Parallel File Systems and Impact of Blade Systems

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

This talk gives an explanation of the design and typical usage of parallel file systems. It will also provide a rich feature list you should take into account when selecting a parallel file system. Some of these features are important for selecting a parallel file system on blade systems. A discussion of this topic will also identify possible options for SAN based file systems. Additionally, the talk will include several examples and lots of experiences with parallel file system administration.

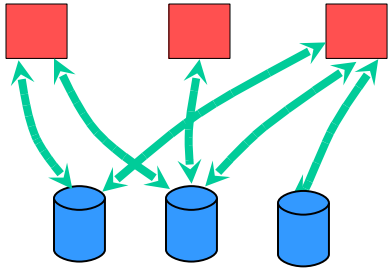
Outline

- Design and typical usage of parallel file systems
- Important features of parallel file systems
- Impact of blade systems on parallel file systems

Introduction

» **What is a distributed file system?**

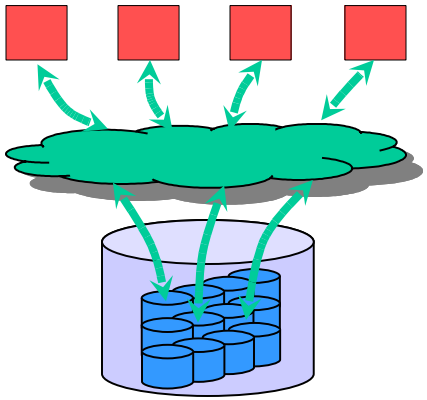
- **File system data is usable at the same time from different clients**



Applications see separate file systems

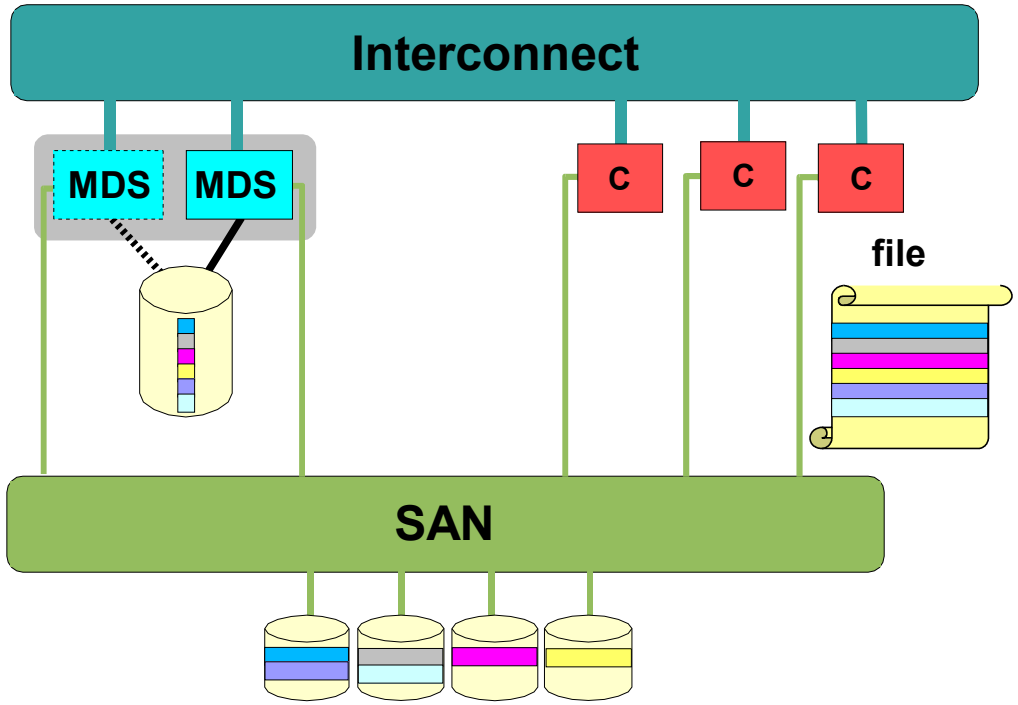
» What is a parallel file system (PFS)?

