Evaluation of Switching Equipment for the LHCb Readout Network

LHCB Technical Note

Issue:	1
Revision:	0
Reference:	LPHE 2005-009 / LHCB 2004-041 DAQ
Created:	20 March 2004
Last modified:	05 May 2004

Prepared By: A. Barczyk, B. Jost, CERN, N. Neufeld, LPHE EPFL

Abstract

The choice of switch equipment to be purchased for use in the LHCb readout network shall be based on careful evaluation of performance and features of candidate hardware. This will be carried out in the framework of the existing LHCb DAQ test bed. The different criteria as well as planned tests to verify if a given switch meets them are presented.

Document Status Sheet

1. Document Title: Evaluation of Switching Equipment for the LHCb Readout Network			
2. Document Reference Number: LHCb 2004-041 DAQ			
3. Issue	**4. Revision**	**5. Date**	**6. Reason for change**
Draft	0	31 March 2004	First draft version
1	0	05 May 2004	First release version

Table of Contents

List of Figures

1. Introduction

This document describes the requirements imposed upon the switching equipment to be used in the LHCb DAQ system, the details of which are given in [1] and [2]. As all the details are given in these references, we refrain from repeating them here. Figure 1 summarises the design layout of the system. Custom built front-end electronics modules generate data frames, with one or two Gigabit Ethernet connections per front-end module. Data rates at this stage are well below the bandwidth offered by Gigabit Ethernet, thus the high number (~500) of connections is reduced to ~100 using edge switches concentrating the incoming traffic. The core readout network routes then the data frames to one of the processing farms, about 100 in number.

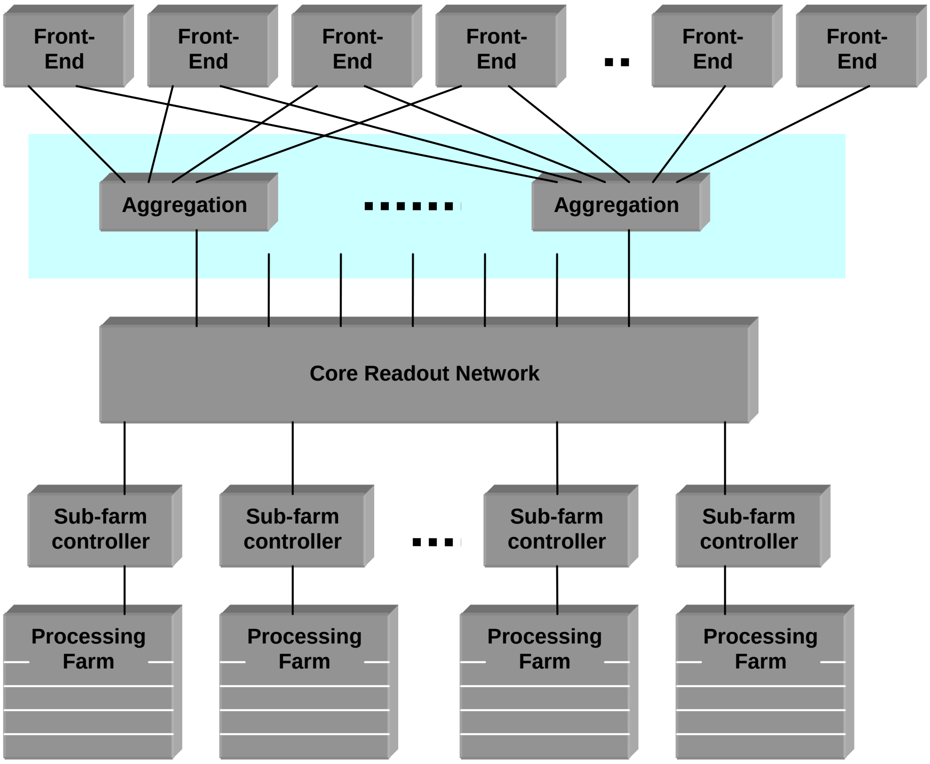

Figure 1: The layout of the LHCb readout network.

Probably the most important feature of the LHCb data acquisition system is the synchronisation of the frames entering the system. The frames are expected to be dispatched on all front-end modules within a time span of a few hundreds of nanoseconds, every ~25 microseconds (called *cycle* in the following). At each cycle, the destination of all frames is one of the sub-farm controller PCs (SFC), which act as a gateway to the attached processing farm. This will clearly result in an over-commitment of the corresponding output port of the core readout network.

2. Requirements for the switching equipment

In order to define the requirements, it is important to distinguish between the two network layers in the data acquisition chain. The first layer, called *aggregation layer*, serves to reduce the number of Gigabit Ethernet links and enhance the link utilisation, while the second layer is intended to route the data to a single destination common to all packets within a transmission cycle. The different purposes lead necessarily to different requirements, and shall be described below. The common features for both layers are:

- Low switching latency: due to the nature of the task at hand, the time between the frame dispatch at the front-end to the reception in the sub-farm controller has to be kept at minimum.

