Results of Benchmark Tests upon a Gigabit Ethernet Switch

LHCb Technical Note

Issue:	Draft
Revision:	1
Reference:	LHCb DAQ 2001-nn
Created:	13th August 2001
Last modified:	17th August 2001

Prepared By: **Jean-Pierre Dufey, Beat Jost, Nikolaus Neufeld and Konrad Paszkiewicz**

Abstract

A short report is given detailing the results of benchmark tests performed on the "Fast Iron" 4000/8000 series of Gigabit Ethernet switches from Foundry Networks. This switch, or one similar in basic design, is under consideration for use in the LHCb DAQ Readout Network. The tests were carried out using Alteon Tigon 2 based network cards running on standard PCI buses. The results presented here are intended for use in simulations to enhance and refine the proposed network topology, in particular the scalability of the central switching network.

We find that the switch contains usable output buffer space of approximately 1MB per blade, capable of storing approximately 1024 frames. Additional tests were carried out to determine typical packetisation lengths and also latencies within the switch, network cards, and backplanes. These tests show structures typical of packetisation at 64 Byte intervals with no apparent Head of Line Blocking. Typical latencies per frame introduced by the switch are dependent on the length but are on the order of 0.5 µs.

Table 1 Document Status Sheet

1. Document Title: Results of Benchmarks for a Gigabit Ethernet Switch			
2. Document Reference Number: LHCb DAQ 2001-nn			
3. Issue	**4. Revision**	**5. Date**	**6. Reason for change**
Draft	1	13 Aug 2001	First version

Table of Contents

List of Figures

List of Tables

1 Introduction

LHCb imposes a somewhat heavier demand upon the RN than many experiments because the data processing algorithms at level 2 depend upon all information from relatively few sub-detectors, rather than on data containing fragments of information on all sub-detectors. Even if one takes into consideration the technology likely to be available at the planned time of installation, the estimated 4 to 6 GB/s in the Readout Network (RN), is challenging. When one takes into consideration possible future upgrades of the detector and/or LHC beam and the consequent inevitable increase in data rate, the need to ensure a reliable, upgradeable DAQ infrastructure becomes a necessity.

Evidently it is of great interest, to optimise the topology of the RN to ensure that the available bandwidth is used efficiently and leaves significant headroom for any unforeseen overheads. To test these various topologies it is necessary to simulate the conditions the system is likely to encounter. However, in order to simulate such topologies it is of course necessary to have information regarding the performance of the network components in use.

The central switch is the crucial element when the scalability of the system is concerned. This is clear from the basic layout from the LHCb DAQ system, where it is evident that the switch is the only central, shared component, all others being independent.

Unfortunately we do not have a first principles insight in the inner workings of commercial switches, nor the accompanying detailed technical specifications. Such information is usually kept secret by the manufacturers. We have therefore to rely on measurements and educated guesses to obtain the parameters for a simulation model of the switch.

This report details the results of benchmark tests upon a FastIron 4000/8000 Foundry Gigabit Ethernet Switch, which are designed to enable the accurate simulation of the Readout Network using such a switch.

Section 2 provides some more detailed background on the proposed topology and its advantages together with some expected data rates, this section is only intended as a brief introduction to the subject of topologies and is included mainly to underline the importance of the switch and may well provide the reader with too little information in which case they are advised to seek out the references. Whilst Section 3 sets out the methodology used to benchmark the switch together with some details of the ‘in house’ developed LHCb Fragment Generator and section 4 presents the results while section 5 gives a summary of the results. No conclusions as such are presented here – these measurements are only half the picture and one must wait for the results of the topological simulation before one can know whether the proposed topology can perform to expectations.

2 Background & Requirements

2.1 An Overview of The Proposed Ethernet Readout Network Topology

To provide a general overview of the DAQ system and the data rates involved Figure 1 may prove helpful.

