Data management at PETRA III

(and partially DESY FS-*)

data acquisition, storage and access

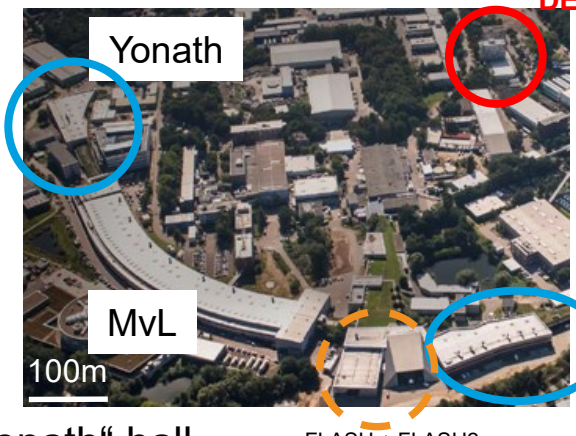
A. Rothkirch
DESY Photon Science, FS-EC (Experiment control)

and many colleagues from FS-EC, DESY Central IT and the beamlines

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

PETRA III Beamlines



„Max v. Laue" hall

P01 - High-Resolution Dynamics
P02.1 - Powder Diffraction and Total Scattering
P02.2 - Extreme Conditions
P03 - Micro- and Nanofocus X-ray Scattering
P04 - Variable Polarization XUV Beamline
P05 - HZG/DESY: Imaging
P06 - Hard X-ray micro/nano probe
P07 - HZG: High energy materials science
P08 - High resolution diffraction
P09 - Resonant scattering and diffraction
(+MX in 2023)
P10 - Coherence applications
P11 - Bio-Imaging and Diffraction

P12 - EMBL: BioSAXS
P13 - EMBL: Macromolecular crystallography I
P14 - EMBL: Macromolecular crystallography II
(EMBL is managing itself, but we're in touch/exchange)

Deutsches Elektronen-Synchrotron
Helmholtz-Zentrum Geesthacht Centre for Materials and Coastal Research (Hereon since a while)
European Molecular Biology Laboratory

„Ada Yonath" hall

P21 - Swedish Materials Science Beamline (SMS)*
P22 - Hard X-ray Photoelectron Spectroscopy
P23 - In-situ and Nano-diffraction
(+ Hierarchical Imaging for Materials Sciences and Biology – Laminography; by Karlsruhe Institute of Technology [KIT], early 2023)
P24 - Chemical Crystallography
*P25 (*Bio-Medical Imaging, Powder Diffraction & Innovation Beamline */ in prep.; 2023/24)*

„Paul P. Ewald" hall

P61 High-Energy wiggler beamline
P62 Small angle X-ray scattering
P63 (combined XAS/XRD/SAXS beamline for operando studies of batteries, catalysts etc. OPERANDOCAT / in prep.; 2024)
P64 - Advanced X-ray Absorption Spectr. (QEXAFS)
P65 - Applied X-ray Absorption Spectr. (class. EXAFS)
P66 Time-resolved luminescence spectroscopy

Diverse Environment

**Various kinds of research* **Different techniques* **Multiple kinds of analysis*

** Plenty of devices resulting manifold data types/sizes etc.*

**Note P21: P21.1 Broad band diffraction & P21.2 Diffraction & Imaging*

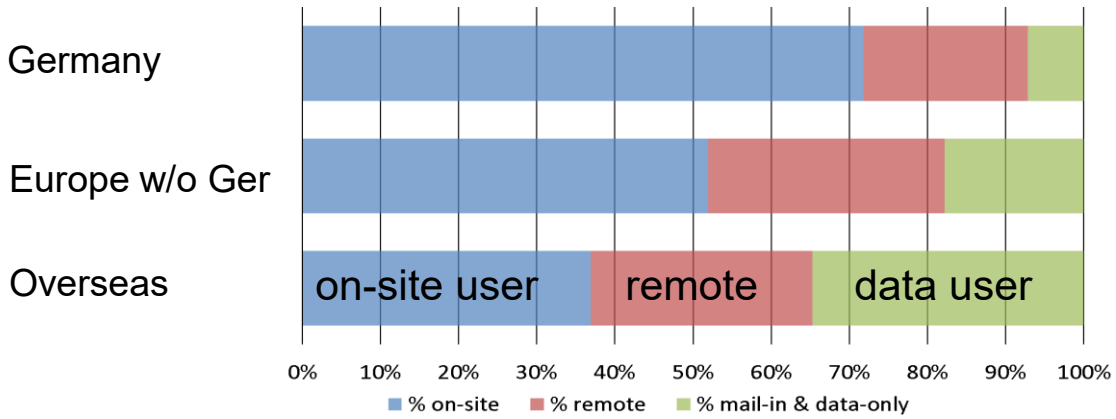
PETRA III
User Statistics

Comprised from H.-C. Wille, O.H. Seeck and C.G. Schroer, Photon Science Meeting, Aug 31 2022

