Real Time Conference 2007

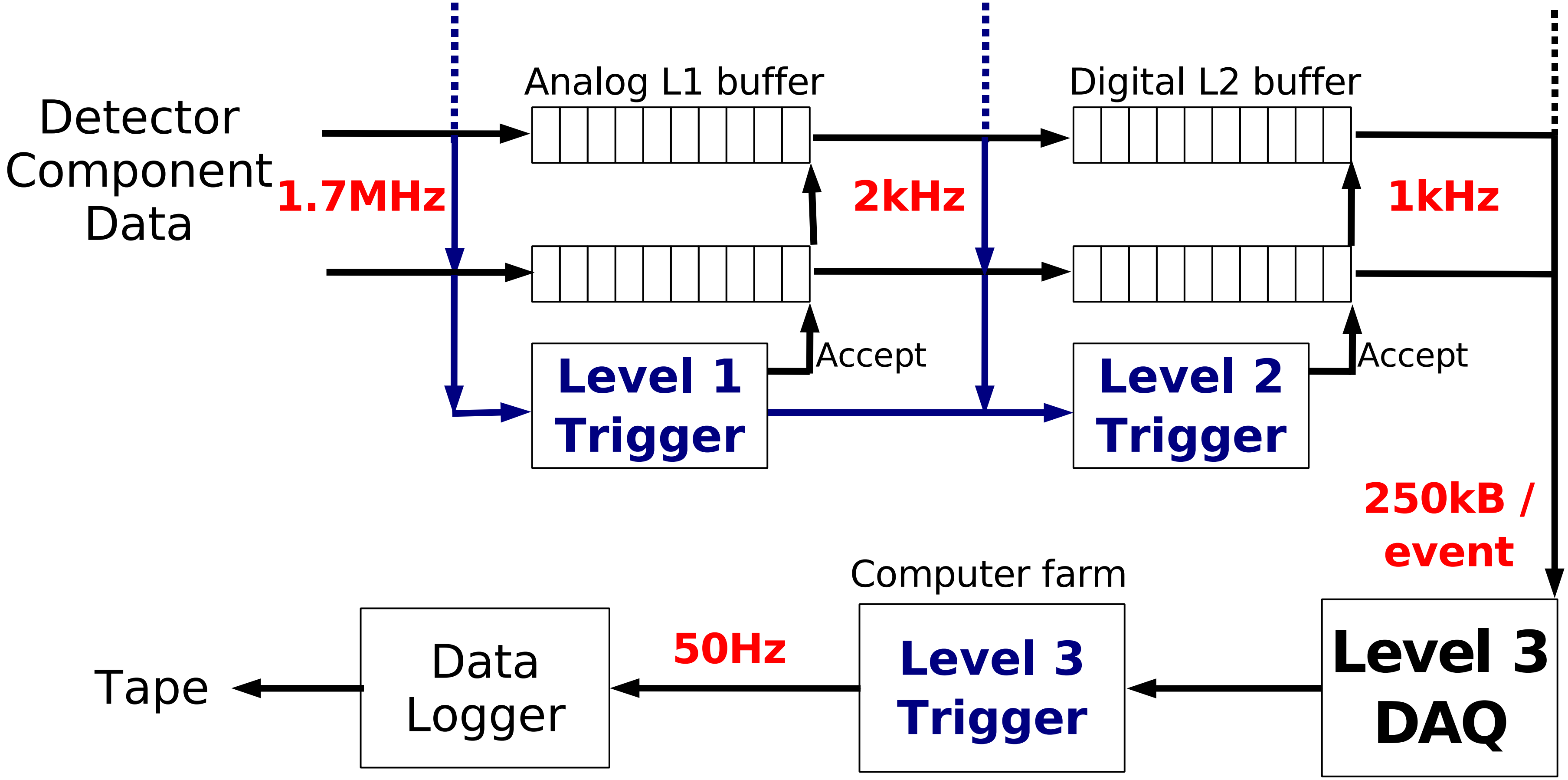
Fermilab, May 4, 2007

The DØ L3DAQ system: operation and upgrades

- Overview and design requirements
- System components
  - Commodity hardware
  - Operation and data flow control
- Upgrades and current performance
- Summary & conclusions

Arán García-Bellido
for the DØ L3DAQ group:
Brown University
FNAL-CD
University of Washington

