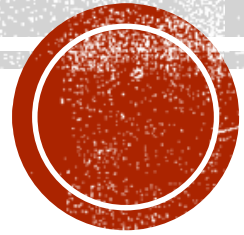
MPLS VPN TROUBLESHOOTING SUGGESTIONS BASED ON EXPERIENCE IN OPERATIONAL MPLS BACKBONE



Nafeez Islam

Assistant Manager,

NovoCom Limited,

Bangladesh.

ABSTRACT

- MPLS VPN configuration itself can be a challenge, but often problems occur much later during the operational phase. This paper does not describe initial MPLS configuration methods, rather a few suggestions on some elements required to be checked during MPLS VPN failure in the operational period.

Successful configuration of Layer 3 or Layer 2 MPLS VPN has been done. Months have passed after provisioning and services running smooth.

But many days after the service had become operational and VPN working completely fine, suddenly client complains about getting end to end CE communication down. Havoc is unleased.

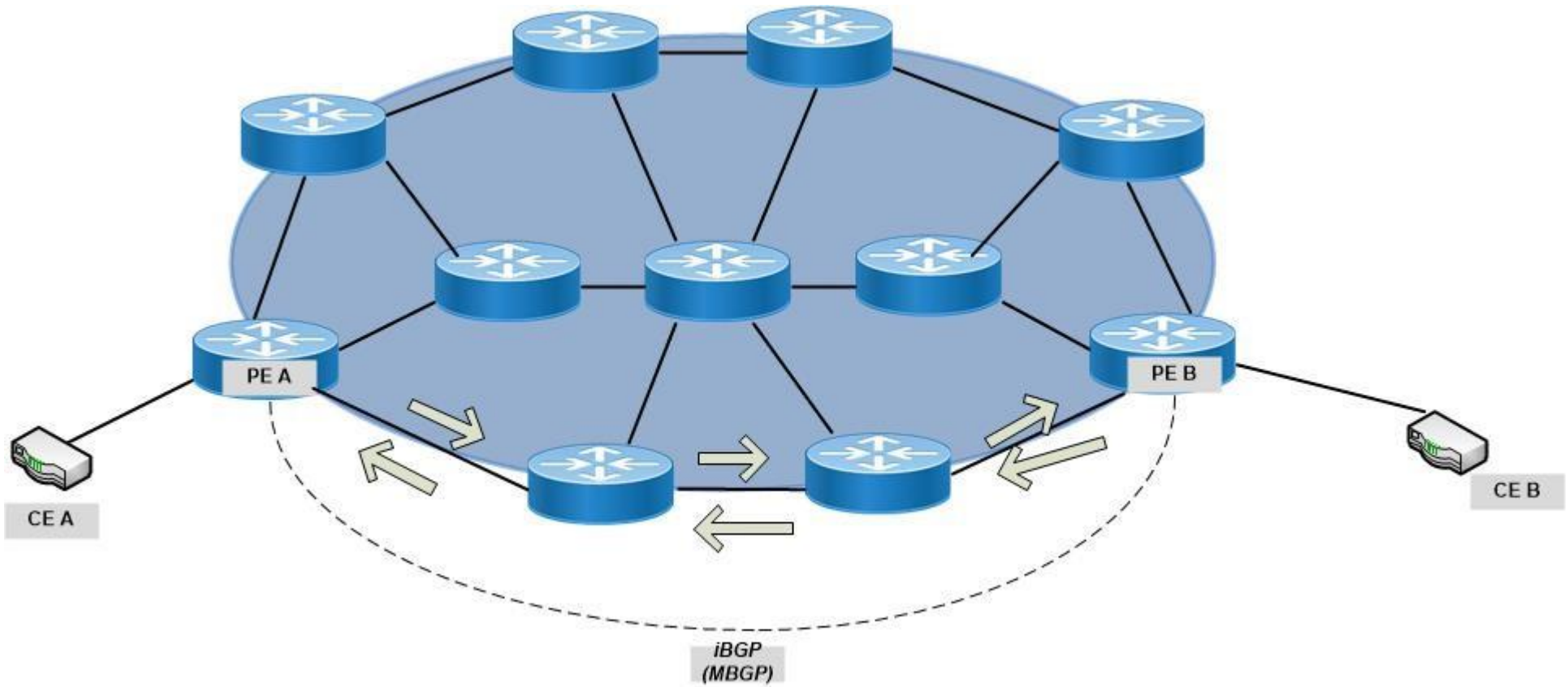
SYMPTOM: RECEIVING PREFIXES IN VRF, BUT NOT GETTING SERVICE/ ICMP RESPONSE

- Most of the time this case takes place because the VPN traffic is going through a path which the traffic is not supposed to use.

