CDF/DOC/LEVEL3/CDFR/5516
December 19, 2000

Letter of Intend for Run IIb:

Upgrade of the Event-Builder and Level-3 Trigger PC Farm

K. Anikeev, G. Bauer, A. Bolshov, J. Friedman, I. Furic, B. Iyutin, A. Korn,
I. Kravchenko, N. Leonardo, M. Mulhearn, S. Rakitin, Ch. Paus,
L. Rosenson, K. Sumorok, S. Tether, J. Tseng, F. Wuerthwein
Massachusetts Institute of Technology

Abstract

An upgrade of the CDF Event-Builder and the Level-3 Trigger PC farm for the Tevatron Run IIb is proposed. Several upgrade scenarios and their implications are discussed.

1 Introduction

During the Run IIa data taking period, the Tevatron is expected to deliver 2 fb^{-1} to the CDF experiment, a factor of 20 more integrated luminosity than in Run I. In the Run IIb period further data will be collected to come up with a total data sample corresponding to 15 fb^{-1}.

For the Technical Design Report of the CDF II detector, the Level-2 trigger provides an accept rate of 300 Hz. This design goal is based on a peak luminosity of $2 \cdot 10^{32}\text{cm}^{-2}\text{s}^{-1}$. According to the Tevatron upgrade plans a maximum peak luminosity of roughly $5 \cdot 10^{32}\text{cm}^{-2}\text{s}^{-1}$ is reached by 2006. Naive scaling predicts a Level-2 accept rate of at least 750 Hz.

However, the Event-Builder data throughput depends not only on the Level-2 accept rate but also on the event size which is closely related to the detector occupancy. The CDF silicon detectors provide the single largest part of the event data and their occupancies depend on the threshold requirements with a given signal to noise ratio. Ideally the occupancy is smaller than 5 percent which corresponds to 750,000 channels $\times$ 0.05 $\times$ 2 Bytes = 75 kBytes per event. Due to radiation damage the signal to noise ratio decreases during the Run II data taking and depending on the threshold tuning the occupancy might change. To allow for useful pattern recognition the occupancy cannot get larger than 10 percent corresponding to 150 kBytes per event. This increases the total event size of about 250 kBytes to 325 kBytes. To account for these type of effects a factor of two on top of the original 250 kBytes event size is applied as a safety margin. This implies a bandwidth of 500 kBytes $\cdot$ 750 Hz = 366 MBytes/s for the Event-Builder for Run IIb. The completion of the upgrade is proposed for 2005 to comply with the expected Tevatron luminosity.

The maximum bandwidth theoretically achievable with the existing Run II system is 15 channels $\times$16 MBytes/s = 240 MBytes/s. In practice about 60 percent of this theoretical limit has been acchieved in benchmarking tests simulating data sizes corresponding to the expected detector occupancy. It is likely that the performance in such tests further improves to about 80 percent of the theoretical limit when tuning the system. The performance when processing real data depends crucially on the load balancing among the various ATM switch inputs.

To be conservative three different Run IIb options are described ranging from simple maintenance of the existing system, to upgrading it to a sustainable bandwidth of at least 400 MBytes/s.

Estimating the CPU power needs in the Level-3 Trigger PC farm is more uncertain than estimating the bandwidth needs for the Event-Builder. For the Level-3 PC farm the achievable performance increase per PC is discussed.

2 Limitations of the Existing System

When analyzing the present setup of the CDF Event-Builder and Level-3 PC farm two different types of limitations are identified. There are throughput limitations and there are processing power limitations. Processing power limitations are trivial: computing power is made available by purchasing CPUs while the Level-3 trigger algorithms in the end have to be adjusted to the available CPU.

Four possible throughput-type limitations are distinuguished

- data throughput at the Event-Builder input (VRB crate)
- Event-Builder control network performance
- data throughput at the PC farm entry which are the converter nodes
- data throughput internally in the processing farm; fast ethernet
- data throughput at the Level-3 output to the Consumer Server and Data Logger

The Event-Builder limitations for the data throughput are about 240 MBytes/s given by the 15 ATM input connections (OC3 standard). With event sizes of 250 kBytes the control network therefore has to function properly up to 1 kHz, which has been shown.

Both, 240 MBytes/s data throughput and 1 kHz event rate, are upper limitations and the real performance will be less and further depend on more parameters like the load balancing per ATM input, the fluctuations of the event fragment length and so on. With the present experience a sustained rate of 600-800 Hz of 250 kBytes events should be achievable without further extensions.

The input into the Level-3 PC farm on the other hand is less problematic since there is one more output OC3 connection a total of 16. The fact that the converter node cointains many PCI cards – SCRAMNet, ATM, quadrupel fast ethernet – has so far not caused any problem. The two fast ethernet output connections have a sustained rate through the converter node of more than 20 MBytes/s while the OC3 input connection provides 16 MBytes/s maximum input.

For the subfarm internal network the maximum rate of 16 MBytes/s is subdivided to at least 8 PCs which makes this a comfortable rate for the 24 port fast ethernet switch which is used.

Without rejection one output node with 2 fast ethernet connections to the subfarm and further 2 fast ethernet connections to the switch which connects to the Consumer Server and Data Logger is able to handle one complete subfarm. Assuming a reasonable event rejection rate an output node could handle the input of more then 2 subfarms on its own. No limitations are expected in that layer. Extension to include more PCs is trivial since fast ethernet switches up to 64 ports exist on the market. Concerning the CS/DL side the Level-3 PC farm software has been designed to communicate with a farm implementation of this device. The extension is trivial and hardware can be bought off-the-shelf.