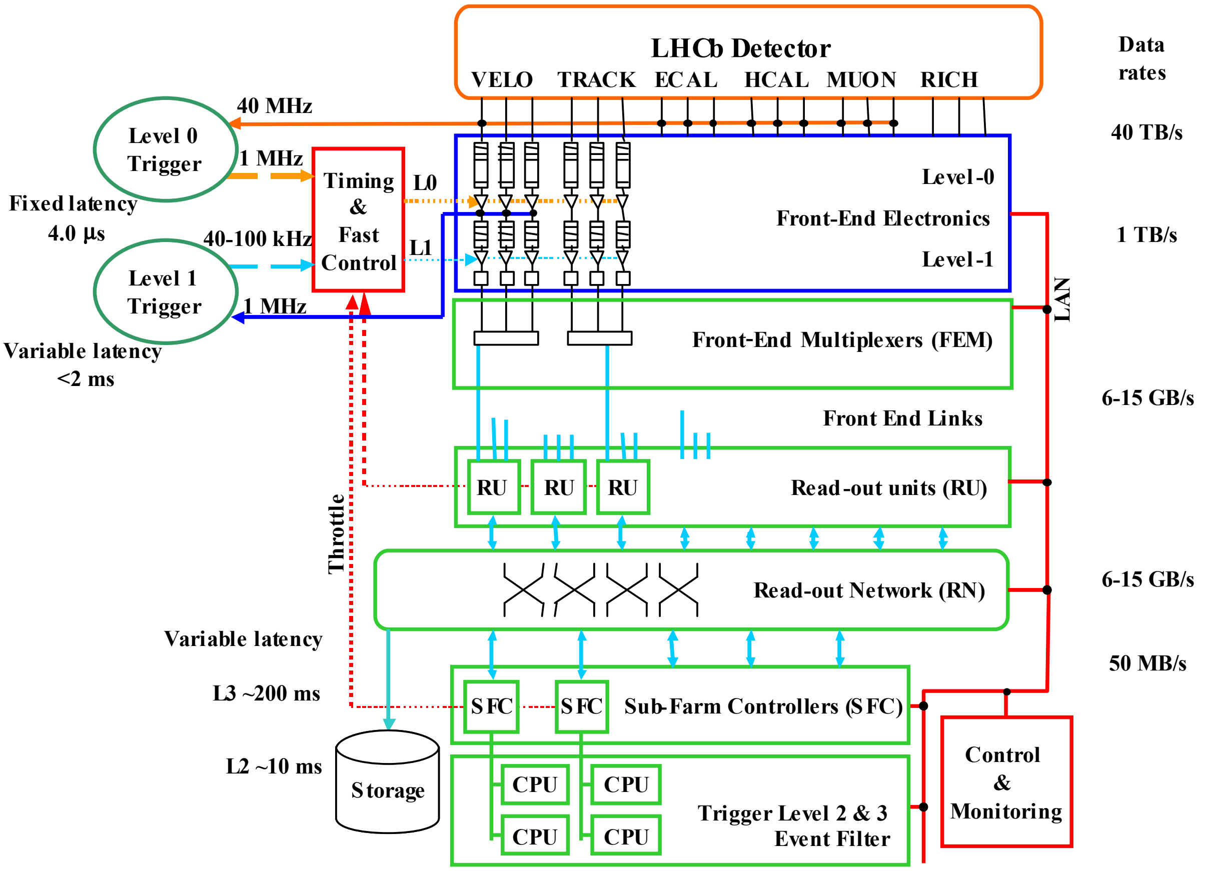


Figure 1 Basic Architecture of the LHCb DAQ System

A typical Readout Network could consist of some 128 RUs connected via Foundry switches to a series of some 128 Sub-Farm Controllers (SFCs). Ordinarily nxn switches would be connect in what is called a ‘Banyan’ network (Figure 2

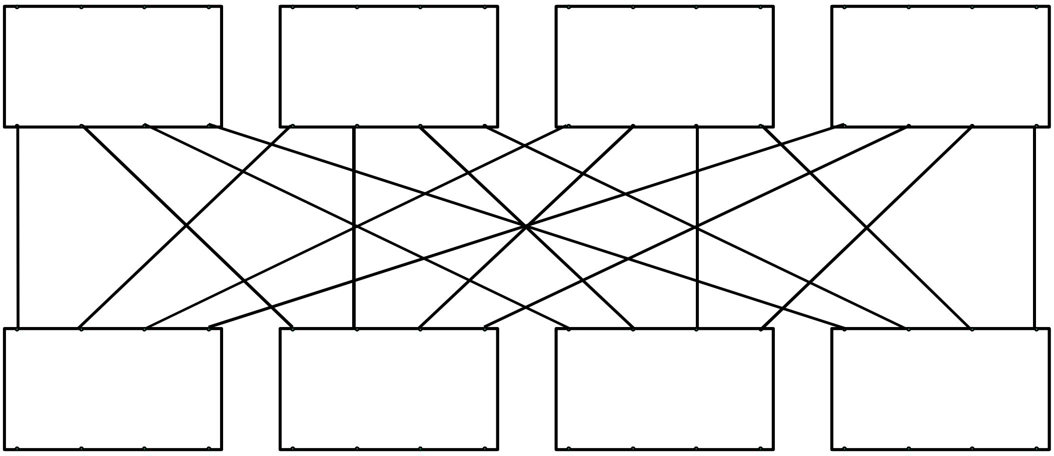


Figure 2 : Banyan configuration for a readout network. Each interconnecting line represents 15 links.

2.2 Real World Effects

To ensure that a specific switch topology can work, it is of course necessary to perform small-scale tests on such vital components as the switch in order that one may accurately model the larger, full-scale behaviour of the topology within a computer simulation [2].

A network switch's performance can vary considerably with both packet size and arrival rate. Even across switching across different blades within the same switch can have a major effect. Buffer sizes and associated flow control may also drastically alter the dynamics of a such a multi-gigabit speed network. Thus the measurement of these parameters is crucial in building a robust and accurate simulation of the future Readout Network. The next section details the methodology by which the measurements of the above parameters were made and provides some brief technical detail on the test-bed switch and the Network Interface Cards (NICs) employed.

3 Benchmarking Software & Methodology

3.1 Hardware Equipment

3.1.1 The Network Interface Cards

These consisted of two standard copper based Gigabit Ethernet NICs produced by Netgear and based on the Alteon Tigon 2 chip. They have two onboard processors running at approximately 86 MHz. To facilitate these tests, the firmware was rewritten to allow one processor to handle incoming frames, and the other outgoing frames. The process of rewriting the firmware was undertaken very much in parallel with the measurements, as (for instance) it would often be found that the sending and/or receiving of frames by the NICs was not sufficiently fast enough to prevent their loss. This specialised firmware was designed specifically to be used in conjunction with the LHCb Fragment Generator software (see below) Additional firmware changes had to be made in order to perform histogramming on board the NIC and to enable a specialised form of flow control. The specialised firmware provides a clock resolution of almost 1/20 μs making for greater accuracy in the measurements.

3.1.2 The Fast Iron Switch

The switch used to obtain the results presented here is a Level 2 Gigabit speed, Foundry 4000 4-Blade model [4], with a claimed ability to switch at line speed independent of packet length, together with 2MB of on board buffer RAM. Only a single blade with 4 active ports was available for testing. A similar 8000 series switch housed in the CMS DAQ test set-up fully configured with 8 blades was also briefly used in an attempt to determine how the performance of the switch varied at higher loads than were achievable in this set-up.

3.2 The Fragment Generator Software

Developed using Python; a C based scripting language, this GUI software allows the user to view and set various registers within the NICs. This allows the user to set various bits controlling packet destinations, numbers and sizes or the type of measurement being made (i.e. ping, pong, or ping-pong), in addition to viewing various statistics on roundtrip time differences and numbers of packets received, amongst countless others. For the purposes of the measurements presented here the most important registers were of course those detailing the timestamps at which the packets are produced/received, the number and size of packets and that controlling the rate of packet production. The software is also capable of producing basic histograms detailing the variation in the timestamps. Additionally it provides a facility to automatically script a series of measurements and output the details (timestamps, histograms etc).

