CDF/DOC/COMP_UPG/PUBLIC/4100
DØ Note 3197
TN ####

Joint DØ/CDF/CD Run II Data Management Needs Assessment

Drew Baden, Liz Buckley-Geer, Joel Butler, Jim Cochran, Mike Diesburg,
Ken Fidler, Stu Fuess, Irwin Gaines, Krzysztof Genser, Umesh Joshi,
Wolfgang Kowald, Stephan Lammel, Mark Leininger, Lee Lueking, Frank Nagy,
Krzysztof Sliwa, Vicky White, Stephen Wolbers

June 4, 1997

Abstract

The charge issued to the Run II Needs Assessment Group by the Computing Division Run II Planning Group is addressed. We have estimated the hardware needs of CDF and DØ for data storage and CPU processing and provide estimates for the cost of the Run II program. We have also examined data set contents and sizes, and the access patterns employed by the two experiments during Run I in an effort to better understand the possible needs for Run II data management and access.

Contents

1 Overview

The Run II Data Management Needs Assessment Group was established by the Joint CD/DØ/CDF Run II Computing Group to evaluate information obtained from Run I and estimate projected needs for Run II in the area of data storage, management, and access. The goal of this document is to provide a summary of the projected hardware needs for both experiments and to a dictionary of terms commonly employed in describing the Run I effort. Data rates and storage requirements based on the current understanding of the proposed experiments will be reviewed and ranges of values will be discussed for each experiment. More details of the Run II computing plans can be found for CDF in the Technical Design Report [1] and computing upgrade notes [2] and for DØ in the Run II Computing and Software Planning Document [3], although many of the numbers have been adjusted to reflect more recent understanding of the needs. In addition to understanding the needs in terms of data sizes and rates, an attempt has been made to understand the actual patterns of data usage. Detailed summaries for several specific Run I analysis efforts are included in the appendix of this note. The software and computing operations needs will be addressed in a future document.

2 The Charge

Data management includes the hardware and software required to support all aspects of CDF and D0 computing beginning from the moment that raw data becomes available at the downstream end of the DA to the moment at which it is delivered to the physicist for analysis. The charge to this working group is:

1. Specify in very general terms the requirements – data volumes, access speeds, latency requirements, storage for temporary data, CPU requirements, and permanent storage – for all phases of the analysis. (These have been written down many times but this is an opportunity to get a single consistent list and for the two collaborations to understand any differences between them);
2. Create a glossary of terms which are used by the two collaborations to describe the datasets, processes, activities,and existing systems in this area to facilitate communication. Of course, any agreement on standardization of terms is most welcome;
3. Specify the kind of software needed at each stage to catalog the data flow, to track all the datasets, and to access the storage; and
4. State, to the extent possible, the human resource needs (or more like the practical limits of the resources available) to manage the system.

Among the issues that need to be considered are:

1. The need to "fall back" and redo various aspects of the processing. How far back does one have to be able to go? What are the requirements on redoing the step (how fast, how much effort).
2. What data must be accessible on disk? What data must be accessible robotically? At what rate (peak, average) does the data need to be accessed.
3. Where in the processing will streaming occur? Who are the customers for the streamed sets? How often will the streamed sets be accessed and do they need to be retained indefinitely? If so, on what kind of storage?
4. Can the data be recorded directly to robotic storage over the network? What precautions must be taken to insure that no data is lost due to network, robot, and software down time?
5. What measures need to be taken to make sure that problems in the downstream portions of the processing pipeline do not hold up the upstream portions?
6. How will dst data be accessed and reduced? What is the access model (independent of the implementation of the data model) for analysis (highly central, highly distributed, work-group)? How much 'private' storage is required for each physicist/analyst under the chosen model? How much data storage? How much IO bandwidth is needed per analyst?

7. What data (type, quantities) needs to be delivered to remote users? If the network alone will not suffice as the delivery medium, what other media are acceptable/required?

8. What software is required by the user to access the data at each level of storage?

Many other issues will undoubtedly occur to you and should be added to the list. Because this is an area where there have been (economically) unrealistic expectations in the past, the working group should at least consider whether stated requirements have a chance of being translated into affordable systems based on their understanding of current and projected costs. We would expect one or a series of reports defining the issues and requirements. A goal would be to have a preliminary report in two months.

3 Run II Data Management Needs

Both CDF and DØ have established initial estimates of basic computing needs for Run II data. Their current best estimate is summarized in Table 1. The numbers in this table are based on several assumptions and are thus to be considered preliminary. However, the numbers reflect the scale of the Run II computing challenge. The CDF picture is taken from a public CDF note which deals with the Run II data handling needs [2]. The numbers are based on a 75 Hz peak data logging rate with a 50% accelerator and experiment on-time (averaged over a year) and a 75% data logging rate variation within a store. A small amount of data is replicated similar to Run I for monitoring and fast data access. The DØ numbers are estimated assuming that the average DAQ rate will be a flat 20 Hz. This implies that the actual online rates will be much higher and, for comparison, the peak rates have been calculated employing the same assumptions made by CDF.

The above numbers assume that the third level trigger systems achieve significant rejections. The event sizes for both detector RAW data and reconstructed information are not precisely known at this early stage. Some minor replication of data is assumed to allow for efficient access. No explicit allowance has been made for any re-processing of data and storage of the same events in multiple reconstructed versions. The CPU assessment is based upon a scaling of resources utilized during Run I as well as information from some early Run II software, since it is not certain how CPU intensive the new reconstruction and analysis software will actually be.

3.1 Anticipated DAQ Rates

For both experiments, the Run II DAQ rates are yet to be completely determined, however, it appears easy to provide rates far in excess of what practical off-line resources will be able to handle. For D0, Level 2 output is expected to be around 1000 Hz. From this, the Level 3 will need to provide a maximum reduction factor of 50 to achieve 20 Hz. Burst rates as high as 100 Hz may be anticipated and these will be buffered and sent to FCC. An event size of 250 kB is probably minimum although this already assumes some inflation from level 3 processing. A low cost (few $10K) Level 3 upgrade path to 40 Hz may be implemented over the course of the run.

The DAQ system of CDF is designed for a level 2 output rate of 300 Hz. The level 2 trigger is more sophisticated than in Run I and leaves less room for additional rejection in level 3. Based on Run I experience and possible improvements in level 3 algorithms, the rejection rate could be as high as a factor of 4. This yields a 75 Hz peak data logging rate with an event size of 250 kB. Data logging will be based on Fibre Channel (FCAL) allowing a bandwidth far exceeding the target output rate of 18 MB/s required for the 75 Hz.

Category	Parameter	DØ	CDF
DAQ Rates	peak rate	53 Hz	75Hz
	equivalent DC rate averaged over the year	20 Hz	28Hz
	average event size	250 kB	250 kB
	average data rate	5 MB/s	7 MB/s
	level 2 output	1000 Hz	300 Hz
	maximum data logging bandwidth	scalable	80 MB/s
Data Storage	number of events	600 M/year	900 M/year
	RAW data	150 TB/year	250 TB/year
	reconstructed data STA(DØ)/DST(CDF)	75 TB/year	135 TB/year
	physics analysis data DST(DØ)/PAD(CDF)	50 TB/year	79 TB/year
	μDST	3 TB/year	-
	total data volume	280 TB/year	464 TB/year
CPU:	reconstruction/event	1000-2500 MIPS·s/event	1200 MIPS·s/event
	reconstruction	34,000 - 83,000 MIPS	56,000 MIPS
	analysis	60,000 - 80,000 MIPS	90,000 MIPS
Software	event database	required	required
	process control	required	required
	storage management	required	required
	secondary databases	required	required

Table 1: Summary of estimated Run II needs for CDF and DØ. Storage needs are on a yearly basis.

3.2 Data Storage Requirements

The detector and DAQ designers for both experiments have provided best estimates for the raw event size, but it is still quite early and many new detector elements are being included in both systems which may increase the size, especially at the beginning of Run II. The reconstruction programs are in early stages, so only estimates based on Run I experience can be used for the amount of information generated in the reconstruction programs. The data sizes are also a function of luminosity and number of bunches as is shown in Table 2 for various data types at CDF for Run I.

The storage numbers in Table 1 assume very little data 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 data replication overhead sometimes as much as a factor of two. For example, in Run I, the output from the reconstruction farms provided an event type which included the raw detector information packaged with the additional reconstructed information. It is not certain what the overheads might be if the reconstructed information and raw data are maintained separately, and links provided to the raw data. It is hoped that this factor can be reduced by streamlining the approach of Run I, or by using a new more efficient data storage model.

The following assumptions are used by CDF to arrive at the data volume in Table 1. The reconstructed event size is estimated to be 50% of the RAW data size or 125 kByte/event based on Run Ib experience. Run II PAD size is estimated to be around 25% of the RAW data size, or 63 kB/event. It is assumed that data access can be provided efficiently without significant data replication. RAW data are not stored with the reconstructed data and DSTs are not split into datasets (as was done in Run I). PAD data will be split into a large number of datasets, introducing some small replication (about 5%). For fast access an "express-line" system analogous to Run I containing less than 10% of the data is assumed. Such a configuration was very successful in Run I in expediting important analyses, in monitoring of the detector, and in debugging of the reconstruction programs. This "express-line" subset will be processed with each new version of the reconstruction program. Access to both latest and previous versions of reconstructed data needs to be provided for debugging purposes. These assumptions lead to an effective stored event size of about 520 kB.

One model considered by DØ envisages event clustering, or streaming, at the output of the reconstruction farm and based on Level 3 information. The number of streams will be limited to a few, and most will be of 50kB/event size. However, the model assumes that the W/Z and possibly a few other physics analyses, around 20% of the total triggers, will require the maximum information, or 250kB/event. We further assume that events will be provided randomly at the 1-2% level with the full 250kB/event information and this sample will be employed for debugging the reconstruction program and for background studies. It is understood from the Run I experience that there will be a streaming overhead of about 30% due to the overlap of events in various streams. With this model we can calculate an average need for reconstructed data storage of about 150kB/event and the total need, including the raw data of about 400kB/event. These assumptions are reflected in the storage numbers indicated in Table 1. The startup profile will probably be somewhat different in order to debug the new detector and reconstruction code, however the event rate will presumably be lower as well.

There has always been a need for re-processing at least part of the data. Many new features will be present in the detectors and reconstruction programs which may require a certain measure of this in Run II for both experiments. Both experiments will have entirely new central tracking systems and in addition DØ will have a central magnetic field. Much of the CDF reconstruction

$\mathcal{L}$ ($\times 10^{31}$)	Raw (kB)	DST (kB)	PAD (kB)
1.49	116	169	31
1.77	132	186	32
1.92	133	196	34
2.37	139	207	36
2.83	140	213	39

Table 2: Growth of data set sizes with instantaneous luminosity for CDF (6 bunches).

code will be new and practically all of that for DØ will be rewritten. There is little doubt that much of the early data will be processed more than once, and each version may need to be saved, adding to the storage needs. It is assumed at DØ that the re-processing overhead will be at the 10% level. The CDF numbers assume re-processing of the "express-line" data but storage for only two versions of this data. CDF also assumes that media used for storage of earlier reconstructed versions can be recycled towards the end of the run.

