Parallel File Systems Compared

Roland Laifer
Computing Centre (SSCK)
University of Karlsruhe, Germany
Laifer@rz.uni-karlsruhe.de

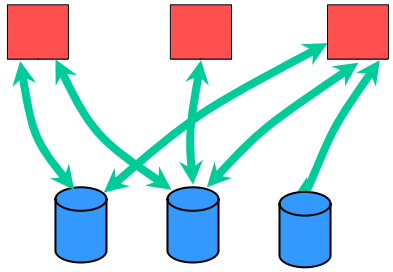
Outline

- **Parallel file systems (PFS)**
  - **Design and typical usage**
  - **Important features**
  - **Comparison of the most important products**
  - **Deployment at the computing centre**

Introduction

What is a distributed file system?

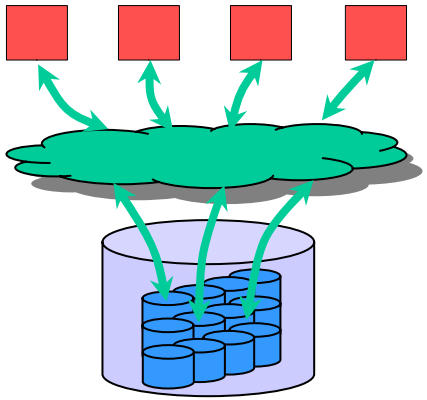
- File system data is usable at the same time from different clients



With multiple servers applications see separate file systems

What is a parallel file system (PFS)?

