Storage Strategy for HPC Users

Philipp Martin

IT Center

Agenda

- I/O in HPC
  - Parallel Filesystems
  - Striping
- RWTH File Systems: Overview
  - Architecture: Big Picture
  - $HOME & $WORK
  - $HPCWORK
  - $BEEOND
- RWTH File Systems: Best Practices
  - Usage Guidelines

Storage Parameters

- Performance
  - Bandwidth [GB/s]: How quickly can I move raw bytes?
  - Metadata [IOPS]: How quickly can I perform file operations?

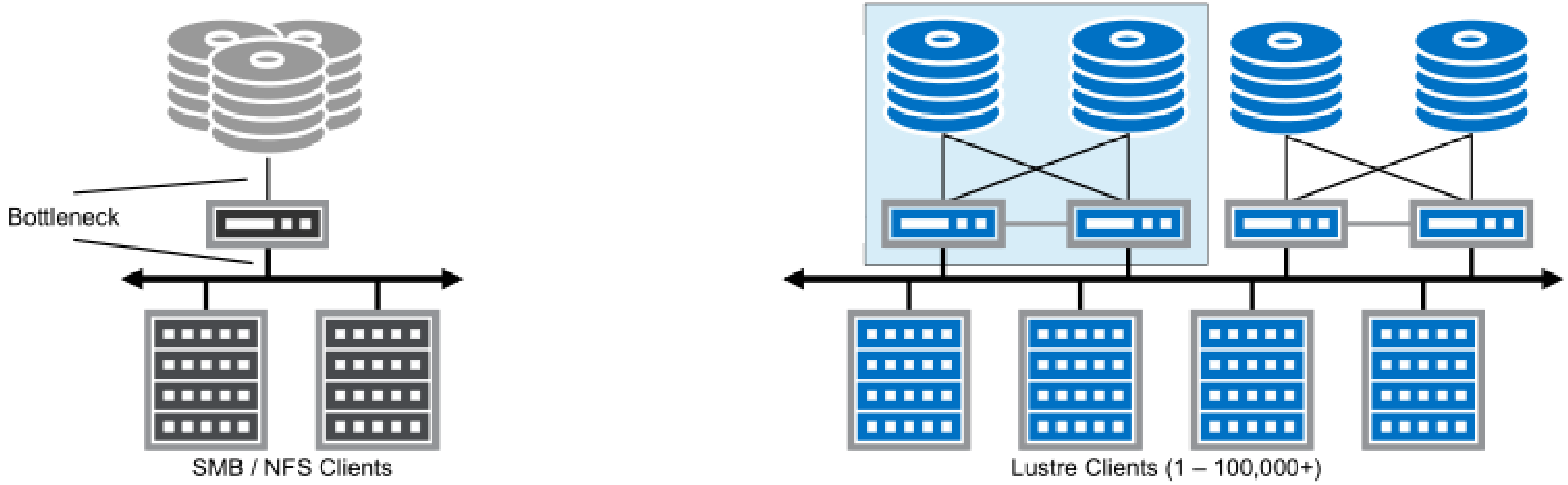
Storage Parameters

- Performance
  - Bandwidth [GB/s]: How quickly can I move raw bytes?
  - Metadata [IOPS]: How quickly can I perform file operations?
- Reliability
  - Uptime: How often is the system unreachable?
  - Snapshots: Protection against accidental deletion
  - Backups: Protection against system failures

Storage Parameters

- Performance
  - Bandwidth [GB/s]: How quickly can I move raw bytes?
  - Metadata [IOPS]: How quickly can I perform file operations?
- Reliability
  - Uptime: How often is the system unreachable?
  - Snapshots: Protection against accidental deletion
  - Backups: Protection against system failures
- Capacity
  - Total size in bytes
  - Total number of files

