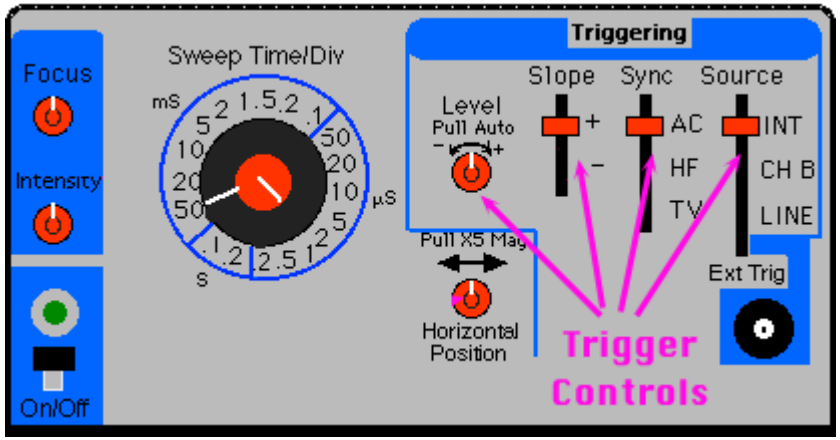


Level 3 Trigger Leveling

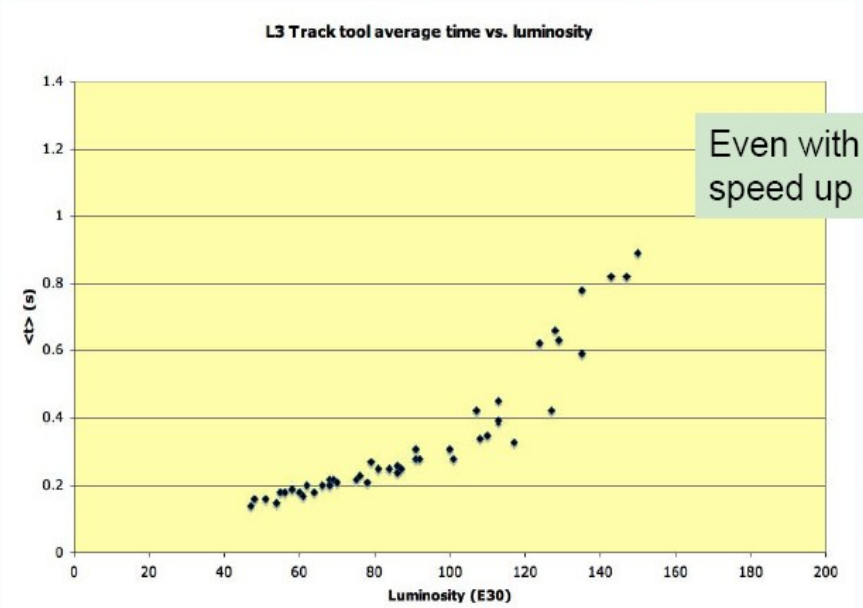
D. Cutts(Brown),
A. Garcia-Bellido(UW),
A. Haas(Columbia),
G. Watts(UW)

The Problem

Events take much longer to filter in Level 3 at high luminosity

Original goal was 50 nodes/1kHz = 50ms/event
We're now talking about >***1s/event*** at high luminosity
We'd need ***>1000 farm nodes*** **to keep up at high luminosity**
Simply increasing the farm power (as we have been), will not be sufficient!

