DØ Computing and Software Operations and Upgrade Plan

DØ Collaboration

May 22, 2002

Acknowledgments to Bonnie Alcorn, Iain Bertram, Amber Boehnlein, Chip Brock, Mike Diesburg, Dave Fagan, Stu Fuess, Nick Hadley, Alan Jonckheere, Qizhong Li, Lee Lueking, Harry Melanson, Wyatt Merritt, Dugan O'Neil, Don Petravick, Ruth Pordes, Serban Protopopescu, Jianming Qian, Heidi Schellman, Dane Skow, Terry Wyatt, Jae Yu

CHAPTER 1 – Introduction

In this document, we present the DØ Computing and Software operation and upgrade plan for 2003-2007. This period essentially covers Run 2a and 2b. The first years of the plan will be covered in the most detail. In later years, we present options for those cases where the best choice is to remain flexible to take advantage of changing hardware and lower costs.

The earlier plan for DØ software and computing, as reviewed by the Von Rüden committee, has been successfully carried out. We are taking data, storing it, and analyzing it. The first results based on Run 2a data have been shown at conferences. This earlier plan covered the period from 1997 to the present. It included writing data to robotic storage in the Feynman computing center, cataloging the data and providing access transparently from disk or robotic storage via the Sequential Access by Metadata system (SAM), using a large symmetric multi-processor machine (DØmino) to provide user access to large amounts of disk and computing power with high I/O capacity, and converting our software (and physicists) from Fortran to C^{++}. The plan also made substantial use of collaboration resources remote from Fermilab. For example, we have succeeded in generating essentially all Monte Carlo events for the experiment in off-site farms, as we proposed. The Run 2a model has been quite successful and, in most cases, we intend simply to scale up the systems, such as the reconstruction farm and mass storage system, to meet the needs. SAM has provided an extremely robust approach to data handling, as it is flexible in terms of hardware deployment and location with a transparent interface to the users. This gives us the ability to supplement DØmino with a large number of commodity processing nodes and to increase the role of computing resources offsite by deploying a set of regional centers to provide analysis and reprocessing capabilities.

This document details the equipment spending to cover both the operation of our existing system and upgrades to it necessitated by an increase in the data taking capabilities of the detector and an increase in the complexity of the events we will take. DØ is now capable of writing the equivalent of 25 Hz average to tape. We expect this capability to increase to the equivalent of 50 Hz average by 2005. Simultaneously, the luminosity is expected to increase from the current value of 2×10^{31} cm^{-2}sec^{-1} to 5×10^{32} cm^{-2}sec^{-1} again by 2005, with a corresponding increase in the complexity of the events. We have used the laboratory's luminosity profile from Steve Holmes' January 2002 talk to HEPAP as an input to this report.

For the purposes of making estimates, assumptions about the data rates have been made. Run 2 consists of two phases, where Run 2a covers the time between now until the shutdown for the installation of the new silicon tracker and trigger upgrades. For Run 2a, we retain previous assumptions from the 1997 plan about data rates and size per event for data tiers. Run 2b is the 4-year period after the upgrade, with 2005 providing the transition. We take the average data acquisition output rate in Run 2b as 50 Hz, twice the current nominal rate, corresponding to 100+ Hz peak output rate. We assume we collect data at the average rate for all instantaneous luminosities as in our experience the triggers are opened to fill all available bandwidth. Ideally, we will require MC generation to produce data samples comparable to half the collider data rate.

As mentioned, the DØ Computing and Software model relies on contributions from the DØ collaborating institutions. For example, the Monte Carlo production of the complete chain of generation, detector simulation, digitization, reconstruction, and trigger simulation takes place offsite at the remote production centers. In addition, we expect that many groups will pursue analysis at their home institutions. The DØ Regional Analysis Center Working Group is studying requirements and potential organizations to facilitate remote analysis. We are investigating the feasibility of supporting the re-reconstruction of collider data at the remote centers.

In addition, the CLuEDØ desktop cluster is composed of machines contributed by the institutions and is managed by members of DØ contributing institutions. Institutions have provided project disk on DØmino, and we anticipate that model will continue. We also anticipate that institutions will contribute to CLuB, the CLuEDØ batch facility, or the Central Analysis Backend (CAB) on DØmino, both of which are Linux analysis farms.

In chapters 2-6 of this document, we present the details of the plan for the various components of DØ Computing and Software, together with our needs, usage patterns and assumptions. Chapter 7 shows a proposed budget and summary.

CHAPTER 2 – Simulation And Reconstruction

2.1 Monte Carlo

The generation of Monte Carlo events in D∅ involves multiple stages and many executables. To integrate all processes, it was decided early on that all programs will use the D∅ event data model (EDM) to carry data in memory and the D∅ object model (D∅ OM) to store persistent event data. In addition, all code was to be organized in independent packages running in a standard D∅ framework. A major implication of these decisions is that the code must be written in C^{++}, or at the very least embedded in C^{++} driving routines.

The first step in Monte Carlo event generation involves the simulation of a physical process, a proton antiproton collision producing a particular final state. Quite a few programs exist that do this and the challenge is to ensure that any of them can be used in D∅ simulation. Almost all existing event generator programs have been written in Fortran. Fortunately, the FNAL CD division maintains code (StdHep) to store the output of the most commonly used in a standard common block format. Hence, all D∅ needed to do was to write a C^{++} wrapper that converts the StdHep Fortran format to C^{++} classes that satisfy the EDM requirements.

After simulating a reaction, the next step is to trace the particles through the D∅ detector, find where they intersect active areas and simulate their energy deposition and secondary interactions. For this, D∅ uses the CERN program Geant v3.21, which is also written in Fortran. A C^{++} wrapper is used to read files produced by the event generators, and to write out the output of Geant in D∅OM format. This executable is called d0gstar. All subsequent steps in the event simulation are handled by programs written almost entirely in C^{++}.

After the particles from the simulated reaction have been traced through the detector, the energy deposition needs to be converted to the form that the real data takes when processed through the D∅ electronics. One also needs to include detector inefficiencies, noise (from the detector and electronic readout), and to take into account the fact that more than one interaction may occur during a beam crossing. Furthermore, some portions of the detector (like the calorimeter) remain sensitive to interactions over a period of time that includes more than one beam crossing. These effects are handled by the D0Sim program. In addition to simulating the data readout electronics, D0Sim is also necessary to simulate the trigger electronics and the effects of the trigger on data selection. This is taken care of by a separate program, D0Trigsim. The program D0TrigSim contains simulation code only for the level 1 trigger. The level 2 and level 3 triggers consist of filtering code running on processors specially designed for this purpose, and thus the same code running in the level 2 and level 3 processors runs in D0TrigSim. The output of D0Sim and D0TrigSim is in the same format as the data recorded by the D∅ data acquisition system, but contains additional Monte Carlo information to make it possible to correlate detector information with the original generator information.

2.2 Reconstruction

The DØ Offline Reconstruction Program (RECO) is responsible for reconstructing objects that are used to perform all DØ physics analyses. It is a CPU intensive program that processes either collider events recorded during online data taking or simulated events produced with the DØ Monte Carlo (MC) program. The executable is run on the offline production farms and the results are placed into the central data storage system for further analysis. The program uses the DØ Event Data Model (EDM) to organize the results within each event. EDM manages information within the event in the form of *chunks*. The Raw Data Chunk (RDC), created either by the Level 3 trigger system or the MC, contains the raw detector signals and is the primary input to RECO. The output from RECO is many additional chunks associated with each type of reconstructed object. RECO is designed to produce two output formats which can be used for physics analyses, and which are optimized for size. The Data Summary Tier (DST) contains all information necessary to perform any physics analysis, and is designed to be 0.150 MB per event. The Thumbnail (TMB) contains a summary of the DST, and is designed to be 10 KB per event. The TMB can be used directly to perform many useful analyses. In addition, it allows the rapid development of event selection criteria that will be subsequently applied to the DST sample. Currently, a root-tuple intended primarily for RECO debugging is generated; however, support for this format will end in July 2002 as it is costly to produce both in computing time and storage.

RECO is structured to reconstruct events in several hierarchical steps. The first involves detector-specific processing. Detector *unpackers* process the RDC by unpacking individual detector data blocks. They decode the raw information, associate electronics channels with physical detector elements and apply detector specific calibration constants. For many of the detectors, this information is then used to reconstruct *cluster* (for example, from the calorimeter and preshower detectors) or *hit* (from the tracking detectors) objects. These objects use geometry constants to associate detector elements with physical positions in space. The second step in RECO focuses on the output of the tracking detectors. Hits in the silicon (SMT) and fiber tracker (CFT) detectors are used to reconstruct *global tracks*. This is one of the most CPU-intensive activities of RECO, and involves running several algorithms. The results are stored in corresponding track chunks, which are used as input to the third step of RECO, vertexing. First, *primary vertex* candidates are searched for. These vertices indicate the locations of ppbar interactions and are used in the calculation of various kinematical quantities (e.g. transverse energy). Next, displaced *secondary vertex* candidates are identified. Such vertices are associated with the decays of long-lived particles. The results of the above algorithms are stored in vertex chunks, and are then available for the final step of RECO – *particle identification*. This step produces the objects most associated with physics analyses and is essential for successful physics results. Using a wide variety of sophisticated algorithms, information from each of the preceding reconstruction steps is combined and standard *physics object* candidates are created. RECO first finds electron, photon, muon, neutrino (missing ET) and jet candidates, that are based on detector, track and vertex objects. Next, using all previous results, candidates for heavy-quark and tau decays are identified. Additional physics object identification is planned (e.g. K_s, Λ, J/ψ, W, Z, etc.) and will be added as the reconstruction algorithms become available.