3.3 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 including many jets or tracks requiring a great deal of CPU time. We are in the process of learning to write C++ code and are still 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 power to be MIPS. There are a few benchmark units which have been employed as a metric for CPU processing power including the MIPS, TINY (FNAL units), CU (CERN Units), and SPEC (Standard Performance Evaluation Corporation). The various units are only roughly comparable since the benchmarks depend on the relative weighting of floating point verses integer operations, hardware configuration, as well as many other factors. For simplicity, we will primarily use MIPS and assume 1 MIPS = 1 TINY = 1 SPECint92 = 1/40 SPECint95.

In Run I one expected an average of 1 minimum biased event per crossing for each $6 \times 10^{30} cm^{-2} s^{-1}$ of instantaneous luminosity. The number of events per crossing for Run II will depend on whether there are 36 or 99 bunches in the machine, and this will be discussed in more detail later. Figure 1 shows the average number of interactions as a function of Luminosity for 6, 36 and 99 bunches in the machine [4]. 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 and trigger configuration. Table 3 shows reconstruction times per event measured in Run I.

We can make a preliminary estimate of the CPU needed for reconstruction based on our Run I experience. A typical DØ event required between 400 and 1000 MIPS·s, and a typical CDF event required 500 to 700 MIPS·s for luminosities ranging between $0.3 \times 10^{31} cm^{-2} sec^{-1}$ and $1.4 \times 10^{31} cm^{-2} sec^{-1}$. CDF's estimate is for a 20% increase in the per event CPU requirements for Run II, which pushes their need to about 1200 MIPS·s/event for the execution time, DØ has estimated a lower limit on its need assuming 1000 MIPS·s/event.

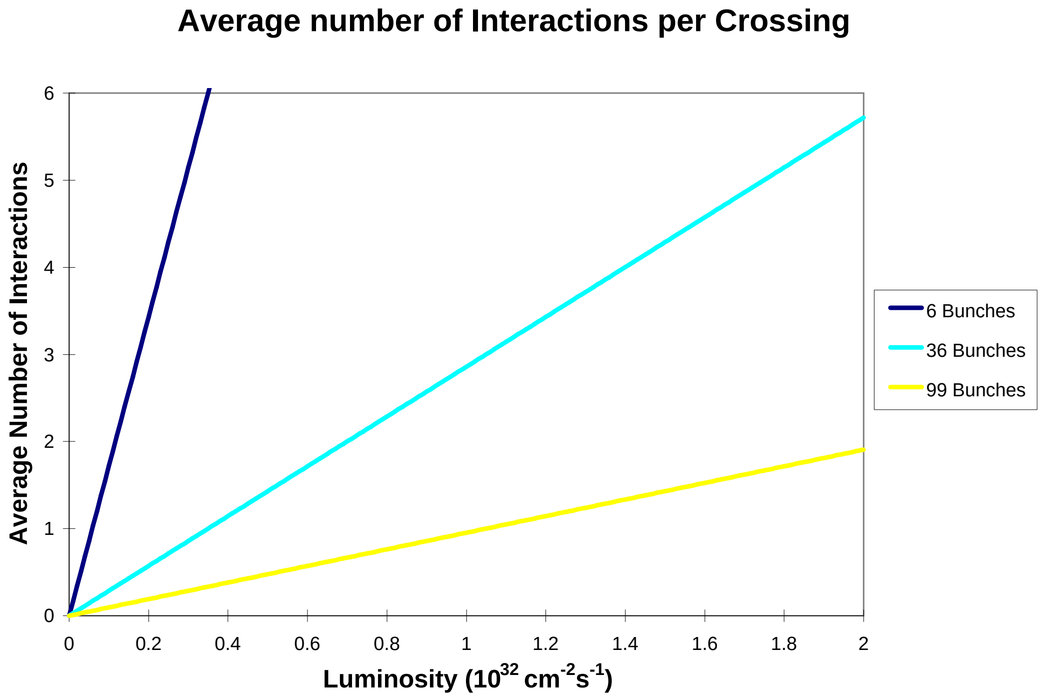

Figure 1: Average number of Interactions per crossing as a function of luminosity for 6, 32 and 99 bunches in the accelerator.

$\mathcal{L}$ ($\times 10^{31}$) $cm^{-2}sec^{-1}$	Avg. Interactions per Crossing	CPU MIPS·s/event	
		DØ	CDF
0.07	.1	460	-
0.3	.5	-	527
0.55	1	540	587
0.8	1.3	-	626
1.0	1.6	-	700
1.40	2.3	1000	-

Table 3: Run I reconstruction CPU needed as a function of instantaneous Luminosity. There were 6 bunches in the accelerator throughout the run.

	Est. DØ Reconstr. MIPS			Est. CDF Reconstr. MIPS		
MIPS·s/event	20 Hz	75% effic.	25% added	28 Hz	75% effic.	25% added
800	16,000	21,000	27,000	22,000	29,000	36,000
1000	20,000	27,000	**34,000**	28,000	37,000	46,000
1200	24,000	32,000	40,000	34,000	45,000	**56,000**
1500	30,000	40,000	50,000	42,000	56,000	70,000
2000	40,000	53,000	66,000	56,000	75,000	93,000
2500	50,000	67,000	**83,000**	70,000	93,000	116,000

Table 4: Estimates of reconstruction farm CPU, in MIPS, required to maintain 20Hz and 28Hz processing rates, for DØ and CDF respectively, over a range of possible CPU·seconds/event. Values are shown assuming 75% farm utilization efficiency, and also for an additional 25% for other miscellaneous processing. The boldfaced numbers are those used for the need range.

Projections for DØ Run 2 reconstruction CPU needs have also been done based on the performance of an early version of the fiber tracker program (referred to as TRF++) written in C++. It is important to bear in mind that this is preliminary code with no provisions for noise, inefficiencies, multiple scattering, non-uniform magnetic fields or other realities which will add complexity. The best estimate currently available for the entire reconstruction effort is to assume that TRF++ represents approximately 1/3 of the CPU which will be required for the complete reconstruction, the other major components being the silicon tracking, the calorimetry and the muon tracking. Figure 2 shows the estimated CPU times as a function of the number of interactions per crossing. Obviously, the number of bunches in the accelerator is an extremely important factor in determining the amount of CPU required to perform the Run II reconstruction. Figure 1 shows the average number of interactions per crossing as a function of luminosity for various bunch configurations. We can estimate, from this, that for 36 bunches in the machine, operating at an instantaneous luminosity of 1.6×10^{32} the average number of interactions per crossing would be 4. Multiplying by a factor of three the reconstruction time indicated in Figure 2 for 4 interactions would lead to a need for around 2400 MIPS·s/event.

The estimates above indicate a broad range of possible reconstruction farm CPU needs. There are, however, additional factors which should be considered when calculating these values. The DØ Run I experience indicated that a large fraction, say 25%, of the total reconstruction effort will be needed for Monte Carlo, calibration and other miscellaneous data. Also, it is difficult to achieve complete operating efficiency for a large farm and 75% was typical in Run I, although 90% utilization was attained after significant system tuning. Possible CPU requirements are shown in Table 4; the ranges in this table are indicative of the large degree of uncertainty in the estimates at this point. The CPU need ranges indicated in Table 1 are highlighted with boldface in Table 4.

3.4 Analysis CPU Requirements

The analysis CPU needs for Run II can be estimated by scaling up the Run I usage. 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 they were low on analysis CPU. It is reasonable to therefore estimate that the Run II DØ analysis computing will be larger by a factor of 20, or

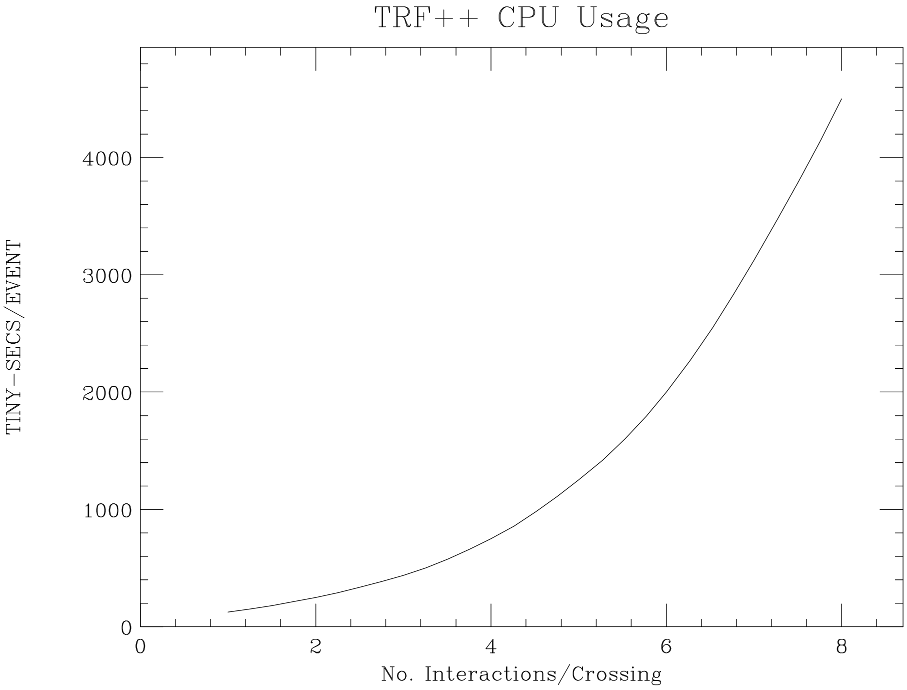

Figure 2: DØ CPU processing time per event for fiber tracking portion of the Run II reconstruction. This measure is based on an early version of the code written in C++. Full event processing time is estimated to be a factor of 3 higher. We assume 1 MIPS = 1 TINY.

in the range of 60,000 to 80,000 MIPS. The CDF estimate uses the 3600 MIPS used for Run I analysis, scaled to the increased number of Run II events, a factor of 30, to give a range of 100,000 to 120,000. Perhaps with efficient usage, 90,000 MIPS will be sufficient, but this value is a lower limit.

