06-09-2021

White Box: GRNET Data Centre Use Case

Grant Agreement No.:	856726
Work Package	WP6
Task Item:	Task 1
Dissemination Level:	PU (Public)
Document ID:	GN43-456-300

Authors: Theodore Vasilopoulos (GRNET, Lefteris Poulakakis (GRNET), Xavier Jeannin (RENATER), Ivana Golub (PSNC), Tim Chown (Jisc)

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 856726 (GN4-3).

Abstract

This document presents how white box can be used as a switch or a router to implement a data centre network in Greece by the Greek NREN GRNET, providing the results of the first implementation in production.

Table of Contents

Executive Summary 1

1 Introduction 2

2 Use Case Description 3
- 2.1 High-Level OAV Approach 3
- 2.2 Performance and Features Requirements 5
- 2.3 Reliability / Redundancy Requirements 6
- 2.4 Cost Requirements 6

3 Solution 7
- 3.1 Maintenance 8
- 3.2 White Box Switch Validation 8
- 3.3 Leaf, Spine, Data Centre Gateway 9
- 3.4 Provided Services 9
- 3.5 Automation 11

4 Results in a Production Network 12
- 4.1 Specific Features Validation 12
  - 4.1.1 MC-LAG failure scenarios 13
  - 4.1.2 MAC Mobility 16
- 4.2 East-West Traffic Performance Validation 18

5 Conclusions 19

References 20

Glossary 21

Table of Figures

Figure 1.1: Global data centre topology - spine-leaf folded Clos topology 4

Figure 3.1: Spine switch, Edgecore AS7712-32X 7

Figure 3.2: Leaf switch, Edgecore AS5812-54T 8

Figure 3.3: Data centre use case topology 10

Figure 4.1: Multi-Chassis LAG (MC-LAG) with peer link between leaf01 & leaf02 and dual-homed server01 13
Figure 4.2: MC-LAG uplink failure scenario 14
Figure 4.3: Test results 15
Figure 4.4: MAC mobility of a VM from server_rack07 to server_rack05 16

Executive Summary

This document describes a white box technology use case for a small-scale data centre deployment for GRNET, the Greek National Research and Education Network operator. It shows how a white box switch can be used in a data centre fabric, including requirements, design, and results obtained from the devices installed in a production network. The motivation for a solution based on white box was to have better vendor independence, more flexibility (the Network Operating System (NOS) and the hardware can be changed independently), cost savings, and being able to use open-source solutions.

The data centre design is based on an IP Clos EVPN/VxLAN fabric and on a spine-leaf folded Clos topology. The network devices selected to implement this design are the Edgecore AS7712-32X (equipped with the chipset forwarding Tomahawk 3.2Tbps) for the spines and the Edgecore AS5812-54T (equipped with the chipset forwarding Broadcom Trident II+ 720Gbps) for the leaves. The Network Operating System (NOS) chosen was Cumulus Linux, recently acquired by NVIDIA.

As this company also owns the chipset supplier, Mellanox, the strategy of NVIDIA is to provide some Cumulus features only on the Mellanox chipset (for instance the EVPN Multihoming feature). The fallback was to use MC-LAG to implement active-active server connectivity and some control plane redundancy features. Moreover, from release 4.4 (July 2021), Cumulus Linux will support only NVIDIA Spectrum-based Application-Specific Integrated Circuit (ASIC) platforms [CMS], putting the white box based on Broadcom ASICs at its end of life within the next few years. However, since the software and hardware are decoupled in a white box, GRNET can either keep white box switch hardware and try another NOS (e.g. Pluribus), or try out white box switches with cheap NVIDIA chipsets and keep the switches based on Broadcom for another use case.

The network devices were implemented for production but, for now, only a few servers were deployed in the data centre to validate the design. The deployment of the other servers for real production will be done later. The solution successfully passed all the validation tests (feature and performance tests). This is the beginning of this small data centre that will gradually gain momentum as new computing requests emerge from GRNET's users and customers.

1 Introduction

This report focuses on a data centre use case implementation for a white box platform. For more details on white box (WB) use in R&E scenarios, see the review conducted by the Network Technology Evaluation task (WP6 T1) in the GN4-3 project on the applicability of new types of white box devices in such use cases [D6.3].

Traditionally, the data centre network was a combination of hardware and software components provided by the same vendor. The white box switch introduces the concept of decoupling software and hardware components. As a result, a customer can choose their own combination from a variety of Network Operating Systems (NOSs) and commodity hardware solutions.

The white box switches have significant advantages such as:

- Freedom of choice (vendor independence)
- Flexibility –option to replace either the NOS without changing the hardware and vice versa;"
- Cost savings
- Being able to use open-source solutions

Over the past few years, the switching market has made significant progress in producing powerful ASICs in terms of throughput, buffer size, programmability, and features. Many traditional vendors also use these commodity ASICs to build their branded hardware. Therefore, the white box networking concept would be of significant interest to NRENs and R&E networks.

In this context, GRNET wanted to assess the use of white box network devices for its data centre implementation. After a preliminary study, GRNET built an implementation of the first phase of its data centre based on white boxes. The use case description and requirements are presented in Section 2, the solution implementation in Section 3, and the prototype production results are commented on in Section 4.

2 Use Case Description

2.1 High-Level OAV Approach

GRNET is already operating three medium-to-large scale data centres in three different Points of Presence (PoPs) based on an IP Clos EVPN/VxLAN fabric [FABRIC_EVPN]. This project aims to implement a new data centre using the same architecture as deployed in other medium-to-large scale data centres but based on white box and non-hardware vendor NOS. The data centre network provides L2/L3 connectivity services for:

- Cloud resources of the public academic community
- Cloud resources of the public sector
- GRNET private cloud infrastructure (internal company resources)
- Server colocation services for academic and public sectors

