Education

Scale-Out Storage

Nick Kirsch
Director of Product Management, Isilon Systems

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- 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.

Abstract

Scale-Out Storage

- This is a overview of scale-out storage systems and their underlying file system technologies - primarily focused on network-attached storage systems. In this presentation, scale-out will be defined in contrast with scale-up storage systems, the market, user, and technology needs driving new class of storage systems will be explained, as well as a survey of open-source and commercial implementations available today.

Refer to Other Tutorials

Check out SNIA Tutorial:

Storage Tiering and the

Impact of Flash on File Systems

Check out SNIA Tutorial:

The File Systems Evolution

Check out SNIA Tutorial:

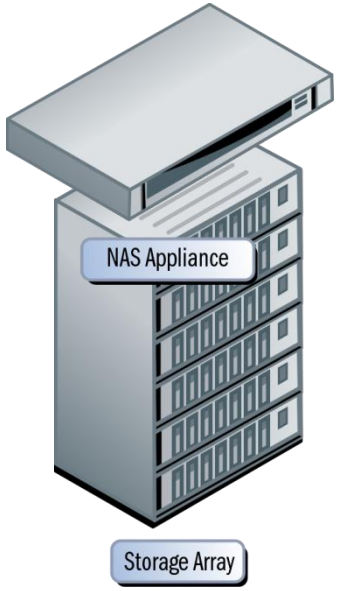
Aspects of Deduplication

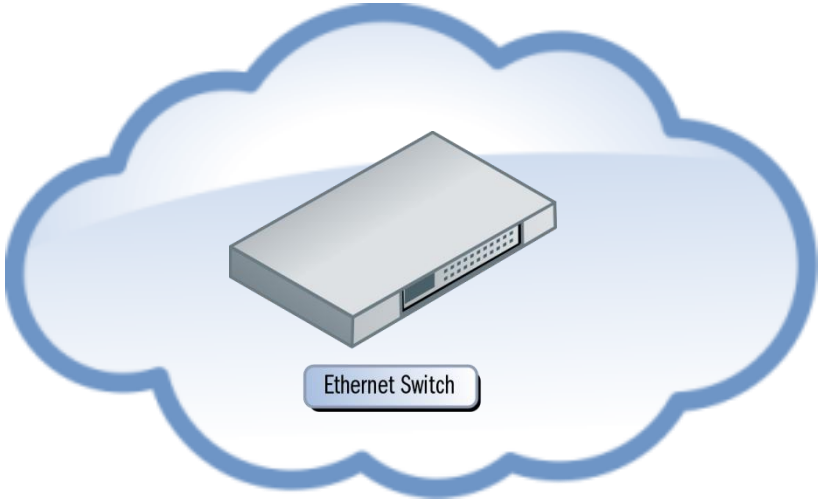
Check out SNIA Tutorial:

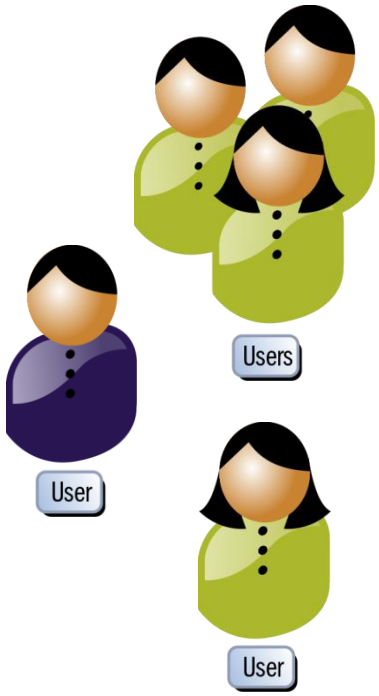
File Systems for Object Storage Devices

Network-Attached Storage

- Simplicity – No Client Configuration Required
- Network Transparency – Standard Protocols
- NAS Server is typically called a **Head**
- NAS Servers typically deployed as a **Pair**

