ATLAS TDAQ DataFlow Network Architecture Analysis and Upgrade Proposal

Stefan Stancu*, Matei Ciobotaru†‡, Krzysztof Korcyl§ and Brian Martin†
*University of California, Irvine, Irvine, CA 92697-4575, Email: Stefan.Stancu@cern.ch
†CERN, 1211 Geneva 23, Switzerland, Email: Matei.Ciobotaru@cern.ch
‡"Politehnica" University of Bucharest, Bucharest, Romania
§IFJ-PAN, Krakow, Poland, Email: Krzysztof.Korcyl@ifj.edu.pl

ATL-DAQ-CONF-2005-034
25 October 2005

***Abstract*— The real-time operation of the ATLAS DataFlow system is highly dependent on the performance of the Gigabit Ethernet network interconnecting its components (≈800 end nodes). After examining the functional and performance requirements of the network, several design alternatives (with respect to traffic repartition on core devices and available concentration technologies) are presented and analyzed. We introduce the use of 10 Gigabit Ethernet as a flexible and simple technology for concentrating traffic. Network testing equipment as well as discrete model simulations are used to assess the performance of various implementation options. Based on performance, fault tolerance and flexibility considerations a preferred architecture is proposed for implementation.**

I. Introduction

The ATLAS TDAQ (Trigger and Data Acquisition) system relies on a three layer trigger to reduce the initial 40 MHz event rate to 200 Hz, before transferring the event data to mass storage. The typical event size is 1.5 Mbyte.

The DataFlow system encompasses the level two trigger[1] (LVL2) and the event builder (EB). The LVL2 uses an RoI (region of interest) based mechanism to analyze the data validated by the first level trigger, while the EB gathers all the scattered fragments of the LVL2 validated events. The DataFlow system is implemented using a large number of PCs interconnected by a high bandwidth Ethernet network. The performance of the network is crucial, because approximately 90 Gbit/s (40 Gbit/s for LVL2 and 50 Gbit/s for EB) must be transfered reliably for a proper system operation. The suitability of Ethernet, as well as the specific features required by the DataFlow system have been presented in [1].

Both technology and our understanding of the TDAQ system have evolved since the initial network design [2]. 10 Gigabit Ethernet (10GE) is particularly well suited for traffic aggregation and is now competitively priced. Different implementation options are possible. We use a combination of a calibrated discrete event simulation model of the TDAQ system [3] (or only sub-parts of the model), together with measurements from the GETB network tester [4] to compare between them.

The work of Krzysztof Korcyl was supported by KBN Grant No. 620/E-77/SPUB-M/CERN/P-03/DZ 110/2003-2005.

[1]Excluding physics algorithms.

II. DataFlow network requirements

Fig. 1 illustrates the block diagram of the DataFlow system [5]. LVL1 validated events are received at up to 100 kHz and buffered in 1600 readout buffers distributed over approximately 150 readout systems (ROSs). After performing the level two rejection, full events are sent further to the Event Filter at approximately 3.5 kHz.

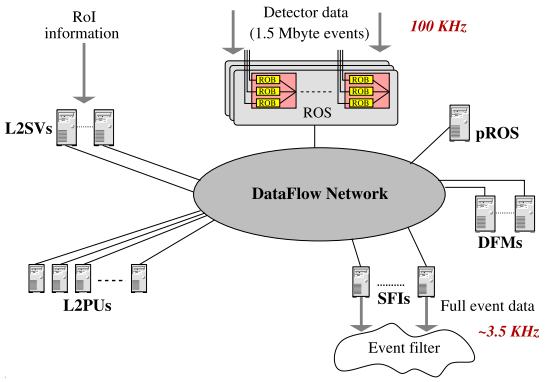


Fig. 1. The baseline DataFlow implementation

The L2SV (Level 2 Supervisor) performs the load balancing of the event processing task among the L2PUs (Level 2 Processing Units). The L2SV forwards the RoI information received from LVL1 to an L2PU having enough free resources. The processing unit successively requests RoI information from the ROSs and analyzes it until a decision as to whether the event should be accepted or rejected is reached. The LVL2 decision is sent back to the L2SV. A detailed analysis record of the validated events is passed further to the pROS (pseudo ROS). The L2SV forwards the level 2 result to the DFM (Data Flow Manager). If the LVL2 decision has been favorable, the DFM assigns an SFI (Sub Farm Input) to request and gather up the entire event data from all the ROSs (including the pROS). The rejected events identifiers, as well as those of the fully built ones, are grouped and sent via a multicast message to all the ROSs in order to clear their memory. The SFI buffers the completed events and passes them further to the Event Filter via a second network interface. As most of the traffic is request-response based the system can recover if occasional packet loss occurs. However the loss recovery

mechanism relies on timeouts, which have a strong penalty on the applications' performance.

