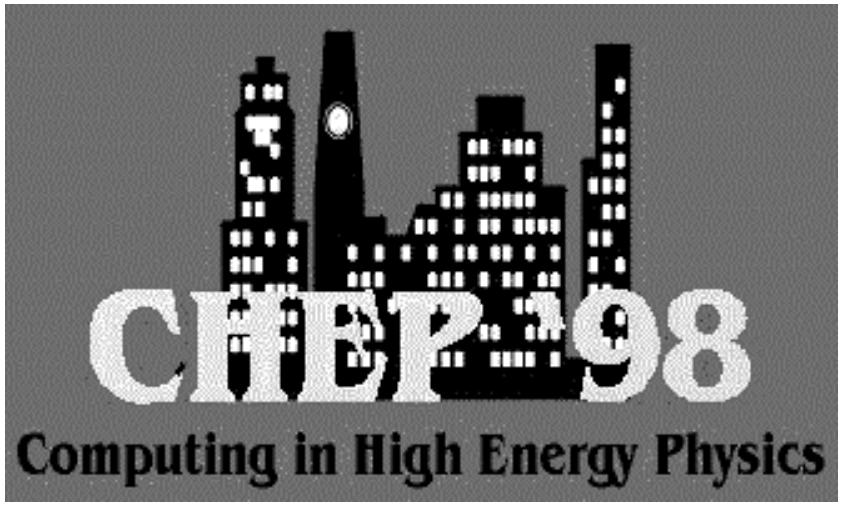


Highlights from CHEP98

J.Harvey

August 31 - September 4, 1998

Hotel Inter-Continental

Chicago, Illinois, USA

Sponsored by

Argonne National Laboratory

Some details…..

- 419 participants (~50% USA)
- 310 talks
  - All prepared electronically
  - 50% available via web by time conference started
  - ~50% presented electronically

http://www.hep.net/chep98/index_papers.html

Parallel Sessions

Session A - Data Analysis and Presentation
Session B - Data Acquisition and Control Systems
Session C - Mass Storage and Data Management
Session D - Farms, Commodity Computing,
Networks and Communication
Session E - Tools
Session F - Algorithms and Methods

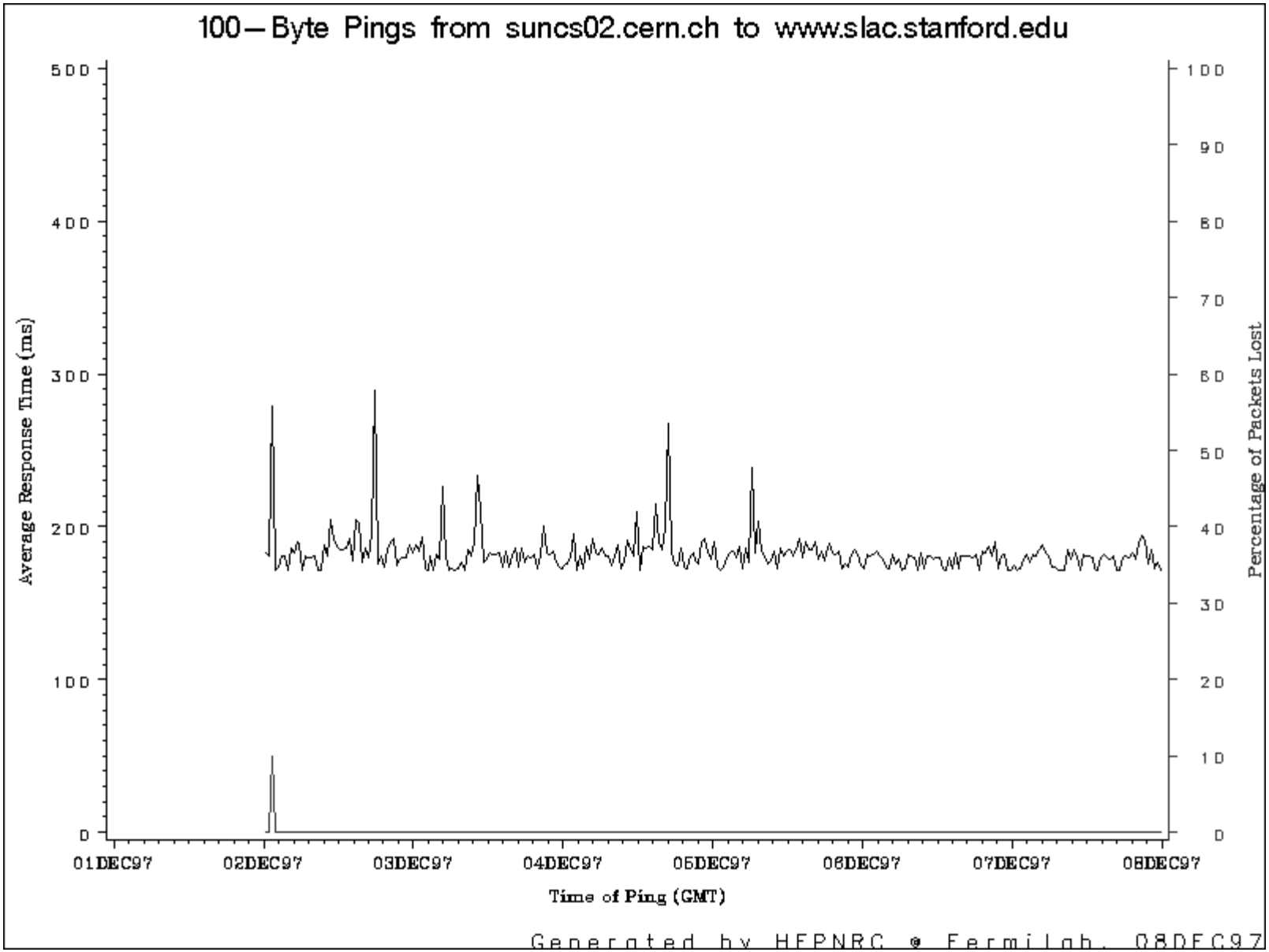
Experiments

❒ DESY	**HERA-B**		50TB/yr	‘98/’99
❒ KEK	Belle			‘99
❒ SLAC	**BaBar**		300TB/yr	‘99
❒ BNL/RHIC	BRAHMS,**PHENIX**,PHOBOS,**STAR**		1.5PB/yr	‘99
❒ Fermi Lab	**CDF** and **D0**	Run II	500TB/yr	‘00 - ‘02
		Run III	5PB/yr	‘03 - ‘05
❒ CERN	ALICE, **ATLAS**, **CMS**, LHCb		5PB/yr	‘05 -

