SIMULATIONS AND PROTOTYPING OF THE LHCB L1 AND HLT TRIGGERS

T. Shears*, University of Liverpool, Liverpool, UK
J. Anderson, T. Kechadi, R. McNulty, A. Smoker, University College Dublin, Dublin, Ireland
A. Barczyk, J. P. Dufey, B. Jost, N. Neufeld, CERN, Geneva, Switzerland

Abstract

The Level 1 (L1) and High Level Triggers (HLT) for the LHCb experiment are software triggers which will be implemented on a farm of approximately 1800 computers, connected via a Gigabit LAN with a bandwidth capacity of 7.1 GB/s and containing some 500 Ethernet links. The architecture of the readout network must be optimised to maximise data throughput, control data flow and minimise errors. We report on the development and results of two independent software simulations which allow us to evaluate the performance of various network configurations. We also describe the construction of two hardware testbeds of the LHCb L1 and HLT trigger system, which allow microscopic and macroscopic study of network and switch behaviour.

INTRODUCTION

LHCb, due to start taking data in 2007, is an experiment designed to study the matter – anti-matter asymmetry of the universe. CP violating decays of B hadrons must be identified and separated from other processes at the trigger level. As the bunch crossing rate is large (40MHz), a three level trigger system has been devised to reduce the volume of events written to disk to no more than 200 per second. The initial trigger, Level 0, is implemented in hardware and reduces the event rate to 1MHz. The following two levels, Level 1 and the High Level Trigger, are implemented in software to allow progressively more complex algorithms to identify events of physics interest.

A readout network is required to transmit data from the detector front ends to the computing farms where the trigger algorithms run. The network must be robust, have low latency, and be able to deliver the bandwidth requirements of the LHCb trigger system. This paper summarises current studies to simulate and prototype this network.

THE LHCB TRIGGER SYSTEM

A schematic diagram of the proposed trigger architecture is shown in Figure 1. Data must be sent from some 126 (323) detector front end sources for L1 (HLT) to 94 destinations. These destinations are computer clusters where both sets of trigger algorithms will run. Further details of the architecture can be found in [1].

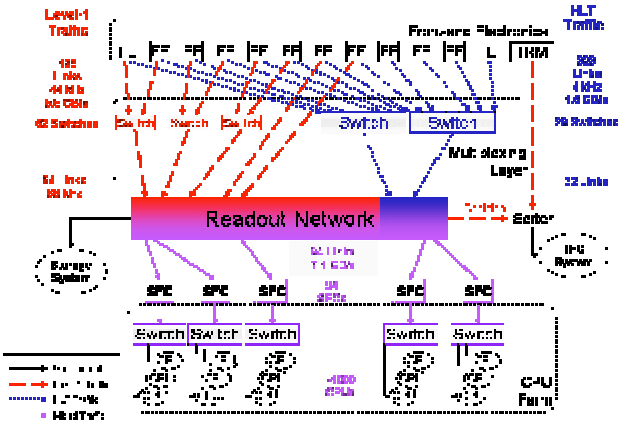

Figure 1: The architecture of the LHCb L1 and HLT system.

Event fragments from individual sources are small, so 25(10) L1 (HLT) events are packed together prior to transmission to reduce the message rate in the system. Typical event sizes for L1 and HLT are 5kB and 40kB respectively. With a L1 input rate of 1MHz the total bandwidth required for the system is 7.1GB/s. The L1 system alone requires a bandwidth of over 5GB/s; 125kB must be received each 25μs by one of the destinations.

NETWORK SIMULATION STUDIES

Two complementary approaches are used to simulate networks that are candidates for the readout network and evaluate their performance. We perform detailed software simulations based on parameterisations of data queuing, packing and switch performance, which allow us to understand the response time of a network at full scale. In parallel we perform hardware simulations where the performance of a real network or network component is examined at the scale of our testbeds. The latter approach is essential where details of switch fabric and logic are complex or unavailable and where parameterisation would be impossible. The data obtained may be input back to software simulation to extrapolate to full scale, or used as a performance evaluation in its own right.

Software Simulations; Custom Simulation

Our most mature software simulation is a custom built implementation in C that models the Banyan type network detailed in [1], where one layer of multiplexors

LHCb-PROC-2004-027
27/09/2004

and two layers of switches connect sources to destinations. Details of data packing at the sources, of data routing and queuing through the switches, and queuing and load balancing at the cluster destinations are all parameterised. Variable data sizes, or actual simulated L1 data for physics processes, may be sent from the sources.

The simulation has allowed many studies of network behaviour. For example, the inset plot in Figure 2 shows the time taken for a destination to reach a Level 1 decision. The main plot in Figure 2 translates this information into the fraction of events that would be affected if the available processing time was reduced. As an example, less than 0.5% of events exceed a cut-off of 30ms.

