RUN II COMPUTING

Amber Boehnlein, FNAL, Batavia, IL 60510, USA for the CDF and DØ Collaborations

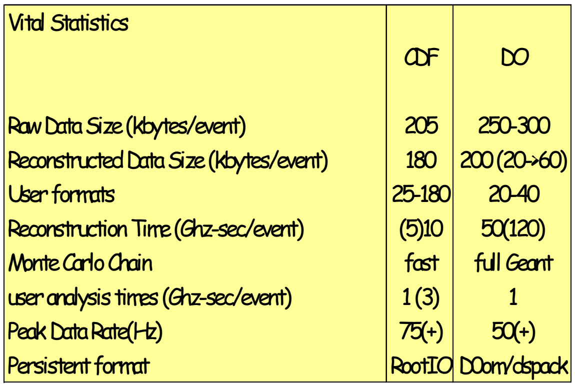
Abstract

In support of the Tevatron physics program, the Run II experiments have developed computing models and hardware facilities to support data sets at the peta-byte scale, currently corresponding to 500 pb-1 of data and over 2 years of production operations. The systems are complete from online data collection to user analysis, and make extensive use of central services and common solutions developed with the FNAL CD and experiment collaborating institutions, and make use of global facilities to meet the computing needs. We describe the similarities and differences between computing on CDF and D0 while describing solutions for database and database servers, data handling, movement and storage and job submission mechanisms. The facilities for production computing and analysis and the use of commodity fileservers will also be described. Much of the knowledge gained from providing computing at this scale can be abstracted and applied to design and planning for future experiments with large scale computing.

DESCRIPTION OF COMPUTING

The fundamental features of Run II computing are similar for CDF and DØ. Events are collected on online systems and the raw data is transferred into robotic storage, and from that point handled using data management services. The primary reconstruction of the data occurs on farms of LINUX compute nodes. Monte Carlo data is produced for both experiments at remote institutions and is stored on tape at FNAL. Both experiments support large central analysis (CA) clusters of LINUX PCs (called CAF for CDF and CAB for DØ) that has access to disk cache and user controlled space as well as supporting analysis and production activities at non-FNAL institutions.

The experiments differ in detail in terms of the amount of time required to perform the reconstruction of data. In the case of DØ, the reconstruction program is relatively slow, caused in part by because the detector has a longitudinally segmented calorimeter and because the tracking volume has relatively few layers, and relatively high occupancy. In addition, the DØ Monte Carlo generation uses full GEANT, reconstruction and trigger simulation. Because of this, computing for the DØ experiment has a relatively strong focus on production activities, performing as many tasks as possible on the production platforms as common services. Such tasks include re-reconstruction, performed at remote sites and post-processing fixing. The primary analysis format used by DØ is the 30 kb/event "thumbnail", which contains final physics objects and some supporting information. CDF, in contrast, has a relatively fast reconstruction code, and a parameterized Monte Carlo and users frequently access the full event data (150 kb/event), and re-reconstruct or otherwise manipulate the data on the analysis platform, leading to a more analysis oriented focus. Individual groups on both experiments tend to do final analysis on customized root based formats. It should be noted that as Run II continues, both experiments are constantly evaluating their needs with respect to the data formats and simulations. DØ has developed a fast parameterized MC, is working aggressively with the FNAL computing division on speeding up the reconstruction and is implementing a root based format that is hoped to find wide spread common use. CDF is considering implementing a smaller (~60 kb/event) event format.



Vital Statistics	CDF	DO
Raw Data Size (kbytes/event)	205	250-300
Reconstructed Data Size (kbytes/event)	180	200 (20->60)
User formats	25-180	20-40
Reconstruction Time (Ghz-sec/event)	(5)10	50(120)
Monte Carlo Chain	fast	full Geant
user analysis times (Ghz-sec/event)	1 (3)	1
Peak Data Rate(Hz)	75(+)	50(+)
Persistent format	RootIO	DOom/dspack

Figure 1 Defining Characteristics of CDF and DØ data.