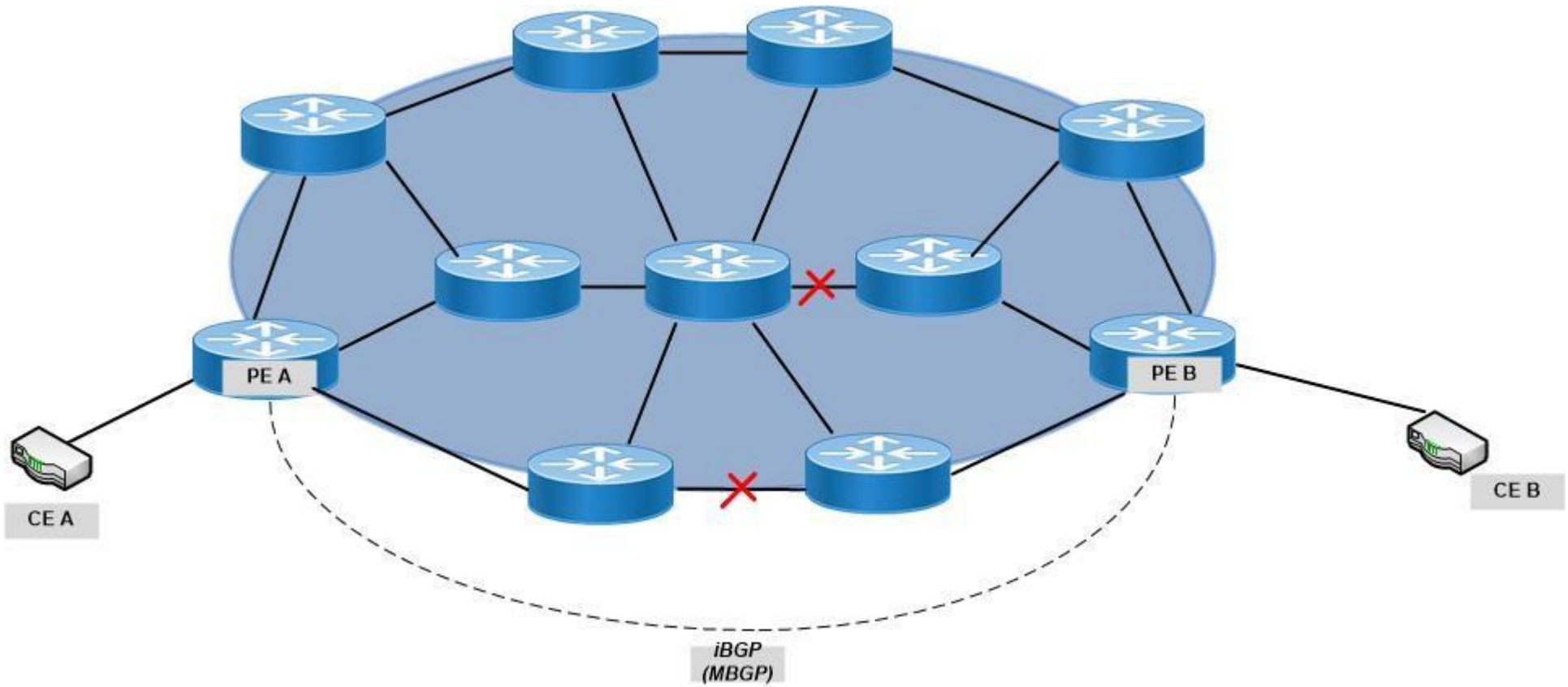
- A service provider's backbone can be quite large and may consist of hundreds of routers.
- Often at some part of the backbone or at a particular link, the required MPLS related configurations are not there. It could be by mistake or sometimes intentional.
- It was not creating problems so far because the VPN traffic was not going through this particular path at the time of configuration.

- Only when there is a change in the IGP routing table and traffic starts going through that problematic link, the data services go down and we come to know that there is a problem.
- In large backbones it is often difficult to predict what will be the end to end path of the traffic based on the IGP metrics after link cuts.

NORMAL FLOW OF VPN TRAFFIC AFTER CONFIGURATION



LINK CUTS, PATH OF VPN TRAFFIC CHANGES



- The first task would be to find the particular link which is causing the problem.
- So, find out the way the VPN traffic is taking in your network.
- Keep in mind that your IGP and normal routing is still working fine.
- Go to one of the provider edge routers PE A. Now traceroute to the Loopback of the other end PE B.

WHY TRACE THE LOOPBACK IP OF PE B?

