Overview of the CDF Run II Data Handling System

L. Buckley-Geer[1], *S. Lammel*[1], *M. Leininger*[1], *T. Watts*[2]

[1] Fermi National Accelerator Laboratory, Batavia, Illinois 60510, USA
[2] Rutgers University, Piscataway, New Jersey 08854, USA

Abstract

The Collider Detector at Fermilab (CDF) collaboration records and analyses proton anti-proton interactions with a center-of-mass energy of 2 TeV at the Tevatron. During the next collider run the experiment expects to record about 1 PetaByte of data, an increase by more than an order of magnitude in data volume compared to the existing data of the experiment.

This paper gives an overview of the new data handling system. The design builds upon successful strategies used in the past but eliminates shortcomings encountered with the old system. The core of the central analysis system will be a pool of over 20 TBytes of data disks. Logically behind this disk pool will be an automated tape library with a storage capacity to hold the full 1 PByte of Run II data. Multi-processor compute systems will be located around this storage pool to provide the required analysis power of over 3,000 SPECint95. Desktop systems will be integrated into the analysis system via a shared login area and user authentication. The data handling software falls into four categories: software to manage computing resources, software for the data disks and for the mass storage subsystem, software to organize and manage the meta-data information, and the I/O modules in the analysis driver for accessing data.

Keywords: data handling, data storage, data access, analysis system, data catalogue, resource management, tape handling, CDF, Run II, overview, CHEP2000

1 Introduction

The Collider Detector at Fermilab (CDF) [1] is a large multi-purpose particle detector at the Fermilab Tevatron. The experiment records and analyses proton anti-proton interactions at a center-of-mass energy of $\sqrt{s} = 2\,\mathrm{TeV}$. The next collider run of the Tevatron, Run II, is scheduled to start in March of 2001. During its two year duration the CDF collaboration plans to record 1 PByte of data. This data volume is more than 20 times the volume of the previous data recording, Run I. Moreover, the data handling and analysis challenge is even larger: With the upgrade of the detector and trigger system a significantly higher purity of good events is expected. And with the growth of the collaboration over the last years and the extended physics potential of the detector an increased analysis activity with respect to Run I is anticipated.

The Run I data analysis strategy was very successful and the strategy for Run II is an evolution from it. In Run I events were sorted online into three data streams based on physics criteria. After reconstruction the events were organized into about 20 primary datasets. All data were stored on Exabyte 8mm tapes. Users extracted the events of interest to their analysis from those datasets creating new, smaller and more analysis specific datasets. This filtering of events and also information per event was repeated until the data sample could be stored on disk. Disk space was managed by the physics groups of the experiment. CPU resources were managed by batch systems on a user basis.

In Run I a DEC/VMS cluster of over 20 machines with the core of the cluster being multi-processors nodes and a large SGI/IRIX multi-processor system with about 20 CPUs formed the central analysis system. Each of the two systems had about 500 GBytes of disk space for data, access to a common Storage Tek (STK) silo with a 1.2 TBytes capacity, and tens of Exabyte 8mm tape drives to provide access to the full 50 TBytes of Run I data in the tape vault. A small IBM/AIX system added extra analysis power towards the end of the run.

While the VMS system was the main analysis system at the beginning of Run I (1991), the SGI system serviced the majority of the analysis activity towards the end of the run. The VMS system was an upgrade from the previous run and was very robust. The SGI was a new system acquired for Run I to meet the increased analysis needs of the experiment. The system was upgraded on a regular basis during the last years to increase the data disk space and to maximize its CPU capacity.

2 CDF Experiment

The Tevatron is the world's highest energy collider. It is located in the Fermi National Accelerator Laboratory [2], near Chicago. A new pre-accelerator, the Main Injector [3], was built during the last years to enable higher luminosity proton anti-proton collisions in the Tevatron. Run II, the first collider run with the Main Injector is scheduled to start in March of 2001, i.e. in about one year. The beam crossing time will be reduced from the previous $3.5\,\mu\mathrm{s}$ to $396\,\mathtt{ns}$ and later to $132\,\mathtt{ns}$. The planned luminosity for Run II is $2 \cdot 10^{32}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. The goal is to accumulate an integrated luminosity of $2\,\mathrm{fb}^{-1}$ during the two year duration of the run. For Run II the center-of-mass energy of the collisions will be increased slightly from 1.8 TeV to 2 TeV.

CDF is a rather old experiment. It recorded its first data in 1985. To handle the higher luminosity the central tracking system of CDF will be replaced completely [4]. The new tracking system will consist of a next-generation silicon vertex detector, an intermediate silicon detector, and a new open-cell drift chamber. The old gas proportional chamber-based calorimeters in the end-cap and forward region are being replaced with scintillating tile calorimeters. New muon chambers are added for increased coverage. All front-end electronics will be replaced to handle the shorter bunch-spacing.

