DØ Run II Data Management and Access

Lee Lueking for The DØ Collaboration

Fermilab, Batavia, Ill, U.S.A.

During the Run II data taking period at Fermilab, scheduled to begin in 1999, DØ plans to accumulate at least 200 TB of raw and reconstructed data per year. Data access patterns observed in the Run I experience have been examined in an attempt to establish an efficient data access environment. The needs and models for storing and processing the upcoming data are discussed.

Key words: Mass Storage; Data Access; D0 Experiment; Data Management

1 Introduction

During the Run II data taking period at Fermilab scheduled to begin in 1999, D0 plans to accumulate raw and reconstructed data at a rate of at least 200 TB per year. The basic overview of the management and processing of this data is shown in Figure 1, with many of the major elements indicated. Our experience with access patterns used in the run I analysis is valuable as a guide in designing the Run II system, although the much larger data volumes will force streamlining the model and the use of new techniques. We are currently discussing the details of the computing and data storage model, which will be tailored to these needs.

2 Storage and Computing Needs

DØ has established initial estimates for its basic needs and these are summarized in Table 1. The numbers in this table are a snapshot of estimates based on current expectations with many assumptions and are, of course, subject to change. These numbers are estimated assuming that the average DAQ rate will be 20 Hz. This implies that the actual online rates will be higher and we assume the standard Snowmass year of 10^7 seconds, and multiply the average rate by π to obtain a peak rate.

Preprint submitted to Elsevier Preprint *22 May 1997*

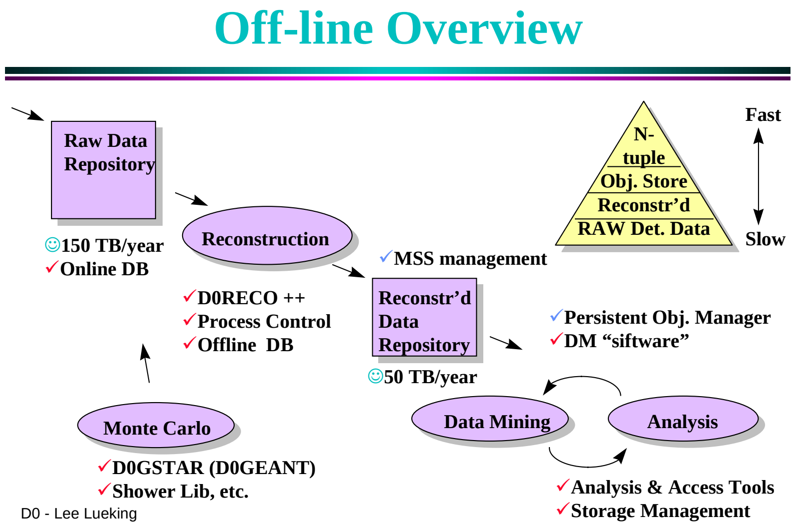

Fig. 1. Summary of DØ data processing, management and access.

The needs are considered to be the base level as several assumptions are made which may not be realistic. First, it is assumed that the rejections of the Level 3 systems would achieve high factors. Second, the event sizes for the raw detector data and the reconstructed information are not well known at this early stage. Third, minimal replication of data at any level is included as a goal to achieve a realistic storage budget. No allowance has been made for any re-processing of data and storing multiple versions of data. It is not known how CPU intensive the new reconstruction and analysis software will actually be.

Anticipated DAQ Rates - It appears easy to provide DAQ rates far in excess of what practical offline resources will be able to handle, but the target for level 3 is 20 Hz with 250 KB/event. The Level 2 output is expected to be to 800-1000 Hz. From this, the Level 3 will provide a reduction factor of 50 to the expected 20 Hz with burst rates as high as 100 Hz. High rates will be buffered and sent directly to the Computing Center at 20 Hz. The 250 KByte is probably the minimum event size although this already assumes some inflation from L3 processing information. The online system will allow DAQ rates much higher than those shown in the table and there will, undoubtedly, be considerable pressure to store data at higher rates than the storage budget might allow.

Category	Parameter	Rate
DAQ Rates	Peak Rate	62 Hz
	Average Rate	20 Hz
	Event size	250 KB/event
	Average data rate	5 MB/s
	Level 2 out	800 Hz
Data Storage	EVENTS	600 M/yr
	RAW	158 TB/yr
	DST	32 TB/yr
	μDST	3 TB/yr
	n-tuple + analysis	3 TB/yr
	Total	196 TB/yr
CPU:	Reconstruction	1200-2400 MIP·sec/event
	Reconstruction	30,000 - 60,000 MIP
	Analysis	30,000 - 40,000 MIP

Table 1
Summary of estimated Run II needs for DØ.