The current version of RECO (p10.15.01) requires about 15 seconds per event to process recently obtained collider events (on a 500 MHz benchmark machine). This time breaks down for each major step as follows - detector: 2 seconds, tracking: 8 seconds, vertexing: 0.2 seconds, particle identification: 3 seconds. MC studies indicate that these times will grow significantly as the instantaneous luminosity of the accelerator (and thus the number of interactions per event) increases. For example, an increase of a factor of 14 is observed in tracking times when going from 2 to 5 interactions per event. In addition, the current efficiency for finding tracks in busy environments (i.e. jets) is low (50 – 70%), and improving the efficiency may require more CPU time. These issues are of significant concern, and efforts are ongoing to speed up existing algorithms and develop new, faster ones. However, it is not yet clear how successful these developments will be. For planning purposes, we do not assume a speedup of the reconstruction. Using current time estimates and breakdowns for Monte Carlo and data, we estimate the reconstruction time per event for various instantaneous luminosities. We expect that these estimates to be low as the number of interactions per crossing was taken from the straight calculation of the luminosity without accounting for the fact that the trigger will bias the event selection to higher multiplicity events. We know from the Monte Carlo studies that the processing time increases dramatically for physics enriched samples. Therefore, we assume a reconstruction processing time of 50 sec/event for Run 2b.

Instantaneous Luminosity ($cm^{-2}sec^{-1}$)	Estimated Reconstruction processing time (500 MHz processor) (sec/event)
9e31	25
20e31	35
50e31 (396 nsec crossing)	80
50e31 (132 nsec crossing)	32

Table 2.1 shows the estimated reconstruction time for various points of instantaneous luminosity. We used the p10 measurements for data, added the particle ID times for Monte Carlo Z events, and scaled based on the known tracking performance as a function of number of interactions.

CHAPTER 3 – Data Handling And Storage Needs

The Sequential Access via Meta-data (SAM) data handling system, jointly developed in the Computing Division within the Online and Database Systems department (ODS/CD) and DØ departments, is a software system that oversees the functions of cataloging data (files and events, and associated metadata regarding production conditions), transferring data in and out of mass storage systems, transferring data among different computer systems (whether connected via local or wide area network), allocating and monitoring computing resources (batch slots, tape mounts, network bandwidth, disk cache space), and keeping track at the user process level of file delivery status. The bookkeeping functions of the SAM system are handled by an ORACLE database, which is accessed via a client-server model utilizing CORBA technology. SAM can be interfaced to different mass storage systems and to different batch schedulers. Files are stored in SAM using interfaces that require appropriate metadata for each file. The files are organized, according to the metadata provided, by data tier (that is, raw, reconstructed, and various summary formats), and by production information (program version which produces the data, etc.). The SAM system also provides file storage, file delivery, and file caching policies that permit the experiment to control and allocate the computing resources. Tape resources can be guaranteed to high priority activities (data acquisition and farm reconstruction); high usage files can be required to remain in the disk cache; different priorities and allocations for resource usage can be granted to groups of users.

The SAM system uses three key concepts, “dataset”, “project”, and “snapshot”, in the delivery of files to the user process. The user first creates a dataset, which is a description of a set of conditions to define a list of files from those files cataloged in the SAM database. Then, the user runs a project, which consists of a user process to which the files in the dataset are delivered. The actual files that the dataset definition corresponds to, at the time the project is run, constitute a dataset “snapshot”. The SAM database keeps track of dataset definitions, of the projects run, and of the snapshots that correspond to the projects. Monitoring, for a system of this complexity, is clearly important as well. SAM provides web pages that indicate the health of various elements of the system, and web interfaces to the information in the SAM database. The most important SAM tool seen by the users is the dataset definition editor. It is accessible via either a web interface or a command line interface, and can also access some of the information in other ORACLE databases linked to the SAM database via run number, e.g., the run configuration database. In summary, the SAM system gives users access to all the files created by the DØ experiment (both detector data and simulation data), in a very flexible and transparent manner – the user does not need to know where the files are physically stored, nor worry about exactly how they are delivered to his/her process. SAM also permits the experiment considerable flexibility in apportioning its computing resources. A more extensive discussion can be found in the documents listed in the bibliography.

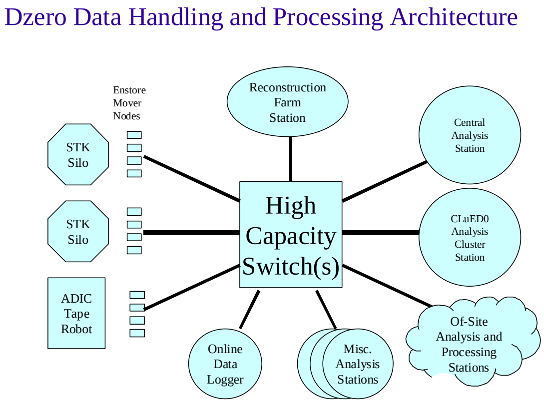


Figure3. 1. All Dzero data storage, processing and analysis systems are connected through high speed networks.

The general hardware architecture currently implemented for data handling is shown in Figure 3.1. It is a network-based approach and is extremely modular and scalable.

There are currently over two dozen operational production SAM stations deployed at Fermilab and remote institutions including the online data logger, the FNAL reconstruction farm, the Central Analysis system (DØmino), a large cluster of Linux desktop machines called CLuEDØ, an analysis and calibration station, and test stations. Six major processing centers have been using these stations for two years to send Monte Carlo data to the central tape storage system at FNAL.

◆**Fermilab (5 stations)**	**Batavia, IL**
◆**Imperial College (2)**	**London,UK**
◆**IN2P3**	**Lyon, France**
◆**Lancaster**	**Lancaster, UK**
◆**Munich**	**Munich, Germany**
◆**NIKHEF**	**Amsterdam, NL**
◆**Prague**	**Prague, Czech Republic**
◆**Wuppertal**	**Wuppertal, Germany**
◆**Boston University**	**Boston, MA**
◆**University of Arizona**	**Tucson, AZ**
◆**U. Texas, Arlington (2)**	**Arlington, TX**
◆**U. Oklahoma, Langston**	**Langston, OK**
◆**Indiana University**	**Bloomington, IN**
◆**Louisiana Tech**	**Ruston, LA**
◆**University of Kansas**	**Lawrence, KN**
◆**Michigan State University**	**East Lansing, MI**

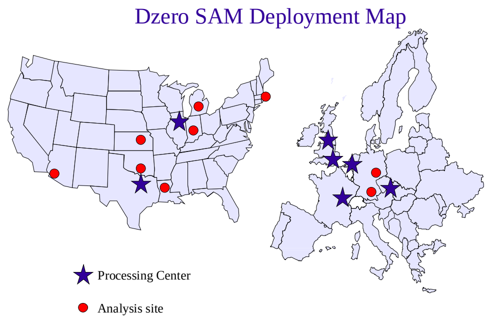


Figure 3.2 Some DØ Locations where SAM Stations are deployed. The number increases regularly.

Although the system is working quite effectively now, we have plans to improve and streamline the operation by 2005. The system will move toward a less centralized model with more station autonomy and independence from the central database at Fermilab Each station (or possibly site) will have its own information services that track operational information for the station, such as cache history and project activity. A global information service will access the activities for all stations and monitor the overall health and activity of the station network.

This decentralization will remove the current single-point-of-failure inherent in the central database and greatly improve performance of the system as it is scaled to the world at large. It is also part of the natural progression of the system toward a "standard" Grid system. We will soon be using components from the Globus toolkit, and job scheduling using Condor. We plan to provide standard interfaces to our data that will include those used by Storage Resource Manager (SRM), an emerging standard in the grid world. Compliance with standards is vital, as our collaborating institutions have computing resources shared by multiple experiments.

3.2 Hierarchical Storage

DØ's data management system relies heavily on Hierarchical Storage Management (HSM) systems for archival storage. The principal HSM used by SAM is Enstore, developed at Fermilab by the Integrated Systems Department of the Fermilab Computing Division (ISD/CD), and largely influenced by DØ requirements. Enstore is deployed at Fermilab and Lancaster University.

ISD is working collaboratively with DESY to provide a disk cache and buffering system (dCache) that acts as a front-end buffer to the tape robot. This will provide direct interfaces to the cache through standard protocols like ftp and GridFTP, allowing any SAM station worldwide to access data directly from a dCache server without going through a specially configured Fermilab SAM station. Additionally, data that are being stored to tape will be available on disk for a short while for reading, allowing the reconstruction farm and the analysis jobs access without tape reads.

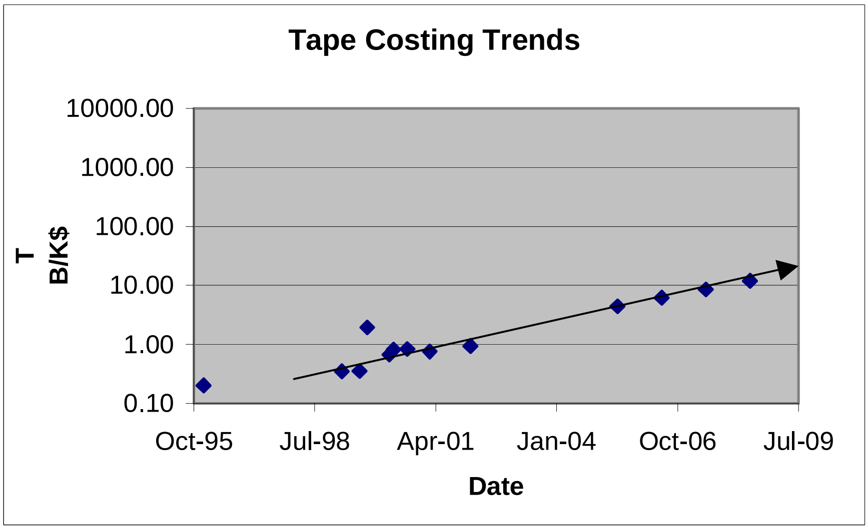
3.3 Hardware Storage Strategies and Costs

