The DØ Level-3 Trigger / Data Acquisition System

Dave Cutts [a], Gordon Watts [a], Jan Hoftun [a], and Ray Zeller [b]

[a] *Physics Dept., Brown University, Providence, RI, 02912, USA*

[b] *ZRL, Bristol, RI, 02809, USA*

We describe the Level-3 trigger / data acquisition system being built for the DØ experiment at Fermilab, to be used in the Tevatron's high luminosity Run 2. We focus on the principal changes to the readout paths used to transfer data from the digitizer crates to multiport memories associated with commercial processing nodes. These new custom components use a Fibre Channel based interface for transmission and operate with a simple protocol that provides an inexpensive solution to parallelizing the data flow.

Key words: DØ; trigger; data acquisition; farm

1 Introduction

The upgrade to the DØ high level trigger and data acquisition system builds on the strength of the existing processor farm architecture and the associated data path components in forming the Level-3 trigger/data acquisition system. The principal changes to the Collider Run 1 system are (1) the replacement of the existing processors with the best low cost, high performance commodity systems available in '98 - '99, and (2) a further parallelization of the data paths into the processor farm. This note will focus on the data path upgrade. Related material and other details concerning the upgrade of the DØ Detector for Collider Run 2 may be found on our web pages[1].

We have described previously[2] the trigger/data acquisition system used in DØ's Run 1. In this system, data flowed from buffer-driver boards (VBDs) in each of 81 digitizing VME crates directly over 8 parallel 48 Mbyte/s data paths, to 8 input ports of multiport memory boards (MPMs) associated with a selected filter processor node. Readout of the VME crates associated with a specific sub-detector was handled with one of the data paths; this data cable was connected to a specific multiport channel in each of the 48 nodes. Normally only one node's MPM channels were enabled to receive an event, although in calibration and "shadow" operation multiple selected nodes received the same event. At any time only data blocks associated with one event were readout. Although the Run 1 data paths had a peak bandwidth of 384 Mbytes/s, they were limited to about 20% occupancy by backplane delays in the VME digitizing crates and by delays in node selection, which was done sequentially upon receipt of a hardware trigger, before readout commenced.

The nominal data handling requirements for the Run 2 Level-3 trigger system are a 1 kHz input rate from the digitizing crates of 250 Kbyte events, and an output rate to the host cluster and recording systems of 20 Hz; the corresponding operational numbers from Run 1 were 160 Hz to Level-3 (of 450 Kbyte events) and 5 Hz to the host. The 1 kHz input requirement for Run 2 is based on estimates[3] of trigger rates for high p_T physics at 2×10^{32} cm$^{-2}s^{-1}$. We consider these nominal data handling limits as somewhat soft, and designed a Level-3 system that could provide additional bandwidth capability to handle either unexpected but important high-rate physics or non-physics (detector or beam related) sources of additional data load. Over the last year there have been discussions of significantly higher trigger rates for the later stages of Run 2, and for Run 3 or TeV33. Naturally we are interested in tackling this challenge comprehensively and in creating a framework to meet the needs of Run 2 which would accommodate further increases in bandwidth in the future.

For Run 2 we may have upwards of 100 VBDs in the digitizing crates, not to mention nearly 400 MPMs distributed throughout the filter processing farm. These buffer memory boards remain well matched to the I/O capabilities of the digitizing crates and processor buses. We will augment the data paths between the VBDs and MPMs to permit data blocks to flow freely, including those from different events. As shown in Figure 1, the Run 1 architecture is one of multiple sources and a single sink (albeit with 8-fold parallelism). For Run 2 we will operate as multiple sources and multiple sinks. Commercial components exist for supporting such data flow; these are being adapted by other groups (as for the LHC). DØ will instead extend the same dedicated, low overhead data flow design used in Run 1, making use of new commercial technology (such as Fibre Channel transceiver chips) where appropriate. This strategy is cost-effective yet permits significantly higher bandwidth, and is easily scaled to handle rates beyond those now anticipated for Run 2.