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



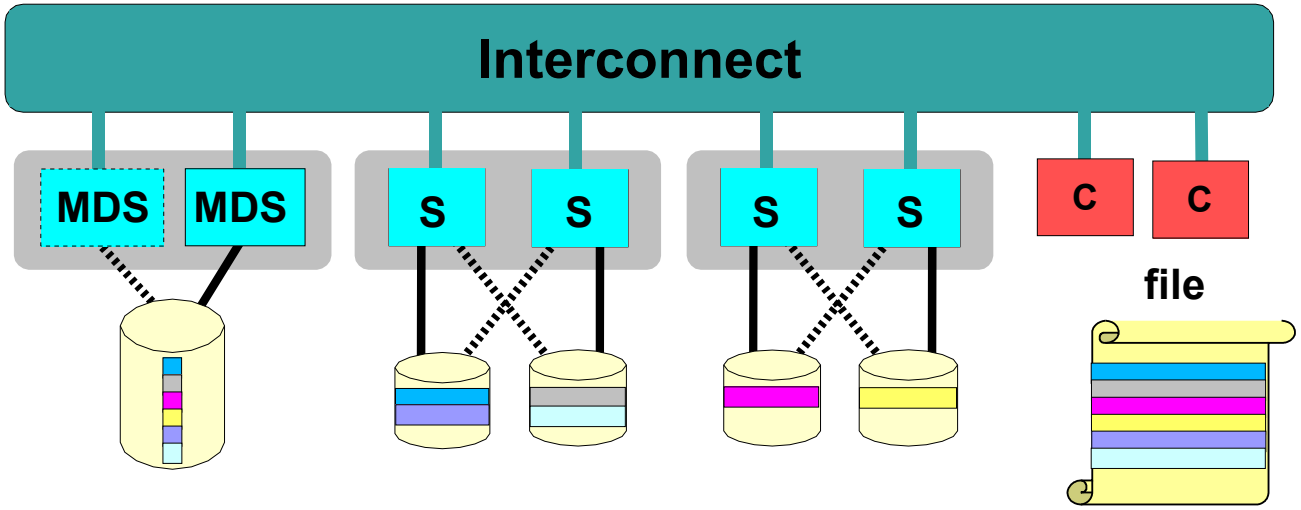
Applications typically see one file system

SAN based parallel file systems



- Striping over disk subsystems
- Traditionally needs a storage area network (SAN)
- Examples:
  - ADIC SNFS, SGI CXFS, RedHat GFS, IBM GPFS (without NSD servers)

Network based parallel file systems

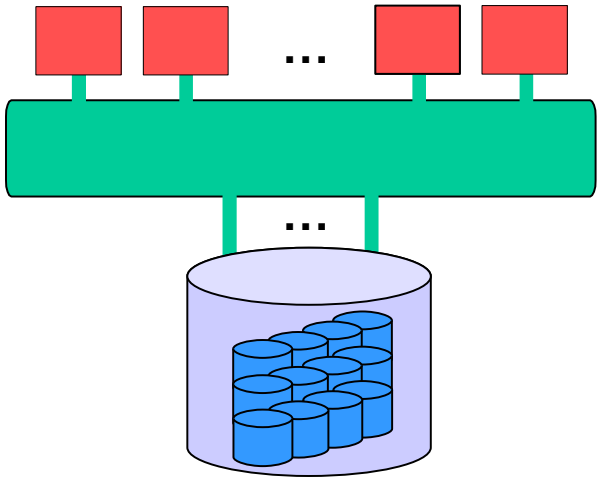


- Striping over servers
- Uses low level and fast communication over interconnect if possible
- Examples:
  - Lustre, IBM GPFS (with NSD servers), Panasas ActiveScale Storage Cluster

Current trends

- **Storage needs increase and disk costs decrease steadily**
  - **Storage systems are rapidly growing**
  - **Storage consolidation in order to reduce administrative costs**
    - **Also allows to dynamically allocate storage**
- **Multiple clients need access to the same data**
  - **Distributed file systems are required**
- **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**

