HPC.NRW

aiXcelerate 2021: Part I – File I/O

HPC.NRW Competence Network

Overview of I/O Technologies on CLAIX

Philipp Martin (RWTH)

aiXcelerate 2021: Part I – File I/O

Detailed Agenda for Session

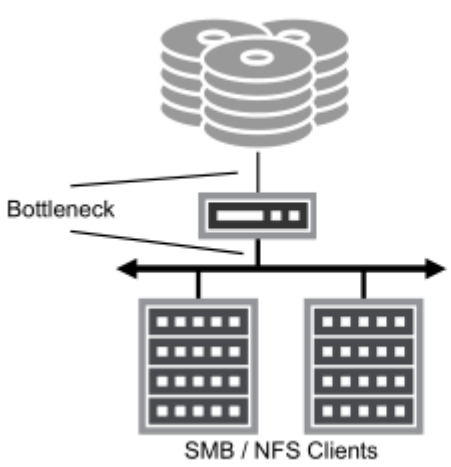
HPC.NRW

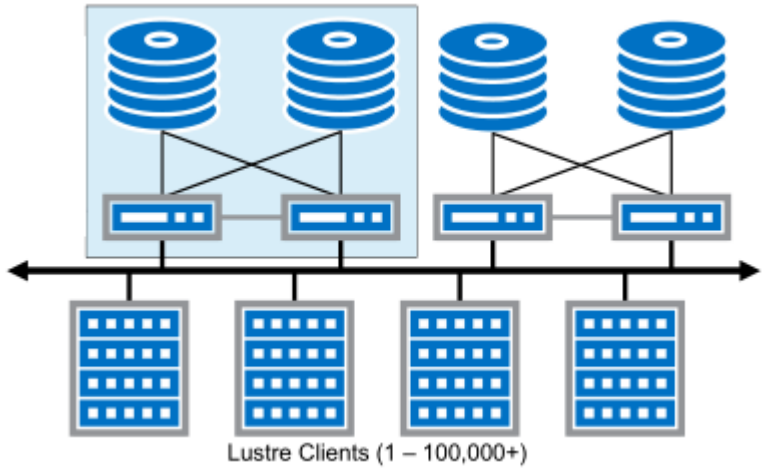
- What is a file system?
  - Hardware
  - Software
- What file systems are available on CLAIX?
  - $HOME
  - $WORK
  - $HPCWORK
  - $BEEOND
  - $TEMP
  - (cvmfs)
- Tuning file system parameters
- Outlook: New file systems in 2022

What is a file system?

HPC.NRW

- "File System" may refer to hardware or software or both
  - Hardware: HDDs vs SSDs (vs NVMEs etc.)
  - HPC hardware basic building blocks are the same as in your computer
  - Software: different solutions for different needs
  - Cluster file system software builds on "normal" software





