A new data-centre for the LHCb experiment

Abstract. The upgraded LHCb experiment, which is supposed to go into operation in 2018/19 will require a massive increase in its computer facilities. A new 2 MW data-centre is planned at the LHCb site. Apart from the obvious requirement of minimizing the cost, the data-centre has to tie in well with the needs of online processing, while at the same time staying open for future and offline use. We present our design and the evaluation process for different site options.

Loïc Brarda, Beat Jost, Daniel Lacarrère, Rolf Lindner, Niko Neufeld, Laurent Roy, Eric Thomas

1. Introduction

LHCb is planning a major upgrade of the detector and in particular the read-out system [1]. The current LHCb trigger and data acquisition use, like all LHC experiments, a high p_t hardware trigger based on calorimeter and muon-detector information to select a number of bunch-crossings for further processing in a farm of industry standard servers. In LHCb this so-called "L0-trigger" selects 1 million events out of the 40 million crossings per second. This mechanism comes at cost for the physics program. For B-hadron decays into other hadrons more than half of the interesting events are lost. Moreover the current detector can only accept a certain number of interactions per second ("luminosity"). To maximize the physics reach after 2017 the detector will be upgraded and the limitation from the "L0-trigger" will be lifted. The resulting readout-system allows to read out the *entire* detector at bunch-crossing rate. Moreover the detector shall be capable to sustain a luminosity of up to $2 \times 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$.

2. The 40 MHz Data Acquisition

For the data acquisition system this "trigger-free" read-out system has important consequences, in particular it makes it much larger. The key parameters are summarized in Table 1. Two

# of input links	10000
DAQ bandwidth per input link	3.2Gbit/s
average total event-size	100 kB
total bandwidth for the DAQ	32 Tbit/s
output bandwidth	2 Gigabyte/s

Table 1. Key parameters for the LHCb 40 MHz DAQ

.

important facts should be noted:

(i) the data produced by a bunch-crossing need to be "zero-suppressed" directly on the detector to reduce the number of input links from the detector.

(ii) it is planned to stage the deployment of the data acquisition and event-filter farm in three equal parts. The numbers presented in this paper are always for the full system.

The input links from the detector are custom technology made to withstand the radiation and magnetic field around the detector. The data from these links need to be re-packaged to be transmitted over a standard local area network technology.

Two such technologies are actively studied: Ethernet with speed-grades of 10, 40 and 100 Gbit/s and InfiniBand in its QDR (Quad Data Rate, equivalent 32 Gbit/s) and FDR (fourteen data rate, equivalent 52 Gbit/s) versions. The most important cost-driver of the DAQ is the large network, which connects the data-sources to the compute units of the event-filter farm. It is in this farm where the interesting event data for physics will be selected. The events are independent so they must simply be distributed in load-balancing way to all of the compute units.

3. Requirements on the farm

The event-filter farm must be capable of processing 40 million events[1] per second, out of which 20000 are retained for permanent storage.

The selection of the events requires the unpacking of the data from the detector, their combination with calibration and alignment information about the detector elements and the progressive reconstruction of higher-level physical quantities from the raw data[2].

It is well possible that the final compute unit, which is the destination of the events will consist of a combination of a industry-standard server and a co-processor card such as a GPU. If anything such a system will be more compact and overall cheaper than a system purely based on industry-standard servers, because otherwise it could not offset the additional complexity it introduces in the software architecture. This is why for the purpose of this paper we will assume the "worst" case of an event-filter farm composed exclusively of industry standard servers.

Currently about the equivalent of 1200 servers [3] are needed to process 1 million events per second. This obviously depends strongly on the algorithms, which are run and on the acceptable output rate. Nevertheless, taking the current algorithm as a starting point and assuming that 40 times more events will need to be processed during the upgrade, one would end up with about 5000 servers. Most of the cores today are of the Intel Westmere type. Until 2017 there should be at least 4 Tik-Tok steps which should "from past experience" reduce the number of servers by roughly a factor 10. The remaining factor four must come from an increase of the farm in size. Obviously there are a lot of (possibly wrong) assumptions here. To name just two, the higher luminosity will lead to events with a higher average number of hits. Some of the algorithms have a quadratic or even worse dependence on the number of hits. This should increase the time needed for a decision, on the other hand, many of the events which were previously rejected in hardware, will have a topology which makes it very fast to reject them in software as well and so this should save overall computational resources.

