Computing & Storage for on-site experiments - Petra3, FLASH and EuXFEL

Martin Gasthuber, ASAP[3], FS-EC and Eu-XFEL Team
BNL, September 24, 2018

the bigger instruments…

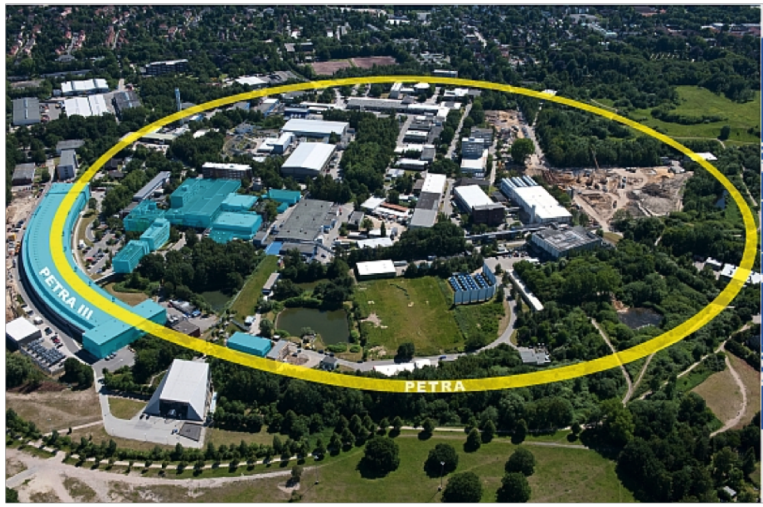
- currently three on-site accelerators are in operation
  - Petra 3
  - FLASH
  - EuXFEL

- all three share some of the computing infrastructure and have much of the architecture and ‘mode of operations’ in common
  - same team at central IT complemented by teams at each instrument (and vice versa ;-)

Petra 3

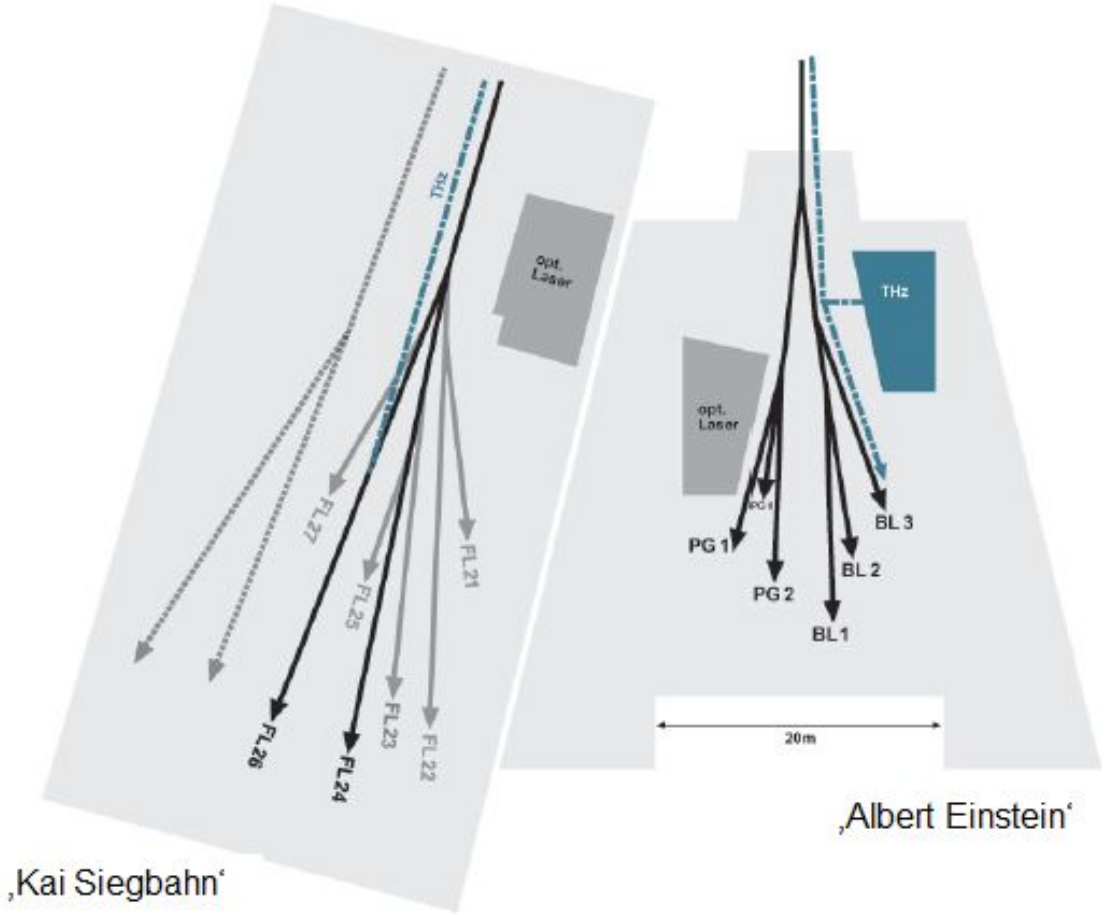
oldest workhorse, recently extended, inherited from particle physics era

- built 1978 for HEP Experiments
- Since 2009: 14 beamlines in operation
- Since 2016: 10 additional beamlines



FLASH 1+2

- Linear accelerator
- started as test facility for TESLA technology
- since 2016 using ASAP^3 for storage & analysis
- 12 beamlines (2 undulator lines) – one active per undulator



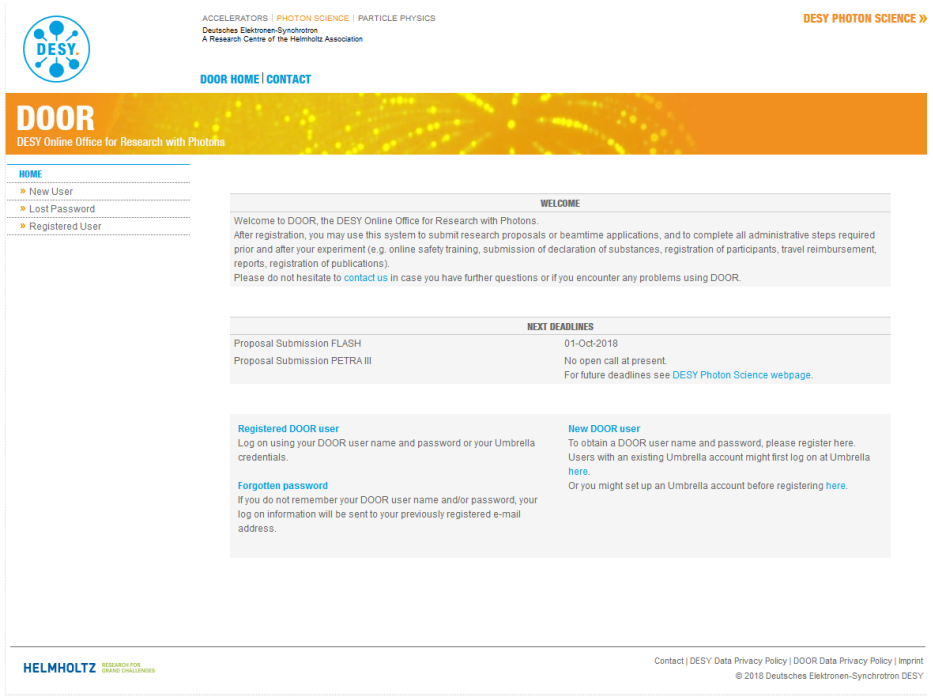
FLASH experimental halls

DOOR

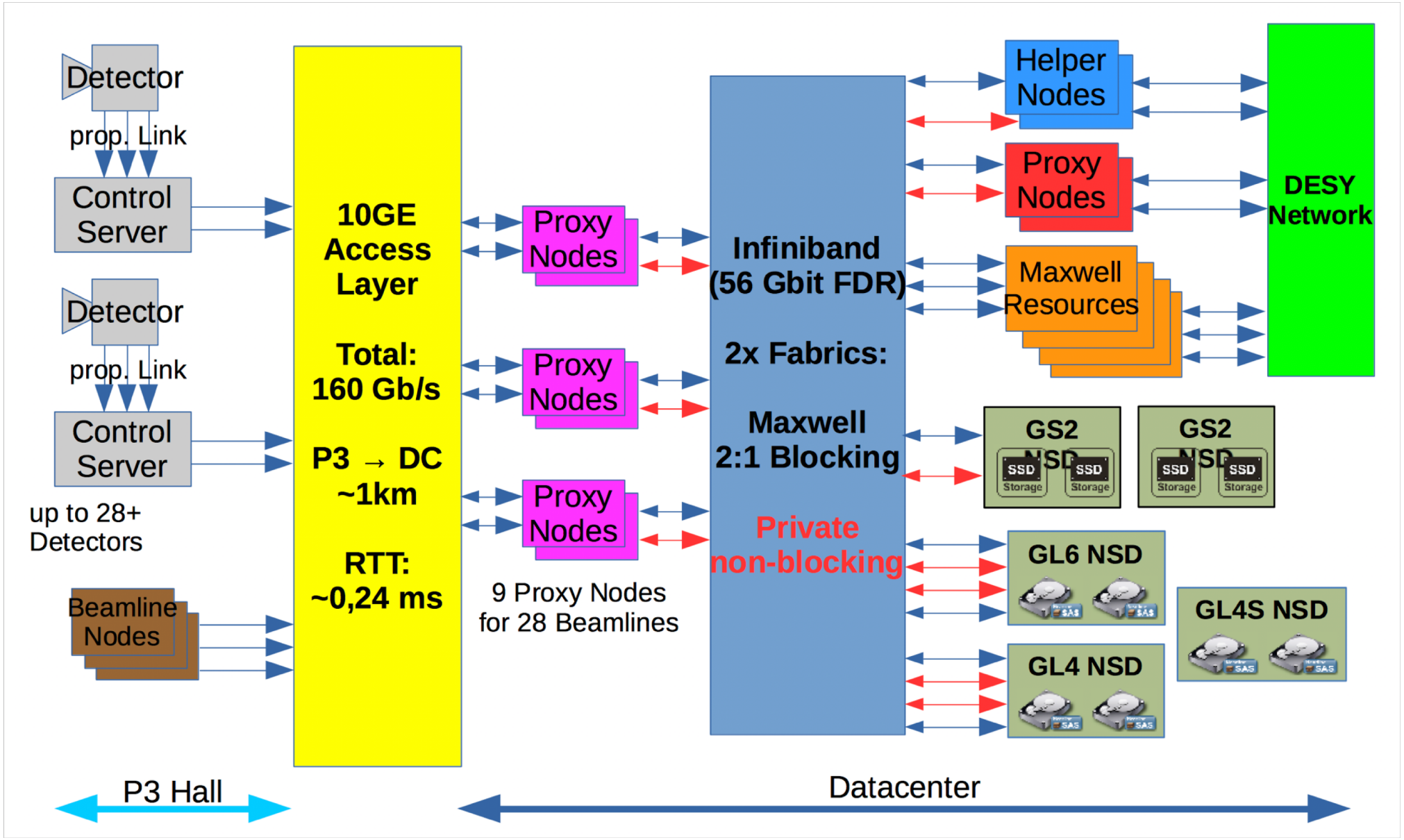
DESY Online Office for Research with Photons

