Available on CMS information server

CMS CR 2003/04

CMS Conference Report

April 10, 2003

A Simulation of an Event Building Network for the CMS High Energy Physics Experiment

S. Aziz[a], L. Berti[b], V. Brigljevic[c], G. Bruno[c], E. Cano[c], A. Csilling[c], S. Cittolin[c], S. Erhan[d], D. Gigi[c], F. Glege[c], M. Gulmini[c,b], J. Gutleber[c], C. Jacobs[c], M. Kozlovszky[c], H. Larsen[c], M. Litmaath[c], I. Magrans[c], G. Maron[a], F. Meijers[c], E. Meschi[c], S. Murray[c], V. O'Dell[a], A. Oh[c], L. Orsini[c], L. Pollet[c], A. Racz[c], D. Samyn[c], P. Scharff-Hansen[c], P. Sphicas[c,e], C. Schwick[c], I. Suzuki[a], N. Toniolo[b], and L. Zangrando[b]

[a] Fermi National Accelerator Laboratory, Batavia, Illinois, USA
[b] Laboratori Nazionali di Legnaro dell'INFN, Legnaro, Italy
[c] CERN, European Organization for Nuclear Research, Geneva, Switzerland
[d] University of California, Los Angeles, California, USA
[e] University of Athens, Greece

ABSTRACT

A simulation of the event building network of the Data Acquisition System of the CMS experiment at the Large Hadron Collider at CERN has been developed. The simulation of this highly complex system allows the validation of the system design and the optimization of its performance. The correctness of the simulation model is verified using measurements from test set-ups and a forecast for the full-scale system is made.

Presented at *SCI2003*, Orlando, Florida, USA, July 27-30, 2003.

1. INTRODUCTION

The Large Hadron-Collider (LHC) at CERN is currently being constructed with the start of operation expected for 2007. It will help to answer the question of the origin of mass of particles. It will collide protons at an unprecedented centre of mass energy of 14 TeV. The energy density reached in these collisions is comparable to the energy density present 10^{-18} seconds after the Big Bang - the genesis of the universe. Under these extreme conditions unstable particles will be produced that will in turn decay into other particles leaving a clear signature of their existence. The particles will be recorded and analyzed by dedicated experiments like CMS (Compact Muon Solenoid experiment), consisting of complex detector components to detect and characterise charged and neutral particles with high precision. The analysis of the collision debris will facilitate the reconstructing of the underlying physical process. Filtering out the processes of interest (or "events") is analogous to finding the needle in the haystack.

The proton beams at the LHC cross each other 32 million times per second and the data size of an event at CMS is estimated to be about 1 MB. This would result in a data stream of 32 TB/s if every collision was fully analysed. To reduce the data stream, a sophisticated "trigger" strategy is being developed. Only a few detector components are used to identify potentially interesting events and trigger the readout of all detectors. This will reduce the rate to 100 kHz resulting in a data stream of 100 GB/s. This data stream is managed by the data acquisition (DAQ) system of the CMS experiment. It requires a large and high performance network serving over 500 individual data sources. A central component of the DAQ system is the event builder (EVB), which reads data fragments from the sub-detectors in response to each trigger and assembles these into full events. The data is transported to a computer farm ("filter farm"), which will filter the incoming events in order to reduce the 100 kHz event rate to about 100 Hz for storage.

Different designs of the EVB have been studied using prototypes and simulation. The simulation is based on the Ptolemy package [1], developed at the University of California at Berkeley, which provides a framework for the simulation of complex and heterogeneous systems. The high costs do not allow building a full-scale prototype and thus the simulation of EVB designs is an important tool for validation. A successful simulation of a small-scale prototype EVB improves understanding and enables convincing predictions to be made about the scaling to the full system.

The Event Builder

The design of the EVB is shown schematically in Figure 1 and a summary of EVB acronyms is given in Table 1. The data are first recorded in the front-end drivers of the sub-detectors and transported to Readout Units (RUs). One RU will collect data from up to eight data sources. In turn the RU will forward the data to the Builder Units (BUs). Each BU will receive data from 64 RUs thus forming a complete event. The right hand side in Figure 1 shows in the "front view" one Readout Unit (RU) Builder. The RU Builder consists of a 64×64 switch connecting RUs with BUs. The front-ends are connected via multiple 8×8 switches to the RU Builders. On the left hand side the 8×8 switches, staggered perpendicular to the RU Builder, are shown. They constitute with the interface to the front-ends and the inputs to the RU the "Front-End Driver (FED) Builder".

Table 1 List of event builder acronyms.

