Computing Experience from CDF and D0

Stephen Wolbers
Fermilab
CHEP2003, March 24, 2003

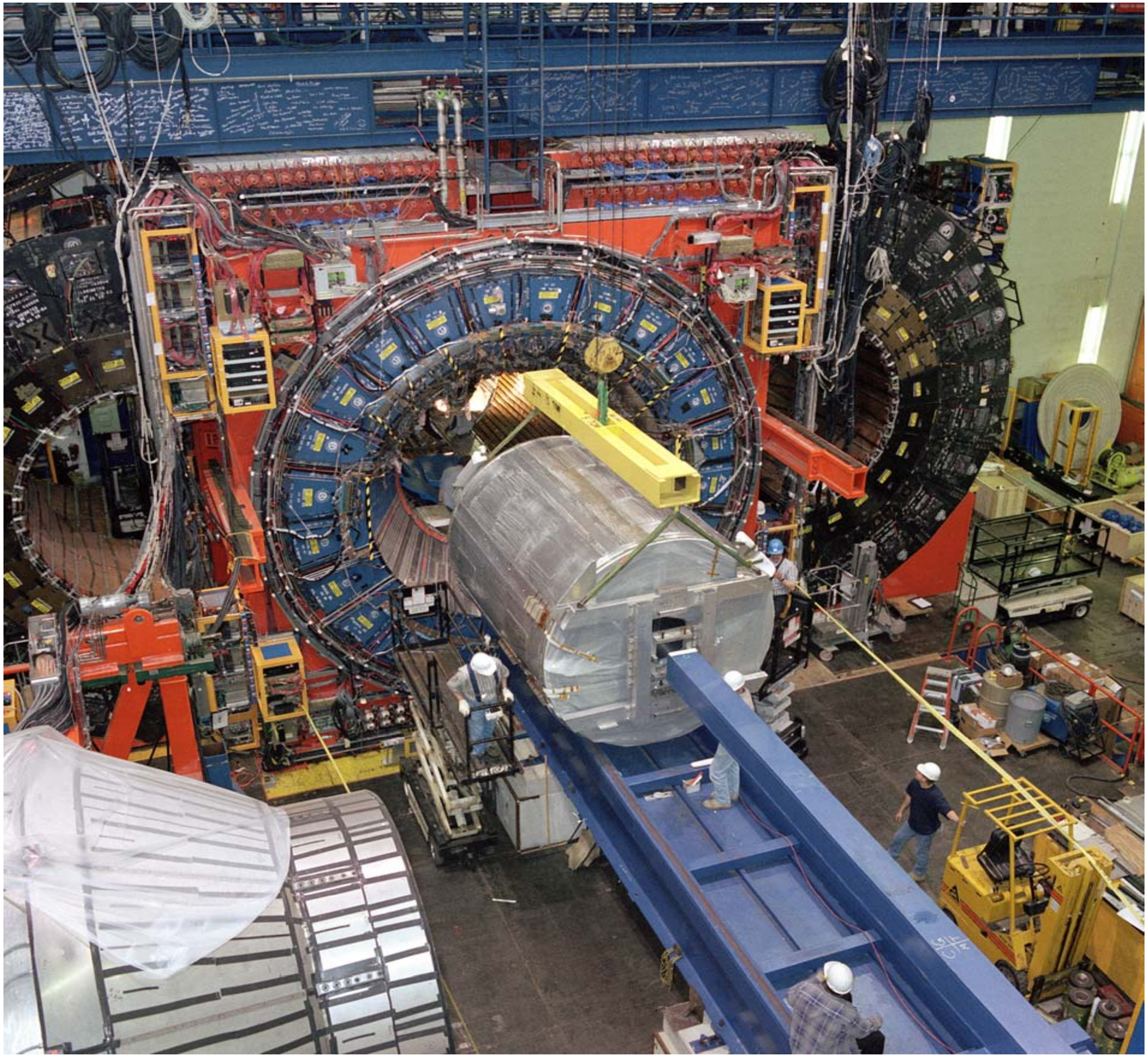
Outline

- Run 2 overview
- Computing experience/issues
- Mid-course corrections
- Future directions
- Conclusions

My viewpoint/bias: Deputy Head of FNAL CD 1997-2003, technical work on CDF production farms.

Run 2 Overview

- Run 2 officially began March 1, 2001.
- Planning for Run 2 computing started many years ago and work has been continuous.
- Luminosity lower than expected, but it is increasing and will continue to increase.
- Both CDF and D0 have been taking data steadily at high rates limited primarily by DA capability. (Triggers are adjusted as luminosity increases.)
- Big detectors, large collaborations, many challenges.
- Computing is a big issue – essential for physics.
- Run 2 has many years to go.