Data Storage - There is large uncertainty in the per event data sizes. The detector and DAQ designers have provided best guess estimates for the raw event size, but it is still quite early and many new detector elements are being added which may inflate the size. The reconstruction program is in an early stage, so only estimates based on Run I experience can be provided for the amount of information added to each event in offline processing.

The storage numbers in Table 1 assume minimal replication, but this strategy may not provide the needed access. In the past, analysis access to data was improved by tiered data types and by streaming or splitting data in each tier according to trigger or filter attributes. This approach has required a large replication overhead, sometimes as much as a factor of two. It is hoped that this factor can be reduced significantly by streamlining the approach of Run I, or by establishing a new data model which will use storage more efficiently.

There has always been a need for re-processing at least part of the data. Many new features will be present in the detector and reconstruction programs. We will have an entirely new central tracking system with a central magnetic field. Most of the reconstruction code will be new and there is little doubt that much of the early data will be processed more than once, and each version will likely be saved, adding to the storage needs.

Reconstruction CPU Requirements - There are several factors which will influence the reconstruction CPU requirements of Run II. Our experience from run I has given us a good understanding of the reconstruction time as a function of the number of interactions per crossing in the accelerator. The reconstruction time is strongly trigger dependent with events which include many jets or tracks requiring a great deal of CPU time. We are in the process of learning to write C++ code and are trying to understand how well its performance compares to the FORTRAN used in Run I. Most of the current C++ compilers are not well optimized and it is believed that it will require a few more years for them to mature. To clarify the discussion, we will use the standard unit of CPU processing to be the MIPS·Second. For simplicity, we will assume 1 MIPS = 1 SPECint92 = 1/40 SPECint95 = 1/10 CERN units.

We can make a naive estimate of the CPU needed for reconstruction based on our Run I experience as well as on the performance of new code being developed. It is well understood from Run I that both CPU and event size are strongly influenced by the complexity of the event which is determined by the number of interactions; with this taken into account, an estimate of 1000 MIPS·sec/event can be made. We can assume that a large fraction, say 25%, of the total reconstruction effort will be needed for Monte Carlo, calibration and other miscellaneous data. It is difficult to achieve complete operating efficiency for a large farm and it is reasonable to assume 75%, although 90% utilization might be attained after significant tuning. With these factors included, we can estimate a total of 30,000 MIPS will be required to maintain a 20 Hz rate. Projections for DØ Run II reconstruction CPU needs have also been done based on an early version of the Fiber Tracker program written in C++. From this we estimate, for 36 bunches in the machine operating at an instantaneous luminosity of 1.6×10^{32}, a need for 60,000 MIPS.

Analysis CPU -The analysis CPU needs for Run II can be estimated by scaling up the Run I usage by a factor of 10. The Run I need for DØ was met by about 1500 MIPS with an additional 1500 MIPS arriving near the end of the run. It is believed that these additional resources represented a much needed addition and that, on average over the Run I period we were low on analysis CPU. It is reasonable to therefore estimate that the Run II DØ need will be 30,000 to 40,000 MIPS of compute power.

3 Run I Dataset Sizes and Access Patterns

The strategy employed in Run I by D0 for delivering data relied on a traditional tiered approach. The following chain illustrates the tiers used to provide data to the analyses: RAW $\rightarrow$ STA $\rightarrow$ DST $\rightarrow$ μDST. A summary of dataset sizes for Run Ib is given in Table 2. These sizes establish the scale for ac-

DØ			
Data Type	Size/evt	Total	Comment
RAW	500 KB	30 TB	
STA	250 KB	15 TB	Reco-able
Stream STA	250 KB	10 TB	10 Streams
DST	50 KB	2 TB	Semi Reco-able
Stream DST	50 KB	3 TB	14 Streams
μDST	5 Kb	0.2 TB	Virtual Streams
ntuple			Analysis dep.
Total		55 TB	42 M events

Table 2
Summary of typical physics dataset sizes for Run IB.

cess times, and thus largely influenced the access patterns of analysis. We are studying the patterns employed in Run I for several physics analyses as part of our Run II planning.

To provide quick access to STA and DST type information, event streaming to smaller datasets was performed. This streaming was done on several platforms employing filter codes provided by the physics groups which were linked and run together in production processing. There was about a 30% overlap in the streams for the DSTs. Each stream was typically 5% of the whole sample except for high Pt leptons and QCD jets which were much larger. This 5% number was a target determined to be the sample size which could be accessed using reasonable resources in the time frame of two to four weeks. The μDSTs were kept on a central cluster and read over the network to the analysis clusters. They were virtually streamed into datasets, however some groups preferred to filter selected events to local disks rather than deal with the network. This strategy was straightforward, but was complicated by multiple reconstruction code versions.

