Network performance measurements as part of feasibility studies on moving an ATLAS Event Filter to off-site Institutes

Krzysztof Korcyl[1], Razvan Beuran[2,3], Robert Dobinson[2], Mihail Ivanovici[2,3], Marcia Losada Maia[2,4], Catalin Meirosu[2,3], and Grzegorz Sladowski[5]

[1] Institute of Nuclear Physics, Radzikowskiego 152, 31-342 Krakow, Poland
Krzysztof.Korcyl@ifj.edu.pl
[2] CERN - European Laboratory for Nuclear Research, CH- 1211 Geneve 23, Switzerland,
{Razvan.Beuran, Bob.Dobinson, Mihail.Ivanovici, Marcia.Losada.Maia, Catalin.Meirosu}@cern.ch
[3] "Politehnica" University of Bucharest, Faculty of Electronics and Telecommunications, B-dul Iuliu Maniu 1-3, sector 6, Bucuresti, Romania
[4] Federal University of Rio de Janeiro, Cidade Universitaria, Rio de Janeiro, Brazil
[5] Cracow University of Technology, Warszawska 24, 31-155 Krakow, Poland
gregs@plusnet.pl

Abstract. In this paper we present a system for measuring network performance as part of the feasibility studies for locating the ATLAS third level trigger, the Event Filter (EF), in remote locations [6]. Part of the processing power required to run the EF algorithms, the current estimate is 2000 state off the art processors, can be provided in remote, CERN-affiliated institutes, if a suitable network connection between CERN and the remote site could be achieved. The system is composed of two PCs equipped with GPS systems, CERN-designed clock cards and Alteon Gigabit programmable network interface cards. In the first set of measurements we plan to quantify connection in terms of end-to-end latency, throughput, jitter and packet loss. This will be followed by running streaming tests and study throughput, IP QoS, routing testing and traffic shaping. Finally, we plan to install the event filter software in a remote location and feed it with data from test beams at CERN. Each of these tests should be preformed with the test traffic treated in the network on the "best effort" basis and also when the traffic is sent via a 'dedicated' channel. The description of the system initially deployed in CERN-Geneva/Switzerland and Cracow/Poland is followed by results from the first measurements.

[6] Work supported in part by the grants of the European Union IST-2001-32243 and the Polish State Committee for Scientific research, SPUB nr 620/E-77/SPUB-M/CERN/P-03/DZ/295/2000-2002 and SPUB-M nr 620/E-77/SPB/5.PR UE/DZ 465/2002-2004

1 Introduction

ATLAS is one of four experiments planned at the Large Hadron Collider at CERN from 2007 onward. The primary proton-proton interaction rate will reach 10^9 interactions per second, this has to be reduced by seven orders of magnitude prior to data recording. To achieve such a reduction a three-level trigger system is being designed. The first, fully synchronous with the LHC collider reduces the initial rate of 10^9 interactions per second to 100 kHz. The second level, asynchronous, based on farms of commodity processors, executes sequences of trigger algorithms on selected data from the detector's buffers and reduces the rate of accepted events by an additional two orders of magnitude. Events classified as interesting by the second level trigger are sent to the Event Builder (EB), where the detector data scattered over thousands of buffers are combined together. The event data aggregated in the EB are sent to the third level trigger - the Event Filter (EF) - where the simplified off-line reconstruction algorithms, running on farms of processors, will reduce further the trigger rate by another order of magnitude. The final 100 Hz rate of events with average size of 2 MB/event will be sent to permanent storage. The task of the EB is to decouple further processing from detector buffer's occupancy. It alleviates requirements for event processing latency. The events do not need to be processed fast, however they need to be processed with the same rate as they arrive: 2 kHz. With an optimistic assumption of 1 second processing time, this requires access to at least 2000 processors.

A large processing farm will be built at CERN, where the experiment will take place, however using distributed resources would reduce the necessary local investments. Such option could be considered as there will be substantial computing power installed in many national centers for off-line analysis. Some of these resources should become accessible via the Grid technology - the Crossgrid project investigates this possibility.

