CDF/DOC/COMP_UPG/PUBLIC/7290

CDF Run II Annual Computing Plan and Budget, FY-05

Stefano Belforte, Ian Fisk, Suen Hao, Jason Harrington, Liz Sexton–Kennedy, Art Kreymer, Ashutosh Kotwal, Stephan Lammel, Elliot Lipeles, Dmitrie Litvinsev, Rick Snider, Rick St.Denis, David Tang, Andreas Warburton, and Steve Wolbers

Oct. 21st, 2004
Version 2.02

The CDF Run II computing plan is updated annually to incorporate changes in requirements, luminosity forecasts, technology development, and data analysis patterns. This note contains the edition for fiscal year 2005. Estimates of FY04 are replaced with the actual numbers and projections for the following three fiscal years, FY-05, FY-06, and FY-07 are presented. The required budget to meet the physics goals of the experiment is estimated.

Contents

1 Introduction

Run II of the Fermilab Tevatron started in March of 2001. So far the collider has delivered over $600\,\mathrm{pb}^{-1}$ of proton antiproton interactions. The CDF experiment is recording the most interesting of those interactions, reconstructs the events, and analyses them. In fiscal year 2004 CDF almost doubled the recorded luminosity. Compute systems need to keep up with the ever increasing data and analysis demands. The capacity of the system is increased when needed as to benefit most from technological advances.

We have compared the plan of this fiscal year (as projected last year) with the actual computing upgrades made during the year. We include what we learned from this into the plan of the next three fiscal years. We have updated our computing plan, estimated data storage, processing, and analysis requirements, developed a procurment plan, and estimated the budget to implement it. However, our understanding and projections of the analysis needs are quite incomplete and while we are committed to the long term plan described in this document, the individual projections should be taken with a grain of salt. We will update the CDF Run II computing plan again in a year or before, shall significant changes occur.

1.1 Requirements Model

For the computing planning of fiscal year 2005 we use the requirements model [1] developed for the FY-04 planning. The study uses three models: a"baseline" update of an old model [2], and a "single-user" and a "multi-user" model that introduced a new scaling behavior to the requirements. The CDF requirements we will use for our budget and procurement plan come from the "multi-user" model. Updating of the parameters used in the models as well as the models themselves has not been possible this year. While we have good usage and utilization information for the interactive system, usage statistics has only recently been collected for the CAF batch system and the records of the last months are empty due to a software glitch. We will present here some updated tables, figures, and text from the FY-04 study.

CDF is increasing its online event logging capability. The upgrade has a significant impact on offline computing requirements. It is designed to allow CDF to avoid deadtime at high luminosities and to maximize the physics program of the Tevatron by writing additional data that will increase the precision of many measurements. One particular measurement driving the upgrade, B_s mixing, is one of the most challenging and important that CDF is expected to make. The upgrade includes two changes that will increase the event logging rate from the detector: implementing raw data compression in level-3, and an upgrade to the data logger bandwidth from $20\,\mathrm{MB/sec}$ to $40\,\mathrm{MB/sec}$. In summer of 2004 a first step on raw data compression was made, reducing the event size by about 10%. Figure 1 shows the raw data event size as function of instantaneous luminosity with this first compression. Before the Tevatron shutdown in August 2004, a test was made with a parallel logger reaching $35\,\mathrm{MB/sec}$. The upgrade anticipate additional compression and an increase to $60\,\mathrm{MB/sec}$ in FY-06.

We employ two basic approaches to arrive at estimates for the various computing resources required by the CDF experiment. We assume the analysis CPU requirements factorize into two types of basic analysis behaviors. The first requirement is a high p_{T} dataset

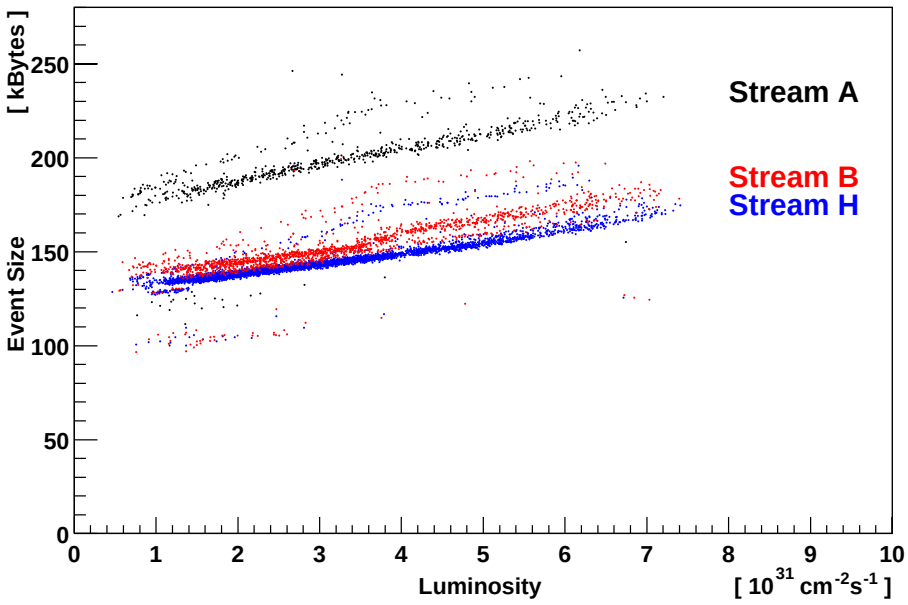

Figure 1: Raw data event size after summer 2004 for the express, A, (largest event size), high-p_{T} lepton, B, and two-track, H, (smallest event size) streams.

analysis that scales with integrated luminosity, call them "dataset-A" requirements. The second requirement is for analysis of extremely large datasets, to study bottom quarks and other high statistics physics, call them "dataset-B" requirements, with requirements that scale with the total events logged to tape. Here we assume that 400 nb of level-3 cross section goes into dataset-A, while the balance of the logging bandwidth goes into dataset-B. We then calculate the resources required to allow 200 users to analyse 5 nb of dataset-A in a single day, and 15 users to process all of dataset-B over the course of 25 days. The dataset-A assumptions are an update to what was assumed in [2] to model our requirements earlier. The dataset-B requirements have been added to model the load on our systems caused by the additional events that are logged starting in FY-04 due to the decreased event size and the anticipated increase in bandwidth capabilities of the data logger.

Some of the basic assumptions used in the model calculations are shown in Table 1. Included are the integrated luminosity delivered by the Tevatron, average initial luminosity of a store, the bandwidth of the data logger, average event size, and peak and average data recording rate. The Tevatron luminosity values correspond to the "design" values [3] quoted by the Beams Division.

The experiment rarely operates at the peak event logging rate. More typical values are 60% to 80% of the effective peak rate. We will assume that the average logging rate is about 70% of the effective peak rate.

The event size is the measured average during a set of runs in 2004 at luminosities around $6 \times 10^{31}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. We have ignored a known luminosity dependence in the size that is ex-

Fiscal Year	03	04	05	06	07	08	09
Delivered Luminosity (1/fb)	0.2	0.35	0.6	1.5	1.7	2.0	2.1
Integrated Luminosity (1/fb)	0.33	0.68	1.2	2.7	4.4	6.4	8.5
Initial Luminosity ($10^{31}/\mathrm{cm}^2\mathrm{s}$)	5.5	6.2	10.5	22.4	27.5	27.5	27.5
Data Logger Bandwidth (MB/s)	20	20	35	60	60	60	60
Average Event Size (kB)	220	150	150	170	170	170	170
Peak Event Rate (Hz)	80	130	230	360	360	360	360
Average Event Rate (Hz)	50	80	170	250	250	250	250

Table 1: Operating parameters and basic assumptions used in the requirements model as function of fiscal year.

pected to produce an approximately linear 40% increase in the data size between 10^{31} and $10^{32}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. For FY-05 the event size is still based upon the raw data compression factor observed in run 167024 during which the trigger table with data compression and bank dropping was tested.

2 Computing and Analysis Model

A conceptual view of the major computing elements and data-flow at CDF at FNAL is pictured in Figure 2. Although incomplete, Figure 2 presents some of the main themes of CDF computing. Raw data is acquired online and is written to a *write disk cache* before being archived in a *tape robot*. The raw data is read by the *production farms*, either by triggering a cache-to-cache copy or directly from the tape robot, where it is reconstructed and the resulting *reconstructed* data is written back to the tape robot. In both cases, there are caches that decouple the *production farm* from the tape robot. The production farms use calibration constants replicated from the online *database* to the offline database and any other replicas (all shown as one database for simplicity). The reconstructed data is read primarily by *batch CPU* via a *read disk cache*. Some of the reconstructed data, and the majority of secondary datasets from the reconstructed data, are also stored in the disk cache with relatively large cache lifetimes where they are accessible by the batch CPU. The batch CPU produces secondary datasets and root *N-tuples* and writes them to output disk and also the tape robot via other *write disk caches* (distinct from read disk caches). The batch CPU makes extensive use of the offline database and its replicas. The batch CPU also analyzes the N-tuples on the static disk. *Interactive CPU* and *user desktops* are used to debug problems, link jobs, and send them to the batch CPU which is the workhorse of CDF analysis. The user analysis farm is exclusively batch. Users desktops can also obtain data from the tape robot via read disk caches, write them back to the tape robot via write disk caches (not shown), and transfer N-tuples and results back to their desktops from the interactive and batch CPU. User desktops and interactive CPU make use of the offline DB and its replicas.

In this model physics groups are encouraged to utilize the batch CPU to produce secondary datasets and write them to static disk and the tape robot. Users are encouraged to produce N-tuples on the batch CPU and transport them back to the desktop for further analysis, but also have the option of utilizing the batch facilities for subsequent re-analysis of the N-tuples. Users have access from their desktops and the interactive CPUs to the datasets on the CAF output disks. The interactive CPU provides a controlled environment for debugging and job submission. The upgrade of the interactive CPU is discussed in Sec. 3.

Offsite resources contribute to this picture by adding additional CPU and disk caches. However, we do not expect to be using offsite tape archiving facilities at this point. The tape robot at FNAL thus serves the role of central storage facility for all official CDF data. In contrast, we do not require a copy of user level data to be stored centrally at FNAL, nor do we require tape storage prior to general open use of the data in CDF. More details on our future vision of bluring the distinction between offsite and onsite computing are discussed in Section 9.5.1.

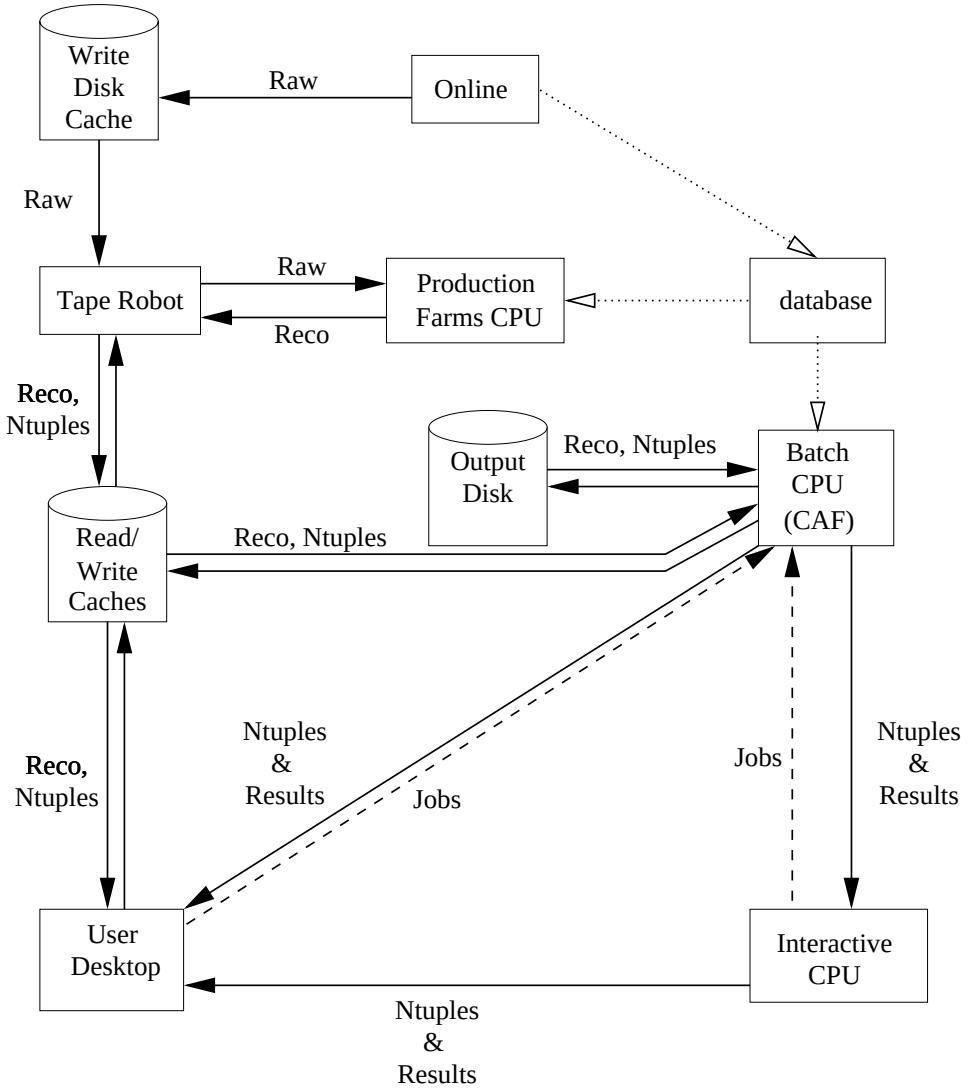

Figure 2: A simplified picture of the CDF Computing Model. Major computing elements in boxes, raw and reconstructed data flow indicated by solid lines, database constant flow indicated by dotted lines, and job flow indicated by dashed lines.

3 Interactive Systems

The CAF discussed in the following section is a batch computing engine that satisfies the majority of CDF's CPU and file serving needs. The CAF is supplemented by an interactive computing system. As of July 2004, the CDF interactive computing system consisted of fcdfsgi2, a 64 x 300 MHz, SGI SMP with roughly 45 TBytes of disk; fcdfhome, a NetApp serving 269 GBytes of disk for user home areas; and fcdfspool, a NetApp serving 645 GBytes of disk for user spool. Two 8 x 700 MHz Intel SMPs running Fermi Linux 7.3.2, fcdflnx2 and fcdflnx3, provide interactive computing for users and offline operations within a reference environment. About 380 Linux/Intel computers in the CDF trailers provide the bulk of interactive computing capacity for the experiment. Node cdfsga, a 28 x 194 MHz, SGI SMP with roughly 3 TBytes of disk, continues to be used for Run I analysis.

