Ethernet for the ATLAS Second Level Trigger?

R.W. Dobinson, S. Haas and B. Martin,
CERN
1211 Geneva 23
Switzerland
and
M. Dobson and J.A. Strong,
Royal Holloway and Bedford New College,
University of London,
Egham, Surrey, UK

ABSTRACT

The use of Ethernet in implementing the ATLAS second level at the CERN Large Hadron Collider (LHC) is considered. Tests carried out on a large network testbed using 100 Mbps DS links and switches are used as a reference point. The evolution and present status of the IEEE 802.3 Ethernet standard is reviewed and technical problems in exploiting the technology at the LHC are explored.

INTRODUCTION

The required network performance of the ATLAS second level trigger has been largely achieved using IEEE1355 serial point-to-point link and switch technology. However, this does not offer a commercially viable solution on the timescales of LHC. The IEEE 802.3 Ethernet standard is rapidly evolving to higher levels of performance and functionality. The possible use of the standard for building the second level trigger system for the ATLAS LHC experiment is discussed within the framework of its present status and future prospects.

THE ATLAS SECOND LEVEL TRIGGER

The ATLAS second level trigger [1] involves the reduction of an incoming rate of 100 kHz by about a factor of 100. This requires the processing of data coming from about 1500 readout buffers by a similar number of processors. The simplest model is one of farming of individual events to single processors [2], although more complicated schemes are also being considered based on the concurrent processing of different sub-detectors [3]. Most proposed schemes depend on the efficient functioning of an interconnect fabric that links the readout buffers to the processors and scales in both size and performance.

The incoming data volume per event is estimated to be 1 Mbyte of which typically 10% must be processed by the second level trigger. This yields a requirement for a sustained data transfer rate of 10 Gbyte/s between buffers and processors. In addition, studies have shown that the CPU loading of the processing nodes and buffers, as well as the trigger decision time, depends strongly on the communications overhead incurred in inter node message passing [4] [5].

THE MACRAME SWITCHING TESTBED

As part of the European Union funded Macramé R&D project a very large switching fabric has been constructed and evaluated at CERN [6] [7]. Up to 1024 end nodes can be preloaded with a predetermined traffic pattern and the response of the network in terms of throughput and latency determined as a function of different topologies and traffic conditions. The overhead for sending a packet is less than 1μs. Topologies studied include Clos, grid and torus networks. Figure 1 shows the construction of a 256 Clos network.

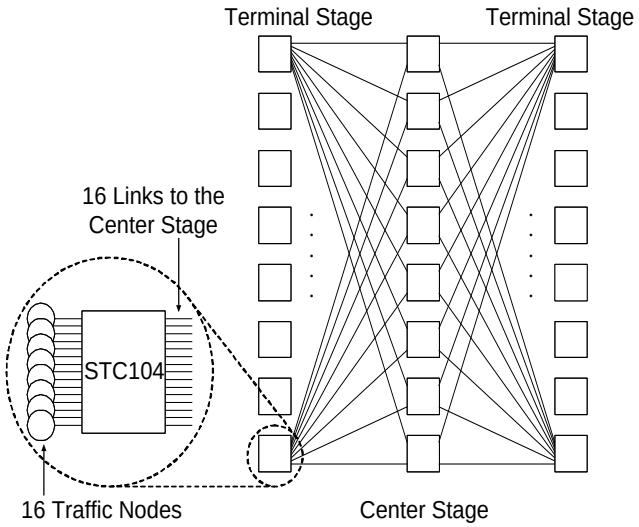

Figure 1: A 256 Clos Network

The interconnect technology used is IEEE 1355 [8] Data-Strobe (DS) links running at 100 Mbps and switched through 32 port STC104 packet switches [9]. DS links provide full duplex serial data transmission with low level flow control. The STC104 uses worm hole routing, whereby a packet is switched from an input port to an output port as soon as the destination address in the packet header has been received. This results in very low switch propagation times, about 1 μs.

Results obtained with the Macramé testbed show that very large switches can be built with very high reliability, no transmission errors have been detected in running the system over periods of several days.

Figure 2 shows the results obtained for the throughput of Clos and grid networks of different sizes. Systematic traffic involves fixed pairs of nodes sending to each other, for random traffic, nodes choose a destination from a uniform distribution. Clos networks offer the best performance, saturating at about 50% of the maximum node link bandwidth for random traffic. They show good scalability in terms of network throughput versus size.