File Systems on CLAIX

HPC.NRW

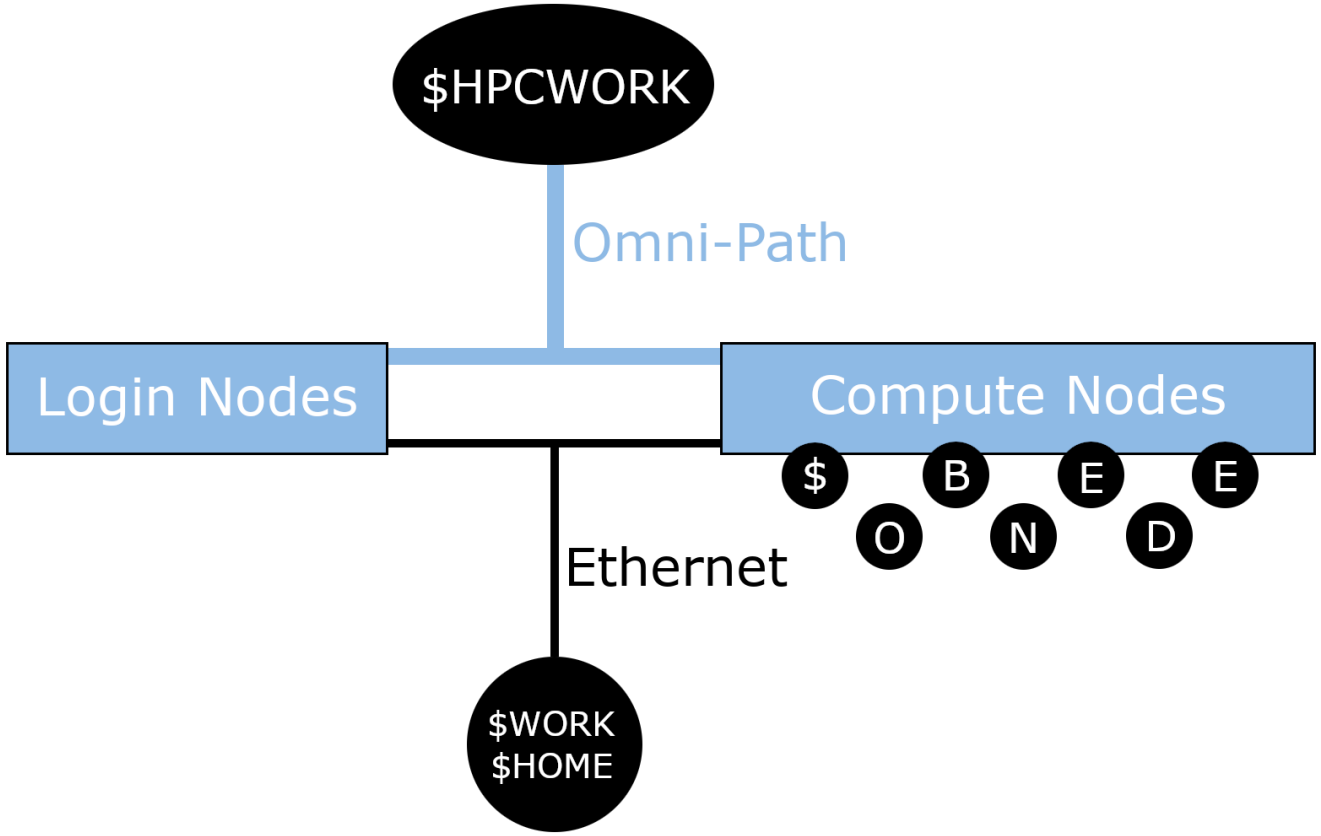
- Parameters to consider
  - Performance
    - Bandwidth
    - Metadata
  - Reliability
    - Uptime
    - Snapshots
    - Backup
  - Capacity
    - Total size in Bytes
    - Total number of files