Typical PFS usage (1): Cluster file system

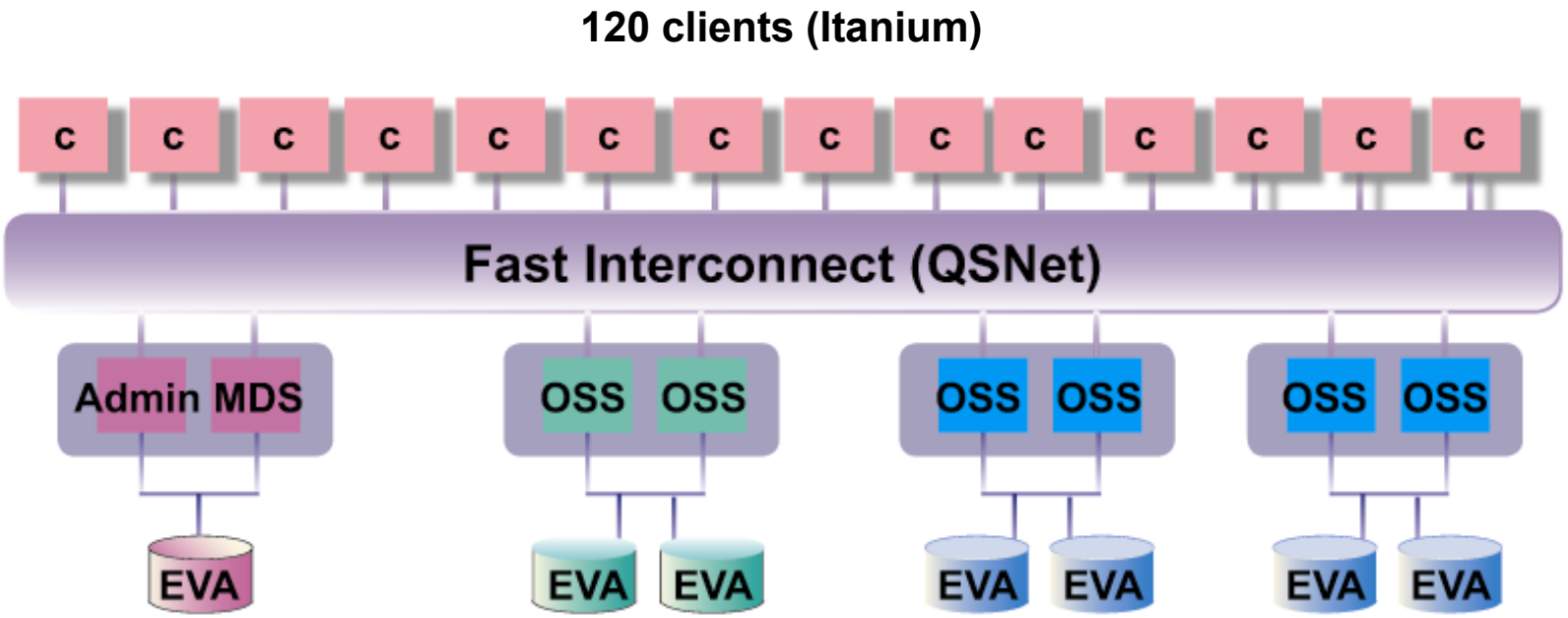


Cluster nodes

Interconnect or SAN

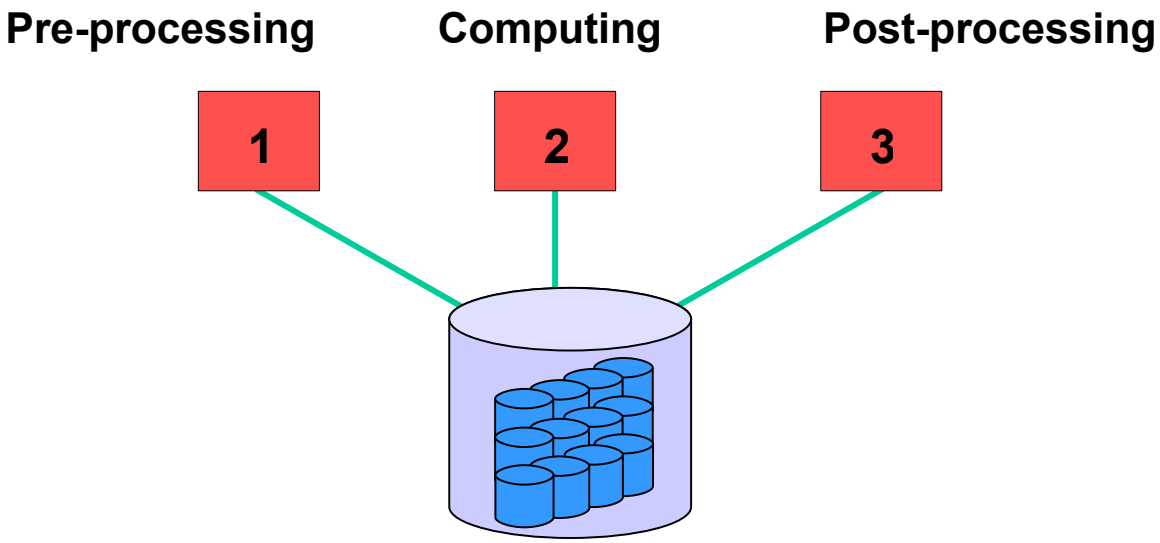
- File system and cluster usually from same vendor
  - Good parallel file system is important for cluster selection
- Benefit is increased throughput, scalability and easy usability

