e-Jornadas 2020

PORTUGAL INCoDe.2030

Novas arquiteturas tecnológicas de redes

Patrocinadores **Platina**

Microsoft cisco warpcom paloalto Check Point SOFTWARE TECHNOLOGIES LTD ELSEVIER

Patrocinadores **Ouro**

ACCUCOMS

Patrocinadores **Prata**

AIP Publishing IOP Publishing SPRINGER NATURE emerald PUBLISHING Westcon

Organização

FCT Fundação para a Ciência e a Tecnologia Computação Científica Nacional FCCN

Agenda

In this session we will:

- Make an introduction to network models evolution
- Display all the building blocks for a scalable transport network
- Show how to improve Control-Plane scalability and intelligence
- Illustrate how our access rings will behave in access topologies
- Design the popular services and new ones
- Exemplify how our traffic will flow in different scenarios
- …and we will answers your questions!

RCTS 100 Project
RCTS 2.0

Intro

Network Models Evolution

RCTS 1.0 -> We're here!

IP

- Simple
- Not scalable

MPLS

- Increased complexity
- Scalable
- Service driven

EPN 4.0 Unified/Seamless MPLS

- Stable Core
- IGP Demarcation
- BGP-LU
- End to End Services
- Very complex

EPN 5.0 Agile Carrier Ethernet

- New transport
- Reduced complexity (no LDP & RSVP)
- Fast Convergence built in natively
- Potential to be programmable

RCTS 2.0 -> …heading here!

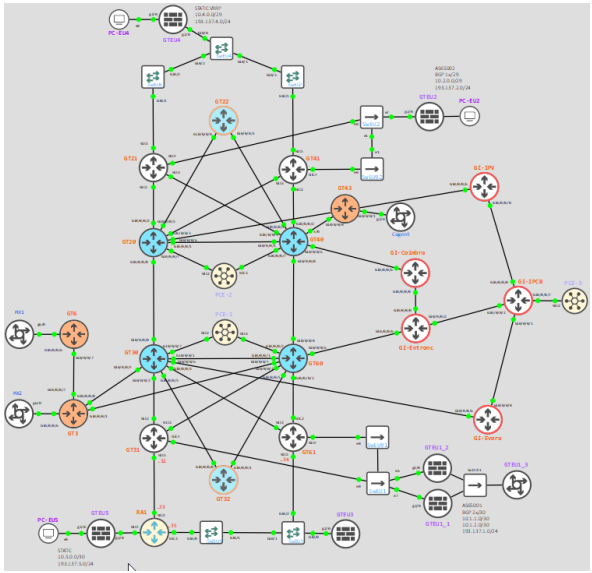
Compass Metro Fabric

- Reduced complexity (no BGP-LU)
- Inter domain reachability achieved trough a controller/path computation engine
- Truly programmable

Step by Step

This is where we want to go, but how do we get there?

- Design a Architecture (and document… HLD)
- Acquire equipment
- Prepare a TestBed
- Define the Services (and document… LLD)
- Prepare/Plan the migrations (divided by phases)
- Execute
- Verify
- Monitor it
- Automate it



RCTS 100 Project
RCTS 2.0

Transport

RCTS 2.0 Building Blocks

- **Transport Layer** based on **Segment Routing** as Unified Forwarding Plane
- **Service Layer** for Layer 2 (EVPN) and Layer 3 VPN services based on **BGP as Unified Control Plan**
- **SDN - Segment Routing Path Computation Element (SR-PCE)** to provide **simple** and **scalable** inter-domain transport connectivity and Traffic Engineering and Path control
- **Automation and Analytics**
  - NSO for service provisioning with Netconf/YANG data models
  - Streaming Telemetry to enhance visibility and monitoring

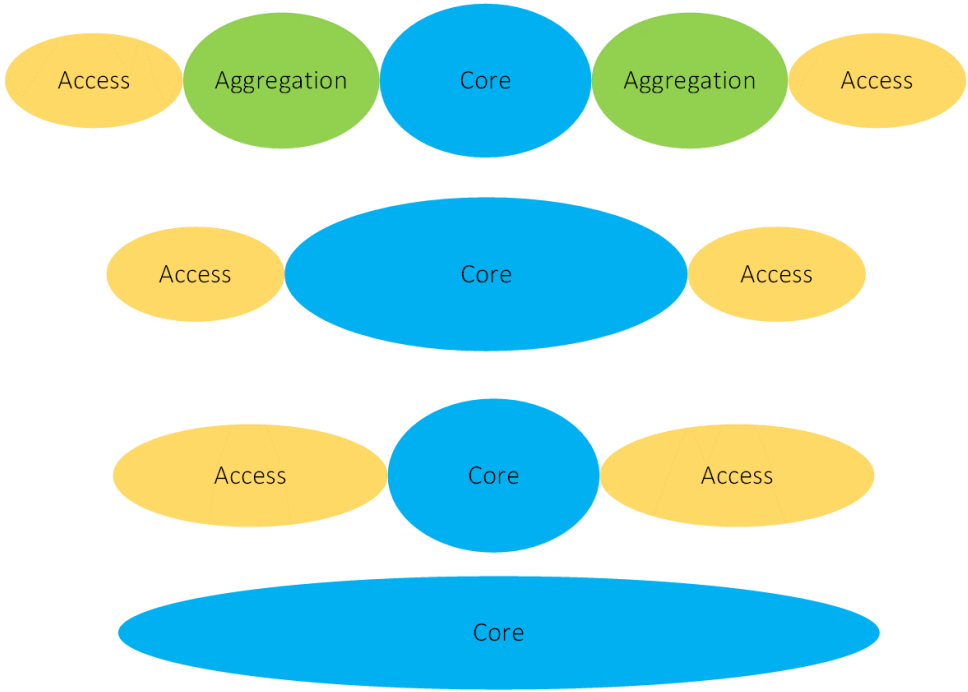
Transport – Design Options

The Design option are:

- Full Model containing all layers. Suitable for large SP.
- Extended Core. Core and access layers with a reasonable number of devices in the core.
- Standard Core/Access. Very simple core with only the main PE´s and P routers.
- Core Only. Only 1 IGP period.

To provide network scale and stability, the Architecture Fabric is structured into multiple IGP Domains. We believe the most suitable for RCTS is the Extended Core.

We will build and present options for both Extended and Core Only Architectures.

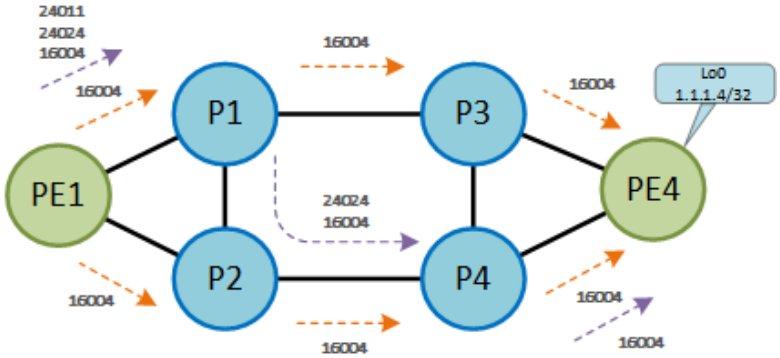


SR Mechanic – SID´s 101

- Configured under IGP routing protocol
- Requires: Enabling SR & Configuring Prefix-SID
- Two basic building blocks distributed by IGP:
  - Prefix Segments
  - Adjacency Segments

SRGB 16,000 – 23,999

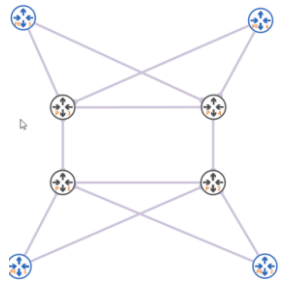
Adj-SID´s >= 24000



- Result: No LDP Needed for label distribution && much easier stack of labels to manage

Intra-Domain Routing and Forwarding

- The Fabric is based on a fully programmable transport. The foundation technology used in the transport design is SR with a MPLS based Data Plane in RCTS 2.0 and a IPv6 based Data Plane (SRv6) in a future RCTS 3.0
- We propose the use of IS-IS as the Core IGP protocol and different instances of ISIS as the Access IGP protocol
- In the research made, it looks SR is made for ISIS
- Segment-Routing embeds a simple Fast Re-Route (FRR) mechanism known as Topology Independent Loop Free Alternate (TI-LFA)



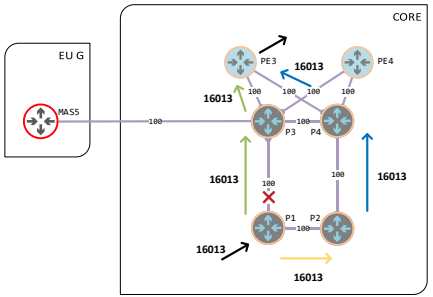
TI-LFA – How good it is?

