A FDDI BASED PROCESSOR FARM FOR THE ZEUS THIRD LEVEL TRIGGER

Stefan Polenz [a,1,2], Robert S. Orr [a,1], Vladimir Rybnikov [b,2]

[a] *Department of Physics, University of Toronto, 60 St George St, Toronto, Ontario, M5S 1A7, Canada, e-mail: polenz@desy.de*

[b] *Nuclear Physics Institute, Moscow State University, Vorobjevy Gory, 119899 Moscow, Russia*

To meet the requirements of computing power and bandwidth the ZEUS Third Level Trigger (TLT) is realized in multiple FDDI loops. Each loop consists of UNIX workstations and one VME/FDDI board for data input from VME memory. For the data transport via FDDI from the VME board to the UNIX stations a "reliable" UDP protocol for high throughput was developed. The FDDI loops are connected by a network switch to the TLT data sink. The software is well separated into different process groups to provide easy maintenance of the physics filter code, and to allow simple portability to other data transport media. The entire process system is self synchronizing.

Key words: DAQ Systems; Trigger Systems; Processor Farm;

1 Introduction

ZEUS is one of the HERA ep-collider experiments at the DESY laboratory in Hamburg. The requirements of the ZEUS trigger system are met by implementing three distinct trigger layers. Under current beam conditions the First Level Trigger accept rate is 500 Hzand the Second Level Trigger rate is 80 Hz. At the HERA design luminosity the TLT input rate is forseen to be 100 Hz.

The TLT system described in this paper is part of a project to completely replace the original FNAL Branch Bus system, which has been operating successfully since 1992 [1].The motivation for this upgrade was the demand for more cpu power required by the ever increasing complexity of the reconstruction algorithms used. The dataflow bandwidth (80 kBytes events at 100 Hz) has remained approximately the same.

The new system uses FDDI as the data transport link; FDDI has ample bandwidth and interfaces have become available "off the shelf" for workstations and VMEbus.

2 Hardware Configuration

On a positive decision of the ZEUS Second Level Trigger the Event Builder (EVB) gathers all the component data and writes the whole event into VME memory (Fig. 1). Two memories are read out

[1] Now at: Department of Physics & Astronomy, University of British Columbia, 6224 Agricultural Rd, Vancouver B.C., V6T 1Z1, Canada
[2] At: DESY, ZEUS-Collaboration, Notkestr. 85, D-22603 Hamburg, Germany, phone: +49-40-8998-3443, fax: +49-40-8998-3092

Preprint submitted to Elsevier Preprint *3 March 1997*

over VMEbus by one PME-FDDI-1 board (manufactured by Radstone Technology PLC) on which WindRiver's operating system VxWorks is running.

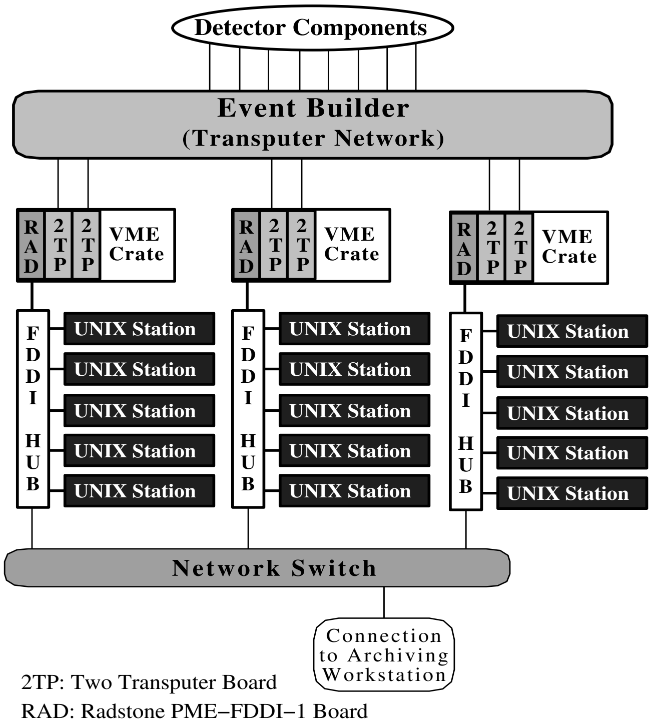

Fig. 1. The TLT Hardware

On request, the PME-FDDI-1 board sends events via the FDDI loop to the target Silicon Graphics Challenge S UNIX station. The workstations are connected to the FDDI loop by a SAS interface; each station contains a R4400 MIPS processor. The three FDDI loops are connected to a network switch, thus only outgoing packages to the archiving workstation cross the switch.

