Chapter 12

Event Readout and Control System

The BTeV data acquisition system (WBS 1.9) consists of two parts, the Readout System used to transfer data from the detector to on-line processing and archival storage, and the Detector Control System which provides data quality monitoring and ensures that all BTeV components operate within design specifications.

A review of the design requirements is followed by detailed discussions of the architecture of the readout system, the data acquisition software, and the detector control system.

12.1 System Overview

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 (crossing interval of 396 ns), and the front-end boards transmit data for every crossing. The average event size has been estimated to be less than ≈200 KBytes. These estimates were obtained using a full Geant based simulation of minimum bias interactions, with allowance for uncertainty in background and detector noise. 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 and suitable design margin increases the bandwidth needed to approximately 800 GBytes/s.

All these data coming out of the detector need 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 less than 500 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.

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 16 GBytes/s will be required.

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

How BTeV will meet the requirements outlined above is 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 Combiners (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 the Data Combiners to Level 1 Buffers (L1B) and to the first level trigger (pixel and muon data only) in the counting room. Each link will operate at 2.5 Gbps.

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

4. Eventbuilder Network: A segmented switching network to combine data from the L1 buffers and 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.

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

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 are below predetermined thresholds and are 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 have not already been compressed at the detector. All data presented to the Readout System Electronics are 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 further reduction in event size. 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) to receive 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 Level 1 Buffers receive data from the Data Combiners and from the input stage of the Level 1 Trigger. The data are held until a trigger decision is made, and then either discarded or forwarded to the Level 2 Trigger. The Level 1 Trigger must return a decision for all crossings within a specified maximum latency, even if processing for those crossings is not complete.

The Level 1 Buffers must accept data that are 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 Level 1 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 Level 1 Buffers must stop accepting input data and must allow data already in memory to be output without being overwritten. A Level 1 Buffer must be capable of generating a substitute 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 Data Network

For each event accepted by the Level 1 trigger system, all necessary data from all detector subsystems must be combined and delivered to processors in the second level trigger system. The Data Network must be capable of delivering a combined rate of 10^{11} bits per second to the Level 2 trigger system.

The Data Network must be capable of routing data from any Level 1 buffer to any Level 2 trigger processor (if multiple readout paths are implemented, Level 1 buffers in one path need not connect to Level 2 processors in a different path, but there must be a way to transfer data at lower speed between Level 2 processors in different paths)

The Event Building software must provide buffer space for at least 16 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 and other 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 and Control system generates signals that are synchronous to the accelerator clock. It is assumed that only the front-end electronics, Data Combiners and Global Level 1 Trigger will require synchronous timing, and that all other components of the Readout System and Trigger System operate asynchronously. The clock signal will be 7.5 MHz (132 nsec). This is three times higher than the crossing rate of 2.5 MHz because of technical reasons having to do with the structure of the accelerator as described in Sec. 12.3.2. The Timing and Control 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 and Control 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 at the next synchronous clock. If the next value of the counter at the time of the synchronous clock is incorrect, 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 DAQ 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 instances. 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 hetergeneous 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 in the primary partition.

12.2.11 Trigger and Detector Managers

The Level 1 and Level 2/3 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 Staging

Events passing the Level 2/3 Triggers will be transmitted to a Data Staging system. This Data Staging system holds data for an extended period pending transfer to permanent storage and allows reprocessing during idle periods of the L2/L3 farm.

The Data Staging 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 locally attached disk drives. These may be used to buffer data during short power or network interruptions. The network supplying data to the Data Staging system, and the Data Staging system itself, must have excess bandwidth capacity to offload the accumulated data in a reasonable period of time. The Data Staging system must support storing similar events as a collection, based on trigger type.

12.2.13 Slow Controls

The BTeV Slow Control system is used to monitor and set controls/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 Network

The Control 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 C0 detector building. All information must be electronically accessible over the standard network.

12.2.16 Databases

Databases are used throughout the system to provide access to configuration parameters 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.

For high-current (greater than 10 amps operating or 50 amps rated current), low-voltage (less than 50 volts) supplies powering the digital circuitry, the safety requirements for high current power distribution systems must be followed. These are detailed in the Fermilab ES&H 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. The 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 "single datapath" architecture faces several problems:

- It requires a switching network with 400 ports (200 for data buffers and 200 for L2/L3 processors).
- It generates data packets with average length of a few hundred Bytes. Short data packets are not handled efficiently by most networking equipment.
- Level 1 Accept and Event Routing messages must be broadcast to every buffer module - at a rate of up to 100 KHz.

The effects of the message size on the network efficiency can be seen in Figure 12.1. 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 (DCB) modules, which immediately multiplex packets from many front-end boards to form larger packets (200-400 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 interval appears to be 3 microseconds (8×396 ns), with a corresponding $8\times$ decrease in the packet processing overhead.

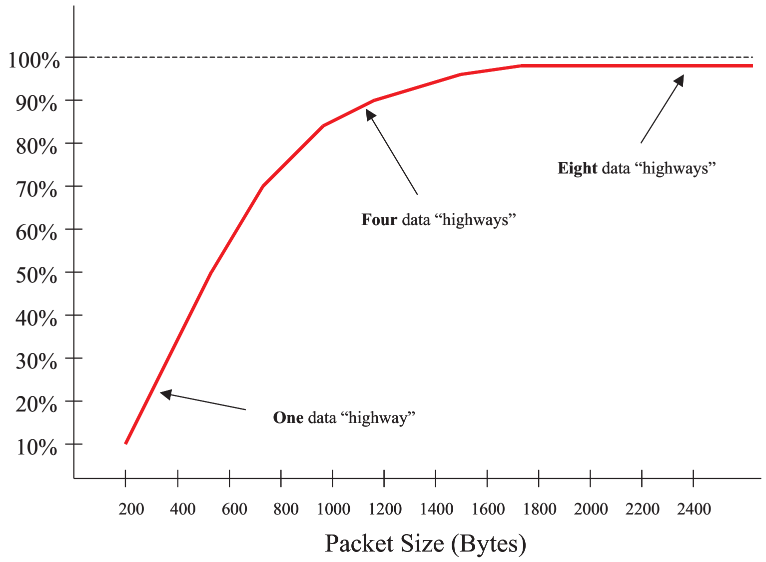

Figure 12.1: Typical Gigabit Ethernet Efficiency.

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 are either processed by the first level trigger system and sent to Level 1 Buffers or sent directly to the buffers. The decision of the first level trigger is then transmitted to the Level 1 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, the control message traffic is also reduced 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 additional Gigabit Ethernet switches. This way it will still be possible, for calibration and test purposes, to route data from any particular bunch crossing to any Level 2/3 farm node - just not with the full DAQ bandwidth. A detailed view of the BTeV readout system including timing, detector control and monitoring can be found in Figure 12.2.

The Readout and Control System design will accommodate an average bandwidth of approximately 500 GBytes/s during steady-state operation. Additional margin is provided for inefficiencies in front-end data balancing and link utilization, and for noise in the detectors.