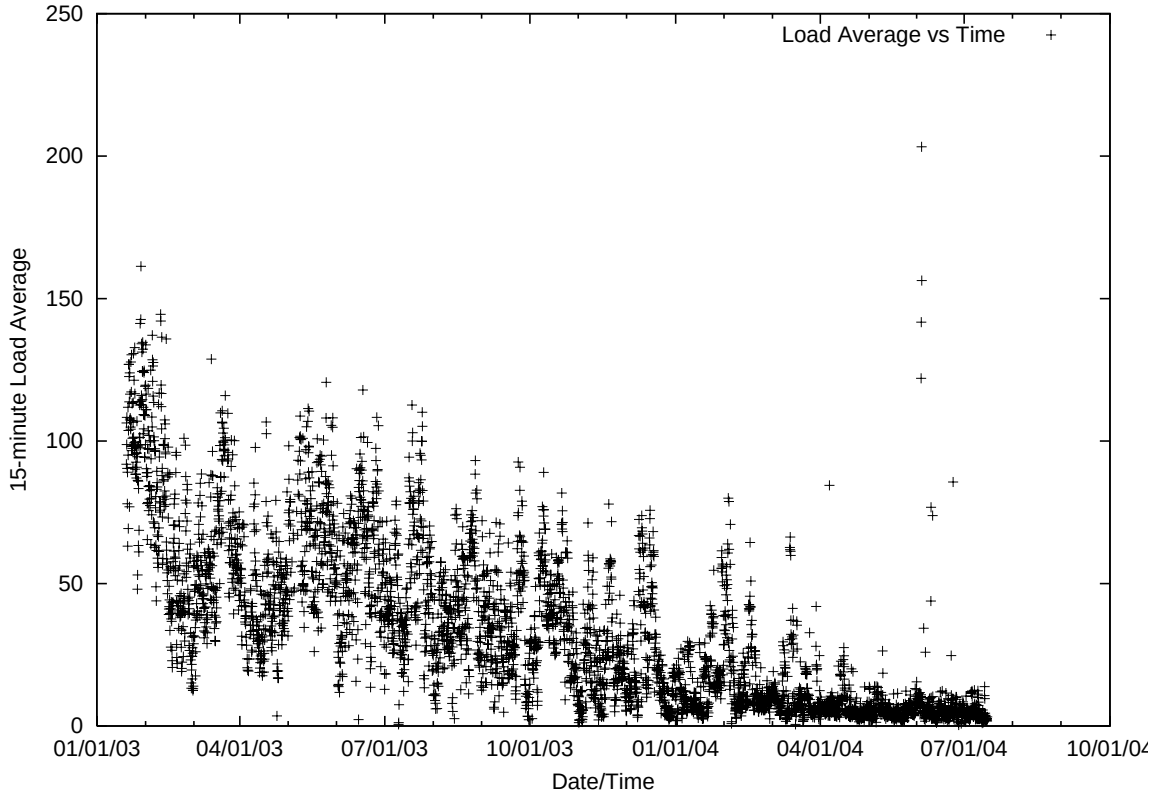

Figure 3: CPU load of fcdfsgi2.

The aging fcdfsgi2 and cdfsga cannot be maintained indefinitely. Further, their relatively slow processors are increasingly avoided by CDF users. Node cdfsga serves a unique role supporting Run I analysis on the IRIX 6.2 operating system on which Run I analysis code functions. It was previously estimated that support would continue through FY04; however, it does not appear that the machine will be decomissioned until FY05. Node fcdfsgi2, on the

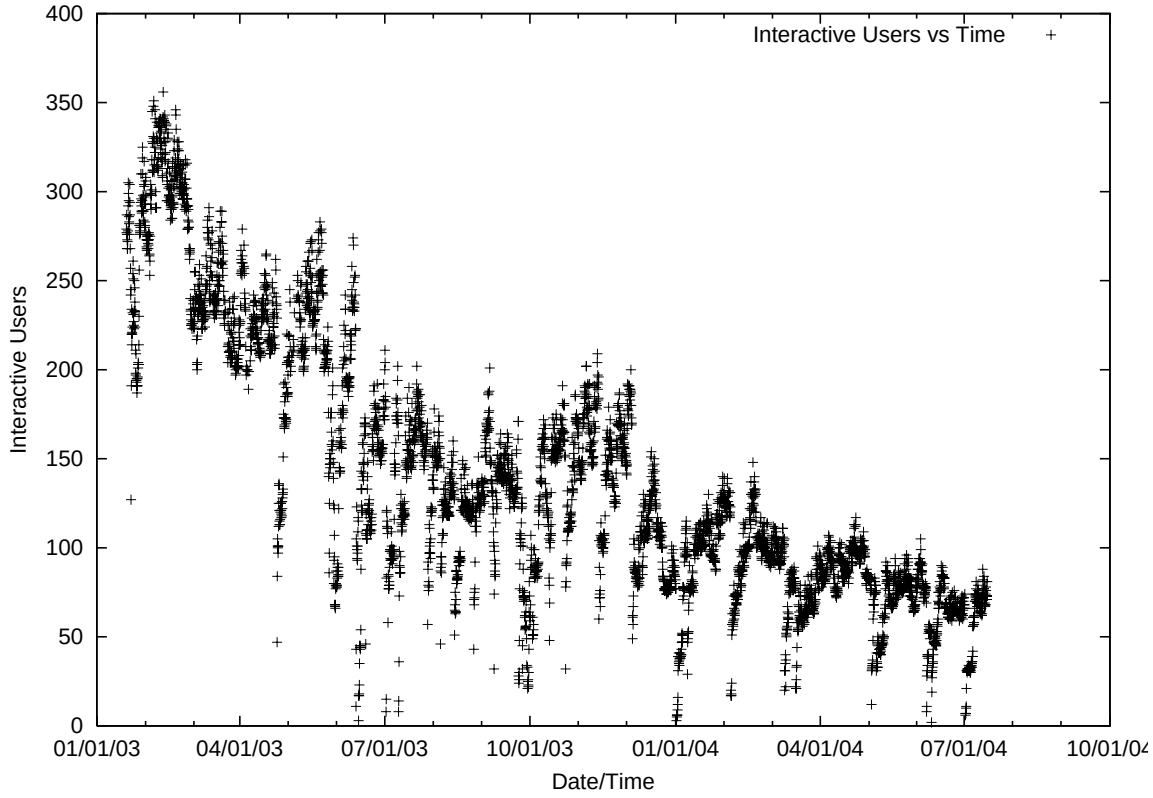

Figure 4: User logins on fcdfsgi2.

other hand, was downsized from 128 to 64 CPUs early in FY04 and shows a continuing decline in use. Primarily, the reason for this is that the machine has no current Run II software. Development and debugging have migrated to Intel/Linux also for binary compatibility with the CAF. Increasingly, fcdfsgi2 is used to share files on its common disk and to serve these files to faster Linux machines via rootd. This fact, coupled with the $201k/year SGI maintenance contract, supports the conclusion that fcdfsgi2 should be fully decomissioned at the earliest possible time. As outlined in the following paragraphs, we expect decomissioning to take place in October, or November, 2004.

To meet the interactive computing needs of CDF within a reference environment using Intel/Linux, a pool of interactive nodes is being constructed. As illustrated in the schematic, there are 3 logical pieces involved.

The first piece is a NIS cluster for account management. This will be constructed using three CAF Stage I dual AMD machines. One machine will act as the NIS Master where accounts are added and modified. NIS maps are created here and dispatched to the two redundant servers. Clients will use broadcast requests to spread queries across the available servers. Barring infrastructure delays, this can be deployed by mid-August. Similar techniques are used to manage accounts on the CDF trailer desktops, so testing should be limited.

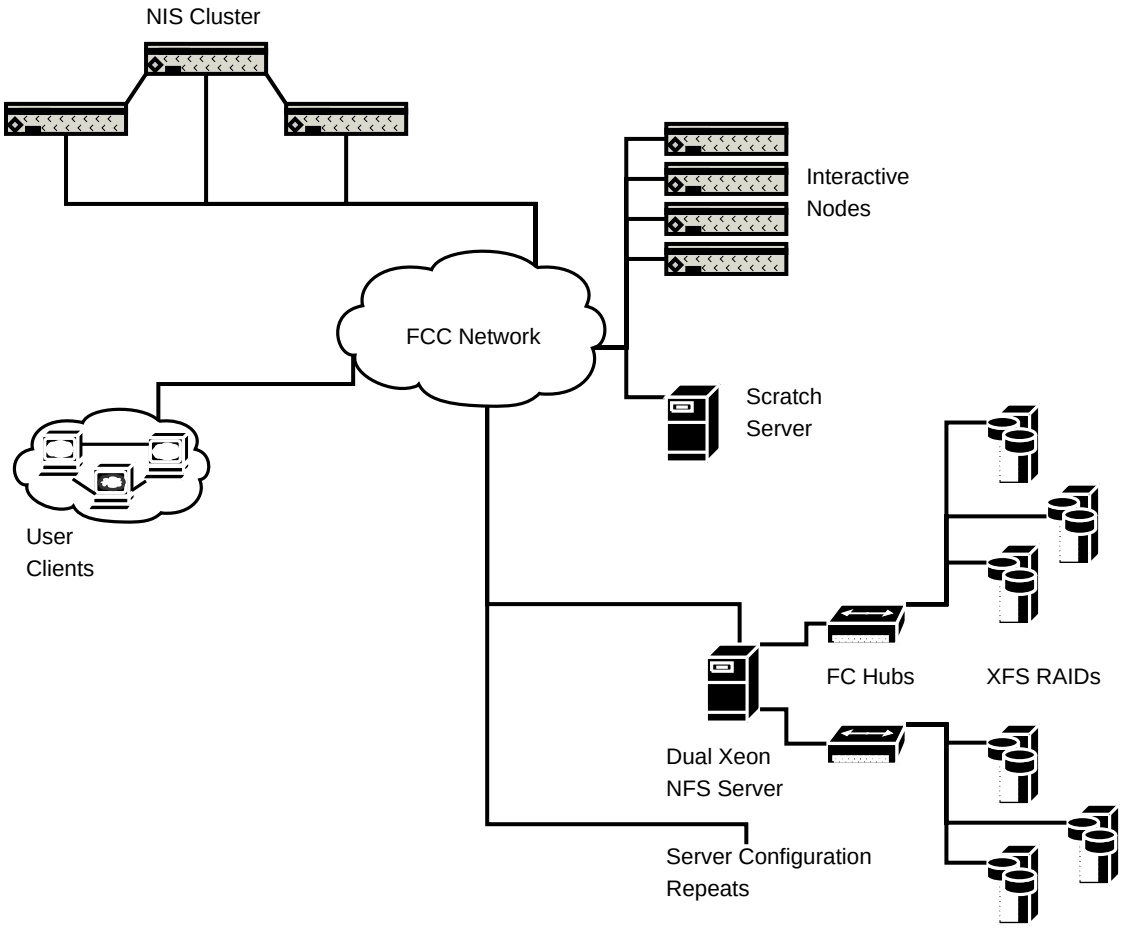

Figure 5: Disks of the interactive login pool.

The second piece is the interactive pool. By design, the pool can be populated by any machines we wish to add. The only constraint is that a machine needs to operate with our reference Intel/Linux distribution, Fermi Linux 7.3.2. To initiate the project, four worker nodes from CAF Stage I have been contributed. These are 2 x 1.4 GHz AMD machines which will be upgraded to have 4 GBytes of RAM each. Home areas will be served by fcdfhome, spool from fcdfspool and a common scratch area from a 2 TByte CAF file server. Experiments with automated load balancing uncovered too many incompatibilities with Fermilab Strong Authentication requirements, so load balancing will be left to the users. Barring infrastructure delays, machine upgrades should be completed and systems deployed by the end of August. User load should be ramped up over a period of time to test stability of the hardware, but the nodes have been burned in; therefore, the primary concerns are sufficient CPU power, memory and network throughput.

To reduce the networking load on the interactive nodes, we recommend creating a small cluster of dedicated rootd servers in order to avoid running rootd on the interactive nodes.

The v2.0 rootd server avoids authentication by allowing anonymous, read-only access, so this should be easily handled in a Linux Virtual Server (LVS) configuration. A prototype can be constructed as needed.

The final piece is a bank of file servers which will inherit the disk arrays from fcdfsgi2 and serve the volumes to the interactive nodes. Presently, we have six 2 x 3.2 GHz Xeon machines with 4 GBytes of RAM, dual gigabit NICs and two available 64-bit PCI slots each. These machines will support XFS, so the existing Fibre Channel host bus adapters (HBA) can be moved from fcdfsgi2 to the new servers and be able to immediately serve the volumes to the interactive nodes. There is, however, a complication to the migration process: XFS for Linux supports only one version of XFS at or below a certain block size. Many of the existing filesystems on fcdfsgi2 use the incorrect XFS version while others use an incompatible block size. At the time of writing, 103 data volumes were in use. Of those, 26 were the correct version of XFS; however, only 17 used a compatible block size. The remaining 86 volumes will need to be recreated for use with Linux. Of the various rebuilding schemes, it seems that building new filesystems and copying data from old to new locally on fcdfsgi2 is optimal. It is estimated that filesystem migration will take 8 weeks. As racks of disk arrays are rebuilt, the HBAs they are connected to can be moved to the Linux servers. Some planning is still required for the filesystem migration, so we will not be able to begin until early August. This puts completion in early to mid October. These estimates support the earlier fcdfsgi2 decomissioning estimates.

The largest interactive computing system at CDF is the collection of desktops in the CDF Trailers. Currently, 384 Linux/Intel desktops are managed by 1.5 full-time CD system administrators. The availability of this and other off-site interactive resources greatly reduces the demand for central interactive computing facilities, and therefore reduces the required size of the interactive pool. The vast majority of desktops are part of clusters owned by collaborating institutions. Some of these clusters are loose-knit collections of "independent" PCs while others have dedicated file servers (serving home areas and data volumes) and compute nodes. There are, also, a small number of specially designed clusters managed by institutional and PPD personnel, e.g. the MIT cluster and the ATOM cluster. The growth of such clusters is limited by the available power and cooling infrastructure.

The interactive computing budget includes miscellaneous expenditures not otherwise included in other categories. For example, in FY04, a new system ($30k) and disk chassis ($15k) was purchased to replace the global CDF code server, cdfpca. There will be similar miscellaneous expenditures in subsequent years. The total cost of interactive computing is estimated at $100k per year for FY05 through FY07.

4 CAF Batch System

The work horse for CDF user analysis is a computing cluster presently consisting of $\sim$ 1500 CPU's, adding up to a total of 3.2 THz of CPU cycles, accessing $\sim$300 TB of disk space.

The CAF implements a vertical slice of the services ultimately anticipated for SAMGrid. In particular it has chosen solutions for management of input and output sandboxes, monitoring at system and user level, user interaction and diagnostics, and has provided a model for sharing of computing resources.

In this Section we focus on the services the CAF provides, and briefly describe some outstanding development issues, as well as human resource requirements. Accounting information of resources consumed is presented in a separate section.

For implementation issues we refer the reviewer to the extensive online documentation at cdfcaf.fnal.gov. In particular, the design document, the DCAF installation guide, and the CAF software documentation. Between these three documents and the CAF User Guide all aspects of the CAF are exhaustively documented except for CAF operations, and details on the CAF based on Condor. CAF based on Condor is discussed in CDF note 7088, and an initial draft for an operations document exists at "http://cdfcaf.fnal.gov/doc/cafcondorOperations/". In addition, we maintain an electronic knowledge base on operational issues regarding site installations, CAF, Condor, dCache, and more at the joint CDF/CMS Tier-2 center at UCSD. [4]

4.1 CAF services

The CAF grew out of the need to maximize the amount of computing we can provide for CDF at more or less fixed cost both in terms of hardware as well as human capital to operate the system. Fiscal pressures as well as the scale of the CDF computing challenge lead to a large batch based cluster of commodity PC hardware.

A user compiles, builds, and debugs their application on their desktop anywhere in the world. To do so we provide low bandwidth access to all CDF data files from anywhere in the world interactively. They then submit their job to the CAF by declaring their binaries, as well as a shell script to run them, a directory structure that contains both, and the level of parallelization desired. The CAF user interface forms a gzipped tar archive and sends it for execution to the CAF cluster. At the CAF site as many instances of the user tar archive are submitted to the batch system as defined by the user at submission time. At execution time, the archive is unpacked, and the user's shell script is invoked with whatever input parameters declared at submission time. One of the input parameters is an integer to distinguish between different instances of the same archive. It is then up to the user to implement the details of the parallelization based on this integer.