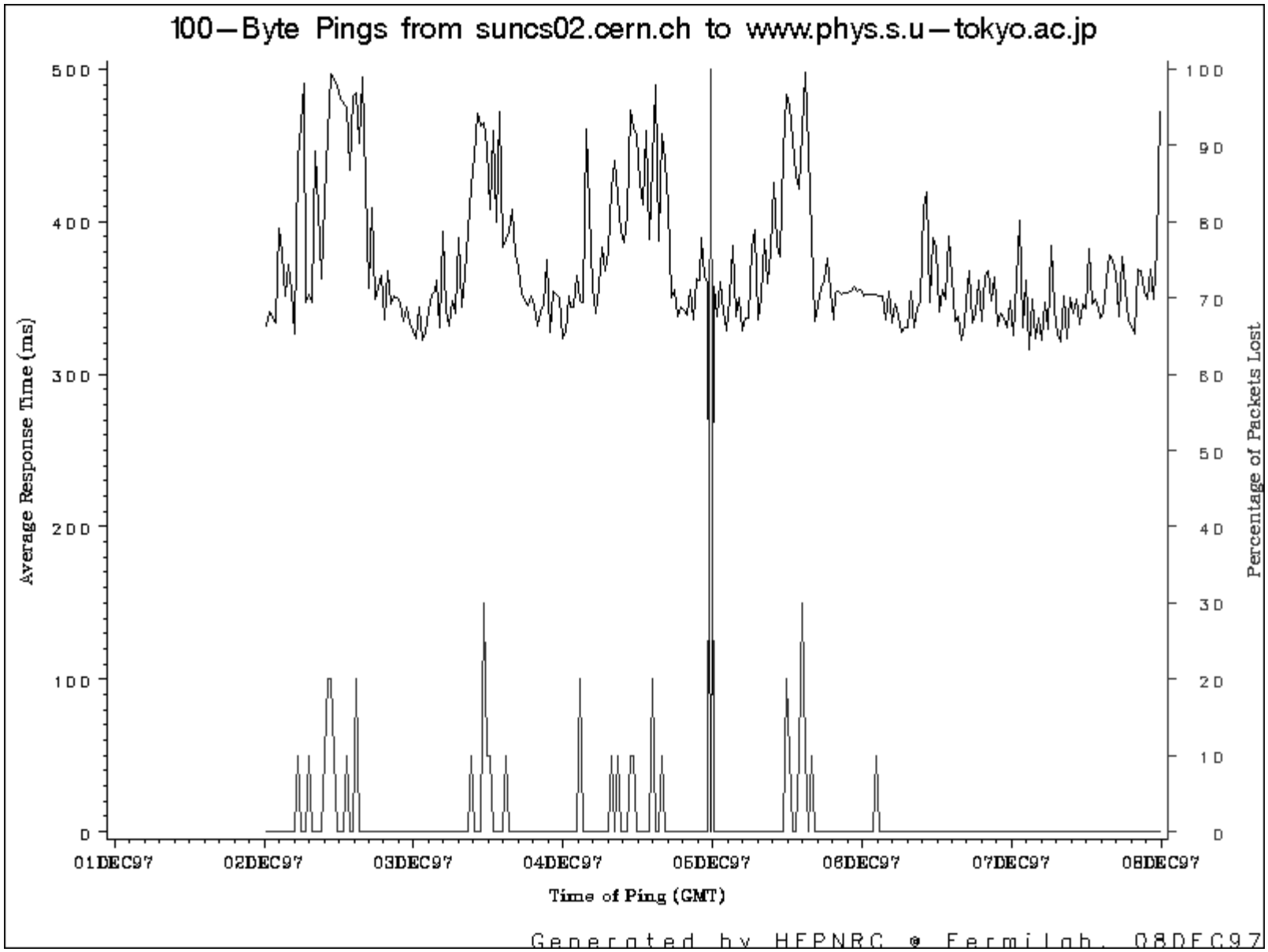
Networking needs and prospects

- ICFA Networking Task Force (NTF) setup to evaluate the status of networking and to make recommendations
- Hundreds of computers (mainly in institutes) test the quality of their connections to tens of sites (mainly accelerator labs).
- Data stored and made available for Web access:
  - http://www.hep.net/cgi-bin/graph_pings.pl
  - http://sitka.triumf.ca/net/nodes.frameset.html
- Perceived quality of service depends on
  - Packet Loss Rate
    - results from congestion, email always works…Telnet doesn't
    - <1% excellent, <2.5% good, <5% OK, >12% unusable
  - Round trip time
    - good ~30 msec, intercontinental ~300 msec, problem cases >500 msec