The DØ data acquisition system



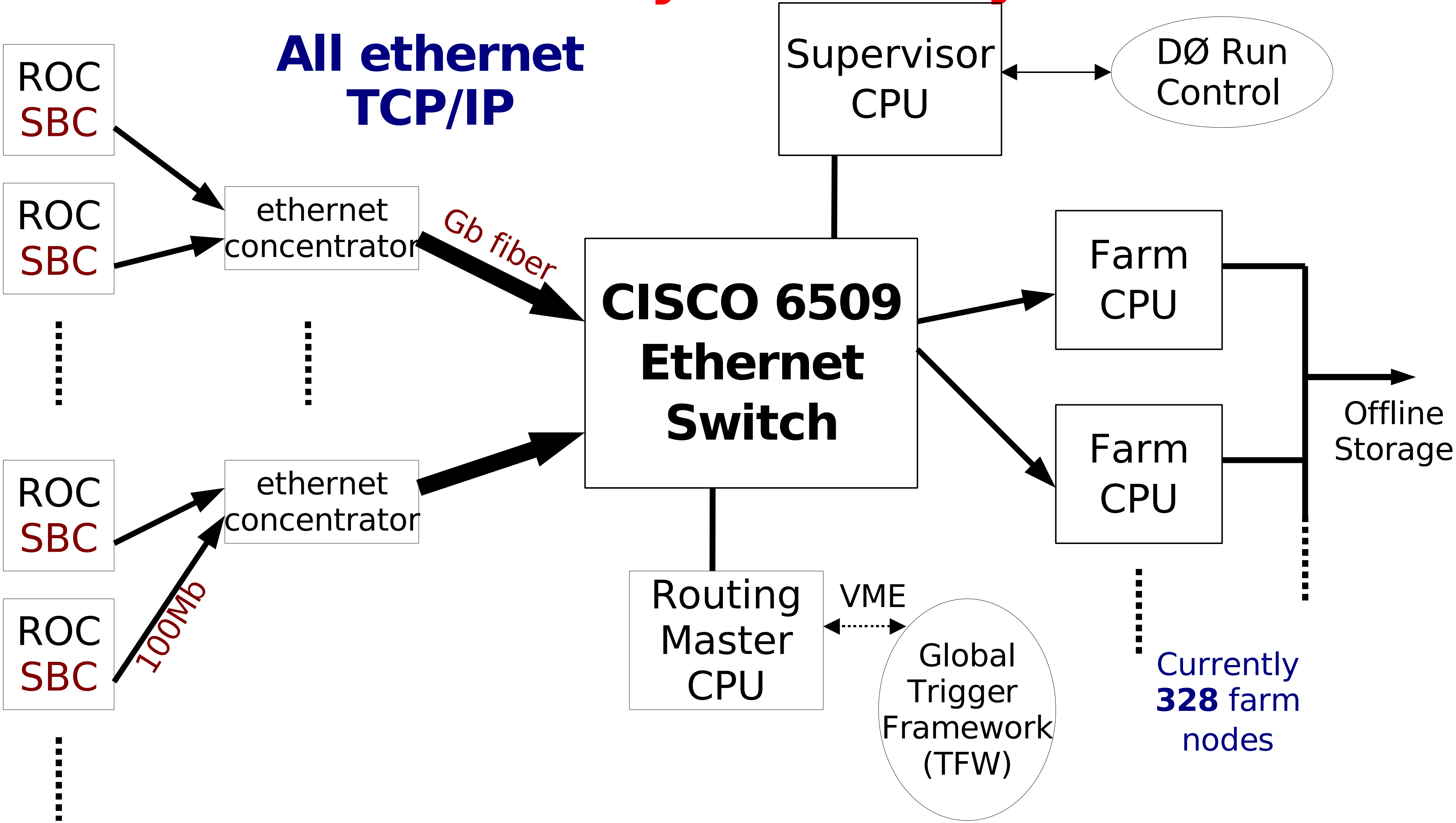
- Levels 1 and 2 are custom hardware
- L3/DAQ system is fully based on commodity hardware

  Transfer event fragments from readout crates to L3 farm, where full event is available and triggered on with offline-like algorithms