To use home based computing equipment efficiently will require very high performance networking at an affordable price. Assuming an average event size of 2 MB moving half of the EF events to the remote sites will require 2 GB/s bandwidth. The GEANT network (Gigabit European Academic Network) and it's successors will be good candidates to carry such traffic. We need to estimate the impact on event latency of moving the Event Filter to remote locations and on the performance of the whole trigger system.

We plan to make measurements to investigate asymmetry in the QoS parameters of the network in both directions. The routes may not be the same but also the traffic competing for services may experience more congestion in one direction than in the other. We plan to measure the one-way latency (as opposed to taking a half of the Round Trip Time (RTT)), packet loss and packets reordering. The assesment of the asymmetry in the QoS parameters is important for our feasibility studies because the Event Filter traffic is highly asymmetric with the bulk of data sent from CERN to Cracow.

2 Setup for measurements

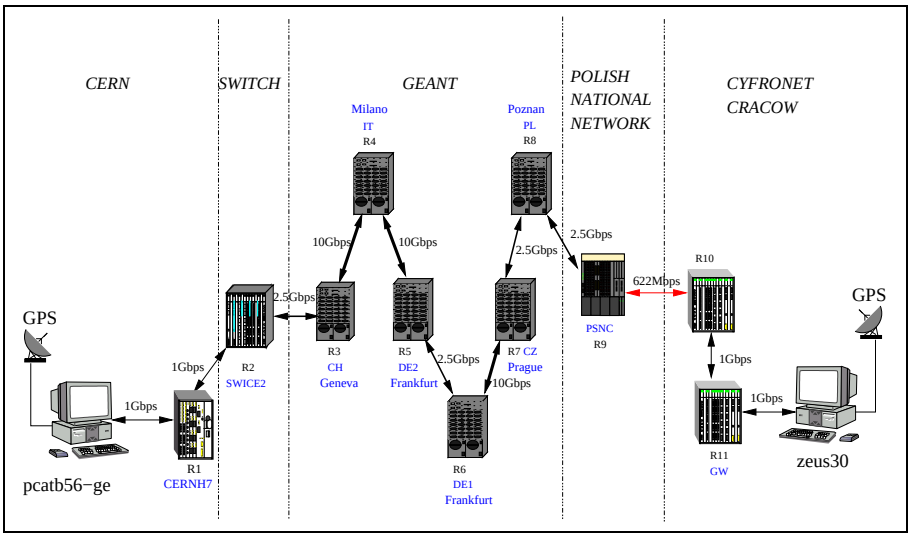

Fig. 1. Network path between CERN-Geneva-Switzerland and Cyfronet-Cracow-Poland

The setup to perform the network performance measurements is presented in Fig.1. Currently the system is installed at CERN-Geneva and in Cyfronet-Cracow. The system is composed of two PCs equipped with GPS systems [1], CERN-designed clock cards and Gigabit Ethernet Alteon programmable network interface cards [2]. The GPS system is used to synchronize time between the two PCs located more than 1000 km apart. The clock card is used to produce precise time stamps used to tag packets traversing the network. The programmable NIC is used to generate traffic according to traffic pattern descriptors as well as receive the time tagged packets and to make an on-line computation of latency and packet loss and throughput.

GPS system and clock cards The usual way to measure network latency and avoid synchronization problems between two nodes is to measure the Round Trip Time (RTT) of packets sent from one node and returned back by a partner node. The network latency is assumed to be half of the RTT value on assumption that the time from sender and receiver and from receiver back to sender contribute equally to the RTT. This may not be true for two nodes separated by more than 1000 km and connected to the Internet.

