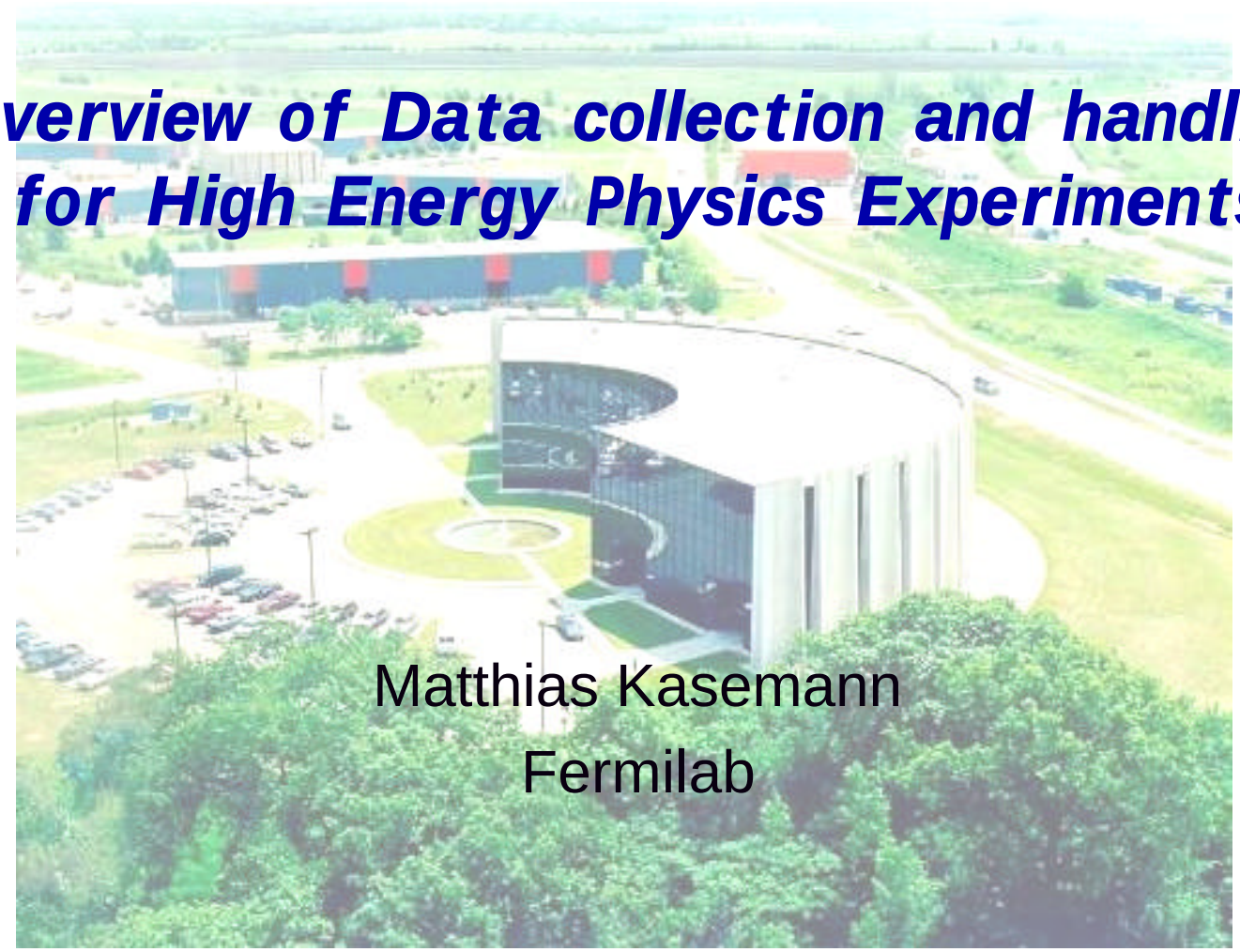
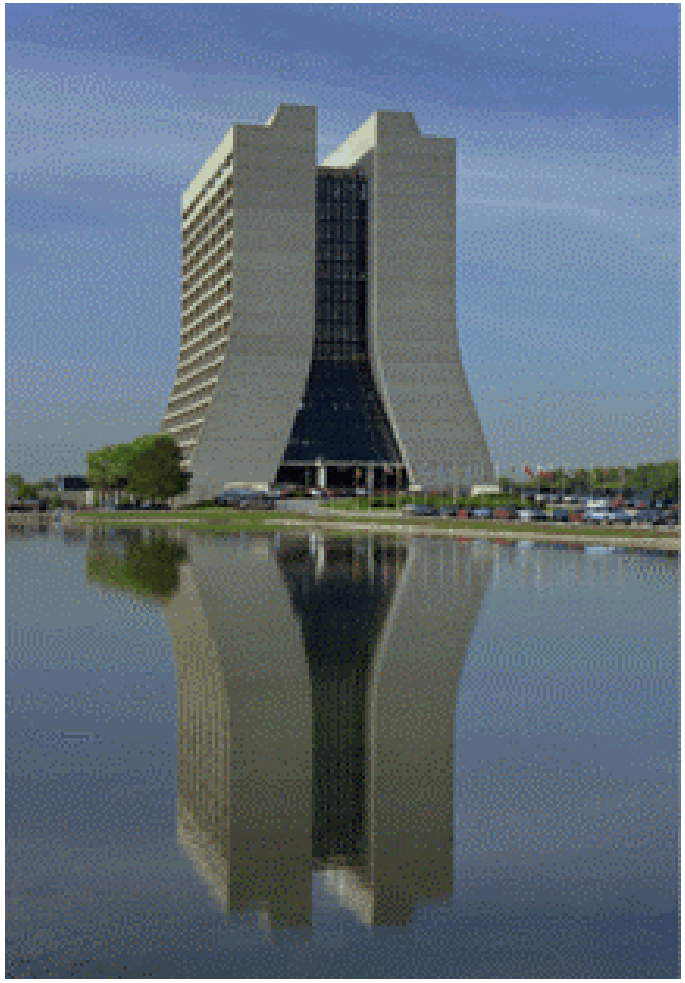


