High Performance Gigabit Ethernet Switches for DAQ Systems

Artur Barczyk, Jean-Pierre Dufey
for the LHCb collaboration
CERN, 1211 Geneva 3, Switzerland
Email: Artur.Barczyk@cern.ch, Jean-Pierre.Dufey@cern.ch

***Abstract*— Commercially available high performance Gigabit Ethernet (GbE) switches are optimized mostly for Internet and standard LAN application traffic. DAQ systems on the other hand usually make use of very specific traffic patterns, with e.g. deterministic arrival times.**

Industry's accepted loss-less limit of 99.999% may be still unacceptably high for DAQ purposes, as e.g. in the case of the LHCb readout system. In addition, even switches passing this criteria under random traffic can show significantly higher loss rates if subject to our traffic pattern, mainly due to buffer memory limitations. We have evaluated the performance of several switches, ranging from "pizza-box" devices with 24 or 48 ports up to chassis based core switches in a test-bed capable to emulate realistic traffic patterns as expected in the readout system of our experiment.

The results obtained in our tests have been used to refine and parametrize our packet level simulation of the complete LHCb readout network.

In this paper we report on the results of our tests, and present the outcome of the simulation using realistic switch models.

I. Introduction

The DAQ system of the LHCb experiment [1], starting from the Level 1 trigger, is based on Gigabit Ethernet technology [2]. A schematic view is shown in Fig. 1.

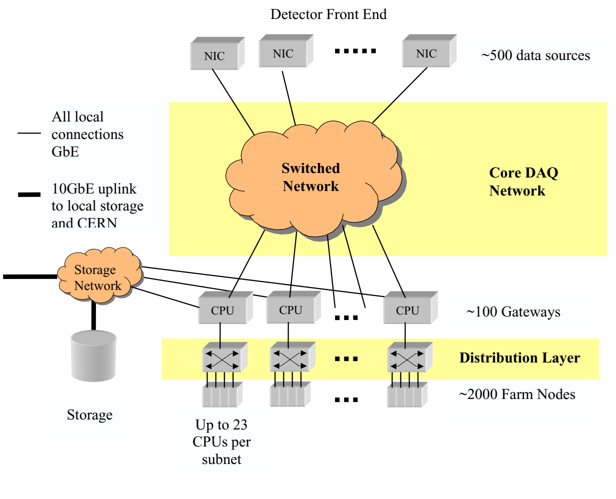

Fig. 1. Schematic view of the LHCb readout system.

Data is directly embedded in IP packets, omitting the use of a transport protocol. No packet retransmission is foreseen. We therefore put strict requirements on the reliability of the transmission and thus on the performance of the switching equipment along the data path.

There are two software trigger levels: Level 1 and High Level Trigger (HLT), both algorithms running on the same trigger farm of $\sim$2000 CPUs. We identify two flows in the system: a high-priority, low latency flow associated with the Level 1 trigger data, and a low-priority flow with no latency constraints relating to the HLT.

Detector Front-End modules inject data into the system upon reception of a trigger signal distributed by the LHCb Trigger and Fast Control (TFC) system. Since each module contains fragments of the same event, this leads to a very particular traffic pattern, where all source ports receive data frames for a single destination port (event building traffic), synchronised to within $\sim$100 ns[1] The nominal frame input rates are 40 kHz and 4 kHz for the Level 1 and HLT, respectively.

The response of the switch to this deterministic arrival time pattern depends on the performance of the switching fabric, the scheduler implementation, internal flow control capabilities and the buffer memory allocation strategy in the switch.

II. LHCb Readout Network Models

The LHCb readout network, as proposed in the Technical Design Report, has been designed as a multistage network, with an aggregation layer meant to reduce the number of ports to the core and enhance the link utilisation, and a core part built of switches forming a fully-connected network. With the advent of high port density, high performance switches, we added two scenarios:

- a medium sized chassis-based core switch, still keeping the aggregation layer, and
- a single high-density chassis based switch with the required number of ports in a single unit.

Common to all these scenarios is the fact that our traffic pattern requires high amount of buffer memory in the switches.

III. Test Bed

High performance switches from various manufacturers have been examined in the test bed built for this very purpose.

We use 16 network processors (IBM NP4GS3) to generate and analyse the traffic. Hosted on a PCI card, each of the

[1] The jitter on the arrival time is given only by the framing time in the FE module's NIC.

NPs provides 3 Gigabit Ethernet ports. These programmable devices allow us to generate the precise timing of each frame. In addition, we use the GPIO pin on the processor boards to synchronise the traffic between all the ports.

