ATL-DAQ-CONF-2006-010
23 June 2006

NETWORKS FOR ATLAS TRIGGER AND DATA ACQUISITION

S. Stancu, M. Ciobotaru, University of California, Irvine, Irvine CA 92697-4575, USA
C. Meirosu*, L. Leahu*, B. Martin, CERN, 1211 Geneva 23, Switzerland

Abstract

The ATLAS experiment will rely on Ethernet networks for several purposes. A control network will provide infrastructure and operational services, and two dedicated data networks will be used exclusively for transferring the event data within the High Level Trigger and Data Acquisition system. This article presents the networking architecture solution for the whole ATLAS TDAQ and illustrates sample performance evaluation results, meant to validate the design concepts.

INTRODUCTION

The ATLAS TDAQ (Trigger and Data Acquisition) system uses 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. While the first level trigger is constructed in dedicated hardware, the second and third level triggers are implemented using distributed systems built of a large number of interconnected PCs.

Fig. 1 illustrates the block diagram of the TDAQ system. The events validated by the first level trigger are pushed at 100 kHz into ≈ 1600 read-out buffers (ROBs) distributed over ≈ 150 ROS (Read-Out System) PCs. The second level trigger analysis task is distributed over approx 500 L2PUs which use an RoI based mechanism to retrieve from the ROSs and analyze the meaningful event data (typically 2% of the entire event). Following a level two accept (approx 3.5 kHz), an event builder system distributed over approximately 100 SFIs (Sub Farm Inputs) is used to gather the scattered event data from all ROSs. A few supervision applications (SVs) perform the event information book-keeping and load balance the tasks on the components of the L2 (level two trigger) and EB (event builder) sub-systems. The SFIs buffer the events and provide them further to the third level trigger (also denoted as Event Filter – EF) on demand. Approximately 1600 Event Filter processors (EFPs) perform the last and most thorough step of the trigger analysis. The EF validated events are temporarily stored on a handful of SFOs (Sub-Farm Outputs) and finally sent to mass storage at approximately 200 Hz.

The trigger and data acquisition system relies directly on two dedicated "data" networks (see Fig. 1): the *FrontEnd network* (used for the ROS, L2 and EB subsystems interconnection) and the *BackEnd network* (provides the communication for the EF sub-system). In addition, all the TDAQ components are connected to a *Control network* meant to provide infrastructure and operational services (e.g. shared file systems, databases, run-control).

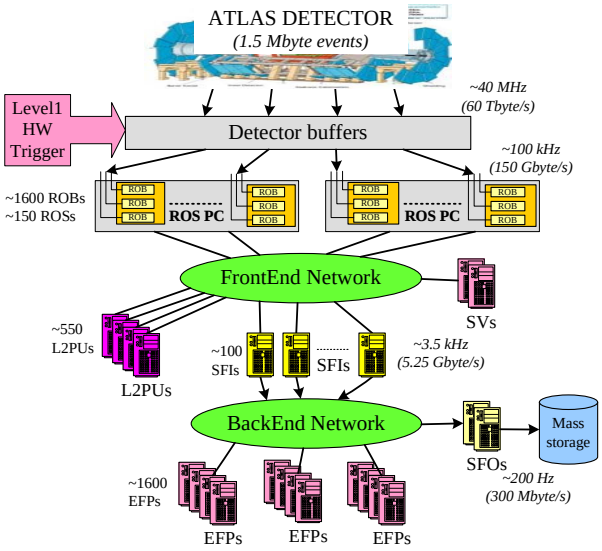

Figure 1: ATLAS TDAQ system block diagram.

Following an overview of the technology and equipment choice for building the ATLAS networks, we shall describe the architectural solution for each of the three TDAQ networks. Since high availability is a key feature of the entire TDAQ system during data-taking periods, results from a proof of concept performance/resiliency test will be presented. Finally, aspects related to the network management and installation will be exposed.

TECHNOLOGY AND EQUIPMENT

The TDAQ choice of the Ethernet technology for the networking infrastructure is justified not only by the good price–performance ratio (most Ethernet products are commodity), but mainly by the fact that it is multi-vendor, and we foresee long term support for it [1].

End-nodes will be preponderantly equipped with copper GE (Gigabit Ethernet – 1000BaseT) interfaces. The network will be built using switching/routing equipment that has become "standard" for large enterprise networks:

- chassis-based devices with over 300 Gbit/s full duplex switching bandwidth. Built-in redundancy (power

* Also affiliated with "Politehnica" University of Bucharest, Bucharest, Romania

supplies, switching fabric) is a common feature for those devices. Typically, the interface modules (line-cards) contain either 40 GE ports or 4 Ten Gigabit Ethernet (10GE) ports.

