A new era for central processing and production in CMS

E Fajardo\textsuperscript{1}, O Gutsche\textsuperscript{2}, S Foulkes\textsuperscript{3}, J Linacre\textsuperscript{4}, V Spinoso\textsuperscript{5}, A Lahiff\textsuperscript{6}, G Gomez-Ceballos\textsuperscript{7}, M Klute\textsuperscript{8}, A Mohapatra\textsuperscript{9}

\textsuperscript{1} Uniandes, Colombia
\textsuperscript{2} Fermi National Accelerator Laboratory, USA
\textsuperscript{3} Fermi National Accelerator Laboratory, USA
\textsuperscript{4} Fermi National Accelerator Laboratory, USA
\textsuperscript{5} INFN, Italy
\textsuperscript{6} RAL, UK
\textsuperscript{7} Massachusetts Institute of Technology, USA
\textsuperscript{8} Massachusetts Institute of Technology, USA
\textsuperscript{9} University of Wisconsin, USA

E-mail: efajardo@cern.ch\textsuperscript{1}, gutshe@fnal.gov\textsuperscript{2}, sfoulkes@fnal.gov\textsuperscript{3}, linacre@fnal.gov\textsuperscript{4}, vincenzo.spinoso@ba.infn.it\textsuperscript{5}, andrew.lahiff@stfc.ac.uk\textsuperscript{6}, guillelmo.gomez.ceballos@cern.ch\textsuperscript{7}, klute@mit.edu\textsuperscript{8}, ajit@hep.wisc.edu\textsuperscript{9}

Abstract. The goal for CMS computing is to maximise the throughput of simulated event generation while also processing event data generated by the detector as quickly and reliably as possible. To maintain this achievement as the quantity of events increases CMS computing has migrated at the Tier 1 level from its old production framework, ProdAgent, to a new one, WMAgent. The WMAgent framework offers improved processing efficiency and increased resource usage as well as a reduction in operational manpower.

In addition to the challenges encountered during the design of the WMAgent framework, several operational issues have arisen during its commissioning. The largest operational challenges were in the usage and monitoring of resources, mainly a result of a change in the way work is allocated. Instead of work being assigned to operators, all work is centrally injected and managed in the Request Manager system and the task of the operators has changed from running individual workflows to monitoring the global workload.

In this report we present how we tackled some of the operational challenges, and how we benefitted from the lessons learned in the commissioning of the WMAgent framework at the Tier 2 level in late 2011. As case studies, we will show how the WMAgent system performed during some of the large data reprocessing and Monte Carlo simulation campaigns.

1. Introduction

The Compact Muon Solenoid (CMS) is one of the four detectors operating at the LHC, collecting and processing data taken from proton-proton collisions at a center-of-mass energy of 7 TeV in 2011 and at 8 TeV in 2012. CMS computing has the task of processing this data as quickly and reliably as possible, as well as producing the Monte Carlo (MC) simulations needed for the end-user analysis. To achieve these objectives a computing model was designed in which computing resources were hierarchically distributed around the world [1].
At the top of this tree lies the Tier-0, a unique computing center located at CERN. The Tier-0 is responsible of storing a cold archival copy (not to be accessed frequently) of the RAW data out of the detector. Additionally it is responsible for the repacking and prompt reconstruction of data, classification of data into Primary Datasets (PDs) according to their physics content, and distribution of these PDs among sites at the next Tier level (Tier-1).

The Tier-1 level is composed of seven sites located in France (IN2p3), Germany (KIT), Italy (CNAF), Spain (PIC), Taiwan (ASGC), the United Kingdom (RAL), and the United States (Fermilab). Each Tier-1 site is assigned “custodiality” of one or more of the RAW data PDs, which are transferred from the Tier-0 and stored as hot copies (for frequent access). The Tier-1 sites are then used to re-reconstruct (ReReco) their custodial PDs.

Each Tier-1 site is also assigned custodiality of a portion of the MC simulations that are produced at the Tier-2 level. The Tier-1 sites then run the redigitization (ReDigi) and ReReco of their custodial MC samples. In case the resources are available, the production of the MC simulations can also run at the Tier-1 sites.