The CDF Collaboration

North America

3 Natl. Labs
28 Universities

2 Universities

Europe

1 Research Lab
6 Universities

1 University

4 Universities

2 Research Labs

1 University

1 University

1 University

Asia

4 Universities
1 Research Lab

1 University

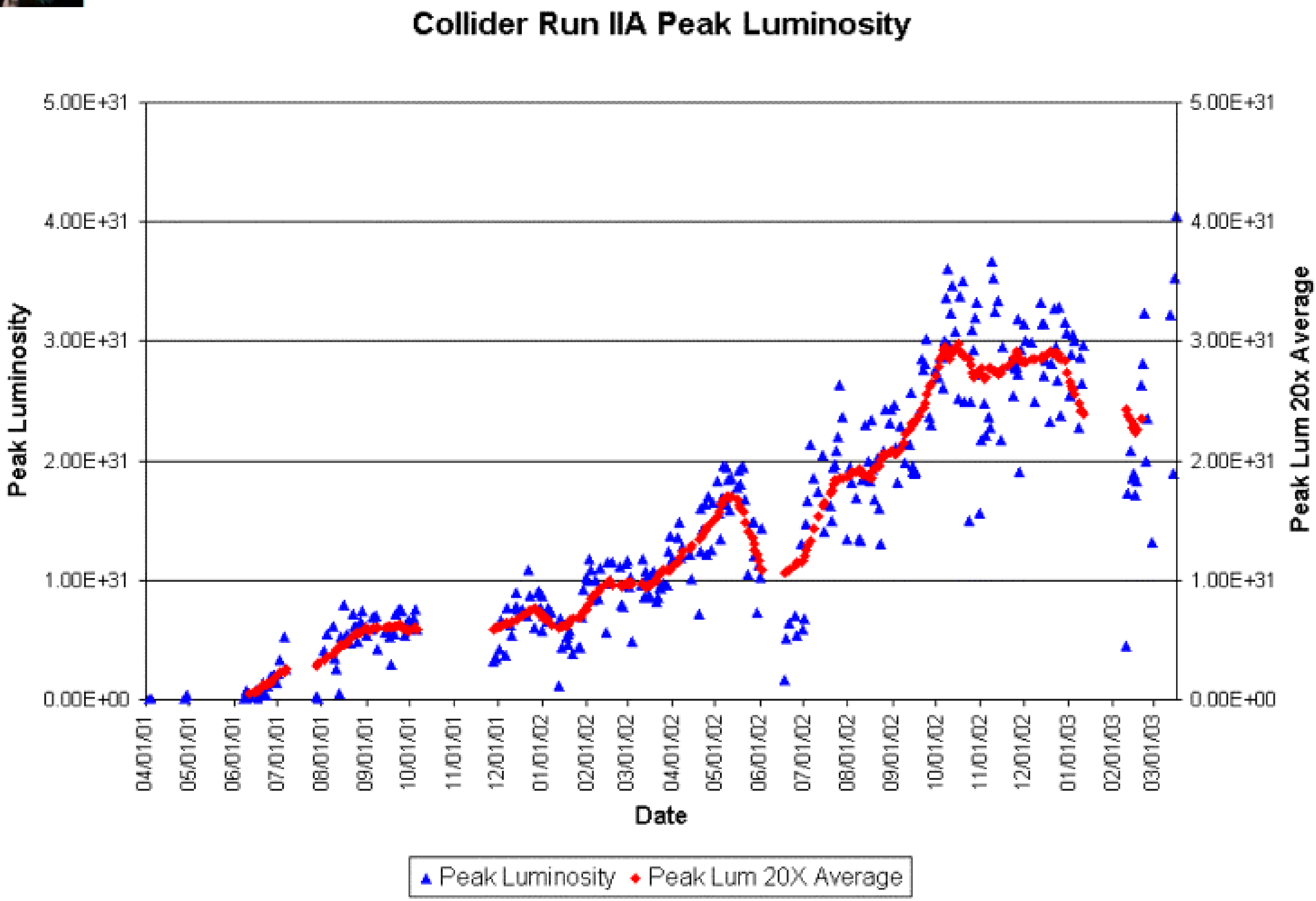
3 Universities

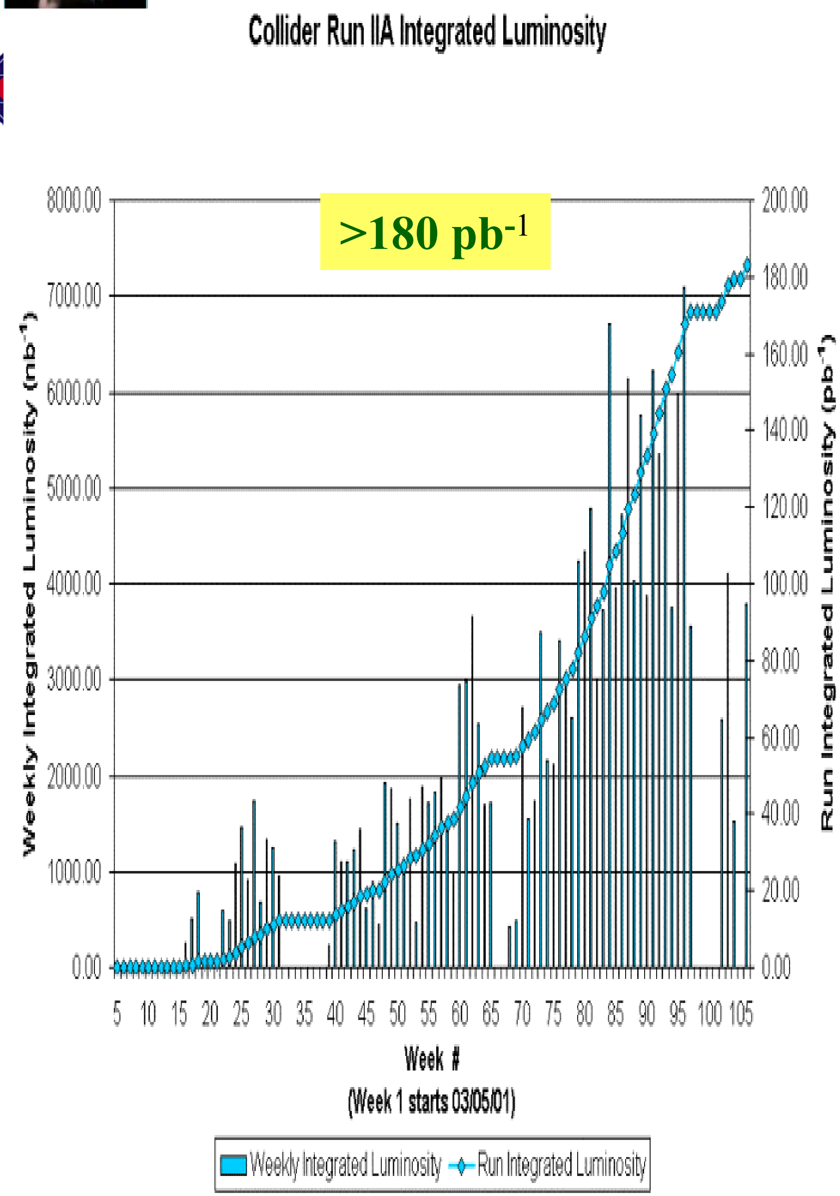
12 countries

59 institutions

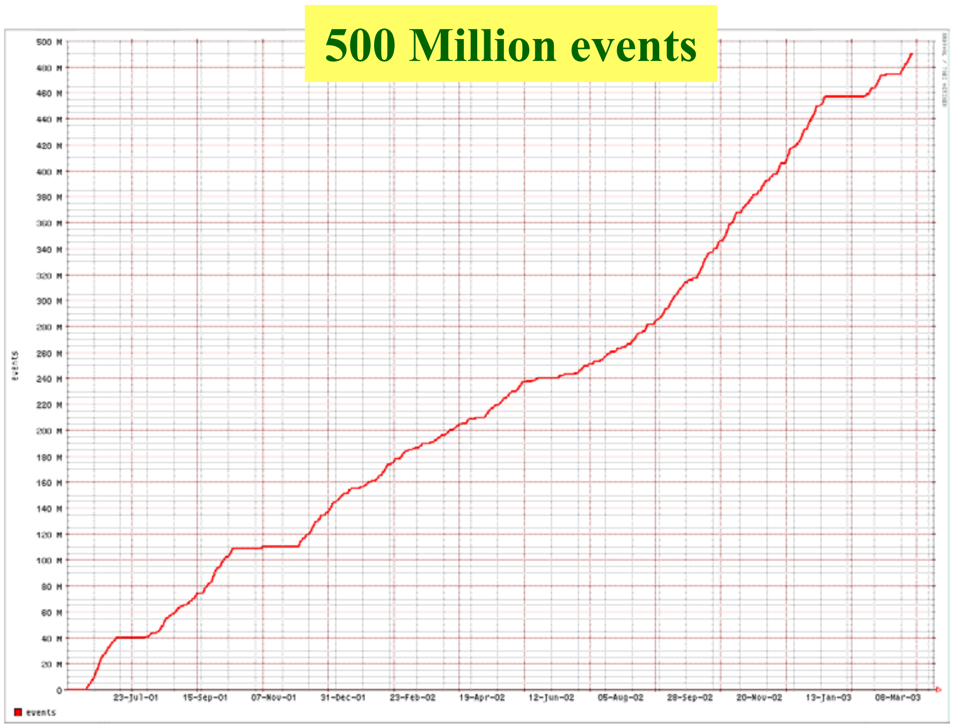
706 physicists

Stephen Wolbers,


Collider Run IIA Peak Luminosity
5.00E+31
4.00E+31
3.00E+31
2.00E+31
1.00E+31
0.00E+00
Peak Luminosity
Peak Lum 20x Average
04/01/01
05/01/01
06/01/01
07/01/01
08/01/01
09/01/01
10/01/01
11/01/01
12/01/01
01/01/02
02/01/02
03/01/02
04/01/02
05/01/02
06/01/02
07/01/02
08/01/02
09/01/02
10/01/02
11/01/02
12/01/02
01/01/03
02/01/03
03/01/03
Date
Peak Luminosity
Peak Lum 20X Average



CDF Raw Events



Run 2 Computing History

- Early work/planning (goes back to 1995-1996):
  - Computing projects in both collaborations.
  - Computing Division efforts.
  - Joint projects Computing Division/CDF/D0.
- Planning and Reviews:
  - DMNAG report (1997)
  - Software Needs Assessment (1997)
  - Von Ruden external reviews:
    - 1997(2), 1998, 1999(2)

Short History (2)

- The Run 2 joint computing project has so far:
  - Designed hardware and software systems for Run 2 data storage, processing and analysis.
  - Defined and ran joint CD/D0/CDF projects.
  - Procured, installed, integrated and operated Run 2 offline computing systems.
  - Spent approximately $18M on equipment over 5 years (2-6-2-4-4).
- Framework, code management, code distribution
- Reconstruction packages
- Online Systems
- Analysis Systems

What was successful?

- Joint Projects
- Data volume was more or less correct.
- Reconstruction farms.*
- Linux.
- Open source.
- ROOT.
- PCs.
- C++/C++ gurus. Liz Sexton
- Networks & TCP/IP.*
- Mass Storage/Enstore.* Don Petravick
- SAM data handling system. Lee Lueking
- Offsite Monte Carlo production (D0).
- Commodity computing.
- Analysis Model.
- Code management/distribution. Art Kreymer
- Reviews.*
- Moore's Law.*

* More Details will be given

Farms

- Farms (PC clusters) for event reconstruction and Monte Carlo are extremely successful.
- CPU power is plentiful and cheap.
- Networking is adequate, local resources are sufficient (memory and disk).
- The big issues are power and cooling and space!
- Maintaining these big systems is a non-trivial effort:
  - Hardware/OS/Networks/NIS/NFS.
  - Reconstruction Code.
  - Database and mass storage connections and tuning.
  - Operations – Software and coordination with collaboration.
    - It is easy to mess this up!

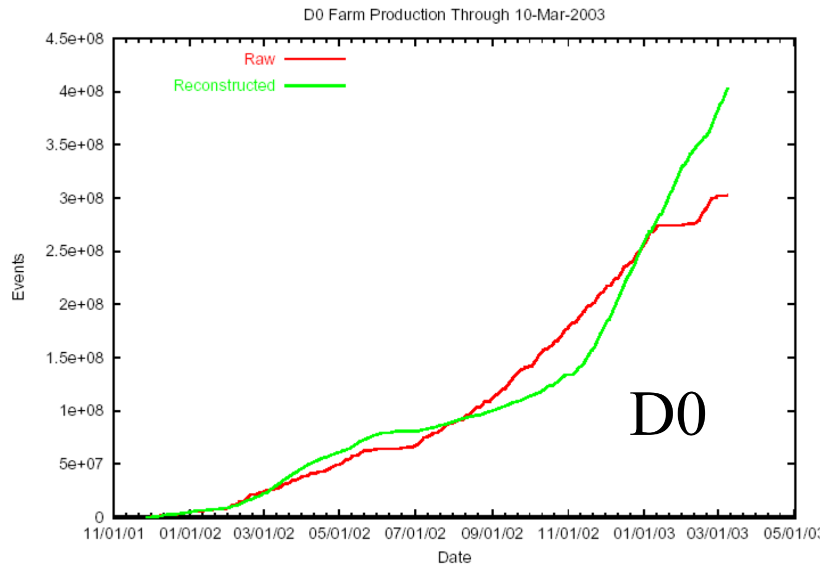
Farms – Run 1 vs Run 2

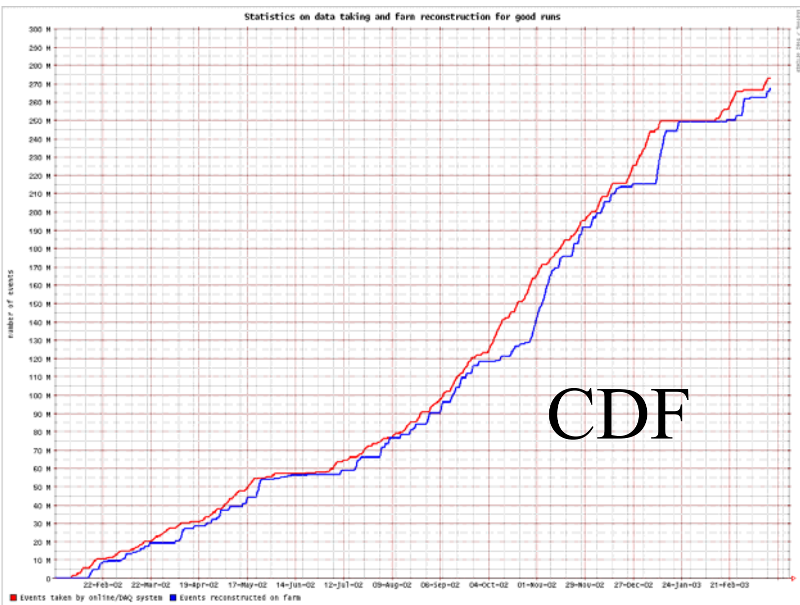
Run 1	Run 2	
• 30 MHz UNIX WS	• 1.7 GHz PC	(X60)
• 16-32 Mbyte memory	• 1-2 Gbyte memory	(X60)
• Shared 10 Mbit	• Switched 100 Mbit	(X100)
• 10 Mbyte executable	• >200 Mbyte executable	(X20)
• 5-20 seconds/event	• 2-10 seconds/event	(X30)
• 5-7.5 Hz	• 50-75 Hz	(X10)
• 100 CPUs	• 800 CPUs	(X8)

