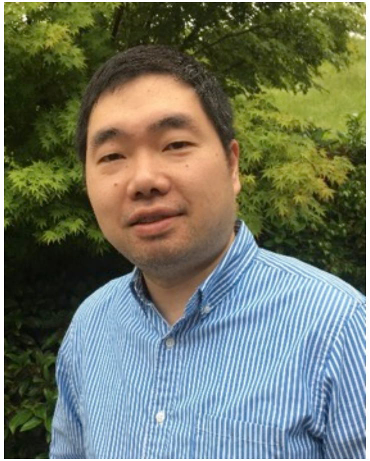
SNIA ESF | ETHERNET STORAGE

# RoCE vs. iWARP
## A Great Storage Debate

**Live Webcast**
**August 22, 2018**
**10:00 am PT**

# Today's Presenters



**John Kim**
**SNIA ESF Chair**
**Mellanox**

**Tim Lustig**
**Mellanox**

**Fred Zhang**
**Intel**

# SNIA-At-A-Glance

# SNIA Legal Notice

- The material contained in this presentation 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.
- 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.

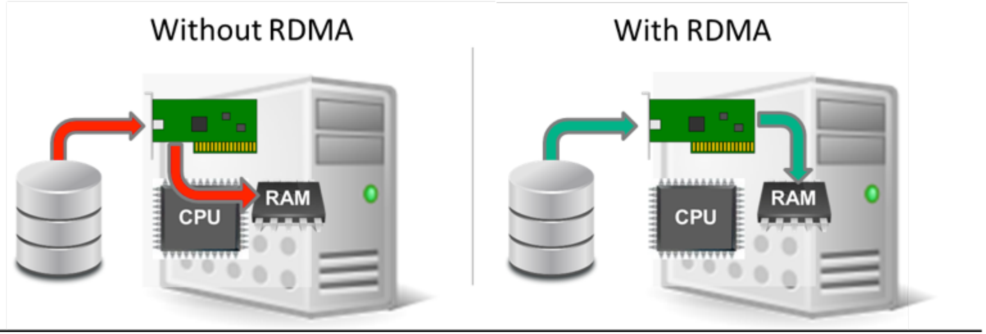
  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

# Agenda

- Introductions  John Kim – Moderator
  - What is RDMA?
- Technology Introductions
  - RoCE – Tim Lustig, Mellanox Technologies
  - iWARP – Fred Zhang, Intel Corporation
- Similarities and Differences
- Use Cases
- Challenge Topics
  - Performance, manageability, security, cost, etc.

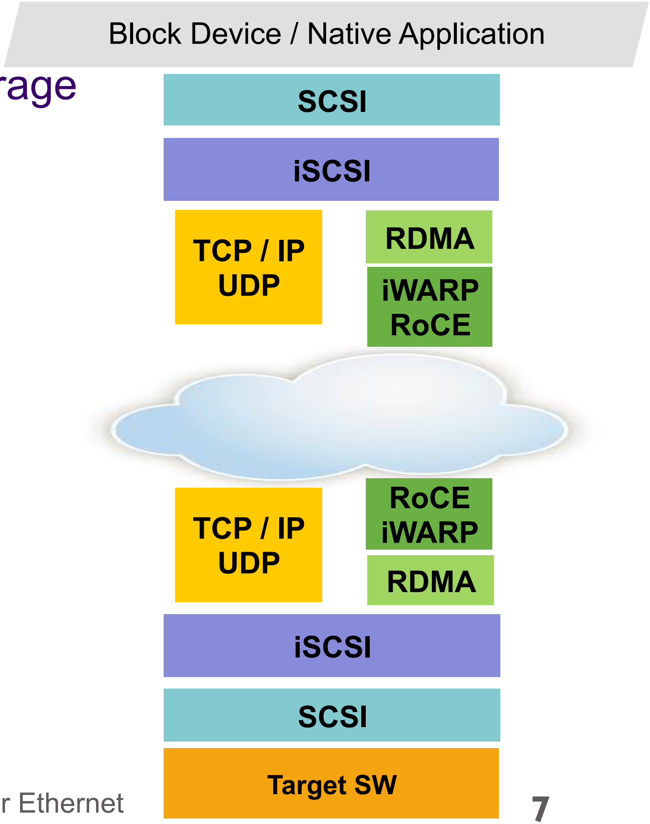
# What is RDMA?