- Users Mar 15th to Dec 22nd 2021
  [including *on-site visits, mail-in services, remote access, data-only users]*
  - Around 3000 unique users
  - Around 6500 user visits *(here: all beamtime participants incl. inhouse)*

ca. 55 % Germany
ca. 37% Europe w/o Ger
ca. 8% Non-Europe

- Kinds of access (excluding internal staff)



  - In general „data-only“ and „mail-in“ is appreciated by visiting users (no need to handle experiment)
  - Users from overseas have preference for remote access/mail-in.
  - An experiment often involves all three types of access
  - Indian beamtimes successfully supported by a permanent Post-Doc since 2021

Experiment / data life cycle

Apply for an experiment

Experiment preparation

- Integrate brought in equipment
  (i.e. unknown accounts)

Start of the experiment

- access to storage space(s)
- access for functional account & users

Data acquisition

- variety of formats, sizes and speed
- different amounts of data
- different operating systems

Activities during the experiment

End of the experiment

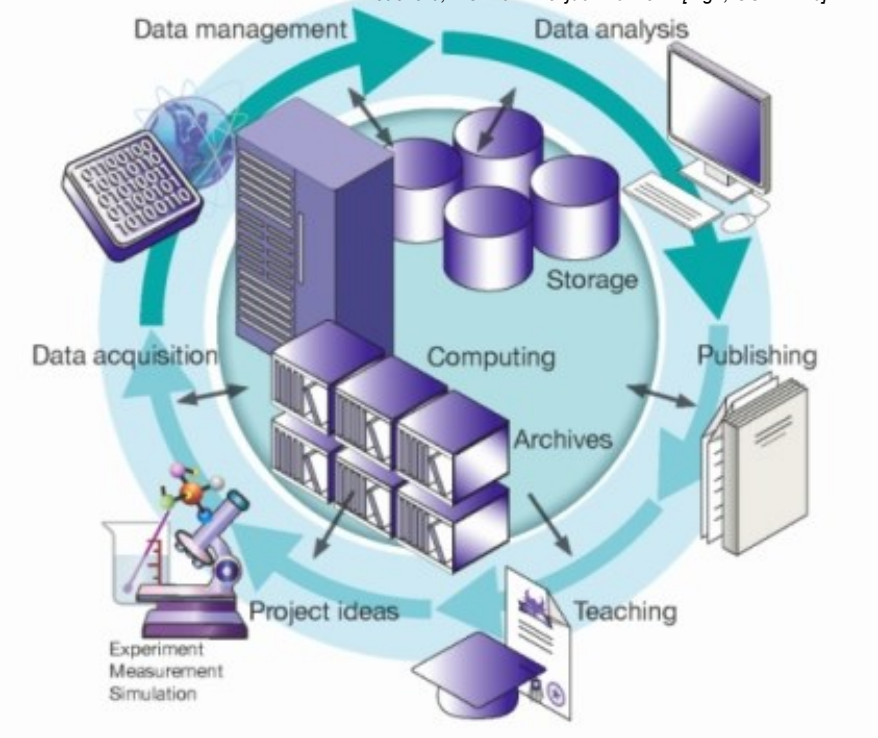
- Data not accessible for next user group

Data access past the experiment

- Offline analysis on- and off-site, download opt.

Data archival & more

Image src: Data-Life-Cycle-Labs-DLCL-model, L. Pouchard, DOI:10.2218/ijdc.v10i2.342 [Fig2, CC BY 4.0]



Relies on DESY Compute Center infrastructure

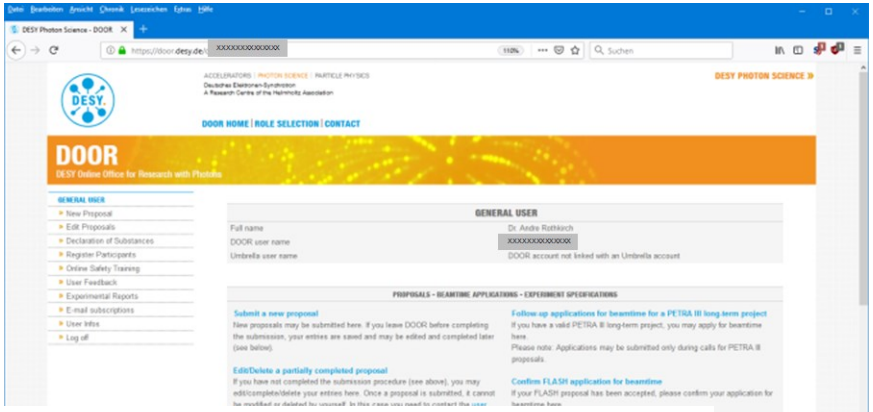
Common issues: space, performance & reliability