- Distributed file system with parallel data paths from clients to disks

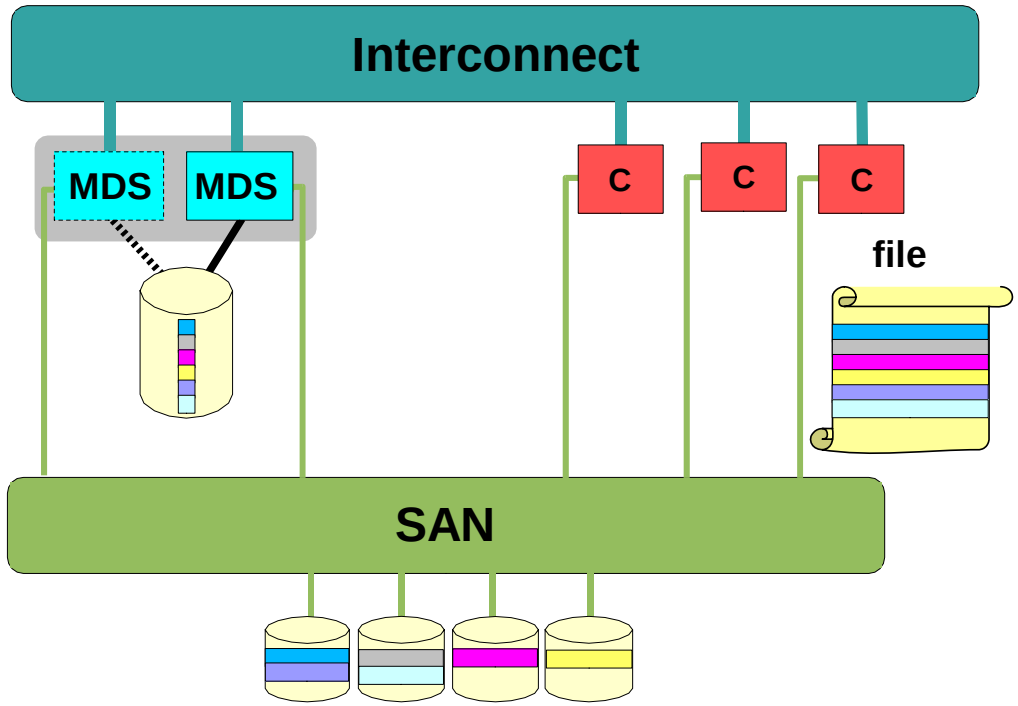


Even with multiple servers applications typically see one file system

Current trends

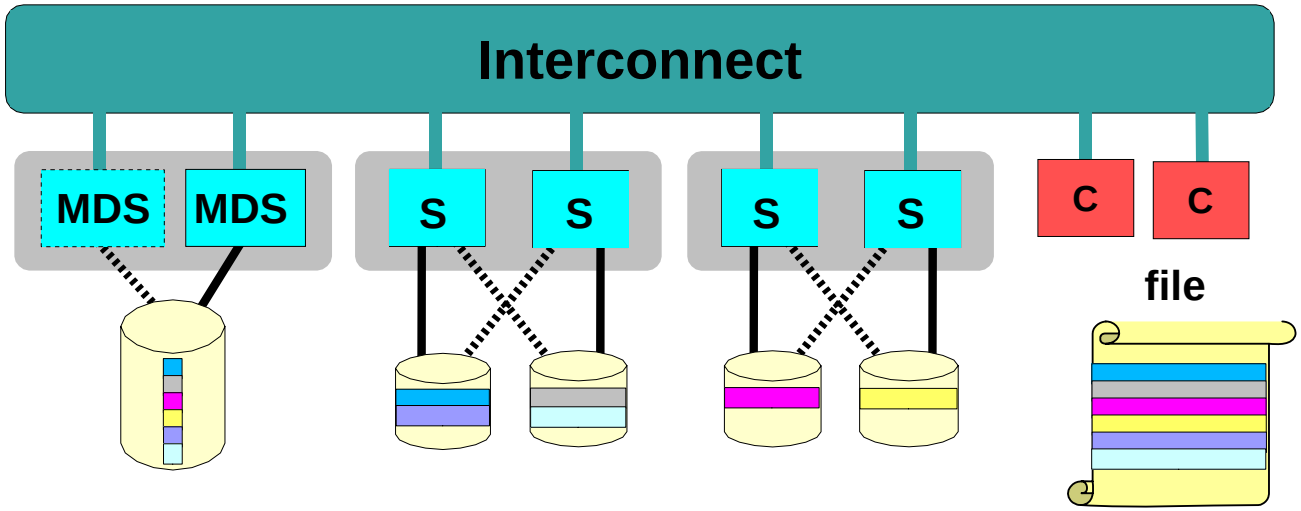
- **Storage needs increase and disk costs decrease steadily**
  - **Storage systems are rapidly growing**
    - **Trend towards RAID6 because of growing chance of multiple disk failures**
  - **Storage consolidation in order to reduce administrative costs**
    - **Also allows to dynamically allocate storage**
    - **New trend to have one parallel file system for multiple clusters**
- **Nearly no improvement in disk access times**
  - **Increased speed by striping data over multiple disks/disk subsystems**
- **Frequently need for high transfer rates**
  - **Trend towards parallel file systems**
    - **Several new parallel file systems were recently developed**
    - **Existing parallel file systems were greatly enhanced**
- **Number of clients in HPC systems is heavily increasing**
  - **Scalability becomes more and more important**

SAN based parallel file systems



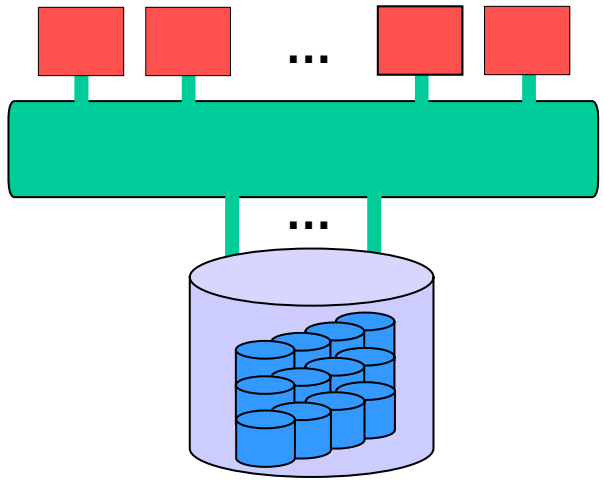
- Striping over disk subsystems
- Needs a storage area network (SAN)
  - Traditionally FC based, alternatives are iSCSI or InfiniBand protocols
- Examples:
  - ADIC SNFS, SGI CXFS, RedHat GFS, IBM GPFS (without NSD servers)

