Storage Strategy for HPC Users

Introduction to High-Performance Computing 2024

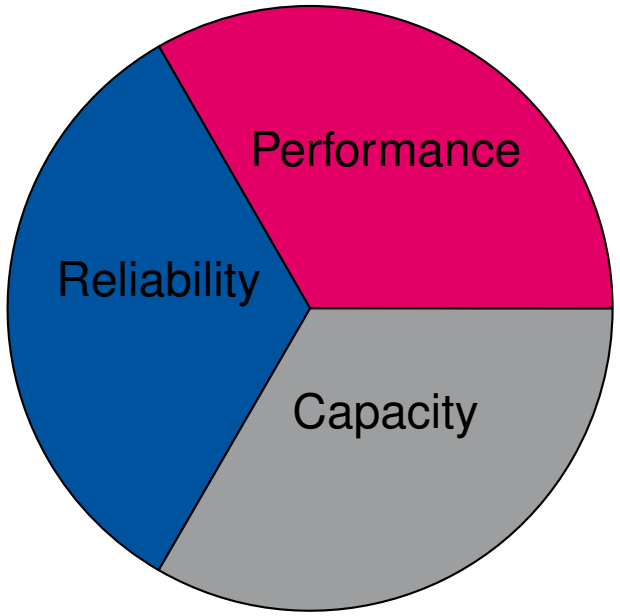
Philipp Martin

IT Center | RWTH AACHEN UNIVERSITY

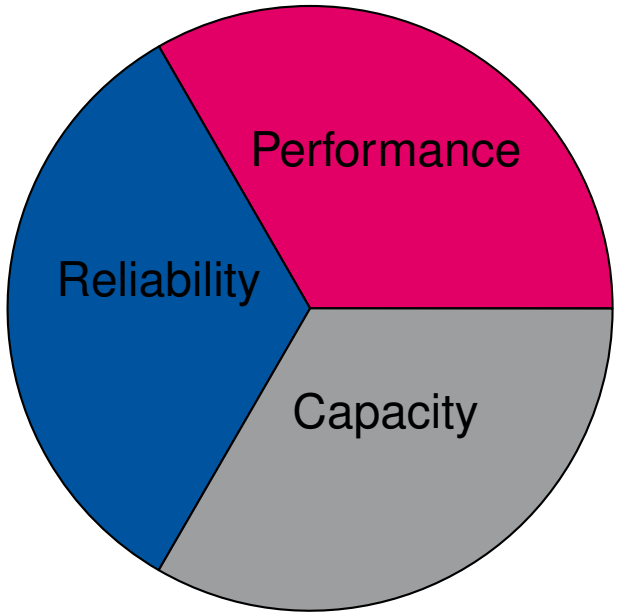
Agenda

- I/O in HPC
  - Why do we need different filesystems?
  - Parallel Filesystem Technology
- I/O on CLAIX-18
  - Overview
  - Usage Guidelines
  - We can help!
- I/O on CLAIX-23: An Outlook

Why do we need different filesystems?

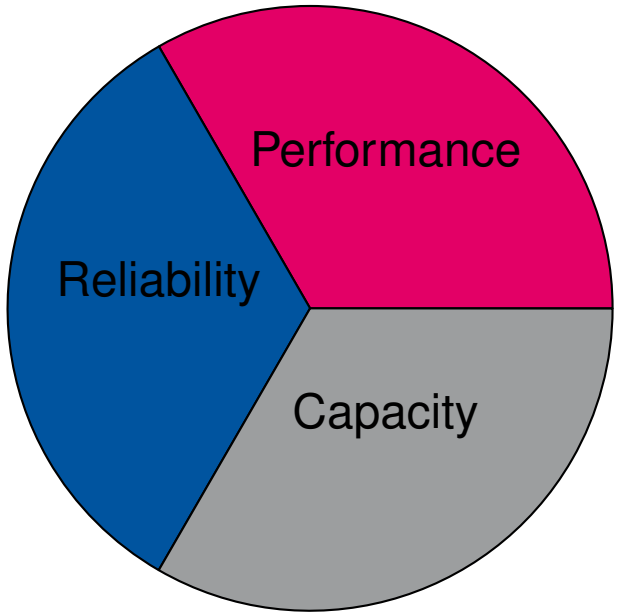


Why do we need different filesystems?



