ASAP3 - New Data Taking and Analysis Infrastructure for PETRA III

2015 J. Phys.: Conf. Ser. 664 042053

(http://iopscience.iop.org/1742-6596/664/4/042053)

Journal of Physics: Conference Series **664** (2015) 042053 doi:10.1088/1742-6596/664/4/042053

ASAP3 - New Data Taking and Analysis Infrastructure for PETRA III

M Strutz[1], M Gasthuber[1], S Aplin[1], S Dietrich[1], M Kuhn[1], U Ensslin[1], G Smirnov[1], B Lewendel[1] and V Guelzow[1]

[1]DESY, Notkestr. 85, 22607 Hamburg, Germany

E-mail: marco.strutz@desy.de

Abstract. Data taking and analysis infrastructures in HEP (*High Energy Physics*) have evolved during many years to a well known problem domain. In contrast to HEP, third generation synchrotron light sources, existing and upcoming free electron lasers are confronted with an explosion in data rates driven primarily by recent developments in 2D pixel array detectors. The next generation of detectors will produce data in the region upwards of 50 Gbytes per second. At synchrotrons, data was traditionally taken away by users following data taking using portable media. This will clearly not scale at all.

We present first experiences of our new architecture and underlying services based on results taken from the resumption of data taking in April 2015. Technology choices were undertaking over a period of twelve month. The work involved a close collaboration between central IT, beamline controls, and beamline support staff. In addition a cooperation was established between DESY IT and IBM to include industrial research and development experience and skills.

Our approach integrates HPC technologies for storage systems and protocols. In particular, our solution uses a single file-system instance with a multiple protocol access, while operating within a single namespace.

1. Introduction

With a circumference of 2.3 km, PETRA III [1] at DESY (*German Electron Synchrotron*, [2]) is the biggest and most brilliant synchrotron light source in the world. Since the end of 2012 all 14 beamlines are available for users.

2. A Changing Landscape

After data taking, data was put on local commodity media by users, recently on USB3 hard-drives. As upcoming data exceeds by far the 1-disk-capacity, it is not possible to use single hard disks any-more. Also, it it not possible to have a proper data management, access control or archiving mechanism in place for such external media. Furthermore, data rates become higher outperforming specifications for transportable media devices. As a result traditional workflows will not work for future detectors. This effects the whole chain of data handling.

Up to now most of the data pipe-lining happens inside the Computer Center. Experiment PCs and offices desktop PCs are connected to the Computer Center by 1GE to 10GE. Data permissions and delegation for data being produced at the beamlines are handled by a dedicated Data Portal.

The data-processing chain starts with a detector for each beamline, producing many files per seconds with a specific file size.

	Detector	OS	Network	Workload	Data rate	10 minutes run
1	Pilatus 6M (25Hz)	Debian	1x10GBE	• 25 files per second, 25MB each	625 MByte/s	~366GB
2	Pilatus 6M (100Hz)	Debian	1x10GBE	• 100 files per second, 7MB each	700 MByte/s	~410GB
3	PerkinElmer 6XRD 1621 AN(ES)	Windows 7	1x10GBE	• 15 files per second, 16MB each • Each file is accompanied by one ~700 byte file • Appends to a log file, once per minute	240 MByte/s ~10.5 KByte/s ~0 KByte/s	~141GB ~6MB ~0KB
4	PCO Edge (ROI)	Windows 7	1x10GBE	• 900 files per second, 150KB each	~131 MByte/s	~77GB
5	PCO Edge (100Hz)	Windows 7	1x10GBE	• 100 files per second, 8MB each	800 MByte/s	~469GB
6	Lambda (9Gps)	Debian	1x10GBE	• Continuous data stream of 9 GBit/s • Data is stored in 10GB files	1.125 GByte/s	~675GB
7	Lambda (36Gps)	Debian	4x10GBE	• Continuous data stream of 36 GBit/s • Data is stored in 10GB files	4.5 GByte/s	~2700GB

Figure 1. Data Rates from 6 operational and upcoming detectors at PETRA III

Typical data-rates (Figure 1) from the detector to the data storage are for instance 131 MByte/sec (900 images per second, each 150 KByte). Eiger [3], PCO Edge [4] and LAMBDA [5] are next generation detectors differing in frame rate, data rate and the operating systems under which they are managed. As example, data rates of these detectors will be orders of magnitudes higher, so instead of having 131 MByte per seconds per beamline one can expect like from 4.5 up to 10 GBytes/sec.