12.3.1 Timing and Control

The BTeV Timing and Control System (TCS) generates and distributes two synchronous signals, a 7.586 MHz Clock and a command "Sync". The TCS clock reference is a narrow range VCXO locked to the accelerator RF. Each DCB subsystem receives the Clock and

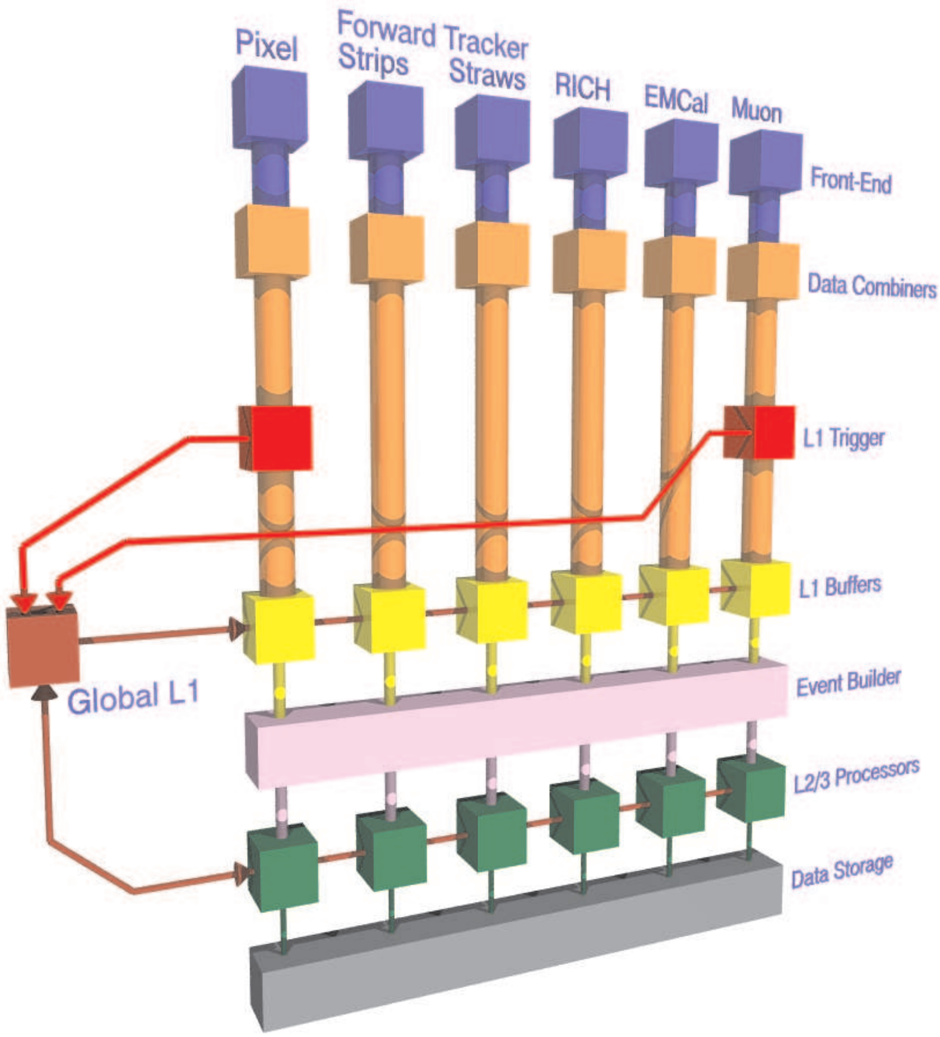

Figure 12.2: Block Diagram of the BTeV Readout System showing 8 parallel Highways.

Sync signals directly from the TCS (via optical links). A PLL at the DCB subrack is used to regenerate and phase the clock. The resulting clock jitter at the front-end module is expected to be less than 50ps. The Sync signal is used to identify a specific clock for systemwide synchronization. Timing accuracy of the Sync signal need only be sufficient to meet setup and hold times relative to the leading edge of the clock. The main purpose of the Sync signal is to reset all crossing counters, but it can be used for any predetermined synchronous operation.

The DCB modules perform all fine-grain timing and clock phase adjustments. The adjustments can be done independently for each front-end link to compensate for differing cable lengths. Static timing information such as the bunch fill pattern is kept in the DCBs.

The crossing counter in the DCB counts input clock cycles, not actual beam crossings.

There are 159 of these clocks in each beam rotation (turn), so the low byte of the crossing counter is modulo 159. The remainder of the crossing counter is a count of the number of turns. The total counter length is at least 52 bits, allowing unique identification of all crossings over the life of the experiment.

Control messages and commands (start/stop/calibrate) are distributed to the DCBs via Ethernet messages. These Control messages may come from either the Detector Manager (detector specific), from Run Control (global) or from the TCS. Messages are asynchronous, containing both the command and clock number, and need not be time ordered. The DCBs synchronize the control messages to the requested clock frame and forward the information to the front-end modules as necessary. Since data are sent from the detector to the DCBs on every crossing, no fast trigger signals are required.

The TCS will periodically send a synchronization check message (via Ethernet) to the DCBs. On the next Sync, the DCB compares the value of its crossing counter to the value sent from the TCS. If different, the DCB resets its counter to the proper value and returns a timing error message (also via Ethernet). Data from that DCB for the period between synchronization checks is then suspect. A similar check is performed between the DCB and its associated front-end modules, to the resolution of the front-end crossing counters (typically 8 bits or less).

Each DCB contains an independent timing state machine and crossing assignment table. In normal running, the content of this table is identical for all DCBs. It can be varied if the system is partitioned. The table holds beam crossing and calibration assignments for each of the 159 clocks per turn for a period of up to 32 turns. Regular calibration events which occur at a frequency of once per 32 turns or higher can be preprogrammed in the table. Events occurring less frequently can be requested via the standard Ethernet "command @ clock" protocol and are synchronized to the specified clock by the DCB.

A simplified block diagram of the timing system is shown in Figure 12.3.

The BTeV timing system re-uses existing hardware such as clock decoders and timing generators currently being developed for the Tevatron BPM project. Other hardware components such as VME CPUs and network switches are available commercially. The optical fan-out cards are a new but fairly simple design.

12.3.2 Front-End Interface

We assume a model where all digitization (and data reduction where appropriate) is performed on the front-end modules. Data are then transmitted on serial links to the 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.

The system clock is synchronized to the accelerator RF and is distributed to the front-end modules as a separate, unencoded signal. The clock is $3\times$ the beam crossing rate to account for the non-integral length of the abort gaps with respect to the crossings.

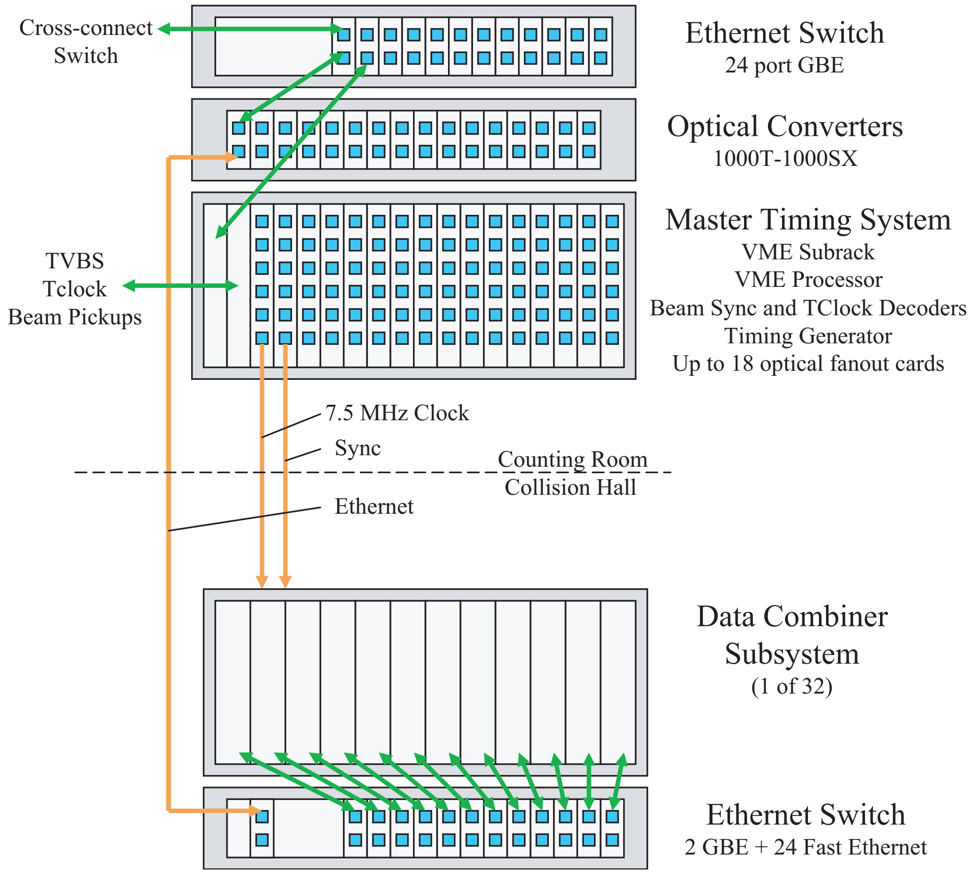

Figure 12.3: Timing and Control Distribution.

