Monitoring and Visualization of the Large Hadron Collider Optical Private Network

Patricia Marcu, David Schmitz
German Research Network (DFN)
c/o Leibniz Supercomputing Centre
Boltzmannstr. 1, 85748 Garching, Germany
{marcu, schmitz}@dfn.de

Andreas Hanemann
German Research Network (DFN)
Alexanderplatz 1, 10178 Berlin
Germany
hanemann@dfn.de

Szymon Trocha
Poznan Supercomputing and
Networking Center (PSNC)
Noskowskiego 12/14, Poznan, Poland
szymon.trocha@man.poznan.pl

***Abstract*—The Large Hadron Collider (LHC) experiments at CERN result in vast amounts of data which are interesting for researchers around the world. For transporting the data to 11 data distribution centers, an optical private network (OPN) has been constructed as the result of a collaboration of several academic networks. The multi-domain nature of this collaboration poses new challenges, in particular to network monitoring. This has been addressed by adopting the multi-domain tool set perfSONAR. In the implemented solution, the network users can access network monitoring information from various measurement points through the LHCOPN Weathermap visualization tool. This paper details the tool itself and its operation within the LHCOPN using perfSONAR tools. It serves as a blueprint for the support of future data-intensive large-scale projects.**

I. Introduction

The significant increase in the availability of high-speed research networks has led to the deployment of large-scale distributed computing environments that are able to serve a large number of geographically separated users by exchanging large amounts of data. Direct communication between sites and computing facilities is now necessary in many working environments. This results in greatly expanded requirements for high-speed, dedicated networks that cross multiple domains. The European Research Network GÉANT [1] and National Research and Educational Networks (NRENs) in Europe are high-capacity telecommunication networks which are based on optical technologies and components that provide wavelength-based services to the research community.

A representative project is the provisioning of the networking infrastructure for the Large Hadron Collider (LHC) at CERN in Switzerland. Its research experiments are expected to produce 15 petabytes of data per year. Therefore, a multi-domain LHC Optical Private Network (LHCOPN) was established [2], dedicated to support data exchange. The LHCOPN consists of Tier-0 and Tier-1 centers connected by End-to-End (E2E) links. These E2E links connect organizations (Tier-1 centers) that are located in different countries and cross the shared network infrastructure of different providers (GÉANT, NRENs) towards the Tier-0 centre at CERN.

One of the most important and difficult issues related to this dedicated network is network management. The monitoring and troubleshooting optical networks and individual E2E links is challenging. Researchers all over the world are increasingly using dedicated optical paths to create high-speed network connections, and different groups of users may want to use monitoring applications for specific research purposes. They need access to network measurement data from multiple involved network domains, visualize network characteristics and troubleshoot related issues [3].

A quick overview and visualization of the network status is necessary to establish demarcation points that help distinguish network issues within LHCOPN from those in the sites. The deployment of monitoring tools and the use of common services should provide a unified network information view across all domains [4]. Typical off-the-shelf solutions do not provide such functionality.

This paper is organized as follows. Section II includes a description of requirements for multi-domain network monitoring. In section III we describe the state-of-the-art of existing solutions for large scale network monitoring. Section IV is introducing the measurement methods in LHCOPN. In section V we present our system design, and section VI provides a detailed description of our implementation of the LHCOPN Weathermap software. Section VII provides conclusions and future work.

II. Requirements

The monitoring of the LHCOPN, i.e. Tier-0 and Tier-1 centres, and the links between them poses several new challenges:

a) Multi-domain monitoring: The LHCOPN itself is based on resources that are supplied by several academic networks such as GÉANT, European NRENs, Internet2, ESnet and Canarie. Therefore, a solution has to be found to collect monitoring data from all these self-administered networks to form a joint view of the resulting network.

b) Monitoring of different layers: While academic networks have been used to monitor the network layer, the LHCOPN requires E2E links on the data link layer to be monitored. These are based on heterogeneous technologies, as the different participating networks use different technologies. E2E links are formed by combining technologies such as SDH and SONET, native Ethernet or Ethernet over MPLS, where

each domain is dependent on the data that it can retrieve from the network management system of its vendor.

c) Joint view of all metrics: In the visualization a view has to be formed by combining E2E link and IP-related monitoring data and by linking these data in a suitable manner. In doing so, it must be considered that there are also several data sources on the IP level, in particular the retrieval of SNMP data from routers and the results of active measurements.