As beamlines at PETRA III are generally not bound to a specific detector the underlying data management need to be able to cope with the heterogeneity.

3. The Challenge

The development of detectors at 3rd generation light sources is currently outpacing experimental method and data acquisition. Single clients will produce 0.5 GBytes/sec and the next generation is already at frame rates of 2 kHz for 4MB files. For 30 beamlines they provide possible aggregated peak rates of up to an average of 50 GBytes/sec.

Also, measurements last from a few hours to a few days resulting in many single data sets up to tens of TBs each. From next generation detectors we also expect multi GBytes/sec spread over many 10GE connections.

Furthermore there is a very dynamic experimental setup with inherent burst nature and a very heterogeneous environment regarding technology, social context and requirements.

4. Phase Change

On the one hand, we need to rethink previous common practices where storage systems were used as FIFOs, where faster and faster disks could have solved many speed constraints or where file systems were used as central data entry point.

On the other hand there are more and more facilities for light sources with same detectors in place targeting the same user communities challenging same or similar issues regarding upcoming data taking, like ERSF [6], Diamond [7] and PSI [8].

5. New Architecture

5.1. Limitations

Data is produced in the experiment hall. As space is very limited local storage was no option. All network and storage hardware needed to be placed in the Computer Center which is round about one kilometre away from the experimental hall.

Most of the beamlines are equipped with 10 Gigabit Ethernet ports. Regarding operating systems we were faced with a broad mix of Windows and Linux flavours. Some even operated with unsupported versions. Also, for data-acquisition the scientists are using shared accounts per beamline.

5.2. Why no commercial off-the-shelf *hardware*

Because ensuring data integrity was a main goal we assessed various solution, even build our own COTS-SSD-cache (COTS, *commercial off-the-shelf*). Finally we decided to go with GPFS [9] (*General Parallel File System*) to take benefit of its native RAID (*Redundant Array of Independent Disk*) feature [10]. This not only dramatically shortens rebuild-time under load but also provides a prioritized rebuild. GPFS Native RAID uniformly spreads or declusters user data, redundancy information, and spare space across all the disks of a declustered array. Instead of using hardware RAID controller, the logic is implemented fully in software. As soon as a disk fails the system is still able to guarantee data-taking while a background job re-establishes the integrity of all parity information.

5.3. Architecture For the Production System

The ASAP3 storage (Figure 2) is divided into two GPFS file-systems; the beamline file-system and the core file-system. The beamline file-system is optimised for the ingestion of data in high speed bursts, while the core file-system has been optimised for capacity and parallel concurrent access. Data taken during a beamtime is first written to the beamline file-system.

