Future trends in distributed infrastructures – the Nordic Tier-1 example

O. G. Smirnova[1,2]

[1] Lund University, 1, Professorsgatan, Lund, 22100, Sweden

[2] NeIC, 25, Stensberggata, Oslo, NO-0170, Norway

E-mail: oxana.smirnova@hep.lu.se

Distributed computing infrastructures have a long history, starting with pools of PCs and clusters three decades ago, briefly passing a stage of HPC Grids, and now converging to fully virtualised facilities. The Nordic Tier-1, prototype of which was launched back in 2003 to serve CERN needs, is a small-scale model of a distributed infrastructure, and thus an interesting case for studying experience and future trends in a controlled environment. The talk will present an overview of the current trends in distributed computing and data storage from the perspective of the distributed Nordic Tier-1, covering a variety of aspects, from technological to political ones.

Keywords: distributed computing, Grid, CERN, LHC

1. Introduction

The Worldwide LHC Computing Grid (WLCG [Bird, 2005]) is a global infrastructure made up of computing, storage and network resources committed to CERN by the member states. The purpose of WLCG is to provide the necessary storage capacity and computing power in order to collect and process data produced by the Large Hadron Collider (LHC [Evans, Bryant, 2008]). The sheer volume and rate of these data mean that it is currently technologically impossible to handle it at CERN, therefore already at the turn of the century CERN decided to rely on a number of data centres worldwide to store and process the data, and collectively this infrastructure became known as WLCG. In certain aspects WLCG is a precursor of modern cloud data centres, but by the virtue of being dedicated to the LHC experiments, it should be seen as a part of the LHC infrastructure, albeit external. Data centres comprising WLCG differ in size and in service levels. Twelve largest sites are designated Tier-1 centres, providing custodial storage of raw and derived data with the help of tape archives; Tier-1 centres also provide round-the-clock service and support, and are interconnected by a dedicated Optical Private Network, LHC-OPN.

The success of the LHC studies proves the validity of the WLCG concept. However, the LHC project foresees a series of upgrades, leading to ever increasing volumes and rates of produced data. At the same time, computing and data storage technologies evolve as well. In order to accommodate the ever-increasing data, WLCG will have to benefit from new technologies, and this has to be done in the situation when investment possibilities are limited.

This paper studies the case of the Nordic Tier-1 as a part of the WLCG infrastructure, and analyses implications of the new requirements stemming from the increasing LHC data quantities. The Nordic Tier-1 is a distributed infrastructure itself (see Section 2), thus it replicates many of the WLCG challenges and is therefore a valid model. Section 3 gives an overview of the most recent requirements and maps them to the operational costs of the Tier-1. Challenges and future trends are discussed in Section 4, and Section 5 presents conclusions.

2. The Nordic Tier-1 Centre

The Nordic Tier-1 centre joined WLCG operations in 2006, following three years of conceptual studies and prototype testing. The concept is quite unique: unlike other Tier-1 facilities, it is spread across four countries: Denmark, Finland, Norway and Sweden. The Nordic region is strongly interlinked and has most advanced network connectivity, supported by the NORDUnet A/S, the only cross-national research network provider. A positive experience of NORDUnet lead to the idea of establishing a cross-national computing and data center for WLCG, and the then emerging Grid technologies offered solutions for such a challenge. Since each Nordic country is rather small, it turned out to be unfeasible to concentrate all the hardware and human resources in one place. Instead, every country pledged to contribute its share to a common solution, and several national computing centres were interconnected using Grid middleware, being coordinated by a dedicated staff hosted originally by NORDUnet. As a consequence, the MoU with WLCG was signed by all four contributing countries. Since 2012, Tier-1 is a part of the Nordic e-Infrastructure Collaboration (NeIC), an initiative hosted by the Nordic organization for research cooperation and research infrastructures (NordForsk), funded through the Nordic Council of Ministers. It is important to note that NeIC supports only a few key Tier-1 positions and covers the LHC-OPN costs, while the majority of system experts and other personnel, as well as the computing and storage infrastructure itself, is funded through various national research projects. While these organizational details may seem to be not immediately related to the technology, they have rather straightforward consequences: the disparate nature of local policies and funding cycles requires a very resilient and fault-tolerant design of all the systems and operations. The Nordic Tier-1 is designed to withstand reduced capacity at any sub-site in a manner transparent to the

end-users; local data caching, quick re-allocation of resources (both hardware and human) and relocation of services are among key features of this Tier-1. A principal scheme of the Nordic Tier-1 is shown in Fig. 1: seven computing centers are equipped with ARC [Ellert, Grønager, …, 2007] middleware for computational tasks, and dCache [Ernst, Fuhrmann, …, 2001] software for storage management. ARC is deployed with large local caches, and dCache is deployed in a distributed manner, with a single entry point but distributed storage pools (disk and tape). All the necessary for WLCG authorization, accounting and monitoring services are enabled as well. Internal network relies on national research network providers, and the links to CERN and other Tier-1s – on NORDUnet.