Overview of Data collection and handling for High Energy Physics Experiments

Matthias Kasemann
Fermilab

Fermilab Mission Statement (see Web)



Fermi National Accelerator Laboratory advances the understanding of the **fundamental nature of matter and energy** by **providing leadership and resources** for qualified researchers to conduct **basic research** at the frontiers of high energy physics and related disciplines.

Fermilab operates the world's highest-energy particle accelerator, the Tev-atron. More than **2,200 scientists from 36 states and 20 countries** use Fermi-lab's facilities to carry out research at the frontiers of particle physics.

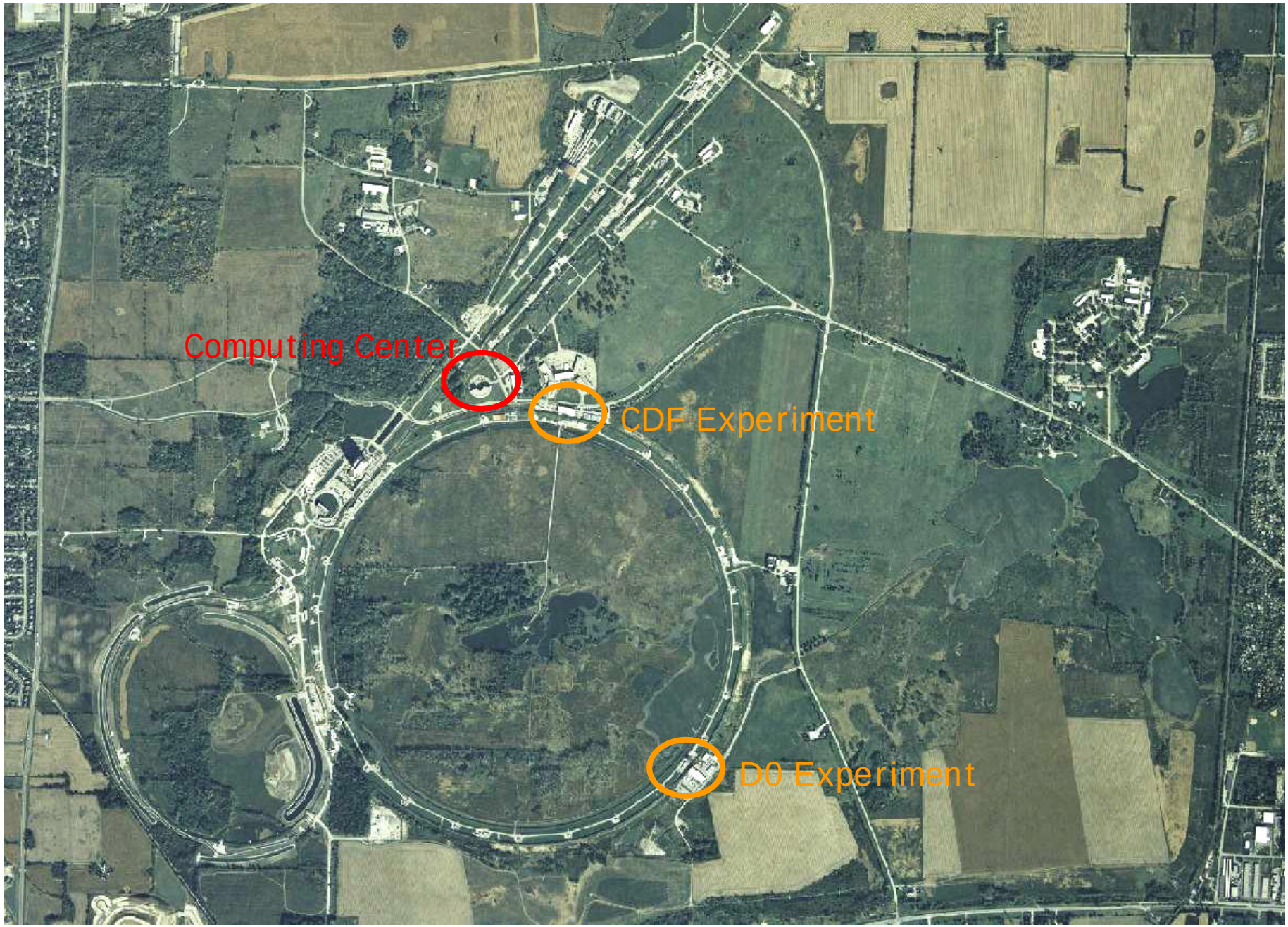
Fermilab Community: Collaborations

- Fermilab is an open site (no fences) and acts as a host to the many universities and institutions pursuing research here.
- Given that, the culture of the lab is very university-like, which is one of its big strengths for scientific research.
- Collaborations:
  - 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

Collaborations

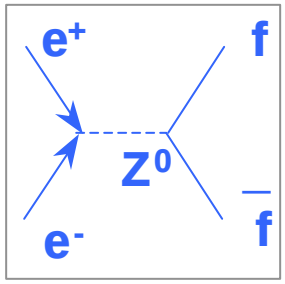
Fermilab

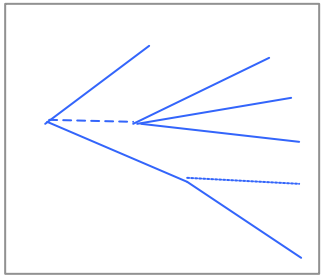
- Detectors are designed and built by large collaborations of scientists and technicians
  - Many tens of institutions (mainly universities)
  - Many hundreds of people
  - Many countries
- Important Run2 Milestone:
  - CDF and D0 (Σ=800**+** scientists) started data taking 03/01/2001 after 4 years of preparation
  - Unique scientific opportunity to make major HEP discovery – **don't miss it**!!