After the user shell script terminates the CAF creates a tar archive of the user working directory on the local node in the cluster, and copies it to a location defined by the user at submission time. In principle, the output location may be anywhere in the world. In practice we provide 50 GB scratch space per user inside the CAF. This scratch space may be accessed transparently using a set of environment variables defined by the CAF for the user. The user may access their scratch space via ftp and rootd from outside the CAF, and via ftp, rsh, rcp,

fcp, and rootd from inside the CAF. We refer to this as *icaf* to indicate that the intended use is as staging area for CAF output, much like imap for email.

The CAF is thus receiving one tar archive with the application, and sending out as many tar archives as there are instances of the user application requested at submission time. An intelligent user will thus copy or delete all files from their working directory before exiting their shell script except for log and core files that they want back.

While the CAF is fundamentally a batch based system, we were unwilling to sacrifice the core functionality provided by an interactive system. We thus implemented not only the usual batch functionality of *submit, stat, kill*, but also a core set of services that allow a user to watch jobs as if they were running on a local desktop instead of a remote cluster. Among these services are *ls, tail, top*, and *debug*. The first three allow the user to obtain information about the local environment in which a given instance of a job is executing without the need to know where that environment is located. The user need only specify the instance and submission ID to get this information. The debug service allows the user to attach a gdb session to a running executable. To do this, the user needs to specify the Unix PID in addition to section and job id. The user may look up the latter on the CAF monitoring pages. Among other details, the web based monitoring provides CPU time consumed for all processes spawned by any instance of a user's job while it is running.

Once all instances of a given submission have terminated, the CAF will parse a set of CAF logfiles created for this submission, and write a summary report to be emailed to the user. The objective with this email report is to provide the user with a quick overview of how well their submission completed. The body of the report provides sufficient information for the user to determine which instances have failed, as well as the reason for failure if known. It is thus very easy for a user to go back and debug individual instances by either inspecting the core and log files they received back with the output tar archive, or by running a specific instance interactively through a debugger. The report will soon include I/O monitoring, presently deployed only on our testcaf.

We consider the CAF services to be in their final form except for minor modifications of *stat* reporting. The one remaining service that we may develop in the future is a concatenation option. It is not unusual for a user to request 1000 instances or more at submission time. The hooks for concatenation exist but we believe that progress in data handling is required, i.e. storage and management of the intermediate results, before concatenation can be implemented in a sensible fashion.

We expect data handling to mature within FY04 and may thus revisit concatenation in FY05.

4.2 CAF future directions

We believe that the CAF's long term value lies in its services provided to the user, as well as its monitoring. The lasting intellectual value is thus in concept rather than implementation. Implementation while its cardinal weakness is also a crucial strength. It is entirely home brew with no standards other than kerberos used in its implementation. This allowed us to build

the first system in little more than 6 months, and will lead us to replace the present implementation as GRID standards mature and reach the same level of reliability and functionality as the CAF has today.

The challenge for the CAF is to replace part of its present implementation with emerging grid standards without sacrificing existing services or reliability. At the same time we want to extend functionality and morph the single CAF at FNAL into a cluster of CAFs across the globe that are connected via grid standards and global resource brokering in the context of SAMGrid and the Open Science Grid. [5]

At the core of our thrust to base the CAF on standards is a re-implementation of the CAF with Condor as the underlying batch system. This re-implementation is complete except for the "hierarchical fair share" capability that we built around FBSNG in oder to support groups of owners.

We were originally expecting the Condor team to provide this functionality. However, at present it is unclear if this will happen on a timescale that is satisfactory. We might therefore need to spend additional human resources in FY05 in order to complete the transition to Condor by developing a sufficient level of hierarchical fair share ourselves to support groups.

In addition, we are interested in morphing the CAF into the kind of VO specific job management layer that both ARDA and OSG envision. This implies changing our infrastructure such that we can use Condor glide-in to exploit LCG/EGEE/OSG resources. This work would likely be done in a collaboration between INFN and UCSD-CMS, in order not to interfer with CAF operations at CDF. CMS users of the Tier-2 center at UCSD already use the CAF infrastructure to access the shared CDF/CMS cluster. It is thus perfectly reasonable to expect closer ties between CDF and CMS computing infrastructure development in the future.

The computing requirements of the CAF increase with increasing luminosity and trigger rate as described by the CDF computing model. Table 2 shows the expected CAF resources as a function of time. Infrastructure and budget limitations prevent all the CAF equipment from being housed at Fermilab. CDF has begun implementing a distributed computing model in which 25% of the computing resources are located off-site in 2004, which is expected to grow to 50% of resources by 2005.

Fiscal Year	Total Need (THz)	CPU Speed (GHz)	Off-Site (THz)	New On-Site (# CPU)	Retire On-Site (# CPU)	Total (# CPU)	Cost On-Site ($M)
03	1.5	2.2	-	2* 159	0	1.5k	0.31
04	2.7	2.8	0.7	2* 200	2* 31	2.3k	0.49
05	7.2	3.9	3.6	2* 190	2* 200	3.6k	0.42
06	16	6.2	8.0	2* 386	2* 66	8.1k	0.85
07	26	9.9	13	2* 332	2* 367	12k	0.73

Table 2: CAF annual procurements for on- and off-site resources.

5 Processing Farm

The CDF Production Farm provides the experiment with production-level reconstruction processing of all raw data using a cluster of about 200 dual Pentium nodes, or about 770 GHz of CPU at present. The farm fulfills two goals: to provide fully reconstructed data using a standard executable and calibrations within a few days of data taking, and to re-process some or all of the data with updated executables or calibrations as dictated by the needs of the physics program.

In this section, we first describe the farm architecture, followed by a discussion of the resource limitations exposed by the most recent rounds of re-processing. We then discuss an upgrade to the farm system that is intended to better optimize resource utilization and increase the flexibility of the processing model. The upgrade will use SAM to track file metadata and to assist in managing job submission in an optimal way.

5.1 Farm Architecture

The CDF Production Farm consists of three servers and 191 worker nodes. One of the servers, cdffarm0, provides the core FBS batch system and a MySQL database used for job and file tracking. The second server, cdffarm2, runs control daemons for resource management and job control for each data stream. The disk space within the farm, which is a collection of IDE drives on all worker nodes, is virtualized through a "dfarm" file system hosted on cdffarm0. The present dfarm capacity is 23 TB. Both cdffarm0 and cdffarm2 are dedicated to these tasks. The third server machine, fnpcc, hosts a java server and a web server that run the user interface and other monitoring operations, as well as interactive services for farm operators.

Table 3 shows the quantities and types of worker nodes currently in the farm. Each of the 192 worker nodes is configured in one of two ways, either as an I/O or a processing node. Typically, there are 16 I/O nodes. One set of I/O nodes stages data from tape and distributes input files to the processing nodes. A second set collects output files from processing nodes, concatenates them as needed into files with a size appropriate for tape storage and then write them to tape. All I/O nodes are 2.6 GHz dual Intel P4 machines and utilize optical Gigabit network connections in order to saturate the I/O channel to tape. The I/O node configuration also includes a pnfs filesystem that provides direct access to the Enstore data archive. The remaining nodes, with a total processing power equivalent to about 680 GHz (Intel P3

Year	Numbers	Type	P3 equivalent GHz
2001	64	P3/1.00 duals	128
2002	32	P3/1.26 duals	81
2002	32	AMD/1.67 duals	107
2003	64	P4/2.6 duals	450*

Table 3: Past production farm procurements. These are the nodes currently in use. (* scaled by 1.35 to Intel P3 equivalent).

equivalent), are configured as processing nodes. These machines are dedicated to running the reconstruction programs.

The production farm control software provides a complete chain of data processing from the retrieval of data from the Enstore tape archive to the storage of reconstructed data in the tape archive. Farm processes are defined by the specified input dataset. The work-flow for each dataset is handled by a "farmlet" that has a series of daemons that perform data and job management functions, and that track file status and job history through the MySQL database. All farmlets work independently and share all farm resources on a first-come-first-served basis. This scheme can lead to unintendend interferences, particularly with regard to tape drive utilization. Each of the farmlets, for instance, can access only a single tape drive for input data staging, and one for each output dataset. The Enstore queuing system does not distinguish between input datasets, however, and will allow a single farmlet to create a large backlog in the tape request queue, thereby blocking access to tape by other farmlets despite the availability of open tape drives. While not an issue for raw data processing, this feature can seriously limit throughput for data re-processing, where the input data volume is large, without occasionally substantial human intervention.

Processing jobs are dispatched in units of a "file-set", a pnfs sub-directory of 10 files, each with a typical size of 1 GB. Eight of the I/O nodes are input stagers, copying data from tape to a local scratch area, then into dfarm. The staged raw data files are first dispatched to workers running the reconstruction executable; the output files are written into dfarm. A raw data file can have multiple output files each containing data that satisfy different sets of online triggers. The output file sizes, consequently, vary from about 20 MB to 1 GB. The output I/O nodes, or "concatenators", collect the products of each output dataset in dfarm and concatenate the files into the final output files. The history information in the database is used to ensure that all events in an output file form a contiguous time period during data taking. Exceptions are allowed for files that fail to process due to multiple abnormal terminations.

5.2 Farm Capacity

Processing tasks on the production farm falls into three main categories: raw data processing during normal data taking, data re-processing and specialized tasks. The impact of each must be considered when discussing the capacity of the production farm.

During periods of normal data taking, the policy of the experiment is to process the raw data within three days of data taking. This requirement stems in part from the need to provide rapid feedback to the detector operations group for the purpose of monitoring data quality. Due to various latencies in providing calibrations and other routine operational delays, the farm must have sufficient capacity to process raw data at an average rate that significantly exceeds the peak logging rate of the experiment. We calculate the farm capacity required to keep up with data taking by multiplying the peak logging rate of the experiment by the CPU per event. A contingency factor of 50% is included in the CPU per event in order to allow for increases in processing time with luminosity or more complex versions of the executable. We further assume an 75% utilization efficiency for the farm (which includes routine operational delays).

Data re-processing is required from time to time because the quality of the reconstruction programs improves with time. Periodic re-processing allows the entire dataset or some fraction thereof to benefit from these improvements. The problem of re-processing differs from that of raw data taking in that the rate of data input to the farm is limited only by the I/O capacity to the tape archive, and far exceeds the rate at which the detector can generate data. During Run I, CDF re-processed each year about 30% of the available data toward the later portions of the run, with somewhat higher fractions near the beginning. For the current estimates, we assume a re-processing fraction of 30% in FY-05, and 20% for all subsequent years. The drop is justified based upon improved reconstruction and production management as the run progresses.

In previous years, we have estimated the total required farm CPU by assuming that some fraction of the entire dataset must be re-processed over the same period during which the raw data is logged. The same contingency and utilization factors are assumed for the re-processing. As we will discuss later, this model is not well matched to actual experience and is one topic that we hope to address with the production farm upgrade. We nonetheless use this model for the current budget estimates because it builds in some reserve capacity in the farm is in fact required in order to accommodate data re-processing, and because the cost of the farm in the end is relatively modest.

The final set of tasks performed by the farm include such items as the beam axis calculation and other calibration constants required by the reconstruction programs. Other tasks may include the processing of datasets for special purposes as requested by the physics groups that are too large to be efficiency conducted without the automatio provided by the farm. The load from these sources is negligible and are typically easily accommodated without explicitly including them in the calculation.

Figure 6 shows the cumulative number of events logged by the experiment and the number of events processed on the farm as a function of time. The above calculation clearly provides sufficient capacity to maintain pace with data taking with allowances for delays in processing.

5.3 Farm Procurement Plan

Using the model outlined in the previous section, we estimate the total required capacity of the farm as a function of time. The results are shown in Table 4. To estimate the cost, we assume the purchase of Intel P4 equivalent dual processor machines at a unit cost of \$2.2k. The processor speed is assumed to double every 18 months, starting with a speed of 2.5 GHz P3 equivalent in FY-04. (The FY-04 speed includes the observed 10% increase in processor speed over the machines available in FY-03.) We then constrain purchases within the farm to increments of two racks, where we have assumed 40 U of usable rack space (the current standard for new purchases) and 1 U nodes. The final procurement plan is tabulated in Table 4 for the next three fiscal years. An increase in the event logging rate in FY-05 drives an increase in the required CPU for the farm. The decreased re-processing fraction in FY-06 compensates for a much smaller relative increase in the logging rate in FY-06. The larger data volume then drives the requirement up again in FY-07.

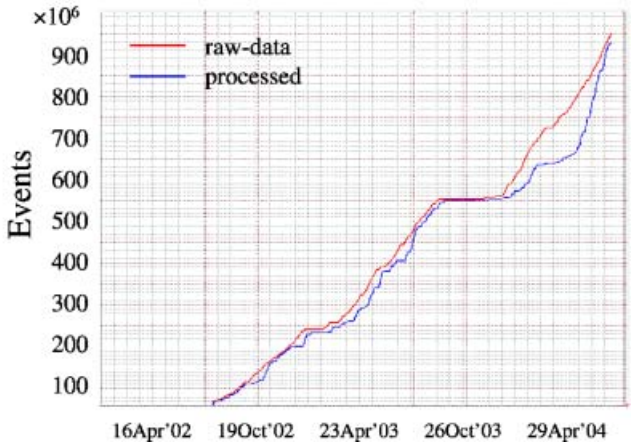

Figure 6: The cumulative number of raw data events logged (red) and the number processed on the production farm (blue). The flat periods are major machine shut-downs. The lag in raw data logging during 2004 occurred when the COT was partially disabled in order to slow aging effects. The capacity of the farm was sufficient to keep up with data taking during the ensuing high-luminosity period.

5.4 Farm Upgrade

Experience with the existing farm has revealed a number of deficiencies in the architecture of the processing system.

- There is insufficient control over resource allocation to allow full utilization of CPU and I/O resources, particularly under heavy re-processing loads.
- Scaling issues with the existing job state and file tracking scheme have lead to significant under-utilization of farm resources in some circumstances, and may limit further expansion of the farm.
- The system design lacks error recovery protocols and is therefore not robust against many common errors. Simple job failures or re-submissions typically require a significant effort from operators in order to restore the state of job and file tracking metadata. This issue is compounded by infrastructure failures that result from the poor scaling properties of the system under heavy loads.
- The lack of a work-flow management framework makes it difficult to alter the processing model in response to changes in requirements.
- The demands upon the farm are episodic, driven by occasional periods of re-processing, when the load is extremely high, or machine shutdowns when no raw data is logged and the load is essentially zero. A significant fraction of the farm remains idle over the course of a year.

FY	Need (THz)	New CPU (#)	Retire CPU (#)	CPU Speed (GHz)	Total (THz)	Total Cost ($M)
03	480	2* 64	2* 73	2.2	525	0.19
04	1100	2* 80	2* 64	3.0	1100	0.24
05	1400	2* 80	2* 64	3.9	1500	0.18
06	1200	0	2* 64	6.2	1300	0
07	2600	2* 80	2* 64	9.9	2600	0.18

Table 4: Production farm procurement. Numbers in FY-03 to FY-04 are actual and FY-05 to FY-07 are estimates.

As a result of these and similar issues, maximizing the throughput on the farm has required adapting the characteristics of the input data stream to the limitations of the farm rather than having the farm processing model evolve with the needs of the experiment. A current requirement for re-processing, for instance, is that the data be split and concatenated in advance. A change in the splitting scheme is not easily handled unless the new split is a strict subset of the existing split. There is no adaptation currently available that can allow more uniform utilization of farm CPU cycles.

