White Box: GRNET Data Centre Use Case

Lefteris Poulakakis (GRNET)
Theodore Vasilopoulos (GRNET)
Giannis Korakis (GRNET)
Christos Argyropoulos (Elastic)

GÉANT Infoshare: White Boxing for NREN use cases
Wed 8th December

Agenda

- Project objectives
- White Box switches overview
- Data Centre topology overview
- Validation tests and results
- Conclusions

Objectives

- Build a new small-scale Data Centre (DC) to host cloud resources for our customers as well as for GRNET internal resource needs.
- Reduce cost.
- Improve independence from vendors.
- Get acquainted with White Box switch concept in general.

White Box switches

Traditionally, the Data Centre network is 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.

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

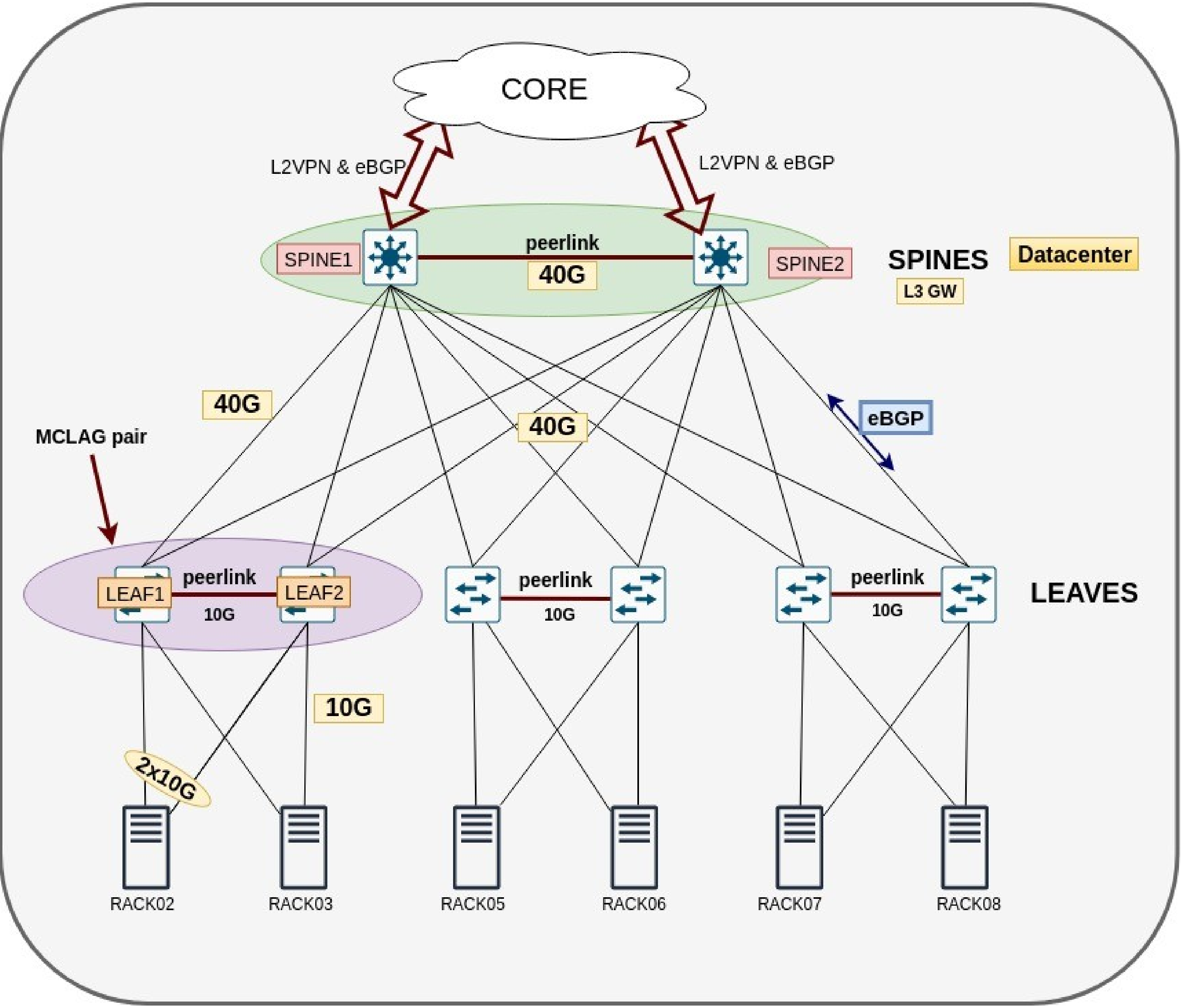
Datacenter use-case design decisions

GRNET is already operating three medium-to-large scale DCs in three different Points of Presence (PoPs) based on:

- IP Clos topology (Spine-Leaf architecture)
- EVPN/VxLAN protocol
- Ansible for configuration management.

Decided to use the same architecture principles but based on White Box switches and non-hardware vendor NOS to provide L2/L3 connectivity to the cloud resources.

Topology



‘**All active links and switches**’ setup for redundancy.

Each server is dual-homed to a pair of leaves with an LACP bond of 2x10Gbps interfaces.

Pair of leaves create **MCLAG peering** through a connection between them (peerlink).

MCLAG feature provides redundancy and load-balancing on hosts.

Each leaf is connected to a couple of spines using 40G physical links.

L3 termination on Spines with virtual gateway redundancy

Underlay (packet forwarding): **eBGP unnumbered (RFC5549).**
Spine-Leaf neighbourship with IPv6 link local addresses.
No need to setup IPv4 point-to-point addresses.

Overlay control plane:
EVPN is used for MAC/IP advertisement of hosts across the DC.

Overlay dataplane: **VxLAN**
Leaves perform VxLAN tunnel encapsulation/decapsulation.

External communication of DC:
Double physical links,L2VPN and double eBGP peerings from each spine towards Border Routers.

NOS acquisition

NOS: Cumulus Linux (version 4.3)

- Widely used in Data Centre solutions
- Native Linux distribution based on Debian-Buster
- NOS that runs on switches from multiple vendors
- Acquired by NVIDIA (March 2019)