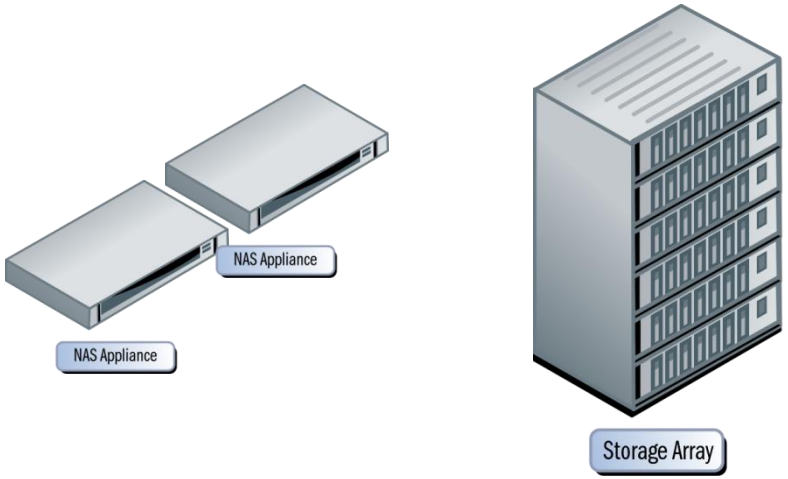






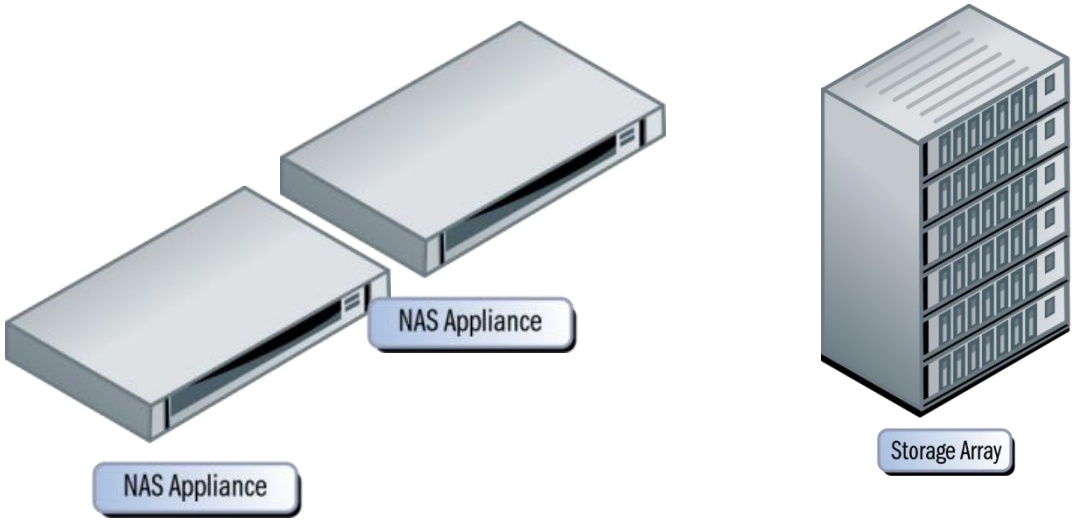
"Scale-Up"

- Deploy More Powerful NAS Pairs
- Storage Can Be Retained and Expanded
- Volume Performance Limited to Single NAS Head



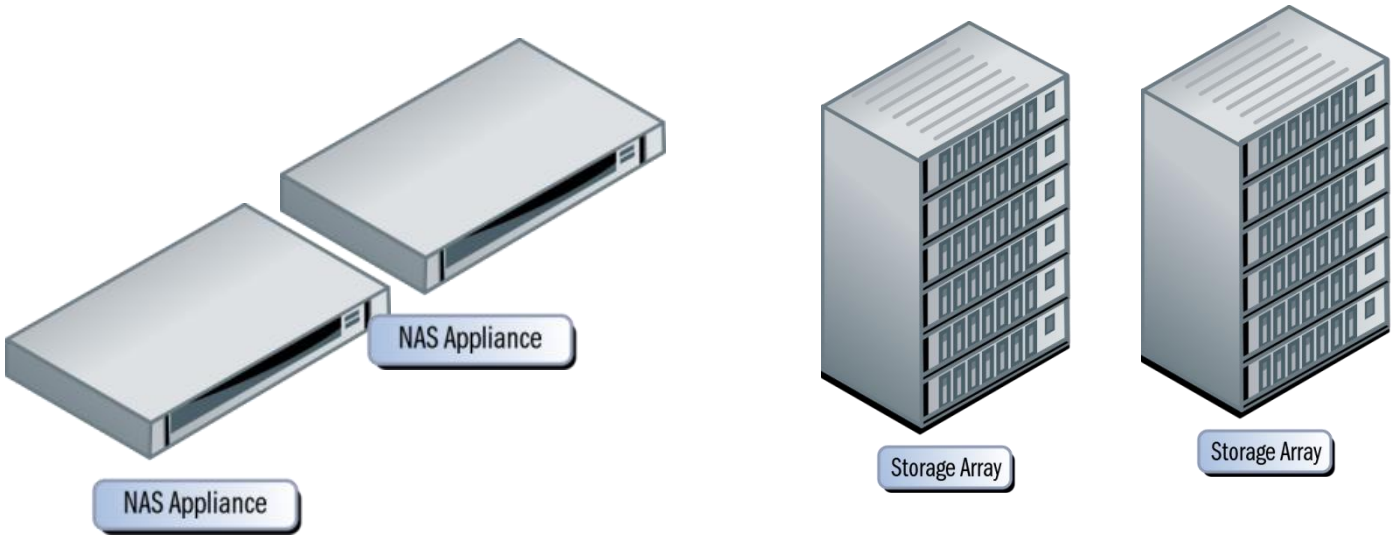
"Scale-Up"

- Deploy More Powerful NAS Pairs
- Storage Can Be Retained and Expanded
- Volume Performance Limited to Single NAS Head



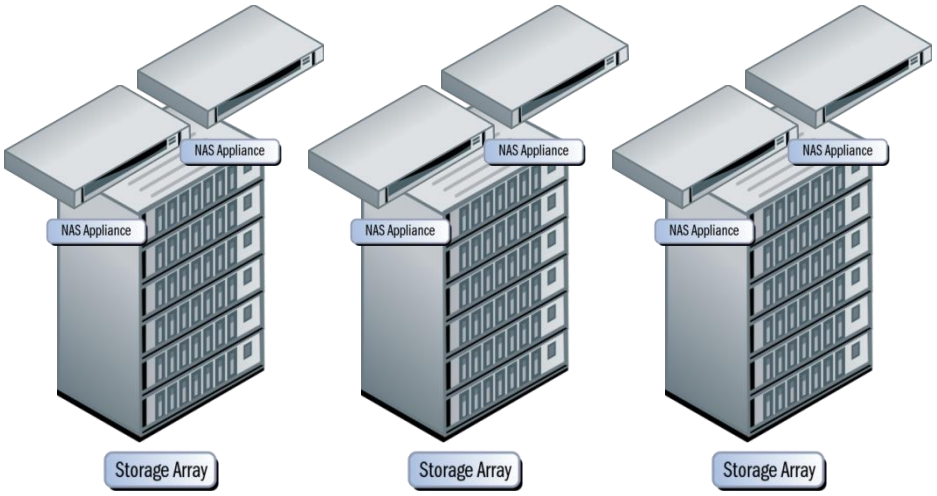
"Scale-Up"

- Deploy More Powerful NAS Pairs
- Storage Can Be Retained and Expanded
- Volume Performance Limited to Single NAS Head



