CDF/DOC/COMP_UPG/PUBLIC/5914

# CDF Plan and Budget for Computing in Run 2

Version 3
May 16, 2002

*Edited by*
Robert M. Harris
*Fermilab Computing Division*

*Contributions from*
William Badgett, Stefano Belforte, Phil Demar, Richard Jetton, Kevin McFarland, Don Petravick, David Tang, Jeff Tseng, Steve Wolbers and Frank Würthwein

**Abstract**

We discuss the plan for CDF computing in run 2 with an emphasis on the budget requirements necessary to meet the physics goals over the next 3-4 years. We consider primarily those areas that require continuing hardware purchases: central analysis facilities, data handling, reconstruction farms, networking and databases.

# Contents

# 1 Introduction

This note presents an overview of the CDF computing plan for the next few years. It is intended as documentation for the Directors Review of Run 2 Computing on June 4 - 6, 2002. Computing expenditures are needed to move, store and analyze increasing amounts of data expected from the Tevatron. We note that the requirements depend largely on the integrated luminosity, and we have adopted a "plan for success" strategy based on table 1, the official luminosity goals of the lab. Run 2 is normally divided into two running periods: run 2A (2 $fb^{-1}$) and run 2B (13 $fb^{-1}$) with a six month shutdown between the two.

| Fiscal Year | 02 | 03 | 04 | 05 | 06 | 07 | 08 |
|---|---|---|---|---|---|---|---|
| Delivered Luminosity ($fb^{-1}$) | 0.3 | 0.9 | 1.3 | 1.6 | 3.5 | 3.7 | 3.7 |
| Total Luminosity ($fb^{-1}$) | 0.3 | 1.2 | 2.5 | 4.1 | 7.6 | 11.3 | 15.0 |

Table 1: The luminosity goals of the Tevatron for Run 2 [1].

We have written a detailed budget plan, estimating our requirements, what we plan to procure in each fiscal year, and providing estimates of how much it will cost. The long range plan is grounded in our current direction, but given the rapid pace of technological advance, and what we will learn in the course of run 2, the probability of the details being correct drop quickly as a function of time. For example, the technological choices just two years ago were large SMP machines, expensive disks, and commodity tape drives in our central systems. That direction has now changed significantly, and we are deploying commodity computing, inexpensive disks, and data center quality tape drives. The change was a result of our experience with the needs of CDF computing in run 2, which required a system designed with more CPU, more disk and more robust tape storage systems than was previously anticipated. Fortunately we were able to take advantage of advances in Linux based commodity computing that were not previously foreseen. Therefore, the reader should not consider this detailed plan and budget for the next 3-4 years as cast in stone. We humbly submit it as our understanding at this time.

# 2 Computing and Analysis Model

A conceptual view of the major computing elements and data-flow at CDF is pictured in Figure 1. Although incomplete, figure 1 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 in the tape robot is read by the *production farms* where it is reconstructed and the resulting *reco* data is written back to 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*. Currently the batch jobs are on fcdfsgi2, a 128 processor SGI, but in the near future the jobs will primarily be on a user analysis farm of PCs. Some of the reconstructed data, and the majority of secondary datasets from the reconstructed data, are also stored on *static disk* where it is accessible by the batch CPU. The batch CPU produces secondary datasets and root *ntuples* and writes them to the static 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 ntuples 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. Currently on fcdfsgi2 there is no distinction between batch and interactive CPU, but the user analysis farm is exclusively batch. Users desktops can also obtain data directly from the tape robot via read disk caches, write them back to the tape robot via write disk caches (not shown), and transfer ntuples 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 ntuples 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 ntuples. Users have access from their desktops and the interactive CPUs to the datasets on static disk. The interactive CPU provides a controlled environment for debugging and job submission.

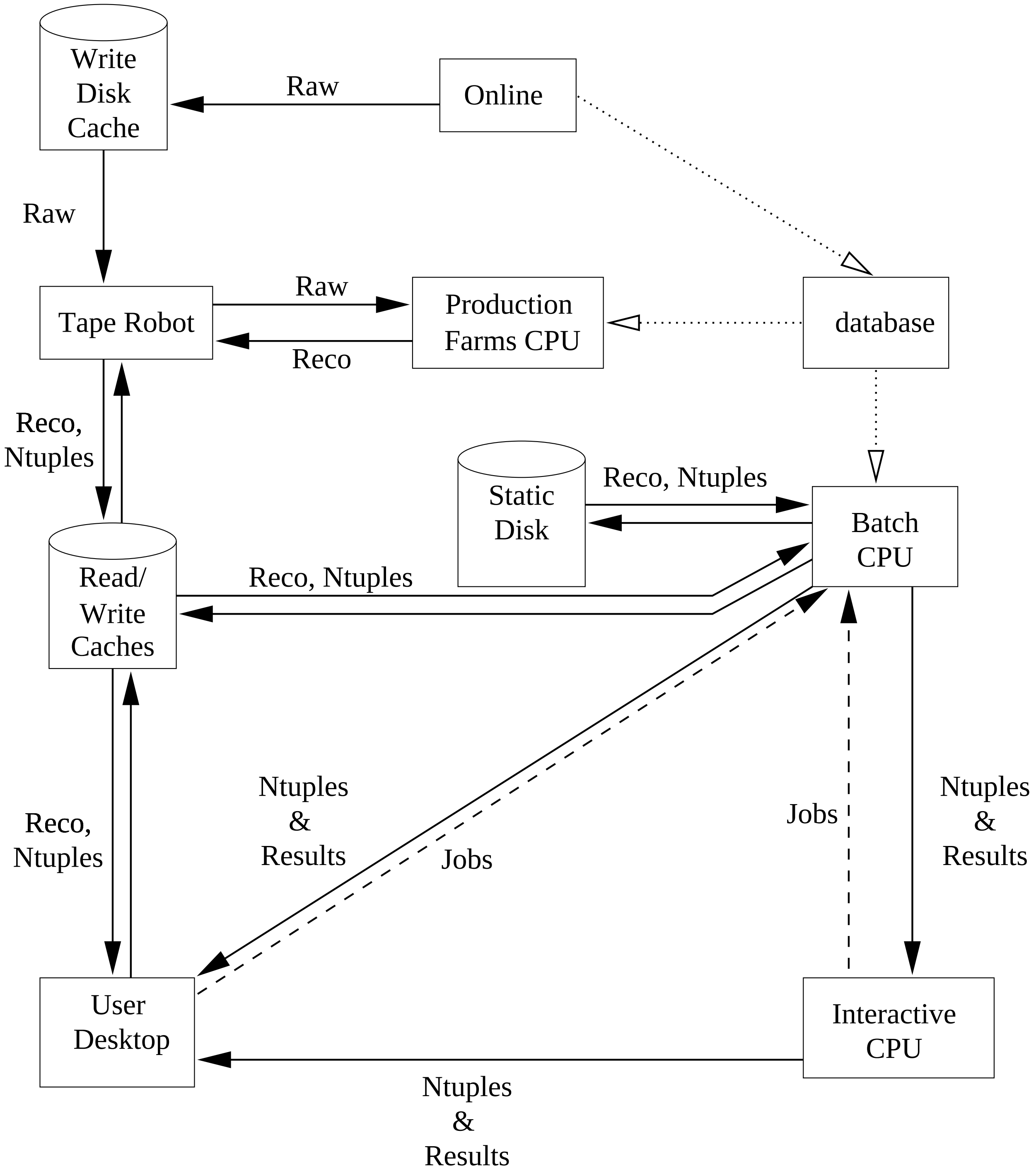


Figure 1: A simplified picture of the CDF Computing Model. Major computing elements in boxes, raw and reconstructed data flow indicated by solid lines, database constant flow indicated by dotted lines, and job flow indicated by dashed lines.

# 3 Requirements

The total computing requirements as a function of fiscal year depends primarily on the integrated luminosity. We model this with a linear dependence and correct for commissioning effects. In table 2 we present the slope and offset of this relation.

| Requirement | Offset | Slope |
|---|---|---|
| Batch CPU (GHz) | 66 (GHz) | 1100 (GHz/fb$^{-1}$) |
| Static Disk (TB) | 32 (TB) | 125 (TB/fb$^{-1}$) |
| Read Cache (TB) | 12 (TB) | 35 (TB/fb$^{-1}$) |
| Write Cache (TB) | 5 (TB) | 10 (TB/fb$^{-1}$) |
| Disk I/O (GB/s) | 368 (MB/s) | 1100 (MB/s-fb$^{-1}$) |

Table 2: The offset and slope of the linear relation between integrated luminosity and computing requirements. There are corrections in FY 2002 for commissioning effects (see text). The farms and tape archive requirements are not listed here because they required a more complex model than a linear dependence on luminosity.

In table 2 the slope of the batch CPU requirement comes from the physics needs [2] determined by the review of our central analysis facilities [3]. That review estimated that for 2 fb$^{-1}$ the needs were in the range 1-2 THz. The upper estimate of 2 THz permits each of 200 users to be able to process their 5 nb data sample in a single day with a typical current rate of 10 events per second on a GHz processor:

$$2\ THz = \frac{200(users/day)\ 5(nb)\ 2(fb^{-1})}{10\ (ev/sec - GHz)}\ \frac{10^6(nb^{-1}/fb^{-1})\ 10^{-3}(THz/GHz)}{10^5(sec/day)}$$

The lower estimate of 1 THz allows for some improvement in software speed. We have used the mid-point of that range (1.5 THz), and added 1/3 contingency and a 10% tax for R&D, giving the 1100 GHz/fb$^{-1}$ tabulated. The batch CPU offset of 66 GHz are 26 duals contributed to the stage 1 CAF by non-Fermilab institutions, discussed in the next section.

