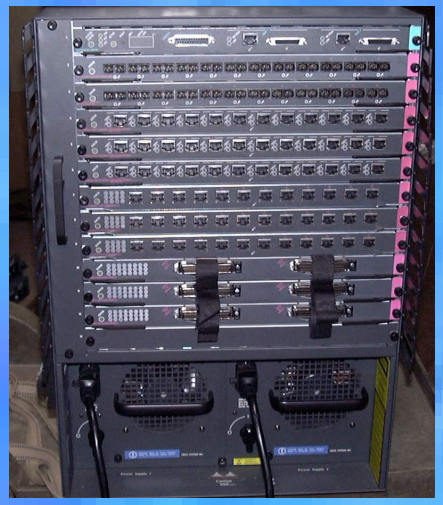
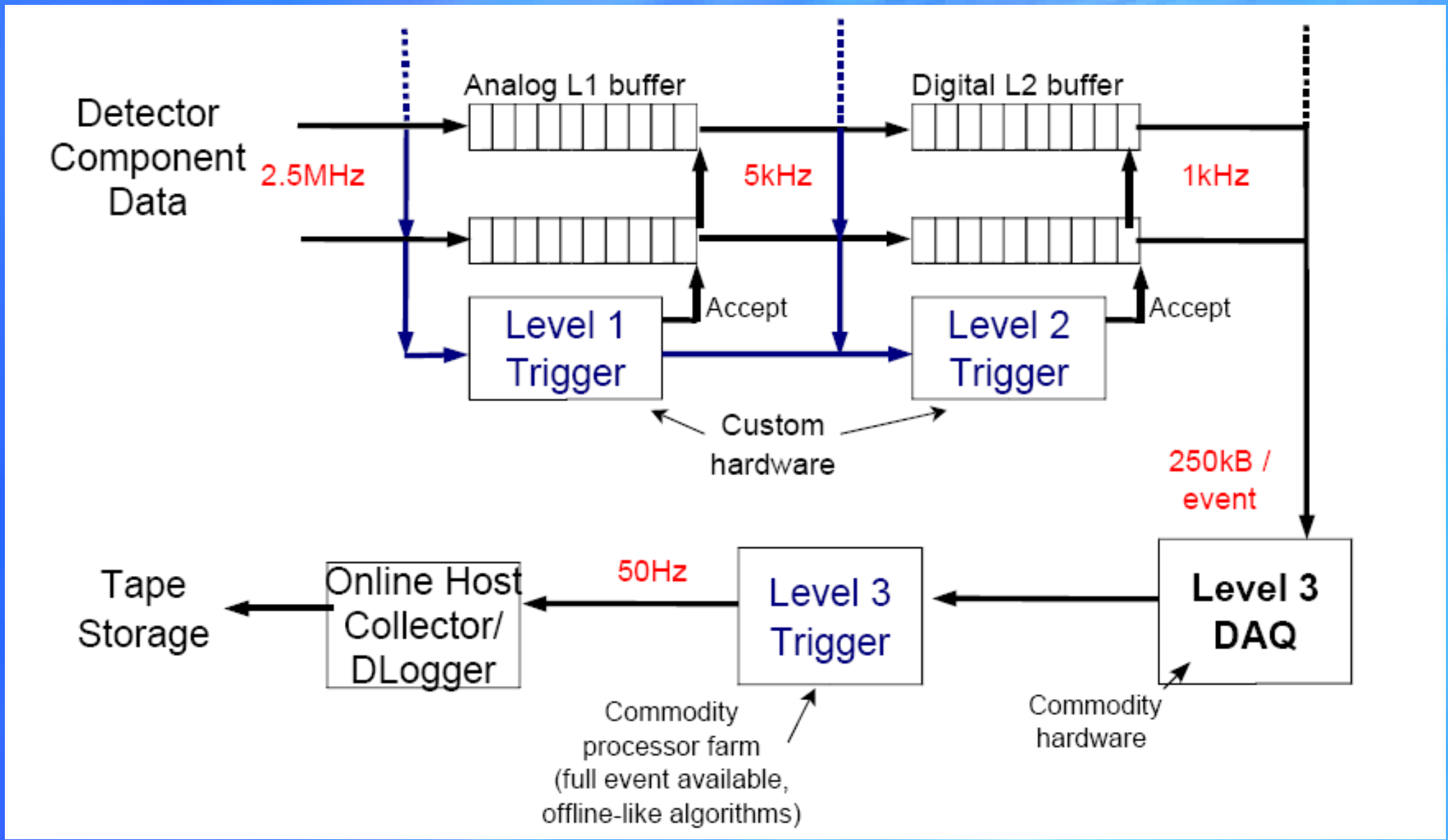
The DØ DAQ



