IPv6-Only Backbone

The journey from a single datacenter to a multi-datacenter backbone without using a single IPv4 for the core routing.

MAEHDROS
INTERNET SERVICES

Presentation of Maehdros

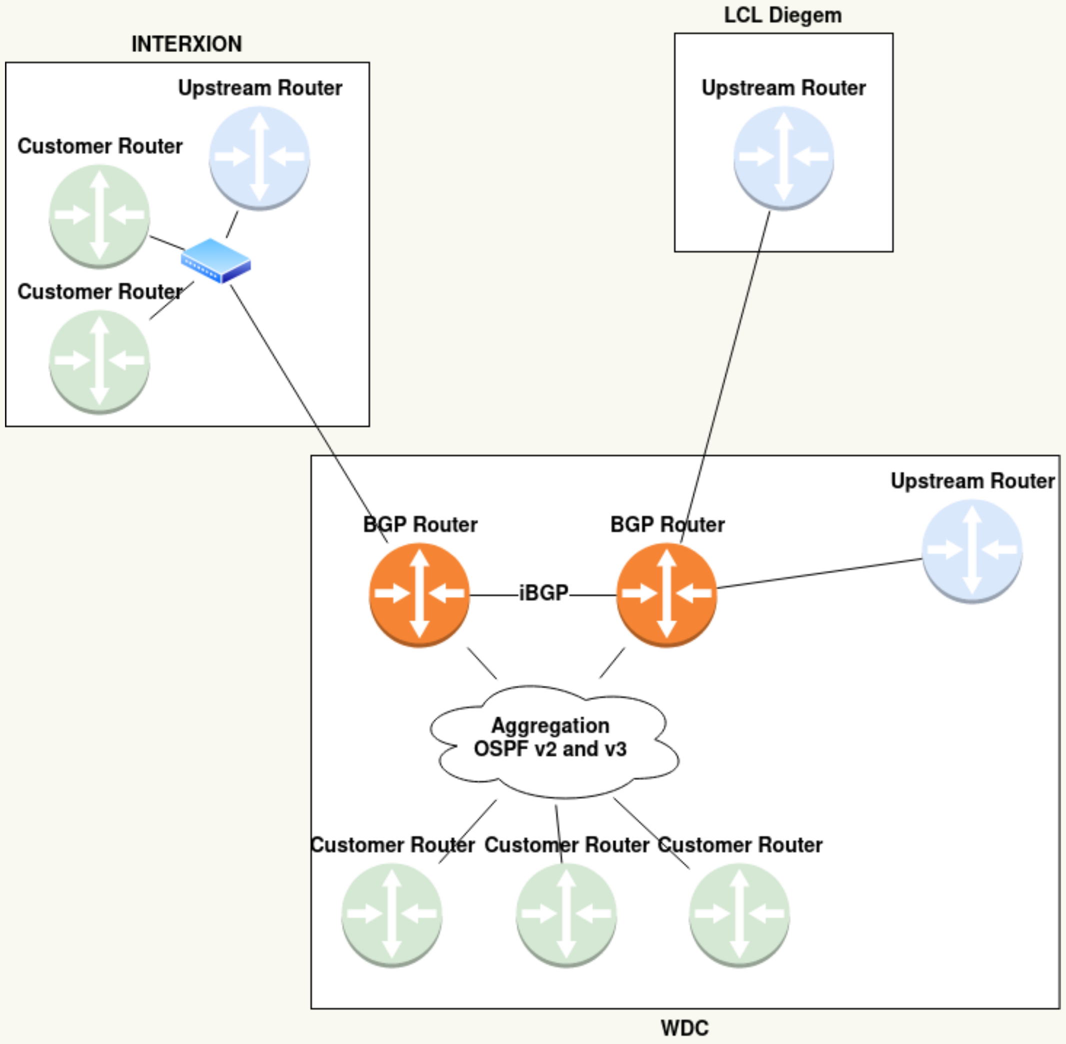
- Hosting company since 2004
  - Mostly Managed Services
- ISP Since 2009
- RIPE : ORG-MS76-RIPE
- AS Number 49677