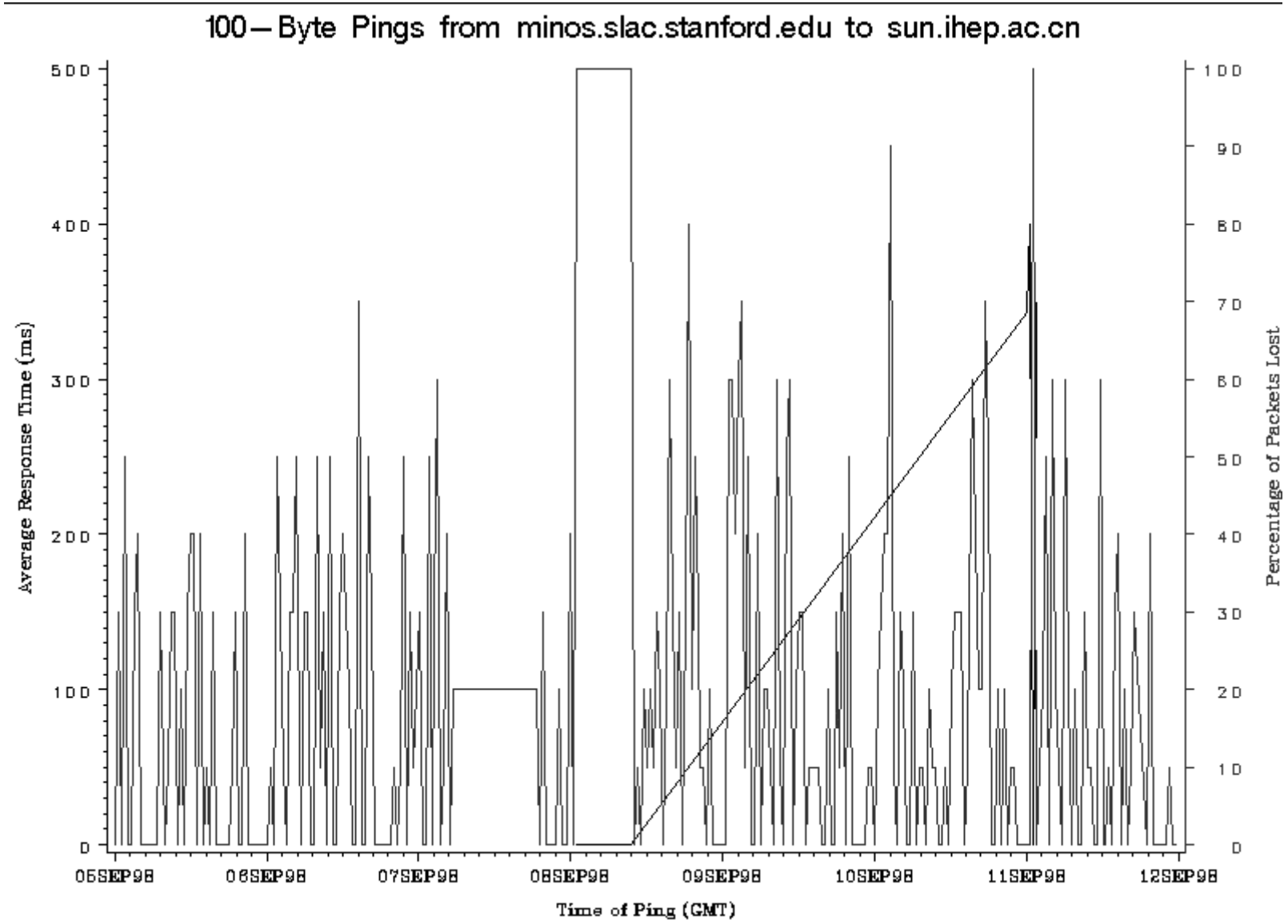
CERN to SLAC


100—Byte Pings from suncs02.cern.ch to www.slac.stanford.edu
Average Response Time (ms)
Percentage of Packets Lost
500
400
300
200
100
0
100
90
80
70
60
50
40
30
20
10
0
01DEC97
02DEC97
03DEC97
04DEC97
05DEC97
06DEC97
07DEC97
08DEC97
Time of Ping (GMT)
Generated by HEPNRC @ FermiLab, 08DEC97

CERN to Tokyo



SLAC to China



Performance Summary

	PLR(%)	RT(msec)	Comment
Fermi-Austin	0	30	National/Perfect
Bologna-Florence	0	30	..
KEK-Osaka	0	30	..
CERN-Lund	0	60	Internat./Perfect
FNAL-DESY	<1	150	..
CERN-KEK	<1	330	..
CERN-ITEP Moscow	3.5	500	Internat./Problem
DESY-SantaCruz	10		US institute
CMU-IN2P3	10		congested TA link
KEK-Texas	12		congested link
FNAL-Brown U.	16		changed I. Provider
SLAC-Beijing	20		+Argentina, A,NZ

Outlook

- Use data to help fix and to predict bottlenecks and other performance problems.
- Questionnaire to experiments - need factor 10 growth every 3-4 years to meet the needs of LHC.
- Improve the international connections e.g. need extra bandwidth for ICFA traffic over Atlantic (October w'shop).
  - 2 new cable systems
    - move from 2.5 Gbps to 10 Gbps
    - move to Wavelength Division Multiplexing (x100)
  - Project Oxygen - global optical fibre cable network
    - 16,000 km, 100 landing points, 16x more bandwidth/cable than At.X
    - pricing independent of destination
    - full commercial service beginning in 2002

DANTE Dai Davies

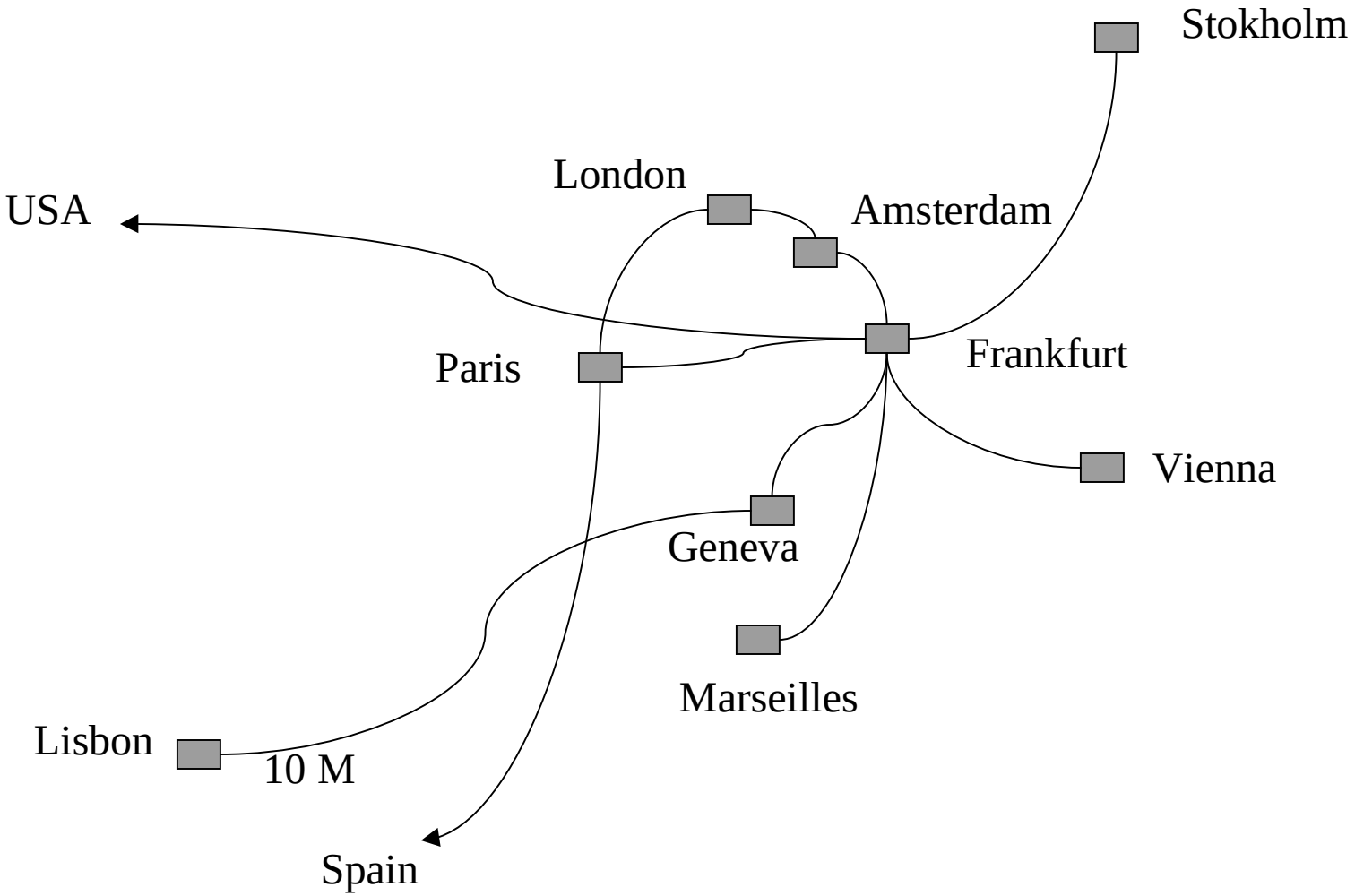
- TEN-155 Pan European network
  - ➢ managed by consortium, co-funded by EU
- In past economics driven by monopoly market - now improving following deregulation