To overcome problems with synchronization of geographically separated nodes we use a Global Positioning System (GPS) to provide a universal time reference. The GPS signal is freely available everywhere on Earth and a GPS receiver unit

that has enough satellites in view, is able to give the Coordinated Universal Time (UTC) with accuracy in order of 100 ns. We are using off-the-shelf GPS receivers with the satellite signal received via exterior aerials and connected to the PCI bus of the computer. Each GPS card produces two output signals: 10 MHz clock and the PPS (Pulse-Per-Second) issued at every change of the UTC second with accuracy of 250 ns. These two signals are used by the CERN-designed clock card. The 10 MHz clock is multiplied on the clock card by 4 and used to update an internal counter. The 1 Hz PPS signal is used to reset the counter to a known value at each occurrence of the PPS signal ($40 * 10^6$ is added to the value of the counter at the previous PPS). Thus the clock synchronization system is based on two key points: the ability to reset counter at the same time (common start of the counters) and the ability to count exactly at the same rate (keep the counters synchronized). In this way the time difference between two synchronized systems is negligible (less than 500 ns) even after several days of running. This synchronization was verified for 2 systems on CERN site seeing the same satellites.

Programmable NIC The Alteon programmable network interface card gives us the possibility to create a flexible network traffic generator and measurement tool. We have developed a set of software routines together with a GUI [3], [4] which we use to prepare descriptor files with traffic pattern we want to apply to the network.

Prior to starting any tests, each card receives a traffic description table containing the full IP and Ethernet headers of the packets to be generated, the size of the Ethernet packets and the time between two consecutively sent packets. We thus have full control over all the fields in the IP and Ethernet headers (including Type of Service at the IP level and Priority from the VLAN field of the Ethernet packet). TCP and UDP were not implemented due to the increased overhead associated with these protocols. UDP requires the computation of a control sum on the data being transmitted and this would be too expensive for the processor on-board the NIC. The TCP requires the maintaining of a full history (sliding window) of packets being transmitted and received and the on-board memory is limited to 1MB. Also, TCP would be too computationally intensive to achieve Gigabit speed with the current on-board processors. Our traffic generator generates streaming traffic at the raw IP level. The content of the packets can be considered as being random.

The outgoing packets are time stamped with time values synchronized to the clock card. The NIC communicates with the clock card 128 times per second over the PCI to get its current value of the time. These readings are used to adjust the NIC's on-board counter which is used to mark outgoing packets. The packets are also marked with sequence numbers to allow packet loss calculation and out-of-order packet detection.

The NIC receiving traffic from the network keeps synchronization with its associated clock card. The on-board counter is used to calculate the latency of incoming packets. The on-board processor builds a real time latency distribution.

The histogram is transferred to the PC's host processor after the completion of the tests. This avoids sending individual packets to the host processor.

3 Measurements

For our tests we use the existing network infra-structure (see Fig.1) with two exceptions. The PC at CERN has been attached directly to the CERN router connected to the GEANT (bypassing the CERN internal network). The PC in Cyfronet has been attached to the router connected to the Polish national network (bypassing the Cracow metropolitan network). The tests were performed with the test traffic treated in the network on the "best effort" basis and also when the traffic was sent via an allocated channel of 100 Mbps on the Polish network (between Poznan and Cracow). In the tests aimed at latency and packet loss measurements we trasmitted IP packets generated by the Alteon cards between the two test sites. The payload used in these tests varied between 48 and 1500 bytes. The traffic pattern was with either constant bit rate (CBR) or Poisson inter packet time distributions.

The list of QoS parameters we measure is presented below:

- one-way average latency
- packet loss
- average throughput
- jitter
- latency histogram
- inter packet arrival time histograms
- IP packets re-ordering and quantification

4 Results

The summary of results from our measurements is presented in Table 1.

	CERN to Cracow	Cracow to CERN
with allocated channel	average latency almost constant zero packet loss (for loads smaller than 90 Mbps)	higher fluctuations in average latency small packet loss (grater for 64 bytes) (for loads smaller than 90 Mbps)
without allocated channel	average latency almost constant very small packet loss (for all rates)	not reproducible

Table 1. Comparison of the network QoS parameters between two directions

The measurements allowed us to observe the asymmetry in the network QoS parameters. The routes from CERN to Cracow and back pass through the same routers but the transfer latency fluctuates more and the packet loss is higher