Network based parallel file systems



- Striping over servers
- Uses low level and fast communication (RDMA) over interconnect if possible
- Examples:
  - Lustre, IBM GPFS (with NSD servers), Panasas ActiveScale Storage Cluster

Typical PFS usage (1): Cluster file system

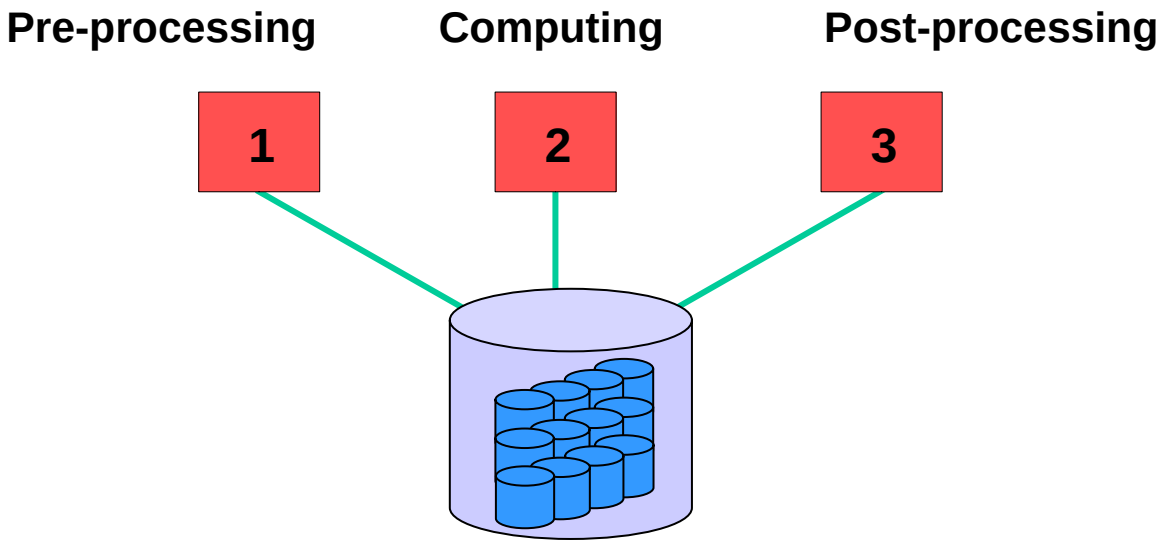


Cluster nodes

Interconnect or SAN

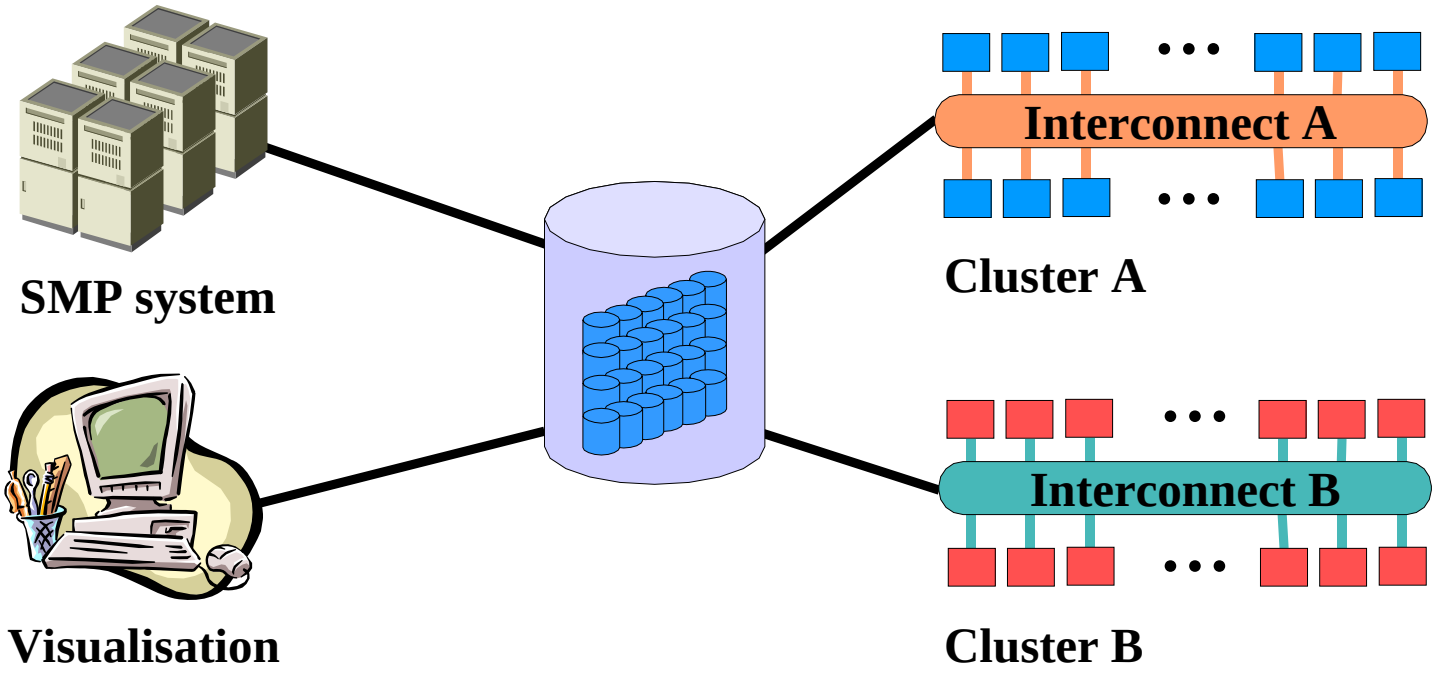
- File system and cluster usually from same vendor
  - Good parallel file system is important for cluster selection
- Benefit is increased throughput, scalability and easy usability

Typical PFS usage (2): Workflow file system



- Typical customers: Oil & gas, digital media
- Usually moderate number of heterogeneous clients
  - SAN based PFS are used in most cases
- Benefit is accelerated workflow and easy usability

Typical PFS usage (3): Global file system



- New concept with additional requirements:
  - Lots of clients (scalability)
  - Minimal downtime
- Examples at LLNL, ORNL, NERSC, TeraGrid, DEISA

Common PFS properties

- **Throughput performance mainly depends on available hardware**
  - **Most PFS can reach accumulated > 10 GB/s for sequential read/write**
    - **More than 100 GB/s have been demontsrated with GPFS and Lustre**
- **Metadata performance of one file system is limited**
  - **Maximum is usually 5000-10000 operations per second**
    - **May be lower for deletes or if files are stored in a single directory**
- **Possible configurations without single point of failure**
  - **Requires dedicated hardware and failover support of software**
- **User level security is not available**