Cumulus Linux Architecture

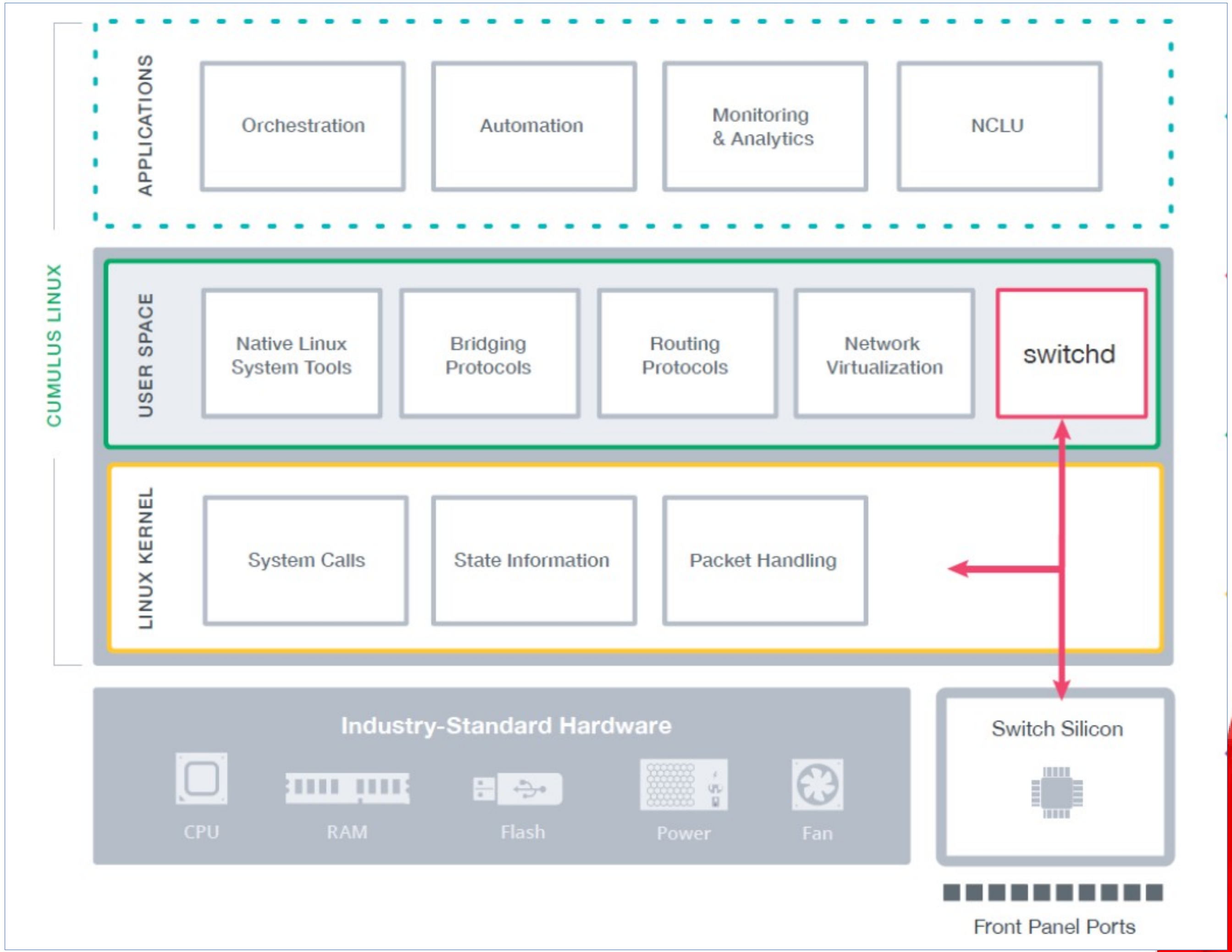
- Process called ‘**switchd**’ peers with the kernel and directly with network ASICs and normalizes the networking model

- Uses **FRRouting** protocol suite for all routing functions

- Cumulus authored **NCLU tool**: A command line interface that simplifies the networking configuration process for all users.

- Almost any Linux tool can be used for configuration and management

- Linux commands and configuration files are always available.



HW Equipment acquisition

Spine: Edgecore AS7712-32X

- 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

HW Equipment acquisition

Leaf: Edgecore AS5812-54T

- 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

Equipment acquisition- Cost

- The solution should not exceed the cost of the budget required for the previous network data centre implementations of GRNET using traditional vendors.

- The cost assessment was made thanks to the TCO calculator [TCO] previously published by the GN4-3 WP6 T1 white box team.

- The cost of the Spine Edgecore AS-7712 with the Juniper QFX-10k-36Q were compared.

- White Box solution was about 10% cheaper.