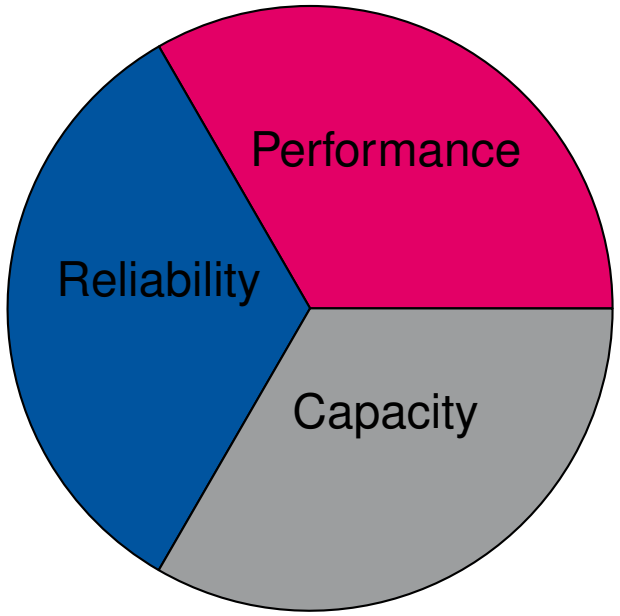
- Performance
  - Bandwidth [GB/s]: How quickly can I move raw bytes?
  - Metadata [IOPS]: How quickly can I perform file operations?
  - Better performance means better hardware

Why do we need different filesystems?



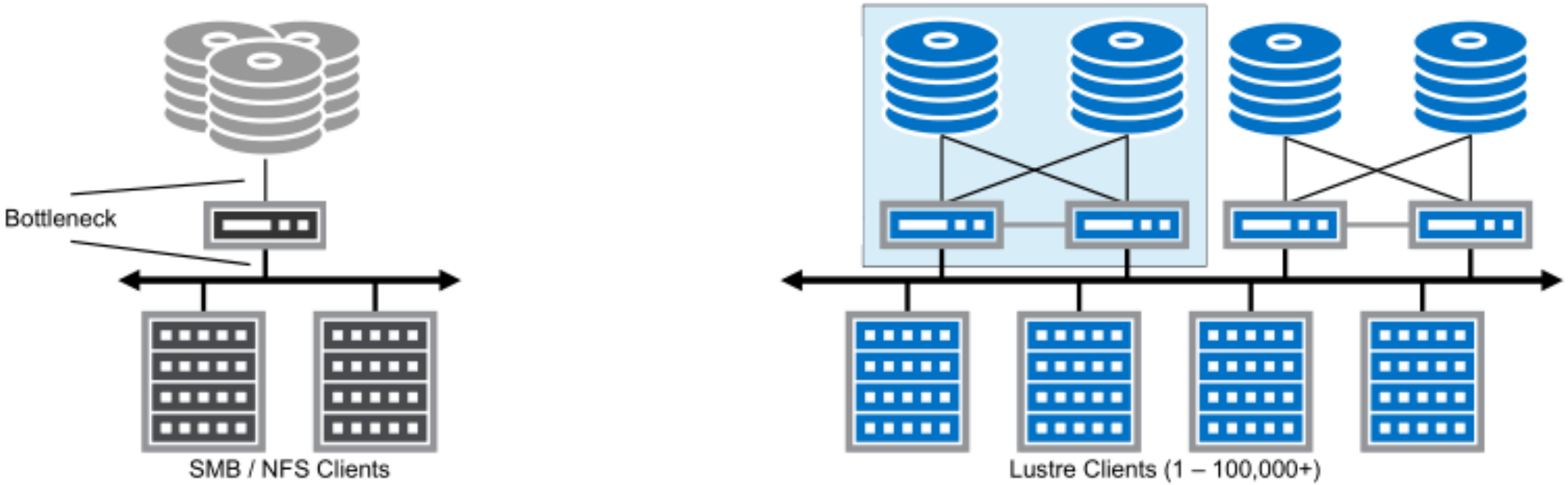
- Performance
  - Bandwidth [GB/s]: How quickly can I move raw bytes?
  - Metadata [IOPS]: How quickly can I perform file operations?
  - Better performance means better hardware
- Reliability
  - Uptime: How often is the system unreachable?
  - Snapshots: Protection against accidental deletion
  - Backups: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

Why do we need different filesystems?