RU	Readout Unit	FED	Front-End Driver
BU	Builder Unit	FU	Filter Unit
FRL	Front-End Readout Link	RM	Readout Manager
RUI	Readout Unit Input	BM	Builder Manager
BDN	Builder Data Network	RCN	Readout Control Network

There are several advantages to this two-stage EVB as compared to a single-stage EVB that would be based on a single switching network: The decoupling of RU Builder and FED Builder allows a staged deployment of the DAQ. It is foreseen to install only half of the RU Builder at the start of data taking in 2007. However the FED Builder cannot be staged. There is no interdependence between RU Builders thus making the system more robust against failure of a single RU Builder.

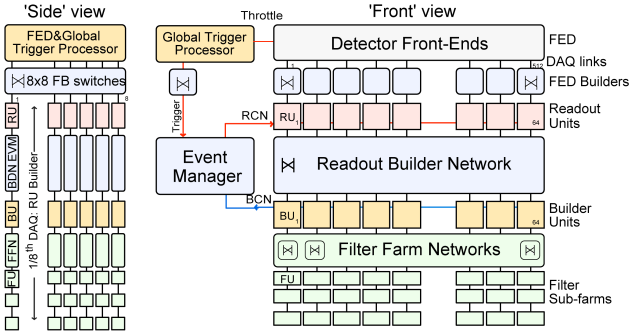

Figure 1 Schematic of the Event Builder. The "front view" of one Readout Unit (RU) Builder is shown on the right hand side. The RU Builder consists of a 64x64 switch connecting RU Inputs with BUs. The front-ends on top are connected via multiple 8x8 switches to the RU Builders. The 8x8 switches staggered perpendicular to the RU Builder are shown on the left hand side. They constitute with the interface to the front-ends and the RU Inputs the "Front-End Driver (FED) Builder".

FED Builder

The elements of the FED Builder within the EVB are shown in Figure 2. The Global Trigger Processor [2] identifies interesting events in the detector and invokes via trigger signals the readout of the FEDs and is part of the FED Builder domain. The FED Builder connects the FEDs via the Front-End Readout Links (FRLs) to the Readout Units (RUs) via the Readout Unit Inputs (RUIs). The fragments from the FRLs are assembled into super-fragments (s-fragments) in the RUIs. The typical fragment size is 2 kB. Thus an s-fragment, composed of eight fragments, has a typical size of 16 kB. The FED Builders assure that s-fragments from the same event will be sent to the same RU Builder. At the maximum trigger rate of 100 kHz, each port of the FED Builder must sustain a throughput of about 100 kHz times 2 kB, i.e. 200 MB/s. The design foresees the use of two links per port in a two-rail network to provide the necessary performance. In this study, different distributions of the fragment size were simulated to investigate the sensitivity of the FED Builder performance to variations and correlations among the fragments.

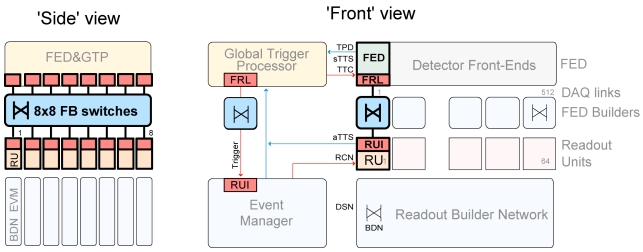

Figure 2 Elements of the FED Builder in the EVB.

RU Builder

The RU Builder is the second stage of the EVB. It receives the s-fragments from the FED Builder and forwards the s-fragments belonging to the same event to one BU, which assembles the full event and forwards it to the Filter Units (FU) for further processing. The full DAQ system has eight RU Builders. Each RU Builder will have to sustain an event rate of up to 12.5 kHz. This is equivalent to 12.5 kHz·16 kB=200 MB/s for each RU/BU node.

The fragments are exchanged over the Builder Data Network (BDN). The Event Manager (EVM), consisting of a Readout manager (RM) and Builder Manager (BM), controls the event building process over the Readout Control Network (RCN) and Builder Control Network (BCN). It is part of the RU Builder domain. The control networks and the BDN may be implemented as a single physical network, or as two separate networks.

Network Technology

Two network technologies for the EVB, namely Gigabit Ethernet and Myrinet, are considered in this study. While Gigabit Ethernet is widely used in commodity computing and is well established, Myrinet is mainly found in the high performance computing domain. Myrinet, a product of Myricom [3], is a Gb/s technology offering a full suite of both network interface cards (NICs) and switches at relatively low cost. A Myrinet network is composed of switching elements and network interface cards connected by point-to-point bidirectional links. The effective link speed is currently 2 Gb/s. The possibility to use one or two rail networks will be given with a new generation of NICs comprising two 2 Gb/s ports on a single NIC, effectively doubling the bandwidth.

