Fermi National Accelerator Laboratory

FERMILAB-Pub-89/260

A Scalable Parallel Open Architecture Data Acquisition System for Low to High Rate Experiments, Test Beams & All SSC Detectors*

Ed Barsotti, Alexander Booth, Mark Bowden, and Carl Swoboda
Fermi National Accelerator Laboratory
P.O. Box 500
Batavia, Illinois 60510

and

Nigel Lockyer and Rick VanBerg
University of Pennsylvania
Philadelphia, Pennsylvania 19104

December 1989

* To be published in IEEE Trans. Nucl. Sci.

A Scalable Parallel Open Architecture Data Acquisition System For Low To High Rate Experiments, Test Beams & All SSC Detectors*

Ed Barsotti, Alexander Booth, Mark Bowden & Carl Swoboda, Fermilab,
Nigel Lockyer & Rick VanBerg, University of Pennsylvania

Abstract

A new era of high-energy physics research is beginning requiring accelerators with much higher luminosities and interaction rates in order to discover new elementary particles. As a consequence, both orders of magnitude higher data rates from the detector and online processing power, well beyond the capabilities of current high energy physics data acquisition systems, are required. This paper describes a new data acquisition system architecture which draws heavily from the communications industry, is totally parallel (i.e., without any bottlenecks), is capable of data rates of hundreds of GigaBytes per second from the detector and into an array of online processors (i.e., processor farm), and uses an open systems architecture to guarantee compatability with future commercially available online processor farms. The main features of the system architecture are standard interface ICs to detector subsystems wherever possible, fiber optic digital data transmission from the near-detector electronics, a self-routing parallel event builder, and the use of industry-supported and high-level language programmable processors in the proposed BCD system for both triggers and online filters. A brief status report of an ongoing project at Fermilab to build the self-routing parallel event builder will also be given in the paper.

Introduction

Several months ago Gil Gilchriese of the SSC started the Task Force on Electronics, Triggering and Data Acquisition for Experiments at the SSC. Members of its steering committee include chairman Andy Lankford (SLAC), Ed Barsotti (Fermilab), Robert Downing (Illinois), Fred Kirsten (LBL), Jon Thaler (Illinois), Yoshijuki Watase (KEK) and Rudy Bock (CERN). The Data Acquisition part (i.e., system architecture, data flow, event building, online processing & data storage) of the committee, headed by Ed Barsotti, has held two meetings at LBL attended by personnel from high-energy physics laboratories, universities and industry. At the first meeting, Fermilab personnel presented a new data acquisition architecture which could meet or exceed SSC requirements. This architecture has been extended throughout the Task Force meetings and at numerous other discussions and meetings. For example, the idea of a standard IC to interface to all front-end subsystems, the Data Collection IC, was presented by Rick VanBerg of the University of Pennsylvania. Industry participated heavily in presentations and discussions telling the Task Force what they were doing and/or planning to do which might help the HEP community solve the data acquisition problems at the SSC.

The heart of the new data acquisition system architecture is a truly parallel event builder which allows data rates from the detector to an online farm to increase to orders of magnitude greater than is possible in existing data acquisition systems. Data rates from the detector with this new architecture can exceed hundreds of GigaBytes per second. The following sections:

- List requirements, etc. for Fermilab's proposed BCD detector, and SSC's proposed bottom physics and 4π detectors
- List goals of a new data acquisition system
- Introduce the new proposed data acquisition system architecture
- Explain the need for system modeling and simulations to be done before system design begins
- Describe a self-routing parallel event builder development project at Fermilab
- List areas where industry can significantly contribute in building a system based on this architecture

New Data Acquisition System... Requirements & Architecture Goals

Data acquisition systems for colliding beam and fixed target experiments are limited in bandwidth to small numbers of MegaBytes per second of data into an online processor farm). For example, the following summarizes data rates for several detector data acquisition systems beginning with L3 at CERN and the Colliding Detector at Fermilab (CDF) through the large solenoid detector at the SSC. BCD refers to the proposed Bottom Collider Detector at Fermilab. It is proposed that a fraction of this experiment will run at Fermilab in the mid-1990s and the full experiment will run at the SSC. MB/s equals MegaBytes/second; GB/s equals GigaBytes/second. Filter implies online reconstruction:

	Rates Into Processor Farm		
Detector	Current	Goal	Processor Farm Use
L3	—	8 MB/s	Filter
CDF	1-2 MB/s	10 MB/s	Filter
D0	—	40 MB/s	Filter
BCD	—	30 GB/s	Trigger & Filter
SSC (4π)	—	10-100 GB/s	Filter
SSC (B)	—	100+ GB/s	Filter

The 30 GB/s and 100 GB/s data rates listed above for the BCD and SSC detectors, respectively, are only very crude upper level estimates. The actual data rates depend upon several factors such as the physics goals of the experiment and how effective Level 1 and Level 2 (i.e., prompt) triggers are at reducing the raw data rate from the detector.

* This work performed under the auspices of the United States Department of Energy.

Note the very large three orders of magnitude increase in data rate requirements between D0 and BCD. This is due in part to the 2.5 MHz BCD interaction rate and also to the fact that special combinatorial logic, sequencers, etc. are not being proposed for the usual '2nd level trigger system' in BCD. Because of the increasing availability of very high-performance industry-supported processors, it is being proposed that the BCD do only prompt triggers in the order of a microsecond with 'special' hardware. All other triggers and all filters would be done in online processor farms. Doing this will provide a more flexible, more powerful and much simpler data acquisition system. However, this architecture does inherently require higher data rates into processor farms. The SSC, with its 100 MHz interaction rate and even with prompt trigger systems near the detector, still require up to and possibly exceeding 100 GB/s data rates from the detector. Additional specifications for the BCD and SSC B & 4π detectors are given in Table 1 below. Zero supression & multiple event buffering is assumed on all front-end channels.

Table 1
BCD and B & 4π SSC Detectors
...Mini-Specifications

	Fermilab BCD	SSC B	SSC 4π
Crossing Rate:	400 nsec	16 nsec	16 nsec
Interaction Rate (& Luminosity):	0.5 MHz ...(10^{31}) 2.5 MHz ...(5×10^{31})	10 MHz ...(10^{32})	100 MHz ...(10^{33})
Subsystems:	Vertex Tracking TRD (partial) RICH EM Cal.	Vertex Tracking TRD RICH EM Cal. Muon	Vertex Tracking TRD RICH EM Cal. Muon Hadron
Event Size	150 KB	500 KB	1 MB
Total # Of Channels:	5×10^6	$>10^7$	$>10^7$
Prompt Trigger Time:	1-3 msec	5 msec	1-3 msec
Desired Prompt Trigger Rejection:	10-50	50	1000-10000
Data Rate From Detector (GB/s):	6-30	100	10-100
Event Rate Into Online μP Farm (Events/second)	(50-250)K	200K	(10-100)K
Required VAX CPU Equivalents:	(50-200)K	1-2M	(0.5-2.0)M

The architecture goals in any new data acquisition system are often quite obvious and often not adhered to. Orders of magnitude more front-end channels, higher data rates and higher online processing power requirements than that in existing data acquisition systems dictate a far more formal approach to systems engineering than ever before done in high-energy physics. Industry-accepted large project management techniques must be used from the onset of the project. System, subsystem and component requirements documents must be written very early in the project and kept up to date and adhered to throughout the project. System modelling and simulations are critical to projects with the aforementioned requirements and must begin early to prove out the architecture. Hardware and software standards must be used whenever possible. Some additional goals of a new data acquisition architecture are as follows:

- Highly parallel architecture (i.e., no 'funnels' through which data must pass)
- Easy expansion for more data sources and/or higher data rates
- Data compaction at the front-end
- Local multiple event buffering
- Simple event building
- Standard data structures throughout
- Many GigaBytes/second data rate capability
- Minimize data movement throughout the system
- Simple control structure (i.e., very few messages to keep data flowing from the detector)
- Use commercially-available hardware and software products and industry-supported standards wherever practical and cost effective
- Build the system with an 'open systems architecture' approach wherever possible such that new technically-advanced commercially-available products can be used in the system without major modifications to the system

Proposed Architecture

Figure 1 shows a simplified block diagram of the proposed data acquisition system architecture. In this example, event fragments are transmitted from the detector, in parallel, over 1024 high speed serial fiber-optic links. These links operate at a typical throughput of 100 MegaBytes per second (MB/s) for a combined data rate of 100 GigaBytes per second.

