Chapter 12

Event Readout and Control System

The BTeV data acquisition system consists of two parts, both critical to the success of the experiment. The purpose of the Readout System is to transfer detector data to archival storage, interfacing with the various triggering components as needed. The Detector Control System provides data quality monitoring and ensures that all BTeV components operate within design specifications.

This document is structured as follows: a review of the design requirements is followed by detailed discussions of the architecture of the readout system and the data acquisition software. The detector control system is covered in the next section. A short description of the counting and control room infrastructure is given at the end of this document.

12.1 System Overview and Requirements

Event rate and event size are the key parameters in the design of any data acquisition system. For BTeV the event rate will be 2.5 MHz at a crossing interval of 396 ns, or as high as 7.6 MHz for a crossing interval of 132 ns, as the front-end boards transmit data for every bunch crossing. The average event size has been estimated to be less than $\approx$200 KBytes for a crossing interval of 396 ns and an average of 6 interactions per crossing, (or $\approx$75 KBytes at a crossing interval of 132 ns and an average of 2 interactions per crossing). These estimates were obtained using a full Geant based simulation of minimum bias interactions. The uncertainty in background and detector noise is allowed for in these estimates as the actual numbers quoted above are about three times those given by the simulation for an average of 6 interactions per crossing (and 2 interactions per crossing for 132 ns crossing time). Multiplying the event size and event rate leads to an estimate of the throughput required at the first stage of the readout system. Adding the expected protocol overhead increases the bandwidth needed to approximately 0.8 TBytes/s.

The BTeV design luminosity is 2×10^{32} cm^{-2}s^{-1}. At this luminosity we expect on average six minimum bias events per crossing if the Tevatron operates with a bunch spacing of 396 ns. The BTeV event rate will increase from 2.5 MHz to 7.6 MHz should the Tevatron operate with the 132 ns bunch spacing. The overall data rate, however, will remain more or less the same as the average number of interactions per bunch crossing and hence the average event size will decrease by the same factor. As we will see later in this document, larger data blocks at a lower frequency are a benefit for a data acquisition system.

All this data coming out of the detector needs to be stored in a buffer system while the first level trigger processes the event. The average processing time of the BTeV Level 1 trigger is about 700 microseconds, but due to the asynchronous nature of the algorithms significantly longer delays are possible and need to be considered in the design of the buffer system. For BTeV we require a buffer depth of at least 100 ms, corresponding to a total buffer memory size of 50 GBytes.

Events accepted by the first trigger level are forwarded to a large processor farm for further analysis and eventually sent to a mass storage device. The throughput required for the network fabric connecting the buffer system and the processor farm is determined by the fraction of events that pass the first trigger level. For a Level 1 accept rate of 2% an aggregate bandwidth of 12.5 GBytes/s - not including protocol overhead - will be required. (The event size for L1 accepted events are slightly larger than the raw input event size.)

Two additional trigger levels implemented in software reduce the event rate by a factor of 20 yielding a total trigger suppression factor of 1 in 1000. The size of the output stream is further reduced to approximately 200 Mbytes/s by reformatting the event and by replacing some of the raw detector information with processed quantities.

The basic system architecture of the readout system is illustrated in Figure 12.1.

How BTeV has chosen to implement the requirements outlined above will be described in detail in the following sections, which also include discussion on the readout software and the detector control system, as well as the counting and control room infrastructure. Overall, the BTeV Readout and Controls system incorporates the following major components:

1. Data Combiner (DCB): A uniform input receiver/multiplexer for all BTeV front-end boards. The Data Combiner will also distribute control, monitoring and timing information to and from the front-end modules.

2. Optical Links: A high speed, low overhead optical network to transfer data from several thousand front-end sources to buffers (L1B) and the first level trigger (pixel and muon data only) in the counting room. Each link will operate at 2.5 Gbps or higher.

3. L1 Buffer (L1B): Large capacity buffer memory to hold data until a trigger decision is made.

4. Eventbuilder Network: A segmented switching network to combine data from the L1 buffers and to deliver it to the Level 2/3 processor farm for further analysis.

5. Data Storage: Events accepted by the trigger system will be transmitted via optical links to a permanent storage system located in the Feynman Computing Center.

6. Timing System: A "fast" control and timing distribution network for precise system synchronization. The timing signals are synchronous to the accelerator clock.

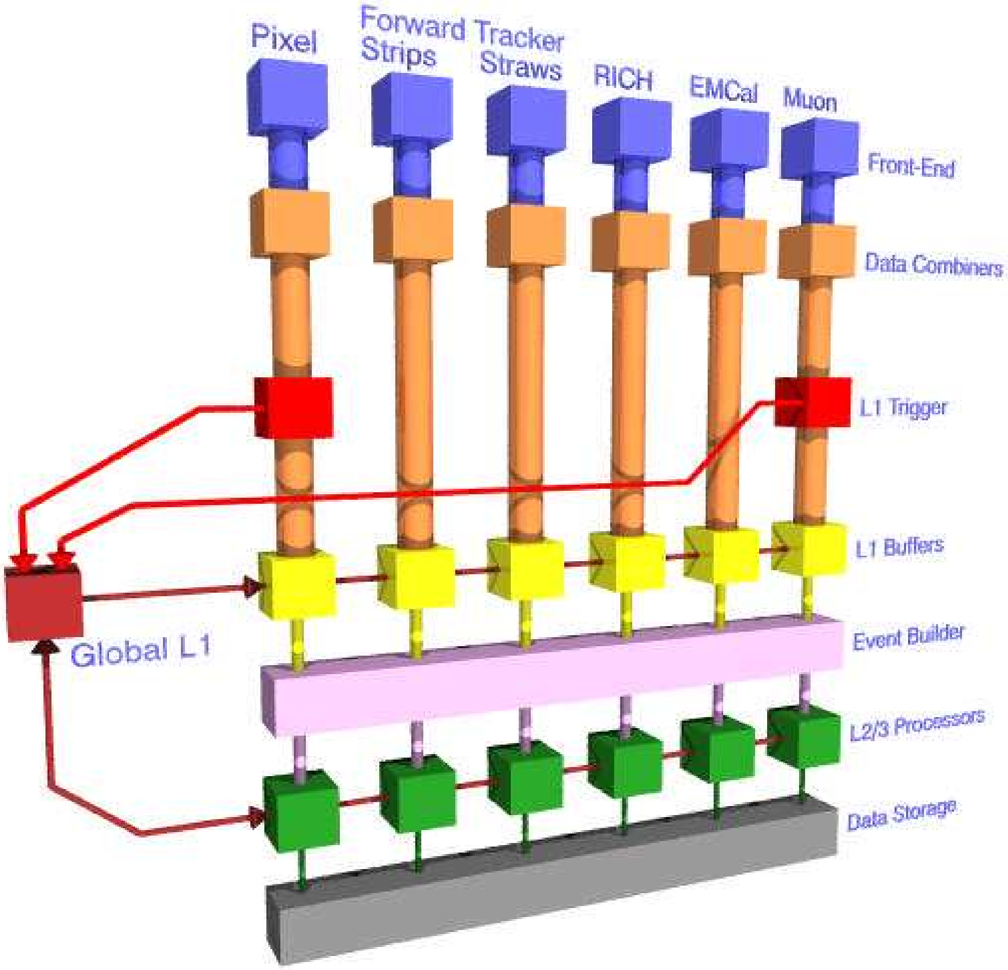

Figure 12.1: Data Acquisition Block Diagram.

7. Configuration and Partioning Subsystem: Software to download, initialize and partition all system components. The partioning subsystem provides the ability to have multiple, concurrent and independent runs with their own user defined trigger requirements and resource list.

8. Run Control Subsystem: Software to control and monitor the operation and overall dataflow of the system.

9. Databases: A system to store and access operating parameters, maintain a time history of all system variables, and store and access parameters necessary for trigger algorithms at all levels.

10. Detector Control (DCS): The Detector Control System includes the software and hardware to set and monitor all system environmental parameters. It incudes an interface

to the Tevatron control system as well as a connection to Fermilab's fire and safety system.

11. Infrastructure: Counting and control room infrastructure, operator and user interfaces.

12.2 Readout and Controls System Requirements

This section describes Readout and Control System requirements that are necessary to achieve the goals of the BTeV experiment. Note that "event" in this document refers to data from a single crossing, regardless of the number of interactions in the crossing. It is assumed throughout this document that data from a single interaction are contained within one event.

12.2.1 Rate Requirements

Most of the data produced by the detector is below predetermined thresholds and is suppressed in the front-end electronics. Approximately 4×10^{12} bits per second are transmitted by the front-end electronics and must be processed by the Readout System. The Readout System may provide compression for data that has not already been compressed at the detector. All data presented to the Readout System Electronics is expected to be in digital form. The combined acceptance of the first level trigger is expected to be 2% of all crossings. The data size of these accepted events will be larger than the size of the incoming raw events, and additional data will be generated by the first level triggers. The Readout System must buffer all data received from the detector for the period of the first level trigger decision and must be capable of delivering a data rate of 10^{11} bits per second to the second level of the trigger system.

The combined acceptance of the second and third level trigger is expected to be 5% of crossings accepted by L1. The Readout System writes data from the third level trigger system to permanent storage. Some of the output data may be summarized, resulting in a reduction in event size of 32%. The Readout System must be capable of delivering a data rate of 2×10^9 bits per second from the third level trigger processors to the data storage system. The Readout System must also be capable of delivering data at a reasonable rate from the data storage system to the processors, allowing use of the processors for offline reconstruction when the detector is not operating or L2/L3 has excess computing resources.

12.2.2 Excess Capacity and Scalability

Readout System bandwidth requirements are based on the sum of estimated data rates for each of the sub-detectors. The Readout System must permit an increase in capacity of at least a factor of two in data throughput at every level (starting with the data combiners since it's unlikely that the links to the front-end boards will scale in number or throughput.) without a redesign of the architecture.

12.2.3 Readout Electronics

The Readout Electronics must respond to Run Control commands and must provide error and status information to the Error Handling/Recovery and Status Monitoring Systems. The Readout Electronics must continue to operate in the presence of faults, such that only data from the failed component is affected. Error detection must be sufficient to automatically identify and isolate failed components. At every stage of the readout chain a synchronization mechanism shall be provided that relates event fragments to crossing number.

The Readout System will provide a standard component (Data Combiner) for receipt of digital data from front-end modules. The Data Combiner will also distribute control, monitoring and timing information, as well as configuration information to the front-end modules.

The Data Combiner must be capable of performing data compression on any uncompressed data received, and must provide sufficient local buffering to smooth data rates on the output data links. It must be remotely resetable and reconfigurable under all conditions not involving hardware failure of the module.

The First Level Buffers receive data from the Data Combiners and various stages of the First Level Trigger. The data is held until a trigger decision is made, and then either discarded or forwarded to the Second Level Trigger. The first level trigger must return a decision for all crossings within a specified maximum latency, even if processing for those crossings is not complete.

The First Level Buffers must accept data that is not in crossing order, but may impose a requirement on sources that all data be grouped by crossing (i.e., data from crossing n may arrive either before OR after data from crossing m, but may not arrive both before AND after.) The First Level Buffers must extract framing information (crossing number and end-of-record) from received data packets for use in identifying and routing the data. On command, the First Level Buffers must stop accepting input data and must allow data already in memory to be output without being overwritten. A First Level Buffer must be capable of generating a null data packet with the proper crossing identifiers when its data source is disabled or malfunctioning.

Data is transmitted from the detector to the counting room on short reach optical links. These links must operate within the specified error rate over a distance of at least 100 meters at 2.5 Gbps or higher. The data link protocol must provide error detection and automatic resynchronization on packet boundaries.

12.2.4 Highway Switch and Event Building

For each event accepted by the first level trigger system, all necessary data from all detector subsystems must be combined and delivered to processors in the second level trigger system. The Highway Switch must be capable of delivering a combined rate of 10^{11} bits per second to the second level trigger system.

The Highway Switch must be capable of routing data from any first level buffer to any second level trigger processor (if multiple readout paths are implemented, first level buffers in one path need not connect to second level processors in a different path, but there must be a way to transfer data at lower speed between second level processors in different paths)

The event builder software must provide buffer space for at least 32 full events, so that L2/3 processors are not idle due to event request latency. Data from a single interaction must be contained in a single event, which is the data from a single crossing. The event building software will reside on the L2/L3 trigger hardware and take up not more than 10% of the CPU resources. The event building software must be able to associate event fragments from a given crossing without error.

12.2.5 Timing and Control

The Timing system generates signals that are synchronous to the accelerator clock. It is assumed that only the Data Combiners and associated front-end electronics will require synchronous timing, and that all other components of the Readout System and Trigger System operate asynchronously. The clock signal will be 7.6 MHz (132 nsec). This is larger than the crossing rate of 2.5 MHz because of technical reasons to do with the structure of the accelerator as described in Sec. 12.3.2. The Timing System must provide a clock synchronized to the accelerator, and must distribute this clock independently to all Data Combiners. The clock source must have no more than 200 psec of jitter (P-P). The Timing system must deliver at least one independent synchronous signal to each Data Combiner for the purpose of aligning commands to specific clock edges.

Any front-end electronics (or Data Combiner) containing a crossing counter must support a command that synchronizes the counter to zero at the next synchronous clock. If the next value of the counter at the time of the synchronous clock is not already zero, a synchronization error must be reported for that front-end or Data Combiner. Each subsystem has a local manager which communicates with Run Control and directs the operation of components in that subsystem. The manager consists of a standard processor and associated software, along with the electronics necessary to distribute synchronous and asynchronous control signals within the subsystem.

12.2.6 Firmware

Components of the Readout Electronics will include embedded software in the form of FPGA firmware and microcontroller code. The embedded software should comply with the standards defined in the BTeV Software Standards document wherever possible. This code will be developed using application specific tools including compilers, debuggers, and diagnostics. All firmware (source and object code) must reside in a software repository that will be used to keep track of different versions of the firmware as it is being developed. The version number of the firmware that is used to process data must be managed in such a way that the firmware version that was used to process data can always be identified. Processes to regularly verify code and run standard datasets must be included. The development software and operating environment necessary to recreate the last implemented version of firmware for each component must be archived. Any unique hardware platforms or keys used in the firmware development process must also be identified and tracked.