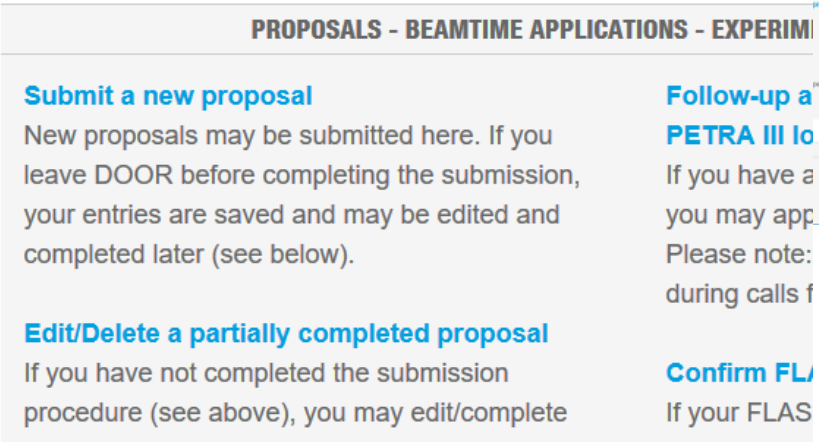
Digital User Office

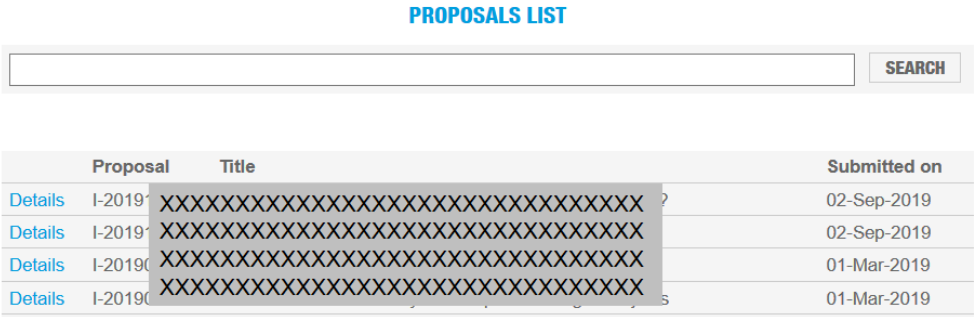
J.P. Kurz (EC), D. Unger (PS), U. Lindemann (IT)

The Digital User Office DOOR facilitates

- Proposal submission
- Peer reviews
- Beamtime scheduling
- Declaration of substances/ List of participants
- Miscellaneous administrative tasks.
- DOOR is based on DUO (PSI). It is a common activity between the FS department and central IT
- Generation of unique ID per BT "**Beamtime Application ID**"







The DOOR user portal

A new storage system / concept for PETRA III
(invented 2015)

- Invention of a managed storage system
  - Directory structure based on facility & beamtime:
    - Access rights (3 groups per beamtime: <AppId>-dmgt, -part, -clbt)
    - Archiving/Portal (4 user roles)
    - Migration/staging
  - Directory-specific policies (e.g. rw, ro or archiving)
  - Control the data 'visibility'/accessibility (*note: BLs have functional accounts*)
    - Temporary storage ("BL-FS") to cope with data from various sources (guest equipment, detector PCs)
    - Limit the visibility of the temporary storage to beamtime and beamline
  - Start-/Stopbeamtime to create temporary and permanent directory structure
  - ACLs for permanent storage ("Core-FS")

- IBM GPFS Storage Server (IBM Spectrum Scale & Elastic Storage Server)

  [General parallel filesystem (GPFS) is a high-performance clustered file system]
  - IBM 5146-GS1: ~55 TB; 2.5" 10K rpm **HDD** (1.2 TB) or 2.5" **SSD** (400 GB or 800 GB). *(initial 2015)*
  - IBM 5146-GL4: ~700TB; 3.5" **NL-SAS HDD**s (2 TB or 4 TB). Note: 5146-GL6: like GL4, but 6 x DCS3700

  Meanwhile several times expanded and first systems already replaced
  - Currently BL-FS 220TB SSD (+ temporary 220 TB HDD) and Core-FS ca. 13 PB HDD total *(last update 2022)*

Start/Stop a Beamtime [by BL staff]

***startBeamtime** --beamtimeId <beamtimeID> --beamline <beamline> [more options]*

- Instantiates beamline (BL-FS) and core filesystem,
  i.e. creates filesets with predefined directory top-level structure and rules/constraints
- BL-FS: NFS3 + SMB (and Hidra) based on whitelist
  - fixed mount point /gpfs/current at every beamline
  - recommended drive letter for Win
  - Hidra: data passing via ZMQ
- Core-FS: NFS4 + SMB and Access control list (ACLs)
- Ingests information of BT into gamma-portal
- Creates 3 unix groups per BT: -dmgt, -part, -clbt
- Copies list of participant from DOOR into ACLs list in portal &
  checks for registry accounts and - if existent - fills unix-groups created given the DOOR role *(leader, pi[=applicant] or participant)*
- *Within limits: allows allocation of compute resources for e.g. auto-processing*
  (i.e. P11 MX, P06 tomo/ptycho, P…)

Start/Stop a Beamtime [by BL staff]