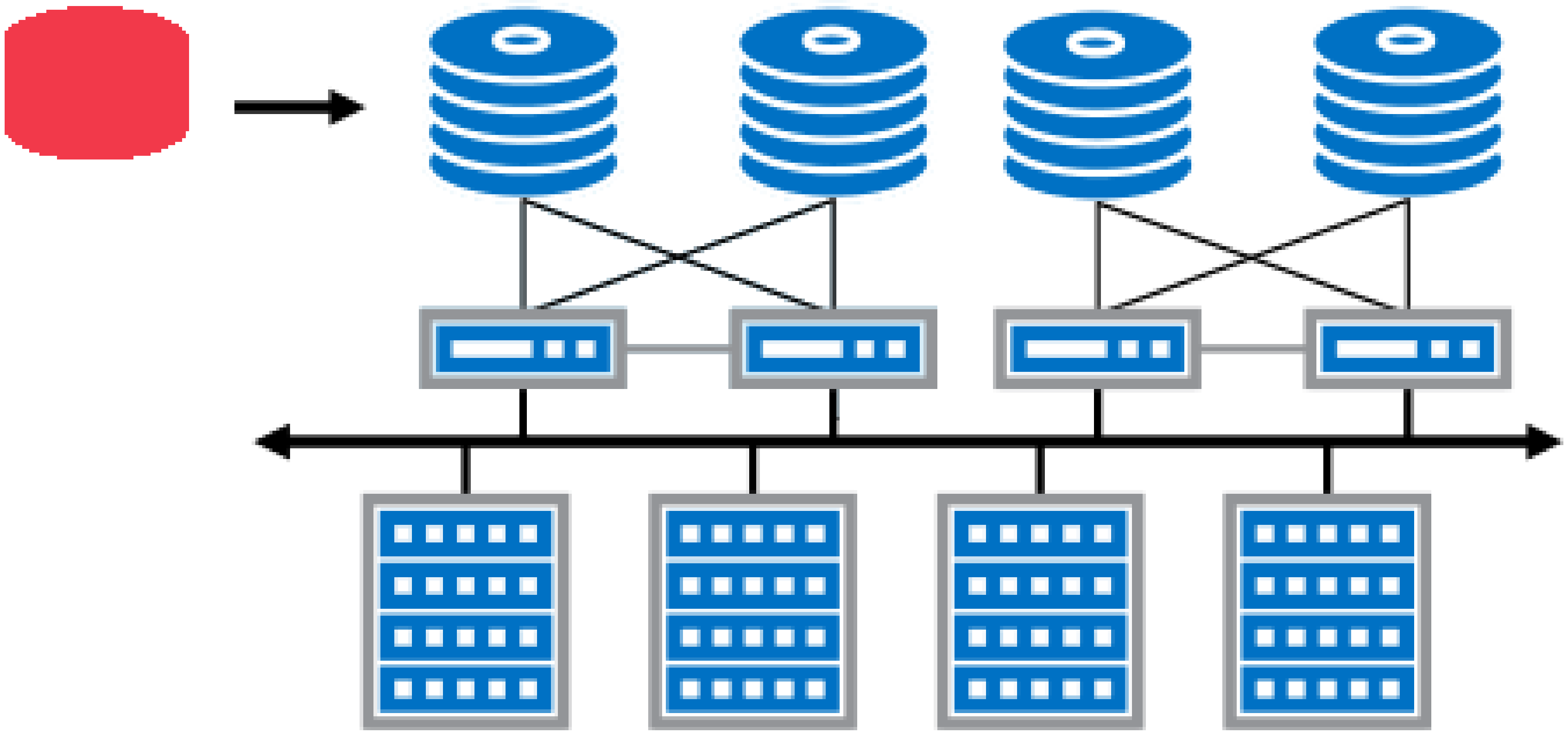
Parallel Filesystems



Taken from: https://wiki.lustre.org/images/6/64/LustreArchitecture-v4.pdf

