GPFS for Data Taking and Analysis at Petra III and European XFEL.

Martin Gasthuber
Co-Author: Stefan Dietrich, Manuela Kuhn, Uwe Ensslin, Janusz Malka
DAQ@LHC, 12.04.16

HELMHOLTZ ASSOCIATION

LSDMA

DESY

About DESY

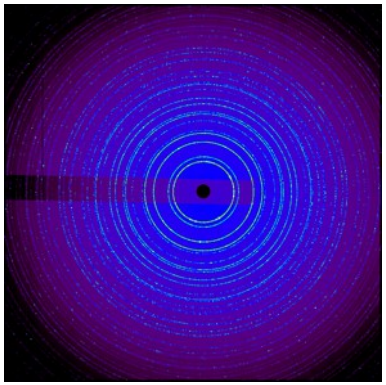
- founded December 1959
- Accelerator Center
  - Research, build and operation
- Research topics
  - Particle Physics (HEP)
    - DORIS, PETRA, HERA, now LHC
  - Photon Science
    - Petra III, FLASH soon EuXFEL
  - Astro Particle Physics
    - IceCube, CTA, ...
- 2 Sites
  - Hamburg
  - Zeuthen (Brandenburg), near Berlin
- ~2300 employees, 3000 guest scientists annually

Hamburg

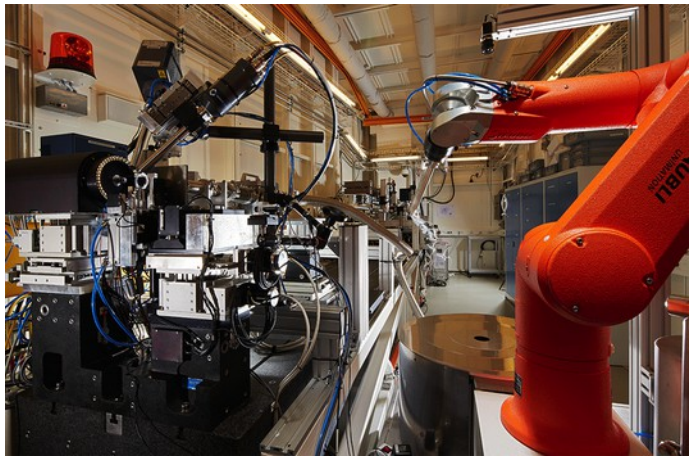
Zeuthen

PETRA III

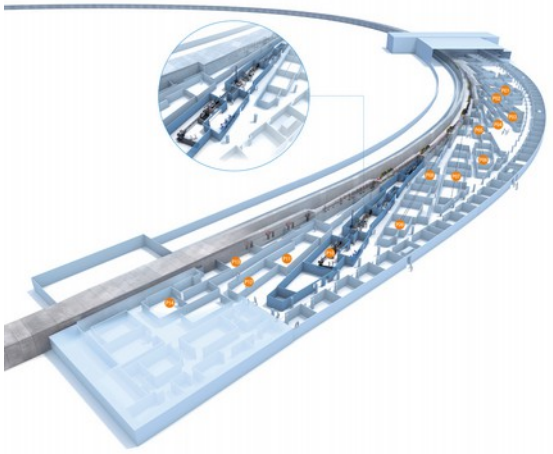
- Ring accelerator
  - 2.3 km circumference
- X-ray radiation
- Since 2009: 14 beamlines in operation
- 2016: 10 additional beamlines starting operation (extension)



Sample raw file



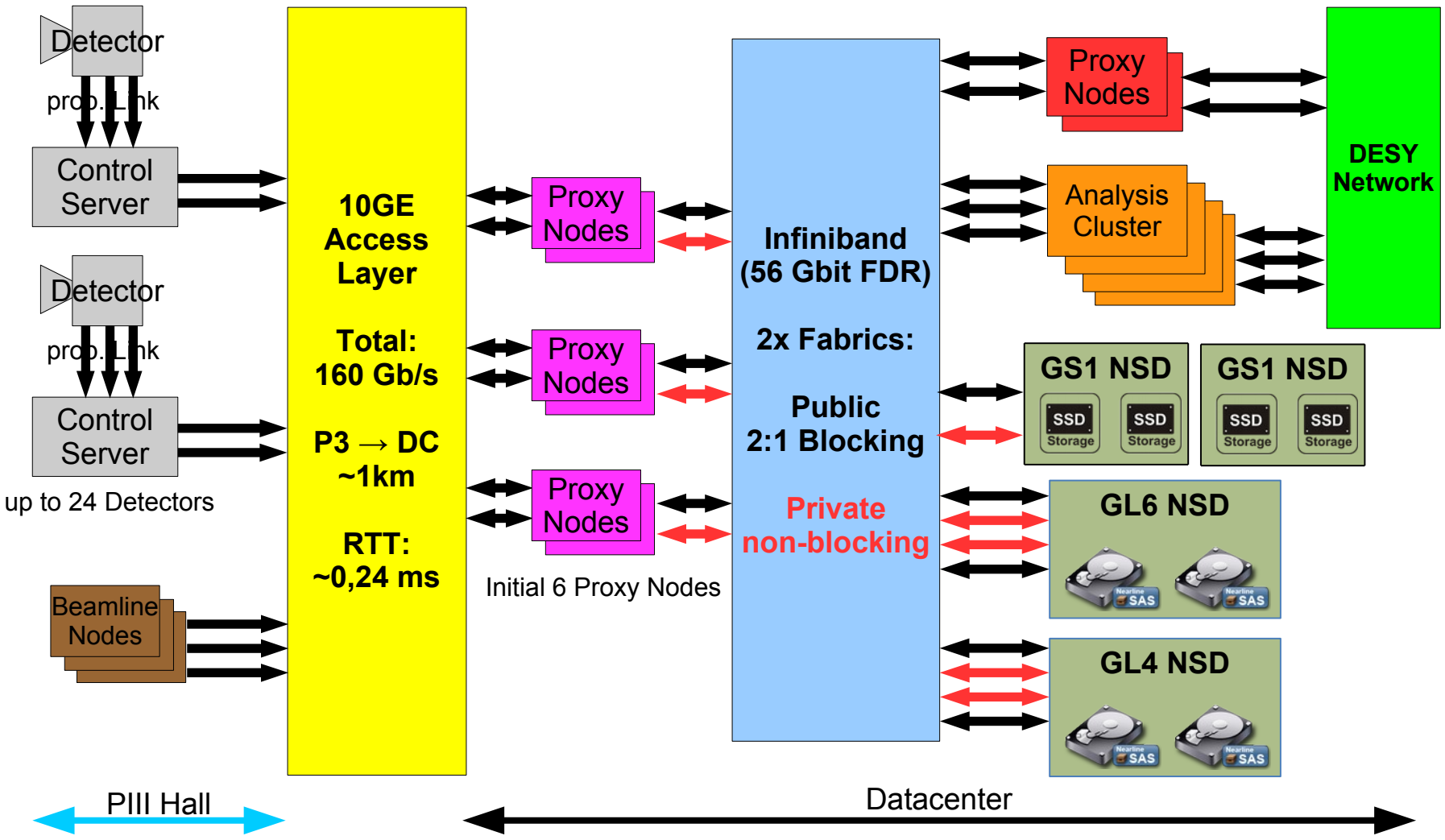
Beamline P11
Bio-Imaging and diffraction



