Fermilab

Computing Division

# *Welcome to Fermilab*

## Large Scale Cluster Computing Workshop

**May 22-25, 2001**

**Matthias Kasemann**

# The Fermilab Scientific Program

Fermilab Computing Division

## Physics of the Weak Energy Scale

- Look for Supersymmetry.
- Make precise measurements of t and W mass.
- Measure $B_s$ mixing, CP parameters.
- Search for Higgs with increasing luminosity.

## Neutrino Masses and Mixing

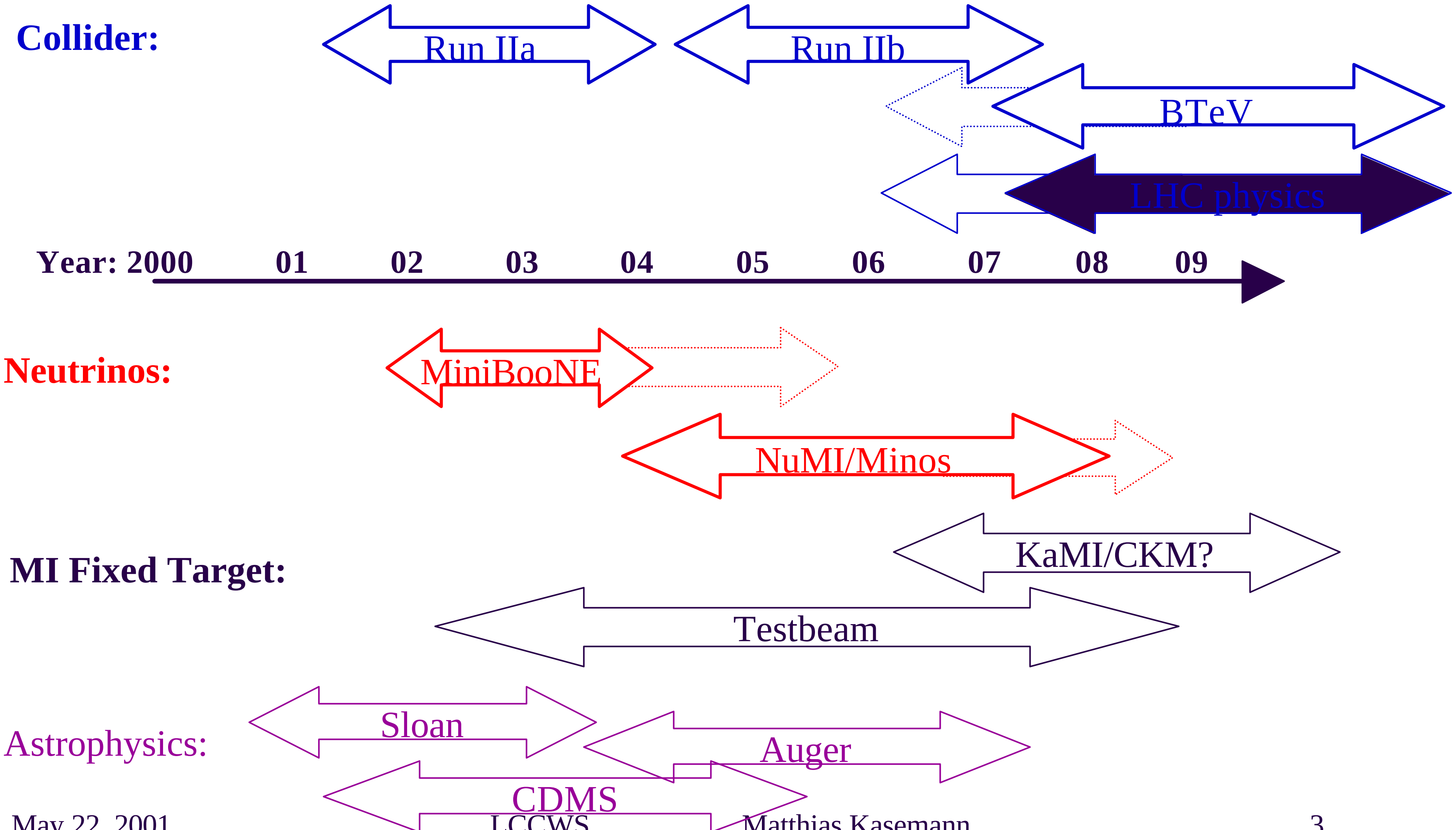
- Study atmospheric neutrino range with MINOS.
- Make a definitive check of the LSND with MiniBoone.

## Particle Astrophysics

- Search for dark matter with CDMS.
- Study highest energy cosmic rays with Auger.
- Discover new astro sources with Sloan Digital Sky Survey.