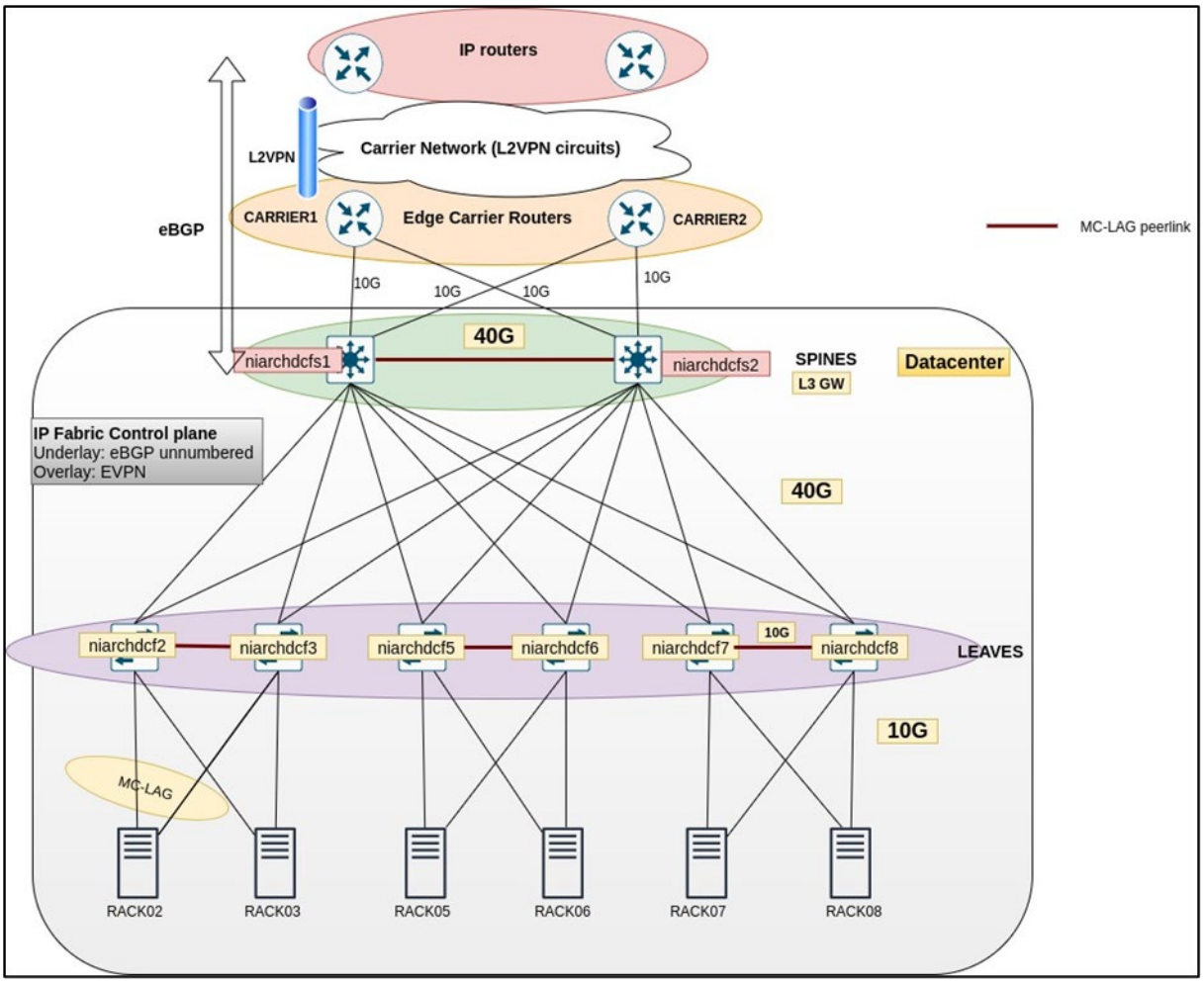

Figure 1.1: Global data centre topology - spine-leaf folded Clos topology

The data centre use case aims to provide a new small-scale data centre for GRNET. Figure 1.1 shows the spine-leaf folded Clos topology model used in this data centre. The protocols IP and eBGP unnumbered implement an underlay L3 network between leaves and spines. The BGP unnumbered standard, specified in RFC 5549, uses extended next hop encoding (ENHE) and no longer requires an IPv4 prefix to be advertised along with an IPv4 next hop. This means that it is possible to set up BGP peering between whitebox switches, and exchange IPv4 prefixes without having to configure an IPv4 address on each switch; the interfaces that BGP uses are unnumbered. The next hop address for each prefix is an IPv6 link-local address, which is assigned automatically to each interface. Using the IPv6 link-local address as a next hop instead of an IPv4 unicast address, BGP unnumbered means there is no requirement to configure IPv4 addresses on each interface [BGP]. VxLAN tunnels are deployed between leaves to provide Layer 2 overlay data plane connectivity. EVPN based on eBGP manages the MAC address learning and exchange. The protocol eBGP provides an “all-active links” backbone topology as well as routing information that carries traffic across the data centre.

Each Top-of-Rack (ToR) switch has 10G ports for server connections and multiple 40 Gbps ports for the uplinks. Using the eBGP ECMP feature, each TOR is connected to a couple of core/spine switches which also act as data centre routers providing L3 services. MC-LAG, Multichassis link aggregation, enables a client device to form a logical LAG interface between two network devices, two MC-LAG

peers. An MC-LAG provides a loop-free at layer 2 network, redundancy, and load balancing between the two MC-LAG peers. On the server side, with MC-LAG implementation, each server is connected to two different TOR switches providing an active-active setup for better redundancy and system throughput.

The objective of the project is the deployment of a data centre network that can take advantage of the above protocols and features, with a decentralised control plane and enough bandwidth capacity to fit user needs. It is also preferable to avoid vendor proprietary or non-standardised technologies to avoid vendor lock in.

2.2 Performance and Features Requirements

To fulfil the project requirements, the following set of features were identified as necessary capabilities for the deployment:

- RIB capacity of 16k/8k routes (IPv4/IPv6)
- MAC address table capacity 32k addresses
- Jumbo frames support (9022 bytes Ethernet MTU, 9000 bytes IP MTU)
- Switching/Forwarding:
  - Switching/forwarding capacity of:
    - >6.4Tbps and >2Bpps (spines switches)
    - >1.4Tbps and >1Bpps (leaf switches)
  - Buffer size >12Mbyte
- Routing in/out of VxLAN tunnel (RIOT capability)
- Bandwidth connection, number of switch ports:
  - >=32 40GBASE and/or 100GBASE QSFP28 capable ports with 4x10G or 2x50G or 4x55G breakout option (spine switches)
  - >=48 RJ-45 10GBASE-T/1000BASE-T ports (leaf switches)
  - >=6 optical QSFP 40GBASE ports (leaf switches)
- Support all basic management plane protocols (SSH, SNMPv2, NTP, etc)
- Support a scripting language (Python, Perl), at least one automation tool (Ansible/Puppet), and configuration rollback capability
- Control and data plane features - match current production data centre capabilities:
  - BGP, BFD, ECMP
  - EVPN
  - MAC Mobility
  - MC-LAG load balancing over dual-homed servers
  - VXLAN
  - LACP
  - LLDP
  - ACLs (L3 and L2)

These requirements were used for the request for proposal that was published to set-up this data centre.

