

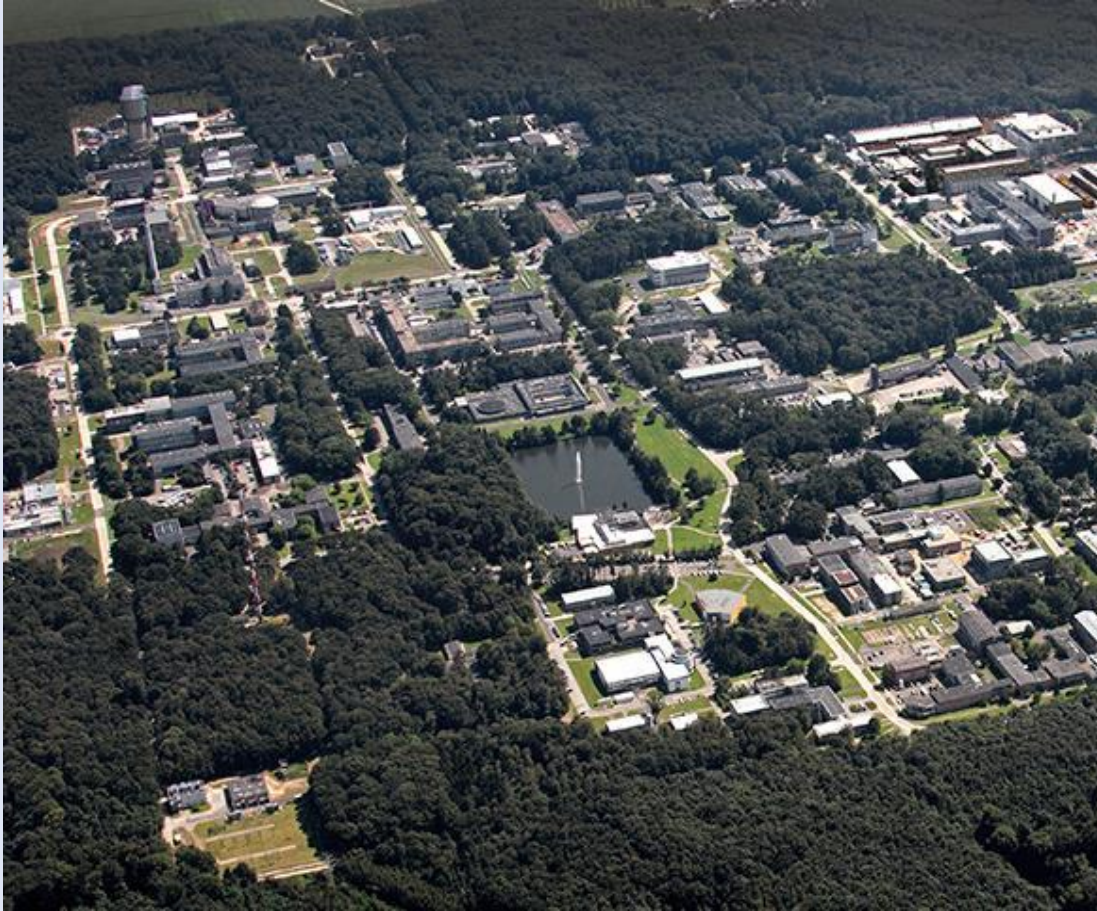


NEW STORAGE FOR JUELICH (UPDATE) EXABYTE FOR EXAFLOP

19. MARCH 2025 | STEPHAN GRAF, OLEG TSIGENOV, MARTIN LISCHWESKI & **SALEM EL SAYED**

FORSCHUNGSZENTRUM JÜLICH

Research and Development on 2.2 Square Kilometres



AT FIRST GLANCE

Founded in 1956



Revenue:
€ 987 million in
2023



Research priorities:
information, energy, bioeconomy



Research campus
with 14 institutes and
18 branch offices in
Germany and abroad
7400 Employees
114 countries



Shareholders:
Federal Republic of
Germany (90%),
federal state of
North Rhine-
Westphalia (10%)

Member of the
Helmholtz
Association



~300 Employees
(~280 FTE)



210 Scientists



22 PhD Students (+ externals)
27 Students (Bachelor/Master)



fz-juelich.de/jsc

JÜLICH SUPERCOMPUTING CENTRE

also some JSC

JSC

also some JSC



HPC, Cloud and Data Systems and Services Division at JSC

JÜLICH SUPERCOMPUTING CENTRE

- **Supercomputer operation for**

- Centre – FZJ
- Region – RWTH Aachen University
- Germany – Gauss Centre for Supercomputing (GCS)
John von Neumann Institute for Computing (NIC)
- Europe – EuroHPC, PRACE, EU projects

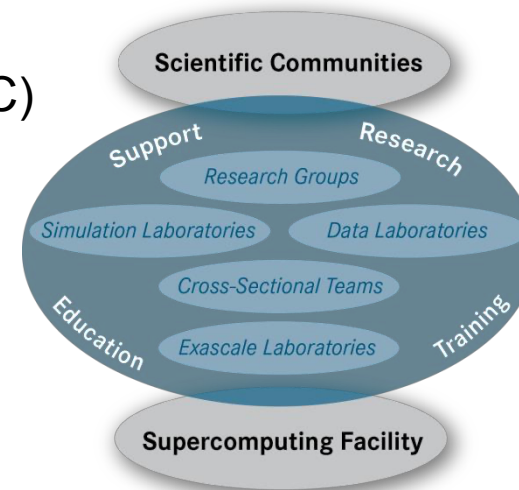
- **Application support**

- Unique support & research environment at JSC
- Peer review support and coordination

- **R&D work**

- Methods and algorithms, computational science, performance analysis and tools
- Scientific Machine Learning and AI with HPC
- Computer architectures, Co-Design, Exascale Labs together with IBM, Intel, NVIDIA

