Connection between Beamlines and the Compute+Storage infrastructure.

Experiment Control and Data Handling at PETRA III

(and partially DESY FS-*)

André Rothkirch
DESY Photon Science, FS-EC (Experiment control group)

and/in collaboration with colleagues from FS-EC, DESY Central IT and the beamlines and …

HELMHOLTZ

DESY.

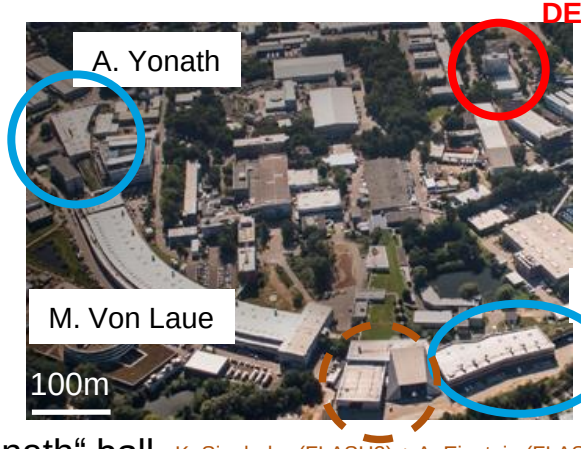
Experiment control - FS-EC tasks/supports

- Digital user office
- Experiment control
  - Design and development of tools and procedures as requested by our users.
  - Deployment, adaptation and extension of collaboratively developed software
    - Control software & HW access layers: Tango, Sardana, Taurus, NeXus.
  - Development and implementation of new methods
  - Hardware integration
  - Consulting, coaching and trouble shooting
- Data management
  - data flow, data access & more
- Online/offline computing, data processing, science software
  - Consulting, coaching (resource access; slurm), trouble shooting
- Electronics (incl. development, e.g. PiLC)
- Operating systems, PC hardware
- User administration

Today put focus

!!! in particular in close collaboration or with or operated/ maintained by Central IT !!!

PETRA III Beamlines



DESY Compute Center

„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 (now **Hereon**)
European Molecular Biology Laboratory

„Ada Yonath" hall K. Siegbahn (FLASH2) + A. Einstein (FLASH)

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.; Sept. 2025 likely beam in OH)*

„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.; likely 2026)
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; timings etc.*

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

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

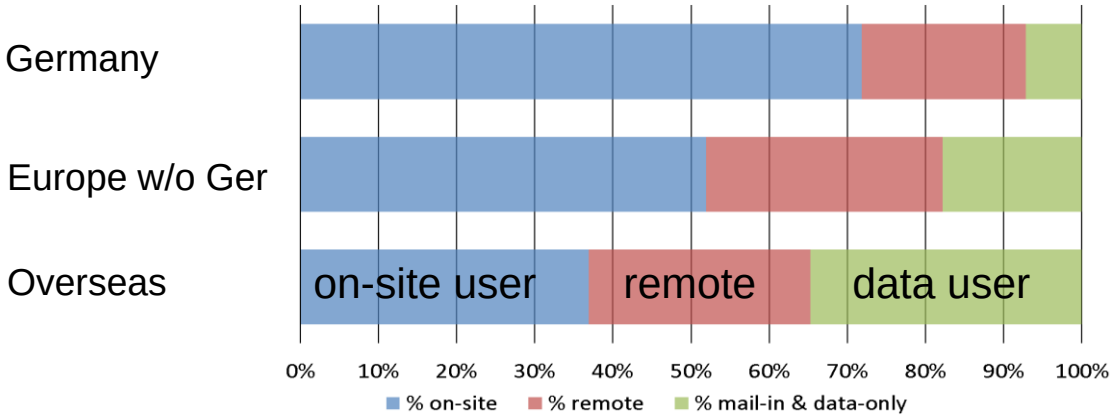
PETRA III

User statistics

- 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 control system requirements

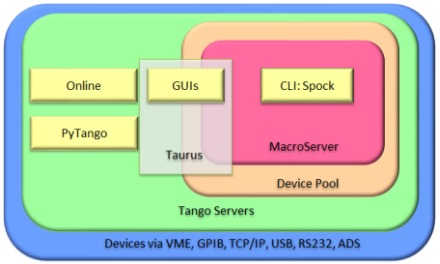
- Supported by an international community
  - collaborative code development saves resources and improves code quality
  - Sustainability & re-use of code (e.g. Tango Servers)
  - communication on related issues
  - standardised user environments
- Performance, reliability
- Client-server structure, multi-client access
- HW access layers & Control software
- Synchronisation of devices
- Flexibility: support various experimental techniques
- Customizable
  - allow for discipline-specific applications
  - allow for extensions by users
- Cope with fast detectors resp. more ‘useful’ photons allow for faster scans
  - Generate higher data rates (or higher acq rates of smaller amount of data)
    - ***Challenges: Reduction; Transfer; Storage***
    - ***Design of concepts to cope with high data rates and large data volumes***
    - ***Fast feedback***
    - ***Analysis***