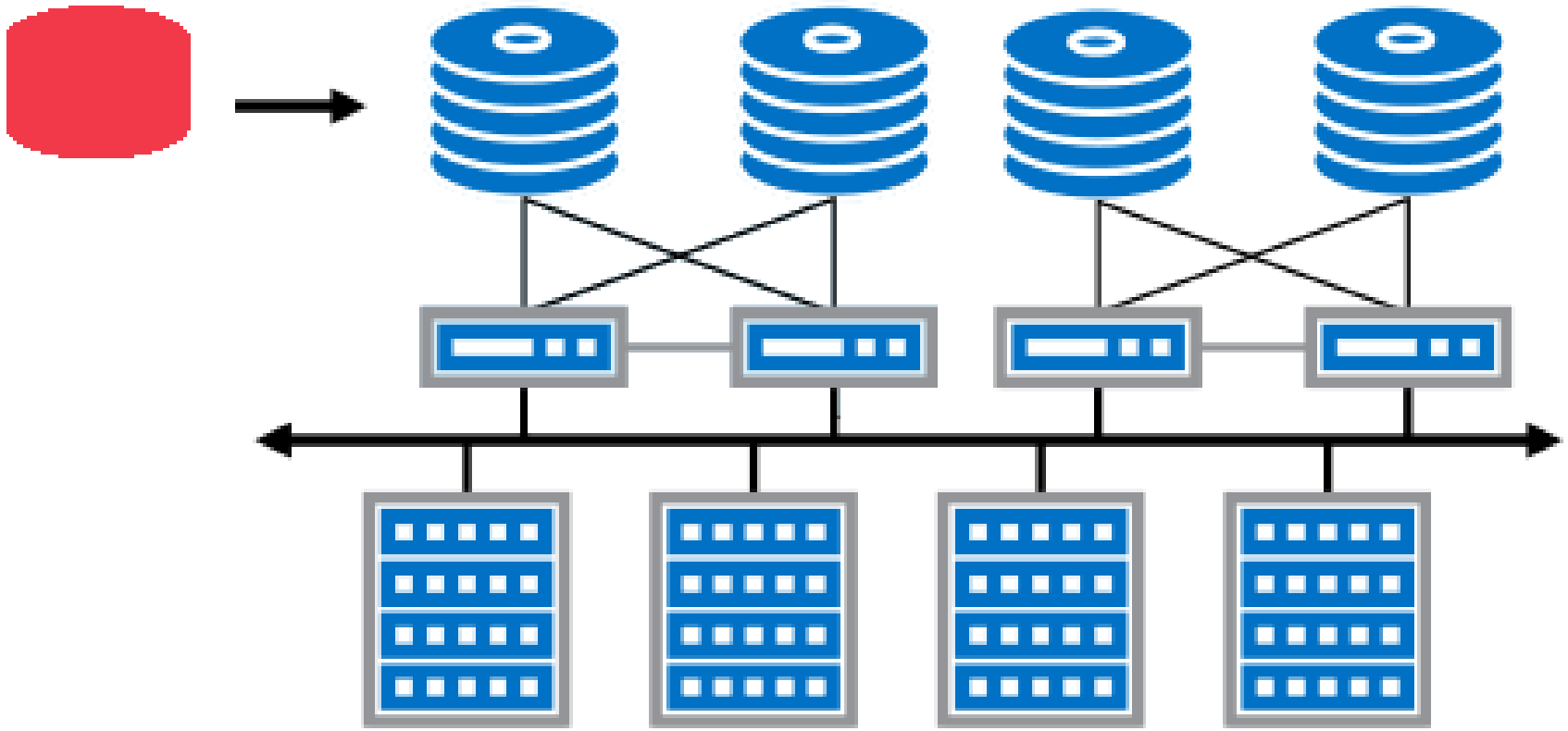
- Performance
  - Bandwidth [GB/s]: How quickly can I move raw bytes?
  - Metadata [IOPS]: How quickly can I perform file operations?
  - Better performance means better hardware
- Reliability
  - Uptime: How often is the system unreachable?
  - Snapshots: Protection against accidental deletion
  - Backups: Protection against system failures
  - Better reliability means redundancies
- Capacity
  - Total size in bytes
  - Total number of files
  - Higher capacities mean more hardware