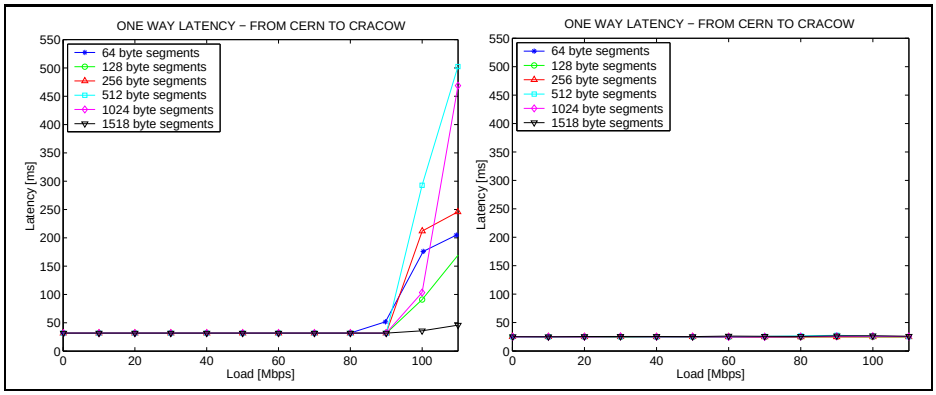

Fig. 2. Average latency as a function of load for packets transmitted from CERN to Cracow. The plot on the left shows results from measurements collected with the test traffic sent over the 100 Mbps allocated channel; the plot on the right for traffic treated in the network on the "best effort" basis.

in the direction from Cracow to CERN. This asymmetry is visible in the measurements with the test traffic treated by the network on the "best effort" basic as well as in the tests with the dedicated 100 Mbps channel. We do not have quatitative results for the test traffic sent from Cracow to CERN on the "best effort" basis as they were not reproducible. During our tests there was no IP packet re-ordering detected.

The plots with the average latency as a function of load for the traffic sent from CERN to Cracow in the "best effort" case and with the dedicated channel is presented in Fig. 2. The transfer latency for the "best effort" case is almost constant. The latency plot for the case with the dedicated channel shows that when approaching the 100 Mbps limit, the latency deteriorates. The similar degradation in performance when approaching the limit is observed for the loss packet ratio as a function of load, see Fig. 3. The rate of lost packets is zero up to 80 Mbps even for the smallest packets. However, reaching the allocation limit creates a large increase in lost packets. This is due to the contract of the channel allocation, where any excess of load above the limit is dropped by the switches/routers. The traffic send from CERN to Cracow shows a very small packet loss in the wide range of tested loads, up to 110 Mbps.

The transfer latency and the packet lost plots for traffic sent from Cracow to CERN with the dedicated channel are presented in Fig. 4. The average latency plot shows higher fluctuations comparing with the average latency of packets sent in the oposite directions. The packet loss is small with exception of 64 bytes packets where the loss increases for a load bigger than 50 Mbps. We want to emphasis here, that despite of having the allocated channel we were observing packet loss on the traffic sent from Cracow to CERN. For the same case with allocated channel there was no loss for traffic sent in the opposite direction.

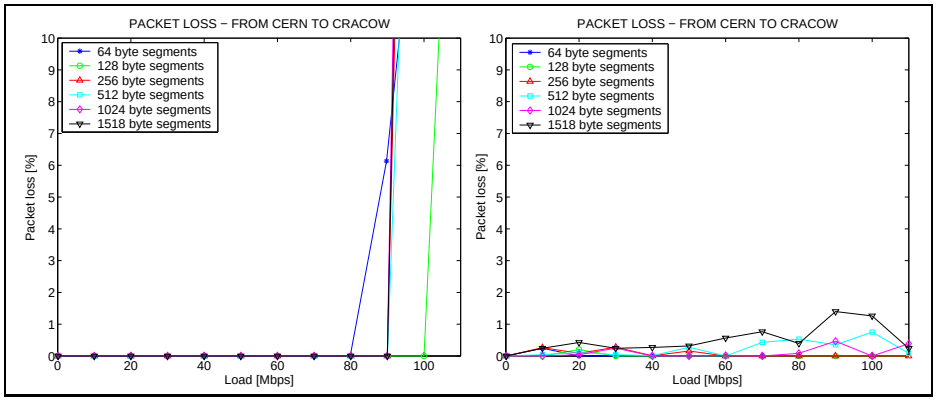

Fig. 3. Packet loss ratio as a function of load for packets transmitted from CERN to Cracow. The plot on the left shows results from measurements collected with the test traffic sent over the 100 Mbps allocated channel; the plot on the right for traffic treated in the network on the "best effort" basis.