RCTS 2.0

LFA-Only

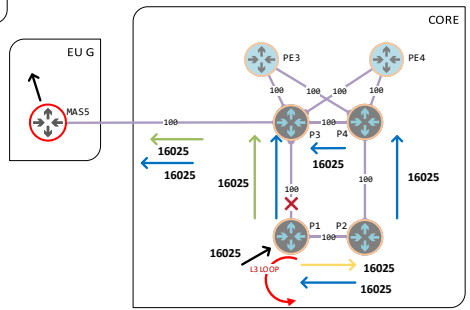
Works with “regular” IGP. Path is protected by “backup” path.

Traffic from P1 to PE3 (active path + backup path)



Where does it fail?

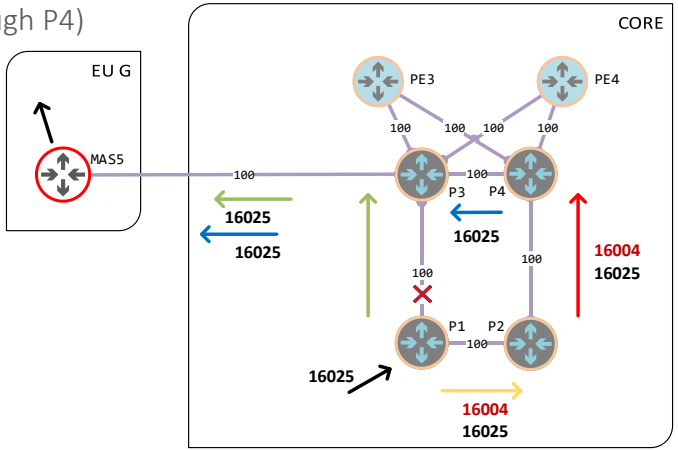
Traffic from P1 to MAS5 have a good backup path but NH P2 thinks the “broken” is still their best!



TI-LFA

Set another MPLS label at ingress P1 and P2 forwards labels to P4 instead of MAS5.

Traffic from P1 to MAS5 (active path + backup path through P4)



Just enable SR and configure TI-LFA under each IGP interface.

RCTS 100 Project
RCTS 2.0

Transport – RR's and SR-PCE's

PCE Controller Summary

Segment Routing Path Computation Element (SR-PCE):

- Runs as a features in a IOS-XR node
- Collects topology from BGP, ISIS, OSPF and BGP Link State
- Computes Shortest, Disjoint, Low Latency, and Avoidance paths
- Deploys tunnels == sends labels stacks
- North Bound interface with applications via REST API

Feeding PCE/XTC

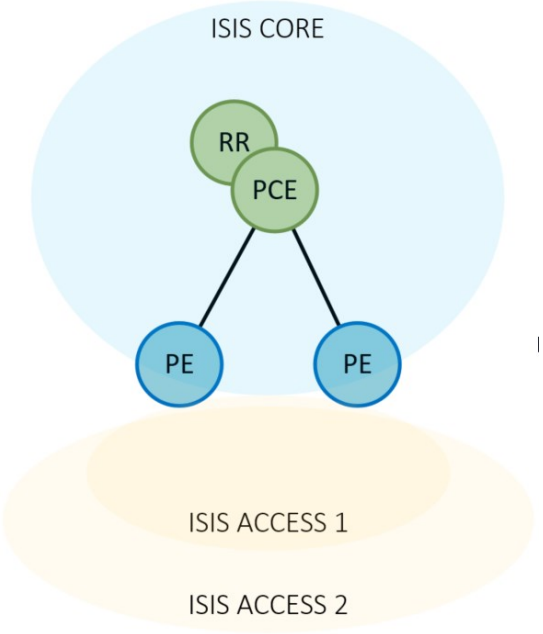
Feeding Link State Topology - Core:

1

- Activate the PCE feature
- PCE participates in the Core IGP
- Since the PCE already has the topology of all the Core, we just need to send the IGP information to BGP inside the PCE

Feeding Link State Topology - Access:

- Establish a PCE(server)-PCC(client) session 2
- Send topology info from the IGP to BGP inside each PE 3
- In the BGP session with the RR´s send the Topology 4



Humm, Controller and SRTE must be hard o.O

1

PCE/RR

```
pce
 address ipv4 193.136.5.11
router isis 1930
 distribute link-state
router bgp 1930
 address-family link-state link-state
```

2

PCC/PE

```
segment-routing
 traffic-eng
  pcc
   source-address ipv4 172.25.1.1
   pce address ipv4 193.136.5.11
    precedence 100
   !
   pce address ipv4 193.136.5.12
    precedence 200
```

3

PCC/PE

```
router isis ACCESS-1
 distribute link-state
```

4

PCC/PE

```
router bgp 1930
 neighbor 193.136.5.11
  address-family link-state link-state
```

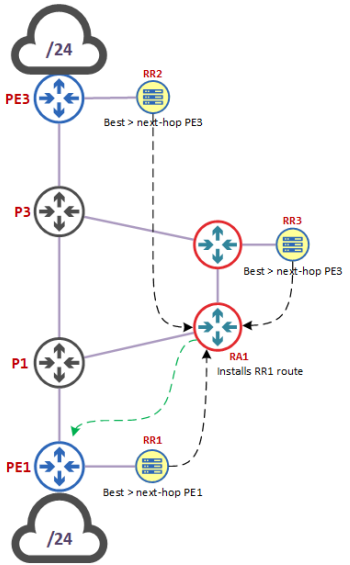
PCE/RR

```
router bgp 1930
 neighbor 172.25.1.1
  address-family link-state link-state
```

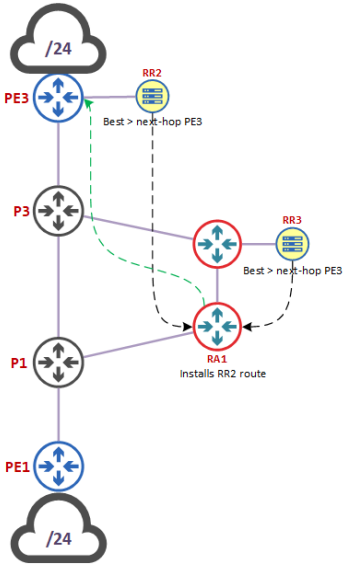
RR´s – Placement

The Problem:

- When a RR learns the same prefix from 2 neighbors and all “usual” attributes are equal then the lowest IGP metric to the neighbor comes into play
- The RR install´s the 2 routes in the BGP table but announces to a third router the best, which?
- The route which was learned from the “closest” neighbor <u>from the RR´s point of view!!!</u>
- Which leads to unideal traffic forwarding



65001, (received & used)
172.25.1.1 (metric 3) from 172.20.0.1 (172.20.0.1)
Origin IGP, metric 0, localpref 100, valid, internal, best
Originator: 172.25.1.1, Cluster list: 172.20.0.1
65001
172.25.3.3 (metric 4) from 172.20.0.2 (172.20.0.2)
Origin IGP, metric 0, localpref 100, valid, internal
Originator: 172.25.3.3, Cluster list: 172.20.0.2
65001
172.25.3.3 (metric 4) from 172.20.0.3 (172.20.0.3)
Origin IGP, metric 0, localpref 100, valid, internal
Originator: 172.25.3.3, Cluster list: 172.20.0.3