Parallel Filesystems

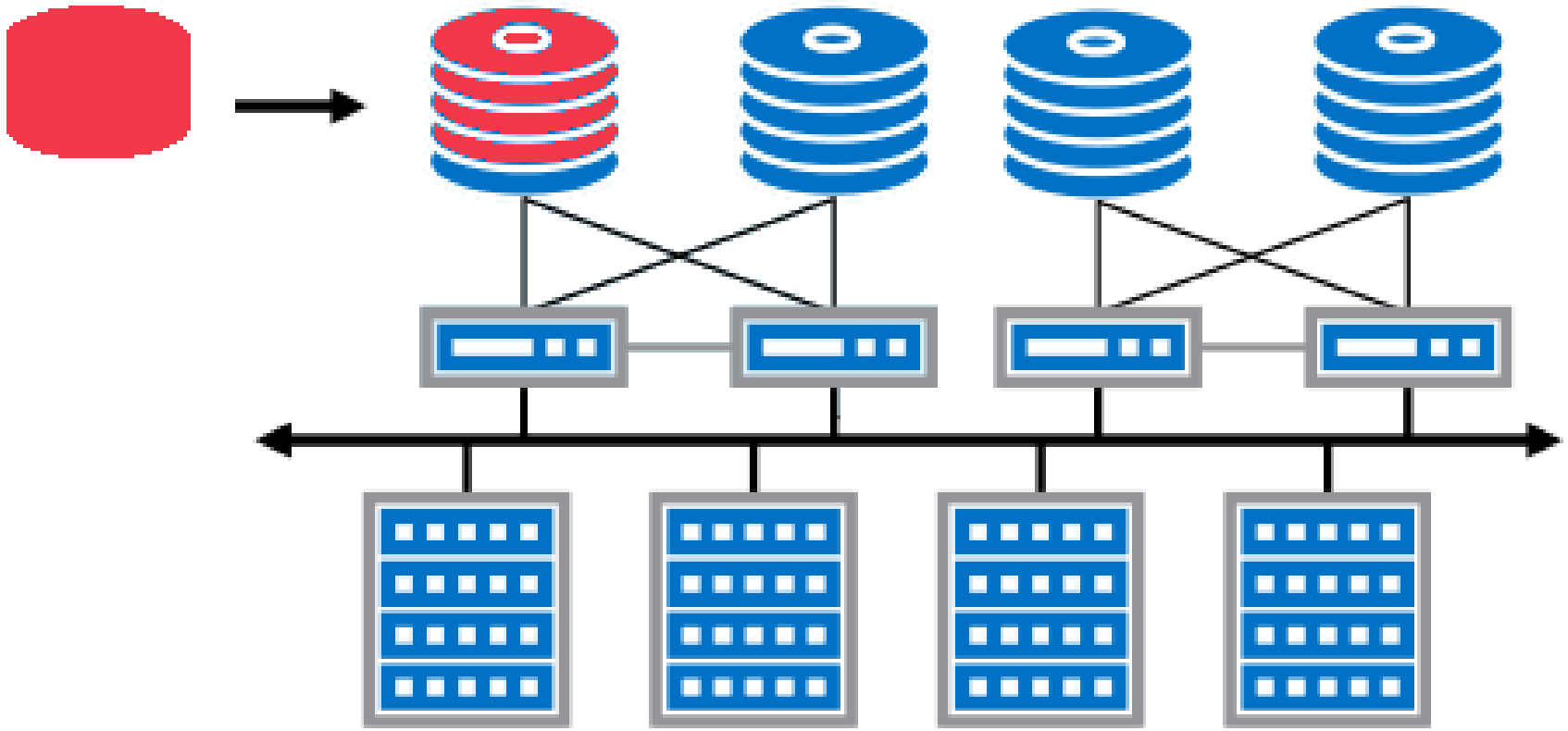


Taken from: https://wiki.lustre.org/images/6/64/LustreArchitecture-v4.pdf

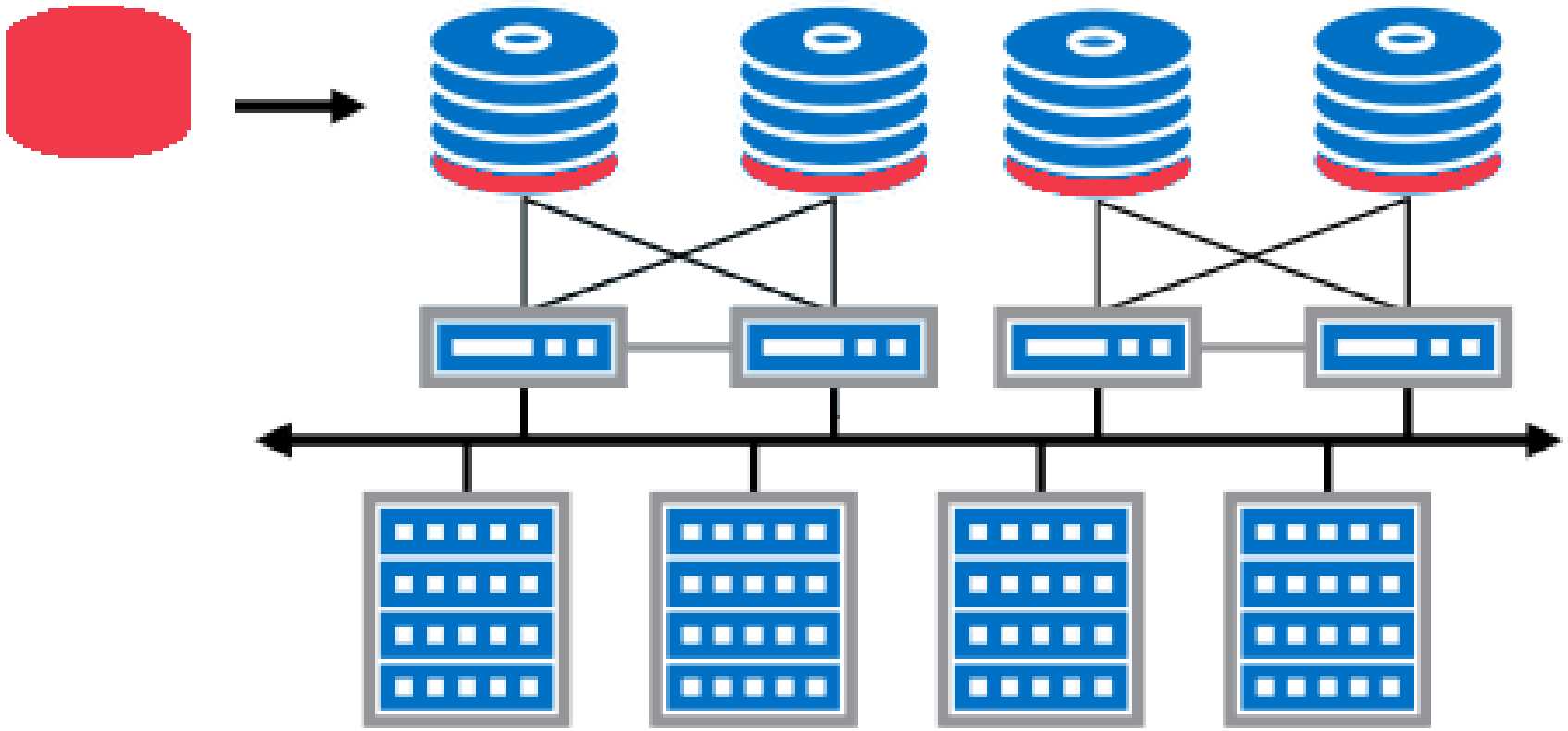
Striping



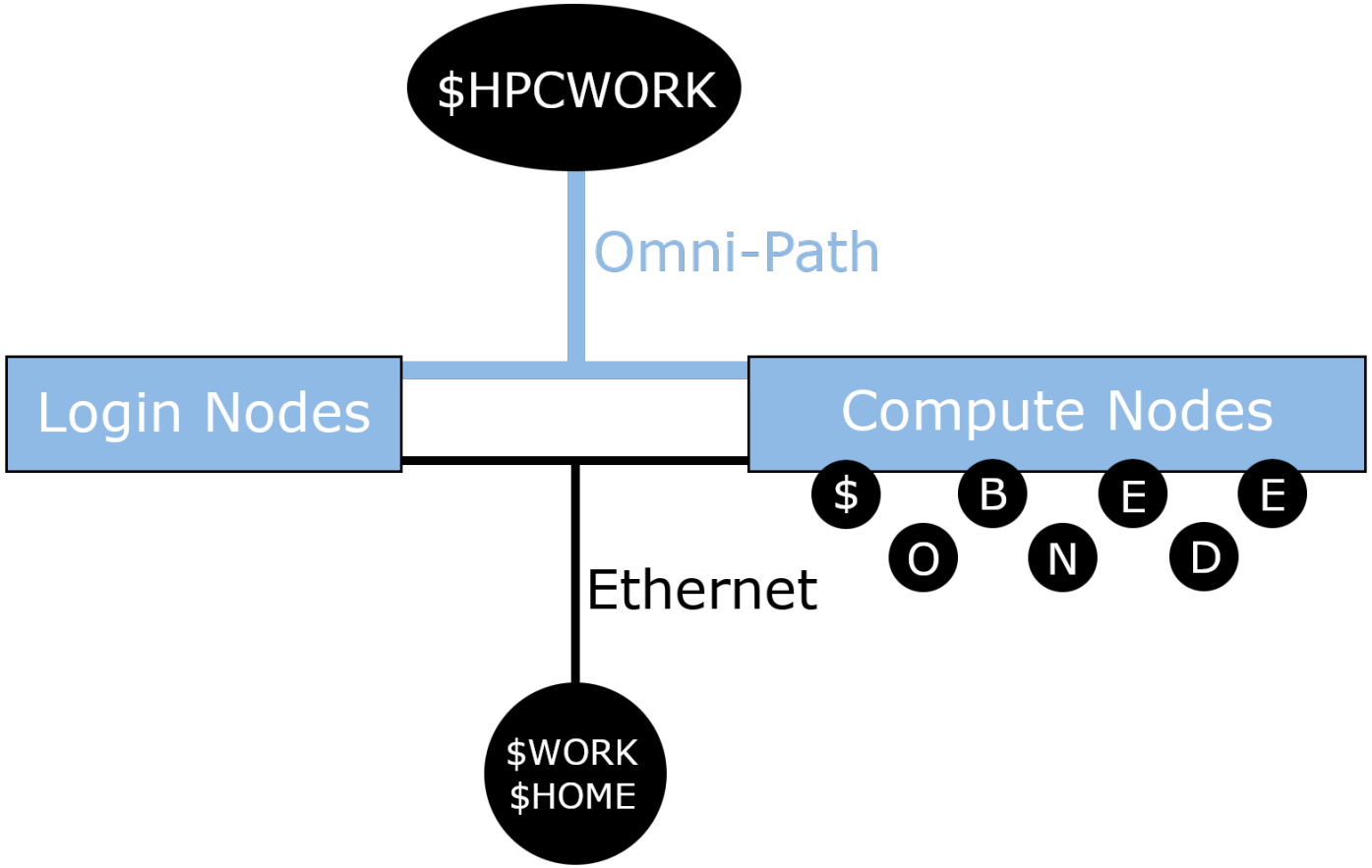
Striping



Striping



Overview

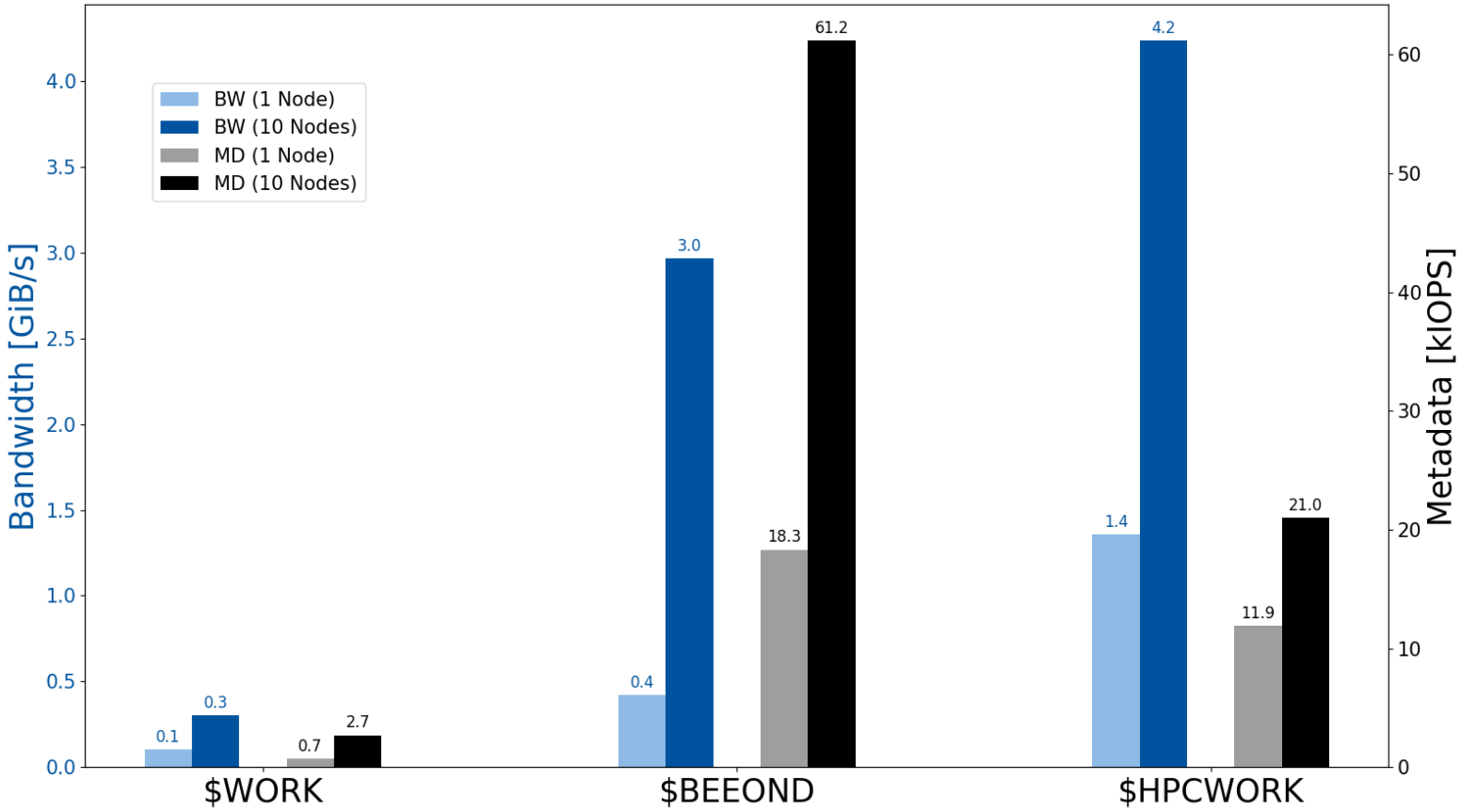

$HPCWORK
Omni-Path