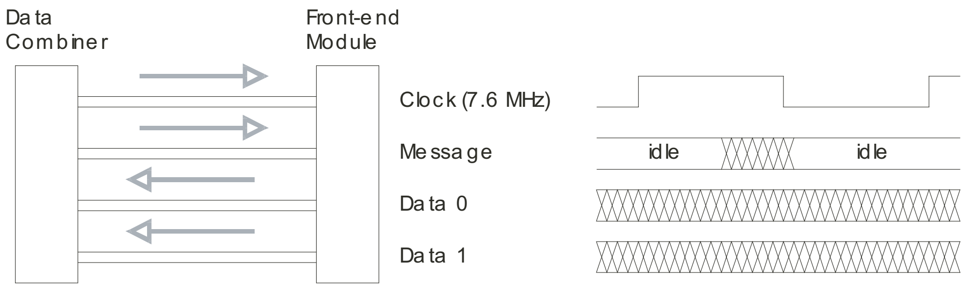

Figure 12.4: Front-End Interface.

In addition to the system clock, there will be a single 0-160 Mbps serial link providing control information to the front-end. The control link will be framed by the system clock, such that any control word received by the front-end will have a guaranteed setup and hold time with respect to a specific rising edge of the clock. The control link is limited to 160 Mbps so that it can be decoded using a simple 4× oversampling receiver.

As a baseline, we are assuming that each front-end data link will operate at a rate of either 140 Mbps (Pixel and Forward Silicon detectors) or 600 Mbps (all other detectors). These limits are influenced by the ASIC processes and by the cost of high-speed data cables and connectors.

The transmit clock for the 600 Mbps data links is derived from the crossing clock using a 80× PLL. The receive clock is derived in the same way and phased aligned to the incoming data.

The transmit clock for the 140 Mbps data links is sent independently as a pair of bi-phase 70 MHz signals and does not require a PLL at the front-end. For data reception, the DCB may use either the source synchronous clock returned by the front-end or the original transmit reference clocks.

The least expensive interconnect for the front-end to Data Combiner link is Category 6 network cable. This is shown to operate at the 600 Mbps rate over distances of at least 5 meters. We will also examine the reliability of the standard RJ45 connectors, which would further reduce cable costs if acceptable.

CAT-6 cable contains 4 twisted 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 12.4). Shielded cables are used to provide a common-mode return path.

If the output bandwidth of a front-end module exceeds the capacity of a single port (2 × 600 Mbps), additional ports may be used. 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 576 Data Combiner modules, packaged in groups of 12 to form 48 Data Combiner Subsystems (Figure 12.5). This grouping is designed to match the output channel count of the DCBs (8 channels each) to the channel count of the optical links (12 channels each). A DCB Subsystem backplane implements the 12 × 8 to 8 × 12 "shuffle" network.

A DCB multiplexes data from up to 48 serial links at 600 Mbps or 144 serial links at 140 Mbps. The DCB output links provide a combined bandwidth of 16 Gbps. The input bandwidth is oversubscribed, but with variations in occupancy the average front-end link utilization is expected to be less than 50%, and most applications will use fewer than the maximum number of input links.

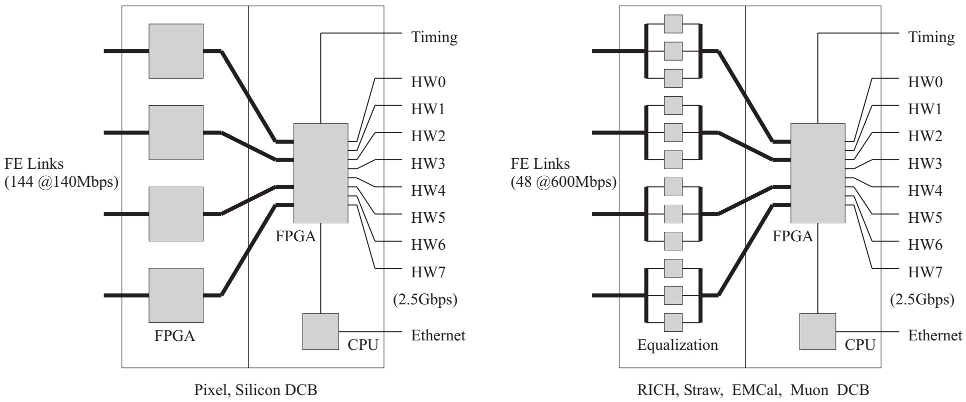

Figure 12.5: Data Combiner Block Diagram.

Crossing data are typically distributed to highways in a "round-robin" sequence. The effective crossing rate for each highway is therefore one eighth of the beam crossing rate, or about 215 KHz. The DCB crossing assignment table allows specific highways to be enabled/disabled. It also allows skipping of highways in a uniform pattern (e.g., 01234567, 12345670, 23456701,...70123456) to avoid any event size resonance between crossings and 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 the Level 1 Trigger, or directly to a Level 1 Buffer.

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 individual crossing timestamps in each input packet and inserting a single timestamp (crossing count) 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.

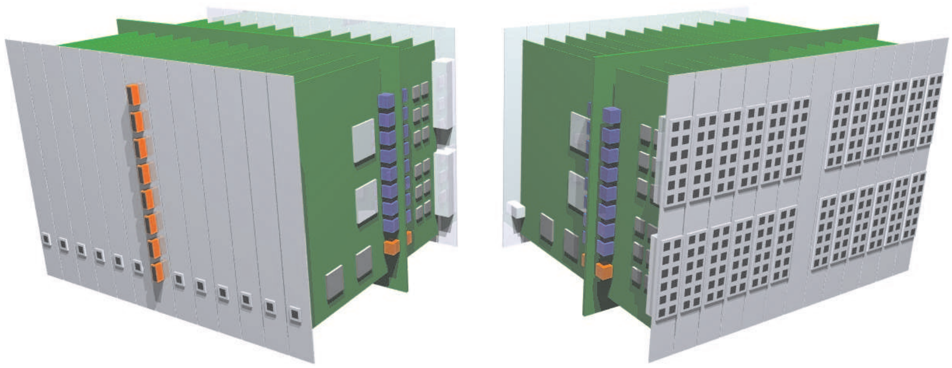

Figure 12.6: Data Combiner Implementation.

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 Level 1 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 exceed 1 KRad/year. At these levels the DCB is not expected to suffer from total dose effects, but will undoubtedly incur single event upset (SEU) to the FPGA configuration memory and data buffers on a regular basis.

To mitigate these effects, the DCB FPGA configuration memory is scanned continuously to detect errors. When an error is found, the on-board processor reloads the FPGA. The processor flash memory is less susceptible to upset, but will also be scanned continuously. Logic and data redundancy are used wherever practical to reduce temporary effects of upsets.

The DCB processor also handles the Ethernet communication, interpreting timing requests from the TCS and formatting data snapshots for the Detector Managers. A possible DCB implementation is shown in Figure 12.6.

12.3.4 Optical Links

The serial outputs of the Data Combiner boards are 2.5 Gbps copper links (2.0 Gbps unencoded data rate). The maximum distance from the Data Combiners (located near the detector) to the Level 1 Trigger System and the Level 1 Buffers (located in the Counting Room) is 60 meters.

This exceeds the distance considered acceptable for reliable high speed data transmission over copper cables, so the electrical signals are converted to optical on the DCB backplane.

Figure 12.7: Parallel Optical Transmitter and Fiber.

The most cost-effective links for this application are based on unidirectional parallel optical transmitters and receivers (Figure 12.7). A total of 384 of these 12 channel optical links provide a maximum data capacity of $\approx$ 1 TByte/sec. The optical receiver plugs directly into the DCB backplane and Level 1 Buffer module/Level 1 Trigger interface, using a 10×10 BGA connector. The use of optical links also provides the benefit of electrical isolation between the Collision Hall and the Counting Room.