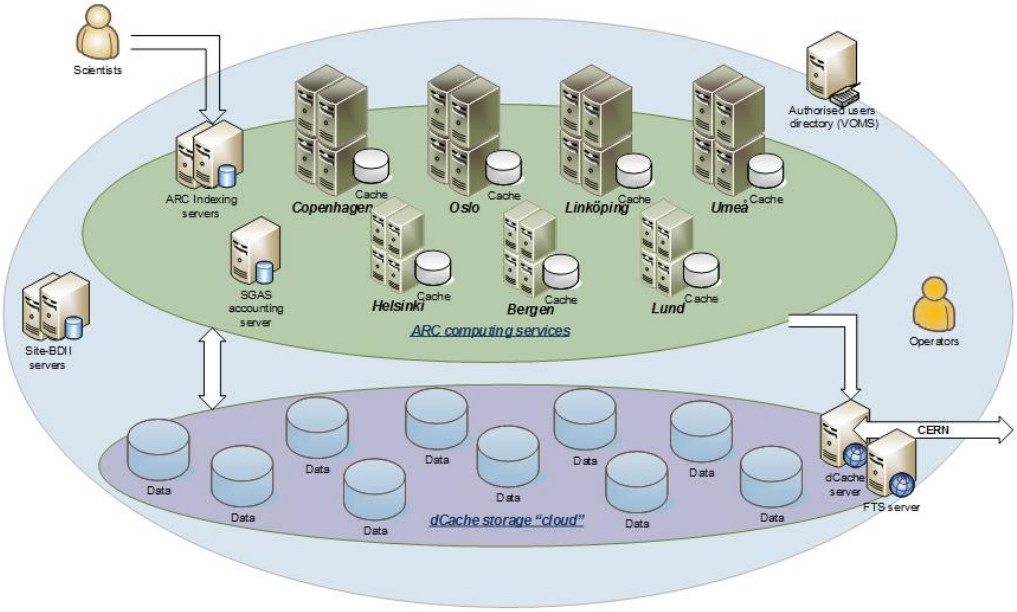

Fig. 1. A principal scheme of the Nordic Tier-1 infrastructure components

The Nordic Tier-1 supports two of the LHC experiments: ALICE and ATLAS. The targets are to provide 9% of the required ALICE Tier-1 capacity, and 6% of the ATLAS Tier-1 capacity. Internally these targets and the respectively committed resources are split between Denmark, Finland, Norway and Sweden, proportionally to the number of involved scientists in each country.

3. LHC experiments requirements and resources

The LHC experiments update their requirements regarding computing and storage resources annually, based on the LHC accelerator parameters and physics program. Originally, it has been estimated that the continuous improvements of the accelerator will lead to a ~20% annual increase in requirements. At the same time, technology improvements were expected to drive down the equipment and data processing costs at approximately the same rate, resulting in a "flat budget" model being adopted by the WLCG. Nevertheless, as can be seen in Fig. 2, the accelerator performance in 2016 improved at a much higher rate, causing a very substantial hike in the requirements.

It should be stressed here that LHC is currently designed to operate until 2037, with gradually improving parameters. Integrated luminosity is expected to increase by an order of magnitude by 2023, and by another order of magnitude by the end of the LHC program.

The better than expected performance of LHC coincides with slower than expected improvements in computing and data technologies. The Moore's law does not apply any more in the traditional sense: the original 18-month cycle became closer to 3 years, and it requires heavy code re-writing to

achieve the stated performance gain. Similarly, traditional storage technologies are approaching their limits, and the new ones are still too costly. A common trend for all the relevant components is diminishing market competition due to a decreasing number of manufacturers. Fortunately, network bandwidth keeps growing, which may allow WLCG to decouple computing and storage sites in a manner similar to the Nordic Tier-1 setup. Still, I/O and infrastructure layers need to be adjusted to make use of the high bandwidth.