12.2.7 Test and Maintainability

Components must include built-in test structures such that all internal functions of the module and the interfaces to upstream and downstream components may be tested with minimal use of external test equipment. Sufficient numbers of spares must be assembled to allow the Readout System to be maintained by module replacement. All programmable components must be "in-circuit" reprogrammable. If there is no permanent data link to the component, the programming interface must be accessible without removing the component from the system.

12.2.8 Readout, Control and Monitor Software

The software required to operate the BTeV detector can be classified in three categories. The first category includes run management and flow control software and support for the partitioning of the readout. Data quality monitoring, configuration, alarms and counting room displays and interfaces are part of the second category. Control system software to monitor voltages, temperatures and similar applications are covered by the last software category.

All software that is designed, or purchased to implement the system control functions must comply with BTEV software standards. Software infrastructure, in particular configuration and downloading of detector constants, shall not introduce more than 5% loss in data taking efficiency.

12.2.9 Run Management

Run Management software is necessary for starting/stopping and organizing all components for data taking. The Run Management software must provide a central facility for system start, stop and automatic error recovery and must provide appropriate monitoring/diagnostic information on DA performance for shift personnel through data taking periods.

The Run Management software must archive run conditions for viewing offline and must provide a central facility to process various component failures and to provide automated mechanisms for recovery where possible. Run management software must provide an interface to change and track changes to run parameters. It must support multiple, independent runs. The Run Control Host must have access (through the control network) to all other subsystem managers/controllers in the system.

12.2.10 Partitioning

During commissioning phases of both the detector and components of the L1 and L2/L3 processing farms, multiple runs will need to happen in parallel using different sets of resources. Some resources, may however, be shared (data switches, the global level 1 trigger, etc.). Partitioning is the ability to provide concurrent, independent runs with their own user defined trigger requirements and user defined resources.

The partitioning mechanism must be able to commission sub-detectors without relying on other sub-detectors to be operational. It must support hetergeonous L2/L3 hardware and OS/software versions. A single partition must be able to support the entire BTeV detector (ie, normal running). Resources must only be reserved for write access by a maximum of one partition. The granularity of a resource should be the smallest unit that does not impact other resources. Not all of a resource needs to be functional for it to be included in a partition. Resources must be capable of being shared across partitions and all affected partitions must be notified when shared resources are modified.

Support of secondary partitions or of parasitic triggers in the same partition cannot adversely affect the throughput of the physics trigger through the primary partition.

12.2.11 Trigger and Detector Managers

The first and second level Trigger Managers are currently viewed as part of the Trigger subsystems. However, they perform the same function as other subsystem managers and may benefit from a common implementation. The Detector Manager provides control/monitor fan-out and fan-in for the Data Combiners associated with a specific subdetector. It also allows standalone local control and monitoring of the subdetector. The Detector Managers may be implemented using the same basic hardware and software for all subdetectors, but may also include detector-specific software. The Detector Manager receives and processes all control messages from the Run Control system and returns status information. It also controls the interface between the general timing system and individual subdetector Data Combiners.

The Detector Manager must allow standalone operation of a complete subdetector. This includes control and monitoring of both Run Control and Slow Control functions and emulation of synchronous signals from the Timing system. It must also be capable of reading (at a significantly reduced rate) any data which would normally be transmitted over the Readout System data links.

The Detector Manager must be capable of locally displaying all subdetector alarms, in addition to passing this information to the Slow Control Host.

12.2.12 Data Storage

Events passing the second and third level triggers will be transmitted via optical links to a permanent storage system located in the Feynman Computing Center. Local storage may also be used to hold data for reprocessing during idle periods of the L2/L3 farm.

The storage system must accept data at an average rate of 2×10^9 bits per second. It must simultaneously supply data at an average rate of 2×10^9 bits per second for offline analysis. The L2/3 processors will have local attached disk drives, these may be used to buffer data during short power or network interruptions at Feynman. The network supplying data to the data storage system and the data storage system itself must have excess bandwidth capacity to offload this accumulated data in a reasonable period of time. An interrupt capacity of 30 seconds in any 1 hour period is sufficient. Data storage must support storing similar events based on trigger type as a collection.

12.2.13 Slow Control

The BTeV Slow Control system is used to monitor and set control/alarms on the detector and in the off-detector electronics (pressures, temperatures, high voltages, etc.). Interface to the main BTeV slow control system must be through a common SCADA package.

The Slow Control system must provide a data path which is independent of the Readout System data path, and/or must remain operational when the Readout System is off-line. Slow control data and alarms must be archived at a rate appropriate to the functions being monitored, such that the state of the system is fully defined for later analysis in the offline code. The Slow Control Host must provide a centralized alarm display for all subsystems.

12.2.14 Control and Data Network

The Control and Data Network provides a general-purpose interconnection for all other subsystems in the BTeV experiment. The Network must provide sufficient bandwidth for efficient database access, download, monitoring, slow control and run control functions. It must support a broadcast capability.

12.2.15 Control Room

The Control Room should be implemented as a remote facility even if located in the detector building. All information must be electronically accessible over the standard network.

12.2.16 Databases

Databases are used throughout the system to provide accesses to configuration parameter and to log status information. There may be several global databases as well as local databases associated with each subsystem. As the architecture and user needs develop, requirements will be established for uptime and reliability, accessibility, performance, scalability, and longevity. Data taking can not be adversely affected by offline process access.

12.2.17 Test Stands

To the extent possible, all BTeV electronics will include built-in self-test features. "Test stands" for these individual components will consist mainly of a small power source and a means of connecting the component to a standard desktop PC. For larger system tests, a test stand which simulates the actual operating environment (full system rack) will be necessary. An attempt will be made to minimize the number of components designed solely for test purposes.

12.2.18 Safety and Security

The Readout and Control System does not pose safety concerns beyond the usual and customary issues associated with high-current low-voltage digital electronics:
If high-current (greater than 10 amps operating or 50 amps rated current), low-voltage (less than 50 volts) supplies power the digital circuitry, the safety requirements for high current power distribution systems must be followed. These are detailed in the Fermilab ESH Manual, Occupational Safety And Health section on Electrical Safety.

A hazard analysis sheet must be completed and signed by any person who will be working with any low-voltage, high-current system, circuit board, or other electronic device. The internal wiring of a commercially manufactured piece of equipment is exempt as detailed in the FESHM section reference above. The reference provides guidance on load connections, ribbon cables, multiple conductors and mechanical components. Safety of people or equipment cannot rely solely on computers or software. The BTeV Readout and Control system must conform to the Fermilab Computer Security Protection Plan. Readout and controls system must operate when cut off from the Fermilab network. Network architecture must allow for rapid isolation from the rest of the Fermilab network.

12.3 Technical Description

For the implementation of the BTeV readout system we have chosen to slightly modify the DAQ architecture outlined in the previous section. When it comes to the actual implementation, this architecture faces several problems:

- Connecting up to 200 buffers to a similar number of edge switches in the L2/L3 trigger farm requires a very large and very expensive network switch.
- At the planned interaction rate, a front-end board will produce an average data packet of less than 10 bytes per crossing. This will result in a large number of very short "messages" - something that is not well handled by commercial networking equipment.
- Level 1 accept and Event Routing messages have to be broadcast to every buffer module - at a rate of almost 100 kHz.

The effects of the message size on the network efficiency can be seen in Figure 12.2. In order to increase networking efficiency and to reduce the complexity of the event-builder fabric, as well as the number of control messages, we have arranged the BTeV DAQ hardware in eight independent "highways". The highway design starts with the Data Combiner modules, which immediately multiplex packets from many front-end boards to form larger packets

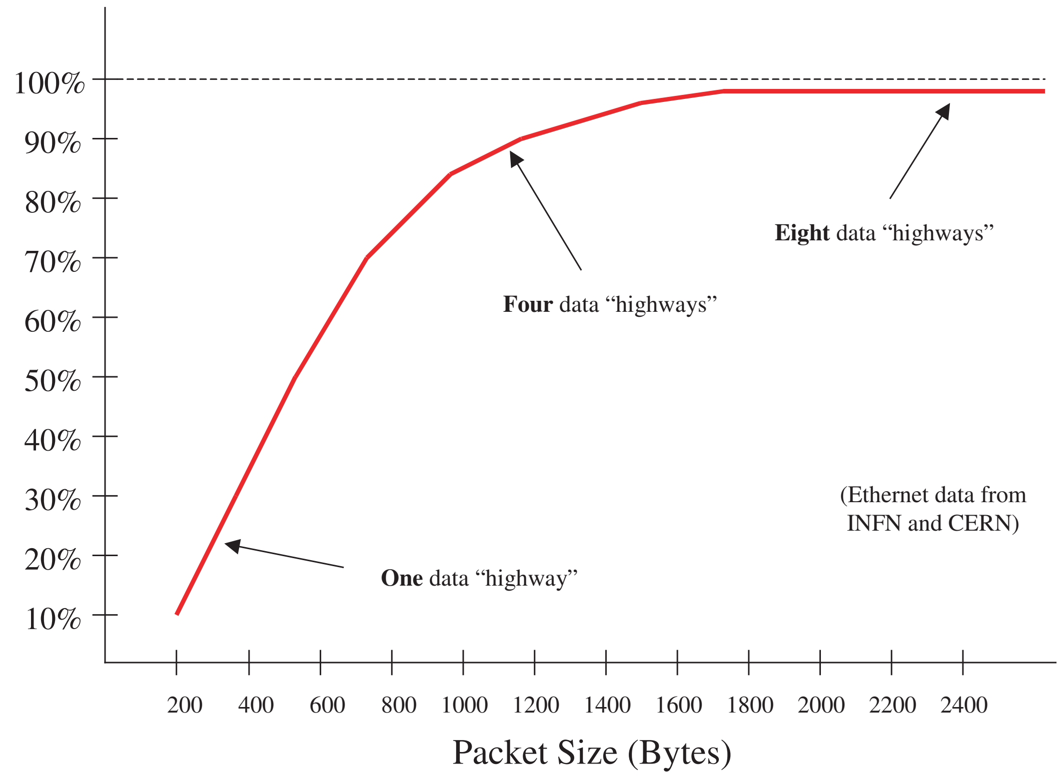

Figure 12.2: Typical gigabit Ethernet Efficiency.

(200-300 bytes). These larger packets are then distributed uniformly to one of 8 output links, each connected to one of the 8 highways. From the viewpoint of a single data acquisition highway, the crossing time appears to be 3 microsecond (8×396nsec), with a corresponding $8\times$ decrease in the packet processing overhead and index table size.

Dividing the system into highways provides the same advantages for data management in the Level 1 Trigger processors and may support a rudimentary level of partitioning.

The DCBs are configured to send all data from one bunch crossing to a single highway. Within each highway, the data is either processed by the first level trigger system and sent to Level 1 Buffers or it is sent directly to the buffers. The decision of the first level trigger is then transmitted to the L1 Buffers, which forward the data to the second level trigger processors. Since the Level 1 Buffers receive "accept" decisions only for events on their particular highway we have reduced the control message traffic by a factor of 8.

We extend the highway model to the trigger farm and assign one eighth of the farm nodes to each highway. This approach allows us to replace the large event-builder switch with eight smaller switches - one for each highway. The highways will be interconnected via a ninth Gigabit Ethernet switch. This way it will still be possible, for calibration and test purposes, to route data from any particular bunch crossing to any particular Level 2/3 farm node - just not with the full DAQ bandwidth. However, we consider this a small price to pay

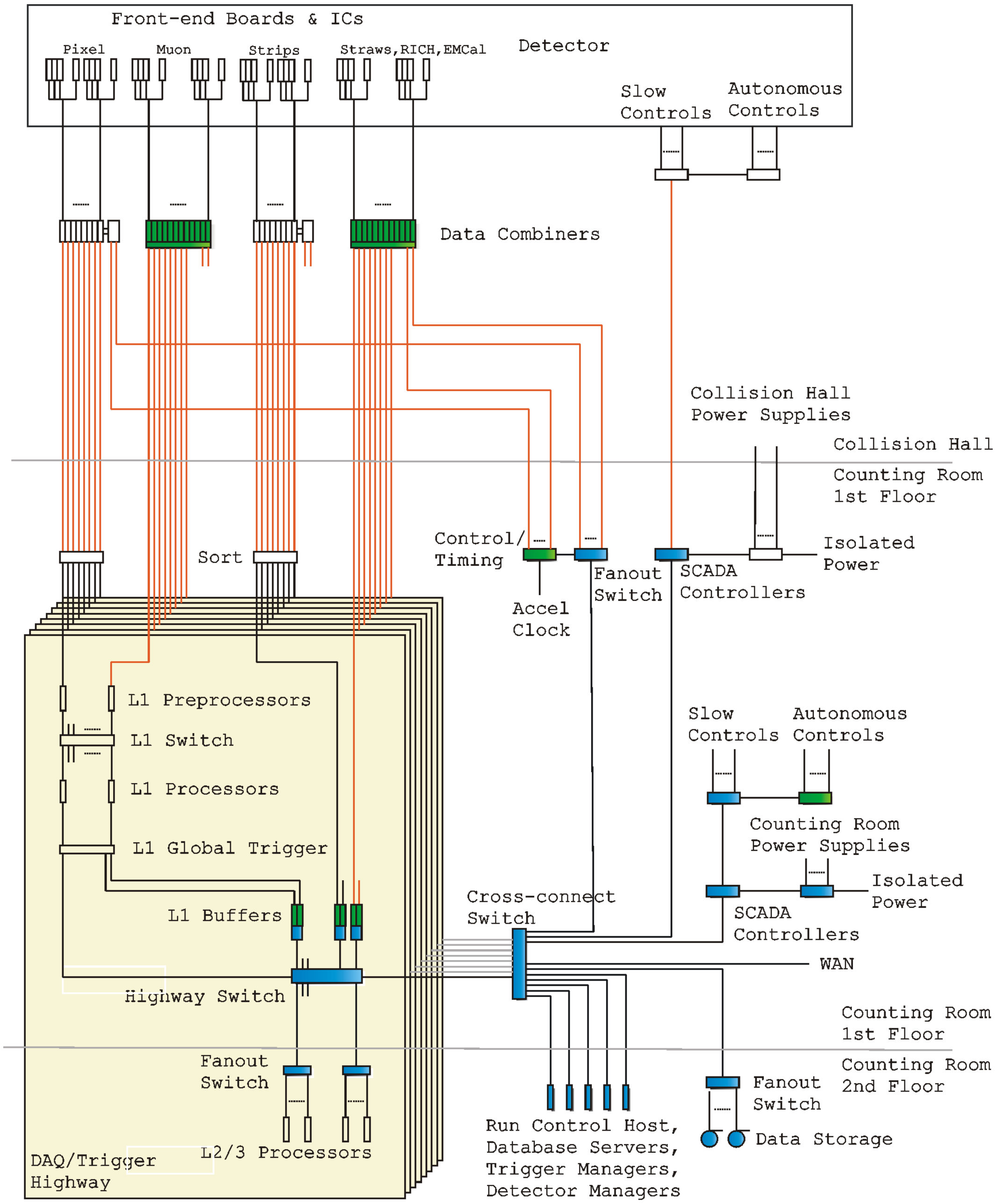

Figure 12.3: Block Diagram of the BTeV Readout System showing 8 parallel Highways (represented by the colored stacked rectangles).