Farm production

Each experiment has collected over 1 billion raw data events and reconstructs events for analysis within a few days of data collection. CDF processes data, uses information to get calibration data and then reprocesses, and recently processed 400 million events in six weeks with a peak daily production on the CDF farm of 30 Million events (as shown in Figure 2). DØ's peak production on the FNAL farm is 25 Million events in a day, but as the reconstruction time per event for DØ is a strong function of instantaneous luminosity, 15 Million events per day is more common. In autumn 2003, DØ reprocessed 500 Million events, including processing 100 Million events at remote farms including WestGrid in Canada, NIKHEF, IN2P3, UK sites and GridKa. Both experiments produce Monte Carlo events at remote facilities. Figure 3 shows the production for the DØ

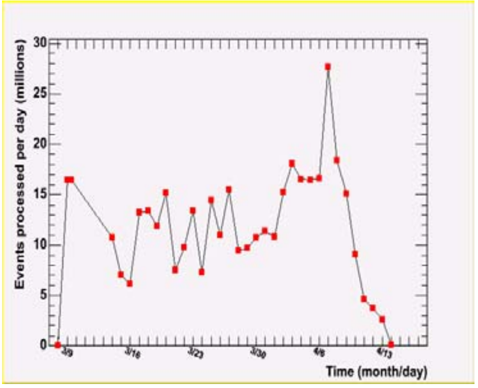


Figure 2 Reconstruction processing for CDF by day, showing a peak throughput of 30 M.

experiment, divided by production site for 2004. For the CDF experiment, 60% of the MC is generated at a facility in at the University of Toronto and 30% in the United Kingdom with the rest contributed by facilities at University of California at San Diego and in Italy.

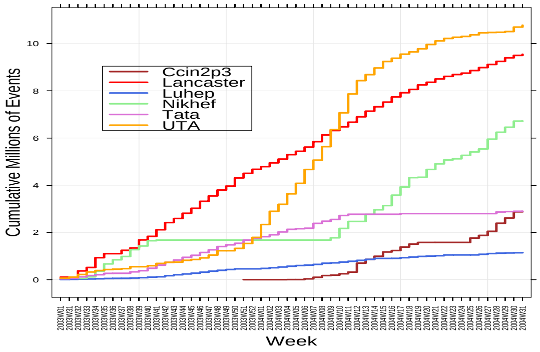


Figure 3 Monte Carlo event production for the DØ experiment integrated for the past year. Contributions from different sites is shone in the various colors

Data Handling Services

Sequential Access via Metadata (SAM) services were developed as part of the CD-Tevatron Joint project with initial design work taking placed 7 years ago. SAM has been in production for DØ for over 4 years. CDF has been using these services for over a year on the remote analysis facilities, and is currently deploying SAM on the CDF analysis facility (CAF) and reworking production processing scripts to use SAM for data management for farm production.

Services provided include comprehensive meta-data to describe collider and Monte Carlo data, consistent user interfaces via command line and web interfaces, local and wide area data transport mechanisms, providing transparent global access to the data. SAM supports adapters to most common batch systems. (PBS, Condor, lsf, site-specific batch systems)

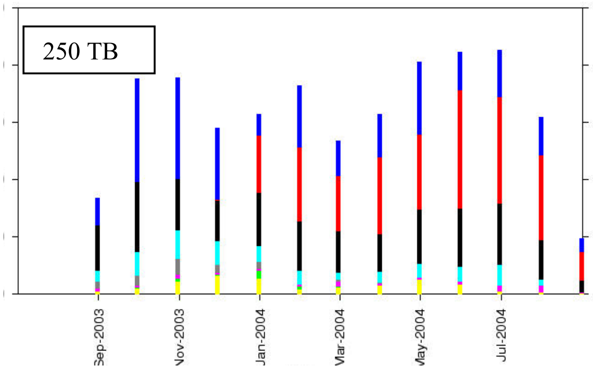


Figure 4 Monthy usage of SAM at DØ in terabytes for the past year. Blue, red and black shows the usage on the central analysis platforms. The yearly consumption was 50 Billion events in 2.1 petabytes. The top of the vertical scale is 250 TB

