



# The CDF Production Computer Farms

**Stephen Wolbers, Fermilab**

**October 4, 2001**



## Questions from Pierre

1. What is the status of 4.0.0x Production?
2. How is Express handled?
3. How are datasets out of production handled?
  - Does a tape have to be filled before the data is available?
  - Is it possible to modify datasets easily?
4. What are I/O issues?
  - When does CPU become a problem? (as opposed to I/O)
  - Can PADs be made out of production?
5. What are the long-term plans for production?
6. Anything else general about the farms?

# CDF Run 2a Farm Computing



- Goal: CPU for event reconstruction of about 5 sec/event on a PIII/500 MHz PC (Each event is 250 KB).
- Assuming 20 MB/sec peak (approx. 75 Hz)
  - Requires 375 PIII/500 processors to keep up
  - Faster machines -> Fewer processors required
    - So 180 PIII/500 duals will suffice.
    - Or 90 PIII/1 GHz duals.
    - Or 45 2 GHz duals.
- Requirement is reduced by accelerator/detector inefficiency and increased by farms inefficiency.



## Status of CDF Farms Hardware

- 154 PC's are in place.
  - 50 PIII /500 duals (couple of years old)
  - 40 PIII /800 duals (1 year old)
  - 64 PIII /1 GHz duals (recent arrivals)
  - ▶ Equivalent to 488 PIII /500 CPU's
  - ▶ Compare to 375 estimated need



