

Hunting Quarks with Linux

Don Holmgren
Fermi National Accelerator Laboratory
dholm@fnal.gov

O'Reilly Linux Conference
August 23, 1999

... and a cast of dozens!

Fermilab:

Jon Bakken, Andy Beretvas, Jim Fromm, Chih-Hao Huang, Robert Kennedy, Jim Patrick, Don Petravick, Ron Rechenmacher, Connie Sieh, G.P. Yeh, Dan Yocum

Massachusetts Institute of Technology:

Gerry Bauer, Christoph Paus, Paraskevas Sphicas, Steve Tether, Jeff Tseng

University of Rochester:

Benjamin Kilminster, Kevin McFarland, Kirsten Tollefson

Academica Sinica, Taiwan

Paoti Chang, Yen-Chu Chen, Ping Yeh

Outline

I. Introduction to Fermilab

II. Linux at Fermilab

III. Linux and the CDF Level-3 Trigger

1.

Introduction to Fermilab

High Energy Physics

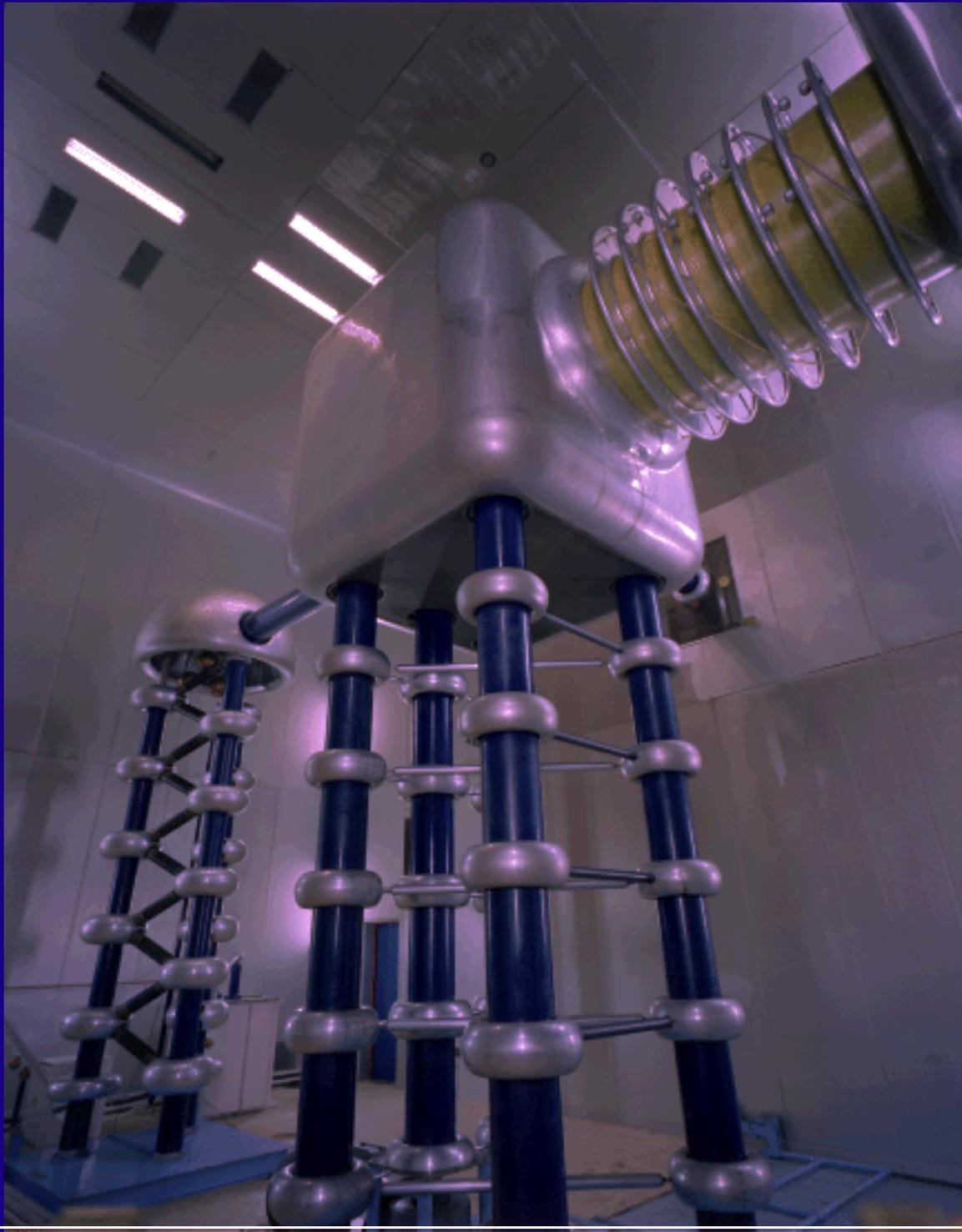
- The study of elementary (fundamental) particles and their interactions
- Experimental tool (microscope) is the particle accelerator

