LHC Tier-0 to Tier-1 High-Level Network Architecture

Produced by the LHC high-level architecture group
with contribution and comments from many.

Authors:
Erik-Jan Bos (SURFnet),
Edoardo Martelli (CERN), Paolo Moroni (CERN)

Editor:
David Foster (CERN)

Version   2.0
Date   July 30, 2005

1. Introduction

The Large Hadron Collider (LHC) is being built at CERN in Geneva, Switzerland. The large amounts of data to be produced by the LHC are scheduled to be sent to data processing and storage centres around the world.

The data source is called "Tier 0" (T0) and the first level processing and storage is called "Tier 1" (T1). The "Tier 2" (T2) sites are typically universities and other scientific institutes that depend on services from one or more T1 sites.

The entire large-scale scientific instrument can be depicted as the collection of:
1. The LHC and its data collection systems;
2. The data processing and storage units at CERN, i.e. T0;
3. The data processing and storage sites called T1;
4. The data processing and storage sites called T2;
5. Associated networking between all T0, T1, and T2 sites.
Figure 1 shows this in more detail.

This document proposes the high-level architecture for the LHC network in so far as it concerns the Tier-0 and Tier-1 centers. This is only part of the overall computing solution for LHC which includes the Tier-2 centers. It is assumed for the purposes of this document that Tier-2 connectivity is dealt with elsewhere. The reader is referred to Appendix D for T2 networking considerations.

The aim of this high-level architecture is to be inclusive of technologies, while still proposing concrete directions for further planning and implementation or Tier-0 to Tier-1 connectivity.

With respect to T0-T1 networking this document proposes a detailed architecture based on permanent 10G light paths. These permanent light paths form an **Optical Private Network (OPN)** for the LHC instrument.

Other definitions:

LHC Network Traffic: The data and control traffic that flows between T0, the T1s, and the T2s.

LHC prefixes: IP address space allocated by the T0 and T1s and assigned to the machines connected to the LHC-OPN.

Light path: (i) a point to point circuit based on WDM technology or (ii) a circuit-switched channel between two end points with deterministic

behaviour based on TDM technology or (iii) concatenations of (i) and (ii).

Examples of (i) are:
- STM-64 circuit;
- 10GE LAN PHY circuit
Examples of (ii) are:
- a GE or 10GE channel carried over an SDH/SONET infrastructure with GFP-F encapsulation;
- an STM-64/OC-192 channel between two points carried over an SDH/SONET infrastructure

NREN: While the abbreviation means National Research and Education Network, this commonly used term is used throughout this text as the collective name for a network that service either the research community or the education community or both.
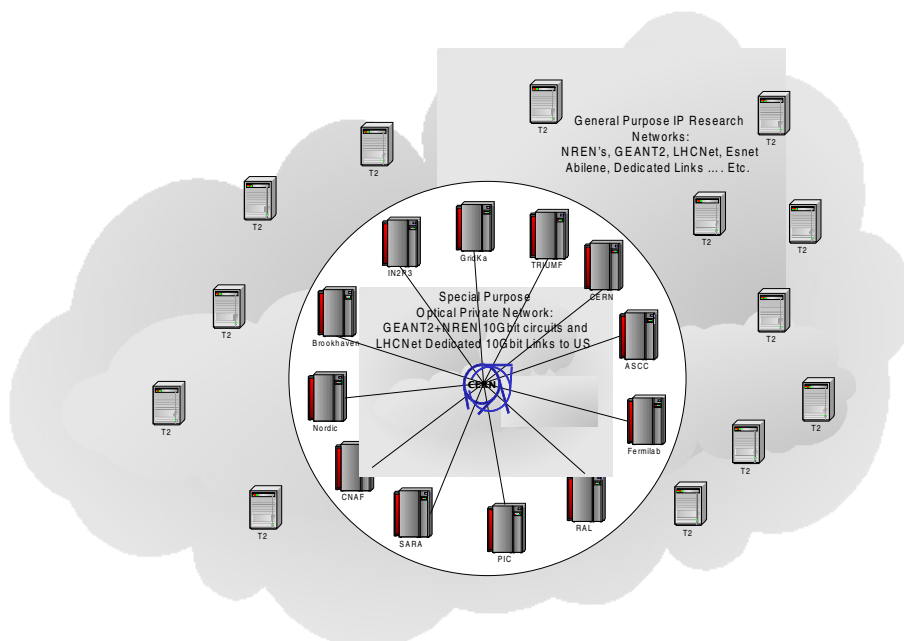

Figure 1. The large-scale scientific instrument.