Most of the analysis efforts were performed using primarily DST or μDST data sets with emphasis being placed on creating ntuples to study distributions or for final event selection. Access to STAs was primarily through an event selection facility which determined the event location from database information, then staged the tape and retrieved the data. Picking events from particular streams greatly expedited this, requiring fewer tape mounts. The primary constraints were centered in the following areas:

(i) **Time -** An effort which required 1 to 2 months to access information, such as going through all the DST data, was considered major but manageable.

A typical time to go through all μDSTs was about 2 weeks, making this a fairly common procedure to produce ntuple data sets. Many efforts used the virtual streams to go through the μDSTs which was somewhat faster. Tape access was, in general , very slow. The 8mm tapes read rates are 500Kb/sec max, but usually averaged between 100 and 200KB/sec accounting for operator mounts, tape positioning and other inefficiencies of serial media.

(ii) **Disk Space -** The combined Run I μDST data set was stored on disk on a file server requiring about 200 GBytes. A few local subsets of this were maintained on the analysis cluster to expedite the analysis operation. A handful of DST or STA events were selected and maintained on local disk for detailed studies of events (like event display), trigger and background studies. Most ntuple samples were maintained on disk.

(iii) **Software Configuration -** Careful preparation and testing of production software was very time consuming. Minor trigger changes online or bug fixes to the reconstruction required weeks or months to be incorporated into the final filtering and analysis efforts. It was understood that strict versioning control was required but perhaps some streamlining would have helped to more easily accommodate changes.

4 The Computing and Storage Model

The computing model being established is based on the analysis needs of the physics groups and the costs and constraints of the projected technologies. With the information from section 2, estimates can be made with regard to the costs for the computing effort needed in Run II. It is assumed that the cost of computing will decrease as the technology improves and commercial market increases. The costs for networking are falling more slowly than other areas of computing and this strongly influences the model. Although the price to performance ratio for disk storage has decreased faster than serial media over the last 10 years, there may still be a factor of 4-10 separating them during Run II. This will force most of the storage to be to tape. There is always the possibility that some breakthrough storage technology might emerge which is fast and cheap, but it is unlikely to develop a strong enough track record to be relied on for Run II needs.

Providing access to this quantity of data will require a largely centralized computing and data storage model. We can estimate that our approximate analysis bandwidth for Run I was in the region of 25-30MB/s. Our goal is to perform the Run II analysis, on a data set roughly 10 times larger, on a similar time frame and thus we can estimate a need for 250-300MB/s bandwidth to the data. This will be easily achievable on any SMP machine likely to be employed in Run II. So, we are anticipating SMP machines attached through high speed links to centralized storage.

Ideally, we would like all data to be in a random access object database and need only one copy of each data object. This is not practical as we will need to store the data in a mixture of random access and serial media. The best we can do may be to attempt to retain the flexibility of an object store by keeping a small part of the data on disk and have access to the other data in robotic or operator mounted storage. We feel that the cost for disk may be around \$50/GB and we may be able to afford enough to store a data set comparable to the μDST for Run I, which would scale to about 3 TB/year. We are still deciding on the exact format for this data. One consideration might be an object database, but it is not clear what compromises need to be made to pursue such a solution. For example, such a solution might not be sufficiently compact to fit the disk budget.

Although the cost for tertiary storage is much less than disk, it will still constitute a major expendature. In an effort to reduce these costs we are considering eliminating much of the redundancy required in Run I by elimination of the STA data type, and defining a more general DST. This may cause additional overheads to the reconstruction CPU since information will be lost and will need to be remade for some event samples. The overheads caused by streaming may also be unacceptable and a very streamlined approach involving virtual streaming or a small number (3 or 4) of streams might be considered. Robotic storage is still quite expensive, and it may not be cost effective to maintain all data sets in this type of storage at all times. We may, for example, rotate raw data stores through the robotics on a few-month time scale. It may be necessary to provide some inexpensive media for archiving infrequently accessed data.

5 Conclusion

The overriding questions for the Run II data management and access are: "How much data can we take?", "How much data can we afford to keep and process?" and " Are there clever ways to squeeze the storage budget and not compromise the physics?" It is clear that the DAQ rates will be constrained by our ability to afford data storage and processing, and not by any intrinsic architectural bottlenecks. Delivering even the conservative amounts of data stated in the above discussion in the timely fashion needed for analysis may prove difficult unless careful planning and a few new strategies are involved.