3.3 The Benchmarking Set-ups

The first measurements given here detail the performance of a direct NIC-NIC connection, in order that one may identify any effect due to the NICs, from the measurements carried out on the switch. The second set of measurements carried out involve two NICs connected together via the switch on the same blade. Then there is an attempt to determine any added latency, which might occur when the two NICs are connected on different blades and whether there is a dependence on packet length. Finally there is an outline of a simple method to estimate the size of the switch output buffer by the use of flow control.

By arranging the measurements in this way, packetisation (i.e. the delays caused by allocating memory for the internal switch packets) within the switch should manifest itself as a step-like function in the results. Latencies (as a function of packet size) through the switch may also be calculated by subtracting the direct measurements. Also latencies across different blades may also be calculated by subtracting the both the direct and the single blade measurements.

Throughout each of the tests, comparisons were made to the statistics provided by the switch to ensure that the switch received incoming packets correctly and that those packets re-transmitted by the switch were correctly received by the other NIC. It is worth mentioning that among several 100 millions of packets created, transmitted and received in the course of these measurements not a single packet showed a bad checksum, indicating a bit error.

3.3.1 Direct, NIC – NIC

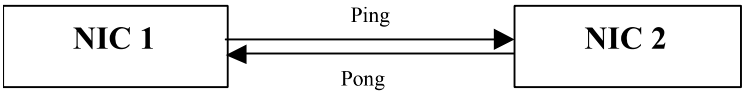


Figure 3 Set for the direct NIC-to-NIC measurements

The NICs were installed on a standard PC via PCI and connected back to back using a standard twisted pair copper Cat5 cable. Flow control was switched off. Measurements of the ping-pong time difference were made at 1-Byte intervals between the minimum (64 Bytes) and maximum (1500 Byte) frame size. The measurements were averaged over 1 million packets per frame size at a rate of 20 kHz (to avoid losing packets in the NICs). The scripting feature of the Fragment Generator software was invaluable for such a large number of measurements.

3.3.2 Through a Single Switch Blade

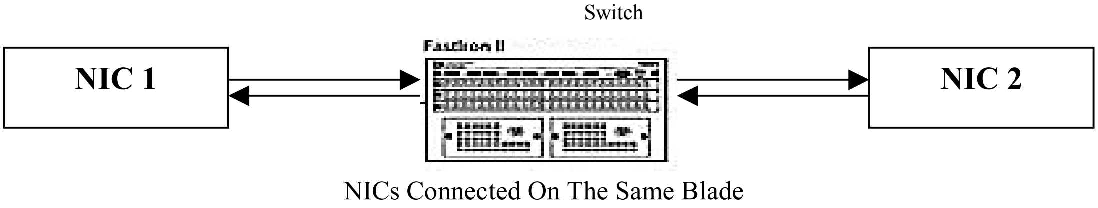
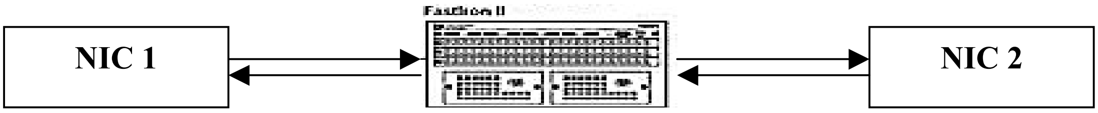


Figure 4 Setup for the measurements with the NICs connected to the same blade of the switch.

Measurements are identical as those in the direct case; except that the NICs are connected together using untwisted copper Cat5 via ports on the same switch blade. Flow control is again switched off on both the NICs and the switch.

3.3.3 Across Two Switch Blades



NICs Connected On Different Blades

Figure 5 Setup for the measurements with the NICs connected on different blades of the switch.

Since our Foundry 4000 has only one blade, it was not possible to carry out these measurement. Instead, the measurements were carried out using the system loaned to us by the CMS DAQ group. The set-up, in terms of cabling and software was the same as for the single blade case detailed above. The main quantity of interest in this measurement, i.e. the additional delay imposed by being forced to switch via the back-plane, can be obtained by a relative measurement as compared to the single blade case.

4 Results

4.1 The Benchmark Results

To reiterate, in the results presented below each measurement is averaged over 1 million packets, transmitted at a rate limited 20kHz to avoid overloading the receiving NIC. Measurements were taken at 1 Byte intervals between 64 and 1532 Byte frames. Additionally the NIC onboard clocks used to measure the timestamps had a resolution of 1/17 μs in the Direct and Single blade tests, while the Twin Blade and Buffer tests were carried out with only a 1μs resolution. The measurements through the switch have not had the direct NIC-NIC measurements subtracted.

A fit to the data was made for each of the three sets of latency measurements using the following chi-squared type formula:

$$y(x) \;=\; a + b \cdot x + \mathrm{int}\left(\frac{x\text{- }1}{c}\right) \cdot d$$

Where:

x is the Frame size in Bytes
a is the Constant overhead (due to cabling, turnaround times, minimum switching time etc)
b is the Latency per Byte (extra time for each additional Byte within a packet)
c is the packetisation quantum
d is the jump at the packetisation boundary