- **Education and training**

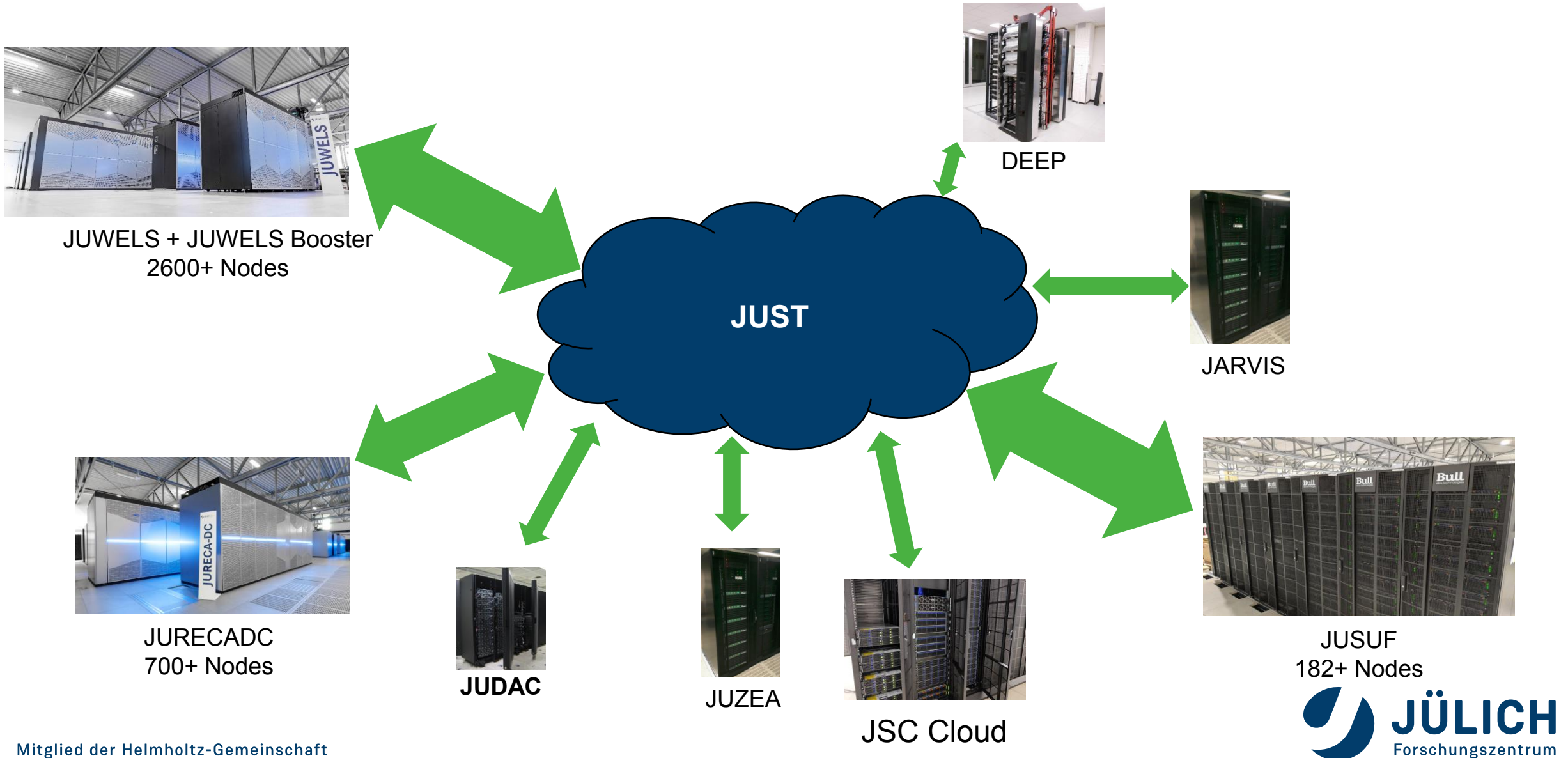


**First Exascale
supercomputer
in Europe**

DEEP



JUELICH STORAGE (JUST)



