Network Resiliency Implementation in the ATLAS TDAQ System

S. N. Stancu, A. Al-Shabibi, S. M. Batraneanu, S. Ballestrero, C. Caramarcu, B. Martin, D. O. Savu, R. V. Sjoen, L. Valsan

Abstract—**The ATLAS TDAQ (Trigger and Data Acquisition) system performs the real-time selection of events produced by the detector. For this purpose approximately 2000 computers are deployed and interconnected through various high speed networks, whose architecture has already been described. This article focuses on the implementation and validation of network connectivity resiliency (previously presented at a conceptual level). Redundancy and eventually load balancing are achieved through the synergy of various protocols: link aggregation, OSPF (Open Shortest Path First), VRRP (Virtual Router Redundancy Protocol), MST (Multiple Spanning Trees). An innovative method for cost-effective redundant connectivity of high-throughput high-availability servers is presented. Furthermore, real-life examples showing how redundancy works, and more importantly how it might fail despite careful planning are presented.**

I. Introduction

THE events recorded by the ATLAS [1] detector are filtered in real-time by a three-level trigger. Fig. 1 illustrates the block diagram of the second and third level trigger of the TDAQ system. The events (approximately 2 Mbyte average size) validated by the first level are buffered in the ROSs (Read-Out Systems). The second level trigger, performs an RoI (region of interest) based event analysis at 100 kHz. An event builder farm composed of SFI (Sub-Farm Interface) applications builds and stores the level-2 validated events at a rate of approximately 5 kHz. The third level trigger (denoted as Event Filter) retrieves full events from the SFIs, analyzes them and passes only the validated ones to the SFOs (Sub Farm Output) at an average rate of approximately 300 Hz. The SFOs provide intermediate disk storage before the permanent recording of the events.

All these functions are performed by two thousand computers inter-connected by high speed Ethernet networks, whose architecture and operational model were described in [2] and [3] respectively. The two dedicated data networks (*FrontEnd* and *BackEnd*) and the *Control* network are implemented using a total of five routers and more than 100 switches, all relying on 10GE (Ten Gigabit Ethernet) and copper GE (Gigabit Ethernet) technology.

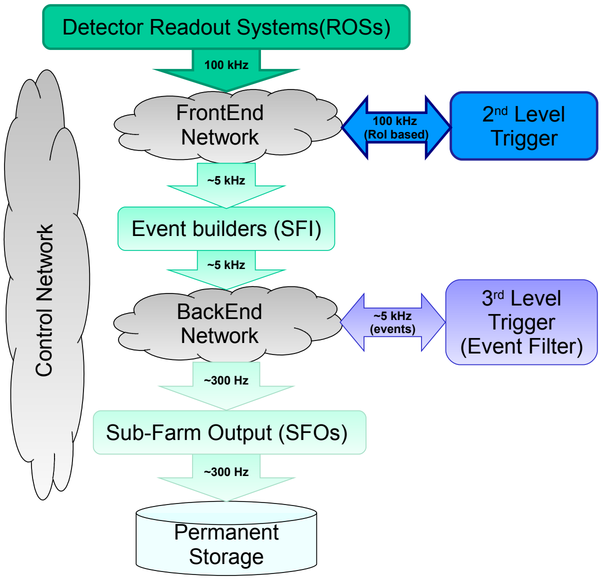


Fig. 1. TDAQ system block diagram

This paper details how the foreseen redundant network links are used to achieve resiliency: what protocols are used, how they are configured and how they have proven to behave in practice. In particular a new (to the best of the authors' knowledge) cost-effective method for connecting high-throughput servers directly to routers is presented.

II. Routers Interconnections

Fig. 2 illustrates the inter-connections between the TDAQ routers, together with the connection to the rest of the networks from the ATLAS experimental site. The Control network has two routers, interconnected by a trunked link made of two 10GE links. One of the routers (the bottom one in the figure) has a backup role: under normal conditions it carries no traffic at all, but it can take over the entire functionality of the primary one (the top one in the figure). Each router from the data networks (currently two routers in the FrontEnd network and one router in the BackEnd network) connects to both Control routers through GE links. These connections to the dedicated data networks are only used for management purposes, e.g. SNMP (Simple Network Management Protocol) monitoring of devices or providing access to a central DHCP [4] service.