2 Components of the DØ Level-3 Data Acquisition

The components of the DØ data acquisition are shown schematically in Figure 1. Both the VBDs and the MPMs exist, and provide the sources and sinks for the data flow. New modules which operate generically as “collectors” and “distributors” are inserted into the data path to achieve the necessary parallelism. We also discuss the Event Tag Generator, not directly in the readout path,

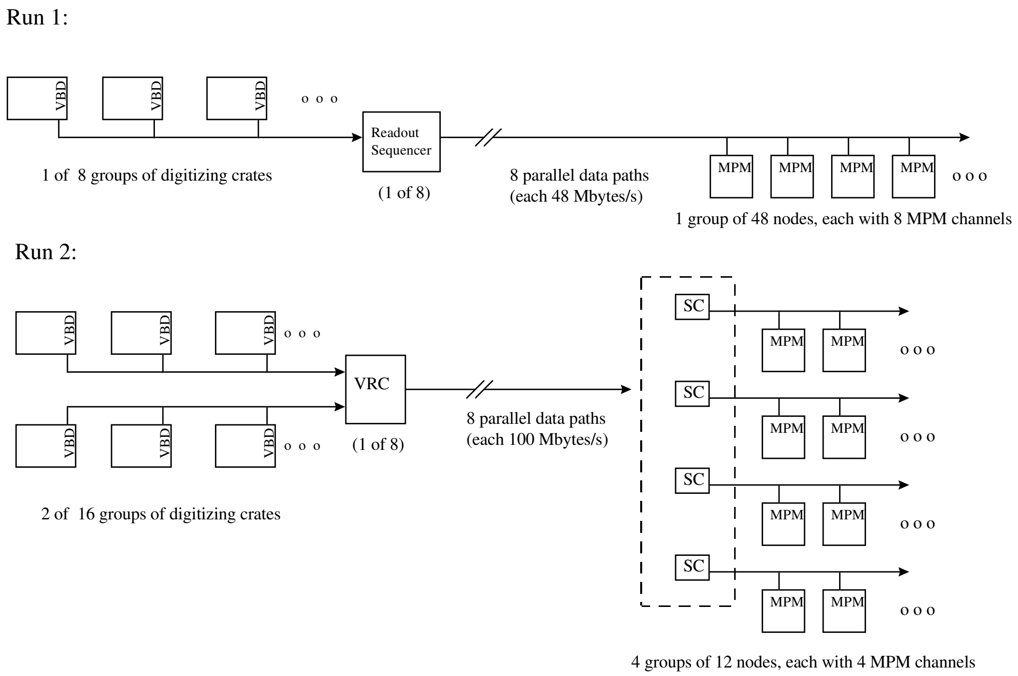

Fig. 1. Comparison of the DØ Run 1 data acquisition, with one stage of the Run 2 system.

which operates somewhat like the Run 1 Supervisor system[2] in assigning events to specific Level-3 nodes. Figure 2 provides a more detailed diagram of a specific path through the data acquisition system.

- **VME buffer driver modules (VBDs)** The VBD[2] provides the connection between data buffers in a VME digitizing crate and a data path to the filter farm. A VBD has dual buffers, one of which can be collecting data on the VME backplane [at 30 Mbyte/s for large blocks] while the other transfers a block onto an external 48 Mbyte/s datapath. For Run 2 the upgrades to the onboard list processors, performed on the VBDs used in the high load crates, in Run 1, will be completed on all modules.
- **VBD readout concentrators (VRCs)** As in the Run 1 design[2], readout of each group of VBDs is performed by circulating a token; while in possession of the token a VBD writes its output buffer onto the data path. The VRCs replace the Run 1 Sequencers[2], which coordinated event-specific readout. In Run 2, the output will occur whenever data is available, independent if their event number. Data blocks flow to each VRC over two 48 Mbyte/s data cables, from two groups of VBDs; in the VRC the blocks are buffered in an internal FIFO and output on a 100 Mbyte/s Fibre Channel (FC) port. Each VRC is located in the electronics counting house close to the two groups of associated VME crates.
- **L3 Segment Controllers (SCs)** As shown in Figure 1, the Level-3 farm is subdivided into separate segments with different data paths, to increase the system bandwidth. A Level-3 segment is a collection of processing nodes

