The D-Zero Level 2 Trigger for Run II Physics[1]

R. Hirosky
for the D-Zero Collaboration
University of Illinois, Chicago, Department of Physics, Chicago, Illinois 60607

Abstract

For the next collider run at the Fermilab Tevatron, the D-Zero experiment (D∅) will employ a three level trigger scheme. The rate into the second level trigger (L2) is expected to be 10 kHz. This trigger must further reduce the rate by a factor of 10 while introducing an overall system deadtime of less than 5% and maintaining trigger efficiencies realized in the previous run. Level 2 is the first trigger stage to combine information from the various detector sub-systems. Preprocessors will format each subdetector's data and pass this information to a "Global" crate that makes the event decision. The Preprocessor and Global crates contain custom built 500MHz Alpha CPU cards based on the DEC PC164 motherboard. We give an overview of the Level-2 trigger including the data flow through system, queuing simulations, and the current implementation status of the various software and hardware components of the system.

I. Introduction to the D-Zero Trigger

Present upgrades to the Fermilab Tevatron will provide more than thirty times the instantaneous luminosity than was available in the previous collider run. The expanded data set will greatly increase the physics reach of the Tevatron deeper into studies of top production and decay, SUSY models, QCD, electroweak and new phenomena. To fully realize the potential of the data, D-Zero's trigger system will be upgraded to function efficiently in the new accelerator environment where the time between successive beam crossings will decrease from a few microseconds to 132ns.

Triggering will be accomplished in three stages. Each detector system will have a first level (L1) trigger, based on fast frontend readouts of detector channels. This will apply threshold cuts to geometric detector objects (i.e. hits, energy depositions) and tag coincidences. Events passing L1 will be passed to the second level (L2) trigger that will combine L1 channels in individual detectors to produce proto physics objects. Each detector subsystem will have its own L1 and L2 systems and will pass lists of objects to the L2 Global component which will combine physics objects from the various subsystems to provide selections based on event characteristics (Fig. 1). Events passing L2 will initiate a full, high precision detector readout pass to the third level trigger (L3), a computer farm of NT-based machines, for further processing and selection, resulting in an output rate to tape of 50Hz or less. This paper describes the L2 trigger system.

[1]We would like to acknowledge the support of the National Science Foundation.

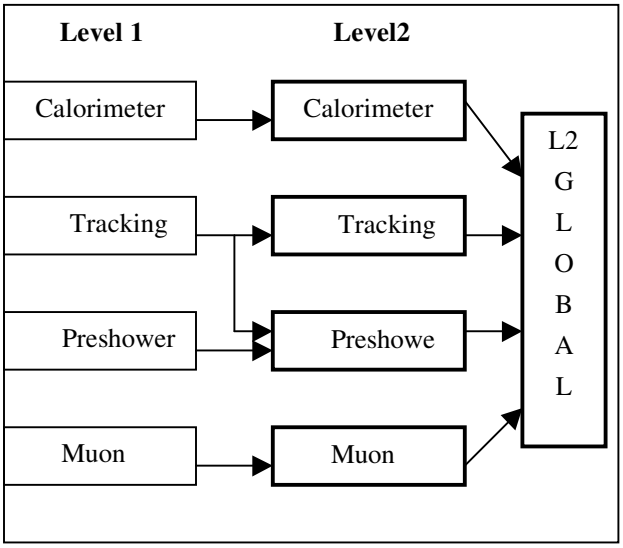

Figure 1: Block diagram of L1 and L2 trigger components.

II. Overview of L2 Trigger

The Run II D∅ Level 2 triggering system is required to provide a factor of 10 rejection while maintaining electron, muon and jet efficiencies obtained with the Run 1 Level 3 software trigger [1]. The system is to impose deadtimes of less than 5% at a 10 kHz input rate while tagging events passing one or more of 128 selection criteria

The L2 design was constrained by the front end data buffering system which requires that trigger decisions be made available in the same order as events passed by the L1 trigger. This imposed a maximum of 16 events in the L2 system awaiting trigger decisions (beyond 16 generates deadtime). The basic parameters of the trigger design indicate that decisions from L2 are required on average every 100μ *sec*. The system architecture was guided by extensive queuing simulations [2] before choosing an architecture for the system. To minimize the I/O connections, it was early decided to restrict specialized preprocessing of detector information to dedicated preprocessors.

