
CERN

# HEPiX Fall 2019

# CERN Computer Centre Network evolution

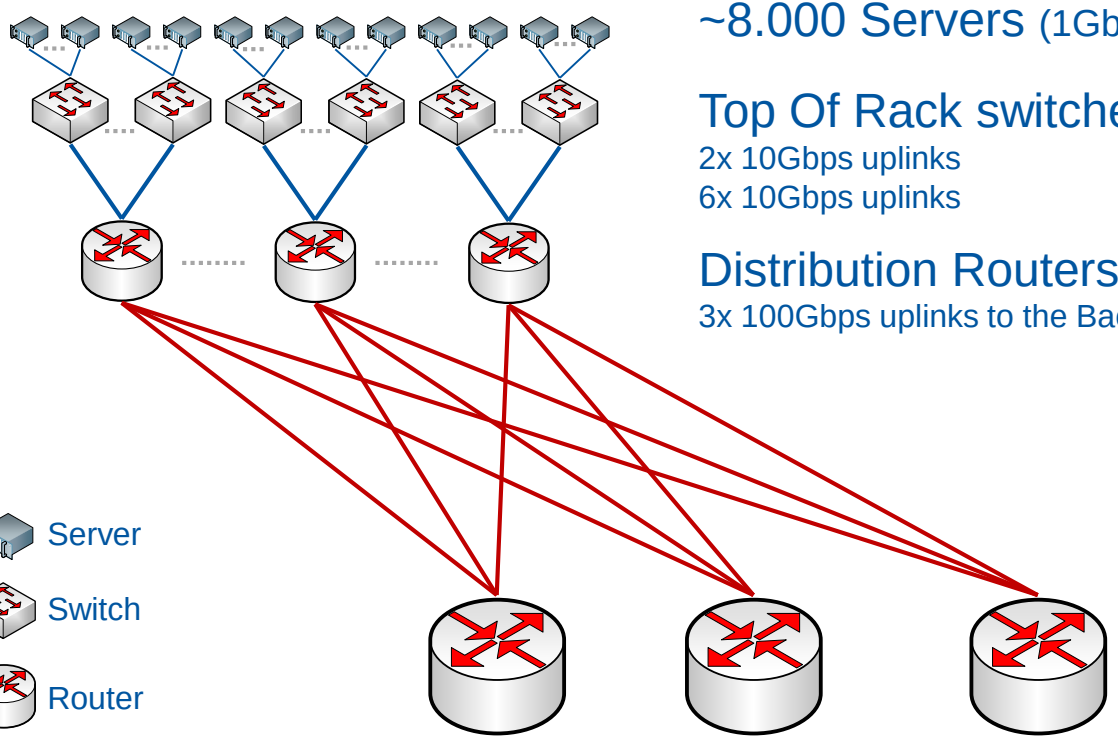
Vincent DUCRET
vincent.ducret@cern.ch

# Agenda

- Former computer centre design
- Core network change
- Connection to the experiments
- Wigner Computer centre
- Distribution network change
- Main issues
- Next steps
- Q&A