The total number of each component type, as well as the bandwidth of all messages passed through the DataFlow network are summarized in Table I. Most of the traffic in the network is represented by the flow of event data from the ROSs to the L2PUs and SFIs (the two highlighted rows in Table I).

TABLE I
DATAFLOW SYSTEM BANDWIDTH REQUIREMENTS

Message	Senders		Receivers		Total BW [Gbit/s]
	no.	BW/node [Mbit/s]	no.	BW/node [Mbit/s]	
L2SV to L2PU	10	160.00	500	3.20	1.600
L2PU to ROS	500	4.00	140	14.29	2.000
ROS to L2PU	**140**	**285.71**	**500**	**80.00**	**40.000**
L2PU to L2SV	500	0.40	10	20.00	0.200
L2PU to pROS	500	0.06	1	28.00	0.028
L2SV to DFM	10	0.32	1	3.20	0.003
DFM to SFI	1[a]	2.80	90	0.03	0.003
SFI to ROS	90	4.36	140	2.80	0.392
ROS to SFI	**140**	**336.00**	**90**	**653.33**	**47.040**
SFI to DFM	90	0.03	1	2.80	0.003
DFM to ROSs	1	4.00	140	4.00	0.004

[a] A single DFM can cope with final system rates, but approximately 35 DFMs will be required for parallel calibration of different sub-detectors.

By design the load on the links shall be inferior to 60% of their capacity [2], in order to minimize the probability of packet loss due to buffer overflow during temporary traffic peaks. The only applications which approach this bandwidth are the ROSs and the SFIs. This is why we can afford to use a multilayer network topology:

- a *concentration layer* aggregates the traffic to/from L2PUs.
- a *central layer* offers non-blocking bandwidth to the ROSs, SFIs and the up-links from the concentration layer.

We shall first present the options for implementing the central layer (also denoted as *network core*), followed by a description and evaluation of several traffic concentration techniques.

III. NETWORK CORE

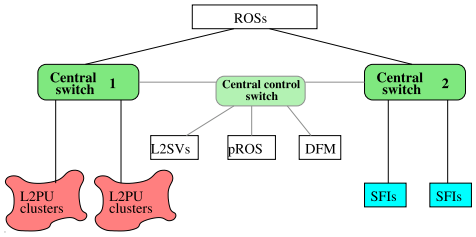
Due to the particularities of the traffic flow, the core of the network can be distributed over several devices. The only constraint is maintaining the connectivity between each L2PU/SFI and all the ROSs. This is achieved by using a separate ROS network interface for each central switch. Taking into account the ROS output bandwidth requirements (see Table I) as well as the size of the large switches currently available on the market, the use of two central switches seems the most suitable solution.

As the system will be deployed gradually, a single central switch will suffice in the initial stage. Later, a second central switch will be needed to accommodate the full size system. This approach considerably reduces the dimension of the central switches and widens the range of candidate devices to products from most major manufacturers (devices with 250-300 Gbit/s total switching capacity). Moreover, the use of two central switches can improve the fault tolerance of the final system.

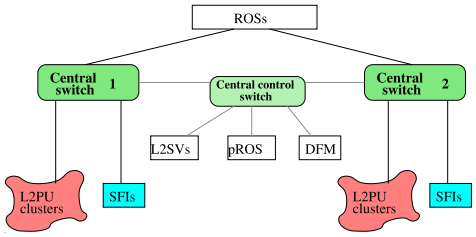
As illustrated in Fig. 2, the use of two central switches gives the liberty of mixing or separating the level two and event building traffic. Control applications (L2SV, DFM, PROS) with low bandwidth requirements may be concentrated before connecting to the central switches using a central control switch. The use of multiple central switches relies on the assumption that each ROS sends the event data through the interface connected to the same central switch as the L2PU/SFI receiving it; this can be enforced using VLANs. All these aspects will be detailed in the following subsections.