Control software & hw access layer

TANGO
Connecting things together
consortium 11 members

Tango (developed at ESRF)

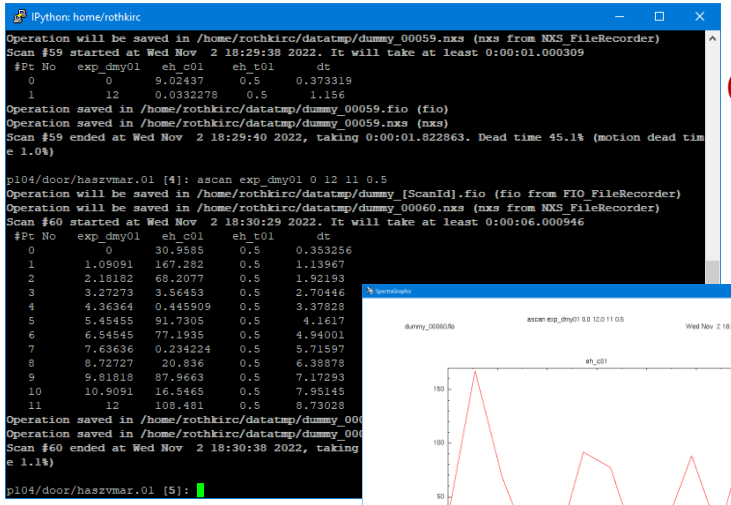
- Tango as the hardware access and transport layer is well established
- Device API: attributes, commands and properties
- Multi-client access
- Reliability & speed OK, note that 2D data have direct data paths
- Generic applications: astor, jive, pogo
- User extensions: PyTango scripts, virtual counters and motors
- > 50 institutes involved; DESY contribs: 284 servers (T. Nunez, Y. Yu, J. Kotanski, H. Blume)



Online: Experiment control program, DORIS era

EC User Interfaces – Sardana

(DESY contributions in development and maintenance by T. Nunez)



CLI

- Sardana features
  - Python
  - **Spock (CLI)**:
    - an ipython implementation
    - well established syntax
  - **Scripting engine**
    - Macros (system, DESY, user), hook mechanism for customizing macros
  - **Taurus: GUIs**

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 & protocols
- scans, makros, viewing, processing, interpreting data

End of the experiment

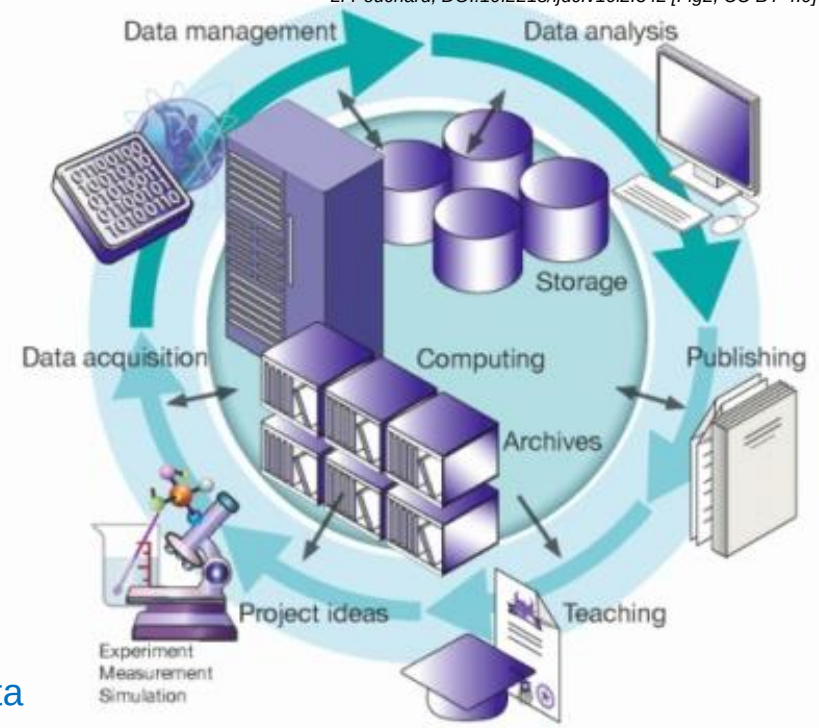
- Data not accessible for next user group