October 4, 2001

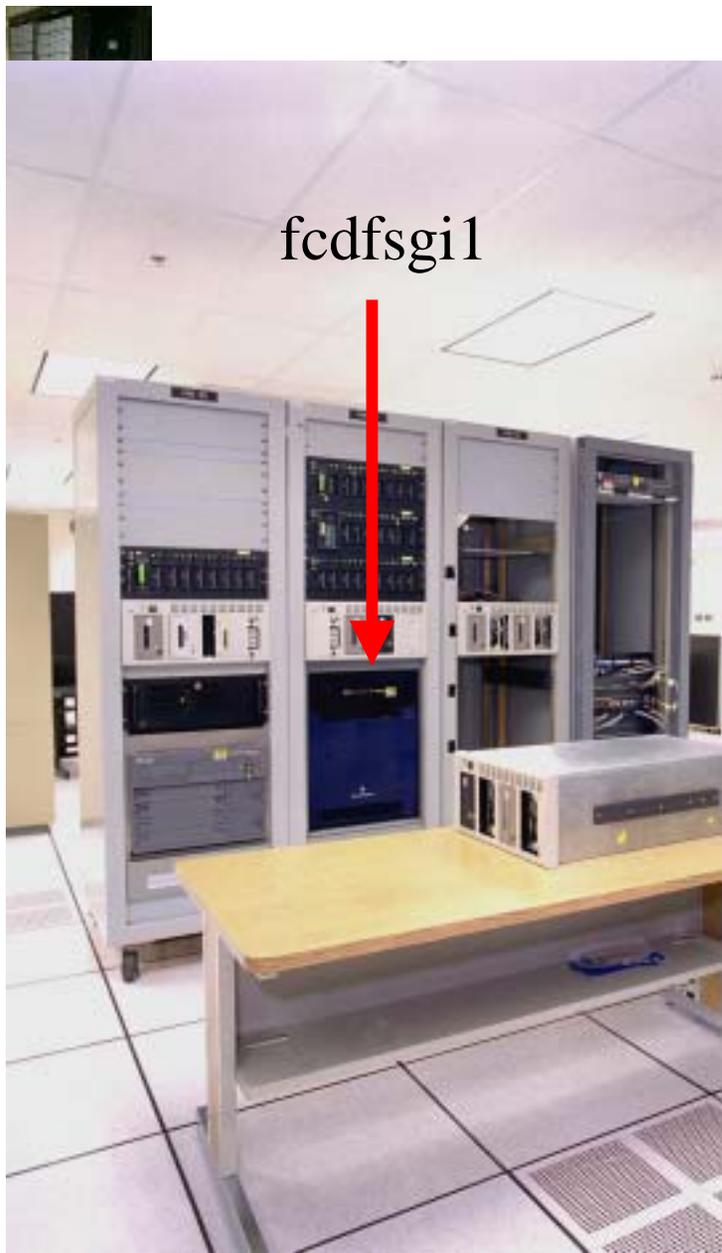
Stephen Wolbers, CDF

5

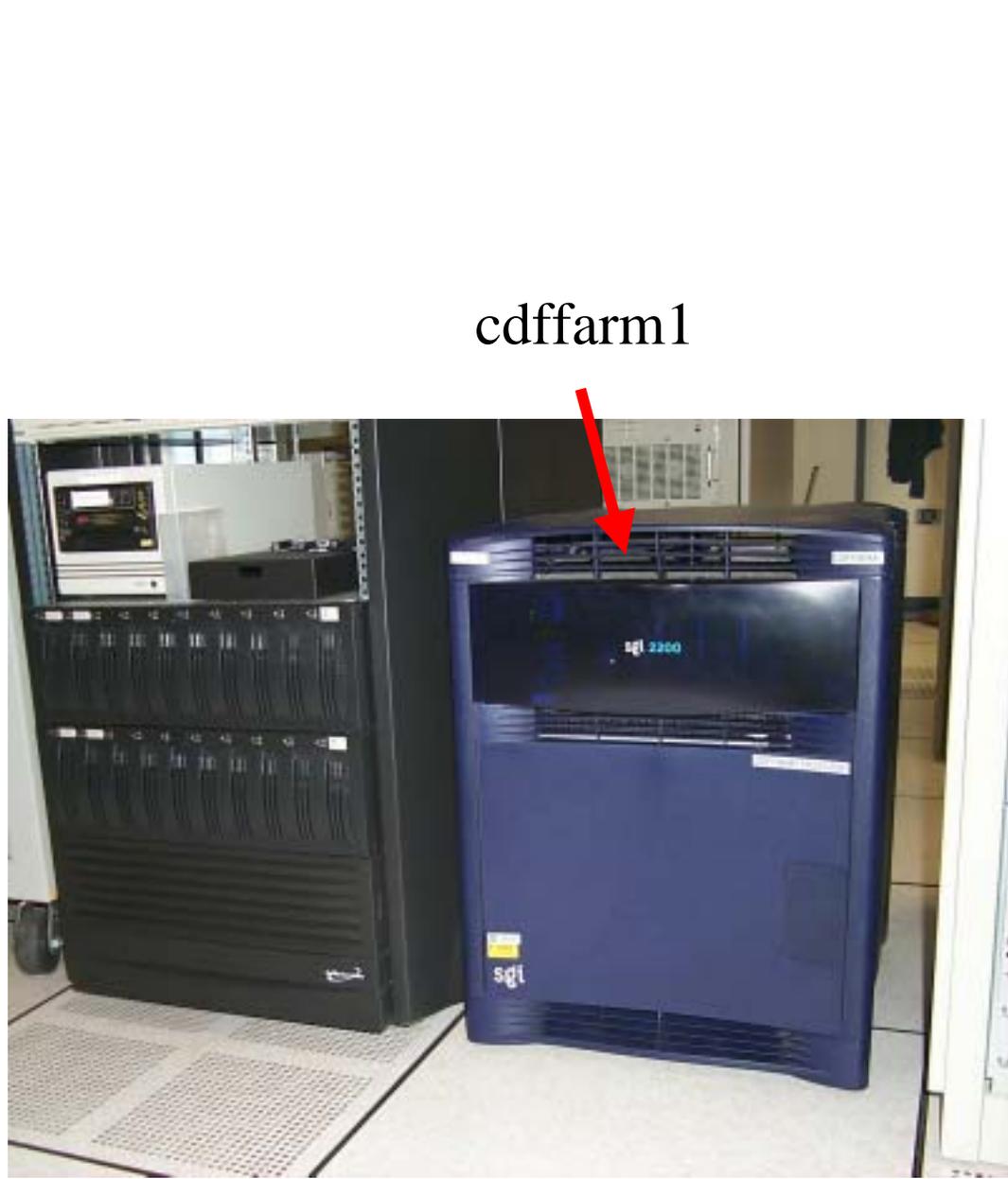


October 4, 2001

Stephen Wolbers, CDF



October 4, 2001



Stephen Wolbers, CDF



# 1. Status of 4.0.0x Processing

- Took some time to start
  - Executable needed modifications
  - I/O system wasn't ready (couldn't stage tapes properly, a few other problems)
- Processing begun on 9/20/01
- Stopped on 9/21/01
  - Problems with DIM/Kahuna/stager
- Processing restarted 9/27/01
- Stopped 10/1/01
  - Problems with executable (crashes) and splitting (maybe)