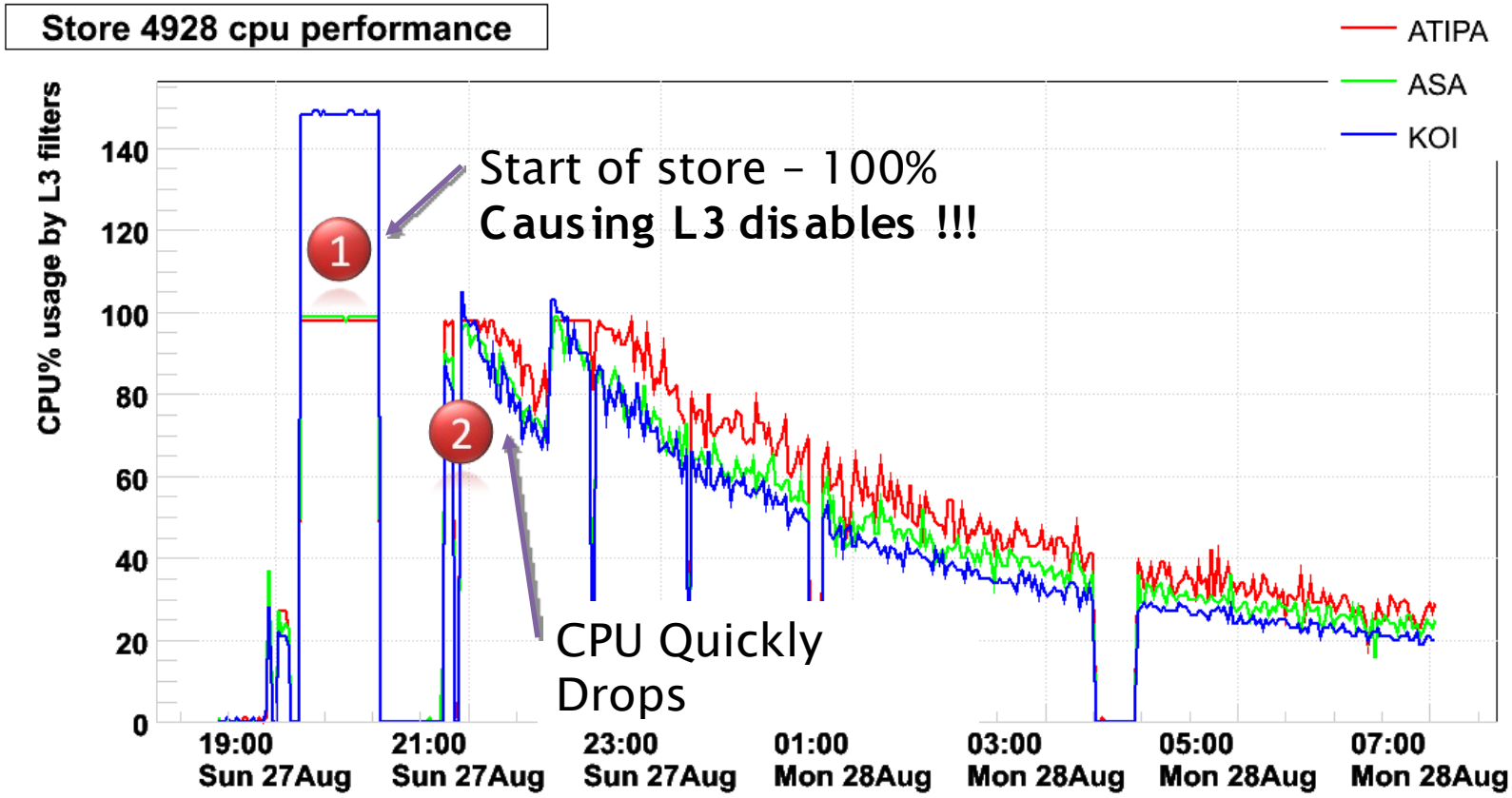
- Present L3 farm is at 99% CPU usage at 150E30
  - New nodes, trigger list design will help, but tracking time increases non-linearly with luminosity



Even with a factor of 3 speed up over p17

Thanks to Rick

Idea: Use Idle CPU Cycles



1. Cache events on L3 Farm for later processing
2. Drain cache using unused CPU cycles

Make the CPU usage a level 100%

*New Monitor Plots being put together by Aran…

Driving The Solution is…

Initial Part of store with steeply falling Luminosity

Less events later in the store free up the farm to process excess events at the start. The current farm must process every single event as it arrives, so the input rate is limited to the CPU available at the start of the store. We get around this with short runs and quick drops in prescale sets.

Initial Part of store with steeply falling CPU time/event

An event early in the store takes longer to process than an event later in the store, thus freeing up more farm CPU time. We currently deal with short initial runs and prescale set changes.

