“More reliable than an airline*”

* Or the GRID

Fermilab

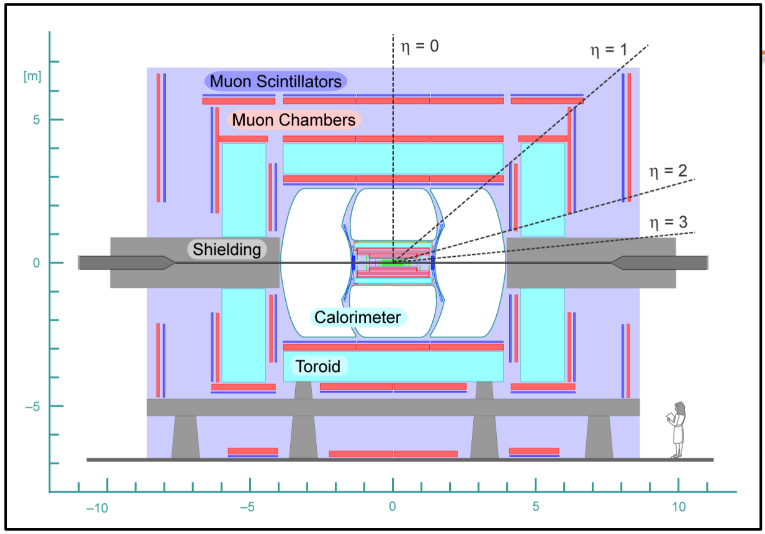
UNIVERSITY OF WASHINGTON

DØ LEVEL 3 TRIGGER/DAQ SYSTEM STATUS

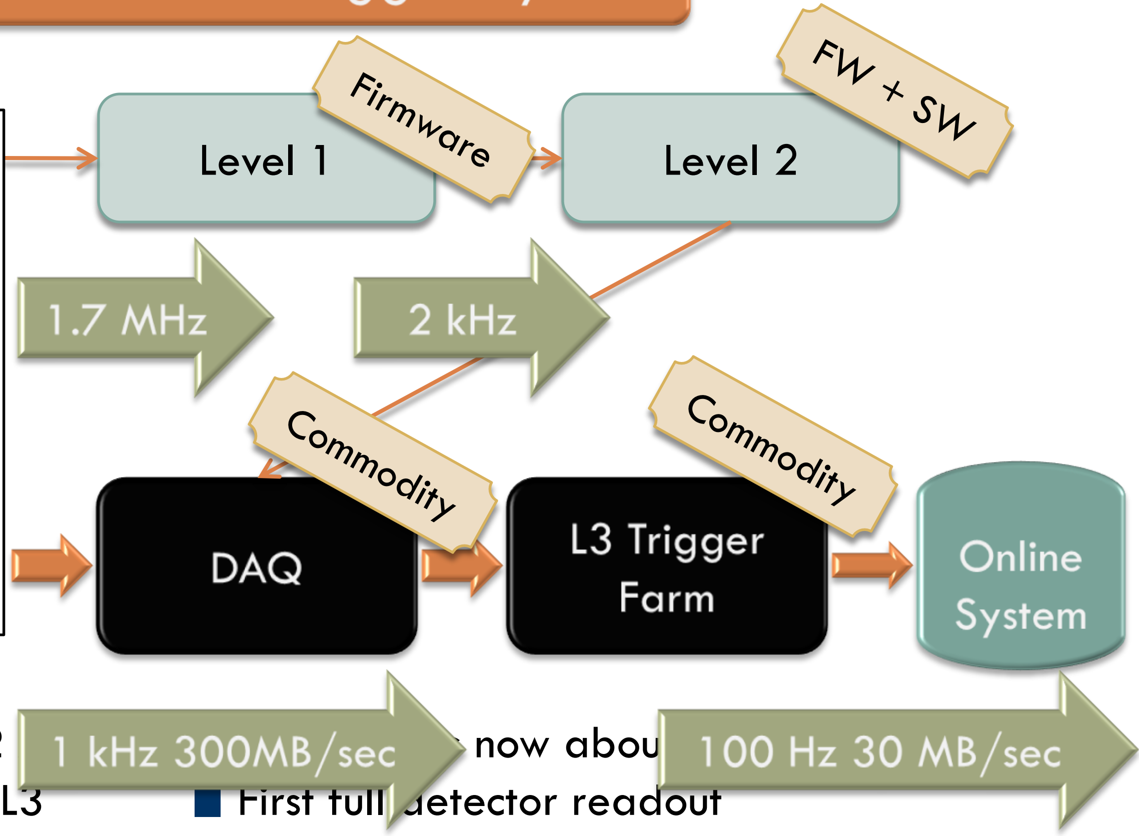
CHEP'07 VICTORIA, BC

G. Watts (for the DØ L3/DAQ Group)

Overview of DØ Trigger/DAQ



Standard HEP Tiered Trigger System



- Full Detector Readout After Level 2
- Single Node in L3 Farm makes the L3 Trigger Decision
- Standard Scatter/Gather Architecture