3 Data Handling Strategies

The data handling strategy for Run II is rather simple: Since the Run I approach worked quite well, we plan to continue along a strategy of successive data filtering. The storage subsystem will stay closely coupled to the analysis CPUs of the system. However, shortcomings have also been recognized in the Run I implementation and will be addressed in the design of the new system. Additionally, the large increase in computing needs of CDF, together with changes in the data persistency mechanism, advances in computing technology, and user requests, make it impossible to re-use most of the the Run I implementation.

Storing both raw and reconstructed data on permanent storage is a necessity. With the large amount of data and a limited computing budget of the experiment, the use of serial media, i.e. tape, is unavoidable. In Run I, access to tape resident data was error prone and considered to be very painful by the users[1]. However, the relatively inexpensive Exabyte 8mm technology allowed CDF to store and analyse all of its data in Run I. For Run II we plan to shield the users from all tape handling by decoupling tape reading and writing from data analysis. With this strategy users deal

[1] In Run I, user analysis jobs read and wrote data directly to and from tape. The software used to read/write data was rather simple and single errors could caused job failures.

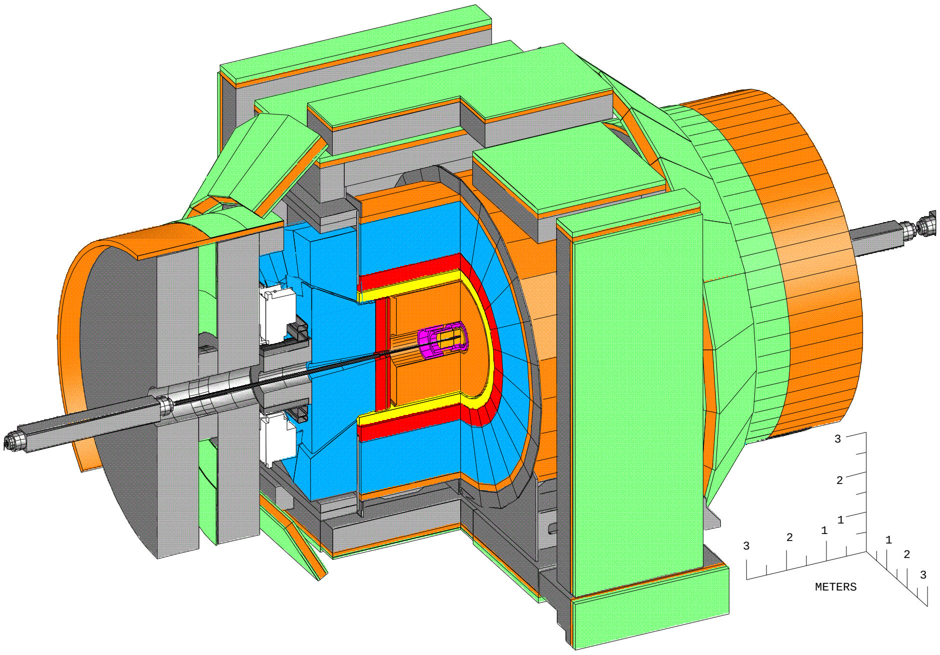

Figure 1: Schematic view of the CDF detector in Run II configuration.

only with disks. The tape subsystem can be taught to automatically retry and thus achieve high reliability with even a less-than-perfect tape technology. Our strategy is to have the tape subsystem be as inexpensive as possible but to be fully automated.

The minimum disk space required for CDF Run II was estimated, based on Run I data storage and access numbers, to be 20 TBytes. In the past, analyses have still been limited by the amount of available disk space. With disk prices dropping fast, we would like to maximize the data disk space. Thus the design of the central analysis system should allow major disk upgrades limited only by budgetary constraints.

Computer systems at home institutions and the desktop systems of the physicists provide a significant resource. The data handling system should allow convenient use of those systems. This requires easy installation of all the handling software on systems other than the central system and the ability to export and import data.

One of the design strategies is to facilitate sharing of datasets among users. This would result in a sharing of otherwise duplicated effort, a speed up of analyses, a better storage space utilization, and thus effectively more space being available to the analyses. The data handling system will extend the tools for organizing and storing data and meta-data of primary datasets to secondary and tertiary datasets created by physics groups and users.

Our data handling strategy calls for no global resource optimization. Instead, analysis jobs will be scheduled individually based on the available resources and resource usage history of the user and physics group. Status and progress information will be maintained inside each job. Analogous to Run I, there will be physics group specific resources with their allocation and management delegated to the physics group.

We want to explore the full physics potential of the detector in the most expeditious and convenient way. The data handling system should be robust and impose no unnatural procedures.

