The DØ Level 3
Data Acquistion System

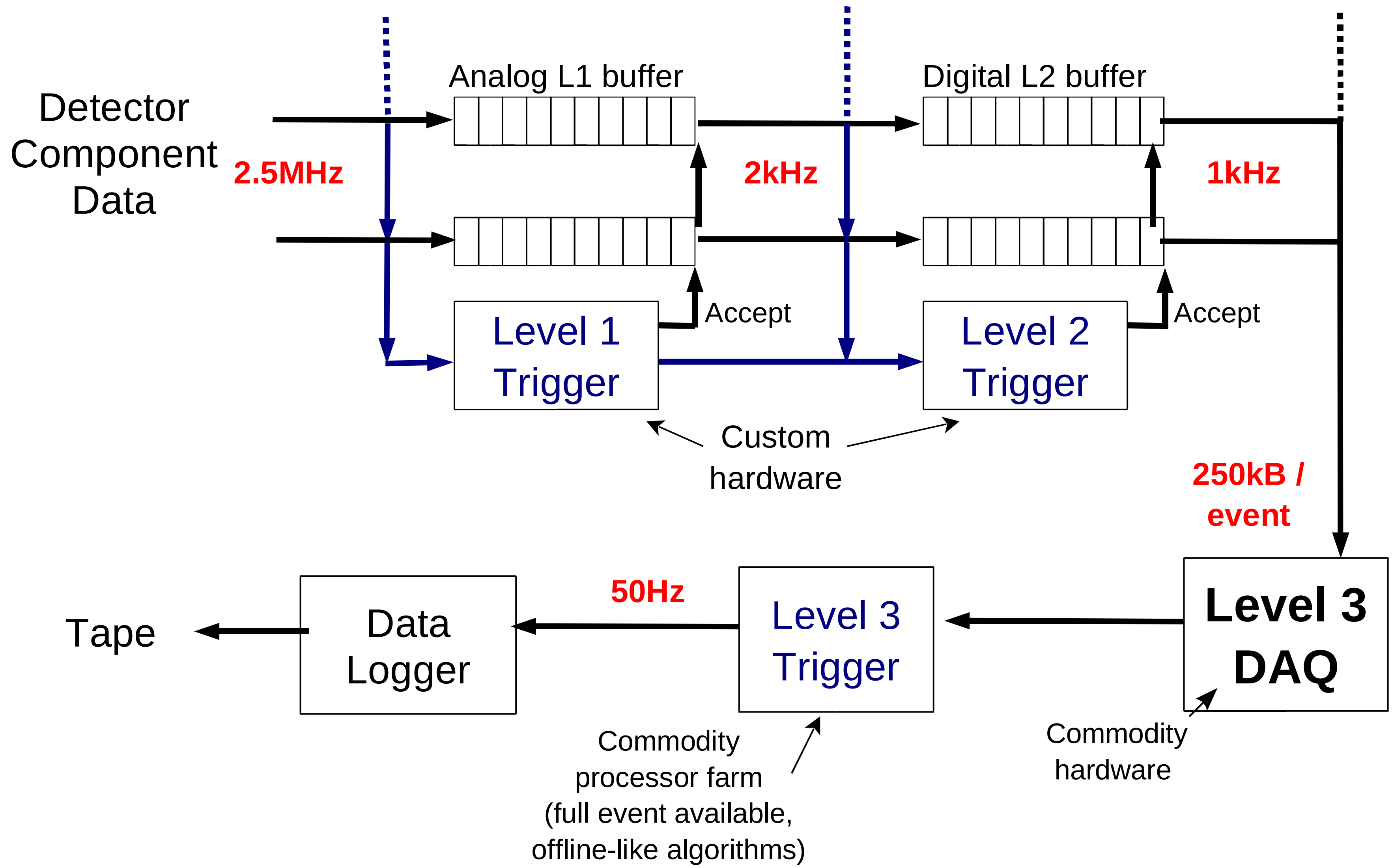
Doug Chapin
Brown Univsersity

For the D0 L3DAQ Group
Brown University
