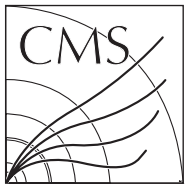
Available on CMS information server

CMS CR -2015/074



The Compact Muon Solenoid Experiment

Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

15 May 2015 (v2)

Online data handling and storage at the CMS experiment

Jean-Marc Andre[5)], Anastasios Andronidis[2)], Ulf Behrens[1)], James Branson[4)], Olivier Chaze(2), Sergio Cittolin[4)], Georgiana-Lavinia Darlea[6)], Christian Deldicque[2)], Zeynep Demiragli[6)], Marc Dobson[2)], Aymeric Dupont[2)], Samim Erhan[3)], Dominique Gigi[2)], Frank Glege[2)], Guillelmo Gomez-Ceballos[6)], Jeroen Hegeman[2)], Andre Holzner[4)], Raúl Jimenez-Estupiñán[2)], Lorenzo Masetti[2)], Frans Meijers[2)], Emilio Meschi[2)], Remigius K. Mommsen[5)], Srecko Morovic[2)], Carlos Nunez-Barranco-Fernandez[2)], Vivian O'Dell[5)], Luciano Orsini[2)], Christoph Paus[6)], Andrea Petrucci[2)], Marco Pieri[4)], Attila Racz[2)], Penelope Roberts[2)], Hannes Sakulin[2)], Christoph Schwick[2)], Benjamin Stieger[2)], Konstanty Sumorok[6)], Jan Veverka[6)], Salvatore Zaza[2)], Petr Zejdl[5)]

Abstract

During the LHC Long Shutdown 1, the CMS Data Acquisition (DAQ) system underwent a partial redesign to replace obsolete network equipment, use more homogeneous switching technologies, and support new detector back-end electronics. The software and hardware infrastructure to provide input, execute the High Level Trigger (HLT) algorithms and deal with output data transport and storage has also been redesigned to be completely file- based. All the metadata needed for bookkeeping are stored in files as well, in the form of small 'documents' using the JSON encoding. The Storage and Transfer System (STS) is responsible for aggregating these files produced by the HLT, storing them temporarily and transferring them to the T0 facility at CERN for subsequent offline processing. The STS merger service aggregates the output files from the HLT from 62 sources produced with an aggregate rate of 2GB/s. An estimated bandwidth of 7GB/s in concurrent read/write mode is needed. Furthermore, the STS has to be able to store several days of continuous running, so an estimated of 250TB of total usable disk space is required. In this article we present the various technological and implementation choices of the three components of the STS: the distributed file system, the merger service and the transfer system.

Presented at *CHEP2015 21st International Conference on Computing in High Energy and Nuclear Physics*

[1)] DESY, Hamburg, Germany

[2)] CERN, Geneva, Switzerland

[3)] University of California, Los Angeles, Los Angeles, California, USA

[4)] University of California, San Diego, San Diego, California, USA

[5)] FNAL, Chicago, Illinois, USA

[6)] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Online data handling and storage at the CMS experiment

J-M Andre[5], A Andronidis[2], U Behrens[1], J Branson[4], O Chaze[2], S Cittolin[4], G-L Darlea[6], C Deldicque[2], Z Demiragli[6], M Dobson[2], A Dupont[2], S Erhan[3], D Gigi[2], F Glege[2], G Gómez-Ceballos[6], J Hegeman[2], A Holzner[4], R Jimenez-Estupiñán[2], L Masetti[2], F Meijers[2], E Meschi[2], R K Mommsen[5], S Morovic[2], C Nuñez-Barranco-Fernández[2], V O'Dell[5], L Orsini[2], C Paus[6], A Petrucci[2], M Pieri[4], A Racz[2], P Roberts[2], H Sakulin[2], C Schwick[2], B Stieger[2], K Sumorok[6], J Veverka[6], S Zaza[2] and P Zejdl[5]

[1] DESY, Hamburg, Germany
[2] CERN, Geneva, Switzerland
[3] University of California, Los Angeles, California, USA
[4] University of California, San Diego, California, USA
[5] FNAL, Chicago, Illinois, USA
[6] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