Appendix C contains a glossary and many of the terms used in this document are explained here.

2.Network background information

The LHC Network is designed to move data and control information in the context of the LHC experiments. This data traffic will consist of the raw and derived data and control information exchanged among the machines connected to the LH–OPN that will have visibility outside the local Tier.

The 12 envisaged T1s are enlisted in Table 1.

T1 name	T1 location	NRENs involved
ASCC	Taipei, Taiwan	GÉANT2 – ASnet
Brookhaven	Upton, NY, USA	ESnet – LHCnet
CERN	Geneva, Switzerland	(not applicable)
CNAF	Bologna, Italy	GÉANT2– GARR
Fermilab	Batavia, IL, USA	ESnet – LHCnet
IN2P3	Lyon, France	GÉANT2 – RENATER
GridKa	Karlsruhe, Germany	GÉANT2 – X–WiN
SARA	Amsterdam, The Netherlands	GÉANT2 – SURFnet6
NorduGrid	Scandinavia	GÉANT2 – NORDUnet
PIC	Barcelona, Spain	GÉANT2 – RedIRIS
RAL	Didcot, United Kingdom	GÉANT2 – SuperJANET
TRIUMF	Vancouver, BC, Canada	GÉANT2 – NetherLight – CA*net 4

Table 1. The 12 envisaged T1s

The resources available at the T1s will not be all the same and therefore the average network load is expected to vary. In addition, the anticipated peak load is an important factor as it is this peak load that the network should be capable of sustaining. While the computing models continue to be refined, this is becoming clearer.

For the moment the agreed starting point is the provisioning of at least one 10 Gbit/s transmission path between each T1 and T0. However, this is not to be considered as the requirement for the future of LHC data processing. This is considered as a reasonable starting point for LHC startup in 2007.

3.Responsibilities

The responsibility of providing network equipment, physical connectivity and manpower is distributed among the cooperating parties.

CERN will provide the interfaces to be connected to each T1's link termination point at CERN. Furthermore, CERN is available to host T1's equipment for T0–T1 link termination at CERN, if requested and within reasonable limits. If this is the case, the T1 will provide CERN the description of the physical dimensions and the power requirements of the equipment to be hosted.

The planned starting date for the production traffic is June 2007, but T1s are encouraged to proceed with the provisioning well before that date, already within 2005. Nevertheless, they must normally be ready at full bandwidth not later than Q1 2006 to be able to participate in Service Challenge–4 (SC4). This is important as the Service Challenges now underway need to build up towards the full capacity production environment exercising each element of the system from the network to the applications. It is essential that the full network infrastructure is in place, in time for testing the complete environment.

Every T1 will be responsible for organising the physical connectivity from the T1's premises to the T0, according to an MoU[1] between the T0 and the T1s.

Every T1 will make available in due course the network equipment necessary for the termination point of the corresponding T1–T0 transmission path at the T1 side, in agreement with the local connectivity plans of the corresponding NREN.

4.LHC networking strategy

The proposed networking strategy is to use at least one dedicated 10 Gbit/s light path between T0 and each T1. This 10G light path terminates at networking gear at the T0 and inside or as close as possible and desirable to the T1. This "far" end-point, as seen from T0, is likely to be either at the T1 itself, at the NREN serving the T1 or any other intermediate service provider. The networking gear on both ends of a light path should be capable of speaking BGP4, and an eBGP peering will be set-up between the BGP speaker at T0 and the BGP speaker at the first hop. Figure 2 shows the proposed topology in more detail.

