Routing in LHCONE

Sándor Rózsa
Network Engineer
California Institute of Technology
Summer 2011 Joint Techs – Fairbanks, AK

Overview

- What is LHCONE
  - LHCONE services
  - LHCONE multipoint service
  - Pilot topology
- Route servers
- Redundancy
- Routing policies
- Configuration details for the participant (connector) routers
- L2 Security
- Summary & Conclusions

LHCONE

LHC Open Network Environment

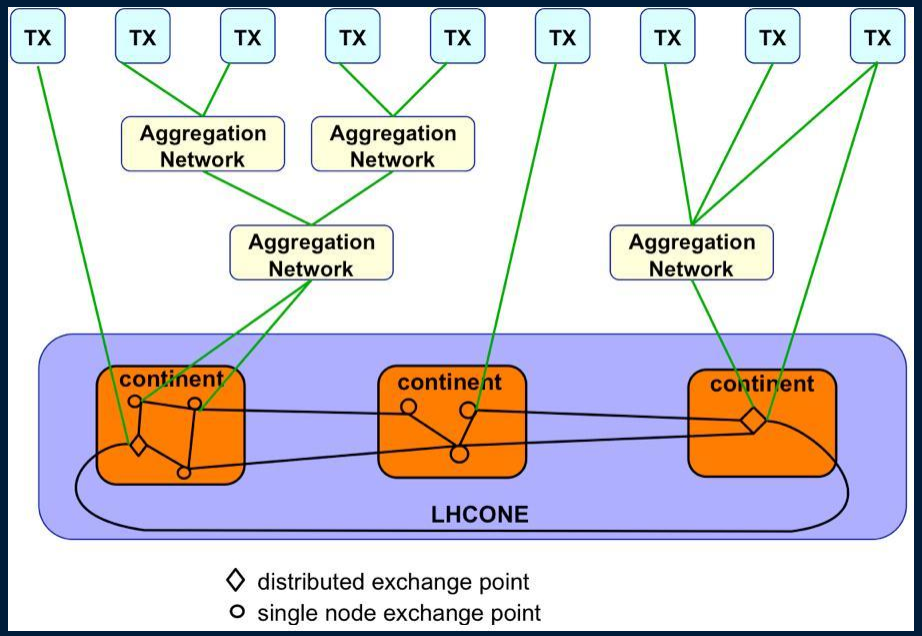
- Provides interconnection between LHC T1/T2/T3 sites
- Provides secure environment for high volume LHC T1-T2, T2-T2 and T2-T3 data transport
- LHCONE is not intended to replace LHCOPN but rather to complement it

There were several workshops/meetings with focus on LHCONE

- June 2010 – Transatlantic Networking for LHC Workshop at CERN
- January 2011 - LHCT2's Technical Meeting at CERN
- February 2011 – LHCT2's Technical Meeting and LHCOPN meeting in Lyon
- June 2011 – LHCONE and LHCOPN joint meeting - Washington