  - **Root on all clients has full file system access**
- **Linux kernel dependencies**
- **NFS or CIFS gateways to connect unsupported clients**
- **POSIX file system semantics**

Main PFS differences (1)

- **Scalability, i.e. number of supported clients**
  - **SAN based file systems are often limited to 100 clients**
- **Heterogeneity**
  - **Supported operating system and kernel versions**
    - **SAN based file systems often support more operating systems**
- **Reliability**
  - **Number of installed systems of similar size**
    - **Expect software problems if file system is new**
  - **Quality of software support**
- **Costs**
  - **Supported storage subsystems and disks**
  - **Requirement for special or additional hardware**
  - **Software and maintenance**
  - **Complexity of administration**

Main PFS differences (2)

- **Metadata and lock management implementation**
  - **Is most critical and complicated part of each PFS**
    - **Usually a PFS is not well suited for mail or database servers**
    - **For MPI-IO parameters have to be carefully chosen**
- **Network options**
  - **Supported networks, protocols and speed**
    - **Examples: GigE, 10 GigE, 4x DDR InfiniBand, 4 Gb FC, iSCSI**
  - **Support for multiple network connections or low level gateways**
- **Adequate backup solution**
  - **Very fast or parallel restore is required**
  - **Snapshots help to create consistent backup or to restore user data**
- **HSM support**
  - **Usually a PFS supports only a dedicated HSM system**
  - **Archiving by users is an alternative to HSM**