Our overall disk strategy is to provide enough disk for both a model of putting all frequently used secondary datasets on static disk, while simultaneously providing a large enough disk cache for a heavily cached model of data access. Both models will likely need to coexist, and CDF's analysis capabilities will be enhanced by providing for both. While we will try to put the majority of the datasets on static disk, a large disk cache will still be necessary for those datasets that the physics groups are not able to fit on disk, for secondary and tertiary datasets that are in the process of being formed and identified as needed on static disk, and for those users accessing large datasets outside of the organized physics group effort. A large disk cache is also a safety net in case estimates of static disk needs are too low, which is almost always true, and caching must be done more heavily than anticipated. We can also afford this approach because we have moved from costly fiber channel attached SCSI disk to commodity network attached IDE disk. The risk incurred in this transition is taken for the commensurate reward of plentiful quantities of cheap disk.

In Table 2 the slope of the static disk requirement is derived from the total disk space requests of the CDF physics groups. We scaled to 1 fb$^{-1}$the disk space requests of the physics groups for 100 $pb^{-1}$ and 400 $pb^{-1}$ [4]. For every fb$^{-1}$, the physics groups requested 62 TB of disk space for data assuming 100 KB event sizes. The physics groups also requested 33 TB of disk space for Monte Carlo and 30 TB for other purposes for every $fb^{-1}$, giving a total of 125 TB/$fb^{-1}$. If we allocated to each physics group the amount of disk they requested the percentage of disk by physics group would be: Bottom 40%, QCD 29%, Top/Electroweak 17% and Exotics 14%. For comparison, the expected size of all the CDF reconstructed primary data is roughly 100 TB/$fb^{-1}$ for 100 KB events, so the slope of 125 TB/$fb^{-1}$ could alternately be understood as what is required to place all the reconstructed data on disk and have an additional 25 TB/$fb^{-1}$ for other uses. The offset of the static disk requirements is the roughly 32 TB currently available for static datasets and R&D.

In Table 2 the slope of the read cache disk requirement comes from the physics needs document [2], where it was estimated that a read cache of 70 TB was needed for 2 $fb^{-1}$. The offset of the read cache is our current 12 TB disk cache on fcdfsgi2. The write cache requirements are typically 30% of the read cache requirements, giving the slope tabulated, and the offset is the combination of our current write caches for the online data logger (2.5 TB), the farms (2 TB) and the users on fcdfsgi2 (roughly 0.5 TB). The slope of the disk I/O requirement comes from the the CPU slope of 1100 GHz/$fb^{-1}$ times 1 MB/sec/GHz for a typical user analysis job (running at 10 Hz on 100 KB events on a 1 GHz processor). The offset of the disk I/O requirement, 368 MB/s, is the requirement for the stage 1 CAF discussed in the next section, plus additional requirements to support anticipated contributions from remote institutions to the CAF in FY02. The data archiving and farms requirements are handled separately in sections 5 and 6 since they do not have as simple a dependence with the integrated luminosity, but instead depend primarily on the rate we log data.

The requirements that depend on the luminosity are corrected for commissioning effects. We require an additional 33% for FY 2002 for all disk requirements and for the batch CPU requirements, needed for the commissioning of this new central analysis facility. Combining the integrated luminosity expectations of table 1 with the luminosity slopes and offsets of table 2, and adding in the commissioning effects we arrive at the integrated computing needs in table 3. The archive volume, archive I/O, and reconstruction farms CPU needs are from more complex analysis in sections 5 and 6, but are summarized in this table for convenience. The integrated needs drive the purchasing plans and the two are compared in the remainder of this document.

| FY | 02 | 03 | 04 | 05 | 06 | 07 | 08 |
|---|---|---|---|---|---|---|---|
| Batch CPU (THz) | 0.51 | 1.5 | 2.9 | 4.7 | 8.6 | 12 | 17 |
| Farm CPU (THz) | 0.37 | 0.70 | 0.76 | 1.3 | 2.1 | 2.5 | 2.8 |
| Static Disk (TB) | 82 | 180 | 340 | 540 | 980 | 1400 | 1900 |
| Read Cache (TB) | 26 | 54 | 100 | 160 | 280 | 410 | 540 |
| Write Cache (TB) | 9 | 17 | 30 | 46 | 81 | 120 | 160 |
| Disk I/O (GB/s) | 0.81 | 1.7 | 3.1 | 4.9 | 8.7 | 13 | 17 |
| Archive Volume (PB) | 0.3 | 0.7 | 1.1 | 1.7 | 2.9 | 4.1 | 5.3 |
| Archive I/O (MB/s) | 65 | 190 | 130 | 480 | 350 | 470 | 590 |

Table 3: The integrated computing needs by the end of each fiscal year as a function of fiscal year.

# 4 Central Analysis Facility

In figure 1 we define the Central Analysis Facility (CAF) [5] as primarily the batch CPU, the static disk, and the interactive CPU. The CAF review determined that for analyzing the data from 2 fb$^{-1}$ we needed an order of magnitude more computing power than could be provided by large SMP machines within our budget. For example, to satsify just the FY02 requirements for batch CPU of 0.5 THz with Sun or SGI SMP machines would cost roughly \$10 million. Instead the committee recommended a technological direction based on commodity PC Linux farms and network attached IDE disk [3].

## 4.1 Batch CPU and Network Attached Disk

The procurement plan for the CAF is summarized in Table 4 and Table 5. As of May 2002 the protoCAF, the proto-type of the user analysis farm consisting of 16 duals and a 1.4 TB fileserver, had been running for two months in the CDF trailers before being moved to FCC.

Stage 1 of the CAF on the second floor of FCC is being commissioned in May 2002, intended for steady operations in June, to be used for analysis of data for Summer conferences. The CPU consists of 43 1.3 GHz duals from Fermilab at a cost of \$108K and 26 duals contributed by non-Fermilab institutions at roughly a cost of \$60K. The static CAF disk starts with 1.4 TB from a single fileserver contributed by MIT at a cost of \$10K. Stage 1 in table 5 consists of 21 TB of static disk from 11 fileservers. The filseservers cost \$10K each and currently have a usable capacity of 1.92 TB per fileserver from a dozen usable 160 GB disks arranged in two filesystems per fileserver. In the near future we expect the 1 TB installation limit per filesystem to be overcome, and then the same fileservers will have a 2.24 TB capacity using seven disks per filesystem. Four of the eleven fileservers were contributed by a non-Fermilab institution at a cost of \$40K. Also included in the Stage 1 entry for CPU in table 4 is the purchase of a code server for the CAF with 2 TB of disk for \$19K. For spending clarity we are adding the $60 + 10 + 40 = \$110K$ from non-Fermilab institutions into both the FY2002 cost and the FY2002 budget. We are hoping for roughly \$360K in non-Fermilab contributions to the CAF in FY02.

| Stage | FY | Needs (THz) | Duals Bought | Duals Total | Speed (GHz) | CPU (THZ) | Total (THz) | Cost (\$M) |
|---|---|---|---|---|---|---|---|---|
| 1 | 02 | 0.1 | 69 | 69 | 1.3 | 0.18 | 0.18 | 0.19 |
| 2 | 02 | 0.5 | 160 | 229 | 1.8 | 0.58 | 0.76 | 0.40 |
| 3 | 03 | 1.5 | 192 | 421 | 2.5 | 0.96 | 1.72 | 0.48 |
| 4 | 04 | 2.9 | 192 | 613 | 3.5 | 1.34 | 3.06 | 0.48 |
| 5 | 05 | 4.7 | +240 -229 | 624 | 5.0 | 2.40 | 4.7 | 0.60 |

Table 4: CAF CPU Procurement plan. The stage, fiscal year,total batch CPU needed, dual CPUs bought, dual CPUs total, speed of each CPU, incremental CPU added at that stage, total CPU available, and cost of incremental CPU.

Stage 2 of the CAF, to be procured in fiscal 2002, will scale the CAF to the level needed for analyzing data for winter conferences in early 2003. Roughly 160 duals (CD/OSS has evaluated 1.5 - 2.0 GHz nodes) and 72 TB of network attached disk will be added. The hardware will be procured after some experience with stage 1 has been attained by July

| Stage | FY | Needs (TB) | Needs (GB/s) | Servers Bought | Servers Total | Server (TB) | Disk (TB) | Total (TB) | Total (GB/s) | Cost ($M) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 02 | 30 | 0.37 | 11 | 11 | 1.9 | 21 | 21 | 0.6 | 0.12 |
| 2 | 02 | 82 | 0.81 | 32 | 43 | 2.2 | 72 | 93 | 2.2 | 0.35 |
| 3 | 03 | 180 | 1.7 | 32 | 75 | 3.5 | 112 | 205 | 3.8 | 0.35 |
| 4 | 04 | 340 | 3.1 | 32 | 107 | 5.5 | 176 | 381 | 7.0 | 0.35 |
| 5 | 05 | 540 | 4.9 | +32 -43 | 96 | 8.7 | 278 | 558 | 8.0 | 0.35 |

Table 5: Static Disk Procurement plan. The stage, fiscal year, static disk space needs, file server bandwidth needs, fileservers bought and obsolesced(-), fileservers total, the server capacity, incremental disk added at that stage, total disk available, total file server bandwidth available, and cost of incremental disk added.