Architecture document

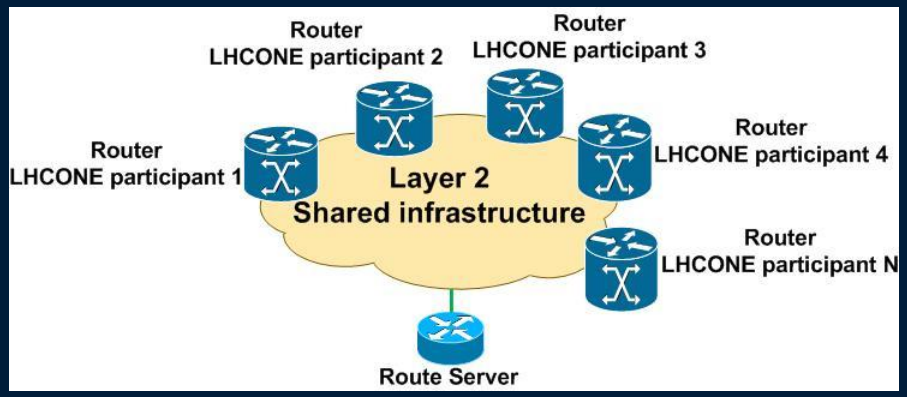
- Was adopted in March 2011



LHCONE services

LHCONE - 4 Services

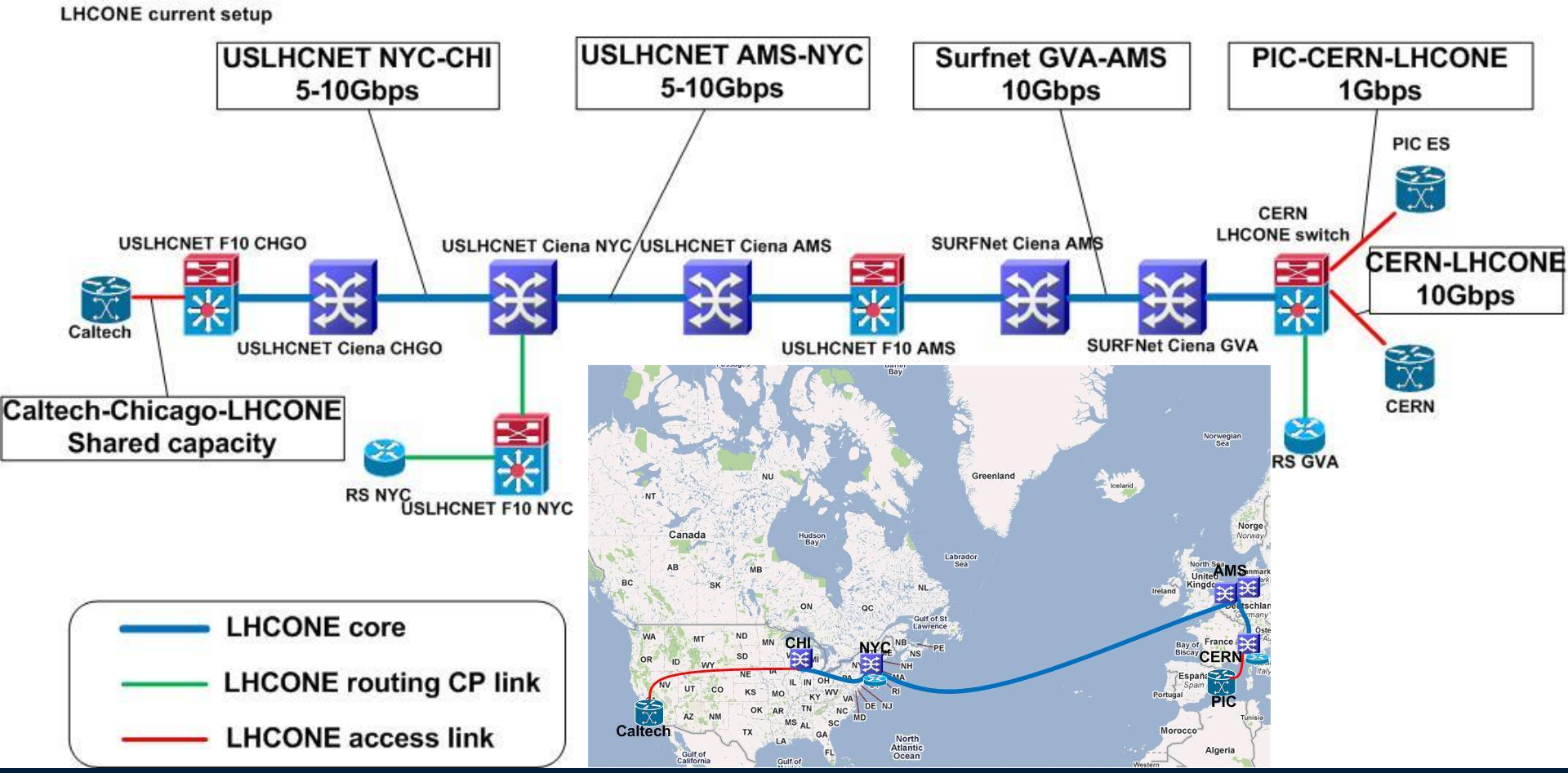
- **Static lightpath (point-to-point) service**
- **Dynamic lightpath service**
- **Multipoint service**
  - Interconnects LHC T1/T2/T3 routers over a shared Layer 2 infrastructure
  - Using route servers for third party route announcements; facilitates configuration
- **Monitoring service**



