/PUBLIC/DOC/COMP_UPG/CDF-6639
Version 1.3

Implications of increased data logging rate on the CDF Run 2 computing plan and budget

F.D. Snider
Robert M. Harris
Fermilab

August 23, 2003

Abstract

We discuss the computing resources required by CDF in light of recent proposals to increase the event logging rate via reductions in raw data event sizes and increases in the Consumer Server Logger bandwidth. Details of the models used to make resource projections are presented, along with comparisons to data from the past year and a half of data taking. We then provide cost estimates for the main components of the computing system as a function of time under various sets of model assumptions.

1 Introduction

This note discusses the CDF Run 2 computing plan and budget within the context of proposed changes to the CDF data acquisition system and operation policies aimed at upgrading the event logging rate of the experiment. Since the demand for computing resources at CDF depends primarily upon the volume of data that needs to be analyzed and the absolute rate at which it is accumulated, a permanent increase in the event logging rate or data volume requires careful evaluation of the implications on the overall computing system. The basis of the discussion will be the CDF computing plan as documented in CDF-5914, from May, 2002 [1]. Most of the models presented in this note are derived from those in CDF-5914, with some changes to better model the current use of the system. More importantly, the past year and a half have provided invaluable operating experience with the system, and a wealth of data against which to test the model predictions. By comparing the two, we can determine the extent to which the model captures the behavior of those features needed for estimating the resource requirements, and hence the costs of the upgraded system.

We will address two different scenarios: the "baseline" plan, and the "upgrade" plan. The former is essentially unchanged from the plan documented in CDF-5914.

The upgrade includes two changes designed to increase the event logging rate from the detector: implementing raw data compression in Level-3, which should result in an increase of about 50% in the event logging rate, and an upgrade to the Consumer Server Logger (CSL) bandwidth from 20 MB/sec to 40 MB/sec. Both the baseline and upgrade paths anticipate an increase to 60 MB/sec in FY06 following previously scheduled DAQ upgrades. These changes and the fiscal years in which we assume they are implemented are shown in Table 1.

Table 1: The assumed changes to the CSL bandwidth and raw data event size, and the resulting peak event logging rate as a function of fiscal year in the baseline and upgrade scenarios. The observed peak event rates from the experiment are typically about 90% of the theoretical maximum value.

	Fiscal year	02	03	04	05	06	07	08	09
Baseline	CSL rate (MB/sec)	20	20	20	20	60	60	60	60
	Raw data compression	No	No	No	No	No	No	No	No
	Implied peak event rate (Hz)	80	80	80	80	120	120	120	120
Upgrade	CSL rate (MB/sec)	20	20	20	40	60	60	60	60
	Raw data compression	No	No	Yes	Yes	Yes	Yes	Yes	Yes
	Implied peak event rate (Hz)	80	80	120	240	360	360	360	360

Given the logging rates, we will then employ two basic approaches to arrive at estimates for the various computing resources required by the experiment. The first method is applied to those systems that are required to keep pace with the average or integrated data logging rate during periods of machine operation. These calculations are typically highly predictive of the needs with relatively few uncertainties.

The second approach is used for systems that depend at least in part upon the usage pattern of physicists conducting analyses. In this case, we adopt a simplified model of the typical user analysis. Such an approach necessarily introduces significant uncertainties in the results since the behavior of a large pool of unrestricted, adaptive users is difficult to predict or model. While the details and motivations for the model are presented in Sect. 4, we can summarize the general idea as follows: we first assume that a data rate corresponding 400 nb of Level-3 cross section goes into a dataset "A", while the balance of the logging bandwidth goes into a dataset "B". We then calculate the resources required to allow 200 simultaneous users to analyze 5 nb of dataset-A in a single day, plus an additional 15 simultaneous users to process all of dataset-B over the course of 25 days. The motivation for these numbers is discussed in Sect. 4.

We restrict our analysis to the systems that dominate the size of the computing budget: the production farm, Central Analysis Facility (CAF) both in terms of CPU and network attached disk, and the tapes and tape drives used in the data archive. After calculating the resource requirements for these systems, including contingencies and losses from inefficiencies, we apply cost models to determine the needed procurement profile for each. We then estimate the overall budget by adding fixed annual expenditures for the major elements of the computing infrastructure that are not covered by the models. No attempt is made to determine how these smaller infrastructure expenditures scale with the proposed changes.

Ultimately, we will propose a new baseline computing plan based upon the upgrade scenario. The content of this document, however, should not be construed as defining the CDF computing plan or budget. It is instead intended to describe the details of the models used to predict resource needs and costs, the results of which will be used as input to the actual plan and budget. For a full description of the new baseline plan, the reader should consult Ref. [2].

The balance of the note describes in detail the models used to evaluate the major resources required to analyze data from the experiment in a timely fashion, and the cost of procuring those resources. Where possible, predictions from the models are compared to data from the 2003 winter conference season. We begin in the next section by describing the common assumptions and parameters required throughout the various model calculations. We then describe the requirements for the production farms. This is followed by a detailed analysis of the Central Analysis Facility (CAF), which forms the backbone of the analysis system for many if not most physicists on the experiment. Next, we discuss the key elements of the Data Handling (DH) system: the tape robot, media and tape drives. Finally, we summarize the various costs and isolate those which depend only upon the logging rate, and are therefore laregly model independent, from those that require strong model assumptions, and therefore introduce sometimes significant uncertainties in the results.

2 Basic assumptions and parameters

Some of the basic assumptions common to many of the model calculations are shown in Table 2. Included are the integrated delivered and acquired luminosities, average and peak accelerator operating efficiencies, CDF data logging efficiency, typical event sizes, average logging rates and the number of events logged assuming the average and peak logging rates.

The machine efficiency represents all effects that contribute to down-time in data taking, such as shutdowns, beam studies, etc., over the course of a year. Historically, a value near 0.3 is typical for a normal year. This value falls during periods of accelerator commissioning, motivating values of 0.15 during FY02 and FY06. During periods of active accelerator operations, the efficiency is approximately 60%. Support for this value is presented in Sect. 3

The luminosity profile corresponds to the "design" profile [3] and represent the more optimistic of the two profiles quoted by the Beams Division. The accelerator and logging efficiencies are typical observed values. Dips in the machine and logging efficiencies correspond to periods of commissioning immediately following long shutdowns, time often associated with reduced operating efficiency.

The (uncompressed) baseline event size is the measured averaged during a set of runs in 2003 at luminosities below 4×10^{31} cm^{-2} sec^{-1}. We have ignored a known dependence on the instantaneous luminosity that is expected to produce an approximately linear 40% increase in the uncompressed data size between 10^{31} cm^{-2} sec^{-1} and 10^{31} cm^{-2} sec^{-1}. The compressed event size is calculated based upon the raw data compression factor (0.675) observed in run 167024 during which the trigger table with data compression and bank dropping was tested. Again we have ignored possible changes in the compression ratio with instantaneous luminosity.

Table 2: Basic assumptions and operating parameters vs. time. The event sizes refer to raw data out of Level-3.