now abou

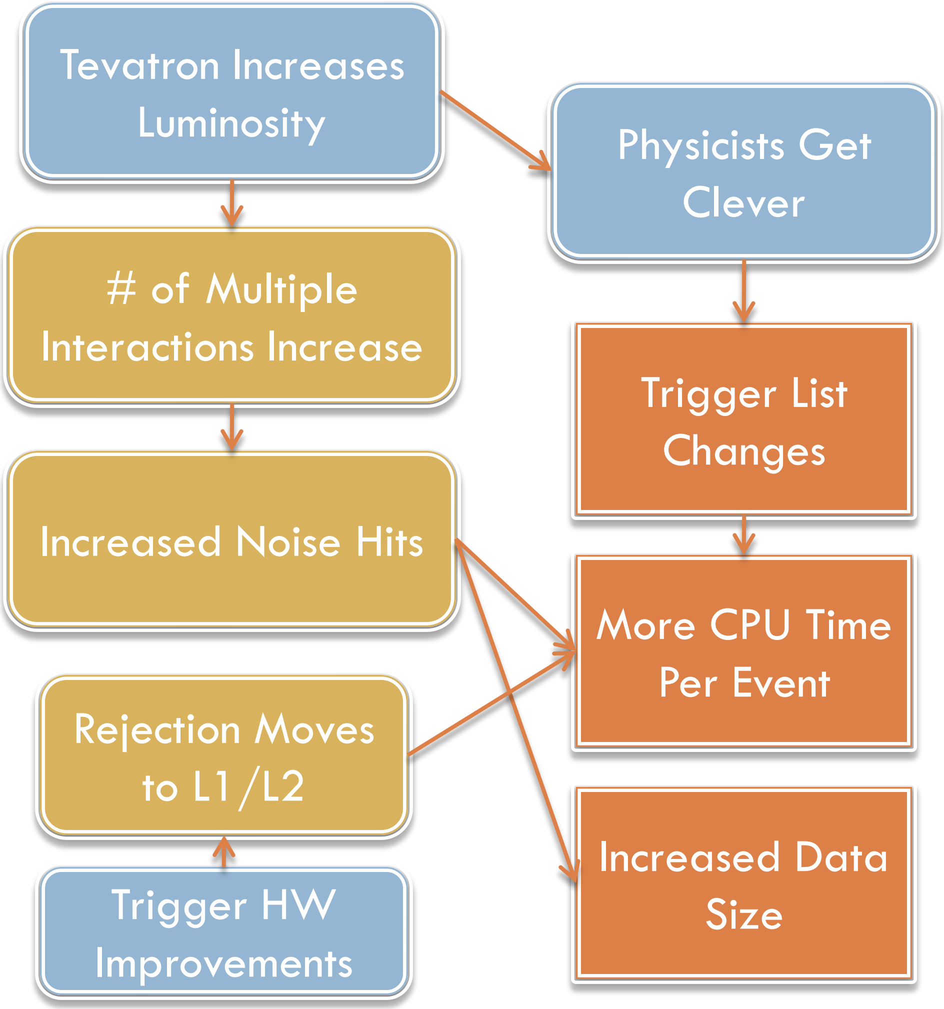
- First full detector readout
  - L1 and L2 use some fast-outs

Overview Of Performance

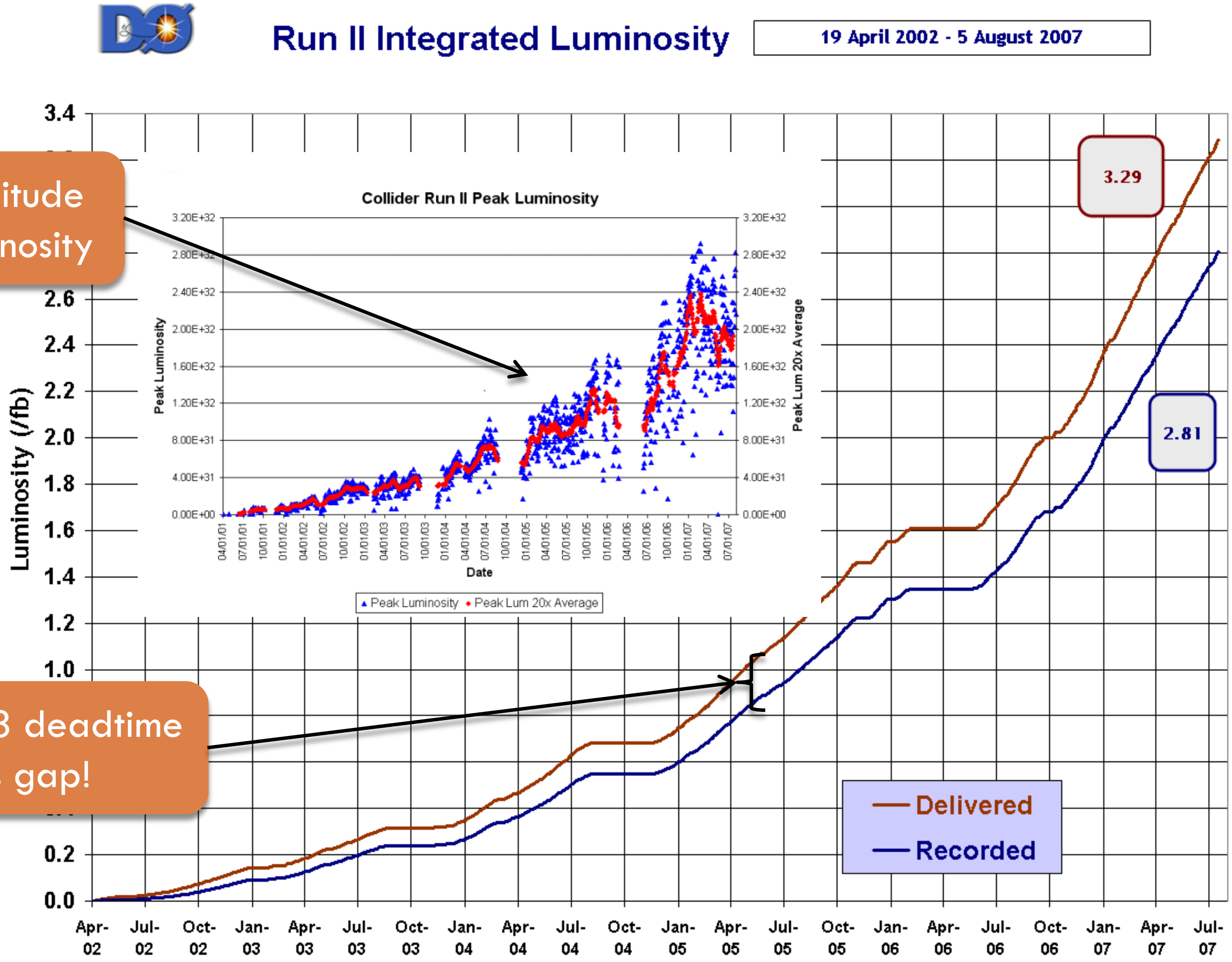
System has been fully operational since March 2002.