Switches for Myrinet are based on a switching chip that is a pipelined 16-port crossbar, supporting wormhole (also known as "cut-through") routing of packets. Packets can be of arbitrary size. Network link-level flow control guarantees the delivery of packets at the expense of an increased potential for blocking in the switches. The NIC has a RISC processor whose firmware can be programmed to interact directly with the host processor for low-latency communications and with the network links to send, receive and buffer packets.

Traffic shaping

Even assuming the existence of a NxN switch with crossbar-like connectivity, the data traffic pattern in the Event Builder where all sources send data to one destination implies blocking or loss of data unless the switch provides output queues large enough to store the equivalent of an entire DAQ event. Such a switch would be prohibitively expensive. For an efficient use of a switch without large output buffers, the switching capability of the crossbar must be supplemented with some packet-level algorithm that provides an arbitration mechanism for the output link. The procedure used to share this link, as well as to ensure that all the other switch links are used concurrently, is referred to as "Traffic-shaping".

Unlike the case of the FED Builder, where the necessary network performance is reached by either leaving the network traffic unmodified or by cutting fragments into multiple packets, the RU Builder uses traffic shaping algorithms to optimize switch utilization. Since the number of nodes is eight times larger in the RU Builder compared to the FED Builder, the congestion in this multi-staged network would become unacceptably large if no traffic shaping algorithms were applied. The Barrel Shifter scheme [4] is employed for a Myrinet based RU-Builder. Destination-driven traffic shaping is used for the RU Builder with Gigabit Ethernet technology.

Simulation

The Ptolemy package provides a framework for the simulation of complex and heterogeneous systems. For the implementation of the EVB simulation, the discrete event (DE) domain was used. The DE domain in Ptolemy is designed for time-oriented simulations of systems such as communication networks. Actions are described as the exchange of "Particles" between objects. These objects, called "stars" in the Ptolemy terminology, can receive Particles over one or many ports and trigger the emission of new Particles.

In the simulation, the EVB is modelled as functional units implemented as "stars" that will be described in detail in the following sections. The EVB communication protocol of the different units is modelled in analogy to the real EVB system described in detail in reference [5]. A schematic of the structure of the simulation with its components and the relation between them are shown in Figure 3.

Either the FED Builder or the RU Builder can currently be simulated. FED and RU Builders combined in a single system have not been simulated since the RU will have sufficient memory to be decoupled from statistical fluctuations of the FED Builder. The performance of the full system is governed by the least performant component of the FED Builder - RU Builder system. The individual components of the EVB are now discussed in detail.

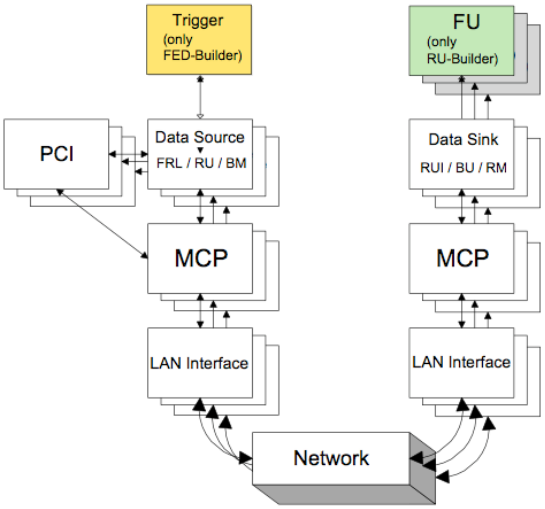

Figure 3 EVB components and their relations.

Data Sources

The data chain begins with the RU or FRL respectively. Components in the readout chain prior to the RU/FRL are not implemented. The data source is modelled with a generator of fragment sizes. The probability distribution can be freely chosen from predefined analytical functions or histograms and can be different for each RU/FRL. It is also possible to vary the correlation of fragment sizes between data sources.

The flexible design of the data source allows a rich set of scenarios to be studied. Besides correlation, other aspects of data conditions such as unbalanced fragment sizes and pathologically large fragments are studied. The imbalance of inputs is parameterized with the imbalance ratio *R* defined as the ratio of the maximum average fragment size and the minimum average fragment size of an ensemble of data inputs. These modelled non-uniformities of data size distribution are motivated by the different characteristics of sub-detectors and the topology of events.