- Design requirements: Input 1kHz, with 250kB/event, output 50Hz

  Currently we operate normally at 300kB/event, output 100Hz

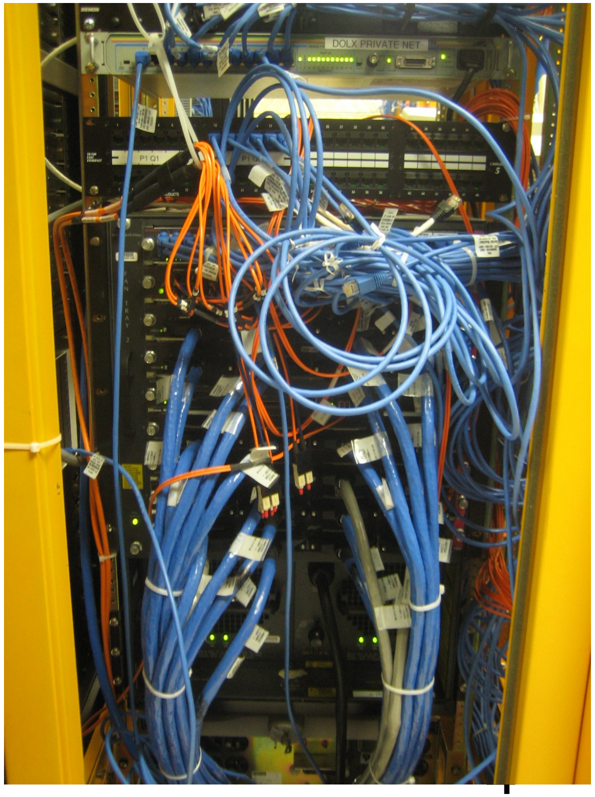
L3DAQ: system layout



- 63 total readout crates (ROC) and 5 Gb ethernet concentrators
- One single board VME computer (SBC) per crate
- 1-20kB data per crate in 1-10 modules

Components (all commodity hardware)

- **SBCs**: VMIC 7750, Pentium III 933 MHz
  - 128MB RAM, 128MB CompactFlash
  - VME to PCI Universe II module
  - Dual 100Mb ethernet (Intel eepro)
  - 3 with heavy load with 1000Mb ethernet
- **Routing Master**: VMIC 7850, P4M 1.7GHz
- **Farm nodes**: 328 total, all dual processor
  - Hyperthreaded Xeon 2.8 GHz (160)
  - Dual core AMD Opteron 1.8GHz (48)
  - Dual core Xeon 2.3 GHz (120)
  - Single 100Mb ethernet
- **CISCO 6509 switch**:
  - 16 Gb/s backplane
  - 9 module slots, currently full
  - 8 port Gb (fiber or copper)
  - 112MB shared output buffer per 48 ports



L3 DAQ operation

Partitioning: Simultaneous runs

- Allocate groups of nodes to each run

Flow control

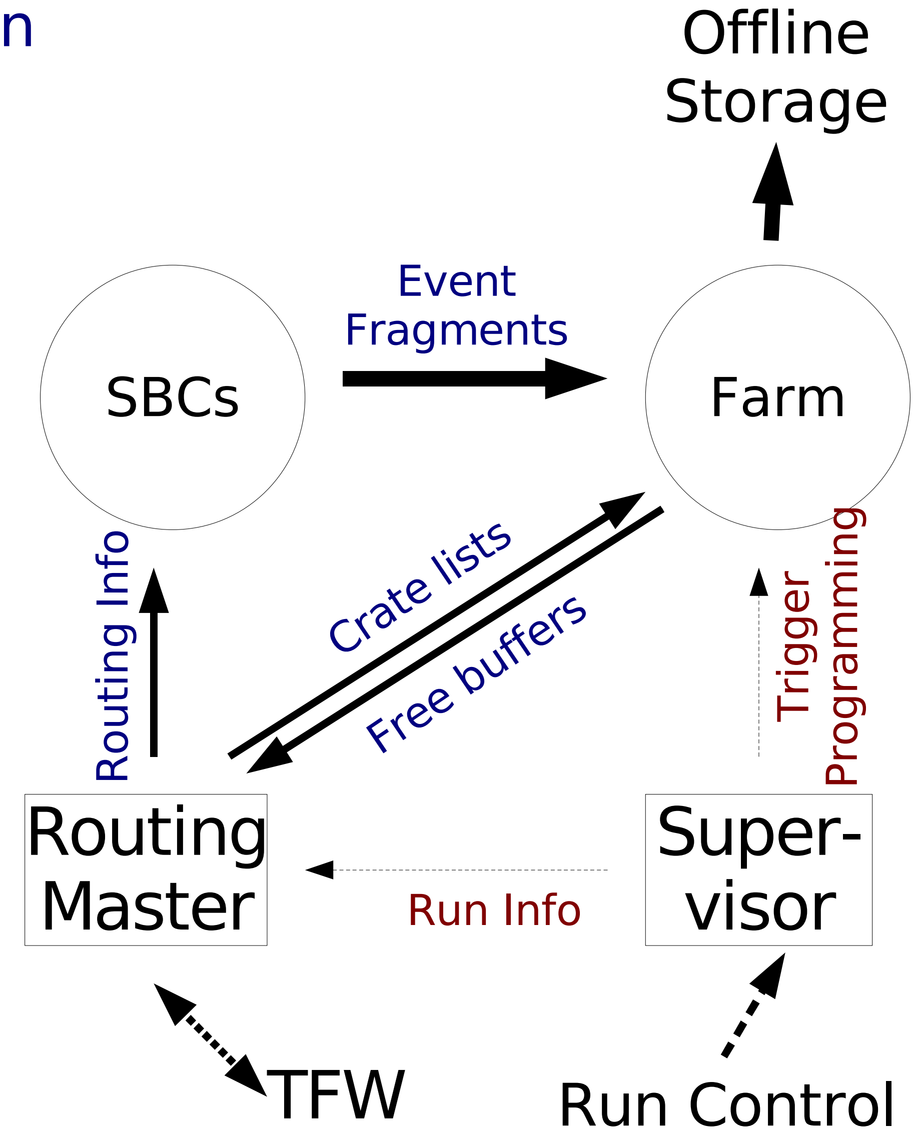
- Tune TCP settings to limit the amount of in-flight data
- Avoids packet-loss in switch
- Advertised buffers in nodes limit number of in-flight events
- Disable triggers if farm fills up