Manuscript received May 28, 2010.

S. N. Stancu and S. M. Batraneanu are with University of California at Irvine, Irvine, California, USA and "Politehnica" University of Bucharest, Bucharest, Romania

A. Al-Shabibi is with CERN, Geneva, Switzerland and University of Heidelberg, Heidelberg, Germany

S. Ballestrero is with CERN, Geneva, Switzerland and University of Johannesburg, Johannesburg, South Africa

C. Caramarcu is with Horia Hulubei National Institute of Physics and Nuclear Engineering - IFIN HH, Bucharest, Romania

B. Martin and D. O. Savu are with CERN, Geneva, Switzerland

R. V. Sjoen is with Department of Informatics, University of Oslo, Oslo, Norway and CERN, Geneva, Switzerland

L. Valsan is with CERN, Geneva, Switzerland and "Politehnica" University of Bucharest, Bucharest, Romania

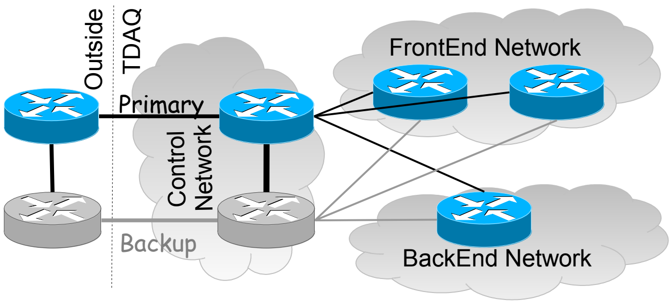


Fig. 2. TDAQ routers interconnections

A. Routing Configuration

The OSPF [5] routing protocol is enabled on all the routers. A separate area is used for the TDAQ networks, and all the inter-router links are part of it. Each router is configured to re-distributes its "connected" routes (i.e. routes to networks directly provided by the routers), so as long as there is at least one valid link to a subnetwork, it will be known in the entire area.

The two Control routers (primary and backup) are connected using two 10GE links to a pair of primary–backup routers from outside the TDAQ area, and the interface is done using static routes, in order to minimize any potential interference:

- On the outside routers, all TDAQ network prefixes are statically routed to both the primary link (with a low cost) and the secondary link (with a high cost).
- On the TDAQ side, each router has two default gateways: a low cost one pointing to the primary link, and a high cost one pointing to the backup link.

Thus, this setup implements a failover mechanism: the primary link is used as long as it is operational, while the higher cost backup link would become used only in the event of the primary link failure.

One of the potential problems of static routing are loops. An IP [6] packet coming from outside TDAQ and addressed to an inactive TDAQ subnet, will be sent back-and-forth between the "outside" and the TDAQ routers until its TTL (time to live) is exceeded: the outside routers will send it to TDAQ based on the destination IP matching a TDAQ network prefix; the TDAQ routers will not have a route to the inactive subnet, and will send it back to the default gateway. To avoid this problem the TDAQ routers have been configured to route all TDAQ network prefixes to the *null* interface with the highest possible cost.

B. Practical Experience

No failure has been experienced on any of the inter-router links. Nevertheless, tests have been performed while commissioning the network. The primary link between "outside" and TDAQ, as well as inter-router links from within the TDAQ areas have been shutdown, and traffic was rerouted on the alternate paths within a few seconds.

III. Edge Switches Connectivity – VRRP

Fig. 3 illustrates two racks of computers, whose edge switches are connected to both Control routers. VRRP [7] is used to make each computer host think it talks to a single virtual router, which is implemented either by the master router ($R1$ in the figure), or by the backup router ($R2$ in the figure) in the event of the master's failure.

