DOI: 10.26314/GARR-Conf18-proceedings-02

A (rock and) rolling ecosystem for next gen DaaS

Ivan Andrian, George Kourousias, Iztok Gregori, Massimo Del Bianco, Roberto Passuello

Elettra Sincrotrone Trieste

Abstract. The requirements of large computing infrastructures are dictated by the applications that should be supported. Albeit this is a very general definition, that is what happened when Elettra designed and implemented its current ICT infrastructure serving the needs of beamline experiments. In practical terms, the network, computing, and storage requirements were defined by the specific needs of each beamline-laboratory of the institute and this lead to a complex ecosystem of different software, hardware, workflows and policies. This paper describes the past, present and future of this mixture of technology that for simplicity we refer to as "infrastructure". The next-generation is not only going to be defined by the institute's need but also from those proposed by a larger informal consortium of similar entities, formed by participating in multiple very large scale projects with several partners. The core and common part of these projects in relation to the infrastructure can be outlined as following: a common cloud-based ecosystem that support Data Analysis as a Service covering data acquisition & management, policies, HPC, storage, networking, remote access and all the related specific technologies. This common core sets a road-map for the ICT infrastructure of Elettra for the forthcoming years.

Keywords. Data Analysis, Distributed Storage, Big Data, CEPH, Kubernetes

1. Growing up a baby storage

The basic needs of any modern scientific experiment always include some flexible way of storing all the data produced by detectors, control devices and related instruments. This brings to consider at first the storage systems that will manage the collected data; the evolution of the particular solution now in service at Elettra is presented here by borrowing the story of a fictious human being.

1.1 Infancy

When the SOFA Storage System (Andrian et al. 2017) was born at Elettra, its parents were, of course, very proud of the newborn. At the same time, they were aware of the huge responsibility they had towards their baby, whose life, since the very first day, was expected long, brilliant and plenty of success. In fact, children are expected to grow up, fortify, become more clever and get as much notions and lessons as years go by. SOFA was no exception: weighting only 480TB (raw) at its first cry, it quickly doubled to 960TB trying hard to break the next barrier of the Petabyte.

1.2 Adolescence

Strength and resilience are two characteristics that every being should gain with time and effort. That's why the initial "factor 2" data replication policy evolved to a safer "factor 3".

This guarantees data integrity against hardware failures of high impact (like two complete cluster nodes put offline). Maturity also brought a new, more stable, release of CEPH, called Luminous, that superseded the initial simple and buggy Hammer. The upgrade of the FileStore approach of using the OSDs (relying on a basic filesystem like XFS on a disk partitioned as usual) to the more modern and efficient BlueStore (direct access to block device by CEPH) is something that requires virtually replacing every single disk in the cluster. This would have led to a lengthy and performance inefficient process. That is why it was decided that every new OSD in the cluster would be set up with BlueStore, while all the previous FileStore-defined OSD's would be converted with low priority during times when the storage cluster is not used intensively by the users.

1.3 Maturity

The Elettra 2.0 project (WWW 1) has been approved in 2017 and its financing secured for five years starting from 2018. The upgrade plan includes a new storage ring design that can be adapted to the existing Elettra storage ring tunnel, building and infrastructure. The new magnetic lattice will increase the brilliance and coherence of the X-ray beam by a factor of 20. The upgrade to Elettra 2.0 will enable new science and the development of new technologies. In particular, the science case supporting the Elettra storage ring upgrade represents a major step forward in synchrotron research, which complements and integrates the new possibilities offered by FERMI, the Free Electron Laser already in operation. As expected, this will have a direct impact on IT and will be at the same time a great challenge. The estimated data grow ratio for the next three years is currently around 4PB/year (Pugliese 2013). That's why the capacity expansion of SOFA has been designed as a rolling process: plan to add new storage nodes every year in order not to run out of storage for experimental data.