3.5 Considerations for Online Streaming

If it is possible for the online systems to provide rates which are significantly higher than the off-line processing can handle, then one might consider writing streams with differing levels of processing priority. In Run I, CDF wrote three streams, A, B, and C. Streams A and B contained high-p_{T} triggers such as top, W, Z as well as b-physics triggers such as $J/\psi \rightarrow \mu\mu$. Stream C contained lower priority b-physics triggers, as determined by the collaboration. While Stream A was processed in real time, Stream B was processed within about a week, and stream C was deferred until the end of Run I when the production farms had completed the higher priority streams. D0 procured off-line resources to process the entire data set in almost real time.

Of course the issues for Run II involve not only off-line processing resources, but many others as well. Even if the farm work is deferred for a large portion of the data until after the main running period, there will be a heavy commitment for storage. Also, there may be critical detector and/or running condition problems which are unobserved until it is too late for correction if this effort is deferred. It is common for such deferred streams to lose priority when the run ends and gathering manpower to process them may be difficult. There is also no guarantee that there will be a significant length of time between Run II and Run III, in which case this data might not be processed. These factors may make this option less attractive.

3.6 Simulation Requirements

The simulation needs should scale with the amount of detector data accumulated and we assume they will represent around 10% of the total needs for storage and processing. Much of the major CPU intensive simulation, i.e. full GEANT, work has been moved to outside institutions at DØ for Run I, and it is hoped that this will also happen in the upcoming run. The simulation work at CDF was performed on central systems during Run I, and large off-site processing is currently not part of the Run II plan. A large portion of the events needed will be done with fast Monte Carlos employing shower libraries or parameterized modeling of events which significantly reduce the CPU needs, but still require storage. There will be a need for a centralized catalog to track the outputs from all of these efforts and maximize the potential use among physics groups within each collaboration.

3.7 Software Requirements

It is understood that there are large software needs for Run II. The areas include database management systems for calibrations, alignment, and possibly event data. For the large amount of storage, specialized management software will be needed. Process control for data processing will be important and software for this needs to be provided. Additional software costs associated with storage management are likely.

4 Run II Cost Estimates and Spending Profile

With the information from the previous section, some estimates can be made on the costs for the computing effort needed in Run II for CDF and DØ. It is assumed that the cost of

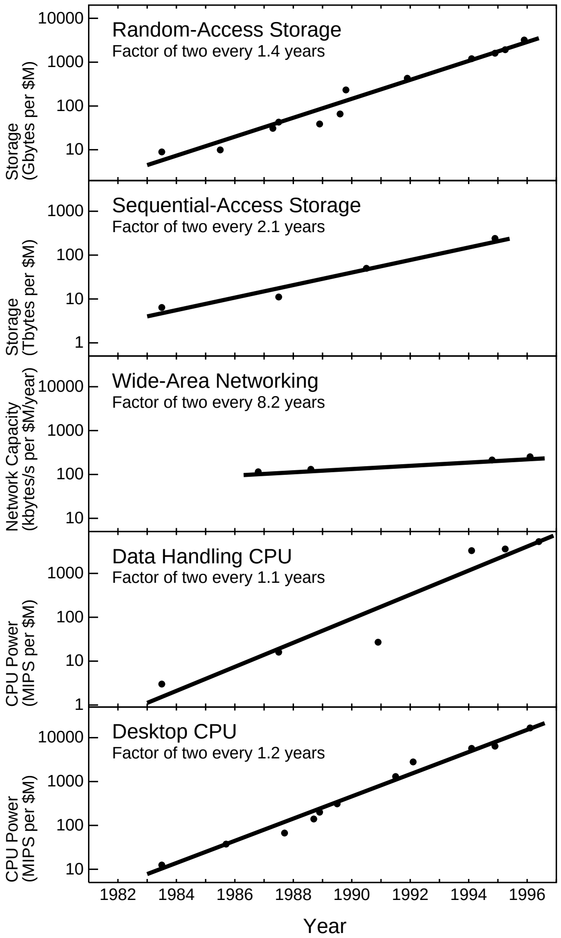

Figure 3: Performance per cost ratio as a function of year for key areas of computing. It should be noted that some of the data indicate a "stair-step" function rather than a line as a result of new or breakthrough technologies. The performance to price ratio for Run II purchases may not be as high as the straight lines might indicate.

computing will, in general, decrease as the technology improves and commercial market for these items increases. This trend is shown for various areas of the computing industry in Figure 3[5]. Although the costs for disk storage have decreased faster than serial media over the past 10 years, there is still an order of magnitude separating them, assuming robotic tape handling. This will force most of the storage to be on serial media, and tapes and tape robotics will constitute one of the major outlays for Run II. There is always the possibility that some breakthrough technology might emerge which is more economical, but it is unlikely to develop a strong enough track record to be relied on for Run II needs.

The costs for networking are falling more slowly than other areas of computing and this influences the computing model. Providing access to the huge amount of data requires sophisticated caching to keep frequently accessed data close to the analyzing CPU. We estimate the data bandwidth for Run I analysis to be a few MB/s. Run II datasets will be on the order of 20 times larger and thus require an analysis bandwidth of order 100 MB/s if the same computing model is used. Such bandwidth does not require high speed networking to desktop systems.

Ideally, one would like all data to be available in random access storage with fine granularity and no duplication. This, however, is not practical as we will need to store the majority of the data on serial media due to cost issues.

4.1 Disk Storage

Disk storage is required for the most frequently accessed data. Especially in the later phases of an analysis, data are accessed repeatedly with similar patterns. The most efficient use of disk storage depends strongly on the computing model chosen. The CDF disk space requirement is based on the Run I model and experience. The frequently assessed PAD data need to be directly accessible. This requires about 28 TB of disk space for Run II analysis (or 14 TB per year). It is estimated by DØ, based on scaling up their Run I experience, that the amount of disk storage will be about 7 TB/year. This space would be used to store their μDST, project and temporary data such as staging, streaming, et cetera.

Based on price/performance development during the past years, we estimate the costs for 1 GB of disk space between \$90, near the beginning, and \$40, near the end, of Run II. For the estimates in this note we assume a cost of \$60 per GB.

4.2 Serial Media and Robotics

There are many considerations in the area of serial media. This includes not only media costs but also acquisition and maintenance costs for the drive, media capacity, media and drive reliability, and availability of robotics. Neither experiment has made a serial media decision at this time and both are watching for market trends and technological developments. The Computing Division is evaluating potential technologies. For the estimates in this note, we assume that we can buy reliable serial media with capacity per cartridge of 50 GB or more in 1999 for \$2/GB.

The serial media cost estimate would be incomplete without a cost estimate of the drives and the data handling computing. Purchase costs of drives vary by almost two orders of magnitude depending on the technology. Maintenance costs for some technologies can be significant when compared to the pure media costs. As long as the serial media selection is open no detailed estimate can be made, but for this note we consider two extreme scenarios, 1. 8mm AME and, 2. STK Redwood, and take the average cost of the two technologies. For two years and 920 TB, i.e. CDF Run II, this translates to \$1.4M or \$680k per year. (For DØ these numbers are scaled by 280 TB over 460 TB, i.e. \$410k per year).

Both experiments assume DAQ data recording will no longer done at the experiment, but rather at the Feynman Computer Center. The data recording activity must be allocated some computing power, which may come from the analysis systems, production systems or a dedicated processors. To account for these additional costs we make a simple estimate by scaling up the drive costs by 15%.

If all data is maintained in robotic storage, it would require 460 TB/year for CDF and 280 TB/year for DØ . Although having all the serial media served by robotic systems is desirable, it may exceed the Run II computing budget. CDF has estimated the amount of "warm" data, i.e. data that needs to be accessed for analysis but that cannot be kept on disk, to include PAD information for all events and level 3 reconstruction and trigger information. A minimum of 200 TB of storage with moderate latency is required (or 100 TB per year). DØ assumes that robotic storage for about one-quarter of the entire set is required to keep only the most needed data. This would then reduce robotic hardware, for them, by a factor of four although the media costs would remain the same.

Item	Capacity	Price
Base 3494 Robot	expandable up to 15 frames	\$60K
Frame	up to 400 cartridges	\$23K
Drives	10MB/sec	\$30K
Cartridges	20GB (soon tba)	\$75

Table 5: IBM 3494 Robot Price Breakdown.

The price breakdown for one specific system, an IBM 3494 currently in use at Fermilab, is shown in Table 4.2. The price for a such a system, maximum capacity and equipped with 8 drives, would be \$650k. The media cost (6000 cartridges) would add another \$450k (estimated prices when media becomes available in summer of 1997). In this configuration the IBM 3494 would provide 120 TB of storage. We expect the media costs to drop by a factor of less than two by 1999. However, the cost for the robot itself is not expected to drop significantly during the next few years. For the estimates in this note we consider the costs of the IBM 3494 to be representative and take \$5.50/GB as the cost for robotics in Run II. Media costs need to be added to the above if the serial media used inside the robot is different than the serial media used for bulk data storage (i.e. serial media cannot be moved physically in and out of the robot). Such a scenario is possible as serial media for bulk data storage is driven by cost per capacity whereas the robotic serial media may have additional specifications and higher reliability to function with increased demand of the robot.

4.3 CPU Requirements

It is well understood that the cost for analysis CPU is significantly higher than reconstruction CPU. Both estimates are made assuming the current trend for CPU price/performance decreasing by a factor of 2 every 1.1 years, as it has historically. It is estimated that a multiprocessor system providing 30,000 MIPS will be available by 1999 for about \$1.5M; this would indicate a cost of around \$40/MIPS. Based on recent reconstruction farm acquisitions and projecting to 1999, it estimated that the price for this type of CPU might be as low as \$20/MIPS. However, since a large part of the farm needs to be in place before the run begins, we will assume a cost of \$25/MIPS. There is work underway currently to investigate the use of commodity PCs for use in farm processing. This effort has been very encouraging and may provide reconstruction CPU at even lower cost.

4.4 Software

The breakdown for in-house, freeware and commercial software is still being analysed. It is known that commercial solutions for some needs, for example DBMS, may require site license or other similar agreements with the vendors. These licenses, and maintenance fees may be significant, \$100K's, but if they free up vital manpower for other less mundane tasks this may be a wise investment. We continue to search for the best solutions to match these needs. It is clear that manpower will be required for software during the evaluation, implementation processes, as well as for maintenance.

Category	Parameter	DØ	CDF
Serial Media	estimated need	280 TB/year	460 TB/year
	unit cost	$2 /GB	$2/GB
	total cost	$560k/year	$950k/year
Robotics	estimated need	100 TB/year	100 TB/year
	unit cost	$5.50/GB	$5.50/GB
	total cost excluding media	$550k/year	$550k/year
Disk Storage	estimated need	7 TB/year	14 TB/year
	unit cost	$60/GB	$60/GB
	total cost	$430k/year	$860k/year
Tape Drives and I/O CPU	need to handle	280 TB/year	460 TB/year
	unit cost	$1.7/GB	$1.7/GB
	total cost	$470k/year	$780k/year
Reconstruction CPU:	estimated need	60,000 MIPS	56,000 MIPS
	unit cost	$25/MIP	$25/MIP
	total cost	$1,500k	$1,400k
Analysis CPU:	estimated need	70,000 MIPS	90,000 MIPS
	unit cost	$40/MIP	$40/MIP
	total cost	$2,800k	$3,600k
Software	event database	to be determined	to be determined
	process control	to be determined	to be determined
	storage management	to be determined	to be determined
	secondary databases	to be determined	to be determined

Table 6: Summary of Estimated Run II costs for CDF and DØ.

