



# Fermilab Science and Computing

**Stephen Wolbers**

**Fermilab**

**May 10, 2001**



## What do we do at Fermilab?

- Study the fundamental nature of particles and their interactions.
- This is done with particle accelerators and with large and sophisticated detectors.
- About 2100 employees, over 2000 scientific visitors.
- Lots of data, lots of computing.



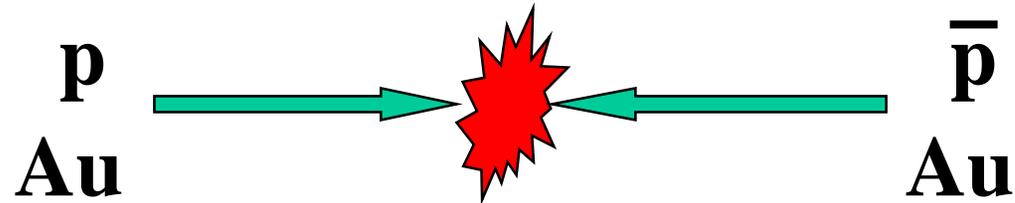
May 15, 2001

Stephen Wolbers, IBM Tucson Visit

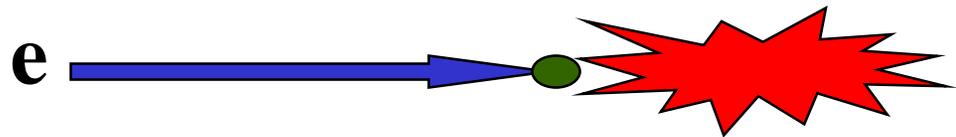


## Collisions Simplified

- Collider:

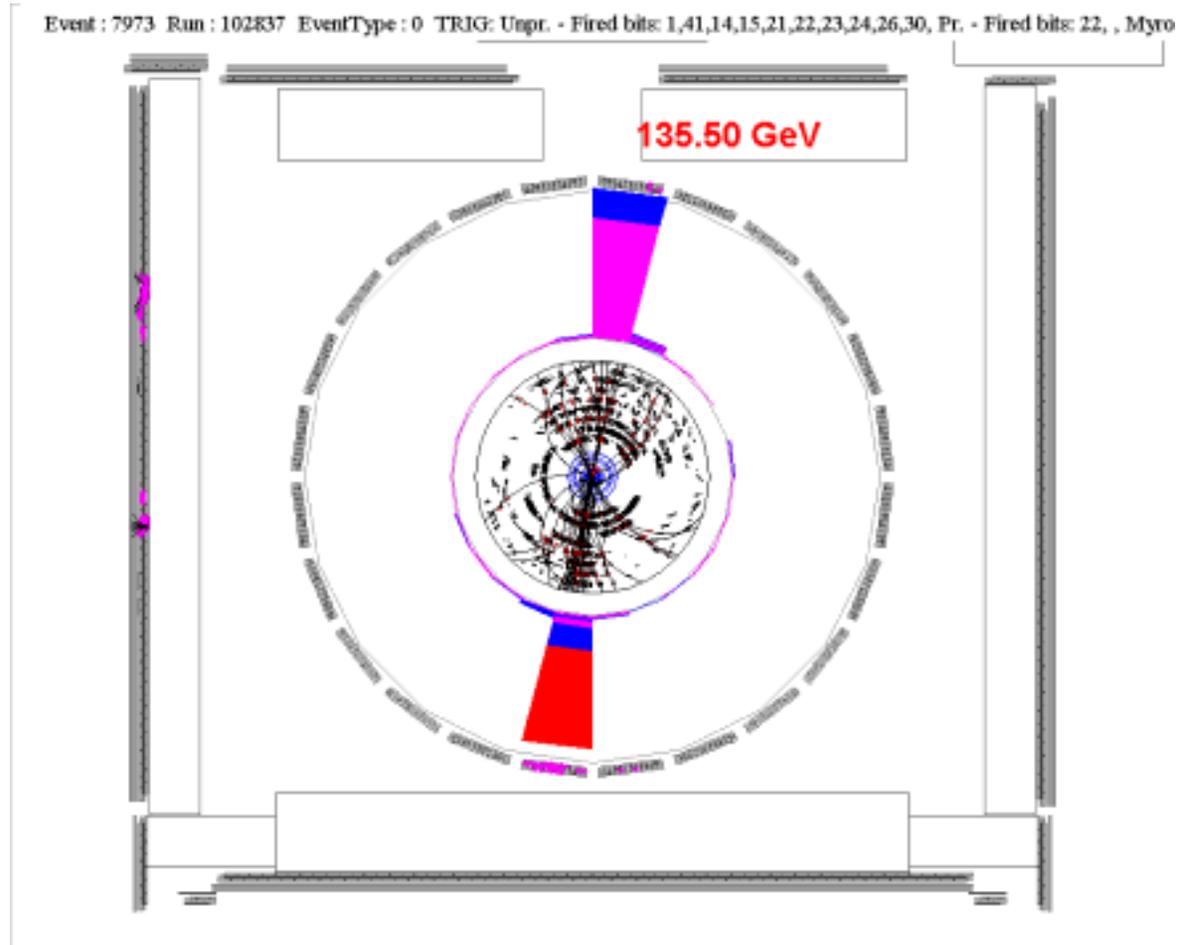


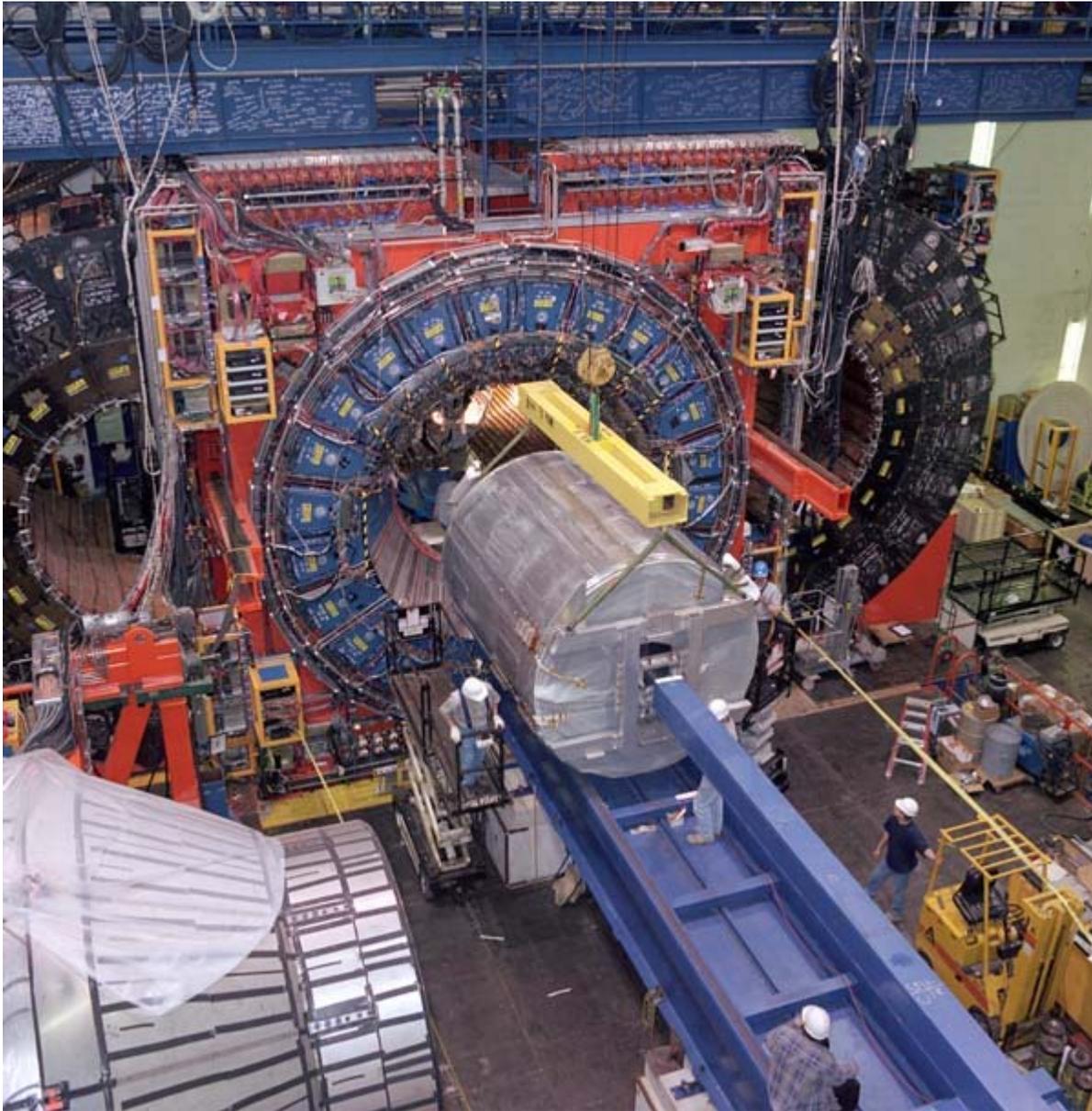
- Fixed-Target:





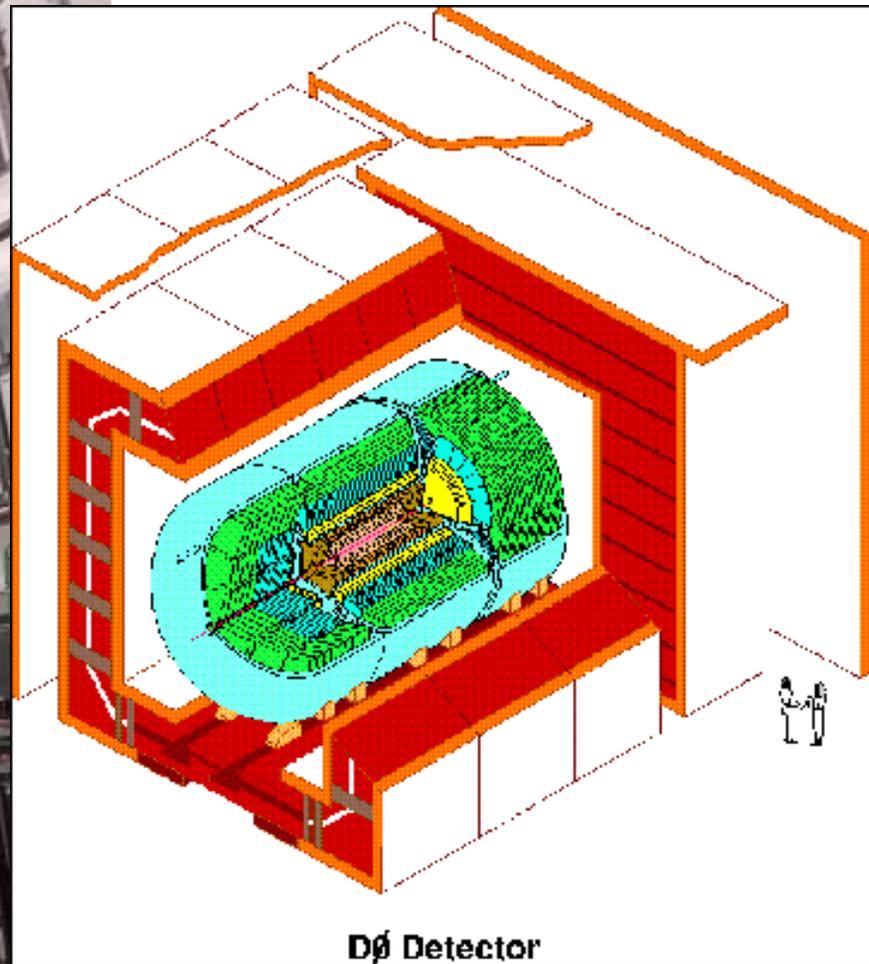
# Collisions – CDF/Fall 2000





May 15, 2001

Stephen Wolbers, IBM Tucson Visit

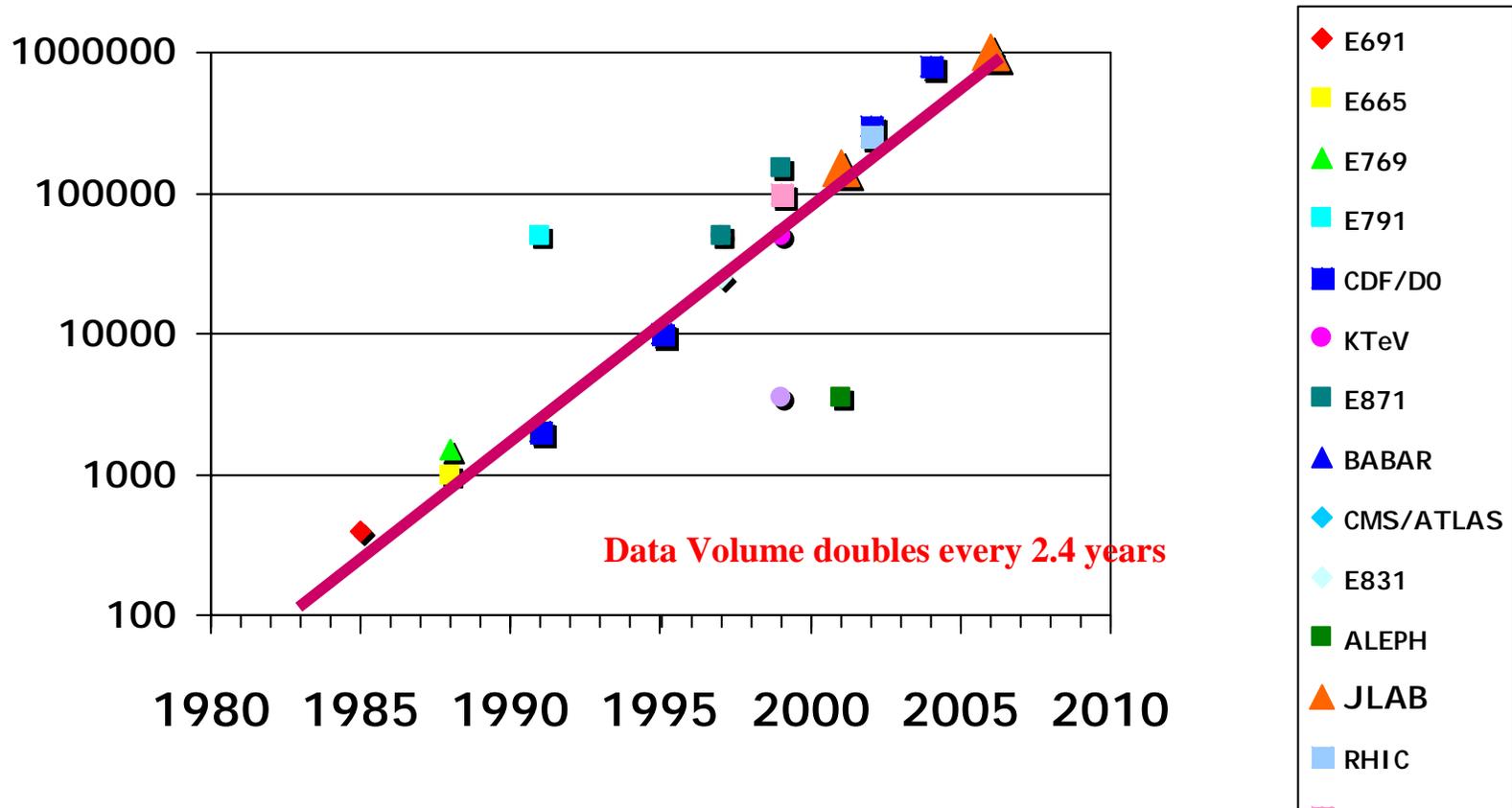


May 15, 2001

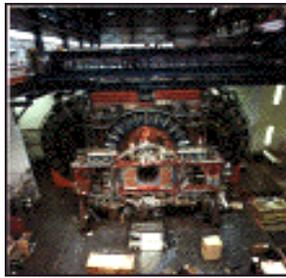
Stephen Wolbers, IBM Tucson Visit



# Data Volume per experiment per year (in units of $10^9$ bytes)



# RunII Data Flows



15 MBps

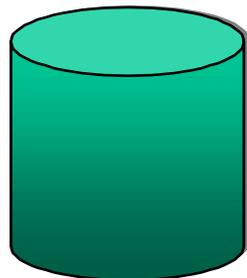
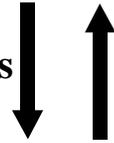


