



## **U.S. CMS Software and Computing**

### **Progress Report for the 3rd Quarter FY2005**

Technical, financial and management status is reported for the period of April 1<sup>st</sup> to June 30<sup>th</sup>, 2005.

#### **Technical Status**

U.S. CMS Software and Computing efforts were driven by the major milestones and activities for FY05:

- Reliable and automated transfers of data between the CERN Tier-0 and USCMS Tier-1 centers at Fermilab.
- Participation in the LCG service challenges
- Preparation, development and integration of software and computing systems

#### **USCMS Facilities Project**

The USCMS Tier1 facilities group finished ordering their FY05 computing equipment. This included:

- Worker nodes: 280 SuperMicro nodes with dual Opteron 248 processors. Previous FNAL orders for computing equipment had all the nodes arriving at one time. A new method was tried this time, with great success. Rather than ordering all 280 nodes at once, an order for 40 nodes (the number of nodes that fit in one rack) was placed and 7 more rack options, each containing 40 nodes, was specified at the same time. This had many benefits including limiting the FNAL liability for receiving nodes that didn't work as we expected from our demo units and it allowed both FNAL and the vendor to spread out the work over manageable units. KOI won the bid process and delivered nodes to us within a month of confirming the order. We aggressively tested the nodes and had them in full production for the users within 2.5 weeks. We have ordered 3 more rack options for delivery, now that the 1st rack worked properly, and

plan on ordering 3 more options by the middle of September. After this order is place, we plan on retiring our oldest nodes.

- Data disk: We received ~30 TB of Nexsan Atabeast Storage Arrays and put them into dCache production service. We placed another order for ~30 more TB of the same unit and await delivery. This will bring us to the desired ~100 TB of dCache data disk by the end of this fiscal year.
- User disk: We continue to work with IBRIX and now have a fully deployed user disk system with user and group quotas. The system has worked well, with IBRIX responding quickly to new (small) problems. The IBRIX system is deployed on all production nodes in the native client mode, and via NFS on user desktops. One unexpected issue that did arise during this period was the filesystem's performance in a saturated network environment. Our monitoring showed only 3/4 saturation, but closer inspection showed a network oversubscription due to the network aggregation algorithm. When more bandwidth was added (going from 4 to 10 GB), the filesystem performed as expected. We are still analyzing the effect of network oversubscription - something we are trying to avoid of course.
- User disk: In addition to the very high quality STK-LSI disk that IBRIX uses, we are starting to deploy medium quality disk (Infortrend, Nexsan) with IBRIX user disk. We will not back up this disk, and it will be completely under the physics groups control without enforced quotas, a feature the users desire.
- UAF: We introduced condor batch submission to the production workers from the UAF. This is allowing us to begin to phase out the FNAL FBS batch system to something more standard. We have also introduced CISCO 6509 load balancing based on their IOS. This works well, including kerberos and AFS, with modern versions (SL3) of ssh, but has some troubles with older (RH7.3) versions. We have enabled cryptocard access for the older version of linux still on many user desktops. We expect to phase out the older FBS interactive load balancing by November.
- Data transfers: Much work was performed to transfer data from CERN to the USCMS center at FNAL to the 7 USCMS Tier2 sites in an automated way. The underlying data transfer was gridftp. These multiple streams were controlled by the dCache SRM which were in turn managed by the CMS PhEDEx data handling system. USCMS participated in the LCG SC3 data challenge on its production system used by all users. We typically have 30 TB/day read rates and 2-3 TB/day write rates.

## 1.2 Tier-2 Facilities Project

The US-CMS Tier-2 program has made substantial progress. We are in the initial stages of the deployment program after several years in research and development dominated activities.

In September of 2004 a request for proposals was submitted to the US-CMS collaboration. The proposal requested that sites outline a program of work to deploy Tier-2 facilities in time for start of the experiment using up to \$250k of hardware funds and 2FTE of site operations. As outlined in the CMS computing model, a nominal Tier-2 center is expected to offer approximately 1000 kSpecInt2000 of processing and 200TB of storage. In all US-CMS received 10 proposals. All proposals indicated that they would be able to deploy at least the nominal facilities with the project support proposed. Several proposals also included substantial university contributions in addition.

In November of 2004 a Tier-2 evaluation committee was formed consisting of lab and university representatives. Finding representatives knowledgeable in computing, but who has not proposed a Tier-2 computing center was challenging, but possible. The Tier-2 committee evaluated a series of metrics and committee representatives made site visits to all of the applicants during the month of December.