JUST HISTORY

JUST 4

2013

Capacity: **(20 PiB Raw)**

JUST 5

2018

Capacity: **50 PiB (75 PiB Raw)**

21 x DSS240
1 x DSS260
7500 10TB Disks
5 GPFS Manager
2 CES Server
2 TSM Server

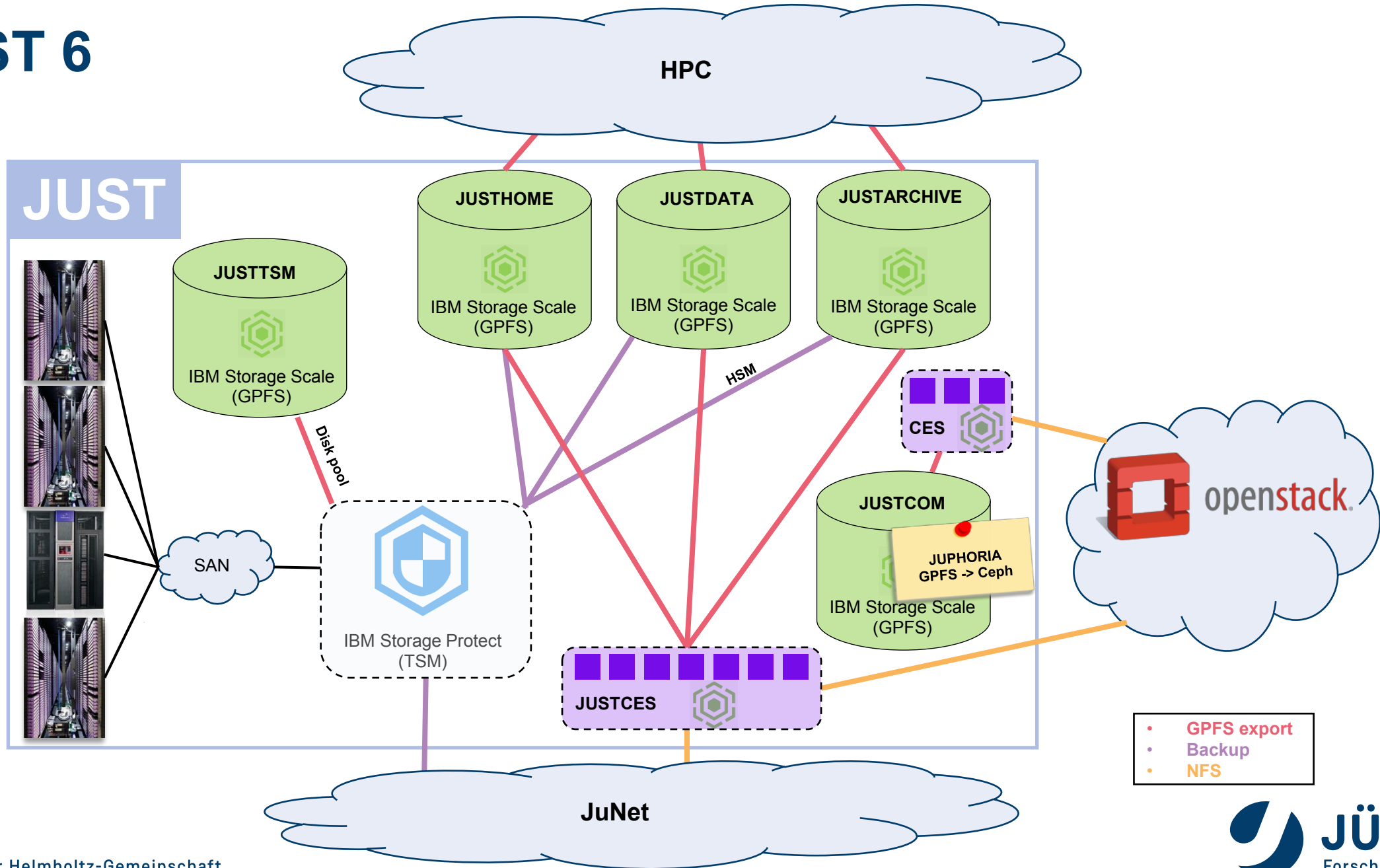
JUST 6

2024

Capacity: **110PiB (154 PiB Raw)**

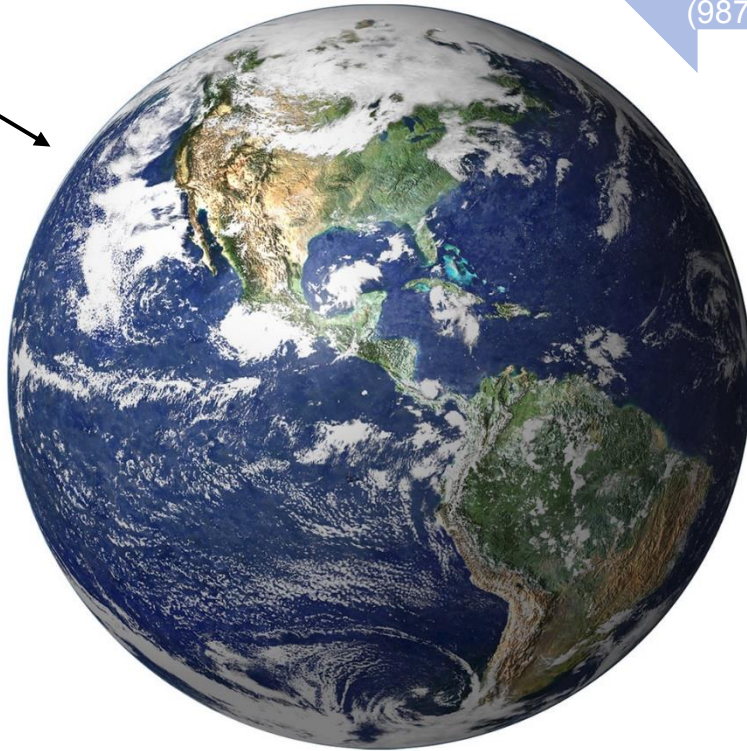
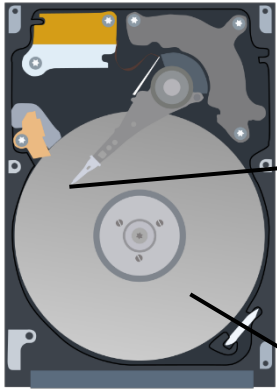
11 x SSS6000
1 x ESS3500
7007 22TB HDD
7 GPFS Manager
7 CES Server
7 TSM Server

JUST 6

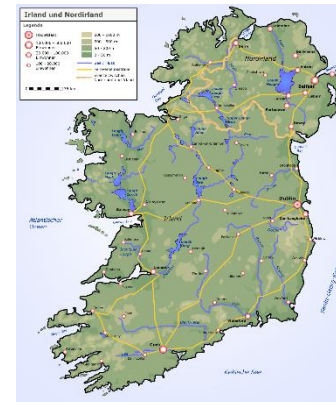
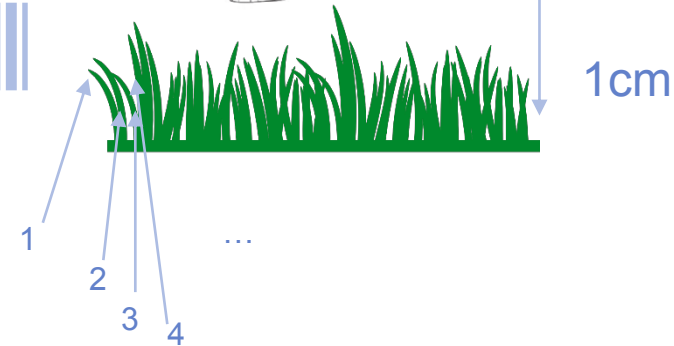


DO NOT FORGET WHAT IS BENEATH

In disks we trust!



