LHCONE prototype planning

Version 1

Authors:
DANTE: Roberto Sabatino, Richard Hughes-Jones, Xavier Martins-Rivas, Matthew Scott
DFN: Klaus Ullmann, Karin Schauerhammer
GARR: Enzo Valente, Marco Marletta, Massimo Carboni
RENATER: Dany Vandromme

Background

This document summarises the discussion held between DANTE, DFN, RENATER and GARR to develop a prototype network infrastructure that meets the networking needs of the LHC ATLAS and CMS experiments. It is developed in accordance with the LHCONE architecture document.

Introduction

With the successful operation of the LHC accelerator and the start of the data analysis has come a re-evaluation of the computing and data models of the experiments. The goal of LHCONE is to ensure better access to the most important datasets by the world wide HEP community and hence improve the data analysis. To effect this, several steps have been taken including:

- flattening the strict Tier1-Tier2-Tier3 hierarchy MONARC model so that any site may connect with any other;
- revisiting the implementation of strategic data placement;
- reinforcing the use of dataset caching – rather like the caching of web pages.

Although the operation of the LHCOPN remains the same, moving data between Tier0-Tier1 and Tier1-Tier1, the steps mentioned above have altered the traffic patterns to the extent that substantial data transfers between major sites are regularly being observed on the national and GÉANT IP backbones, often lasting for several days. This traffic is not just a few high bandwidth flows but often made up of many small flows – the users have learnt how to use gridftp very effectively.

These changes have been discussed at the LHCOPN meetings and at a dedicated meeting in January 2011 between the experimenters and the NRENs. Several suggestions were presented and discussed on how to find a way forward. At the LHCOPN meeting in Lyon during February a draft architecture
document for LHCONE was constructed taking into account the user requirements and design input from the academic network community. The main principle was to separate the traffic, thus avoiding degraded performance both to the LHC community and to the other users of the academic IP networks. Doing nothing is not an option as it would almost certainly result in a serious impact on all network users. Also at the Lyon meeting it was suggested to build a prototype involving a limited number of sites in parallel with deriving a suitable implementation document.

This current document builds on discussions between DANTE, and some of the NRENs that serve the Tier2 community in Europe who were present at the meeting in Lyon. It presents the options available to provide a suitable prototype of an integrated, unified solution that satisfies the users requirements while at the same time provides the flexibility to allow the European NRENs to meet their national constraints and to allow high bandwidth connectivity to the US and Asia. It should be noted that there are several open questions, but it is clear that this infrastructure must be operationally viable and financially sustainable in the NREN-GÉANT construct. Also that there needs to be a smooth migration path from a prototype to a permanent operational environment.

This paper references v2.1 of the LHCONE architecture paper.

**LHCONE services**

LHCONE foresees 4 services as follows:

1. Shared layer 2 domains (private VLAN broadcast domain)
2. Point to point layer-2 connections
3. Dynamic circuits with BW guarantees
4. perfSONAR based measurements and archive

**Scope of prototype**

In the LHCOPN meeting held in Lyon on 10-11 February 2011 it was agreed that a prototype for LHCONE should be established in a short time frame. It should be based on supporting the top-10 T2 sites and all T1s involved in ATLAS and CMS. These top-10 sites are in the process of being finalised and the information available at present is as follows:

<table>
<thead>
<tr>
<th>Atlas top10</th>
<th>CMS top10</th>
<th>T1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DESY-HH, Hamburg, Germany</td>
<td>GridKa, Karlsruhe, Germany</td>
<td></td>
</tr>
<tr>
<td>GRIF, Orsay, France</td>
<td>IN2P3, Lyon, France</td>
<td></td>
</tr>
<tr>
<td>Glasgow, UK</td>
<td>RAL, Oxford, UK</td>
<td></td>
</tr>
<tr>
<td>Roma1, Rome, Italy</td>
<td>CNAF, Bologna, Italy</td>
<td></td>
</tr>
<tr>
<td>IFAE, Barcelona, Spain</td>
<td>PIC, Barcelona, Spain</td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td></td>
</tr>
<tr>
<td>Tokyo University, Tokyo, Japan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGLT2, Ann Arbor, Michigan, USA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicago, Illinois, USA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t.b.d., Canada</td>
<td>CERN, Geneva, Switzerland</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NDGF, Nordunet, Copenhagen</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SARA, Amsterdam, Netherlands</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ASCG, Taiwan</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FNAL</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BNL, Brookhaven</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NN, Canada</td>
<td></td>
</tr>
</tbody>
</table>

The precise objectives of the prototype have not yet been set out by the experiments. Our planning assumption is based on delivering an infrastructure that can support the above services between the top-10 sites for ATLAS and CMS and can scale to support all Tier1/2/3s.

