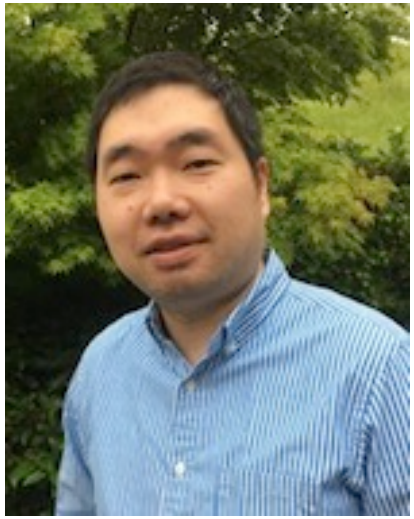
SNIA NSF | NETWORKING STORAGE

Networking Requirements for Ethernet Scale-Out Storage

Live Webcast
November 14, 2018
10:00 am PT

Today's Presenters



Fred Zhang
Intel

Saqib Jang
Chelsio Communications

John Kim
SNIA NSF Chair
Mellanox

SNIA-At-A-Glance

SNIA NSF | NETWORKING STORAGE

SNIA-At-A-Glance

170
industry leading organizations

3,500
active contributing members

50,000
IT end users & storage pros worldwide

Learn more: **snia.org/technical** **@SNIA**

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

- Scale-out Storage Overview
  - What is scalability? Scale-out vs. Scale-up
  - Driving forces behind the growth of Scale-out Storage
  - Different types of Scale-out Storage
- Network requirements for Scale-out Storage
  - East-west traffic, TCP Incast, speed matching
  - Low latency inter-node communications
- All flash Scale-out Storage considerations
- Key takeaways

What Is Scalability? Scale-out vs. Scale-up

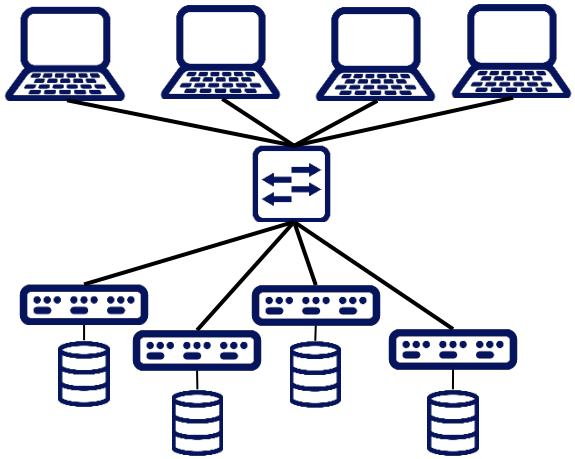
SNIA NSF | NETWORKING STORAGE

- Scalability: capability to expand to support increasing workload (performance/capacity/users)

- Scale-out: add more systems to one cluster
  - Each additional system adds performance and/or capacity
  - Manage multiple systems as one cluster or virtual system
  - May have cluster network

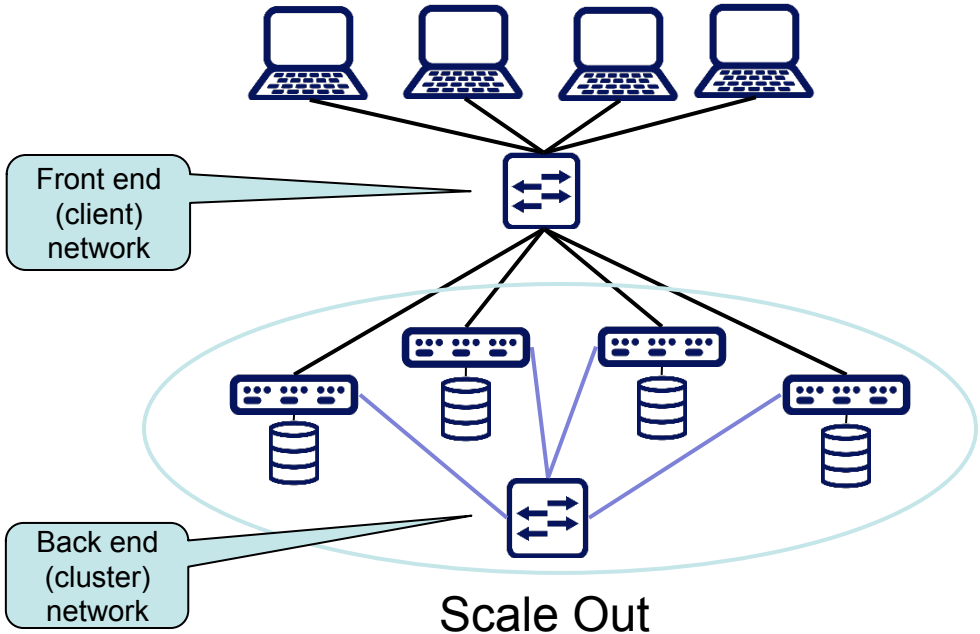
- Scale-up: add more capability to a single system
  - Add more CPUs, faster CPUs, more drives, more memory
  - When reach system maximum, add more individual systems
  - Manage systems separately