- Remote Direct Memory Access
  - DMA from the memory of one node into the memory of another node without involving either one's operating system
- Performed by the network adapter itself, no work needs to be done by the CPUs, caches or context switches
- Benefits:
  - High throughput
  - Low latency
  - Reduced CPU utilization



# RDMA as a Transport

- Block storage networking technology and networked file storage
  - SCSI protocol running (usually) on TCP/IP or UDP
  - SMB Direct, NFS v4
  - Storage Spaces Direct
- RDMA supported by native InfiniBand*, RoCE and iWARP network protocols
- Standardization (RoCE by IBTA, iWARP by IETF)
  - RFCs 5040, 5041, 5044, 7306, etc.
  - RoCE first available: 2003 Windows / 2005 Linux / 2006 VMware
  - iWARP first available: 2007
- "iSCSI" usually means SCSI on TCP/IP over Ethernet




*Almost always over Ethernet

SNIA ESF | ETHERNET STORAGE

# ROCKY

## RoCE – Tim Lustig, Mellanox

# What is RoCE?

- RoCE (RDMA over Converged Ethernet)
- The most popular RDMA implementation over Ethernet
  - Enables highest throughput, lowest latency and lowest CPU overhead for RDMA
  - Designed for enterprise, virtualized, cloud, web 2.0 and storage platforms
  - Increases performance in congested networks
  - Deployed in large data centers
- Proven, most widely deployed RDMA transport
  - Server efficiency and scaling to 1000s of nodes
  - Scales to 10/25/40/50 and 100G Ethernet support and beyond

# RoCE Overview

- RoCE v1
  - Needs custom settings on the switch
    - Priority queues to guarantee lossless L2 delivery
    - Takes advantage of PFC (Priority Flow Control) in DCB Ethernet


- RoCE v2 (lossless) – Improved efficiency
  - RDMA transport paradigm depends on a set of characteristics
    - No dropped packets
    - Arbitrary topologies
    - Traffic class types

# RoCE Support

SNIA | ETHERNET
ESF | STORAGE

## DCB – Data Center Bridging

- DXBX – Data Center Bridging Exchange
- ECN – Explicit Congestion Notification
- PFC – Priority Flow Control
- ETS – Enhanced Transmission Specification

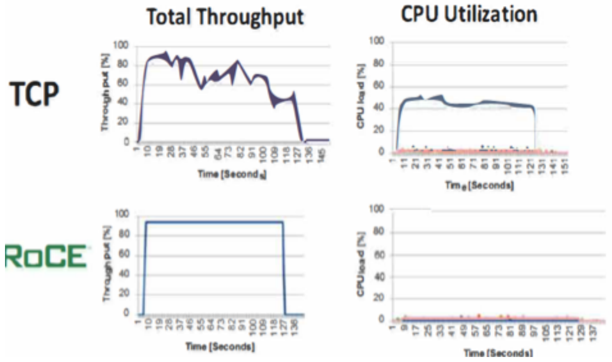
| Ethernet | IEEE 802.1x |
|---|---|
| Congestion Notification | Yes (802.1az) ECN, DCB |
| Lossless | Yes (802.1Qbb) PFC |
| Classes of Service | Yes (802.1Qaz) ETS |

# Wide Adoption and Support

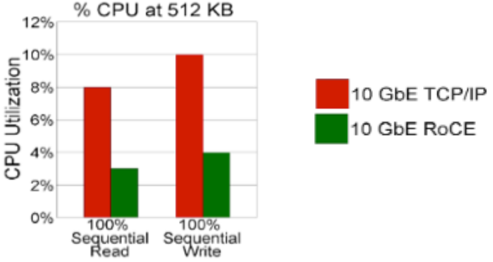
- VMware
- Microsoft SMB 3.0 (Storage Space Direct) and Azure
- Oracle
- IBM Spectrum Scale (formerly known as IBM GPFS)
- Gluster, Lustre, Apache Spark, Hadoop and Ceph
- Software-Defined Storage (SDS) and hyperconverged vendors
- Nearly all NVMe-oF demonstrations, designs, and customer deployments are using RoCE