Software

- Linux OS on SBCs and farm
- C++ and shell scripts

Monitoring: Server architecture

- Data format is XML
- Heavily multithreaded to handle large number of sources and displays



Event buffering

Routing master

- Buffer 10 event tags (routing info) before sending to each SBC to minimize ethernet overhead
- Without buffering:
  63 crates × 1kHz = 60,000 packets/s

SBC

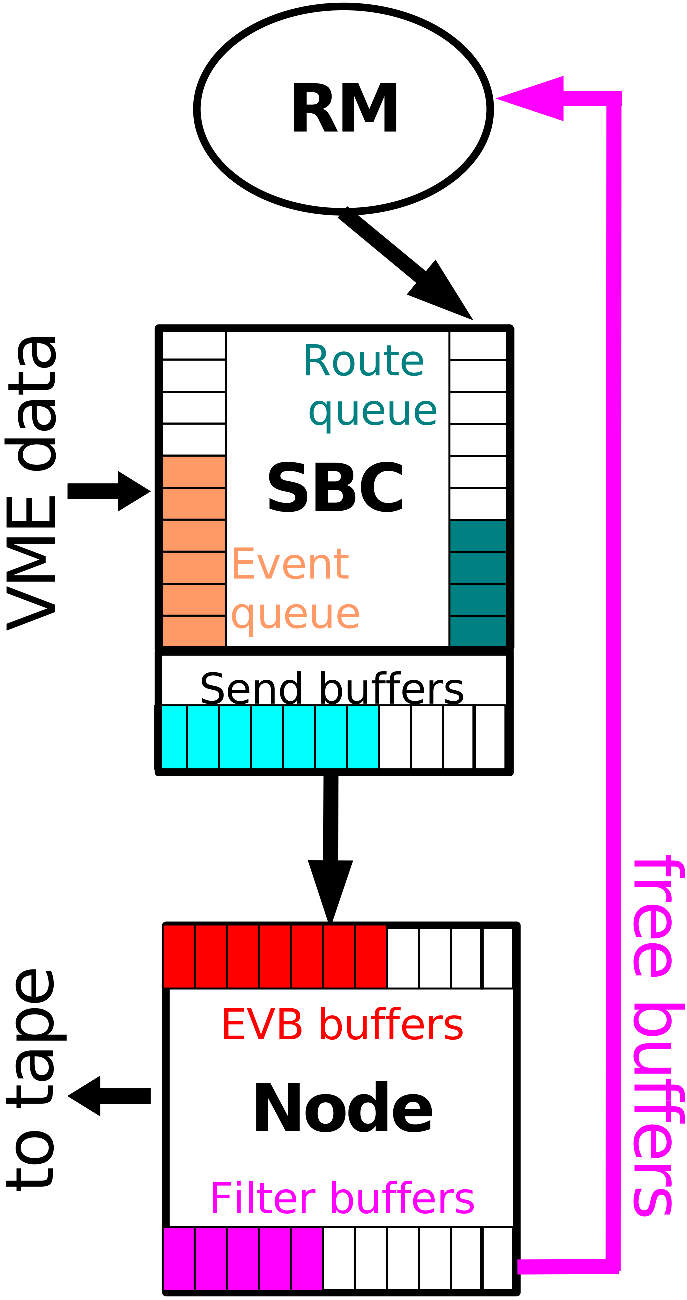
- Buffer 50 event fragments before routing
- 10 for RM event tag buffer and 40 for TFW FIFO depth
- Large (1MB) TCP/IP send buffer

