The CMS Event Builder Demonstrators based on GigaEthernet Switched Network

M.Bellato[1], *L. Berti*[2], *M. Gulmini*[2], *G. Maron*[2], *N. Toniolo*[2], *G. Vedovato*[2], *S. Ventura*[1], *X. Yang*[2]

[1] INFN – Sezione di Padova, Via Marzolo 8, 35100 Padova, Italy
[2] INFN – Laboratorio di Legnaro, Via Romea 4, 35020 Legnaro, Italy

Abstract

The event builder system is one of the most challenging items of the data acquisition of the CMS experiment. It will require a very efficient 1000 nodes switched network with a bandwidth in the order of Tbps and the ability to assemble event fragments at rate closed to 100 kHz. Several switch technologies and different architectures are currently being evaluated. This paper describes the demonstrator that has been set up to study a 4x4 event builder based on a gigaethernet switched network

keywords Event builder, Switched Network, GigaEthernet

1. Introduction

The LHC 40 MHz collision rate will produce in the CMS[1] detector 10^9 interaction per second, corresponding to 100 TByte/s of information. This input rate is reduced, by the level 1 trigger system, to 100 kHz before being handled by the DAQ system. The system is based on a Tbps switched event builder network connecting 512 readout units (RU's) to 512 builder units (BU's), each of which connected to a computer farm. The RU's read out data from the detector elements at a first-level trigger rate up to 100 kHz and buffer the event fragments for up to 1 s. The expected average event size is about 1 MByte with a corresponding fragment event size of 2 kByte.

The computing power provided by the on-line farms (software trigger) should be sufficient to insure the lowering down of the final accepted event rate to about 100 Hz. According to the link speed available, two different software trigger strategies have been proposed[2]. In the case of 1 Gbps links a "multi-step" event building is used. The first step of the filter, that occurs at the full rate of 100 kHz, moves only the 25% of the event data from the RU's to a BU with an expected rejection factor of 10. The remaining 75% of the data of the selected events (that now occur at 10 kHz) are then moved to the second step of the event builder. This mechanism reduces the required event builder throughput of a factor 3 at the cost of control complexity and increased latency. If 5 or 10 Gbit/s links will be available, the event builder could be performed in "single step", where all the event information will be moved from RU's to BU's.

As already underlined[3], the event building traffic is highly systematic, as multiple sources (RU's) compete for the same destination (BU) causing congestion on the switch link output buffers. The effect of this congestion depends on the switching technology used, varying from a reduced throughput to a loss of data.

The whole event flow control from RU's to BU's is handled by an event manager which is also required to broadcast the first level trigger information to all RU's. Different protocols have been proposed and tested. Appropriate control traffic schemes have been included in the protocols to avoid output link congestion, like barrel shifter, rate division, credit based protocol, etc. Two types of information are thus being carried by the event builder network: data and control. The network can be then logically split into event builder data network and event builder control network. According to the technology used, the two networks can be implemented in the same physical network or two different networks can be used.

Small scale prototypes (demonstrators) are thus being developed to evaluate the network technologies and to study the more appropriate architectures.

2. The Event Builder Demonstrator

A 4x4 gigaethernet (GE) event builder demonstrator has been set up. The demonstrator configuration is shown in fig.1. A GE switch (Intel 6000) connects 4 RU's to 4 BU's. 7x7 configuration was also possible, but problems found in the switch limited the test to 4x4. An event manager (EVM) controls the system, providing also the broadcast of the level 1 trigger information. The event builder protocol used[2] is presented in fig. 2

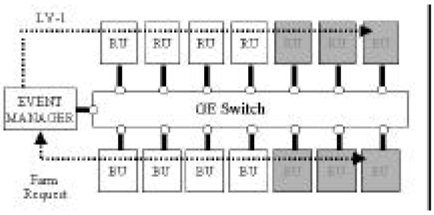


Figure 1

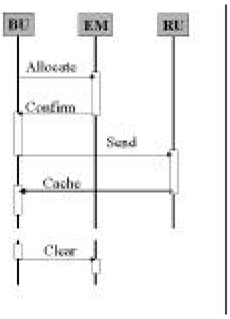


Figure 2

The demonstrator is based on 450 MHz Pentium III based commodity PCs running vxWorks 5.4 RTOS. PCI bus has 33 MHz and 32 bits capability. Intel pro1000 nic are used to connect all the nodes to the GE switch. We exploit the full duplex features of the GE to implement both the builder data network and the builder control network on the same physical switched GE network. Optimized software for EVM, RU and BU has been written according to the event builder protocol adopted.

Event fragments are moved from RU's memories to the BU's memories. The fragments are also fully merged into the BU memories. Every control message is encapsulated into one ethernet frame.

Switch output buffer overflow is avoided using a "credit based" mechanism, where the BU basically limits the request of new events according to level of occupancy of its input queues. A traffic shaping mechanism is not used, but the event builder protocol adopted (each BU asks, with a round robin scheme, event fragments to each RU; no broadcast is used) should "shape" the traffic automatically, avoiding congestion in the output links.

A direct access to the ethernet frames is used. A VxWorks Enhanced Network Device driver (END) has been written for the Intel Pro 1000 nic. END drivers exploit the low latency communication layer provided by vxWorks Scalable Enhanced Network Stack (SENS).