- Current Status
  - Stream A is finished (to run 127412) : 740K events
  - Stream J is finished (to run 127409) : 6.0M events
  - Stream B : 2.0M events
  - Stream G : 500K events
- 4.1 next?



## 1. 4.0.0x

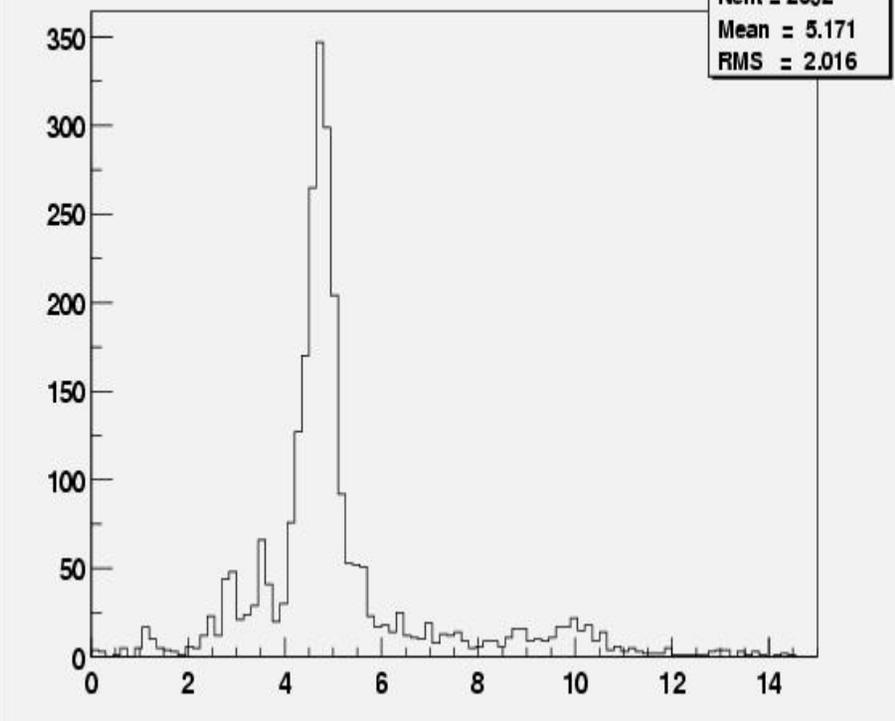
- Why did it take so long to process August/September data?
  - CDF ran short of tapes so raw data properly took priority
  - When that was solved there was a big change in the works: 3.18.3 to 4.0.0.
  - It was decided by CDF to wait for 4.0.0.
  - Then it took time to get 4.0.0 into good shape for production, eventually leading to 4.0.0i



