The 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
15-17 September 2011, Prague, Czech Republic

Monitoring Large Scale Network Topologies

Ciprian Dobre [1], Ramiro Voicu [2], Iosif Legrand [3]
[1] University POLITEHNICA of Bucharest, Spl. Independentei 313, Romania, ciprian.dobre@cs.pub.ro
[2] California Institute of Technology, Pasadena, CA 91125, USA, Ramiro.Voicu@cern.ch
[3] CERN, European Organization for Nuclear Research, CH-1211, Geneve 23, Switzerland, Iosif.Legrand@cern.ch

Abstract— **Network monitoring is vital to ensure proper network operation over time, and is tightly integrated with all the data intensive processing tasks used by the LHC experiments. In order to build a coherent set of network management services it is very important to collect in near real-time information about the network topology, the main data flows, traffic volume and the quality of connectivity. A set of dedicated modules were developed in the MonALISA framework to periodically perform network measurements tests between all sites. We developed global services to present in near real-time the entire network topology used by a community. The time evolution of global network topology is shown in a dedicated GUI. Changes in the global topology at this level occur quite frequently and even small modifications in the connectivity map may significantly affect the network performance. The global topology graphs are correlated with active end-to-end network performance measurements, done using the Fast Data Transfer application, between all sites. Access to both real-time and historical data, as provided by MonALISA, is also important for developing services able to predict the usage pattern, to aid in efficiently allocating resources globally.**

Keywords—monitoring, large scale networks, topology.

I. Introduction

An important part of managing global-scale distributed systems is a monitoring system that is able to monitor and track in real time many site facilities, networks, and tasks in progress. The monitoring information gathered is essential for developing the required higher level services, the components that provide decision support and some degree of automated decisions and for maintaining and optimizing workflow in large scale distributed systems. Especially the network related aspects as topology monitoring can be very valuable in current LHC era when large amounts of data are expected to be transferred over the network.

In a distributed environment, a large number of monitoring events are generated by the system components during the execution or interaction with external objects (such as users or processes). Monitoring such events is necessary for observing the run-time behavior of the large scale distributed system and for providing status information required for debugging, tuning and managing processes. However, correlated events are generated concurrently and can be distributed in various locations, which complicates the management decisions process.

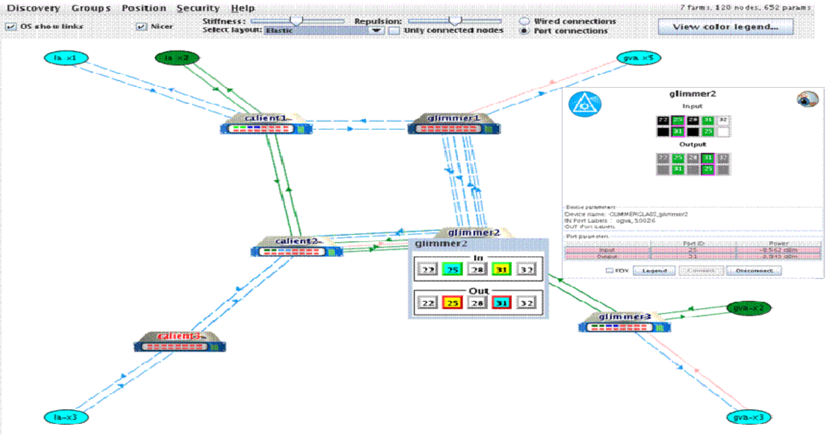
We present a set of services developed in the context of the MonALISA framework for monitoring and controlling large scale networks such as the ones supporting the LHC experiments. The framework and services are currently used in production in several of the largest Grid systems worldwide. This monitoring framework is one of the most complex platforms, which offers a large set of monitoring services and supports resource usage accounting, optimization, and automated actions. MonALISA service provides a distributed framework for monitoring and optimizing large scaled distributed systems. It is based on a Dynamic Distributed Service Architecture and is largely applicable to many fields of data-intensive applications. Specialized distributed agents are used for global optimization and guidance in large systems. A key point is its scalability coming for the distributed multi-threaded engines that host a variety of loosely coupled, self-describing dynamic services, capable of dynamic registration and discovery. It is implemented mainly in Java and is using JINI and WSDL technologies.

The rest of the paper is structured as follows. Section 2 presents the MonALISA monitoring framework. In Section 3 we present the monitoring services for large scale networks, together with solutions for the representation of network topologies at different OSI layers. In Section 4 we present a real-world use-case for monitoring network topology in case of one of the largest network supporting the LHC experiments at CERN. Finally, in Section 5 we give conclusions and present future work.