The programming model is based on 4 components: a cross-assembler to generate NP4GS3 specific executables, a device driver for the processor board, a client-server application for communication with the network processor, and a Python scripting library to define the traffic pattern used in a test. We have developed the last two components specifically for the LHCb switch test bed.

IV. Switch Performance Tests

We have identified two main issues crucial to the performance of the LHCb readout network: the amount of egress buffer memory and its allocation strategy, and the switching fabric and scheduler performance. Correspondingly, our two main tests concentrate on examining the amount of memory available to a single port, and the frame loss when subject to aggregation traffic. Additional evaluations such as switching latency measurement, full-mesh test and high-statistics port-to-port data transfers, were carried out to estimate the frame loss rate not related to over-subscription.

A. Memory Layout

Different switches provide varying amounts of buffer memory. While the "nominal" memory sizes can be easily obtained from the manufacturer, we have found out that by far more interesting is the allocation strategy. Buffer memory is usually organised according to Quality of Service (QoS) profiles, which, for untagged, unprioritised traffic might result in significant memory space being kept unused. In Fig. 2 we show one example of measured queue depth as a function of frame size. The full (black) line shows the number of queued frames, while the dashed (blue) line indicates the corresponding amount of memory used.

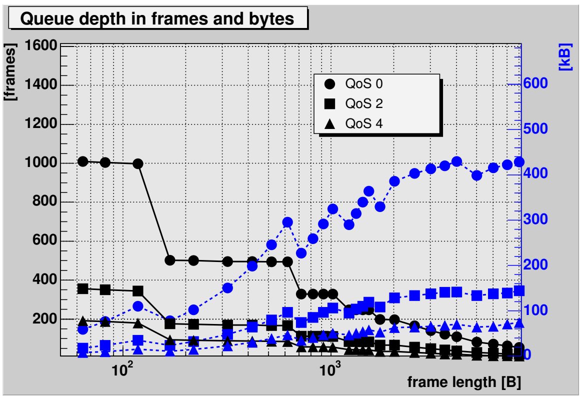

Fig. 2. Example of measured queue depth as function of frame size for 3 different Quality of Service settings.

Three different QoS priorities have been compared, with the priority assigned on port basis. We observe two important points: the first one is the effect of memory allocation strategy according to frame length which leads to sub-optimal memory utilisation for short frames. The second interesting point is the fact that queues are shorter for higher priority traffic. This is not surprising, since higher priority queues are serviced either more frequently (Weighted Round Robin), or as long as they contain data (Strict Priority Queuing), thus less high priority frames need to be queued wrt. low priority ones.

B. Switching Latency

The switching latency has been measured for two reasons: latency budget for the Level 1 trigger, and to obtain realistic parametrisation of the simulation of our system. We observe that under heavy load, typical latencies are under $20\mu s$ for single unit devices, and below $200\mu s$ for chassis based switches.

In Fig. 3 we show an example of the measured latency for frame sizes in the range defined by the Ethernet standard of 64-1518 Bytes. Two measurements were perfromed: one with only a single frame being forwarded, and one where the switch was subject to 80% load on all other ports. We observe an increase in latency with increased load, as well as less deterministic behaviour.

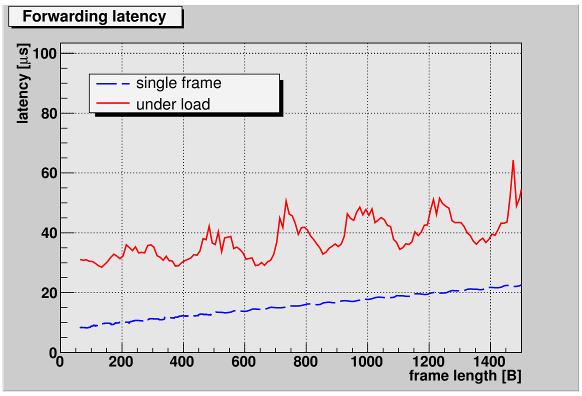

Fig. 3. Example of measured forwarding latency.

C. Traffic Aggregation

The switches in evaluation were subject to our aggregation layer traffic pattern. Several combinations of input to output port assignments were tested. No problems were observed even under 100% output link load on the high perfomance switches.

V. Switch Internals

An important input to our simulation studies has been the internal layout of the switch. Even the relatively small, 48 port switches are based on building blocks of typically 10-12 ports interconnected through the main switching fabric, an example being shown in Fig 4.