Fiscal year	02	03	04	05	06	07	08	09
Integ. lum. (pb^{-1})	0.08	0.3	0.68	1.35	2.24	3.77	6.14	8.56
Integ. lum. acquired (pb^{-1})	0.06	0.25	0.6	1.2	2	3.38	5.51	7.69
Avg. acc. operating effic.	0.15	0.3	0.3	0.3	0.15	0.3	0.3	0.3
Peak acc. operating effic.	0.3	0.6	0.6	0.6	0.3	0.6	0.6	0.6
Average logging effic.	0.7	0.9	0.9	0.9	0.9	0.9	0.9	0.9
Baseline scenario								
Event size (kB)	220	220	220	220	220	220	220	220
Average event rate (Hz)	25	50	50	50	150	150	150	150
Average events/year (G)	0.12	0.47	0.47	0.47	0.71	1.42	1.42	1.42
Average total events (G)	0.12	0.59	1.06	1.53	2.24	3.66	5.07	6.49
Peak total events (G)	0.38	1.13	1.89	2.64	3.77	6.04	8.3	10.57
Upgrade scenario								
Upgrade event size (kB)	220	220	135	135	135	135	135	135
Upgrade avg. event rate (Hz)	25	50	85	170	250	250	250	250
Average events/year (G)	0.12	0.47	0.8	1.6	1.18	2.36	2.36	2.36
Average total events (G)	0.12	0.59	1.39	3	4.17	6.53	8.89	11.25
Peak total events (G)	0.38	1.13	2.26	4.53	6.23	9.62	13.02	16.42

The average event logging rate is typically observed to be between 60% and 80% of the peak logging rate for a run. We use values that are 70% of the maximum effective peak logging rates from Table 1. The number of events per year is calculated from the logging rate and the average operating efficieny.

In addition to these time dependent parameters, we several assumptions that do not vary with time. Table 3 shows the values for some of these, including the size of production output events after compression (again ignoring luminosity dependence), the production event processing time and the "typical" user analysis processing time. The production output event size is the measured value over a random set of runs. The origin of the event processing times is discussed in more detail in later sections. Here we note only that the user analysis time is one of the least well-known quantities in this analysis. A large error on this value, however, should appear as a disagreement in the calculted and observed network throughput on the CAF, a point to which we will return in Sect. 4. Additional assumptions specific to the various systems are discussed in the relevant sections.

Table 3: Some important time independent parameters used throughout the various model calculations. The event processing times are quoted as the time required on a 1 GHz P-III equivalent processor.

Compressed production event size (kB)	180
Production event processing time (GHz-sec)	3.3
"Typical" user event processing time (GHz-sec)	0.2

3 Production farms

The size of the production farm is driven primarily by two requirements. First, it must process all data from the detector within a few days of being logged. Second, it must re-process some fraction of previously logged data as updates to the production executable become available or for other special projects. The production farm model, therefore, has two components, one that scales with peak event logging rate, and another that scales with the total number of events logged. Additional contributions are needed to account for operational delays, testing and R&D.

The model used here is essentially the same as that used in CDF-5914, with some changes based upon running experience over the past year. To estimate the required capacity, we first calculate the total number of events processed per fiscal year, which is the sum of the data logged per fiscal year and the number of previously logged events that are reprocessed. The former value is given by multiplying the the peak event logging rate by the average running efficiency, typically 30% (see Table 2). Historically, only 30% of the data has required re-processing. We take this value for normal detector operations, with larger values near commissioning periods. There is some uncertainty in the re-processing fraction. Through this fiscal year, the re-processing fraction has been near unity. We have assumed here that the need for re-processing will decline in the future. The average for the later portions of Run 1 was about 0.3 [4].

To obtain the required farm capacity, we multiply the total number of events processed during a year by the CPU required per event, 3.3 GHz-sec (see Fig. 1) times a contingency factor of 1.5. The latter allows for increases in processing time with later versions of the executable or with increased instantaneous luminosity. We then divide by 0.75 to take into account the average farm utilization efficiency (Fig. 3). Contributions to this factor include operational delays from calibrations, executable problems, etc. Finally, we multiply by the ratio of the peak accelerator operating efficiency, typically about 60% (Fig. 2), to the average. This factor ensures that the average processing rate is sufficient to keep up with data taking during periods of stable machine operation.

The estimated production farm capacity as a function of fiscal year is compiled in Tables 4 and 5 for the baseline and upgrade scenarios.

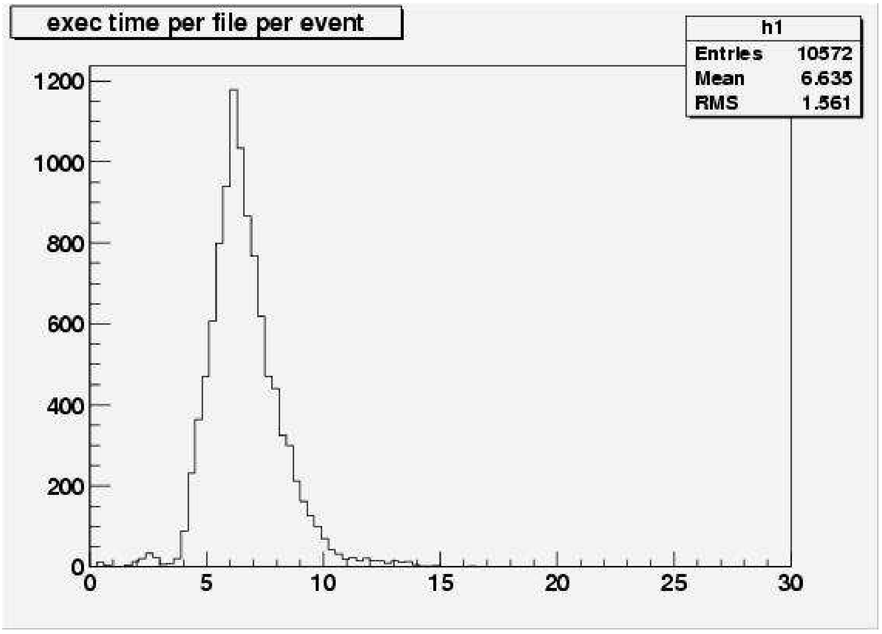

Figure 1: Event execution time for 4.8.4g1 ProductionExe on a 500 MHz farm node.

By examining the performance of the existing production farm over the past year, we can assess whether the farm, and therefore the model, provides the needed capacity to keep up with data taking. Figure 4 shows the number of logged and processed events as a function of time from Feb. 2002 through May 2003. From the relative slope of the two curves, we see that the farm processes events significantly faster than the data logging rate, a fact which allows the farm to keep up with data taking. Periods during which no processing occurred were typically due to problems with the production executable or calibrations. We conclude that the model adequately estimates the required capacity for the farm.

A comment on the re-processing model is in order at this point. Note that by combining the number of logged events with the number to be re-processed as the model input, we implicitly assume that re-processing occurs on the same time scale as accelerator operations, i.e., about six months. If re-processing were limited to special requests on small to moderate sized datasets, then this re-processing model would likely be acceptable. If, on the other hand, wholesale re-processing of all logged data or large sections thereof is required, then it is doubtful that a processing time of six months is acceptable.