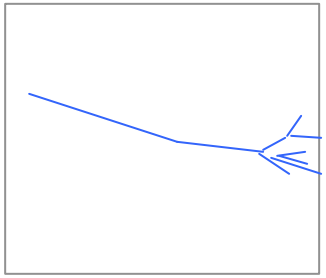

Computing Center
CDF Experiment
D0 Experiment

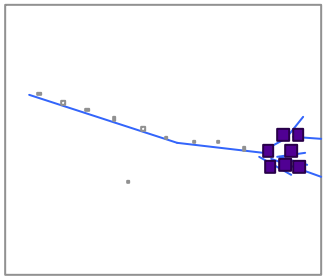
From Physics to Raw Data: what happens in a detector

Fermilab









250Kb – 1 Mb

2037 2446 1733 1699
4003 3611 952 1328
2132 1870 2093 3271
4732 1102 2491 3216
2421 1211 2319 2133
3451 1942 1121 3429
3742 1288 2343 7142

Theoretical Model of Particle interaction

Fragmentation, Decay

Interaction with detector material

Multiple scattering, interactions

Detector response

Noise, pile-up, cross-talk, inefficiency, ambiguity, resolution, response function, alignment, temperature

Raw data (Bytes)

Read-out addresses, ADC, TDC values, Bit patterns

Particle production and decays observed in detectors are Quantum Mechanical processes. Hundreds or thousands of different production- and decay-channels possible, all with different probabilities.

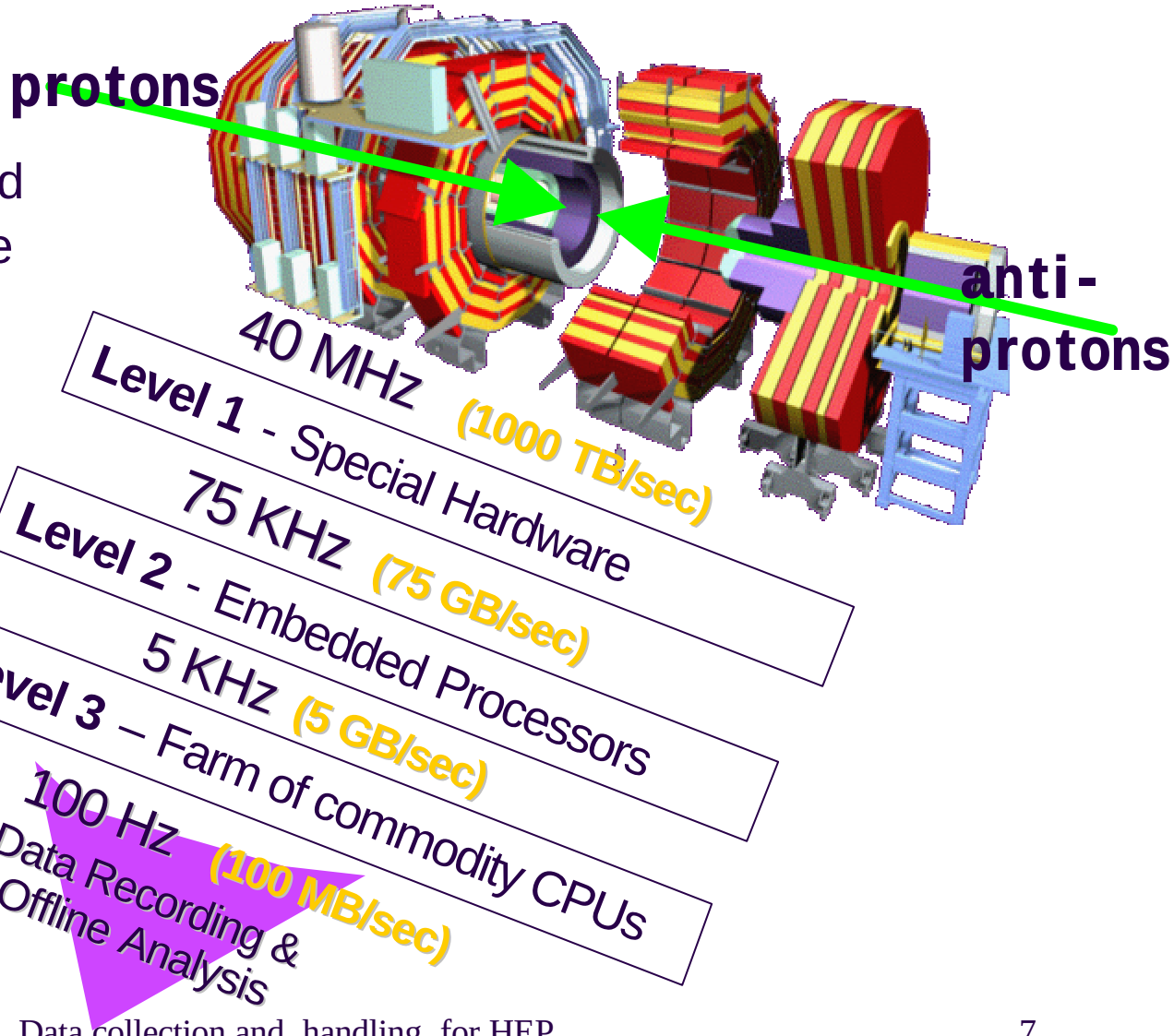
In the end all we measure are probabilities!!