Each such block can perform local switching, so that it is possible, by switching some of the traffic locally, to set up a non-blocking system, even if the switch itself is intrinsically blocking. Each ASIC has its own buffer memory of up to 1 MB, shared between its ports. Such chipsets, consisting of N-port switch on a chip and a corresponding fabric chip are available from Broadcom [3], Marvell [4] and other

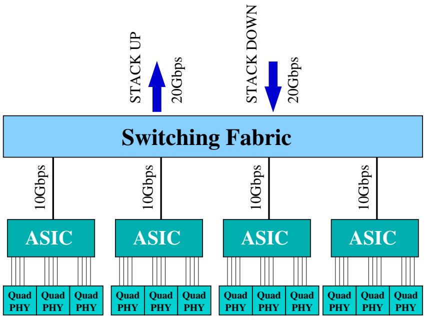

Fig. 4. An example of a 48 port switch architecture with 2 stacking connections of 20 Gbps each.

manufacturers, and are often found in 1U rack-mountable edge switches.

The internal structure of such a switch is reflected in the measured forwarding latency as shown in Fig. 5. We can clearly distinguish traffic switched within an ASIC (dotted line), through the fabric, but still within the same switch (full line), and traffic switched across two switches connected through a stacking link of 20 Gbps bandwidth (dashed line).

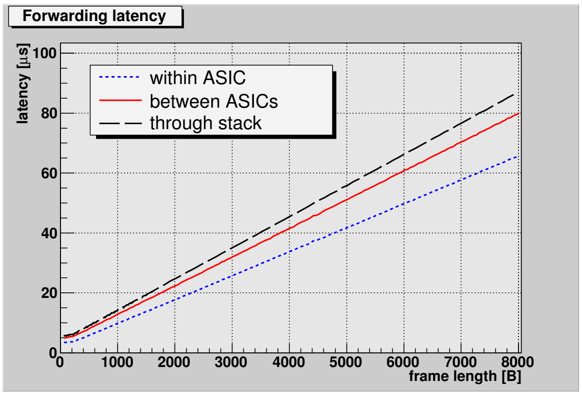

Fig. 5. Measured forwarding latency of a 48 port GbE switch.

VI. Simulation

The packet level simulation package developed to verify the design of the LHCb readout network has been extended to reflect the internal architecture of the switches. Using the output of the LHCb physics simulation as source of frame timing as sizes, we were able to obtain important parameters such as the total transfer time of the event, as well as buffer memory utilisation.

Fig. 6 shows the layout of the complete simulation model, including the aggregation layer and the core network.

Full-mesh topology has been chosen for the core network, as shown in Fig. 7. Each switch contains ports connected to the aggregation layer, as well as ports connected to the gateway PCs. Interconnections between the core network switches are made up of aggregated links with a throughput of 3 Gbps each.

Physics Monte-Carlo simulation samples have been used to obtain realistic packet sizes as well as frame arrival times.

The simulation confirmed that the transfer time is dominated by the queuing time. In a multi-stage network with a fully-connected core, the transfer time for a Level 1 trigger event is below 3.5 ms, and below 4 ms for a HLT event.

We refined the simulation model by including the internal structure of a switch as described in Sec. V. We also make use of the 20 Gbps stacking connections between the aggregation layer and the core network, as shown in Fig. 8.

Further enhancements include higher bandwidth on internal connections in the core network, made possible by the free ports otherwise used to interconnect it to the aggregation layer. Four ports per aggregated link are used, rising the bandwidth per internal connection to 4 Gbps. Aggregated links of 2 GbE ports each are also used in this scheme for the connection to the destination hosts. In addition, we use a priority scheme with two levels, prioritizing the latency-critical Level 1 trigger traffic.

This refined model with the above modifications reduces the transfer time of the Level 1 data to well below 1 ms, show in Fig. 9. The memory utilisation per unit of 12 ports remains below 400 kB.

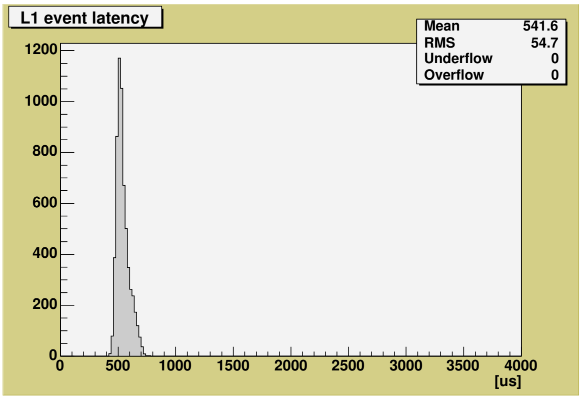

Fig. 9. Level 1 event latency.

VII. Conclusions

We have evaluated several high-performance GbE switches for suitability in the LHCb readout network. Important insights have been gained about the memory allocation, fabric performance and the internal layout of the devices. The data has been used to create a realistic packet level simulation model of the complete system. The outcome of the simulation provides satisfactory results, demonstrating that the readout network can be built using commercially available switches, and critical issues such as Level 1 Trigger traffic latency and memory utilisation in the switches are well under control.