# RoCE Benchmarks

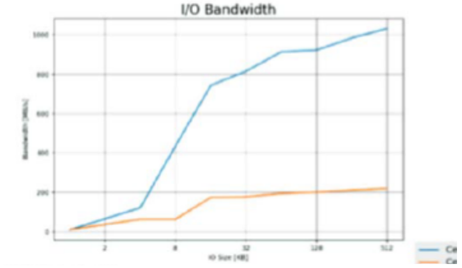
## TCP vs RoCE

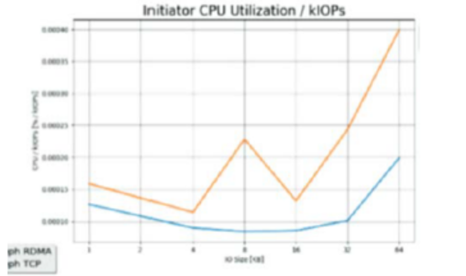



## MSFT SMB 3.0

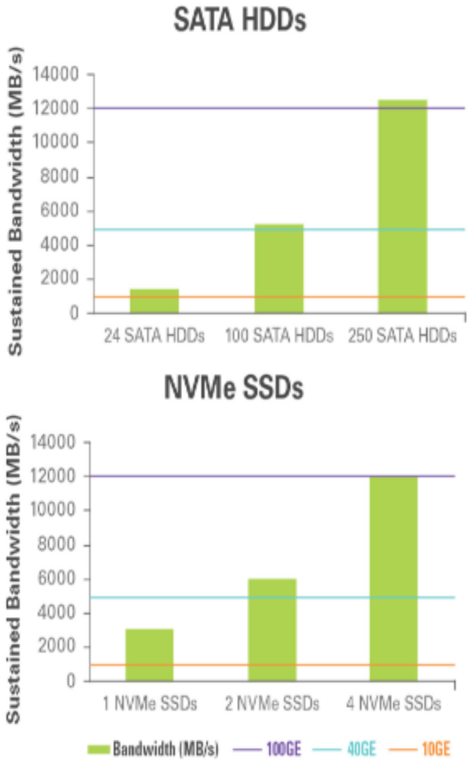

## Ceph



# RoCE Future-Proofs the Data Center

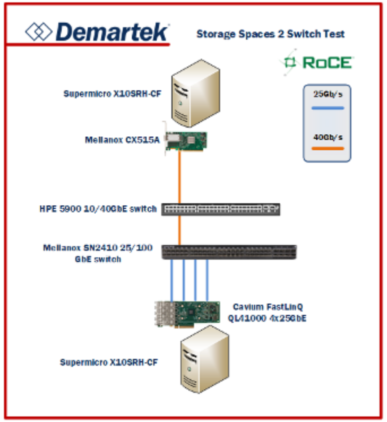
- Transform Ethernet networks and remove network, storage and CPU bottlenecks
  - Support for NVMe eliminates throughput and latency bottlenecks of slower SAS and SATA drivers
    - A NVMe SSD can provide sustained bandwidth of about 50 HDDs
  - RoCE extends NVMe to NVMe-oF
    - Access remote storage systems similarly as locally attached storage
  - Solid State NVM is expected to be 1,000 times faster than flash
    - 3D XPoint, Optane



# Additional Resources

SNIA | ESF ETHERNET STORAGE

- RoCE Initiative.org (http://www.roceinitiative.org/resources/)
- Demartek – (https://www.demartek.com/)





## RoCE Deployment Guide 2018 Edition

The second edition of the RoCE Deployment Guide reflects both a growing industry interest in RoCE technology's network acceleration capabilities as well as the increasing number of RoCE-capable product offerings to support it.

DOWNLOAD NOW

# iWARP – Fred Zhang, Intel

# What is iWARP

- iWARP is: ~~Internet Wide-Area RDMA Protocol~~

  iWARP

- iWARP is NOT an acronym
- iWARP can be used in different network environments: LAN, storage network, Data center, or even WAN

http://www.rdmaconsortium.org/home/FAQs_Apr25.htm

# What is iWARP

- iWARP extensions to TCP/IP were standardized by the Internet Engineering Task Force (IETF) in 2007. These extensions eliminated three major sources of networking overhead: TCP/IP stack process, memory copies, and application context switches.