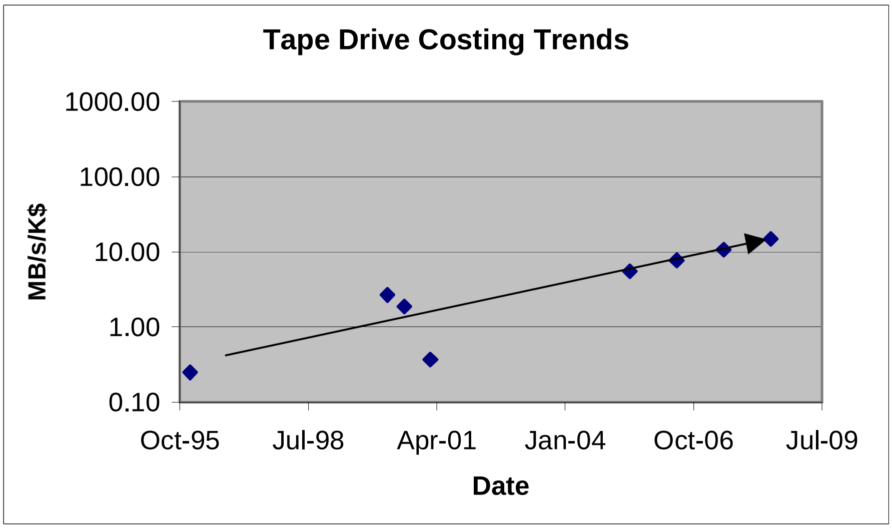
Tape technologies are constantly improving and the densities of data continue to increase. DØ has good experience with two tape technologies in the current run -- STK 9940 drives and media used with an STK Powderhorn robot and the first generation IBM LTO drives and media used in the ADIC AML/2 robot. We will project these products' development schedule into Run 2b and look at possible data storage costs. The constantly declining costs of IDE-type disk will undercut the price for tape, if one considers only the media cost, late in Run 2b, but the operational and deployment costs related to disk are expected to surpass those of tapes for some time.

ISD has projected the costs of drives and media in 2003 and beyond based on information from our current vendors. Recent price trends projections support this and are shown in the charts below (Figure 3.3). STK will double the capacity of the 9940 drives near the end of 2002, while maintaining the same cartridges and media, with the introduction of the 9940B. They have a roadmap for a new drive technology in 2004 that will require new media, for which the capacity per cartridge is not known. IBM and other LTO consortium members plan to sell a drive in early 2003 with cartridge capacity of 200 GB requiring new media, and IBM is working on a technology with a 400 GB (new media) cartridge by end of 2004.

We project that storage technologies will continue along their current declining price trends with tape capacity per unit cost doubling every 2 years, and disk capacity doubling every 18 months. Table 2 below indicates possible tape cartridge capacity and cost per

GB storage. The tables begin in CY 2003 where we are reasonably confident in the numbers and project to CY 2009 assuming the capacity of cartridges doubles every 2 years, and the cost per cartridge remains constant. The Commodity Off The Shelf (COTS) disk drive numbers assume the capacity per drive doubles every 18 months and the price per drive remains constant. We use CDF's experience for cost of drives and associated components for RAID network-attached disk (see CDF Plan and Budget for Computing in Run 2) to give a realistic appreciation of the actual cost of the storage system, rather than the raw cost/GB of IDE disk.





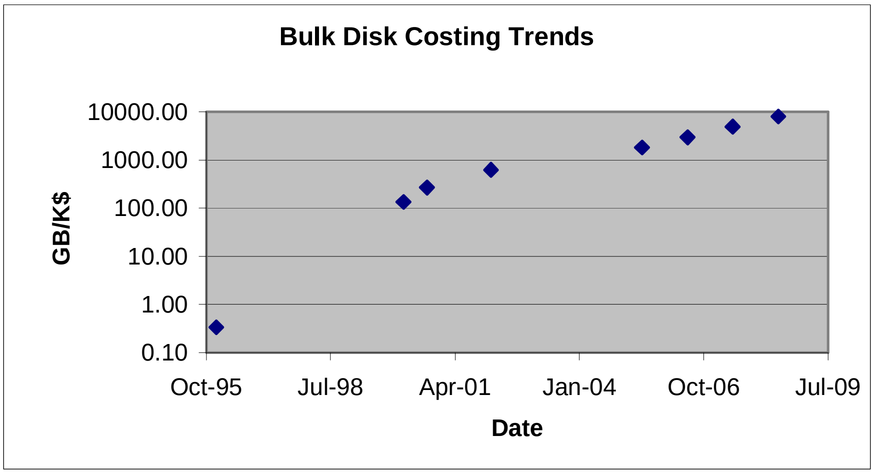


Figure 3.3 Charts showing tape and drive costing trends. The chart on the top shows TB/k$ vs. time, and the one in the middle is MBps/k$ vs. time. The lower chart shows bulk disk costing trends in GB/k$ vs time.

	2003		2005		2007		2009	
	GB	$/GB	GB	$/GB	GB	$/GB	GB	$/GB
STK	120	0.65	250	$0.30	500	0.15	1000	0.07
LTO	200	0.50	400	$0.25	800	0.12	1600	0.06
Disk (COTS)	200	4.00	800	$1.30	3200	0.40	12800	0.25

Table 3.1 Cartridge/Disk capacity, tape/disk cost

It is possible that the cost for bulk disk storage will be comparable to tape sometime during Run 2b. The ISD/CD department is working to understand how disk farms might be deployed to replace tapes, and tape robots. The experience to date indicates that the reliability of such repositories is not sufficiently high to quickly move in this direction. Also the effort involved in commissioning such disk drive facilities is still high. However, it is apparent that we will transition toward major disk storage facilities, especially for smaller files produced near the end of the detector data processing chains. As the costs of disk falls having RAID-5, or even mirrored data sets, becomes cost effective. There is no question that, even near the beginning of the run when primary copies of data still reside on tape, files will be replicated on (SAM /GRID) caching resources at Fermilab and many regional data centers. The sum of these disk caches may eventually represent storage comparable to the permanent tape storage supplied for primary and secondary data sets.

	2003		2005		2007		2009	
	MB/s	k$each	MB/s	k$each	MB/s	k$each	MB/s	k$each
STK	20	30	40	30	80	30	160	30
LTO	20	11	40	11	80	11	160	11
Disk (COTS)	10	0.200	40	0.200	160	0.200	640	0.200

Table 3.2 Tape drive read/write rates and cost per drive.

Regardless of the use of commodity disks for data storage, it is clear that there will be one or two major tape technology transitions needed during the run due to the five year time period considered. Each transition will involve replacing tape drives, upgrading Enstore tape mover nodes, increasing network throughputs, and possibly copying old data to new media.

3.4 Robotics and Tape Drives

DØ has access to an ADIC AML/2 robot with LTO drives and an STK Powderhorn Silo with 9940 drives for data storage. Capacities for these devices are summarized in Table 3.3. As tape cartridge capacities increase, the option to copy old data to new media exists and has to be weighed against the cost for additional robotic storage, or the decision to "shelve" certain older data sets. As new robots or silos are added, floor space in Feynman Computing Center and the needed infrastructure must be identified and planned for. Our expectation is that a significant amount of processing will occur at remote centers, especially MC production and secondary data creation, and any reprocessing, and this will not be stored at Fermilab.

	Tape Slots/unit	Drive Slots/unit	Mounts /hour	K$/unit
STK-Powderhorn	5500	20		75
ADIC-AML/2	3500	20	150	300

	2002		2003		2005		2007		2009	
	TB	MB/s	TB	MB/s	TB	MB/s	TB	MB/s	TB	MB/s
STK	300	200	660	400	1320	800	2640	1600	5280	3200
ADIC/AML2	330	200	700	400	1400	800	2800	1600	5600	3200

Table 3.3 The current robot performance specifications are shown in the top table. The projected robot capacities as a function of time are shown in the lower table. The table assumes twenty drives per robot. For the ADIC/AML2, a "unit" is a single quadro-tower. DØ's AML2 robot has 3 such units.

To make estimates of tape and disk storage needs, we identified possible data tiers, with size per event. The total amount of storage needed is related to the number of events collected, and we assign a factor that represents the number of times an event is stored for each data tier. Each raw event is stored once, for example. We allow for some reprocessing storage in the estimate, and assume a large amount of derived data will be archived. The assumptions are shown in Table 3.4

		size		tape factor	disk factor
sizes	raw event	0.25	MB	1	0.001
	raw/RECO	0.5	MB	0.2	0.001
	data DST	0.15	MB	1.2	0.1
	data TMB	0.01	MB	2	1
	data root/derived	0.01	MB	8	0
	MC DØGstar	0.7	MB	0.1	0
	MC DØSim	0.3	MB	0	0
	MC DST	0.15	MB	1	0.2
	MC TMB	0.02	MB	3	0.5
	FAST MC	0.02	MB	2	0.5
	MC rootuple	0.02	MB	0	0

Table 3.4 shows the event sizes and stored data for tape and central analysis disk cache. The columns labeled "tape factor" and "disk factor" show the number of events on tape and disk, for each tier relative to raw data.

The tiers are defined as follows: raw/RECO is the data tier for which the raw data is kept with the reconstructed output. Such samples are useful for trigger and reconstruction studies and some types of physics analysis such as W mass determination, which will need more information than the DST can provide. The data summary tier (DST) is expected to have sufficient information to allow some limited re-reconstruction of high level physics objects. We assume that slightly more DST than raw data will be stored on tape to allow for some reprocessing. The thumbnail (also called the TMB or the micro-DST) is a physics summary format, and is presumed to be the starting point for the most user analysis. We assume that DST level reprocessing will produce additional copies of the TMBs which must be concurrently stored. We anticipate that most derived data sets will be subsets of the thumbnail, and based on Run I experience, we allow for a large amount of these sets to be stored on tape. The amount of the Monte Carlo tiers which must be stored require trade offs between tape costs and the ability to re-reconstruct and to re-run the trigger simulation and to simulate different instantaneous luminosities. Here we assume a generous amount of MC DST storage. The MC TMB will be twice as large as the comparable collider data tier because more information is stored. The disk storage percentages listed are used as a guide to determine the size of the disk cache on DØmino to support analysis. We assume that there is one primary TMB sample on disk on the central analysis system and the derived data sets are kept on physics group project disk. With these assumptions, the total storage needs are shown in Table 3.5.