PFS products (1): Lustre

» Status

- – **User base is rapidly growing**
  - **E.g. SSCK, U of Dresden, LLNL, PNNL, Sandia, ORNL, CEA, TITECH, PSC**
- – **Roadmap, FAQs and source code from Cluster Filesystems Inc. (CFS)**
  - **http://www.clusterfs.com/**
- – **Lustre products available from many vendors**
  - **CFS, HP, Cray, Bull, LinuxNetworx, Transtec, Sun, DDN**

» Pros and Cons

- + **Runs pretty stable**
  - **Experiences at SSCK: http://www.rz.uni-karlsruhe.de/dienste/lustretalks**
- + **Open source**
- + **Scalable up to 10000's of clients**
- + **High throughput with multiple network protocols**
  - **InfiniBand, Quadrics, Myrinet, TCP/IP**
  - **LNET routers provide gateways with high performance and failover**
- – **Currently supports only Linux clients**
  - **Patchless client will hopefully remove kernel dependencies**

PFS products (2): IBM GPFS

» Status

- Large user base
  - E.g. DEISA sites, FZK, NERSC, SDSC, LLNL, TeraGrid
- IBM GPFS Concepts, Planning, and Installation Guide provides good introduction
  - http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.gpfs.doc/gpfsbooks.html
- Also available via OEMs, e.g. LinuxNetworx

» Pros and Cons

+ Runs pretty stable
+ Offers many useful features
  - Snapshots, data and metadata replication, online disk removal
+ Scalable up to 1000's of clients
+ Feature list is permanently improved
- Currently supports only Linux and AIX clients
- Limited support for different network protocols
  - InfiniBand RDMA support is still missing

PFS products (3): Panasas ActiveScale Storage Cluster

» **Status**

– **Medium user base**
  - **E.g. U of Cologne (RRZK), LANL, Walt Disney, Paradigm**

– **Further information**
  - **http://www.panasas.com/products_overview.html**

» **Pros and Cons**

\+ **Easy installation and administration**

\+ **Supplies good performance for random IO**

\+ **Offers additional useful features**
  - **Snapshots, dynamic load balancing**

\+ **Scalable up to 1000's of clients**

– **Currently supports only Linux clients**

– **Supports only Gigabit Ethernet**
  - **Throughput per client is limited to 80-100 MB/s**

– **Needs dedicated storage hardware from Panasas**

PFS products (4): ADIC StorNext File System (SNFS)

» **Status**

– **Medium user base**
  - **E.g. FZK, CGG, Digital FilmWorks, Air Force Research Lab**

– **Further information**
  - **http://www.adic.com/stornext**
  - **ADIC is now owned by Quantum**

» **Pros and Cons**

\+ **Support for many different clients**
  - **Linux, Irix, Solaris, Windows 2000/XP/2003, MAC OS X, AIX, HP-UX, UNICOS**

\+ **Good HSM and backup integration**

\+ **Easy installation**

\+ **Offers additional useful features**
  - **Snapshots, data replication, guaranteed bandwidth, multipathing**

– **Scalable up to 128 clients**

– **Needs a storage area network**

PFS products (5): SGI CXFS

» **Status**

– **Medium user base**
  - **E.g. LRZ, U of Dresden, SARA, NASA, BBC, Ford, John Deere**

– **Further information**
  - **http://www.sgi.com/products/storage/tech/file_systems.html**

» **Pros and Cons**

\+ **Support for many different clients**
  - **Linux, Irix, Altix, Solaris, Windows 2000/XP/2003, MAC OS X, AIX**

