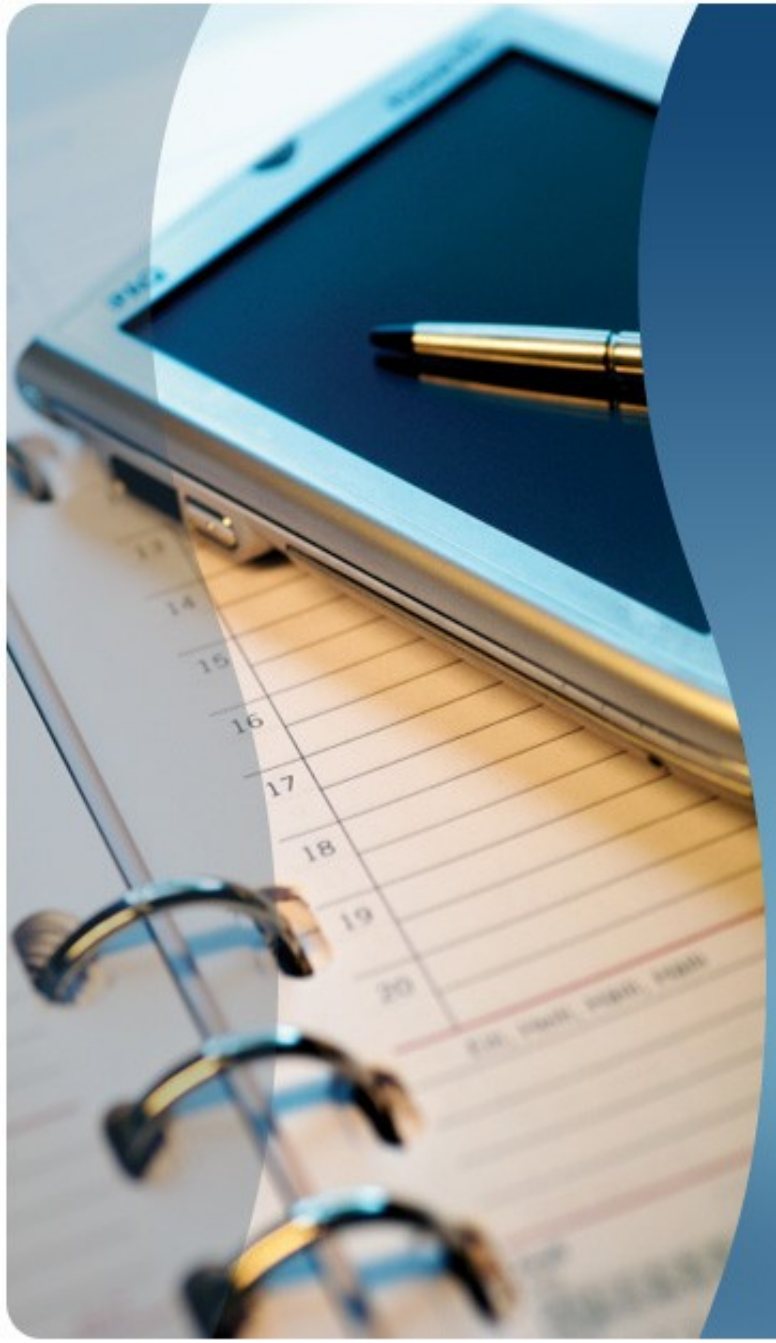
Parallele Dateisysteme für Linux und Solaris

