

# Bandwidth Improvements to the CDF Data Acquisition System

Frank Chlebana (Fermilab)

All Experimenter's Meeting

Apr 2 2007

- DAQ Overview

- Readout Deadtime

  - *TDC DSP Processing Limit*

  - *Data Size Limitations*

- Busy Deadtime

  - *Data Size Limitations*

  - *Event Builder*

  - *Data Logger*

# Trigger and Data Acquisition System

The online “trigger” is used to select an event rate of about 75 Hz from the 2.5 MHz (396 ns crossing) beam crossing rate.

## L1 Trigger (25 → 35 KHz)

Calorimeter, Muon, Forward Detectors and Tracking triggers (XFT)

Typically about 60 L1 triggers

## L2 Trigger (300 → 800 Hz)

Calorimeter, Muon and Impact parameter triggers (SVT)

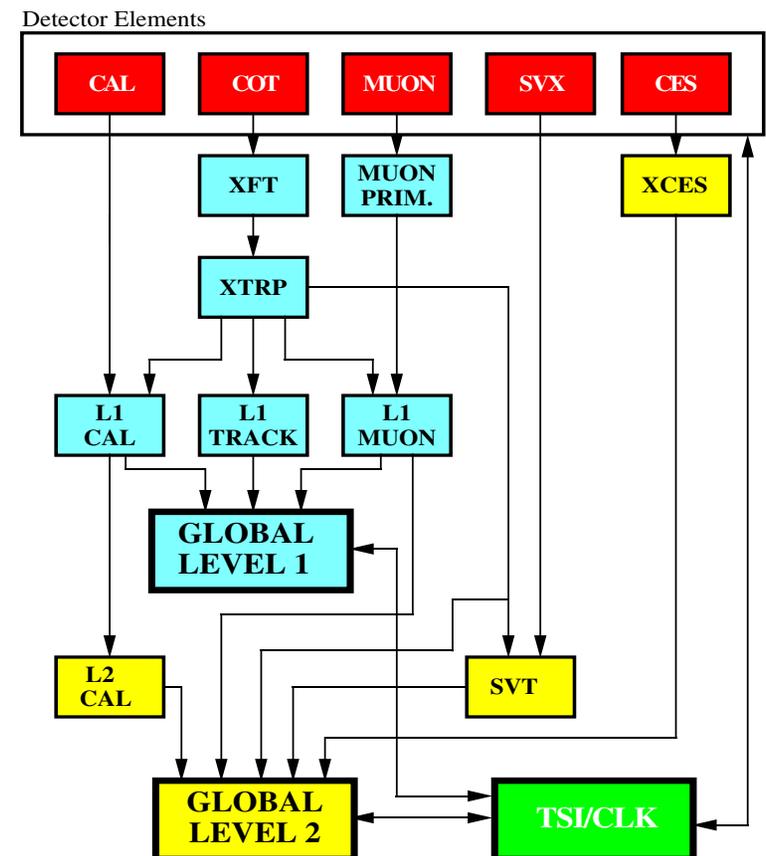
Typically about 130 L2 triggers

## L3 Trigger (24 → 100 MB/s)

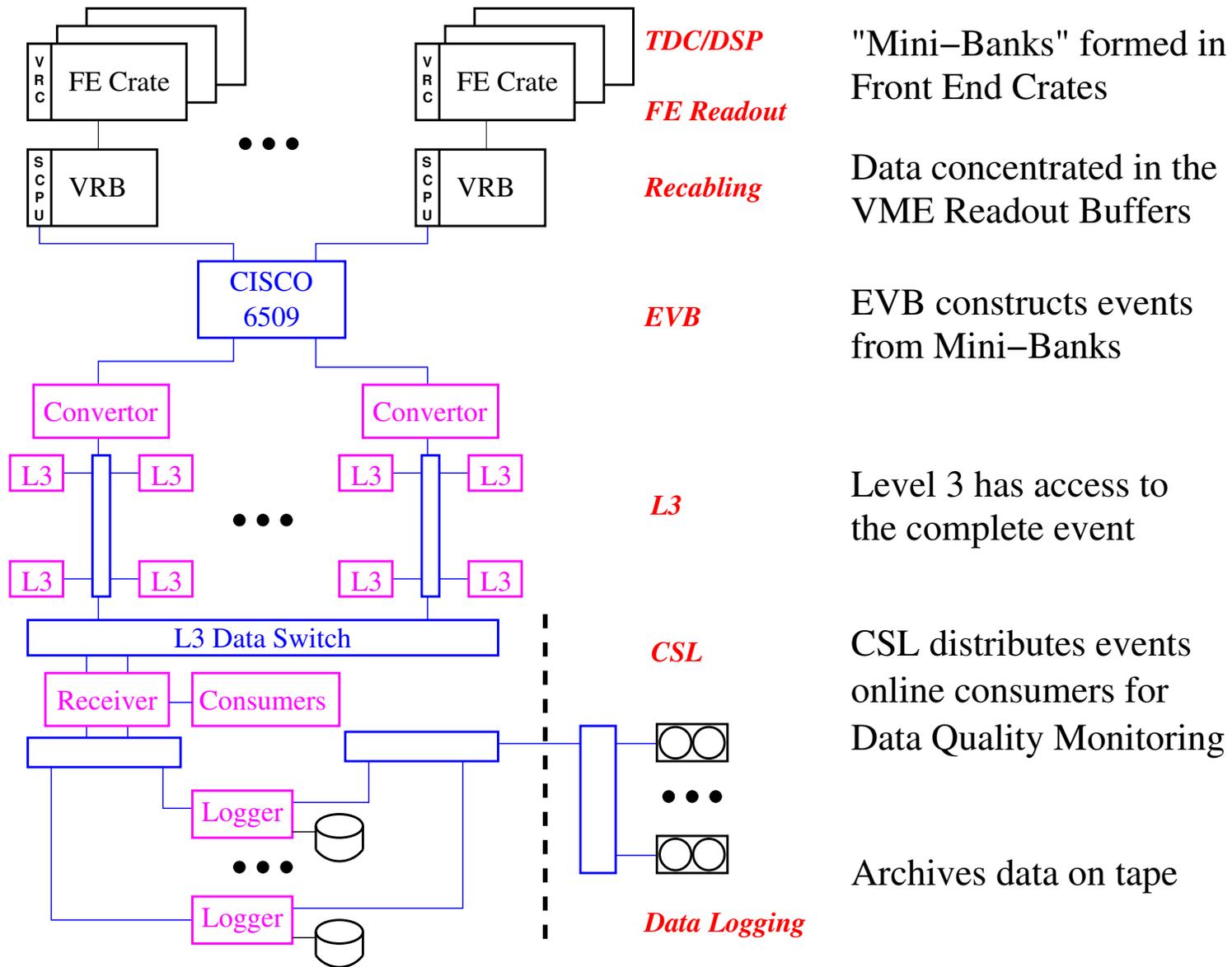