Andy Haas
Columbia University

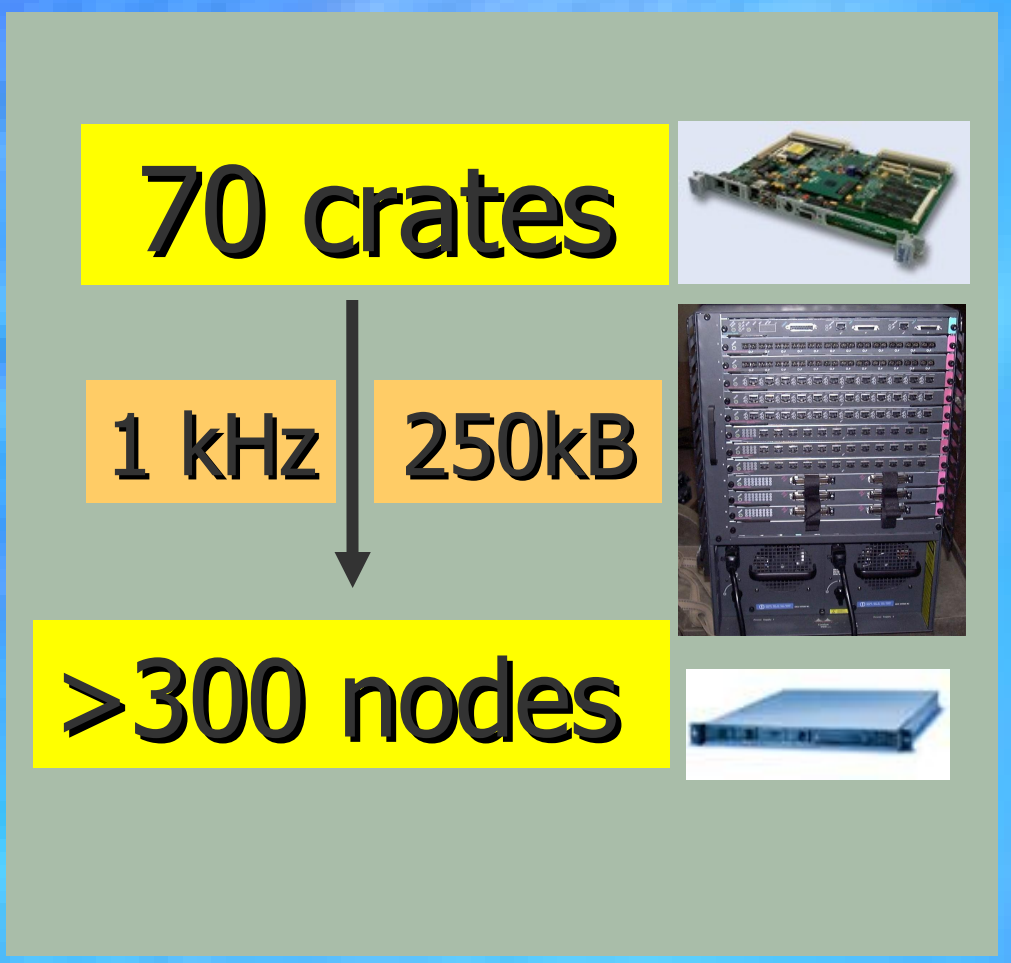
January 24, 2008

Trigger Overview



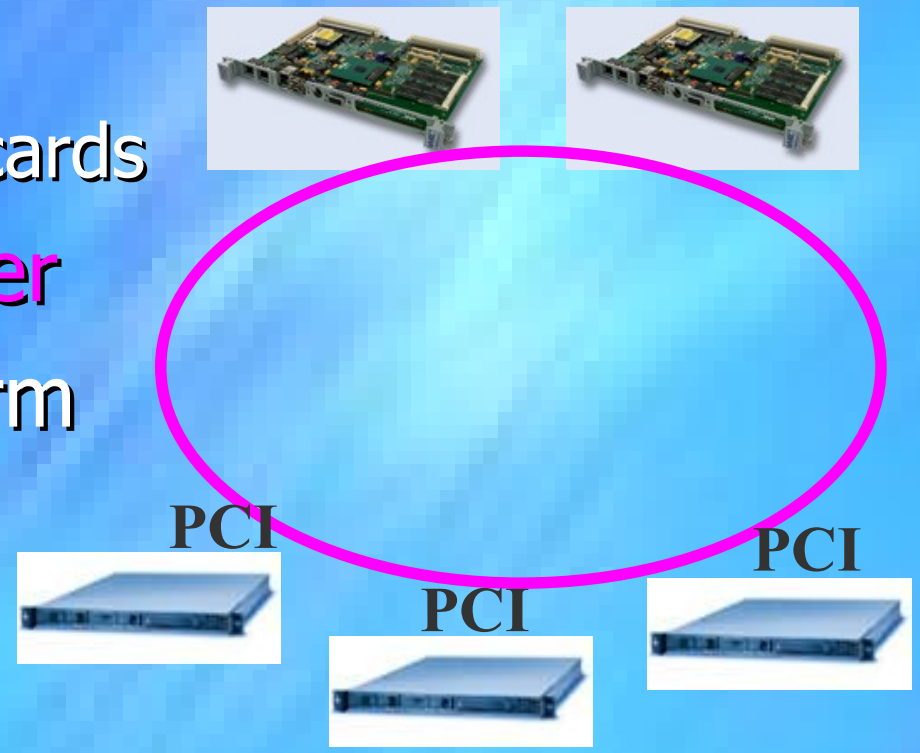
Level 3 DAQ Overview

- Gathers raw data from the front-end crates following each Level-2 trigger accept
- Assembles the event fragments in a farm node, for filtering by Level-3 algorithms



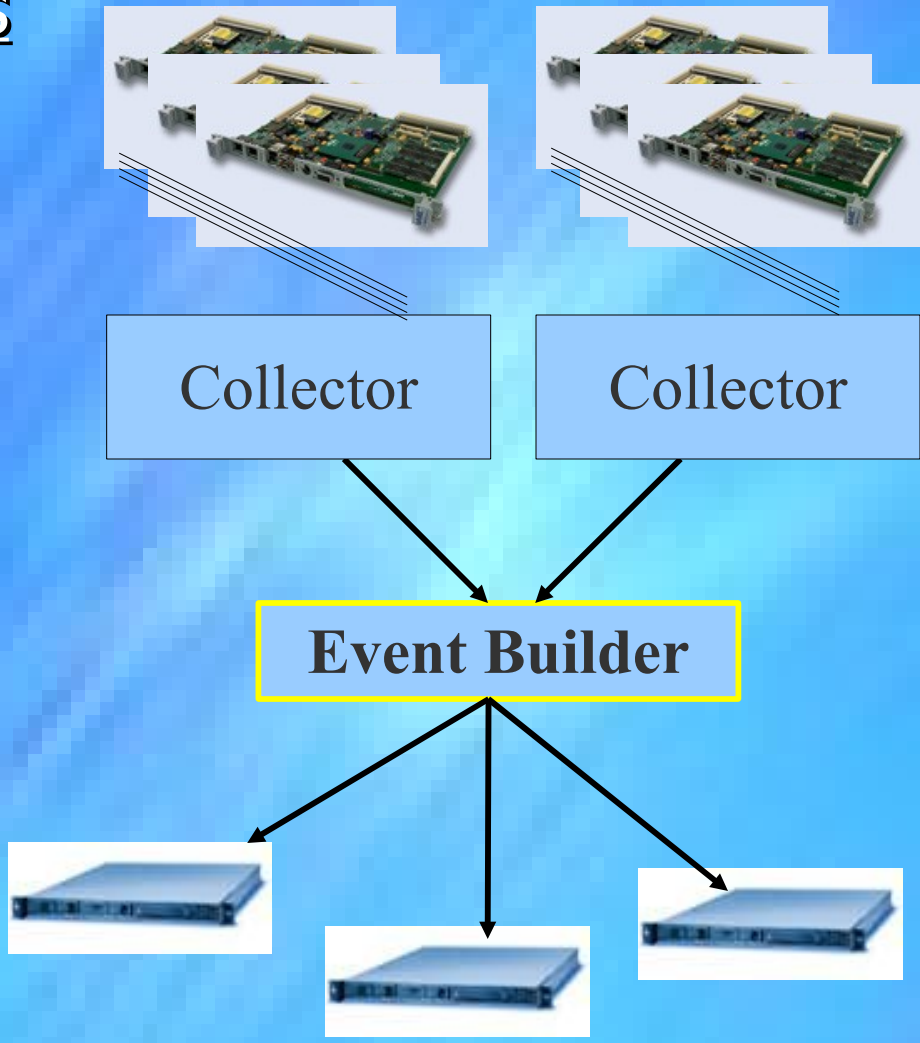
Original "Fiber / Token" DAQ Design

- Upgrade of the Run I DAQ
  - Reused Run I crate-readout cards
- Event data circulated on fiber
- Custom PCI card in each farm node assembles events



Problems

- 2001 – beam was turned on
- PCI cards still not working
  - Firmware problems
  - Non-disclosure agreements



- Temporary system:
  - data from each floor sent to a single event-builder machine over Ethernet
  - ~50 Hz event rate into Level 3 filter farm!
  - Collectors and Event Builder running Windows NT

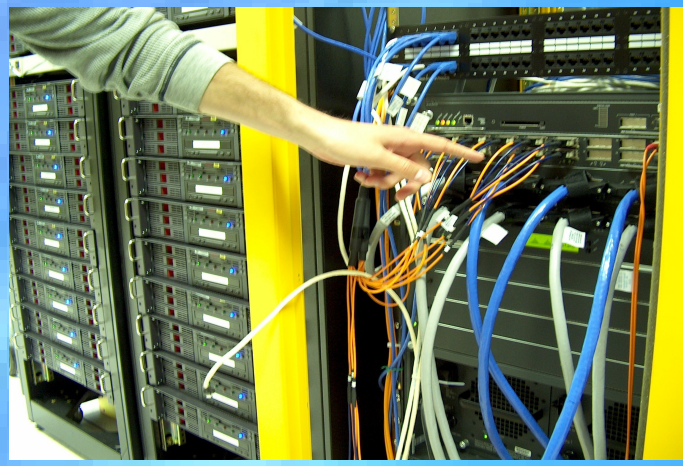
What to do?

- Here's $1M
- Find a way to read out the detector at 1000 Hz
- Try to be done in 6 months
- Don't interfere with current data taking of 50 Hz