800 Mach
(987840 Km/h)

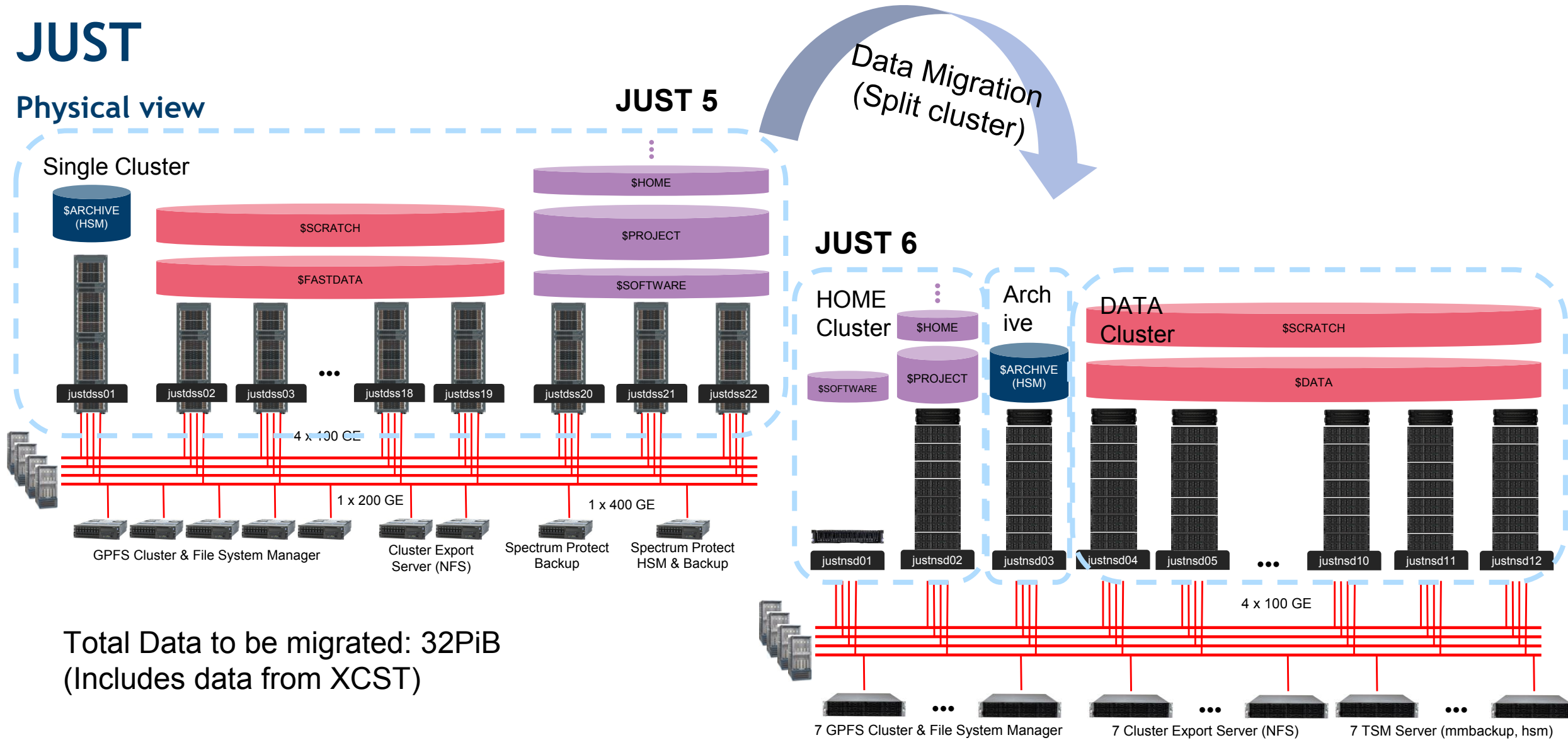


Errors < 10
Area of Ireland

JUST *a* MIGRATION

JUST

Physical view



MIGRATION OPTIONS

For Storage Scale

- Do not migrate
 - Pro: It's a green field
 - Con: Data is not moved (That is a user problem 😊)
- Restripe
 - Pro: All happens within GPFS with minimal downtime
 - Con: Some file system configurations cannot be changed
- Data copy
 - Con: Needs downtime to check equivalence
 - Main issue: ``cp``, ``rsync``, ... etc. are single process
 - Alternatives: script around ``rsync``, ``mpifileutils`` MPI-based alternatives to ``cp`` → ``dcp``, ``rm`` → ``drm``, ...
- AFM
 - Pro: New file system and data transferred by GPFS
 - Con: JSC has no experience (What could go wrong?)

Filesystem	Usage	#inodes	Dest. Filesystem	GPFS Cluster	Migration method
arch	500TB (27PB)	22M	arch	justarchive	nsd restripe
arch2	500TB (21PB)	9M	arch2	justarchive	nsd restripe
scratch	9,5PB	420M	scratch	justdata	will be newly created
fastdata	6,6PB	111M	data1	justdata	AFM
largedata	15,6PB	111M	data1	justdata	AFM
largedata2	6,5PB	39M	data1	justdata	AFM
home	28TB	56M	home	justhome	copy
project	3,6PB	702M	project1	justhome	AFM
hpcmon	6TB	3M	hpcmon	justhome	copy
usersoftware	2TB	20M	usersoftware	justhome	copy
*_software	16TB	105M	software	justhome	copy
Total (AFM + Copy)	32.35PB	1147M			

COPY: MPIFILEUTILS

Setup

- **mpifileutils:**
 - MPI-based alternatives to ``cp`` → ``dcp``, ``rm`` → ``drm``, ``rsync`` → ``drsync``, ...
 - **Open source** (BSD 3-Clause License) <https://github.com/hpc/mpifileutils>
- **Configured** a separate cluster (5 Nodes)
 - 1x AMD Epyc 9334 32 Cores; 2.7GHz; 210W TDP
 - 384GB memory: 12x 32GB DDR5-4800 ECC-reg.
 - 2 x NVIDIA Dual-Port ConnectX-7 HCA; NDR200/HDR/HDR100 IB; 200/100 GbE
- **Install** required packages and dependencies using *EasyBuild*: *OpenMPI* and *mpifileutils*
- **Export** clusters and mount file systems
 - **Avoid any root squashing**
- **Initiate transfer** under root (OpenMPI is not happy with that)