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



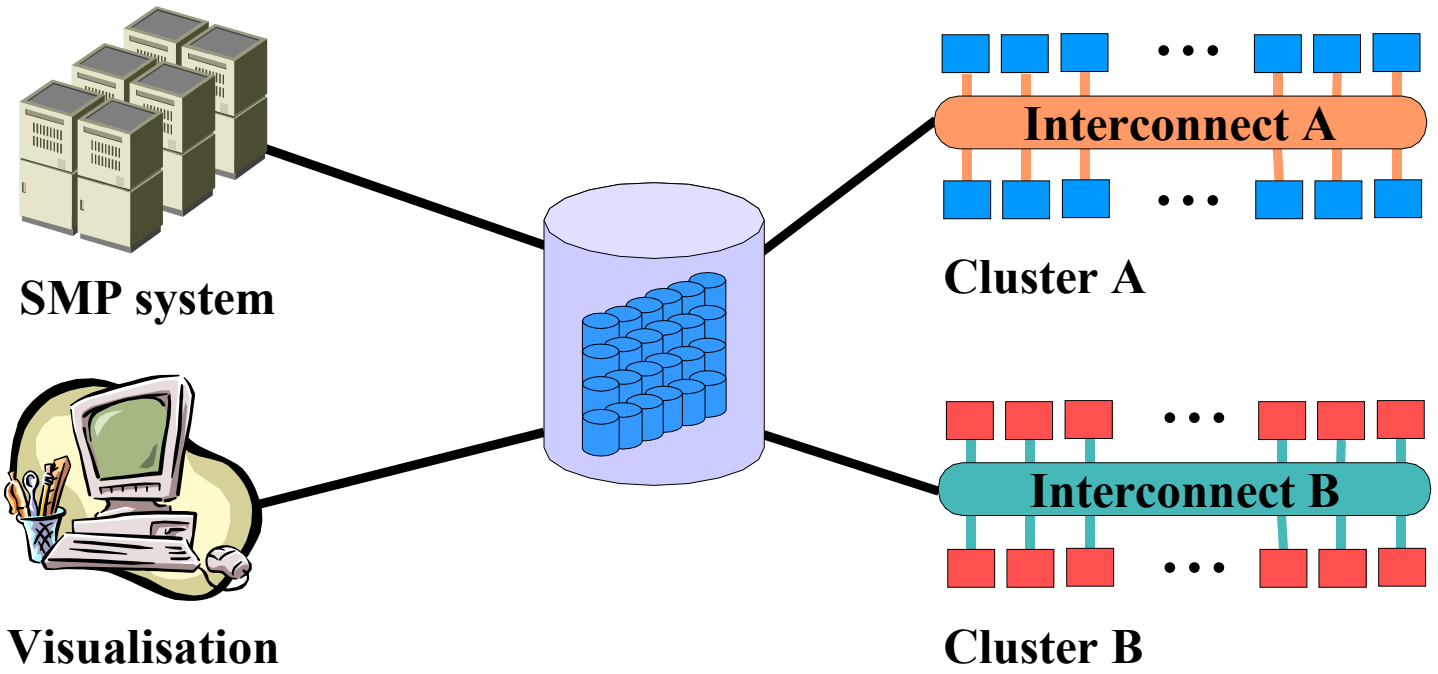
	$HOME	$WORK
Capacity	3.8 TB	7.6 TB
Write performance	240 MB/s	480 MB/s
Read performance	380 MB/s	760 MB/s