- Small, but motivated team of ~6 people
  - One excellent CD engineer

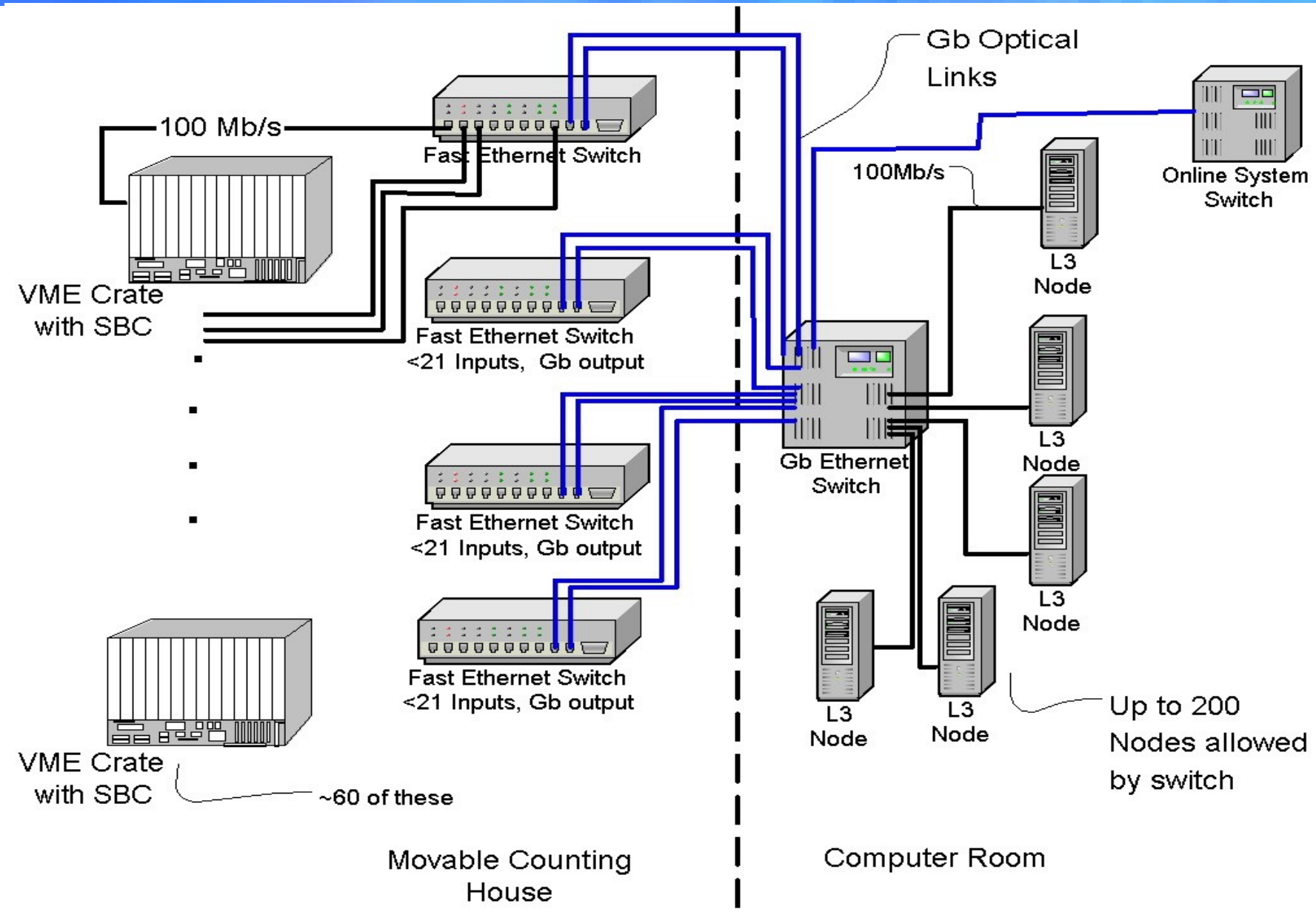
Commodity System

- Commodity hardware
- Commodity software



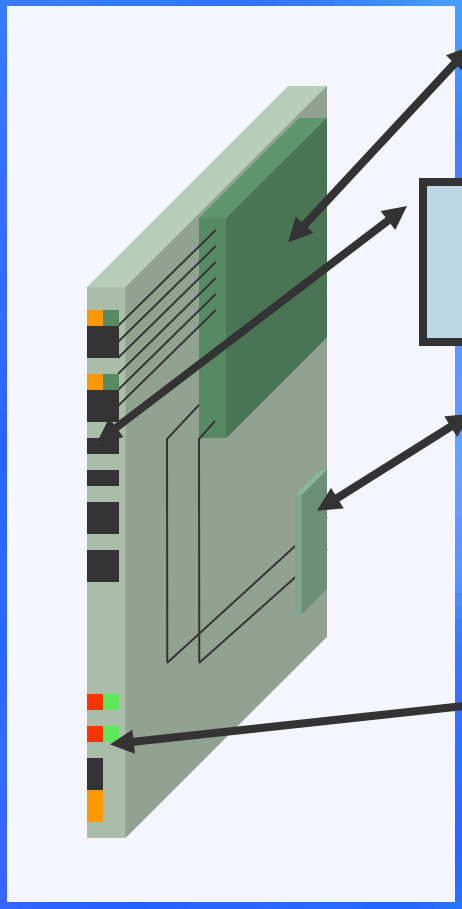
We chose a good mix of hardware and software and built a system that easily met the 250KB @ 1KHz (=250MB/sec) requirement

Ethernet DAQ System



The Single-Board Computer (SBC)