Tests are made with both fixed and variable size event fragment. Variable size event fragments are generated with a gaussian distribution centered to the expected CMS event fragment average size (2 kByte).

3. Point to Point Measurements

This tests have the aim to figure out the basic performance of the GE NIC and the switch.
A single source sends fixed size packets to a single destination through the switch. Fig. 3 shows the measured throughput and the corresponding rate enabling or disabling the standard ethernet flow control mechanism. The sender saturates the receiver causing a packet loss up to 16%. No packet are lost if the flow control is enabled.

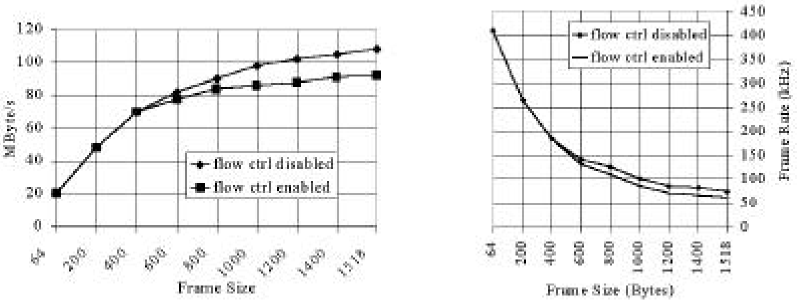


Figure 3

Round trip time (rtt) has been also measured. In this case, a single source sends a fixed size packet to a destination and then waits an acknowledge packet of the same size of the sender from it. Smaller packets yield a rtt around 20 µsec.
The intrinsic limits of the GE technology have been investigated in the following way. 4 data sources located in one switch send fixed size packets to 4 different destinations located to another switch. The two switches are connected via a single GE link. The inter-switche link is saturated (120 MByte/s) with frame size bigger than 800 bytes. Packet rate closed to 1200 kHz are obtained with small frame size (64 bytes).

4. Event Builder Measurements

The achieved 4x4 event builder performances are reported in the following tests.
Fig. 4 shows the receiving bandwidth as seen by a single BU and the event builder aggregate rate, including the effect of the level 1 information broadcasted by the EVM on the same GE network. All the BUs are well balanced as shown in figure 4. Fig. 5 shows the event builder performance varying the configuration from 1x1 to 4x4. In these cases the fragment size is fixed. Fig. 6 reports the effect on the builded event rate of a gaussian size distribution, with the fragment size varying up to 50% of the average size of 2 Kbytes.

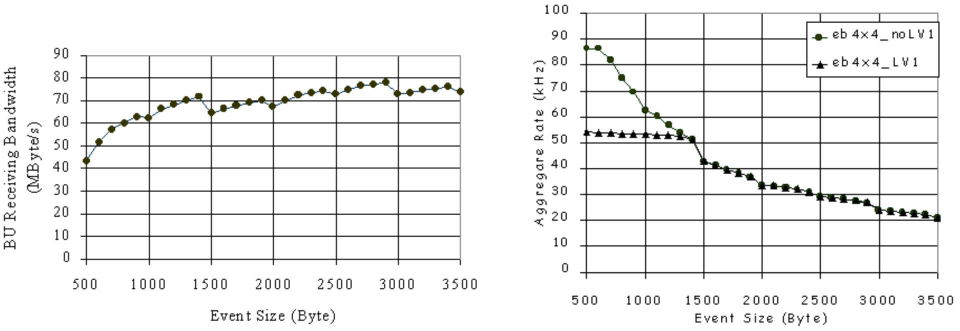
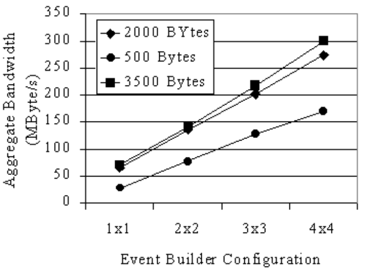


Figure 4



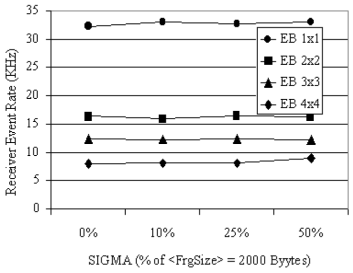


Figure 5

Figure 6

4.1. Conclusions

The CMS event builder demonstrator based on GE has been presented. No performance constrains are introduced by the GE technology. The switches used are fully non blocking. For fragment size of 2 kBytes (fixed) an event builder rate of 30 kHz has been reached with a 4x4 configuration. No substantial performance degradation has been observed for variable size event fragments. Both data and control network can be implemented in the same physical GE network. Level 1 trigger information can be broadcasted using the same GE network, slowing down the aggregate event rate only for event size smaller than 1500 bytes. The BUs are at the moment the overloaded elements: bandwidth in excess of the present PCI bus I/O capability and more processing power is needed . The system scalability is fair up to 4x4. It should be extended to 16x16 in the near future using 1000 baseT copper links. Detailed simulations are required to study the large multistage switch needed to implement the CMS event builder.

References

[1] The CMS collaboration, The Compact Muon Solenoid, CERN, Technical Proposal, N. 7, LHCC 94-38, Dec 1995

[2] CMS TriDAS DAQ Event Builder, System Design Description, draft, August 1999

[3] G. Antchev et al. , The CMS Event Builder Demonstrator based on Myrinet, RT99, Santa Fe, USA