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

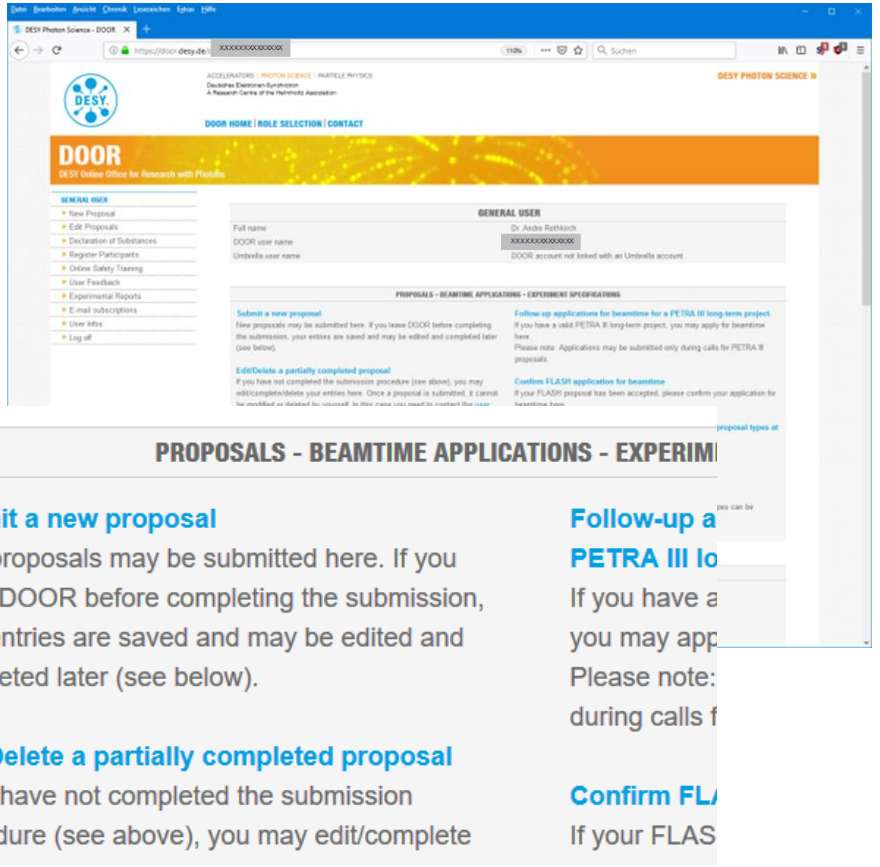
Common issues: space, performance & reliability

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

Close collaboration of Central IT and FS-EC

- Invention of a managed storage system *(1st concept implemented in 2015)*
  - 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 & constraints (SSD based)
    - Limited visibility of the temporary storage to beamtime and beamline
  - Start-/Stopbeamtime to create temporary and permanent directory structure
  - ACLs for permanent/on longer terms storage ("Core-FS") (HDDs)

