Design and Performance of the DZero Data Acquisition System

G. Briskin[1], M. Clements[1], D. Cutts[1], K. Kakaz[2], S. Mattingly[1], G. Watts[2], R. Zeller[3]
[1]Dept. of Physics, Brown University, 182 Hope Street, Providence, RI, 02912, USA
[2]Dept. of Physics, University of Washington, P.O. Box 351560, Seattle, WA, 98195, USA
[3]Zeller Research Limited, Bristol, RI, 02809, USA

Abstract

The D0 Run 2 data acquisition system features a novel design permitting complex event routing in a high speed, data driven architecture. Data blocks associated with multiple events flow freely from ~70 front-end digitization crates. Eight separate readout concentrators funnel these blocks onto optical fiber paths to the Level 3 trigger processor farm. The processor farm is divided into segments, each of which is fed by a segment bridge. Data blocks flow to each bridge in turn, where they are either routed to a Level 3 node (or nodes) in the segment, or passed onto the next segment bridge. Blocks that do not flow to a processor node are returned to the readout concentrator. Thus the system has a built-in mechanism for rate limitation. The Event Tag Generator (ETG) and segment bridges facilitate the event routing decisions in real time. The routing is based on the trigger bits from the hardware trigger. This and the availability of nodes in each segment determine to which node(s) the event is routed. We present operational and performance details of this system, designed for comfortably more than the nominal 1 kHz input; being modular and parallel, it can be easily upgraded.

I. INTRODUCTION

The D0 Collaboration operates a high-energy physics detector situated at the Fermilab Tevatron proton-antiproton collider. For the upcoming Tevatron Run 2, the D0 detector has approximately 800,000 channels that will be read out at a nominal rate of 1 kHz [1]. These 800,000 readout channels will provide an average of 250 KB per event, which means a total data rate of 250 MB per second that will have to be collected by the data acquisition system. With upgrades to some of the D0 digitizing electronics, the readout rate could be raised as high as 4 kHz and simulations of the data acquisition system show that the system can be expanded easily to accommodate readout rates as high as 10 kHz.

Two levels of hardware/firmware triggers reduce the 1 MHz primary interaction rate to the nominal 1 KHz of full digitization. The data for each selected interaction event flows from 60-80 readout crates to a single Farm Node where the event is assembled and the Level 3 trigger analysis determines if the event should be kept and forwarded to a tape storage system. As shown in Figure 1, the data from each readout crate enters the data acquisition system via a VME Buffer Driver (VBD) board in the crate. The VBDs send the data blocks to VBD Readout Concentrators (VRC), each of which collects data from up to 32 VBDs. The VRCs forward the data to Segment Bridges (SB) which route the data to an Intel-based Farm Node where the Level 3 trigger filters are applied [2]. The Segment Bridges' routing is based on an event tag that is created by the Event Tag Generator (ETG), which receives from the hardware trigger framework the event number and trigger bits resulting from the Level 1 and Level 2 triggers. A Level 3 Supervisor node handles the configuration of the data acquisition system.

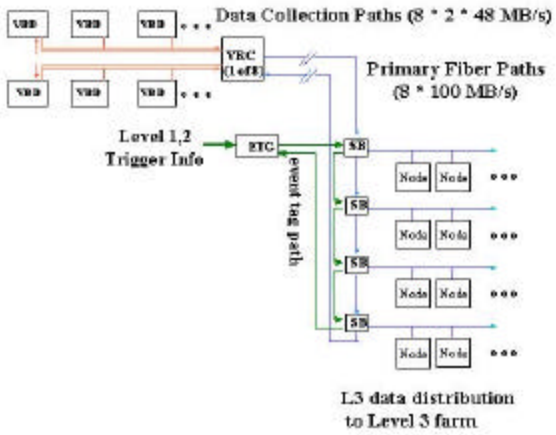

Figure 1: The D0 Data Acquisition System

The VBDs send their data blocks to the VRCs via a Data Cable consisting of a 64-conductor twisted pair "twist-n-flat" copper cable as well as a similar 26-conductor cable. The Data Cable is operated with FutureBus transceivers and has a bandwidth of 48 MB per second. There are up to 16 VBDs on each Data Cable and each VRC has two Data Cables attached to it, making each VRC a data sink for 96 MB per second. There are 8 VRCs making for a total VBD to VRC bandwidth of 768 MB per second (see Figure 1). The 8 VRCs each send data to the 4 Segment Bridges via an optical fiber recirculation loop that has a bandwidth of 100 MB/s, giving a total VRC to SB bandwidth of 800 MB/s. If the first SB in the loop does not route the data to a Farm Node on its segment, the data passes to the next SB. If the data makes it to the last SB and is not taken by that SB, the data returns to the VRC for recirculation. Data flows from a SB to its segment's Farm Nodes via 4

copper 100 MB/s cables, yielding a total SB to Farm Node bandwidth (from 4 SBs) of 1.6 GB/s. The event tag created by the ETG, which determines the SBs' routing decisions, is sent to the SBs via a separate recirculation loop as shown below.