# *Fermilab HEP Program*

Fermilab
Computing Division



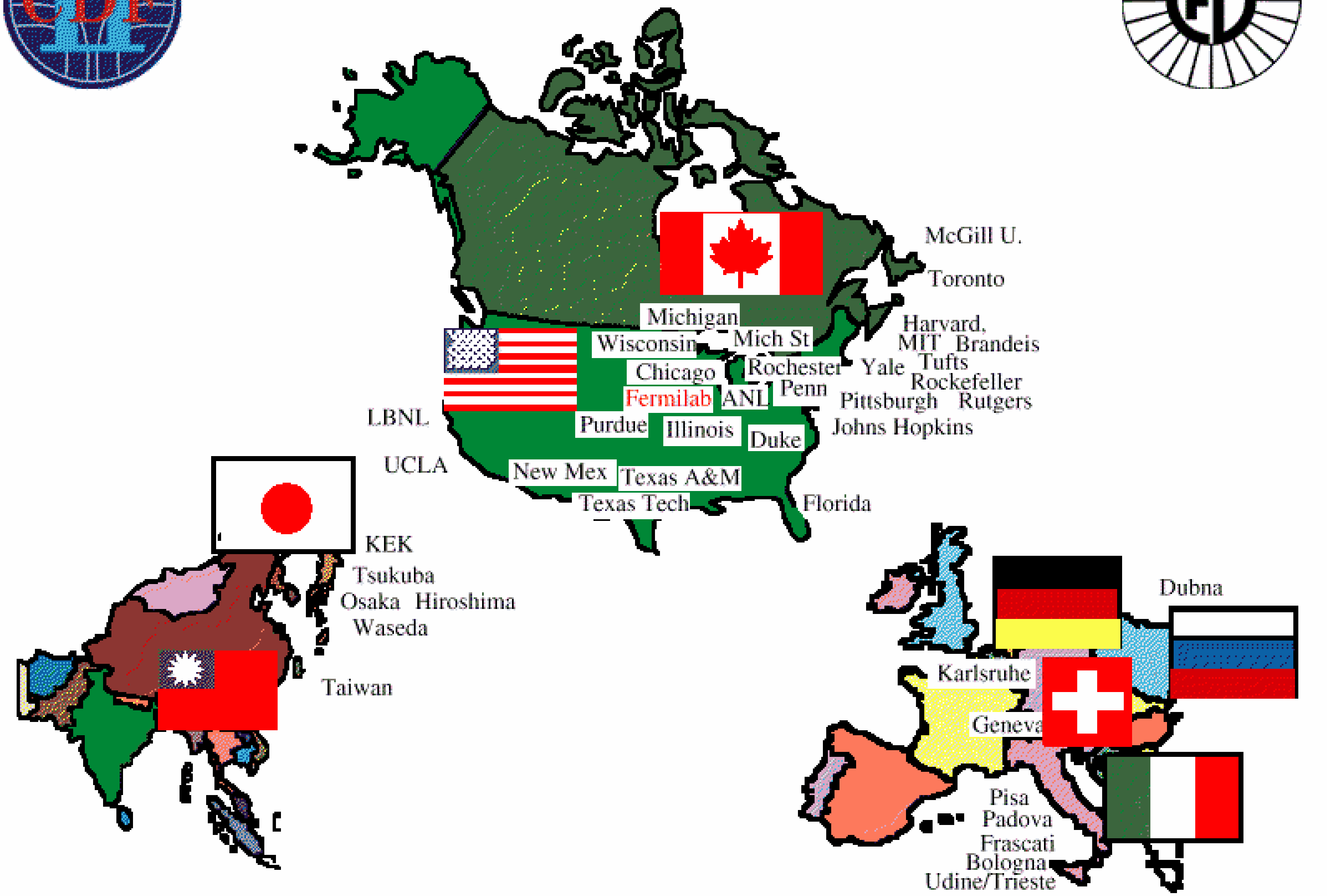
# *World-wide Collaborations at Fermilab*

Fermilab
Computing Division

- 2,716 Physicists work at Fermilab
- 224 institutions from:
  - 38 states (1,703 physicists)
  - 23 foreign countries (1,014 physicists)
- 555 graduate students
- (probably a similar number of postdocs)
- It is interesting to note that only 10% of CDF and D0 physicists work for Fermilab

Status: 1999

# The CDF Collaboration



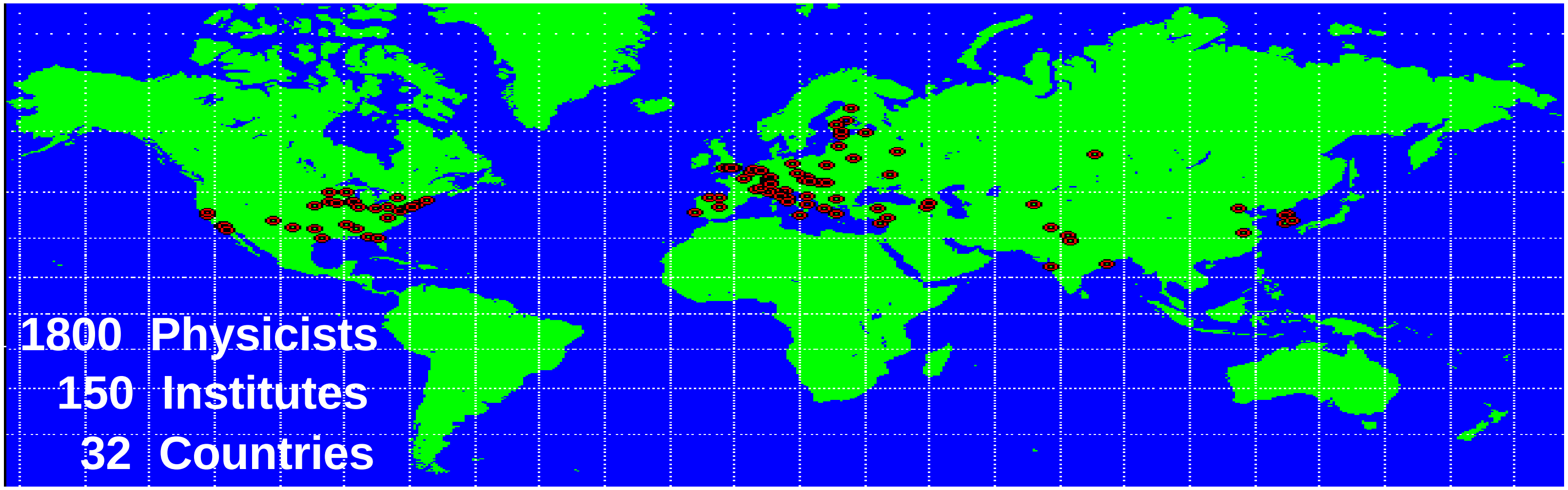

The CDF Collaboration consists of 490 physicists from 41 institutions representing 8 countries