Fermilab Aerial View

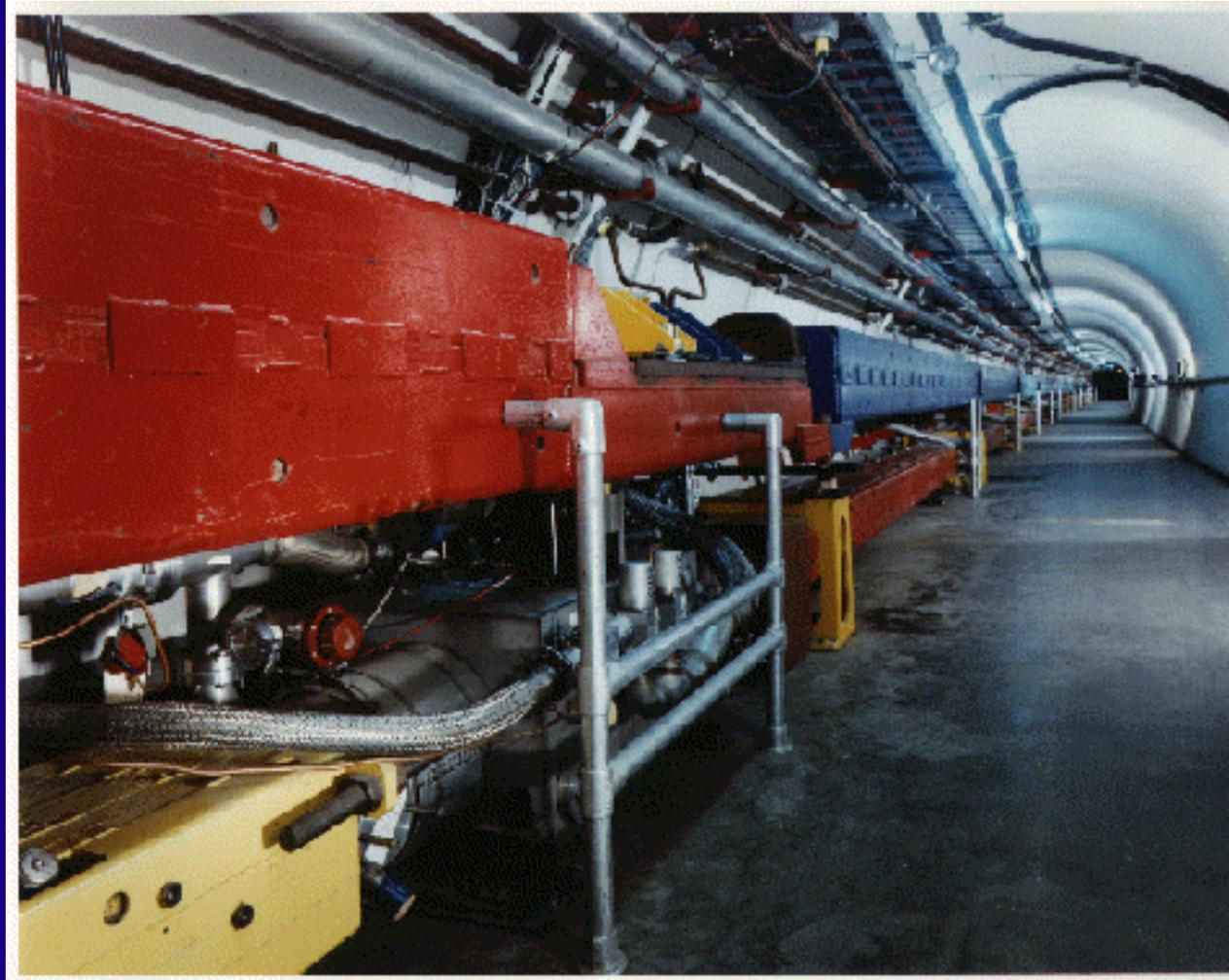


Wilson Hall

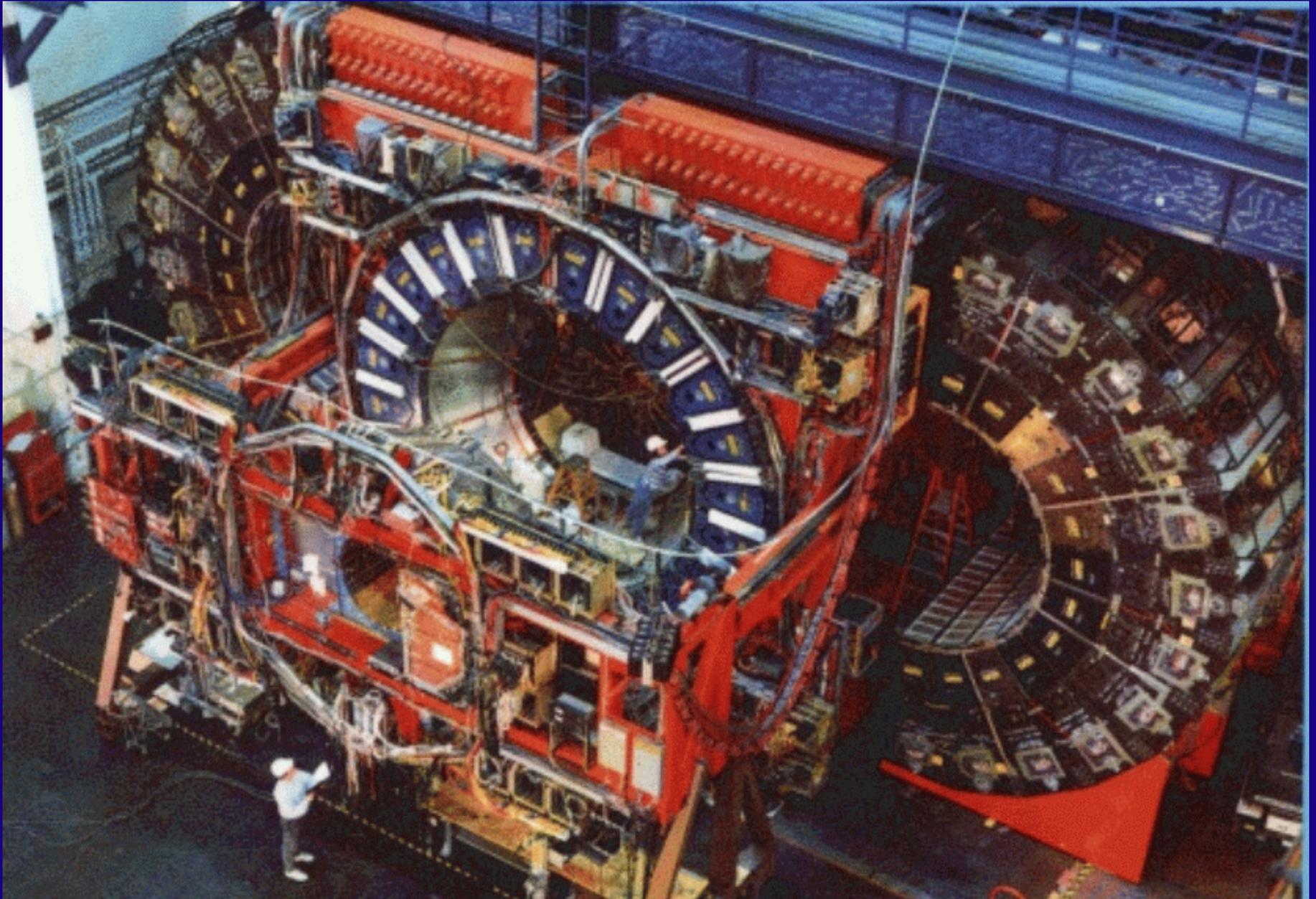




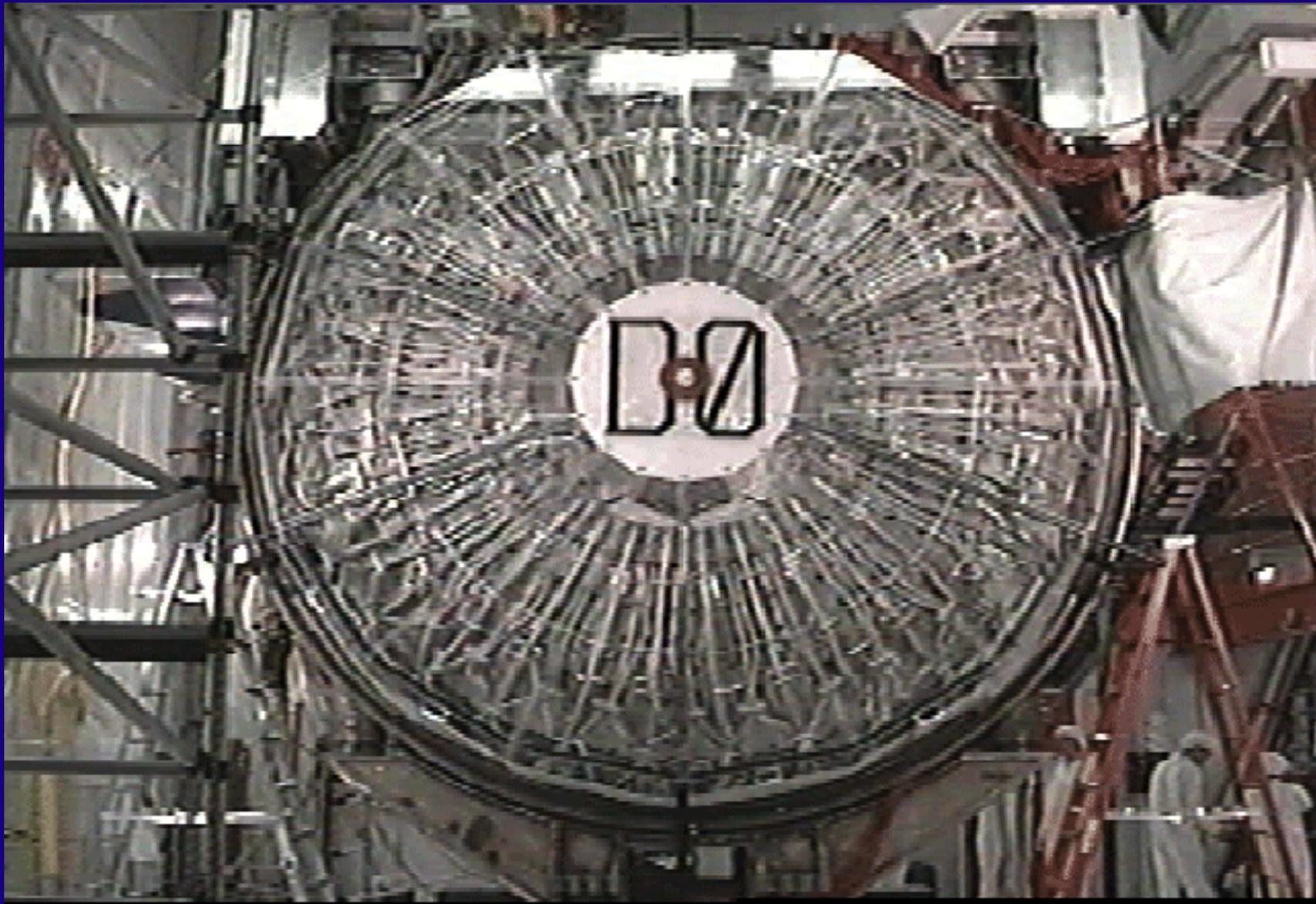
Accelerator Tunnel



CDF Detector



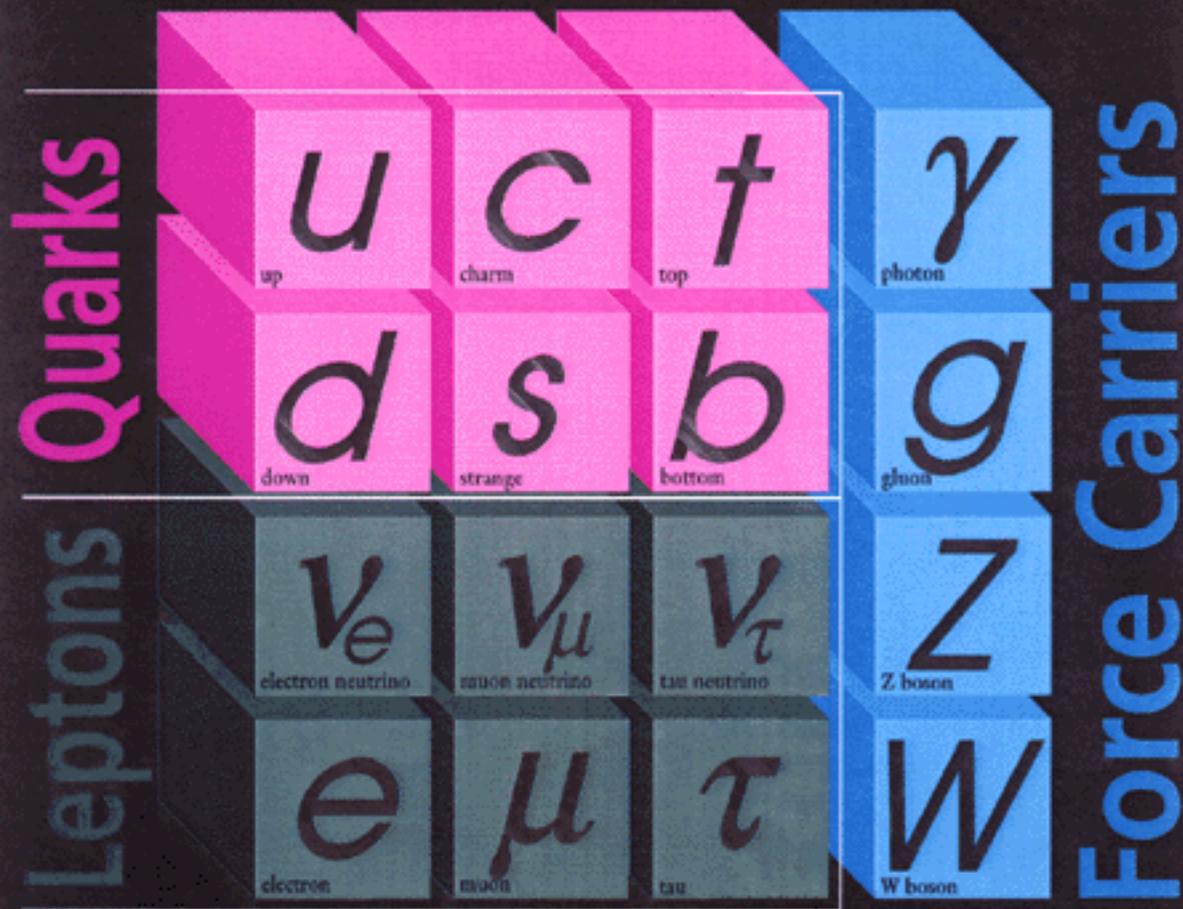
D0 Detector



Some Fermilab Numbers

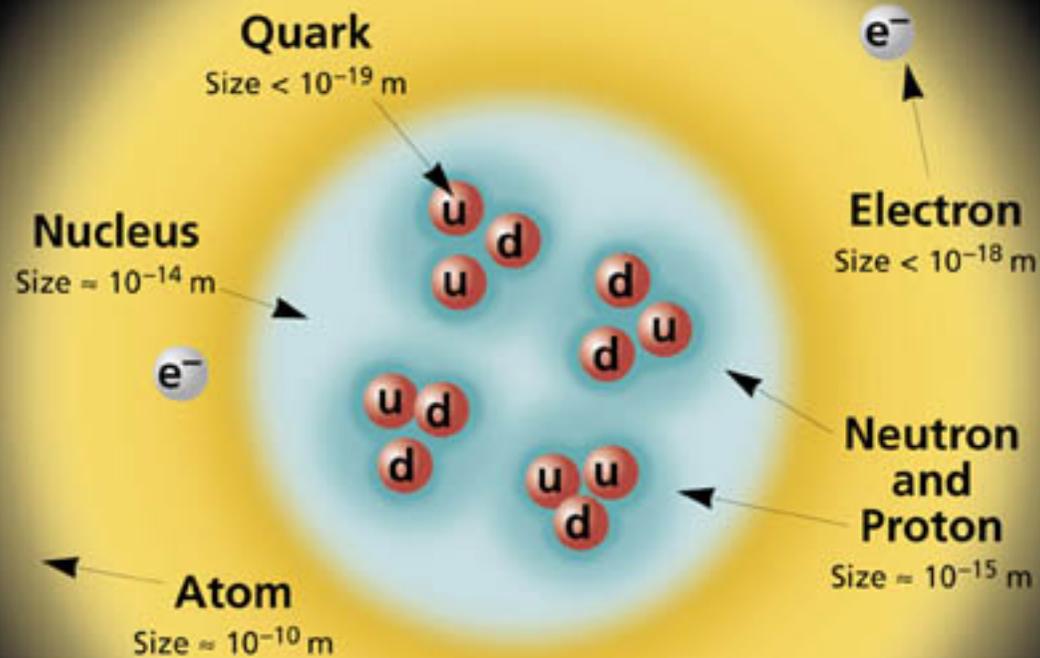
- Energy: Protons and Anti-Protons accelerate to 1 TeV (10^{12} electron-volts)
- Electricity cost: \$1,500,000 per month (we get a 50% discount!)
- Collisions/second: 1,000,000
- Raw data rate: 1 terabyte/second (terabyte = million megabytes)
- Data written to tape in Run II (planned): 2 Petabytes (petabyte = million gigabytes)
- Employees: 2000

ELEMENTARY PARTICLES



I II III
Three Generations of Matter

Structure within the Atom

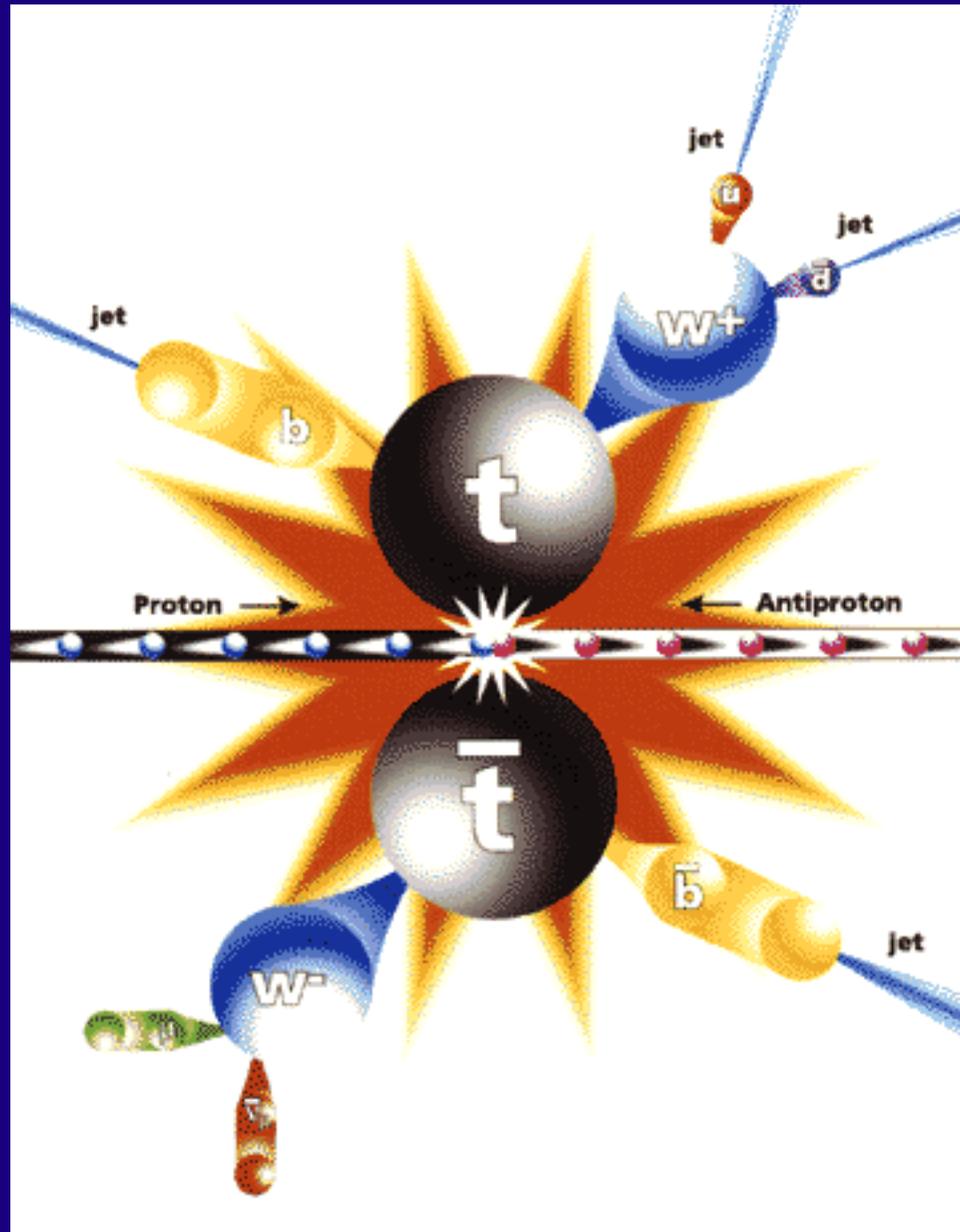


If the protons and neutrons in this picture were 10 cm across, then the quarks and electrons would be less than 0.1 mm in size and the entire atom would be about 10 km across.