4.0.0i.eps



exec time per file per event





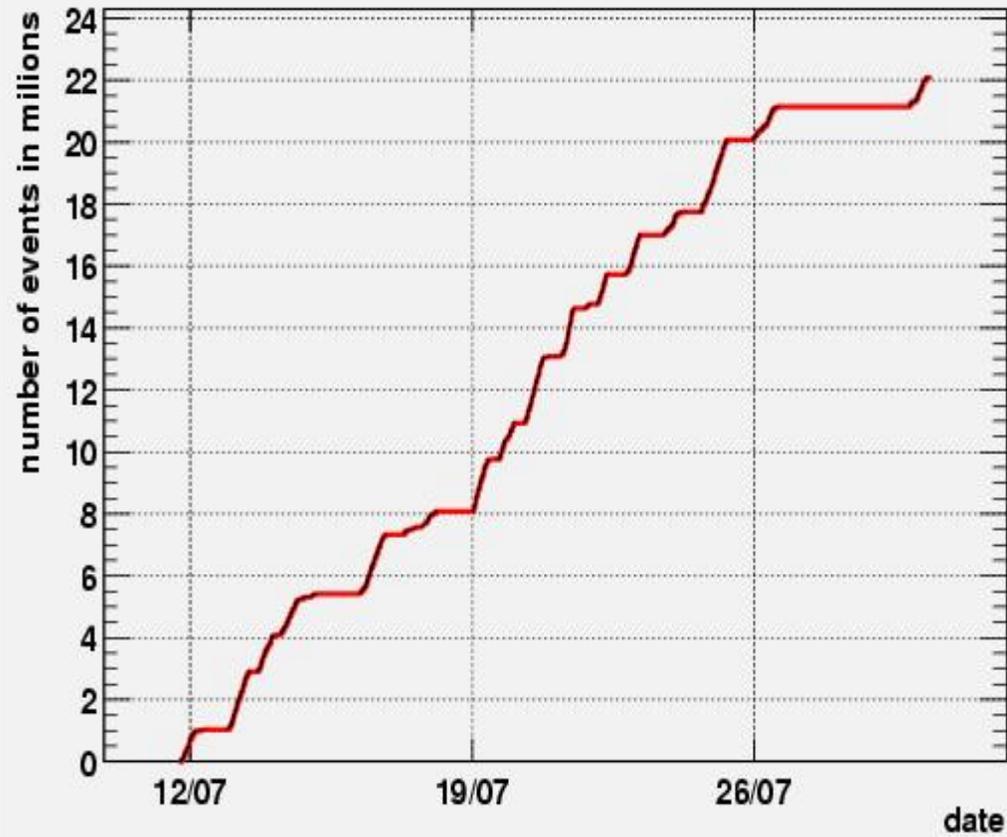
## June-July 2001 Data

- **First substantial data taken in Run 2**
  - Approximately 34 million events (“good runs”).
    - Approximately 8.5 TB of data.
    - Approximately 10 TB of output data.
    - ~3.5 seconds/event on PIII/500
  - I/O system was still not fully operational at this time, and this led to a backlog of data.
  - Long accelerator downtime (unplanned) allowed the farms to catch up with the backlog of data.
  - Code modifications (mainly due to detector changes) were common.
  - Procedures for providing proper calibrations were tuned up during this time.
  - Full splitting into many output datasets was implemented.



## June/July Processing

3.18.0.eps



3.17.1.eps





## 2. Express

- **Currently**
  - All stream A files are copied to a LOOK area on fcdfsgi2.
  - These files are immediately processed on fcdfsgi2 using the latest stable version of ProductionExe with the correspondent calibration constants.
  - The output goes to another LOOK area on fcdfsgi2, /cdf/data01/stream\_a.
    - Note that these files do NOT go onto tapes and are replaced when the disk space is needed to store new output files or the output files from the Farms Stream A processing are generated.
  - All Stream A data is processed on the farms when the newest best available ProductionExe is ready. The output files go onto tapes and are logged in the DFC.



## 2. Express

- **Issues for expressline:**
  - Is the current processing strategy optimal?
    - Do people get the data quickly enough?
    - Are the calibrations good enough?
  - Should an intermediate expressline be developed?
    - Process filesets (not files) as soon as they are available.
    - Needs some changes to the DH system to implement.



## 3. Datasets

- Currently, there are 4 streams and 7 datasets
- Streams:
  - A: Express (aphysr)
  - B: Electron/Photon (bphysr)
  - G: Jets (gphysr)
  - J: Muons (jphysr)
- Datasets:
  - Express (aphysa,b,c,...)
  - Electron (btst1a,b,c,...)
  - Photon (btst2a,b,c,...)
  - Jet (gtst1a,b,c,...)
  - Et (gtst2a,b,c,...)
  - Minbias (gtst3a,b,c,...)
  - Muon (jphysa,b,c,...)



