



DØ Computing Model
February 16, 2004

This document presents an updated computing plan for the DØ experiment covering FY2004-2006 and includes scope and cost estimates for hardware. Our approach in planning and costing for computing has been to estimate the total requirements as if all hardware were placed at Fermilab. Remote facilities that are made available to DØ then receive credit to DØ's Operations Common Fund in reflection of the savings that they generate. The total costs quoted in this document therefore do not reflect either the expected actual cost to Fermilab or the expected actual cost at the remote installation, which may be subject to many local factors. They reflect the cost of a "virtual center" at Fermilab that will never actually exist, but if it did, would be able to carry out all of our computing needs locally.

The cost estimates in this document reflect several changes relative to the version presented in September 2003. We have seen improved performance in the most recent versions of the reconstruction, decided to expand the thumbnail data tier to include tracking information while proposing to drop the DST format for collider data, and have added the concept of Common Sample reprocessing on the thumbnail. In addition, the 2004 DØ experiment planning includes a full reprocessing of the data from the raw tier in a six month time period. This knowledge has been folded into our cost estimates.

Originally we used the laboratory's luminosity profile from Associate Director Steve Holmes' January 2002 talk to HEPAP as an input to our planning. For this update, we have based projections mainly on our experience so far in Run II. The number of events we are writing to tape is essentially independent of accelerator luminosity, though we have factored in the expected increase in the complexity of the events as the instantaneous luminosity increases.

² Running at 132 nanosecond crossing time is no longer within the scope of the Tevatron program.

Reco timing

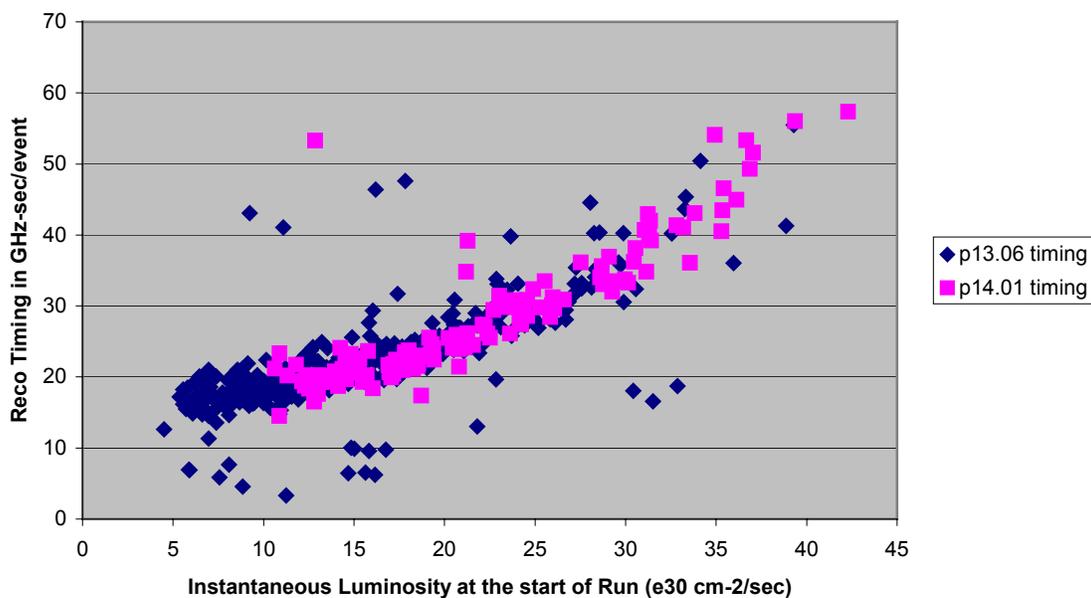


Figure 1 shows the measured reconstruction time as a function of instantaneous luminosity. The blue points show the time for p13 and the pink points show the time for p14.03. As can be seen, the timing and shape for these two versions is the same.