Scale-out vs. Scale-up Comparison



Scale Up

Multiple individual systems



Scale Out

Multiple systems managed as one cluster

Scale-out Storage Overview

SNIA NSF | NETWORKING STORAGE

- Refers to multiple types of storage technology
- Software layer presents an aggregate view of storage resources across multiple nodes
- Storage resources may be added to system to provide near-linear growth
- Storage resources are treated as building blocks rather than monolithic entities

Drivers of Scale-out Storage

SNIA NSF | NETWORKING STORAGE

- Data growth
  - Beyond one single system
  - Unstructured data/objects
- Compliance requirements
  - Retention, disaster recovery
- Performance needs
  - Machine learning / AI
  - Analytics
  - Parallel processing
  - Distributed access
- Acquisition costs
  - Might leverage less costly hardware
- Operational costs
  - Many individual systems
  - Scaling performance and capacity at different rates
  - Data locality/access

Scale-Out Storage Driver: Dramatically Improved TCO

- Reclaiming stranded DAS capacity
- Converge data and storage traffic onto IP/Ethernet networks*
- Leverage Ethernet economics and commodity servers*
- Leverage commodity servers*

*Many—but not all—scale-out storage systems use Ethernet for the cluster and client networks. Some scale-out storage solutions run on commodity servers.

Types of Scale-out Storage

- Clustered storage appliances
- Parallel File System
- Object storage
- Distributed big data, e.g. Hadoop
- Hyperconverged infrastructure

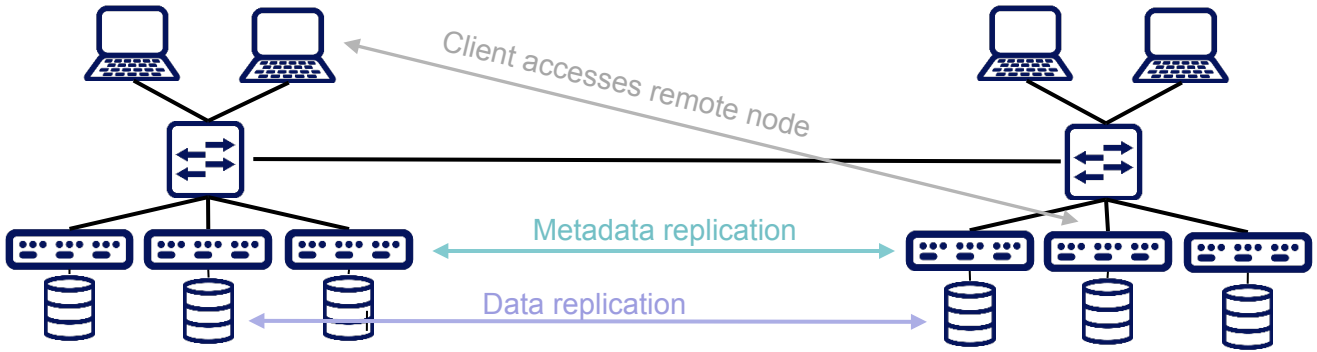
Scale-Out Flash Storage

- Centralized flash management
  - Present distributed flash as one system image
  - Easier flash disaggregation
- Popular with large customers
  - Hyperscalers, service providers, large enterprise
- NVMe-oF
  - Enables networked flash performs similar to local flash
- Storage is faster
  - Often for database/ transactional workloads
  - Higher system performance expectations
- Need faster network
  - More bandwidth, lower latency
  - More likely to mix network speeds