In January, the committee presented its final report and recommendations to the US-CMS manager for Software and Computing. Of the ten applying sites, 3 were existing prototype centers: the University of California, San Diego; the University of Florida; and Caltech. These sites joined with the University of Wisconsin and applied for additional Shared Cyber Infrastructure support through a proposal called DISUN (Data Intensive Science University Networks). This proposal was matched with support for the software and computing program. Three other sites, University of Nebraska, MIT, and Purdue University, were selected as new Tier-2 centers.

A very successful kickoff workshop was held at UW Madison in May to launch the production phase of the Tier-2 program. Tier-2 centers are contributing in an active role in the LCG service challenges, and preparing for the 2006 CMS data challenge. The existing centers and the new centers are part of the official CMS event MC production, and are participating in the deployment of the Open Science Grid.

### **1.3 Grid Services and Interface with the Open Science Grid (OSG)**

This section covers continued support of Grid3, interoperability and collaboration with the LCG and EGEE grid infrastructures, development of Grid services, integration of the OSG Integration Testbed, and contributions towards OSG production. We interviewed and hired three people who started at the end of the quarter.

Grid3 continues in operation. US CMS uses Grid3 resources for continued simulation production, enables the agreed upon use of about 10% of US CMS resources to US ATLAS for the “Rome meeting” physics production, and supports opportunistic use by other users.

Last year we achieved interoperability between the Tier-1 and the LCG compute clusters. We have merged the Grid3 and LCG information schema and providers and additionally installed them at the Caltech and the University of Florida Tier-2 sites. Jobs can be successfully

submitted from the LCG User Interface to execute on these resources.

The Virtual Organization Management System (VOMRS), developed for US CMS and iVDGL last year and in use on Grid3 since that time, was accepted by the LCG for their own deployment. Work continues in collaboration with the LCG deployment and security groups to update VOMRS to meet LCG requirements on handling revoked certificates and to interface it to the CERN HR database.

A first release of Condor-C, which enables a distributed workload management system by allowing jobs in one machine's job queue to be moved to another machine's job queue, was delivered as a contribution to the EGEE gLITE middleware as part of their workload management infrastructure.

Development projects for Grid Services are invariably joint with other stakeholders and the Computing Division. The current phase of the joint Runjob joint project was completed and effort reduced to the small amount needed for support and maintenance.

The first release of the new Virtual Organization role based user accounting mapping and authorization modules were delivered to the Open Science Grid Integration Testbed and supported for deployment on all Compute Elements. These services extend the existing Grid3 VO management service to enable facility administrators to control the local accounts of grid users across the resources they manage, to allow VOs to dynamically manage the relative priority of jobs of different types based on the role of the user submitting the job, and to include administrative interfaces and utilities to provide a coherent system. US CMS will use these components to allow the Tier-1 and Tier-2s control of user accounts to meet local security requirements and to manage prioritization of resources across the grid for CMS production and analysis jobs.

## **1.5 Distributed Tools**

We continue to work on MCPS for production replacement of MCRunjob, and collaborate with RefDB people on summary process reengineering to allow control over the summary formatting and content from the RefDB, and also to disconnect the summary creation process to allow non-production users to create production quality summaries for private samples. The FNAL RunJob common project is successfully complete.

We continue with development and integration work on MCPS for general users, including deployment at Tier-2 sites. We continue to refine the MCPS features based on feedback from these sites and user testing. Documentation and user education efforts are also underway.

In the area of Workload Management, we are developing tools to provide delayed abstract planning for MCPS jobs so that they can tailor themselves to the execution when they arrive, look up software installations, and stage in any data files needed.

Effort has been applied to move intensive Pool XML catalog processing for large datasets to the batch jobs themselves to reduce overhead in job creation. Designs for a “Publication On Demand” service for constructing datasets in an automated fashion has been done, and some prototyping of a web service based system has taken place.

Development of a JobMon plug-in for ShREEK is complete. Any ShREEK based task can now have a JobMon monitor added to it very easily; an MCPS interface has also been created. Testing at FNAL and UCSD has been successful; MCPS jobs at both sites can be interactively monitored via the JobMon services.

PhEDEx has been successfully installed at 4 Tier-2 sites, and effort has been expended towards a release candidate for OSG by integrating components on the OSG integration testbed. PubDB has been deployed at one Tier-2 site.

## **1.6 Core Software and Support, CSS**

USCMS has taken a leadership role in the re-engineering of the CMS software framework, following the framework review of last year and the re-organization of the CMS software effort. The status of the project as a whole is that we are on schedule in phase 1. A demonstrator has been delivered and in March of '05 the coding of the final framework classes has started and is expected to continue until July of '05

Two Software developers at Fermilab, Jim Kowalkowski and Marc Paterno, have been the lead designers of this project. As the lead designers, they have been important in integrating in the participation of volunteer developers from international CMS. They have also been helping with the project planning and generation of the detailed task list. Many of the classes in the CoreSW subsystem have been written by them.