2.3 Reliability / Redundancy Requirements

The following resilience features were included:

- Control plane redundancy is provided by the MC-LAG implementation on ToR switches and EVPN Virtual Gateway redundancy for the L3 gateway spine switches.
- To enable redundancy for data centre connection to the Internet, two point-to-point Layer 2 circuits are implemented between each spine and the IP routers. Two eBGP peerings are established between each spine and the IP routers.
- Hot-swappable power supplies and fan units are required.

2.4 Cost Requirements

The solution must not exceed the cost of the budget required for the previous network data centre implementation using traditional vendors. The cost assessment was made using the TCO calculator [TCO] previously published by the GN4-3 WP6 T1 white box team. The cost of the edgecore AS-7712 and the Juniper QFX-10k-36Q were compared. It is not possible to include the cost computation in this document for legal reasons. However, the white box solution was about 10% cheaper at the time.

3 Solution

GRNET initiated the procurement procedure for eight white box switches for a small data centre. The requirement was to build an EVPN/VXLAN-based data centre network and match the services already provided in the other GRNET production data centres based on traditional vendor devices.

The Request For Proposal (RFP) procedure resulted in the procurement of two AS7712-32X Edgecore switches, capable of acting as spines, and six AS5812-54T Edgecore switches, capable of acting as leaves.

Spine - Edgecore AS7712-32X (see Figure 3.1):

- 32-Port 100G QSFP28
- ASIC Broadcom Tomahawk 3.2Tbps
- Intel Atom® C2538 CPU
- ONIE software installer
- dual 110-230VAC 650W PSUs
- 6 Type C Fan Modules with power-to-port airflow

Figure 3.1: Spine switch, Edgecore AS7712-32X

Leaf - Edgecore AS5812-54T (see Figure 3.2):

- 48-Port 10GBASE-T with 6x40G QSFP+ uplinks
- ASIC Broadcom Trident II+ 720Gbps
- Intel Atom® C2538 CPU
- ONIE software installer
- dual 110-230VAC 400W PSUs
- 5 Type D Fan Modules with power-to-port airflow

Figure 3.2: Leaf switch, Edgecore AS5812-54T

The Network OS chosen is Cumulus Linux 4.3 as it is widely used in data centre solutions, and based on open-source components for switchport management (ifupdown2) and routing software (Free Range Routing (FRR) [FRR]).

3.1 Maintenance

GRNET signed a contract with the local integrator who agreed to be a single point of contact for both hardware and software support. For any case that needs a hardware replacement the integrator is responsible for the replacement and the shipment of the hardware.

For the software part, the integrator can act as a relay to Cumulus for our requests. In reality, GRNET can directly open technical cases to Cumulus and had already opened a case (for a really minor issue), which was managed and solved appropriately.

3.2 White Box Switch Validation

The basic concept of validation is based on feature tests. First, the individual L2/L3 switch features related to configuration and management are considered. The second aspect of the validation is related to the data centre IP CLOS leaf-spine topology environment. As a result, the main aspects of validation are:

- Configuration and management (Linux and industry-standard model CLI, NCLU, SSH configuration in a Cumulus Linux NOS)
- Programmability and automation (Netconf, Ansible)
- Underlay and overlay network configuration: BGP, BFD, ECMP, EVPN, MAC Mobility, MC-LAG load balancing on a dual-homed server, VXLAN, LACP, LLDP, ACLs (L3 and L2).
- Fundamental scenario tests (intra-inter VLANn routing, MAC mobility, Load Balancing)
- Performance validation (throughput test, VM mobility convergence time)

Regarding performance tests, the following tests are conducted:

- Throughput capacity
- Latency
- VM mobility convergence time

3.3 Leaf, Spine, Data Centre Gateway

An IP CLOS Leaf-Spine physical topology is used. Each leaf switch is interconnected with the two spines of the fabric.

An EVPN-based fabric [FABRIC_EVPN] is based on an MP-BGP overlay - usually eBGP - to propagate EVPN family advertisements. Usually, an underlay IGP or point-to-point eBGP between all the adjacent devices of the fabric is used to advertise loopback addresses.

Overlay and underlay BGP configuration can be a complex and a massive part of the network configuration. Network automation is intensively used to realise this. Cumulus Linux supports BGP auto-configuration (BGP unnumbered), eliminating the need for massive configuration, address management, and complex automation abstraction.

The configuration of the data centre router is based on the Configuration Examples page on the NVIDIA website [CMS-EVPN-CFG].

3.4 Provided Services

The EVPN multihoming feature is supported by Cumulus Linux from 4.2 as a standards-based replacement for MC-LAG in data centres deploying Clos topologies. Cumulus supports EVPN multihoming only on Spectrum ASIC-based switches [EVPN-2]. Due to the fact that leaf switches are equipped with the Broadcom Chipset, the alternative was an MC-LAG based solution to achieve an active-active multihoming setup on the server side.