- Mechanically supported in the crate by a custom 9U "Extender" board

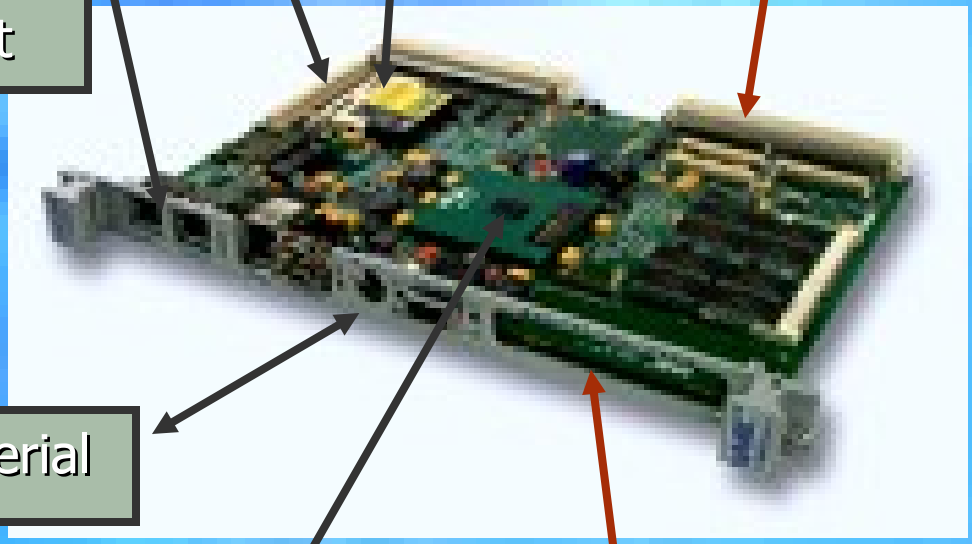


SBC

Front-panel connections

J3 connector

Status lights



128 MB RAM

128 MB Flash ROM

VME interface

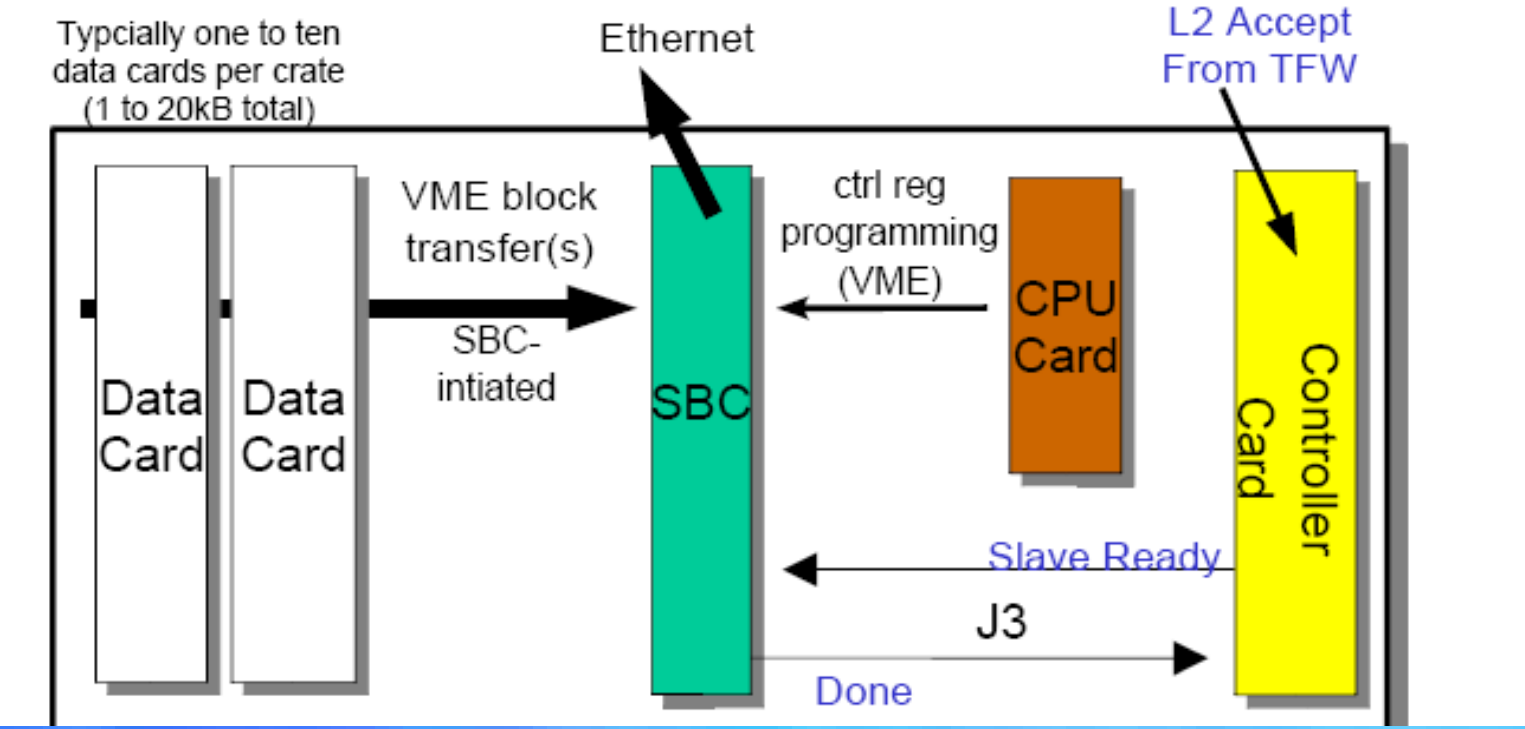
Dual 100Mb Ethernet

VGA, Serial

933 MHz CPU

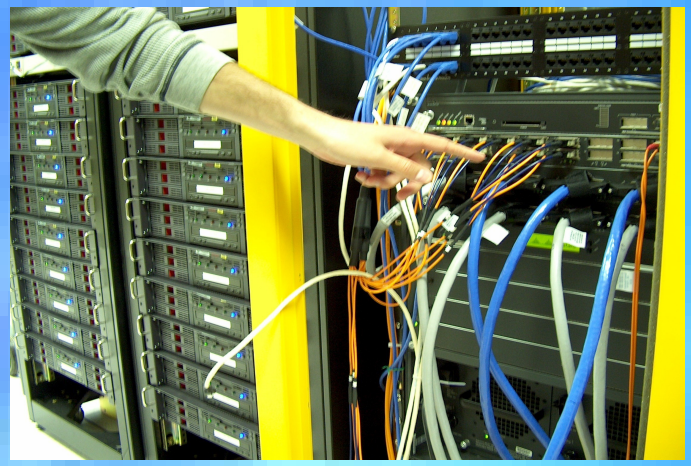
"PMC" slot (filled with BVM I/O module)

Crate Readout



Cisco Ethernet Switches

- 6509 (single central switch)
  - 16 GB/s backplane
  - 112MB/48 ports of output buffering
  - 9 module slots – can each support:
    - 48 100mb/s ports
    - 8 Gb fiber or copper ports



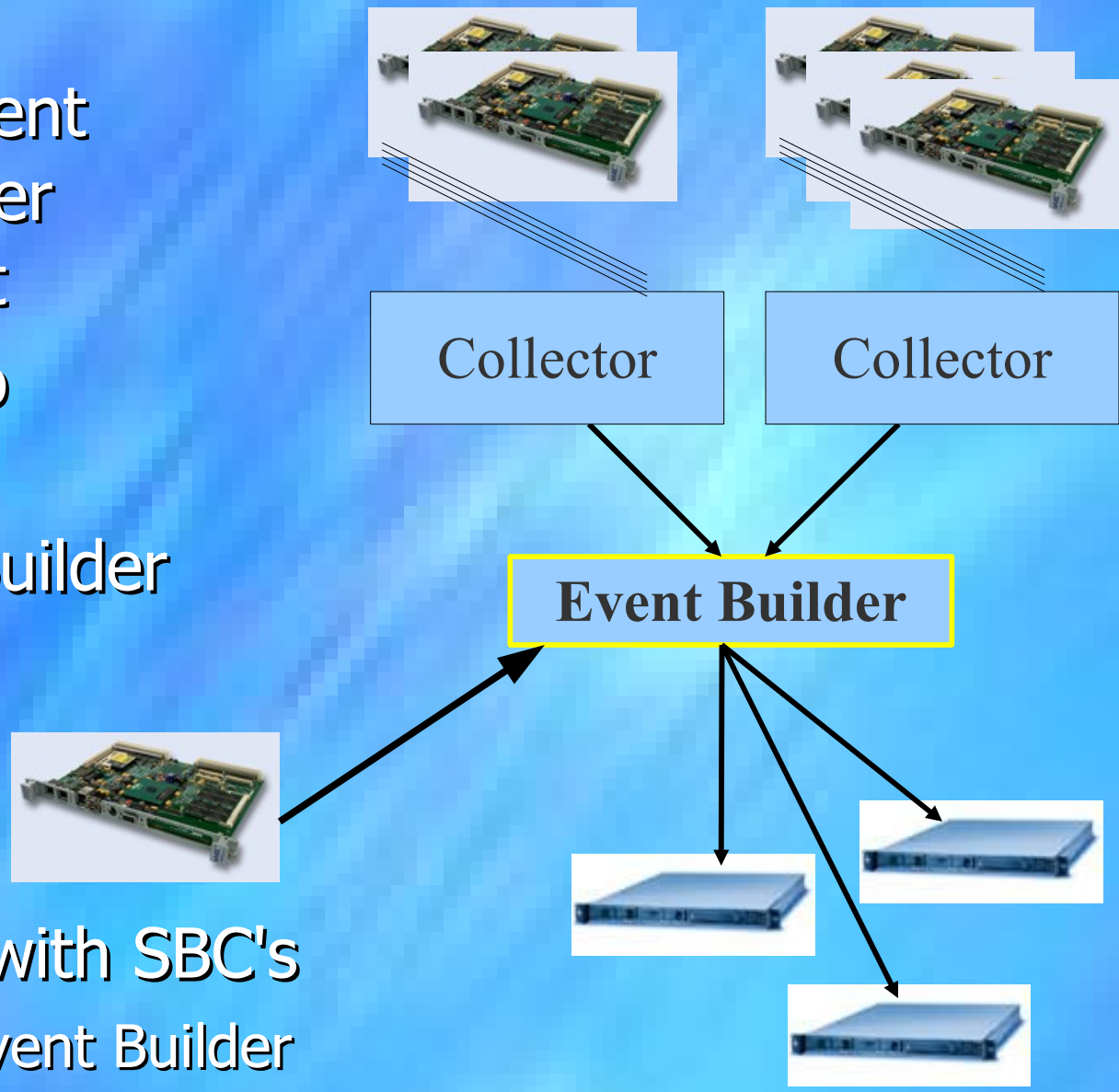
- 2948G (currently 5 of these in the system)
  - "concentrator" switch
    - Combines data from 10 100mb/s inputs in a single Gb fiber
      - No packet loss possible
  - Dual GBICs

Farm Nodes

- Currently ~300 Nodes
- Dual, Quad CPU
- Dual Core
- 1 GB Ram / core
- Dual Ethernet

- Easily expandable
  - (Just buy a new rack)

- Intermediate system:
  - data from each floor sent to a single event-builder machine over Ethernet
  - ~50 Hz event rate into Level 3 filter farm!
  - Collectors and Event Builder running Windows NT
  - Some crates installed with SBC's
    - directly send data to Event Builder
    - Linux -> Windows NT