	Run 2a 2 years	Run 2b 4 Years
Total Number of Event	1.58×10^9	6.31×10^9
TAPE data accumulation (TB)		
raw event	394.20	1971.00
raw/reprocessing	157.68	788.40
data DST	283.82	1419.12
data TMB	31.54	157.68
data rootuple	126.14	630.72
MC DØGstar	110.38	551.88
MC DØSim	0.00	0.00
MC DST	236.52	1182.60
MC TMB	94.61	473.04
FAST MC	63.07	315.36
MC rootuple	0.00	0.00
total storage (TB)	**1,498**	**7,490**
DISK data accumulation (TB)		
raw event	0.39	1.58
raw/reprocessing	0.79	3.15
data DST	23.65	94.61
data TMB	15.77	63.07
data rootuple	0.00	0.00
MC DØGstar	0.00	0.00
MC DØSim	0.00	0.00
MC DST	11.89	47.22
MC TMB	15.77	63.07
FAST MC	15.77	63.07
MC rootuple	0.00	0.00
total storage (TB)	**84**	**366**

Table 3.5 above shows the total data storage required for assumptions listed above for Run 2a and Run 2b

For Run 2a, DØ needs approximately 1.5 PB of robotic storage, assuming that the final data formats and data rates are achieved. SAM enables DØ to use both robots transparently to the applications accessing the data, such as the online logging, farm operations and user analysis. DØ has an option to buy a second STK robot, giving DØ over 2 PB of storage with the current generation of drives and media. However, it has not been demonstrated that the LTO tape drives/media are operationally reliable for large data sets, although test writing the RECO output to LTO tape drives are underway at this time. In addition, the next generation of 9940 tape drives will be available shortly for testing, and if that drive meets specifications, the two STK robots will be sufficient for the Run 2a needs, with the Monte Carlo data remaining in the AML2 robot. For the purposes of costing robot storage for the next three years, we have assumed that we will purchase the second silo in FY-2003.

The tape cost for the storage scenario outlined in Tables 3.4 and 3.5 is approximately $500,000 per year assuming that half of the data is stored in the STK robot with 9940b drives (in 2003 and 2004) and half the AML2 robot and the evolution of technology proceeds as assumed.

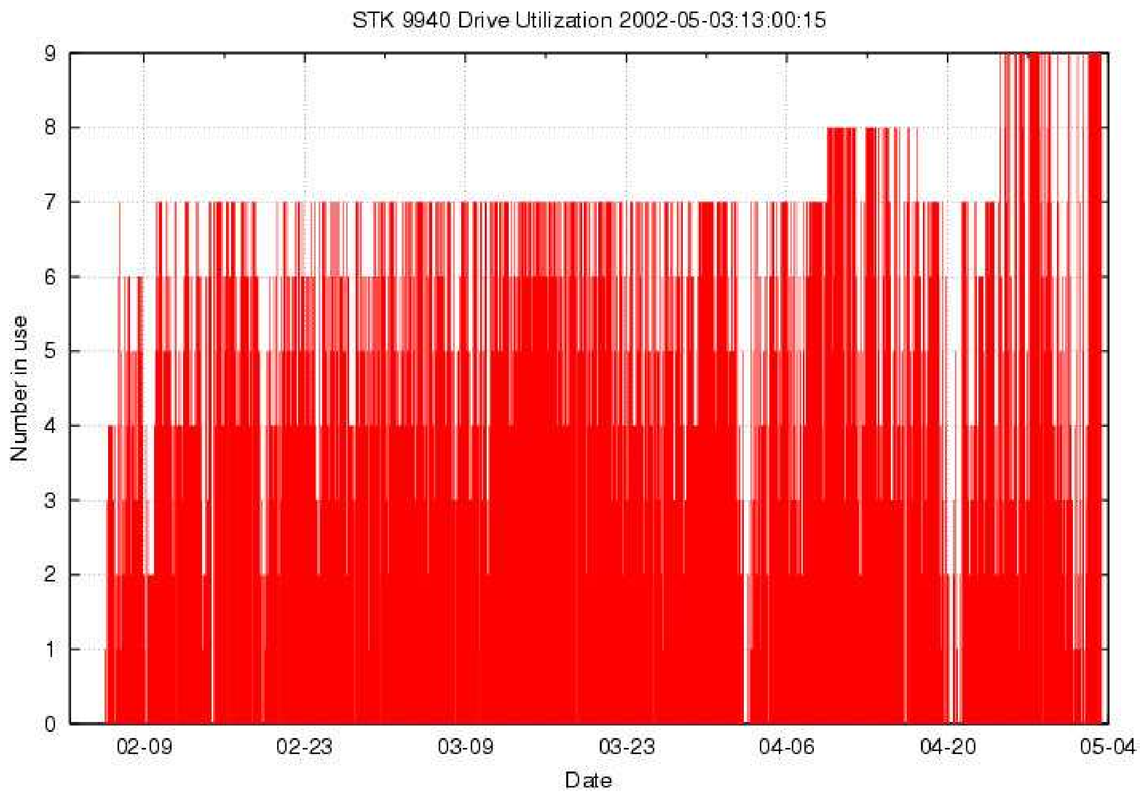


Figure3.4 shows the load on the nine existing 9940 drives.

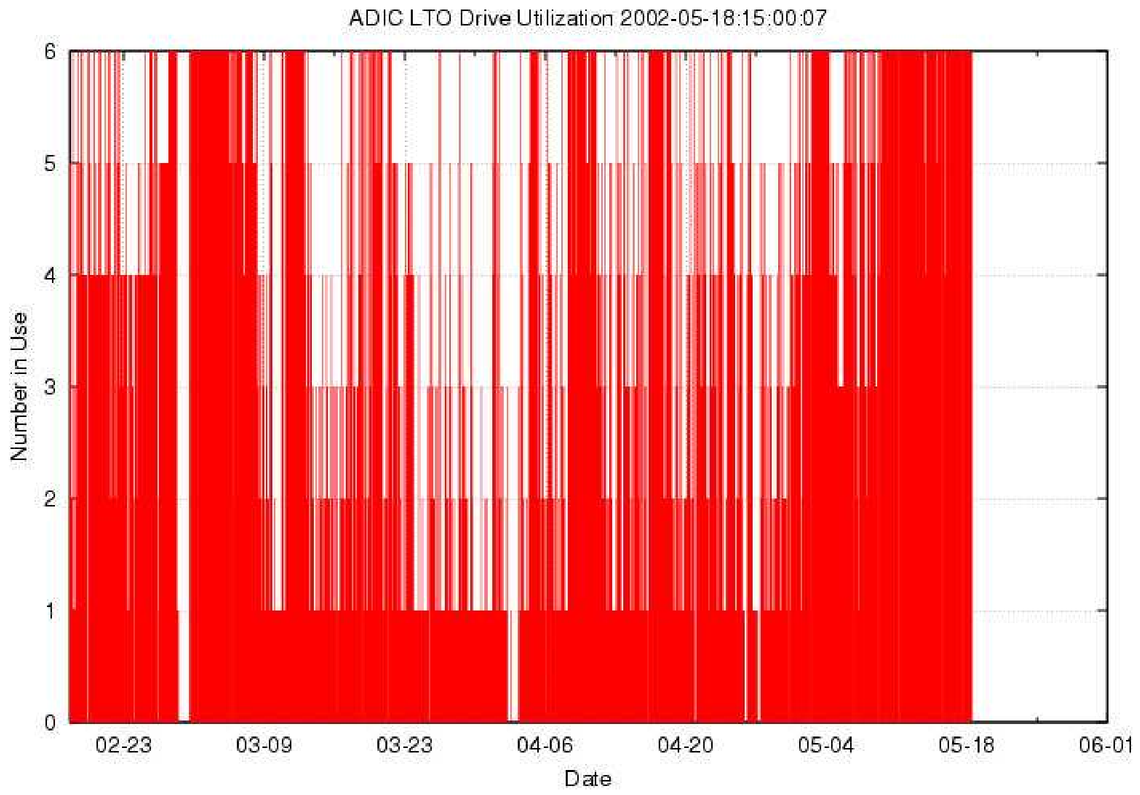


Figure 3.5 shows the load on the six existing LTO drives. The start of farm output going to LTO drives as a test starting on May 11 is clearly seen.

DØ currently owns nine 9940 tape drives and six LTO tape drives as shown in Figure 3.4 and 3.5. The drives are heavily in use, supporting online operations, reconstruction farm operations, Monte Carlo farm operations and user requests, including event picking. Given the accelerator duty cycle and the fact that DØ collects data at about half the eventual rate, these usage plots indicate that tape drives are a constraint on the capabilities of the current system. Event picking in particular is a costly operation as there are a number of fixed time operations incurred with mounting and dismounting the tapes. We estimate it takes 300 seconds to retrieve a single 1 Gbyte file (as is done for picked events) from tape with a 10 Mbyte/sec drive when all latencies are counted.

The computing model for DØ relies on tape for access to DSTs as only a small fraction of DSTs are disk resident at any one time. A bare needs estimate of the number of 10 Mbyte/sec drives needed for Run 2a operations follows

- 3 required to support online operations
- 3 required to support Farm operations
- 2 required to support incoming MC
- 8 to spool through the DST sample in 3 months in 2004
- 8 to support secondary analysis stations such as the regional centers

- 4 to support user requests.

Event picking will have to be administratively controlled as it could consume an infinite number of drives.

The bare estimate is 28 drives. STK 9940 drive and mover nodes costs $30K and the LTO drives and mover nodes cost $11K, so it would certainly be beneficial if the LTO drives meet our specifications. However for the purposes of costing the system, we assume that 9940 drives will have to be purchased. We assume that 15 drives will be purchased in 2003, and an additional 15 drives will be necessary in 2004 to support analysis activities.

For Run 2b, we have assumed the purchase of two STK robots and 20 drives per year, which should nominally meet our storage needs assuming future generations of drives and media are available in 2005. However, we will have more information on this in future years.

CHAPTER 4 – Computing Systems

This section describes the production and analysis systems for DØ, located at FCC, at DØ and worldwide.