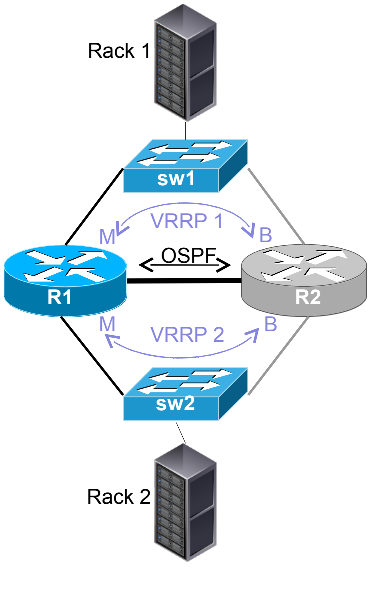


Fig. 3. Edge switches redundant connectivity using VRRP

A. VRRP Configuration and Operation

Only the configuration for Rack1 will be described, the one for Rack2 being identical. Let Rx_{swy} denote the interface of router Rx connected to the switch swy. Interfaces $R1_{sw1}$ and $R2_{sw1}$ provide the same subnet (*subnet_1*) to sw1, and implement a VRRP instance inside this subnet. The two router interfaces exchange VRRP protocol messages through sw1, and the router with the highest priority (R1 in the figure) becomes master. The master router implements the "virtual" router, with IP address *vrrp_ip1* and MAC (Media Access Control [8]) address *vrrp_mac1*, and all hosts in Rack1 are configured to use *vrrp_ip1* as default gateway.

When all the links are operational the two racks communicate through the primary router (sw1–R1–sw2). If link R1–sw1 goes down, the VRRP instance running on $R2_{sw1}$ will no longer see the master router, and will become master itself. Hosts in Rack1 continue to talk through the default gateway (*vrrp_ip1* and *vrrp_mac1*), but this gateway is now implemented by R2. The communication between the two racks will be asymmetric: Rack1 will talk to Rack2 through sw1–R2–sw2, while Rack2 will talk to Rack1 through sw2–R1–R2–sw1 (since OSPF is configured to redistribute connected routes – see Section II-A – when $R1_{sw1}$ goes down because of the link failure, R1 will know to route traffic to subnet_1 to R2). Under certain circumstances asymmetric traffic may lead to flooding [9] so this is only an emergency operation mode.

In the TDAQ Control network a single VRRP instance is used for each subnet (R1 is always the master, R2 the backup), providing redundancy, but no load balancing. In case higher bandwidth is required, two VRRP instances could be ran in each subnet, providing both redundancy and load balancing, with the caveat of asymmetric traffic [9].

B. Practical Experience

No practical failures have occurred to date on any edge switch uplink, but the replacement of the cable management kit on the primary router allowed us to perform a test in realistic conditions. The TDAQ system was idle at the time of the intervention, but the hosts from all the racks were up and continued to run despite the fact that all the edge switches' uplinks to the primary router were shutdown.

The TDAQ system installation was incremental in terms of both networking equipment and servers. We will further present two issues that were encountered as the system grew in size, the backup Control router being deployed only in one of the last stages.

1) Non-conform VRRP Implementation: By means of proxy ARP (Address Resolution Protocol) [10] the routers have been configured to make the entire Control network (divided in class C networks at the router level) look like a single class B subnet for the end-hosts. This is common practice on local area networks and is aimed at simplifying the network parameters configuration of the end-hosts. When deploying VRRP with proxy ARP enabled, we discovered that when a host issued an ARP request meant to be proxied by the router, it received back two ARP replies:

- a legitimate one from the master router, with the proper VRRP MAC address.
- a spurious one, from the backup router, with the MAC address of the router's physical interface

Depending on which ARP reply was received first, the host would either behave correctly (correct ARP received first), or would use the backup router (spurious ARP received first). This nondeterministic load balancing, coupled with the introduction of asymmetric traffic made us postpone the VRRP deployment until the manufacturer suppressed the spurious ARP reply. It is worth noting that a simple test of disabling the primary link could have been deceiving, as all connectivity would have been preserved: the traffic using the correct MAC would have flipped onto the backup link to the new master, and the one using the spurious MAC address would have continued to run on the backup link.