4.2 Direct, NIC-NIC Results

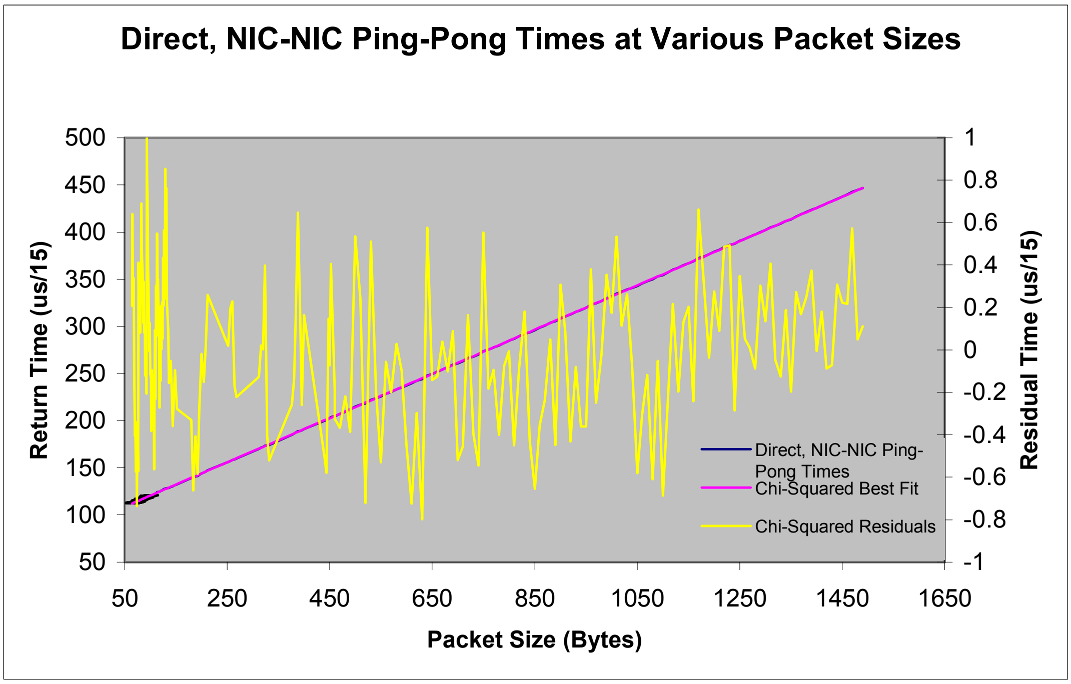


Figure 6 Results of the direct NIC-to-NIC measurements.

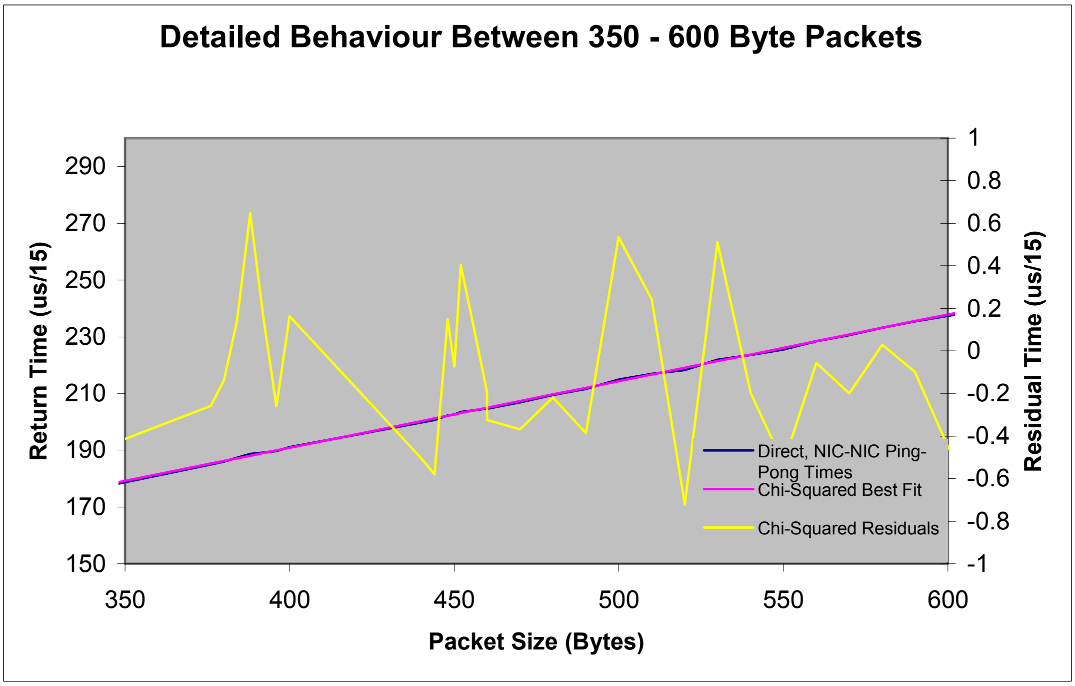


Figure 7 Close-up of the latency measurement together with the fitted function in the size-range of 350-600 Byte frames.

The values of the fitted parameters are summarized in Table 2

Constant Overhead	5.706 µs
Latency	0.0139 µs/Byte
Packetisation Quantum	0 Bytes
Quantum Step	0 µs/64 Bytes

Table 2 Results of the fit to the date of the latencies for direct NIC-to-NIC connections

Comments

The above plots indicate that the NICs are able to send and receive packets with an inherent latency, which is linearly dependent upon the packet size. The additional small-scale structures, which can be seen in the 'detailed' plot in Figure 7, can be attributed to errors in measurement due to counter wrap-a rounds and various delays caused by histogramming. Their effect is negligible and may be ignored.

The constant overhead in the NIC appears to be quite high, and rises relatively steeply compared to the latencies within the switch. Examining both plots one can see that as expected no packetisation appears to be present within the NICs. This makes the task of measuring any packetisation within the switch considerably easier.

