GUUG-Frühjahrsfachgespräch 2008

The File Systems Survey

Christian Bandulet
Principal Engineer
Data Management Ambassador
Sun Microsystems Inc. (Frankfurt, Germany)

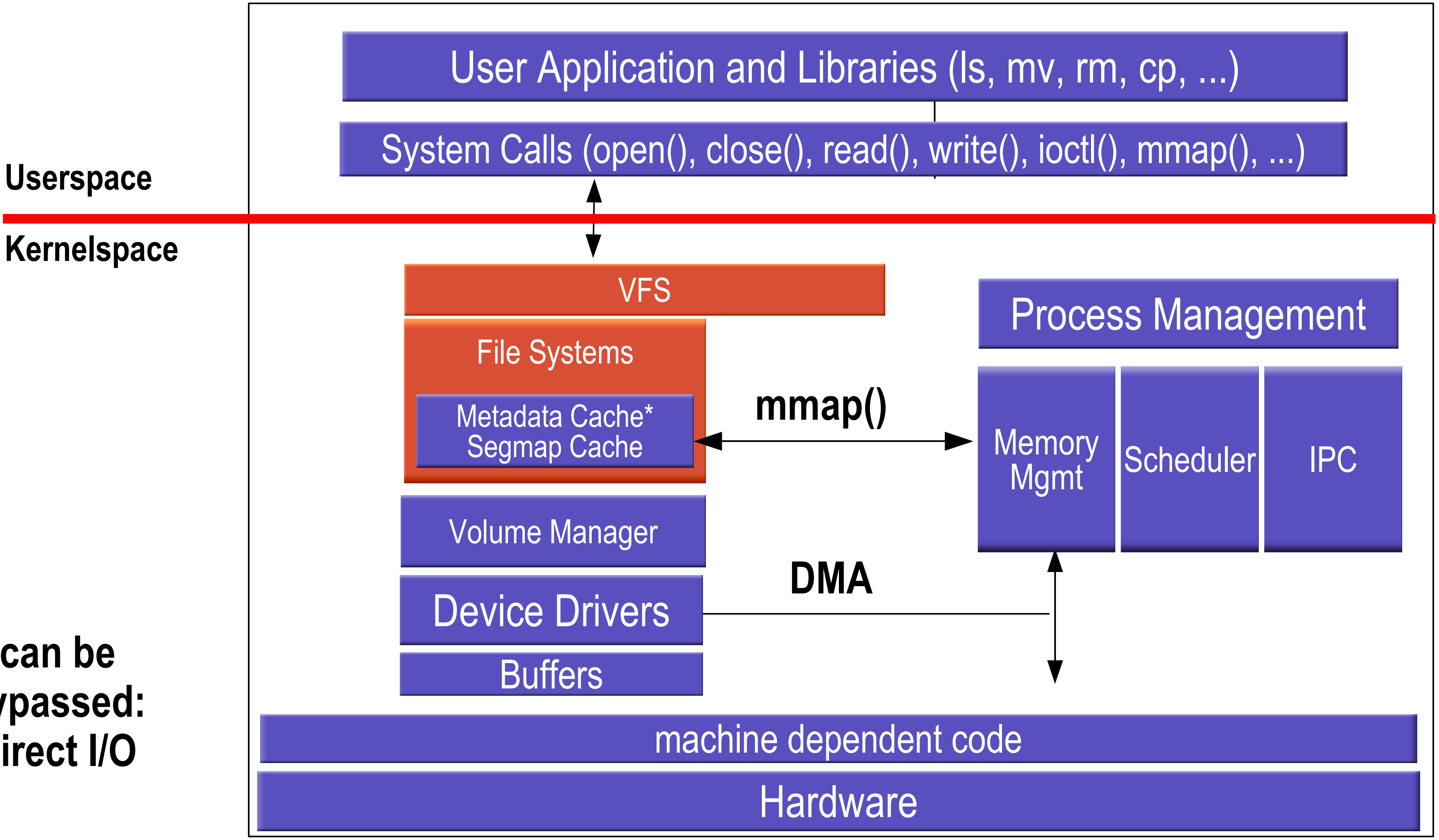
Agenda

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

Agenda

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

File System & Operating System



*can be bypassed: Direct I/O

Agenda

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

http://en.wikipedia.org/wiki/List_of_file_systems#Network_file_systems

- local/Disk File Systems
  - > # ADFS – Acorn's Advanced Disc filing system, successor to DFS.
  - > # BFS – the Be File System used on BeOS
  - > # EFS – Encrypted filesystem, An extension of NTFS
  - > # EFS (IRIX) – an older block filing system under IRIX.
  - > # Ext – Extended filesystem, designed for Linux systems
  - > # Ext2 – Second extended filesystem, designed for Linux systems.
  - > # **Ext3** – Name for the journalled form of ext2.
  - > # **FAT** – Used on DOS and Microsoft Windows, 12, 16 and 32 bit table depths
  - > # FFS (Amiga) – Fast File System, used on Amiga systems. This FS has evolved over time. Now counts FFS1, FFS Intl, FFS DCache, FFS2.
  - > # FFS – Fast File System, used on *BSD systems
  - > # Fossil – Plan 9 from Bell Labs snapshot archival file system.
  - > # Files-11 – OpenVMS filesystem
  - > # GCR – Group Code Recording, a floppy disk data encoding format used by the Apple II and Commodore Business Machines in the 5¼" disk drives for their 8-bit computers.
  - > # HFS – Hierarchical File System, used on older Mac OS systems

http://en.wikipedia.org/wiki/List_of_file_systems#Network_file_systems

- local/Disk File Systems (cont'd)
  - # HFS Plus – Updated version of HFS used on newer Mac OS systems
  - # HPFS – High Performance Filesystem, used on OS/2
  - # ISO 9660 – Used on CD-ROM and DVD-ROM discs (Rock Ridge and Joliet are extensions to this)
  - # **JFS** – IBM Journaling Filesystem, provided in Linux, OS/2, and AIX
  - # LFS – 4.4BSD implementation of a log-structured file system
  - # MFS – Macintosh File System, used on early Mac OS systems
  - # Minix file system – Used on Minix systems
  - # **NTFS** – Used on Windows NT, Windows 2000, Windows XP and Windows Server 2003 systems
  - # NSS – Novell Storage Services. This is a new 64-bit journaling filesystem using a balanced tree algorithm. Used in NetWare versions 5.0-up and recently ported to Linux.
  - # OFS – Old File System, on Amiga. Nice for floppies, but fairly useless on hard drives.
  - # PFS – and PFS2, PFS3, etc. Technically interesting filesystem available for the Amiga, performs very well under a lot of circumstances. Very simple and elegant.
  - # **ReiserFS** – Filesystem that uses journaling
  - # Reiser4 – Filesystem that uses journaling, newest version of ReiserFS
  - # SFS – Smart File System, journaled file system available for the Amiga platforms.
  - # UDF – Packet based filesystem for WORM/RW media such as CD-RW and DVD.

http://en.wikipedia.org/wiki/List_of_file_systems#Network_file_systems

- Local/Disk File Systems (cont'd)
  - # UDF – Packet based filesystem for WORM/RW media such as CD-RW and DVD.
  - # **UFS** – Unix Filesystem, used on older BSD systems
  - # UFS2 – Unix Filesystem, used on newer BSD systems
  - # UMSDOS – FAT filesystem extended to store permissions and metadata, used for Linux.
  - # **VxFS** – Veritas file system, first commercial journaling file system; HP-UX, Solaris, Linux, AIX
  - # VSAM
  - # **WAFL** – Used on Network Appliance systems
  - # XFS – Used on SGI IRIX and Linux systems
  - # **ZFS – Used on Solaris 10**

http://en.wikipedia.org/wiki/List_of_file_systems#Network_file_systems

- Distributed/Network File Systems

> * 9P The Plan 9 and Inferno distributed file system
> * **AFS** (Andrew File System)
> * AppleShare
> * Arla (file system)
> * Coda
> * CXFS (Clustered XFS) a distributed networked file system designed by Silicon Graphics (SGI) specifically to be used in a SAN
> * Distributed File System (DCE)
> * **Distributed File System** (Microsoft)
> * Freenet
> * Global File System (GFS)
> * **Google File System** (GFS)
> * IBRIX Fusion™
> * InterMezzo
> * Isilon OneFS™
> * **Lustre**
> * **NFS**
> * OpenAFS
> * Server message block (SMB) (aka Common Internet File System (**CIFS**) or Samba file system)
> * Xsan (a storage area network (SAN) filesystem from Apple Computer, Inc.)
>

http://en.wikipedia.org/wiki/List_of_file_systems#Network_file_systems

- Special Purpose File Systems
  - > # acme (Plan 9) (text windows)
  - > # archfs (archive)
  - > # **cdfs** (reading and writing of CDs)
  - > # cfs (caching)
  - > # Davfs2 (**WebDAV**)
  - > # devfs
  - > # ftpfs (ftp access)
  - > # fuse (filesystem in userspace, like lufs but better maintained)
  - > # **GPFS** an IBM cluster file system
  - > # JFFS/JFFS2 (filesystems designed specifically for flash devices)
  - > # lnfs (long names)
  - > # LUFS (replace ftpfs, ftp ssh ... access)
  - > # nntpfs (netnews)
  - > # OCFS (Oracle Cluster File System)

http://en.wikipedia.org/wiki/List_of_file_systems#Network_file_systems

- Special Purpose File Systems (cont'd)
  - > # ParFiSys (Experimental parallel file system for massively parallel processing)
  - > # plumber (Plan 9) (interprocess communication – pipes)
  - > # **procfs**
  - > # romfs
  - > # specfs (Special Filesytem for device files)
  - > # SquashFS (compressed read-only)
  - > # sysfs (Linux)
  - > # tmpfs
  - > # wikifs (Plan 9) (wiki wiki)
  - > # **pvfs** (Parallel Virtual File System)
  - > # pvfs2 (Parallel Virtual File System, 2nd generation)

Some Technologies and Products…

IBM AFS WebNFS Cisco FileEngine Apple Xsan VERITAS CFS

ISO9660 Coda DiskSites FilePort RFS

DB2 PolyServe Matrix Server Oracle OPS/RAC Samba FineGround

IBM Lustre FS

CXFS

HP TruCluster/CFS ADIC StorNext FS Redhat GFS

OpenAFS

Sybase

Informix XPS

IBM SANergy Sanbolic MelioFS

Tacit Networks Ishared PVFS WebNFS DFS

SMB

OSD Isilon IQ OneFS Nuview StorageX pNFS

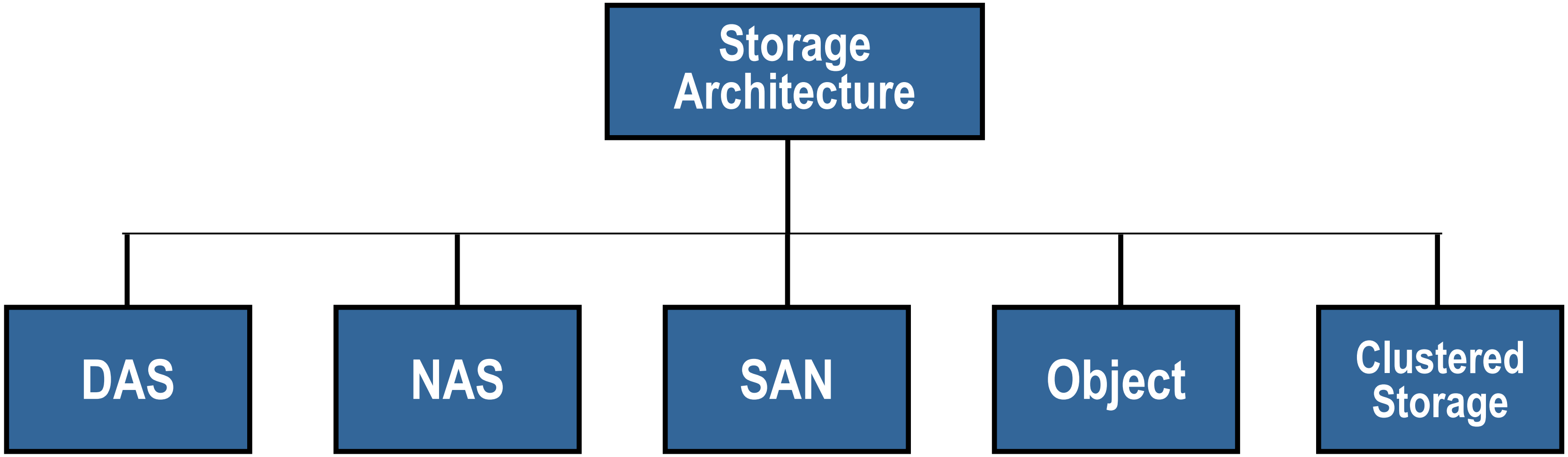
More than 100 products exist on the planet !!

© Copyright: christian.bandulet@sun.com

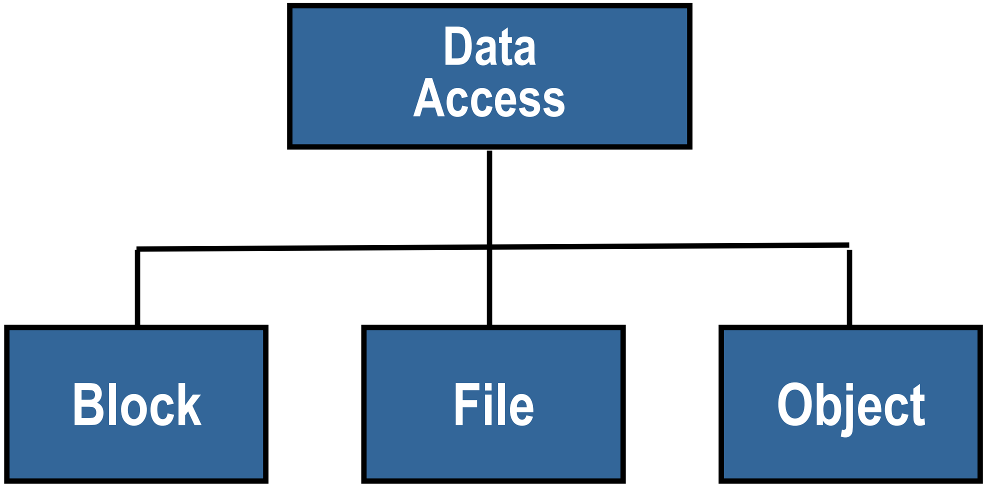
Source: www.snia.org

FS & Storage Architectures

File systems can run on arbitrary storage architectures:



Data Access Taxonomy



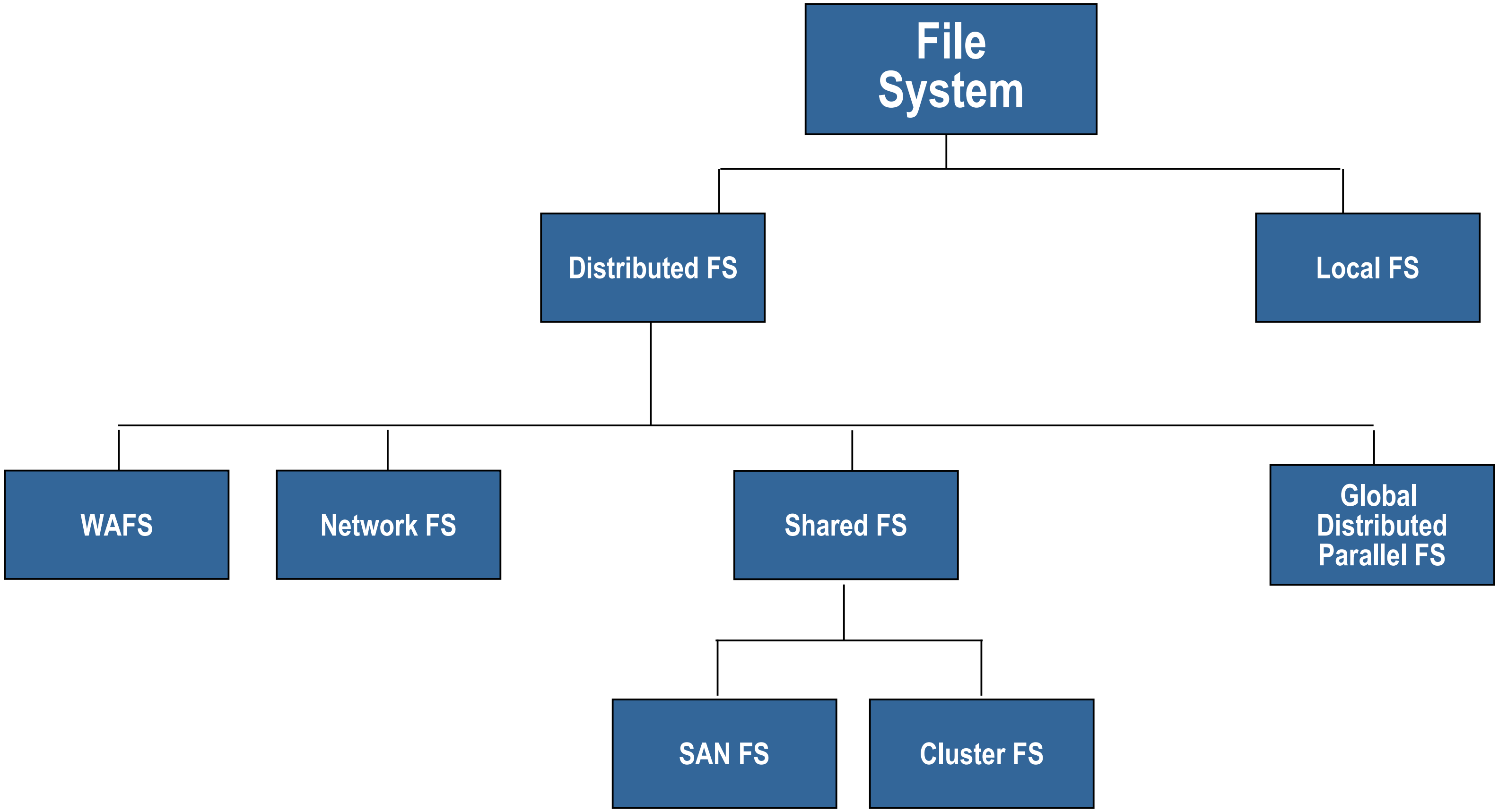
Data Access

Block

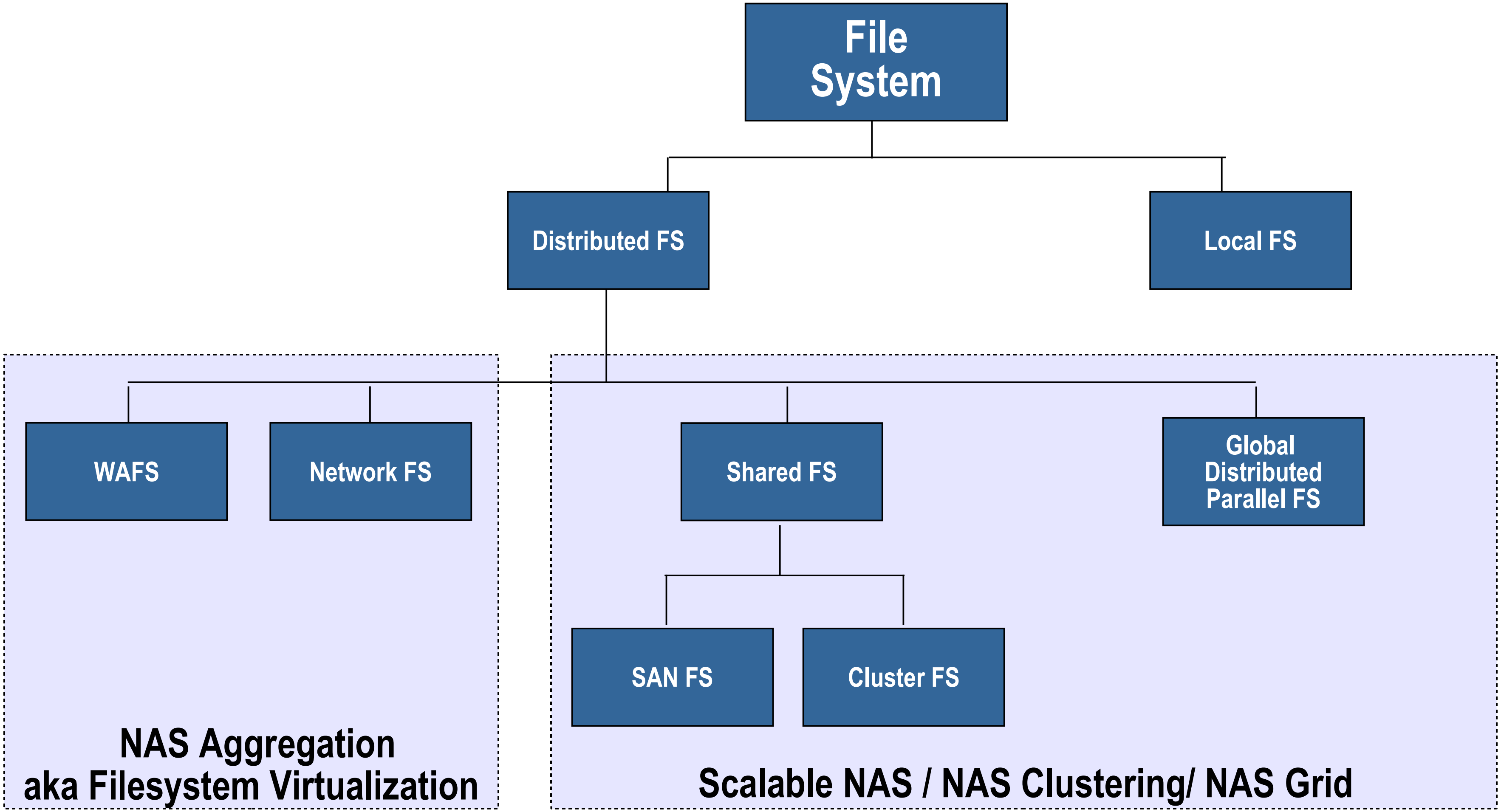
File

Object

File System Taxonomy

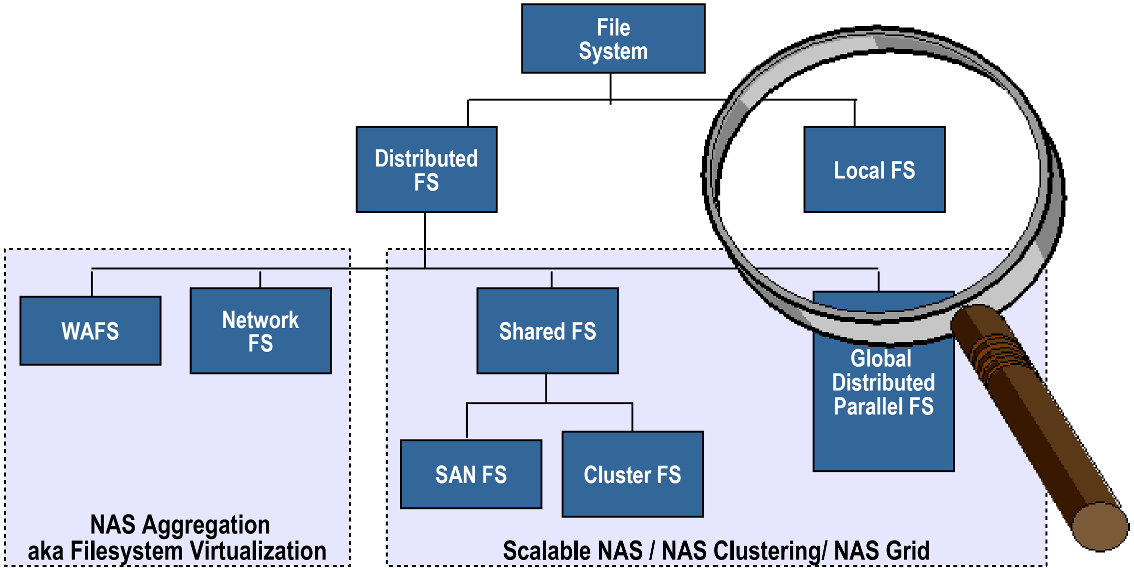


File System Taxonomy


File System
Distributed FS
Local FS
WAFS
Network FS
Shared FS
Global Distributed Parallel FS
SAN FS
Cluster FS
NAS Aggregation
aka Filesystem Virtualization
Scalable NAS / NAS Clustering/ NAS Grid