4.5 Cost Summary and Spending Profile

Based on the anticipated need and projected computing prices, a summary of estimated costs is provided in Table 6. Roughly, the indicated needs are the average based on the ranges in Table 1. In the area of software, the exact needs and costs are not yet determined. We also include a possible spending profile for the fiscal years 1998 through 2000 in Table 7. Spending in 1998 will be restricted to areas which are critical for detector commissioning and startup preparation. This includes I/O systems for reconstruction farms, infrastructure for data handling, and robotics. Most of the farm CPU will be purchased in FY 1999 and FY 2000 to leverage the cost-performance advantages. It will be mandatory to acquire key infrastructure elements and base data handling and robotics in FY 1998 to allow enough integration time. Buying those components a year before the run is expected to impose little cost-performance disadvantage.

Included in the spending profile are also miscellaneous and base computing costs which cover items such as infrastructure and desktop needs specific to each experiment. Their estimate is based on Run I costs.

Item	Experiment	FY98	FY99	FY00	Total
Serial Media*	DØ	$0.010M	$0.560M	$0.550M	$1.12M
	CDF	$0.010M	$0.990M	$0.900M	$1.90M
Robotics	DØ	$0.200M	$0.550M	$0.350M	$1.10M
	CDF	$0.200M	$0.550M	$0.350M	$1.10M
Disk	DØ	$0.040M	$0.400M	$0.420M	$0.86M
	CDF	$0.040M	$0.800M	$0.880M	$1.72M
Tape Drives and I/O CPU	DØ	$0.200M	$0.490M	$0.250M	$0.94M
	CDF	$0.300M	$0.860M	$0.400M	$1.56M
Reconstruction CPU	DØ	$0.100M	$1.000M	$0.400M	$1.50M
	CDF	$0.100M	$0.850M	$0.450M	$1.40M
Analysis CPU	DØ	-	$1.400M	$1.400M	$2.80M
	CDF	-	$1.800M	$1.800M	$3.60M
Miscellaneous	DØ	$0.200M	$0.300M	$0.300M	$0.80M
	CDF	$0.200M	$0.300M	$0.300M	$0.80M
Total	DØ	$0.750M	$4.700M	$3.670M	$9.12M
	CDF	$0.850M	$6.150M	$5.080M	$12.08M
Total Equipment		$1.580M	$9.300M	$7.300M	$18.18M
Total Operating		$0.020M	$1.550M	$1.450M	$3.02M

Table 7: Possible spending profile for CDF and DØ Run II computing. * Serial media is part of the operating budget.

5 Conclusion

The overriding questions and concerns of the Run II data management needs assessment group are: “How much data will the two experiments take?”, “How much data can we afford to process and store?” and “Are there clever ways to squeeze both storage and CPU budgets without compromising the physics?”

We have made a careful study of the Run II computing needs based on experience gained in Run I, anticipated DAQ rates for Run II, and projected costs for computing. It is clear that the DAQ rates will be constrained by our ability to afford data storage and not by any intrinsic architectural bottlenecks. Storing, processing, and delivering the large amounts of data stated in the above discussion will be expensive even with costs for computing which are expected two or three years from now.

A Run I Data and Analysis

A.1 Data Types

The strategy employed by both D0 and CDF for delivering data for physics analysis used the traditional tiered approach where the amount of information stored per event decreases at each tier. The following chains illustrate the tiers used by each experiment to deliver data to the analyses:

- CDF: RAW → DST → PAD
- DØ: RAW → STA → DST → Micro-DST

Below are included the names and typical size of each event for each event type. Additional information is included in Appendix B.

1. Data Type 1. Raw detector information.
   - **CDF, Name: RAW , Size:130kB/event** - Includes all detector information.
   - **DØ, Name:RAW , Size:600kB/event** - Includes all detector information.
2. Data Type 2. Reconstructed information.
   - **CDF, Name:DST, Size:190kB/event** - Includes all reconstructed information plus all detector information. This is the most complete level of data information and allows a complete re-run of the reconstruction. The DST is output by the CDF production executable on the farm and the data is resident on tape.
   - **DØ, Name:STA, Size:250kB/event** - Includes all reconstructed information. This is the most complete level of data information and allows a complete re-run of the reconstruction without the need for database access. The STA is output by D0RECO on the farm and the data is resident on tape.
3. Data Type 3. Reduced set of reconstructed information.
   - **CDF, Name:PAD, Size:30-40kB/event** - This level of data is the most compressed version that is output from the CDF production. It has enough information to allow a partial re-reconstruction of the event, notably a re-reconstruction of calorimeter based quantities and refitting of SVX tracks using different alignment. The PAD data is resident on tape.
   - **DØ, Name:DST, Size:50kB/event** - This is the intermediate level of data information and will allow a partial re-reconstruction of the event, notably a re-reconstruction of calorimeter based quantities. The DST is output by D0RECO on the farm and the data is resident on tape.This DST eliminates the need for the special QCD microSTA.
4. Data Type 4. Thumbnail sketch of event information.
   - **CDF** - None.
   - **DØ, Name:μDST, Size:5kB/event** - This format is a streamlined, tightly packed, and then gzip compressed version of the DST.

DØ				CDF			
Data Type	Size/evt	Total	Comment	Data Type	Size/evt	Total	Comment
RAW	500 kB	30 TB		RAW	130 kB	8.3 TB	
STA	250 kB	15 TB	Reco-able	DST	190 kB	12 TB	Includes RAW
Stream STA	250 kB	10 TB	10 Streams	Split DST	190 kB	16 TB	27 Datasets
DST	50 kB	2 TB	Semi Reco-able	PAD	32 kB	2 TB	
Stream DST	50 kB	3 TB	14 Streams	Split PAD	32 kB	2.7 TB	39 Datasets
μDST	5 kB	0.2 TB	Virtual Streams	-	-	-	no equivalant
ntuple			Analysis dep.				Analysis dep.
Total		55 TB	42 M events			41 TB	64 M events

Table 8: Summary of Physics dataset sizes for Run IB.

A.2 Run I Dataset Sizes and Access Patterns

A summary of dataset sizes for Run Ib is given in Table 8. These sizes set the scale for access times, and thus largely influence the access patterns of analysis. A significant effort was spent trying to understand the patterns employed in Run I for several physics analysis. Following are summaries of several of these analyses for DØ and CDF, with the detailed reports included as Appendicies. The reports include answers to the following questions:

1. What events do you look at? a.data, b. support/calibration/background, c. simulation.
2. How do you select those events?
3. What processing steps are needed for the event selection?
4. What processing steps are needed for the event analysis?
5. How many times and how often do you access the events?
6. Which pieces of each even are needed?

These questions were interpreted in slightly different ways by various analysis authors, but provide a basis to understanding the trends.

A.2.1 Dataset Sizes for D0

To provide quick access to STA and DST type information event event streaming to smaller datasets was performed. Table A.2.1 indicates the nature of each stream and the sizes for both STA and DST, A "-" means the stream did not exist in this format This streaming was performed on several platforms using a package referred to as "omni_filter". It employed filter codes provided by the physics groups which were linked and run together in production processing.

Originally there was a Really Good Events stream which was 40% of total and which was abandoned in favor of the high Pt leptons. There is about a 30% overlap in the streams for the DSTs. Each stream was typically 5% of the whole sample - except for leptons and QCD jets which are much larger. This 5% number was a target determined to be the size of data set which could be accessed using reasonable resources in the time frame of two weeks to a month.

In summary, the techniques employed at D0 were directed at delivering data to analysis efforts quickly at the expense of having multiple copies of each event. This strategy was straightforward,

Name	Description	% of total	STA (TB)	DST (TB)
ELF	Electron like	27	-	0.54
HLP	High Pt lepton	15	3.0	0.30
B1M	Single μ (b-physics)	11	-	0.22
TBG	Top background	10	-	0.20
QCR	Rapidity gap	7	1.4	0.14
MU1	Single μ	6	1.2	0.12
QGA	Photons	4	-	0.08
TPJ	Top $\rightarrow$ jets	3	0.6	0.06
B2M	$\mu\mu$	4	0.8	0.08
PI0	π^0	3	-	0.06
NPA	New phenomena	3	0.6	0.06
MIN	Minimum bias	1	0.2	0.02
BSM	Small Angle μ	2	0.4	0.04
MET	Missing E_t	6	1.2	-
TAU	Tau	3	0.6	-

Table 9: D0 Dataset Streaming.

but was complicated by the fact that DØ ran at least 3 versions of production in Run I. Analysis examples were provided for several physics analyses and detailed reports are included in unrefined form as Appendix D.

A.2.2 Access Patterns for D0

All of the major datasets were available by tape staging on the D0 File Server (DØFS) cluster. The μDSTs were kept on D0FS and read over the network to the analysis clusters. They were virtually streamed (the file, record and byte offset are kept for each event, and a streaming map allows quick access) into datasets, however some groups copied the selected uDST events to their own workstations rather than deal with network bottlenecks. The STA and DST were physically streamed on tapes and people "picked" events from unstreamed STA or DST datasets if necessary. Most of access was to the uDST because it is easier. The top group used all the uDSTS and selected their own sample rather than use the virtual streams. They went over the sample approximately twice.

The Run I metadata (data about the data) was contained in a production database (a relational database) and FATMEN (a hierarchical database) which, together, occupied a little over 2 Gb. These provided access to event data at the granularity of run and event range to about 500 events. The D0 Direct Access to Data (D0DAD) catalog provided information about trigger and filter as well as access at the single event level of granularity; this database was about 4Gb.

In summary, most of the analyses efforts were performed using primarily DST or μDST data sets with emphasis being placed creating ntuples with essential event information needed to study distributions or for final event selection. For many of the analysis, the entire DSTs sample of their data set was accessed from once every couple of months, to twice a year. Access to STAs was primarily through the pick event facility which selected an average of about 20K events on 10K tapes each month over the period of heaviest data anlysis (centered toward the end of the run 1b period, near the middle of the run). Picking events from particular streams greatly expedited this by requiring fewer tape mounts, but this was largely transparent to the user. Development of some inteligent software to minimize tape mounts further could make this even more efficient. Most of the efforts performed calorimeter energy (CAFIX) or particle ID (CLEANEM, CLEANMU) corrections to the data in the process of transferring to ntuple.