Farm node Event Builder (concatenates fragments)

- 20 buffers (event processing)
- Advertise a maximum of 6 free buffers to RM

6509 switch

- 7 slots (each with 112MB shared output buffer for 48 nodes)
- 6 buffers * 48nodes/module * 300kB/event = ~86 MB max in transit through each module



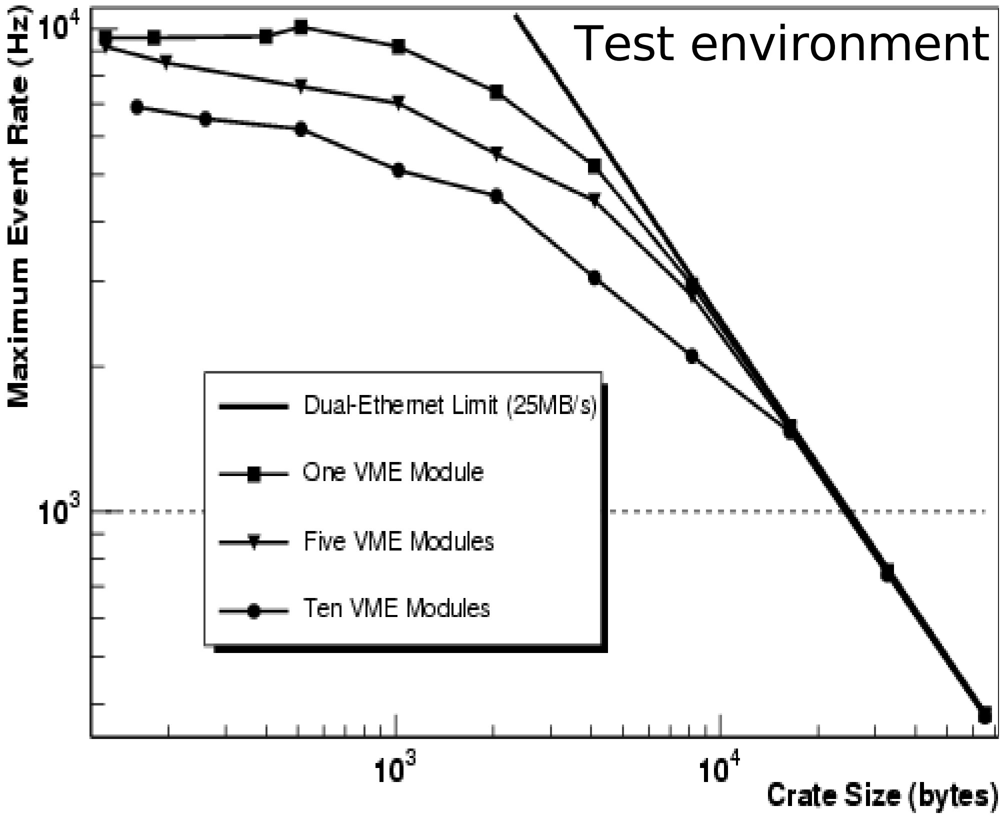
SBC performance

SBC Operation

- Very reliable hardware
  - 1 replacement/year
- Customized Linux kernel
  - Executes the VME reads
  - Configurable depending on crate type
  - Event fragment buffering
- User level process matches route info to fragments and sends to node

Have 3 different regimes based on crate payload:

- single-ethernet if crate size <10MB/s
- dual-ethernet if crate size is <20MB/s
  - two connections from each farm node
  - toggle sending between connections
- Gb-ethernet if crate size is >20MB/s
  - Three crates have peaks of ~200MB/s