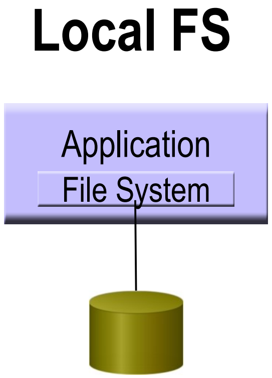
Agenda

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



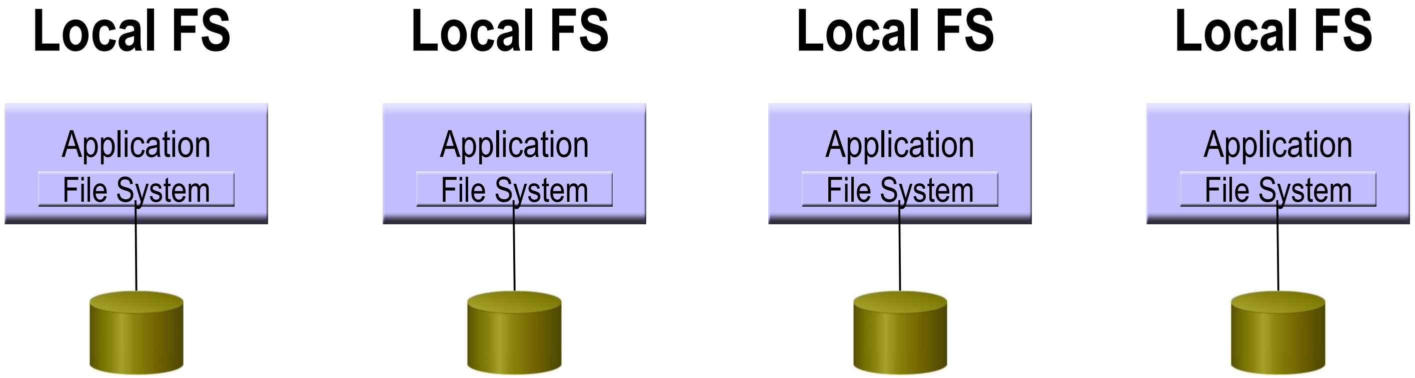
File System
Distributed FS
Local FS
WAFS
Network FS
Shared FS
Global Distributed Parallel FS
SAN FS
Cluster FS
NAS Aggregation
aka Filesystem Virtualization
Scalable NAS / NAS Clustering/ NAS Grid

Local FS



- **Co-located** with application server

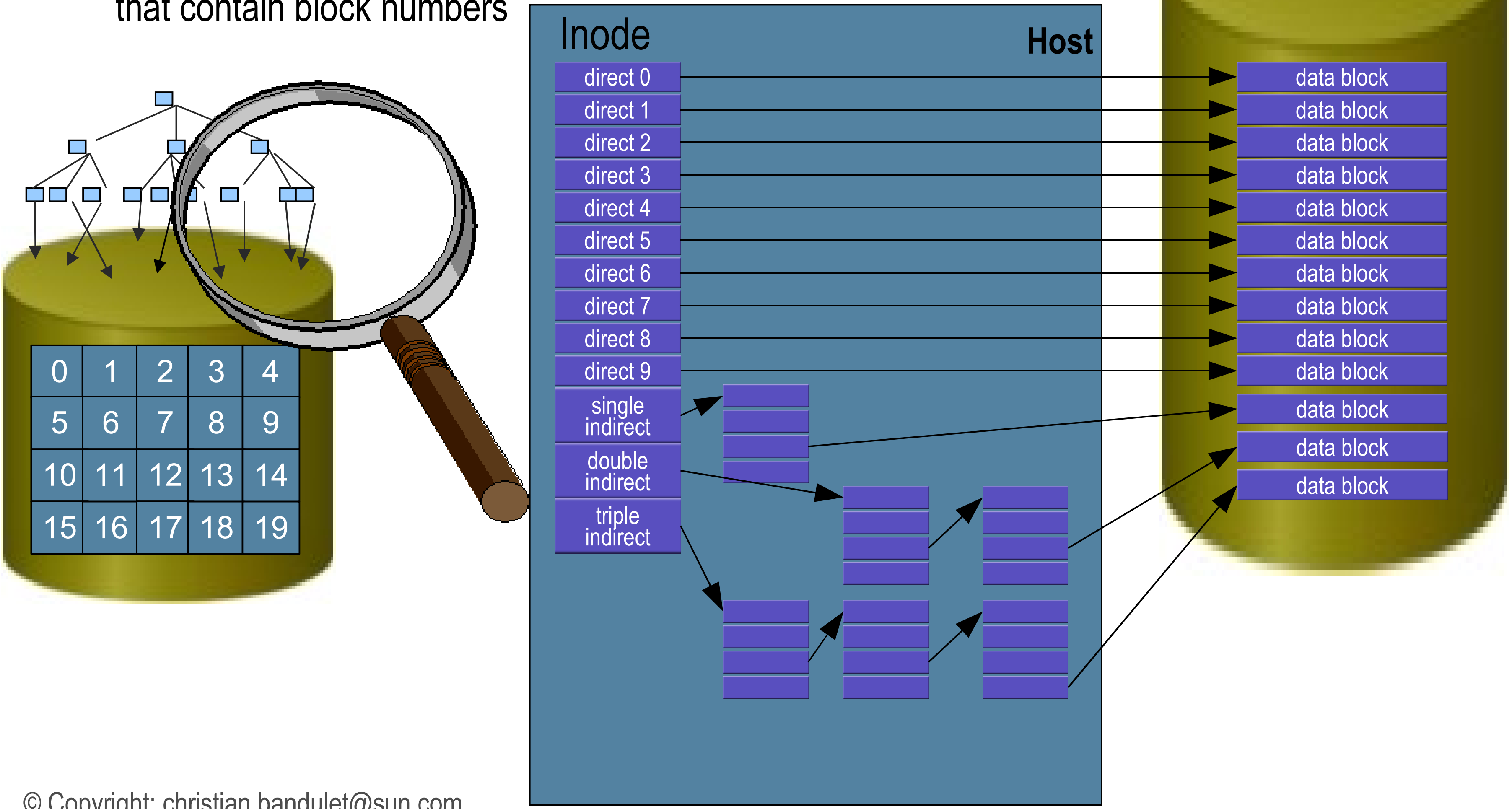
Local FS



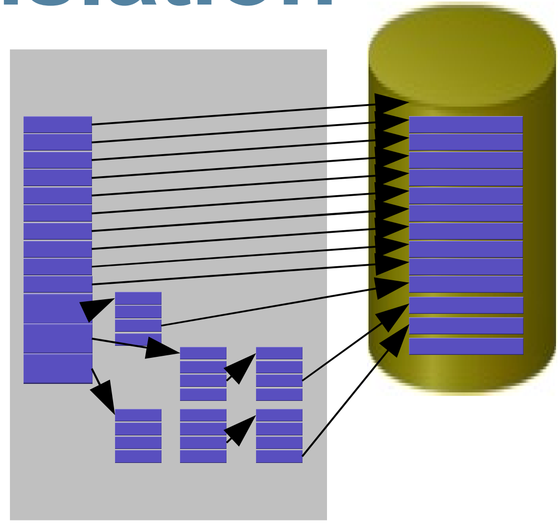
- **Islands of storage** (limited data sharing)

Traditional File System - Inode

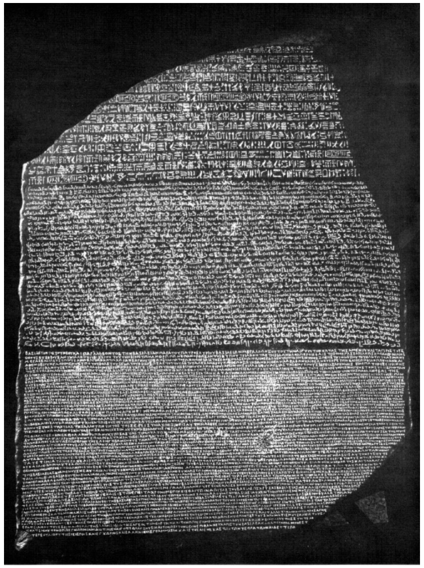
- The inode contains a few block numbers to ensure efficient access to small files. Access to larger files is provided via indirect blocks that contain block numbers



Logical to Physical Translation



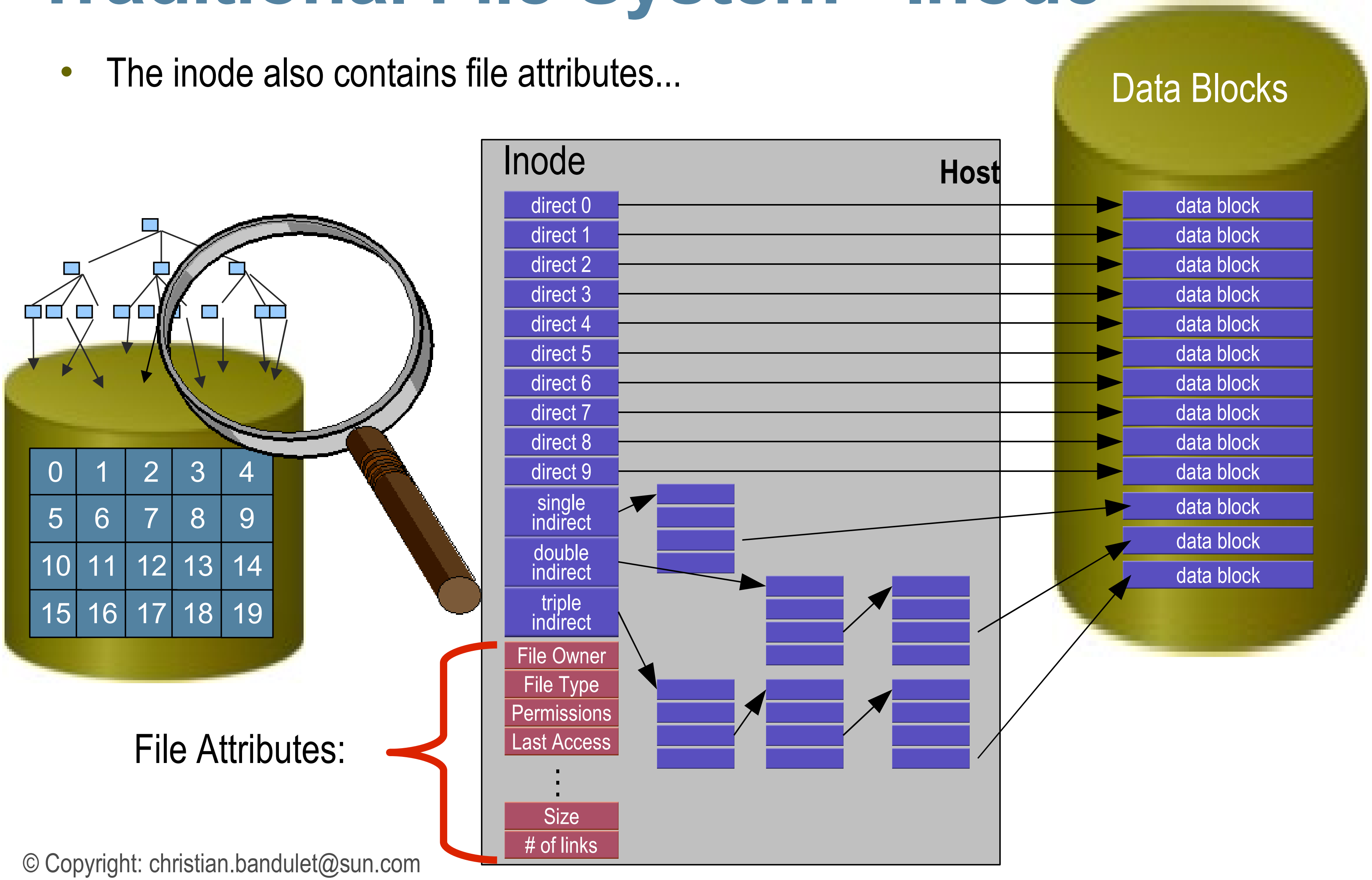
Hieroglyphs: 3100 B.C - 400 A.D



Rosetta Stone: was created in 196 BC,
discovered by the French in 1799 at Rosetta,
a harbor on the Mediterranean coast in Egypt,
and translated in 1822 by
Frenchman Jean-François Champollion

Traditional File System - Inode

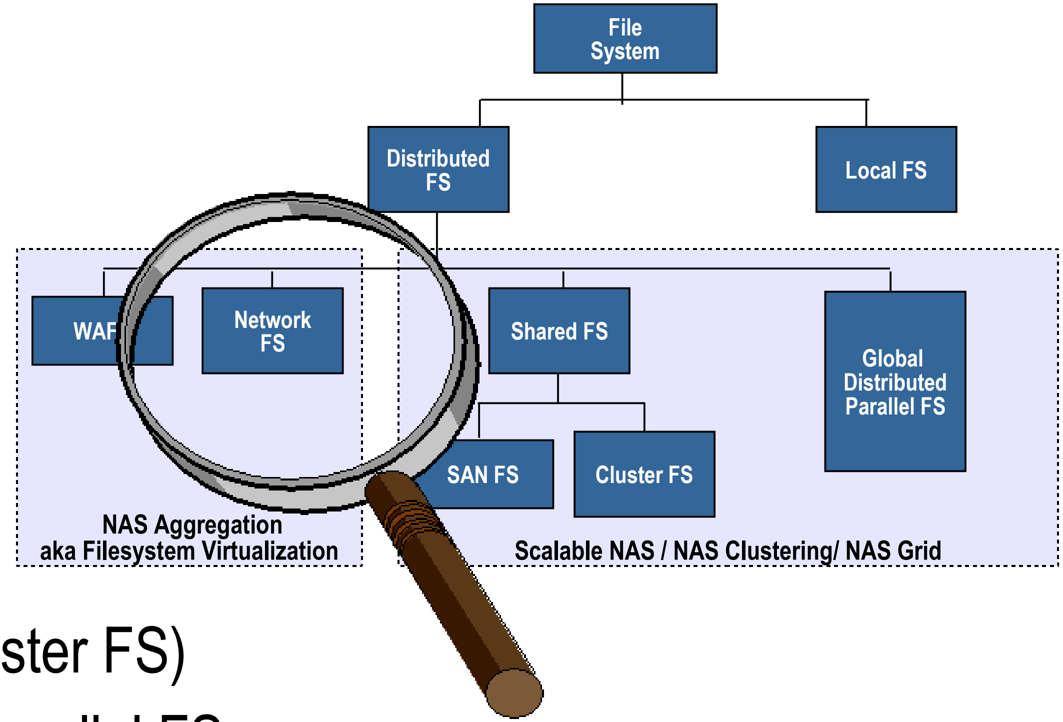
- The inode also contains file attributes...



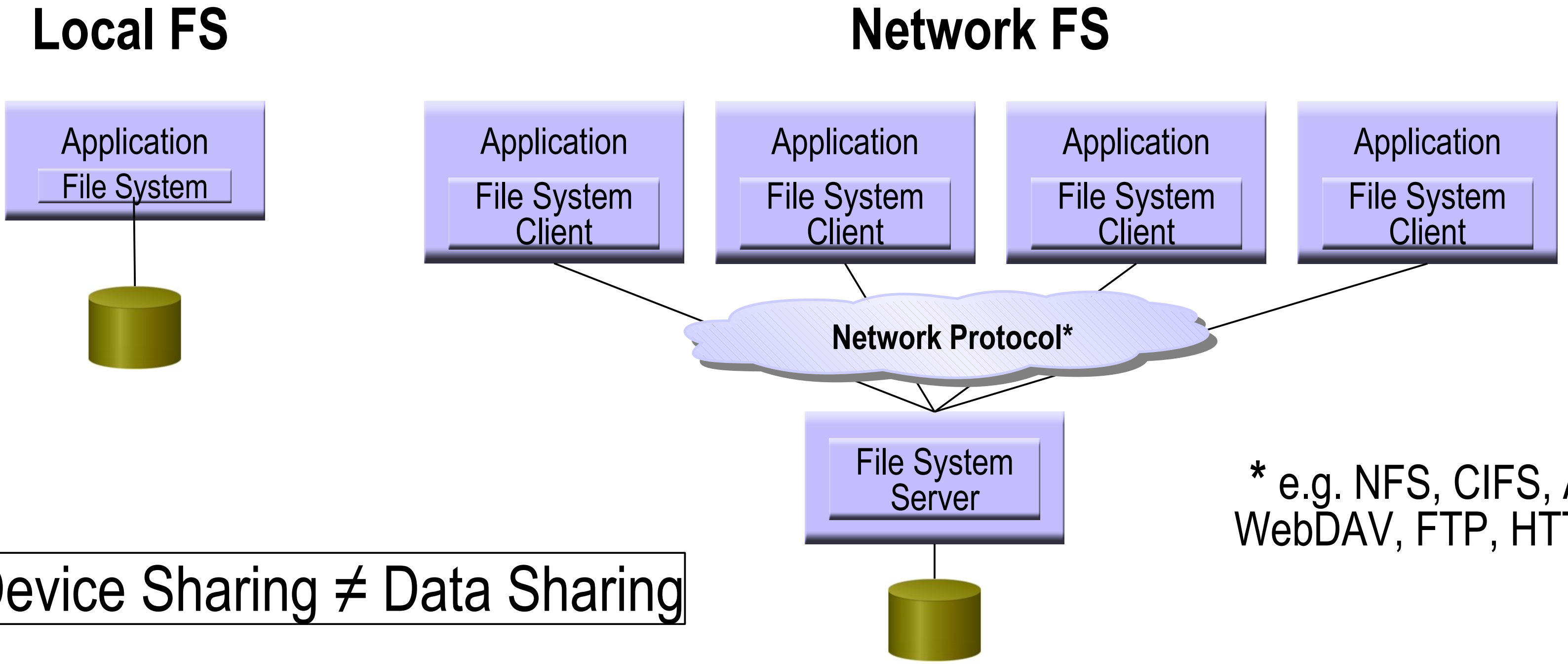
0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19

Agenda

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


File System
Distributed FS
Local FS
WAFS
Network FS
Shared FS
Global Distributed Parallel FS
SAN FS
Cluster FS
NAS Aggregation aka Filesystem Virtualization
Scalable NAS / NAS Clustering/ NAS Grid

Network Files System – aka Proxy FS

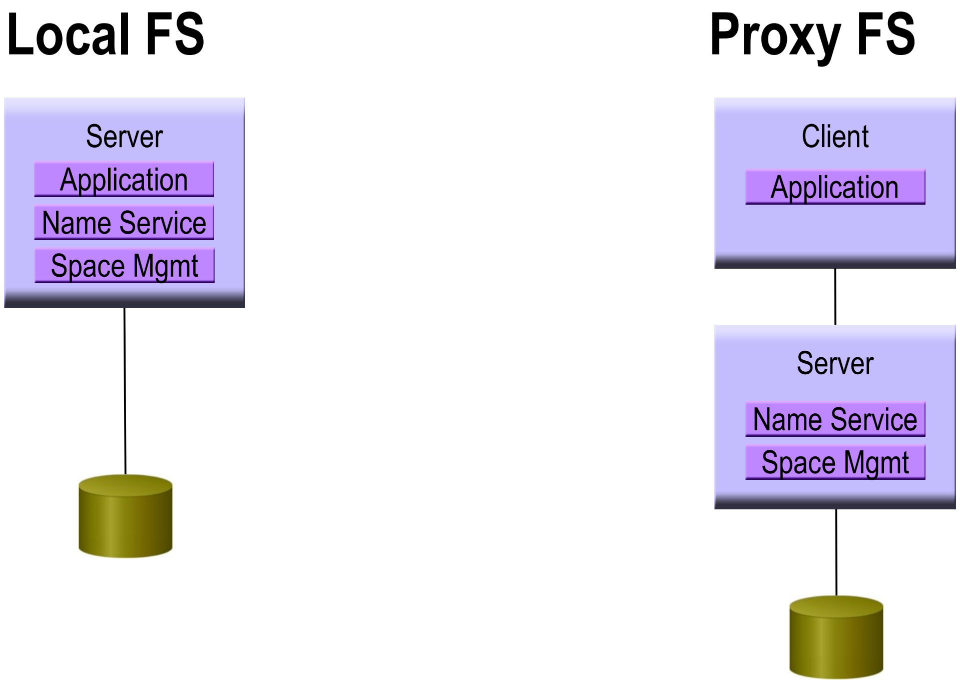


* e.g. NFS, CIFS, AFP, WebDAV, FTP, HTTP, ...