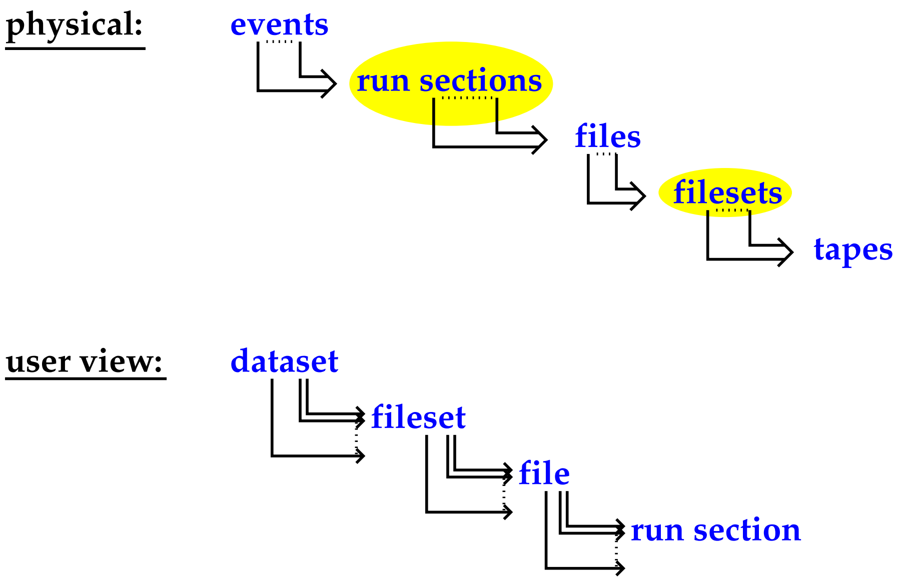

Figure 2: CDF Run II data organization: physical and user view.

4 Data Organization

The data will be organized hierarchically. Users will access data mainly by datasets. Datasets are collections of events with similar physics properties and same information per event. Event information can contain raw detector data, reconstructed physics data, selected analysis data, etc. At the other end of the hierarchy run sections describe short periods of data taking. Each event belongs exactly to one run section; knowing the run section of an event will allow the data in any dataset to be located.

In the physical view all events of a run section belonging to the same dataset are written into the same file. A file will contain events of multiple run sections. Run sections are rather short periods of data taking of about 30 sec (or 2,000 events). Files will be written with a size of about 1 GByte. Files of the same dataset are then clustered together into filesets. Each fileset will have a size of about 5 to 10 GBytes and be stored in a separate partition on tape. The size of the tapes is expected to be between 50 and 100 GBytes.

In the view of the user a dataset consists of one or more filesets. Filesets contain multiple files with each file having all the events of that dataset belonging to the run sections in the file. During normal analysis a user will access the data by dataset name and its breakdown into filesets and files will be of no concern.

In Run II the raw data will be logged to magnetic disk at the experiment in a handful of datastreams. The datastreams are collections of primary datasets to enable easier access and prioritized processing (and re-processing) of data. Primary datasets are defined online, based on level 3 trigger [5] information. The raw data are spooled to tape in the Feynman Computer Center (FCC). The production system will reconstruct the raw data and will split each datastream into its primary datasets. Physics groups are expected to make secondary and tertiary selections based on level 3 trigger and reconstruction information.

5 Central Analysis System

The central analysis system has to provide most of the analysis power required by the experiment. The minimum compute power required for the upcoming run of the CDF experiment was estimated, based on Run I experience, to be 3,000 SPECint95. The commissioning of the central analysis system started in autumn of 1999. The system will be extended and upgraded over four or more years and may even transition into the Run III analysis system. With such a long lifetime a more modular design becomes important, so that aging components can be replaced and new technologies can be integrated as they become available or mature.

The design of the system is based on the idea of a tight storage and CPU connection. The central storage pool will consist of local and shared data disks with a capacity of at least 20 TBytes, an automated tape library with a capacity of 1 PByte, and shelf resident tapes in case the capacity of the tape library is exceeded. The compute system consists of a small cluster (order 10 nodes) of moderate to large sized multi-processor machines. The machines are clustered loosely, sharing only the user login and cluster spool area (scratch disk space is local on each machine) and authentication information (i.e. password, etc.).

A sketch of the hardware architecture with the major components is shown in Figure 3. The following subsections discuss each of the components in more detail.

5.1 Network Fileserver and User Authentication

An analysis system with more than two or three standalone compute nodes is not attractive from a user point of view. Keeping analysis software, passwords, etc. manually synchronized or splitting the work over multiple system adds unwanted overhead. Even with work or physics group specific association, collaborative working will be degraded. To enable the option of an analysis cluster we need to share at a minimum user files and passwords efficiently and reliably. Several vendors offer clustering software. Unfortunately most are restricted to homogeneous clusters of that vendor and are proprietary. If we want to have a cost competitive compute system and make use of Intel/Linux systems in the central analysis cluster then such software is not an option.

Fermilab and CDF have a lot of experience with the Sun's Network File System (NFS). In the past we used computers to NFS-serve filesystems and used version 2 of NFS. Performance issues and reliability, especially with SGI servers, made clustering based on NFS worrysome. During the last years dedicated NFS fileservers have matured to a point where they can achieve a performance similar to filesystems on locally attached disks. In addition, NFS version 3 has become available on most platforms.

