LHCONE Multipoint Service Description 3-12-12

The intent of this document is to lay out the current understanding of the LHCONE Multipoint Service that will be a part of the LHCONE effort and the resources that will be available to that service. Further it will attempt to clarify the roles of the various exchange points in the US and Europe and how those facilities might interact. Where there are questions about how this service operates or how the resources interact these will be highlighted if not addressed. Where there are gaps in the service or in the available resources these will be identified.

The goal is to describe how a user might be able to take advantage of one of this service offering to achieve the scientific goals they have.

At a very high level this activity can be seen as a dedicated program specific service that exists as an overlay on the worldwide R&E infrastructure, with multiple and varied ways to access the service.

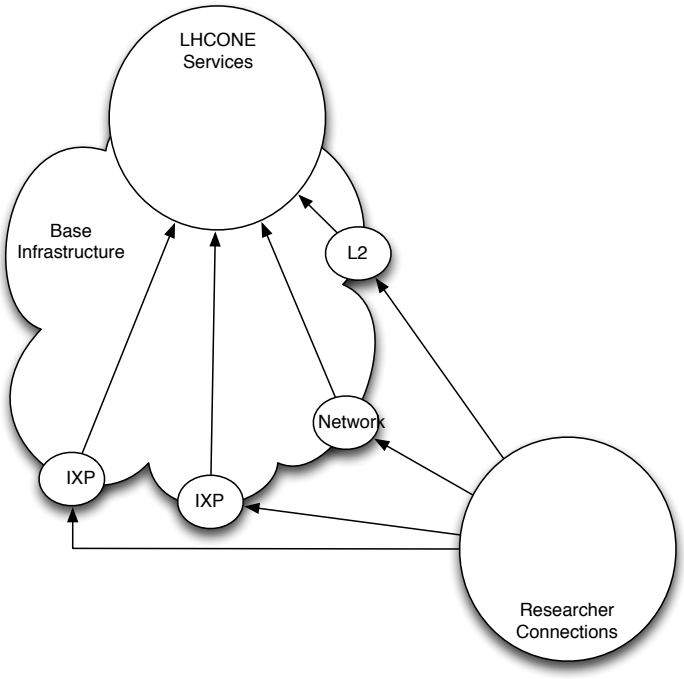

Figure 1: LHCONE Multipoint Service

The base infrastructure is the overall collection of networks and interconnects between networks, e.g. it could be Internet2, ESnet, GEANT, CANARIE, CERN, USLHCnet and all the various trans-continental links. Overlaid on that is the service or services that will be thought of as the LHCONE network. A researcher might

connect to either an IXP e.g. MAN LAN, StarLight, NetherLight, CERNlight or WIX or make a connection to any participating networks. Those connections might be at Layer 2 or utilize other options the network makes available.

What this is intended to illustrate is that the infrastructure exists to bring the actual researchers to the LHCONE services.

LHCONE Multipoint Connection Service:

Description:

The Mulitpoint Connection service allows participants to move traffic between one another as needed.

The initial approach to implementing this service is as a (Virtual) Private, multidomain, IP network restricted to the participants in LHCONE.
It is based on the resources of the general purpose global R&E infrastructure with the ability to add dedicated resources as needed.

Within a single administrative domain, it is possible to implement a shared broadcast domain using a specific IP prefix or it can be implemented via a VRF. In some cases VRFs are created because it is a virtual routing instance on shared routers, in others – e.g. the 6506 located at StarLight - it is entirely dedicated.

The primary difference between how this implementation functions and how a large scale shared layer 2 structure functions is that there are routed boundaries between portions of the shared structures and as a consequence there is a requirements for the exchange of routing information across those boundaries. This information will be exchanged using BGP.

Technical Description of VRF Implementation:

The basic functionality of this service is to provide access to a set of interconnected best effort IP networks. Traffic will be balanced across the resources that this service uses, e.g. trans-oceanic links, through the exchange of IP prefixes using BGP and the application of BGP policy (e.g. prepending or meds)

The following diagrams are intended to show the various resources that will be available to the service, how the control planes (BGP in this case) are structured, how the exchange points (GOLEs) are structured, and how all the resources are interconnected.

The general structure of the Multipoint Service implemented using VRFs and focusing on the relationship between the VRFs and the IXPs is shown in Figure 2:

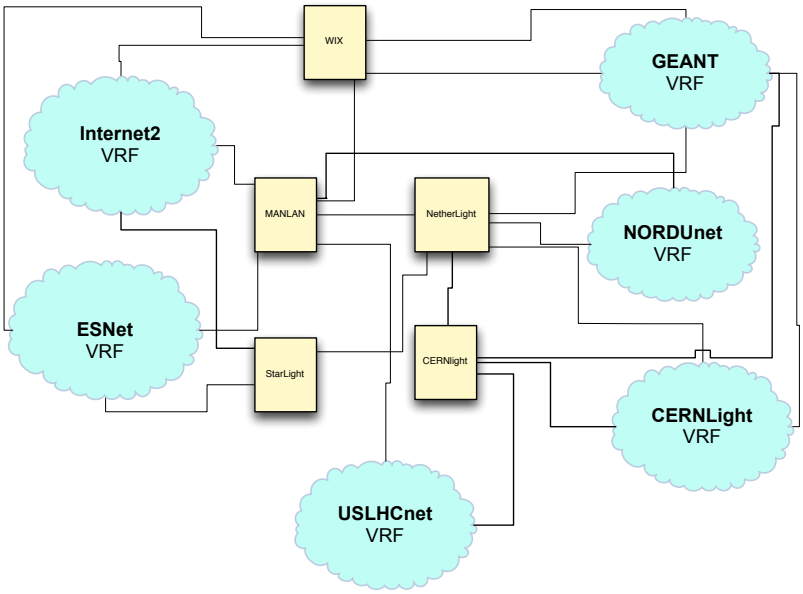

Figure 2: General Structure of Multipoint Service using VRFs

As shown there will be a number of VRFs. These VRFs will connect to one another to exchange traffic either directly, or via the use of exchange points. Some of these GOLEs (Open Exchange Points) will be interconnected using existing Trans-Oceanic resources, others will interconnect using existing R&E infrastructure.

As shown in Figure 3, participating LHCONE sites can connect to an exchange point or to a participating network, or both.

Community	Type	Meaning	Notes
65001:XXXX	Operational	prepend 1x to ASxxxx	Mandatory

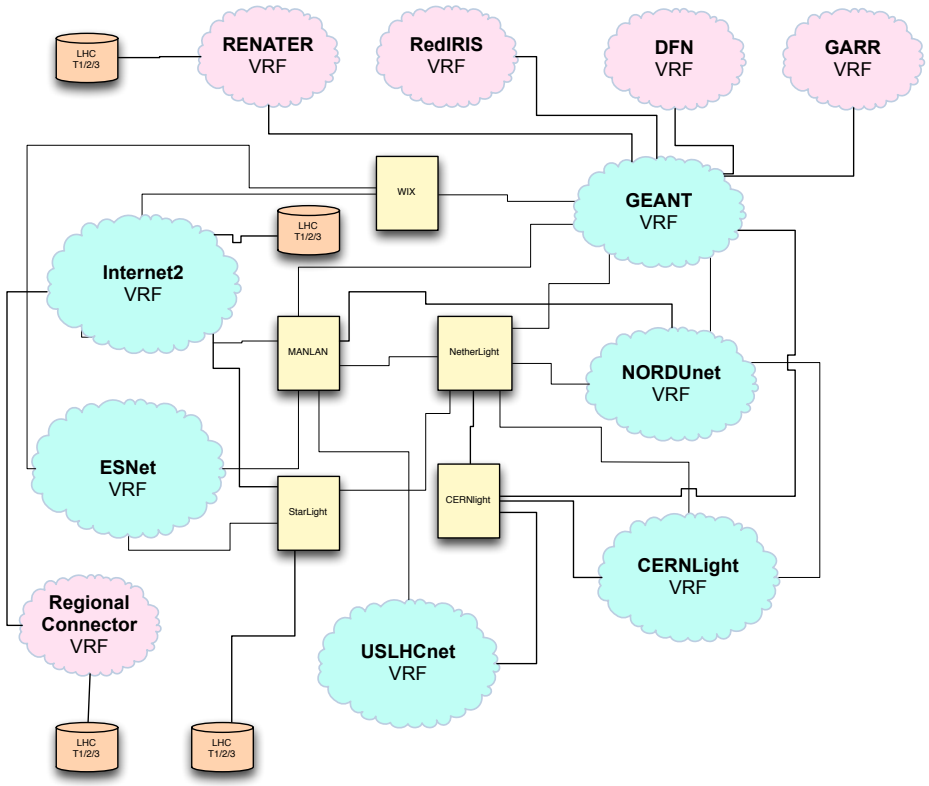

Figure 3: General Structure of Multipoint Service including Connectors and sites using VRFs

For current purposes only the US to Europe connections are shown, Asia and South America will need to be added at a later date, as those connections become better understood. The exact size of the links is not important to the architecture, with the caveat that the traffic across those links will need to be balanced to fully utilize all the existing paths.

BGP Policy

The LHCONE Multipoint Service VRF Implementation has BGP policies that will be either suggested or required.