\+ **Good HSM and backup integration**

\+ **Offers additional useful features**
  - **Guaranteed bandwidth**

– **Scalable up to 64 clients**

– **Needs a storage area network**
  - **InfiniBand is also supported**

– **Needs dedicated hardware for MDS**

PFS products (6): RedHat GFS

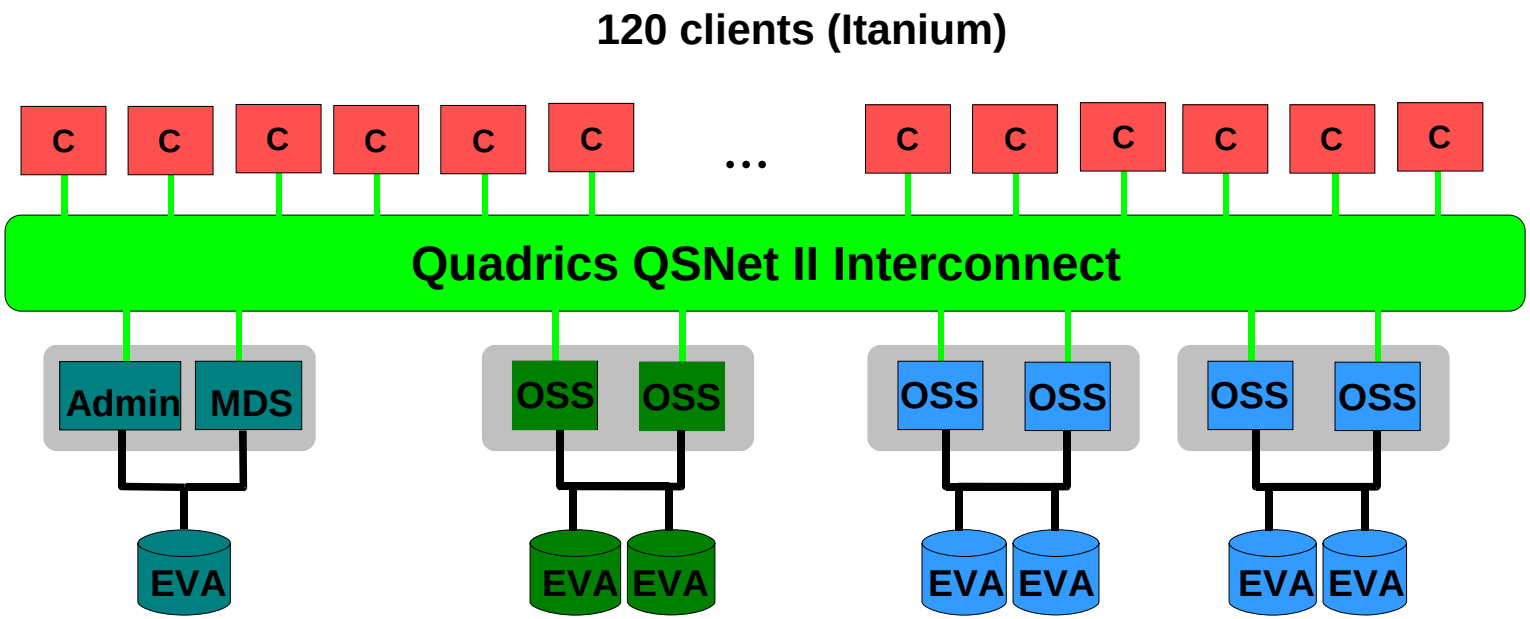
» Status

- – Possibly small commercial user base
  - E.g. Secure-24, CD-adapco
- – Further information
  - http://www.redhat.com/software/rha/gfs/