4.1 Local Farms

The current DØ farm system consists of an SGI O2000 used as an I/O server node and 122 dual processor Intel systems used as worker nodes. The SGI node is used to buffer all output back to the Enstore storage system via SAM. The I/O node is an 8-processor system with 2GB of memory and 930GB of disk space. The I/O node also does merging of small files before sending them to the storage system. Two gigabit Ethernet interfaces and one 100Mb/s interface provide network connectivity for the I/O node. One gigabit interface is in the Enstore subnet and used only for routing files to offline storage. The other gigibit interface is in the farm subnet and is used only for buffering files from the worker nodes. The 100Mb/s interface is in the DØ offline subnet and used for all other outside connections to the system. All the interfaces are connected to a central Cisco 6509 switch.

The workers include forty 500MHz Pentium-III, fifty 750MHz Pentium-III, and thirty-two 1 GHz Pentium-III nodes. The total capacity of this system is approximately 80,000 SpecInt2000s. It is expected that this will be expanded in the next few months with 128 dual processor 2GHz nodes. This will bring the total capacity to near 250,000 SpecInt2000s. In the current configuration, twenty of the 500MHz nodes are dedicated to input file staging and global tracking test processing. All worker nodes are connected via 100Mb/s interfaces to the same Cisco 6509 switch as the I/O nodes.

The farms are run using a series of scripts that control job submission, execution, and monitoring. The scripts are written in c-shell, python and javascript. Job submission is done via a web interface which allows any user to make a processing request by specifying a dataset defined in the SAM DB, the required version of the production release to use for processing, what type of processing to do, and a suggested running priority. Farm operators may modify these requests and/or approve them for running from the same web page. The web page also displays the current approval, running and completion status of each request. All SAM datasets submitted for production are defined in such a way that only unprocessed files are included in the processing request.

The current version (p10.15.01) of the DØ reconstruction program requires approximately 13 CPU seconds per event on a 500MHz node to process current (March '02) data. An additional 5 seconds is used for the production of root-tuple files (which will be replaced by TMB files in July 2002) and another 1-2 seconds for merging of files. The total current expenditure is roughly 20 CPU seconds per event on a 500MHz processor. The exact CPU requirements for farm processing in 2b will depend both upon what processing is done (reco, root-tuple generation, merging, splitting, etc) and the complexity of the data itself. It would be unrealistic to presume that the current processing times will be maintained at the increased luminosity expected in 2b. At the risk of being overly optimistic we will assume that the total farm processing time per event in Run 2b can be held to 50 500MHz-CPU seconds per event. We will also assume any reprocessing as improved versions of the reconstruction program become available would be provided by the regional centers.

The farm systems will need to be operational as soon as the detector is capable of running at full rate and thus are not amenable to a phased-in purchase after the start of high rate running. As stated in the introduction, we assume that the full processing power needs to be online at the beginning of Run 2b in the fall of 2005. We can only expect 1.5 doublings of CPU power between now and the latest possible purchase of 2b farm systems. Consequently we will assume that 3GHz, 4GHz, and 6GHz Pentium-4 (P-4) systems will be available for purchase on '03, '04, and '05 respectively. Recent historical trends have seen new generation systems introduced at a roughly constant $2500 per dual processor unit.

The performance of a P-4 node can be estimated from the SpecInt ratings of currently available units and scaling the P-4 GHz rating. SpecInt2000 ratings for 500MHz P-III and 2.0GHz P-4 are 216 and 640 respectively. We will assume the SpecInt rating of 4.0GHz P-4 is 1280, or 5.9 times faster than a 500MHz P-III.

Folding together a 50 Hz average data rate, 50 seconds of 500MHz equivalent processing time per event and a presumed 70% efficiency for farm operations and reprocessing needs leads to an estimated need for ~600 P-4 nodes, assuming the purchase is spread over 3 fiscal years as shown in the spreadsheet. The total cost with this spending profile is ~$1M.

The I/O node will have to be scaled up to handle the increased load of 900 processors. From current experience we will estimate that a four-processor 4GHz P-4 system with 1-2TB of disk will service about 100 worker nodes. We estimate such a system will cost ~$25K in 2004. This will add $125K to the above farm total cost. It may be advantageous to distribute this I/O processing across a larger number of smaller nodes, but we assume the total cost for the necessary I/O functions to be about the same, i.e. an additional $25K for every 100 nodes.

The P-4 nodes will require an equal number of network connections. The 6509 switches currently in use can accommodate eight 48-port 100Mb boards. Thus at least two 6509 switches will be required for farm connection. This will require purchase of an additional 6509 switch and boards at an estimated cost of ~$100K. This cost is included in the networking cost estimate.

The existing farm nodes will not play a role in Run 2b. All farm systems are purchased with a standard 3-year warranty. After the warranty period is up it is presumed that nodes will not be repaired (except for trivial repairs) but simply decommissioned when they fail. All current nodes (except the 2 GHz nodes yet to be purchased) will be out of warranty by 2004. The 2 GHz nodes represent only about 15% of the above estimated need. We assume that about 10% of the base cost of the farm will be needed each year to replace out –of -warranty nodes lost through normal attrition, i.e. ~$150K per year.

Average Rate:	50
Farm Efficiency:	70%
Misc. Processing:	10%
Reprocessing:	0%
Cost/node:	2,500
I/O Cost/100 nodes	25,000

CPU	SpecI2000
3GHz	960
4GHz	1280
6GHz	1920
10GHz	3200
15GHz	4800

FY05 Target Spending Fraction:		20%		30%		50%		Total	
Execution	500MHz CPUs at	FY03, 3GHz Nodes		FY04, 4GHz Nodes		FY05, 6GHz Nodes		Target	
Time	Beginning of Run	No. Nodes	Cost	No. Nodes	Cost	No. Nodes	Cost	No. Nodes	Cost
50	3929	55	162,500	82	230,000	138	395,000	275	787,500
75	5893	82	230,000	124	360,000	207	592,500	413	1,182,500
100	7857	110	325,000	165	462,500	276	765,000	551	1,552,500

Table 4.1 shows farm purchase scenarios. Three RECO time/event possibilities are shown. At this time, we anticipate 50 sec/event for the published instantaneous luminosity guidance. This estimate does not explicitly call out the reprocessing contribution to be supplied by the institutions at the regional centers; however, the difference between 50 sec/event cost and the 75 sec/event cost gives a reasonable estimate of the cost of 50% reprocessing.

The purchase of nodes in 2003 and 2004 would allow a flat processing capability of roughly 40 Hz given the projected increase in RECO time as a function of luminosity as shown in Table 2.1.

4.2 Remote Farms

Over the next five years it is expected that Remote Production facilities will provide significant processing power for the DØ collaboration. It is envisaged that there will be three major tasks to be carried out by these sites:
a)Monte Carlo (MC) Production
b)Secondary reprocessing of the data.
c)CPU intensive user analysis jobs.
Each of these modes of operation will have different requirements. The simplest is MC production; it is essentially self-contained and does not require database access. Reconstruction of data for analyses requires database access and careful bookkeeping for specific binaries. User analyses jobs require that we can run a generic binary with any appropriate input.
The current remote production sites are located at Boston University, CCIN2P3, Lancaster University, Nikhef, Prague, and the University of Texas, Arlington. These sites provide 450 750 Mhz CPUs that are currently used for MC processing. Future sites are planned at Manchester University, Oklahoma, University College, Dublin and Karlsruhe.

The minimum required capability of these production facilities must be sufficient to meet all of the needs for MC production. For a standard 750 MHz CPU the time per event of the various stages of full plate level MC processing are:

Process/Time per Event	***Generation***	***Detector Simulation***	***Digitization***	***Reconstruction***	***Analyze***	***Total***
0.5 Events Overlaid, Plate Level Geant (sec/event)						
WW inclusive	0.8	280	20	19	4.5	325
Technirho	0.8	300	20	21	5	345

Table 4.2 shows current Monte Carlo chain generation time per event on a 500 MHz machine for plate level samples.

|At the digitization stage each event will be overlaid by a zero bias event (random sample of the detector) to simulate noise and additional soft interactions. Each event will be processed several times at different instantaneous luminosities. Because of this the average time per event for each plate level event will 550 seconds.

We would like to generate about half as many Monte Carlo events as we collect data events. Using the same assumptions as in the farm production profile, table 4.3 shows the cost and number of nodes which the regional centers would have to purchase to meet this need assuming a mix of plate level and fast simulation; 140 seconds corresponds to roughly one-quarter of the events using full simulation and the other three quarters using fast simulation.

Average Rate:	25
Farm Efficiency:	70%
Misc. Processing:	0%
Reprocessing:	0%
Cost/node:	2,500
I/O Cost/100 nodes	25,000

CPU	SpecI2000
3GHz	960
4GHz	1280
6GHz	1920
10GHz	3200
15GHz	4800

FY05 Target Spending Fraction:		20%		30%		50%		Total	
Execution Time	500MHz CPUs at Beginning of Run	FY03, 3GHz Nodes		FY04, 4GHz Nodes		FY05, 6GHz Nodes		Target	
		No. Nodes	Cost	No. Nodes	Cost	No. Nodes	Cost	No. Nodes	Cost
100	3571	60	172,500	67	190,000	74	210,000	201	572,500
140	5000	83	207,500	94	232,500	104	260,000	281	700,000
180	6429	107	267,500	121	300,000	134	332,500	362	900,000

Table 4.3 Resources needed at regional centers for Monte Carlo simulation

It is understood that the remote production facilities will in many cases be shared with other experiments and will not be able to upgrade operating systems purely to meet DØ software requirements.

4.3 DØmino and the Central Analysis Back-end

DØmino is an SGI Origin 2000 system comprising 176 R12000 (300 MHz) processors, with an attached data cache of ~ 30 TB fiber channel disk, with RAID disk for system needs and user home areas. DØmino's current role is to provide a centralized, stable, and uniform work environment providing interactive and batch services for on and off-site users. DØmino provides very high I/O and effective data transfer capacity into the petabyte-scale HSM and provides network capability unmatched in the industry. DØmino's current configuration has eight Gigabit Ethernets that are configured for interactive usage, data movement enabling the effective retrieval of data from sequential storage media (tape), data sharing and movement of these data to other secondary analysis systems (on and off site), and data movement into a cluster of Linux nodes for so-called back-end computing.

