JUST

JUELICH STORAGE CLUSTER

22. NOVEMBER 2018 I STEPHAN GRAF (JSC)

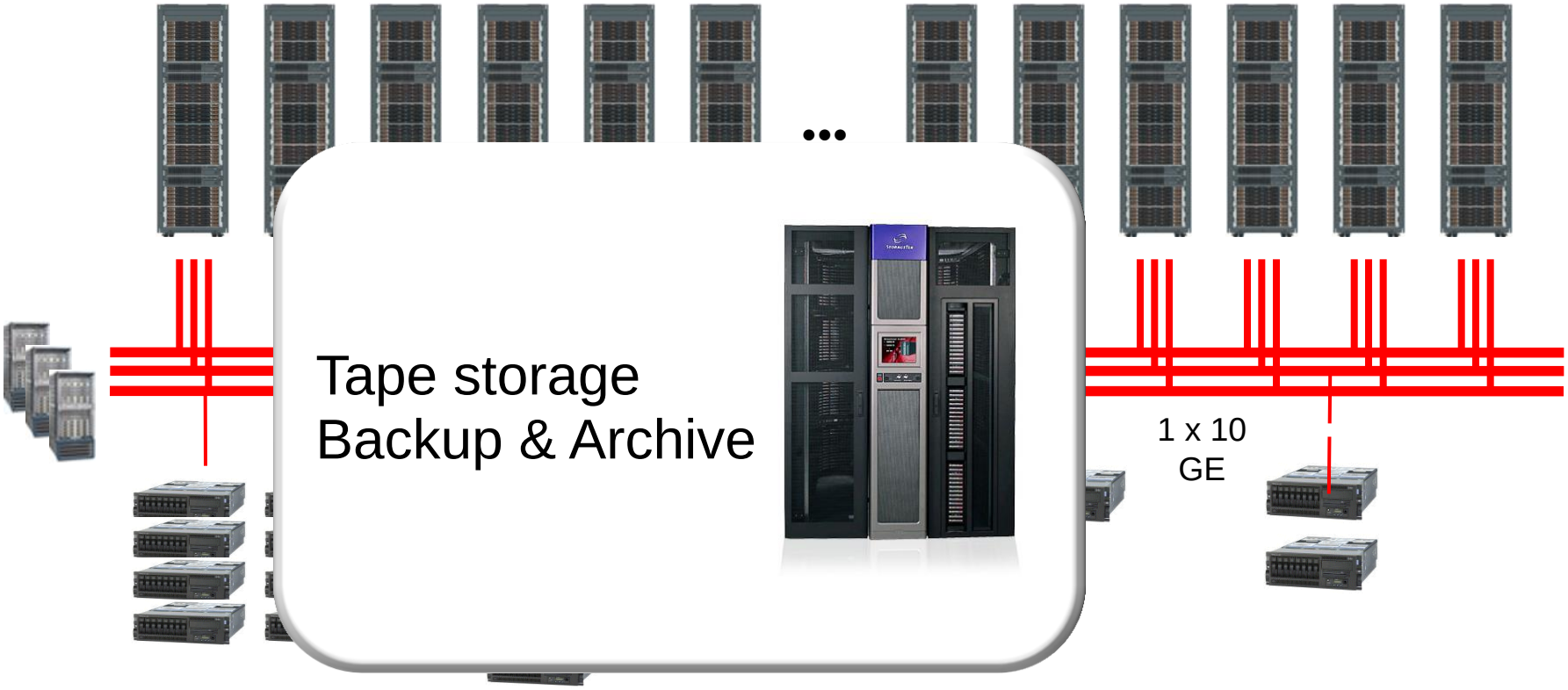
JUST CLUSTER

Key Characteristics

- Separate cluster including a lot of disk capacity for our users
- 5th generation
- Filesystem *Spectrum Scale* (GPFS) from IBM
  - Parallel access
  - POSIX compliant
  - Bandwidth optimized
  - End-to-End data integrity
- 75 PB gross capacity
- No user login
- Cross mounted on HPC systems