The Trigger Model (FED-Builder)

The trigger in the simulation generates trigger signals with a Poisson time distribution with adjustable mean. External objects can throttle the trigger. It monitors the trigger efficiency which is defined as $\varepsilon = N_{acpt}/N_{tot}$ with N_{acpt} the number of trigger signals issued in the unthrottled state and N_{tot} the total number of trigger events. The trigger connects to all FED builders in case of multiple FED builders.

The FRL Model (FED-Builder)

The FRL represents the first element in the data flow of the FED-Builder simulation. The FRL is connected to a trigger. The FRL will throttle the trigger if it cannot keep up with the rate of incoming data due to its finite memory. The PCI bus access is modelled with priority queues and a simple arbitration for different access modes. Bus access times are parameterised with constant set-up times and effective transmission speeds taking into account wait cycles. The interface to the network is modelled with the NIC model described later. No traffic shaping is applied in the FED builder.

The RUI Model (FED-Builder)

The RUI model is used to study the performance of super fragment building and memory usage. The RUI collects the data fragments created by the FRLs. It assembles super-fragments and forwards them to the RUs. The simulation models the collecting of fragments and the assembling of the super-fragments. The RUI is implemented without a detailed model of the PCI bus. This is justified by the fact that the forwarding of the super-fragments to the RUs (with a PCI bus speed of 512 MB/s) is fast compared to the time needed to collect fragments (effective link speed of 200 MB/s).

The RU Model (RU-Builder)

The RU features components similar to the FRL. It includes the models of the PCI bus and the NIC, but it does not connect to a trigger. The data source is modelled in the same way as in the FRL. In the RU Builder simulation super-fragments are assumed to be always available for forwarding i.e. the RUI saturates the RU. Limitations in the performance of the RU builder are thus not caused by the FED builder, but by genuine RU properties and represent the saturation limit.

The BU Model (RU-Builder)

Similar to the modelling of the RUI, the PCI bus is not included in the simulation of the BU. The omission of the PCI bus is justified by the fact that the effective link speed between the RU and the BU will be slow compared to the transmission speed of super-fragments from the receiving NIC to the BU over the PCI bus. Furthermore, it is assumed that the PCI bus will not severely limit the transmission speed of control messages. The BU can be operated with and without the EVM. When the EVM is absent, the BU will act as a pure sink. Optionally the assembly and shipping of complete events to the FUs can be invoked

The EVM Model (RU-Builder)

The EVM model consists of an RM, BM, trigger and an RCN/BCN. The RM has the task of tagging triggered events with so called Event IDs and distributing the information to the RUs with minimal latency. The Event ID is attached, or allocated, to the event for the time the event data is being collected. The BM serves allocated Event IDs to the BUs. The BUs request event data from the RUs using Event IDs. An allocated Event ID becomes available again as soon as it is released by the BU, i.e. all event data has been collected.

The control networks and the BDN can either share the same or each have their own separate physical network. The configurable EVM parameters include the Event ID table size, packing factor of control messages, bandwidth of the control network and trigger rates.

The FU Model (RU-Builder)

The FUs are modelled as a simple extension to the BU. The FUs are connected to the BU over an idealized network. Each FU has settable processing time distributions and memory size. Typically eight FUs were connected to each BU.

The NIC Model (Myrinet)

Before data can be sent to or received from the network, they have to be copied from the host memory into the memory of the NIC or vice versa. Thus, a model of the NIC is included in all models which are connected to the Myrinet network: FRL, RUI, RU, BU and EVM.

The copy process uses both DMA and programmed I/O modes to transfer data via the PCI bus. The latency of the transfer process is modelled in detail taking into account the finite bandwidth of the PCI bus, its usage, and processing times. In case of the RUI and the BU models, the communication between NIC and RUI/BU is processed directly, bypassing the DMA and Programmed I/O modelling.

The NIC is modelled as two stars, the MCP star and the NI star. The MCP star models the custom firmware running on the RISC processor in the NIC and the NI star simulates the actual hardware interfacing to the network. The latencies introduced by the NIC are given by the processing times for distinct tasks. These processing times are given as parameters to the NIC model and have been determined from test-bench set-ups. Thus the simulation requires no "free" parameters, beyond those obtainable from point-to-point measurements.

The NIC model can simulate two versions of the Myrinet hardware, LANai9 and LANai10, the difference being that LANai9 has one 2 Gb/s port whereas LANai10 will provide two 2 Gb/s ports and will allow higher processing speeds relative to LANai9 hardware due to faster memory, a faster processor and an improved DMA architecture. Since the LANai10 is not yet available, the anticipated performance has been estimated by reducing the processing times by a factor two and adding a second port in the simulation.