12.3.5 Level 1 Buffers

The Level 1 Buffer (Figure 12.8) receives partially multiplexed event data from Data Combiners and Level 1 Trigger Processors on 24 input links (at 2.5 Gbps each). The links may be optical (two 12 channel parallel optical receivers) or copper (one 12× Infiniband cable). Copper links are used for the Level 1 Trigger to Level 1 Buffer interconnection. 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.

Groups of eight input links are routed to a single FPGA containing the deserializers and buffer controller. Each FPGA controls two banks of standard DDR SDRAM. There are three of these FPGA/memory blocks on each Level 1 Buffer module, with a combined memory bandwidth of 16 GBytes/sec. The memory is partitioned into 8 independent circular buffers of 100 MBytes each. Remaining memory is used to store the crossing index table. The circular buffers are sequential write/random read, so that the Level 1 Trigger decisions do not have to be time-ordered.

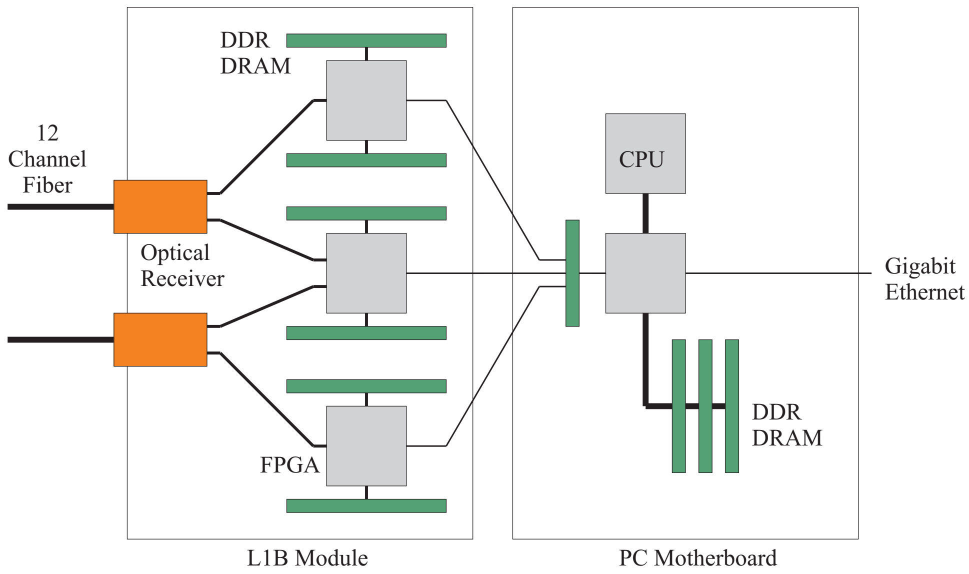

Figure 12.8: Level 1 Buffer Block Diagram.

The data stream in each channel is examined to locate crossing boundaries. Data are written to the circular buffer and a pointer for the associated crossing number is written to a lookup table. The circular buffer has a capacity of approximately 100-200 thousand crossings (depending on link occupancy). This corresponds to a minimum of 500 milliseconds of available Level 1 Trigger decision time, long in comparison to most existing first level trigger systems.

Level 1 processing will timeout and automatically accept the data if the processing time approaches this limit. The buffer size can be expanded at minimal cost if trigger simulations warrant. With low-cost 512 MByte DIMMs, the total Level 1 Buffer size for the readout system is 576 GBytes.

When the Level 1 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.

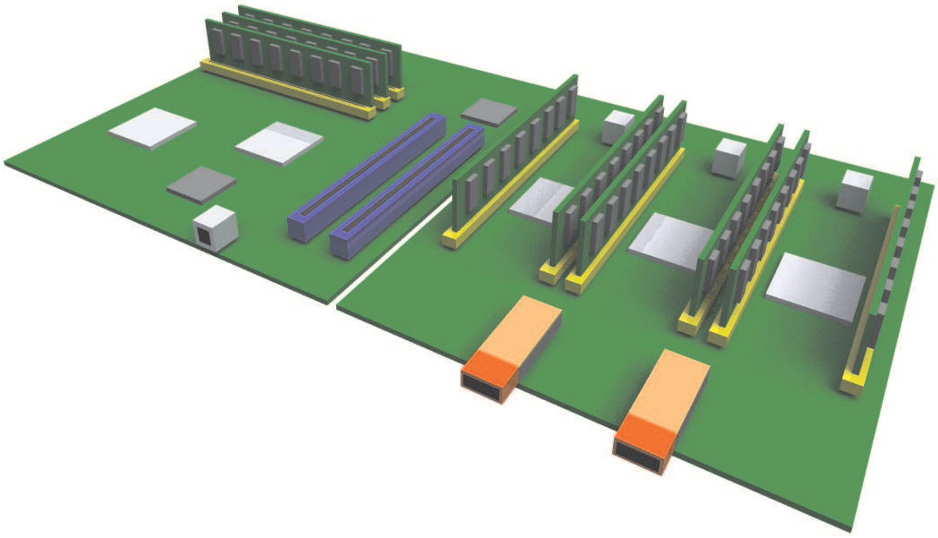

Figure 12.9: Level 1 Buffer Implementation.

The serial link receivers, input circular buffers and multiplexing logic are implemented on the Level 1 Buffer module. The Level 1 Buffer module is paired with a standard PC motherboard to form a Level 1 Buffer Subsystem. The PC provides the Gigabit Ethernet port and the memory for the output buffer. The complete Level 1 Buffer Subsystem appears as a standard TCP server on the L2/3 network. There are a total of 192 Level 1 Buffer subsystems, 24 in each highway. Additional Level 1 buffers are implemented in the Level 1 Trigger processors. A likely implementation of the Level 1 Buffer is shown in Figure 12.9.

12.3.6 Network

Data from the Level 1 Buffers in each highway are transferred to L2/3 Processors. A single L2/3 Processor receives all of the data for a particular crossing, and the final stage of event building is done in the L2/3 processor.

A network of Gigabit Ethernet switches connects the Level 1 Buffers and the L2/3 Processor Farm. The primary switch in each highway provides 72 Gigabit Ethernet ports with the following assignments:

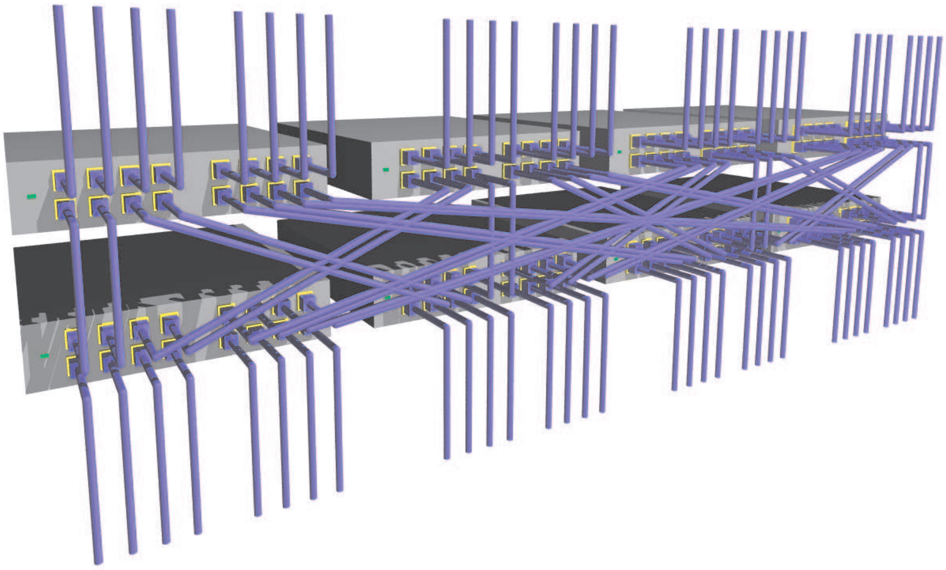

Figure 12.10: Possible Implementation for Highway Data Network.

The 36 L2/3 Fanout ports connect to a second group of switches located in individual L2/3 Processor racks. Each of these fanout switches serves up to 7 L2/3 Processors. With this configuration, 252 L2/3 Processors may be attached to each highway.

The primary highway switch can be implemented using a single 72 port switch or a number of smaller switches (e.g., six 24-port switches, arranged in two stages as shown in Figure 12.10).