‘96	2 Mbps	Circuits	220 k$/Mbps/yr
‘97	34 Mbps	ATM VP	165 k$/Mbps/yr
‘98	155 Mbps	SDH	33 k$/Mbps/yr

- Platform for IP service and quality of service
  - ➢ challenge is managerial : quality defines cost
- US connectivity : now 45 Mbps, future 155 Mbps ? (issue is cost-sharing)
- Future : TEN-155 for 3 years, with plans for 622 Mbps and 2 Gbps in a 4 year framework

DANTE

TEN-155 Pan European Network

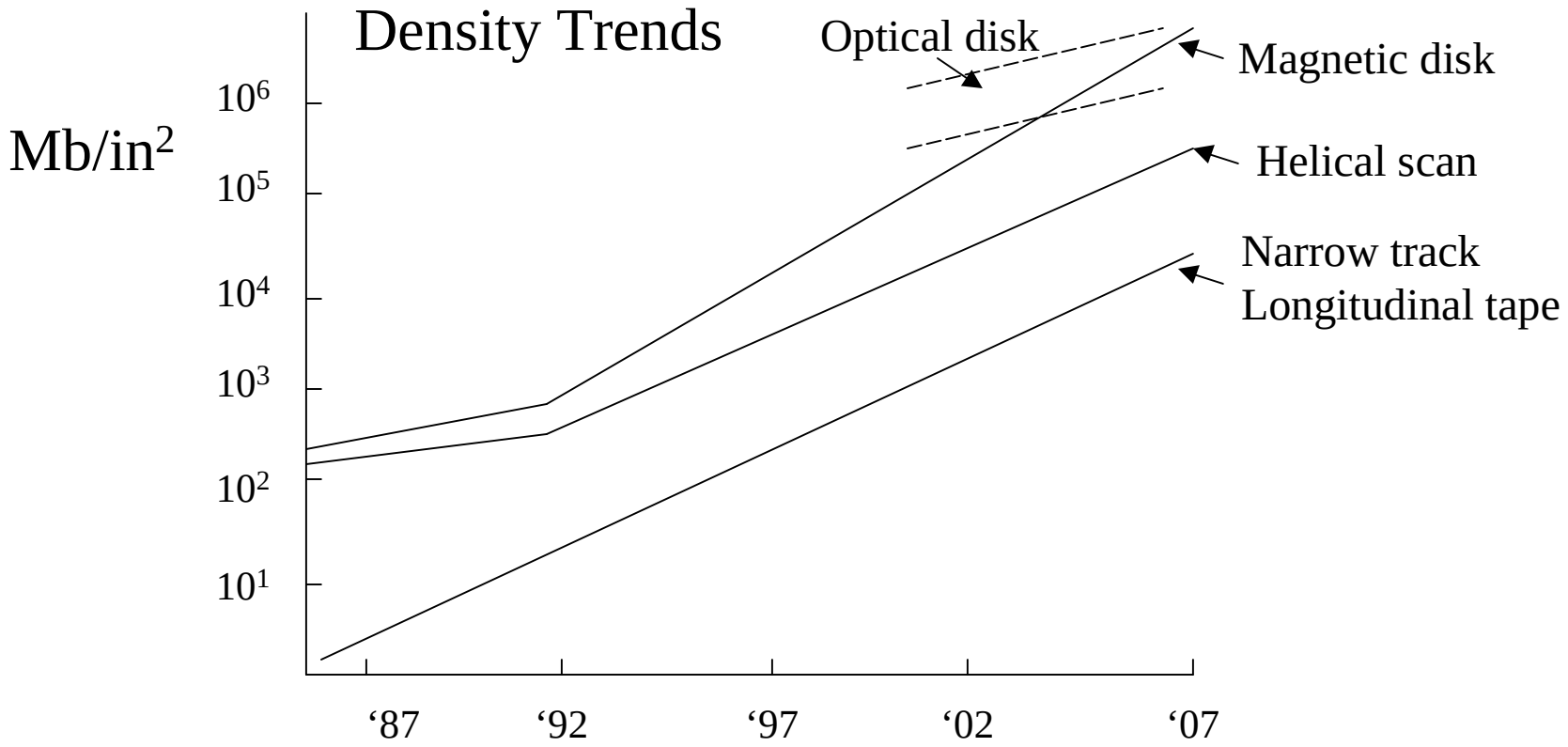


Data Storage Strategies

Gary Sobel/ StorageTek

- Storage needs moved quickly from TB to PB
- By end of this year needs will be 5-6 TB/ day (imaging applications)
- '02 : some customers needs will be 1 PB/day (ExaB/yr)
  - 7.3 M x 50 GB cartridges
  - 1000 transports @ 11 MB/s
  - 4 acres of real-estate
  - huge power bill
- "Caught by surprise"

Density Trends



- Magnetic disk outpacing all storage technologies (60% per year, will continue)
- By '03, 300 GB capacity on 3.5 " : 30 Gb/in^2
- Super paramagnetic limit reached in ~'03 (thermal energy destroys magnetic after 1day-1 year)
- Tapes give volumetric storage advantage

Product Trend Tape Product Family

(N.B. Internet transmission of talk turned off)

	Capacity(GB)	Speed(MB/s)	When
Redwood	**50**	**11.1**	**Now**
PT1	**100**	**10**	**3Q99**
PT2	**150**	**20**	**3Q01**
PT3	**300-450**	**40**	**1Q03**
PT4	**750-1100**	**50**	**1Q05**
PT5	**2000**	**60**	**1Q07**

- Increase track density to minimise amount of tape (9μm)
- ATLAS,CMS ~3 tapes per day & 2 drives (100 MB/s)
- LHCb 1 tape /day & 1 drive
- ALICE would need 40 drives to achieve 2 GB/s

PC Computing - Farms

35 talks on PC-related Computing (compared to 7 at CHEP97)

P-Pro - Pentium Pro P-II - Pentium II

- DESY
  - HERMES — 10 dual P-Pro — Linux
  - ZEUS — 20 PCs
  - HERA-B 2/3LT — 100 P-II — Linux
  - HERA-B 4LT — 10(goal ~150) — Linux
  - ZEUTHEN — 40 PCs — Linux
- RHIC
  - Production — 40 dual P-II — Linux
- CERN
  - PCSF — 8 dual P-Pro, 33 dual P-II — NT
  - NA48 — 24 PCs — Linux
- KEK
  - Belle — Linux

