The Heart of the Data Center

Scalable and Reliable Network Design Guide: A Practical Approach

Fiordoliva Federico (INAF - Osservatorio Astronomico di Roma)

Project Objectives

Unified Layer 2 Management for Project Servers

Key Requirements

- Link Redundancy (Bonding/Aggregation).
- Support for Mixed Connectivity (Fiber and Copper) and speeds from 1 to 40 Gbps.
- High Availability and Fault Tolerance (Distribution and Aggregation areas).
- Advanced Layer 2 Security.
- In-service (Runtime) Updates.

Hardware/Software Selection

- **Modular Switches (SFP, SFP+, QSFP+)**
- **Logical Stacking** [1]
- **Supported Protocols**
  - DHCP Snooping
  - UAC (User Access Control)
  - MSTP/RSTP
  - EVPN-VXLAN
  - LACP

Aggregation Stack

- 3 Modular Chassis / Units
- 12 QSFP+ Ports
- 96 SFP/SFP+ Ports
- Dual Power Supply (Redundant PSU)
- Up to 1.44 Tbps [Switching Capacity]

Distribution Stack [TOR]

- 8 Units/Appliances
- 32 QSFP+ Ports
- 32 SFP/SFP+ Ports
- 384 1000 BASE-T Ports
- Dual Power Supply (Redundant PSU)
- Up to 500 Gbps [Switching Capacity]

Architecture: Virtual Chassis Stack

- 3 Separate VCs (Virtual Chassis)
  - Aggregation: DMZ
  - Distribution: DMZ-A, DMZ-B

Architecture: Virtual Chassis Stack

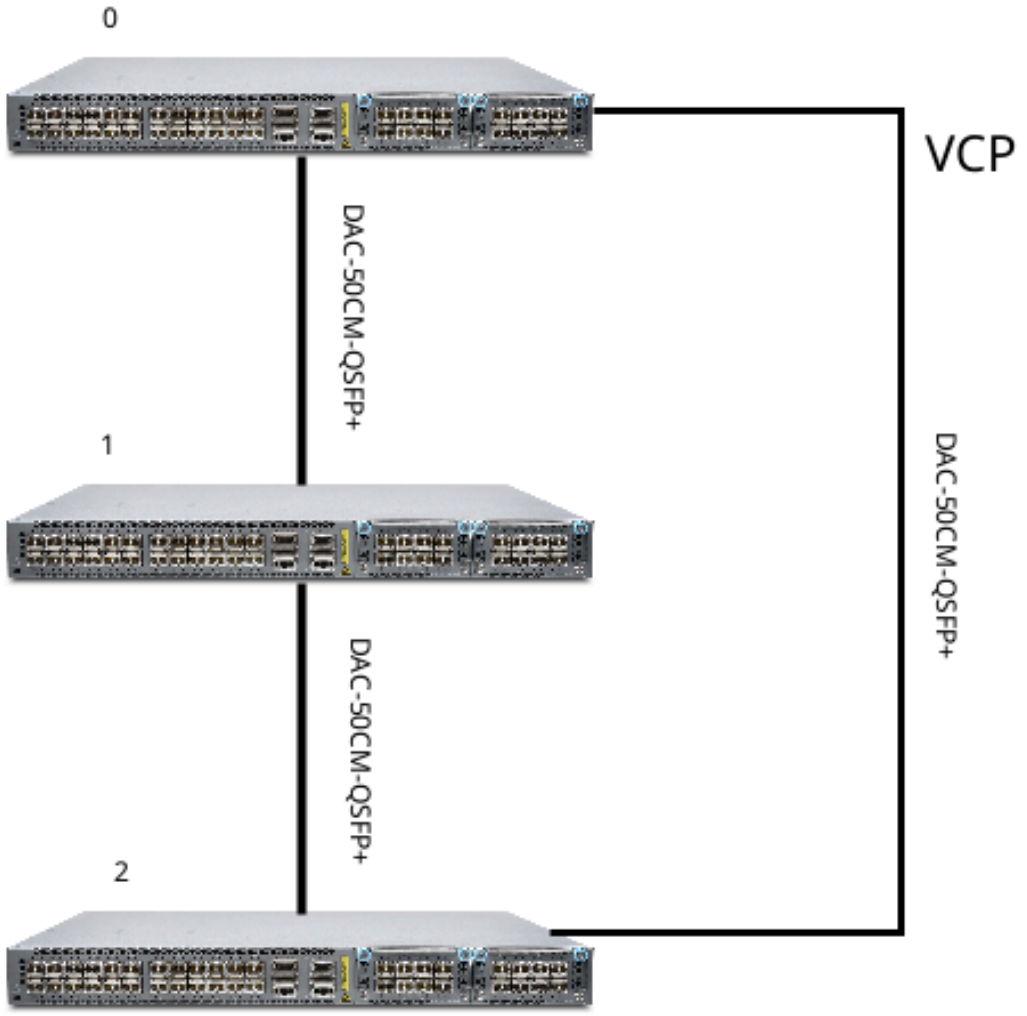
???