DOOR

- Workflow system for experiments processes at DESY Photon Science (PETRA III, FLASH)
  - proposal submission, internal and external review, beam time scheduling etc.
- Web application
  - php-based
  - Permanent further development of DUO clone (originated at PSI, Switzerland)
  - CentOS Linux Server, Apache
- Data stored in central Oracle Database Server
- Role-based System
- Other DUO clones are in use at EXFEL (UPEX), Hamburg and MAX IV (DUO), Lund, Sweden

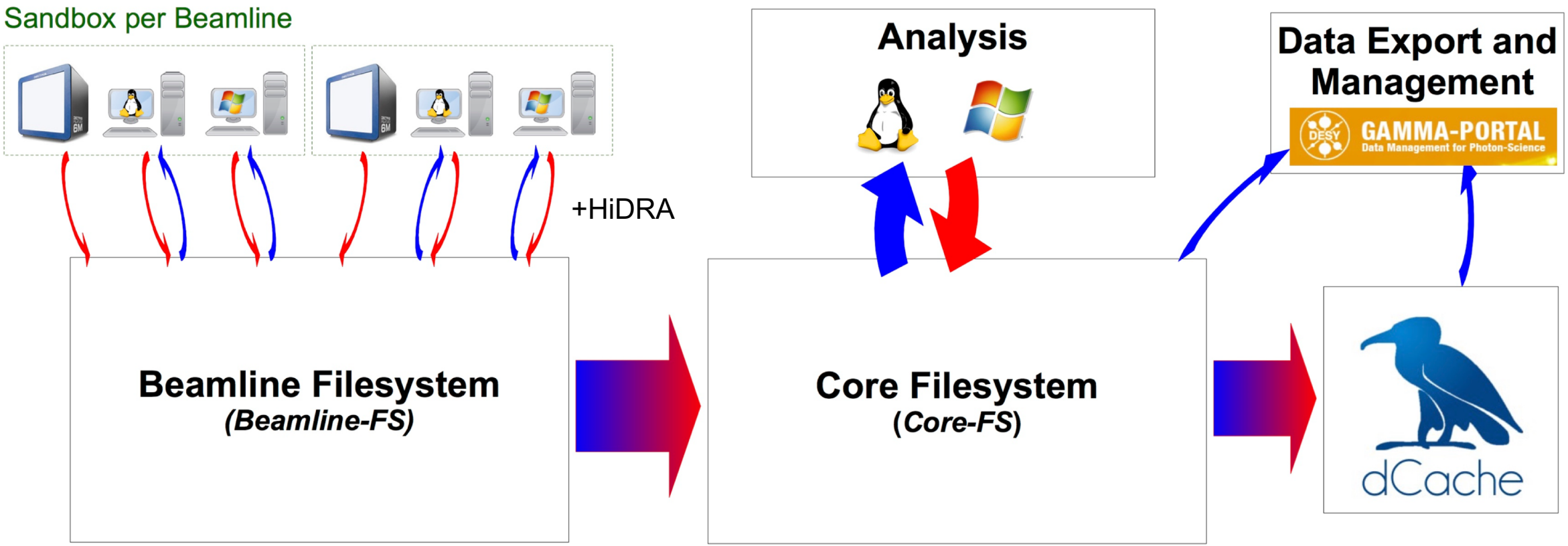


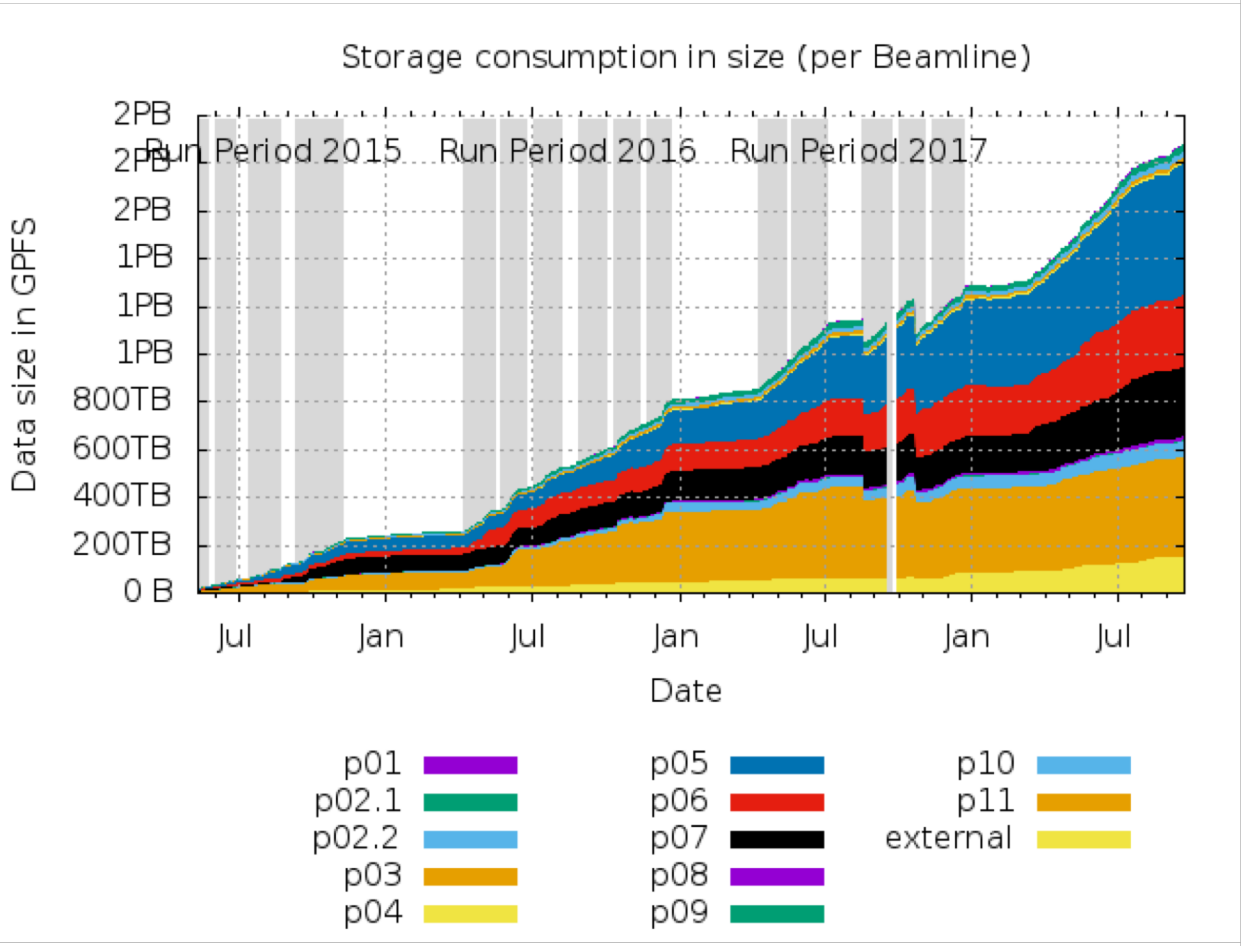
layout – networks, hardware

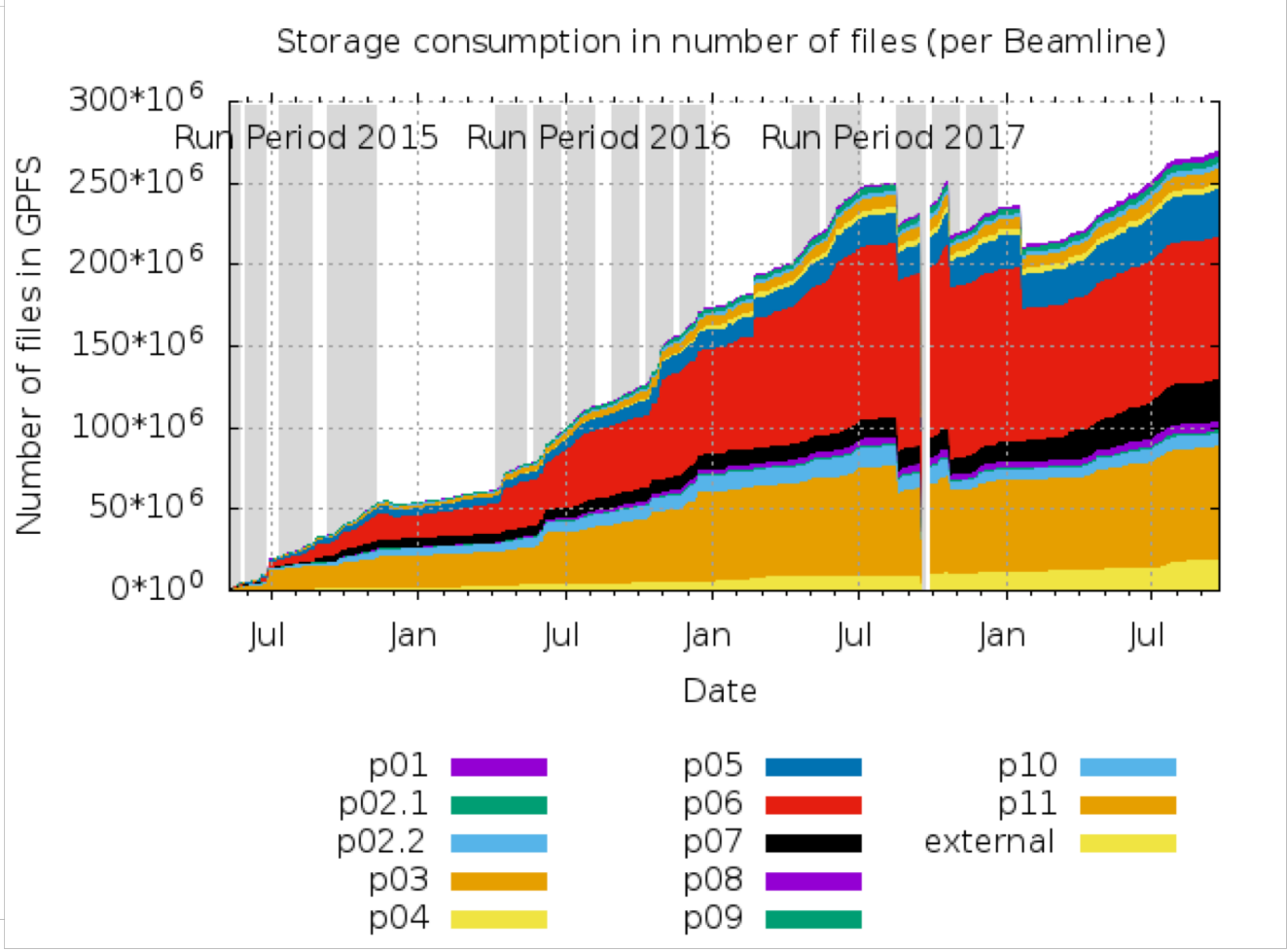


915 disk drives
96 SSDs
~540 Nodes

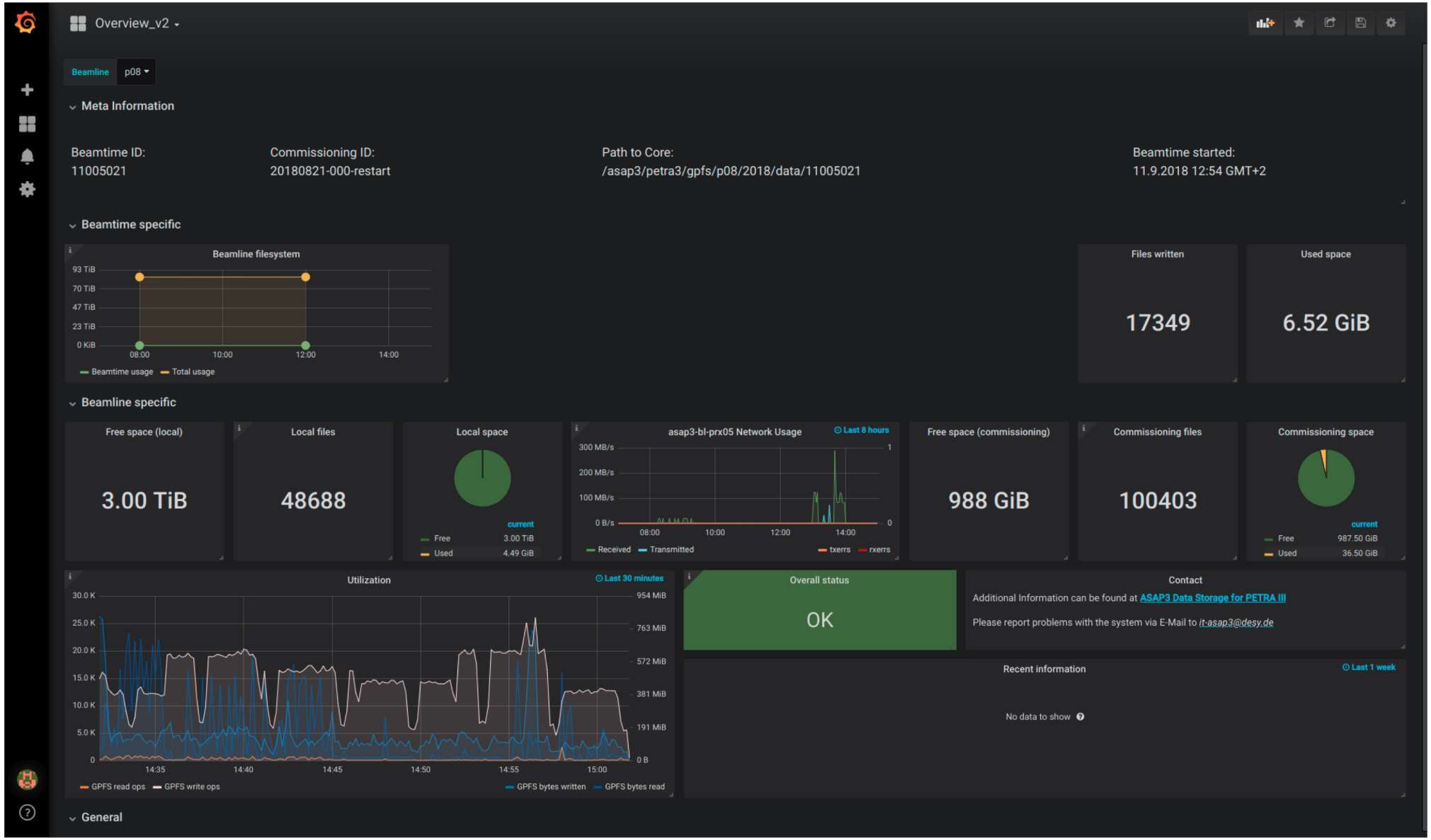
Overview – dataflow & services




Storage consumption in size (per Beamline)
Run Period 2015
Run Period 2016