LHCONE multipoint service

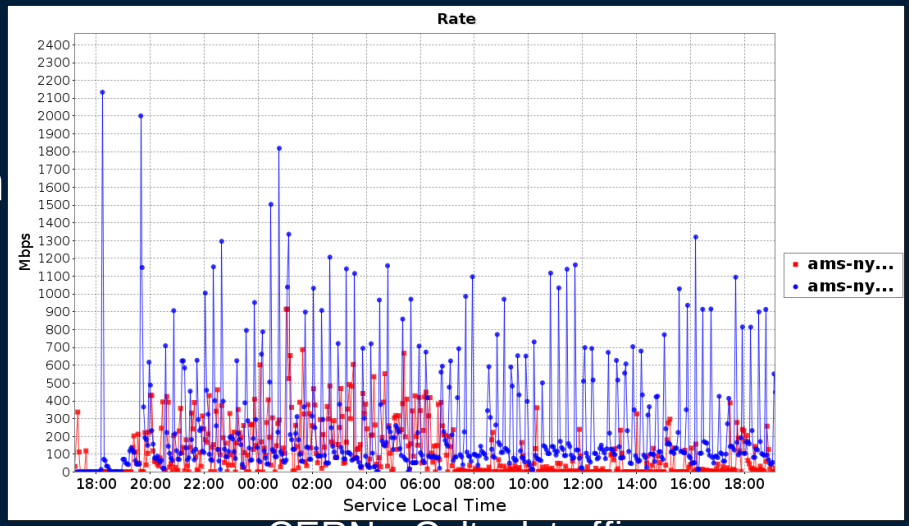
Current setup

- LHCONE pilot today – CERN - Caltech – PIC

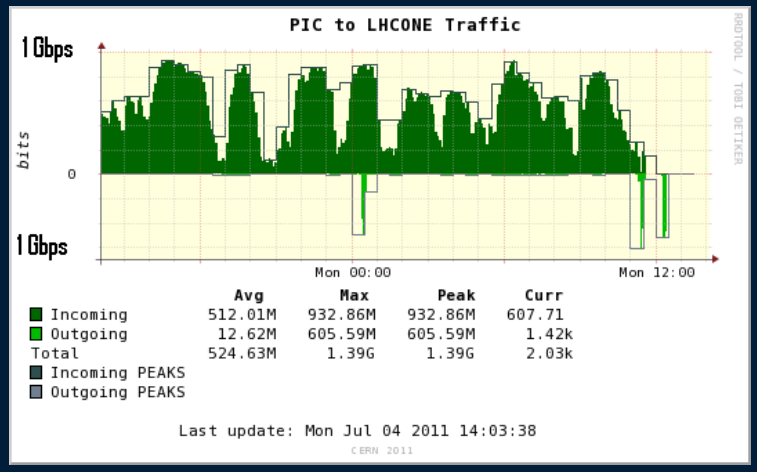


LHCONE multipoint service

- **This is already functional**
  - Pilot participants: CERN, Caltech
  - PIC joined in end of June
- **Both IP connectivity options are available**
  - IPv4: Currently used by LHC sites
  - IPv6: For future use
- **AS20641**
- **IPv4 prefix: 192.16.156.0/22**
- **IPv6 prefix: 2001:7f8:1c:3000::/64**



CERN – Caltech traffic



Traffic from/to PIC (plot source CERN)

Route servers

- It's a BGP daemon
- Peers with the routers of the LHCONE participants
- Centralized network entity providing all the routes available in the network
- The RS does not appear as next-hop for the clients
  - User traffic stream does not pass through the route server
  - Control plane entity
- We tested route servers on the following software/hardware platforms
  - OpenBGPd
  - BIRD
  - Force10

OpenBGPd

Installation

- OpenBSD – www.openbsd.org
- OpenBGPd – www.openbgpd.org

Features

- Easy to install
- Stable
- Multi Threaded
- Single config file for BGP IPv4 and IPv6

Drawbacks

- Extra element in the network

Evaluation results

- **It** is able to handle the full IPv4/IPv6 routing tables
- It was installed on the 31st of March 2011. It is stable since then.