Data reduction and recording: here CMS in 2006

Fermilab

On-line System

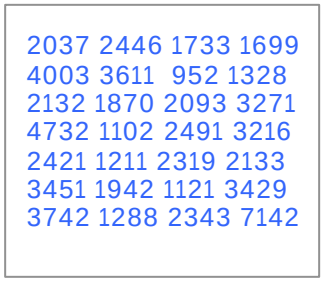
- Multi-level trigger
- Filter out background
- Reduce data volume
- 24 x 7 operation

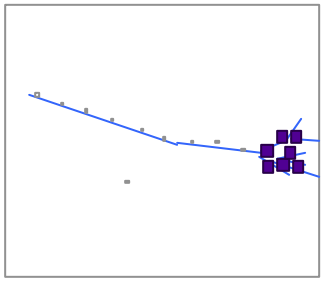


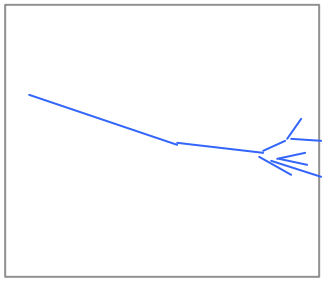
From Raw Data to Physics: what happens during analysis

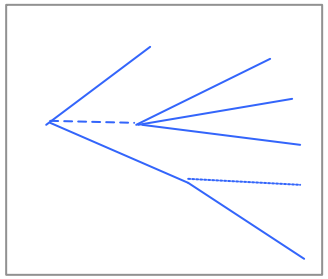
Fermilab

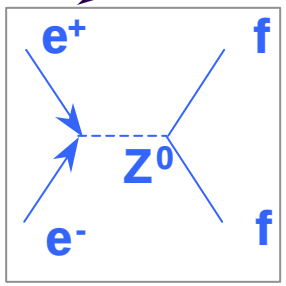
250Kb – 1 Mb | 100 Kb | 25 Kb | 5 Kb | 500 b











Raw data

Convert to physics quantities

Detector response

apply calibration, alignment,

Interaction with detector material

Pattern, recognition, Particle identification

Fragmentation, Decay

Physics analysis

Basic physics

Results

Reconstruction

Analysis

Simulation (Monte-Carlo)