Architecture: Virtual Chassis Stack

- **Acts as a single logical switch**
- **Zero-touch expansion: new switches require no manual configuration**
- **Support for Multi-Chassis LACP (802.3ad/802.1AX)**
  - All units share a unified switching table
- **Single configuration file for simplified management, backup, and recovery.**

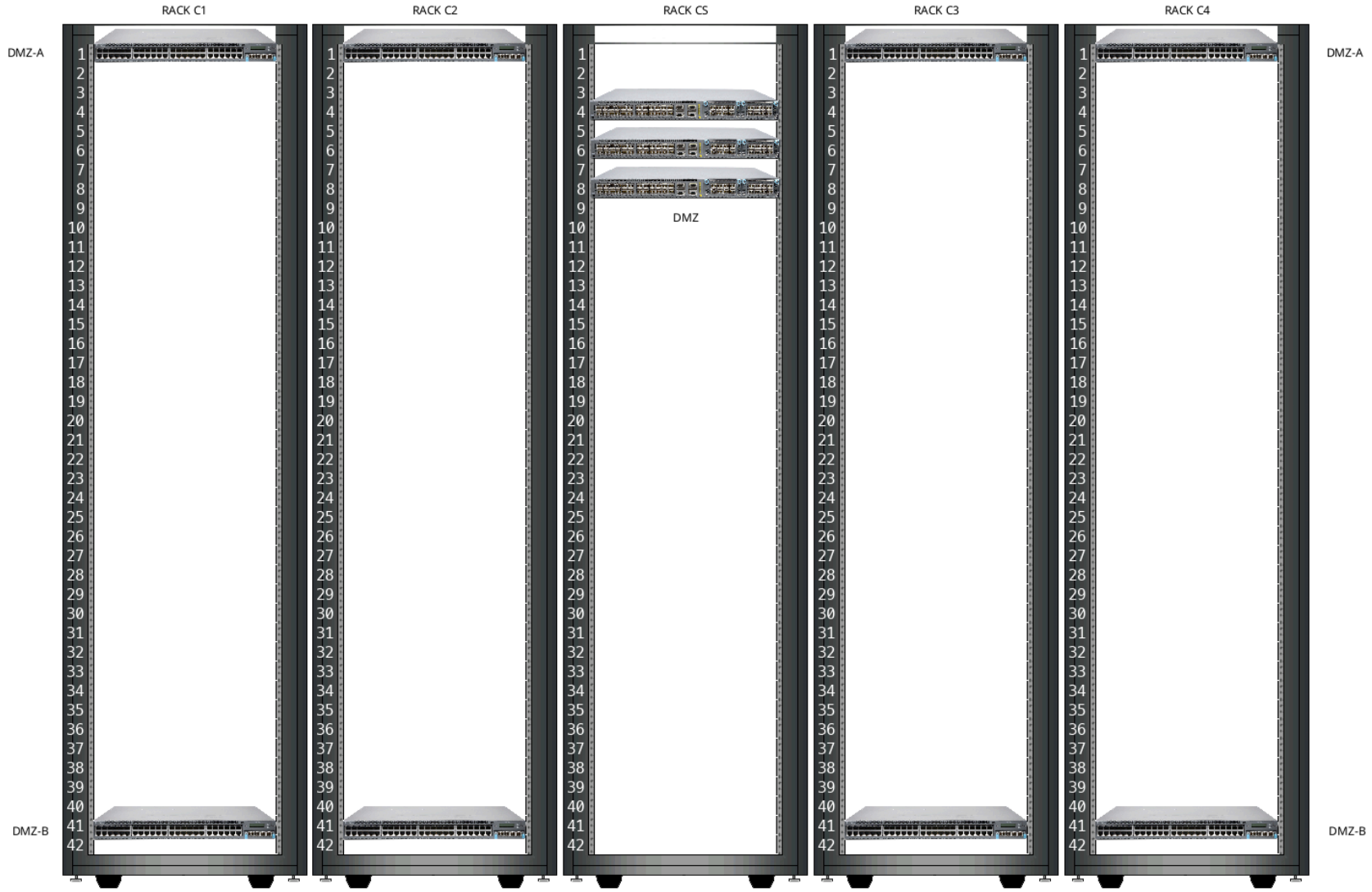
DMZ

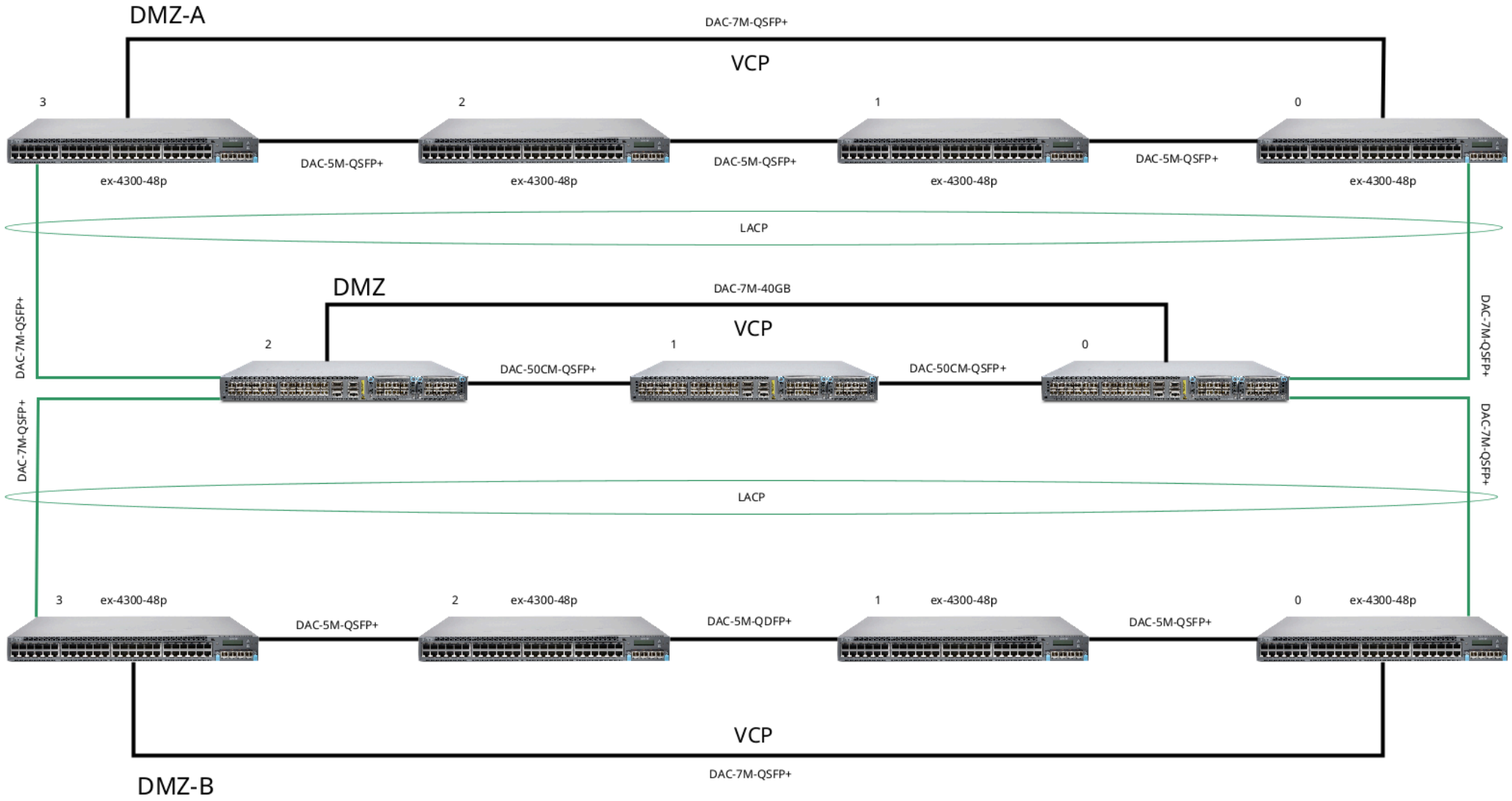
- Ring Topology / Ring Connection
- Direct Attach Cables (DAC)[2]
- Switch ID