IPv4

```
group "RSv4" {

        announce all
        set nexthop no-modify
        neighbor $PEERCernV4 {
                descr "CERN peering"
                remote-as $ASCern
                tcp md5sig password md5psswd
        }
        neighbor $PEERUltralightV4 {
                descr "Ultralight peering"
                remote-as $ASUltralight
                tcp md5sig password md5psswd
        }
neighbor $PEERUslhcnetv4 {
                descr "USLHCNET peering"
                remote-as $ASUslhcnet
                tcp md5sig password md5psswd
        }
}
```

IPv6

```
group "RSv6" {
        announce all
        set nexthop no-modify
        neighbor $PEERCernV6 {
                descr "CERN peering-IPv6"
                remote-as $ASCern
                tcp md5sig password
          md5psswd
        }
        neighbor $PEERUltralightV6 {
                descr "Ultralight peering-IPv6"
                remote-as $ASUltralight
                tcp md5sig password
          md5psswd
        }
neighbor $PEERUslhcnetv6 {
                descr "USLHCNET peering-
          IPv6"
                remote-as $ASUslhcnet
                tcp md5sig password
          md5psswd
        }

}
```

BIRD

Installation

- On Debian VM installed from the BIRD repository with the command apt-get
- VM 1024MB of RAM, 1 virtual processor
- Separate daemons/config files for IPv4 and IPv6 BGP
- http://bird.network.cz/

Features

- Easy to install
- Stable

Drawbacks

- Extra element in the network
- Single threaded

Evaluation results

- It is able to handle the full IPv4/IPv6 routing tables
- It was installed on the 1st of June 2011. It is stable since then.

IPv4

```
protocol bgp CERN {
        description "CERN v4 peering";
        local as 20641;
        neighbor 192.16.156.10 as 513;
        rs client;
        hold time 240;
        startup hold time 240;
        connect retry time 120;
        keepalive time 80;
        start delay time 5;
        error wait time 60, 300;
        error forget time 300;
        path metric 1;
        default bgp_med 0;
        default bgp_local_pref 0;
        password “MD5Password";
        export all;
}
```

IPv6

```
protocol bgp CERN6{
        description "CERN v6 peering";
        local as 20641;
        neighbor 2001:7f8:1c:3000::513:1 as 513;
        rs client;
        hold time 240;
        startup hold time 240;
        connect retry time 120;
        keepalive time 80;
        start delay time 5;
        error wait time 60, 300;
        error forget time 300;
        path metric 1;
        default bgp_med 0;
        default bgp_local_pref 0;
        source address 2001:7f8:1c:3000:0:2:641:2;
        password “MD5Password";
        export all;
}
```

Force10

Force10 E600 switch/router

- E600 Terrascale
- FTOS version: 8.3.2.0
- The functionality has been achieved by manipulating the community and next-hop attributes for each LHCONE client

Features

- Existing infrastructure can be used
- No extra elements in the network

Drawbacks

- The AS number of the RS stays in the AS path - does not provide all the route server functionalities

E600 config

```
neighbor 192.16.156.10 remote-as 513
 neighbor 192.16.156.10 description LHCONE-F10-RS
 neighbor 192.16.156.10 route-map AS513-IN in
 neighbor 192.16.156.10 route-map LHCONE-RS-OUT out
 neighbor 192.16.156.10 ebgp-multihop 255
 neighbor 192.16.156.10 maximum-prefix 100
 neighbor 192.16.156.10 soft-reconfiguration inbound
 neighbor 192.16.156.10 no shutdown
 neighbor 192.16.157.11 remote-as 32361
 neighbor 192.16.157.11 description LHCONE-F10-RS
 neighbor 192.16.157.11 route-map AS32361-IN in
 neighbor 192.16.157.11 route-map LHCONE-RS-OUT out
 neighbor 192.16.157.11 ebgp-multihop 255
 neighbor 192.16.157.11 maximum-prefix 100
 neighbor 192.16.157.11 soft-reconfiguration inbound
 neighbor 192.16.157.11 no shutdown
route-map AS513-IN permit 5
 set community 513:20641
route-map AS32361-IN permit 5
  set community 32361:20641
route-map LHCONE-RS-OUT permit 10
 match community comm-caltech-lhcone
 set next-hop 192.16.157.11
route-map LHCONE-RS-OUT permit 10
 match community comm-cern-lhcone
 set next-hop 192.16.156.10
```

RS conclusions

- **OpenBGPd and BIRD provide all the necessary features**
- **Force10 - Does not provide all the route server functionalities**
  - The AS number of the RS stays in the AS path
- **Our choice would be BIRD**
  - Can be easily installed on XEN virtualized servers