Because the dataflow is predominantly in one direction on each switch port, switches with "oversubscribed" (partially blocking) fabrics are satisfactory.

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

The eight highway switches are cross-connected to allow communication between highways at lower speeds (e.g., to read data from sequential crossings for calibration purposes). A

Connection	**Number of Ports**
L1 Buffers	26
L2/3 Fanouts	36
Global L1	3
Control Switch	3
(reserved)	4

Table 12.1: Network Port Allocation

separate Control Switch provides the central network interconnection point for the Detector Managers, Run Control processor, Slow Control processors and Database Server.

This switch also connects to the data storage system, allowing L2/3 processors in any highway to write accepted events to any storage device. The Control Switch also provides the appropriate security features for connection to the external network.

Control Switch assignments are:

Host	
Run Control	1
Detector Managers	6
Trigger Managers	2
Database Server	2
DCBs	2
Data Storage	4
Slow Control Processors	2
External	1
L2/L3 Manager-I/O Host	2
Data Switches	8

12.3.7 Event Identification, Event Building and Distribution

The basic crossing identification is derived from the 7.5 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 Level 1 trigger processor and input stage of the Level 1 Buffers, the resolution must be at least 24 bits (2 million crossings), and for events that pass the Level 1 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 12.11.

The same switching network is used for both data and control, with control messages taking a small fraction ($< 5\%$) of 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.

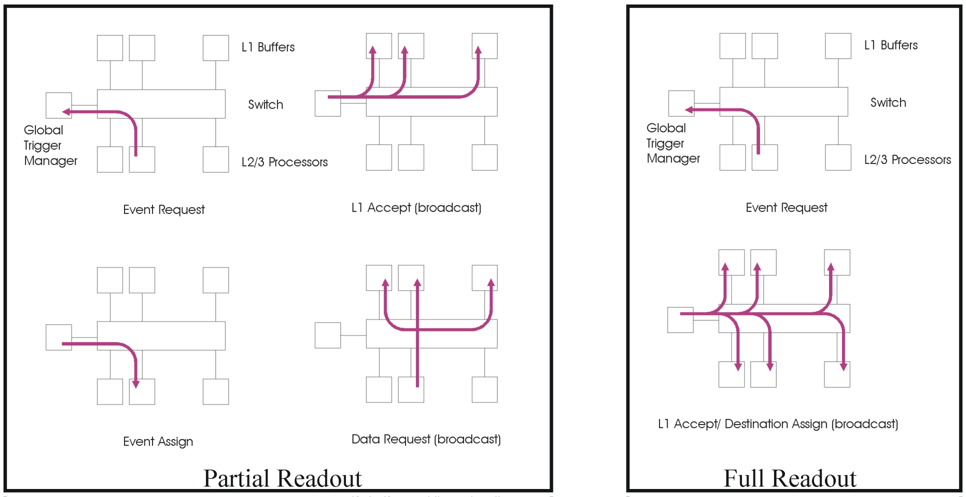

Figure 12.11: Basic Readout Control Messages.

L2/3 processors make event requests to the ITCH (Information Transfer Control Hardware). 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 Level 1 Buffers telling them to transfer that event to their output buffers. The Level 1 Trigger and ITCH have approximately 500 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 ITCH 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 Level 1 Buffers. The baseline design assumes that the data request is a simple broadcast to all Level 1 Buffers in the highway, but the L2/3 processor has the option of requesting data from a subset of buffers, analyzing those 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 Level 1 Buffers in the highway. Alternatively, the event assignment/routing information can be included in the Level 1 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 Level 1 Buffer approaches the capacity of the buffer (>95%), a warning message is sent to the Global Trigger controller. The ITCH 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 are 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 200-400 bytes. These are sent to an Level 1 Buffer where data from up to 24 DCBs are merged. At this stage an event is split into fragments of 2-8 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 percent 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 system. The implementation has not been finalized but the baseline conceptual design is as follows. Each L2/3 worker node collects and keeps the output data from one run segment before transferring it to the DAQ staging storage. For a 10 minute run segment, the average run segment data size is about 80 MB per L2/3 worker node and contains about 1000 events (crossings). As the data for each run segment is sent to the DAQ staging storage, if requested the run segment data may also be written to the local L2/3 worker harddisk for caching for future use. The run segment data is also saved to local disk on the L2/3 worker node if there are problems transferring the data to the DAQ staging storage. Since the network data transfer rate has sufficient capacity the data compression (*e.g.* gzip), may take place in the DAQ staging storage nodes. Each L2/3 worker node communicates with the Global Trigger controller to keep track of run segment data for the event book-keeping.

The DAQ staging storage has capacity for many months worth of data taking. This staging storage acts as both a buffer to the data archival storage and as a large data cache for possible further processing. To increase the efficiency of data transfer to the data archival storage, run segment data will be combined into larger files before the transfer, and as already mentioned, data compression could be done at this stage. Further processing of the run segment data is possible on the L2/3 worker nodes during idle periods. This further processing could include calibration and alignment processing; compiling monitoring data; splitting of data into different data streams; combining run segment data into larger files; and the offline L4 fast charm and beauty monitoring. Included in the further processing is communication with the database service for the event catalogue book-keeping.

A sizable DAQ staging storage is included as part of the DAQ. This staging storage is enough to hold all the data from the Stage 1 running, including some additonal data either from calibration runs or from extra prescale data (or data with less data reduction). The seven months of running for Stage 1 (which include one month of IR commissioning) will produce about 1 petaByte of data that need to be archived. Extra capacity is included to account for additional data that would likely be taken at this time for calibrations and tests.

For the DAQ Staging Storage and for the data archival storage we are looking at mass storage systems such as the Fermilab disk-based dcache system in use by Run II experiments at Fermilab. The current dcache development path for Run II will meet most, if not all, of the requirements for the BTeV mass storage system. There is still some time before we need to decide on which mass storage technology to use. Given the rapid advances in this sector, we decided not to specify details of the implementation of the DAQ Staging Storage and the data archival storage system at this time. Instead we have allocated funds in the baseline for the hardware purchase of the first part of the DAQ Staging Storage in FY08, and the remaining in FY09.

12.3.9 Common Electronics Features

All components in the readout system are designed to operate at a single source 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, Level 1 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 data acquisition software supports independent, possibly 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 Level 1 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 ensure 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 Level 1 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 Level 1 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).

The Detector Manager 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 control switch.

A Control Supervisor is responsible for controlling and monitoring the slow control subsystems 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 (e.g., high voltage systems, cooling systems and calibration pulsers) need to be controlled. Thousands of parameters (e.g., power supply voltages, temperatures, gas mixtures) need to be monitored at regular intervals. These monitoring and control tasks will be performed by the BTeV Detector Control System (DCS). A schematic block diagram of this system is shown in Figure 12.12.

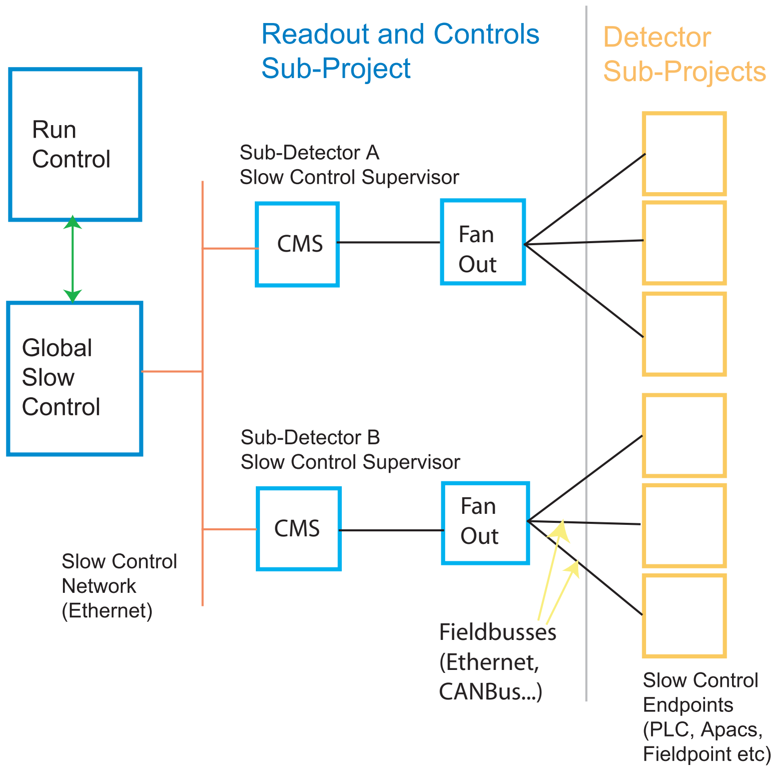

Figure 12.12: Block Architecture of the Detector Control System.