Roland Rambau
Principal Engineer GSE
Sun Microsystems GmbH

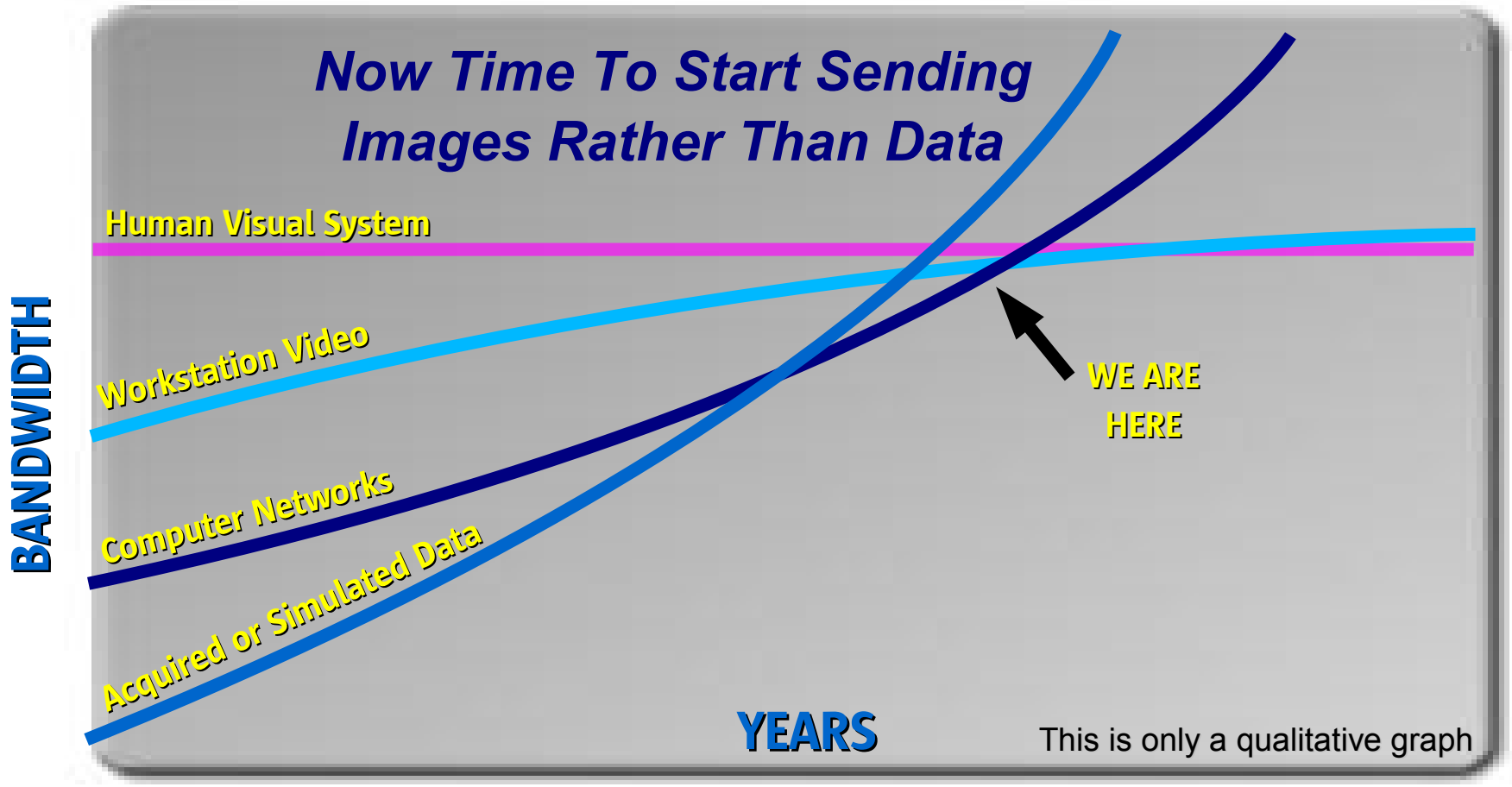


Agenda

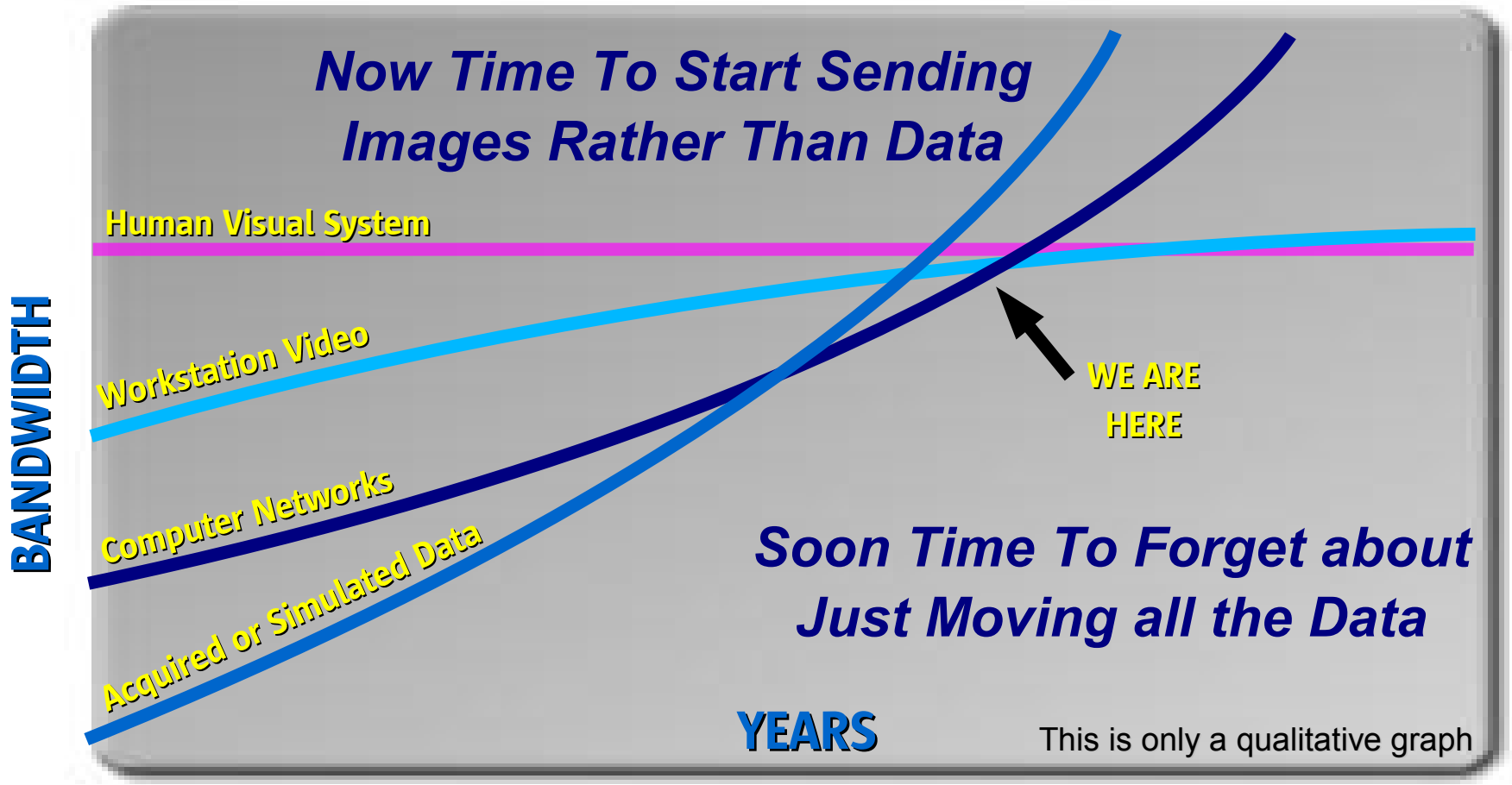
- kurze Einführung
- QFS
- Lustre
- pNFS

(Sorry ...)