Typical PFS usage (2): Workflow file system



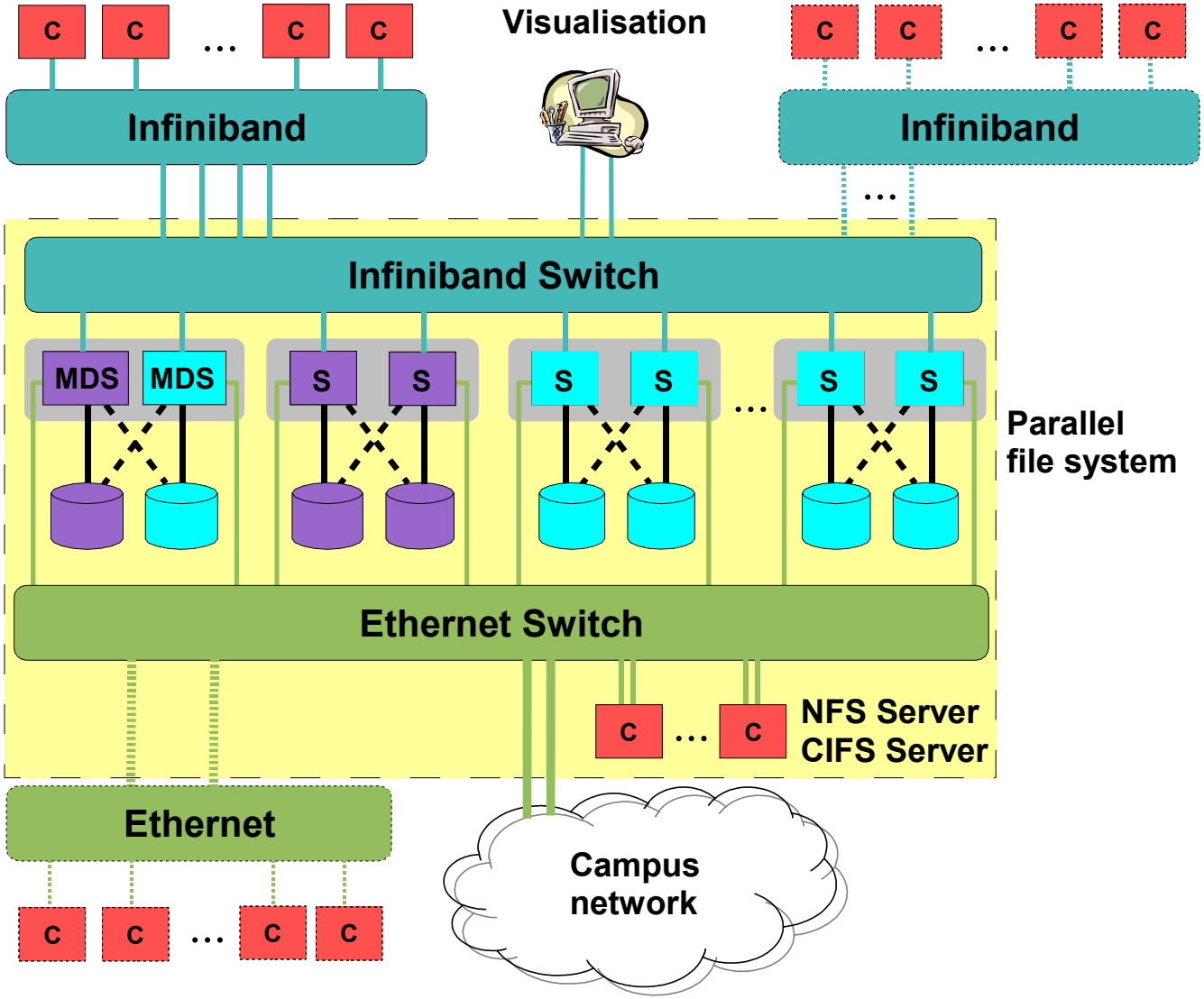
- Typical customers: Oil & gas, digital media
- Usually moderate number of heterogeneous clients
  - SAN based PFS are used in most cases
- Benefit is accelerated workflow and easy usability

Typical PFS usage (3): Global file system



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

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



PFS features (1)

- **Reliability and resilience**
  - Number of installed systems of similar size
    - Expect software problems if file system is new
  - Quality of software support
  - No single point of hardware failure
  - Failover support of software
  - Replication of data and metadata
  - Supported RAID levels, e.g. RAID6
- **Scalability**
  - Number of supported clients
    - SAN based file systems are often limited to 100 clients
- **Performance**
  - Metadata, throughput and sometimes even random access
  - Guaranteed bandwidth

PFS features (2)