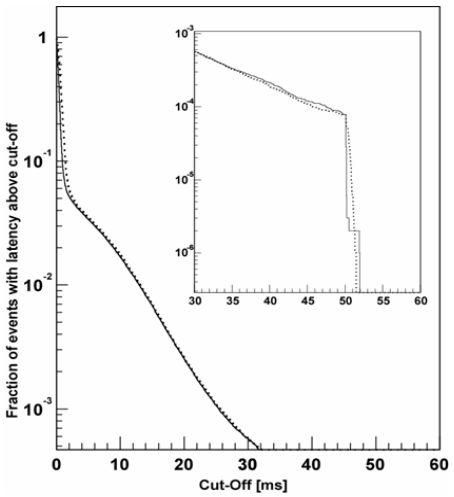

Figure 2: Fraction of events exceeding the maximum processing time as a function of time cut-off. The solid and dotted lines show results with and without data packing included, respectively. The time taken for a level 1 decision to be reached with no cut-off imposed is shown in the inset plot. (Note that simulated L1 algorithm processing times are limited to 50ms in the model, which gives the sharp drop around 52ms.)

Software Simulation: Ptolemy Simulation

An alternative simulation, based on Ptolemy [2] and implemented in java, is being developed to provide a more flexible method of evaluating network performance. Ptolemy allows the construction of network components, and indeed whole networks, from the combination of "actors" (implemented as java classes) within a graphical user interface. As configuration is straightforward, we expect that different network configurations or components can be modelled easily.

A version of the Banyan network of [1] has been implemented in Ptolemy, and switch buffer occupancies cross-checked against the custom simulation for a variety of traffic patterns and source data sizes at L1. So far all tests show that the custom and Ptolemy based simulations give similar results, thus increasing our confidence in the veracity of each simulation. Work is ongoing to extend the Ptolemy based simulation to evaluate other networks.

Hardware Simulations; Switch Testbed

Software simulations are only reliable if underlying assumptions and parameterisations regarding network components are accurate. This becomes a particular issue when including larger, more complicated, switches in simulations as details of switch behaviour, particularly under LHCb-style traffic patterns, may not be available.

We have addressed this concern by developing a testbed designed to measure switch properties. The testbed consists of 24 programmable data sources, each connected with a 1Gb/s link to a switch. The sources can be synchronised to within o(100)ns, which ensures that the simultaneous arrival of LHCb L1 input data can be replicated. Python based scripts are used to configure and run the system.

Switch buffer occupancy, latency, error and packet loss rate can all be measured in this testbed. As an example, Figure 3 shows the measured buffer size for one switch under test. The number of frames needed to fill the buffer as a function of frame size are shown, and translated into an effective buffer size of 4MB, which is expected for this particular switch.

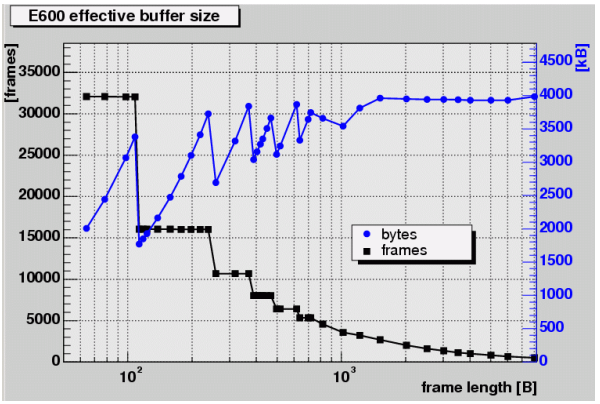

Figure 3: Effective buffer size of switch (blue points), and number of frames needed to fill buffer (black points), shown as a function of frame size.

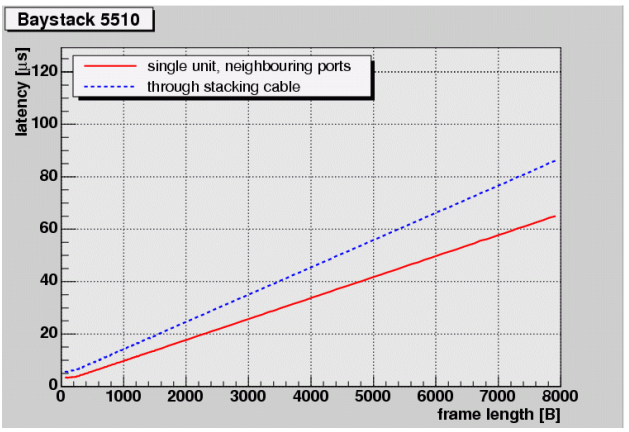

Figure 4: Switch forwarding latency as a function of frame size for transfers between ports on the same ASIC (solid line), and between ports on different ASICs (dotted line).

