Fermi National Accelerator Laboratory

FERMILAB-Conf-98/348-E
CDF

ATM Based Event Building and PC Based Level 3 Trigger at CDF

J. Fromm et al.
For the CDF Collaboration

Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

December 1998

Published Proceedings of *International Conference on Computing in High Energy Physics (CHEP '98)*, Chicago, Illinois, August 31 - September 4, 1998

ATM Based Event Building and PC Based Level 3 Trigger at CDF

J. Fromm, D. Holmgren, R. Kennedy, J. Patrick, D. Petravick, R. Rechenmacher, G.P. Yeh

Fermi National Accelerator Laboratory

G. Bauer, K. Kelley, P. Ngan, C. Paus, T. Shah, P. Sphicas, K. Sumorok, S. Tether, J. Tseng, D. Vučinić

Massachusetts Institute of Technology

B. Kilminster, K. McFarland, K. Tollefson

University of Rochester

Abstract. Baseline event building and Level 3 trigger processing at the CDF experiment at Fermilab is specified to process about 300 events/s with an average event size of about 150 KB. First, the event must be assembled from fragments originating from approximately 16 readout sources, with individual average fragment sizes ranging from 12 to 16 KB. Then the Level 3 processor-based trigger performs reconstruction and filtering tasks requiring in excess of 45000 MIPS of CPU power on the assembled events. We present a distributed, scalable architecture using commodity components: VME-based CPU modules for the readout, an ATM switch for event buildling, and Pentium-based PC's running Linux for event processing. Pentium-based PC's are also used for event distribution throughout and collection from the Level 3 processors via multiple 100 Mbps Ethernets. We report on preliminary studies conducted at CDF with a small-scale prototype. This architecture is also a possible candidate for the CMS experiment at LHC.

INTRODUCTION

The Collider Detector at Fermilab [1] (CDF) is a general purpose particle detector which has taken over 100 pb^{-1} of data at the Fermilab Tevatron since 1987 and is scheduled to take data again in 2000, accumulating well over 10 pb^{-1} per week. To take advantage of the high luminosity of the upgraded Tevatron, a three-level trigger

hierarchy has been preserved from previous data runs, where each succeeding level filters events on the basis of increasingly refined reconstructions of objects within the event. In this way the first two trigger levels will reduce the event rate from 7.6 million events/s to about 300 events/s (up to 1000 events/s). The Level 3 trigger, implemented as a "farm" of computers analyzing the whole event record, will further reduce that rate to roughly 30 events/s (up to 75 events/s) which can then be recorded for offline analysis. The amount of computing resources brought to bear on Level 3 processing is specified to be such that the processing time per event is on the order of seconds, rather than hundredths of seconds or less which characterize the first two trigger decisions.

This article concerns the development of the subsystems for Run II which assemble the event for Level 3 processing (the "event builder") and then deliver the whole event to a single computer for analysis. It must therefore assemble and deliver events at the specified input rate, 300 events/s, though it is also desirable that it be able to operate up to the Level 2 limit at 1000 events/s. The approximately 16 distinct event fragments will each contain on average 12 to 16 KB of data, with the total being around 150 to 250 KB per event. The aggregate data throughput of the system must therefore be at least 44 MB/s, with up to 245 MB/s desirable.

Such high throughput is readily available with commercial network technology. The Run II event builder will be based on an ATM switch; such technology is also being investigated for use in the CMS experiment, where the number of inputs and outputs would be on the order of 500 each [2]. The use of inexpensive Pentium-based personal computers, organized into "subfarms" hanging off of event builder output ports, is also being investigated for the purpose of satisfying the sizable computing requirements for Level 3.

EVENT BUILDER

The event builder test system is shown schematically in Figure 1. Event data enters the system through the Scanner CPU's (SCPU's) and is sent through the event network, which in this sytem is the ATM switch, to the Receivers (RECV's). The flow of data is controlled by the Scanner Manager, which communicates with the Scanners and Receivers via the separate command network.