The NIC Model (Gigabit Ethernet)

The Gigabit Ethernet NIC is modelled as a simple store-and-forward unit. It introduces a fixed latency to the packet transfer in order to account for system overheads.

Single Switch Elements (Myrinet)

Implementations of single crossbar switch elements with 16 ports corresponding to Myrinet Xbar16 switches have been made by others [6]. The implementation follows closely the Myrinet network technology using wormhole routing, back-pressure flow control and small slack buffers. A round-robin token arbitration scheme has been added.

Single Switch Elements (Ethernet)

The switch model used is based on an architecture with internal buffer memory, output queues and a high-performance cross point backplane. The switch consists of 8 line cards, each equipped with 8 full-duplex Gbit ports. The line cards are connected via an 8 Gbit/s backplane to each other. The line cards have 2 MB shared memory, equivalent to about 1000 Ethernet packets. Incoming packets are dropped if the available resources are not sufficient to store them.

2. SIMULATION OF THE FED BUILDER

The FED-Builder operates in a pure "push" mode, incoming fragments on the FRL side are pushed to their destination on the RUI side and no traffic shaping is employed. The switch utilization can be improved dividing fragments into packets of a Maximum Transfer Unit (MTU), so large fragments get divided into multiple smaller packets. This reduces the effect of output port blocking by large fragments, since other senders can deliver their packets interleaved and continue with the next destination.

Comparison with Data

The throughput of a simulated LANai9 FED-Builder without MTUs for different distribution types is compared to test-bench data in Figure 4. The distributions have a log-normal shape with the mean fixed to 2 kB. They differ in width and the imbalance ratio R. In general, a good agreement between data and simulation is observed.

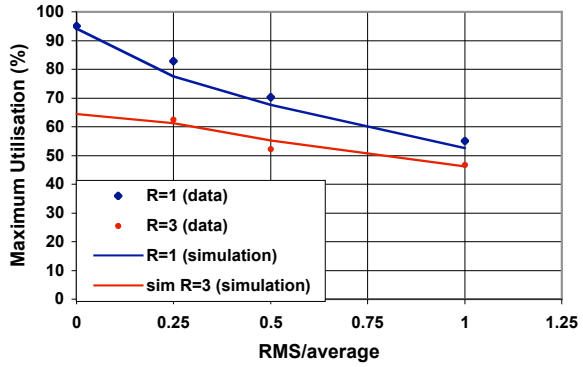

Figure 4 Throughput for different distribution types for data and simulation for the FED Builder.

Starting from the good description of the current test-bench hardware, different scenarios were simulated to investigate the influence of data conditions, the performance of LANai10 hardware, and the usage of MTUs. The fundamental "figure-of-merit" is the maximal trigger rate the FED Builder can accept without throttling the trigger.

Trigger Rate and Throughput

The trigger efficiency was scanned as a function of the trigger rate. A drop in efficiency is expected if the product of trigger rate and average fragment size exceeds the maximum throughput of the system. The result of the simulation is shown in Figure 5 for LANai9 and LANai10 hardware with and without the usage of MTU for a log-normal fragment size distribution with a mean of 2 kB and an RMS of 2 kB. The trigger efficiency starts to degrade at a trigger rate, that coincides well with the expected rate from the simulation of the maximum throughput.

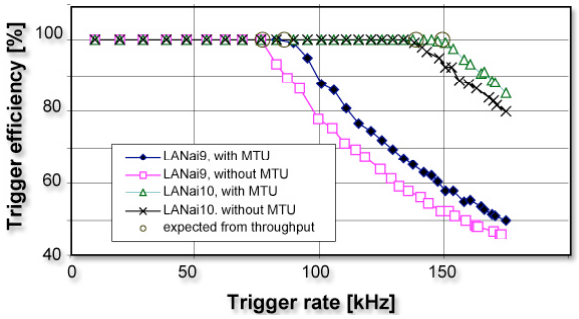

Figure 5 Trigger efficiency vs. trigger rate for different hardware configurations. The open circles mark the from the saturation throughput expected maximal trigger rate.

The result of having multiple FED Builders in the system has been studied with a system of 64 single 8x8 FED builders corresponding to the full system. It is observed that the maximum trigger rate is not sensitive to the number of FED Builders in the system. The trigger efficiency of a system of multiple FED Builders is not simply the product of the single FED Builder efficiencies, because the throttling of the trigger affects all FED Builders and hence the systems are tightly coupled.