Until better numbers are available the planning will thus be based on 5000 servers for the event-filter farm.

By a "server" we understand a dual-socket machine, which provides at least 1 GB of RAM per hardware-thread[4]. It is expected that this will stay the most cost-effective platform for our type of application and that such a server will need at most one half of a rack-unit. In the following section we will derive from this the requirements for the farm as a whole.

[1] It should be noted that only 30 million out of these 40 contain interesting physics data.

[2] The raw data normally give only information about a charge deposit and/or arrival time at or from a specific location of the detector

[3] Equivalent in performance to a dual-socket system using the Intel 56xx technology.

[4] By hardware-thread we mean any processing unit which the operating system scheduler can treat as independent.

max # of servers	5000
max # of useful Us for servers	2500
number of Us for switches	2 Us for 36 servers
Us for patch-panels per rack	3
depth of the racks	min 1000 mm
number of power feeds per rack	2
total usable IT power	2 MW
minimum power per server	350 W

Table 2. Main parameters for the data-center for the LHCb Event-Filter-Farm (EFF)

4. Requirements on the data-centre

The new data-centre has the requirements listed in Table 2. Other requirements are a low PUE[5] value, a potential usefulness of the centre after the expected life-time of LHCb (10 years) and that the initial capital expenses are compatible with the overall budget of the LHCb upgrade (about 50 MCHF).

4.1. Non-requirements

In this section we discuss a few simplifications which our application allows us to make and which we use to cut costs.

LHCb has very good experience with running the servers with the root filesystem on an NFS server. The local disk is only used as a scratch space and for a swap partition. The disk-activity is low, so the risk of damage to the installation due to a power-failure is very low [6]. The detector itself is not on safe power either, such that a power-outage terminates the data-flow immediately. Finally the event processing time is very short, so the loss in the event of a cut is minimal. All these facts together allow operating the bulk of the compute power without UPS backup and this will hence not be required for the new farm.

Very little central storage is required, amounting to at most half a petabyte. This can be easily hosted in a single dedicated rack and the majority of the racks can thus be optimized for high-power compute density, saving floor-space and helping with cooling efficiency.

Since the data-centre is designed to our specifications, we can dimension the wattage per sever such that it allows to maximize the CPU performance per unit of money. Given the relatively short life-time of the facility and the rather low duty-cycle, capital expenses should be optimized before operational expenses.

5. Implementation options

The current LHCb event-filter farm is housed in the UX85A area at Point 8. This site is 100 m under-ground in an area accessible only for personnel under radiological supervision and hence off limits to the technicians of contractors. Moreover it re-uses rack infrastructure originally conceived for Fastbus electronics, which had to be adapted to house servers. There is a limitation in cooling and electrical power of ~ 500 kW. In particular floor-space and cooling capacity cannot be easily extended underground. All these reasons make it both necessary and very desirable to build a completely new data-center for the upgrade.

In addition as the network will be considerably larger, additional space for the network will most likely be needed in the underground area. Currently the network occupies nine racks.

[5] Power Usage Efficiency, defined as the ratio of total power provided to the data-center to the power used by the IT equipment.

[6] None has been observed in four years of operation, during which there were numerous power-cuts.

In the following we will discuss the various options to build a data-center according to the rough specifications given above.

5.1. Brick and mortar
It is commonly said that a brick and mortar data-centre costs around 10 million USD per MW of IT capacity, for example see [2], where it should be kept in mind that these figures normally assume full battery backup, fly-wheels and similar. Since our overall budget for this project is well below 10 MCHF it can still be justifiably asked, how we can even consider such an option.