2) Proxy ARP Limitation: As previously described, proxy ARP was enabled in the initial installation stage, in order to simplify the end-hosts configuration. As the system grew in size, sporadic TCP [11] connection failures were observed when a large number of hosts (above 1000) were simultaneously trying to contact the same pool of a few servers. After obtaining the precise error of the communication failure ("no route to host"), it was discovered that the router could not proxy all the ARP requests fast enough. This limitation was enforced by a denial of service protection mechanism that was reducing the rate of broadcast messages passed to the router's processor. In order to avoid any further scalability problems due to this limitation, proxy ARP has been disabled and the network configuration of the hosts has been adapted to match their specific class C subnet. The problem hasn't been observed ever since.

IV. High-Throughput High-Availability Servers

The TDAQ system has a number of approximately seventy infrastructure and monitoring servers that require high throughput. Since part of them are essential for the operation of TDAQ, we were confronted with the need of providing reliable high-throughput connectivity for these servers.

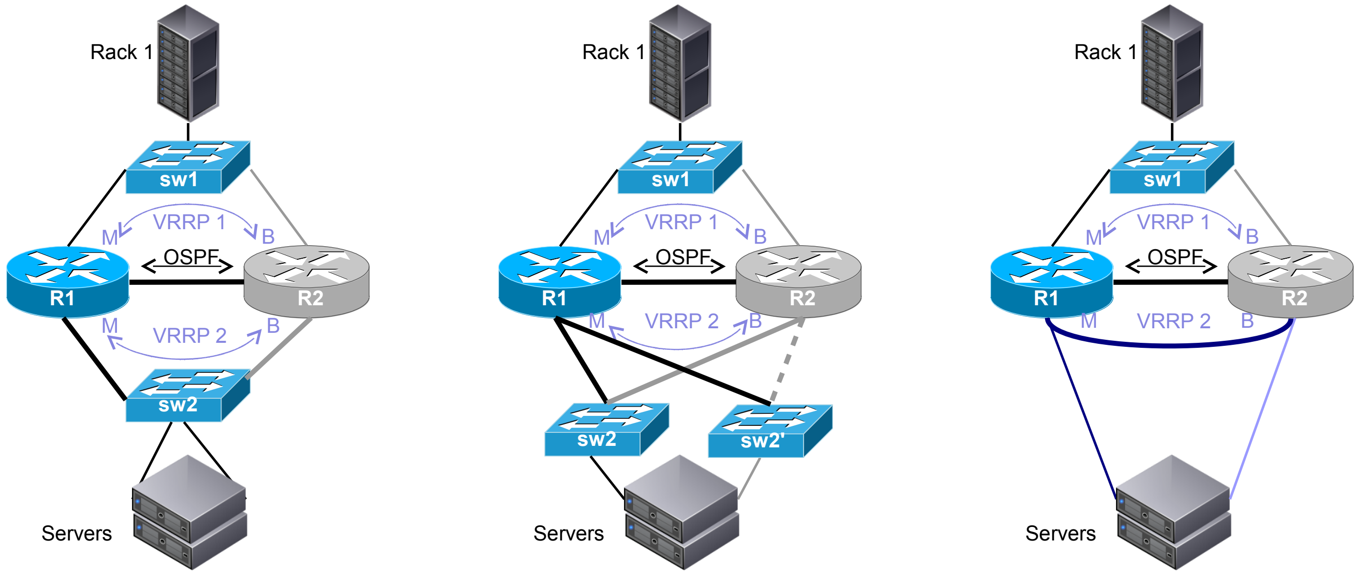
A. Achieving High-Throughput Redundant Connectivity

Since all the servers are equipped with two on-board Ethernet ports, we use both ports and bond them in a single Linux interface [12]. Fig. 4 illustrates the available options for connecting the servers (VRRP is configured on the two routers as previously explained in Section III):

a) *Single edge switch with 10GE uplinks.* The servers are connected to an edge switch through a bond interface that can be configured to load balance the traffic on both links. The edge switch is a single point of failure in this setup.