Redundancy in LHCONE multipoint service

The following issues should be addressed by redundancy

- **Participant's link failure**
  - Participant can use multiple links to the LHCONE core
- **Route server failure**
  - Participants peer with multiple route servers
  - Multiple route servers address this issue
- **Inter-regional link failure**
  - Multiple regional inter-regional links
    - No STP
    - LAG can be used
    - Currently Layer 1 protection provided by USLHCNET
    - We foresee Layer2 multipath (SPB, TRILL) once it is standardized and available
  - If all the links fail between two regions
    - LHCONE can be partially functional due to multiple geographically dispersed route servers

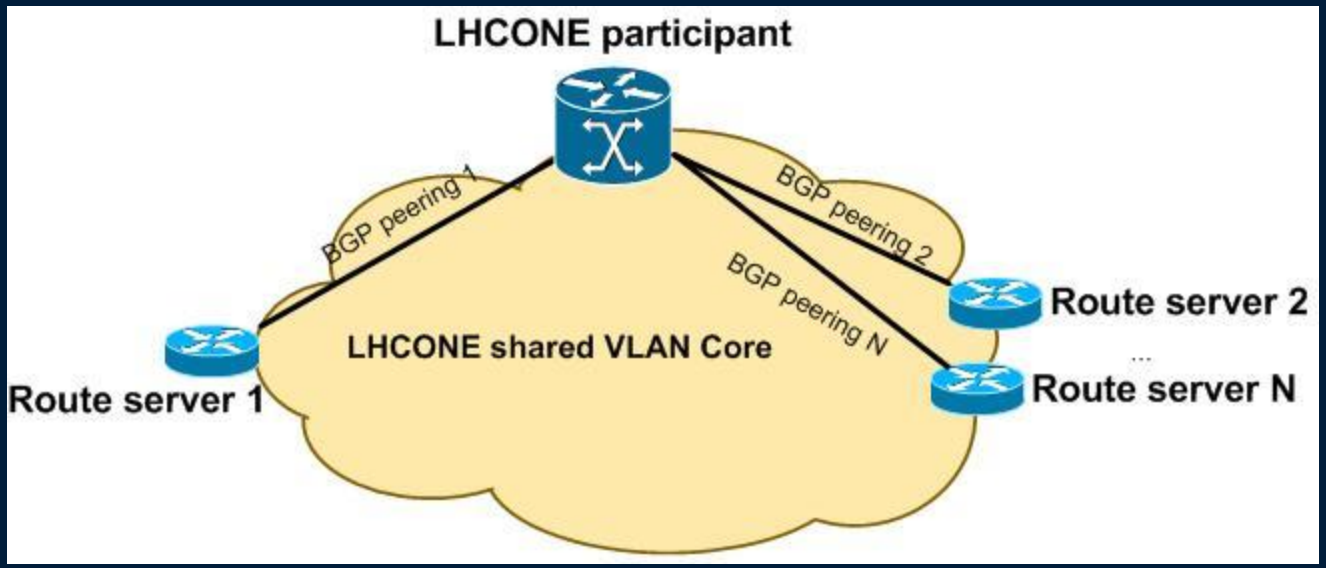
Redundancy
Route server failure

Components might fail

All the route servers provide the same routing information

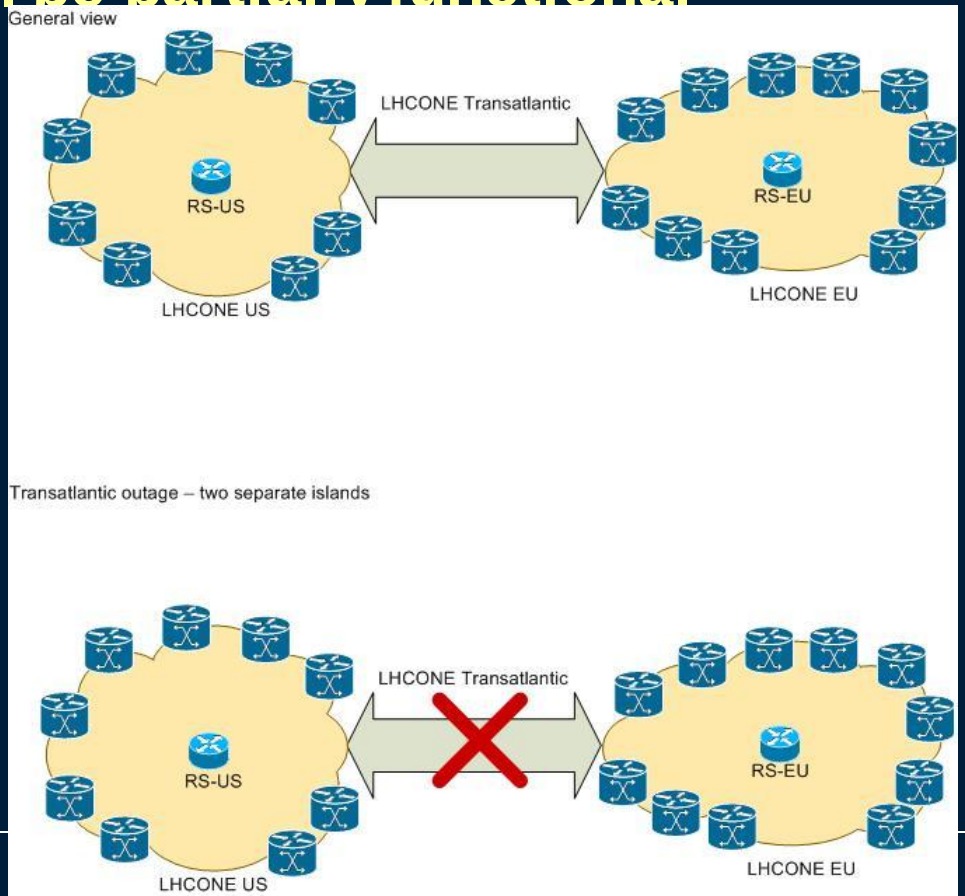
LHCONE participant's perspective

- Participant's router peers with all available route servers
- Participant's router receives the same prefixes from all the route servers
- If one RS fails the participant will have access to the routing prefixes via an alternate RS.



Redundancy
Link failure in the core

- Major (E.g. Transatlantic) outage – loss of connectivity between regions
- Due to the geographically dispersed route servers the network will be partially functional



Routing policies

General policy

- Participants announce all (and only) their LHC related prefixes. Currently there is no restriction on the prefix length.
