Aug. 16, 1999

DØ Physics Analysis Project Servers

Report of the Global Computing Model Working Group

(Gilvan Alves, John Butler, Dave Cutts, Tom Diehl, Mike Diesburg, Paul Grannis, Lee Lueking, Pierre Petroff, Heidi Schellman, Mark Strovink)

The model for DØ computing recommended by the GCM in August 1998 included three layers:

- The Central Analysis server now expected to be about 90kMIPs, 30 TB disk, 750 TB robotic data store, with 3 GB Ethernet interconnection;
- the Project Servers (PS), then suggested to be about 5 separate isolated servers each with about 5kMIPs and 1 TB disk,
- and the desktop machines.

The charge for the 1999 discussions was to formulate the needs and applications for the Project Servers. The GCM held two meetings at the Seattle workshop and a meeting at Fermilab on July 22. Talks presented included:

(Roger Moore) A Linux server operating at Michigan State with Redhat Linux 6.0 and VMWare NT emulator. The NT and Linux systems are co-resident; the NT performance was quoted at 95% (CPU) and 60% (disk access) relative to that of a bare NT machine.

(`http://ettin.pa.msu.edu/linux/talks/LinuxCluster.ppt (.pdf)`)

(Greg Landsberg) Potential applications for the PS (see below), and discussion of use DVD RAM jukeboxes for cheap data storage.

(`http://brserver1.hep.brown.edu/users/Greg/gcm.pdf (.ps, .ppt)`)

(Rich Partridge) R&D on a data mining application for events and objects using commercial relational databases on an NT platform. The hope is to improve the time required to access and analyze physics objects by one or two orders of magnitude.

(`d0server1/users/grannis/POD_GCM_seattle.ppt`)

(Heidi Schellman) Comments on Project Server disk management; what needs to be backed up and what needs fast access. It was noted that the primary data store resides with the CAS/robot systems, so that event data at the PS can be considered secondary. However, user data (mail/memos/logs; source codes, perhaps small ntuples) residing in user directories will need backup services.

(`www-d0.fnal.gov/ schellma/workshop99/GCC99_19990627.ppt`)

(Lee Lueking/Don Petravick) Storage area networks (SAN) projects. Computing Division has been pursuing ways to provide shared access to disks from multiple hosts. A SBIR project in this direction was recently terminated. The GFS project at Minnesota, based on open source Linux looks promising, and is aimed at multiple platforms. It is believed that GFS could be the basis for 'modest' SAN's of the sort envisioned for DØ Project Servers. DØ should indicate its interest for CD development.

A strawman proposal prepared in Seattle was discussed in the July 22 meeting. The resulting recommendations and comments from the GCM to the DØ Computing and Software leaders and the spokesmen are given below.

Recommendations and Discussion:

The specification of Project Servers (roughly 1 Terabyte disk, and of order 5 kMIPs) was discussed several times, and found to be approximately matched to the applications envisioned. We stress that if the applications were to vary significantly upwards or downwards, then the rationale for the PS layer would be weakened. The desirability of the Project Server layer is based on the existence of applications requiring datasets near the geometric mean of desktops and the CAS.

1. Project Server desirabity:

Many applications have been identified that would be suitable for a Project Server. Many of these could be executed on the CAS as well, but there are several advantages to contemplating the additional PS layer. These include:

1. Flexibility – the PS, based upon commodity PCs, offer superior cost effectiveness for analysis of restricted data sets, where the full connectivity to the extended data set available on the CAS is not needed. Thus, we could provide a framework for computing where new types of applications could be addressed with relatively less cost.
2. Some DØ physicists propose specific applications for our analysis computing using NT, which is not currently supported by Fermilab Computing Division for physics analysis. SLAC has recently developed a Java Applications Studio physics simulation and analysis package (`http://www-sldnt.slac.stanford.edu/nld/documentation/`) that operates in the NT environment. The Project Servers would offer the mechanism for exploring and exploiting such applications.
3. The PS would serve as useful templates for analysis computing at DØ institutions or regional centers.
4. Examples of applications that have been suggested for PS are typically those that

require repeated cycles on a limited data set, or those that employ non-supported platforms:

Applications suggested include:

- EM energy calibration
- Jet scale calibration
- SMT/CFT alignment
- trigger simulation
- Jet out-of-cone studies

- b-tag algorithm development
- most ntuple based analyses
- B physics analyses
- special reconstructions (kinks, K_S, long-lived particles)
- "hot physics" chases
- data mining based on commercial database programs

- Minbias event overlay; NOISY
- pdf grid generations
- W and top mass templates
- fast MC
- JETRAD, DYRAD etc.
- basic MC event generations, GEANTing

These proposed applications appear to fit reasonably well to the PS base specifications.