Figure 4 shows the switching latency for another switch. Latency is linear as a function of data size sent by the sources, and also negligible with respect to the L1 budget of 50ms - normal Ethernet frames would be sent within 20µs.

These results can be used as input to software simulations in order to extrapolate the behaviour to LHCb L1 and HLT scales and evaluate subsequent network performance. Additionally, any errors or packet loss observed during tests can be added to the simulation to make it more realistic.

Hardware simulations; Large Scale Testbed

As well as measuring the low-level response of switches, it is vital to study the overall behaviour of the readout network at scale in hardware using prototypes which attempt to replicate both the number of front-ends/destinations and the actual data volumes of the LHCb trigger.

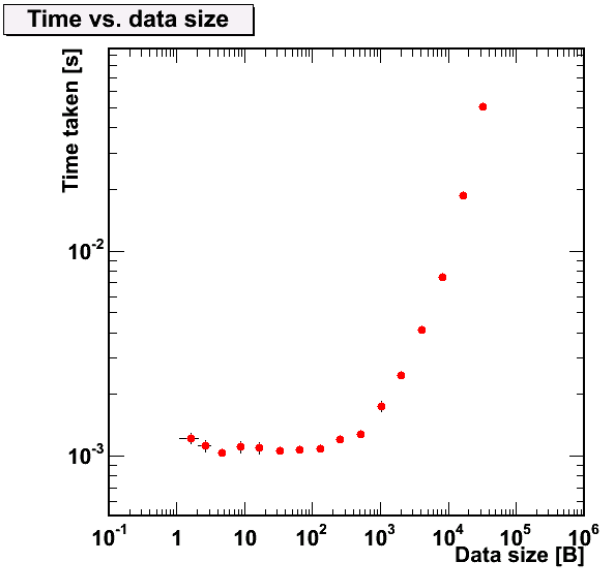

Figure 5: Time taken for 50 computer nodes to receive data, shown as a function of data size sent from each source. The increased time for data sizes above 1kB is due to the bandwidth limitation of the computer network interface cards used in this test.

A testbed which aims to study macroscopic network performance has been developed using the 940 node MAP2 supercomputer at the University of Liverpool. One hundred of the nodes are used and connected to a large (168 port) switch with 500 Mb/s of bandwidth per link. Each node can be configured as either a data source or destination. The test programs used to stress the network are implemented using MPI (Message Passing Interface[3]). These tests stress the network behaviour by using appropriate MPI synchronisation models between the sources and destinations. The time taken for destinations to receive data, for different traffic patterns and data source sizes, is recorded.

An example test, which mimics LHCb L1 data traffic patterns, uses half of the nodes as sources, and half as destinations. All sources send synchronously to each destination in turn, and the time taken for all destinations to receive recorded. Figure 5 shows this time as a function of data size sent by each of the sources. Results can be compared to LHCb L1 where each source sends approximately 1kB, and one of the destinations receives data every 25µs. We would expect 50 destinations to receive data in 1.25ms, which is consistent with the testbed results.

CONCLUSIONS

The Level 1 and High Level Triggers for the LHCb experiment are software triggers which will be implemented on a farm of approximately 1800 computers, connected via a Gigabit LAN with a bandwidth capacity of 7.1GB/s and containing some 500 ethernet links. The architecture of the readout network must be optimised to maximise data throughput, control data flow and minimise errors. We have developed a complementary suite of software and hardware simulations to predict, compare and evaluate network behaviour. Results from these simulations will enable us to choose the most appropriate readout network scheme for LHCb.

.

ACKNOWLEDGEMENTS

We would like to thank the organisers of the conference for providing a forum in which to present this work, and the Royal Society for their financial support. We also thank Prof. T. Bowcock and Dr. A. Washbrook for their help in configuring the large scale testbed on the Liverpool MAP2 supercomputer.

REFERENCES

[1] LHCb Trigger System Technical Design Report, CERN/LHCC 2003-31

[2] http://ptolemy.eecs.berkeley.edu/ptolemyII/

[3] See, for example:

William Gropp, Rusty Lusk, “Tuning MPI Applications for Peak Performance”, Pittsburgh, (1996);

William Gropp, Ewing Lusk, Anthony Skjellum, “Using MPI, portable parallel programming with the Message Passing Interface”, The MIT Press, Cambridge, Massachusetts, London, England, (1994);

Pacheco S. Peter, “Parallel Programming with MPI”, San Francisco University, Morgan Kaufman Publishers, Inc., San Francisco, California, (1992).