- Trigger software written by large collection of non-realtime programmer physicists.
  - CPU time/event has more than tripled.
- Continuous upgrades since operation started
  - Have added about 10 new crates
  - Started with 90 nodes, now have over 300, none of them original
  - Single core at start, latest purchase is dual 4-core.
- No major unplanned outages

An Overwhelming Success



24/7



Over order of magnitude increase in peak luminosity

Constant pressure: L3 deadtime shows up in this gap!

Basic Operation

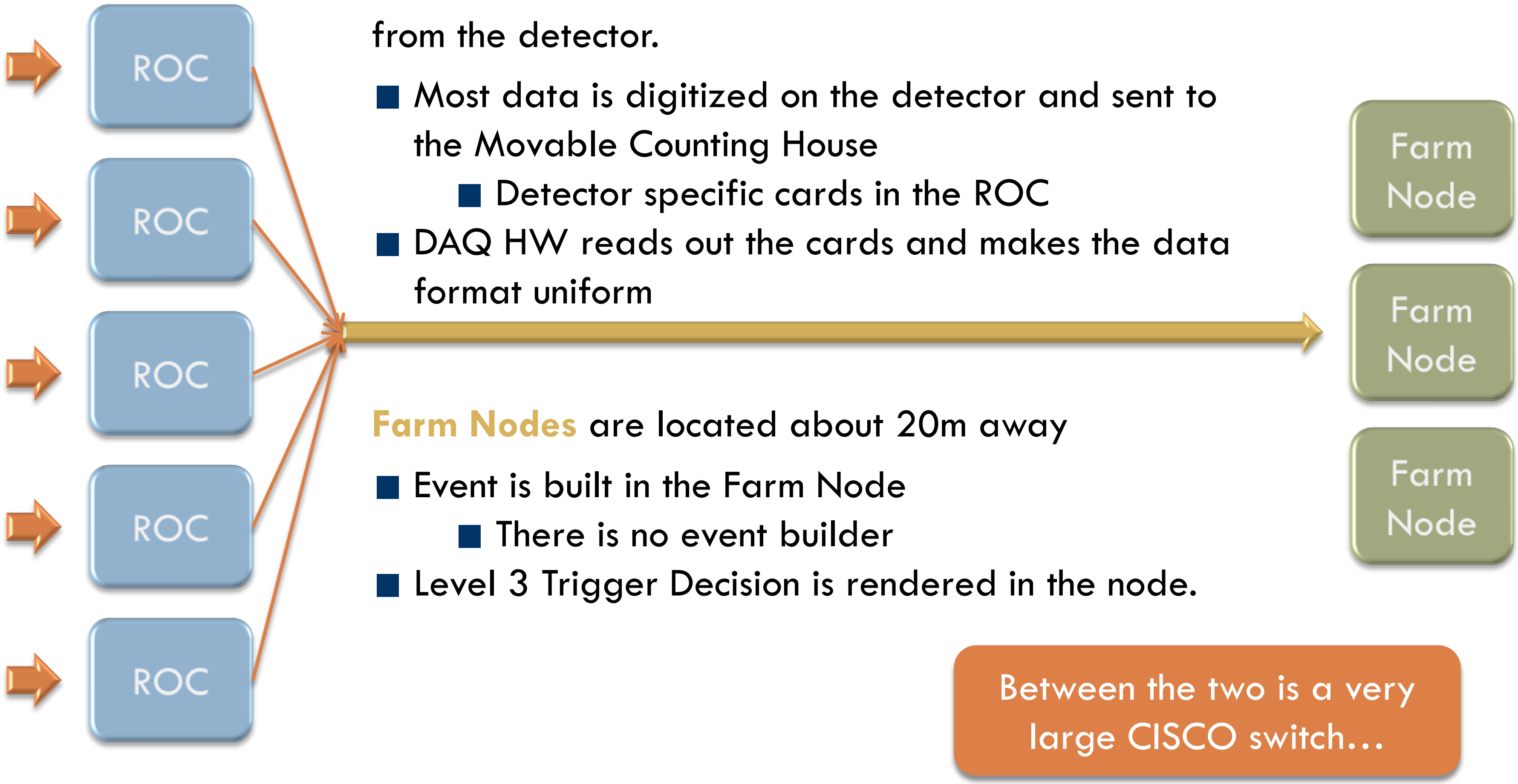
Data Flow

- Directed, unidirectional flow
- Minimize copying of data
- Buffered at origin and at destination

Control Flow

- 100% TCP/IP
- Bundle small messages to decrease network overhead
- Compress messages via configured lookup tables

The DAQ/L3 Trigger End Points



Read Out Crates are VME crates that receive data from the detector.

- Most data is digitized on the detector and sent to the Movable Counting House
  - Detector specific cards in the ROC
- DAQ HW reads out the cards and makes the data format uniform

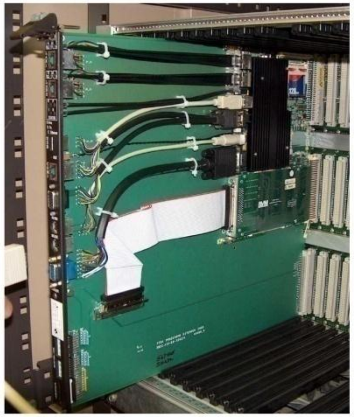
Farm Nodes are located about 20m away

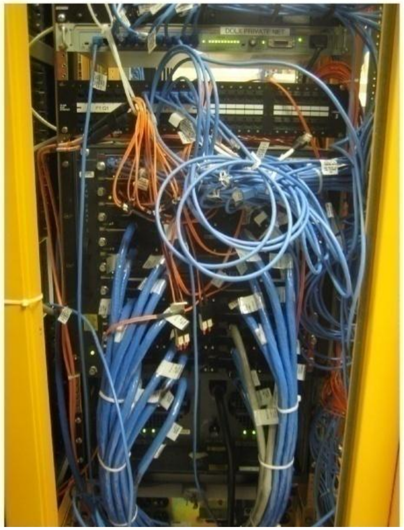
- Event is built in the Farm Node
  - There is no event builder
- Level 3 Trigger Decision is rendered in the node.