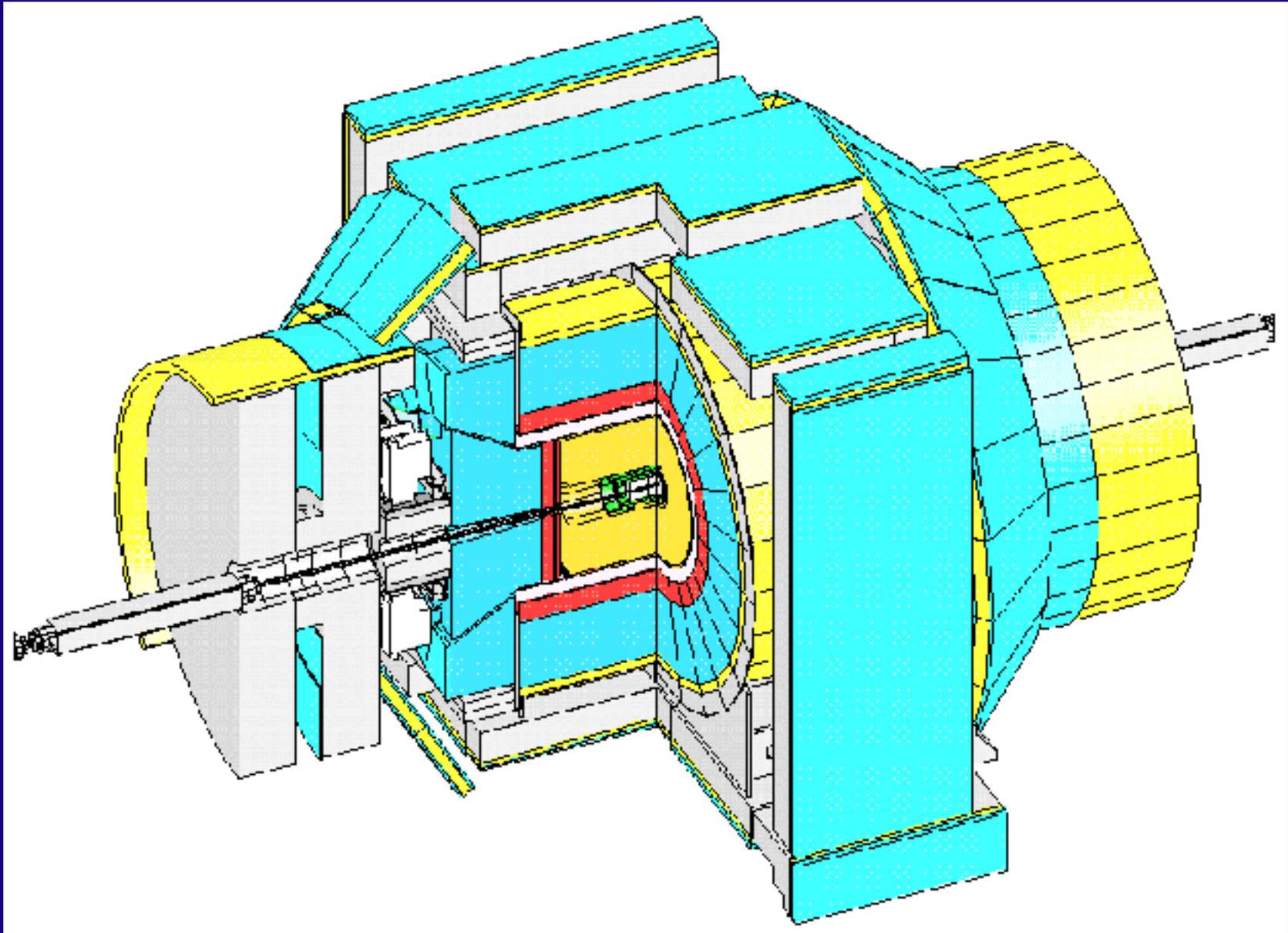
Jobs of a Detector

- Watch proton-antiproton collisions (“events”)
- Determine trajectories and energies of particles which emerge
- “Reconstruct” the events (translate 1 MByte of raw data into physics quantities)

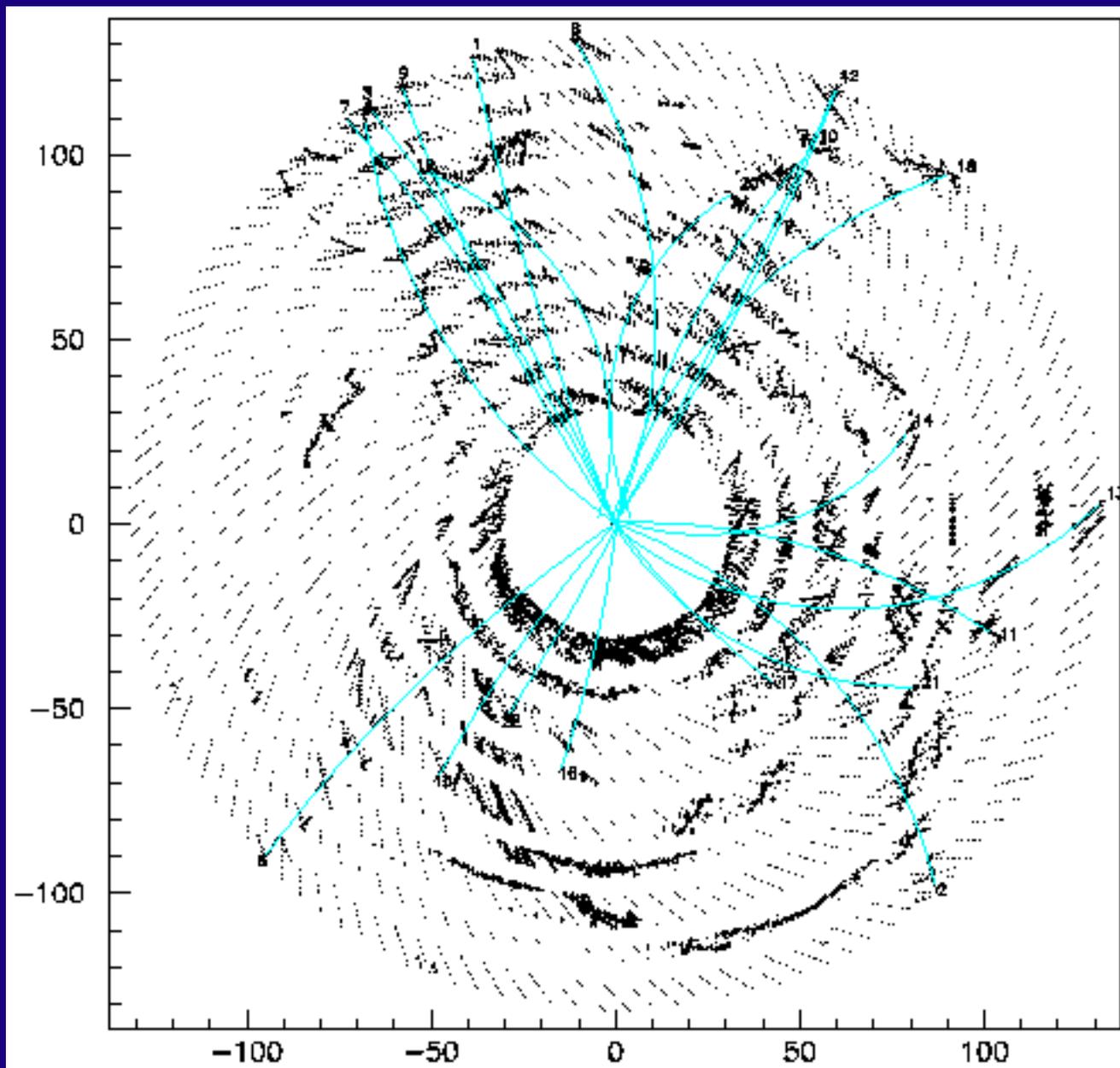
A Top-Quark Event



The CDF Detector



Track Reconstruction



e + 4 jet event

40758_44414

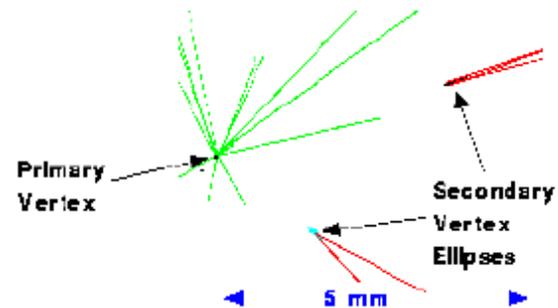
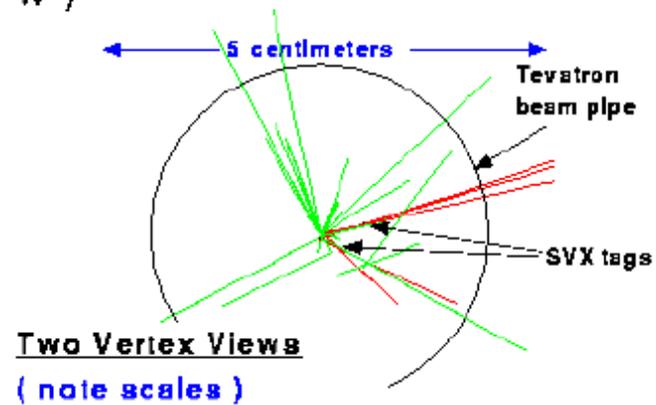
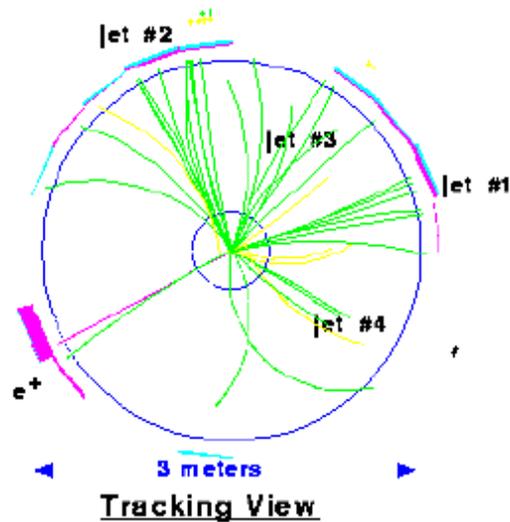
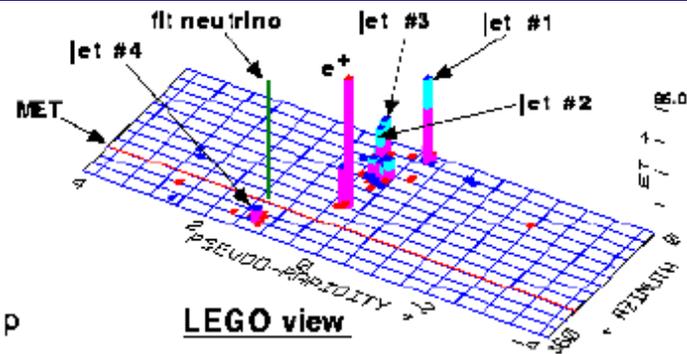
24-September, 1992

TWO jets tagged by SVX

fit top mass is 170 +/- 10 GeV

e^+ , Missing E_T , jet #4 from top

jets 1,2,3 from top (2&3 from W)



CDF Detector Numbers

- 750,000 channels of data
- Around a terabyte (10^{12}) bytes per second
- Track has 63,000 gold plated wires, each about 10 feet long
- On average, about 100 particles emerge per collision, recorded for about 1 microsecond
- Storage costs - 0.2 cents/megabyte for tape - constrain amount of data we'll record
- Initial costs of CDF detector: \$100 Million
- In Run I, about 2 dozen "Top" events out of 100 million events stored
- 1 "Top" event per trillion collisions

II.

