The DØ Level 2 Trigger [1]

James T. Linnemann
Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824 linnemann@pa.msu.edu

Abstract

The DØ Level 2 (L2)trigger is a subsystem of the overall DØ trigger [1] which has the responsibility of reducing the event rate from 10 KHz to 1 KHz, while introducing less than 5% deadtime. The system is organized as a number of preprocessors and a global processor. Each preprocessor digests information from a single detector system and produces a condensed list of the objects. The global processor receives the object lists, matches across detectors, and applies selection criteria to the objects. The front end system contains buffers to hold 16 events awaiting L2 decisions. The preprocessors are implemented as either arrays of DSP's or fast serial processors; the global processor is a fast serial processor, a Dec Alpha 21164 of 500 MHz (or better) on a VME board with a 128 bit input bus. The board is evolved from a commercial PCI-based PC card design. The overall L2 design has been guided by extensive queuing simulations.

I. Constraints, Simulations, and Architecture Choices

The Run II DØ Level 2 (L2) triggering system [2] requirements are to provide a factor of 10 in rejection while maintaining electron, muon and jet efficiency of the Run I Level 3 (L3) software trigger [3], with a deadtime of less than 5% at 10KHz input rate. In addition, the system will tag events passing one or more of 128 selection criteria (with conditions defined in an ASCII trigger definition file), a level of potential complexity also comparable with the L3 trigger from the first run.

Several constraints were imposed by design choices made before specification of L2. The front end data buffering system requires that trigger decisions be available in the same order as events passed by the L1 trigger, and imposed a maximum of 16 events in the L2 system awaiting trigger decisions (beyond 16 generates deadtime). Reuse of the VME drivers for sending events to L3 for final readout also limited the use of the VME bus for other purposes. As the drivers do not relinquish mastership during event transfer, time-critical communication cannot use VME in crates containing these drivers.

The basic parameters indicate that decisions are required on average every 100μ sec. We performed extensive queuing simulations [4] before choosing an architecture for the system. The chief simulation tool was RESQ [5], which incorporates a mixture of queuing theory and direct simulation. The chief results were checked in simple cases by Fortran programs and using results from queuing theory for simple subsystems [6]. To minimize the I/O connections, it was early decided to restrict specialized preprocessing of detector information to preprocessors.

For combining information across detectors, a farm architecture was considered, but was found to generate enormous deadtime in spite of what appeared to be adequate capacity, because of the requirement that decisions be announced in the order of arrival of events. All farm nodes would be idled when an event requiring a long processing time appeared: most nodes would be left holding events whose fate had been decided, but which could not be announced until a decision was reached on a previous event. Instead, a single faster global processor with a FIFO for holding events for decisions was chosen.

The final architecture can be thought of as a stochastic pipeline of two steps: preprocessing each detector's data in parallel, and combining the high-level information 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 $50 - 75\mu$ sec for processors and global alike. It is critical that the preprocessors not be restricted to event-synchronous operation, lest the processing time distribution in the preprocessors resemble a long-tailed "worst of n" distribution. Sensitivity analysis also indicated the importance of minimizing variation of decision time from event to event (thus imitating a true pipeline), and the importance of providing the maximum number of buffers (16) in front of each transfer point in the system. This simplifies data flow control in L2, allowing any busy signals to be generated by the front end, allows design of each data transfer in confidence that there is always a place ready to receive the data, and allows the system itself to decide where buffers are needed. However, one must then monitor the buffer utilization in each location to understand the dynamics of the system under load.

II. Global Processor

As shown in Figure 1, preprocessors are foreseen for electromagnetic clustering, jet finding, track sorting, preshower hit finding, and muon tracking. We will discuss in detail the global processor, the calorimeter preprocessors, and the muon subsystems as their designs are further advanced.