MCLAG vs EVPN Multihoming

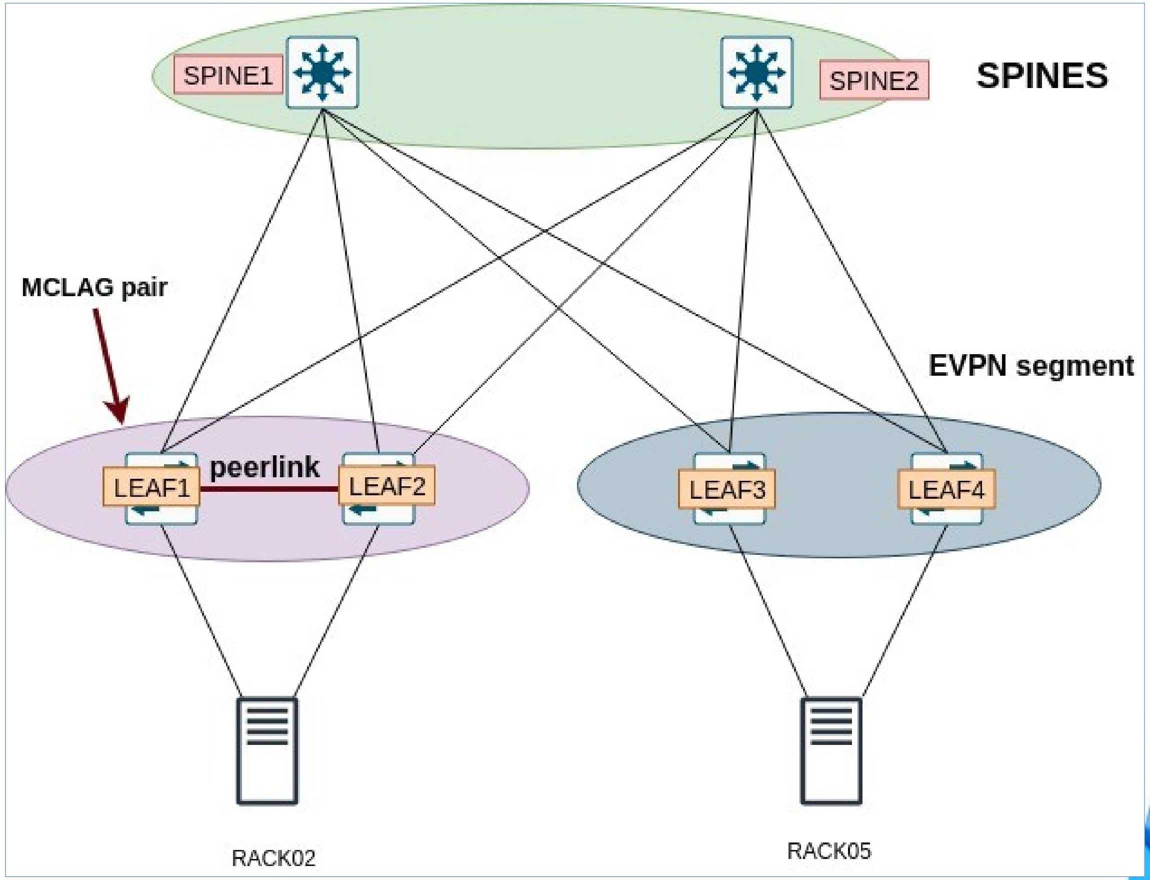
MCLAG: uses peer link between switches to determine host sharing and primary-secondary negotiation.

EVPN-MH: uses the same ESI value (Ethernet Segment Identifier) through EVPN to determine dual-homed hosts.

EVPN has advantages against MCLAG:

- Eliminates the need for peer links or inter-switch links between the top of rack switches.
- Peer link capacity must increase proportionally to the traffic served to underlying servers.
- Allows multi-vendor interoperability

GRNET uses EVPN-MH to production DCs



Cumulus release 4.2 does not support EVPN-MH for Broadcom Chipset. Used MCLAG to overcome the problem.

Testing

Configuration and management

- NCLU
- Ansible
- Support Documents

Protocol features and failure scenarios

- eBGP unnumbered, eBGP towards core routers
- ECMP and load balancing across data centre links
- EVPN MAC/IP advertisements
- DHCP relay
- L3 GW redundancy
- MCLAG failure scenarios
- ACLs

- **Performance measurements (iperf3)**
- TCP/UDP traffic (Up to 20Gbps = LAG capacity)
- Latency

