From PETRA III to XFEL - Online & Offline Storage System based on Common Grounds

Stefan Dietrich
Co-Author: Martin Gasthuber, Manuela Kuhn, Janusz Malka, Uwe Ensslin
HEPiX Spring 2016 Workshop
DESY Zeuthen (DE), 2016-04-20
Agenda

> Architecture Review and Introduced Changes
> Operational Issues
> ZeroMQ for Detector Data Transfer
> Current Activities for XFEL
> Summary and Outlook
Recapture

> Current and future detectors exceeded capabilities of storage system

> SPEED project: DESY and IBM collaboration to setup new system

> Solution based on IBM Spectrum Scale and Elastic Storage Server (ESS)
  - Data ingest via SMB/NFS/ZMQ
  - Separation between beamline (online) and offline operation (analysis etc.)
  - SSD burst buffer
  - Declustered software RAID on ESS

> Running in production since April 2015

> See HEPiX Spring 2015 talk “ASAP3: New data taking and analysis infrastructure for PETRA III“ for more details!
ASAP³ Architecture

- Detector
- Control Server
- Proxy Nodes
- Proxy Nodes
- Proxy Nodes
- Proxy Nodes
- Helper Nodes
- Proxy Nodes
- Maxwell Resources
- GS1 NSD
- GS1 NSD
- GL4 NSD
- GL6 NSD

10GE Access Layer
- Total: 160 Gb/s
- P3 → DC
- ~1km
- RTT: ~0.24 ms

Infiniband (56 Gbit FDR)
- 2x Fabrics:
- Maxwell 2:1 Blocking
- Private non-blocking

Cutom Link
- Datacenter
- Detector
- Control Server
- Beamline Nodes
- P3 Hall

up to 24 Detectors

6 Proxy Nodes for 12 Beamlines
1 Detector Specific Proxy
Changes in the Architecture

> Consolidated FDR InfiniBand Fabric in new racks

- ASAP3 now uses the Maxwell fabric
- Maxwell Fabric: 2:1 Blocking, with 3 top and 8 leaf switches (192 ports)
- Private Fabric: 1 switch for redundancy, only critical nodes connected (36 ports)
- No congestion observed so far

> Replaced GSS24 with ESS GL6

- No longer mixed building blocks
- Administration and support easier
- Additional capacity used for home filesystem for Maxwell analysis cluster

> Major changes for the Gamma Portal

- UNIX group based ACLs instead of user ACLs
- Easier to handle for adding/removing people
- Deprecated container build feature, now data download via FTPS
First Production Period

> Statistics for 2015
  - 369 beamtimes stored
  - 104 commissioning runs
  - Total volume: ~300 TB and ~70 Mio. files

> Overall user experience: good!
  - No beamtime loss due to lacking space
  - BL scientist: More time for experiment, sample preparation and user support
  - Reconstruction faster and more stable

> Shutdown has been used to update systems
  - ESS 4.0.2 and Spectrum Scale 4.2.0.2

> PETRA III runtime for 2016:
  April – December
ZeroMQ for Data Transfer

> First two use cases implemented in Python

> “Vacuum Cleaner”

- Data picked up on the detector and send through ZMQ to GPFS
- Currently in testing phase, first use case for the PerkinElmer (Windows) detector

> Live Viewer and Online Data Analysis

- Send images to a receiver for display/monitoring/analysis at beamline
ZeroMQ Data Transfer: Dectris Eiger 4M

> Next gen. detector: Dectris Eiger 4M
  - Capable of 30 Gb/s @ 2000 Hz

> Data access via HTTP

> Directly outputs compressed or uncompressed HDF5 container

> Compressed
  - Detector limited to 500 Hz
  - Output size varies, e.g. 500 Hz with 500 Frames, ~31 MB HDF5/s

> Uncompressed
  - Outputs 4 GB HDF5 container every second
  - ~750 Hz possible
  - Single 10GE link and internal buffer size not sufficient
ZeroMQ Data Transfer: Architecture

> Dedicated Beamline proxy for initial setup

> Data Receiver queries for new files and downloads them via HTTP GET
  
  ▪ Single stream sufficient for compressed containers
  ▪ 16 streams used for uncompressed files, can use the whole 10GE link

> Online Data Analysis (OnDa) access files via metadata from GPFS

> Display results and change experiment parameters with little delay!
Beamline Dashboard

> Display relevant information on Dashboard for user
> Metrics provided by Icinga and Elasticsearch
> Grafana used for visualization
Operational Issues

> Most of the critical encountered GPFS issues were related to InfiniBand

- Problems with Mellanox OFED on ppc64, e.g. Blueflame, MSI-32
- Firmware bugs in the InfiniBand HCAs
  > High load caused the HCA to reset, but did not recover properly
  > Resulted in 11 corrupted files, half of them could be restored
- Root cause has been identified and solved by IBM and Mellanox

> GPFS Deadlocks

- Long running waiters causing the cluster to become unhealthy
- Partially solved by new GPFS releases or efixes
- Ongoing, still experiencing deadlocks in erratic periods