Another Fermilab software developer, Bill Tanenbaum, is implementing many of the essential components of this project. He is also administrator for the international CMS repositories associated with this project. He is writing most of the new POOL services package which provides even input and output.

Another Fermilab developer, Walter Brown, was involved with the initial prototype work of the parameter set/ job configuration system, which was delivered in April.

In addition a developer at Cornell, Chris Jones has, been contracted to work with the framework team. There is also a sizeable effort outside the US, in particular at CERN and in Italy, and now also in France, contributing to the CMS framework re-engineering effort.

In addition, in the software support there has been a lot of work associated with migrating the software build tool SCRAM to the new version, and with moving to a new version of the Linux OS distribution, Scientific Linux v3.

Natalia Ratnikova, who is a computing professional at Fermilab, has also been working on improving the DAR package for binary releases, with the help of a summer student.

## 2.0 Core Applications and Support, CAS

Since the last quarterly report, the CAS portion of the USCMS S&C has been reorganized. Tanenbaum, Paterno and Brown now report to the Software Support area. The efforts of Tuura, Eulisse and Wildish, although tracked by CAS, are reported to the Applications Services area for Tuura and Distributed Computing Tools Area for Eulisse and Wildish. The efforts of the remaining engineers (Case, Xie, Wilkinson, Muzaffar, Osborne and Stavrianakou) will be reported here.

Xie, a developer at CERN employed by Princeton, continued her work on the POOL Filecatalog. She developed a plugin mechanism for a split catalog, and implemented a MySQL backend for the split catalog. She continued work on the RDBMS interfaces in POOL.

Wilkinson, a developer at Caltech, continued work on the Calorimetry Framework. Several new classes were added to EcalPlusHcalTowerBase, to ease its use. This class is becoming widely used by the Ecal and Hcal developers. The Calorimetry rewrite is continuing, with the intention to use as much of the CommonDet structure as possible. The Ecal and Hcal code has been decoupled, allowing one to be used without the other.

Case, a developer at UC Davis, Wilkinson and Xie are all contributing in this area. Case is working together with Muon developers to produce a design for the Muon Barrel database. Wilkinson is working together with Ecal/Hcal on their design. Both are working with the Conditions DB group led by Lueking. In addition, Case and Xie are working together on the POOL RAL, as well as together with the LCG Conditions Database groups.

Muzaffar, a developer at CERN employed by NEU, fixed a number of bugs in the Ignominy package (dependency checking) and improved a number of algorithms. Ignominy was also upgraded to the latest version of SCRAM (the CMS build system). This was a significant effort, as SCRAMV1 has changed significantly from the earlier versions.

Case also supplied help on DDD for users and developers.

Osborne, a developer at CERN employed by NEU, and Muzaffar ran a 3 day workshop on IGUANA, to review the current applications, focusing on the most critical performance issues, 2D visualization and additional services needed. Numerous bug fixes were implemented, and many improvements made, notably in the LEGO plots and track visualization. During this time IGUANA and IGUANACMS were moved to the latest SCRAM version.

### 3.0 Project Office

DOE funding guidance for Tier-1, Grid and CAS efforts stays about stable for FY05. It was not until June that NSF guidance was firmed up. Thus for the first eight months of FY05, until the start of the new funding period, the current effort funded through the NSF stays constant, providing CAS effort and one FTE system management effort at each of the US CMS prototype Tier-2 centers, Caltech, UCSD and U. Florida.

For the DOE fraction of the S&C funds, in the following table we show the funding allocations (BCWS), the allocated effort, the reported effort and the actual invoiced costs (ACWP). We still see a large lag time between the effort being spent and the actual accounting.

