ASAP^3: New Data Taking and Analysis Infrastructure for PETRA III.

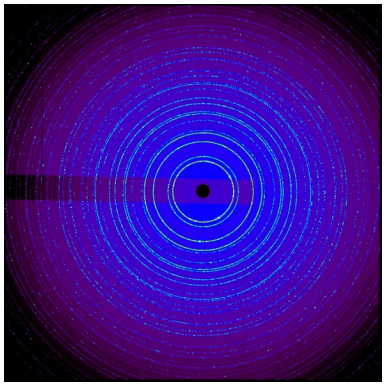
Stefan Dietrich
Co-Author: Martin Gasthuber, Marco Strutz, Manuela Kuhn, Uwe Ensslin, Steve Aplin
HEPiX Spring 2015 Workshop
Oxford University (UK), 2015-03-25

HELMHOLTZ
GEMEINSCHAFT

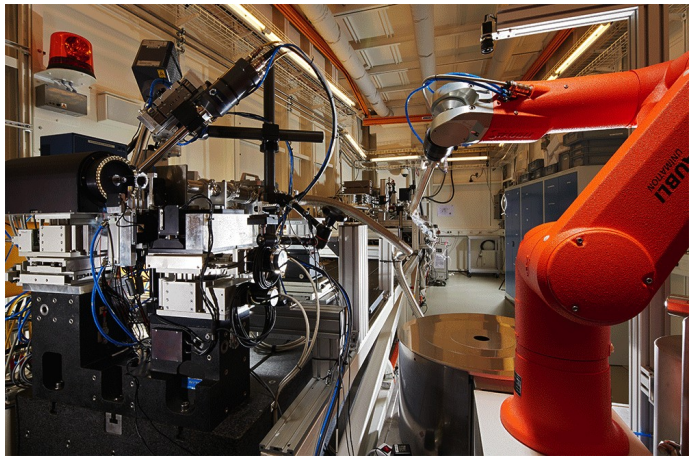
DESY

PETRA III

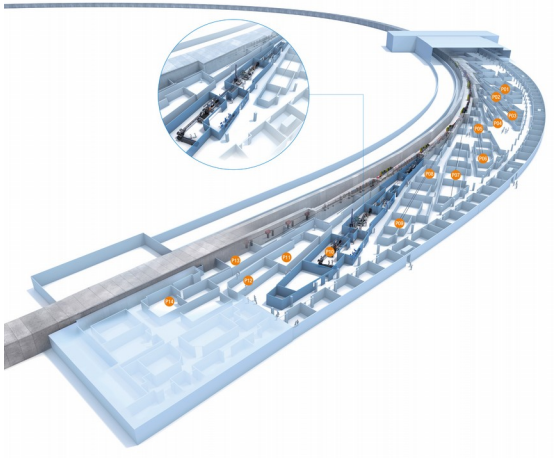
- Ring accelerator
- X-ray radiation
- Since 2009: 14 beamlines in operation
- Since February 2014: Shutdown for new extension



Sample raw file

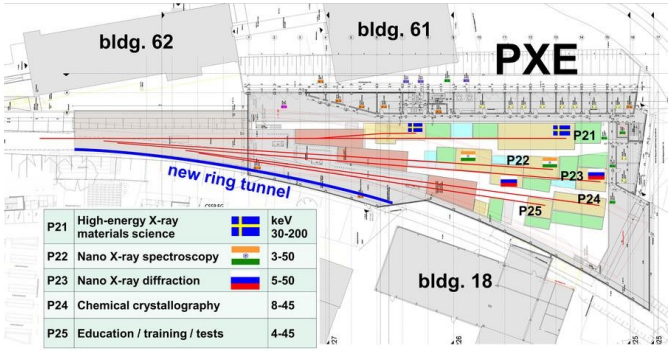


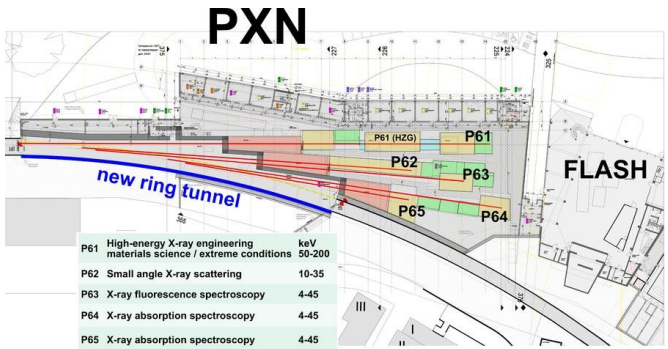
Beamline P11
Bio-Imaging and diffraction