The final architecture can be thought of as a stochastic pipeline. Preprocessing of each detector's data occurs in parallel and the high-level information is combined in a second, global stage. Each stage has a nominal time budget of 100μ *sec*; the simulations give low deadtime for processing times in the 50-75μ *sec* range for Preprocessors and Global alike. Simulations indicated the importance of minimizing variation of decision time from event to event, imitating a true pipeline, as well as the importance of

providing the maximum number of buffers (16) in front of each transfer point in the system. With this configuration we gain simplicity in data flow control in L2 and busy signals can be generated by the front end. This also allows guaranties that each data transfer has a buffer ready to receive it and allows the system itself to decide where buffers are needed.

Each L2 preprocessor receives L1 data from one or more subsystems. The Preprocessors currently in development are for the calorimeter, tracking system, PreShower system, and muon system. A Preprocessor for reconstruction of displaced vertices in the silicon vertex detector is under study. The system components are illustrated in Figure 1.

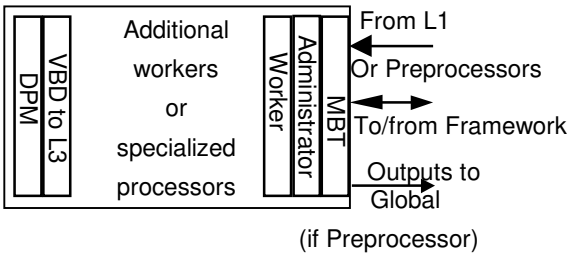

Figure 2: Standard Level 2 crate.

III. L2 CRATE

All DØ L2 systems will occupy a standard L2 (9U VME for physics) crate. The crate, shown in Figure 2, contains a VME bus and a custom-built 128 bit Magic Bus on the J3 connectors (a handshaking bus capable of 320MB/s). Each crate contains the following devices: dual-port-memory (DPM), output to Level3, a Magic Bus Transceiver card, for fast I/O, and at least two processors. The DPM is used for the download of run-time parameters and the reporting of monitoring data to the trigger control computer (TCC). A VME Buffer/Driver card is used to write data to the L3 system. The two processors are referred to as the "Administrator" and the "Worker". The above devices are the minimum required for L2 crates. Where appropriate, additional specialized hardware for data conversion or processing can be used.

A Worker node applies trigger algorithms to the L1 data channels. The Administrator does all event processing and local trigger control tasks that do not involve the application of the trigger algorithm. These include, but are not limited to, verifying the data integrity, controlling communication with the trigger framework including notification of L2 decisions, controlling the output of monitoring data, and controlling the readout of events to Level3. The separation of tasks is an additional level of parallelism.

A. Alpha Processor in VME Card

The design of the Alpha processor board is based on the layout of the PC164 card sold by Digital Semiconductor. Each card contains a Digital 500MHz 21164 Alpha CPU. The processor can execute 2-4 instructions per cycle. The board includes Digital's 21172 PCI interface, providing a 33 MHz, 64-bit input path to memory (maximum rate of 267 MB/s). Added elements to the PC164 design include: a 64bit PCI to VME interface based on the Tundra Universe chip [3]; a block transfer controller connecting to the 128-bit MBus; a fast I/O port for registers for control or monitoring (the "Fred" port), and a bi-directional 64-bit programmed I/O interface between PCI and Mbus. The DØ Alpha cards are closely related to the those being developed for the CDF experiment.