COPY: MPIFILEUTILS

Transfer

- **Requires scripting:**
Wraps a `dsync` command
- Set environment for OpenMPI
- **No `--exclude` option:**
 - Skip `.mmSharedTmpDir`, `.mmbackup`,
`.snapshots`, `.afm` and `.ptrash`
 - Start a `dsync` for each highest directory
that does not contain the skip list
- Transfer can happen in production,
but final transfer needs a maintenance

```
$ mpirun -H host1:32,host2:32,... --map-by core --npernode \  
dsync --progress 10 --xattrs all --verbose -P --delete \  
/just5/path2dir /p/path2dir  
[2024-06-10T11:07:23] Walking /just5/path2dir  
[2024-06-10T11:07:25] Walked 85241 items in 1.343 seconds (63450.212 items/sec)  
[2024-06-10T11:07:25] Walking /p/path2dir  
[2024-06-10T11:07:27] Walked 85241 items in 2.146 seconds (39718.219 items/sec)  
[2024-06-10T11:07:27] Comparing file sizes and modification times of 68832 items  
[2024-06-10T11:07:27] Item Rate : 68832 items in 0.009183 seconds (7495961.087041  
items/sec)  
[2024-06-10T11:07:27] Deleting items from destination  
[2024-06-10T11:07:27] Removed 16 items in 0.102 seconds (157.436 items/sec)  
[2024-06-10T11:07:27] Copying to /p/path2dir  
[2024-06-10T11:07:27] Items: 16  
[2024-06-10T11:07:27] Directories: 0  
[2024-06-10T11:07:27] Files: 16  
[2024-06-10T11:07:27] Links: 0  
[2024-06-10T11:07:27] Creating 16 files.  
[2024-06-10T11:07:27] Copying data.  
[2024-06-10T11:07:28] Copy data: 100.170 MiB (105036380 bytes)  
[2024-06-10T11:07:28] Copy rate: 273.398 MiB/s (105036380 bytes in 0.366 seconds)  
[2024-06-10T11:07:28] Sync completed in 0.004 seconds.  
[2024-06-10T11:07:28] Setting ownership, permissions, and timestamps.  
[2024-06-10T11:07:28] Sync completed in 0.003 seconds.  
[2024-06-10T11:07:28] Data: 100.170 MiB (105036380 bytes)  
[2024-06-10T11:07:28] Rate: 213.086 MiB/s (105036380 bytes in 0.470 seconds)  
[2024-06-10T11:07:28] Updating timestamps on newly copied files  
[2024-06-10T11:07:28] Completed updating timestamps  
[2024-06-10T11:07:28] Completed sync
```

COPY: MPIFILEUTILS

Transfer statistics

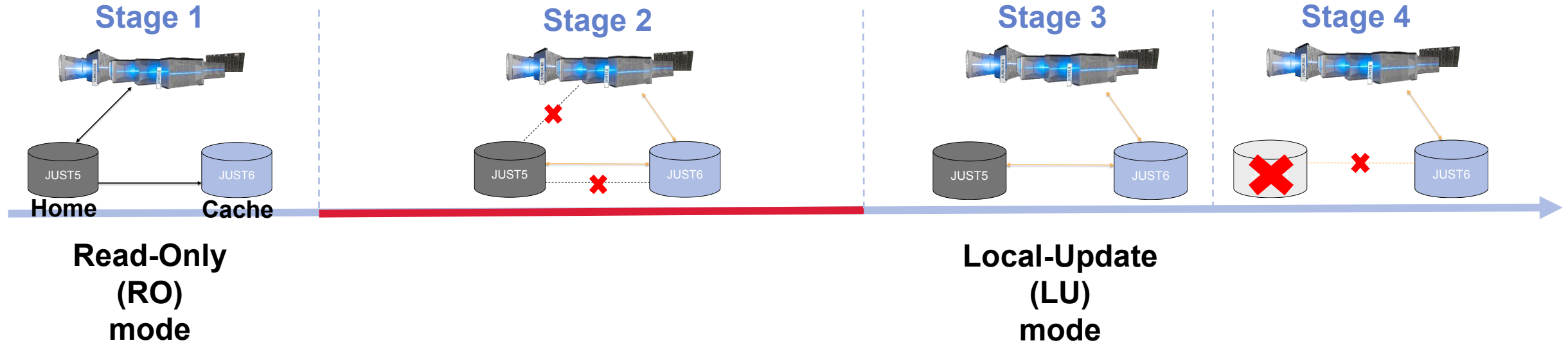
	JUST5 → JUST6 home software user_software hpcmon	JUST6 → JUST6 user data restructure requests
Transferred Data Size	32 TiB	492 TiB
Rate max	10 GiB/s	46 GiB/s
Items (Files/Folders)	160 Million	5.9 Million
Max items per second	399,576 /s	198,842 /s
Total runtime	~28 h (Most during production)	

- Above results are scrapped from logs across several runs and do include:
 - Errors
 - Missing data
 - Faulty runs
 - ...

AFM: 4 STAGE MIGRATION PLAN

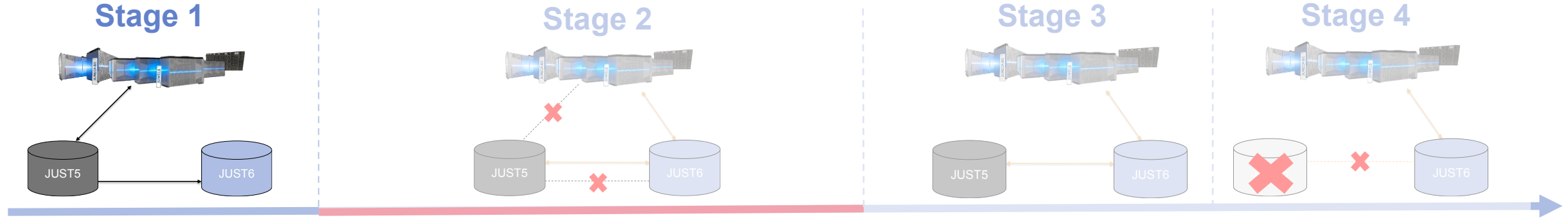
- **AFM:**

- Enables sharing of data across clusters
- Create association between two GPFS-Clusters or to a NFS data source