PETRA III Extension

- Extension for PETRA III
- 2 new experiment halls
- 10 new beamlines
- Bigger and faster detectors...
- Planned operational start: April 2015





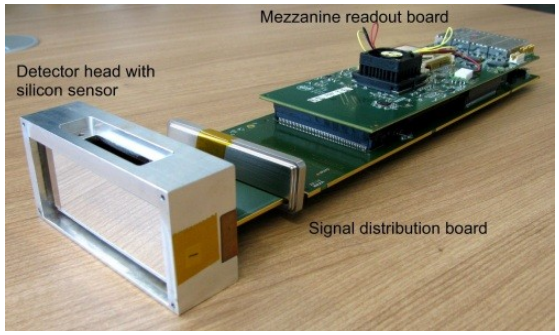
New Challenges

> New detectors achieve higher data rates:

- Pilatus 300k: 1,2 MB Files @ 200 Hz
- Pilatus 6M: 25 MB files @ 25 Hz
  7 MB files @ 100 Hz
- PCO Edge: 8 MB files @ 100Hz
- PerkinElmer: 16 MB + 700 Byte files @ 15 Hz
- Lambda: 60 Gb/s @ 2000 Hz
- Eiger: 30 Gb/s @ 2000 Hz

> Old storage system hit limits

> New storage system has to be installed during PETRA III shutdown!



Limitations

- Datacenter is ~1 km away
- Low space in experiment hall and at beamline
  - Local storage is no option
- 10 Gigabit Ethernet available
- Mix of operating systems:
  - Windows
  - Multiple Linux distributions
  - Sometimes unsupported versions
- Shared accounts for data-acquisition per beamline
- Time, personal and money is limited

Requirements for New Storage System

- High performance for single clients > 1GB/s
- Handle data peaks (Burst Buffer)
  - Data-acquisition has bursty nature
  - First measurement, change sample, second measurement and so on
  - Duration: minutes, hours, days
- Protection between beamlines
  - Competition, data must not be readable from every PC
  - Data must not be readable by next scientist at beamline
  - Data-acquisition of a beamline should not interfere other beamline

We Are Not Alone...