- First Hosting Provider to officialy support IPv6 for customer (Jan 2011) Thanks to Eric Vyncke and Olivier Bonaventure
- Relatively small : Extremely small IPv4 range allocated by RIPE
- IPv4 allocated : one **/21** and a **/22** (3072 ip addresses)
- IPv6 allocated : **2a02:21d0::/32**

Structure of Network

- Hosting Infrastructure means :
  - Mostly outgoing traffic
  - Larges subnets /25 with few waste of address
- ISP Infrastructure means :
  - Mostily incoming traffic (that balances with Hosting Traffic)
  - Many very short linknet, high rate of wasted address (Previously at least 2 addresses were wasted per customer)

Network Topology : Before

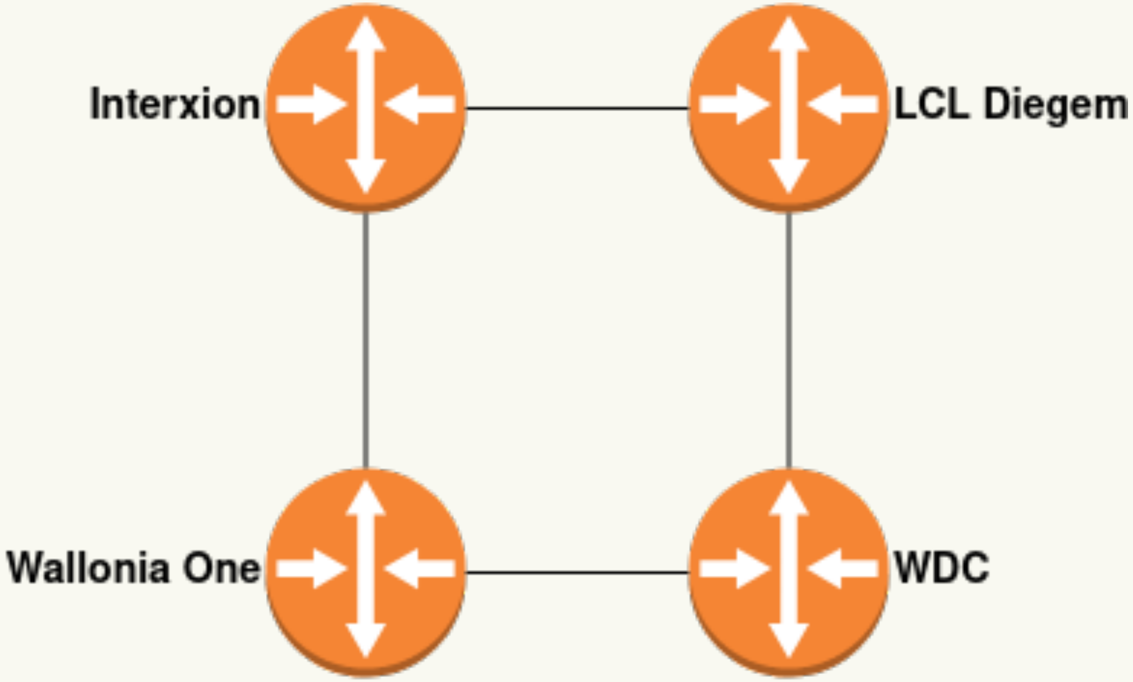


- Globally redundant,but not that much (mostly for Hosting)
- ISP Customer connexion not redundant (in case of ITX failure)
- No Layer 3 outside WDC (all routers are in WDC)