100 Mbps



20 MBps

400 MBps



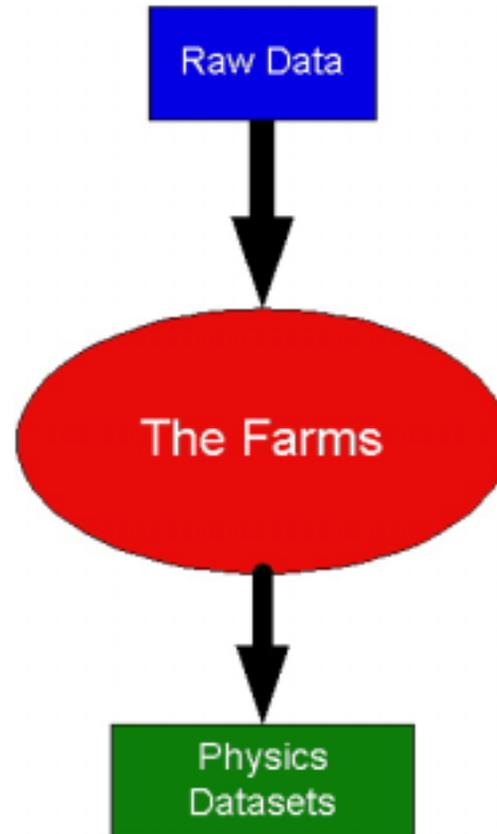
Stephen Wolbers

HEP-CCC

June 25, 1999

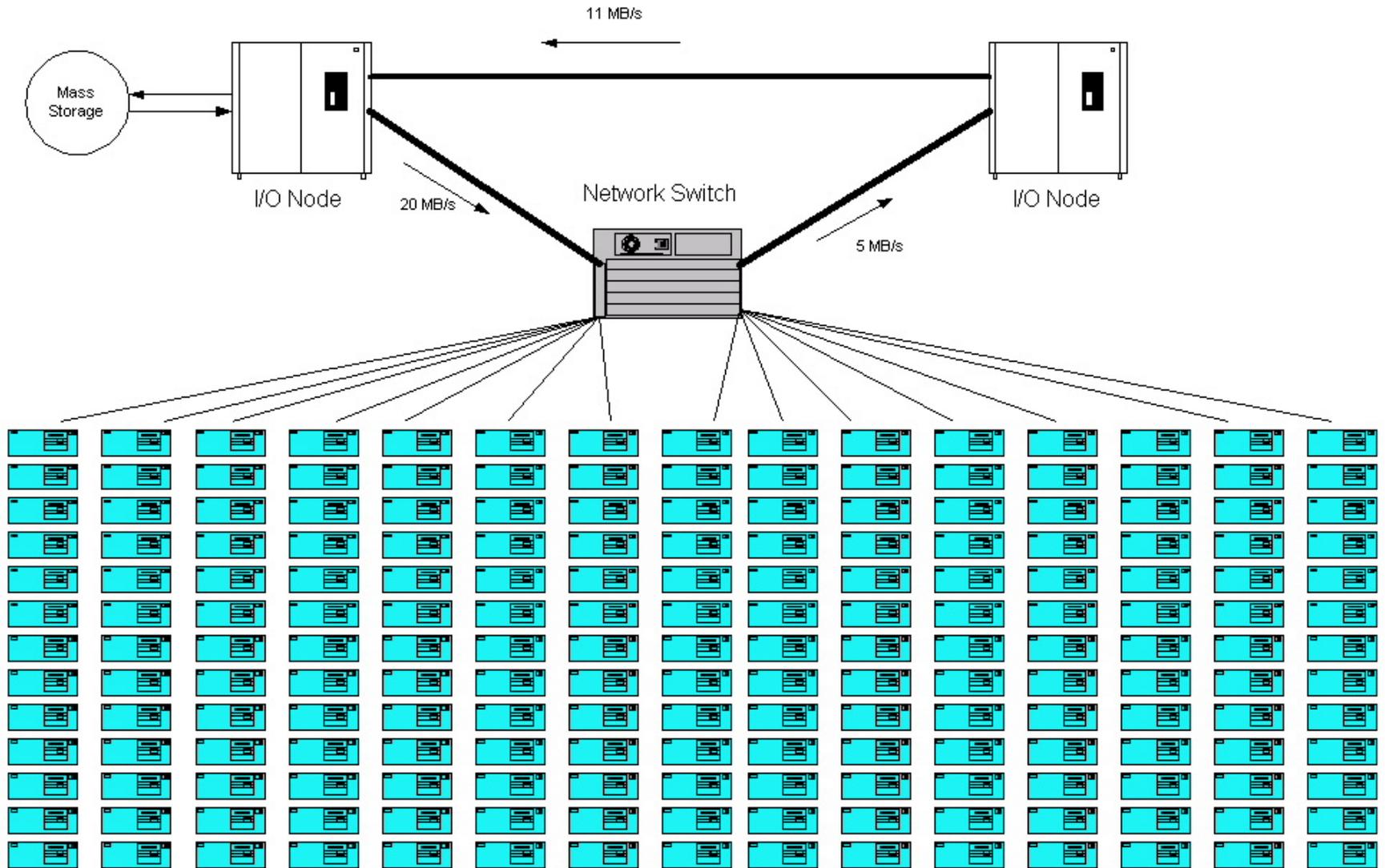


## Simple Model





# Run II CDF PC Farm





May 15, 2001

Stephen Wolbers, IBM Tucson Visit

12



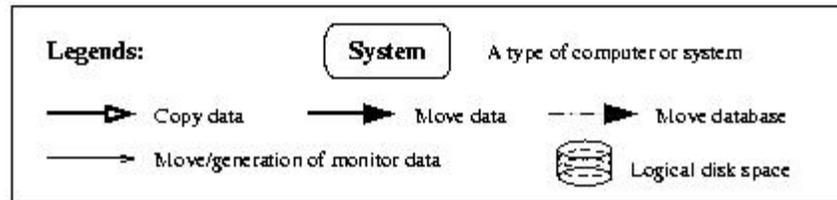
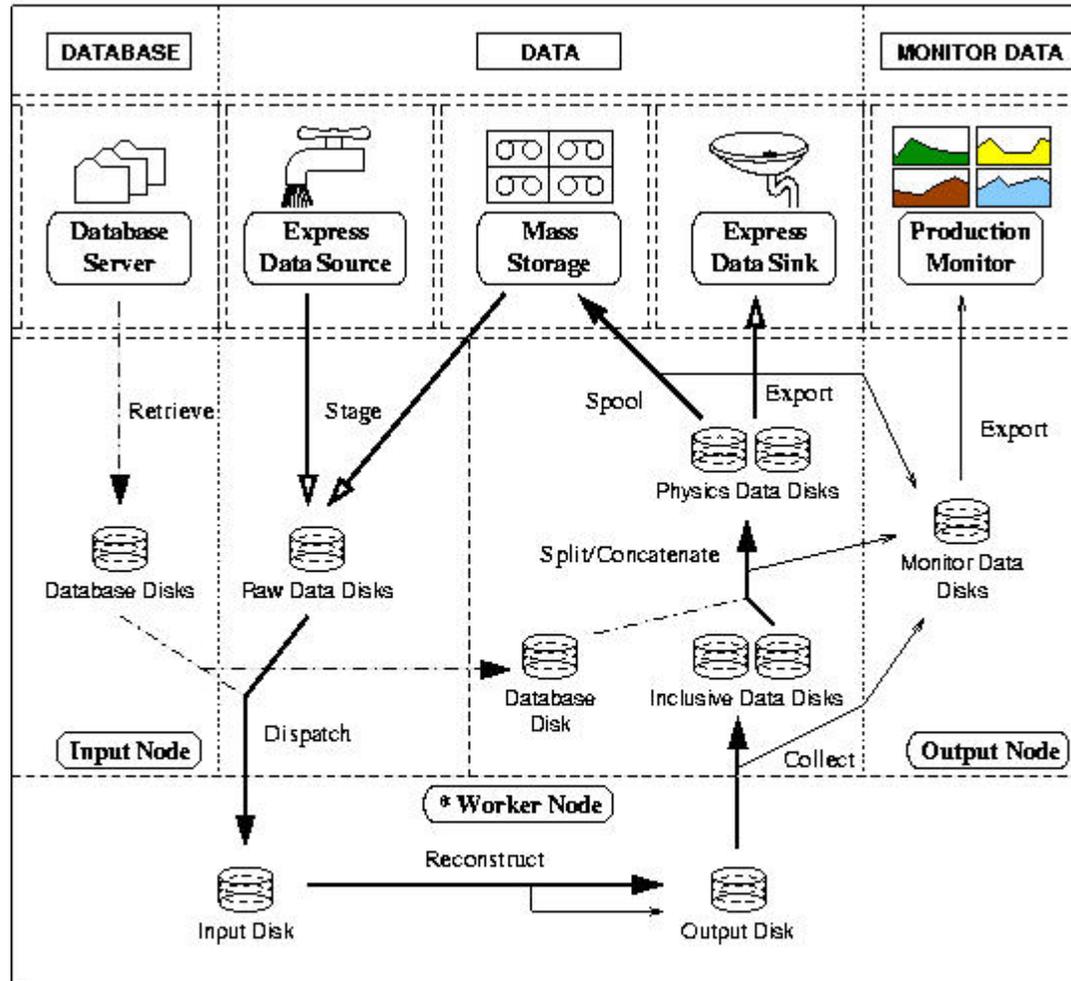
May 15, 2001

Stephen Wolbers, IBM Tucson Visit

13



# Conceptual Model of Run 2 Production System



## Analysis Computing - Run 2a

- CDF and D0 have both acquired large Silicon Graphics O2000 multi-processor systems for large analysis tasks.