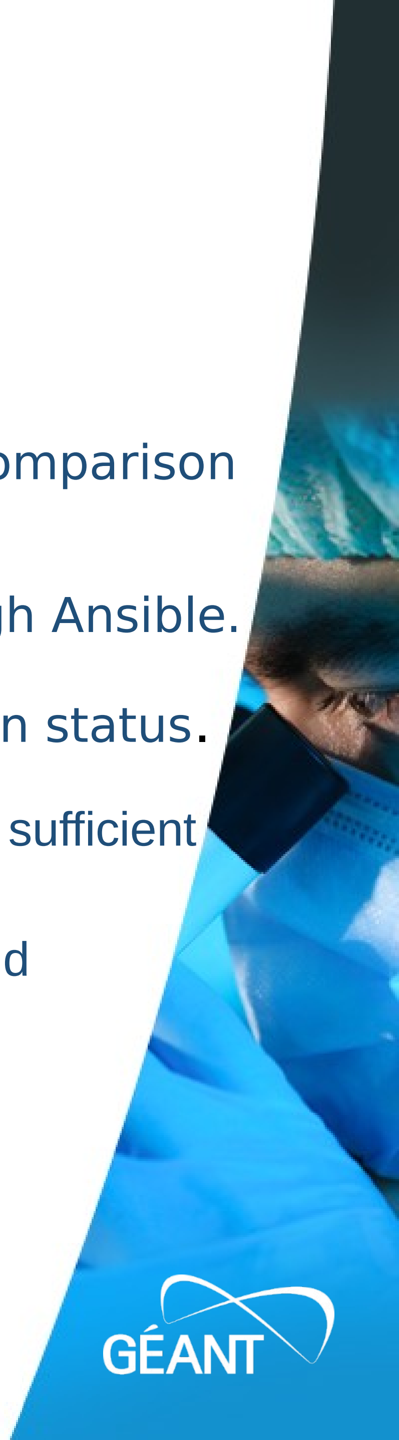
Special feature validation

- MAC mobility

Results and validation

Configuration and Management

- NCLU ‘net show’ commands sometimes give irrelevant and misleading output in comparison with configuration files.
- More reliable to directly change the **configuration files** either manually or through Ansible.
- The native Linux commands are available and always display the real configuration status.
- Documentation for Cumulus NOS has a lot of configuration scenarios and examples but is not sufficient enough for troubleshoot.
- Very active community (via a Slack channel) that can be of significant help on configuration and troubleshooting.



Results and validation

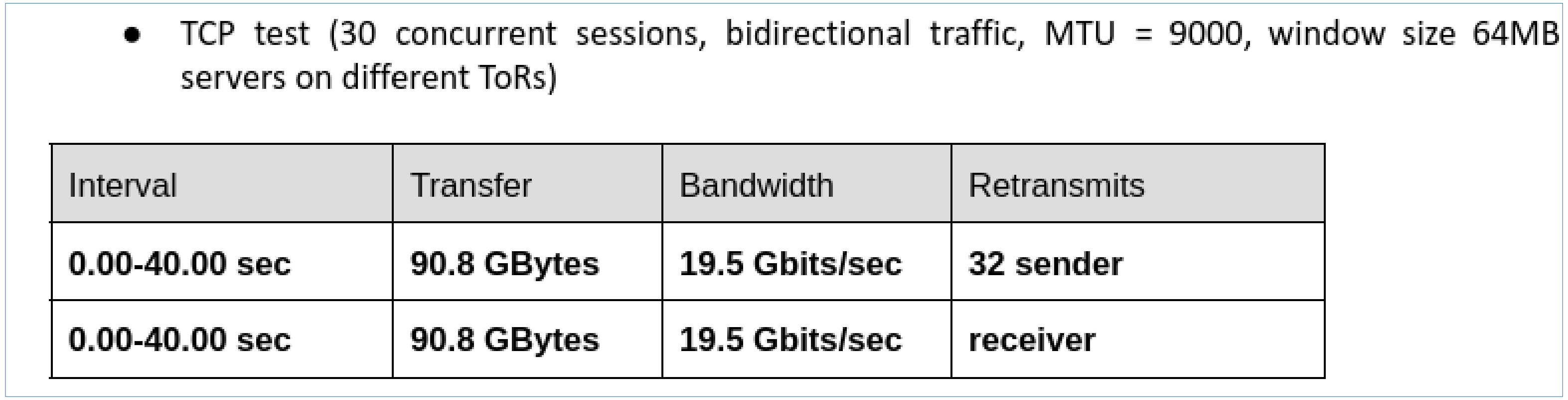
Protocol features and failure scenarios

- All fundamental protocol features were successfully tested.
- Despite the disadvantages against EVPN-MH, MCLAG feature tests, performance tests and failure scenarios proved that **MCLAG functions properly**.