*Note that LHCONE is a value-add to the LHC and NREN community and to avail of this value-add a set of requirements are put upon NRENs and T1/2/3 centres. In any case full IP connectivity between all T1/2/3s will continue to be provided by NRENs and GEANT as normal.*

The planning for the prototype requires infrastructure changes within GEANT and NREN networks and co-ordination with ESNET and Internet2. It also requires co-ordination with ASCG, TIFR and sites in Tokyo and Beijing.

It must be noted that to support services #2 and #3 above no infrastructure changes are required in GEANT and NREN networks and global connections. Service #2 is already offered via GEANT Plus services extending into NRENs and US. Service #3 is essentially the automation of service #2 along with bandwidth guarantees and is currently under pilot preparation in GEANT/NRENs (i.e. Autobahn pilot).

Hence the prototype will focus on implementing infrastructure to deliver services #1 and #4 above as the current infrastructure deployed on GEANT does not support them.

**Mapping of roles of NRENs and GEANT to LHCONE architecture**

Within Europe there should be a number of GEANT PoPs that act as *distributed exchange points* as defined in the LHCONE architecture paper. The NRENs will provide national infrastructures to
aggregate the T1/2/3s, in effect aggregate networks that act as connectors to LHCONE (using the terminology from the LHCONE Architecture paper).

The European distributed exchange points will have connections to US based (distributed or not) exchange points and will accommodate connectors from China and India.

It is possible that there will be single exchange points in Europe too, that connect to GEANT’s distributed exchange points (e.g. netherlight), for T2s not connected to NREN/GEANT, if any.

One or more Route Servers will be deployed in Europe to facilitate exchange of BGP information between the connectors to the LHCONE infrastructure

**NREN technical considerations**

NRENs have several means of developing national infrastructures to collate the T1/2/3s in their respective countries. This paper does not address how this is done but it sets out the requirements that in order to comply with the LHCONE architecture the NREN must be able to present a separate interface to GEANT which carries the related traffic. This requirement applies also to the T1/2/3 centres where they will have to be able to distinguish between a connection to the NRENs’ normal infrastructure and the infrastructure in support of LHCONE

**GEANT technical considerations**

The equipment currently deployed within GEANT does not support the full set of LHCONE services in their current configuration.

The ALU MCC 1678 is able to support service N.2, which is an existing operational service. Service N.3 is under pilot preparation within NRENs and GEANT at present. Service N.1 is not supported by the ALU MCC1678. It is expected that the current equipment tender for GEANT will deliver equipment able to support this service in ~ April 2012.

GEANT’s Juniper T1600 routers have VPN functionality that can emulate service N.1 (i.e. VPLS) and hence the GEANT routers can be configured to act as *LHCONE distributed exchange points*.

For this to be effective, it is necessary that the NRENs are configured to act as aggregate networks or connectors and present a logically, or physically separate interface towards the GEANT routers. This means the NRENs need to present to GEANT’s PoPs one of:

- separate physical port (to Juniper router or ALU MCC)
- logical separation by means of a VLAN or LSP on the existing IP interface

**Prototype design**

The design of the prototype has 3 main components:

- intra-European design
trans-atlantic design
connectors from Asia

Intra-European design

A set of GEANT PoPs will act as distributed exchange points to which the NREN aggregate networks will connect to. NRENs will have to differentiate the T1/T2/T3 traffic destined for LHCONE from the normal IP traffic. To do this, as stated above, NRENs will need to use either a physically separate interface or a logically separate one (i.e. VLAN or LSP).

In order to support stable and reliable BGP information exchange and routing between the connectors, a route server should be installed in Europe and maintained by GEANT.

In the long term we expect that switching equipment resulting from the imminent equipment tender will provide the functionality required to support the LHCONE architecture in a cost efficient manner. This will entail deployment of carrier-grade devices offering E-LAN services and EPL services with dedicated 10Gbps links or a share of 100Gbps links between them.

In the short term, for the purposes of the prototype, two options exist for making the GEANT PoPs act as distributed exchange points:

- use of VPLS on GEANT’s Juniper T1600 routers
- purchase and deployment of Ethernet switches in the GEANT backbone such as Juniper EX4500

use of VPLS on Juniper routers

In this case a set of distributed exchange points is established by configuring a Ethernet VPN (using VPLS technology) over the existing routers. Looking at the available connectivity, (40Gbps CH-DE-NL and 20Gbps NL-UK-FR-CH and 20Gbps CH-IT) connectivity for the Ethernet VPN is shared with GEANT’s normal IP connectivity hence no additional capacity is required on GEANT.