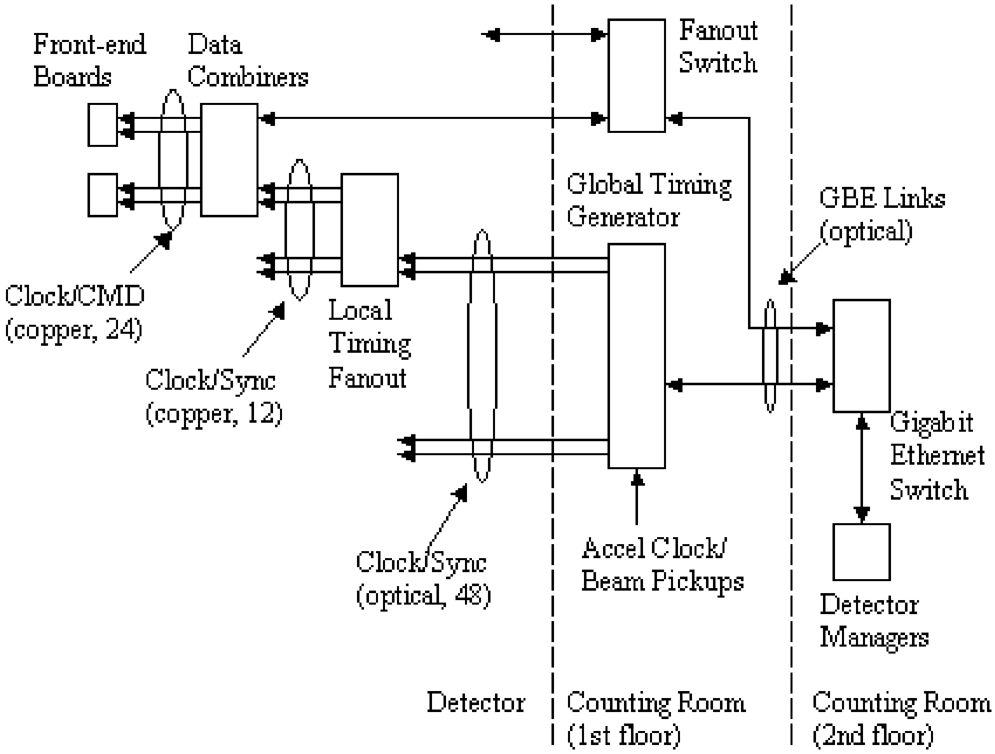

Figure 12.4: Timing and Control Distribution.

for the advantages offered by the highway approach. A detailed view of the BTeV readout system including timing, detector control and monitoring can be found in Figure 12.3.

The design of the Readout and Control system provides additional margin for inefficiencies in data balancing and link utilization, noise in the detectors, and limited expansion. The peak design rate of the system is 800 GBytes/s, of which approximately 500 GBytes/s is usable bandwidth during steady-state operation.

12.3.1 Timing and Control

The Global Timing Generator (GTG) provides signals synchronized to the accelerator clock. If the accelerator is off-line, a local oscillator can be selected. The GTG also maintains the global crossing counter which is used (along with the sync signal) to align and check downstream crossing counters in the Data Combiner (DCB) and front-end modules. Each GTG fanout module drives clock and sync signals to 12 DCB modules. A simplified block diagram of the timing system is shown in Figure 12.4.

Control messages may come from either the Detector Manager (detector specific) or Run

Control (global). Messages are asynchronous, containing both the command and crossing number, and need not be time ordered.

Each DCB synchronizes messages to the selected crossing clock. Messages that must be transmitted on a regular basis are stored in a table in the DCB. Phase adjustments can be made at the DCB clock input and outputs to compensate for variations in GTG to DCB and DCB to front-end cable lengths.

12.3.2 Front-End Interface and Data Combiner

We assume a model where all digitization (and data reduction where appropriate) is performed on the front-end modules. Data is then transmitted on serial links to a Data Combiner Board. Control and timing information is transmitted from the Data Combiner Boards to the front-end modules. These signals are all implemented using differential copper links, preferably within the same physical cable.

Consideration was given to encoding timing and control, so that only one transmit and one receive link would be required for each front-end module. Unfortunately, the serial to parallel latency of the deserializer varies between manufacturers and also varies for selected components following each power cycle. This results in a possible system clock skew between front-end modules, which cannot be removed by tuning. To avoid this problem, we plan to distribute the system clock as a separate, unencoded signal synchronized to the accelerator. The clock is 3× the crossing rate to account for the non-integral length of the abort gaps with respect to the crossing rate.

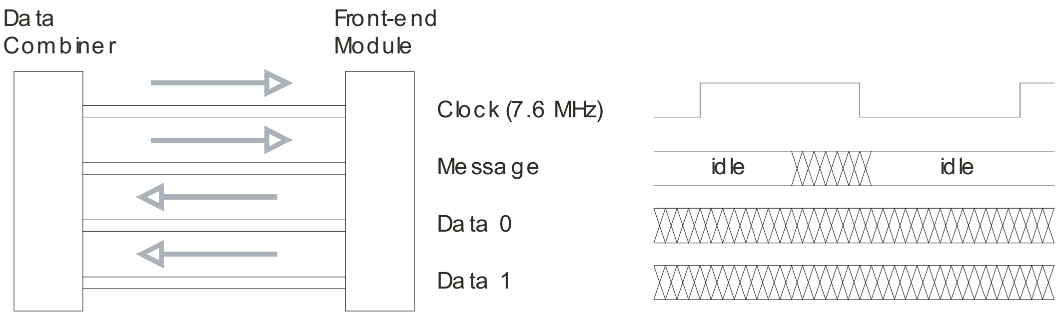

Figure 12.5: Front-End Interface.

In addition to the system clock, there will be a single serial data link providing control information to the front-end module, and one or more serial links for data generated by the front-end module. The control link will be framed by the system clock, such that any control word received by the front-end module will have a guaranteed setup and hold time with respect to a specific rising edge of the clock.

Front-end modules should consider driving the serial data link reference and output framing clocks from a local oscillator as the accelerator synchronized system clock is not always sufficiently stable (especially during beam ramp).

As a baseline, we are assuming that each front-end module data link will operate at a rate of ≈600 Mbps. This limit is influenced by the current cost of high-speed data cables and connectors. There are several announced serial standards (PCI Express, Serial RapidIO, Serial ATA II) operating at link speeds of 2 Gbps or higher, and if one of these becomes economically feasible prior to implementation, we will likely move to that standard.

The least expensive cabling for the front-end to Data Combiner connection is standard CAT-6. This is shown to operate reliably at the 600 Mbps rate over distances of at least 5 meters. We will also examine the reliability of standard RJ45 connectors, which would further reduce cable costs if acceptable. If space is limited, higher density cables with up to 48 signal pairs (6 front-end ports per cable) may be used.

Each cable/port contains 4 differential pairs. Two of these are used for the system clock and serial control link to the front-end module, leaving two pairs available for serial data links from the front-end to the Data Combiner (Figure 12.5).

If it is possible for the output bandwidth of a front-end module to exceed the capacity of a single connection (2 × 600 Mbps), a second or third port may be added. From the viewpoint of the Data Combiner, each port will be considered a separate logical front-end module. Clock and control signals are duplicated in each port.

12.3.3 Data Combiner

The Readout and Controls subproject is supplying ≈250 Data Combiner (DCB) modules for use in the RICH, Straw, EMCal and Muon readout paths. DCB modules are packaged in groups of 12 to form 20 Data Combiner Subsystems (Figure 12.6). This grouping is designed to match the output channel count of the DCBs (8 channels each) with the channel count of the optical links (12 channels each). A DCB Subsystem backplane implements the 12 × 8 to 8 × 12 "shuffle" network. Additional Data Combiners (with application specific features) are used in the Pixel and Strip readout paths, and are supplied by those subprojects.

A DCB multiplexes data from up to 48 input serial links (24 ports) with a total bandwidth of 30 Gbps. With variations in occupancy, the average front-end link utilization is expected to be less than 50%, so the DCB output links provide a bandwidth of 20 Gbps (one 2.5 Gbps serial connection to each of the 8 system highways).

Crossing data is distributed to highways in a "round-robin" sequence. The effective crossing rate for each highway is therefore one eighth of the detector crossing rate, or about 320 KHz. The DCB will include an option to enable or disable specific highways and to allow skipping of highways in a uniform pattern (e.g., 01234567, 12345670, 23456701,...70123456). This feature will be used only if there is an apparent resonance between the accelerator and the default distribution of crossings to highways.

For each crossing, the data from all inputs are concatenated to form a single packet with an average size of a few hundred bytes (depending on sub-detector). This packet is transmitted to a Level 1 Buffer or Level 1 Trigger.

The data concatenation is performed by programmable logic in the DCB, so it should be possible to do some additional data reduction at this stage, for example removing the

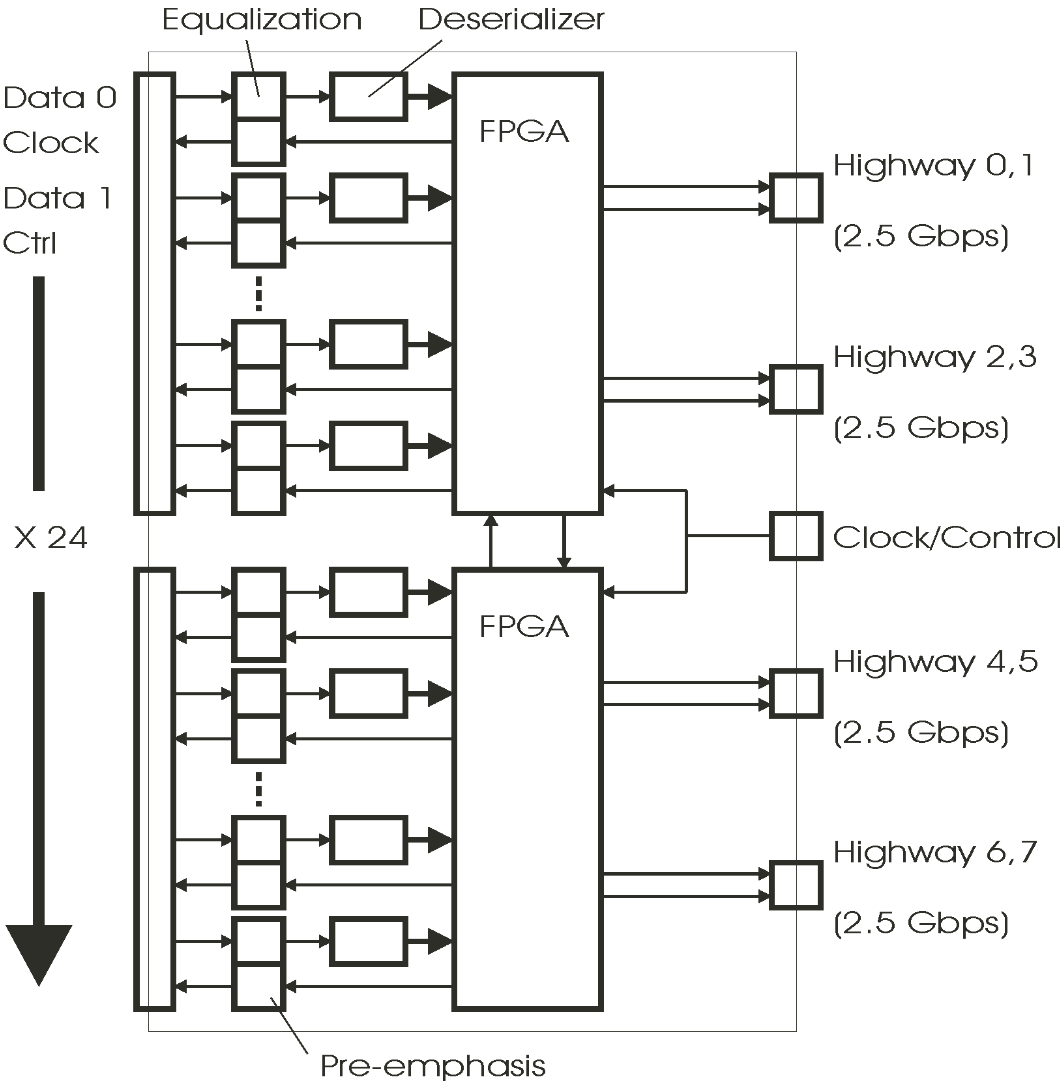

Figure 12.6: Data Combiner Module and Subsystem.

individual crossing timestamps in each input packet and inserting a single timestamp in the output packet. Internal packet formats will vary somewhat between sub-detectors and it may not be feasible to implement this additional data reduction in all DCBs.