B. Magic Bus Transceiver Card

Data transferred to and from the Magic Bus pass through a VME card named the Magic Bus Transceiver (MBT) Card. The high speed 128 bit input path is equipped with a buffer for up to 16 events. The input transport takes place on Cypress Hotlinks [4] (20MB/s serial data paths). Chip sets handle serial to parallel conversion at either end. Each MBT card provides 8 such input paths. The MBT control logic waits for all inputs for an event to arrive, places the input event on the MBus, and generates a MBus broadcast to the Alpha cards. Control registers are visible to MBus so that the Alpha cards can control the MBT card. There are two output paths that are targets of the 64 bit MBus programmed I/O on the Alpha Cards. A 128 bit wide path will transport the L2 event decision to the L2 hardware framework. In addition, an MBT card has 2 Hotlinks output paths. These are intended to allow Preprocessors to send their output to L2 Global.

IV. GLOBAL PROCESSOR CRATE

In the L2 Global Processor crate when data arrives via the MBus it is broadcast to both an Administrator node and a Worker node. The data includes high-level information from the L1 trigger framework such as which trigger conditions require evaluation and whether the event has been marked for any kind of monitoring or additional processing. The Administrator is responsible for coordinating assignment of the event buffers, L3 readout, and sending monitoring information to a trigger control computer (TCC). The Administrator verifies the data. The Worker analyzes events, reporting the result (a mask of 128 pass/fail bits) to the administrator and the L2 hardware framework. If the event passes, the framework controls readout of the event data to L3. For events marked as "monitoring", statistics, error messages, and processing time buffers are collected and written via VME to the dualport memory. This is accessible to the TCC, which will serve the data to monitoring comsumers. To avoid long latencies caused by collisions with L3 readout on the VME bus, communication between the Administrator and Worker takes place on the MBus.

V. CALORIMETER PREPROCESSOR

The calorimeter preprocessor will apply clustering algorithms to the L1 calorimeter data. The L1 trigger reports calorimeter transverse energy (E_T) in (.2 x .2, η x ϕ) towers, additionally L1 will report a bit mask designating towers above fixed E_T thresholds for use as starting seeds in

clustering algorithms. Worker nodes will be included for clustering of electromagnetic objects, jets, and missing transverse energy. The L1 data arrives on 10 input links (via MBT cards) and will transport 2.6KB/event of tower transverse energy data, including electromagnetic towers and full tower sums. The tower masks will require an additional 0.3KB of data. The Preprocessor will also receive L1 qualifier information, telling whether the event was a monitoring event, whether the preprocessor is required to run, and other high-level control information.

The electromagnetic clustering algorithm considers 1 X 2 tower regions, applies a threshold, and cuts on neighboring energy and on the fraction of electromagnetic energy in the cluster. The jet algorithm clusters a 5 x 5 tower area and reports total E_T, the jet location, and the E_T-weighted RMS width of the cluster. The missing transverse energy algorithm returns a vector sum of all tower E_T values and may apply a variety of threshold and fiducial cuts in the calculation.

The calorimeter preprocessor will use the same basic design as the L2 Global processor. Data are broadcast to the administrator and Worker Alphas. When the algorithms are completed and the result passed to L2 Global the administrator awaits a message on the Serial Command Link (SCL) module [5] (from the framework) indicating whether an event has passed in L2 Global. Its buffer structure is somewhat more complex than that of L2 Global, as an event has more requirements to pass before its location is freed.

VI. Preshower and Central Tracking

Both the PreShower and central tracking preprocessors will be configured with an Administrator-Worker set of Alpha processor boards as in the calorimeter preprocessor. The L2 central tracking trigger (CTT) will use a single Alpha worker and will assemble list of tracks from the Central Fiber Tracker [6] and report them to L2 Global in a P_T ordered list. The PreShower preprocessor will have worker nodes for both central and forward PreShower detectors [6].