DMZ-A

DMZ-B


RACK C1
RACK C2
RACK CS
RACK C3
RACK C4
DMZ-A
DMZ-B
DMZ
DMZ-A
DMZ-B

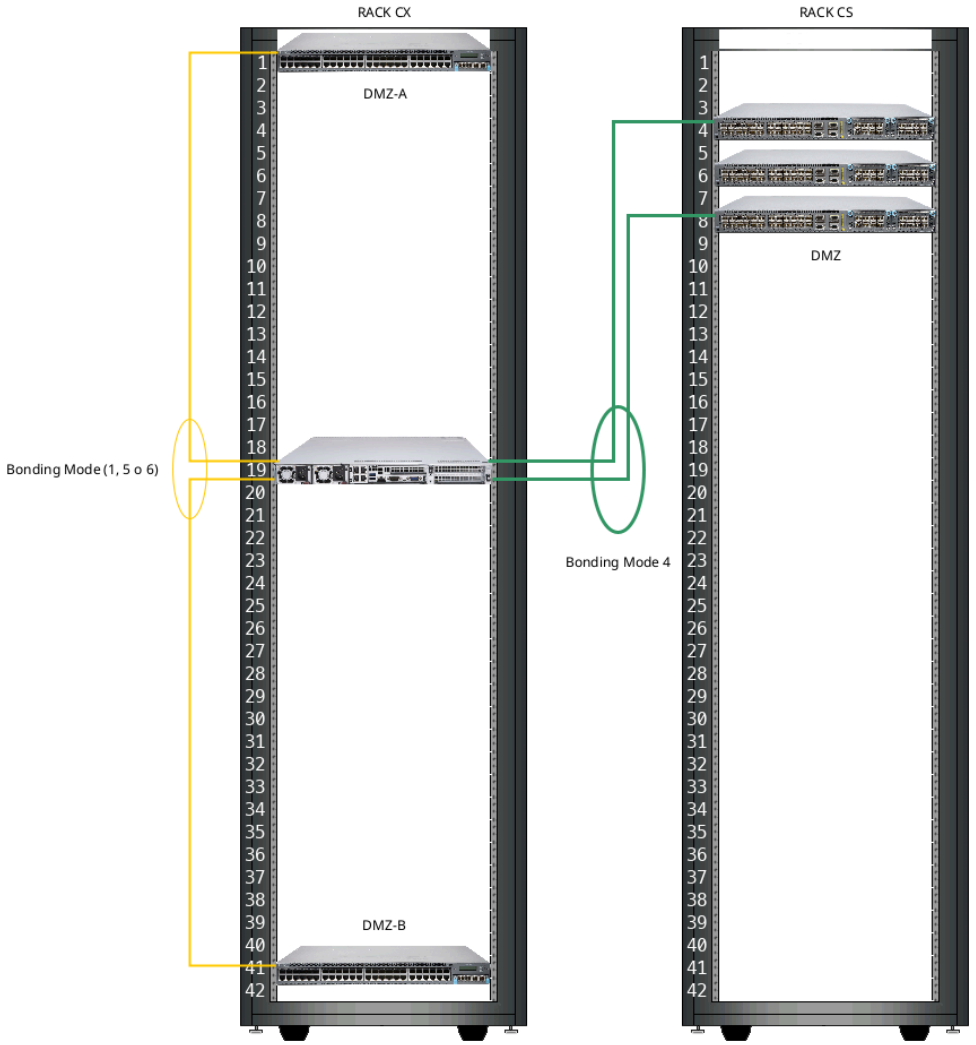

DMZ-A
DAC-7M-QSFP+
VCP
3
2
1
0
DAC-5M-QSFP+
DAC-5M-QSFP+
DAC-5M-QSFP+
ex-4300-48p
ex-4300-48p
ex-4300-48p
ex-4300-48p
LACP
DMZ
DAC-7M-40GB
VCP
DAC-7M-QSFP+
DAC-7M-QSFP+
2
1
0
DAC-50CM-QSFP+
DAC-50CM-QSFP+
DAC-7M-QSFP+
DAC-7M-QSFP+
LACP
3
ex-4300-48p
2
ex-4300-48p
1
ex-4300-48p
0
ex-4300-48p
DAC-5M-QSFP+
DAC-5M-QDFP+
DAC-5M-QSFP+
VCP
DAC-7M-QSFP+
DMZ-B

Server Requirements

- At least 2 Copper ports (RJ45)
- At least 2 SFP/SFP+ ports

Server Connectivity / Links

- Copper: Connected to DMZ-A and DMZ-B using Bonding Mode 1, 5, or 6
- Fiber: LACP (802.3ad) connected to the DMZ Aggregation Stack



Scalability

- Ability to expand each TOR stack up to 8–10 units (covering up to 8–10 racks).
- The DMZ stack is also scalable up to 8–10 units to accommodate additional TOR stacks.

Protocols and Security

Link Aggregation Control Protocol [LACP] – 802.3ad / 802.1AX

Physical Interface Aggregation (LACP)

LACP

Primary Purpose:

- Increase total bandwidth and provide redundancy (fault tolerance).

Simplicity:

- Dynamically manages the addition or removal of links in real-time.

Load Balancing:

- Distributes traffic across available physical links based on hashing algorithms (MAC, IP, or Port-based).

Rapid Spanning Tree Protocol [RSTP] – 802.1w

Loop Prevention

RSTP

Rapid Convergence::

- Transitions from Blocking to Forwarding state within seconds (typically 1–2s).

Backward Compatibility:

- Fully backward compatible with the legacy 802.1D (STP) standard.

Port Roles:

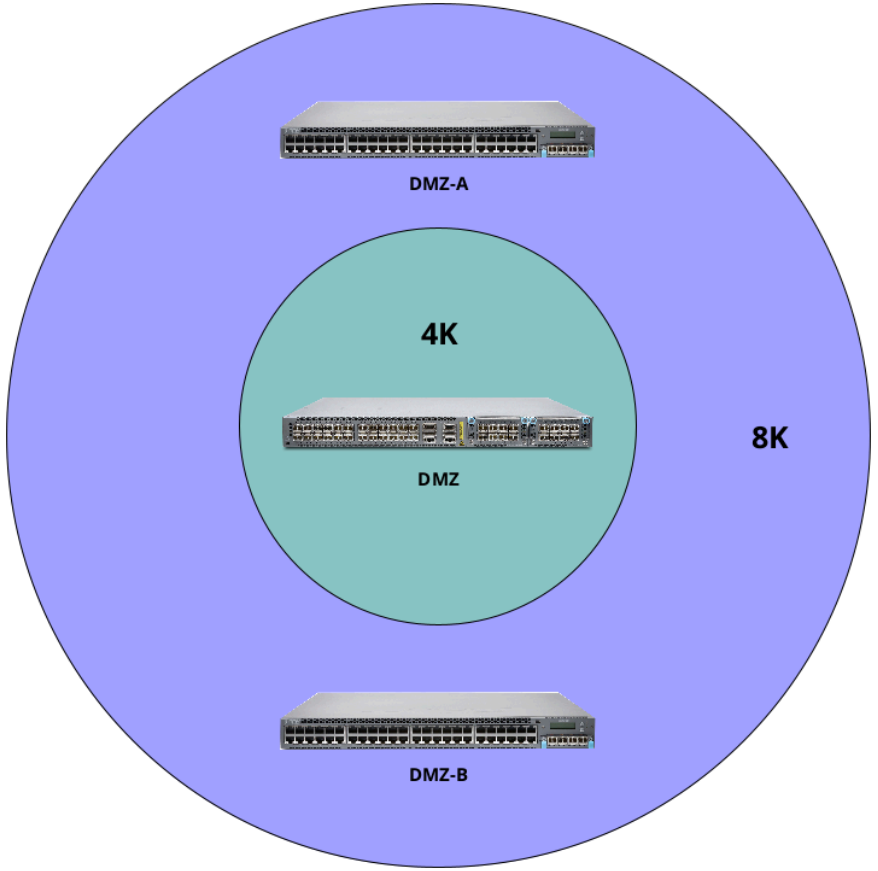
- Introduces Alternate (Root Port backup) and Backup (Designated Port backup) roles for instantaneous failover.

