



Education

# The File Systems Evolution

Christian Bandulet, Sun Microsystems

# SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
  - ◆ Any slide or slides used must be reproduced without modification
  - ◆ The SNIA must be acknowledged as 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 nor should be construed as legal advice or opinion. If you need legal advice or legal opinion please contact an attorney.
- The information presented herein represents the Author's personal opinion and current understanding of the 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.**

## ➤ The File Systems Evolution

- ◆ File Systems impose structure on the address space of one or more physical or virtual devices. Starting with local file systems over time additional file systems appeared focusing on specialized requirements such as data sharing, remote file access, distributed file access, parallel files access, HPC, archiving, security etc.. Due to the dramatic growth of unstructured data files as the basic units for data containers are morphing into file objects providing more semantics and feature-rich capabilities for content processing. This presentation will categorize and explain the basic principles of currently available file systems (e.g. local FS, shared FS, SAN FS, clustered FS, network FS, WAFS, distributed FS, parallel FS, object FS, ...). It will also explain technologies like NAS aggregation, NAS clustering, scalable NFS, global namespace, parallel NFS, storage grids and cloud storage. All of these files system categories are complementary. They will be enhanced in parallel with additional value added functionality. New file system architectures will be developed and some of them will be blended in the future.

# Check Out Other Tutorials



**Check out SNIA Tutorial:  
DFS Over CIFS**



**Check out SNIA Tutorial:  
Storage Tiering for File &  
NAS Systems**



**Check out SNIA Tutorial:  
NAS and iSCSI Technology  
Overview**



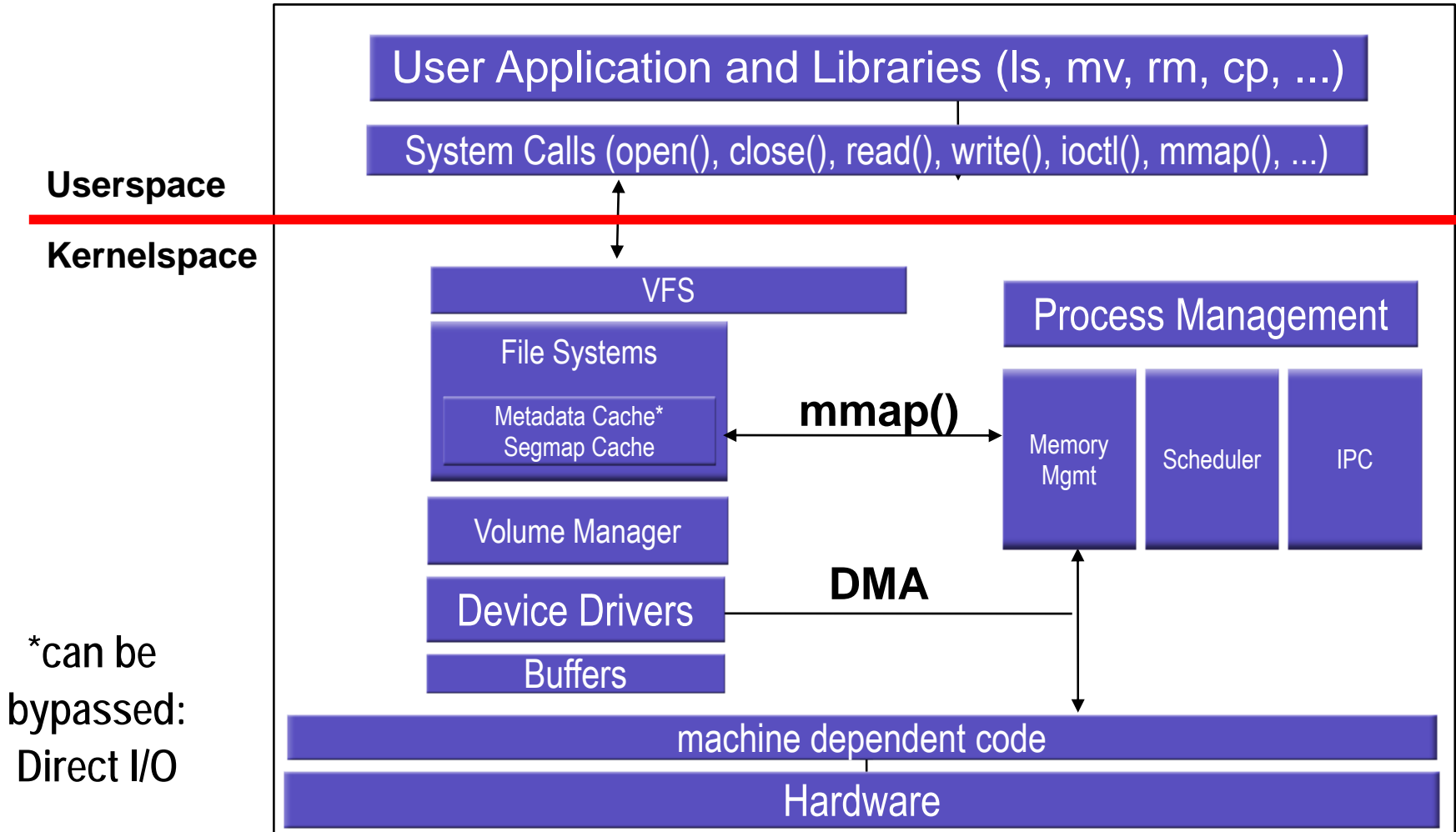
**Check out SNIA Tutorial:  
Find and Select the Right File  
Storage for Your Application**



**Check out SNIA Tutorial:  
Scaling NFS Through pNFS**

- **File System Basics**
- File Systems Taxonomy
- Local FS
- Shared FS/Global FS
  - ◆ SAN FS, Cluster FS
- Network FS
- Scalable NAS / Scalable NFS
- Wide Area FS
- Distributed FS
- File Virtualization
- Distributed Parallel FS
- NAS Cluster / NAS Grid
- FS Future Developments

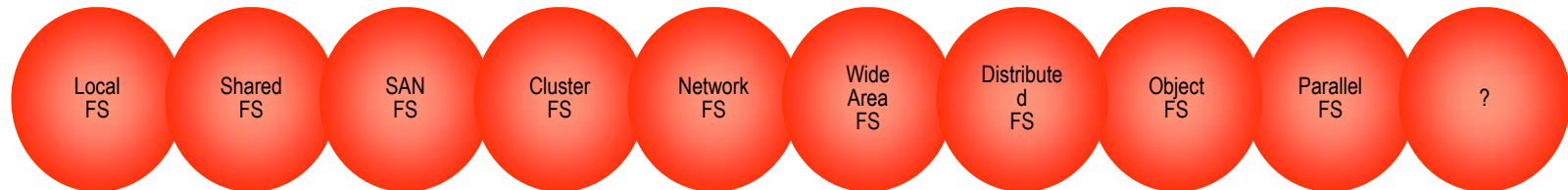
# File System & Operating System



- File System Basics
- **File Systems Taxonomy**
- Local FS
- Shared FS/Global FS
  - ◆ SAN FS, Cluster FS
- Network FS
- Scalable NAS / Scalable NFS
- Wide Area FS
- Distributed FS
- File Virtualization
- Distributed Parallel FS
- NAS Cluster / NAS Grid
- FS Future Developments

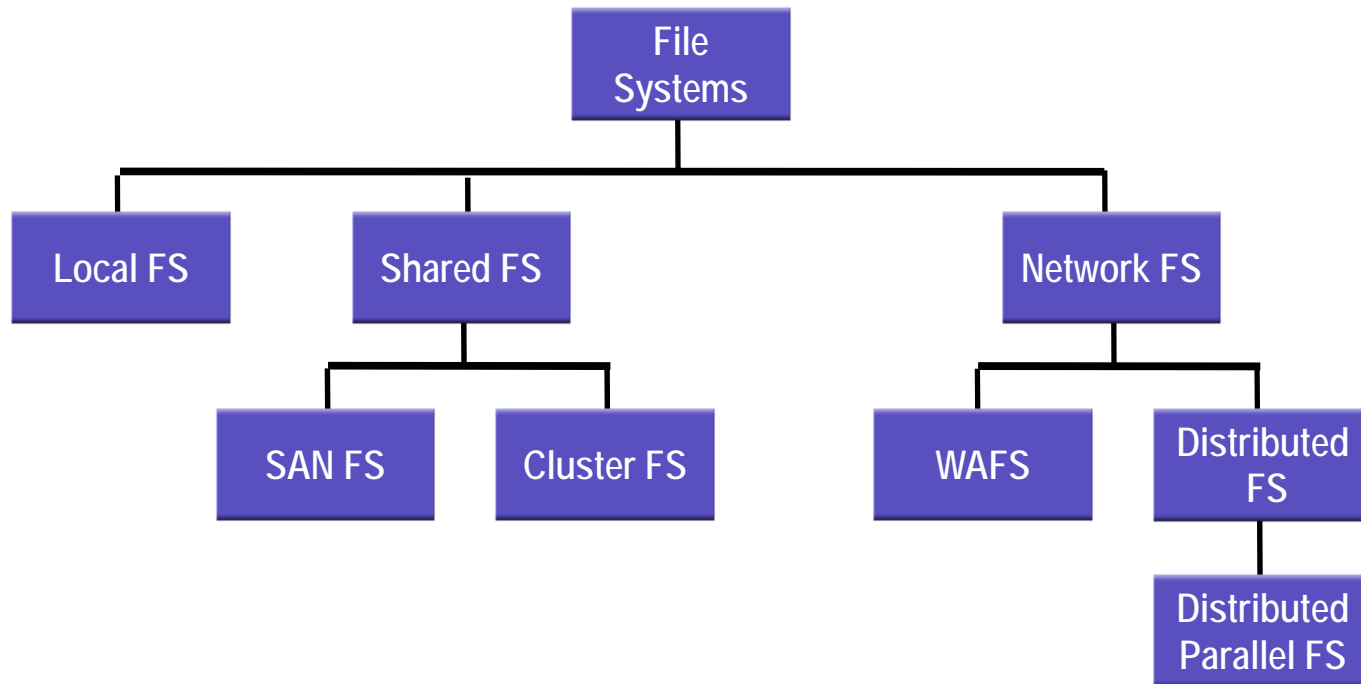
# The File Systems Evolution

- File systems evolved over time
- Starting with local file systems over time additional file systems appeared focusing on specialized requirements such as data sharing, remote file access, distributed file access, parallel files access, HPC, archiving, etc.

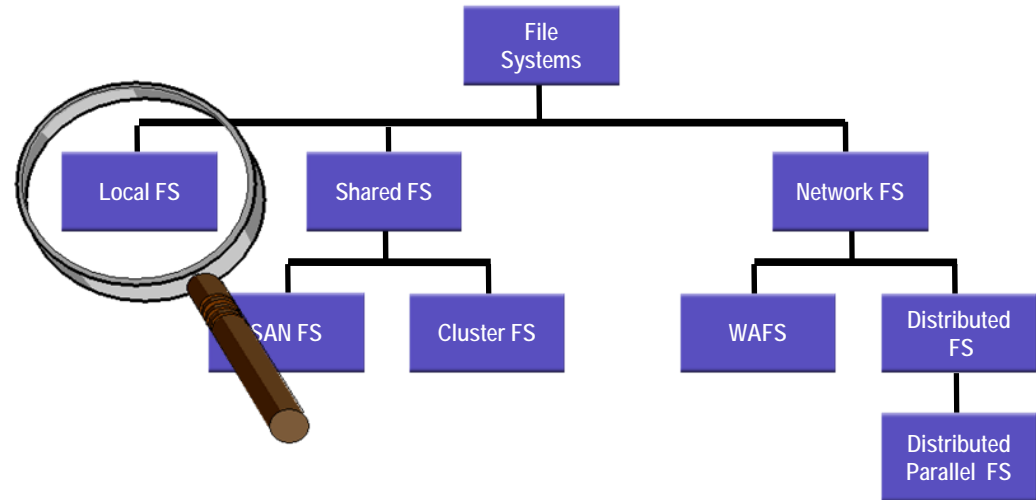


Note: The picture above does not reflect the exact sequence in which the files system types appeared. Some of them actually appeared in parallel. It is also not the intention to indicate that a new file system replaces its predecessors. Instead they are targeting complimentary objectives.

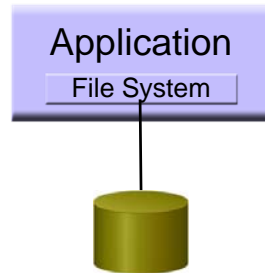
# File System Taxonomy



- File System Basics
- File Systems Taxonomy
- **Local FS**
- Shared FS/Global FS
  - ◆ (SAN FS, Cluster FS)
- Network FS
- Scalable NAS / Scalable NFS
- Wide Area FS
- Distributed FS
- File Virtualization
- Distributed Parallel FS
- NAS Cluster / NAS Grid
- FS Future Developments



## Local FS



➤ FS is **co-located** with application server

## Local FS



## Local FS



## Local FS

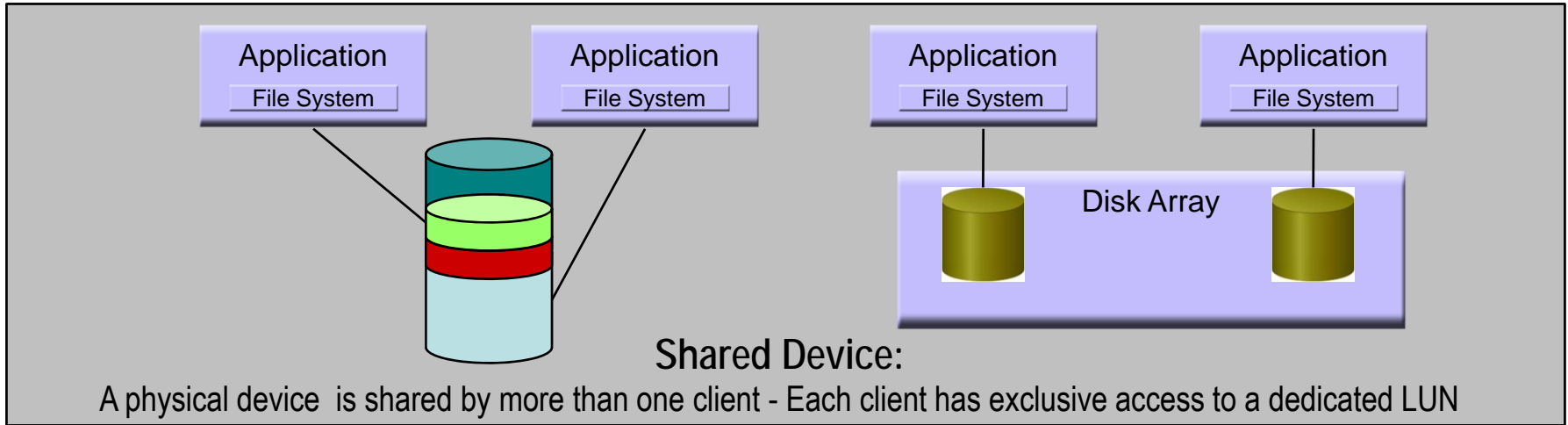


## Local FS

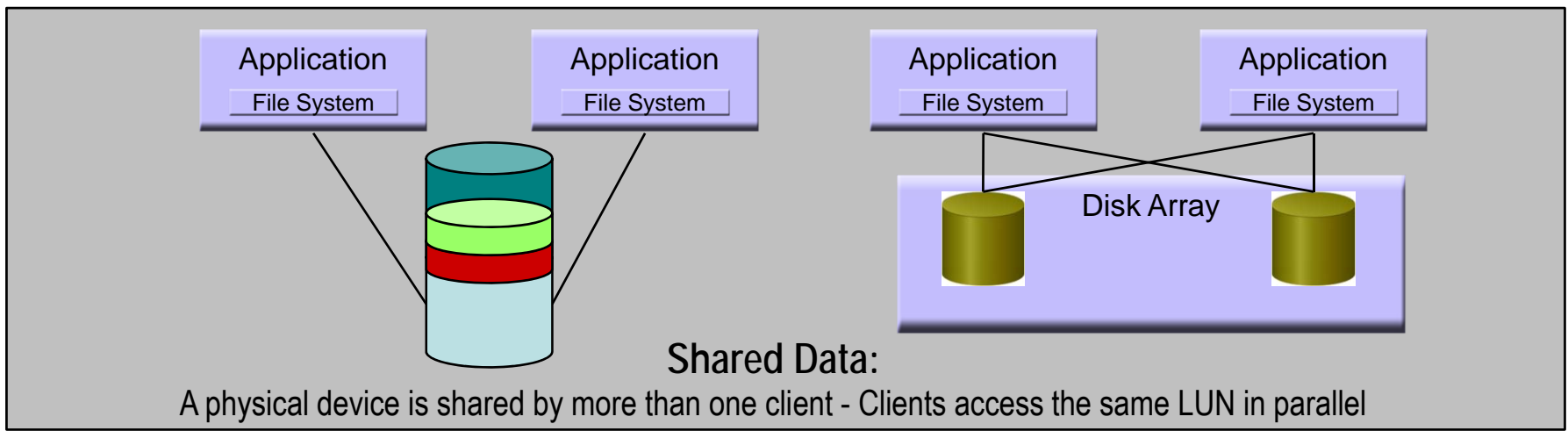
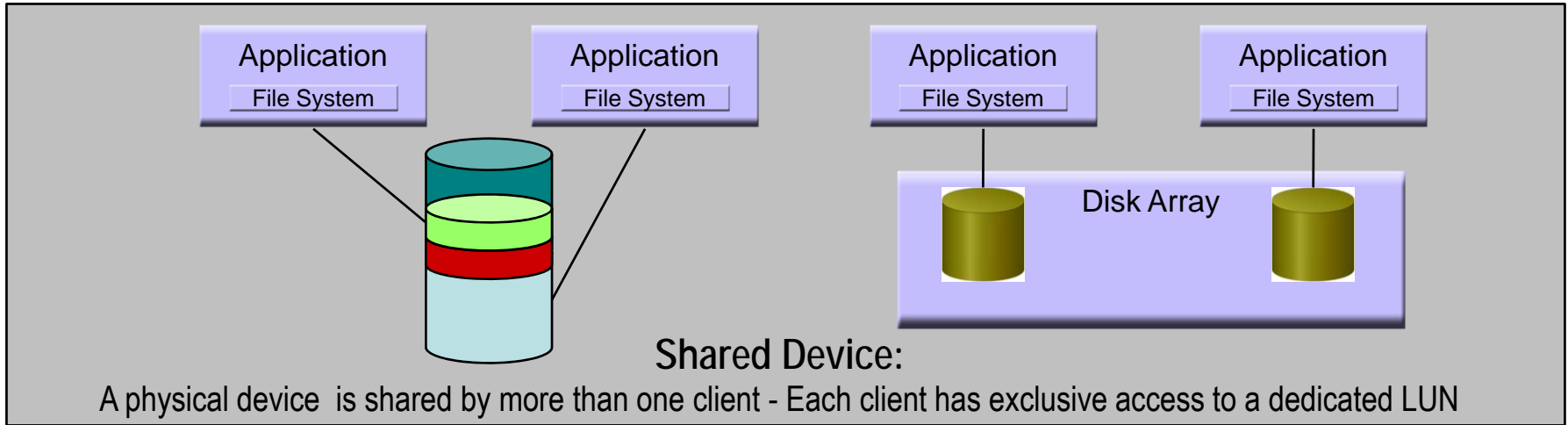