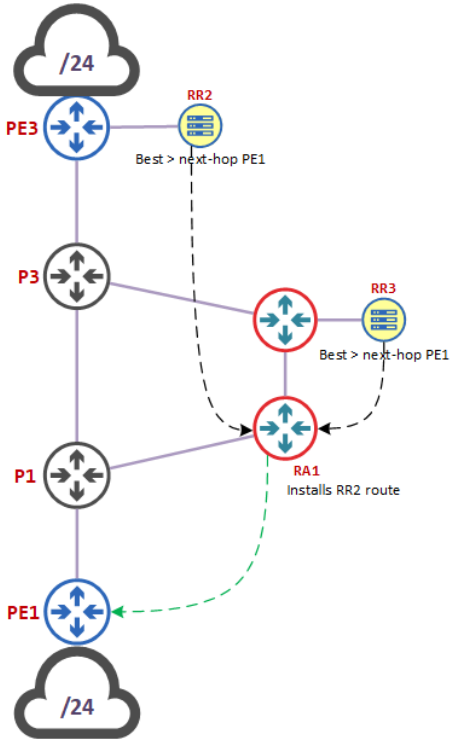
65001, (received & used)
172.25.3.3 (metric 4) from 172.20.0.2 (172.20.0.2)
Origin IGP, metric 0, localpref 100, valid, internal, best
Originator: 172.25.3.3, Cluster list: 172.20.0.2
65001
172.25.3.3 (metric 4) from 172.20.0.3 (172.20.0.3)
Origin IGP, metric 0, localpref 100, valid, internal
Originator: 172.25.3.3, Cluster list: 172.20.0.3

RR´s – Optimal Route Reflection

With the feature **optimal-route-reflection** active:

- The routes passed from the RR´s to the clients are independent of the IGP metric of the RR´s.
- RR´s pass the best route based on the IGP metric between the source and destination of the traffic.
- In the given example the metric is also equal from RA1 to PE3. In this case the lowest originator/router-id is used.
- Which is also equal. The tie-breaker is the router-id/cluster-id of the RR´s.
- Independently of which of the RR´s route RA1 will install, the traffic will be forward to PE1, the closest router.

```
65001, (received & used)
  172.25.1.1 (metric 3) from 172.20.0.2 (172.20.0.2)
    Origin IGP, metric 0, localpref 100, valid, internal, best
    Originator: 172.25.1.1, Cluster list: 172.20.0.2
 65001
  172.25.1.1 (metric 3) from 172.20.0.3 (172.20.0.3)
    Origin IGP, metric 0, localpref 100, valid, internal
    Originator: 172.25.1.1, Cluster list: 172.20.0.3
```




```
router bgp 1930
address-family ipv4 unicast
  optimal-route-reflection pe1 172.25.1.1
  optimal-route-reflection pe2 172.25.2.2
!
neighbor 172.25.1.1
  use neighbor-group RR-Clients
   address-family ipv4 unicast
    optimal-route-reflection pe1
```

BGP Add-Path

Overrides BGP default behavior of sending only the best route

Simple send more than 1 route (with different NH) for the same prefix.

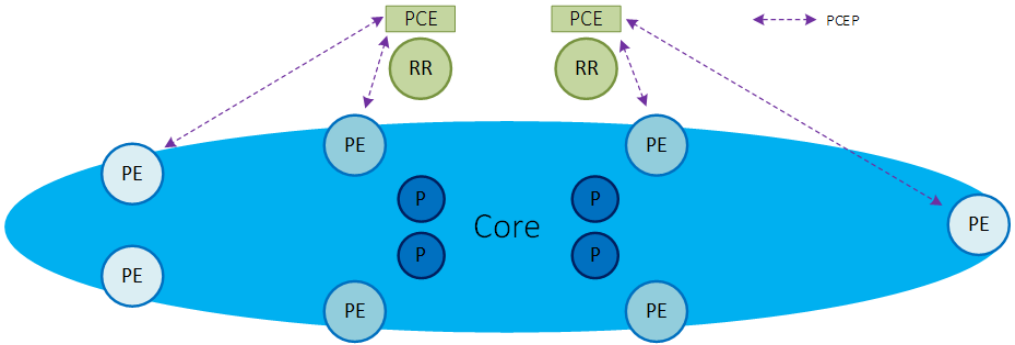
Add-Path vs ORR

ORR:

- Resource intensive in the RR
- Needs network convergence in case of failure

Add-Path:

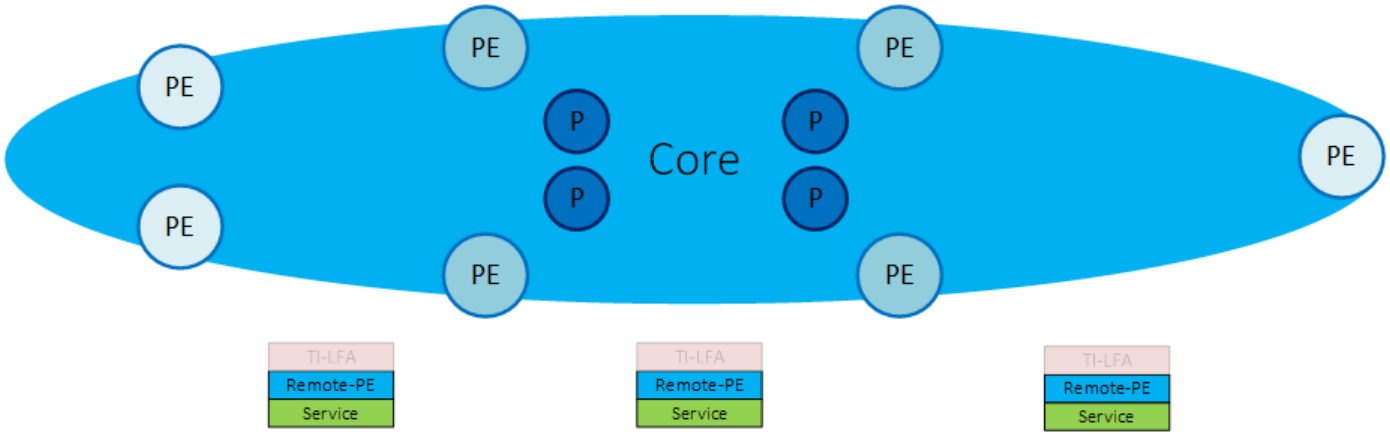
- Doubles the BGP table in PE's per RR
- Faster Convergence in case of failure



RCTS 100 Project
RCTS 2.0

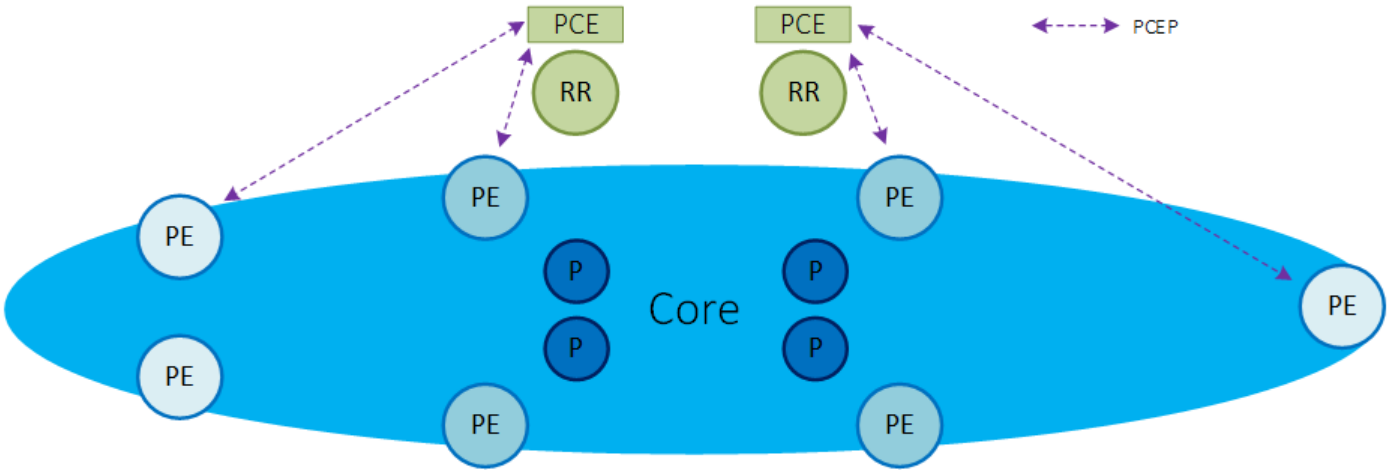
Transport – Core Only Model

Inter-Domain Forwarding… or Not!!!