III. State of the art

The issue of multi-domain monitoring is not only a challenge in the context of the LHCOPN, but also in the general operation of networks. In 2004 a collaboration of the GN2/GN3 project (between Internet2, ESnet, RNP and others) was started to jointly develop a communication protocol and tool set under the name perfSONAR [5], [6]. This development has made necessary by the limitations of existing tool sets which were tied to single domain monitoring and limited to metrics that can be monitored. Such limitations apply e.g. to the MonALISA [7] tool set. Apart from being used in the LHCOPN, the perfSONAR tools are also used within the networks that participate in the collaboration and in other locations, as the software is open source.

The introduction of hybrid networks and the possibility to deliver E2E links that involve multiple domains has led to the need to monitor these links. Therefore, a special tool called E2EMon (E2E Monitoring Tool) [8] has been developed over the recent years. Every domain which provides a segment of such an E2E link needs to have an E2EMon Measurement Point (MP) in place which retrieves data from the local network management system to provide status information for the link segment. Due to the heterogeneity of technologies on the layer below IP, it is only useful to provide an operational and administrative up/down status for each link segment. The status data for the segments are used for E2EMon to calculate status data for the whole E2E link.

For IP-level monitoring, the HADES, BWCTL and RRD MA tools are of interest, as they provide relevant metrics and are integrated into the perfSONAR framework. They can therefore easily become part of an overall management solution.

- HADES (Hades Active Delay Evaluation System) [9] uses dedicated hardware boxes to perform active tests in the network to measure delay, jitter, packet loss and traceroute (with respect to IPPM recommendations [10]). For precise timing, GPS antennas are installed in addition to the hardware boxes. Networking Time Protocol (NTP) can also be used but with less precision.
- BWCTL (Bandwidth Test Controller) [11] carries out throughput tests with TCP or UDP.
- RRD MA/SQL MA (Round Robin Database Measurement Archive/Structured Query Language Measurement Archive) are tools that provide archived measurement data. Typically, they store data retrieved via SNMP from routers to provide information about link utilization, interface errors and output drops.

There are already several tools which can visualize perfSONAR measurement data [12]. One of them is perfsonarUI which can be used for troubleshooting by allowing a direct interaction with perfSONAR measurements.

All principles of the perfSONAR protocol will be taken into account in the customization for the LHCOPN, especially its *multi-domain-monitoring* feature. Furthermore, this will be done with respect to the *different layers monitoring* also developed within the GN2/GN3 projects. The missing requirement for this customization is the *joint view on all metrics*. Also, a global overview of the LHCOPN was needed, in which different layer views coexist. This customization was achieved by a dedicated version of the Customer Network Management (CNM) tool [13] (in a browser-based version).

IV. Measurements in LHCOPN

To understand the metrics displayed in the LHCOPN Weathermap, it is necessary to know how the measurements are carried out. The deployment of the perfSONAR measurement tools at each Tier-1-centre is therefore shown in Figure 1.

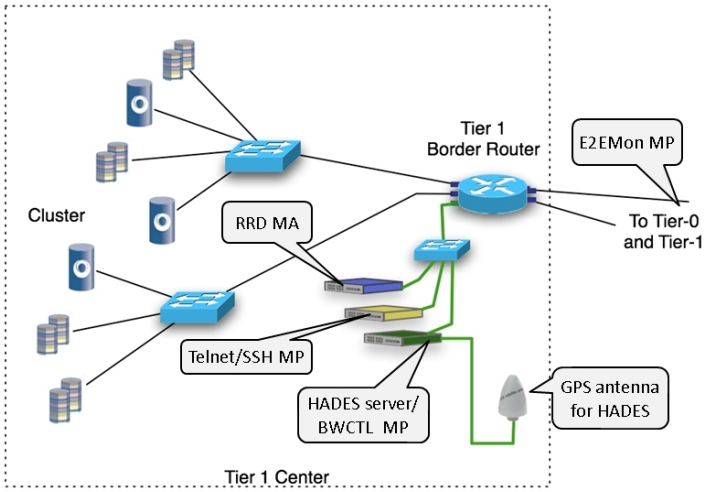

Fig. 1. Tier-1 site configuration

Three measurement servers are located at each centre. One of the servers is used to host an RRD MA to collect utilization, interface errors and output drop data related to the router located at the Tier-1 centre.

