

US-CMS Milestone Completion Report

Title: Achieving a 25% Computing System at US-CMS Tier-1 and Tier-2 centers

Year: 2005

Expected Completion Date: November 2005

Actual Completion Date: November 2005

Description of Milestone: This milestone describes the goal to reach 25% scale computing centers at the US Tier-1 and Tier-2 centers. In some areas this is 25% of the total capacity and in other areas it is 25% of the total complexity. The CPU is typically given as 25% capacity because CPU is on a flatter improvement projection and CPU is needed for simulation. Disk storage is typically listed in complexity. Disk storage is on a steeper performance ramp so waiting to purchase is a larger savings in cost.

The goal of the scaling milestone is to maintain the facilities on an achievable operations ramp that allows procurement, commissioning and deployment. We wish to continually increase the scale of facilities and facility services, but on a procurement path that allows discovered problems to be found and fixed. There is always a trade off between bringing equipment into service for operations and buying late. In US-CMS we think we have found a reasonable balance, though the procurement ramp is steep in some places.

Conclusions:

The Tier-1 Center

The FNAL Tier-1 center finished 2005 by completing the first year of a three-year procurement cycle in preparation for the beginning of the experiment. The processing and storage milestones of 25% capacity and complexity were met on time using the budgeted deployment effort.

In 2005 280 dual AMD Opteron 268 worker nodes were procured in addition to the 180 dual Intel Xeon nodes already in operation. This raises the total facility capacity to approximately 1000kSI2k of processing spread over 1000 batch slots. The new nodes have 1GB per processor, which is currently enough to support the CMS application. In the future it may be necessary to procure systems with 2GB per batch slot.

Also in 2005, 3 NEXSAN disk enclosures with 13TB of usable space each were deployed within the dCache system. Each enclosure is served by 2 data serving hosts. This raises the total deployed disk cache at FNAL to approximately 100TB. The NEXSAN systems are expected to be the primary disk device procured in 2006 to satisfy the 600TB goal. The total number of disk servers at FNAL is approximately 20, which easily satisfies the 25% complexity milestone.

The grid services for interfacing to processing, monitoring, information providing, and storage have been deployed and supported. The scale at which they are operated is not

yet 25% of the experiment goals, but the basic service functionality is beginning to mature.

US-CMS is operating the Tier-1 procurement, deployment, commissioning, and operations with approximately 5FTE of effort. This does not include grid and facility interface development, the facility service development, or the effort spent in user support.

The Tier-2 Centers

For the Tier-2 centers the CPU goal was estimated at 60-80 dual CPU worker nodes constituted about 25% of the system. In the case of disk-based storage, the Tier-2 goal was to address and manage the complexity of running large storage arrays, rather than providing sheer capacity. Operating 200 TB of disk will be one of the major challenges of Tier-2 operations, and even operating 10% of that capacity provides a significant challenge today. Thus, we set a goal of operating at least 20 TB of disk at each site for this milestone. (The amount of disk actually available for storage is often reduced by choices of configuration, e.g. RAID arrays.) By the end of the quarter, all seven sites had purchased the hardware needed to meet these goals, and all but MIT (which received funding later than the others) had deployed it. While all sites were asked to meet this specification, each chose to go about it in a way that best matched their own local environments, so that they could optimize available resources and existing relationships with IT organizations and computing vendors. Here is a description of the computing systems at each of the Tier-2 sites:

Caltech: This site bought 30 dual dual-core 2.2 GHz Opteron nodes, each with 4 GB of RAM, this year, and already had 32 Intel Xeon 2.4/2.8 GHz nodes, each with 1 GB of RAM. All nodes were bought from Acme Micro Systems. This puts 62 nodes with 184 CPU's at the disposal of the Tier-2 program. Each of the 30 new nodes carries four 300 GB SATA drives that are part of a RAID device made with a four-port 3ware 9500 4LP SATA controller. Each node contributes approximately 1 TB of storage space into a dCache pool. An additional 10 TB of dCache space is available from the pre-existing cluster.

Florida: This site serves a community conducting iVDGL/GriPhyN and Ultralight R&D efforts, local CMS analysis, and CMS simulation production. The Tier-2 facility now has 31 TB of RAID storage, with 25 TB configured in dCache pools, and the balance serving as scratch and home space. In the fall, Florida purchased 84 nodes of dual-core dual Opteron 2.2 GHz servers, configured with 4 GB RAM and 0.5 TB of disk per node. 60 of these are allocated primarily to CMS activities, 20 for general OSG use, and the remainder for an interactive analysis farm. Dedicated CMS resources amount to 511 kSI2K. All of this equipment was bought with iVDGL/GriPhyN and local DOE funds; no Tier-2 funds have yet been spent. The systems were purchased from Rackable, and Florida has been delighted with that vendor.