Full offline reconstruction

Typically about 182 L3 triggers

### RUN II TRIGGER SYSTEM



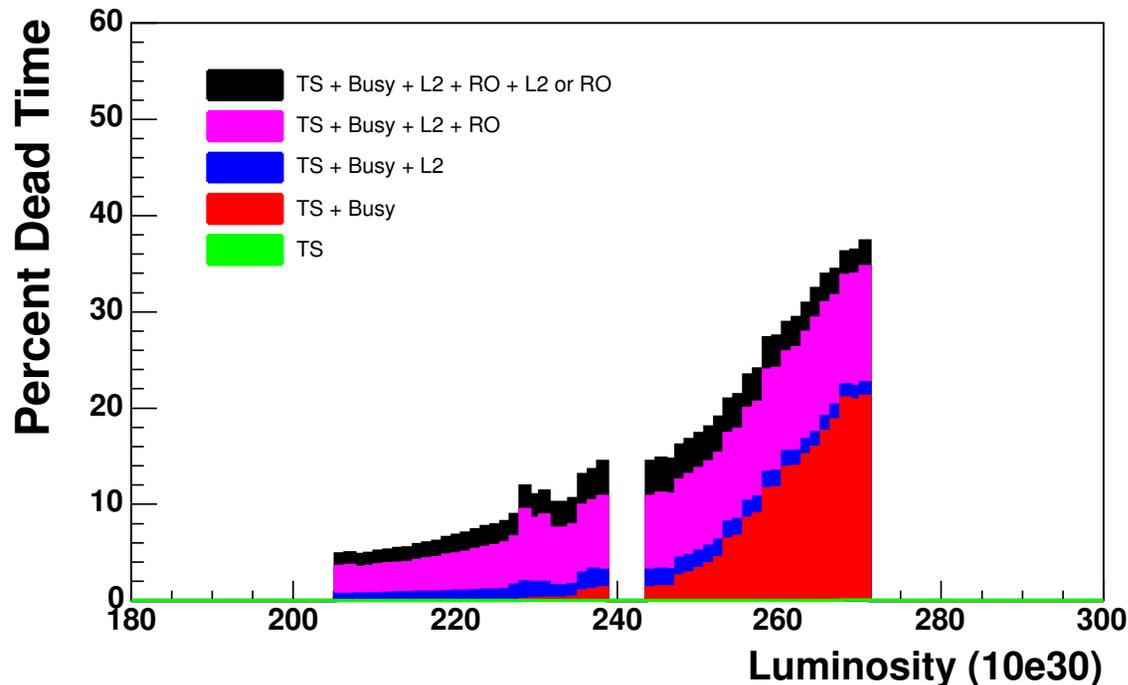
# Overview of CDF DAQ



*Bottlenecks can occur at different stages*

# Dead Time

At high luminosities we can trigger at rates higher than the DAQ can handle leading to *high deadtimes*.



## Readout Deadtime:

Unable to readout event in FE crate fast enough (*high occupancy...*)

## Busy Deadtime:

Unable to send data to the VRBs (*VRB readout, L3 processing, CSL...*)

Trigger rates are controlled by the trigger configuration

Cut values and prescales are determined by the Physics Groups

→ *Trigger changes involved extensive study and can take a long time to validate...*

# Readout Deadtime

We have two main types of Front End Crates...