The second server is the HADES box which conducts one-way delay [14], IP delay variation (jitter) [15], packet loss [16], and traceroute tests with any other HADES box in the LHCOPN every minute. It is connected to a GPS antenna for precise timing.

In addition, the HADES box hosts a BWCTL MP which is used for throughputs with the Tier-0 centre every 8 hours. The BWCTL MP is hosted on a different interface to avoid interference with HADES measurements. The third server is used for the Telnet/SSH MP, a tool that allows configuration data to be retrieved from the routers. It is only mentioned here for completeness, but does not carry out any regular measurements.

In addition to these measurements on the IP level, data is also collected for the E2E links. The links start at the Tier-0 centre or at one of the Tier-1 centers (backup links), end at one of the Tier-1 centers and typically cross several administrative domains (e.g. GÉANT, European NRENS, Internet2 or ESnet). The status of each link is then calculated based on the NMS data from each domain involved.

V. Tool Design

The measurements that are carried out have to be displayed in a suitable manner which means in this case that a trade-off between correct display and usability has to be made. For example, HADES measurements are not directly located on the routers, so that delay data is not exactly measured at the location of utilization measurements. Events on the short link between router and HADES box can lead to wrong interpretations.

Even more difficult considerations have to be made for E2E link status data and its relation to IP metrics. By default the IP data in the network uses the direct way via an E2E link. However, if the E2E link fails (including the optical protection), then the IP protection performs a rerouting. Although IP packets are still transferred, they take another route on the optical level. Therefore, it is necessary to clearly distinguish between optical and IP level.

For this reason, a data model is introduced in the following that covers all topology information per network layer and all necessary topology mapping information for the LHCOPN Weathermap.

With respect to the requirements stated in Section II, the operators of the LHCOPN should have a global view on their network. They should also have layer-related, location-related and metric-related views on the LHCOPN. An E2E view is needed to check the availability of the E2E links involved. Therefore, different layers (topologies) have been defined: E2E link, HADES, BWCTL and Router Topology.

A. E2E Link topology

To satisfy the LHCOPN requirement for multi-domain monitoring accessed through a global view, a layer based on the E2E link has been specified to form the main view. This layer gives an overview of the whole LHCOPN respectively on dedicated E2E links involved in the LHCOPN. For each link that is displayed in the topology two kinds of abstractions can be involved. A link represented here can be an E2E Link or it can be an E2E link plus another E2E link which serves as optical (1+1) protection. The other kind of abstraction that is involved, is that for each E2E link the status is derived from data retrieved from multiple NMS. In the following a detailed description of the E2E Link topology, whose representation in the LHCOPN Weathermap is shown in Figure 2, is given.

The topology consists of *abstract nodes* and *abstract links*. Abstract nodes represent the Tier-0 and Tier-1 LHCOPN locations and are named accordingly. They abstract the exact location where measurements are conducted in order to allow easy linking of this topology to the other topologies. The abstract links are non-directed (i.e. bidirectional) links between the abstract nodes.

The metric used for this abstract layer is the *aggregated status* for each abstract link. It is computed every 5 minutes from the E2EMon status of all associated E2E links. For a single E2E link the status is retrieved in the E2EMon system by polling all E2EMon MPs every 5 minutes.

The four status values of the abstract links are computed as follows:

- DOWN: if one associated E2E link is down
- WARNING: if no associated E2E link is down and at least one has the status warning
- UNKNOWN: if no associated E2E link is down or indicates a warning and if the status of at least one of the associated E2E links is unknown (could not be measured or the measurement could not be obtained)
- UP: if the status of each associated E2E link is up

The list of E2E links associated with an abstract link may include links which are currently missing (unknown) from the accessible E2E Link topology.

Currently, the rules do not take into account the case where one associated E2E link is missing in the known E2E link topology, and its backup link exists already. In this case, the aggregated status for the abstract link is computed only from the non-missing E2E link.