In this scenario, the aggregation follows the access point model where each router emulates a port in a distributed switch model.

For the provision of end-to-end circuits between T2 sites, the end sites will request via their NRENs a GEANTPlus service instance following the established procedures hence no specific infrastructure needs to be established
In this example, prototype sites in Spain are connected to the distributed exchange point in GEANT’s Paris PoP. This can be achieved by utilising a GEANTPlus instance between Madrid and Paris or via a 10Gbps GEANTLambda instance. Alternatively, and perhaps preferably, the VPLS setup could be extended to include the GEANT PoP in Madrid.

*use of Juniper Ethernet switches*
In this case a set of native Ethernet switches (e.g. Juniper EX-4500) are deployed inside GEANT POPs and are interconnected by dedicated 10Gbps links. The NREN aggregate networks need to connect to this infrastructure by means of a separate dedicated physical port. It requires purchase of native Ethernet switches and deployment of 10Gbps links between them. As in the example using VPLS, sites for example in Spain could use remote connections to one of the distributed exchange points in Milan or Paris (for example there is a 10G Madrid-Milan that carries very little traffic)

The Pros&cons of these two approaches are discussed further below
Trans-atlantic design

It is necessary to bring the state of the current trans-Atlantic capacity plans into perspective and see how they fit with LHCONE.

At present GEANT avails of 40 Gbps of capacity to US, 20 for IP and 20 for GEANT Plus. These circuits are funded on a 50-50 basis by GEANT and US partners.

The current setup is as follows
The current plan is to increase capacity to 60Gbps, 40 for IP and 20 for GEANTPlus as depicted below. This is again funded 50-50 by GEANT and US partners. The US funding party is now the ACE project led by Jim Williams.
For the purposes of LHCONE the links to Washington are not useful as no connector is based near Washington, at least for the prototype phase. The useful locations are New York (MAN-LAN) and Chicago (Starlight).

A new VLAN in support of LHCONE should be setup on the IP links between NYC and Amsterdam, sharing capacity with the general IP connectivity on those links. Alternatively one of the two links may be dedicated to LHCONE. An additional 10Gbps link, dedicated to LHCONE, should be provided by GEANT between Chicago and a GEANT PoP (for example in Geneva, Paris or London). The cost of such a link is estimated at 200 kEuros/p.a.

The links to the MCCs (in red) in London and Paris are used to support GEANTPLus services to US sites, supporting service N.2 of LHCONE.

Connectors from Asia
This is for further analysis

Cost analysis
Establishment of the LHCONE infrastructure will incur a number of GEANT related costs that will have to be covered. Other national costs will be covered by the respective NRENs
As a basic assumption, it is assumed that all costs that are directly attributable to LHCONE are shared amongst the beneficiaries. How the costs are shared will be determined by the CSWG (Cost Sharing Working Group).

Costs that are directly attributable to LHCONE include:

- cost of dedicated transatlantic links (e.g. Chicago-Geneva) procured by DANTE
- cost of dedicated 10Gbps links within GEANT fibre footprint interconnecting the distributed exchange switches
- cost of access links and ports to LHCONE for an NREN that connects to a remote PoP
- cost of dedicated HW (switches, route servers)

In addition to the directly attributable costs, there are costs related to the use of existing shared resources, which include:

- utilisation of shared capacity on transatlantic links and GEANT backbone links
- utilisation of routers to implement Ethernet VPN (VPLS)
- NOC costs

These shared resource costs will also be taken into account by the CSWG once LHCONE is operational. For the prototype they may or may not be taken into account.

The cost analysis below assumes that for the prototype phase only the directly attributable and additional costs need to be considered.

For the prototype phase we make the following assumptions:

- 6 distributed exchange switches deployed in UK, NL, DE, CH, IT and FR GEANT PoPs
  - Shared GEANT capacity between them (VPLS) or
  - Dedicated 10Gbps links between them (native Ethernet) with dedicated switches
- A remote 10Gbps connection from ES to connect PIC and the T2 site involved in the prototype to FR or IT. An option for this is to utilise the GEANT ES-IT 10Gbps links which is scarcely utilised at present
- A dedicated 10Gbps link from Chicago to a GEANT PoP
- Deployment of 1 or 2 route servers
- PerfSONAR Monitoring systems (costs TBD)

The table below shows estimated additional costs for delivering the LHCONE prototype infrastructure via the two options described earlier. For the VPLS based solution, it is assumed that the current shared IP connectivity is sufficient to carry the expected traffic for the duration of the prototype and until 2012. If traffic levels increase, some of the GEANT links may need to be upgraded further at a cost of 60K per link per year.
Note that pricing of a link from Chicago to a GEANT PoP is indicated at ~200K. In reality this depends upon the GEANT location and the choice of provider. 200K should be seen as an indicative average price rather than actual price we may need to pay.