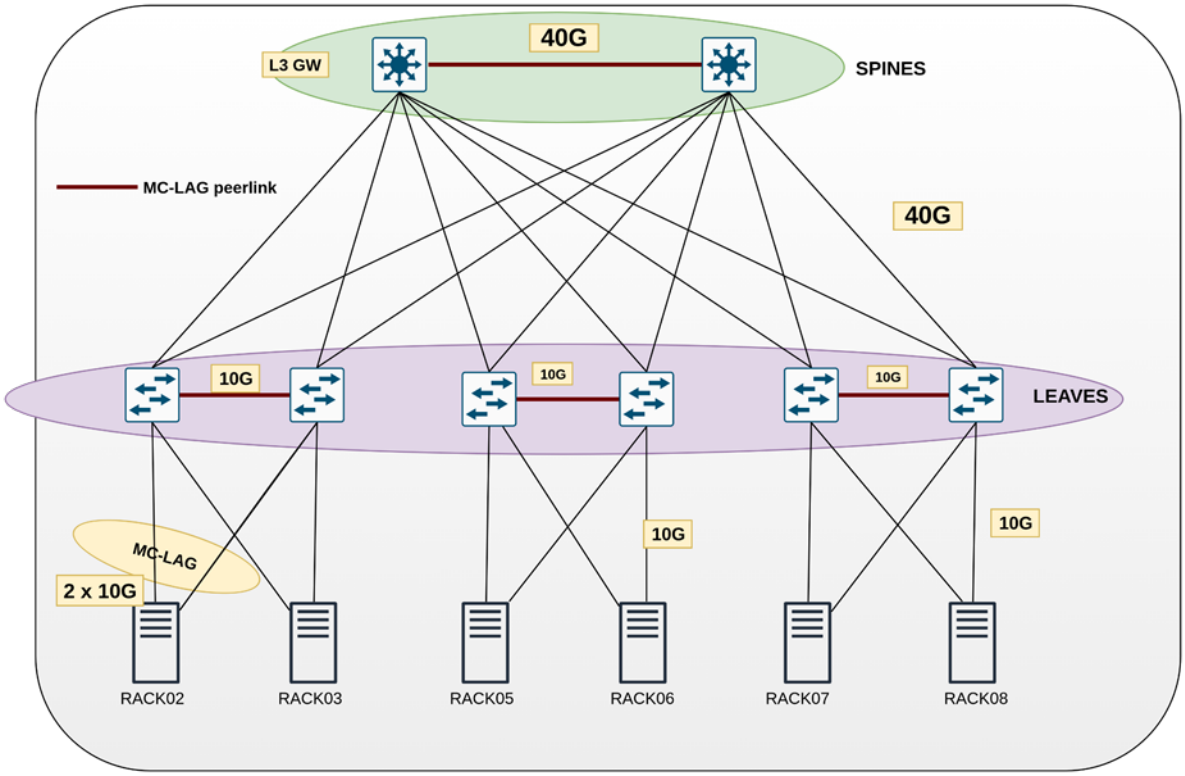

Figure 3.3: Data centre use case topology

The diagram in Figure 3.3 shows the Bare Metal Servers connecting to the fabric with an LACP bond with two of the fabric's leaves. Each pair of leaves that host the same servers create an MC-LAG peering through the peer-link between them. From the hosts' point of view, each of the links in its bond is connected to the same system, and so the host uses both links. This way a redundant pair of ToR switches is created, offering failure tolerance and increased bandwidth due to the active-active setup (load sharing).

Plain L2 bridging, without any routing support, is one of the services to be validated. The EVPN fabric must be able to act as a plain L2 switch, enabling private LAN services.

The centralised routing setup is similar to the existing data centre setup for GRNET. In a centralised setup, two devices (spines) are assigned as the L3 gateways of the fabric. This approach simplifies the configuration to only two devices but can cause additional east-west traffic in the data centre, especially on a large scale. However, a distributed architecture brings L3 routing down to Leaf switches, which simplifies traffic flow but adds configuration load and forces the leaves to support VxLAN L3 routing functionality. A distributed architecture defines two models for L3 routing: symmetric and asymmetric. The symmetric model routes and bridges on both the ingress and the egress leaves. The asymmetric model allows routing and bridging on the VXLAN tunnel ingress (ingress leaves), but only bridging on the egress leaves [EVPN].

The design choice for the data centre was to use the centralised setup and enable L3 routing on the spines. This choice is based on the requirement but also on simplicity and familiarity with the setup. Cumulus Linux supports the different flavours of decentralised setup, and this is considered to be tested and validated as part of the future work.

3.5 Automation

An advantage of white box solutions, especially Linux-based, is the variety of tools developed for Linux server administration that can be natively used for switch management. Especially in a data centre case, this may be very useful, as the same tool can manage both network elements and bare-metal servers' network configuration.

Additionally, a white box device can support network automation solutions (Puppet, Saltstack, etc.), allowing a good level of scalability.

Taking into consideration the tools used by GRNET for its own network (Ansible) and server (Puppet) management, each having a quite extensive code base, the decision was made to use Ansible for the Cumulus Linux OS management.

Even though this option has some disadvantages, it was judged as crucial to use the same toolset already used for the rest of the GRNET network The benefits of this option are:

- Ansible is running in a remote connection mode without a local initiated NETCONF session.
- No external module installation and maintenance are required, since the white box switch is managed as a native Linux server where core Ansible modules are available.

4 Results in a Production Network

This section presents the performance and feature tests that were conducted to validate the white box data centre for production.

4.1 Specific Features Validation

The following features have been tested and validated:

- eBGP neighbourship (BGP unnumbered feature - RFC 5549)
- eBGP peering with external router
- eBGP ECMP routes
- EVPN neighbourship
  - Virtual Tunnel End-Point loopback announcement throughout the fabric.
- EVPN remote MAC-learning from remote hosts on same VLAN
- Intra/Inter -VLAN connectivity
- DHCP relay towards external DHCP server
- Virtual Router Redundancy (VRR)
- The 2 spines provide L3 gateway redundancy
- Bond interface configuration
- LACP bypass [lLACP-BY]
  - LACP bypass helps during the OS server installation, allowing the remote OS to be downloaded (PXE process) through one link of the LAG declared on the leaf but not configured on the server as the OS is not yet installed.
- MC-LAG peering
- MC-LAG Load Balancing on host connected to two different leaves
- MC-LAG failure scenarios
  - The detail is presented in Section 4.1.1 below.
- MAC mobility (RFC7432)
  - The feature allows a host or end-station (as defined by its MAC address) to move from one Ethernet segment to another (more details are given below).
- ER-SPAN
  - This enables the mirrored packets to be sent to a monitoring node located anywhere across the routed network.

- NCLU (Network Command Line Utility)
  - This provides a CLI to allow the configuration of the switch to be changed. NCLU was found convenient to manage the configuration but it appears to be more reliable to directly change the configuration files either manually or through Ansible. Moreover, it has been observed that NCLU 'net show' commands sometimes give irrelevant and misleading output in comparison with configuration files. The native Linux commands always display the real configuration status.

Cumulus Linux is based on a daemon at its heart, 'switchd'. It communicates between the switch and Cumulus NOS, and all the applications running provided with Cumulus Linux. Fundamental changes to the configuration require a restart of the switchd daemon/service. It has been observed that sometimes after the restart, switchd could not restart correctly. In that case, it is possible to manually restart the daemon at Linux level via a console connection. The console connection is mandatory because the hardware can be rebooted remotely without manual intervention.