➤ **Islands of storage** (no data sharing)

# Shared Device vs. Shared Data

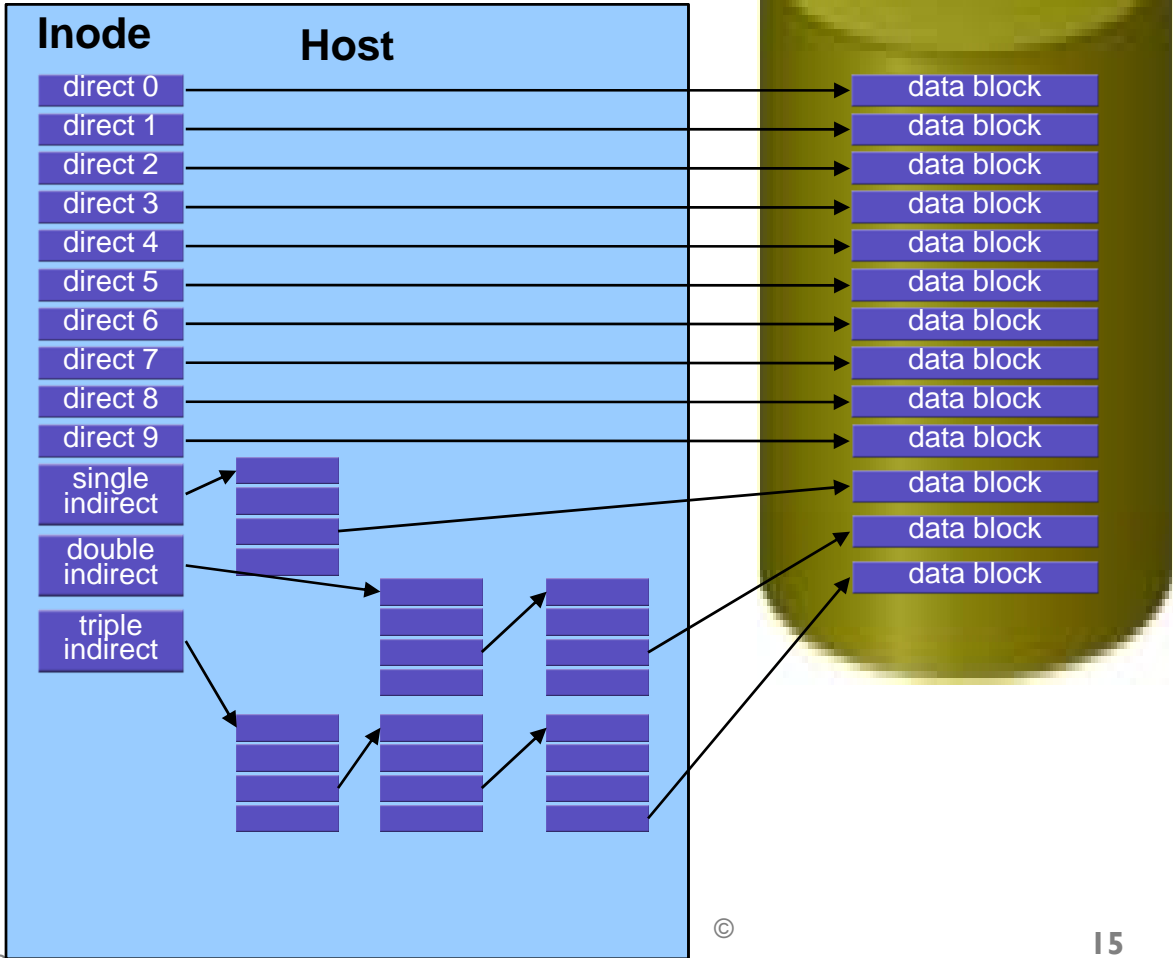
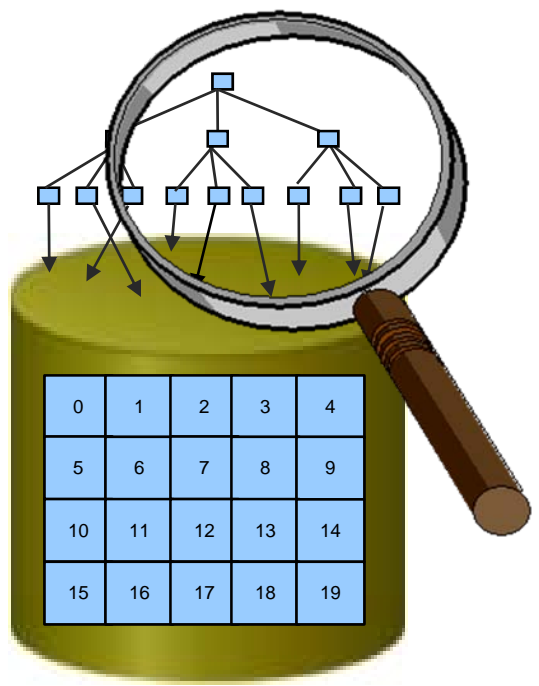


# Shared Device vs. Shared Data



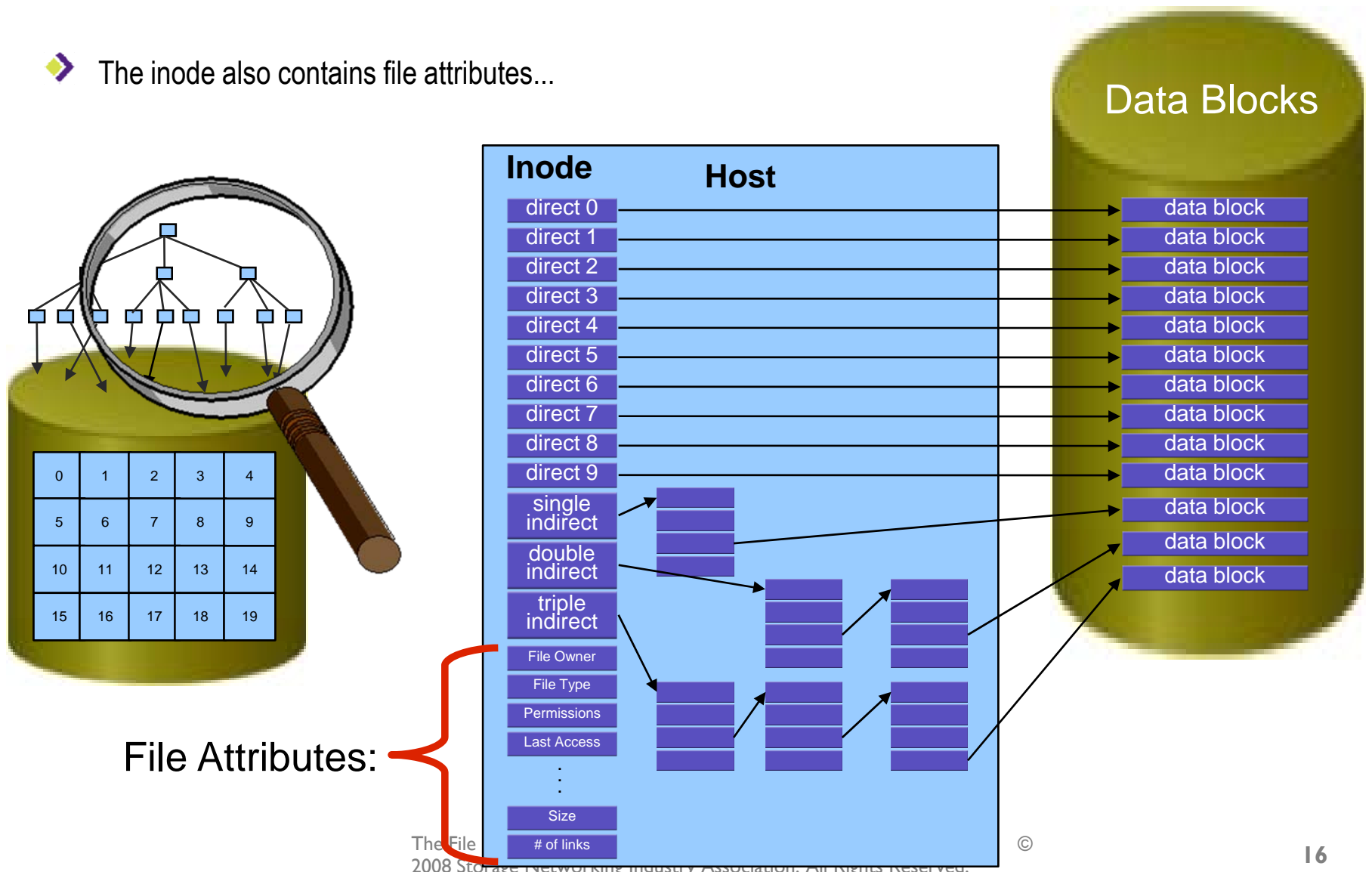
# Traditional File System - Inode

➤ When a file system is created, data structures that contain information about files are created. Each file has an inode and is identified by an inode number (often referred to as an "i-number" or "inode") in the file system where it resides.

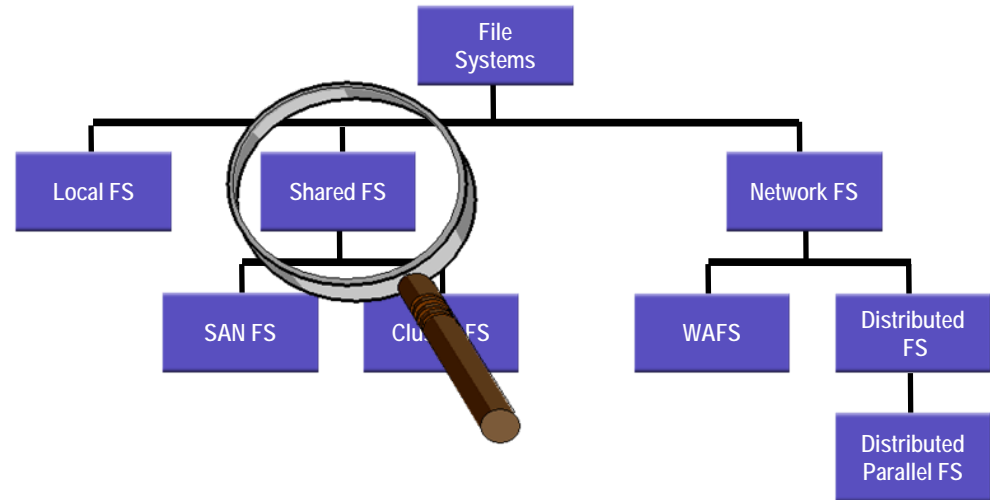


# Traditional File System - Inode

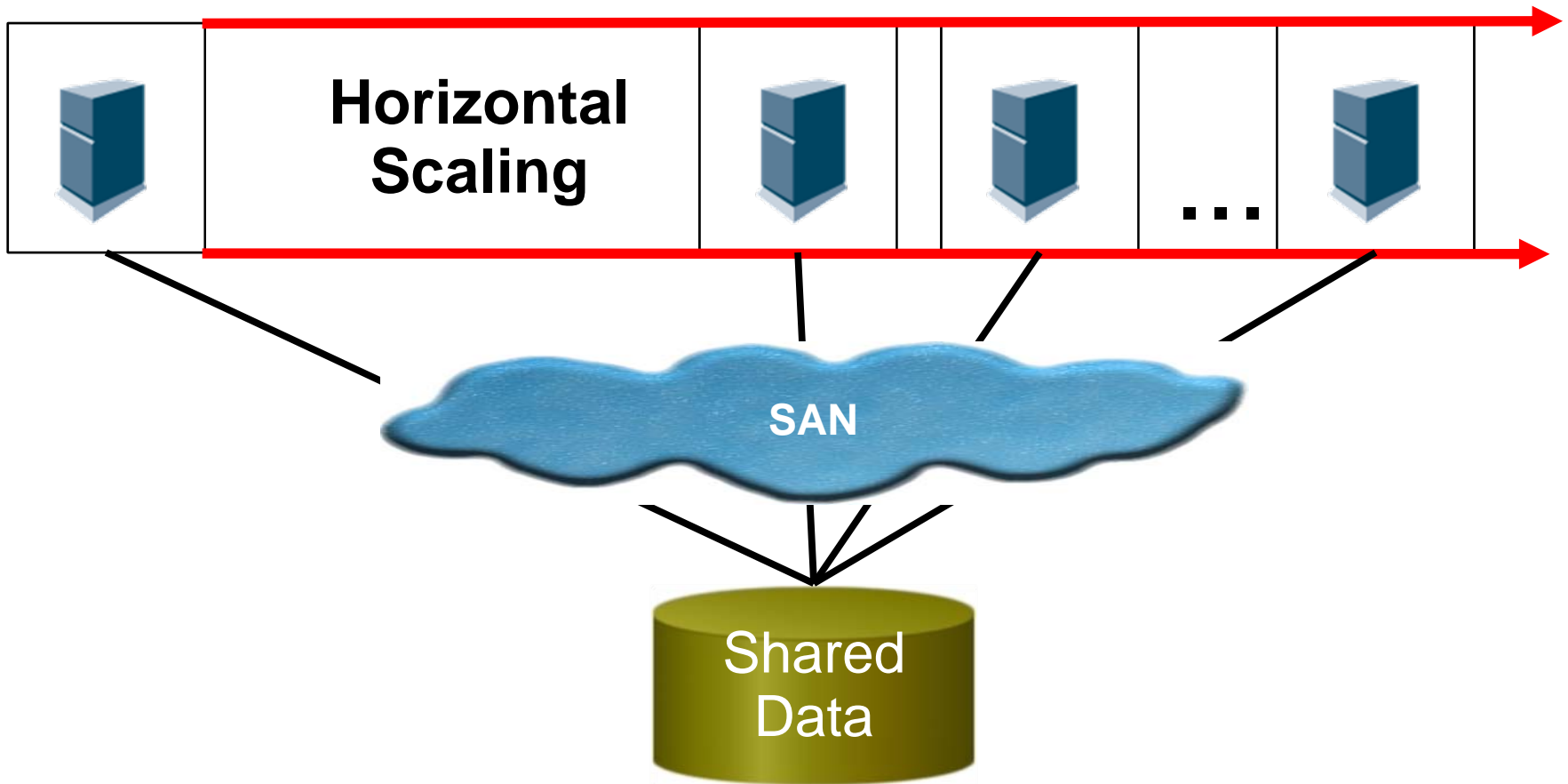
➤ The inode also contains file attributes...



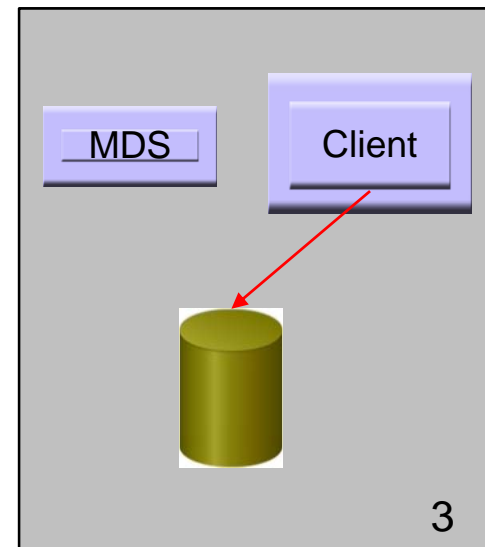
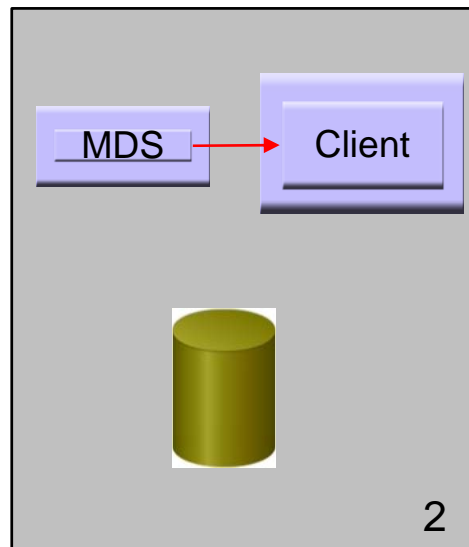
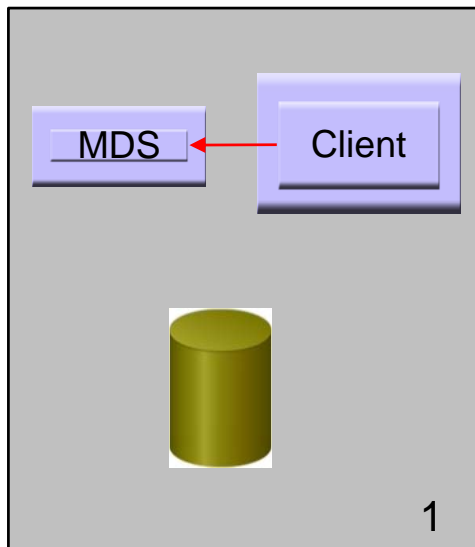
- File System Basics
- File Systems Taxonomy
- Local FS
- **Shared FS/Global FS**
  - ◆ **SAN FS, Cluster FS**
- Network FS
- Scalable NAS / Scalable NFS
- Wide Area FS
- Distributed FS
- File Virtualization
- Distributed and Parallel FS
- NAS Cluster / NAS Grid
- FS Future Developments

