Testing Ethernet Networks for the ATLAS Data Collection System

F. R. M. Barnes, R. Beuran, R. W. Dobinson, M. J. LeVine, *Member, IEEE,* B. Martin, J. Lokier, and C. Meirosu

Abstract-- **This paper reports recent work on Ethernet traffic generation and analysis. We use Gigabit Ethernet NICs running customized embedded software and custom-built 32-port Fast Ethernet boards based on FPGAs to study the behavior of large Ethernet networks. The traffic generation software is able to accommodate many traffic distributions with the ultimate goal of generating traffic that resembles the data collection system of the ATLAS experiment at CERN. Each packet is time stamped with a global clock value and therefore we are able to compute an accurate measure of the network latency. Various other information collected from the boards is displayed in real time on a graphical interface. This work provides the tools to study a test bed representing a fraction of the 1600 ATLAS detector readout buffers and 600 Level 2 trigger CPUs using a combination of the Fast Ethernet boards and the Gigabit Ethernet NICs.**

Index Terms **Networks, Switches, Modeling, Measurement**

I. Introduction

ATLAS is one of the four experiments being implemented for the Large Hadron Collider at CERN, which is expected to begin operation in 2006. The ATLAS intersection will produce collisions at rates up to 40 MHz. Selected collisions are accepted by the Level 1 (LVL1) trigger, based on summary information from selected detectors, at rates of up to 100 kHz. Events selected by LVL1 are then passed to a farm of 500 - 700 Level 2 (LVL2) CPUs [PCs running Linux]. A CPU which is assigned an event requests data (Regions Of Interest or ROIs) from a fraction of the approximately 1600 readout buffers (ROBs) based on information received from the LVL1 system. On average these ROIs constitute 5 per cent of the entire event image (2 MByte).

The data corresponding to entire events accepted by LVL2 (approximately 2-5 kHz) must be transported from the ROBs to the higher level trigger (Event Filter).

The LVL2 CPUs, the Event Filter CPUs, and the ROBs will be interconnected by an Ethernet switch fabric. The transaction rate and the data rate expected will impose stringent requirements on the network and on the attached nodes (LVL2 and Event Filter CPUs and ROBs). The ATLAS LVL2 architecture is described elsewhere in [1-3].

There are many open questions regarding the network topology and the protocols to be used. The behavior of switches at high occupancy needs to be determined, including measurements of latency, numbers of lost packets, reliability of multicasts, and the utility of VLANS.

In the work reported here, we measure the properties of network switches and switch fabrics with traffic generators created for this purpose. For Gigabit Ethernet (GBE) we make use of Alteon [4] network interface cards (NICs) running customized software as well as custom-built 100 Mbit Ethernet (FE) traffic generators/testers, described in detail later.

II. ATLAS Trigger System Emulation

The aim of the work described here is to develop tools which will allow direct measurements of the behavior of a moderate-sized (approximately 1/10 scale) subset of the ATLAS trigger system emulated with traffic generators in both GBE and FE.

III. Model Validation

Behavior of the full-scale network fabric will be determined by computer models of the network. These models, described previously in [1] and elsewhere in [5] in this conference, are based on precise measurements of individual switch behavior, and generate predictions for the behavior of moderate-sized systems which can be compared directly to the measurements of systems identical to those being modeled.

The measurements of individual switch behavior are carried out using the techniques described here, as are the measurements of the behavior of the moderate-sized systems. This is an iterative process, where the measurements result in refined models and model parameters; the refined models are then tested against new measurements.

The traffic generators described in this paper are capable of generating traffic covering a wide spectrum of frame sizes and rates, so that the entire operating space of the switches can be covered by the measurements.

IV. The Need for Customized Traffic Generators

An accurate model requires measurements of latency, throughput and packet loss collected on a large-scale test bed. Standard PCs are not able to produce Ethernet traffic at full line speed for GBE nor precise traffic characteristics for FE traffic. A global clock is required in order to perform timing measurements; clock synchronization over the network is not sufficiently accurate for these measurements.

Manuscript received June 21, 2001; revised November 29, 2001.