The primary constraints were centered in the following areas:

1. **Time** - An effort which requires 1 to 2 months to access information is considered major, but managable. Some analyses required DST information to create their ntuples, going through all the DSTs required this kind of effort. A typical time to go through all μDSTs was about 2 weeks making this a fairly common procedure to produce ntuple data sets. Some efforts used the virtual streams to go through the μDSTs and others preferred creating physical subsets of the events interesting to them. Picking STA events was time consuming, though the facility was tremendously improved over the course of the run and a great deal of attention was placed on minimizing the number of tape mounts by pooling user requests and allowing the user to specify particular STA streams in which the hit rate of finding their events would be high. Tape acces is, in general , considered very slow and this is of course true. 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 involved with serial media.

2. **Disk Space** - The combined Run I μDST data set was stored on disk on D0FS requireing 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. Many ntuple samples were maintained on disk the largest being the QCD jet sample which was made from μDSTs, it was 200GB,took only two weeks to create. One QCD analysis (BFKL on 630GeV data) had about 15 millionevents and required the DST to make the ntuples, this was only done about once every 3-6 months.

3. **Software Configuration** - Careful preparation and testing of production software was a major impediment. Minor trigger changes online or bug fixes to the reconstruction might require several months to trickle down to the final filtering and analysis efforts. It was understood that strict versioning control was required but, perhaps some streamlining would have helped to make changes to online trigger info automatically be transferred to offline.

A.2.3 Dataset Sizes for CDF

In Run Ib CDF recorded 64 million events in the main data stream (Stream B) which were processed in real-time and made available to the physicists within about 2 weeks of the data being taken. In addition, 3.7 million events out of the total were also recorded in a separate "express line" stream (Stream A) and processed within hours. This stream contained the high-transverse momentum leptons necessary for the top quark search. It was also used to monitor data quality during the run.

To provide quicker access to DST and PAD type information, physical event streaming to smaller datasets was performed, for both Stream A and B. Table V indicates the nature of each stream and the sizes for both DST and PAD, A "-" means the stream did not exist in this format. There were 27 DST datasets and 39 PAD datasets. The DSTs from the main data stream are not heavily used in analysis. The total data volume for Run I was 41 TB of which

16 TB was due to inclusion of raw data on the inclusive and split DSTs. The information about where tape a file is located is stored in FORTRAN indexed files.

Copies of selected datasets were made for distribution to remote institutions. In Run Ib about 17000 8mm tape copies were made. About 3600 of these were kept on site at Fermilab for use by physicists in the CDF Portakamp complex and the rest were sent offsite.

A.2.4 Access Patterns for CDF

In Run I CDF followed a model where the data and the CPU are tightly coupled. There was no high-bandwidth access from desktop systems to the central data stores.

The production datasets are taken by the physics groups and futher split or compressed. This can be a lengthy step and difficult to recreate if a problem is discovered later. This is an I/O intensive job and is usually performed on the central systems (usually by a small number of people) where there is efficient access to large production datasets on 8mm tape. The resulting datasets are stored on disk or if they are too large for disk they are placed in the STK tape robot.

The users doing physics analyses perform further selections of these datasets, normally on the central systems. The result of these event selections is typically an ntuple. These ntuples are usually small enough to be transferred to the desktop system for further analysis. Most physics analyses start from the PAD data. However a small number of analyses such as the W-mass measurement require the DST data. All analyses may require the full DST for a small number of events. The amount of data used from the 32 kByte PAD event ranges from 1 Kybte for exotic searches to about 10 kByte for a typical b-quark analysis.

Analysis examples were provided for several physics analysis. Detailed reports are included in unrefined form as Appendix C. We summarize some of the findings below. Obviously the access patterns vary depending on the analysis.

The QCD analyses typically look at a large number of events but only a small amount of information per event. They create general purpose QCD ntuples from the production PAD datasets (typically 100 tapes -streams QJ0B, QJ7B,QJ5B,QJ2B). These were created three times in Run I. These ntuples are then accessed many times.

The b-physics group created "BPADS" which were 30% smaller than the standard production PAD. The input to this process was 8 streams (MUCB, MULB, ECLB, PSIB, BHMB, BLMB, BYSB, BEMB - about 40% of all the data) and the output was 12 streams. They reapplied the Level 3 trigger requirements on the production output. They also compressed the jet information and further compressed the calorimeter and track information. The BPADS were stored in the STK tape robot. These files were used as the starting point for all b-physics anaylses. One of the major headaches in creating these datasets was the inadequate cataloging tools.

An example of an exotic physics analysis that had very different access patterns between Run 1A and Run 1B is the SUSY search using low P_T dileptons. In Run 1A the analysis had to use 1000 DST tapes containing low P_T electrons and muons to select a sample. This was very painful and took a long time. In contrast, in Run 1B there was a level 3 trigger for this sample and the necessary information was kept on the PAD format so the sample could be selected from 60 PAD tapes. The Run II situation may well be more like Run 1A where we are not sure exactly what data samples or detector information we need to do do a certain piece of physics.

Name	Description	% of total	DST (TB)	PAD (TB)
Stream A				
ECMA	High Pt Central Electrons	18	0.12	0.02
EPMA	High Pt Plug Electrons	4	0.03	0.004
EWZA	W and Z events	16	0.11	0.02
MUOA	High Pt Central Muons	18	0.12	0.02
PSIA	$J/\Psi \rightarrow \mu\mu$	39	0.26	0.04
EXOA	Missing Et and exotic triggers	6	0.04	0.007
QJTA	High Pt Jets	4	0.03	0.004
QSEA	Total transverse energy triggers	3	0.02	0.003
QPHA	Photon triggers	5	0.03	0.006
Stream B				
QJTB	Jets	11	0.9	-
QSEB	Total transverse energy triggers	3	0.3	-
PHOB	Photon triggers	20	1.7	-
MUCB	High Pt Central Muons	6	0.5	0.09
MULB	Low Pt Central Muons	5	0.4	0.07
MUFB	Forward Muons	11	0.9	0.13
ECMB	High Pt central electrons	4.5	0.4	0.07
ECLB	Low Pt central electrons	11	0.9	0.16
EMPB	Plug electrons	13	1.1	0.2
PSIB	$J/\Psi \rightarrow \mu\mu$	8	0.7	0.1
BHMB	High mass $\mu\mu$ pairs	7	0.6	0.04
BXXB	Low mass $\mu\mu$ pairs, Υ, $B \rightarrow e\mu$	5	0.4	-
XOXB	Missing Et	5	0.4	0.09
XTAB	Taus	7	0.6	0.1
XMBB	Minimum bias	4	0.3	0.06
XDLB	Dilepton triggers for exotics	8	0.7	0.1
QJ0B	Jets ET 100 GeV	6	-	0.08
QJ7B	Jets ET 70 GeV	1	-	0.01
QJ5B	Jets ET 50 GeV	2	-	0.01
QJ2B	Jets ET 20 GeV	2	-	0.03
QJFB	Two forward jets	2	-	0.03
QSJB	Top $\rightarrow$ multijets	2	-	0.03
QSEB	Total Et $>$ 175 GeV	2	-	0.02
GHIB	Photons Et 23 GeV, isolated	7.5	-	0.1
GLOB	Photons Et 10 GeV, isolated	3.2	-	0.04
GNIB	Non-isolated photons	3	-	0.04
GGMB	Diphotons	3	-	0.04
GMUB	Photon + muon	4	-	0.06
BLMB	Low mass $\mu\mu$	3	-	0.04
BYSB	$\Upsilon \rightarrow \mu\mu$	1	-	0.01
BEMB	$B \rightarrow e\mu$	1	-	0.01

Table 10: CDF Dataset Streaming.

B Data Type Bank Contents

B.1 Data Type 1. Raw detector information.

- **CDF, Name: RAW , Size:130kB/event** - Includes all detector information.
- **DØ, Name:RAW , Size:600kB/event** - Includes all detector information.

B.2 Data type 2. Reconstructed information.

- **CDF, Name:DST, Size:190kB/event** - Includes all reconstructed information plus all detector information. This is the most complete level of data information and allows a complete re-run of the reconstruction. The DST is output by the CDF production executable on the farm and the data is resident on tape. It contains the following information:

CTCS DEDX QTRK QVTL	Tracking (incl. compressed info)	17 kBytes
VTVZ EVTX	Vertex information	0.5 kBytes
QSVC SVXS SVXR SVXL	SVX tracking (incl. compressed info)	13 kBytes
SPTE SETA QTOW	Calorimeter (incl. compressed info)	8 kBytes
CMUO CMIO CMTS CMTL CMUS	Central muons	2.5 kBytes
FMUO FMUS FMUL	Forward muons	2 kBytes
FMUR FMSX FMUX		
METS QMET	Missing Et	0.4 kBytes
PESS CESS CASS CASL CESL	strip chambers and pre-radiator	6 kBytes
PESL CCSL CESQ CPRQ PESQ		
PEAQ PHWQ FHXQ FEAQ		
CLSL CALL CALS	Clustering information	7 kBytes
JETS DETS L2JT	Jets	2 kBytes
ELES ELCE ELCA ELTS	Electrons	1.8 kBytes
ELTL ELEC CATL CENX CEMO		
TAUO TTRS TCAE TAUR	Taus	0.2 kBytes
BBCQ BBLU TTLC TNNC	Trigger/luminosity information	2 kBytes
TAGZ TCMC TBMQ TCMQ CFWQ TL2Q		
TL3Q		

- **DØ, Name:STA, Size:250kB/event** - Includes all reconstructed information. This is the most complete level of data information and allows a complete re-run of the reconstruction without the need for database access. The STA is output by D0RECO on the farm and the data is resident on tape. It contains the following information:

CAEQ	compressed calorimeter hits	20 kBytes
DHIT	compressed CDC hits	25 kBytes
FHIT	compressed FDC hits	40 kBytes
THIT	compressed TRD hits	8 kBytes
VCHT	compressed VTX hits	41 kBytes
MUOH	compressed MUO wamus hits	12 kBytes
SAMH	compressed MUO samus hits	12 kBytes
DTRK	reconstructed CDC tracks	7 kBytes
FDCT	reconstructed FDC tracks	1 kBytes

VTXT	reconstructed VTX tracks	2 kBytes
TRGR	trigger bank	20 kBytes
L2EM	level 2 EM bank	3 kBytes
CAID	calorimeter hot cell bank	0 kBytes
JUTL	jet utilities bank	0 kBytes
	"base"DST	20 kBytes
	Total	210 kBytes

B.3 Data type 3. Reduced set of reconstructed information.