File Systems on CLAIX

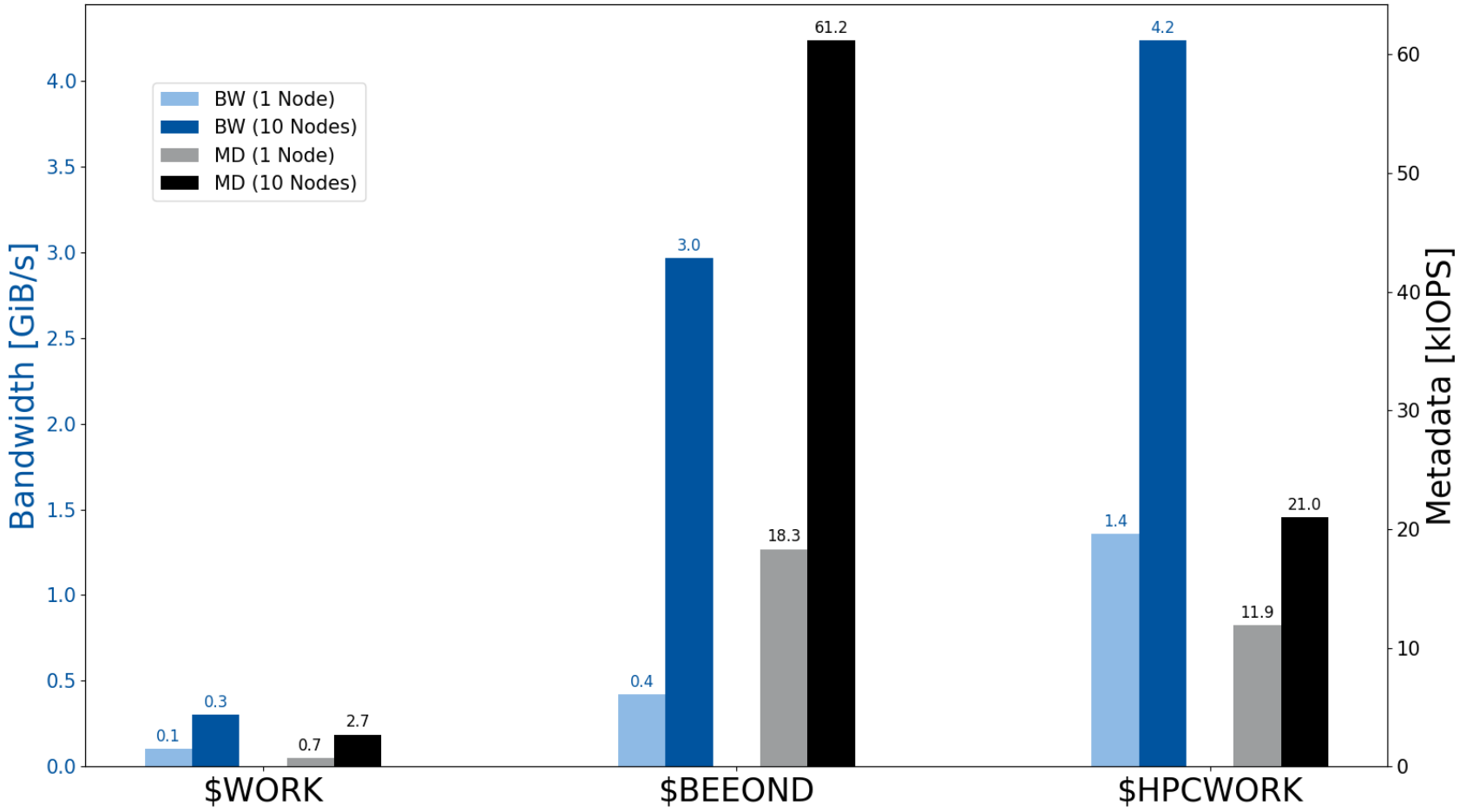
HPC.NRW

Access	Technology	Cap. Quota	File Quota	Backup	Pros	Cons
$HOME	NFS	150 GB	-	Tape (off-site)	-reliable -backup	-limited performance
$WORK	NFS	250 GB	-	Snapshots	-reliable	-limited performance
$HPCWORK	Lustre	1000 GB	50 000	None	-Bandwidth -Capacity	- Less reliable
$BEEOND	BeeGFS	-	-	None	-Performance	-temporary -memory usage
$TMP	XFS	-	-	None	-Performance	-temporary -isolated

File Systems on CLAIX



File Systems on CLAIX



File Systems on CLAIX

HPC.NRW

- $HOME and $WORK
  - Currently use an Isilon System with HDDs
  - Are connected via 1 GB Ethernet and NFS / CIFS
  - Scaled-up version of a home-style NAS
  - Bottlenecked easily
  - Very reliable due to redundancies and simple architecture
  - Capacity Quotas: 150 / 250 GB
  - Backup on tape for $HOME, Snapshots for both
  - New $HOME in 2022