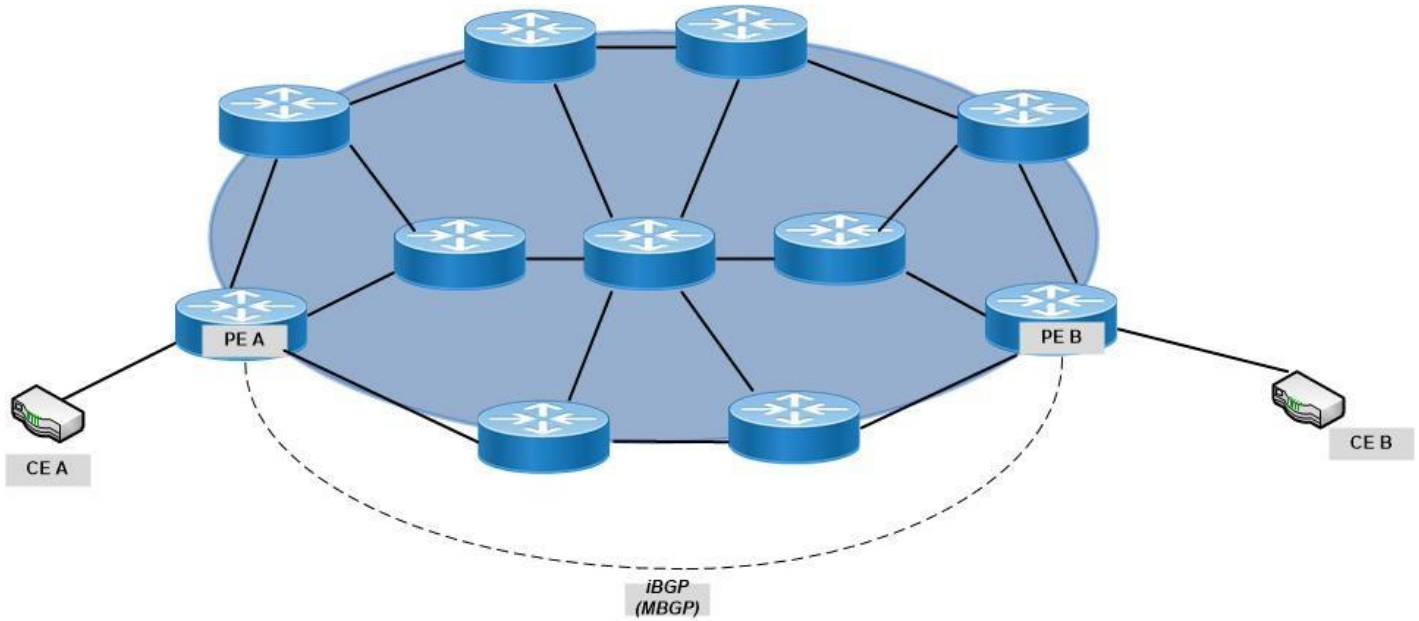
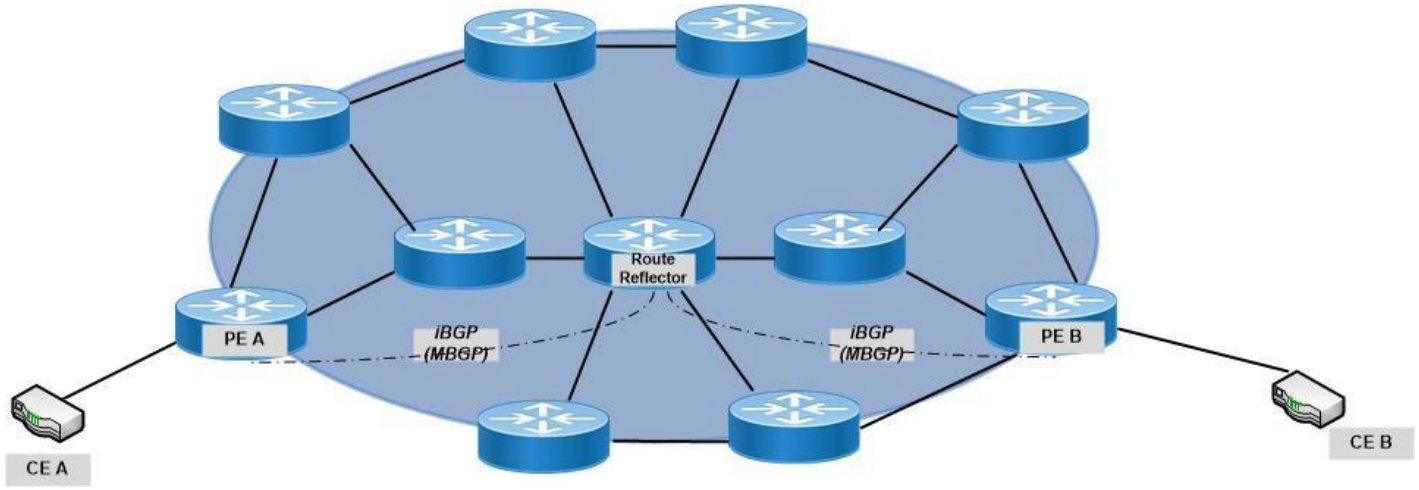