A final comment about the troubleshooting process, during validation: some features did not work properly (MC LAG for instance) and nothing was found in the log file. The team was not very experienced in Cumulus and, despite the documentation, it was difficult to identify the root cause of the issue. This left the team with only the option to reboot the switch.

4.1.1 MC-LAG failure scenarios

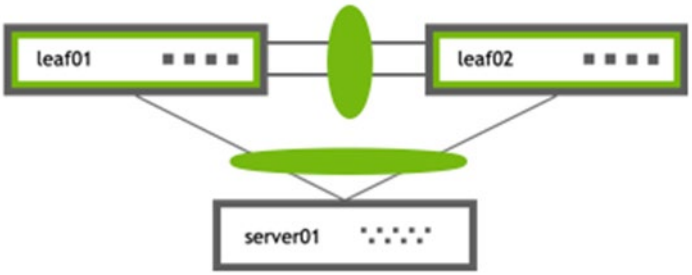

Figure 4.1: Multi-Chassis LAG (MC-LAG) with peer link between leaf01 & leaf02 and dual-homed server01

- **Primary peer failure**

  When the peering relationship is established between a pair of TOR switches, one switch is granted the primary role and the other the secondary role. When a MLAG-enabled switch is in the secondary role, it does not send STP (Spanning Tree Protocol) BPDUs on dual-connected links between the leaf and the server.

  When the primary peer reboots or shuts down, the secondary peer realises that the primary is in 'not active' state ('clagd-backup-ip' is not active) and transitions to the primary role. When the original primary switch is powered back on, it resumes the primary role (pre-emptive functionality).

- **Secondary peer failure**

  When a secondary peer reboots or shuts down, traffic passes only through the primary peer, as expected.

- **Peer link failure (between leaves)**

  When the peer link fails, the peer with the secondary role shuts down all the MC-LAG member interfaces attached to it. The traffic is traversing only through the primary peer switch.

- **Uplink failure (between leaf and spine)**

  According to the MC-LAG feature, the peer link carries very little traffic compared to the bandwidth sent by the data plane. The only traffic going across the peer link is traffic from the 'clagd process' and some LLDP or LACP traffic. The traffic received on the peer link is not forwarded out of the dual-connected bonds. In the scenario seen in the image below, when the uplink failure occurs on a leaf, the traffic that previously passed through the uplinks now traverses through the peer link.

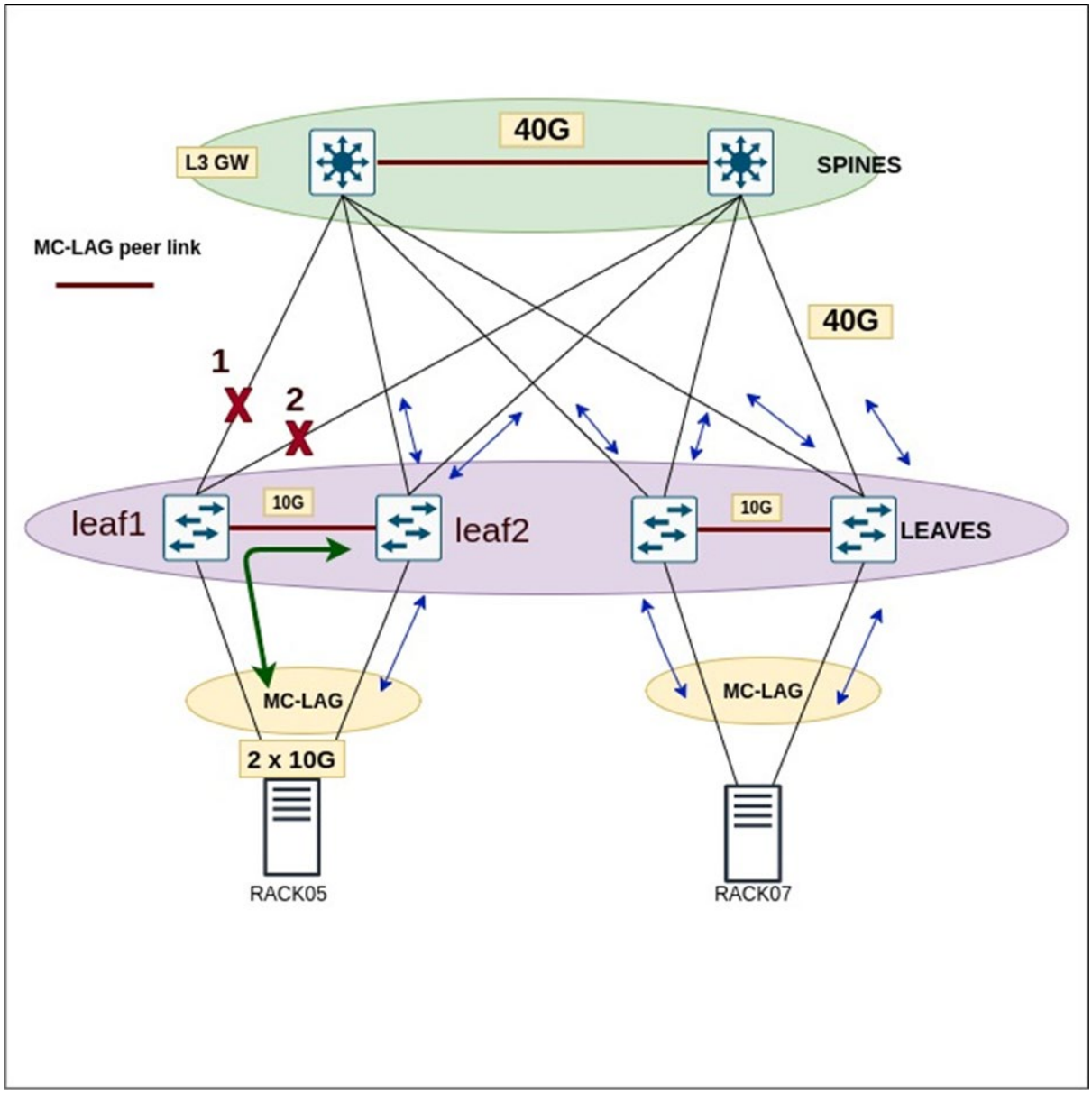

Figure 4.2: MC-LAG uplink failure scenario

As shown in the above figure, the scenario was to shut down leaf1 first, link No1, and then additionally shut down link No2. As such, a uplinks failure was simulated on leaf1. The test results are shown below.