- Static routes: The source and destination addresses will be set by the Experiment Control System (ECS). Therefore, the switch should allow for at least 200 static routes.

- Unidirectional traffic: The data flow is directed strictly from front-end modules to the processing farm. Only datagram traffic is in use, no acknowledgement frames are transmitted in the “reverse” direction. This eases to some extent the requirements on the capacity of the switching fabric.

- Zero loss: Because no reliable protocol is in use, it is important that no packets are lost, even if the output ports will be overcommitted for short periods of time. The amount of buffer memory needed is currently under study. The switch shall be non-blocking, in particular show no Head-of-Line blocking behaviour.

2.1. Aggregation switches

In the aggregation layer, the switches perform a multiplexing function in order to enhance link utilisation, while at the same time reduce the number of input ports to the main readout network. This is considered similar to the functionality of an edge switch with an uplink. However in our case, all input ports are Gigabit Ethernet due to latency constraints, but the data delivered by the front-end electronics boards amount to between 3 and ~40% of single link load per port. The aggregation factors will therefore vary between 2:1 and up to 30:1 (the maximum can be lower, depending on the number of ports offered by the switch in use).

Data frames delivered by the front-end modules are synchronised to within a few hundreds of nanoseconds, i.e. at each cycle all input ports will receive data destined for the same output port. The switching fabric has to cope well with this situation. The input/output port assignment is strictly static, even though the destination address will change with every cycle. The cycle time will be determined at the setup of the data acquisition run, and can be expected to be around ~25 µs.

2.2. Readout network (core switch)

The purpose of the core readout network is to route the incoming traffic to one of the sub-farms for data processing. Three key differences distinguish this layer from the aggregation layer:

- Data will arrive at the input ports in “packet trains”, as a direct result of the aggregation discussed above. Since all these frames will go to the same destination, either the switching fabric has to offer enough bandwidth, or some amount of input buffering has to be available.

- All data arriving in one cycle has to be routed to the same output port, which will change with every cycle.

- Two streams will be mixed within the core network: a high priority, low latency flow, and a low priority flow with no latency constraints. Note: the zero loss constraint holds for both flows!

We consider two possible architectures for the main readout network: a single core switch with the necessary number of ports and switching capacity, or a multilayer network built of several switches with moderate number of ports, arranged e.g. in a Banyan topology.

Clearly, the output ports will be overcommitted for the duration of several milliseconds, which puts some requirements in terms of output buffer size. The amount of memory needed is currently under study, and depends on the architecture of the switch (output / virtual output /central queuing).

Prioritisation of the flows shall not lead to packet drop in the low priority queue – the low latency data is expected to “overtake” the low priority data, but not to cause packet loss.

2.3. Interoperability

Interoperability of the equipment is an important issue. Many manufacturers provide a wide product palette ranging from core switches for the backbone, down to edge switches for office connectivity. The structure of our setup is not different to this approach, thinking of the core network as the backbone and the aggregation layer as edge. Clearly a single vendor solution provides best means to guarantee best interoperability of the equipment. Although desirable from this point of view, it is not clear whether it can be maintained throughout the long life-time of the experiment (10 years and more). High value will be put therefore on the compliance of the equipment to well defined standards.

3. Tests to be performed

The aim of the tests to be carried out is threefold:

1. to verify whether a candidate switch is suitable for the purposes of our data acquisition system,

2. to gather parameters for use in the simulation of the complete LHCb DAQ system. This is an important point insofar as the tests can be performed using a small number of ports as compared to the size of the full system. In particular, if the main readout network shall be implemented as a multi stage switching network, the simulation, if fed with the correct parameters, shall allow us to extrapolate to the full size of the system, and so to choose the right topology.

3. to allow for a qualitative comparison of switches from different manufacturers.

3.1. Forwarding latency

There are two main components to the forwarding latency of a switch: the routing table look-up and the data transfer inside the switch.

Thanks to the static nature of the routing tables in our setup, and its limited size (only ~100-200 destinations), the first component can be expected to be small. It depends however on the overall load on the switch ports, since the routing tables are usually a shared resource. In our setup we therefore test two conditions: forwarding time between two ports, with all other ports idle, and the forwarding time with all ports under 100% load.

The frame transfer within the switch depends on the size of the frame, and whether the packet is handled locally in the blade, or travels through the backplane. The tests therefore include a scan over the full range of allowed Ethernet frame lengths, i.e. 64 to 1518 Bytes[1]. In the case of stackable switches, we also measure the forwarding time between ports in different modules, or between two line cards for chassis based models.

3.2. Packet loss rate

3.2.1. Full-mesh test

Independently of the intended use of a candidate switch, we perform a full-mesh test to obtain a measure of the packet loss rate. In this test, all ports are connected to traffic generators, synchronisation is not an issue. The link load for all ports is set at 100%; the frame size is varied between 64 and 1518 Bytes (or up to 9018 Bytes if Jumbo frames are supported by the switch). High statistics data is collected to obtain the packet loss rate.