E-mail: lavinia.darlea@cern.ch

Abstract. During the LHC Long Shutdown 1, the CMS Data Acquisition (DAQ) system underwent a partial redesign to replace obsolete network equipment, use more homogeneous switching technologies, and support new detector back-end electronics. The software and hardware infrastructure to provide input, execute the High Level Trigger (HLT) algorithms and deal with output data transport and storage has also been redesigned to be completely file-based. All the metadata needed for bookkeeping are stored in files as well, in the form of small documents using the JSON encoding. The Storage and Transfer System (STS) is responsible for aggregating these files produced by the HLT, storing them temporarily and transferring them to the T0 facility at CERN for subsequent offline processing. The STS merger service aggregates the output files from the HLT from ∼62 sources produced with an aggregate rate of ∼2GB/s. An estimated bandwidth of 7GB/s in concurrent read/write mode is needed. Furthermore, the STS has to be able to store several days of continuous running, so an estimated of 250TB of total usable disk space is required. In this article we present the various technological and implementation choices of the three components of the STS: the distributed file system, the merger service and the transfer system.

1. Introduction

The DAQ (Data AcQuision) system at CMS has been upgraded (DAQ2) during the 2013–2014 long shutdown to an innovative file–based concept [1], which has been successfully implemented and tested during the last part of 2014 and beginning of 2015. Figure 1 depicts the last part of the DAQ2 chain. The full event building in DAQ2 is performed on Builder Units (BU) which are forwarding the event data to Filter Units (FU) nodes running the High Level Trigger software. BU and FU nodes are connected using 40 Gbit Ethernet link on the BU side and either 10 Gbit

(new generation of FU nodes) or 1 GBit (legacy HLT nodes) on the FU side. The builder unit and corresponding FU nodes form an HLT appliance. There are 62 appliances in the DAQ2 system, consisting of between 12 and 18 FU nodes each. The events selected by the HLT are forwarded to the Storage and Transfer System (STS). To establish data transfer of events built on the BU to the HLT running on the FU and from the HLT output to the STS, a file–based approach is used. FU nodes use NFS v4 to mount a ramdisk partition on a BU which serves as a large (240 GB) buffer of fully built events (raw data) for the HLT input files and a RAID-0 configuration of 4 spinning disks hosted by the BU for the HLT output files.

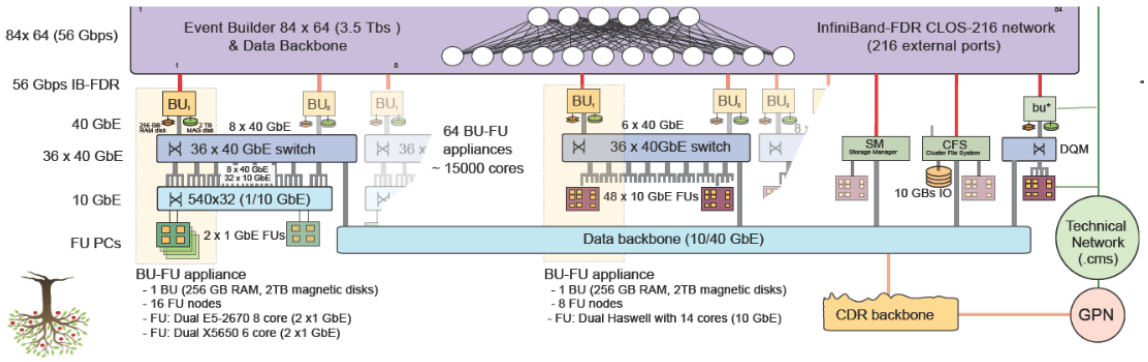


Figure 1. Storage and Transfer System in the DAQ chain.

In CMS, a Luminosity Section (LS) is defined as a quantum of data taking controlled by the Trigger Control and Distribution System (TCDS) and used later for the accounting of effective integrated luminosity. An LS is defined as a fixed time span lasting a predefined number of LHC orbits and treated as a unit. The LS is currently set to 23s. HLT processes executing on FU nodes output their data at the end of a LS in multiple output streams, also providing accounting of processed events and events selected by the HLT in JSON metadata files. In addition to the main data stream for physics analysis (corresponding to roughly half of data volume), there are streams for detector monitoring and calibration, event display and online Data Quality Monitoring (DQM). There are several merging steps performed in the Filter Farm and the STS, where data and metadata are aggregated and check-pointed at the end of each LS. This article discussed the STS. Its role is to aggregate the data from the HLT, provide storage as temporary buffer and transfer the data to Tier-0 at CERN, as well as to local online clients.

2. Storage and Transfer System Role and Requirements

The STS consists of three main components which are strongly correlated and interact in order to achieve the overall purpose of the STS.