# CMS Computing Challenges

Fermilab

Computing Division

- Experiment in preparation at CERN/Switzerland
- Strong US participation: ~20%
- Startup: by 2005/2006, will run for 15+ years




**Major challenges associated with:**

**Communication and collaboration at a distance**

**Distributed computing resources**

**Remote software development and physics analysis**

**R&D: New Forms of Distributed Systems**

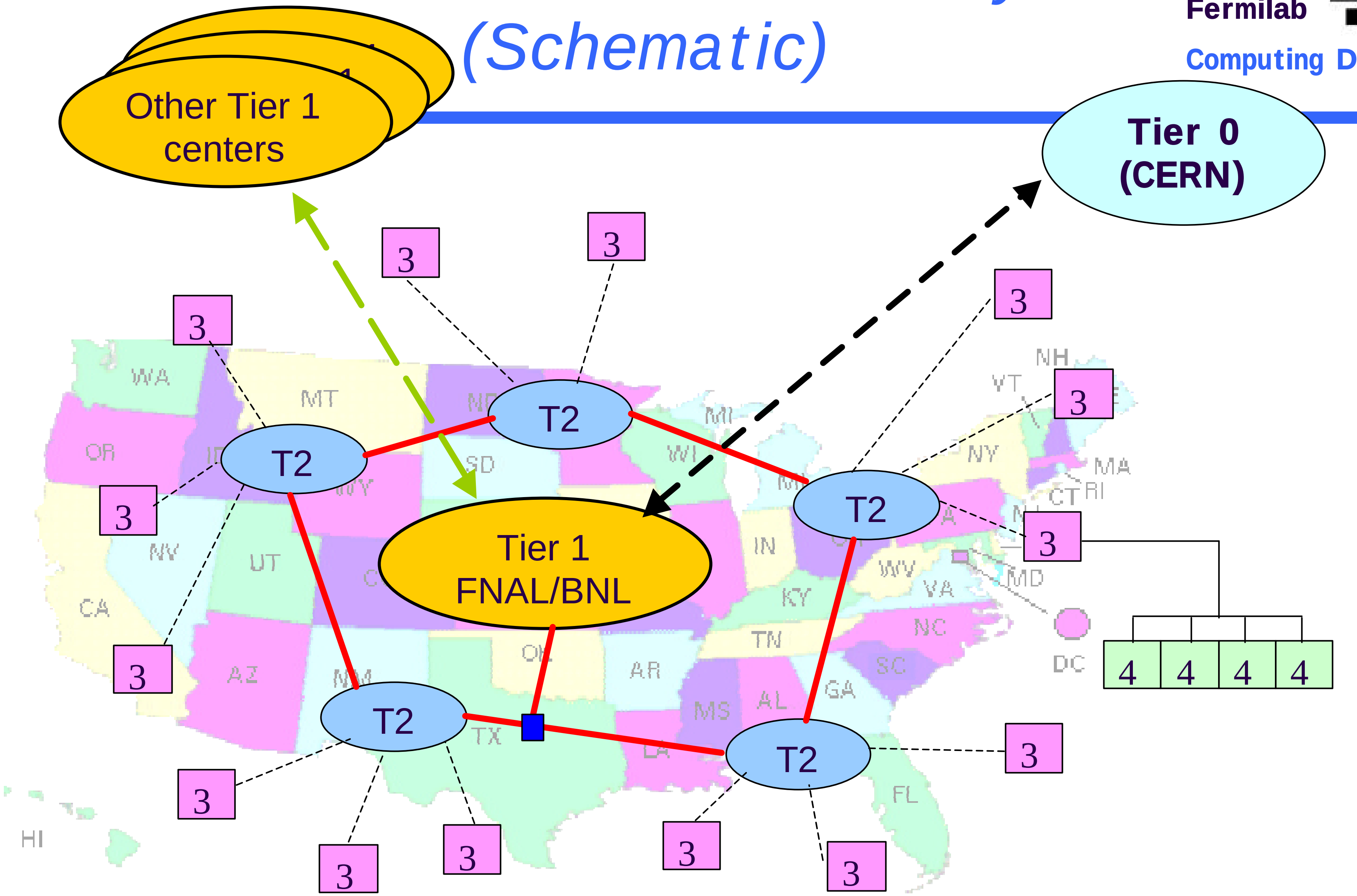
# CMS Computing Solution: A Data Grid

Fermilab
Computing Division

- Deploy computing resources as hierarchical grid
  - Tier 0 ⇒ Central laboratory computing resources (CERN)
  - Tier 1 ⇒ National center (Fermilab / BNL, other countries)
  - Tier 2 ⇒ Regional computing center (university)
  - Tier 3 ⇒ University group computing resources
  - Tier 4 ⇒ Individual workstation/CPU
- We call this arrangement a "**Data Grid**" to reflect the overwhelming role that data plays in deployment
- LHC data volume / Current experiments: factor 2-4
  - CDF: ~ 450 TB / year
  - Compass: ~ 300 TB / year of RAW data
  - STAR: ~ 200 TB / year of RAW data