1.4 Memory

As time goes by, data will pile up in memory and its future use and need could not be easily predictable. SOFA is a disk-based, multi replicated object storage for fast read and write access. Storing huge amounts of "sleepy" data would be uneconomic to say the least. That's why a new extension of the system for "nearly online" data is under procurement. According to the new defined data retention policy, experimental data will be stored in the fast online system (CEPH) for three years. After that period it will be moved to a slower technology, still allowing users to retrieve it locally (filesystem paradigm) or even from remote sites (through the VUO interface). Tape cartridges have proved to be cheaper that disks per same data size, with the only side effects of longer data access times and the need of using different protocols and tools for the data interface. However, also these systems have evolved over time and now abstraction layers are available as "gateways" to data. This means that it is possible to create a virtual filesystem or even an object storage interface that doesn't care if the final destination is based on disks, flash memories, tapes or future technologies still to be released. The exact final solution for SOFA has not been selected yet. At the moment of this writing there are various candidates both hardware (tape libraries) and software (abstraction layer). The tape libraries currently under analysis, all based on the LTO-8 standard and with an initial raw target of 5PB meant to be extended yearly, are:

- DELL EMC2 ML3
- Quantum Scalar i6
- Spectra Logic T950(v)
- Spectra Logic Spectra Stack
- IBM (model tbd)

The software abstraction layer is of no less importance. This need itself is nothing new: in the past years the GRID community, for example, has developed a complete solution based mainly on IBM GPFS, IBM TSM (now Spectrum Pro) and a custom-made software, GEMSS (Pier Paolo Ricci et al. 2012), whose main components are presented in figure 1. Even if this solution works perfectly for the purposes and contexts it was designed for, it is too demanding for a new isolated solution like ours. IBM licenses would probably not fit in the budget themselves and not all the features of this solution are necessary for our requirements.

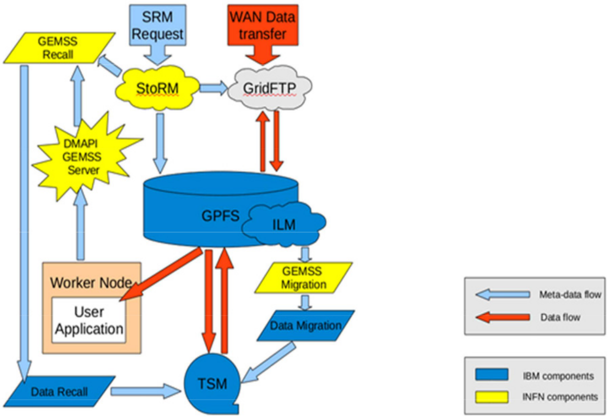


Fig. 1 Scheme of the GEMSS components

After a market analysis, which took into consideration both commercial and Open Source software, four candidates were selected:

- LizardFS – Open Source, (WWW 2) – is a distributed Filesystem that can operate both on disks and on tapes. Optional commercial subscriptions are available to secure the needed support against bugs and incidents, but are mandatory to get the development required to cope with the latest tape technologies (LTO-8, LTFS, etc.). Access to the data is possible by using a userspace mount on Linux systems (POSIX-compatible filesystem);
- OpenArchive – Open Source, (WWW 3) – is a mature and at a state-of-the-art archive solution that supports disks and tapes. Access to the data is made through protocols like NFS, CIFS and FTP;

- QStar Archive Storage Manager – Commercial, (WWW 4) – virtualises the access to storage devices of a wide spectrum (Disk Array, Object Storage, Tape Libraries, Optical Disk Libraries, WORM and Cloud), offering NFS and S3 interfaces;
- Spectra BlackPearl appliance – Commercial, (WWW 5) – which offers access only through S3 APIs and dedicated clients.

Again, the final choice hasn't been taken yet, but the most likely candidates so far are OpenArchive and Qstar ASM. The completion of this project extension to SOFA is planned to be done by first half of 2019.

2. Data Analysis As A Service (DaaS)

Data analysis is becoming a key factor for the success of modern experiments. This also relates to the huge growth in size of the dataset produced nowadays. Processing of Big Data needs capable HPC clusters and smart algorithms that can reduce data without losing information and processing time. Access to the data is also critical, since HPC requires high throughputs that can be guaranteed only inside a local area network (at Elettra, for instance, the beamline workstations connect to the cluster no faster than 10Gb/s, the in-cluster communication is now at 40Gb/s and will be upgraded to 100Gb/s). At the same time, experiments need to take place in various different facilities around the world: the data generated cannot be easily and effectively moved among laboratories for a later analyses. Remote access to a local set of data, tools, and processing is pursued by several projects Elettra is part of.