There will be available a set of BGP communities that sites can implement if they desire. The communities are defined as:

65002:XXXX	Operational	prepend 2x to ASxxxx	Mandatory
65003:XXXX	Operational	prepend 3x to ASxxxx	Mandatory
65010:XXXX	Operational	do not announce to ASxxxx	Mandatory
65011:0000	Operational	do not announce to any Tier-1	Mandatory
65012:XXXX	Operational	do not announce except to ASxxxx	Optional
(tierx-org-as):65101	Informational	Prefix originated by a Tier-1	Mandatory, set by the VRF
(tierx-org-as):65102	Informational	Prefix originated by a Tier-2	Mandatory, set by the VRF
(tierx-org-as):65103	Informational	Prefix originated by a Tier-3	Mandatory, set by the VRF
(tierx-org-as):65151	Informational	Prefix originated by a TierX in Europe	Mandatory, set by the VRF
(tierx-org-as):65152	Informational	Prefix originated by a TierX in Asia	Mandatory, set by the VRF
(tierx-org-as):65153	Informational	Prefix originated by a TierX in America	Mandatory, set by the VRF
(tierx-org-as):65201	Informational	Tier-X supporting ALICE	Optional, set by the Tier-X
(tierx-org-as):65202	Informational	Tier-X supporting ATLAS	Optional, set by the Tier-X
(tierx-org-as):65203	Informational	Tier-X supporting CMS	Optional, set by the Tier-X
(tierx-org-as):65204	Informational	Tier-X supporting LHCb	Optional, set by the Tier-X
(transit-as):65111	Informational	Prefix learned over MANLAN	Optional
(transit-as):65112	Informational	Prefix learned over Starlight	Optional
(transit-as):65113	Informational	Prefix learned over Netherlight	Optional

(transit-as):65131	Informational	Prefix learned over Atlantic link x	Optional
(transit-as):65132	Informational	Prefix learned over Atlantic link y	Optional
(transit-as):65133	Informational	Prefix learned over Atlantic link z	Optional

(tierx-org-as):65201	Informational	Tier-X supporting ALICE	Optional, set by the Tier-X
(tierx-org-as):65202	Informational	Tier-X supporting ATLAS	Optional, set by the Tier-X
(tierx-org-as):65203	Informational	Tier-X supporting CMS	Optional, set by the Tier-X
(tierx-org-as):65204	Informational	Tier-X supporting LHCb	Optional, set by the Tier-X

They are defined as optional and must be set by the Tier-X because each VRF should not have the responsibility to know which VO is served by which TierX and it is unrealistic to make it mandatory.

VRFs will not accept private ASes and private IP addresses..

Exchange Points/GOLEs:

The role of an IXP in the LHCONE Multipoint Service VRF implementation is to provide a neutral location where VRFs that are connected to the IXP can exchange both control information and data. In this case, the control plane information is BGP. Note that while the GOLE is open, there may be policies that are applied to the control plane information exchange. It is also possible that an individual site or a set of organizations might obtain a port on at an IXP.

Within the LHCONE Multipoint Service VRF implementation, the intent is that the connected and participating VRFs will have a full mesh of BGP peerings within any given IXP. The various lines in this diagram are control plane (i.e. BGP) connections between the various participants. This figure is for illustrative purposes and is not meant to exclude or preclude other participants.

Note that some of these connections within the IXP are over Trans-Atlantic links.

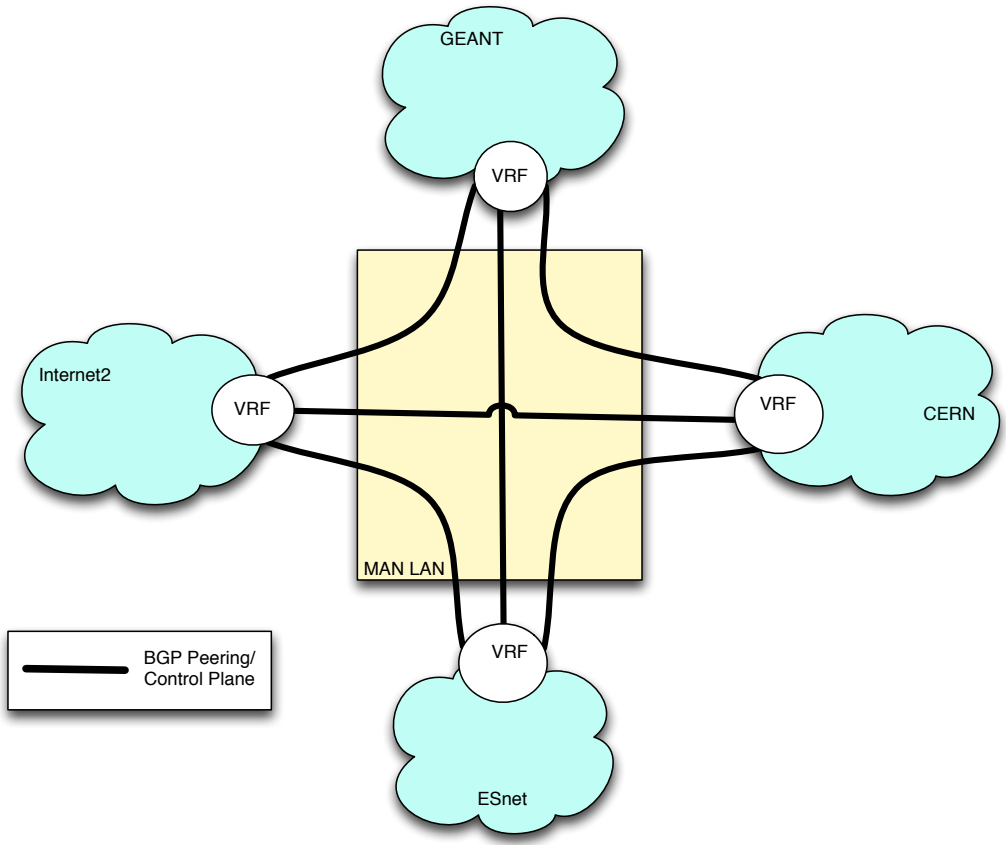

Figure 4: Control Plane (BGP) between VRFs at an IXP

Some sites, e.g. campuses with a direct connection to the IXP, will not be brought into the full mesh of peering. Should a campus or other organization have a link into the IXP they will be expected to peer with a “Patron” organization, resulting in:

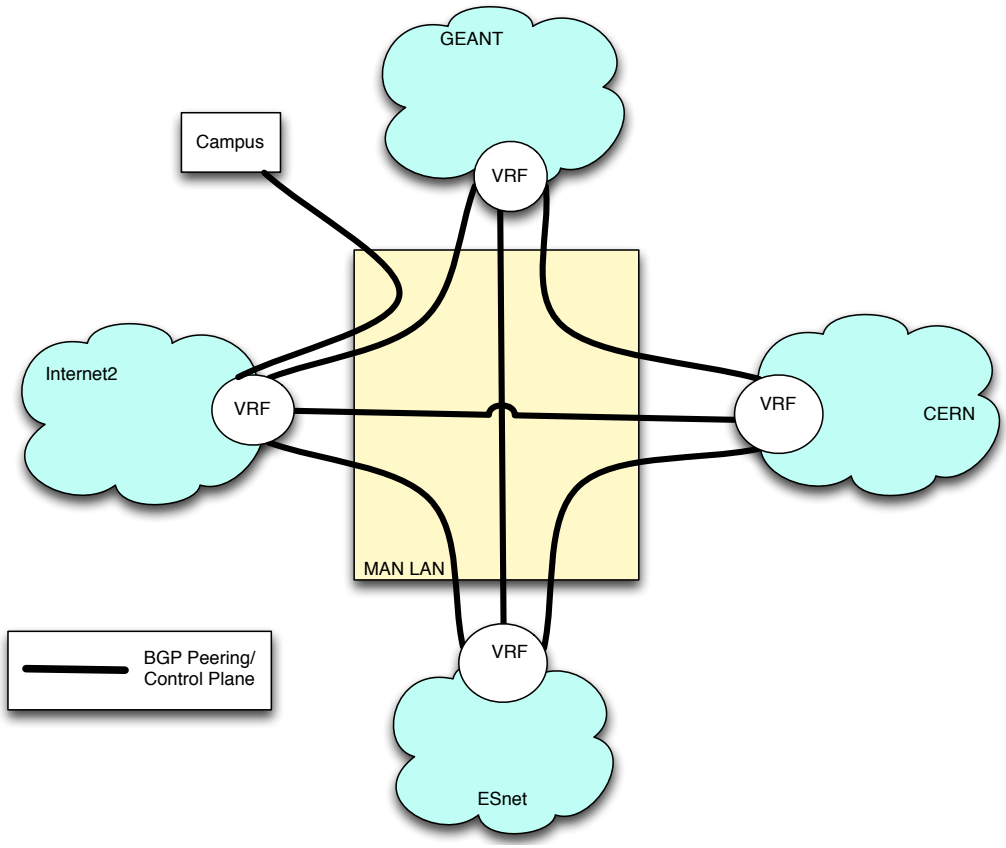

Figure 5: Patron Organizations and the Control Plane

In Figure 5, the campus hands prefix information to Internet2's VRF and in turn gets routing information from that VRF. In order to prevent non-optimal data paths from being used, a technique called 3rd party routing is employed. This has the effect of separating the control plane from the data plane. For this to happen the IXP must create a shared VLAN within the IXP for itself, the other VRFs and any connector it is acting as a "Patron" for. Where 3rd party routing is not required there is no need for the shared VLAN.

There are several IXPs that will be participating in this service. This list may certainly expand over time, but initially these will be looked at:

- MAN LAN
- WIX
- StarLight
- NetherLight
- CERNlight.

MAN LAN and WIX are represented here:

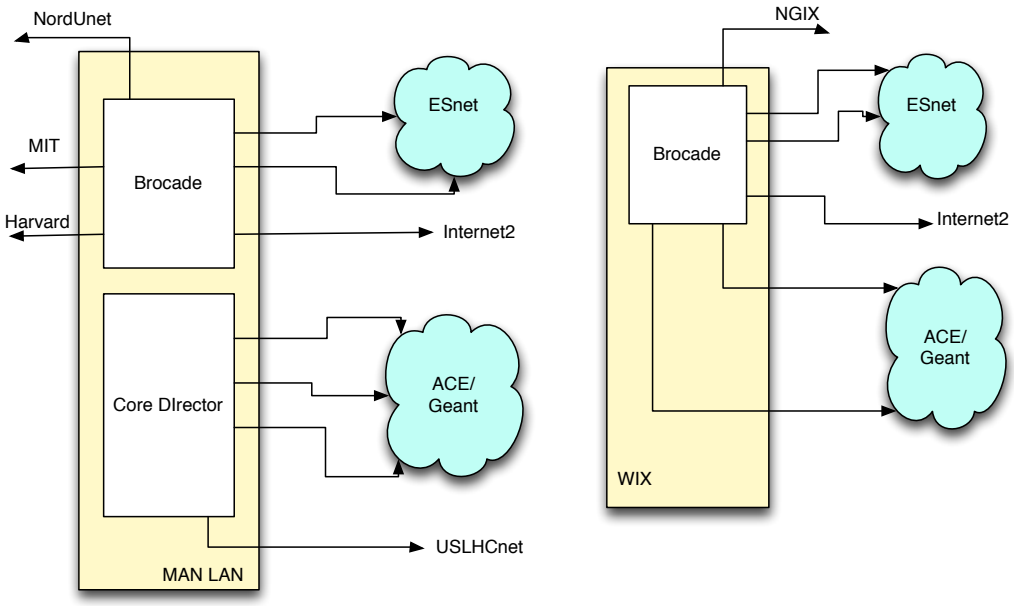

Figure 6: MAN LAN and WIX

Both MAN LAN and WIX will support 3rd party routing if participants ask for that service.

Starlight is represented here:

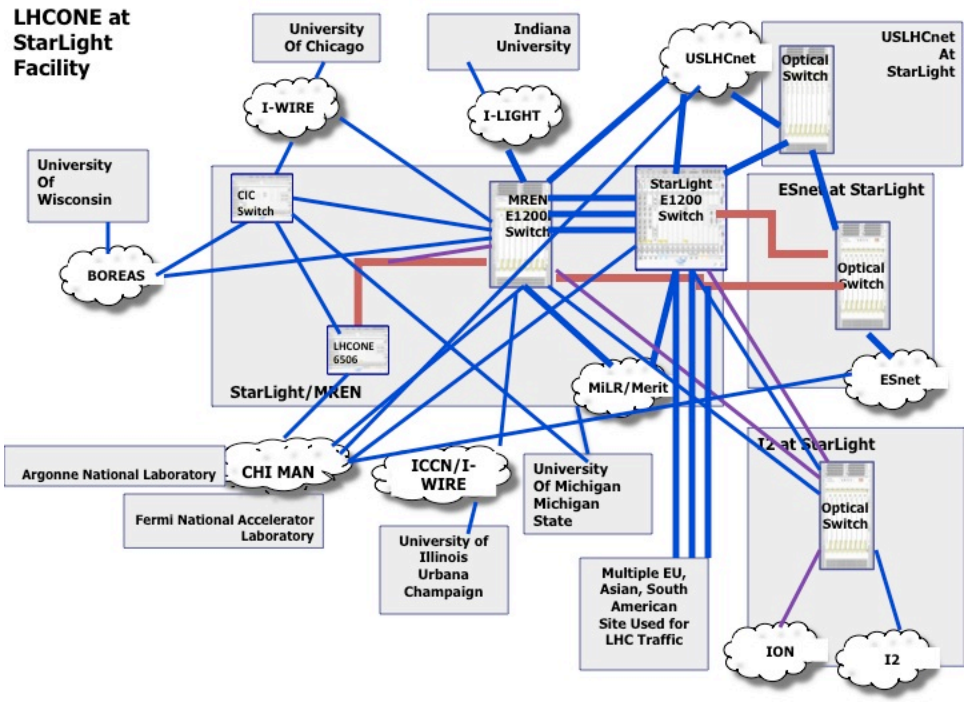

Figure 7: StarLight

Netherlight is represented here:

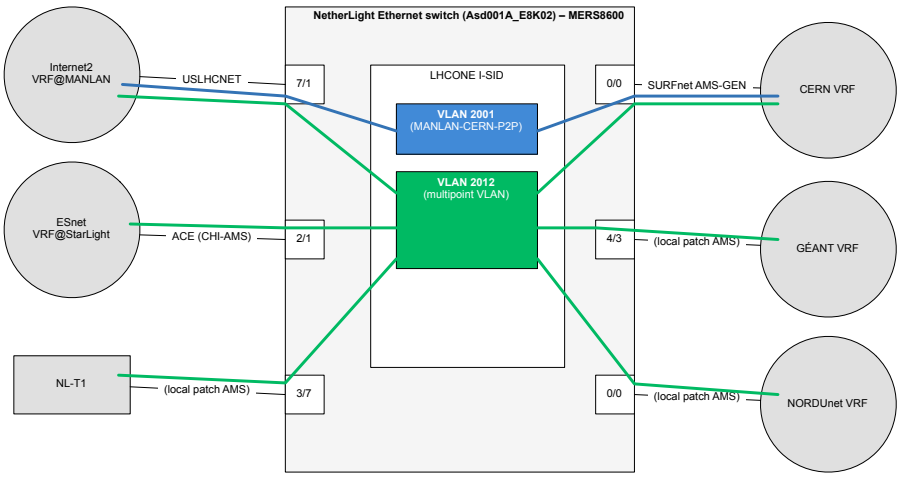

Figure 8: Netherlight

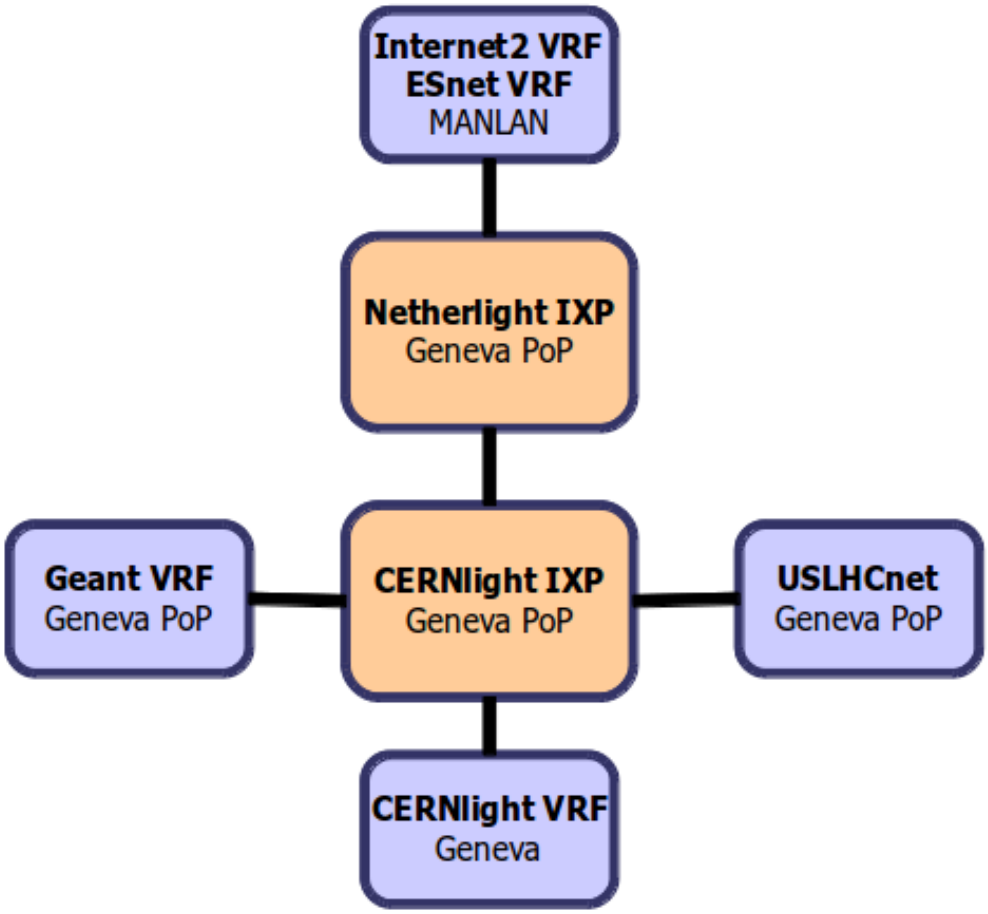

Figure 9: CERNLIGHT

Interconnecting IXP's:

IXPs will generally connect directly via Layer 2 (e.g. MAN LAN and NetherLight). A similar situation may well exist between MAN LAN and WIX or StarLight and MAN LAN **or NetherLight and CERNlight.**

In these cases there will be a circuit between two IXPs, and the data path between a site connected at IXP-A and a site connected at IXP-B may be directly through the two IXPs, but the control of the path is via their respective "patrons".

In the picture below, VRF 1,VRF 2 and VRF 3 would establish a BGP peering across the IXP's and the layer 2 path between them allowing traffic to flow between the sites connected to the VRFs. Notice that they do this on a shared vlan. There are however other vlans operating at the IXP's as well.