# *LHC Data Grid Hierarchy (Schematic)*

Fermilab
Computing Division



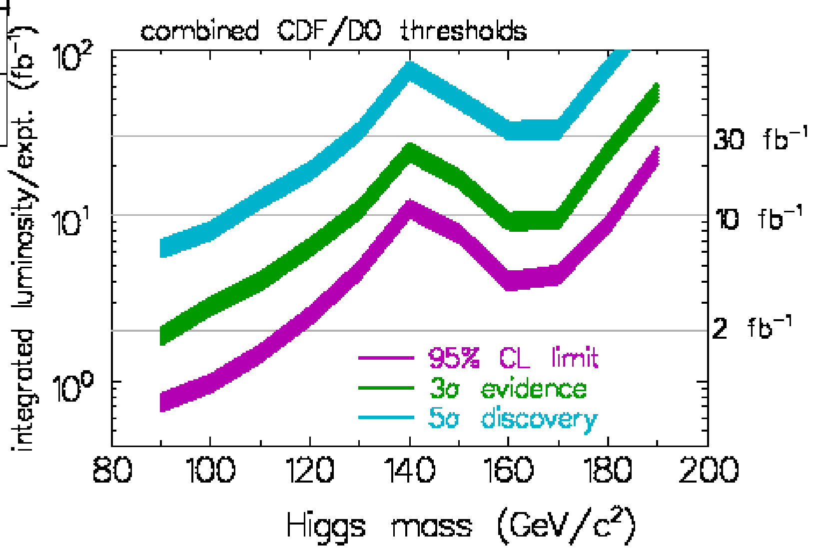
# *"Run 2 Science: The hunt is on"*

## *FermiNews, Vol 24, March 2, 2001*

Fermilab
Computing Division

- Run 2: A long run starts in 2001.
  - Further upgrades will continue increasing the luminosity. They require a **big** effort.

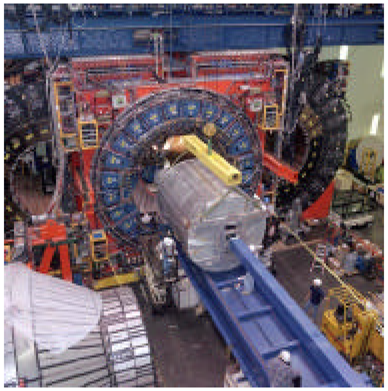
| Run | Dates | Integrated Lumi ($fb^{-1}$) |
|---|---|---|
| I | 1993-1996 | 0.1 |
| IIa | 2001-2003 | ~2 |
| IIb | 2004-2007 | ~15 |



# Run 2a: Data Flows

Fermilab

Computing Division

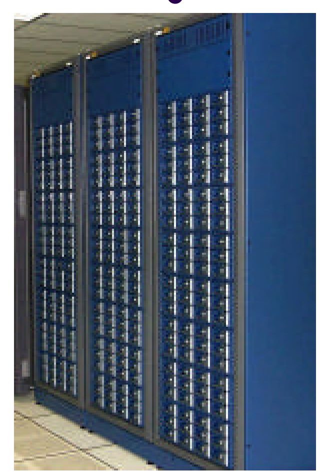


15-20 MBps

100 Mbps

400 MBps

20 MBps



# Run 2a Data Volumes (estimates, from 1997 planning)

Fermilab
Computing Division

| Category | Parameter | D0 | CDF |
|---|---|---|---|
| **DAQ rates** | Peak rate | 53 Hz | 75 Hz |
| | Avg. evt. Size | 250 KB | 250 KB |
| | Level 2 output | 1000 Hz | 300 Hz |
| | maximum log rate | Scalable | 80 BM/s |
| **Data storage** | # of events | 600M/year | 900 M/year |
| | RAW data | 150 TB/year | 250 TB/year |
| | Reconstructed data tier | 75 TB/year | 135 TB/year |
| | Physics analysis summary tier | 50 TB/year | 79 TB/year |
| | Micro summary | 3 TB/year | - |
| **CPU** | Reconstr/event | 25 - 65 SI95xsec | 30 SI95xsec |
| | Total Reconstruction | 2000-4000 SI95 | 2000-4000 SI95 |
| | Analysis | 2000-4000 SI95 | 2000-4000 SI95 |
| **Access for analysis** | # of scientists | 400 - 500 | 400 - 500 |

# *Run 2a Equipment Spending Profile*

**(Total for both CDF & D0 experiments)**

Fermilab
Computing Division

- **Mass storage: robotics, tape drives + interface computing.**
- **Production farms**
- **Analysis computers: support for many users for high statistics analysis (single system image, multi-CPU).**
- **Disk storage: permanent storage for frequently accessed data, staging pool for data stored on tape.**
- **Miscellaneous: networking, infrastructure, ...**

| Fiscal Year | MSS | Farms | Analysis | Disk | Misc | Total (both) |
|---|---|---|---|---|---|---|
| Spent in FY98 | $1.2M | $200K | - | $200K | $400K | $2M |
| Spent in FY99 | $2.2M | $700K | $2M | $800K | $300K | $6M |
| Spent in FY00 | $450K | $350K | $100K | $300K | $800K | $2M |
| Budget FY01 | *$450K* | *$350K* | *$2.14M* | *$690K* | *$70K* | *$4M* |
| *Plan for FY02* | *$500K* | *$1.2M* | *$2.16M* | *$610K* | *$30K* | *$4.2M* |
| *Total Needs* | *$4.8M* | *$2.8M* | *$6.4M* | *$2.6M* | *$1.6M* | *$18.2M* |
| | | | | | | |
| *Continuing Operations (FY2002 and beyond)* | | | | | | *$2M* |