The DCB logic includes a "snapshot" function to capture a specific crossing and send that data through the network connection to a Detector Manager. This provides an alternate low speed path for commissioning of sub-detectors prior to installation of L1 Buffers and data highways. It also allows cross-checking of data sent through the main data acquisition path.

The DCBs are located near the detector, but should not be placed in areas where radiation levels are expected to exceed 0.5 KRad/year (the limit for most commercial integrated circuits). A possible implementation is shown in Figure 12.7.

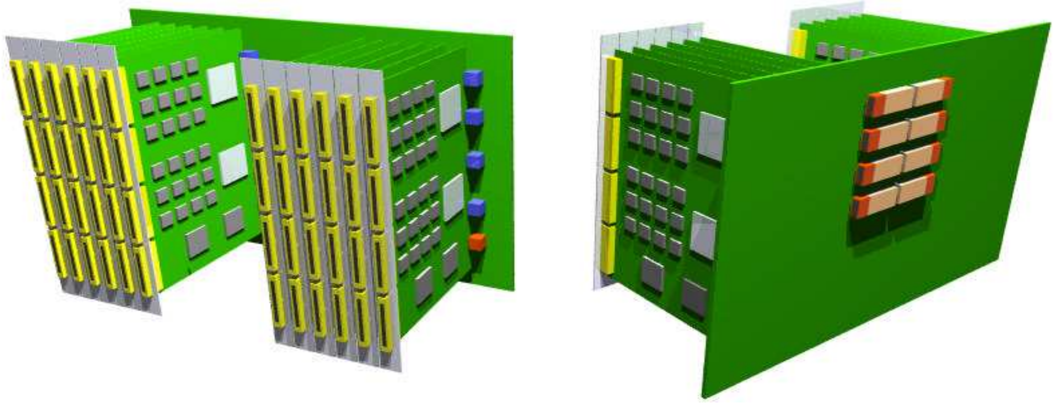

Figure 12.7: Data Combiner Module Implementation.

12.3.4 Optical Links

The serial outputs of the Data Combiner are 2.5 Gbps copper links. The maximum distance from the Data Combiners (located near the detector) to the L1 Trigger System and the L1 Buffers (located in the Counting Room) is 70 meters. This exceeds the distance considered acceptable for reliable high speed data transmission over copper cables. Conversion to optical links also provides the benefit of electrical isolation between the Collision Hall and the Counting Room.

Figure 12.8: Parallel Optical Transmitter and Fiber.

The most cost-effective links for this application are based on unidirectional parallel optical transmitters and receivers (Figure 12.8). A total of ≈250 of these 12 channel optical links provide a maximum data capacity of ≈ 800 GByte/sec. At the end of the each link is a 12 channel optical receiver which plugs directly into the L1 Buffer module or L1 Trigger interface.

There are 28 conduits connecting the Collision Hall to the Counting Room. With 6-8 inner-ducts per conduit, each inner-duct will contain only 2 parallel fiber cables for easy installation or removal.

12.3.5 L1 Buffers

The L1 Buffer (Figure 12.9) receives partially multiplexed event data from a number of sources, mainly Data Combiners and L1 Trigger Processors. Up to 48 sources are attached to each L1 Buffer subsystem. Each source is independent and no assumptions are made about crossing order for events arriving within or across channels, other than the requirement that all data associated with a specific crossing on a specific channel be grouped together.

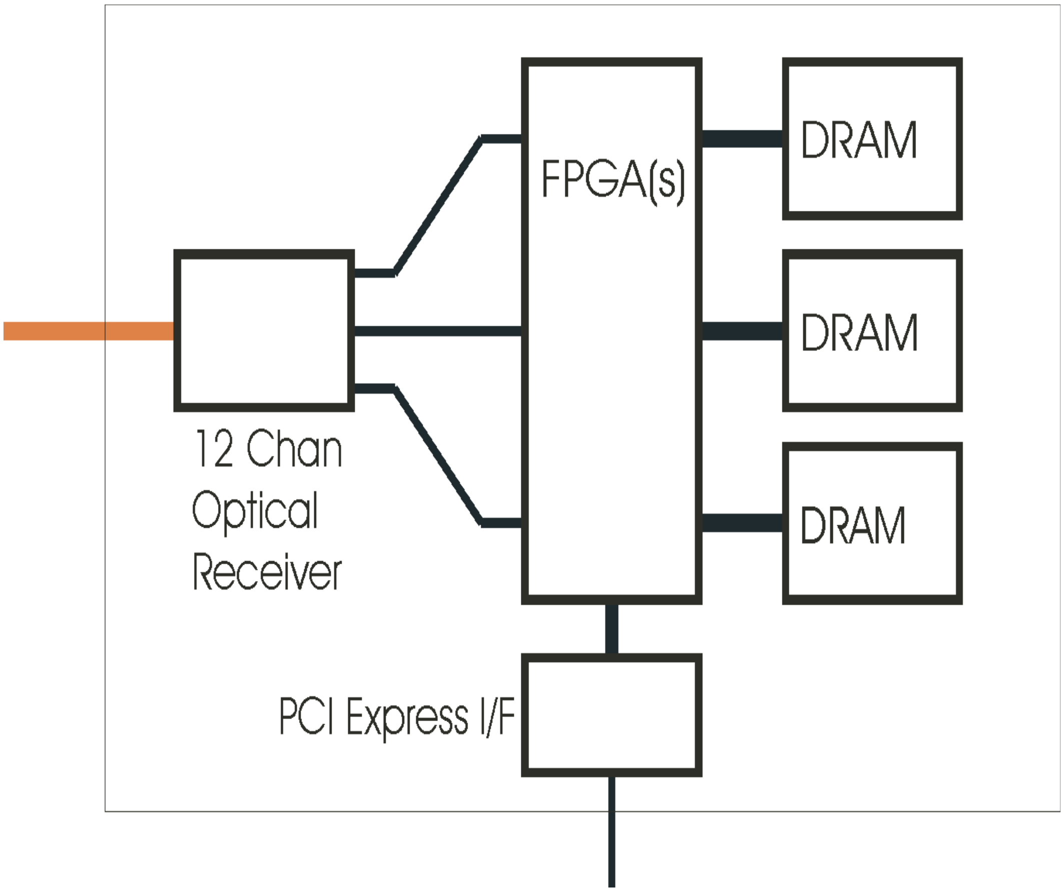

Figure 12.9: Level 1 Buffer Module and Subsystem.

The data stream in each channel is examined to locate crossing boundaries. Data is written to a circular buffer and a pointer for the associated crossing number is written to a lookup table. The circular buffer is implemented in DRAM, with a capacity of approximately 100-200 thousand crossings (depending on link occupancy). This corresponds to approximately 500 milliseconds of available L1 Trigger decision time, long in comparison to most existing first level trigger systems. L1 processing will simply timeout and automatically accept the data if the processing time approaches this limit. The buffer size can also be expanded at minimal cost if trigger simulations warrant.

When the L1 Buffer Controller receives an L1 accept message, it concatenates data from each of the 24 input buffers and copies that data to the output buffer. The output buffer has a capacity of approximately 100,000 accepted events (>20 seconds of continuous data), although individual events can be held indefinitely pending a L2/3 processor request.

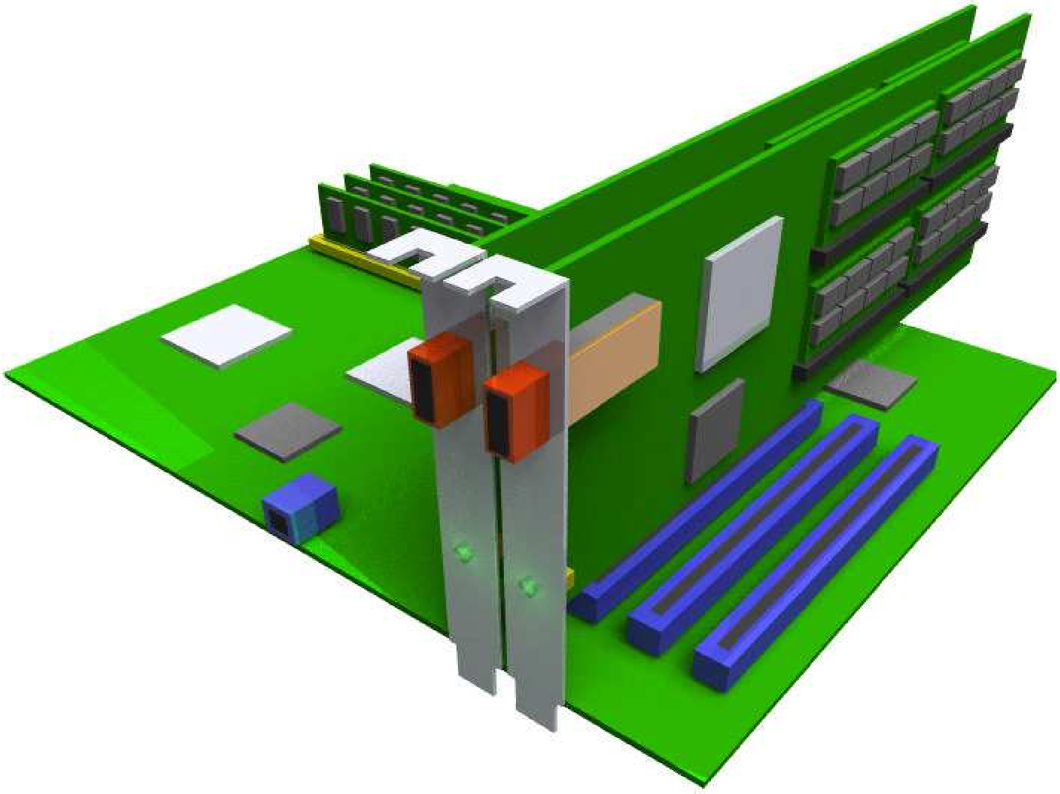

Figure 12.10: Possible Implementation using a PC-style Motherboard.

The serial link receivers, input circular buffers and multiplexing logic are implemented in the L1 Buffer module. There are a total of 320 L1 Buffer modules in 80 L1 Buffer subsystems. The Readout and Controls subproject is supplying all L1 Buffers for the system, including those used in the Pixel and Strip datapaths. Both the Data Combiner and the L1 Buffer Subsystems will include a CPU for monitoring and configuration. A possible implementation is shown in Figure 12.10.

12.3.6 Network

Data from the L1 Buffers in each highway is transferred to L2/3 Processors. A single L2/3 Processor receives all of the data for a particular crossing (32 separate blocks), and the final stage of event building is done in the L2/3 processor.

A network of Gigabit and Fast Ethernet switches connect the L1 Buffers and the L2/3 Processor Farm. The primary switch in each highway provides 64 Gigabit Ethernet ports with the following connections:

Connection	**Number of Ports**
L1 Buffers	24
L2/3 Fanouts	24
Global L1	1
Cross-connects	2
(reserved)	5

Table 12.1: Network Port Allocation

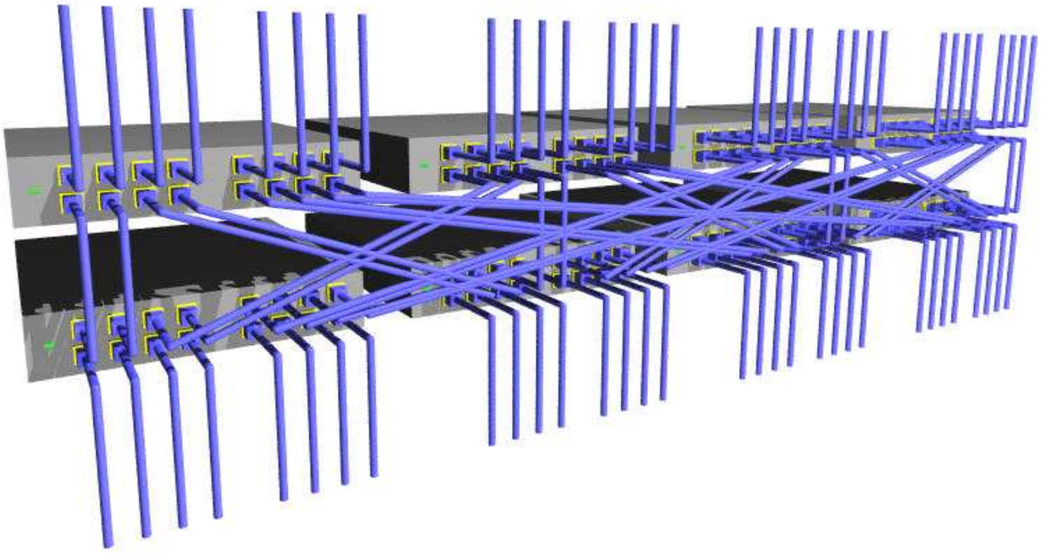

Figure 12.11: Possible Implementation for Highway Network Switch.

The 24 L2/3 Fanout ports connect to a second group of switches located in individual L2/3 Processor racks. Each of these fanout switches supports 2 Gigabit Ethernet connections and 24 Fast Ethernet connections. With this configuration, 288 L2/3 Processors may be attached to each highway. The number of processors is easily expanded by adding fanout switches.

The primary highway switch can be implemented using a single 64 port switch or a number of smaller switches (e.g., sixteen 8-port switches or eight 16-port switches, arranged in two stages as shown in Figure 12.11). Because the dataflow is predominantly in one direction on each switch port, switches with "oversubscribed" (partially blocking) ports may be satisfactory.

With sufficient internal buffering, external traffic shaping is generally not required. If the internal buffering is not sufficient, simple traffic shaping using fixed packet sizes, rotation, and starting offset (barrel shift algorithm) can be used to avoid blocking.