Objectives

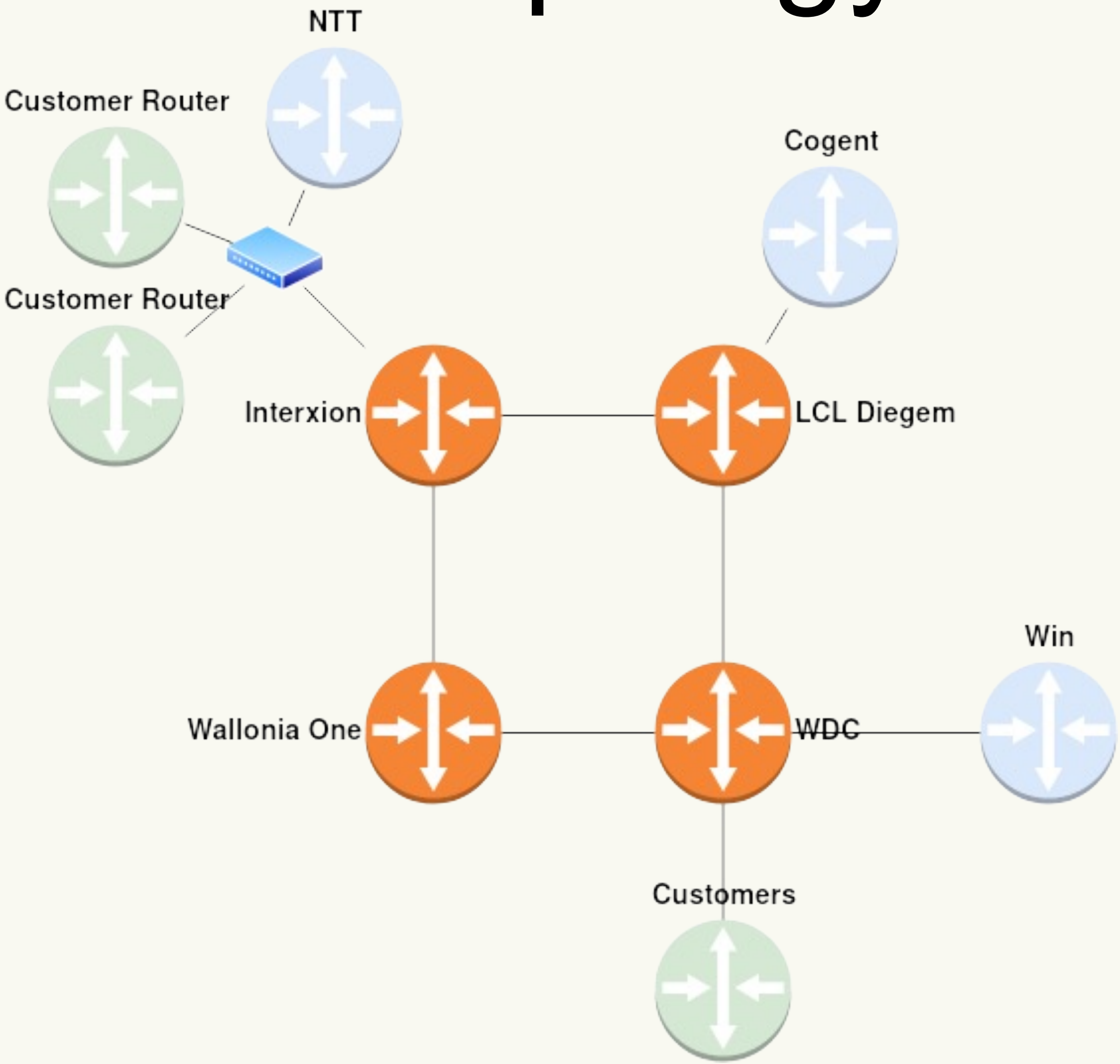
- Increase Redundancy
  - Incidents on the backbone should not impact our ISP customers
- Allow for easier hosting of infrastructure
  - Layer 3 capabilities in every datacenter
- Better scalability
  - Allow multiple datacenter to get connectivity from layer 1 providers (Sofico, Eurofiber, Voo, ...)
  - Waste as few IP addresses as possible