File Systems on CLAIX

HPC.NRW

- $HPCWORK
  - Parallel file system based on Lustre on HDDs
  - Omni-Path network
  - Quota of 1 TB and 50 000 files
  - Architecture uses multiple metadata servers and file servers
  - Very fast for large files
  - Less reliable due to more complicated architecture
  - No backup or snapshot capabilities
  - Different systems for CLAIX-16 and CLAIX-18
  - New $PROJECT in 2022

File Systems on CLAIX

HPC.NRW

- $TMP
  - The local SSDs of each compute node
  - Around 80 GB for CLAIX-16, 400 GB for CLAIX-18
  - Very fast but not connected
- $BEEOND
  - Connects the local SSDs of all current compute nodes
  - Requires exclusive nodes
  - Request with `#SBATCH --beeond`
  - Good Metadata and Bandwidth performance
  - Temporary!
  - Uses compute node resources
  - Unstable above ~50 Nodes
  - Tunable

File Systems on CLAIX

HPC.NRW

Access	Technology	Cap. Quota	File Quota	Backup	Pros	Cons
$HOME	NFS	150 GB	-	Tape (off-site)	-reliable -backup	-limited performance
$WORK	NFS	250 GB	-	Snapshots	-reliable	-limited performance
$HPCWORK	Lustre	1000 GB	50 000	None	-Bandwidth -Capacity	- Less reliable
$BEEOND	BeeGFS	-	-	None	-Performance	-temporary -memory usage
$TMP	XFS	-	-	None	-Performance	-temporary -isolated

File Systems on CLAIX

HPC.NRW

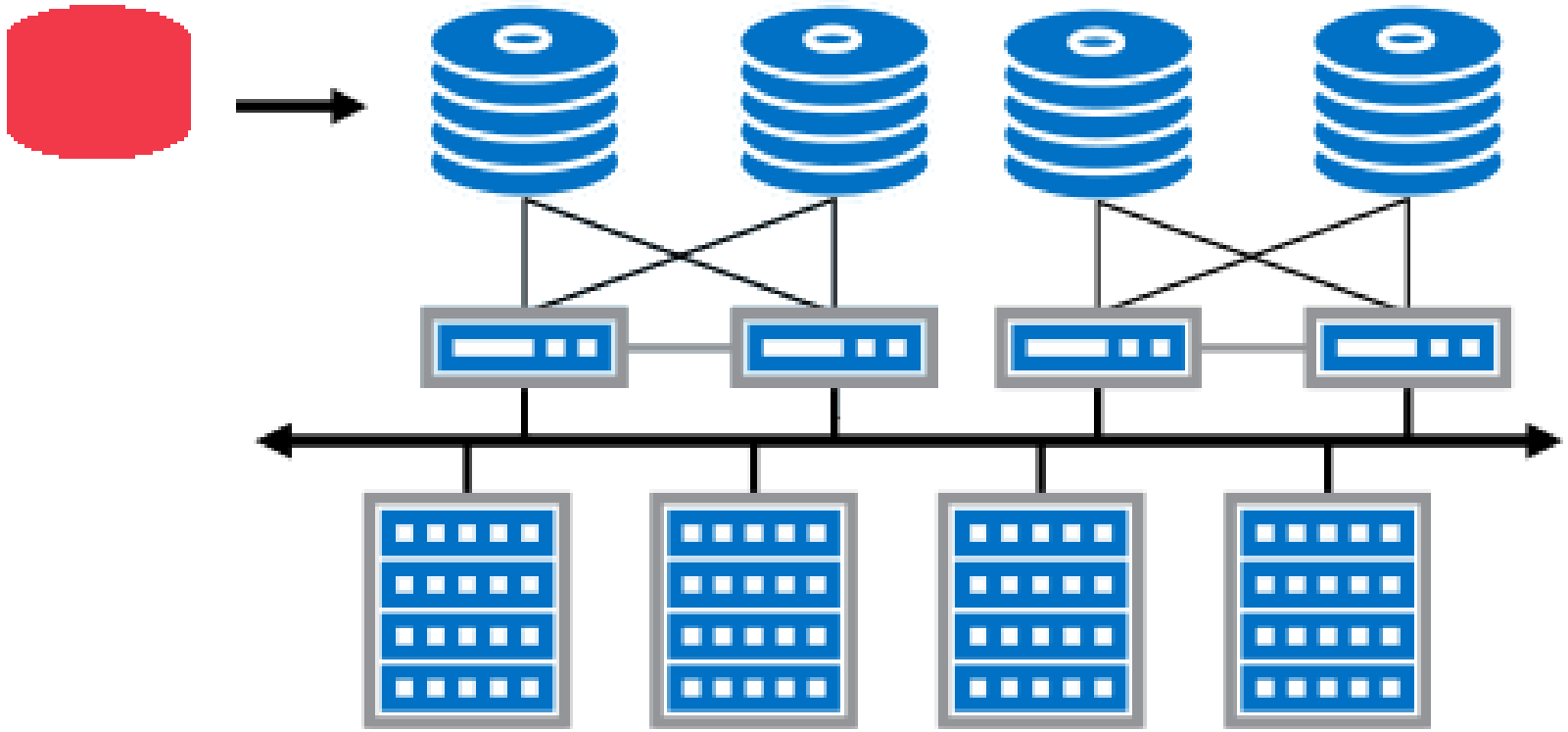
- Use $HOME to store important data or results
  - Do not use for computation
- If possible, use $BEEOND
  - Consider time of copying files
  - Use for up to ~20 Nodes
  - Especially for metadata heavy work
- Otherwise, use $WORK for metadata heavy applications
- … and $HPCWORK for bandwidth heavy applications
- If unsure, ask us!