Fermilab
University of Washington

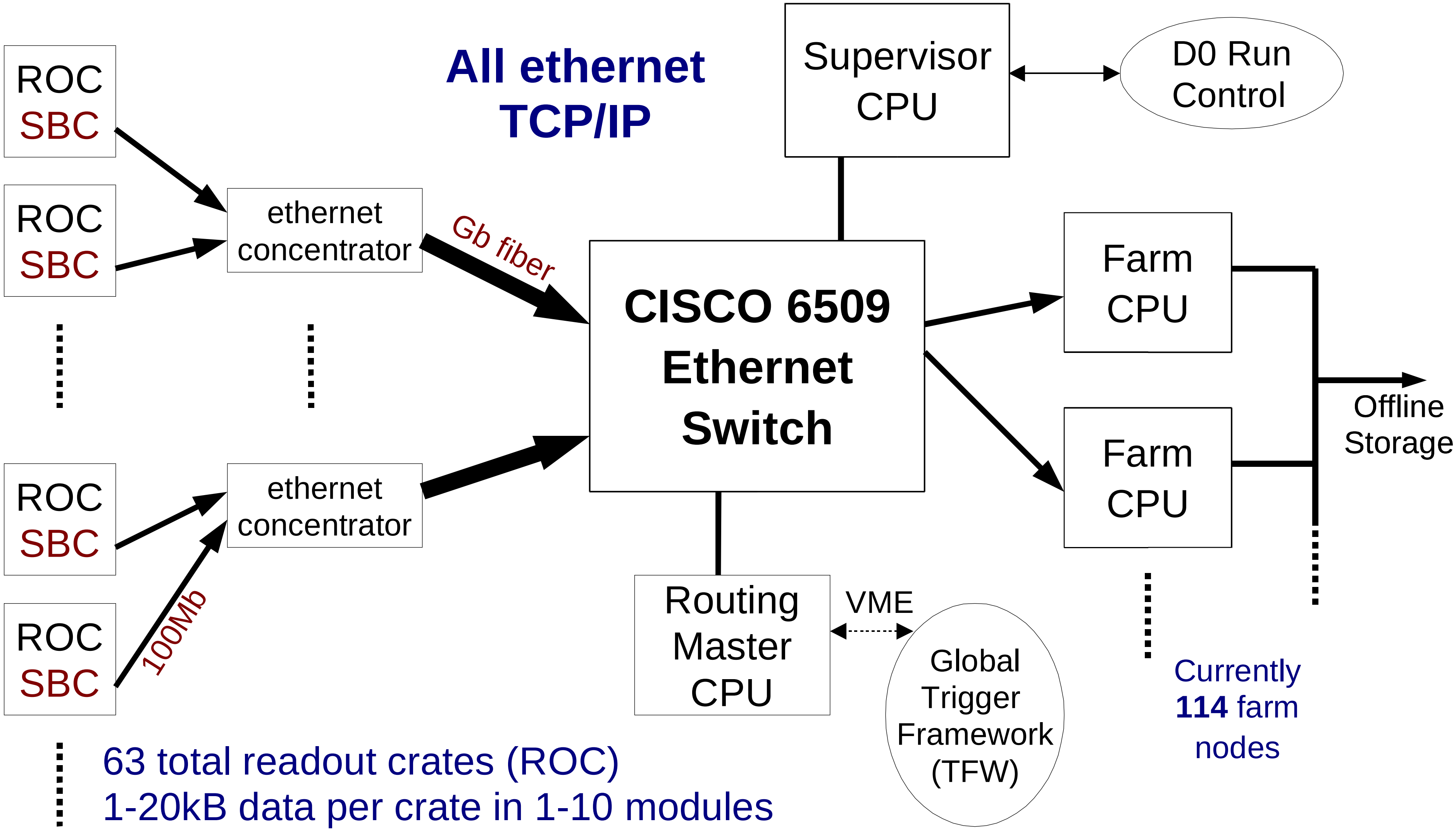
Outline
System Overview
Performance
Running Experience
Future Plans