There are several factors in our favor. The land is for free, and we can use a site which is in principle fully technically equipped both with high-power feeds and a large cooling plant. The site is industrial, close to the airport, and apart from complying with standard environmental protection and safety codes no additional constraints on the building are foreseen. The facility is single purpose and security other than ordinary theft protection is not a concern as no confidential data are treated or stored.

Our original idea was to colocate the data-center with a planned new building for a control-room and offices. For the cooling we decided to use heat-exchangers mounted at the back of the racks to cool the exhaust-air[7]. This concept has been developed and is currently used by all LHC experiments for their data-centers. Compared to room air-conditioning and using Computer Room Air Conditioning (CRAC) units this was considered a progressive design 10 years ago. For the current LHCb data-centre which is installed underground any cooling system relying on the exchange of air would have been very impractical anyhow. Measurements at the time have shown that about 90 to 95% of the exhaust heat of the servers will be absorbed by the water, the rest has to be absorbed by conventional air-conditioning.

Moreover such a system is attractive at LHCb's experimental site, because electrical and cooling power are available from much larger facilities made for the Large Hadron Collider itself. Unfortunately, while two MW of additional electrical power can be added at a moderate cost (the price of the transformers and the necessary cabling), the additional cooling power requires a new refrigeration plant.

The PUE of the current water-cooled solution has been estimated by the CERN Cooling and Ventilation group to be about 1.3.

5.2. Remote hosting
Remote hosting is a logical consequence of cheap network bandwidth and the strive to use the most cost effective services, wherever they are available. For LHCb remote hosting could mean two things:

(i) Using a data-centre somewhere off the CERN sites
(ii) Using an existing data-centre at CERN, specifically the IT data-centre in building 513 on the Meyrin site.

In both scenarios it is assumed that only the infrastructure but not the servers or the network equipment are rented.

5.2.1. Remote hosting off the CERN site This option is interesting also because it is what the CERN IT department has adopted as its baseline for the future. In this scenario the costs would therefore be the rental cost for the data-center space, the cost for electricity and the rental-cost of the data-path from the LHCb-site to the remote hosting site.

[7] The water temperature in these exchangers is 13 degrees Celsius

5.2.2. Data-transport using a small number of fibers In both options of remote hosting the data will need to be transported over a very small number of fibers. For off-site hosting the reason for this is obvious, for CERN-site hosting the reasons will be discussed in the next section.

The current high-speed, long-distance standards (like LR4) transmit 25 Gbit/s on a single color. Without multiplexing this would mean that for the 32 Tbit/s of LHCb 1280 fibre-pairs are needed. Using multiple wave-lengths ("colors") on the same fibre avoids this problem. 400 colors and more can be achieved today in practice[8], bringing down the number of required fibers to an acceptable level. We have conducted a study with one of the major suppliers of such solutions. The details are under Non-Disclosure-Agreement NDA, however the equivalent cost per 10 Gbit/s, even assuming an aggressive price compression until 2017 is estimated to be around 7000 USD (in 2011 equivalent currency).

Another option would be the use of colored interfaces directly in output ports of the LHCb DAQ network equipment and then use passive optical multiplexing equipment to put these wavelengths on a fibre. However the colored interfaces are quite expensive compared to the standard ones (up to a factor three), long-distance (single-mode) interfaces are required and such a passive infrastructure offers practically no monitoring. It seems unrealistic to operate such an infrastructure over ten years without trained personnel and the cost is still prohibitively high.

5.2.3. Remote hosting on the CERN site All remote hosting solutions outside the CERN-site incur operational costs for the rent of the facility and the rent of the data-path. These costs could be avoided if a suitable facility could be found on one of the CERN sites (Prevessin or Meyrin). We have investigated this scenario under the assumption that a suitable fraction of the computing centre in building 513 were at our disposal. This is the only facility at CERN, which has already today the necessary basic infrastructure.