A. Alpha in VME Card

The difficulty of the task of the global processor depends on the amount of data it receives and the complexity of analysis to be performed on the data. Our current estimates call for .5KB of information to be provided by preprocessors on an average

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

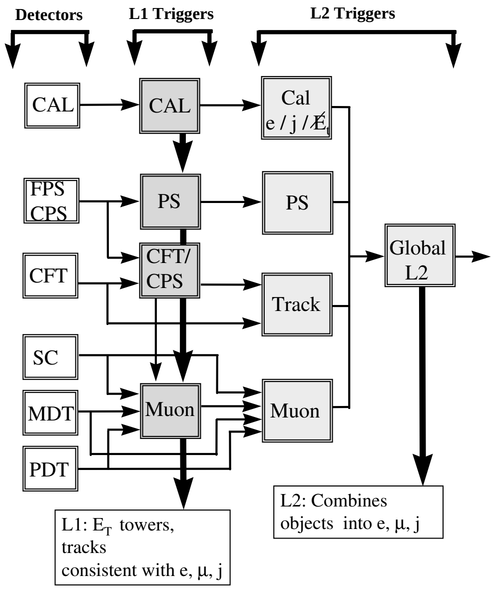

Fig. 1 L1 and L2 trigger elements. The horizontal arrows denote information flow.

event. Since the architecture guarantees that the most complex processing is localized in the preprocessors, the remaining task is simpler than that performed in the Run I L3, so of order 50K instructions/event (100 instructions per byte) seems adequate. This can be achieved by a 500 MHz processor such as the RISC processor made by Digital, the Alpha 21164 [7]. The more difficult problem is to get the data to the processor. We found that the processor card being developed at the University of Michigan by Myron Campbell for the CDF experiment was largely compatible with our requirements, and the decision that both DØ and CDF would use the same product for their global L2 trigger processor.

The card being developed is shown in Figure 2. The design is based on the layout of the PC164 card sold by Digital Semiconductor. The processor can execute 2-4 instructions per cycle. The board includes Digital's 21172 PCI interface, which provides a 33 MHz, 64-bit input path to memory, for a maximum rate of 267 MB/s. The card under development adds several elements to the PCI bus of the core workstation design: a 64-bit VME interface, the Tundra Universe chip [8]; a block transfer controller between a 128-bit I/O bus in the P3 connector of the VME crate (the "Magic Bus" (MBus), a handshaking bus capable of 320MB/s); a fast I/O port for registers for control or monitoring (the "Fred" port), and a bidirectional 64-bit programmed I/O interface between PCI and MBus. The programmed I/O may be used for putting test data on the MBus, for inter-processor communication, for output via MBus, or for controlling MBus input devices such as the MBus Transceiver card described later. The first two additions (VME and MB block input) are designed, tested, and added to the board layout. The second two additions (registers and MB programmed I/O) are under design, and intended for a second round prototype.

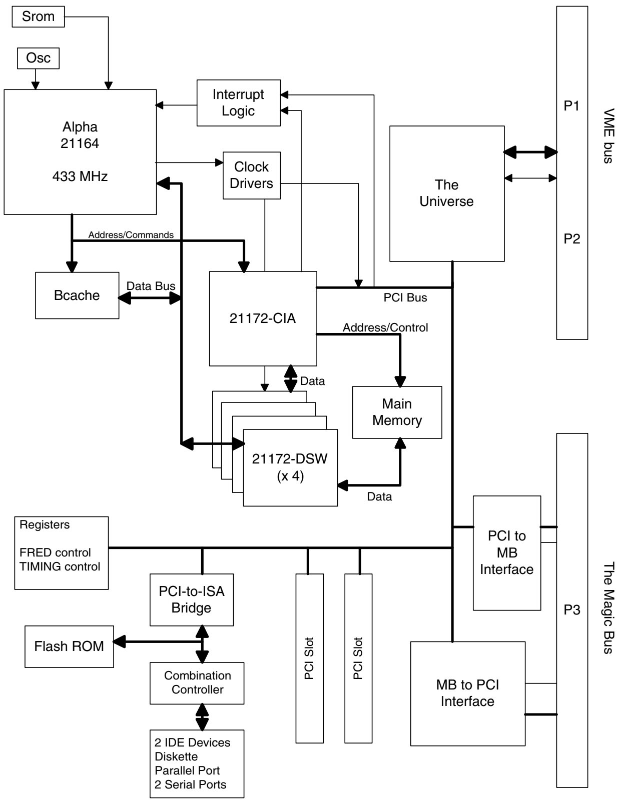

Fig. 2 The Alpha in VME Card.