B. HADES topology

The HADES [9] boxes have been deployed at the Tier-0 and Tier-1 LHCOPN locations to provide QoS measurements. The HADES topology is made up of the abstract nodes, so it can easily link them to E2E measurements and directed (uni-directional) *HADES links* between them. A HADES link is determined by its source and targeted abstract nodes. As HADES links correspond to pairs of abstract nodes, they are identified by ordered pairs of abstract locations.

HADES measurements are run as a full mesh between all nodes. To prevent users from being overloaded with too much data, the visualization in the Weathermap is limited to measurements that relate to paths where E2E links exist.

The metrics on the HADES layer are IP performance metrics (IPPM) computed for each HADES link (one way delay [14], IP delay variation (jitter) [15] and packet loss[16]) as well as the hop list/count metric. As the links are directed between two different HADES end points A and B, the metrics exists for $A \rightarrow B$ and $B \rightarrow A$. All these metrics have a time resolution of 5 minutes. The HADES metrics are stored in a HADES Measurement Archive (based on the SQL MA) which are used to store and publish historical monitoring data produced by the HADES Measurement Points.

C. BWCTL topology

Similar to HADES, BWCTL verifies available bandwidth from each endpoint to other points to identify throughput problems. In the LHCOPN BWCTL, nodes are included within the HADES boxes by using a second interface card.