The Tier-2 level consists of more than 50 sites around the world located in universities and small institutes. The computing resources of the sites are used partly for the production of MC simulations and partly for end-user analysis work. The last level, the Tier-3 doesn’t have any pledged resources for the experiment central production. Even though central production can use these resources opportunistically.

CMS Computing organizes all of this work through a team of experts distributed worldwide: the Workflow Team and the Coordinators. In order to increase the reliability and throughput of the system, in early 2011 the Workflow Team moved from the Production Agent (ProdAgent) framework to the new Workload Management System (WMAgent) framework. This change was made in several steps. First, at the beginning of 2011 all Tier-1 level work was transferred to the new framework: data ReReco and MC ReDigi and Rereco. The Tier-2 level work (MC production) was still done using the ProdAgent until late 2011, when the production of MC simulations also migrated to the WMAgent framework.

This paper describes the operational benefits of the change of framework, the actual deployment of the WMAgent, the organization of the Workflow Team and its achievements during 2011/12 operating the new framework.

2. ProdAgent Vs WMAgent

There were some operational issues encountered during the operation with ProdAgent [4], of which 100% accountability had the highest priority. Therefore a change from a message-based system (ProdAgent) to a state-based system (WMAgent) was designed. Once this and other issues were addressed, a plan to migrate, commission and deploy the newly WMAgent in production was executed. To some extent the WMAgent is the evolution of the ProdAgent, but although they share some behavior, the basic idea on how the work gets distributed is totally different. In this section we will show the basic differences which make the highest impact on operations. For more details on the design and technical differences between ProdAgent and WMAgent, see [2] and [3] respectively.

2.1. Manpower

A WMAgent can run several requests (workflows) in parallel while also keeping several in queue. On the other hand the ProdAgent can only run one request at a time and once it is done the operator had to inject a new one. Hence with the migration of framework the operator task also changed: From individually running specific requests to monitoring a fully integrated system as a team. Therefore in order to fully cover the production and processing needs with the ProdAgent it was necessary to have five experts operating the infrastructure full time. Each of those experts was in charge of monitoring their own ProdAgent installations and running the
work that was assigned to them. Instead, with the WM Agent it suffices to have two full time experts (Workflow Team Leaders), and several part time Operators. Thus the costs of operating the system were reduced from 5.5 expert FTE with the Prod Agent, to only two expert FTE (Team Leaders) plus 1.5 service FTE (Operators) with the WM Agent.

2.2. Coordinators Task

Alongside a reduction of manpower came a change in the Coordinator’s tasks. In the Prod Agent framework, the Coordinators would receive, via email or a Hypernews forum, the configuration of the work to be injected. They would then decide at which sites to run the work, thus assigning it to one of the five different operator teams which was each allocated different zones to run their assignments (one Tier-1 with several Tier-2 sites associated with it). In the WM Agent framework, requests and their configurations are injected into the central Request Manager directly by the requestor, allowing full accountability and bookkeeping. The Coordinators then assign each request to appropriate sites. For data ReReco and MC reprocessing a single Tier-1 is chosen. On the other hand for MC production one custodial Tier-1 site is chosen alongside all of the Tier-2 sites which have stable data transfer link to the selected Tier-1 (more details in [5]).

In the current version one request cannot be split on several WM Agent instances. Hence the coordinator also assigns the request to one WM Agent instance (there are separate WM Agent instances for reprocessing and MC).

The request and assignment process is done with the Request Manager, with no need for interchange of email for configuration or assignment and thus less human intervention and fewer errors. Finally, when the Workflow Team signals the request as closed-out, denoting that the output datasets are complete and available at their custodial Tier-1 site, the Coordinator just needs to announce it to the CMS community. In contrast, with the Prod Agent there was a continuous exchange of information (email, web forums, and an e-log) about the status of a given request.