B. L2 Global Processor Crate

Figure 3 shows how DØ intends to use the Alpha cards in the L2 Global Processor. Data arrives via the MBus and is broadcast to both an administrator node and a worker node. Included in the data is 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 processing. The administrator is responsible for coordinating assignment of the event buffers, and coordinates L3 readout and sending monitoring information to a trigger control computer (TCC). The administrator looks at the data only for consistency checks.

The worker analyzes events and reports the answer (a mask of 128 pass/fail bits) to the administrator and the L2 hardware framework, which uses the information to control readout of the main event data to L3. When an event has been marked as "monitoring", statistics, error messages, and processing time buffers are collected and written via VME to the dualport memory (DPmem) accessible to the trigger control computer (TCC), which will serve the data to monitoring comsumers. All handshakes between the administrator and worker take place on the MBus to avoid long latencies caused by collisions with L3 readout on the VME bus. The "Fred" register ports on the alpha boards report for scaling or logic analyzer viewing items like the current processing phase, and the current number of events in the buffers. Similar information may be reported by the MBus transceiver cards described in the next section.

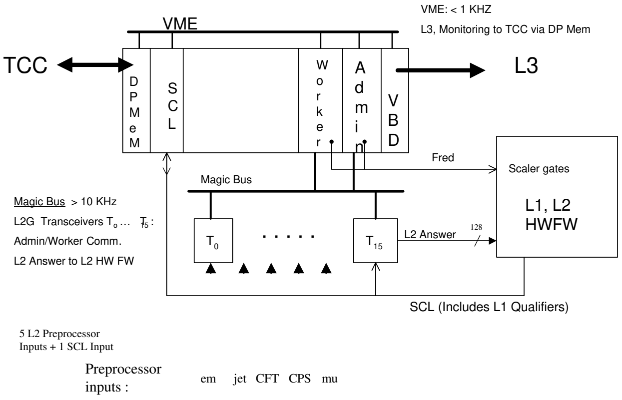

Fig. 3 L2 Global Processor Crate.

C. Transceiver Card

The I/O connected to the Magic Bus pass through a VME card dubbed the Magic Bus Transceiver (MBT) Card , whose block diagram in shown in Figure 4. 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 [9], which are 20MB/s serial data paths, probably implemented in optical fiber, with chip sets to do serial to parallel conversion at either end. Each MBT card has 8 such input paths. The MBT control logic obtains MBus mastership to do broadcast block transfer input to the Alpha cards. The MBT control logic waits for all inputs for an event to arrive, places the input event on the MBus, and generates MBus broadcast source addresses. The logic allows MBus arbitration between sources to give the Alpha cards time to carry on their handshaking in the course of processing events. The control registers are visible to MBus so that the Alpha cards can control the MBT card. There are two output paths, both using only 64 bits of MBus, as they are targets of the lower-performance MBus programmed I/O on the Alpha Cards. The 128 bit wide path is intended to transport the L2 Global event decision back to the waiting L2 hardware framework. In addition, an MBT card has 2 Hotlinks output paths. These are intended to allow L2 Calorimeter Preprocessor (next section) to send its outputs to L2 Global.

D. Software Development Environment

A developer kit is available at very modest cost which provides not only the CAD layout files for the card, but documentation and a software kit consisting of C header files and libraries allowing one to download, run, and debug code compiled and linked on an Alpha Unix workstation. All but the debugger are also available from NT. The environment is sufficient, but austere: there is no operating system, and weak support for dynamic memory allocation.

We currently are designing the control software and believe we understand the communication paths and messages, the algorithms for input and buffer control, and how we intend to monitor the system. We have run timing simulations for algorithm code and its steering code, and believe we can meet our timing goals.