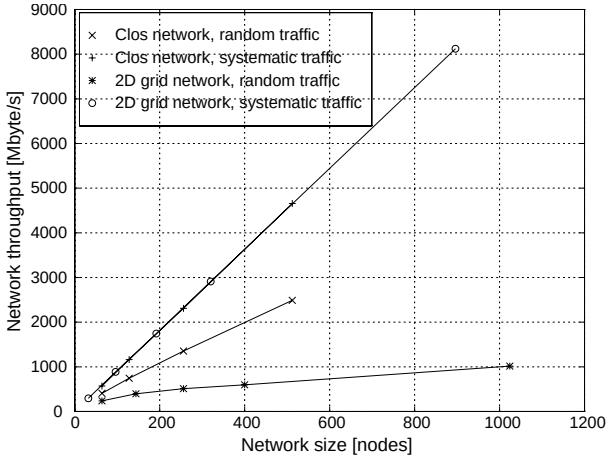

Figure 2: The throughput versus network size for Clos and grid networks under systematic and random traffic with 64 byte packets

Figure 3 shows that the overall packet latency through a network remains low up to conditions near saturation.

ATLAS second level trigger data patterns for different subdetectors have been transmitted through the testbed at rates approaching and in most cases exceeding the 100 kHz required [6]. In the tests carried out individual buffers and processors are emulated by terminal nodes on the network. The traffic applied to the network is characterised by the collection of data from three to twenty buffer nodes by one processor node. Several processors may be receiving data concurrently during the same event. At high rates the system throughput is not limited by the available link bandwidth but by network congestion.

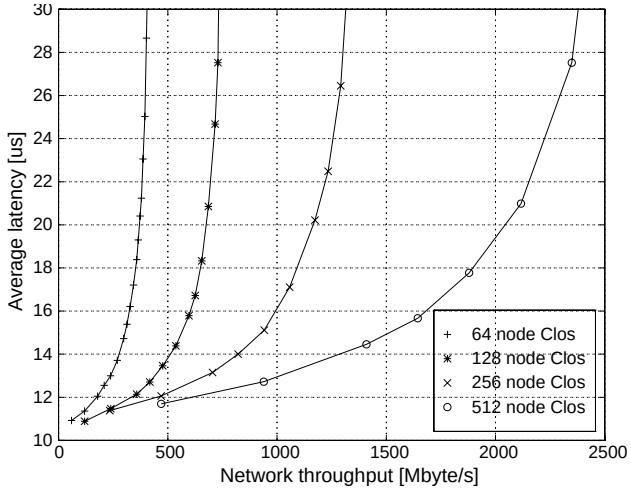

Figure 3: Latency versus throughput for Clos networks under random traffic with 64 byte packets

Destination contention is inherent to this type of traffic in which packets queue for access to the processor node. This causes congestion at the destination port and blocked packets continue to occupy switch ports in the terminal stage as well as in the center stage and, depending on packet length, all the way back to the buffer node. These queued packets occupy ports in the switch fabric and reduce the paths available to other packets. The presence of a stalled packet in the buffer node means that the packet behind it cannot be sent even if its destination is free. This is referred to as head-of-line blocking.

The presence of multiple paths across a Clos network between any two points reduces congestion. The STC104 supports a locally adaptive routing scheme which allows packets to be sent down any free output link in a programmed set of consecutive links. This improves performance by ensuring that there are no packets waiting to use one link when an equivalent link is idle. A set of links used to access a common destination can therefore be logically grouped together, increasing the aggregate throughput to the destination. This grouped adaptive routing allows efficient load-balancing in multi-stage networks.

Using grouped adaptive routing with ATLAS trigger traffic can give significant performance improvements. For example, for one of the subdetectors, the achievable event rate was increased from 50 kHz when using deterministic routing to 125 kHz when adaptive routing was applied.

The work carried out with the Macramé testbed has shown that very large multi-stage switching fabrics, an example of which is shown in Figure 1, can be constructed using 100 Mbps DS links. Furthermore this type of interconnect has demonstrated its ability to handle ATLAS level 2 type traffic. Therefore it is natural to consider building similar large network architectures based on switched 100 Mbps Fast Ethernet.

THE EVOLUTION OF ETHERNET

The original implementation of Ethernet used a 10 Mbps coaxial cable shared between multiple nodes using CSMA-CD packet transmission (IEEE 802.3). This system required half duplex operation. Subsequent developments involved twisted pair wiring to a central hub, while retaining the original concept of multiple nodes on a shared bus or segment (IEEE 802.3i).