p14.06 RECO time vs Luminosity

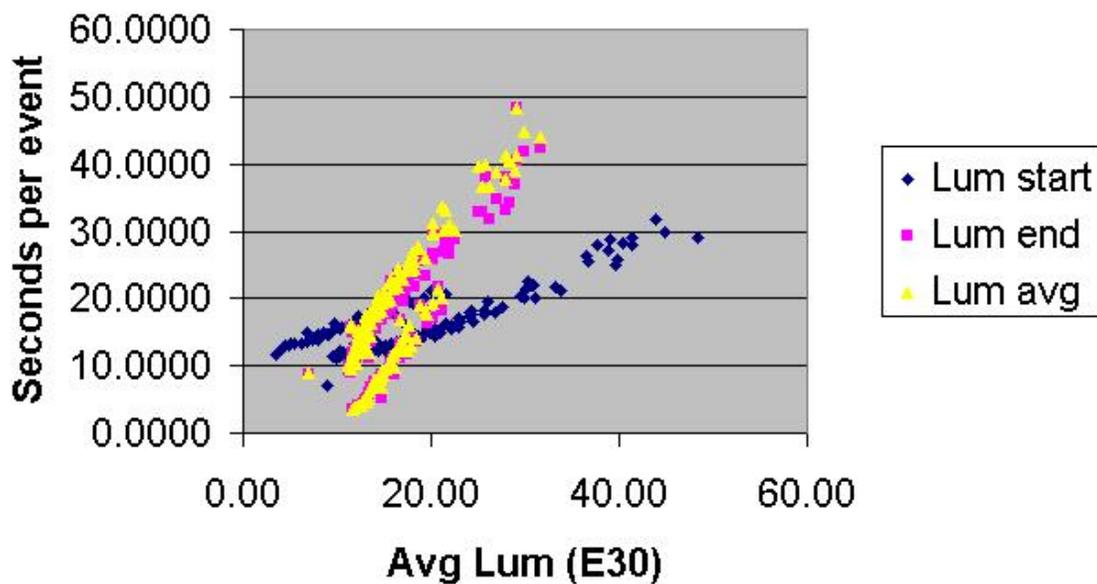


Figure 2 shows the measured reconstruction time as a function of instantaneous luminosity, shown for the initial, average and ending luminosity. The blue points show the timing as the function of the starting luminosity, the pink points as the function of the ending luminosity, and yellow as a function of the average luminosity. The bimodal behavior is believed to be the result of different prescale sets. The reconstruction estimates have been adjusted based on the improved performance of p14.06 relative to p14.03.

Based on the p14.06 timing numbers and assuming that the luminosity profile will not change significantly in 2004, we take 25 GHz*sec per event as the average reconstruction time in 2004. In 2005, we assume an average of 40 GHz*sec/event, and in 2006, 60 GHz*sec/event. Figures 1 & 2 are provided as justification for the change to the assumptions about Reco performance.

data assumptions

rates	average event rate	16	Hz		
	raw data rate	5	MB/s		
	Geant MC rate	1.6	Hz		
		size		tape factor	disk factor
sizes	raw event	0.25	MB	1	0.01
	raw/RECO	0.5	MB	0.2	0.01
	data DST	0.15	MB	0	0
	data TMB	0.05	MB	6	2
	data root/derived	0.04	MB	9	1.5
	MC D0Gstar	0.7	MB	0.1	0
	MC D0Sim	0.3	MB	0	0
	MC DST	0.3	MB	1	0
	MC TMB	0.05	MB	3	1
	PMCS MC	0.02	MB	2	0.5
	MC rootuple	0.02	MB	0	0

Table 1 Event size and stored data for tape and central analysis disk cache is shown. The columns labeled “tape factor” and “disk factor” show the fraction of events on tape and disk for each tier relative to raw data. The above tape and disk factors should be taken as representative—different assumptions apply the FNAL virtual center, the FNAL realized center, and different remote centers. These tables are directly from the planning spreadsheet, which has some artificial distinctions between formats.

The tiers are raw/RECO (RAW), DST and thumbnail (TMB). The raw/RECO (RAW) data tier includes the raw data and the reconstructed output. These samples are useful for trigger and reconstruction studies and those analyses which need more information than the DST provides. The thumbnail is a physics summary format, and is presumed to be the starting point for most physics analyses. An extension to the TMB will include hit information, allowing for limited reprocessing, and the DST tier will be phased out for collider data starting in p17. The expansion of the TMB will lead to an event size of 50 Kbytes per event. The current version of the TMB contains extensive calorimeter information, and the Common Samples group has run a reprocessing pass over the p14 data set. We thus allow for two disk resident sets of TMBs. These changes affect the number of file servers needed for central analysis.

The total disk storage corresponds to about 115 TB/year at FNAL. 30% was added for contingency, and 10% (about 12 TB) set aside for dCache. We assume that the current cost of disk servers is \$20,000 for 5 TB of disk and assume a doubling every 18 months.

File Server Cost Estimate

cost/fileserver	10,000
Network cost/16 FS	10,000
Contingency	40%

Year	Capacity(TB)
2003	2.5
2004	3.5
2005	5.5
2006	8.7

Data Volume	FY03		FY04		FY05		FY06	
	No. FS	Cost						
114.79	64	690,000	45	480,000	29	310,000	18	200,000

Table 2 shows the cost for analysis and dCache file servers. The contingency corresponds to 30% on the disk estimate and 10% to account for the dCache servers.