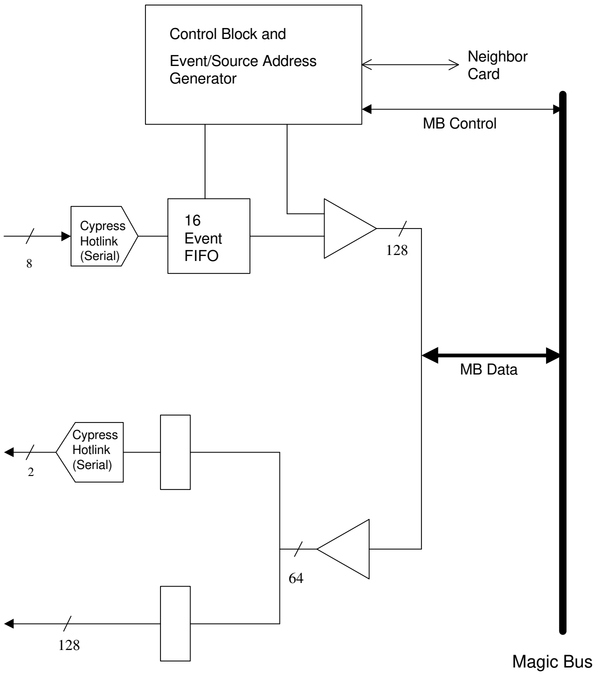

Fig. 4 Magic Bus Transceiver (MBT) Card .

III. Calorimeter Preprocessor

The calorimeter preprocessor will have the task of applying clustering to the L1 calorimeter data (.2 × .2 in $\eta \times \phi$), since the L1 trigger is tower-based. The electromagnetic clustering algorithm considers 1×2 tower regions, applies a threshold, and cuts on neighboring energy; the jet clustering considers a 5×5 tower area. It is intended to use the same basic fabric as the L2 Global processor, as shown in Figure 5.

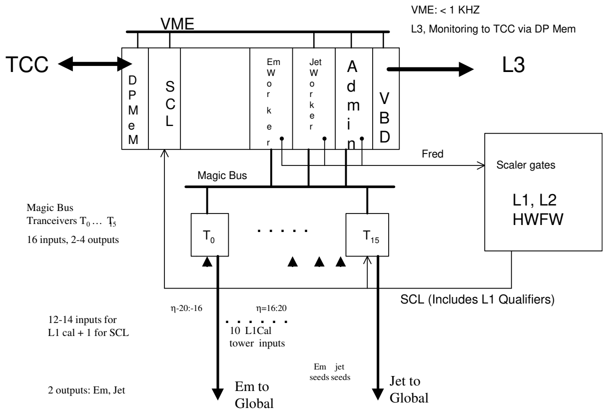

Fig. 5 L2 Calorimeter Preprocessor crate.

In this application, the demands on the input system are higher: 10 input links transport 2.6KB/event of tower transverse energy data, including electromagnetic cells (EM) and the EM+HAD(hadronic) tower sums. Another 2 links transport .3KB of bitmaps indicating which towers contain L1 candidates. A final link contains the L1 qualifier information, telling whether the event was a monitoring event, whether the preprocessor is required to run, and other high-level control information.

Data are broadcast to the administrator and two worker

Alphas: EM and Jet. In this preprocessor application, the administrator must await the message on the Serial Command Link (SCL) module indicating whether an event has failed in L2 Global, or whether it is to be read out to L3, so 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. The administrator also has to communicate with both workers for each event, and one has to choose whether to keep the workers event-synchronous. Finally, the relative bus loading on VME and MBus is reversed from the situation in L2 Global, as MBus is heavily loaded, but the VME load to L3 is relatively light: only about .1KB per passed event (1MB/s) flow on VME, while MB I/O exceeds 3KB/event or 30MB/s. Even with this loading, the latencies for inter-processor communication is still expected to be acceptable. With all I/O accounted for, both processors are expected to meet their time budgets.