In order to support demands for higher throughput, networks may be partitioned into two or more independent segments. Segments are joined in pairs using bridges (IEEE 802.1D). Traffic local to a segment is restricted to that segment whereas traffic destined for remote segments is passed across bridges until it arrives at its final destination. Bridges are given the ability to automatically learn the required routing of packets for nodes in the network. The learned routing scheme must not introduce unwanted duplicate packets and must avoid loops

Recently there has been an evolution of Ethernet to 100 Mbps (Fast Ethernet, IEEE 802.3u). This has been accompanied by a move away from shared segments towards point-to-point links connected by switches. Point-to-point links improve performance by avoiding collisions and offer full duplex operation. Packet based flow control may be implemented using pause packets to signal congestion on a switch port and thus avoid losses due to buffer overflow (IEEE 802.3x).

The latest development in what has been a very rapid extension of the standard over the last few years is the emergence of Gigabit Ethernet (IEEE 802.z) supporting data transmission rates of 1 Gbps. Although not yet formally ratified, the Gigabit Ethernet standard is well advanced and products are becoming available.

There are numerous books and articles on Ethernet. Reference [10] gives a short and informative review of the historical development of the standard.

THE POSSIBLE USE OF ETHERNET IN THE ATLAS SECOND LEVEL TRIGGER

Perceived advantages

The installed base of Ethernet is enormous, in its 10 Mbps form it has totally dominated the desk-top market of PCs, workstations and servers. It is extremely unlikely that it will be dislodged from this position as *the* commodity interconnect. Ethernet equipment is routinely bundled by computer manufacturers in their products and there are numerous component suppliers.

The highly competitive market assures low prices. Network interface cards (NICs) and switch ports for the 10 Mbps version are as low as $25 and $80 respectively, the equivalent numbers for Fast Ethernet are $60 and $200. Prices of Gigabit products, which are just now being introduced, are high, several thousands of dollars. However, experience with Fast Ethernet showed very rapid price decreases in the years following initial introduction, similar reductions of about 30% a year are likely to occur in the Gigabit market.

The prognosis for IEEE 802.3, the view of many IT suppliers, is that 100 Mbps switched Fast Ethernet will dominate the market place within a few years and that the Gigabit standard will fall in price sufficiently rapidly to allow its widespread use on the desk top.

When choosing an interconnect for LHC applications it must be remembered that the start up date for the machine is 2005 and that the lifetime of equipment installed is likely to be in excess of a decade. The advantages of using Ethernet, a standard interconnect whose future use and evolution can be seen into the next century, are therefore considerable.

Given that the cheapest way to buy computing power for LHC applications might be through the use of PC products, then triggering solutions using parallel computing systems totally synthesised from commodity components can be envisaged. However, this is not to say the use of Ethernet in trigger systems is inevitable. Its use in some areas, in particular in high performance trigger systems, is not without problems and these we will try to address briefly in the rest of this paper.

Baseline measurements for Fast Ethernet

In order to establish a baseline, a series of measurements have been carried out using two directly connected 200 MHz Pentium PRO PCs running LINUX. Different size messages were transferred between application programs in the two PCs using the standard TCP/IP socket mechanism. The results are shown in Figure 4.

For small message sizes, 10 bytes or less, such as would be used for control in the ATLAS second level trigger, the data throughput is very poor due to the software overheads. A measurement of CPU utilisation at this point showed nearly 45% occupancy. Longer messages are less heavily penalised but even for messages of 1kbyte, which correspond to typical data transfers between buffers and processors, the achieved throughput is only 88% of the maximum and a measurement of the CPU showed that it was still 25% loaded.