II. The MonALISA Monitoring Framework

MonALISA (Monitoring Agents in A Large Integrated Services Architecture) [1] is a globally scalable framework of services jointly developed by Caltech and UPB. MonALISA is currently used in several large scale HEP communities and grid systems including CMS [2], ALICE [3], ATLAS [4], the Open Science Grid (OSG) [5], and the Russian LCG sites. It actively monitors USLHCNet production network as well as the UltraLight R&D network [2]. MonALISA also is used to monitor and control all the EVO reflectors, and to help to optimize their interconnections [8].

As of this writing, more than 300 MonALISA services are running throughout the world. These services monitor more than 60,000 computing servers, and thousands of



☑ OS show links ☑ Nicer
Stiffness: Repulsion:
Select layout: Elastic □ Only connected nodes
○ Wired connections ● Port connections

glimmer2

Device parameters
Device name: GLIMMERGLASS_glimmer2
IN Port Labels : ogva_10026
OUT Port Labels

Port parameters

	Port ID	Power
Input	25	-8.562 dBm
Output	31	-9.049 dBm

□ FDX

Figure 1. Layer 1 topology: Monitoring and autonomous controlling optical switches.

concurrent jobs. More than 3.5 million parameters are currently monitored in near-real time with an aggregate update rate of approximately 50,000 parameters per second.

A large set of MonALISA monitoring modules has been developed to collect specific network information or to interface it with existing monitoring tools, including: SNMP modules for passive traffic measurements and link status; Active network measurements using simple ping-like measurements; Tracepath-like measurements to generate the global topology of a wide area network; Interfaces with the well-known monitoring tools MRTG, RRD [7]; Available Bandwidth measurements using tools like pathload; Active bandwidth measurements using Fast Data Transfer (FDT) [9]; Dedicated modules for TL1 [10] interfaces with CIENA's CD/CIs, optical switches (Glimmerglass and Calient) and GMPLS controllers (Calient) [11].

In the MonALISA framework the overall status of the dispersed systems being monitored is provided by either a GUI client or through specialized web portals. For the dedicated modules and agents used to monitor and control Optical Switches the GUI client of MonALISA provides a dedicated panel. This panel facilitates the interaction between users and the monitored Optical Switches. It offers to the end user a number of features such as complete perspective over the topology of the monitored optical networks or the possibility to monitor the state of the Optical Switches or the possibility to dynamically create new optical paths.

The tremendous interest in optical networks led the Internet Engineering Task Force (IETF) to investigate the use of Generalized MPLS (GMPLS) and related signaling protocols to set up and tear down lightpaths. GMPLS is an extension of MPLS that supports multiple types of switching, including switching based on wavelengths usually referred to as Multi-Protocol Lambda Switching (MPλS). In order to meet the expectations of future network technologies in the prototype system we made the first step towards integrating emerging light path technologies. We implemented the monitoring module and control agent that provide an interface between MonALISA and the Calient's GMPLS-based control plane.The described system, part of MonALISA, is currently used in production to monitor and control a CALIENT Optical Switch located at California Institute of Technology in USA and another GLIMMERGLASS Optical Switch located at the European Center for Nuclear Research, in Switzerland. The dedicated monitoring modules use the TL1 language to communicate with the switch and they are used to collect specific monitoring information. The state of each link and any change in the system is reported to MonALISA agents. The system is integrated in a reliable and secure way with the end user applications and provides simple shell-like commands to map global connections and to create an optical path / tree on demand for any data transfer application. A schematic view of how the entire system works is shown in the figure 2.

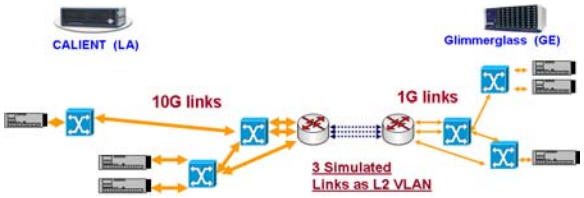


Figure 2. The system used to monitor and control Optical Switches and to create on demand optical path used in production.

The implemented prototype system is able to create dynamically an end to end light path in less than one second independent of the number of switches involved and their location. It monitors and supervises all the created connections and is able to automatically generate an alternative path in case of connectivity errors. The alternative path is set up rapidly enough to avoid a TCP timeout, and thus to allow the transfer to continue uninterrupted.