MIT: This site will build a linked cluster of Tier-2 machines plus nodes for CMS heavy-ion physics and for CDF which are funded by other sources. Three air-cooled racks have

been installed, and water-cooled racks are being evaluated for future expansion. Seven servers, all dual Opteron 2.2GHz machines, were bought to host various services. One with 7.2 TB disk is an NSF server, and one serves as the dCache DB server, and the others are used for the remaining services. There are 20 worker nodes, all with dual dual-core Opteron 1.8 GHz CPU's and 120 GB system disks and 2.4 TB data disks. The data disks are configured in a RAID array with an effective disk space of 1.9 TB per node, for a total of almost 40 TB across all nodes. The worker nodes alone provide about 120 kSI2K of processing power.

Nebraska: This site currently has 16 TB of SATA drives in 3 FC enclosures, purchased from Zzyzx at a cost of approximately \$2000/TB. This disk is served by four dual Opteron servers. Nebraska also has 16 nodes with 400 GB SATA drives in them that are further available to dCache. Four 4 TB internal servers (dual dual-core Opteron) have been purchased from TeamHPC for a total of approximately \$2000/TB (disk + server). All are or will be on a gigabit switch. This will result in about 40 TB total storage, and allow testing of three distinct approaches to maximizing performance/price. In addition, the main computing cluster (red.unl.edu) was purchased and installed in summer 2005. It consists of 64 dual dual-core 275 Opteron nodes with 4 GB RAM and dual gigE per node. It was purchased this from TeamHPC for a total of \$260K. The system has been stable and performed well. This system should be of a capability of roughly 360 kSI2k.

Purdue: This site began the year with 50 dual-processor Dell 1750 systems, which carry 3.06 GHz processors. Purdue purchased an additional 64 Dell 1425sc systems with 3.2 GHz processors. In total, this is about 295 kSI2K of computational resources. The storage systems consist of six Apple Xserve RAID enclosures served by Dell 1850 systems, plus additional Dell machines to support the necessary software infrastructure. A total of 35 TB of storage is available to the project, of which about 10% is reserved for applications and user directories.

UC San Diego: This site hosts a major computing cluster for CDF, and in doing so has demonstrated its ability to run a large production facility with close to uninterrupted service, as will be required for CMS. As CMS cannot yet saturate the available computing power, UCSD has focused on maintaining their CDF operations while also establishing the infrastructural hardware required for the broad range of expected Tier-2 services. Their cluster comprises 72 dual CPU compute nodes, and close to 40TB of disk space. The bulk of the disk space is organized in resilient dCache, plus a 5TB fileserver for user space, and software installations. In addition, UCSD has 20 nodes for infrastructure, ranging from CMS tools like phedex and pubDB, to OSG infrastructure like a set of OSG CEs (testbed and production), SE, GUMS, Clarens server, 8 nodes for dCache infrastructure, an NFS server, the Rocks headnode, an interactive login node, two versions of Edge Service Framework, as well as a CDF headnode. In addition, we host FroNtier squids for both CDF and CMS, as well as a SAM installation for CDF.

Wisconsin: This site is part of a campus grid project called GLOW, which has already amassed a substantial set of resources. These amount to about 45 TB of storage space on

Apple Xserve RAID systems and access to about 300 CPU's on average. Given these existing facilities, Wisconsin purchased a small number of new machines this fall. These include six dual 2.8 GHz Xeon machines with 4 GB of memory, an additional 7 TB Apple RAID with a Xeon system as host, and seven other servers for service tasks. The site has also placed an order for 46 dual-dual 1.8 GHz Opteron systems, with 45 TB of PATA disk; these should be commissioned in January 2006.

In total, the seven Tier-2 sites have acquired 1628 CPU's (many in dual-core Opteron processors) and 316 TB of disk for data storage in dCache pools. Thus, an "average" Tier-2 site is running the equivalent of 116 dual-CPU systems and 45 TB of disk -- easily exceeding our benchmark for the capacity/complexity milestone.

Based on the size of the facilities that we want to operate, we anticipate that each site will require two FTE for operations. At the moment, these are the estimated number of FTE that each site is currently using:

Caltech -- 2.0
Florida 1.75
MIT 2.0 (plus small amounts of help from various others)
Nebraska 1.5
Purdue 1.75
UCSD 2.0
Wisconsin 2.0

Based on their experience so far, the staff at the sites believe that two FTE are sufficient for operations. However, some have expressed concern that it is just enough, with little room for error, and others are concerned that the funds available will not be enough to hire two FTE in expensive labor markets.