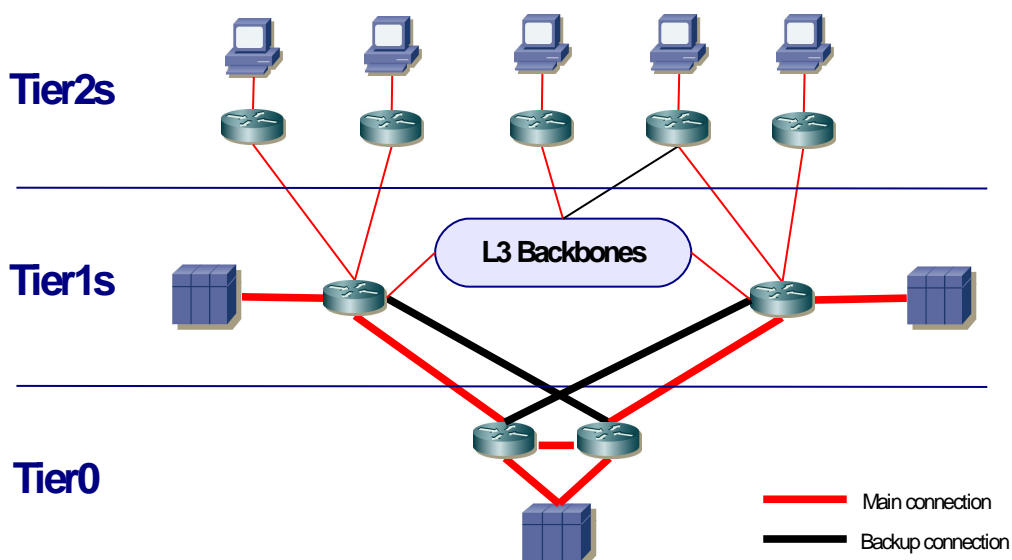[1] See: http://lcg.web.cern.ch/LCG/C–RRB/MoU/MoU.pdf

Figure 2. Network architecture

IPv4 is the network protocol chosen to provide communications among the upper layers applications at the first stage; other network protocols like IPv6 can be considered in the future. Every Tier is encouraged to ensure that support of an MTU size of at least 9000 bytes will be possible on the entire path between T0 and the T1.

From the architecture point of view, every T0–T1 link should handle only production LHC data (LHC Network Traffic). The routing among the T0 and T1s sites will be achieved using BGP.

T1–T1 traffic via the T0 will also be allowed, although T1s are encouraged to provision adequate other connectivity by means of direct lightpath, tunnel over IP/MPLS networks, transit over IP backbones, or any other suitable technology.

In order to provide high reliability, it is recommended to foresee a backup using a separate fibre path for every T0–T1 link. The backup connectivity can be achieved using additional Light paths. Unless there is a well defined plan to provide enough bandwidth, the use of backup Layer 3 paths across NRENs and Research Backbones is discouraged as this might heavily interfere with the other research and general purpose Internet connectivity of T0 or the T1s[2].

---

[2] The LHC Network Traffic between T0 and each T1 is expected to be such that normal service of the routed Internet infrastructures would be disrupted by having one or more LHC Network flows over this, By talking to the experiments it became apparent that in case of a temporary failure of a particular T0–T1 link it's best to wait for the repair of such link or to rerouted through another part of the LHC-OPN, but not to use the routed Internet to avoid disruption and complaints. The computing models of the experiments foresee sufficient buffers to sustain temporary outages.

It is known that a number of Tier-1's may be interconnected using cross-border fiber in the near future a a mechanism to provide backup as well as direct T1-T1 connectivity. In case a T1-T1 connection is already known to exist today or in the near future, the T1(s) involved are encouraged to make this known to the LHC networking community.

On T0's networking equipment, for connecting the T1's access links, 10GE LAN PHY ports will be available. Ports of flavour 10GE WAN PHY or STM-64/OC-192 can be negotiated between CERN and an individual T1 on request.