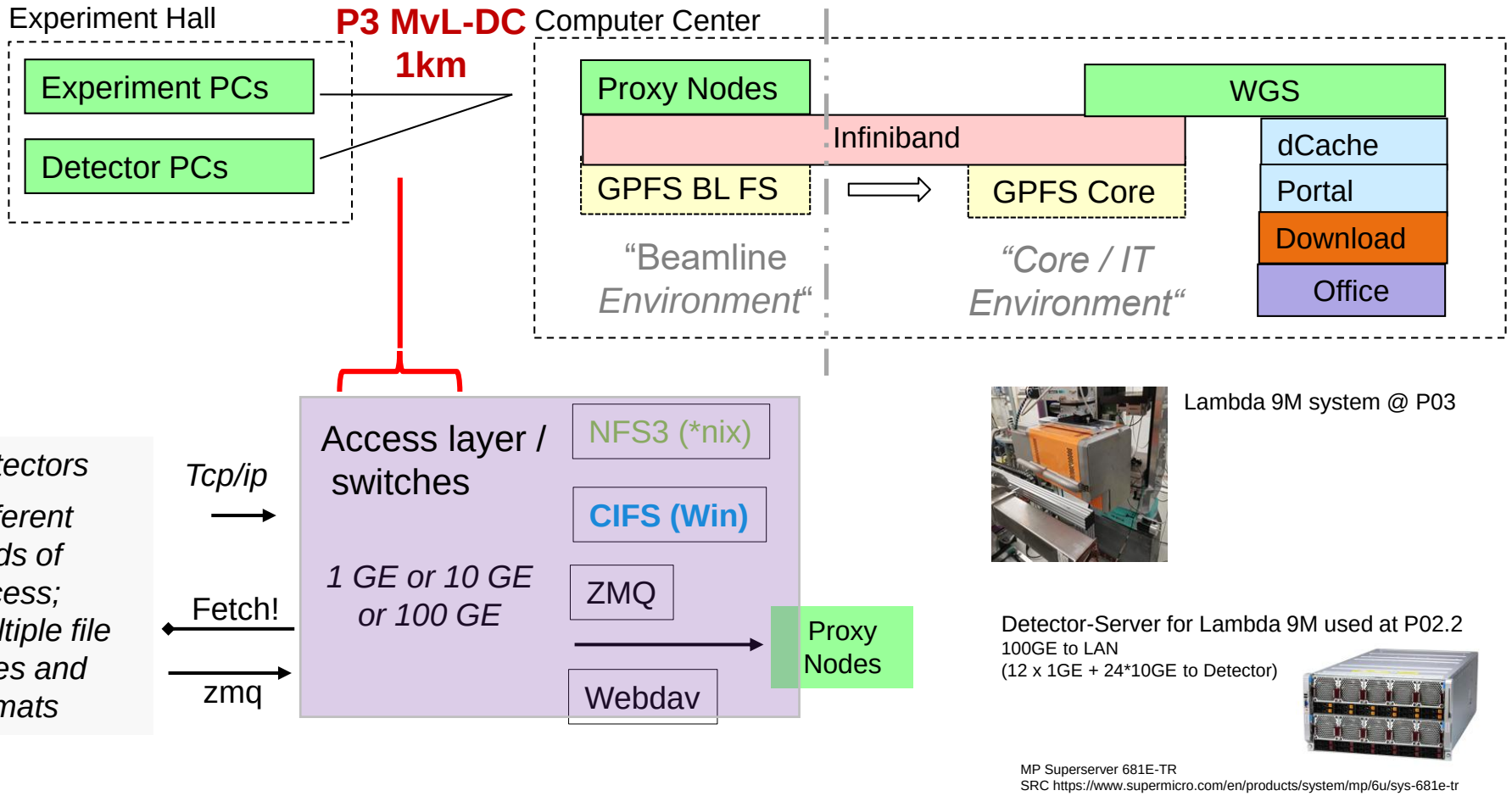
- IBM GPFS Storage Server (IBM Spectrum Scale & Elastic Storage Server)
  [General parallel filesystem (GPFS) is a high-performance clustered file system]
  - IBM 5146-GS1: ~55 TB; 2.5" 10K rpm **HDD** (1.2 TB) or 2.5" **SSD** (400 GB or 800 GB).
  - IBM 5146-GL4: ~700TB; 3.5" **NL-SAS HDD**s (2 TB or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

  (initial 2015)

  Meanwhile several times expanded and first systems already replaced
  - Currently BL-FS 440TB SSD and Core-FS ca. 17 PB HDD total

  (last update 2025)

DESY Photon Science GPFS storage system (schematic)



Distances max vary depending on Experiment Hall (M.v. Laue, A. Yonath, P. Ewald., A. Einstein, K. Siegbahn or ….)

Start/Stop a Beamtime *[by Beamline staff]*

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

- Instantiates beamline (BL-FS) and core (Core-FS) filesystem, i.e.
  creates file sets with predefined directory top-level structure and rules/constraints
- BL-FS: NFS3 + SMB (and Hidra) based on whitelist (hostname/IP)
  - 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 list/ACLs 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 Beamlie staff]*

Predefined directory toplevel structure and rules/constraints

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

"Beamline world"

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]

Since 2021:
"passthrough mode", 10TB
(i.e. data becomes 0 Byte in BL-FS [NOT in core], is no more editable and ignored in transfer)

Beyond/outside "Beamline world"

Permanent storage **(on ACLs)**

"GPFS Core file system"

GPFS - native on Analysis nodes
- or by NFS 4 (typ. krb5) & SMB (office)