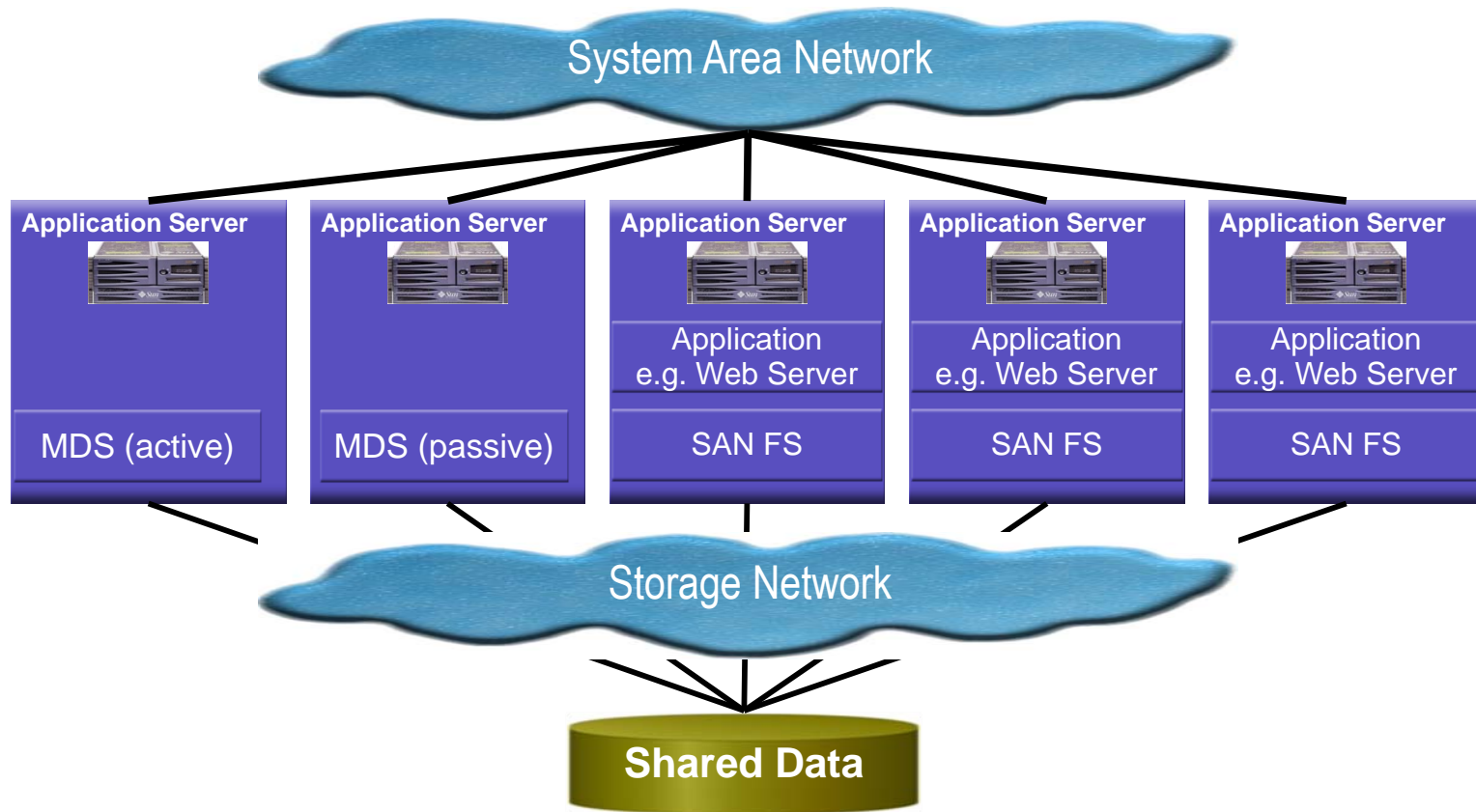


# Shared FS – Scale-Out



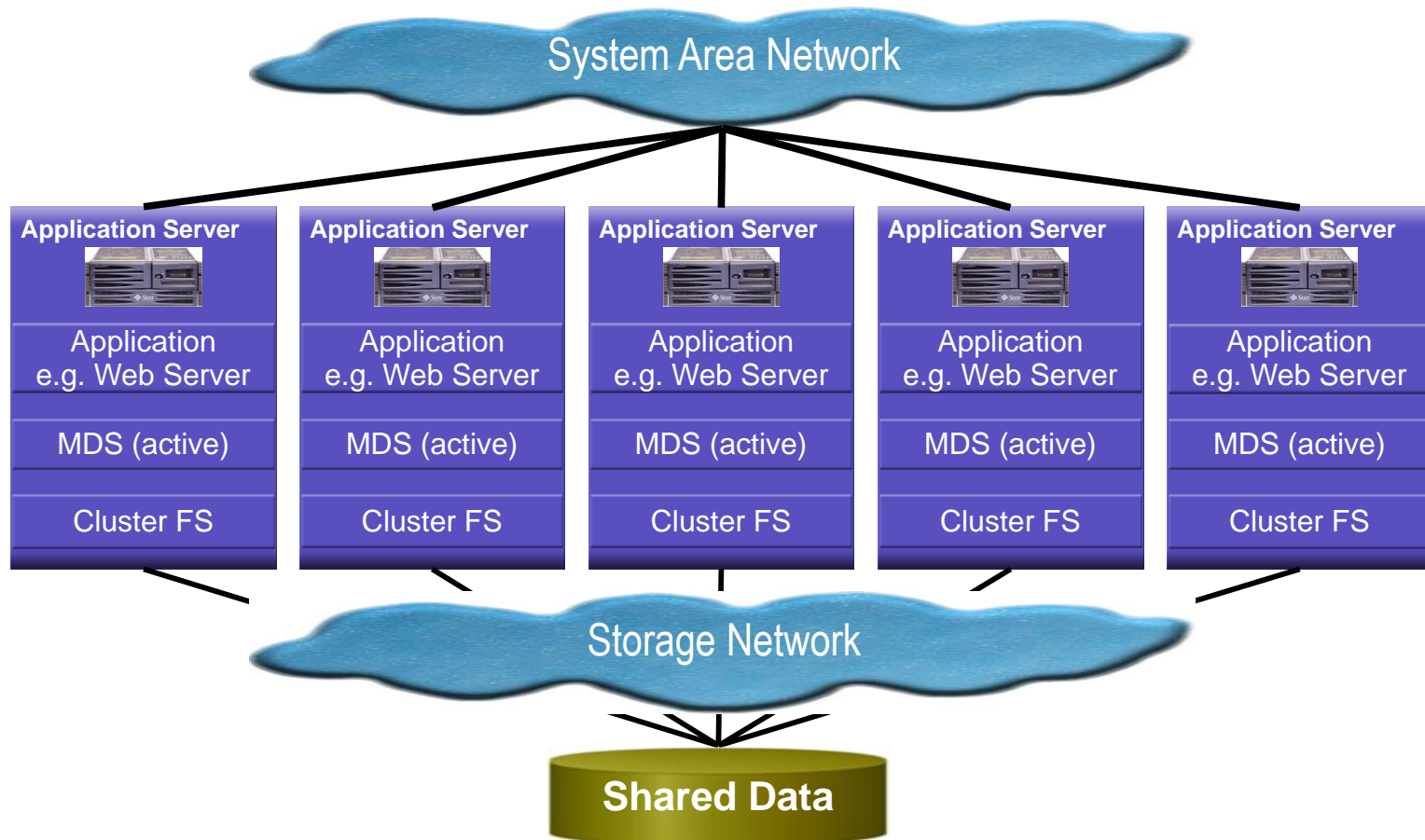
- Separate Metadata Server (MDS)
- Separation between logical and physical placement
- File access is a three-step transaction...





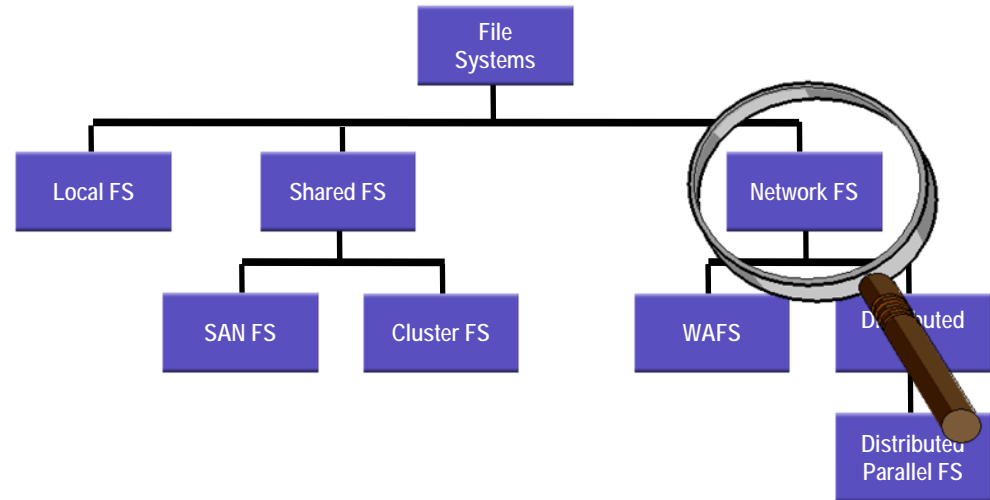
- MDS is not part of each node (i.e. master/slave - asymmetric)
- Heterogeneous with unlimited number of nodes
- Unlimited distance between nodes

# Shared FS / Global FS – Cluster FS



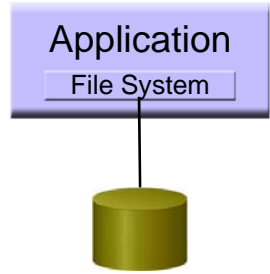
- MDS is part of each client (cluster) node; (i.e. peer-to-peer - symmetric)
- **Homogeneous with limited number of nodes**
- **Limited distance** between (cluster) nodes

- File System Basics
- File Systems Taxonomy
- Local FS
- Shared FS/Global FS
  - ◆ SAN FS, Cluster FS
- **Network FS**
- Scalable NAS / Scalable NFS
- Wide Area FS
- Distributed FS
- File Virtualization
- Distributed Parallel FS
- NAS Cluster / NAS Grid
- FS Future Developments

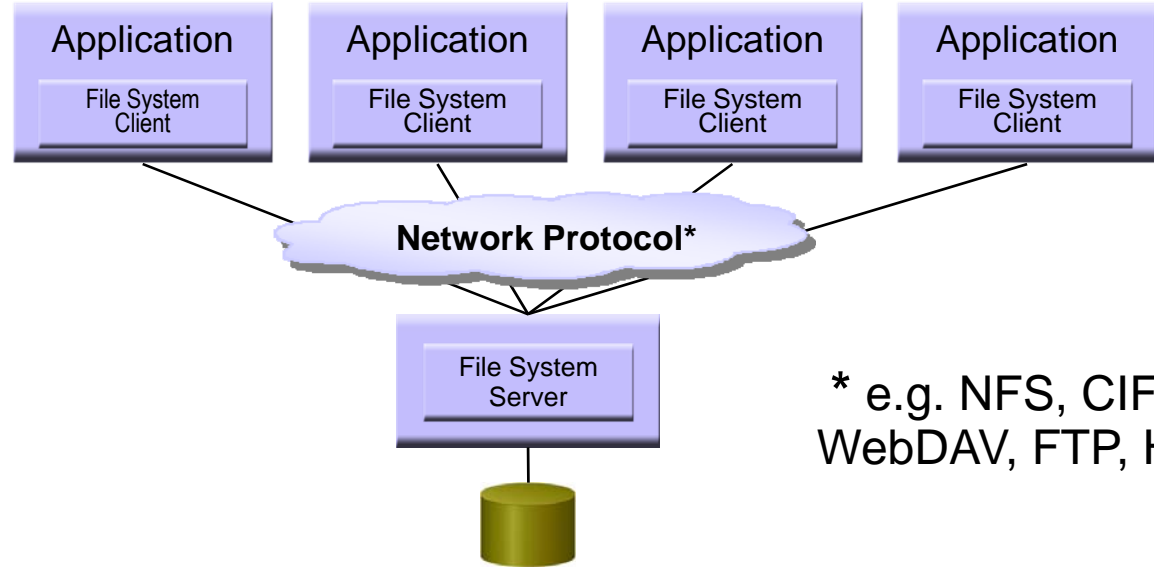


# Network Files System- aka Proxy FS

## Local FS



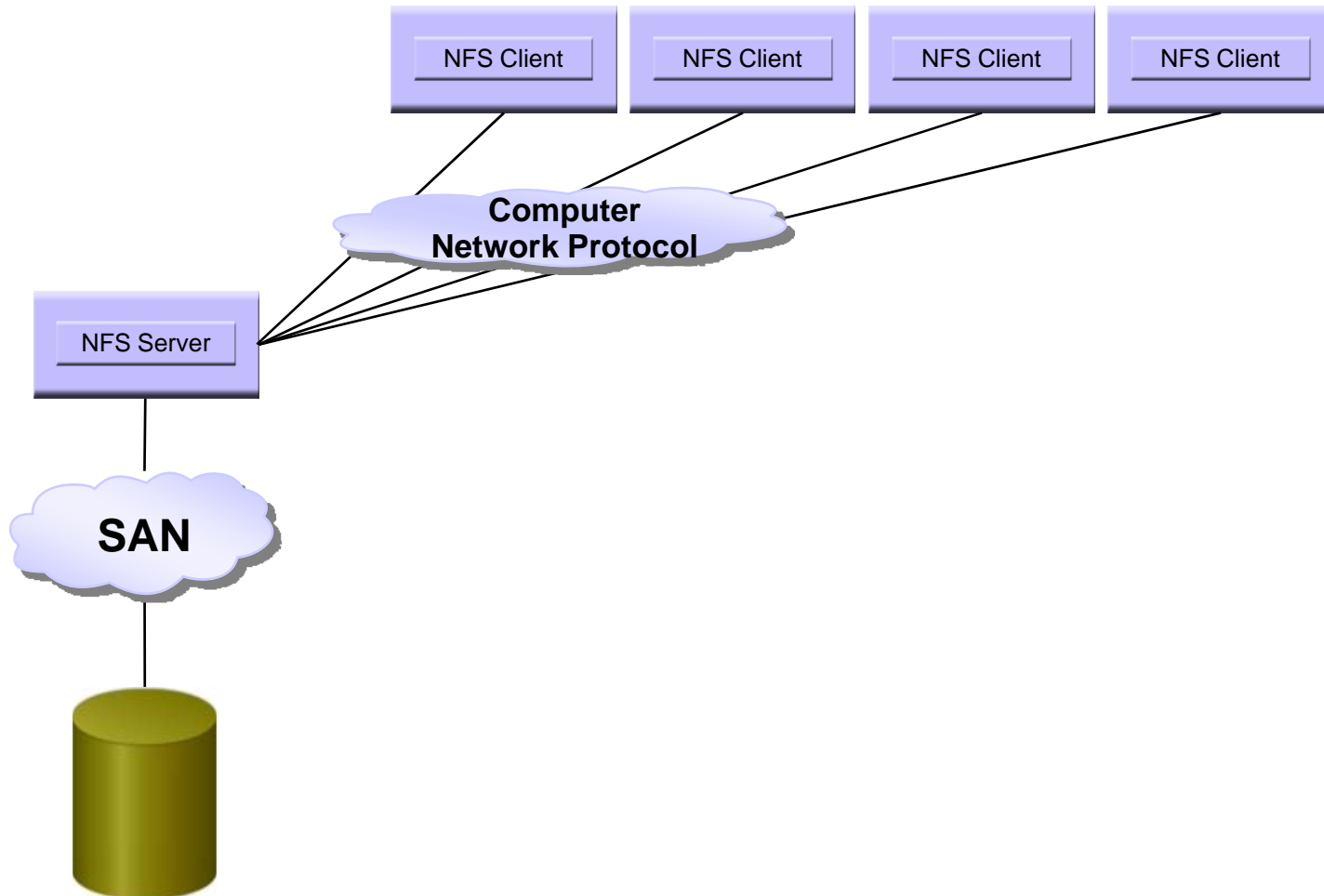
## Network FS



\* e.g. NFS, CIFS, AFP, WebDAV, FTP, HTTP, ...

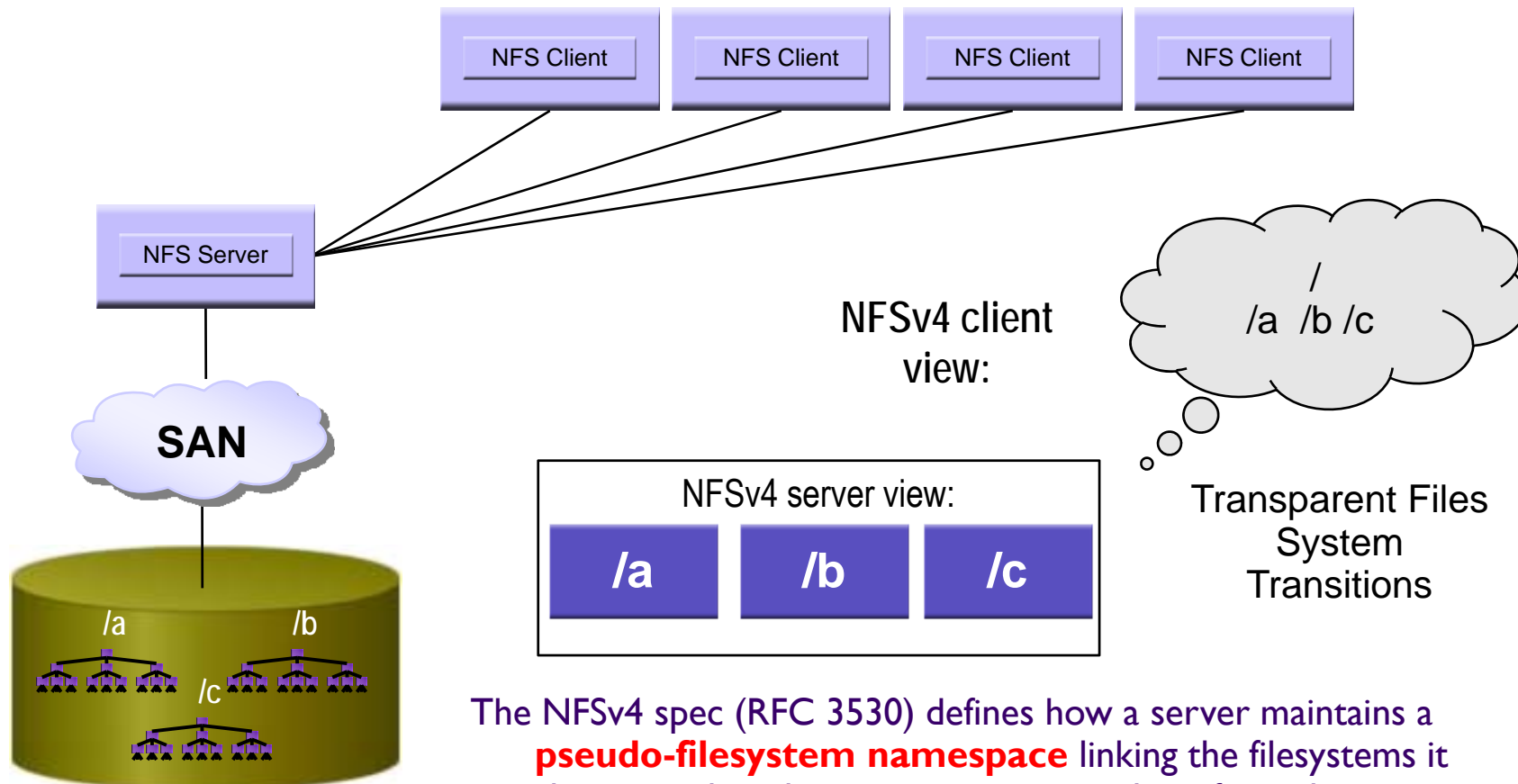
➤ A network file system is any file system that supports **sharing of files over a computer network protocol** between a client and a server

