STORAGE ELEMENT PERFORMANCE OPTIMIZATION FOR CMS ANALYSIS JOBS

G. Behrmann[1], J. Dahlblom[2], J. Guldmyr[2], K. Happonen[3] and T. Lindén[3]

[1] *Nordic Data Grid Facility, Kastruplundgade 22, DK-2770 Kastrup, Denmark*
[2] *CSC -IT Center for Science, P.O. Box 405, FI-02101 Espoo, Finland*
[3] *Helsinki Institute of Physics, P.O.B. 64, FIN-00014 University of Helsinki, Finland*

Tier-2 computing sites in the Worldwide Large Hadron Collider Computing Grid (WLCG) host CPU-resources (Compute Element, CE) and storage resources (Storage Element, SE). The vast amount of data that needs to processed from the Large Hadron Collider (LHC) experiments requires good and efficient use of the available resources. Having a good CPU efficiency for the end users analysis jobs requires that the performance of the storage system is able to scale with I/O requests from hundreds or even thousands of simultaneous jobs.

In this presentation we report on the work on improving the SE performance at the Helsinki Institute of Physics (HIP) Tier-2 used for the Compact Muon Experiment (CMS) at the LHC. Statistics from CMS grid jobs are collected and stored in the CMS Dashboard for further analysis, which allows for easy performance monitoring by the sites and by the CMS collaboration. As part of the monitoring framework CMS uses the JobRobot which sends every four hours 100 analysis jobs to each site. CMS also uses the HammerCloud (HC) tool for site monitoring and stress testing and HC has replaced the JobRobot. The performance of the analysis workflow submitted with JobRobot or HC can be used to track the performance due to site configuration changes, since the analysis workflow is kept the same for all sites and for months in time. The CPU efficiency of the JobRobot jobs at HIP was increased approximately by 50 % to more than 90 %, by tuning the SE and by improvements in the CMSSW and dCache software. The performance of the CMS analysis jobs improved significantly too. Similar work has been done on other CMS Tier-sites, since on average the CPU efficiency for CMSSW jobs has increased during 2011. Better monitoring of the SE allows faster detection of problems, so that the performance level can be kept high. The next storage upgrade at HIP will consist of SAS disk enclosures which can be stress tested on demand with HC workflows, to make sure that the I/O-performance is good.

Finnish CMS Tier-2 hardware setup

The Helsinki Institute of Physics (HIP) CMS Tier-2 resources (T2_FI_HIP) in Finland are distributed over two locations, the University of Helsinki Kumpula campus and CSC - IT Center for Science. The storage is based on the dCache system and located at CSC with the main Linux cluster Jade (768 cores), which replaced the Sepeli cluster in 2010. Jade is shared only between CMS and ALICE. On the Kumpula campus there is a PhEDEx- and a Frontier-server and the Linux clusters Korundi (400 cores) and Alcyone (860 cores), which replaced the Ametisti cluster. Korundi and Alcyone are shared between the HIP and the Departments of Chemistry, Computer Science and Physics. An approximately 10 km long Optical Private Network (OPN) connects the storage at CSC to the Kumpula campus. The 246 TB dCache storage hardware consists of a Hitachi AMS 1000 and AMS 2500 Fibre Channel SAN and a Sun Fire X4540. The system will be upgraded with sixteen HP D2600 SAS diskshelves with 384 TB of raw disk. A Sun Fire X4540 and a D2600 disk shelf are used for 56 TB of Lustre workspace.

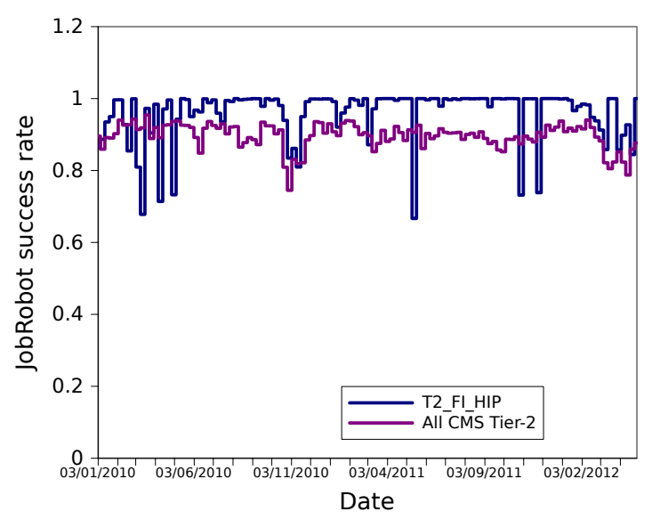
Finnish CMS Tier-2 software setup

- **Storage Element**
  - **dCache** [1] has been used, knowledge and support mostly available for dCache, since the Nordic Tier-1 (NDGF) uses and contributes to dCache
  - Both dCache/xrootd and dCache/dcap doors have been used
  - OpenSolaris on dCache head nodes with ZFS on the AMS LUNs
  - Linux on the Sun Fire X4540 with XFS on software RAID
  - Linux on HP DL360G7 diskservers connected with SAS RAID controllers to the D2600 diskshelves
- **Compute Element**
  - The Nordugrid **Advanced Resource Connector** (ARC) middleware [2, 3]