With the addition of CDF and other experiments to the SAM project, the second generation development has benefited from experience and new perspectives leading to extended and improved functionality. To that end, the metadata schema and database server were updated in 2004 , and there has been an introduction of an interface to dCache in addition to the native SAM cache mechanisms. A second generation Monitoring and Information Server (MIS) prototype moves away from log file monitoring, providing more real time monitoring enabling better diagnostics and reducing the burden on the experts as routine operations can be supported by experiment shift crews.

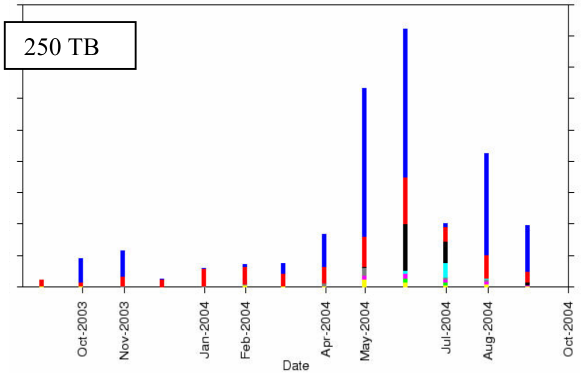


Figure 5 Monthly usage of SAM at CDF in terabytes for the past year. Blue shows CAF station testing, and red shows the Gridka usage. The top of the scale is 250 TB

From an operational point of view between Oct 2003-Sept 2004, DØ has consumed 2.1 petabytes of data, and 50 billion events through all SAM stations (shown in Figure 4), while CDF has consumed 1.5 PB in 12B events (Figure 5). On the newer analysis nodes (called CABSRV1), DØ has recently deployed 20 TB of cache disk in a pass through mode, which has reduced the amount of data pulled in from tape for analysis. The current process wait times can be seen in Figure 6 in which the cached files are delivered quickly, 60% within 20 seconds although presumably faster as the 20 seconds is the precision of the information and practically all requests from tape are satisfied within 5 minutes. As a note, the performance of the system was quite robust; with inadequate cache, 2/3 of the requests for files involved staging files form tape; even so, 90% of the tape requests were satisfied with 10 minutes.

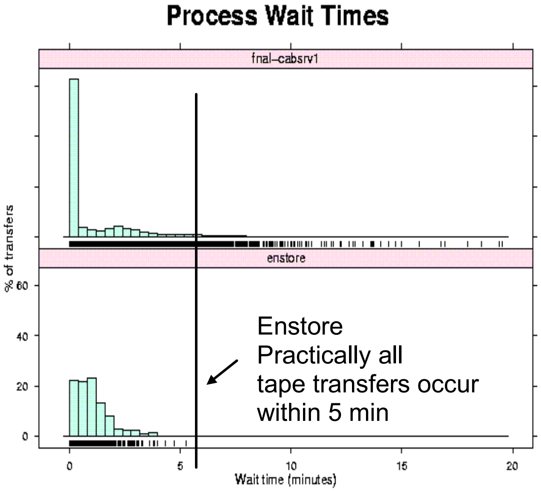


Figure 6 Process wait times for CABSRV1. The upper plot is "intra-station" (cache) transfers, and the plot is transfers from tape. The line shows the 5 minute mark, and the vertical scale is the % transfers, with 60% max.

In addition to the data handling services, SAMGrid provides Job and Information Monitoring (JIM) services as well. These services have recently been deployed in production for the DØ MC production, (see Figure 7) enabling a common submission site controlling 10 execution sites, and are being tested for reconstruction. JIM is built on GLOBUS, and was migrated to VDT.