Between the two is a very large CISCO switch…

Hardware

- ROC's contain a Single Board Computer to control the readout.
  - VMIC 7750's, PIII, 933 MHz
  - 128 MB RAM
  - VME via a PCI Universe II chip
  - Dual 100 Mb ethernet
  - 4 have been upgraded to Gb ethernet due to increased data size
- Farm Nodes: 328 total, 2 and 4 cores per pizza box
  - AMD and Xeon's of differing classes and speeds
  - Single 100 Mb eithernet
- CISCO 6590 switch
  - 16 Gb/s backplane
  - 9 module slots, all full
  - 8 port GB
  - 112 MB shared output buffer per 48 ports

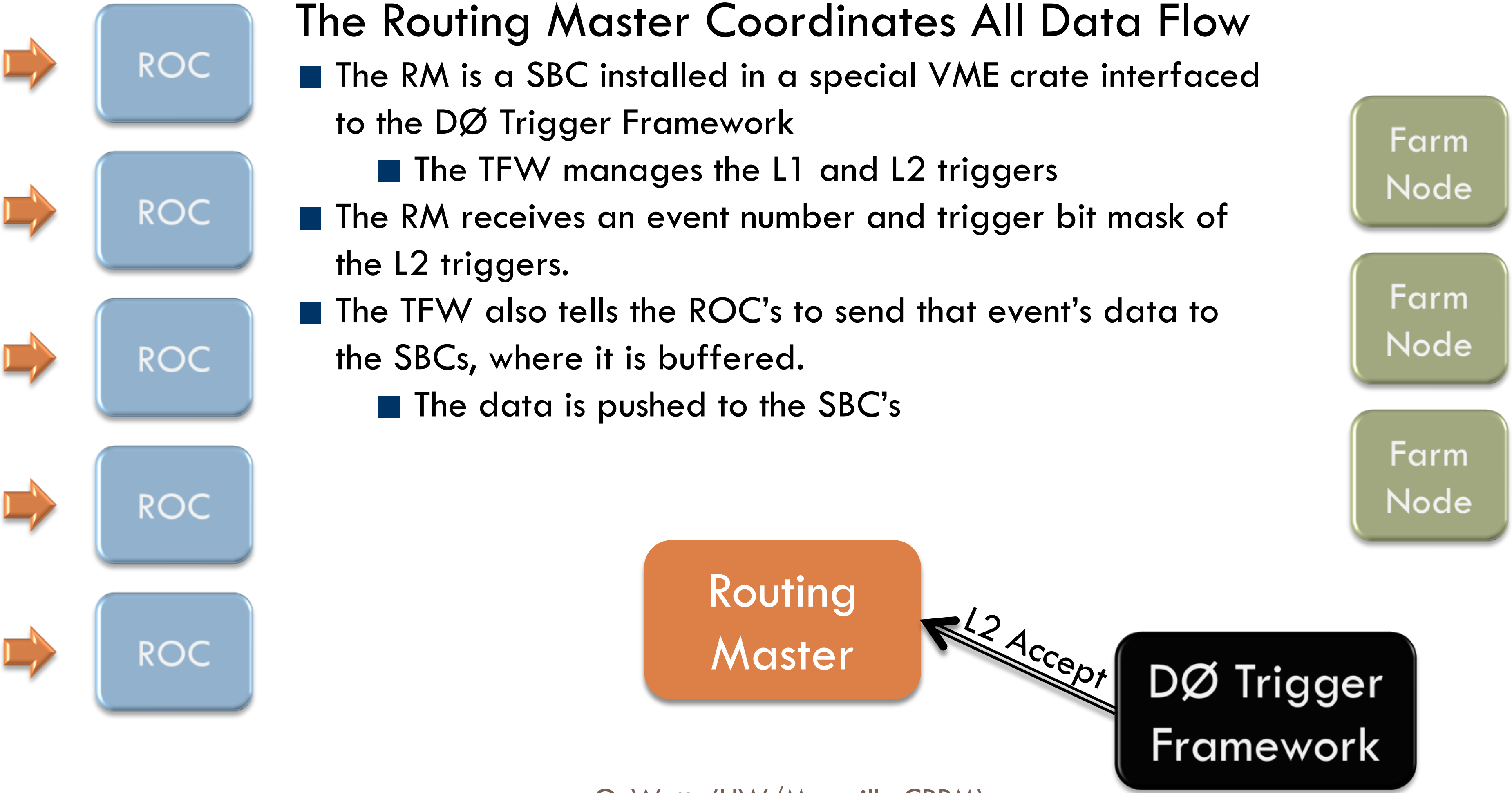




Data Flow

The Routing Master Coordinates All Data Flow

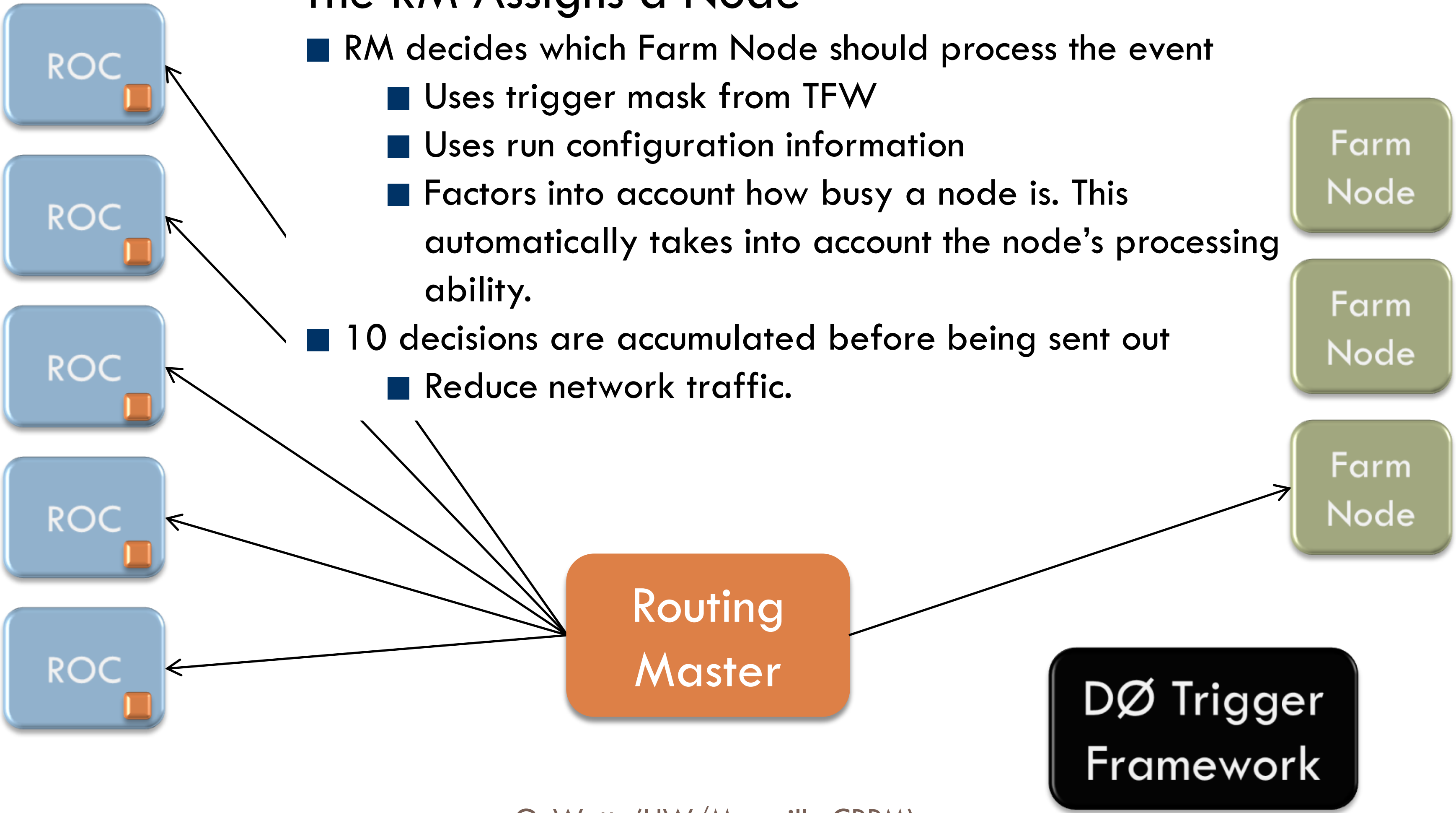