To satisfy the demands of data intensive grid applications it is necessary to move to far more synergetic relationships between applications and networks. Currently, even the most complex scientific applications are simply passive users of the existing network infrastructure. The main objective of the VINCI (Virtual Intelligent Networks for Computing Infrastructures) project is to enable users' applications, at the LHC and in other fields of data-intensive science, to effectively use and coordinate shared, hybrid network resources, to correlate them with available processing power in order to dynamically generate optimized workflows in complex distributed system (Figure 3).

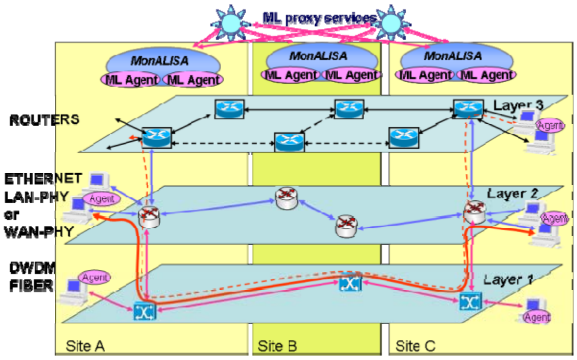


Figure 3. A schematic view of the functionality to provide dynamically an efficient end to end path to data intensive applications. The VINCI system is optimizing the path allocation using as much as possible Layer 1 or Layer 2 segments.

VINCI is a multi-agent system for secure light path provisioning based on dynamic discovery of the topology in distributed networks. For this project we are working to provide integrated network services capable to efficiently use and coordinate shared, hybrid networks and to improve the performance and throughput for data intensive grid applications. This includes services able to dynamically configure routers and to aggregate local traffic on dynamically created optical connections.

The system dynamically estimates and monitors the achievable performance along a set of candidate (shared or dedicated) network paths, and correlates these results with the CPU power and storage available at various sites, to generate optimized workflows for grid tasks. The VINCI system is implemented as a dynamic set of collaborating Agents in the MonALISA framework, exploiting MonALISA's ability to access and analyze in-depth monitoring information from a large number of network links and grid sites in real-time.

III. MONITORING AND REPRESENTATION OF NETWORK TOPOLOGIES AT DIFFERENT OSI LAYERS

We present monitoring and representational services developed considering various network topologies and the differences posed by network equipments operating at various OSI levels. In large-scale networks, such as USLHCNet and UltraLight, we found devices at ever OSI layer.

A. The Physical Network Layer Topology

A set of specialized TL1 modules are used to monitor optical switches (Layer 1 devices) from two major vendors: Glimmerglass and Calient. We were able to monitor the optical power on ports and the state of the cross-connects inside these switches.

Based on the monitoring information an agent is able to detect and to take informed decisions in case of eventual problems with the cross connections inside the switch or loss of light on the connections. The MonALISA framework allows one to securely configure many such devices from a single GUI, to see the state of each link in real time, and to have historical plots for the state and activity on each link. It also allows to easily manually create a path using the GUI. In Figure 1 we present the MonALISA GUI that is used to monitor the topology on the Layer 1 connections and the state and optical power of the links. The same GUI can be used to request an optical path between any two points in the topology. All the topology related information are kept distributed, every MonALISA service having its own view of the network. Every agent computes a shortest path tree based on Dijkstra's algorithm. The convergence in case of problem is very fast, as every agent has the view of the entire topology.

B. Layer 2 Network Topology / Circuits

The *USLHCNet transatlantic network* has evolved from DOE-funded support and management of international networking between the US and CERN. USLHCNet today consists of a backbone of eight 10 Gbps links interconnecting CERN, MANLAN in New York, and Starlight in Chicago. The core of the USLHCNet network is based on Ciena Core Director CD/CI multiplexers which provide stable fallback in case of link outages at Layer 1 (the optical layer), and full support for the GFP/VCAT/LCAS [13] protocol suite.

For the Core Director (CD/CI) we developed modules which monitor the routing protocol (OSRP) which allows us to reconstruct the topology inside the agents, the circuits (VCGs), the state of cross connects, the Ethernet (ETTP/EFLOW) traffic, the allocated time slots on the SONET interfaces and the alarms raised by the CD/CI (see Figure 4).