Run Period 2017
Data size in GPFS
2PB
2PB
2PB
1PB
1PB
1PB
800TB
600TB
400TB
200TB
0 B
Jul
Jan
Jul
Jan
Jul
Jan
Jul
Date
p01
p02.1
p02.2
p03
p04
p05
p06
p07
p08
p09
p10
p11
external


Storage consumption in number of files (per Beamline)
Run Period 2015
Run Period 2016
Run Period 2017
Number of files in GPFS
300*10^6
250*10^6
200*10^6
150*10^6
100*10^6
50*10^6
0*10^0
Jul
Jan
Jul
Jan
Jul
Jan
Jul
Date
p01
p02.1
p02.2
p03
p04
p05
p06
p07
p08
p09
p10
p11
external

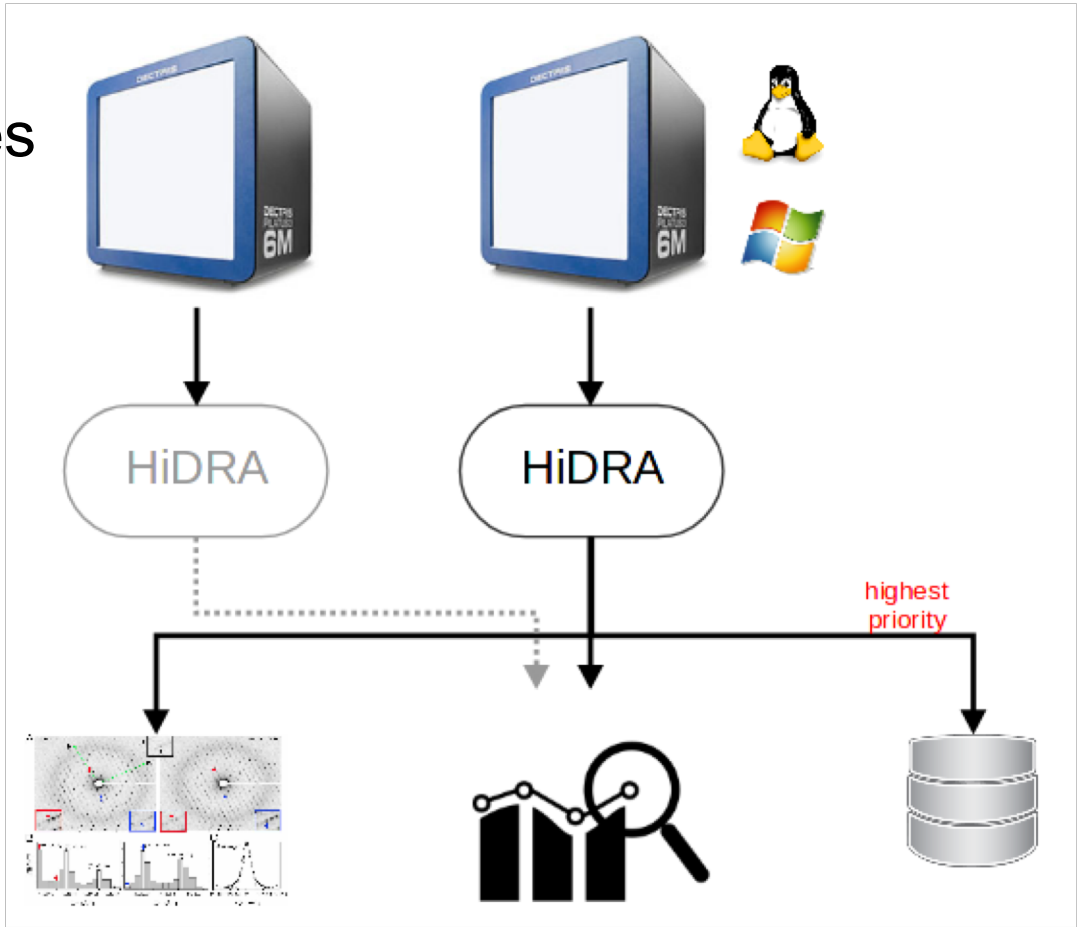
Dashboard – per beamline



HiDRA

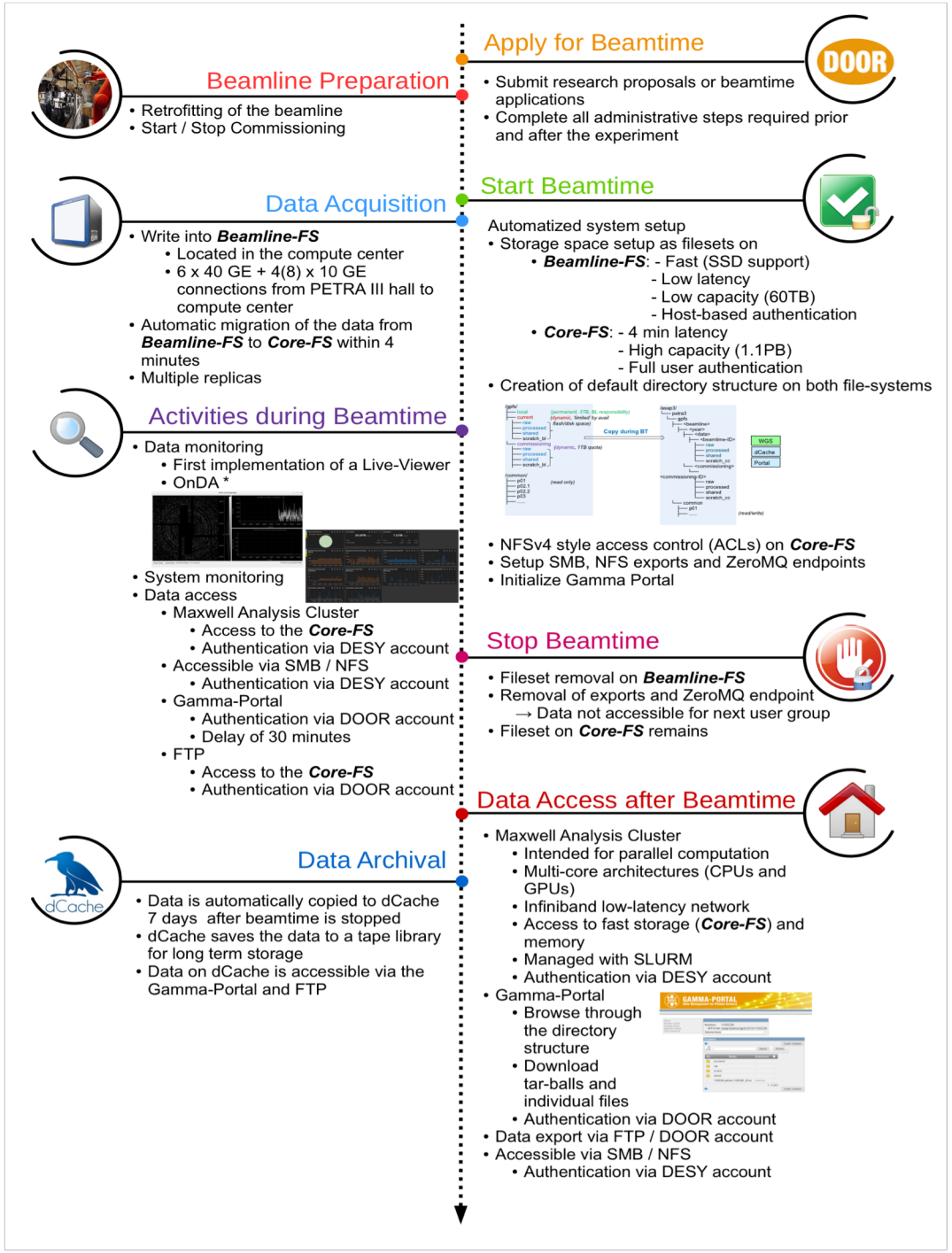
- High Data Rate Access
- Generic tool for high speed data multiplexing based on Python and ZeroMQ
- Actively used by FLASH and Petra 3 beamlines
  - data transfer from detector to GPFS
  - online monitoring & analysis
    - i.e. CFELs ONDA