<b>FY05 Funds and Costs</b>	FY04 total	<b>FY05 total</b>	FNAL	Universities	Caltech	Princeton	UCD	Wisc	Virginia	Iowa
<b>Funding Allocated</b>										
CAS Personnel	\$1,498k	<b>\$1,059k</b>	\$636k	\$422k	\$0k	\$270k	\$152k			
UF Personnel	\$3,306k	<b>\$4,406k</b>	\$3,664k	\$742k	\$480k			\$171k	\$66k	\$25k
Tier-1 Equipment	\$1,363k	<b>\$1,100k</b>	\$1,100k	\$0k						
Tier-2 Equipment	\$187k	<b>\$0k</b>		\$0k						
Project Office, Reserve	\$667k	<b>\$982k</b>	\$982k	\$0k						
<b>Total Allocated</b>	\$5,471k	<b>\$7,547k</b>	<b>\$6,382k</b>	<b>\$1,164k</b>	<b>\$480k</b>	<b>\$270k</b>	<b>\$152k</b>	<b>\$171k</b>	<b>\$66k</b>	<b>\$25k</b>
<b>Effort Allocated</b>										
CAS FTE years	10	<b>9</b>	4	5	1	2.5	1			
UF FTE years	19.7	<b>31</b>	23	8	3			3.5	0.5	1.25
<b>Effort Spent (9 months)</b>										
CAS FTE years		<b>7.0</b>	4.4	2.6		1.9	0.8			
UF FTE years		<b>23.3</b>	17.0	6.3	2.3			2.6	0.5	0.9
<b>ACWP in k\$, as invoiced</b>										
CAS Labor		<b>\$1,111k</b>	\$710k	\$401k	\$40k	\$225k	\$136k			
UF Labor		<b>\$2,140k</b>	\$1,988k	\$151k	\$0k				\$54k	\$25k
T1 Equipment		<b>\$427k</b>	\$427k	\$0k						
T2 Equipment		<b>\$0k</b>	\$0k	\$0k	\$0k					
Project Office		<b>\$148k</b>	\$148k	\$0k						
<b>Total</b>		<b>\$3,826k</b>	<b>\$3,273k</b>	<b>\$553k</b>	<b>\$40k</b>	<b>\$225k</b>	<b>\$136k</b>	<b>\$0k</b>	<b>\$54k</b>	<b>\$25k</b>

With the start of the new NSF funding period and the start of the DISUN project, the new Tier-2 centers start to receive funding according to a startup plan. The funding allocation of the NSF program is shown in the following table, as approved by project oversight in the PMG meeting. The Tier-2 centers receive funding through the iVDGL project, the DISUN project and the USCMS project (Research Program, RP). The table shows the allocated funds from these sources for the RP period nominally starting in May.

	Caltech	UFL	UCSD	UW	MIT	Purdue	UNL	NEU	UCLA	tot
iVDGL eq	\$20k	\$20k	\$20k							\$60k
iVDGL labor	\$130k	\$130k	\$130k							\$390k
<b>tot iVDGL</b>	<b>\$150k</b>	<b>\$150k</b>	<b>\$150k</b>	<b>\$0k</b>	<b>\$0k</b>	<b>\$0k</b>	<b>\$0k</b>	<b>\$0k</b>	<b>\$0k</b>	<b>\$450k</b>
RP eq				\$180k	\$125k	\$188k	\$188k			\$680k
RP labor					\$125k	\$188k	\$188k			\$500k
RP PO									\$39k	
RP CAS labor								\$370k		\$370k
RP S&C reserve									\$461k	\$461k
<b>tot RP</b>	<b>\$0k</b>	<b>\$0k</b>	<b>\$0k</b>	<b>\$180k</b>	<b>\$250k</b>	<b>\$375k</b>	<b>\$375k</b>	<b>\$370k</b>	<b>\$500k</b>	<b>\$2,050k</b>
Tier-2C eq	\$173k	\$173k	\$173k	\$0k						\$518k
Tier-2C labor	\$203k	\$203k	\$203k	\$272k						\$880k
Tier-2C PO									\$103k	\$103k
<b>tot Tier-2C</b>	<b>\$375k</b>	<b>\$375k</b>	<b>\$375k</b>	<b>\$272k</b>	<b>\$0k</b>	<b>\$0k</b>	<b>\$0k</b>	<b>\$0k</b>	<b>\$103k</b>	<b>\$1,500k</b>
<b>tot funding</b>	<b>\$525k</b>	<b>\$525k</b>	<b>\$525k</b>	<b>\$452k</b>	<b>\$250k</b>	<b>\$375k</b>	<b>\$375k</b>	<b>\$370k</b>	<b>\$603k</b>	<b>\$4,000k</b>

The NSF funding periods cross fiscal year boundaries, while in the past we have reported within fiscal year periods. In the future we will reconcile the different funding periods in our reporting.

A change request for applying \$63k of management reserve to Tier-1 equipment procurements was approved by project oversight. Those funds are for the acquisition of more servers adding additional storage. A change request of \$70k was approved for the University of California, San Diego to purchase new equipment to upgrade their local area network. Change requests for; Virginia Tech to add \$8k, and University of Iowa to add \$75k and Princeton University to add \$95k, all applying FY05 S&C management reserve for additional effort or adjust small cost changes, were also approved.