The BTeV DCS system will manage the slow control needs of all BTeV sub-detectors including the trigger system, pixel system, silicon strip and straw trackers, the RICH detector, electromagnetic calorimeter and the muon system, but also detector specific components of the C0 collision hall building, the analysis magnet SM3, and the electronics' racks protection system.

Additional 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).

BTeV management has assembled a slow controls task force that was charged to collect slow control related information from every detector component as well as the groups responsible for the C0 collision hall infrastructure. The task force has submitted its report which will be guiding the readout and controls group (i.e. WBS 1.9) in the design of the detector control system. The total number of personnel working on detector control is intended to increase and to comprise members with an appropriate degree of specialization. As an additional organizational step the task force recommended the formation of a detector control group consisting of members from the readout and controls group (WBS 1.9) and liasons from each sub-detector.

This group will be responsible for developing a complete and practicable detector control system. It is envisioned that this group will subsequently take over management of the construction phase as well as overseeing the DCS during the running of the experiment.

The BTeV detector control system will be based on a commercial software package implementing the supervisory, control and data acquisition (SCADA) standard. To minimize development costs and to maximize maintainability, we will use commercially available controllers, interface modules and software throughout the entire detector control system wherever possible.

We anticipate that the BTeV DCS will have a hierarchical structure to exploit the modularity available in (some) SCADA sytems. Figure 12.12 shows a possible DCS architecture.

For each sub-project (e.g., detector components, analysis magnet, collision hall, etc.) a complete SCADA system will be implemented on the Control Supervisor PC. This system is self contained and sufficient to monitor and to operate each sub-detector slow control system in stand-alone mode (e.g., during commissioning).

A schematic diagram of a Control Supervisor is shown in Figure 12.13. The services of the central control system will be implemented using the same software architecture. This simplifies not only software development and maintenance but allows us to move modules such as graphical user interfaces between the local component control system and the main user consoles in the counting room.

The connection between the global control system and the sub-detector control system is shown in Figure 12.14. The readout and controls subproject (WBS 1.9) is responsible for the purchase, installation and maintenance of the detector control host computers, the maintenance of the "control supervisor" software for each sub-detector and the slow control network (Ethernet). The DCS network connects the control supervisor computer to the central control system that will provide central services such as archiving, alarm reporting and user consoles.

Some common software used by multiple detector groups will be written and maintained by the DAQ group for general use by the experiment. The endpoint hardware and any associated local controllers are detector specific in most cases and are the responsibility of

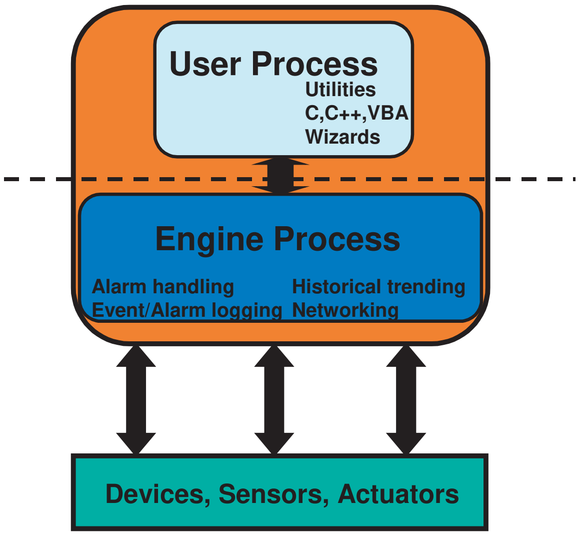

Figure 12.13: Block Diagram of a Control Supervisor.

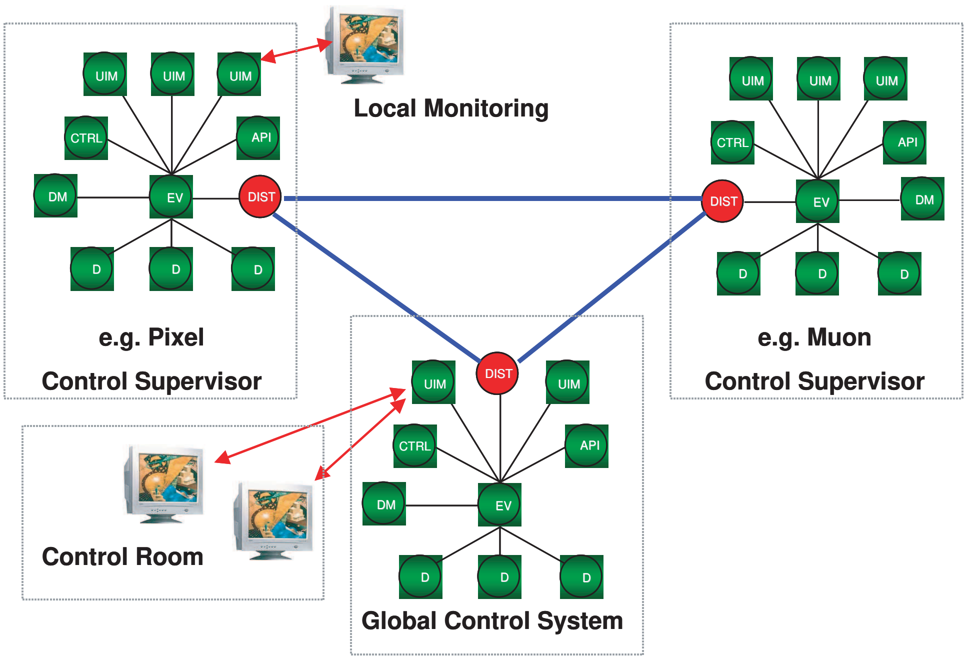

Figure 12.14: Distributed Detector Control.

the individual sub-detector groups. Again, subsystem specific critical protection hardware (i.e. the hardware that protects the detector systems from real damage such as over-current protection, interlocks, etc,) are the responsibility of each detector sub-system.

	Hardware Channels		**Software Channels**	
System	Fail safe	Normal	HV & LV	All Other
1.1 Magnets	19	6	8	1
1.2 Pixels	388	54	4740	0
1.3 RICH	19	212	772	171
1.4 EMCal	34	280	1440	28
1.5 Muons	4	203	1216	56460
1.6 Straws	0	169	744	0
1.7 Strips	238	21	504	0
1.8 Trigger	0	0	0	96086
1.9 DAQ	0	0	0	5000
1.10 Building, Racks	920	86	28	61

Table 12.2: Estimated numbers of monitored hardware parameters and software generated control data.

The overall detector control data quantity and transfer rates are driven by the quantity of detector control parameters needed to be monitored and the rate at which they need to be monitored. It is anticipated that a significant cost driver of the detector control system will be the need for failsafe mechanisms to monitor critical components of the detector.

An extensive initial accounting of detector sub-systems and the collision hall building of parameter types, quantities, sensor types, parameter readout frequencies and parameter "criticality" has been completed

Table 12.3.13 lists the current estimates of monitored DCS channels.

12.3.13.1 SCADA Selection

BTeV will need to select a supervisory, control and data acquisition (SCADA) system for its detector control system. Examples of such systems are PVSS II by ETM (selected by the four LHC experiments) and iFix by Intellution/General Electric (selected by CDF and MINOS).

A potential SCADA system will need to be evaluated on features like its database functionality, scalability (i.e., number of channels it can accommodate), the number of readout crates it can manage, its scripting capability, its alarm handling functionality, etc. The evaluation and ultimate selection of this kind of complex software will require detailed comparisons of its features and the creation of test installations.