The Central Analysis Back-end (CAB) for DØmino is designed to augment the aging processors with commodity computing. To be effective, the configuration must take full advantage of the resources provided by the server machine while providing a simple interface for batch jobs. The basic configuration uses two dedicated network interfaces. One network provides the home and product file systems, while the other is dedicated to serving disk data and providing for data movement to local cache. The user interface to access these machines requires only the specification of the SAM station name during job submission. The underlying batch system is PBS, which allows nearly identical specifications as LSF (which is used for job submission on DØmino itself) and which can also be configured to allow job submission from desktop nodes. An estimation of the amount of analysis computing required can be found in Chapter 7.

With the analysis backend CAB, DØmino will meet our needs until at least 2005, as was originally planned. Options for replacing or upgrading DØmino are under discussion.

4.4 CLuEDØ Desktops

All Linux desktop machines at DØ are managed as part of the CLuEDØ cluster (Clustered Linux Environment at DØ). CLuEDØ is first and foremost a Linux desktop cluster. It is the primary network interface for its users. However, it is also a code development platform and processing farm and has become invaluable to the experiment in these capacities. This section describes each of these aspects of CLuEDØ from the standpoint of both basic functionality and technical design.

CLuEDØ currently consists of approximately 170 nodes located in each of the five buildings at the DØ site plus a small number in Wilson Hall and in the Feynman Computing Center. The design of the cluster can support continued growth of Linux desktop use at DØ throughout the 5-year planning period.

CLuEDØ is an institute-based cluster. Allocation of CLuEDØ resources is managed by a group of administrators contributed by member institutes. There are no physics groups in the cluster, only institute groups. System resources are allocated based on institute contributions to hardware and management of the system. CLuEDØ has been designed around the principle "many hands make light work". There are no fulltime system administrators in CLuEDØ. The cluster is run by a group of volunteers each giving approximately 0.2FTE to the project. Currently we have approximately 20 volunteers. This is just adequate for maintenance of the existing cluster. The successful running of the cluster also owes a great deal to the support of the DØ Task Force (DØ/CD). Technical support for home directories (including backup) and DØ code are important aspects of the cluster design.

In addition to acting as the primary desktop interface for many users and a fast code development platform, CLuEDØ is a powerful processing farm for the experiment. There are currently in excess of 200 available batch queues and both the number available and processing power per machine increase steadily over time. CLuEDØ uses PBS (Portable Batch System) to manage batch resources on the cluster. PBS was chosen because it is the most widely used, flexible, free batch system available for Linux. The configuration splits shares in the batch system by institute. Institute shares are weighted by their contribution to CLuEDØ in terms of hardware, admin manpower, etc.

In order to access DØ data on CLuEDØ, SAM must be made to deliver files to the desktops. This is currently working in a testing mode on the cluster but is expected to reach production status soon. The design is to transfer files either from tape or from the central analysis SAM cache to a large central CLuEDØ disk cache with good network access. From here the files can be transferred to small local caches configured on each desktop machine. An interface between SAM and PBS exists which allows SAM to influence the scheduling of jobs based on file delivery.

The technical and management structure design of CLuEDØ are both well suited for continued growth of Linux desktop use at DØ. There is no hard limit on how many machines can be accommodated by the model.

One issue of concern is administrative continuity. Manpower contributed primarily by postdocs and grad students is, by its very nature, unstable. The positions occupied by our administrators are generally short-term thus we maintain a large number of administrators trained to do many tasks. However, the cluster would benefit greatly from the hiring of an individual who could maintain at least 0.5FTE on CLuEDØ system management. Currently DØ does not employ any computing professionals directly involved in day-to-day linux desktop support. As CLuEDØ continues to grow to cover most desktop systems at the experiment it would be strongly in the interests of the experiment to have someone in place whose primary job is linux desktop support. As will be discussed in the analysis CPU section of this document, DØ is also building a CLuEDØ backend compute farm onsite at DØ. The new hire could be fully occupied with 50% load on each of these projects.

4.5 - Analysis CPUs (CLuB)

DØmino is a high-bandwidth machine capable of serving large datasets (10's of TB) to SGI or linux CPUs for processing. The CLuEDØ desktop linux cluster has some capacity to process small datasets (eg. 100GB). The DØ computing model allows for an intermediate level of processing on datasets of the order of 1Tb. CluB (CLuEDØ Backend) provides processing for intermediate datasets while also allowing seamless integration with the DØ desktop environment.

CluB is a farm of rack-mounted linux PCs being installed on the second floor of the DØ assembly building. It consists of two types of machines: 1) disk servers 2) CPU servers. The disk servers are dual processor machines with large disks based on IDE RAID arrays. Currently these machines support 1.2TB in a 4U case. The nodes are also dual processor in 1 or 2U cases. The machines are contributed by DØ member institutes, and resources are allocated based on contribution. A small "seed" of the cluster is being purchased by DØ/CD consisting of networking infrastructure and test disk and CPU nodes. Further CPU and disk servers will be purchased from the vendor directly by institutes and sent to Fermilab. DØ Computing and Software will also make a small yearly contribution to CluB.

CluB will not support interactive logins to the nodes but rather will provide access via the batch system (PBS). Data delivery will be on a private data network served by SAM to a large central cache and transferred to CPU nodes via rcp. Batch output will be stored on the disk servers and will be accessible by rcp/scp/ftp and via NFS from CLuEDØ nodes on the interactive network.

Individual CluB nodes will not be supported on a 24-hour basis. Daytime coverage will be provided primarily by shift workers chosen from a pool of administrators contributed by DØ institutes. The cluster would benefit greatly from the hire of an individual capable of maintaining 0.5FTE on CluB support. Since the role of coordinating installations, shifts, etc. will require more time than a single volunteer admin may be willing to contribute, this role is best filled by a computing professional.

4.3 Remote Analysis Computing

DØ has a strong commitment to remote analysis, and fully expects regional analysis centers to provide computing resources beyond the base computing resources located at FCC. Indeed, we anticipate that non-FNAL resources should provide approximately half of the analysis computing and all reprocessing. Following the model of other experiments, we imagine the most effective way to deploy such services and to assure adequate support personnel is to concentrate them in a number of regional centers. Of order 5-10 such centers would allow a critical mass at each. DØ is in the process of developing this model. Our current plans are ambitious, with the idea of serving out the thumbnail and derived data sets to the regional centers, and to support desktop analysis at all remote institutions. Remote contributions could include supplemental production capabilities beyond Monte Carlo generation. The regional analysis centers are expected to provide some of the following services:

- Code Distribution
- Batch Processing
- Data Delivery
- Data Reprocessing
- Database access
- MC Production and Processing
- MC Data Storage

Among the above listed services, data reprocessing and batch processing services would require significant computing power, which supports a model with some number of regional centers. A document describing this model is in preparation.

It is assumed that remote production jobs will make full use of the current DØ Grid project. Use of the Grid is not specific to remote analysis, but rather is a coherent part of the overall computing plan. The general principles for analysis at the regional centers include attempting to move the jobs to the data which require a grid queuing system to make this available at all sites. It is likely that we will need Mass Storage close to processing centers to store the results of production activities.

The Fermilab-resident component of DØ will have its own regional center in the form of the CluB system, described above.

CHAPTER 5 –Infrastructure

5.1 Networking needs at FCC and DØ

DØ is currently connected to the Feynman Computing Center via Gigabit Ethernet carried over three pairs of fibers and we are adding six more pairs. We assume these connections can be driven at full rate by the network hardware on each end giving us a total capacity of 9Gb/s bandwidth. This should be adequate to support the online logging needs (~80 Mb/sec peak rate), CLuEDØ/CluB, and interactive and NFS traffic between DØ and FCC.

The local network infrastructure at DØ will require some enhancements. The most straightforward way to increase the bandwidth to DØ users is to replace the current hubs with switches and Gigabit uplinks to the DAB Cisco 6509. There are currently 21 hubs in DAB. One could probably expect to cut this number in half when replacing the hubs with switches. That would imply installation of at a cost of ~$80K for 11 Catalyst 2948 type switches and an additional 16-port Gigabit card for the DAB 6509 switch. Additional Gigabit cards can be used to improve service to the satellite buildings. We expect to purchase an additional Cisco 6509 switch at DØ with several Gigabit cards to service CluB as well as an additional Cisco 6509 switch will be needed at FCC for the expected farm and central analysis expansion. The estimated cost of the switch chassis and blades is $100K. In addition, a second switch in DAB will need blades, estimated to cost $60K.

It is not yet possible to predict when 10Gigabit Ethernet will be viable option for increasing bandwidth on the Fermilab backbone. Endpoint connections for 10Gigabit Ethernet currently cost ~$30K each. The higher capacity of 10Gigabit Ethernet also requires that the associated network hardware will have to be upgraded. We presume this is effectively replacing the 6509 switches with the next generation of equipment. Replacement cost will be similar to original costs of these units, i.e. $100K. Full cost of converting to 10Gigabit Ethernet for the FCC to DØ link will likely be in the $200K range for all associated equipment.

It is presumed that the bulk of any data transferred from FCC to DØ area would go to the CLuB nodes and, hence, those nodes would account for most of the connectivity requirements at DØ. However, by 2006 the demands of the desktop systems might exceed the 100Mb/s connections that are currently available. If Gigabit Ethernet to the desktop is necessary, it would require major improvements to the wiring infrastructure in the DØ buildings as well as wholesale replacement of the network infrastructure. Such an upgrade of the desktop connectivity would cost approximately $400K.

5.2 Fermilab Connectivity to the Outside World

Fermilab currently has an OC3 (155Mb/s) connection to ESnet. This will be upgraded within the next year to an OC12 (622Mb/s) connection. However, this will not significantly improve the available bandwidth between Fermilab and most of DØ's collaborating institutions. This is true for two reasons. First, the ESnet backbone itself is currently only an OC12 connection. This connection might be upgraded to OC48 (2.45Gb/s) in roughly 12-18 months. Unfortunately, most of DØ collaborating institutions are not directly connected to Esnet. Hence improving our connection to ESnet will not significantly impact our connectivity to these institutions.