In the current test system a FORE Systems ASX-1000 non-blocking ATM switch [3] is used as the event network. "Non-blocking" refers to the fact that the switch's internal bandwidth accomodates the maximum input bandwidth. The output ports are buffered to the depth of 13 K cells in order to accomodate temporary overloads of the bandwidth of a single output link. The switch is currently equipped with sixteen 155 Mbps input/output ports but is expandable up to 64. These ports are connected to PowerPC-based VME CPU's running VxWorks 5.3. In general, 200 MHz Motorola MVME2603's are used as Scanners, and 66 MHz MVME1603's emulate Receivers, though these roles can be changed simply by downloading different software onto the individual nodes. The ATM interfaces are

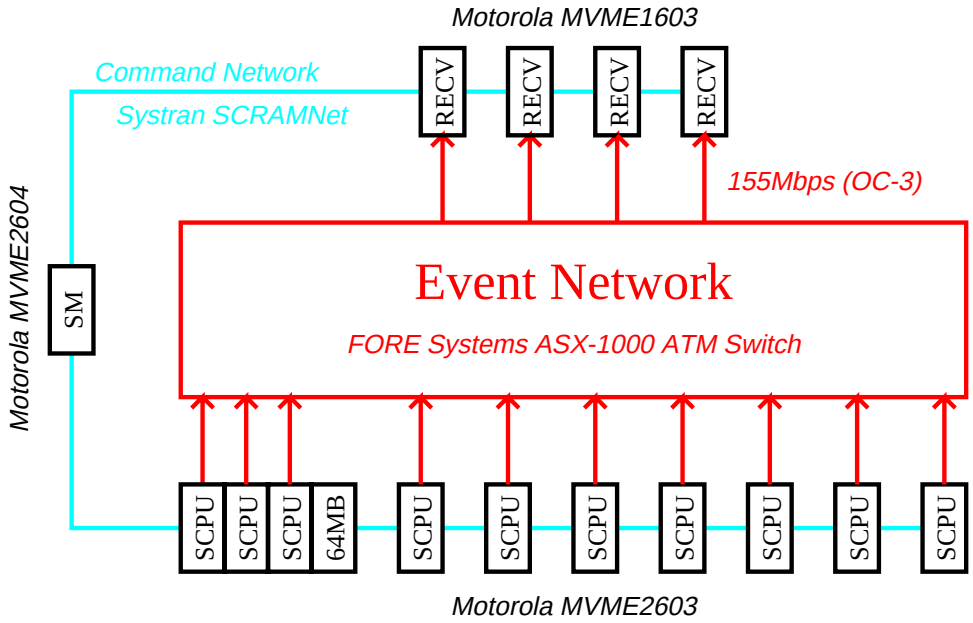

FIGURE 1. The event builder test system. The Scanner CPU's (SCPU's) and Receivers (RECV's) are Motorola MVME2603 and MVME1603 VME-based computers, and are connected to the Fore Systems ASX-1000 ATM switch with 155 Mbps (OC-3) optical links. Three of the MVME2603's reside in a single crate with a 64 MB VME memory. The Scanner Manager (SM) is a Motorola MVME2604 which communicates with the other computers via a separate command network implemented by a Systran SCRAMNet ring.

Interphase 4515 PMC-ATM adapters with 1 MB on-adapter RAM. The driver has been developed in-house, and implements only bare AAL5 without any higher level protocols. Since AAL5 by itself does not guarantee data delivery, system tests look not only at throughputs but also at losses.

The command network is a Systran SCRAMNet ring of VME reflective memories. This network provides for the fast, reliable transfer of small messages. The Scanner Manager is in general a 200 MHz MVME2604, though in some older tests it is a 66 MHz Radstone RS603; again, a different Scanner Manager can be used simply by downloading the software onto a different computer. In the current system, messages are received by polling the reflective memories for new ones.