JUST – PHYSICAL VIEW

21 x DSS240 + 1 x DSS260 → 44 x NSD Server, 90 x Enclosure → +7.500 10TB disks



8
TSM Server
Power 8

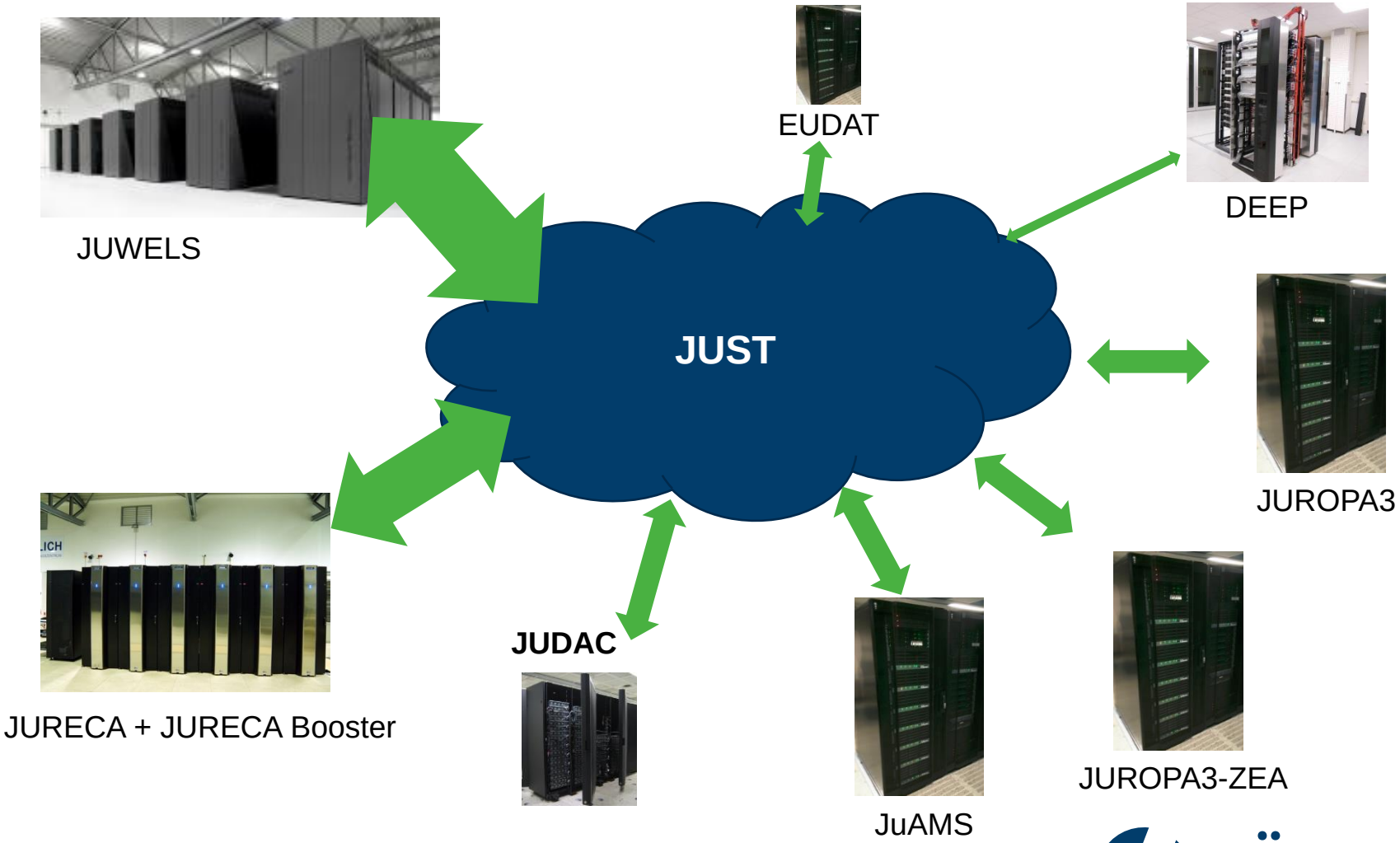
5
GPFS
Manager

Cluster Export
Server (NFS)