To evaluate the viability of such an approach, we acquired a small evaluation unit from Network Appliances in 1998. Tests of performance and reliability of this unit were very positive. In 1999 CDF bought two Network Appliances F740 with about 100 GBytes of disk space each. The configuration of the two fileservers is identical; however, they are not clustered in failover mode but are setup standalone. The first one serves the clusterwide user login area, the second fileserver provides the spool area. The fileservers have dedicated point-to-point network connections to all compute nodes in the central analysis cluster and a connection to the general purpose network for the desktop systems. The two fileservers have been running problem and maintenance free since their commissioning. We expect to upgrade the disk space later this year to increase user quotas for the login area to half a GByte.

Fermilab has selected Kerberos version 5 from MIT [6] for user authentication to avoid passwords being sent unencrypted over the network. A pilot project started in 1999 [7] and is being used on the machines in the CDF Run II central analysis system. The local password file is used for all other user account configuration information. Deployment of Kerberos on the desktop systems

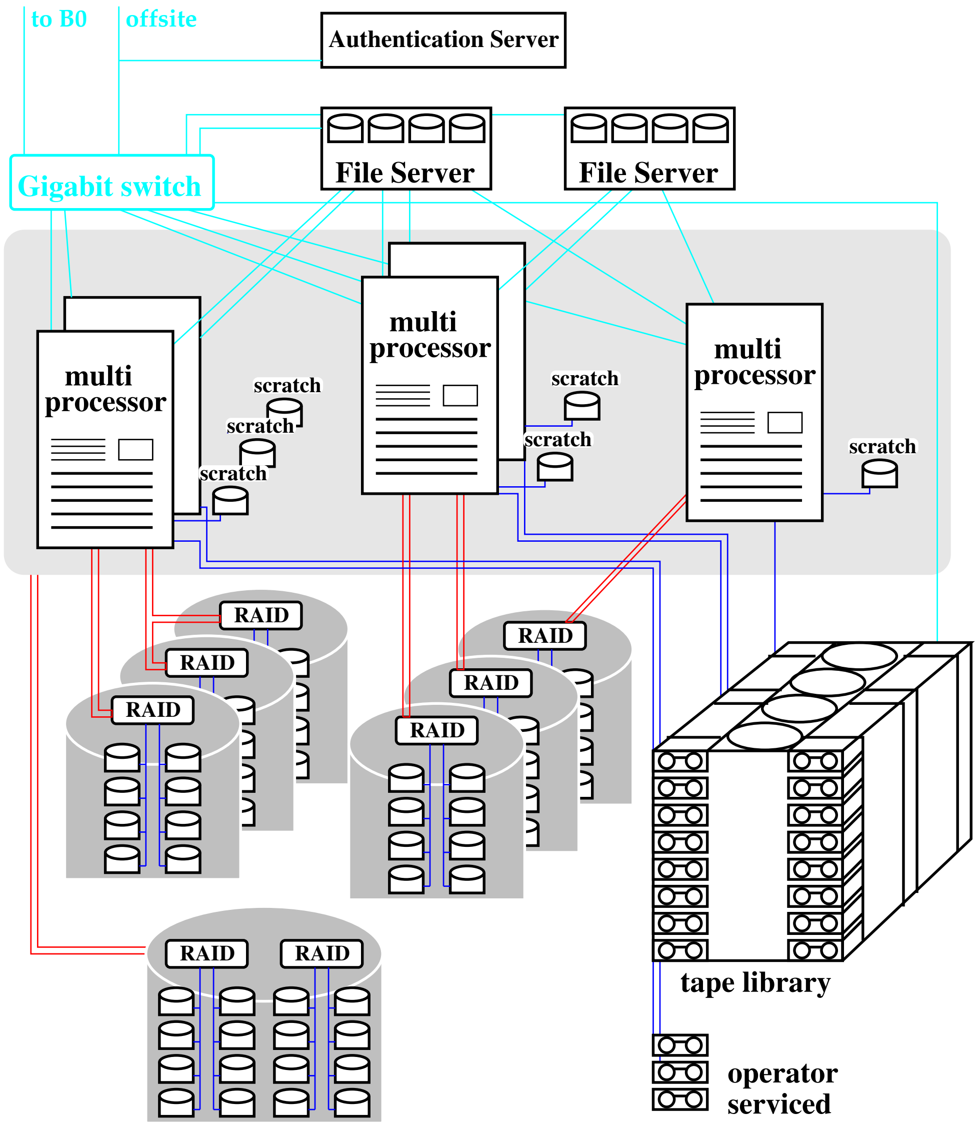

Figure 3: Sketch of the CDF Run II central analysis system.

started in 2000.