Device Sharing ≠ Data Sharing

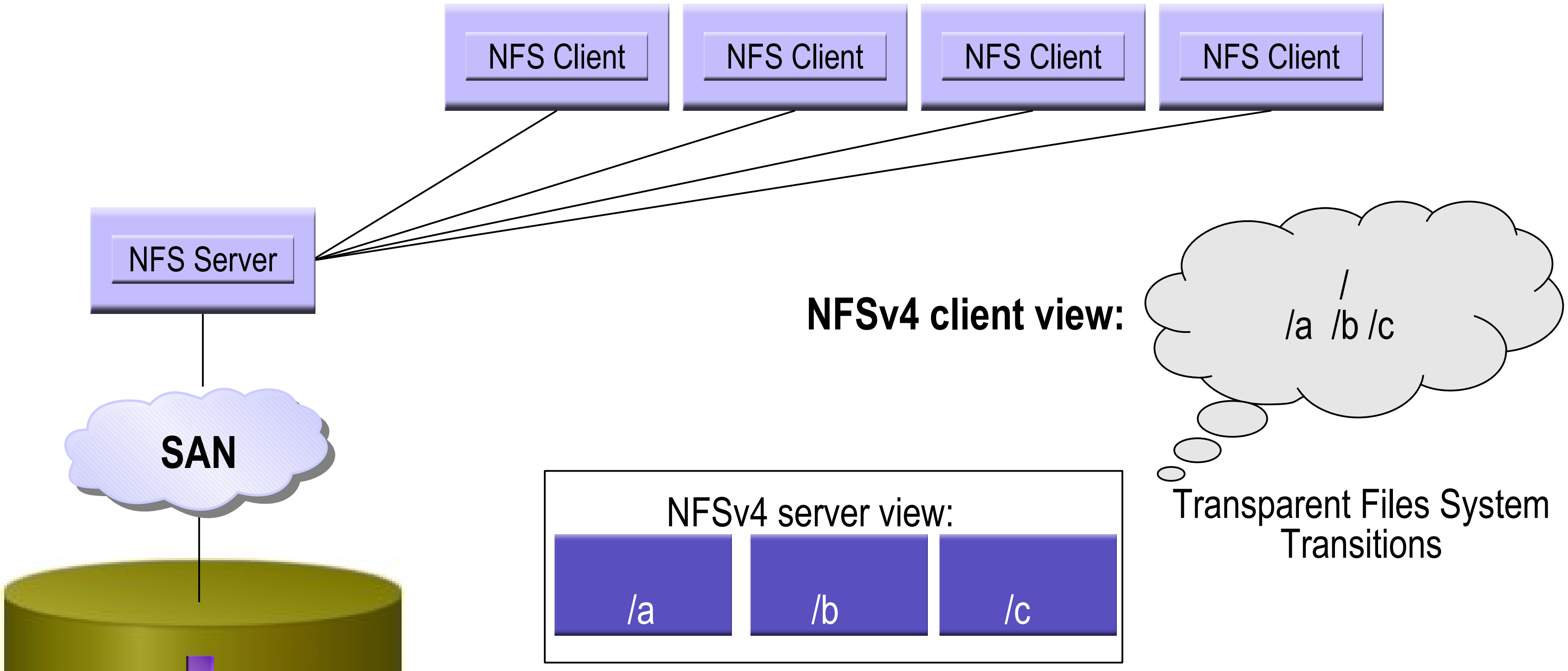
- A network file system is any computer file system that supports **sharing of files over a computer network protocol**

Local FS and Proxy FS



NFSv4 Single-Server Namespace Extension

Server Pseudo FS – aka Shared Name Space



NFS Client

NFS Client

NFS Client

NFS Client

NFS Server

SAN

NFSv4 client view:

/
/a /b /c

NFSv4 server view:

/a

/b

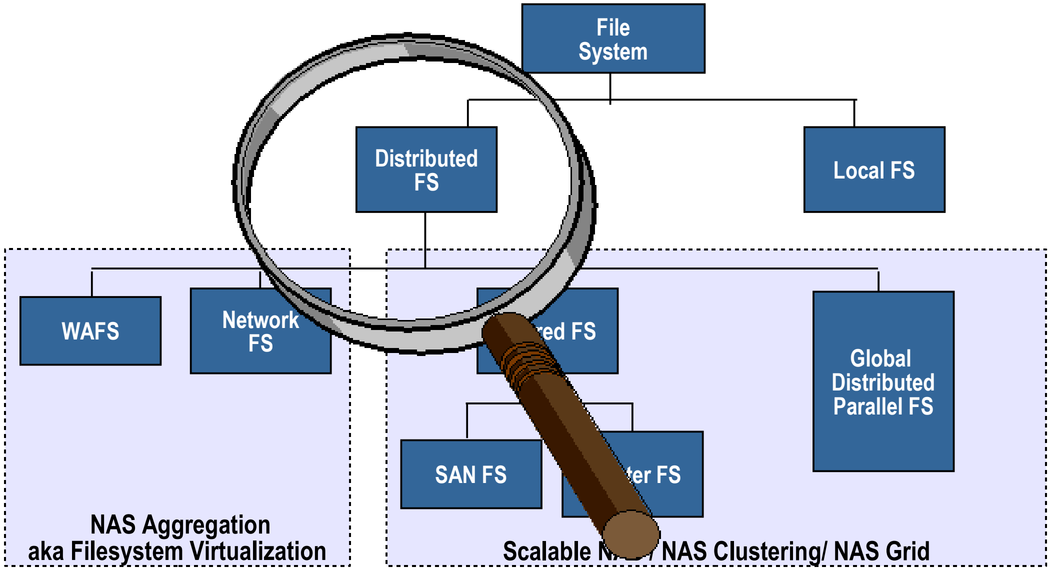
/c

Transparent Files System Transitions

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 filesystems on the server transparently.

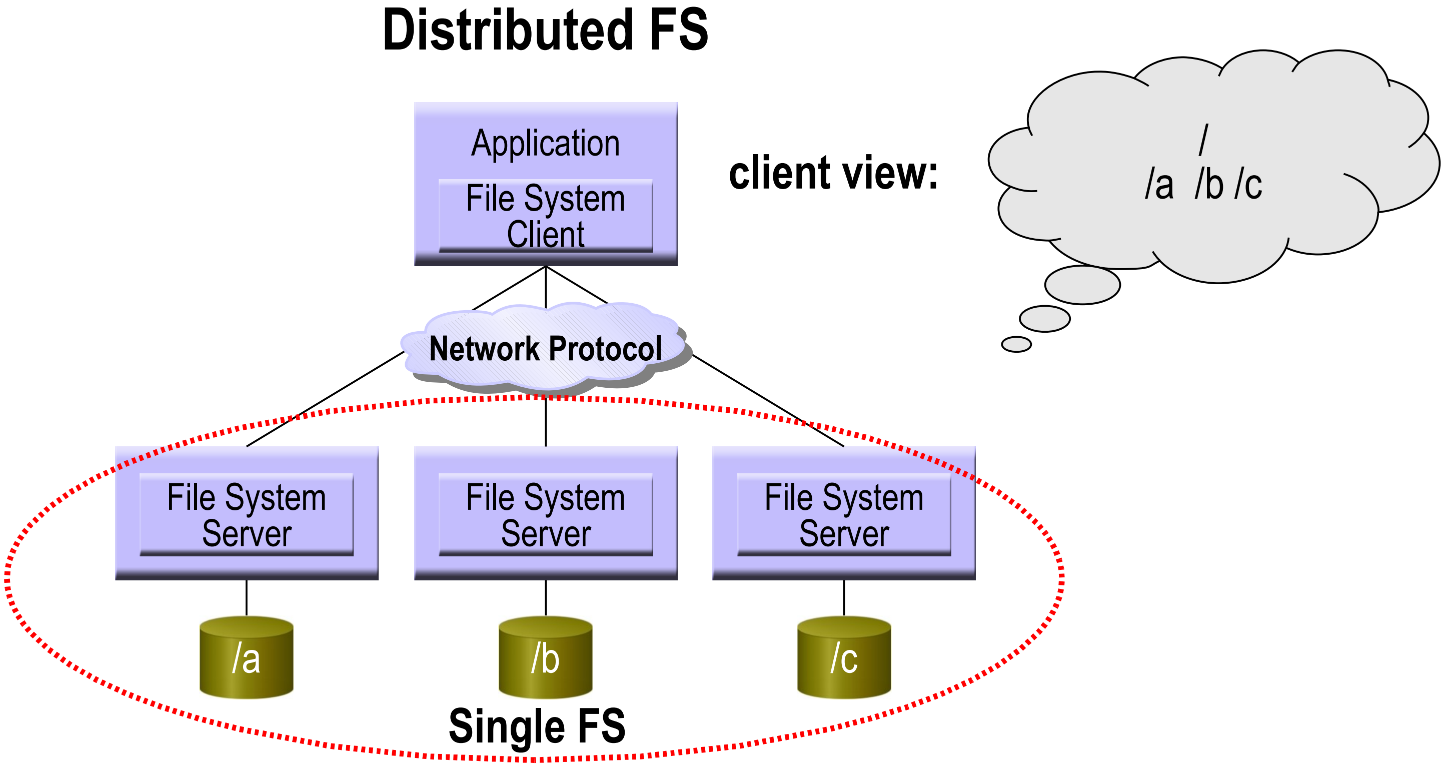
Agenda

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


File System
Distributed FS
Local FS
WAFS
Network FS
red FS
Global Distributed Parallel FS
SAN FS
ter FS
NAS Aggregation
aka Filesystem Virtualization
Scalable N / NAS Clustering/ NAS Grid

© Copyright: christian.bandulet@sun.com

Distributed File System (DFS)

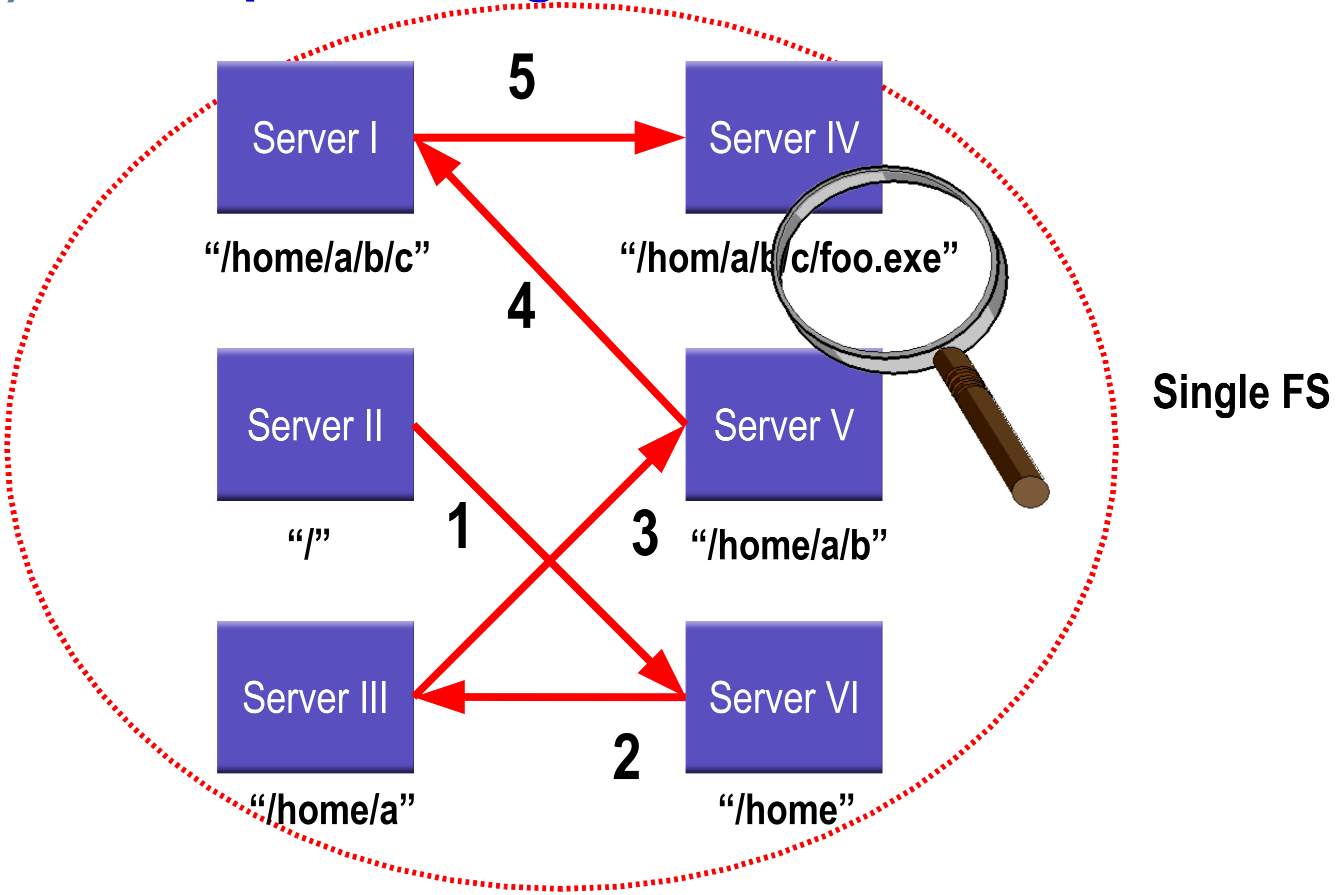


- **A distributed file system is a network file system** whose clients, servers, and storage devices are dispersed among the machines of a distributed system or intranet (≠ Parallel FS)

Distributed File System (DFS)

Andrew FS (AFS), www.OpenAFS.org, Coda

“read /home/a/b/c/foo.exe”

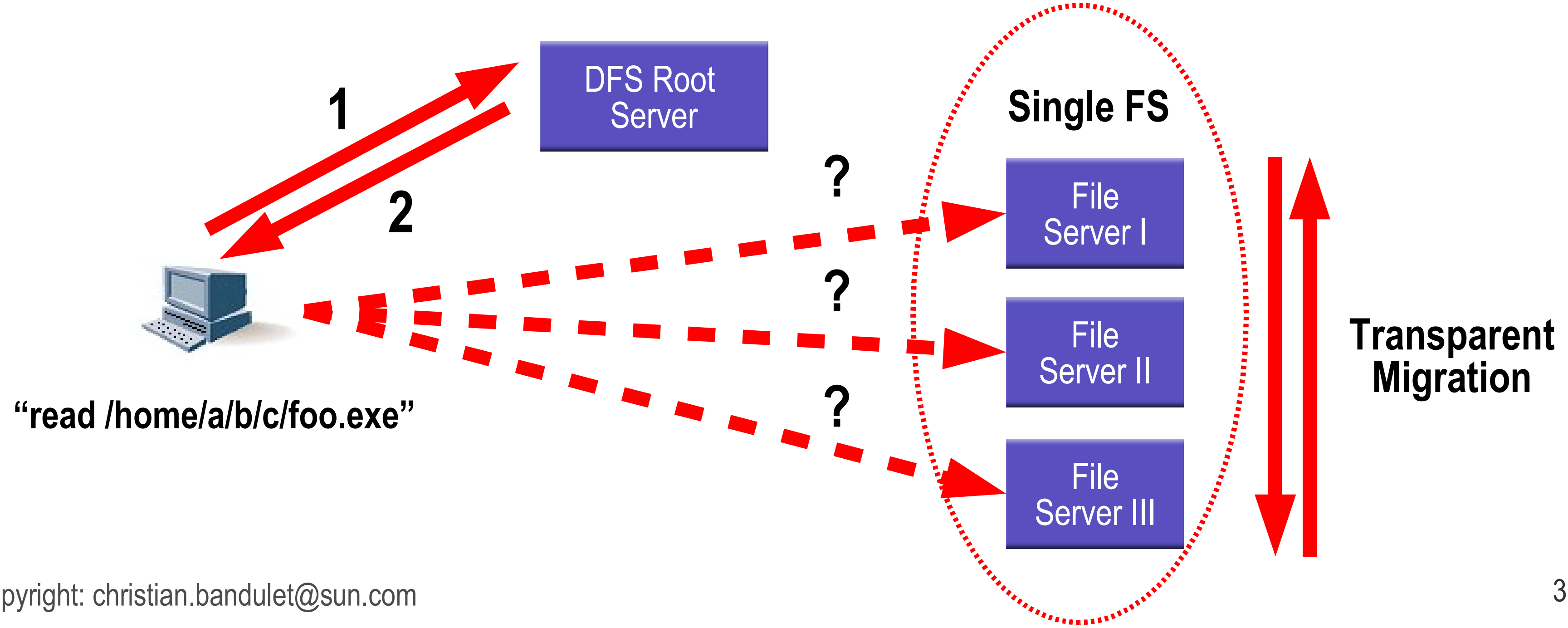


- Using Ethernet as a networking protocol between nodes, a DFS allows **a single file system to span across all nodes** in the DFS cluster, effectively creating a unified logical namespace for all files.

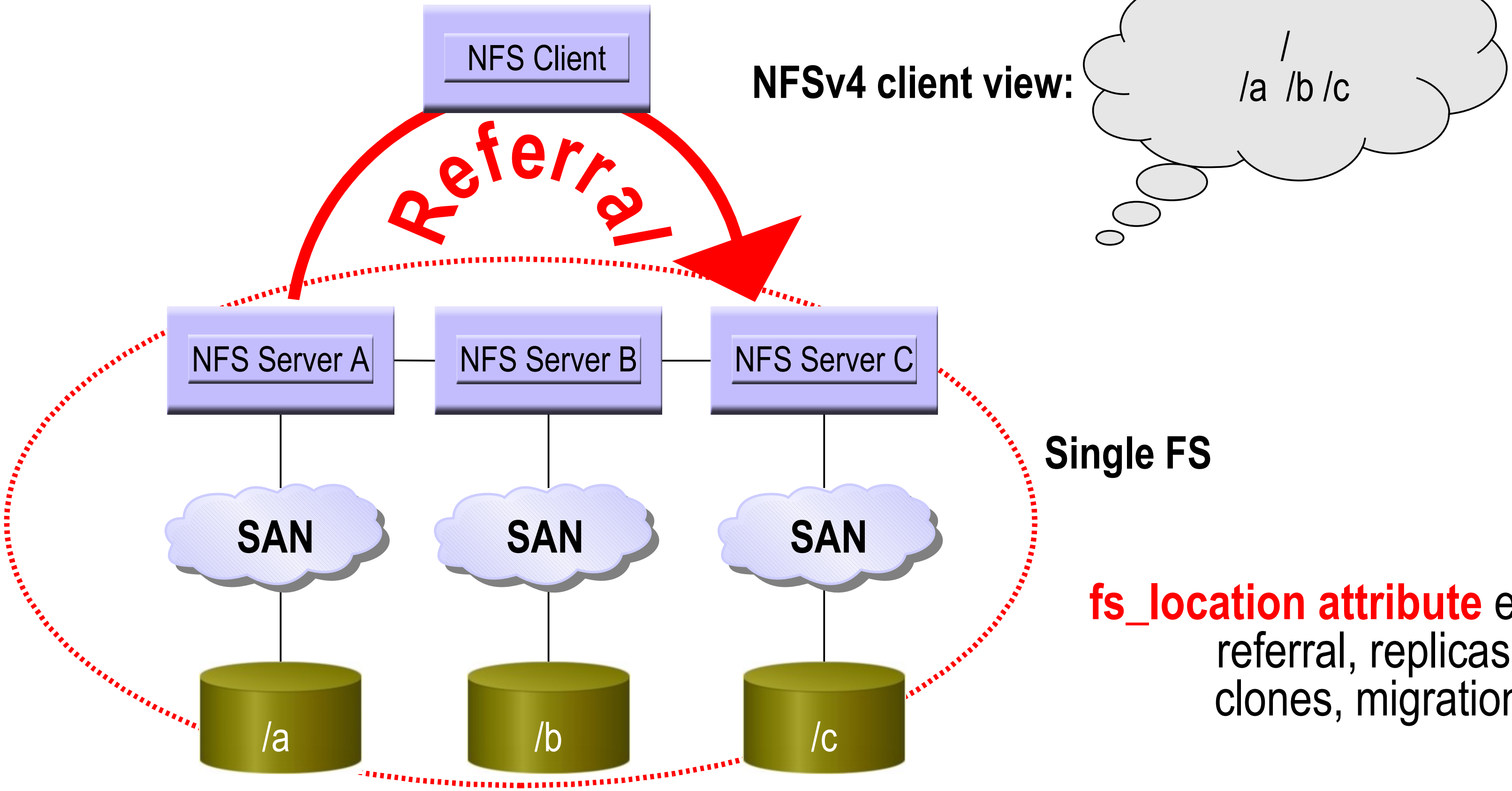
Distributed File System (DFS)

MS Distributed File System (DFS)

- Uniting files on different computers into a **single namespace**
- With DFS, you can make files distributed across multiple servers appear to users as if they reside in one place on the network
- Users no longer need to know and specify the actual physical location of files in order to access them.
- **Logical file location is de-coupled from physical location**



NFSv4.1 – Multi-Server Name Space

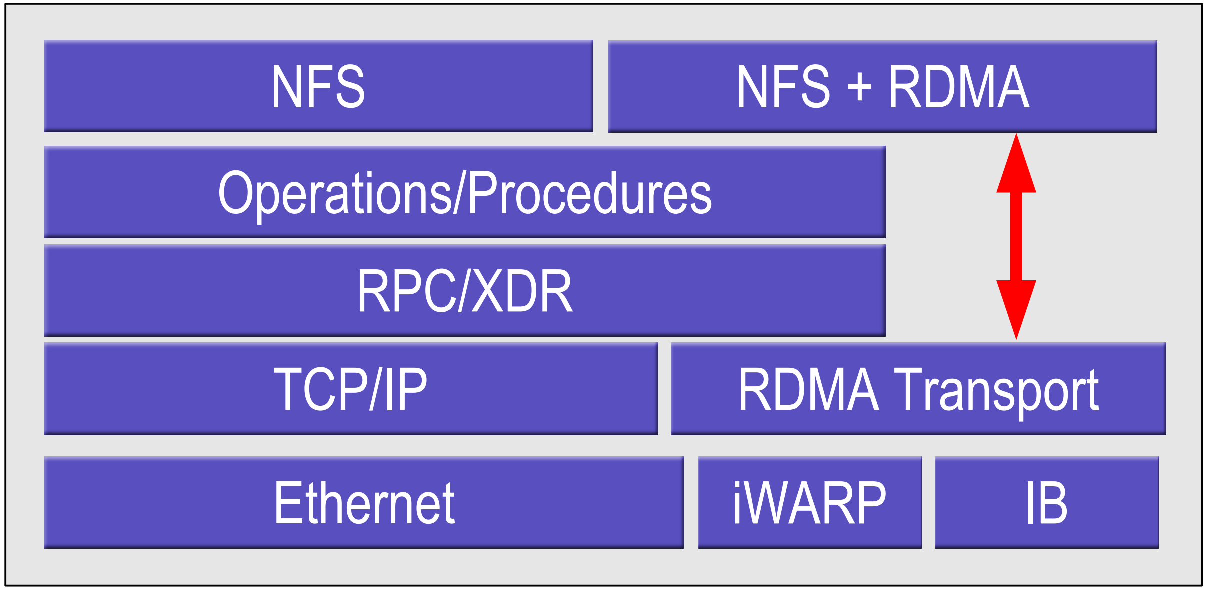


fs_location attribute enables: referral, replicas, clones, migration

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 an **enterprise namespace**