diamond

ESRF
The European Synchrotron

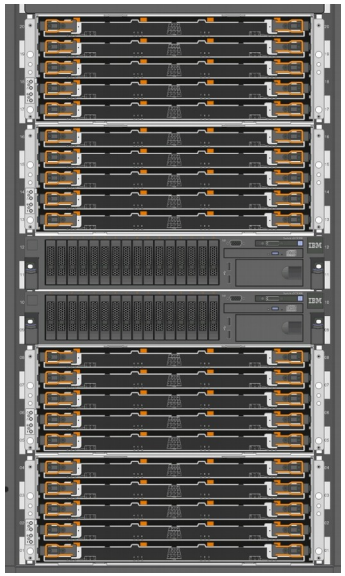
PAUL SCHERRER INSTITUT
PSI

DESY

DESY & IBM Collaboration

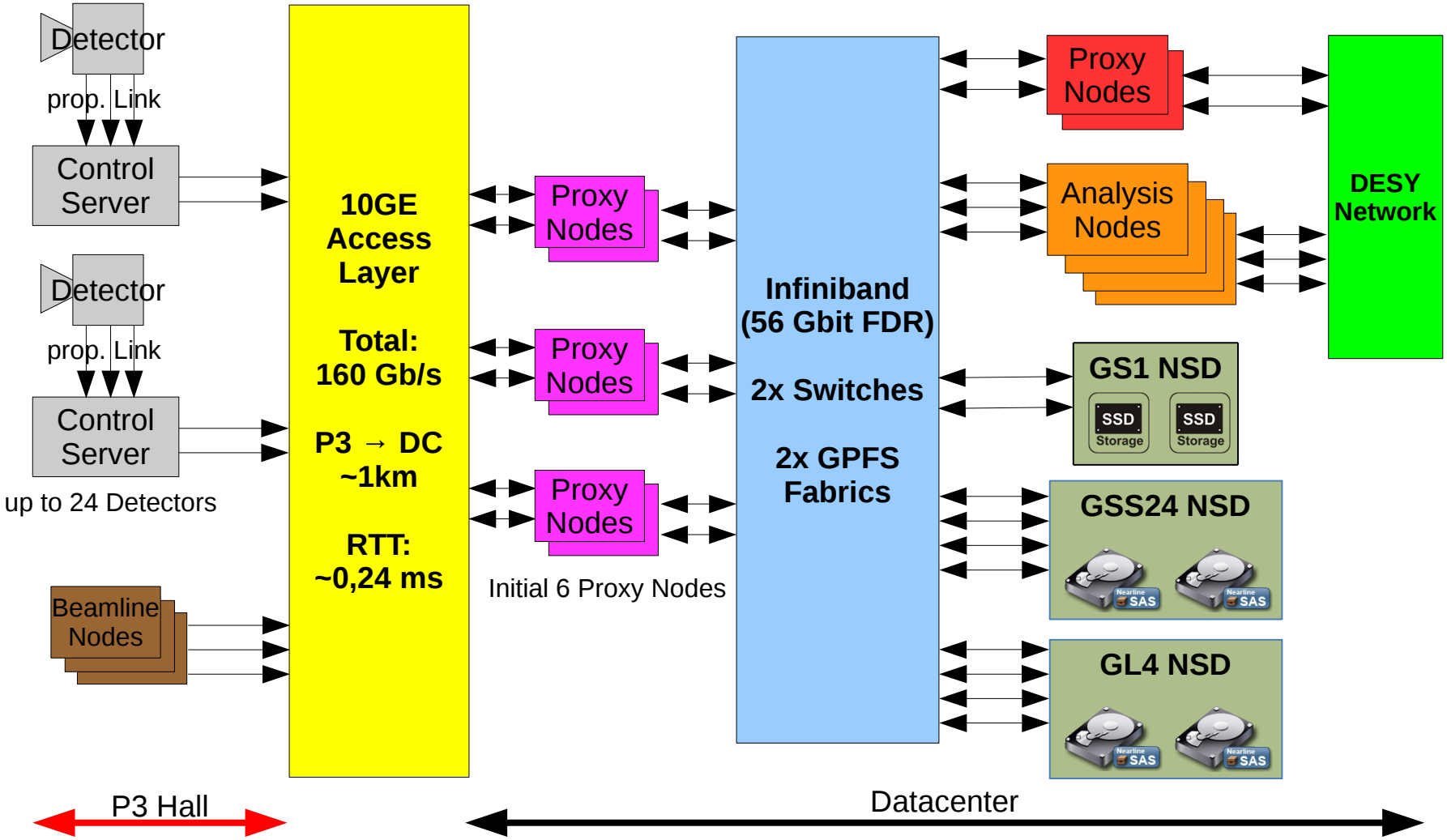
- Collaboration with IBM within scope of SPEED project
- Timeline for SPEED project: June 2014 -> March 2015
- For DESY: Access to experts from development, research and support
- 6 people from DESY, 3+ from IBM
- Solution based on GPFS and Elastic Storage Server (ESS)
  - GPFS 4.1.0-6
  - ESS supports “GPFS Native RAID”
- Initial invest: 1x GSS24 (232x 3TB NLSAS)
- Loan:
  - 1x ESS GL4 (232x 4TB NLSAS)
  - 1x ESS GS1 (24x 400GB SSD)
- So far good performance and stability
- Convert GSS24 → ESS GL4

