Online data storage service strategy for the CERN computer Centre

G. Cancio, D. Duellmann, M. Lamanna, A. Pace
CERN, Geneva, Switzerland

Abstract. The Data and Storage Services group at CERN is conducting several service and software development projects to address possible scalability issues, to prepare the integration of upcoming technologies and to anticipate changing access patterns. Particular emphasis is put on:

- very high performance disk pools for analysis based on XROOTD[1]
- lower latency archive storage using large, cost and power effective disk pools
- more efficient use of tape resources by aggregation of user data collections on the tape media
- a consolidated system for monitoring and usage trend analysis

This contribution will outline the underlying storage architecture and focus on the key functional and operational advantages, which drive the development. The discussion will include a review of proof-of-concept and prototype studies and propose a plan for the integration of these components in the existing storage infrastructure at CERN.

1. Introduction

The CERN data centre has currently 30 PB of disk and 50 PB of tapes and it is growing at a rate of more than 20 PB/year driven by High Energy Physics (HEP) experiments that will push the CERN data centre within the Exabyte scale within the lifetime of the Large Hadron Collider (LHC). This growth is followed with similar trends in several other scientific disciplines where computing is one of the strategic components for their future research.

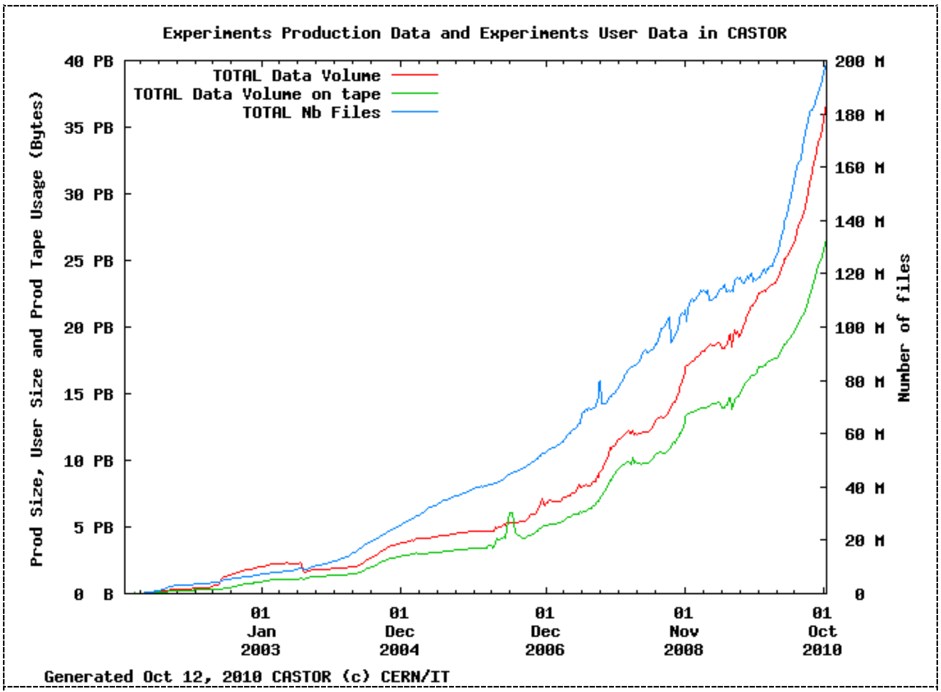

Figure 1. Storage volume evolution at CERN.

Published under licence by IOP Publishing Ltd

Today, there are several on-going research activities in the CERN IT department ensuring that we are prepared to respond to future challenges, namely building architectures aiming to deliver scalable and reliable low cost services in the areas of networking, processing and data.

The next generation data management and storage services are the main aim of this paper and current ideas and strategies will be discussed in the remaining part of this document.

2. Present solution for Data and Storage services

The current set of solutions is based on a highly scalable and low-cost architecture. All hardware is based on commodity components, assembled by industry following detailed specifications and integrated in a data management framework to deliver a comprehensive set of data and storage services to the HEP scientific community, in particular the LHC experiments.

The present strategy for the CERN data centre minimizes the cost per storage volume and it ensures an access time well below 0.5 s for any data stored on disk and below a few minutes for data archived on tape. We estimate the current architecture to be scalable to at least half a billion files without a significant drop in performance.

3. Future requirements and expected limitations of the present architecture for Data and Storage services

While sufficient scalability with a low cost approach has been achieved, the availability of new technological solutions offers the opportunity to improve other areas where requirements have only been partially met.

Below is the list of requirements that we need to satisfy in the coming years in order to support CERN's physics programme. For each requirement, we indicate its relative importance, how well it is currently met and how well this could be met in a future system.