| Extension | Solution | Benefit |
|---|---|---|
| Offload TCP/IP | Offloads the TCP/IP process from the CPU to the RDMA-enabled NIC (RNIC) | Eliminates CPU overhead for network stack processing |
| Zero Copy | iWARP enables the application to place the data directly into the destination application's memory buffer, without unnecessary buffer copies | Significantly relieves CPU load and frees memory bandwidth |
| Less Application Context Switching | iWARP can bypass the OS and work in user space to post the command directly to the RNIC without the need for expensive system calls into the OS | Can dramatically reduce application context switching and latency |

# iWARP Protocols

**Remote DMA Protocol (RDMAP)**
RDMA Ops: RDMA read/write and send

**Direct data placement over Reliable Transports (DDP)**
Message segmentation and reassembly
Tagged and untagged buffer models
Direct data placement and in-order message delivery

**Marker PDU Aligned Framing for TCP (MPA)**
Framing and stronger data integrity (CRC)
FPDU alignment/containment (MPA-aware TCP)

**TCP**

**IP**

# Top Tier iWARP Applications

| Application | Category | User/Kernel | OS |
|---|---|---|---|
| **SMB Direct Client/Server** | Storage – Network file system | Kernel | Windows |
| **Storage Spaces Direct** | Storage – Network block storage | Kernel | Windows |
| **NVMe* over Fabrics Initiator/Target** | Storage – Network block storage | Kernel | Linux |
| **NVMe over Fabrics Initiator/Target for SPDK** | Storage – Network block storage | User | Linux |
| **LIO iSER Initiator/Target** | Storage – Network block storage | Kernel | Linux |
| **uDAPL** | Messaging middleware | User | Linux |
| **OFI/libfabric provider for VERBs** | Messaging middleware | User | Linux |
| **Open MPI/Intel® MPI Library** | HPC | User | Linux |
| **NFS/RDMA client/server** | Storage – network file system | Kernel | Linux |
| **rsockets** | Messaging middleware | User | Linux |

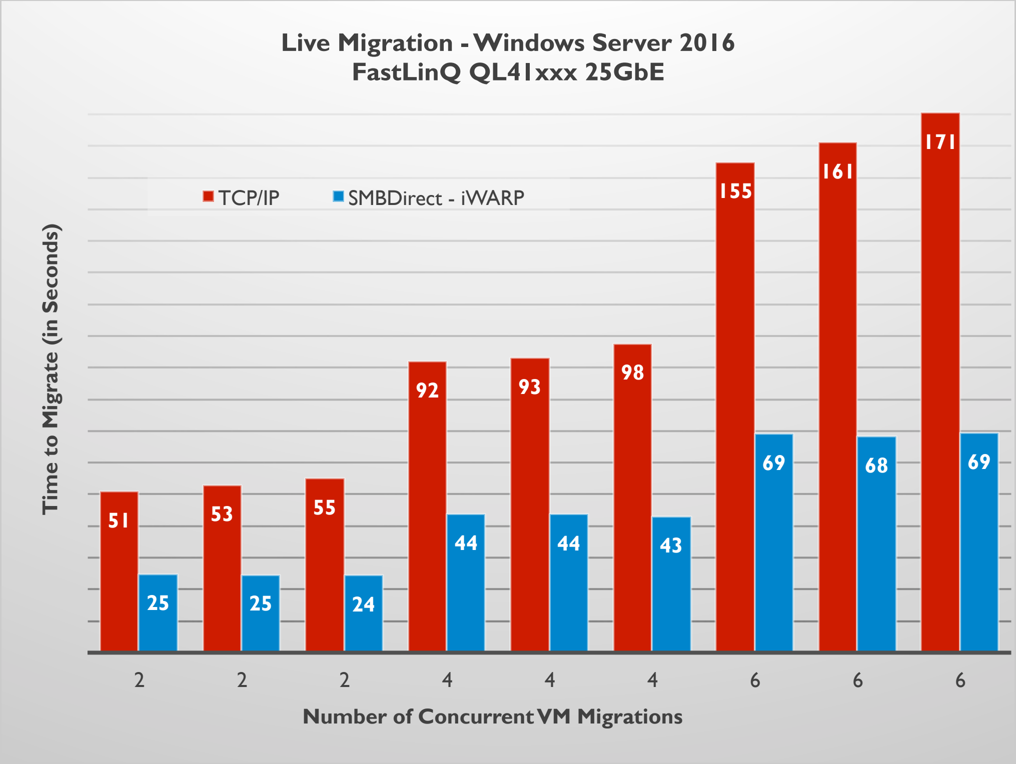
# iWARP Performance

SNIA ESF | ETHERNET STORAGE

- **>1M** IOPs SMB Direct Storage Performance, **1.67x** TCP
  - with Intel Ethernet Connection X722 4x10Gb featured iWARP
  - 4k, 70%Read 30%Write