- Participants will receive the whole (and only) LHCONE routing table from the route servers.

How should this be configured on the participant routers?

- We have tested several routing platforms – config details are publicly available on the www.lhcone.net

LHCONE participant configuration details - Brocade

Thanks to CERN for sharing this config with us

```
router bgp
 local-as 513
 neighbor 192.16.156.1 remote-as 20641
 neighbor 192.16.156.1 description "----> LHCONE
       Route Server EU"
 neighbor 192.16.156.1 ebgp-multihop 255
 neighbor 192.16.156.1 password
       MD5PASSWORD
 neighbor 192.16.156.1 soft-reconfiguration
       inbound
 neighbor 2001:7f8:1c:3000:0:2:641:1 remote-as
       20641
 neighbor 2001:7f8:1c:3000:0:2:641:1 description
       "----> LHCONE Route Server EU"
 neighbor 2001:7f8:1c:3000:0:2:641:1 ebgp-
       multihop 255
 neighbor 2001:7f8:1c:3000:0:2:641:1 password
       MD5PASSWORD

 address-family ipv4 unicast
  neighbor 192.16.156.1 maximum-prefix 100
  neighbor 192.16.156.1 route-map in LHCONE-IN
  neighbor 192.16.156.1 route-map out LHCONE-
       OUT
  neighbor 192.16.156.1 send-community
  no neighbor 2001:7f8:1c:3000:0:2:641:1 activate
 exit-address-family

 address-family ipv6 unicast
 neighbor 2001:7f8:1c:3000:0:2:641:1 activate
 neighbor 2001:7f8:1c:3000:0:2:641:1 maximum-
       prefix 100
 neighbor 2001:7f8:1c:3000:0:2:641:1 route-map in
       LHCONE-IN
 neighbor 2001:7f8:1c:3000:0:2:641:1 route-map
       out LHCONE-OUT
 neighbor 2001:7f8:1c:3000:0:2:641:1 send-
       community
 exit-address-family
!
```