Data flow from detector to analysis

Fermilab

Experiment

15-20 MBps

Permanent storage

100 Mbps

“analysis CPU”

20 MBps

“reconstruction”

400 MBps

“analysis disks”

Run IIa Equipment Spending Profile

(Total for both CDF & D0 experiments)

Fermilab

- 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*

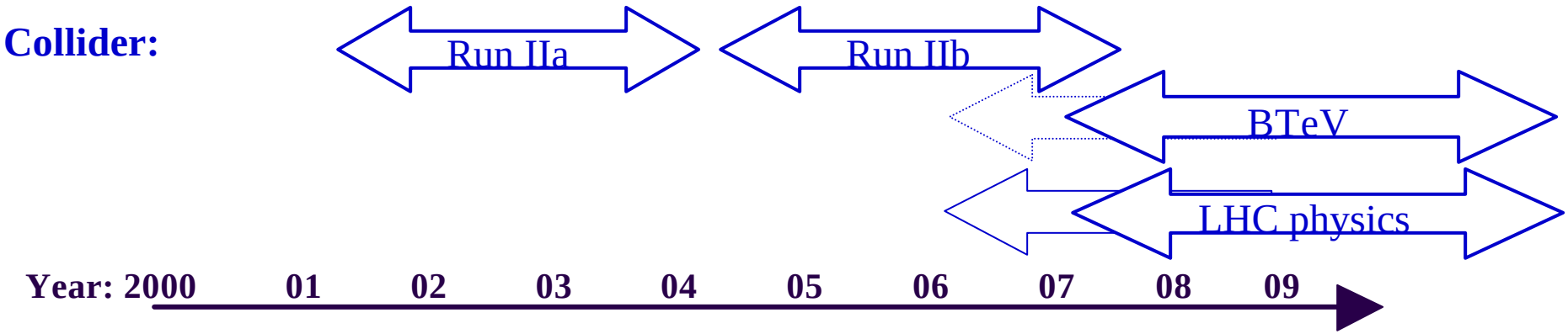
RUN IIa Equipment

Fermilab

- Analysis servers
- Disk storage
- Robots with tape drives

Fermilab HEP Program

Fermilab



Run 2 Data Volumes

Fermilab

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 MB/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

- 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
- 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