b) *Two Edge switches with 10GE uplinks.* Since the two links from each server connect to different switches, the bond interface must be configured in the *active-backup* mode (only one of the two links is used at any time). As each router has two interfaces to the rack, the IP interface can no longer be a physical interface, but a VLAN [13] one. The two edge switches (each of them with uplinks to both routers) will create an Ethernet loop, which must be broken by activating a VLAN aware spanning tree protocol (e.g. MST [13]). This option has no single point of failure, but its cost is rather high, as it involves the use of four 10GE ports out of which only one is forwarding traffic at any time (25% efficiency).

c) *Direct router connection.* The servers are directly connected to the routers: one link to R1, the other one to R2. As described in (b) above, the bond comprising the two links must be configured in the *active-backup* mode. Each router will have a VLAN IP interface for a group of servers (typically the servers hosted in the same rack). If the primary interface of a server fails, the bonding module will rapidly switch to the backup one. Since the rest of the servers have an active primary link, the rest of the network will try to address the server through R1's VLAN interface, which is still active. A connection between the two VLANs on R1 and R2 is needed in order to render the server with the faulty primary link accessible from R1's VLAN interface: R1 – R2 – backup-interface. It is essential that the VLAN link between the routers is redundant. Since the inter-router connection used by OSPF is already redundant, we have chosen to use the same link by including it as a tagged member of all VLANs serving directly connected servers. The link between the two VLAN interfaces of the routers allows the use of VRRP in the same way as for (a) and (b) above, in order to cope with an eventual router failure.
To the best of the authors' knowledge, this method is a novelty. This option has no single point of failure either, but in comparison with option (b) it is clearly more cost-effective: it only uses copper GE ports on the router (with 50% efficiency) and no additional switches. Nevertheless,



(a) Single edge switch (10GE uplinks)

(b) Two edge switches (10GE uplinks)

(c) Direct router connection

Fig. 4. Options for redundantly connecting high-throughput high-availability servers

inter-rack cabling is required, which can be difficult to handle if racks are far apart or the number of servers is high.

The direct router connection method (option c) has been deployed for all critical TDAQ servers, as well as for a dedicated Network Attached Storage unit used ATLAS-wide.

B. *Practical Experience*

The redundant connection was exercised on two occasions: the failure of a server's primary interface (noticed only by the monitoring system and subsequently reported by the shifter) and a planned shutdown of the primary router's copper GE interfaces (also unnoticed, except for a handful of servers that were depending on a file-server with a missing backup link).

V. Redundancy with Load Balancing – MST

Due to its high throughput and low latency requirements, the FrontEnd network is implemented using two core routers [2], but only their Ethernet switching features [8] are used. As depicted in Fig. 5, a full vertical slice of the system (input from one of the two interfaces of each ROS, output to half of the Trigger farms) is associated to each core device.

Since the distance between the ROSs (approximately 150 PCs, located underground) and the rest of the TDAQ system (installed at the surface) exceeds 100 meters, copper GE cannot be used. The initial design of using fibre GE interfaces on the ROSs was abandoned when 10GE became affordable. Instead, a layer of concentrator switches is installed underground in order to aggregate groups of up to 10 copper GE interfaces from the ROSs into one 10GE fibre link. Since a failure of one 10GE link would impact up to 10 ROS interfaces, we rely on the MST protocol [13] for providing both redundancy and load balancing over these 10GE links.

A. *MST Configuration*

All the 10GE uplinks of the ROS concentrator switches plus an additional high speed link (four trunked 10GE links between the two core devices) belong to two VLANs. All the switches to which these links are connected run two MST instances, one VLAN being mapped to the first instance, and the other VLAN to the second instance. The priorities of each core are configured such that Core1 is the *root* of the first MST instance, and Core2 is the *root* of the second instance. The state of each uplink will be *Forwarding* in one MST instance (first instance for Core1 and second instance for Core2) and *Blocking* in the other MST instance (second instance for Core1 and first instance for Core2). Due to its high bandwidth, the trunk interconnecting the two cores remains in the *Forwarding* state in both instances.