- **CDF, Name:PAD, Size:30-40kB/event** - This level of data is the most compressed version that is output from the CDF production. It has enough information to allow a partial re-reconstruction of the event, notably a re-reconstruction of calorimeter based quantities and refitting of SVX tracks using different alignment. The PAD data is resident on tape. It contains the following banks:

LUMD BBLD BBCQ BFLD LRID EVCL EVCD	Luminosity	0.6 kBytes
VTVZ EVTX	Vertex	0.5 kBytes
CCSL CLSL	Clustering information	0.5 kBytes
JETS DETS	Jets	2.6 kBytes
ELES ELEC ELCE CATL	Electrons	1.7 kBytes
CMIO CMUO CMUD CMPD CMXD CSPD CSXD CMCR	Central Muons	6.6 kBytes
FMUO FMCE FMTE FMSX FMUX	Forward muons	1.7 kBytes
QTRK QSVC DEDX SVXK	Tracking	6.4 kBytes
TAUO	Taus	0.05 kBytes
METS QMET	Missing Et	0.4 kBytes
TBMQ TCMQ CFWQ TL2Q TL3Q TAGZ	Trigger	1.4 kBytes
QTOW CESQ CPRQ PESQ PEAQ PHWQ FHXQ FEAQ	Calorimeter	12 kBytes

- **DØ, Name:DST, Size:50kB/event** - This is the intermediate level of data information and will allow a partial re-reconstruction of the event, notably a re-reconstruction of calorimeter based quantities. The DST is output by D0RECO on the farm and the data is resident on tape.This DST eliminates the need for the special QCD microSTA. This data level is the "base" DST augmented with the following banks:

CAEQ	compressed calorimeter hits	20 kBytes
DTRK	reconstructed CDC tracks	6 kBytes
FDCT	reconstructed FDC tracks	1 kBytes
CAID	calorimeter hot cell bank	0
JUTL	jet utilities bank	0
TRGR'	other	1 kBytes
" base" DST		20 kBytes
	Total	50 kBytes

List of all the banks on the" base" DST

CACL	MHTT	ZTRK	PROC	FTRH	HMTE	PTAU
JETS	MUCA	JNEP	PPHO	HMTP	T0TH	TTRH
ESUM	PNUT	MUCD	FDCT	SAMT	SATS	STTH
CASH	MUD1	GLOB	ERMG	DTRH	SATN	MSEG
PMUO	MUOT	HEAD	VFIT	VERH	L2CD	JTCS
MUON	CATD	VTXT	FILT	VTRH	STNB	C2EM
PARH	JAUX	DTRK	MTRH	PELC	STSB	
CAPH	HSTR	VERT	PLV0	L0VT	STSA	
ZFIT	TPRL	T0TD	TRDT	HSUM	STNA	
JPAR	FRES	TSUM	ZTRH	RECO	VCOR	

B.4 Data Type 4. Thumb-nail sketch of event information.

- **CDF** - None.
- **DØ, Name:μDST, Size:5kB/event** - This format is a streamlined, tightly packed, and then gzip compressed version of the DST.

C CDF Physics Analysis Examples

C.1 B-physics - collected by Liz Buckley-Geer

Liz looked at two Run 1B analysis topics, one involving J/ψ and the other inclusive leptons. General comment - the B-group seemed happy to have their data in the STK robot. This is good seeing as we will have much more of the data in a robot in the future.

C.1.1 $B \rightarrow J/\psi X$ where $X = K, K^*, K_s$

1. What events do you look at?
   (a) data
   Stream A J/ψ sample. PAD format.
   (b) support/calibration/background
   No special samples. Use J/ψ sidebands for background
   (c) simulation
   Generate 10^6 $B \rightarrow J/\psi \rightarrow \mu\mu$ using Pythia + QFL (CDF fast simulation). Reject about 90% of generated events that fail cuts.
2. How do you select those events?
   Start with J/ψ sample, apply standard muon finding code, vertex muons to make J/ψ and add additional tracks, apply cuts to fitted parameters.
3. What processing steps are needed for the event selection?
   Muon, tracking, SVX beam position
4. What processing steps are needed for the event analysis?
   Look at soft electron tags in B-decays - needs calorimetry Also muon and tracking

5. How many times and how often do you access the events?
   Make ntuple. Do new strip once every few months. Has run > 10 times on full J/ψ sample. Sample is on local workstation disk.

6. Which pieces of each event are needed?
   Tracking (SVX, CTC)
   Muons
   Beam position
   Luminosity
   Calorimetry
   dE/dx
   Central strip chambers
   Central preradiator
   The last four are needed for soft electron tagging

For Run II this analysis would use all types of particle ID that are available in CDF.

C.1.2 B flavor tagging (soft electron tagging and jet charge tagging)

1. 1. What events do you look at?

   (a) data
   Low pt Inclusive electron, muon and eμ samples. BPAD format (see below) 6×10^6 , 2×10^6 μ, 500,000 eμ

   (b) support/calibration/background
   Some min bias data and jet data (low pt jet $E_T > 20$ GeV) Also use electron and muons samples with looser cuts. Will be switching to Stream C muon triggers for muon fake backgrounds

   (c) simulation
   Pythia + QFL Used to obtain beta/gamma correction factor for lifetime. About 50,000 events per trigger type.

2. How do you select those events?
   Start with muon and electron samples. Filter on trigger lepton. Apply SVX vertex tagging, associate tag with lepton. Will be switching to track based jet finding to avoid biases with muon. About 80,000 e and mu events survive, 15000 eμ events.

3. What processing steps are needed for the event selection?
   Tracking, SVX beam position and alignment, electron, muon, calorimeter jets. item What processing steps are needed for the event analysis?
   Tracking and tagging info Soft lepton tagging - dE/dx, strip chamber, preradiator.

4. How many times and how often do you access the events?
   One analysis uses CW ntuple, the other RW ntuple. Selection to make primary e and μ ntuple done about twice so far. eμ about 5-6 times. Samples have been used 20-30 times to check the effects of cuts and for background studies. Take 4 hours (not including processing time) to read 6×10^6 events if they are already staged from STK onto disk. Typically takes about 1 week to remake ntuple for muons and longer for electrons.

5. Which pieces of each even are needed?
   Tracking (SVX, CTC)
   Vertex
   Electrons
   Jets
   Muons
   Beam position
   Luminosity
   Calorimetry
   dE/dx
   Central strip chambers
   Central preradiator
   The last four are used for soft electron tagging

C.1.3 Creation of BPADS

The B-group created their own data sets from the production output, these are called BPADS.

9 input streams - electron, muon (high and low pt), J/ψ, $\mu\mu$ (high mass and low mass), Υ, eμ

12 output streams - above + $K^{\star}\gamma$, sequential μ, 2 mass windows for J/ψ

Applied L3 trigger requirements again on offline production quantities. Tightened muon cuts.

Made some specific b-banks which were not used very much, mostly due to lack of unpacking routines. Compressed jet info and calorimetry and tracking further.

All events stored in STK robot.

Most of the pain was related to dealing with the CDF production_manager database and it would have been useful to be able to specify a file sequence number on tape in the analysis driver.

C.2 Exotic analyses - collected by Stephan Lammel

Stephan reported on the SUSY dilepton analysis and the missing Et analysis for both Run 1A and Run 1B.

C.2.1 SUSY dilepton - Run 1A

1. What events do you look at?
   (a) data
       Low pt electrons and muons. DST - about 1000 tapes
   (b) support/calibration/background
       trigger efficiency imported single lepton sample - 300 tapes $\rightarrow$ 10 tapes
   (c) simulation
       80 tapes (8500 format)
2. How do you select those events?

3. What processing steps are needed for the event selection?
   Electron/muon information. Tracking, vertex and calorimetry. About 2 tapes output from 1000 input.

4. What processing steps are needed for the event analysis?
   jets, missing E_T

5. How many times and how often do you access the events?
   multiple passes over about 100 events.

6. Which pieces of each even are needed?
   Electron
   Muon
   Tracking
   Vertex
   Calorimetry

C.2.2 SUSY dilepton - Run 1B

A trigger was added in run 1B and the necessary information was available on the PADS. This made things much easier

1. What events do you look at?
   (a) data
   Exotic dilepton PADS - about 60 tapes
   (b) support/calibration/background
   $J/\psi \rightarrow \mu\mu$ PADs (300k events)
   $\Upsilon \rightarrow \mu\mu$ PADs (50k events)
   $Z \rightarrow \mu\mu$ PADs (3K events)
   $Z \rightarrow ee$ PADs (5k events)
   Top $\rightarrow \ell\ell$ (10 events)
   (c) simulation
   See above

2. How do you select those events?

3. What processing steps are needed for the event selection?
   Electron/muon information. About 12 tapes output. item What processing steps are needed for the event analysis?
   Tracking, vertex, jets, missing E_T.

4. How many times and how often do you access the events?
   12 tapes have been read about 15 times so far and ntuple made (1 kByte/event). Background studies - ntuple, 15 versions. Read about 2000 times.

5. Which pieces of each even are needed?
   See above

C.2.3 SUSY missing E_T - Run 1A

1. What events do you look at?

   (a) data
   Missing E_T DSTs - 350 tapes

   (b) support/calibration/background
   Jet 50 DSTs (65 tape → selection → 1000 events
   high-pt e PADs (disk)
   high-pt μ PADs (disk)
   Jet 20 DSTs (128 tape) → selection → 500 events
   Jet 100 DSTs (15 tapes) → selection → 300 events
   Jet 70 DSTs (15 tapes) → selection → 200 events

   (c) simulation
   Signal - 1 million events
   W+jet - 50k events
   Z+jet - 50k events
   top - 20k events

2. How do you select those events?

3. What processing steps are needed for the event selection?
   Calorimetry, Missing Et, jets → 35 tapes vertex and trigger → 5 tapes

4. What processing steps are needed for the event analysis?
   Tracking, muons electrons - less that 1 tape - on disk

5. How many times and how often do you access the events?
   Read 5 tapes 50 times. Read disk sample about 100 times.

6. Which pieces of each even are needed?
   Calorimetry
   Missing Et
   Jets
   Tracking
   Muons
   Electrons
   Vertex
   Trigger