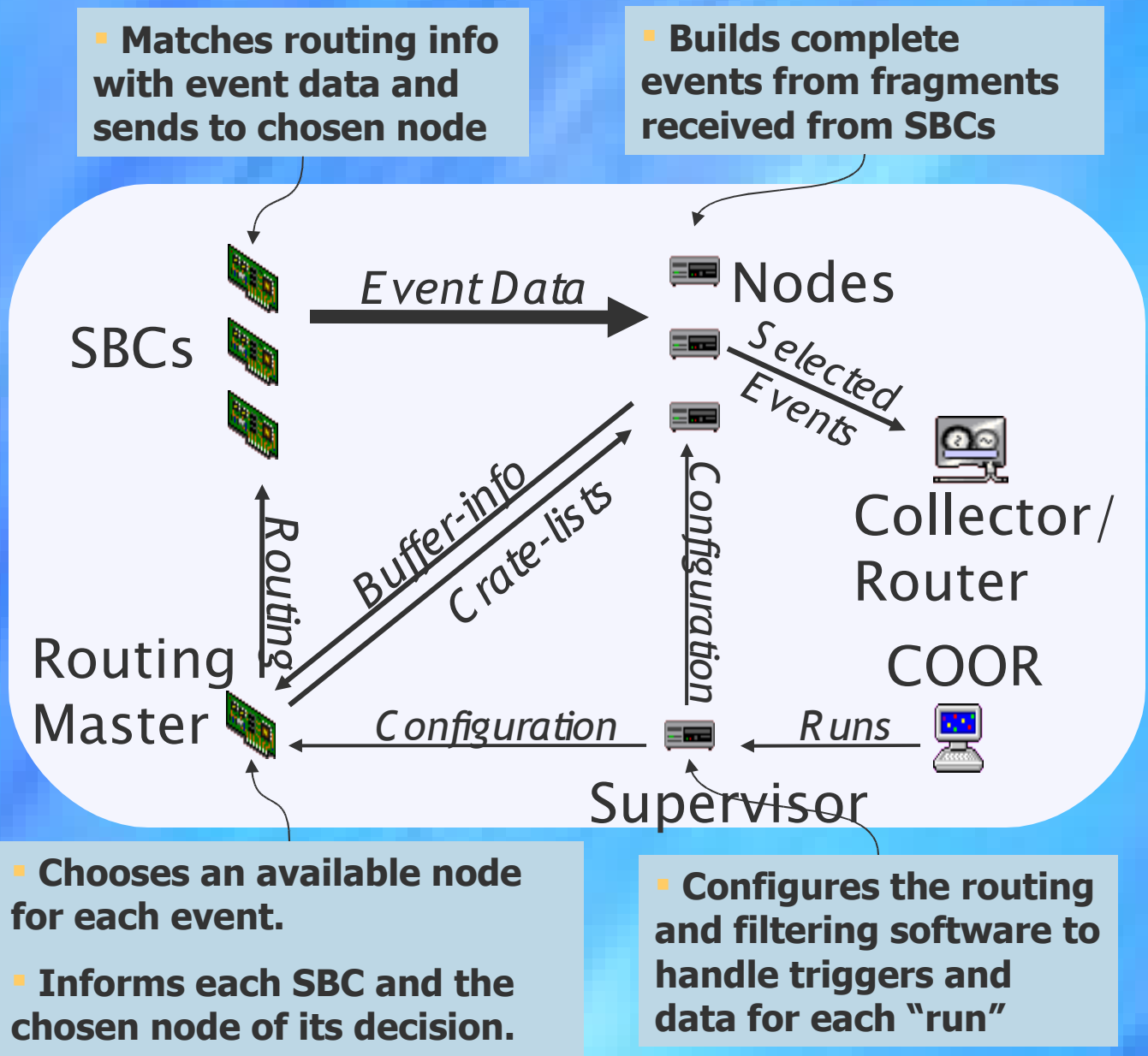


DAQ Software

- C/C++
- TCP/IP
- Linux

- Commodity software libraries: ACE, Xerces, TRACE, Zlib

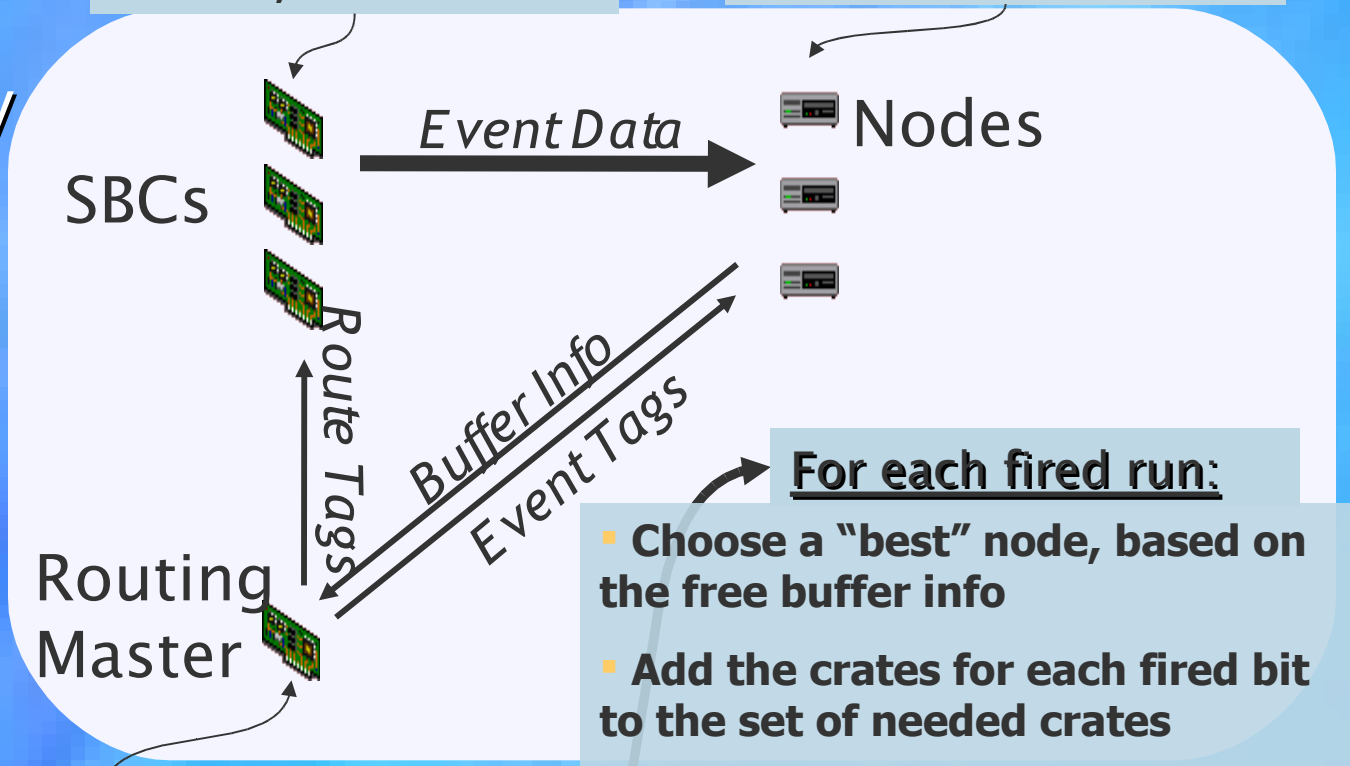
- TCP connections are robust / reconnecting
  - components of the system restarted 'on the fly'



Routing / Event Building

- Performance:
  - ~1 kHz operation
  - ~10 ms event latency
- Simultaneous "runs"

Run
- Set of trigger bits
- Set of crates for each bit
- Set of nodes



For each fired run:

- Choose a "best" node, based on the free buffer info
- Add the crates for each fired bit to the set of needed crates
- Queue a new (or add to an existing) route tag to be sent to each needed crate
- Send an event tag to the node containing the fired bits and the set of needed (expected) crates

- Event 5466 fired on bits 45,48,66
- Consult runs… A run is "fired" if its bits overlap with the fired bits