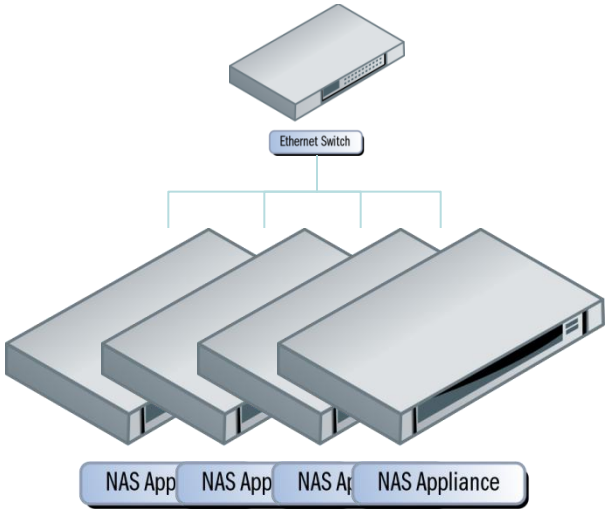
Horizontal Scaling

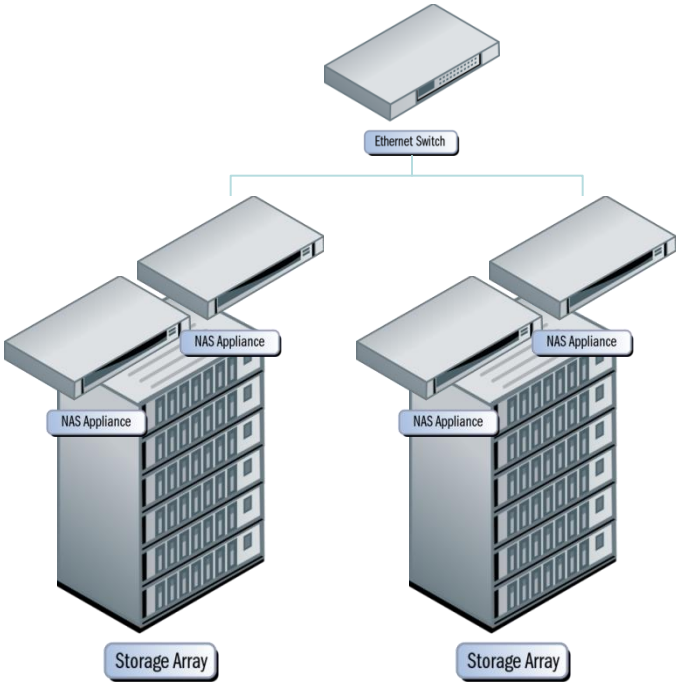
- Deploy Additional NAS Pairs
- Break Workflow Between NAS Pairs
- Duplicate Data Between NAS Pairs
- No Coupling Between NAS Pairs



Scale-Out Systems

- Distributed/Aggregated System
- Tight Coupling Between NAS Servers
- NAS Servers Referred to as **Nodes**





Forces Driving Scale-Out Category

- Deployment/Capital Costs
  - Storage Needs Difficult to Size Initially
  - Difficult to Fully Utilize All Resources
- Management
  - Horizontal Scaling Increases Management Complexity
- Performance
  - Systems Cannot be Scaled-Up Effectively
  - Workflows Cannot be Segmented
- Reliability
  - Redundancy Typically Limited to 2-way