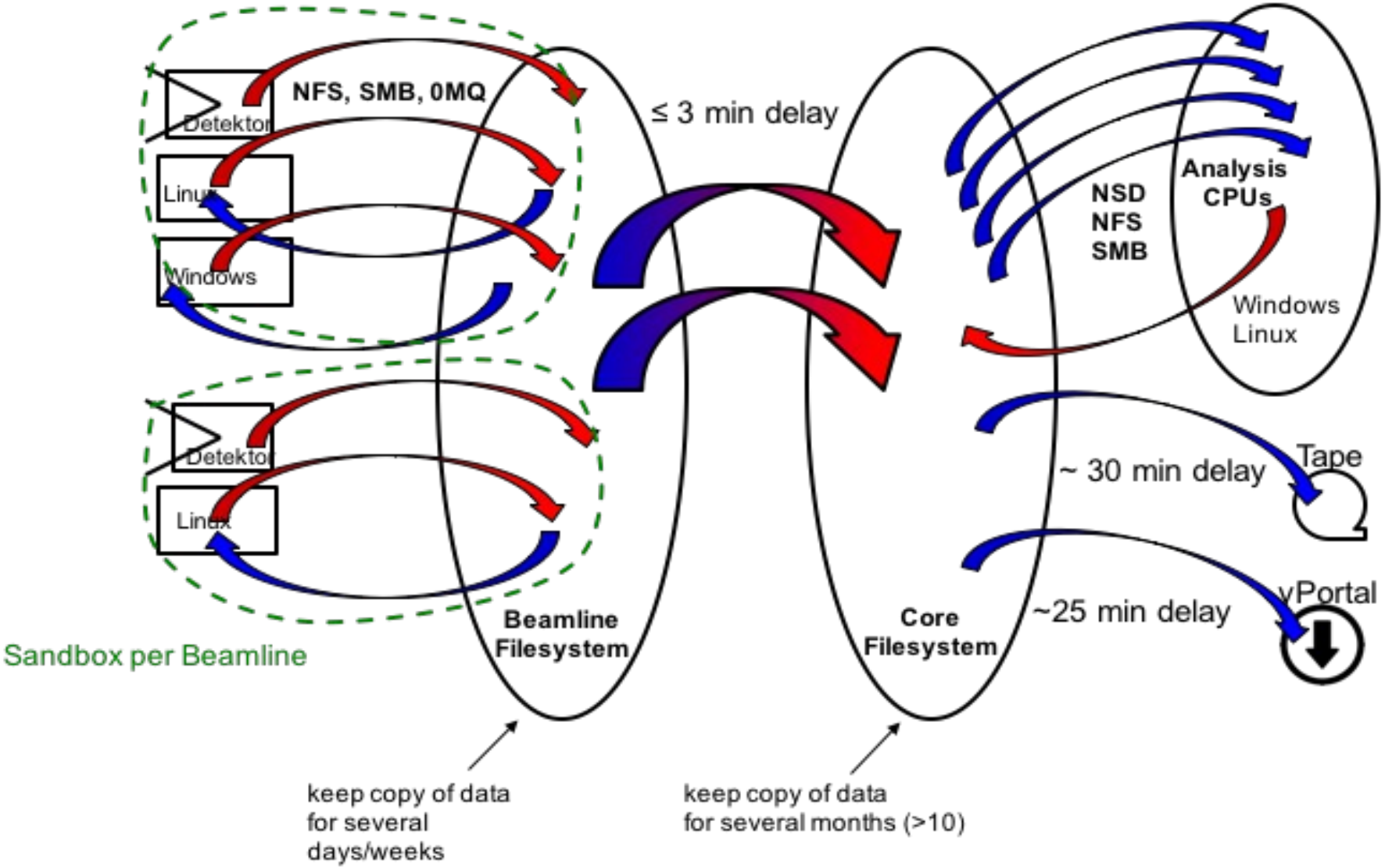
DESY & IBM Collaboration

- Collaboration with IBM within scope of SPEED project
  - June 2014 → March 2015 (core SPEED)
  - PETRA III Restart: March 2015 → October 2015
  - Next period: April 2016 → December 2016 – just continued working ;-)
- Solution based on IBM Spectrum Scale and Elastic Storage Server
  - GPFS 4.1.0-8
  - ESS 2.5.x
  - Currently upgrading to GPFS 4.2.0-1 and ESS 4.0.0 – finished last week
- Multiple ESS building blocks running
  - 2x ESS GS1 (24x 400GB SSD)
  - 1x ESS GL4 (232x 4TB NLSAS)
  - 1x ESS GL6 (348x 4TB NLSAS)

ASAP³ Architecture



from the cradle to the grave



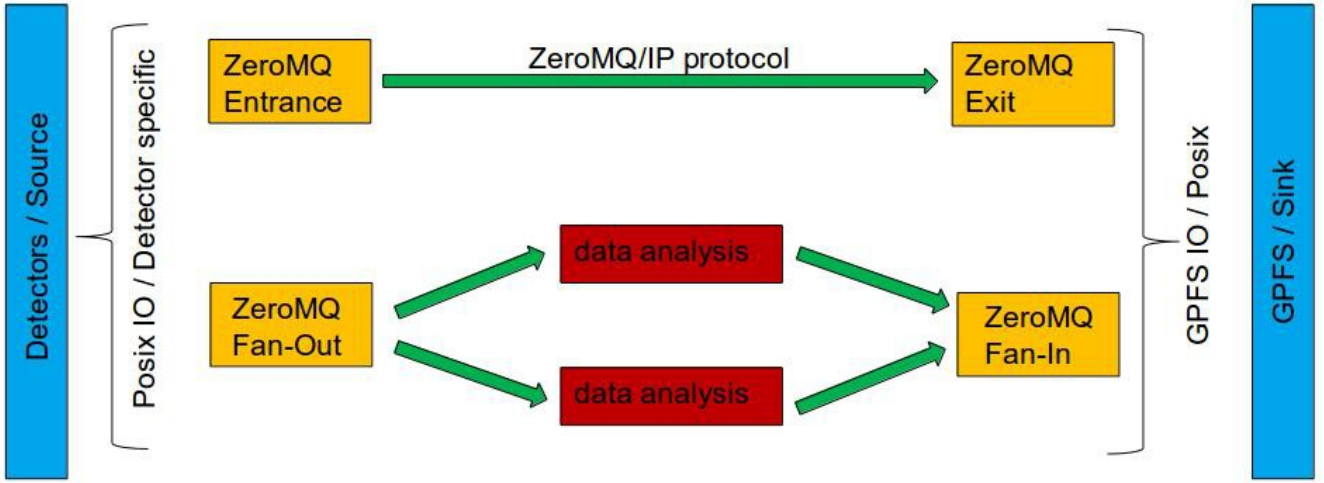
Access Protocols

- Proxy nodes export GPFS for multiple protocols
- Beamline
  - NFSv3 (Kernel) with cNFS
  - SMB, based on Samba 4.2
  - ZeroMQ based
- Core
  - NFSv3 (Kernel) and SMB (Samba 4.2)
    - Custom script required for ID mapping between Linux and Windows
  - After ESS 4.0.0 upgrade: Migration to Cluster Export Services (CES)
    - NFS based on Ganesha (userland NFS server) – support v4 (implemented also 4.1)
  - Native GPFS access on Analysis Cluster (offline)