Event Network Tests

The most basic tests involving the ATM components are those in which N_{send} computers perform uncoordinated rapid-fire packet transmissions to each of N_{recv} receiving computers. The number of packets moving through the switch at a given time is therefore $N_{send} \times N_{recv}$. The driver is called directly, and the control network is ignored. These tests therefore reflect the best possible data throughput

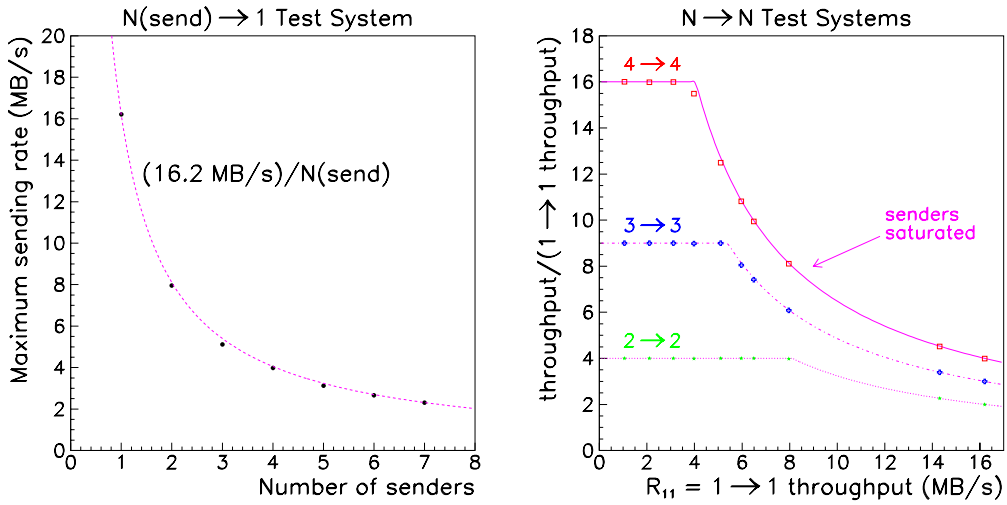

FIGURE 2. Left: maximum sending rate for N_{send} equivalent senders for one to seven senders to one receiver. The curve is the theoretical maximum, above which cells will be lost. The points occasionally lie below the curve because of the coarse-grained rate limit control. Right: relative throughput (to R_{11}, the $1 \rightarrow 1$ data throughput) vs. R_{11}, for $N_{send} = N_{recv}$.

performance.

One obvious issue in this setup with multiple senders and receivers is that if several senders send data to a single receiver faster than it can be received, the ATM switch will simply drop the overflowing cells. However, the ATM adapter can be instructed to restrict its own sending rate to a given receiver by setting a hardware prescale counter. If the maximum reception rate is v_{max}, which has been measured by sending from one computer to another, then one naively expects that the maximum sending rate from equivalent senders will be v_{max}/N_{send}, above which cells will be lost, and indeed this is seen to be the case in Figure 2(left).

In the above "rate division" scheme, a sender's unused bandwidth can then be directed towards other receivers in the system. The total throughput of the system should therefore scale as $N_{send} \times N_{recv}$. This scaling behavior is shown for $N_{send} = N_{recv}$ by the plateaus in Figure 2(right). The falling relative throughput for $R_{11} > v_{max}/N_{rec}$, where R_{11} is data throughput for an $N_{send} = N_{recv} = 1$ system (essentially the rate limit), is due to having saturated the senders' ATM links. In these tests, no cells were lost at any set rate limit. It was also confirmed that the same data was received as was sent.