Balanced and Imbalanced Input Distributions

The dependence of the throughput on the mean of a log-normal distribution is shown in Figure 6 for LANai9 and LANai10 hardware with and without the usage of MTUs. The root mean squared (RMS) of the distribution is set equal to the mean. In this scenario, all inputs have the same fragment size distribution and are uncorrelated. The plot shows the clear performance gain when the two-rail LANai10 hardware is employed, and also demonstrates the beneficial effect of using MTUs.

The performance of the FED-Builder is sensitive to the fragment size distributions of the FRL. The case where a subset of FRLs send more data on average than the others, was studied by varying the imbalance ratio R for a configuration with 4 FRLs above and 4 below the average (referred to as configuration A), and another configuration with 1 above and 7 below the average (referred to as configuration B). For LANai10 hardware, the target throughput of 200 MB/s is sustained for configuration "A" up to R-values of 3.5, while for configuration "B" the limit is reached at R=1.9. Although the configuration "B" seems to be more sensitive to R, it should be noted that in terms of the actual maximum average fragment size the values do not differ by much, namely 3.1 kB for the configuration "A" and 3.4 kB for the configuration "B".

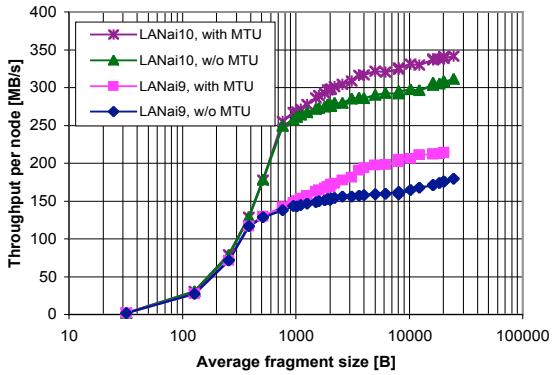

Figure 6 Throughput per node dependence on the mean of the input distribution to the FRLs. The RMS is set equal to the mean of the log-normal distribution.

Correlation of Data Sources

The correlation of fragment sizes between different FRLs is very likely to occur, since the data volume of many detector components is determined commonly by the characteristics of the event in the detector. The sensitivity of the FED Builder performance has been studied by varying the linear correlation coefficient between the FRL data sources for a lognormal distribution with an average of 2 kB and an RMS of 2 kB. The variation of the correlation from 0% to 100% resulted in a negligible influence on the performance for LANai9 and LANai10 hardware with and without MTU.

3. SIMULATION OF THE RU BUILDER

The network of the RU Builder can be realized with Gigabit Ethernet or with Myrinet. Laboratory tests with both network technologies have been pursued and demonstrator set-ups with up to 64 ports have been built. The scope of the RU Builder simulation is to verify results obtained with the test bench set-ups and to predict, on the basis of the reliable description of the test bench data the performance of the final system.

The simulation of the RU Builder closely follows the set-up of the test benches as well as the layout proposed for the final system, which includes the EVM. The FED Builder has not been included in the simulation of the RU Builder. The data source is integrated into the RU model. As explained above, the coupling between the two builders is low, so that each subsystem can be studied independent of the other as it was done with the test benches.

RU Builder with Myrinet

The RU Builder in the current design is a 64x64 port network. The core switching-element is an 8x8 crossbar and thus a 64x64 network must be realized as a multistage network. Different topologies (two stage delta network made of 16 8x8 cross-bars, three stage folded Clos network made of 24 8x8 cross bars) have been investigated for the event builder. When the barrel shifter traffic shaping is deployed, the switch utilization is close to 100% and independent of the network topology used. For random traffic with fixed fragment size a network utilization of about 45% was observed for both topologies.

Results without EVM

Extensive studies of the throughput dependence on the fragment size allowed the verification of the simulation model of the hardware. Special attention has been paid to reproduce the measurement, which can be seen as a benchmark test for the simulation. Especially the preparation of packets in the NIC, which has to keep up with the barrel shifter cycle, is a time critical process since it pushes the NIC to the limit of its processing power and internal memory bandwidth.

The simulation results are compared with the data obtained from the test bench operating with 32 RUs and 32 BUs, without EVM in Figure 7. In this case it is assumed that the BUs will always have allocated Event IDs available. The two simulation curves differ in the assumptions of the memory bandwidth internally available to the NIC processor. If multiple DMAs are ongoing, the remaining memory bandwidth available can limit the execution time of the program running on the NIC processor. To model this, the processing times were increased depending on the number of ongoing DMAs. The lower curve implements this "memory-bandwidth" effect, while the upper curve assumes that sufficient bandwidth is available. The data points are enclosed in between the two curves, the band can be seen as the level of uncertainty in the simulation model. The operation point of the RU Builder will be at 16 kB corresponding to the average size of a super-fragment, where the influence of the limited memory-bandwidth is insignificant.