- **Costs**
  - Complexity of administration
  - Software and maintenance
  - Requirement for special or additional hardware
  - Supported storage subsystems and disks
- **Sustainability**
  - Financial status of vendor
  - Number of installed systems
  - Open source
- **Heterogeneity**
  - Supported kernel and operating system versions
    - SAN based file systems often support more operating systems
  - Available gateway solutions, e.g. NFS or CIFS export

PFS features (3)

- **Application interface**
  - **Full POSIX support**
  - **Support for memory mapped files**
  - **Support for MPI-IO**
- **Network options**
  - **Supported networks, protocols and speed**
    - **Examples: GigE, 10 GigE, 4x DDR InfiniBand, 4 Gb FC, iSCSI**
  - **Support for multiple network connections**
  - **Supported network adapters**
- **Security**
  - **Strong user level authentication and authorization, e.g. with Kerberos**
  - **Strong authentication for administrative tasks, e.g. file system mount**
  - **Support for data encryption**

PFS features (4)

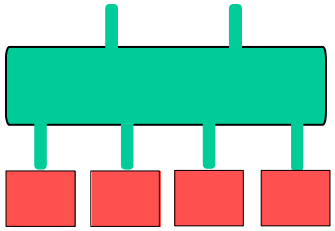
- **Backup**
  - **Backup and restore of user data**
    - **Snapshots might be sufficient**
  - **Backup and restore of complete file systems**
    - **Very fast or parallel restore is required**
    - **Snapshots help to create consistent backup**
  - **Separate important and scratch data**
    - **Quota support helps to limit amount of data**
- **HSM support**
  - **Usually a PFS supports only a dedicated HSM system**
  - **Archiving by users is an alternative to HSM**
- **Administration support**
  - **Performance and health monitoring**
  - **Good documentation**

Impact of blade systems on parallel file systems (1)

- **Limited number or small form factor of adapters**
  - **FC adapters possibly not supported or with limited throughput**
    - **SAN based file systems not possible or with reduced performance**
  - **Fast network adapters possibly unsupported or with limited throughput**
    - **Network based file systems with reduced performance**
    - **Separate GigE for SAN based file systems possibly not available**
  - **Number of storage adapters is restricted**
    - **Servers have to be outside the blade system**
- **Usually servers have to be outside the blade**
  - **i.e. servers must be managed separately**
    - **Affects only metadata servers for SAN based file systems**
    - **Some solutions provide black box anyway, e.g. HP SFS**
    - **Required anyway if PFS is used as global file system**

Impact of blade systems on parallel file systems (2)

- **Integrated FC or network switch uplinks might be oversubscribed**
  - **Either use network option with full bandwidth**
    - **External switches are necessary**
  - **Or use network option with limited throughput**
    - **This might not be supported by the parallel file system**



- **Usually huge number of small clients**
  - **Increased costs for large SANs**
  - **Support for lots of clients is required**
    - **SAN based file systems might not be possible solution**
- **Alternatives for SAN based parallel file systems**
  - **Use storage communication over interconnect if supported**
    - **iSCSI over IP on any type of interconnect**
    - **iSER (iSCSI Extensions for RDMA), e.g. over Infiniband**
    - **SRP (SCSI RDMA Protocol) over Infiniband**

Selecting a PFS

- **Depends on your environment**
  - Number and type of clients
  - Security policy and network environment
  - Available budget and administration staff
  - Existing backup or HSM solutions
  - Blade systems may extend the requirement list
- **Depends on your applications**
  - Requirements for throughput and metadata performance
  - Required application interfaces
- **Depends on other PFS criteria**
  - Reliability
  - Sustainability

Further information (1)

- **Lustre**
  - **Production experiences with HP SFS at SSCK**
    - **http://www.rz.uni-karlsruhe.de/dienste/lustretalks**
  - **HP SFS – a Lustre appliance from HP**
    - **http://www.hp.com/techservers/products/sfs.html**
  - **Roadmap, FAQs and source code from Cluster Filesystems Inc.**
    - **http://www.clusterfs.com/**
- **IBM GPFS documentation**
  - **http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.gpfs.doc/gpfsbooks.html**
    - **Concepts, Planning, and Installation Guide provides good introduction**
- **Panasas ActiveScale Storage Cluster**
  - **http://www.panasas.com/products_overview.html**

Further information (2)

- **ADIC SNFS**
  - **http://www.adic.com/stornext**
- **SGI CXFS**
  - **http://www.sgi.com/products/storage/tech/file_systems.html**
- **Red Hat GFS**
  - **http://www.redhat.com/software/rha/gfs/**