Most of DØ's collaborating institutions do have connections to networks that connect to the Chicago Startap. However, Fermilab does not currently have a direct connection to the Startap. There is an effort under way to provide such a connection. If it is possible to find available unused fiber between Fermilab and the Startap then an OC48 connection could be in place within about 1 year. Without such dark fiber this connection is probably several years away (~2006?). Possible use of fiber owned by ComEd or the CTA is being explored at this time.

A pessimistic scenario would be that Fermilab remains with an OC3 connection throughout this year and then goes to an OC12 connection to ESnet in 2003. That would remain the primary link until an OC48 connection existed to Startap sometime around 2006. An optimistic scenario would be that Fermilab goes to an OC12 connection to ESnet by the Fall of 2002. An OC48 connection to Startap could exist by early 2003. The OC48 connection could be upgraded to OC192 (10Gb/s) whenever funding permits.

As connections are shared by all groups at Fermilab, we expect to get about 1/3 of the available bandwidth. That places practical limitations on DØ's connectivity to our collaborators at about 6, 25, or 100MB/s for OC3, OC12, or OC48 connections respectively. OC48 connections will be necessary to support significant data reprocessing at the regional centers.

5.3 Databases

The offline databases will continue to be stored in the Oracle RDMS, hosted by the Computing Division and managed by the central database administration team. The database hardware and software infrastructure will continue to be upgraded as needed and in anticipation of planned database growth and use. We will continue to rely on a 3-tier application architecture, which makes use of a middle database server layer to isolate the user applications from the details of the database structure and interface, limits the concurrent number of users to the database, provides for common performance and functional enhancements such as caching, transformation of the queried data to a more useable form, and provides for easier management of the overall system. It is expected that the number and variety of database applications will increase as the experiment moves into an operational analysis phase, as well as current applications requiring maintenance and periodic upgrade. The overall database and associated application infrastructure will be enhanced to better support remote analysis as it becomes more widespread and the experiment relies more heavily on the remote institutions processing and analyzing the datasets. Some replication of databases to remote sites is anticipated – although at this time the scope and mechanisms are not worked through. As the DØ grid project proceeds, it can be anticipated that extensions to and modifications of the database infrastructure will be necessary.

5.4 Database Software Infrastructure

A transition from Oracle 8.1 to Oracle 9i will be necessary within the next couple of years. This will enable continued production support from Oracle and to allow us to take advantage of new features and performance in the product. Oracle is now deployed on Linux as well as Solaris and effort is expected to allow production level support on this platform. We will investigate the use of public domain databases for some offsite use – but the preferred model is to make use of Oracle significant offerings to support all database users in DØ.

Additional Oracle license will be required and are expected to cost $50K every other year (with 2002/2003 being an “on” year). Currently the Computing Division pays for the maintenance and support of Oracle and its layered products (designer, oem etc) .

5.5 Database Hardware

It is anticipated that the DØ database disk requirements will grow at the rate of 200-300GB/year - which together with the indexes and backup requirements implies a disk purchase of around 1TB/year ($30-50K) . This reflects only the disk needs of the offline production database. Other disk is required for test, mirror and development machines.

The load on the database server machines will increase with active user analysis and with the size of the dataset. It can be expected that more computing power is needed in the database server machines during 2003 and 2004. Alternative strategies will be explored - such as purchasing of more Solaris or moving to Linux host machines - but it is expected that the total costs will not vary greatly and will average around $60K/year.

Database growth in Run 2b is difficult to estimate but we assume it will not scale strictly with the data rate. The number of detector channels is approximately the same so calibration, trigger, and other configuration info scales only with time (duration of the run). The SAM events table, approximately one-half of the SAM storage, scales with event rate. The number of file records scales with the number of files, so we propose increasing the data file size to 5 Gbytes (compared to the current nominal 1Gbyte files). Other storage needs for SAM are small. Based on these estimates, we will need to purchase a new database system in 2005. This upgrade is expected to cost $300K and include a new expandable RAID array. Traditionally such upgrades have been accompanied by upgrades in database and layered product versions that then occur in parallel with the production services.

Application Name	Estimated size after 2 years
Offline Calibration Top Level	40MB
Offline CAL calibration	90GB
Offline SMT Calibration	80GB
Offline Muo Calibration(MSC,MDT,PDT)	30GB
CFT Offline Calibration	14GB
CPS Offline Calibration	2GB
FPS Offline Calibration	8GB
FPD Offline Calibration	In development
Offline luminosity and streams	200GB
L1,L2,L3 Trigger	2GB
SAM File and Event	700GB
Speakers' Bureau	80MB
Releases Request	100MB
VLPC Calibration	7GB
RCP	2GB
RUN_CTL_COND	105GB
	1.15TB

Table 5.1 Examples of the estimated size of some typical database applications are shown in the table.

We plan to support snapshots of the databases needed for analysis at non-FNAL sites to remove the single point of failure in the data access system. As a start, we plan to copy read-only data, such as calibration, run information, luminosity, and a subset of the SAM tables. The site that is hosting the replica would need to provide technical and operational support for a production system.

Upgrades to the database infrastructure will most likely be required as the merge of SAM with standard grid middleware proceeds. This can be done pragmatically and in stages with effort coming from the specific development and deployment projects.

The database, networking and other infrastructure costs are summarized in Table 5.2. Included in this table are the resources that are necessary to build the DØ code releases.

Infrastructure Costs	2003	2004	2005	2006	2007	***Total***
Databases:						
Server upgrades	$60,000	$60,000	$0	$25,000	$25,000	
non COTS disk and controllers	$60,000	$20,000	$10,000	$10,000	$10,000	
DB system replacement			$300,000			
Software	$50,000	$0	$50,000	$0	$50,000	
DB totals	$170,000	$80,000	$360,000	$35,000	$85,000	$730,000
Networking	$120,000	$120,000	$100,000	$500,000	$100,000	$940,000
Build Machines/web servers	$60,000	$60,000	$60,000	$60,000	$60,000	$300,000
dCache/datahandling servers	$50,000	$50,000	$50,000	$50,000	$50,000	$250,000
Total, fixed cost	$400,000	$310,000	$570,000	$645,000	$295,000	$2,220,000

Table 5.2 shows a summary of the estimated infrastructure costs.

CHAPTER 6 – Analysis Patterns And Needs Estimate

At this time, DØ does not have a large data sample, nor is the output of the reconstruction written in the final formats. The primary analysis efforts have been focused on gaining a basic understanding of detector performance. However, we can look at current access and analysis patterns as a guide to the eventual patterns. Most primary user analysis is done on the available high-level data tier—which is currently the root-tuple generated by the production reconstruction. Physics groups have coordinated efforts skimming through root tuple or reco output to cull samples of interesting events (generating "derived data sets") and to pick samples of raw data events for re-reconstruction studies. There have also been coordinated efforts for specialized reprocessing of small data sets for tracking studies and physics studies.

Extrapolating from these access patterns as well as the experience from Run 1, we will assume that small groups of individuals, coordinated by physics and analysis groups will generate derived data sets for more general use, using DØmino or CAB. Such data sets could include skims of the thumbnails to generate samples suitable for desktop analysis, skims of the DSTs for background studies or analysis for which the thumbnail does not contain sufficient information, or when some limited re-reconstruction is required. Similarly, physics or analysis groups should co-ordinate efforts to obtain large samples of picked events. To regulate the DST access, we expect to provide the "freight train", the process of having rotating DST samples on disk, with the goal of cycling through all DSTs within a few months. In this model, it is assumed that the bulk of the user analysis is done from derived data sets that were generated from the thumbnails, and that many of those data sets in general are small enough that desktop analysis is feasible. Larger data samples of approximately 1 TB can be accessed on CLuB and the other regional centers.

For planning purposes, we assume that all of the analysis computing needed to generate the derived data samples must to be provided by DØmino/CAB to have access to the large disk cache, and consequently is an FNAL supported resource. Generating derived data sets from the thumbnail is not likely to be computational intensive. A recent Electroweak physics group root-tuple skim averaged roughly 0.1 sec/event on DØmino's 300 MHz processors. Ideally, the generation of the derived data sets keeps pace with farm production, but when a problem is discovered, the entire derived data set might need to be regenerated on a fairly short timescale. Generation of the DST derived data sets, on the other hand, is likely to be computationally intensive.

Additional information to consider is that current analysis usage on DØmino and CLuEDØ combined is estimated to be the equivalent of 350 500 MHz processors. That is comparable in size to the current (incomplete) reconstruction production farm, which is sized to handle a steady data collection rate of 15 Hz. Scaling factors would have to be applied for the number of events to process and the number of analyzers and the types of analysis processing, but it is clear that the analysis computing needs are large.

Various analysis categories can be identified and each of these analysis classifications can be assigned one of three categories: a) high resource, but few users, b) medium resources with medium users, c) very low resources but many users.

- Group creation of derived data sets for primary physics analysis (high)
    - DSTs
    - Thumbnails
    - Pick events
- Individual creation of derived data sets (high)
- Background studies and efficiency determination that cannot be done on derived data sets (including) (medium)
    - Trigger studies
    - Generation of turn on curves
    - Detector performance studies
    - Optimization of reconstruction algorithms such as b-tagging
    - Determination of mis-identification probabilities for physics objects
- MC studies
    - Generation of Monte Carlo test samples (high)
    - Generation of fast MC test samples (PMCS) (low)
    - Trigger simulations studies for efficiencies and tuning trigger conditions and algorithms (medium)
    - Reconstruction studies for efficiencies and algorithm development (medium)
- End level user analysis on derived data sets (low)

We can assume that most of the high resource work takes place at the physics group level and has a relatively long lead time over a large amount of data. The medium jobs take place at an analysis topic level over a smaller amount of data, and the user level data takes place on a very small data sample. This leads to an estimate (rounded) of 4 THz for analysis CPU for the Run 2a data sample. In 2002, we are purchasing the first CAB nodes, which will amount to 0.3 THz in analysis computing. That purchase is not included in the cost estimate.