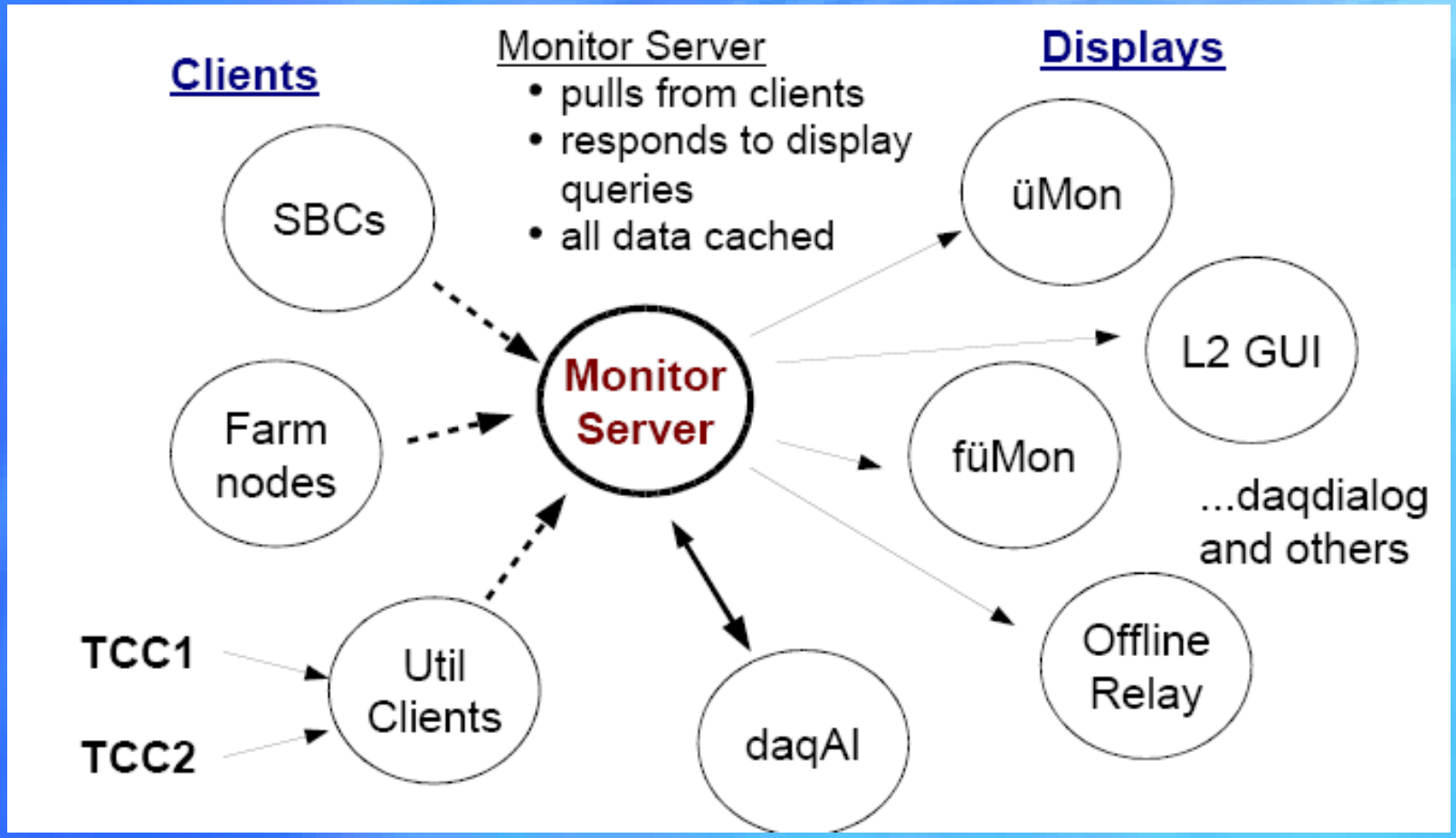
Monitoring and Debugging Tools

- TRACE
  - ~Real-time logging of kernel and user-space programs
  - Can trigger external scope for understanding VME issues

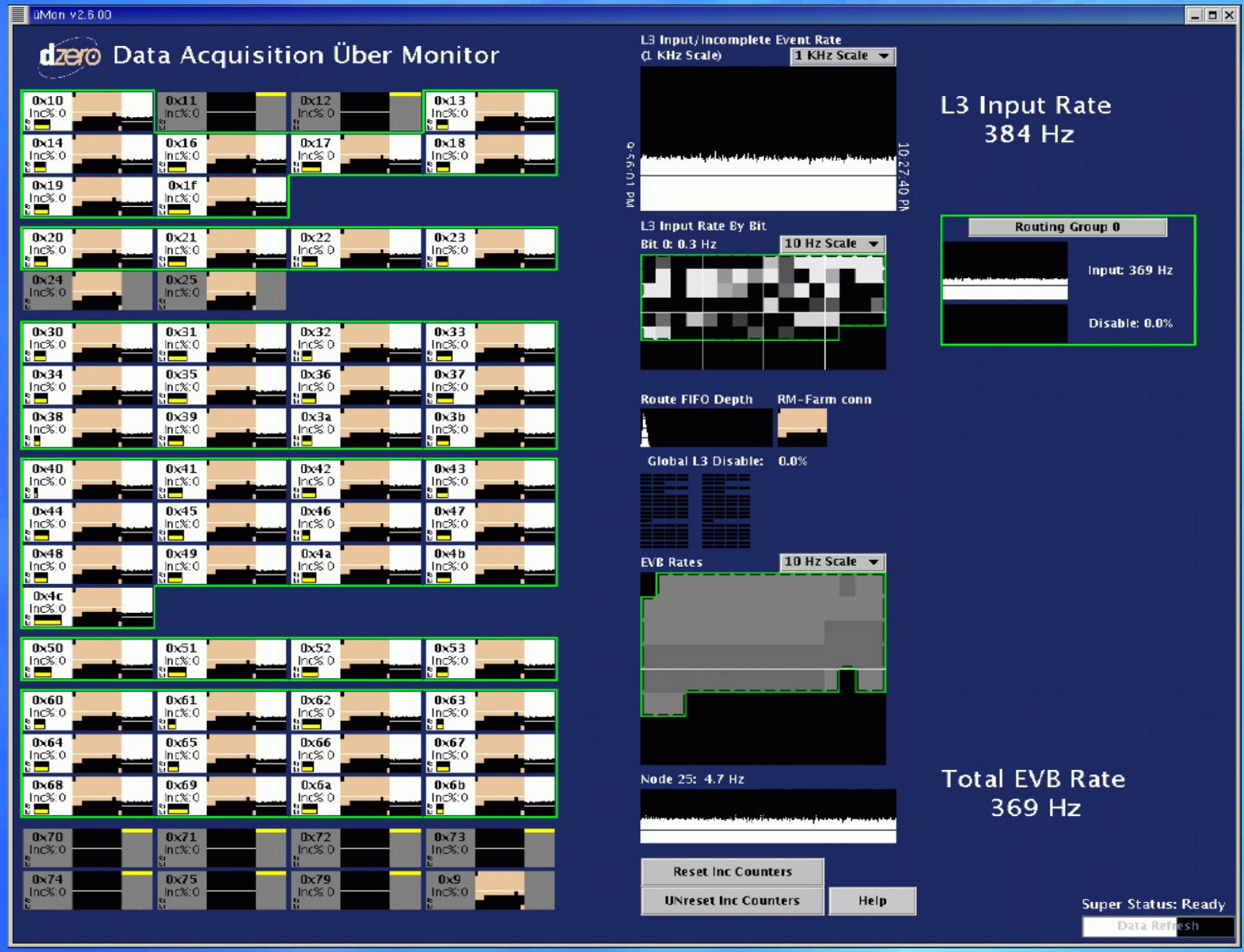
```
entry #   timestamp          id  level  clock tick          message
-------   ---------          --  -----  ----------          -------
 8979   1017631352199255  KERNEL  10    183863277955182   ioctl_bigPhysRel (aft rel) 24=used 13=tail 12=he
 8980   1017631352199212  rm   4    183863277915087    Getting event...
 8981   1017631352199208  rm   4    183863277910827    Readout is enabled
10721   1017631337589517  rm   8    183849719469012      Waiting for global enable event
10722   1017631337589514  rm   4    183849719466324    Check readout disable...
10723   1017631337589511  rm   4    183849719463400    Done routing event
10724   1017631337589217  rm   4    183849719190706    Disabling global trigger
10725   1017631337589216  rm   4    183849719189051    Disabling global trigger for bit 1
10726   1017631337589175  rm   4    183849719151653    Routing event 33178
10727   1017631337589173  rm   8    183849719149222      Event 33178 bits 00000000000000000000000000000002
10728   1017631337589130  rm   4    183849719109894    Success unpack TFW block
10729   1017631337589090  rm   4    183849719072440    Unpacking TFW block
10730   1017631337589081  rm   4    183849719063895    Got event of length 2892
10731   1017631337589076  KERNEL  10    183849719059223   ioctl_bigPhysRel (aft rel) 0=used 12=tail 12=he
```