- Scalability to Exabyte scale: This requirement is essential and it is currently met. Possible future scalability problems that are foreseen within the name service provider are currently been addressed by the next generation name server that is already in prototype phase [5].
- Security: Security is an important requirement, which is only partially met and fine-grained access control may be needed in the future. Today this is not a problem as data is typically stored in large aggregations with shared access control lists and comes from well-identified trusted sources. This may change in the future if the nature of data stored would change or would come from other sciences for which the security requirements are higher. In order to be prepared for these scenarios we are currently switching to the xroot protocol for client access that allows us to support multiple authentication methods (e.g. Kerberos, X.509 certificates, etc.) combined with the fine-grained access control (file level). In order to minimize our development effort and align with established market standards we will consider supporting the NFS 4.x or HTTP based protocol for client access, with equivalent functionalities to the xroot protocol, which is already in use in high-energy physics.
- Accountability and journaling: The accountability requirement is important and fully implemented within the monitoring and loggings subsystems. However, while every action is accounted and logged, we do not have all the tools necessary to reverse or rollback an arbitrary amount of actions (journaling). The necessity of this feature is today mitigated by the existence of the hierarchical file system allowing rollback to the previous copy of the data from tape. However, the journaling requirement will become essential when/if driven by additional cost savings we would keep only one (reliable) copy of the data without additional offline copies stored on tape.

- Global accessibility: All data in the computer centre must be accessible worldwide, from any computer having Internet access. This requirement is currently met and we see no reason why this would become a problem in the future.
- Manageability: Today we have an on-call operational team able to address the complex operation of the service, which requires engineer-level skills. We consider it important to be able to evolve towards an operation framework where the service is more resilient to failures and therefore operational interventions can be performed by less qualified personnel during working hours only without the necessity to have people on call.
- Variable availability, reliability and performance: These are new requirements that will become very important in the future. Today we provide fixed quality of service levels, which satisfy our users but we see this as a growing limitation that requires an evolution to be ready for future challenges. Today we do not have a "reliability layer (in software)" in the service and therefore both the availability and the reliability of the service is determined directly by the hardware reliability. When a hardware component fails, the service fails and we can only increase our reliability/availability by either replicating data or by buying more expensive appliances. In practice, for the vast majority of data, the reliability is directly mapped to the reliability of the storage system, which cannot be easily modulated to match the exact requirements of the data owner. Mismatching this requirement has a direct impact on cost effectiveness because with inappropriate or fixed levels of service we are forced to over provision hardware in order to deliver the defined quality of service.
- Multiple levels of service: This requirement is related to the previous one. Once variable level of availability, reliability and performance are available, then multiple levels of service at different costs become possible allowing the most appropriate offer of storage services with different price/cost tags associated.
- Decoupled from the hardware installed and from its lifecycle: Understanding what functionalities are delivered in hardware and which are delivered in software is essential to us. Today the hardware layer delivers the storage and the reliability figures. The rigidity of the hardware (typical lifetime of three years) and vendor-published reliability figures sometimes have been substantially different from those we measured, impacting the reliability of the whole service which is inherited from the limited reliability of the hardware. This is why we consider it essential to move both the "reliability and performance layers" in software, allowing the hardware to fail at arbitrary rates that are compensated by increased data redundancy provided by a small fraction of additional hardware that is entirely managed in software.
- Low cost hardware: CERN has a long tradition in buying low-cost hardware based on commodity components assembled by industry. This has been a key strategic direction that has allowed the data centre to scale and compete effectively with third party providers. For the future, if the "reliability and performance layers" are also moved in software then we can expect even further gains in the hardware layer. This is interesting because the "software layer" accounts only as a fixed cost, while the hardware layer has a cost for us that is proportional to the data centre size.
- Power efficient: Today, a large fraction of our storage has reached the ultimate power efficiency of zero power consumption when idle (tape storage). Clearly with the continuing cost decrease of hard disks, we see an increasing role of hard disks in the data centre. However, the reduced purchasing cost is quickly lost by the power consumption of the disk in the absence of a proper power management solution. This is an area where we see important possibilities of improvement and, as in the previous requirement, we think that a software solution would allow us to become power efficient at a fixed cost compared to a variable cost solution that we would have if the power management is embedded in the hardware at a cost proportional to the number of boxes.