First idea for the topology



- Circular topology
- Support for one link failure
- All links have equal cost

First idea for the topology



Step 1 : OSPF v3 Area 0

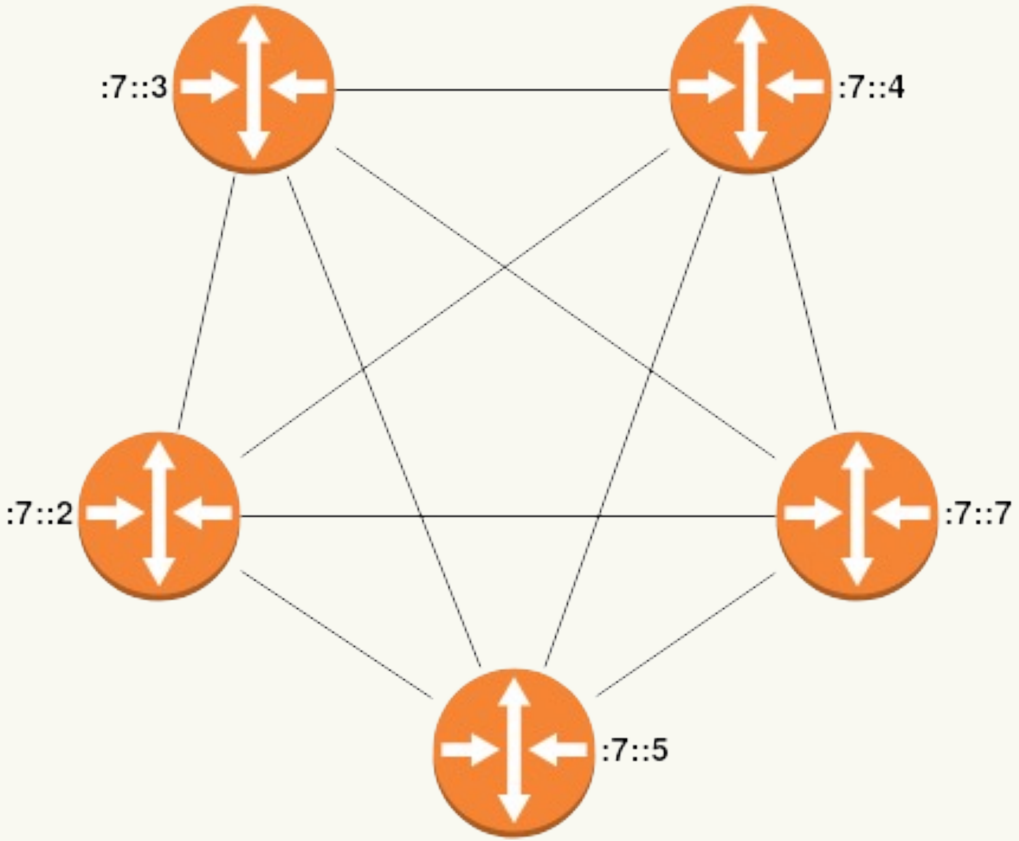
- Each Router in the backbone is an OSPFv3 router present in the area 0
- IPv6 only router ID are x.x.x.x, where x is just the router number
- IPv6 and IPv4 router use their highest IPv4 address as router-id

- Each router has at least two neighbors
- Each router has at least three interfaces in area 0 (Loopback + backbone links)
- We use it only to distribute internal IPv6 routes, required for BGP sessions reachability

```
set protocols ospfv3 area 0.0.0.0 interface 'lo'
set protocols ospfv3 area 0.0.0.0 interface 'eth6'
set protocols ospfv3 area 0.0.0.0 interface 'eth11'
set protocols ospfv3 area 0.0.0.0 interface 'eth4.604'
set protocols ospfv3 parameters router-id '3.3.3.3'
```

Step 2 : BGP over the OSPFv3 routers

- BGP Topology : Full mesh
- Prefix : 2a02:21d0:7::/64
- Total of 5 routers, 20 sessions, still human manageable
- Within the backbone, sessions are establish to IPv6 peers only, between public **loopback addresses (/128)**
- VyOS 1.3: capability extended-nexthop allow for mixed routing table (RIB and FIB) – RFC 5549