Linux at Fermilab

Roles of Computers

- Control of detectors
- Data acquisition
- Data storage
- Filter and sort events (online reconstruction)
- Calibrate, filter, and categorize events (offline reconstruction)
- Analyze events

Clusters at Fermilab

- Events can be analyzed independently (“embarrassingly parallel”)
- So, use clusters (“farms”)
- Also, in past years, a message passing API (**CPS**, like PVM, MPI)
- See Gregory Pfister’s “In Search of Clusters”
- FNAL has operated up to 400 UNIX workstations in a cluster



Linux - “How did *that* sneak in?”

- 1994: Linux Kernel 1.0
- 1995: Sloan Digital Sky Survey (www.sdss.org)
 - Gateway from DOS to Survey CVS code repository
 - Used for camera firmware (Forth) and telescope control codes (Pascal)
- 1996:
 - Proposal, implementation of PCFARM prototype cluster
- 1997:
 - Proposal, implementation of CDF Level 3 prototype filter
- 1998:
 - January - Computing Division announces Linux support
 - First Linux production farm purchased
 - CDF Level 3 filter decision accepts Linux, commodity hardware
- 1999:
 - First Run II PC purchases (150 dual processor systems)

Why Linux at Fermilab?

- Use of commodity hardware is the largest victory
 - Price/performance for HEP passed workstations in '95
 - UNIX is the conservative approach
 - Non-UNIX (Windows NT) is the radical approach
- The easiest of the PC UNIX candidates (Solaris, FreeBSD, Linux)
 - Widest hardware support
 - Most active development
 - Most widely adopted by HEP community
- Source codes were key to our systems development work
 - Instrumentation, device drivers, debugging

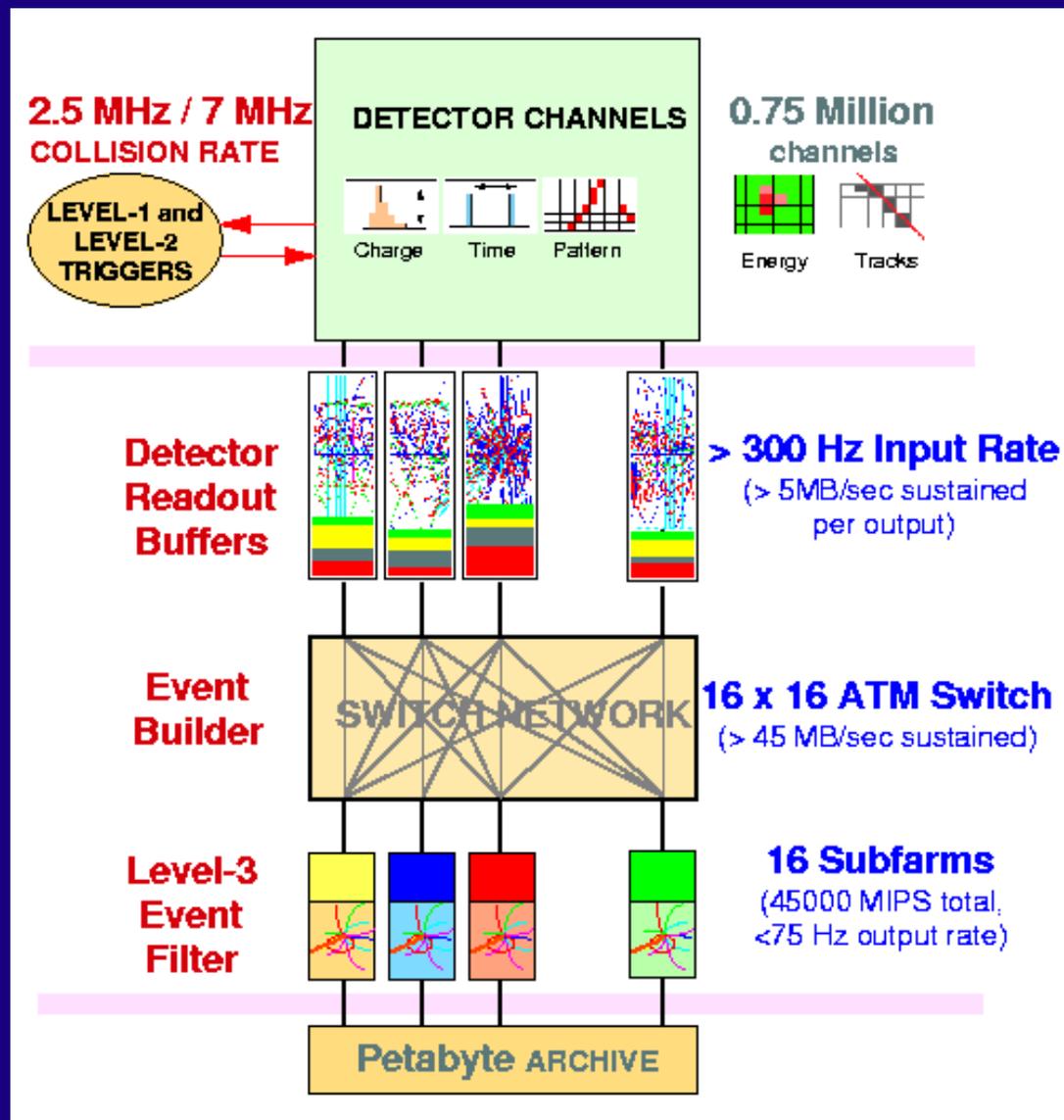
FNAL Linux Support Policy

- FNAL Version of Linux:
 - based on RedHat + updates + patches + Fermi software
 - new releases about every 6 months
 - site installs via network
 - CD-ROM's for laptops, home, collaborators
 - some vendors will pre-install Fermi Linux
 - Intel-Linux *only*
 - security patches from FNAL repository (via **Autorp**m) required nightly

III.

Linux and the CDF Level-3 Trigger

Data Flow Schematic



CDF Run II Specifications for Level-3

Specification	Requirement
Level-2 Peak Trigger Rate	300 - 1000 Hz
Mean Event Size	200 - 250 KB
Maximum Level-3 Output Rate	15 MB/s (75 Hz)
CPU	45,000 MIPS
# 500-MHz Dual-PII Nodes	96

- Output rate is fixed by mass storage budget
- ATM Network switch has 16 OC-3 switch outputs
- Optimistic numbers!

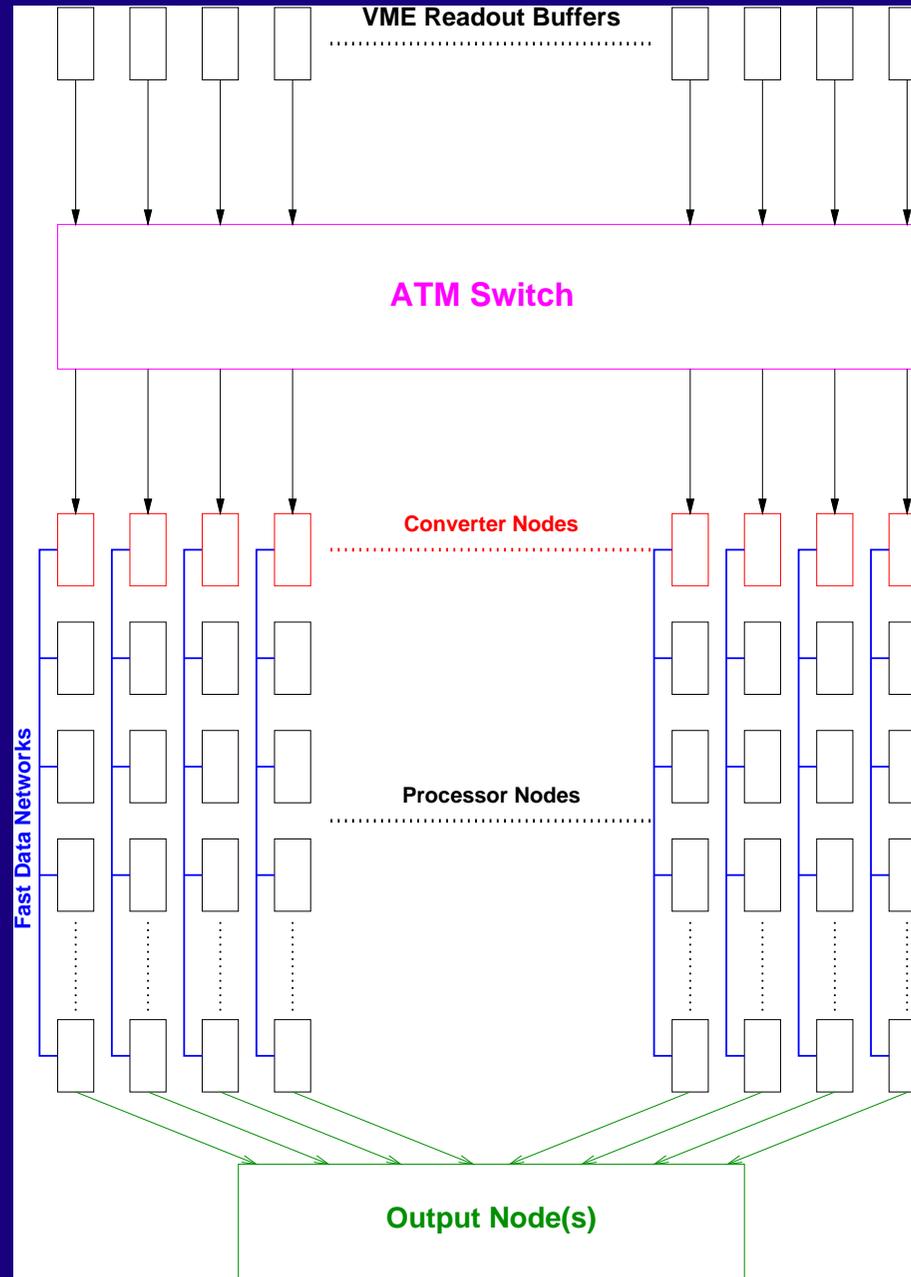
Architecture

- Ideal implementation should:
 - scale
 - use commodity components
- Run I approach:
 - large Silicon Graphics SMP systems in single layer
 - each computer receives, filters, outputs
 - factor of 20 less CPU required
- Not enough CPU in PC's for single layer approach (16 nodes)

“Tiny” MIPS

Processor/Speed	MIPS
Alpha 500 MHz, Linux	322
<i>Pentium III 550 MHz (estimate)</i>	<i>310</i>
Alpha 440 MHz, OSF1	281
<i>Pentium II 500 MHz</i>	<i>280</i>
<i>Pentium II 450 MHz</i>	<i>252</i>
SGI R10000 O200/O2000 196 MHz	225
<i>Pentium II 400 MHz</i>	<i>223</i>
<i>Pentium II 350 MHz</i>	<i>197</i>
SGI R10000 Challenge 190 MHz	192
<i>Pentium II 333 MHz</i>	<i>188</i>
RS6000/43P, 200 MHz, AIX	177
<i>Pentium II 300 MHz</i>	<i>168</i>
<i>Pentium II 266 MHz</i>	<i>147</i>
<i>Pentium Pro 200 MHz</i>	<i>113</i>
<i>Pentium 166 MHz</i>	<i>82</i>