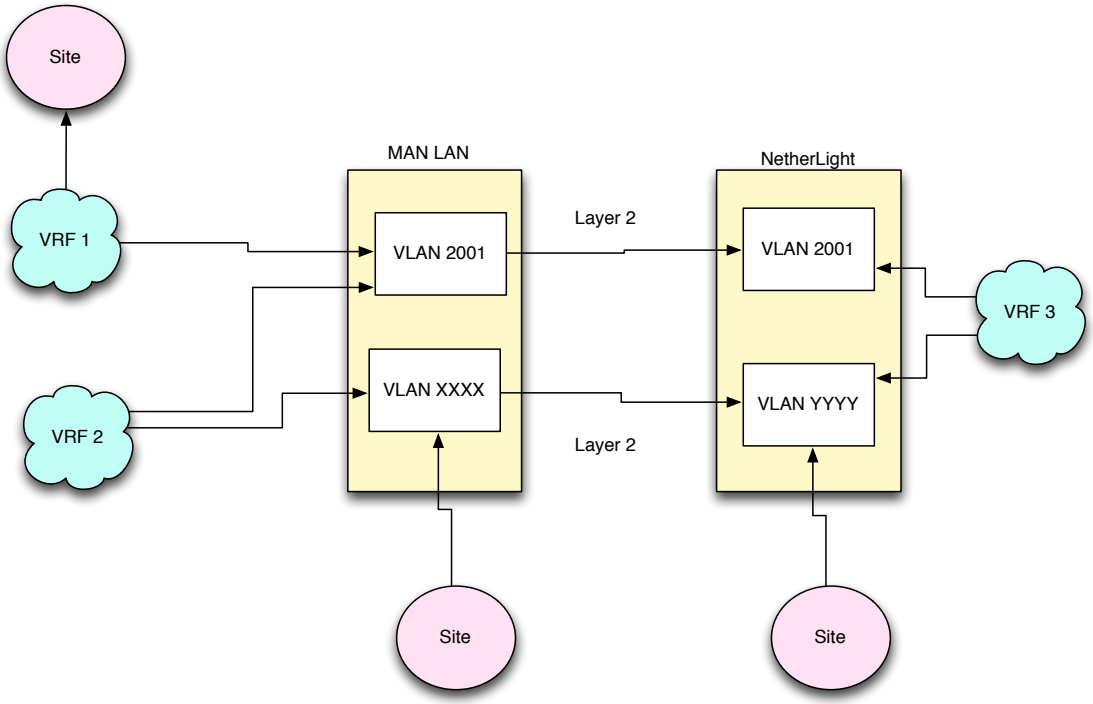

Figure 9: Interconnecting IXPs at Layer 2

Among the questions that arise here are:
1) How do the VRFs connected to the different IXPs establish peering. Are separate VLANs established across the links for each or is some other solution suggested?

Each IXP can decide whether:

- To have a shared VLAN which connects all the VRFs and customers present at the IXP.
- To have p2p vlans between any pair of participants

2) If the site at NetherLight does not have someone willing to do 3rd party routing for them, how do they get connectivity to the LHCONE service.

If there's no patron, an end site should peer with some or all the VRFs present at the IXP, depending on its own needs and anticipated traffic patterns.

An inter-GOLE VLAN comprises a connected broadcast domain between two or more GOLES. Inter-GOLE VLANs require coordination between GOALs to ensure VLAN ID uniqueness. Several LHCONE inter-GOLE VLAN allocation straw-man approaches are offered for consideration.

Approaches

1. **GRAVL Block**
2. **TAG Swapping**
3. **IDC**

GRAVL Block :

Assumptions:

- No IDC or path setup, resource manager or protocol is defined in the LHCONE architecture to work out the VLAN or tag swapping/translation.
- An organization provides a registry service for the GRAVL Block allocations.
- All GOLES will reserve this range of VLAN IDs for LHCONE inter-GOLE paths.
- VLAN IDs not in the GRAVL block are not constrained by LHCONE uniqueness requirements.
- Discrete, Non-unique VLAN IDs may exist within multiple GOLEs as long as they do not belong to the same broadcast domain.

A consistent Global RAnge VLAN (GRAVL) block is reserved exclusively for LHCONE use at each participating GOLE. This set of VLAN ids will subsequently be divided into a dedicated sub-range for each GOAL to allocate as needed. When the GOLEs sub-range is exhausted they request another from the GRAVL Block.

The GRAVL Block is only necessary for inter GOLE VLAN ids that comprise a cohesive broadcast domain between them. Single GOLE, point to point networks should use VLAN ids that adhere to the local GOLE allocation policy.

GRAVL#1 2200-2399	GOLE	VLAN ID RANGE
	GOLE#1	2200-2219
	GOLE#2	2220-2239
	GOLE#3	2240-2259

GRAVL Block allocation

Advantages:

- Conceptually simple.

TAG Swapping:

Assumptions:

- No LHCONE global conventions for VLAN allocation between GOLES.
- VLAN id collisions are worked around using tag swapping aka. VLAN tag translation.
- Manual path setup coordination between GOLES to establish the VLAN tag swapping sequence.

Optical layer transport gear in general has the ability to swap VLAN tags when dropping the circuit at a GOLE.

Advantages:
- No inter-GOLE uniqueness requirements.
- No registration authority.

IDC:
Assumptons:
- Leverage existing IDC technology to manage VLAN id resources during inter-GOLE path setup.
- No LHCONE global conventions for VLAN allocation between GOLES.

Each participating GOLE is required to provision an IDC to manage VLAN id resources for inter-GOLE LHCONE paths.

Advantages:
- Flexible and efficient allocation of VLAN id resources.
- Inefficient pre-allocation of reserved blocks is unnecessary.
- A straightforward production exercise of existing community developed technology.

Resources available:
Overall the set of resources that will be available to this service are represented in this diagram. It is not that critical to understanding the service that all the details of those interconnects be represented.