File Systems on CLAIX

HPC.NRW

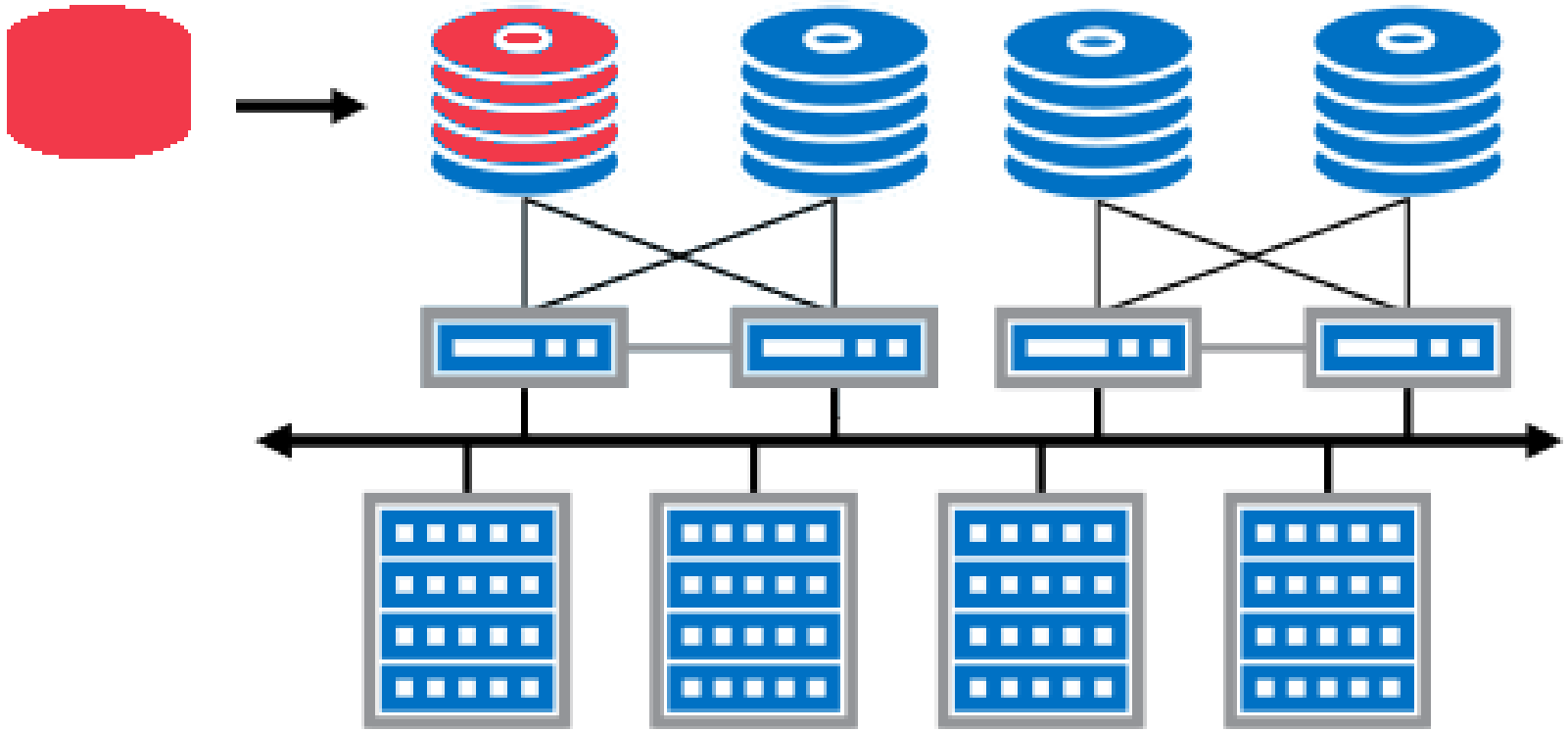
- cvmfs
  - Used to provide software or static data sets
  - Example: Singularity containers, ImageNet data set
- Data Transfer Nodes
  - copy18-1.hpc.itc.rwth-aachen.de
  - copy18-2.hpc.itc.rwth-aachen.de
  - copy.hpc.itc.rwth-aachen.de