VII. Muon Preprocessor

In the L1 trigger, rough coincidences exist between scintillating fiber tracks and muon track candidates. The L2 muon preprocessor [7] must fit muon track candidates using drift time information. Studies with fast serial processors indicated that the raw CPU was adequate to handle typical events, however, high-occupancy events generated deadtime due to the long tails in the processing time distribution. The solution was to apply the fitting algorithm in a massively parallel platform. The choice was arrays of DSP's, each working on distinct geographical regions of the detector. While portions of the available CPU will be relegated to mostly empty geographical sections in a given event, the benefit is an algorithm whose execution time is deterministic, no matter what the occupancy level. The processor cards will be based on the Texas Instrument TMS320-C6201 DSP. Each processor card will contain 5 DSP units. This DSP is internally configured with 8 parallel subprocessors and has a maximum speed of 200MHz. The DSP is rated at 1600MIPS; however, since most code is unable to take full advantage of the parallelism, the typical optimized code runs at 300-500MIPS.

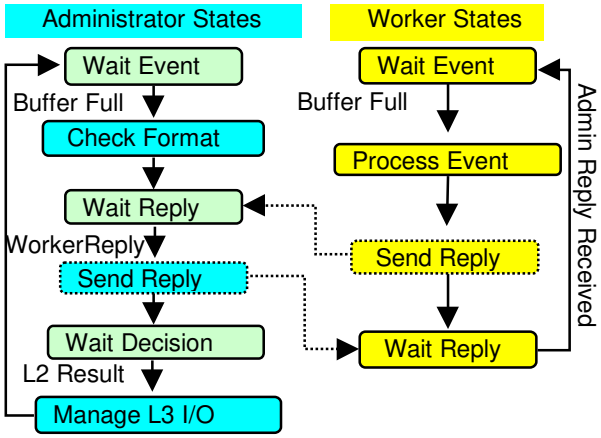

Figure 3: State diagram of Administrator and Worker code

VIII. Alpha Software

The basic data movement code consists of buffer management [8] (Fig.3). The event data are stored in an internal memory buffer that changes state as it moves through the processing stages. These buffers are synchronized between all the alpha cards in a crate because all cards see every event broadcast on the Mbus. A Worker node has the most simple event loop. After initialization it awaits an event which it then processes and informs the Administrator when it is finished. It then waits until the Administrator tells it which buffers to allocate for new events and which event to process next. The Administrator loop must handle readout of the processed events. When at event arrives the Administrator checks the format to verify data integrity and waits for a worker to respond. When a Worker responds the Administrator sends a reply so that the Worker can resume processing and then checks for framework to report back a trigger decision. Finally the Administrator will manage readout to L3, by telling the VBD to read the event, and then update the relevant buffers' states.

IX. Timing Studies

Although L2 crates are not restricted to one Worker, all Workers in a crate must use the same input sources and must operate independently of one another. One Worker must not have knowledge of any event-by-event devices other than the MBT and the Administrator node.

The operational mode of multiple-Worker systems has been studied in a discrete event simulation [9]. The choice is between operating the systems in event synchronous or event asynchronous modes. The event synchronous mode, also known as lockstep, requires the Administrator to ensure that all Workers are processing the same event at the same time. The primary advantage to lockstep mode is the correlation between a given event and a temporal period. The other operational mode, event asynchronous or non-lockstep,

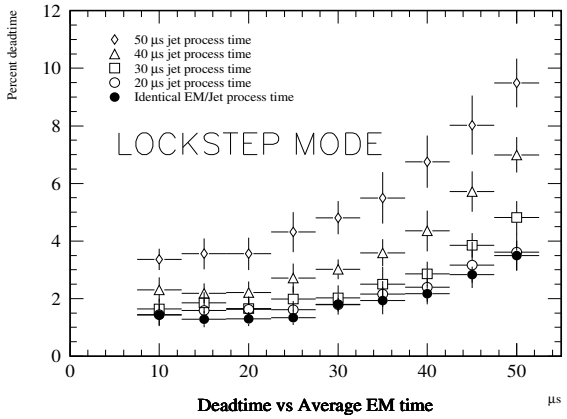

Figure 4: Deadtime, in percent, versus EM processing time, for workers configured in lockstep mode.