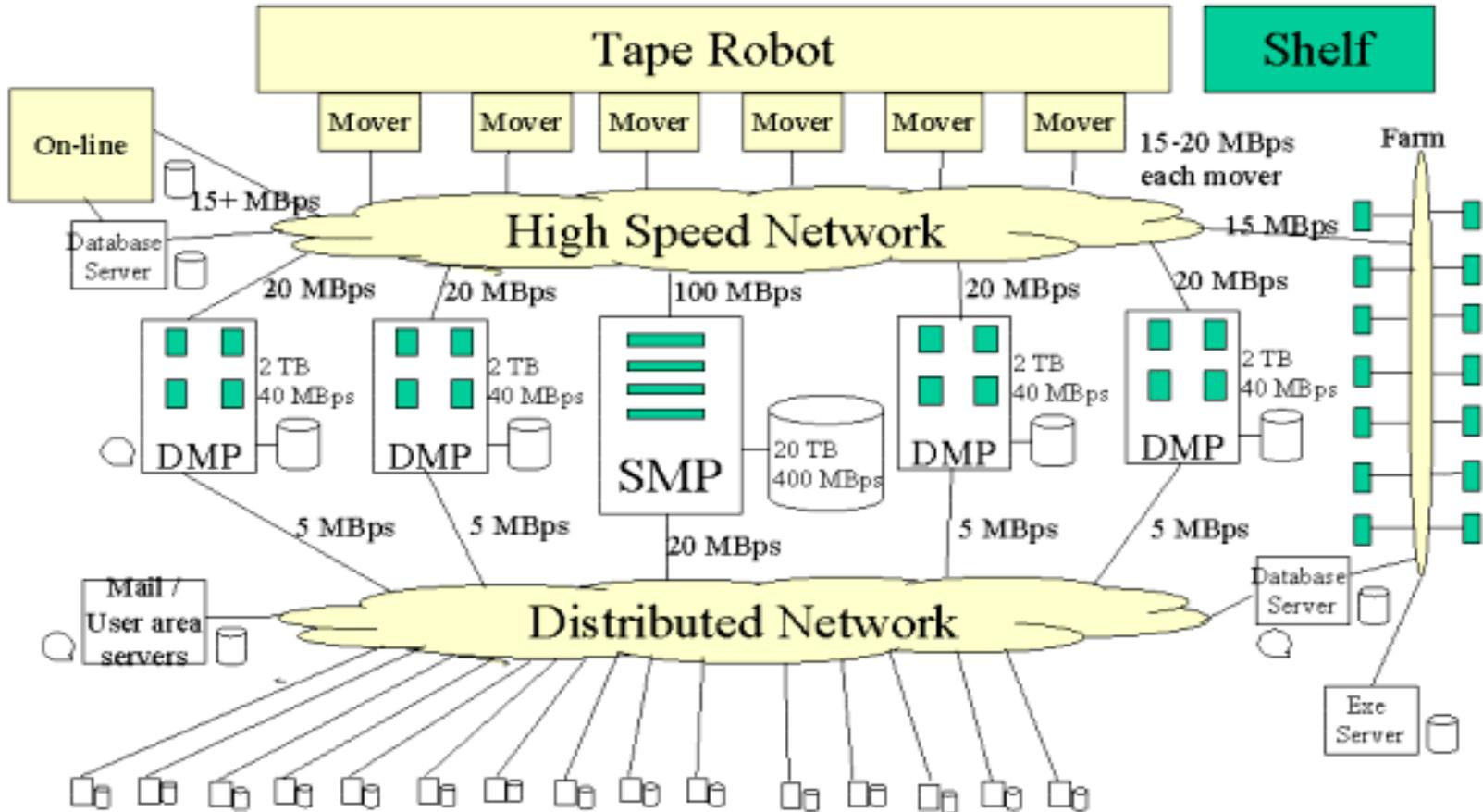


# Analysis Computing

- Each system has:
  - Access to data on tape.
    - D0 access is over the network
    - CDF access is via SCSI connected peripherals
  - Access to disk storage.
    - About 30 Tbytes attached to central systems
    - This will increase, especially as disk prices continue to fall
  - LSF Batch software is used to schedule jobs and manage resources on these systems.



# Proposed D0 Analysis Computing Configuration



400 desktop : 50GB, 0.1MBps (avg), 10 MBps (burst) ea

SMP = Symmetric Multi-processor  
DMP = Distributed multi-processor

○ Tape Backup

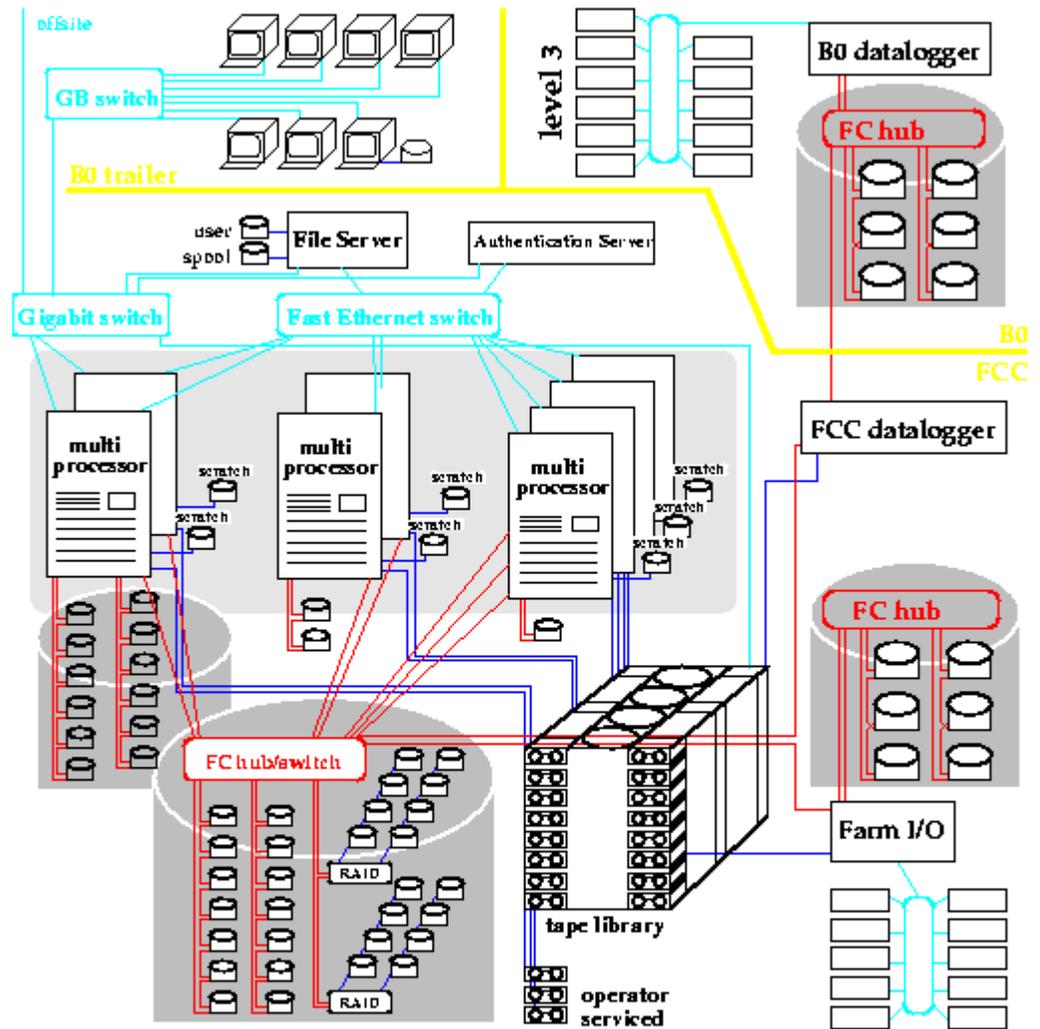
9/21/98

# Data Access Model: CDF



- Ingredients:**

- Gigabit Ethernet
- Raw data are stored in tape robot located in FCC
- Multi-CPU analysis machine
- High tape access bandwidth
- Fiber Channel connected disks





## Analysis Computing

- The large SGI is a conservative (and expensive) solution to analysis computing needs.
- Both collaborations are exploring the use of PC's + EIDE disk + 100 Mbit or 1 Gbit network connection for analysis.
- These projects may lead the way to more cost-effective solutions for the analysis of the large amount of data that will be taken in Run 2a and Run 2b.