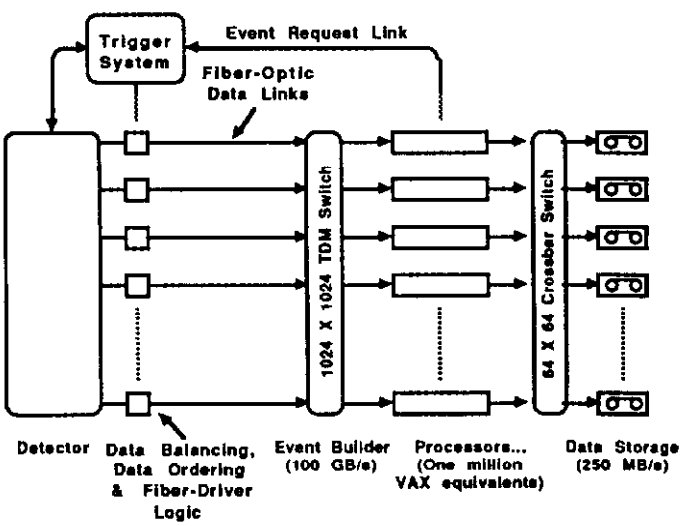

Figure 1
System Architecture Overview

The event building stage is also implemented in parallel using a switching network. The switching network (i.e., event builder) can be any of several variations commonly found in telecommunications and parallel processor system designs. One example is a simple time-division multiplexed (TDM) switch based on time-slot interchange (TSI) buffers and a multi-stage barrel shifter or crossbar, similar to a telephone central switch. The buffer memory can either be consolidated at the switch input and output ports (time-space-time) or distributed through the switch (space-time-space). The output links drive a large array of independent processors. As an alternative, the switching network can be superimposed on the processor array using local node mesh routers and packet buffers (e.g., Intel Touchstone project).

All of these options allow self-routing of data packets. Self-routing eliminates much of the normal system control overhead by permitting each processor to individually control its own data flow (e.g., any processor or group of processors can request an event of any trigger type, an event can be directed to any processor or group of processors, etc.). With the exception of the trigger, there is no central control point in the system. An Event Request Link connects the processors to the trigger system.

A farm of processors with combined performance in the range of one million VAX equivalents is likely for SSC detectors. These processors may include a mixture of standard high-level language programmable and special-purpose devices. Higher levels of integration in processors and memory within the next ten years will reduce the cost of the processor farm to the $10-$20/MIP level required for construction of million VAX equivalent systems.

Accepted events are written to optical or video tape at a much lower rate (a few hundred MegaBytes per second). A second, smaller data switch might be inserted between the online processor farm and online data storage media to allow redirection of specific event types to specific storage devices.

Figure 2 expands on the detector block of Figure 1 illustrating architecture components on and near the detector. In this example, 128-channel front-end ICs of a detector subsystem are linked together and to Data Collection ICs on the detector. These Data Collection ICs are standard interfaces to as many detector subsystems as possible. Using a standard interface across detector subsystems saves substantial resources both during system development and throughout the lifetime of the detector. Data Collection ICs can be multiplexed to other Data Collection ICs until data rates off the detector from various ICs are somewhat balanced. Outputs from Data Collection ICs drive event data tens of meters to areas near the detector but in non-radiation areas.

Highest system throughput is attained when all event data sources to a parallel event builder have similar amounts of data when averaged over hundreds of events. The Data Balancer logic can, via computer control, shift detector data sources to different fiber-optic links and into the parallel event builder.

Because event data transmitted off the detector is not event ordered (i.e., data fragments from an event can be received by the Data Ordering logic before data fragments from earlier events). The Data Ordering logic buffers event data until all the data from a particular event has been received before shipping that event's data to the parallel event builder. The Data Ordering logic also transmits events time-ordered (i.e., event n transmitted then event n+1 transmitted, etc.).