2002. The subsequent stages of the CAF will have to be located on the first floor of FCC, since there is inadequate space on the second floor to support it. This space and power is scheduled to be available in August-September 2002. The two 8-ways for stage 1 will continue to be used for interactive logins for stage 2, most likely augmented with additional scratch disk.

Throughout table 4 we have assumed a Moore's law doubling of processor speeds every 18 months, and a fixed cost of $2.5K per dual. Similarly, throughout table 5 we have assumed a Moore's law doubling of disk capacity per fileserver every 18 months, and a fixed cost of $10K per fileserver. We believe this is conservative since the trend is for disk capacities to more than double every 18 months. In table 4 we replace duals every 3 years, both to fit the CPU in the space available in FCC and for ease of management, so in FY 2005 we replace the duals purchased in FY 2002. Similarly in table 5 we replace fileservers every 3 years. The file server bandwidth capabilities in table 5 come from the total number of fileservers times 50 MB/s currently achieved per fileserver over GBit Ethernet, times the assumed number of GBit Ethernet connections for each type of fileserver in table 16. The purchasing plan each year meets or slightly exceeds the requirements.

## 4.2 Interactive CPU and Legacy Systems

As of May 2002, the run 2 central analysis facility consisted of the legacy systems fcdfsgi2, a 128 processor 300 MHz SGI SMP with roughly 30 TB of disk, fcdflnx1, a four processor 700 MHz Intel/Linux SMP with a TB of disk, fcdfhome, a NETApp serving 0.5 GB of disk per user for the home areas, and fcdfspool a NETApp serving 1 GB of disk per user for spool. Another 15 TB of SCSI disk were available and scheduled to be connected to fcdfsgi2 by May 2002. As of May 2002, two 8-node 700 MHz Intel/Linux SMP boxes are commissioned and serving as interactive CPU with a TB of scratch space each, and a dual Intel/Linux box is being used as a code server.

The interactive CPU plans for fiscal years 2003 - 2005 are still being developed. The two 8-ways will likely be augmented or replaced with other Linux N-ways or with a login pool of cheaper duals and some scheme for sharing disk among them for the convenience of users and ease of management. As the powerful Linux based CAF becomes the predominant means of data analysis, we expect that users are going to become less interested in 300 MHz processors under IRIX on fcdfsgi2 as an interactive system. Further, the lab pays $176K per year for the maintenance contract with SGI for all 128 processors of fcdfsgi2, and the cost scales with the number of processors. The maintenance costs will be a strong incentive to decommission some of the fcdfsgi2 processors by late FY2003. As the interactive CPU power of fcdfsgi2 decreases, we will need a replacement interactive system with reasonable amounts of CPU and local disk.

The spending plan on interactive CPU and associated local disk is shown in table 6. In FY2002 the two 8-node 700 MHz Intel/Linux systems we purchased cost $70K. We did not purchase any local disk for those systems, but borrowed some of the legacy fiber channel connected SCSI disk discussed below. Throughout FY2002 fcdfsgi2 will have it's full complement of 128 processors, and lots of local disk, so the need for additional interactive CPU and local disk is limited. By late FY2003 we plan to decommission 64 fcdfsgi2 processors, increasing the need for Linux based interactive CPU and local disk, and our planned expenditure doubles to $150K. Similarly in FY2004 and FY2005 we continue the decommissioning of fcdfsgi2, and although we expect the cost of CPU to be roughly constant it is likely we will need to buy substantially more local disk to replace obsolete disk that was used on fcdfsgi2.

| FY | Legacy CPU (processors) | Interactive CPU ($M) | Local Disk ($M) | Total ($M) |
|---|---|---|---|---|
| 2002 | 128 | 0.07 | 0.00 | 0.07 |
| 2003 | 64 | 0.10 | 0.05 | 0.15 |
| 2004 | 32 | 0.10 | 0.10 | 0.20 |
| 2005 | 0 | 0.10 | 0.10 | 0.20 |

Table 6: Interactive CPU Procurement plan. The fiscal year, number of processors remaining in fcdfsgi2, interactive Linux based CPU purchased, local disk purchased for that CPU, and total cost of Interactive CPU and local disk.

Not listed in table 6 are the following items for the legacy central analysis facility. The return in FY2002 of a 24 processor sun getting us back $364K, the upgrade of fcdfsgi2 from 64 to 128 processors costing us $372K, the purchase of 15 TB of fiber channel attached SCSI disk for $243K (and $32K for the fibers, hubs and adapters to connect them to fcdfsgi2), and contributions of 2.8 TB of disk on fcdfsgi2 from non-Fermilab institutions totaling $40K. For completeness in discussing the budget for FY2002 we mention that we received $158K in cost transfers from non-Fermilab institutions, and that we did not spend $422K of our FY2001 budget, due to a transition in our technology direction, and $228K has been returned to us by the division now and the other $194K is planned for us before the end of the fiscal year. Summing these numbers gives another $364 + 40 + 158 + 422 = \$984K$ in budget for FY2002 and an additional $372 + 243 + 32 + 40 = \$687K$ in spending on legacy CAF hardware.

# 5 Data Handling

In figure 1 we define Data Handling (DH) as primarily the tape robot, write disk caches and read disk caches.

## 5.1 Requirements

The requirements for data handling are primarily driven by the storage requirements, which are estimated in table 7. The size of the data archive required is driven primarily by the raw data logging rate, which also determines the reconstructed output rate once we take account of event compression and reprocessings, and also drives the creation rate of secondary datasets and the need for Monte Carlo. The run 2B detector upgrade aims at tripling the trigger and DAQ output capability, and in table 7 we assume that the data logging peak rate capability will triple from 20 MB/s to 60 MB/s in FY05. The raw data logging rate tabulated is the product of the peak logging rate and the overall running efficiency of the accelerator and the detector taking into account all down-times. The running efficiency is typically 30% when the accelerator and detectors are performing well and understood, and about half that during the commissioning periods of FY02 and FY05. In table 7 the number of processings on the farms starts out large in commissioning periods and reduces with experience to the typical value of 1.3 processings performed in run 1 ( 30% of the data is reprocessed once). The compression factor, the ratio of farms output event size to raw data event size, is currently 0.7 (250 KB events on input and 170 KB events on output) and the goal is to decrease to 0.4 (100 KB events on output). The secondary datasets and ntuples are roughly half the size of the primary reconstructed datasets. The Monte Carlo dataset rate is calculated assuming 100 KB events produced on the farms at the rate shown in table 13, which scales with integrated luminosity as discussed in section 6, but ends up being similar to the rate of secondary datasets. Adding contributions from raw, reconstructed, secondary and Monte Carlo data gives the integrated archive in table 7. The archive up through FY04 is roughly the $0.5PB/fb^{-1}$ anticipated by previous estimates [6], and beginning in FY05 we anticipate significantly more data, driven by the increased rate capability of the DAQ.

The archive I/O rate requirement is also estimated in table 7. In addition to the logging rates we have discussed above, there are rates for farm reading, rates for user reading, and the rate to copy media to periodically increasing storage densities. The farms reading rate tabulated follows directly from our previous numbers for raw data logging and reprocessings. The current user reading rate out of the robot is on average 1 TB/day, or 10 MB/s, for an archive volume of 0.1 PB. If our model is one of accessing data primarily from the robot, using a disk caching system, then the needed rate may scale with data volume, which is what we have assumed in table 7. The number for the user rate through FY04 is similar to that anticipated previously [6]. Since we hope to replace disk caching with static disk as much as possible for frequently accessed datasets, our rate requirements for the robot will likely be less than what is listed in table 7. The rate requirements for copying the data to larger storage densities follows from the tape media plan in section 5.3. The rate listed is the sum of the rate needed to both copy the data out of the robot to disk and then to copy it back into the robot using faster drives with larger storage densities, and to complete the copy in 3 months time.

## 5.2 Tape Robot & Drives

In ref [7] we discuss the stage 1 tape handling system based on an STK powderhorn robot with 10 STK T9940A drives and using Enstore for networked access.