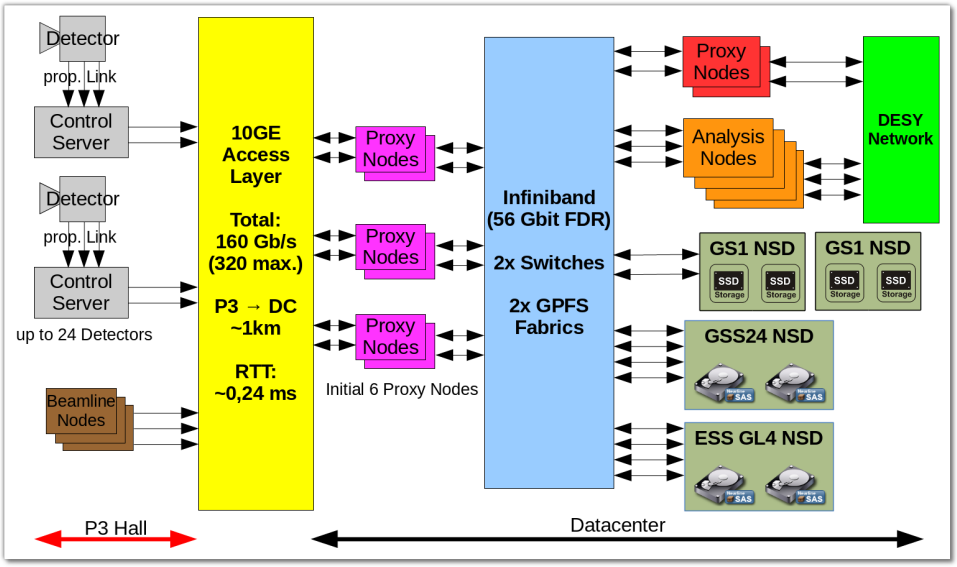

Figure 2. Basic ASAP3 Architecture[1]

[1] *GS1*, *GSS24* and *ESS GL4* are IBM product names. For more information visit `http://www.ibm.com/`

With a delay of minutes data will be asynchronously copied to the core file-system from where it can be accessed and analysed afterwards. While the beamtime is active data can be accessed directly from the beamline file system. This enables any experiment to run a quick analysis of produced data, for instance to align specific detector settings before producing the main bulk of data. Especially experiments with a short time window (like less than an hour) rely on a responsive system. Once the beamtime is stopped corresponding data can only be accessed from the core file-system by then.

The main reasons to split "online" and "offline" area looks familiar to existing architectures, a few points differ in the motivation of doing so, though.

Firstly, the separation of different authentication and authorization concepts. The experiment sections do not use ACL or a user based authentication resp. not any UID/GID management. Hosts are registered for a dedicated experiment network share and from then on any data access initiated from these hosts is granted. While files are being created, the ownership is not of concern and can therefore be managed in any way suited best for the experiment. Nontheless, state of the art ACLs and user authentication are provided at the core file system.

Secondly, the architecture enables a better optimization of the NFS/SMB/GPFS stack with respect to the IO profile produced by detectors.

Thirdly, the architecture provided an automated and controlled migration of data from an experiment (beamline) to the core file-system. This includes ownership changes, ACL inclusion and the preparation of following steps within the "lifecycle management" like archiving to tape and data access from remote clients.

Finally, this approach enables to scale in terms of bandwidth, capacity and IOPS (*Input Output Operations per Seconds*) while being able to apply different optimization and scaling strategies for the beamline and the core file-system.

5.4. The Beamline File-System

The beamline file-system (Figure 3) is dedicated for the data-taking part. Its main function is to face the beamlines and taking out the bursty nature of data generation.
By design it is optimized for high write performance and high IOPS. We decided to use a tiered storage. Tier-0 operates as SSD burst buffer to store round about 10 TB. After a short period of time data is been automatically migrated to Tier-1. It holds up to 60 TB of data on spinning disks.
Because of the "strong" and automated coupling between beamline file-system and the core file-system, and the automated steering of archiving and remote access steps the system must be able to handle a relative high number of GPFS policy runs (a fast scan of namespace with SQL like queries) at high rate.

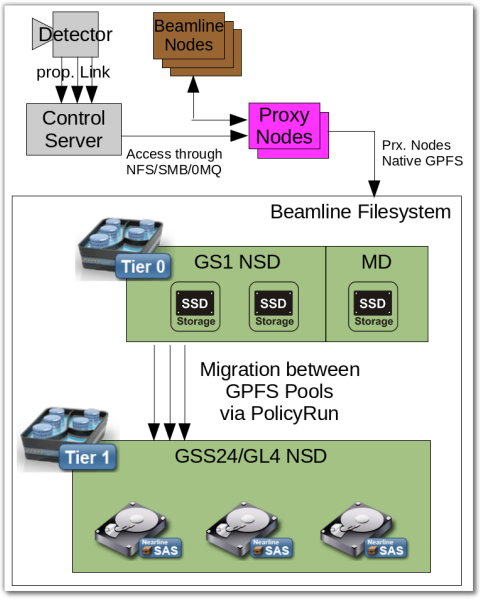

Figure 3. The Beamline File-System.