Cost estimates for production systems are shown in Figure 3. At this time, we anticipate 25 sec/event for the 2004 instantaneous luminosity guidance and assume 40 GHz-sec/event in 2005 and 60 GHz-sec/event in 2006. We assume that the local FNAL farm efficiency is 70%, and that the online rate averages 16 Hz, with a peak rate of 50 Hz to tape, and 30% combined accelerator duty factor and DØ data collecting efficiency. The average of collection rate of 16 Hz is in agreement with the measured collection rate over the past six months. We assume 3 GHz processors in 2004, with 6 GHz processors available in 2006, with a fixed price per machine (dual processors) of \$2200, and an I/O cost of \$25,000 assigned to each 100 nodes. We take into account the existing plant and plan to retire nodes after 3 years when they leave warranty. For reprocessing we assume that the reprocessing centers are 50% efficient (as it is much harder to achieve high efficiency on shared resources). A full reprocessing of the data set from raw using p17 was determined to be essential to meet the physics needs of the experiment. The 180 day duration for such a reprocessing was calculated assuming remote and extensive use of FNAL resources during the Summer '04 shutdown. For the purposes of planning, we assume that the '04 plan is representative of the reprocessing needs for the future.

For simulation, our goal is to generate about three Monte Carlo events for every four collider events collected. Using the same assumptions as in the farm production profile, Table 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 (detailed) and fast simulation. In the detailed simulation, each event is overlaid at the digitization stage by with zero-bias events (random sample of the detector) to simulate noise and additional soft interactions. An average of 170 seconds corresponds to roughly one-quarter of the events using full simulation and the other three quarters using fast simulation.

The experience from the Common Sample group's preparation from the fall reprocessing pass shows that each event takes about 0.75 GHz*sec/event, and that the samples need to be turned around quickly. At this time, the Common Samples group anticipates running the TMB reprocessing continuously, but with producing complete data samples in two months.

Primary Reconstruction Cost Estimate				SUM
Year	2004	2005	2006	
Reco time	25	40	60	
Required CPU	686	1097	1646	
Existing system	670	192	645	
Nodes to purchase	4	168	124	296
Cost	\$8,507	\$395,414	\$297,963	\$ 701,883
#Nodes at FNAL	360	268	296	924

Reconstruction Cost Estimate				
Year	2004	2005	2006	
reco time	25	40	60	
duration	120	120	120	
fraction	100%	100%	100%	
Rate	48.67	48.67	48.67	
Farm eff.	50%	50%	50%	
#nodes	603	724	724	2051
CPU required (GHz)	2433	3893	5840	12167
	\$ 1,477,181	\$ 1,767,617	\$ 1,767,617	\$ 5,012,414

Monte Carlo Cost Estimate

Year	2004	2005	2006	
MC time	170	170	170	
duration	365	365	365	
fraction	15%	5%	5%	
Rate	2.40	0.80	0.80	
Farm eff.	70%	70%	70%	
#nodes	145	36	36	217
CPU required (GHz)	583	194	194	971
	\$ 342,900	\$ 79,475	\$ 52,983	\$ 475,358

TMB Reconstruction Cost Estimate

Year	2004	2005	2006	
reco time	0.75	0.75	0.75	
duration	60	60	60	
fraction	200%	100%	100%	
Rate	194.67	97.33	97.33	
Farm eff.	70%	70%	70%	
#nodes	52	19	13	84
CPU required (GHz)	209	104	104	417
	113,758	\$ 42,659	\$ 28,440	\$ 184,857

Table 3 Resources needed for production, including primary processing, reprocessing, MC production, and TMB reprocessing. The reprocessing of MC events is included in the estimate.

For planning purposes, we used the 2003 experience on CAB to estimate our analysis needs. During the current analysis period in preparation for the 2004 winter conferences, the available analysis CPU was barely adequate. However, there were a number of competing effects. The expansion of CAB was delayed due to problems with the vendor. The TMB reprocessing was run by the Common Samples Group on CAB in Nov-Dec, leaving very little time for users to do analysis in time for the winter conferences, and thus creating an artificially tight peak demand. The Common Sample Group samples did not contain all information needed by the Top group, which lead to some replication of the samples. Additionally, some of the user applications are known to use excessive amounts of memory, causing inefficient use of CPU, especially on the older CAB nodes. At this time, we do not make any major adjustments to the analysis CPU estimate, but do assume that the TMB reconstruction will either require additional analysis computing or will handled as a production activity either at FNAL or at a remote center.

In addition to the costs outlined above, there are infrastructure costs associated with running the experiment. The primary sources of these are for database machines, disk and servers, networking, miscellaneous machines such as those used to build the DØ code base, and web servers. We also often have to pay for other infrastructure costs such as purchasing raid arrays for Enstore. Table 4 shows our estimated infrastructure costs.

	2004	2005	2006
Databases			
servers	\$30,000	\$30,000	\$30,000
disk	\$30,000	\$30,000	\$30,000
Networking	\$120,000	\$80,000	\$100,000
Machines	\$60,000	\$60,000	\$60,000
Totals	\$240,000	\$200,000	\$220,000

Table 4 Cost estimate for infrastructure.