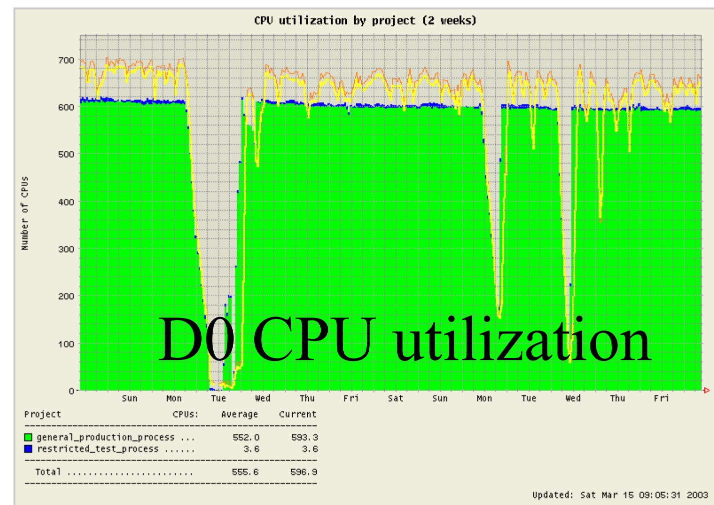
A tremendous increase in capability

CDF and D0 Reconstruction

Bottom Line: Keeping up with Raw data and Reprocessing

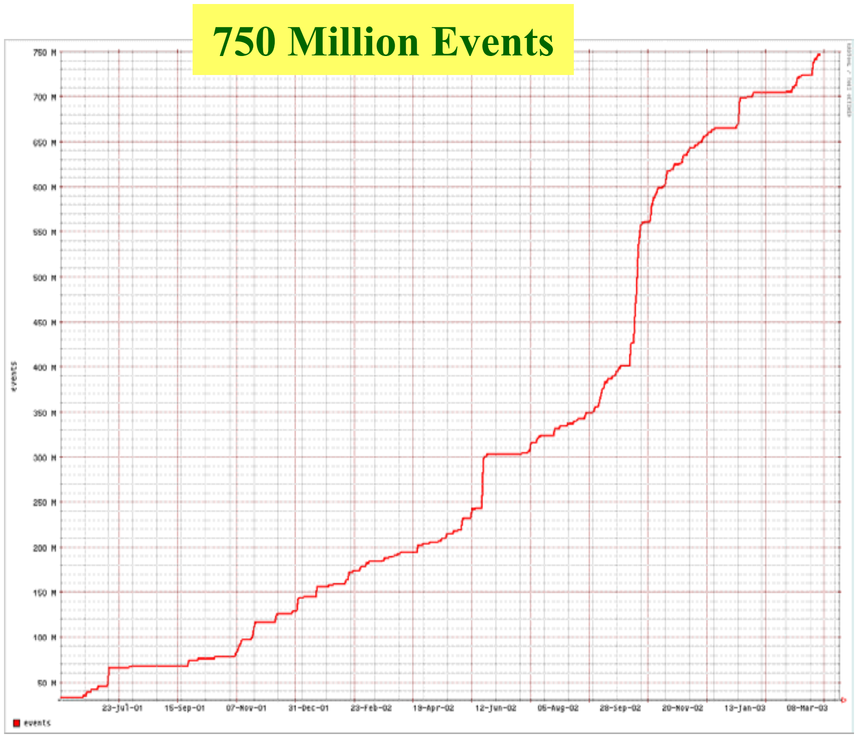






Versions of code/CDF

I (Commissioning)	3.11.0g	9,775,297
I (Commissioning)	3.12.0	9,866,564
II (1x8)	3.14.0	2,539,217
III (36x36)	3.15.0c	5,113,927
III (36x36)	3.16.0	5,113,927
IVa (June, 2001)	3.17.1	12,322,465
IVb (June, 2001)	3.18.0	21,993,586
V (August, 2001)	4.0.0i	9,882,486
V (August, 2001)	4.1.0	38,857,497
V (August, 2001)	4.2.0a	8,685,950
VI (December, 2001)	4.2.0b	6,445,760
VI (December, 2001)	4.2.0c	26,944,076
VI (December, 2001)	4.3.1	14,519,033
VII (2002 data)	4.3.1b	10,369,500
VII (2002 data)	4.3.2	10,109,874
VII (2002 data)	4.3.2a	38,690,220
VII (2002 data)	4.5.2	125,161,272
VII (2002 data)	4.8.0	16,572,810
VII (2002 data)(through Jan,2003 shutdown)	4.8.4a	257,886,088
TOTAL		630,849,549

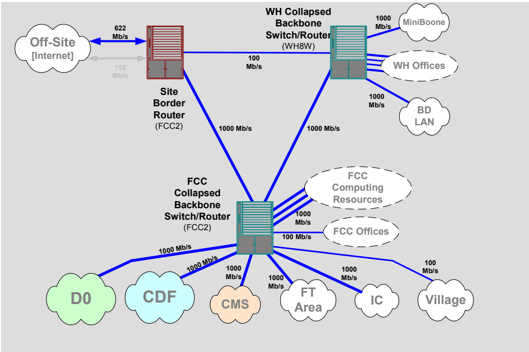


Includes Reprocessing


CDF Farms

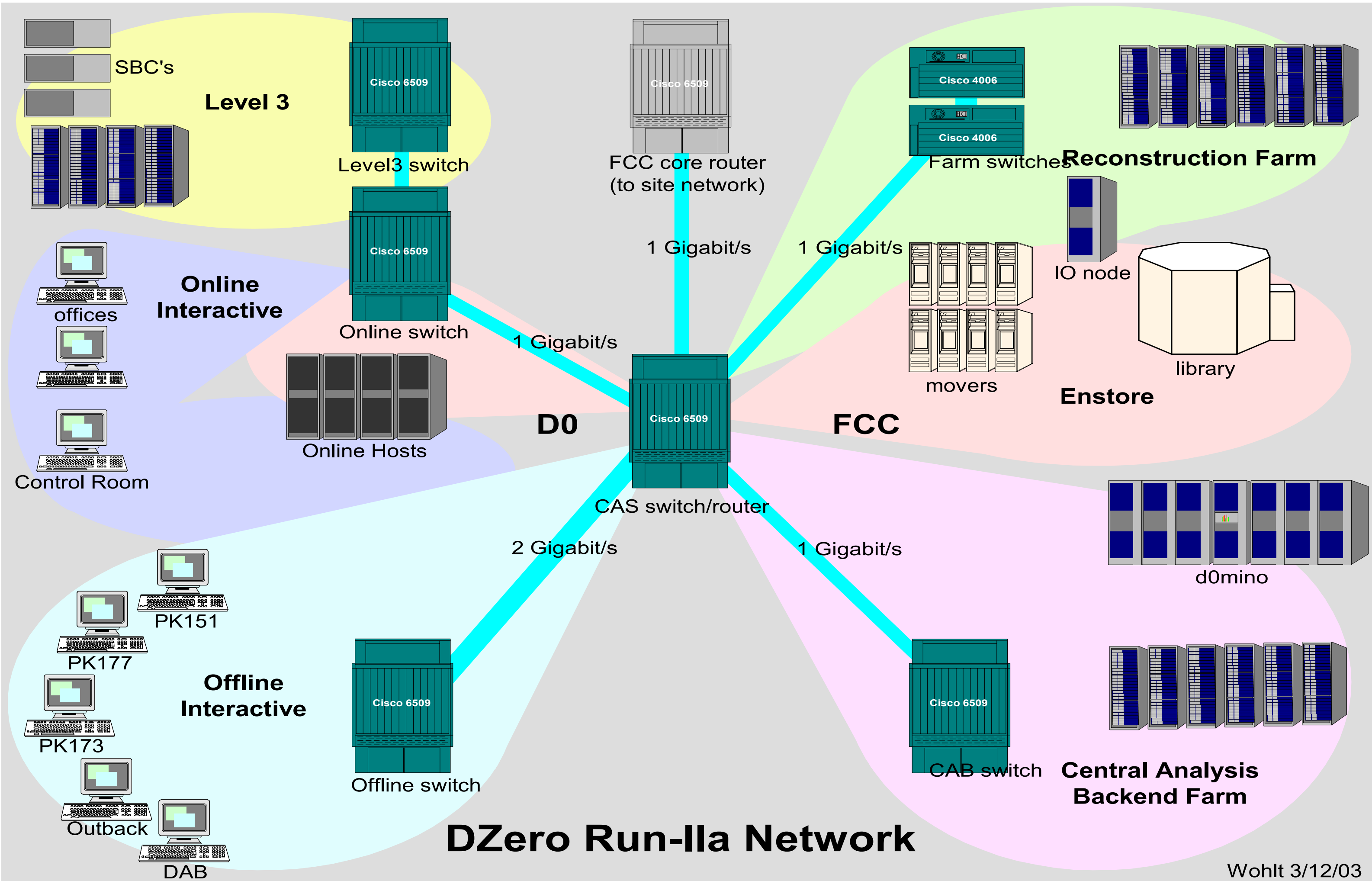

D0 Farms

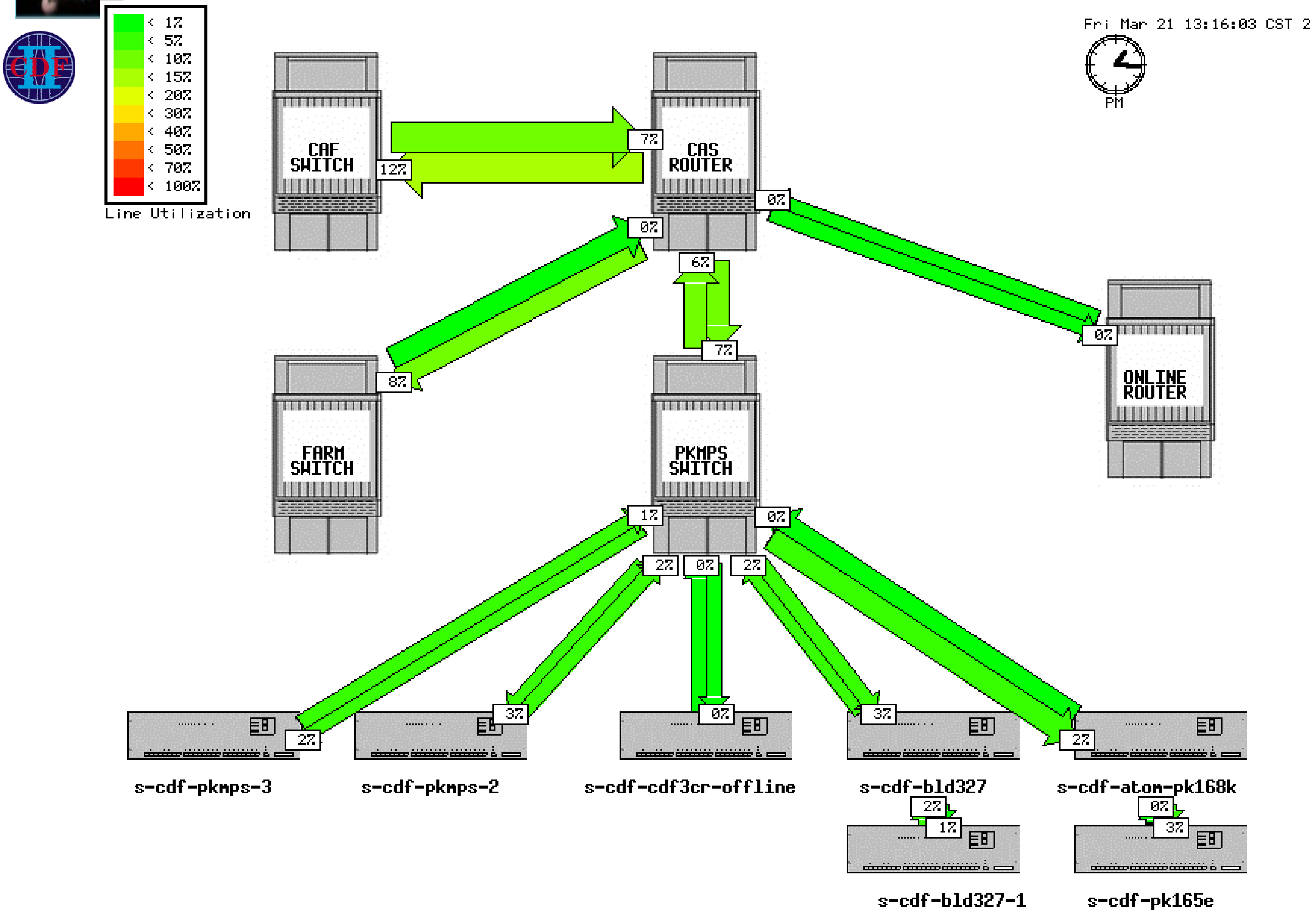
Networks



- LAN
  - Very large and growing LAN for both experiments:
    - Connection of central systems.
    - Desktops.
    - Data movement drives the requirements ever higher.
  - A tremendous success.
  - Issues of scaling need to be solved.
    - Especially switch-to-switch.