# Network File System Protocol (NFS)



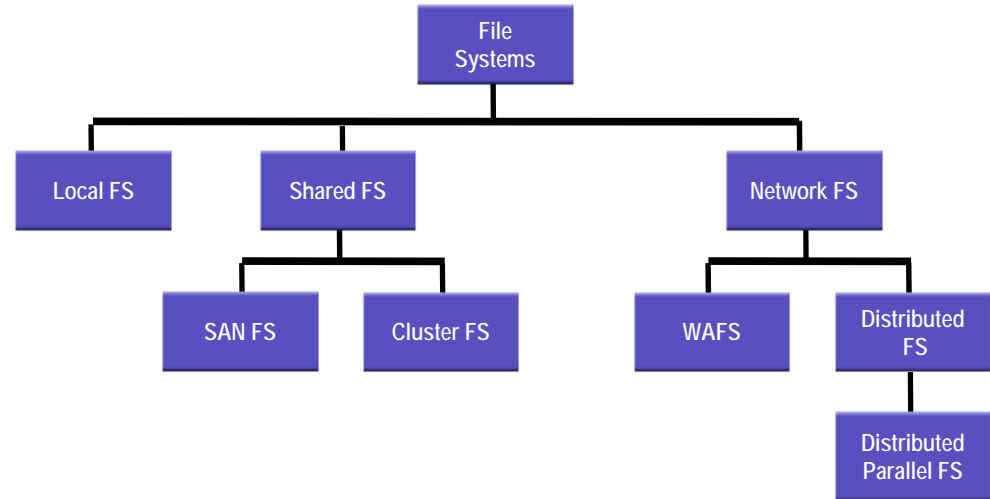
# NFSv4 Single-Server Namespace

## Server Pseudo FS – aka Shared Name Space



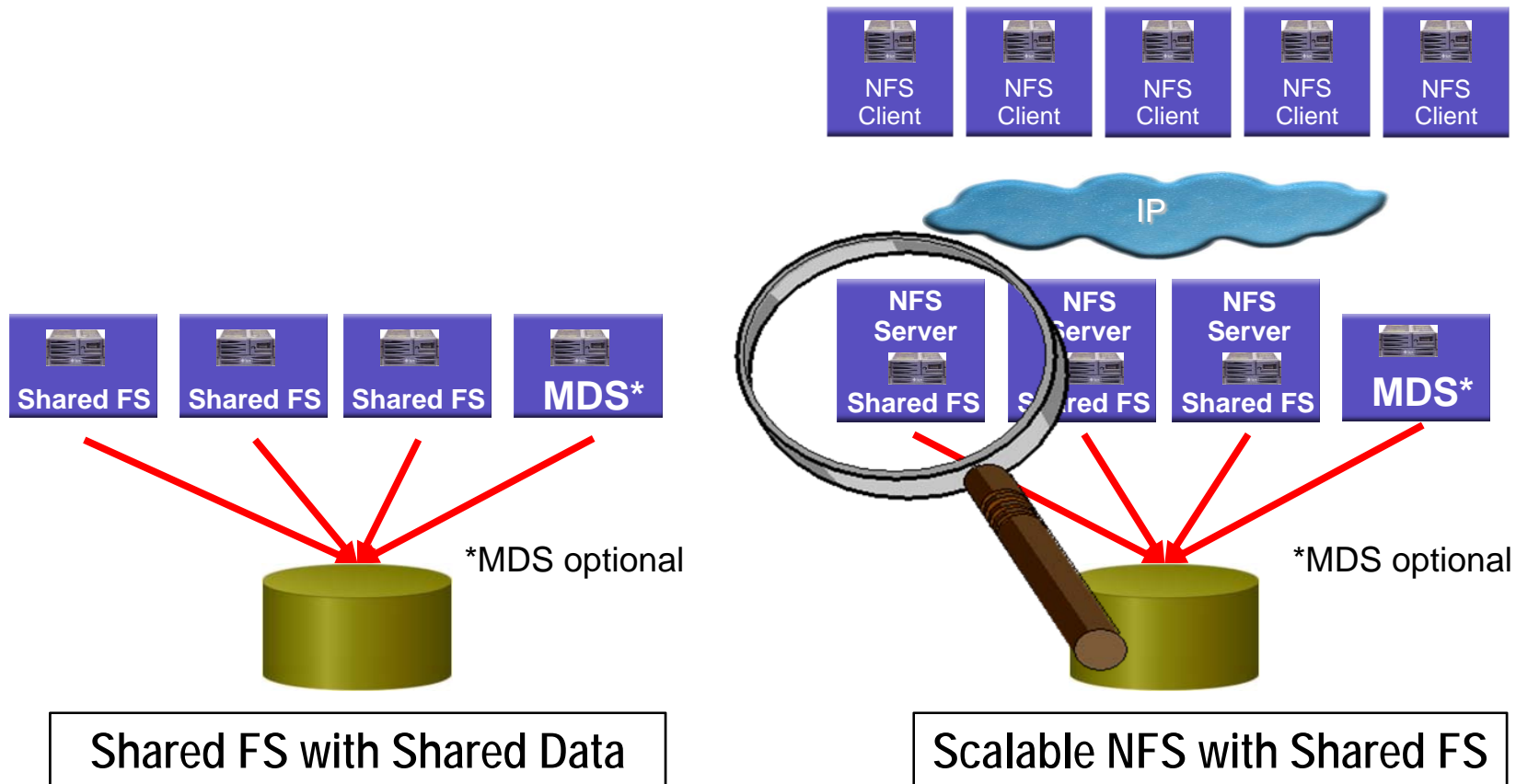
The NFSv4 spec (RFC 3530) defines how a server maintains a **pseudo-filesystem namespace** linking the filesystems it shares, so that clients can navigate to them from the server root. Many clients rely on this "single-server namespace" to be able to access all file systems on the server transparently.

- File System Basics
- File Systems Taxonomy
- Local FS
- Shared FS/Global FS
  - ◆ SAN FS, Cluster FS
- Network FS
- **Scalable NAS / Scalable NFS**
- Wide Area FS
- Distributed FS
- File Virtualization
- Distributed Parallel FS
- NAS Cluster / NAS Grid
- FS Future Developments

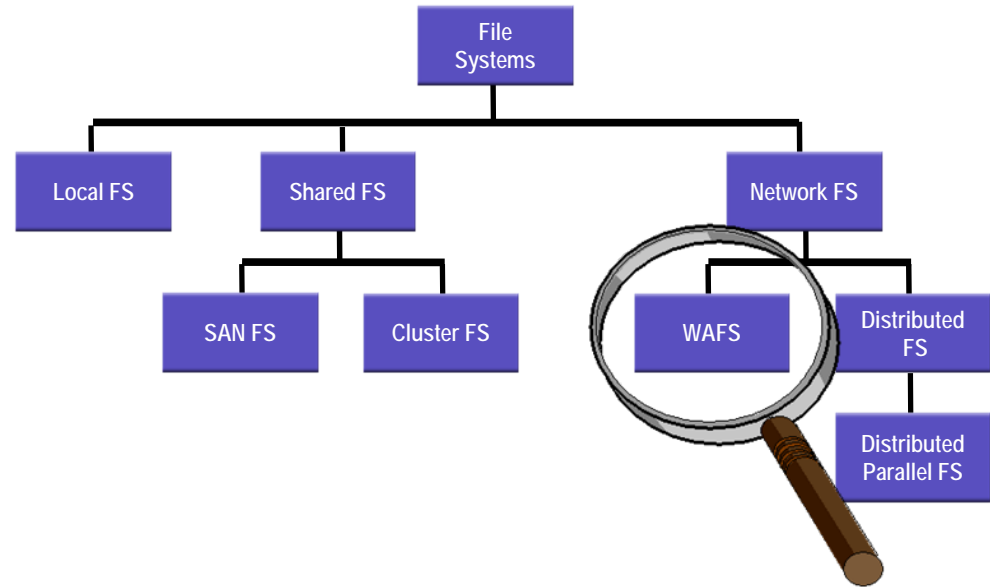


# Scalable NAS (NFS & Shared FS)

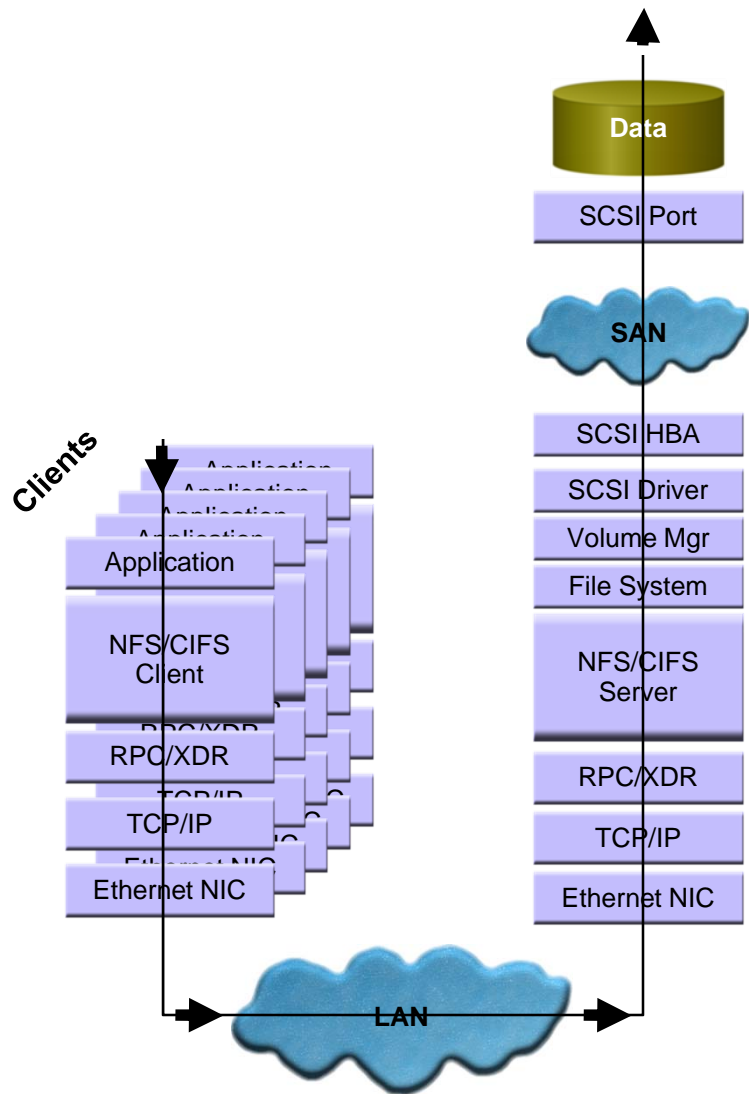
## ➤ Export NFS from Shared FS



- File System Basics
- File Systems Taxonomy
- Local FS
- Shared FS/Global FS
  - ◆ SAN FS, Cluster FS
- Network FS
- Scalable NAS / Scalable NFS
- **Wide Area FS**
- Distributed FS
- File Virtualization
- Distributed Parallel FS
- NAS Cluster / NAS Grid
- FS Future Developments

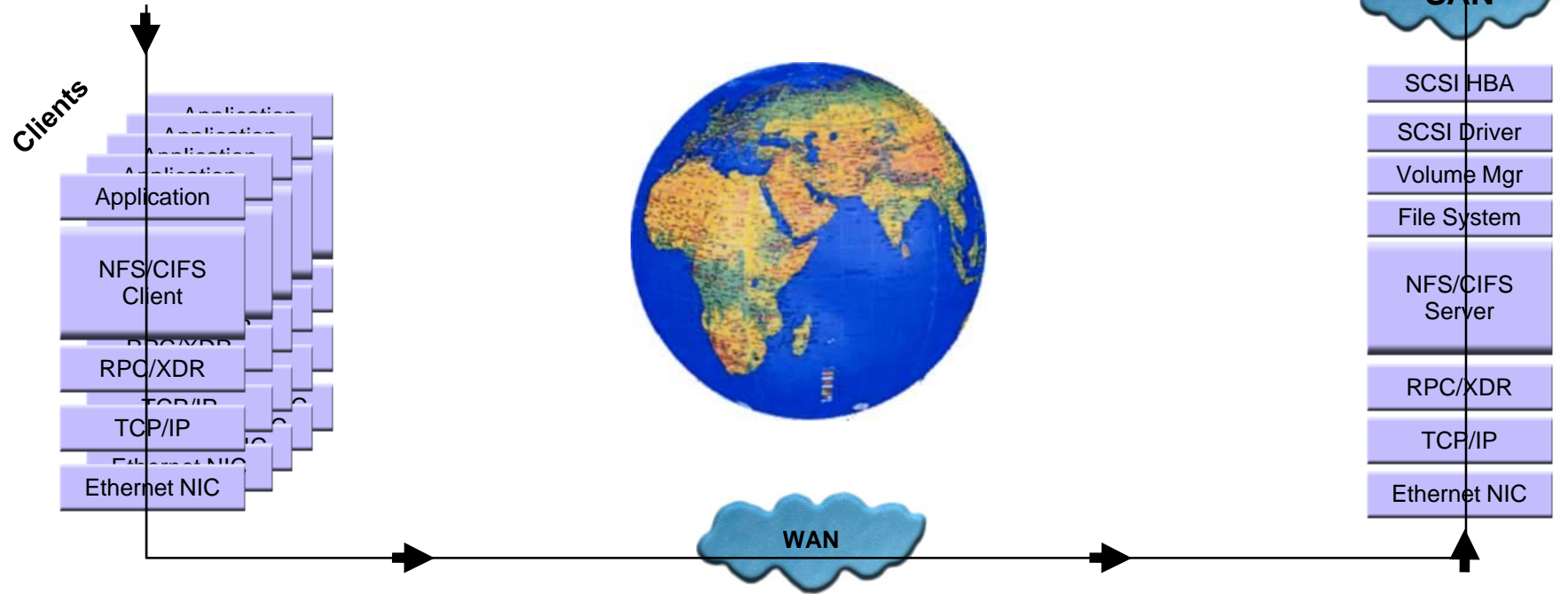


# Network FS Stack



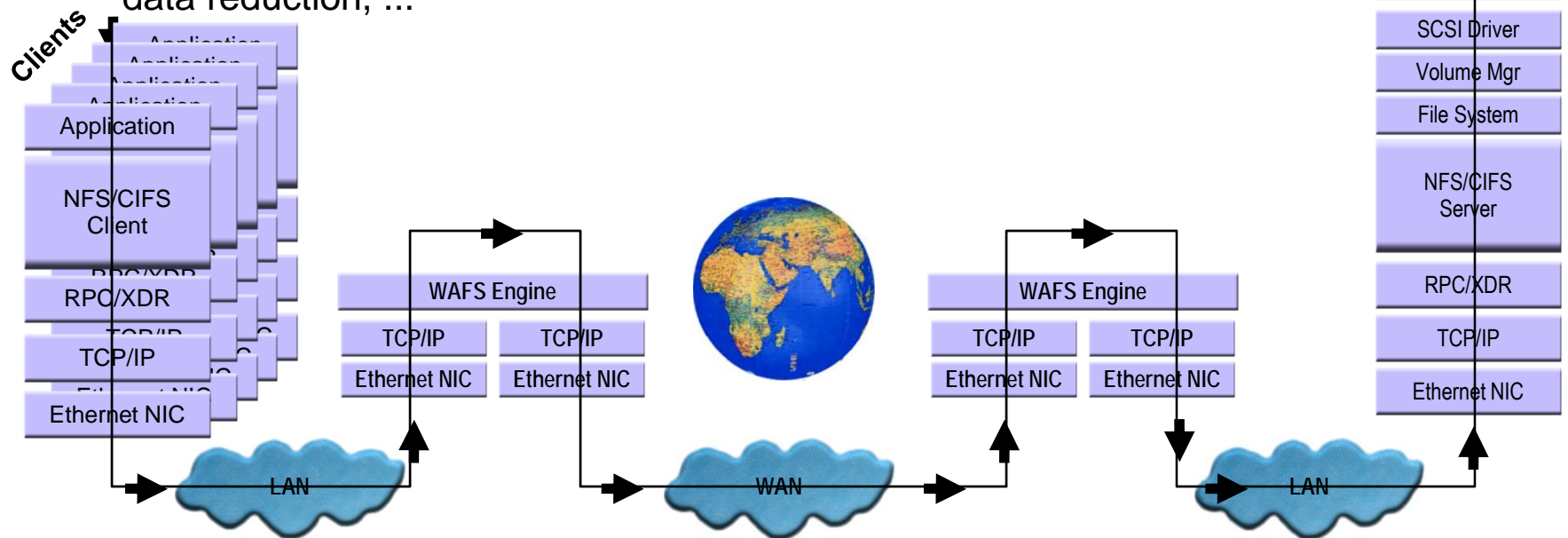
# Network FS in a Distributed World

- Consolidating file and storage resources into the data center eases management, administration, cost, and compliance
- Global file sharing and collaboration
- Remote office consolidation and optimization
- **Most application and file access protocols perform poorly over the WAN**