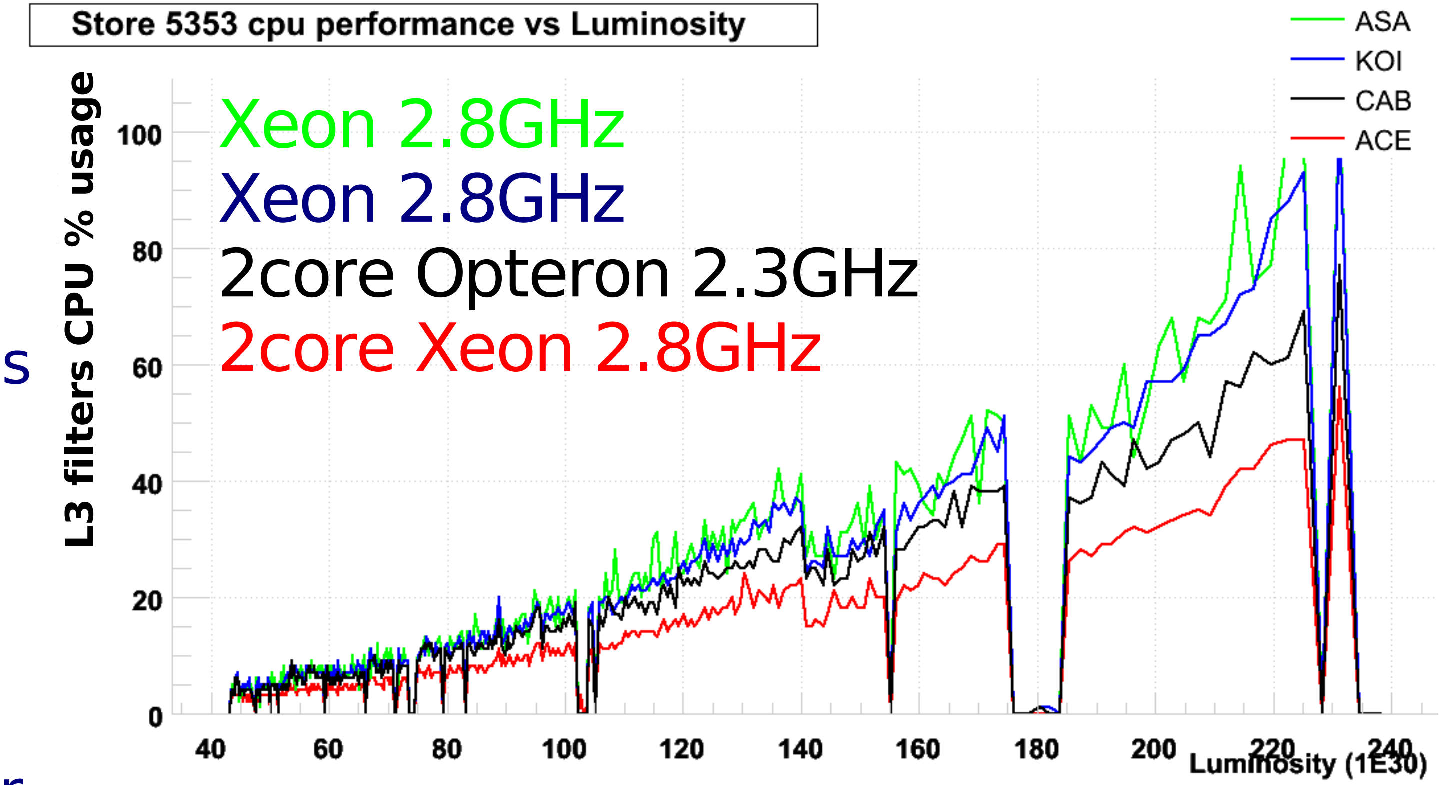
Test environment

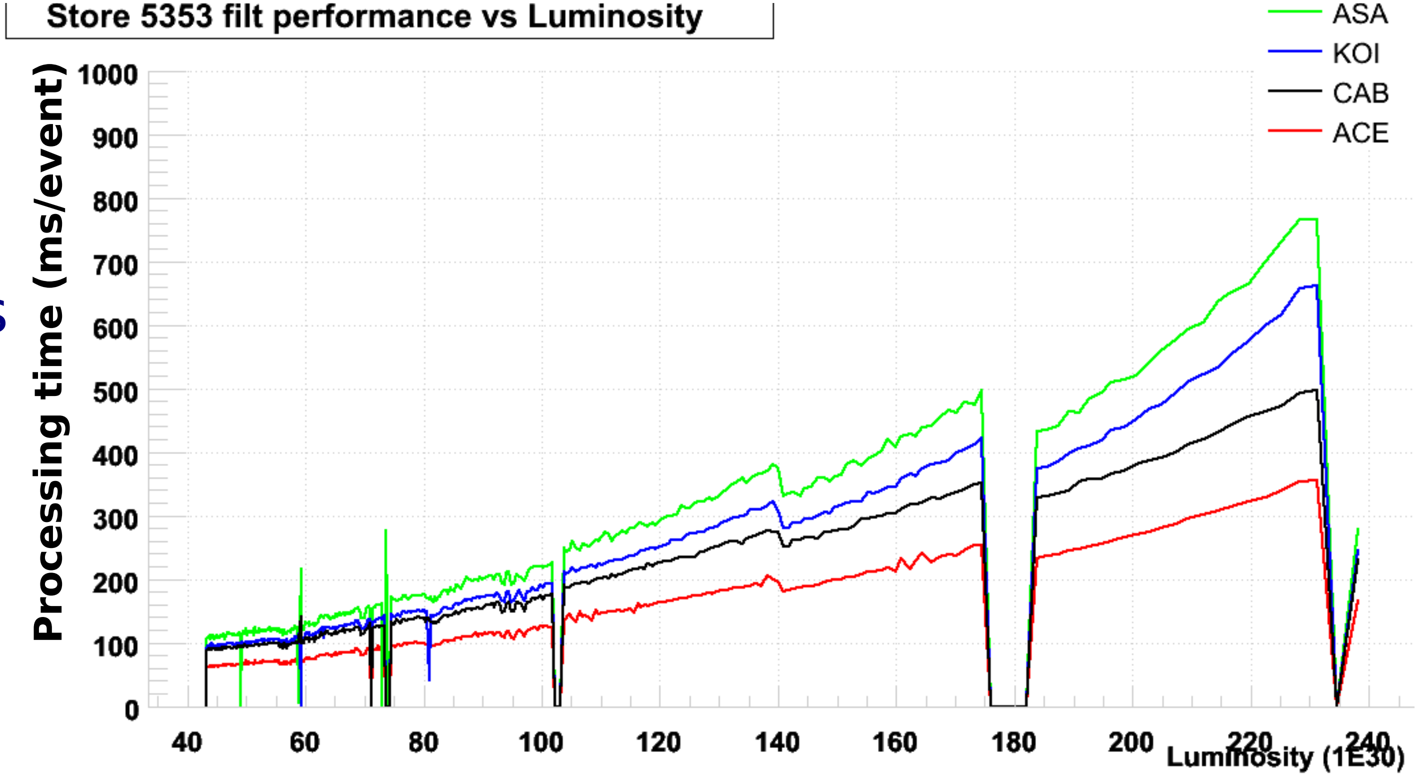
Limits

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

Nodes performance

- Dual processor hyperthreaded nodes have three L3 filter processes running (18% more efficient than two L3 filters)
- Dual processor dual core nodes have four L3 filters running
- Scaling with luminosity differs
- Memory bandwidth is also a factor
- Dual core dual processors are more robust at highest luminosities





Arán García-Bellido (UW)

Farm running experience

Farm node hardware breaks often

- Minor problems: few/week
- Warranty service: around one machine/month