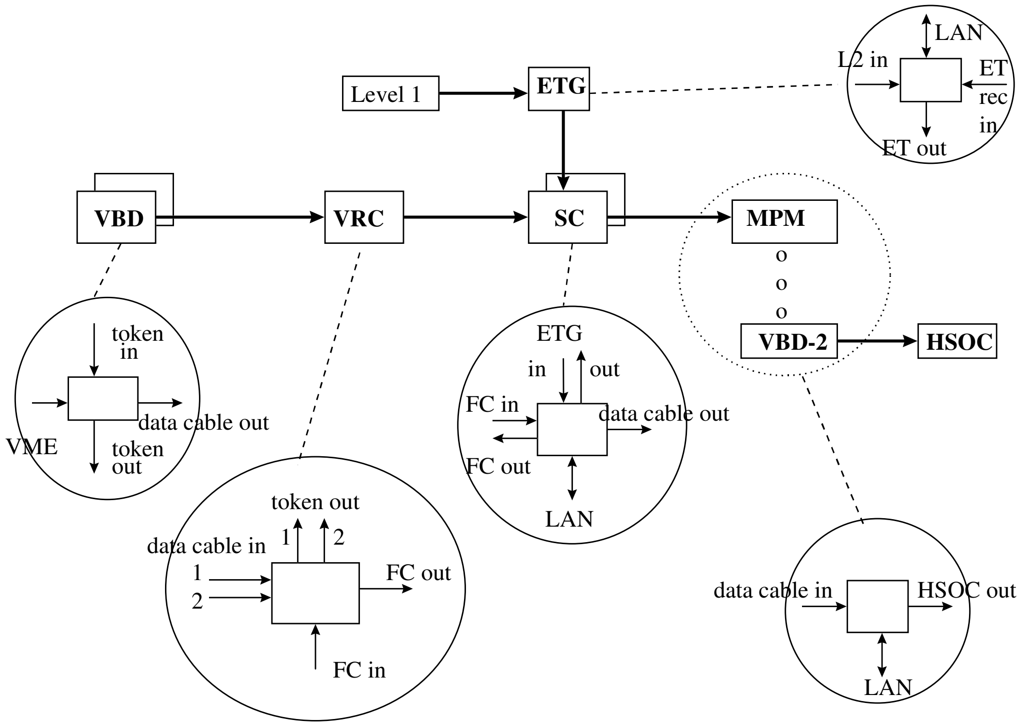

Fig. 2. Details of the DØ data acquisition components.

which share the same data cable paths, with each data cable bussed to a specific MPM channel for every node in the segment. In terms of the existing cable layout, this segmentation is very easy to achieve. As indicated in Figure 2, each SC receives data blocks from two VRCs via Fibre Channel, and either absorbs the blocks in an input FIFO or passes them via FC to the next SC servicing a different Level-3 segment. Data that cannot be handled by any segment is returned to its source VRC via FC. To balance the data flow from the two VRCs into each SC card servicing 48 Mbyte/s MPM channels, there should be four segments of the Level-3 farm, as shown in Figure 1.

– **Multiport Memories (MPMs)** As in Run 1, the MPMs are the receiving elements of the data link between the digitizing electronics and the filter processing farm. The MPM[2] is a VME board with two independent channels; each channel has an external port capable of receiving data at 48 Mbytes/s into its own 2 Mbytes of memory. This memory is multiported; it can be accessed via the external port, VME, VSB, or a special function port. A bus adapter provides a link from the memories in the VME crate to a commercial processor system. The Run 1 MPMs will receive a small modification in the form of a daughter card to permit address recognition, so that data blocks from multiple events can share a specific segment path. Additional improvements to be included on the MPM daughter card are under consideration, such as the ability to receive data at higher rates than 48 Mbyte/s, together with a single level of buffering, for rate smoothing.

- **VME-PCI Adapter** The adapter provides a high speed DMA link between MPM boards and the Level-3 node's filter processing system. The adapter permits simultaneous transfer on the data for an event into private processor memory, while filter analyses on a previous event (or events) are in progress.
- **Event Tag Generator (ETG)** This new module facilitates part of the node assignment function that was handled by the Supervisor processor, in Run 1. The ETG receives from the Level-2 trigger an event number and a list of specific crates to be readout. The ETG correlates this information with previously loaded Level-3 node configuration data and constructs an event tag which it circulates to all segment controllers. The event tag includes the event number plus additional information which can be used to target the event's data blocks to a specific node, group of nodes, or segment. In normal running, the ETG will generate tags such that the event data will be distributed by the SC's to an available Level-3 node, as described below. Based on our experience in Run 1, the ability to direct particular triggers to a specific node, or group of nodes, is very useful. This specific node assignment may be such that the event data is received by a single node in a specified group, or in fact by all the nodes in a group.

3 Data Flow

A data block moves through three principal stages from a VME digitizing crate to a Level-3 filter processing node:

- transport from the VBD buffer to a VRC's input FIFO,
- movement out of the VRC's FIFO into a FIFO in a SC, and
- transfer from the SC to the targeted MPM in a Level-3 node.

The readout of the data blocks from the VBDs is controlled via tokens, circulated by the VRCs. The tokens do not contain event number information, and the VBDs do not perform event number matching as they did in Run 1. Rather, the VBDs are configured in a free-run mode, and data blocks for any event flow as soon as the VBD receives a readout token. The VBD provides the data block with a header containing its event number which is used later in the system to direct the data block to the appropriate Level-3 node's MPM.