## PC analysis computing -- examples



May 15, 2001

Stephen Wolbers, IBM Tucson Visit

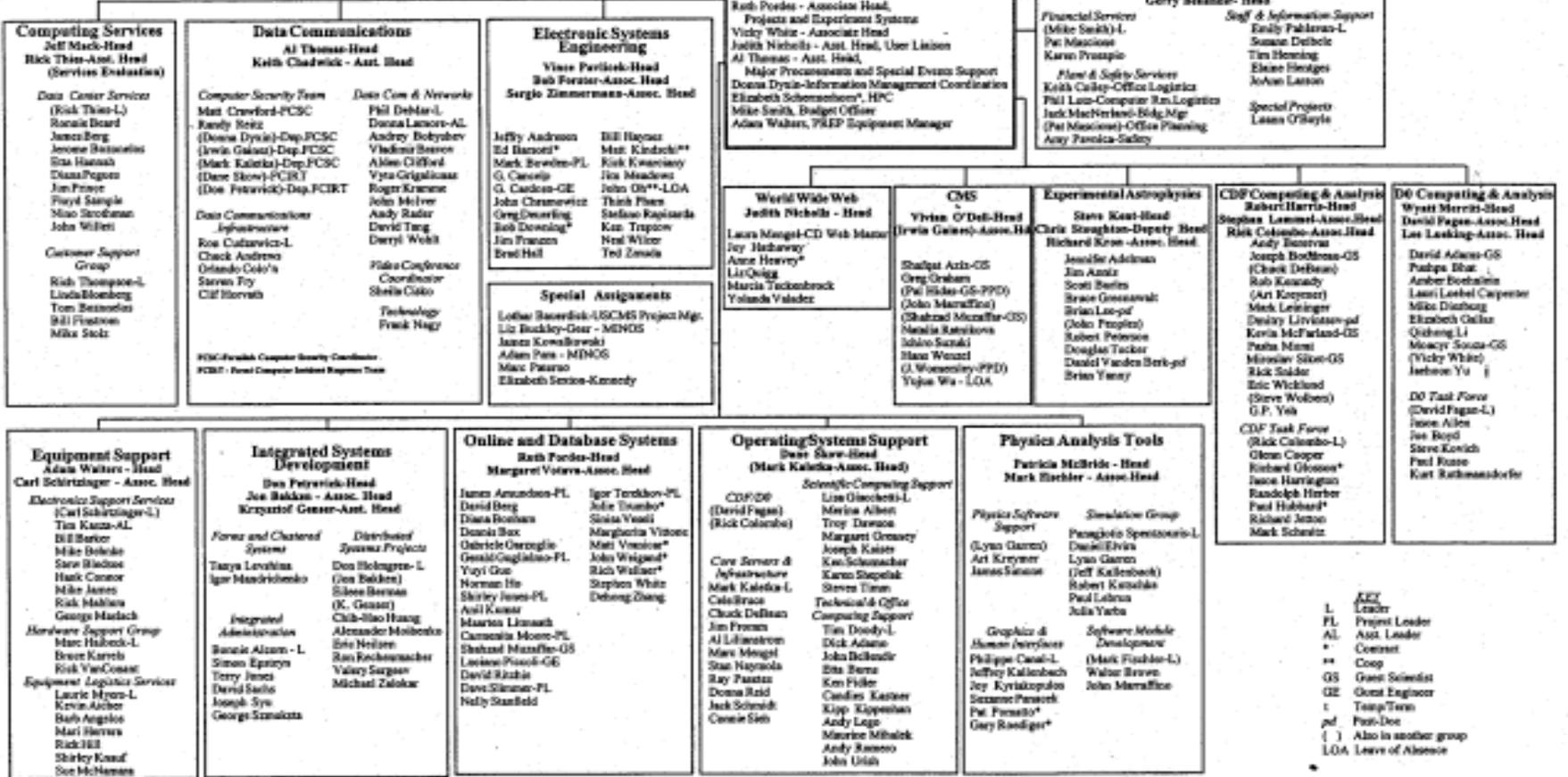
20



# Computing Division Organization Chart

APRIL 9, 2001

Approved: *M. Kammann* 4/10/01  
 Matthew Kammann - Head, Computing Division Date



- L KEY
- PL Leader
- AL Asst. Leader
- \* Content
- \*\* Coop
- OS Guest Scientist
- GE Guest Engineer
- t Temp/Term
- pd Part-Time
- ( ) Also in another group
- LOA Leave of Absence



## Storing Petabytes of Data in mass storage

- Storing (safely) petabytes of data is not easy or cheap.
  - Need large robots (for storage and tape mounting).
  - Need many tapedrives to get the necessary I/O rates.
    - Tapedrives and tapes are an important part of the solution, and has caused some difficulty for Run 2.
  - Need bandwidth to the final application (network or SCSI).
  - Need system to keep track of what is going on and schedule and prioritize requests.