<table>
<thead>
<tr>
<th></th>
<th>Ethernet VPN (VPLS) solution</th>
<th>Native Ethernet Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>One-off cost</td>
<td>Annual recurring cost</td>
</tr>
<tr>
<td>10Gbps Chicago-GEANT</td>
<td>200K</td>
<td>200K</td>
</tr>
<tr>
<td>7xDedicated 10Gbps links</td>
<td>n/a</td>
<td>420K</td>
</tr>
<tr>
<td>6x dedicated Ethernet switches</td>
<td>n/a</td>
<td>120K</td>
</tr>
<tr>
<td>Route server (x2)</td>
<td>15K</td>
<td>15K</td>
</tr>
<tr>
<td>Remote access for ES sites</td>
<td>0 (uses ES-IT link)</td>
<td>0</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>15K</strong></td>
<td><strong>200K</strong></td>
</tr>
</tbody>
</table>

**pro and cons analysis of VPLS vs. Ethernet switches**

<table>
<thead>
<tr>
<th></th>
<th>Ethernet VPN (VPLS)</th>
<th>Native Ethernet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pros</strong></td>
<td>• Less expensive for prototype • VPLS provides fail-over and removes ethernet loops • Has OAM • More flexibility for NREN connections (VLAN, LSP, separate port)</td>
<td>• matches long term implementation • simpler technical setup • total separation of the IP platform</td>
</tr>
<tr>
<td><strong>cons</strong></td>
<td>• need for staff training and establishment of operational processes and development of OSS tools • no real separation from the IP platform.</td>
<td>• more expensive for prototype • lengthy lead times for delivery of dedicated circuits • need to manage ethernet loops • No OAM • Potential conflict with GEANT equipment tender • Fewer options for NREN connections (needs separate port) • Need to establish operational processes and develop OSS tools</td>
</tr>
</tbody>
</table>
From this table it appears that for the prototype phase of LHCONE the Ethernet VPN solution is more suitable. It is considerably cheaper and provides greater flexibility for the NREN connections. It is cheaper simply because with the VPLS solution we envisage sharing the currently deployed IP capacity rather than implementing dedicated 10Gbps links as forseen in the long term. With the current traffic levels we are confident that both normal and LHCONE traffic can be supported, for some time, at least for a prototype phase until 2012. When the GEANT equipment tender is complete and the new switching equipment rolled out, it will be possible to move to a native solution, which will have functionality that overcomes the limitations of equipment that we could purchase now within the public tender boundaries (i.e. OAM, loop prevention,...). At that time we would also expect capacity requirements to have grown and probably exceed the current shared IP connectivity on GEANT used by the VPLS solution.

Operational aspects

The operational aspects relate to the operation of the technology to support LHCONE and service N.1. In the case of using VPLS on the existing Juniper T1600 routers, whilst the technology is in general proven, it has never been offered within GEANT. There is therefore currently a gap in technical expertise, operational processes and monitoring functionalities which will all take time to overcome. The operational model will also need to be enhanced to support route servers. For the prototype phase the GEANT Noc will be offering service on a best effort basis until the tools, processes and the team are up to speed with the requirements associated with an operational service.

Issues to resolve for prototype

It is not clear at present if JANET and REDiris, which both host sites foreseen for the prototype will accept the LHCONE architectural model and if they are willing or able to present to GEANT a separate interface. If these two NRENS do not intend to do any traffic separation for the T1/2/3s then their role in the prototype cannot be supported.

The LHCONE architecture foresees the deployment of a multi-domain monitoring infrastructure. The candidate for this is perfSONAR. There is ongoing work in GN3 to have a more stable and reliable SW base for perfSONAR that can be deployed.

Next Steps

For the Prototype we need to check that it meets the use-cases and includes the ATLAS and CMS sites provided by the experiments.

Both for the LHCONE prototype and for the evolution to a full infrastructure there is a need to present the current ideas and to discuss the implications with all the NRENS that serve the Tier2 community in Europe. Perhaps a good way to do this would be to circulate the paper and organise a VC.
It would then be most helpful if the NRENs could discuss the proposal with their LHC sites to ensure that it meets their requirements, this might well include the other two LHC experiments ALICE and LHCb.

This could be followed by a workshop which would include members from the NRENs and country representatives of the Tier 2 sites.