End to end services natively – no label swapping

SR-TE with PCE

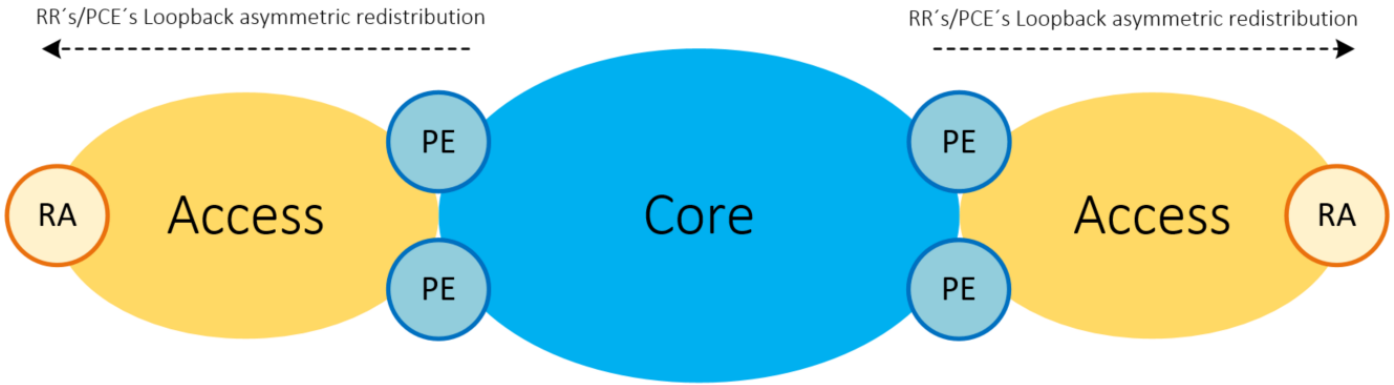


Ability to steer traffic based on several requisites (low latency / custom)

RCTS 100 Project
RCTS 2.0

Transport – Extended Core Model

Inter-Domain Redistribution

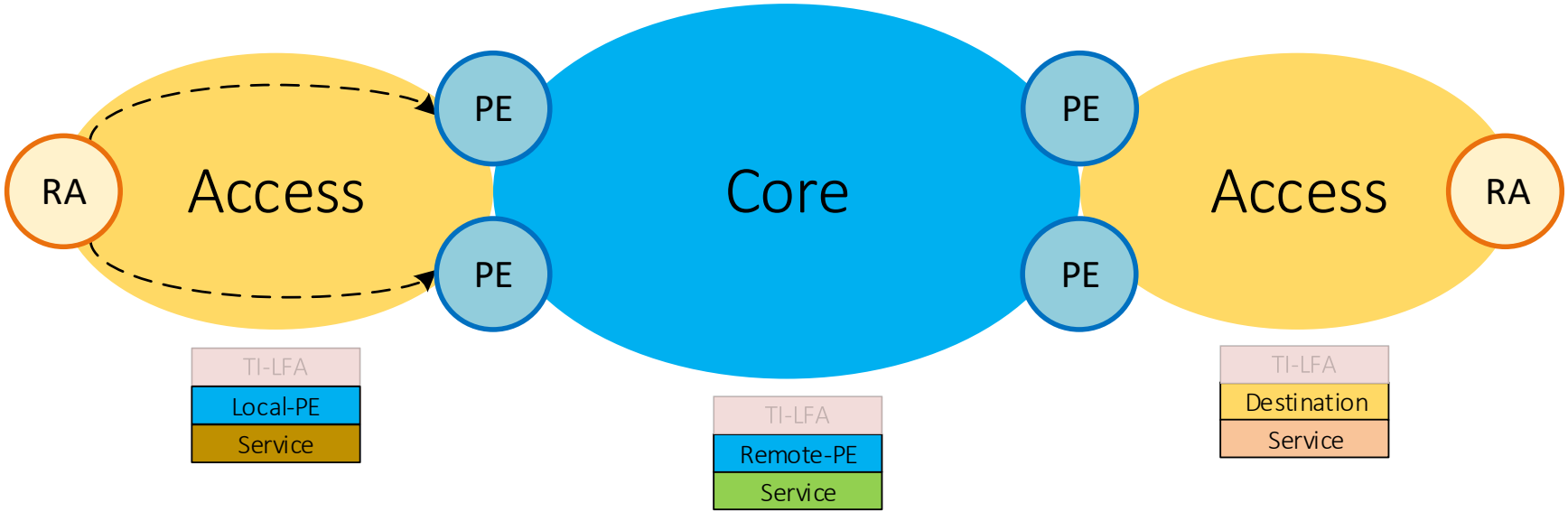


Asymmetric redistribution, thus it won't cause any L3 routing loop in the network.

Always the same address´s, thus does not affect scalability in the Access IGP Domain

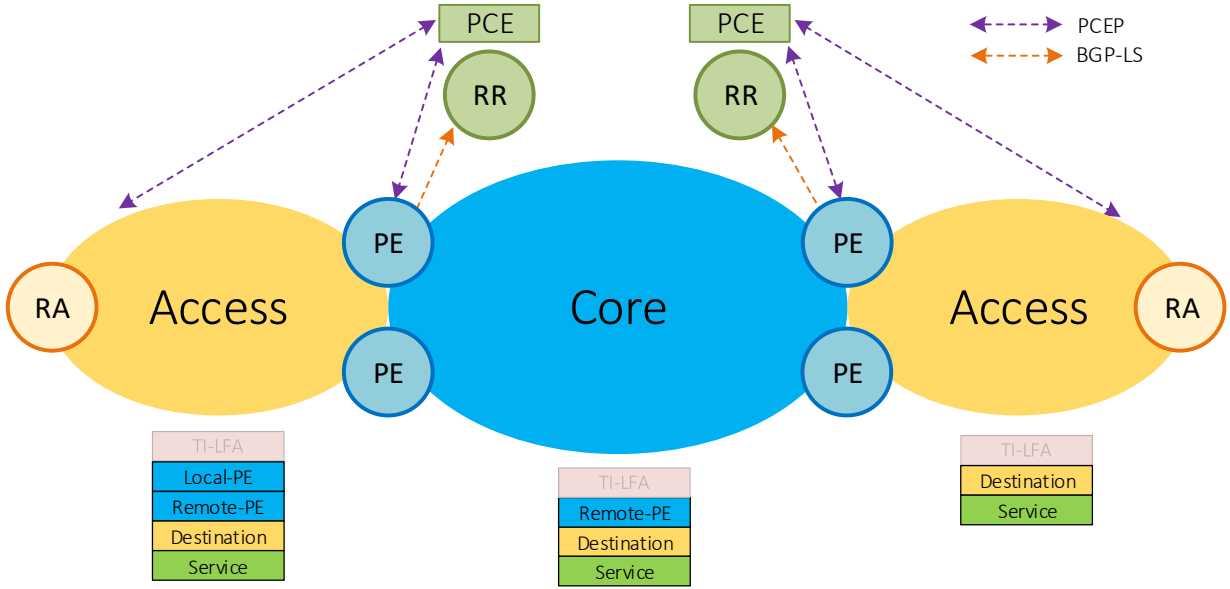
```
Router static
 address-family ipv4 unicast
  193.136.5.8/30 null 0 description RR-PCE-ROUTES
prefix-set RR_XTC-LOOPBACKS
  193.136.5.9/32,
  193.136.5.10/32
route-policy CORE-TO-ACCESS
  if destination in RR_XTC-LOOPBACKS then
    pass
  endif
end-policy
router isis ACCESS-RING-1
 address-family ipv4 unicast
  redistribute static route-policy CORE-TO-ACCESS
```

Inter-Domain Forwarding – no PCE!