In order to address these issues, we have undertaken a complete re-design and upgrade of the farm control system. The goals of the upgrade are to improve the resource management capabilities of the system, provide robust error recovery and to migrate to an infrastructure that will allow production jobs to be inter-operable on other platforms and the production farm to be a computing element within a GRID computing model for CDF. As a short term goal, we intend to make farm jobs inter-operable on remote systems under the control of CDF by end of the Fall 2004 shutdown, and inter-operable on machines that are not under the control of CDF by the end of 2005 or 2006.

5.4.1 Proposal for a SAM-Based Farm

The cornerstone of the new farm system will be SAM, which will provide data handling and file metadata services. Job state information will be recorded in SAM as well. The daemons that currently monitor job status and manage work-flow will be replaced by SAM projects, existing batch submission tools and project or job management threads. Input datasets for projects are defined prior to submission using highly flexible database queries that can be applied to any cataloged data. Error recovery is provided by existing SAM project recovery tools. False starts can simply be re-submitted as new projects, thereby avoiding the lengthy and labor intensive cleanup required to restore the farmlet daemons and metadata to an appropriate initial state. Resource allocation can be handled in part via batch management tools and the use of resource class ads to target jobs that require special services to the specific nodes that can provide those services. Splitting input datasets into multiple projects submitted in parallel offers a second mechanism by which to control resource utilization.

The new farm system will be developed on the CAF platform. Since production jobs currently require little if any CDF-specific services on the current production farm, there will little need to rely on CAF or CDF-specific features in the new implementation, a fact that should simplify the task of GRID enabling production jobs. Once production jobs are demonstrated to operate on the CAF at full scale, the existing farm will be converted to the new system, including the deployment of CAF software on the farm. The CAF infrastructure will in principle add the farm to the pool of computing available to the experiment for user analysis.

There are several important benefits to this plan:

- It will allow production processing to proceed uninterrupted should the development schedule slip into a period of data taking, regardless of whether the deployment on the CAF or the migration to the farm is late.
- The CAF becomes the platform for future development and testing of farm software, thereby allowing these activities to proceed in parallel with on-going farm activities.
- The farm becomes available for opportunistic use by jobs submitted to the CAF, which will improve the utilization of farm resources.
- The virtual boundary established between the CAF and the farm allows the experiment to dynamically increase the size of the farm into the CAF should the need for re-processing exceed the capacity of the farm.
- Since SAM is a common tool across all Run II experiments, the farm system will benefit from efforts to extend SAM to the GRID via JIM.
- Since the CAF is ubiquitous across CDF, efforts to migrate either general CAF or production processing to the GRID will benefit the other.

The plan outlined above will provide a flexible architecture in which to conduct farm operations. As importantly, the services provided by the CAF are essentially the same as those needed on the GRID. A proper structuring of the production system on the CAF infrastructure can therefore offer a number of simple development pathways by which the system can evolve to operate in a GRID environment. A more detailed technical plan by which to reach this final goal is a high priority for the coming year.

6 Data Handling and SAM

The CDF DH (Data Handling) system is comprised of user application interfaces (DH modules in AC++), SAM, dCache, and Enstore. One may think of these four elements as user API, "data handling", cache management, and archival storage. SAM's role in the CDF DH system is to control data movement, and to record this movement in the metadata catalogue.

Support for the elements of the DH system is divided among several entities: The DH modules are the responsibility of CDF, and are supported jointly by Rutgers University and the CDF project within the Run II department in Fermilab-CD. SAM is a joint CDF and D0 project, recently joined by Minos, with database and GRID support from CD-CSS. On the CDF side, SAM is a major responsibility of our UK collaborators, with Glasgow contributing a co-leader to the SAM joint project. Routine operation of CDF dCache is the responsibility of CD-Run II, with development support from CD-CCF. Operation of CDF Enstore system is the responsibility of the CCF department.

The last two years have seen significant changes in the DH system. The dCache product was integrated into CDF DH and commissioned during FY-03. The focus since that time has been to prepare CDF for production use of SAM and to lay the groundwork for retirement of the existing DH system. In the coming year, we expect to begin introducing GRID functionality into the DH system.

The remainder of this section is organized as follows: We first discuss archive related costs, as well as the model used to predict them. Costs for cache disks are discussed in Section 6.2. This is followed by a discussion of DH operations and performance. We conclude with future directions.

6.1 Data Archive

The tape achive consists of three components: the automated tape library, the tape drives that provide I/O to the archive and the tapes that fill it. In this section, we will discuss the requirements relevant for each of these components and discuss the plans for meeting those needs.

6.1.1 Data Archive Requirements

The tape archive must accommodate the raw data from the detector, the primary production datasets, secondary datasets and Monte Carlo data, all of which are EDM-based root files. This accounting neglects the volume contributions from tertiary datasets or other highly compressed files created by the physics groups, since these sources are expected to be relatively small.

For budgetary purposes, we adopt the 'upgrade' option of CDF 6639, version 2.0 (Implications of increased data logging rate on CDF Run II Computing plan and budget). We echo here the broad outline, noting which options are being chosen, and giving the bottom line results.

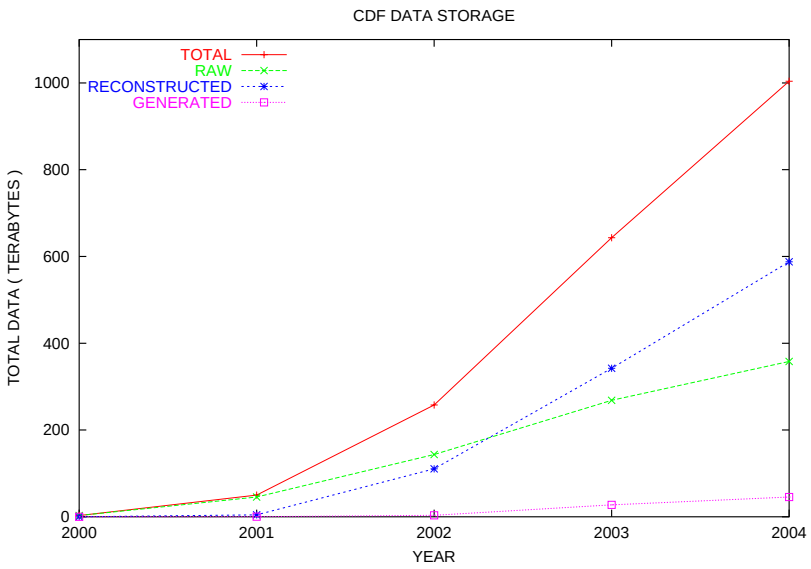

Figure 7: Volume of raw, reconstructed and simulated data stored in the tape robot as a function of time. The total volume is shown in red.

The volume of raw data in the archive is calculated from the average level-3 trigger rate, the event size and the integrated time the experiment is logging data. We assume a combined, overall accelerator and detector efficiency of 0.3 over the course of a year. The average level-3 trigger rate is a function of the peak data logging rate from the experiment. Upgrades to the peak logging rate proposed in FY-03 will be implemented as planned starting in FY-05, almost doubling the rate from 20 MB/s to 35 MB/s. The rate increases again in FY-06 to 60 MB/s, where it remains for the rest of Run II. The raw data event size is determined by the level of raw data compression and the instantaneous luminosity. A more compressed raw data format intended for FY-04 deployment has been delayed until FY-05 pending improvements to the trigger. An additional contingency factor applied in estimating the budget (see Sect. 6.1.2) covers the increased event size due to changes in the instantaneous luminosity.

The volume of reconstructed data is taken from the total number of events logged and the production event size. The number of events logged follows from the average event logging rate and the integrated time during which the experiment is logging data.

Production re-processing increases the estimated volume by an additional factor of two in FY-04, falling to 0.3, the historical value, in most subsequent years. Secondary and MC datasets are assumed to contribute a volume equal to about 50% of the size of the production output before re-processing.

Table 5 shows the estimated archive volume. The quantity of data added in each category and the total volume agrees with the observed data volume, plotted in Fig. 7, within about 10%.

The total I/O demands on the robot will determine the number and type of tape drives that are required. To estimate the I/O to the archive, we sum the contributions from all sources:

Fiscal year	04	05	06	07
Event logging rate (Hz)	85	170	250	250
Raw data (TB)	64	198	354	354
Production output (TB)	230	305	496	673
Secondary datasets (TB)	115	153	248	337
Annual archive (TB/year)	409	656	1098	1364
Total archive (TB)	930	1660	2758	4121
Raw data (MB/sec)	12	21	18	36
Farms I/O (MB/sec)	25	31	23	63
Analysis I/O (MB/sec)	168	419	909	1422
Archive I/O (MB/sec)	205	471	950	1521

Table 5: Tape archive volume. Farm I/O for FY-05 and later is for reprocessing and writing only, assuming a tapeless input path for raw data. CAF I/O is not adjusted downward yet to allow for expanded disk read buffer capacity. The requirements for FY-05 are taken from this table.

writes of raw data, farms output, re-processing, secondary datasets and Monte Carlo storage, and reads for production, secondary dataset creation and general analysis. The contribution from the completed tape migration from 9940A to 9940B tapes was small. The load due to future media conversions is not specifically included.

Data moving in or out of the archive is staged to disk first in order to adapt the I/O rate of external data consumers or producers to the I/O rate of the tape drives. This staging step implies that the archive need only provide the average read and write rates in order to keep pace with demand. To obtain the bandwidth required by raw data logging, for instance, we multiply the peak logging rate by the operating efficiency during peak periods (typically 0.6). The data rate required to write output from the production farm into the archive is obtained by multiplying the raw data write rate by the ratio of production output to raw data event sizes.

At present, raw data processed on the production farm is written to the archive, then read back to the farm directly from tape, requiring corresponding tape drive capacity. During FY-05, we expect to move raw data to the farms directly via dCache read buffers. Likewise, we expect to move the farms output directly into the dCache read buffers. We have adjusted the required tape read rates accordingly.

To estimate the archive I/O required by user analysis, we take the total estimated read rate on the CAF and multiply by the cache miss rate. Experience indicates that about 10% of the file requests on the CAF result in cache misses that require reads from tape (see Sect. 6.3).

The results of these estimates as a function of fiscal year are presented in Table 5. The observed tape drive I/O, shown in Fig. 8, is in fair agreement with the estimates. In FY-04 raw data and farms output account for write rates of about 2 TB/day. Rates are sustained at around 4 Tb/day in early 2004 while reprocessing older data, with peaks up as high as 5

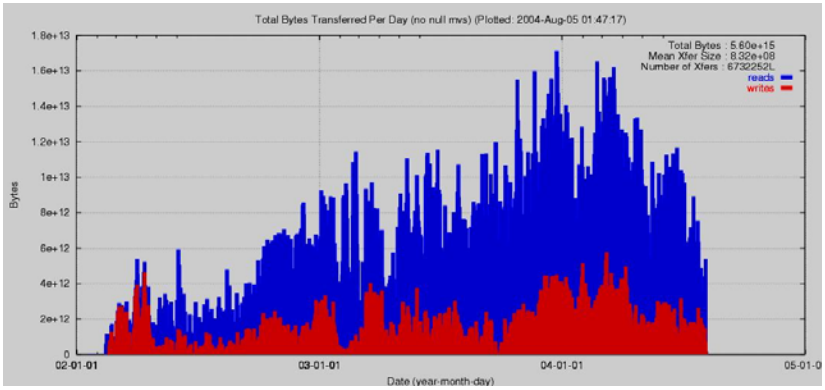

Figure 8: Tape I/O (TB/day) as a function of time. Much of the load in early 2004 was due to the (possibly final) reprocessing of all CDF data.

TB/day on occasion as secondary datasets are being written. Read rates peaked at around 16 TB/day, limited by tape drive availability.

To determine the number of tapes needed to provide the required archive capacity, we consider not only the size of existing tapes, but also anticipated changes in tape technology and available densities. Such developments occur over long time scales and require careful planning of technology evaluation, deployment and possibly density migrations.

A migration of CDF data from the old 60 GB 9940A density to the 200 GB 9940B density was completed in FY-04. The process was performed over 18 months at low priority in order to avoid interferring with normal tape operations, and in order to avoid the purchase of additional expensive tape drives. The process re-cycled about 6000 existing tapes and avoided the purchase of about 4000 tapes over the past two years, which would have been an expense of roughly $300k.

In the 2003 plan, we expected to migrate to an as yet unspecified technology "X" in FY-05 with twice the density of the existing 9940B tapes. This new technology would require the purchase of new tapes, so tape re-cycling will not be an option. To date, these tapes are not yet available, [1] so the "X" tape technology deployment is assumed to occur in FY-06. The cost of density migration is not included in the budget.

To calculate the number of tapes needed, we take the estimated archive volume each fiscal year and divide by the tape cartridge capacity. The requirements are shown in Table 6. For FY-04, we estimated a tape consumption rate of about 40 tapes per week averaged over the entire year. Figure 9 shows the volume of data written during the last few months of FY-04. The tape consumption rate during the last weeks of the plot is about 50 per week. We expect the observed rate to exceed the annual estimate by about 5 tapes per week since it occurred during a period of active data taking.

[1] One currently available candidate, the LTO-2, is only as dense as 9940B tapes with only a modest media cost advantage ($55 versus $75 per tape), making migration of existing tapes a money loser. STK has indicated that they may provide evaluation units of 9940C by October, 2004.

Fiscal year	04	05	06	07
Capacity added (TB)	409	656	1098	1364
Tape capacity (GB)	200	200	400	400
Cartridges added	0	3280	2745	3410
Migration needs	0	0	4150	0

Table 6: Media requirements.

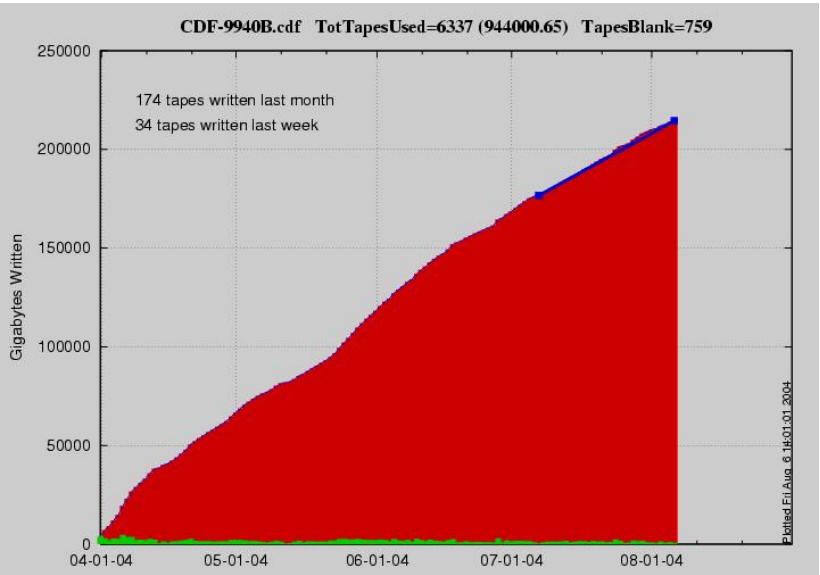

Figure 9: Recent tape usage rates by CDF. We use up to 50 tapes per week.

6.1.2 Data Archive Procurement Plan

To calculate the number of tape drives needed to operate the experiment, we take the estimated I/O bandwidth to the archive and divide by the I/O capacity of the drives. We then multiply the result by a contingency factor of two to take into account tape drive contention, separation of reads and write, down-times, etc. We ignore any constraints on the total number of drives that can be used by the robots and issues such as the mixing of drives types within a single robot.