- Typically hard drives and CPU fans
- FNAL Computing Division in charge of maintenance

Software must assume nodes will crash/be unavailable

- Supervisor process reassigns nodes dynamically
- Farm nodes initiate connections to RM and SBCs
- Version of L3 filter software to run is set manually

Upgrade of the farm: from 82 to 328 nodes and beyond

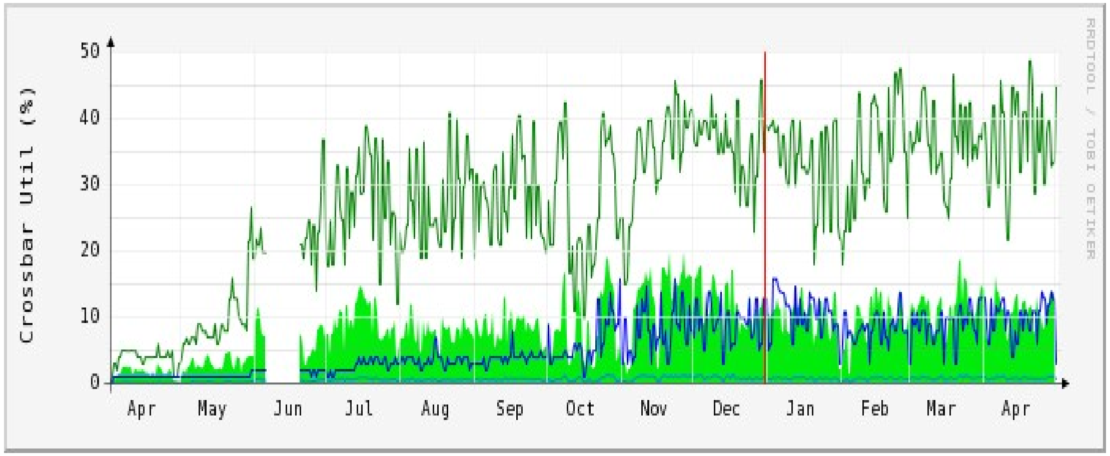
- FNAL CD experience is very valuable
- Strict vendor requirements
- Purchase fully assembled racks with on-site service from vendor
- Copying new versions of the L3 filter executables (300MB with rsync) to 328 nodes is painfully slow!