3 Process Design

The TLT processes are devided into the control, transport, filter and monitor process groups which are independent of each other, apart from Interprocess Communication (see Fig. 2). All processes in the system are event driven, i. e. only on an external message or signal does a process start or continue execution. Further the processes are ordered in a hierarchy as shown in Fig. 2. The top process (TLT Run Control) is the communication interface to the Central Run Control (CRC) and handles all state transitions of the TLT. The states are: *IDLE*, where the TLT is available but inactive, *READY*, in which case all processes are configured for the next data taking, and *ACTIVE* for actual data taking. The next level in the hierarchy are the "Branch Controllers" which in turn control all the "Analyzers" and the PME-FDDI-1 board on the FDDI loop.

All processes of the Control Process group deal with the state transitions. When starting up, all control processes wait for messages from the appropriate processes in the hierarchy; by exchanging state information messages they are able to synchronize. A restarted process can join the system in any state or state transition. The Analyze-Controller is the parent process of the Input and Output Tasks and of the Physics Analyzer (in case of a multi processor machine it can control an accordingly higher number of Physics Analyzers). While in the *ACTIVE* state it requests events from the Input Task, prompts the Physics Analyzer to deal with the event and if required sends a command to the Output Task to write the event to the archiver.

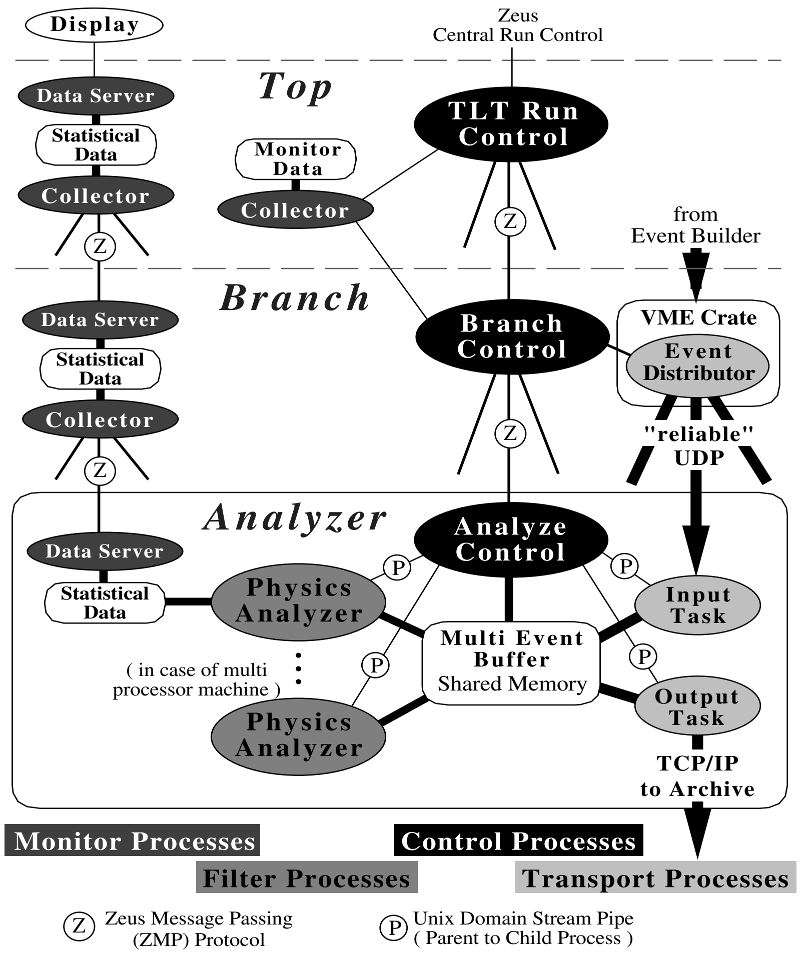

Fig. 2. The TLT Process Scheme

The jobs of the Transport Process group handle the data flow from the EVB to the archiving workstation. They are the only processes which use the FDDI. The Event-Distributor, which runs on the PME-FDDI-1 board, sends on request events to the Input Tasks. On the "Analyzer" the events are written to a multi-event-buffer from where, after a positive decision of the physics filters, they are sent to the archiver by the Output Task. The Input and Output Tasks are spawned at the transition to *READY* and terminated at the transition to *IDLE*.

The only process in the Filter Process group is the Physics Analyzer. While being analyzed, an event stays in the multi-event-buffer, the results of the physics filter are appended only if the event is accepted. For monitoring purposes, the filter job keeps track of various statistical data during a run. This information is written to shared memory, where it can be accessed by monitoring processes.

The Monitor Processes do not have any impact on other processes, but they provide the only method of monitoring. There are three different kinds of monitoring processes, the Display (for presenting on the screen), the Server (to provide the information of the shared memory on request) and the Collector process. The latter e. g. adds up all histograms from all "Analyzers" of one branch and writes the result again into shared memory. In this way it is very easy to collect monitoring data branchwide and then again TLTwide. In addition, the Display can present data from every level of the hierarchy just by connecting to the appropriate Server.