R. Beuran, R.W.Dobinson, M.J. LeVine, J. Lokier, B. Martin, and C. Meirosu are with CERN, 1211 Geneva 23, Switzerland. (telephone (41) 22 767 3908, email: Micheal.LeVine@cern.ch)

F. R. M. Barnes is with the University of Kent, Canterbury, UK

M. J. LeVine is also supported by Brookhaven National Laboratory under U.S. Department of Energy contract DE-AC02-98CH10886

J. Lokier is also with Degree2, Ltd.

R. Beuran and C. Meirosu are also with Politehnica University of Bucharest, Bucharest, Romania

V. The Fast Ethernet Tester

This board, custom-built at CERN, implements 32 full-duplex 100 Mbit/s Ethernet ports utilizing Altera Flex 10K FPGAs to implement the Media Access Controllers (MAC), as well as other functions which will be described here. All of the FPGAs are programmed almost entirely in Handel-C [6], a commercially available high-level language. The decision to build these boards rather than use commercial network testers was partly based on economics; however the most compelling motivation was the ability to generate traffic identical to that to be found in the ATLAS trigger system. This is not achievable in existing commercial testers, which do not generate Poisson-distributed frame intervals, nor can they generate frames in response to incoming frames. The board architecture is shown in Figure 1.

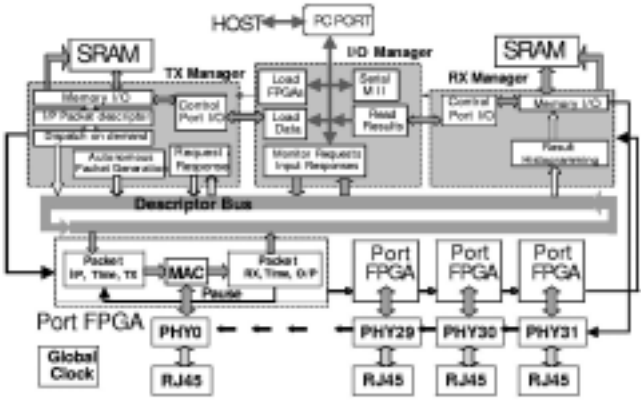

Figure 1. Architecture of the Fast Ethernet tester board

A. Fast Ethernet tester architecture

Connection to the host CPU is implemented using an IEEE 1284 enhanced parallel port connection. Management of this link on the tester board is carried out by an FPGA (IoMan), which passes data and commands from the host to the various other components on the board, as well as managing data destined for the host.

A 25 MHz slotted ring interconnects the 32 MACs, IoMan, as well as two other FPGAs: the packet transmit manager TxMan and the packet receive manager RxMan. The slotted ring is 52 bits wide; 32 bits are used to transport data, while most of the remaining bits are used to transport command and status signals as well as the global 25 MHz clock.

During operation, the function of the slotted ring is to deliver packet descriptors from TxMan to each of the MACs in turn. A descriptor consists of 5 32-bit words, and contains sufficient information for the MAC to generate an outgoing Ethernet frame. The exact contents of the transmit descriptor depend on the measurement being performed. TxMan has 1 Mword (36 bits wide) of private static RAM which is used to store the transmit descriptors generated by the host. The bandwidth provided by the slotted ring is sufficient to allow each MAC to generate the smallest (64-byte) Ethernet frames at line speed.

Incoming packets are processed by the MACs to generate receive descriptors (5 32-bit words) which contain the essential information extracted from the incoming frames. For example, time stamps contained in incoming frames are compared with the current clock to compute a latency. These descriptors, loaded unto the slotted ring by each MAC, are received by RxMan, which uses the information to update counters and histograms of latency, frame size or difference in arrival time between successive packets, depending on the measurement being performed. The histograms are stored in a private static RAM of 1 Mword (36 bits wide), and are ultimately uploaded to the host over the IEEE 1284 link.

The nature of the histograms being accumulated is determined by control registers implemented in RxMan. The nature of the information extracted from incoming frames by the MACs can be re-programmed as needed.

Transmit descriptors are produced by TxMan using one of two means. The descriptors can be generated from information stored in TxMan s descriptor memory, cycling through the descriptors associated with each MAC in either a systematic or quasi-random way. Alternatively, the incoming receive descriptors can be used by RxMan to generate transmit descriptors which are then passed to TxMan. It is the latter mechanism that will be used to generate the ATLAS request-response behavior of the ROBs.

The processing of transmit descriptors to produce outgoing frames is carried out by a CPU implemented in Handel-C in the MACs. Generating a behavior for this CPU requires assembling a program on the host, which is fed as data to the MAC; this is a very quick procedure compared to the process of reprogramming and refitting the MAC. All outgoing frames contain a timestamp based on the board s global 25 MHz clock; the timestamps are thus coherent to 40 ns.

B. Handel-C

Handel-C [6] is a high-level language designed specifically to generate netlists for FPGAs. The syntax of the language is similar to that of C, with certain additions, e.g., variables of arbitrary width. Parallelism is introduced using syntax similar to that of Occam [7], including *channels* to synchronize parallel processes. The language itself is easy for C programmers to learn; however, it is the parallelism that requires a new way of thinking. Unlike processes running on a CPU in parallel, each process on an FPGA executes a statement on every clock cycle.