# RUN 2a Equipment



- Analysis servers, going from 30% to 60% of full systems:
  - D0 delivered, CDF purchase in progress
- Disk storage (65% of full system)
- Robots with tape drives

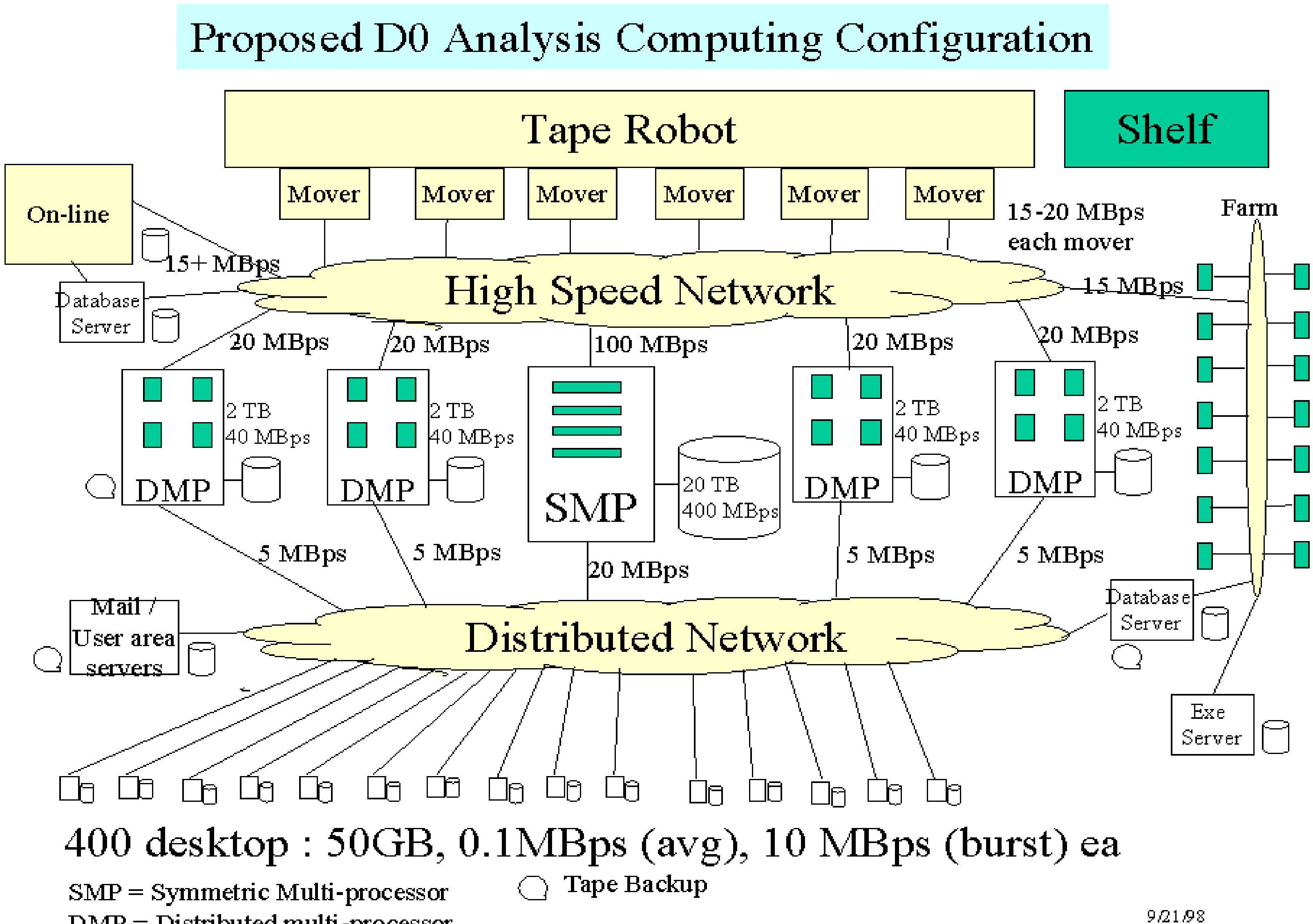
# Status of CDF/D0 Farms

Fermilab
Computing Division

- 88+97 PC's are in place.
  - 48+47 PIII/500 duals
  - 40+53 PIII/750 or 800 duals
- 60 more PC's are on order (PIII/1 GHz duals)
- I/O nodes are ready.
- Integration and testing of the system is complete.
  - 20 Mbytes/sec can be achieved.
- The CDF system is being used to process and reprocess data from the commissioning run
  - about 1.3 Tbytes taken in October, 2000
  - both systems are used to generate and reconstruct simulated data.
- Both are ready and used for raw data reconstruction.