The stage 2 tape robot and drives purchase will be driven by the need to increase the storage capacity. The STK robot has 5500 slots which we use for 60 GB T9940A cartridges for a total capacity of 330 GB. In May 2002 CDF had 100 TB of data and anticipates writing

| Fiscal Year | 02 | 03 | 04 | 05 | 06 | 07 | 08 |
|---|---|---|---|---|---|---|---|
| Peak Logging Rate (MB/s) | 20 | 20 | 20 | 60 | 60 | 60 | 60 |
| Running Efficiency | 0.15 | 0.3 | 0.3 | 0.15 | 0.3 | 0.3 | 0.3 |
| Raw Data Logging (MB/s) | 3 | 6 | 6 | 9 | 18 | 18 | 18 |
| Processings | 2 | 1.5 | 1.3 | 1.5 | 1.3 | 1.3 | 1.3 |
| Compression | 0.7 | 0.5 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 |
| Reconstructed Data (MB/s) | 4 | 5 | 3 | 5 | 9 | 9 | 9 |
| Secondary Data (MB/s) | 2 | 2 | 2 | 2 | 5 | 5 | 5 |
| Monte Carlo (MB/s) | 0.4 | 1 | 2 | 2 | 5 | 5 | 5 |
| Total Logging (MB/s) | 9 | 14 | 13 | 18 | 37 | 37 | 37 |
| Archive/year (PB) | 0.3 | 0.4 | 0.4 | 0.6 | 1.2 | 1.2 | 1.2 |
| Integrated Archive (PB) | 0.3 | 0.7 | 1.1 | 1.7 | 2.9 | 4.1 | 5.3 |
| Farms Reading Rate (MB/s) | 6 | 9 | 8 | 14 | 24 | 24 | 24 |
| User Reading Rate (MB/s) | 30 | 70 | 110 | 170 | 290 | 410 | 530 |
| Copying Rate (MB/s) | 20 | 100 | 0 | 280 | 0 | 0 | 0 |
| Archive I/O (MB/s) | 65 | 190 | 130 | 480 | 350 | 470 | 590 |

Table 7: The data handling archive needs. As a function of fiscal year, the peak online logging rate of raw data, the total running efficiency, the resulting average raw data logging rate, the number of times the raw data is reconstructed, the compression factor going from raw to reconstructed data, the average rate of writing reconstructed data, the average rate of writing secondary data and ntuples, the average rate of writing Monte Carlo data, the total rate of all data written, the total archive written that fiscal year, the total archive in the tape robot, the rate the farms read data from the robot, the rate the users read data from the robot, the rate needed for copying data from one storage density to another, and the total rate into and out of the robot.

between 50 and 150 additional TB between now and October 1, for a total usage of 150 - 250 TB by the end of the fiscal year. We have three options for additional capacity which we list in order of increasing technical risk and decreasing cost.

1. A new STK robot with T9940A drives.

   Our default plan is to purchase another STK powderhorn robot with 10 T9940A drives and situate it in the FCC Mezzanine, for a total cost of $470K. This will give us another 330 TB of storage with 100 MB/s of rate capability, with additional rate capability immediately available from moving drives from the existing robot as the rate needs of that data diminish.

2. A new STK robot with T9940B drives.

   The new T9940B drive from STK, planned for beta testing at Fermilab in May 2002 and for production by July 2002, has a rate and storage capability triple the existing T9940A drives at roughly the same cost. The media is the same as the T9940A media. If testing by ISD indicates this drive is acceptable before August 1, 2002, we would purchase 8 this fiscal year in addition to the STK robot, for an additional storage capacity of 1.1 PB and an additional rate capability of 240 MB/s using GBit interfaces for each drive, at a total cost of $430K.

3. LTO drives in our existing ADIC robot.

   LTO drives are currently being used in the D0 ADIC robot for Monte Carlo data with the intent of using them for collider data later in the year. ISD is in the course of procuring 6 LTO drives to test in the CDF ADIC robot once the AIT-2 drives and tapes have been removed. The LTO drives currently have a rate of 15 MB/s and write to 100 GB cartridges and at $8K per drive cost a fraction of an STK T9940 drive. If ISD's testing reveals that the LTO drives robustness is competitive with STK, we can purchase 25 drives for our ADIC robot at cost of $200K, and run them at 10 MB/s each using existing fast Ethernet interfaces for each drive.

In table 8 we present the procurement plan for robots and drives assuming option 1 above is followed. This is the least risky and most costly option. In stage 1 we reused our run 1 STK silo for run 2 with 10 STK T9940 drives for a total capacity of 0.33 PB and bandwidth of 100 MB/s. In stage 2 in FY2002 when we fill the first robot we simply write new data to a second robot for a total capacity of 0.66PB. Stage 3 occurs in FY2003, at the end of which the total data accumulated is estimated to reach 0.7 PB exceeding the 0.66 PB of the stage 2 system. Accommodating this additional data is straightforward in mid FY2003, since the STK T9940B drives will have been in production for around a year by that time and we can use their higher storage capability. In stage 3 we plan to upgrade the first of our two robots with T9940B drives which can write 200 GB tapes. We move the T9940A drives to the second robot. Most of the data in the first robot will need to be rewritten at the higher densities of 200 GB per cartridge, and other data may be designated as either too old or uninteresting and moved outside the robot. This first robot will have a renewed capacity of 1.1 PB, and the robot procured in stage 2 with the older drives will have a capacity of 0.33 PB, giving us 1.43 PB capacity by the end of FY2003. Then in FY2004 in stage 4 we replace all the T9940 drives in the second robot with 10 T9940B drives, discarding the obsolete T9940 drives, and giving us a total capacity of 2.2 PB keeping us comfortably ahead of the 1 PB expected by the end of 2004. In FY2005 in stage 5 we imagine a T9940C drive with 2 times the storage capacity and bandwidth of the T9940B, a 400 GB cartridge and 60 MB/s drives. As we did in stage 3 for the A to B transition, we buy 10 T9940C drives for the first robot increasing our storage capacity to $2.2 + 1.1 = 3.3$ PB and our bandwidth capacity to $600 + 600 = 1200$ MB/s. In FY06 and beyond we rely completely on the T9940C technology to supply the remaining storage capability we need, and if the data exceeds 4.4 PB we buy a third robot. Note that each year the rate in table 8 easily exceeds the rate requirements in table 7.

In addition to the costs for production systems, table 8 shows $50K per year for R&D on the next generation systems and alternative systems. For example, as we mentioned above, in FY2002 we are purchasing 6 LTO drives to study that option for the ADIC robot.

## 5.3 Tape Media and CDF Operating Costs

The total operating costs for tapes are summarized in table 9. The legacy costs for tapes were $1.3/GB for AIT-2 tapes. The costs for tapes are around $75 per 60 GB tape cartridge for STK 9940A drives, also giving $1.3/GB. In FY02 we copied 0.1 PB of data from our problematic AIT-2 tapes in the ADIC robot onto STK T9940A tapes in the STK robot. We expect as much as 0.3 PB total of data in FY02, of which all will be on 9940A tapes. The total cost includes both AIT-2 and STK T9940A media. In FY02, if we buy more STK

| Stage | FY | Data (PB) | Robots Total | Drives Bought | Drives Total | Storage (PB) | Rate (MB/s) | Cost ($M) | R&D ($M) | Total ($M) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 02 | .1 | 1 | 10 A | 10 A | 0.33 | 100 | 0.25 | - | 0.25 |
| 2 | 02 | .3 | 2 | 10 A | 20 A | 0.66 | 200 | 0.47 | 0.05 | 0.52 |
| 3 | 03 | 0.7 | 2 | 10 B | 20A 10B | 1.43 | 500 | 0.30 | 0.05 | 0.35 |
| 4 | 04 | 1.1 | 2 | 10 B | 20B | 2.2 | 600 | 0.30 | 0.05 | 0.35 |
| 5 | 05 | 1.7 | 2 | 10 C | 20B 10C | 3.3 | 1200 | 0.30 | 0.05 | 0.35 |

Table 8: Robot procurement plan. The stage, fiscal year stage procurement begins, total data accumulated up to the end of that fiscal year, total robots available, drives bought in that stage (A are T9940A drives and B are T9940B drives), drives available after procurement, storage capacity of robots and media in that stage, total rate available, cost of that stage, cost of R&D, and total cost.

T9940A drives in a second STK silo as planned, then in FY03 we will continue to write tapes at T9940A densities for the first half of FY03, writing 0.2 PB of data at a cost of \$1.3/GB. In mid-FY03 we plan to install STK 9940B drives, which will write 200 GB to the same \$78 cartridge, for a cost of \$0.4/GB. We will write the additional 0.2 PB of FY03 data at these higher densities, and then to be able to increase the storage capacity of our two robots we will copy all 0.5 PB off of the 9000 tapes at T9940A densities and onto tapes at the higher T9940B densities. The total cost in FY03 reflects the large costs to support both the T9940A and the T9940B media for the majority of our data, but as a result we benefit in FY04, when only 0.4 PB of tape at a cost of \$0.4/GB needs to be purchased. It is possible that the FY04 cost will even be zero, since we may be able to reuse the 9000 tapes written at 9940A densities. Alternately, for data security purposes, we can save all the 9940A tapes outside of FCC so that if there is a catastrophe that destroys the entire FCC building (fire, tornado, terrorist attack, etc.) our 0.5 PB of data from the first half of Run 2A will not be lost. In FY05 we again purchase a new drive type, STK T9940C, with a hypothetical cartridge capability of 400 GB and a media cost of \$0.2/GB. Only half of the data in FY05 is written to STK T9940B, and then all the data in the archive is copied to STK T9940C, and again the 1.4 PB of 9940B tapes can either be reused (if that is possible) or stored outside of FCC for data security purposes. Then again in FY06, not shown in the table, the media cost drops to roughly \$0.22M, benefitting from the low cost of only writing incremental tape volumes for the year. Thus the operating cost for tapes should oscillate between roughly \$0.5M per year and \$0.2M per year, for an average cost of roughly \$350K per year, in this model of tape technology evolution and use.

| FY | AIT-2 (PB) | 9940A (PB) | 9940B (PB) | 9940C (PB) | Cost ($M) |
|---|---|---|---|---|---|
| 02 | .1 | .3 | - | - | 0.52 |
| 03 | - | .2 | .7 | - | 0.54 |
| 04 | - | - | .4 | - | 0.16 |
| 05 | - | - | .3 | 1.7 | 0.46 |

Table 9: Tape procurement plan. The fiscal year, data written to AIT-2 tapes, 9940A tapes,9940B tapes, 9940C tapes and the total cost that FY for tape purchases.