Compiling a program written in Handel-C involves two steps: the Handel-C compiler produces a netlist which in turn is input to the Altera MaxPlusII [8] fitter. The resulting code is sent over the IEEE 1284 link to IoMan.

Each of the FPGAs is re-programmable from the host. IoMan, whose behavior rarely changes, is programmed using JTAG over a Byte Blaster cable. The remaining FPGAs are re-programmed via a JTAG chain managed by IoMan, using data transported over the IEEE1284 link.

VI. The Gigabit Ethernet Tester

The GBE tester is based on the Alteon [4] Gigabit Ethernet NIC. This card is based on the Tigon chip, which contains two customized MIPS CPUs, a DMA engine, RAM, a GBE MAC, and a PCI interface (see Figure 2). The code supplied by the vendor has been modified to allow us to utilize the NIC as a tester. Time stamps for outgoing packets are determined from a global clock card plugged into the PCI back plane. The two CPUs are utilized for outgoing and incoming traffic, respectively. Finally,

outgoing frames are generated on the NIC, and incoming frames are processed on the NIC, in order to avoid the non-deterministic behavior associated with transfers across the PCI back plane. The software is written in C, and compiled with tools adapted from the gcc compiler.

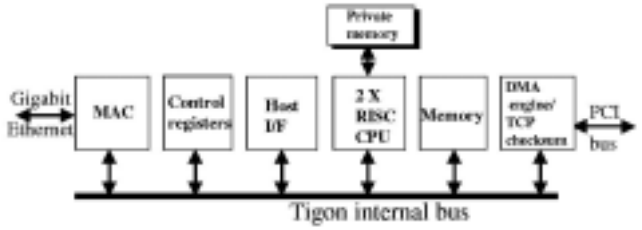

Figure 2. The Alteon NIC architecture

An industrial PC chassis equipped with 12 PCI slots is used to host 8 Alteon NICs and a global clock card.

The absolute time used to time stamp outgoing packets and to calculate latencies for incoming packets is determined by polling a clock card on the PCI bus. The shared nature of the bus internal to the Tigon chip as well as that of the PCI bus limits the precision of the time measurement; the precision achieved is approximately 200 ns.

VII. Performance of the Ethernet Traffic Generators

Traffic generation is accomplished at full line speed for every packet size, both for Fast Ethernet and Gigabit Ethernet. Traffic shape is programmable. At present, constant bit rate (CBR) and Poisson distributed traffic have been implemented. Since the traffic pattern depends on a set of fully configurable traffic descriptors which determine the time of emission, destination address, as well as frame length, generation of arbitrary traffic patterns is possible.

Global clocks determine the precision of time stamping. For the case of the Fast Ethernet tester, the 25MHz clock is global to the board, with a precision of 40 ns. A facility exists to distribute the clock to multiple Fast Ethernet tester boards.

The measured distribution of arrival time differences between successive packets for constant bit rate traffic for the Fast Ethernet tester with ports directly connected (no switch) has a spread of 40 ns, which is the precision of the measurement imposed by the clock frequency.

Figure 3 shows the measured inter-packet arrival time for packets generated and received by the Fast Ethernet tester, with the ports directly connected (no intervening switch). The traffic was generated with a negative exponential distribution. The modulating structure in Figure 3 is due to rounding of the arrival time differences to the nearest microsecond. Figures 4 and 5 show the same distributions generated and received by the Gigabit tester. The total spread in Figure 4 is approximately 200 ns as noted earlier, which is to be compared with 40 ns for the Fast Ethernet tester. The structure in Figure 5 is due to rounding, as noted for Figure 3.

Figures 3 and 5 show the first time bin substantially higher than the expected exponential curve. This is an artifact which arises due to the true negative exponential distribution for inter-packet gaps, down to times corresponding to gaps which are illegal. The MAC always inserts the required inter-packet gap; therefore the first bin represents the sum of all inter-packet arrival times up to the legal minimum for the frame size being used.