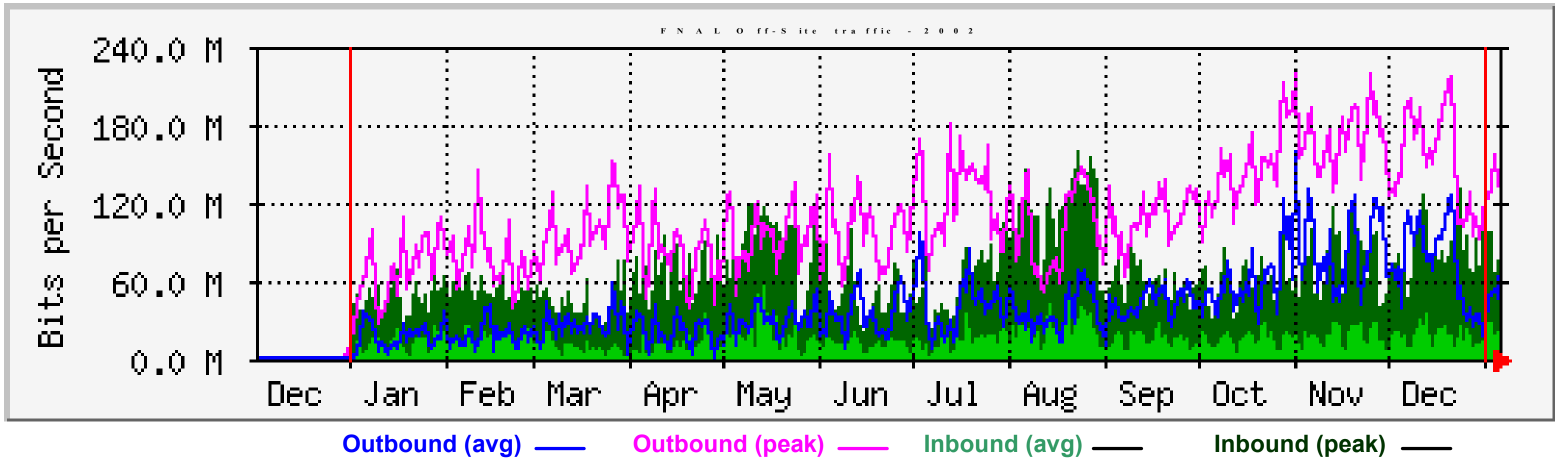
D0 Run-II Network Topology




Computing Division
Line Utilization
< 1%
< 5%
< 10%
< 15%
< 20%
< 30%
< 40%
< 50%
< 70%
< 100%
Fri Mar 21 13:16:03 CST 2
PM
CAF SWITCH
CAS ROUTER
FARM SWITCH
PKMPS SWITCH
ONLINE ROUTER
7%
12%
0%
0%
6%
7%
8%
0%
1%
0%
2%
0%
2%
2%
3%
0%
3%
2%
2%
1%
0%
3%
s-cdf-pkmps-3
s-cdf-pkmps-2
s-cdf-cdf3cr-offline
s-cdf-bld327
s-cdf-atom-pk168k
s-cdf-bld327-1
s-cdf-pk165e

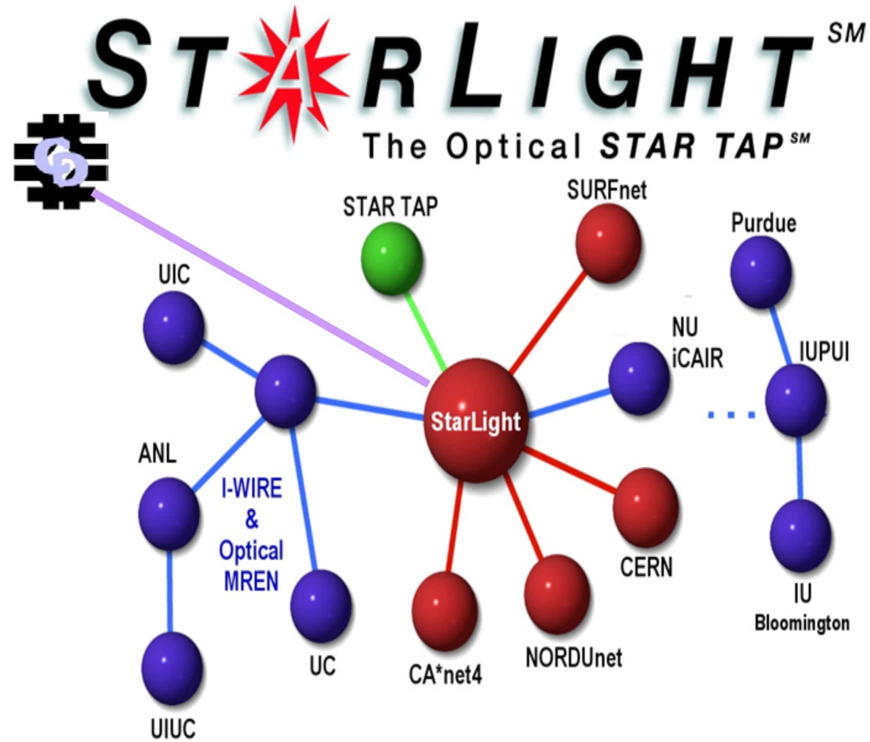
Wide Area Networking

- This is an area which is becoming ever more important (widely distributed computing, grid, data exchange)
- ESNet upgrade to 622 Mbps December, 2002.
- Long-term increase in data rates have been seen.



Upgrades to WAN

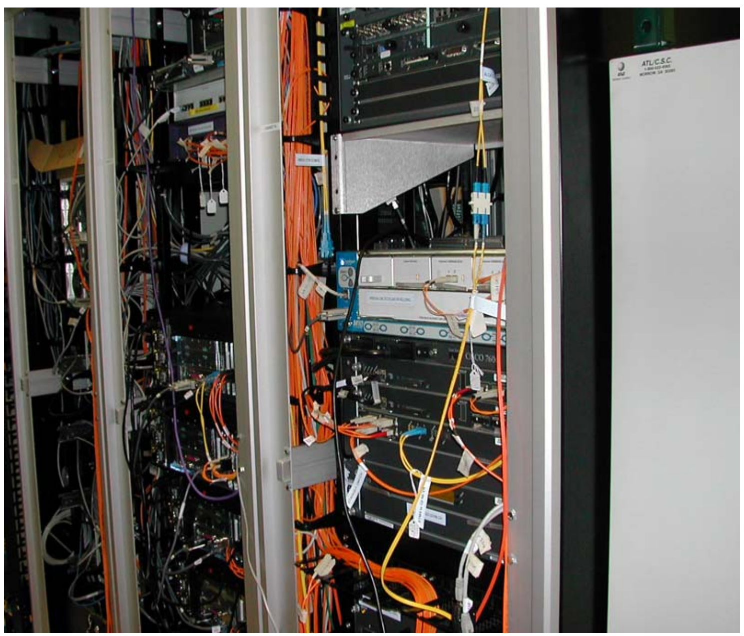
- Increased capacity for offsite connections is very important for Run 2 (and CMS and others).
- One possibility: Connect to Starlight (dark fiber leased from ComEd).
- Investigating many ways to increase bandwidth.

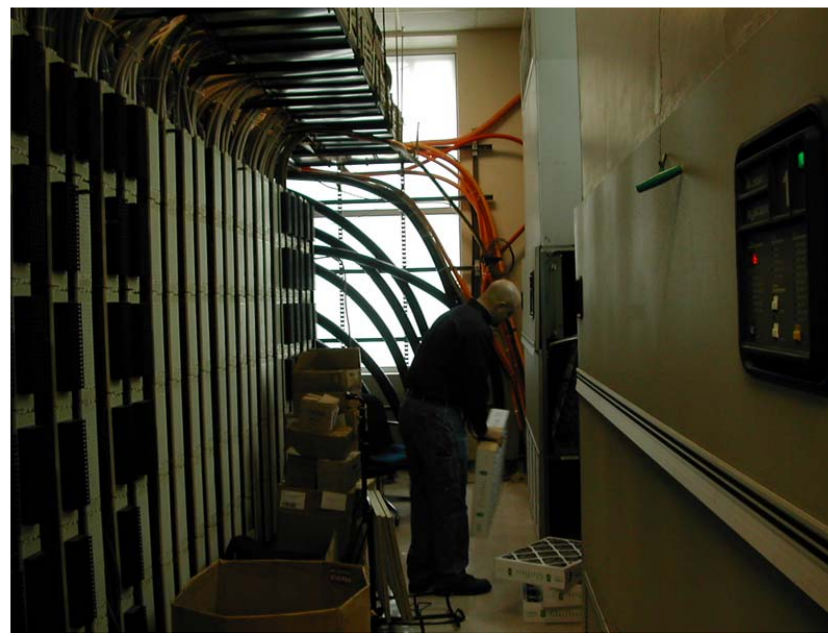


Starlight: optical networking interconnection point downtown Chicago (*710 Lake Shore drive*)