Interval	Transfer	Bandwidth	Retransmits	Comments
0.00-0.50 sec	1.15 GBytes	19.7 Gbits/sec	0	(omitted)
0.50-1.00 sec	1.13 GBytes	19.4 Gbits/sec	0	(omitted)
1.00-1.50 sec	1.15 GBytes	19.8 Gbits/sec	0	(omitted)
1.50-2.00 sec	1.15 GBytes	19.7 Gbits/sec	2	(omitted)
0.00-0.50 sec	1.15 GBytes	19.7 Gbits/sec	0	
0.50-1.00 sec	1.15 GBytes	19.7 Gbits/sec	0	
4.50-5.00 sec	1.15 GBytes	19.6 Gbits/sec	0	
5.00-5.50 sec	1.15 GBytes	19.8 Gbits/sec	0	
5.50-6.00 sec	1.13 GBytes	19.4 Gbits/sec	0	
*** output omitted ***				
10.50-11.00 sec	1.15 GBytes	19.8 Gbits/sec	0	
13.50-14.00 sec	1.14 GBytes	19.6 Gbits/sec	0	
14.00-14.50 sec	1.15 GBytes	19.8 Gbits/sec	0	
14.50-15.00 sec	1.14 GBytes	19.6 Gbits/sec	0	
15.00-15.50 sec	**1.15 GBytes**	**19.7 Gbits/sec**	**0**	
15.50-16.00 sec	**936 MBytes**	**15.7 Gbits/sec**	**0**	**Link-failure No1**
16.00-16.50 sec	**217 MBytes**	**3.65 Gbits/sec**	**30**	
16.50-17.00 sec	**1.12 GBytes**	**19.2 Gbits/sec**	**134**	
17.00-17.50 sec	1.15 GBytes	19.8 Gbits/sec	0	
17.50-18.00 sec	1.12 GBytes	19.2 Gbits/sec	0	
18.00-18.50 sec	1.14 GBytes	19.6 Gbits/sec	0	
18.50-19.00 sec	1.14 GBytes	19.6 Gbits/sec	0	
19.00-19.50 sec	1.15 GBytes	19.7 Gbits/sec	151	
19.50-20.00 sec	1.14 GBytes	19.6 Gbits/sec	0	
20.00-20.50 sec	**1.14 GBytes**	**19.7 Gbits/sec**	**0**	
20.50-21.00 sec	**510 MBytes**	**8.55 Gbits/sec**	**40**	**Link-failure No2**
21.00-21.50 sec	**349 MBytes**	**5.85 Gbits/sec**	**315**	
21.50-22.00 sec	**1.14 GBytes**	**19.6 Gbits/sec**	**0**	
22.00-22.50 sec	1.14 GBytes	19.6 Gbits/sec	0	
22.50-23.00 sec	1.13 GBytes	19.5 Gbits/sec	0	
23.00-23.50 sec	1.13 GBytes	19.5 Gbits/sec	0	
23.50-24.00 sec	1.15 GBytes	19.7 Gbits/sec	0	
*** output omitted ***				
54.50-55.00 sec	1.14 GBytes	19.7 Gbits/sec	0	
55.00-55.50 sec	1.14 GBytes	19.6 Gbits/sec	0	
55.50-56.00 sec	1.14 GBytes	19.6 Gbits/sec	0	
56.00-56.50 sec	1.14 GBytes	19.6 Gbits/sec	0	
56.50-57.00 sec	1.13 GBytes	19.5 Gbits/sec	0	
57.00-57.50 sec	1.12 GBytes	19.3 Gbits/sec	0	
57.50-58.00 sec	1.15 GBytes	19.7 Gbits/sec	0	
58.00-58.50 sec	1.14 GBytes	19.6 Gbits/sec	0	
58.50-59.01 sec	1.15 GBytes	19.4 Gbits/sec	0	
59.01-59.50 sec	1.12 GBytes	19.7 Gbits/sec	3	
59.50-60.00 sec	1.14 GBytes	19.6 Gbits/sec	0	
0.00-60.00 sec	**134 GBytes**	**19.2 Gbits/sec**	**674**	**sender**
0.00-60.00 sec	**134 GBytes**	**19.2 Gbits/sec**		**receiver**

Figure 4.3: Test results

As is apparent, there is slight service degradation for approximately 1 second after each uplink failure (between leaf and spine), with some few TCP retransmits. After the second uplink failure (the first uplink was still down), traffic traverses through the peer link and throughput values are the same as before the failure.

4.1.2 MAC Mobility

It is possible for a given host or VM to move from one Ethernet segment to another. This is referred to as 'MAC mobility' or 'MAC move'. During a MAC move, there are two sets of MAC/IP Advertisement routes, one set with the new Ethernet segment and one set with the previous Ethernet segment. During this VM movement, there is an amount of time that the MAC address would appear to be reachable via both segments. For all the switches in a data centre fabric in the EVPN to correctly determine the current location of the MAC address, all advertisements via the previous Ethernet segment must be withdrawn by the switches in the Fabric (PEs). This is accomplished through the MAC Mobility extended community attribute [RFC_7432]. A leaf detecting a locally attached MAC address for which it had previously received a MAC/IP Advertisement route with a different Ethernet segment identifier advertises the MAC address in a MAC/IP Advertisement route tagged with a MAC Mobility extended community attribute with a sequence number greater than the sequence number in the MAC Mobility extended community attribute of the received MAC/IP Advertisement route (see Figure 4.4).