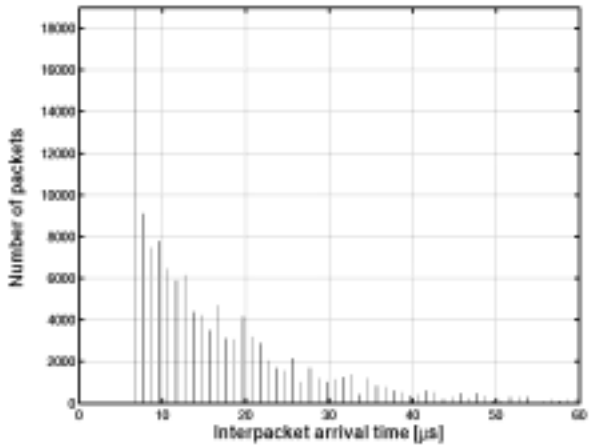

Fig. 3. Inter-packet arrival time for Fast Ethernet Poisson traffic, 64 byte frames, 12 s mean inter-arrival time (approximately 50% load).

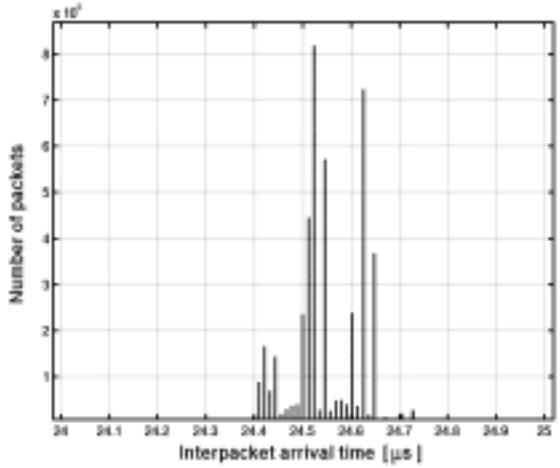

Fig. 4. Inter-packet arrival time for Gigabit Ethernet constant bit rate traffic, 1518 byte frames, 50% load

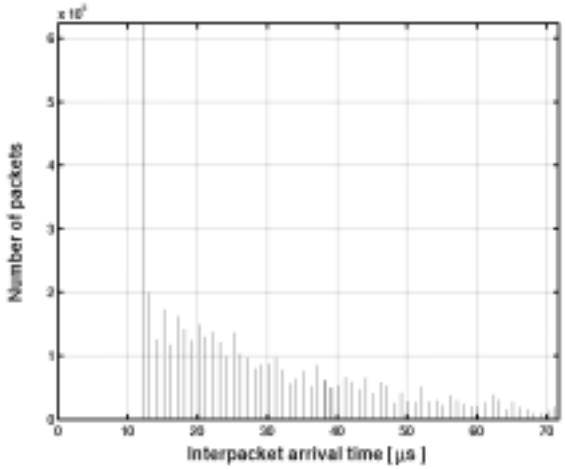

Fig. 5. Inter-packet arrival time for Gigabit Ethernet Poisson traffic, 1518 byte frames, 50% load

Both testers can be run from a graphical user interface capable of controlling and displaying results from up to 32 ports simultaneously.

VIII. Broadcast Traffic

Broadcast traffic is important for the ATLAS high-level triggers; its reliability needs to be understood quantitatively. We have measured the rate of loss of broadcast packets under

different load conditions in the network. The measurements were performed using only broadcast traffic, without flow control. Figure 6 shows the broadcast packet loss through an 8-port Gigabit Ethernet switch as a function of aggregate offered load, for constant bit rate traffic. An ideal switch should transmit 7 times the offered traffic (copies to all other ports). These measurements show that for frame rates above 100,000 frames/sec, the broadcast packet loss starts to become measurable. The switch used was an early generation Gigabit switch, which implements broadcast in software. Obviously this performance would not be acceptable in a network where broadcast needs to be heavily used.

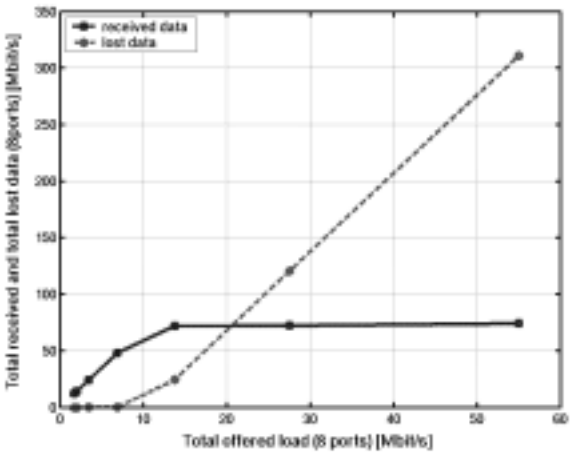

Fig. 6. Broadcast traffic through an 8-port Gigabit Ethernet switch, 64 byte frames, constant bit rate traffic.