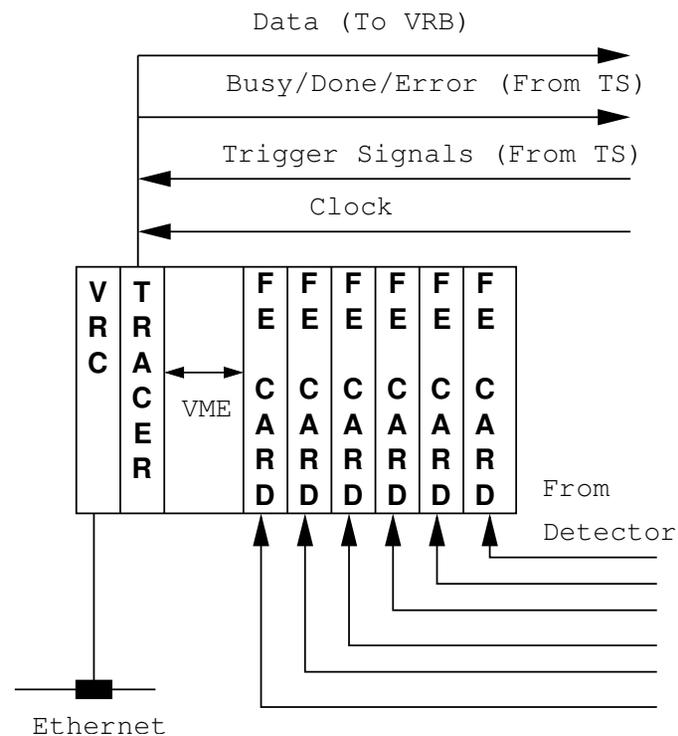
## 1) Crate processor formats the data

Reduce the readout time with faster processors *Motorola MVME 2100* → *MVME 5500*

*Issues: OS version, software compatibility...*

## 2) "Spy mode" readout

Data is processed by the Front End Card (TDCs)



Data is transferred to VRBs as it is read out of the TDCs.

Does not require that the VME Readout Controller (VRC) read the data from the Front End cards *and* write back to the TRACER...

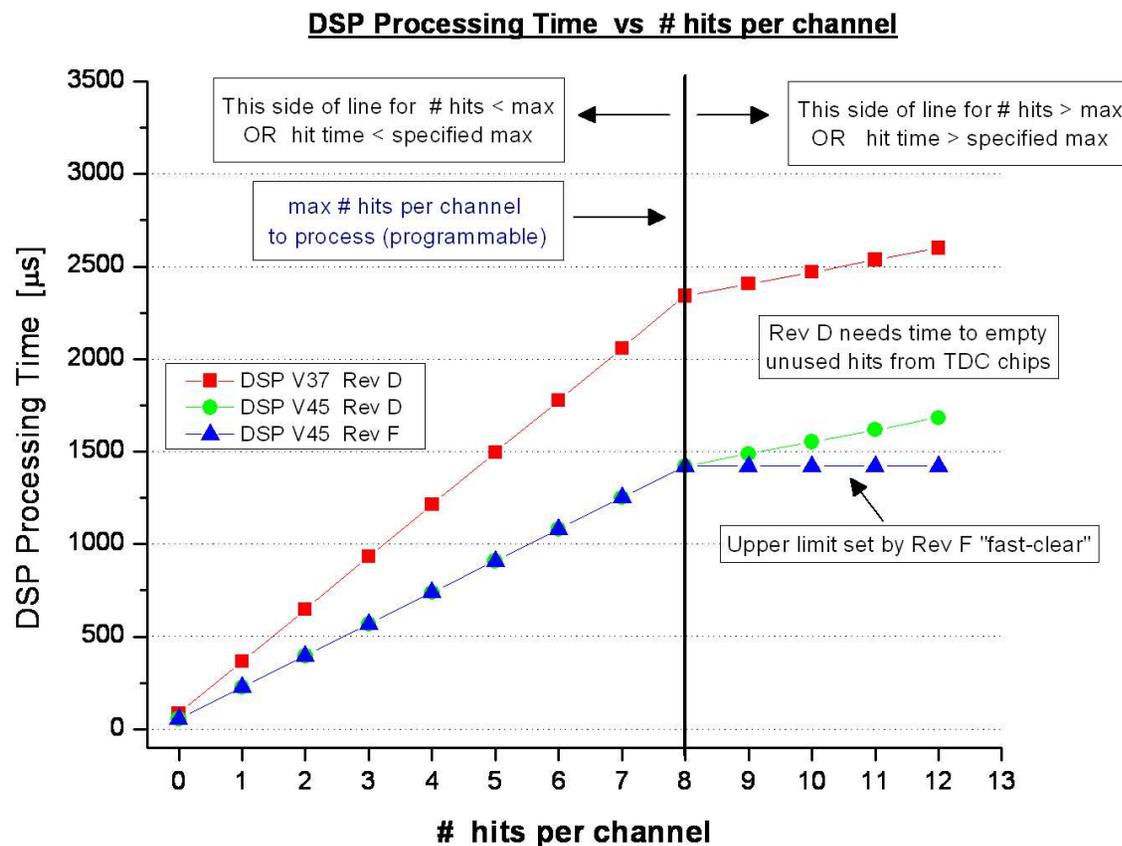
→ *Reduce FE Card processing time...*

→ *Reduce data transport time...*

Recognized that the TDC DSP processing time would limit the L2 accept rate → started planing for a replacement board