# Former design and limitations

# Former computer centre design



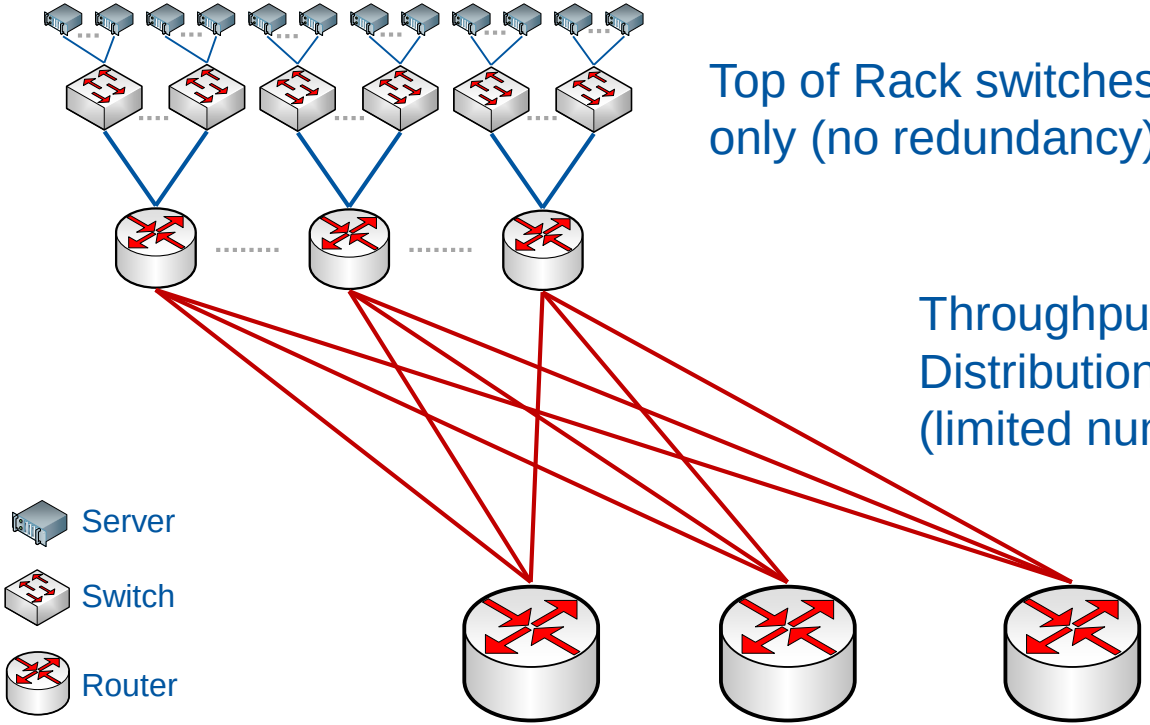

~8.000 Servers (1Gbps or 10Gbps)

Top Of Rack switches (~400)
2x 10Gbps uplinks
6x 10Gbps uplinks

Distribution Routers (Brocade) (x9)
3x 100Gbps uplinks to the Backbone routers

Backbone routers (x3)

# Former computer centre design




Top of Rack switches are attached to one router only (no redundancy)

Throughput capacity is limited between Distribution and Backbone Routers (limited number of 100G ports capacity)

# Goals of the migration

- Increase the overall throughput capacity
  - Provide high density of 100Gbps ports for Distribution/Backbone connections
  - Provide 100Gbps connections to the experiments
  - Provide 40Gbps (or 100Gbps) links between ToR switches and distribution routers

- Increase redundancy and flexibility:
  - Router redundancy for the existing ToR switches
  - Support VxLAN for new features (Tungsten Fabric, SDN, etc…)