NFS RDMA Problem Statement

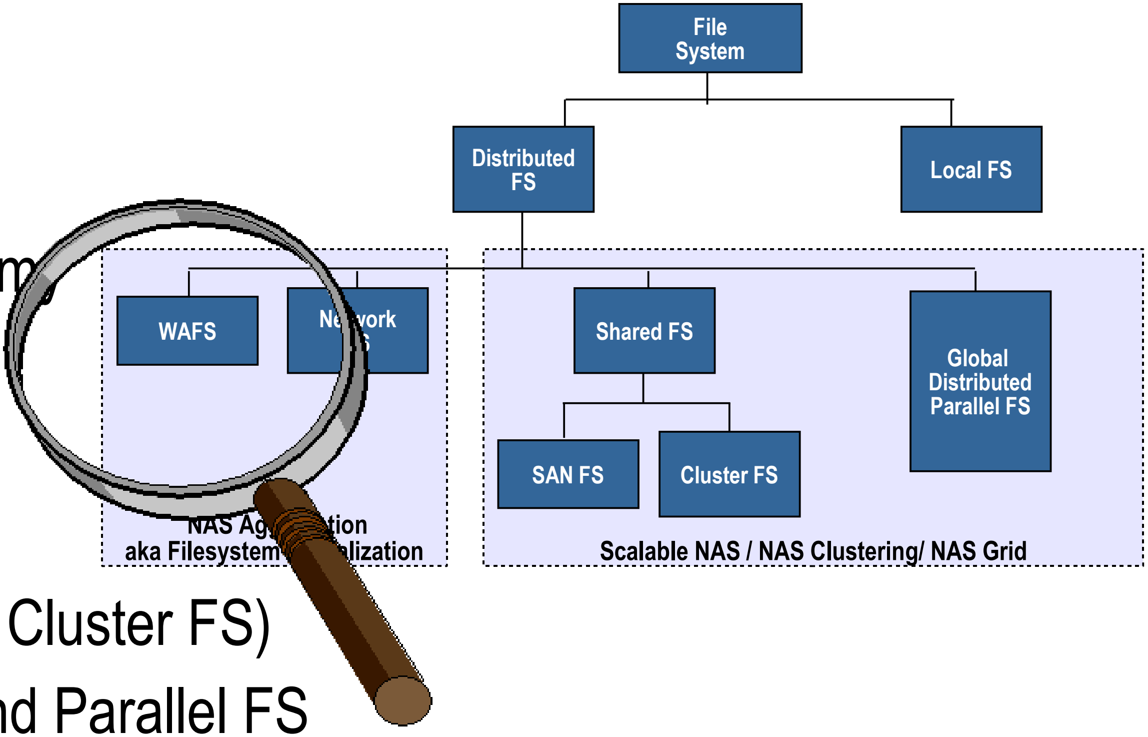
Block Diagram:



- http://ietf.org/html.charters/nfsv4-charter.html
- http://ietf.org/internet-drafts/draft-ietf-nfsv4-nfs-rdma-problem-statement-05.txt
- http://ietf.org/internet-drafts/draft-ietf-nfsv4-rpcrdma-04.txt
- http://ietf.org/internet-drafts/draft-ietf-nfsv4-nfsdirect-04.txt

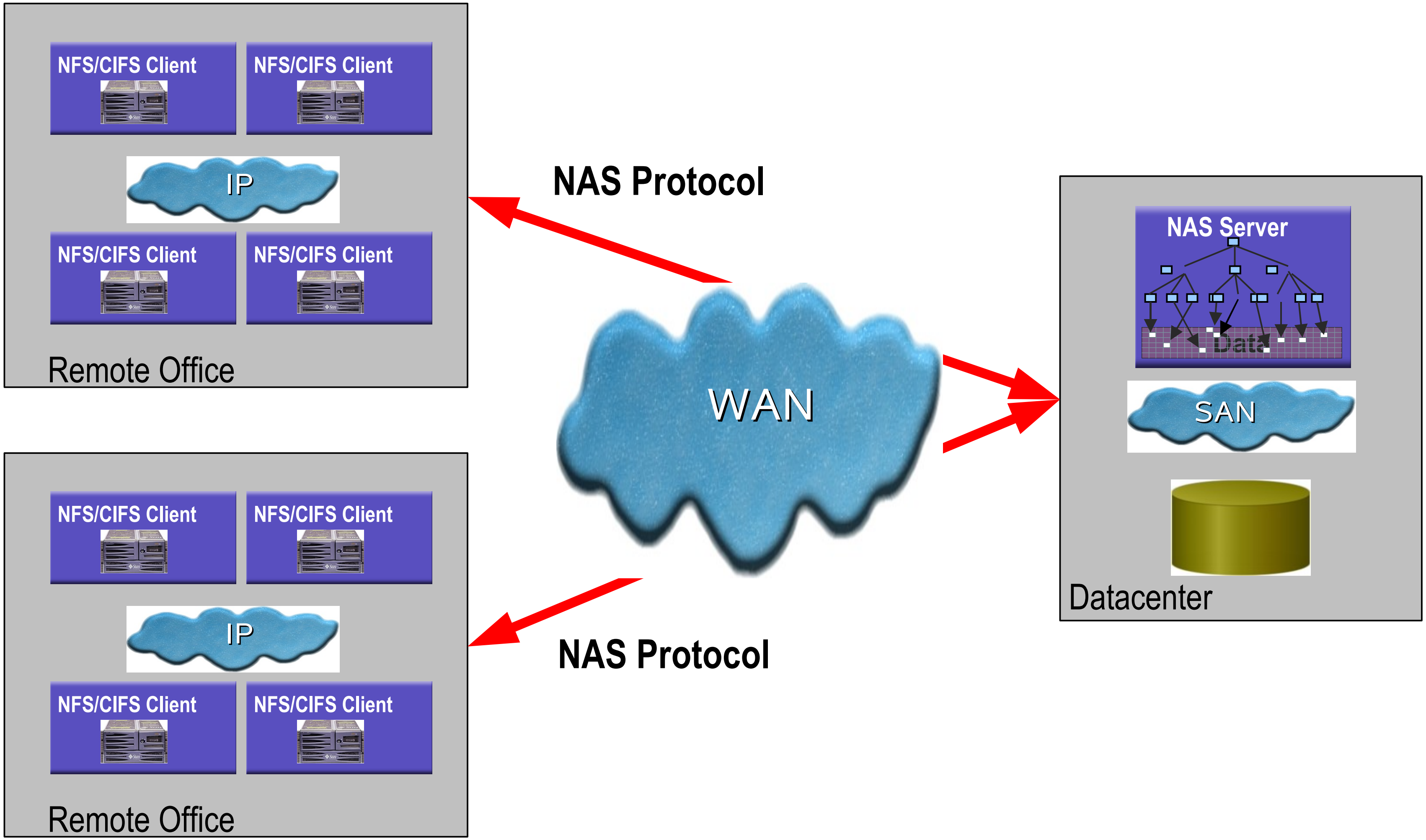
Agenda

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


File System
Distributed FS
Local FS
WAFS
Network FS
Shared FS
Global Distributed Parallel FS
SAN FS
Cluster FS
NAS Aggregation aka Filesystem Virtualization
Scalable NAS / NAS Clustering/ NAS Grid

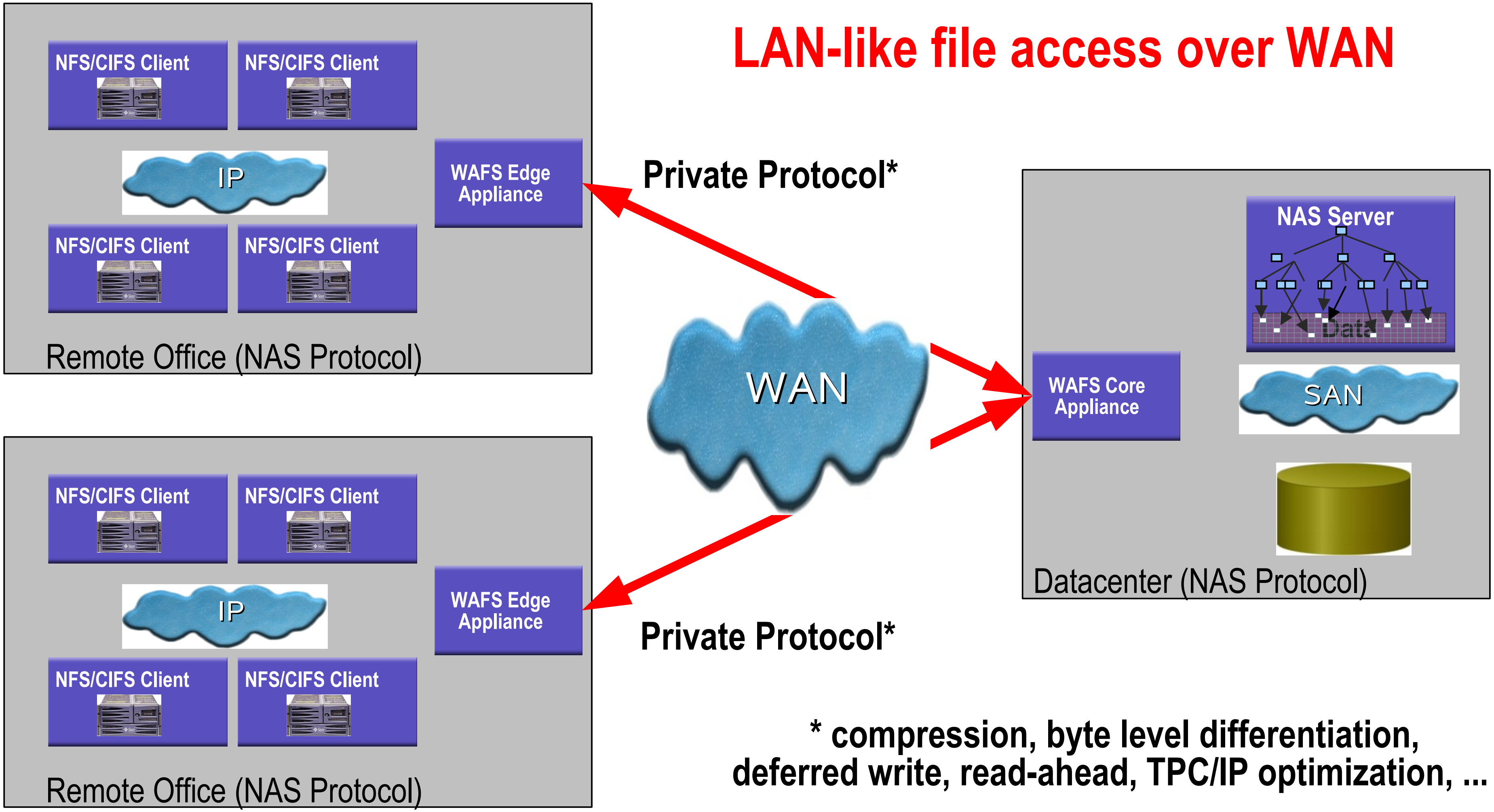
WAFS – aka Network Compression

Problem Statement



WAFS (NAS Aggregation/Virtualization)

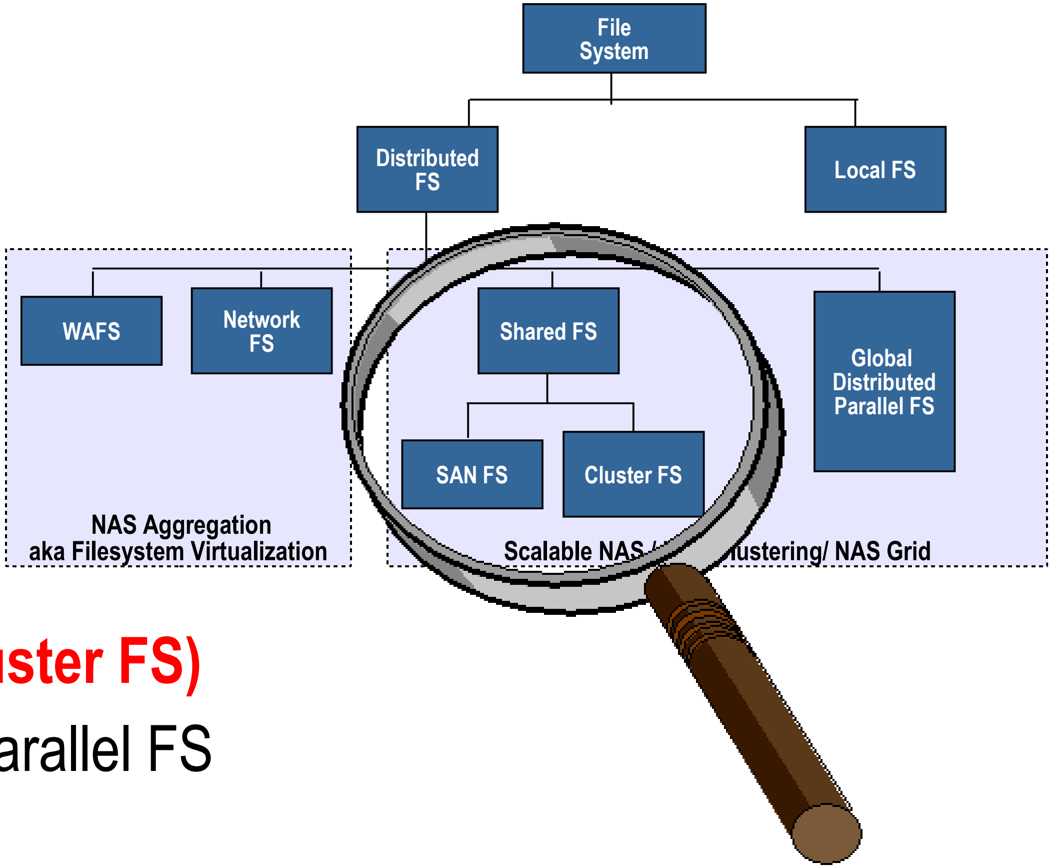
LAN-like file access over WAN



* compression, byte level differentiation, deferred write, read-ahead, TPC/IP optimization, ...

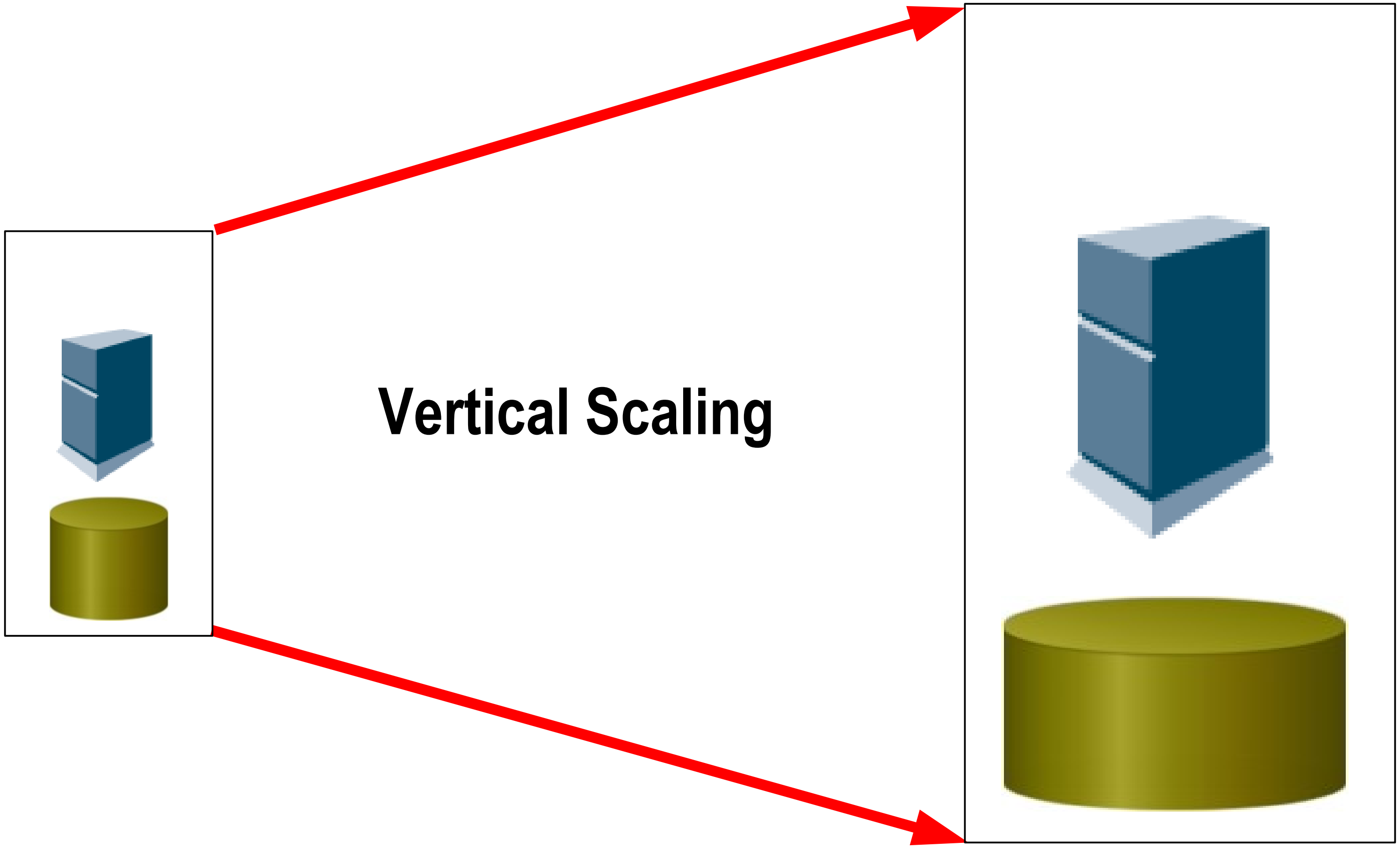
Agenda

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

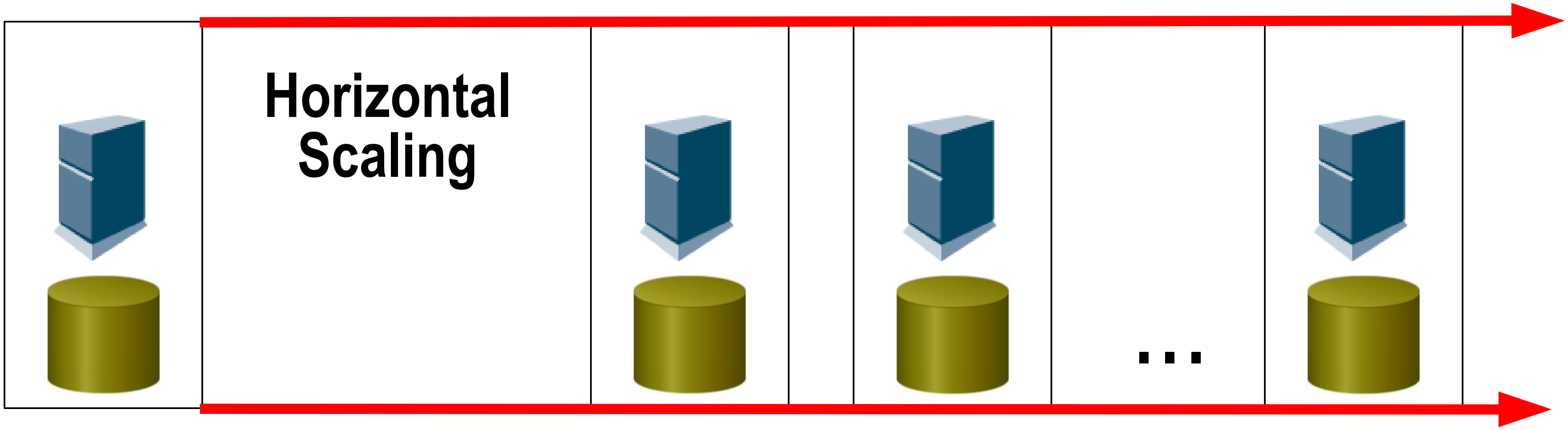


File System
Distributed FS
Local FS
WAFS
Network FS
Shared FS
SAN FS
Cluster FS
Global Distributed Parallel FS
NAS Aggregation aka Filesystem Virtualization
Scalable NAS / ...lustering/ NAS Grid