TDC DSP processing time is sensitive to number of hits/Channel *and* number of hits outside the time window

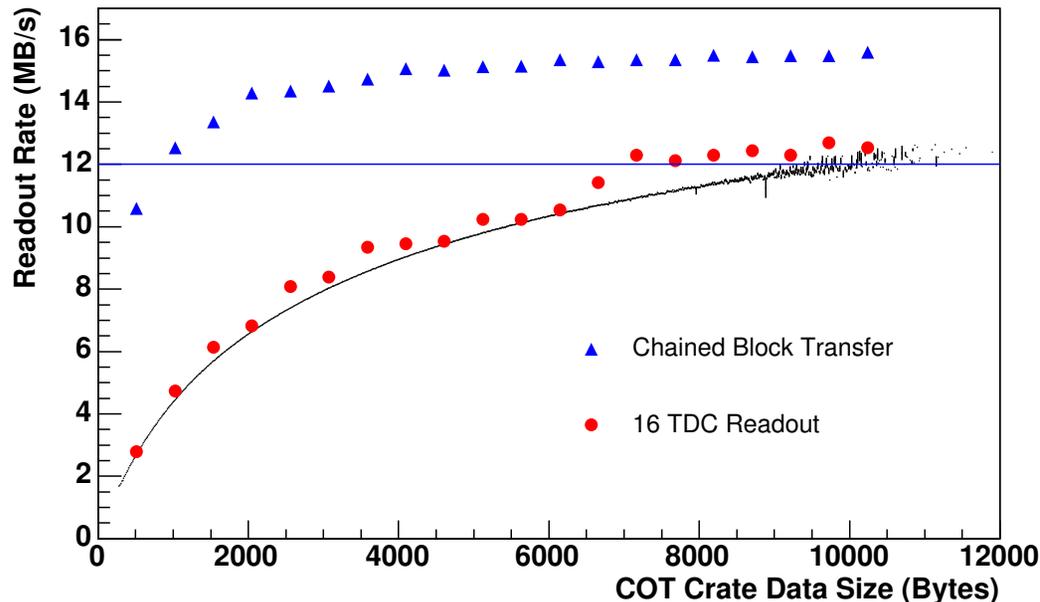
## DSP processing time vs hits/channel



- *Max hits 8 → 4*
- *“Fast Clear” (hardware mod) removes hits outside time window without penalty*
- *Data format → pack the same info in fewer data words*

After reducing DSP processing time, we were limited by how fast we can readout the TDCs over VME

Rearbitrating for the VME bus each time a separate TDC was being accessed (first read data length, then read data)



16 TDCs/Crate

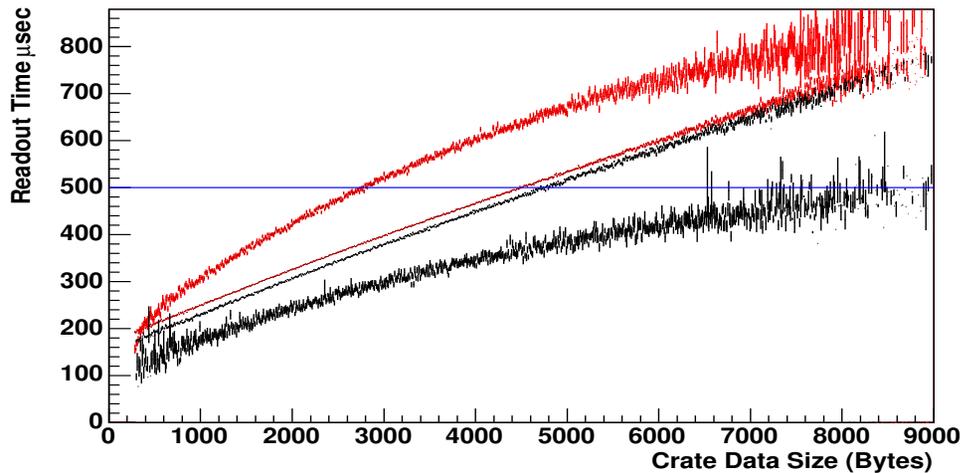
Store event length in separate register → *More mods to the DSP code...*

Read TDC data in second pass using “chained block” transfer

Spy Mode readout: *TRACER* (12 MB/s limit) captures data on VME back plane and transports it to the VRB