- next generation in development
  - online & offline access – one API
  - query – include user key/value
  - scalable, more efficient

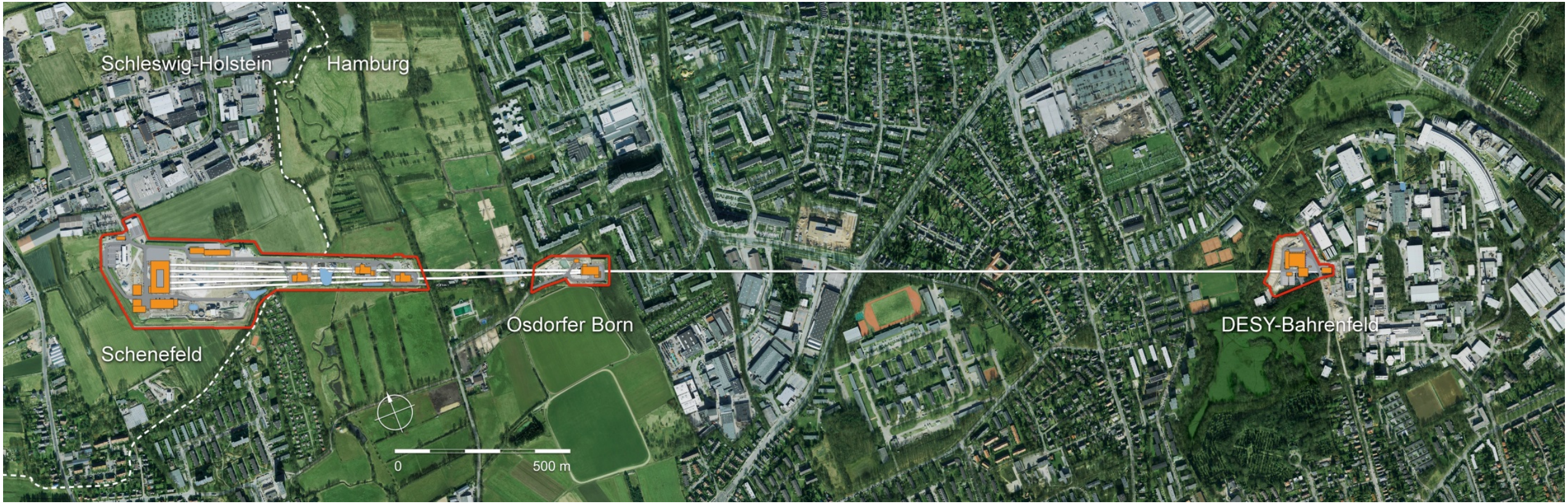


ASAP³ – glued services

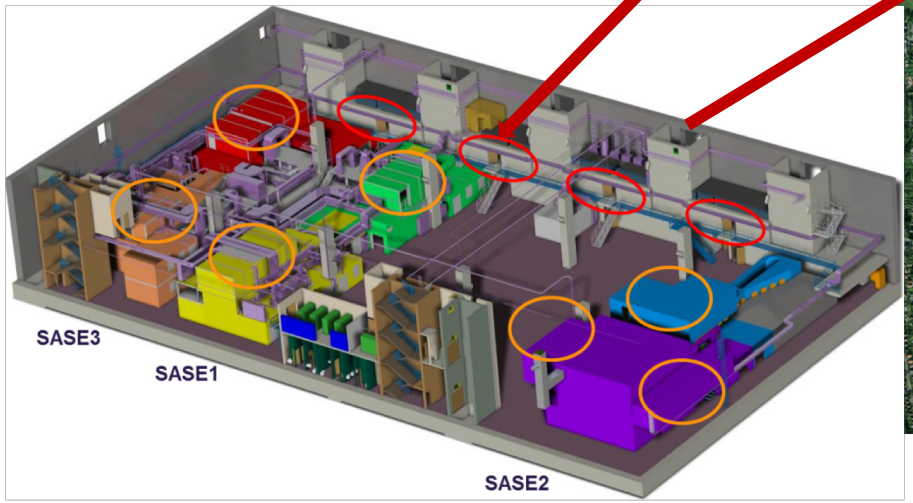
- GPFS
  - policy scans, XATTR, ACLs, multi-protocol access
  - GNR, end to end checksumming
  - fast – although it's a POSIX filesystem – see CORAL
  - startbeamtime/stopbeamtime preparation – automatic setup
- DOOR – beamtime metadata
- Gamma Portal – manage your beamtime data (visibility, ACL, …)
- external access – FTP
- dCache & Tape copy
- HiDRA – online data analysis/live viewer, cover 'first mile'
- dashboard – overview per beamline
- home build InfiniBand monitoring
  - clean and running network – first on the list ;-)
- many scripts 'glueing' – mostly python



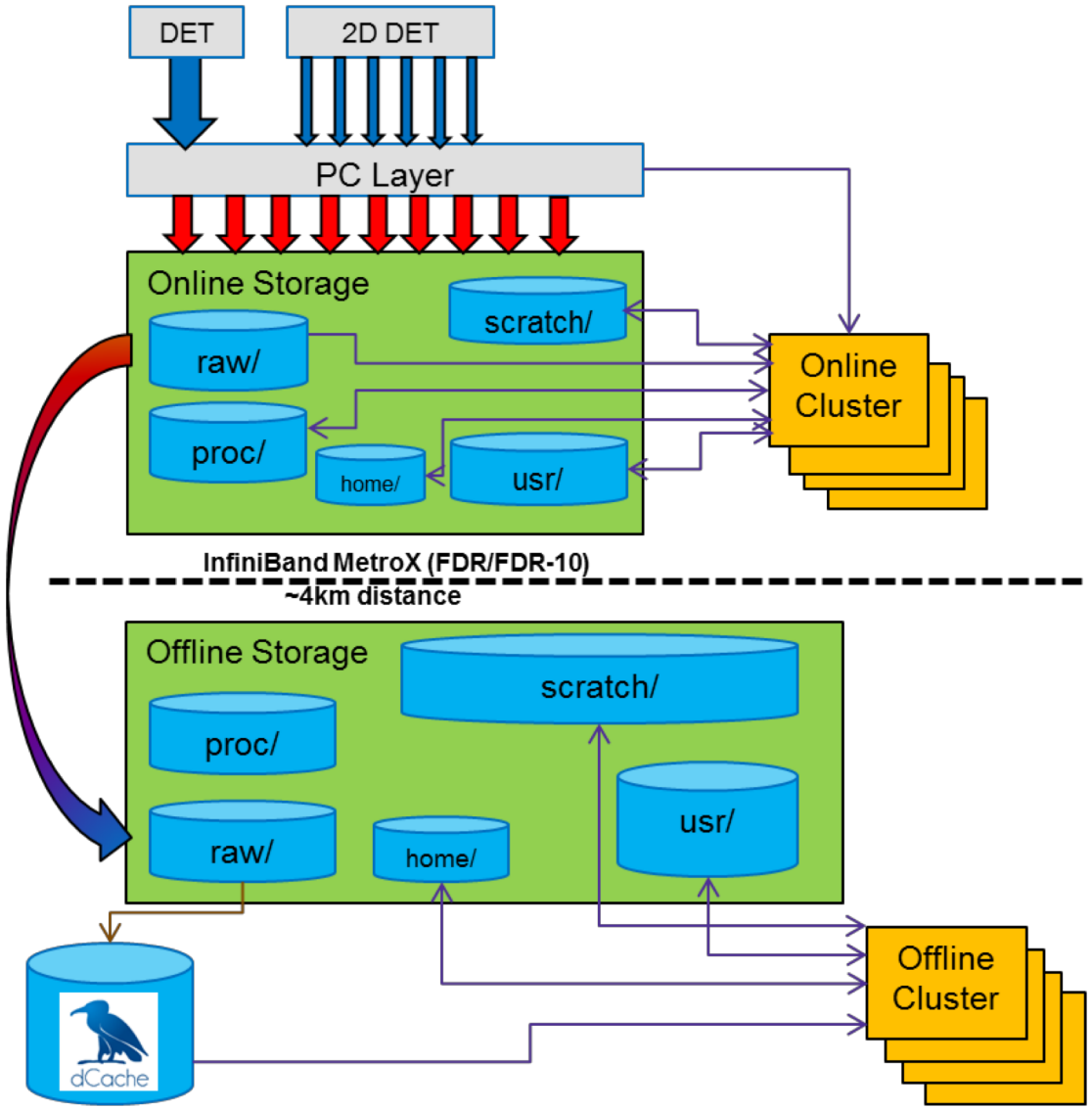
EuXFEL – a new research facility/instrument / September 2017