AFM: 4 STAGE MIGRATION PLAN

Stage 1



✓ Setup JUST6

```
$ mmvdisk vdiskset define ...  
$ mmvdisk filesystem create ...
```

✓ Setup AFM

```
# Connect the two clusters
```

```
# Setup AFM relationship
```

```
$ mmcrfileset <gpfs01> <fileset> -p afmTarget=gpfs:///remote-fs-path/userData --inode-space=new -p afmMode=RO
```

```
$ mmlinkfileset <gpfs01> <fileset> -J <pfad>
```

```
$ mmchfileset <gpfs01> <fileset> -p afmEnableAutoEviction=no
```

✓ Start moving data

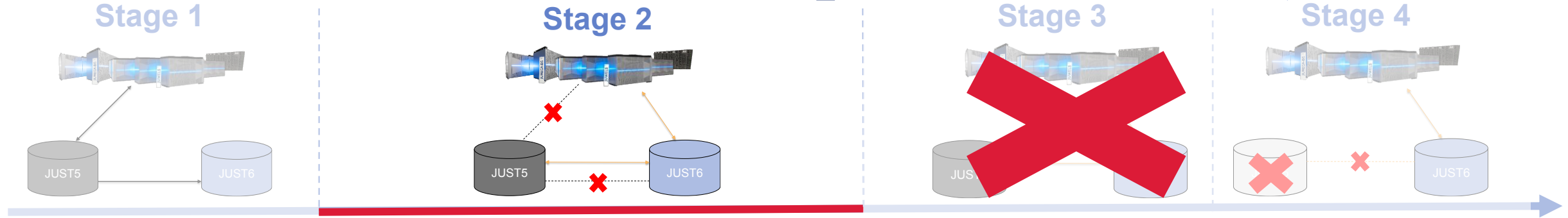
```
$ mmafmctl <gpfs01> prefetch -j <ifsetcache01> --directory </gpfs/gpfs01/ifsetcache01>
```

✓ Check progress

```
$ mmafmctl <gpfs01> checkUncached -j <ifsetcache01>
```

AFM: 4 STAGE MIGRATION PLAN

Stage 2



Planned ³/₂ day maintenance
Day 1/2

- ✓ Unmount on all HPC systems
- ✓ Rerun data move (All data moved) → No **STAGE 3 needed!**
- ✓ Convert AFM to LU-mode

```
$ mmunlinkfileset <gpfs01>  
mmchfileset <gpfs01> <files>  
mmlinkfileset <gpfs01> <file>
```

Day 2/2

**X HPC SYSTEM
WATER LEAK !**



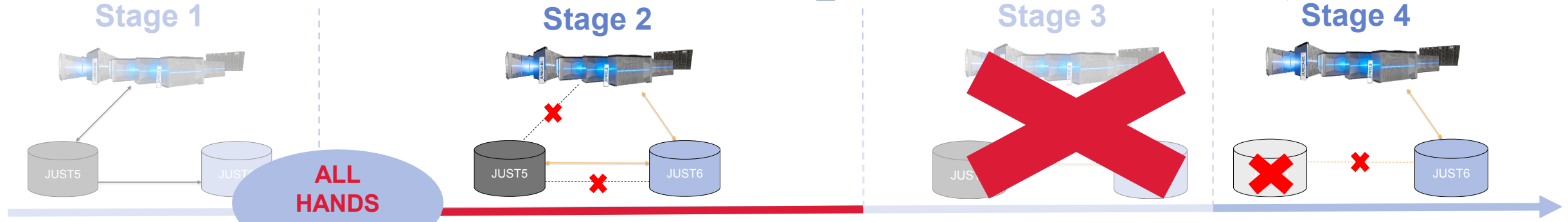
X POWER TURNED OFF



- ✓ Luckily:
Minimal drive loss

AFM: 4 STAGE MIGRATION PLAN

Stage UNKNOWN??!



Day 3/2

- ✓ Further damage control

Day 4/2

- ✓ Finish the setup of JUST6
- ✓ Mount on all available systems

X Users: Data is not up to date

Wrong parameters configured
lead to un-cached files not listed

STAGE 3 WAS NEEDED!!!

```
$ mmafmctl <gpfs01> checkUncached -j <ifsetcache01>
```

Verifying if all the data is cached. This may take a while...

Orphans list file: /var/mmfs...3067762

```
$ mmafmctl <gpfs01> prefetch -j <ifsetcache01> --directory /p/path --force --prefetch-threads 48
```

mmafmctl: Performing prefetching of fileset: <ifsetcache01>

mmafmctl: Listing all files of directory /p/<gpfs01>

...

prefetch successfully queued at the gateway.

```
$ mmafmctl <gpfs01> checkUncached -j <ifsetcache01>
```

mmchfileset(2024-06-11 00:04:56): Listing all uncached files of directory <jfsetcache01>