A. Traffic distribution on the central switches

Two options are available for the traffic distribution on the network core: *separated* LVL2/EB traffic (Fig. 2(a)) and *mixed* LVL2/EB traffic (Fig. 2(b)).



(a) separate LVL2 and EB traffic



(b) mixed LVL2 and EB traffic

Fig. 2. Central core options.

In the initial proposal of two central switches [2], the separation of LVL2 and EB traffic was considered beneficial, as it eliminates all network level interference between the two subsystems (level 2 trigger and event builder). Subsequent measurement and modeling activities have shown virtually no difference between the two options of distributing traffic on the central switches [6]. The network level interference of the two traffic types proved to be completely negligible in comparison to the inherent interference at the ROS level.

The mixed traffic solution presents advantages from two points of view: central switches sizing and fault tolerance. The size of the central switches is perfectly balanced when

traffic is mixed. There is no need to have a larger switch for the subsystem requiring more bandwidth (i.e. the event builder). If LVL2 and EB traffic are separated, the failure of a central switch would bring the TDAQ system to a stop. On the other hand, if traffic is mixed and one of the central switches fails, the system can continue to operate at reduced rate. Depending on the fault tolerance features of individual TDAQ applications, software reconfiguration may be required (masking out the applications connected to the faulty device), but no physical intervention is needed.

In conclusion, the mixed LVL2 and EB traffic solution (Fig. 2(b)) is preferred, because it improves the fault tolerance of the system and keeps the two central switches size balanced.

B. Central control switch

The control applications (L2SVs, DFM and pROS) have low bandwidth requirements and need to connect to both central switches. In order to reduce the port count of the central switches a *central control switch* aggregates all control traffic and distributes it to both central switches (see Fig. 2).

As the control applications are vital for the operation of the TDAQ system, the central control switch must be highly redundant. One may want to take advantage of the fault tolerance of the central switches[2] and connect the control applications directly to them. The downside of this approach is the increased size of the central switches, which narrows down the range of available equipment.

Considering the low bandwidth requirements of the control applications (see Table I), the system performance should be insensitive to adding or removing the central control switch. Modeling results of the full TDAQ system confirmed this hypothesis.

To summarize, the presence of the central control switch is justified by the need to efficiently use the bandwidth of the two central switches, and keep their size to a minimum.

C. VLANs for traffic separation

A requirement for the multi-device core is to force event data traffic to flow directly from the ROSs to the L2PUs/SFIs through a single central switch. If no precaution is taken, some applications may retrieve data from the ROSs via an under-dimensioned path (see Fig. 2): SFI – central switch 2 – central control switch – central switch 1 – ROS . This misuse of the system can be prevented by defining VLANs on the central control switch: one separate VLAN per central switch. All communication between the central switches is cut off, but control applications need access to both VLANs. The Linux operating system running on all the PCs hosting control applications makes this possible by emulating a separate interface per VLAN [7].

IV. Concentration techniques

As described in Section II, a concentration layer is introduced for the L2PU applications. The input bandwidth per processing unit is slightly below 100 Mbit/s (see Table I), so[3] we propose to use a grouping factor of 6 L2PUs per 1 Gbit/s in order to keep the link utilization inferior to 60%. Taking into account the L2PU packaging (30 units per rack), it is most convenient to locate the concentrator switch inside the rack, connect all units to it, and route a number of up-links to the network core.

Multiple links between two Ethernet devices cannot be used "out of the box", as they introduce loops[3]. Three methods can be used to achieve the desired concentration: *trunking*, *VLAN based* concentration and *10 Gigabit Ethernet* (10GE). The first two methods use multiple links between the concentrator and the central switch, plus additional configuration in order to break the loops. The third one uses single 10GE links for connecting to the network core. For the sake of clarity we shall assume that a single 48 GE ports switch (eventually with two 10GE up-links) is available per rack.

The performance of each concentration option has been evaluated in a test setup comprising a concentrator switch and a central switch. The GETB network tester [4] is used to send partially meshed traffic through the lines connecting the two switches. All ports connected to the concentrator switch send 64 byte frames to all ports connected to the central one, while 1518 byte frames are used for the opposite direction flow. For the 10GE concentration option, discrete modeling simulation results are also presented.