Site monitoring and stability

To ensure that the WLCG resources are available and usable a comprehensive set of monitoring tools are in use. WLCG runs Site Availability Monitoring tests on all resources once every hour. These generic tests are used to check the basic functionality of the resources. In addition to this CMS runs its own monitoring tools to check higher level functionality and to ensure that the resources can run the needed applications, for details see these two CHEP 2012 contributions [4, 5].

The stability, reliability and low maintenance needs of ARC CEs has enabled T2_FI_HIP to use several ARC CEs. The redundancy of the CE service ensures that CMS-jobs can be run even if one CE is down. The next Figure plots the success rate for CMS JobRobot jobs since 2010 for T2_FI_HIP and the success rate for the average of all CMS Tier-2 sites.



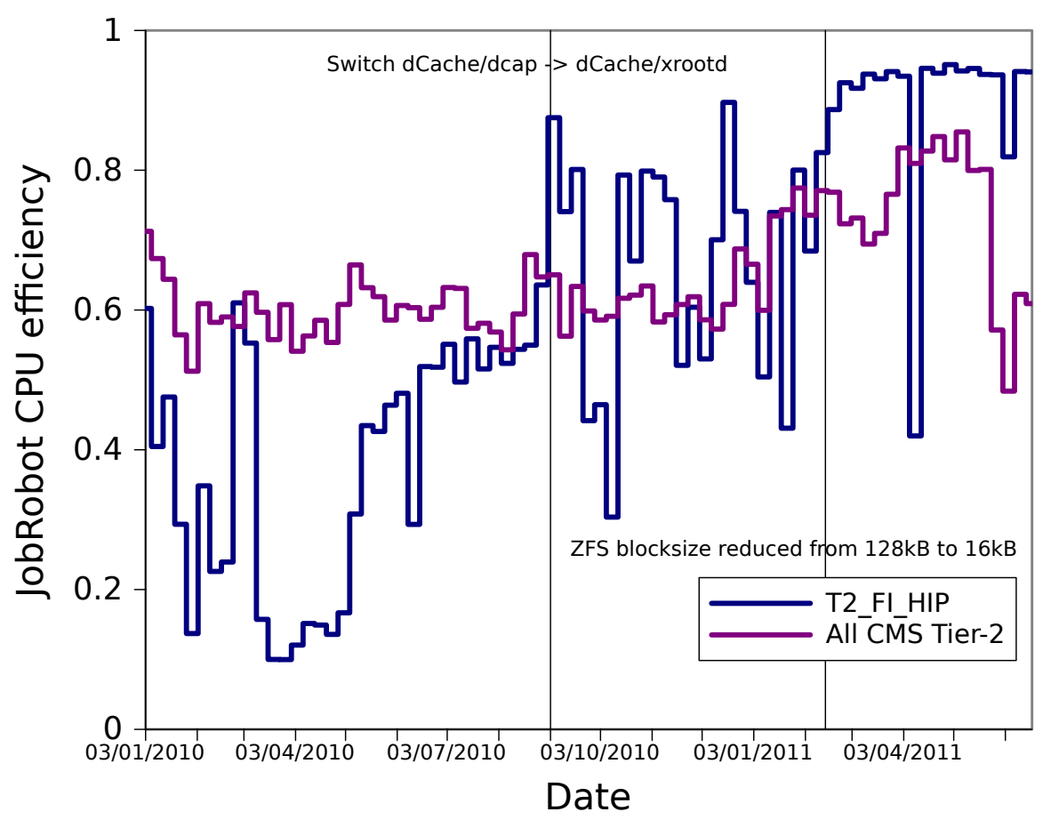
JobRobot success rate efficiency of jobs 2010 January - 2012 April. The success rate of 458 832 jobs at T2_FI_HIP was 95.6 %.