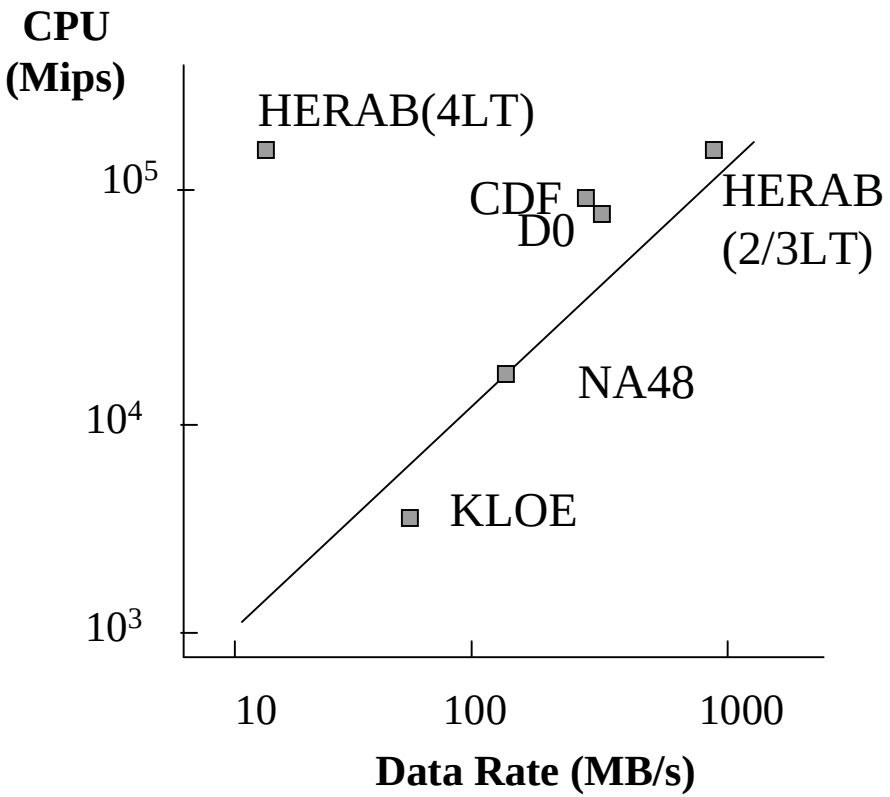
PC Computing - Farms

- Jefferson Lab
  - Production — 50 dual P-II — Linux
- RAL
  - Production — 11 dual P-II — NT
- Fermi Lab
  - E871 — 64 PCs — Linux
  - CDF/D0 — 18 dual P-II(now), ~500(by 2000) — Linux
  - CDF(L3) — PC farm
  - D0(L3) — 16 quad P-II — NT
- NASA Beowulf Project (1994)
  - ~25 farms — up to 126 nodes in each — Linux

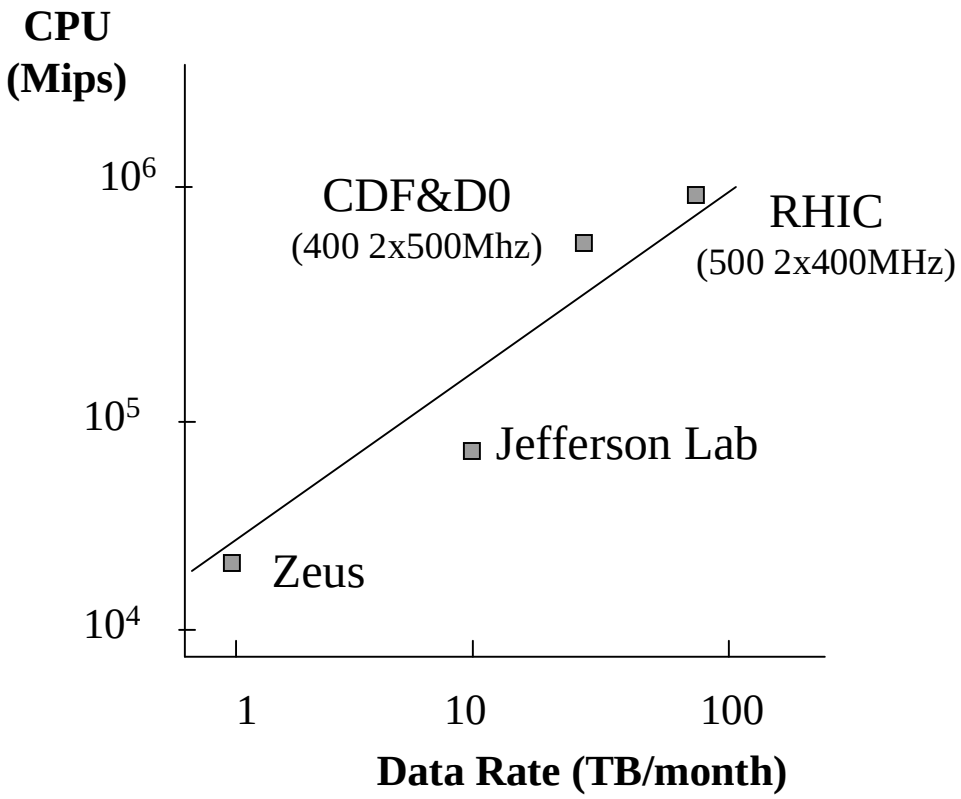
"Do-it-yourself Rocket Science"

Farms

Online



Reconstruction



Linux

- Most farms use Linux
  - ➢ low cost
  - ➢ widely used - “build on previous experience”
  - ➢ open source - “access to OS source code valuable in real-time systems”
  - ➢ software for off-the-shelf clustered PC hardware from Beowulf
  - ➢ “easy to port existing software”
- Performance Figures (CDF Run I data)