Scale-Up

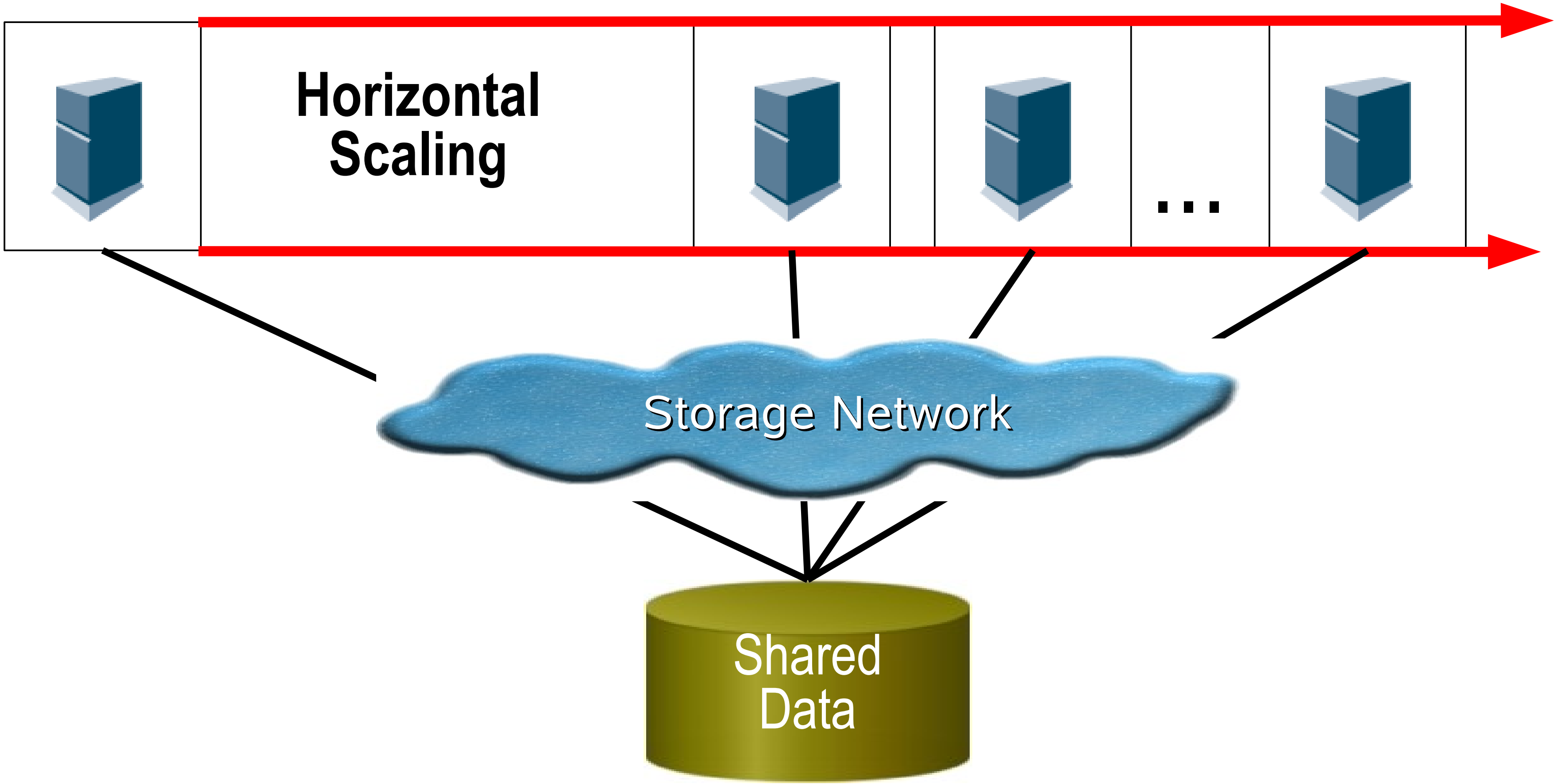


Scale-Out



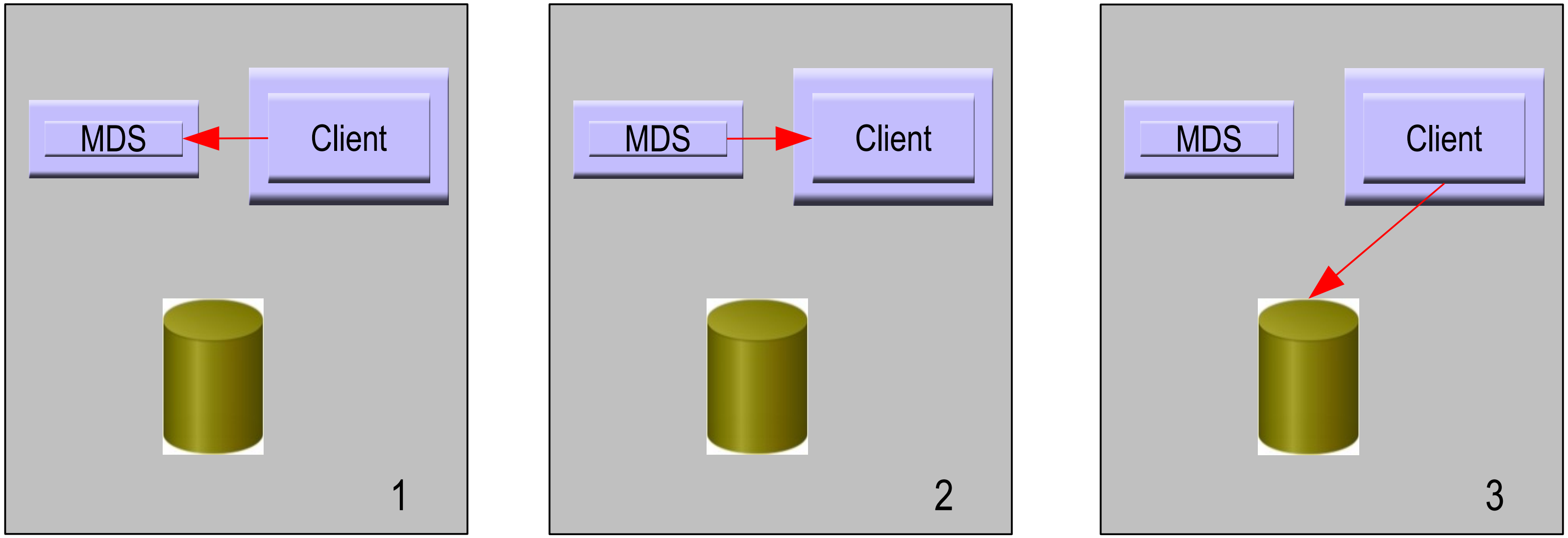
- Creating **islands** of data
- **Replication** of data

Scale-Out with Shared FS

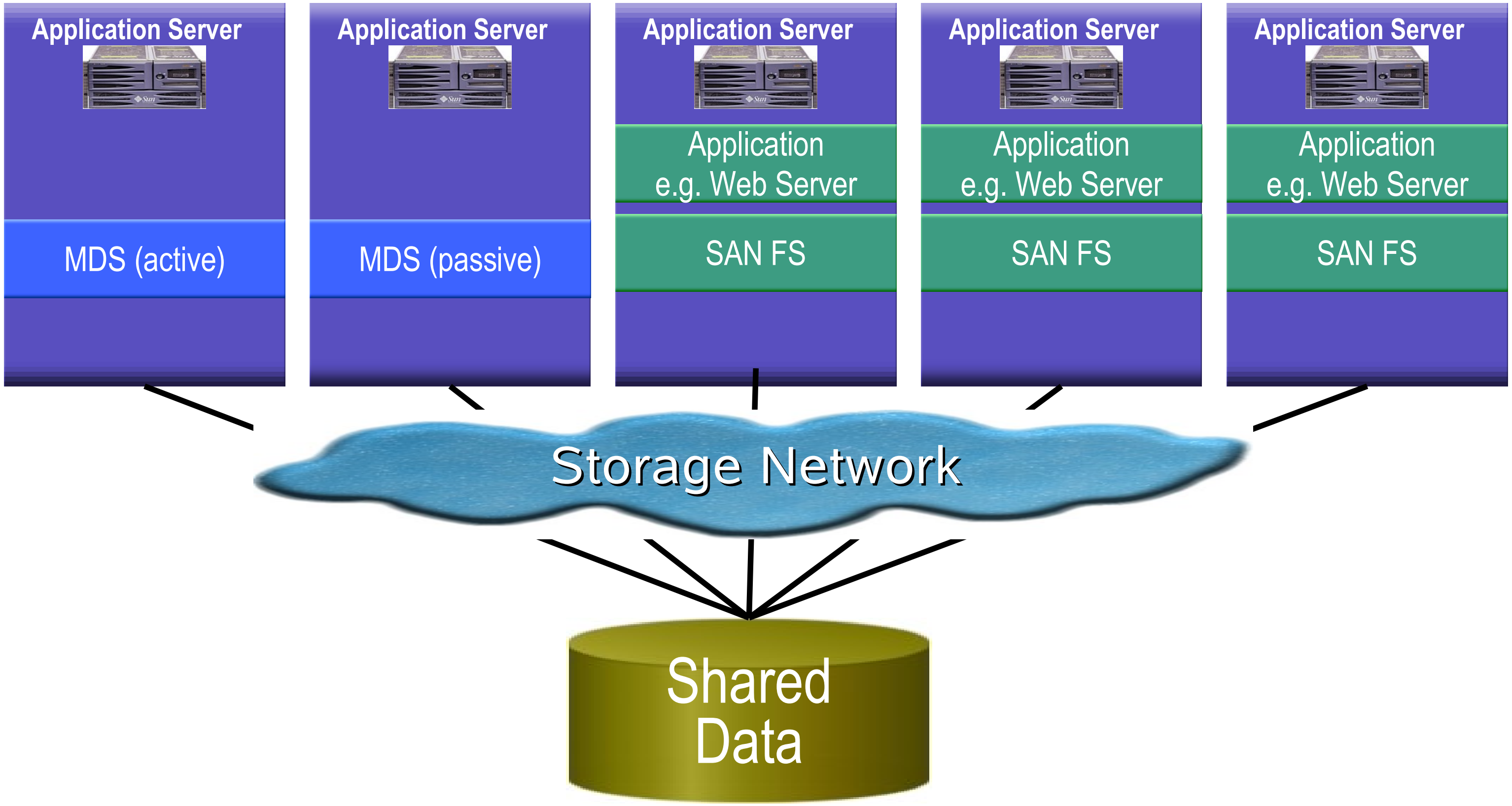


Shared FS & Metadata

- File access as a two-step transaction...

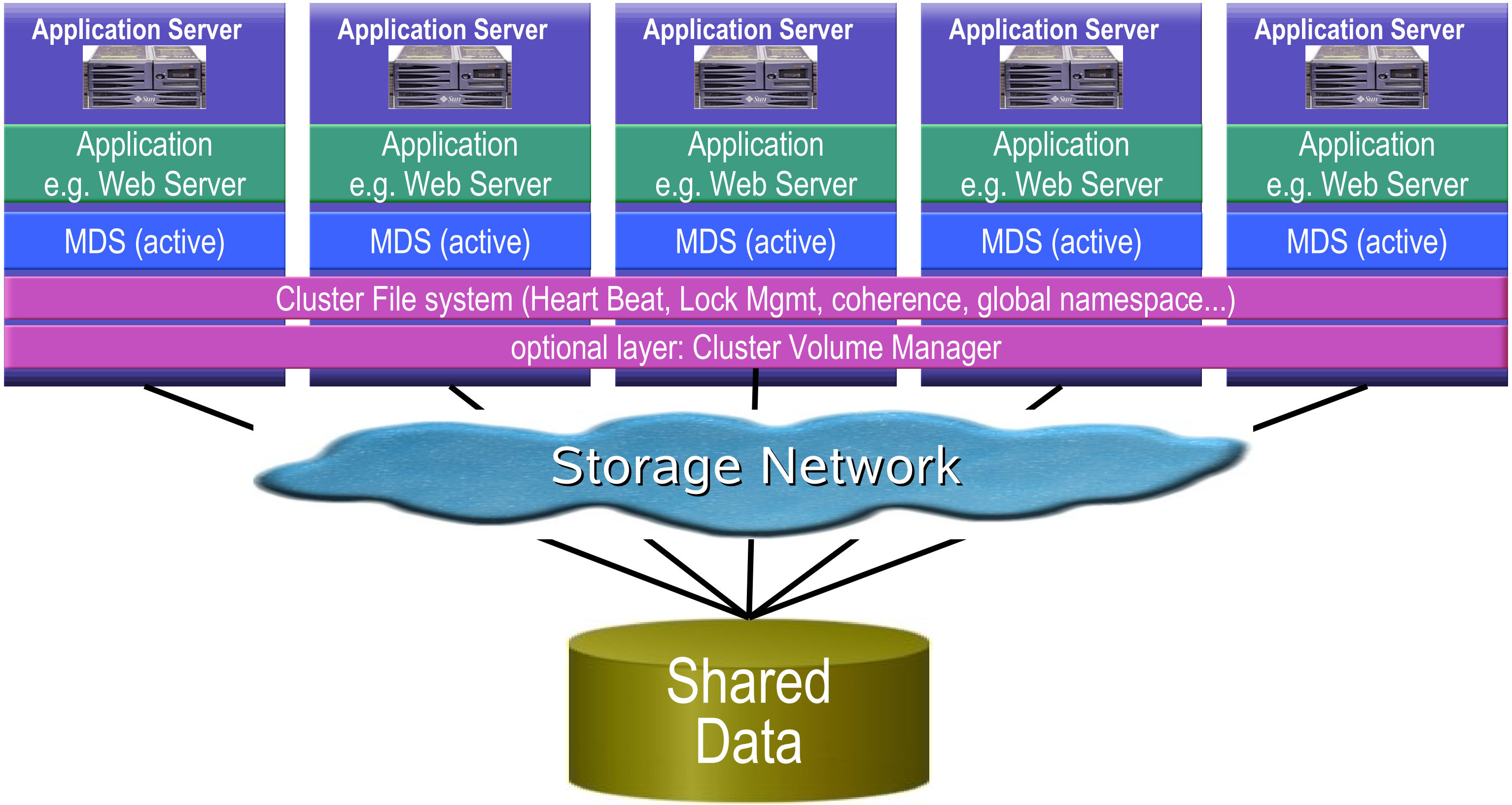


Shared FS – SAN FS



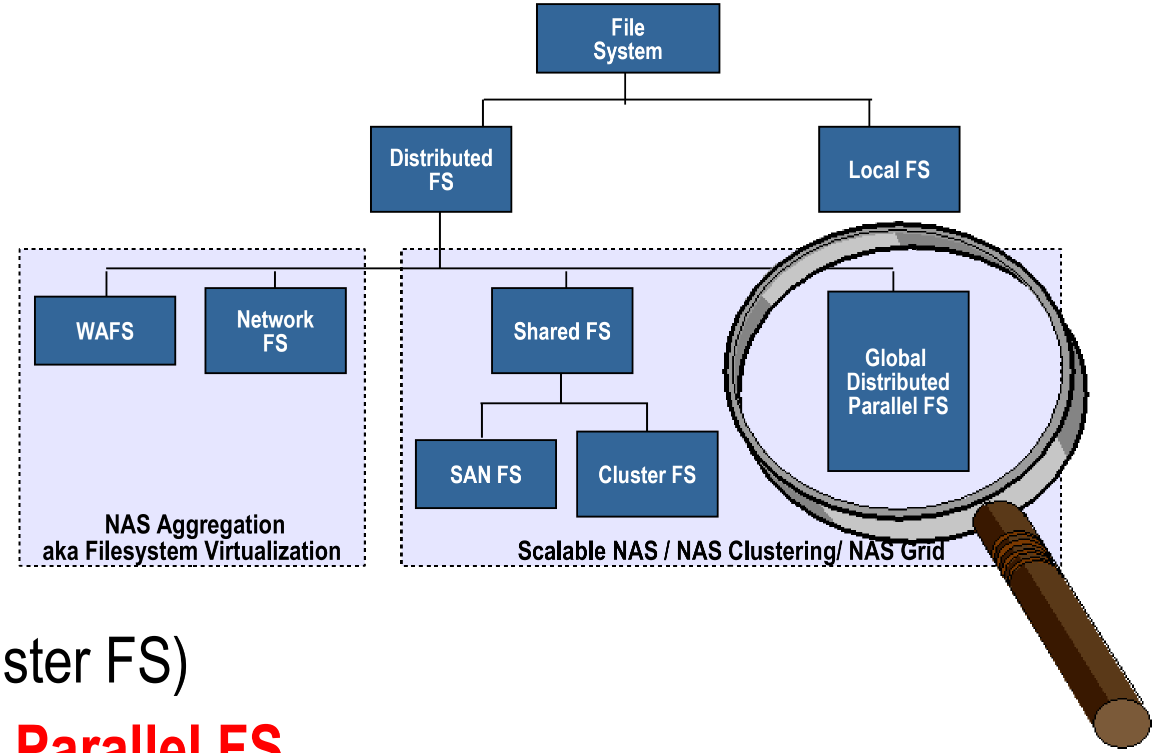
- MDS is part of each cluster node **master slave** (**asymmetric**)
- **Heterogeneous with unlimited number of nodes**
- **unlimited distance** between nodes

Shared FS – Cluster FS



- MDS is part of each cluster node **peer-to-peer** (**symmetric**)
- **Homogenous with limited number of nodes**
- **Limited distance** between nodes

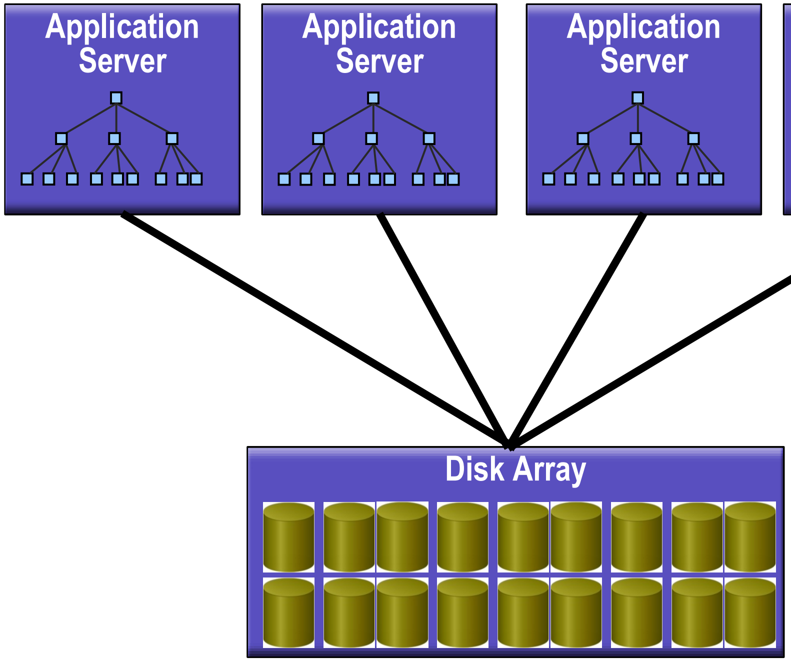
Agenda


File System
Distributed FS
Local FS
WAFS
Network FS
Shared FS
Global Distributed Parallel FS
SAN FS
Cluster FS
NAS Aggregation aka Filesystem Virtualization
Scalable NAS / NAS Clustering/ NAS Grid

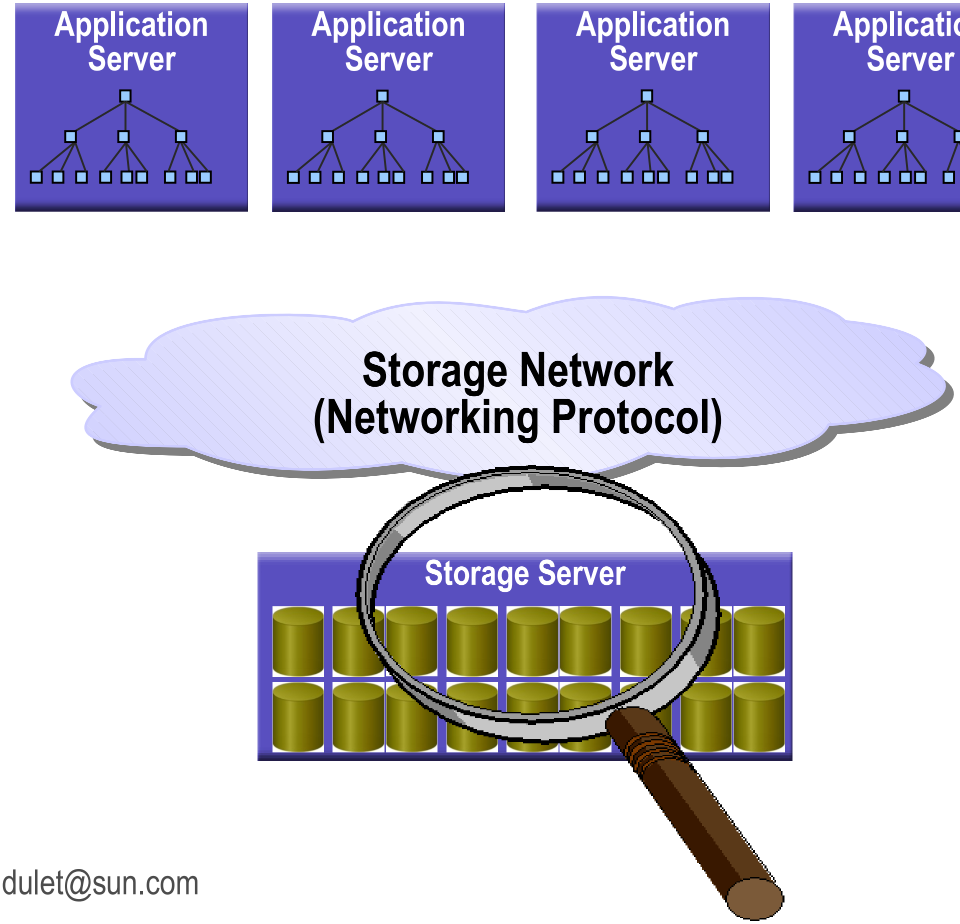
- File System Basics
- File Systems Taxonomy
- Local FS
- Distributed FS
- Wide Area FS
- Shared FS (SAN FS, Cluster FS)
- **Global, Distributed and Parallel FS**
- File System Virtualization
- Scalable NAS
- NAS Cluster / NAS Grid

Global FS (~ Shared FS)