Performance

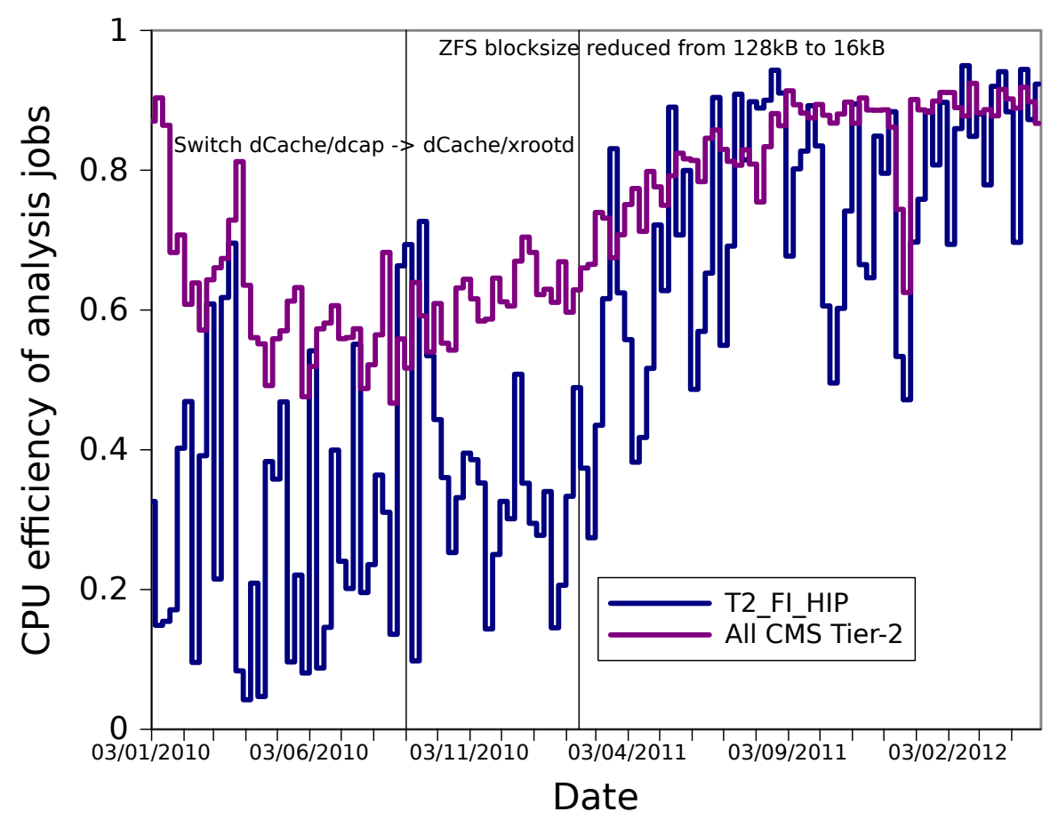
Monte Carlo simulation is mainly a CPU bound problem, so achieving a good CPU efficiency for these kinds of jobs is not difficult. Monte Carlo production jobs spend tens of seconds / event, with very little input I/O.

CPU efficiency depends on the time needed to process an event. Analysis jobs can spend as little as tens on ms / event, which creates a significant load on the storage system and the jobs can be I/O-bound. Analysis of CMS data requires good performance from the disk system, the network and the application software. The 1 Gb/s network link between the Kumpula campus and CSC became saturated in 2010. A new 10 Gb/s link enabled the efficient usage of the Korundi ARC CE and improved CPU efficiency of the analysis jobs. When the new Jade CE was installed then the old Sepeli CE had to be connected before decomissioning with only a 1 Gb/s network link to the dCache storage, which decreased the CPU efficiency of the analysis jobs temporarily. In September 2010 the default dCache access protocol was changed from **dCache/dcap** to **dCache/xrootd**, which improved the JobRobot CPU efficiency by 30 %, because the xrootd driver of ROOT has less aggressive read-ahead than the dcap driver and therefore does not saturate network and disk bandwidth as quickly, see the next Figure.

The **ZFS**-filesystem in use on the AMS 1000/2500 diskpools uses atomic reads of the set blocksize to be able to protect against bitrot by calculating checksums of all reads. The CMS data is stored as objects in ROOT trees and based on a few examples studies using **DTrace** the majority of the read operations were seen to be 1–5 kB in size. This lead to read amplification saturating the disk system in high load situations and giving poor CPU efficiency, because the blocksize was 128 kB. All CMS data was rewritten with a new ZFS blocksize of 16 kB in February/March 2011, which improved the CPU efficiency of both JobRobot and real analysis jobs, see the next Figures. CMSSW has since then been developed to increase the read size of the objects read to decrease the load on the storage system.