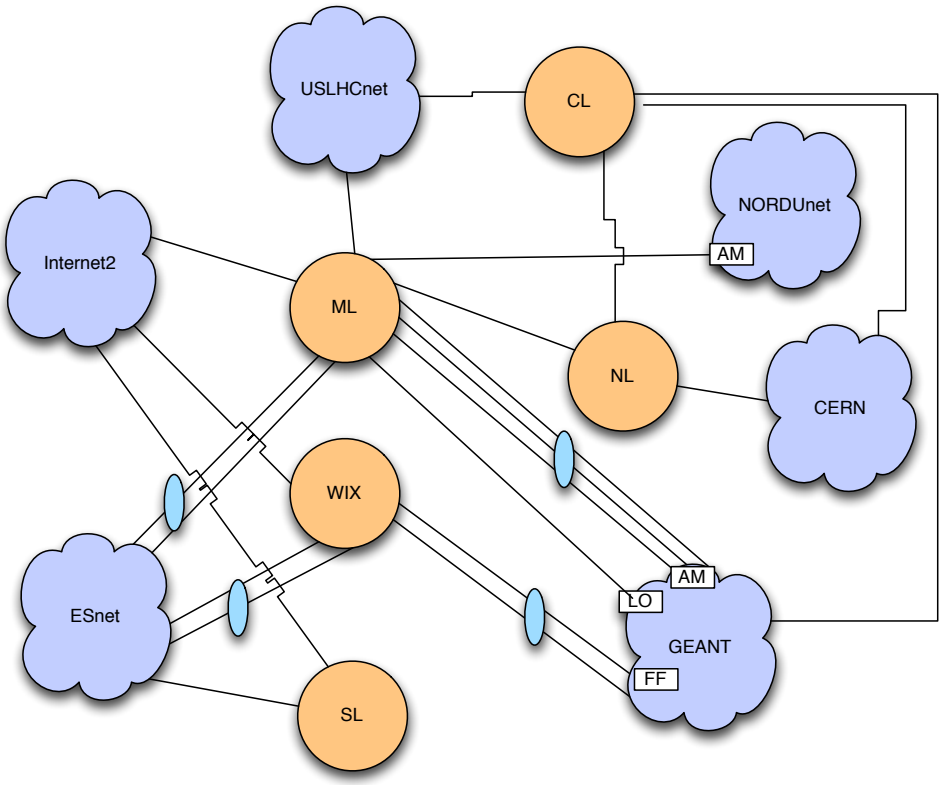

Figure 10: Available Resources

The VRF service will draw upon the set of resources depicted in Figure 10. The BGP peerings and policies will determine the exact design and structure of the network that will be built on these resources as well as the control and data planes.

Figure 9 does not include potential resources for the Point-to-Point service. That would expand the number of links available.

Dynamic Circuits and LHCONE Multipoint Service VRF Implementation
Access to this service will be via Layer 3 or via static or dynamic Layer 2 circuits. Access using dynamic protocols (e.g. IDC, NSI) is strongly encouraged, as that will allow the most flexible evolution of the architecture.

Initially, access via dynamic protocols will need to be via persistent Layer 2 connections, as there is, currently, no way to dynamically create the necessary peerings. Dynamic circuit capabilities to create a path to a port where they can access a participating VRF. Once that path is created a control plane session (BGP peering) will need to be established on the port between the participating site and the VRF.

There are perhaps ways to more fully automate the connectivity between end sites and the VRF instances.

One option is to work with the provider of the VRF instance to preconfigure a number of settings, including:

1) VLAN tag
2) IP address
3) Relevant BGP configuration
4) Prefix information for filters.

The specified VLAN would be used to interconnect the dynamic circuit edge port to the VRF instance. Once the circuit is established, the BGP session will come up and traffic will be able to flow without any further manual intervention. Some consideration and testing of this idea would be needed before it is broadly used. Doing this would require more involvement of the NOC operations staff. It would also require documentation of which peerings might be coming up and going down at intervals – the NOCs are usually given alerts about those actions.

A further refinement of this idea would be to enhance the web interface for provisioning dynamic circuits to include the above information and any other needed information and dynamically build the BGP session. This would be more difficult and longer-term effort.

In both these cases it is important to understand that the dynamic tools are building a circuit to the interface on the VRF edge router that is in the topology file. It is not building a circuit directly to the VRF.

As a part of the configuration that would need to be done during the circuit setup would be to associate the VLAN tag on that circuit with the VRF in the router.

User Access:

User access to this network should not require any special effort on their part. The presumption is that the campuses will in some manner, determined by the nature of their connection to one of these backbones, advertise the set of prefixes that are relevant to the LHC community on their campus. The traffic will be routed along the correct paths. This does require that at some point between the edge of the VRFs and the user there is a diversion of the traffic from the normal R&E paths to the LHC VRF.

Users can only inject packets with source address belonging to the prefixes they announce. VRFs should consider implementing uRPF checks.

Instructions and basic definitions are being developed and located at:

https://twiki.cern.ch/twiki/bin/view/LHCONE/LhcOneHowToConnect

Measurement / Monitoring / Diagnostic Coordination
The proposed LHCONE Diagnostic Service has been detailed in a separate document.