	Jobs	Data Set (%)	Duration	Processing Time(500MHz)
High	6	30%	12 weeks	5 sec/event
Medium	50	10%	4 weeks	1 sec/event
Low	150	1%	1 week	0.1 sec/event

Table 6.1 shows a scenario for analysis usage based on consideration of known types of analysis.

The series of tables shown below shows the cost estimate for analysis CPU using the above assumptions shown over the entire anticipated sample. The first table in the series shows some of the input parameters. The total number of events is taken from the initial assumptions about the rate. The cost estimate table is split between 2004 and 2005 to separate Run 2a and Run 2b.

Total Data Sample:	7.89E+09		
Offline Efficiency:	70%		
Contingency:	0%		
Analysis Type:	Short	Medium	Long
Time/event:	0.1	1.0	5.0
% of Data Sample:	1%	10%	30%
Duration (Days):	7	30	90
Number of Jobs:	150	50	6

Total Event Fraction:		10%		10%	
Analysis Type	THz CPUs at End of Run	FY03, 3GHz Nodes		FY04, 4GHz Nodes	
		No. Nodes	Cost	No. Nodes	Cost
Short	1.40	31	77,500	23	57,500
Medium	10.87	244	610,000	183	457,500
Long	6.52	146	365,000	110	275,000
Total:	18.79	421	1,152,500	316	865,000

20%		30%		30%		Total	
FY05, 6GHz Nodes		FY06, 10GHz Nodes		FY07, 15GHz Nodes		Target	
No. Nodes	Cost	No. Nodes	Cost	No. Nodes	Cost	No. Nodes	Cost
31	77,500	28	70,000	18	45,000	131	327,500
244	610,000	220	550,000	146	365,000	1037	2,592,500
146	365,000	132	330,000	88	220,000	622	1,555,000
421	1,152,500	380	1,025,000	252	680,000	1790	4,875,000

Tables 6.2-6.4 show a profile for the acquisition of analysis computing sufficient to analyze the entire Run 2 data set. This excludes any desktop resources and does not include the retirement of older machines, factors that are likely to offset each other. We assume the same processing efficiency factors as for farm production, although that is likely to be an overestimate in the case of an analysis system. For that reason, we do not include the 2002 analysis computing purchase in this estimate.

CHAPTER 7 – Budget Summary

This chapter summarizes the projected equipment spending, and provides some information about the assumptions used to make the projections. The Laboratory has provided guidance of $2M per year.

7.1- Computing Systems

As described, DØmino is meeting DØ's needs quite well for access to large disk cache and as a network interface. The processors on DØmino are becoming obsolete (300 MHz processor), however with the analysis backend CAB, DØmino will meet our needs until at least 2005, as was originally planned. DØmino was commissioned in 1999 with an anticipated five-year service life. Consideration must therefore be given to replacing or upgrading DØmino on the timescale of 2005. Replacing DØmino in kind with another SMP machine of similar scale with fewer, but more current processors would cost approximately $2M. This is cost prohibitive, and to stay within the guidance, would require scaling back the rest of the system to an unacceptable level. However, there is as yet no completely demonstrated alternative using all commodity components that can completely replace DØmino's functions. At this time, our strategy is to explore commodity solutions by deploying CLuB and CAB. After gaining a year's experience with these systems, we will be in a better position to design an all-commodity system for our analysis patterns, and to understand the costs of such a system in terms of support, reliability, and data handling. In this document, we make no assumptions about mechanism for replacing DØmino, but allocate $200K per year in Run 2b to provide disk cache and servers. The purchase of commodity analysis CPU is already in the model and is directly estimated.

The farm and analysis computing needs were outlined in the respective sections. Note the analysis estimate includes an assumption of institution contributions to meet the estimated needs. The production farm needs are estimated for 50 Hz data collection rate with a reconstruction time of 50 sec/event, with any reprocessing occurring at the regional centers.

An additional person in the DØ task force could assist in the administration of CLuEDØ and provide expertise to bring up CluB. As CluB is a possible test bed for a DØmino replace, stable and continuous administration in addition the institution volunteers is highly desirable.

We also need a user backup system. We estimate this would cost $100K if existing robotic storage, such as part of the ADIC/AML2 robot ,can be used.

7.2 - Robotic Storage and Disk Cache

The estimated roadmap for tape and disk storage is detailed in Chapter 3. To summarize, we plan to purchase a second STK silo and populate it with 15 drives in 2003 and purchase an additional 15 drives in 2004. For Run 2b, we assume the purchase of two robot and drives per year.

In Run 2a, we plan to add an additional 15 TB of disk per year to DØmino. This would primarily be added to the SAM cache and used for TMB storage and DST operations. Institutions can contribute project disk space as well as additional disk for CLuB. For Run 2b, the issue of disk purchases is tied up with the replacement of DØmino as all fiber channel disk would have to be replaced with inexpensive IDE disk and the means to serve that disk. We assume $200K per year for that need in Run 2b.

As a note, mathematical models such as queuing simulations for computer systems would be an excellent tool for understanding scaling issues in commodity based systems and to guide the understanding how best to allocate resources, and to anticipate hot spots. The DØ computing model is complicated, and tradeoffs must be made within fixed resources. Mathematical models have long been used in trigger/DAQ design and we encourage the Computing Division to develop and support such modeling tools for computing systems.

7.3 - Infrastructure costs

The costs associated with the databases are detailed in Chapter 5, and include the cost of database machines, disks and controllers, backups and software. The networking costs are also detailed in Chapter 5, and include expanded links between the DAB, the trailers and Outback, DAB and FCC, and additional switches for DAB and the farms. In Run 2b, the upgrades can be substantial, including a 10 Gb backbone from FCC to DØ, and Gb to the desk tops. There are additional costs to support the releases with Linux build machines and disks, to provide web servers, and to supply servers for small special purpose SAM stations and for dCache machines. These costs are summarized in Table 5.2.

Table 7.1 shows the projected spending from 2003 to 2007. This is an overall estimate that assumes substantial contributions from DØ institutions. A project disk backup system is included in the 2003 estimate.

DØ Total Cost Estimate (assuming institution contributions)						Total
	2003	2004	2005	2006	2007	***2003-2007***
Infrastructure Costs	$400,000	$310,000	$570,000	$645,000	$295,000	$2,220,000
Analysis Including Institution Contributions	$1,152,500	$865,000	$1,152,500	$1,025,000	$680,000	$4,875,000
FNAL CLuB Contribution	$50,000	$50,000	$50,000	$50,000	$50,000	$250,000
Reconstruction	$225,000	$325,000	$575,000	$150,000	$200,000	$1,475,000
Disk cache	$150,000	$100,000	$200,000	$200,000	$200,000	$850,000
Robotic storage	$75,000	$0	$150,000	$150,000	$150,000	$525,000
Tape drives	$450,000	$450,000	$300,000	$600,000	$600,000	$2,400,000
Backup facility	$100,000					
Sum	$2,602,500	$2,150,000	$2,997,500	$2,820,000	$2,175,000	$12,745,000

Table 7.1 shows the overall DØ Total cost (excluding MC production)

Table 7.2 shows the base level computing equipment that must be purchased by FNAL and installed in FCC, subtracting out the proposed institution contributions. That base level of functionality includes the infrastructure costs, robot storage and tape drives, initial reconstruction, and a base level of analysis computing, which would first be targeted at generating derived data sets. Some approximate spending numbers for 2002 are supplied as a guide.

DØ institutions are expected to purchase approximately half of the analysis nodes and to supply any reprocessing capability at the regional centers, costing an estimated $2.5M over 5 years as well as supplying the equipment for the MC processing (an estimated $0.7M over 3 years) and any mass storage needed to support these activities. In such a model where substantial amounts of raw or reconstructed data will have to be sent to the remote centers, excellent FNAL connectivity to the outside world is critical for success.

DØ Cost Estimate, FNAL contributions							Total
	2002	2003	2004	2005	2006	2007	***2003-2007***
Infrastructure Costs	$400,000	$400,000	$310,000	$570,000	$645,000	$295,000	$2,220,000
Analysis CPU	$400,000	$635,000	$462,500	$412,500	$610,000	$597,500	$2,717,500
FNAL CLuB Contribution	$30,000	$50,000	$50,000	$50,000	$50,000	$50,000	$250,000
Reconstruction	$400,000	$162,500	$230,000	$395,000	$150,000	$200,000	$1,137,500
Disk cache	$0	$150,000	$150,000	$200,000	$200,000	$200,000	$850,000
Robotic storage	$400,000	$75,000	$0	$150,000	$150,000	$150,000	$525,000
Tape drives	$200,000	$450,000	$450,000	$300,000	$600,000	$600,000	$2,400,000
DØmino Memory	$150,000						$0
Backup facility		$100,000					
Sum	$1,980,000	$2,022,500	$1,652,500	$2,077,500	$2,405,000	$2,092,500	$10,250,000

Table 7.2 shows the FNAL equipment contribution to DØ computing and software.

7.4 - Conclusions

In conclusion, we present the equipment funding plans for DØ for the next 5 years. We intend to continue the basic model, scaling to meet anticipated needs. We find that the lab guidance will provide an adequate basic level of functionality; however, meeting all anticipated needs will require external contributions from DØ institutions in the form of Regional Analysis Centers, one of which, CluB, will be located at DØ. DØ is currently developing a detailed plan for Regional Analysis Centers, which will require excellent network connectivity.

Bibliography

[1] A. Baranovski, et. al, “SAM Managed Cache and Processing for Clusters in a Worldwide System” , to be presented at Cluster 2002, September 2002 in Chicago, IL, and published in the Conference proceedings.
[2] The Globus Project, http://www.globus.org
[3] The Condor project home page, http://www.cs.wisc.edu/condor/.
[4] Storage Resource Manager, http://sdm.lbl.gov/srm
[5] The Enstore home page, http://www-isd.fnal.gov/enstore
[6] FBS is the Fermilab Batch System, http://www-isd.fnal.gov/fbsng