CHEP
27 September 2004

D0 Data Acquisition System



L3DAQ: Commodity-Based System



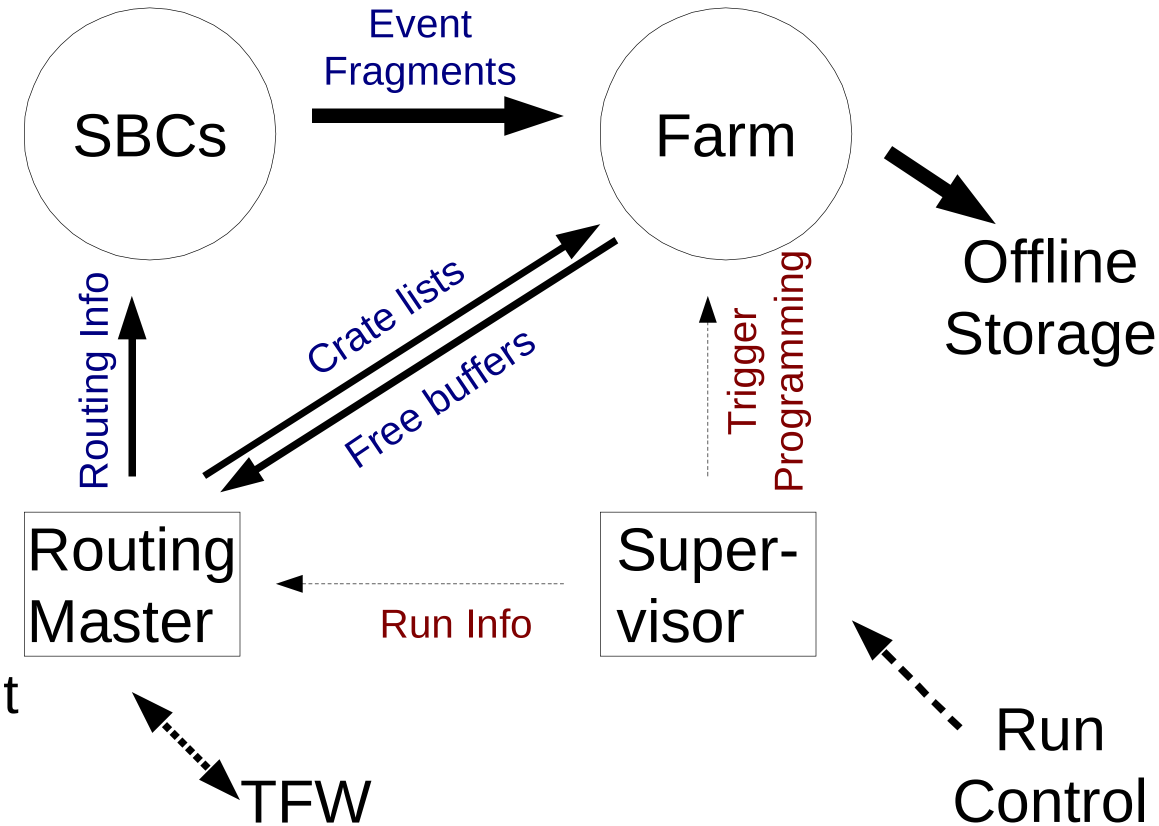
63 total readout crates (ROC)
1-20kB data per crate in 1-10 modules
One Single board VME computer (SBC) per crate

L3DAQ Operation

Partitioning
- Simultaneous runs
- choose destination node based on triggers

Flow Control
- TCP settings limit amount of data in-flight
  - avoid packet loss in switch
- Token scheme limits number of events in-flight
- Supress triggers if farm fills up