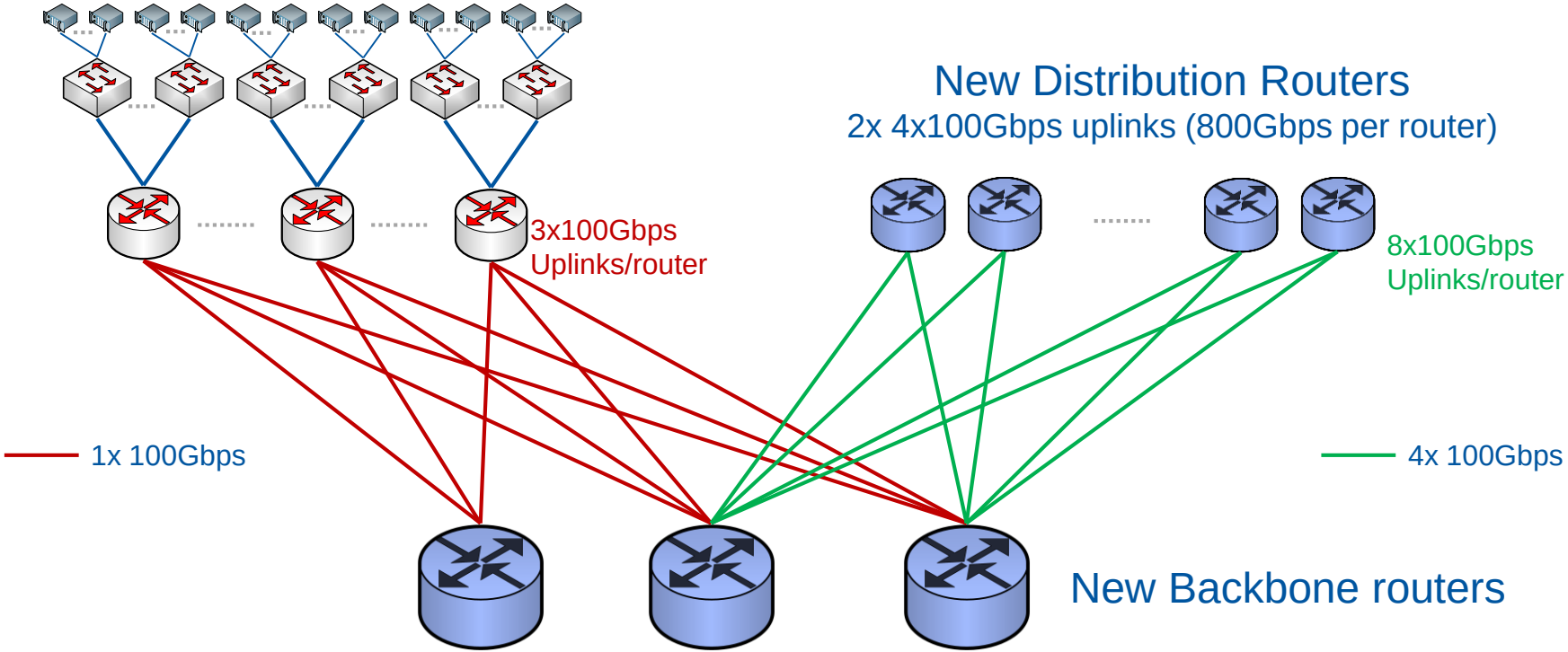
# Computer Centre
# Core network change

# Core network change

From:
33 Rack Units
Max 64x 100Gbps ports

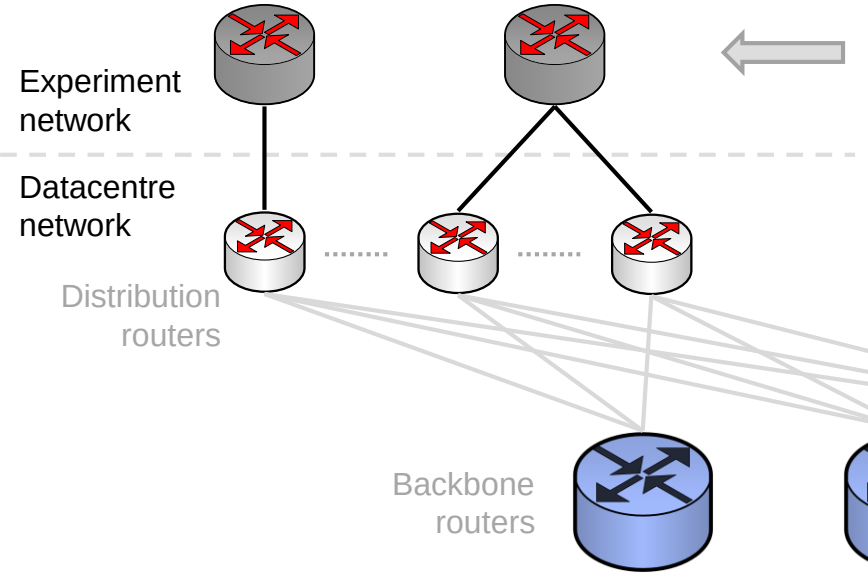
To:
13 Rack Units
Max 240x 40/100Gbps ports