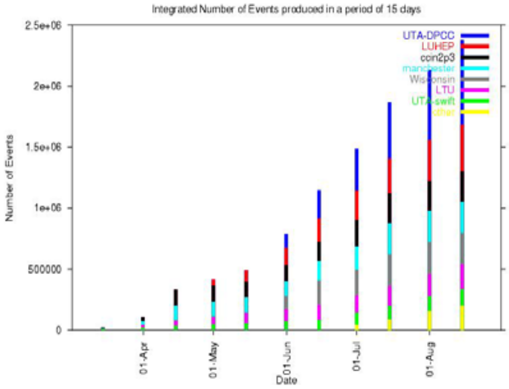


Figure 7 Commissioning of SAMGrid for MC production from April-August 2004. The vertical scale is number of events produced, and each color represents a new execution site. At the time, 2.5 Million events had been generated.

Storage

The Run II experiments have nearly 2 petabytes of data in robotic storage. All of the CDF data is on 9940b media in an STK powderhorn silo. In addition to STK, DØ also maintains an ADIC AML2 robot with LTO I and LTO II drives. The data movement to tape has exceeded 20 TB per day at peak, and the June 2004 daily storage is shown in Figure 8. The number of mounts per day for DØ on the ADIC robot is shown in Figure 9, where 3000 tapes are routinely mounted in a day. The tape system has proven to be remarkably robust, and the loss of data due to problems in the robotic system is negligible for both experiments.

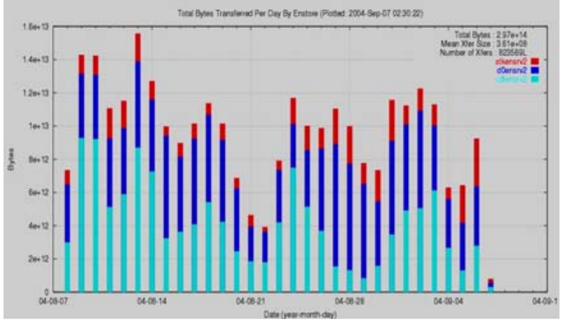


Figure 8 Daily data movement into robotic storage for June 2004. The light blue histogram corresponds to CDF, the dark blue to DØ, and red to all other users, top of scale is 20 TB.

Dcache, developed jointly by DESY and FNAL is in use at FNAL for CDF, CMS, D0, LQCD, and as a front end to the general robotic system. WAN (gridftp) and LAN (dccp) transport mechanisms are supported, as well as an SRM grid interface. The CDF experiment makes extensive use of dCache, with 150 TB of disk space deployed. Typically, 25 TB of data with no or few errors is moved by dCache per day with a peak of 60 TB—which was the standard daily movement with the addition

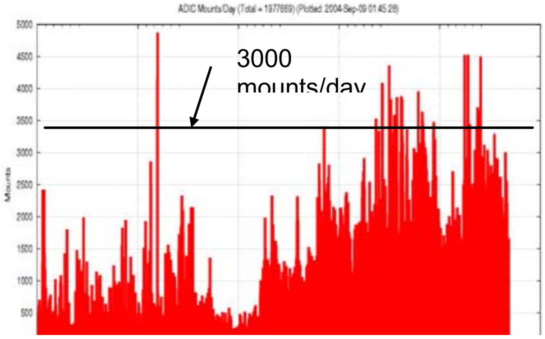
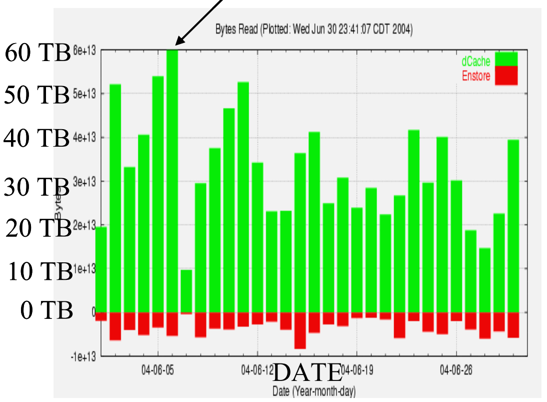


Figure 9 Number of mounts per day for the DØ ADIC robot. Due to this mount rate, the second arm of the robot has been activated.

of a SAM test that also moved 30 TB. The June 2004 daily dCache movement for CDF is shown in Figure 10.