The selection of the SCADA system is a high priority item the implementation and installation of such a system is highly correlated with overall detector installation/commissioning.

12.3.13.2 Sensor and Readout Module Selection

Significant cost reduction, ease of maintenance, and increased reliability of the detector control system will be achieved if common sensor and readout module solutions among the detector sub-systems can be identified.

To this end, the existing data from the initial polling of the sub-systems will be used by the detector controls group to identify sets of parameters common to multiple sub-systems amenable to measurement by a common sensor. Afterwards, specific sensors and their corresponding readout modules will be identified.

Existing sensor and readout module solutions at D0/CDF/MINOS and the four LHC experiments will be considered as possible solutions. For cost minimization, we will attempt to minimize the number of field bus technologies to be used for communication among the SCADA system and the readout modules.

Readout modules (and sensors) for all detector sub-systems must be consistent with the field bus choices.

12.3.13.3 Software development

After the SCADA software has been selected, a significant amount of supervisory and local control software (scripts and GUIs) will need to be written to implement the detector control system.

We envision that the full range of the software will be written by a combination of software engineers and physicists. We propose that the detector control liaison for a particular subsystem communicate to the appropriate software engineer, for a representative number of detector control channels, important design features like the parameters that need to be monitored, the various alarm conditions, actions to be taken, etc., and that the engineer create a prototype GUI or script.

A suitable training environment (e.g., individual or detector wide workshops and test-stands) will be set up to teach the liaison how to extend the GUIs/scripts to the full complement of detector control channels for a given detector sub-system. Templates, examples and test installations would be provided by the central detector control effort. The detector control liaison would then be responsible for any subsequent programming.

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)

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 12.15 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 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 are 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 collision hall.

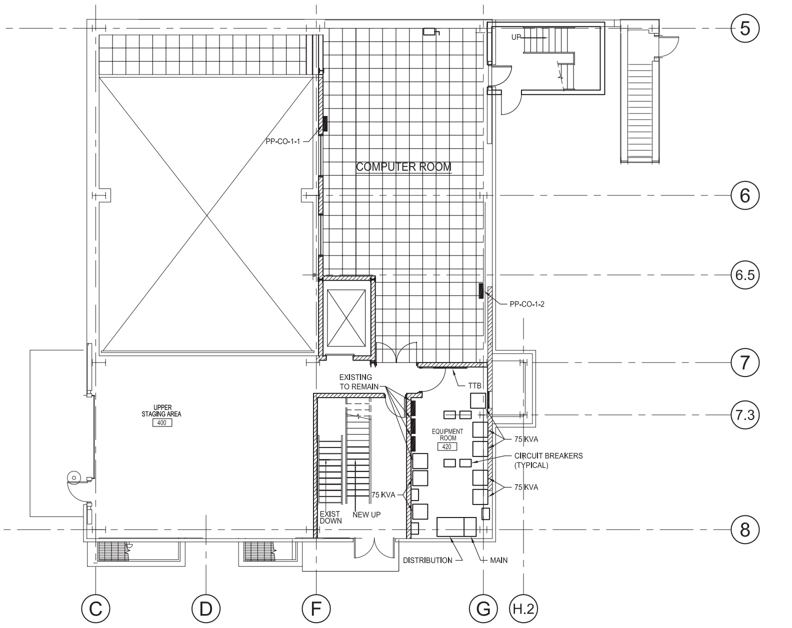

Figure 12.15: First floor Counting Room at C0.

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

Subsystem	**SubRack Estimate**	**Rack Estimate**
DAQ Electronics	48 DCB subracks	12
	192 L1Bs	6
	48 Data Switches	2
	30 PCs (detector manager,slow control)	2
	Management System, disk and database server	6
DAQ TOTAL		28

Table 12.3: Rack Estimates for WBS 1.9

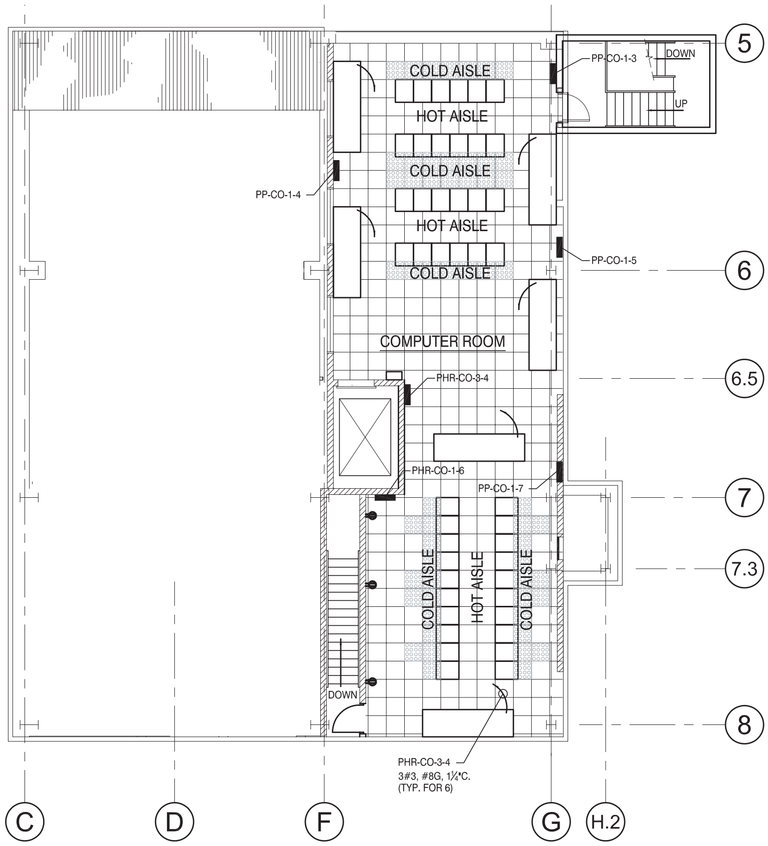

Figure 12.16: Third Floor Counting Room at C0.

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 12.16 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 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).

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 12.17). This board included three ASDQ integrated circuits and a medium density FPGA. 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.

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 standard CAT6 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 and IEEE1394 cables to determine if the signal characteristics and connector reliability are suitable for our application.

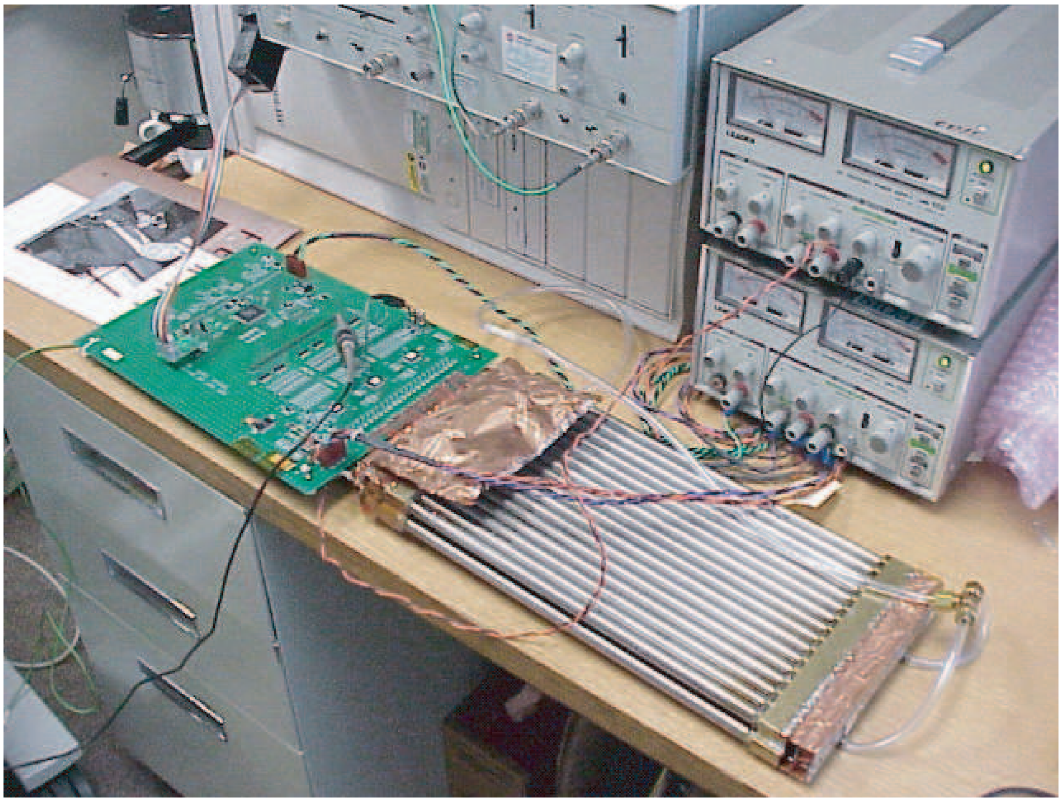

Figure 12.17: ASDQ Test Board with Prototype Muon Detector.