3 Event-Builder Upgrade Scenarios

In this section three upgrade scenarios are described, one of which is simply to maintain the status quo. For each upgrade scenario, both the human resource needs, as well as the equipment costs are discussed.

3.1 Maintaining the Status Quo

As discussed above, the existing system is likely to sustain a Level-2 accept rate of $600-800$ Hz for an event size of 250 kBytes. To maintain this level of performance until 2008 requires to plan for hardware failures. Some of the hardware as well as some software will be obsolete, and thus no longer commercially available.

In the following the hardware of the Event-Builder is listed, and issues of obsolescence and potential replacement needs are discussed.

- The ATM cards in the SCPU crates are obsolete already. Enough spares have been purchased to be prepared for a long run.

- ATM cards in the Linux PCs are currently going out of production. This process is being watch closley and sufficient spares will be purchased.

- The ASX 1000 ATM switch has a huge installed base and vendor, Marconi, guarantees that they will back this product indefinitely.

- In the previous run SCRAMNet used for the control network was very stable after the initial shakedone. Some more spares have to be bought to extend the system into the Run IIb.

- The scanner CPUs are commercial MVME 2600 single board computers with a large installed base. They will be available until the beginning of Run IIb and depending on the failure rate more modules have to be purchased.

The main worry of the software is the ATM driver on the Linux PCs. Already now the operating system on the converter nodes is running an older version to be able to use the driver which had to be significantly changed from its original version. It is expected that the operating system, on the farm is updated at least once or twice and it might be necessary to adapt the driver to the newest Linux release.

Between now and 2008 roughly $100k are needed for spare parts, and at least one FTE year on software development to maintain the system at 600-800 Hz Level-2 accept rate at 250 kBytes event size.

3.2 Upgrading the ATM Network

The simplest upgrade path for Run IIb is to switch from the present OC3 to OC12 ATM links and buy something like the new switch ASX 4000 from Marconi. Switch ports based on the OC12 standard provide 4 times the bandwidth of the corresponding OC3 ports corresponding to a theoretical limit of close to 1000 MBytes/s, $60-80\%$ of which ought to be achievable in practice if there were no other bottlenecks.

The system limitations are then likely to be not the ATM switch but instead:

- VRB readout bandwidth via VME: 1000 MBytes/s spread evenly over 15 VME crates results in a 67 MBytes/s bandwidth requirement within each VME crate. This can not be achieved with present VME hardware and protocols. Benchmarks of the D64 read/write bandwidth for a MVME2600 came close to 50 MBytes/s. The useful bandwidth is probably as low as 30-40 MBytes/s resulting in 450-600 MBytes/s sustained throughput rate for the Event-Builder.

- The present system is limited to about 1 kHz Level-2 accept rate due to the maximum message passing rate over the SCRAMNet ring. To go beyond this the message passing protocol needs to be upgraded.

- The Converter nodes presently feed their subfarms via two fast ethernet links on a 4 port card. The output bandwidth can be doubled by making use of all four ports without additional hardware. The total bandwidth through the converter nodes is then up to 48 MBytes/s which matches the performance of the ATM connection.

A sustained Event-Builder bandwidth of $450-600$ MBytes/s is achievable in this scenario. Past experience has shwon that in the Event-Builder the usage of the ATM switch is sufficiently different from its commercial use in the telephony industry. This implied that commercially available drivers have to be modified in a non trivial way. This requires significant development which is expected to be the largest software development effort in this scenario.

The cost of the upgrade including the new switch and all necessary network interface cards is estimated to be \$500k but should be re-evaluated in the future.

The total software development for the two ATM drivers, plus changes to the SCRAMNet message passing protocol are estimated to amount to $2-3$ FTE years.

3.3 Gigabit Ethernet Alternative

In principle the ATM hardware could be abandoned and gigabit ethernet using TCP/IP could be used. The performance increase is in the order of a factor of 7-8 at the switch depending somewhat on the development of this technology in the near future.

However as shown above the Event-Builder bandwidth is limited by the bandwidth inside the DAQ VME crates. To overcome this the number of DAQ crates would have to be increased, or the crates would have to be upgraded with faster backplane technology like RACEway, CompactPCI or similar.

This upgrade seems not to be cost-effective given that the expected bandwidth needs are not in excess of the $450-600$ MBytes/s. A gigabit ethernet Event-Builder thus provides the same bandwidth as the ATM upgrade but on the other hand requires significant amount of software development as large parts of the Event-Builder software would have to be re-written. This option will only be considered if the collaboration decided for some reason that the $450-600$ MBytes/s is an insufficient data transfer rate.

4 Level-3 PC Farm Upgrade

At present it is impossible to provide reasonable estimates of the CPU needs for the Level-3 PC farm over the course of Run IIb. Instead it is pointed out that the performance/price ratio has roughly doubled every 1.5 years in the past. A replacement of the obsolete PCs in 2004 will therefore probabaly quadruple the present processing power. A total budget of about \$500k is proposed for a Level-3 PC farm upgrade.

5 Conclusions

We herby express our interest in upgrading the Event-Builder bandwidth by a factor of about 3-4, and the Level-3 CPU performance by a factor of about 4 for Run IIb. This requires about \$1 million in hardware costs, and roughly 3-4 FTE years of software development. We propose to complete the upgrade by 2005.