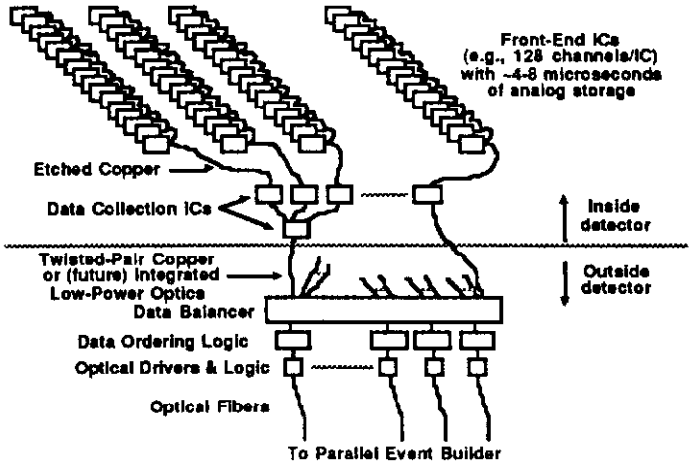

Figure 2
On & Near-Detector Logic

System Architecture Modelling & Simulations ...Why Do It?

Simulation conveys the idea of reproducible experiments and extensive exploration of the design space. Modeling conveys the idea of one or more computer programs which represent the design. In much the same way as structured analysis and structured design (SASD), modeling and simulation allows system designers to consider a large complex homogenous mass as a collection of well defined smaller pieces which individually are more easily studied and modified.

With the ever increasing complexity of detectors and their associated data acquisition systems it is important to bring together a set of tools to enable system designers, both hardware and software, to understand the whole system including the behavioral aspects and the interaction of different functional units within the system. For complex systems, human intuition is inadequate since there are simply too many variables for system designers to begin to predict how varying any subset of them affects the total system. On the other hand, exact analysis, even to the extent of investing in disposable hardware prototypes, is much too time consuming and costly. Simulation bridges well the gap between physical intuition and exact analysis by providing a learning vehicle in which the affects of varying many parameters can be analyzed and understood. In this way much time can be saved in the design process and one has significantly increased the probability of understanding not only the system as a whole but also the interaction of different sub-systems.

The purpose of modelling and simulating a switch-based data acquisition system is to provide a learning vehicle

whereby the system designer can experiment with different architectures and control mechanisms to enable us to better understand the design. Modeling and system simulations assist system designers in determining throughput for different configurations, identifying potential bottlenecks, interfacing to "physics data" simulations, identifying busiest channels, selecting proper buffer sizes, determining the number of processors required, determining data rates, etc.

A partial list of simulations which should be done as this data acquisition system architecture is being developed is given below:

- Functional simulations of normal system operation
- Effects of variations in .....
    - Number of detector channels
    - Event size
    - Event distribution
    - Front-end buffer size
    - Number of data links
    - Data link transfer rate
    - Packet lengths
    - Routing algorithms
    - Number of processors
    - Average processing time
    - Processing time distributions
    - Processor buffer sizes
    - Event request delays
- Architecture improvements resulting from system simulations
- Effects of errors .....
    - Data errors
    - Header errors
    - Routing errors
    - Buffer overflow
    - Source failures
    - Processor failures
    - Switch failures
- Architecture improvements resulting from diagnostic simulations

Self-Routing Parallel Event Builder Project ...Status Report

A prototype Self-Routing Parallel Event Builder, based on the simple TDM switch, is being developed at Fermilab with SSC Generic R&D funds. This system will contain 64 channels, each operating at a nominal 20 MB/s rate for a combined throughput of approximately one GB/s. The switch is packaged in a single 9U Eurocard crate and is expandable for higher data rate or additional data source requirements. The detector and processor farm are both emulated by test modules which transmit and receive simulated event data at full bandwidth.

The test system is monitored by a Solbourne UNIX workstation running a standard real-time graphical interface package (DataViews). Links to an expert system (NEXPERT) for diagnostics will also be investigated. The architecture will be extensively modelled and simulated using Verilog-XL. The initial modelling and simulations will be on large system blocks such as the event builder, the online processor farm, etc. Data for the simulations will come both from pseudo-random data as well as from physics Monte-Carlo simulations. Later modelling and simulations will break the larger blocks into several smaller ones and will add diagnostics to each block. The system will have the ability to switch between simulations and the prototype hardware from a common user interface.