A. Trunking based concentration

The trunking standard (IEEE 802.3ad) allows the aggregation of multiple link segments in order to obtain a higher bandwidth interconnection between devices. In order to maintain the 6 L2PUs per 1 Gbit/s concentration ratio, we use five GE links between the concentrator and the central switch and aggregate them in a trunk (see Fig. 3(a)).

Unfortunately there is no standard mechanism for load balancing frames on the individual links within a trunk. The only requirement of the 802.3ad standard is to preserve the frames order. Depending on the load balancing algorithm implementation, some links may be over-utilized, while others have spare capacity or are even idle.

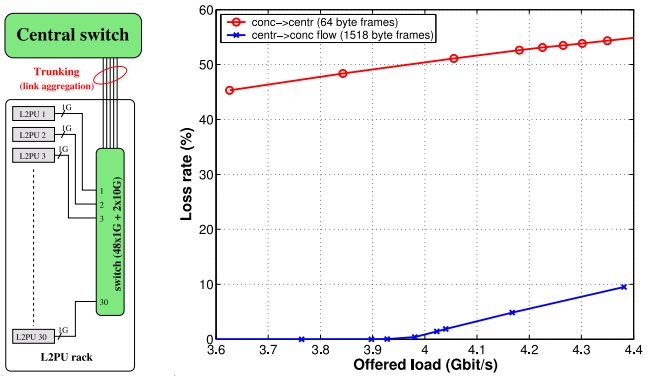
In our test setup we used two devices from the same manufacturer, aggregated four GE links in a trunk (maximum trunk size configurable on the concentrator switch) and expected to achieve a throughput close to 4 Gbit/s. As illustrated in Fig. 3(b), the 64 byte frames flow experiences 50% loss when the intended load on the trunk is 4 Gbit/s. Only two out of the four aggregated links forward traffic, while the other two are idle. On the reverse path, the 1518 byte flow nearly reaches 4 Gbit/s throughput with no loss.

B. VLAN based concentration

VLANs can be used to split a physical switch into several smaller logical switches. We define five disjunct VLANs which

[2]Central switches are chassis based devices with redundant power supplies, fans, management modules and often a redundant switching fabric.

[3]Loops are forbidden by the Ethernet standard as they allow broadcast traffic to infinitely circulate and consume all available bandwidth.

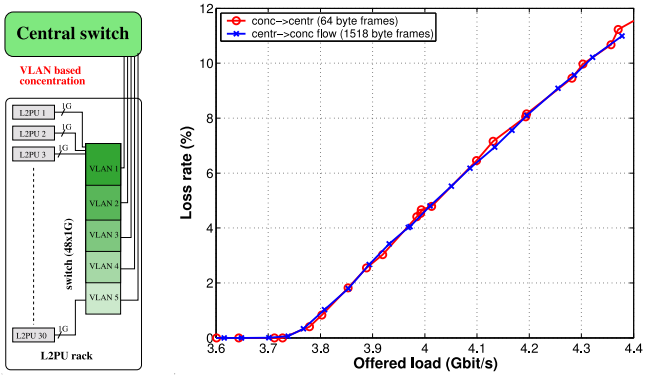


(a) topology (b) test result

Fig. 3. Trunking based concentration

partition the switch: each VLAN concentrates 6 L2PUs and connects to the network core via a GE up-link (see Fig. 4(a)).

In the test setup we used four links between the two switches in order to keep the test results comparable to the ones obtained for trunking. Four VLANs were defined on the concentrator switch: three of them concentrate five tester ports each, while the fourth one aggregates traffic from only four ports. The test results are depicted in Fig. 4(b): a throughput of approximately 3.8 Gbit/s is achieved for both traffic directions. Due to uniform traffic distribution on all the tester ports, the up-links corresponding to the five-tester-ports VLANs are loaded more than the up-link of the four-tester-ports VLAN. For an aggregated throughput in the (3.8, 4.75) Gbit/s range, the three heavier loaded links are saturated, while the fourth one has spare capacity.