The VRC acts as a bridge between the VBD readout loop and the circulation of the data block to the segment controllers of the Level-3 farm. The data block flows on the Fibre Channel (FC) output port to the 100 Mbyte/s "long haul" cable from the moving counting house (MCH) to the distributor in the fixed counting house (FCH). This primary FC data path is daisy chained to each of the segment controllers serving a specific MPM channel. The FC output of the last segment controller in the chain returns to a FC input of the VRC. This provides a recirculation mechanism for data blocks which are not accepted by a segment controller during their transit around the loop. Both the data cable inputs on the VRC and the FC inputs on the segment controllers have multi-stage FIFO buffering. The recirculation design includes built-in rate control, as the controller does not circulate a new readout token to either of its VBD loops unless the FIFO for recirculation data blocks is empty. Arbitration logic on the controller determines if the next data block sent on the Fibre Channel path is from the recirculation FIFO or the data cable FIFO. This logic also checks for various error conditions and generates flags appropriately, as in the case when data blocks have been recirculated too many times.

In a similar fashion, the segment controllers act as a bridge between the flow of data blocks from the MCH and their flow to the Level-3 farm. As described earlier, the Level-3 farm is divided into separate segments; and for each segment there are segment controllers, with an individual card for each individual data cable and associated MPM channel. As the data blocks arrive in an input FIFO queue, the segment controller makes a comparison of its event number with those contained within its copy of the event tag list, to determine if the data block is targeted for its segment. The segment controller also checks the fill status of a FIFO holding data blocks for transfer to the Level-3 farm segment. Data blocks that are targeted for the segment move to this FIFO; others are transferred via FC to another segment controller, with status bits in its header set appropriately. Data blocks in the FIFO will then move onto the data cable, with the node address being set on the control lines so that the block is copied into the targeted MPM. The controller maintains a list of the data blocks received which it uses to determine when the transfer of all the blocks on that MPM channel have been sent; it toggles a control line to signal the MPM of this completion.

The node assignment for a new event is made through the interaction of the segment controllers and the ETG. When the hardware trigger is received, the ETG generates an event tag. As described above, the ETG creates header bits that described the intended segment controller action. This tag is then sent on a recirculation loop to each of the segment controllers. Event tags which have not been preassigned by the ETG to a specific node(s) or segment can be accepted by any segment controller with space in its node/tag list. The controller uses data on hardware lines from each Level-3 system to determine node availability and make the assignment of a new event to a node in its segment. Watchdog times on each segment controller count the elapsed time after the start of a new event as well as the time interval between packets for the same event.

Shown also in Figure 2 are elements of the data path for events which pass a Level-3 node's filter analysis. This path involves transfer from a VBD in each node to a MPM associated with a high speed output controller (HSOC).

4 Summary

The DØ Level-3 system described here has been extensively modeled[4]. This work has verified different aspects of the design, demonstrating for example the natural throttle provided by data block recirculation, which occurs only when the data flow reaches system capacity. Simulations have also illustrated the inherent scalability of the design, in which the addition of data path components provides straightforward bandwidth increments.

The Run 1 trigger processor nodes were based on successive generations of VAXstations (through VS-90's) running VAXELN. For Run 2 the Level-3 filter nodes as well as the data acquisition readout control systems will be based on commercial processors running Windows-NT. Development work involving NT Level-3 systems is underway.

In summary, the Run 2 Level-3 trigger and data acquisition is a natural upgrade to DØ's Run 1 data acquisition and high level trigger processor farm. Both the replacement of the processors and the upgrade of the readout/control system make use of the existing structure and components of the data paths. High performance commercial processors with the best price/performance will fit easily into this framework. The modifications to the data acquisition system adds additional parallelism in stages as increased performance is required. Control and filter framework software make efficient use of the Level-3 resources. The Level-3 trigger and data acquisition are well matched to the needs of DØ.

This work is supported in part by U.S. Department of Energy Grant DE-FG02-91ER40688.

References

[1] `http://d0sgi0.fnal.gov/hardware/upgrade/upgrade.html`

[2] Dave Cutts, *et. al.* Proceedings of CHEP'92, Annecy, p. 262 (1992) *Operation of the DØ Data Acquisition System*, and references cited therin.

[3] J. Qian, ed., Fermilab Pub-96/357-E, *The DØ Upgrade, the Detector and Its Physics.*

[4] Gordon Watts, *et. al.* CHEP'97 *Simulation of the DØ Level-3 Trigger System.*