Monitoring

2 Cluster
Management

CENTRALIZED STORAGE

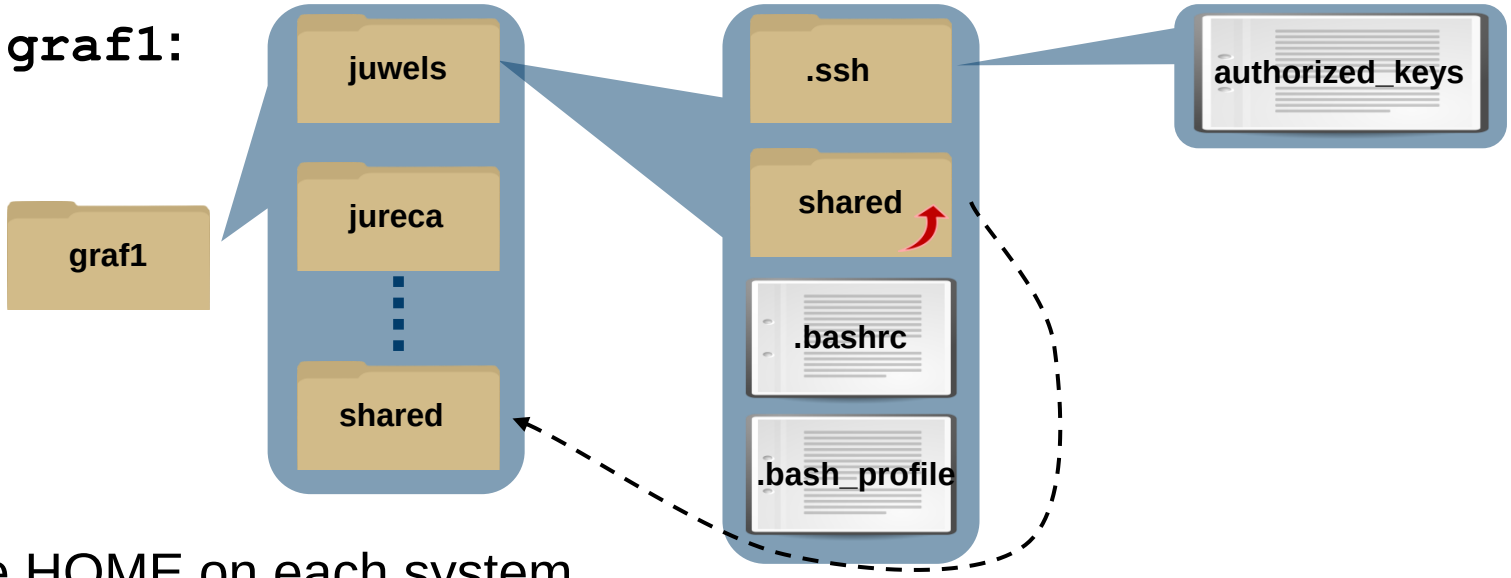


NEW HPC USAGE MODEL IN JÜLICH

- Major changes:
  - **Only one account per user** (today: one account for every project)
  - Distinguish between ***Compute*** and (new) ***Data Projects***
- **Starting in December 2018**
- Migration to new model in big maintenance
  Friday 30th of November – Tuesday 4th of December
  - All user data will be moved in new file system structure
  - LDAP will be replaced by new designed and populated LDAP
  - Dispatch infrastructure will be replaced to handle the new Usage Model

USER DIRECTORY (HOME)

- Path: **`/p/home/juser/<userid>`**
- **Small quota per user: 10 GB + 10.000 files**
- Data is in **Backup**
- Store your personal data (System profiles, SSH Key, …)
- For user **`graf1`**:



- Separate HOME on each system
  e.g. on JUDAC: **`$HOME`** = **`/p/home/jusers/graf1/judac`**
- Link to **`shared`** folder

PROJECT REPOSITORY

- Data repository for the compute project
  Path: **`/p/project/<group>`**
  e.g: **`$PROJECT`** = **`/p/project/cjsc`**
- Default Quota: 16 TB / 3 Mio inodes (files)
- Data is backed up
- User must be joined to project to get access
- Lifetime depends on project time span
  ➡ longterm storage/archiving can be realized by a ***data project***
- Data owner is the project

BUT TODAY…

HOME is User and Project Directory

- Login directory
  - Profiles
  - SSH keys
  - …
- Path: `/home[a,b,c]/<group>/<userID>`
  `$HOME = /homeb/zam/zdv124`
- Store your project related data
- **Quota per group: 10 TB + 3 million files**
- Data is in **Backup**

until end of
November 2018

JÜLICH Forschungszentrum

SCRATCH DIRECTORY

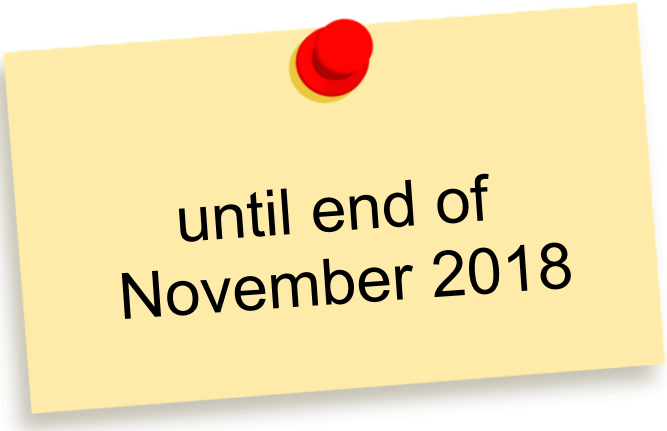
- Bandwidth optimized
  - JUST is capable of >400 GB/s
  - HPC to JUST bandwidth depends on network design
    - JURECA: 100 GB/s
    - JURECA Booster: 200 GB/s
- Belongs to compute project
- Path: **`/p/scratch/<group>`**
  `$SCRATCH` = **`/p/scratch/cjsc`**
- Temporary files, checkpointing
- **Quota per group: 90 TB + 4 million files**
- **No Backup**
- **!!!Data deleted after 90 days without access!!!**
- Empty directories are deleted after 3 days