Software
- Linux
- C++
- Many shell scripts

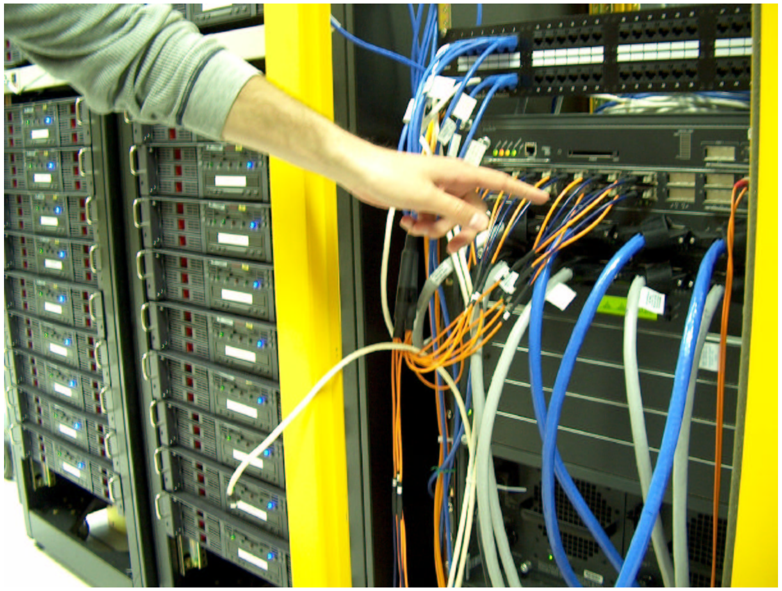


Serving the Needs of D0 since May 2002

Components

SBCs

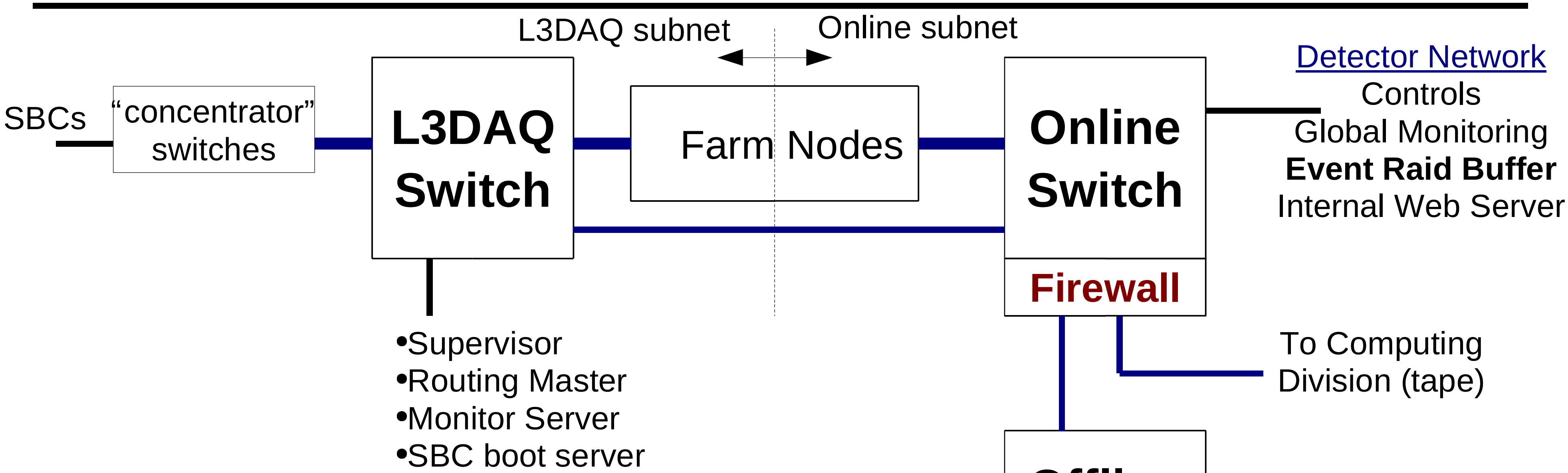
- VMIC-7750, PIII 933MHz, 128MB RAM
- 128MB CompactFlash
- VME UniverseII
- Dual 100Mb ethernet (Intel EEPRO)

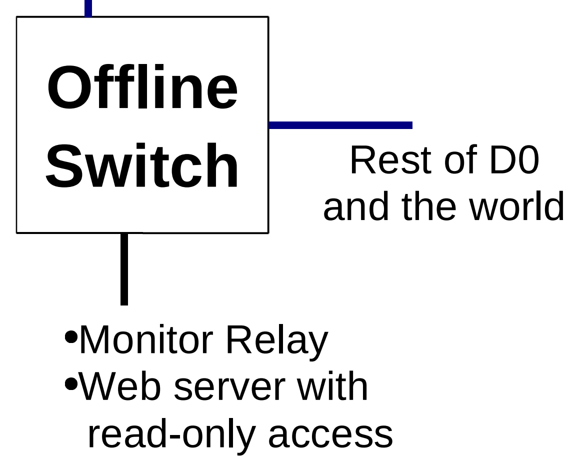


Farmnodes

- Dual Processor
  - PIII 1GHz (48)
  - AMD2000 1.6GHz (34)
  - Xeon 2.8GHz (32) NEW!
- Dual 100Mb ethernet

Network Topology





Security - Firewall

- Kerberized (or cryptocard) ssh only to Online
  - ssh port-forwarding feature is useful
- Firewall holes for
  - External monitor server relay
  - External web server
    - read-only access to some NFS disks