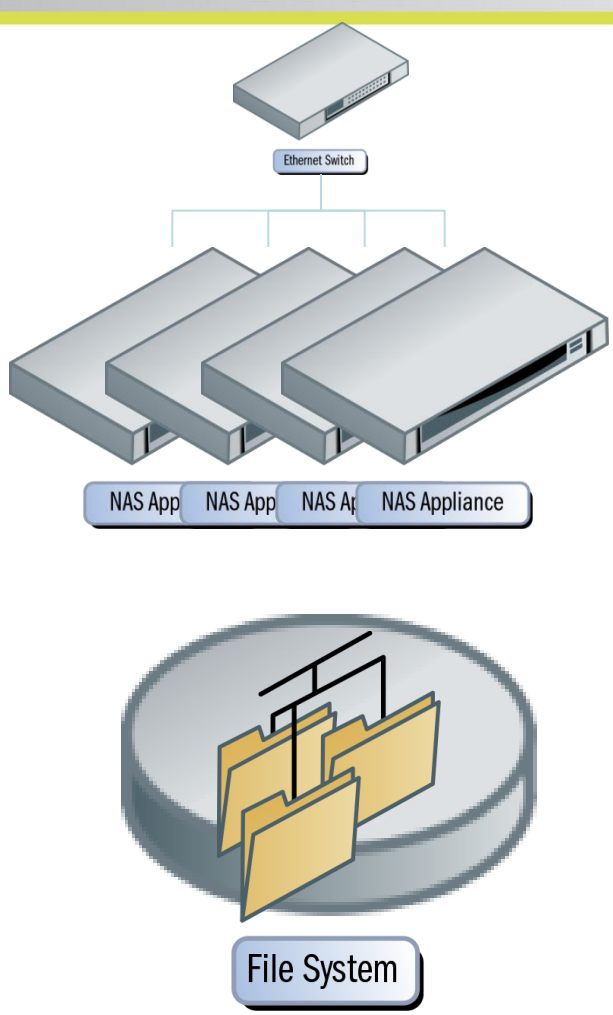
Scale-Out: Architecture

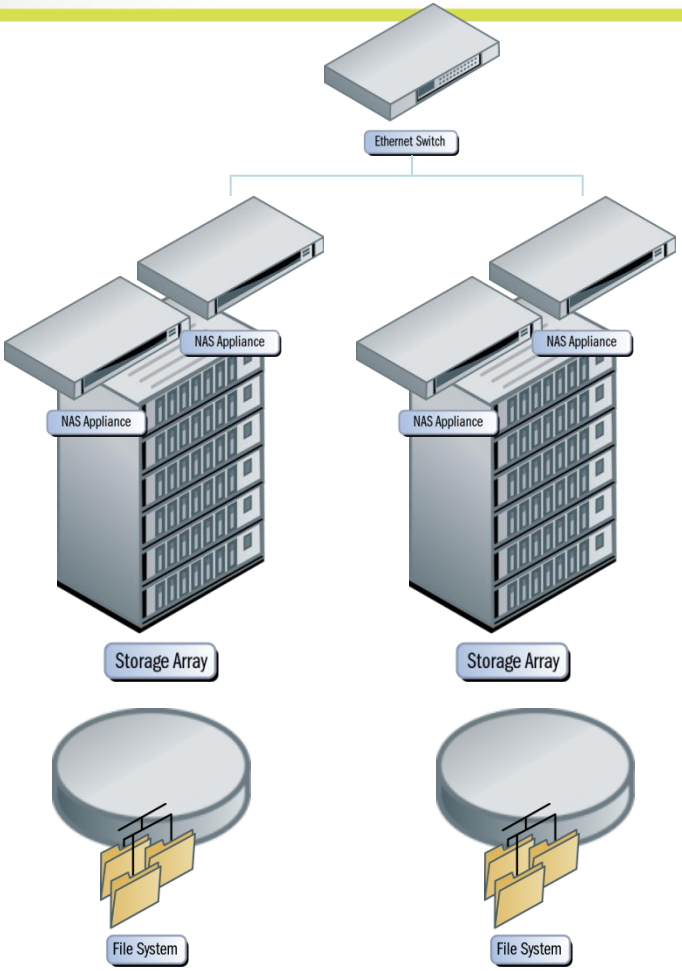
- Distributed System
  - Degree of Coupling Varies
- Single Namespace
- High Availability and Data Protection
- Management Simplicity
- Investment Protection
- Ease of Scale
- Architecture choices drive many details!

Scale-Out: Volumes

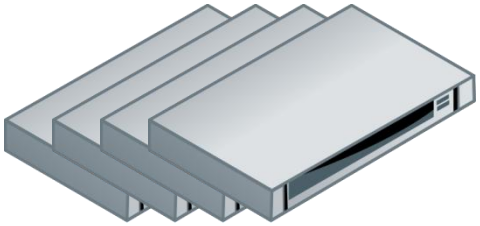
- Single Namespace Presentation to Clients
  - No Client Software Required For Presentation
  - No Client Setup Required for Presentation
- Scale-Out Volume
  - Single Volume
  - Multiple Volumes
- Storage Efficiency
- Locking Semantics
- Multi-Protocol Semantics