## 3. Datasets

- The definitions of streams and datasets is an input to the farms, it is not a problem for the farms to have them change.
- Currently a tape's worth of data must be available to get the farm output to tape.
  - This means that each dataset being created must have disk space on fcdfsgi1 for a full tape plus enough to store data while the tape is being filled and checked.
  - This is estimated to be 2.5 times a full tape or 125 GB for each dataset that is being filled.
  - For the current 7 datasets, this is 875 GB, which now is available on fcdfsgi1 (for the past week or so).



## 4. I/O and CPU

- I/O is the issue on the farms.
  - The full farm can easily push more data to the DH system than it can handle easily (as well as pull more data from the system than it can handle).
  - As the ProductionExe slows down and the output gets smaller, this becomes less and less a problem.
  - A large capacity, smooth running fcdfsgi1 will help.
  - There is plenty of CPU for the moment for reconstruction and simulation.



## 4. I/O and CPU

- PADs
  - Making PADs out of production is not a problem.
  - In fact, it is what the system was designed to do.
  - The farms welcome such a step as soon as possible.



## 5. Long-term plans

- Ideally, the farm should keep up with data-taking.
- This means first that Expressline is reconstructed as soon as possible (depends on how fast the data and calibrations are available).
- All streams should be reconstructed quickly.
  - “Quickly” means:
    - Input tape for a stream must be filled.
    - Tape is staged on fcdfsgi1.
    - Farm reconstructs and splits the output.
    - Output is concatenated and saved to disk on fcdfsgi1.
    - Output for a dataset is written to tape when a full tape’s worth is available.



## 5. Long-term plans

- **Can we run all streams at once?**
  - At the moment, yes. There is enough output disk space on fcdfsgi1 available.
  - With more datasets the answer is probably no.
- **Possible options**
  - Rotate through streams (this is what we have done)
  - Get more disk space on fcdfsgi1 (DH work)
  - Share datasets on tapes (DH)



## 6. A short Introduction to the CDF Production Farms

- Design Goals
- Architecture
- Hardware and Interface to Datahandling
- Software and Control

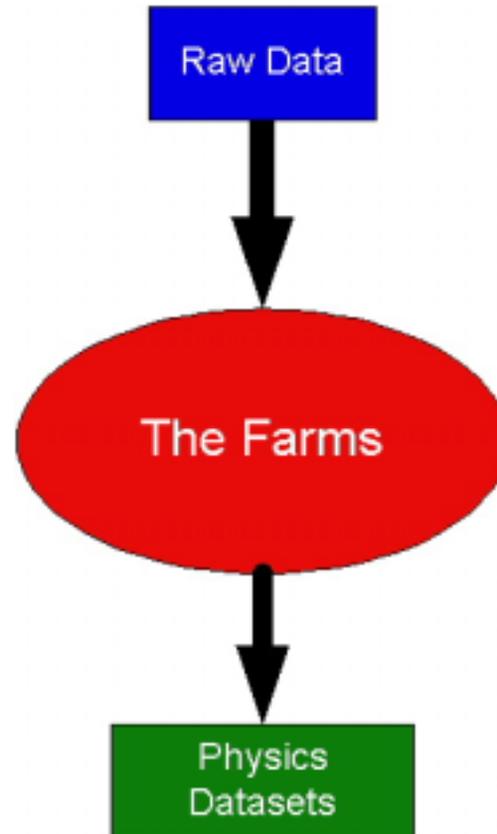


## CDF Offline Production Farms for event reconstruction - Design Goals

- The CDF farms must have sufficient capacity for Run 2 Raw Data Reconstruction.
- The farms also must provide capacity for any reprocessing needs and Monte Carlo.
- Farms must be easy to configure and run.
- The bookkeeping must be clear and easy to use
- Error handling must be excellent.