Monitoring and Debugging Tools

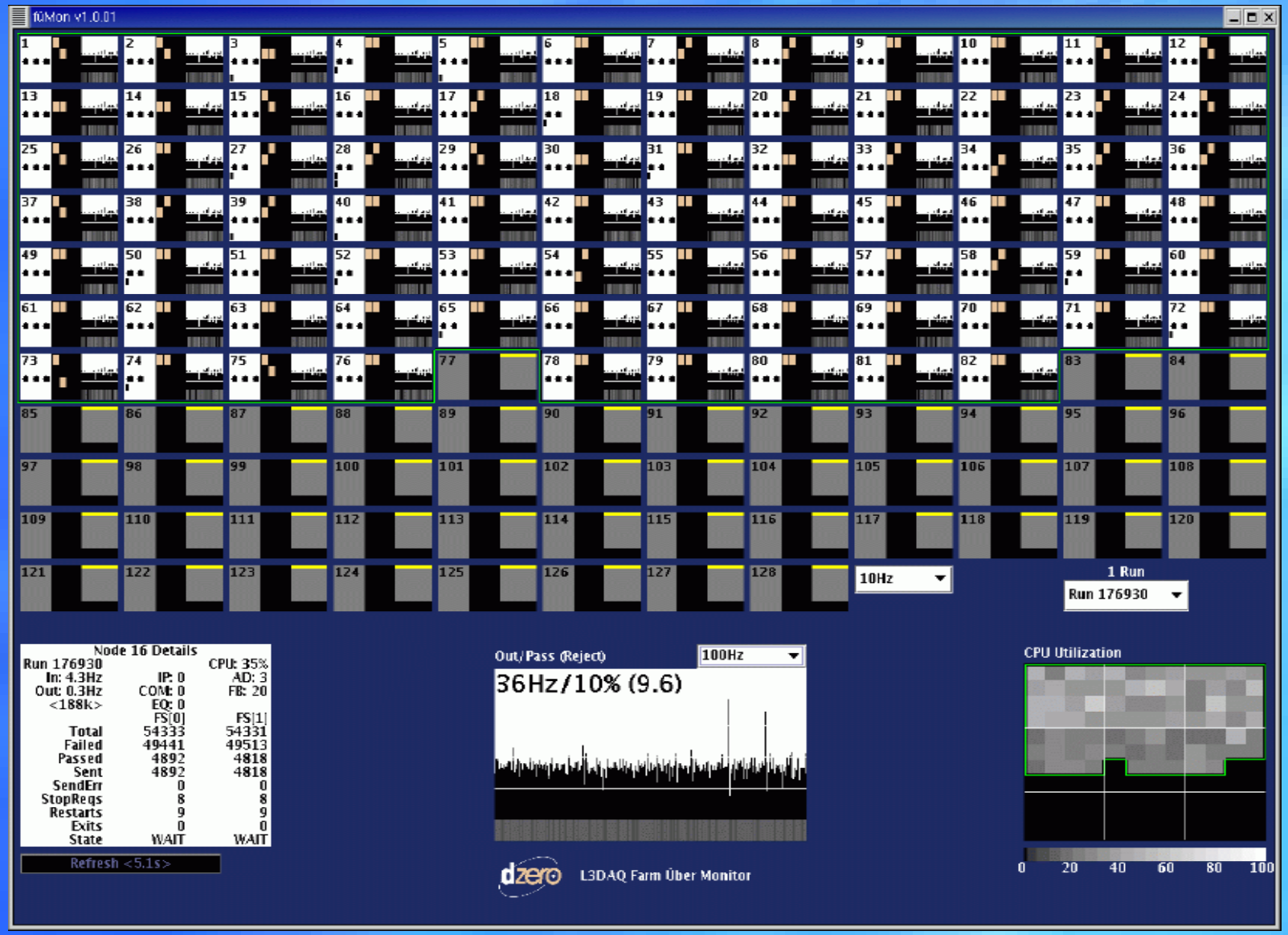
- Crucial for development and key to long-term support and usability



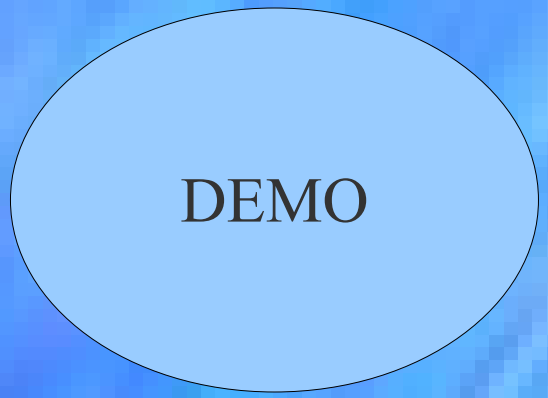
uMon : Control Room SBC Display



fuMon : Control Room Farm Node Display



L3x Qt : Expert-level DAQ Display



Lessons Learned

- VME systems integration is *delicate*
- 200 ms dropped packet problem
  - TCP not tuned for 'real-time' applications by default
    - TCP_RTO_MIN parameter and others need tuning
  - Detailed knowledge of TCP necessary
- Expert understanding of Linux kernel and TCP tools needed
- Moving parts break
  - Farm nodes: fans, disks

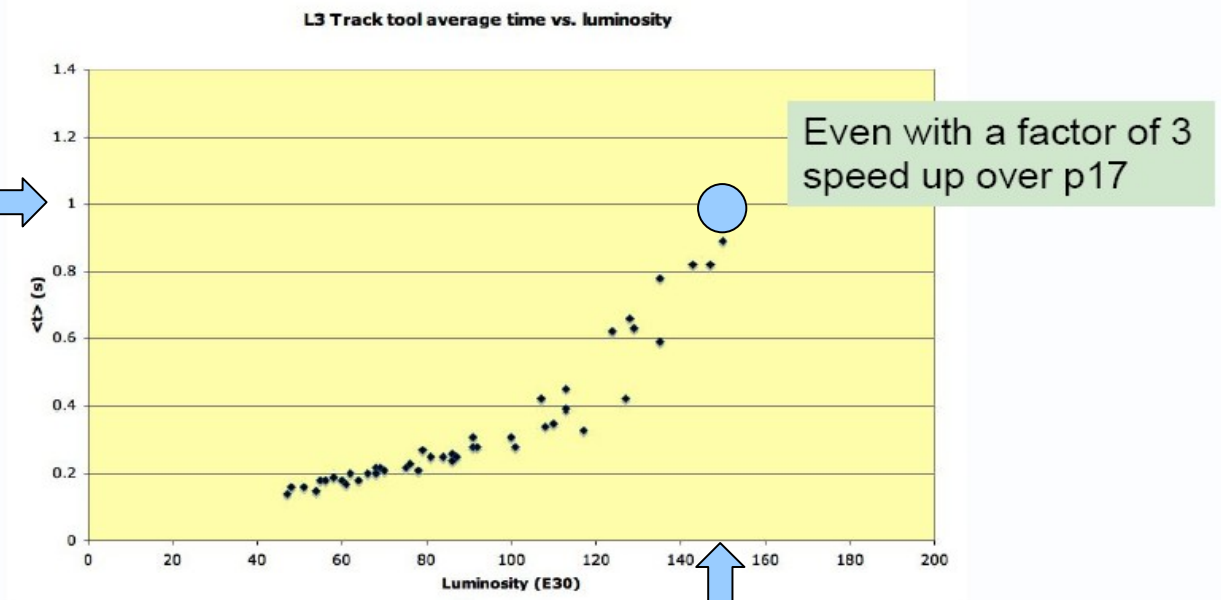
Later Upgrades

- The system is expandable and flexible to handle any DAQ needs or problems that arise
  - Gigabit Ethernet has been added on overloaded crates
  - Many more nodes have been added
  - Node maintenance taken over by Computing Division
  - "Lazy" reconnects
    - nodes give up after a while
    - start trying again after being signalled of a system change
  - Event caching...