File Systems on CLAIX

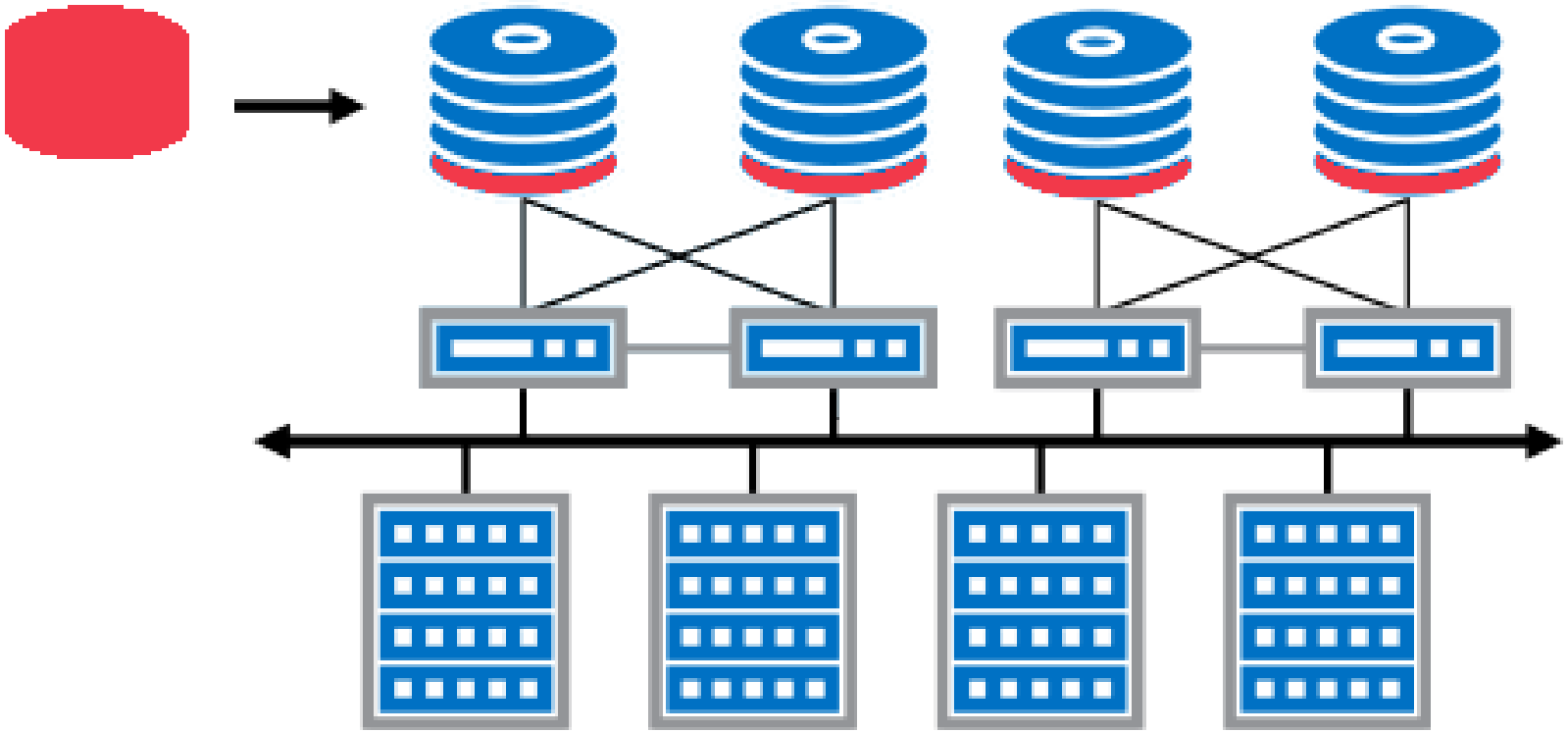


File Systems on CLAIX

HPC.NRW



File Systems on CLAIX



Tuning File System Parameters

HPC.NRW

- $BEEOND
  - Theoretically allows for very flexible configurations
  - Contact us!
  - Set striping: `beegfs-ctl --setpattern --numtargets=<N> --chunksize=<C> <FILE>`
- $HPCWORK
  - Set striping: `lfs setstripe -C <N> -S <C> <FILE>`

Outlook for 2022

HPC.NRW

- New $HOME and $PROJECT
- Intelliflash and Lustre
- SSDs!
- Quotas etc. not yet finalized

Schedule – I/O

HPC.NRW

Monday, December 6

Start	End	Topic	Speaker
09:00	10:00	Does I/O matter to me?	RWTH
10:00	10:15	Break	
10:15	11:15	Overview of I/O Technologies on CLAIX	Philipp Martin (RWTH)
11:15	11:30	Break	
11:30	12:30	Using Darshan for I/O Analysis	Radita Liem (RWTH)
12:30	14:00	Lunch Break	
14:00	15:00	Using Score-P & Vampir for I/O Analysis	Marc-André Hermanns (RWTH)
15:00	15:15	Break	
15:15	16:45	BYO: Preparation of benchmarks and job submissions for user codes	

Tuesday, December 7

Start	End	Topic	Speaker
09:00	10:00	I/O Libraries: Overview and MPI-IO	Marc-André Hermanns (RWTH)
10:00	10:15	Break	
10:15	11:00	I/O Libraries: HDF5	Sebastian Lührs (Forschungszentrum Jülich)
11:00	11:15	Break	
11:15	12:00	I/O Patterns Best Practice	Radita Liem (RWTH)
12:00	13:30	Lunch Break	
14:00	15:00	BYO: Review benchmark results	
15:00	15:15	Break	
15:15	16:45	BYO: Lightning talks about take-aways	