Command Network Tests

In order to build an event using the "rate division" method, the Scanner Manager broadcasts one SEND_EVENT message to all the Scanners; each Scanner then sends

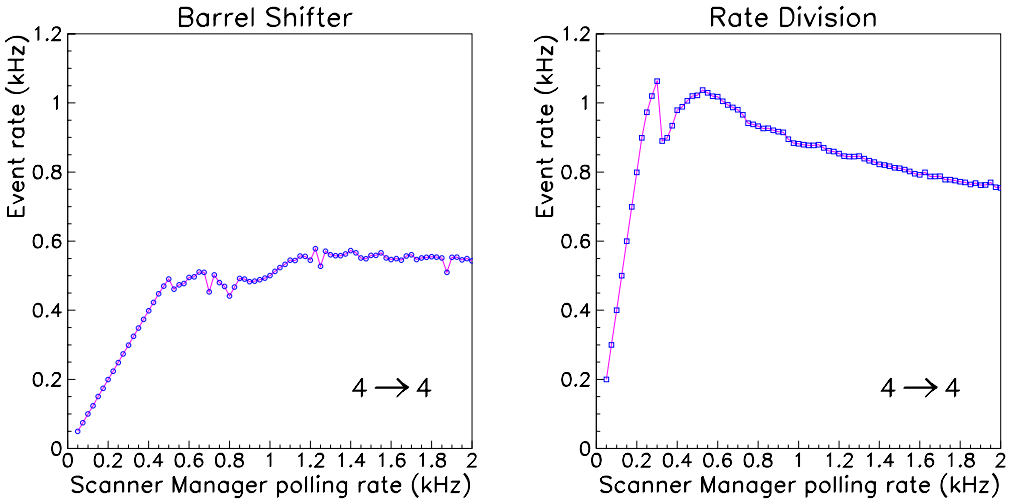

FIGURE 3. "Event" (messages only) throughput as a function of the rate at which the Scanner Manager polls for messages, for "barrel shifter" (left) and "rate division" (right) control methods. The Scanner Manager is a Radstone RS603 in these tests.

at its allocated rate, after which it sends its acknowledgement back to the Scanner Manager. Multiple events are built concurrently as in the event network tests. This "rate division" method is in contrast with the "barrel shifter" method, which in all forms requires each Scanner to be informed one at a time via the command network when it is to send its data at the full rate. At CDF, this latter method has been implemented with the Scanner Manager sending the individual SEND_EVENT commands, interleaving events being sent to different Receivers. Thus, the "barrel shifter" method incurs substantial control overhead from generating and passing these messages.

The "rate division" and "barrel shifter" methods can be compared by running the event builder system without actually passing any data through the event network. In this case, an "event" is simply a complete round of control messages. These tests therefore measure the best possible (non-empty) event throughputs for the two methods. The results for a $4 \rightarrow 4$ system are shown in Figure 3, where the Scanner Manager is a Radstone RS603. The "barrel shifter" plateaus around 450 events/s, which satisfies the Run II target at 300 events/s, but not the 1000 events/s desired. The "rate division" method, on the other hand, reaches 1000 events/s, albeit without sending any actual data. However, direct measurements show that the CPU is quickly saturated in the "rate division" test; tests with the MVME2604 as Scanner Manager have shown event rates well in excess of 1000 events/s even when sending data through the event network.