Some Critical Qualitative Trends

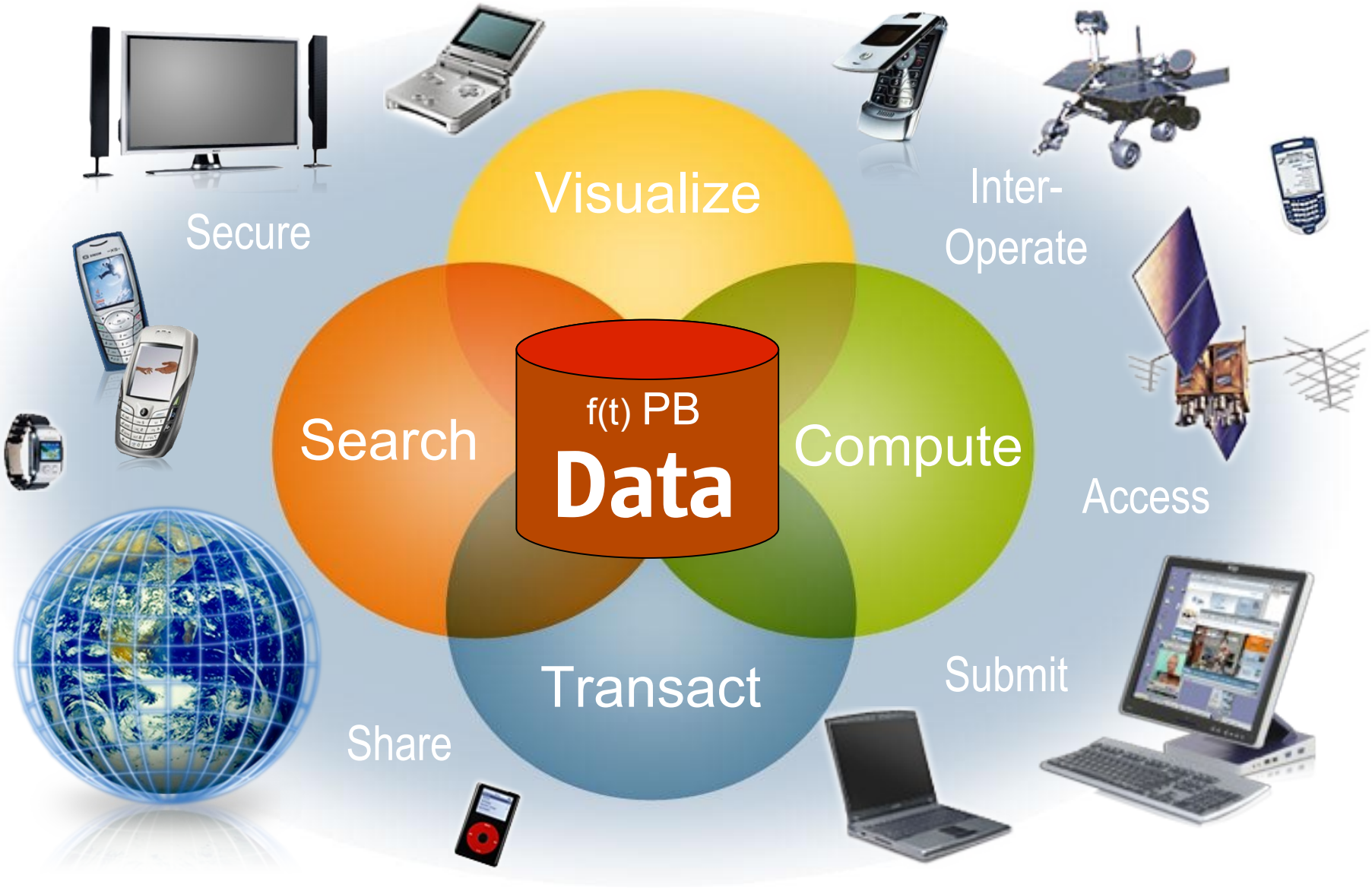


Some Critical Qualitative Trends

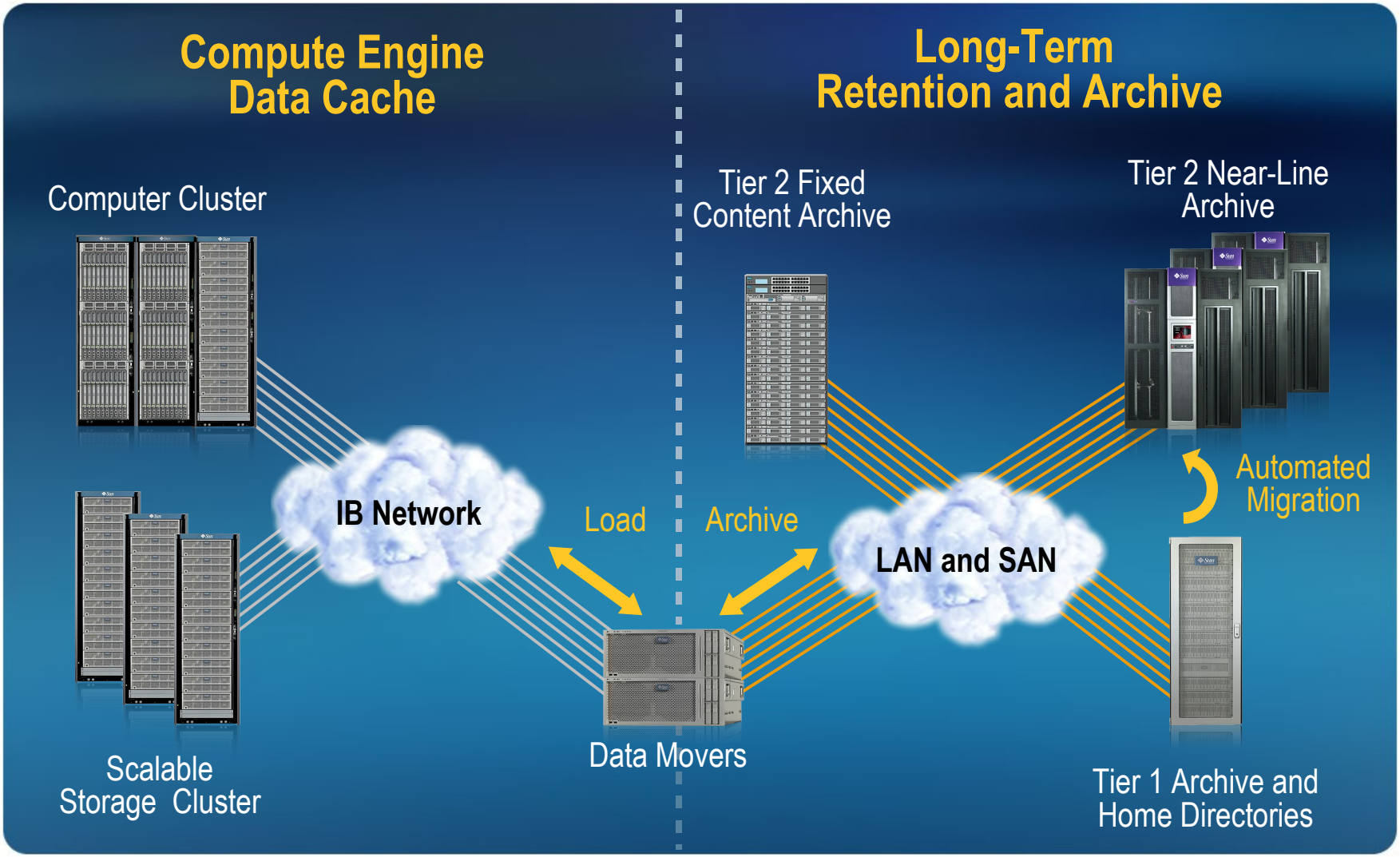


- InfiniBand DDR x12 is about **3.5 PB/week** per direction
  - GbE is about 3 PB/year

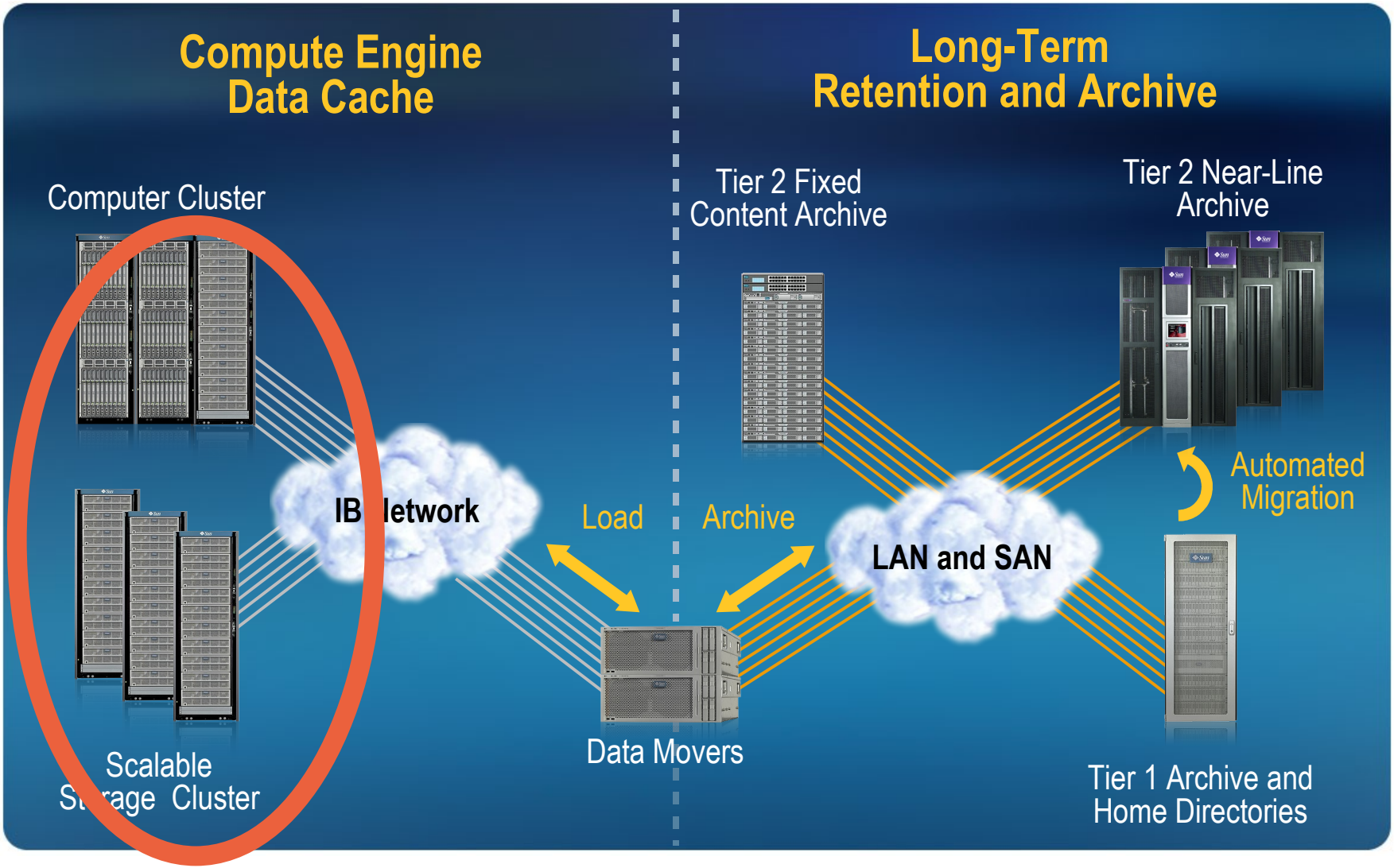
High Performance Computing Is Data Centric



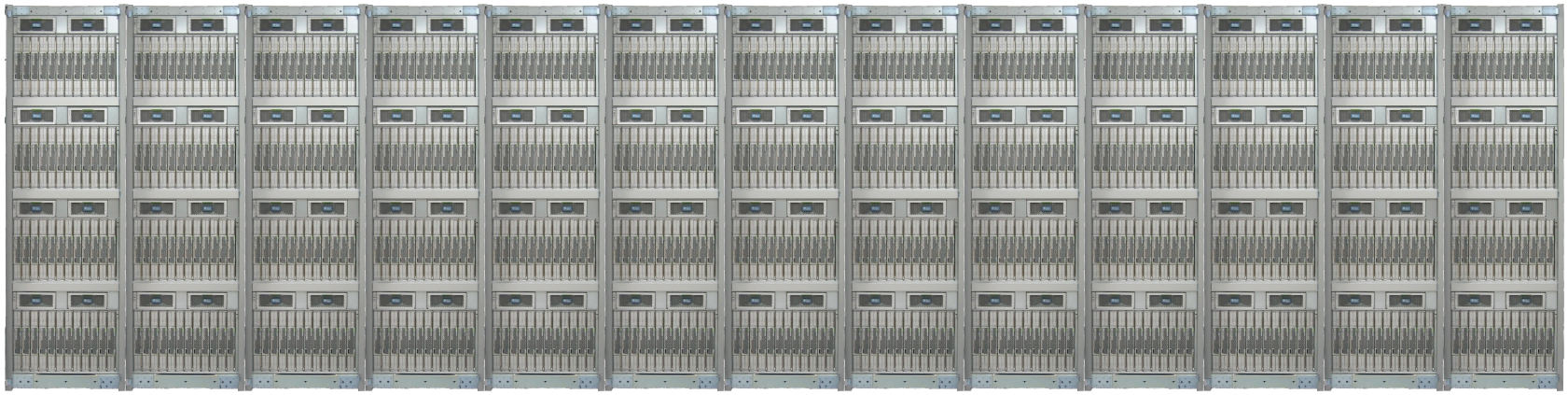
Sun HPC Storage Solutions



Sun HPC Storage Solutions



Basic distributed file system diagram



many Clients (100s-100000s)