4.3 Results Through A Single Switch Blade (Port to Port)

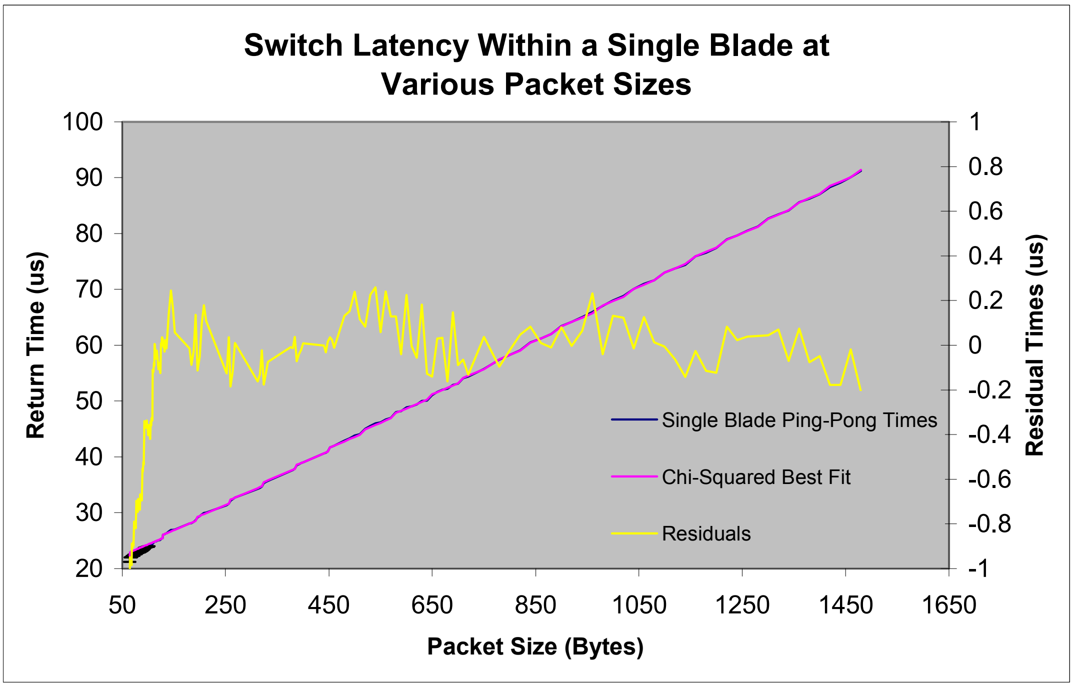


Figure 8 Latency measurements for traffic through a switch where the two NICs are connected to the same blade. Clearly visible is an additional regular structure as a function of packet size.

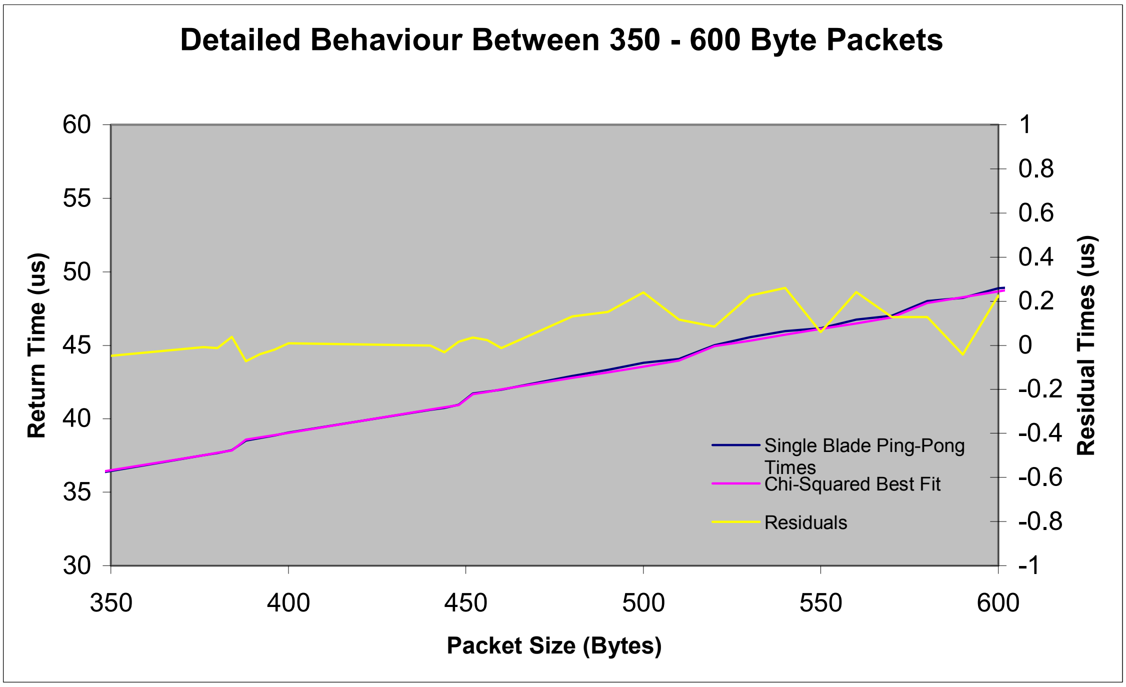


Figure 9 Close-up plot of the latency measurement across the switch (through the same blade) for frame sizes between 350 and 600 Bytes. The 64-Byte sub-structure is clearly visible.

The results of the fit to the data are summarized in Table 3

Constant Overhead	1.166 µs
Latency	0.038 µs/Byte
packetisation Quantum	64 Bytes
Quantum Step	0.035 µs/packet

Table 3 Results of the fit to the data for the latency measurements throu a single blade of the switch.

Comments

When the NICs are connected together over a single switch blade it can be seen that the latency the switch adds scales linearly with packet size, in a similar fashion to the latency within the NICs. However, the switch itself appears to add considerably more latency per Byte than the NIC's sending and receiving routines. One also notes the small residuals indicating the high goodness of fit.

One can clearly see a 64 Byte packetisation effect within the switch, (the delay in allocating and filling another block of memory for an internal switch packet) even though it is relatively small. This small packet step can cause measurement difficulties if one does not have a sub microsecond clock resolution with which to measure the NICs.