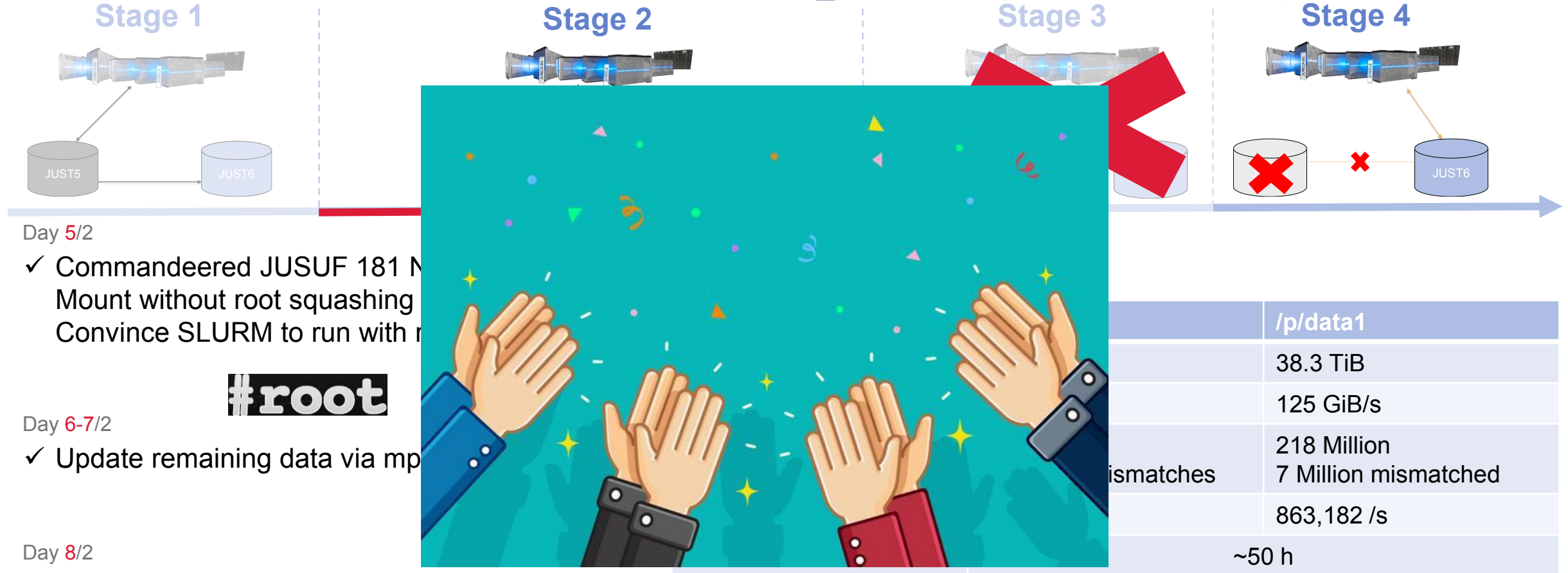
Verifying if all the data is cached. This may take a while...

All the data is cached.

But it wasn't !

AFM: 4 STAGE MIGRATION PLAN

Stage UNKNOWN??!



- ✓ Commandeered JUSUF 181 M
- Mount without root squashing
- Convince SLURM to run with r

#root

- ✓ Update remaining data via mp

✓ **WENT BACK TO PRODUCTION**

Filesystem	Usage	#inodes	Dest. Filesystem	GPFS Cluster	Migration method
arch	500TB (27PB)	22M	arch	justarchive	nsd restripe
arch2	500TB (21PB)	9M	arch2	justarchive	nsd restripe
scratch	9,5PB	420M	scratch	justdata	will be newly created
fastdata	6,6PB	111M	data1	justdata	AFM + copy
largedata	15,6PB	111M	data1	justdata	AFM + copy
largedata2	6,5PB	39M	data1	justdata	AFM + copy
home	28TB	56M	home	justhome	copy
project	3,6PB	702M	project1	justhome	AFM + copy
hpcmon	6TB	3M	hpcmon	justhome	copy
usersoftware	2TB	20M	usersoftware	justhome	copy
*_software	16TB	105M	software	justhome	copy
Total (AFM + Copy)	32.35PB	1147M			



JUPITER - THE ARRIVAL OF EXASCALE IN EUROPE



Member of the Helmholtz Association



EuroHPC
Joint Undertaking



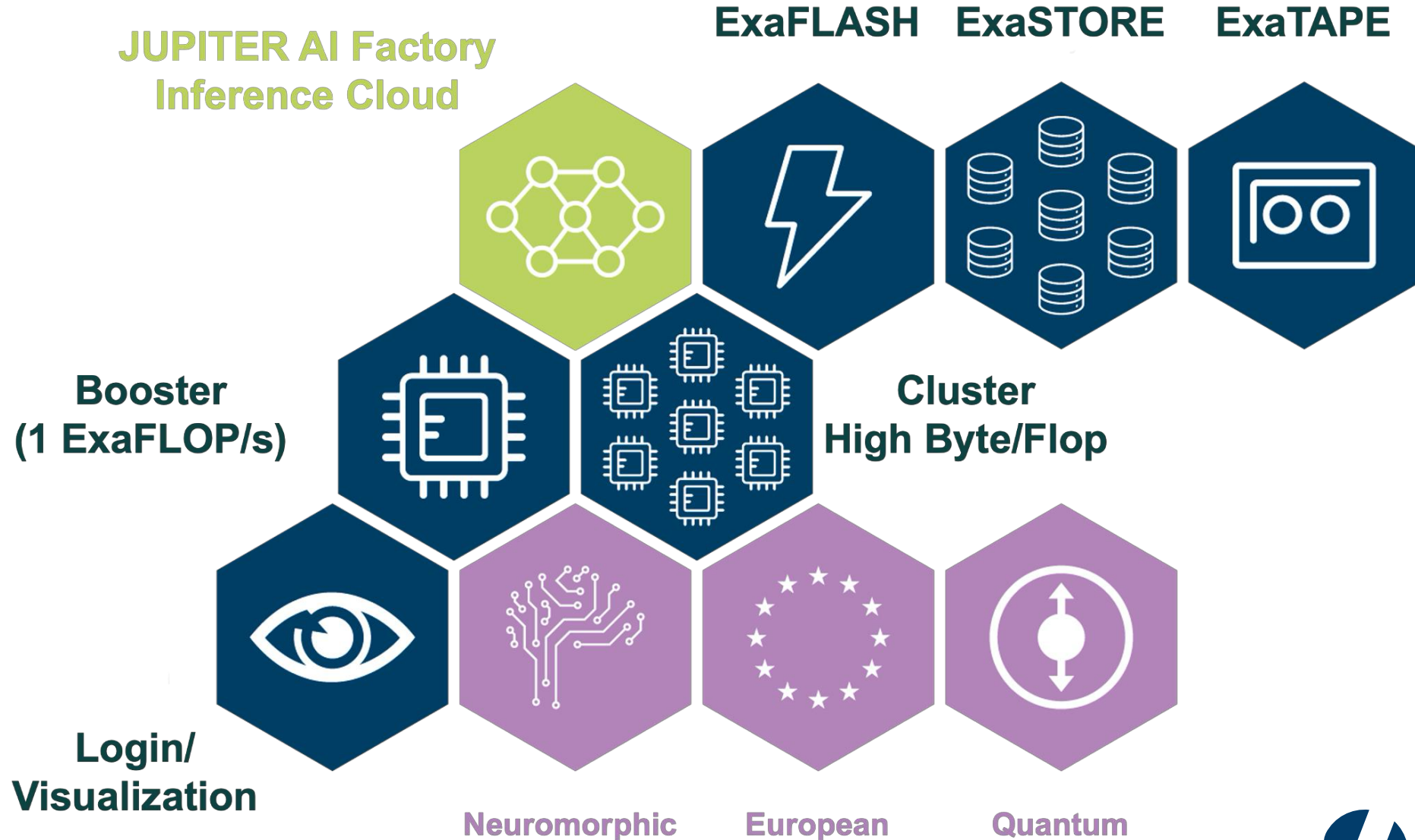
Bundesministerium
für Bildung
und Forschung