In case of an uplink failure (e.g. ros-swA to Core1) the other uplink (ros-swA to Core2) will change its state from *Blocking* to *Forwarding* in the first MST instance. The hosts connected to Core1, which were accessing the ROSs through Core1 – ros-swA, will now access them through the remaining uplink: Core1 – Core2 – ros-swA. Nevertheless, the total bandwidth to ros-swA will be halved, as its remaining uplink is shared by traffic from both MST instances.

B. *Practical Experience with MST*

Until now only partial failures have been experienced on the running system. On one occasion a fibre had to be cleaned in order to correct occasionally observed reception errors. This process was fairly transparent for the users: the link was shutdown, MST enabled the alternate path, the fibre was cleaned, the link was re-enabled and MST reverted back to the original path. The total communication downtime during this procedure was of the order of 200 ms. On two other occasions a faulty memory location on a line-card caused a 0.04% packet

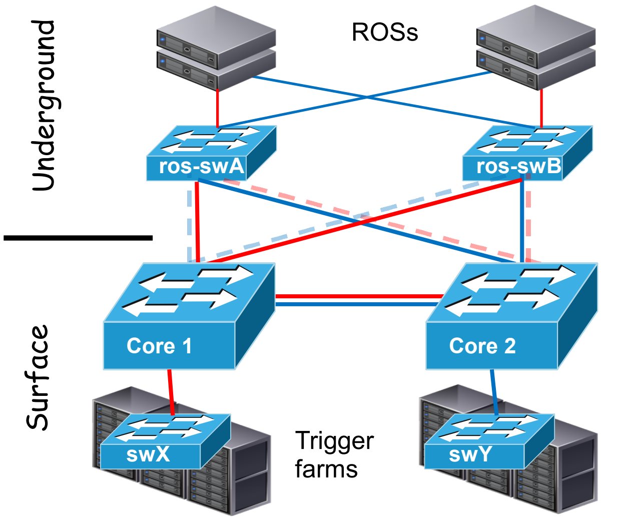


Fig. 5. FrontEnd network: redundancy and load balancing on the ROS switches uplinks using MST

loss for two 10GE ports, impacting the functioning of the TDAQ system. The rapid fix of the problem was to simply shutdown the problematic ports, and rely on the alternate paths (which for the time being provide sufficient bandwidth, as the system is not running at maximum speed).

VI. Trunking for High Throughput

Link aggregation [14] is heavily used in the BackEnd network both on networking equipment and on Linux computers. Since the throughput and latency requirements for the BackEnd network are more permissive (only full events are transferred at a lower rate in comparison to the FrontEnd network), it is implemented as an IP routed network with a single core router. The trigger farms are aggregated at the rack level using edge switches which connect to the core router through two trunked GE links. All SFIs and SFOs directly connect to the core router, either through a single GE link, or through two trunked GE links (most of them).

A. Load Balancing Configuration on Trunks

Using a trunk load balancing algorithm that preserves the frame order (i.e. fully compliant to [14]) gave sub-optimal results depending on the number of active trigger racks and their allocation to SFIs and SFOs. Since only full events (2 Mbyte each) are transferred through the network, frame reordering does not introduce a strong performance penalty. Practical experience has showed that the use of load balancing algorithms that fully exploit both uplinks provides a higher event throughput in the BackEnd network, despite the fact that it introduces out-of-order frames. Thus the router is configured to use a random load balancing on the trunked links, while the bonded interfaces on the SFIs and SFOs use a round-robin load balancing mode.

Apart from providing throughput higher than 1 Gbit/s at an affordable price, trunks offer builtin redundancy. To improve the fault tolerance of the system to failures of the router line-cards, the links of each trunk are distributed over different line-cards.

B. Practical Experience with Trunking

Apart from one partial failure of a group of ports on a line-card, the use of trunking caused sporadic link autonegotiation failures on the links belonging to bonded Linux interfaces.

1) Partial Line-card Failure: A group of 12 ports on the router, belonging to the same MAC chip, was sending corrupted frames. All the trunks that had a port in this group became unusable, as the corrupted frames sent by one of their links were dropped at the receiver side. The trunk only takes into account the link status, so the faulty link continued to be used.