BUT TODAY …

The scratch file system is named WORK

- Same characteristics as $SCRATCH (same disks)
- Path: `/work/<group>/<userID>`

  $WORK = /work/zam/zdv124



JÜLICH Forschungszentrum

ARCHIVE REPOSITORY

Data Project

- Filesystem consist of 2 tiers: disks (cache) and tapes (long term)
- Path: `/arch[2]/<group>`
  $ARCHIVE = /arch/zam
- Archive your results
- Only available on login nodes

BUT TODAY…

- Path: `/arch[2]/<group>/<user>`
  $ARCH = /arch/zam/zdv124

```
[zdv124@judac01:/arch/zam/zdv124> ls -lisah
total 320K
  407977  128K drwx------   2 zdv124 zam  64K May 18 10:01 .
  407555  128K drwxr-xr-x 316 root   sys  64K May 24 15:00 ..
18062260   64K -rw-r--r--   1 zdv124 zam    5 Sep  2  2011 datu.txt
12920848    0 -rw-r--r--   1 zdv124 zam  12G Jun  3  2015 Vervet_s0050_tiff.tgz
```

- Recall per file is expensive (2 minute mount time ~ 200 MB/s)
  → **use tar balls > 1TB**
- **Avoid renaming of directory structures** (may trigger huge recalls)

JÜLICH Forschungszentrum

FASTDATA REPOSITORY

Data Project

- High Bandwidth (close to $SCRATCH)
- Data project proposal must point out it's requirements for **FASTDATA**
- Path: **/p/fastdata/<group>**
  $FASTDATA = **/p/fastdata/zam**
- **Quota per group: depends**
- **Files are in Backup**

LARGEDATA REPOSITORY

Data Project

- Separate storage cluster (XCST)
- High Capacity (disk based)
- Data project proposal must point out it's requirements for *LARGEDATA*
- Path: **/p/largedata/<group>**
  $DATA = **/p/largedata/zam**
- **Quota per group: depends**
- **Backup realized by file system snapshots**

FILESYSTEMS - SUMMARY

File System	Description	Project Type	Characteristics
home	Users HOME File Systems		User Quota: 10GB/2Mio Files Files in Backup
project	Compute Project File System	Compute	Group Quota: 16TB/3Mio Files Files in Backup
scratch	Compute Project Scratch File System	Compute	Group Quota: 90TB/4Mio Files
fastdata	High Bandwidth and large Capacity File System	Data	Group Quota: depends Files in Backup
largedata	Large Capacity (Disk based)	Data	Group Quota: depends Snapshot based Backup
arch arch2	Archive File System (Tape)	Data	Group Quota: 500K Files Files in Backup Migration to tape

JUDAC – JUELICH DATA ACCESS

Data access and transfer cluster

- All HPC user can login on judac:

```
ssh <userid>@judac.fz-juelich.de
```

- Independent from HPC systems (e.g. in maintenance)
- Purpose: data transfer in & out the HPC filesystems
  - scp, rsync
  - Grid Tools
  - UNICORE FTP
- For more information go to → JUDAC Web Page

HINTS & TIPPS

- Restore files from backup: **`adsmback`**
  - **available only on JUDAC**
  - Calls IBM TSM Backup/Restore GUI
- Quota usage information: **`jutil`**
  - Project group quota info:
    `jutil project dataquota -g <project>`
  - User quota info: **`jutil user dataquota -u <user>`**
- SSH/SCP usage
  - Multiple external (scripted) access can be classified as an attack
    → Firewall will block external IP
- Take care of your files
  - No special characters in filenames (newline, tab, escape, …)

AND FINALLY

- For any problem (accessing files, access rights, restore, quota, data transfer, …) contact JSC application support:

  sc@fz-juelich.de

- If you want to optimize your application IO:

  Attend the ***Parallel IO*** **Training** @ JSC

  - HDF5, MPI-IO, SIONLIB, …

Questions?

JÜLICH Forschungszentrum