Owned by Northwestern University

STARLIGHT





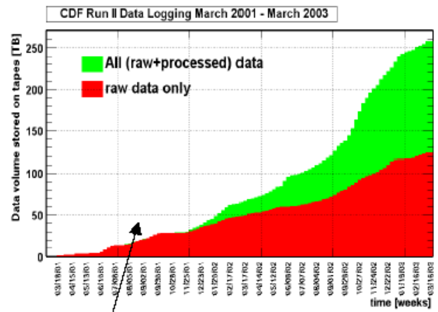
Mass Storage

Mass Storage

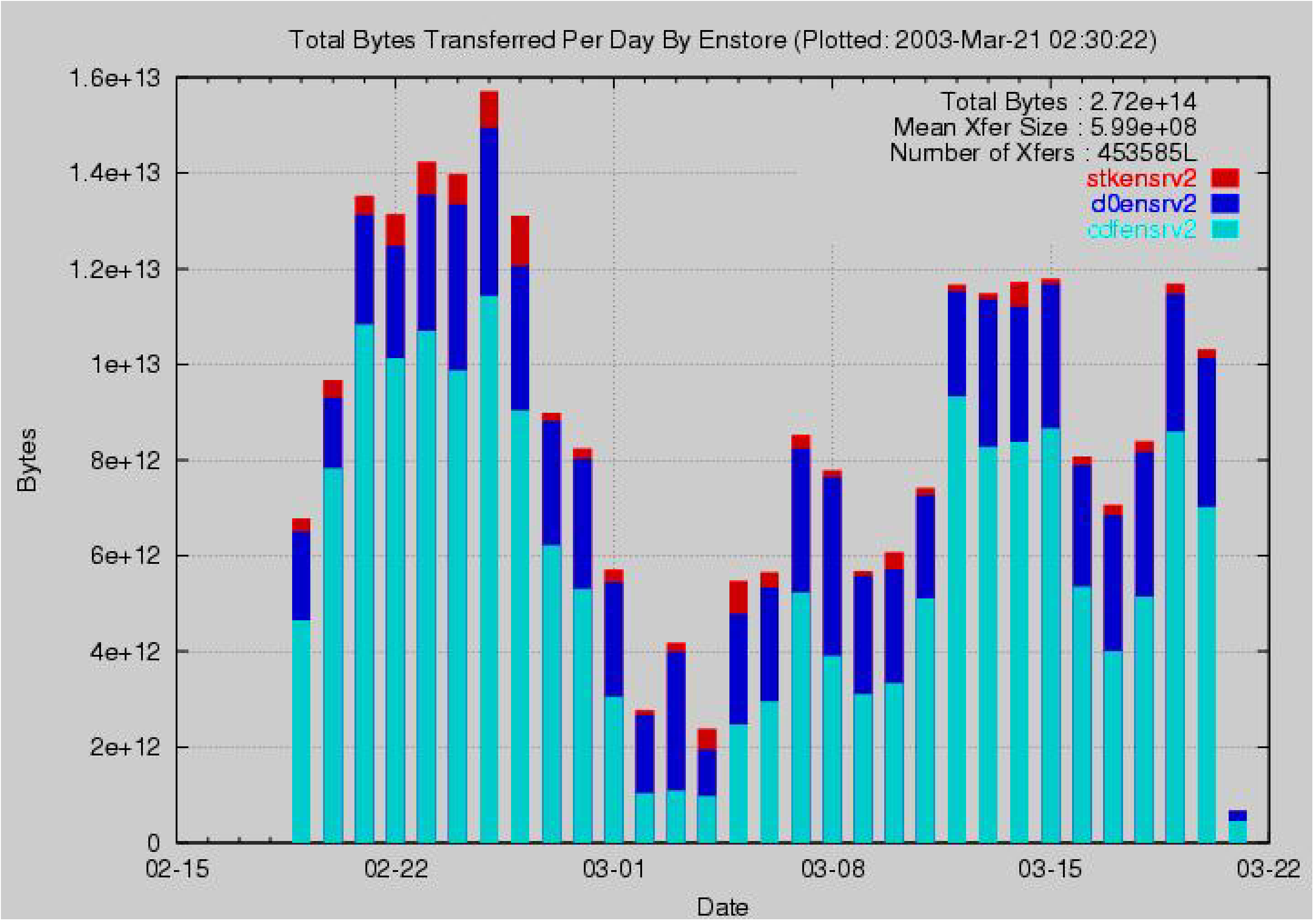
- Mix of STK 9940a (60 Gbyte cartridges) and LTO 1 (100 Gbyte cartridges).
  - Allows competition – no single vendor.
- Currently migrating to STK 9940b (200 Gbyte cartridges) and investigating LTO 2.
- Software layer – Enstore.

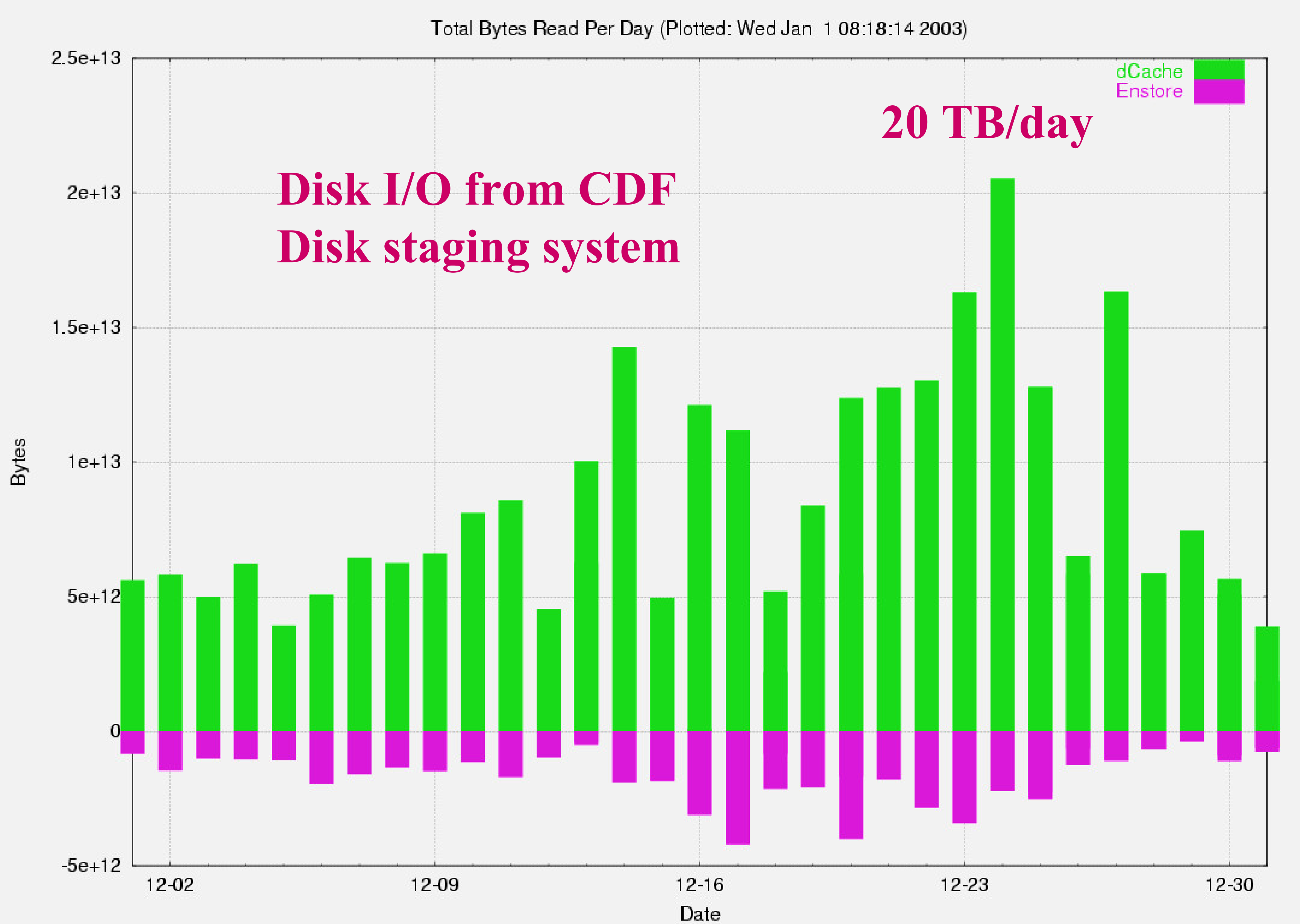
Don Petravick

- Total Run 2 data on tape : 554 Tbyte.
  - Grows by ~2 TB/day.
  - Sometimes shrinks.
    - Older farms output.



Total Tape I/O per day



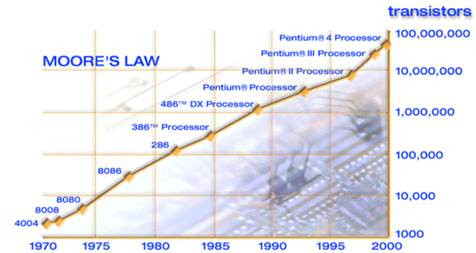

Total Bytes Read Per Day (Plotted: Wed Jan 1 08:18:14 2003)
dCache
Enstore
20 TB/day
Disk I/O from CDF
Disk staging system
Bytes
2.5e+13
2e+13
1.5e+13
1e+13
5e+12
0
-5e+12
12-02
12-09
12-16
12-23
12-30
Date

Reviews

- Reviews are good! (up to a limit)
- They allow for:
  - Organization and coordination of plans/thinking.
  - Collection of important planning information.
  - Outside/new look at computing issues.
    - BaBaR, JLAB, CERN, DESY, other
  - Access to new resources.
  - Attention of the Directorate/other higher level people.

Moore's Law

- Moore's Law is essential to modern HEP computing capabilities.
  - We heavily rely on it.



- It is not a substitute for:
  - More efficient, faster code.
    - 50% more CPU not the same as faster code, same amount of CPU
  - Smaller datasizes.
  - Thinking before doing.
- The increased computing drives the science and vice-versa.
- Everyone would benefit from optimizing wherever possible.
- (Except those that give CHEP talks and need to show pictures of large amounts of computing stuff.)

What did not work (or needed change)?

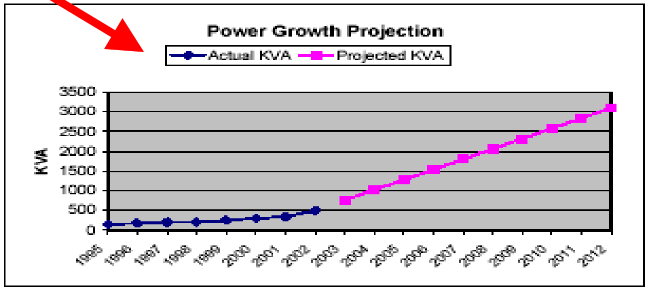
- Mass storage.*
- SMP and Analysis computing.*
- Scaling systems too quickly.
- Commodity PC and Disk.*
  - Hard to get and keep something that really works.
  - Linux kernels, large disk systems, RAID systems.
  - Disks.
    - Same issues as with PCs or most commodity equipment.
- Windows NT/2000/Commercial Software.
  - KAI (soon to be gone), analysis tools.
- Fibrechannel/SAN.

* More Details

What did not work (2)?

*More Details

- Data handling (CDF).*
- Power and cooling of computer centers.
- Database performance.
- Reconstruction code speed.
- Lack of adequate monitoring.
- Procurement latency.
- wbs project management.



Mass Storage

- 8mm tape and flexible robots were chosen as Run 2 mass storage systems.
  - Many reasons for this, all correct at the time.
- Significant problems with performance, robustness and capabilities of 8mm tape and its integration into ADIC robots.
- Solution: STK/9940 and ADIC/LTO, coupled to the Enstore mass storage software (and Dcache).
  - Rapid deployment, building modifications.
  - Successful major modification in mid-project.
- Technology and competition gave us a solution.