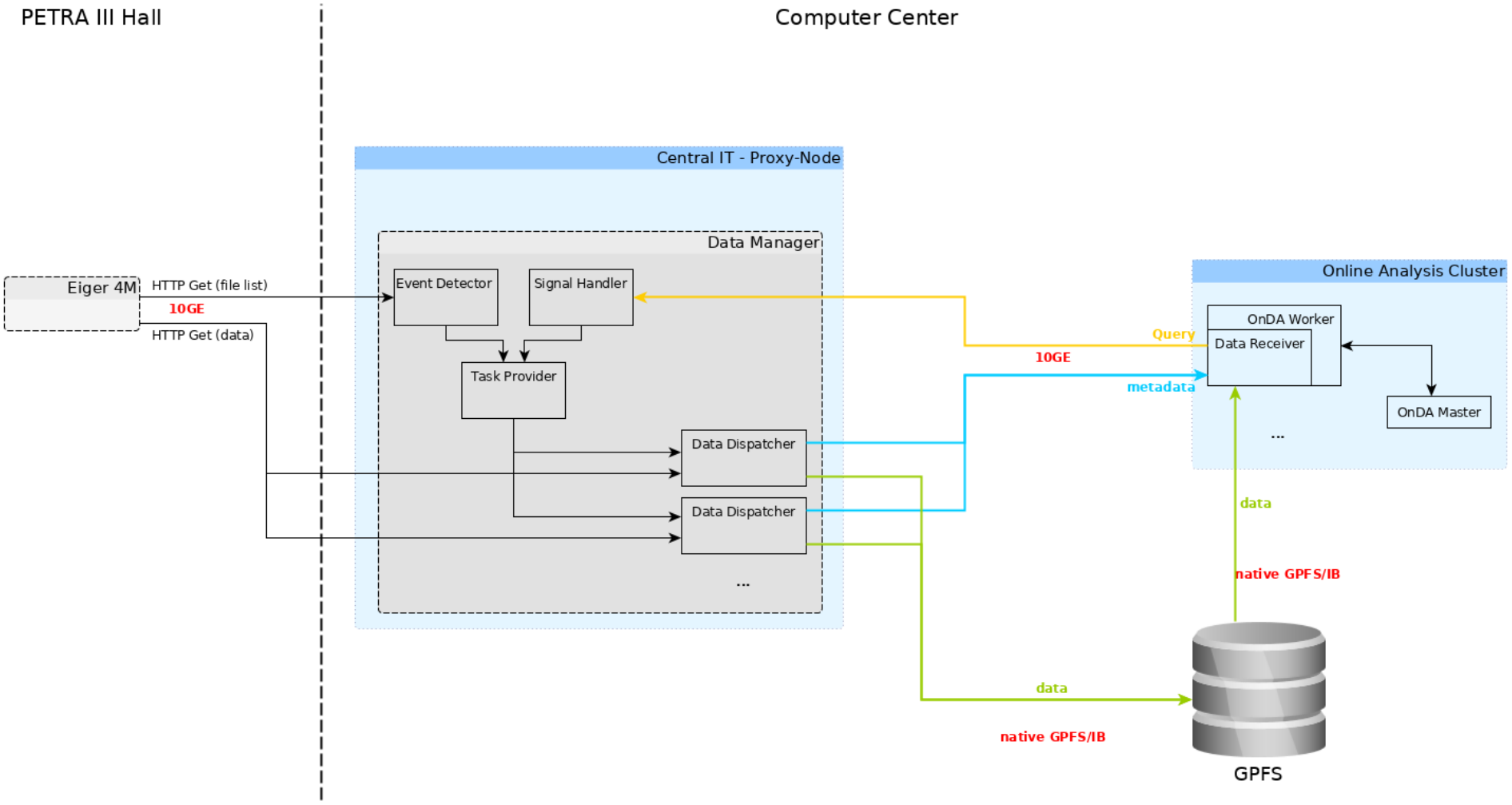
DESY

ZeroMQ and Data Ingest

- currently building environment to support next generation detectors
- target: aggregate ~30GB/sec
- first mile prototypes currently under development and testing
  - “vacuum cleaner”: pick up data from detector and send it to GPFS
  - Live Viewer: send images to a receiver, for display/monitoring at beamline

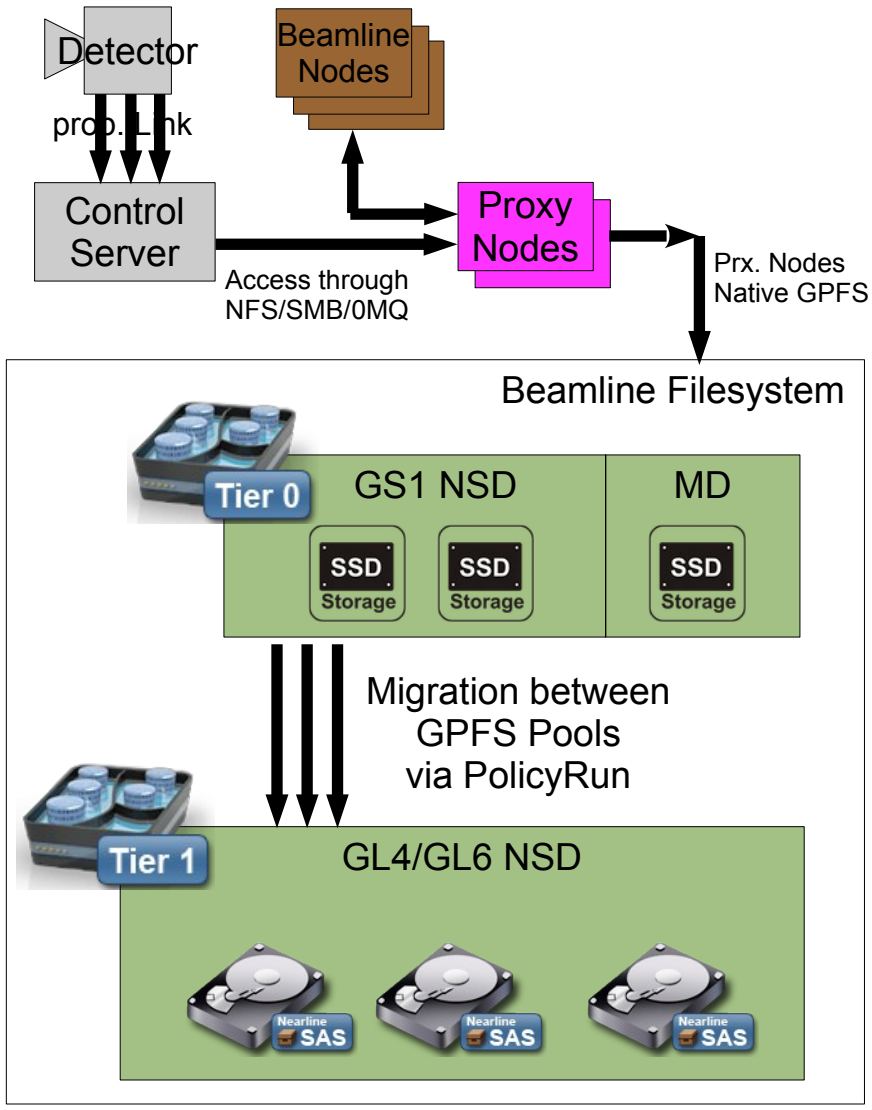


ZMQ – more complex first mile – include online analysis



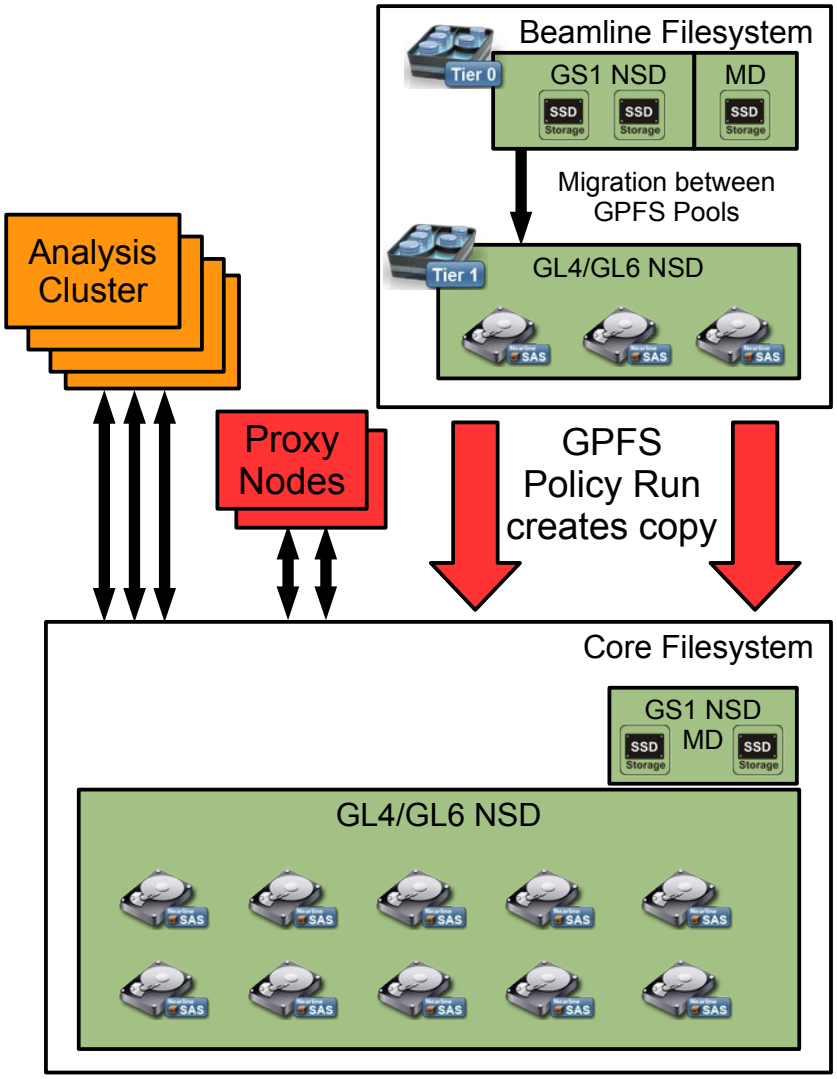
Beamline Filesystem

- "Wild-West" area for beamline
- Only host based authentication, no ACLs
- Access through NFSv3, SMB or ZeroMQ
- Optimized for performance
  - 1 MiB filesystem blocksize
  - Pre-optimized NFSv3: ~60 MB/s
  - NFSv3: ~600 MB/s
  - SMB: ~300-600 MB/s
- Tiered Storage
  - Tier 0: SSD burst buffer (< 10 TB)
  - Migration after short period of time
  - Tier 1: ~90 TB capacity