Figure 5 shows the elapsed time in sending a message between the two application programs in the two Pcs. It is

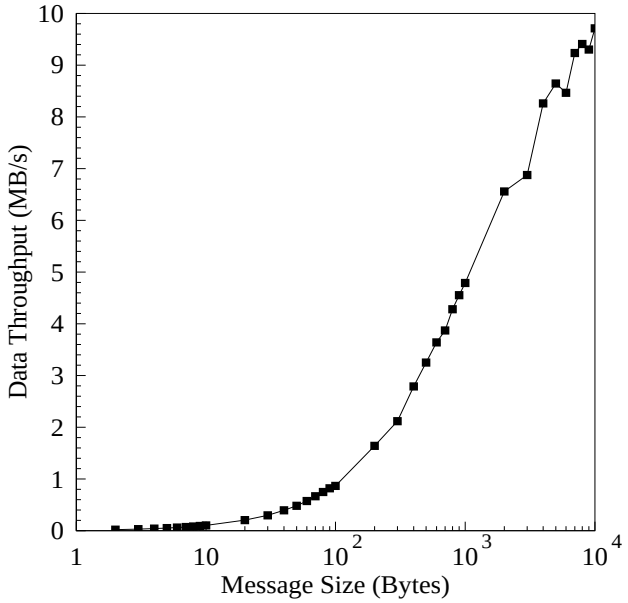

Figure 4: The throughput between two directly connected PCs using TCP/IP sockets

defined as half the round trip time for an echoed message. Results for directly connected PCs are compared with those obtained with a store and forward type switch. In both cases the curves are linear, with an increasing divergence between them, showing the switch latency is data size dependent. The fixed overhead in sending a message due to the overheads in the two PCs is 92 μs.

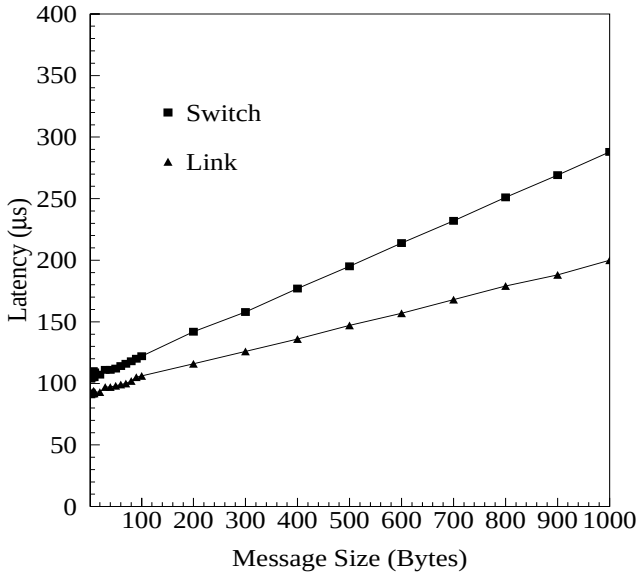

Figure 5: Elapsed time for message transfer for a direct connection (Link) and a switched connection (Switch)

The switching fabric

The ability of commercial Ethernet switches to dynamically learn the required routing of packets between source and destination nodes is one which is valuable in setting up and administrating large networks.

However, the learning process imposes topology constraints prohibiting loops in the network, thus only a single connection is permitted between any pair of adjacent switching elements in the network.

Packet transmission between nodes attached to the same switch is limited only by the internal capabilities of that switch, whereas the communication performance between nodes on different switches is bounded by the single link restriction. Therefore dynamic learning precludes the construction of multi-stage networks with Clos style interconnection schemes.

There are three ways round this problem. Firstly, the automatic learning process can be disabled and the switch routing tables pre-loaded; the tables are then static as opposed to dynamic. Secondly, the problem of inter-switch bandwidth limitation can be addressed by using higher speed links; using Gigabit Ethernet helps alleviate the problem. Lastly, several physical connections between two switches can be treated as a logical group and administrated as such, thus effectively increasing the aggregate inter-switch bandwidth. This option is already offered by several manufacturers, although each has their own proprietary way of doing this.

The use of store and forward packet routing in most Ethernet switches will increase the network latency significantly compared to the worm hole (or cut through) routing used in the STC104. In the former case the whole of a packet must be buffered before it is transferred from the input to the output switch port as seen in Fig 5. For a minimum length packet this delay time is 6.4 μs for maximum length packets it is 151.8 μs.

Although some Ethernet switches are supplied with a cut through option, the majority use store and forward. The reasons for this are two fold. Firstly, the store and forward approach allows packet integrity to be checked before being passed through a switch and packets in error may be rejected. Secondly, installations may include both 10Mbit and 100Mbit segments, as well as the new Gigabit technology. As gaps in Ethernet packet are not permitted, transmission between segments running at different speeds requires packets to be buffered. Therefore, in a market dominated by store and forward switches, a relatively long latency through a large network seems to be unavoidable. The implications of this are discussed later.