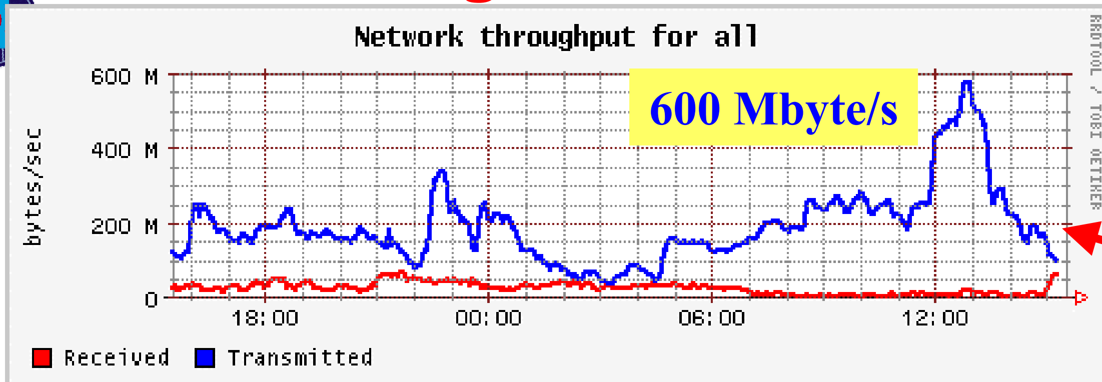
Analysis Computing/SMP/PC

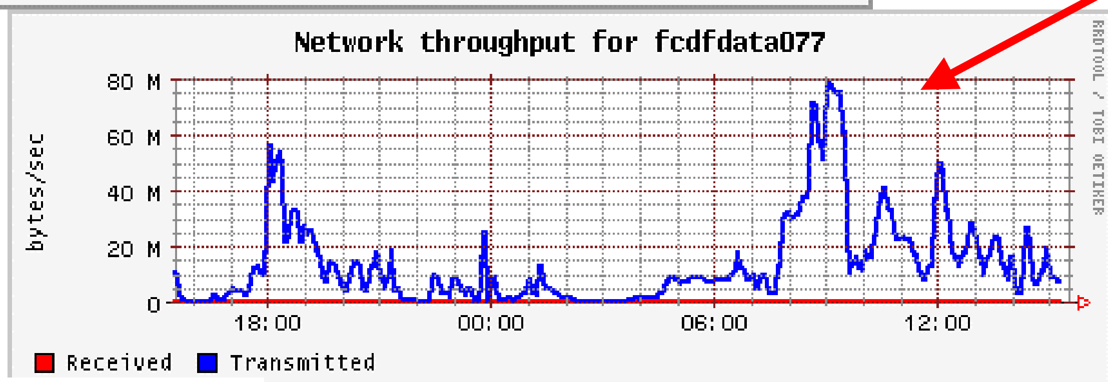
- Analysis Computing in Run 2 is rapidly moving from big SMPs to PC based systems.
  - Driven by cost, performance, capabilities.
  - SMPs remain as fileservers and common shared environments.

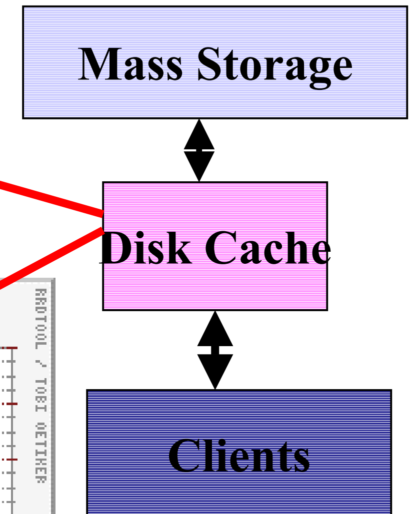
Analysis Computing

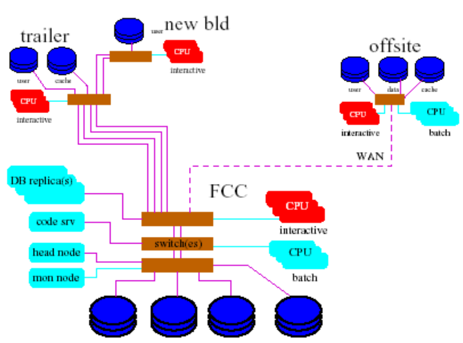
- Tough problem: Many clients, unpredictable behavior, short timescales (conferences), large peaks of load.
- Last problem to be addressed – DA, online, production, Monte Carlo all dealt with first.
- In addition, analysis computing normally competes to some extent with production activities:
  - Tapedrives
  - Network
  - Staging systems
- Analysis is distributed worldwide – should be integrated in the design.
- In the end the analysis computing is where the physics results happen – Very Important!

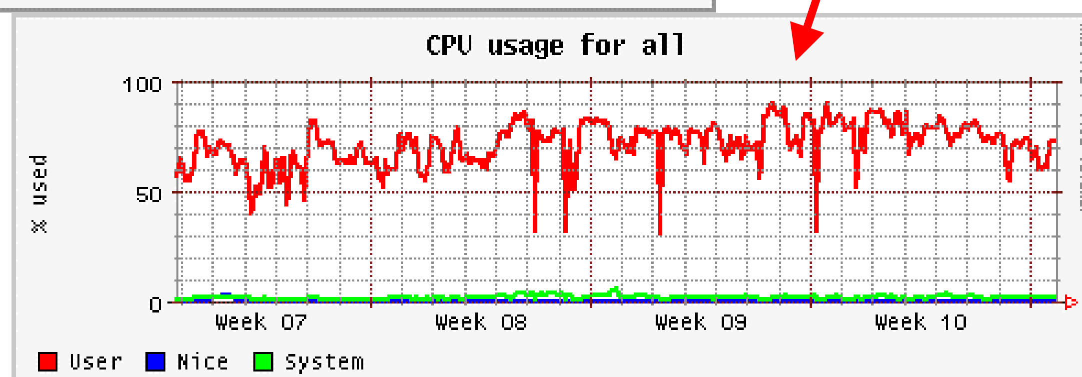
Moving a lot of data on CDF CAF











Frank Wuerthwein

Commodity Computing

- Commodity computing is a challenge.
  - Much variety, constant change.
    - Even within individual purchases.
- Attempt to solve:
  - Common evaluation of equipment.
  - Common specification of configurations.
- Likely to remain an issue.
- This can be a huge effort!

Data handling/CDF

- CDF data handling model became unsupportable.
  - Direct attached tapedrives.
  - SMPs.
  - Fibrechannel/SAN.
  - Staging software.
- Solution: move to common CD/CDF/D0 tools.
  - Enstore/STK/9940.
  - Dcache.
  - PC analysis systems (CAF).
  - SAM/grid.
- Effort to change was not small.
  - Tape copying.
  - Software modifications.
  - System tuning.

Gabriele Garzoglio, Fedor Ratnikov
Rob Kennedy, Dmitry Litvintsev

What was not expected?

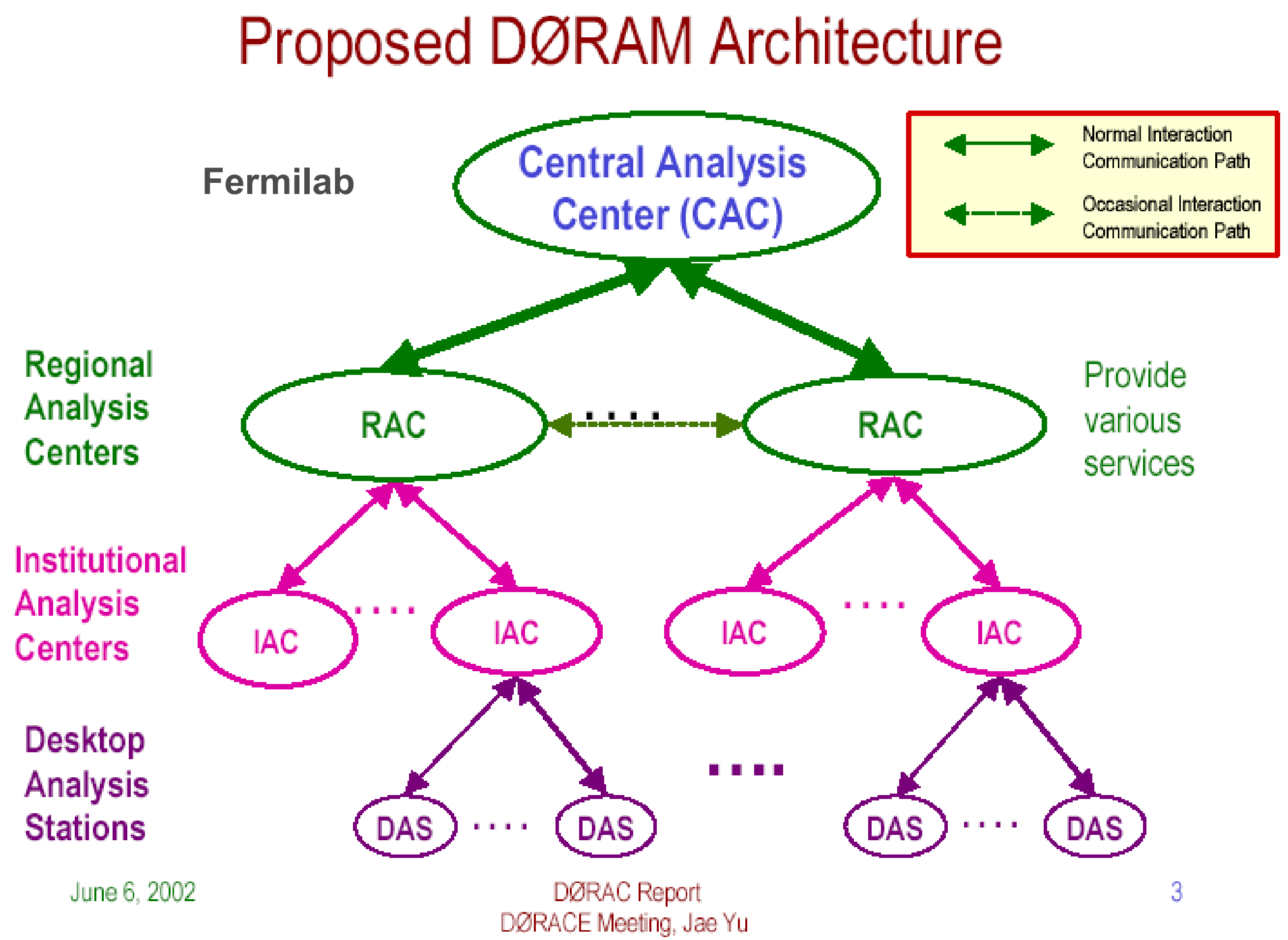
- Distributed computing, the grid.
- Really cheap PCs and disks and Linux.
- Long ramp up of the collider.
- Commissioning of the detectors and triggers.
- Duration of Run 2.
- Timescales for finishing software projects.