<u>**Recommendation 1:**</u>

We endorse the plan for DØ Project Servers, to be used for applications on limited data sets of the order of 1 TB.

2. Responsibility for Project Servers:

The DØ needs include flexibility to address new styles of analysis computing that may arise. The collaboration has expressed willingness to participate in some way in providing the Run II analysis capability. The Fermilab Computing Division resources are presently mainly directed to provision of the reconstruction farms, the primary data storage, the Central Analysis Server, and the various infrastructures associated with data retrieval, sequential access of data, etc. There is benefit for DØ to be able to retain a measure of control over its computing environment, so as to respond to specific analysis needs as they arise, and to ensure a good connect to computing within the collaboration. The benefit of having some portion of Run II analysis computing under the primary control of DØ brings the responsibility of the collaboration to undertake to provide the resources for it.

Recommendation 2:

DØ should take the primary responsibility for Project Server analysis computing.

3. Implementation:

Implementation of the PS layer will require careful discussion on several issues, within DØ and with Fermilab CD, before a firm decision to proceed can be taken. These include specification of possible PS systems and security and maintenance issues.

We endorse the principle "if you have the root privilege for a system, you inherit the system maintenance responsibility".

Items needing study by the DØ Computing/Software group and/or Fermilab CD include:

- A detailed specification of Linux and NT Project Servers should be made and subjected to the range of potential applications for suitability. Cost estimates of these specified systems should be made.
- A detailed proposal for serving data from the PS primary store to PS users should be made.
- There should be a definite proposal regarding PS data backup policy. Dataset types to be flagged for system backup and those regarded as secondary should be established.
- The positioning of PS systems within the Fermilab security firewall should be established.
- An estimate of necessary system management to be provided by DØ and by CD should be made. The cost of both elements of system management should be estimated.
- The principle adopted in 1998 that the Project Servers should be independent of each other (no fast data serving from one server to another) should be retained to the extent that disk cross-mounts degrade the system performance. Independence of the several PS systems will mean that DØ must develop some new styles for analysis computing; users desiring to use the data sets on a server will need to take themselves to machines connected to that server.

  The potential for Storage Area Networks capable of giving multiple connections to a

PS data set from workstations within the PS, or across servers, should be explored. The project envisioned by the Computing Division to develop effective sharing of disks across the Project Server nodes gives an attractive prospect for improving the ability of multiple users to work on common data sets, and should be pursued.

Recommendation 3:

The DØ Computing/Software leaders should, in conjunction with Fermilab Computing Division, prepare responses to the implementation issues listed above. In view of the need for collaboration review and approval of PS plans, we ask for responses by Oct. 1 1999.

4. Project Server support:

To retain the flexibility of the Project Servers and the responsiveness to DØ needs, we believe that the DØ institutions will need to bear a major part of the costs of the system. Implicit in the recommendation that DØ retain control of much of the PS layer is the obligation to assume the major share of its cost and maintenance.

The estimated cost of a project server (5kMIPs, 1 TB disk, etc.) in next years currency is roughly $100,000 per server. For the purpose of estimation, we guess that in full analysis mode, DØ will want about 10 such servers. With approximately 500 physicists on the DØ masthead, this means that we will need to contribute about $500 per person per year over a four year period to provide the Project Servers.

Recommendation 4.1:

Each DØ institution should supply $500 per person on the masthead per year for the years 2000 - 2003 toward the Project Server analysis computing.

This recommendation extends and formalizes the recommendation last year that DØ support its physics analysis through a tax. (Last year the GCM recommended that for each desktop provided at Fermilab, an institution would give $1000 for project disk of the institution's choice and $1000 toward common project disk.) The current recommendation replaces that of last year, and recognizes that the existence of flexible analysis computing is in the interests of all in the collaboration.