Table 7 shows actual and projected drive procurements through FY-07. The current archive uses STK T9940B drives, with a maximum I/O rates of 30 MB/sec.

During FY-05, we will seek ways to reduce the number of tape drives needed by the experiment. The motivation for this effort is two-fold. First, the cost of drives is a large fraction of the total computing budget. Reducing the number required potentially frees funds for other uses or allows us to meet our budget guidance. Second, we are reluctant to spend substantial sums on a tape technology that is about to be replaced with a much more effective technology.

FY	Needs (MB/s)	Robots Total	Drives Bought	Drives Total	Date Avail	Storage (PB)	Rate (MB/s)	Cost ($M)
03	190	2	3B	13B	1/04	0.64	400	0.20
04	410	2	5B	18B	7/04	1.0	540	0.13
05	940	2	13B	31B	7/05	2.0	930	0.43
06	1900	2	16X	31B+16X	7/06	3.0	1900	0.48
07	3000	2	19X	31B+35X	7/07	4.0	3000	0.57

Table 7: Robot procurement plan. Numbers in FY-03 and FY-04 are actual and FY-05 to FY-07 are estimates.

The specific steps we will pursue are:

- Implement a tapeless (one-pass) data path for farm input and output. This step eliminates the need to read raw data and production output from the archive.
- Use dCache write pools to decouple raw data sources from Enstore, and allow optimization of data transfer into Enstore. (This will be needed in any case for next generation 60 MB/sec drives.)
- Expand dCache read pools in order to reduce the cache miss rate, and therefore on the need to read tapes.
- Reduce the average event size of production events from the present 150 kB. A production event size of 60-80 kB would dramatically reduce the demand for tape and tape I/O capacity. More of the production output would fit on disk, which would further reduce the cache miss rate and the need for tape I/O.

There are other approaches that can reduce the need for tape drives. The most simple is to make better use of the drives currently in service. The 9940B drives are capable of 30 MB/s. In practice, however, we rarely achive much more than 20 MB/s writing. Software upgrades plus tuning such parameters as file size can help improve the effective data throughput.

We could also move to adopt a new tape technology earlier than outlined above. All of the following drives are expected to ship late in 2004 or early in 2005:

- STK 9940C - 500 GB - 60 MB/sec
- LTO-3 - 400 GB - 40 MB/sec
- SDLT 4 - 600 GB - 70 MB/sec
- SAIT - 1000 GB - 60 MB/sec

While some of these could be installed in the existing STK robots, it will take about nine months to evalute and certify any of them once they become available. It therefore seems unlikely that we could actually deploy a new technology substantially earlier than FY-06.

To obtain the cost of tapes needed, we first multiply the expected archive volume and number of tapes from Table 6 by a contingency factor of 1.2. Assuming a cost of $75 per cartridge for both current and technology "X", we obtain the actual and projected media costs shown in Table 8. The cost of density migration has not been included.

FY	Archive Volume (PB)	9940A Tapes (PB)	9940B Tapes (PB)	"X" Tapes (PB)	Tape Cost ($M)
03	0.40	.22	.24	-	0.18
04	0.98	-	-	-	0.00
05	2.0	-	.59	-	0.22
06	3.3	-	-	1.3	0.25
07	4.9	-	-	1.6	0.31

Table 8: Tape procurements. The fiscal year, data written to 9940A tapes, 9940B tapes, X tapes and the total cost that FY for tape purchases. Numbers in FY-03 to FY-04 are actual and FY-05 to FY-07 are estimates.

Presently we have written to roughly 6,500 of the 7,500 tapes in the CDFEN silos, with about 3,500 slots available to be filled with new tapes. The projected data logging in FY-05 will fill most of the available 11,000 tape slots in the two existing STK Powerhorn 9310 silos. If higher density tapes become available in FY-06, it is unlikely that density migration can prevent the need for an additional robot.

To deal with continued demand for archive space, we are considering the following range of possible actions, listed in approximately the order of preference:

- Remove a large fraction of raw data tapes either to cold storage or to alternate robotic capacity at Fermilab. Once the reconstructed datasets are available for a given file, CDF should normally have no further need of raw data tapes, except for rare technical reasons. This could free up nearly a third of the slots.
- Remove or re-cycle old data tapes which no longer have physics value. We have started the process of reviewing older datasets.
- Consolidate tapes to increase the average tape utilization. We may be able to recover more than 10% of used tapes by concatenating partially filled tapes.
- Expand at least temporarily into the new CMS robot. This step requires formal service agreements for which we have in principle.
- Purchase a third (used) STK silo.
- Expand into existing AML libraries. This requires service agreements, and support for an additional robot technology with which we had operational difficulties in the past.

6.2 Network Attached Disk

The basic plan for disk is to store as much processed data on disk as possible while also providing sufficient space for staging, data caching, data validation, and Monte Carlo data storage. In addition to these uses, some disk is required to store N-tuples or other analysis data samples coordinated by the physics groups.

During FY-04, about 150 TB of dCache pools were deployed. About another 50 TB of disk were used for CAF staging or dedicated to local storage for specific university groups.

The majority of the analysis resources have gone to the large B physics datasets. For planning purposes we can scale disk requirements directly with data logging rates. Table 9 shows the estimated disk space needs and cost for FY-05 and beyond, and the actual volume and cost in FY-03 and FY-04.

FY	Need (TB)	New Server (#)	Retire Server (#)	Server Size (TB)	Additonal Space (TB)	Total Space (TB)	Total Cost ($M)
03	180	18	0	5	90	204	0.34
04	320	8	0	8	64	300	0.14
05	490	19	42	13	180	480	0.29
06	720	18	21	20	250	730	0.27
07	1100	11	18	32	210	940	0.17

Table 9: Disk procurement plan at Fermilab. Numbers in FY-03 to FY-04 are actual and FY-05 to FY-07 are estimated needs.

We do not currently have a good model of the relationship between the total disk space in dCache and the cache miss rate. In the coming months, we hope to improve this understanding in order to better optimize the balance between the amount of disk space and the tape and tape I/O requirements. As previously discussed, reductions in the production event size may change this balance and reduce the need to scale data handling services to still higher levels.

6.3 Data Handling Operations and Performance

The dCache and Enstore sysetms typically handle an I/O load of about 20 TB to 40 TB per day, as shown in Fig. 10. The fraction of data read from tape, shown is red, is usually about 10% of the total data volume delivered. Based on special load tests and experience with real user loads, we estimate the existing system can provide acceptable file delivery service at about 80 TB/day, and 4 TB/hour. We have already seen sustained loads of 60 TB per day.

To maximize cache hits, thus minimizing DH related inefficiencies on the CAF, we partition the dCache system into several pool groups, based on the expected access patterns. The usage load in each of these groups has minimal impact on the other groups.

1. “Volatile”: regular cache, any datasets not mentioned below.

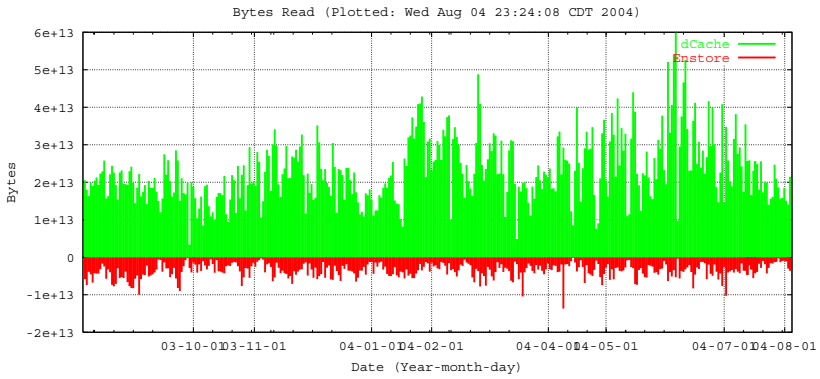

Figure 10: Number of bytes read per day from dCache. Data starts on January 1st, 2003 and the plot ends at late August 2003. The spike of 55 TB/day was a deliberate load test prior to declaring dCache "in production" at CDF.

2. "Golden": secondary datasets that are most relevant for a conference season. We guarantee that those are always on disk by providing sufficient disk space to keep up with new data coming in. We arrive on the list of golden datasets in collaboration with the CDF physics coordinator.

3. "Raw Data, Raw Data stream A, and Big Buffer ": some datasets, especially raw data streams, are either so large or so infrequently used that the number of times a file is accessed while in cache is rather small. This cache pool thus functions more like a FIFO buffer than an actual cache.

4. "Little Buffer ": some deprecated datasets should be accessable on a limited basis, allowing only a few files to be accessed and with very limited disk space allocated. We have set aside six pools for a total of about 4 TBytes for this group, and tightly restricted the number of tape drives available.

The data handling system issues a warning nto users who attempt to access large datasets that are not yet on disk. We require such activities to be coordinated with the DH operations group so that the data can be pre-staged, thereby minimizing loss of CPU time on the CAF due to tape latencies. SAM users get some level of automatic pre-staging because a user declares their dataset at CAF submission time rather than at runtime. Eventually, SAM will automate all pre-staging activity.

6.4 Future Directions

We discuss here the issues of SAM migration, dCache write pools, durable cache, and tape dispersal/replication.

6.4.1 SAM migration

There are many reasons for CDF to adopt SAM:

- Combined development and maintence of DH software with D0, reducing costs by eliminating redundant solutions.
- A clear path to GRID supported tools.
- DH support for offsite computing. This is discussed in detail in section 9.
- Improved operational efficiency of the CAF at FNAL, as discussed above.
- Flexible creation of derived datasets. SAM dataset definitions are created directly by users and groups, the traditional CDF datasets are (mostly) just Enstore file families tracked via the database and file naming conventions. File families are useful administrative tools crucial to efficient tape utilization, but not nearly fine grained enough to track the full range of physics analysis activities.
- Standard, automatic tools to track the processing of files so that partially completed projects can be recovered in spite of occasional hardware and software failures. This is particularly valuable for Farms production and when producin large secondary datasets.

At the time of the 2003 review, CDF had joined the joint SAM project, was working to merge the CDF/D0 database schemas, had started to commission modest offsite analysis facilities and was running the "Predator" process to keep the SAM database tables up to date with the official DFC tables. Since then, the SAM project has continued as a joint project with contributions from CDF and D0, Minos, and CD. CDF has made many contributions to the SAM effort:

- Upgrades and fixes to several SAM components
- An entirely new test-harness framework now in production for all experiments,
- Station and client installation tools useable by all experiments.
- Versioned releases of the SAM station and client software
- Installation procedures for SAM dbservers.
- Installation procedures for SAM web servers.

The final merged SAM database schema has been adopted in production by all experiments.

An interim interface from SAM to dCache was put in place in 2004. This has been heavily tested, and should be adequate for initial production deployment. The longer term interface, via SRM, is being designed and should be available during 2004.

CDF is now preparing to store all new data into the SAM tables and freeze the old DFC tables, thereby encouraging SAM migration. Before SAM is used as the primary production

DH mechanism in CDF, a few tasks must be completed. We are following a plan similar to the one used to bring dCache into production. First, we are providing CDF software releases that allow SAM to be chosen as an option in open beta testing. Once it is proven that SAM can satisfy CDF needs reliably, we will make SAM the default. We are strongly committed to making SAM the preferred choice.

We have established three primary SAM deployment goals for the summer of 2004. They are, in order of priority:

- Unrestricted, production availability of SAM on CAF. The target for this is Sep. 1, 2004. We are moving into open beta testing in mid-July. Preliminary tests of the tota project load, file rates and data rates were successful. We are performing the final load tests now.

- Storage of all summer MonteCarlo production via SAM. All offical CDF MonteCarlo production is being done offsite. Standard tools for SAM file storage have been tested by the MonteCarlo managers. Deployment is pending demonstration that this data remains available to non-SAM users.

- Deployment and pinning of specific high interest datasets on remote computing facilities (dCAF's). Over 10 Terabytes of data is already available on remote sites via SAM. The Karlsruhe site has routinely analyzed data at a sustained 3 TB/day, with peaks to 7 TB/day. Several other sites have been running at up to a TB/day. For comparison, the typical CDF dCache rates were about 25 TB/day. The standard CDF job submission tool (CafGui) supports submission of jobs to the dCAF's, and we are improving documentation of the dataset availablity. We expect offsite analysis to increase substantially this summer and fall.

We now have standardized installation and operations procedures, and have a good support infrastructure in place for remote computing.

6.4.2 Write Caching

One of the longstanding issues with CDF computing infrastructure is the need for a tapeless or one-pass data path. The historical term "tapeless datapath" may be misleading. The tapeless datapath eliminates unnecessary tape reads. It continues to write data to tape at least as frequently as before.

At present, a user does not see full results of freshly logged data until after it has been written to and read from tape at least twice: raw data is written to tape, read for input to the production farm, production output is then written to tape and finally read into dCache for user access. We do operate a tapeless "Express" production system, but with only on a small fraction of the data.

The tapeless data path moves raw data files to the farms and from farms output to the dCache read pools that serve the CAF/DCAF without re-reading the data from tape. Technically, files are written to small write-pools, from which they are both logged to tape and copied automatically to the larger read pools.

Creating a tapeless data path by adding write caches at the output of both data logger and the production farm will make much more efficient use of the limited number of expensive tape drives we have. It will generally expedite production farm operations and improve CAF operational efficiency.

In order to provide ample time for final calibrations (two to three weeks) before farm analysis, we plan to provide a one month raw data read pool. After the data logger upgrade, this will require at least 40 TB of disk, a small fraction of the total dCache pools. Likewise, special purpose high-availability write caches are small in size and cost compared to the overall cache system, and were not included in the budget estimates. The use of dCache write pools has not yet begun. We expect to commission write pools in the fall of 2004.

6.4.3 Durable Cache

Apart from write caching in front of the robot, we also have a clear need for better support of non-archival, but durable storage for individual user data. In a typical analysis, a user starts with some secondary dataset produced in a coordinated fashion by a physics group. The output of this processing on the CAF will generally be a quite sizable collection of relatively small output files. The user thus needs to store these files temporarily for validation, further analysis and possibly concatenation. In general, this processing step is done more than once in order to fix some oversight or the other. Old versions may be deleted to conserve disk space.

An ideal storage system for this use case is disk resident only, and supports deletion as well as reservations and quotas. At present, we support this activity by providing user scratch space inside the CAF. This solution, however, does not scale well, especially if groups of users organize themselves to produce common datasets.

We need to fully virtualize the user scratch space, and then guarantee that datasets are spread randomly across many pieces of hardware. We are presently discussing an implementation of these ideas based on a SAM-dCache combination. While some of the details of such a system still need to be worked out, CDF, D0 and US-CMS have all expressed interest in this functionality, and DESY and CCF are interested in implementing it.

6.4.4 Data Replication

There are specific plans being made within the Computing Division for dispersal of the archival tapes, presumably to reduce the risk of catastrophic loss of all the Run II data. At present, we do not know whether such a dispersal is justified, especially since there appear to be several relatively inexpensive alternatives. One alternative to dispersal is to copy tapes to a low-cost medium. DVD-R, for example, costs of order \$100k per PB and is dropping quickly. Recorders write to double layer DVD-R at a sustained 16 MB/sec. It may be possible to make a complete copy of existing data in months for off-site storage. Concerns about the long term archival quality of these media may not be so relevant in this disaster recovery context: the dispersal plan contemplates 50% loss of data.