3. Production Setup

The manpower reduction mentioned above was possible due to the change in the framework and the setup. With the Prod Agent there were autonomous instances, which required autonomous teams. Instead, in the new system the WM Agent instances are all connected to single central Request Manager and Global Workqueue.

In this framework a normal request follows the path described in Figure 1. When a new request is injected into the Request Manager, Physics Management can approve it or reject it. If it is approved the Coordinator assigns the request as was mentioned in the section above. Afterwards the Global Workqueue acquires the request and divides it into smaller blocks of work. Then it distributes them to the WM Agents. Later it creates the jobs out of the blocks of work that it acquired and submits the jobs to the sites, through the GlideIn System [6] (with the Prod Agent the grid submission used a combination of gLite, condor router, and GlideIns). Then different information about the jobs is uploaded to different CMS services. The job status is uploaded to Dashboard [7] and the job output (files) information is uploaded to PhEDEx (Physics Experiment Data Export) [8] and DBS (Data Bookkeeping System) [9].

The change of framework, infrastructure, and grid submission allowed an increase in efficiency. Since the number of parallel running jobs achieved in central production greatly rose from the Prod Agent to the WM Agent (Figure 2). A record high of 50k parallel running jobs was achieved with the WM Agent framework.

3.1. Actual setup

Our current setup comprises seven WM Agent instances located on both sides of the Atlantic (Fermilab and CERN), connected to a single Request Manager (ReqMgr) and Global Workqueue
Figure 1. WMAgent set-up and operational workflow as a production service.

Figure 2. Number of simultaneously running jobs, for both central processing and Monte Carlo production activities (yellow and red colors, respectively). Number of jobs submitted via ProdAgent (left) and WMAgent (right) are displayed.

located on the cmsweb.cern.ch cluster at CERN. This cluster also holds the frontend for other CMS critical services such as PhEDEX and DBS. Four of the seven WMAgents are used for MC production: two handle bulk requests, one handles high priority requests, and the last one is used for overlap when new WMAgent versions are installed. The four agent instances are connected to a GlidelIn frontend located at CERN which is served by up to three factories.
located at Indiana, UCSD and CERN [10].

The Tier-1 processing setup is similar, except that one bulk processing WMAgent suffices (a WMAgent running on a machine with 48GB of memory can comfortably handle up to 20k parallel running jobs). The three Tier-1 processing WMAgents are connected to a frontend at Fermilab which is served by another factory at FNAL. This setup allows a high level of redundancy which mostly avoids a single point of failure. In the case of one WMAgent failing, the resource utilization would not be deeply affected because the other ones can automatically ramp up.

4. The Workflow team
The people in charge of operating the above mentioned infrastructure is the Workflow Team. It consists of people spanning different time zones from Europe to California who are employed by more than five different institutions. The advantage of the worldwide distribution of manpower is almost round the clock surveillance of the system. However the task for coordinators and team leaders to keep a coherent group of people working together towards the same objective becomes harder. In this section we will present the team tasks, the advantages of the highly distributed characteristics of the team and the challenges that come along with it.

4.1. Team tasks
There are three main team tasks: keep the infrastructure working (debug WMAgent problems), debug site specific problems, and mark requests as ready to be announced. Most of the monitoring is done through a web application: the Global Monitor. It provides the status of the different requests (new, assigned, running, completed, closed-out, announced), the number of pending, running, successful, and failed jobs of a request (Request Monitor), the status of the agents and their components (Agent Monitor), and the number of parallel running, failed and successful jobs grouped by site (Site Monitor).

Infrastructure monitoring The objective of this task is to keep the system up and running. This mainly consists of checking in the Agent Monitor the components of the different agents marked as down. If a component is down the operator proceeds to log in to the machine and check the logs for the given component, restart it, and make an entry on the team e-log about the incident. If the problem persists then one of the team leader checks and diagnoses the situation, and if needed contacts the WMAgent support (developers). But keeping up the infrastructure is not limited to just restarting some components: it also involves checking the load on the machines and if necessary clean up disk space. An alert system is being developed in order to further automatize this task and it will be soon commissioned for production [11].