Overall performance

Recent typical store

- Start at ~900Hz input with 300kB, finish at ~300Hz with 250kB

Cisco switch



- Max utilization peaks at ~50% in the module with all the Gb connections
- All other modules peak at 35%

Routing Master

- Decision is made and sent in under 1 ms
- CPU usage at 1 kHz is ~50%, maxed out at 1.4 kHz in a test environment

SBC operation

- Crates with 20kB frag. size result in ~80% CPU utilization at 1 kHz
- RAM memory could be a problem if many more nodes added

Upgrades & new ideas

Farm upgrades

- Phase out old nodes when warranty expires
- New more powerful nodes added at current market standard
- Processing needs are difficult to predict long-term
- Evaluation of current “power” as a function of luminosity helps extrapolate future needs
- May need new slot(s) for CISCO 6509 switch

SBC upgrades

- VMIC 7805 with Gb ethernet was tested and works fine
- New model could replace old SBCs with dual ethernet

New ideas (very preliminary)

- Trigger leveling: store events in the node hard drive at the beginning of the stores and process them when the pressure on the farm is less, an hour or so later
- Share farm with offline Monte Carlo production

Conclusions

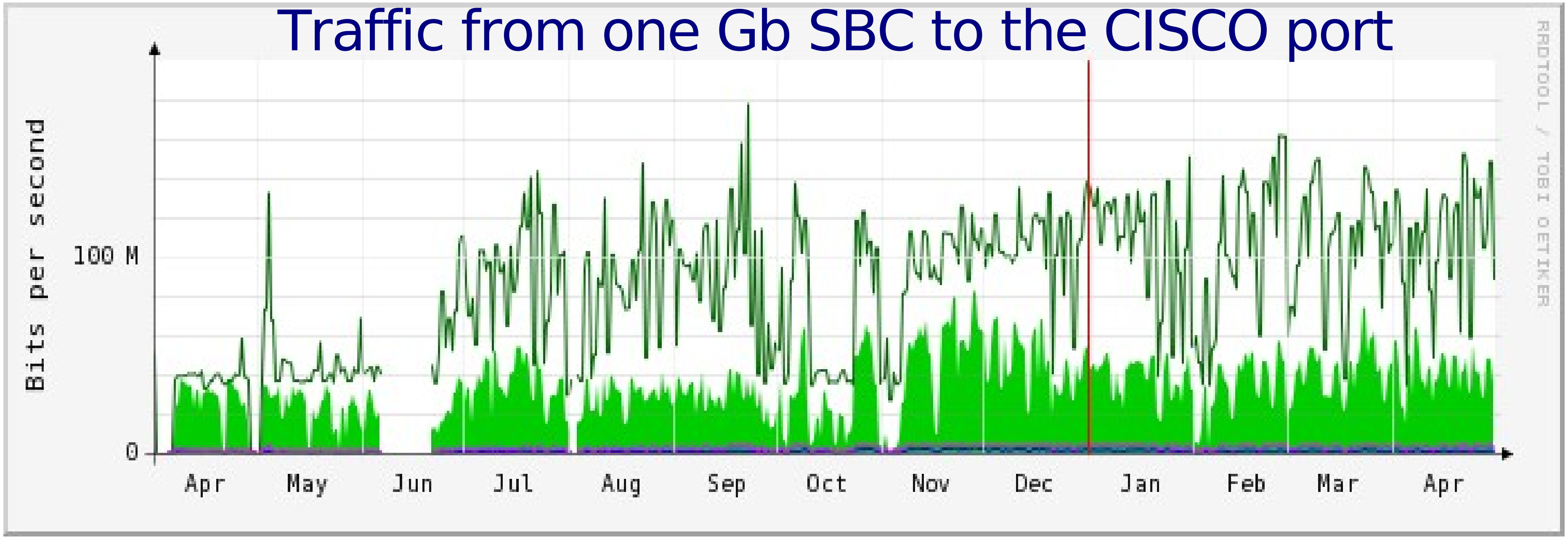
- DØ L3 DAQ built from commodity hardware
- 63 VME sources to 328 node processor farm
- Input: up to 350kB events at 1kHz (or 350MB/s)
- Based on Ethernet and TCP/IP communication
- Stable, reliable, expandable:
  - Successfully expanded from 80 to 328 nodes
  - Two-core chips in use, curb the processing time
  - Were able to double the output rate (50 to 100Hz)
- More upgrades straightforward
  - Replace subset of farm or add new ones
  - Front-end SBCs replacement if needed
- Keep improving to meet the needs of DØ

Extra Slides

More information:

http://www-d0online.fnal.gov/www/groups/l3daq/

SBC with Gb link



Traffic from one Gb SBC to the CISCO port

Green: Incoming traffic

Dark green: Peak incoming traffic