More technical details about the IP addressing and the routing configuration can be found in Chapter 8.

5.Security considerations

It is important to address security concerns already in the design phase, especially because the expected amount of network traffic can't be treated by the security equipment already in-place. Some Tiers might decide to connect the data processing and storage units directly to the LHC-OPN, bypassing the existing firewalls, and thus opening some potentially security breaches. The measures that are going to be suggested in this chapter are intended to protect the access to the LHC-OPN resources via the LHC-OPN network from any unauthorized source. However, if the LHC-OPN resources are also connected to the general purpose network, they might be accessed by anyone else, according to the security policies of the owner.

The architecture proposed will create a private network that will connect several domains. It is important that the Security Officers of all the domains involved are aware of all the potential risks, and that they formally agree with the security implementation proposed.

In a general purpose environment, ACL-based network security does not guarantee sufficient protection. However, because of the relatively low number of expected LHC prefixes and of expected (Grid) applications, it is considered possible in the LHC networking case that appropriate ACLs can sufficiently reduce the risks involved with unrestricted Internet reach-ability at reasonable cost.

By means of ACLs, the devices on the LHC-OPN will be kept as protected as possible from external access: Assuming that TCP and UDP port numbers involved in the expected applications are known, only traffic between source-destination pairs of LHC-OPN addresses and on specific port numbers will be allowed between T0 and T1s, whereas the default ACL behavior will be to discard any other packets. Obviously, some

exceptions are most likely to be required: packets belonging to routing protocols and basic troubleshooting tools (ping, traceroute, etc.), monitoring tools (SNMP), etc. Other exceptions might be considered for specific applications for which regular port numbers are not known by default.

Attention should be paid to the configuration of client (high-numbered) TCP ports: while the principle of applying the knowledge of application ports is still applicable in theory, this would considerably increase the size of every ACL; if possible, it is recommended to take advantage of Cisco-like "permit tcp established" ACL statements. This is not possible for UDP, so for UDP applications (if any) to work, a compromise should be agreed between the number of the corresponding ACL entries and the security risk of opening all high-numbered UDP port numbers. Further input is required from the physics applications in order to refine the UDP situation.

Input ACLs will be applied at the T0 side on the interfaces towards the T1s. Similarly, T1s are heavily encouraged to apply input ACLs on their interfaces facing the T0. If a T1 is not in the position to apply input ACLs towards the T0, the T0 might consider applying output ACL towards that T1.

It is important to secure all the access to the LHC-OPN network: every Tier must localize every connection with the general purpose network and secure it against abuse and leaking of traffic. Special care must be taken with the backup connection, especially if it is not a dedicated link.

The details of the security model are now the subject of a separate working group (Appendix B).

6.Further details on networking

This chapter describes the networking situation on a per continent basis, as each of the three categories that are elaborated on below have their own specifics.

6.1 The European situation

In Europe, DANTE is building the GÉANT2 network[3], in the context of the EU funded GN2 project. GÉANT2 will be an optical and packet switching network, also known as a "hybrid network". The GÉANT2 network will connect 34 countries through 30 national research and education

[3] More information on GÉANT2 can be found on-line at the URL below: http://www.geant2.net/

networks (NRENs), using multiple 10Gbps wavelengths. In a number of countries, especifically for those NRENs listed in Table 1, GÉANT2 will have the possibility to deliver light path services right from the start, between CERN and the NREN of that country.

It is planned that GÉANT2 will deliver the connectivity between CERN and all the European T1s, except for CERN in its role as T1 and for IN2P3 which has dark fibre into CERN. For all other European T1s, GÉANT2 will provide one 10G light path per T1 between the T0 and the NREN involved.

For T1–T1 links inside Europe, a choice of GÉANT2 light paths and cross border dark fibres, lit by two adjacent NRENs, can and will be used. To establish this topology, DANTE, the NRENs involved and the LHC community need to work together to come up with an optimal solution from which the LHC experiments can benefit most both technically and financially.