Scale-Out: Namespace/Volume

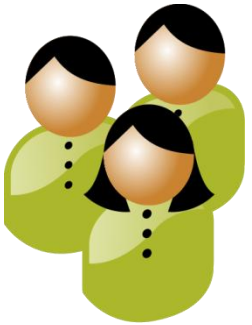




Scale-Out: Namespace

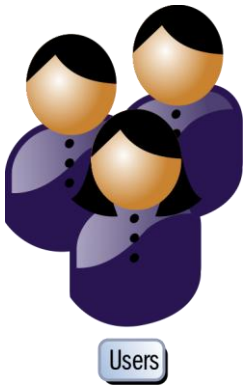


Scale-Out NAS

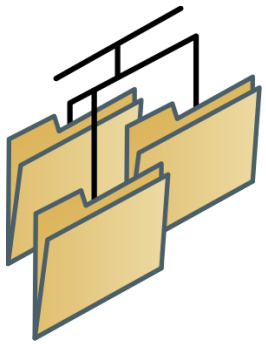


Users

User



Users

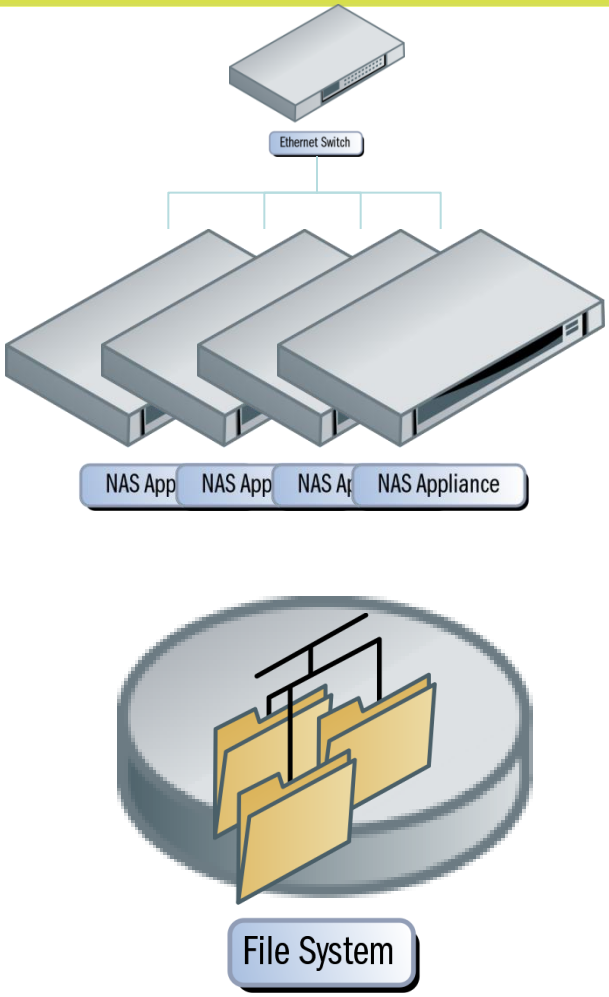


File System

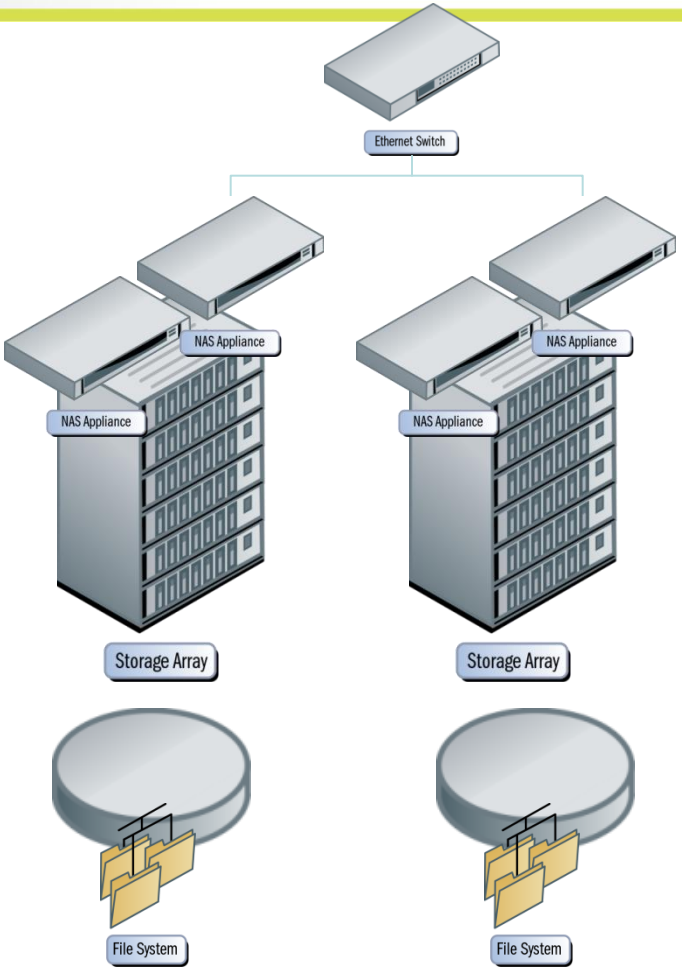
Scale-Out: Data Protection

- Must Sustain Both Node Failures and Drive Failures
- Distributed Data Protection
  - Mirror Blocks Between Nodes for Redundancy
  - Generate FEC Blocks and Distribute Between Nodes
- Conventional HW/SW RAID Techniques
  - Use RAID techniques within Node
  - Node typically looks like a NAS Pair
- Data Protection Granularity
  - File-Level – different files can have different protections
  - Block-Level – protection is dictated at volume level