# Accelerate Live Migration with iWARP

SNIA ESF | ETHERNET STORAGE




## FastLinQ QL41xxx iWARP

Reduces Live Migration Time by 58%

Highly Predictable Migrations

## Benefits

Shorter Maintenance Windows

Adaptive Load Balancing – SLAs

Less Flight time = Less Risk

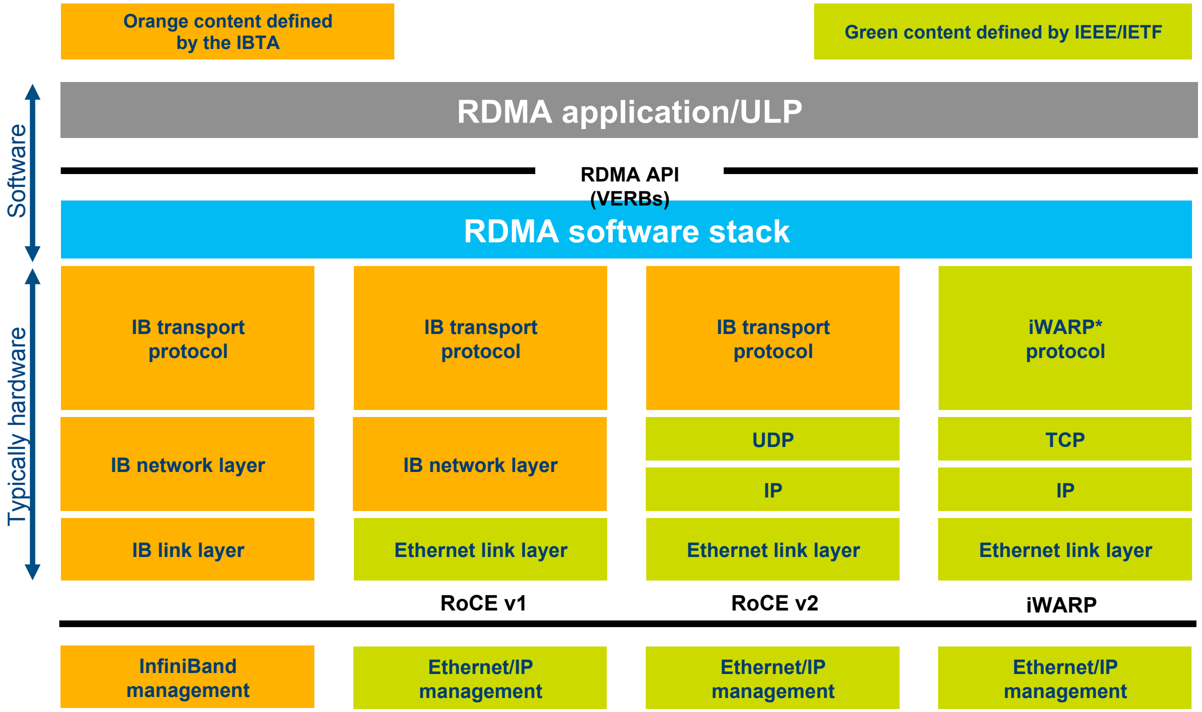
# Similarities and Differences

# RoCE vs. iWARP Network Stack Differences

SNIA ESF | ETHERNET STORAGE

## RoCE Vendors

- Marvell
- Broadcom
- Mellanox Technologies
- Xilinx
- Kazan Networks




## iWARP Vendors

- Marvell
- Chelsio Communications
- Intel
- Kazan Networks

RoCE portion adopted from “Supplement to InfiniBand Architecture Specification Volume 1 Release 1.2.1, Annex A17: RoCEv2,” September 2, 2014

# Key Differences

| | RoCEv2 | iWARP |
|---|---|---|
| **Underlying Network** | UDP | TCP |
| **Congestion Management** | Rely on DCB | TCP does flow control/ congestion management |
| **Adapter Offload Option** | Full DMA | Full DMA and TCP/IP* |
| **Routability** | Yes | Yes |
| **Cost** | Comparable NIC price; Best practice is DCB on switch; DCB configuration experience | Comparable NIC price; Integrated with Intel Platform(4x10Gb); No requirement on switch; |