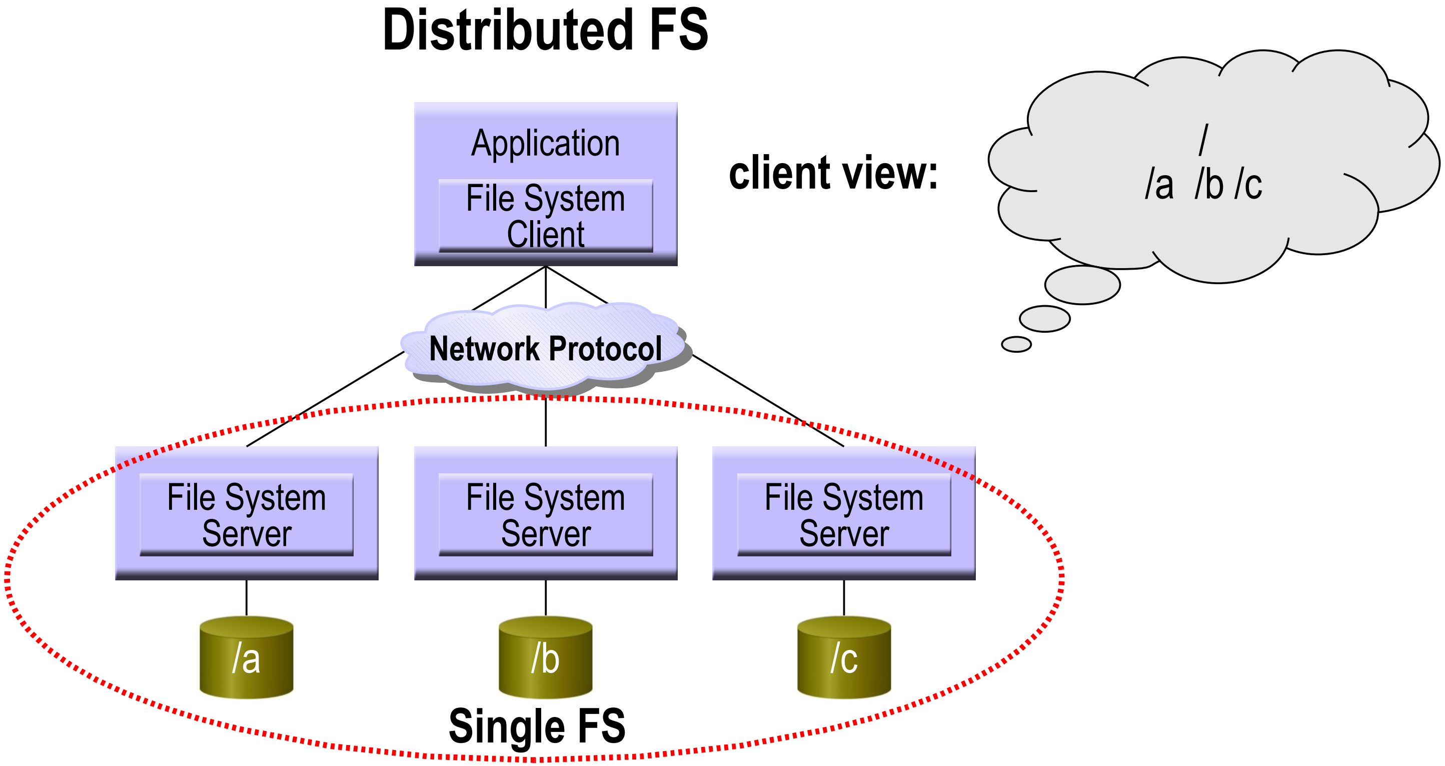
Data Sharing



Global & Network File System

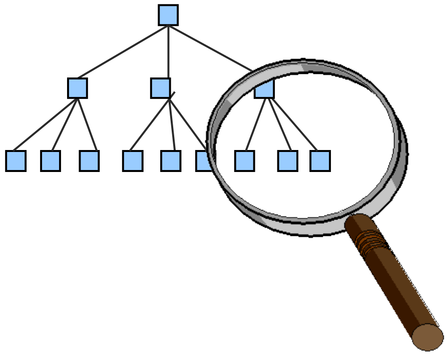


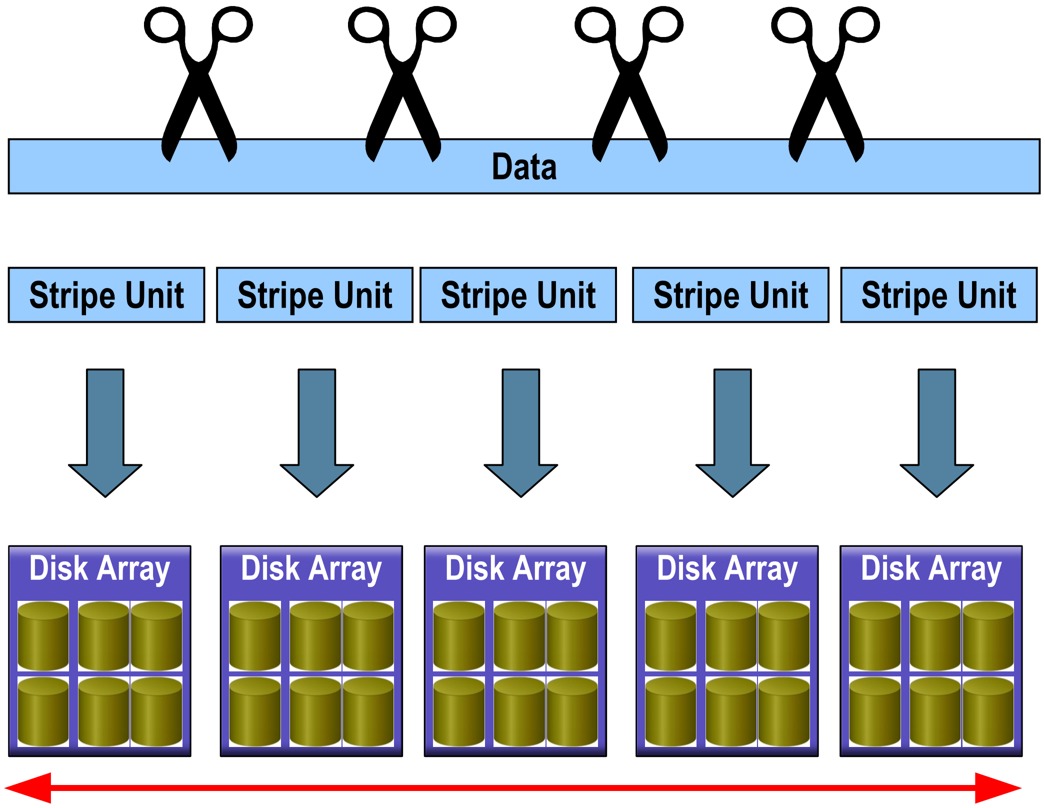
Distributed File System (DFS)



- **Files** are distributed across file servers

Parallel Data Access – RAID 0,5





Data

Stripe Unit

Stripe Unit

Stripe Unit

Stripe Unit

Stripe Unit

Disk Array

Disk Array

Disk Array

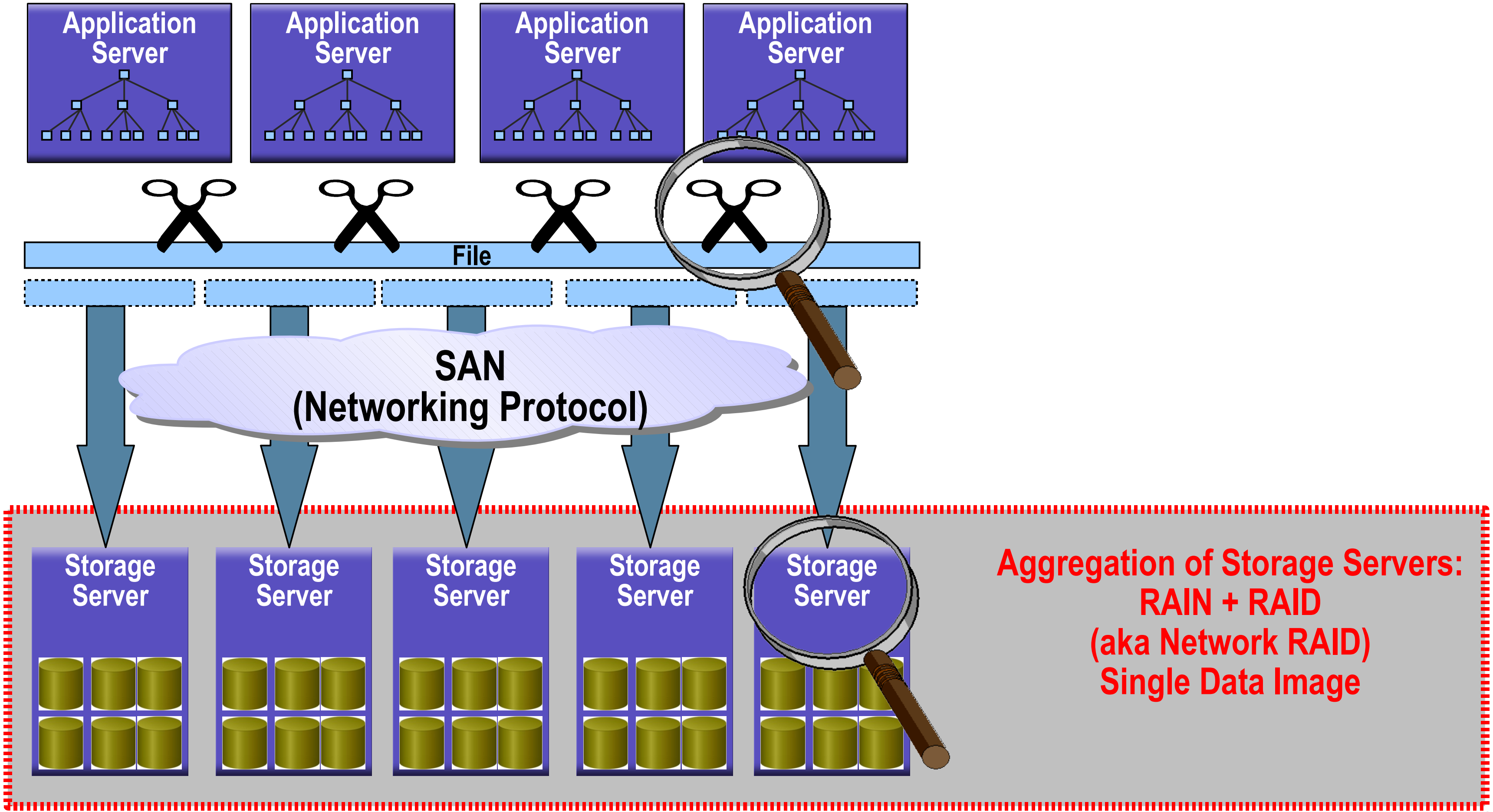
Disk Array

Disk Array

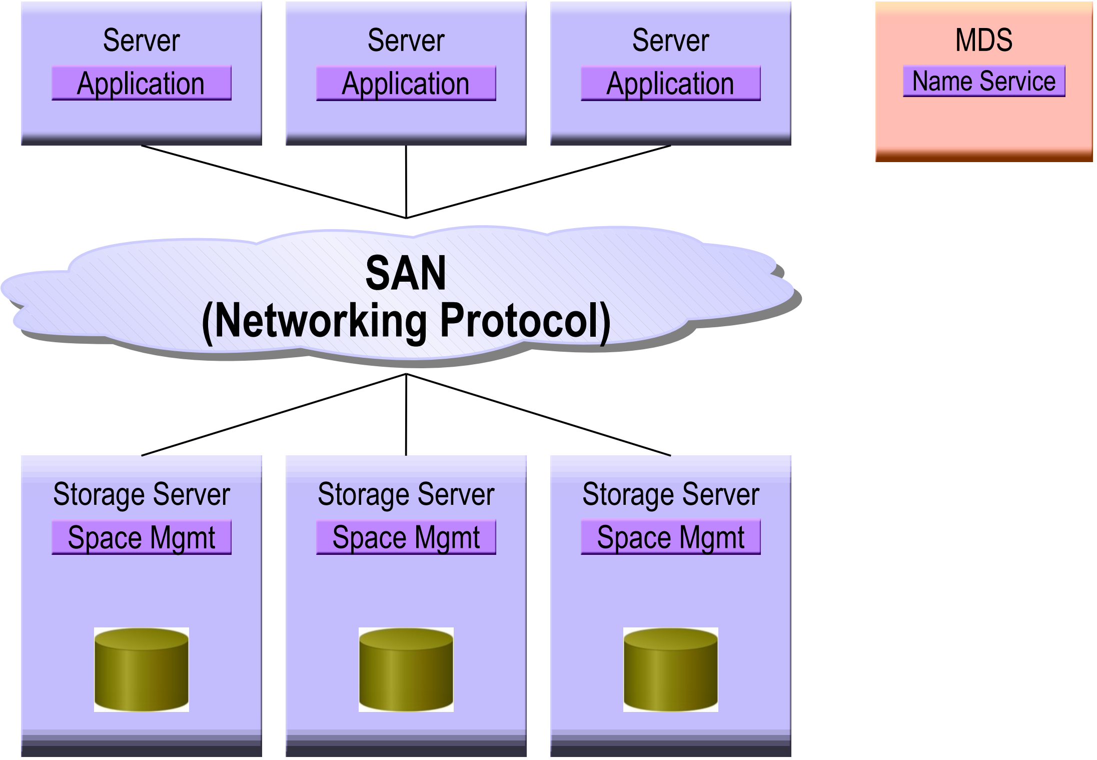
Data segments are striped across storage devices

Global,Distributed & Parallel File System

File Segments distributed across storage nodes



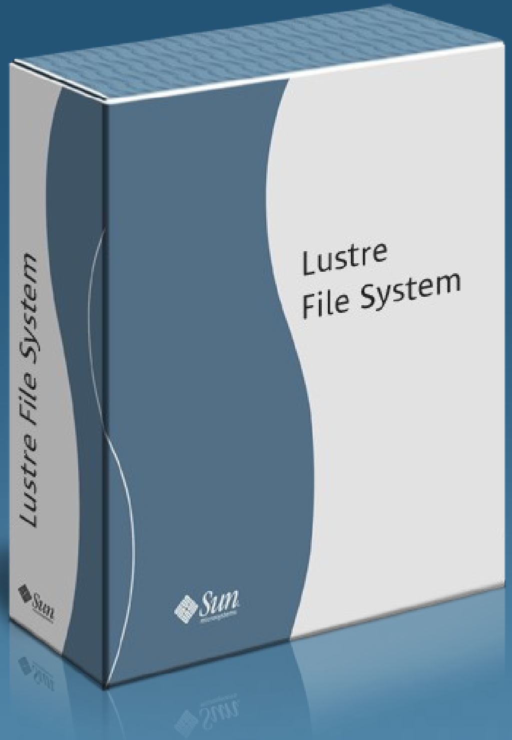
Global Distributed Parallel FS



Lustre™ Cluster File System

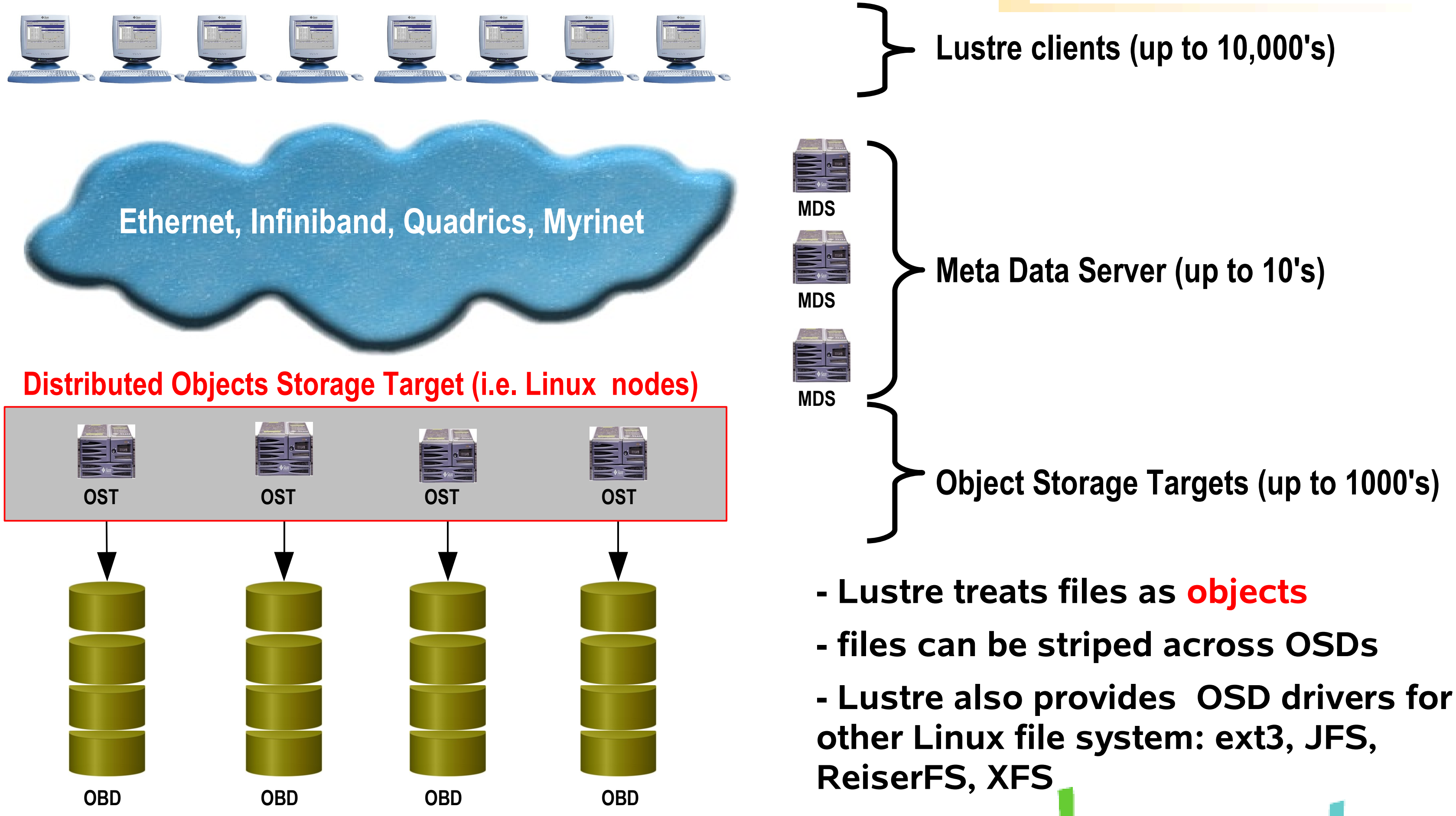
World’s Largest Network-Neutral Data Storage and Retrieval System

- The worlds most scalable parallel filesystem
- 10,000’s of clients
- Proven technology at major HPC installations:
  - Tokyo Tech, TACC (Sun), LANL, LLNL, Sandia, PNNL, NCSA, etc.
- **70% of Top10 run Lustre**
- **50% of Top30 run Lustre**
- **15% of Top500 run Lustre**



Lustre Global, Distributed & Parallel FS

Cluster File Systems, Inc

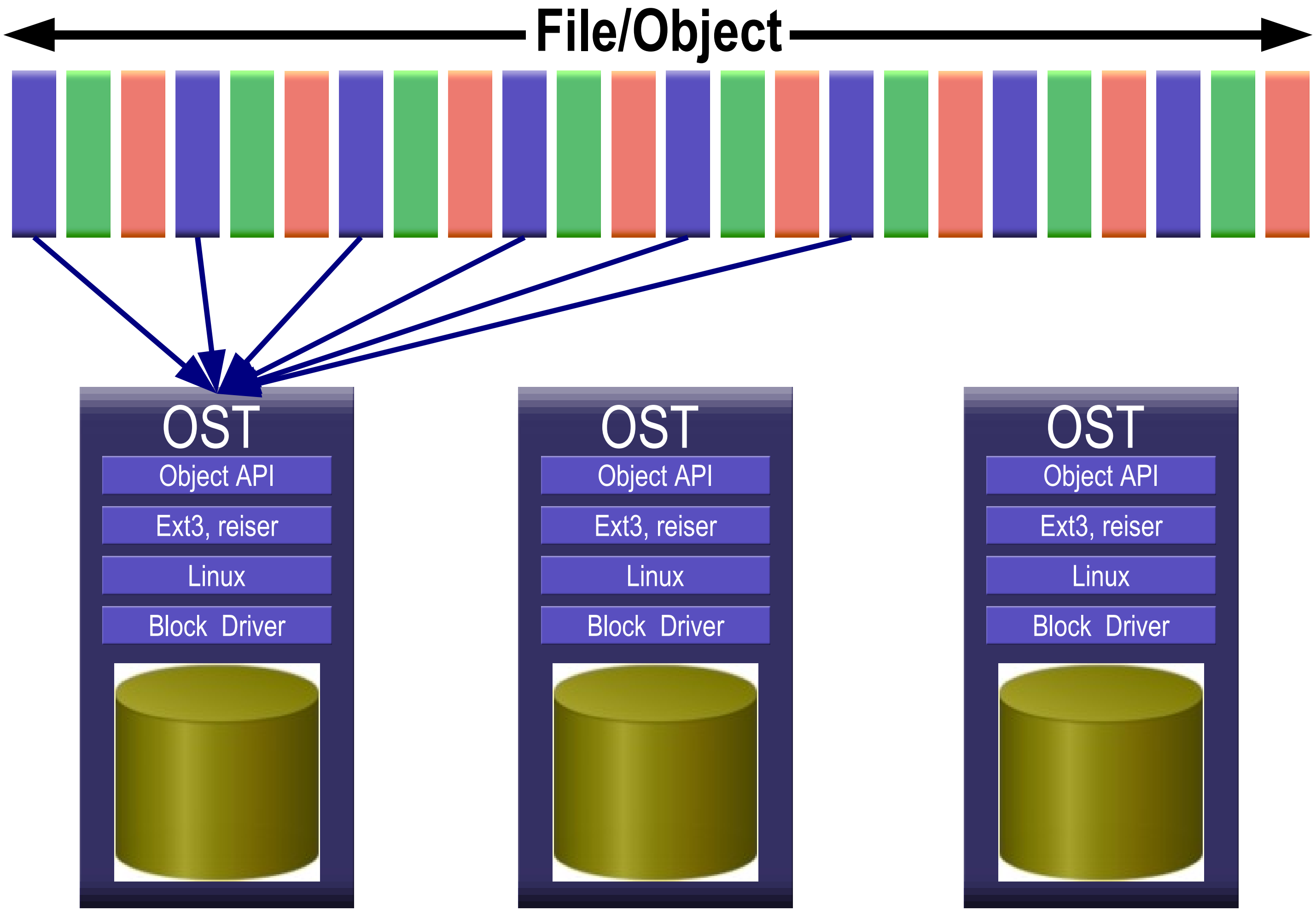


Lustre clients (up to 10,000's)
Ethernet, Infiniband, Quadrics, Myrinet
MDS
MDS
MDS
Meta Data Server (up to 10's)
Distributed Objects Storage Target (i.e. Linux nodes)
OST
OST
OST
OST
Object Storage Targets (up to 1000's)
OBD
OBD
OBD
OBD