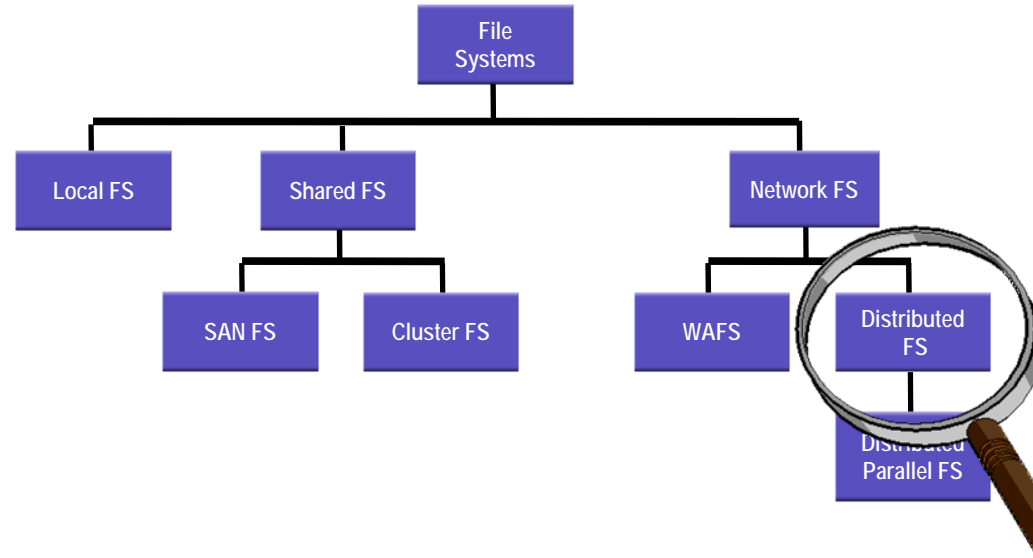


# Wide Area FS

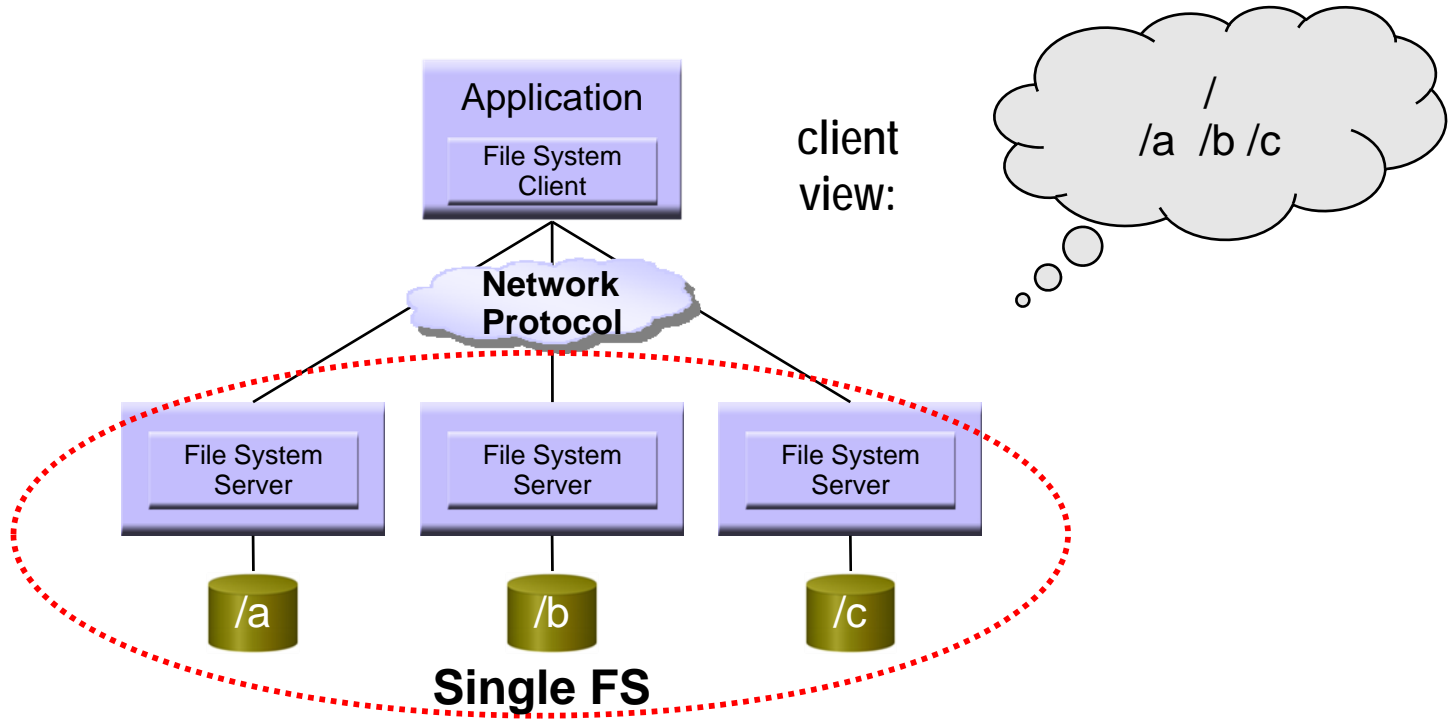
- Wide Area File Services – aka Wide Area File System (actually not a FS!)
- Protocol-specific optimization: HTTP, NFS, CIFS, WebDAV, FTP, TCP/IP, ...
- Application-specific optimization: email, document management, SQL, ...
- Intelligent caching: read-ahead, deferred write, coherency, ...
- Data compression: file-aware differencing, data aggregation, I/O clustering, dictionary-based compression (de-duplication), cross-protocol data reduction, ...



- File System Basics
- File Systems Taxonomy
- Local FS
- Shared FS/Global FS
  - ◆ SAN FS, Cluster FS
- Network FS
- Scalable NAS / Scalable NFS
- Wide Area FS
- **Distributed FS**
- File Virtualization
- Distributed Parallel FS
- NAS Cluster / NAS Grid
- FS Future Developments

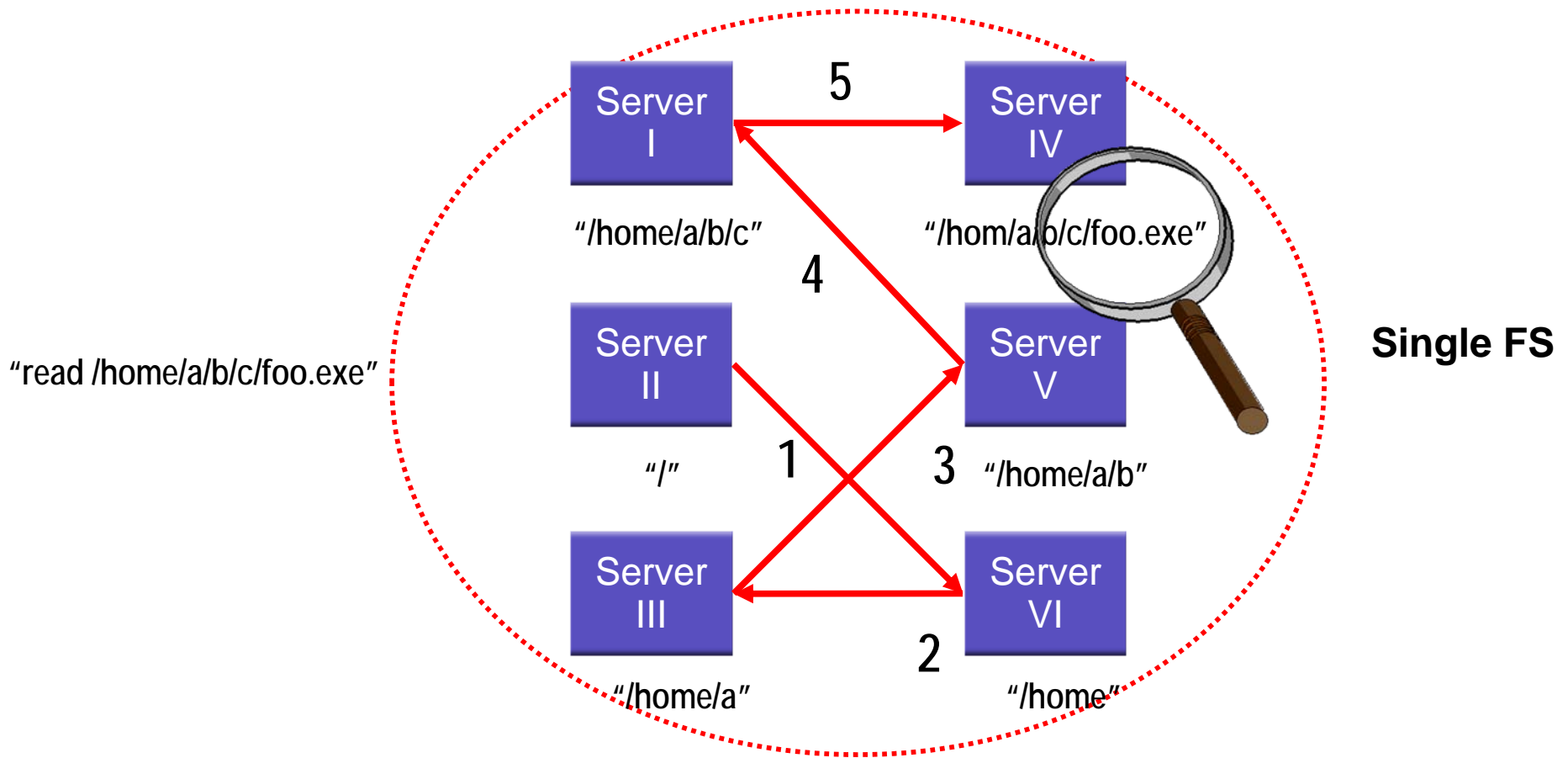


# Distributed File System (DFS)



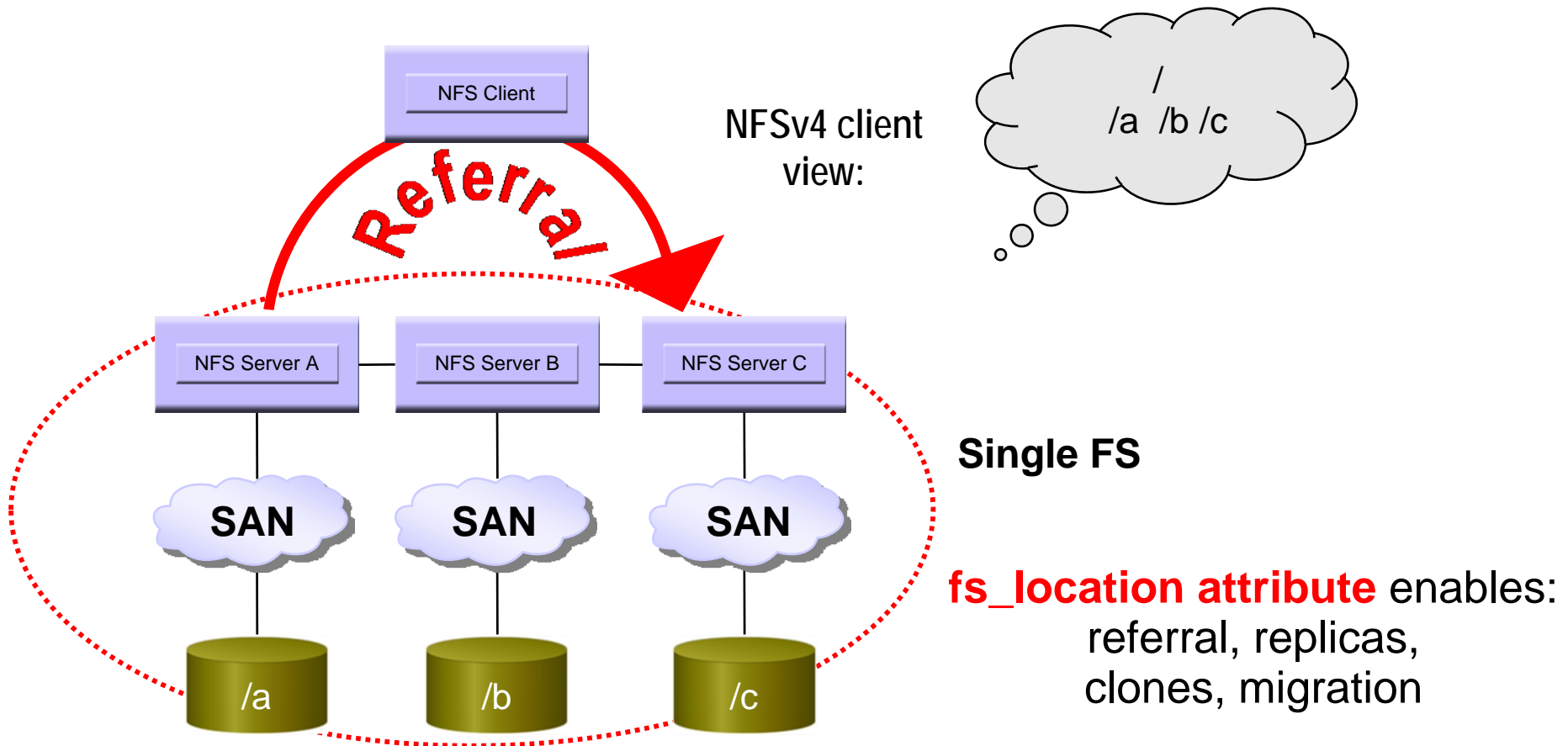
➤ **A distributed file system is a network file system** whose clients, servers, and storage devices are dispersed among the machines of a distributed system or intranet ( ≠ Parallel FS)

# DFS Logical Data Access Path



- Using Ethernet as a networking protocol between nodes, a DFS allows **a single file system to span across all nodes** in the DFS cluster, effectively creating a unified **Global Namespace** for all files.

# NFSv4.1 – Multi-Server Name Space

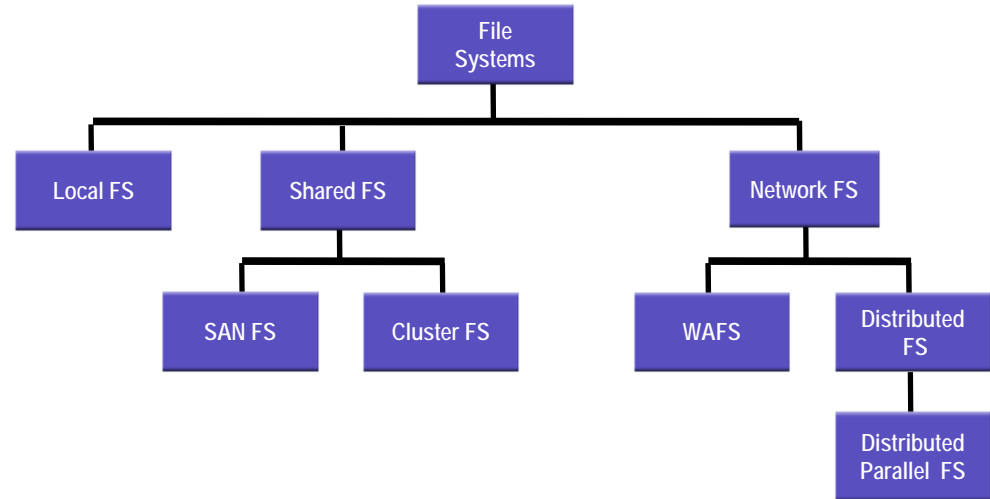


**Single FS**

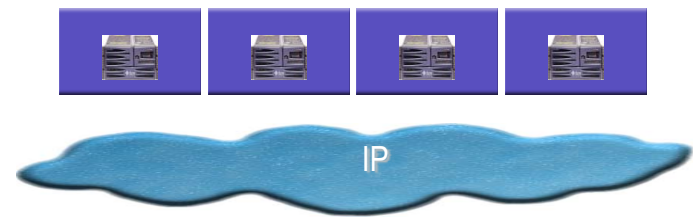
**fs\_location attribute** enables:  
referral, replicas,  
clones, migration

NFSv4.1 supports attributes that allow a namespace to extend beyond the boundaries of a single server through **location attributes**. A server can inform a client that data it seeks lives at another location; this is called "**referral**", and referrals can be used to construct a **Global Namespace**

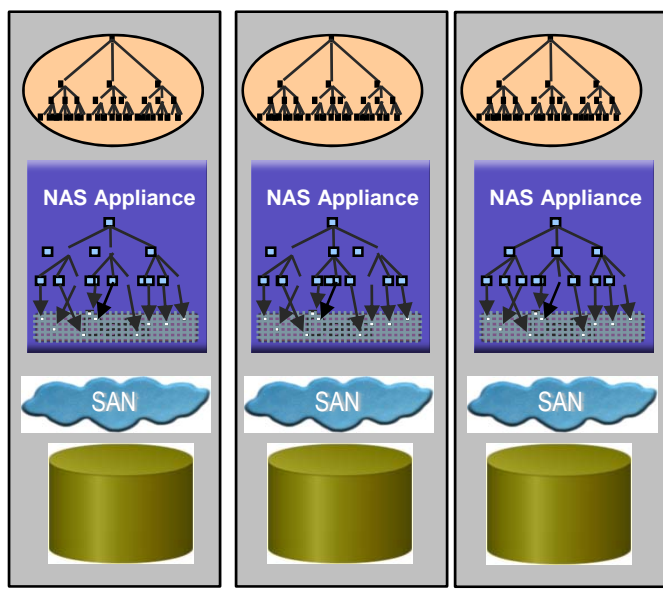
- File System Basics
- File Systems Taxonomy
- Local FS
- Shared FS/Global FS
  - ◆ SAN FS, Cluster FS
- Network FS
- Scalable NAS / Scalable NFS
- Wide Area FS
- Distributed FS
- **File Virtualization**
- Distributed Parallel FS
- NAS Cluster / NAS Grid
- FS Future Developments



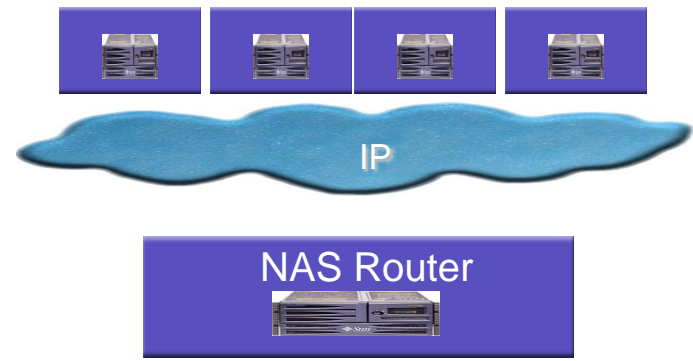
# Network Attached Storage (NAS)



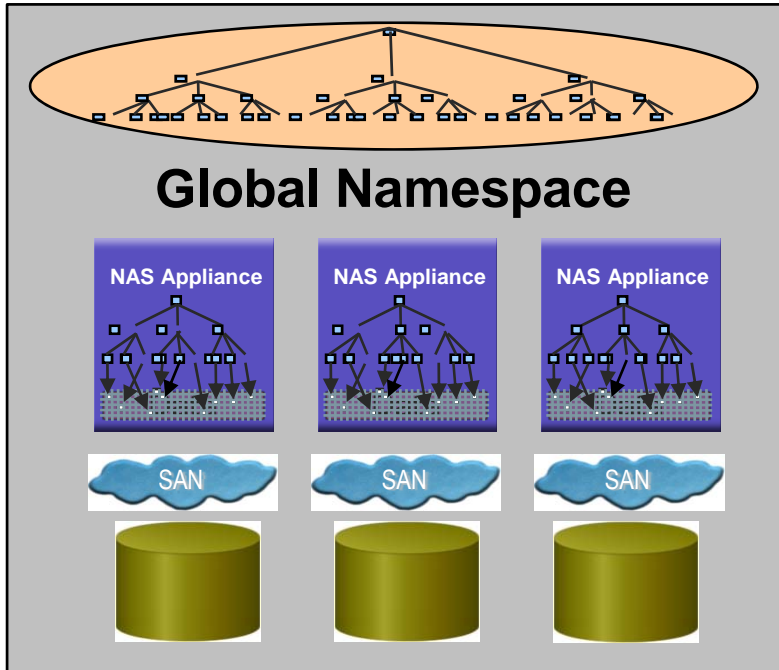
- Storage Islands
- Separated Namespaces
- Multiple mount points



