THE DØ VIRTUAL CENTER AND PLANNING FOR LARGE SCALE COMPUTING*

Amber Boehnlein, FNAL, Batavia, IL 60510, USA

Abstract

The D0 experiment relies on large scale computing systems to achieve her physics goals. As the experiment lifetime spans, multiple generations of computing hardware, it is fundamental to make projective models in to use available resources to meet the anticipated needs. In addition, computing resources can be supplied as in-kind contributions by collaborating institutions and countries, however, such resources typically require scheduling, thus adding another dimension for planning. In addition, to avoid over-subscription of the resources, the experiment has to be educated on the limitations and trade-offs for various computing activities to enable the management to prioritize. We present the metrics and mechanisms used for planning and discuss the uncertainties and unknowns, as well as some of the mechanisms for communicating the resource load to the stakeholders.

INTRODUCTION

In order to correctly account for in-kind contributions of remote computing, D0 uses the concept of a Virtual Center in which all of the costs are estimated as if the computing were located at solely at FNAL. In contrast to other such models in common use, D0 accounts for contributions based on a percentage contribution to a task rather than strictly on money spend on hardware. This gives incentive for achieving the maximum efficiency of the systems as well as encouraging active participation in the computing model by collaborating institutions. This method of operation leverages a common tool and infrastructure base for all production-type activities.

PLANNING

In order to plan for computing for a multi-year experiment, we have to define a model and develop a set of tools to explore scenarios for spending the computing budget and making best use of offsite resources. The Run II computing program is reviewed yearly, and the detailed documentation is made available [1]. As part of the planning exercise, the Run II experiments are expected to demonstrate and justify computing costs for equipment consistent with a guidance provided by FNAL. The focus of this aspect of the review is on equipment fund allocations, with some minor focus on operating fund used to purchase tape. Maintenance and facility costs are not included. In addition to the FNAL equipment budget, DØ also has access to other monies for computing; remote in kind computing contributions at sites external to FNAL, sites which may not be in the US. University contributed systems used for desktop computing at FNAL is an additional contribution. Our first pass at developing a budget model, starting at 2002, was targeted towards exploring the best use of the FNAL equipment budget. To the greatest extent possible, we use our knowledge about the past use of the system, including the efficiency at which the systems run and past hardware costing trends to estimate future need and equipment cost. In particular, we use the past data collection rate, past knowledge of the amount of data consumed on the systems and processing time per event to estimate the future needs for fileservers, analysis and production and analysis compute node needs. Table 1 shows the estimated data collection, where we expect to increase the average output rate in 2006, due to the increased instantaneous luminosity. Mass storage estimates focus on the amount of tape to buy and insuring that we have enough slots in the existing robots to accommodate them. We do not yet have a good data rate driven model for estimating the number of drives to purchase based on fundamental parameters, although we intend to develop such a model. The use of the tape drive plant at peak is used an indicator that more drives are needed (or not). Another large spending category is infrastructure, which covers equipment for database support, interactive use, networking, and I/O support for the farm and analysis cluster. We estimate of the networking costs based on the number of compute nodes and fileservers that we estimate that we need to purchase. For other items, we look at the aging elements of the system and work out strategies to for replacing them. As most of the initial systems were purchased at roughly the same time prior to the start of Run II, many of them are now due to be replaced. We have been budgeting for the past several years to replace various elements; however, in the face of demands for compute cycles, the infrastructure updates are often postponed. This is to be expected--the nominal equipment computing budget for the Run II experiments was $2M/year, however starting in 2003, the budget was reduced to $1.5/year (reduced to 1.3M due to an overrun in 2002), and in 2004, the budget was $1.4/year, but with additional calls to support the new High Density Computing Facility at FNAL. Recent infrastructure purchases include replacing the SUN database machine in 2003, with reusing the disk from the