C.3 Information collected by Wolfgang Kowald

```
TOP

DATA :

   I  : Stream A inclusives  : STRA_5F         DST  : 143 tapes (600 GB)
```

```
                                                    (trailer,vault)

   II : Stream B jet samples : QJ2B_5P,QJ5B_5P
                               QJ7B_5P,QJ0B_5P PAD  : 90 tapes  (180 GB)
                               XMBB_5P              : enough

      FNALD, Trailer workstations
      Vaulted/Library tapes
      Started with split datasets. Restart on inclusives.

 MC   :

   ttbar 19 different masses : 30 GB ( stripped 6 GB )

   VECBOS                    : ???

-----------------------------------------------------------------------------

 I :

 First stage selected datasets (Selection : L3 trigger bits) :      PADs

   EIA : 13 GByte, 500,000 events
   MUA : 16 GByte, 650,000 events
   PWA :  5 GByte
   EWA :  6 GByte
   EZA :  8 GByte

 Second stage (Selection : good high pt lepton )                    PADs

   Incl. electrons : 5 GByte, 125,000 events
   Incl. muons     : 4 GByte,  89,393 events

 Third stage (Selection : isolation, trigger, good run,
                          Z mass window, no dilepton, MET > 20 )    PADs

   Incl. W's       : 3.6 GByte, 81,300 events
   Incl. Z's       : 3.8 GByte,

 Fourth stage (Selection : jet activity )                           PADs

   W+jets          : 0.5 GByte, 10,658 events

=============================================================================

 Inclusive jets/ dijets
```

```
   DATA : QJ2B_5P,QJ5B_5P,QJ7B_5P,QJ0B_5P      70(92) tapes (140 GByte)
          XMBB_5P                            ??(68) tapes

      FNALD
      Vaulted tapes
      Ran tree times on dataset. ( Error in ET definition )


   MC   : ??
```

```
  Selected datasets (Selection + Analysis) :                    NTUPLES

   NTUPLE : 1.3 GByte

   Used for inclusive jet and dijet analyses.
```

==

```
  Total Et cross section

  DATA : QSEB_5P                                   40 tapes (80 GByte)

       cdfsga, trailer workstation
       Vaulted/Library tapes
       Ran once
       Selection and cuts tuned on disk files from 1A
```

```
  Selected datasets ( Selection + Analysis : JETS, METS,          NTUPLES
                    VTVZ, HATFLT )

   NTUPLE : Run 1A - 50 Mbytes Run 1B - 0.2 Gbytes
```

===

```
  Charged Higgs

   DATA : X0XB_5P,(X0X1_3P)       73 1B tapes, 41 1A tapes   (~ 120 GByte)

      Trailer Alpha cluster
      Library tapes
      Cross checks with top samples
```

```
MC :

   100 pb**-1 (6 alphas for 2-3 months at Rutgers)

--------------------------------------------------------------------------------

First stage (Selection : TAU0, MET,                              PADs
             Ran production in order to fix tracking)

   PADs : 36 GByte (8 tapes)


Second stage (Selection + Analysis : ? )                         BINARY

   BINARY : 2.6 GByte ( ~700000 events)
   Ran two to three times on PADs.

Third stage (Selection + Analysis : ? )                          NTUPLES

   NTUPLES : ?


Comments:
             - Tapes drives on FNALD very bad. Used local drives.
             - Binary files manageable. Good planning necessary.
             - Checks with top sample important.
             - Most exotics used top sample.

===============================================================================

W/Z + jets


DATA : ELEX_3P,MUOX_3P tapes
       1B Incl. lepton top sample on disk
                               ( 13 GByte + 16 GByte )

   Duke (1A)
   Trailer cluster (1B)
   Ran 2 to 3 on each dataset (tracking, EM corrections ...)

--------------------------------------------------------------------------------

Selected Z datasets (Selection + Analysis : most banks )        PADs,NTUPLE

   z-ee   : PADs        .520 GByte
```

```
    Z-uu   : PADs        .153 GByte ??
    Z-ll   : NTUPLE      .015 GByte

 Selected W datasets (Selection + Analysis : most banks )       PADs,NTUPLE

    incl e : NTUPLE        .5  GByte
    W      : PADs         5.0  GByte
    W      : NTUPLE       0.07 GByte
```

==

NUMBER OF TAPES AND DATASET SIZES FOR RUN Ib DERIVED FROM TAPE COPYING INFORMATION (- : not tapes were copied) :

Dataset		tapes	GByte	x20
ECMA_F	:	45	90	1800
EPMA_F	:	45	40	800
EWZA_F	:	45	90	1800
EXOA_F	:	45	50	1000
MUOA_F	:	45	120	2400
PSIA_F	:	61	250	5000
QJTA_F	:	45	30	600
QSEA_F	:	46	30	600
QPHA_F	:	45	30	600
STRA_F	:	143	580	11600
ECMA_P	:	38	17	340
EPMA_P	:	38	17	340
EWZA_P	:	37	19	380
EXOA_P	:	40	-	-
MUOA_P	:	37	24	480
PSIA_P	:	37	48	960
QJTA_P	:	25	-	-
QSEA_P	:	40	-	-
QPHA_P	:	40	7	140
STRA_P	:	40	140	2800

Dataset		tapes	GByte	x20
BHMB_F	:	332 ^	1200	24000
BXXB_F	:	332 v	"	"
ECLB_F	:	288	1130	11600
ECMB_F	:	209	440	8800
EPMB_F	:	229	900	18000
MUCB_F	:	493 ^	1500	30000
MUFB_F	:	493 |	"	"
MULB_F	:	494 v	"	"

```
   PHOB_F  :  384      -              -
   PSIB_F  :  169     660          13200
   QJTB_F  :  335 ^ 1060          21200
   QSEB_F  :  335 v    "              "
   XDLB_F  :  172     690          13800
   XMBB_F  :  178     200           4000
   XOXB_F  :  338 ^  790          15800
   XTAB_F  :  338 v    "              "
   STRB_F  : 2107      -              -
---------------------------------------------
   BHMB_P  :   39 ^  150           3000
   BLMB_P  :   39 v    "              "
   BEMB_P  :   34 ^   40            800
   BYSB_P  :   34 v    "              "
   ECLB_P  :   45     180           3600
   ECMB_P  :   19      70           1400
   EPMB_P  :   46     180           3600
   GHIB_P  :   48 ^  170           3400
   GLOB_P  :   48 |    "              "
   GNIB_P  :   46 v    "              "
   GGMB_P  :   34 ^   80           1600
   GMUB_P  :   34 v    "              "
   MUCB_P  :   77 ^  310           6200
   MULB_P  :   77 |    "              "
   MUFB_P  :   69 |    "              "
   MUTB_P  :   18 v    "              "
   QJOB_P  :   87 ^  170           3400
   QJ2B_P  :   88 |    "              "
   QJ5B_P  :   88 |    "              "
   QJ7B_P  :   87 |    "              "
   QJFB_P  :   87 v    "              "
   QSEB_P  :   34      70           1400
   QSJB_P  :   34      -              -
   PSIB_P  :   24     100           2000
   XDLB_P  :   71 ^    -              -
   XOXB_P  :   73 |    "              -
   XTAB_P  :   73 |    "              -
   XMBB_P  :   68 v    "              -
   STRB_P  :  357      -              -
------------------------------------------------------------------------------
```

D DØ Physics Analysis Examples

Some representative DØ analyses have been examined ranging from large event sample QCD-type to small sample search-type activities.

Summary of Analysis Access patterns

D.1 SUSY Gauginos to Tri-leptons and MEt (Doug Norman)

1. What events do you look at?

I have used primarily collider data in the form of DST's and UDST's. These DST's and UDST's consist of the entire collider data set. I process smaller subsets of these complete data sets. These data subsets are defined by loose cuts and can consist of several thousand events. I also use much smaller (100) events in the form of STA's. These events most closely match the search criteria of my analysis and are used in event displays. The DST and UDST are processed to make ntuples that are stored on disk.

For trigger studies I have used large STA event samples, because the needed information was not available on the DST's. (this is a pain)

For background studies I have also used raw data (minimum bias data). I store 1000 such events on disk.

For simulations I require a physics generator that produces particle information, a detector simulator that produces raw data (typically very large files) which usually require data bases (very large showerlibrary), and a reconstruction program that produced DST's. These DST's are then turned into ntuples; there is not enough disk space to store all the dst's that we need much less the geant files. Everything but the ntuples get written off to tape.

2. How do you select those events?

The events are selected using D0DAD files which contain a "list" of run and event numbers of those events that pass a specific (typically loose) event selection (filter). My ntuple maker, for example, when using the D0dad file will only process of the entire data set only those events that pass the filter that I designed. The D0dad files are created in a process using several filtering algorithms, one d0dad file per algorithm. This way we avoid having to run through the entire data set to get to just the events that we want. This helps if we are required to process the data more than just a couple of times.

Sta's are selected using a PICK utility which require run and event information. One or more events can be "picked" at a time. These picked events are staged to a disk where one can copy them to ones own storage area and use in, for example, events displays.

The raw data I've selected using logbook information and staged to disk using the Fatmen utility.

3. What processing steps are need for the event selection?

A) Prefiltering to make d0dad files (this can require more than one step)

B) creation of ntuple which can be used to select events for STA picking based on physics selection cuts

C) STA picking

D) raw or STA data staging to disk from tape (or other)

4. What processing steps are need for the event analysis.

A) raw data needs to be reconstructed to make STA's and DST's

B) DST's which contain physics objects (electrons, muons, jets...) with some set of quality descriptors usually need to have these objects corrected for energy scale corrections (CAFIX), or have additional descriptors calculated (CLEANEM,CLEANMU...) when processed to make ntuples. This additional processing doesn't usually lead to the create of other DST's.

C) Ntuples are examined using PAW and selection proceeds with the use of COMIS routines.

For simulated data the steps are : ISAJET (for example), GEANT, TRIGGER SIMULATION, RECONSTRUCTION, NTUPLE + PAW

How many times and how often do you access the events?

Ntuples: too many to remember

DST's: 6-12 times, once every one-two months

STA's: 6-12 times, once every one-two months

raw: 1 or 2 times , very seldom

Simulated data (dsts): Many times

6. Which pieces of each event are needed?

Not sure how to answer this. I've had to use trigger info, raw calorimeter information, processed data (electrons, jets, tracks, muons, missing Et, processed calorimeter cell data...), ISAJET data (particle info). In short everything.

D.2 SUSY di-lepton (John Womersley)

1. What events do you look at? a. data b. support/calibration/background c. simulation

a. all events with photon triggers; about 2M for run I, from about 5 L2 filters plus special runs b. was done using the same data sample as a. c. plate level (detailed GEANT) simulations for detailed understanding of photon efficiencies + shower library GEANT for topological cuts

2. How do you select those events?

Events are selected by D0DAD file based on filter name

3. What processing steps are needed for the event selection?

An ntuple maker is run. The ntupler runs CLEANEM, CAFIX etc. and saves all events passing very loose cuts. The Ntupler is run on microDST files.

4. What processing steps are needed for the event analysis?

Ntuples are analysed on d0cha using PIAF running on d0sgi0.

5. How many times and how often do you access the events?

Ntupler was rerun O(10) times. Ntuples were accessed few hundreds of times.

6. Which pieces of each even are needed?

We use: PELC and PPHO banks (electrons, photons) EM cluster shape information and energy by layer Tracks, vertices, and hits in road in the CD Jets (various cone sizes) Missing ET Muons Trigger and microblanking info.