Predefined directory toplevel structure and rules/constraints

***startBeamtime** --beamtimeId <beamtimeID> --beamline <beamline>*

Temporary storage (**on IPs**)

Limited to Beamline & Beamtime ("BL-FS")

GPFS with NFS 3 / SMB

/gpfs/
- local *(permanent, 3TB, 'emergency buffer')*
- current
  - raw
  - processed
  - shared
  - scratch_bl

(dynamic, 'limited' by avail flash/disk space)

Copy during BT (one way) [inotify runs on proxies; delete is accounted]

Permanent storage (**on ACLs**)

"GPFS Core file system"

GPFS - native on Analysis nodes
- or by NFS 4 mapping / SMB (office)

/asap3/
- <facility> (e.g. petra3)
  - gpfs
    - <beamline>
      - <year>
        - <data>
          - <beamtime-ID>
            - raw *(read onlyexcl. dmgt)*
            - processed *(read/write)*
            - shared *(read/write)*
            - scratch_cc *(read/write; NOT archived)*

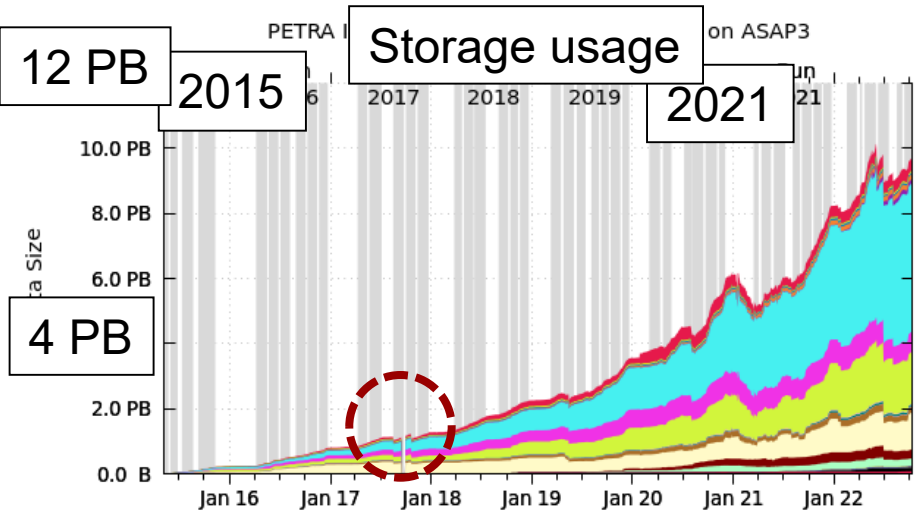
WGS
dCache
Portal
Download

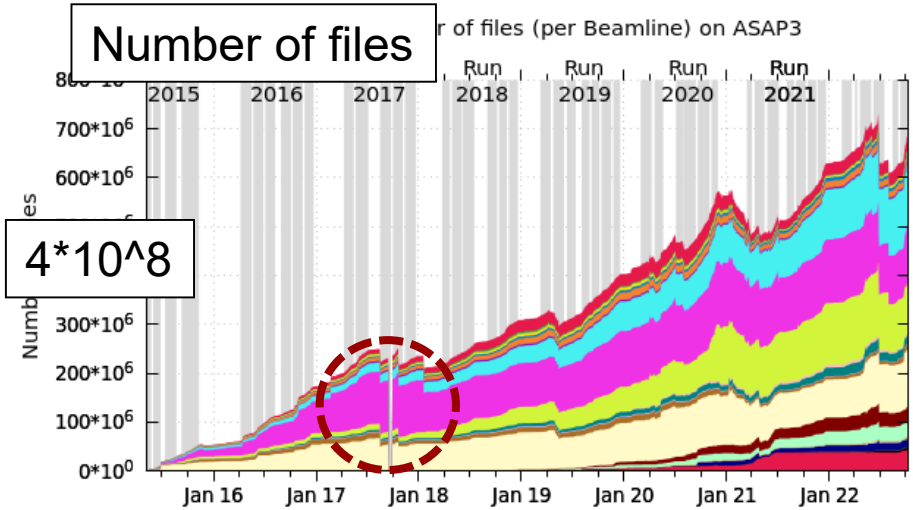
Since 2021: passthrough mode, 10TB

***stopBeamtime** --beamtimeId <beamtimeID> --beamline <beamline> makes given BT invisible for BL*

a beamtime can not be restarted

GFPS "permanent" storage (GPFS core) – P3 usage





P3 2015: ~0.200 PB total (30 TB/month)
P3 2016: ~0.460 PB (55 TB/month)
P3 2017: ~0.580 PB (75 TB/month)
P3 2018: ~1.0 PB ~750 TB (90 TB/month) [+ 250 TB FLASH+Ext.]
P3 2019: ~1.4 PB (as of 3.12.) ~1.3 PB (140 TB/month) [+85 TB FLASH+Ext]
P3 2020: ~3.0 PB (~1.9 PB [+90 TB FLASH+Ext] as of 13.11)
P3 2021: ~3.8 PB (~ 400TB/month)
P3 2022: ~3.4 PB (status as of 10.10.22)