Tapes are not the only CDF operating costs, since we typically spend \$100K each year on miscellaneous costs of operations (racks, cables, installations, misc. cards, consultants, etc.) and \$50K each year for desktops for Fermilab physicists at CDF (paid for by the computing division regardless of the physicists division). Combining the average years tape cost of about \$350K, with the operating costs of \$100K, and \$50K for desktops, we arrive at roughly \$500K per year for CDF operating expenses, with the reminder that there are significant year-to-year variations depending on tape technology.

### 5.4 Read Disk Caches

Read caches will be used wherever there are substantial data analysis needs. This includes the batch CPU and the interactive CPU in the CAF and the desktops in the CDF trailers. The disk caches need to be network accessible fileservers. As of May 2002 a prototype system [8] had been tested and was working at CDF, with a production system planed for June. The prototype used software from CDF, and dCache from DESY and CD/ISD, to access 3 network attached disk pools of 450 GB. In May 2002 a different prototype using SAM as well as dCache had been used to access data on a similar disk cache, with a more complete test planned for June 2002. Both DH systems, when they reach production stage, will use the same fileservers for read caches as are being purchased for the CAF.

The read disk cache procurement plan is summarized in table 10. As of May 2002 CDF has a legacy read disk cache of 12 TB of directly attached disk on fcdfsgi2 managed by the disk inventory manager. Our experience has shown this cache size to be adequate for current needs, with a cache lifetime of roughly a week, but the technology will need to be replaced with networked attached disk of about the same size in FY2002. We plan to do this in two steps. In Stage 1 we procure 4 fileservers and use the 8 TB of disk for the first operational system in June 2002. This read cache should be sufficient to replace the 12 TB of directly attached disk cache on fcdfsgi2, which can be reused as static disk by the physics groups. The stage 1 read cache will also serve the data caching needs of the stage 1 CAF. A second purchase in FY2002 will increase the total read cache to 26 TB. All purchases will be combined with the CAF static disk purchase already discussed.

| Stage | FY | Needs (TB) | Servers Bought | FS Size (TB) | Disk (TB) | Total (TB) | Cost (\$M) |
|---|---|---|---|---|---|---|---|
| 1 | 02 | 12 | 4 | 1.9 | 8 | 8 | 0.04 |
| 2 | 02 | 26 | 8 | 2.2 | 18 | 26 | 0.08 |
| 3 | 03 | 54 | 8 | 3.5 | 28 | 54 | 0.08 |
| 4 | 04 | 100 | 8 | 5.5 | 44 | 98 | 0.08 |
| 5 | 05 | 160 | +10 -12 | 8.7 | 60 | 159 | 0.10 |

Table 10: Read disk cache procurement plan. The stage of the CAF, fiscal year, total read cache needs, fileservers bought, file server size, incremental disk added, total disk available, cost of disk added. Not shown in the total for stage 1 is 12 TB of legacy disk cache available to users to meet their needs.

### 5.5 Write Disk Caches

As of May 2002 the DH system consisted of a legacy ADIC tape robot and a fully operating STK robot. As of May 2002 all data is being written directly to the STK robot using Enstore, but the raw data is also being written to ADIC as a safety measure during a transition period we expect to complete in May. Two TB of dual ported disk are simultaneously available to the online data logger and fcdfsgi1 via the CXFS file system and a fiber between B0 and FCC. The online data logger writes to the dual ported disk, and fcdfsgi1 logs the raw data on that disk into the STK robot, and also the ADIC robot during the transition period. We intend to replace the logging function of fcdfsgi1 with a network accessible write disk cache.

| Stage | FY | Needs (TB) | Servers Bought | FS Size (TB) | Disk (TB) | Total (TB) | Cost ($M) |
|---|---|---|---|---|---|---|---|
| 2 | 02 | 9 | 4 | 2.2 | 9 | 9 | 0.04 |
| 3 | 03 | 17 | 3 | 3.5 | 11 | 20 | 0.03 |
| 4 | 04 | 30 | 3 | 5.5 | 16 | 36 | 0.03 |
| 5 | 05 | 46 | +3 -4 | 8.7 | 17 | 53 | 0.03 |

Table 11: Write disk cache procurement plan. The stage of the CAF, fiscal year, total write cache needs, fileservers bought, file server size, incremental disk added, total disk available, cost of disk added.

The write cache procurement plan for write disk cache is shown in table 11. The write cache purchases do not begin until stage 2 of the CAF when they begin to be needed to archive the large quantities of secondary and tertiary data being written on the CAF. Tables 5, 10 and 11 all have the same fileservers and 3 year obsolescence cycle.

## 6 Production Farms

### 6.1 History

The historical growth of the farms is presented in table 12. As of May 2002 the CDF production farms are 169 duals, 2 I/O nodes, server nodes, consoles, and a network switch. The system is running well and can process data effectively. In April 2002 the farms switched their I/O from the ADIC robot to using the Enstore/STK robot. The farms will need regular CPU increments to replace old machines and to increase CPU capacity. In addition, some upgrade to the networking will be necessary if the farms are to handle significantly increased rates.

### 6.2 Requirements

In table 13 we estimate the farms requirements. Starting with the peak logging rate and the running efficiency we estimate the event/year and the total events of raw data. The number of re-processings of all data on the farms starts out large in commissioning periods and reduces with experience to the value of 0.3 re-processings performed in run 1 ( 30% of the data was reprocessed). However, here we have assumed that this reprocessing is done every fiscal year, working on the total backlog of events, so the farms have enough capacity

| FY | Nodes (duals) | Total (duals) | PIII Type (GHz) | PIII CPU (GHz) | Total (GHz) | PIII CPU (SpecInt95) | Total (SpecInt95) |
|---|---|---|---|---|---|---|---|
| 1999 | 50 | 50 | 0.5 | 50 | 50 | 2000 | 2000 |
| 2001 | 23 | 73 | 0.8 | 36 | 86 | 1472 | 3472 |
| 2001 | 64 | 137 | 1.0 | 128 | 214 | 5120 | 8592 |
| 2002 | 32 | 169 | 1.3 | 80 | 294 | 3226 | 11818 |

Table 12: Farms procurement history. The fiscal year, nodes added, total nodes, speed of each node, total speed of nodes added, and total speed of all nodes in GHz, the speed of nodes added and the total speed of all nodes in SpecInt95.

in any one fiscal year to do the reprocessing needed. From this we estimate the total events processed each year on the farms, which starts out close to a billion and increases to around 6 billion by the end of run 2. The measured reconstruction time is now 5 seconds on a 500 MHz PIII. We increase this by a factor of 2 to accommodate potential increases in reconstruction time with multiple interactions, giving a requirement of 5 GHz-sec/event. Assuming an up-time of 75%, each 1GHz processor on the farms can process 5 million events per year. Dividing the processed events per year by 5 million gives the tabulated number of GHz required to satisfy the average rate needs of the farms. To allow the farms to keep up with typical data rates that occur over a few days time, a factor of 2 more CPU is required than the average CPU necessary. The factor of 2 comes from comparing the expected up-time of the system over multiple stores, roughly 60%, with the canonical uptime over a fiscal year, roughly 30%. Multiplying the average rate needs of the farms for data processing by a factor of 2 gives the data CPU needed in table 13.

| Fiscal Year | 02 | 03 | 04 | 05 | 06 | 07 | 08 |
|---|---|---|---|---|---|---|---|
| Peak Logging Rate (Hz) | 80 | 80 | 80 | 240 | 240 | 240 | 240 |
| Running Efficiency | 0.15 | 0.3 | 0.3 | 0.15 | 0.3 | 0.3 | 0.3 |
| Events/Year (Billions) | 0.4 | 0.8 | 0.8 | 1.2 | 2.4 | 2.4 | 2.4 |
| Total Events (Billions) | 0.4 | 1.2 | 2.0 | 3.2 | 5.6 | 8.0 | 10.4 |
| Re-processings | 1 | 0.5 | 0.3 | 0.5 | 0.3 | 0.3 | 0.3 |
| Processed/year (Billions) | 0.8 | 1.4 | 1.4 | 2.8 | 4.1 | 4.8 | 5.5 |
| Data CPU average (GHz) | 160 | 280 | 280 | 560 | 820 | 960 | 1100 |
| Data CPU needed (GHz) | 320 | 560 | 560 | 1100 | 1600 | 1900 | 2200 |
| MC events/year (Billions) | 0.14 | 0.41 | 0.59 | 0.73 | 1.6 | 1.7 | 1.7 |
| MC CPU estimate (GHz) | 50 | 140 | 200 | 240 | 530 | 570 | 570 |
| Total CPU needed (GHz) | 370 | 700 | 760 | 1300 | 2100 | 2500 | 2800 |

Table 13: The farms CPU requirements. As a function of fiscal year, the peak online logging rate of raw data, the total running efficiency, the resulting raw events each year, the total raw events up to that year, the number of reprocessings excluding the first reconstruction pass, the total number of events reconstructed each year including all reprocessings, the required CPU to process the average rate of raw data, the CPU needed for raw data including the ability to keep with rates larger than average, the number of Monte Carlo events produced per year on the farms, the required CPU to produce this Monte Carlo, and the total required CPU on the farms.