5.2 Compute Subsystem

The CPU requirement to analyse CDF's Run II data is estimated based on Run I usage scaled by the increase in data volume. Thus the estimate of 3,000 SPECint95 does not include the expected increase in analysis activity due to event purity and complexity, and size of the collaboration. We expect this not to be a problem as a significant part of the Run I usage was used for Monte Carlo generation which can be off-loaded onto a PC farm like the production system. Although a fully populated SGI Origin 2000 can now provide the 3,000 SPECint95 in a single system, it will not have much upgrade capability and will be rather expensive. Large scale multi-processor systems (SMP or NUMA) are still supercomputing equipment and will, most likely, become less and less affordable for high energy physics. Moderate sized compute systems are significantly more cost effective as they are commodity workgroup servers. While uni-processor and small sized compute systems are even less expensive, data connectivity and management of the systems makes them a less attractive solution. We expect the compute system to be a cluster of order ten nodes of large and moderate sized servers. The cluster can be potentially heterogeneous. CDF currently supports its reconstruction and analysis software on SGI/IRIX, Sun/Solaris, and Intel/Linux.

In 1999 the first large compute node was bought in a limited bid between SGI and Sun. The 64 processor Origin 2000 was commissioned at the end of last year. It provides about one third of the estimated CPU requirement. The machine has a small amount of local SCSI disk space to hold a copy of the CDF reconstruction and analysis software. It has 100 Mbit/sec Ethernet point-to-point connections to each fileserver and a 1000 Mbit/sec Ethernet connection to the main network switch of CDF, a Cisco 6509.

A small Sun E4500 was also commissioned in 1999 to provide a Solaris development platform. The machine has 4 processors, local SCSI disk space for the CDF reconstruction and analysis software and 100 Mbit/sec Ethernet connections to the fileservers and main network switch.

5.3 Storage Subsystem

The core of the central analysis system is a pool of over 20 TBytes of disk space. This disk space will be used to cache data from tape and to assemble new datasets. Some of the datasets will be accessed very frequently and are expected to be kept disk resident for several months.

SCSI is a low overhead, efficient, flexible channel protocol. It works very well and has broad support. On the other hand, standard or parallel SCSI has a very limiting physical layer. Most people encounter these limitation as maximum length of the SCSI bus or cable. With the large number of devices in the disk pool of the central analysis system, we decided to use Fibre Channel (FC) [8] as the interconnect. Using the SCSI protocol over Fibre Channel then provides an efficient channel style protocol and a physical layer that possess little limitations. With the use of Fibre Channel we get also a network character that we can use to share disks between nodes in the compute system. Sharing disk devices requires coordination beyond multiple initiator support in the SCSI interface of the device. In the case of data sharing a shared or sharable filesystem is also required. For the static datasets that will be kept disk resident for several month, read-only access to the data from the compute subsystem will be sufficient. If the data is written into ISO-9660/High-Sierra filesystem format the data can be read by all nodes in the compute subsystem independent of operating system. A small FC network between part of the disk pool and the compute nodes will provide a storage area network (SAN) for the static datasets.

The total I/O bandwith from/to disk was estimated based on the Run I data access patterns and measurements to about 1.5 GBytes/s. Even with the largest size disks such a bandwidth is not a problem from the device side. However, the interconnect between disk pool and compute system requires tens of Fibre Channel connections, about one for every half to one TByte of disk space. CDF procured last year the first part of the disk space, about 8 Tbytes, to be available for detector commissioning work, well before the start of the physics run. For increased robustness of the storage subsystem in the event of disk failures and to provide higher than single disk I/O rates to user analysis processes without software disk striping a hardware RAID solution was selected. Each of the 22 independent units has one Chaparral K7413 RAID controller and about a third of a TByte of RAID level 3 space. The 22 units are connected over several FC loops to the compute subsystem via small hubs.

The total data volume of CDF Run II is expected to be over 1 PByte. The media cost for this data volume is significant and the use of commodity tape technology could result in large savings for this subsystem. With a tape technology also affordable at remote sites, data export and import will be simplified. The Fermilab serial media working group has not yet made a recommendation for Run II but is is thinking along the direction of commodity tape technology.

CDF has planned the central analysis system for what is called lights out operation. To realize unattended operation of the system, all data must be accessible without operator assistance. In 1998 CDF acquired an automated tape library from EMASS, now ADIC. The storage elements inside this AML/2 library can accomodate tape media of practically any form factor. In 1999 the library was upgraded to 4 quadro-towers. In this configuration the library can hold 1 PByte of data, assuming 50 GBytes 8 mm style media or similar density. The library has space for over 200 tape drives. All tape drives will be directly attached to the compute nodes via parallel SCSI. The required 24 hour average I/O bandwith to/from tape for CDF Run II was estimated to be about 150 MBytes/sec. A moderate sized pool of commodity type tape drives (with about 5 MBytes/sec each) can easily provide this bandwidth. Each compute node will have access to a small pool of tape drives and a few SCSI buses per node will suffice. A small number of tape drives are currently installed in the library and connected to the compute subsystem. The drives, Sony AIT and Exabyte Mammoth, are used for testing and integration.