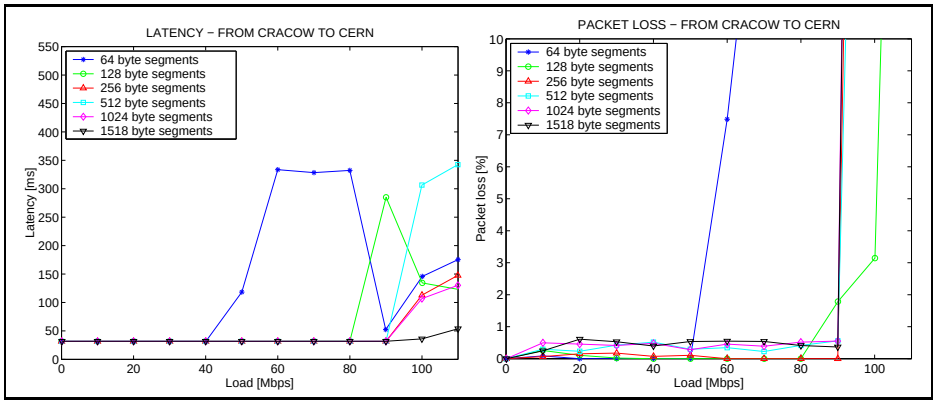

Fig. 4. The plot on the left shows results from the latency measurements as a function of load with the test traffic sent over the 100 Mbps allocated channel from Cracow to CERN. The plot on the right for shows packet loss ratio as a function of load for traffic sent over the 100 Mbps allocated channel from Cracow to CERN.

After network quantification in terms of QoS we begun tests with higher level communication protocol TCP/IP. We made preliminary streaming tests with TCP out-of-the box. Our results are summarized in Table 2

	with allocated channel	without allocated channel
average duration	579.66 s	567.18 s
standard deviation	12.34	28.37
average rate	13.8 Mbps	14.1 Mbps

Table 2. Streaming tests with out-of-box TCP/IP. 1 Gbyte of data transferred from CERN to Cracow using custom application.

5 Conclusions

The first set of tests showed the importance of measuring one way packet behavior as a function of load. In order to further understand our results a better understanding of layer 1 and layer 2 details is being sought.

We plan to continue streaming tests in both directions and tune the TCP/IP parameters to make better use of the allocated bandwidth. In the next step we will install the Event Filter software in Cracow and send real, experimental data from CERN to Cracow for processing. In parallel we will install similar systems in some other institutes. Measurements between CERN and the Niels Bohr Institute in Copenhagen will commence shortly.

References

1. GPS167PCI GPS Clock User's manual / Meinberg Funkuhren
2. Alteon WebSystems, Tigon/PCI Ethernet Controller rev 1.4, Aug 1997. Available: www.alteonwebsystems.com
3. "Testing and Modeling Ethernet Switches and Networks for use in ATLAS High-Level Triggers";
   Dobinson, R W; Haas, S; Korcyl K; Le Vine, M J; Lokier, J; Martin, B; Meirosu, C; Saka, F; Vella, K;
   in: IEEE Trans Nucl Sci.: 48 (2001) no. 3 pt. 1 pp607-12
4. "Testing Ethernet networks for the ATLAS data collection system";
   Barnes, F R M; Beuran, R; Dobinson, R W; Le Vine, M J; Martin, B; Lokier, J; Meirosu, C
   in: IEEE Trans Nucl. Sci.: 49 (2002) no. 1 pp.516-20