The use of the common fund would be targeted to costs related to the Project Servers. This includes initial equipment cost, licenses and costs related to system management and maintenance, including personnel as necessary.

The principle of a 'common fund' for experimental support in large collaborations has long been standard at CERN and other Laboratories. It is now being used more frequently within U.S. collaborations as well. The tax recommended seems at least commensurate with the common good that we all receive through the operation of a forefront detector.

The tax would ordinarily be fulfilled either through cash transfers to a budget code established at Fermilab for DØ Project Servers. Contributions to computing hardware or system management, in kind, could be foreseen, if clearly approved by DØ as a part of Project Server clusters.

We envision that the Fermilab physicists should participate in this common fund. This principle will require careful discussion and negotiation, in which we should recognize the unique position of our Fermilab colleagues, through whom so much of the DØ infrastructure is already provided. One example of a Fermilab contribution might be the overall system management needed to maintain and backup the Project Servers.

The ability of DØ institutions to pay the common fund tax may vary from institution to institution, and over time. It will be necessary that the collaboration establish mechanisms to oversee the contributions to the common fund, to rule on exceptions, or institute special case solutions. We advocate that the common fund be administered through the Institutional Board and a specially formed 'Financial Committee' representing the management of major groupings within the DØ collaboration. These bodies would need to set up procedures to handle special cases, to assess the yearly contributions, and to fix the specific hardware, software and personnel costs to be supported through the common fund. It will be essential that there be widespread acceptance of the principle of the common fund, and that its cost be seen to be distributed equitably across the collaboration.

The Fermilab Directorate should be involved in establishing any plan for imposition of a common fund, as this has not been standard practice at the Laboratory.

We also note that this common Project Server fund offers a way for new groups applying to DØ to contribute materially to the experiment after the completion of the upgrade. The collaboration may wish to require larger contributions to the common fund, in lieu of prior contributions to the upgraded detector, from such institutions.

Recommendation 4.2:

The DØ Computing/Software group should bring its proposed implementation for Project Servers and the cost per masthead person per year to the Institutional Board for its October 1999 meeting. The IB should rule on the acceptance by DØ of the plan, and take the responsibility for setting up the panel to regulate the common fund procedures.

5. Initial PS complement:

The full range of applications of the Project Servers and the relative merits of computing there or on the Central Server is not yet fully explored. Our advocacy of the Project Server layer is based largely upon the flexibility and cost-effectiveness that it gives, and the chance for DØ to move quickly to provide new computing methods for problems as they arise.

In the initial analysis of Run II data, it is likely that the CAS could provide the CPU/disk/connectivity needed. However, even in the early analysis phase, there are many tasks that would also be well suited for PS computing. We believe that since there are suitable applications (geometry, calorimeter calibrations, trigger simulations, energy/momentum corrections etc.), and since the collaboration needs to gain experience to guide future analysis computing development, that we obtain experience with PS systems starting at the beginning of Run II. It may be that the initial PS servers need not have the full 1TB/5kMIP capability in the early days, but we advocate that the initial PS systems be implemented with capability substantially above that in a desktop machine, so as to understand better the role of this intermediate level.

The utility of Linux-based analysis computing is already demonstrated through the use of such platforms for Monte Carlo data analysis, and the implementation of Linux in some of our institutions. Thus initial implementation of a Linux PS system is entirely natural and can serve as a template for exploitation elsewhere in the collaboration.

The use of NT servers for analysis computing is not yet widespread, but there are a variety of suggested applications that benefit from the availability of commercial software. Applications at other laboratories using NT may have utility for DØ computing. DØ needs to gain some experience in this sector to guide future development. In this area, DØ is ahead of the Fermilab Computing Division, and our experience can help shape future CD directions. We note the existence of substantial NT computing power within DØ, and understand that this could be the basis for an initial NT Project Server.

<u>Recommendation 5:</u>

We recommend that two project servers be acquired for the start of Run II. This implies a purchase by spring 2000.

We recommend that the initial PS implementation have one NT server and one Linux server.

6. Future of GCM:

The GCM has now been in existence for 15 months and has delivered broad recommendations to the collaboration on DØ analysis computing. The detailed implementation of these plans now rests with the Computing/Software group and spokespersons.

Recommendation 6:

We recommend that the GCM working group be disbanded.