2.1. The merger system

The role of the merger system is to aggregate the output of the filter units as to obtain one data file per LS per stream. It functions on two different levels:

- mini–merger: aggregate the selected events coming from the FUs at the BU level, such as to obtain 1 data file per BU/FU appliance per LS per stream
- macro–merger: collect and merge all the outputs of the mini–mergers such as to obtain 1 file per LS per stream

The merger system needs to provide meta–data files for each data file that is produced. The meta–data is used by the online monitoring system, described in [2], but also to check data consistency throughout the several stages of the online data processing.

The merger system needs to provide various specialisations of "affregation" methods. Even if most of the data streams "aggregation" translates to "concatenate", there are special data files types that need special tools for their merging – typically the DQM (Data Quality Monitoring) streams will contain histograms, that need to be merged using dedicated functions. The handling of the meta–data files are another special case that the mergers need to implement in order to provide bookkeeping throughout the life cycle of the data.

A requirement for the merger system concerns the latency: a maximum delay of 2LS (46s) between the time when the FUs have delivered their selected events and the time when the macro–merger has completed its task is considered acceptable.

2.2. The transfer system

Once the macro–merger has aggregated all the data into the required format, these data files need to be transferred to various locations for further offline processing. Typically, the data destinations can be:

- Tier0 at CERN: all the physics streams, as well as most of the sub–detectors data need to be transferred to Tier0, from where it can be picked up for offline reconstruction
- dedicated sub–systems areas: special sub–systems, such as DQM, EventDisplay and prompt calibration, need to process their final output online, so they provide dedicated areas where the transfer system has to move the respective files after their macro–merging is complete
- local: in specific cases the data can be temporarily stored locally for debugging purposes

The main requirement for the transfer system is to send the data to Tier0 at a speed of 1GB/s.

2.3. The storage system

In order for the merger and transfer systems to perform their duties an appropriate infrastructure is needed. The proposed solution is a distributed file system that is visible to the BU nodes and is exposed to Tier0 via a dedicated link. The storage system serves as output layer for the mini–mergers, input and output for the macro–merger and input for the transfer system. It needs to provide the aggregated bandwidth for all these operations:

- mini–mergers output: the BU/FU appliances are expected to provide an aggregated traffic of 2GB/s into the distributed file system;
- macro–merger input and output: the macro–merger needs to process the output of the mini–mergers online, which means read at 2GB/s and write at 2GB/s;
- transfer input: the transfer system reads and transfers the merged data at 1GB/s;
- overall: the design of the storage system has to ensure a total sustained bandwidth of 7GB/s of parallel read/write.

3. Storage and Transfer System Implementation

3.1. Merger system

Multiple strategies have been considered to fulfill the requirements of the merger system. Two of them are currently available and will be described as follows.

The first implementation is the Additive (A) option: it follows the standard logic that has been described in the previous section: there is one mini–merger process running on each BU which outputs one data file and its peer meta–data (per BU per LS per stream) into the

distributed file system. The macro–merger picks up these files and aggregates them into the final files, which are then exposed to the transfer system. This implementation is robust and relatively easy to debug.

The second implementation, which is currently in use, is the Copyless (C) option: it takes advantage of the fact that the outputs of the mini and macro–mergers share the same physical support. Thus the mini–mergers write in parallel in the final "macro–merged" file in a carefully arbitrated manner, while the macro–merger only checks for the completion of this file and exposes it to the transfer system. This implementation comes with the huge advantage of reducing the required bandwidth to 3GB/s by eliminating one read and one write operation from the macro–merger. It is also extremely fast due to the parallel writing into the same file. However, option C is presumably more sensitive to corruption, so the arbitration mechanism has been implemented with particular care.

3.2. Transfer system

Unlike the merger system, the transfer system is not a new component in the CMS DAQ chain. A decision was taken to continue relying on the transfer system which has been used during the Run1 period. However, due to the several changes in the data input format and in the storage infrastructure, a number of features have been added:

- destination: the transfer system is capable of identifying the final destination of data for each stream and per run
- bookkeeping: a new logic has been implemented in the hand–shake protocol between the CMS site and Tier0, mainly to account for the single unified data storage

3.3. Storage system

A number of distributed file systems have been evaluated, such as GPFS, EOS, oneFS and Lustre. It was concluded that the most suitable solution given the requirements was Lustre [4]. Figure 2 depicts the Lustre functional concept.