Keeping Up

- Events take longer to filter at high inst. luminosity
  - Original goal was 50 nodes/1kHz = 50ms/event
  - Now taking *>1 sec./event* at high luminosity

- Present L3 farm is at 99% CPU usage at 150E30
  - New nodes, trigger list design will help, but tracking time increases non-linearly with luminosity

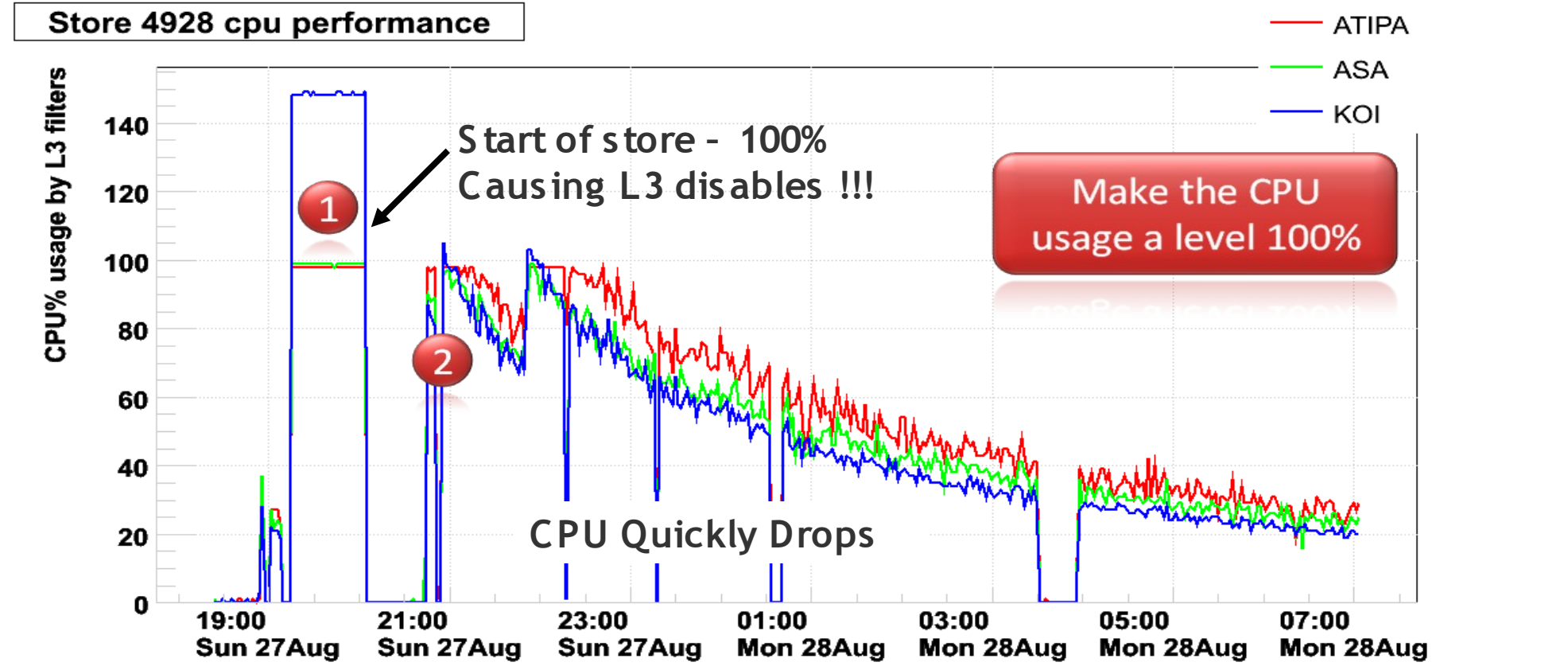


1 sec./event

Even with a factor of 3 speed up over p17

150 e30

Using Idle CPU / Event Caching



1. Cache events on farm nodes for later processing
2. Drain cache during unused CPU cycles at lower luminosity

- Assume a 2 hour run and a standard lumi. profile:
  event caching equivalent to ~50% more farm power

Conclusions

- DØ developed a commodity, Ethernet-based DAQ system for Run II
  - On time, and under budget
  - Very reliable performance for past 7 years
- System has been easily upgraded to meet constantly increasing demands
  - Commodity hardware continues to improve with time
  - C++ software very flexible

Conclusions

- The DØ DAQ was a good example for the future
  - CDF upgraded to Ethernet DAQ in ~2004
  - Almost all LHC experiments use commercial, Ethernet DAQ

- From the ATLAS TDAQ TDR:

“

Experience has shown that custom electronics is more difficult and expensive to maintain in the long term than comparable commercial products.

The use of commercial computing and network equipment, and the adoption of commercial protocol standards such as Ethernet, wherever appropriate and possible, is a requirement which will help us to maintain the system for the full lifetime of the experiment.

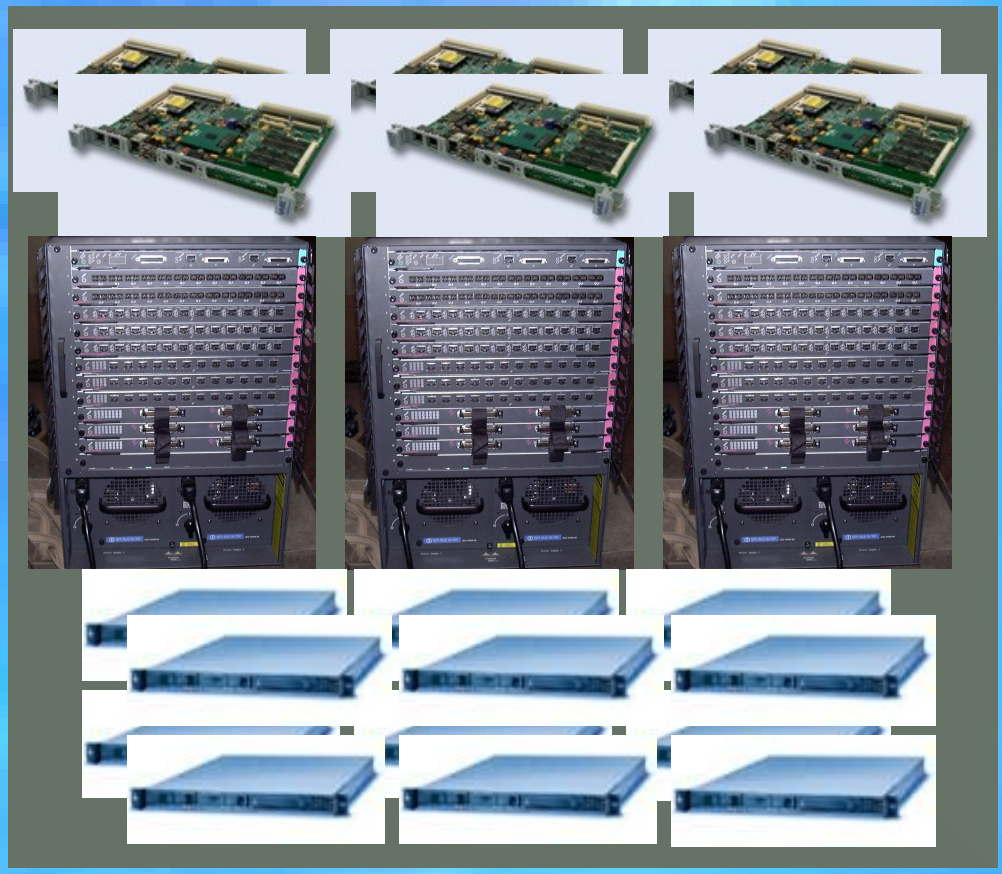
The adoption of widely-supported commercial standards and equipment at the outset will also enable us to benefit from future improvements in technology by rendering equipment replacement and upgrade relatively transparent.

”

Backup

Why use Commodity?

- Dependable, supported, widely-used, high-performance
- Very little schedule risk
- Well understood budget
- Expense is comparable to a custom system
- Easily upgradeable as technology improves



Installation in 2002

Feb. 15:
VMIC 7750 SBCs and BVM I/O cards ordered.

Mar. 9:
Cisco 6509 switch and 2948Gs installation complete.

May 5:
Ethernet cabling complete.
First assembled SBCs arrive.

Apr. 24:
Extender boards parts and I/O modules arrive in bulk.

May 24:
Tracking crate SBC installation done.

Feb. 20:
First full integrated online software test.
Supervisor upgraded.

Apr. 1:
41 SBCs arrive!

May 8:
First floor SBC installation done.
VRBC multi-buffer works.

June 15:
48 farm nodes.
Full-rate DAQ complete!

Feb. 1:
Routing Master installed in the TFW.
SBCs in a few crates, sending to Segment Bridges.

Mar. 26:
Software transition complete!
"Virtual SBCs"

Jun 4:
SBC installation complete!

Layered Event Building

- Reading out > ~100 VME crates requires more complicated event building
- Put DAQ 'simplexes' together to form a multi-layered, expandable system
- The output of each simplex becomes an input for the next layer
- Could do ~10000 SBCs to 10000 Nodes with 2 layers

- **Get info from Emperor and pass to SBCs**

- **Tell DAQ nodes which event node to use**
- **Advertise total free buffers to the Emperor**

- **Emperor, for each event:**
  - **Pick an Event Node Group (ENG) with the most free buffers**
  - **Inform the NM and RMs of the choice**