Debugging site problems Given the distributed nature of the CMS computing model the debugging and monitoring of site specific problems becomes a communication challenge. Site problems can be divided into two types: jobs cannot enter a site or jobs can run but fail. The first one is solved by communicating the problem to the GlideIn factory support and debugging the problem with them. The second one involves the monitoring of failed jobs in the Global Monitor together with the information of Dashboard. Once the nature of the problem is found—a site problem or a configuration problem—it is then correctly propagated, to the site through the Savannah and GGUS ticket systems [12] in the first case or to the requestor in the latter one. Once the situation is understood the request is either aborted (user configuration problems) or the submission to the site is stopped until the problem is understood and solved.
Closing-out requests  Once a request is finished, all jobs have either succeeded or failed. At this moment the operator checks that the event count of the request is higher than 95% of the requested events for MC production and reprocessing and 100% for real data. If it is lower, then another request must be created to recover the failed jobs; this task normally involves communication with the coordinators. Once the event count is correct, the transfer subscription (to the corresponding custodial Tier-1 site) is monitored to be at 100%. Just after the operator agrees that the event count and transfers are fine, he sets the status of the request on the ReqMgr as closed-out.

Commissioning of Sites and Infrastructure  The final task which is mainly performed by the Team leaders is to tightly work with the Integration team in commissioning activities. These ones involve commissioning of new sites for MC production, as well as testing new versions of ReqMgr, Global Workqueue and WMAgent.

4.2. Benefits of a highly distributed heterogeneous team
There are several benefits that come along with the increase of headcount and distribution of manpower around the globe. First, almost 24/7 monitoring of the system is available since the members of the team span from Europe up to the western United States. Second, a high level of redundancy is available to cover unavailability, e.g. Conferences, training courses, other non-research duties, holidays, sickness, and even maternity leaves. Third, it allows for a high level of iteration with the different sites located around the world: an operator in California has higher chances of iterating on a problem in Asia than an operator in Europe; likewise an operator in Europe has higher chances of fixing a problem in Russia.

4.3. Challenges of a highly distributed heterogeneous team
The worldwide nature of team imposes some challenges. Besides the inherent difficulties of keeping a coherent team working towards the same objective, the fact that they are located in different countries and continents, and are employed by up to five different institutions makes it even harder. Coordination of the team is made through the e-log, a simple web-based Logbook [13]. There is also a high turnover of persons, since the service work for the experiment is done by graduate students and postdocs who do not work full time. Hence up-to date documentation of operations becomes indispensable. Currently the team has a twiki page that serves as a knowledge base and also as a starting point for newcomers. Although it is not complete it has met the team demands as it was demonstrated that new members could become operational in less than one month.

5. Achievements 2011/12
Besides the record high of 50k parallel running jobs mentioned above, the Workflow Team has ensured consistently smooth operations throughout 2011-2012. The processing activities of 2011 with the WMAgent have met the CMS demands. All of the data from 2011 was re-reconstructed and up to eight partial re-reconstruction passes were also done. Hence, the total number of re-reconstructed data events is more than twice the total number of events recorded at the Tier-0 (Figure 3). This is a result of several ReReco passes for the same input datasets with different software versions.

The rate of MC event production and reprocessing has increased as can be seen in Figure 4. This coincides with when the MC production moved to the WMAgent at the beginning of 2012, when all operations were grouped in a single coherent team of goals and practices.

Finally, the Workflow Team achieved a 90% reliability in the success/failure rate of the production jobs and more than 95% in reprocessing, as shown in Figure 5.
Figure 3. Number (in millions) of AOD events prompt out of the Tier-0 (left), ReReco reprocessed event by Workflow team at the Tier-1 sites (right), both in year 2011.

Figure 4. MC Production and Reprocessing in 2011/2012: Number of events per month.