- The RM is a SBC installed in a special VME crate interfaced to the DØ Trigger Framework
  - The TFW manages the L1 and L2 triggers
- The RM receives an event number and trigger bit mask of the L2 triggers.
- The TFW also tells the ROC's to send that event's data to the SBCs, where it is buffered.
  - The data is pushed to the SBC's



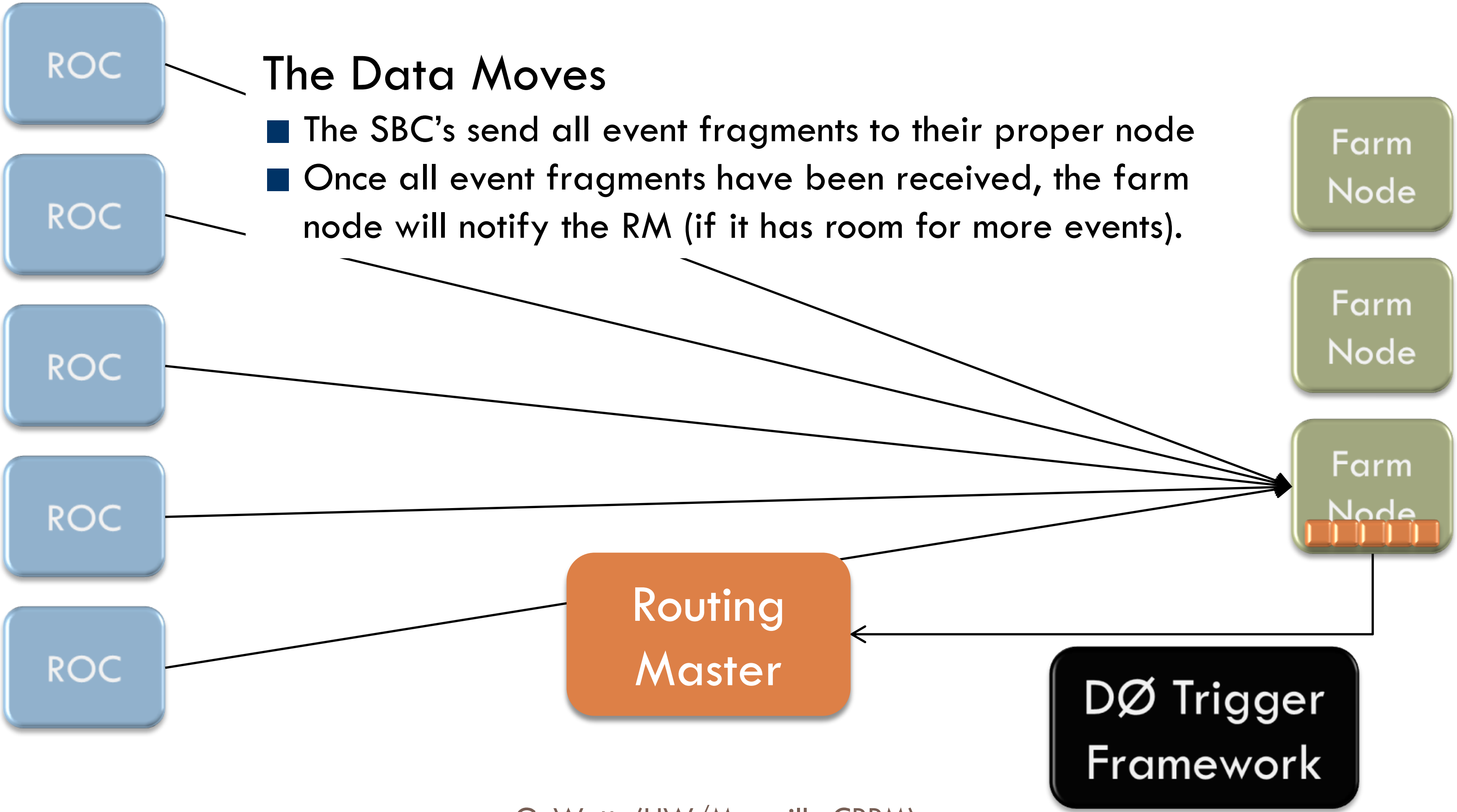
Data Flow

The RM Assigns a Node

- RM decides which Farm Node should process the event
  - Uses trigger mask from TFW
  - Uses run configuration information
  - Factors into account how busy a node is. This automatically takes into account the node's processing ability.
- 10 decisions are accumulated before being sent out
  - Reduce network traffic.