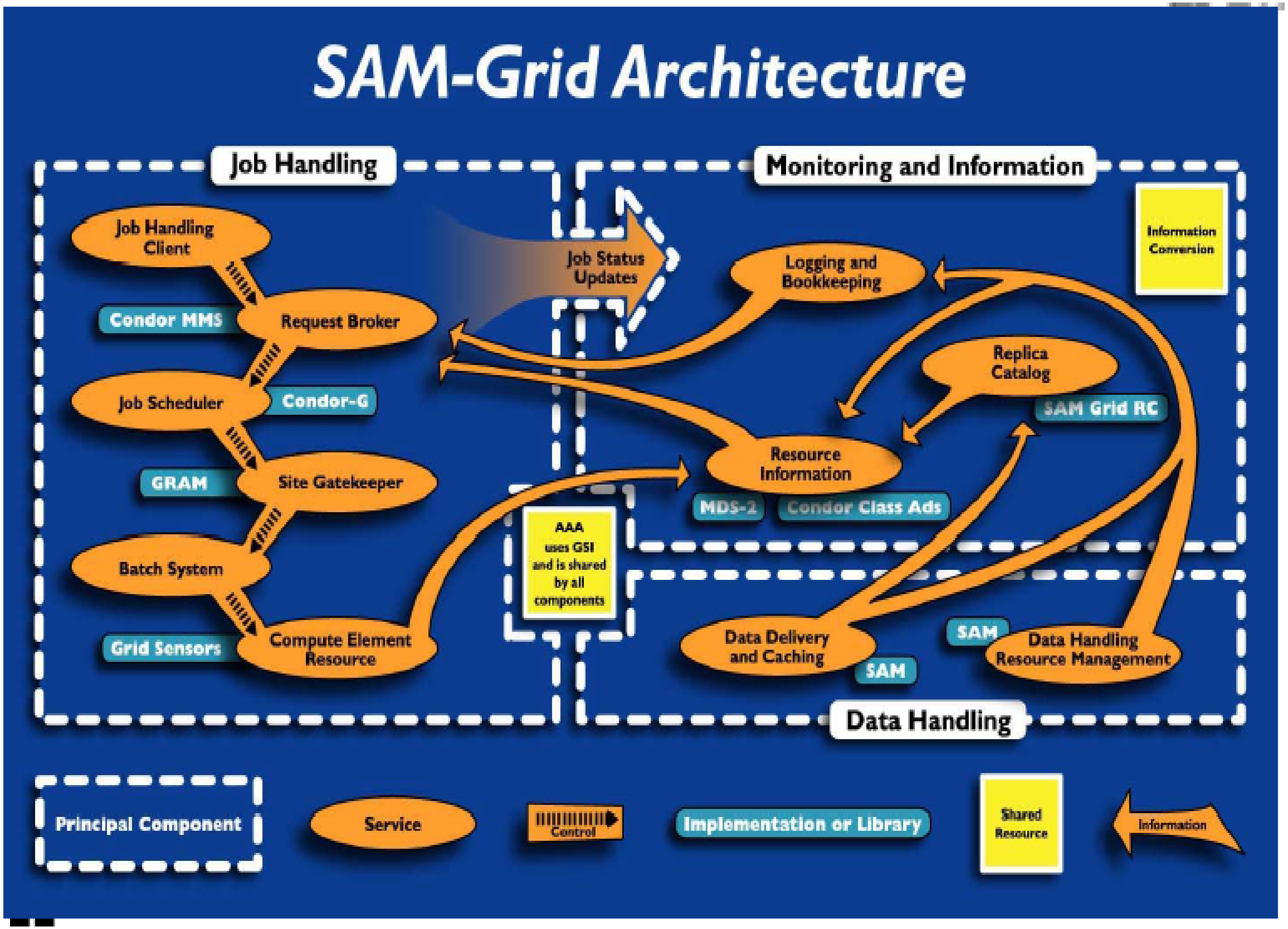
Distributed Computing and the grid

- CDF and D0 are both highly distributed collaborations, with many physicists and computing resources.
- Making use of that potential has been an issue from the start for D0 (Monte Carlo generation) and for CDF more recently.
- These efforts are going to grow.
- The SAM data handling system, used from the beginning by D0 and recently by CDF, is being used/modified for grid/distributed computing.

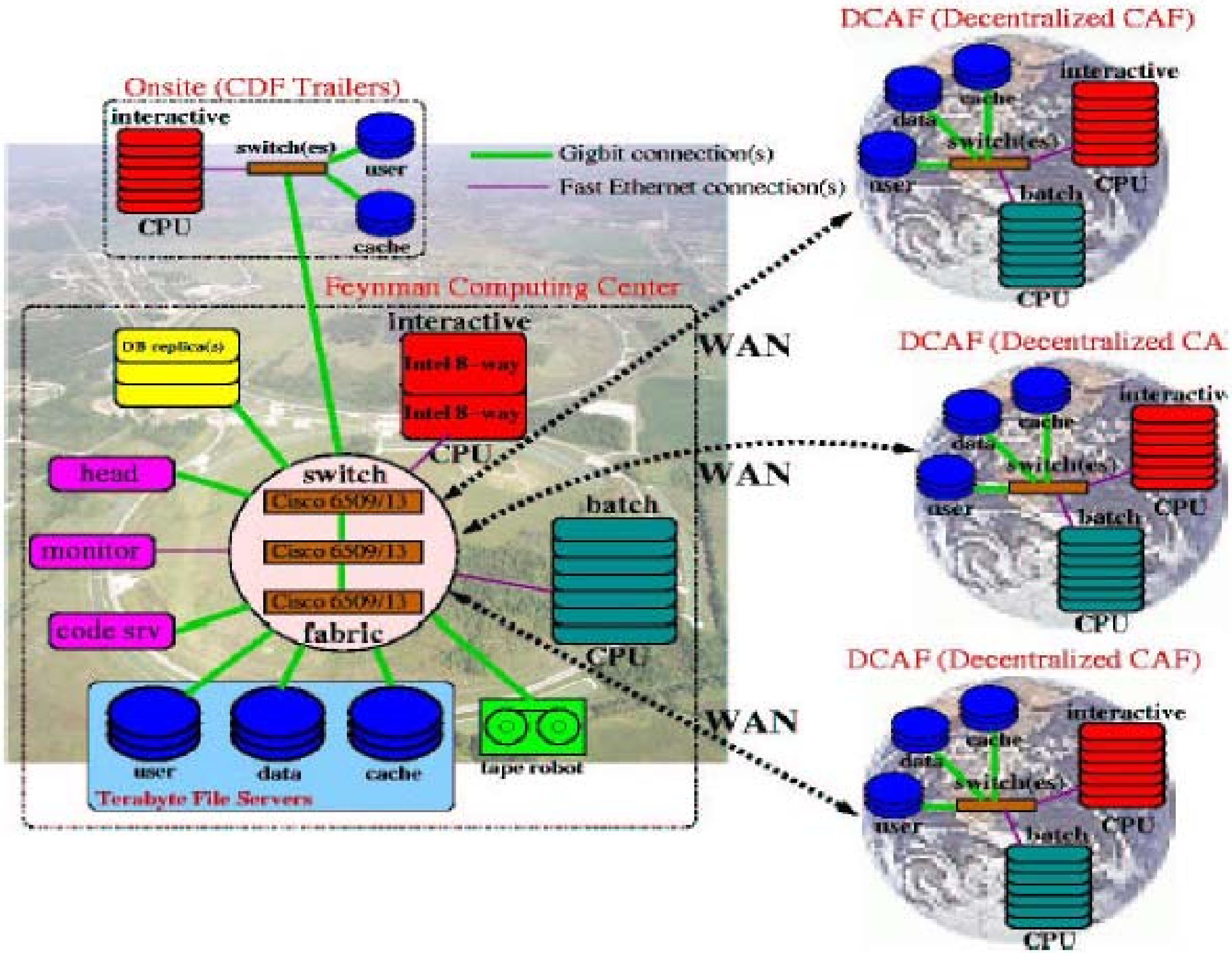
Igor Terekov, Stefan Stonjek, Fedor Ratnikov, Lee Lueking

D0 regional analysis centers




SAM-Grid Architecture
Job Handling
Monitoring and Information
Job Handling Client
Condor MMS
Request Broker
Job Scheduler
Condor-G
GRAM
Site Gatekeeper
Batch System
Grid Sensors
Compute Element Resource
Job Status Updates
Logging and Bookkeeping
Information Conversion
Replica Catalog
SAM Grid RC
Resource Information
MDS-2
Condor Class Ads
AAA uses GSI and is shared by all components
Data Delivery and Caching
SAM
SAM
Data Handling Resource Management
Data Handling
Principal Component
Service
Control
Implementation or Library
Shared Resource
Information

Future CDF Direction



Lessons

- Plan for change, don't be afraid to make major modifications.
- Make realistic requirements estimates, taking into account experience as well as wishes.
- Optimize as much as possible.
  - Hardware can make up for excesses.
  - But it is easier to run less stuff and/or get things done more quickly and easily and with less storage, memory, etc.
- Don't buy too early.
- Test but be prepared for many surprises.
- Scale systems carefully and slowly.
- Be flexible, even late in the process.
  - But in 10-20 year experiments it is not clear when "late" is.

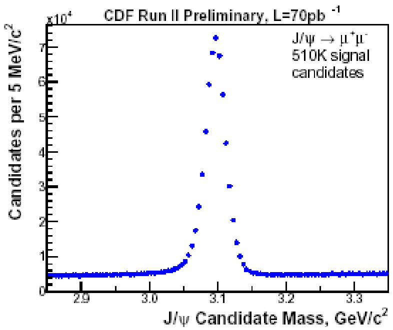
Lessons (2)

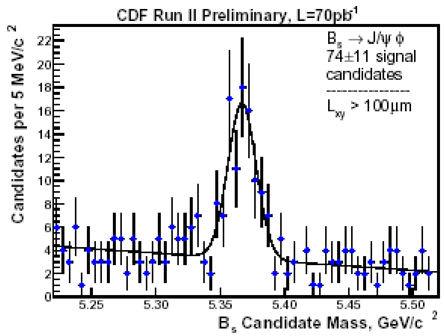
- Since all development (H/W and S/W) cannot be provided at one time, choose a path that gives necessary features and performance at any given time without impacting the overall system development (easy to say!)
  - This may require unoptimized systems at first, maturity coming later.
- Have a core group be responsible for infrastructure.
  - But be sure they are listening to the collaboration and others.
- Make good use of reviews.
  - Focus the project.
  - Get attention of management.
  - Get resources.