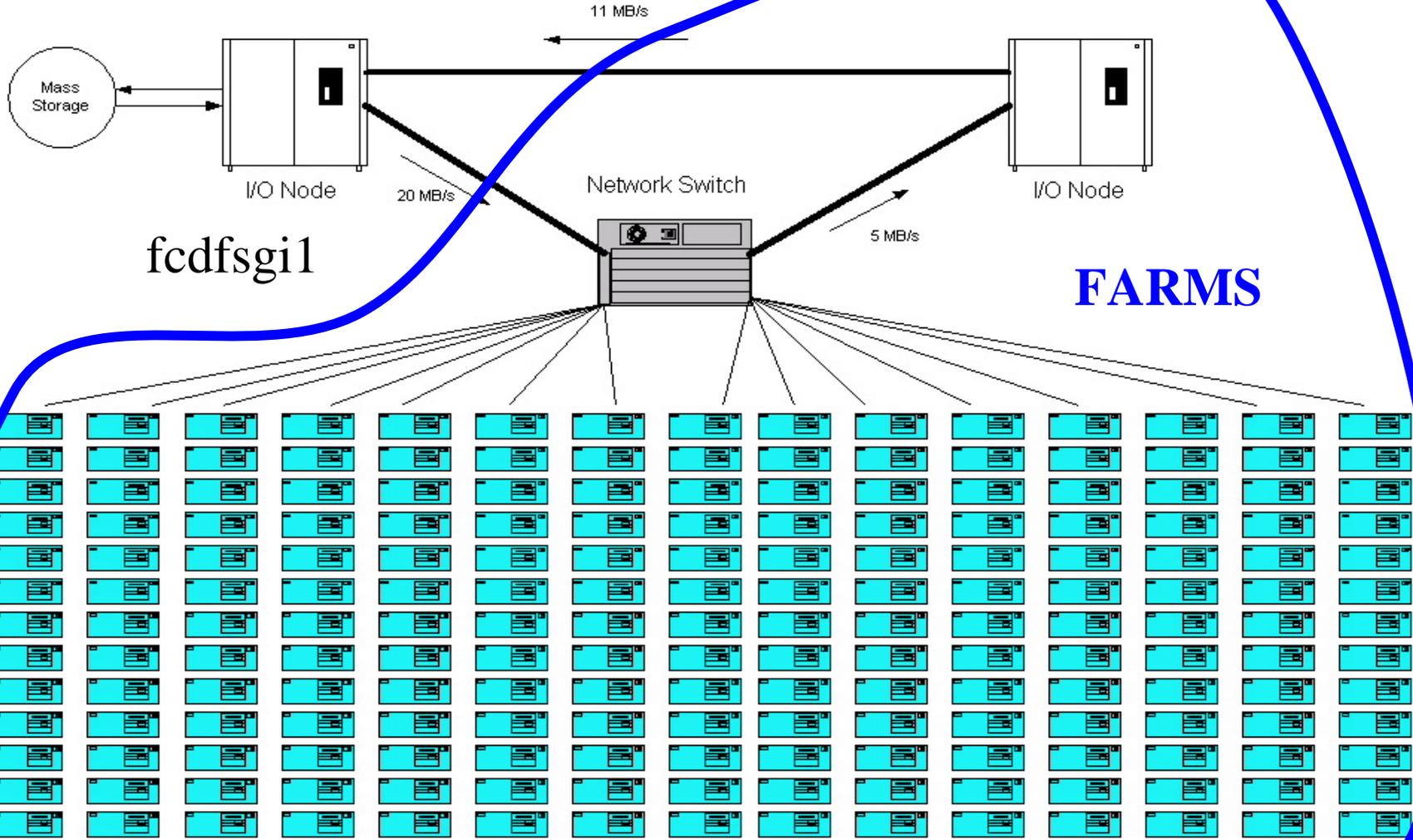


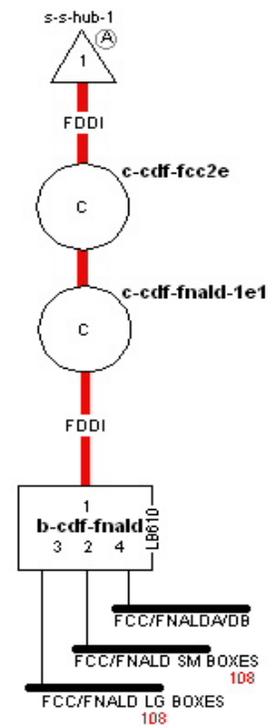
