High Energy Physics at Fermilab: Petabyte Scale Scientific Data Distribution

Dr. Michael Diesburg
Computing Division
Fermi National Accelerator Laboratory
P.O. Box 500
Batavia, IL 60510
diesburg@fnal.gov
Phone: 630-840-2679

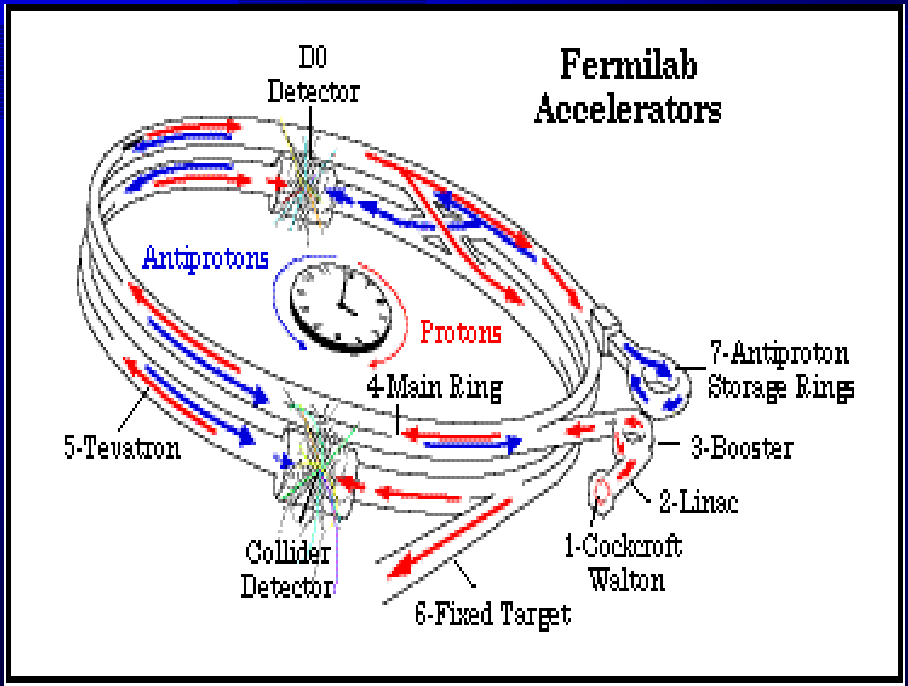
Presented at THIC in Englewood, CO June 27-28, 2000

What Is Fermilab?

- Fermilab operates the world's highest energy particle accelerator, the Tevatron, located outside Chicago near Batavia, IL
- 2,200 Scientists from 36 states and 20 countries use Fermilab facilities to do research on the fundamental nature of matter and energy.

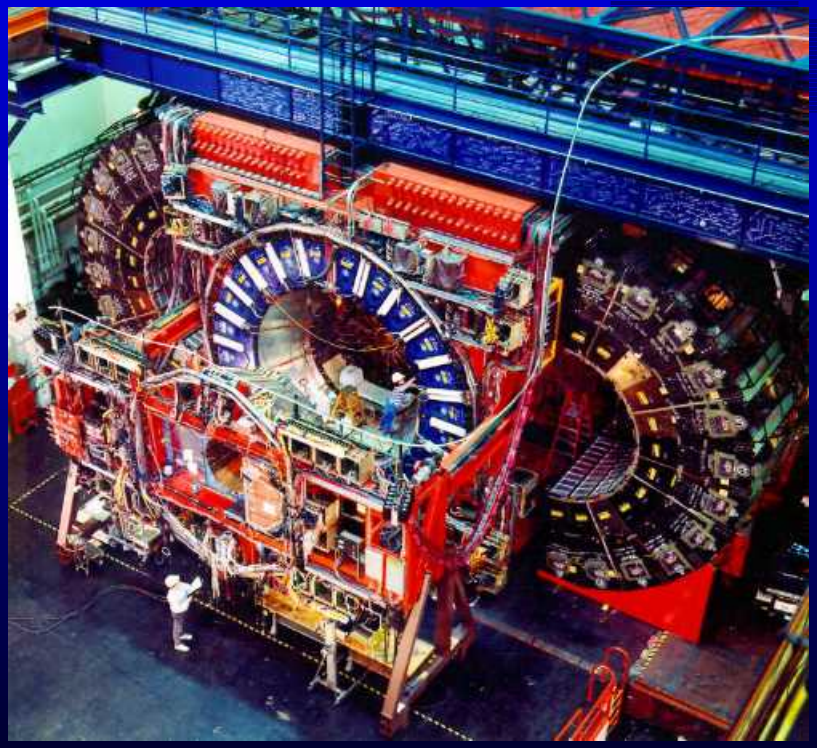
What Is Fermilab?