6. Conclusions
CMS Computing has met the demands of the experiment’s community by achieving reliable and large scale production and processing operations. The WMAgent showed its reliability and scalability features to efficiently deliver the needed MC samples for physics analysis. Finally, the Workflow Team had a highly efficient operation, up to the point that 800 million MC events could be produced in a month. The future looks promising since the Release Validation operation of new versions of CMSSW is being commissioned to use the WMAgent, hence new operators will join the Workflow Team. The team also expects to improve its operations based on more documented procedures and a less steep learning curve for new collaborators.
Figure 5. Success/failure rate of MC Production (above) and reprocessing (below).

Acknowledgments
We would like to thank the DMWM team for providing us the tools: Request Manager, GlobalWorkQueue and WMAgents from which our operations have greatly benefited. Then we would like to mention the great GlideIn support we have had during this year. We also acknowledge all the CMS Computing Operations and especially the Workflow Team and the Coordinators for their dedicated work on making this operation as successful as it is. Finally we would like to show our greatest gratitude to the international funding agencies for their support on this work.

References
[1] Bonacorsi D and the CMS Computing project 2011 Journal of Physics: Conference Series 331 072005 URL http://stacks.iop.org/1742-6596/331/i=7/a=072005
[2] Evans D, Fanfani A, Kavka C, van Lingen F, Eulisse G, Bacchi W, Codispoti G, Mason D, Filippis N D, Hernández J and Elmer P 2008 Nuclear Physics B - Proceedings Supplements 177-178 285 – 286 ISSN 0920-5632 proceedings of the Hadron Collider Physics Symposium 2007 URL http://www.sciencedirect.com/science/article/pii/S0920563208000418
[3] Wakefield S, Ryu S, Evans D, Metson S, Foulkes S, Norman M and Maxa Z 2012 Journal of Physics: Conference Series CHEP 2012
[4] Adelman-McCarthy J, Gutsche O, Haas J D, Prosper H B, Dutta V, Gomez-Ceballos G, Hahn K, Klute M, Mohapatra A, Spinoss V, Keira D, Caudron J, Liao J, Pin A, Schul N, Lentdecker G D, McCarty J, Vanelderen L, Janssen X, Tsypianov A, Barge D and Lahiff A 2011 Journal of Physics: Conference Series 331 072019 URL http://stacks.iop.org/1742-6596/331/i=7/a=072019
[5] Kaselis R, Magini N, Sartirana A, Kreuzer P, Flix J, Gutsche O, Klute M and Pilerov S 2012 Journal of Physics: Conference Series CHEP 2012
[6] Siligoi I, Bradley D C, Holzman B, Mhashikar P, Padhi S and Wurthwein F 2009 Computer Science and Information Engineering, World Congress on 2 428–432
[7] Andreeva J, Campos M D, Cros J T, Gaidioz B, Karavakis E, Kokoszkiewicz L, Lanciotti E, Maier G, Ollivier V, Nowotka M, Rocha R, Sadykov T, Saiz P, Sargsyan L, Sidorova I and Tuckett D 2011 Journal of Physics: Conference Series 331 072001 URL http://stacks.iop.org/1742-6596/331/i=7/a=072001
[8] Magini N, Ratnikova N, Rossman P, Sánchez-Hernández A and Wuliksh T 2011 Journal of Physics: Conference Series 331 042036 URL http://stacks.iop.org/1742-6596/331/i=4/a=042036
[9] Afaq A, Dolgert A, Guo Y, Jones C, Kosyakov S, Kuznetsov V, Lueking L, Riley D and Sekhri V 2008 Journal of Physics: Conference Series 119 072001 URL http://stacks.iop.org/1742-6596/119/i=7/a=072001
[10] Sfiligoi I, Dost J, Zvada M, Butenas I, Wuerthwein F, Kreuzer P, Teige S, Quick R, Hernández J and Flix J 2012 Journal of Physics: Conference Series CHEP 2012

[11] Maxa Z 2012 Journal of Physics: Conference Series CHEP 2012

[12] Antoni T, Bosio D and Dimou M 2010 Journal of Physics: Conference Series 219 062032 URL http://stacks.iop.org/1742-6596/219/i=6/a=062032

[13] Team W 2012 Workflow team e-log URL https://cmslogbook.cern.ch/elog/Workflow+processing/