A second alternative is simply to re-locate the raw data tapes to a remote offline location. The production output files all contain a copy of the raw data with sufficient information to allow re-processing by production. Since data loss is typically extremely low, there is not need for the re-located tapes to be accessible via automated libraries, thereby further reducing the storage cost. Manually handling cases in which the production tapes fail is not expected to be a significant cost.

7 Databases

CDF currently utilizes Suns for online databases and a combination of Sun and Linux boxes for offline databases. CDF database hardware setup is listed in Table 10. The online production machine, b0dau35, and development and integration machine, b0dau36, are identical Enterprise 4500 servers with two 400 MHz processors each. The machines are dedicated to running Oracle database server. The online production database is behind a firewall. It is divided into several applications: Trigger, Hardware, Run, Calibration and Slow Control (MCS). Write access to these applications requires running on privileged nodes located in FCC. The content of the online production database is replicated to offline production database via Oracle read-only replication. The offline production machine is fcdfora4 which is a Sun V880 with 8 900 MHz processors, 32 GB RAM with about 1 TB of fiber channel disk drives, and Gigabit Ethernet.

Besides replicated applications the offline production database hosts Data File Catalog and SAM [2] applications. Fcdfora1, is older Sun Enterprise E4500, hosts offline development and integration Oracle instances. The access to the offline production database is not restricted. Since summer 2002 the content of online production database (except slow control) and Data File Catalog are replicated to more powerful machine, fcdflnx1, a quad PIII 700 MHz machine with 4 GB of memory and about 1 TB of disk space on SCSI RAID arrays.

The amount of data used by existing application is constantly monitored. Disk space usage as the function of time is used to make projections of space needed for application in the future. Example plots for online and offline production databases are shown in Figure 11.

name	OS	CPU	RAM	Disk	Oracle
b0dau35	Solaris 2.7	2×400 MHz USparc	1 GB	500 GB	9.2.0.5.0
b0dau36	Solaris 2.8	2×400 MHz USparc	4 GB	1.2 TB	9.2.0.5.0
fcdfora4	Solaris 2.8	8×900 MHz USparc	32 GB	1 TB	9.2.0.5.0
fcdfora1	Solaris 2.8	2×400 MHz USparc	1.25 GB	500 GB	9.2.0.4.0
fcdflnx1	RH AS	4×700 MHz PIII Xeon	4 GB	1 TB	9.2.0.4.0

Table 10: Database hardware and software configuration

7.1 Database Replication Hardware

The online data logger and production farms need continuous access to the offline production database in order to log and reconstruct raw data. At the same time, until September 2002, the offline production database was also accessed by users running analysis jobs primarily in read only mode. Increased analysis activity accompanied by substantial growth of CPU power led to several incidents when database and system resources could not handle the demand. This

[2] Sequential Access through Metadata

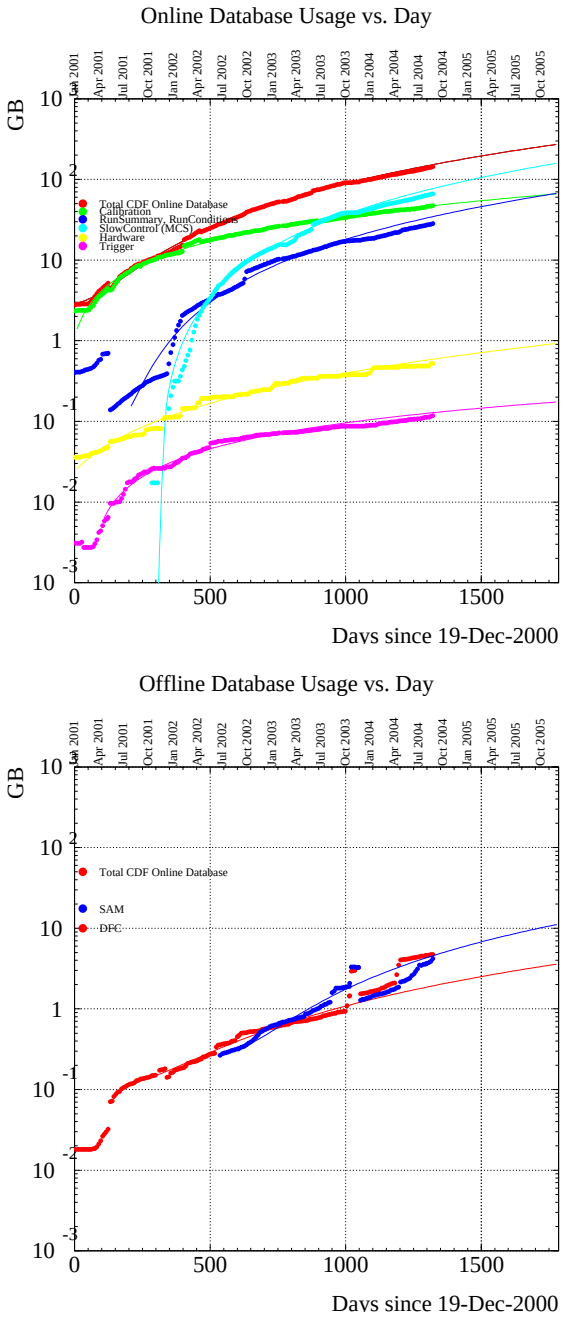

Figure 11: CDF DB space usage on online and offline production Oracle instances. Plot for offline instance shows only read/write applications: SAM and DFC.

issue was addressed by developing a strategy of distributing of databases via replication on site. The replica copies of the database are read only instances accessed by the majority of users; and the production farms and online data logger have exclusive access to the primary offline production database. As pure read-only databases the backup costs are minimal.

The first replica, cdfrep01, hosted by fcdflnx1, running Red Hat Advanced Server Linux operating system, was implemented in the summer 2002. The replication proceeds via Oracle read-only replication from the originating either online or the offline production databases depending on the application. Replication development work using Oracle streams replication is performed in the existing fcdfdata012 machine. Cdfrep01 is used by the CAF and all other users. The offline production machine has become isolated from disturbance by general users and is exclusively accessed by online data logger and production farms. There is a fail-over to fcdfora1 in case of emergency or maintenance work on fcfdlnx1.

In FY-04 CDF database group acquired an 8-way Dell server machine, later 2 TB of SCSI RAID array disks were added to it. Currently the work is on the way to commission this machine as a replacement for the overloaded fcdflnx1.

FY	DB CPU (n-ways)	DB Disk (TB)	Cost ($M)
03			0.15
04	2	4	0.07
05	6	1	0.05
06	2	2	0.03
07	2	2	0.03

Table 11: Database CPU and disk procurement plan. The fiscal year, the number of n-way Linux boxes purchased that year for DB machines, the TB of disk purchased and the cost.

7.2 Database Replication software

During the year 2004 the CDF DB group and CD/DSG groups have been working on implementation of Oracle streams replication. Oracle streams allow the data propagation to proceed in sequential mode, thus avoiding firewall issue, and unloading the source database machine. Streams also allows automatic propagation of DDL changes to replica sites. In addition, streams replication would make it practical to consider maintaining offline database copies based on Oracle at remote institutions, for example in Japan, the U.K. and Italy, for which latency and speed of access to the databases can be limiting factors. Although initial experience with Oracle streams has not been very encouraging. During accelerator shutdown we will perform final evaluation of this product.

7.3 Support of computing at remote sites

CDF has expressed interest in the n-tier database access currently used and developed by D0 collaboration and Fermilab CD. Although n-tier access will not resolve Oracle licensing

issue it will provides a local caching or "secondary sourcing" of data from offline production database resulting in more efficient use of computing resources at the remote sites. CDF database group participated in formulation of requirements to the design of n-tier DB access system for CDF.

The FroNtier DB is a new system for the distribution of frozen database content which utilizes standard Web tools connected into a multi-tier topology. At the moment, it is CDF's best candidate for providing for remote database access from DCAFs.

An example of how FroNtier DB would handle a request for a typical calibration table is shown in Fig. 12. The gain comes from the use of the Squid server to cache the response of the FroNtier servlet. The fact that the client has been code-generated ensures that all requests for a specific calibration table produce one and the same HTTP string, and thus only the first such request actually reaches the Tomcat and causes both Tomcat and Oracle to perform work. All subsequent requests are handled solely by the Squid, which simply delivers a cached already prepared response.

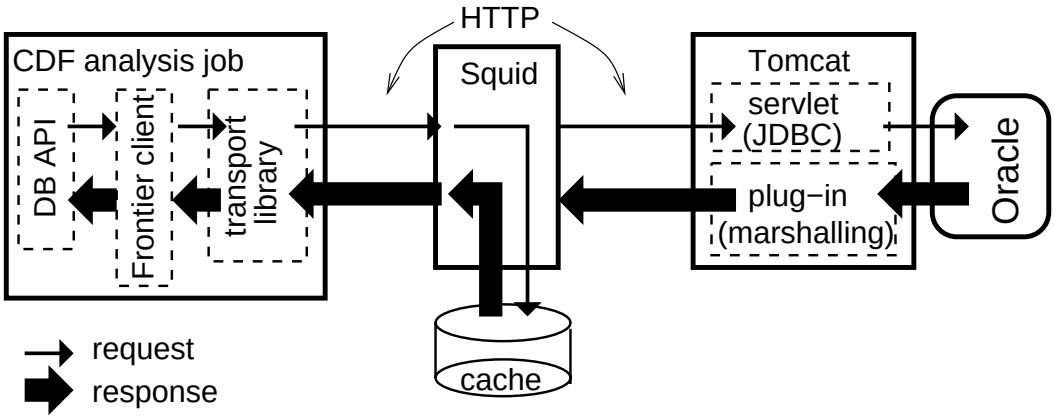

Figure 12: A sketch of the execution of a FroNtier DB request. When the DB API discovers that it needs a certain calibration table, it calls the FroNtier client. Just like an Oracle (OTL) client, the FroNtier client is code-generated. The FroNtier client specifies the order and the types of the fields, but it delegates the details of the data decoding to the FroNtier transport library, which uses libCURL to send a HTTP request and retrieve the response. The request passes through a Squid server, and reaches the Tomcat server which runs a FroNtier servlet. The FroNtier servlet uses JDBC to query Oracle (cdfofprd). Oracle's response is passed onto a plug-in specific to this table, which also has been code-generated along with the FroNtier client and which calls subroutines from the FroNtier transport library to ensure data consistency between the servlet and the client. The response of the FroNtier servlet is cached in the Squid, so that every other request for this calibration table will retrieve the cached response.

The essence of the proposed implementation is the wide-spread deployment of Squid servers, preferably as close to clusters of worker nodes as possible. We assume that every remote DCAF will have at least one local Squid, and that many university groups will elect to have

local Squids as well. We allocate two Squids to be close to large clusters of CAF worker nodes, and allow two Tomcat/Squid pairs to connect directly to Oracle (cdfofprd). The latter two Squids will perform load-balancing, and provide redundancy in case of failure of one of the Tomcat/Squid pathways. True to its name, the system will have N-tiered topology: the remote Squids will connect to one of the two Squids with Tomcats in order to utilize already cached data.

For the FY05 budget, we request six server-class machines. A machine with two 3 GHz Xeons, 4 Gb of memory, and 180 Gb of disk is about $7k. Therefore we request $45k for FroNtier DB system.

7.4 DB budget

The existing load from users' jobs running on Fermilab CAFs is well handled by the replica machines. The offline production machine serves exclusively the production farm and is loaded lightly. Therefore with exiting DB setup supplemented with load balancing between offline production and cdfrep01 we should manage to handle ever increasing CDF load during the lifetime of the experiment. The FroNtier solution to be deployed this fall would allow us to shift load from expensive machines running Oracle servers to commodity Linux boxes running FroNtier components. Approximate breakdown of database spendings are given in Table 11. Starting 2006, after the hardware for FroNtier system is bought, we forsee only maintenance costs.

8 Networking

8.1 CDF Networking

The CDF network topology has become more complicated in 2004 with the increasing size of the CAF computing resources and the growing number of boundary conditions associated with physical infrastructure. CDF has moved a large number of CAF worker nodes to New Muon in response to power and cooling limitations in the Feynman Computing Center (FCC).

The original CAF resources fit in one switch. With the growing number of resources, they were located in two switches across two floors of FCC. The move to house worker nodes in remote buildings further increases the distribution. Beneficial occupancy of the High Density Computing Facility (HDCF) is expected early in September of 2004, which will host the majority of the CAF worker nodes. The plan of the FNAL computing division is to locate all new worker nodes to HDCF and data servers in FCC. This separation of worker nodes and data servers requires a more careful assessment of network topology.

The heart of the CDF offline computing network is the CAS switch, a Cisco 6509, located in FCC2.

- It has 4 10 GBit connections, one port is reserved for the site up-link and the other three are connected to CAF switches: one located in FCC1, one in FCC2, and one in New Muon.
- Fcdfsgi2 is currently connected to this switch via 5 Gbit connections, 1 for interactive use and 4 for Enstore.
- The CDFEN Enstore robot currently has 10 Fast Ethernet (FE) connections to the offline switch for the movers for T9940A drives, and 13 GigE connections to the offline switch for the movers for T9940B drives.
- The stage 1 CAF file servers use 15 GBit connections and the CAF stage 1 worker nodes use 67 FE connections.

The worker nodes and disk servers from the later CAF stages have dedicated switches. The stage 2 CAF has its own 6513 switch in FCC1 and is connected to the CAS switch over one 10 GBit link. The stage 2 CAF switch connects 76 CAF file servers via GBit connections and connects 217 worker nodes via FE connections. The stage 3 CAF switch, located in FCC2, connects 22 data servers connected over dual 1 GBit copper ports and 88 worker nodes connected over single gigabit copper ports. The last CAF segments are physically located in New Muon and connected to a 6509, which connects to the CAS switch over 1 10 GBit fiber. Figure 13 shows a simplified view of the CAS and CAF switches and where the CAF elements are connected.