For Monte Carlo we are currently planning on generating around 1.25 million events/day for each $\text{fb}^{-1}$ of integrated luminosity that year. This requirement came from an early estimate of how many event per day was reasonable to produce on the farms, and afterwords we added a luminosity dependence, so this estimate has large uncertainties. The time to produce a Monte Carlo event a few months ago was around 15 seconds on a 500 MHz PIII, giving 7.5 GHz-sec/event. We hope that recently measured execution times of 40 seconds per event can be reduced to 15. Again assuming an up-time of 75%, and the old rate of 7.5 GHz-sec/event, each 1GHz processor on the farm can produce 3 million events per year, and the estimated CPU required for MC starts out at 50 GHz in FY02 and increases to around half a THz by the end of run 2. The total requirements in the last row of table 13 is the sum of the data and MC needs.

### 6.3 Purchasing Plan

The plan for future acquisitions is presented in table 14. For the costs we have assumed $2.5K for each PC, $12.5K for annual network upgrades, and $37.5K for miscellaneous annual expenditures (this includes power installation, cabling, disks, consoles, etc.).

| FY | Needs (GHz) | Nodes (duals) | Total (duals) | Type (GHz) | CPU (GHz) | Total (GHz) | CPU (SI95) | Total (SI95) | Cost ($M) |
|---|---|---|---|---|---|---|---|---|---|
| 2002 | 370 | 32 | 201 | 1.8 | 116 | 410 | 4608 | 16426 | 0.10 |
| 2003 | 700 | +68-50 | 219 | 2.5 | 290 | 700 | 11600 | 28026 | 0.22 |
| 2004 | 760 | +32-23 | 228 | 3.5 | 188 | 888 | 7488 | 35514 | 0.13 |
| 2005 | 1300 | +54-64 | 218 | 5.0 | 412 | 1300 | 16480 | 51994 | 0.19 |

Table 14: Farms procurement plan. The fiscal year, total CPU needed, nodes added, total nodes, speed of each node, total speed of nodes added, and total speed of all nodes in GHz, the speed of nodes added and the total speed of all nodes in SpecInt95, and the cost including incidentals (see text).

The farms currently handle 20 MByte/s input and an equivalent output. This was accomplished with the CDF DH system, fcdfsgi1, cdffarm1, gigabit Ethernet, etc. The farms recently changed to an Enstore/STK system which easily achieves over 30 MB/s. To increase to larger rates is straightforward and not expensive. The switch-to-switch connection is GBit Ethernet and it could be trunked or upgraded to 10 GBit. The tape drives can be upgraded to 9940b, giving 30 MByte/s per drive. The nodes which collect the ouput of the farms and send that data to Enstore have will Gbit/s connections – 4 nodes to start. This can be increased if needed. One could easily imagine increasing the overall throughput of the system to 80 MByte/s by making these upgrades.

## 7 Databases

CDF currently utilizes Suns for offline database machines. The production machine, fcdfora1, and the development and integration machine, fcdfora2, are identical Enterprise 4500 servers with two 400 MHz processors each. The machines are dedicated to running Oracle database server version 8.1.7. The database is divided into several applications: Trigger, Hardware, Run, Calibration and Slow Control originate on the online database system and are replicated to the offline database via Oracle basic replication; writing to these applications requires running on privileged nodes located in FCC. The Data File Catalog and SAM applications originate on the offline database and are not currently replicated to any other locations. The online database is controlled behind a firewall, while the offline database has no access restrictions. Access restrictions to the offline database may be increased as replicas are implemented.

## 7.1 Offline Production Database

The offline production machine fcdfora1 is using 300 GB out of 500 GB of available disk, and given the design of the system it cannot be easily expanded to meet future disk and CPU needs. In less than two years the database size will exceed capacity. The machine also has severe disk I/O limitations which have caused serious degradations in performance of the database. The cost of fixing these limitations would be equivalent to replacing the entire system with new and faster Linux boxes with significantly larger disk space.

Therefore, in FY2003, CDF plans to replace fcdfora1 and fcdfora2 with Linux boxes at a minimum of 1.6 GHz and 3 Tb of disk, and at least one gigabit Ethernet interface. We have run the Oracle database server on a smaller scale workstation, including replication from online, without problems in the past. As this machine will have a read/write database, we need to do comprehensive backups, including staging and caching backups to disk and storing archives to tape.

## 7.2 Database Replica Hardware

The online data logger and the production farms need continuous access to the offline production database in order to log the raw data and reconstruct it. However, the offline database is also accessed by the general pool of users running analysis jobs primary in read only mode. Over the last year, several incidents have occurred where the database and system resources could not handle the demand. The situation will worsen as data continues to accumulate, the CAF system is put into place, and SAM begins to be used extensively. For this reason, we plan to implement a series of replica copies of the database on site and at remote institutions. The new replica copies will be read-only instances; the production farms and the online data logger will then have exclusive access to the primary offline production database.

In the early summer of 2002, we plan to implement a complete replica of the offline production database on the existing 700MHz quad P3 machine fcdflnx1 with 4 GB of memory, in order to get the system up and working quickly. This replica would be used by the CAF and all general users requiring read only access. This scheme achieves the immediate goal of isolating the production farms and online data logger from disturbance by general users. This solution requires incremental costs in FY2002 for the addition of about 1Tb of disk space to fcdflnx1 with concomitant additional SCSI interfaces, bringing the total to approximately 2 Tb. As a pure read-only replica, backup costs will be minimal. The fcdflnx1 machine should continue to reside in FCC and have a gigabit Ethernet connection. The replication would proceed through Oracle basic replication from the originating database, either online or offline depending on application. Replication development will take place on the existing cdfpcb machine for the near future.

In FY2003, we plan to purchase a new permanent replica host, cdfrep00, a 1.6 GHz quad

P4 machine with a minimum of 16GB of memory, in order to address the expanding pool of users in the CAF and elsewhere. Again, this machine will have a minimum of 2Tb of disk and at least one gigabit Ethernet interface. At this time, fcdflnx1 may then become a replication development machine or continue to function as a secondary replica. Remote sites may then choose to implement their own local replicas that contact cdfrep00 only, with no direct contact to the primary offline or online databases. Expressions of interest have been made by TTU and Glasgow, with more to follow. Note that remote institutions are expected to provide their own hardware and at least one responsible contact person to maintain the local replica, although assistance will be available from Fermilab.

In FY2004, judging by the demand on the replica server, we will most likely need to implement two additional replicas on site. The first replica cdfrep00 would continue to serve the CAF; the new replicas cdfrep01 and cdfrep02 would serve the CDF trailer community and the remote institutions, respectively. The new machines would be the relatively equivalent to cdfrep00 at the time, presumably significantly faster and with more memory. We may rearrange the old and new machines in order to optimize load leveling. In FY2005 we will replace and upgrade the older production database systems as necessary. The 4-ways may become n-ways as necessary at any step in procurement plan, and the disk amount grows with time. All details in 2004 and 2005 will depend on actual running experience with the earlier Linux boxes.

## 7.3 Database Replication Software

The current plan for replicas on-site and remote, is to use the included Oracle replication wrapped with some convenience scripts and utilities. For the initial replica, cdfrep00, we will be using Oracle version 8.1.7 basic replication. Basic replication cannot be extended to remote nor trailer sites since it does not work with the firewall restrictions around the online database. Basic replication also does not scale easily with many replicas – the administrative overhead would be prohibitive.

To extend the replication capabilities to remote and non-FCC sites, we will need the Oracle Streams functionality in their version 9.2 release. This release is scheduled for approximately May or June 2002, with Sun and Linux versions being released first. Oracle Streams allows the data propagation to proceed in a sequential or leap-frog manner, thus avoiding the firewall issue. Streams also allows automatic propagation of DDL changes to replica sites, easing administrative issues.

During June and July, the CDF database group will work on testing and developing features of Oracle v9.2, including a toy database replication setup. Upgrading a major Oracle version represents a significant risk in terms of potential down-time, so production installation will wait until there is a significant accelerator down-time after the version 9.2 evaluation period.

An ongoing project has been the conversion of the Oracle database schema and data into a MySQL database. The primary reason for porting to a freeware database is to save the Oracle licensing fees; this may not be an issue (see below). The MySQL solution has not been ruled out, but a number of serious problems indicate that it will not be available in the near term. On the positive side, Oracle Streams has expanded support for third party generic replication, which may be useful in solving some of the MySQL problems.

| FY | DB CPU (n-ways) | DB Disk (TB) | Cost ($M) |
|---|---|---|---|
| 2002 | 0.5* | 2 | 0.02 |
| 2003 | 3 | 6 | 0.15 |
| 2004 | 2 | 6 | 0.10 |
| 2005 | 2 | 9 | 0.10 |

Table 15: 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. (* peripherals only)

### 7.4 Oracle Licenses

As of May 2002, the Oracle license has not been renegotiated due to difficulties on the Oracle and DoE side. The existing carry-over license extends until the new contract takes effect. This license allows a maximum of 135 concurrent user sessions across all the Fermilab Oracle database instances, and extends to remote institutions affiliated with Fermilab/CDF. Current usage is well within this limit. The future total number of concurrent users is difficult to predict, and could very well exceed the limit. In the offline code, steps have been taken to make the connections more efficient, and more can be done. However, additional concurrent user license units may in the end have to be purchased. As of a quote from March 2002, these units cost $1625 per user session per year (price includes support and discount). Final numbers will become available when the DoE completes negotiations with Oracle on May 15, 2002.