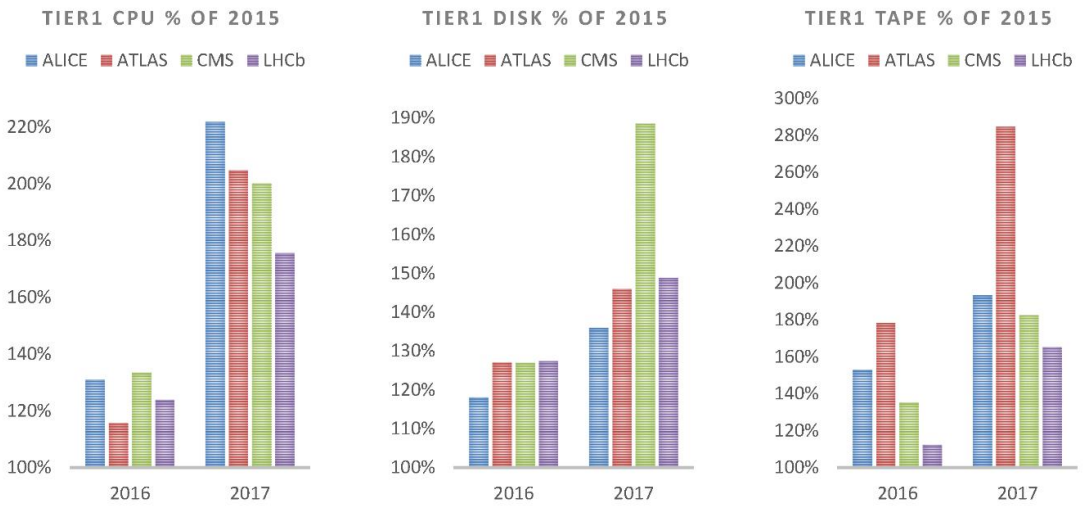

Fig. 2. Annual increase in LHC experiments' requirements as compared to 2015. The hike in 2017 well exceeds the anticipated 20% increase due to significantly improved accelerator performance.

As a result of the opposite trends outlined above, LHC experiments expect severe shortage of resources. For example, ATLAS expects to be short of computing power by almost an order of magnitude by 2023, if the existing approaches will not change [Campana, 2016]. Unavailability of prompt computing power will aggravate shortage of storage media.

It is sometimes being argued that distributed infrastructures create significate overheads, and that consolidation of resources may lead to optimization, lower costs and ultimately more resources for the LHC. The Nordic Tier-1, being a distributed infrastructure itself, conducted an internal evaluation of such consolidation scenarios [Flix, 2016]. As can be seen in Fig. 3a, the evaluation showed that the major cost component for the Nordic Tier-1 are personnel costs (around 50%), followed by the hardware and network costs.

Since there is no single Nordic site comparable to a Tier-1 in terms of size and services, a model was used by Flix to estimate single-site costs, shown in Fig. 3b. An interesting conclusion is that while some savings can be achieved, the hardware resource costs might even increase. While some overheads can indeed be reduced, various in-kind contributions offered by the centres will be lost in a single-site case. Overall, no significant gains are expected to come from resource consolidation, at least not in the Nordic countries context.

4. Challenges and opportunities

It is now clear that the "flat budget" model of the WLCG will not cover the ever-growing LHC requirements, unless computing and storage resource consumption is decreased significantly. There is no one aspect that can quickly lead to such a decrease: derived data and Monte Carlo simulation typically use up most processing power and storage capacity, and none of these can be easily reduced. All the LHC experiments are investigating ways of optimizing their algorithms. Since most LHC applications are serial, re-writing them as parallel code is an obvious option, but it is a lengthy process, requiring, among others, making use of vector registers, instruction pipelining, multiple instructions per cycle, improving data and code locality, using hardware threading and reducing memory consumption per core, and eventual gains are difficult to assess

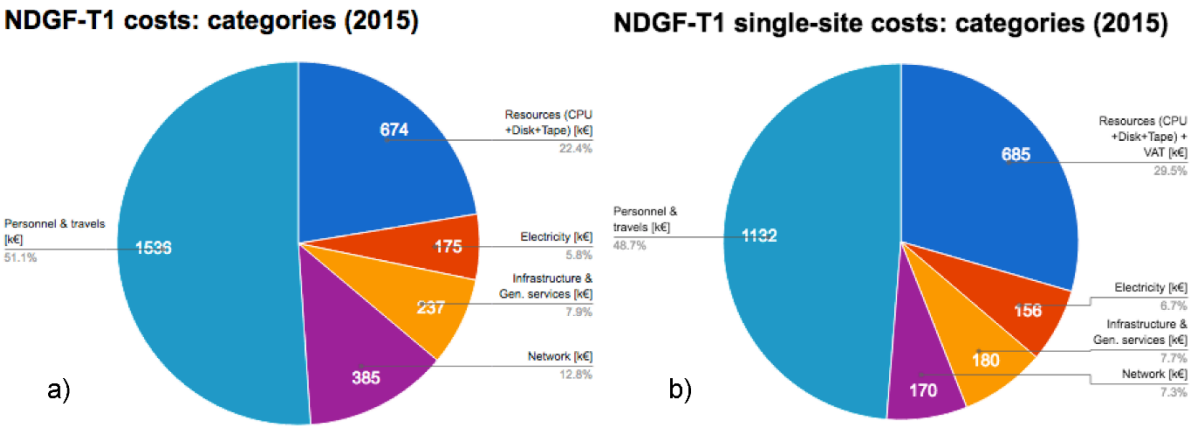

Fig. 3. Cost categories of the Nordic Tier-1, figures taken from the evaluation report [Flix, 2016]. The left panel (a) shows the current situation, while the right panel (b) shows a single-site model.