	2003	2004	2005	2006	SUM
FNAL Analysis CPU	\$505,400	\$339,000	\$522,000	\$337,000	\$1,198,000
Primary Reconstruction	\$200,000	\$8,507	\$395,414	\$297,963	\$701,883
Re-Reco (machine cost)	\$611,128	\$1,477,181	\$1,767,617	\$1,767,617	\$5,012,414
Re-Reco (value)	\$152,782	\$738,590	\$883,808	\$883,808	\$2,506,207
Monte Carlo	\$140,392	\$342,900	\$79,475	\$52,983	\$475,358
TMB Reconstruction		\$113,758	\$42,659	\$28,440	\$184,857
File Servers/disk	\$262,000	\$480,000	\$310,000	\$200,000	\$990,000
Mass Storage	\$280,000	\$230,000	\$100,000	\$500,000	\$830,000
Remote Analysis					
Infrastructure	\$244,000	\$290,000	\$210,000	\$230,000	\$730,000
FNAL Total	\$1,491,400	\$1,347,507	\$1,537,414	\$1,564,963	\$4,449,883
Virtual Center Total		\$2,542,755	\$2,543,357	\$2,530,194	\$7,616,306
Virtual Center (2003)		\$2,404,000	\$1,995,500	\$1,922,000	\$6,321,500

	2003	2004	2005
FNAL Analysis CPU	\$505,400	\$339,000	\$522,000
Primary			

Reconstruction <i>Re-Reco (machine cost)</i>	\$611,128	\$1,477,181	\$1,767,617
Re-Reco (value)	\$152,782	\$738,590	\$883,808
Monte Carlo	\$317,016	\$342,900	\$79,475
TMB Reconstruction		\$113,758	\$42,659
File Servers/disk	\$262,000	\$600,000	\$380,000
Mass Storage	\$280,000	\$230,000	\$100,000
<i>Remote Analysis</i>			
Infrastructure	\$244,000	\$290,000	\$210,000
FNAL Total	\$1,491,400	\$1,467,507	\$1,607,414
Virtual Center Total		\$2,662,755	\$2,613,357
Virtual Center (2003 estimate)		\$2,404,000	\$1,995,500

Table 5 Final cost estimate for FNAL and the virtual center. The FNAL spending for 2003 is also shown, as is the estimated value for 2003 for reprocessing and MC generation. For comparison, the estimated cost of the virtual center using the September 2003 assumptions is shown.

Table 5 shows the total estimates for 2004-2006. The FNAL total includes the base infrastructure, primary reprocessing and the central analysis facility, including fileservers and compute nodes; these items are denoted in yellow. The Virtual center total is the sum of the FNAL cost plus the reprocessing costs (including TMB reprocessing which is called out separately), and the MC production cost. Those items are highlighted in violet, as is the total. Desktop analysis resources (such as those in use at IN2P3 and GridKa Centers) are not counted as part of this version of the planning, and are left blank. For the purposes of comparison, the virtual center estimate obtained during the 2003 planning exercise is shown in teal. The primary changes are the improved performance of the reconstruction (decreasing costs), the assumption of reprocessing 100% of the data (increasing costs), the duration of the reprocessing, and the addition of the TMB reprocessing (increasing costs). The duration of the reprocessing has several effects on the cost estimates. The cost of the hardware is calculated based on the reconstruction time, duration and efficiency of the reprocessing, however, the value calculated to the experiment is based on the time that hardware is in use, on the assumption that there will be other consumers of the hardware when available. In the 2003 estimate, we assumed

that the duration of the reprocessing would be 90 days—and that we could use the MC hardware for reprocessing. However, with a six month turn-around in the reprocessing, we no longer assume that the MC machines can be redirected. In addition, our experience during the fall 2003 reprocessing was that the physics groups expect MC production at a constant level.

The change in the Reconstruction time estimates combined with the expanded TMB leads a different allocation of spending in 2004 for farm nodes relative to worker nodes, and leads to increased costs in 2005 relative to previous estimates, in part because the current farm can keep up with the expected data rate and reco time, leading to relatively more spending in 2005 as the legacy nodes are decommissioned while processing time increases due to increased luminosity.

Also shown in Table 5 is the estimated value to the experiment of the Remote contributions to the reprocessing of 100M events, and the production of 25M MC events. This value is calculated as if the hardware were purchased in late 2003. The total value is estimated to be \$300,000, again, making an assumption that the value of the computing is related to the amount of time it is in use, and that personnel costs are not included. For 2003, as several facilities were “christened” during the reprocessing, a case could be made that the value is underestimated. In addition, one could make a case, particularly for the MC facilities, one should use 2002 as the benchmark year for estimating the value as the hardware had to be available in January 2003.