The eight highway switches are cross-connected through a ninth switch. This switch allows communication between highways at lower speeds (e.g., to read data from sequential crossings for calibration purposes). It is also the central network interconnection point for the Detector Managers, Run Control processor, Slow Control processors and Database Server. Finally, this switch connects to the data storage system, allowing L2/3 processors in any highway to write accepted events to any storage device. Switch assignments are:

Host	
Run Control	1
Detector Managers	6
Trigger Managers	3
Database Server	2
DCBs	12
Data Storage	4
Slow Control Processors	8
Cross-connects	16
External	1
L2/L3 Manager-I/O Host	4
(reserved)	7

12.3.7 Event Identification, Event Building and Event Distribution

The basic crossing identification is derived from the 7.6 MHz clock. There are 159 clocks per accelerator revolution, with 36 crossings. The low order 8 bits of the crossing number identify the clock "tick" within a "turn". Higher order bits identify the turn number within a run segment. All crossing identifiers are cross-referenced to the date and time.

Each level of the readout system needs to track crossing numbers only at the resolution required to uniquely identify the data in that level. For front-end modules and Data Combiners, this is typically 8 bits (1 turn) since the buffer depth in these components is limited. For the global L1 trigger processor and input stage of the L1 Buffers, the resolution must be at least 24 bits (2 million crossings), and for events that pass the L1 Trigger, a resolution matching the maximum length of a run segment (>32 bits) is necessary.

12.3.7.1 Event Distribution

The basic software interface between the Global Level 1 trigger, the Level 2/3 trigger farm and the data acquisition system is shown in Figure 12.12. The same switching network is used for both data and control, with control messages taking a small fraction ($< 5\%$) of

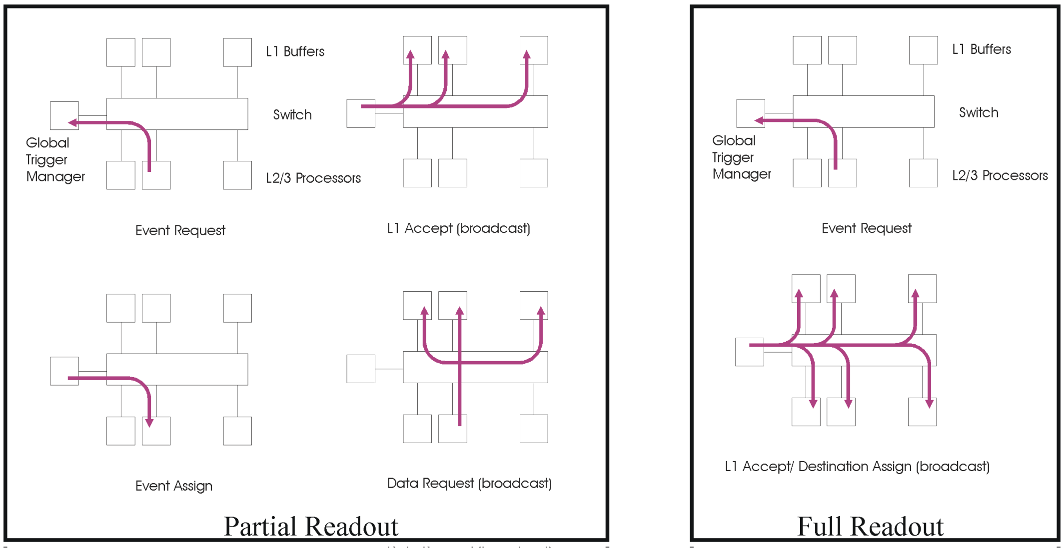

Figure 12.12: Basic Readout Control Messages.

the total bandwidth. Messages are asynchronous and buffer sizes are such that there are no significant real-time requirements placed on the delivery of these messages.

L2/3 processors make event requests to the global trigger system controller. These requests may be generic or they may specify certain trigger types. An event request message may ask for more than one event.

When an event passes the Level 1 trigger, a level 1 accept message is broadcast to all L1 Buffers telling them to save that event. The L1 Trigger and Global Trigger controller have approximately 400 ms to make this decision and transmit the message. Once an event is saved, there is no time limit on assigning that event to a processor or requesting the data.

The global trigger controller maintains a list of event requests and accepted events. It may assign events to L2/3 processors in the order that requests are received, or it may try to balance network traffic by distributing events uniformly across the L2/3 network ports. The assignment message is sent to the L2/3 processor, and there may be more than one event assignment in each message.

The L2/3 processor then requests data from the L1 Buffers. The baseline design assumes that the data request is a simple broadcast to all L1 Buffers in this particular highway, but the L2/3 processor has the option of requesting data from a subset of buffers, analyzing that data, and then making additional requests in any order it chooses. The L2/3 processor must eventually make a request (to send or delete data) to all L1 Buffers in the highway.

Alternatively, the event assignment/routing information can already be included in the L1 accept broadcast to the buffers. Upon receipt of this message the buffers push the data to the selected L2/L3 node.

If the number of saved events in an L1 Buffer approaches the capacity of the buffer (>95%), a warning message is sent to the Global Trigger controller. The Global Trigger controller must then stop issuing L1 accepts until it receives a message indicating that the buffer is again ready ($< 90\%$). Each 5% change in buffer utilization represents approximately 1 second of continuous L1 accepts in the absence of any L2/3 activity, so again there is no significant real-time requirement on these messages.

12.3.7.2 Event-Building

All the data from a single bunch crossing is called an event. An event can include multiple hadronic interactions. The process of combining the data fragments from individual front-end modules into one record is called event-building. A BTeV event will be built in three steps. First, each DCB combines data from 24 front-end modules into a larger event fragment of typically 200 - 300 bytes. These are sent to an L1 Buffer where data from up to 24 DCBs are merged. At this stage an event is split into 24 fragments of typically 3-8 KBytes. These are sent via the switching network to a node of the L2/L3 trigger farm where the final event-building step will be done in software. We have performed benchmark tests and found that only a few per cent CPU time is used by this last event building step.

12.3.8 Data Logging

Events accepted by the Level 3 trigger are sent to data logging processors and temporarily stored on disk local to the Level 3 node itself before being transferred to persistent storage in the Feynman Computer Center. In the event of a loss of services to Feynman, these local disks will buffer data until the services are once again available. Level 3 nodes will provide enough disk storage to buffer the accepted events for several days. These events would be immediately accessible for continued processing. A centralized hard disk storage system attached to the cross-connect switch at C0 can provide the same capability with only a small increase in overhead.

In the baseline system event files on local disks are then forwarded to a tape robot in the Feynman Computer Center. Data that has been written to local mass storage may later be transferred back to the L2/3 farm for further analysis during periods of low processor utilization. During longer interruptions, data that has been written to long-term storage (e.g., the tape system at Feynman) can be accessed and spooled to local disk storage for further processing and analysis.

The data logger sub-system is attached to the L2/L3 farm using the individual highway switches and the cross-connect switch. Given the high trigger rejection and the decrease in average event size, this utilizes approximately 1-2% of the data switch bandwidth while avoiding the need for a separate data-logging network.

We have modeled our data logging system based on the experience gained by the Run II upgrades for the current Tevatron experiments. The BTeV system will use Enstore, the mass storage system infrastructure supported at Fermilab. A current canonical Enstore system consists of a few (5) PCs to provide services, plus an additional “mover” PC for each tape drive. All PCs are commodity boxes running Linux. By the time tape drives are purchased (2007), it is anticipated that 40Mb/sec drives will be readily available, at a sustained data rate of 200Mb/sec, a minimum of 6 drives should be sufficient.

However we have still some time before we need to decide on which mass storage technology to use. It should be noted that the current economic trend favors optical disk media (such as DVD-RW) over traditional magnetic tape by a factor of approximately 3X. It still needs to be evaluated if space, power, and reliability requirements can be met with these alternate media choices.

12.3.9 Common Electronics Features

All components in the readout system are designed to operate at a single supply voltage, which is either 48 volts DC (unregulated) or 120 volts AC, depending on the component.

All connections between components are point-to-point serial links.

To the extent possible, all components will include built-in self-test (BIST) features to allow standalone and in-situ testing. Links will include pseudo-random bit sequence (PRBS) generators and checkers, so bit-error rates can be determined for all links in parallel. Modules will include data realistic pattern generators at all inputs to test internal functionality.

Modules attached to the network (DCBs, L1 Buffers) are identifiable by the network MAC number, which is also printed on each module. Other components are identified by printed label, and in the case of link cables, by labels at both ends regardless of cable length.

12.3.10 Software Infrastructure

On a larger scale, the data acquisition software has the same requirements of most HEP experiments, namely,

- **Readout:** Moving the data from the front-end boards to the archival system, passing through the various trigger levels along the way .
- **Run control:** Managing a period of data taking. This includes configuring and initializing hardware and software systems as necessary, starting and stopping the acquisition period, monitoring the overall data flow, and archiving run information that will be needed for offline analysis.

Because of the complexity and large number of electronics modules of the detector and trigger systems, various components will need to be tested in parallel in order to bring up the machinery efficiently. It is therefore essential that the DA software supports independent, possible concurrent, readout streams over partial configurations.

Run control itself can be divided into several components. The core services on which all other components are based are:

User Interface software is the common graphical user interface and libraries for all readout system packages. It is designed to present a common appearance across applications and platforms. Exceptions may include user interfaces for the Slow Control system and Network Monitoring software that are part of integrated commercial packages.

Process Management is the software needed to start the online data acquisition software and verify that it remains active during a run. The system may be restarted from various operating states, ranging from "cold" start to several levels of system reset. The Process Management software determines what other processes need to be loaded, initialized, or reset, and ensures that they are properly synchronized.

The initial release of the Process Management software will operate in a single node environment (i.e., individual Detector Managers) with a basic command line interface. Subsequent releases will add support for multiple nodes, graphical user interface, and connection to the central online databases. The final release will add capabilities to restart individual failed components.

The Message Passing System is the software that interconnects all other processes, either locally or across the network, using a common message format. The initial release will operate with a single server to route all messages. Subsequent releases will expand to support a multi-tier architecture to handle a large, distributed system.

Electronics Support software is needed to configure embedded processors in various readout system components, such as Data Combiners and L1 Buffers, which require operating systems (real-time LINUX) and network interfaces. Routing tables in the network switches must be configured for each highway, and for the system cross-connects.

Error Handler software is needed to log and present component and system errors to operations. Logs will also be used for diagnostics and triage as problems occur. Additionally, certain errors or series of errors may generate automated responses and possible recovery. This is essential as the sheer number of components comprising the detector will insure that some sort of failure will happen quite frequently.

12.3.11 Readout Software

Run Control is the high-level process responsible for initializing, starting and stopping the readout sequence. Run Control does not directly control data flow on an event-by-event basis.

Data Acquisition Monitoring software is used to monitor and display information about data flow in the system, including data rates, buffer utilization and overall load balancing. The initial release will provide a text based interface and run on a single Detector Manager. Subsequent releases will cover data highways and the full data acquisition system, and will include a graphical interface and interface to the Run History database.

Configuration Management software is responsible for the selection, verification, and download of readout system constants and operating parameters. It is closely related to the global system Process Management software. The initial release will run on a single Detector Manager, with subsequent releases adding multiple node coverage, a graphical interface and interface to the Run History database.

Partitioning software provides the virtual segmentation of readout system components into one or more quasi-independent paths. Ideally, single partitions may cross data highways and multiple partitions may exist in the same highway. In reality, the segmentation options may be more limited due to sharing of components and bandwidth limitations between highways.

The Run Control host and all Detector Managers communicate with the readout system through a common network switch. This means that all partitions running on these machines have access to any component in the system, regardless of data highway. Commands may be sent to selected subsets of the readout system at the level of individual DCBs and L1 Buffers.

Although the Level 1 Trigger is not partitioned, the global trigger manager within a highway can select events by type for assignment to L2/3 processors in specific partitions.

Control Room Logbook software will be accessible from all Run Control and Detector Manager user interfaces. The logbook will be a standard software package used in previous systems.

A software Event Builder resides on each Level 2/3 processor and performs the final stage of event building, combining Ethernet packets from the 32 L1 Buffers into a single event. Some consideration was given to implementing this operation in hardware or in a separate sub-farm manager, but initial tests have shown that the required overhead in the L2/3 processors is not significant.

A similar software event building function is included in each Detector Manager, for use in assembling test data directly from DCBs within a sub-detector at low rate.

The Data Quality Monitor is a set of routines to histogram, view and archive data for comparison of results in different trigger conditions and system configurations,

Data Logging software controls the transfer of accepted events from the L2/3 farm to mass storage. Approximately 200 MBytes/sec of data passing all levels of the online trigger system will be recorded for offline analysis.

12.3.12 Detector Support

12.3.12.1 Detector Managers, Control Supervisors

There are several control processors in the system designated as Detector Managers and Control Supervisors, one of each for every major sub-detector, although we will consider merging the two if sufficient performance is available in a single CPU box. These computers perform many of the same functions as the global Run Control processor, but are meant to allow parallel independent operation of the sub-detectors. This is a logical distinction only, since all Detector Managers are connected through the main network switch.

A Control Supervisor is responsible for controlling and monitoring the slow control systems of a detector component, and a Detector Manager is used for initialization of any readout electronics attached to that sub-detector. A Detector Manager can send commands to sub-detector DCBs and front-end modules, and can read raw data at low speed directly from the DCBs or processed events from the trigger farm.