Step 2 : BGP over the OSPFv3 routers

- IPv4 BGP routes have the loopback of their originating router as the next-hop
- OSPFv3 is used to find the best path to the next-hop.
- In case of link failure, OSPFv3 reconverges before BGP notices the failure

```
B>  1.0.166.0/24 [200/0] via 2a02:21d0:7::7 (recursive), 6d03h47m
  *                      via fe80::a236:9fff:fe2c:7020, eth11, 6d03h47m
B>* 1.0.167.0/24 [20/1531] via 194.53.172.33, eth4.501, 3d04h54m
B>  1.0.168.0/24 [200/0] via 2a02:21d0:7::7 (recursive), 6d03h47m
  *                      via fe80::a236:9fff:fe2c:7020, eth11, 6d03h47m
B>  1.0.169.0/24 [200/0] via 2a02:21d0:7::7 (recursive), 6d03h47m
  *                      via fe80::a236:9fff:fe2c:7020, eth11, 6d03h47m
B>  1.0.170.0/24 [200/0] via 2a02:21d0:7::7 (recursive), 6d03h47m
  *                      via fe80::a236:9fff:fe2c:7020, eth11, 6d03h47m
B>  1.0.171.0/24 [200/0] via 2a02:21d0:7::7 (recursive), 6d03h47m
  *                      via fe80::a236:9fff:fe2c:7020, eth11, 6d03h47m
```

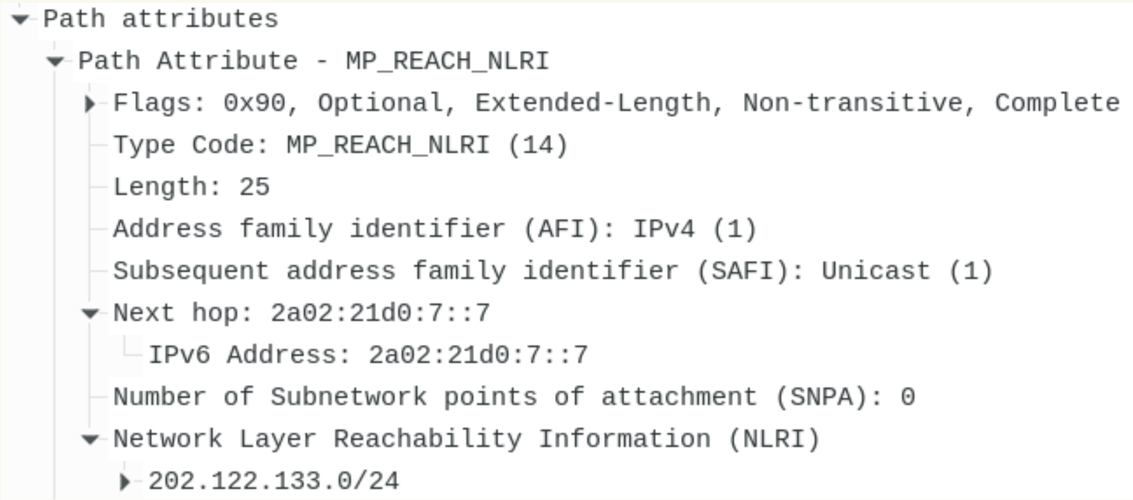
Step 2 : Show ip bgp summary

Session uptime is more than a year

```
Neighbor              V          AS MsgRcvd MsgSent   TblVer  InQ OutQ  Up/Down State/PfxRcd
194.53.172.1            4       5406 27141742  585958        0    0    0 11w4d19h        91907
194.53.172.2            4       5406 27303679  586074        0    0    0 16w1d05h        91907
2a02:21d0:7::2            4      49677  588703 506918702        0    0    0 13w4d05h            4
2a02:21d0:7::4            4      49677 170702966 507041776        0    0    0 01y05w6d            0
2a02:21d0:7::5            4      49677 168744538 507041776        0    0    0 01y05w6d       480483
2a02:21d0:7::7            4      49677 173879396 507041776        0    0    0 01y05w6d       480491
81.20.71.65             4       2914 311346479  581709        0    0    0 21w5d09h       874740
2001:728:0:5000::141      4       2914 211319327  581719        0    0    0 21w5d09h NoNeg
2001:7f8:26::a500:5406:1 4       5406 202581752  617253        0    0    0 11w4d19h NoNeg
2001:7f8:26::a500:5406:2 4       5406 264930631  630901        0    0    0 12w0d19h NoNeg
```

Step 3 : RIB

- Routes are exchanged through BGP-MP sessions.
- BGP RIB is not enough to route packets without OSPFv3
- Extended next-hop basically only allows BGP to send an **IPV6 next-hop** for an IPV4 prefix