*“The Devil is in the details...”*

# TDC Readout Improvements Over Time..

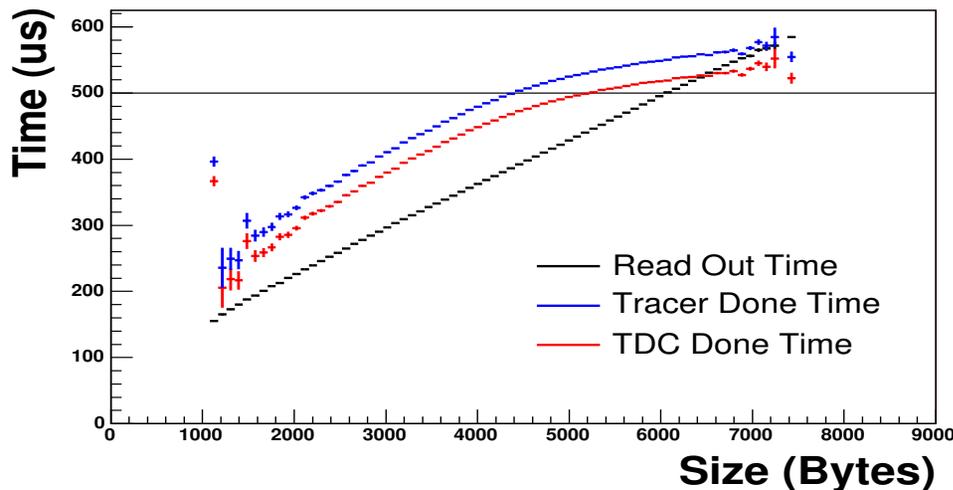


Red: V37

→ DSP processing is the limiting factor

Black: V45

→ Readout over VME is the limiting factor



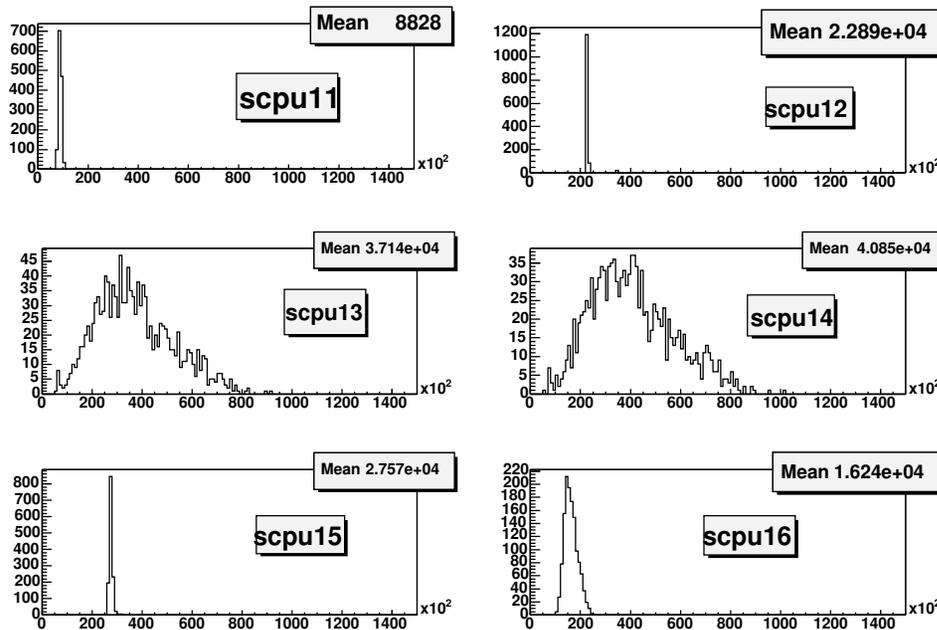
● Optimize DSP code

● Change data format to reduce size

● Optimize FE readout, use chained block transfer...

# Busy Deadtime (Event Fragment Size)

A useful feature of the CDF DAQ is the ability to “partition” the detector so different subsystems of the CDF Detector can be worked on simultaneously.



Initially had 6 DAQ VRB crates, associated with sub-detectors so that the EVB and be partitioned.

*As Luminosity (occupancy) increased, deadtime resulted from an imbalanced distribution*

→ Recabled input to VRBs to balance the load

*We can still “partition” the detector by bypassing the EVB using the “Software EVB”. Much reduced bandwidth but still good for development.*

# Busy Deadtime (Event Builder Rate Limit)

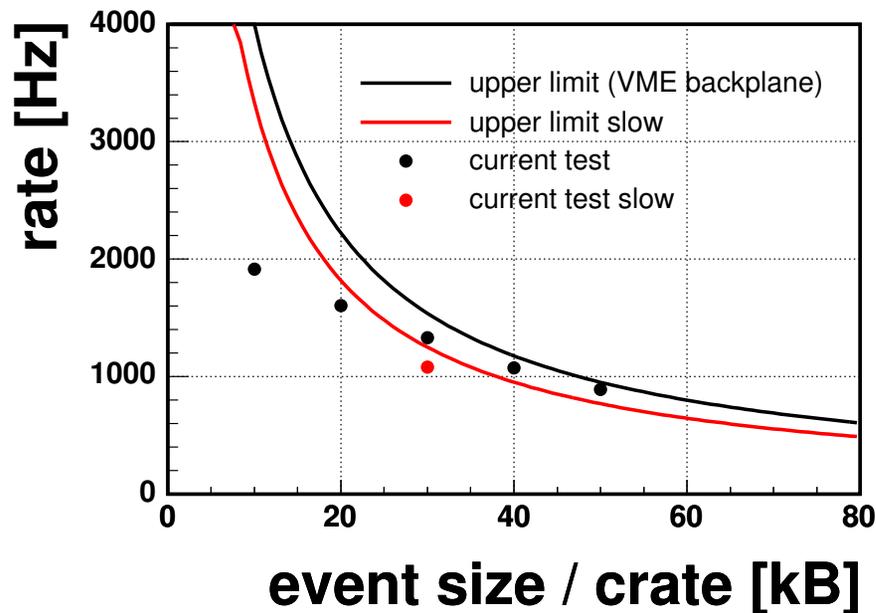
Initial design goal 300 Hz → actual rate limit 400 Hz.