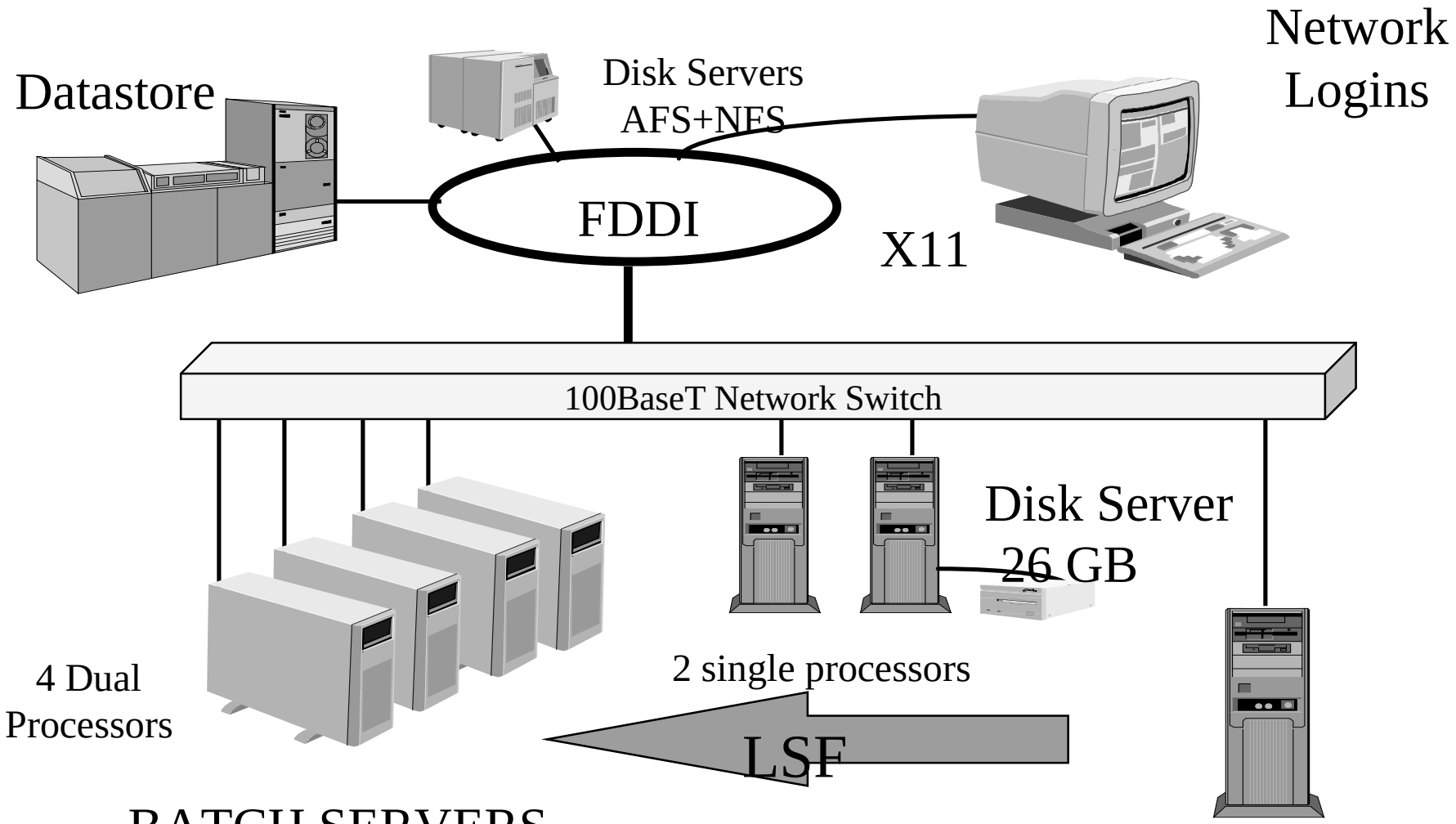
CPU/clock(MHz)	***CPU time(sec)***	***CPU ratio***
R4400/200	229	1
P5/166	272	0.85
P6/200	161	1.4

Dual P-Pro (SMP) gave results twice as fast as for a single processor i.e. performance equivalent to R10000 processor

Price/Performance ratio a ***factor of 3 better*** than for R10000 (SGI SMP)

NT

- For desktop, NT and Linux are both popular
  - e.g. RAL has 1400 PCs (1000 run NT)
- Disadvantages of NT
  - license costs for remote client (e.g. LSF)
  - cannot link mixed object code
  - no file-system links (make copies to working directory)
  - not UNIX
- Advantages of NT
  - NT has a large acceptance outside HEP (e.g. commercial enterprises based on NT) and therefore future looks more secure
  - technical software developed on NT, available on UNIX later
    - LSF, AFS, NAG library, Objectivity
  - not UNIX


Network
Logins
Datastore
Disk Servers
AFS+NFS
FDDI
X11
100BaseT Network Switch
Disk Server
26 GB
4 Dual
Processors
2 single processors
LSF
BATCH SERVERS
P200 NT4
NT Farm
FRONT END
NT 3.51 (Multi-user)

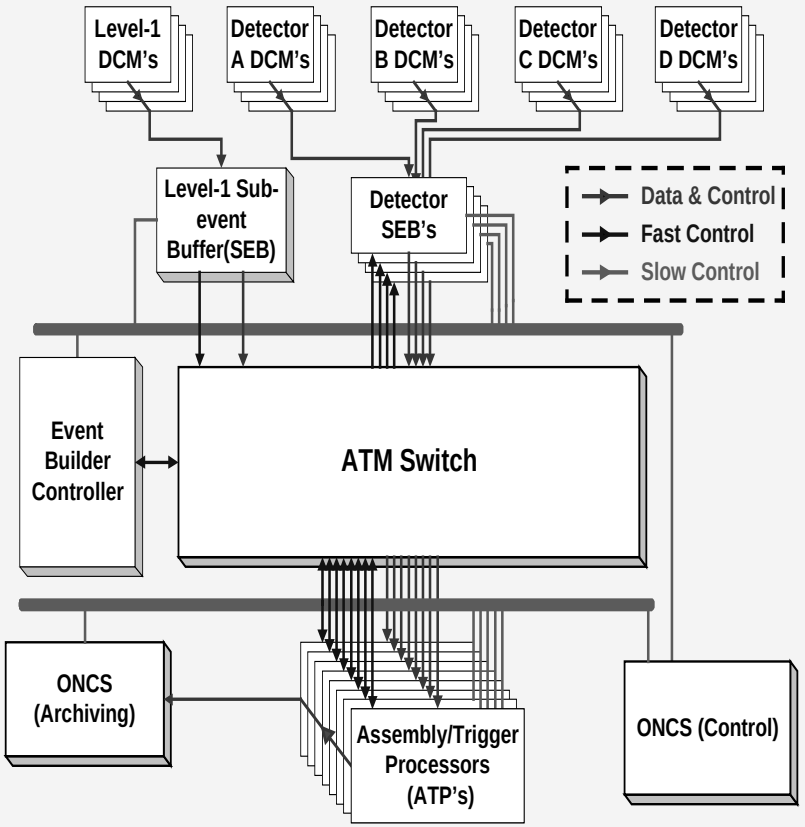
PC Computing - Conclusions

- Moving from UNIX farms to PC farms (in HEP and elsewhere)
- NT/Intel can deliver a good service (“but still waiting flood of users”)
- In ‘99 will see many more farms and with more nodes (100-1000)
- By CHEPY2K, PC computing will be main source of CPU, both on- and off-line.

DATA ACQUISITON

PHENIX - Event Builder Components

- Data Rate 200-200 MB/s
- Plan for x10 increase 2 GB/s
- Sub-event Buffer(SEB)
- Assembly/Trigger processor
  - Receives order from Controller to “pull” the event data from the relevant SEBs into its memory
- Controller
  - Coordinates activities of SEB and ATP via message-passing mechanism



PHENIX - Technical Choices

- Primary considerations
  - Performance requirements
  - Scalable
  - Commercial products
  - Clear upgrade path
- ATM satisfied these criteria
  - Switch-based architecture is widely used and scalable.
  - Available ATM switches can deliver bandwidth needed
  - Flow control is handled in the switch, lightening load on software developers!