Leveraging SSDs for the file-system metadata operations results in big improvements with respect to the policy runtime and the overall load on the metadata system.

The beamline file-system is treated as "wild-west" area. Authentication is achieved host-based without involving ACLs.

Controlled by policies data get automatically migrated to the core file-system after a couple of minutes. For data access we offer NFSv3 and SMB exports.

5.5. ZeroMQ

Also, we are investigating a message passing approach to move data from detectors directly into GPFS. We have a working prototype implementation in place which is based on ZeroMQ [11].

When SMB (*Server Message Block*) often struggles to handle high ingest rates of some detectors, our implementation is able to cope with it and also is able to saturate a 10 GE link. Our initial model acts as a data vacuum cleaner. The detector writes data to a local ramdisk from where we pick new files and move them through multiple TCP streams to an endpoint. This endpoint then passes incoming data to GPFS.

We expect ZeroMQ to be widely used for the "first mile" of the dataflow from the detectors to the first entry into stable storage (the GPFS instance and beamline file-system in our case). Initially, the setup covers the point-to-point data link between both endpoints, ensuring the maximum and non-blocking transfer of data at speed of up to 1 Gigabyte per second (10GE link speed equivalent). The development in this area will soon be enlarged to cover more expected use-cases like: (1) fast data processing - copy data stream - process - feed data back, (2) fan out - process - fan in, to cover more (and parallized) processing chains and (3) having parallized data streams to improve storage system aggregate input data rate and to leverage multiple 10GE links.

We also want to include monitoring, online data analysis and repacking of data (i.e. HDF5/Nexus). As detector manufactures started to integrate ZeroMQ into their software as well we would even more gain from this approach.

Our long-term goal is to have a framework which enables a user to define a data flow model, for instance by modeling data sources and sinks and their cross correlation. It should only confront the user with essential building blocks necessary to model the desired data flow. The framework then should set-up and run the underlying software and network configuration accordingly.

5.6. Performance optimization

At the early stage of beging we have measured and collected various performance numbers as baseline for optimizations. Like for example NFS clients writing into the beamline file-system achieved a data rate of 60 MBytes/sec.

The process to boost the performance for the NFS and SMB paths into the system from the experimental stations where made in several session with GPFS and SMB specialists. Apart from minor changes in the SMB layer, the biggest improvement has come from parameter tuning at the GPFS layer.

This included increasing READ and WRITE sizes for clients, and increasing daemon count for NFS server. Two options having had the biggest impact are the log file size of GPFS and the file-system block size. Both settings were especially relevant for the high file creation rate the detectors are demanding.

5.7. The Core File-System

The core file-system (Figure 4) is optimized for parallel access and storage.
In contrary to the beamline file-system, it reflects the "ordered world". Full user authentication is enforced by NFSv4 ACLs so scientists need to have a valid DESY account to access their data.
Data can be accessed through NFSv4.1 [12] (Ganesha), SMB or native GPFS. In particular providing a SMB export was challenging as we need to map Windows SIDs (*Security Identifier*) to Unix GIDs (*Group ID*) and UIDs (*User ID*) and haven't had an existing service in place. To achieve the mapping we are now using the idmap backend of Samba [13] by using a customized script which queries our Linux and Windows directory services to construct a valid id-mapping.
While receiving data from the beamline file-system a single UID and GID is set for all files. Whether users are allowed to access them is handled by ACL rules. For each beamtime, metadata is produced and stored into a database. This allows to fetch correct accounts for a beamtime and setup the proper ACLs during the copy process.

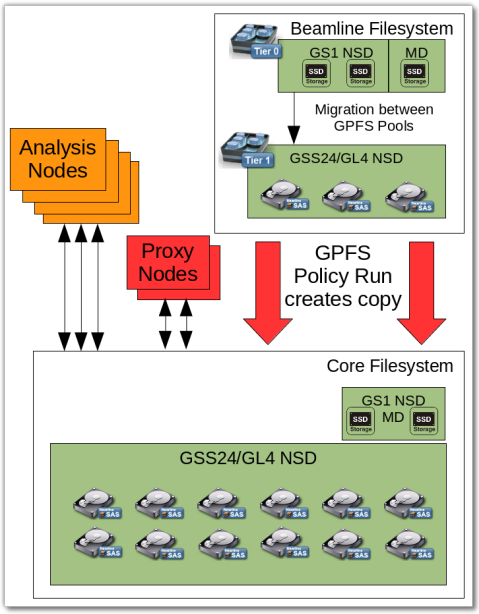

Figure 4. The Core File-System.