12.3.13 Detector Control System (DCS)

In an experiment the size of BTeV, several hundred devices need to be controlled (e.g., high voltage systems and calibration pulsers). Several thousand parameters need to be monitored at regular intervals (e.g., power supply voltages, temperatures, gas mixtures). These tasks will be performed by the BTeV Detector Control System (Figure 12.13). Additional

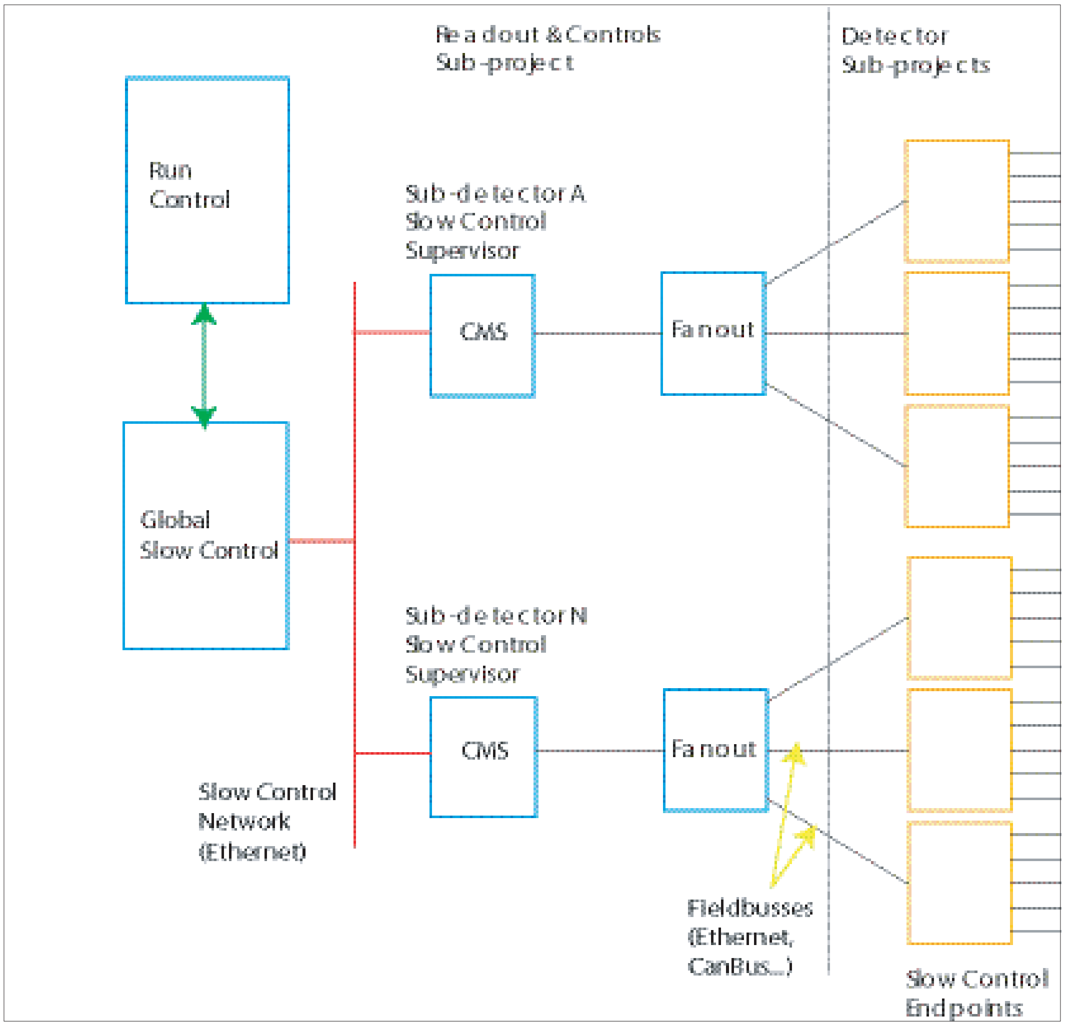

Figure 12.13: Block Architecture of the Detector Control System.

functionality of the DCS includes user interfaces/consoles and archiving of environmental parameters and detector conditions. Safety critical functions are explicitly not included in the Detector Control System (with the exception of monitoring/archiving operations). The BTeV detector control system will be based on a commercial software package implementing the SCADA standard. We have begun evaluation of two such packages: IFIX by Intellution, which is currently used by the CDF experiment and PVSS-II by ETM which is the software of choice for all the CERN LHC experiments.

In order to decrease development costs and to improve maintainability we will use commercial controllers, endpoints and software throughout the entire slow control system wherever possible. The Readout and Controls subproject is responsible for the slow control host computers (the Control Supervisor, one for each major sub-detector) and the slow control network (Ethernet). The endpoint hardware and any associated local controllers are detector specific in most cases and are the responsibility of the individual sub-detectors.

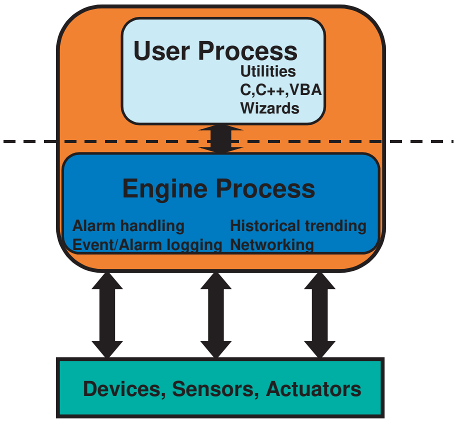

Figure 12.14: Block Diagram of a Control Supervisor.

A slow controls development system will be commissioned at an early stage to provide higher level control and interface software, and to make recommendations for general-purpose digital and analog endpoint hardware.

The Global Detector Control will have a hierarchical structure taking advantage of the modularity available with (some) SCADA systems. For each sub-detector a complete SCADA system will be implemented on the Control Supervisor PC. This system is self contained and sufficient to monitor and to operate the sub-detector slow control system. A schematic diagram of a Control Supervisor is shown in Figure 12.14. As shown in Figure 12.15, the sub-detector control system will be connected to the Global Control System, another SCADA based system, which serves the user consoles in the Control Room and manages the central archive and error logging/alarm systems.

12.3.14 Databases

The BTeV system accesses a number of databases, with varying mass storage and real-time requirements. Solutions based on both commercial and freeware database servers will be considered. Standard APIs will be developed for use by other components of the readout and controls system, trigger system, and individual clients. For the commercial option, an intermediate database access interface may exist between the applications and the main database.

Database applications will be written for run history, luminosity monitoring, readout hardware configuration, trigger system configuration and detector/front-end calibration, as well as a generic application for use by other subprojects. A production equipment database application is also needed to track the status of front-end, readout, trigger and networking hardware in the system (more than 10,000 modules, plus cables).

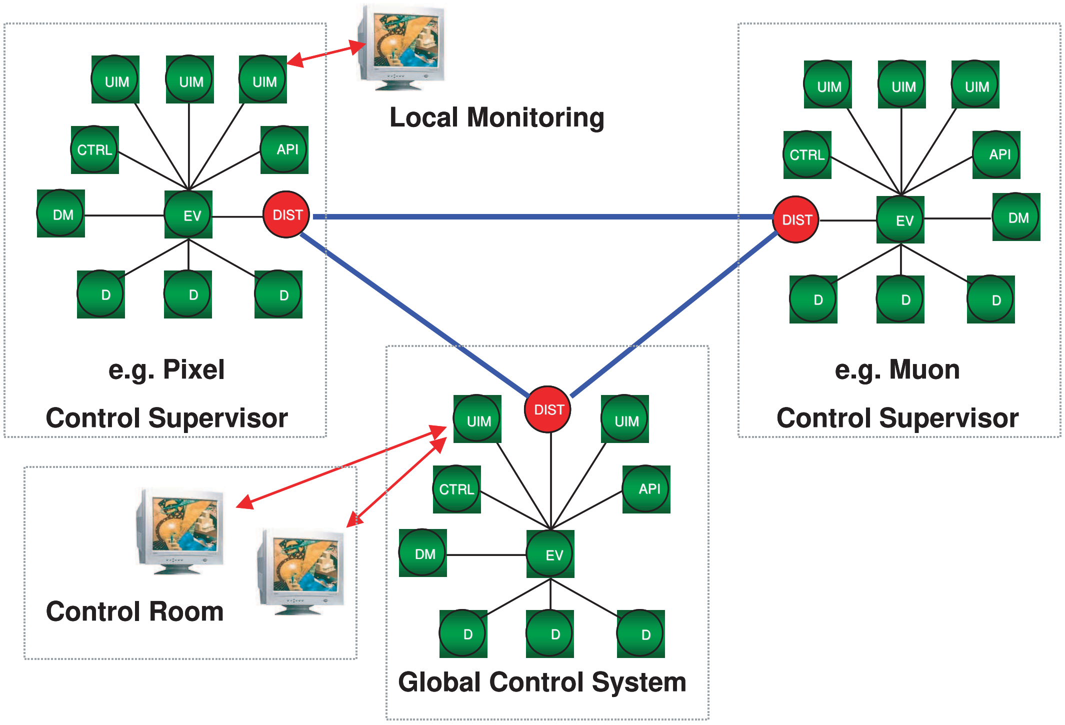

Figure 12.15: Distributed Detector Control.

An extensive evaluation period is anticipated to define requirements for database access, response time, up-time, partitioning, sizing, backup and failover.

12.3.15 Test Stand and Test Beam Support

The readout and controls subproject is responsible for limited support of test stands and test beams, including development of general purpose drivers and software for use with the Pixel system PCI readout card.

The funding profile places delivery of most of the production readout hardware near the end of the project, so it is unlikely that these components will be available for test beam use. We expect to provide some software support for existing data acquisition hardware used in test stands and test beams, but do not plan to develop any hardware specifically for test purposes.

12.3.16 Integration Test Facility

The integration test facility will house a complete vertical slice of the readout and control system along with a subset of the first and second level trigger and front-end electronics. It will include resources necessary to test all system components at full operating bandwidth during both the development and production phases.

12.3.17 Infrastructure: The BTeV Counting and Control Rooms

The BTeV Control and Counting rooms will be located in the C0 building. The Counting Room houses the readout electronics, the run control, database server and detector control system computers as well as the L2/L3 trigger farm. User consoles, alarm panels etc. will be located in the Control Room. Figure 12.16 is a section of the C0 layout showing the counting room area. The Counting Room will be subdivided into three floors. The first floor will house the trigger and DAQ electronics while the L2/L3 farm uses about two thirds of

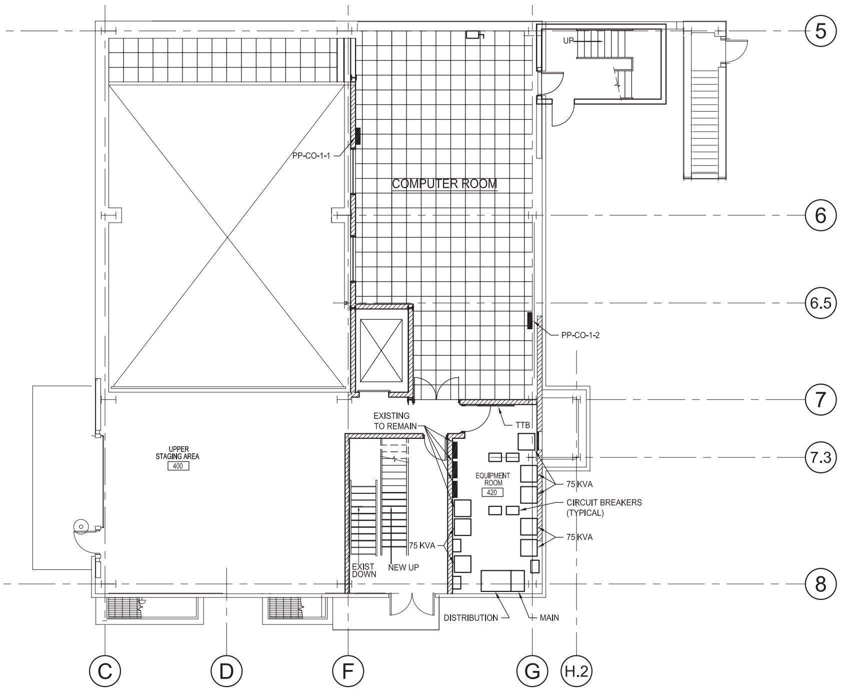

Figure 12.16: Layout of the first floor Counting Room in the C0 Collision Hall.

the third floor. The second floor will house the control room.

12.3.17.1 Rack Count

We have estimated the number of racks needed to house the BTeV trigger and DAQ electronics. We assume that all connections to the detector area are optical and that the data sent from the detector is all digital (with the possible exception of some test and debug signals). Furthermore, we assumed that all the electronics for the detector components will be located in the pit.

We estimate that the DAQ and trigger electronics will require about 180 kW of clean, electrical power. The L2/L3 farm will need 375 kW - not counting the power for extra fans etc.

12.3.17.2 Rack Dimensions, Floor Layouts

The electronics in the counting room will be mounted in standard 19″ racks. i.e. outside dimension/width of 22″. Depending on issues such as airflow (front-back vs. bottom-top) and cable routes the racks will be 36″ or 42″ deep. For the purpose of this document we assume a rack footprint of 24″x42″. While racks can be placed in a row, we must allow door-door clearance between rows. The minimum clearance would be 48″ between rows, to keep the doors from banging into each other. To allow for easier access, e.g. for a scope cart we assume a row-row spacing of 54″. While we would prefer the same spacing between the row of racks and the wall we have to reduce this to 30″ in order to fit 50 racks into the counting room. One 30″ wide mid-row cross walk has been included. Figure 12.17 shows a possible layout for the first and third counting room floors. Space in the counting room will be tight. Providing sufficient cooling in particular for the third floor with the L2/L3 farm will be a challenge and is part of the study presented in the C0 outfitting project.

12.4 R&D

This section describes current and past R&D efforts by the Readout and Controls group.

12.4.1 Architecture

Most of the Readout and Controls R&D effort to this point has been devoted to defining the overall system architecture. During the pre-construction phase, we will do preliminary design of critical sections of each of the major system components, to ensure that the required functionality can be accomplished at or below the projected costs.