# D0 Computing System

Fermilab
Division



# D0 data handling: SAM - Grid enabled

Fermilab
Computing Division

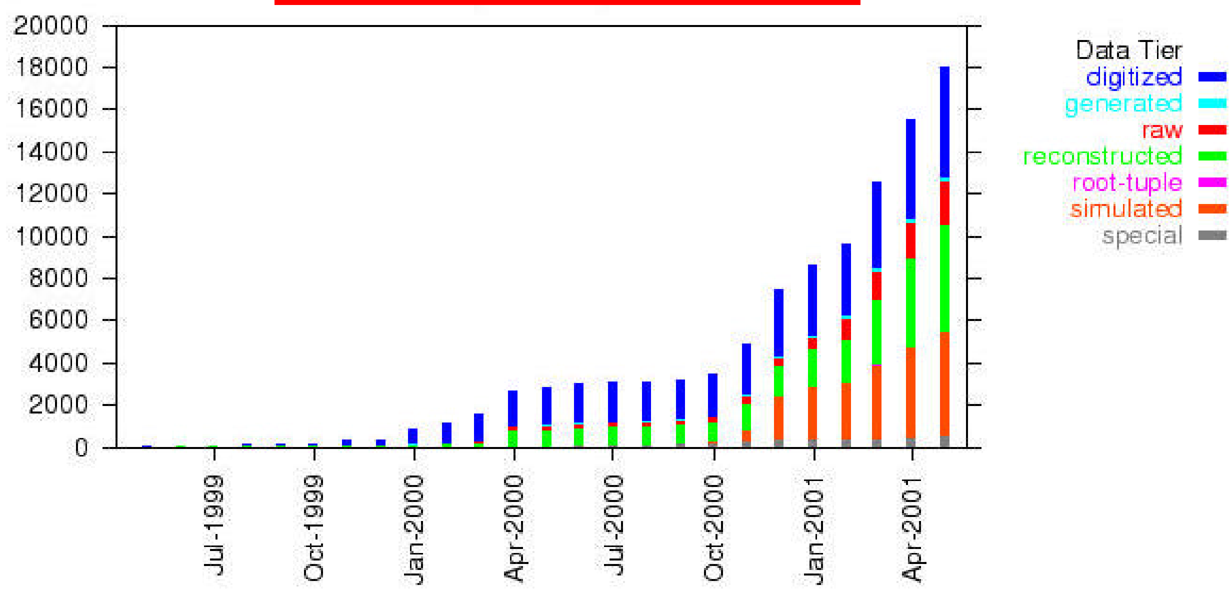
## Fermilab Stations

- ◆Central Analysis
- ◆Online
- ◆Farm
- ◆Linux analysis stations (3)

## Remote Stations

- ◆Lyon (IN2P3)
- ◆Amsterdam (NIKHEF)
- ◆Lancaster
- ◆Prague
- ◆Michigan State
- ◆U. T. Arlington