The CAF worker nodes are being concentrated in HDCF, while the CAF disk servers are located in separate switches in FCC1 and 2. The disk servers need to connect to the tape movers and the CAF worker nodes. The CAF worker nodes only need good connections to the disk servers. The current network topology, in which all satellite switches are connect to

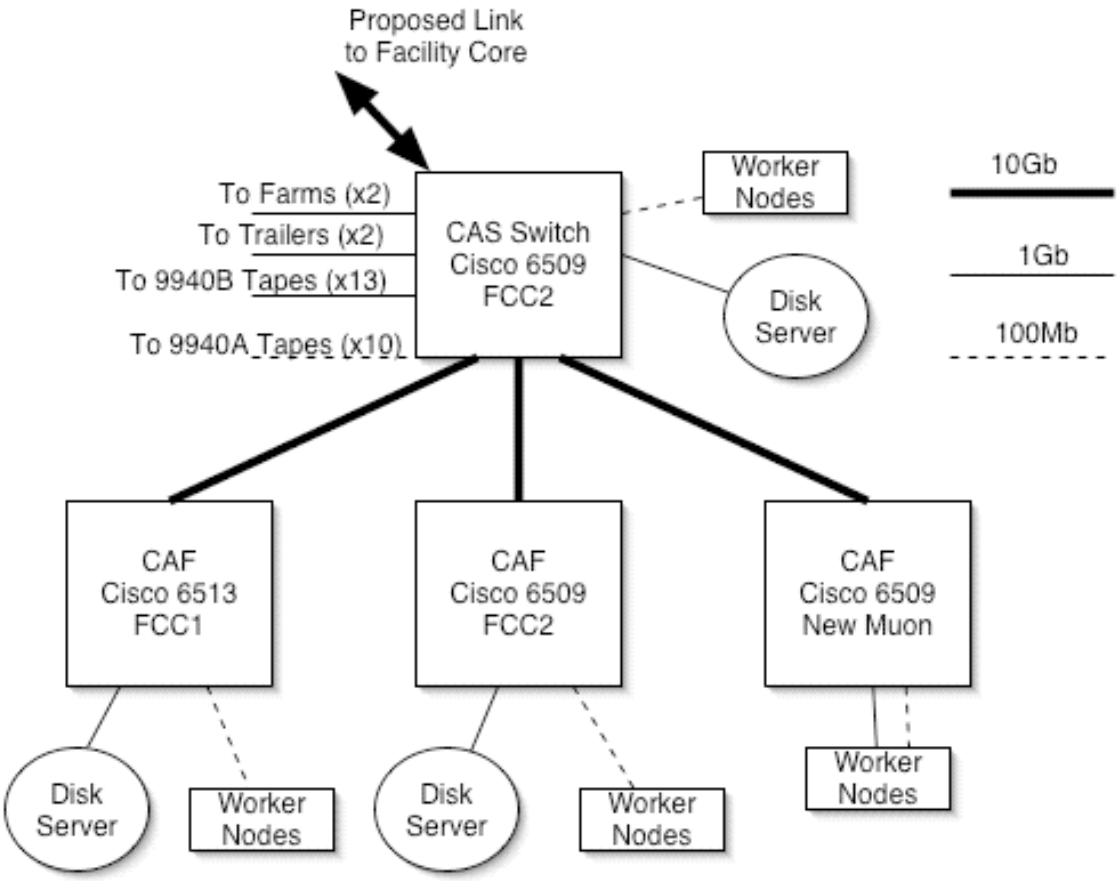

Figure 13: A simplified view of the current CDF Local Area Network in 2004.

the CAS switch, does not reflect the structure of the data flow in the system. A more appropriate structure might be to consolidate the disk servers at the center of the network diagram and try to reduce the percentage of time a worker node has to connect through the CAS switch to connect to a disk server.

A diagram of a possible network setup is shown in Figure 14.

Ideally in the diagram above the disk servers currently located in the CAS switch would move to the now central CAF switch. The tape movers for the 9940B drives would also move to the CAF switch to minimize the distance between the disk servers and the data from Enstore.

Most of the network changes in FCC require reorganizing the network switches on the second floor of FCC, but do not require large network equipment acquisitions. There is sufficient space in the gigabit copper blade in the FCC2 CAS switch to accommodate the 2004 disk server acquisitions, the new 2004 tape movers, and maintain space for early 2005 equipment procurements. There are sufficient blade slots available to move 16 port fiber blades from the CAS switch to accommodate CAF1 disk servers and existing 9940B tape movers.

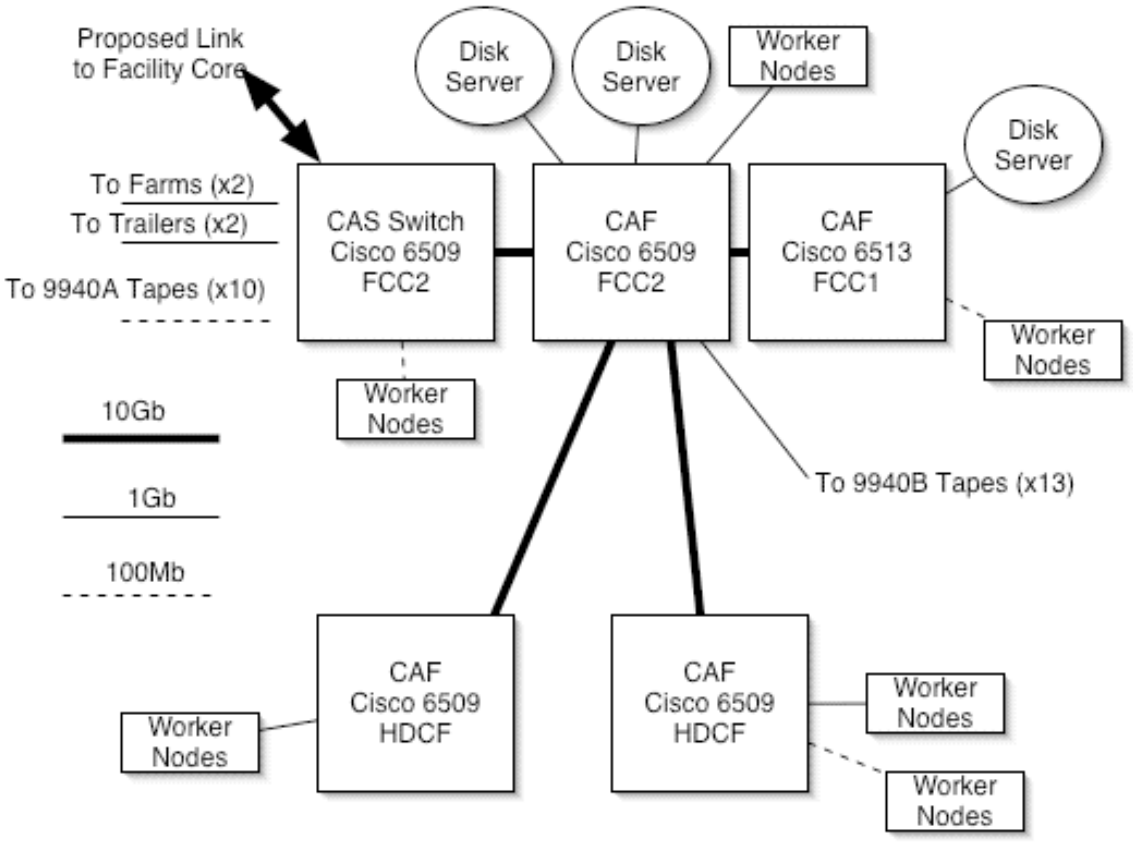

Figure 14: A simplified view of the proposed changes to the CDF network architecture.

In order to support the new CAF work nodes and the new farm nodes planned for HDCF. CDF needs to purchase a new Cisco 6509 switch with 160 copper ports and 4 10 GBit fiber ports for up-link. One 10 GBit GBIC is needed in FCC2 to accommodate the additional 10 GBit up-link from HDCF.

For each of FY05, FY06, and FY07 we plan on purchasing additional CAF worker nodes and file servers with the necessary network connections. The proposed 6509 for HDCF has sufficient capacity for estimated FY05 CAF worker nodes acquisitions. Additional worker node purchases in FY06 and 07 will require an additional switch. The switches in FCC have sufficient capacity for disk server purchases in FY05, 06, and 07. In table 12 we estimate the cost of this networking by assuming a Moore's law like decrease: networking costs that drop by a factor of 2 every 18 months. In practice networking costs have dropped much more slowly than Moore's law. One issue that was not foreseen in previous computing plans was the need to duplicate network infrastructure in satellite buildings.

FY	FCC Cost ($M)	Trailer Cost ($M)	Total Cost ($M)
03	0.23	0	0.23
04	0.19	0	0.19
05	0.18	0.07	0.25
06	0.06	0.06	0.12
07	0.04	0.04	0.08

Table 12: LAN procurement plan. The fiscal year, cost of Fermilab computing center networking, cost of CDF trailers networking and total cost.

8.2 Trailer LAN

The networking in the trailers has not been upgraded in many years, and the networking group has recommended an upgrade for each of the past 4 years. The available network resources have consistently been used for CAF and computing center needs. The CDF trailers LAN currently supports 100 Mb/s connections to the majority of CDF offices, and this lags behind the current network capabilities of desktop Ethernet cards which are 1 Gb/s, and restricts the data transfer rates for existing file servers in the trailers. Currently multiple satellite switches are used to extend the ports available on the trailers 6509 switch, in an architecture that lowers the bandwidth capacity of many offices. The infrastructure in the trailers is currently primarily fiber. The new office building is wired with copper.

There is one 16 port gigabit fiber module in the FCC CAS switch that can be relocated after fcdfsgi2 is retired. This module should move into the trailer central switch, where it could be used to provide up to 16 ports of gigabit connections. This will require upgrading some of the fiber to copper switches located in trailer offices, and this should be reserved for groups operating small clusters and disk servers in the trailer areas.

All the CDF 65 series switched have been upgraded to the newest supervisory module except the 6513 used in the trailers. This switch currently has two trunked gigabit links to the CAS switch, which will be oversubscribed. This switch also serves as the up-link from the new CDF office building switches to the CAS switch in FCC. In order to provide a 10Gb up-link from the trailer switch an upgrade is needed. This year CDF should provide a limited number of gigabit ports and upgrade the switch. This will allow the gigabit infrastructure in the offices to grow in the future.

With the copper infrastructure in the new CDF building, providing gigabit is somewhat easier. For a small initial investment with the possibility of upgrade in the future, a Cisco 3750 seems like an appropriate choice. It is possible to get a 16 1 GBit copper ports with a 10 GBit up-link. Up to 8 additional 3750 switches with 24 1 GBit copper ports can be chained together using 30 GBit links to provide upgrades in the future.

8.3 WAN

In FY03 the OC3 connection between Fermilab and ESNET was upgraded to OC12 with a capacity of 622 Mb/s. In 2004 FNAL purchased a fiber connection to the StarLight hub in Chicago. This has provided 2 1 GBit and 1 10 GBit research networks to the lab. While the main traffic for the site will continue to go through the ESNET connection, research projects and schedule-able data transfers can use the higher performance fiber connection. The port on the CAS switch 10 GBit blade reserved for the up-link should be installed with a 10 GBit GBIC.

8.4 Proposed Budget For 2004

The proposed purchases for this year along with their estimated costs are given in Table 13.

Description	Quantity	Cost	Total Cost
WS-C6509, Cat 6509 Chassis, 9slot	1	$6,175	$6,175
WS-CAC-3000W, Catalyst 6000 3000W AC Power Supply	2	$1,820	$3,640
WS-SUP720, Catalyst 6500 Cisco 7600 Supervisor 720	1	$16,970	$16,970
WS-C6K-9SLOT-FAN2, Catalyst 6000 Fan Tray	1	$322	$322
MEM-C6K-CPTFL64M, Cat6500 Sup720 Compact Flash Mem 64MB	1	$260	$260
WS-X6704-10GE, Catalyst 6500 4-port 10-Gigabit Ethernet Module	1	$12,100	$12,100
WS-X6748-GE-TX, Catalyst 6500 48-port Fabric Enabled 10/100/1000 GE Module	4	$9,100	$36,400
XENPAK-10GB-LR, 10GBASE-LR Serial	2	$2,600	$5,200
ws-c3750g-16td-s 3750 Networking Switch	1	$14,000	$14,000
XENPAK-10GB-LR, 10GBASE-LR Serial	2	$2,600	$5,200
WS-CAC-3000W, Catalyst 6000 3000W AC Power Supply	2	$1,820	$3,640
WS-SUP720, Catalyst 6500 Cisco 7600 Supervisor 720	1	$16,970	$16,970
WS-C6K-9SLOT-FAN2, Catalyst 6000 Fan Tray	1	$322	$322
MEM-C6K-CPTFL64M, Cat6500 Sup720 Compact Flash Mem 64MB	1	$260	$260
WS-X6704-10GE, Catalyst 6500 4-port 10-Gigabit Ethernet Module	1	$12,100	$12,100
XENPAK-10GB-LR, 10GBASE-LR Serial	2	$2,600	$5,200
Total			$140,559

Table 13: Proposed network procurements for 2004.

This does not include the cost of small fiber to copper switches in the trailers, but total cost for enough switches to fully utilize the 16 available gigabit ports should be about $10k.

If any networking budget is available, CDF should purchase a 48 port gigabit module for the upgraded trailer 6513 at a cost of $9,100. This will allow all the offices currently connected with copper ports to upgrade to gigabit this year. Otherwise, the module will be included in the FY05 budget.

8.5 Proposed Networking Plans for 2005 and 2006

After the additional switch procurement this year for HDCF, CDF has sufficient networking capacity for a 2005 hardware procurement similar in size to the 2004 procurement. The new switch in HDCF can accomodate 160 additional systems. The three switches in FCC should be able to accomodate a reasonably large procurement of disk servers and central infrastructure, especially as the older equipment attached to lower density network blades is retired and replaced.

The network budget for new infrastructure in 2005 will mainly be applied to network blades for HDCF, where an additional 3 are needed; upgraded gigabit infrastructure for offices, both the trailers and the new office building, which more gigabit fiber modules can be used from the CAS and CAF switches and additional copper gigabit switches are needed; and upgrades of the modules in the CAS and CAF switches in FCC to replace low density blades with higher density copper gigabit modules.

The item that needs to be watched in 2005 is the over-subscription on the 2 10Gb links between FCC and HDCF. The plan of the computing division to to locate equipment that requires uninterruptible power, like disk servers, in FCC and high power density equipment, like worker nodes, in HDCF. Currently 2 10Gb links are proposed. As CDF moves from 30-40TB per day of data served to analysis applications to 80TB of data served per day, the 10 gigabit links will begin to see high utilization. It is possible to add additional 10Gb links between HDCF and FCC by taking advantage of open 10Gb ports on the CAS and CAF switches and adding multiple routes between an HDCF CAF switch and FCC. This will require a reorganization of the subnet used in the CAF and may require the acquisition of routing modules for the CAF switches that host disk servers. The utilization of the current links should be monitored and an upgrade should be reserved as an option.

In 2006 there will be additional CAF acquisitions for both HDCF and FCC. In HDCF there will not be any network ports available and another 6509 (or 2006 equivalent) will be required. The networking capacity in FCC should be sufficient, provided some of the current worker nodes hosted there are retired. By 2006 the separation of disk servers and worker nodes between FCC and HDCF will be complete and if the CDF networking between the buildings has not been upgraded to multiple 10 GBit links it will probably need to be.

9 Offsite Computing

9.1 Status and Perspective

Offsite computing is by now a de-facto important reality of CDF the computing environment. During FY-04 we moved ahead along the plan outlined one year ago which still provides our blueprint for having a full GRID-like environment by the end of FY-05.

The motivations and history that led to that plan still stand as described in the 2003 version of this document and will not be repeated. We recall here the step-by-step procedure that we indicated then and use it as guideline for measuring development.

In practical terms, we expect CDF to extend from Fermilab to incorporating computing facilities outside the laboratory in a step-by-step process:

1. allow easy usage of remote facilities for code development, data analysis and MC generation by single institutions
2. move off site (part of) organized MC production
3. move off site (part of) single user MC production spreading it uniformly across all institutions
4. move off site analysis of secondary and/or tertiary data sets, by duplicating CDF data at remote institutions and giving everybody access to it
5. exploit large off site CPU capability for interactive analysis and reprocessing
6. develop formal agreement and define the "price" of each service, only after it has been demonstrated to work

We focused FY-04 efforts on technical aspects of making remote sites capable of providing the services. The main issue for FY-05 will be the operational aspects, and thus understanding of performance and usage is needed to arrive at a "pricing" that reflects services rendered in Step 4 and 5 above.

Overall FY-04 has been very successful and now the situation with respect to the 6 points in the above list is as follows (more details are given later on):

1. This has actually been in place for many years by means of a very successful code distribution tool that still provides the foundation of our GRID effort. Since 2003 some institutions have started to use SAM to transparently import data needed for analysis, and this is now the main operation mode of, *e.g.*, German CDF collaborators.
2. By now virtually all organized MC production is performed offsite. Automation of the process and of the data import is now in a development/test phase, we plan to use SAM for data import back to Fermilab tape storage, file concatenation and bookkeeping. MC samples produced offsite are already cataloged using SAM metadata. Use of the RUNJOB tool for process control and of JIM for automatic brokering across several sites are options to be explored.