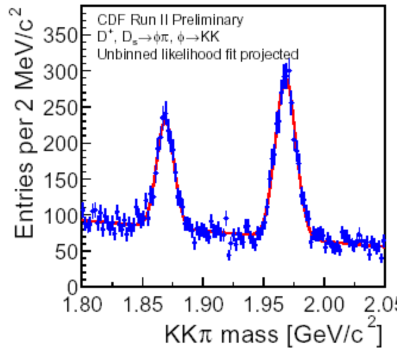
Lessons (3)

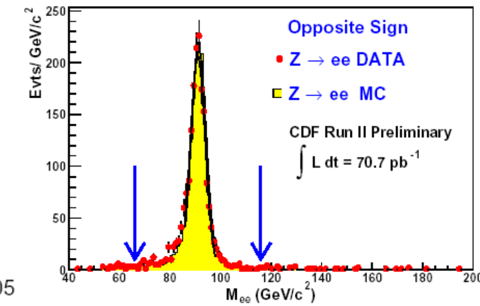
- Joint projects are hard, but worth it.
  - Require coordination with (many) other parties.
  - But the long-term support and features are worth it.
- Two coordinators/leaders are better than one.
  - Complementary strengths.
  - Better coverage.
  - Ideas can be bounced off of each other.
- Databases are very important.
  - They connect everything.
  - Provide some reward for working on them.
  - Use freeware DB?

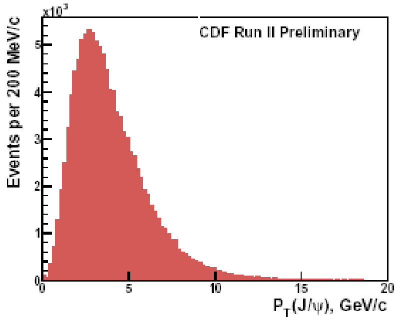
Physics!

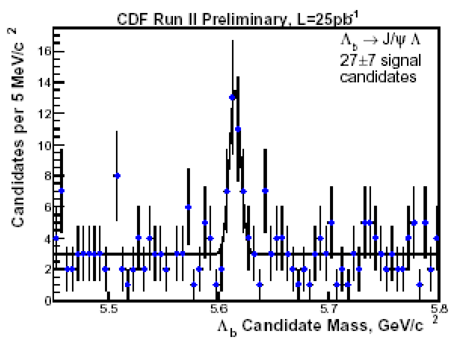

CDF Run II Preliminary, L=70pb⁻¹
J/ψ → μ⁺μ⁻
510K signal candidates
Candidates per 5 MeV/c²
J/ψ Candidate Mass, GeV/c²

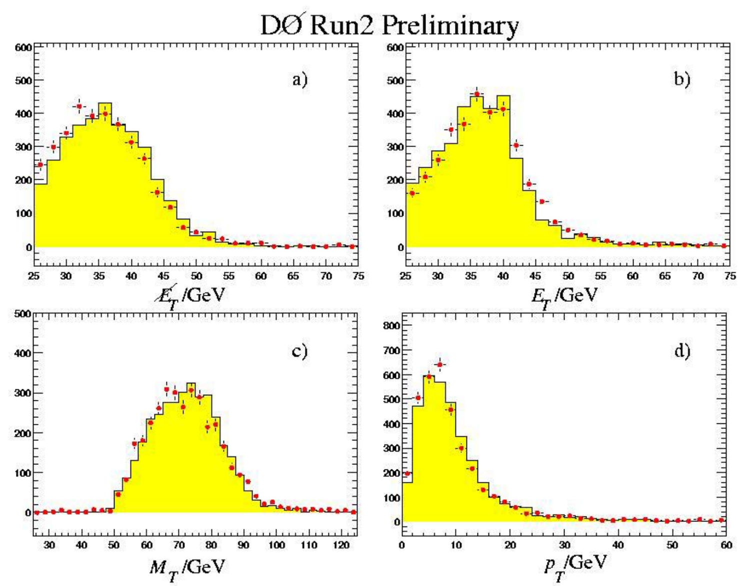

CDF Run II Preliminary, L=70pb⁻¹
Bs → J/ψ φ
74±11 signal candidates
Lxy > 100μm
Candidates per 5 MeV/c²
Bs Candidate Mass, GeV/c²

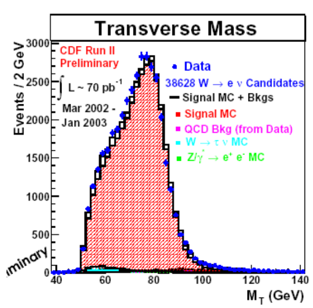

CDF Run II Preliminary
D⁺, Ds→φπ, φ→KK
Unbinned likelihood fit projected
Entries per 2 MeV/c²
KKπ mass [GeV/c²]

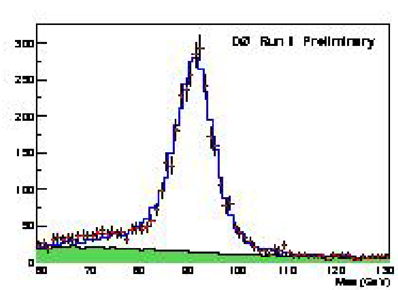

Opposite Sign
Z → ee DATA
Z → ee MC
CDF Run II Preliminary
∫ L dt = 70.7 pb⁻¹
Evts/ GeV/c²
Mee (GeV/c²)


CDF Run II Preliminary
Events per 200 MeV/c
PT(J/ψ), GeV/c


CDF Run II Preliminary, L=25pb⁻¹
Λb → J/ψ Λ
27±7 signal candidates
Candidates per 5 MeV/c²
Λb Candidate Mass, GeV/c²


DØ Run2 Preliminary
a)
b)
c)
d)
E̸T /GeV
ET /GeV
MT /GeV
pT /GeV

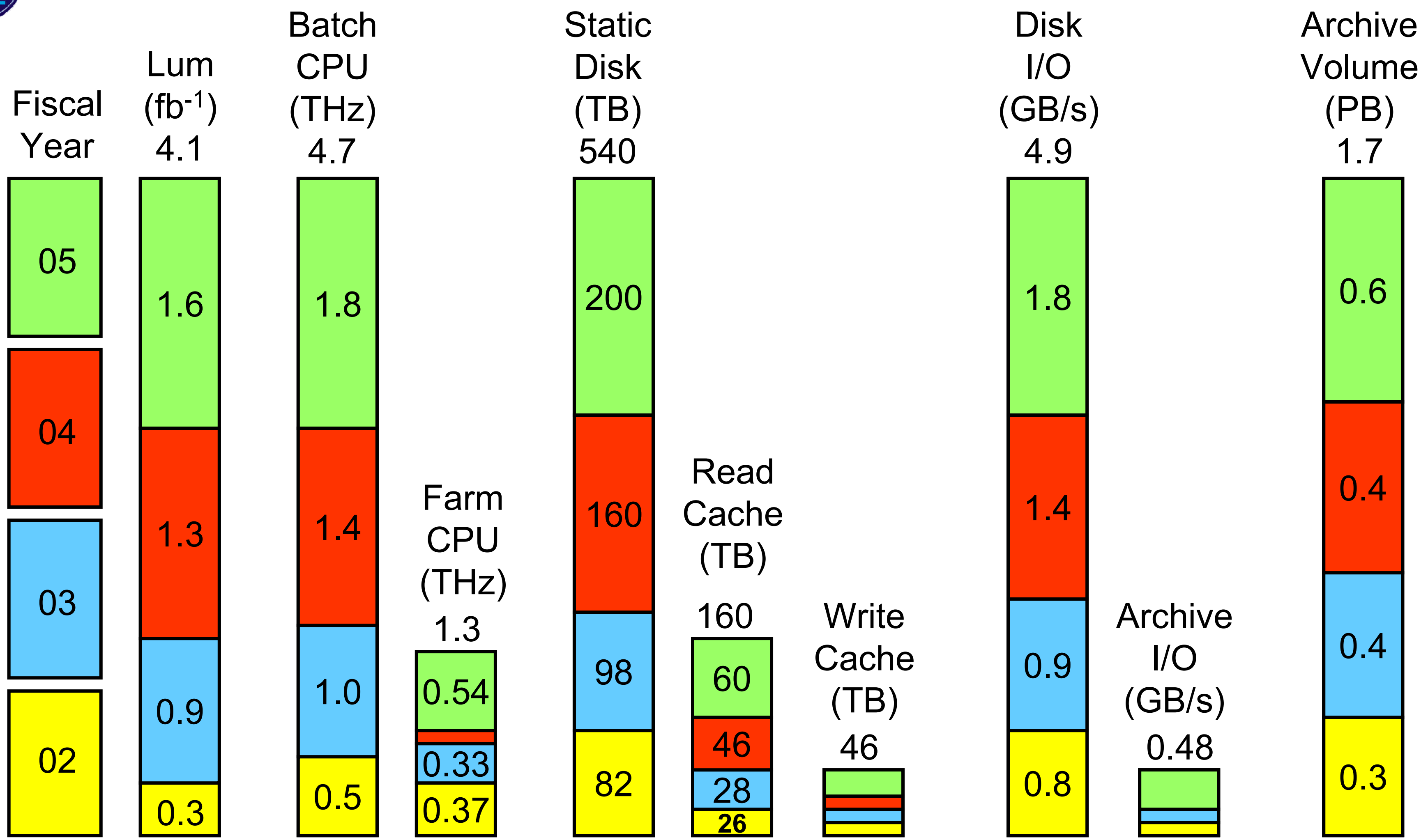

Transverse Mass
CDF Run II Preliminary
∫ L ~ 70 pb⁻¹
Mar 2002 - Jan 2003
Data
38628 W → e ν Candidates
Signal MC + Bkgs
Signal MC
QCD Bkg (from Data)
W → τ ν MC
Z/γ* → e⁺ e⁻ MC
Events / 2 GeV
MT (GeV)


DØ Run II Preliminary

On to the future

- Run 2 is ramping up, has a long life.
  - Code maintenance.
  - Hardware upgrades and increments.
  - Technology changes.
- BaBar, RHIC, JLAB, LHC, others all facing similar issues.
- Grid is happening.
- The temptation (often the necessity) is to solve only our own problems.
  - It makes sense to try to align efforts to provide maximum effort and sharing.
  - It is somewhat difficult to align running experiments with LHC experiments.

Required Computing FY02 – FY05 (CDF)



Summary

- Run 2 Computing is a great success.
- This was a large effort, involving many people over many years.
- The data has been processed and analyzed quickly to produce physics results.
- But the computing systems aren't perfect:
  - Entire systems were replaced.
  - Development had to (and has to) be consistent with data-taking and data-analysis needs.
- There is a huge amount of data on the way.
  - The luminosity continues to increase.
  - The data rate will increase in Run 2b.
  - The physics demands it and the technology allows it.

Thanks

I wish to thank many many people who worked on Run 2 computing over the years. It was truly a large and distributed effort.

I also wish to thank those who gave me suggestions and ideas for the talk – Wyatt Merritt, Amber Boehnlein, Heidi Schellman, Liz Buckley-Geer, Rob Harris, Frank Wuerthwein, Liz Sexton, Don Petravick, Ruth Pordes, Matthias Kasemann + many others who have worked on Run 2 over the past many years.