One example of architecture optimization is the implementation of data highways which effectively split the readout system into eight parallel data acquisition and trigger paths. This was first proposed by the Trigger group as a way of simplifying the Level 1 Trigger hardware and has obvious advantages for the rest of the system as well. It reduces the

Subsystem	**Card Estimate Rack Estimate**	
DAQ Electronics	24 DCB subracks, 4 per rack	6
	80 L1Bs, 8 per rack	10
	8 Network Switches, 12U each	4
	30 PCs (detector manager,slow control)	2
L1 Trigger Electronics	Pixel and muon L1 trigger	28
L2/L3 Farm	768 dual-CPU units - 1 U modules. (32/rack)	24
	Management System, disk and database server	6
COUNTING ROOM TOTAL		80

Table 12.2: Rack Estimates for WBS 1.9

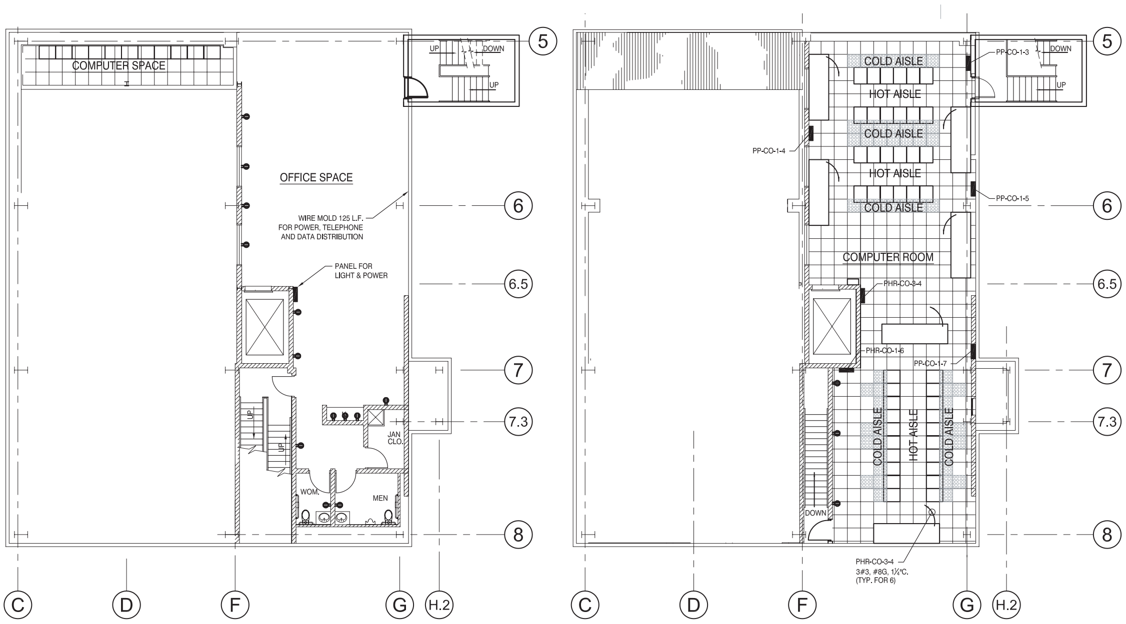

Figure 12.17: Possible Layout of the BTeV Counting Room (Floor 2 and 3).

overhead involved in processing data packets at every stage in the system, by increasing the size of each packet and reducing the number of packets. It also provides a better match for commercially available network switches (and in fact, allows commercial switches to be used, when they would otherwise be too inefficient). Coincidently, a decision to split the CMS readout system into eight parallel paths was made at the same time as the BTeV decision (although this does not extend to the Level 1 Trigger).

A unique aspect of the BTeV system is the decision to transfer all data off the detector at the full crossing rate. This requires a substantial infrastructure for data transport and buffering, but also greatly extends the available L1 Trigger decision time. The result is a very sophisticated first level trigger, which is implemented mainly in software and therefore easily adapted and improved.

12.4.2 Front-end

In cooperation with the Muon group, we implemented a test module to study the effects of mixing fast digital and sensitive analog components on the same circuit board (Figure 12.18). This board included three ASDQ integrated circuits and a medium density FPGA (Altera APEX). It was connected to a prototype Muon plank with high voltage applied. In addition to the ASDQ readout logic in the FPGA, code was added to intentionally generate digital noise both on and off the chip. The results were encouraging, with very little digital noise showing up in the ASDQ signal. We believe that mixing analog and digital circuitry on the same front-end board, with proper isolation, is an acceptable approach. This assumes that the front-end board is located in an area where the radiation levels do not pose a risk to the digital components.

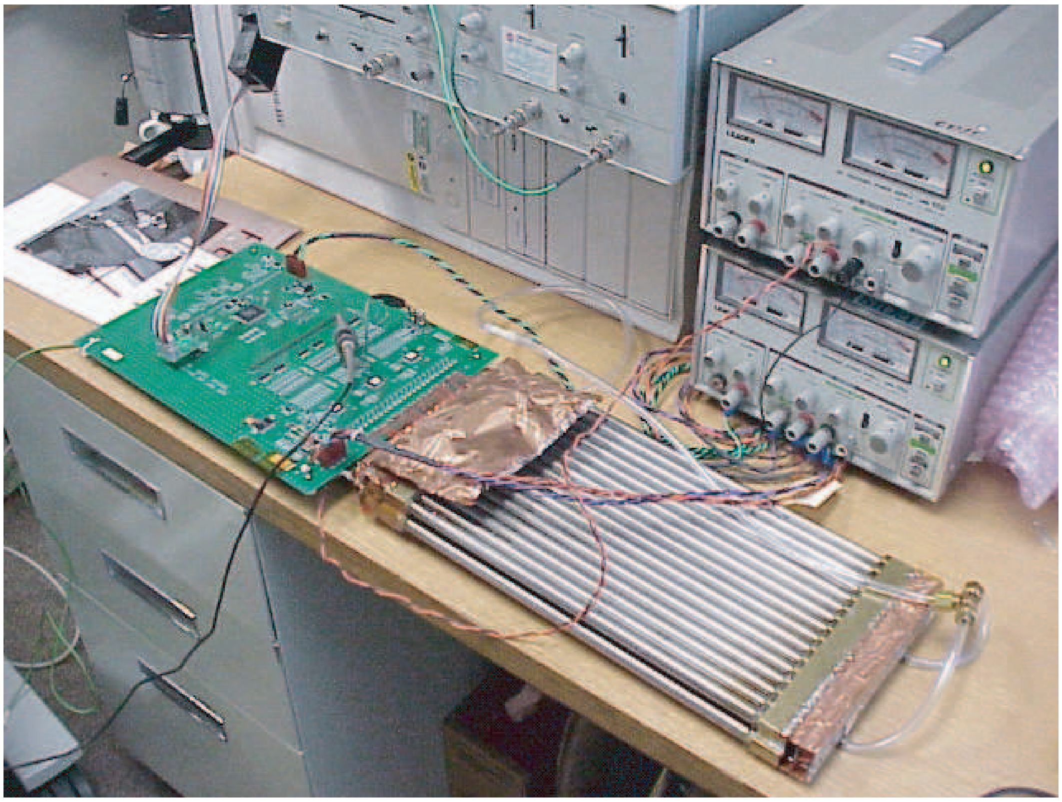

Figure 12.18: ASDQ Test Board with Prototype Muon Detector.

A second area of Front-end research involves the standard interface to the Data Combiner. The baseline design is an individual cable to each front-end module, with two differential signals in each direction (8 wires). The signals from the Data Combiner to the front-end are a crossing clock and a serial control link. The control link messages are framed by the clock, so that commands can be synchronized to a specific crossing. Two serial data links connect the Front-end to the Data Combiner. One or both of these may be used for data output.

Because of the high number of front-end modules, we are looking at ways to minimize the cost of this interface. The use of low cost standard cables is one approach. A shielded CAT 6 network cable includes the necessary four differential links and has been demonstrated to operate at LVDS levels and speeds up to 620 Mbps over distances of at least 5 meters (this is the Starfabric standard physical layer). We plan to test this cable, along with USB 2.0 and IEEE1394 cables to determine if the signal characteristics and connector reliability are suitable for our application.

A new method of distributing clock signals has also been under investigation. The baseline design uses the traditional clock fanout tree, with individual lines adjusted to provide the same clock phase at each front-end module. The alternate approach makes use of a single cable tapped at each front-end module. A pulse (or encoded digital signal) is transmitted down the cable and is reflected at the end. Circuitry in the front-end module then calculates the average of the incident and reflected pulse times, which is identical to the time the pulse reached the end of the cable, regardless of the tap position.

For either clock distribution method, we plan to move much of the timing intelligence as close as possible to the front-end. Only the clock itself and a single synchronous reset signal will be distributed systemwide. All other timing information will be transmitted asynchronously or generated locally and then synchronized by the Data Combiners or front-end modules.

Finally, we are investigating the use of commercial FPGAs as TDCs. The deserializers built into the latest generation of FPGAs are ideal for this application, again provided that the FPGA is not located in an area of significant radiation. The FPGAs include all necessary buffering and processing logic, and can be reprogrammed for specific applications. Simulations have been performed using manufacturers device models to show feasibility, and two PCI test modules will soon be assembled. The first module uses a Lattice Semiconductor FPGA with eight built-in high-speed deserializers. It should be capable of 320 psec per bin resolution, at a cost of approximately $50/channel. This will be followed by a second PCI test card using an Altera Stratix FPGA with up to 64 deserializers. This part is capable of 1.2 nsec per bin resolution at a cost of less than $5 per channel.

12.4.3 Serial Links

The BTeV readout architecture will use many high-speed serial links to deliver data from the front-end to the first and second level Trigger systems. We have built test modules to study the bit-error rates and distance capability of these links using both optical and electrical drivers. The previously mentioned eight channel TDC demonstration card can also be used as a standard multi-channel serial data link. We also plan to test parallel optical transmitter and receivers as they become available. These parts provide the lowest overall cost per link.

A number of standard high-speed serial link protocols are currently in development (PCI 3.0, Serial RapidIO, Serial ATA). As standard interface components become available we will try to integrate them into the serial link testing.

12.4.4 Built-in Test

The cost to design and program a test fixture for an average printed circuit board is approximately $30K. This provides a one-time test of the manufactured product. We plan to develop a set of standard integrated self-test capabilities to be used in all readout electronics hardware that will eliminate the need for production test fixtures, and allow in-situ testing during system operation.

An example is bit-error rate testing of serial links. Successful operation of BTeV will require error rates of 10-15 or better. It would take approximately two weeks of testing to

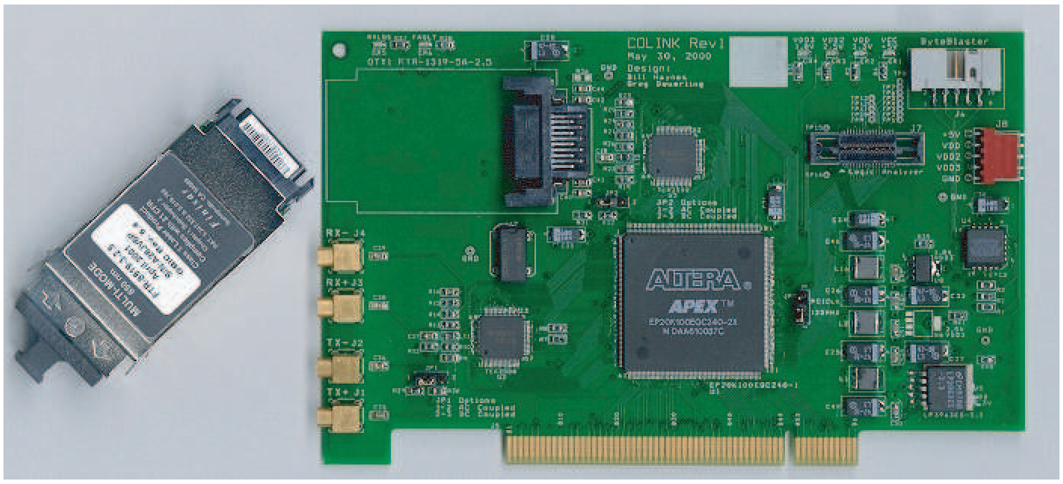

Figure 12.19: 2.5 Gbps PCI to Serial Link Card (Optical or Electrical Interface).

Figure 12.20: Parallel Optical Links (typically 12 channels per link @ 2.5 Gbps per channel).

verify that rate on a single link, multiplied by ≈20,000 links, and the results would mean very little if the interconnecting cable is not the same one used in the final system. The same test can be performed, in system, on all links simultaneously using the pseudo-random bit sequence generation and checking logic built into many new link interface integrated circuits.

12.4.5 Embedded Processing

The initial designs of the Data Combiner and L1 Buffer subsystems include a commercial PC motherboard as the control interface (and also as the output data interface for the L1 Buffer). We plan to specify the development environment for these processors during the pre-construction R&D phase, and begin porting the required kernel software.

The FPGA devices that will likely be used in these subsystems have built-in embedded processors, which also require development tools.

12.4.6 Network

The readout and control network will consist of eight large Gigabit Ethernet switches (one in each data highway), plus a cross-connect switch and a number of Gigabit to Fast Ethernet fan-out switches. A demonstration switch containing 12 Gigabit and 48 Fast Ethernet ports has been purchased for use in developing and testing the network control software and drivers. In addition to these tests, we also expect to benefit from the many years of switch and network research already conducted at CERN.

Each highway switch in the final system will handle 64 Gigabit connections. We plan to compare the performance of a single 64 port switch to that of a network built from several smaller switches (e.g., eight 16 port switches). At current prices, the network based on the smaller switches may be significantly less expensive.

The L2/3 processors connect to the Fast Ethernet fan-out switches, and the final assembly of event data takes place in the processors. A study of the software overhead required to do this final event assembly was conducted by Ohio State with the conclusion that no hardware acceleration (using either a special interface card or a separate processor) would be necessary.

Detector Control System The slow controls network will be Ethernet based, using commercial SCADA control software. During the pre-construction phase, we plan to acquire a development license for the high-level software and begin evaluation of components. Recommendations will be published for general-purpose digital and analog I/O modules and a simple slow controls application will be created.