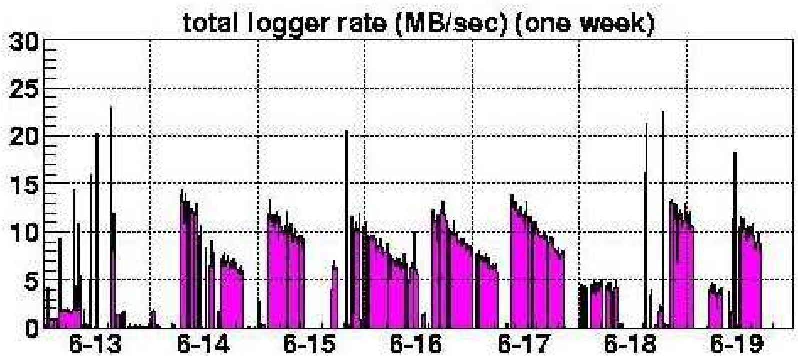

Figure 2: The CSL data logging rate in MB/sec for a seven day period of sustained accelerator operation. We can see from this plot that an average peak logging efficiency of 60% is a reasonable value to assume. We also note that the peak logging rate of 20 MB/sec was never attained, and that the average was close to about 12 MB/sec. The luminosity during this period was below 4×10^{31} cm^{-2} sec^{-1}.

One possible alternative would be to re-process during accelerator shutdowns when there is no load from data logging. The need for re-processing, however, would only coincidentally align with such times. Consequently, we believe this portion of the model will need to change as long as a large fraction of the data requires re-processing, as appears to be the case given recent experience.

To accommodate such activity on the farms with the current software would require a significant increase in the size of the farm, much of which would remain idle when not engaged in re-processing. A solution better integrated into the balance of computing in the experiment is likely required to adequately address this issue, a discussion which is well beyond the scope of this note.

Table 4: Production farm requirements under baseline scenario.

Fiscal year	03	04	05	06	07	08	09
Peak event rate (Hz)	80	80	80	240	240	240	240
Running efficiency	0.3	0.3	0.3	0.15	0.3	0.3	0.3
Events/year (10^9)	0.75	0.75	0.75	1.13	2.26	2.26	2.26
Total events (10^9)	1.1	1.9	2.6	3.8	6.0	8.3	10.6
Re-processings	1.0	0.5	0.3	0.2	0.3	0.3	0.3
Processed/year (10^9)	1.89	1.7	1.55	1.89	4.08	4.76	5.43
Average CPU (GHz)	395	355	324	395	853	995	1137
Required CPU (GHz)	790	711	648	790	1706	1990	2275

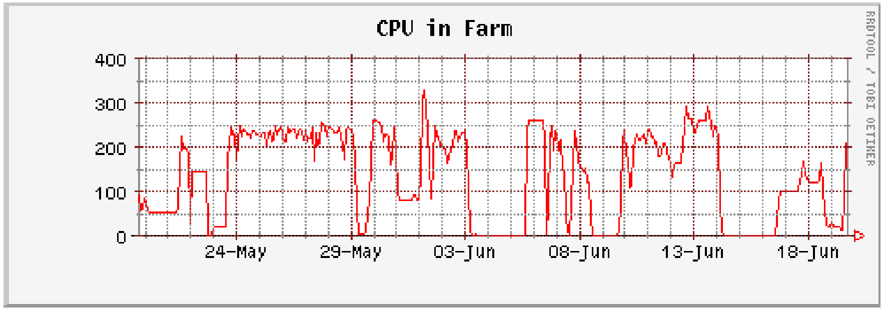

Figure 3: Production farm CPU utilization (in number of active CPU's) for a one month period of active data taking. The plateau in the first third of the plot is near approximately the maximum number of CPU's available for production processing. The average efficiency of 75% is obtained by analyzing such data over long periods of running.

Table 5: Production farm requirements for upgraded event logging rate and CSL bandwidth.

Fiscal year	03	04	05	06	07	08	09
Peak event rate (Hz)	80	120	240	360	360	360	360
Running efficiency	0.3	0.3	0.3	0.15	0.3	0.3	0.3
Events/year (10^9)	0.75	1.13	2.26	1.7	3.4	3.4	3.4
Total events (10^9)	1.1	2.3	4.5	6.2	9.6	13.0	16.4
Re-processings	1.0	0.5	0.3	0.2	0.3	0.3	0.3
Processed/year (10^9)	1.89	2.26	3.62	2.94	6.28	7.3	8.32
Average CPU (GHz)	395	474	758	616	1315	1528	1742
Required CPU (GHz)	790	948	1516	1232	2630	3057	3483

3.1 Production farm cost estimate

To determine the cost of the farm, we assume that the farm consists of dual CPU Linux boxes at a cost of $2.2k each in constant dollars. The CPU power is assumed to increase by a factor of two every 18 months, starting at 2.2 GHz for FY2003. Machines are retired after about three years of service in order to limit the total number of machines in the farm. The results are shown in Tables 6 and 7 as a function of fiscal year and event logging rate.[1]

[1] The reader is reminded at this point that the tables show the model predictions, which may differ somewhat from actual purchases or the proposed purchasing plan for the farms, which is described in Ref. [1]. One important reason for this is that the timing of purchases relative to fiscal year boundaries, especially in the first years, may produce apparent deficits or excesses with respect to the annual averages from the model.

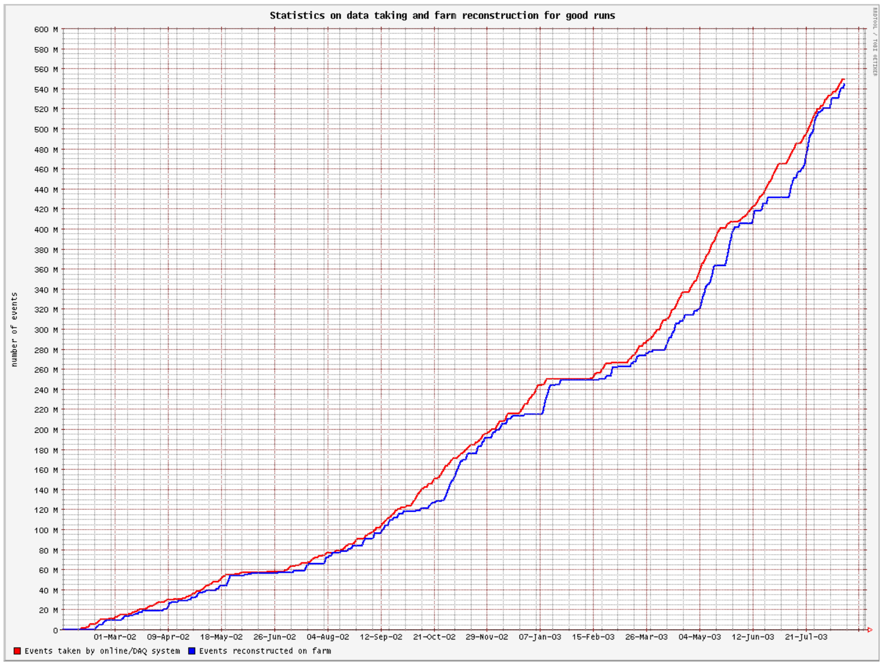

Figure 4: Number of good events logged (red) and processed on the production farm (blue) as a function of time from February, 2002 (the start of physics data taking) through August, 2003.

4 Central Analysis Facility

For the purposes of this note, the CAF consists of batch CPU and network attached disk.[2] We assume the primary use of these machines will be production of EDM-based secondary and tertiary datasets, and EDM-based user analysis on the batch systems. While we can reasonably estimate the CPU needed for secondary and tertiary dataset production, it is far more difficult to estimate the CPU requirements for user analysis since it depends critically upon usage patterns and analysis models, both of which are highly dependent upon the available resources, the ease of accessing those resources and the type of data upon which users rely.