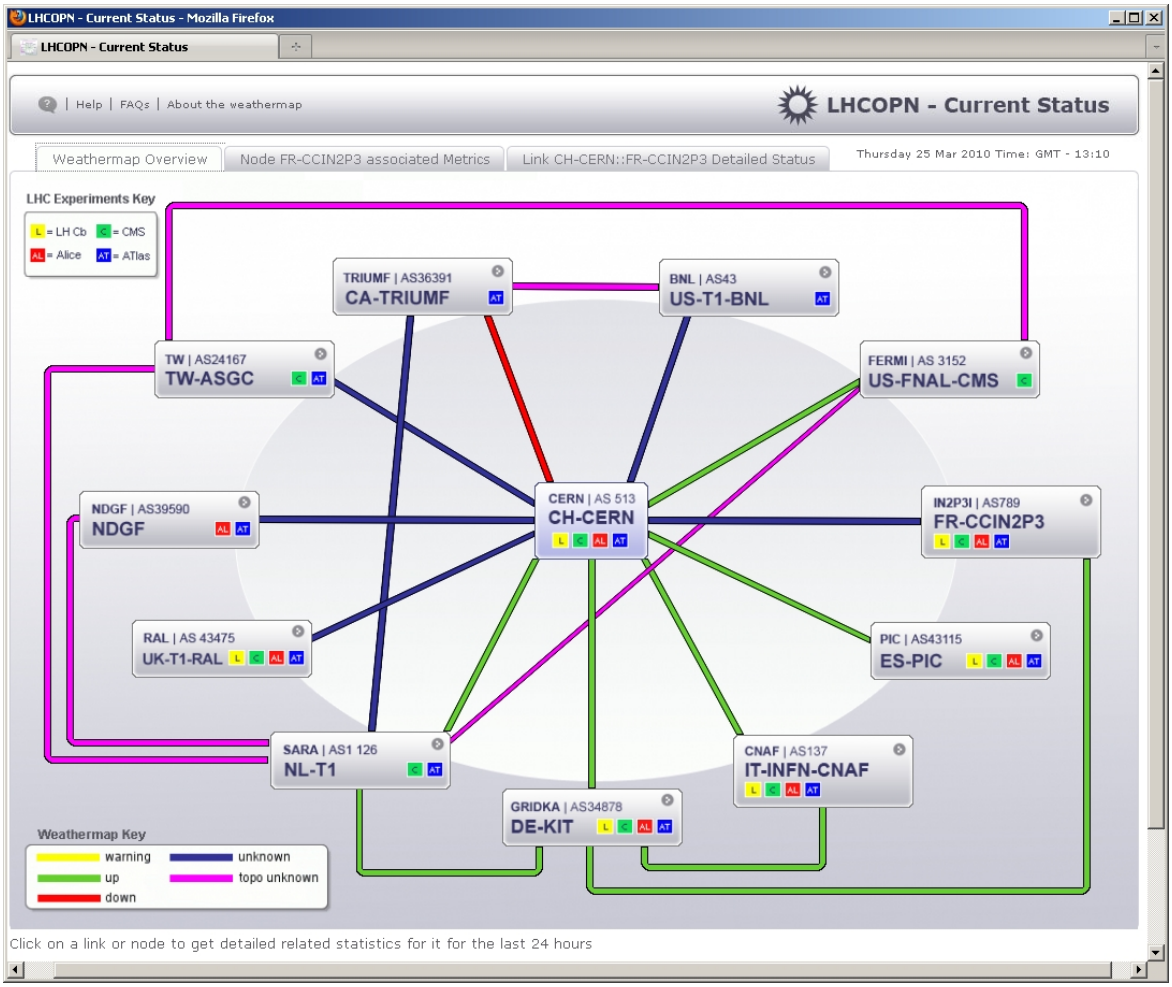

Fig. 2. A view on the E2E Link Topology Tab in the LHCOPN Weathermap tool

Each BWCTL end point address is associated with an abstract node. This is a 1:1 mapping but the IDs of the BWCTL end point and of the abstract node are not the same (BWCTL IP addresses vs. location names).

On this layer, the needed metrics are minimum, medium and maximum BWCTL throughput (stored in the SQL MAs). As the BWCTL links are directed between two different BWCTL end points A and B, these BWCTL metrics exist for both directions $A \rightarrow B$ and $B \rightarrow A$.

D. Router topology

Information about the IP links between two different IP interfaces is needed to determine the status of the links between two IP interfaces within the LHCOPN (these are VLANS in their terminology).

The IP topology consists of the abstract nodes which relate here to IP interfaces and *IP links* (pairs of IP interfaces). One IP link corresponds to a VLAN in the LHCOPN terminology. The current assumption is that one abstract link is associated with one IP interface pair only. This means that, if one abstract link has two or more E2E links, they both contribute (in an aggregated manner) to the same IP link (back-up link or bundle of links). Also, one E2E link can contribute to a single VLAN only.

The metrics used on the IP topology are utilization, input errors and output drops for each end point of an IP link. These metrics have a time resolution of 5 minutes.

VI. Implementation highlights

A. Data retrieval, filtering and integration

The data retrieval is concerned with the fetching and updating of topology information, topology mapping information, metric mapping information (see section V), as well as the actual metric data fetching. The measurements in LHCOPN as well as metric data fetching were outlined in Section IV.

The *abstract topology* and its associated E2E links have to be imported from a structured, static configuration file provided by LHCOPN users or, in future, from an online configuration file.

The *E2E Link topology* is fetched from the E2EMon export interface together with their individual E2E link states and have to match the ones (associated with abstract links) specified above.

Fig. 3. A possible representation of the link status tab

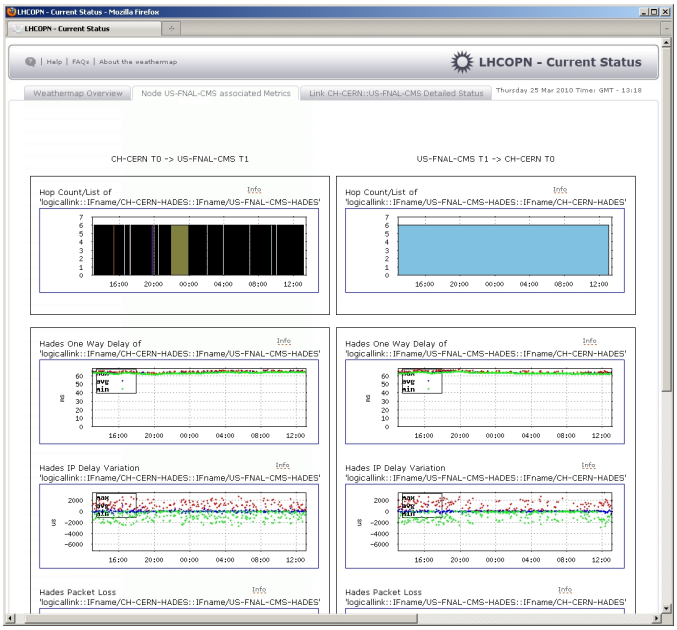

Fig. 4. A possible representation of the node status tab

The *HADES topology* is imported from metadata of the LHCOPN HADES MA. The topology mapping (abstract node 1:1 HADES node) is trivial, and has to be altered to show links that are interesting to the Weathermap (links that correspond to an abstract link).