6.2 The North American and Canadian Situation

The US HEP-community is arranging for two 10G waves between T0 and selected locations in the US, namely MAN LAN in New York City and StarLight in Chicago. The connectivity to the US and within the US is a combination of initiatives, mainly the LHCNet initiative led by Caltech and the ESNet initiative.

The Canadian HEP community is currently working with CANARIE to engineer a 10G solution between Vancouver, BC and Geneva which is anticipated to be ready by September 2005.

6.3 Networking to Taipei

ASNet has procured bandwidth between Taipei and Amsterdam, through StarLight in the US. This link currently is an OC–12 into Amsterdam, but is likely to be upgraded over time to OC–48 and OC–192, if the bandwidth to be procured from carriers into Taipei allows ASNet to do so.

Also ASNet has constructed a Point of Presence (PoP) in Amsterdam, at SARA. Between Amsterdam and Geneva, ASNet is looking for options to connect.

7.Operations

It is clear that all entities contributing to the LHC-OPN have a level of responsibility in ensuring the smooth operation of the networking. Fault detection, diagnosis, resolution and reporting are all complex functions that require disciplined coordination and good communication channels among the parties involved. Similarly, day to day configuration of the infrastructure to add new locations or functionality also requires coordination. This chapter proposes a "Keep It Simple" approach by introducing the LHC-OPN "thin" NOC which has a well defined role as described below.

It seems likely that the network equipment of the LHC-OPN will be procured, owned and managed by several responsible parties, as defined in the "Responsibilities" chapter. The LHC-OPN NOC will have at least read-only access at least to the T0 and T1 devices of the LHC-OPN (read-only access to other intermediate devices can also be considered, whenever agreed). Through this access, the NOC will proactively monitor the status of the infrastructure.

The NOC will provide a single point of contact for the users at the T0 and the T1s of the LHC-OPN for fault reporting and correction. It will liaise with all parties contributing to the infrastructure in order to diagnose faults and to ensure they are resolved. The NOC will not resolve configuration and equipment faults, but will rely on the intervention of the appropriate partner in the overall infrastructure. The NOC will issue periodic reports to the LHC user community on the resolution of the faults, the network utilization and other information relevant to the LHC-OPN infrastructure.

The details of the operational model are now the subject of a separate working group (Appendix B).

8. IP addressing and BGP Routing

In order to effectively manage some network security and routing, it is essential to aggregate as much as possible the IP addresses used in the context of LHC network traffic.

The following guidelines are to be followed:
- Every T1 and the T0 must allocate publicly routable IP address space to the machines that need to be reached over the T0–T1 links (the "LHC prefixes", as defined in the Introduction).
- LHC prefixes should be aggregated into a single CIDR block for every T1; if this is not possible, only a very small number of CIDR blocks per T1 would be acceptable.
- LHC prefixes should be dedicated to the LHC network traffic.
- If a T1 can't dedicate a prefix but need to use its whole address space, then it will be its own responsibility to inject in the LHC–OPN only the traffic destinated to the LHC–OPN resources.
- LHC prefixes can be carved as a CIDR block from a T1's existing allocations or obtained as new allocation from the appropriate RIR through already established channels.
- LHC prefixes cannot be from RFC1918 and related (like RFC3330) addresses.
- T0 will allocate /30 prefixes for the addressing of the T0–T1 links, i.e. the links that connect to CERN up to the first BGP speaker in the path.
- Every T1 and T2 interested in exchanging traffic directly with the T0 is required to provide the T0 with the list of its LHC prefixes. T0 will maintain a global list of all LHC prefixes and inform T1s and T2s about any changes.

External BGP peerings will be established between T0 and each T1. More precisely, the T1 peer is the BGP speaker directly connected to the T0 on behalf of a specific T1, e.g. an NREN connecting a T1 or the BGP-capable gigabit Ethernet switch of the T1.

