Network performance measurements as part of feasibility studies on moving 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], Catalin Meirosu[2,3], and Grzegorz Sladowski[4]

[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] 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 [5]. 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 Gigabit Alteon programmable network interface cards. We plan to perform three types of measurements: quantifying connection in terms of end-to-end latency, throughput, jitter and packet loss, running streaming tests and study throughput, IP QoS, routing testing and traffic shaping and finally installing the event filter software in a remote location and feeding it with real on-line data produced at test-beams at CERN. The description of the system initially deployed in CERN-Geneva/Switzerland and Cracow/Poland is followed by results from the first measurements.

1 Introduction

The trigger system for the ATLAS experiment is composed of three layers. 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

[5] 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

farms of commodity processors, executes sequences of trigger algorithms fetches interesting data from detector's buffers and reduces the rate of accepted events by another 10^2. Selected events are sent to the Event Builder, where the detector data, scattered over thousands of buffers, are combined together. The event data aggregated in the Event Builder are sent to the third level trigger - the Event Filter (EF) - where the 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 the permanent storage.

The task of Event Builder to decouple further processing from the detector buffer's occupancy reduces the contraints on event processing latency. The events do not need to be processed in a time critical way, however they need to be processed with the same rate as they arrive: $\sim$2 kHz. With the optimistic assumption of 1 second processing time, this requires access to at least 2000 processors. A large processing farm will be built at CERN, however using distributed resources would reduce the necessary local investments. Some of these resources, installed for off-line analysis, should become accessible via the Grid technology - the Crossgrid project is investigating 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 and moving half of the Event Filter events to the remote sites will require 2 GB/s bandwidth. The GEANT network and its successors will be a good candidates to carry such traffic. We need to estimate the impact of moving the EF to the off-site institutes on the performance of the whole trigger system.

2 Setup for measurements

The Fig.1 shows the system in its initial setup used to commision and test on CERN-intra network. Currently the system is assembled at CERN/Geneva and in Cyfronet/Cracow. The system is composed of two PCs equipped with GPS systems [1], CERN-designed clock cards and 1 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 patterns as well as receive the tagged packets and to make on-line analysis of latency.

GPS system and clock cards We use the GPS system as a time reference for one-way latency measurements because the traffic flow from CERN to Cracow and back may not be routed via the same paths making the standard RTT measurements inadequate. We use the 10 MHz clock and the PPS (Pulse Per Second) signals from the GPS card. The signals are fed into the home-made clock cards to reset them synchronously and correct their deviation. We measure time difference less than 500 ns between the two synchronized systems after several days of running.

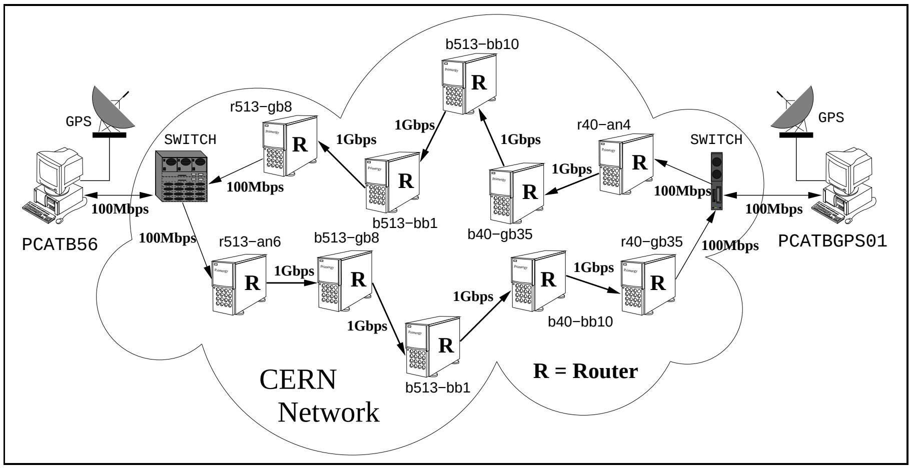

Fig. 1. System shown is an initial test setup installed at CERN site prior to being used for Trans European network connections

Programmable NIC The use of the Alteon programmable network interface card gives us the possibility to create a flexible network traffic generator and measurement tool. 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 [3], [4]. The time stamps, synchronized with clock card, are added at the very last moment. At the receiving NIC, the on-board processor collects traffic performance statistics including a latency histogram. The histogram is transferred to the host processor after completion of tests avoiding possible bottleneck on PCI with high rate transfers.

3 Measurements

We plan to run series of practical end-to-end tests using the existing network infra-structure i.e. passing from CERN to Cracow via the GEANT backbone and national and regional networks. This would quantify the present network and it would help us to identify our needs as well as current restrictions and bottlenecks. We plan to conduct three types of tests over the next few months:

1. Quantifying the connection from CERN to Cracow in terms of end-to-end latency, throughput, delay/jitter and loss with raw IP packets.
2. Run streaming tests using TCP/IP and perform studies on IP QoS, routing testing, traffic shaping.
3. Install prototype Event Filter software in Cracow and send from CERN online real data collected in the test-beam runs.

All three types of tests will be performed when accessing the network on the ”best effort” basis, then repeated with a certain bandwidth preallocated.

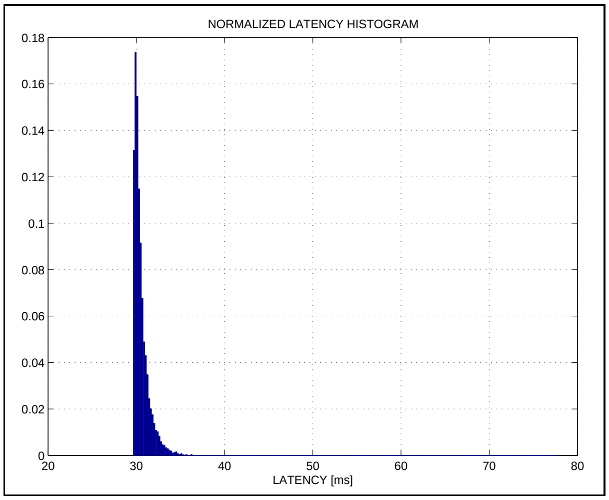

Fig. 2. Latency distribution for packets sent from CERN and collected in Cracow.

4 Results

The system was put into operation a couple of weeks ago. We tested it and learn how to operate using intra-CERN network. Very recently we performed the first data transfers between CERN and Cracow. Preparations are well advanced for running similar tests between CERN and the Niels Bohr Institute in Copenhagen.

The recorded latency of CERN to Cracow transfers is presented in Fig 2. The shape of the histogram can indicate that at the time of measurement the GEANT network between CERN and Cracow was very lightly loaded. To measure the one-way latency we were sending 100 packets per second. As the work is in progress any new measurements will be added to the presentation.

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