## Robots and tapes



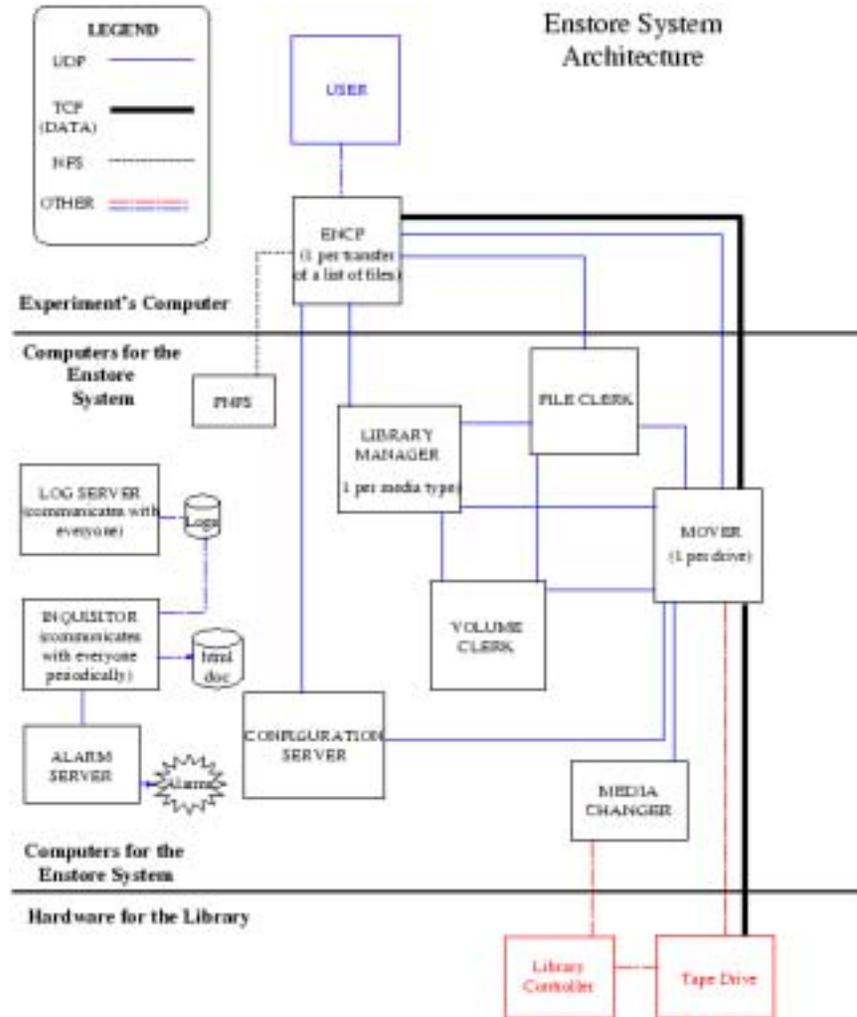
May 15, 2001

Stephen Wolbers, IBM Tucson Visit

23



# Enstore Software System





# Tapedrives and tapes

- Tapedrives are not always reliable, especially when one is pushing for higher performance at lower cost.
- Run 2 choice is Exabyte Mammoth 2.
  - 60 Gbytes/tape.
  - 12 Mbyte/sec read/write speed.
  - About \$1 per Gbyte for tape. (A lot of money.)
  - \$5000 per tapedrive.
- AIT2 from SONY is the backup solution.
- The robotics which exist can handle most any tapedrive technology.
- Given the Run 2 timescale, upgrades to newer technology will occur.
- Finally, Fermilab is starting to look at PC diskfarms to replace tape completely.

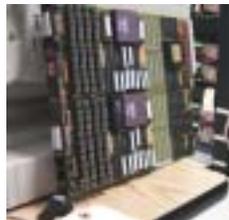


May 15, 2001

Stephen Wolbers, IBM Tucson Visit

26

# Old and New Lattice Gauge Computing at Fermilab



May 15, 2001

Stephen Wolbers, IBM Tucson Visit



## Run 2b at Fermilab

- Run 2b will start in 2004 and will increase the integrated luminosity to CDF and D0 by a factor of approximately 8 (or more if possible).
- It is likely that the computing required will increase by the same factor, in order to pursue the physics topics of interest:
  - B physics
  - Electroweak
  - Top
  - Higgs
  - Supersymmetry
  - QCD
  - Etc.



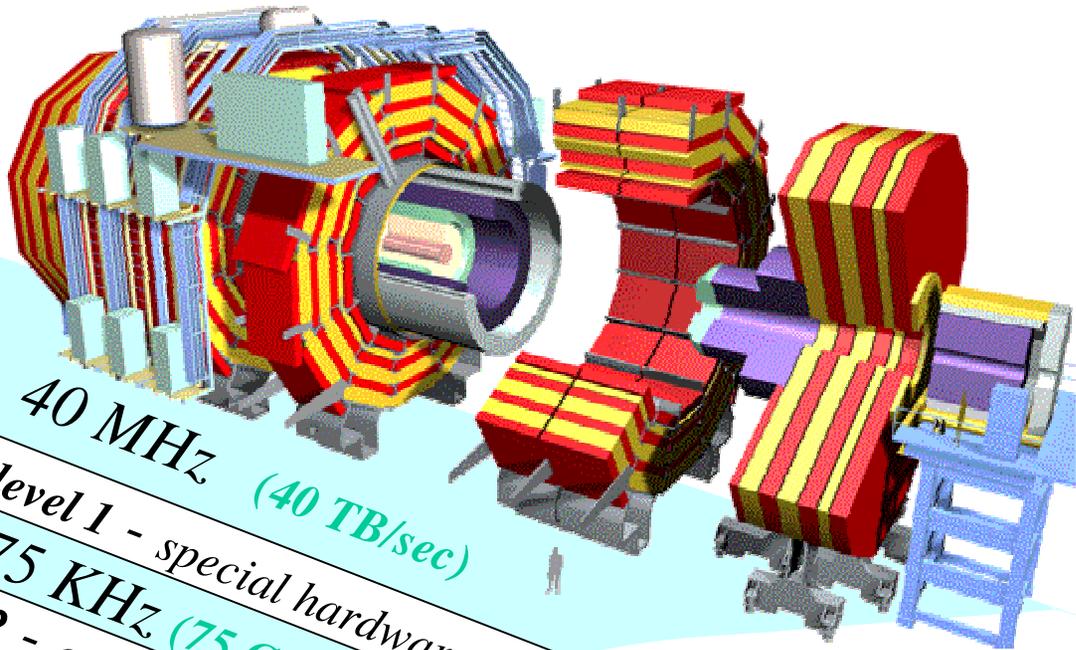
## Run 2b Computing

- **Current estimates for Run 2b computing:**
  - 8x CPU, disk, tape storage.
  - Expected cost is same as Run 2a because of increased price/performance of CPU, disk, tape.
  - Plans for R&D testing, upgrades/acquisitions will start next year.
- **Data-taking rate:**
  - May be as large as 80 Mbyte/s.
  - About 1 Petabyte/year to storage.



## LHC Computing

- LHC (Large Hadron Collider) will begin taking data in 2006-2007 at CERN in Switzerland.
- Data rates per experiment of 100 Mbytes/sec.
- 1 Pbyte/year of storage for raw data per experiment.
- World-wide collaborations and analysis.
  - Desirable to share computing and analysis throughout the world.
  - GRID computing may provide the tools.