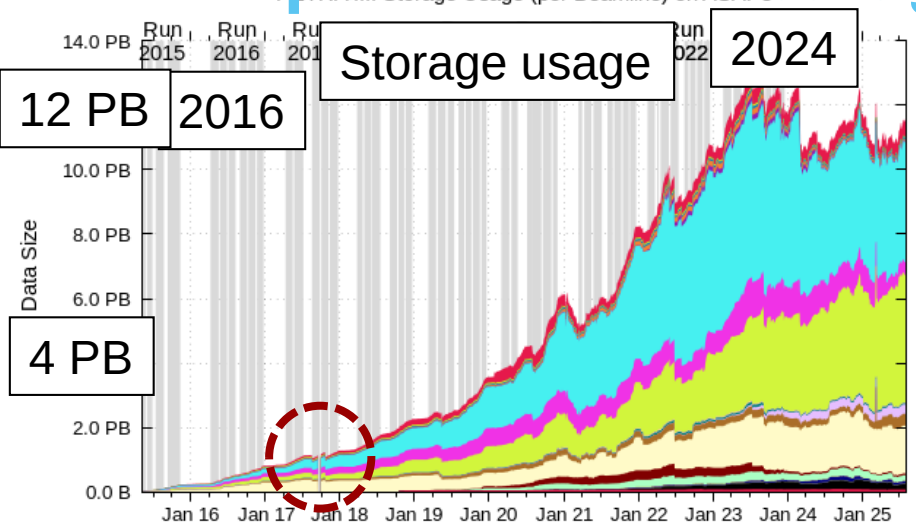
/asap3/
- <facility> (e.g. petra3)
  - gpfs
    - <beamline>
      - <year>
        - <data>
          - <beamtimeID>
            - raw *(read only excl. dmgt)*
            - processed *(read/write)*
            - shared *(read/write)*
            - scratch_cc *(read/write; NOT archived)*

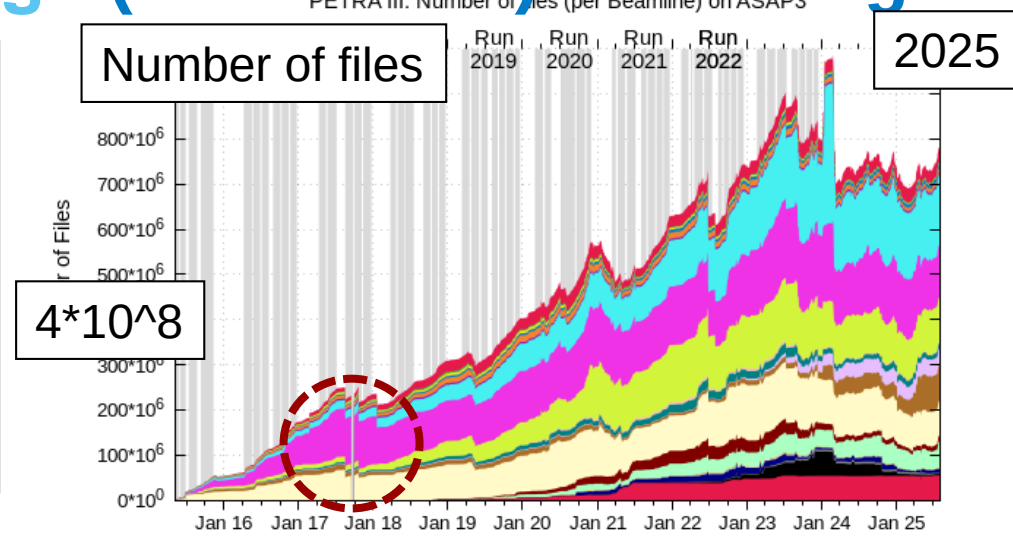
WGS
dCache
Portal
Download

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

note: 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, finally 4.0 PB)
P3 2023: ~4.6 PB
P3 2024: ~4.2 PB
P3 2025: ~2.3 PB (status as of 06.08.25)

New detectors [Eiger+Lambda 2M&pool]

New detectors [Eiger+Pilatus3 2M]

More BLs & efficient use

Amount per BT: 100MB to 100+ TB (in ~2-3 days)

Currently 4-5 PB / year at PETRA III

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]

Archiving/portal

Current data „lifetime“ and archiving

(init as of Feb 16, 2022)

- 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, currently after 180 days if not used or up to 720 days)
  - 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 dwell time prolongation possible (if constraints allow)
- Data export and Management → gamma-portal (+ altern. Globus [ro])
- So far nothing deleted (from tape)

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

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

Implementation of a data policy according to the rules worked out by PanData similar to other sources in Europe is envisaged (F.A.I.R.)