Oracle's new pricing structure may allow or even require that a server based pricing scheme be used. Alternatives to increasing the license limits include implementing a middle tier server between the users and the database, or continuing with implementation of a freeware solution for some or all replicas. Limited manpower puts strict limits on software solutions, and may end up being more expensive than buying additional licenses.

## 8 Networking

### 8.1 LAN

The scaling of a user analysis farm, network attached disk, and network attached tape drives to meet CDFs storage and CPU requirements in run 2 all rely heavily on the local area network. We are planning for significant growth in the CDF LAN in the coming years. Figure 2 shows a simplified and incomplete diagram of the current and anticipated CDF LAN with major components.

The heart of the current CDF network is the offline switch, a CISCO 6509 in FCC2. 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. The stage 1 CAF fileservers use 15 GBit connections and the CAF stage 1 worker nodes use 69 FE connections to the offline switch. The farms will write to Enstore via a single GB/s connection between the offline switch and the farms switch, another CISCO 6509. The farms worker nodes are each connected to the farms switch via

FE, except for the worker nodes that will serve as output nodes, currently 4, which have Gbit connections. The farms stager and raw data logging node fcdfsgi1 is connected to the farms switch via Gbit, and directly mounts the dual ported disk with b0dau32 via Gbit connections. The offline switch is currently connected to the FNAL switch, and then to the outside world, via a single Gbit connection. The offline switch is also currently connected to the trailers switch, a CISCO 6509, via a Gbit connection. The desktops in the trailers are connected to the trailers switch using FE connections, although sometimes indirectly via satellite switches.

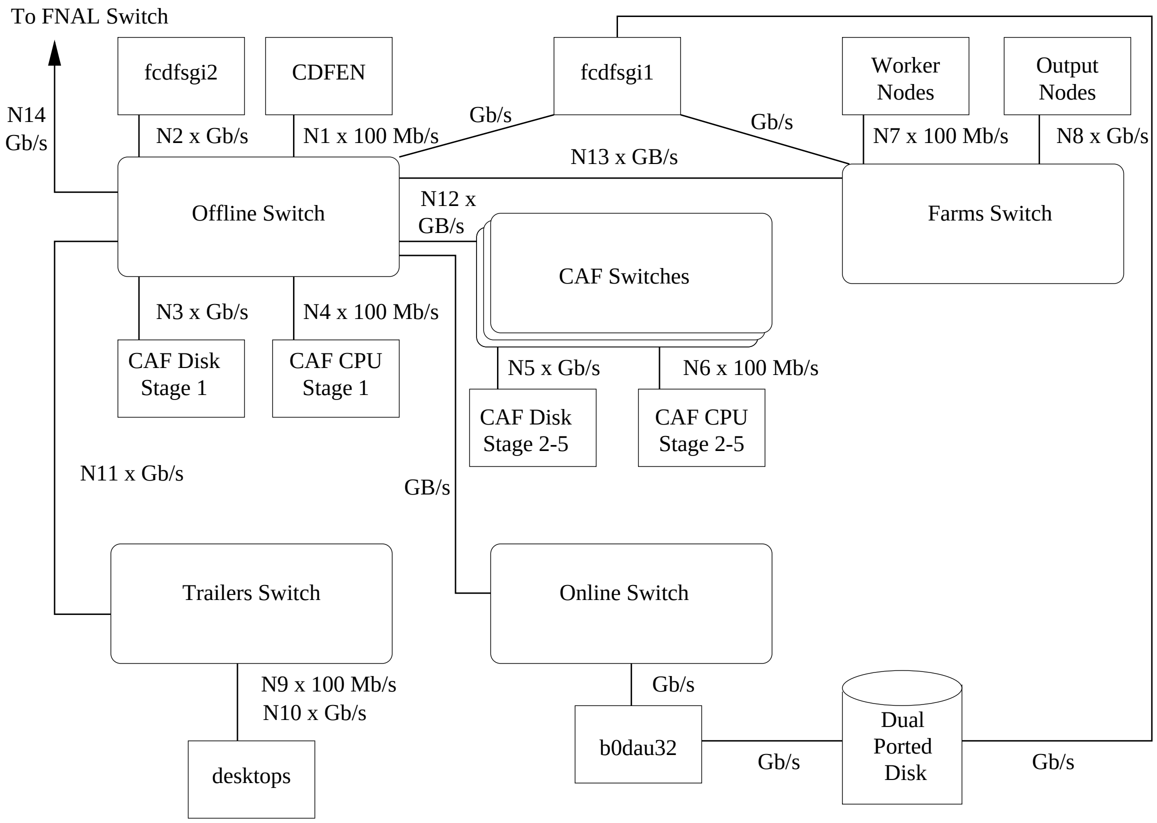


Figure 2: A simplified view of the CDF Local Area Network. Major computing elements in boxes, switches in rounded boxes, and Ethernet, GBit, or multi-GBit connections are shown by lines. The Nx, where x = 1-14, are the number of connections per major computing element.

The needs of the CAF and disk cache are driving the LAN plans shown in table 16. For stage 1 in FY2002 we purchased 2 FE modules and 1 GigE module and attached them to the offline switch to support N4 = 69 and N3 = 16 in Fig. 2. We also will upgrade the switch from 16 Gbit/s capacity to 256 GBit/s capacity. For stage 2 in FY2002 we plan to purchase 4 FE modules and 3 GigE modules and attach them to a new CISCO switch in FCC1 to support N6 = 160 and N5 = 44. Similarly for each FY from 2003 to 2005 we will purchase another 4 FE modules to support 192 additional nodes per year, and an additional Ethernet switch to support growth, each switch with at least 256 GBit/s capacity and at least 7 ports available to us. Each year we are also buying enough GigE modules, currently with 16 ports each, to support the fileservers with 1 or 2 GigE connections each. In FY2004 the increased disk size requires 2 GigE connections per fileserver. Each year between FY2003 and FY2005 we are adding progressively more networking capability, with better switches, and doing it at roughly the same $100K cost incorporating a Moore's law growth.

| Stage | FY | FE Ports Bought | FS bought | GigE Ports/FS | GigE Total | Switch Bought | Cost ($M) | Misc ($M) | Total ($M) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 02 | 96 | 16 | 1 | 16 | - | 0.04 | 0.07 | 0.11 |
| 2 | 02 | 192 | 44 | 1 | 60 | 1 | 0.14 | 0.00 | 0.14 |
| 3 | 03 | 192 | 43 | 1 | 103 | 1 | 0.10 | 0.15 | 0.25 |
| 4 | 04 | 192 | 43 | 2 | 189 | 1 | 0.11 | 0.14 | 0.25 |
| 5 | 05 | 240 | +45 -60 | 2 | 219 | 1 | 0.07 | 0.18 | 0.25 |

Table 16: LAN procurement plan. The stage of the CAF, fiscal year, number of Fast Ethernet ports bought, number of fileservers bought, the gigabit Ethernet ports required per fileserver, the total number of gigabit Ethernet ports required, the number of switches bought, the cost of of all this CAF related networking, the cost of miscellaneous networking primarily for the CDF trailers and total cost.

Networking costs for the farms are discussed in section 6 and included in the farms total costs.

In addition to the plan shown explicitly in table 16 for CAF and disk cache driven networking, there are miscellaneous sources of networking costs to support the N values and switches in Fig. 2. We have assumed that this will require roughly an additional $140K - $180K per year as shown in table 16 and discussed below. The largest costs will be for upgrades to the networking for the CDF trailers. The increasing use of network attached tape and disk storage in FCC, and the increasing number of users analyzing data in the CDF trailers, should increase the networking requirements both in the trailers and between the trailers and FCC. 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 some offices. Since there is essentially no more room on the existing 6509 switch for additional ports, more switches will need to be added to support further expansion. Each switch in the trailers needs to be capable of supporting 1 or more gigabit connections in FY 2003, both to support the possibility of gigabit fileservers in the trailers and to allow multiple gigabit links between switches as necessary. The current plan is to purchase either a 6509 or 6513 in FY 2003, fabric enabled to support 256 Gb/s on its backplane, which will be needed to lay the foundation for N10 gigabit connections to multiple fileservers. The older 6509 with a 16 GB/s backplane will continue to support the N9 fast Ethernet connections to the desktop.

In FY2005 we anticipate upgrading the older 6509 to a new switch with an even higher backplane capacity, to support the future possibility of ten Gbit connections to fileservers, and the existing 6509 with the 256 GB/s backplane should support any Gbit connections to the desktop that could be common at that time. Each switch will likely cost around $40K, and there will be comparable costs for more Gbit ports as well.

When the cost of gigabit Ethernet connections (currently around $2K each including both the port and the media converter) drops to around the current level of FE connections ($500 each), we will need to begin supporting gigabit Ethernet in the trailers, and N10 which is now approximately zero will start to increase. Assuming a drop by a factor of 2 each year in cost, the GigE upgrade in the trailers would need to begin in FY 2004. We could imagine buying GigE connections for 100 desktops or fileservers in FY 2004 at the cost of $50K, and assuming another factor of 2 drop in cost the following year, buying GigE connections for another 200 desktops or fileservers in FY2005 at the cost of $50K.