D0: Data Volume collected

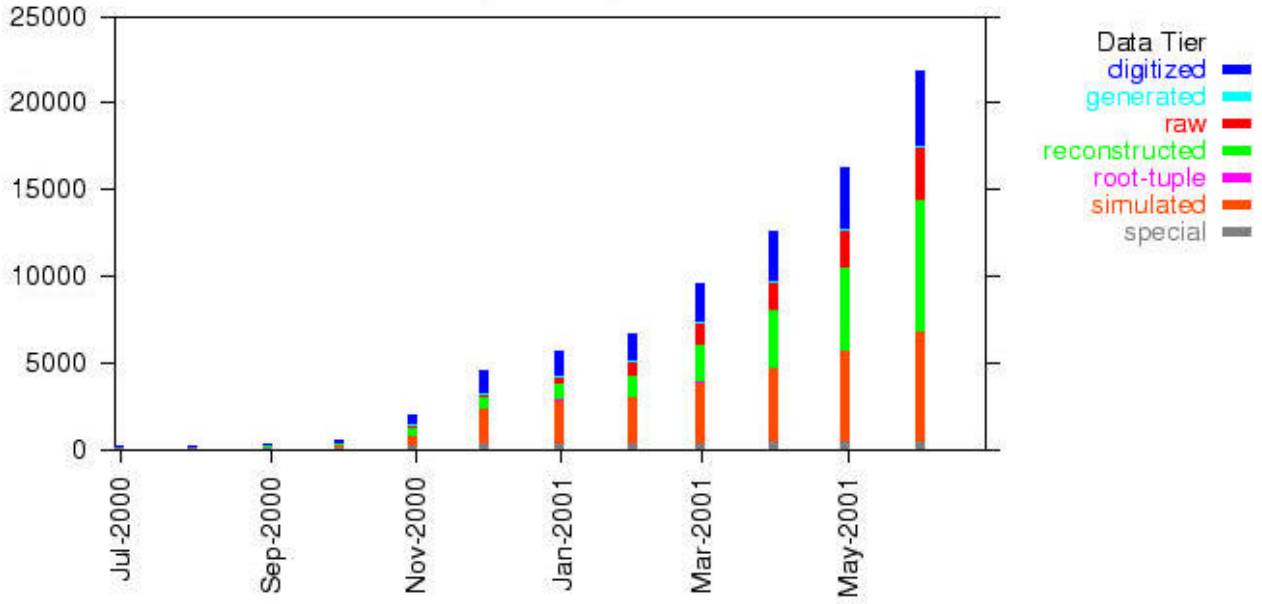
Fermilab Stations

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

Remote Stations

- Lyon (France)