60 TB read by CDF clients on 06 June 2004
(no discernible file delivery errors)



CDF dCache – June 2004
nBytes Read Per Day

Figure 10 The data movement for dCache for the CDF CAF for June 2004, showing a daily peak of 60 TB in green, the red shows the data transferred from the tape system for the same day.

Analysis systems

Both experiments support 200 users on the FNAL based analysis systems at peak times. The types of analysis range from reconstruction and track fitting using the full framework (typically supporting B physics) to root-tuple based analysis, and some user MC generation. DØ supports post-processing "fixing" of processed data as a common activity, although some of that activity is moving to the production platform. Both CDF and DØ procure the worker nodes used in the analysis clusters, configure them, and burn them in prior to making them available for general use, consistent with the best practices of the FNAL Central Services farm group. This procedure identifies problems in a timely way, and typically leads to smooth integration of increments to the clusters.

The CDF Analysis Facility (CAF) is a LINUX based system with 3.25 THz of worker nodes and 150 TB disk cache and 150 TB of group controlled space. There are two sets of systems, one running the Farm Batch System (FBS), the other running Condor, however there is a common user interface. 83% of the user analysis jobs use an average of 1Ghz sec/event, the other 17% have a mean of 3 Ghz sec/event. The CAF system tends to be fully utilized. A snapshot of the processes on the FBS side of the CAF for the past year is shown in Figure 11.

DØ is phasing out an SGI Origin 2000 that is called Central Analysis, and there are two linux systems, running PBS as the batch system, referred to as the Central Analysis Backends (CAB and CABSRV1). The CAB system has 2 THZ worth of worker nodes and 21 TB of fileserver disk deployed as SAM Cache on CABSRV1 with 20 TB local disk cache on the worker nodes and 70 TB user controlled space deployed on the desktop cluster. As can by seen in Figure 12, the use on the system has increased dramatically in the past 18 months when the original worker nodes were put in service. The queues correspond to long and short SAM jobs and for non-SAM (noted as Medium) usage where non-SAM batch usage is typically the generation of MC test samples or root based analysis. In January 2004, the new worker nodes were made available to users, and a sharp turn on of usage can be seen corresponding to a heavy analysis period. The capacity of the system effectively doubled and the turn on was completely smooth for scaling up the batch system and the demands on the SAM system. This can be seen in Figure 13 as well which details the number of events per week consumed on various analysis platforms. Dark purple shows the consumption on CABSRV1, showing the fast turn on in January 2004. The peak weekly usage on CAB (shown in light purple) was 1.4 B events. Another peak is seen in the CABSRV1 plot, at the end of the period shown, corresponding to deployed fileserver cache mentioned in an above section. As a note, green shows the consumption at Gridka, blue on DØmino and yellow on CLuEDØ. Analysis system usage can vary, but in general, 1 billion events are processed per week, at an average of 1 GHz*sec/event.

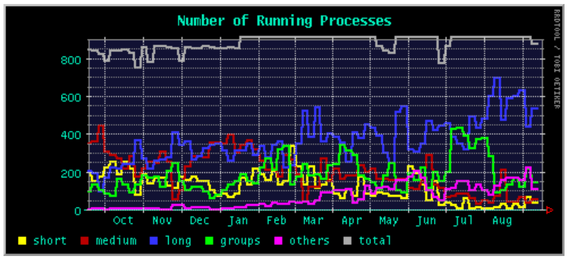


Figure 11 The number of user processes running on the FBS side of the CAF for the past year. The vertical scale is number of running processes. The system is typically in full use. The colored lines indicate various batch queues while the white line at the top indicates the sum of the running processes.