The operational status for the Force10 ports and all the Ciena CD/CI alarms are recorded by the MonALISA services. They are analyzed and corresponding email

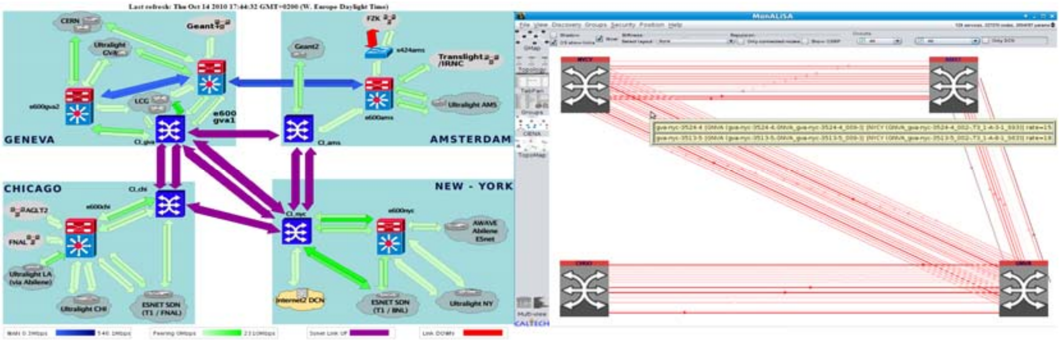


Figure 4. A network weathermap (left) and the layer 2 topology for the dynamic circuits (right).

notifications can be generated based on different error conditions. We also monitor the services used to collect monitoring information. A global repository for all these alarms is available on the MonALISA servers, which allows one to select and sort the alarms based on different conditions. The link status information is very sensitive information for the SLA (Service Level Agreement) with both the experiments and the link providers. Because of this very strict monitoring requirement the monitoring had to have almost 100% availability. We achieved this monitoring each link at both ends from two different points. The NOCs in Europe, Geneva and Amsterdam, are cross-monitored from both locations, and the same in US. In this way we monitor each link in four points and with special filters this information is directly aggregated in the repository. For redundancy and reliable monitoring we keep at least two instances of repositories running, one in Europe and one in US. For the past two years we manage to have 100% monitoring availability inside USLHCNet.

C. *Layer 3 Routed Network Topology*

For the routed networks, MonALISA is able to construct the overall topology of a complex wide area network, based on the delay on each network segment determined by tracepath-like measurements from each site to all other sites, as illustrated in Figure 5.

For any LHC experiment such a network topology includes several hundred of routers and tens of Autonomous Systems. Any changes in the global topology are recorded and this information can be easily correlated with traffic patterns. The time evolution of global network topology is shown a dedicated GUI. Changes in the global topology at this level occur quite frequently and even small modifications in the connectivity map may significantly affect the network performance.

IV. A REAL USE CASE FOR TOPOLOGY INFORMATION

The Alice Grid infrastructure uses MonALISA framework for both monitoring and controlling. All the resources used by AliEn [14] services: computing and storage resources, central services, networks, jobs are monitored by MonALISA services at every site.

A. *Bandwidth measurements between Alice sites*

The data transfer service is used by the ALICE experiment to perform bandwidth measurements between all sites, by instructing pairs of site MonALISA instances to perform FDT memory-to-memory data transfers with one or more TCP streams.

The results are used for detecting network or configuration problems, since with each test the relevant system configuration and the *tracepath* between the two hosts are recorded as well. The MonALISA services are also used to monitor the end system configuration and automatically notify the user when these systems are not properly configured to support effective data transfers in WAN. In Figure 6 we present the results recorded from one site to all the others.

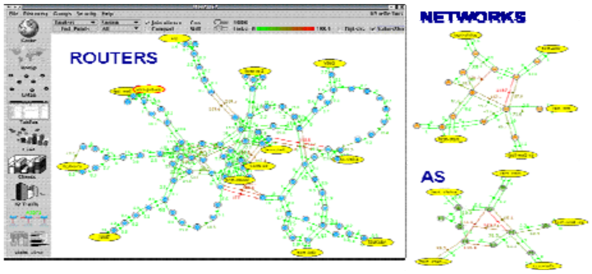
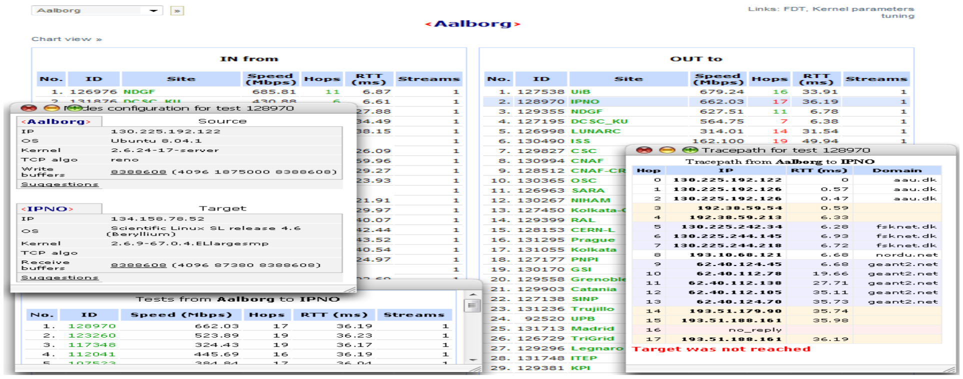