The 400 Hz limit results from *message passing latency*

## EVB Upgrades for Run II (300 → 1000 Hz):

- Use GigE Switch instead of ATM
- Replace Scamnet With GigE (TCP) for message passing  
→ *greatly improves the message passing latency*
- Replace SCPU (VxWorks) with faster processors (VMIC 7805) running Linux  
→ *GigE Network Interface*
- Optimize/rewrite Trigger Manager software

## Upgraded EVB exceeds the design goal of 1000 Hz



As the event size (Luminosity) increases, time to readout the VRBs will limit the trigger rate.

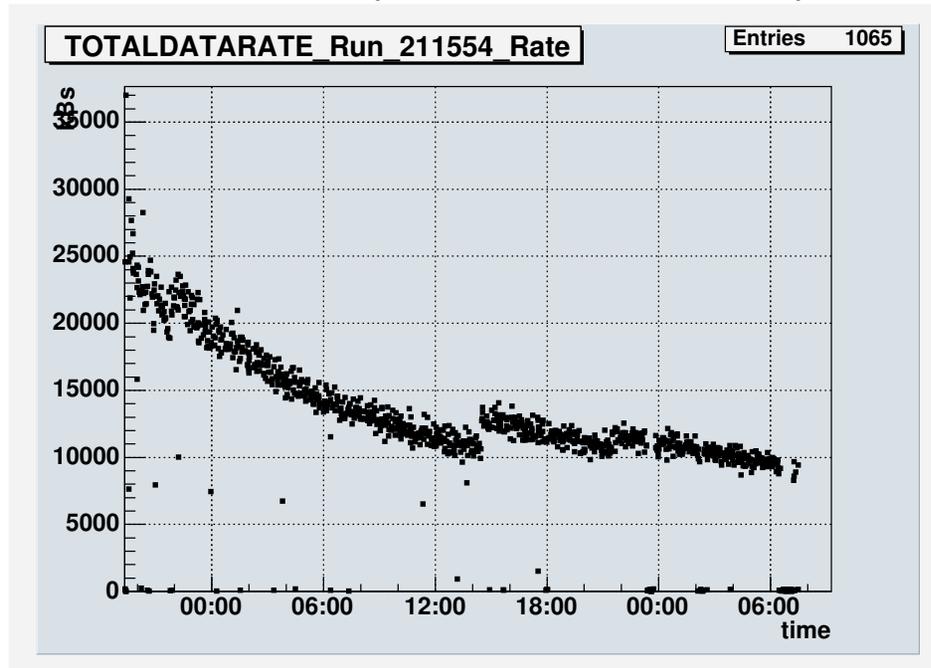
→ *Bandwidth limit comes from the VME transfer speed (shown as the solid curves)*

In order to raise the bandwidth limit, we increased the number of DAQ VRB crates from 6 to 12

*Next Problem...*

# Upgrades to the Consumer Server/Data Logger (CSL)

Run 211554 ( $\mathcal{L} = 176 \times 10^{30}$ )



Old CSL limited to  $\sim 24$  MB/s

→ *Can write to the buffer disk at a higher rate, but limited by the tape writing speed*

→ *Cannot keep up with a high sustained logging rate*

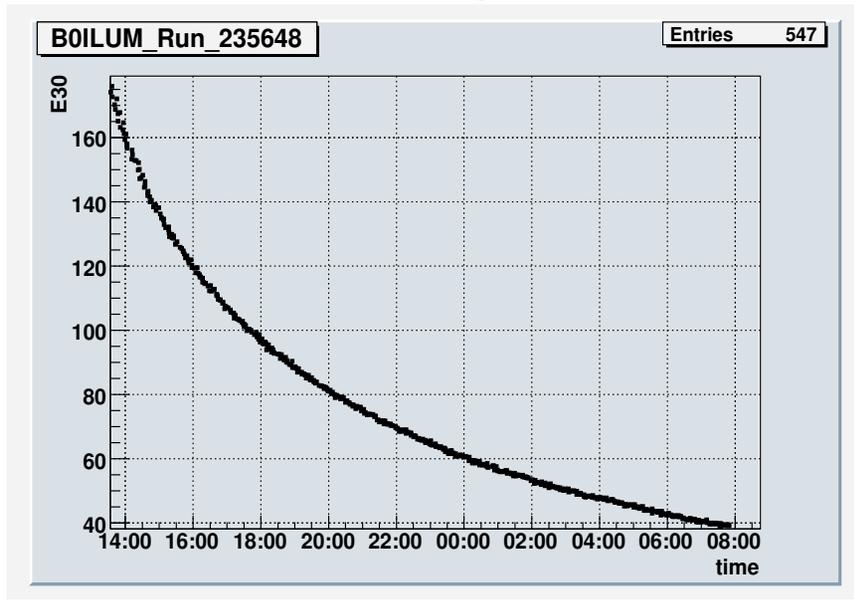
Aging equipment, difficult to support

CSL upgrade design target is 80 MB/s

EVB Rate	1000 Hz	1000 Hz	1200 Hz
Rejection at L3	4-5x → 250-200Hz	4x → 250 Hz	4x → 300 Hz
Event size	200 KB	250 KB	250 KB
Required Bandwidth	40-50 MB/s	63 MB/s	75 MB/s