The longer the run and the higher the starting luminosity the more effective trigger leveling will be

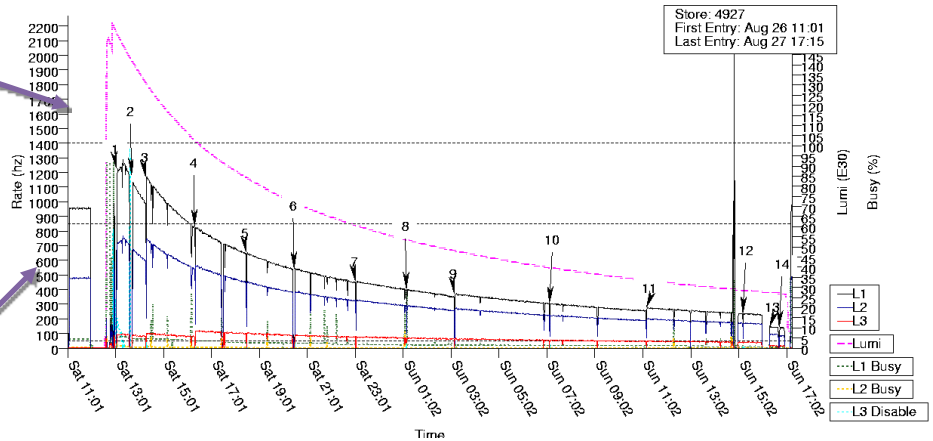
➡ Will Not Address 1 kHz Input Rate Limit!

Two bottlenecks: Event Rate and CPU limit.
Event rate is also limited by the L3 *readout* speed.

Simulation Of Performance

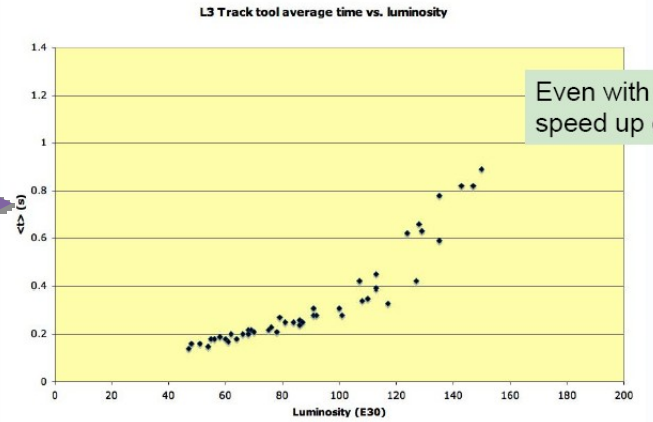
Use falling exponential model of **luminosity**

Use falling exponential model of **trigger rate**



Use *rising* exponential model of **CPU required per event**

- 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



Even with a factor of 3 speed up over p17

Thanks to Rick

What Does CPU Buy You?

Currently can run at 160E30, 800 Hz input rate, with 100% busy CPU

If you only reduce your CPU usage per event by 20% (no trigger leveling) simulation predicts:

171E30

And with 40% reduction in CPU it would get you to about 180E30.

Running at 950 Hz

I pulled all the parameters for this simulation from a store that had what looked like a max input rate of 800 Hz.

Marco wants to run with 950 Hz as the nominal input rate.

With the current farm, the max luminosity you could expect to run at would be **150E30**, compared to 160E30 at 800Hz.

Adding Trigger Leveling

The length of the run now becomes important – the longer the run, the more extra CPU cycles will be available to process the initial events.

The store I looked at had a 2 hour initial run with the 130-150 prescale set. Keep the 2 hour time for now…

The Current farm could get up to 180E30, the same as a 40% reduction in CPU usage (172E30 at 950 Hz input rate).

→ ***Trigger Leveling is worth a 40% increase in farm power (for a 2 hour run)***

The dual-core, dual-chip (4 CPU) machines *now being added to the farm* are thought to add 40% to the current farm's power. Without trigger leveling that will get you up to 180E30, and with trigger leveling that will get you up to 205E30 (or 195E30 @ 950 Hz).

The 4 Hour Run

Increase the simulation to 4 hours. This allows more time to process the events that come in at the beginning. Note that your prescale set still has to keep you to 800 Hz!