The metadata itself cannot be altered afterwards, but users might be added through the *Gamma-Portal* [14] to a beamtime.

To ensure data redundancy, a 3-fault-tolerant Reed-Solomon code from the underlying GPFS native RAID is used. A GPFS block is partitioned into 8 data strips and 3 parity strips, which are distributed across all physical disks.

5.8. Network Infrastructure

Nodes of the PETRA III hall are connected by 10GE ports to the Computer Center, using four dedicated 40GE switches. Inside the computing center our ASAP3 infrastructure operates on InfiniBand FDR (56Gb/s). Six proxy nodes acting as protocol gateways, each connected both to 10GE and to InfiniBand. By using InfiniBand we gain performance by its underlying low network latency which heavily improves metadata lookups.

6. Monitoring

To be able to answer questions about the health state of the system and to see what is going on, we are using different tools and services such as checks for Nagios/Icinga and RZ-Monitor (a home grown solution) for hardware monitoring.

To collect performance metrics and log files we have been developing an in-house system (Figure 5) which currently uses Elasticsearch [15] as data backend. To correlate multiple events we collect log-files, common system metrics and GPFS specific metrics and ingest them into an Elasticsearch instance. On top of Elasticsearch, we use Kibana [16] as tool to navigate through and to create customized dashboards. Our goal is to provide individual views for beamline scientists, for system operating and for extended debugging purposes.

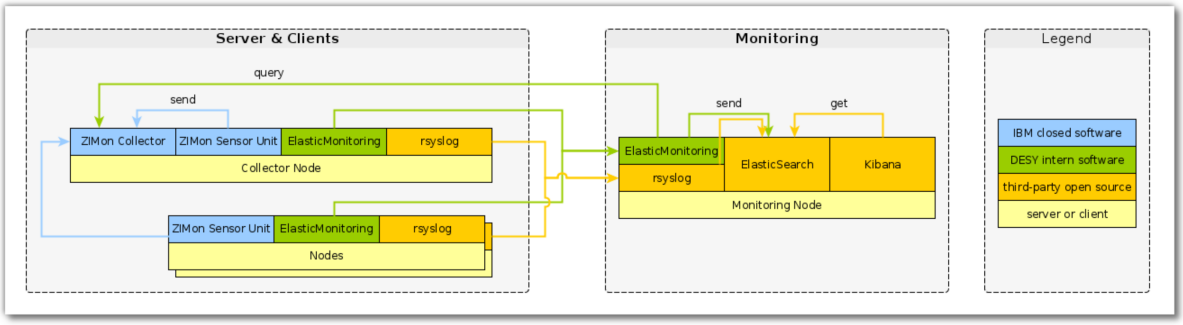

Figure 5. Architecture for collecting GPFS metrics and log-files.

7. Data Life-Cycle

Our data life-cycle is mainly steered by tags (XATTR) assigned to directories and files. These tags are set automatically by GPFS policy runs, although at any time it can be done manually as well. Thereby files get automatically published so users are able to download them from remote facilities. Also, we are using tags to handle copy-to-tape workflows integrating in our dCache [17] services. Within GPFS tags determine the dataflow across Tier-0, Tier-1, beamline- and core file-system.

8. Different Ways For Users to Access their Data

Accessing data is coupled to the type of account a user authenticates against our services. We distinguish between DESY user accounts and DOOR (*DESY Online Office for Research with Photons*, [18]) accounts.

DOOR accounts can be seen as lightweight accounts dedicated for scientists of external institutes which do not necessarily need to register for a DESY account. From a security perspective DOOR accounts are weaker, as no id-card needs to be provided by a user.

In contrast, to register for a DESY account a user need to be identified in person, e.g. by providing an id-card.