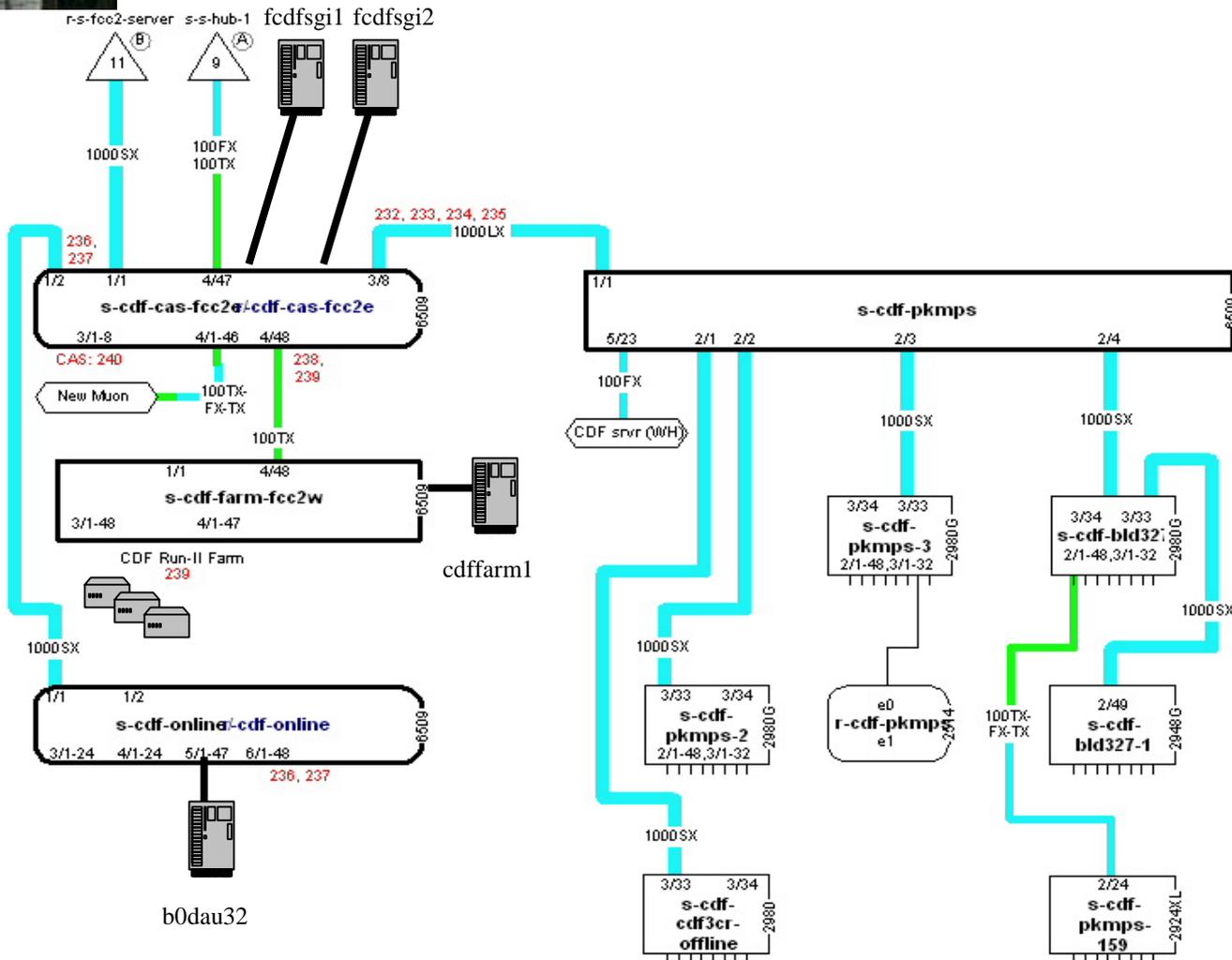
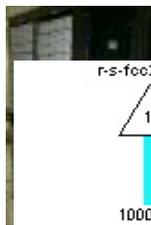
## Simple Model