Again the value of the solution hinges on the cost of transporting the data. As has been discussed in section 5.2.2 the cost of multiplexing technology is thought to be prohibitive on the time-scale for this project. Since the distance is relatively short (about 3 to 4 km depending on the exact path), in principle it can be thought of installing the required fibers. We have done a study assuming the worst-case, in which a single fibre-pair carries only 10 Gbit/s. This amounts to 8000 fibers, including spares. The civil engineering work for laying such an amount of fibers has been estimated to be more than 10 MCHF. In addition there are serious technical difficulties in routing such a huge amount of fibers on the Meyrin site, which has a long legacy of installations and in bringing them into an existing building with an enormous number of existing connections. For all these reasons this solution must be discarded.

Building a new data-centre somewhere else on a CERN-site other than Point 8 itself is evidently subject to a similar cost for the installation, which as shall be seen, is more than estimated for a on-site container solution.

6. Modular data-centre - containerized data-centre

Comparing various cooling solutions it became quickly clear that to achieve a very low PUE, the PUE must be the primary criterion for any aspect of the design, from power-density to rack-arrangement. For example free cooling is most easily achieved with a single rack-row, or at maximum two. This leads to a rather long, narrow floor-plan, which is far from ideal for the simultaneous use as an office-building as foreseen in the original LHCb plan. Also, there is a policy at CERN to avoid mixing office and utility use of buildings. Finally the re-use of a facility at Point 8 after the completion of LHCb is doubtful.

[8] Much more has already been demonstrated in laboratory setups.

For all these reasons we started to look for a separate, more modular solution for the data-centre. Studying reports such as [2] convinced us that a "container"-based data-centre is ideal. These offer very efficient integrated cooling solutions and can easily be deployed and after use re-sold or moved. We have done detailed studies with two major suppliers to convince ourselves of the feasibility of a centre to our specifications within our budget. Without preempting the mandatory tender process, it is clear that affordable solutions do exist.

One obvious advantage is that efficient free cooling is standard in these products. Following recommendations in [3] extra cooling power will only be required during a very small number of hours of a year in the Geneva area (estimated around 5%). ASHRAE has released revised recommendations for cooling facilities in data-centers, which will allow to run at up to 40 degrees C [4]. This will completely eliminate the need for any additional cooling[9].

An important requirement for us will be that the containers are supplier-neutral. Since the money for servers will come in several installments, following the needs of the experiment, it seems unwise and impractical to lock into a specific server supplier.

7. Conclusion

The upgraded LHCb experiment will require a massive amount of computing power, which cannot be hosted in the existing infrastructure. Considering the rather modest budget for the overall upgrade of the experiment, it is crucial to find the most cost-effective solution for housing this computing power. We have studied several options: co-location with an office-building, various remote hosting options and finally a solution based on a container data-center. This last option fits our requirements particularly well, promises the lowest cost and a potentially very good energy efficiency. Moreover it offers the best investment protection for CERN. The results of this study will be the basis for a call for tender in 2017.

References

[1] LHCb Collaboration 2011 Letter of Intent (LoI) for the LHCb Upgrade [Online]. Available: http://cdsweb.cern.ch/record/1333091?ln=en

[2] Bramfitt M and Coles H 2011 Modular/Container Data Centers Procurement Guide: Optimizing for Energy Efficiency and Quick Deployment [Online]. Available: http://hightech.lbl.gov/documents/data_centers/modular-dc-procurement-guide.pdf

[3] ASHRAE TC 9.9 2008 Thermal Guidelines for Data Processing Environments [Online]. Available: http://www.ashrae.org

[4] the green grid 2012 Updated Air-Side Free Cooling Maps: The Impact of ASHRAE 2011 Allowable Ranges [Online]. Available http://www.thegreengrid.org//~media/WhitePapers/WP46UpdatedAirsideFreeCoolingMapsTheImpactofASHRAE2011Allow

[9] A small caveat is that not all major server manufacturers have yet validated their entire product line for this kind of environment, however there is strong push in the market to do so.