# Core network change



# Connections to the experiments

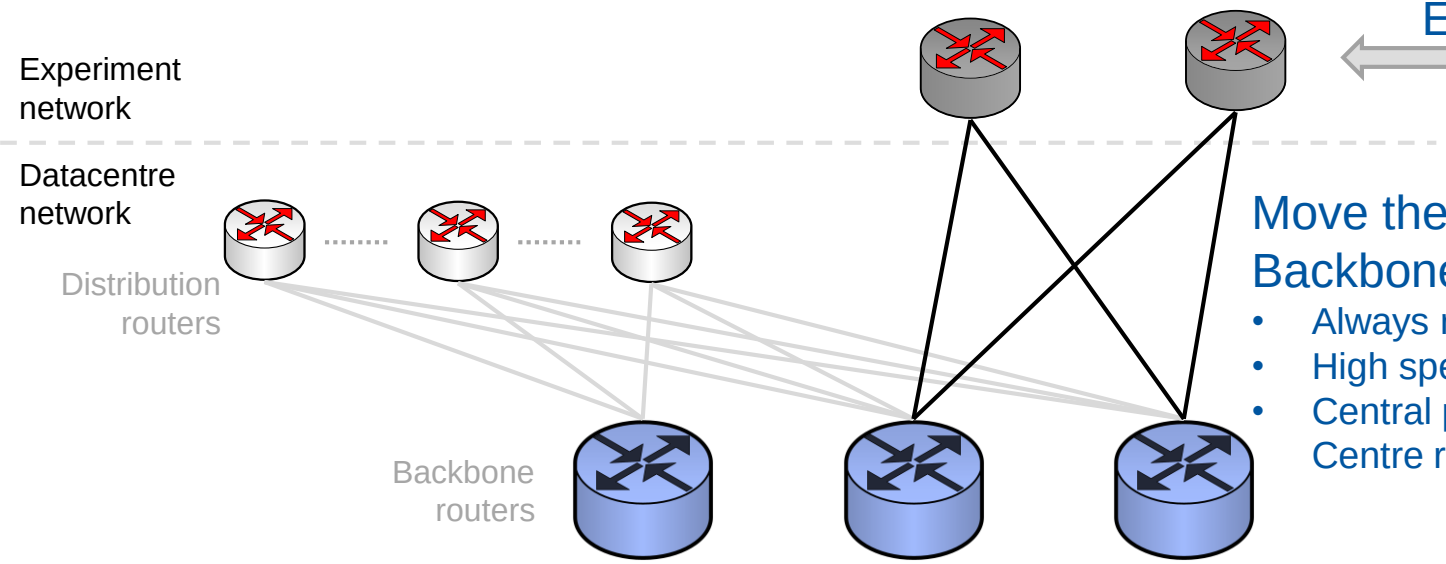
# Connections to the experiments




## Experiment routers (CDR links)

- Links to ALICE, ATLAS, CMS, LHCb, ProtoDUNE
- Attached to the distribution router
- Redundant or not
- Various speeds (from 2x10Gbps to 2x40Gbps)

# Connections to the experiments




## Move the connections to the Backbone routers

- Always redundant
- High speed (40/100Gbps and more)
- Central point of access to Computer Centre resources