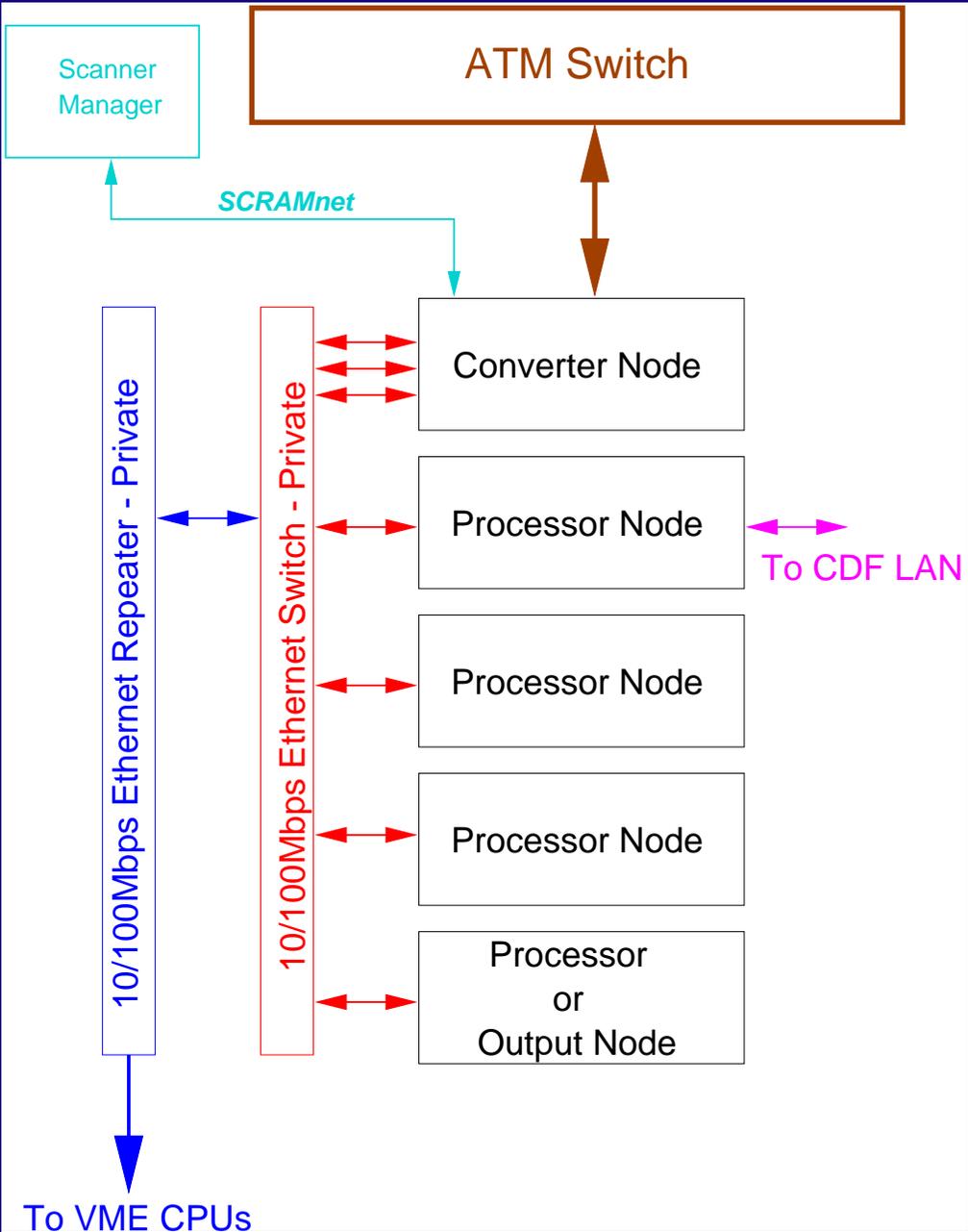
Run II Level-3 Architecture



Architecture

- distribute CPU load to multiple PC's
 - “converter” nodes get event fragments from ATM
 - full events sent to “processor” nodes
 - use multiple fast ethernets to keep up with OC-3 (16.5 MB/sec)
- separate I/O and filter jobs
- to scale CPU, add more processors
- to scale I/O:
 - increase ATM switch width, or
 - increase ATM rate (OC-3 to OC-12)

Prototype - Fall, 1997





Level 3 Prototype Hardware

- Converter nodes:
 - Initial: 200 MHz Pentium Pro, “FX” Chipset
 - Final: 350 MHz Pentium II, “BX” Chipset
 - ForeRunnerLE ATM Interface
 - Adaptec quad-“tulip” ethernet
- Processor nodes:
 - Dual 200 MHz Pentium Pro
 - “Tulip” fast ethernet

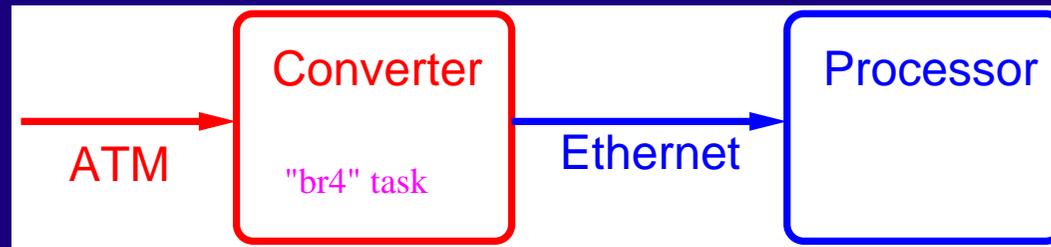
ATM and Linux

- ATM support (“pre-alpha”, v0.31) from EPFL (Werner Almesberger)
 - well written, documented, and supported
 - AAL0 and AAL5 very easy to get up and running
- NIC’s evaluated:
 - SMC ATM Power155
 - IDT evaluation board (“Nicstar” chipset)
 - FORE Systems ForeRunner 200E
 - FORE Systems ForeRunnerLE - also based upon “Nicstar”
- Performance:
 - All cards delivered wire speed at minimal CPU utilization
 - Restriction on VPI common - only VPI = 0 (sometimes also 1)
 - Nicstar cards allow large VPI values

ForeRunnerLE Device Driver

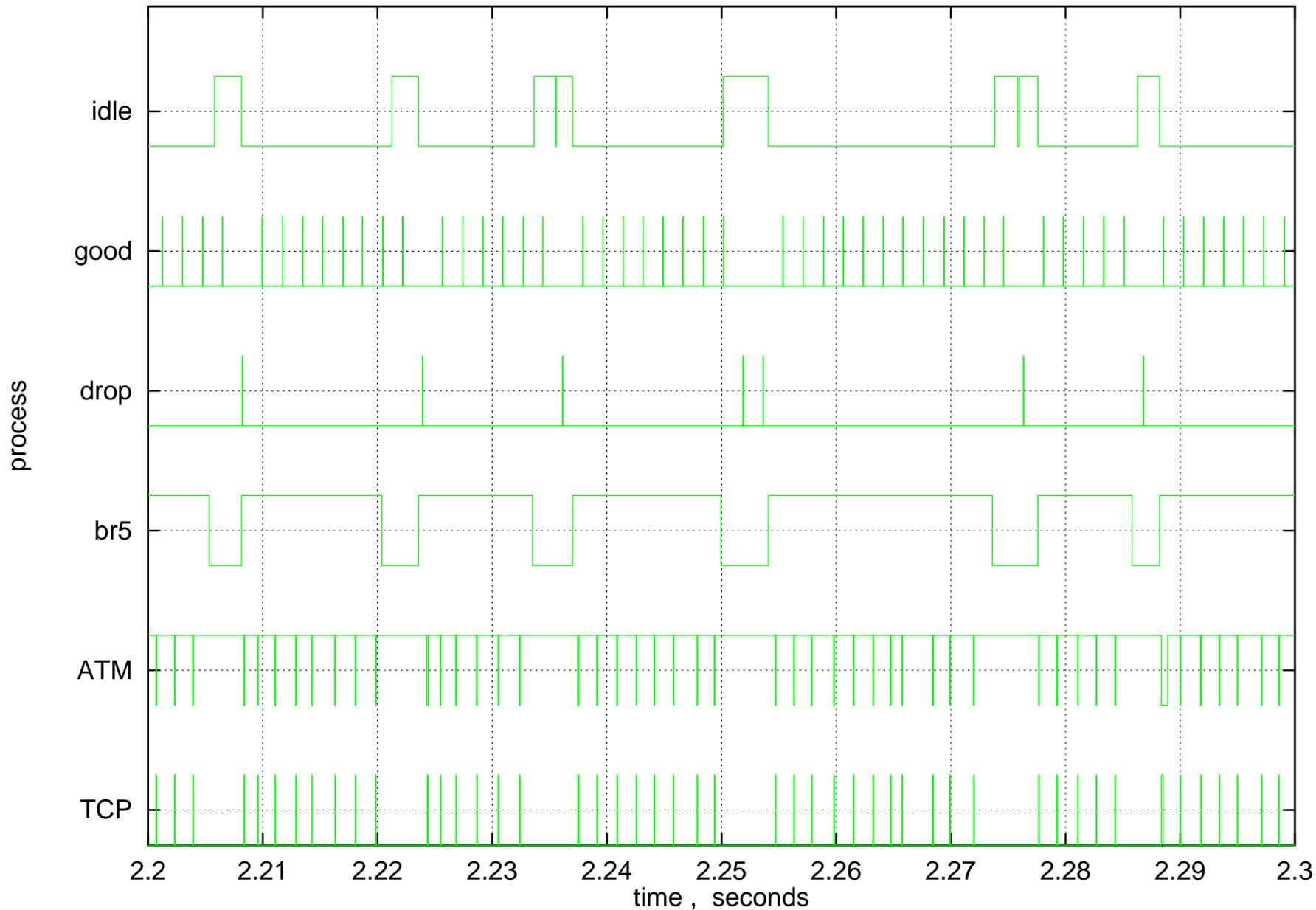
- EPFL ATM included IDT driver (*Nicstar*, from Matt Welsh)
IDT engineering card, *not* ForeRunnerLE, supported
ForeRunnerLE also *Nicstar* based
- Choice dictated by VPI/VCI restrictions:
VPI designates Level-3 destination, Level-2 source
VCI designates buffer
- Modifications to IDT driver:
Hidden “latch” to enable DMA on ForeRunnerLE
Debugging
VPI / VCI configuration
need_resched

“Toy” Level 3 - Version 1



- Open loop
 - ATM free running
 - Use rate division and/or interpacket delay to adjust rate
- Observed Problems
 - Dropped AAL5 packets upon startup
 - Sensitivity to other tasks (packet drops)

AAL5 Packet Drops



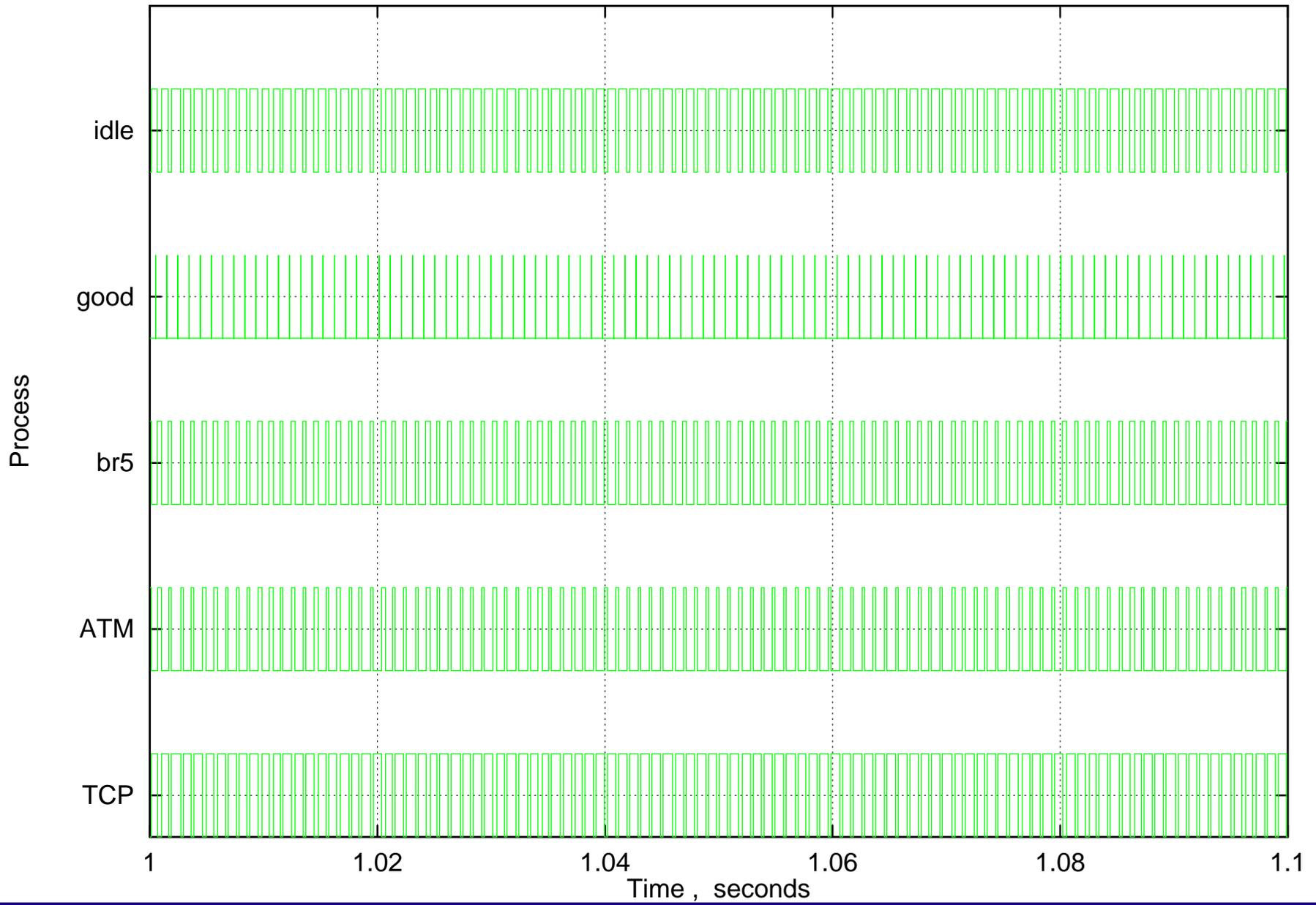
TRACE Facility

- Circular buffer in kernel containing traces:
 - Time stamped using Pentium cycle clock
 - Kernel can write traces
 - User-space can write traces
 - Minimal system load
 - Maskable per process/kernel/class
- Analysis package
 - Octave (*Matlab* clone) based
 - Oscilloscope-style time plots of system activity
 - Statistical analysis