During a running experiment DOOR accounts are restricted to access data within the beamline file-system using SMB or NFSv3 exports. This enables the user to mount the beamline file-system to their workstation. Nevertheless, heavy analysis tasks are meant to be executed on dedicated workgroup server which have native access to GPFS. Although data gets migrated to the core file-system automatically DOOR accounts have no access to the core file-system.

Data of the core file-system is only accessible by valid DESY accounts. We provide exports for SMB and NFSv4.1. Workgroup and analysis server are connected by native GPFS clients. These server should be uses by the scientists, either to run analysis jobs during an experiment or after its data taking had been finished.

A third way of data access is provided by the Gamma-Portal. Supporting DOOR and DESY accounts users are able to login remotely and to download their data through HTTP. Individual files can be downloaded where simultaneous file transfers are also supported. In addition files can be bundled by the Gamma-Portal and provided as single tar-archive.

9. Summary and Outlook

Since April, 27th 2015 the ASAP3 infrastructure is running in production and providing scientists and integrated solution to handle their data life-cycle while our current architecture still offers multiple options for scaling.

Although building upon GPFS it is not just "another" GPFS instance. We have strongly customized and tuned the system to align to the scientists workflows. To give one example, we started with 60MB/sec write performance for NFSv3 (small files, single stream). Now we are reaching the MB/sec. Furthermore, our system heavily uses the integrated GPFS policy engine to automate as much workflows as possible and to reduce IOPS on file-system operations. During the project both IBM and DESY were benefiting of a close collaboration. So far, GPFS is able to handle the data rate and the GSS/ESS delivering good performance and stability.

We also have experienced the challenging combination of detectors running on Windows. Fortunately, message passing with ZeroMQ is a viable option for data transfer and a promising alternative to attach detectors to an underlying storage system while having full control over the data stream.

While PETRA III has just started their data taking again the next research facility *XFEL* [19] needs to have an initial system ready by 2016. XFEL requires up to 100PB and produces much faster data rates up to 50GB/sec. We are already elaborating possible solutions where ASAP3 can act as blueprint. Initial tests follow and hardware is already in place.

References

[1] PETRA III, http://photon-science.desy.de/facilities/petra_iii/index_eng.html
[2] DESY, German Electron Synchrotron, http://www.desy.de/index_eng.html
[3] Eiger detector, www.dectris.com/EIGER.html
[4] PCO Edge detector, http://www.pco.de/scmos-cameras/
[5] LAMBDA (*Large Area Medipix Based Detector Array*) detector, http://photon-science.desy.de/research/technical_groups/detectors/projects/lambda/index_eng.html
[6] ESRF, The European Synchrotron, http://www.esrf.eu
[7] Diamond Light Source, UKs national synchrotron science facility, http://www.diamond.ac.uk
[8] PSI, Paul Scherrer Institut, http://www.psi.ch
[9] GPFS, General Parallel File System, http://www-01.ibm.com/support/knowledgecenter/SSFKCN/gpfs_welcome.html
[10] GPFS Native RAID - Declustered Array, IBM Knowledge Center,http://www-01.ibm.com/support/knowledgecenter/SSFKCN_4.1.0/com.ibm.cluster.gpfs.v4r1.gpfs200.doc/bl1adv_introdeclustered.htm
[11] ZeroMQ, Messaging Library, http://zeromq.org
[12] NFS v4.1 Ganesha, User-Mode file server for NFS, https://github.com/nfs-ganesha/nfs-ganesha/wiki
[13] Identity Mapping (IDMAP) of Samba, https://www.samba.org/samba/docs/man/Samba-HOWTO-Collection/idmapper.html
[14] Gamma Portal, Data Management for Photon Science, https://gamma-portal.desy.de
[15] Elasticsearch, https://www.elastic.co/products/elasticsearch
[16] Kibana, https://www.elastic.co/products/kibana
[17] dCache, http://www.dcache.org
[18] DOOR, DESY Online Office for Research with Photons, https://door.desy.de/door/
[19] The European XFEL, http://www.xfel.eu/overview/in_brief/