- "pizza box" devices with 40 to 60 Gbit/s full-duplex switching bandwidth. These switches come in a 1U or 2U format and have 24/48 GE ports and an optional module with up to two 10GE interfaces.

With the appropriate software support, these devices can operate either at layer 2 (Ethernet switches) [2] or at layer 3 (IP routers) [3].

Since the number of end-nodes to be interconnected is large (e.g. ≈3000 for the control network) and most of them have modest bandwidth requirements [4] compared to their network interface speed (GE), it is natural to group them into clusters using "pizza box" switches as concentrators. The network core is composed of chassis-based devices, receiving GE or 10GE up-links from the concentrator switches, and direct GE connections from the applications with high bandwidth requirements.

During data-taking periods the TDAQ system is expected to operate round the clock (24/7). The TDAQ networks design takes into account this high availability requirement. The next subsection will give a brief overview on the available techniques for building resilient Ethernet networks.

Resilient Ethernet networks

The only way to build a resilient system is to introduce redundancy into it. For the case of Ethernet networks redundancy can be introduced either at the component level (e.g. devices with redundant power supplies and switching fabric), or at the network level (deployment of additional devices/links in order to provide alternate paths between communicating nodes).

Several protocols are available for achieving the network level redundancy [5]:

- *Layer 2 protocols* – trunking and spanning tree. Link aggregation (trunking) [6] enable multiple physical aggregated links to appear as if they were a single logical link. The Multiple Spanning Tree (MST) protocol [7] is the most efficient protocol for achieving topology level redundancy in a Layer 2 network. If multiple paths are present in the topology, only one of the paths is allowed to be active, while the others are kept in stand-by mode. Used in conjunction with VLANs [8], MST allows the efficient use of multiple paths (one path can be active in one VLAN, and a stand-by path in another VLAN). See [9] for more details.

- *Layer 3 protocols.* While dynamic routing protocols, such as RIP [10] or OSPF [11] have "built-in" support for multiple traffic paths, static routed environments are sensitive to the failure of the default gateway. The VRRP [12] eliminates this single point of failure, by emulating a "virtual router" (made up of multiple devices) which acts as the default gateway.

TDAQ NETWORKS ARCHITECTURE

In this section we will detail the architecture of each of the three TDAQ networks, and also address the possibility of moving processing power between different stages of the trigger in order to optimize its efficiency.

Control network

The terminology for this network originates from the fact that it is used to support the run-control traffic for TDAQ applications. In addition, the control network will provide various infrastructure (e.g. shared file systems) and TDAQ specific (database access, operational monitoring) services.

A total of approximately 3000 end-nodes are interconnected by the Control network (see Fig. 2). The redundant network core can be implemented either as a single device with built-in full redundancy, or two interconnected devices (still with a certain redundancy degree, e.g. power supplies). As the main traffic in the control network is constituted by the flow of data between the infrastructure service PCs and the end-nodes, the former ones are connected directly and redundantly to the core. The end-nodes (e.g. L2PUs, EFPs) are clustered at the rack level using concentrator switches with two up-links connected to the core.

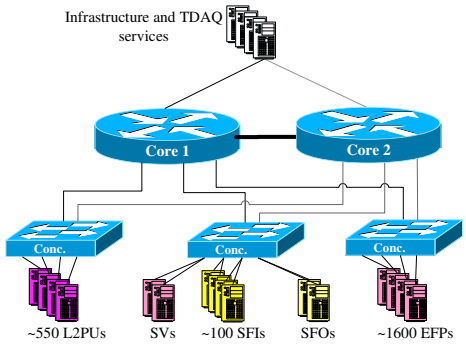

Figure 2: control network diagram.

Since there are no special performance requirements, and the network comprises a large number of end-nodes, it will be operated at layer 3. Static IP routing should suffice, since the network topology is simple. The core devices shall act as IP routers, and have one sub-net per concentrator switch. Thus the layer 2 (Ethernet) broadcast domains are small (one domain per rack), and eventual problems (high rate broadcasts, flooding) remain local to the rack level sub-net.

FrontEnd network

The FrontEnd network (also denoted as DataFlow network) must provide a cross sectional bandwidth of the order of 100 Gbit/s (half for L2, half for EB) with a minimum

of packet loss. While the design approach of this network has been detailed in [1, 13], we shall only give a brief presentation of the network architecture (depicted in Fig. 3).