Results and validation

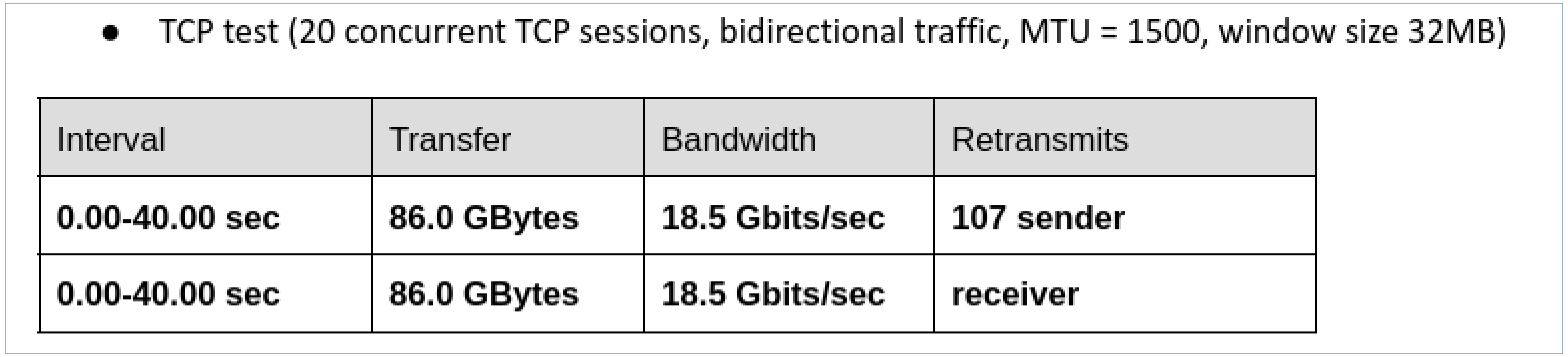
Performance measurements

- Sufficient for regular operation and many failure scenarios.