D.3 Di-muon (David Fein)

1. What events do you look at? a. Data - OUr analysis covered the dimuon data. Specifically we looked at dimuons in the central region, but are hoping to expand this to include an analysis which covers a muon in the central with a muon in the forward region. b. support/calibration/background - we did directly use some 'special' runs which were used to determine cosmic ray signatures and efficiencies. These were very helpful. c. simulation - We used the standard simulation and didn't find any drawbacks here.

2. How do you select those events? The events were selected using one of the physics streams (B2M). We had a significant amount of disk space available so we copied the dsts to a private area and were able to work on them from there. We found this was necessary due to the number of times we wanted to access the events.

3/4. What processing steps are needed for the event selection/analysis? After the dsts were copied, we used standard D0 softward and tools to select our events and make ntuples from this smaller data sample.

5. How many times and how often do you access the events? Initially, the events were accessed quite often while the analysis matured. This is why we transferred the events to a private area. This part of the analysis would have taken a significant amount of time if we were unable to do this.

6. Which pieces of each even are needed? We found that the information provided on the dsts were sufficient.

D.4 QCD jets (Bob Hirosky)

Here are a few comments about Run II data management from my QCD perspective. I have mainly been working with physics data in two forms uDST and DST. For many jet's analyses uDST information is sufficient and we are able to ntuple the entire run 1 dataset in about two weeks time, this is OK as we have been remaking the ntuples only about twice/year. So I am pleased w/ our present uDST scheme. If there will not be enough disk space in run 2 to keep uDST's on disk, then I would advocate some sort of physical streaming of uDST's. I think the goal should be that one can run through all uDST data for any analysis on the time scale of a couple of weeks if this is possible. I have also been accessing much DST level data via FATMEN. For the data used in the diffractive analyses; this task takes several months. Physical streaming of DST level data is ESSENTIAL for analyses like this that require cell level information. *general complaint* - clearly there are two sides on this issue In run 1 there was a great deal of inertia involved in changing the production farm code. In some cases this adversely affected certain analyses, since bug fixes were delayed for long periods. I hope we have more flexibility in run2. * This inertia also affected the streaming of data at DST level. Namely, after a trigger list was changed, it could take months for the new trigger to be aded to the proper stream in omnifilter. This can effectively make some data inaccessible for practicle purposes. I suggest we addopt the following scheme for offline (dst) streaming.

1) we define a number of fixed offline dst streams 2) each stream chooses events passing a certain trigger mask 3) this mask goes into the trigger name in a uniform way via the trigger meister/TCB - so when a new trigger is added to the list, the data's offline stream is automatically set. this mask can also choose the level of the stream (ie dst or STA equivalent) w/ this scheme there will be no delay in streaming waiting for the production code to be updated - certain streams had this functinalitly in run 1, but not all.

*) special streams can also be assigned to physics groups for special needs ie, I can imagine trigger which will need 10's of picobarns of data. This is too much Lum for special runs, but it could be taken in a few weeks of running globally. It would be very usefull to have this data go to something like a qcd_special offline stream for analysis. ie for a certain run range this stream whould be known to contain the "special trigger" info only - this offers all the convenience of having special run data in it's own private set of tapes, but it allows a trigger to accumulate a reasonable amount of luminosity.

D.5 QCD jets BFKL and Low energy data (Terry Heuring)

Most of the data for BFKL and 630 GeV analyses were taken during special runs which dictated how we processed and accessed our data. Since we were not part of the standard data taking, our data was typically located on a small number of contiguous tapes. Due partly (but certainly not solely) to the fact that we wrote a huge amount of data in a short time (15-25Hz), we sometimes experienced delays in reconstruction since the farm was designed with the standard daq rate in mind (1-2Hz). This also had the effect of reducing our event size down as we read out only those detectors we planned to use in our analyses.

Once reconstruction was accomplished, we would systematically turn the DSTs or uDSTs into ntuples and base our analysis on those. Periodic changes in calibrations would cause us to have to remake our ntuples and hence reaccess the DSTs, but this was not frequent perhaps occurring once every 3-6 months.

This scenario may be repeated at some level for the upcoming run and should be considered when designing the computer systems to be used.

D.6 Central Rapidity Gap dijet (Tracy Taylor-Thomas)

1. What events do you look at? a. data Central rapidity gap dijet events from global triggers and special runs (QGT, Low luminosity runs) For global triggers, use streamed DST's - QCR stream For special runs, use DST's

b. support/calibration/background 1)Background are dijet events with a large jet eta separation. 2) calibration - zero bias data from DST's c. simulation

2. How do you select those events? Use streamed DST's - QCR stream for most analysis If streamed DST's not available (i.e. for special runs), use DST's

3. What processing steps are needed for the event selection? A D0Project was made to convert the large number of streamed DST's into a more managable number of QCD Microdst's (QMD) and QCD Microsta's (QMS). These data types keep calorimeter tower(QMD) or cell(QMS) information. Our group uses both the QMD and QMS data sets. As an example, the QCR data is on close to 100 streamed DST tapes. The QMD data made from these DST's fits on 19 tapes.

Other QCD analysis groups also use these condensed data types. The jet shapes subgroup uses the QMS data for jet triggers (the QCJ stream of DST's). This data takes around 150 tapes in DST form and 77 tapes in QMS form.

The D0Project was only run for the global triggers. The special runs are picked by filter from DST's.

4. What processing steps are needed for the event analysis? Reco with CAFIX. Ntupling (from QMD or QMS data if available). Ntuples are typically made by running through data from FATMEN. For the larger data sets, this is very time-consuming (on the order of a few months).

5. How many times and how often do you access the events? The smaller data sets (special runs and/or subsets of global data) are accessed around 3 times on average. The larger data sets are accessed less frequently. Events are accessed occasionally using the PICK_EVENTS utility.

6. Which pieces of each even are needed? Calorimeter tower and/or cell information, tracking, jet, and level 0 information.

Comments: Calorimeter tower and cell information is important for several QCD analyses. Because this information was dropped from the standard D0 Microdst for Run 1b, we had to create our own data types through a D0Project. The D0Project turned streamed QCD DST's (both QCR and QCJ stream) into QCD Microdst's and Microsta's. This processing took several months scattered over the duration of Run 1b. Using the D0Project to create our QCD Microdst's and Microsta's greatly reduced the number of tapes one has to use for a particular analysis. However, if FATMEN is used to access these tapes, using the full data set is painfully slow. We are starting to use the robotic tape device on the Challenge to access the data on these tapes. While this method is significantly faster than using FATMEN, I am not convinced that the speed will remain as high when more people start using this method (i.e. - if one had FATMEN to oneself, it would be very fast too).

It would be useful in Run II to either have all the information groups need on the standard data type (which would mean more storage space necessary to hold it all) or a faster way to access data on tape.

D.7 W Mass analysis (Eric Flattum)

1. What events do you look at?

a. data Events with two high pt electrons or a high pt electron with MEt and relatively loose electron identification requirements. The ideal data set would be sta's or dst's with raw detector information so recalibration could be done on the data set once the calibration had been done.

b. support/calibration/background This is where Run 1 data had serious draw backs. Because of calibration studies that showed we needed to correct the track center of gravity and electron position we could not treat the background data set the same way as the signal data set. For instance we could not recompute the MEt in the background data using the new vertex position because the calorimeter information was not available. Again having dst's is very important for this measurement.

c. simulation It would have been nice to have large sets of W and Z GEANT events so the W mass measurement technique could have been shown to work on the Monte Carlo before it was applied to data. This is something that Run 2 W mass people should keep in mind.

2. How do you select those events? With as loose a kinematic and electron ID requirements as possible. The limiting factor was disk space to keep the data.

3. What processing steps are needed for the event selection? The information from RECO was fine.

4. What processing steps are needed for the event analysis? Essentially we RECO'd the data ourselves. This correcting the CDC track center of gravity and finding the new vertex. With this new vertex the electron energy was recalculated using a windowing algorithm and the MEt was recalculated in a similar vein.

5. How many times and how often do you access the events? When the analysis was in full swing; several times a month.

6. Which pieces of each even are needed? For the W mass analysis you can't have enough information. You never know, maybe the TRD as some point would have been useful.

Could you provide a brief outline of the data reduction steps used in the W mass analysis. The questions may be used as a guideline and any information about what datasets were used, how long it took to access certain data sets and what constraints influenced the analysis approach would be useful.

As far as the data reduction went the simple kinematic and loose electron ID cuts reduced the data to a resonable size. But with sta's or dst's the size space used was a disk or two. The problem is with the background or control data. Since there are 1M events or more you need a reasonable way of accessing their dst's. For this data set was is needed is calorimeter and tracking information. We never even attempted to access this data. Getting the signal dst's was enough work. It took a month or two of going through the streamed data and picking events. Actually the fastest was to pick events over the X-mass break.

D.8 Top (Jeff Bantly)

1. What events do you look at? all three

a. data b. support/calibration/background c. simulation

2. How do you select those events?

3. What processing steps are needed for the event selection?

answer to 2,3 combined

data - events were streamed by the Top group by making a list of all MDS files and then passing them through our own filters. Events were written out to separate streams for each channel. Selection of events was as minimally restrictive as possible without filling available streaming disk space. Typically took a couple of months to do all of the available files. Also, STAs were picked for re-recoing and when non-MDS info was needed. for example, for proper main ring corrections to the calorimeter.

support/calibration/background - used as needed to determine good particle ID and to determine detector efficiencies. events obtained by streams or special runs.

simulation - MC was generated on almost every local and remote platform available to D0. location depended on number and nature of events being generated.

Clearly huge numbers of events were involved in all of these categories.

4. What processing steps are needed for the event analysis?

events were analyzed with results stored in ntuples typically

5. How many times and how often do you access the events?

we streamed the entire data stream twice. we streamed smaller subsets probably several times. Streamed events were probably analyzed many times as ntuples added more terms and/or new versions of CAFIX etc.

6. Which pieces of each event are needed?

everything. if MDSs weren't sufficient then DSTs or STAs were used.

References

[1] CDF Run II Technical Design Report.

[2] Liz Buckley-Geer, Stephan Lammel, Rob Plunkett, "CDF Run II Data Handling", CDF/DOC/COMP_UPG/PUBLIC/4072, February 26, 1997.

[3] DØ Run II computing Planning Board, "Run II Computing and Software Plan for the DØ Experiment", January 31, 1997.

[4] Jeff Bantly, et al,"A Proposal for the DØ Upgrade Luminosity Monitor",D0 Note 2860, February 23, 1996.

[5] From the CERN Technology Monitoring Group, available in CMS CTP (CERN/LHCC 96-45), or at http://cmsdoc.cern.ch/ftp/CMG/CTP/index.html.