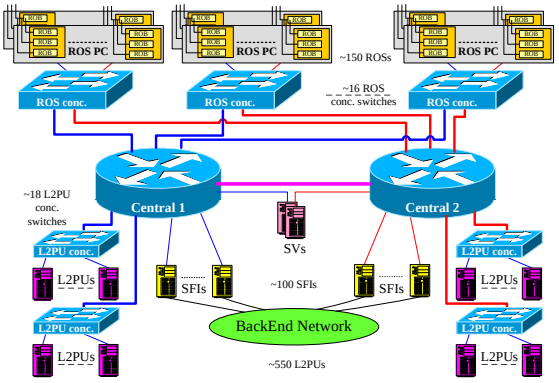

Figure 3: FrontEnd network diagram.

The use of two chassis switches interconnected through a high speed link (e.g. two aggregated 10GE lines) for the FrontEnd network core improves the fault tolerance of the system (the system still operates at half rate in the event of the failure of one of the core devices).

The L2PUs are concentrated at the rack level: 30 L2PUs connect first to an aggregation switch which is further linked to the core through a 10GE up-link. The SFIs and the SVs applications connect directly to the core devices. While the ROSs are located down the ATLAS detector pit, the rest of the components are housed in a surface building. Since the distance between the ROSs and the central switches is higher than 100m, copper GE cannot be used. Instead of using fibre GE NICs (Network Interface Cards) on the ROSs it is more convenient to introduce an additional layer of "concentrator" switches located in the proximity of ROSs. ROSs are equipped with copper GE interfaces connected to the concentrator switches. Each concentrator switch has 20 GE inputs from the ROSs and two 10GE up-links, connecting each to a core device.

We plan to operate this network at layer 2. In this case loops appear in the network between the ROS concentrator switches and the two core devices. This problem can be solved by using VLANs and MST. Two VLANS are defined on all ROS concentrator switches and the 2 core devices, and MST is configured to maintain only one VLAN active per central switch. This provides good redundancy: for example if one of the links connecting the Central-1 device to a ROS concentrator switch fails, the MST will re-converge and the L2PUs/SFIs attached to the Central-1 switch will be able to reach the ROSs from the affected concentrator switch through the *Central-1 – Central-2 – ROS concentrator switch* path.

BackEnd network

The second dedicated data acquisition network must interconnect ≈1700 end-nodes and withstand a throughput of approximately 50 Gbit/s. As illustrated in Fig. 4, we chose to implement the network core using a single device with built-in full redundancy (power supplies, switching fabric, supervisor modules). The Event Filter processing units are concentrated at the rack level. An aggregation switch connects the EFPs from the rack to the network core through a redundant link (link aggregation of two GE lines). The rest of the applications (SFIs and SFOs), which require a higher throughput, are connected directly to the core.

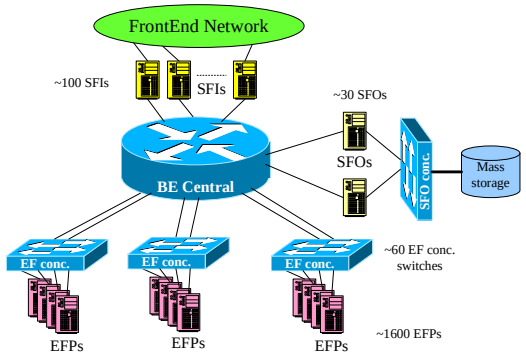

Figure 4: BackEnd network diagram.

Similar to the Control Network, the BackEnd network shall be operated at layer 3 (IP routing at the core, with one sub-net per concentrator switch) in order to restrict Ethernet broadcast domains to the rack level.

Interchangeable processing power

Since it is difficult to foresee *a priori* the efficiency of the triggers, it is desirable to have the ability to rapidly move processing power between the level-2 trigger and the level-3 trigger (Event Filter). This feature can be achieved by using processor racks with a data switch connected both to the FrontEnd network core (via a 10G up-link) and the BackEnd network core (via two aggregated GE lines). The connectivity of the rack to one or the other networks can be modified at ease by software enabling/disabling of the desired up-links.

SAMPLE RESILIENCY TEST

Since the FrontEnd network resiliency relies on the appropriate usage of multiple links in conjunction with VLANs and the MST protocol, we have performed an evaluation test of this technique. Two 10GE links have been used to interconnect two devices. The links are part of two VLANs (VLAN1 and VLAN2) and the MST is configured to maintain active VLAN1 on link-1 and disable it on link-2 and vice-versa. Traffic generators [14] have been used to stream data across the links within both VLANs.