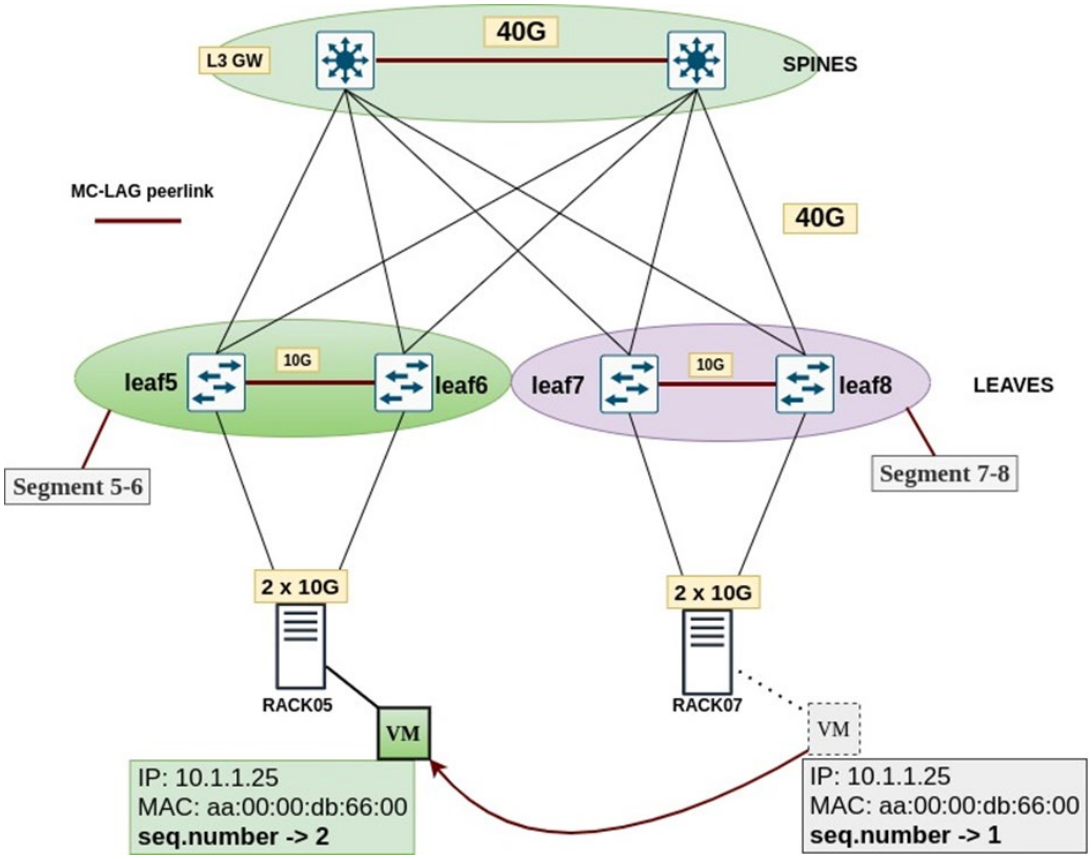

Figure 4.4: MAC mobility of a VM from server_rack07 to server_rack05

For the purpose of the test, a VM was created on server_rack07 and moved via a cloud orchestration tool (Ganeti cluster) towards server_rack05. The VM IP is 10.1.1.25 and MAC aa:00:00:db:66:00. The results are explained in steps below:

1. The VM is on server_rack7. This is shown by the EVPN MAC learning on leaf5.
   *root@**Leaf05**:mgmt:/home/user# net show evpn mac vni 200 mac aa:00:00:db:66:00*
   *MAC: aa:00:00:db:66:00 *** VM MAC address ****
   *Remote VTEP: 10.10.10.70 *** MAC advertised from both leaf7 & leaf8 (Segment 7-8) ****
   Sync-info: neigh#: 0
   *Local Seq: 0 **Remote Seq: 1 *** VM is on remote server_rack7 with sequence No 1******
   Neighbours:
   10.1.1.25 Active
   fe80::a800:ff:fedb:6686 Active

2. The VM is migrated to server_rack5. This is shown on both leaf5 and leaf7 MAC learning.
   *root@**Leaf05**:mgmt:/home/user# net show evpn mac vni 200 mac aa:00:00:db:66:00*
   MAC: aa:00:00:db:66:00
   *Intf: bond1(62) VLAN: 200 ***** MAC is now found under bond1 interface on leaf5 ******
   Sync-info: neigh#: 0
   Local Seq: 2 Remote Seq: 1 *** Local sequence number is incremented by 1 ***
   Duplicate detection started at Wen Aug 11 03:40:28 2021, detection count 1
   Neighbours:
   10.1.1.25 Active
   *fe80::a800:ff:*fedb:6686 Active

   *root@**Leaf07**:mgmt:/home/user# net show evpn mac vni 200 mac aa:00:00:db:66:00*
   MAC: aa:00:00:db:66:00
   *Remote VTEP: 10.10.10.50 *** MAC is advertised from remote Segment 5-6 ****
   Sync-info: neigh#: 0
   Local Seq: 1 Remote Seq: 2 *** leaf7 shows that VM went to server_rack5 ***
   Neighbours:
   10.1.1.25 Active
   fe80::a800:ff:fedb:6686 Active

The above NCLU command results shows that the VM was successfully migrated from server_rack7 to server_rack5. On leaf5, the MAC Mobility extended community attribute has a sequence number 2 which is greater than the sequence number 1 that leaf7 (and leaf8) advertises. There is also a message for duplicate address detection which is normal and has no impact since the sequence number decides the actual VM location. As a result, all white-box switches withdraw MAC advertisement of segment 7-8 and forward traffic on segment 5-6.

During the migration no ICMP packet loss was observed. Logs from leaf switches show that the MAC mobility convergence time was 5.8 ms

4.2 East-West Traffic Performance Validation

This section details the performance validation tests and results across the data centre. For this purpose, several TCP and UDP throughput tests were run using the iperf3 traffic generator tool. The typical testing paths were between servers in the fabric either hosted on the same or a different pair of ToR switches (leaves).

- TCP test (30 concurrent sessions, bidirectional traffic, MTU = 9000, window size 64MB servers on different ToRs).

Interval	Transfer	Bandwidth	Retransmits
0.00-40.00 sec	90.8 GBytes	19.5 Gbits/sec	32 sender
0.00-40.00 sec	90.8 GBytes	19.5 Gbits/sec	receiver

TCP throughput with Jumbo frames is very close to maximum port aggregation speed (2 x 10G ports).

- TCP test (20 concurrent TCP sessions, bidirectional traffic, MTU = 1500, window size 32MB).

Interval	Transfer	Bandwidth	Retransmits
0.00-40.00 sec	86.0 GBytes	18.5 Gbits/sec	107 sender
0.00-40.00 sec	86.0 GBytes	18.5 Gbits/sec	receiver

TCP throughput with lower MTU size is lower as expected but still at a satisfactory rate.

- UDP throughput in the same leaf switch (server_rack5 to server_rack6), 20 concurrent streams with 19 Gbps maximum bandwidth, 8948 Bytes datagram size.