# Run II CDF PC Farm







## Summary

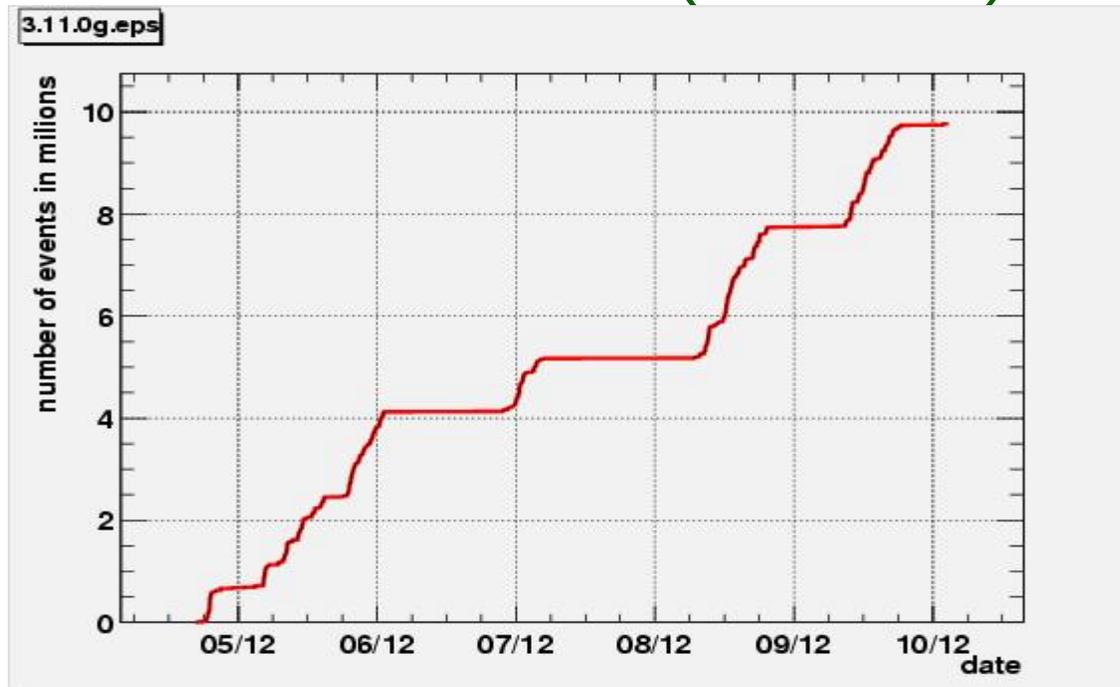
- The farms have a huge amount of CPU.
- Most (not all!) problems have been:
  - Tapes (lack of)
  - I/O capacity and functionality
  - ProductionExe not ready
  - ProductionExe crash rate too high
  - Calibrations not ready
- Progress is being made
  - More disk on fcdfsgi1 means a smoother operation with more I/O capacity
  - ProductionExe 4.0.0 is available and running
  - Farm software is constantly improving
  - 3-3.5 Million events/day can be sustained through the farm at this time (using only about ½ of the CPU)



# Early Processing Experience

## • Commissioning Run (October, 2000)

- Ran 4 weeks after data was collected.
- 9.8 Million Events, 730 GB input, 1080 GB output.
- CPU/event = 1.2 seconds (PIII /500)





## Lessons from Commissioning Run

- Data size was not an issue. Farms could easily keep up.
- I/O was problematic. It was easy to flood the system, fill disk buffers, etc.
- Reconstruction code was an issue. Modifications were common, leading to occasional delays.

# Early Processing Experience



## April 2001 Data

- First 36x36 bunch collisions.
- Ran about 1 week after data was taken.
- 5.1 Million Events, 1.2 TB input, 1.6 TB output
- CPU/event = 1.0 seconds (PIII /500)