The *BWCTL topology* is imported from metadata in the LHCOPN BWCTL SQL MA. The mapping between BWCTL nodes (BWCTL IP addresses) and abstract nodes is statically configured.

Potential LHCOPN IP interface address pairs are imported from the metadata of various LHCOPN RRD MAs and then need to be altered according to the abstract link to IP interface address pair mappings specified above.

B. Visualization

To meet the requirements of LHCOPN users, a visualization consisting of three tabs has been designed: *Overview Map Tab*, *Metric Tab* and *E2E Link Tab*.

1) Overview Map Tab: In this tab (see figure 2) a map consisting of abstract nodes and abstract links (described in section V-A) is shown. This overview map indicates the current status using four colors: RED, YELLOW, GREEN and BLUE for the current abstract link status DOWN, WARNING, UP and UNKNOWN (defined in Section V-A).

In addition to that, the fifth status (MAGENTA) does not represent a value of the metric's aggregated status, but instead indicates that there is a serious mismatch in the topology mapping concerning the abstract link; namely that all associated E2E links (from the point of view of the abstract topology) are unknown to the E2E topology (from the point of view of the E2E topology). This status is called topology unknown and indicates that no aggregated status for the abstract link could be computed.

The content of the other two tabs (Metric tab and E2E Link tab) is shown when clicking either on an abstract link or an abstract node in the map. So by clicking and choosing the selected abstract element, data corresponding to this abstract link or node is loaded in the metric tab and E2E link tab.

2) The Metric Tab: The metric tab shows statistical graphs of metrics associated to particular abstract links.

If an abstract link in the overview map is selected, data for this specific link is shown. If a Tier-1 abstract node is selected, the abstract link from the Tier-0 abstract node (CERN) to this selected Tier-1 abstract node location is selected. If the Tier-0 abstract node (CERN) is selected, data for all abstract links from CERN to any Tier-1 is displayed.

In the metric tab, 24-hour metric graphs of various metrics of the network layers (see section V) are presented for the chosen abstract link. The list of visualized metrics is different depending whether the selected abstract element is a node or a link.

Metrics for an abstract link: Selecting an *abstract link* in the overview map displays the following metrics for this link (see figure 3) in the Metric tab:

- The graph of the E2E aggregated status associated with the abstract link itself. This is based on the data model in section V-A and is visualized in the previously mentioned status colors.
- The RRD MA metrics graphs (see Section IV) for the single IP link associated with the abstract link. These are visualized for both IP interfaces at both end points of the IP link.

All the metrics are measured and updated every 5 minutes.

Metrics for an abstract node: Selecting an *abstract node* in the overview map (all Tier-0 to Tier-1 abstract links related to the selected abstract node) displays statistic graphs of the

following metrics for the chosen abstract links are shown (see figure 4):

- The *Hop count metric graph* is divided into differently colored areas, indicating different routes.
- The *HADES metrics* are visualized as scatter plot graphs (values are dots), each with a 5 minute time resolution. For one way delay and jitter the minimum, medium and maximum is needed.
- *BWCTL metric graphs* are visualizing the minimum, medium and maximum BWCTL throughput.

3) E2E Link Tab: The metric tab specified in the previous section, shows metrics on different (network) layers in a more end-to-end like fashion between the Tier-0/Tier-1 locations. In addition to this, the E2E link tab presents a section status view for the focused abstract link. This is done by wrapping the HTML page for the E2EMon section status for each E2E link associated to the focused abstract link in the map overview tab.

If the selected abstract element in the map overview tab is an abstract link, the E2EMon segment status is shown for all E2E links associated to this.

If the selected abstract element in the map overview tab is an abstract Tier-1 node, the E2EMon segment status is shown for all E2E links associated with this focused abstract link.

C. Client and Access Point

The client is implemented as a dynamic HTML page with some java script code used for the tabbing interface.

To speed up the access, some graphical parts of the content are created and cached in advance:

- The current 24-hour statistic plot of any network element necessary as specified in section V, are usually updated on a 5 minutes basis.
- The overview map which includes the link status color is updated every 5 minutes.
- Internal to the HTML dynamic generation scripts, additional data base content caching is performed to speed up access further.