> Installation Toolkit

- xCAT is required for installation and update of ESS
- Requires ded. DHCP and DNS, runs “only” on POWER8
- For now, we use single xCAT machine to install all ESS systems
From PETRA III to European XFEL

> Europe scale project
> 1.2 billion Euro for construction, 11 participating countries
> Construction started 2009, expect regular operations in 2017
> Ultra-short X-ray flashes
  - 27,000 times per second
  - Billion times higher brilliance than conventional X-ray radiation sources
  - Make movies while atoms build molecules
Beamlines and Site Structure

- electron tunnel
- photon tunnel
- undulator
- electron switch
- electron bend
- electron dump

Linear accelerator
for electrons (10.5, 14.8, 17.5 GeV)

SASE 2
0.05 nm - 0.4 nm

SASE 1
0.05 nm - 0.4 nm

SASE 3
0.4 nm - 4.7 nm

MID
Materials Imaging and Dynamics

HED
High-energy Density matter Scattering

SPB
Single Particles, Beams & Sources

FXE
Fermilab Fast Experiments

SQS
Small Quantum Systems

SCS
Solid State Physics & Scattering

Schenefeld
Hamburg-Osdorf
Hamburg-Bahrenfeld

Stefan Dietrich | ASAP³: Status Update and XFEL Activities | 2016-04-20 | Page 13/20
DAQ Rates and Volume

- Train contains 1 – 2700 pulses
- Detector sync with train
- Size and volume depending on detector and pulses per train
- Directly outputs HDF5 files
- 1 – N trains per HDF5 file
  - 1 GB up to >10 GB
- Initially: 1 train per file
  - Every PC-layer nodes outputs 1 GB file per 1.6 seconds
  - Increase volume by having more pulses and trains/file

<table>
<thead>
<tr>
<th>Detector</th>
<th>Data/Pulse</th>
<th>Data/Train</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mpxl 2D camera</td>
<td>~2 MB</td>
<td>~1 GB</td>
<td>~10 GB/s</td>
</tr>
<tr>
<td>4 Mpxl 2D camera</td>
<td>~8 MB</td>
<td>~3 GB</td>
<td>~30 GB/s</td>
</tr>
</tbody>
</table>
Online and Offline Data Flow

---

**Train Builder**
- Reshuffles picture modules into whole picture
- Pictures shuffled in trains
- Sends single trains per channel

**PC-Layer**
- Data analysis for monitoring
- Data Reduction, e.g. FPGA compression
- Veto
- File creation in memory and online filesystem

**Online Cluster**
- 10-80 nodes
- Online data analysis and re-calibration

**Transfer Online → Offline Storage**
- Evaluation: MetroX or 40GE Ethernet
- Evaluation: GPFS AFM or custom scripts

**Offline Storage**
- Shared across multiple SASE
- Data arrives after delay, stored on GPFS
- Copy data to dCache for long term archival
- Raw data access only from dCache (TBD)
- Offline cluster stores calibrated data on GPFS
- Additional analysis from calibrated data
Challenges for GPFS

> Handling large bursts for longer periods
  - ~30 GB/s for 30 minutes
  - Memory based storage for handling bursts?

> Quality of service for online filesystem
  - Online cluster not allowed to disturb PC-layer
  - Losing trains in the worst case → very bad
  - Data transfer from online to offline filesystem also eligible for QoS

> Long range InfiniBand with Mellanox MetroX
  - Token management issues due to latency?

> “Mixed mode” enclosures
  - SSDs with NLSAS drives in single enclosure
  - Motivation: Increase IOPS, lower cost, space constrains
Initial Test Setup for XFEL

> 2x ESS GL4
  - ~900 TB raw capacity with 4 TB drives
  - 1x GL4 for online storage in Schenefeld, up and running
  - 1x GL4 for offline storage in DESY HH datacenter

> 2x Mellanox MetroX TX6100
  - Evaluation equipment provided by Mellanox
  - 3x long range fibre uplinks
  - 6x IB FDR links to local switch

> EDR InfiniBand infrastructure
  - Clients will stay on FDR for now

> First tests with QoS from GPFS
Summary and Outlook

> Users are happy with the new infrastructure
> GPFS proved stable enough
  ▪ However, nasty deadlocks decrease cluster reliability and availability
> Expected XFEL data rates will be challenging
> (Finally) start migration to Cluster Export Services (CES)
  ▪ Core cluster will receive capability for high available NFSv4.1 and SMB
  ▪ Beamline cluster will stay on custom setup
> Prepare ZeroMQ data transfer for general production
> Evaluating FPGA compression on POWER8 for Lambda
> Performance regression tool under development
Questions?
Backup: New Detectors and Changed Setups

> New detectors achieve higher data rates
  - Lambda (60 Gb/s @ 2kHZ), Eiger (30Gb/s @ 2kHZ), AGIPD

> New experimental setups
  - CFEL: Crystallography