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 have been 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.

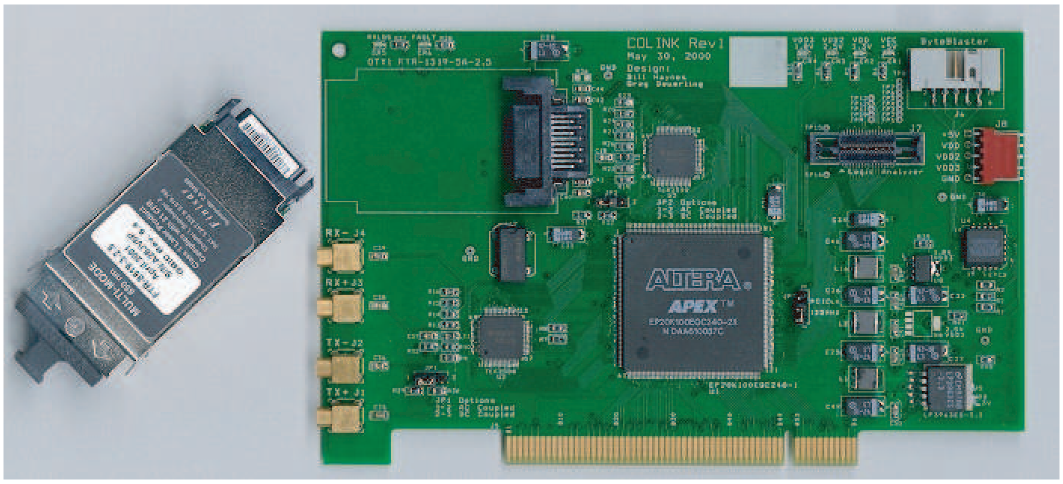

Figure 12.18: 2.5 Gbps PCI to Serial Link Card (Optical or Electrical Interface).

We also plan to test parallel optical transmitter and receivers as they become available. These parts provide the lowest overall cost per link.

Figure 12.19: Parallel Optical Links (12 channels per link @ 2.5 Gbps per channel).

A number of standard high-speed serial link protocols are currently in development (PCI Express, 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 to 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 verify that rate on a single link, multiplied by $\approx$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 Network

The readout and control network will consist of eight 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.

Each highway switch in the final system will handle 72 Gigabit connections. We plan to compare the performance of a single 72 port switch to that of a network built from several smaller switches (e.g., six 24 port switches). At current prices, the network based on the smaller switches may be significantly less expensive.

The L2/3 processors connect to 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.

12.4.6 Detector Control System

The slow control network will be Ethernet based, using commercial SCADA control software. 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 control application will be created.

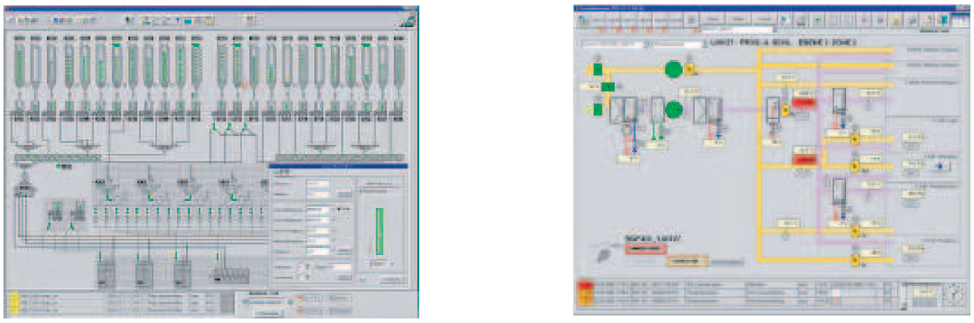

Figure 12.20: Slow Controls Interface Examples using PVSS.

12.4.7 Readout Software

The overall software design documents will be completed, with evaluation of software languages, development environments, and middleware that will be used through the construction project.

12.5 Production, Testing and Integration

This section describes the production, testing and quality assurance plans for the BTeV data acquisition and controls systems. The data acquisition system (DAQ) is responsible for transporting the data from the detector to the first level trigger system, 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. 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.

12.5.1 Read-Out Electronics

The BTeV read-out electronics consists of Data Combiner Boards, optical links, Level 1 Buffer modules and the Timing/Control system.

12.5.1.1 Data Combiner Boards

The Data Combiner Boards 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 are planned before production starts. 576 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 at 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 of these modules. A database will be included in the production and test plan so that a test record and shipping log of all modules will be accessible on the web.

12.5.1.2 Timing and Control System

The timing and control system distributes synchronous information such as the bunch crossing signal to the front-end electronics modules. It provides control signals to ensure 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 and Ohio State. One prototype step is planned before production. 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 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 will be commercially developed. One prototype step is 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 will be designed to receive data from up to 24 Data Combiners 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 are planned before production starts. 192 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 ensure 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 ensure 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 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 ensure compatibility and for software development purposes two test labs will be set-up at Fermilab and at Ohio State. Only hardware and software modules that pass these compatibility tests will be utilized in the experiment.

12.5.4 Databases

The BTeV online system will use databases to store configuration information, to archive environmental conditions and run parameters, as a 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 depends to a great extent 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 L1 Buffer modules and the Level 2/3 trigger farm. The fabric will be constructed of commercial network switches using Gigabit Ethernet technology. Before purchasing the production units we allow 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 Feynman Computing Center 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 Standard precautions (e.g. avoidance of electrostatic discharges) will be required for the transport of electronics modules.

Personnel Required Most of the readout and controls equipment can be maneuvered by hand.

Time Required A few daya will be needed for the transportation of the electronics modules, PC's and other equipment. Transportation of equipment from Ohio State will require two additional 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, *i.e.* is no longer needed for the insertion of detector components, we will install the subracks that house the Data Combiners.

For each component, cables need to be installed to connect the front end modules to the Data Combiners - about 5,000 cables in total. The connection to the counting room is provided by 384 optical fiber bundles (each with 12 fibers). Before we can run these bundles we will install innerducts 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 150 cables will connect Data Combiners 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, control processors and networking. 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. Installation steps include setting up the control room furniture, the network infrastructure as well as the computer/operator consoles. The detector control system requires equipment to be installed in different locations. Most of the monitoring and control system in the detector hall will be installed by the sub-detector groups. Network (Cat5) 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 computers (Detector Manager and Control Manager/Supervisor) still needs to be defined. 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 other than equipment to access second level racks and overhead cable trays.

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 that these modules and cables do not 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 5000 readout cables (front end to DCB)	$\approx$1000 hours
Install 384 optical fiber bundles	$\approx$300 hours
Install 3000 network cables	$\approx$500 hours
Install 150 timing cables	$\approx$150 hours
Install $\approx$576 DCB modules	$\approx$10 hours
Install $\approx$192 L1B modules	$\approx$10 hours
Install detector control cables	$\approx$150 hours
Install PCs and workstations ($\approx$30)	$\approx$100 hours

12.6.4 Testing at C0

1. Infrastructure Tests
   (a) Utilities - Leak test cooling water systems
   (b) Safety Systems - Evaluate 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.