# File Virtualization

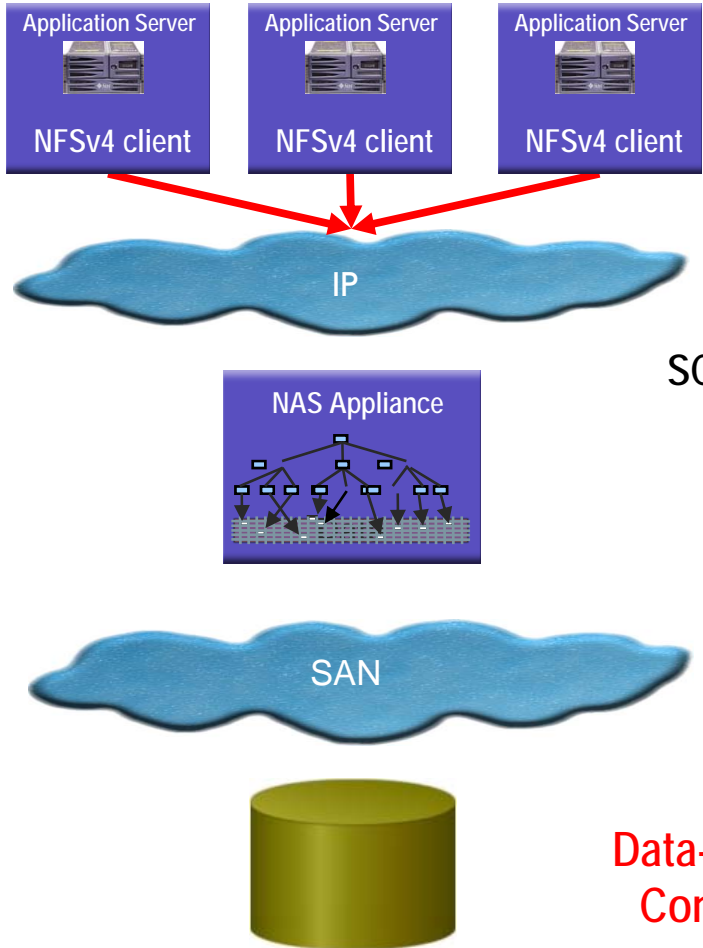


- In-Band Solution
- Aka NAS Aggregation
- NAS router

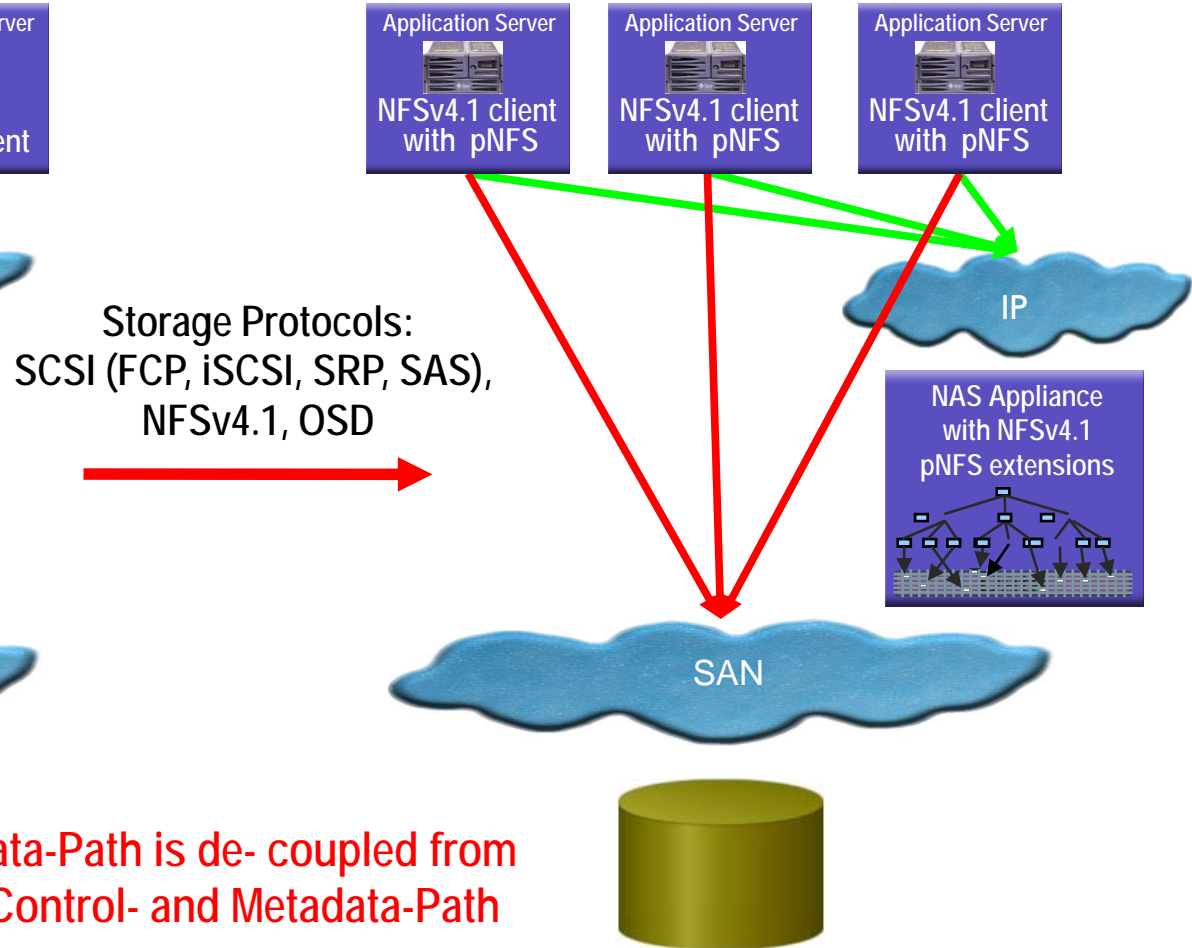


# FS Virtualization – NFS4.1 pNFS

## In-Band NAS:



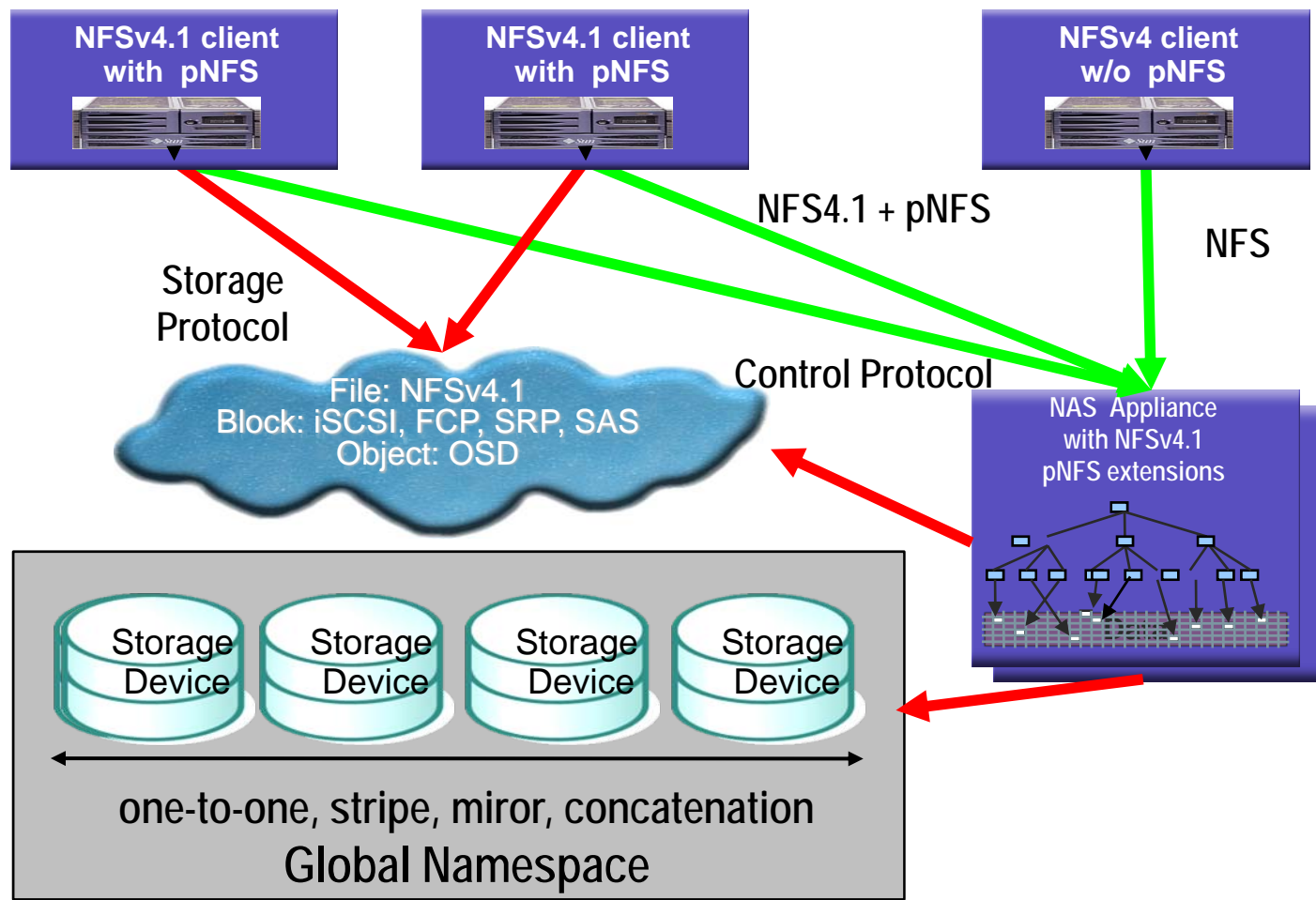
## Out-of-Band NAS:



Storage Protocols:  
SCSI (FCP, iSCSI, SRP, SAS),  
NFSv4.1, OSD

**Data-Path is de-coupled from  
Control- and Metadata-Path**

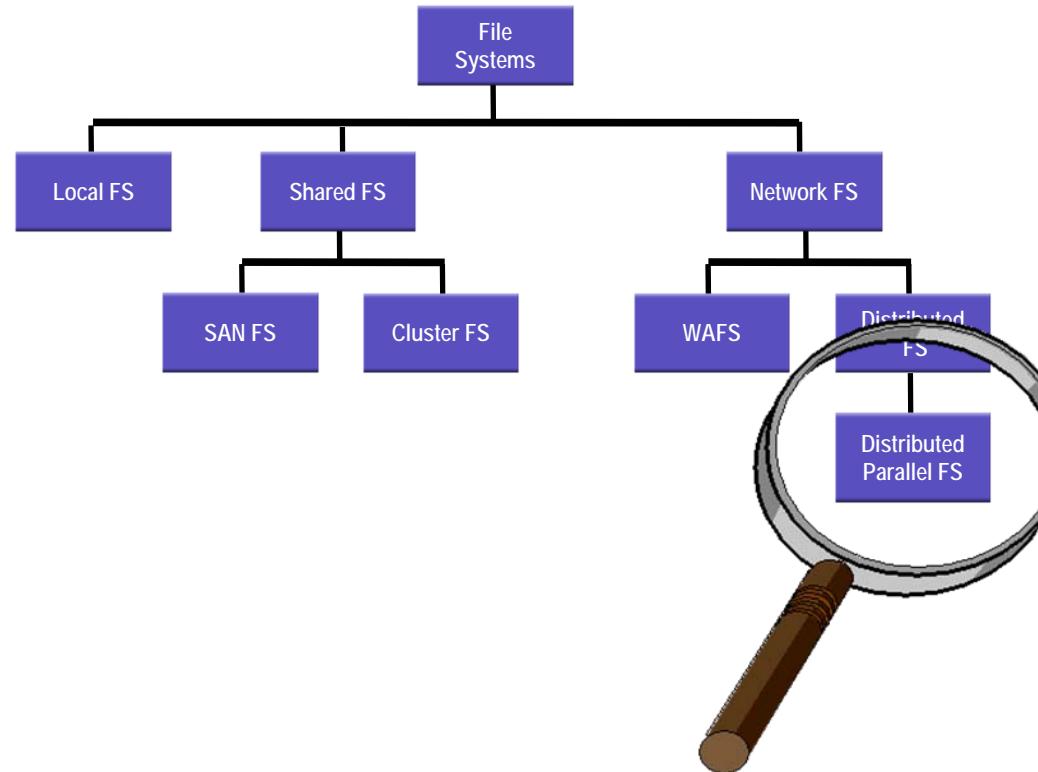
# FS Virtualization – NFS4.1 pNFS



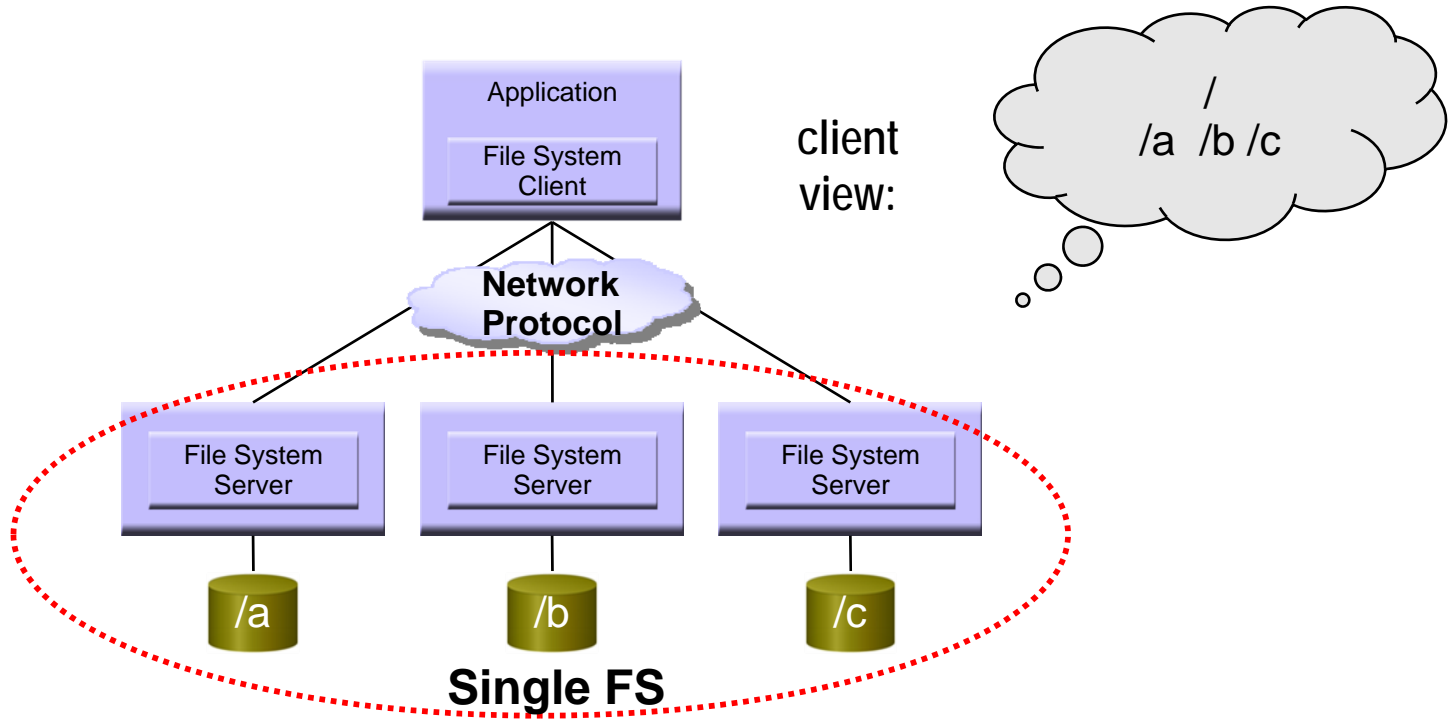
MDS acts as proxy for clients not pNFS enabled

Metadata Server (MDS) creates Global Namespace

- File System Basics
- File Systems Taxonomy
- Local FS
- Shared FS/Global FS
  - ◆ SAN FS, Cluster FS
- Network FS
- Scalable NAS / Scalable NFS
- Wide Area FS
- Distributed FS
- File Virtualization
- **Distributed Parallel FS**
- NAS Cluster / NAS Grid
- FS Future Developments



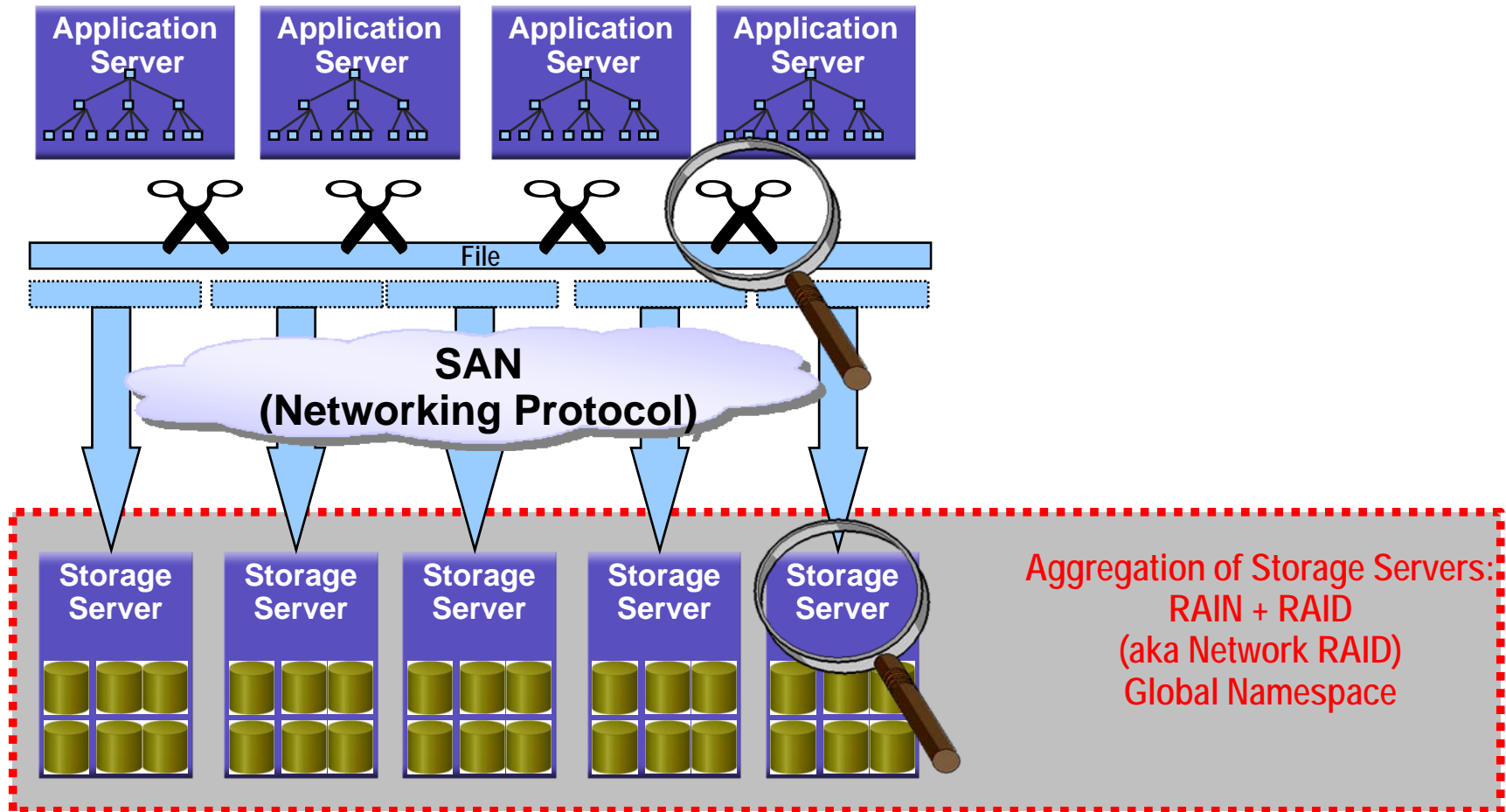
# Distributed File System (DFS)