There is currently a growing interest in using "free" opportunistic resources, such as e.g. research allocations at supercomputer centres and private clouds or volunteer computing. Such resources so far can accommodate only few selected workflows of the LHC experiments, but this can be improved in several ways. By enhancing functionality of Compute Elements one can make more efficient use of traditional batch systems, expand WLCG to disk-less sites, and reduce the number of small sites by pooling their resources in a volunteer-computing style.

Certain gains can be made in terms of data storage, and a number of optimization approaches are being implemented, such as limiting lifetime of low-priority data sets, reducing numbers of disk copies in favor of increasing use of tape storage, enabling storage federations and data streaming. The roles of different centres in the LHC data hierarchy are being re-considered, consolidating functions when possible. Much can be learned from Big Data technologies, and several conceptual approaches making use of object storage are being developed, along with increased use of analytics and deep learning techniques.

Despite the ever-increasing use of non-conventional resources in WLCG, close to 90% of LHC computing still relies on Grid technologies. Moreover, most opportunistic resources are accessed via Grid interfaces. Here, one can point out growing popularity of ARC, which now powers approximately 20% of the WLCG resources. One of the reasons of its popularity is that it was designed for time-shared HPC resources, and interfaces not just to Grid and HPC facilities, but also to for volunteer computing resources via BOINC. It supports all relevant batch systems and storage services, including Amazon's S3. ARC's built-in caching is a convenient solution for disk-less sites, and accompanies distributed dCache infrastructures. A distributed dCache itself is also gaining popularity as a way to implement storage federations and transparently consolidate storage resources. The approaches pioneered by the Nordic Tier-1 are being replicated in more and more distributed sites.

5. Conclusions

The Worldwide LHC Computing Grid is a dynamic distributed infrastructure, successfully facilitating provision of storage and computing services to the LHC experiments. The sharply increasing resource demands caused by an outstanding performance of the accelerator lead to new challenges, a part of which must be addressed by the experiments, and a part by the WLCG. The infrastructure exists now for more than a decade, and underwent several incremental changes during its life time. The original strictly hierarchical multi-tiered model was based on the assumption of limited bandwidth; today the

limitations are elsewhere, and the hierarchy is being replaced by a mesh-like structure aimed to optimize processing times. The "jobs to data" paradigm gives way to data streaming and caching, and data federations make even more use of the network bandwidth. Virtual computing centres using cloud resources become a reality, and while cloud cycles may come with a price tag, savings in personnel costs can be more significant. Indeed, an evaluation of the Nordic Tier-1 showed that hardware resources constitute just 22% of operational costs of the centre. Therefore, while some consolidation of resources in large centres or private clouds is inevitable, the WLCG will remain a distributed infrastructure in the foreseen future, keeping Grid services as interfaces to various resource types. While computing-related challenges can be addressed by improved application algorithms and by opportunistic usage of non-dedicated facilities, storage requires dedicated long-term effort, and optimization possibilities are rather limited. A full transition to object-based storage is not currently foreseen, but several test studies are being conducted in that technology. The overarching goal for distributed computing infrastructures, like WLCG or the Nordic Tier-1, is to meet the user requirements while optimizing resource consumption and reducing operational costs.

References

Bird I. LHC Computing Grid Technical Design Report // CERN-LHCC-2005-024. 2005.

Evans L., Bryant P. LHC Machine // Journal of Instrumentation. — 2008. — Vol. 3, No. 08. — P. S08001.

Ellert M., Grønager M., Konstantinov A., Kónya B., Lindemann J., Livenson I., Nielsen J. L., Niinimäki M.. Smirnova O., Wäänänen A. Advanced Resource Connector middleware for lightweight computational Grids // Future Gener. Comput. Syst. 2007. Vol. 23, No. 2. P. 219-240.

Ernst M., Fuhrmann P., Gasthuber M., Mkrtchyan T., Waldman C. dCache, a distributed storage data caching system // In Proceedings of CHEP 2001: international conference on computing in high energy and nuclear physics. — 2001. — Vol. 34, No. 48 — P. 241-244.

Campana S. The ATLAS computing challenge for HL-LHC // To appear in Proceedings of CHEP 2016: international conference on computing in high energy and nuclear physics. — 2016.

Flix J. Investigating options for future Nordic WLCG Tier-1 operations. — 2016. URL: https://neic.nordforsk.org/2016/07/05/nordic-model.html.