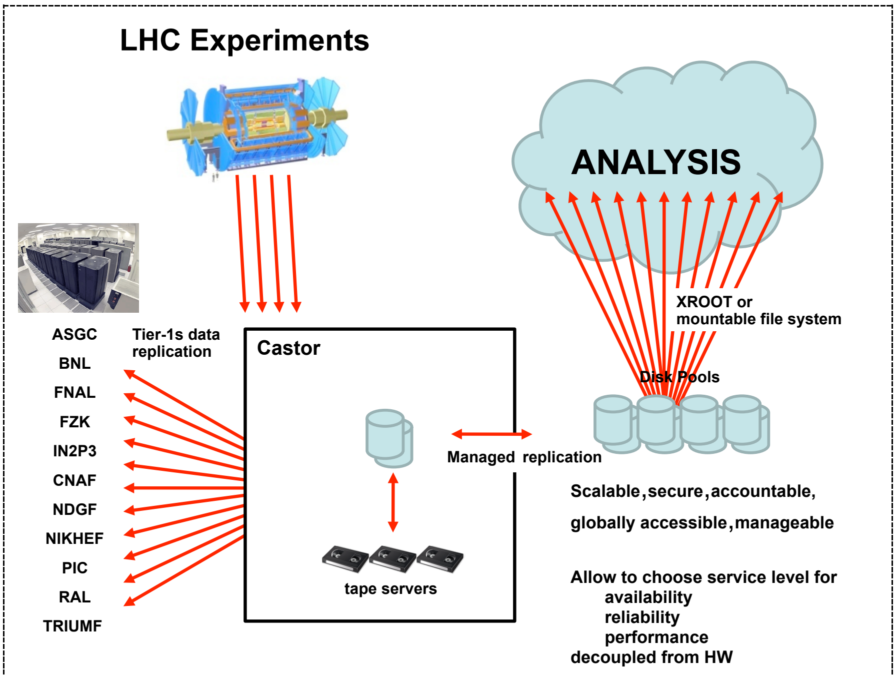

Figure 2. CASTOR archive and analysis disk pools at CERN.

4. Areas for future developments

Following the list of requirements to fulfil in the coming years, there are two areas where we are planning research and development activities.

The *first is the area of namespace management* where a high performance distributed name service provider is required to handle the catalogue and the entire metadata of one billion files at rates exceeding 5000 requests/second. These developments are considered strategic in order to be able to guarantee future scalability and the possibility to integrate even the most demanding client access protocol, which includes the possibility of seeing the entire storage as a client-mounted file system, globally available with local cache.

The *second area of development is the implementation of variable reliability*, which is linked to variable performance. The motivation for this activity is that we want to have the flexibility to better match customer requirements in terms of service availability, data reliability, access performance and cost without forcing them into a "one size fit all solution". In addition, we want to increase our internal service availability and reliability with additional redundancy so that, in case of failure, the service would continue to operate at reduced performance within the service level agreement. This will allow us to have operational interventions done asynchronously during working hours and with grouped interventions only. Today we have a prototype, which implements variable reliability and performance by configurable data replication. This approach works very well but it is inefficient in terms of storage cost because every replica requires an additional 100% of the original storage volume. This explains why we consider introducing additional coding algorithms and implementing pluggable solutions for dispersed storage using encoding techniques such as Reed-Solomon[2] error correction, or RAID-6 over the network, or double / triple parity, or LDPC[3,4] (low-density parity-check) in storage that would be scattered in multiple independent servers.

5. Current strategy

Today, we procure hardware components at minimum cost and implement a maximum of services as independent software components. The main arguments in favour of the independent software component is their modularity, for example, it is easier to modify the encoding algorithm from simple replication to dispersed LDPC storage if this does not require hardware changes.

This strategy may not necessarily be in contradiction with the all-in-hardware approach since we understand that once the three components (storage, reliability layer, power management) have been optimized, they can be consolidated and implemented on well-defined industrial standards delivering an evident gain by embedding these services within the hardware.

In order to validate this approach, we consider it essential to have the software solution being explored in parallel with a pure hardware one in order to understand if the rigidity imposed by the hardware approach is compensated by a lower cost or by increased performance at equal reliability.

A global indicator for the cost of ownership (weighting hardware costs with operational effort, power requirement and availability targets) should be defined and used during the investigation phase. Realistic workloads coming from our users communities should be used to guide the investigation.

In addition, all comparisons require a complete interoperability between the software and the hardware solution at the lowest network intra-storage level in order to be able to understand if the performance difference that the hardware may reveal is related to the efficiency of the encoding algorithm or to the additional processing power that is embedded within the storage.

In conclusion, we consider the software investigation essential, and this requires the identification of open encoding algorithms that ensure the interoperability between the "flexible" software solution based on commodity hardware and "optimized" hardware that has been customized for performance/reliability/cost. We think that a comprehensive research programme is necessary to identify the set of storage appliances parameters eventually converging into future industrial standard (hence that could be used as the generic base for the data centre storage).

[1] http://xrootd.slac.stanford.edu/papers/Scalla-Intro.pdf

[2] Plank J S, A tutorial on Reed-Solomon coding for fault-tolerance in RAID-like systems, *Software – Practice & Experience, 27(9) (1997) 995-1012*

[3] Gallager R, Low-Density Parity-Check Codes, *IRE Transactions on Information Theory 8 (1962) 1-21*

[4] Gaidioz B, Kolblitz B and Santos N, Exploring High Performance Distributed File Storage Using LDPC Codes, *Parallel Computing 33 (2007) 5-264*

[5] Peters J A and Janyst L, Exabyte Scale Storage at CERN, *These Proceedings*