4.4 Result Across Two Switch Blades (Port to Port)

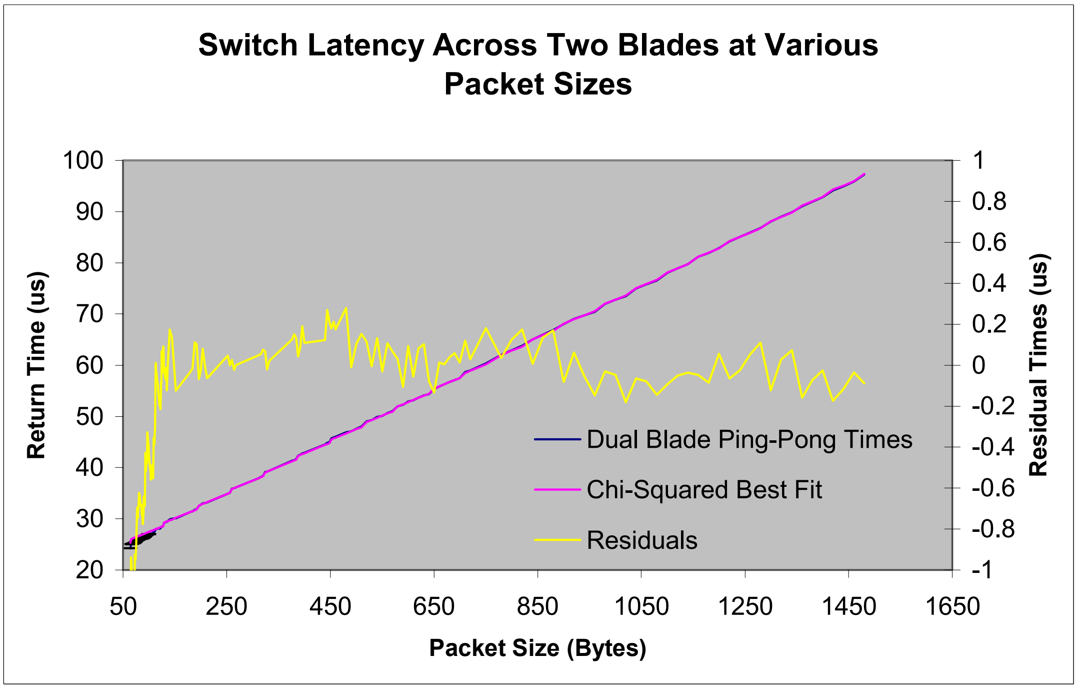


Figure 10 Round-trip latency measurements between ports connected to two different switch blades.

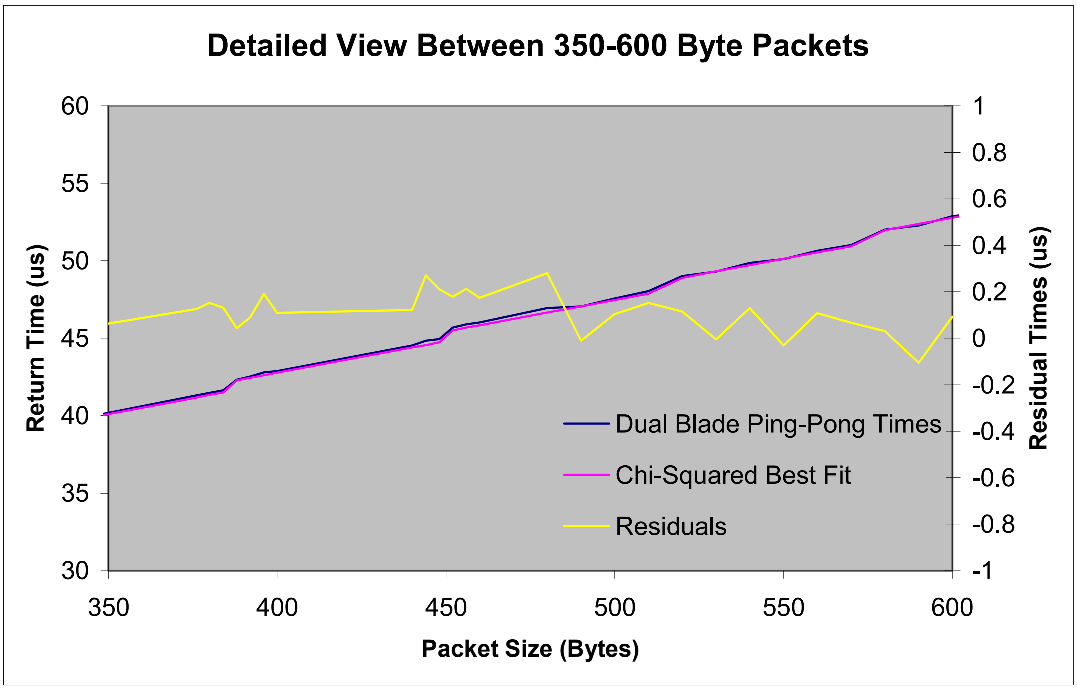


Figure 11 Close-up of the round-trip latency for frame sizes between 350 and 600

Bytes

The values of the parameters after fitting are summarized in Table 4

Constant Overhead	1.338 µs
Latency	0.041µs/Byte
Packetisation Quantum	64 Bytes
Quantum Step	0.0362 µs/packet