Login Nodes
Compute Nodes
$
B
E
E
O
N
D
Ethernet
$WORK
$HOME

Overview

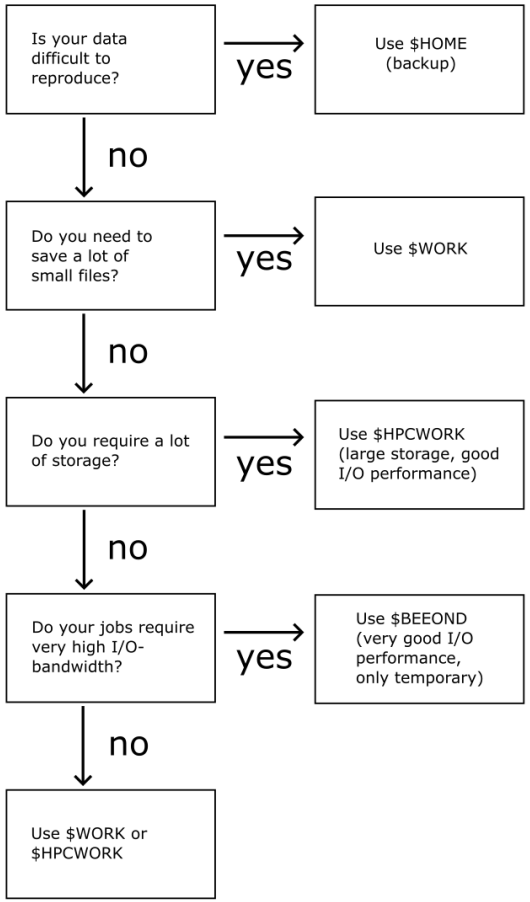
Access	Filesystem	Cap. Quota	File Quota	Backup	Pros	Cons
`$HOME`	NFS	150 GB	-	Tape (off-site)	- reliable - backup	- limited bw. - limited quota
`$WORK`	NFS	250 GB	-	Snapshots	- reliable	- limited bw.