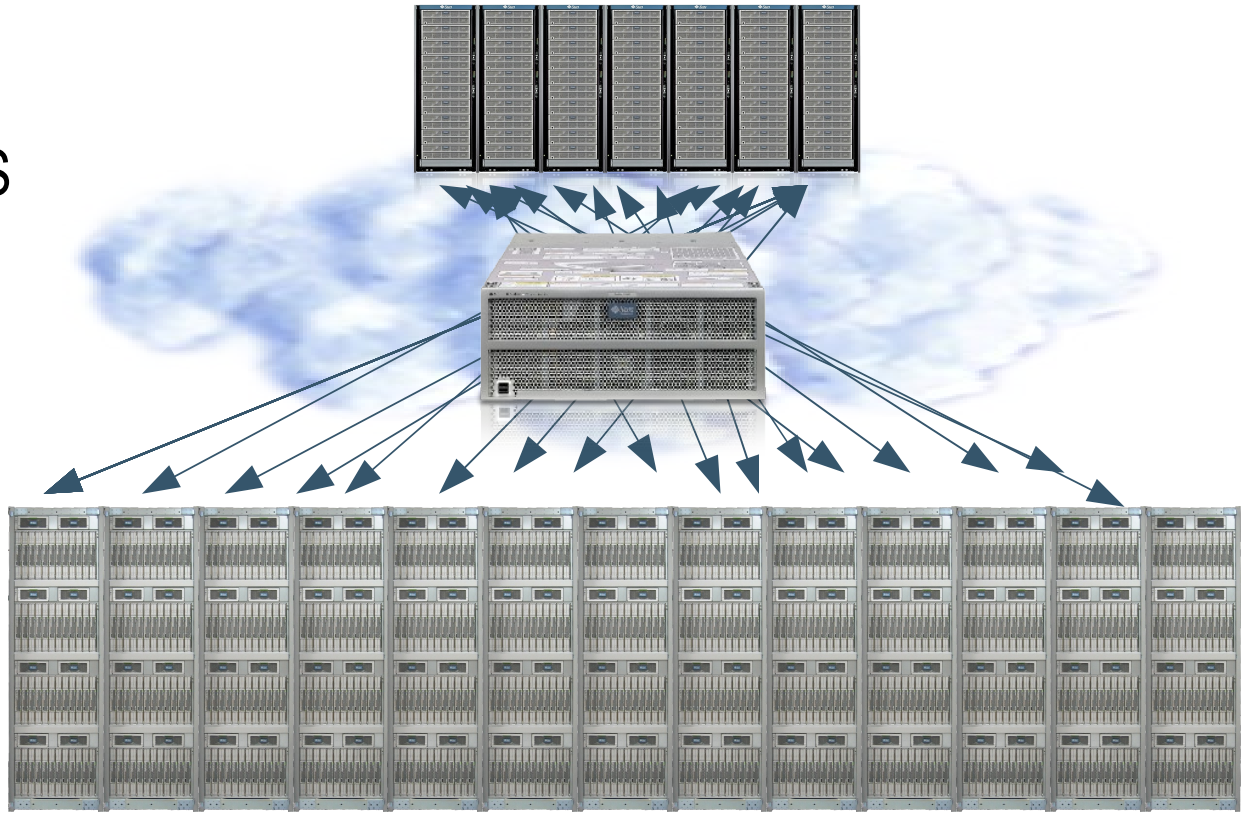
Survey of Distributed File Systems

There are 3 kinds of file systems for clusters:

- Proxy file systems
  - Name services, space management, and data services brokered through a **single server**.
  - This is then not a parallel file system
- SAN file systems
  - Name services and space management on the **metadata server; direct parallel access to the data** devices.
- Object storage file systems
  - Name services on the metadata server(s); **space management on the storage servers**; direct parallel access to the storages servers.

PROXY File Systems

- Name services, space management, and data services brokered through a single server. For example:
  - NFS
  - CIFS
  - Sun PxFS
  - etc.

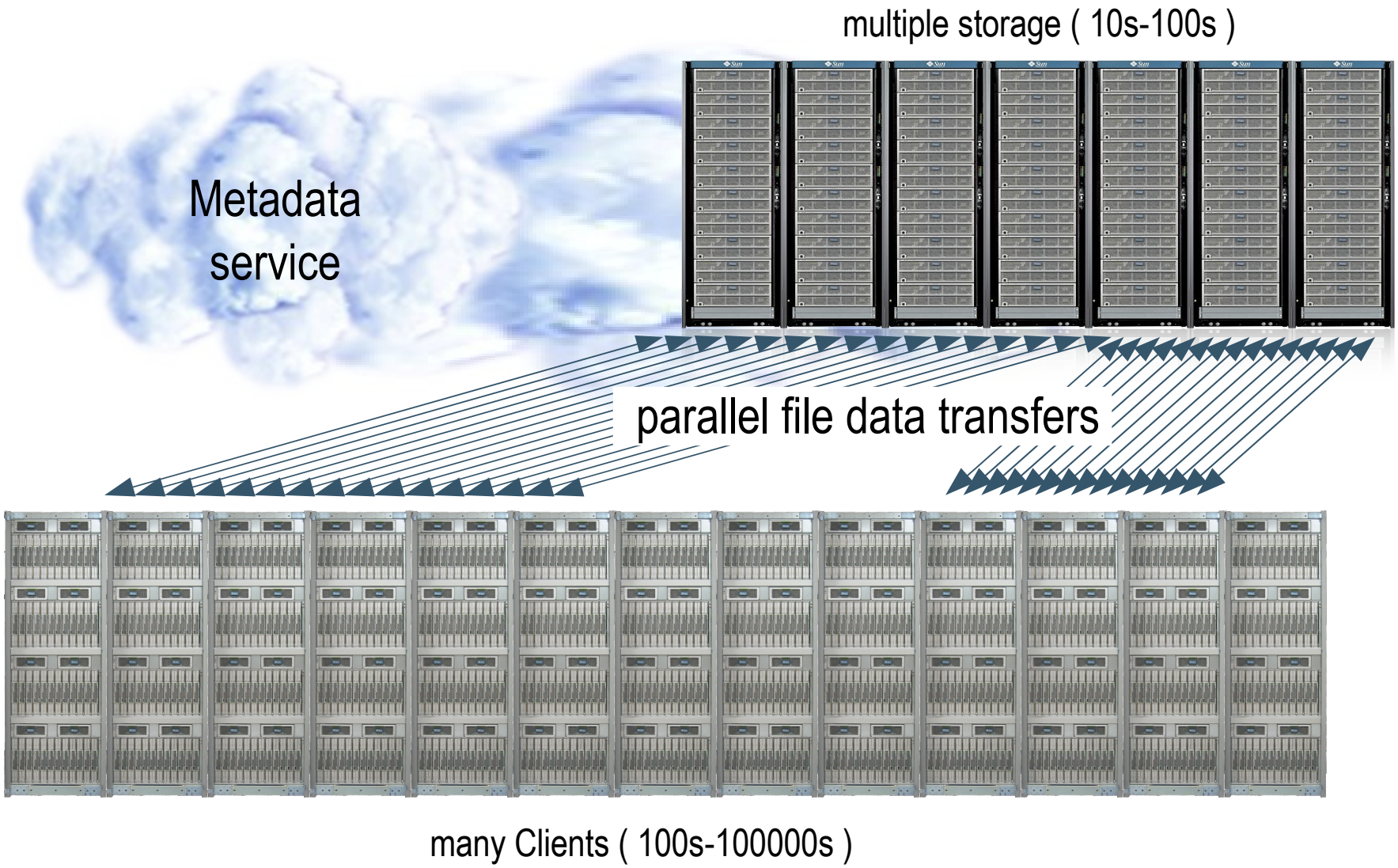


many Clients (100s-100000s)

Basic NFS numbers

- some people claim NFS is ~40 MB/s max (on GbE)
  - used to be true very long time ago
- we have seen 235 MB/s already years ago, from single file, untuned (on 10GbE)
- today we see 980 MB/s in the lab using a new NFSrdma implementation

Basic Parallel file system diagram



SAN File Systems

- Name services and space management on the metadata server; Clients directly access the data devices.For example:
  - Sun QFS Shared File System
  - IBM GPFS and IBM SANergy
  - ADIC StorNext
  - SGI CxFS
  - Redhat GFS, Polyserve, IBRIX, etc.
- SAN file systems with HSM:
  - SAM-FS on Sun Shared QFS
  - StorNext HSM on ADIC
  - DMF on SGI CxFS