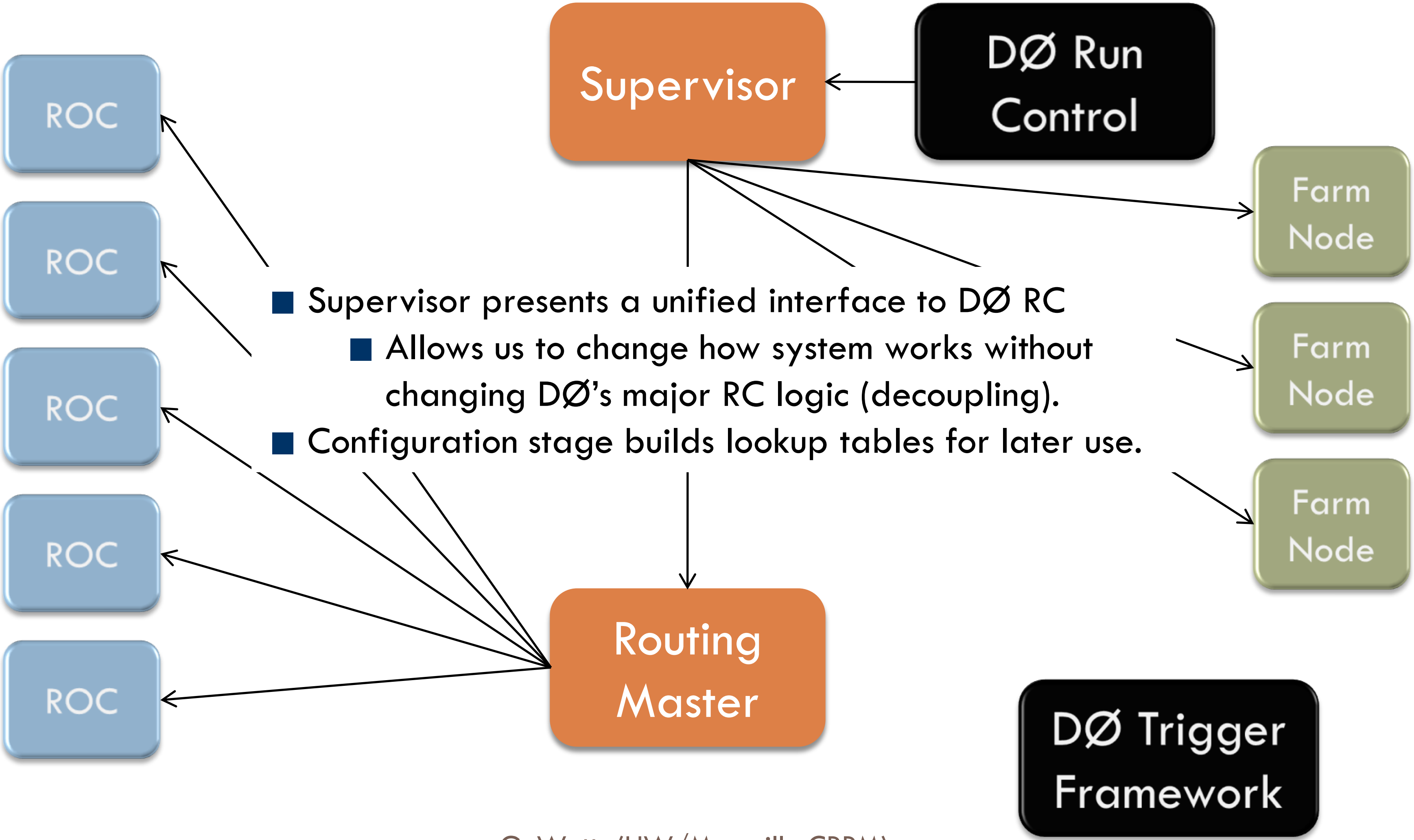
Data Flow



The Data Moves

- The SBC's send all event fragments to their proper node
- Once all event fragments have been received, the farm node will notify the RM (if it has room for more events).

Control Flow



- Supervisor presents a unified interface to DØ RC
  - Allows us to change how system works without changing DØ's major RC logic (decoupling).
- Configuration stage builds lookup tables for later use.