`$HPCWORK`	Lustre	1000 GB	50 000	None	- bandwidth - capacity	- less reliable
`$BEEOND`	BeeGFS	400 GB p.N.	-	None	- metadata - bandwidth	- temporary
`$TMP`	XFS	400 GB p.N.	-	None	- metadata - bandwidth	- temporary

IO on CLAIX-18

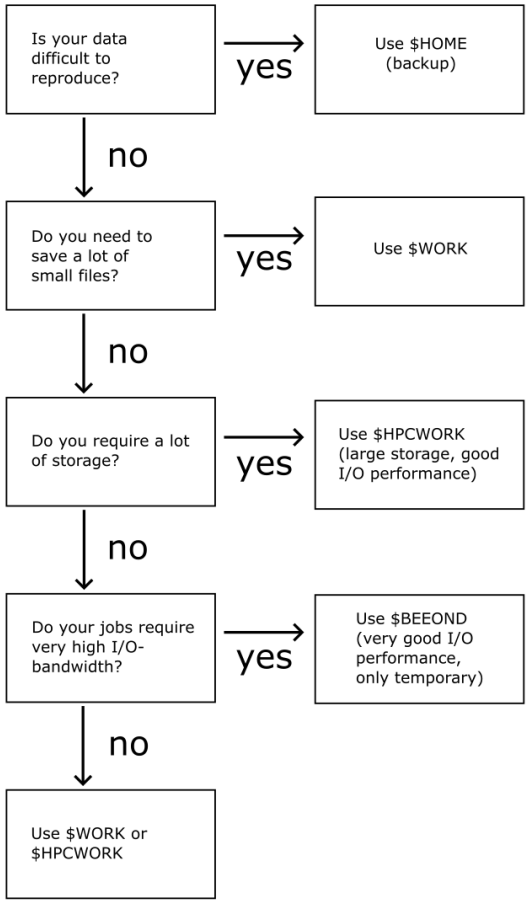
Overview



Filesystem Choice



Filesystem Choice



Additional Considerations

- `$BEEOND` scales with the number of nodes in your job
  - ca. 400 GB per Node
- For large file transfers, use the copy nodes!
  - `{copy,copy18-1,copy18-2}.hpc.itc.rwth-aachen.de`

How to use `$HOME` and `$WORK`

Snapshot Demonstration

How to use `$BEEOND`

Slurm Script Demonstration

What we can help with!