RSTP [Configuration]

Root Bridge Selection:

- Setting the correct priorities ensures the proper functionality of the protocol (Selection of the `ROOT BRIDGE`).
  - Common priority values: `4k, 8k, 16k, 32k`
- Edge Mode Ports:
  - Configured for end-devices (servers/workstations) to ensure fast convergence (immediate transition to Forwarding state).
- Security and Stability:
  - BPDU Guard: Blocks unauthorized topology changes and loops by disabling Edge ports upon receipt of a BPDU..

RSTP



DHCP Snooping

- Blocks packets from unauthorized DHCP servers
- Simple yet powerful security measure

DHCP Snooping

Implementation

- Untrusted (Active): Enabled on EDGE ports to block rogue DHCP offers.
- Trusted (Disabled): Disabled on authorized ports (uplinks, known DHCP servers) to allow legitimate traffic.

Link Layer Discovery Protocol [LLDP] 802.1AB

"Know your neighbor"

LLDP

- Device visibility: Real-time info on connected devices.
- Faster Troubleshooting: Immediate identification of physical link issues.
- DCIM Automation: Seamless integration with systems like NetBox or Nautobot.
- Consistency Checks: Automated verification of configuration alignment across the fabric.

LLDP

```
root@dmz> show lldp neighbors
....
Local Interface    Parent Interface    Chassis Id          Port info          System Name
xe-0/0/12          ae10                7c:c2:55:87:76:e4   ens3f0np0          virt10.oa-roma.inaf.it
xe-1/0/12          ae10                7c:c2:55:87:76:e4   ens3f1np1          virt10.oa-roma.inaf.it
....
```

LLDP

`show lldp neighbors interface xe-0/0/12`

```
....
Local Interface    : xe-0/0/12
Parent Interface   : ae10
Local Port ID      : 609
Ageout Count       : 1

Neighbour Information:
Chassis type       : Mac address
Chassis ID         : 7c:c2:55:87:76:e4
Port type          : Mac address
Port ID            : 3c:ec:ef:dd:15:10
Port description   : ens3f0np0
System name        : virt10.oa-roma.inaf.it

System Description : Debian GNU/Linux
....
```

LLDP

`show lldp neighbors interface xe-0/0/12`

```
....
Organization Info
       OUI      : Ethernet Bridged (0x0080c2)
       Subtype  : VLAN Name (3)
       Info     : VLAN ID (1000), VLAN Name (vlan1000)
       Index    : 1

Organization Info
       OUI      : Ethernet Bridged (0x0080c2)
       Subtype  : VLAN Name (3)
       Info     : VLAN ID (1002), VLAN Name (vlan1002)
       Index    : 2

Organization Info
       OUI      : Ethernet Bridged (0x0080c2)
       Subtype  : VLAN Name (3)
       Info     : VLAN ID (1555), VLAN Name (vlan1555)
       Index    : 3

Organization Info
       OUI      : Ethernet Bridged (0x0080c2)
       Subtype  : VLAN Name (3)
       Info     : VLAN ID (1556), VLAN Name (vlan1556)
       Index    : 4
....
```

Maximum Trasfer Unit [MTU] - Jumbo Frames

- Set MTU to 9000 on all infrastructure inter-switch links (ISLs).
- Set MTU to 9000 on switch ports dedicated to high-throughput traffic (Storage).
- Set MTU to 9000 on server interfaces involved in high-throughput operations (Storage/SAN).

EVPN-VXLAN

Standard-based protocols essential for Disaster Recovery policies and geo-replicated environments.

VXLAN (data-plane)