- Amsterdam (Netherlands)
- Lancaster (UK)
- Prague (Czech R.)
- Michigan State
- U. T. Arlington

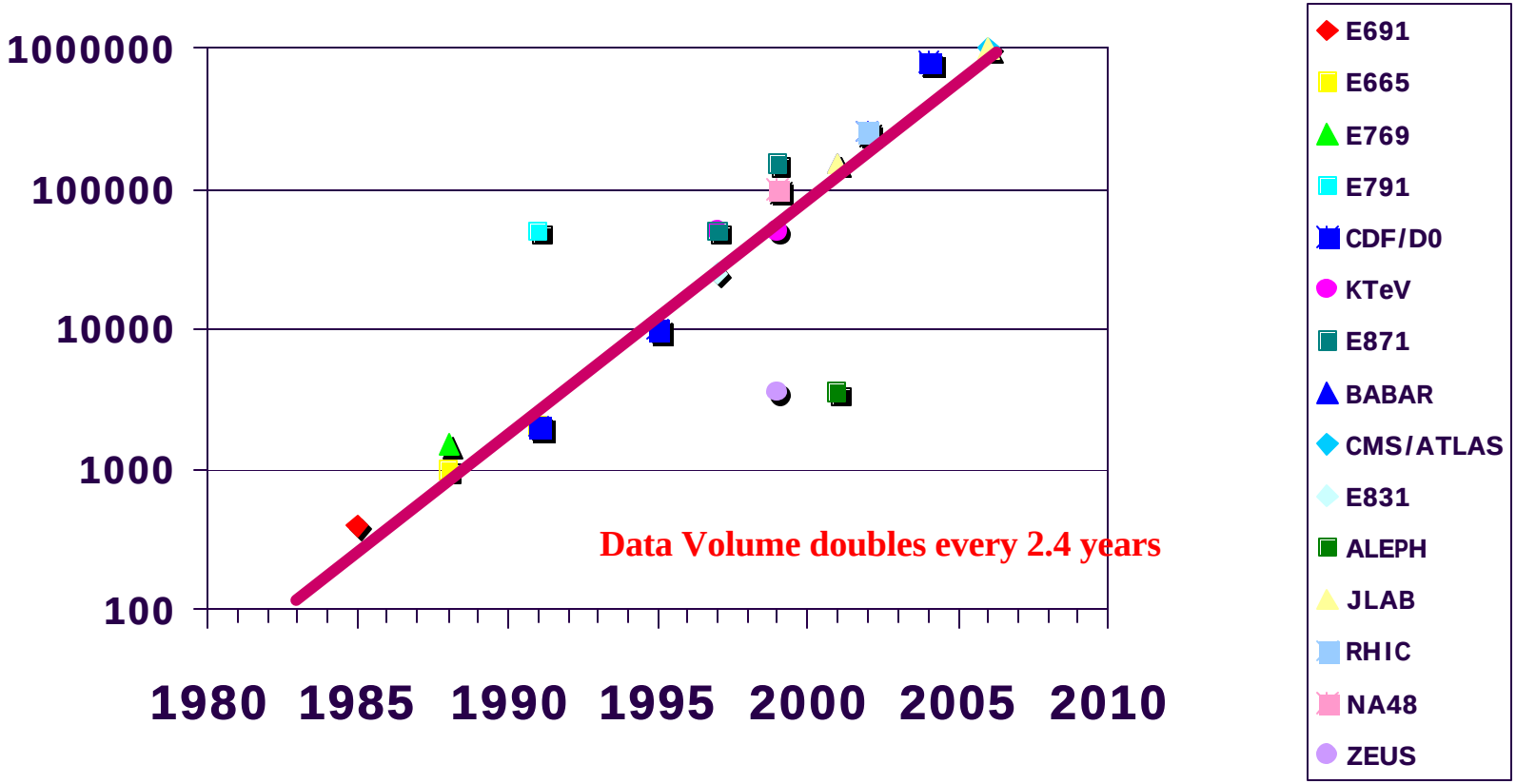
Data Volume [GB]