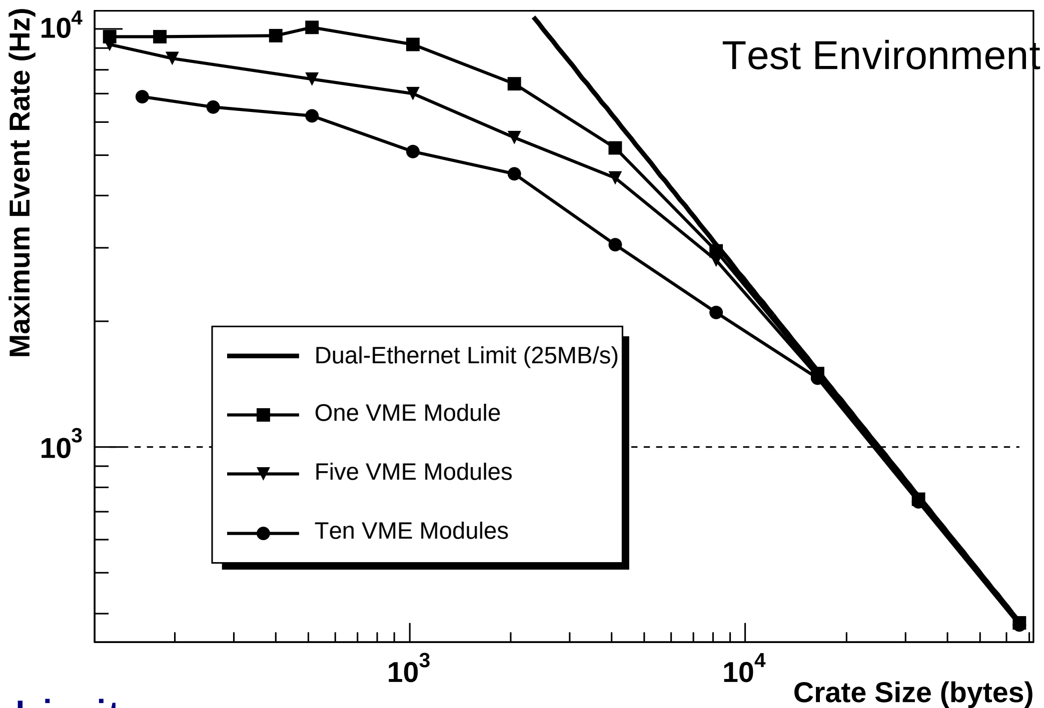
SBC Performance

SBC Operation

- Custom kernel module
  - VME reads
  - Event fragment buffering
- User-level process
  - matches route info to fragments
  - Sends to node(s)
    - no explicit data-copying

Dual ethernet operation

- On crates with large payload
- two connections from each farmnode
- Bind connection to interface
- toggle sending between connections



Limits

- Reach dual-ethernet limit for crate size > 20kB
- VME overhead is main limit for < 20kB
- CPU limited near 10kHz
  - D0 design is 1kHz

SBC Running Experience

Board quality is good

- No problems in over three years
- keyboard/video connectors – very useful

CompactFlash disk is **very** slow

- only used to store config information
- network boot and RAM disk much faster
  - NFS mounted software/log area

Some kernel/driver/EEPRO issues

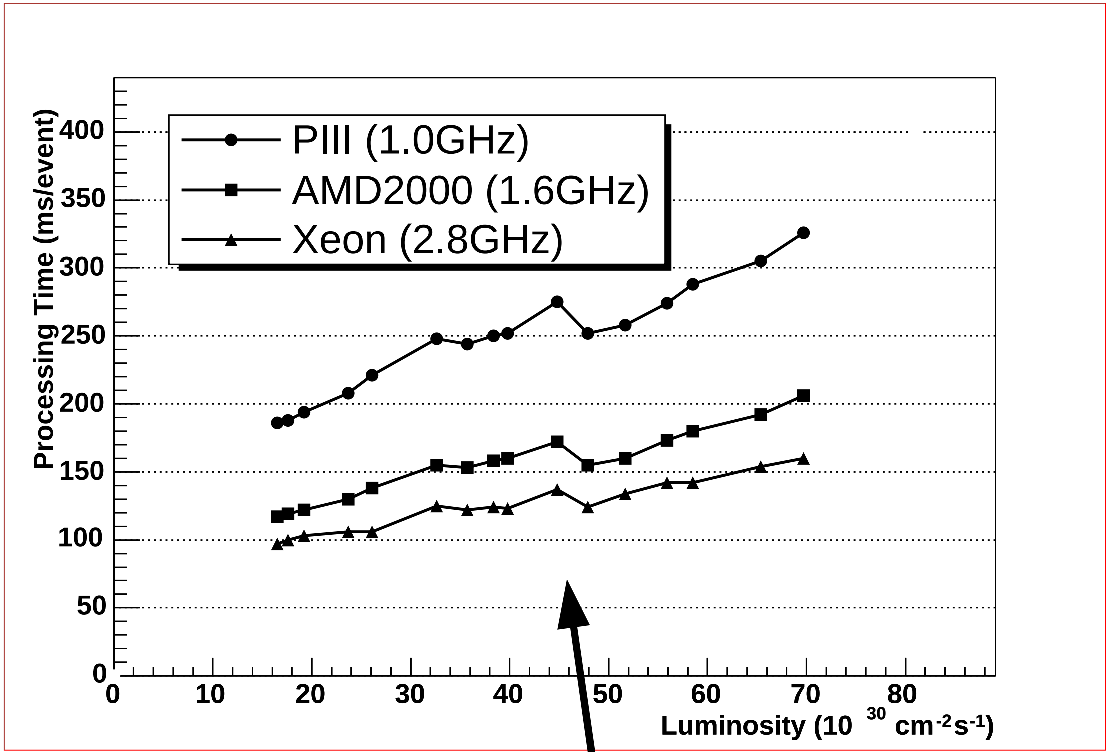
- Mysterious packet loss with some driver/kernel combos:
  - 2.4.20 and non-Intel (Becker) driver
  - 2.4.26 and Intel driver
- Using 2.4.26 and non-Intel (Becker) driver successfully