ESS GS1



ESS GL4/GSS24

New Architecture

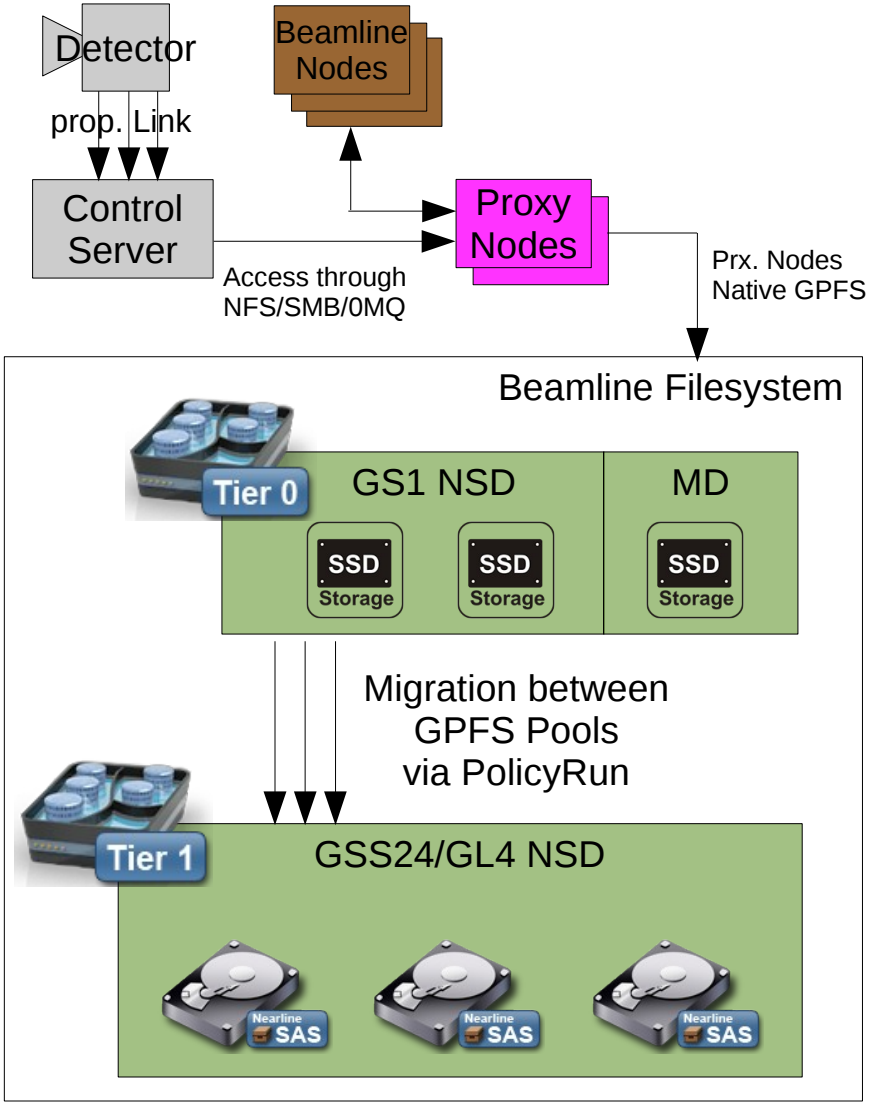


Access Protocols

- Proxy nodes export GPFS for multiple protocols
- Beamline
  - NFSv3 (Kernel)
  - SMB, based on Samba 4.2
  - ZeroMQ
- ZeroMQ: Messaging library
  - Available for multiple languages
  - Multiple message patterns available (PUSH/PULL, REQ/REPLY)
  - One-way tunnel from detector to GPFS
- Core
  - NFSv4.1 (Ganesha)
  - SMB, based on Samba 4.2
  - Native GPFS