IT infrastructure at XFEL and DESY

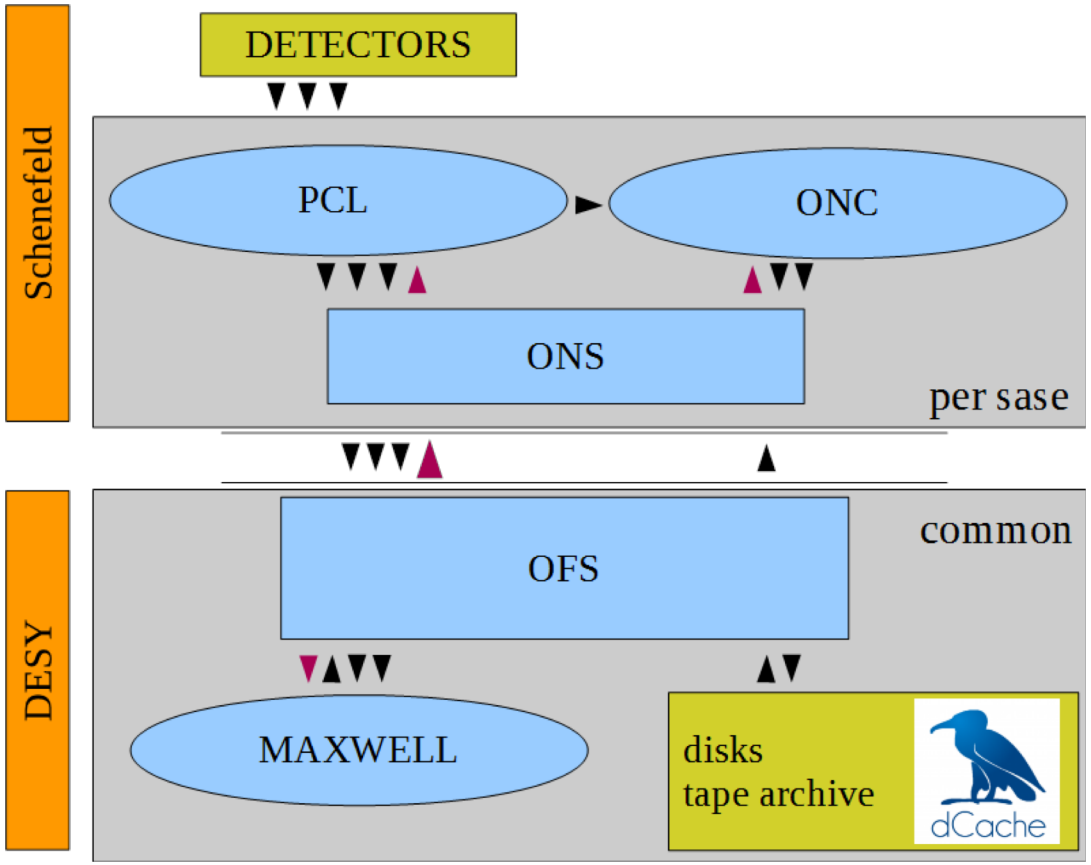


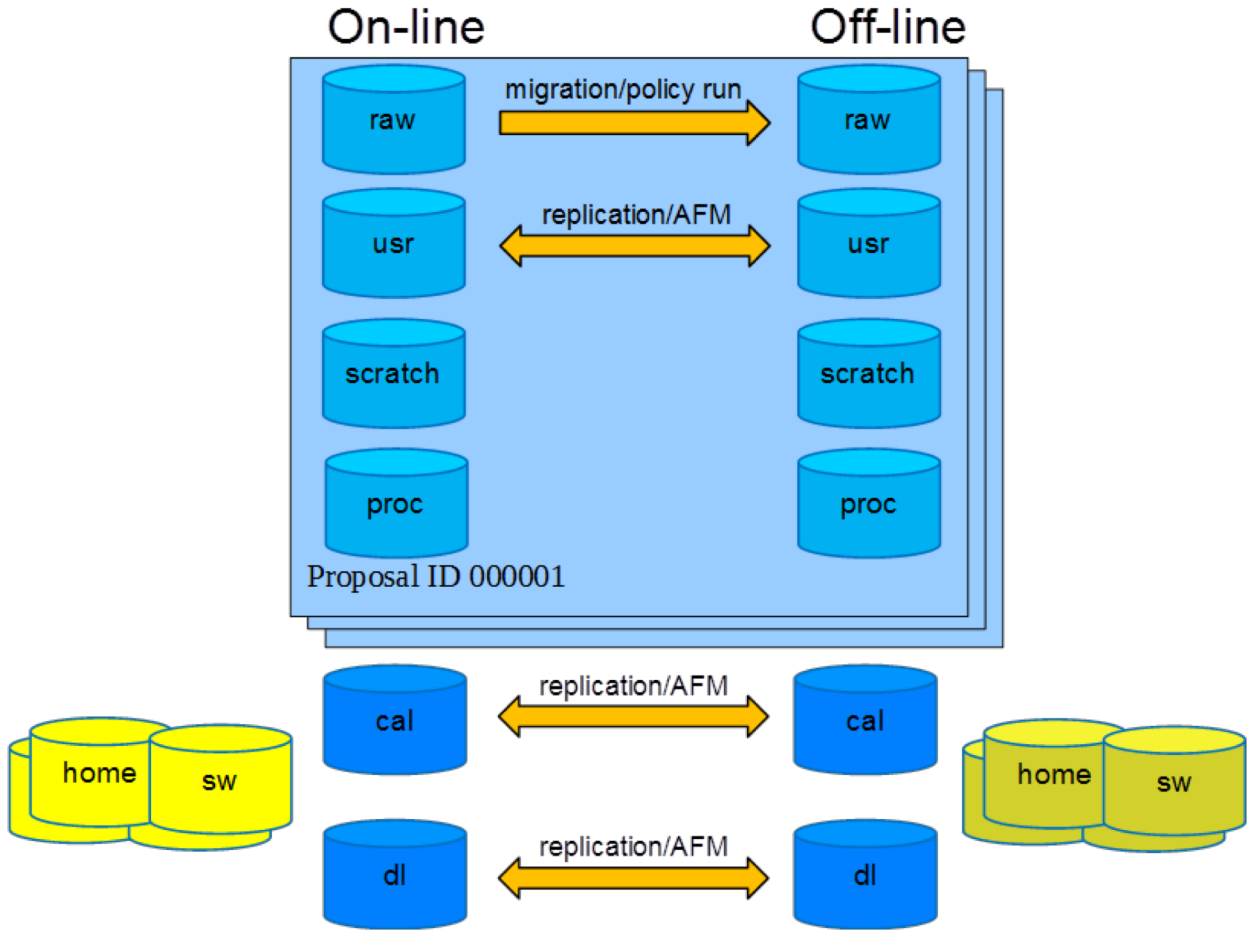
DAQ & Offline architecture

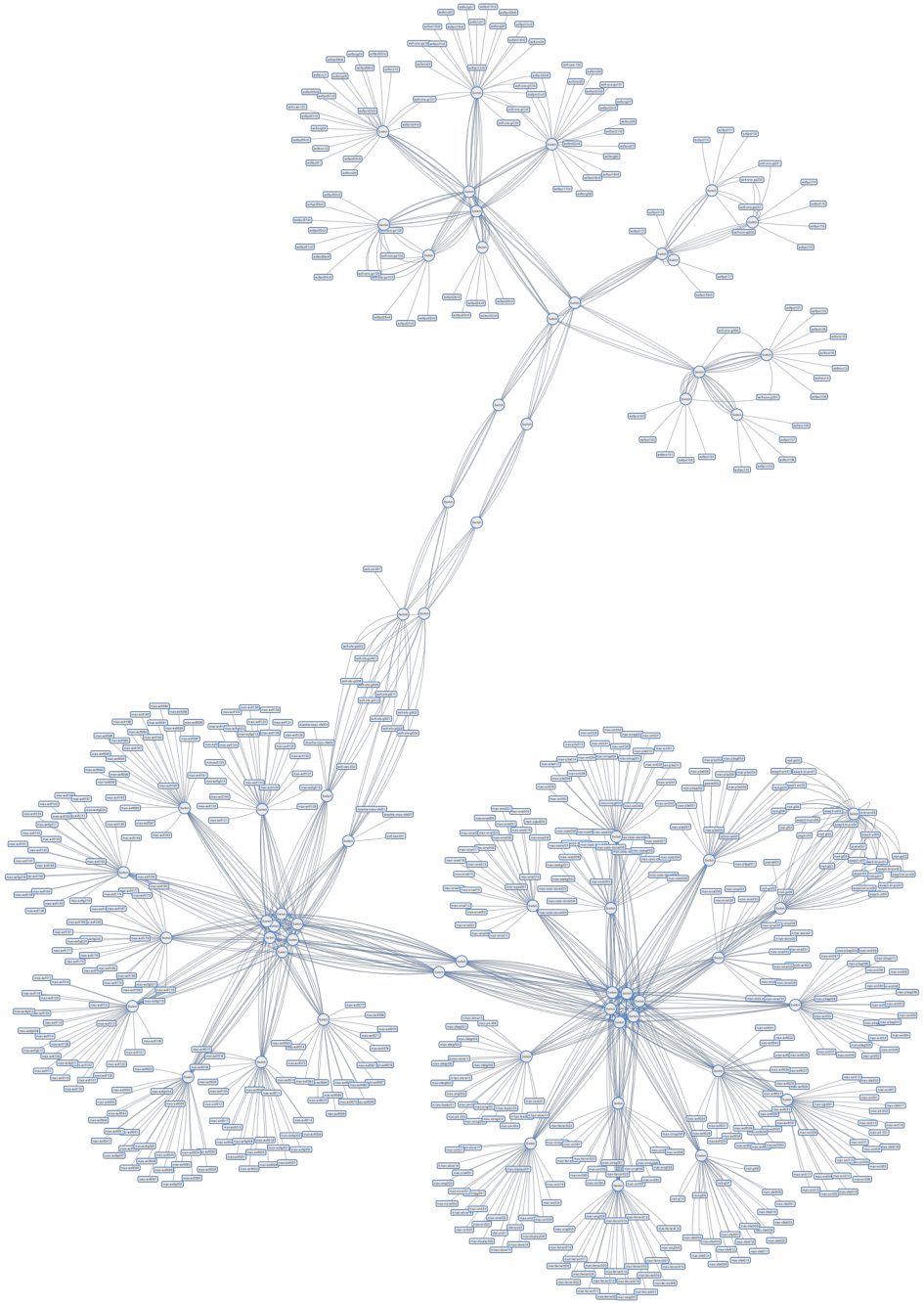


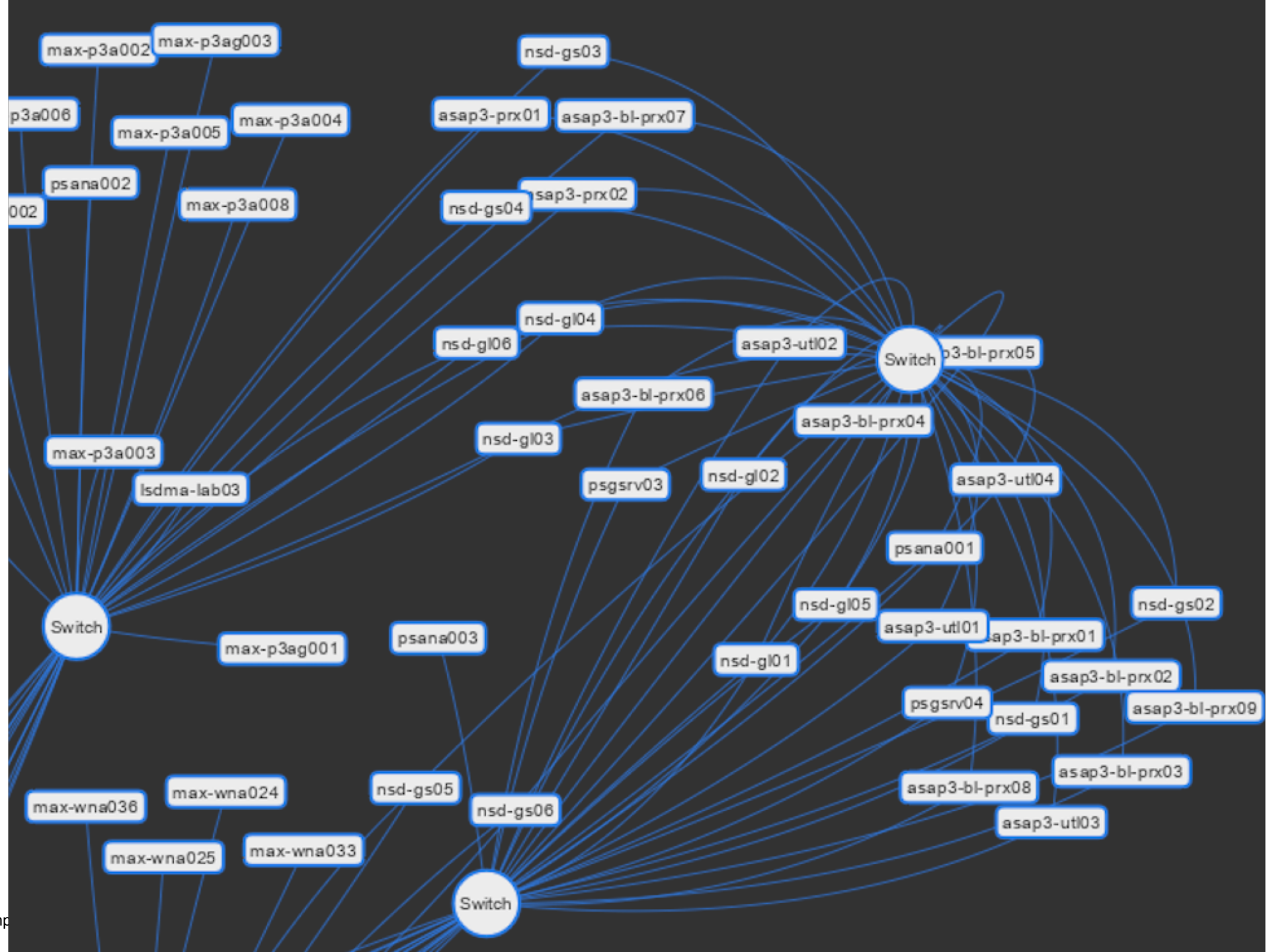
- PC Layer
  - Pixel reordering
  - Data reduction, FPGA based compression
  - Veto
  - File creation on the online storage (HDF5)
- Online cluster
  - CPU and GPU nodes
  - Online data analysis (fast feedback)
  - Calibration and data correction
- Long-Haul Infiniband Metro-X
  - Supports 6 long haul ports (FDR-10 40Gb/s)
  - Scalable by multiple switches
- Transfer online to offline storage
  - Custom scripts with policy runs
  - GPFS AFM
- Offline storage
  - Shared across experiment stations
  - Data migrated to offline storage after quality checks
  - Copy data to dCache, ACLs
  - Raw data access only from dCache
  - Calibrated data stored on GPFS
  - User analysis based on calibrated data