- 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**

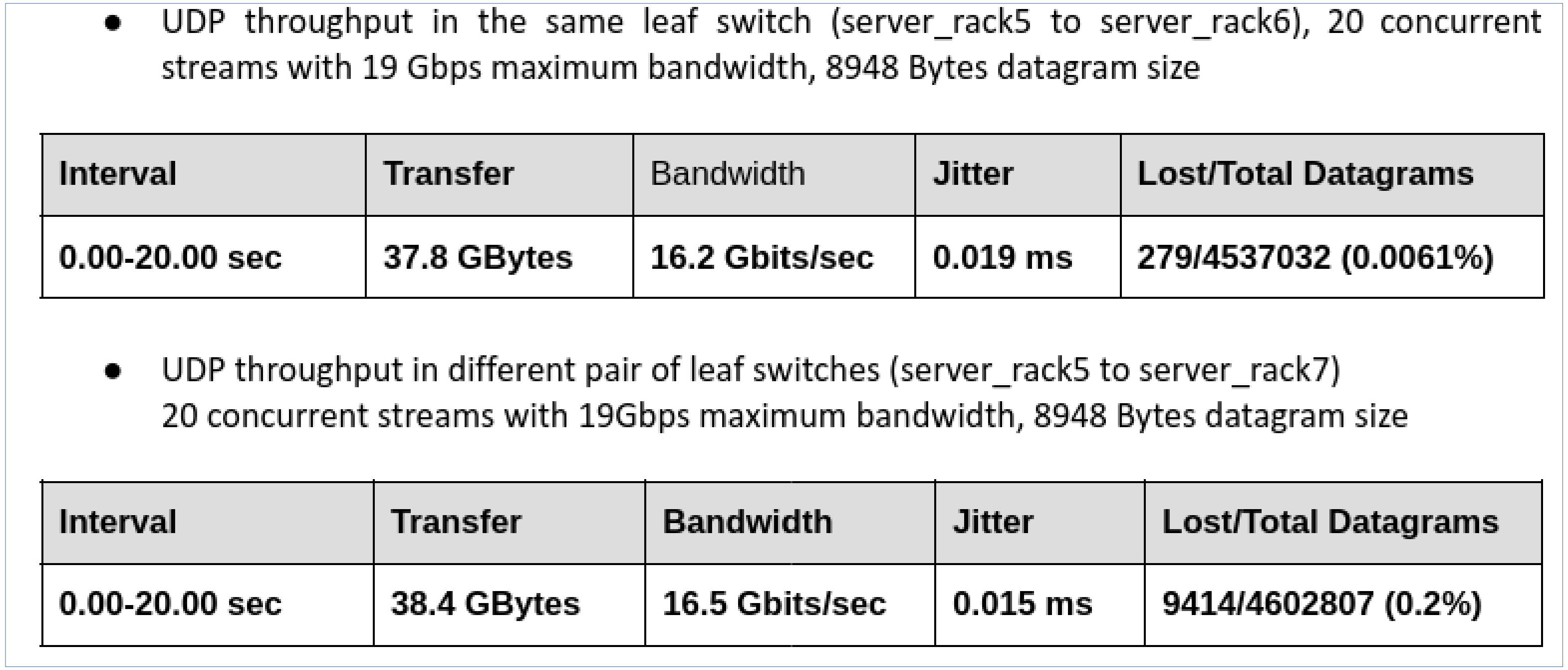


- TCP test (20 concurrent TCP sessions, bidirectional traffic, MTU = 1500, window size 32MB)