40 MHz (40 TB/sec)  
level 1 - special hardware

75 KHz (75 GB/sec)  
level 2 - embedded processors

5 KHz (5 GB/sec)  
level 3 - PCs

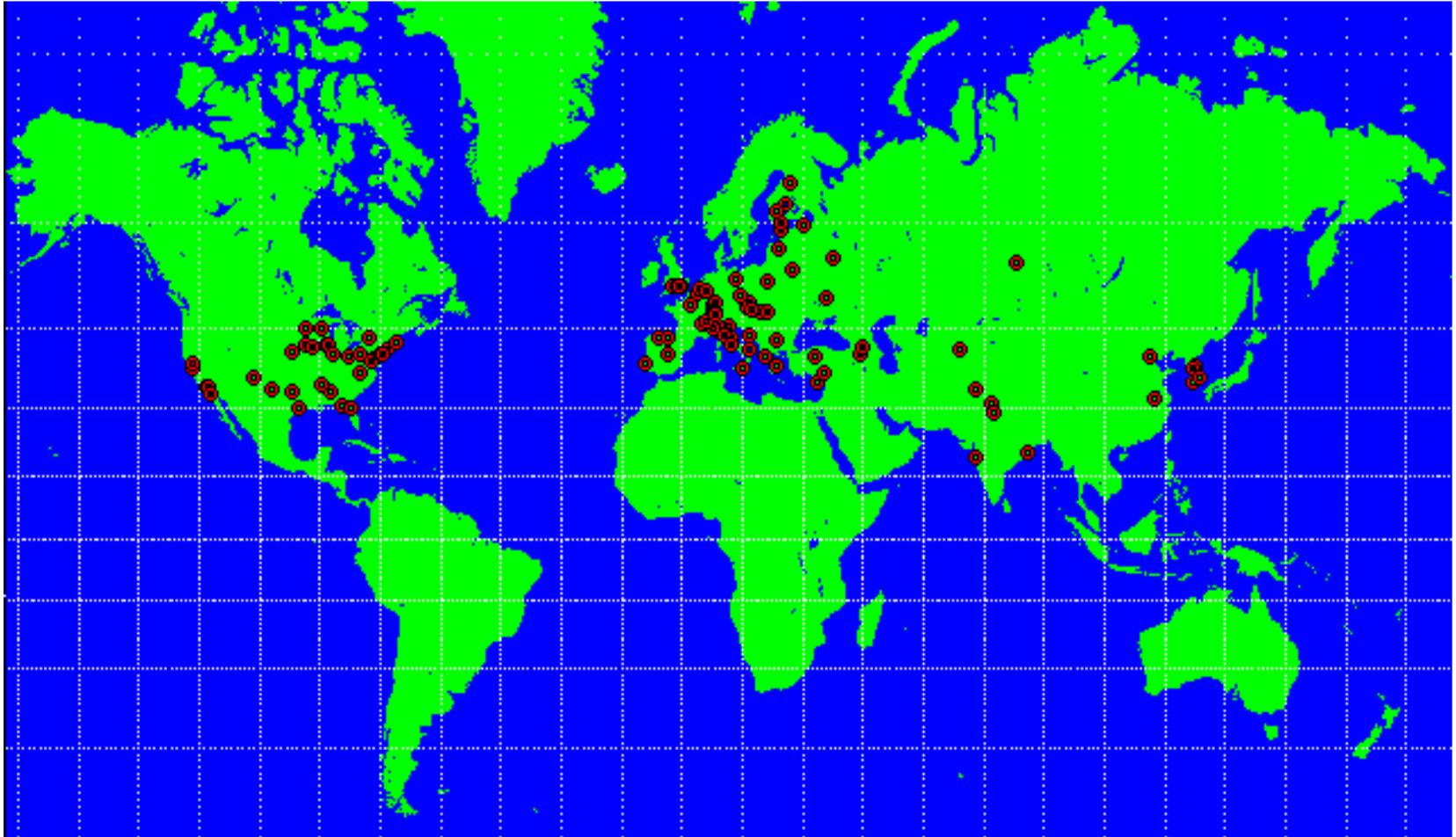
100 Hz  
(100 MB/sec)

data recording &  
offline analysis



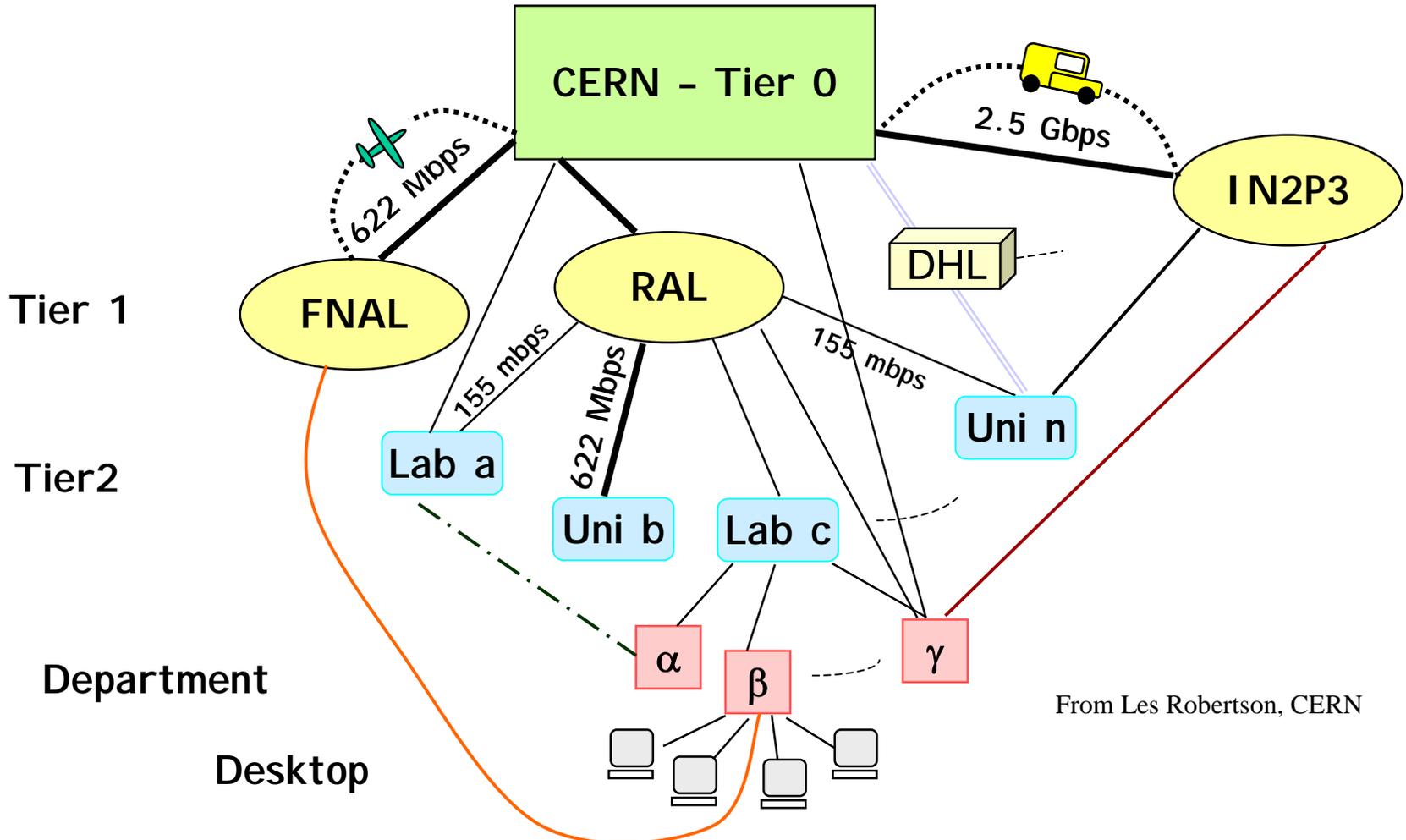


*World Wide Collaboration*  
⇒ *distributed computing & storage capacity*





# CMS/ATLAS and GRID Computing



From Les Robertson, CERN



# What does the Grid do for you?

Les Robertson

- You submit your work
- And the Grid
  - Finds convenient places for it to be run
  - Organises efficient access to your data
    - Caching, migration, replication
  - Deals with authentication to the different sites that you will be using
  - Interfaces to local site resource allocation mechanisms, policies
  - Runs your jobs
  - Monitors progress
  - Recovers from problems
  - Tells you when your work is complete
- If there is scope for parallelism, it can also decompose your work into convenient execution units based on the available resources, data distribution