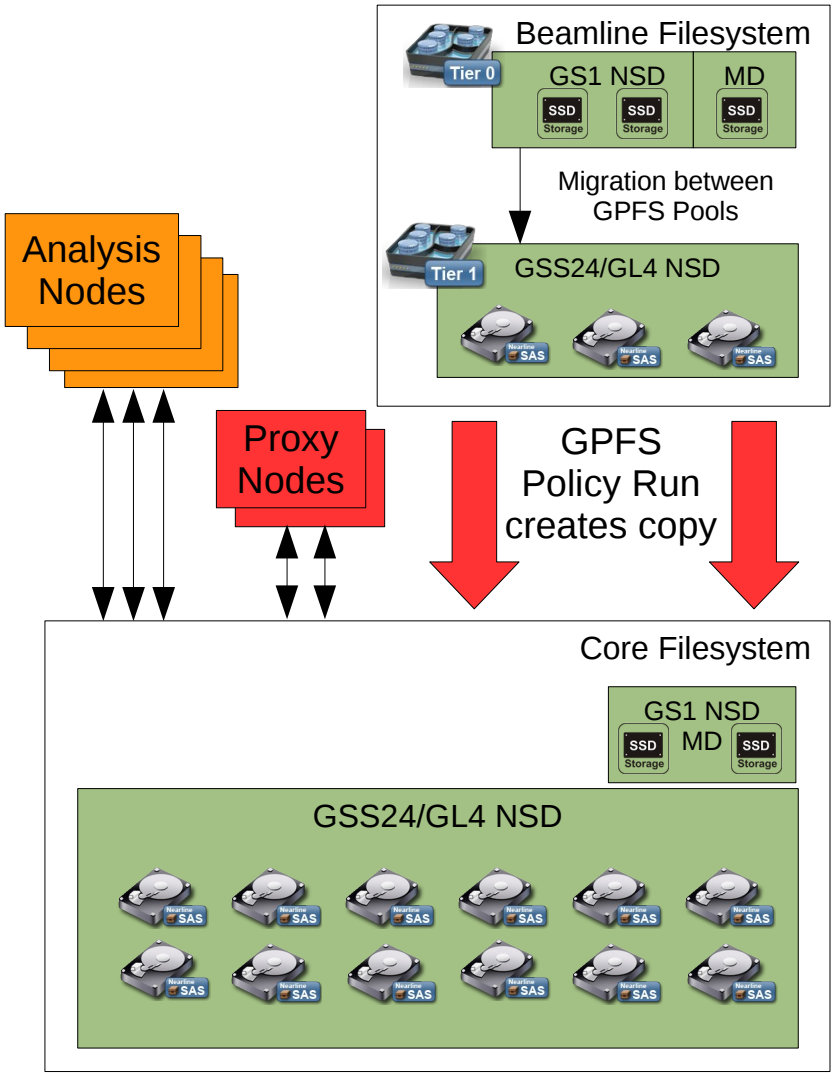
Beamline Filesystem

- “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: ~60 TB capacity

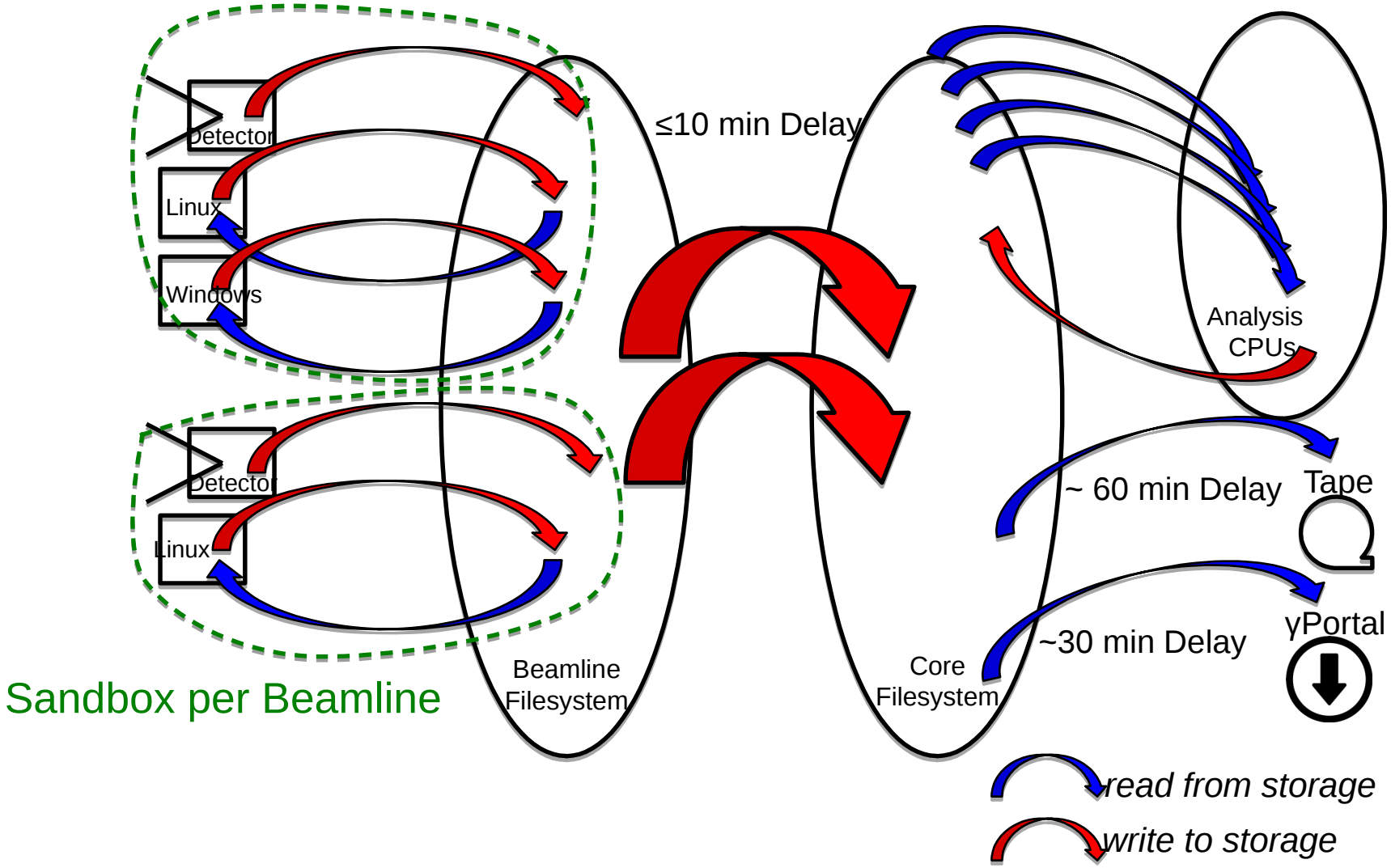


Core Filesystem

- “Clean world”
- Full user authentication
- NFSv4 ACLs
- Access through NFSv4, SMB or native GPFS
- GPFS Policy Runs copy data
  - Beamline → Core Filesystem
  - Single UID/GID
  - ACL inheritance gets active
  - Raw data set to immutable
- 8 MiB filesystem blocksize
- Fileset per beamtime



Dataflow from Detector



Gamma-Portal

- Download of data through web browser
- Login with DOOR accounts
- Folder and files will be “tagged”
  - Setting extended attributes (XATTR)
- GPFS Policy Runs create list with tagged folders/files
- Default folder structure
  - raw: Raw data
  - shared: Data for the portal
  - processed: Processed files
  - scratch: Scratch Space
- XATTR allow different folder structure for power users
- Gamma Portal uses list for display