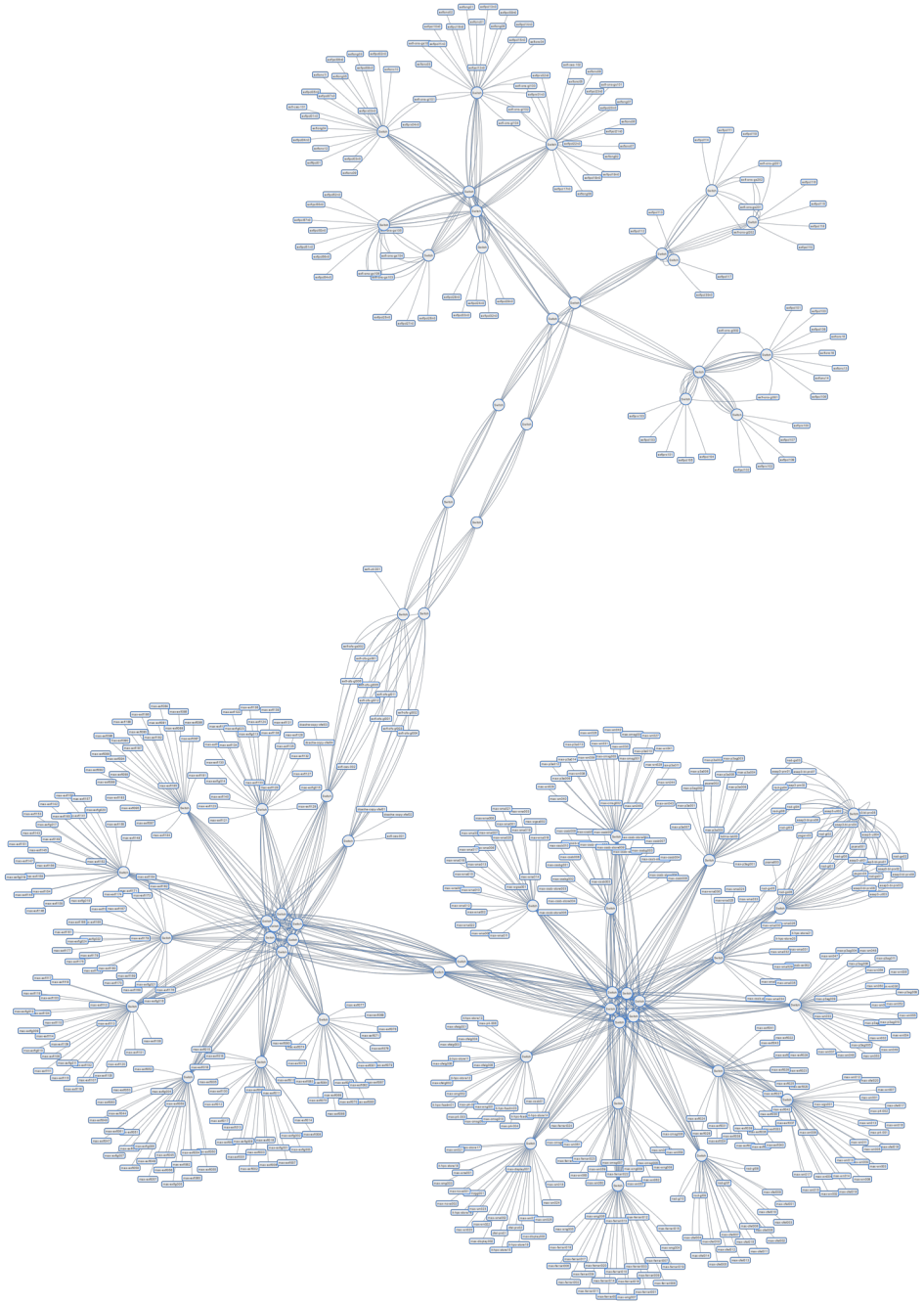
File systems and data placement

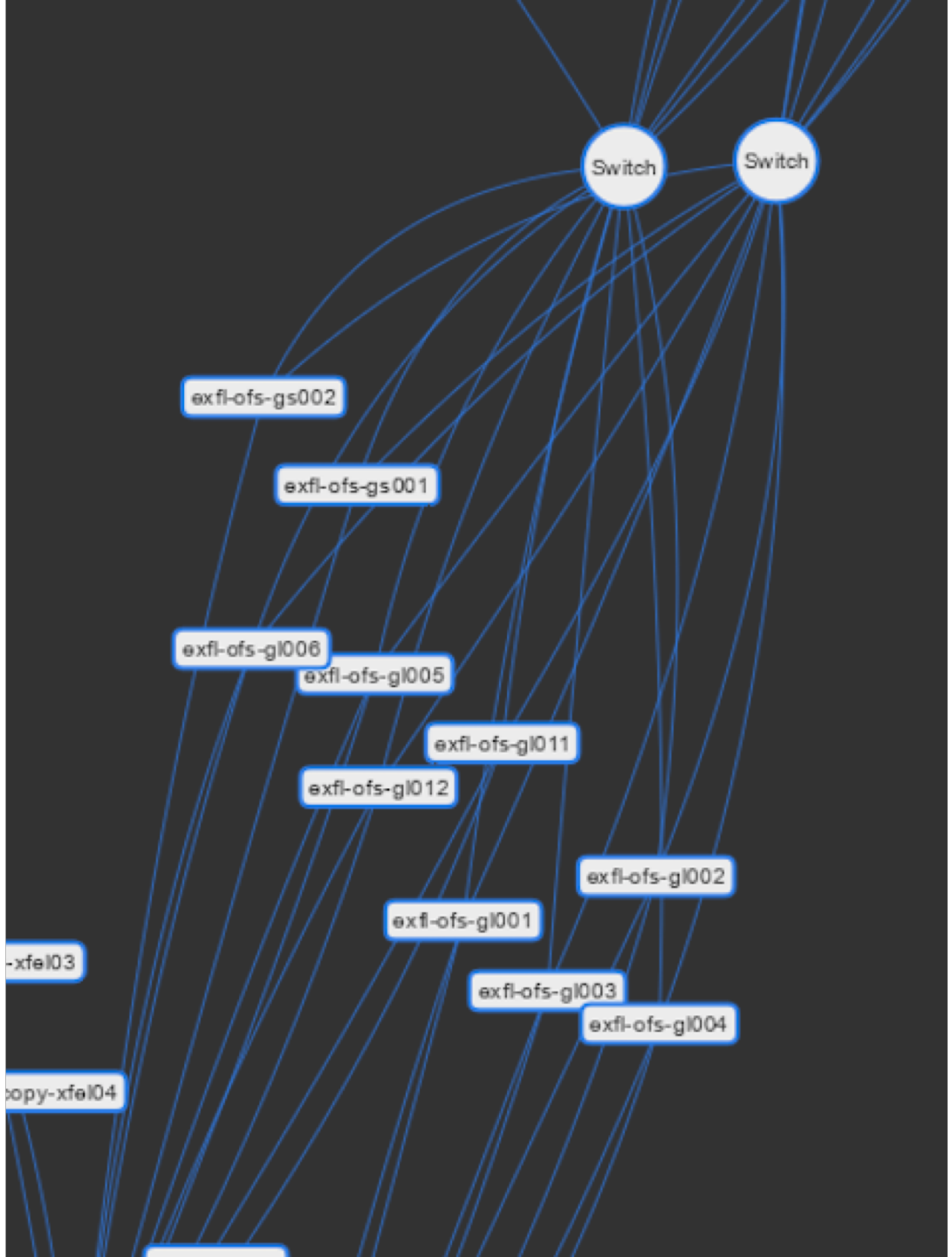







max-p3a002
max-p3ag003
nsd-gs03
p3a006
max-p3a005
max-p3a004
asap3-prx01
asap3-bl-prx07
psana002
002
max-p3a008
nsd-gs04
sap3-prx02
nsd-gl04
nsd-gl06
asap3-utl02
Switch
p3-bl-prx05
asap3-bl-prx06
asap3-bl-prx04
nsd-gl03
max-p3a003
lsdma-lab03
psgsrv03
nsd-gl02
asap3-utl04
psana001
nsd-gl05
nsd-gs02
Switch
max-p3ag001
psana003
asap3-utl01
ap3-bl-prx01
nsd-gl01
asap3-bl-prx02
psgsrv04
asap3-bl-prx09
nsd-gs01
asap3-bl-prx03
max-wna024
nsd-gs05
asap3-bl-prx08
max-wna036
nsd-gs06
asap3-utl03
max-wna025
max-wna033
Switch