4 Protocols

For the data transport between the PME-FDDI-1 board and the UNIX stations a "reliable" UDP (RUDP) protocol was developed [2]. This was necessary because the TCP performance on the PME-FDDI-1/VxWorks 5.2 platform was inadequate. The RUDP exploits fully the hardware features of the PME-FDDI-1 board. The events are split up in UDP packages, each package preceded by a header. By that means the destination can determine that an event was received correctly. On failure, the missing part of an event is sent again.

The ZEUS Message Passing (ZMP) protocol was developed to unify message passing for the ZEUS Data Acquisition system. It is reliable, non-blocking, connectionless (UDP based), platform independent (VMS, UNIX and OS9) and provides brokerage of named services. Inside the TLT it is used by the Monitor and the Control Processes.

5 Performance

With the RUDP protocol a sustained rate of 5.5 MByte/sec from the PME-FDDI-1 board to the workstations was achieved. The PME-FDDI-1 board itself is able to write at a rate of 7 MByte/sec from its internal DRAM to FDDI, but it is limited by the data transfer into the DRAM over VMEbus. For event transfers from the workstations to the archiever, a rate off 11.5 MByte/sec was meassured using TCP/IP. With the EVB writing out event rates of up to 30 Hz per VME memory at average event sizes of 80 kBytes, one PME-FDDI-1 board keeps up reading out two memories. This yields a 60 Hz event rate per TLT branch. The cpu power of the workstations, running the current filter code, is sufficient to handle 100 Hz event rate overall.

6 Conclusions

The system is extremely resilient, allowing processes to restart during data taking. It is, without exception, based on standard network protocols. This ensures good portability. Any demand for more bandwidth or cpu power is easily satisfied by adding more TLT branches (FDDI loops) or more workstations to the system. The separation of the filter process from the other TLT processes allows for easy modifications of the physics algorithms.

7 Acknowledgements

The authors would like to thank the ZEUS Central Data Acquisition (CDAQ) group, especially C. Youngman and T. Voss, for their support and encouragement. The ZMP protocol is provided by the CDAQ group. And we thank all the many members of the collaboration who contributed to the filter algorithms over the last years. This work is supported, in part, by NSERC, Canada.

References

[1] D. Bailey et al., The Zeus Third Level Trigger Hardware Architecture and the Control System, *Proceedings CHEP conference* Annecy (1992) 171.

[2] A. Kalagov, V. Rybnikov, Internal Communication, ZEUS–Note 95–81, 95–102 and 95–152.