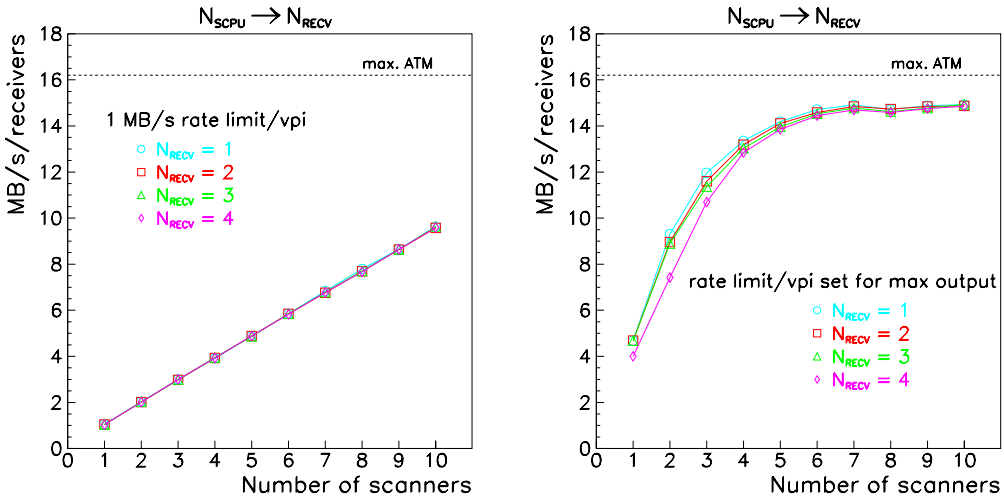

FIGURE 4. Data throughput per receiver vs. the number of Scanners, for rate limits set at 1 MB/s (left) and v_{max}/N_{SCPU} (right). In the 1 MB/s case, all four curves lie on top of one another.

System Tests

The concern in using the "rate division" method is that data may be lost due to collisions at the switch outputs. It is clear from the event network tests, however, that data loss can be eliminated. It is also interesting to see if the scaling behavior evident in the event network tests can be observed in systems involving the command network.

The present tests attempt to simulate Run II conditions by allowing the event fragments to vary in size according to a truncated gaussian distribution with a mean of 16 KB and standard deviation of 4 KB, with limits at 8 KB and 32000 bytes. This level of variation was typical for tracking detectors in Run I. It is expected that in real data-taking conditions the fragment sizes of a given event would be somewhat correlated in size, but the variations in these tests are uncorrelated.

Figure 4(left) shows the data throughput per Receiver as a function of the number of Scanners in the system. The rate limitation has been set to 1 MB/s, which is appropriate for a system with 16 Scanners. No data loss is observed. The linearity in N_{RECV} is demonstrated by the fact that all four curves for the different values of N_{RECV} lie on top of one another. The throughput also nearly scales with the number of Scanners, with a slight degradation due to the random fragment size variations. The transfer time of an event is determined by its largest fragment, creating inevitable idle times in the system; if the fragments are fixed in size, the slowdown disappears.

If instead the rate limitation on each of the Scanners is set so that the Receiver bandwidths are nearly saturated, $v_{limit} = v_{max}/N_{SCPU}$, the data throughput

plateaus around 14.5 MB/s as shown in Figure 4(right). The average data throughput does not reach the maximum link speed, 16.2 MB/s, again because of the idle times from the random fragment size variation; if the fragment sizes are fixed to large values, the data throughput can reach the full link speed. Again, no data loss is observed in any of these tests. It is evident that in such a controlled network environment, the overheads of high-level protocols to guarantee packet delivery are not needed.

VME Readout Tests

In the above tests, the Scanners simply allocate blocks of main memory and send them over the event network. In the Run II system, the Scanners will be reading data out of VME modules called VME Readout Boards (VRB's). Since both the VME and ATM operations utilize a single PCI bus in the Scanners, there is concern that the simultaneous transfers will reduce the event builder performance. Initial tests, with one Scanner reading a VRB prototype, have indicated that MVME2603's equipped with the Universe II PCI-VME bridge do not degrade the throughput of the event builder. Future tests with multiple VRB's in each crate are planned.

LEVEL 3

The use of networked farms of relatively inexpensive workstations has been standard for offline processing of events for several years, and it is natural for this evolution to continue by utilizing the increasingly powerful personal computers available on the consumer market instead of high-end workstations [4]. It is also natural that these farms should begin to penetrate the world of online computing as well. Such a solution has been proposed to provide the CPU power required for Level 3 trigger processing.