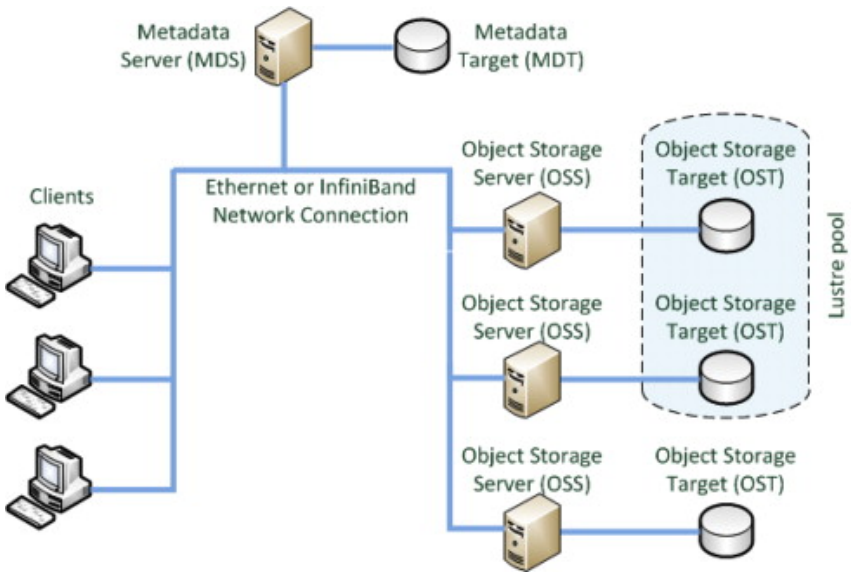


Figure 2. Lustre FS architecture.

The storage system is divided in 2 logical components: the data storage and the meta–data storage. The data storage consists of the physical storage itself, the Object Storage Targets

(OST), and the servers that exposes it to the clients, the Object Storage Servers (OSS). The meta–data storage has a similar structure, with a storage device (Metadata Target, MDT) which is exposed via a Metadata Server (MDS). The role of the MDS is to orchestrate the usage of the OSTs and to direct the clients requests to the appropriate OSS.

In terms of servers, a choice was made to use DELL R720 both as OSS and MDS. The hardware that has been chosen for the storage itself (OSTs and MDT) is E–Series devices from NetApp:

- MDT controller: 1 E2724 with 16 drives of 1TB each, partitioned in one RAID6 volume group with an additional 8 hot spare disks bay
- OSTs controllers: 2 E5560 with 60 disks of 2TB each, partitioned in 6 RAID6 volumes, for a total of 240TB raw space
- OST expansion shelves: 2 DE6600 with 60 disks of 2TB each, partitioned in 6 RAID6 volumes, for a total of 240TB raw space

Figures 3 and 4 depict the OST storage devices as installed in the racks at CMS facility at CERN, Cessy.

Figure 3. Front OST.

Figure 4. Disk shelves.

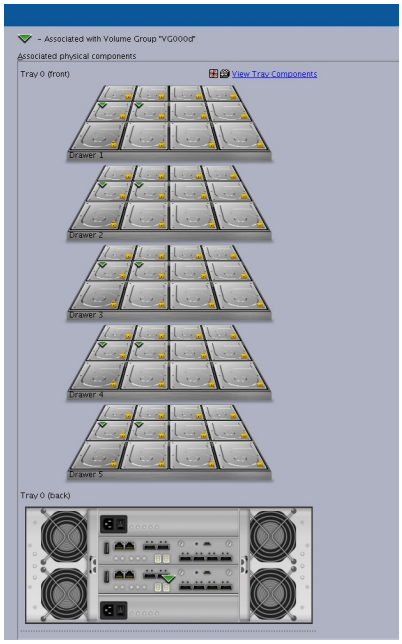
4. Status and Conclusion

4.1. Merger system

The merger service has been used in cosmic runs for more than 6 months, of which it has been proven to be very stable and reliable. Although, at the time of writing this article, there have been no collisions in Run2, the mergers were tested in beam/splashes conditions in April 2015 and there were no worrying signs for the upcoming collisions. A typical cosmic run is shown in figure 6: it can be seen that both the minimergers and the macromerger are keeping up with the actual data taking and as soon as a new LS is available from the DAQ system, the minimergers start. For more details on the layout of the monitoring page see [2].