Figure 5. MonALISA real time view of the topology of WANs used by HEP. A view of all the routers, or just the network or “autonomous system” identifiers can be shown.

B. *Automatic storage discovery for Alice*

Using the monitoring information from trace-like measurements, derived information is computed in the repository, associating the Autonomous System (AS) number to each of the nodes in a network path. The repository also runs other monitoring modules that provide global values and one of them periodically queries AliEn for the list of defined storage elements and their size and usage according to the file catalog. Then periodic functional tests are performed from the central machine to check whether the basic file operations (add, get, remove) are successful. The entire software and network stacks are checked through these tests, thus the outcome should be identical for all clients trying to access the storages.

Aggregating the monitoring and test results, a client-to-storage distance metric is computed and used to sort the list of available storage elements to a particular client. Then the closest working storage elements is selected either to save the data or, in case of reading, sorting the available locations based on this metric, trying to read from the closest location. The algorithm associates to each storage element a list of IP addresses representing known machines from its vicinity.



IN from

No.	ID	Site	Speed (Mbps)	Hops	RTT (ms)	Streams
1.	126976	NDGF	685.81	11	6.87	1
2.	131876	DCSC_KU	430.88	6	6.61	1
					27.88	1
					34.49	1
					38.15	1
						1
					26.09	1
					59.96	1
					29.27	1
					23.93	1
						1
					21.91	1
					29.97	1
					40.07	1
					42.44	1
					43.52	1
					40.54	1
					24.97	1
						1

Nodes configuration for test 128970

<Aalborg>	Source
IP	130.225.192.122
OS	Ubuntu 8.04.1
Kernel	2.6.24-17-server
TCP algo	reno
Write buffers	8388608 (4096 1875000 8388608)
Suggestions	

<IPNO>	Target
IP	134.158.78.52
OS	Scientific Linux SL release 4.6 (Beryllium)
Kernel	2.6.9-67.0.4.ELlargesmp
TCP algo	
Receive buffers	8388608 (4096 87380 8388608)
Suggestions	

Tests from **Aalborg** to **IPNO**

No.	ID	Speed (Mbps)	Hops	RTT (ms)	Streams
1.	128970	662.03	17	36.19	1
2.	123260	523.89	19	36.23	1
3.	117348	324.43	19	36.17	1
4.	112041	445.69	16	36.19	1
5.	107523	384.84	17	36.04	1

OUT to

No.	ID	Site	Speed (Mbps)	Hops	RTT (ms)	Streams
1.	127538	UiB	679.24	16	33.91	1
2.	128970	IPNO	662.03	17	36.19	1
3.	129355	NDGF	627.51	11	6.78	1
4.	127195	DCSC_KU	564.75	7	6.38	1
5.	126998	LUNARC	314.01	14	31.54	1
6.	130490	ISS	162.100	19	49.94	1
7.	129827	CSC				
8.	130994	CNAF				
9.	128512	CNAF-CR				
10.	130365	OSC				
11.	126963	SARA				
12.	130267	NIHAM				
13.	127450	Kolkata-C				
14.	129399	RAL				
15.	128153	CERN-L				
16.	131295	Prague				
17.	131055	Kolkata				
18.	127177	PNPI				
19.	130170	GSI				
20.	129558	Grenoble				
21.	129903	Catania				
22.	127138	SINP				
23.	131236	Trujillo				
24.	92520	UPB				
25.	131713	Madrid				
26.	126729	TriGrid				
27.	129296	Legnaro				
28.	131748	ITEP				
29.	129381	KPI				