» Pros and Cons

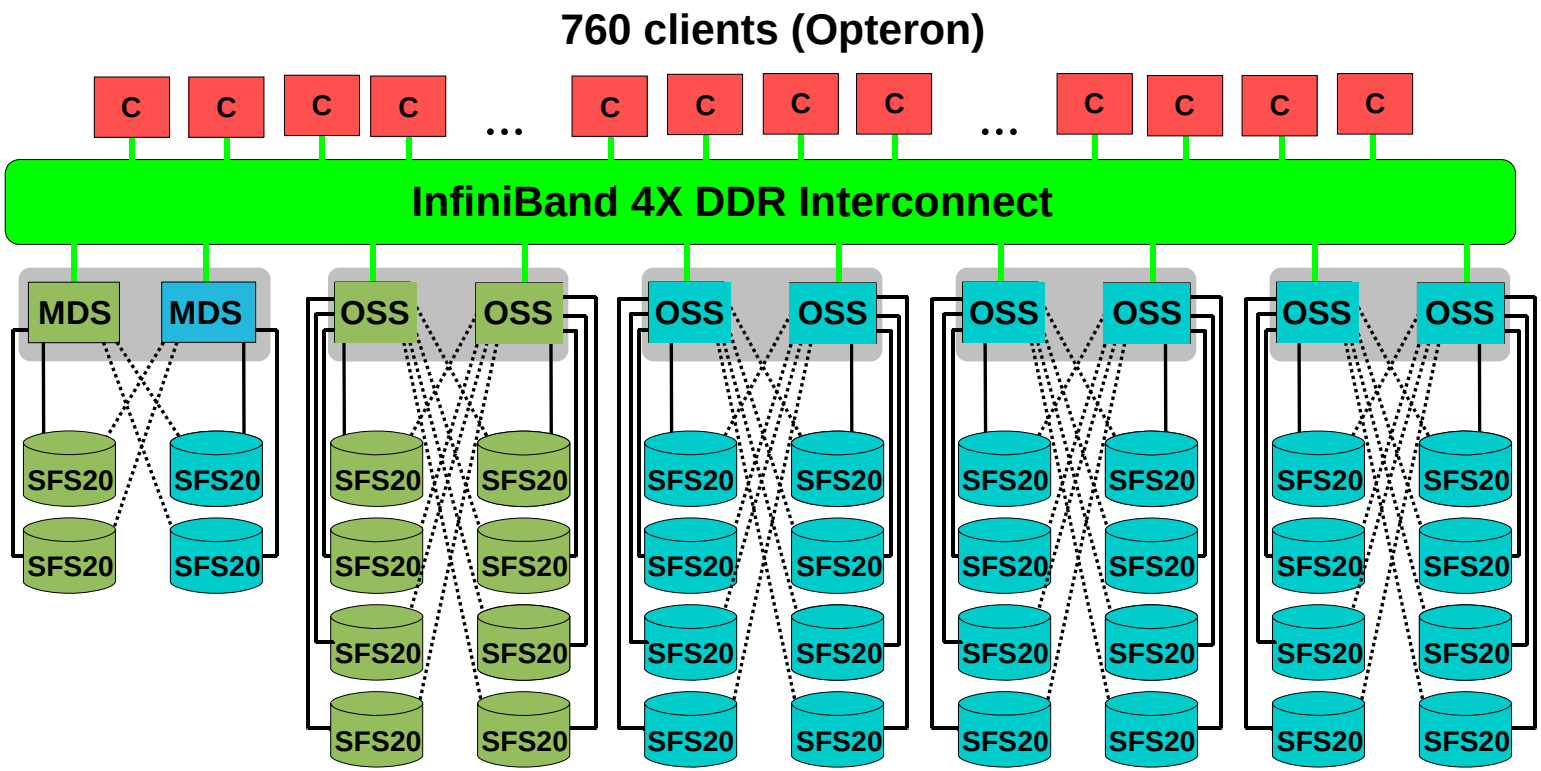
- \+ Tightly integrated with RedHat Linux
- \+ Open source
- \+ Supports Oracle RAC database clustering
- \+ Scalable up to 256 clients
- – Supports only Linux clients
- – Needs a storage area network
- – Needs HP Serviceguard for HA solution
- – Not sure if stability is good
  - Lock manager was redesigned due to performance problems

Example: HP SFS/Lustre at SSCK's HP XC6000



	$HOME	$WORK
Capacity	3.8 TB	7.6 TB
Write performance	240 MB/s	480 MB/s
Read performance	380 MB/s	760 MB/s

Example: HP SFS/Lustre at SSCK's HP XC4000



	$HOME	$WORK
Capacity	8 TB	48 TB
Write performance	360 MB/s	2100 MB/s
Read performance	600 MB/s	3600 MB/s

Example: SSCK's plan for a global parallel file system