- Use PCI-based processors Off-the shelf PCs (high performance, widely used)
- Running Windows-NT 4
  - All ATM hardware guaranteed to work on NT
- Full OO implementation of all aspects of system from data formats to messages

DAQ

- Many examples of solutions for parallel event building :
  - Euroball use Fibre Channel
  - CLEO III use Fast Ethernet
  - CDF use ATM
  - KLOE use FDDI
  - STAR use SCI

SOFTWARE TOPICS

Database Panel

- ODBMS (Objectivity) tried in ATLAS, BaBar, CMS, STAR...
- Disappointment at the impact of the Standards Body (ODMG)
  - hope was to reduce dependence on single vendor and to spur market
  - no-one adheres to it…will companies survive?
- Transient and persistent models of data
  - shield users from having to know how data are stored
  - allows evolution to different storage mechanisms
  - complicates the object model : converters, links, hash tables
- 70% work seems to be implementation dependent
  - schema management, data protection security, admin/monitoring tools
- Worries about scalability (>>10^9 objects) and about integration with mass storage system
- Performance OK and cost reasonable

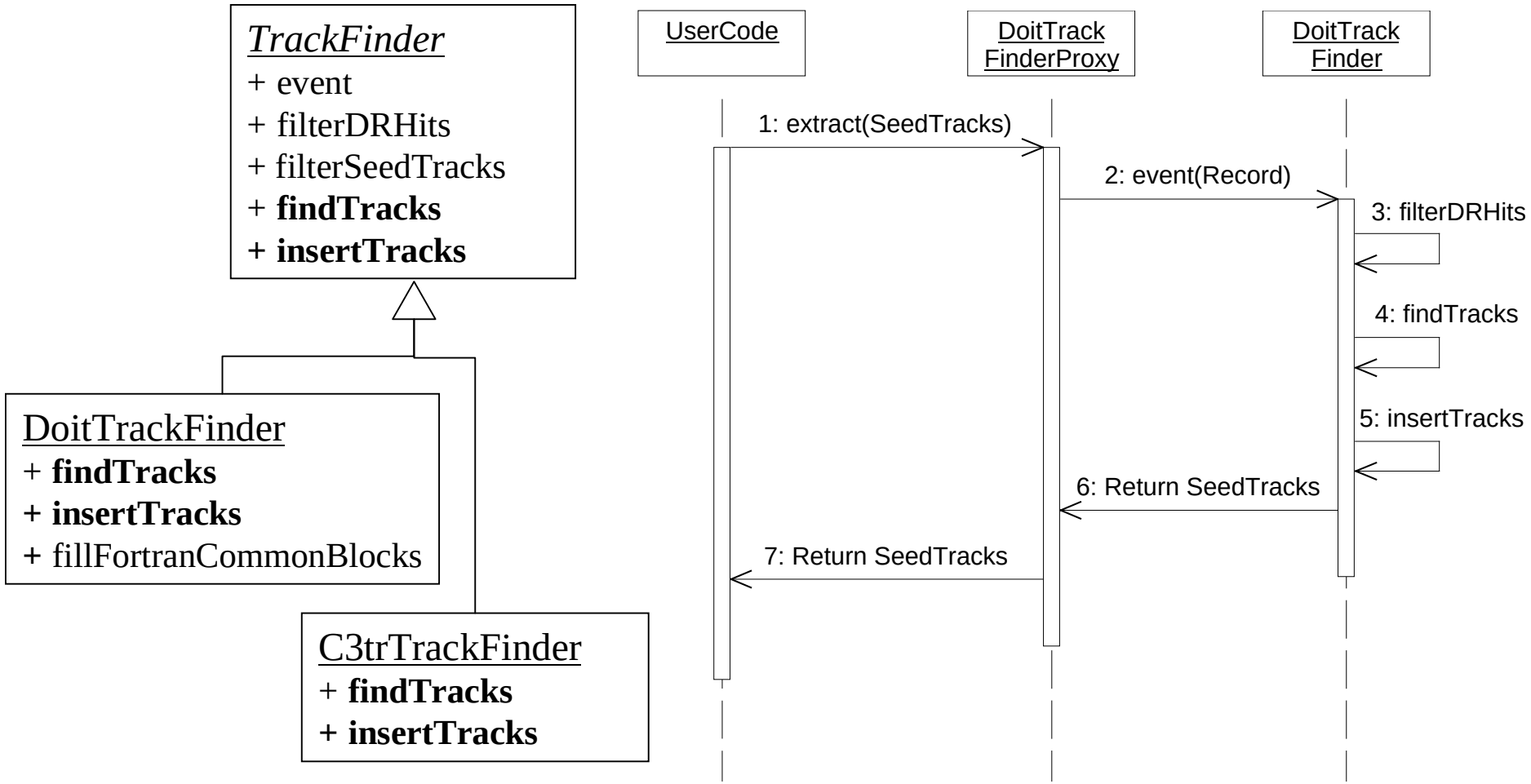
Database Panel

- CDF came to different conclusion
  - wanted to keep control of what is on disk
  - wanted to avoid problems due to queries having unforeseen effect
  - will use the ROOT I/O storage system (if support issue can be resolved)
- Trends
  - use of ODBMS (Objectivity) for :
    - "Conditions DB", "Calibration DB", Event Store, ...
  - BaBar believe took right approach and are "just about ready", but need performance improvements i.e. clustering, indexing and parallel iteration
  - significant use of ROOT as an alternative (CDF, D0, PHENIX)
  - mass storage - HPSS

Software Tools and Algorithms

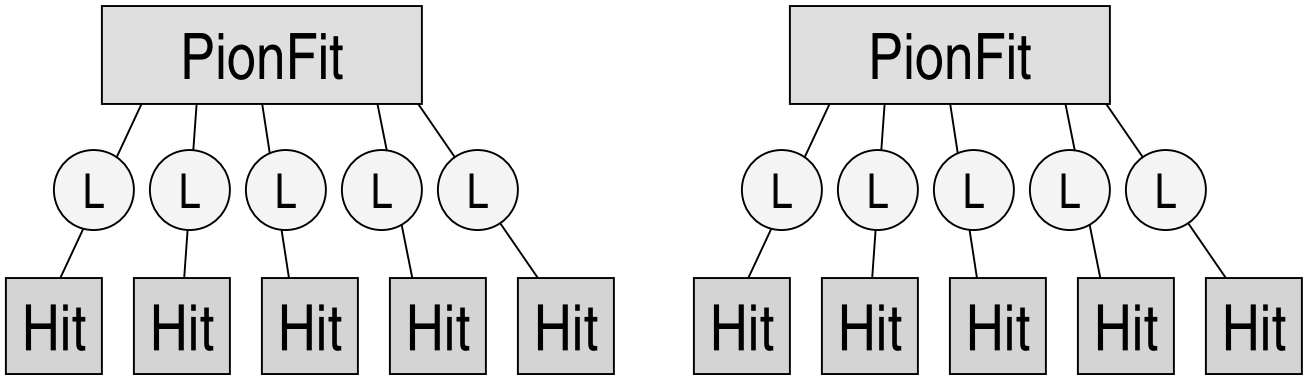
- OO programming in C++, CORBA, STL
- Importance of Analysis and Design stressed
- Importance of “packages” for linking, release management and documentation - part of the design
  - “Large Scale C++ Software Design, John Lakos”, Addison Wesley,’96
- Many examples of mature designs presented :
  - Track reconstruction for CDF's silicon tracking system
  - D0 object-oriented tracking software
  - The Tracking Infrastructure for CLEO III
  - BaBar's Object-Oriented Tracking System
  - TRF++: an object-oriented framework for finding tracks
  - Particle Identification Framework for the BaBar Experiment
  - An Object Oriented Design and Implementation of Vertex Finding for the D0 Detector