Archiving/portal (cont.)

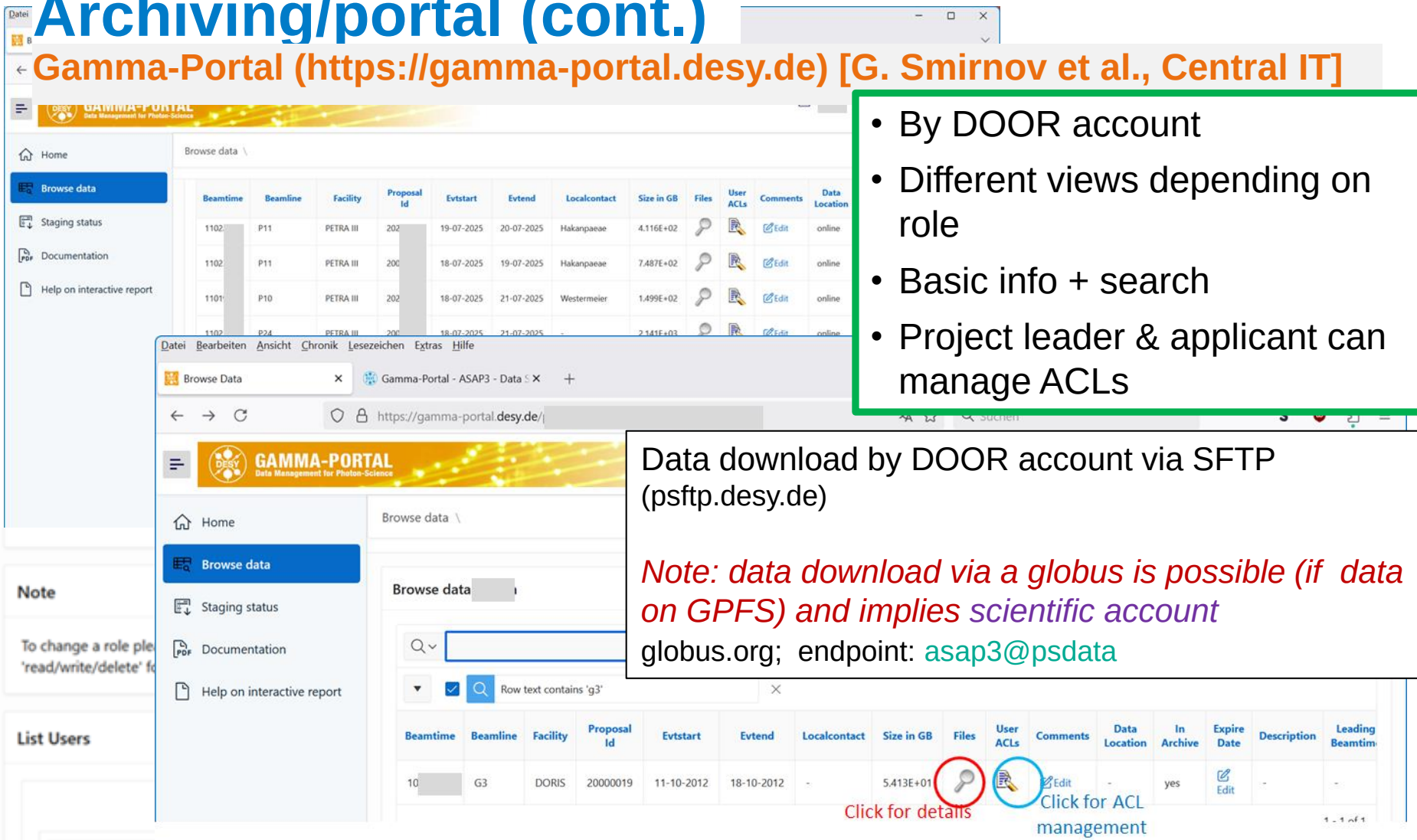
Gamma-Portal (https://gamma-portal.desy.de) [G. Smirnov et al., Central IT]

- By DOOR account
- Different views depending on role
- Basic info + search
- Project leader & applicant can manage ACLs

Data download by DOOR account via SFTP (psftp.desy.de)