In the following, we will adopt two different analysis models to estimate the required CPU power of the CAF. As a check, we will compare the model predictions against the CAF utilization observed during the 2003 winter conference season. We also attempt to bound the CAF capacity needed by applying simple scaling rules to the existing CAF, which is assumed to have been adequate during the 2003 winter conferences. This exercise will also provide a feel for the overall uncertainty the predictions.

[2]Interactive machines, although technically defined as part of the CAF [1], are assumed to be used for job development, debugging and management, tasks which are not within the scope of this analysis.

Table 6: Baseline procurement plan for the production farm. Nodes are retired after three years of service.

Fiscal year	02	03	04	05	06	07	08	09
Required (GHz)		790	711	648	790	1706	1990	2275
Added nodes		+116-73	+5-56	+9-56	+37-116	+34-5	+8-9	+14-37
Speed (GHz)		2.2	3.5	5.6	8.8	14.1	22.4	35
Cost ($1k)		255	11	20	81	75	18	31

Table 7: Upgrade procurement plan for the production farm. Nodes are retired after three years of service.

Fiscal year	02	03	04	05	06	07	08	09
Required (GHz)		790	948	1516	1232	2630	3057	3483
Added nodes		+116-73	+39-56	+66-56	+13-116	+59-39	+26-66	+9-13
Speed (GHz)		2.2	3.5	5.6	8.8	14.1	22.4	35
Cost ($1k)		255	86	145	29	130	57	20

4.1 Batch CPU model

For many analyses, the size of datasets scales approximately with integrated luminosity. This behavior is particularly evident for the high P_t datasets, many of which are not pre-scaled. The user analysis model for the CAF assumes that user datasets are a fixed cross section. (Further discussion of similar analysis models can be found in Refs. [5] and [6].) We then require that the CAF provide the capacity to allow 200 simultaneous users (from among about 400 physicists with active analyses) to process 5 nb datasets in a single day assuming an average event processing rate of 0.2 GHz-sec. The latter number is based upon an informal survey of analysis jobs on the CAF [7]. An independent, although low statistics survey by one of the authors produced the identical estimate. We will have further remarks on this quantity shortly when we examine the network throughput on the CAF. To arrive at the total CPU power needed in the CAF, we correct the capacity calculated above by the average CPU utilization and add an additional 30% contingency factor. The average CPU utilization observed during the winter conference period was about 80%[8].

4.1.1 Checks of luminosity-scaled analysis model

To test the model, we estimate the CPU needed for the winter conferences based upon the data accumulated until that time. The model predicts a required capacity of 380 GHz; the actual size of the CAF was 939 GHz, only one third of which provided access to the most popular datasets. The CPU utilization on the other two-thirds was about

50%. After removing the one third of the capacity that was unused, we find that about 630 GHz provided all the processing needed during the winter conferences. The typical number of users was about 100. The model prediction is low, despite assuming a larger number of users, but did not include any component that scaled with running time.

Another prediction of the model is the network activity associated with event I/O. Assuming that most users were reading production output with an event size of 180 kB, we estimate a network read rate of about 210 MB/sec during the 2003 winter conferences. The observed rate during this period is shown in Fig. 5, and is consistent with the model prediction.

We can also turn the calculation around and attempt to deduce the event reconstuction rate from the observed network throughput. Taking the observed throughput to be 250 MB/sec and assuming that all CAF users are analyzing data, we deduce an event reconstruction time of about 0.6 GHz-sec (i.e., 0.6 seconds per event on a 1 GHz P-III processor). This value is likely high since the CAF usage included Monte Carlo jobs, which have zero input, and user-run production jobs, which have a much lower throughput. Although we do not yet understand the relationship between the observed network I/O rate and the CPU utilization on the CAF, we can conclude that the assumed event processing time of 0.2 GHz-sec is within a factor of three of the actual average value for user analysis jobs on the CAF, and possibly much closer.

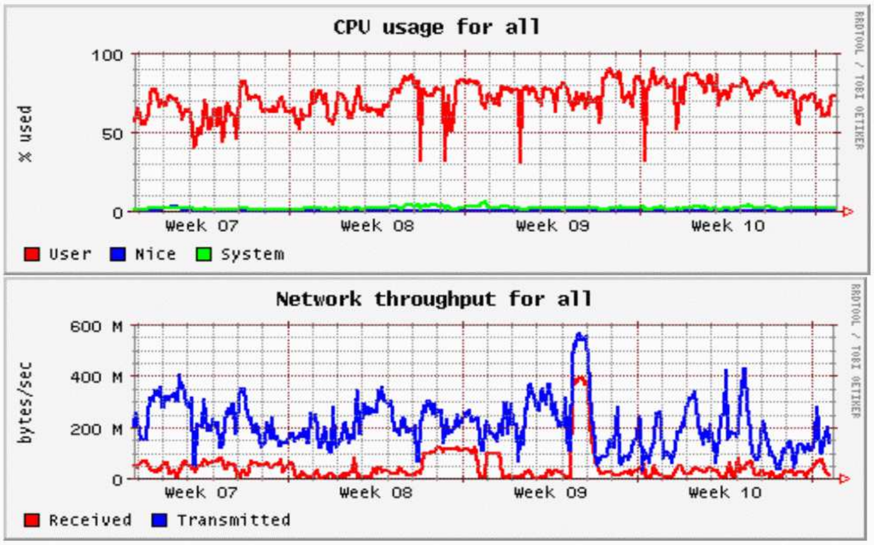

Figure 5: CPU utilization and network throughput on the CAF for a one month period during the winter conference season. The spike in the network activity was associated with tests of the data handling system.

It is instructive simply to scale the existing CAF, which was adequate for analysis needs during the 2003 winter conferences, by either integrated luminosity or integrated number of events logged (i.e., running time), and compare the result to that of the analysis-based model. Since the actual dataset has components with both types of scaling properties, we would expect that these two extrapolations represent extremes in the needed capacity of the CAF. Table 8 shows the results of this scaling exercise as a function of time. Note that in the near term, the analysis model predicts a lower CPU requirement than either naive scaling model. This is the same discrepancy we observed above between the estimated and actual utilization of the CAF during the winter conferences. After the first year, however, the analysis model predictions are within the range of values between the luminosity and run-time scaled extrapolations.

Collectively, these tests provide some confidence that the luminosity-scaled analysis model yields predictions that are within a sensible range of reality given the existing logging rates. We conclude that the luminosity-scaled model can be used to estimate the approximate cost of the CAF within the baseline model.

Table 8: Comparison of the baseline projections for the size of the CAF (in 1 GHz PIII equivalents) with scaling models based upon the size of the CAF used for the 2003 winter conferences. The luminosity and run-time scaling rules bound the expected scaling behavior.