Another interesting picture showing typical mergers latencies is given in figure 7. It shows the actual delays (measured in seconds) between the different stages of the merging process: the upper plot shows the average delay of the mini–mergers with respect to the time when the FUs have delivered their selected events and the lower plot shows the time between the mini–mergers finishing their aggregation and the macro–merger delivering the final files to the transfer system. In this particular sequence the total delay was not higher than 10s, and this is quite representative for the general mergers behaviour.

It can be stated that the mergers have been proven to be very reliable and steady throughout the tests that have been performed, both with and without beams.



A significant amount of effort has been invested in configuring the storage in high–availability mode. This is ensured at a number of levels, both hardware and logical. All the devices which are involved in the configuration are dual powered from normal and UPS sources. All the servers are configured in active/passive pairwise fail-over mode. The volume partitioning has been done in such a way as to ensure full shelf failure redundancy, as shown in figure 5.
Additionally, the Lustre file system is exported via both data networks available in the CMS private network, the InfiniBand (56Gb) and the 40GE. However, even if it can be mounted via either of the networks, the fail-over between the two of them is not automatic as of now. Possible solutions are being investigated.

Figure 5. Volumes partitioning.

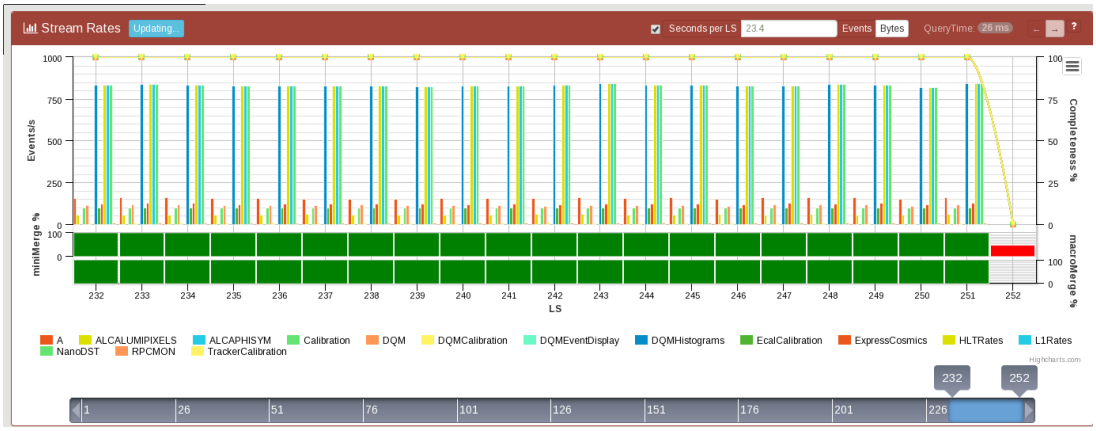


Figure 6. Mergers monitoring sample.

4.2. Transfer system

The transfer system has been successfully upgraded to transfer and account for DAQ2 merged files. At the end of April 2015, its Tier0 output has also been migrated from the old CASTOR destination to the new EOS one. Figure 8 shows a typical transfers monitoring sample. The first table gives an overview of the latest runs, while the second one provides detailed numbers related to file sizes, bandwidth usages in and out of the transfer system and status of files in Tier0. The monitoring page helps identifying possible issues and delays, such as the run 239785 missing one file.

As of beginning of May 2015 there is work in progress for the transfer system, mainly for throughput benchmarking and optimization, but also for having the whole system fully managed in the system administration central management system, puppet [3].

Figure 7. Mergers delays sample.

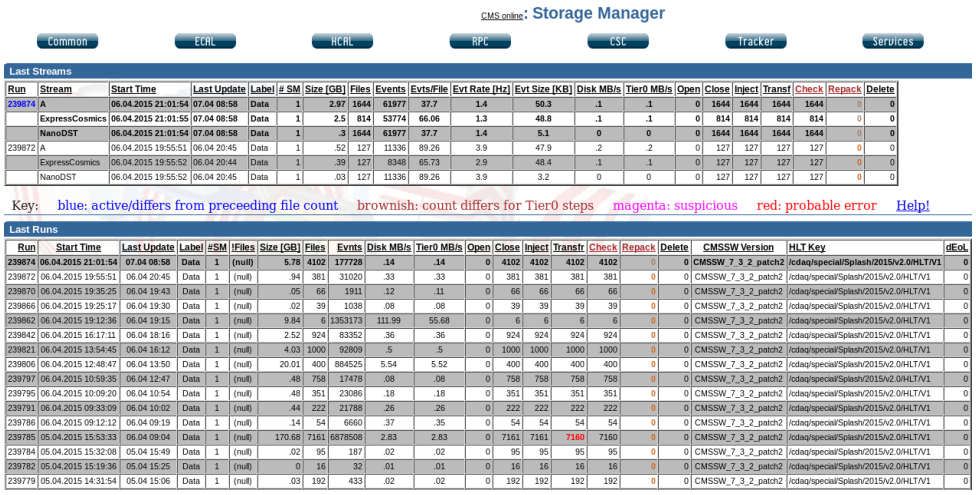


Figure 8. Transfer system monitoring sample.