- Lustre treats files as **objects**
- files can be striped across OSDs
- Lustre also provides OSD drivers for other Linux file system: ext3, JFS, ReiserFS, XFS

lustre

Lustre: File/Object striped across 3 OST's



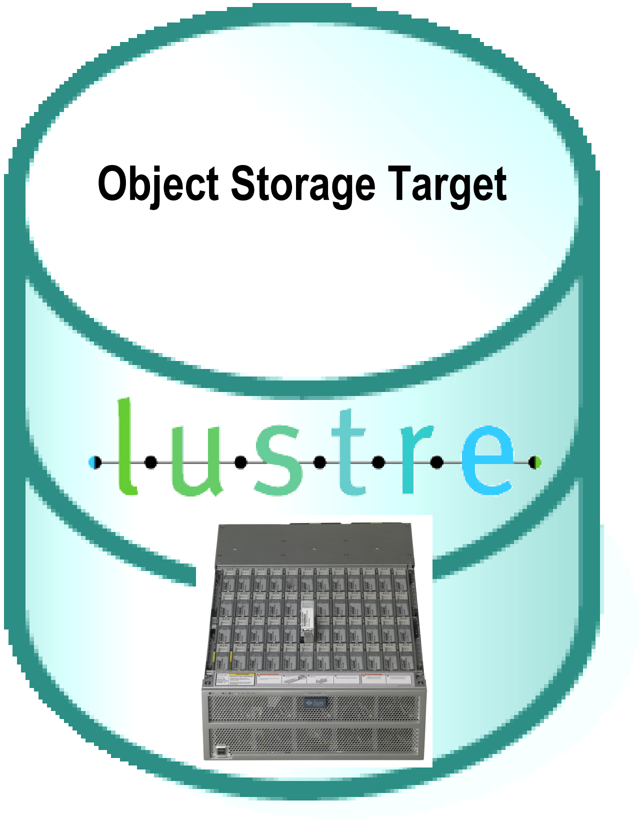
Lustre & Thumper

Object-Based
Cluster File system Target

+

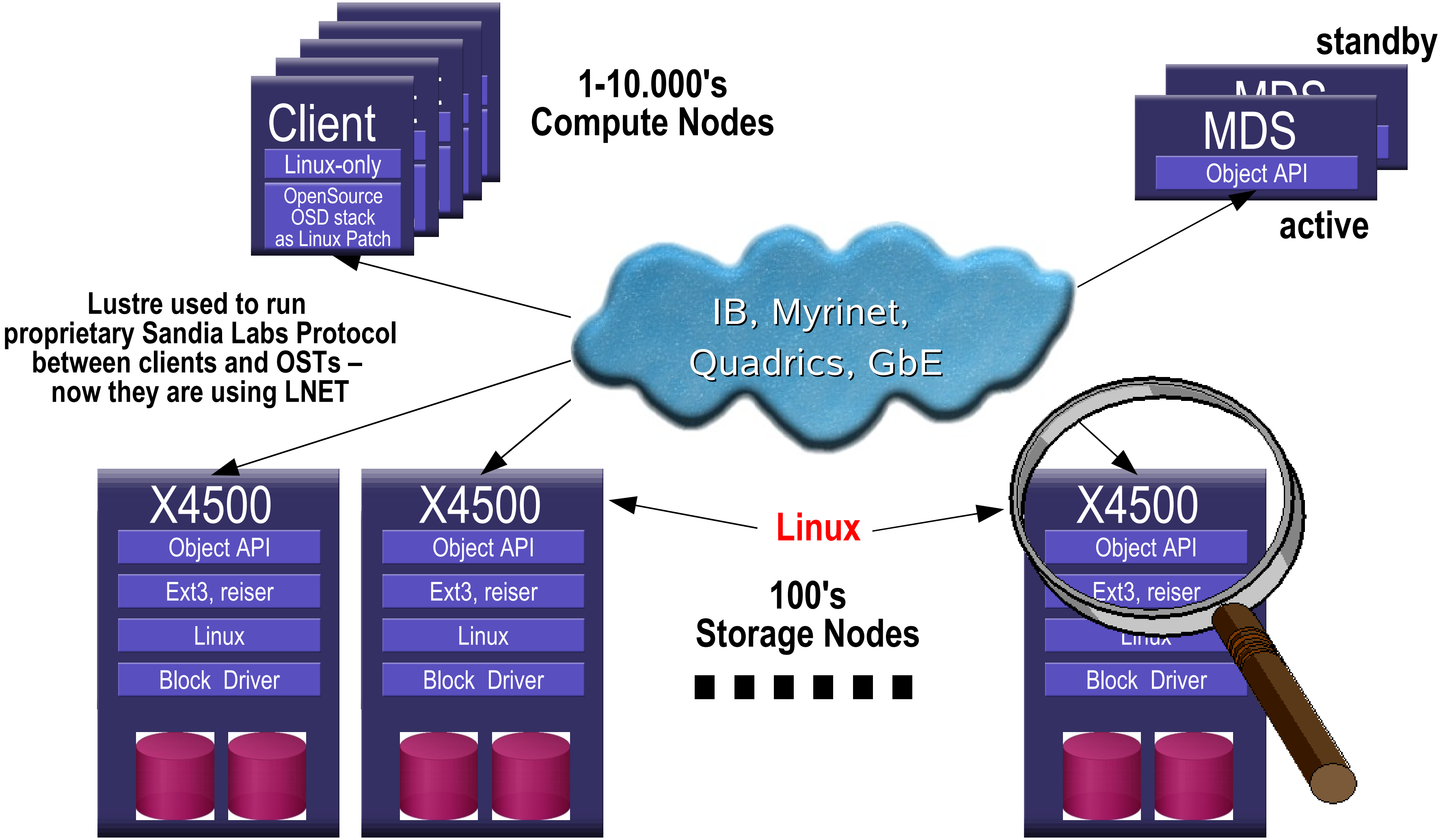
X4500 – aka Thumper
24TB on 4U

=



Object Storage Target

TITech & TACC



Client

Linux-only

OpenSource OSD stack as Linux Patch

1-10.000's Compute Nodes

standby

MDS

MDS

Object API

active

IB, Myrinet, Quadrics, GbE

Lustre used to run proprietary Sandia Labs Protocol between clients and OSTs – now they are using LNET

X4500

Object API

Ext3, reiser

Linux

Block Driver

X4500

Object API

Ext3, reiser

Linux

Block Driver

Linux

100's Storage Nodes

X4500

Object API

Ext3, reiser

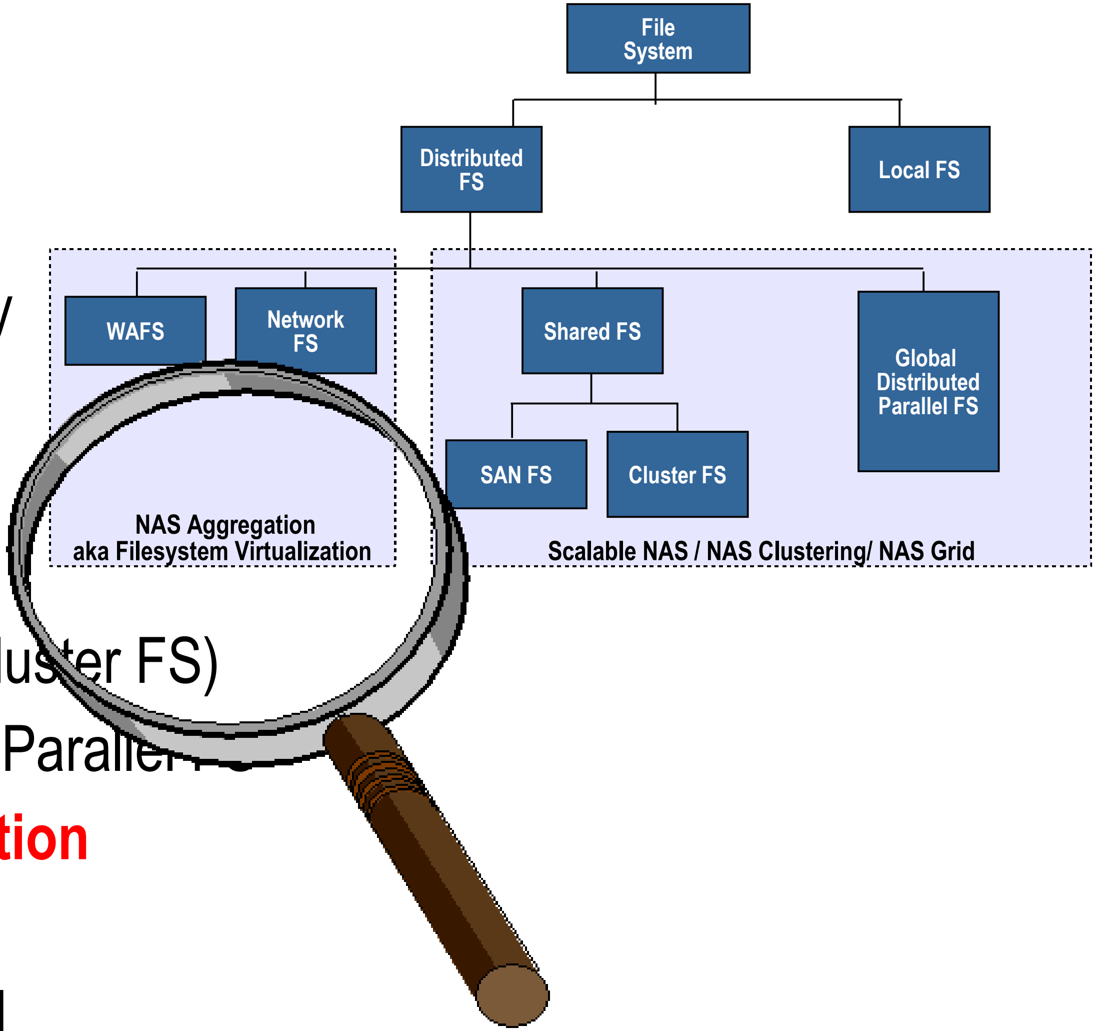
Linux

Block Driver

Heterogeneous Block Storage Devices

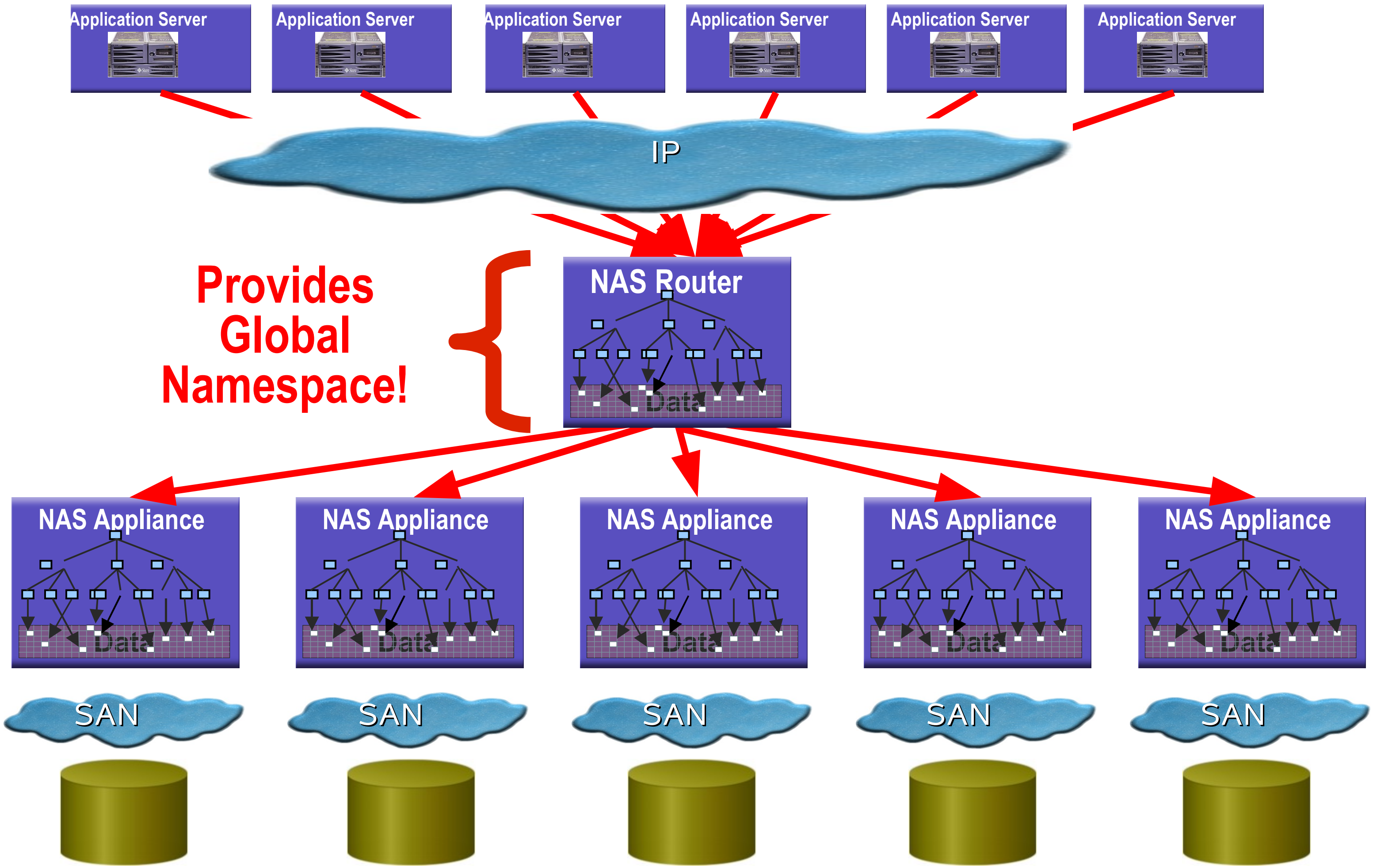
Agenda

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


File System
Distributed FS
Local FS
WAFS
Network FS
Shared FS
Global Distributed Parallel FS
SAN FS
Cluster FS
NAS Aggregation
aka Filesystem Virtualization
Scalable NAS / NAS Clustering/ NAS Grid

FS Virtualization – NAS Aggregation

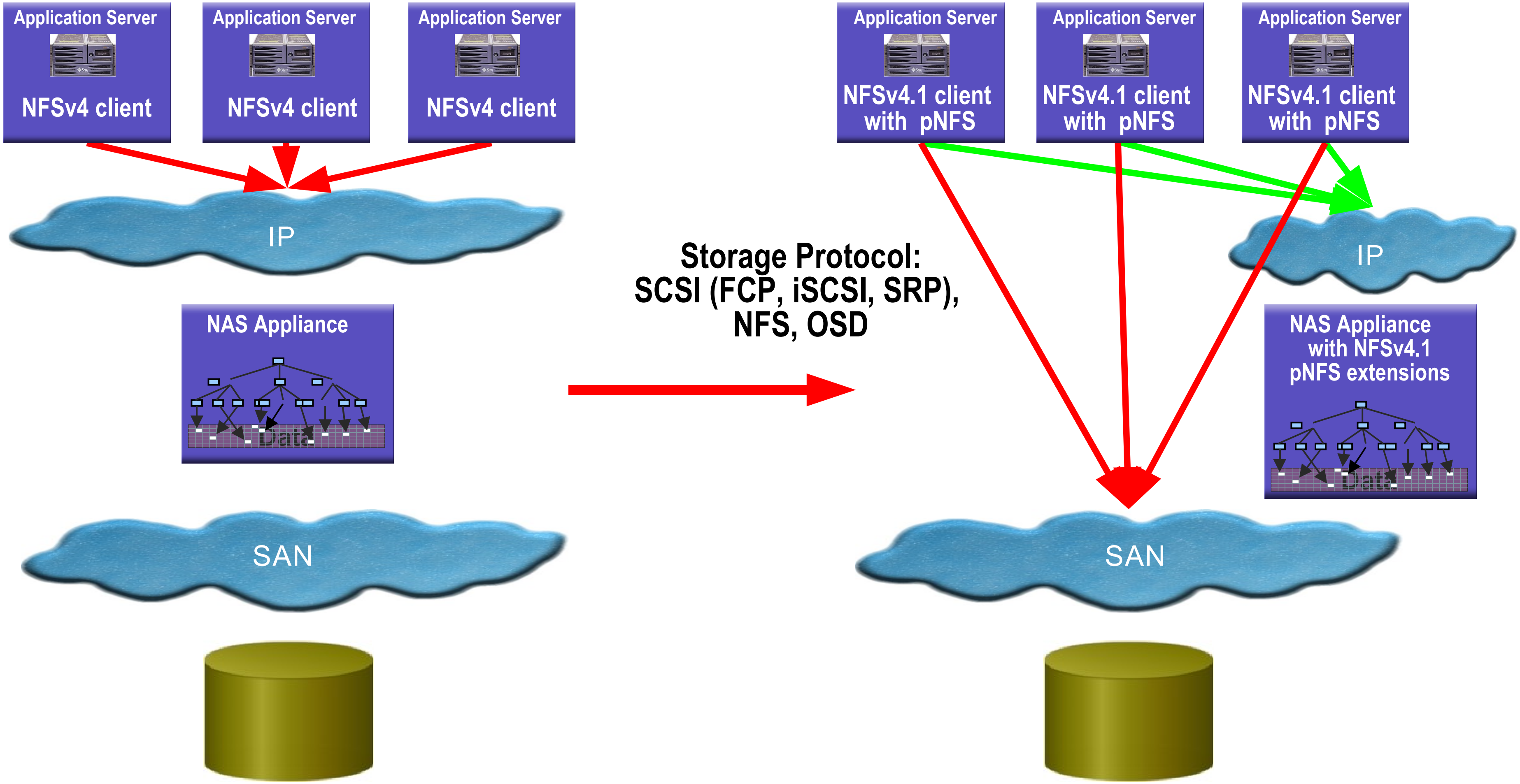
In-Band



FS Virtualization – NFS4.1 pNFS

In-Band NAS:

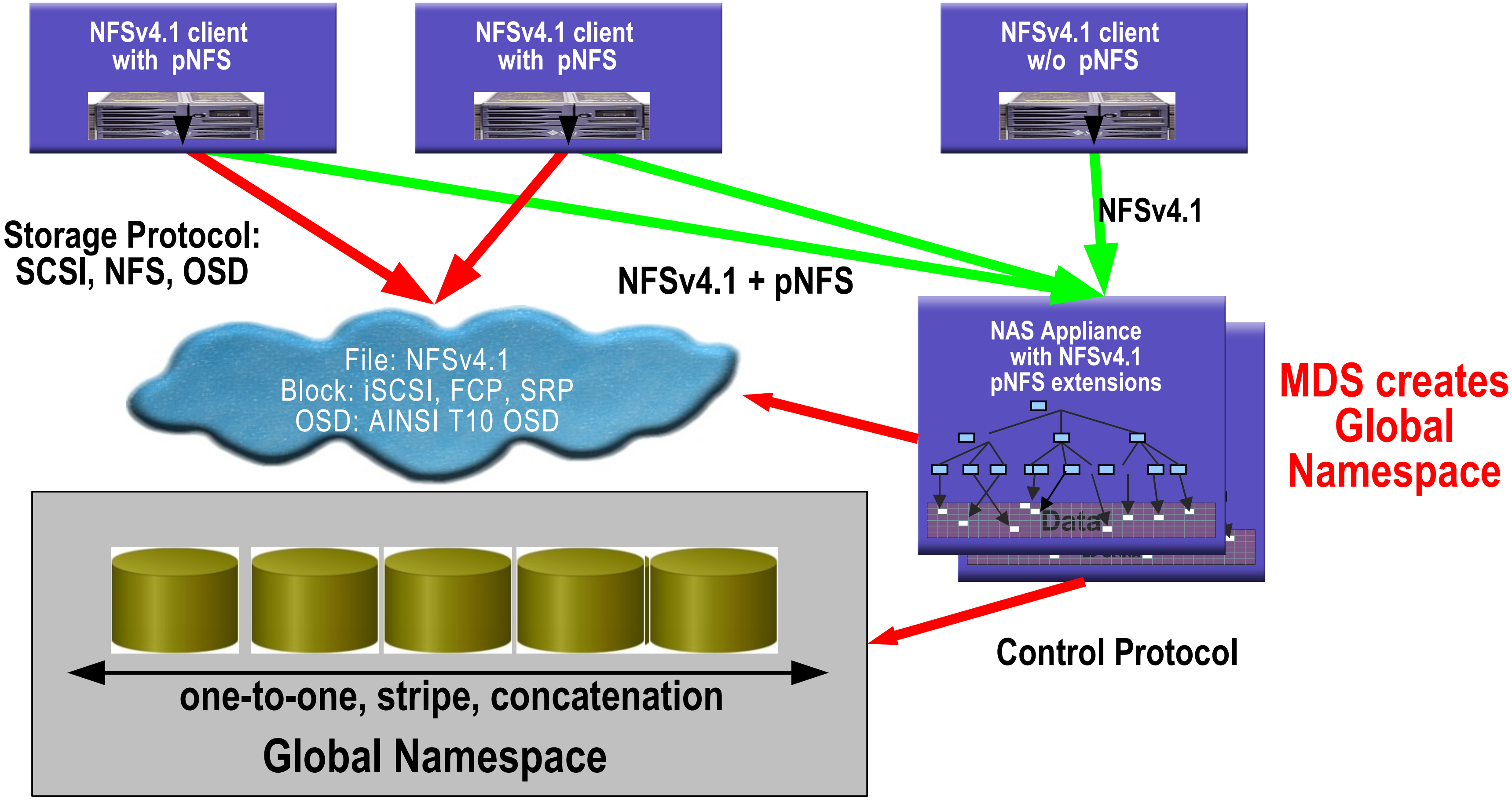
Out-of-Band NAS:



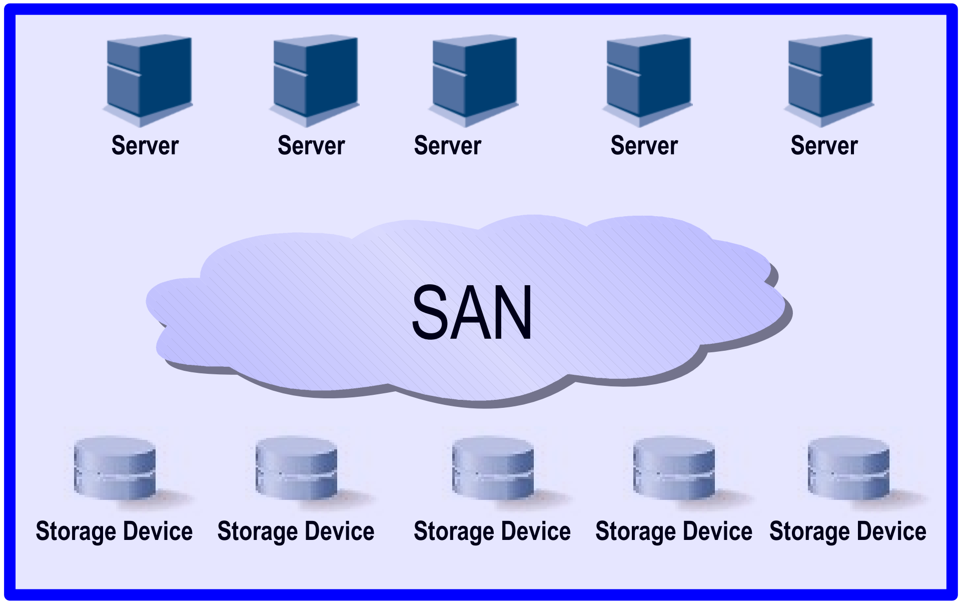
Storage Protocol:
SCSI (FCP, iSCSI, SRP),
NFS, OSD