Interval	Transfer	Bandwidth	Retransmits
0.00-20.00 sec	37.8 GBytes	16.2 Gbits/sec	0.019 ms

- UDP throughput in different pair of leaf switches (server_rack5 to server_rack7), 20 concurrent streams with 19Gbps maximum bandwidth, 8948 Bytes datagram size.

Interval	Transfer	Bandwidth	Retransmits
0.00-20.00 sec	38.4 GBytes	16.5 Gbits/sec	0.015 ms

UDP throughput is slower than TCP with the condition that the percentage of lost datagrams remains acceptable.

Another test was conducted to check the latency between the servers, which showed that the latency across servers in a different pair of leaf switches (server_rack5 to server_rack8) was 0.149ms (average).

5 Conclusions

The results in Section 4 show that all the fundamental functionalities and features were achieved for the data centre use case. Moreover, the white box switches performance results are sufficient for regular operation and can handle several failure scenarios. The network devices are now implemented in production. Currently, this is the beginning of this small data centre that will gradually gain momentum as new computing requests emerge.

Although there is a significant learning curve for engineers to get familiar with the Cumulus Linux environment and troubleshooting, the NVIDIA networking docs web page [NVIDIA-1] provides multiple use case scenarios and configuration examples. The troubleshooting documentation was not sufficient, but there is a very active Cumulus community (via a Slack channel) that can be of significant help with configuration and troubleshooting.

GRNET's design decision was to purchase White Boxes with the Broadcom chipset (June 2020), taking into account the support of Cumulus NOS at that time. The fact that NVIDIA acquired Mellanox (March 2019) and later Cumulus (June 2020) had a substantial impact on further deployment of Cumulus NOS. The EVPN Multihoming feature (used by GRNET in all production data centres), developed from Cumulus in the 4.2 release, is supported only for Mellanox ASICs (Spectrum A1, Spectrum 2 and Spectrum 3) [EVPN-2]. Therefore, GRNET turned to MC-LAG as an alternative solution. The MC-LAG feature has disadvantages compared to EVPN-MH [EVPN-2] and brings inconsistency with other GRNET production data centre deployments. Despite the disadvantages, feature and performance tests proved that MC-LAG functions properly, as expected for a widely used feature in the network industry.

More significantly, beginning from release 4.4 (July 2021), Cumulus Linux supports only NVIDIA Spectrum-based ASIC platforms [CMS]. This release removes support for Broadcom-based networking ASICs. Broadcom-based ASICs will continue to be supported throughout the life of the Cumulus Linux 3.7 and 4.3 releases. The white box, acquired by GRNET, based on Broadcom ASICs, will reach end of life within the next few years. The software and hardware are decoupled in a white box, which provides independence from these market changes. As a result, GRNET can either keep white box switch hardware and try another NOS (eg. Pluribus), or try out white box switches with NVIDIA chipsets and keep the Cumulus NOS (keeping in mind that the NVIDIA chipsets are relatively low-cost). The switches based on Broadcom (Edgecore) acquired could be easily reused for another use case.

References

[D6.3]	Deliverable D6.3 *White Box Evaluation* https://www.geant.org/Projects/GEANT_Project_GN4-3/GN43_deliverables/D6-3_White-Box-Evaluation.pdf
[BGP]	BGP unnumbered on Cumulus Networks operating system https://docs.nvidia.com/networking-ethernet-software/cumulus-linux-42/Layer-3/Border-Gateway-Protocol-BGP/
[CMS]	Cumulus Linux release version 4.4 https://docs.nvidia.com/networking-ethernet-software/cumulus-linux-44/Whats-New/
[CMS-EVPN-CFG]	NVIDIA EVPN Configuration Examples https://docs.nvidia.com/networking-ethernet-software/cumulus-linux-44/Network-Virtualization/Ethernet-Virtual-Private-Network-EVPN/Configuration-Examples/
[EVPN]	VXLAN routing with EVPN: asymmetric vs. symmetric model https://cumulusnetworks.com/blog/asymmetric-vs-symmetric-model/
[EVPN-2]	EVPN Multihomin https://docs.nvidia.com/networking-ethernet-software/cumulus-linux-42/Network-Virtualization/Ethernet-Virtual-Private-Network-EVPN/EVPN-Multihoming/
[FABRIC-EVPN]	IP FABRIC EVPN-VXLAN REFERENCE ARCHITECTURE https://www.juniper.net/content/dam/www/assets/reference-architectures/us/en/ip-fabric-evpn-vxlan-reference-architecture.pdf
[FRR]	FRR routing Project https://frrouting.org/
[ILACP-BY]	LACP Bypass https://docs.nvidia.com/networking-ethernet-software/cumulus-linux-43/Layer-2/LACP-Bypass/
[NVIDIA-1]	NVIDIA networking docs webpage https://docs.nvidia.com/networking-ethernet-software/
[RFC_7432]	RFC7432 - Section 15 - MAC mobility https://datatracker.ietf.org/doc/html/rfc7432#section-15
[TCO]	White box TCO calculator https://www.geant.org/Projects/GEANT_Project_GN4-3/GN43_deliverables

Glossary

ACL	Access Control List
ASIC	Application-Specific Integrated Circuit
BGP	Border Gateway Protocol
BPDU	Bridge Protocol Data Unit
CLI	Command-Line Interface
DHCP	Dynamic Host Configuration Protocol
eBGP	External Border Gateway Protocol
ECMP	Equal-Cost Multi-Path Routing
ENHE	Extended Next Hop Encoding
EVPN	Ethernet VPN
FRR	Free Range Routing
GRNET	Greek National Research and Education Network
IGP	Interior Gateway Protocol
IP	Internet Protocol
LACP	Link Aggregation Control Protocol
LAN	Local Area Network
MAC	Media Access Control
MC-LAG	Multi-Chassis Link Aggregation
NCLU	Network Command Line Utility
NOS	Network Operating System
OS	Operating System
PoP	Point of Presence
PXE	Preboot Execution Environment
R&E	Research and Education
RFC	Request for Comments
RFP	Request For Proposal
RIB	Routing Information Base
STP	Spanning Tree Protocol
TCO	Total Cost of Ownership
TCP	Transmission Control Protocol
ToR	Top-of-Rack
UDP	User Datagram Protocol
VLAN	Virtual Local Area Network
VPN	Virtual Private Network
VRR	Virtual Router Redundancy
VXLAN	Virtual Extensible Local Area Network
WB	White Box