



Special Prescaling at Level One

David Saltzberg
(UCLA)

Trigger Hardware Meeting

May 8, 2003



The boundary conditions (as I understand them)

- We want to take as many B (and D?) triggers as possible:
 - while keeping L2 deadtime / 5%
 - At $\mathcal{L}=1\times 10^{32}$, there are ~ 70 kHz B candidates to take at L1 that are worth consideration by the SVT.
 - At $\mathcal{L}=1\times 10^{32}$, there are only ~ 10 kHz to take at L1 for high p_T physics program.
- L2 deadtime when all four buffers are full.
- Typical L2 latency is about 37 μ sec
 - maximum throughput is $1/37 \mu\text{sec} = 27$ kHz
 - Only 4 buffers \rightarrow lower throughput since processing times vary and L1 triggers arrive stochastically
 - Current limit is ~ 20 kHz (ie, 10kHz of B's)



Buffer MC studies

- Using dead.f

- I know Modsim is better, but this is what I had to work with

- simulates L2 deadtime

- Uses timing plots from Peter Wittich

- Includes:

- 1) random L1 accepts,
- 2) front-panel loading times from “clients”,
- 3) pipelining in interface boards,
- 4) MB transfer timing,
- 5) DMA transfer into alpha,
- 6) alpha event pipelining,
- 7) TS handshake time,
- 8) fluctuations on above using Γ fcn

- Seems to get deadtime and typical buffer usage profile ~right for example runs

- Not suprising since basic timing is very simple:
wait for silicon → MB transfer → process → talk to TS



is dead.f reasonable?

- Check with recent run: 162663

- ~1.5% observed L2 deadtime

- dead.f predicts 1.9%

- buffer usage (for taken events) predicted vs. observed:

buffer	obs (%)	pred (%)
0	53	58
1	29	28
2	13	10
3	5	3

- Have tested runs with more deadtime (L2 torture) and also got reasonable agreement.

- Still, Modsim will do better and has been better tested.



deadtime is due to fluctuations

- **<buffer occupancy>**
 - at 20kHz L1A $\langle \text{buff.occ} \rangle = 1.3$
 - Poisson prob. of 4 or more full = 4.6%
 - which agrees with 5% deadtime.
 - means we are not using buffers very efficiently. I.e., we are prescaling B,D triggers heavily because we are afraid of the expected Poisson fluctuations of 1.3 up to ≥ 4



Special Scaling

- Instead of a flat prescale (or dynamical prescale which is a special case of a flat prescale)...
 - set prescale to 1
 - Only take event if 3 buffers are empty
 - (seems optimal choice so far)
 - Ie, “sneak in the B’s when they cause ~no
deadtime



Physics Payoff

- Assume 10kHz high- p_T physics & can ask different questions:
 - For same 10kHz B candidates, what is L2 deadtime?
 - is 1.7% (instead of 5%)
 - For 5% L2 deadtime, how many more B candidates?
 - better question
 - hard to make it 5% dead!
 - if 120kHz of available B's can get 4.6%
 - gives 21kHz B candidates into L2.
 - ie, at theoretical limit: $1/\text{latency}$.



Extra bonus, Uses alpha pipelining & Shortens L2 latency

- <Buff occ.>: 1.3 → 2.4
 - Means alpha pipelining has more opportunity to kick in
 - L2 latency: 37 μ sec → 34 μ sec
 - this improvement is limited by silicon latency
 - ➔shortening silicon latency would have a linear improvement in L1A rate since CPU time is mostly absorbed in pipelining
 - Because all improvements are “in addition” to high- p_T physics, shortening times would be a super-linear benefit to B&D physics
- Side effect: less “slack” for readout to not incur deadtime.
 - Modsim could weigh in on this better.
 - Presumably, once we have good events found by L2, readout bandwidth will follow. (i.e. ‘We should have such problems.) Anyway, this is just getting us back to the rates we originally designed for in IIa.
 - Muon board will offset this effect somewhat



Actual improvement?

- Improvement to B “L1 triggers” considered by L2 is \sim a factor of 2.
- But does that translate into $\times 2$ more B’s for physics?
 - Are the extra events we’d get as pure?
 - Do we run out of B cross section?
 - L2A (readout) bandwidth issue?
- This trick cannot be applied to certain prescaled “calibration” triggers which want to sample same instantaneous luminosity conditions
- Ought this trick be applied to some QCD and other physics events too?
 - give each trigger a “rank” where rank=# of L2 buffers that must be free to accept it.



Context for hardware improvements

- Under this regime:
 - 4 μ sec off of CPUtime takes (~40% off) deadtime from 3.7% \rightarrow 3.1% and an extra 1kHz L1A
 - 4 μ sec off of silicon latency instead (~15% off) 3.7% \rightarrow 2.9% and an extra 2kHz at L1A.
 - Difference is due to pipelining. Ie, L2dec “sneaks” its processing under the silicon latency time for the next event. Higher buffer occupancy helps this.
- Ie, both matter, but for fractional improvements (ie, corner cutting & tricks), pushing on silicon offers more payback.



Conclusions

- What is proposed is essentially a “load-leveler” for level-2.
 - Improves throughput to close to $1/\text{latency}$.
- Implementing this special prescaling cannot hurt and should help at least somewhat.
- if MC simulation is right, and if all L1 B's are created \sim equal, and if we can read them out, improvement could be about a factor of 2 and/or less deadtime for high- p_T physics