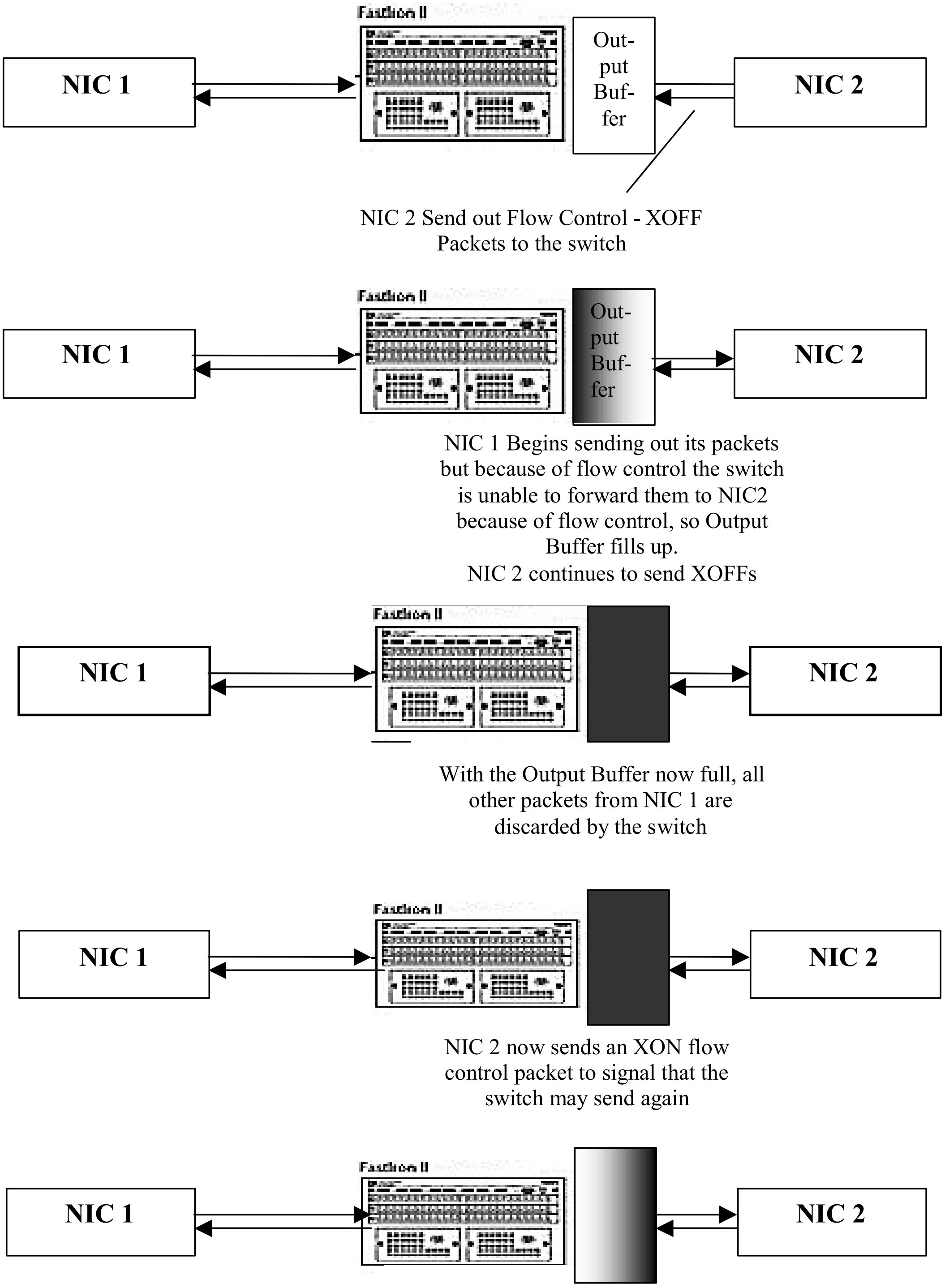
Table 4 Fitted parameter values for the measurements across different switch blades.

Comments

Switching across different blades over the backplane does not appear to drastically alter the performance of the Foundry switch. In fact the backplane merely adds an extra constant overhead, rather than one, which scales with packet size. By subtracting the overhead from the single blade measurement from the overhead here, one calculates this 'backplane' overhead to be 0.172 µs. All other features appear more or less identical to the case when one switches over a single blade. Note however that these results were obtained in the loaned CMS setup, where the clock resolution was only around 1 µs.

4.5 Output Buffer Size Measurements

The set-up here is identical to the case where the NICs are connected together on the same blade, with the exception that this is a simple 'ping' measurement and that flow control is now active on both the NICs and the switch. But the actual measurement performed here was not of ping-pong times, but of the number of packets dropped. The order of events proceeds thus:



NIC 2 Send out Flow Control - XOFF
Packets to the switch

NIC 1 Begins sending out its packets
but because of flow control the switch
is unable to forward them to NIC2
because of flow control, so Output
Buffer fills up.
NIC 2 continues to send XOFFs

With the Output Buffer now full, all
other packets from NIC 1 are
discarded by the switch

NIC 2 now sends an XON flow
control packet to signal that the
switch may send again

The Switch may now drain the Output
buffer by sending to NIC 2. Once
complete the number of packets sent
by NIC 1 and received by NIC 2 may
now be compared.

Figure 12 Sequence of actions for measuring the output buffer size

Approximately 1000 packets of both 64 and 1024 Bytes were transmitted separately for this buffer measurement. By comparing how many packets the switch receives and how many are sent out one may obtain an estimate for the size of the output buffer. This is done by taking the number of packets transmitted by the switch after flow control is switched off and multiplying it by the packet length. In addition by checking different packet sizes one may see if there is a limitation on the total number of frames that the buffer can hold in addition to any overall limitations imposed by the size of the buffer.

One may also check the receive buffer on NIC 2 using the Fragment Generator, to check precisely which packets are lost first by the buffer