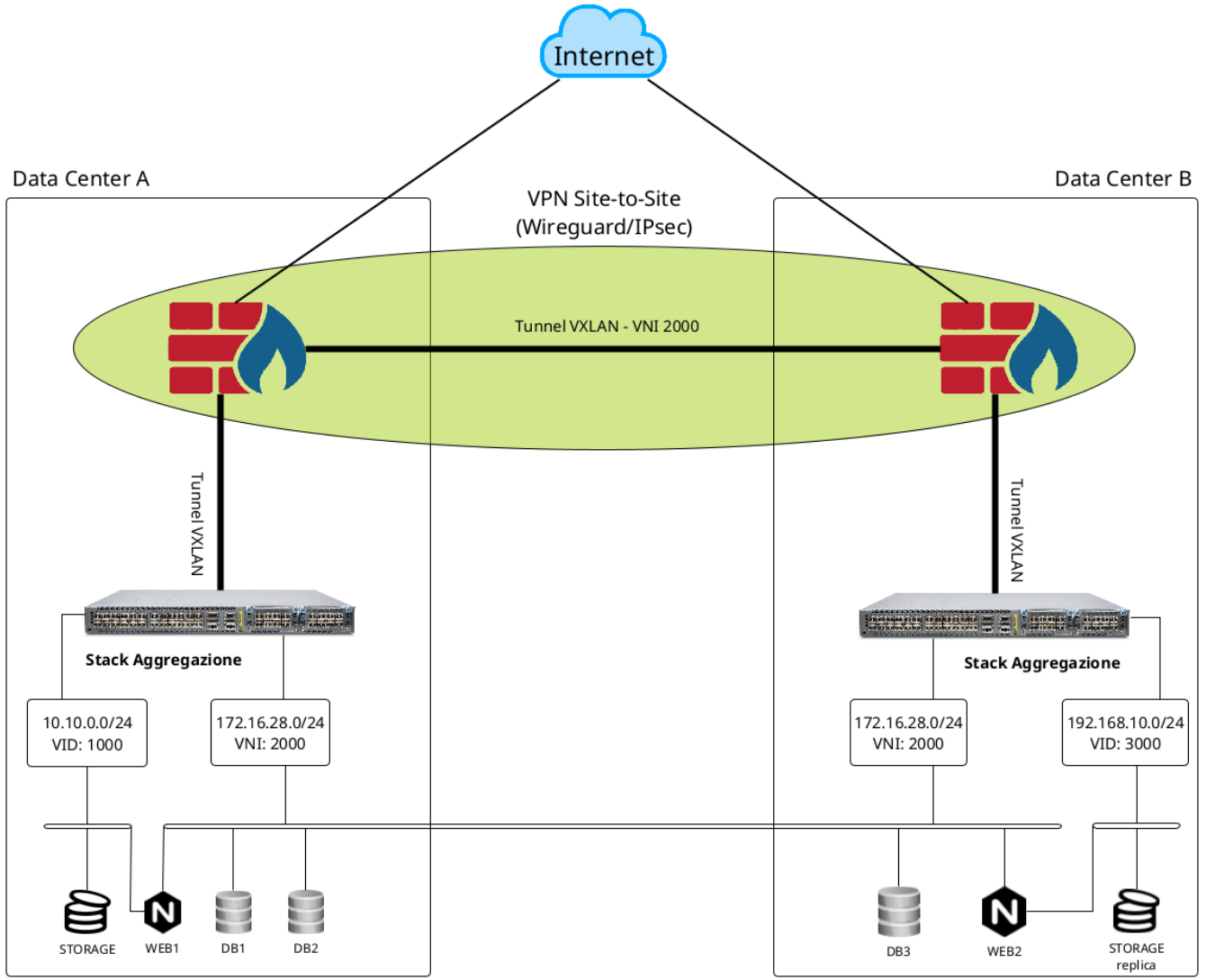
- L2 over L3 Encapsulation: Encapsulates Layer 2 traffic over a Layer 3 network.
- Scalability: Increases the limit from 4,096 VLANs to 16 million VNIs (Virtual Network Identifiers).

EVPN (control-plane)

The "brain" of the network: it manages the distribution of IP-MAC bindings.

Disaster Recovery

- DB Replication
- Asynchronous Storage Replication[3]
- Off-site Backup Replication (e.g., via Tape Library) for geo-redundancy.


Internet
Data Center A
Data Center B
VPN Site-to-Site
(Wireguard/IPsec)
Tunnel VXLAN - VNI 2000
Tunnel VXLAN
Tunnel VXLAN
Stack Aggregazione
Stack Aggregazione
10.10.0.0/24
VID: 1000
172.16.28.0/24
VNI: 2000
172.16.28.0/24
VNI: 2000
192.168.10.0/24
VID: 3000
STORAGE
WEB1
DB1
DB2
DB3
WEB2
STORAGE
replica