Tracepath for test 128970

Tracepath from **Aalborg** to **IPNO**

Hop	IP	RTT (ms)	Domain
0	130.225.192.122	0	aau.dk
1	130.225.192.126	0.57	aau.dk
2	130.225.192.126	0.47	aau.dk
3	192.38.59.54	0.59	
4	192.38.59.213	6.33	
5	130.225.242.34	6.28	fsknet.dk
6	130.225.244.145	6.93	fsknet.dk
7	130.225.244.218	6.72	fsknet.dk
8	193.10.68.121	6.68	nordu.net
9	62.40.124.45	6.68	geant2.net
10	62.40.112.78	19.66	geant2.net
11	62.40.112.138	27.71	geant2.net
12	62.40.112.105	35.11	geant2.net
13	62.40.124.70	35.73	geant2.net
14	193.51.179.90	35.74	
15	193.51.188.161	35.98	
16	no_reply		
17	193.51.188.161	36.19	

Target was not reached

Figure 6. Inter-site bandwidth test results. Tracepath is also recorded.

V. Conclusions

Network monitoring is vital to ensure proper network operations over time, and MonALISA was successfully used to provide its monitoring services to control a vast majority of the data intensive processing tasks used by the LHC experiments. In order to build a coherent set of network management services it is very important to collect in near real-time information about the network traffic volume and its quality, and analyze the major flows and the topology of connectivity.

In this paper we presented the capabilities of MonALISA framework towards monitoring and representing large scale networks at different OSI layers. We also present a very useful use case where informed automatic decisions based on monitoring information can improve reliability and increase overall performance of the system.

Acknowledgment

This work was supported by the Department of Energy and National Science Foundation within the DoE grant No DE-FG02-08ER41559, by the National Science Foundation within the UltraLight grant, contract No PHY-0427110. The work is co-funded by national project “DEPSYS - Models and Techniques for ensuring reliability, safety, availability and security of Large Scale Distributes Systems'”, Project CNCSIS-IDEI ID: 1710, and by national project “TRANSYS – Models and Techniques for Traffic Optimizing in Urban Environments", Contract No. 4/28.07.2010, Project CNCSIS-PN-II-RU-PD ID: 238.

References

[1] MonALISA official website, (2011), Last retrieved February 24, 2011, from: http://monalisa.caltech.edu/..

[2] CMS Experiment official website (2011), Last retrieved March 1, 2011, from: http://cms.cern.ch.

[3] ALICE Experiment official website (2011), Last retrieved February 26, 2011, from: http://aliweb.cern.ch

[4] Atlas Experiment official website (2011), Last retrieved February 12, 2011, from: http://atlas.web.cern.ch

[5] OSG official website (2011), Last retrieved February 25, 2011, from:http://www.opensciencegrid.org

[6] Costan, A., Dobre, C., Cristea, V., Voicu, R., (2008). *A Monitoring Architecture for High-Speed Networks in Large Scale Distributed Collaborations*, In Proc. of the 7th International Symposium on Parallel and Distributed Computing, ISPDC'08, pp. 409 – 416, Krakow, Poland.

[7] RRD official website (2011), Last retrieved February 26, 2011, from: http://www.mrtg.org/rrdtool

[8] Legrand, I. C., Newman, H. B., Voicu, R., Cirstoiu, C., Grigoras, C., Dobre, C., Muraru, A., Costan, A., Dediu, M., Stratan, C., (2009), *MonALISA: An agent based, dynamic service system to monitor, control and optimize distributed systems*, In Computer Physics Comm., 180(*12*), December 2009, pp. 2472-2498.

[9] FDT official website (2011), Last retrieved February 26, 2011, from: http://fdt.cern.ch.

[10] TL1 – Transaction Language 1 Generic Requirements Document GR-831-CORE: http://telecom-info.telcordia.com/site-cgi/ido/docs.cgi?ID=SEARCH&DOCUMENT=GR-831.

[11] Calient Technologies official website (2011), Last retrieved February 26, 2011, from: http://www.calient.net

[12] GMPLS – General Multi-Protocol Label Switching Architecture RFC3945.

[13] ITU-T Rec. G.7042, “Link Capacity Adjustment Scheme (LCAS) for Virtual Concatenated Signals,” Feb. 2004.

[14] Bagnasco S, Betev L, Buncic P, Carminati F, Cirstoiu C, Grigoras C, Hayrapetyan A, Harutyunyan A, Peters A J and Saiz P 2007 AliEn : ALICE environment on the GRID, J. Phys.: Conf. Ser. 119.