Gbytes added




*Since time began (for Run2):*

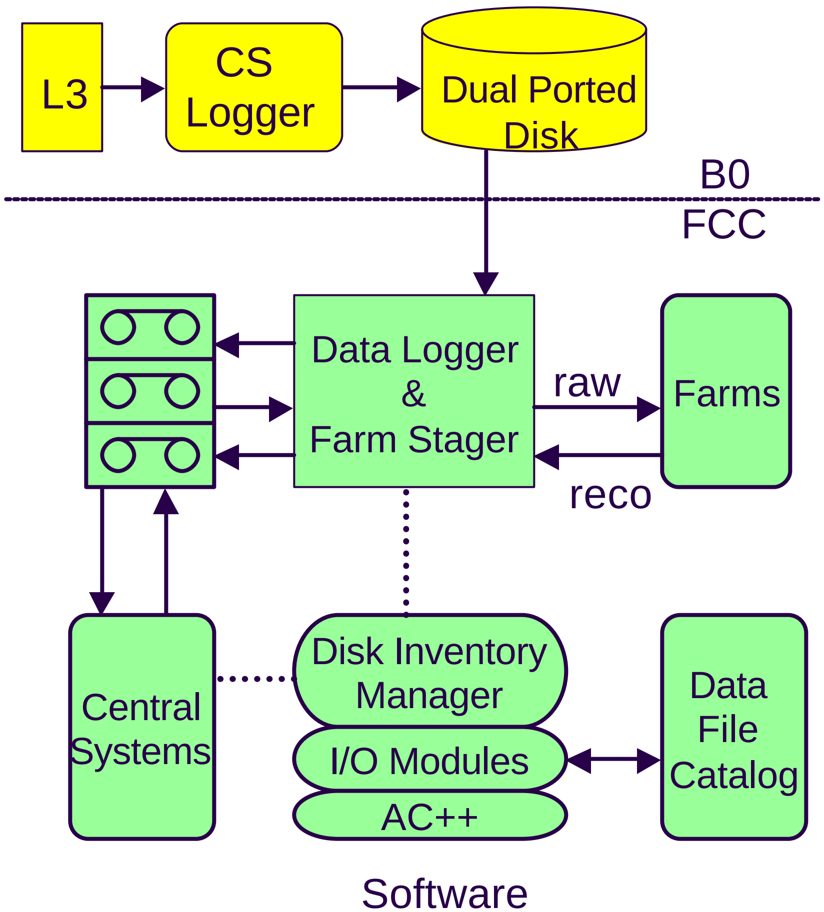
*~100M events*

*~120M files*

*~20 TBytes*

# CDF: Performance Milestones

Fermilab
Computing Division

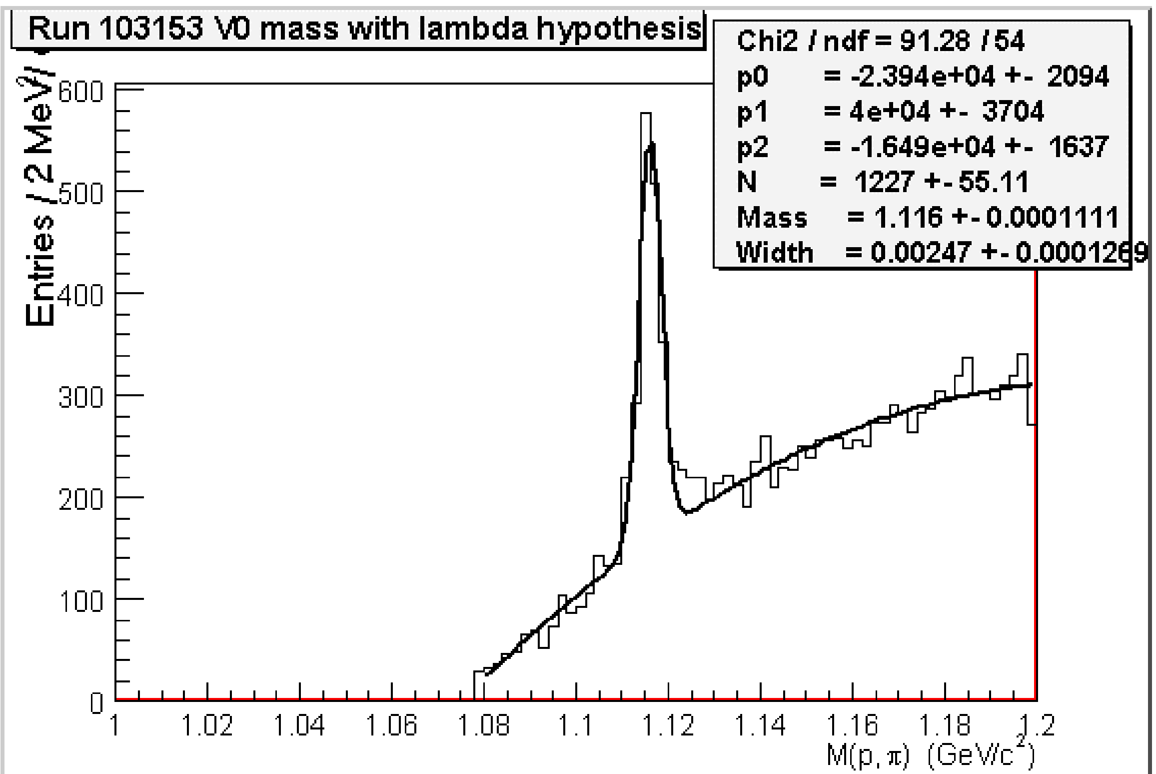


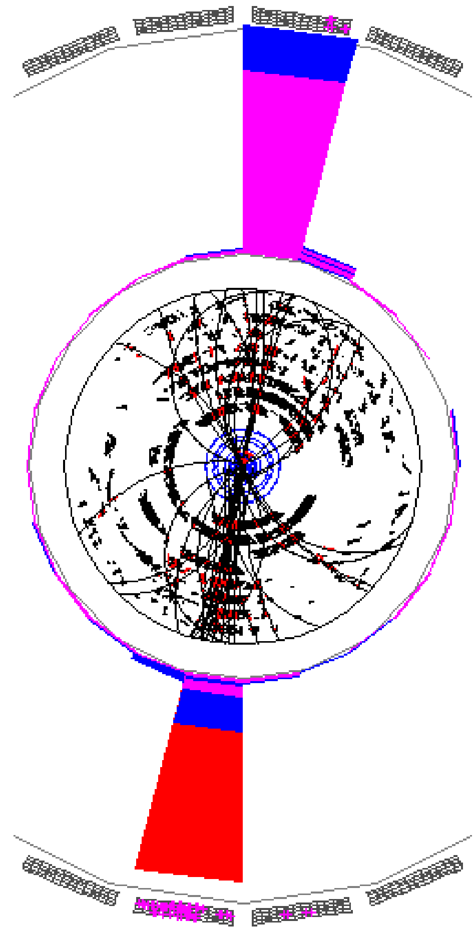

- Mock Data Challenge II May-July 2000.
  - Rate test of system: 12 MB/s.
    - Limited by hardware availability.
  - Single components: 20 MB/s.
- Commissioning: October 2k
  - Recorded 10 million events.
  - Logged 3 TB of data to tape.
  - Reconstructed data on farms.
    - $1^{st}$ pass with Calorimetry & COT
    - $2^{nd}$ pass with SVX & Database
      - 3 s / evt reco (debug & non-opt)
- Start of Run2: March 1
  - Start of physics data taking: ~June 1

# CDF Commissioning Run Data

Fermilab
Computing Division

- Functioning tracking (e.g. reconstructed $\Lambda \rightarrow p\pi$ below).
- Functioning calorimetry (e.g. event with two 200 GeV jets below).



# First planning for Run 2b

Fermilab
Computing Division

- First Run 2b costs estimates based on scaling arguments
  - Use predicted luminosity profile
  - Assume technology advance (Moore's law)
  - CPU and data storage requirements both scale with data volume stored
- Will be able to refine estimates after 1 year of Run 2a experience
- Data volume depends on physics selection in trigger
  - Can vary between 1 – 8 PB (Run 2a: 1 PB) per experiment
- Have to start preparation by 2002/2003

# LCCWS: Cluster computing now

Fermilab Computing Division

- Large compute clusters are reality and used now in HEP
  - For raw data and Monte Carlo production
  - For accelerator calculations
  - For theoretical physics calculations (Lattice QCD, …)
  - Particle Astrophysics Data processing
- Cluster computing is an obvious solution when
  - ratio of CPU/IO-bandwidth is high
  - Parallel computing can benefit from granularity of events
  - Cross-node communication requirements are low
  - Resource allocation is easy
    - Within one group or tightly managed otherwise

# Cluster Computing: Some Questions (1)

Fermilab
Computing Division

- Should or can clusters emulate a mainframe?
  - Resource allocation, Accounting, Monitoring
  - single vs. heterogeneous environment
  - System administration
- How much can the compute models be adjusted to make most efficient use of cluster computing?
- Were is it more cost-efficient **not** to use cluster computing?
- What is the total cost of ownership for clusters?
- How can a cluster be build based on incidental use of desktop resources (à la seti@home)?