Scale-Out: Data Protection

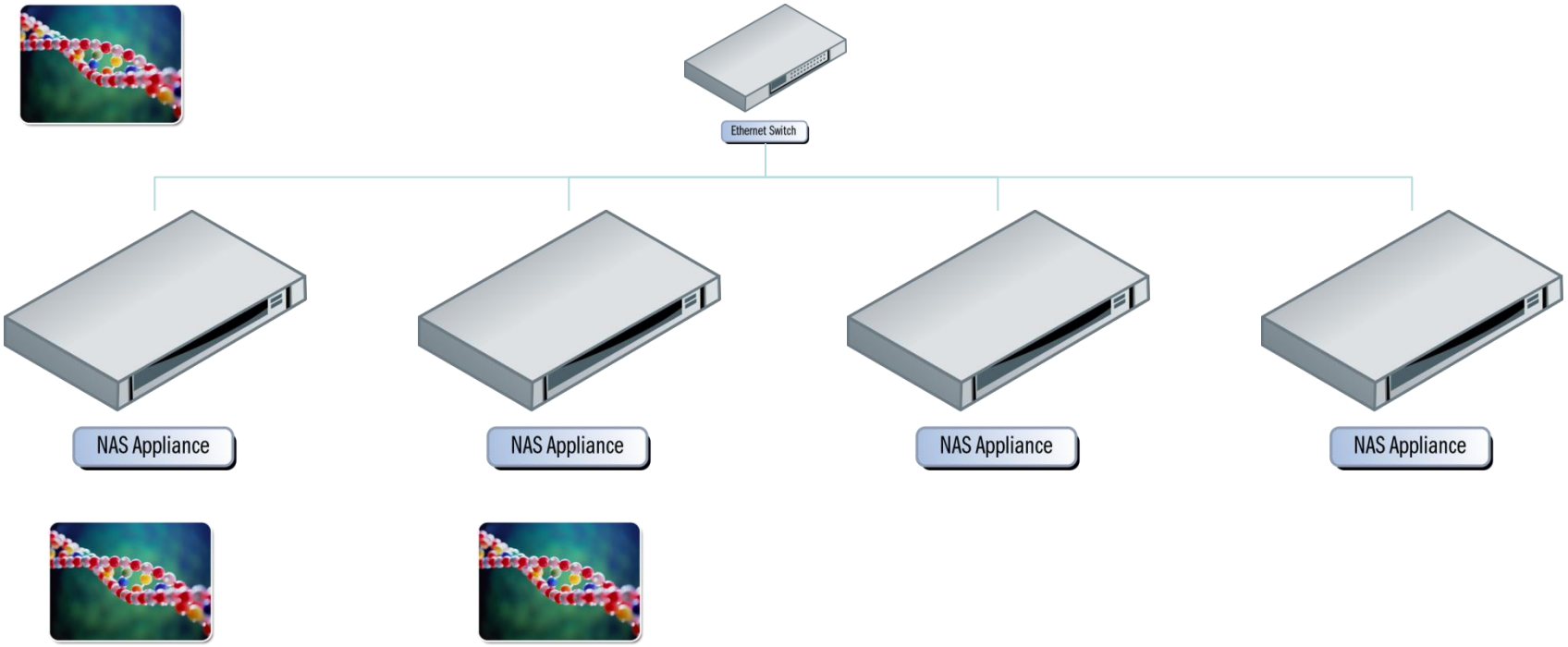


Distributed Protection



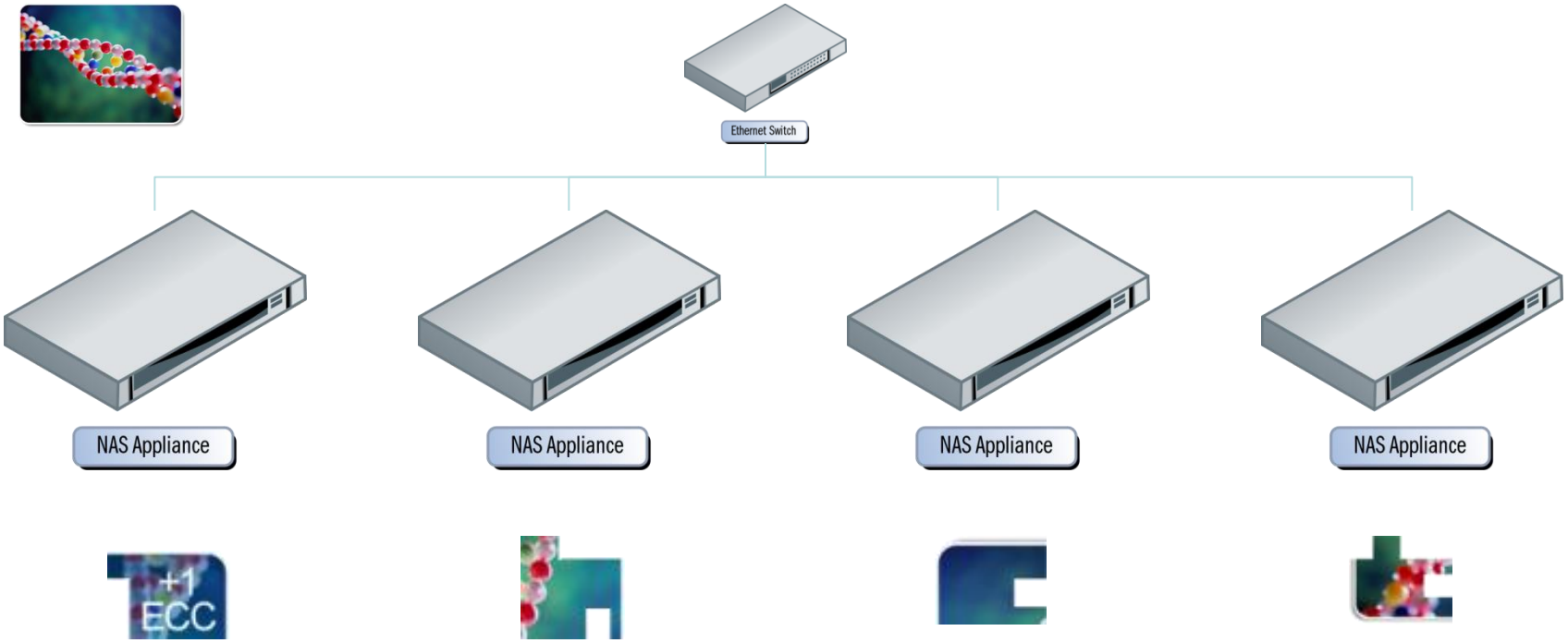
Conventional Protection

Scale-Out: Data Protection



Mirror Blocks Between Nodes for Redundancy

Scale-Out: Data Protection



Stripe files with FEC (forward-error-correction) protection

Scale-Out: Hardware

- All-In Appliance Based Architectures
  - Disk, CPU and Memory Fully Contained Nodes
  - Expand By Adding Appliances
  - Near-Commodity and Custom Chassis
- Traditional Head/Shelf Appliance Architecture
  - Paired NAS Head + Disk Shelves
  - Expand By Adding Pairs and/or Shelves
  - Near-Commodity and Custom Chassis
- DIY/BYO
  - Commodity Server + Disk Shelves
  - Highly Flexible Arrangements

Scale-Out: Storage Interconnect

- Distributed Systems Require Fast-Interconnects
  - High Throughput
  - Low-Latency
- Interconnect Typically Private, Self-Managed