IX. Quality of Service

Quality of service refers to the distinction between different traffic streams in order to guarantee priority for one of them. In the measurements performed here, VLAN (Virtual LAN) priorities are used to label the traffic. Flow control was turned off for these measurements. The measurements of packet loss whose results are shown in Figure 7 were performed using two Alteon NICs sending to a third Alteon NIC via a commercial Gigabit Ethernet switch. One sending node has been assigned high priority and the other low priority. The figure shows clearly that, while the aggregate offered throughput is less than the line speed of the destination port, both streams are delivered with no losses. Beyond this point, however, the low priority traffic is observed to be suffering losses while the high priority traffic is delivered without losses up to 90 per cent of capacity for the high priority stream.

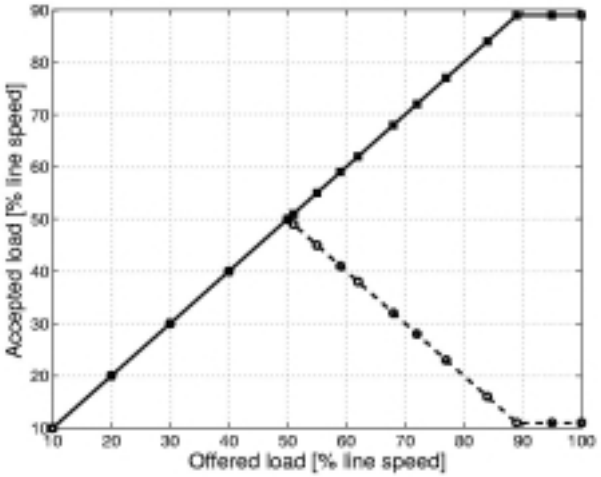

Fig. 7. Quality of service measurements for a Gigabit Ethernet switch. Two ports (one high priority, the other low priority) send to a third port. Accepted throughput is plotted as a function of the total offered throughput.

X. Fast Ethernet Tester as ROB Emulator

The Fast Ethernet tester can be made to emulate the network traffic generated by the Readout Buffers (ROBs). These devices receive requests originating in the LVL2 CPUs. They will respond to these requests with one or more frames containing data from their respective detectors. The FE tester can be reprogrammed to pass the relevant information from the incoming request frame to the RxMan FPGA which will then generate a transmit descriptor which is passed to the TxMan FPGA. TxMan will pass this descriptor to the MAC instead of fetching a descriptor from its descriptor memory. The descriptor contains sufficient information to generate the correct type, size, and delay for the outgoing response packet; the data payload, however, will contain no useful information except for the time stamp recovered from the incoming request packet. Copying this time stamp, which was inserted by the LVL2 CPU requesting the data, allows the CPU to measure directly the latency of the request-response process.

XI. Future Developments

A test bed consisting of eight FE testers (256 emulated ROBs) plus 64 PCs acting as supervisors and LVL2 farm nodes, together with a switch fabric, will provide a model of approximately 15 per cent of the entire ATLAS LVL2 system. In addition to measurement of packet loss and latency in the network, access to the MACs allows histogramming of internal queue occupancies under different conditions; this allows a quantitative measurement of the elasticity which exists in the system.

The Gigabit Ethernet tester will be expanded to multiple computer chassis in order to increase the number of ports. The FE and GBE testers will be integrated under one steering program, which includes a single graphical user interface as one option. The ROB emulator functionality described above is in the process of being implemented. The test bed described will be used to evaluate various candidate messaging strategies as well as switch fabric topologies.

REFERENCES

[1] R. W. Dobinson, S. Haas, K. Korcyl, M. J. LeVine, J. Lokier, B. Martin, C. Meirosu, F. Saka, and K. Vella, Testing and modeling Ethernet switches and networks for use in ATLAS high-level triggers, *IEEE Trans. Nucl. Sci*. vol. 48, No. 3, 2001.

[2] "ATLAS high-level triggers, DAQ and DCS," Tech. Proposal CERN/LHCC/2000-17, Mar. 2000.

[3] J. Bystricky and J.C. Vermeulen, "Paper modeling of the ATLAS LVL2 trigger system," ATLAS Internal Note ATL-DAQ-2000-030, April 2000.

[4] Alteon, Inc. http://www.alteonwebsystems.com/products/acenic/

[5] K. Korcyl, P. Golonka, and F. Saka, Modeling large Ethernet networks for the ATLAS high-level trigger system using parameterized models of switches and nodes , this conference, submitted for publication.

[6] Celoxica, Ltd. http://www.celoxica.com

[7] *Occam 2 Reference Manual*, Prentice Hall International Series in Computer Science, C. A . Hoare, Ed., Cambridge, 1988.

[8] Altera, Inc. http://www.altera.com