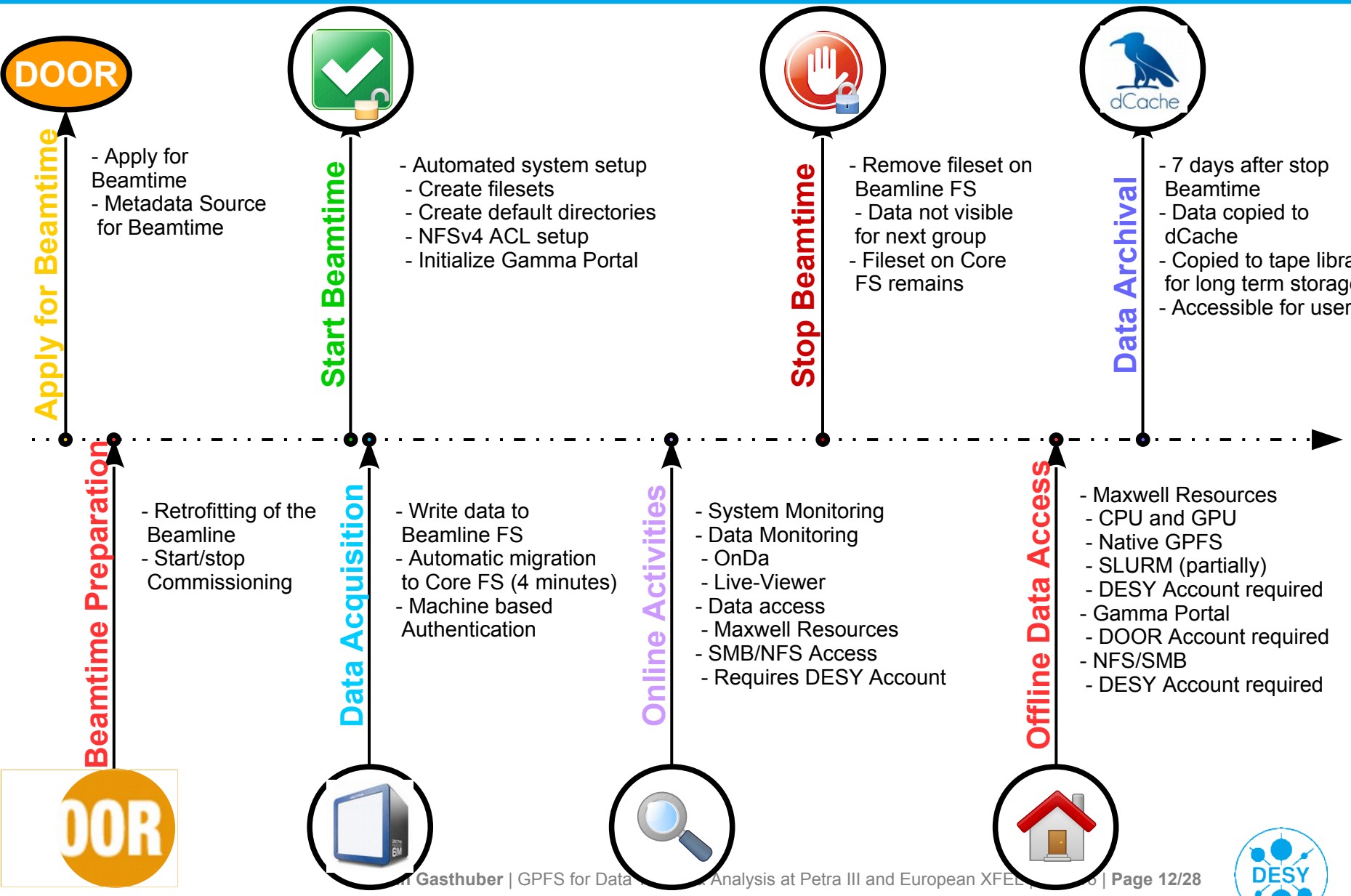


Core Filesystem

- "Clean world"
- Full user authentication
- NFSv4 ACLs
- Access through NFSv3, SMB or native GPFS
- GPFS Policy Runs copy data
  - Beamline → Core Filesystem
  - Single UID/GID
  - ACL inheritance gets active
  - TBD: raw data set to immutable
- 8 MiB filesystem blocksize
- 2 snapshots per day
- Fileset per beamtime



The User's View of the System

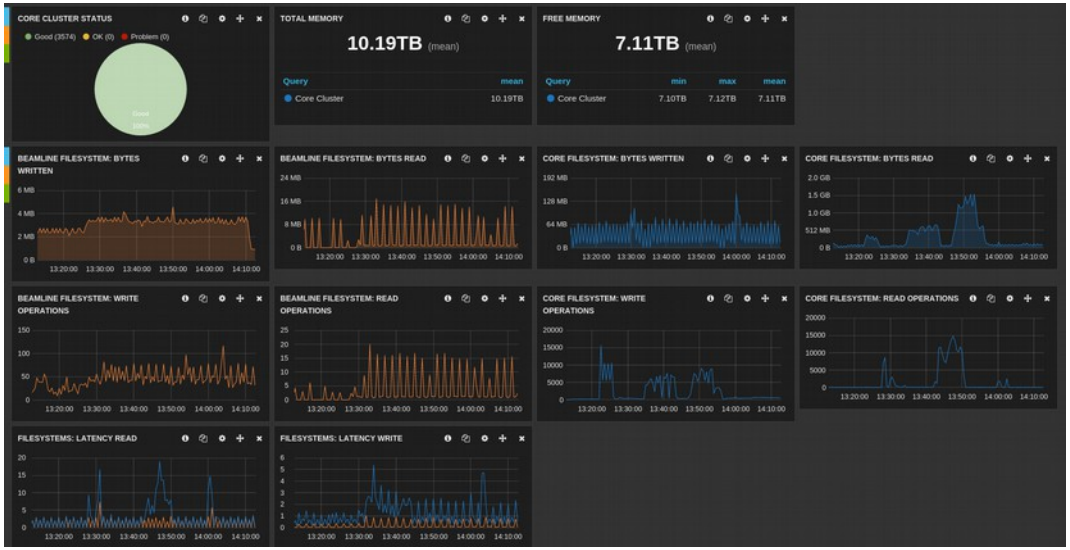


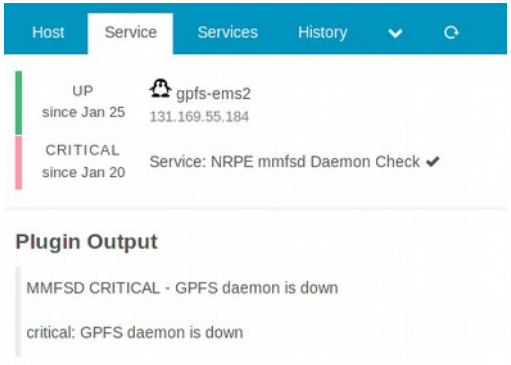
Performance Tuning

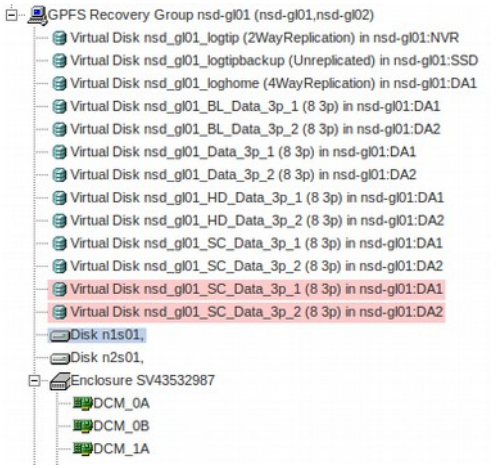
- impressive speed-ups after several tuning sessions (days)
- getting the right expert on hand (days) was crucial
- biggest performance impacts for our workload
  - GPFS Filesystem Blocksize: 1M for BL, 8M for Core
  - nfsPrefetchStrategy: Improve prefetching for NFS access
  - logBufferSize: “Journal size”
  - pagepool: Increase available memory for caching for GPFS
  - NFS clients: increase rsize/wsize
- still an important topic
  - Developing tools to measure performance over time – nearly done
  - Required to spot performance regressions - done
- finding the right expert at IBM is possible…
  - ...but not always easy ;-)