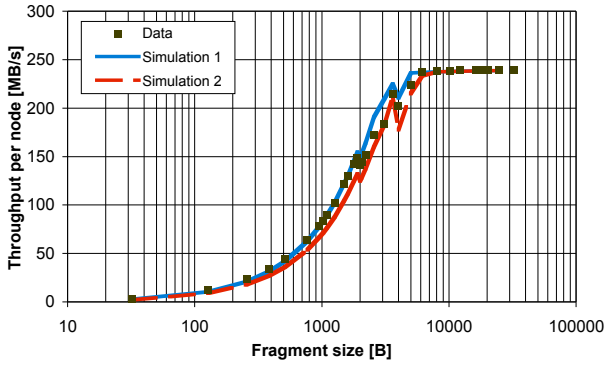

Figure 7 Throughput per node as a function of the fragment size for a 32x32 network and LANai9 hardware. The curves represent the simulation result for two different performance assumptions. The symbols show the data points obtained with the test set-up.

Results with EVM

The effect of the addition of an EVM has been also studied. The final system will have an EVM with an RM on the BU side and BM on the RU side of the switch. The addition is expected to affect the performance of the RU builder since, firstly, one port of the builder is used for the EVM, and, secondly, the distribution of Event-ID, or rather a shortage of them, can delay the sending of fragments and thus degrade the performance. The first cause of degradation is inherent to the design and is irreducible. The effect of the second cause depends on the number of available resources and the servicing times.

The simulation is compared to data obtained with the test bench set-up in a configuration of 31 RUs, 31 BUs and one BM on the RU side and no packing of control messages (see Figure 8). The degradation of the performance due to the presence of the BM is consistent with what is expected from a barrel shifter.

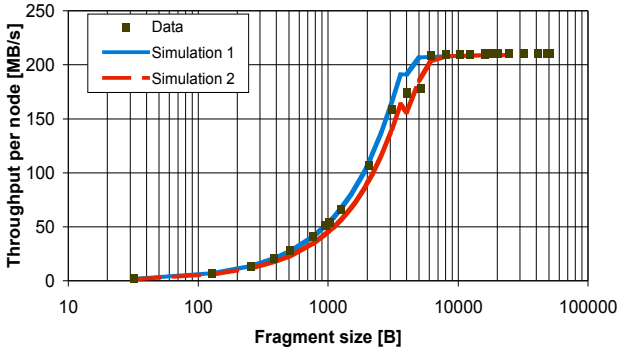

Figure 8 The throughput per node vs. fragment size for a 31x31 RU Builder with BM is compared for simulation results and data.

The performance of the final 64x64 system has been simulated based on the parameters describing the 32x32 test bench and is shown in Figure 9 for a configuration with and without BM. The scaling of the system is as expected for a barrel shifter. The performance with BM is slightly better than for the 32x32 configuration since the fraction of ports used to transfer event data is 63/64 compared to 31/32. The performance of the anticipated LANai10 hardware using only one rail is shown in the same figure. The plateau is now attained already at smaller fragment sizes because the NIC is now faster and has time to pack more fragments in a barrel shifter packet.

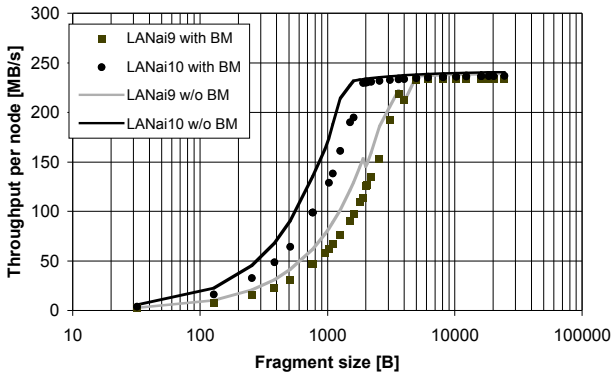

Figure 9 Throughput vs. fragment size for a 63x63 RU Builder with BM for LANai9 and LANai10 hardware.

RU Builder with Ethernet

The simulation result of a RU Builder with Gigabit Ethernet is shown in Figure 10. For fragment sizes below 16 kB, where no packet loss is observed, there is a good agreement between data and simulation. For larger fragment sizes packet losses start to occur resulting in a drop of throughput. The refinement of the simulation parameters and the modelling is in progress to reproduce the behaviour for large fragment sizes.