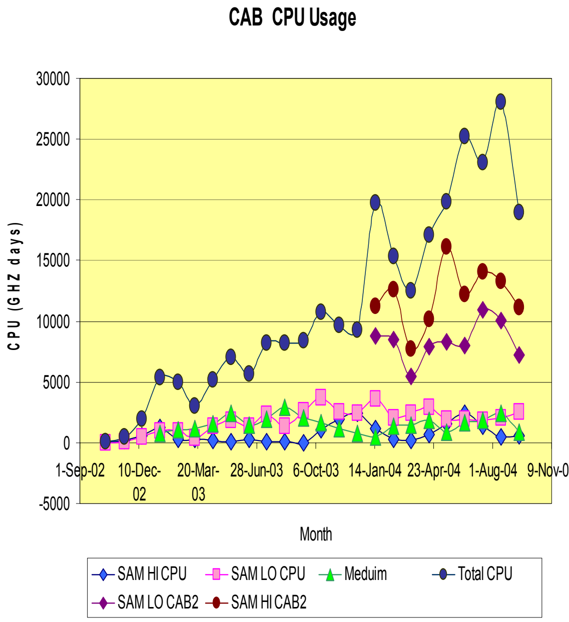


Figure 12 Use of the DØ analysis clusters CAB and CABSRV1 from September 2002 Through September 2004, plotted monthly. The vertical scale is in GHz days/month. The colored lines indicate various queues with the total shown in dark blue. Note the introduction of CABSRV1 in January, 2004.

Both experiments make extensive use of remote facilities. There are 31active DØ SAM stations and 24 CDF SAM stations outside of FNAL, and 35 % of the CDF analysis resources are available outside of FNAL, typically in the form of dCAFs (de-centralized Analysis Facilities. For DØ, 10% analysis projects are run on the remote stations, however, for DØ, the emphasis is on production activities. For CDF the scale of offsite analysis resources available is roughly 30% of the FNAL CAF.

I would like to extend my thanks to all of my Run II colleagues who contributed to this talk. By necessity, this talk focused on some areas, ignored many, and did justice to none. For more information, please see the papers in these proceeding co-corresponding to posters and talks that relate to Run II computing

·Databases and FroNtier [204] [205]
·Networking [359] [369]
·Enstore and dCache [107] [190] [464] [471]
·SAMGrid—performance, monitoring, Metadata [335] [451] [455][462][400] [38][293][481]
·CDF Posters –Monitoring the CAF [390] [484]
·DO Posters
reprocessing [362], Virtual Center [372]
Interfacing to other Grids [55] [58]

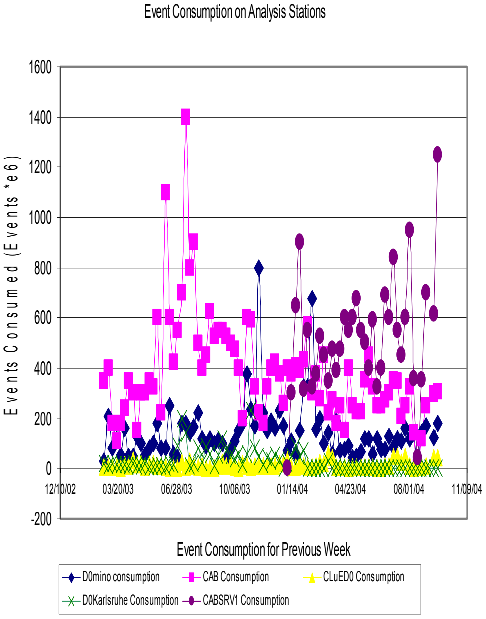


Figure 13 Weekly consumption of events on DØ analysis platforms in millions of events. Light purple is CAB, dark is CABSRV1.