*Provide enough capacity so that the CSL is not the bottleneck in the system*

## Plot of Luminosity vs Time

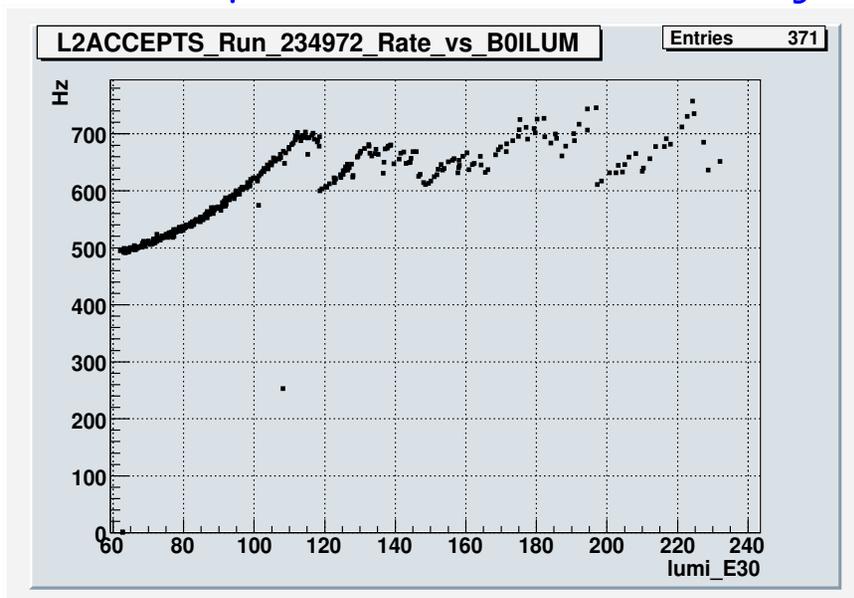


During the course of the store the luminosity drops

→ *Trigger rates decrease*

Can make use of the available bandwidth by *dynamically* changing prescales during the run.

## L2 Accept Rate vs Luminosity



→ *Puts a bigger demand on downstream data logging*

Tape writing need to be able to keep up with the data logging rate

# CSL Upgrades

## Retain original software design and port onto new architecture

- Current software structure serves our purpose
- Limited resources for any major rewrite

## Increase data throughput

- Use modular distributed logging architecture

## Increased buffer capacity

- Buffer increased from 3.7 TB to 24 TB
  - At 80 MB/s we have more than 3 days of buffering capacity
- Plenty of time to react to down stream problems*

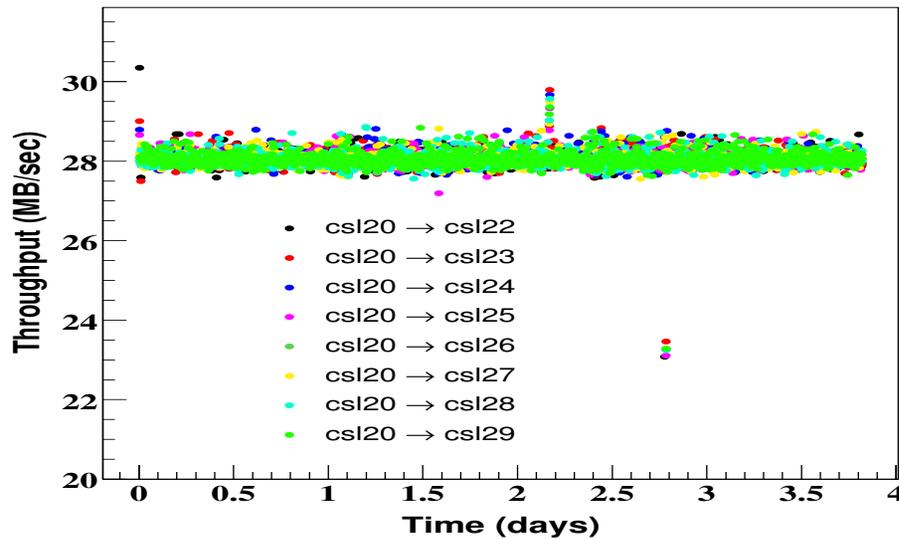
## Improve maintainability and robustness

- Replace aging equipment with new servers/disk arrays
- Use Linux instead of IRIX
- Redundant hardware with hot spares
- *Automatically bypasses failed hardware*