References

[1] LHCb Collaboration, "LHCb Online System, Data Acquisition and Experiment Control, Technical Design Report", CERN/LHCC 2001-040, LHCb TDR 7, 19 December 2001.

[2] IEEE, "802.3 IEEE Standard for Information Technology, Part 3: Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications", Piscataway, NJ, 2002.

[3] Broadcom Corporation web site [Online]. Available: http://www.broadcom.com/

[4] Marvell web site [Online]. Available: http://www.marvell.com/

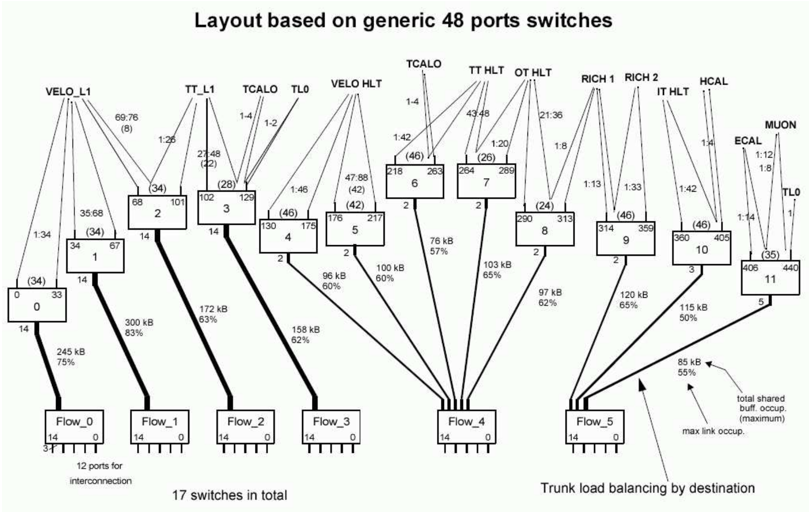

Fig. 6. System layout using generic 48 port GbE switches.

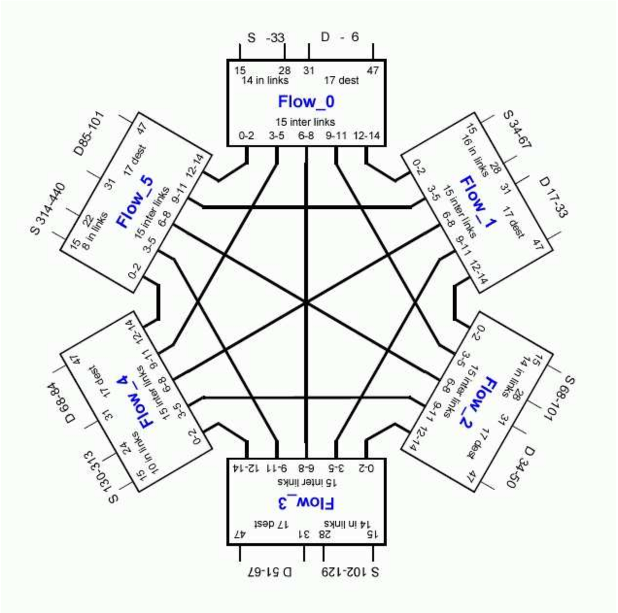

Fig. 7. Core network topology using 48 port GbE switches.

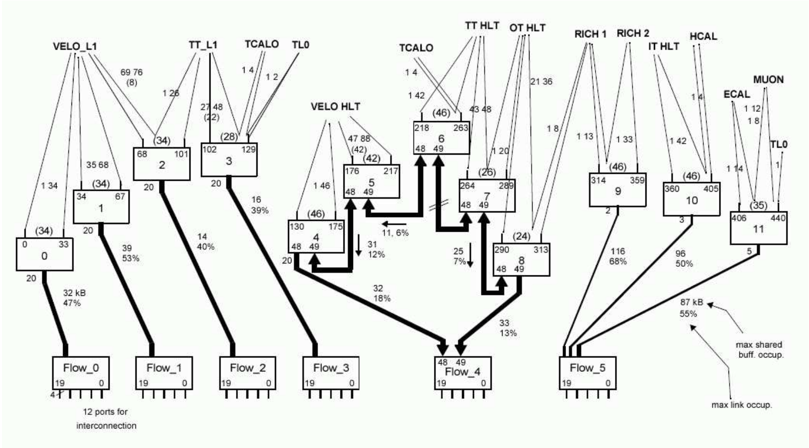

Fig. 8. System layout using stackable switches.