New detectors [Eiger+Lambda 2M&pool]
New detectors [Eiger+Pilatus3 2M]
More BLs & efficient use

Amount per BT: 100MB to 100+ TB (in ~3 days)

Currently ca. $7*10^8$+ files in total (2015/2016 ~ 1, doubled in 2017)
- new detectors resulting larger files
- Partially NeXus/HSF5 files (Lambda, Eiger, [AGIPD])
- Partially compressed data

(Start to) delete data from GPFS permanent disk storage in 2017/2nd half
[on GPFS capacity left, last data access; data kept on tape; staging on request]

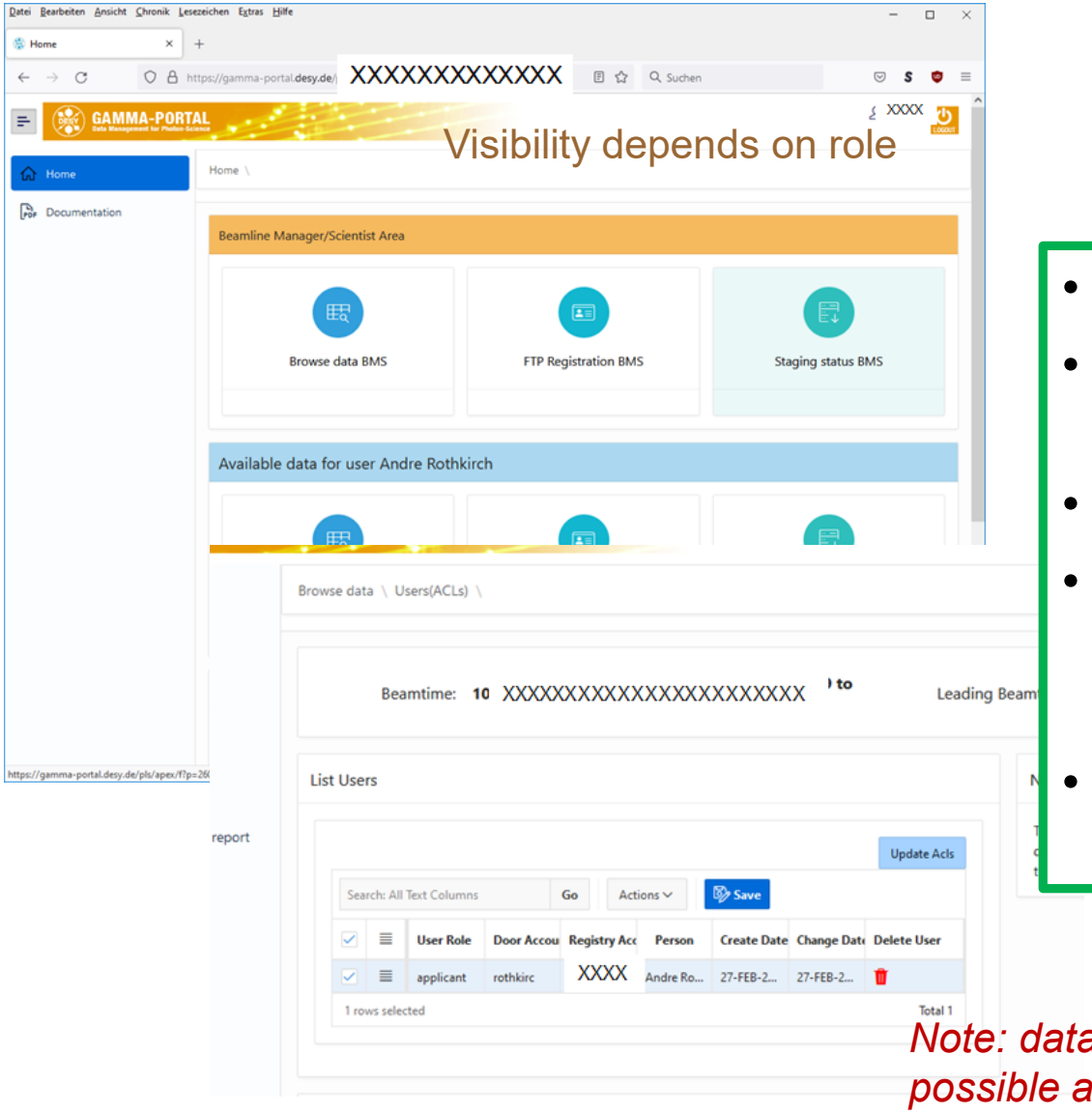
Current data „lifetime“ and archiving

- Data is kept in GPFS core (HDD) for analysis
  - Removed from disk 180 days after BT stopped (visiting users) (larger hold times for in-house)
  - Data removed from disk if capacity runs short
  - 2 snapshots a day for max. 21 days to cope (mainly) accidentally deleted files
- Data is copied into long term storage “DESY dCache”
  - ~7 days after stopBeamtime was given
  - Two tape copies per BT
  - Delta is created before removal from GPFS disc
- Data stage (restore from archive) on request to FS-EC
- Data export and Management
  - gamma-portal for access management
  - ftp/TLS + Globus to outside (ro) if on GPFS
- So far nothing deleted

Tape robot @ DESY IT [src: Manuela Kuhn, talk GridKa School 2016, Karlsruhe

- *Open question(s): How long preserved? Open access? ……*

Gamma-Portal (htts://gamma-portal.desy.de)



- By DOOR account
- Different views depending on role
- Basic search
- Enable FTP export Data download by dedicated FTP client then
- Project leader & applicant can manage ACLs

Note: data download via a globus is also possible and implies scientific account

Gamma-Portal (cont.)