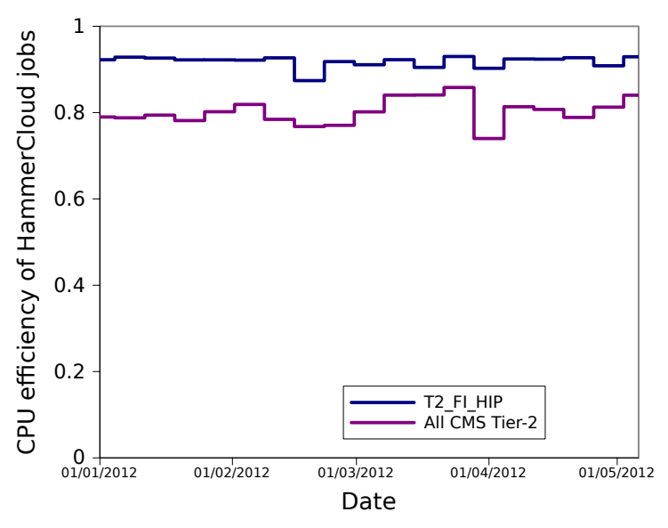


JobRobot jobs CPU efficiency 2010-01 - 2011-06.

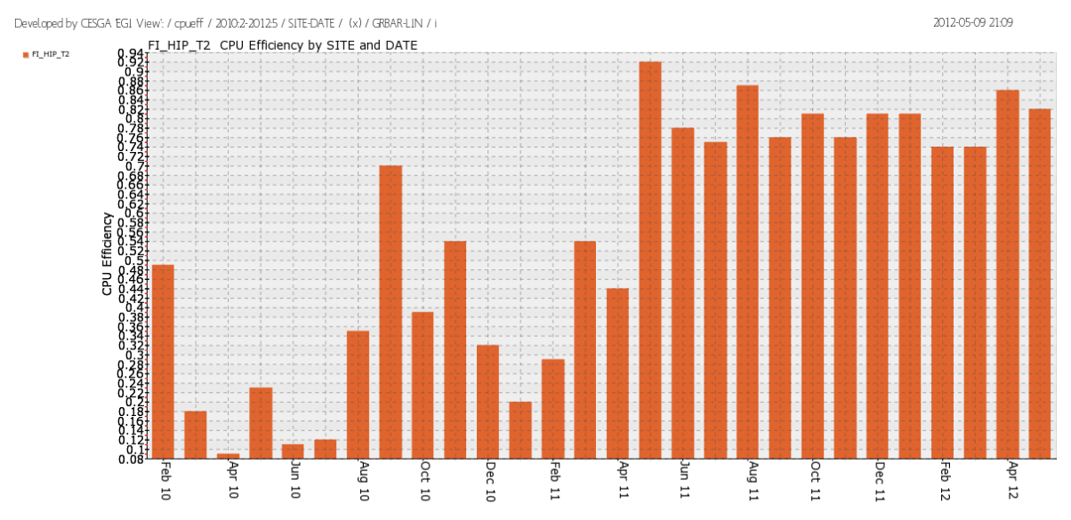


Analysis jobs CPU efficiency 2010-01 - 2012-04.

The dCache version was upgraded frequently in 2011 to have the latest dCache/xrootd bugfixes included and the RAID system was tuned to do less checks to decrease the internal I/O activity on the system. The number of times dCache/xrootd tries to open a file was increased from 3 to 100 and with these changes the analysis jobs run faster and more reliably than before. The CPU efficiency is very good and stable for the HC-jobs, see the next Figure to the left. The next Figure to the right shows the average CPU efficiency for all CMS jobs at T2_FI_HIP computed from the APEL database. The CPU efficiency has been good since the performance problems were solved, but with room for improvement.

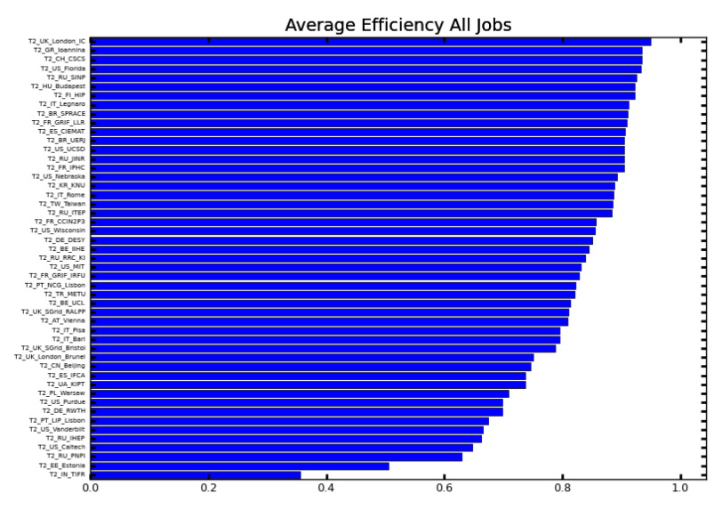


The average CPU efficiency of CMS Tier-2 HC jobs January - May 2012.