Performance

Single Board Computers

Farm Nodes

Data Buffering

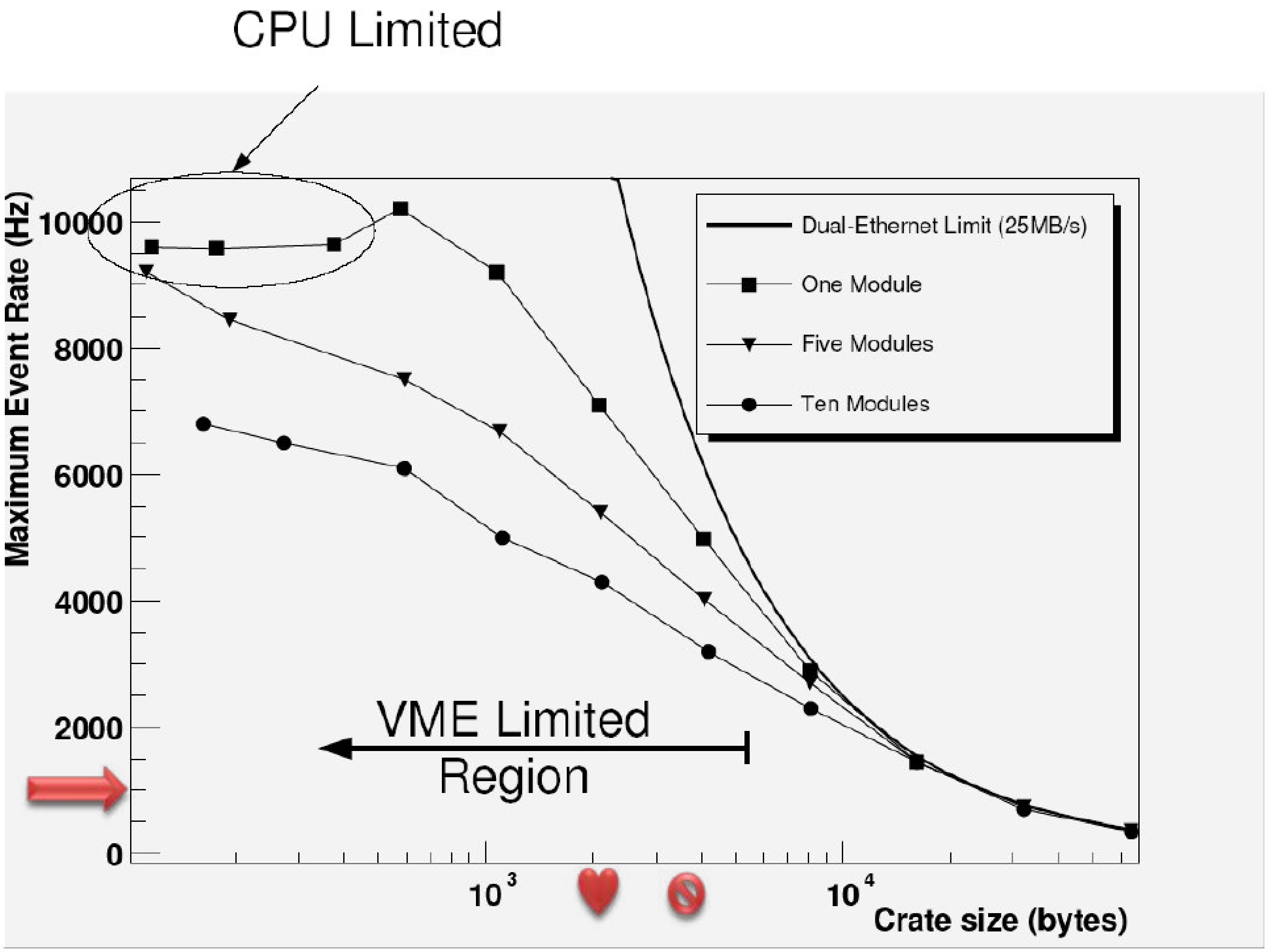
Single Board Computers

- Most Expensive and Reliable Hardware In System
  - We Replace about 1/year
  - Often due to user error
- Runs Stripped Down Version of Linux
  - Home brew device driver interacts with VME
  - User mode process collects the data, buffers it, and interacts with the RM
  - Code has been stable for years
  - Minor script changes as we update kernels infrequently.
- 3 networking configurations
  - <10 MB/sec: Single Ethernet port
  - <20 MB/sec: Dual Ethernet ports
    - Two connections from each farm node
  - > 20 MB/sec: Gb Ethernet connection
    - 3 crates have peaks of 200 Mb/sec
- Problems
  - Large number of TCP connections must be maintained
  - Event # is 16 bit; recovering from roll over can be a problem if something else goes wrong at the same time.

Single Board Computers

At 1 kHz CPU is about 80% busy

Data transfer is VME block transfer (DMA) via the Universe II module



Farm Nodes

- Run Multiple Copies of Trigger Decision Software
  - Hyper threaded dual processor nodes run 3 copies, for example.
  - The new 8 core machines will run 7-9 copies (only preliminary testing done).
  - Designed a special mode to stress test nodes in-situ by force-feeding them all data.
    - Better than any predictions we've done.
- Software
  - IOProcess lands all data from DAQ and does event building.
- FilterShell (multiple copies) runs the decision software
- All levels of the system are crash insensitive
  - If a Filter Shell crashes, new one is started and reprogrammed by IOProcess – rest of system is non-the-wiser.
- Software distribution 300 MB – takes too long to copy!