Monitoring

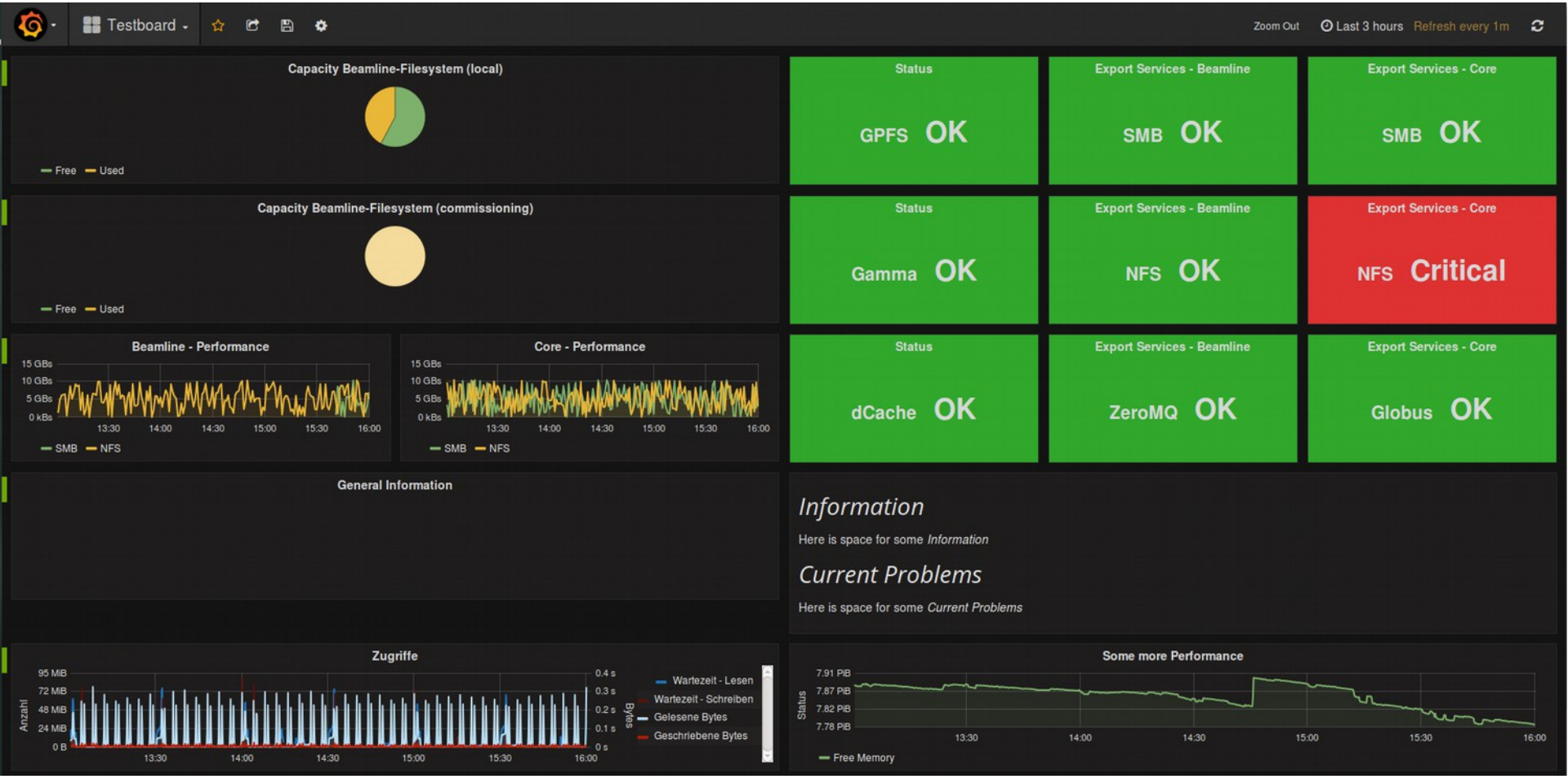
- Data collected from ZIMon dumped to Elasticsearch
- Allows correlation of performance metrics with logs
- Dashboards
  - View only: for beamline, staff or public areas
  - Expert view: Overview for the administrator
- Service and Hardware Monitoring with Icinga and RZ-Monitor (home grown)





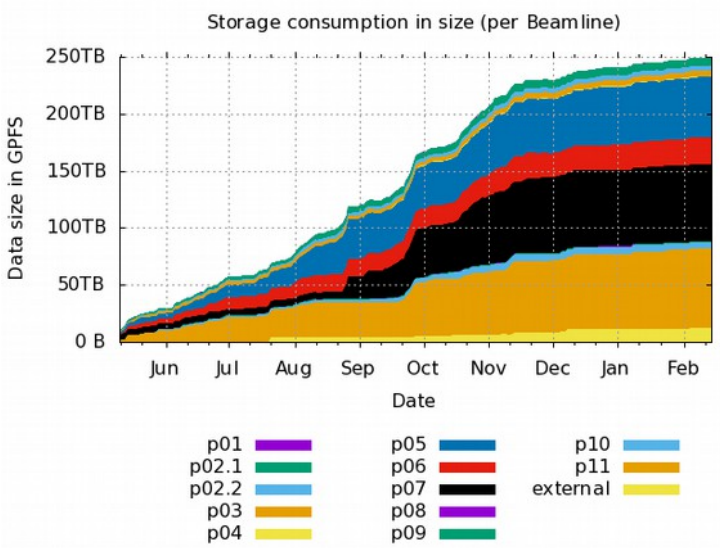


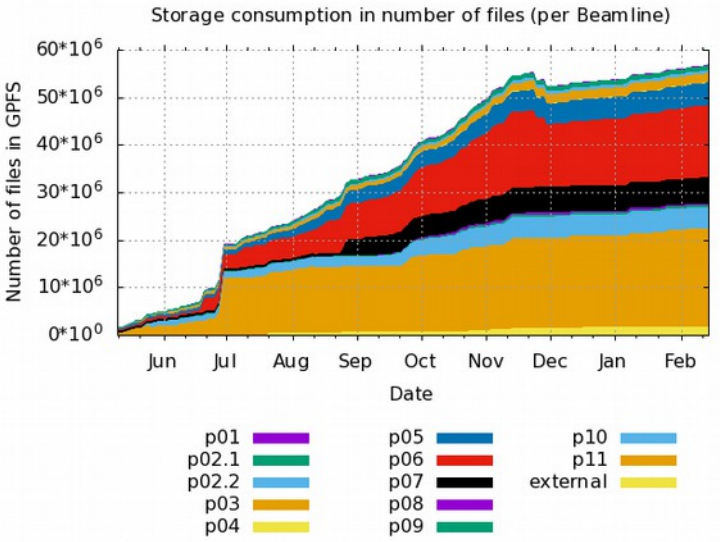
Dashboard – initial version



Experiences from the first period

- Overall user experience: good!
  - BL scientist more time for experiment, sample preparation and user support
  - no beamtine loss due to lacking space
  - reconstruction a lot faster and more stable
  - over reliance: “runs with blind trust”
- Overall GPFS and ESS stability: good!
  - good stability and performance
  - 3 near-critical issues – all IB/OFED related
  - deadlock/waiters – consequences of tight coupling – newer versions improved, more to come
- First detectors of new generation being installed during current shutdown
  - 3x Lambda Modules
  - 1x Eiger 4M