FS Virtualization – NFSv4.1 pNFS

Out-of-Band

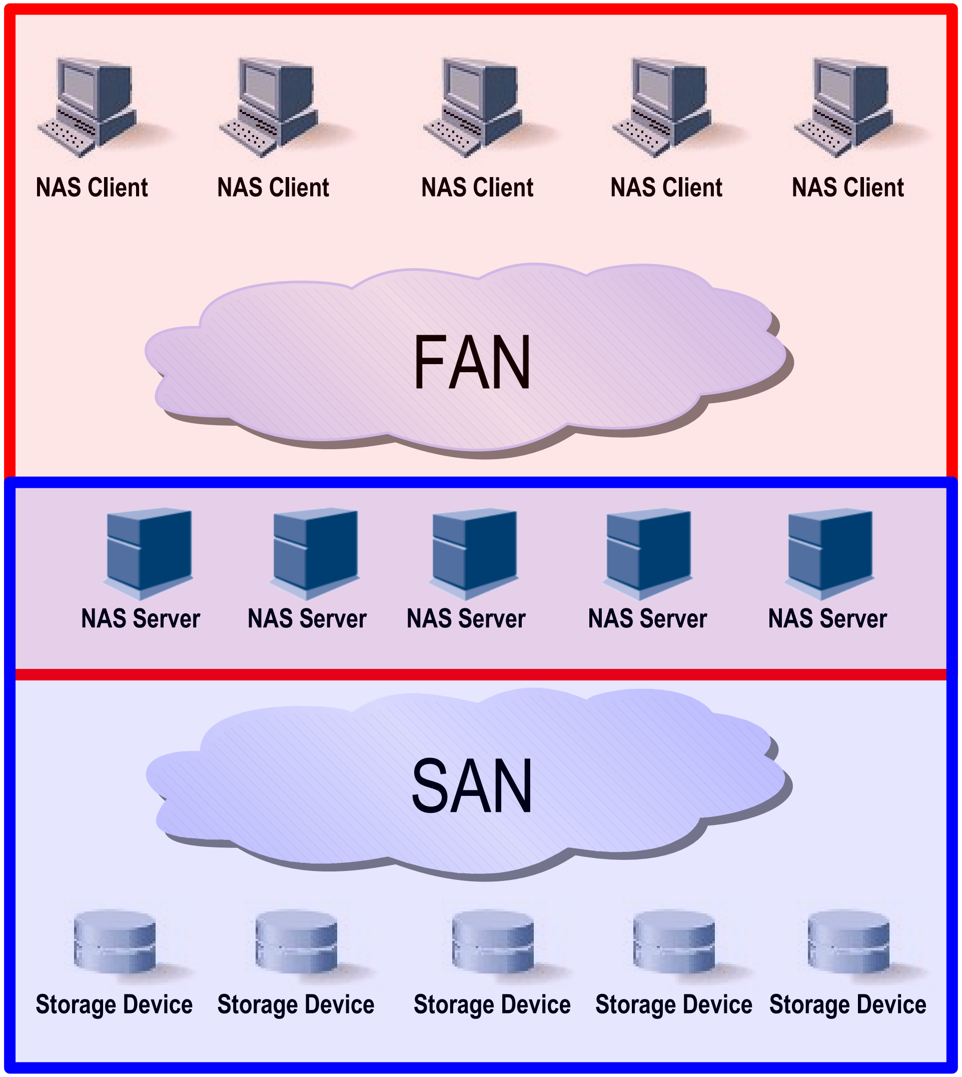


FS Virtualization – File Area Network



Global Address Space
(WWN, 24-bit fabric addresses,
nameserver),
zoning, routing, ...

FS Virtualization – File Area Network

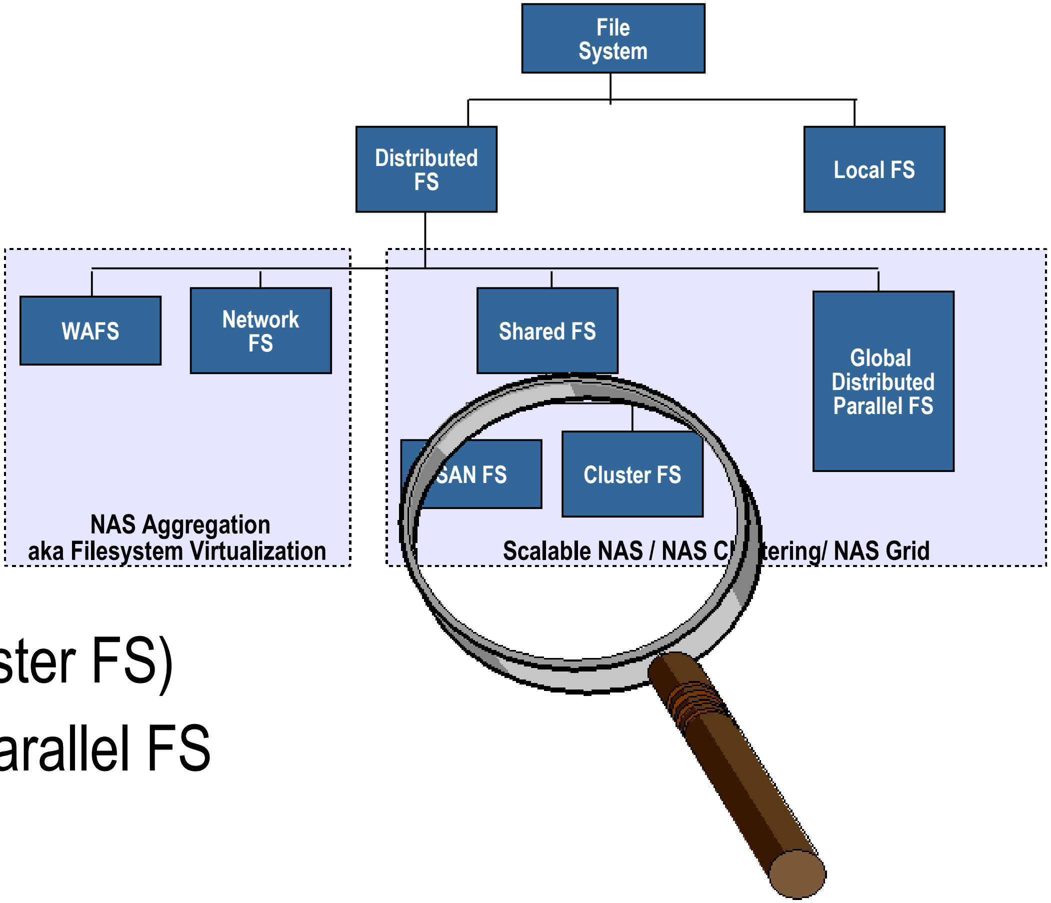


Global Namespace,
load-balancing, network compression (WAFS), data protection (security, replication, ...), SLA/ILM (migration, retention), ...

Global Address Space
(WWN, 24-bit fabric address, nameserver),
zoning, routing, ...

Agenda

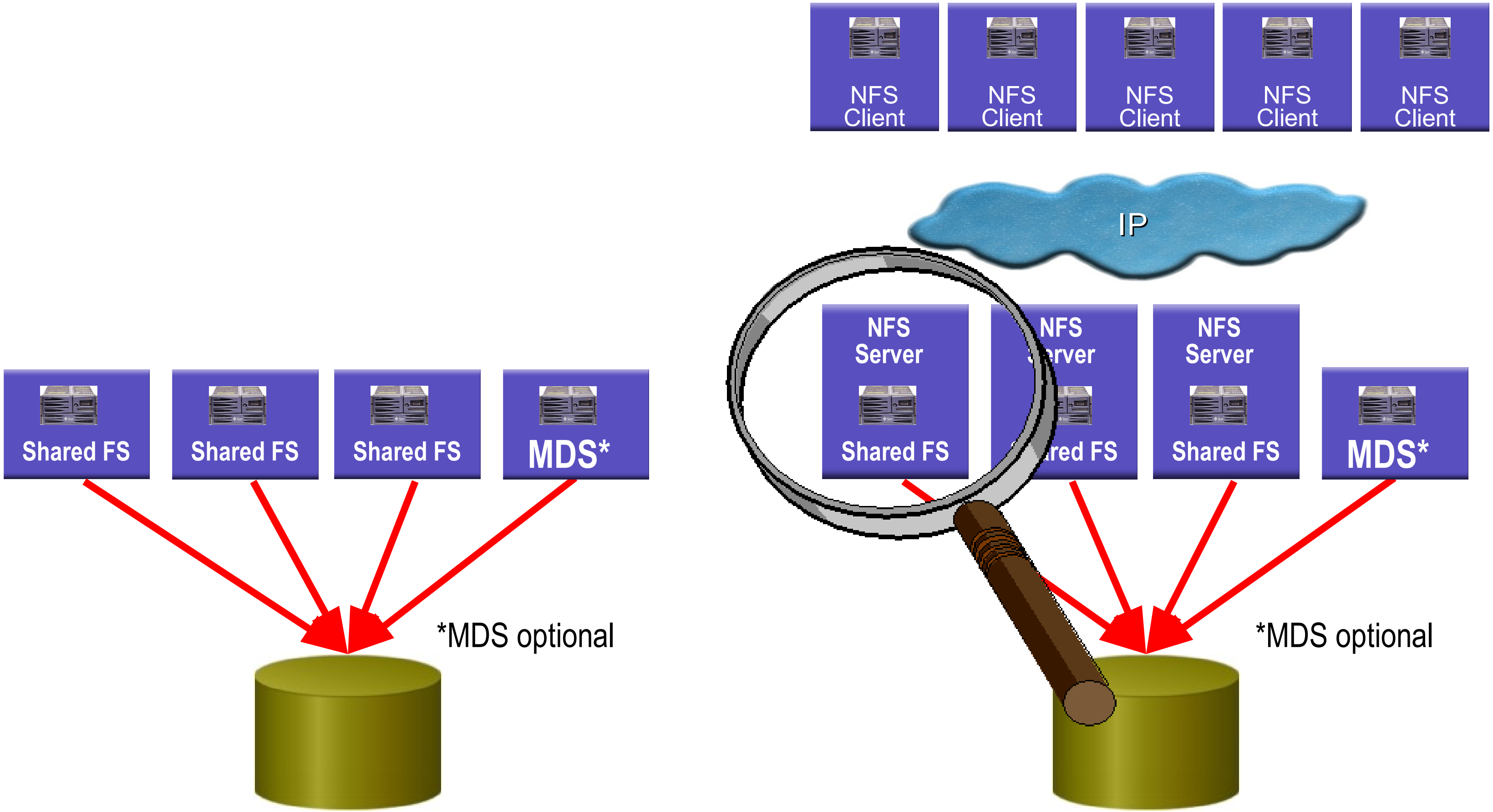
- File System Basics
- File Systems Taxonomy
- Local FS
- Distributed FS
- Wide Area FS
- Shared FS (SAN FS, Cluster FS)
- Global, Distributed and Parallel FS
- File System Virtualization
- **Scalable NAS**
- NAS Cluster / NAS Grid



File System
Distributed FS
Local FS
WAFS
Network FS
Shared FS
Global Distributed Parallel FS
SAN FS
Cluster FS
NAS Aggregation aka Filesystem Virtualization
Scalable NAS / NAS Cl tering/ NAS Grid

© Copyright: christian.bandulet@sun.com

Scalable NAS (NFS & Shared FS)

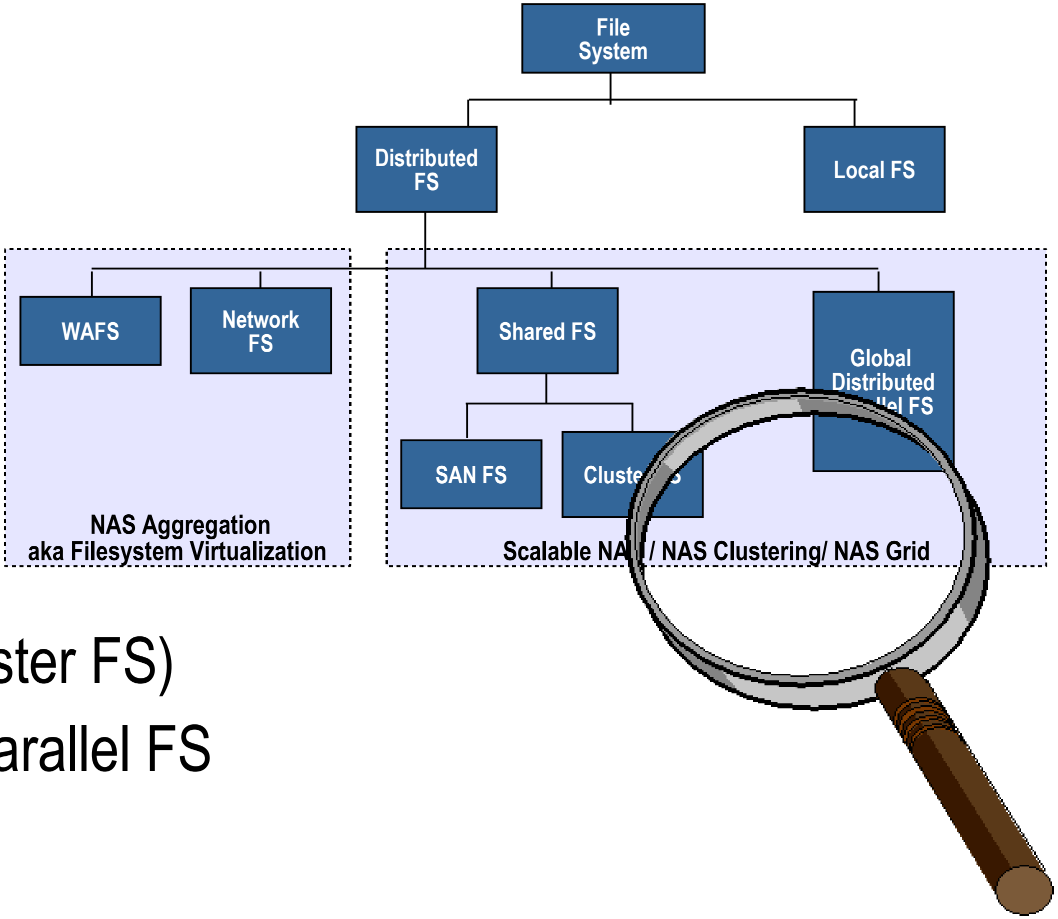


Shared FS with Shared Device

Scalable NFS with Shared FS

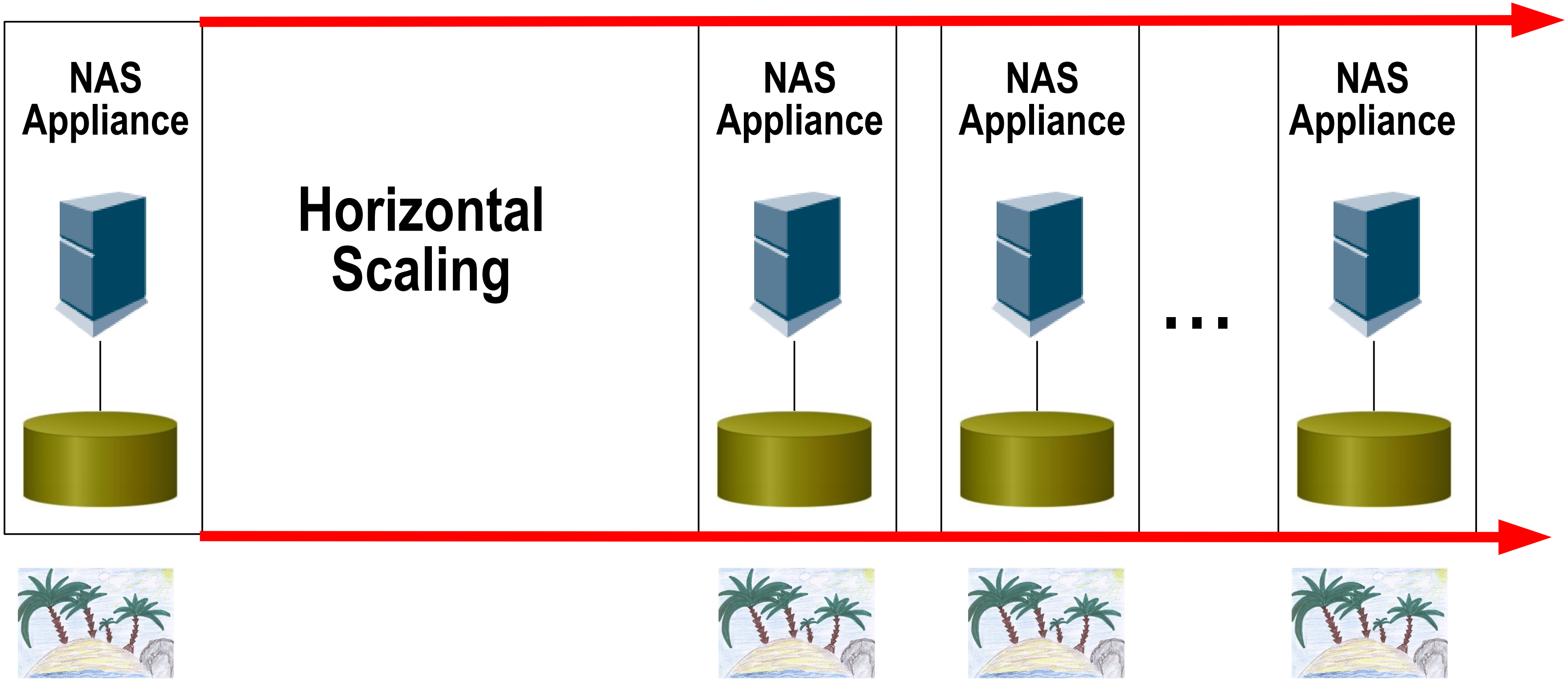
Agenda

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


File System
Distributed FS
Local FS
WAFS
Network FS
Shared FS
Global Distributed ...llel FS
SAN FS
Cluste... FS
NAS Aggregation aka Filesystem Virtualization
Scalable NA... / NAS Clustering/ NAS Grid

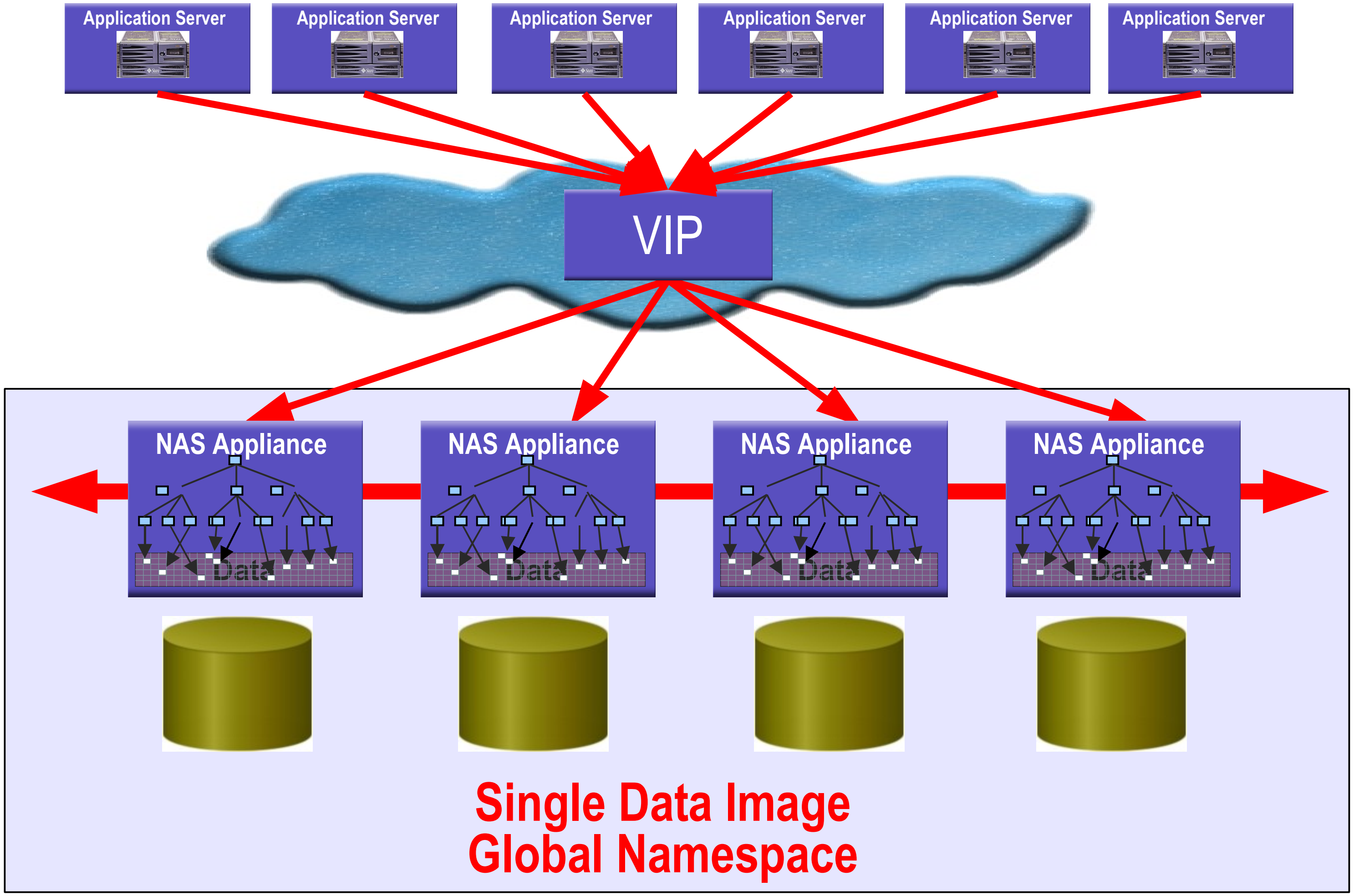
© Copyright: christian.bandulet@sun.com

NAS Scale-Out Problem Statement



- Creating **islands** of data
- **Replication** of data

NAS Cluster / NAS Grid



Single Data Image
Global Namespace

GUUG-Frühjahrsfachgespräch 2008

The File Systems Survey

Christian Bandulet
Principal Engineer
Data Management Ambassador
Sun Microsystems Inc. (Frankfurt, Germany)