5.4 Desktop and Remote Systems

Desktop systems of the physicists in the B0 trailers can be part of the analysis cluster and share authentication, login, and spool area with the central analysis system. We hope that this will result in a better integration and thus utilization of those resources, and allow to offload CPU intensive but data un-intensive tasks from the central analysis system. The desktop system is expected to play a particular role in the late stage of the analysis when small data samples can be copied onto the desktop for repeated analysis.

The new desktop systems that have come into the B0 trailers during the last two years for Run II have all been PCs. The Fermilab Computing Division supports Linux on Intel PCs. A customized Fermilab distribution based on the RedHat distribution is provided to users and used on the CDF Run II desktop systems. Most of the desktops have a similar configuration and are managed by a local group of system managers. The desktops are connected via 100 Mbit/sec Ethernet to a local network switch which has a 1000 Mbit/sec uplink to the main network switch of CDF.

Remote systems can be built analogous to the central analysis system. Configurations will range from a single desktop up to a cluster of multi-processor systems with tape libraries. A single moderate sized multi-processor system should satisfy the analysis needs at most home institutions and eliminate the need for clustering and fileservers. Most storage subsystems at remote sites will probably consist of only data disks with one or two tape drives for data import/export or to backup new datasets. The use of a commodity tape technology at Fermilab would allow to directly import/export tapes from/to Fermilab.

5.5 Dataloggers

The CDF Run II datalogger will have two parts: A datalogger in B0 will receive events from the level 3 trigger system. About 75 events per second will be sent to the datalogger at peak luminosity. Events will be sorted by datastream and each datastream will be written to its own set of files. To simplify bookkeeping and luminosity counting, files will be closed and opened at run section boundaries. Events at the file boundary that are out-of-order (due to the parallelism in the trigger system) will be sorted and written into either of the two files depending on their collision time. The CDF online group has acquired a SGI Origin 2200 as B0 datalogger. The data disks of this machine will have a dual Fibre Channel interface.

The datalogger in the Feynman Computer Center will read the files from those disks and write them to tape inside the library. A single-mode fibre is used to connect the second FC port of the disks in B0 with the datalogger in FCC. We plan to use the cluster filesystem from SGI, CXFS, to share files on those disks between the two datalogger machines. The machine in B0 will run the meta-data server while the FCC machine will be setup as a weightless client. Data needs to be archived at about the same speed at which they are recorded by the B0 datalogger. At peak luminosity this will be about 20 MBytes/sec.

5.6 Production System

The dataloggers record signal information from the detector. Those raw data have to be converted to physics quantities before the events can be analysed. The reconstruction of the events is very CPU time intensive. A farm of PCs running Linux was found to be the most cost effective solution for this processing. Two I/O nodes will serve on the order of 200 PCs. The first I/O node has the connection to the central storage pool and will fetch raw data files onto its local disks. The files will be distributed to the processing nodes, the PCs, and stored there on a local disk during processing. The processing nodes will write the reconstructed events to the local disk and when the processing of the raw data file is complete will send the output file to the second I/O node. The I/O nodes and worker nodes will be connected through a switch with Fast Ethernet connection to each worker node and Gigabit Ethernet connection to the I/O nodes.

While the raw data are organized into a handful of datastreams, the production output should be organized along datasets. There will be around 8 datasets per datastream for a total of about 50 datasets. Splitting into dataset will be done in the second I/O node. While the size of a raw data event is approximately 250 kBytes, the reconstructed information per event is much smaller, around 65 kBytes. With the smaller event size and the dataset splitting, several files need to be concatenated in order to write files close to the target filesize of 1 GBytes. The concatenated files are then send to the storage subsystem and archived on tape.

The CDF Run II production systems are described in detail in Ref [9].

6 Data Catalogue

The main location of data information is the CDF Data Catalogue. The implementation is based on a relational database. The CDF meta-data are organized into a hierarchy of tables containing dataset, fileset, file, and run section information. The data catalogue also contains general bookkeeping information such as tape volume contents and allocation, and physics related bookkeeping information like data quality and luminosity information. A set of secondary tables will keep processing, filter, and selection information that is common among various datasets.

A software package called DataFileDB was written to query, update, and manage the information in the CDF data catalogue. The package provides a database independent interface to the information. CDF has selected Oracle as database management system for the main data catalogue at Fermilab. MSQL and flat file based implementations are supported for remote sites and for private instances. Physics group and user owned instances will be available. Their design is currently under discussion.

A small Java based web interface is being developed as GUI interface. In the first version of the software query access will be provided to the four main tables (dataset, fileset, file, and run section information) of the CDF data catalogue. Query access to secondary tables is planned for a later version and update/management access through this interface is under discussion.

The CDF data Catalogue is described in more detail in Ref [10].