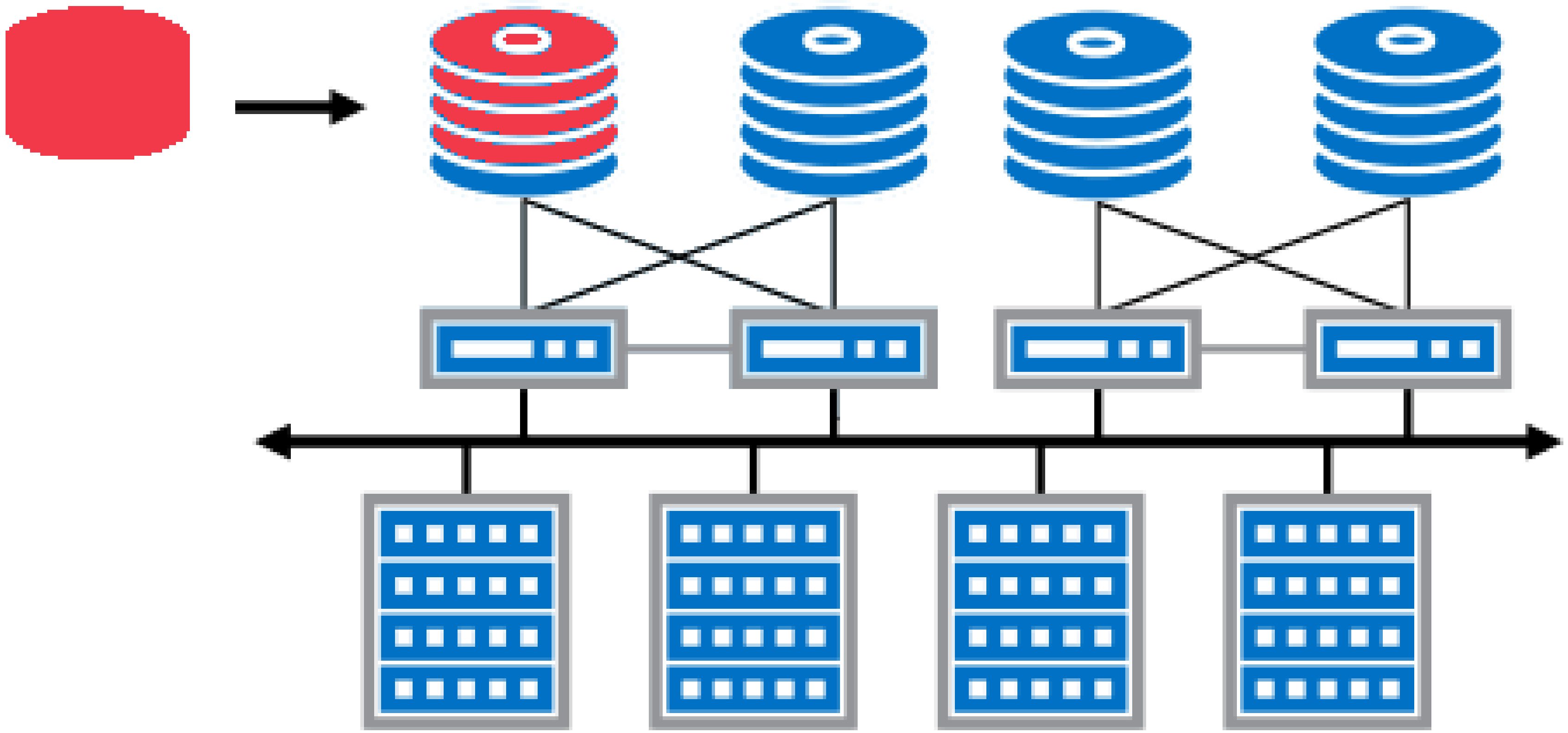
I/O in HPC

Striping



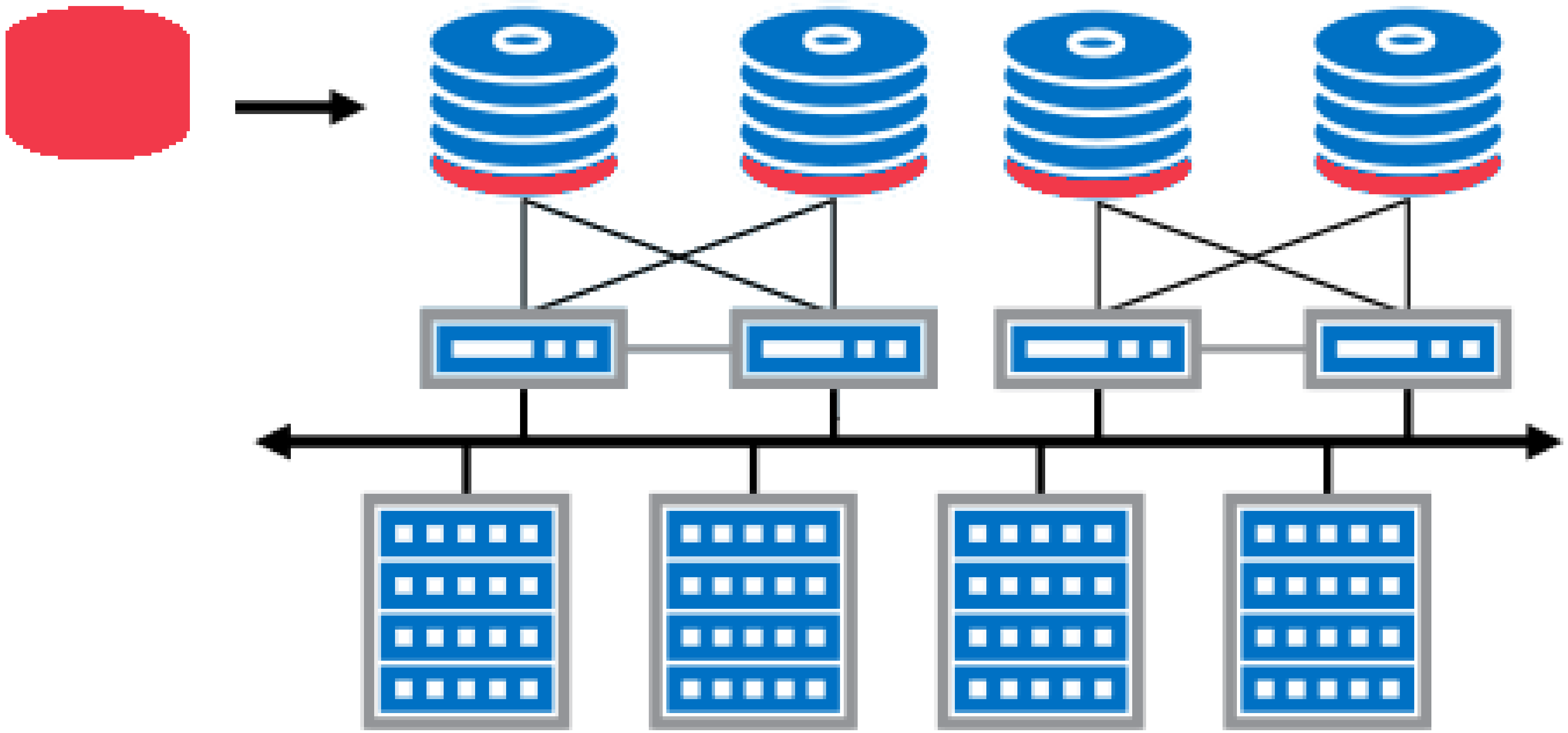
I/O in HPC

Striping



I/O in HPC