ACL management and rights on GPFS core (implies DESY Reg. account)

- Four roles exists (hierarchical)

1. Leader	}	<AppId>-dmgt
2. Applicant	}	
3. Participant	→	<AppId>-part
4. Collaborator	→	<AppId>-clbt

Assoc. to BT groups (LDAP) created with startBeamtime

- Only one leader possible
  - can add/remove users
  - can change roles
  - r/w/d to raw/processed/shared/scratch (raw in future time limited?)
- Applicant
  - can add users
  - can change roles over lower hierarchy / advance role up to applicant
  - can remove users (only of lower hierarchy)
  - r/w/d to raw etc.
- Participant
  - r on raw; rwd on processed/shared/scratch
- Collaborator
  - read only access

Note:
FTP/TLS download by DOOR account (ro)
Globus export is ro as well and needs Desy account

Developments since 2015

- Invention of dedicated detector net for demanding detectors („detector net“)
  - Single 10GE link for detector PCs to DESY LAN (Cu Base-T)
  - Special subnet 192.168.138.* (i.e. LAN only, NO internet)
  - Dyn. VLAN NOT supported -> detector has to be in given subnet to use 10GE plug
  - Ethernet (RJ45) wall plugs at experiment are marked by purple frames
- Currently in progress: Invention of PDAQ Network
  - Separated from Office network
  - Option for 100GE, e.g. Eiger2 (installed at 4 beamlines already)
    (100G QSFP28 LR4,10 km, λ1296-1309 nm)
- GPFS used now at PEX, FLASH & FLASH2, “Special instruments” [e.g. labs] & more
- GPFS for external BTs (i.e. experiments by PS staff carried out not at Desy or Xfel)
- GPFS for groups/research teams started
  (no external users, only group ACLs, no portal, different “scheduling”)
- GPFS capacity expanded multiple times, from on 2017/18 each year
  currently BL-FS ~220T SDD (+220T HDD) / core ca. 13000 T ~ 13 P)
- GPFS replacements to keep state (i.e. HW out of warranty)
- Invention of HiDRA and Lavue & further developments
  (well received by our Beamlines)
- Copy procedure improved/modified in 2018 (inotify on proxies)
- Procedure to remove data from GPFS permanent storage (HDD) in 2017
- Passthrough entered into force in 2021 for all Beamlines

Further findings and action taken

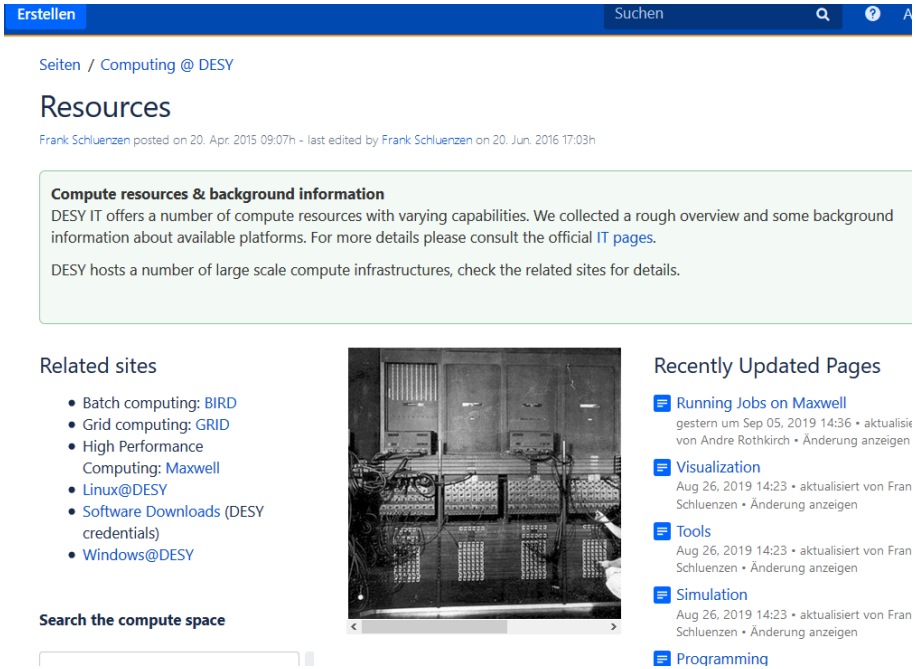
- More and more (external) users can‘t take data home
  - Lack storage resources (e.g. already “some” TB+ per BT challenge first users)
  - Lack basic compute resources or particular analysis software
- Staff as well as external users ask for central ‚interactive‘ compute resources (ssh / FastX whenever one likes)
- Staff as well as external users have growing needs for “large” compute power meeting demands like e.g.
  - amount of RAM, number of cores and/or nodes, amount of GPUs
  - faster remote GUI (opengl stuff like e.g. Avizo)
  - **Exclusive** resource usage