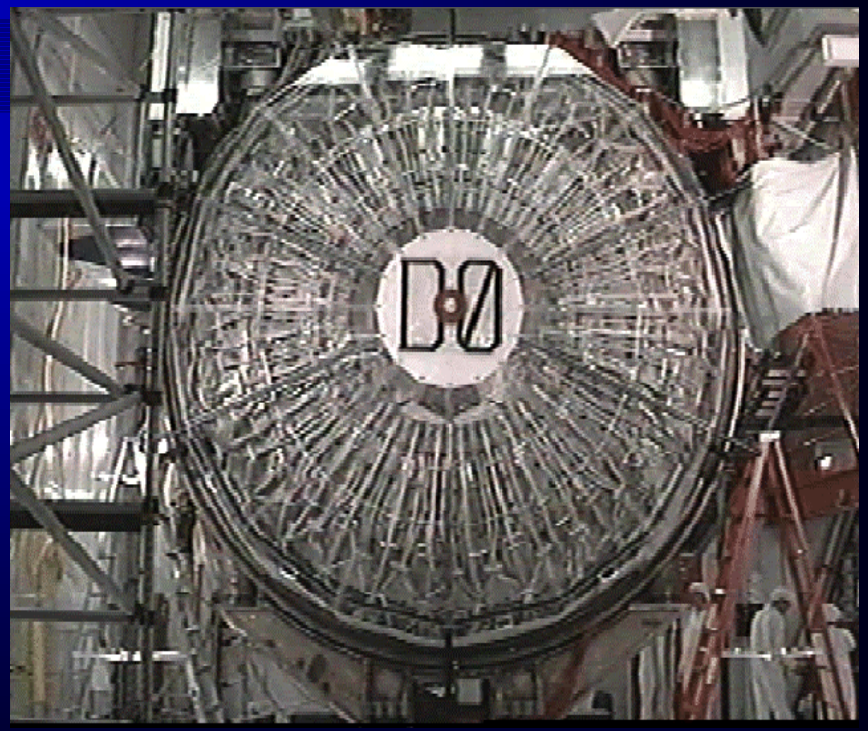
- Protons and anti-protons collide inside two huge collider detectors on the accelerator ring.
- The collisions produce bursts of secondary particles: quarks, electrons muons, neutrinos, W bosons and other exotica



The Detectors

- The detectors must record all information about the byproducts of the collisions: particle trajectories, energy, and charge
- The detectors are about 3 stories high, contain ~1,000,000 channels of electronic read-out and weigh about 5000 tons each





The Data Problem

- The data problem begins with the collisions
- Collisions occur in detectors at rate 5-10MHz
- Each collision produces about 250-500KB of information
- Data is collected 24 hours/day, 7 days/week for periods of 1-2 years
- If we tried to record it all...