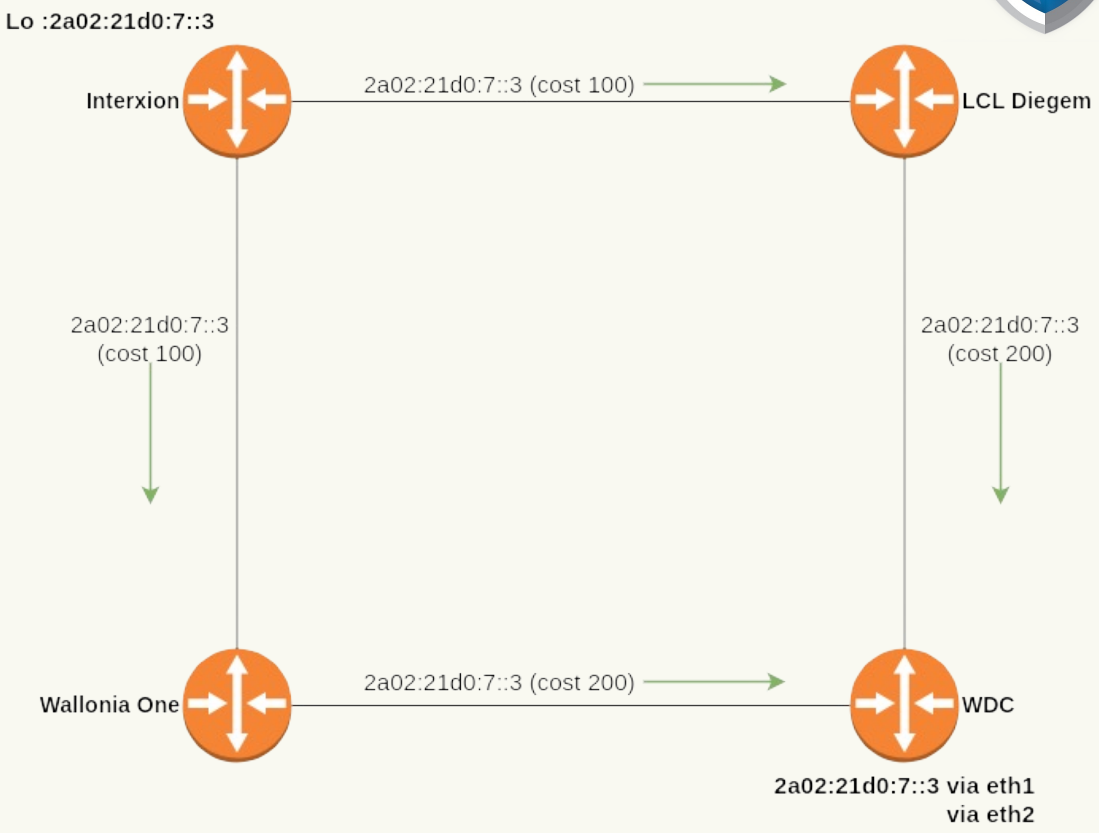


```
▾ Path attributes
  ▾ Path Attribute - MP_REACH_NLRI
    ▸ Flags: 0x90, Optional, Extended-Length, Non-transitive, Complete
      Type Code: MP_REACH_NLRI (14)
      Length: 25
      Address family identifier (AFI): IPv4 (1)
      Subsequent address family identifier (SAFI): Unicast (1)
    ▾ Next hop: 2a02:21d0:7::7
        IPv6 Address: 2a02:21d0:7::7
      Number of Subnetwork points of attachment (SNPA): 0
    ▾ Network Layer Reachability Information (NLRI)
      ▸ 202.122.133.0/24
```

Example of BGP UPDATE for an IPV4 route
Notice : Adress family is for an IPv4 route (202.122.133.0/24) and the next hop is an IPv6 host.