(a) topology (b) test result

Fig. 4. VLAN based concentration

The slightly unbalanced distribution of tester ports on VLANs highlights a drawback of this concentration method: there is no possibility to share the load between the up-links of the same concentrator switch.

C. 10GE up-link concentration

This technique takes advantage of the 10GE up-link ports available on the recent generation “pizza-box” Ethernet switches. A 10GE line is used to connect a concentrator switch to the network core. The central switches are chassis-based devices, and nowadays all such models support 10GE line-cards. Moreover the price per bandwidth is similar for 10GE and 1GE line-cards.

As one can notice, the use of one 10GE up-link for 30 L2PUs provides twice the required bandwidth. While preserving the L2PUs rack encapsulation, two workarounds to achieve the 6 L2PUs to 1 Gbit/s concentration factor are available:

- the use of two-to-one oversubscribed ports on the central switches. Even if the link can run at 10GE, the switching bandwidth of the central switch port is only 5 Gbit/s.
- the concatenation of two concentrator switches: a concentrator switch connects through a 10GE line to another concentrator switch, which further uses its second 10GE port as an up-link to the network core.

The concatenation of two concentrator switches has a slight impact on the fault tolerance of the system: if the up-link to the central switch fails, the number of L2PUs lost by the TDAQ system doubles, as both concatenated switches are disconnected from the network core. Fortunately losing a fraction of L2PUs is not critical for the system; it will continue to operate at a lower rate. The 60 L2PUs which would be lost in case of the up-link failure represent slightly more than 10% of the total number of L2PUs. The corresponding decrease in the LVL2 accept rate is tolerable for short repair periods.

A single concentrator switch connected with a 10GE link to the central one was used in our test setup (see Fig. 5(a)). The results from Fig. 5(b) show that approximately 9 Gbit/s can be transfered from the central switch to the concentrator one, and virtually 10 Gbit/s on the reverse path. The discrete event simulation results (included in Fig. 5(b)) are in agreement with the experimental results: the 1518 byte frames flow experiences a higher loss rate than the 64 byte frames one. It is intuitive to expect a higher loss rate for the 1518 byte frames flow, as the congestion on the 10GE up-link port is stronger (the number of ports sending 1518 byte frames is three times higher than that of ports sending 64 byte frames).

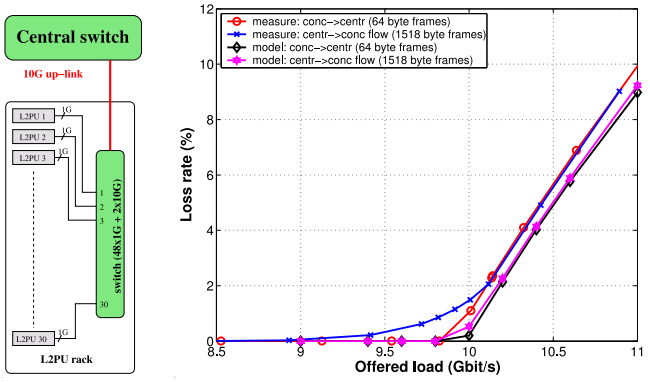
The 10GE concentration option seems the most appropriate. It is the simplest and the most efficient in using the network core bandwidth, as shown by both test and modeling results.

V. Architecture Upgrade Proposal

Fig. 6 depicts the architecture we propose for implementation: the network core is distributed over two large switches, each of them handling a mixture of LVL2 and EB traffic; a separate central control switch is used to aggregate the control traffic, while the LVL2 processing units are concentrated using 10GE concatenation.

A. Full system modeling: VLAN vs. 10GE based concentration

The discrete model simulation of the full size TDAQ system showed similar performance for the VLAN based and the