* This may be close to reality. Marco is pushing for run trasitions at inst. lum. thresholds, to adjust prescales. First run: 230->170e30 : 3.3 hours

The current farm w/out trigger leveling remains unchanged at a max of 160E30. With trigger leveling the current farm could now handle 205E30 (or 195E30 @ 950 Hz).

*Trigger Leveling is like **doubling** the farm power (for a 4 hour run)*

With the additional 40% increase in farm power from new nodes, this would get you to 250E30 (or 225 @950 Hz).

What Kind Of Trigger Leveling?

All events must be processed by the end of the run, no spill over from run-to-run. Minor modifications to L3, no changes to lumi system, CR, or DL. Some changes to how the control-room looks at things...

Event CPU cycles between stores can be used. Extensive changes to L3, the online system, lumi, offline, databases, etc. But is *only* a software obstacle.

Event Latency

Events are cached in the L3 Nodes (in memory, and on disk!)

- The easiest thing for us, and certainly for the rest of the system, is to keep them time-ordered, i.e. FirstInFirstOut

This means there will be a latency: the time between L1 trigger and arrival at the online system!

- What does this mean for the CR, DL, etc.?
  - Probably no problem
- What does this mean for the Luminosity system?
  - No problem

For a 2 hour run with current farm you can expect at worst 10 minute delay. For a 4 hour run, the delay could be as long as 45 minutes!!

The delay is maximal near the middle of the run...

Event Latency

- What does this mean for the Examines & CR data quality monitoring?

Si, CFT, etc., do online data checks (tick/turn number matching, etc.). Could that sort of thing be moved into the event builder?

Could a small section of the farm be dedicated to running in *copy* mode and passing a subset of current events, routing them to the distributor and having them ignored by the data logger. Those would be used for monitoring.

The generation of any run pausing alarm from an examine is bound to be problematic. Geoff Savage will look at a catalog of these.

Early Run End

Worst case scenario:
1 hour into a planned 2 hour trigger leveled run,
captain decides to end the run.

Would have to wait ~10 minutes for the farm cache to be processed...

Often runs are stopped early soon after they start, when the bad condition of the beam or detector is realized; short runs will have few
events cashed anyway.

Prescales are tuned now so that the initial luminosity doesn't overload the cpu capacity of the farm.
With L3 leveling, prescales will be tuned so that the queues are drained at the planned run-edge.

Event Losses

Node Crash (not filter shell crash!!...)

Worst case:
Could loose 8 minutes divided by 200 nodes worth of events.
Events are randomly distributed.

Is current lumi DAQ good enough to catch this when it happens?

No! But it's not a big deal.

Software: L3

We are already doing Trigger Leveling!

Nodes contain an internal queue of events which is probably never more than ¼ a second deep (3 events).

*4 events/second * 2500 seconds = 10000 events*
*10000 events * 2MB/event = 20000 MB = 20GB*
A lot of memory... but not a lot of disk space!

To first order turning on trigger leveling is just making this queue deeper. Queue depths of 8 minutes or so are within the Linux address space capability. Longer queue times will require authoring a backing store in a disk file (instead of swap space). Probably take someone in our group one week of full time work to get this up, running, and debugged.

But there could be synchronization issues required for the DL/CR etc. which would require more work from us. The routing master would have to be more clever about which nodes it sends events to...

Monitoring

Features will be added to FuMon

Number of free / full buffers in each node queue
Processing rate of queue in each node
Estimated time until empty queue for each node

A total farm number of free / full buffers
Processing rate of total farm queue
Estimated time until last empty node queue

Time stamps for events as they are being sent out of L3?

Tests

We have already done some work on L3 node code,
may test this afternoon!

Additional monitoring variables of FilterShell queue sizes

Codes fixes to allow for more event buffers

We will try increasing the number of buffers
(on a couple nodes)

Tests will be done while we're commissioning the 48 nodes
borrowed from CAB

Conclusions

We (L3DAQ group) think this is worth a try
... and no showstoppers have been found so far.

You get a huge performance gain in L3 filtering power *needed* for running at high luminosity.
(About a *doubling* of farm power... depending on run conditions... worth nearly $1M)

It should be pretty easy to implement.
(Tests will begin soon...)

There are downsides in terms of control-room flexibility and simplicity...