Fiscal year	03	04	05	06	07	08	09
Baseline	960	2240	4510	7520	12700	20730	28920
Luminosity scaled	1686	3957	7960	13278	22420	36581	51041
Run-time scaled	1494	2490	3485	4979	7967	10954	13942

4.1.2 Run-time scaled analysis model

While the luminosity-scaled model appears to capture the current computing behavior of the experiment, this is unlikely to be true for the added event logging bandwidth, since these datasets by definition grow with running time (where we will use “running time” and “integrated number of logged events” interchangeably). We therefore treat the increased event logging rate separately using a run-time scaled analysis model. As before, the results will depend critically upon the analysis model we adopt. Unlike the case with the luminosity scaled models, however, no tests with available data exist to check whether the model predictions are even within an order of magnitude.

To address this problem, we adopt two “plausible” analysis models with greatly differing computing demands. The models are plausible in that they appear to the authors to represent a reasonable mode of operation for an analysis on a large dataset, which by its nature requires more care and planning in the use of computing resources than does an analysis on smaller datasets. The first is a multi-user model similar in form to the luminosity scaled analysis model used above. We assume a total of 15 simultaneous users (among perhaps 60 within the B group working with these datasets), each of whom requires 25 days to process the entire dataset. All bandwidth above 400 nb at the average instantaneous luminosity is assumed to contribute to the run-time

Table 9: Comparison of the estimated CAF computing needed using two different analysis models in which the computing scales with running time. The baseline component is subtracted so as to isolate the run-time scaled contribution. The multi-user model estimate is comparable to the baseline estimate, while the single-user model estimate falls well below it. Note that run-time scaling behavior starts in FY04.

Fiscal year	03	04	05	06	07	08	09
Avg. inst. lum (10^{31} cm^{-2} sec^{-1})	2.3	4.0	7.1	19	16	25	26
Avg event rate (Hz)	50	85	170	250	250	250	250
Event rate into dataset (Hz)	41	69	140	180	190	150	150
CPU multi-user (GHz)	0	1467	4482	6340	10283	13466	16605
CPU single-user (GHz)	0	122	374	528	857	1122	1384
Baseline CPU (GHz)	960	2240	4510	7520	12700	20730	28920

scaled dataset. The run-time scaling behavior begins in FY04 with the start of raw data compression. All other parameters are the same as before.

The second model represents a highly managed data reduction scheme more similar to the production of secondary datasets. A single user is assumed to access the dataset, with 20 days required to process the entire dataset. Such a model would allow a physics group to produce a highly compressed dataset several times a year (although it seems unlikely they would want to do so more than two or three times per year). All other parameters are the same, including the composition of the dataset.

Table 9 compares the results from these two models after subtracting the baseline component. In the multi-user model, a small number of people create a CPU demand that is comparable to the baseline demand from the rest of the experiment. A managed analysis policy for this large dataset could greatly reduce this demand, as evident from the single-user model. For reference purposes, we will estimate the cost of the CAF for both of these extreme models.

As a crude sanity check, we can again check the estimated size of the CAF in the upgrade scenario against the bounds from the simple scaling of the existing CAF. Table 10 contains the estimated CAF capacity for the single-user and multi-user models, and those obtained from the size of the 2003 winter CAF scaled by integrated luminosity and integrated number of events in the upgrade scenario. The single and multi-user models in this table are the sum of the luminosity-scaled analysis model (the baseline estimate in Table 9) and the single-user or multi-user model contributions (also in Table 9). The multi-user estimates lie just within the upper bound of the simple scaling estimates, while the single-user estimates lie near the lower bound.

4.1.3 Batch CPU cost estimate

To estimate the cost of the computing required under each of these scenarios, we assume the CAF consists of dual CPU Linux boxes, each of which costs \$2.2k in

Table 10: Comparison of the combined CAF size estimates in the upgrade scenario (in 1 GHz PIII equivalents) with scaling models based upon the size of the CAF used for the 2003 winter conferences. The luminosity and run-time scaling rules bound the expected scaling behavior.

Fiscal year	03	04	05	06	07	08	09
Single-user	955	2364	4884	8052	13560	21848	30303
Multi-user	955	3709	8992	13863	22985	34193	45524
Luminosity scaled	1686	3957	7960	13278	22420	36581	51041
Run-time scaled	1494	2988	5975	8216	12697	17178	21660

Table 11: Cost of CPU for the CAF under the baseline model. The capacity in FY02 is 914 GHz.

Fiscal year	03	04	05	06	07	08	09
Needs (THz)	0.960	2.240	4.510	7.520	12.700	20.730	28.920
Nodes added	66	184	264	187	230	245	162
Speed (GHz)	2.2	3.5	5.56	8.84	14.06	22.36	35.55
CPU added (THz)	0.290	1.287	2.937	3.307	6.468	10.955	11.517
Cost ($1k)	145	405	581	411	506	539	356

constant dollars. The processor speed is assumed to correspond to a Pentium III at 2.2 GHz in FY03, increasing by a factor of two every 18 months. Nodes are retired after 3 years of service. The capacity in FY02 is assumed to be 667 GHz, corresponding to the CDF-owned portion of the system. The procurement plan and cost of the baseline, upgrade multi-user and single-user models are shown in Tables 11–13. The estimated expenditures in FY03 are artifically low because the model averages needs and expenditures within a year and ignores purchasing lead times. These deficiencies lead to edge effects in the first year of the cost model.

4.2 Network attached disk

The basic plan for disk space attached to the CAF is to store as much processed data on disk as possible, while providing sufficient space for staging, data caching (if required), data validation and MC data storage. In addition to these uses, some disk is required to store ntuples or other highly compressed data samples coordinated by the physics groups. Here, we will only address the portion associated with secondary dataset storage and production.

During the 2003 winter conference season, a total of 56 TB was dedicated to the CAF for data storage. This quantity met the goal of storing almost all heavily used

Table 12: Cost of CPU for the CAF under the single-user analysis model. The capacity in FY02 is 914 GHz. Run-time scaling behavior begins in FY04.

Fiscal year	03	04	05	06	07	08	09
Needs (GHz)	955	2364	4884	8052	13560	21848	30303
Nodes added	66	201	287	196	246	256	168
Speed (GHz)	2.2	3.5	5.56	8.84	14.06	22.36	35.55
CPU added (GHz)	290	1406	3192	3467	6918	11447	11944
Cost ($1k)	145	442	631	431	541	563	370

Table 13: Cost of CPU for the CAF under the multi-user analysis model. The capacity in FY02 is 914 GHz. Run-time scaling behavior begins in FY04.

Fiscal year	03	04	05	06	07	08	09
Needs (GHz)	955	3709	8992	13863	22985	34193	45524
Nodes added	66	394	535	292	422	384	232
Speed (GHz)	2.2	3.5	5.56	8.84	14.06	22.36	35.55
CPU added(GHz)	290	2756	5951	5164	11867	17170	16494
Cost ($1k)	145	867	1177	642	928	845	510

secondary datasets on disk. We again consider the disk needed to store the fixed cross section datasets separately from the data associated with increased event logging rates. For the fixed cross section datasets, we depart from the normal scheme of luminosity scaling because the quantity of disk appear prohibitively large, especially in later years. We therefore elect to scale the winter fileserver capacity by the total estimated number of raw data events (see Table 2).

For the events from the run-time scaled dataset, the above prescription becomes prohibitively large. We therefore limit the size of the disk-resident portion to the number of events that can be processed in seven days on the CAF, under the assumptions for either the single-user or multi-user analysis models described in Sect. 4.1.2

The required fileserver capacity shown in Table 14 is the sum of two previous estimates. If we assume the baseline demand grows with integrated luminosity rather than data volume, then the FY04 estimate increases by about a factor of 1.5, increasing steadily to a factor of 4 by FY09. Note that all the disk in Table 14 is managed by the data handling system.