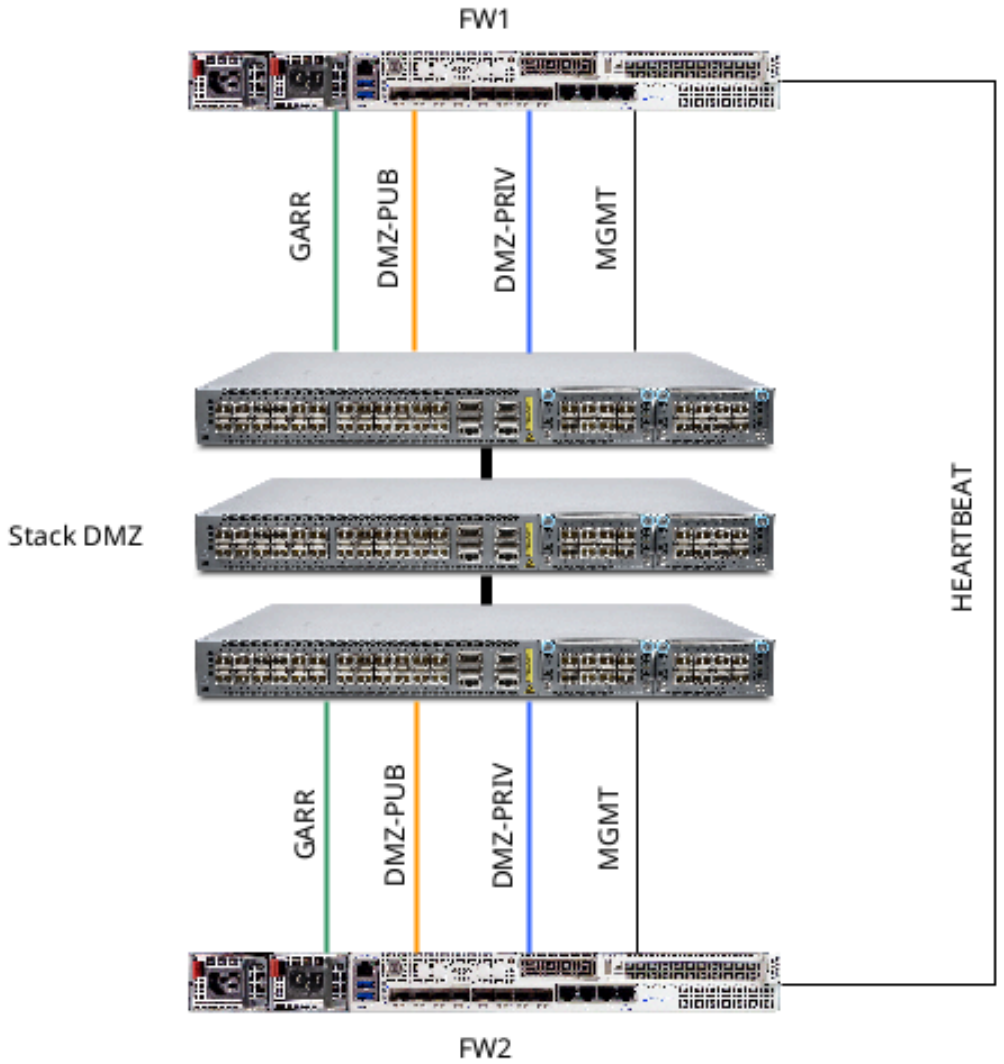
Disaster Recovery

- How can certain services migrate in real-time?
  - GSLB (Global Server Load Balancing): Directs traffic to the most available or geographically closest data center.
  - BGP (Border Gateway Protocol): Manages dynamic routing to announce network prefixes from different locations.

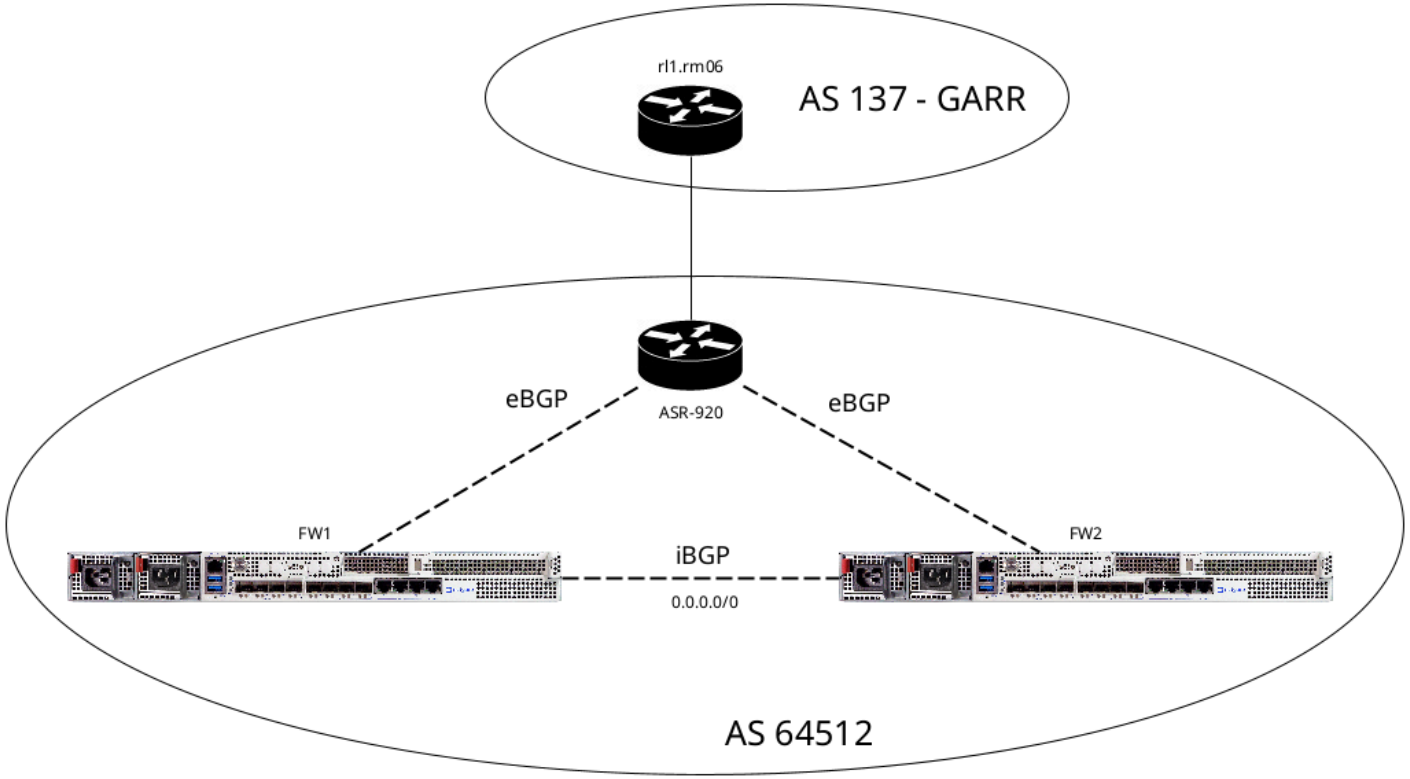
Firewalling

pfSense HA Pair (High Availability):