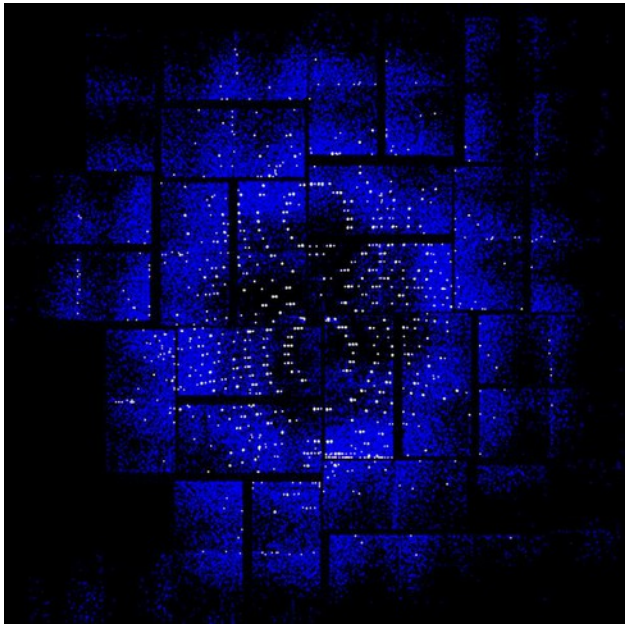
things we hope(d) would work, but…

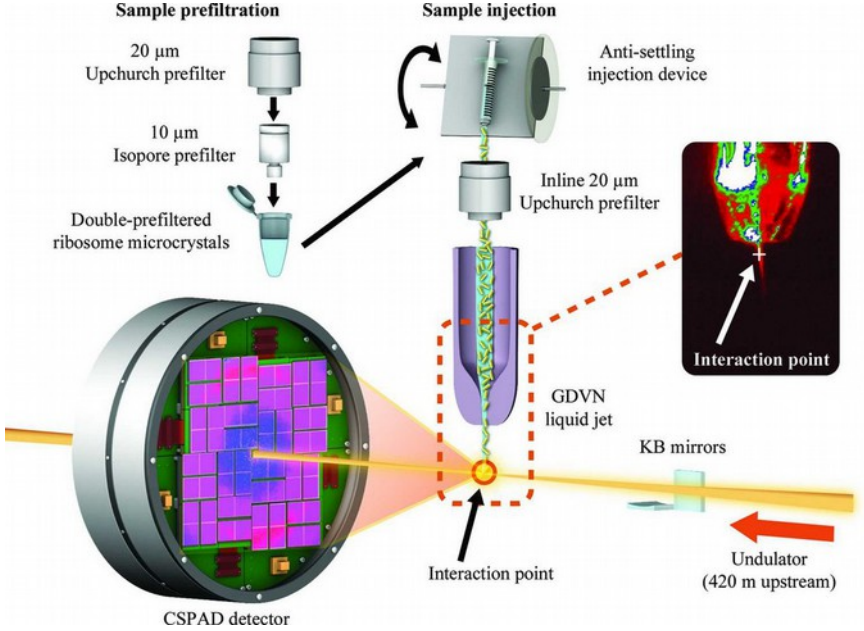
- current architecture result of process during last months
- detectors as native GPFS clients
  - old operating systems (RHEL 4, Suse 10 etc.)
  - inhomogeneous network for GPFS: InfiniBand and 10G Ethernet
- Windows as native GPFS client
  - more or less working, but source of pain
- Active file management (AFM) as copy process
  - no control during file transfer
  - not supported with native GPFS Windows client
  - cache behavior not optimal for this use case
- self-made SSD burst buffer
  - SSDs died very fast (majority), others by far too expensive (NVMe devices) but survived !
- remote cluster UID-mapping

DESY

Future Detectors and Experimental Setup

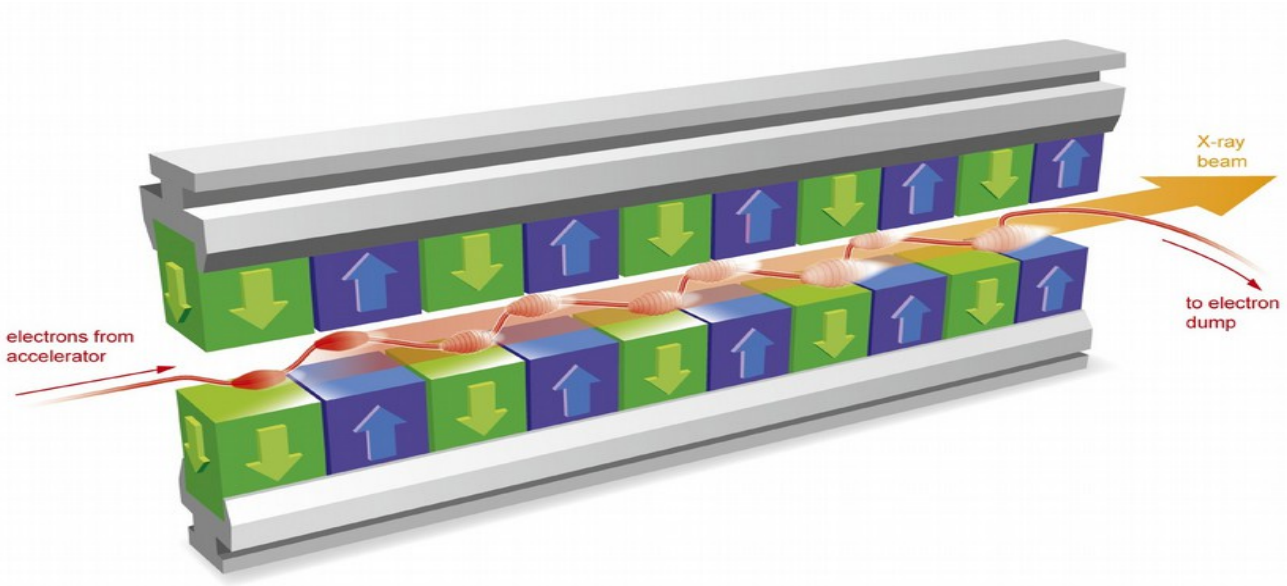
- New detectors achieve higher data rates
  - Lambda (60 Gb/s@ 2kHZ), Eiger (30Gb/s @ 2kHZ), AGIPD
- New experimental setups
  - CFEL: Crystallography



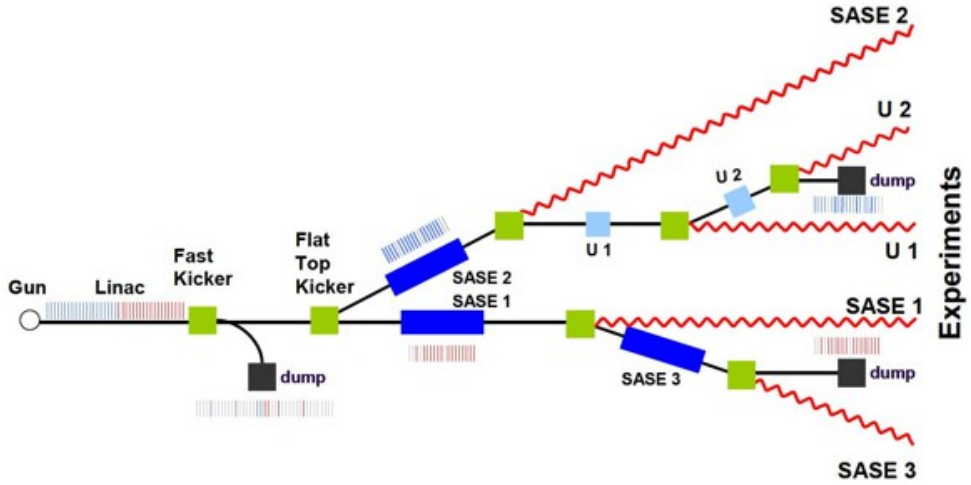


European XFEL – http://www.xfel.eu

- Europe scale project, >1.2 B Euro, 11 member states, construction started 2009, expect regular operations in 2017
- ultrashort X-ray flashes - 27 000 times per second with a brilliance that is a billion times higher than that of the best conventional X-ray radiation sources.
  - make movies while atoms building molecules