There is currently N11=1 Gbit link between the trailers and FCC which, will need to be increased. The cost of increasing the link to a total of 2 Gb/s or 4 Gb/s is minimal, only the cost of the additional Gbit ports in both the offline 6509 and the trailers 6509. The fibers between FCC and the trailers are in place. Between now and FY 2004 we anticipate increasing the the number of GBit links up to a maximum N11 = 4. At the beginning of FY 2004, eighteen months from now, we anticipate that the cost for ten Gbit Ethernet will have dropped from roughly $50 K currently, for a single module with a single port, to around $25 K for a module with two or more ports. At that point it will be cost effective to buy a ten Gbit module for the offline switch and another for the trailers switch, and be able to support multiple ten GBit connections between FCC and the trailers. We also anticipate the need for ten Gbit connections between the trailers switches and the CAF switches in FY2005.

The number of FE connections between the offline switch and the CDFEN robot, N1, is currently 10: one FE for each mover node connected to a drive in the CDFEN robot. In section 5 we discuss the plan to add another robot in FY2002, with 10 additional 9940A drives and also 10 FE connections. These FE connections will either go to the new switch for stage 2 of the CAF, also purchased in FY2002, or will go to the original offline switch. If they go on the offline switch then there must be sufficient bandwidth on the up-link between CAF and Offline switch to handle the user archive read rate on the CAF, estimated in table 7. The purchase in FY2003 and FY2004 of T9940B drives suggests gigabit connections to take advantage of 30 MB/s drive rates, and these would likely go on the CAF switches purchased in those fiscal years respectively, requiring a total of 20 gigabit ports by 2004. In 2005 we would have to consider 2 gigabit connections per drive if we wanted to take advantage of the the full 60 MB/s hypothesized for T9940C.

The number of Gbit links between the offline switch and the CAF switches, N12, required for our bandwidth needs is being studied. There is currently N13 = 1 GBit link between the offline and farms switch which wont need to be increased until perhaps we begin the increased data taking rates possible in FY2005. Finally, the number of Gbit links between the offline and FNAL switch, N14, will need within the next few years to support the additional off-site bandwidth discussed in the next section. The costs mentioned are roughly the same as the money allocated for miscellaneous networking expenditures in table 16.

## 8.2 WAN

The CDF institutions have been surveyed on networking and data distribution plans [9]. Most US university groups are connected to Internet 2 and most groups have been thinking of using network transfers to get the data and avoid tapes if possible. Roughly 10% of the institutions think that tape may be used as well, and only one group plans on having a system integrating tapes and disks like Fermilab. The average disk space at remote institutions in January 2002 was 400 GB, and the anticipated average disk space in July 2002 was 1.5 TB. These results indicate appreciable storage capabilities at remote institutions and heavy usage by CDF of the WAN.

One approach to estimating the WAN requirements is to assume the storage capacity of the remote institutions will drive the amount of data they copy from Fermilab. The relevant storage medium at CDF institutions is disk, since tape is primarily used for backup purposes. The average group is expected to have 1.5 TB of disk space now, and it is likely that amount of disk will double every year with the decreasing price of cheap IDE disk. A large fraction of this disk will likely be devoted to reconstructed secondary datasets copied over the network from Fermilab. Assume each of 50 institutions copies a dataset that occupies half their disk from Fermilab every 3 months, and recopies more complete, selected, compressed and refined data every 3 months for the course of the run we get the WAN requirements profile in table 17. Coincidentally, the rate required in Mbit/s is identical to the remote disk storage in TB because the number of seconds in 3 months is the same as the number of Mbits in a TB: 8 Million.

| FY | Disk (TB) | Rate (Mbit/s) |
|---|---|---|
| 2002 | 50 | 50 |
| 2003 | 100 | 100 |
| 2004 | 200 | 200 |
| 2005 | 400 | 400 |

Table 17: WAN requirements by storage. As a function of FY, the anticipated remote disk storage that will be re-populated every 3 months from Fermilab via the WAN, and the required rate at the end of the FY to re-populate it.

Another approach to estimating the WAN requirements is to assume that the data volume at Fermilab will drive the rate without any limit from the resources at the remote institutions. Taking this approach we note that the rate of production of reconstructed data is expected to be 2.5 MBytes/s throughout the run, and each of 50 institutions could want a direct pipeline to this data to analyze it. Then the rate needed is 1 Gbit/s.

The most bandwidth intensive estimate of WAN requirements is to assume that users will access data from Fermilab transparently into local data caches, pulling it over the network as needed every time the data is absent from their local cache. Assuming this, INFN alone estimates that 10 MBytes/sec/user x 20 users x 20% cache misses = 40 MBytes/sec. Scaling this to roughly 100 active remote users for the entire collaboration would give 200 MBytes/sec, or 1.6 Gbit/s.

Since typically 30% of the connection rating is achieved, the current OC3 connection with a rating of 156 Mbit/s could become strained by the end of this fiscal year. Fortunately

Fermilab is in the process of installing an OC12 connection with a rating of 622 Mb/s that should be in place by this Summer. We congratulate the lab on this timely upgrade. While necessary for the anticipated load in FY 2003, the OC12 connection may be inadequate for CDF needs alone by FY 2004, so we strongly recommend that the lab pursue it intended upgrade to an OC48 rated at 2.4 Gb/s by FY2004. Roughly 18 months after that, in FY2006, we foresee the need to upgrade to an OC192, rated at 10 Gb/s.

# 9 Summary and Conclusions

| FY | Batch CPU ($M) | Inter. CPU ($M) | Farm CPU ($M) | DB ($M) | Tape Robot ($M) | CAF Disk ($M) | Cache Disk ($M) | Net-work ($M) | Legacy Sys. ($M) | Total ($M) |
|---|---|---|---|---|---|---|---|---|---|---|
| 2002 | 0.59 | 0.07 | 0.22 | 0.02 | 0.77 | 0.47 | 0.16 | 0.25 | 0.69 | 3.24 |
| spent | (0.19) | (0.07) | (0.11) | (0.00) | (0.25) | (0.12) | (0.04) | (0.12) | (0.69) | (1.59) |
| 2003 | 0.48 | 0.15 | 0.22 | 0.15 | 0.35 | 0.35 | 0.11 | 0.25 | - | 2.06 |
| 2004 | 0.48 | 0.2 | 0.13 | 0.10 | 0.35 | 0.35 | 0.11 | 0.25 | - | 1.97 |
| 2005 | 0.60 | 0.2 | 0.19 | 0.10 | 0.35 | 0.35 | 0.13 | 0.25 | - | 2.17 |

Table 18: CDF computing equipment purchasing plan. The fiscal year, batch CPU for the CAF, interactive CPU and its local disk, production farm CPU, databases, tape robot & tape drives, network attached disk for the CAF, read and write cache disk, networking, legacy CPU and disk systems, and total procurements.

The equipment costs for computing in run 2 is summarized in table 18, including costs to participating CDF institutions for hardware they contributed. Our FY2002 equipment budget from Fermilab was $2M, but as mentioned in section 4 we also received $110K in CAF contributions and $984K in budget carry-overs, hardware returns, and legacy system contributions, bringing our total budget to $3.09M. Further, we are anticipating between $0.15M and $0.36M more from non-Fermilab contributions, giving us a total budget of between $3.24M and $3.45M which should cover our spending plans for FY2002, listed in table 18. Even with these contributions it will be tight in FY2002, since we have transferred around $100K from equipment to operating to compensate for over-spending our tight operating budget, and we will likely need to transfer at least $50K more. Not shown in table 18 is roughly $500K per year for operating costs including tapes, discussed in section 5.3.

The procurement costs for FY 2003 to FY 2005 are roughly the $2M per year budget guidance that the lab has given. Although an effort has been made to keep the budget within this guidance, it is also clear that we are buying a dramatic improvement in the computing, storage and data transfer capability of CDF, which should be sufficient for our needs throughout the first half of run 2.

In conclusion, increases in luminosity drive the computing requirements for run 2. The CPU and disk needs scale with integrated luminosity. The tape archive and farms needs scale mainly with DAQ logging capability. Computing procurements are planned to meet the needs. We plan to procure hundreds of cheap PCs to access large static datasets stored on inexpensive network attached IDE disk. Fast network attached tape drives will store and retreive data from tapes that will each hold hundreds of GBs. The budget is not dominated by any single system, and the procurement costs are roughly within the budget guidance the lab has given us.

# References

[1] S. D. Holmes, "Run II Performance and Plans", presentation to HEPAP on January 29, 2002.

[2] CAF Review Committee, B. Ashmanskas, et al., "Physics Analysis Computing needs assessment CDF CAF Review Fall 2001", CDF Note 5787, December 14, 2001.

[3] CAF Review Committee, B. Ashmanskas, et al., "Final Report CDF CAF Review Fall 2001", CDF Note 5802, December 7, 2001.

[4] CDF Joint Physics Group Meeting, February 15, 2002.

[5] T. Kim, M. Neubauer, F. Würthwein et al., "CAF design document - Stage 1", CDF 5961, May 14, 2002.

[6] D. Baden et al., Joint D0/CDF/CD Run II Data Management Needs Assessment (DAMNAG), CDF Note 4100, March 20, 1997.

[7] R. M. Harris et al., "A Tape Handling System for CDF in Run 2", CDF Note 5822, January 11, 2002.

[8] Technical Review of CDF Data Handling Software Infrastructure, R. Harris, R. Kennedy and J. Tseng, CDF 5917, May 2002, in preparation.

[9] K. Sliwa, "CDF Networking Needs", presentation to CDF Institutional Computing Representatives Board, January 25, 2002, and also CDF note 5836 in preparation.