...in the past 12 Months

Data Volume per experiment per year (in units of Gbytes)

Fermilab



Improving our knowledge: better experiments in HEP

Fermilab

Desired Improvement		Computing Technique
Higher energy	→	Accelerator Design/simulation
More collisions	→	Acc. Design and controls
Better detectors	→	Triggers (networks, CPU), simulation
More events	→	Disk, tape, CPU, networks
Better analysis	→	Disk, tape, CPU, networks, algorithms
Simulation	→	CPU, algorithms, OO
Theory	→	CPU, algorithms, OO

How long are data scientifically interesting? "Lifetime of data"

Fermilab

- 1. Month after recording:
  - Verification of data integrity
  - Verification of detector performance and integrity
- 6-12-24 months after recording:
  - Collect more data
  - Process and reconstruct "interpret the bits"
  - Perform data analysis
  - Compare to simulated data
  - Publish!!
- >2 years after recording:
  - Data often superseded by more precise experiments
  - Combine results for high statistics measurements and publish!!
  - Archive for comparison and possible re-analysis
- >5 years after recording:
  - Decide on long-term storage for re-analysis

Tape storage history: the last 10 years

- 1990: “New tape retention policy”
  - Maximize accessibility of tapes actively used
  - Provide off-site storage for data which have finite probability of being needed in the future
  - Default retention period in FCC set to 5 years, extended on justified request
  - Disposition of redundant and obsolete tapes decided together with experiment spokespeople (based on scientific value)
- 1992: “Too many tapes, need room to store new data”
  - Tapes not accessed within years moved to off-site storage
  - Tapes retrievable within few working days
- 1998: New “Fermilab Tapes Purchasing Policy”
  - Tapes intended for “long term” storage are purchased and owned by FNAL
  - Tapes cannot be removed from FCC/storage by experimenters + users
- 1999 – 2000: remove 9-track round tapes from FCC archive
  - Data needed in the future moved to off-site storage
  - Disposition of redundant and obsolete tapes decided together with experiment spokespeople (based on scientific value)

Relevant questions for tape storage:

- All questions wrt. Storage or Disposal are discussed and procedures established in concurrence with **DOE records manager in FAO/CH**
- In order to satisfy DOE requirements for disposal we request written approval from spokesperson:
  - What is stored on the individual tape? (raw data, summary data, etc)
  - When were the tapes written?
  - Do subsequent summary tapes exist (are tapes redundant)?
  - Do you foresee future research needs for these tapes?
  - Are software and computers still available for reanalysis or is it feasible to write new software to do so?
  - Do you have any reason to retain the tapes? Explain…