Different Network Traffic for Scale-Out Storage

SNIA NSF | NETWORKING STORAGE

- Distributed Data Access
  - Stripe for performance
  - Distributed read / write
  - Distributed compute
- Coherence/Protection
  - Replication / backup
  - Object erasure coding
  - Metadata consistency



Different Networking Needs

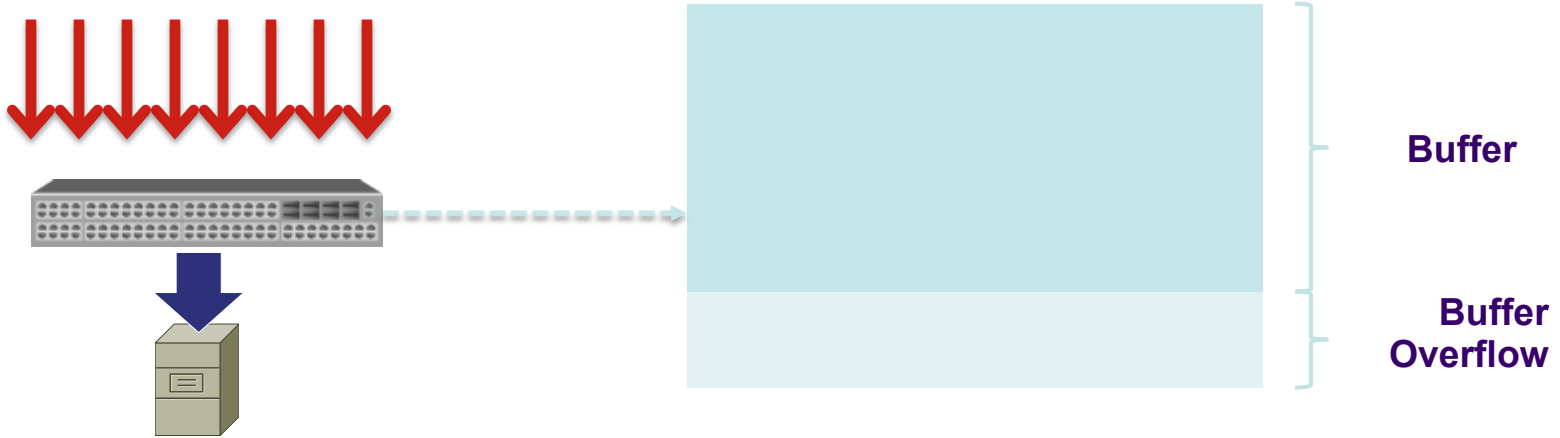
- Internal/cluster network
  - For replication, coherence, monitoring/heartbeat
  - Ethernet, InfiniBand, PCIe, or proprietary
- External network for client access
  - Ethernet, Fibre Channel, or InfiniBand
- Performance needs for data access
  - High Bandwidth for large files/objects, sequential I/O
  - Low Latency for Random I/O, databases, analytics, metadata

Networking Requirements: Massive East-West Traffic

- DAS on compute nodes is a pool of storage resources
- Any-to-any connectivity for storage traffic
- Concept of locality so as to minimize network traffic
- New layer of east-west traffic

Understanding TCP Incast

- Synchronized TCP sessions arriving at common congestion point (all sessions starting at the same time)
- Each TCP session will grow window until it detects indication of congestion (packet loss in normal TCP configuration)
- All TCP sessions back off at the same time



Networking Requirements for Ethernet-based Scale-Out Storage

Networking Requirements: TCP Incast

SNIA NSF | NETWORKING STORAGE

- Many-to-one communication problem that can occur in data networks
- A single request for data can result in responses from many storage nodes simultaneously oversaturating host connection
- When many hosts access the same storage device simultaneously, it creates Incast at storage device

Networking Requirements: Storage-Networking Speed Mismatch

SNIA NSF | NETWORKING STORAGE

- Enterprise networks mix 1, 10, 25, 40, and 100Gbps
  - Example: server connected at 10Gbps but storage system connected at 40Gbps
  - Host request sent at 10Gbps, responses come back at 40Gbps
- Massive speed mismatch can cause buffer exhaustion
  - Increases risk of Incast and congestion
  - Full buffers → dropped packets → slower app. performance
  - Commonly referred to as “Slow Drain”