LHCONE participant configuration details – Cisco

```
no bgp enforce-first-as
 neighbor 192.16.156.1 remote-as 20641
 neighbor 192.16.156.1 description ---> LHCONE
       RS EU<---
 neighbor 192.16.156.1 shutdown
 neighbor 192.16.156.1 ebgp-multihop 255
 neighbor 192.16.156.1 password 7 MD5Password
 neighbor 192.16.156.1 update-source Vlan3000
 neighbor 192.16.156.1 version 4

 neighbor 2001:7F8:1C:3000:0:2:641:1 remote-as
       20641
 neighbor 2001:7F8:1C:3000:0:2:641:1 description
       ---> LHCONE-IPv6 RS EU <---
 neighbor 2001:7F8:1C:3000:0:2:641:1 shutdown
 neighbor 2001:7F8:1C:3000:0:2:641:1 ebgp-
       multihop 255
 neighbor 2001:7F8:1C:3000:0:2:641:1 password 7
       MD5Password
 neighbor 2001:7F8:1C:3000:0:2:641:1 update-
       source Vlan3000
 neighbor 2001:7F8:1C:3000:0:2:641:1 version 4
```

```
address-family ipv4
  no neighbor 2001:7F8:1C:3000:0:2:641:1 activate
  neighbor 192.16.156.1 activate
  neighbor 192.16.156.1 soft-reconfiguration
       inbound
  neighbor 192.16.156.1 route-map LHCONE-IN in
  neighbor 192.16.156.1 route-map LHCONE-OUT
       out
 exit

 address-family ipv6
 neighbor 2001:7F8:1C:3000:0:2:641:1 activate
 neighbor 2001:7F8:1C:3000:0:2:641:1 soft-
       reconfiguration inbound
 neighbor 2001:7F8:1C:3000:0:2:641:1 route-map
       LHCONE-IN in
 neighbor 2001:7F8:1C:3000:0:2:641:1 route-map
       LHCONE-OUT out
```

LHCONE participant configuration details – Force10

```
no bgp enforce-first-as
 neighbor 192.16.156.1 description ---> LHCONE RS EU <---
 neighbor 192.16.156.1 route-map LHCONE-IN in
 neighbor 192.16.156.1 route-map LHCONE-OUT out
 neighbor 192.16.156.1 ebgp-multihop 255
 neighbor 192.16.156.1 maximum-prefix 100
 neighbor 192.16.156.1 send-community
 neighbor 192.16.156.1 password 7 MD5Password
 neighbor 192.16.156.1 soft-reconfiguration inbound
 neighbor 192.16.156.1 no shutdown
 neighbor 2001:7f8:1c:3000:0:2:641:1 remote-as 20641
 no neighbor 2001:7f8:1c:3000:0:2:641:1 activate
 neighbor 2001:7f8:1c:3000:0:2:641:1 description ---> LHCONE-IPv6 RS EU<---
 neighbor 2001:7f8:1c:3000:0:2:641:1 ebgp-multihop 255
 neighbor 2001:7f8:1c:3000:0:2:641:1 send-community
 neighbor 2001:7f8:1c:3000:0:2:641:1 password 7 MD5Password
address-family ipv6 unicast
 neighbor 2001:7f8:1c:3000:0:2:641:1 activate
 neighbor 2001:7f8:1c:3000:0:2:641:1 route-map LHCONE-IPV6-IN in
 neighbor 2001:7f8:1c:3000:0:2:641:1 route-map LHCONE-IPV6-OUT out
exit-address-family
```

Layer 2 Security

Layer 2 security threats

- Broadcast storm
  - Broadcast storm protection
- Layer 2 loops
  - Loop protection
- Unauthorized devices
  - Only declared MAC addresses allowed

We also need

- Max number of allowed MAC addresses should be limited
- Quarantine VLAN for newcomers

Currently configured Layer 2 security features

Shared VLAN configuration details

```
interface Vlan 3000
 mtu 9252
```

Access interface configuration details

```
interface TenGigabitEthernet 0/1
 switchport
 mtu 9252
 mac learning-limit 2 station-move
 storm-control broadcast 1 in
 storm-control unknown-unicast 1 in
```

Summary & conclusions

- **LHCONE multipoint service is operational**
  - There are three pilot participants: CERN, Caltech and PIC
  - Inside the multipoint domain the routing is managed by the route servers
  - Multiple route servers provide redundancy in case of failures
    - Route server failure
    - Major link failure in the core
- **Several route servers were tested – OpenBGPd, BIRD, Force10**
- **Configuration details were documented and are available for the participants (connectors)**
- **Layer 2 security threats were identified – workaround was proposed**

Questions

- www.lhcone.net