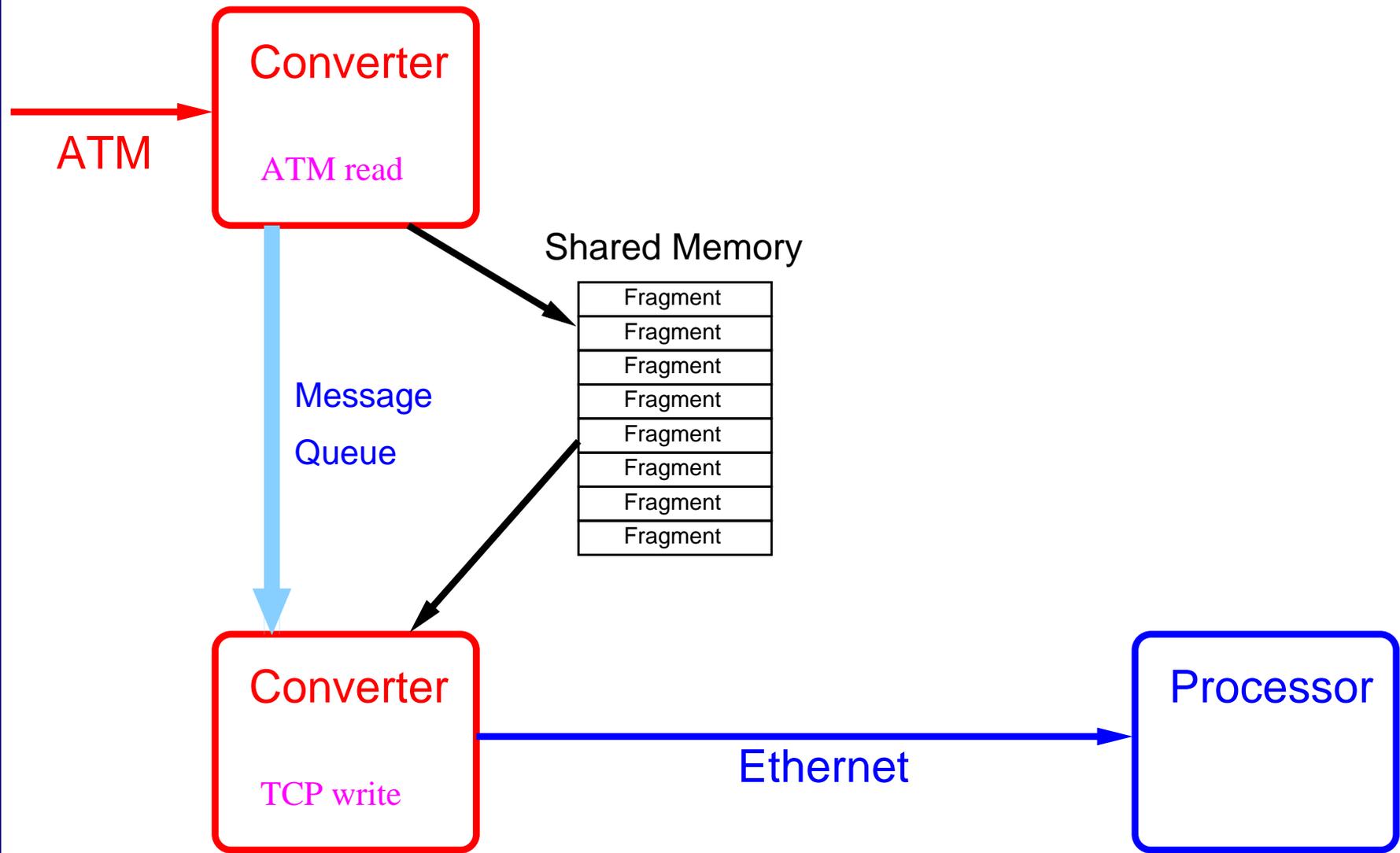
“Toy” Level 3 - Problem Solutions

- Disable Jacobson “Slow Start” in TCP/IP stack
- Enlarge socket buffers
- Use real time queues (*sched_setscheduler()* system call)
- Bump kernel global *need_resched* in ATM driver

No AAL5 Drops with Real Time Queue



Toy Level 3 - Version 2



“Toy” Level 3 - Version 2

- Better simulation of Level-3 software framework
Shared memory, message queues
- Observed problems:
Couldn't increase rate beyond 12 MB/sec
But, on different brand of PC saw 14.5 MB/sec!

I/O Performance of PC Hardware

- We thought of these possible bottlenecks:
 - CPU utilization
 - PCI bus bandwidth
 - Memory bandwidth
- Memory bandwidth turned out to be the real limit
 - PCI (*VMETRO*) bus analyzer critical to debug
 - Saw CPU utilization increase, PCI efficiency decrease at near rate limit
- The fix: use faster memory

Memory Bandwidth - Why So Critical?

- Many copies from ATM to Ethernet:
ATM interface DMA to kernel buffer (ATM *skbuff*)
skbuff to user-space buffer
user-space buffer to kernel buffer (Ethernet *skbuff*)
skbuff to multiple *skbuffs* (fragmenting)
skbuffs DMA to Ethernet interface
Total of 8 moves on memory bus
- 15 MBps rate requires 120 MBps memory bandwidth!
- SDRAM arrived just in time

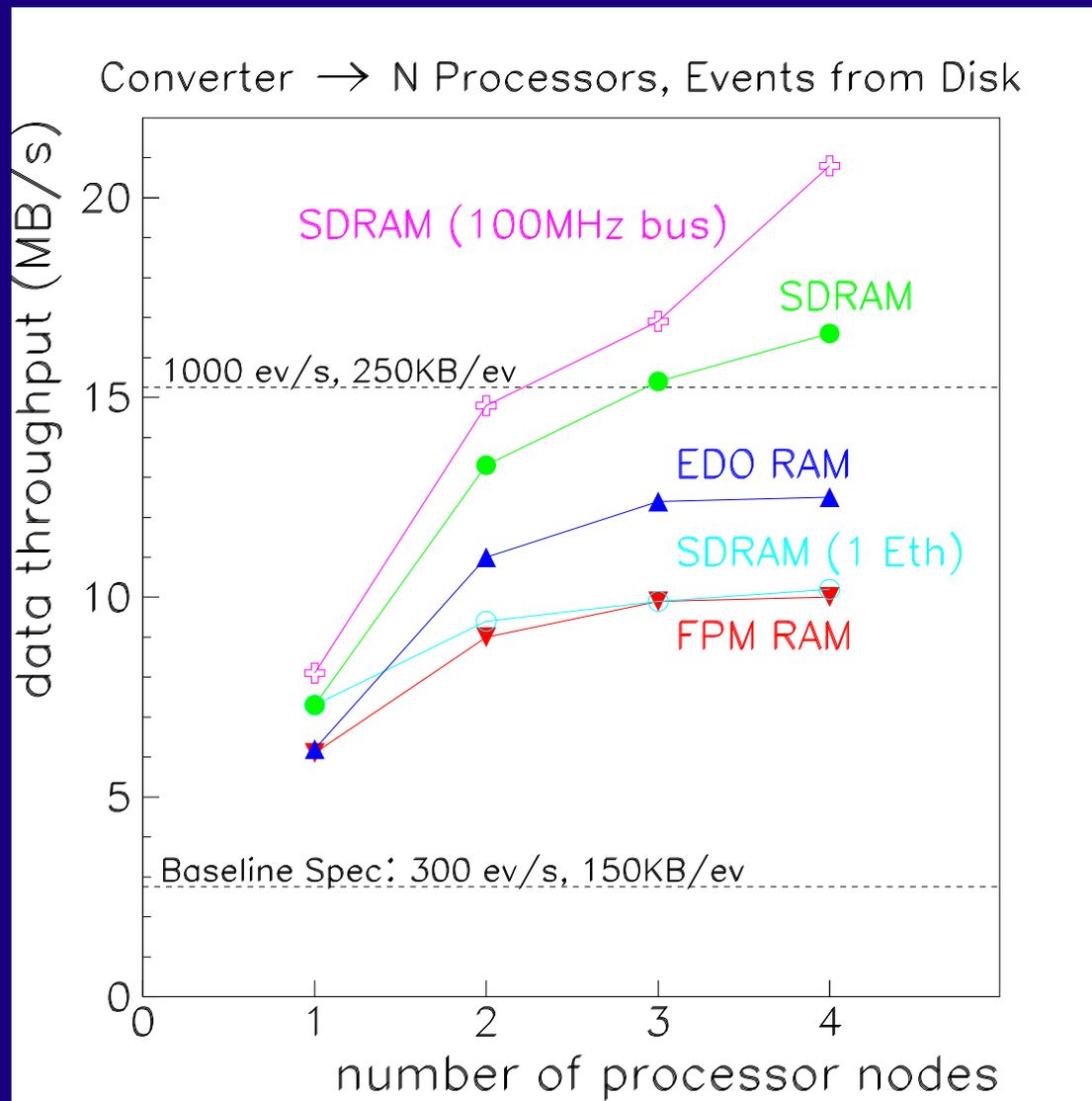
Memory Bandwidth (*STREAMS* Benchmark)

- Memory evolution (moving **doubles** to **doubles**):
 - FPM (Pentium Pro): 80 MByte/sec
 - EDO (Pentium Pro): 110 MByte/sec
 - SDRAM (Pentium II “LX”): 170 MByte/sec
 - SDRAM (Pentium II “BX”): 270 MByte/sec