DataViews is written in C and runs on most 32-bit workstations. It has two main components; DV-Draw and DV-Tools. DV-Draw enables users to create and modify color pictures of all the components of a DAQ system. DV-Tools allows the user to specify the dynamic interactions of all components both on each screen and between screens, and allows integrating these displays into application programs. The interaction utilities provide complete programmer control over what input and output will be available, where on the screen it will occur, and how the end user will interact with the system (menus, locator devices, keyboards, etc.). Color thresholds can be set to give the end user a quick visual prompt when critical limits have been reached or breached.

A software bridge between DataViews and NEXPERT has been developed which allows application programmers to automatically link expert systems to graphic system interfaces. (i.e., the expert system's knowledge can dynamically control the appearance of the graphics display, and in this bidirectional relationship, the graphics user interface can control rule-based processing in NEXPERT).

Verilog-XL is a hardware description language and interactive simulator which allows mixed mode simulation. It enables the designer to integrate different design levels: architectural, microcode, behavioral, register transfer as well as gate and switch levels.

The high-speed parallel nature of this system requires increased emphasis on error detection and data validation. Because data is heavily pipelined, there may be as many as several thousand events and several million event fragments buffered in the event builder at any instant. In addition, there is very little correlation between the order of arrival and the order in which event fragments are routed through the switch. Although this operation is completely transparent, it implies that a switching error may affect many thousands of events rather than the few events buffered in a traditional system.

Many of the features of this test system are made possible only by recent commercial developments, for example;

- low cost fiber-optic tranceivers and high-speed encoders, from AMD and Gazelle, to implement the serial data links
- digital signal processors and video DRAM for TSI buffer memory and control
- high-density crossbar integrated circuits (LSI Logic) for the central switch

Operation of the prototype Self-Routing Parallel Event Builder can be described using Figure 3 which illustrates a four channel system. There are four Input TSIs which buffer data from data sources S1-4 and four Output TSIs which buffer data to destinations D1-4. The TSIs are interconnected by a four channel barrel shifter (or crossbar). An event is transmitted from the detector to the event builder as four separate, parallel fragments. Each fragment contains a header word which determines the destination (self-routing event data). This destination is the same for all fragments of a given event. Each Input TSI controller places it's event fragment in the selected destination buffer.