4.2.1 Cost estimate for network attached disk

Tables 15–17 present the procurement plan and cost of the baseline model, and the upgrade under the single-user and multi-user models. Fileservers are assumed to have a capacity of 5 TB in FY03, to double in capacity every 18 months (factor of 1.59 per year) and to cost \$15k each in constant dollars. Servers are retired after three years of service. The additional cost of fileservers in the upgrade plans are about the same because the number of events processed in a single day for the two models is nearly the same.

5 Data handling

We consider the Data Handling (DH) system to consist of the tapes and tape drives that comprise the data archive. Typically, the read and write disk caches are also considered as part of this system. The cost in all scenarios, however, is dominated by those associated with the archive, hence the focus on those components.

5.1 Data archive model

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 assumed to be EDM-based root files. We will ignore the contributions from tertiary datasets or other highly compressed files created by the physics groups, since these sources are expected to be relatively small.

To estimate the archive volume, we multiply the run-averaged logging rate by the average running efficiency within each FY and the event size after compression. All logged events are assumed to go into primary and secondary datasets; the size of these events is equal to the size of production output events (about 180 kB after compression). We have ignored the dependence of the event size on the instantaneous luminosity. In practice, raw data events are expected to increase in size from about 200 kB at 1×10^{31} cm^{-2}s^{-1}, to about 280 kB at 1×10^{32} cm^{-2}s^{-1}. To obtain the needed archive volume, the above size is multiplied by an additional contingency factor of 1.2.

To check the calculation, we estimate the archived data volume used for the winter 2003 conferences, accumulated between February, 2002, and January, 2003, to be about 180 TB. This value follows from an average logging rate of 50 Hz over 10.4 weeks of continuous beam time (52 weeks at a 20% running efficiency). Figure 6 shows the volume of raw and reconstructed data stored in the archive as a function of time. The arrows on the plot indicate the time interval during which the winter conference dataset was collected. A total of about 170 TB was added to the archive during this time, in good agreement with our estimate.

The estimated archive volume as a function of fiscal year for the baseline model is in Table 18 as a function of FY. The total volume in FY02 was 68 TB. For the upgrade model, we assume a raw data compression factor of 0.675 beginning in FY04, and a factor of two increase in CSL bandwidth beginning in FY05. The results for this model are shown in Table 19.

All I/O to the tape robot proceeds through a fixed number of tape drives mounted in the robot. The total I/O demands on the robot will determine the number of tape

Table 14: Estimated fileserver capacity on the CAF for the baseline model, and the single and multi-user run-time scaling models. behavior begins in FY04.

Fiscal year	03	04	05	06	07	08	09
Baseline (TB)	126	210	294	420	672	924	1176
Upgrade: single-user model (TB)	126	293	629	869	1357	1824	2289
Upgrade: multi-user model (TB)	126	285	604	834	1300	1749	2197

Table 15: Procurement plan for fileservers on the CAF under the baseline model. The capacity in FY02 is 84 TB. Servers are retired after three years of service.

Fiscal year	03	04	05	06	07	08	09
Needs (TB)	126	210	294	420	672	924	1176
Units added	8	11	7	6	8	5	3
Capacity/unit (TB)	5	8	13	20	32	51	81
Added capacity (TB)	40	87	88	121	256	254	242
Cost ($1k)	120	165	195	120	165	120	75

Table 16: Procurement plan for fileservers on the CAF under the single-user model. The capacity in FY02 is 84 TB. Servers are retired after three years of service.

Fiscal year	03	04	05	06	07	08	09
Needs (TB)	126	293	629	869	1357	1824	2289
Units added	8	21	27	11	16	9	6
Capacity/unit (TB)	5	8	13	20	32	51	81
Added capacity (TB)	40	167	341	221	511	457	485
Cost ($1k)	120	315	510	195	315	255	135

Table 17: Procurement plan for fileservers on the CAF under the multi-user model. The capacity in FY02 is 84 TB. Servers are retired after three years of service.

Fiscal year	03	04	05	06	07	08	09
Needs (TB)	126	285	604	834	1300	1749	2197
Units added	8	20	25	11	15	8	6
Capacity/unit (TB)	5	8	13	20	32	51	81
Added capacity (TB)	40	159	316	221	479	406	485
Cost ($1k)	120	300	480	195	300	255	120

drives that are required. To estimate the I/O to the archive, we sum the contributions from all sources: raw data logging; raw data read by the farms; production data written by the farms; production data read by the CAF; data cache misses; and secondary datasets written to the archive. We have neglected a few percent contribution from tape copies required to migrate to higher tape densities since these activities can likely be performed parasitically.

Data moving in or out of the archive is generally 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 (see Table 2).

The raw data read rate on the production farm, we take the product of the total number of events processed per year, the raw data event size and the ratio of the peak to average operating efficiencies from Table 2. The write rate is equal to the read rate times the ratio of the production output size (from Table 3) to the raw data size (Table 2). The total I/O rate on the farm depends upon the scenario, and may be a function of fiscal year. At present, raw data processed on the production farm must first be written to the archive, then read back to the farm. This procedure is required due to a technical limitation of dCache. The baseline model assumes that this scheme is used throughout the entire run. The baseline archive I/O rate to the farms is therefore the sum of the farms read and write rates. Recent developments in dCache will allow files to be copied from one staging pool to another, thereby avoiding the need for this read/write cycle, and one additional read of the production output by the CAF. The upgrade estimates assume that this feature is implemented starting in FY04, after which the archive I/O rate to the farm is equal to the production output rate alone.

The archive I/O rate to the CAF, as noted above, includes one contribution equal to the production output rate in the baseline, but not in the upgrade scenario. The archive read rate associated with user analysis is equal to the total read rate on the CAF times fraction of file requests that result in cache misses, which we assume to be 10%. The total event read rate is obtained from the number of events processed per

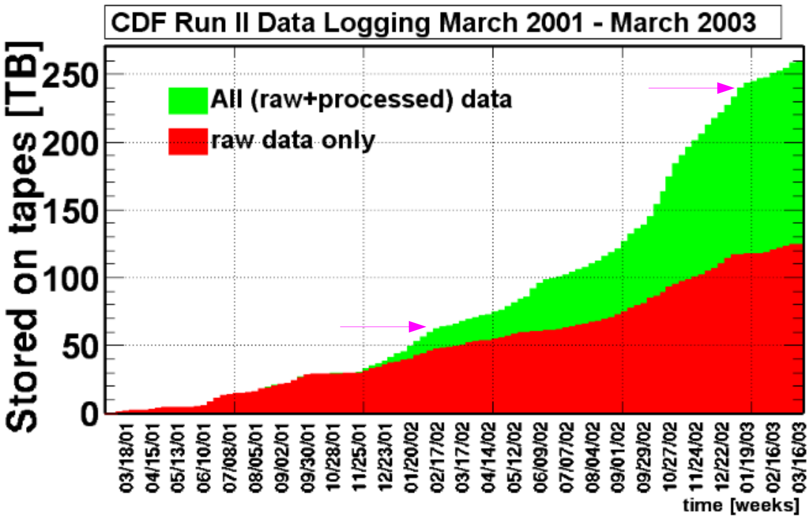

Figure 6: Volume of raw and reconstructed data stored in the tape robot as a function of time. The arrows indicate the time period during which the 2003 winter conference dataset was accumulated. The archive volume grew by about 170 TB during this time.

day (the product of the number of users and the number of events processed per day in a given analysis model, summed over models).