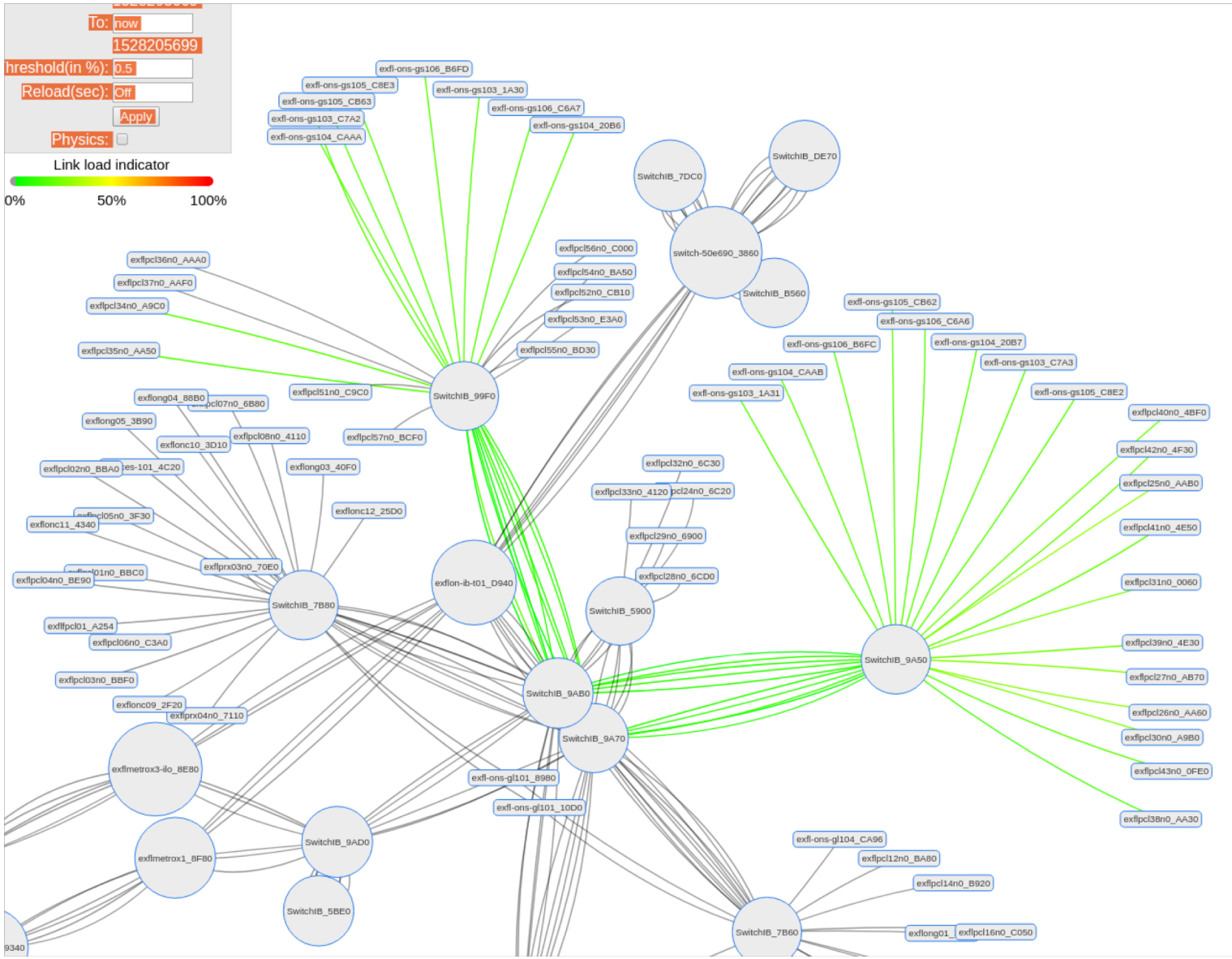


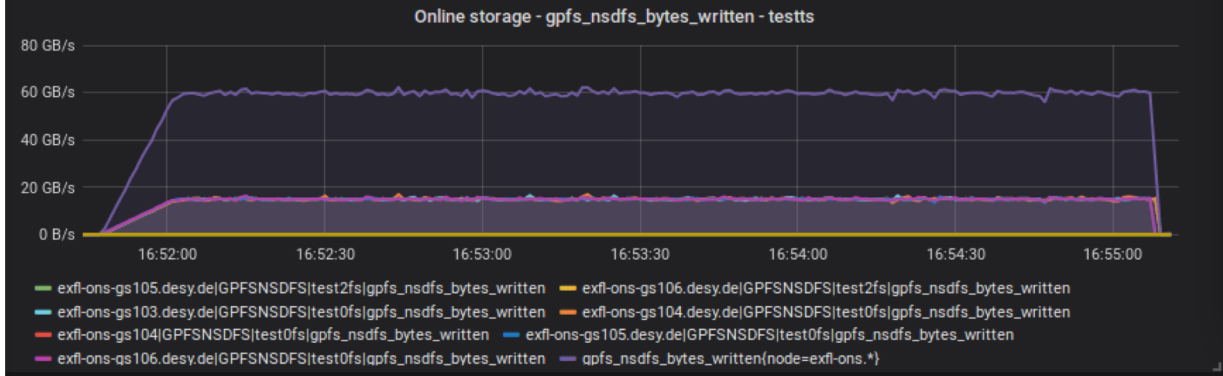

Switch
Switch
exfl-ofs-gs002
exfl-ofs-gs001
exfl-ofs-gl006
exfl-ofs-gl005
exfl-ofs-gl011
exfl-ofs-gl012
exfl-ofs-gl002
exfl-ofs-gl001
-xfel03
exfl-ofs-gl003
exfl-ofs-gl004
opy-xfel04

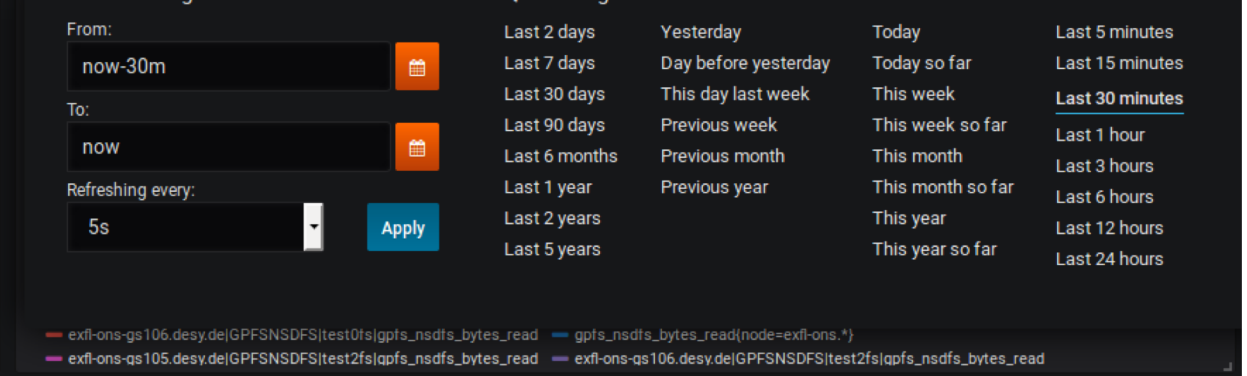
All Flash for DAQ writes (online storage)

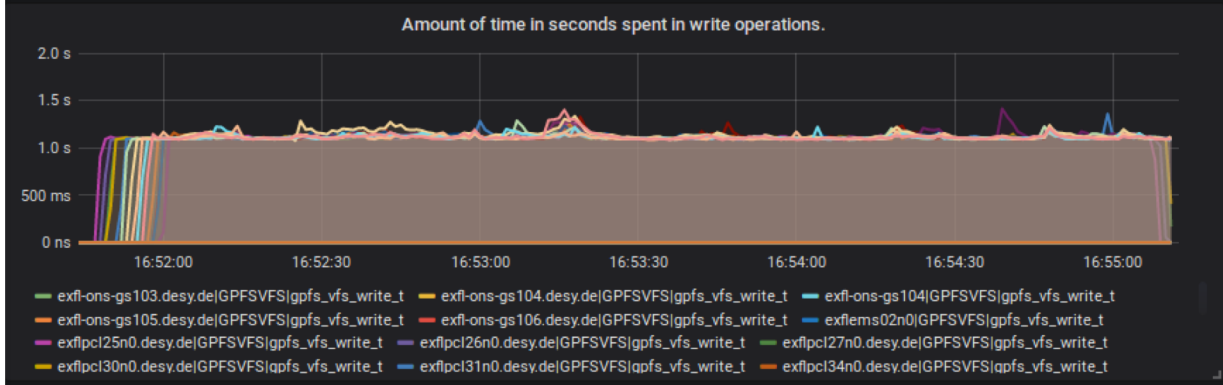
- GS4S
  - 4 x 24 x 4TB (SAS connected)
  - EDR (16 ports)
  - measured 30GBs per system
  - agnostic against ‘crazy IO’
  - limited by Linux SCSI stack and PCIgen3
- Petra 3 & FLASH (ASAP^3)
- EuXFEL (3x)

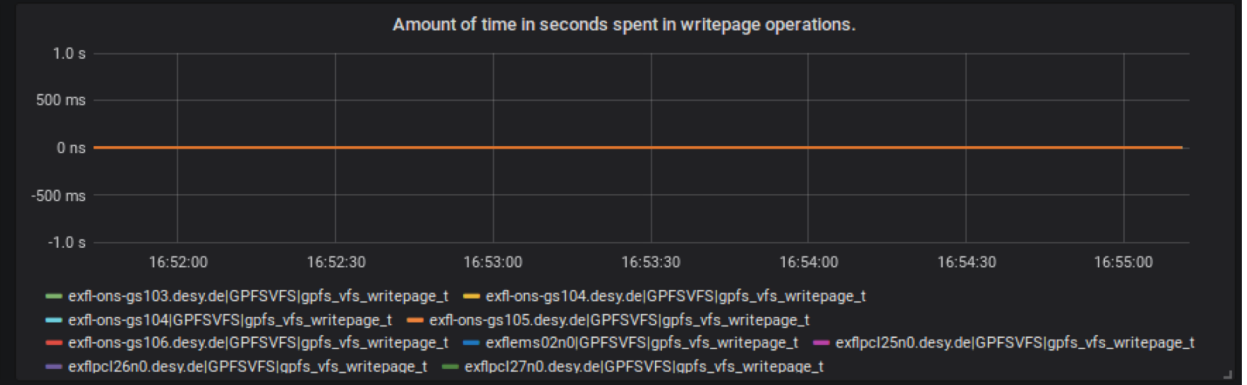
- tests – EuXFEL – 15 nodes each writing 200 x 4GB file every second
  - jitter < 20%