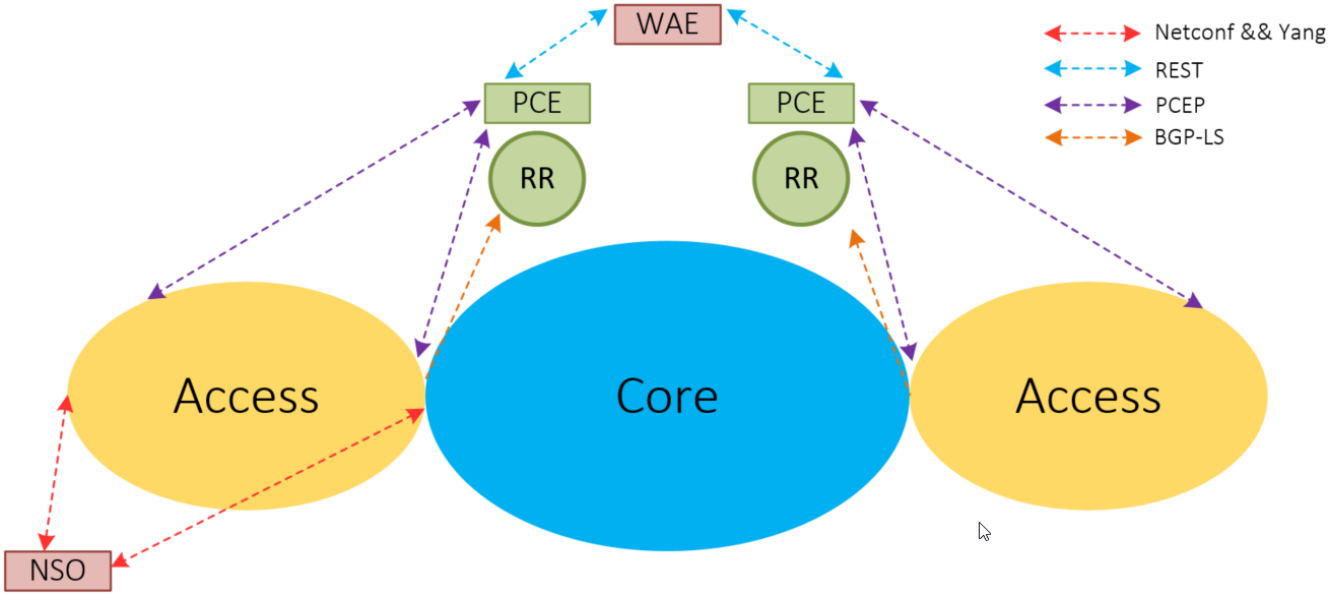
Stitching the services on the Core PE´s with EVPN-VPWS

Inter-Domain Forwarding – PCE!



True end-to-end Services with a centralized controller

Future Network Automation



Model Chosen

Due to management feedback and concerns the model chosen to drive RCTS 2.0 is the CORE Only model.

This doesn’t imply the other model doesn't comply, just that more testing is needed to provide trust and experience.

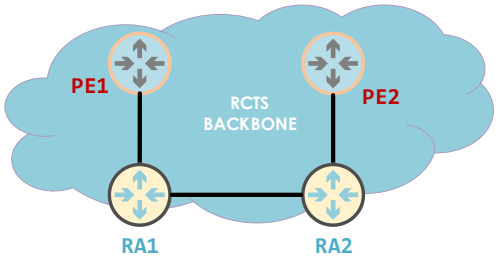
Thus, a ring in the Extended Core model will be deployed in our datacenters, connected to RCTS 2.0 as if it were in production.

RCTS 100 Project
RCTS 2.0

Access

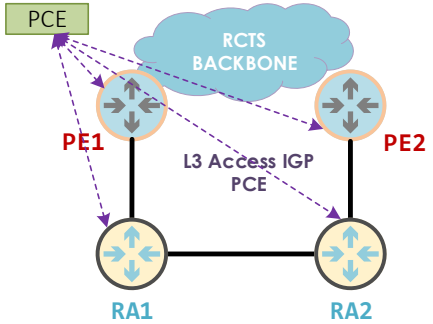
Access Rings – Layer 3 (1)

Core Only



- Simplest and easiest design
- Scalability might become an issue
- Complex and expensive RAs required

Extended Core



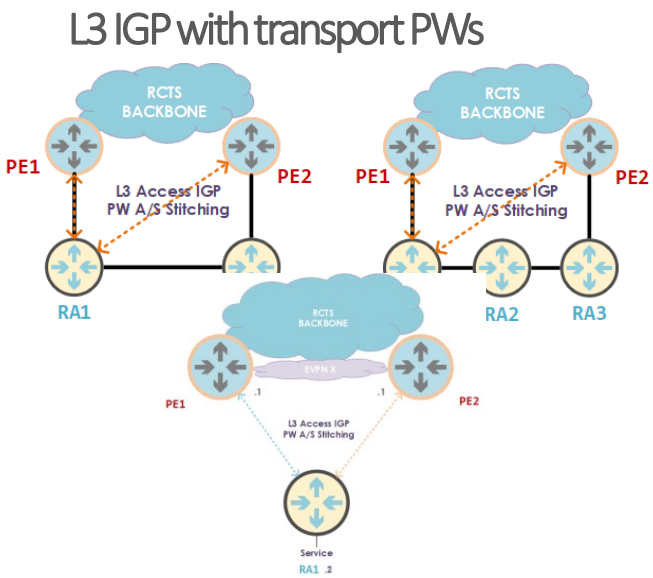
- Still End-to-End services with network segmentation
- RAs need to support Segment Routing and EVPN
- PCEP technology is not mature and does not seem very interoperable

Access Rings – Layer 3 (2)

1) MPLS-LDP with LDP PWs
2) MPLS-RSVP with TE PWs
3) **MPLS-SR with EVPN-VPWS**

As an evolution step, access rings can be deployed with multilayer switches capable to transport services to PEs or even deploy IP services directly:

vlans -> pseudowire
Subinterfaces -> BD or PWHE



- **Topology Independent** – all rings will work as triangles
- Still requires MPLS capable devices
- Complex solution without end-to-end services

Access Rings – Layer 2 (1)

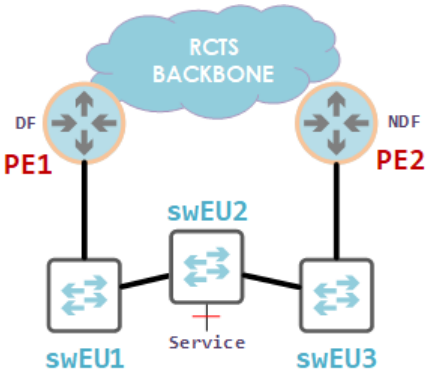
The main goals for L2 access rings:

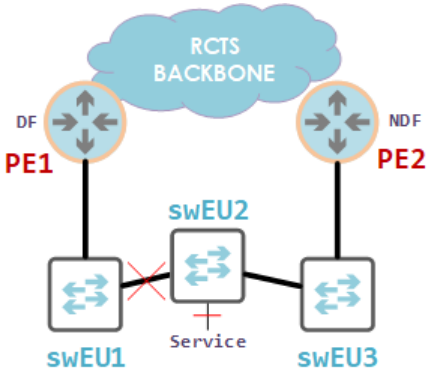
- Provide **equal or greater redundancy** to the services when comparing to RCTS 1.0;
- If possible using the same technology for **all the services** (EVPN);
- Eliminate STP and **physical loops**;

The main problem found:

EVPN **lacks a proper way to detect indirect failures** in Ethernet connections. This characteristic causes problem when modelling all L3 and L2 services for topologies larger than squares (>2 switches in chain);

Ethernet OAM and Ethernet ring protection were tested to fill that gap but the convergence times or the extreme complexity don't seem to fit as a proper approach.