Sun StorageTek QFS Shared File System

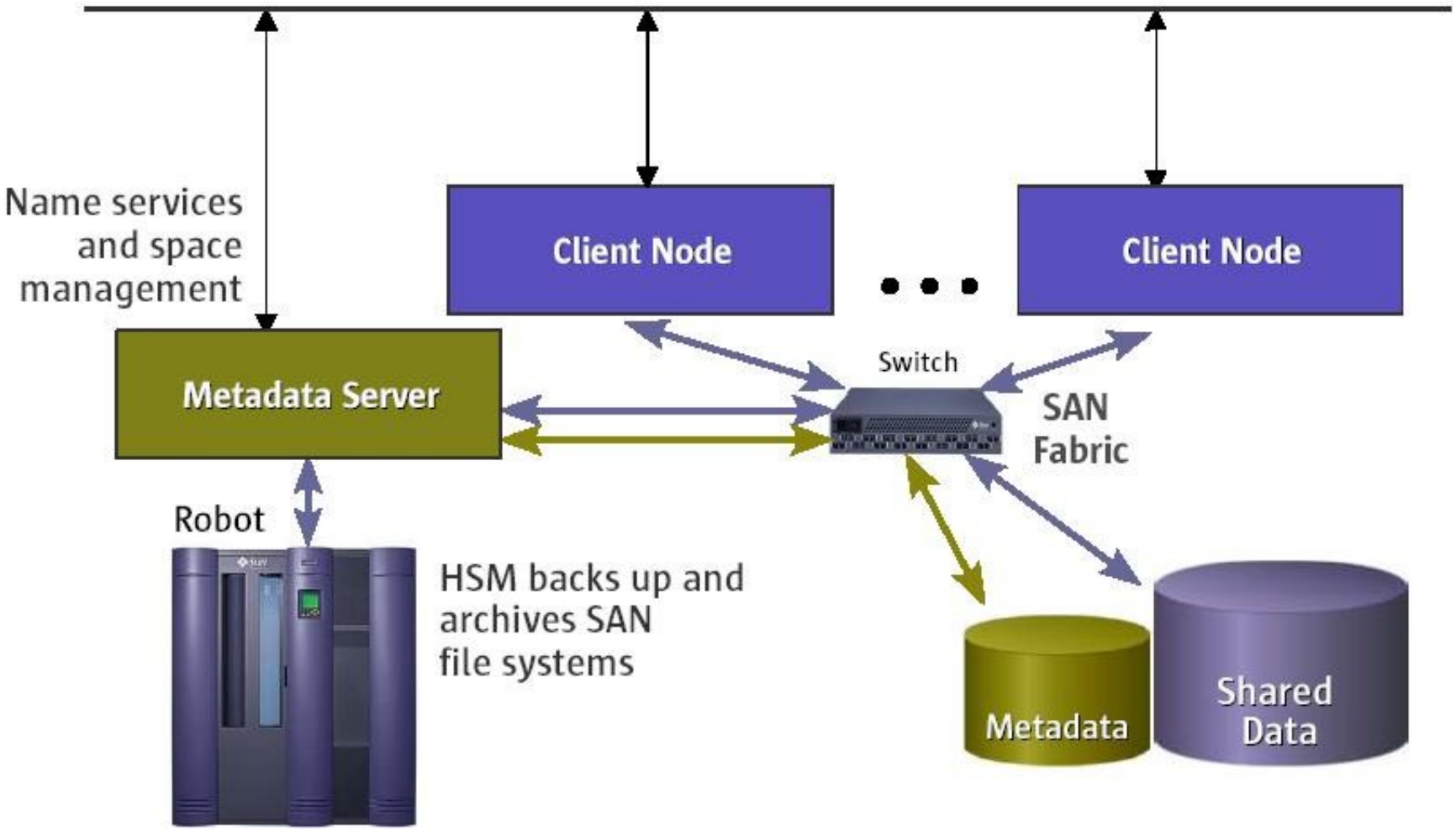
QFS is a high-performance shared file system which can be configured with metadata stored on devices separate from the data devices for greater performance and recovery or combined on a single device depending on the needs of the application.

The file system may be integrated with the SAM software to provide storage archiving services in the shared file system environment.

It supports Solaris and Linux clients. The Metadata server is always on Solaris.

SAN File Systems

Metadata over LAN



Sun StorageTek QFS

SAN File Sharing Innovation

Data Consolidation

- SAN file sharing
- Name services and space management on the metadata server
- Direct access to the data devices

Performance & Scalability

- Tune file system to the application
- Predictable performance
- File system performance scales linearly with the hardware