2) Indeterministic Autonegotiation: At powerup, a high rate of link autonegotiation [15] failures has been observed on the trunked links connected to Linux computers: links coming up at 10Mbit/s or 100Mbit/s and even link speed mismatches between the PCs and the switches. Several remedies have been tried out (replacing the cables, restarting the interfaces) without success. Finally, the interfaces on the PCs were set to advertise only 1000 Mbit/s while autonegotiating, in order to either have a properly working interface or no interface at all. On top of that, a script that tries to bring up the interfaces several times has been deployed. The workaround was successful, but the cause of the problem remains uncomprehended.

Subsequently similar autonegotiation failures have been observed for normal interfaces (not bonded), but with a much smaller frequency. The same workaround has been successfully applied on all the data interfaces of the PCs.

VII. Conclusions

The ATLAS TDAQ system comprises a large number of computers interconnected by several networks, each one with its own specificities. We have deployed a variety of protocols meant to maximize the network resiliency: OSPF for inter-router connections, VRRP for providing redundant connectivity with no load balancing on routed networks, MST for providing both redundancy and load balancing on switched networks, and trunking whenever high throughput was required. In particular, a novel cost effective method for reliably connecting high-throughput servers to routed networks has been implemented.

The large majority of real-life failures experienced to date haven't been text-book failures, but partial ones. Since incomplete failures retain part of the functionality, they typically cannot be detected by the protocols foreseen to use alternate paths in such events. Thus, practical experience has showed that foreseeing redundancy can reduce downtime in case of failures, but is far from guaranteeing 100% uptime. Also, as the system increased in size new problems had to be addressed, e.g. proxy ARP rate limitation or sporadic link autonegotiation failures.

All the protocols described in the paper are open standards (not manufacturer specific), therefore all the presented methods can easily be implemented on any compliant network equipment.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the ATLAS TDAQ collaboration [16] for providing constant support and feedback that guided our work.

REFERENCES

[1] ATLAS Experiment. [Online]. Available: http://atlas.ch/

[2] S. Stancu, C. Meirosu, M. Ciobotaru, L. Leahu, and B. Martin, “Networks for ATLAS Trigger and Data Acquisition,” in *Proc. Computing in High Energy and Nuclear Physics (CHEP’06)*, Mumbai, India, Feb. 2006.

[3] S. M. Batraneanu, A. Al-Shabibi, D. Ciobotaru, Matei, M. Ivanovici, L. Leahu, B. Martin, and S. N. Stancu, “Operational Model of the ATLAS TDAQ Network,” *IEEE Trans. Nucl. Sci.*, vol. 55, no. 2, pp. 687–694, Apr. 2009. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04484220

[4] G. Malkin, “Dynamic Host Configuration Protocol,” RFC 2131, Mar. 1997.

[5] J. Moy, “OSPF Version 2,” RFC 2328, Apr. 1998.

[6] J. Postel, “Internet Protocol,” RFC 791, Sep. 1981.

[7] R. Hinden, “Virtual Router Redundancy Protocol (VRRP),” RFC 3768, Apr. 2004.

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

[9] (2007) Unicast Flooding in Switched Campus Networks. Document ID: 23563. [Online]. Available: http://www.cisco.com.do/application/pdf/paws/23563/143.pdf

[10] S. Carl-Mitchell and J. S. Quarterman, “Using ARP to Implement Transparent Subnet Gateways,” RFC 1027, Oct. 1987.

[11] Information Sciences Institute University of Southern California, “Transmission Control Protocol,” RFC 793, Sep. 1981.

[12] The Linux Foundation. (2009) bonding. [Online]. Available: http://www.linuxfoundation.org/collaborate/workgroups/networking/bonding

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

[14] *Link Aggregation*, IEEE Std. 802.1ax.

[15] *Carrier sense multiple access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications*, IEEE Std. 802.3 – 2008.

[16] The ATLAS collaboration. (2010, May) The ATLAS Trigger/DAQ Authorlist, version 4.0. ATL-DAQ-PUB-2010-002. [Online]. Available: http://cdsweb.cern.ch/record/1265604