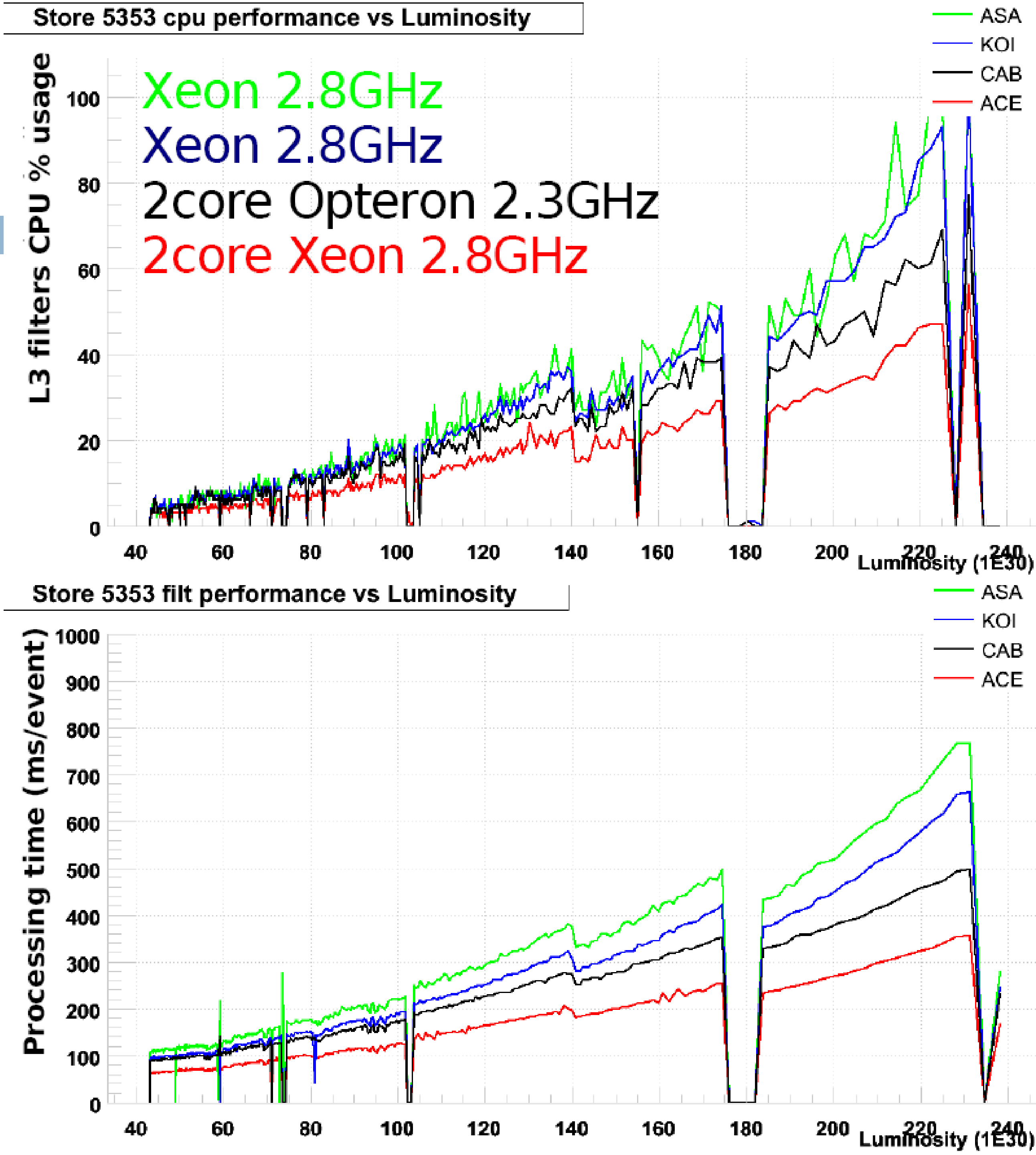
Farm Nodes

- Reliability
  - Minor problems: few/week
  - One/month requires warrantee service.
  - Enlisted help from Computing Division to run Farm
  - Well defined hand-off procedures to make sure wrong version of trigger software is never run.
  - Notice definite quality difference between purchase – tried to adjust bidding process appropriately.
  - No automatic node recovery process in place yet…
- Partition the Run
  - Software was designed to deal with at least 10 nodes
  - Some calibration runs require 1 node – special hacks added.
- Regular Physics uses the whole farm
  - Could have significantly reduced complexity of farm if we'd only allowed this mode of running.
- Network
  - Sometimes connections to SBC are dropped and not reestablished
    - Reboot of SBC or Farmnode required.
  - Earlier version of Linux required debugging of tcp/ip driver to understand latency issues.
- Log Files
  - Need way to make generally accessible

Farm Nodes

- Different behavior vs Luminosity
- Dual Core seems to do better at high luminosity
    - More modern systems with better memory bandwidth

CPU Time Per Event



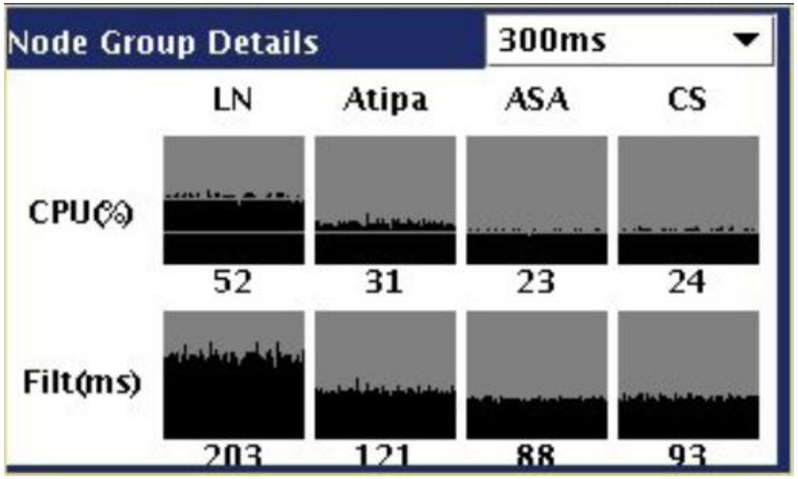
Event Buffering

SBC Buffering

- Event fragments are buffered until the RM sends a decision
  - RM buffers up to 10 decisions before sending them out
- We've never had a SBC queue overflow
- TCP/IP connection for each node
  - If we add lots more nodes, might need more memory

Farm Node Buffering

- RM bases node event decision on size of internal queue
  - Provides a large amount of buffering space
  - Automatically accounts for node speed differences without having to make measurements
  - The occasional infinite loop does not cause one node to accumulate an unusually large number of events.



Node Group Details 300ms

	LN	Atipa	ASA	CS
CPU(%)	52	31	23	24
Filt(ms)	203	121	88	93

Future & Conclusions

Upgrades

Farm Nodes

- Purchase of 8 core machines will arrive in a month
- Discard old nodes when warranty expires
  - 3-4 years: given their CPU power they are often more trouble than they are worth by that time.
- Original plan called for 90 single processor nodes
  - "Much easier to purchase extra nodes than re-write the tracking software from scratch"
- Hoping not to need to upgrade the CISCO switch

SBCs

- Finally used up our cache of spares
  - Purchasing a new model from VMIC (old model no longer available).
- No capability upgrades required

Other New Ideas

- Lots of ideas to better utilize CPU of farm during the low luminosity portion of a store
  - But CPU pressure has always been relived by "Moore's Law".
- Management very reluctant to make major changes at this point

Conclusion

- This DØ DAQ/L3 Trigger has taken every single physics event for DØ since it started taking data in 2002.
- 63 VME sources powered by Single Board Computers sending data to 328 off-the-shelf commodity CPUs.
- Data flow architecture is push, and is crash and glitch resistant.
- Has survived all the hardware, trigger, and luminosity upgrades smoothly
  - Upgraded farm size from 90 to 328 nodes with no major change in architecture.
- We are in the middle of the first Tevatron shutdown in which no significant hardware or trigger upgrades are occurring in DØ.
- Primary responsibility is carried out by 3 people (who also work on physics analysis), backed up by Fermi CD and the rest of us.