[1] If jumbo frames are supported, the maximum is extended to 9018 Bytes.

3.2.2. Synchronised traffic, multiplexing

In this test we want to evaluate the behaviour of the switch under what we expect to be realistic conditions during data taking.

The ports of a switch are split according to functionality, let's say N input ports and M output ports. The traffic generators are programmed to send data synchronised to an external clock, with the clock rate between 30 and 60 kHz. Two test patterns are used:

- Aggregation: a single frame is generated per source at each cycle. The destination assignment is static in this case. The frame size is set to a value corresponding to 100% output link load, and is given by the multiplexing factor (N:M) and the clock rate f.
- Core network: at each cycle, a "packet train" is generated at the input ports, in order to emulate the outputs of the aggregation switches. The destination for all N inputs is assigned in Round-Robin mode among the M output ports with each cycle.

We test several N:M combinations, e.g. for a 24 port "pizza-box" switch, the combinations are 23:1, 22:2, 18:6 (for the aggregation layer test), and 12:12 (for the core network test).

We are well aware that 100% link load will not correspond to normal operating condition; this requirement is meant as stress-test of the equipment. We are interested in a) the loss rate at 100% link load, and b) the maximum link load at which the transmission can be considered loss-less ($<10^{-12}$ packets dropped).

3.3. HOL blocking

While the main readout network switch(es) is/are expected to be fully non-blocking, this requirement is not strict for the aggregation layer switches. A certain degree of over commitment of the switching fabric can be tolerated due to the fact that the input link utilisation will be well below 1 Gb/s. However, Head-of-Line blocking cannot be tolerated.

We do test this criterion by blocking an output port by means of continuous pause frame transmission. The traffic generated on an input port consists of a number of frames for the blocked port, enough to fill this port's output buffer, followed by frames for another, free, port. The analyser attached to the free port is expected to receive all frames transmitted to this destination.

3.4. Buffer size

For switches from manufacturers where we do not obtain information on the structure of the buffers, we will determine it using our test setup.

In its simplest form, one port is connected to a traffic generator, and one to an analysing device. The generator transmits frames carrying an identifier (frame number), while the analyser continuously issues pause frames such as to block the transmission on this port. After a sufficiently large number of frames has been transmitted by the source, the pause frame generation is stopped and the analyser looks for the first missing frame. The frame number, together with the frame size, gives a measure of the queue length and the memory size. Two measurements are necessary: one with the minimum and one with maximum supported frame size. The first measurement determines the queue length, while the second measures the buffer memory size. The tests are extended to use several ports, and aim at determining whether the buffer is shared.

A second way of measuring the buffer size avoids the use of pause frames, but is less accurate: Several input ports send frames to a single destination, thus over committing the output port. Again, the first missing frame gives a measure of the buffer size and queue length.

3.5. Link aggregation and resiliency

While not strictly required by the design documentation of the DAQ system, link aggregation would simplify the setup of the system. We will include verification of the link aggregation features in the tests described in Sec. 3.2. Note that due to the static nature of the setup, (automatic) LACP features are not desirable.

Resilient connections are not envisaged in our setup at this point, but will be investigated in a test.

3.6. Quality of Service

As mentioned in Sec. 2.2, the main readout network will route two flows with distinct priorities. As Layer 3 protocol will be used throughout the system, it is envisaged to use DiffServ (DSCP) for traffic prioritisation, if any. At the present moment, the use of 802.1p based prioritisation is not excluded, though.

The tests described in Sec. 3.2.2 include generation of two streams with two different DSCP values. The bandwidth ratio between the two flows is defined by the data rates of the experiment, and is used in the tests.

3.7. Management functions

All switches used in the DAQ system will have to interface to the Experiment Control System of the LHCb experiment. While this interface is not yet fully defined, it is clear that due to the scale of the setup, the switches have to provide functionality for automated management and monitoring. An important point in the design of the LHCb readout system is the separation of data and control paths. Therefore, a separate management port is highly desirable (out-of-band management).

4. Time requirements

The estimated time requirements to carry the out the tests are about 4 weeks for “pizza box” devices, and at least 2 additional weeks for stackable (if two devices are available) and chassis based switches with two line cards.

This time estimate is based on the fact that some of the tests involve significant amount of data, the different configurations and of course some learning time needed to familiarise ourselves with the particular hardware and management software (if applicable). It is not thought as a strict requirement; if the hardware can be made available for shorter test periods only, the most crucial tests (drop rate, response to our specific traffic pattern) will be performed. The requirement of additional time requested for multi-blade switches seems obvious, since additional tests will be carried out involving transfers across the blades.

5. References

[1] "A common implementation of Level 1 trigger and HLT Data Acquisition", A. Barczyk, J-P. Dufey, B. Jost, N. Neufeld, LHCb 2003-079 DAQ

[2] "LHCb Online Networking Requirements", B. Jost, N. Neufeld, LHCb 2003-166 DAQ