Step 3 : RIB

- OSPFv3 propagates reachability information for BGP sessions and routing
- Equal cost on all links means load-balancing is possible
- Instant failover to the second route in case a link goes down
    - Improved by : BFD protocol and Uplink Failure Detection



Caveats

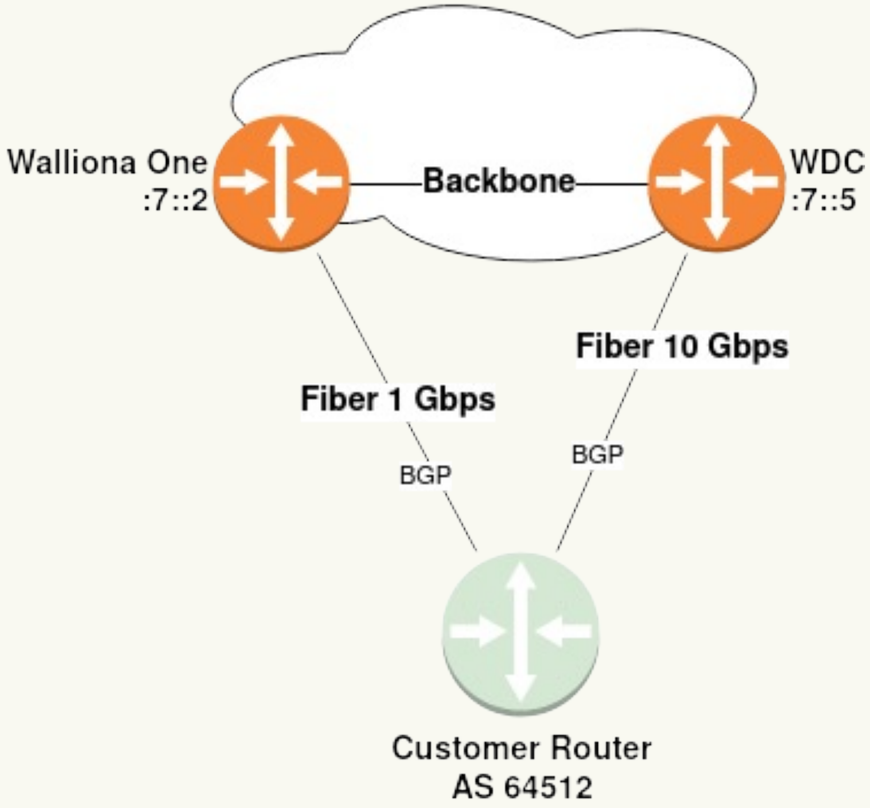
- Backbone router are reachable in IPv6 only. (Can be problematic when being on a mobile connexion without IPv6)
- Traceroute rows fails on the backbone routers (and customers can see it)