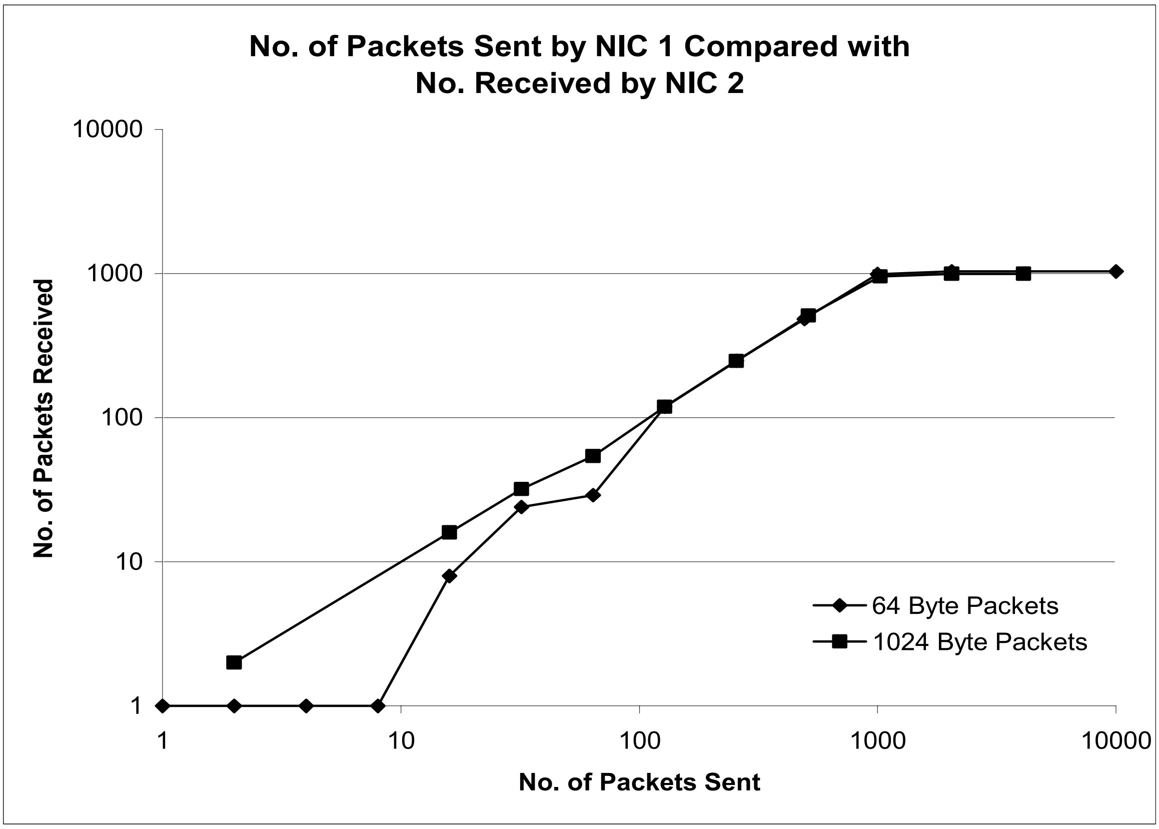


Figure 13 Number of frames received versus the Number of frames sent in the case where the output port was blocked previously.

Comments

Examining Figure 13 one sees the full effects of flow control on the buffer occur at approximately 1MB (1024 1024-Byte packets) and at 1024 64-Byte.packets. This indicates that the buffer is limited in capacity (either physically or due to parameters set within the switch) to 1 MB, and to holding a maximum of approximately 1000 frames within the buffer memory. The 1024 frame limit is per port, whilst the 1MB limit appears to be shared between the 4 ports on each blade. (This limit is 2 MB on the Foundry FastIron 8000, as it has 8 ports per blade, however the 1024 frames/port limit still applies).

There are also additional effects, which appear somewhat strange and could warrant further investigation. It was observed (see the 64 Byte line in Figure 13), that once flow control is turned off by NIC 2, the switch will send out only the first packet. It then appears to drop the next 10 packets either as they are received from NIC 1, or as they are sent out to NIC 2. The rest of the data transfer then appears to take place normally until the limits set out above. This is somewhat peculiar behaviour for which we do not have any explanation. This phenomenon does not appear to depend upon packet length.

Other observations include that the switch, whilst it submits to flow control request from the NICs, it will not actually signal flow control if the buffer for the destination port is full. Instead it simply discards any further incoming packets. Other traffic with different, non-blocked, ports will still pass normally however

5 Summary

To summarise the result of the parameters inferred from the measurements performed with the Foundry Fast Iron 4000/8000 series switch:

Latency as a function of packet size

The fitted formula given in Section 4.1 may be re-expressed in the following form:

$$y(x) = a + b \cdot x + \mathrm{int}\left(\frac{x-1}{c}\right) \cdot d$$

Where:

x is the Frame size in Bytes
a is the Constant overhead (due to cabling, turnaround times, minimum switching time etc)
b is the Latency per Byte (extra time for each additional Byte within a packet)
c is the packetisation quantum
d is the jump at the Packetisation boundary

The results are summarized in the following Table 5

Table 5 Summary of the fitted parameter values for the cases of single-blade and across-blade transfers

Single Blade (Port-to-Port)	a = 0.537 µs
	b = 0.038 µs/Byte
	c = 64 Bytes
	d = 0.035 µs/packet

Across Blades (Port-to-Port)	a = 1.338 µs
	b = 0.041 µs/Byte
	c = 64 Bytes
	d = 0.0362 µs/packet

Packetisation

The Switch has been found to be a packet switch, with an internal packet length of 64 Bytes

Buffer Sizes

With the queue depth current settings the following limits occur at the output of one port:

a. Up to 1024 packets

b. Up to 1 MB data in total

Both limits must be respected at the same time! It is possible that the available buffer can be made to be 2 MB, by setting suitable parameters (to be confirmed). Anyhow, each blade provides a shared output buffer for all 8 ports it serves, of 2 MB total

Head of Line Blocking

The switch has been found not to show any head of line blocking.

Buffer overflow

Upon buffer overflow, for any reason (too many packets, no buffer space etc) the switch will simply discard all incoming packets directed to any of the blocked ports. Traffic to non-blocked ports will still pass.

Flow control

The switch respects flow control, i.e. upon reception of a flow-control packet, it will stop sending to the port from which it has received it. The switch itself will never issue a flow-control packet, but will simply drop packets, when it is low on resources.

With this information we are now in a position to begin simulating a Readout Network employing the optimised topology together with simulated Fast-Iron switches. We now also have the machinery in place to do quick, and reliable measurements on other switches, to evaluate whether they fulfil the requirements of the LHCb DAQ and to obtain input parameter for our simulation.

6 Acknowledgements

We would like to thank the CMS DAQ Group for the kind loan of their equipment.

7 References

[1] B. Jost, *"The LHCb DAQ System"* – Presentation at the DAQ 2000 Workshop, October 2000 Lyon

[2] J. P. Dufey at al. in preparation

[3] N. Neufeld, "The LHCb Event Building Strategy" - Presentation at the IEEE NPSS Real Time 2001, June 4 - 8 Valencia, Spain

[4] See http://www.oti.net/Systems/Foundry/PDF/big_iron.pdf for more details