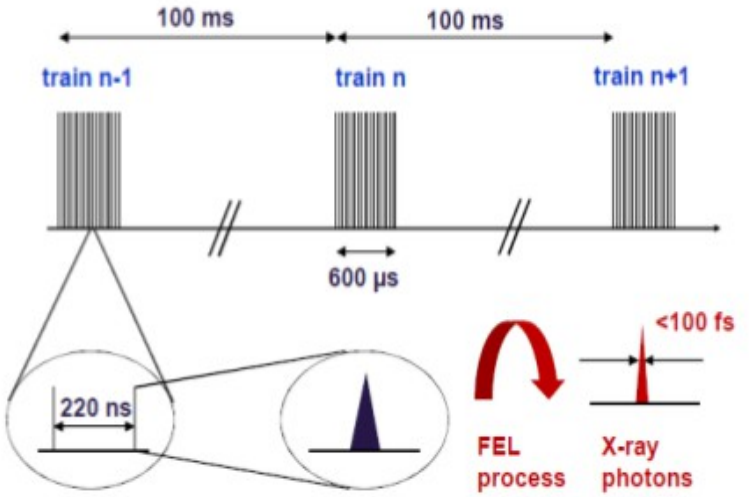


XFEL – beamlines & site structure



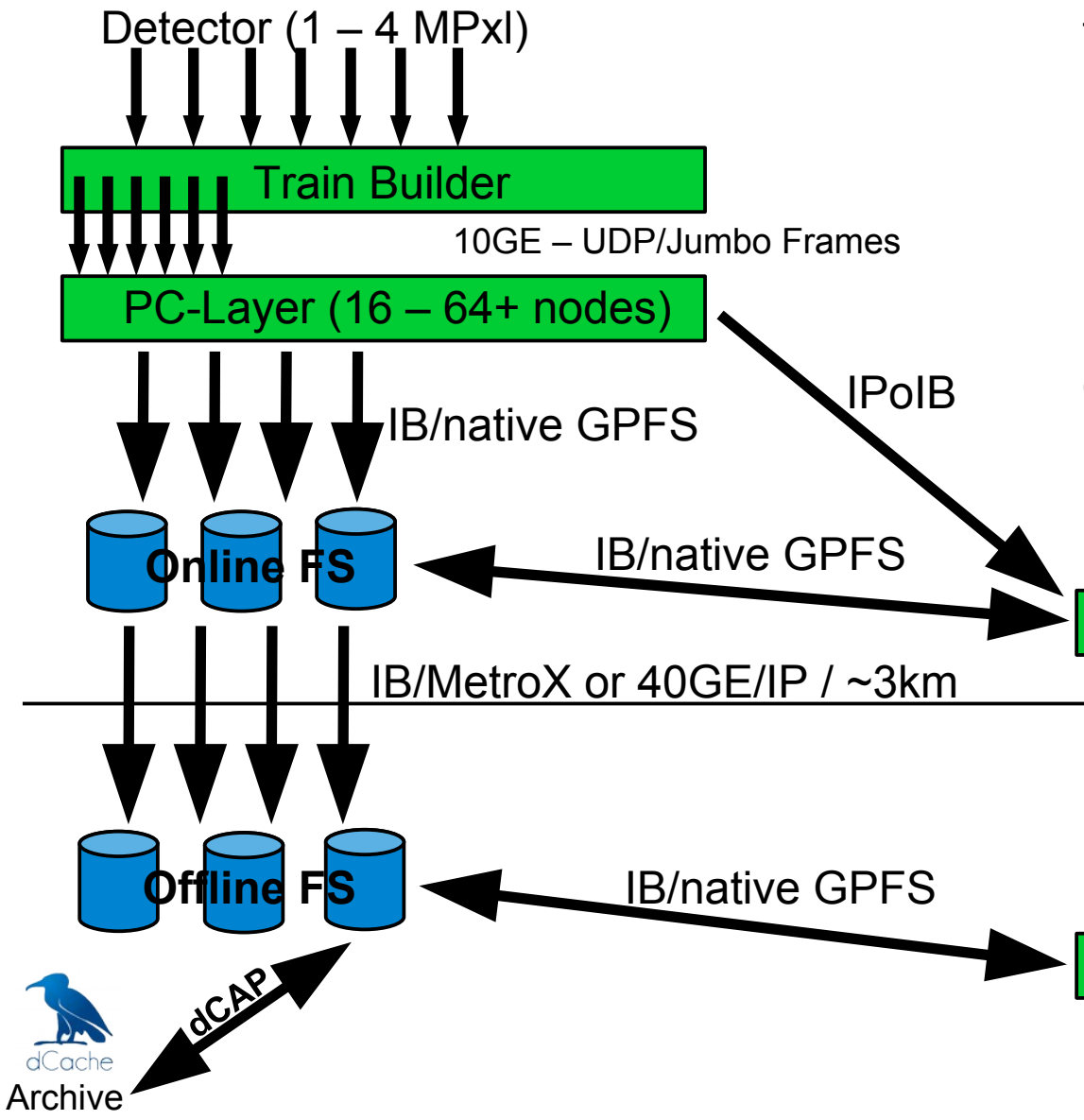
XFEL: DAQ rates and volume

- train contains 1 – 2700 pulses
- detector sync with train
- size and volume - depending on detector and pulses per train
- file format HDF5
- 1 – N trains per file
  - 1GB up to >10GB
- every PC-Layer nodes creates a 1GB file per 1.6 seconds – one train per HDF5 file – or multiple of this



Detector	Data/Pulse	Data/Train	Rate
1 Mpxl 2D camera	~2 MB	~1 GB	~10 GB/s
4 Mpxl 2D camera	~8 MB	~3 GB	~30 GB/s

Online & Offline Data Flow



Train Builder

Reshufles picture modules into whole picture
Reshufles pictures into trains
Sends single trains per channel

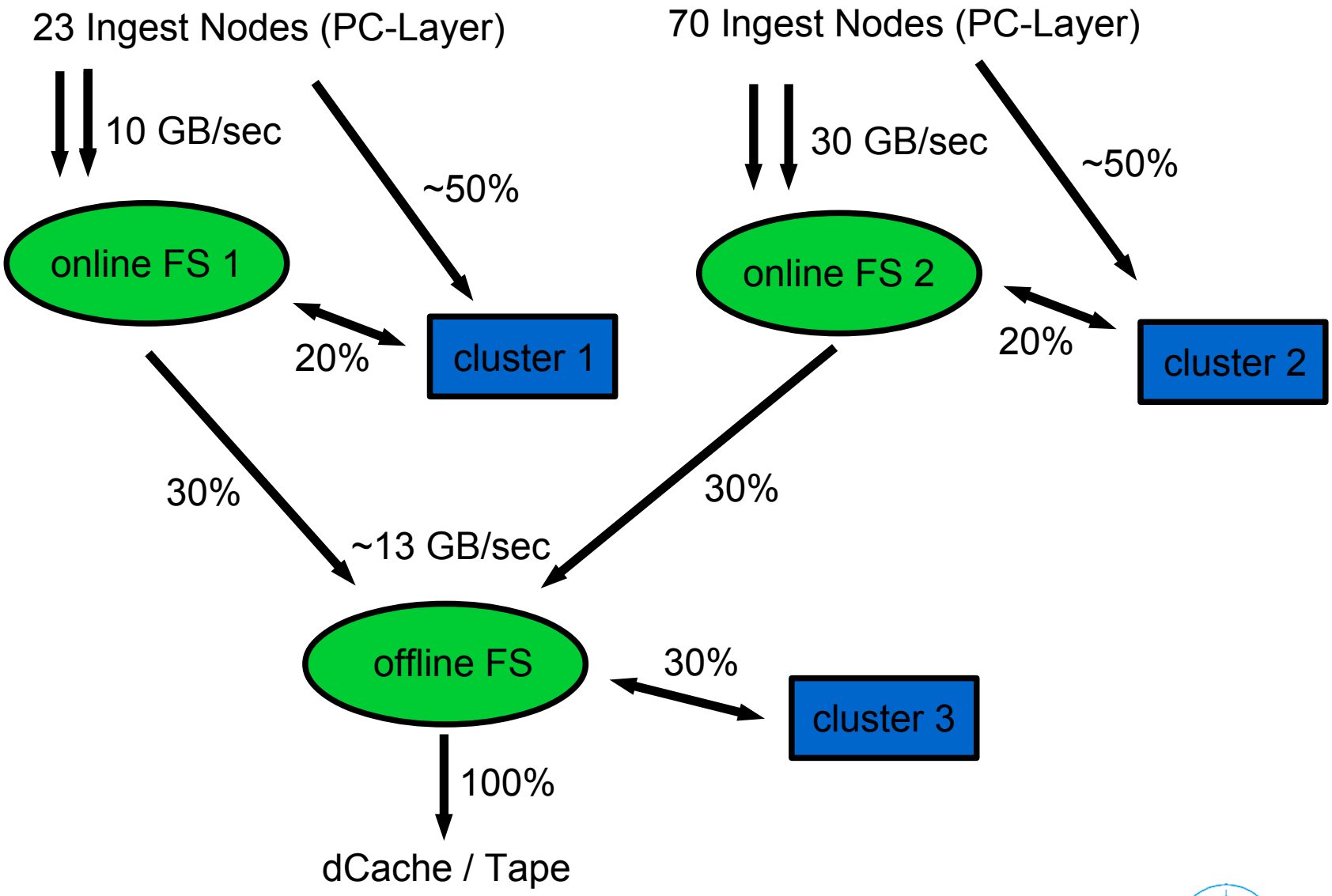
PC-Layer

Data analysis for monitoring
Data Reduction / i.e. FPGA compression
Veto
File creation in memory & Online FS