2.1 CALIPSOplus

This is the main project the activities of which are already ongoing. Its aim is to remove barriers for access to accelerator-based light sources in Europe and in the Middle East. In numbers it counts a 10M EUR budget, 19 partners, 3 associated partners and several observer members. It expects to provide more than 82,500 hours of trans-national access to 14 synchrotrons and 8 free electron lasers plus trans-national access tailor-made to SMEs. Its running time is 2017-2021 and its JRA2 work package (Demonstrator of a Photon Science Analysis Service) is what dictates the DaaS component for Elettra (WWW 6).

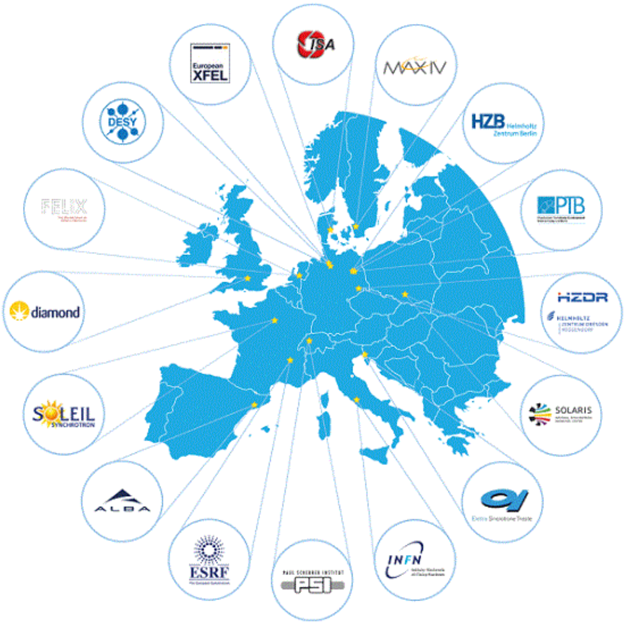
2.2 PaNOSC

This is a very large project of 20M EUR budget involving CERIC-ERIC, where Elettra is a member of. CERIC-ERIC serves as a single entry point to some of the leading national research infrastructures in eight European countries, enabling the delivery of innovative solutions to societal challenges in the fields of energy, health, food, cultural heritage and more. The PaNOSC Photon and Neutron community provides curated Open Data from a large variety of experiments and generates Petabytes of data every year. With specific focus on the European Open Science Cloud (WWW 7) the PaNOSC consortium aims at realising the EOSC concepts for finding and analysing scientific data produced from its facilities.

2.3 LEAPS

The League of European Accelerator-based Photon Sources (WWW 8) is a strategic consortium initiated by the Directors of the Synchrotron Radiation and Free Electron Laser (FEL) in Europe (figure 2). Its primary goal is to actively and constructively ensure and promote the quality and impact of the fundamental, applied and industrial research carried out at their respective facility to the greater benefit of European science and society. As expected computing is one of the main elements of the proposed strategy. The infrastructure described in this paper should be compatible for the first phase of LEAPS related projects.

Fig. 2
The LEAPS
Consortium



3. Storage and data analysis: a story of a marriage

The requirements dictated by the collaboration projects Elettra is part of, suggest that it's not possible any more to think of moving data between laboratories. It is then necessary to bring the Data Analysis tools closer to the data itself. Standardisation of data formats and programs is an ongoing process that every facility is pursuing, but that is not sufficient alone. To speed up this process it is clear that complete ready-to-use packages need to be developed, and an effective and easy way of deployment should be created. Technologies like virtualisation (KVM, ...), containerisation (LXC, Docker, ...), orchestration (Swarm, Kubernetes, ...) can now be thought as the building blocks of a new Data Analysis layer on top of existing HPC/Storage clusters. The blueprint presented in figure 3, for example, details the architecture under development in the CALIPSOplus JRA2 collaboration. When completed, a scientist will be able to connect from remote to the Data Analysis facilities of one of the collaboration members via a dedicated Access Portal, set up a computation environment by pulling ready-made containers from one of the repositories (Github-like services), link the computational nodes (containers) to the Data Catalogue and Storage and finally orchestrate this on-demand computational cluster by way of Kubernetes. An important aspect of this scenario is that it will be performed by way of a self service DaaS infrastructure, with no need of intervention of any data or computing operator/scientist.