Networking Requirements: Low-Latency Communications

SNIA NSF | NETWORKING STORAGE

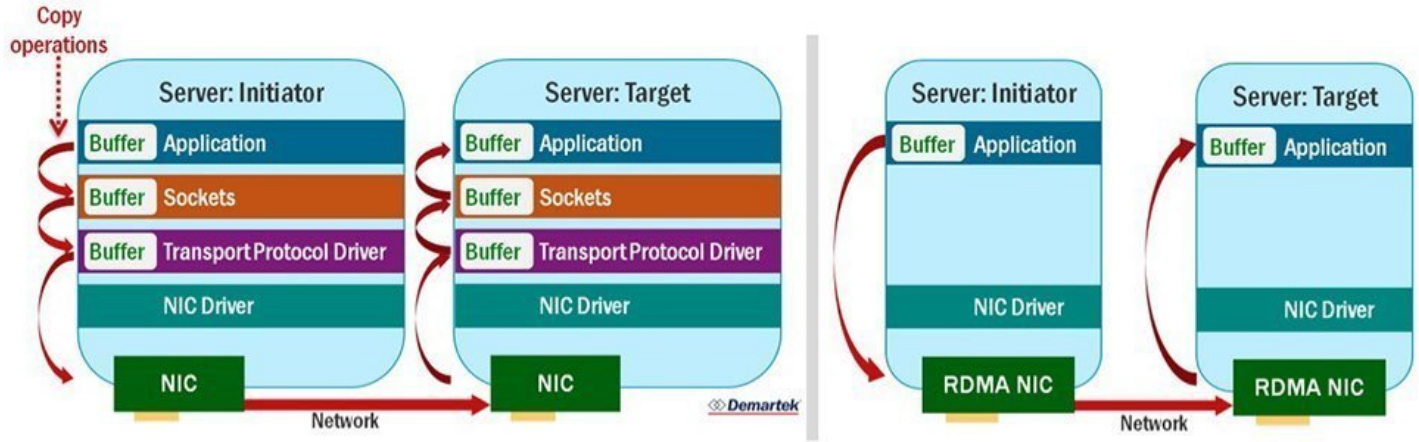
- Moving data between nodes
  - Data protection in order to rebuild after a failure
  - Balance the system as nodes are added or upgraded
- Remote access
  - Requests to node A for data that's stored on node D
  - Metadata access and synchronization
- Initiator access for latency-sensitive workloads
  - Random I/O, OLTP, analytics

Networking Recommendations (1)

SNIA NSF | NETWORKING STORAGE

- Scalable leaf-spine non-blocking switching
  - Deterministic latency, any-to-any non-blocking
  - Address east-west traffic scalability
- Increased buffer sizes in switches
  - Enables lossless delivery in face of speed mismatch
  - Potential solution to TCP Incast
- Use of Data Center Bridging (DCB)
  - Uses Explicit Congestion Notification (ECN)
  - Addresses TCP Incast by preventing buffers from overflowing

Networking Recommendations (2): RDMA Networking



- Direct data movement in and out of server
- Bypass storage software stack and buffer copy operations
- Dramatically reduced latency and improved CPU performance
- NVMe-oF (RDMA) optimal low-latency transport for SOFS

Key Takeaways

- Scale-out storage is increasingly popular
- Networking challenges
  - Generates more/different east-west traffic between nodes
  - Need high bandwidth and/or low latency, especially with flash
  - Incast, Speed mismatches
- Need the right network
  - Higher speed adapters, modern switches
  - Direct data placement DMA technology

Upcoming NSF Webcasts

SNIA NSF | NETWORKING STORAGE

Virtualization and Networking Storage Best Practices

January 17, 2019

Register at: https://www.brighttalk.com/webcast/663/337602

Networking Requirement for Hyperconvergence

February 5, 2019

Register at: https://www.brighttalk.com/webcast/663/341209

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 Networking Storage Forum (NSF) website and available on-demand at www.snia.org/library
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-NSF blog: sniansfblog.org
- Follow us on Twitter @SNIANSF

SNIA NSF | NETWORKING STORAGE

Thanks!