- Infiniband
- 10GbE
- Myrinet

Scale-Out: Node Software

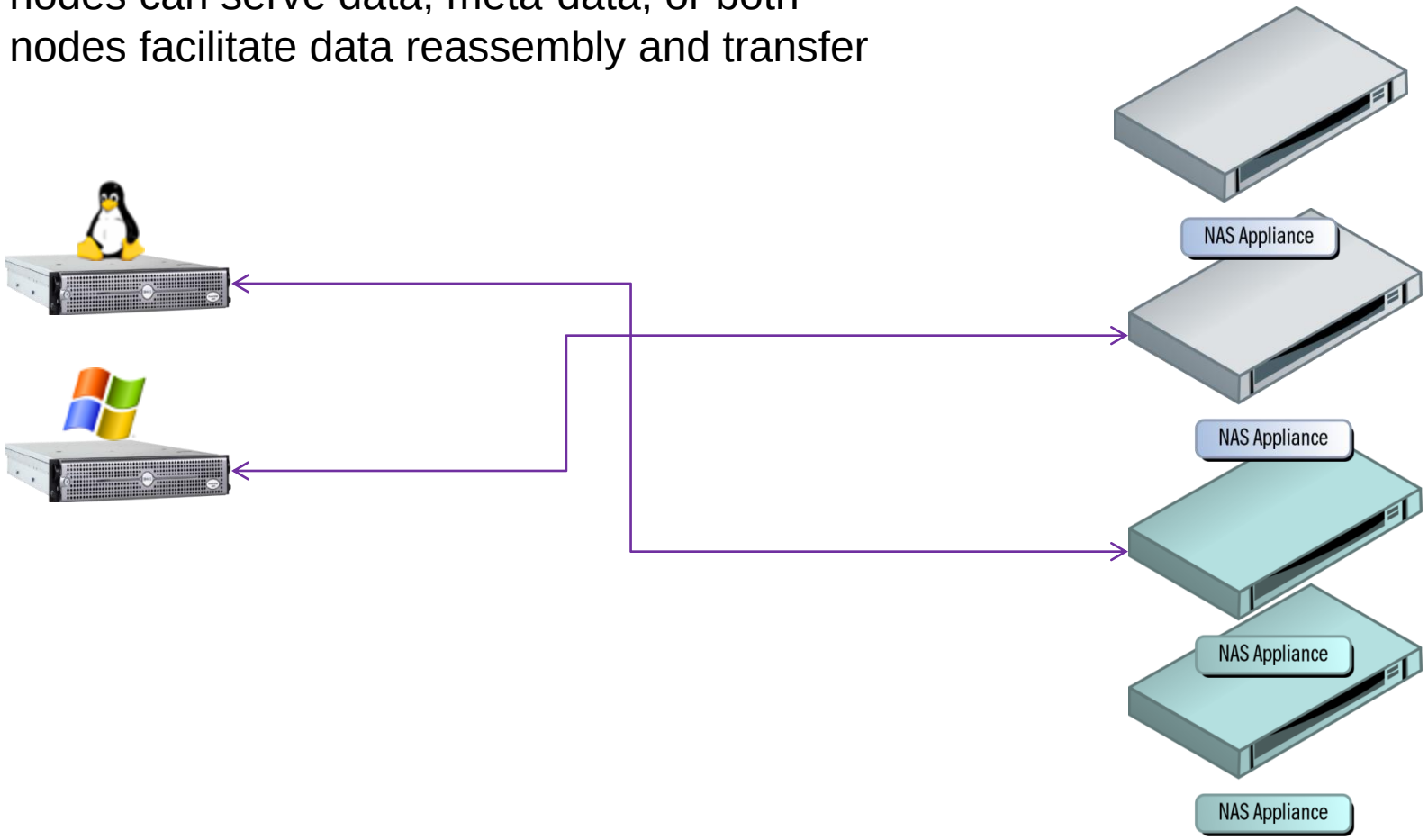
- System Software
  - Identical Software Versions
  - In-Family Software Versions
  - Out-of-Family Software Versions
- Protocol/Data Servers
  - Split Data/Meta-Data Nodes
  - Data/Meta-Data Pod w/ Accessibility
  - Distributed Data/Meta-Data

Scale-Out: Client Access

- NFS/CIFS/iSCSI (all)
  - Can be balanced using Round-Robin, DNS Delegation
  - MPIO/iSCSI Specific Drivers
- Client-Side Drivers (optional)
  - Performance Benefits
  - Load-balancing
- Customer-Aligned Clients (optional)
  - Manually Align Client to Data

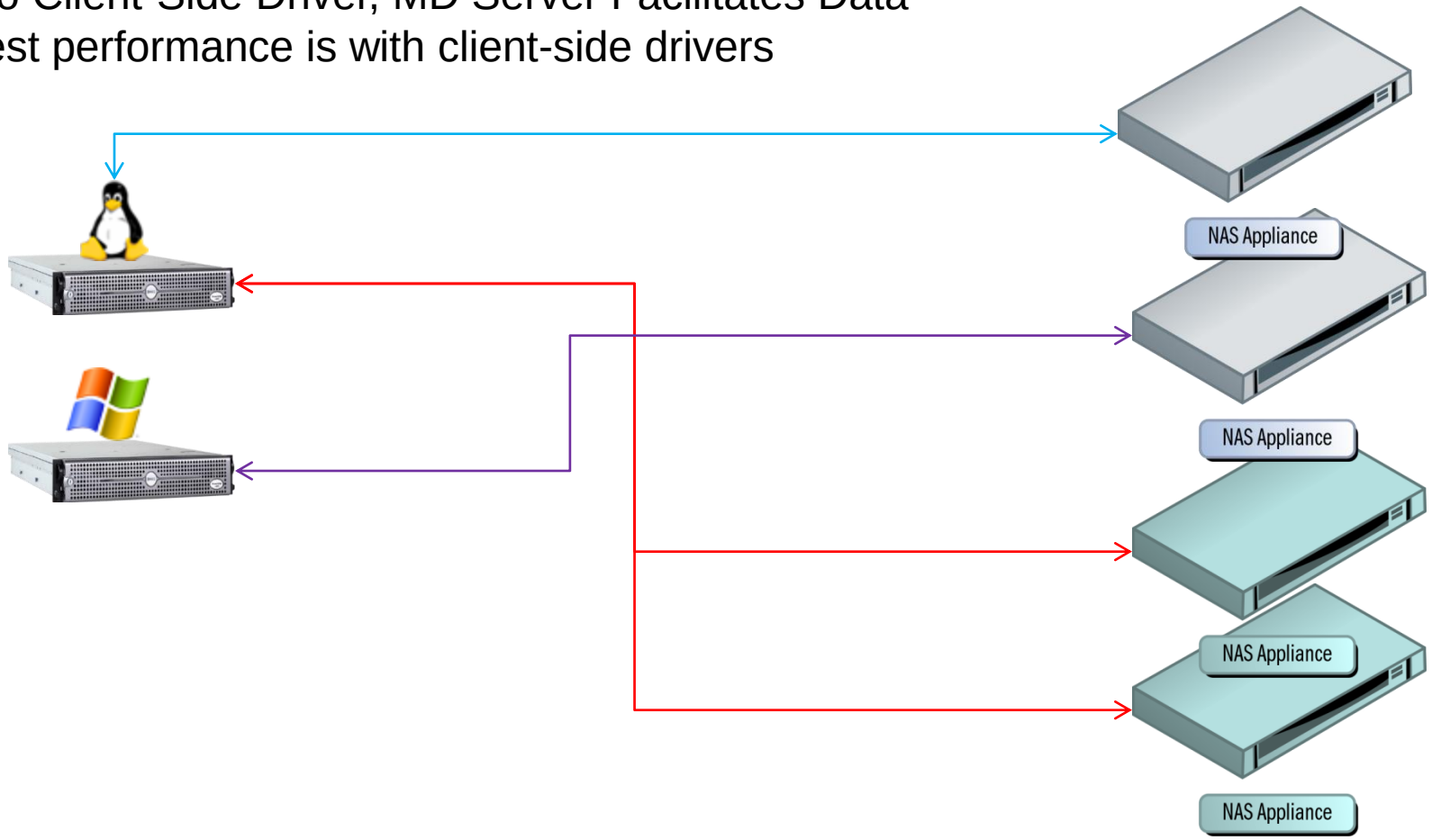
Scale-Out: Shared Data/MD Node

Data is striped across nodes
All nodes can serve data, meta-data, or both
All nodes facilitate data reassembly and transfer



