Overview of CDF Run II Data Handling System

Liz Buckley-Geer, Mark Leininger, Terry Watts
Stephan Lammel

- **Introduction**
- **The CDF Experiment**
- **Strategies for Run II**
- **Data Organization**
- **Central Analysis System**
- **Storage Subsystems**
- **Disk Inventory Manager/Stager**
- **Tape Handling Software**
- **Conclusions**

Introduction

The Collider Detector at Fermilab is a large multi-purpose detector at the Fermilab Tevatron. The experiment records and analyses proton anti-proton interactions at a center-of-mass energy of 2 TeV.

The experiment recorded its first data in 1985. During the last collider run, Run I, (1992 to 1995) about 50 TBytes of data were recorded.

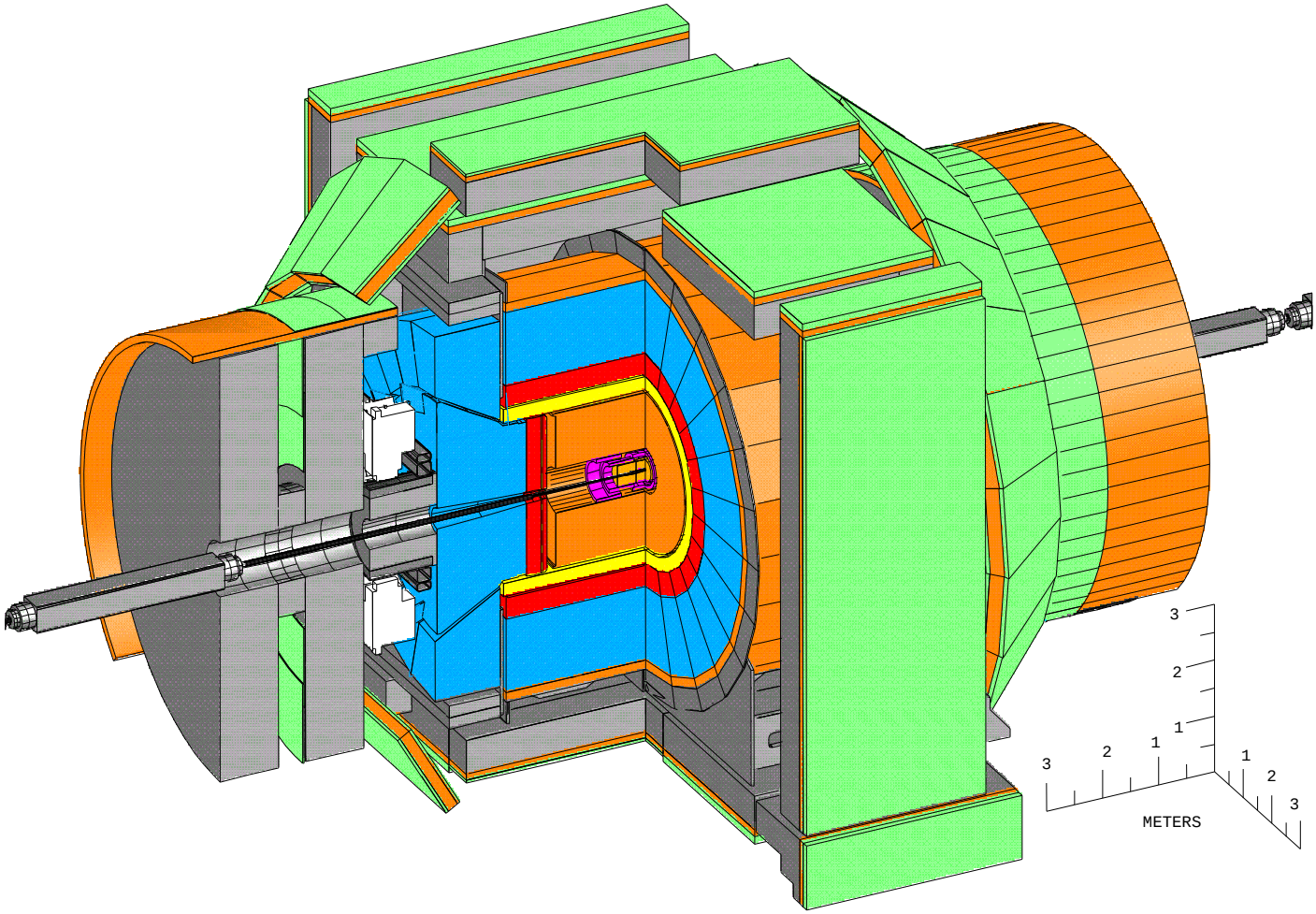
For Run II the Fermilab accelerators are being upgraded to provide a significantly higher luminosity and a 10% increase in center-of-mass energy. The Tevatron will operate at much shorter bunch spacing of 396 ns.