Table 1: Accummulation of data for the DØ experiment, including current accummulation, and outyear projections assuming an increase in data rates in 2006

data samples (events)	Current	2005	2006	2007	2008
events collected	1.00E+09	5.05E+08	9.46E+08	9.46E+08	9.46E+08
total events		1.50E+09	2.45E+09	3.40E+09	4.34E+09
TAPE data accumulation (TB)					
Yearly storage (TB)	757	525	697	763	830
total storage (TB)	757	1,282	1,979	2,742	3,572
disk data accumulation (TB)					
Yearly storage (TB)	45	51	96	96	96
adjusted for format change in 2005	0	43	0	0	0
Yearly adjusted storage (TB)	45	95	96	96	96
total storage (TB)	45	140	236	332	428

older machine and systematically upgrading to higher capacity in order to meet the expanding storage needs. Among the infrastructure drivers for 2005 are replacing the SGI systems used for IO on the analysis system and the production farm facility. In addition we are exploring solutions for home areas and login pools. The current cost estimates for 2005, with out-year projections for 2006-2008 are shown in Table 2. Note that the estimates in the table reflects a bottoms-up estimate, and choices will have to made in order to bring the budget into line with the FNAL guidance of $1.5 M for 2005. As we are improving the reconstruction speed, it is hoped that need for the reconstruction nodes are over-estimated. Guidance for the out-years is unknown, but we anticipate they will not exceed the $1.5M guidance of recent years.

The model has, by necessity, large uncertainties. A primary cost driver for DØ is the speed of the reconstruction. The DØ experiment has a small tracking volume with high occupancy, which leads to performance challenges in achieving highly efficient tracking algorithms particularly for low pT tracks. As the instantaneous luminosity increases, the occupancy increases and the reconstruction time increases exponentially.

Virtual Center

We calculate the value of the "virtual center" that would be required to meet all of DØ's computing needs. The non-FNAL contributions to DØ computing are sizable, and our calculation of the value of the center is designed to recognize that computing contributions can be used to offset contributions to the DØ operating fund. The value of the center is calculated by assuming that all necessary nodes and fileservers with crude networking estimates to do all production activities (including reprocessing and MC production) and FNAL analysis would be purchased in the current year. In this way, for past fiscal years, we can calculate a value based on the number of events delivered and, based on the percentage of a country's contribution to an activity, assign a portion of that value to the country. This has benefits over other accounting models which typically use purchase cost as a metric. By calculating an overall need, and assigning value based on contributions, sites are able to amortize hardware purchases, and sites that run at high efficiency receive extra credit, giving the remote centers a stake in how well the applications run. For future years, we estimate the value in similar way as described above, as a guide to planning.

The calculation of "value" is one that we are still developing. The estimates for compute nodes for processing, reprocessing, and fileservers are handled as described above, but make no allowances for existing systems. The estimates of the infrastructure value are difficult to make for systems that are purchased and function for many years at time. As an example, clearly, the "value" of the legacy Origin 2000 and fibre channel disk, is not the purchase cost of $1.25M, but neither is it obvious that its value is the replacement cost in LINUX compute nodes and IDE fileservers. We currently assign the infrastructure a fixed value of $0K. For the mass storage system, we take the value of the silos to be the (used) purchase cost as they cost about the same, assign the AML2 a comparable value, and pro-rate the value of the drives based on capacity and speed. As an example, a 9940b drive is assigned replacement cost for 2005, and

Table 2: Spending profile for 2003, 2004, and projected spending for 2005-2008 for the FNAL equipment budget. This is a bottoms-up estimate and has not been adjusted for the $1.5M guidance.

	Purchased 2003	Purchased 2004	Purchase 2005	Purchase 2006	Purchase 2007	Purchase 2008
FNAL Analysis CPU	$470,000	$277,000	$417,132	$534,926	$406,376	$350,311
FNAL Reconstruction	$200,000	$370,000	$454,269	$717,742	$443,490	$362,546
File Servers/disk	$111,000	$350,000	$357,000	$356,000	$293,000	$276,000
Mass Storage	$280,000	$254,700	$40,000	$600,000	$300,000	$100,000
Infrastructure	$244,000	$140,000	$547,000	$200,000	$200,000	$200,000
FNAL Total	$1,305,000	$1,391,700	$1,815,402	$2,408,667	$1,642,867	$1,288,856

1/2 that cost in 2006 when the 9940c drives will be available. Currently, analysis at remote facilities is not considered to be part of the virtual center; however, that is anticipated to change once SAM-Grid is fully deployed. Additionally, we assign no value to wide area networking, despite the fact that excellent network connectivity between FNAL and remote sites is crucial to the performance. Table 3 shows the projected value assigned to all activities, with the FNAL based activities shown in green and remote based activities shown in yellow. This scheme has the advantage of rewarding high efficiency and of taking into account the life cycles of equipment. By this method, older equipment which continues in service contributes to the value, a benefit for countries that might be spending most of their equipment money in one or two years. Interesting to note, the value needed to supply DØ's computing needs declines slowly, except for a jump in 2006 due to a data collection rate increase. This is the expected result, if the computing needs are tied to the data collection rate, then as computing becomes cheaper, then for fixed needs, the value goes done. This is in contrast to the annual costs, which can vary dramatically as legacy equipment is replaced.