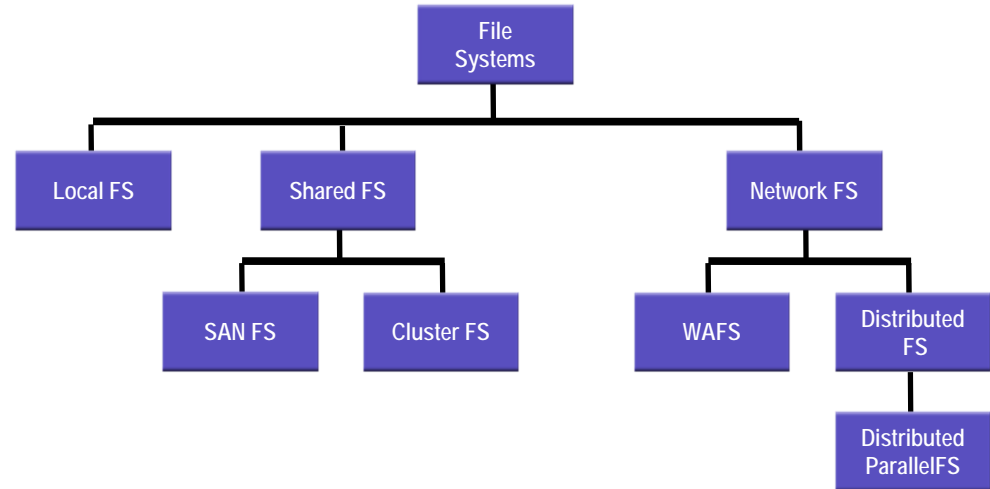
➤ **Files** are distributed across file servers

# Distributed & Parallel File System

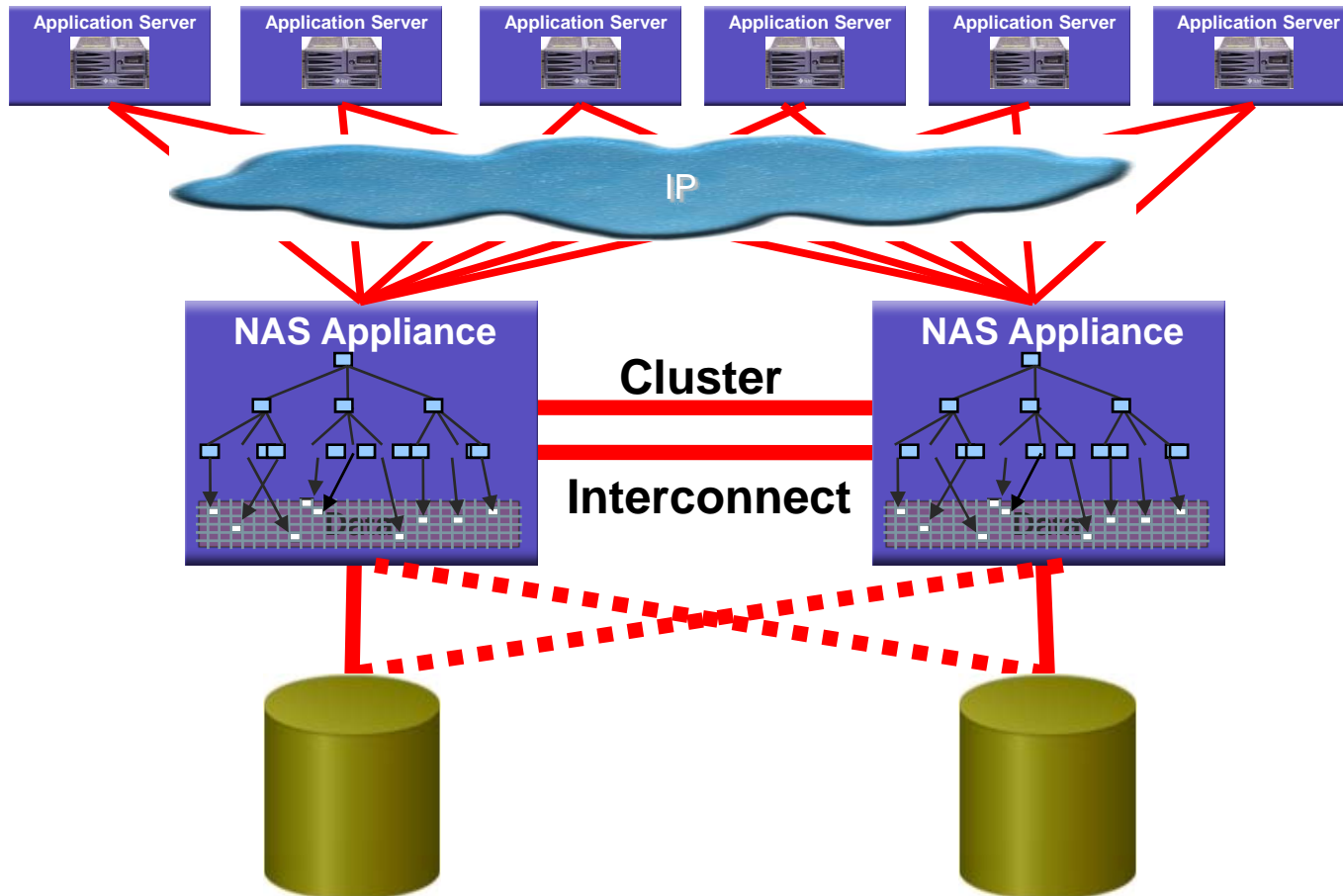
- File Segments distributed across storage nodes – Parallel I/Os



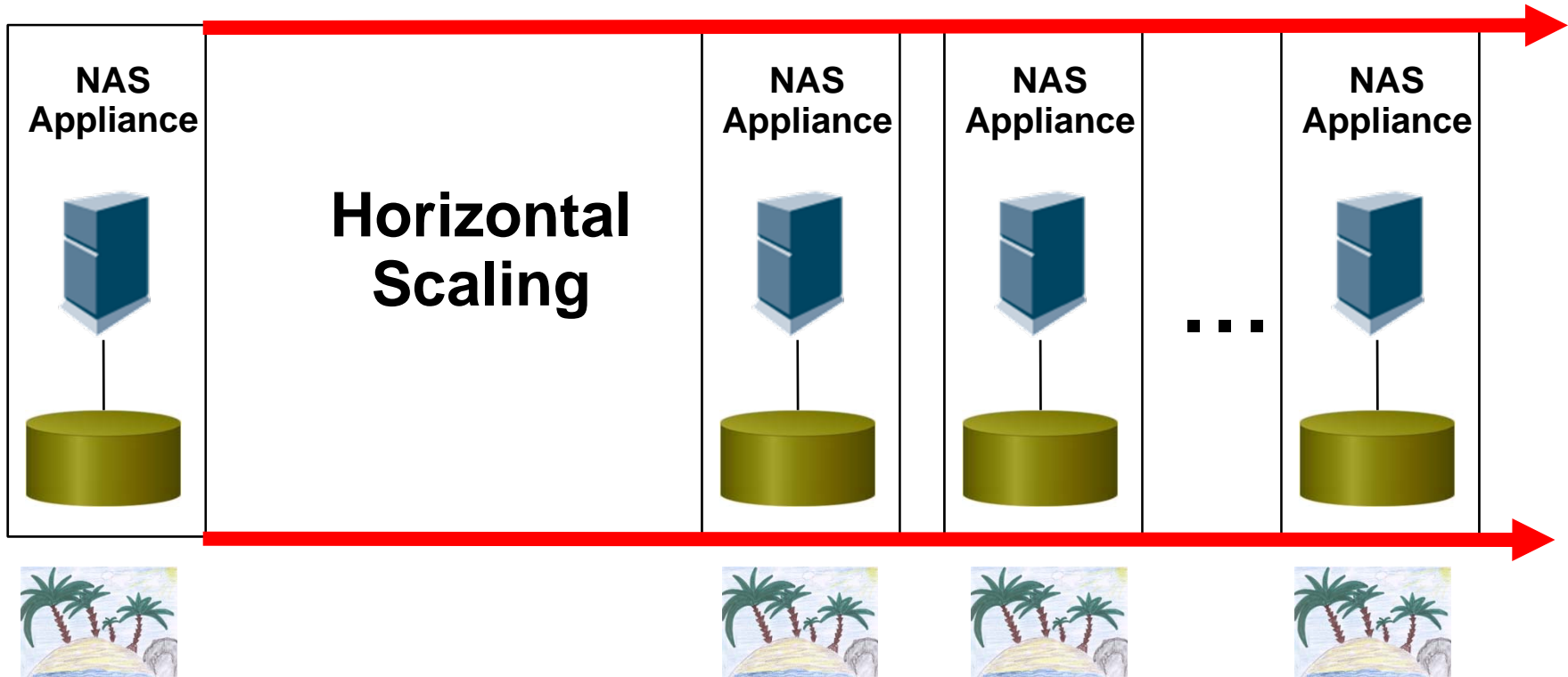
- File System Basics
- File Systems Taxonomy
- Local FS
- Shared FS/Global FS
  - ◆ SAN FS, Cluster FS
- Network FS
- Scalable NAS / Scalable NFS
- Wide Area FS
- Distributed FS
- File Virtualization
- Distributed Parallel FS
- **NAS Cluster / NAS Grid**
- FS Future Developments