- Demands cannot be fulfilled by single concept/ one system fits all
- Different systems are needed for specific computing use cases
- Resource management is needed
- One has to be aware of computing and choose decent resource

Data access and analysis environment
(located @ Computer Center)

- Invention of Scientific accounts (i.e. DESY accounts for external users (‘external’ ≠ industry/commercial) with own namespace ‘psx’
- Provision of interactive resources max-fs-display
  (max-fsc/max-fsg will be shut down early 2023)
- Creation of specific batch resources for PS managed by SLURM
  - Slurm partition ps (inhouse)
  - Slurm partition psx (external [non-commercial] users or use cases)
- Invention of display-servers for processing involving GUI
- Remote access (firewall/tunnel or Web-Browser)

https://confluence.desy.de/display/IS/Resources



Batch (Maxwell cluster)

- *Exclusive resources usage for jobs managed by SLURM*
- *Efficient resource usage (batch queue, resource definitions, optimize costs etc.)*
- *Homogeneous/common environment for ‘all groups’, e.g. rules, IB, GPFS …*

Maxwell cluster batch resource (by IT)

What is the Maxwell cluster?

- **A large number of powerful computers (**named max-<something>)
  - All connected through a fast low latency network (56Gb EDR/FDR [partially faster])
  - All connected to Petra3 GPFS storage (and CFEL, EXFEL, CSSB storage)
  - All connected to dCache ("on demand")
  - All equipped with 256GB up to 1.5TB of memory per node
  - Quite a number of nodes with 1-4 Nvidia P100/V100 GPUs, also some A100 nodes
  - Lots of software pre-installed
- **Main purpose**
  - High Performance Computing
  - Offline Data Analysis
  - Simulations of all kind
  - Remote Visualization
  - Any application which can make use of the special features of Maxwell!

E.g. Ansys, Comsol, Fdmnes (MPI version), Matlab, OpenFOAM, Orca, Quantum espresso, Tensorflow, Xds, Xmimsim, XRT …..

E.g. Conuss less well suited (single threaded/few mem. ….)

- **All jobs are scheduled by the SLURM scheduler (via submission hosts)!**
  - Usually jobs don't have to wait very long
  - But it depends on the jobs requirements
  - and there is no VIP fast lane …
- Interact. nodes aka max-fsc/-fsg, desy-ps-cpu/-gpu or the new FastX3 ones max-fs-display are NOT part of SLURM

Maxwell

- Basic work principle

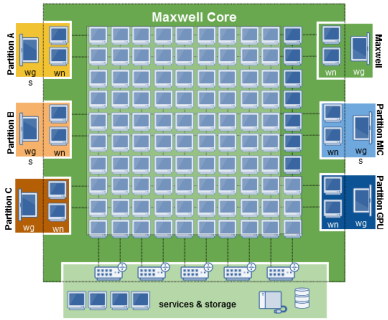
User → **Login/connect** → Job submission node

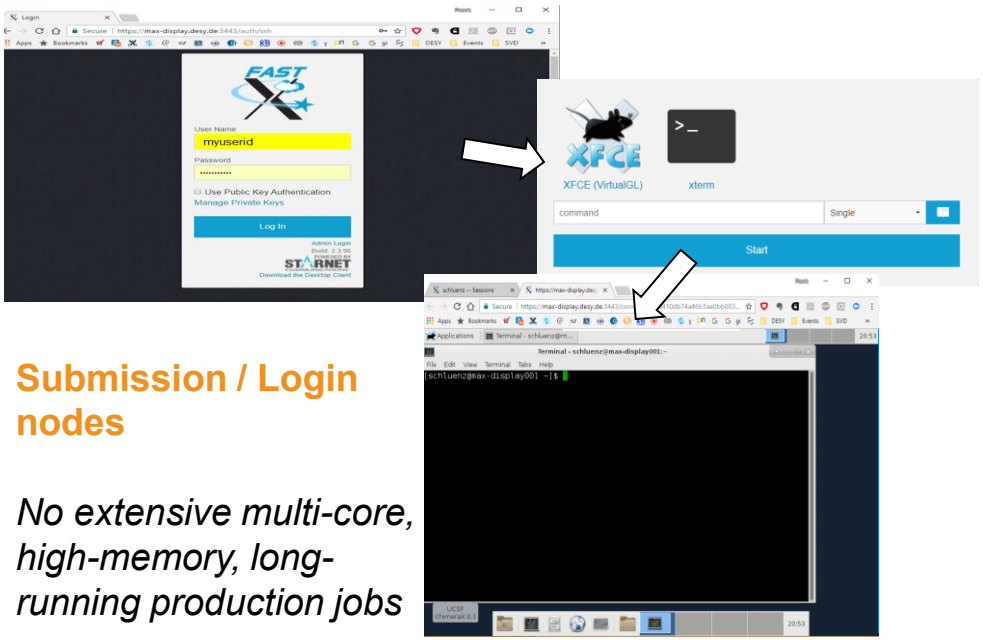
ssh
https

Prepare
Develop
Test
[exec small jobs]