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**

Results and validation

Performance measurements



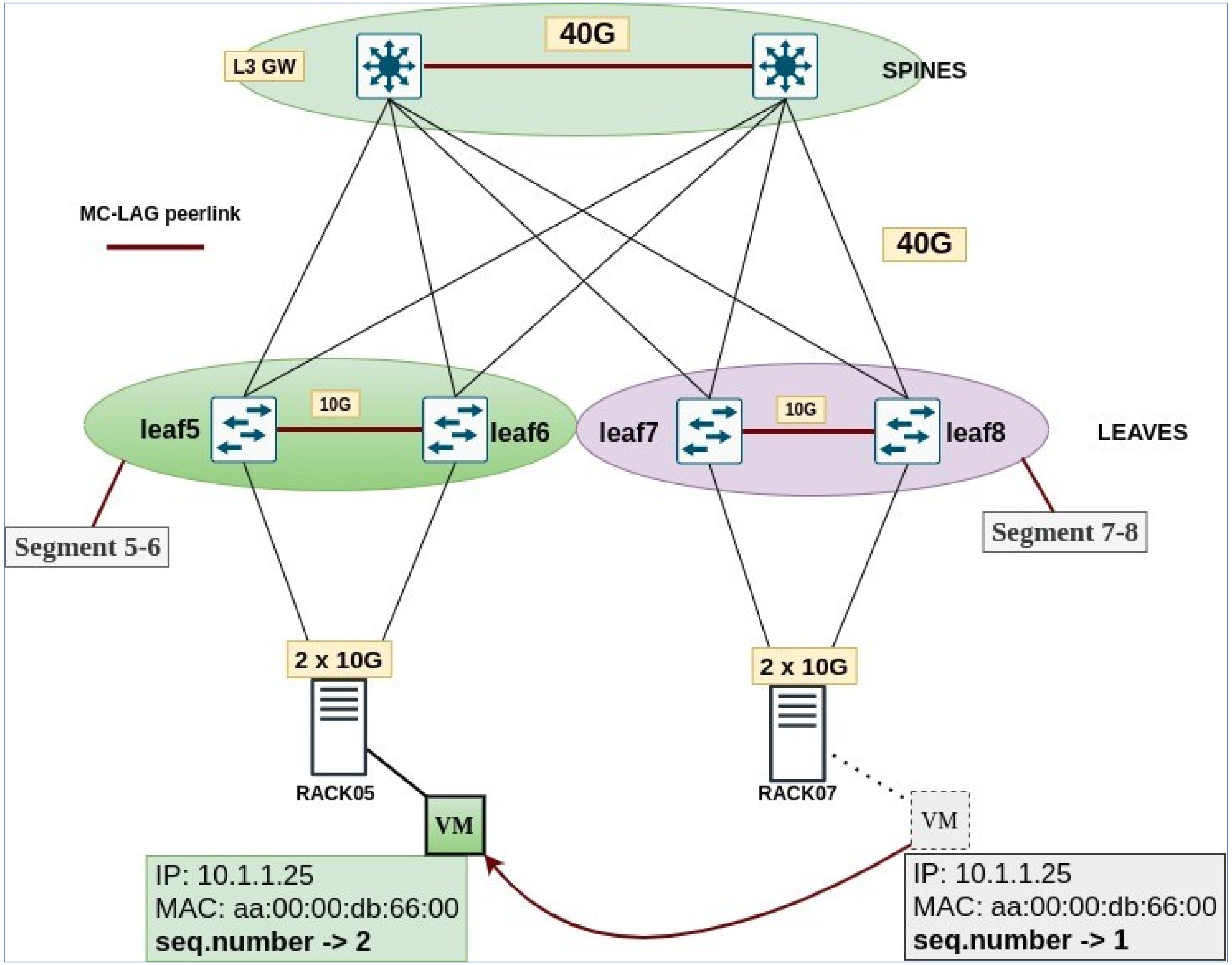
- UDP throughput in the same leaf switch (server_rack5 to server_rack6), 20 concurrent streams with 19 Gbps maximum bandwidth, 8948 Bytes datagram size

Interval	Transfer	Bandwidth	Jitter	Lost/Total Datagrams
0.00-20.00 sec	37.8 GBytes	16.2 Gbits/sec	0.019 ms	279/4537032 (0.0061%)

- 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	Jitter	Lost/Total Datagrams
0.00-20.00 sec	38.4 GBytes	16.5 Gbits/sec	0.015 ms	9414/4602807 (0.2%)

Special Feature Validation:
MAC Mobility



IP: 10.1.1.25
MAC: aa:00:00:db:66:00
seq.number -> 2

IP: 10.1.1.25
MAC: aa:00:00:db:66:00
seq.number -> 1

- Quite often in Data Centres there is a need to move a host/VM from one HW node to another.
- EVPN takes care of the movement with MAC mobility feature(RFC_7432).
- MAC Mobility extended community attribute.

Testing:

- VM movement through Ganeti cluster management tool from Rack07 to Rack05.
- MAC mobility convergence time was **5.8 ms**.

Market changes

- Cumulus NOS initially supported several types of **Spectrum and Broadcom ASICs**

- **March 2019**: NVIDIA acquired Mellanox

- **June 2020**: GRNET acquired Edgecore switches with **Broadcom Chipset**

- **June 2020**: NVIDIA acquired Cumulus

- **July 2020**: Cumulus 4.2 release - EVPN MH deployment only on Mellanox switches with Spectrum ASIC.
  solution: MCLAG

- **July 2021**: Cumulus release 4.4 supports only Mellanox switches.
  Cumulus 4.3 is the last release that supports Broadcom chipset.

- **December 2025**: Cumulus release 4.3 - End of Support.

- GRNET – 4 remaining years of support.

- Either keep Cumulus NOS and try Mellanox switches or keep Edgecore switches and try another NOS.

GÉANT
Networks • Services • People

Thank you

Any questions?

www.geant.org

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).