# Two-Node NAS Cluster (Failover)

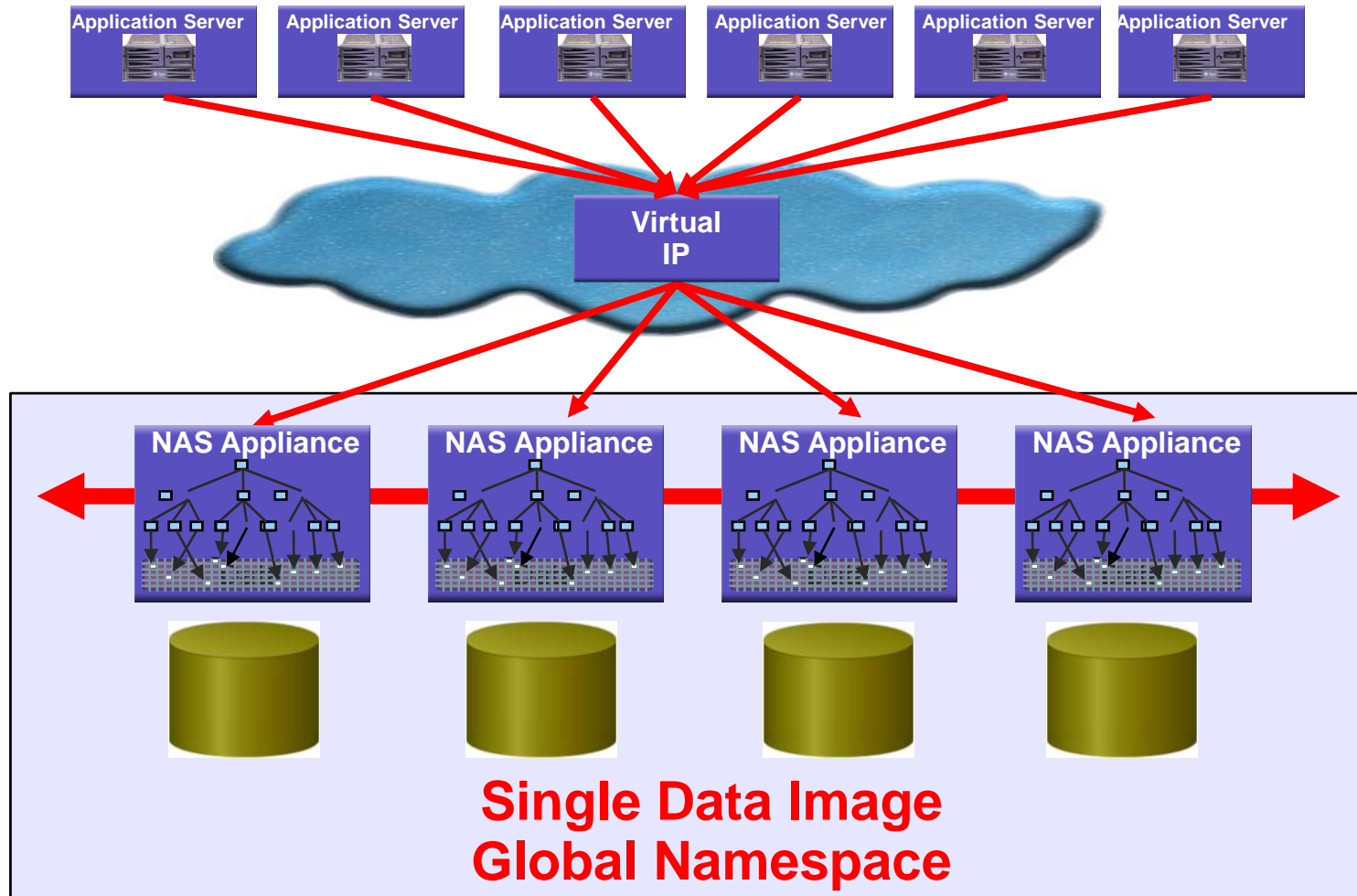


# NAS Scale-Out Problem Statement

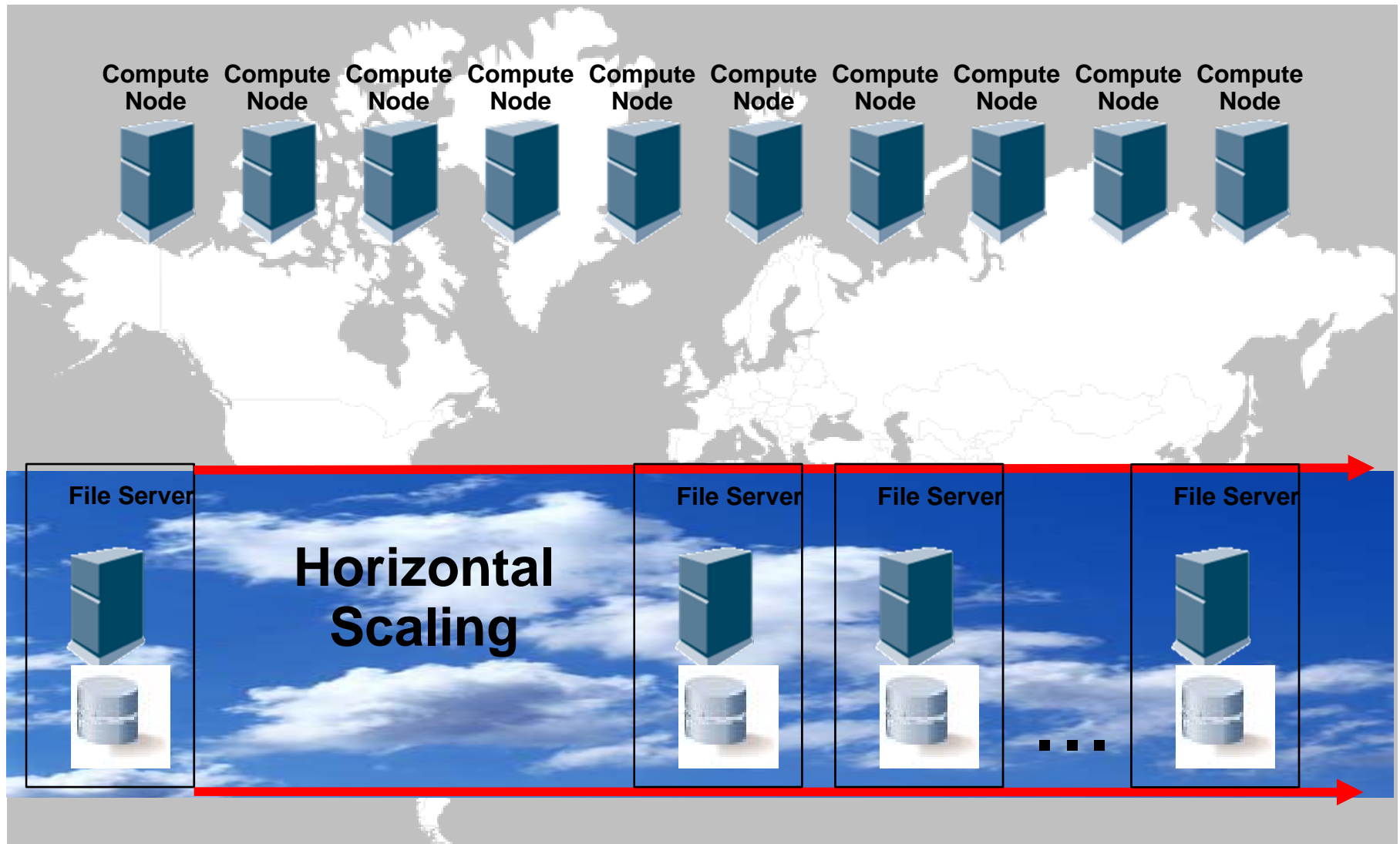


- Horizontal scaling without data replication or creating **islands** of data

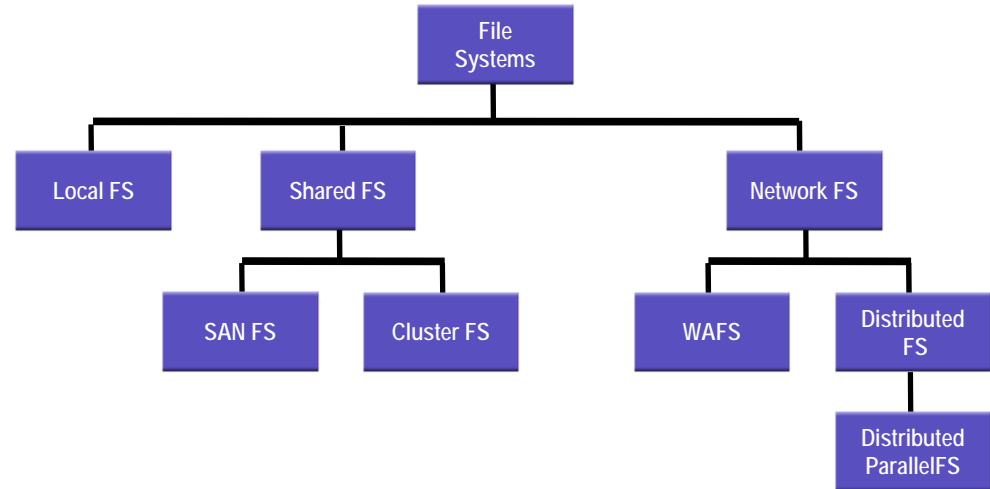
# NAS Cluster / NAS Grid



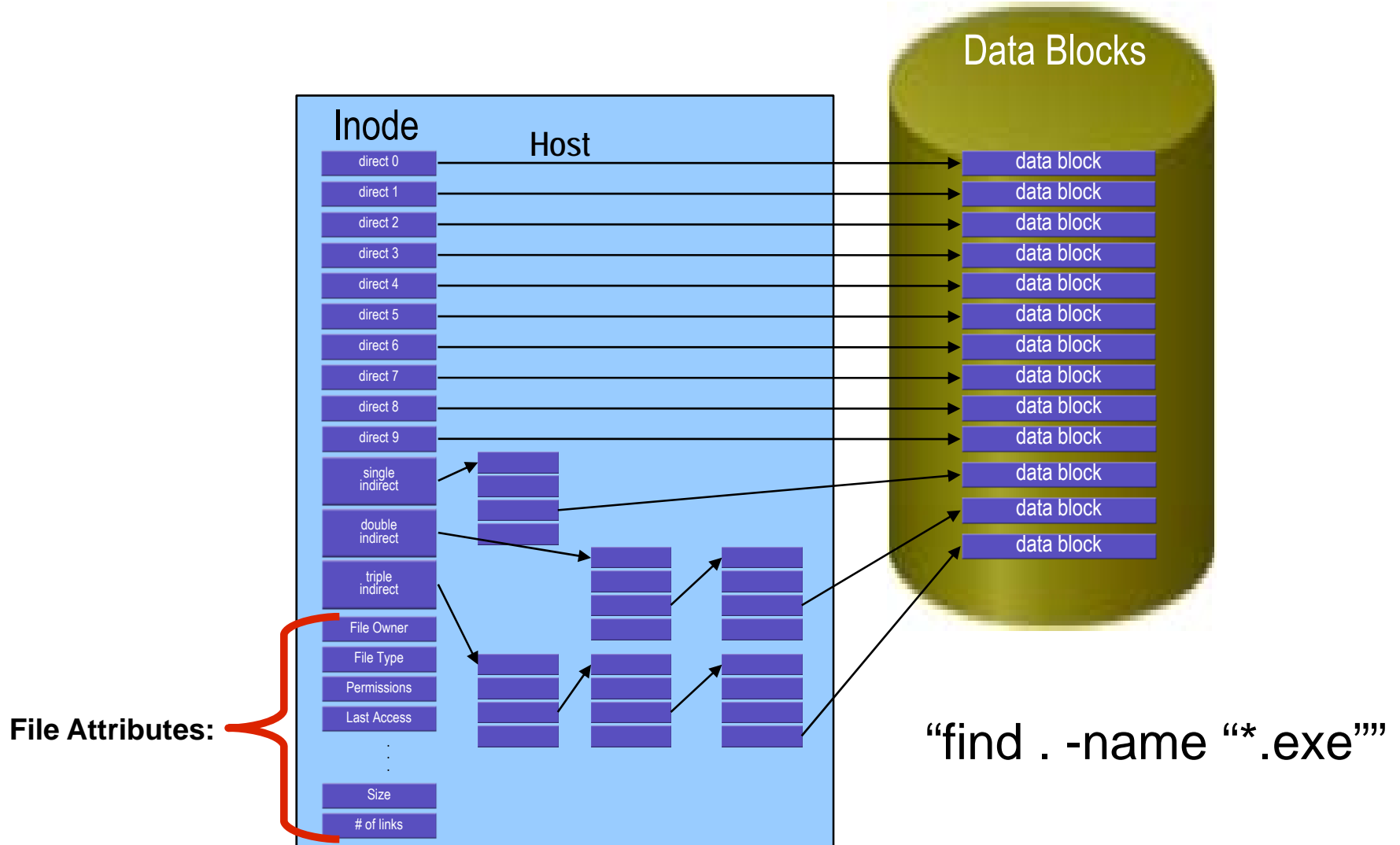
# Cloud Storage/Computing (SaaS)



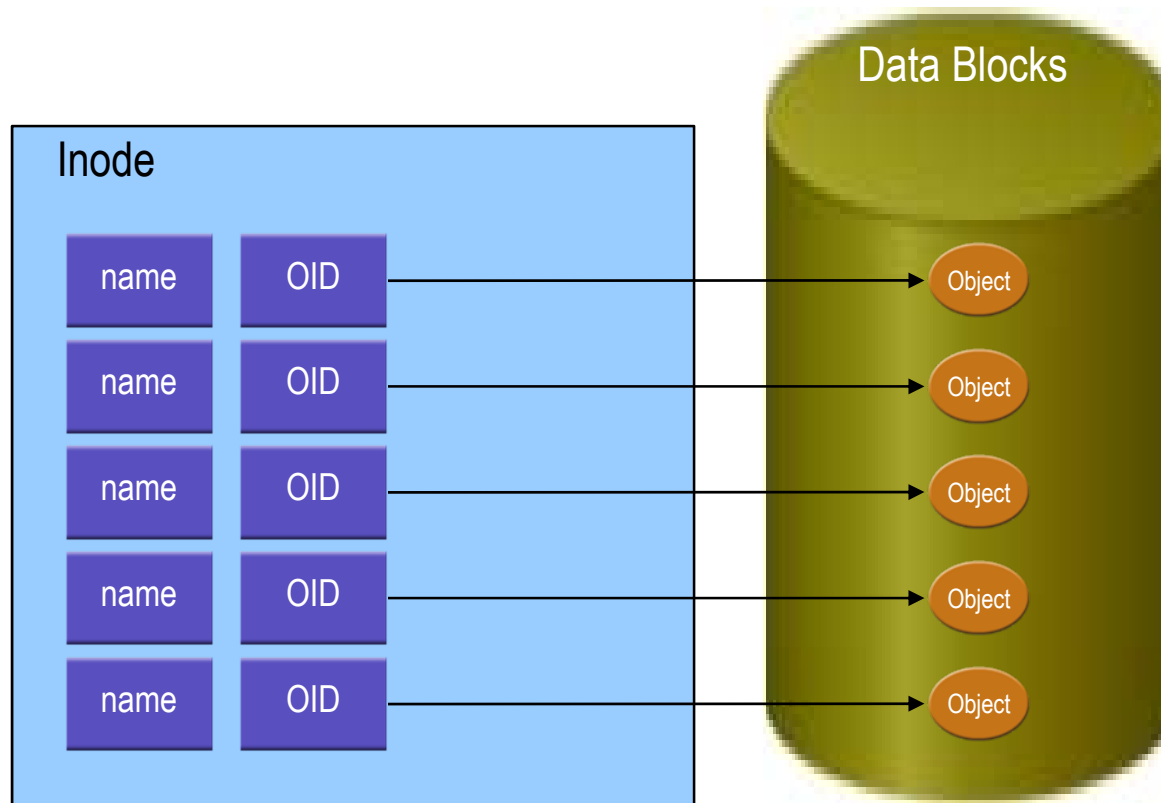
- File System Basics
- File Systems Taxonomy
- Local FS
- Shared FS/Global FS
  - ◆ SAN FS, Cluster FS
- Network FS
- Scalable NAS / Scalable NFS
- Wide Area FS
- Distributed FS
- File Virtualization
- Distributed Parallel FS
- NAS Cluster / NAS Grid
- **FS Future Developments**



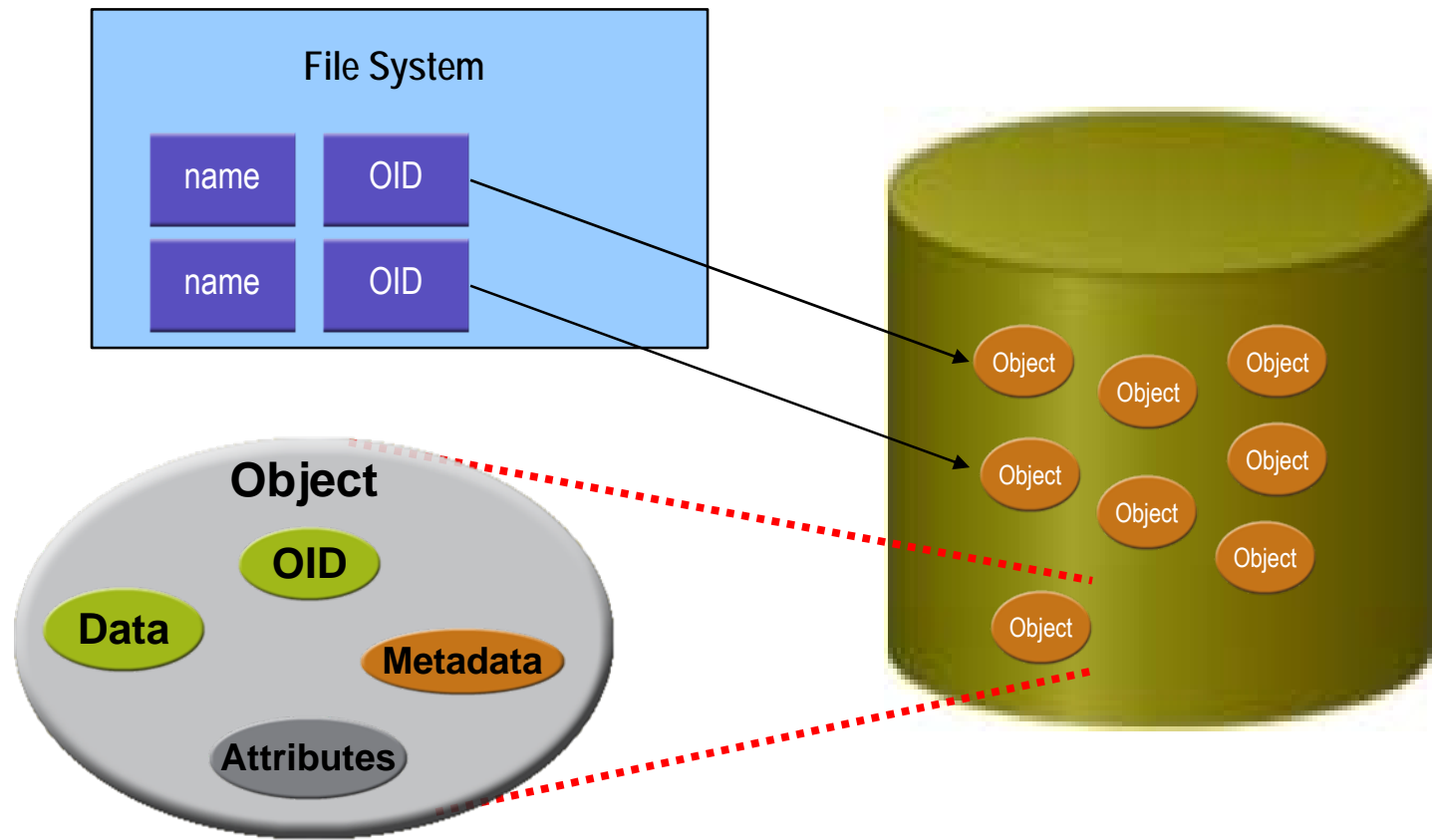
# File Systems & Metadata



# Files Are Morphing Into File Objects



# Files Are Morphing Into Objects...

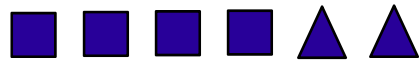


“select \* where customer\_ID < 17 and location = “Frankfurt, Germany””

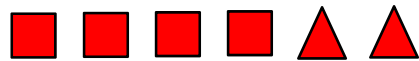
# Aggregation of Storage Servers (RAIN)

## Data Placement

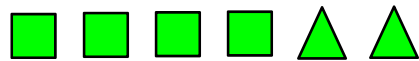
Object 1



Object 2

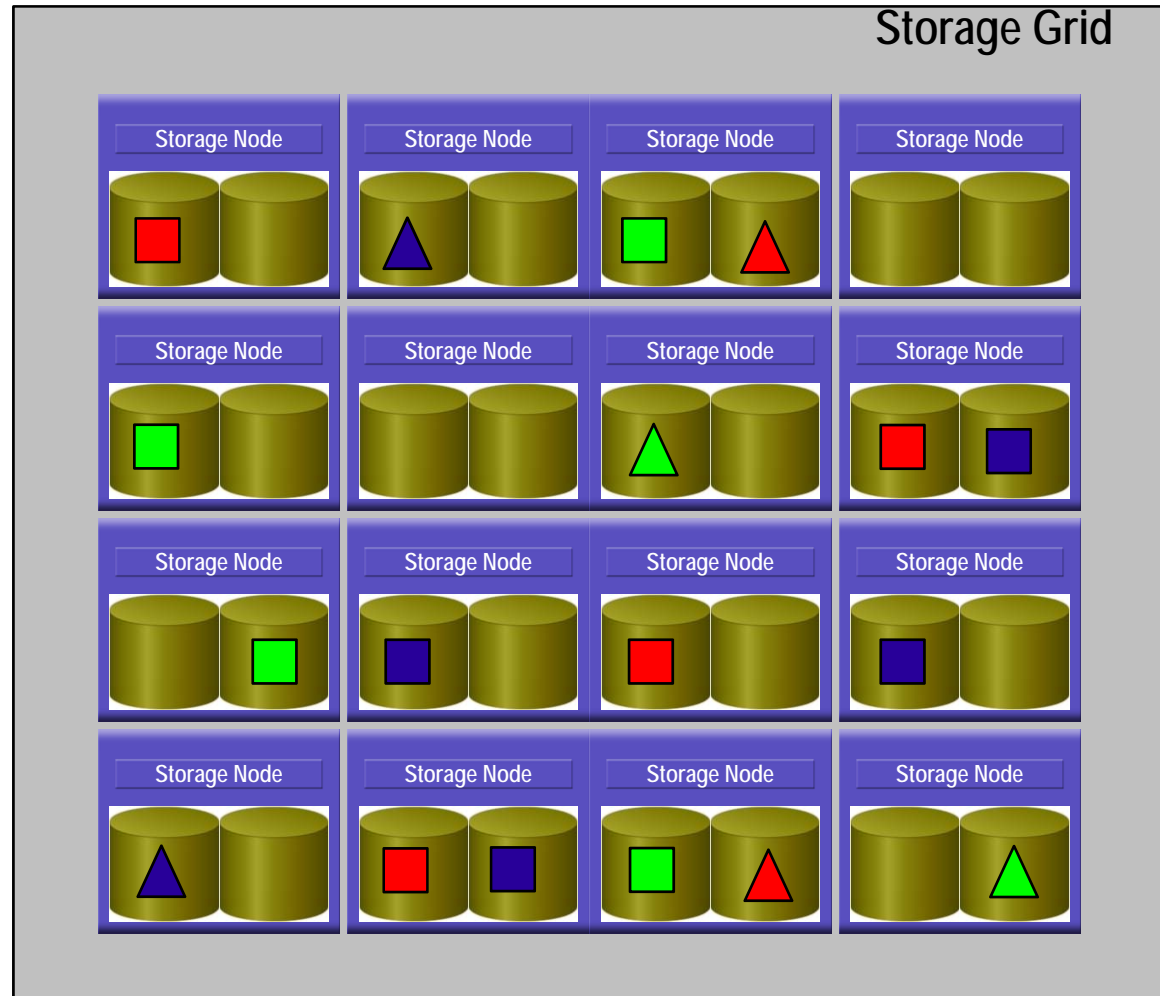


Object 3



□ = Data

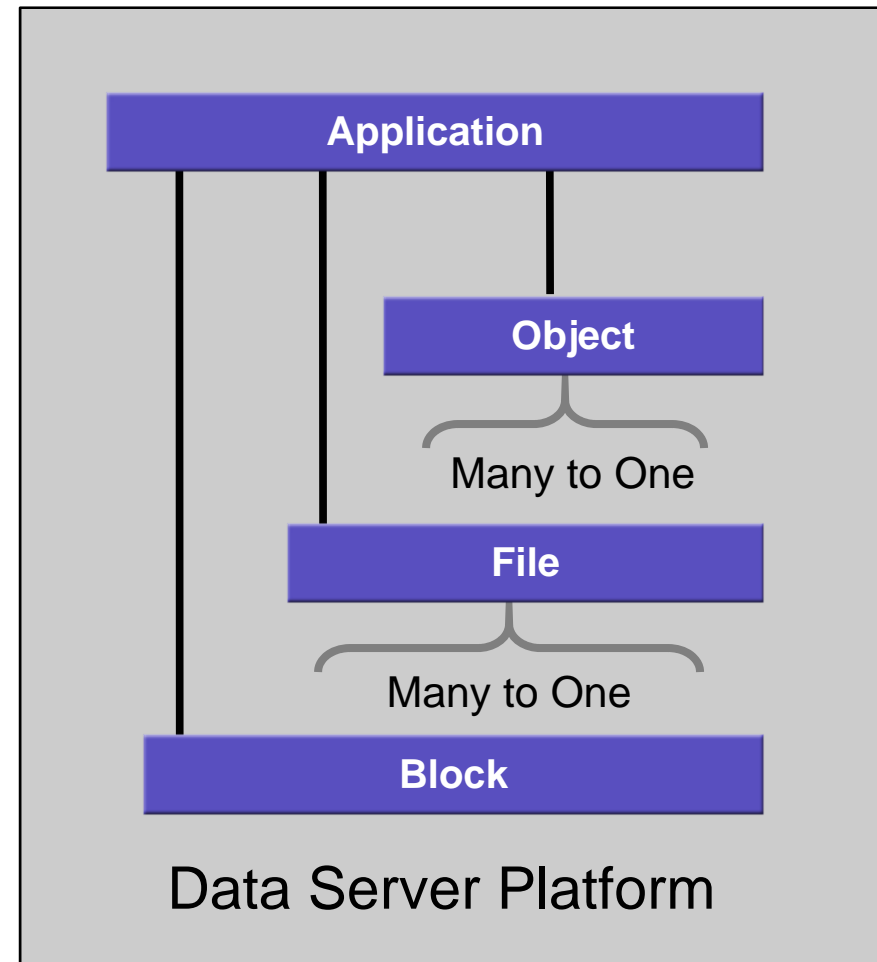
△ = Parity



# Data Serving Hierarchy

3 Levels of Abstraction

- Application may interface with the storage subsystem in anyone of three layers:
  - ◆ **Block** with highest performance and very little meta data
  - ◆ **File** with medium performance and some meta data
  - ◆ **Object** with medium performance and *rich* meta data



- Please send any questions or comments on this presentation to SNIA: [trackfilemgmt@snia.org](mailto:trackfilemgmt@snia.org)

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

**- SNIA Education Committee**

**Christian Bandulet, Sun Microsystems**