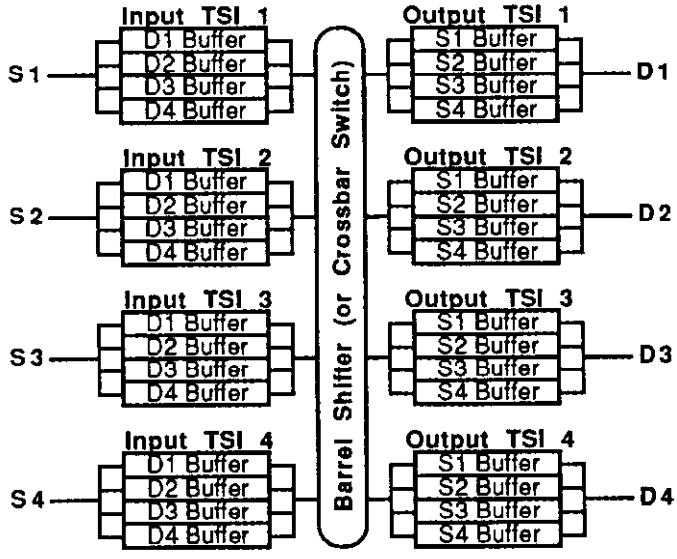

Figure 3
4 X 4 Event Builder Block Diagram

The barrel shifter follows a fixed rotation, continuously connecting each of the four input TSIs to each of the four Output TSIs. Four fixed length data packets are transmitted through the barrel shifter on each step of the rotation. The packets are selected from the Input TSI buffers which correspond to the current Output TSI connections. For example, when the barrel shifter is in its initial state (i.e., S1 connected to D1, S2 to D2, S3 to D3 and S4 to D4), a packet is copied from buffer D1 of Input TSI 1 to buffer S1 of Output TSI 1. At the same time, another packet is copied from buffer D2 of Input TSI 2 to buffer S2 of Output TSI 2. Likewise for channels 3 and 4. On the next step of the switch rotation (i.e., S1 connected to D2, S2 to D3, S3 to D4 and S4 to D1), packets are transferred from buffer D2 of Input TSI 1 to buffer S1 of Output TSI 2, from buffer D3 of Input TSI 2 to buffer S2 of Output TSI 3, from buffer D4 of Input TSI 3 to buffer S3 of Output TSI 4, and from buffer D1 of Input TSI 4 to buffer S4 of Output TSI 1.

At the Output TSI, a full event is assembled by concatenating one event fragment from each of the four buffers as the data is transmitted to the processor. Four packets of data from four different events cross the switch during each packet interval. The bandwidth of data flowing through the event builder matches the bandwidth of data from the detector.

The actual implementation of the TSI uses a single dual-port memory (video DRAM) which is logically divided into N circular buffers. The barrel shifter is implemented using a programmable crossbar configured as one or more independent barrel shifters to allow system partitioning.

In future versions of this parallel event builder, it is likely that integrated opto-electronic switching will lead to increased throughput and reduced cost. Although the prototype event builder is packaged as a standalone unit, it can also be segmented such that the Input TSI buffers are placed at the data source and the Output TSI buffers are part of the processor memory. This eliminates a large amount of unnecessary data movement and serial/parallel data conversion.

Industry...Electronics & Standards

The enormous size and stringent requirements of future high-energy physics data acquisition systems and the amount of laboratory and university physicists, engineers and support personnel needed to totally develop, install and maintain these systems leads system designers to rely more than ever before on both technology and hardware and software standards from industry.

Most of the needs of tomorrow's high-energy physics data acquisition systems from front-end electronics through online storage will be met because of the close match between our needs and the electronics industry's priorities. Examples are as follows:

- Radiation-hardened front-end ICs from the space and defense industries
- Low-cost integrated optics onto both GaAs and Silicon substrates for driving data off the detector from the space, defense and communications industries
- Fiber data transmission support ICs, drivers & receivers and cable from the communications industry
- High-throughput self-routing parallel event builder (switch) from the communications industry (e.g., crossbar switches, optical-GaAs-optical & totally optical switches, etc.) and switching networks superimposed in processor arrays from the computer industry
- Very large online and offline processor arrays from the computer industry
- Advanced standard packaging from the electronics and computer industries
- High-speed and large capacity storage devices from the information, movie and other industries

Because the electronics and computing industries have finally realized that, to greatly help product acceptance, standard packaging is a must for their future products, two emerging packaging and data transmission standards should prove very useful in new high-energy physics data acquisition systems. Futurebus+ has over 100 companies working on new products that should be announced within the next two to three years. Several large semiconductor and computer companies are furiously working to complete the Scalable

Coherent Interface (SCI) standard. This standard can be configured in either a point to point or ring architecture and can transmit data at one GigaByte per second. Software standards are somewhat lagging. The biggest push is in the standardization of Computer Aided Software Engineering (CASE) tools such that these tools quickly meet or exceed the equivalent Computer Aided Engineering (CAE) tools available to hardware engineers.

Summary

The data acquisition architecture presented in this paper has been reviewed by physics laboratory and university personnel throughout the world. These reviews, unfortunately, have only been based on general detector and system requirements discussed at BCD and SSC workshops and at SSC Task Force on Electronics, Triggering and Data Acquisition meetings. Needed in depth data acquisition system requirements will only begin to be fully understood when formal SSC detector collaborations are formed. Only then can detailed system and subsystem requirements discussions between detector system personnel and data acquisition system designers takes place and ensuing system requirements documents be written, reviewed and approved by the proper technical and management personnel.

Acknowledgements

The authors would like to thank Hector Gonzalez, Bill Graves, Ken Treptow, Oscar Trevizo, Gary Moore, Don Walsh and especially all participants of the SSC Task Force on Electronics, Triggering and Data Acquisition for Experiments at the SSC for their valuable input into the development of the data acquisition system architecture described in this paper.