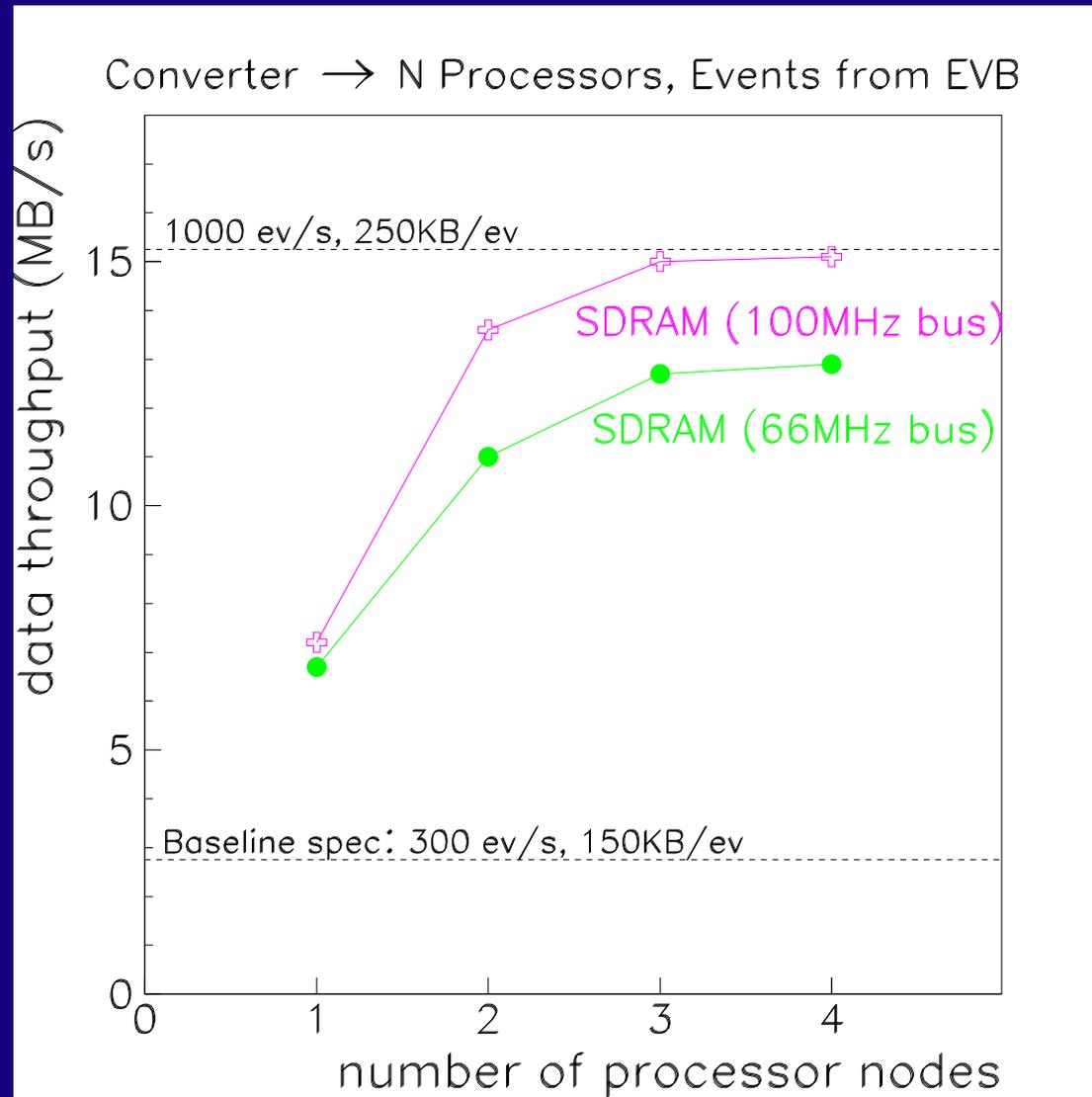
“Toy” Level 3 - Performance

- 12 MBps with FPM memory
- 14.5 MBps with EDO memory
- 16.5 MBps with SDRAM (66 or 100 MHz memory bus)

Throughput of Real Level 3 Codes - Disk Source



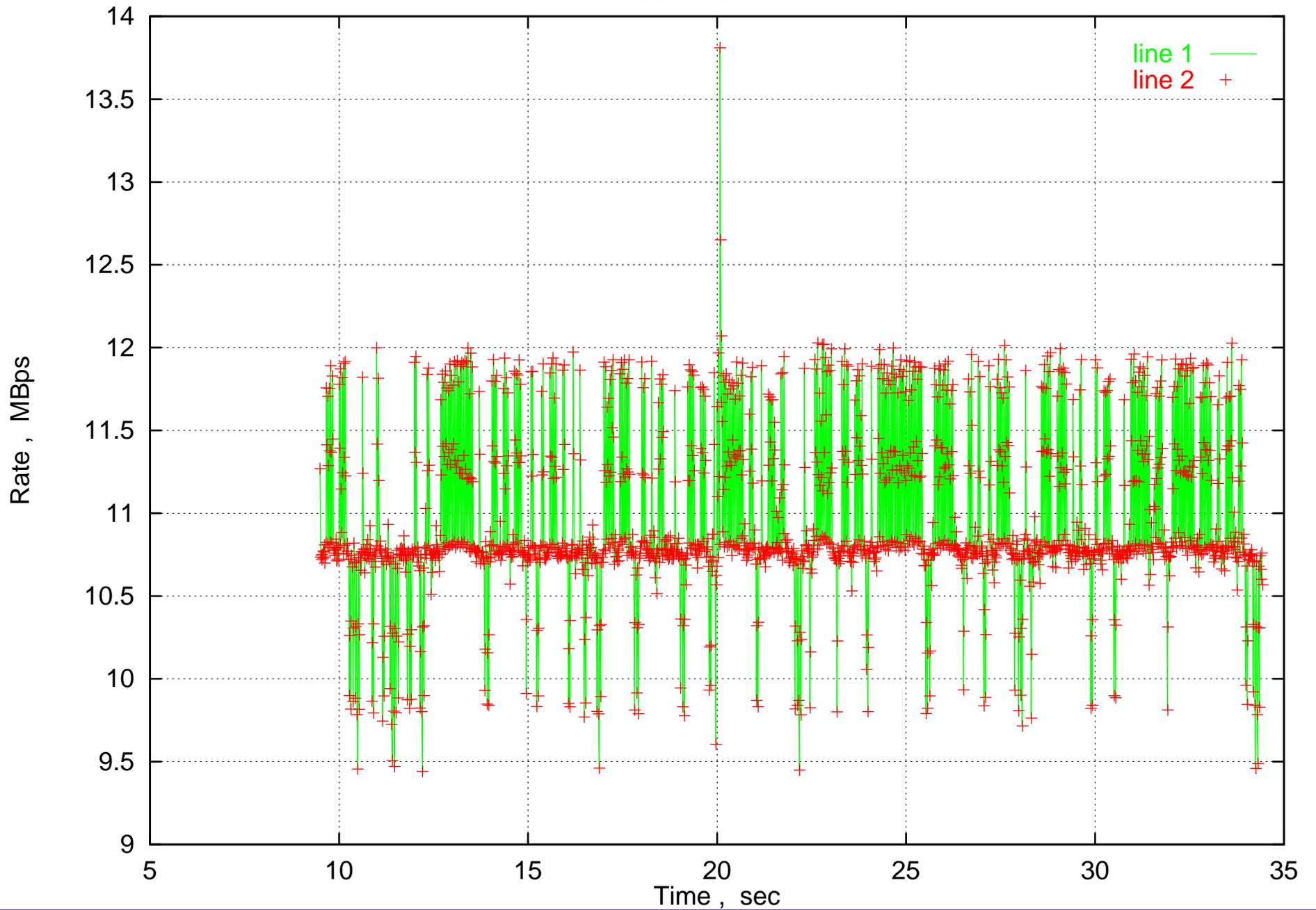
Throughput of Full Level-3 Framework - ATM Source



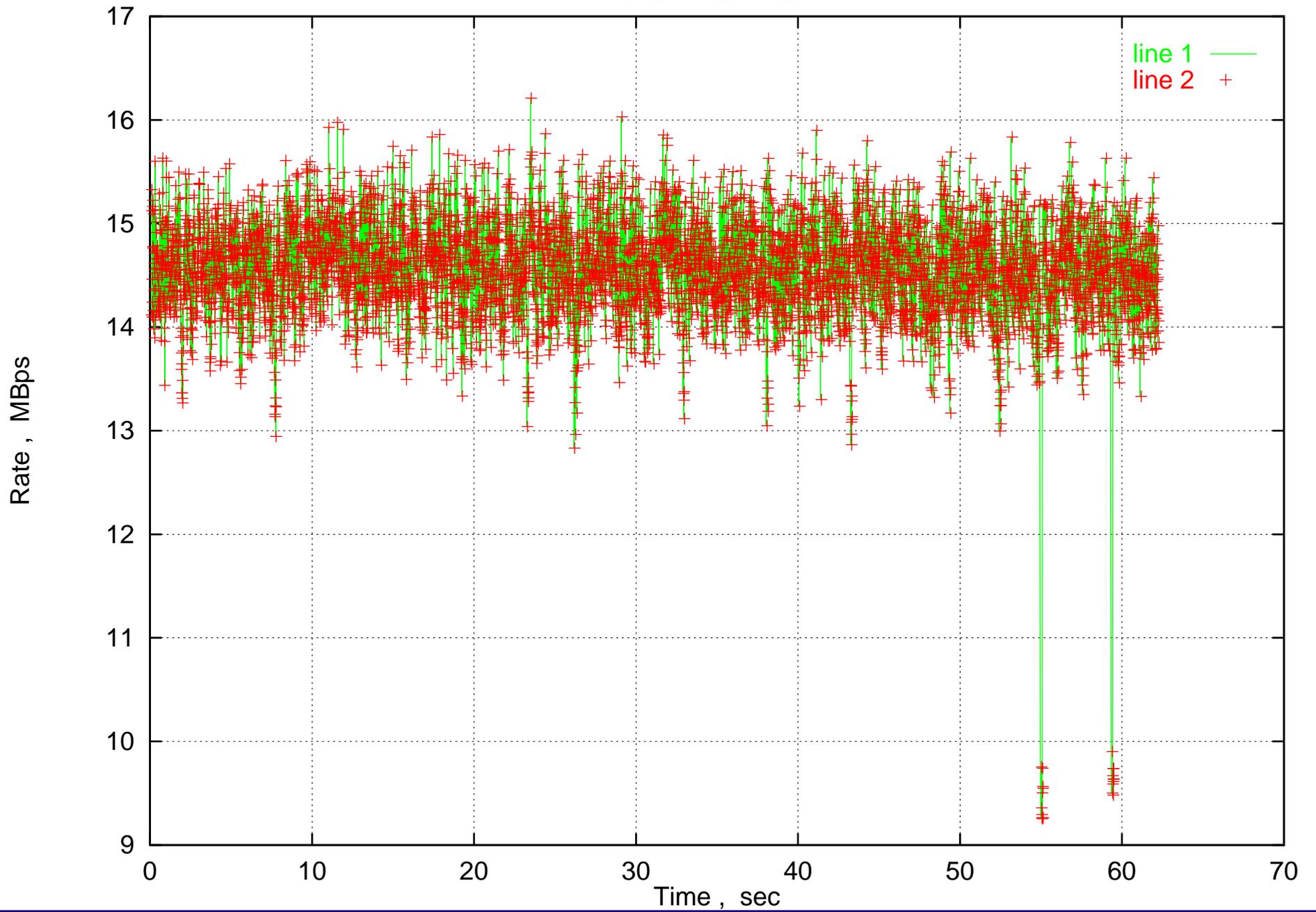
Time Slices

- Level-3 software framework does a lot of polling
- In UNIX, minimum *nanosleep()* is a time slice
- Better results when converter nodes use 1msec instead of 10msec time slice