* Depending on vendor implementations

# Key Differences - RoCE

- Light-weight RDMA transport - RDMA transfers done by the adapter with no involvement by the OS or device drivers
- Based on ECN/DCB (RoCEv1) standards that provide a lossless network and the ability to optimally allocate bandwidth to each protocol on the network
- Scalable to thousands of nodes based on Ethernet technologies that are widely used and well understood by network managers
- Widely deployed by Web 2.0, supported by OS vendors and storage manufacturers
- rNIC demand a slight premium but are becoming commodity NICs

# Key Differences- iWARP

- Built on TCP instead of UDP
- TCP provides flow control and congestion management
  - Can still provide high throughput in congested environment
- DCB is not necessary
- Can scale to tens of thousands of nodes
- Can span multiple hops, or across multiple Data Centers

# Use Cases

# Use Cases - RoCE

SNIA ESF | ETHERNET STORAGE

- **Cloud Computing**
  - Efficient, scalable clustering and higher performance virtualized servers in VMWare, Red Hat KVM, Citrix Xen, Microsoft Azure, Amazon EC2, Google App Engine
- **Storage**
  - Performance increase of 20 to 100% when using RoCE instead of TCP, and latency is typically reduced from 15 to 50% across Microsoft SMD Direct, Ceph and Lustre
- **Big Data / Data Warehousing**
  - Accelerates data sharing/sorting, higher IOPS and linear scaling with exponential growth
  - Ideal for Oracle RAC, IBM DB2 PureScale, and Microsoft SQL
- **Virtualization**
  - VMware ESX and Windows Hyper-V now support inbox drivers to reduced migration time
- **Hyper-Converged (HCI)**
  - Achieve faster performance for storage replication and live migrations
- **Financial Services:**
  - Unleashes scalable CPU performance on low latency applications like Tibco, Wombat/NYSE, IBM WebSphere MQ, Red Hat MRG, and 29West/ Informatica.
- **Web 2.0:**
  - RoCE minimizes response time, maximizes jobs per second, and enables highly scalable infrastructure designs. It’s ideal for applications like Hadoop, Memcached, Eucalyptus, and Cassandra.

# Use Cases - iWARP

- High Performance Computing
  - Low-latency message passing over an Ethernet network
  - Optimized for Open MPI/Intel® MPI
- Storage: Hyper-Converged or Disaggregated
  - Low latency, high throughput
  - Built-in Microsoft SMB Direct, Storage Spaces Direct, Storage Replica
  - Support NVMe over Fabric, Persistent Memory over Fabric
  - Ideally for Hyper-Converged storage due to TCP based flow control and congestion management
- Big Data
  - Accelerates Hadoop MapReduce, SPARK Shuffling
  - Alluxio acceleration
- Virtualization
  - Windows Server Hyper-V
  - Windows VM live migration acceleration

# Summary

| | RoCE | iWARP |
|---|---|---|
| Transport | UPD/IP | TCP/IP |
| Network | Lossless | Standard |
| Adapter | rNIC (soft-RoCE) | NIC |
| Offload | Hardware | Hardware |
| Switch | DCB (resilient RoCE) | Standard |

Our Next Great Storage Debate:

**Centralized vs. Distributed**

September 11, 2018

Register:

https://www.brighttalk.com/webcast/663/332357

# More Webcasts

- Other Great Storage Debates
  - FCoE vs. iSCSI vs. iSER https://www.brighttalk.com/webcast/663/318003
  - Fibre Channel vs. iSCSI: https://www.brighttalk.com/webcast/663/297837
  - File vs. Block vs. Object Storage: https://www.brighttalk.com/webcast/663/308609
- On-Demand “Everything You Wanted To Know About Storage But Were Too Proud To Ask” Series
  - https://www.snia.org/forums/esf/knowledge/webcasts-topics

# After This Webcast

- Please rate this webcast and provide us with feedback
- This webcast and a PDF of the slides will be posted to the SNIA Ethernet Storage Forum (ESF) website and available on-demand at 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: sniaesfblog.org
- Follow us on Twitter @SNIAESF

# Thank You