Note: data download via a globus is possible (if data on GPFS) and implies scientific account

globus.org; endpoint: asap3@psdata



archiving/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 (partially changed) -> 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 or Lambda 9M (installed at 6 beamlines already)
  (100G QSFP28 LR4,10 km, λ1296-1309 nm resp. 100G QSFP28 CWDM4, 2 km, λ1271-1331 nm)
- GPFS used now at PEX, FLASH & FLASH2, “Special instruments” [e.g. labs] & more
  as of Aug 2025: 77 GPFS instances
- GPFS for external BTs (i.e. experiments by PS staff carried out not at Desy or Xfel)
  as of Aug 2025: 432 external BTs scheduled
- GPFS for groups/research teams
  (no external users, only group ACLs, no portal, different “scheduling”)
- GPFS capacity expanded multiple times, from on 2017/18 each year
  (currently BL-FS ~440T SDD / GPFS-core ca. 17000 T ~ 17 P)
- GPFS replacements to keep state (i.e. HW out of warranty)
- Invention of HiDRA and Lavue & further developments
  (well received by our Beamlines) [e.g. Lavue: https://lavue-org.github.io/lavue/stable/index.html]
- Procedure to remove data from GPFS storage (‘GPFS Core’, HDD) in 2017
- Copy procedure BL-FS to Core improved/modified in 2018 -> inotify on proxies
  (i.e. at moment a fast (inotify) and slow (each few minutes) copy is active)
- Passthrough entered into force in 2021 for all Beamlines (i.e. remove data from BL-FS if > 10T)

Further findings and action taken

Growing amounts of data, increasing complexity of experiments (or modelling) and data analysis lead to special requirements regarding software and hardware

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 compute resources

“Wishlist” ;-)

- amount of RAM
- number of cores (and/or nodes)
- GPUs
- GUI (also for opengl stuff)
- **Exclusive** resource usage
- Interactive option (ssh / FastX whenever one likes)
- “Fast” processing

There exists no one system fits all :-(

Maxwell environment as one part to support the research / data analysis [on best effort basis]

maxwell & remote analysis

- Invention of Scientific accounts with own namespace ‘psx’
  (i.e. DESY accounts for external users (‘external’ ≠ industry/commercial)
- Provision of interactive resources max-fs-display
  (max-fsc/max-fsg will be [were] shut down early 2023)
- Creation of specific batch resources for PS managed by SLURM
  - Slurm partitions 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)

Update in summer/autumn 2024:
Partition relayout, distinguish cpu / gpu nodes

- Slurm partitions pscpu + psgpu (inhouse)
- Slurm partition psxcpu + psxgpu (external [non-commercial] users or use cases)
- New partition for ‘partial node’ jobs (“Solaris”)
  e.g. “sbatch --partition=solcpu –cluster=solaris –mem=4G –cpus-per-task=4 --ntasks=1“

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 …*
- *Docu:* https://maxwell.desy.de

SRC: A.R. et al., Satellite Meeting: News from Computing for Photon Science, DESY PS | XFEL Users Meeting 2024

Maxwell & remote analysis

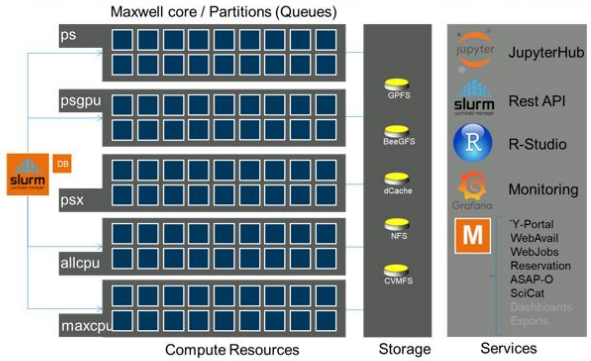
Basic working principle

User → **Login**: https, FastX3client, ssh → Submission / Login / „interactive“ nodes → **Submit job**: sbatch [salloc]

- **max-fs-display.desy.de**
  Dedicated PETRA III / IV, FLASH/FLASH2, Hereon
- **max-display.desy.de**
  All maxwell users including PS

Purpose Remote visualisation
Prepare /Develop/Test, run small jobs*
Job submission to cluster

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



“Maxwell cluster” (=managed by SLURM)

job queue
- *Exclusive resource per job*
- *CPU / GPU nodes*
- *constraints: nodes per job / max duration ..*
- *(depending on ‘partition/HW owner’)*