The goal of Run II is to accumulate an integrated luminosity of $2\,\mathrm{fb}^{-1}$. (Run I yielded $100\,\mathrm{pb}^{-1}$.)

The estimated data volume of Run II is 1 PByte. Although the trigger will be much tighter in Run II, new triggers are added to explore the full physics potential of the experiment.

With an increased size of the collaboration (now over 500 physicists) and more complicated data (due to the multiple interactions at high luminosity) the demands on compute resources for data analysis will increase much more than the 20 fold increase in data volume.

The CDF Experiment



Strategies for Run II

The data handling strategy for Run II is very simple:

- **continue/enhance the Run I system but**
- **eliminate the shortcomings recognized in that system and**
- **integrate new computing technologies that are mature enough.**

The main strategy in Run I analysis was event filtering in multiple stages. The reduction steps were chosen by the user to minimize re-selection risk and effort.

- **User created datasets were not always well documented and known, similar selections were done by different users.**

In Run I primary datasets were created in a very organized way directly after event reconstruction.

- **It would be nice to extend this to secondary/tertiary datasets.**

Documentation of user-created datasets varied from user to user and was kept in logbooks and technical notes.

Propagation or exchange of this information was on an oral basis on the corridors.

- A CDF Data Catalogue will be provided so that all relevant dataset information can be stored in a uniform and organized manner.

In Run I each event was stored on average more than 3 times to ease data access.

- With the increased storage costs for serial media, we cannot afford this approach for Run II.
- The CDF Data Catalogue should allow us to better share secondary and tertiary datasets.

In Run I we tried to avoid or delegate resource management as much as possible. There was no prioritization of analyses but an attempt to accomodate each analysis individually. CPU for analysis work was sufficient. Data disk space was managed by each physics group with mixed success.

- Better batch system for Run II.
- More dynamic disk space management via a disk inventory manager.

While there was sufficient CPU for data analysis, Monte Carlo generation and simulation drained the central anal-

ysis system(s). With the mainframe style approach upgrade options were very limited and painful. During the last year we have also reached the (soft) limit on the number of disks or free SCSI busses on the system.

- Replace single system approach of Run I with a simple cluster.
- Design/plan data disk subsystem with "no" upgrade limitations (or budget only limitation).

The desktop of the physicists had a very significant CPU power in Run I. Still, 99% of all analysis was done on the central systems.

- For Run II we plan to integrate the desktop systems into the analysis cluster and make their use more convenient.

All Run I data is stored on 8mm tapes using Exabyte 8200 and 8500 technology. While this was a very cost effective solution, media and drive reliability made access to tape resident data very painful.

- For Run II we plan to decouple tape reading/writing from data analysis.
- The idea is to have users deal only with disk and hide all tape access.

Data Organization

The CDF data will be organized hierarchically:

- Datastreams containing several datasets are written by the datalogger.
- Production splits these into the primary datasets. Physics groups and users create secondary and tertiary datasets from them.
- Each dataset is written into its own set of files.
- Files are then clustered into filesets.
- The lowest granularity in the CDF data handling system will be run-sections. Those are non-overlapping of order 30 second data taking periods.