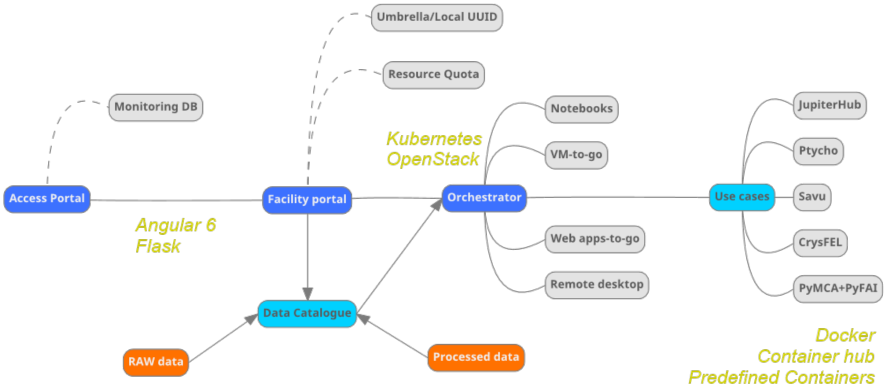


Fig. 3
CALIPSOplus JRA2
DaaS Blueprint

4. Conclusions

The ITC infrastructure may be seen as a complex ecosystem of software, hardware, workflows and policies. These are traditionally designed according to the requirements of the environment where they are going to be used for. In the case of synchrotron facilities like Elettra, they mostly depend on the needs of its beamlines and labs. There is though a special case that we report in this paper. Due to the participation in large international projects and consortiums, which are also considerable sources of funds, the future ICT must take into account their demands. Our ICT infrastructure will be influenced by the direct and indirect participation in CALIPSOplus, PaNOSC and LEAPS, among other projects of smaller size. At the same time there is an ongoing major upgrade project for the Elettra 2.0 new storage ring which will give us new IT-related challenges to face.

References

Ivan Andrian et al. 2017, Life (of Big Storage) in the Fast Lane, Proceedings of 2017 GARR Conference, pp. 117-122, DOI:10.26314/GARR-Conf17-proceedings-22

Roberto Pugliese 2013, "Relazione Tecnica Accompagnatoria Progetto DIESEL 2.0", Internal Elettra document.

Pier Paolo Ricci et al. 2012, J. Phys.: Conf. Ser. 396 042051.

WWW 1, http://www.elettra.eu/lightsources/elettra/elettra-2-0.html

WWW 2, https://lizardfs.com

WWW 3, https://www.openarchive.net

WWW 4, http://www.qstar.com/index.php/archive-manager/

WWW 5, https://spectralogic.com/products/blackpearl/

WWW 6, http://www.calipsoplus.eu

WWW 7, https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

WWW 8, https://www.leaps-initiative.eu

Authors

Ivan Andrian - ivan.andrian@elettra.eu

Ivan Andrian (degree in Information and Computer Science at the University of Udine) started his career as researcher in System Administration and Network Engineering at Elettra Sincrotrone Trieste. Over the years, where he also worked as developer of mixed web and mobile systems, he specialised in Linux distributed systems. He is currently Chair of JACoW.org, an international collaboration project for the Open Access to conference proceedings in the field of particle accelerators.

George Kourousias - george.kourousias@elettra.eu

George Kourousias (PhD) is a researcher in Elettra Sincrotrone Trieste, Scientific Computing team. His research is in X-ray microscopy, X-ray Florescence, CDI/Ptychography, and Data Science including Neural Networks. With more than 40 publications in international peer reviewed journals and a Patent in image compression, he leads AI3, an Elettra project funding, and he is co-supervising PhD students on Imaging and AI. George is in the founding team of Singularity University Trieste Chapter.

Iztok Gregori - iztok.gregori@elettra.eu

Iztok Gregori is a Linux System Administrator in the Elettra ICT Systems and Services team. He has years of experience in High Performance Computing, Virtualization Technologies and Storage infrastructures. He supervises the Elettra Scientific Storage and its computing resources.

Massimo Del Bianco - massimo.delbianco@elettra.eu

Massimo Del Bianco is a Network Engineer taking care of the network at Elettra Sincrotrone Trieste, where he participates in the ICT Systems and Services team as Network and Systems Administrator, also managing the peripheral firewall; he collaborates in managing the virtualisation and storage clusters with particular attention to high bandwith networks optimisation.

Roberto Passuello - roberto.passuello@elettra.eu

Roberto's been a network – and UNIX – System Administrator employed in research institutions since 1996. He's always been a Linux enthusiast and is currently Head of ICT Systems and Services at Elettra Sincrotrone Trieste. He spends his spare time practicing yoga, staring at the Universe and looking after his three kids.