(a) topology (b) test result

Fig. 5. 10GE based concentration

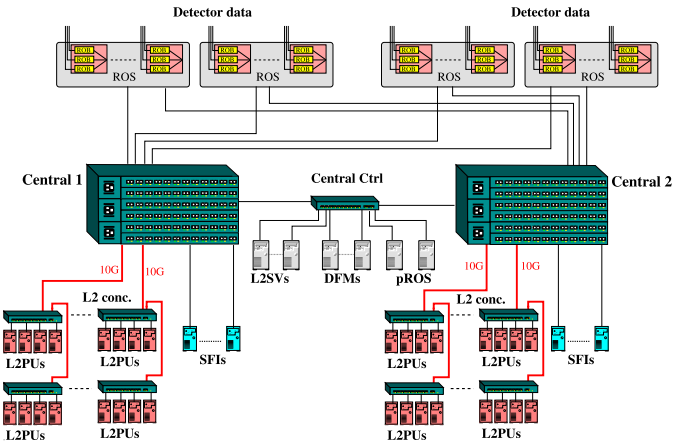


Fig. 6. Architecture upgrade proposal

10GE concentration. The distribution of the switches queue occupancy at the congestion points on the data path from the ROS to the L2PU is presented in Fig. 7. The 10GE concentration alleviates congestion at the network core (output of the 10GE central switch port), but introduces a new congestion point in the concentrator switch (GE port output to the L2PU). However the congestion degree is small in both cases, and the probability of queues growing above 15 elements is inferior to 10^{-6} (see Fig. 7). As long as switches have buffers that can accommodate more than 15 frames, the potential losses at the analyzed congestion points will not cause a system performance drop.

VI. Conclusion

Key aspects of the DataFlow network architecture have been analyzed and optimized. Distributing the network core over two devices, each handling a mixture of level two and event building traffic, improves the fault tolerance of the system, while reducing the size of the required switches.

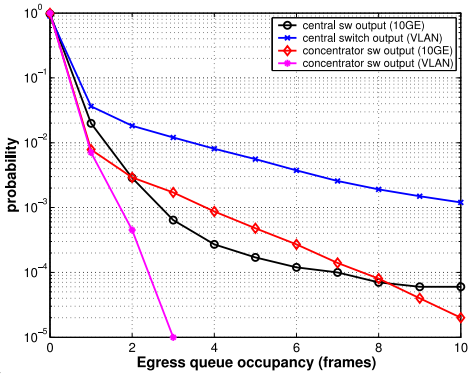


Fig. 7. Egress queue occupancy distribution on a full system model: central switch output to the concentrator, concentrator switch output to the L2PUs.

The use of concentrator switches with 10GE up-links simplifies the process of aggregating L2PUs, guarantees an efficient use of the up-link bandwidth, and improves the flexibility of adding/removing processing power to/from the level two trigger.

The discrete model simulation predicts a smooth operation of the full TDAQ system running on the proposed network architecture. The next validation step is deploying a similar topology in the pre-series testbed.

Acknowledgment

The authors would like to express their gratitude to the ATLAS TDAQ collaboration for providing constant support and feedback that guided our work.

References

[1] S. Stancu, B. Dobinson, M. Ciobotaru, K. Korcyl, and E. Knezo, "The use of Ethernet in the Dataflow of the ATLAS Trigger and DAQ," *ECONF*, vol. C0303241, p. MOGT010, 2003.

[2] H. Beck, B. Dobinson, K. Korcyl, and M. LeVine. (2003, Feb.) ATLAS TDAQ: A Network-Based Architecture. [Online]. Available: https://edms.cern.ch/file/391592/2.2/DC-059.pdf

[3] R. Cranfield, P.Golonka, A. Kaczmarska, K. Korcyl, J. Vermeulen, and S. Wheeler, "Computer Modeling the ATLAS Trigger/DAQ System Performance," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 3, pp. 532–538, June 2004.

[4] M. Ciobotaru, S. Stancu, M. LeVine, and B. Martin, "GETB, a Gigabit Ethernet Application Platform: its Use in the ATLAS TDAQ Network," in *Proc. IEEE Real Time Conference 2005*, Stockholm, Sweden, June 2005, in press.

[5] ATLAS HLT/DAQ/DCS Group, *ATLAS High-Level Trigger Data Acquisition and Controls Technical Design Report*. CERN/LHCC/2003-022, Oct. 2003.

[6] J. Vermeulen *et al.*, "ATLAS DataFlow: the Read-Out Subsystem, Results from Trigger and Data Acquisition System Testbed Studies and from Modeling," in *Proc. IEEE Real Time Conference 2005*, Stockholm, Sweden, June 2005, in press.

[7] B. Greear. (2003, Sept.) 802.1Q VLAN Implementation for Linux. [Online]. Available: http://www.candelatech.com/~greear/vlan.html