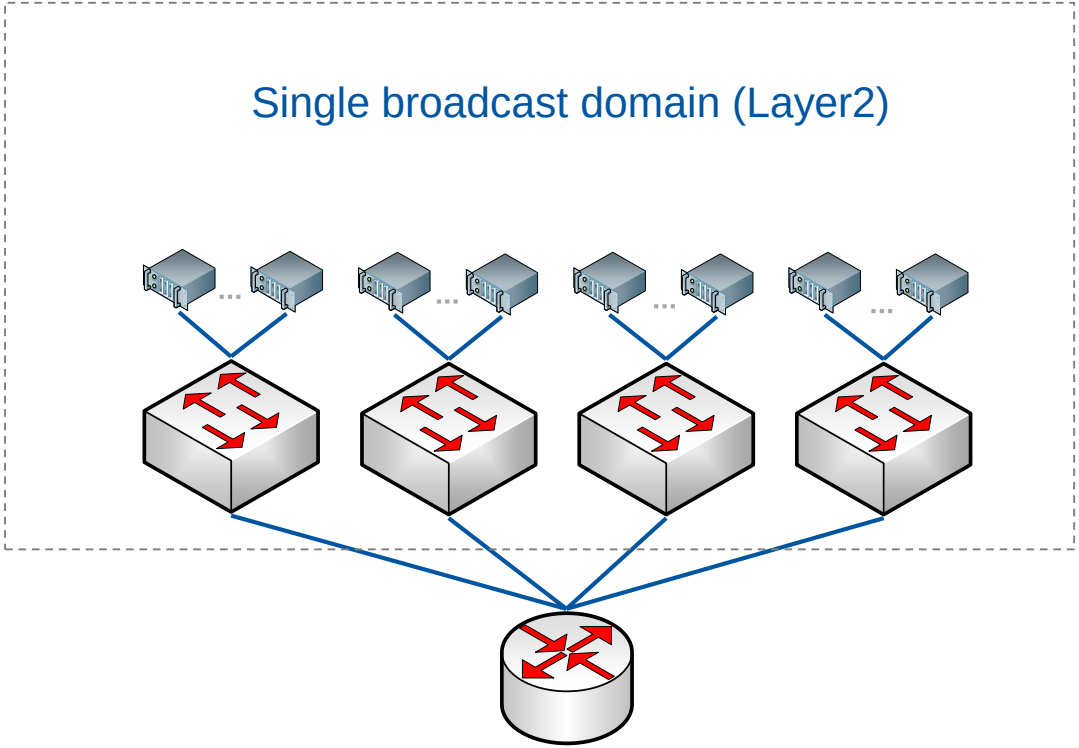
# Wigner Computer Centre

# Wigner computer centre

- Extension located near Budapest (3x100Gbps link with Geneva)
- Contract ending
- Decommissioning during 2019
- Disk servers moved to CERN computer centre
- CPU nodes moved to dedicated physical containers
- More details in:
  - “LHCb containers - Network overview” - Daniele Pomponi
  - “How to provide required resources for Run3 and Run4” - Wayne Salter

# Computer centre Distribution network change

# Distribution network – current option 1

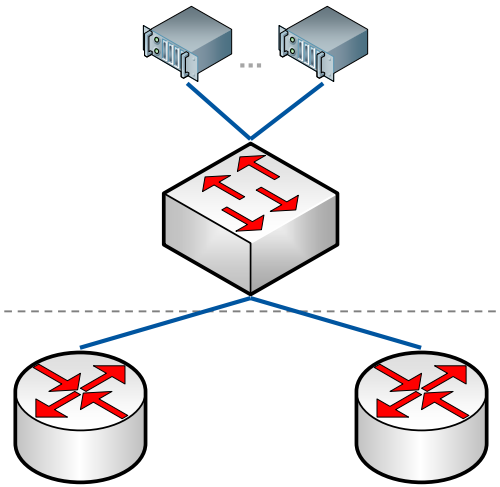



- One broadcast domain with one/several switches and one router
- Large majority of the deployed services (e.g. CPU servers)

- No router redundancy
- No switch redundancy

# Distribution network – current option 2

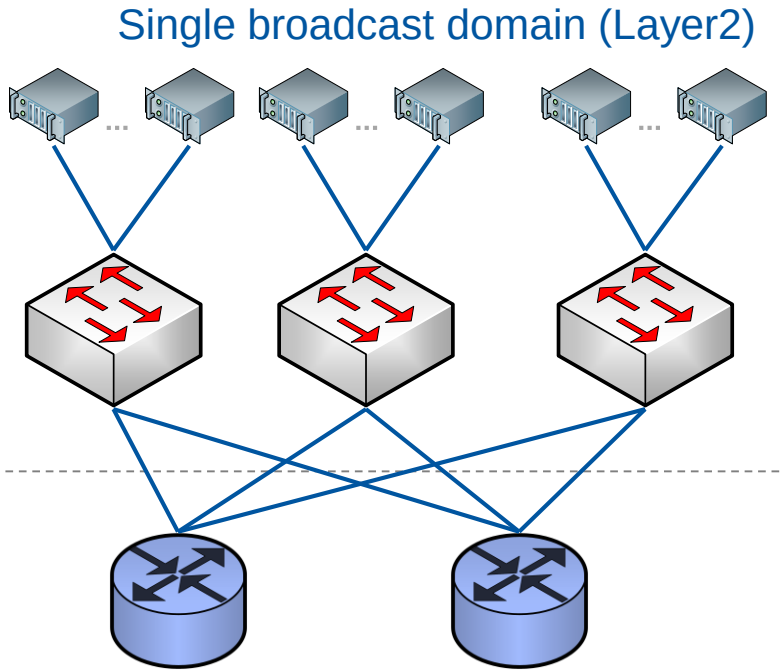
Single broadcast domain (Layer2)