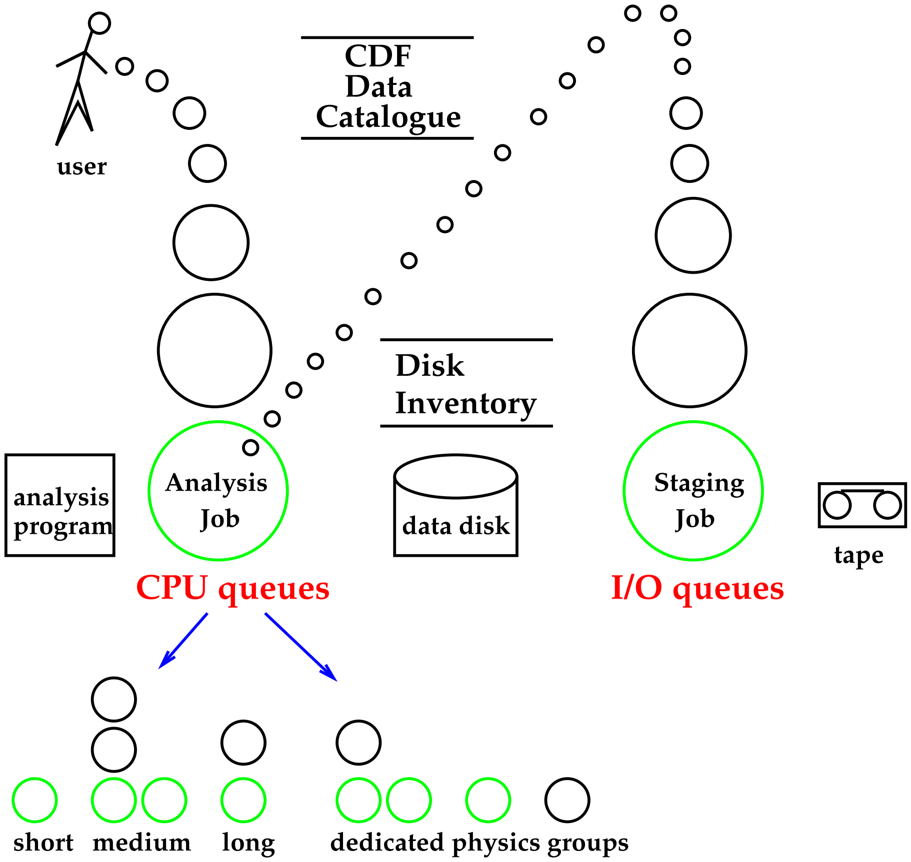

Figure 4: Sketch of the CDF Run II batch system.

7 Resource Management

Identifying the limited resources and being able to manage those is a key issue in the CDF Run II data handling. With the rather large number of collaboration physicists and diverse physics program of the experiment a dynamic management is needed in the central analysis system for most of the limited resources. The important resources to be managed are: CPU time for analysis jobs, disk space for data, and tape drives for transferring data in and out of the mass storage subsystem. The I/O bandwidth to and from the data disk pool is expected to be ample.

A full discussion of the resource management was presented at CHEP2000 [11].

7.1 Batch System

CPU and tape I/O resources will be allocated to users through a batch system. The allocation of CPU cycles among users and groups has long been done by the use of batch queues. CDF plans to use LSF from Platform Computing as a batch system on the central analysis system.

There will be a set of queues based on the CPU requirements of analysis jobs and a second set of queues for the tape I/O staging jobs which are mostly automatically generated by the analysis jobs. There will be CPU queues for physics groups and special projects as well as general purpose use. There will be I/O queues for input, output, and for event picking, as well as for special purposes. The number of jobs allowed to run in the various queues will be controlled dynamically rather than with hard limits. This will provide a more efficient use of resources.

With the amount of CPU consumed by an I/O job being proportional to the amount of data copied to or from tape, management of the CPU in the I/O queues effectively manages the tape I/O. The control of priorities in the scheduling of I/O jobs will then act as the management of the tape I/O resource.

Fairshare scheduling that takes account of the amount of CPU consumed by groups and users and of their quota will be used to ensure that users get a fair share of the CPU and I/O resources. The CPU and I/O usage will be accounted separately and might have different quotas from each other.

7.2 Disk Inventory Manager

Disk space will be managed through quotas, on the login, scratch, and spool areas and data disk space through a disk inventory manager. To simplify a complicated problem, CDF data analysis and processing jobs read and write data only from or to disk; the tape handling software will transfer data between tape and disk.

Some datasets will be accessed more frequently than others and, if small in size, copies can be fixed permanently on disk; such datasets will be chosen by collaboration management. Larger datasets and less frequented datasets will be staged from tape to disk and reside in a shared array of disks for temporary use. A disk inventory manager coordinates the use of all the data disk space of static and cached data.

The quantum of data used for managing disk space is the fileset. Data transfers between disk and tape move all filesets on a tape in one job. It is filesets that are marked static or appear temporarily in shared space; the disk inventory manager manages filesets and keeps no knowledge of datasets as such.

The filesets in shared space remain in place until space is needed to fetch more filesets. The disk inventory manager will use an algorithm based on frequency of use and of time since last use to determine which filesets to remove when space is needed. To ensure that temporary filesets remain in shared space until processed by a requesting user job, a use reservation is maintained by the disk inventory manager until the user job releases the fileset.