This architecture has a number of unique strengths. Arbitration logic in the VRCs gives preference for the transmission of recirculated blocks over newly arrived data blocks from the VBDs. Thus the recirculation loops provide automatic flow control and backpressure, allowing near 100% utilization of the bandwidth of the system; no traffic shaping is necessary. Individual data blocks for different events independently flow asynchronously and in one direction. The data block header contains the event number and crate identification and the routing of the data block is done based on this information. This simplicity allows communications between components of the data acquisition system to be minimal and simple. The sender of a data block does not need to know where the data needs to go. The routing is done based only on Farm Node availability and on the trigger bits associated with an event. Custom network modules hosted by commodity PCs facilitate the data flow. The custom modules work without interference from the PC's CPU, so there is no network load on the CPU. The current design of the system has a total bandwidth of 800 MB per second and with the addition of more VRCs and/or SBs, the bandwidth is easily scaled up.

II. The VME Buffer Driver

The VBD was used during Run 1 of the Tevatron and proved to be very reliable. Each VBD is placed in a digitizing or readout crate and loads into an internal buffer 4-6 KB per event via the VME back plane. When the VBD receives a token from its VRC, it sends the data to the VRC on the 48 MB/s Data Cable. The VBD includes a VME interface with list processing DMA, dual SRAM buffers and an external data cable interface with token arbitration logic. Reading over the VME back plane is done in block mode at 25-30 MB per second and token arbitration time is less than 10 ?s.

III. The Serial Interface Board

The VRC, SB and the Farm Nodes all are built from commodity PCs running Windows NT and hosting one or more Serial Interface Buffer (SIB) boards. The SIB includes one high-speed serial link (one input and one output) using the Inova chipset, which provides 1.32 Gb/s in each direction. In addition, there is a low speed serial LVDS link that has a bandwidth of 50 MB/s in each direction. On the SIBs used as VRCs, there is a fiber link to a separate panel to service two 48 MB/s ports for the Data Cables. The SIBs are PCI cards and include an i960 processor that serves as a bridge to the local PC. The SIB also includes a main control Altera FPGA for controlling the serial links as well as the main 128 MB of SDRAM memory. This main memory is 72 bits wide, time sliced with a 10 ns clock and is a standard sized DIMM that can be exchanged with a 256 MB DIMM should a larger memory be found to be advantageous. An additional pair of smaller FPGAs is used in the VRC version of the SIB to control the two VBD readout channels. The main memory, the FPGAs and the i960 are all connected to a local bus and the host PC can reprogram the FPGAs via mapped local bus registers. The only changes necessary for each different application can be made via firmware changes to the FPGAs.

A. *The VME Readout Concentrator and Segment Bridge*

A single SIB is used to construct a VRC. The data flows from the data crates, each with a VBD, into one of the 8 VRCs and then continues to the SBs through the 100 MB/s Inova connection.

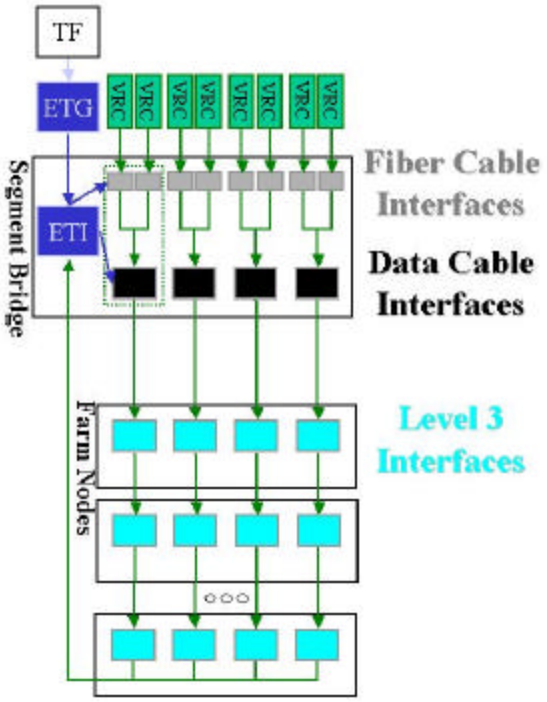

Figure 2: Detailed Data Flow Diagram

A detailed view of the data flow from the VRCs through to the Farm Nodes can be seen in Figure 2. Each VRC recirculation loop sends its data to one of 8 SIBs configured as a Fiber Cable Interface (FCI) board in a SB. Groups of two FCIs forward data to one SIB configured as a Data Cable Interface (DCI) board, which sends the data down a 100 MB/s copper cable to the daisy chained Farm Nodes. Each group of 2 FCIs and 1 DCI is called a Cable Group and is hosted by a single PC.