The organization of the farm is shown in Figure 5. The farm is actually organized as several "subfarms," each subfarm hanging off of one event builder output port. Sixteen such subfarms are envisioned. A Receiver node interfaces with the ATM switch and forwards the data to Processor nodes over an inexpensive intranet such as Fast Ethernet. The Processor nodes execute the Level 3 trigger algorithm and forward the passing events to the data logging services.

All the personal computers in the prototype run the Fermilab-supported Linux operating system, based on the Red Hat 5.0 distribution, which provides a generic UNIX environment [5]. The control software is a straightforward port of the Run Ib control software, which ran under IRIX. The choice of Linux has also enabled optimizations which have proven useful in time-critical online applications [6].

A prototype system has been built with one Receiver node and four Processor nodes in order to test its ability to sink the entire event builder output. Each node is connected to a Fast Ethernet switch; since the maximum bandwidth available at each Ethernet port is about 12 MB/s, two ports are connected to the Receiver

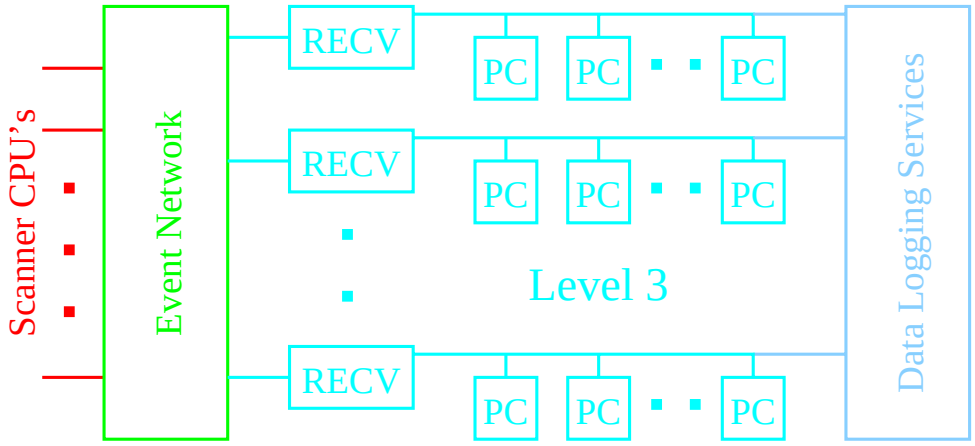

FIGURE 5. Organization of a Level 3 trigger processing farm into multiple subfarms. Each subfarm is connected to the event builder through a dedicated Receiver node.

node. The Processor nodes, purchased in late 1997, contain two 200 MHz Pentium Pro CPU's. The Receiver node started out as a single 200 MHz Pentium Pro, but has been upgraded to improve its I/O performance as shown in Figure 6(left) in tests of the intranet throughput. The latest upgrade, to a 350 MHz Pentium-II system with the BX chipset supporting an improved 100 MHz memory bus, exceeds the throughput requirement for 1000 events/s at 250 KB/event when sending data to three or more Processor nodes (six or more nodes are expected in an actual subfarm), and comfortably exceeds the specified throughput at 300 events/s and 150 KB/event.

Figure 6(right) shows the throughput of the combined event builder/Level 3 prototype system. The 100 MHz bus provides sufficient throughput for use in the real system. It should be noted that the market has continued to push even further increases in computing and I/O power for similar prices per node.

The largest combined event builder/Level 3 system operated so far uses all the MVME2603's and MVME1603's as Scanners sending to the one Receiver node. With 14 Scanners, each sending 16 KB per event (with the 4 KB variation), this system approaches the scale of what is required at CDF for Run II. The Receiver node for this test used SDRAM memory on a 66 MHz bus and sent its data to three Processor nodes, one of which forwarded its data to another ("Output") node in order to simulate the somewhat more complicated network traffic of the real system. With each Scanner sending 1 MB/s, the event rate was 58 events/s at 224 KB average per event. With sixteen Level 3 subfarms this rate translates to over 900 events/s.