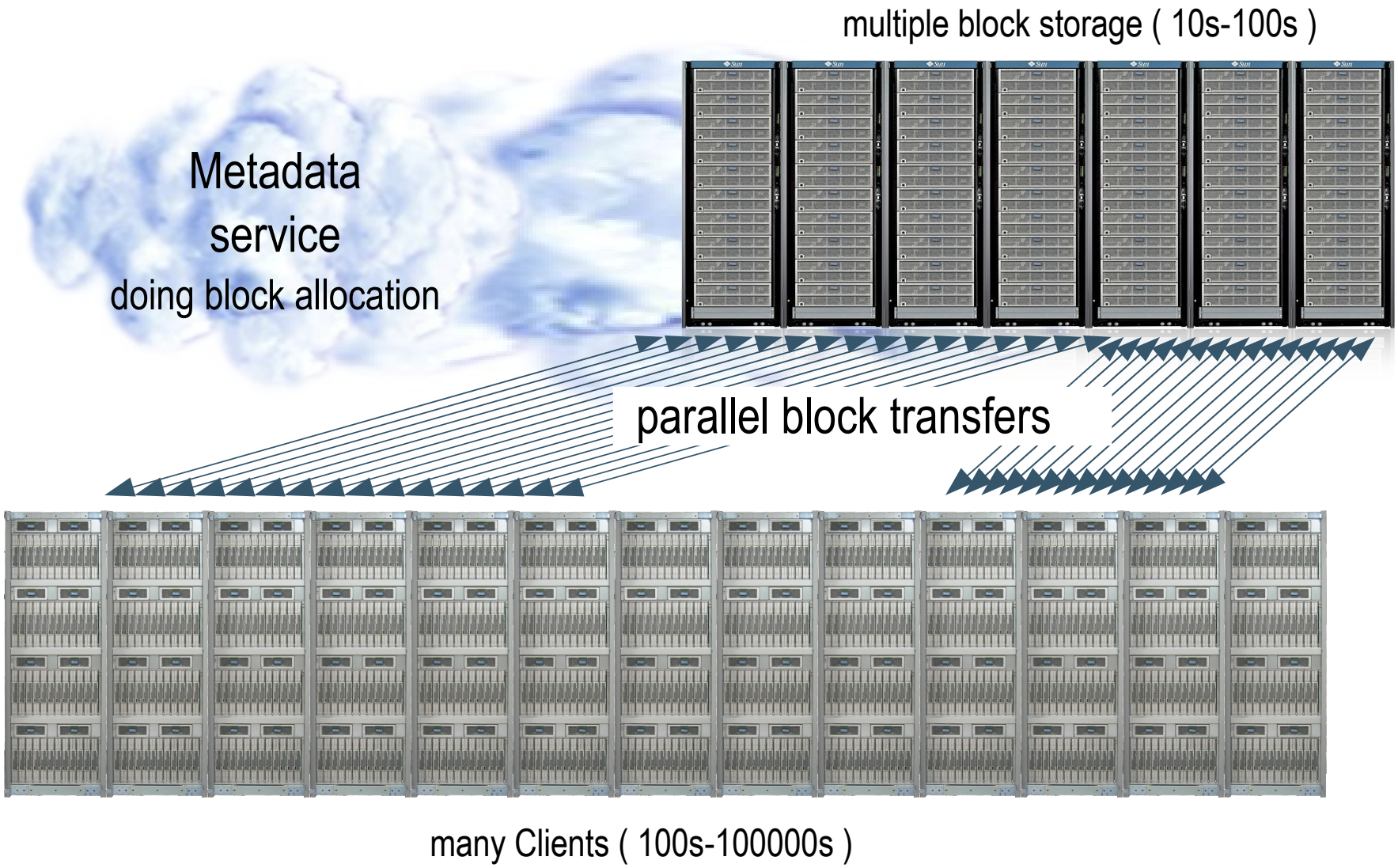
Parallel Processing

- Multi-node read/write access
- 128 nodes supported
- 256 nodes with IB support

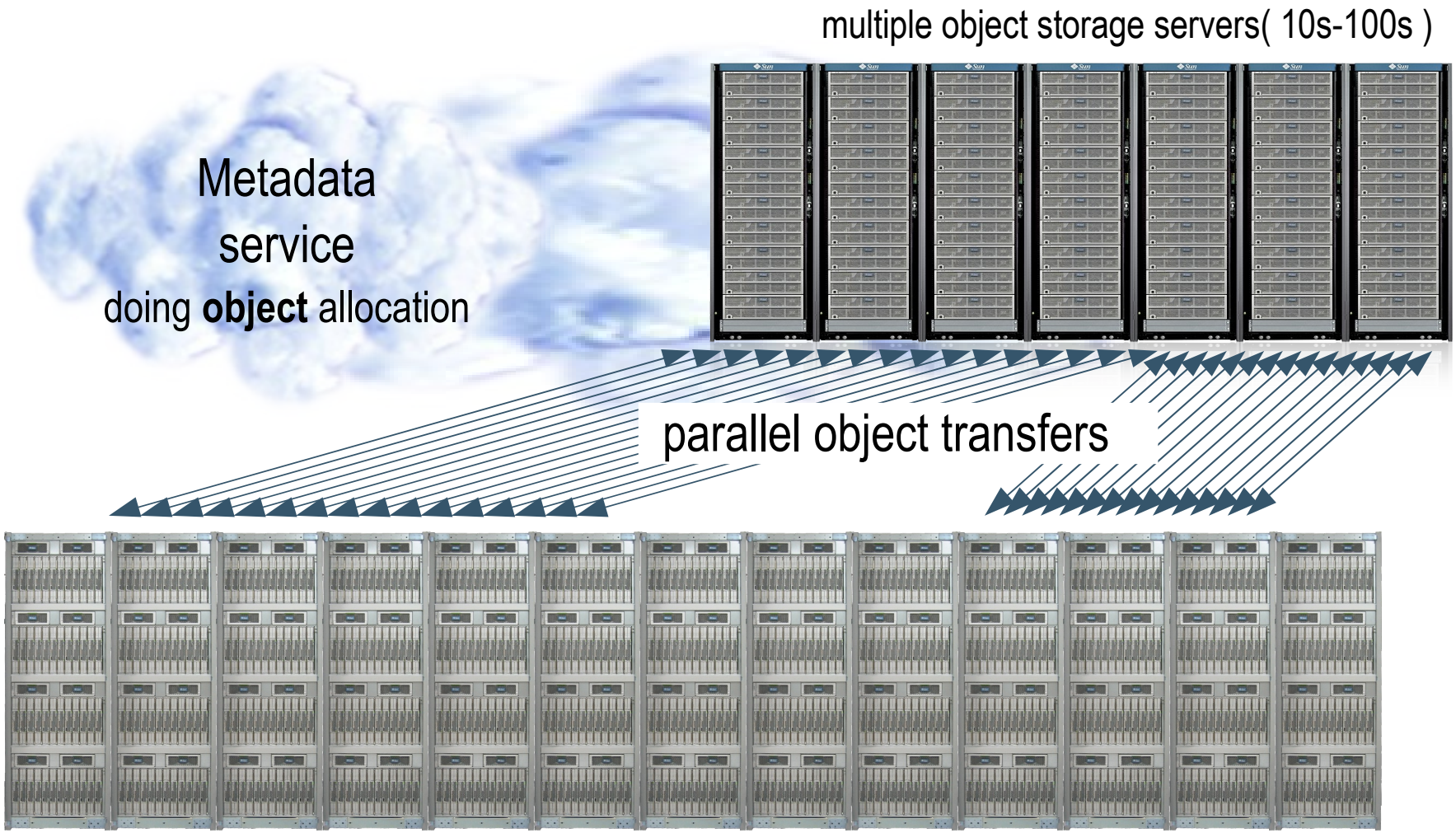
SAM-QFS 4.6 New Features

- Simplified Install & Setup
- Resource Mgt. Info GUI
- Data Integrity Verification
- Archive Copy Retention
- HA-SAM
- Snaplock compatibility for ECMs
- Directory Lookup Performance – Phase 1
- Honeycomb Integration as a disk archive target
- Sun Cluster support - Shared QFS nodes outside the cluster
- Linux Updates – SLES 10 support

Basic SAN file system diagram



Basic Object file system diagram



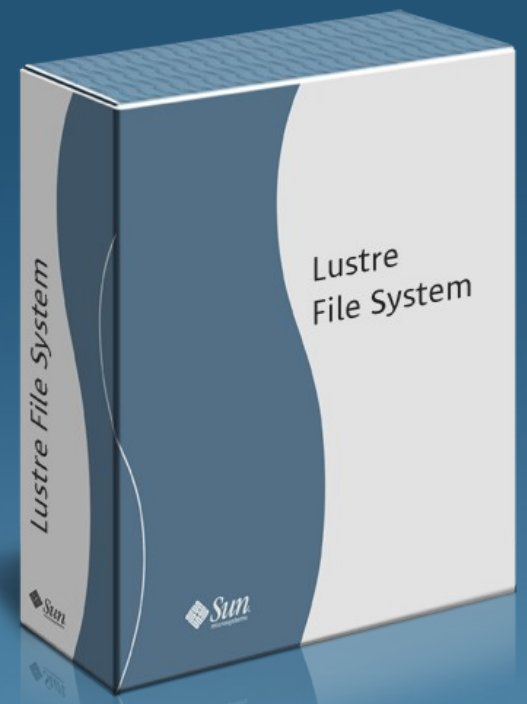
Object Storage File Systems

- Name services on the Metadata Server(s);
  space management on Object Storage devices (OSDs);
  direct access to the Object Storage devices (OSDs)
- Currently available object storage file systems:
  - **Lustre**, from Sun (Cluster File Systems, Inc.)
  - Panasas PanFS
  - IBM SanFS
- No HSM support on the Object Storage file systems
- IBM and Panasas adhers to the T10/OSD standard;
  T10/OSD is a SCSI protocol.
- Lustre uses a proprietary protocol

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.
- 50% of Top30 run Lustre
- 15% of Top500 run Lustre



The First Sun Constellation System Implementation

- The world's largest general purpose compute cluster based on Sun Constellation System
- More than 500 Tflops
  - 82 Sun ultra-dense blade platforms
  - 2 Sun ultra-dense switches
  - 72 Sun X4500 storage servers
  - X4600 frontend and service nodes
- Sun is the sole HW supplier
- Opteron Barcelona based
- started operation february 4th 2008

What is Lustre?

- Lustre is a storage architecture for clusters
  - Open source software for Linux licensed under GNU GPL
  - Standard POSIX-compliant UNIX file system interface
  - Complete software solution runs on commodity hardwares
- Key characteristics
  - Unparalleled scalability
  - Production-quality stability and failover
  - Object-based architecture
  - Open (Multi-vendor and multi-platform)
- Roadmap

Understanding Lustre

- Clients: 1-100000, good target is a couple of 10000
- Object Storage servers: 1-1000, in reality up to 400-500
- Metadata serves: 1-100, in reality only a couple
- Operation:
  - Client goes to the metadata server, get a handle describing where the data is
  - File can be striped over many object servers
  - Client does the I/O
  - There are no locks and additional info on the metadata servers!
  - Max file size: 1.2PB
  - Max file system size: 32PB
  - Max number of files: 2B

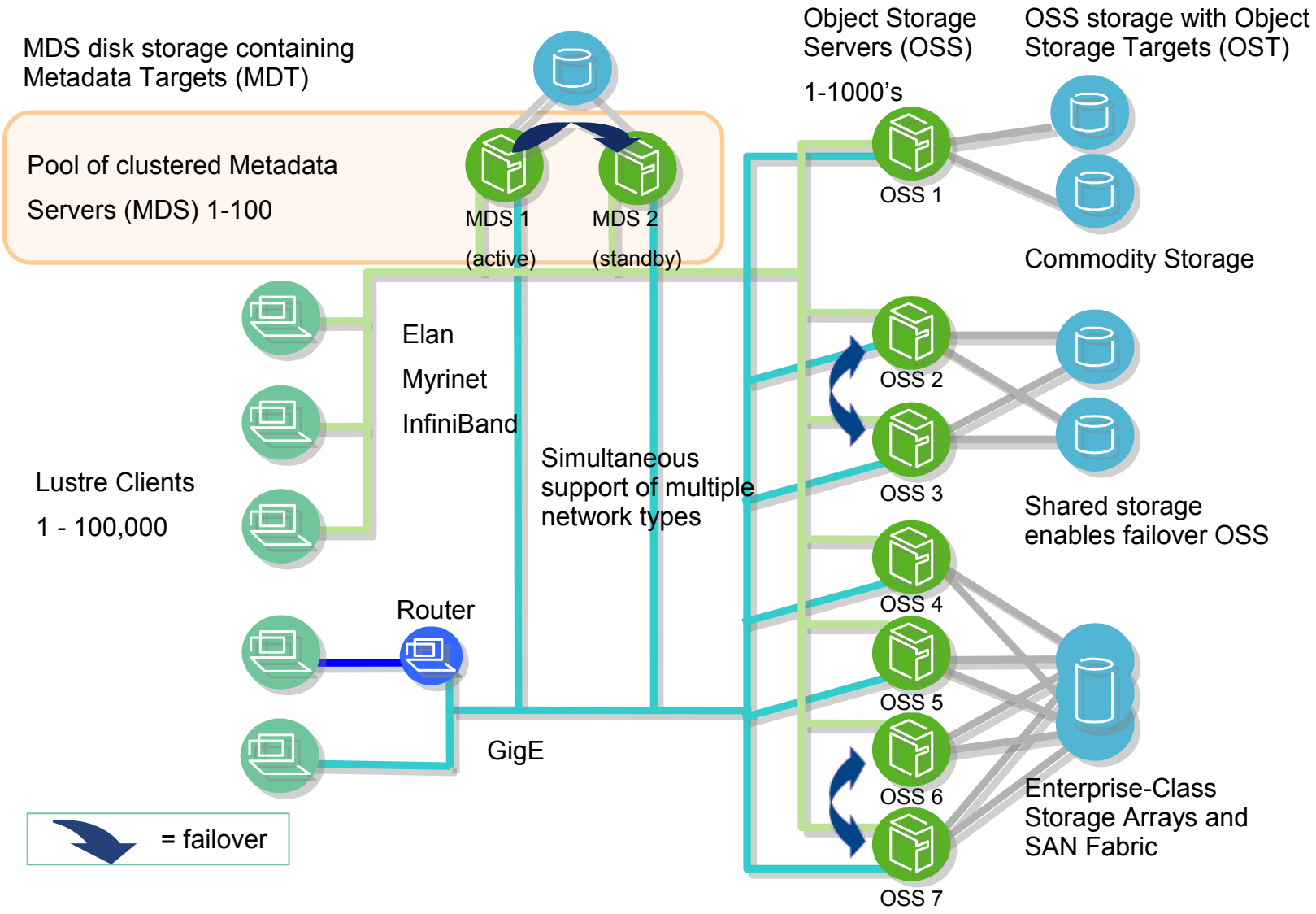
TACC users already reported:

- "Our datasets comprises 3 single real variables per grid point at 4096^3 [which] is 768GB. Our code took about 22 secs to write the files (each processor writes a file) which means a parallel performance of ~35GB/sec."
  - (TACC acceptance test measured 45 GB/s on benchmark)
- Red Storm test: 160 wide stripe (not recommended, use odd, or even prime numbers rather!!!), 10000 clients, 40GB/sec I/O, nice and constant performance for a single file activity

Lustre today

WORLD RECORD #Clients	Clients: 25,000 – Red Storm Processes: 200,000 – BlueGene/L Can have Lustre root file systems
WORLD RECORD #Servers	Metadata Servers: 1 + failover OSS servers: up to 450, OST's up to 4000
Capacity	Number of files: 2Billion File System Size: 32PB, Max File size: 1.2PB
WORLD RECORD Performance	Single Client or Server: 2 GB/s + BlueGene/L – first week: 74M files, 175TB written Aggregate IO (One FS): ~130GB/s (PNNL) Pure MD Operations: ~15,000 ops/second
Stability	Software reliability on par with hardware reliability Increased failover resiliency
Networks	Native support for many different networks, with routing
Features	Quota, Failover, POSIX, POSIX ACL, secure ports

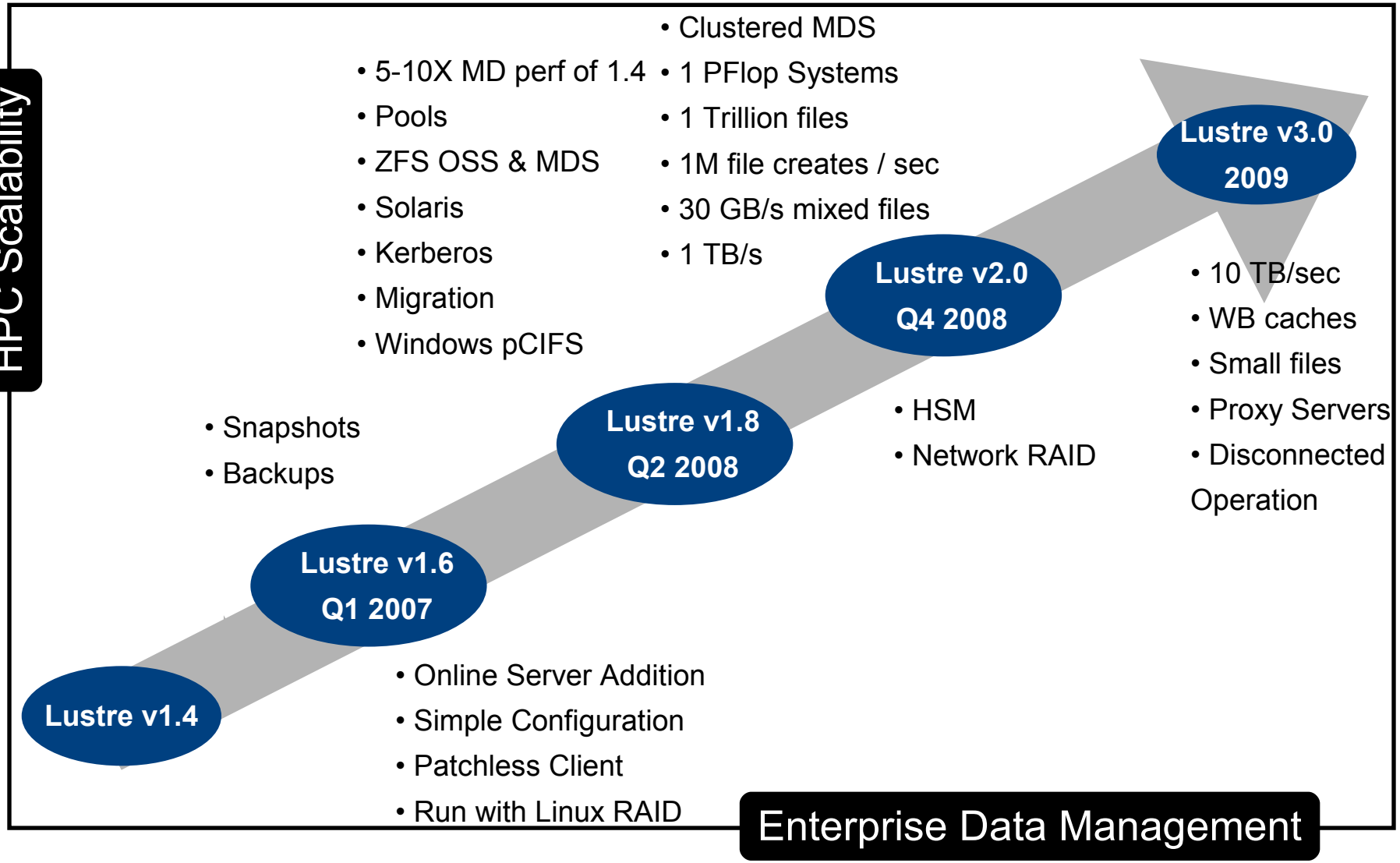
A Lustre Cluster with everything in it:



Recent Improvements

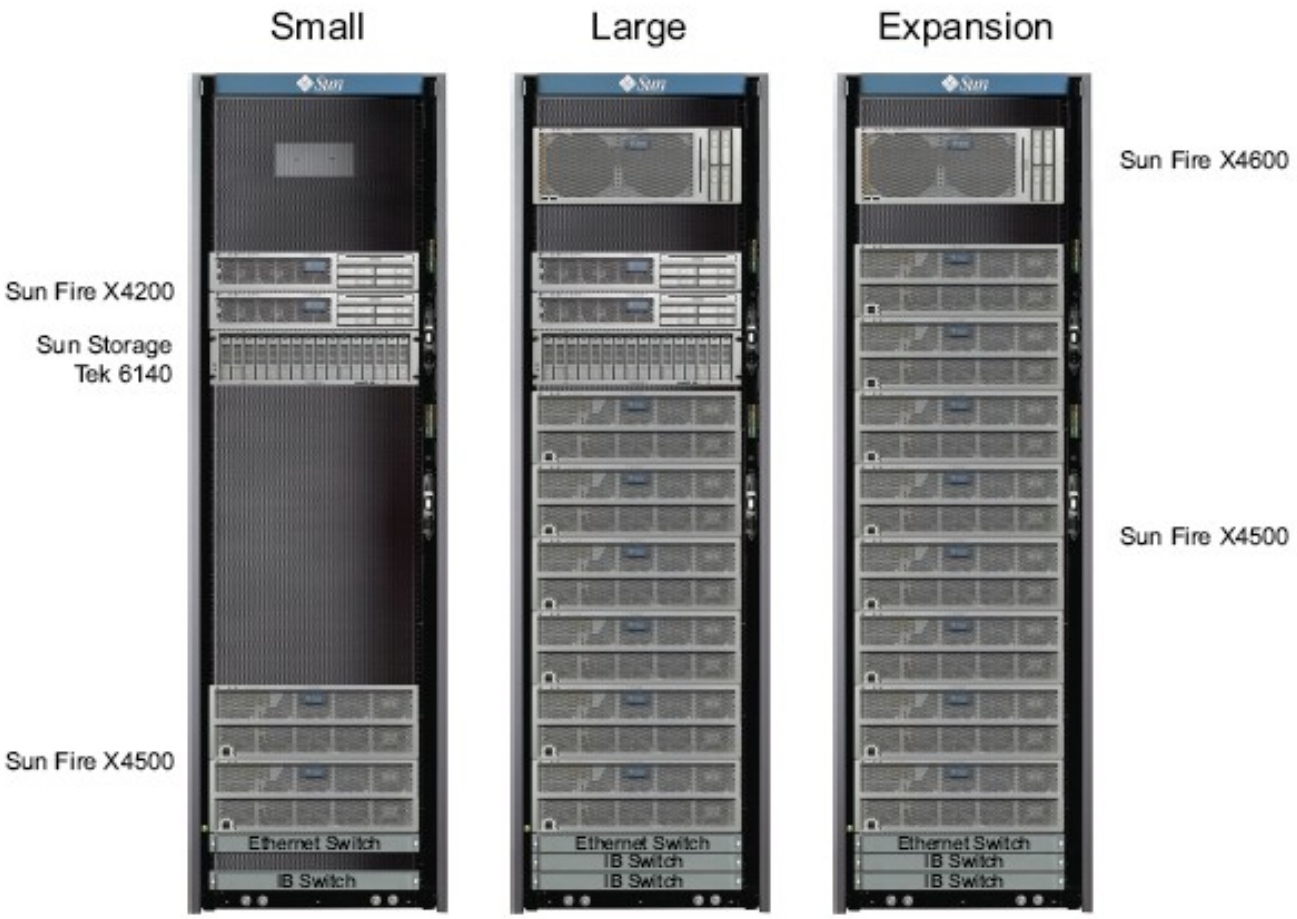
Lustre	Clients require no Linux kernel patches (1.6) Dramatically simpler configuration (1.6) Online server addition (1.6) Space management (1.8) Metadata performance improvements (1.4.7 & 1.6) Recovery improvements (1.6) Snapshots & backup solutions (1.6) CISCO, OpenFabrics IB (up to 1.5GB/sec!) (1.4.7) Much improved statistics for analysis (1.6) Backup tools (1.6.1)
Linux	Large ext4 partitions support (1.4.7) Very powerful new ext4 disk allocator (1.6.1) Dramatic Linux software RAID5 performance improvements
Other	pCIFS client – in beta today