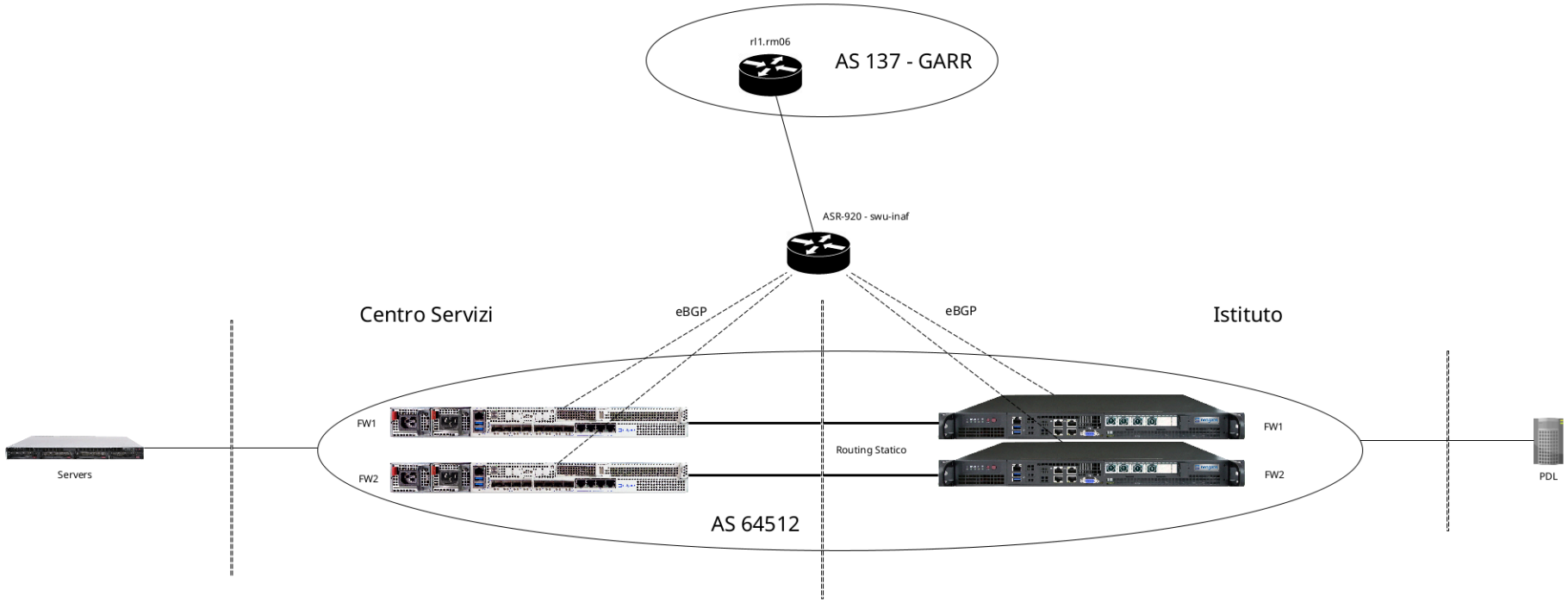
Utilizes a redundant cluster configuration (Active/Passive) to ensure gateway continuity.


FW1
GARR
DMZ-PUB
DMZ-PRIV
MGMT
Stack DMZ
HEARTBEAT
GARR
DMZ-PUB
DMZ-PRIV
MGMT
FW2

BGP


rl1.rm06
AS 137 - GARR
eBGP
ASR-920
eBGP
FW1
iBGP
0.0.0.0/0
FW2
AS 64512

Institute


rl1.rm06
AS 137 - GARR
ASR-920 - swu-inaf
Centro Servizi
eBGP
eBGP
Istituto
FW1
FW2
Routing Statico
FW1
FW2
Servers
PDL
AS 64512

Hardware

- Netgate 8300 MAX
- Juniper EX4600-48F (CS)
- Juniper EX4300-48P (TOR)
- Juniper EX4300-48MP (TOR)

Conclusions

A scalable and easy-to-manage Data Center.
Ready for Disaster Recovery (DR-Ready).

Acknowledgments

Astri Mini Array

CTA+

Stiles

Q&A

Thank you for your attention

Note

1: Some manufacturers limit the stack to 8 units and impose specific distance constraints.
2: Certain vendors support Virtual Chassis (VC) functionality over standard fiber optic links.
3: Storage Disaster Recovery is a complex subject; implementation must be evaluated on a case-by-case basis depending on the specific storage architecture.