The throughput across the connection between the two devices has been recorded for each of the VLANs (see Fig. 5). For the time periods corresponding to the AB, CD and EF segments both links are operational. MST is configured to activate link-1 in VLAN1 and disable it in VLAN2,

and activate link-2 in VLAN2 and disable it in VLAN1. Thus, the throughput for each VLAN equals the line-speed capacity. For the BC and DE time periods one of the links between the two devices is broken. The MST re-converges and enables the remaining active link in both VLANs. The two traffic flows get a fair share of 5 Gbit/s each.

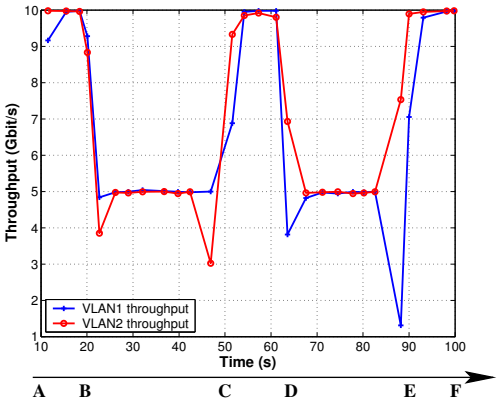

Figure 5: Sample resiliency test using MST and VLANs.

MANAGEMENT AND INSTALLATION

It is good practice not to mix management traffic and normal traffic. If the normal traffic has an abnormal pattern it may overload some links and potentially starve the management traffic. This is why we foresee to use an "out of band" device interface for management/monitoring purposes. While most chassis-based devices have a real "out of band" management interface, the pizza-box devices rarely do. This problem is overcome by isolating a port (e.g. the highest number port) in a VLAN dedicated for management. The management/monitoring can be done via a small dedicated layer 2 network, providing access to the management interfaces of all devices.

Since the ATLAS experiment is foreseen to start taking data in 2007, the equipment installation (including network devices) has already started. Maintaining an accurate image of the active device information proves to be tedious. This is why we have developed a tool which auto-discovers the network topology based on the MAC (Media Access Control) address table information on the switches. A sample auto-generated topology is illustrated in Fig. 6. However, in order to have a proper bookkeeping of the active installation, we need to interface this tool to the installation database.

CONCLUSION

In this paper we have presented the design choice for the three networks of the ATLAS TDAQ system. All networks are implemented in Ethernet technology using devices available from multiple manufacturers. Due to the flexibility and high availability requirements for the network, we have chosen to deploy a modular architecture

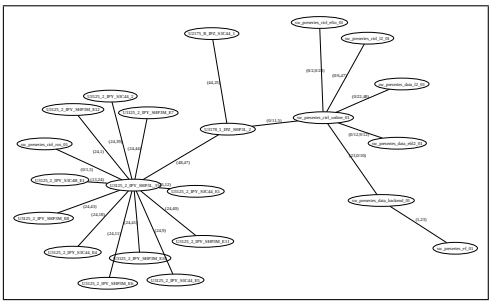

Figure 6: Sample auto-generated network topology.

with aggregation switches at the rack level and redundant links to a network core. Issues related to the active installation bookkeeping and management of devices have been also addressed. For more details on the operation model of the TDAQ networks refer to [15].

ACKNOWLEDGEMENTS

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," in *Proc. CHEP 03 Conference*.

[2] *Media Access Control (MAC) Bridges*, IEEE Std. 802.1d.

[3] J. Postel, "Internet Protocol," RFC 791, Sept. 1981.

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

[5] T. Sridhar. (2004, July) Redundancy: Choosing the Right Option for Net Designs. [Online]. Available: http://www.commsdesign.com/showArticle.jhtml?articleID=25600515

[6] *Aggregation of Multiple Link Segments*, IEEE Std. 802.3ad.

[7] *Multiple Spanning Trees*, IEEE Std. 802.1s.

[8] *Virtual Bridged Local Area Networks*, IEEE Std. 802.1Q.

[9] Cisco white paper. Understanding Multiple Spanning Tree Protocol (802.1s). [Online]. Available: http://www.cisco.com/warp/public/473/147.html

[10] G. Malkin, "RIP Version 2," RFC 2453, Nov. 1998.

[11] J. Moy, "OSPF Version 2," RFC 2328, Apr. 1998.

[12] R. Hinden, "Virtual Router Redundancy Protocol (VRRP)," RFC 3768, Apr. 2004.

[13] S. Stancu, M. Ciobotaru, and K. Korcyl, "ATLAS TDAQ DataFlow Network Architecture Analysis and Upgrade Proposal," in *Proc. IEEE Real Time 2005 Conference*.

[14] 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*.

[15] C. Meirosu, B. Martin, A. Topurov, and A. Al-Shabibi, "Planning for Predictable Network Performance in the ATLAS TDAQ," in *Proc. CHEP 06 Conference*.