→ **Submit job** →

sbatch
[salloc]





Submission / Login nodes

No extensive multi-core, high-memory, long-running production jobs

"Maxwell cluster"
(=managed by SLURM)

job queue
Exclusive resource per job
CPU / GPU nodes

Accesses / work in GPFS core (or beegfs or …)

Technical Implementation

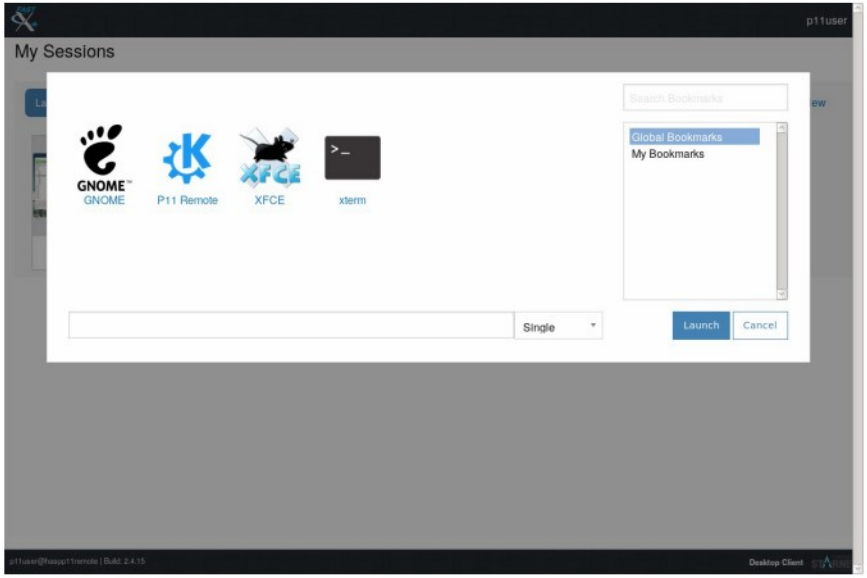
Make it working

FastX

- Commercial software, license available
- Already in use for Maxwell access
- X server in a browser
- Sessions can be shared
- Running on a dedicated host

Tailored X session

- Kiosk mode KDE
- Experiment control GUI
- Beam position monitor / feedback
- Browser (results, cameras, wiki)
- No terminal!



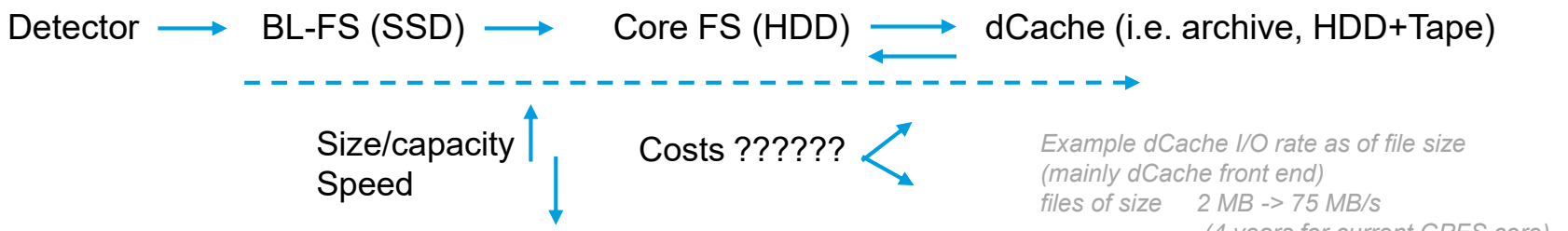
Data acquisition & analysis & management: integration has become increasingly important

detector PCs; network, proxies, storage, compute resources, remote access or download services, PSX accounts

Impossible to scale past/current ways of conducting experiments and data creation rates into near future

Selected items to address

- One can not optimize/update one part in the chain without having impacts on others
  - Storage size(s) depends on incoming data rate, data amount and dwell time(s)



  - Data chain is likely not a one-way path (i.e. re-staging)
- Costs for IT infrastructure have to be considered when purchasing detectors
- Storage and Computing may not scale linearly → technology leaps
- Changes imply coordination of FS and IT
- Technical considerations
  (distances on campus; space/power/cooling; specialized detector HW vs. “huge computer”)
- Legal issues, special terms of use (not become [I]SP; classification by vendors or discounts)

Thx for your attention!

Questions?