- Poor I/O performance
- Figuring out the correct file system
- Figuring out the correct striping settings etc.
- Open a ticket: servicedesk@itc.rwth-aachen.de

I/O on CLAIX-23: An Outlook

- CLAIX-23 is currently in testing
- New interconnect
- New `$BEEOND`
  - ca. 1.4 TB per Node
  - Performance ca. 2x that of CLAIX-18
- New `$HPCWORK`
  - Migration will happen soon
  - New project creation will also be enabled soon

Questions

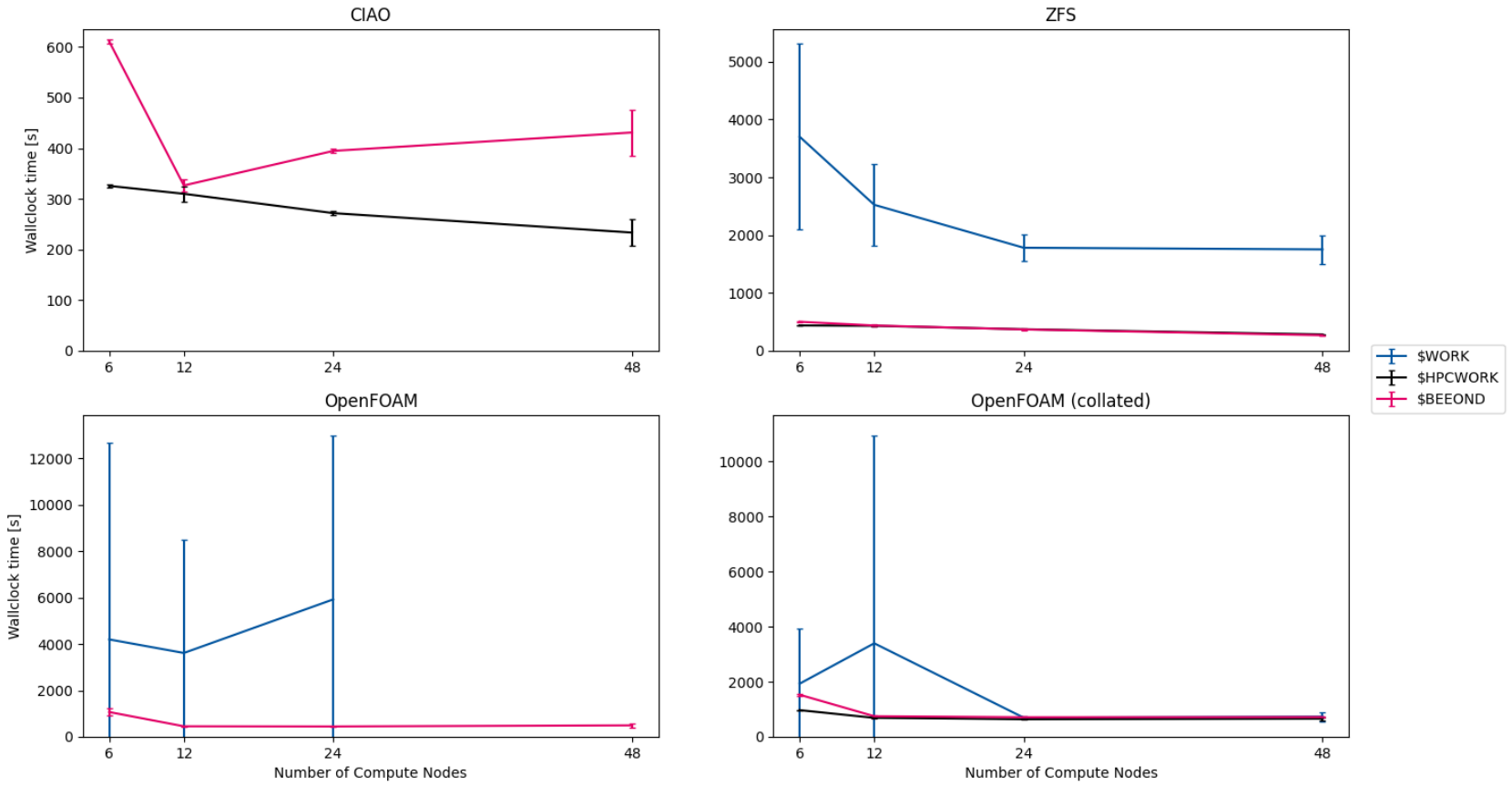
Questions?

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]
Number of Compute Nodes
$WORK
$HPCWORK
$BEEOND

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