Gigabit Ethernet offers further options and enhancements to an interconnection scheme based on 100 Mbps technology. As has been previously discussed, it can act as a high performance network backbone linking together Fast Ethernet switches. In addition, it may be used to interconnect clusters

of readout buffers and processors, thus reducing the overall size of the network required and its latency.

System Design issues

The interface between network nodes and the network is a crucial consideration in building high performance systems. It involves the careful design of both the Network Interface Card (NIC) and the host software.

The commercial exploitation of Ethernet is based on the use of the TCP/IP protocol stack supports general purpose communications over local and wide area networks. The performance of these protocols depends on a number of factors including the host processing speed, host operating system, NIC design and overall implementation strategy. Baseline measurements presented previously suggest that for application in the ATLAS second level trigger the TCP/IP stack should be removed and replaced by a strategy more geared to parallel programming built directly on top of the Ethernet packet level interface.

The goal will be to obtain high data transfer rates between applications programs in different network nodes with minimum elapsed time and host CPU loading. The mechanisms for achieving this are well established [11], but not always implemented in commercial products. They may be summarised as follows:

- minimise the number of host interrupts
- avoid memory to memory copies
- avoid time consuming operating system calls and context switches
- implement light weight communications protocols and a simple application programming interface (API)

In all of the above considerations there is little which is Ethernet specific. Whatever the interconnect used in the ATLAS second level trigger these topics will have to be addressed. However, the use of Ethernet will have to take account of the relatively long network latency. The solution will be to overlap computation and communication in such a way that network latency is hidden. Simply stated this will require that a processing node is kept busy carrying out computation on one or more events while subsequent event data is being acquired. The implementation of such a scheme would be facilitated by an efficient multi- processing kernel, with low context switching overheads, running on the host, closely coupled to a customised intelligent communications controller. The latter being responsible for data collection and management.

CONCLUSIONS

We have briefly reviewed the needs of the ATLAS second level trigger and what has been learned from the Macramé switching testbed about building large high performance networks. The present status and future prospects for Ethernet were summarised and potential advantages for use at the LHC outlined. We conclude that, although there are technical problems to be addressed, none of them appear insoluble given the rapid evolution of the technology and its wide commercial support in the IT industry.

ACKNOWLEDGMENTS

We are grateful for the support of the European Union through the Macramé (ESPRIT project 8603) and ARCHES (ESPRIT project 20693) projects.

REFERENCES

[1] “The ATLAS Technical Proposal”, CERN/LHC/94-43, LHCC/P2,
ISBN: 92-9083-067-0.

[2} D. Calvert et al, “Operation and Performance of an ATM Based Demonstrator for the Sequential Option of the ATLAS Trigger”, Xth IEEE Real Time Conference 97, Beaune-France, September 22-26, 1997.

[3] J.R. Hansen et al, “Local-Global Demonstrator Programme for the ATLAS Second Level Trigger”, Xth IEEE Real Time Conference 97, Beaune-France, September 22-26, 1997.

[4] M.Dobson et al. “Paper Models of the ATLAS Second Level Trigger”, ATLAS Internal Note, draft DAQ note, November 27th, 1997

[5] J.C. Vermeulen et al, “Performance Requirements of Proposed ATLAS Second Level Trigger Architectures from Simple Models”, Xth IEEE Real Time Conference 97, Beaune-France, September 22-26, 1997

[6] M. Zhu et al, “Realisation and Performance of IEEE 1355 DS and HS Link Based, High Speed, Low Latency Packet Switching Networks”, Xth IEEE Real Time Conference 97, Beaune-France, September 22-26, 1997

[7] S. Haas et al, “Results from the Macramé 1024 Node IEEE 1355 Switching Network”, Presented at EMMSEC97, European Multimedia, Microprocessor and Electronics Conference, Florence, Italy, 3-5th November, 1997.

[8] IEEE Std. 1355, “Standard for Heterogeneous Inter-Connect (HIC). Low Cost Low Latency Scalable Serial Interconnect for Parallel System Construction”, IEEE Inc., USA 1995

[9] “The STC104 Asynchronous Packet Switch”, Data Sheet, April 1995, SGS Thomson Microelectronics.

[10] “Switched and Fast Ethernet”, Breyer and Riley, ZD Press, 1996, ISBN 1-56276-426-8.

[11] M.Boosten, R.W. Dobinson, B. Martin, “The Network Interface Bottleneck”, Xth IEEE Real Time Conference 97, Beaune-France, September 22-26, 1997.