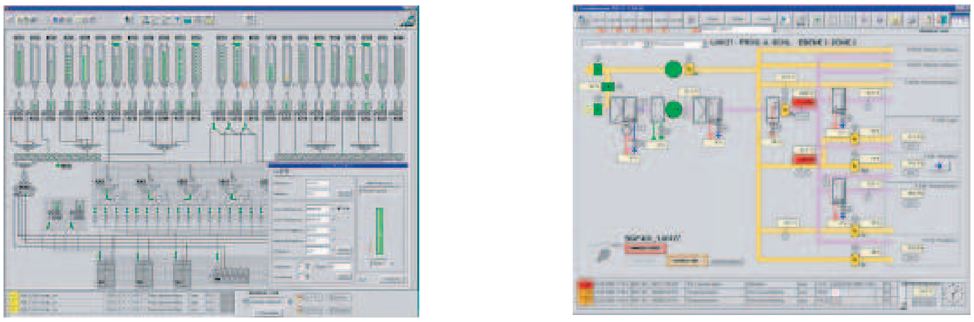

Figure 12.21: Slow Controls Interface Examples using PVSS.

12.4.7 Readout Software

During the R&D phase of the upcoming year, the overall software design documents will be completed which will require evaluation of software languages, development environments, and middleware that will be used through the construction project.

12.5 Production, Testing and Integration

This sectiondescribes the productions, testing and quality assurance plans for the BTeV data acquisition and controls systems. The data acquisition system (DAQ) is responsible to transport the data from the detector to the first level trigger processor(s), to store that data while a Level 1 decision is pending, to forward accepted events to the Level 2/3 trigger farm and finally to send complete events to a mass storage system. The DAQ system has to provide sufficient bandwidth to operate at a bunch crossing frequency of 2.5 MHz and up to 7.6 MHz. The performance of the DAQ system as well as of the other BTeV components is monitored by the detector control system (DCS). In the following sections we list the major production items - both hardware and software - for the DAQ and the DCS systems. Details on the design of the DAQ/DCS systems can be found elsewhere.

12.5.1 Read-Out Electronics

The BTeV read-out electronics consists of Data Combiner Boards (DCBs), optical links, L1 Buffer modules and the Timing/Control system.

12.5.1.1 Data Combiner Boards

The DCBs receive data from front-end electronics modules, combine several input streams into one output stream and transmit the data via optical links to the counting room where the information is stored until the level 1 trigger has reached a decision. The Data Combiner board will be designed at Fermilab. Two prototype steps before are planned before production starts. 256 modules will be needed for the BTeV readout system. Board assembly and initial testing will be done by outside vendors. Full tests will be performed before the modules are installed in C0. Both Fermilab and the Ohio State University have set-ups to carry out this kind of measurement. Based on the prototype experience, the two groups will come up with a standard test procedure for the production DCB modules. A database will be included in the production and test plan so that a test record and shipping log of all DCB modules will be accessible on the web.

12.5.1.2 Timing and Control System

The timing and control system (TCS) distributes synchronous information such as the bunch crossing signal to the front-end electronics modules. It provides control signals to insure that the data pipelines remain synchronized and allows for standard commands such as “Start Run” or “Reset Bunch Crossing Counter” to be distributed. The Timing and Control System will be designed at Fermilab. Two prototype steps before are planned before production starts. Board assembly and initial testing will be done by outside vendors. Full tests will be performed before the modules are installed in C0. Both Fermilab and the Ohio State University have set-ups to carry out this kind of measurement. Based on the prototype experience, the two groups will come up with a standard test procedure for the production

TCS modules. A database will be included in the production and test plan so that a test record and shipping log of all TCS modules will be accessible on the web.

12.5.1.3 Optical Links

The Optical Links provide the connection between the Data Combiner boards and the L1 Buffer system where data for each crossing is stored until a Level 1 decision has been reached. This sub-system consists of Serializer/Deserializer chip sets, the optical transmitters and receivers as well as the optical fibers running from the collision hall to the counting room. The Optical Links system will be developed at Fermilab. Two prototype steps before are planned before production starts. Board assembly and initial testing will be done by outside vendors. Full tests will be performed before the modules are installed in C0. Both Fermilab and the Ohio State University have set-ups to carry out this kind of measurement. Based on the prototype experience, the two groups will come up with a standard test procedure for the production optical modules and links. A database will be included in the production and test plan so that a test record and shipping log of all components of the optical links sub-system will be accessible on the web.

12.5.1.4 L1 Buffer

The L1 Buffer module (L1B) will be designed to receive data from up to 24 DCBs and to store the incoming data long enough for the Level 1 trigger to reach a decision. Accepted events will be sent via a Gigabit Ethernet link to the Level 2/3 farm for further processing. The L1 Buffer board will be designed at Fermilab. Two prototype steps before are planned before production starts. 256 modules will be needed for the BTeV readout system. Board assembly and initial testing will be done by outside vendors. Full tests will be performed before the modules are installed in C0. Both Fermilab and the Ohio State University have set-ups to carry out this kind of measurement. Based on the prototype experience, the two groups will come up with a standard test procedure for the production L1B modules. A database will be included in the production and test plan so that a test record and shipping log of all L1B modules will be accessible on the web.

12.5.2 Data Acquisition Software

12.5.2.1 Software Infrastructure

The software infrastructure for the BTeV DAQ system includes several modules or packages that can be designed and tested in parallel. These include Error Handling and Reporting, Message Passing as well as User Interface Support. These software modules will be designed by groups from Fermilab and The Ohio State University. Standard software coding practices will be implemented to insure that the programs are not only functional but also well documented and easy to maintain. For each package test suites will be included. Collaborative code development tools such as CVS will be augmented by a "release system" that makes it easy for users to obtain a consistent set of the DAQ software libraries.

12.5.2.2 Read-Out Software

The read-out software will be built on top of the infrastructure layer described in the previous section. It is again split into several packages that can be designed and tested in parallel. These include Run Control, Configuration, Partitioning, Eventbuilding, Support for Data Quality Monitoring as well as the data logging sub-system. These software modules will be designed by groups from Fermilab and The Ohio State University. Standard software coding practices will be implemented to insure that the programs are not only functional but also well documented and easy to maintain. For each package test suites will be included. Collaborative code development tools such as CVS will be augmented by a "release system" that makes it easy for users to obtain a consistent set of the DAQ software libraries.

12.5.3 Detector Control System

The Detector Control System (DCS) monitors the performance of the BTeV detector, records environmental data such as barometric pressure and provides an interface to the Tevatron monitor and control system. The data acquisition group provides the control and monitoring software including user interface support and access to the online database. The actual monitoring hardware (sensors, PLCs, power supplies etc) will be provided by the detector components. To insure compatibility and for software development purposes two test labs will be set-up at Fermilab and at Ohio State. Only hard- and software modules that pass these compatibility tests will be allowed in the experiment.

12.5.4 Databases

The BTeV online system will use database to store configuration information, to archive environmental conditions and run parameter, as repository for geometry data needed by the Level 2/3 trigger processes and much more. Database design is a difficult task. Robustness and ease of maintenance of a database depend to a great deal on choosing the right data representation. We will rely on the expertise of the Fermilab database group to develop the system level software. Much of the database application software will be developed by BTeV users.

12.5.5 Control and Data Networks

A core element of the BTeV data acquisition system is a large switched network fabric between the 256 L1 Buffer modules and the Level 2/3 trigger farm. The fabric will be constructed of commercial network switches using a combination of Fast Ethernet and Gigabit Ethernet technologies. Before purchasing the production units we allowed for a prototype phase to test switch performance and to evaluate software protocols. These tests will be performed at Fermilab and The Ohio State University. Based on the prototype experience, the two groups will come up with a standard test procedure for the production switches and the network connections.

12.5.6 Infrastructure and Integration

The readout and controls task includes the infrastructure for the counting room and the control room as well as electronics support for the collision hall. Infrastructure components such as racks, cooling and rack monitoring will be designed by Fermilab during the development phase of this sub-project. Production racks and power supplies will be pre-assembled by the vendor. Final testing including burn in will be done at Fermilab.

12.6 Installation, Integration and Testing Plans at C0

This section describes the Installation, Integration and Testing Plans for the Readout and Controls system.

12.6.1 Summary of Testing Prior to Moving to C0

The entire readout chain will be tested before moving to C0. These tests include front end modules (provided by the detector groups), Data Combiner boards, optical links and the L1 Buffer system. Integration tests will be performed for the Data Combiner to Front End module interface(s), the interface between the L1 Buffer system and the trigger system as well as for the interface between the timing systems and the detector electronics. Included in those tests is not only the hardware but also the software integration of the central run control and configuration systems, user applications and detector component specific components.

12.6.2 Transportation of Readout and Controls Equipment to C0

Equipment Required All readout and controls equipment will be staged at the Fermilab Computer Lab and at the Ohio State University. Equipment will be moved from the Feynman Center to C0 by Fermilab Material Distribution Department Trucks and Drivers. The Ohio State University's motor pool will be used to move equipment from Ohio to Fermilab.

Special Handling Relay Racks will be transported with standard tie-down precautions. Standard precautions (e.g. avoidance of electro static discharges) will be required for the transport of electronics modules. A transportation procedure will be prepared for the transportation operation.

Personnel Required Most of the readout and controls equipment can be maneuvered by hand. The relay racks might require the use of a crane and other equipment to bring them to the first floor of the counting room.

Time Required All relay racks can be loaded and transported in one-half day. Another one-half day will be needed for the transportation of the electronics modules, PC's and other equipment. Transportation of equipment from Ohio State will require two days.

12.6.3 Installation of Level 2 Subproject Elements at C0

Installation Steps Components of the readout and controls system will be placed in the C0 detector hall, the counting room and in the control room (both of which are in the C0 building). Installation of most of the readout and electronics components in the detector hall will be coordinated with the detector sub-groups. As soon as space becomes available, ie is no longer needed for the insertion of detector components, we will install the racks that house the DCBs and the optical switch modules (Note: the exact placement of the DCBs is still under discussion. Some might be located in the detector hall will others will be in the counting room. In the latter case optical switch modules will be used in the detector hall. For each component cables need to be installed to connect the front end modules to the DCB/Optical Switch box about 3,000 cables in total. The connection to the counting room is provided by approximately 256 optical fiber bundles (each with 12 fibers). Before we can run these bundles we will install special inner-ducts in the ducts connecting the detector hall with the counting room. This way we will be able to replace individual fibers should a problem develop. Approximately 300 cables will connect each DCB/Optical Box with the timing system. An installation plan for the readout and controls cabling will be developed in coordination with the detector groups and the overall installation coordinator (wbs 1.10). Installation of readout and controls equipment in the counting room starts with the relay racks, power and cooling. Once these services are available we will install the L1 Buffer system and the Data Combiner modules that are not located in the detector hall. Approximately 3,000 network cables have to be installed between the L1 Buffer system, the switching network and the Level 2/3 farm. Work in the control room can proceed in parallel. Installations steps include setting up the control room furniture, the network infrastructure as well as the computer/operator consoles. Just as the readout system the detector control system requires equipment to be installed in different location. Most of the monitoring and control system in the detector hall will be installed by the sub- detector groups. Network (Cat 5) and field-bus cables will connect these systems to the supervisor components of the control system that are located in the counting room. The precise location of the equipment computer (Detector Manager and Control Manager/Supervisor) still needs to be defined. While most of these workstations will be placed in the counting room some need to be close to the hardware and will reside in the detector hall. The elements of the Global Detector Control System will be split between the counting room (supervisor CPUs) and the control room (workstations with the user interface(s)). Installation of the detector control will be coordinated with the detector group and the installation coordinator (wbs 1.10). An installation plan will be developed.

Equipment Required No special installation equipment is required. A crane might be needed to lower (some of) the relay racks into position.

Special Handling Issues Electronic modules have to be handled with care to avoid damage due to electrostatic discharge.

C0 Infrastructure Required Utilities required at C0: electrical power, water cooling for the relay racks, network connection.

Potential Impact on Other Level 2 Subproject For a system test each component needs to have at least parts of the readout and controls equipment in place. However, care must be taken to avoid that these modules and cables block access to the detector and impede the installation of other components. A detailed cabling schedule will be developed.

Accelerator Impact of Installation None - of course we need access to work in the detector hall.

Safety Issues None (besides standard work place safety)

Personnel Required Riggers for the relay racks, furniture. Electricians, plumbers for the electric and cooling infrastructure (relay racks).

Time Required

Install 3000 readout cables (front end to DCB/Optical box)	≈1000 hours
Install 256 optical fiber bundles	≈300 hours
Install 3000 network cables	≈500 hours
Install 300 timing cables	≈150 hours
Install 22 relay racks, connect to services	≈100 hours
Install ≈240 DCB modules	≈10 hours
Install ≈320 L1B modules	≈10 hours
Install detector control cables	≈150 hours
Install PCs and workstations (≈50)	≈100 hours

12.6.4 Testing at C0

1. Infrastructure Tests
   (a) Utilities - Leak test cooling water systems
   (b) Safety Systems -Test electrical safety
2. Control/Monitoring System

(a) Interface the detector control system to the detector specific control and monitoring system.
(b) Complete integration.

3. Timing/Clock System - Clocks will be needed to do a full readout test

4. Stand-Alone Subsystem Testing
   (a) Mechanical - verify that the system fits together.
   (b) Electrical/Electronics - Repeat internal test program developed previously using the Integration Test Facility.
   (c) Power supplies and network connections.
   (d) Software - Repeat internal test program developed previously using the Integration Test Facility.
   (e) Personnel Required - to be determined.
   (f) Time Required - to be determined.

5. Multiple Subsystem Testing
   (a) Mechanical - None
   (b) Electrical/Electronics - Repeat internal test program developed previously using the Integration Test Facility including tests of the entire readout chain and the detector control system.
   (c) Software - Repeat internal test program developed previously using the Integration Test Facility including tests of the entire readout chain and the detector control system.
   (d) Personnel Required - to be determined.
   (e) Time Required - to be determined.

12.7 Organization

At the time of this writing, the list of institutes participating in the readout and controls task includes Fermilab and The Ohio State University. Staffing is not yet completed and other institutions are expected to join this effort.