IV. Muon Preprocessor

The task of the muon preprocessor[10] is probably the most challenging of any of the elements of the L2 trigger. Already in the L1 trigger, rough coincidences exist between scintillating fiber tracks and muon track candidates. The L2 muon preprocessor must fit muon track candidates using drift time information. In addition, it must avoid generating deadtime in the face of possibly high occupancy levels. Early 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 radical: apply massive parallelism in the form of arrays of DSP's, each working on distinct geographical regions of the detector, all running identical code. In some ways this enormously increases the amount of required CPU power: the tracking algorithm is typically run on mostly-empty geographical sections. However, the benefit is an algorithm whose execution time is deterministic, no matter what the occupancy level. The economics is such that sufficient DSP's can be directed at the problem that this fixed time is within the time budget. Adaptive Systems[11] sells CNAPS PCI cards each equipped with 128 20MFlop/sec DSP's. Estimates from instruction counts suggest 25μ sec/event for the track fit, and another 30μsec/event for I/O. The system shown in Figure 6 exists in an early prototype, with the algorithm in serial and parallel versions running on a single CNAPS card. Other DSP solutions are also being considered. The inputs to the system are again Cypress Hotlinks. The I/O crate is currently under design.

V. Software

Since the selected processors are programmable, the bulk of the L2 trigger effort is ultimately in software. The (smallish) fraction of the algorithm development effort aimed at implementing the algorithms on the target machines does not dominate the effort, demanding though it is. Development and testing of the algorithm is a large part of the effort. Making the algorithm available to physics users in a simulation package to design triggers is a large effort. Other efforts include using a detailed simulation to validate the simulation against the running trigger hardware; performance monitoring and bottleneck diagnosis (design of monitoring data structures, display graphics); and the download of trigger conditions and general control of the trigger system.

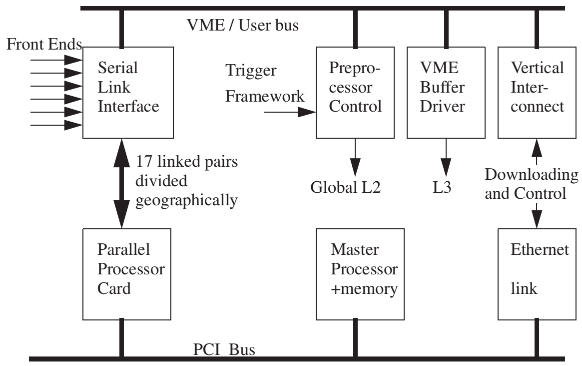

Fig. 6 L2 Muon Preprocessor crates.

VI. Conclusion

The CPU power needed for programmed processing for the L2 trigger is available commercially. The problem remains sufficiently demanding that C coding rather near the machine is required. I/O is the area where special design effort is called for, even when the solution is based on commercially available components.

VII. References

[1] G. Blazey, "The DØ Run II Trigger" in these proceedings.

[2] The L2 trigger documentation can be found on the DØ WWW page *http://d0sgi0.fnal.gov/* under "Technical —DØ Upgrade —Trigger Systems"

[3] S. Abachi *et. al* "The DØ Detector" *NIM, vol. A338, 1994* p. 185

[4] Queuing simulation results can be found on the DØ WWW page. *http://d0sgi0.fnal.gov/* under "Technical— DØ Upgrade —Trigger Systems"

[5] C. Sauer *et. al*, "The Research Queueing Package, Version 2," RA138, RA139 IBM Research Division, San Jose CA.

[6] A. O. Allen, "Probability, Statistics and Queueing Theory," Academic Press, 1978, and H. G. Perros, "Queueing Networks with Blocking," Oxford Press, 1994.

[7] The Alpha Processor and boards are described on the Digital web page. *http://www.digital.com/semiconductor/alpha/alpha.html*

[8] The Universe chip is described on Tundra's web page. *http://www.tundra.com*

[9] Cypress HotLinks products are described on the Cypress web page. *http://www.cypress.com/cypress/prodgate/prod_top.htm*

[10] M. Fortner, "A Massively Parallel Processor for Level–2 Muon Triggers at D0", Proceedings of the AIHENP '96 conference, Lausanne, Switzerland, September, 1996. NIM; see also reference 1.

[11] Manufactured by Adaptive Solutions, Eugene, Oregon, USA