Controls and priorities on the amount of reserved and static disk space by group and user are plannned as a feature of the disk inventory manager.

8 Tape Handling Software

In Run II analysis jobs will no longer access tapes directly, instead they will access all data via disk. The disk inventory manager will coordinate the copy of data between disk and tape. The staging/spooling jobs will be scheduled by the batch system. With this approach the tape handling software is pretty much decoupled from all other data handling and experiment software.

For Run I CDF developed a small tape tool package, mt_tools. The package provides convenient tools to read and write ANSI file structured tapes and has efficient retry in case of tape errors. The package was used extensively with Exabyte 8500 tape technology in both the online and production system where staging disks were used. The software has undergone a major overhaul in 1998 to adapt it to the changed structure of Fermilab's tape software. It is now a rather thin layer on top of the Fermilab Tape Tools (FTT) and Operator Communication Software (OCS).

The FTT package provides a platform and device type uniform tape interface. It was developed at Fermilab toward the end of Run I to collect all knowledge and expertise on device type specific features of tape drives and to provide a uniform low level interface in one package. The package is available for all operating systems that are currently supported by CDF. It handles the full line of Exabyte and Quantum tape drives, both generations of Sony AIT, and various other tape technologies through generic table entries.

The OCS package tracks the status of tape drives and handles operator mount/dismount requests. A little database inside the package stores the configuration information and is used to maintain device allocation, mount status, error rates, etc. for each drive. Access to tape drives can be managed on a per user and group basis. The package provides a common interface to the application to issue tape volume mount and dismount requests, independent of the device type or location. For human operators servicing tape drives a X11 based operator console to view and acknowledge mount/dismount requests is included. CDF has written a little daemon to connect the operator interface of OCS with the library control interface of several stackers and tape libraries. In the case of the EMASS/ADIC AML/2 the DAS/ACI interface is used. This way all mount/dismount requests look the same on the application side, i.e. in mt_tools, independent of the type of operator servicing the tape drive. The architecture of OCS supports setups for individual hosts as well as cluster with tape drives and operators on multiple nodes.

9 I/O Modules for Accessing Data

The analysis driver of CDF for Run II is a joint development with the BaBar experiment at SLAC. The analysis driver has simple file I/O modules to read and write data files. To make use of the new CDF data handling system, specialized input and output modules are being provided.

The input module allows a user to select data by dataset instead of specifying a list of files, etc. The user can veto run sections based on data quality information or restrict a selection to a particular range of data taking. The input module will use the CDF data catalogue to translate the datset names into a list of filesets. This list is then passed to the disk inventory manager. The disk inventory manager will trigger input staging for the filesets not in its disk cache as appropriate. Filesets available in the cache will be reserved for processing and their storage location send back to the input module. The input module coordinates reading of the files in the fileset and individual events are processed by the analysis driver.

The output module is a bit more complex. It has to be able to switch output files automatically if the target filesize is reached and sort out-of-order events based on run section affiliation at such file boundaries. For new datasets that will be managed by the disk inventory manager it has to collect and store the meta-data of the file in the CDF data catalogue and hand over the file to the disk inventory manager to be clustered into a fileset and stored on tape. To do this temporary disk space may need to be alocated from the disk inventory manager.

A discussion of the I/O modules for accessing data was presented at CHEP2000 [10].

References

1 F. Abe *et al.*, Nucl. Instr. and Meth. **A271** 387 (1988), see also `http://www-cdf.fnal.gov/pubcdf.html`.

2 Fermi National Accelerator Laboratory, `http://www.fnal.gov/pub/fermilab_intro.html`.

3 Fermilab Main Injector, `http://www.fnal.gov/pub/main_injector.html`, and `http://www-fermi3.fnal.gov/injector.html`.

4 CDF Collaboration, "The CDF II Detector - Technical Design Report", FERMILAB-Pub-96/390-E (1996).

5 I. Kravchenko *et al.*, "Event Builder and Level 3 Trigger at the CDF" CHEP2000, Paper 224.

6 Kerberos: The Network Authentication Protocol, `http://web.mit.edu/kerberos/www/`.

7 M. Kaletka *et al.*, “Fermilab Strong Authentication Project”, CHEP2000, Paper 306.
8 X3T9.3/Project 755D/Rev 4.2, “Fibre Channel Physical and Signalling Interface (FC-PH)”, Working draft, ANSI, October 8, 1993, The Fibre Channel Association, “Fibre Channel: Connection to the Future”, ISBN: 1-878707-19-1.
9 S. Wolbers *et al.*, “Design and First Tests of the CDF Run 2 Farms”, CHEP2000, Paper 70.
10 P. Calafiura *et al.*, “The CDF Run II Datafile Catalog and Data Access Modules”, CHEP2000, Paper 367.
11 E. Buckley-Geer *et al.*, “CDF Run II Data Handling: Resource Management and Tests”, CHEP2000, Paper 368.