As mentioned earlier, the ETG receives the hardware trigger information and circulates an event tag to the Segment Bridges. The tag is received in a SB by an Event Tag Interface

(ETI) board, which decides to handle the event or to pass the tag to another SB. If it accepts an event, the ETI provides the event number to each FCI in its SB. Then, when the digitized data arrives, the FCI uses a content addressable memory to check the event number in the header of each incoming data block and if the event number matches one in its list, the data block is sent to the DCI. The DCI receives a list of crates to expect for a given event from the FCI and checks off the crates as they flow through. The ETI informs the DCI of the availability of Farm Nodes and the DCI sends the data to the appropriate node. The farm nodes are able to add information asynchronously about their availability to the data flowing down the daisy chain without slowing or clobbering the data. This information is passed to the ETI at the end of the Farm Node daisy chain. No data flows back to the ETI. Each SB consists of 4 Cable Groups and one ETI, and is hosted by a total of 5 PCs. A more detailed view of part of a Cable Group can be seen in Figure 2.

The data flows into the SB from the previous SB in the loop or the VRC (in the case of the first SB in the loop) via the high-speed Inova link on an FCI. The FCI receives event tag information from the ETI through the lower-speed LVDS link. The data either flows on to the next SB or back to the VRC (in the case of the last SB in the loop) or, if the FCI decides to send this data to one of it's Farm Nodes, the data is sent to the DCI via the PCI bus. The DCI then forwards the data to the Farm Nodes through its Inova link via a copper cable. The DCI receives node availability information from the ETI via its lower-speed LVDS link.

B. *The Event Tag Generator*

The ETG receives from the hardware trigger framework a 16-bit event number for each event as well as 128 bits of trigger information. The ETG also has the ability to turn off individual triggers in the trigger framework via 128 bits of veto information. On receipt of the trigger information, the ETG creates an event tag for the event; this tag includes the event number together with an event class, which can be associated with specific Farm Nodes. The ETG uses the 128 bits of trigger information and a content addressable memory (downloaded each run) to make the association between the event's hardware trigger and the group of nodes appropriate to receive the event. This logic allows targeting specific triggers to a specific filter Farm Node, or to a group of nodes.

It also permits the use of secondary nodes to receive specific triggers in a "shadow" mode. Once created, the event tag is circulated to the SBs. The event tag is returned to the ETG for recirculation if, because no nodes are available, none of the SBs can service the event's readout. As with the primary data path, this recirculation loop optimizes the use of the resources - in this case, the Level 3 filter Farm Nodes.

C. *The Farm Node*

Each Farm Node contains 4 SIBs configured as L3 Interface (L3I) boards. Each L3I receives data blocks from a previous Farm Node or the SB's DCI (in the case of the first Farm Node in the chain) and can either take the data and use it to add to an event being built or it can send the data block to the next Farm Node in the chain or back to the ETI in the SB (in the case of the last Farm Node in the chain). The Farm Nodes also can send a READY signal that is circulated by the other nodes in this way; the ETI is informed asynchronously of a Farm Node's availability. Each L3I has 128 MB of memory, so the entire Farm (~100 Farm Nodes) contains enough buffering for more than 1 minute of data taking at 1 kHz. Each Farm Node is an Intel symmetric multiprocessor system. One Level 3 filter process is run on each processor. Passed events are sent to the tape storage system via a dedicated 100 Mb per second Ethernet link.

IV. STATUS

Currently, the system has been fully simulated and been found to be scalable to a 10 kHz (2.5 GB per second) rate. D0 is using a low speed Ethernet-based version of the system for data taking during commissioning of the detector. A driver for the SIB has been written and simple control code using this driver has been minimally tested. The first batch of final hardware is expected to arrive soon and will immediately be tested using a dedicated test stand at Brown University. Monitoring and control structures for this system have been developed and are discussed in a separate paper submitted to these proceedings [3].

V. REFERENCES

[1] G. Briskin, D. Cutts, A. Karachintsev, S. Mattingly, G. Watts, R. Zeller, *The D0 Data Acquisition System and its Operational Control*, Proceedings of CHEP2000, 358 (2000).

[2] A. Boehnlein, G. Briskin, G. Brooijmans, D. Claes, D. Cutts, S. Mattingly, M. Souza, G. Watts, *Data Handling and Filter Frameworks for the D0 L3 Trigger System*, Proceedings of CHEP2000, 353 (2000).

[3] G. Briskin, M. Clements, D. Cutts, K. Kazkaz, S. Mattingly, G. Watts, R. Zeller, *Control of Large Scale Distributed DAQ/Trigger Systems in the Networked PC Era*, DAQ2000 Workshop, NS2000 (this conference)