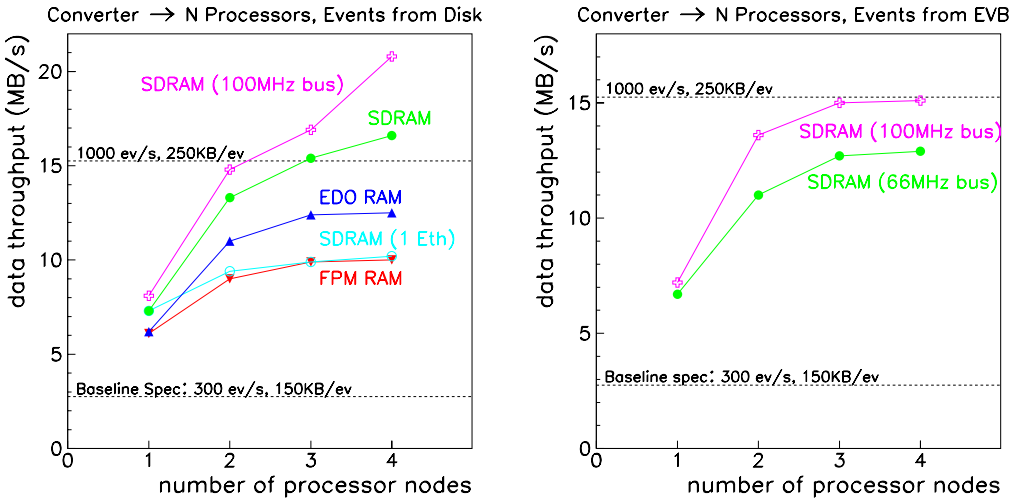

FIGURE 6. Left: data throughput of the Receiver node when sending to multiple Processor nodes. The data is read from disk and stored entirely in cache memory. The different curves reflect different memory and bus upgrades to the Receiver node. A comparison for 66 MHz bus SDRAM systems between single and dual Ethernet throughputs is also shown. Right: data throughput when the data is received from the event builder test system, for 66 MHz and 100 MHz bus SDRAM systems.

CONCLUSION

This article has reported on results from event builder and Level 3 test systems utilizing hardware that is widely available on the market, such as ATM switches and adapters, VME computers, and personal computers. Customization has been confined exclusively to software. Tests of the event builder prototype show the expected behavior regarding rate limitations and scaling, all without cell loss in spite of not using a high-level protocol. The event throughput is well in excess of the 3 MB/s per Receiver required to satisfy the baseline specification. Sustained data throughputs of 14.5 MB/s per Receiver have been observed with fragment size variations similar to that expected to be seen in data. Initial tests with Level 3 prototype hardware indicate that Pentium-based personal computers can also comfortably satisfy the I/O demands of the baseline specification for Run II. Further improvements and upgrades are still possible, and prototype systems are expanding for more thorough tests.

REFERENCES

1. F. Abe *et al.* (CDF Collaboration), *Nucl. Instrum. Methods* A **271**, 387 (1988); F. Abe *et al.* (CDF Collaboration), *Phys. Rev.* D **50**, 2966 (1994); The CDF II Collabora-

tion, *The CDF II Detector: Technical Design Report*, FERMILAB-PUB-96/390-E, October, 1996.
2. The CMS Collaboration, *Technical Proposal*, CERN/LHCC 94-38 (LHCC/P1), December 15, 1994.
3. FORE Systems, ForeRunner *ATM Switch Architecture*, April, 1996.
4. S. Wolbers *et al.*, "Processing Farms Plans for CDF and D0 for Run II," CHEP talk 77.
5. D. Skow *et al.*, "Linux Support at Fermilab," CHEP talk 220.
6. D. Holmgren *et al.*, "Application of PC's and Linux to the CDF Run II Level 3 Trigger," CHEP talk 183.