L3 Farm Performance

Relative Performance

PIII 1GHz:	**1.0**
AMD2000 1.6GHz:	**1.6**
Xeon 2.8GHz:	**2.0**

- Scaling is non-linear with GHz (or AMD rating)
- Real limit probably memory bandwidth
  - most L3 Filter data structures are larger than cache size



Track triggers prescaled

Farm Running Experience

Farm node hardware breaks often

- Typically hard drives and CPU fans
  - Hard drives fail incrementally (frustrating)
- About one machine/month requires warranty service
- Minor problems about once/week
- Correlation with age and component quality
  - AMD2000 nodes are actually worst offenders

Software must assume nodes will crash / be unavailable

- SBCs, RM, and Supervisor: robust connection handling
  - Supervisor reassigns nodes dynamically
  - Farm nodes initiate connections to RM and SBCs
- Our version and control software not as clever (later slides)

Farm Upgrade Experience

Newest group of nodes

- 32 Xeon 2.8GHz (one rack) installed 3 months ago
- Procurement time about 4 months total: desire to delivery
  - Purchase fully assembled racks with on-site service from single vendor

Fermilab Computing Division experience valuable

- Very strict vendor requirements
  - Specify cable quality, certified memory, power/heat limits
  - Require some known good subcomponents (cat5 cable)
- Most problems caught in advance
  - Site visits by vendors!
  - At least one rack fully assembled and inspected at vendor site

L3DAQ Future Plans

A few SBCs to be replaced

- Gb ethernet versions needed
  - tracking occupancies will increase crate size beyond 20kB/crate
- Evaluation SBC ordered – expect no issues
  - Tests by others (CDF) claim 40MB/s sustained

Continuous farm upgrades

- Oldest nodes (PIII 1GHz) now 3yrs old
- Plan to replace oldest set of nodes every year
  - Can react quickly to needs
    - processing needs difficult to predict long-term

Rework versioning and control software

- Easier long-term management
- Less expert involvment
- Improve reliability

Versioning and Control

Current versioning and control system

- Relatively simple shell scripts, remote shell
- Some (not enough!) web interfaces
- Active-user scheme
  - User required to check system state and correct it
    - Specific nodes/SBCs passed to control scripts
  - Many opportunities to make mistakes
    - Especially with farm node failure/recovery frequency
    - Experts modifying farm nodes for tests

Plan underway

- Implement a state-machine on each node/SBC
- User specifies desired state of entire system
  - Stored in central DB – web configurable?
- State-machine runs appropriate scripts to reach state

Versioning and Control

State-machine candidate: **cfengine**

- Common opensource software package used to keep configurations up-to-date for large clusters
  - Set links, modify config files, sync files
- Can run arbitrary scripts
  - On boot, on request, or periodically

Initial software download to farm nodes

- L3 filter software package is ~300MB (180MB gzipped)
  - Executables, libraries, calibration data
- Chained rcp or rsysnc takes several minutes to complete
- Successfully tested **BitTorrent**
  - Balanced p2p filesharing system (like kazaa, napster)
  - Takes full advantage of node-to-node bandwidth
  - Test: 180MB gzip file to 32 nodes in under 45s!

Conclusions

D0 L3 DAQ in operation since May 2002

- Commodity-based system
- Meets/exceeds current needs of D0

Upgrade paths straightforward

- Replace subset of farm every year
- Limited as-needed front-end (SBCs) replacement

Ease-of-use improvments

- Versioning and Control software replacement