- One broadcast domain with one switch and two routers
- Specific service (e.g. DFS, or DNS)
- **Limited to one switch per broadcast domain**

- Router redundancy
- Backup link is passive (vrrp)
- No switch redundancy

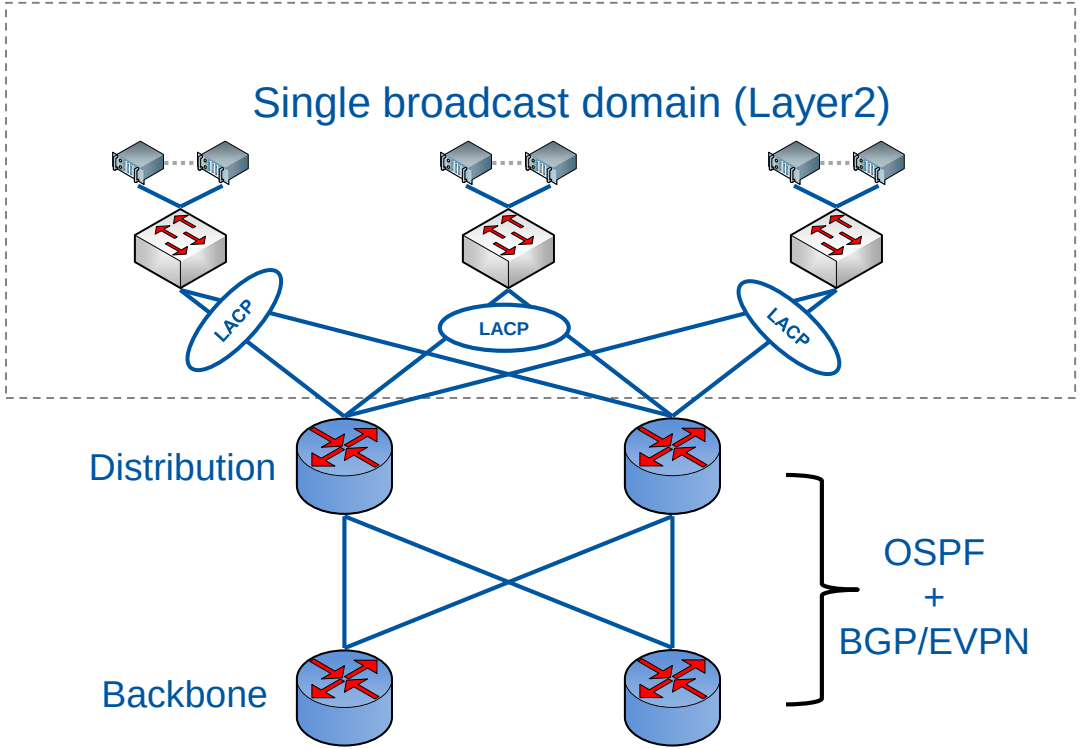
# Distribution network – New setup




- Router redundancy
- NO limitation of 1 switch per broadcast domain
- All links active
- Keep existing Layer2 ToR switches

- Becomes the standard for all network services (homogeneous setup)
- Simplifies network operations and brings more overall redundancy