Gamma-Portal – Beamtime Overview

GAMMA-PORTAL
Data Management for Photon-Science

Browse archive

- Home
- Browse archive
- Staging status
- Migration status
- Data Download

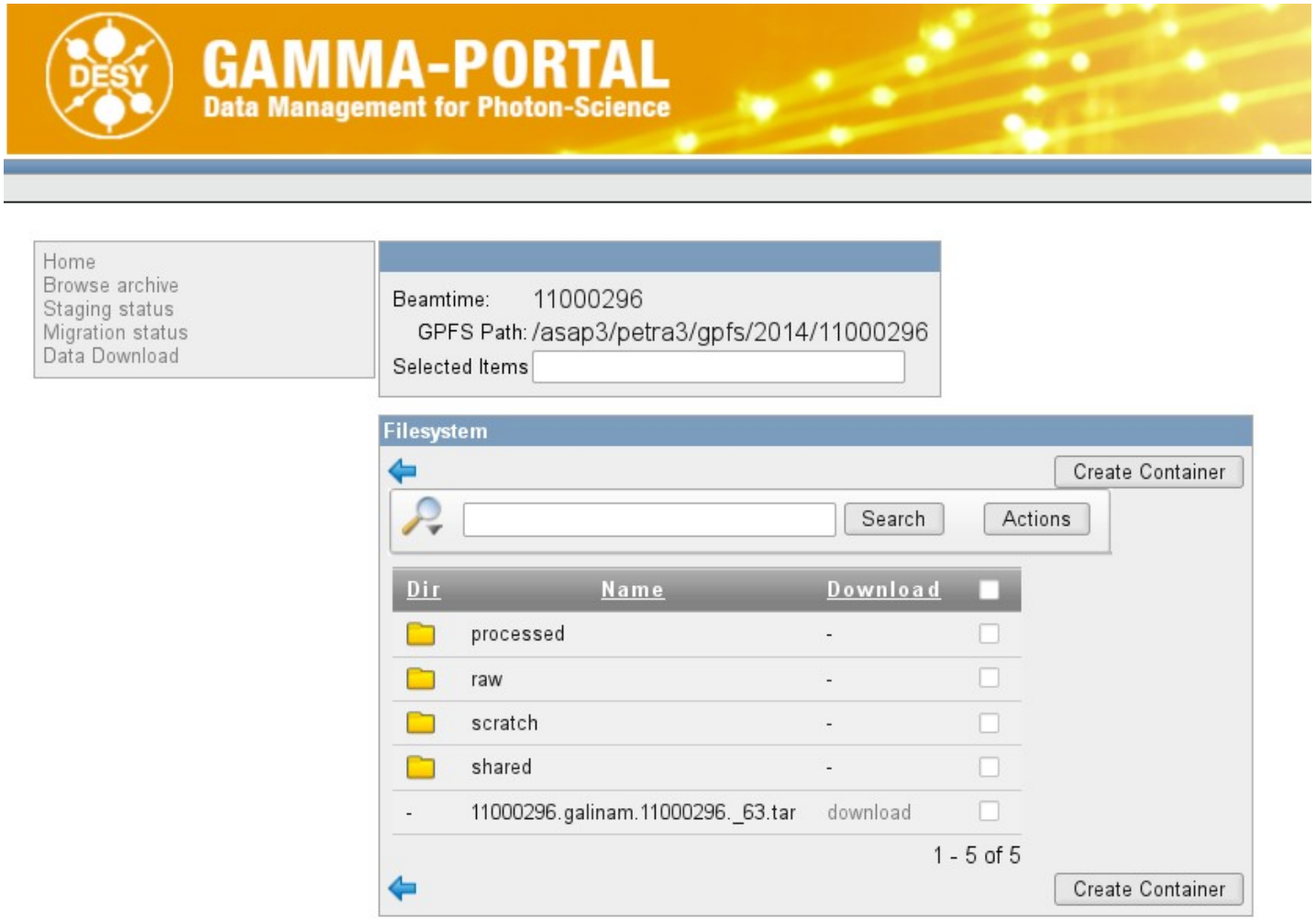
Browse archive

Search

Beamtime	Facility	Proposal Id	Evtstart	Evtend	Localcontact	Size in Mb	Files	Users ACLs	Beamline
10000002	A2	20060072	11-DEC-06	13-DEC-06	Gehrke	960.97			DORIS
10000269	PG1	20000004	01-AUG-09	03-AUG-09	-	960.18			FLASH
10000279	P01	20080031	21-JAN-09	24-JAN-09	Laasch	0.75			PETRA III
10000307	P09	20000025	19-NOV-09	22-NOV-09	-	0.00			PETRA III
10000307	P09	20000025	19-NOV-09	22-NOV-09	-	0.75			PETRA III
10000318	PG2	20000003	23-JAN-10	26-JAN-10	-	0.00			FLASH
10000325	A2	20000001	22-FEB-10	23-FEB-10	-	0.00			DORIS
10000326	P01	20100001	06-APR-10	08-APR-10	Schluenzen	0.00			PETRA III
10000347	X	20100020	12-JUL-10	19-JUL-10	Kurz	0.01			DORIS
10000351	P10	20090008	02-AUG-10	05-AUG-10	Franz	0.00			PETRA III
10000355	P06	20000025	01-AUG-10	02-AUG-10	-	961.73			PETRA III
10000357	P10	20090008	30-JUL-10	31-JUL-10	Franz	46.43			PETRA III
10000395	BL1	20080020	16-FEB-11	19-FEB-11	Duesterer	0.00			FLASH
10000445	P01	20100019	14-NOV-11	16-NOV-11	Drube	6,005.10			PETRA III
10000472	P03	20110005	12-OCT-11	14-OCT-11	-	7,478.78			PETRA III