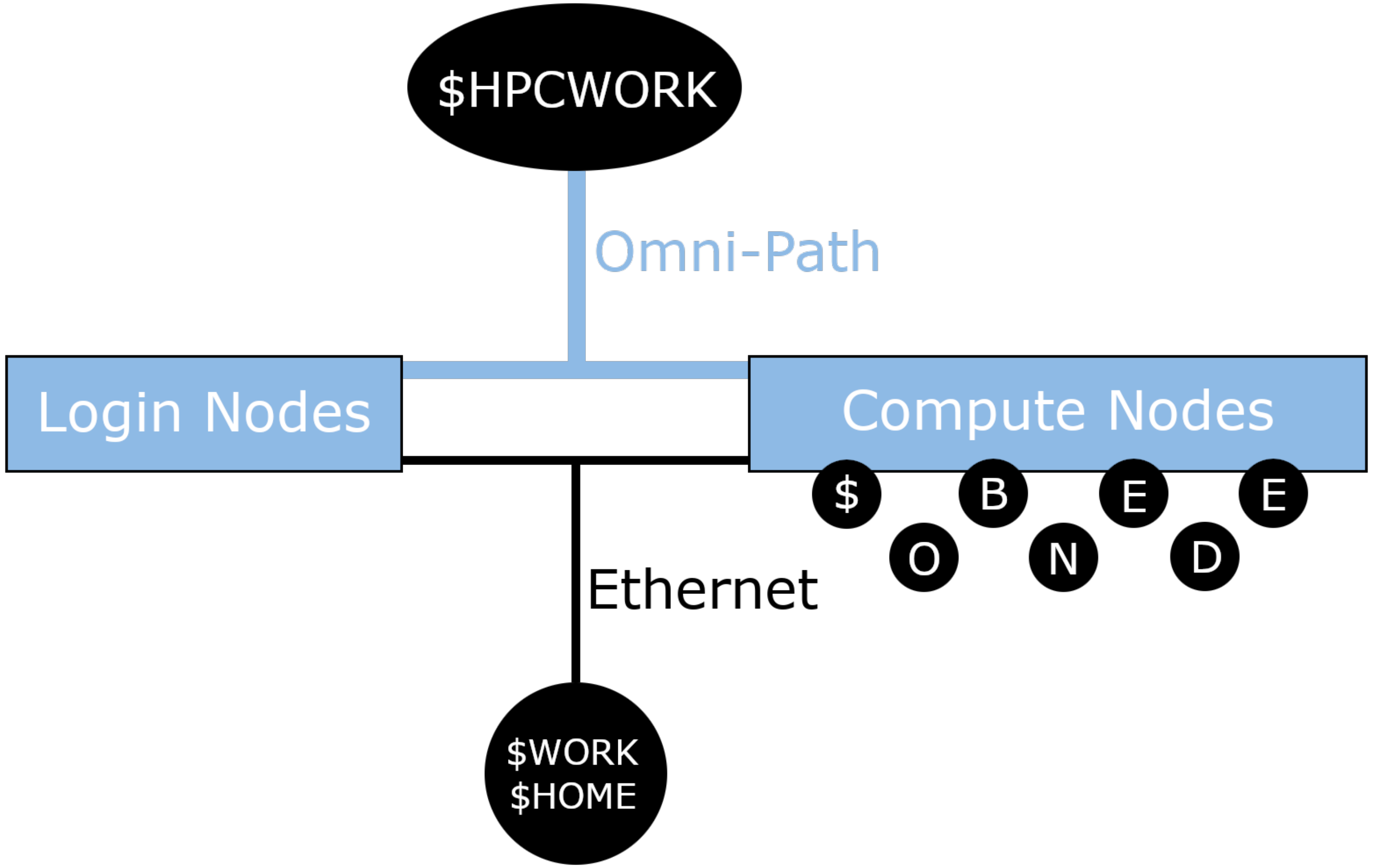
Striping



Agenda

- I/O in HPC
  - Parallel Filesystems
  - Striping
- RWTH File Systems: Overview
  - Architecture: Big Picture
  - $HOME & $WORK
  - $HPCWORK
  - $BEEOND
- RWTH File Systems: Best Practices
  - Usage Guidelines