Advantages

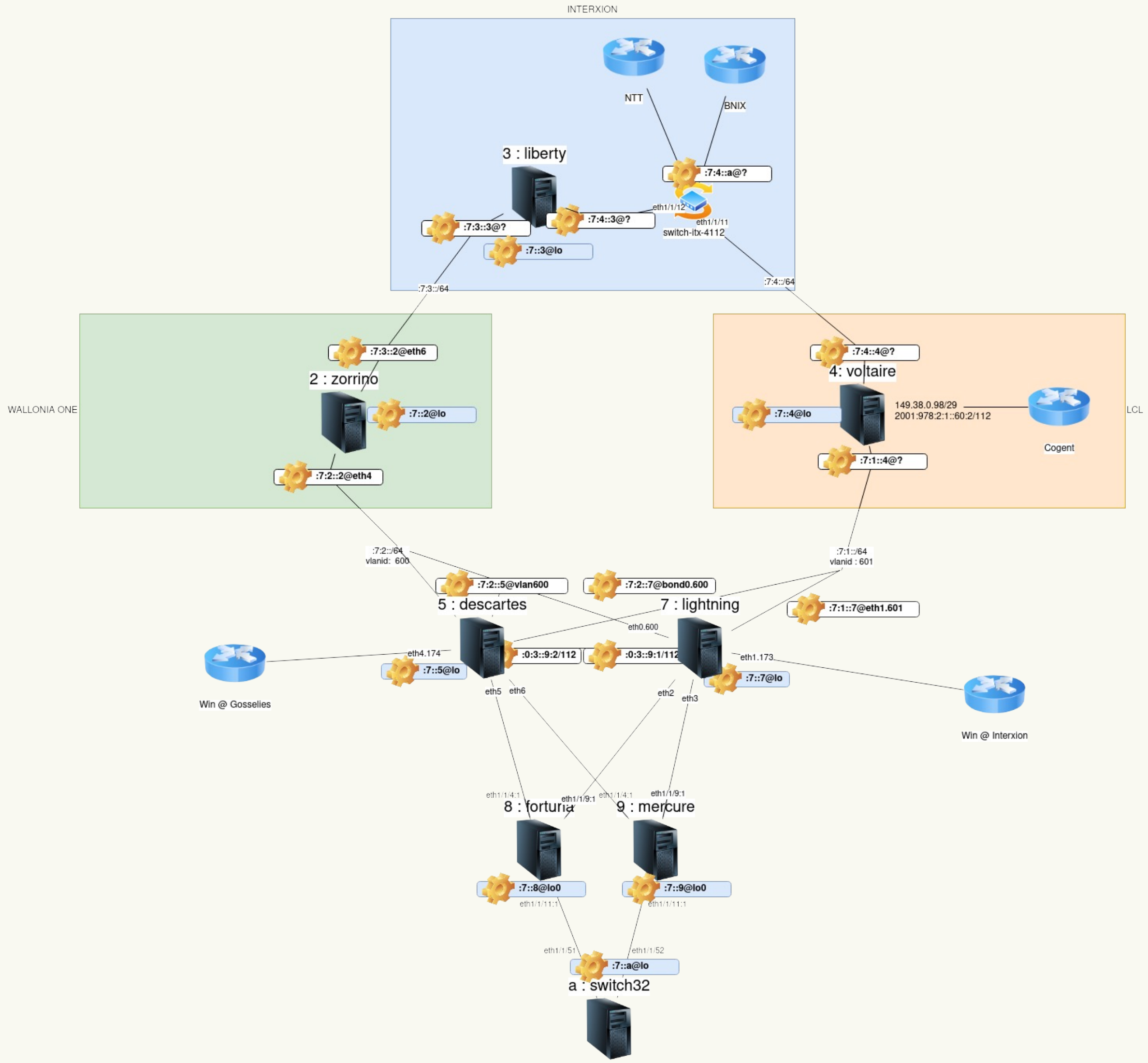
- No wasted IPv4 address
- Fewer protocol to manage (no OSPFv2)
- Redundancy is easy to build, because IPv6 subnet size allows for multiple routers within the same segment

Side advantages

- ISP customer that are physically redundant use the same configuration with private AS to exchange routes between two datacenter
- When used in conjunction with unnumbered BGP, you can even get connected with no public IP address whatsoever (not even IPv6). You basically get IP transit over no address



Real Backbone


INTERXION
NTT
BNIX
3 : liberty
:7:4::a@?
:7:3::3@?
:7:4::3@?
eth1/1/12
eth1/1/11
switch-itx-4112
:7::3@lo
:7:3::/64
:7:4::/64
WALLONIA ONE
:7:3::2@eth6
2 : zorrino
:7::2@lo
:7:2::2@eth4
:7:4::4@?
4: voltaire
:7::4@lo
149.38.0.98/29
2001:978:2:1::60:2/112
Cogent
:7:1::4@?
LCL
:7:2::/64
vlanid: 600
:7:1::/64
vlanid : 601
:7:2::5@vlan600
:7:2::7@bond0.600
5 : descartes
7 : lightning
:7:1::7@eth1.601
eth0.600
eth4.174
:0:3::9:2/112
:0:3::9:1/112
eth1.173
:7::5@lo
:7::7@lo
Win @ Gosselies
Win @ Interxion
eth5
eth6
eth2
eth3
eth1/1/4:1
eth1/1/9:1
eth1/1/4:1
eth1/1/9:1
8 : fortuna
9 : mercure
:7::8@lo0
:7::9@lo0
eth1/1/11:1
eth1/1/11:1
eth1/1/51
eth1/1/52
:7::a@lo
a : switch32