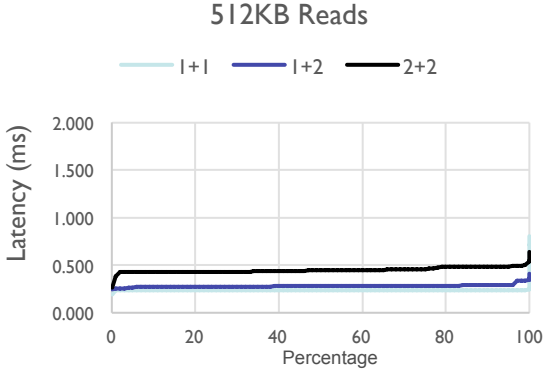
Large IOPs* (512 KB) in Gbps

512KB	Read BW	Write BW
1 Thread		
1 IO	17.82	16.14
2 IO	29.74	23.13
2 Thread		
2 IO/t	37.21	30.61

Client-to-Server Performance



Near line-rate with large IOs
- 37.2 Gbps with 2 threads



Excellent Log Write Performance

- 7.2M Read IOPs*, 512 Byte, single outstanding IO
- 3.3M Write IOPs*, 512 Byte, single outstanding IO

*IOs are not written to non-volatile storage
iWARP used for these results
Test configuration details in backup

Other SMB Uses

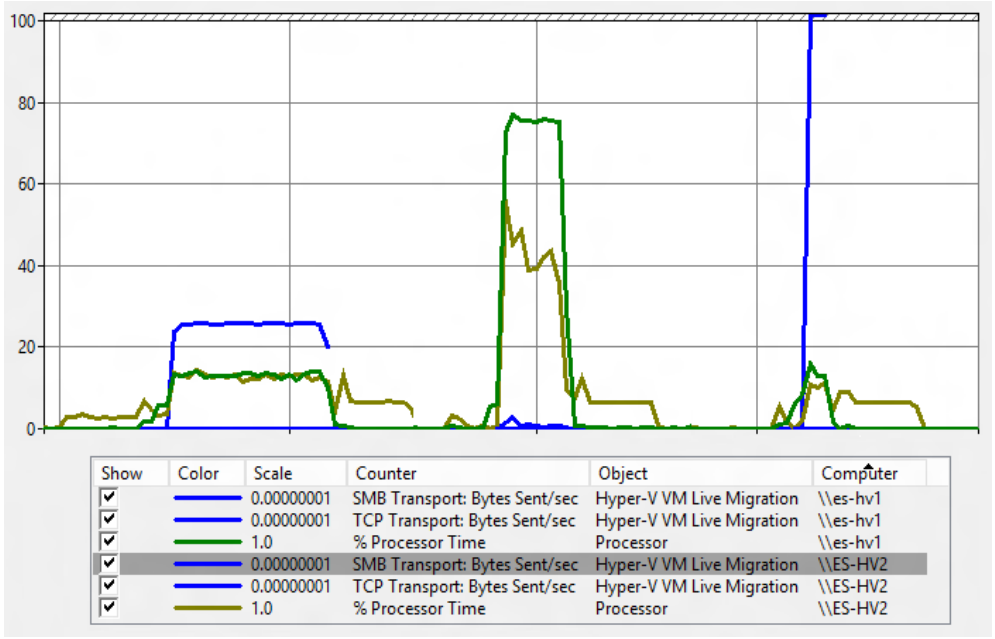
SNIA Ethernet Storage Forum

- SMB as a Transport
  - Provides transparent use of RDMA
- Example: Virtual Machine Live Migration
  - Peer-to-peer memory image transfer
- Bulk data transfer, ultimate throughput
  - Memory-to-memory “unconstrained by disk physics”
  - Very low CPU with RDMA – more CPU to support **live migration**
    - Minimal VM operational blackout
- Other SMB3 benefits to migration:
  - SMB3 availability provides transparent storage failover after migration
    - Minimal VM operational blackout, maximum VM availability and load balancing

VM Live Migration Performance

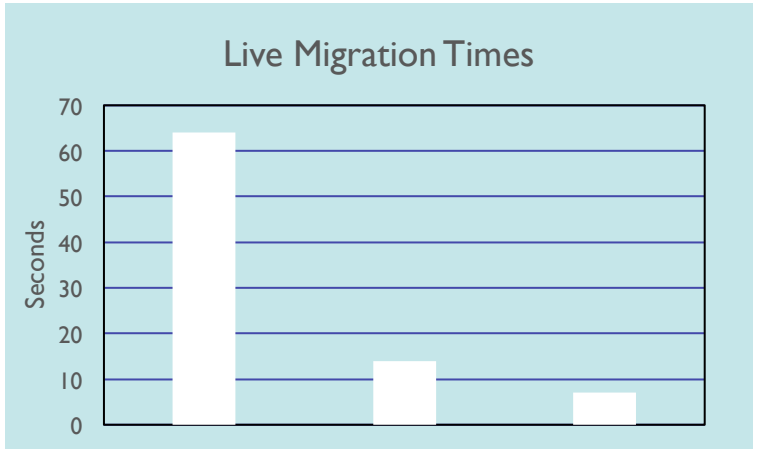
SNIA Ethernet Storage Forum

TCP/IP | Compression | SMB+RDMA



- TCP for Live Migration
  - CPU core-limited, long migration time
- TCP with Compression for Live Migration
  - Even more CPU-intensive, reduced migration time
- SMB Direct for Live Migration
  - Minimal CPU, 10x better migration time
  - Greatly increased VM live availability

Live Migration
2x10 GbE RoCE



SMB3 / SMB Direct / RDMA Resources

SNIA Ethernet Storage Forum

- Protocol documentation
  - http://www.microsoft.com/openspecifications
- SNIA education SMB3 tutorial
  - http://www.snia.org/education/tutorials/FAST2015
- SNIA SDC 2014
  - http://www.snia.org/events/storage-developer/presentations14#smb
- Jose Barreto's blog
  - http://smb3.info/
- SNIA-ESF webcast on NVMe over fabrics
  - https://www.brighttalk.com/webcast/663/132761

After This Webcast

SNIA Ethernet Storage Forum

- This webcast and a PDF of the slides will be posted to the SNIA Ethernet Storage Forum (ESF) website and available on-demand
  - http://www.snia.org/forums/esf/knowledge/webcasts
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-ESF blog
  - http://sniaesfblog.org/
- Follow us on Twitter @ SNIAESF

Thank You

SNIA™
Ethernet Storage Forum