Big Picture

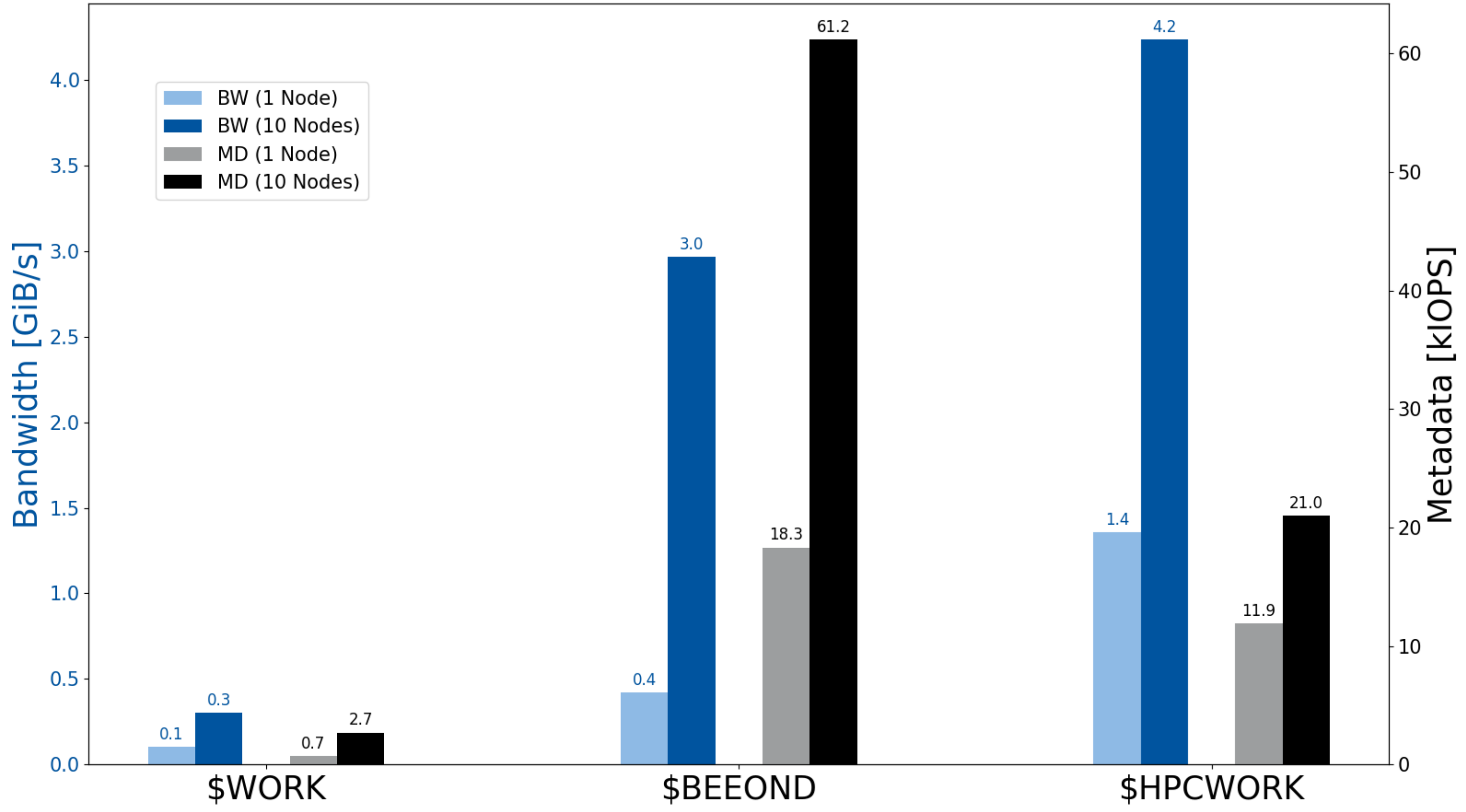

$HPCWORK
Omni-Path
Login Nodes
Compute Nodes
$
B
E
E
O
N
D
Ethernet
$WORK
$HOME

RWTH File Systems

Overview - File System Summary

Access	File System	Cap. Quota	File Quota	Backup	Pros	Cons
`$HOME`	NFS	100 GB	-	Tape (off-site)	- reliable - backup	- limited bw. - limited quota
`$WORK`	NFS	250 GB	-	Snapshots	- reliable	- limited bw.
`$HPCWORK`	Lustre	1000 GB	50 000	None	- bandwidth - capacity	- less reliable
`$BEEOND`	BeeGFS	-	-	None	- metadata - bandwidth	- temporary - memory usage

Overview - IO500 Benchmark Results



$HOME & $WORK

- Use NFS over Ethernet
- Limited bandwidth, but very reliable
- $HOME is backed up to tape
  - $HOME should be used for configuration files and data that cannot be recovered otherwise
  - $WORK should be used for applications that are light on I/O