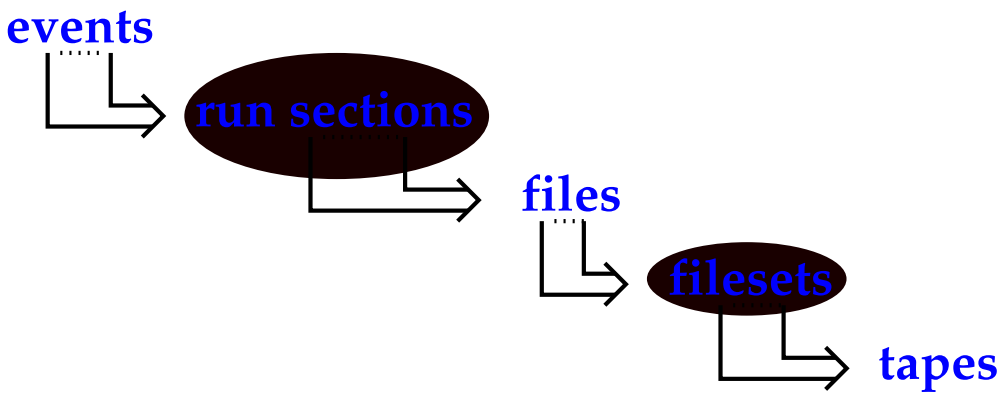
With this approach we are able to eliminate event level entries in the catalogue but can still easily determine luminosities and prescales, and discard events with questionable run conditions in each dataset.

Filesets provide a convenient granularity for data management on both tape and disk.

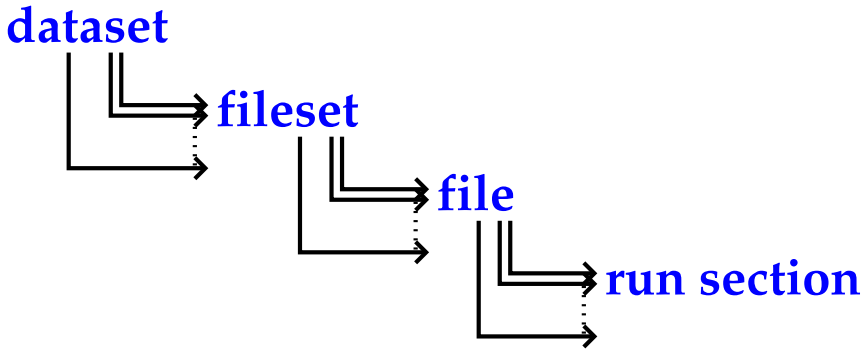
The CDF data catalogue is the key to the Run II data organization.

Data Organization

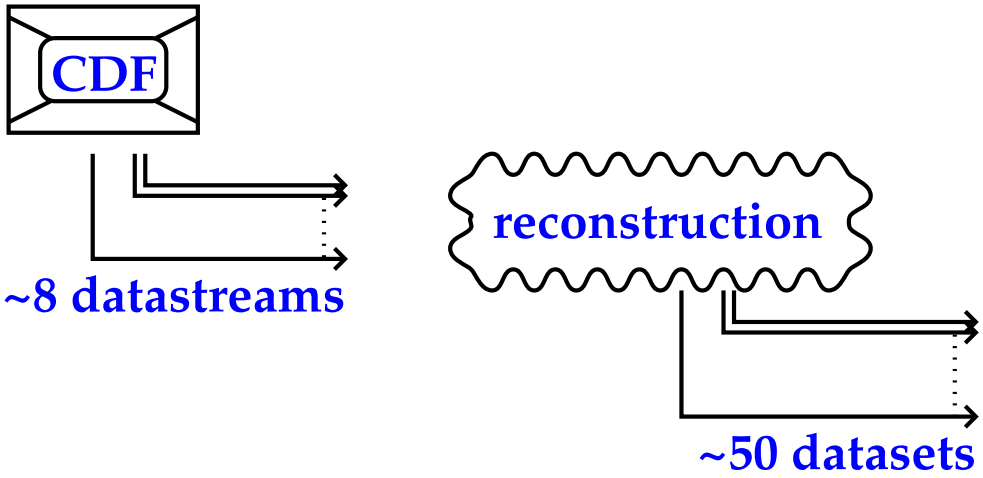
physical:



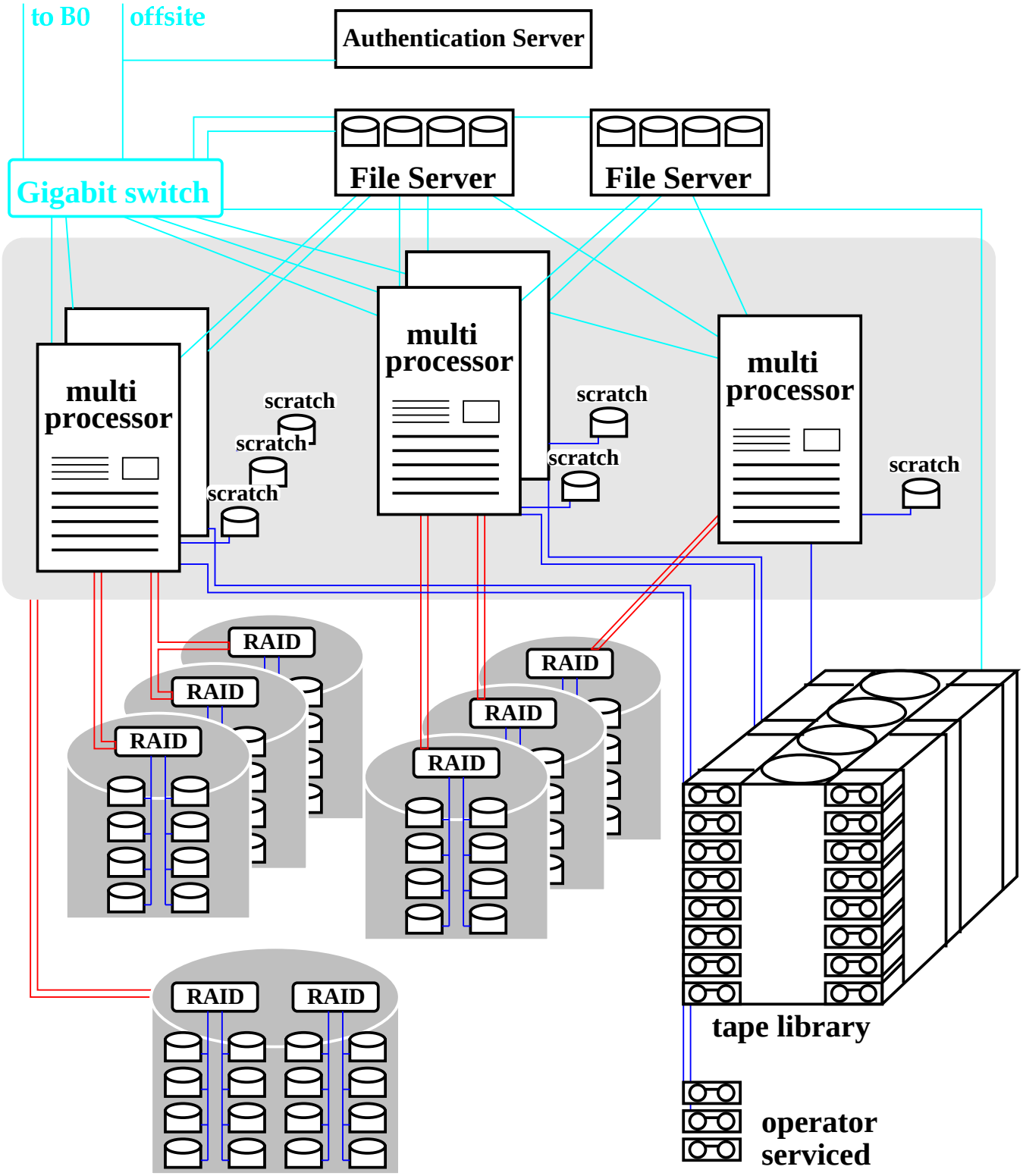
user view:



data flow:



Central Analysis System



Central Analysis System

The central analysis system for CDF Run II is a potentially heterogeneous cluster of mid to large size multiprocessor systems.

The required compute power for CDF Run II was estimated to be about 3,000 SPEC_Int95. The platforms/operating systems currently supported by CDF for offline data analysis are: SGI/IRIX, Intel/Linux, and Sun/Solaris.

The compute nodes in the central analysis cluster will be clustered loosely, sharing only the user login and spool areas and authentication information.

The user home and the cluster wide spool area is realized via two NFS fileservers from Network Appliance.

Fermilab is investigating the use of kerberos for user authentication.

Each machine in the cluster has local scratch space, local data disk space, and tape drives in the tape library.

Global read-only data disk space can be accessed by all machines.

Storage Subsystems

The core of the central analysis system is a pool of over 20 TBytes of data disk space.

The exact split of locally attached read-write versus global read-only data disk will be decided as we go along.

Logically behind this disk pool is a robotic tape subsystem with a storage capacity of over 1 PByte. Shelf resident tape space will be used in case of library overflow.

We have bought the first 8 TBytes of data disks. We decided to use RAID controllers with FC host connection and parallel SCSI device channels.

Tape drives will be attached to the compute nodes via parallel SCSI.

We have not yet decided on the Run II tape technology. Exabyte Mammoth-2 and Sony AIT-2 are candidates.

Since we deferred the selection of the tape technology, we bought a multi-media capable library from EMASS/ADIC last year.

Disk Inventory Manager/Stager

We will use data disk space mainly as big cache of data that is archived on tape (or will be archived to tape).

- A disk inventory manager will keep track of the data that is on disk, on a fileset level.
- It will accept read reservations from users (read lock) and clear them upon request.
- It will trigger staging of tape resident data to disk. (find and recycle space...)
- It will manage temporary disk space for output (write lock).
- It needs to have simple quota and queuing to prevent users from blocking space.

The heterogeneous architecture of the analysis cluster and our desire to share data disks with static data among compute nodes requires features beyond those of current packages.

We wrote some prototype software in autumn of 1998 to investigate and test the neccessary interaction with both analysis jobs and the batch system. We are now completing the implementation.

Tape Handling Software

With data access by datasets, we actually know the access pattern very well once a job has started.

By clustering data of the same dataset during tape writing, we can avoid a file level tape access. Volume based tape access reduces the mount/seek/rewind/dismount overhead and allows us to keep the tape subsystem very efficient without a lot of work.

Since we want to decouple tape access from analysis programs, we need a package to copy data from tape to disk and vice versa. The mt_tools package that we used successfully in Run I did exactly this. It uses the Fermilab FTT and OCS packages for low level tape operations, to track tape drive allocation, and for operator communication.

There was an overhaul last summer adding tape partition support.

We also wrote a little interface between the tape library control software and OCS. This way all mount/dismount requests look the same on the application side independent of the "operator type".

Conclusions

- We have analysed the data handling of the last run to identify successful approaches and shortcomings.
- We have worked out a data handling and analysis system for CDF Run II that should allow us to explore the full physics potential of the new detector.
- We have done extensive R&D, prototyping, and benchmarking of new components and approaches during the last year and a half.
- The core of the analysis system is now commissioned.
- We have prototype/first versions of all data handling software.
- We should have a production quality version of all the data handling software to exercise in the engineering run that will start August 15th.
- We are looking forward to physics data on March 1st 2001.