VII. Conclusion and further work

In this paper the monitoring of the LHCOPN has been explained with a focus on the LHCOPN Weathermap. While the support structure is ready to fulfill its needs, it will only prove its usefulness in day-to-day operations once the LHC experiments are running, and large amounts of data are actually transferred via the network.

Besides continued improvements to the already existing tools, an alarm tool is currently under development. It is designed to be quite flexible in terms of alarm generation, to be suitable for different user needs.

The perfSONAR services used for the LHCOPN and the Weathermap are likely to be relevant to future large scale projects in Europe. A collection of such projects can be found in the roadmap of the European Strategy Forum on Research Infrastructures (ESFRI) [17]. For the Weathermap, this means that different ways of customization to meet the needs of other projects are going to be investigated. For projects that want to use dynamic circuits, the perfSONAR group is already investigating suitable monitoring methods.

Acknowledgments

This work is part of the Multi-Domain Networking Service Activity within the GN3 project which is co-funded by the European Commission.

References

[1] GÉANT, "Géant Homepage," http://www.geant.net/, 2010.
[2] E.-J. Bos, E. Martelli, and P. Moroni, "LHC Tier-0 to Tier-1 High-Level Network Architecture," CERN, Tech. Rep., 2005.
[3] J. W. Boote, E. L. Boyd, J. Durand, A. Hanemann, L. Kudarimoti, R. Lapacz, N. Simar, and S. Trocha, "Towards Multi–Domain Monitoring for the European Research Networks," in *Selected Papers from the TERENA Networking Conference*. TERENA, Oct. 2005. [Online]. Available: http://www.terena.org/publications/tnc2005-proceedings/
[4] M. K. Hamm and M. Yampolskiy, "Management of Multidomain End–to–End Links. A Federated Approach for the Pan–European Research Network 2," in *Moving from Bits to Business Value: Proceedings of the 2007 Integrated Management Symposium*. Mnchen, Germany: IFIP/IEEE, May 2007, pp. 189–198.
[5] A. Hanemann, J. Boote, E. Boyd, J. Durand, L. Kudarimoti, R. Lapacz, N. Simar, M. Swany, S. Trocha, and J. Zurawski, "Perfsonar: A service-oriented architecture for multi-domain network monitoring," in *Proceedings of the 3rd International Conference on Service-Oriented Computing (ICSOC 2005)*. Amsterdam, The Netherlands: ACM, December 2005, pp. 241–254.
[6] "perfSONAR project," http://www.perfSONAR.net.
[7] "MONitoring Agents using a Large Integrated Services Architecture (MonALISA)," http://monalisa.caltech.edu/, California Institute of Technology.
[8] M. Hamm and M. Yampolskiy, "E2E Link Monitoring: System Design and Documentation," GN2 Project, Tech. Rep., 2008. [Online]. Available: https://wiki.man.poznan.pl/perfsonar-mdm/images/perfsonar-mdm/1/12/GN2-JRA4-06-010v240.pdf
[9] "HADES," http://www.win-labor.dfn.de/cgi-bin/hades/selectnew.pl?-config=, WiN - Labor, RRZE, Erlangen.
[10] V. Paxson, G. Almes, J. Mahdavi, and M. Mathis, "Framework for IP Performance Metrics," USA, Tech. Rep., 1998.
[11] "Bandwidth Test Controller (BWCTL)." [Online]. Available: http://www.internet2.edu/performance/bwctl/bwctl.man.html
[12] A. Hanemann, V. Jeliazkov, O. Kvittem, L. Marta, J. Metzger, and I. Velimirovic, "Complementary Visualization of perfSONAR Network Performance Measurements," in *Proceedings of the International Conference on Internet Surveillance and Protection (ICISP)*, vol. 2006. Cap Esterel, France: IARIA/IEEE, Aug. 2006.
[13] "CNM for GN3 project homepage," http://sonar1.munich.cnm.dfn.de/.
[14] G. Almes, S. Kalidindi, and M. Zekauskas, "A One-way Delay Metric for IPPM," USA, Tech. Rep., 1999.
[15] C. Demichelis and P. Chimento, "IP Packet Delay Variation Metric for IP Performance Metrics (IPPM)," USA, Tech. Rep., 2002.
[16] G. Almes, S. Kalidindi, and M. Zekauskas, "A One-way Packet Loss Metric for IPPM," USA, Tech. Rep., 1999.