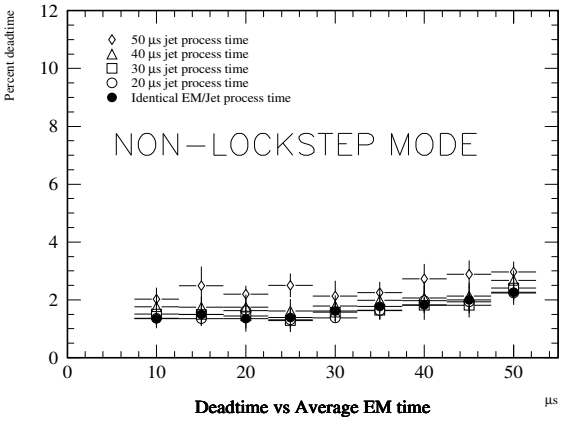

Figure 5: Deadtime, in percent, versus EM processing time, for workers configured in non-lockstep mode.

allows a Worker to start processing the next event as soon as it finishes the current event.

The system used to study the two modes of multi-Worker operation was the L2 calorimeter system. The missing transverse energy processor was assumed to have a constant processing time of $45\mu sec$. The mean time of the EM and Jet Workers' hyperexponential distributions was allowed to vary up to 50 μ *sec*. The results of the simulation are shown in Fig. 4. When both EM and Jet processors have times near the nominal time budget and the system is operating in lockstep mode, the L2 system deadtime approaches ten percent, double the specified maximum. For the same processing time, the non-lockstep system is still within the 5% deadtime specification (Fig. 5). More importantly, it can be seen that the system operating in non-lockstep mode is much more stable with respect to changes in the Worker processing time.

To further understand the source of the deadtime when operating in lockstep mode, the simulation was run with the Jet and EM processing times 100% correlated. The low deadtime results of this test, shown as solid circles in Figs. 3-4, suggest that the deadtime in the uncorrelated case is due to a worst-of-n situation — all Workers must wait for the slowest Worker. Based on the simulation results, multiple Worker Preprocessor crates will be implemented to run in non-lockstep mode. Lockstep mode will still be used for testing and debugging purposes. The question of multiple L2Global Workers, with each calculating a different event, is more complicated. The results from the Workers must be reported in event-arrival order. Fortunately, the Administrator-Worker scheme provides a way to have multiple L2Global Workers operating in an event asynchronous mode with the Administrator reserializing the trigger decisions prior to notifying the trigger framework. This mode of operation was modeled [9] and a clear gain demonstrated utilizing two Workers. If the available mean processing time is held constant, the largest percentage decrease in deadtime comes with the additional Worker without regard to the operational mode. If, however, the question is posed as how to obtain the highest available mean processing time for a constant deadtime, the advantage of operating in an event asynchronous mode becomes apparent. It is worth noting that neither provides a factor of two in available processing time. The only way to achieve that is to lower the input rate, i.e. increase L1 prescales or thresholds. The current plan is to implement L2 Global as a single Worker system with the capabilities of adding multiple Workers in either operational mode.

X. STATUS

The core functionality of the Alpha boards has been verified on several prototype models, final design modifications specified, and the production boards ordered. Prototypes of other cards in this paper are currently being tested. Software studies of physics algorithms are underway, the L2 infrastructure code has been prototyped, and both device driver software and inter-processor communication software is under development. A test of the full L2 system will commence in the fall of this year and the L2 trigger will be ready to run when the data flow in mid 2000.

XI. REFERENCES

[1] S. Abachi et al., "The DØ Detector", *NIM*, vol. A338, p185, 1994.

[2] Queing simulation results are available from the DØ web page. http://www-d0.fnal.gov under Technical-- DØ Upgrade—Trigger Systems.

[3] Documentation for the Universe chip is available at: http://www.tundra.com

[4] Hotlinks, see Cypress web page: http://wwwcypress.com

[5] J. Linnemann, Proceedings of the 10th IEEE Real Time Conference 97, Beaune, France, 22-26 Sept. 1997.

[6] G. Blazey, see above proceedings.

[7] M. Fortner Proceedings of the International Conference on Computing in High-Energy Physics, Chicago, Illinois, 31 August – 4 September 1998.

[8] R. Moore, see above proceedings.

[9] R. Martin, see above proceedings.