# Cluster Computing: Some Questions (2)

Fermilab
Computing Division

- How to use clusters for applications with high I/O requirements?
  - Implications on data model
  - Implications on data access methods
  - I/O bandwidth and latency requirements and solutions
  - Is there a break-even point?

- How to design clusters for high-availability requirements?

# *Compute clusters: the next steps*

Fermilab
Computing Division

- We are facing the same challenge at different labs:

  ***How to get the most computing for the least $$ ?***

  - Lets share our experience…
  - Lets exchange ideas…
  - Can we share solutions?
  - How to entertain joined projects?

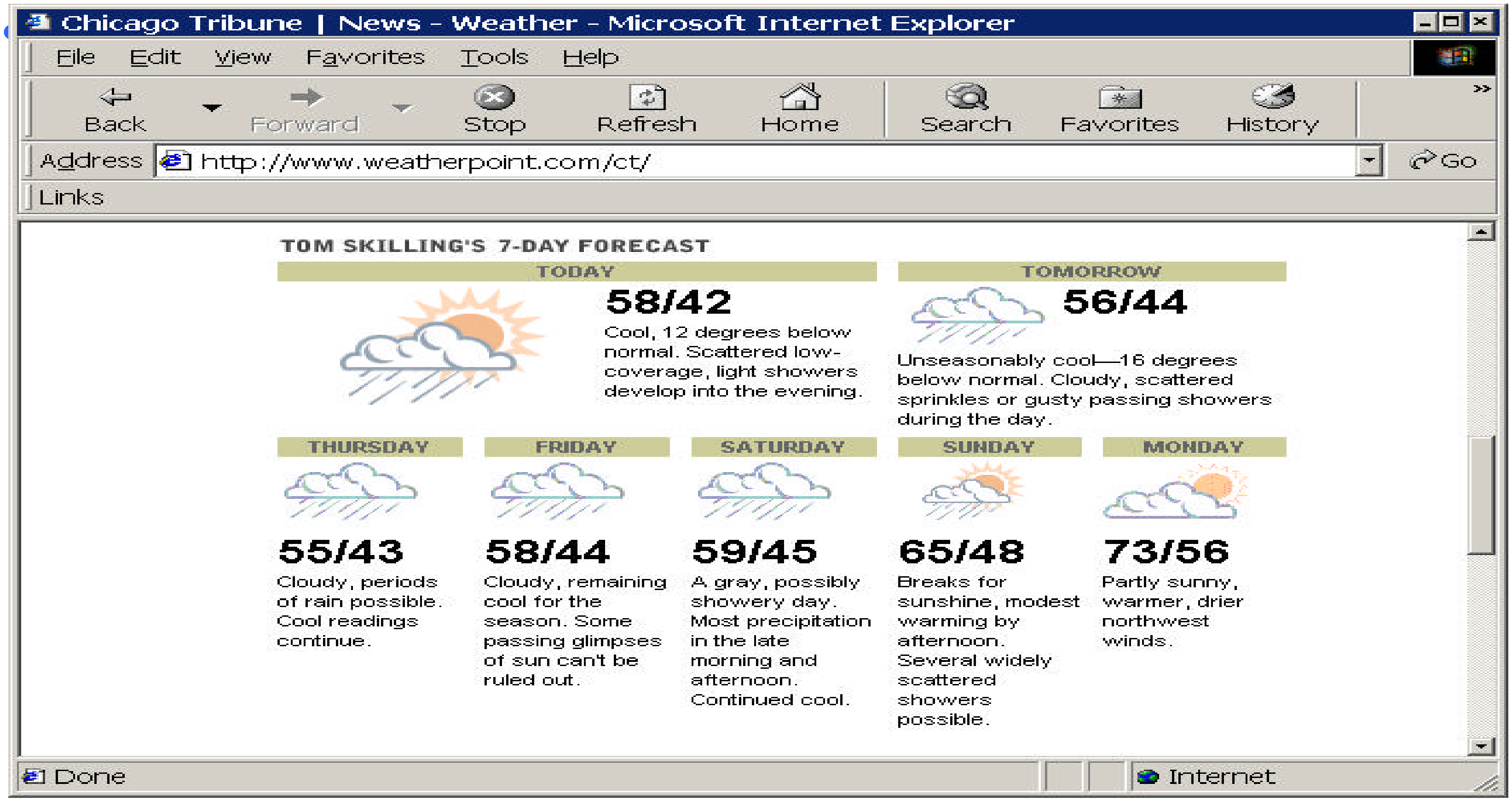
- That is why we are here….

***Welcome !!***

- We need to do global computing for global experiments
  - Grid computing and data distribution is addressing this
  - Not to forget operational aspects…

# *At last: The Weather*

Fermilab
Computing Division




- **The weather will be ideal for good workshop working conditions** *it will rain….;-(*