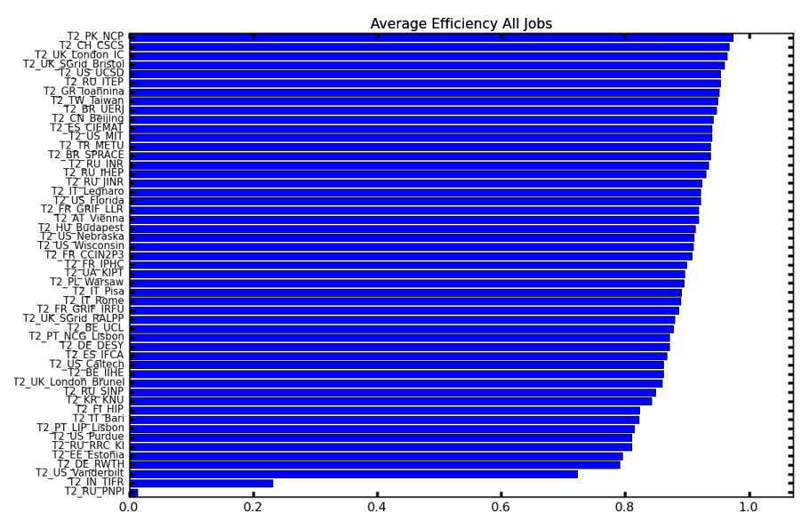


The average CPU efficiency of all CMS jobs at T2_FI_HIP February 2010 - April 2012 from the WLCG APEL database.

For some I/O intensive analysis jobs the T2_FI_HIP CPU efficiency is below the CMS Tier-2 average as can be seen in the following plots comparing the HC-jobs and the analysis-jobs CPU efficiency distribution for CMS Tier-2 sites. This is the case if the event processing time is very short compared to the overhead of opening files or if several jobs access data residing on a single pool and overload it.



HC-jobs CPU efficiency Tier-2 distribution Jan-Apr 2012.



Analysis-jobs CPU efficiency Tier-2 distribution Jan-Apr 2012.

The SAS D2600 disk shelves give a sequential read speed of the order of 800-900 MB/s, while the AMS 1000/2500 pool servers can only deliver 100–200 MB/s. HC-jobs store all details of the job I/O statistics, which is very useful for optimizing storage performance for random access I/O to further increase the CPU efficiency of CMS analysis jobs.

Summary

- The Finnish CMS Tier-2 site has shown good availability and reliability.
- The average JobRobot success rate is 95.6 % since 2010.
- Use of redundant ARC CEs gives reliability.
- The CPU efficiency of analysis jobs was poor in 2010, due to configuration changes, disoptimal filesystem settings and the small CMSSW reads.
- The Storage Element performance was improved in February 2011 by decreasing the ZFS blocksize from 128 kB to 16 Kb and rewriting all data on disk. Since then CMS JobRobot- and HC-jobs run stably with a CPU efficiency exceeding the average of all CMS Tier-2 sites.
- HC-jobs can be used to optimize the the CPU efficiency of CMS analysis jobs for the new SAS disks.

Acknowledgements

This project has been funded by the Finnish Ministry of Education and the Academy of Finland. The CSC personnel is thankfully acknowledged for maintaining CPU and disk resources.

References

[1] dCache homepage, `http://www.dcache.org/`.

[2] M.Ellert et al., *Advanced Resource Connector middleware for lightweight computational Grids*, Future Generation Computer Systems 23 (2007) 219-240.

[3] NorduGrid 2012 conference, May 30-June 1, 2012, Uppsala, `http://indico.hep.lu.se//conferenceTimeTable.py?confId=1185`.

[4] A.Sciaba et al., *Towards a global monitoring system for CMS computing operations*, Talk 182, Track: Distributed Processing and Analysis on Grids and Clouds (track 3), Proceedings of CHEP 2012.

[5] J.Flix et al., *Towards higher reliability of CMS Computing Facilities* Poster 260, Track: Distributed Processing and Analysis on Grids and Clouds