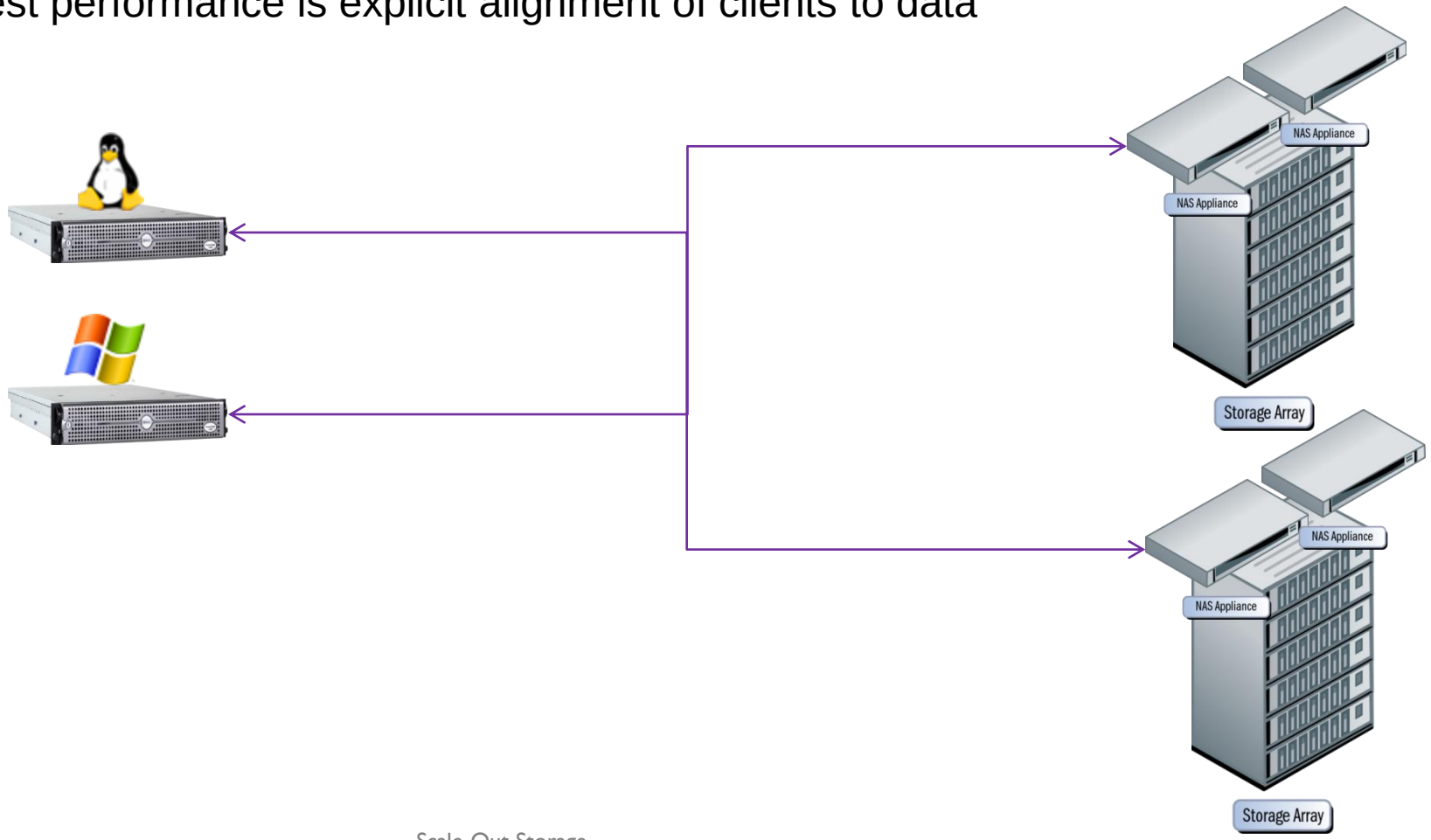
Scale-Out: Dedicated MD Servers

Client-Side Drivers Split Data and Meta-Data Streams
w/o Client-Side Driver, MD Server Facilitates Data
Best performance is with client-side drivers



Scale-Out: Explicit Alignment

All nodes can service data/meta-data
Data not striped across nodes
Best performance is explicit alignment of clients to data



Scale-Out: Performance

- Cache Semantics
  - Globally Accessible Caches
  - Node-Local Caching
  - Block-Indexed Caches
  - File-Indexed Caches
  - Double-Caching
- Throughput and I/O Characteristics
  - Large-Block I/O (throughput)
  - Small-Block I/O
  - Transactional/Latency-Sensitive I/O

Scale-Out: Management

- Dependency on Volume Configuration
- Managing a Single System
- Managing Multiple Systems within a Single System
- Life Cycle
  - Initial Configuration
  - Node Failure/Replacement
  - Storage Expansion
  - Node Addition

Q&A / Feedback

- Please send any questions or comments on this presentation to SNIA: trackfilemgmt@snia.org

Many thanks to the following individuals for their contributions to this tutorial.

- SNIA Education Committee

Nick Kirsch