  Would generate ~100Exabytes/year
- To reduce this to a manageable level the data is examined in real time by a series of filters
- Each filter makes increasingly complex decision as to the “usefulness” of each event
- Only most interesting events are kept
- Final data stream is reduced to an event rate of 20-50Hz
- With the post-processing duplication thrown in, we end up with final data sets that are only ~1Petabyte/year/experiment

The Next Stage of the Problem

- Once the data has been collected, you have to read it back
- Next stage of processing is to deliver the data to a reconstruction farm of processors where the raw detector data is turned into physics objects
- Requires large amounts of CPU, but fortunately the problem is trivially parallel, each event is independent of others and all events are processed in the same way
- Done on large banks of workstations and output written back to tape store

Onward to the Analysis Stage

- After the Reconstruction farm produces physics objects, further event selection and compression produces data sets for final analysis
- Data is grouped in streams of similar type events to facilitate further analysis. Efficient access dictates that location of data in store must be under control of user.

- Final data sets consist of:
  - Raw Data
  - Reconstructed physics objects
  - Data Summary Tapes (DSTs) with most important information
    - 25-50 Streams of DSTs ranging in size form 1-2% of data to 40%
  - Disk resident thumbnail sketch of each event, ~30TB
  - Oracle event catalog of all data sets, ~1TB

Typical Analysis Flow

- User searches the event catalog database for set of events which passes some filtering criteria
- Using this selection of events, sample of interest is further reduced by scanning through thumbnail events on disk
- If needed information is not on thumbnail, user then must make pass over DST events stored on tape
- Pass over DSTs is used to make ntuples from which final histograms are made
- In some cases further information is needed either from raw data or from the reconstructed physics objects
- For some types of analysis sample of relevant events can be reduced enough to allow keeping all raw data for sample on disk
- In either of these last two cases, random access to raw data intape store is required.

Summary of Data Access Needs

- Raw data written to Tape Store at ~20MB/s
- Raw data extracted from Tape store and sent to reconstruction farm at ~20MB/s
- Physics objects written to Tape Store from reconstruction farm at ~10MB/s
- DST data delivered to analysis engines in from Tape Store in streaming I/O mode at ~100MB/s
- Raw events and reconstructed physics objects delivered to analysis engines from Tape Store in random access mode at ~50MB/s
- Disk access to thumbnails and ntuples at ~1GB/s

- With exception of reconstruction farm, all above data is delivered to SGI Origin 2000 system which serves as main compute engine

Summary of Data Access Needs

- Ideally one would deliver data from Tape Store to any of our collaborators located around the world
- Such delivery requires that the storage system not be tied to local systems. We must be able to deliver data at rate across the network
- Delivery of data from tape store world-wide at whatever rate the network can bear

- Some data actually flows into tape store from outside institutions
- Understanding the operation of the detectors requires extensive simulation of events
- Large CPU requirements of simulation dictate marshaling all resources we can find
- Input to Tape Store of Monte Carlo events from outside institutions at ~1-2MB/s

Summary of Data Access Needs

Requirements of the system:

- Ability to do ~200MB/s aggregate I/O to tape store
  - ~150MB/s streaming I/O
  - ~50MB/s random access I/O
- Ability to control location of data in store to assure timely access
- Ability to deliver data remotely across the network
- Ability to import data sets from institutions

- A petabyte requires a lot of media. Cheap is good.

Schedule and Future Plans

- Both the CDF and D0 experiments will being collecting data next year on March 1st, 2001
- Data collection will continue for ~2 years
- At end of 2 years upgrades in accelerator operation will allow increase of 4-5 in raw interaction rates with consequent increases in data storage and access.
- Data collection is expected to continue until startup of Large Hadron Collider (LHC) at CERN
- Each experiment will have collected ~15PB of data

- Analysis of this type of data usually continues until is is superceded by a data set of greater statistical significance or greater scope.
- Likely to be graduate students chewing on this data until ‘07 or ‘08

And in the More Distant Future...

- LHC will begin operation around 2006
- Fermilab will act as US regional computing center for the Compact Muon Solenoid (CMS) experiment
- At turn on CMS will record ~10PB/yr with increases in future years as the LHC is upgraded
- Fermilab will provide compute resources and data access for all US collaborators

Eagerly awaiting advances in data storage technology to allow us to keep pace with constantly increasing demands of scientific research