1 - 15

Gamma-Portal – Beamtime Download

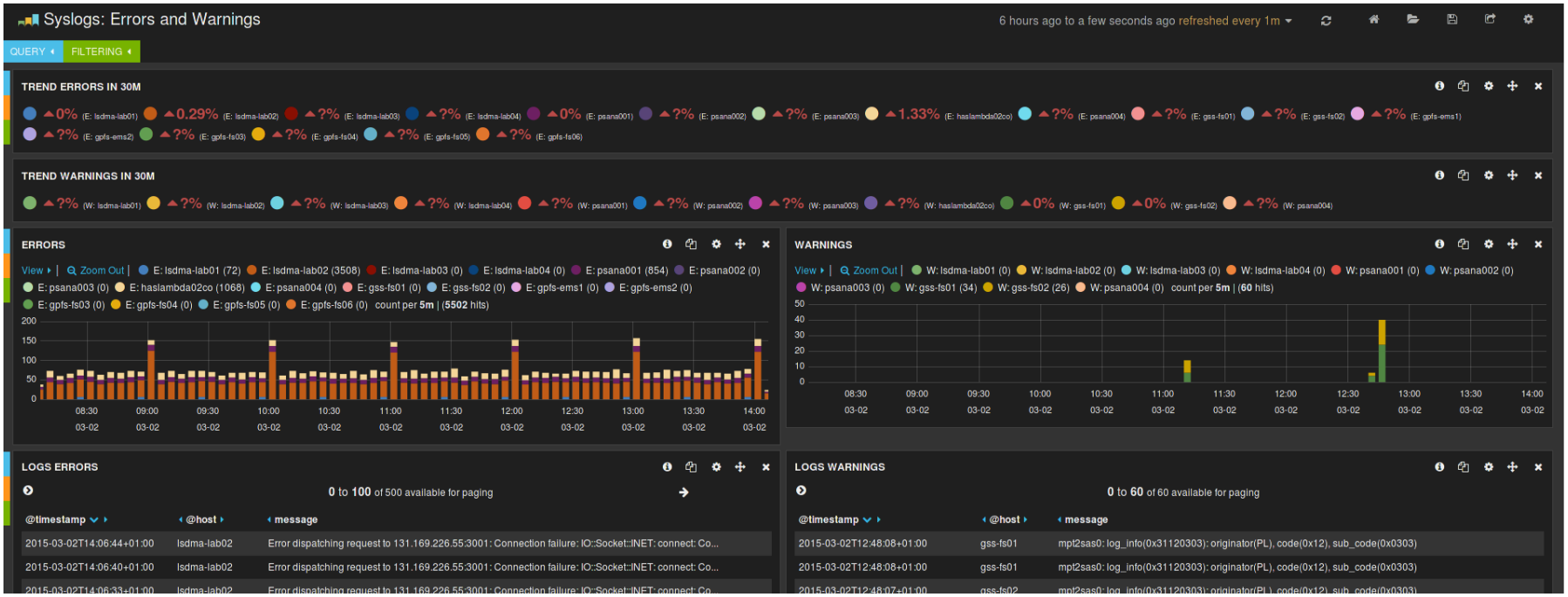


GPFS Monitoring

- Monitoring is an important part
- Hardware Monitoring
  - Checks for Nagios/Icinga and RZ-Monitor (home grown)
- Dashboard view
  - Read-only view for important resources and metrics
  - For people at the beamline or operating
- Expert-/Admin view
  - Correlation between multiple events
  - Error analyses and debugging
  - View current resource usage
  - Planning for resources
- Using Elasticsearch and Kibana
  - Data will be collected through rsyslog and ZIMon (IBM proprietary)

Kibana - GPFS Filesystem Metrics

Kibana – Log Analysis

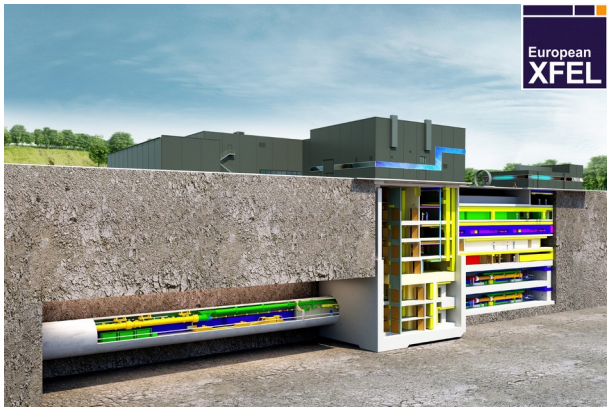


Things we hope(d) would work, but...

- 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 behaviour not optimal for this use-case
- Self-made SSD burst buffer
  - SSDs died very fast, firmware bug according to vendor
- Remote Cluster UID-Mapping

Outlook

- Connection analysis cluster
  - Connection over Infiniband
  - Native GPFS access
  - 18x Linux Nodes, 4x incl. GPUs
  - 2x Windows machines planned
- Further development of ZeroMQ approach
- Next customer XFEL (http://www.xfel.eu/)?
  - Start in 2016/2017
  - Use new storage system as base
  - Expected datavolume: 100 PB



Summary

- GPFS is able to handle the data rate
- GSS/ESS delivers good performance and stability
- Current architecture offers multiple options for scaling
- Detectors with Windows are tricky to handle
- ZeroMQ is viable option for data transfer
- Still some work ahead, not yet everything finished

Questions?

Backup: Data Archival on Tape

- Measurement data has to be archived on tape
- GPFS supports TSM, but not desired
  - Cost reason
- Instead, use dCache for tape archival
- Same process used as in the portal
- Files and folder will be “tagged” for archival
- GPFS Policy Run creates file list
- Files will be copied with DCAP to dCache
  - DCAP retains NFSv4 ACLs during copy
  - Linux tools loose NFSv4 ACL information

Backup: Beamtime Setup

- Measurement time is a beamtime
- Management script for setting up beamtime
  - Fileset creation on beamline and core filesystem
  - Create default folder structure
  - Setup NFSv4 ACLs on core
  - Activate Policy Runs for migration
  - Setup of beamline NFS and SMB exports and ZeroMQ endpoints
- End of a beamtime
  - Filesets and exports will be removed from beamline filesystem
  - Fileset on core filesystem remains
  - Further analysis possible