These are the guidelines for the BGP configuration:
- T0 will use the CERN Autonomous System number (AS513).
- T1s will use the AS number of the entity that provides the LHC prefixes to them or the AS number of their standard upstream NREN.
- Every T1 will announce its own LHC prefixes to T0.
- T0 will announce its LHC prefixes to every peering T1.
- T0 will accept only the LHC prefixes related to a specific T1, i.e.

the T1's own LHC prefixes, plus LHC prefixes of any T1 or T2 for which that T1 is willing to provide transit for.

- T0 will re-announce to all the T1s all the LHC prefixes received in BGP. Nevertheless, since T1s are encouraged to establish direct connectivity among themselves, they can filter out unnecessary LHC prefixes according to each individual T1-T1 routing policy.
- T1 will accept T0's prefixes, plus, if desired, some selected T1's prefixes (see previous bullet).
- T0 and T1s should announce their LHC prefixes to their upstream continental research networks (GÉANT2, Abilene, ESnet) in order to allow connectivity towards the T2s. Special care must be taken by each Tier to not leak out BGP prefixes that belong to other Tiers.
- Usage of static routes is not advisable.
- No default route must be used in T1-T0 routing.
- It is the responsibility of every Tier to make sure that any of its own machines within the LHC prefix ranges can reach any essential service (for instance the DNS system).

The details of the IP Addressing and routing are now the subject of a separate working group (Appendix B).

Appendix A. T2s

T2s' connectivity is beyond the scope of this document. However some indications are given in case T2s need to use the LHC-OPN. T2s usually upload and download data via a particular T1. But it can be necessary for them to reach another T1 or even the T0.
Technically, the transit is easily provided if the T1 announces the T2's prefixes to T0, and all the security barriers are opened for them. But this assumes a "static" allocation of a T2 to a particular T1.

Because of the numerous issues around opening the OPN for the T2's it is assumed that the T0-T2 traffic will be handled by the normal L3 connectivity provided by the research and private network links.

Appendix B. Working Groups

1. **IP Addressing and Routing project-lhcopn-routing@cern.ch**
2. **Security**   **project-lhcopn-security@cern.ch**
3. **Operations**   **project-lhcopn-operations@cern.ch**
4. **Monitoring**   **project-lhcopn-monitoring@cern.ch**

Appendix C. Glossary and abbreviations

ACL	Access Control List
ASN	Autonomous System Number
ASNet	Taiwanese NREN.
BGP	Border Gateway Protocol
CANARIE	Canadian NREN. See: http://www.canarie.ca/
CIDR	Classless Inter-Domain Routing
DANTE	Delivery of Advanced Network Technology to Europe; DANTE's purpose is to plan, build and operate pan-European research networks. See: http://www.dante.net/
DFN	German NREN. See: http://www.dfn.de/
ESnet	The Energy Sciences Network. See: http://www.es.net/
GARR	Italian NREN. See: http://www.garr.it/
GÉANT2	The new European hybrid network interconnecting the NRENs. See: http://www.geant2.net/

IGP	Interior Gateway Protocol
L3	ISO-OSI Layer 3, usually referred to routed IP traffic
LHC	Large Hadron Collider. See: http://lhc-new-homepage.web.cern.ch/lhc-new-homepage/
LHC Network	Network connecting T0 and T1s for the LHC experiments
LHC network traffic	Data exchanged among data centres over the LHC network
LHC-OPN	Large Hadron Collider – Optical Private Network
MoU	Memorandum of Understanding
NOC	Network Operations Centre
NORDUnet	Nordic research and education network. See: http://www.nordu.net/
NetherLight	Open optical exchange in Amsterdam. See: http://www.netherlight.net/
NREN	National Research and Education Network (this term is used throughout this document as a generic term also including a National Research Network such as ESnet)
RedIRIS	Spanish NREN. See: http://www.rediris.es/
RENATER	French NREN. See: http://www.renater.fr/
RIR	Regional Internet Registry
SURFnet	Dutch NREN. See: http://www.surfnet.nl/
T0	Tier 0 site
T1	Tier 1 site
T2	Tier 2 site
UKERNA	UK NREN. See: http://www.ukerna.ac.uk/