CLEO III - Track Finding

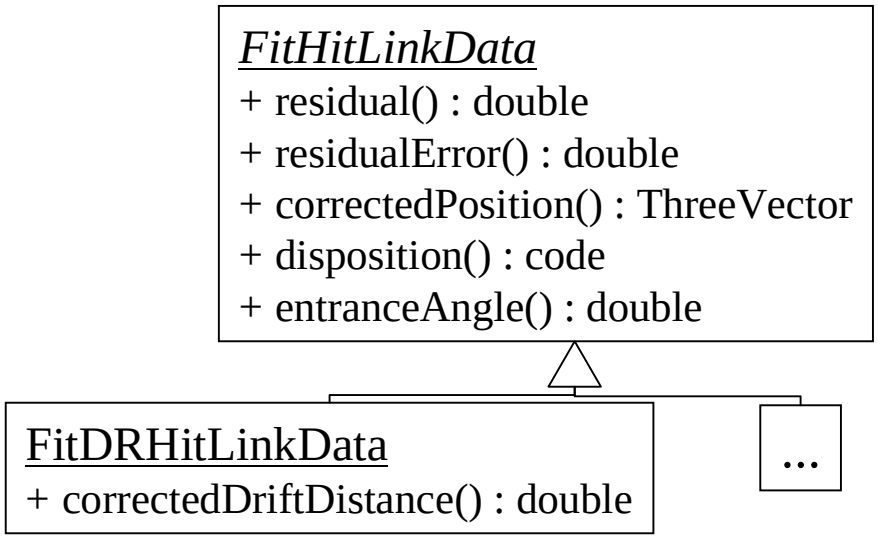


Design Patterns

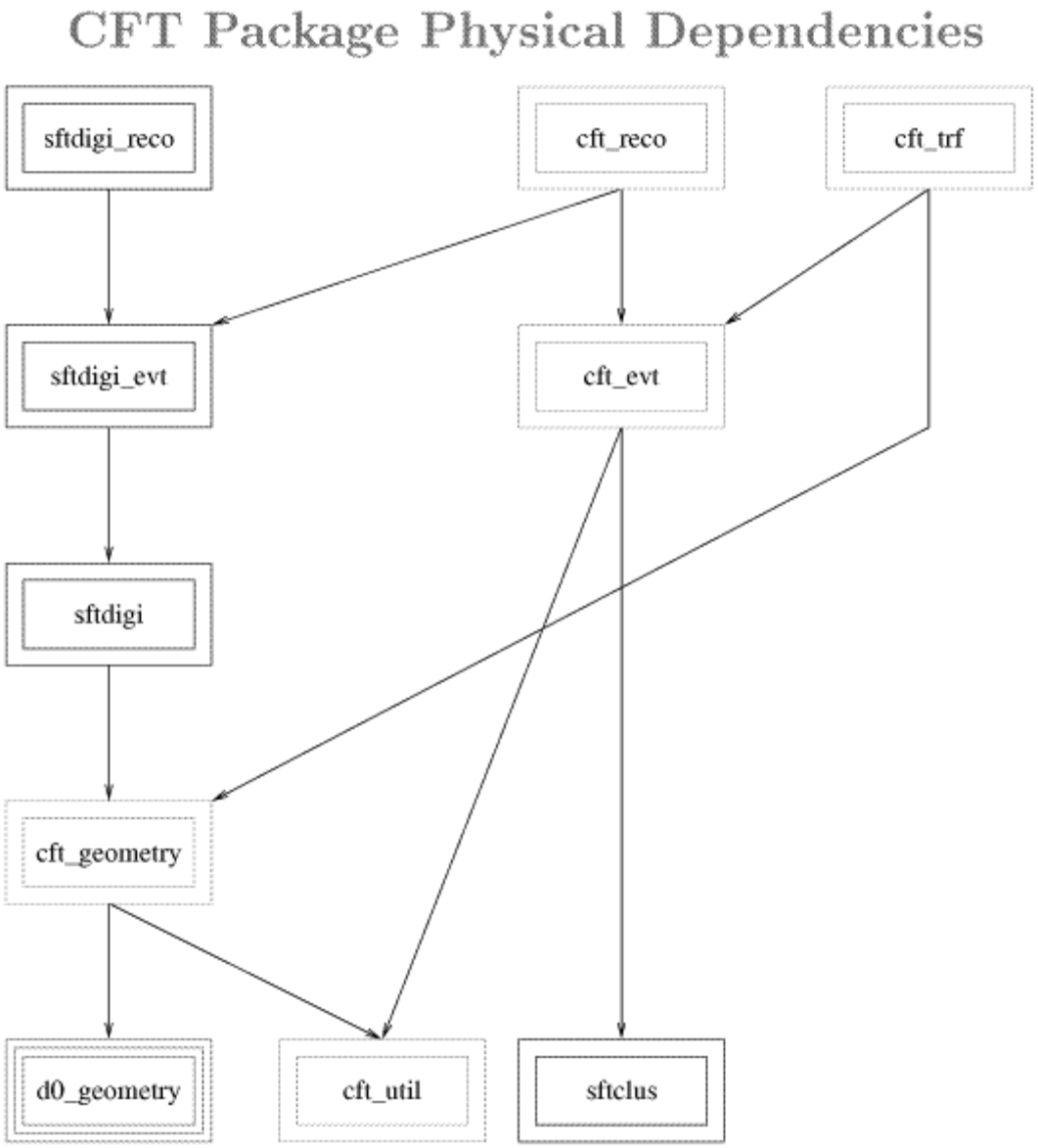
Fit - Hit Lattice (CLEO III)



- Hits are corrected for each mass hypothesis
- Link data natural place for information
- Uncorrected information still available



Packages


CFT Package Physical Dependencies
sftdigi_reco
cft_reco
cft_trf
sftdigi_evt
cft_evt
sftdigi
cft_geometry
d0_geometry
cft_util
sftclus

Analysis Tools

- ROOT widely used as a PAW replacement
  - designed to ease transition to C++
  - (ALICE, CDF, STAR, PHENIX, BaBar)
- Java based tools are close to being useful
  - Java Analysis Studio (Tony Johnson - SLAC)
  - Read and judge yourself (and then download and run)

    http://www.hep.net/chep98/paper98/221/chep98.ppt
- HEPExplorer tools from LHC++ not ready
- Factors limiting acceptance : commercial tools, non-open design
- “There will be no single PAW replacement”

Future CHEP Meetings

- CHEPY2K - Padova (Mazzucato) in Spring 2000
- CHEP'01- Beijing in Autumn 2001
- Future venues proposed Vienna and Lisbon