Ministerium für
Kultur und Wissenschaft
des Landes Nordrhein-Westfalen



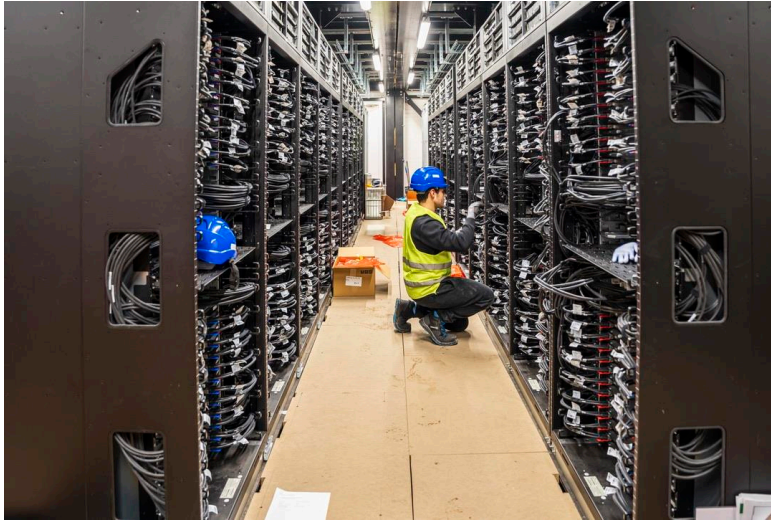
JÜLICH
Forschungszentrum
Shaping Change

JUPITER – HIGH-LEVEL ARCHITECTURE



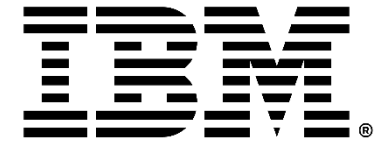
JUPITER / EXASTORE

Construction site



EXASTORE

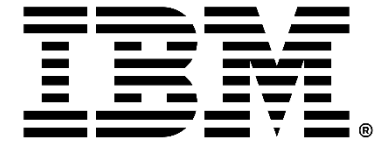
Physical view



	ExaSTORE	JUST6	Ratio (ExaSTORE:JUST6)	
Servers	22 x SSS6000 servers	11 x SSS6000 servers	2	
HDD	91x7x22=14014 x 22TiB	91x7x11=7007 x 22TB	2	
Capacity	300 PiB Raw	150 PiB Raw	2	
Bandwidth	1200GiB/s	600GiB/s	2	
				hernet + Infiniband
				2 x Infiniband per server
				Support Servers
				7 GPFS Cluster & File System Manager

EXASTORE

Physical view



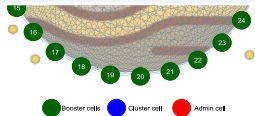
	ExaSTORE + JUPITER	JUST6 + (JUWELS&JURECA)	Ratio (ExaSTORE:JUST6)
Compute	1 exaflops	~60 petaflops	
Bandwidth	1200GiB/s	600GiB/s	
Ratio petaflop per GB/s	1.2	10	0.12



EXAFLASH



	ExaFLASH	ExaSTORE	JUST6	Ratio (ExaFLASH:ExaSTORE :JUST6)
Servers	20 x SSS6000	22 x SSS6000	11 x SSS6000	1.8 : 2 : 1
SSD/HDD	48 x 20 = 960 x 30TiB	91x7x22=14014 x 22TiB	91x7x11=7007 x 22TiB	0.13 : 2 : 1
Capacity	29 PiB Raw	300 PiB Raw	150 PiB Raw	0.19 : 2 : 1
Bandwidth	Write: 2.1 TiB/s Read: 3.1 TiB/s	1200GiB/s	600GiB/s	5.29 : 2 : 1



GPFS Cluster & File System Manager

Support
Servers

DESIRED FEATURES

DESIRED FEATURES

Looking forward “Ask us anything”

- REST API → no root ssh needed (we are investigating)
- Tape optimized backup of HSM migrated data **SPF-I-3049**
- mmbbackup support for ACLs **SPF-I-2996**
- C-Group version 2 support for QoS **GPFS-I-903**
- Reliable Quota **GPFS-I-807**
- Capacity quota support for HSM managed file systems **GPFS-I-1001**
- Prometheus Exporter for zimon metrics **GPFS-I-889** (we are investigating)
- Efficient Deletion of Fileset backup data based on mmBackup per Fileset **GPFS-I-1150**
- Manual ticket creation by mmcallhome/GUI **GPFS-I-1119**
- Command to force Dynamic Pagepool to shrink to minimum size **GPFS-I-1151**



GREETINGS FROM THE STORAGE TEAM