RWTH File Systems

$HPCWORK

- Lustre parallel file system
- Access with RDMA network (Intel Omnipath)
- High bandwidth, but limited metadata performance
- Not optimized for reliability
  - Use for applications that use few large files

RWTH File Systems

$BEEOND

- File system that uses the local SSDs of the compute nodes
- Is **temporary**, i.e. any data that hasn't been copied somewhere else is lost when the job ends
  - Request via `#SBATCH --beeond`
  - Staging currently in progress, likely to be rolled out by end of March

$BEEOND

- File system that uses the local SSDs of the compute nodes
- Is **temporary**, i.e. any data that hasn't been copied somewhere else is lost when the job ends
  - Request via `#SBATCH --beeond`
  - Staging currently in progress, likely to be rolled out by end of March
- Use for applications that use many small files
- Use for jobs that use a moderate number of nodes (< 50)

Agenda

- I/O in HPC
  - Parallel Filesystems
  - Striping
- RWTH File Systems: Overview
  - Architecture: Big Picture
  - $HOME & $WORK
  - $HPCWORK
  - $BEEOND
- RWTH File Systems: Best Practices
  - Usage Guidelines

Usage Guidelines

- Use $HOME to store important data or results, not for computation
  - Limited performance and space
  - Backups (are costly)
  - We have to enforce 100 GB quotas

Usage Guidelines

- Use $HOME to store important data or results, not for computation
  - Limited performance and space
  - Backups (are costly)
  - We have to enforce 100 GB quotas
- If possible, evaluate $BEEOND
  - Decent bandwidth and good metadata performance
  - Warning: temporary (per job) storage!

Usage Guidelines

- Use $HOME to store important data or results, not for computation
  - Limited performance and space
  - Backups (are costly)
  - We have to enforce 100 GB quotas
- If possible, evaluate $BEEOND
  - Decent bandwidth and good metadata performance
  - Warning: temporary (per job) storage!
- When in doubt, use $WORK or $HPCWORK
  - Simple to use
  - $WORK: supports large number of small files
  - $HPCWORK: good bandwidth performance

Usage Guidelines

- Use $HOME to store important data or results, not for computation
  - Limited performance and space
  - Backups (are costly)
  - We have to enforce 100 GB quotas
- If possible, evaluate $BEEOND
  - Decent bandwidth and good metadata performance
  - Warning: temporary (per job) storage!
- When in doubt, use $WORK or $HPCWORK
  - Simple to use
  - $WORK: supports large number of small files
  - $HPCWORK: good bandwidth performance
- Use data transfer nodes for transferring large amounts of data
  - `{copy,copy18-1,copy18-2}.hpc.itc.rwth-aachen.de`

Questions

Questions?