Online FS -> Offline FS

Evaluating: InfiniBand with MetroX or 40GE Ethernet
Evaluating: AFM or custom scripts for data transfer
Main DAQ stream: Highest priority (QoS)
Initial Test Setup: 2x IBM ESS GL4 / MetroX

further abstraction

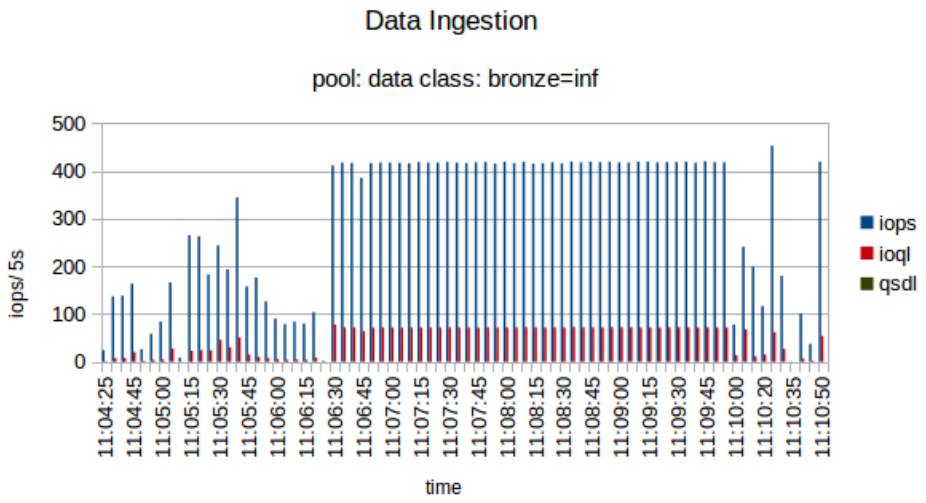


Challenges for GPFS

- Handling large bursts for longer periods
  - ~50 GB/s for 30 minutes
  - Memory (Storage Class Memory) based storage for handling bursts?
- Quality of service for online filesystem (QOS -> next slide)
  - use 'throttling' to implement 'priorities'
- Long range InfiniBand with Mellanox MetroX
- Initial (current) work areas:
  - Flash intermixed with spinning disks in large enclosure – lower costs
  - QOS
  - (L)ight (W)eight (E)vents – lighter than DMAPI

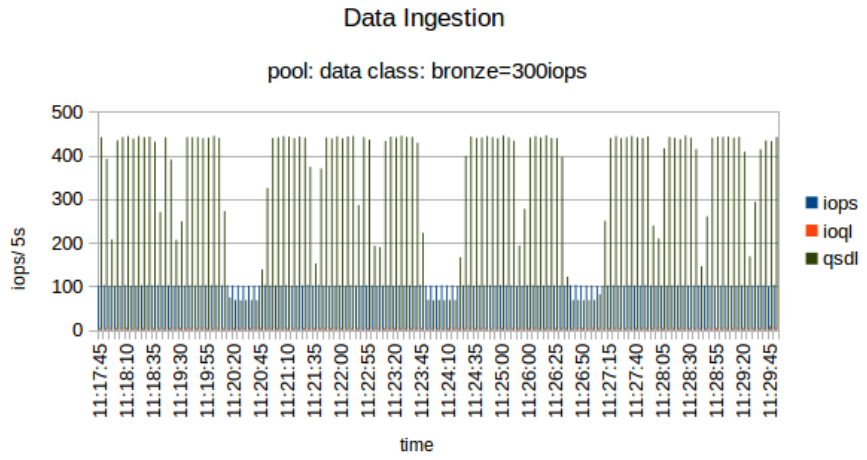
DESY

QOS - observations



no throttling – qsdl == 0
write data (detector simulation)

active throttling – same (IO) process as above



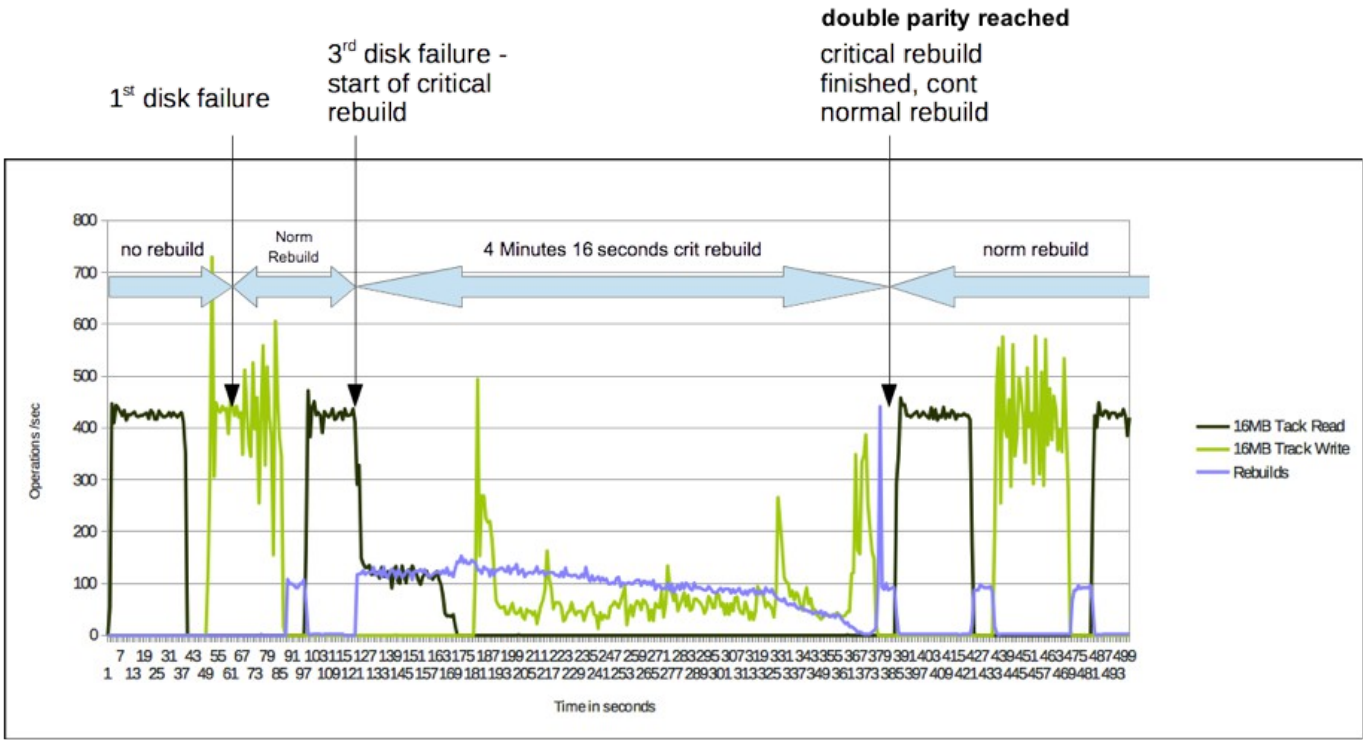
Summary

- GPFS is becoming a core component around data taking
  - fast+reliable+scalable - GPFS Native RAID (GNR) a 'must have'
  - concurrent access with multiple protocols (SMB, NFS, nat. GPFS, Object/Swift)
  - data life cycle management for automated data flow (**policy runs / LWE**)
  - integration with our (sometimes) weird infrastructure
  - powerful developments (sail behind HPC – Coral) – increasing
  - enterprise features – NFSv4ACL, Quota (inode, capacity), XATTR, ...
  - filesets (filesystem in a filesystem)
  - nearly unlimited number of config parameters (tuning potential)
- Development continues, on both sides – EuXFEL preparation
  - joined – Flash Intermix, QOS, LWE, ...
    - Power + FPGA for compression, commodity SSD for read-mostly, faster access data
  - ZMQ/NanoMQ based 'first mile' – faster (RDMA), native GPFS ingest, lightweight

Questions?

fast rebuild in operation – rebuilding req. protection

414TB out of 436TB are in use / T_0 first Disk failed, T_0 +20 - second failed, T_0 +35 – third failed



by courtesy of Sven Oehme IBM/Research Almaden