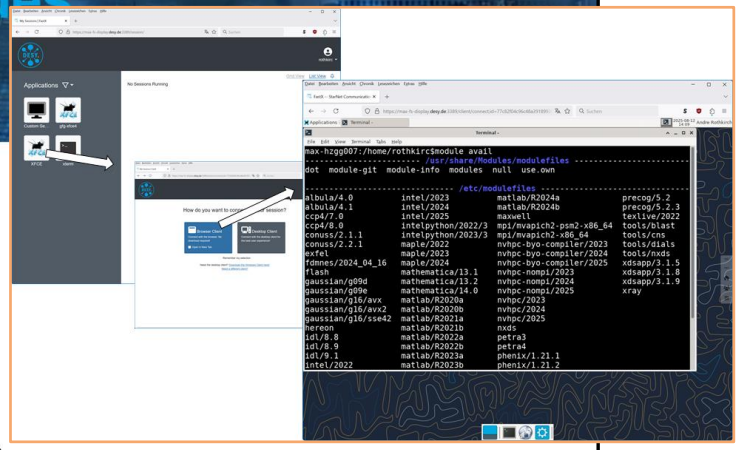
Accesses/work in GPFS core (or beegfs)

NOTE:

retired: max-fsc, max-fsg
retires next: max-nova, likely soon
[replacement for both is max-fs-display])

The interactive/shared/job submission nodes

Load balancer, access & guideline



„Maxwell“ nodes “max-fs-display“ (and max-display.desy.de)

- Personal accounts to use, NO functional ones
- (partial) access by registry resources „max-ps2“ or „max-psx2“
  (which allow usage of slurm partitions ‘ps’/psgpu and ‘psx’, respectively)
- ‘Interactive’, ssh / FastX3 => SHARED resource
- “Local” home(login) directory on GPFS
  (30GB quota, will NOT be extended, NO backup/archive -> have a copy “at home”)
- Access to BT data residing on GPFS (raw / processed / shared / scratch_cc)
- Scratch space(s) avail, typ. SHARED space among all users
  delete files by yourself as soon as possible to free capacity (it should not become ‘data grave’)
- USERS are recommended to account use “policies”/principles of common behaviour in view of the shared resources (selection see next slide)
- Same OS and software as the nodes in the cluster; *same ‘home dir’*
- *(still) Cent OS 7.9.2009*

NOTE: recent and future changes by talk Frank S. and Sven S.

OS update in summer/autumn 2024: Now RHEL 9
Multi Factor Authentication implemented (curr. 2FA) As bonus: self service password reset

Graphical Login

Features and constraints / use cases / policies

max-fs-display.desy.de & max-display.desy.de are primarily for graphical logins and applications requiring GPU hardware acceleration

Features
- load balanced login
- persistent sessions
- hardware acceleration
- session sharing

Use cases
- job submission & monitoring
- graphical applications
- software development /code compilations (also for cuda)
- low-profile computation
  - ~1-4 CPU cores
  - <10% memory (thus typ. < 40GB)
- Very short computational jobs for testing (< ? min)

Mis-Use cases
- compute intensive jobs, in particular GPU jobs
  (primary task of the GPUs is rendering)
- crypto mining
- Unlicensed/illegal software

Monitoring
- heavy resource usage
  - Lowering priority
- continuous heavy load
  - session termination
- Disabled by now (enable again?)
  - email every NN minutes

Some numbers on Maxwell

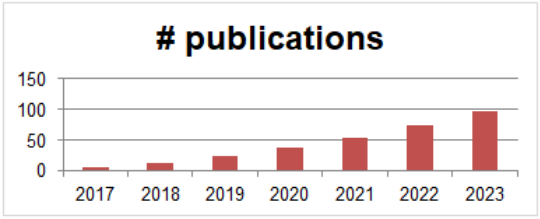
Nodes in ps* + psx*: total 44

SLURM partitions (selection)

Partition	*Users*	*Preemption***
pscpu / psgpu	PS staff (i.e. "hasylab")	no
psxcpu / psxgpu	PS users	no
upex	EuXFEL users	no (?)
allcpu / allgpu	everyone	Yes (requeue)
maxcpu / maxgpu	DESY*	no
short	DESY*	no

* DESY staff with *mpi-like* applications

** Preemption means that a job running in a "preemptable" partition can be terminated by any job in a higher priority partition containing the same nodes. Terminated jobs will either be requeued - the job will go into the queue as if it was never running - or simply cancelled.



Some statistics / numbers in 2023

Maxwell Users	**~2800**
Num. SLURM users	~950
Num. jhub users	~700
Num. pure batch users	~450 (not using jhub at all)
Num. batch jobs	~4.000.000
Num. jupyter jobs	~8.000 (0.2%)
Num. of issues	650 (incl. some admin tickets) [0.0006 per user/day ;-)]
Resolved issues	649

Thank you!!!

Questions?