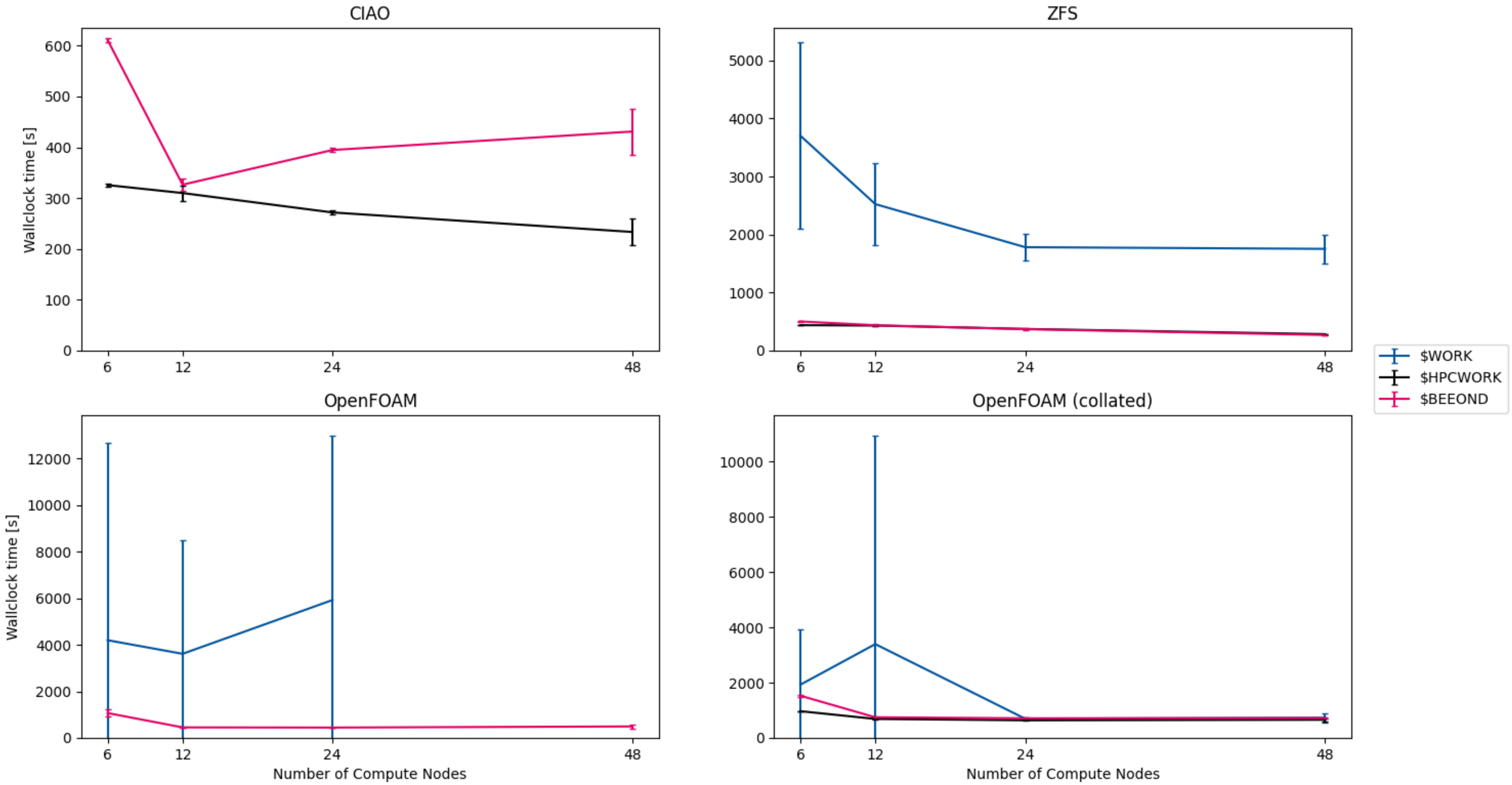
Backup

Backup

Overview - IO500 Benchmark Results

- The IO500 is a benchmark designed to test the I/O capabilities of High-Performance Computing systems
- It uses several different scenarios to test both best and worst cases for bandwidth and metadata performance
- The results are averaged geometrically

Benchmarks


CIAO
ZFS
OpenFOAM
OpenFOAM (collated)
Wallclock time [s]
Wallclock time [s]
Number of Compute Nodes
Number of Compute Nodes
600
500
400
300
200
100
0
5000
4000
3000
2000
1000
0
12000
10000
8000
6000
4000
2000
0
10000
8000
6000
4000
2000
0
6
12
24
48
$WORK
$HPCWORK
$BEEOND

RWTH File Systems

Parameters

- Isilon ($WORK, $HOME)
  - 15 Nodes
  - Each: 35 HDDs (3 TB, 7200 RPM, SATA) and 1 SSD (1.6 TB, SATA)
  - Total: 1.1 PB Net Capacity, 4 GB/s aggregate bandwidth
- Lustre ($HPCWORK)
  - 10 Units
  - Each: 180 HDDs (8 TB, 7200 RPM, SATA)
  - Total: 9.9 PB Capacity, 150 GB/s aggregate bandwidth
- Local disks ($BEEOND)
  - Per compute node: 1 SSD (480 GB, SATA)

CLAIX 16

- Lustre-16
  - 3 PB Capacity, 50 GB/s aggregate write bandwidth, 35 GB/s aggregate read bandwidth