Lustres Intergalactic Strategy



Sun Lustre Solutions

Sun Customer Ready Scalable Storage Cluster



Small
48 Terabytes*
1.6 GB/sec

Large
144 Terabytes*
4.8 GB/sec

Expansion
192 Terabytes*
6.4 GB/sec

* with 500 GB drives

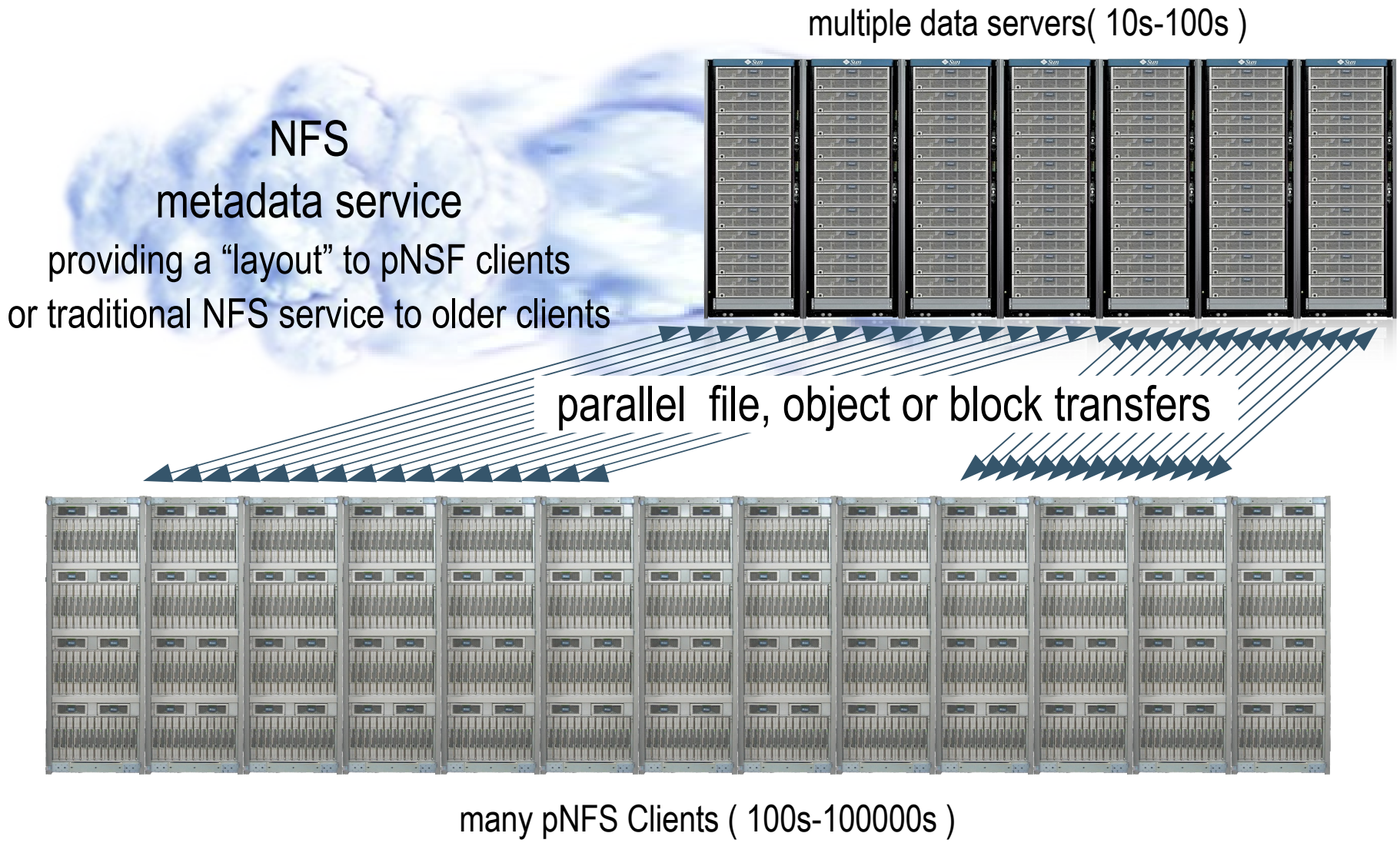
Conclusion

- Lustre is almost 9 years old
- The #1 file system at scale
- The originally elusive numbers are in the bag:
  - >100 GB/sec
  - >10,000 clients
- Cluster growth was seriously underestimated
  - Lustre will scale – 10TB/sec, 1,000,000 clients

What is pNFS?

- Parallel extensions to NFS to improve bandwidth
  - Designed by HPC community
  - Adopted by IETF NFS WG as part of NFSv4.1
    - Draft standard expected consensus 2008 (presently draft 21)
    - final content due next week, complete approvals later this year
- Major features
  - Parallel transfer from multiple servers to single client
  - Global name space
  - Horizontal scale of data storage
  - Exists within mental model of existing NFS community
  - Three types of implementation: *files, objects, blocks*
    - All offer *identical* file semantics to client

Basic pNFS diagram



Architecture

- "Classical" cluster file system architecture
  - Metadata is handled by centralized server (MDS)
  - User data is distributed to many data servers (DS)
- Clients open files by asking the MDS
  - MDS verifies permissions, sends back *layout*
  - Layout describes how file data is organized and where
  - Standard supports RAID-0 and replicated data
    - Striping done at the file level
    - Files have custom organizations, even in same directory
      - /pnfs/file1 might be 5-way striped on s1, s99, s4, s12, s04
      - /pnfs/file2 might be a singleton file on s4

Orthogonal Features

- Exists within NFSv4.0 context
  - Mount, showmount, sharemgr, GSS security, Kerberos, TX are all the same as for traditional NFS
  - Data transfer protocol is the same too
- NFS-over-RDMA
  - Remote DMA allows client/server communication without resorting to XDR/TCP/IP protocol overhead
  - Applies to non-parallel NFS and pNFS
  - New Solaris implementation underway – in onnv Q4CY07
    - Fully standard compliant (ignore old version in Solaris 10)
    - Fast and efficient: 980 MB/sec vs 235MB/sec (10G ethernet) (and 980 MB/sec was hardware limited on x2100m1)

pNFS Implementation

- pNFS is *optional* for NFSv4.1 clients
  - MDS must be prepared to proxy data from DS to client if client cannot handle layouts
  - Not expected to be common
  - However, this required capability may prove “useful”
- Data on all DS shares – global – name space
  - Could be hundreds or thousands of data servers
- pNFS Implementation Status
  - Code is in advanced prototype stage, both Linux and Solaris
  - Source released to OpenSolaris, estimated put back late '08
  - Server code also already in Lustre

Sun HPC Open Software Stack

Sun CRS, Support, Architectural, Professional Services

Developer Tools Distributed Applications	Sun Studio 12 Sun HPC ClusterTools	Free
Management Workload Management Cluster Management	Sun Grid Engine Software Sun Connection, ROCKS, Ganglia	Open, Free
Distributed IO File System, Visualization	Sun Lustre, QFS, NFS, pNFS et al. Sun Visualization System	Open
Nodes Processors and Kernels	AMD 64 Opteron, Windows, intel, ULTRASPARC, solaris, opensolaris	Open
Interconnect	Gigabit Ethernet, Myrinet, Infiniband, and Suns 3456 Port Non-Blocking IB Switch	Open

Vielen Dank!

Roland Rambau
roland.rambau@sun.com