3. This is now reality. As of a few months ago, several CDF users located, *e.g.*, at Fermilab have successfully used several offsite farms located at institutions other then their own, needing no additional help other than a web page and an announcement "that they could".

4. This is also now a reality. In the same situation as previous step, although still used on a smaller scale.

5. This is at present an R&D project.

6. Negotiations with the CDF International Finance Committee on this point started more than one year ago. The politics are moving steadily in this direction and more details are given later in this document. The missing technical step here is a uniform monitoring and accounting mechanism, which we are now developing.

We are at this point reasonably pleased with the status of steps 1.,2.,3. and 4. Those functionalities have now been deployed at several sites and we have moved from development and beta testing into a production and operation phase, with weekly "offsite operation" meetings, transparent access for all CDF users to several offsite farms and a few TB of data replicated offsite for local analysis. Large scale usage of the offsite farms is now limited not by the tools, but only by the very limited size of those installations (both as CPU and data disk) compared with the central CAF, so that most users do not perceive an advantage over using the Fermilab machines. We expect this to change as soon as most popular datasets (*e.g.* inclusive hadronic *B*) are replicated offsite.

9.2 Status of Offsite Resource as of Summer 2004

While some CDF collaborators own CDF-reserved computers, other share access to largish facilities with other experiments. This makes it almost impossible to tell a-priori, *e.g.*, how much CPU power CDF physicists can use offsite. We expect that once a framework will have been clearly defined for usage of those facilities by all CDF collaborators, something like a minimum amount of CPU cycles available can be defined, while most likely accounting of offsite facilities contribution will have to be done a-posteriori, based also on the actual efficiency and effectiveness of each single installation. Besides, many institutions will keep priority of usage for their own members, so there will not be an a-priory guarantee as to how much usage generic CDF users will obtain, but a-posteriori this can be a considerable fraction of those resources.

As we predicted one year ago, committing of computing resources by CDF institutions for general access has been a slow process, which has moved in some place faster, in some place slower than expected, due to local financial policies and the pace of hardware acquisition. Nevertheless, the following table, which is the snapshot of offsite hardware resources available for CDF by summer 2004, shows a clear picture of emerging large offsite facilities and confirms a growing trend with respect to last year's document. In this table we only listed institutions that allow access to all CDF members, with the exception of the German

GRIDKA site where this, while being the institution policy, is still technically problematic since access has to be done via LCG GRID software.

Some sites are geared toward MC production especially as common CDF resources are concerned, so the local amount of disk is somewhat uninformative, we write 0.1TB as storage in this case to mean that those sites do not envision allowing significant data access to CDF users at large, at least at present. Local access policies are constantly evolving and will possibly change in the future as a result of political negotiations, hardware additions and our attempt to enforce a common policy.

Institution	CPU (GHz)	Disk (TB)	Access Gbit/sec	notes
Canada	250	0.1	2	(1) (3) (4)
Canada	1500	0.1	2	(2) (3) (4)
Germany (GRIDKA)	1895	20	1	(2) (4) (6)
Italy (CNAF)	900	30	1	(1) (5)
Japan (Tsukuba)	150	10	1	(1) (5)
Korea	120	1	0.1	(1) (4)
Taiwan	135	3	0.5	(1) (5)
Rutgers	100	4	0.2	(1) (5)
MIT	115	1	1	(1) (3) (4)
UCSD	280	5	1	(1) (4)
UCSD	200	5	1	(2) (4)
TOTAL	> 4000	80	-	

Table 14: Computing resources available offsite to CDF users by summer 2004. Notes: (1) reserved for CDF group (2) shared with other experiments (3) dedicated to MC production (4) allows unrestricted equal access to all CDF members (5) gives priority of usage to local CDF members (6) CDF may use from 10% up of the CPU at GridKa, the indicated disk is CDF-own.

In addition to what is shown in the table, CDF groups in the UK at University of London and Liverpool and in Spain at University of Cantabria and Barcelona/ICFA are setting up local CDF farms that are expected to be accessible to CDF members by end of 2004. Very likley, local users will retain privilege, but we plan to have all those sites embedded in the same common framework for access, monitoring, and accounting.

Winter 2004 has seen the first deployments of large offsite data disk pools with local copies of selected datasets. Now UCSD, INFN, Canada, and Taiwan host several dataset replicas in the few TB range.

Most off site institutions have a high speed local connection to the Internet, so access to those facility can be highly efficient. Experience shows that effective throughput on WAN can be limited by many hard-to-find bottlenecks. Our best experience so far is with the Canada-FNAL link that has shown reliably and consistently $\sim 200\,\mathrm{Mbit/sec}$ for production MC trans-

fer. which seem to represent the maximum presently allowed by the Fermilab-ESNET connection.

9.3 Offsite MC Production

At present, the primary use of off site computing resources for CDF as a whole is in the form of off site MC production, mostly in Toronto, UK, SanDiego and Italy. This is presently done by running MC jobs locally (*i.e.*, no GRID-like remote central control) and importing data back to FNAL by ftp to on site disks and hand write to Enstore. MC production in the summer of 2004 has also seen a increase in complexity, with the implementation of extensive run dependence in the preparation of large-scale samples. A side effect of this improvement is the addition of extra bookkeeping and concatenation steps required to satisfy tape-storage file-size requirements. This additional functionality has required a significant amount of additional human effort. The Canadians have MOU responsibility for coordinating the MC production, as well as providing 1 Million events per day capacity in Canada; above the guaranteed minimum of 1 Million events/day, the Canadian cluster may be, and in past has been, able to produce significantly higher rates of MC production. The Canadian facilities, however, are experiencing greater usage loads by other non-CDF users. For example, during ATLAS data challenge exercises, the capacity of the Canadian farms undergoes a substantial reduction. Other institutions (UK, *e.g.*) are considering the possibility of setting up analogous MOUs for taking up responsibility to produce a fraction of the CDF needed MC events. The MC production group has already produced $>$ 100 Million events in FY-03, about 60% in Canada and 30% in UK. MC production is proceeding at an increasing pace and more than 200M events have been produced in Canada and the UK in FY-04. We expect this to increase more than linearly with luminosity as we improve the accuracy of the simulation and physics analysis of Run II data matures. It is worth noting that as CDF reconstruction code evolves we do not only need to reprocess old data, but also to generate new MC samples with the same reconstruction version as used in analysis.

All MC generated in this fashion is coordinated via the physics groups, each of which has a MC representative in the simulation group where coordination of large-scale MC production occurs; but this by no means represents all the MC generated by CDF. Most analysis work needs very large MC production runs that are tuned to the particular topic and cannot be shared with others. These "single-user" MC samples are, and will be in the future, produced privately by individuals or small groups as needed and will not be managed by the physics groups.

The next steps for general CDF computing offsite has been providing users the means to generate their private MC samples, *e.g.* exploiting CAF installations at remote institutions, which is now possible. While at present significant amounts of user level MC are still produced on the FNAL CAF, which is a poor usage of a system built for good data access and tightly coupled to the main CDF data repository, most of that work can be done easily off site and results copied back to FNAL using current CAF tools. Indeed this has started happening during winter 2004 and we are now in the process of establishing accounting tools to quantify

this usage. Random sampling of offsite farms suggests that so far much more then 50% of their CPU time is being used for MC production.

Work is now undergoing to develop by end of Summer 2004 an "easy to use" user facility that will allow directly storing on FNAL's Enstore tape archive and cataloging into SAM the output of a MC run on a farm worker node running on any CAF around the world. Data will be moved using GridFtp re-using technology already developed for the JIM project. We have identified the major issues there in the need of guaranteeing a minimum file size for efficient tape and metadata operation, so in some cases storage on intermediate disk buffer and file concatenation will be needed. We envisage also to do this under SAM control, but plans are not final yet about the best way to organize the needed bookkeeping into SAM.

9.4 Offsite Data Analysis

In 2003 we presented the plan to expand, as our GRID project matures, remote sites into more general user analysis centers off site. This is now a reality with several sites having an established procedure to import large data sets to local disk caches and make them available to CDF users for analysis, so relieving CPU load and data access congestion from FNAL's CAF. We have experienced that it is more effective to preload specific data sets, lock them on local cache disk, and advertise their availability to users, rather then import data on demand according to random analysis jobs and end with a lot of cache misses. At present about 10TB of data are replicated offsite and advertised to CDF users via web pages. Work is in progress for a more clear and unified presentation of this information. Those datasets range from J/ψ samples to Inclusive Hadronic B, to high p_{T} leptons. Since winter 2004 CDF has formalized this distributed computing effort with the creation of "GRID Operation" meetings and of a political body, the "CDF Computing Resources Coordination Board" where representatives from each institution that contributes to the CDF-Grid can coordinate their resource deployment.

9.5 Toward a CDF Grid

9.5.1 The Vision

The long term vision for CDF computing is that users develop and debug their application at their desktop somewhere in the world. They then submit their job to SAM-grid, specifying a dataset to analyze in addition to the usual CAF information. SAM-grid selects an execution site based on locally available data as well as CPU resources. The user job is queued at the local site, and eventually instances of it start. SAM provides input data, and the user writes out her/his output into a local scratch area. After completion of an instance, the user may declare the files produced to the DH system for storage in a location where the user has sufficient quota to store the files on disk.

In addition, DH provides metadata catalogue services for the user's output data such that it can be reused as input for a future job by anybody in CDF. Data stored in this manner is not backed up to tape, and may be permanently erased by the user who owns the data. Tape

archiving would require in general additional concatenation to achieve file sizes for efficient operations of tape archive resources, and we plan to use SAM to develop tools that allow users to do this effectively.

As organizational principle we expect to be guided by the notion of physics centers rather than regional centers. Assuming we can arrange policies of ownership for resources at the level of "virtual center" it is much more efficient to concentrate a given dataset and the CPU resources required for its analysis in a set of dedicated sites rather than spreading all datasets across all sites more or less evenly. Needless to say, this ideal will require some amount of negotiation and deliberation to be successful. On the other hand this is exactly how present offsite institutions are organizing their computer clusters in a spontaneous way.

9.5.2 The Tools

Technically, we will accomplish this initially via a combination of SAM, dCache, Enstore, CAF, and JIM. At present, dCache read operations, Enstore, and CAF are fully in production. SAM is routinely used offsite and has been deployed for usage also in central CAF. JIM is still in development and test mode, but expected to ramp into production level support in FY05. In particular work is now in progress to use JIM and the RunJob utility to centrally manage distributed MC production across all CDF offsite farms. In addition to this batch based processing, we are also pursuing interactive computing, both in form of a central interactive platform at FNAL discussed in section 3, as well as an interactive GRID.

The interactive GRID computing effort is a collaboration between UCSD, INFN, and MIT. The goal is to build an interactive GRID computing system with response times to queries of order 10s for analyzing O(10GB) ntuples. The system is to be based on Root's PROOF tool by adding an interface to SAM metadata catalogue, and Condor/Globus GRID middleware, a first prototype was developed for SC2003 last November.

As the CDF-GRID takes form and is deployed offsite, we will have to interface to and incorporate tools already in use in those sites, primarily the LCG/EGEE emerging LHC GRID, so that CDF farms in Europe, Taiwan and elsewhere may be run as a sub-sect of larger LHC clusters sharing manpower and resources. Work toward this integration is already started and constitutes next frontier of the CDF Analysis Farm (CAF) development which is by now a joint project of INFN and UCSD.

9.5.3 The Financial Side

The CDF International Finance Committee has been debating at length the formalization of foreign contribution to CDF costs. The current position of the Committee is that the CDF's plan to move 50% of analysis work offsite by 2006 is reasonable and it is a matter of fact that several countries have already contributed resources to that goals and are considering plans to increase their commitment up to the indicated level. On the other hand it is clear that such contributions will be on a voluntary basis, in a best effort spirit, and quite likely there will be no MOU-kind document. At the same time proper accounting of the usage of remote resources is perceived to be fundamental for a fruitful collaboration both as guide-

line for efficient usage and acknowledgment of the contribution. Assuming that the present developments are indeed successful it is conceivable that we start a more global computing cost accounting in FY05.

Up to today, even in lack of such a formalization, CDF has nevertheless received substantial financial contribution by foreign countries, Japan and Italy mainly, who have done so for many years, more recently also UK, Switzerland, Korea and Canada have contributed. In particular concerning computing, Canada has taken on a serious commitment to MC production and is providing very significant computing resources for it, and so is the UK and many US institutions who have produced Monte Carlo samples for CDF data analysis since roughly December 2002. It is CDF's first priority to preserve all positive sides of how things have been working till now, and therefore to be very cautious and careful in defining a brand new policy.

10 Summary and Conclusions

Run II of the Fermilab Tevatron is now in it's fourth year. Luminosity is ramping up, data recording of the CDF experiment is stable, event reconstruction is operating smoothly, and data analyses have yielded a broad spectrum of results and first publications. However, only 10% of the Run II data has been recorded so far. While the most challenging part of Run II is still ahead, the compute systems of the experiments are operational and used very actively. Analysis strategies are being refined and the requirements models of the event tuned up. We have performed a budget estimate for fiscal year 2005 and updated the projections for the following years, FY-06 and FY-07. The equipment costs are summarized in Table 15. For FY-03 and FY-04 the costs are the actual spending. The cost estimates of FY-05 to FY-07 are based on the requirements model and computing plan described in this document to meet the computing need of the experiment.

Fiscal Year	Batch CPU ($M)	Inter. CPU ($M)	Farm CPU ($M)	Data-base ($M)	Tape Robot ($M)	Cache Disk ($M)	Net-work ($M)	Total ($M)
03	0.31	0.08	0.19	0.15	0.20	0.34	0.23	1.5
04	0.49	0.06	0.24	0.07	0.13	0.14	0.19	1.4
05	0.42	0.10	0.18	0.05	0.43	0.29	0.31	1.8
06	0.85	0.10	0.00	0.03	0.51	0.27	0.12	1.9
07	0.73	0.10	0.18	0.03	0.48	0.17	0.08	1.8

Table 15: CDF computing equipment spending summary. Numbers for fiscal years 2003 and 2004 are actual expenditures. For the last three fiscal years the costs are estimates.

At the end of FY-01 the CDF experiment switched its computing strategy to farms of commodity PCs for computing and storage to reduce the computing costs and to meet the ever increasing CPU demand of the experiment. The CAF and disk cache have been build up during the past years, the reason for the significant costs in FY-03 and FY-04. However, the costs to expand CAF and the data disk cache continue to dominate the total computing costs. For the next fiscal year, FY-05, we expect CPU demand of the CDF software and disk demand of the SAM data handling system to continue at or above this years allocation. Work on an interactive cluster started in FY-04. It is being commissioned and expected to grow significantly over the next years. A flat annual upgrade in dollars at $100k is estimated. The item will be adjusted as the system comes into production and we see how it is being accepted and used by the physicists of the experiment.

References

[1] R. Snider and R. Harris, "Implications of increased data logging rate on the CDF Run 2 computing plan and budget", CDF Note 6639 (2003).

[2] R. Harris, et al., “CDF Plan and Budget for Computing in Run 2”, CDF note 5914 (2002).

[3] The Run II Luminosity Upgrade at the Fermilab Tevatron, Project Plan and Resource-Loaded Schedule, June 15, 2003.

[4] http://tier2.ucsd.edu

[5] http://www.opensciencegrid.org