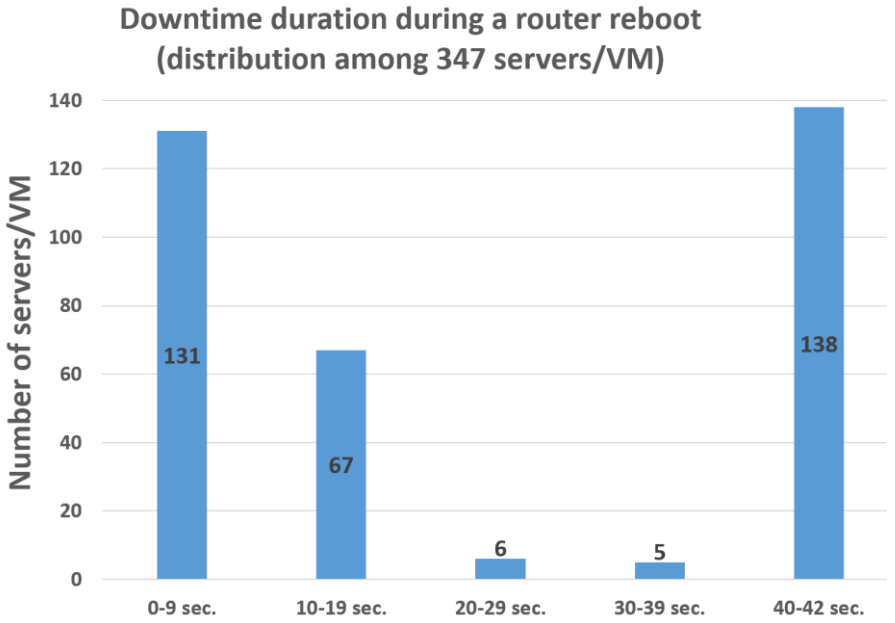
# Distribution network – New setup




- Between Backbone and distribution routers:
  - OSPF underlay (identical to former setup)
  - BGP/EVPN overlay
  - Backbone routers used as route reflectors
- On distribution routers:
  - VxLAN L3 gateway (=default gateway for servers, similar to former setup)
  - VxLAN ESI (Ethernet Segment ID) used to allow LACP across two different router (similar to MC-LAG, but providing more flexibility and using a standard protocol)
  - 1 broadcast domain = 1 VxLAN
  - 1 switch = 1 ESI
- It enables us to migrate 400+ existing basic Layer2 switches and provide them with router redundancy

# Distribution network – New setup

- Primary tests
  - 7 switches (HP and Brocade)
  - 4 broadcast domains (VLAN/VxLAN)
  - 340+ servers/VM
- New services deployed with this setup
  - 18 switches
  - 17 broadcast domains (VLAN/VxLAN)
  - 2x40Gbps uplink for 10Gbps switches
  - 2x100Gbps uplinks for new 25Gbps switches




Acceptable, but study on-going to improve these results

# Challenges and issues

# Challenges and issues

- Optic compatibility between old and new devices (CFP-100G-SR10 vs QSFP28-100G-SR4, etc…)
- New platform and new OS
  - Time to learn
  - Configuration tools to be adapted (still on-going)
  - New bugs…
- New protocols to understand and test (BGP/EVPN + VxLAN)
- Working out the most appropriate design given on our constraints and requirements

# Challenges and issues

From….

```
interface ve XYZ
 ip helper-address A.B.C.D
 ipv6 dhcp-relay destination 2001:1458:xxxx:yyyy::zz
 ipv6 dhcp-relay include-options client-mac-address
```

To….

```
forwarding-options {
    dhcp-relay {
        dhcpv6 {
            relay-agent-option-79;
            group relay6-srv {
                active-server-group dhcp6-srv;
                interface irb.XYZ;
            }
            server-group {
                dhcp6-srv {
                    2001:1458:xxxx:yyyy::zz;
                }
            }
            active-server-group dhcp6-srv;
        }
        forward-only;
        server-group {
            dhcp-srv {
                A.B.C.D;
            }
        }
        active-server-group dhcp-srv;
        group relay-srv {
            active-server-group dhcp-srv;
            interface irb.XYZ;
        }
    }
}
```

# Next steps

# Next steps

- Automate router configuration with our configuration tools
- Migrate existing ToR switches to Juniper routers (with dual router redundancy based on VxLAN)
- Tungsten Fabric (SDN)
  - On-going work on features managed by the Hypervisors
  - Interconnect the “cloud” to the “real” network using a dedicated router acting as an External Gateway in the Tungsten Fabric setup.

# Q&A


CERN