10 msec time slices



1 msec time slice

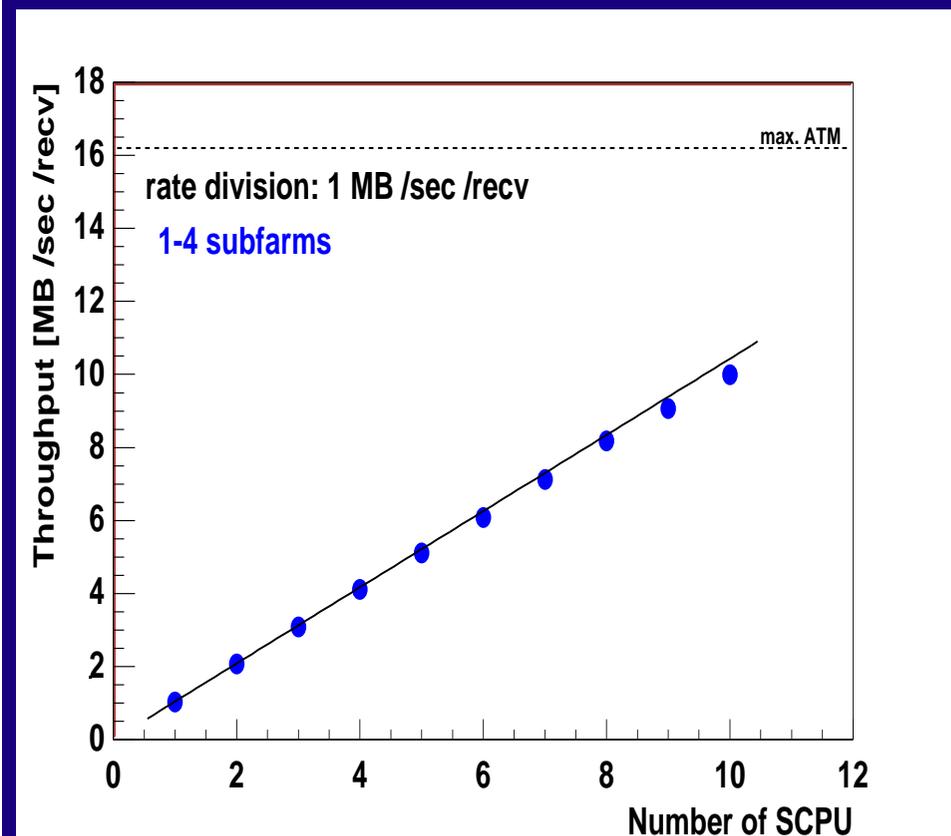
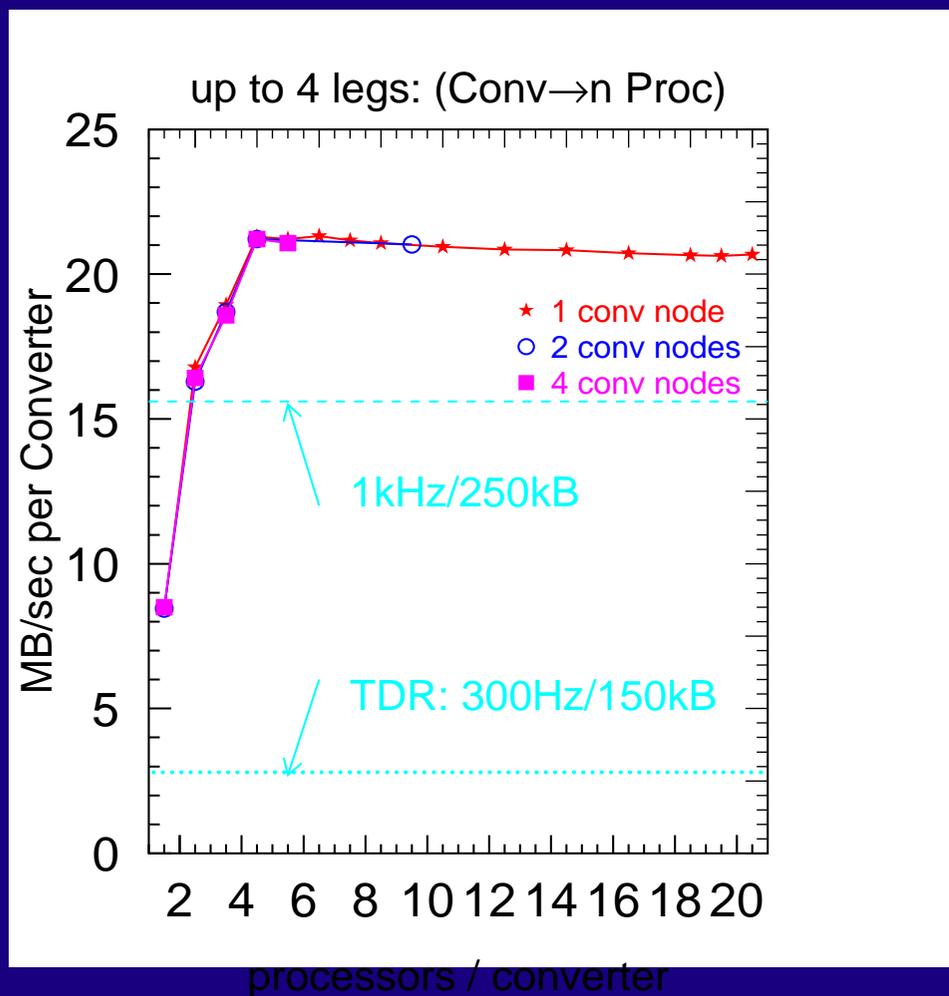


Summary: Linux for Quasi-Real Time

- Modifications to the kernel:
 - TCP/IP “Slow Start” disabled
 - Socket buffer sizes increased (default max 64K in 2.0)
 - Bump *need_resched* in receive IRQ handler
 - Converter node time slice from 10 to 1 msec
- Use built-in Linux real time scheduling (POSIX)
sched_setscheduler()
- Use standard UNIX debugging tools:
 - strace*
 - tcpdump*
 - and also, *TRACE*
- **Never underestimate utility of full access to operating system source code**

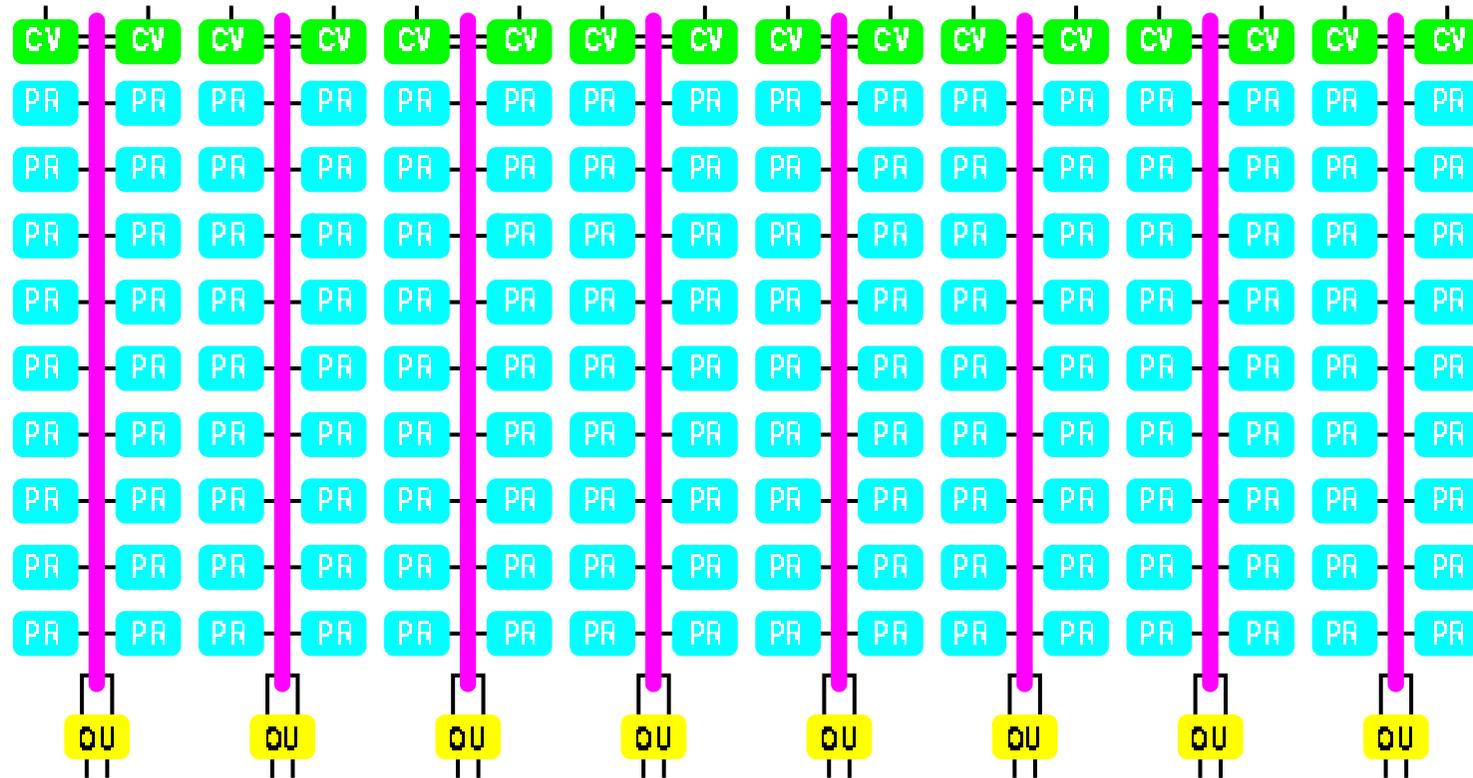
Acceptance of Design

- Based upon results, CDF approved this design for Run II in Spring, 1998
- System was expanded to 32 nodes to test scaling - testing successful





Level-3: Run II Start



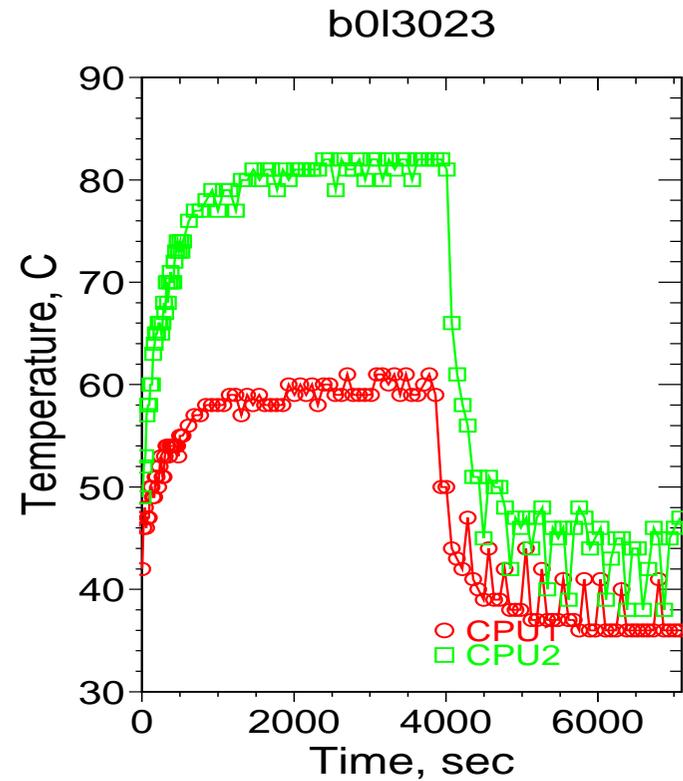
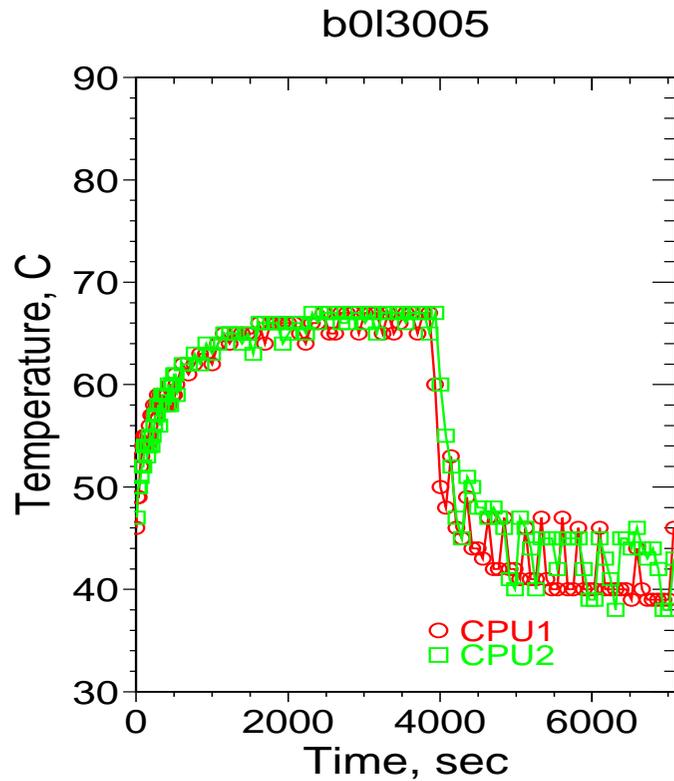
16×9 nodes = 144 processor nodes

Building and Running Large Clusters

- Choose vendors carefully:
 - evaluate test system
 - provide master disk for replication
- Health monitoring:
 - temperatures, fans
 - accessible on many boards via *lm_sensors* package
- Intel's server boards instead use **IPMI**
 - abstracts sensor definitions
 - maintains event logs based upon limit alarms
 - remote (via serial line) resets and power control
 - hardware watchdog timer

Good and Bad Temperature Profiles

- Measure CPU temperatures while running CPU-intensive codes
- “b0l3023” had loose fan:



Conclusions and Futures

- Linux is having a big impact at Fermilab
 - offline reconstructions farms (hundreds of nodes)
 - CDF Level 3 trigger (at least 100 nodes)
 - desktops (300+ nodes now)
- Linux is suitable for soft realtime applications
- Future trigger application - BTeV experiment (2005?):
 - 4000 processors
 - 25 GB/sec rate into Level 3
 - at least a 256 x 256 switch (Myrinet?)
- Future online applications:
 - embedded systems (VME-based realtime computers)
 - hope to provide an alternative to VxWorks
 - will use this year on Motorola PPC boards on *Cryogenic Dark Matter Search*