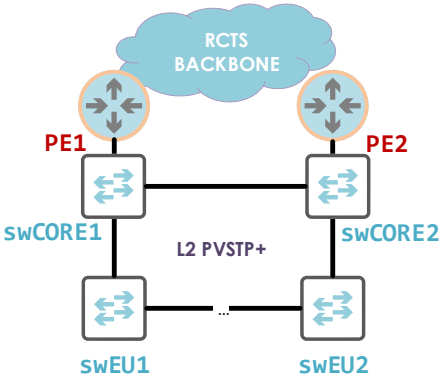




Access Rings – Layer 2 (2)

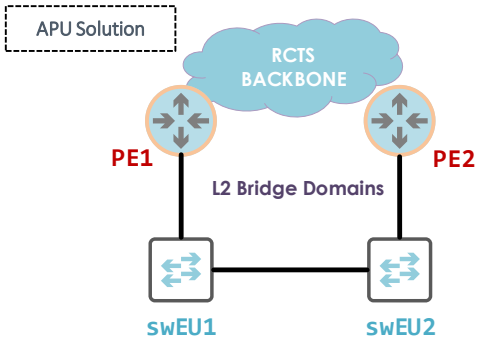
The three following models were tested to implement RCTS planned services:

L2 STP Rings (RCTS1.0)



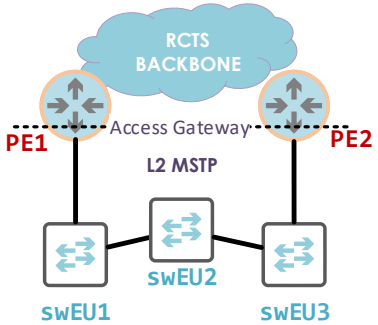
- Compatible with RCTS current technology and equipment
- PVSTP+ is a proprietary protocol (may be changed)
- VLAN and STP management required

L2 Loop-Free Square Rings



- Service redundancy is accomplished through PEs
- VLAN management required but local
- Indirect faults are not detected on L2 leased circuits or backup links

L2 STP-AG Rings



- STP exists only on access rings and it is not propagated through backbone services
- MSTP required for MST-AG
- The main AG technology is MST-AG and Cisco Proprietary
-

Access Rings – The Winner(s)

An effort should be made to create **L3 access** devices, either through extended core or core-only models. These solutions are the ones that can offer true **E2E services** and a complete **abstraction of the underneath physical topology**.

If proper L3 equipment is not available for all the locations, **layer 2 rings** should be the alternative to transport network services in a redundant way.

EVPN access should be deployed for all those services to ensure consistent network design and higher levels of service redundancy. For square topologies, even STP will not be needed. For the larger ones, rings as RCTS1.0 with interconnection between RCTS Data Centres and STP to avoid loops will guarantee a coherent service design.

RCTS 100 Project
RCTS 2.0

Services

Services Disclaimer

What has to change in service configuration?

- Everything that used to rely on VLANs spreading the backbone: **VOIP**, **RCTS Plus** and **Management**.
- These services will be implemented over IP backbone with **MP-BGP**.

Are these changes also mandatory on all the L2 rings?

- Actually No… The rings can be “stitched” at PEs “as it is”: we would still have path protection and keep the same designs.

Do we want to change it?

- VOIP, MGMT and static SERVIP **don’t have PE redundancy**…
- **STP** is not part of EVPN/VPLS best practices and have larger convergence times.

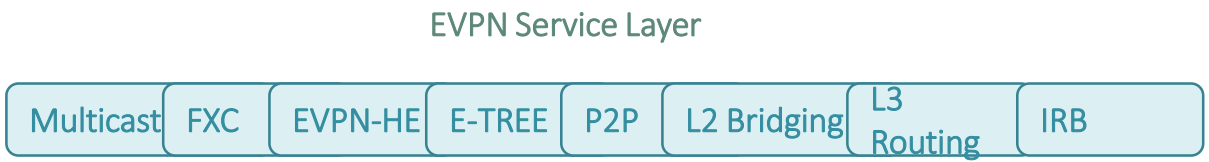
EVPN – The new kid in town

EVPN has been gaining a lot of attention in both SP and DC world has it builds over VPLSs to add **redundancy** and solve **non-optimal BUM** forwarding traffic.

With these additions as well as **a smooth integration with MPLS** data-plane, EVPNs are positioned as the best service option for RCTS.

The service is mainly designed for L2 services but with its IRB extensions can also be used to provided L3 access connectivity.

The control-plane is mainly built with one additional **BGP Address-Family** that should be configured among all the PEs of the network (through RR).

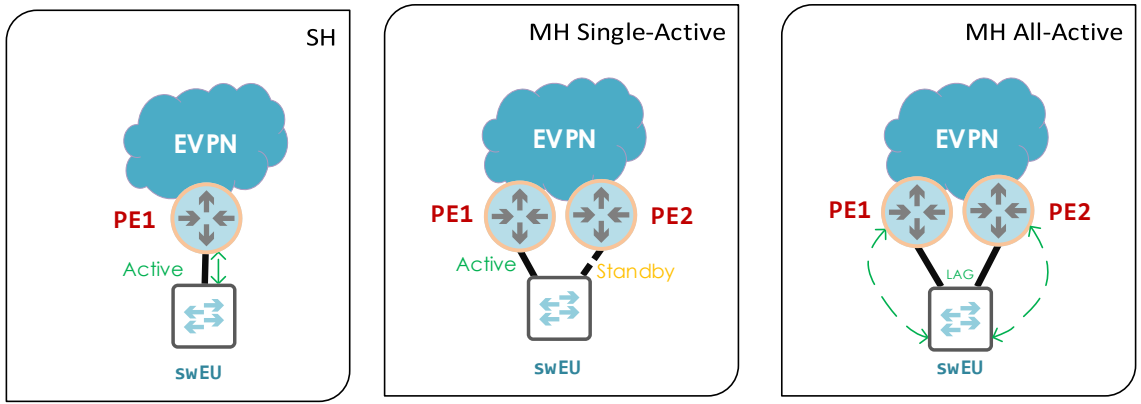


EVPN – Built-In Redundancy

As for what concerns Multi-Homing, EVPN adds two modes:

- Single-Active mode where only one PE is forwarding traffic at a time (**Active/Standby** Solution). Per-vlan hashing might be supported.
- All-Active mode where two PEs can load balance all the traffic at the same time (**Active/Active** Solution).

However, one should note that these resilience mechanisms are designed for distributed PEs or DC gateways and to work in **triangular** topologies – all other topologies should be **adapted carefully**.



EVPN for Access – “APU” Solution (1)

Although EVPN is usually associated with L2 services, it was also designed to be **an integrated routing service** for data centers. In those cases, a **scalable “virtual” gateway** can be added to both PEs on an Ethernet Segment.

This solution is not suitable to replace Internet or L3 VPN services, as the backbone still relies on the regular BGP AFs. However, **the signaling part of Segments between PEs** might be seen as a simpler alternative to VRRP on the access part of the network.

In fact, only a single **GW and MAC address** is needed for both PEs, and signaling happens at the much more **stable backbone interfaces**.

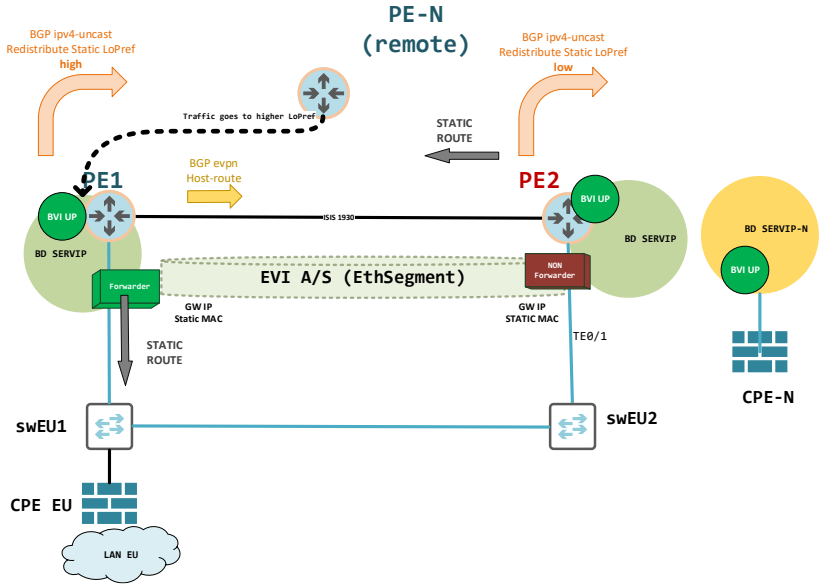
The Active/Stanby case should be the one configured to not only ensure proper QoS treatment (same active PE) but also to guarantee proper fault recovery to indirect link failures.