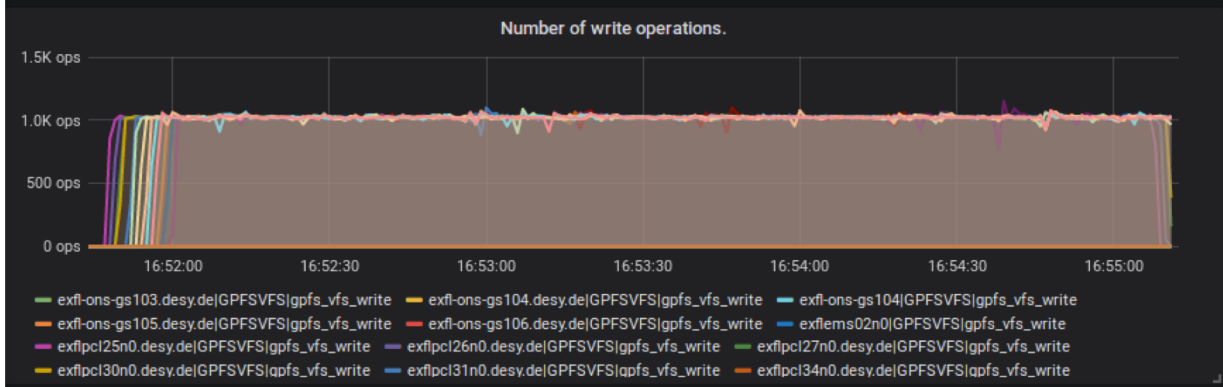


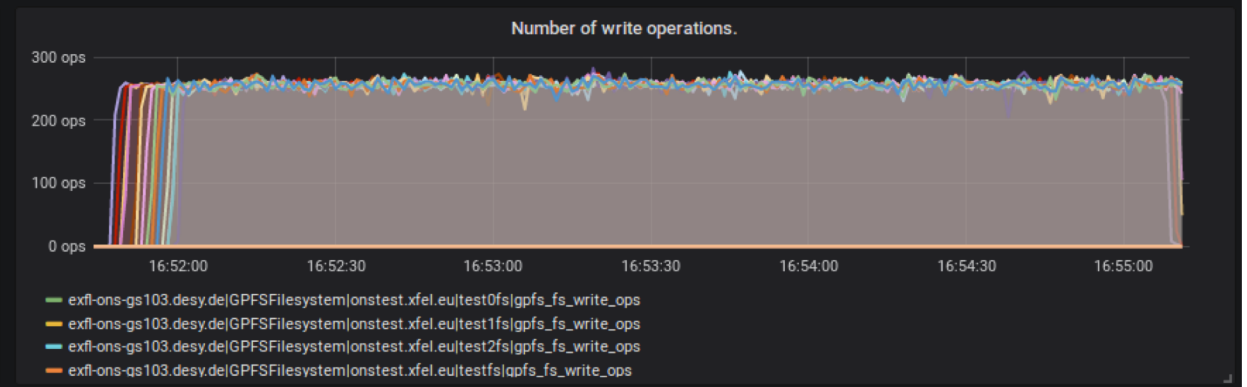

Online storage - gpfs_nsdfs_bytes_written - testts
80 GB/s
60 GB/s
40 GB/s
20 GB/s
0 B/s
16:52:00
16:52:30
16:53:00
16:53:30
16:54:00
16:54:30
16:55:00
exfl-ons-gs105.desy.de|GPFSNSDFS|test2fs|gpfs_nsdfs_bytes_written
exfl-ons-gs106.desy.de|GPFSNSDFS|test2fs|gpfs_nsdfs_bytes_written
exfl-ons-gs103.desy.de|GPFSNSDFS|test0fs|gpfs_nsdfs_bytes_written
exfl-ons-gs104.desy.de|GPFSNSDFS|test0fs|gpfs_nsdfs_bytes_written
exfl-ons-gs104|GPFSNSDFS|test0fs|gpfs_nsdfs_bytes_written
exfl-ons-gs105.desy.de|GPFSNSDFS|test0fs|gpfs_nsdfs_bytes_written
exfl-ons-gs106.desy.de|GPFSNSDFS|test0fs|gpfs_nsdfs_bytes_written
gpfs_nsdfs_bytes_written{node=exfl-ons.*}

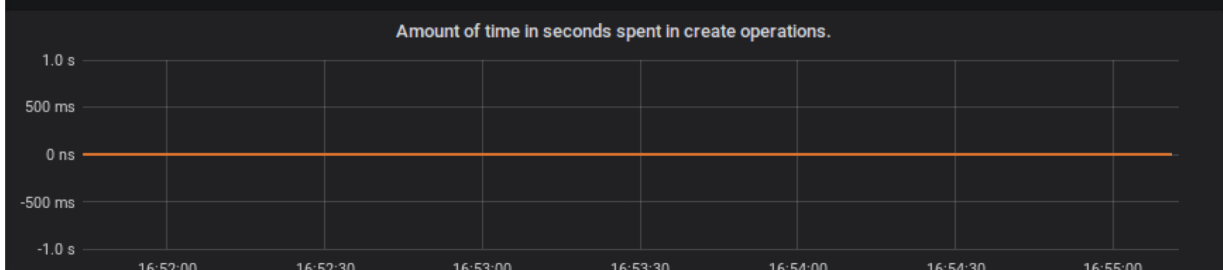

From:
now-30m
To:
now
Refreshing every:
5s
Apply
Last 2 days
Last 7 days
Last 30 days
Last 90 days
Last 6 months
Last 1 year
Last 2 years
Last 5 years
Yesterday
Day before yesterday
This day last week
Previous week
Previous month
Previous year
Today
Today so far
This week
This week so far
This month
This month so far
This year
This year so far
Last 5 minutes
Last 15 minutes
Last 30 minutes
Last 1 hour
Last 3 hours
Last 6 hours
Last 12 hours
Last 24 hours
exfl-ons-gs106.desy.de|GPFSNSDFS|test0fs|gpfs_nsdfs_bytes_read
gpfs_nsdfs_bytes_read{node=exfl-ons.*}
exfl-ons-gs105.desy.de|GPFSNSDFS|test2fs|gpfs_nsdfs_bytes_read
exfl-ons-gs106.desy.de|GPFSNSDFS|test2fs|gpfs_nsdfs_bytes_read

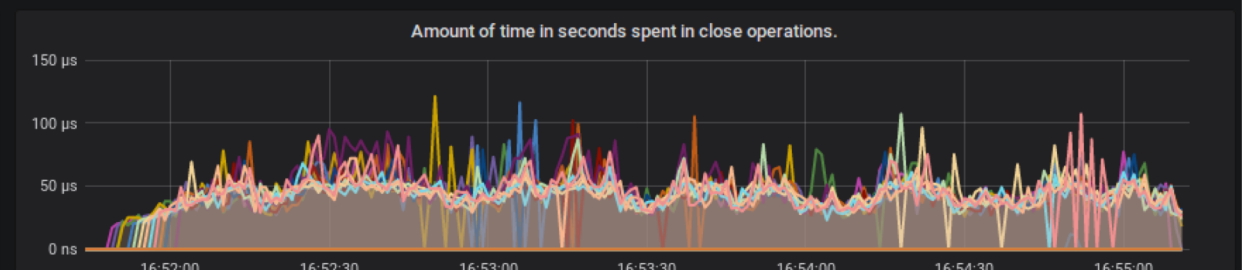

Amount of time in seconds spent in write operations.
2.0 s
1.5 s
1.0 s
500 ms
0 ns
16:52:00
16:52:30
16:53:00
16:53:30
16:54:00
16:54:30
16:55:00
exfl-ons-gs103.desy.de|GPFSVFS|gpfs_vfs_write_t
exfl-ons-gs104.desy.de|GPFSVFS|gpfs_vfs_write_t
exfl-ons-gs104|GPFSVFS|gpfs_vfs_write_t
exfl-ons-gs105.desy.de|GPFSVFS|gpfs_vfs_write_t
exfl-ons-gs106.desy.de|GPFSVFS|gpfs_vfs_write_t
exflems02n0|GPFSVFS|gpfs_vfs_write_t
exflpcl25n0.desy.de|GPFSVFS|gpfs_vfs_write_t
exflpcl26n0.desy.de|GPFSVFS|gpfs_vfs_write_t
exflpcl27n0.desy.de|GPFSVFS|gpfs_vfs_write_t
exflpcl30n0.desy.de|GPFSVFS|gpfs_vfs_write_t
exflpcl31n0.desy.de|GPFSVFS|gpfs_vfs_write_t
exflpcl34n0.desy.de|GPFSVFS|gpfs_vfs_write_t


Amount of time in seconds spent in writepage operations.
1.0 s
500 ms
0 ns
-500 ms
-1.0 s
16:52:00
16:52:30
16:53:00
16:53:30
16:54:00
16:54:30
16:55:00
exfl-ons-gs103.desy.de|GPFSVFS|gpfs_vfs_writepage_t
exfl-ons-gs104.desy.de|GPFSVFS|gpfs_vfs_writepage_t
exfl-ons-gs104|GPFSVFS|gpfs_vfs_writepage_t
exfl-ons-gs105.desy.de|GPFSVFS|gpfs_vfs_writepage_t
exfl-ons-gs106.desy.de|GPFSVFS|gpfs_vfs_writepage_t
exflems02n0|GPFSVFS|gpfs_vfs_writepage_t
exflpcl25n0.desy.de|GPFSVFS|gpfs_vfs_writepage_t
exflpcl26n0.desy.de|GPFSVFS|gpfs_vfs_writepage_t
exflpcl27n0.desy.de|GPFSVFS|gpfs_vfs_writepage_t


Number of write operations.
1.5K ops
1.0K ops
500 ops
0 ops
16:52:00
16:52:30
16:53:00
16:53:30
16:54:00
16:54:30
16:55:00
exfl-ons-gs103.desy.de|GPFSVFS|gpfs_vfs_write
exfl-ons-gs104.desy.de|GPFSVFS|gpfs_vfs_write
exfl-ons-gs104|GPFSVFS|gpfs_vfs_write
exfl-ons-gs105.desy.de|GPFSVFS|gpfs_vfs_write
exfl-ons-gs106.desy.de|GPFSVFS|gpfs_vfs_write
exflems02n0|GPFSVFS|gpfs_vfs_write
exflpcl25n0.desy.de|GPFSVFS|gpfs_vfs_write
exflpcl26n0.desy.de|GPFSVFS|gpfs_vfs_write
exflpcl27n0.desy.de|GPFSVFS|gpfs_vfs_write
exflpcl30n0.desy.de|GPFSVFS|gpfs_vfs_write
exflpcl31n0.desy.de|GPFSVFS|gpfs_vfs_write
exflpcl34n0.desy.de|GPFSVFS|gpfs_vfs_write


Number of write operations.
300 ops
200 ops
100 ops
0 ops
16:52:00
16:52:30
16:53:00
16:53:30
16:54:00
16:54:30
16:55:00
exfl-ons-gs103.desy.de|GPFSFilesystem|onstest.xfel.eu|test0fs|gpfs_fs_write_ops
exfl-ons-gs103.desy.de|GPFSFilesystem|onstest.xfel.eu|test1fs|gpfs_fs_write_ops
exfl-ons-gs103.desy.de|GPFSFilesystem|onstest.xfel.eu|test2fs|gpfs_fs_write_ops
exfl-ons-gs103.desy.de|GPFSFilesystem|onstest.xfel.eu|testfs|gpfs_fs_write_ops


Amount of time in seconds spent in create operations.
1.0 s
500 ms
0 ns
-500 ms
-1.0 s


Amount of time in seconds spent in close operations.
150 µs
100 µs
50 µs
0 ns

on the list

- GPFS events (cluster wide inotify)
  - events managed through Kafka
  - first tests completed

- GNR on network (MeStore)
  - promising lab results
  - smoother scaling at very high utilization
  - potential for – faster/cheaper burst buffer, capacity (disk) configurations

- logbook (digital)
  - everybody has a ‘not well beloved’ system, looking for better alternative
  - everybody acknowledge importance

- easier and scalable ‘online data analysis’ services and configurations

observations – end of 2017

primarily from IT perspective

- doubling detector rates faster than Moore's law (8-20 months)
- more than 1GBs from beamline to storage (faster + more detectors)
  - NFS & SMB ruled out
- online data analysis (aka. fast-feedback-loop)

HiDRA etc.

- primary problem at the origin – getting data out of detector server memory fast enough (continuous mode)
- more Petra III Extensions in operations & FLASH in the game
- first experiments with ‘Online Analysis’ – hand crafted setup (Maxwell nodes access beamline FS)
  - could not simply be established as regular service – IO resource overrun, authentication/authorization, …
  - docker + volume driver + …
- load increasing / beamline & core FS
- connection/cooperation to other labs - ‘capable of development’ – too weak
  - standardization, common practices and developments, …
  - CHEP and HEPiX like conference/workshop missing
- synergies with XFEL systems (long-range InfiniBand, all-flash systems, …)

summary

- stable and performant operation of GPFS
- profit from CORAL developments
  - less locks
  - buffer & rpc code
  - NUMA awareness (network & gpfs & block device threads)
- massive scale of new X GBs detectors – everybody wants that
  - several 10 of these – a little problem ;-)
  - data reduction (or ‘find the good ones’) becomes most important
  - Petra 4 on the horizon – heavy planning activities – next big jump in data rate (another x1k)
- PS is different than HEP – in many ways - computing too ;-)