## Expandable

- Can increase number of logger nodes and disk buffer



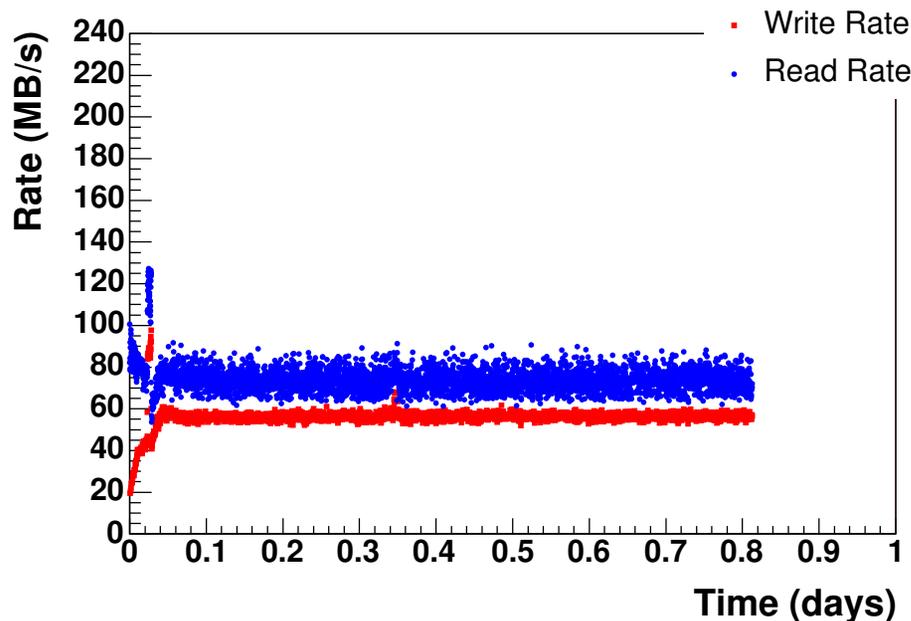


## Network Tests

Data sent from receiver to 8 loggers

→ Total network rate:

$$\underline{8 \times 28 \approx 220 \text{ MB/s}}$$



## Disk IO Tests

Concurrent Read/Write test on 8 logger nodes

→ Total write rate:

$$\underline{8 \times 56 \approx 445 \text{ MB/s}}$$

→ Total read rate:

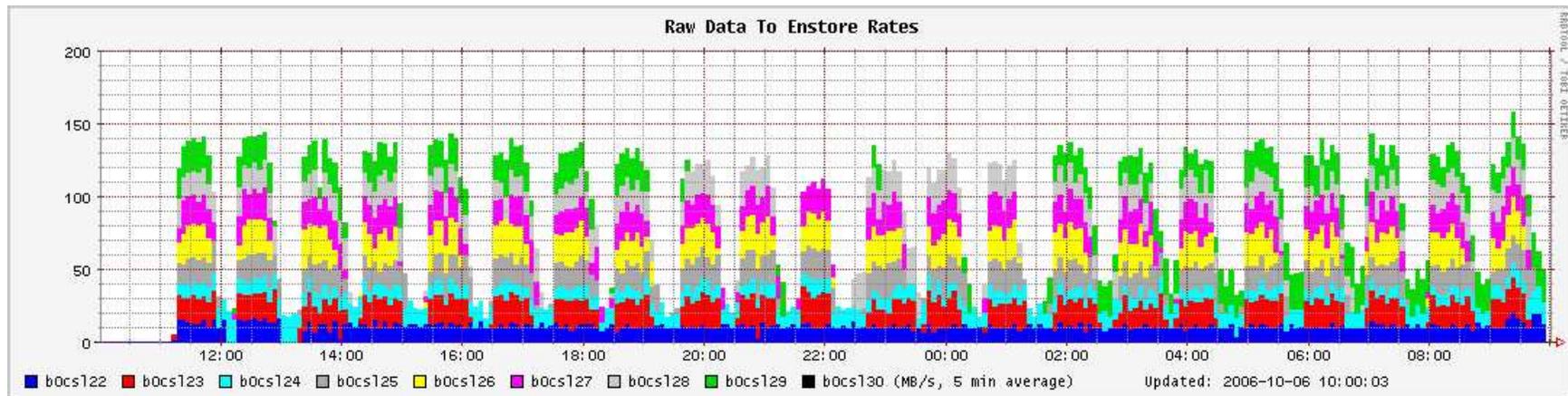
$$\underline{8 \times 74 \approx 590 \text{ MB/s}}$$

→ Can catch up if down stream tape logging is interrupted

→ Can buffer data faster than it's delivery (limited by network)

# Dry Run Tests

Data was sent through the full CSL chain at about 100 MB/s continuously for more than 24 hours.



New design allows running of eight parallel tape writing jobs  
→ significantly increases tape writing capacity

→ *We stressed the system well beyond our target of 80 MB/s*

→ *Adopting a common data logging solution across CDF and DØ in order to reduce the support requirements*

Note: Expected CMS logging rate is 100 MB/s

# Summary

The DAQ requires continuous monitoring in order to identify where we will encounter rate limitations as Luminosity increases  
→ *Require proper tools to understand the performance*

Important to identify potential bottlenecks early on before they become an issue

→ *Some improvements were easy to do, others involved significant coordination across many groups*

→ *Some upgrades took over a year to complete...*

- *Use faster processors as they became available*
- *Optimize code: DSP code, Front End readout...*
- *Optimize data format*
- *Reconfigure hardware, balance data load in VRBs*
- *Major architecture changes, EVB, CSL*