The **main problem** is that EVPNs are designed for triangles with direct links and active PE switchover is forced only by down interfaces. With that, only **topologies without indirect faults** to detect (as optimized “squares”) can be used without relying on physical loops combined with any type of STP.

EVPN for Access – “APU” Solution (2)

So… How does it work for Static IP Services?

1. An operator ensures that the PE connecting to the main link is the Segment Designated Forwarder. In Cisco devices, mod(VLAN/(nº PE) is the used formula.
2. A scalable design in the PEs ensure that the static routes are always advertised with higher LocalPref on main PE.
3. As BVI is always up on both sides, a more specific host route (EVPN AF) is shared from PE1, ensuring that static route on PE2 always points to PE1 (PE2 generated traffic).



EVPN for Access – “APU” Solution (3)

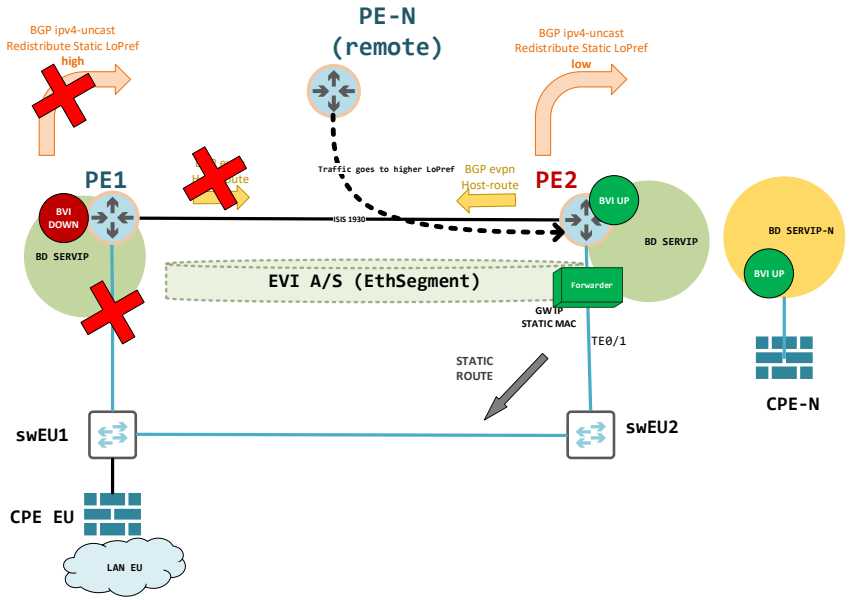
And… How does it handle redundancy?

1. If a direct link fails, PE1 will notice it and immediately switch the state of the BVI:
   a) The static redistribution is **stopped** on PE1;
   b) The host-route is **withdrawn** by PE1;
   c) The PE2 gets the signaling to be DF through the backbone;
2. swEU1 will drop the mac address (interface down) and start flooding to the backup link.
3. Static on PE2 is the same, but the most specific host route is gone and will forward to the directly connected BVI.

CPE never loses ARP (same IP+MAC) and convergence is fast!

Even traffic from PE1 will get to PE2 through the host route.

All services connected to swEU2 will work the same way: DF was always PE2 and switchover is done on a per-VLAN basis.



EVPN for Access – “APU” Solution (4)

Does it get back properly?

It’s the same process to get to initial state. EVPN signaling will get PE1 as DF again and return traffic (remote-PE -> CPE EU) will force the mac-move on swEU1.

What about indirect failures?

Square topology ensures that the only indirect link is the backup one where no convergence is required.

Does it work for directly connected?

The same process applies without static redistribution. Host-route will be more specific than the interconnect network.

How does it behave with STP?

The ring is setup as RCTS1.0 and with PVSTP+ running. Indirect failures on primary path may exist but PEs will not have to switch roles as there is an alternate path to the main PE.

Switchover only happens if PE1 goes down. Reconvergence might be slower as BGP timers will be required.

EVPN for Access – “APU” Solution (5)

What if a PE fails?

The same switchover process happens. The overall timing would not be as short, as backup PE will have to wait the BGP timers to expire to withdrawn the previous EVPN routes and to become the DF.

Is there a split-brain problem with an isolated PE?

In a very unlikely scenario of a PE without uplink to the RCTS backbone, core-group will be configured and used to group all the P-PE interfaces: if all of them are down, main PE will ensure that its Ethernet Segments get to a NON-FORWARDING state. As in the previous scenario, PE2 will become the DF after PE1-RR session is down.

Does it work in L3 access rings?

It’s even better! If pseudowires are deployed instead of VLANs, the regular sinalisation between a multilayer switch and a PE will make the backup tunnel as “standby” – this feature will create an “empty” Bridge Domain on secondary PE that will force the BVI down. All the convergence is the same, but only one IP is propagated at each time throughout the network.

L3 Services summary

All EU service configuration remains the same!

Internet Service

- Redundant BGP sessions are still the recommended approach.
- Static configurations become PE redundant.

Voice Service

- Voice services become path and PE redundant.

Management

- Internal management also become path and PE redundant.

L3VPN

- New available service to provide same private connectivity as RCTS+ through ip routing.

L2 Services summary

All EU service configuration remains the same!

RCTS Plus

- Becomes highly redundant within our backbone
- Faster convergence during failover
- Some existent services might me configured as P2P (VPWS) to avoid mac-learning in our backbone

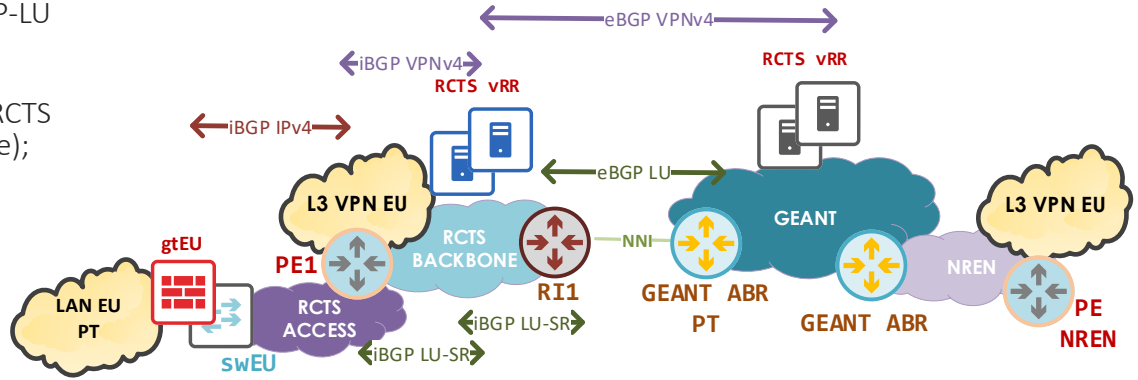
Services – MDVPN

MDVPN is the interconnection option for L3VPN shared with GEANT. The current architecture can be improved with the new design with **MPLS** data-plane and **MBGP**.

SR demands some adjustments on regular Inter-AS Option C connectivity. BGP-LU will have to announce SR indexes (as BGP-SR) to avoid SR/LU label overlapping on RR and to ensure correct prefix mapping.

Option 1: IGP-SR <-> BGP-LU redistribution;

Option 2: BGP-LU extended to RCTS PEs through vRR (represented here);



RCTS 100 Project
RCTS 2.0

Peering and flows

Peering Design

Keep it Simple

Two main ASBR in each region: Porto and Lisbon.

All **NNIs** connected to peering routers.

A third router just for **GigaPIX redundancy**.

Active/Active scenario must be allowed @GEANT.

Full Routing Table (FRT) only exists on RI-PRT and RI-LX1.

Communities will generate the **Partial Routing Table** (PRT), a combination of Gigapix, Geant R&E and Direct Peers prefixes

These prefixes will be leaked to RR´s.