We have already noted in Sect. 4.1.1 that the data throughput estimates for the CAF during the winter conferences are in good agreement with the observed network throughput. As a further test of the I/O estimate, we can compare the total I/O load on the archive with the observed value during the 2003 winter conference season. Figure 7 shows the archive I/O rate during the winter and spring of 2003. The arrow points to the period of the winter conferences, during which we observed an average read rate of 6 TB/day. We estimated a demand of 56 MB/sec, or about 5 TB/day during this period.

Tables 18 and 19 present the results of these estimates as a function of fiscal year.

5.2 Media cost

The archive currently contains tapes of two densities with capacities of 60 GB and 200 GB , written by the STK 4490A and 4490B tape drives, respectively. Both densities utilize the same tape cartridge and are interchangeable between the two drives. Tapes at the lower density are therefore being copied to higher density tapes, and then added to the pool of empty high-density tapes. We assume that this process continues at a rate to meet all space demands until the supply is exhausted or a newer technology arrives, whichever comes first.

A new technology with a capacity of 300 GB per cartridge is expected to be avail-

Table 18: Baseline archive volume and I/O rates. The volume in FY02 is 68 TB. There are no pool-to-pool copies in the baseline model. The archive I/O estimates for the single-user model of the upgrade scenario are within a few percent of the baseline estimates.

Fiscal year	03	04	05	06	07	08	09
Avg. event rate (Hz)	50	50	50	150	150	150	150
Avg. operating effic.	0.3	0.3	0.3	0.15	0.3	0.3	0.3
Raw data event (kB)	220	220	220	220	220	220	220
Production event (kB)	220	135	135	135	135	135	135
Raw data volume (TB)	104	104	104	156	311	311	311
Production output (TB)	85	85	85	127	254	254	254
Secondary datasets (TB)	85	85	85	127	254	254	254
Archive/year (TB)	273	273	273	410	819	819	819
Archive volume (TB)	341	614	887	1297	2116	2935	3754
Raw data (MB/sec)	12	12	12	18	36	36	36
Farms I/O (MB/sec)	48	43	39	48	103	120	138
CAF I/O (MB/sec)	91	188	339	544	907	1427	1957
Archive I/O (MB/sec)	173	262	408	699	1093	1638	2192

able in FY04, making adoption possible starting in FY05. We have elected instead to construct the cost model to migrate to an as yet unspecified technology "X" in FY06 with twice the density of the existing 4490B tapes. The cartridges for this new technology will not be interchangeable with the current cartridges. We further assume that 4490B tapes will be migrated to X-tapes at a rate of 5500 tapes per year (the contents of one tape robot) starting in FY06. We have otherwise ignored total capacity constraints of the tape robot, since additional robots are relatively inexpensive.

To calculate the number of tapes needed, we take the estimated archive volume from Tables 18 and 19 and multiply by a contingency factor of 1.2. The tapes are assumed to cost $75 each in constant dollars. The results for the baseline and upgrade scenarios are shown in Tables 20 and 21. (The tape requirements are the same for the single-user and multi-user models.)

The cost of tape migration is about $200k in FY06 for the baseline model, and $300k over two years for the upgrade model. Migrating to the 4490C technology with the 300 GB capacity would increase the migration cost by about $50k in each year of transition, with additional outlays of $40k ($30k) to $100k ($60k) per year in the upgrade (baseline) scenario in order to keep up with growth of the archive. All amounts are in constant dollars.

Table 19: Archive volume in the upgrade scenario and multi-user model. The volume in FY02 is 68 TB. Pool-to-pool copies begin in FY04.

Fiscal year	03	04	05	06	07	08	09
Avg. operating effic.	0.3	0.3	0.3	0.15	0.3	0.3	0.3
Raw data event (kB)	220	220	220	220	220	220	220
Production event (kB)	220	135	135	135	135	135	135
Avg. event rate (Hz)	50	85	170	250	250	250	250
Raw data volume (TB)	104	119	238	175	350	350	350
Production output (TB)	85	144	288	212	423	423	423
Secondary datasets (TB)	85	144	288	212	423	423	423
Archive/year (TB)	273	407	814	598	1197	1197	1197
Archive volume (TB)	341	748	1562	2160	3357	4553	5750
Raw data (MB/sec)	12	12	24	18	36	36	36
Farms I/O (MB/sec)	48	26	41	33	71	83	95
CAF I/O (MB/sec)	91	282	651	978	1597	2332	3073
Archive I/O (MB/sec)	173	320	716	1099	1754	2451	3204

5.3 Tape drives

The current archive uses the STK T9940A and T9940B tape drives with maximum I/O rates of 10 MB/sec and 30 MB/sec, respectively. All 9940A drives will be retired in FY03. As discussed in the previous section, we elect to migrate to a technology "X" in FY06. We assume these drives to have a maximum I/O rate of 60 MB/sec. The 4490B drives will be retired over two years starting in FY06, with half of the FY05 inventory retired each year.

To calculate the needed I/O capacity, we take the estimated I/O bandwidth from Tables 18 and 19 by a contingency factor of two to take into account tape drive contention, separation of reads and write, down-time and provide a safety margin in the estimates. We ignore the constraint 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. Tables 22–23 show the procurement plan for tape drives as a function of fiscal year for all models.

6 Summary of costs

The cost estimates for the above systems ignore a number of expenditures related to infrastructure, routine maintenance and operations. In order to arrive at a more realistic summary of the total offline budget, we need to include these additional expenditures. The cost summaries in Tables 25–27 show these added expenses as a function of fiscal year, and the sum of all expenses. We assume that these additional costs are inde-

Table 20: Media procurement plan under the baseline model. All 200 GB tapes are copied to 400 GB tapes in FY06. The "cartridges added" row includes only those tapes needed to accommodate new data, while the "migration needs" row covers the tapes required for migrating from the lower density cartridges. The existing inventory of 4490B tapes supplies anticipated needs through 2005.

Fiscal year	03	04	05	06	07	08	09
Capacity needed (TB)	410	737	1065	1556	2539	3522	4505
Tape capacity (GB)	200	200	200	400	400	400	400
Cartridges added	0	0	0	1231	2457	2458	2457
Migration needs	0	0	0	2662	0	0	0
Cost ($1k)	0	0	0	292	184	184	184

Table 21: Media procurement plan under the upgrade scenario. All 200 GB tapes are copied to 400 GB tapes in FY06 and FY07. The "cartridges added" row includes only those tapes needed to accommodate new data, while the "migration needs" row covers the tapes required for migrating from the lower density cartridges.

Fiscal year	03	04	05	06	07	08	09
Capacity needed (TB)	410	898	1874	2592	4028	5464	6900
Tape capacity (GB)	200	200	200	400	400	400	400
Cartridges added	0	0	2371	1795	3590	3589	3590
Migration needs	0	0	0	2750	1935	0	0
Cost ($1k)	0	0	178	341	414	269	269

Table 22: Baseline procurement plan for tape drives. The 9940A drives are retired prior to the FY03 entry in the table. The 4490B drives are retired over two years starting in FY06.