4.3. Storage system

Given the strict functional and performance requirements for the storage system, it has been carefully designed and tested before being deployed in production. The total usable disk space of the Lustre file system is 349TB, which meets the initial requirement of 250TB. The most interesting aspect was the validation of the bandwidth requirements. A first test was performed during the commissioning of the hardware in order to prove that the hardware performance exceeds the initially required bandwidth of 7GB/s in parallel read/write operation. The plot in figure 9 shows the throughput obtained with plain dd commands on one of the two OST controllers. A steady 5GB/s was observed per controller and the controllers load balance was perfectly symmetrical, thus the full system comprising two controllers delivered 10GB/s total throughput over the Lustre file system. One particularity that was observed during the commissioning was that the write processes tended to get a higher priority, thus the obtained 10GB/s were split into ∼3GB/s read and ∼7GB/s write. This suits our use case, because the reads have lower priority than the writes.

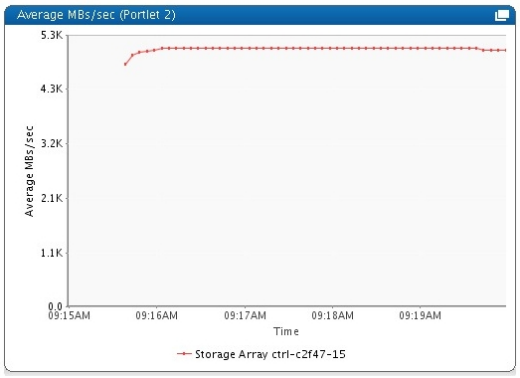


Figure 9. Commissioning acceptance.

The second validation stage consisted in emulating mergers runs with different number of BUs. This kind of testing had the double purpose of testing both the storage system bandwidth and the robustness of the merger system. While figure 10 shows one of the steady emulation runs during which the mergers have been running at an average of 7.5GB/s rate (the figure depicts the rate only on one of the two controllers) for more than four hours, figure 11 shows the results that have been obtained using different number of BUs. The fit function shows an obvious non–linear behaviour with the number of BUs and that a saturation of the bandwidth is expected. However, for the final number of 64 BUs we can still expect more than 8GB/s, which exceeds the initial requirement.

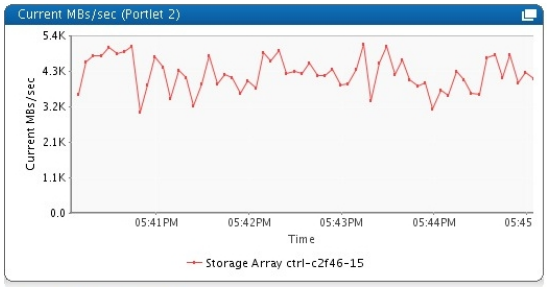


Figure 10. Merger emulation.

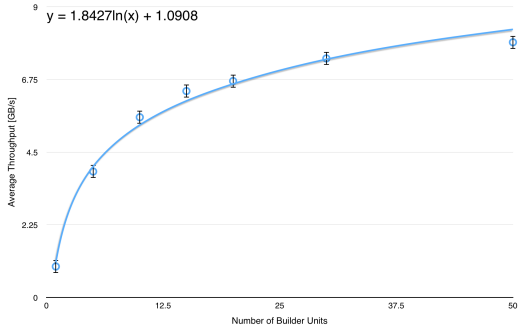


Figure 11. Storage bandwidth benchmarking.

Acknowledgments

This work was supported in part by the DOE and NSF (USA).

References

[1] E Meschi et al., File-based data ow in the CMS Filter Farm, CHEP, Okinawa, Japan, 2015
[2] S Morovic et al., A scalable monitoring for the CMS Filter Farm based on elasticsearch, CHEP, Okinawa, Japan, 2015
[3] https://puppetlabs.com/
[4] http://opensfs.org/lustre/