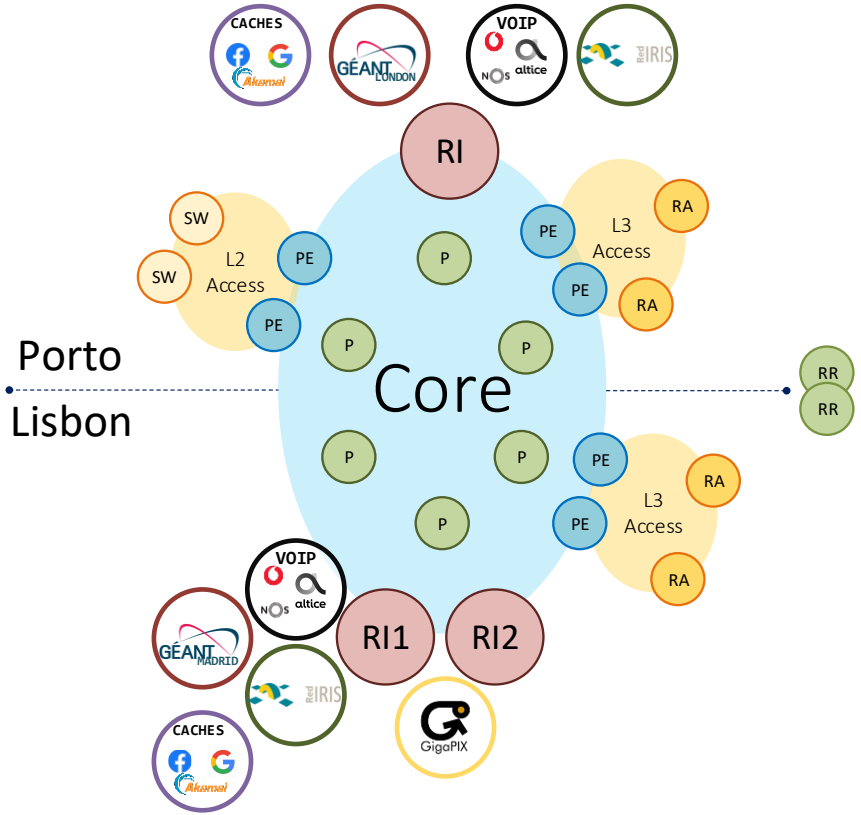
Internal Routing Table (IRT) is the collection of specific RCTS prefixes.

PEs should only know IRT + defaults as default option.

FRT or PRT subsets might be added when needed.

RRs know **all prefixes** from all routers from all AFs.

Ingress route filtering is performed based on communities on each PE.



GEANT Traffic

Hot-Potato Routing

Egress:

All PEs will send traffic to the closest RI due to default route.

No per-destination optimization possible by default.

Ingress:

Aggregated prefixes will be announced on both links.

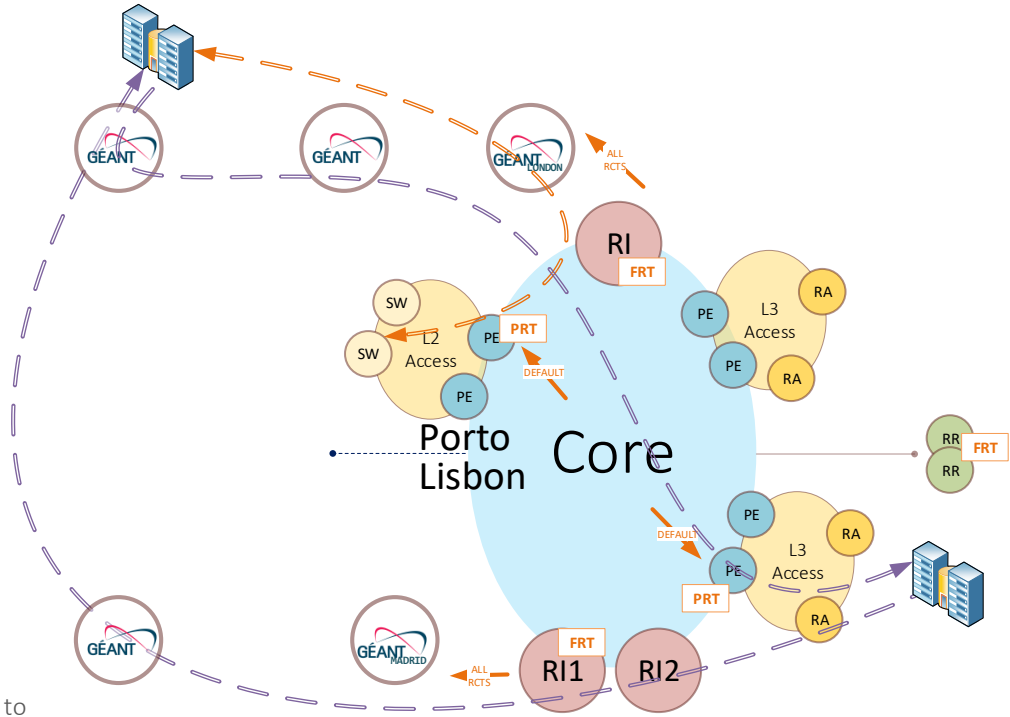
Hot-Potato Routing will also exist in GEANT network.

Redundancy:

If one RI fails, all the connectivity is assured by the other.

If one uplinks fails, traffic might flow to closest RI and then be rerouted to active link (optimization needed).

Simplicity in BGP design is the main advantages. Traffic burden changes in CORE should be measured.



Segmentation of prefixes per zone would be awesome!

RCTS 100 Project
RCTS 2.0

Network Considerations

Topics for another meeting?

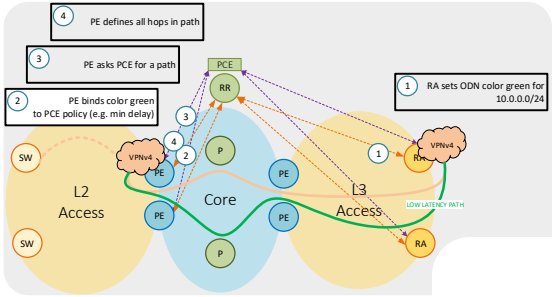
SRv6 and full IPv6 programmability

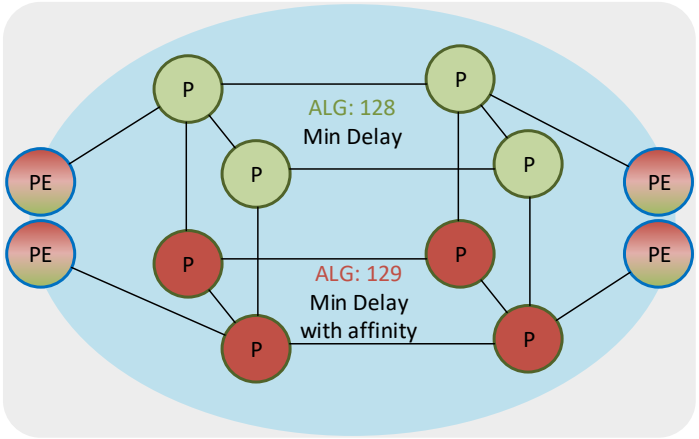
Traffic Engineering and **per-flow** traffic engineering

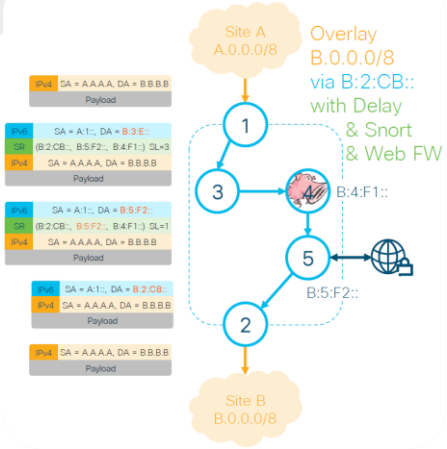
Flex-algo for dual-plane backbones

MTU in MPLS networks

QoS for MPLS services







Thank you!

João Silva joao.silva@fccn.pt

Pedro Lorga pedro.lorga@fccn.pt

Tiago Monteiro tiago.monteiro@fccn.pt