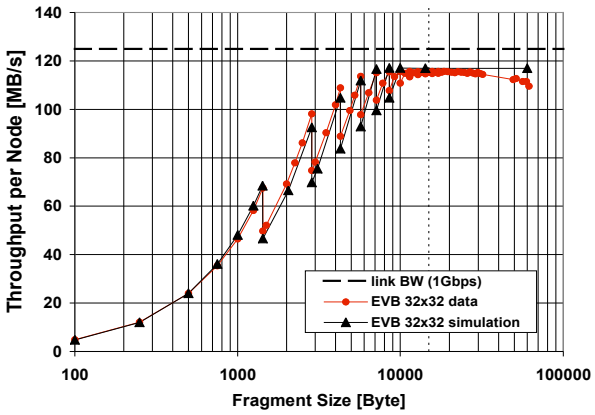

Figure 10 Comparison between data from a test set-up and simulation of the throughput per node for a 32x32 EVB based on Gigabit Ethernet.

4. SIMULATION OF THE FULL EVENT BUILDER

The full system with FED Builder and RU Builder combined has not been simulated. However, both stages have been studied separately considering the missing element as a pure sink or ideal source, respectively. The assumption that both stages are decoupled is justified if the memory in the RU can absorb the fluctuations of the output of the FED Builder. The overall EVB performance is clearly determined by the minimum of the performance of the FED Builder and RU Builder stages.

A prediction is made for the maximum trigger rate as a function of the number of RU Builders, assuming balanced and uncorrelated inputs (see Figure 11). A RU Builder based on a one-rail Myrinet network employing a barrel shifter traffic-shaping algorithm is used. Both a one-rail (LANai9) and two-rail (LANai10) scenario are simulated for the FED Builder. The capacity of the set of RU Builders increases linearly with the number of RU Builders (by construction) and will hold irrespective of the technology used for the RU Builder. For a single RU Builder, a one-rail FED Builder will provide enough performance to reach the 12.5 kHz trigger rate. The FED Builder performance does not scale linearly because of the output blocking. For more RU Builders, only the two-rail based FED Builder can match the capacity of the RU Builders. A maximum trigger rate of 110 kHz can then be reached with the full EVB system.

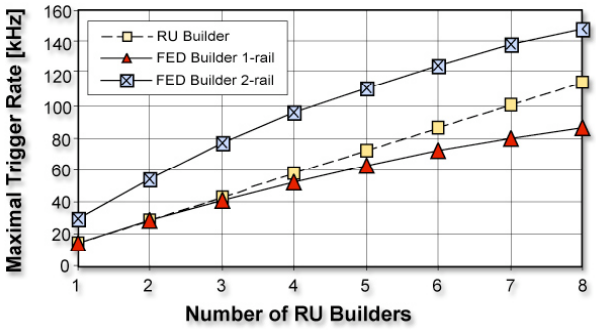

Figure 11 The maximum trigger rate as a function of the number of RU Builders in the DAQ system for a one- and two-rail FED Builder and a one-rail Myrinet Barrel Shifter RU Builder. Inputs are assumed to be balanced and uncorrelated. Inputs are generated according to a log-normal distribution with an average of 2 kB (16 kB) and an RMS of 2 kB (16 kB / 8), for the FED Builder (RU Builder), respectively.

5. SUMMARY

A model of the CMS DAQ system and its components was developed in the Ptolemy framework. The modelled components are verified with data measured from test set-ups of the EVB. A good agreement of data and model is observed. A detailed study of the dependence of the EVB performance on different data conditions shows that the design is robust against a wide range of conditions. A simulation of the full system for RU and FED Builders shows that the requirement of fully efficient operation at a 100 kHz trigger rate can be fulfilled with the current design.

REFERENCES

[1] The Ptolemy Project, see http://ptolemy.berkeley.edu

[2] "CMS The Trigger Systems, Technical Design Report", CMS-TDR 6.1, CERN/LHCC 2000-38(2000)

[3] Myricom Inc., Arcadia, CA, USA, see http://www.myri.com

[4] E. Barsotti, A. Booth and M. Bowden, "Effects of Various Event Building Techniques on Data Acquisition System Architectures", Santa Fe Computing 1990:82-101, http://fnalpubs.fnal.gov/archieve/1990/conf/Conf-90-061.pdf

[5] "CMS Data Acquisition & High-Level Trigger Technical Design Report", CMS-TDR 6-2, CERN/LHCC 2002-26 (2002).

[6] J.-P. Dufey et al., "The LHCb Trigger and Data Acquisition", IEEE Trans. Nucl. Sci., **47** (2000) 86; B. Rensch, "A Few Results from EvtBldg-Sim", Niels Bohr Institute, Copenhagen, 1998.