Fiscal year	03	04	05	06	07	08	09
I/O capacity needed (MB/sec)	345	525	816	1398	2186	3276	4385
Drive I/O rate (MB/sec)	30	30	30	60	60	60	60
Drives added	2	5	9	14	13	19	18
Cost ($1k)	60	150	270	420	390	570	540

Table 23: Procurement plan for tape drives under the upgrade scenario and the single-user model. The 9940A drives are retired prior to the FY03 entry in the table. The 9940B drives are retired over two years starting in FY06. The cost of this model is slightly less than that of the baseline because of reduced archive I/O allowed by pool-to-pool copies.

Fiscal year	03	04	05	06	07	08	09
I/O capacity needed (MB/sec)	345	491	979	1556	2467	3539	4727
Drive I/O rate (MB/sec)	30	30	30	60	60	60	60
Drives added	2	4	12	14	15	18	20
Cost ($1k)	60	120	360	420	450	540	600

Table 24: Procurement plan for tape drives under the upgrade scenario and the multi-user model. The 9940A drives are retired prior to the FY03 entry in the table. The 9940B drives are retired over two years starting in FY06.

Fiscal year	03	04	05	06	07	08	09
I/O capacity needed (MB/sec)	345	640	1432	2198	3508	4902	6408
Drive I/O rate (MB/sec)	30	30	30	60	60	60	60
Drives added	2	9	19	18	21	24	25
Cost ($1k)	60	270	570	540	630	720	750

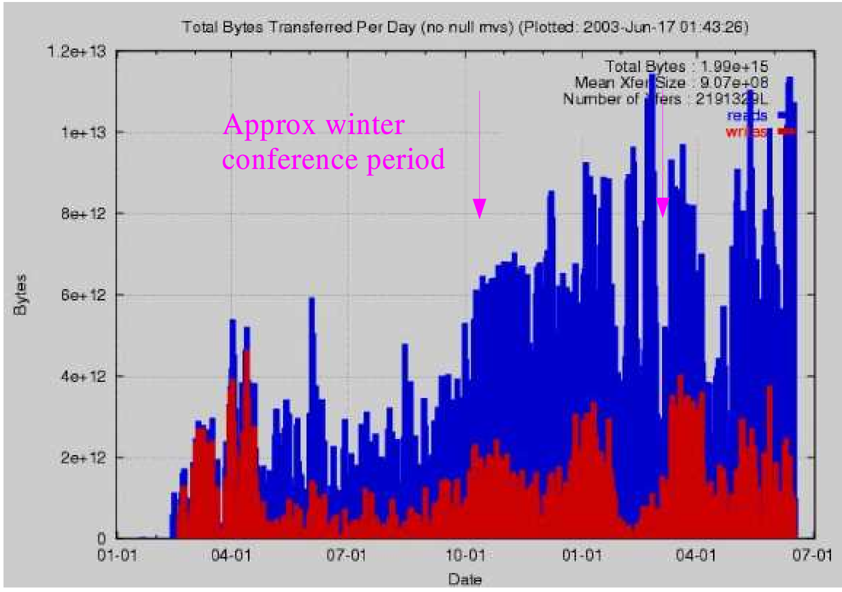

Figure 7: Tape I/O on the archive (TB/day) as a function of time. The arrows indicate the test period from Oct., 2002, through March, 2003. The model baseline model predicted a total throughput of about 5 TB/day.

pendent of the computing scenario, and that there are no significant technical issues associated with scaling infrastructure systems that might significantly change the cost structure. We have also ignored manpower issues. Note that the estimates for FY03 are somewhat low due to edge effects in the cost models.

7 Conclusions

We have developed models to estimate the computing resource needs as a function of fiscal year for those systems that drive the cost of offline computing. Several key predictions of the models show reasonable agreement with computing data from the 2003 winter conference season, although significant uncertainties remain in the estimates, particularly in the needs of the CAF. Assuming, however, that the demand on the CAF is driven by the size of datasets and that the size of the CAF during the winter conferences was adequate to satisfy the demand, we show that our estimates lie within the range determined by the extremes in the scaling behavior of demand.

From the model estimates of demand, we determine the needed capacities and cost of the various components under the baseline and upgrade scenarios. These results are used as input to the CDF computing plan and budget described in CDF-6640.

Table 25: Summary of costs (in units of $1k) for baseline scenario.

Fiscal year	03	04	05	06	07	08	09
Farm CPU	255	11	0	31	112	15	9
Farm misc	50	50	50	50	50	50	50
CAF CPU	145	405	581	411	506	539	356
Fileservers	120	165	195	120	165	120	75
Media	0	0	0	292	184	184	184
Drives	60	150	270	420	390	570	540
DB	150	150	150	150	150	150	150
Networking	250	250	250	250	250	250	250
Misc	100	100	100	100	100	100	100
Total	1130	1281	1596	1824	1907	1979	1714

Acknowledgements

The authors thank the following people for providing input for use in the computing models: Yen-Chu Chen, Robert Kennedy, Jonathan Lewis. Dmitri Litvintsev, Rick Ste. Denis, Alan Sill, Jeff Tseng, Tony Vaiciulis, Steve Wolbers and Frank Wuerthwein.

References

[1] CDF-5914, "CDF Plan and Budget for Computing in Run 2", 2002.

[2] CDF-6640, "CDF Pland and Budget for Computing in Run 2: Second Annual Edition", ed. by R. Harris and F. Wuerthwein, 2003.

[3] Presentation by J. Spalding at the DOE Review of the Tevatron Run II Luminosity Upgrades, July, 2003.

[4] Pekka Sinervo, private communication.

[5] CDF-5785, "Physics Analysis Computing needs assessment, CDF CAF Review, Fall 2001", 2001.

[6] CDF-5802, "Final Report, CDF CAF Review, Fall 2001", 2001.

[7] Frank Wurthwein, private communication.

[8] Frank Wurthwein, "User Analysis Computing at CDF", presentation at the CHEP 2003 conference, March 24–28, 2003.

Table 26: Cost summary (in units of $1k) under the upgrade scenario and the single-user model.

Fiscal year	03	04	05	06	07	08	09
Farm CPU	255	86	123	0	136	33	33
Farm misc	50	50	50	50	50	50	50
CAF CPU	145	442	631	431	541	563	370
Fileservers	120	315	510	195	315	255	135
Media	0	0	178	341	414	269	269
Drives	60	120	360	420	450	540	600
DB	150	150	150	150	150	150	150
Networking	250	250	250	250	250	250	250
Misc	100	100	100	100	100	100	100
Total	1130	1513	2352	1937	2407	2210	1957

Table 27: Cost summary (in units of $1k) under the upgrade scenario and the multi-user model.

Fiscal year	03	04	05	06	07	08	09
Farm CPU	255	86	123	0	136	33	33
Farm misc	50	50	50	50	50	50	50
CAF CPU	145	867	1177	642	928	845	510
Fileservers	120	300	480	195	300	255	120
Media	0	0	178	341	414	269	269
Drives	60	270	570	540	630	720	750
DB	150	150	150	150	150	150	150
Networking	250	250	250	250	250	250	250
Misc	100	100	100	100	100	100	100
Total	1130	2073	3078	2268	2959	2672	2233