Table 3 The projected value for DØ computing for the next four years. The FNAL based activities are shown in green, remote in yellow

	Estimated Value			
	2005	2006	2007	2008
FNAL Analysis CPU	$724,054	$833,811	$817,048	$738,631
FNAL	$820,089	$1,087,730	$773,295	$543,752
File Servers/disk	$560,000	$688,000	$528,000	$560,000
Mass Storage	$1,182,000	$1,201,000	$1,501,000	$1,501,000
FNAL Infrastructure	$0	$0	$0	$0
MC	$128,353	$170,152	$160,390	$85,056
Reprocessing	$1,792,632	$3,317,845	$3,245,506	$2,940,120
Virtual Center Total	$5,207,128	$7,298,539	$7,025,239	$6,368,560

The disadvantage to this scheme occurs when a large production activity is delayed relative to the anticipated schedule. The remote sites might have reserved capacity for a certain activity for a certain time. If the activity can take place in that time frame, the site might not be able to participate. Clearly, the model has to be adjusted in such a case. One way is to allow for "carry-over". Another is to assign a nominal credit based on the available resources and average site efficiency. Fortunately, the Monte Carlo generation is a steady activity, and can be used to offset some of the more targeted activities such as reprocessing. Introducing other steady activities into the accounting, such as user analysis, will also provide a scheduling buffer.

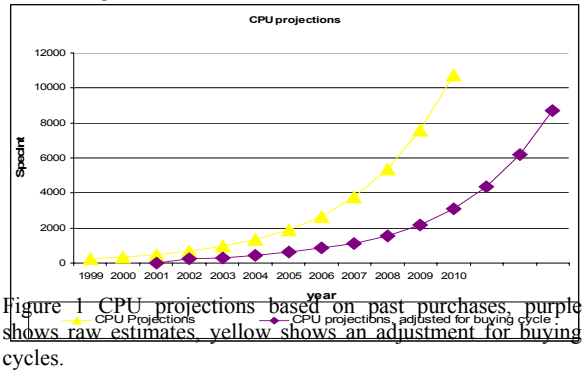

Figure 1 CPU projections based on past purchases, purple shows raw estimates, yellow shows an adjustment for buying cycles.

Another difficulty, independent of the accounting is shown in Figure 1. Figure 1 shows hardware cpu projections shown as a function of time based on fitting a function to past purchases. The raw curve is shown in purple, but as we purchase based on best performance per dollar, an adjustment is applied. While the raw projection itself is dubious, applying the adjustment for buying cycle would lead to considerably different set of constraints.

CONCLUSIONS

DØ has been using data rate based cost project models for planning for several years, and the model has proven useful for understanding the trade-offs associated with a limited computing budget. In addition, by extending these concepts, the collaboration has started to work towards an equitable way of assigning value for in-kind computing contributions.

ACKNOWLEDGEMENTS

DØ thanks the staffs at Fermilab and collaborating institutions, and acknowledge support from the Department of Energy and National Science Foundation (USA), Commissariat \` a l'Energie Atomique and CNRS/Institut National de Physique Nucl\'eaire et de Physique des Particules (France), Ministry of Education and Science, Agency for Atomic Energy and RF President Grants Program (Russia), CAPES, CNPq, FAPERJ, FAPESP and FUNDUNESP (Brazil), Departments of Atomic Energy and Science and Technology (India), Colciencias (Colombia), CONACyT (Mexico), KRF (Korea), CONICET and UBACyT (Argentina), The Foundation for Fundamental Research on Matter (The Netherlands), PPARC (United Kingdom), Ministry of Education (Czech Republic), Natural Sciences and Engineering Research Council and WestGrid Project (Canada), BMBF and DFG (Germany), A.P.~Sloan Foundation, Research Corporation, Texas Advanced Research Program, and the Alexander von Humboldt Foundation.

REFERENCES

[1]http://d0server1.fnal.gov/projects/Computing/Reviews/Sept2004/Index.html contains much more detailed information about the DØ computing model and future plans, and financial agreements.