Diagram :Direct mBGP between PE routers

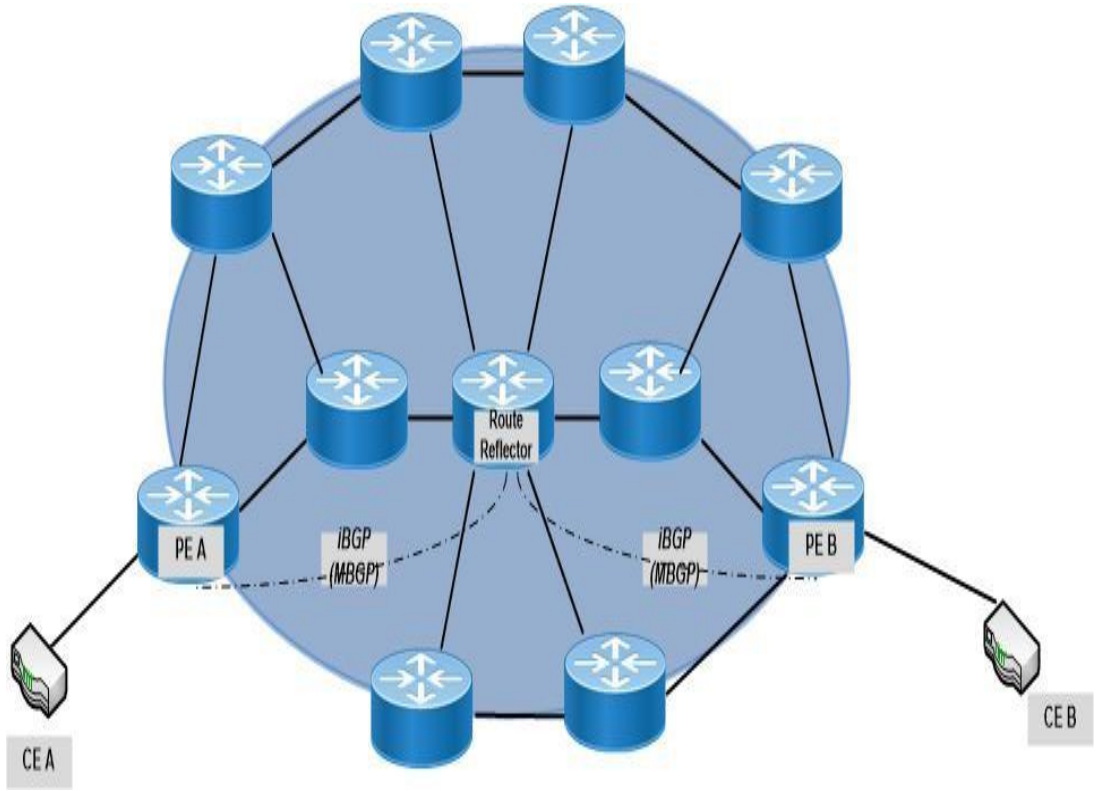
WHY TRACE THE LOOPBACK IP OF PE B?

Diagram :With a Route Reflector



WHY TRACE THE LOOPBACK IP OF PE B?

- If it is a Layer 2 tunnel, e.g VPLS , you probably have configured the tunnel on the loopback IP.
- If it is Layer 3 MPLS, for all the VPNv4 routes the next hop is the loopback of PE B.
- In each labelled packet, the top label (also called transit label) will be the label for loopback of PE B.



- So what path the VPN traffic is taking from PE A depends on what is the best path to reach loopback IP of PE B determined by IGP .
- So the summary is you have to traceroute from PE A to PE B, then again from PE B to PE A. Now you know the exact the path being taken by the data service.
- **After determining the whole path,now go to each of the hops and inspect checkpoints 1,2 and 3 one by one which are explained in following slides.**
- You can also use traceroute mpls command which tells you exactly after which hop the LSP is broken.

Checkpoint 1 :LSP Breaking- Higher MTU than 1500 not allowed on a particular link

- Labels which are imposed add extra bytes in the header. In the label stack, for traffic between P routers there is 1 label, for VPNv4 there will be 2 labels, bottom label and transit or top label ,or even higher for tunnels or MPLS TE.

Why is higher MTU than default 1500 bytes is necessary?

Checkpoint 1 : LSP Breaking- Higher MTU than 1500 not allowed on a particular link

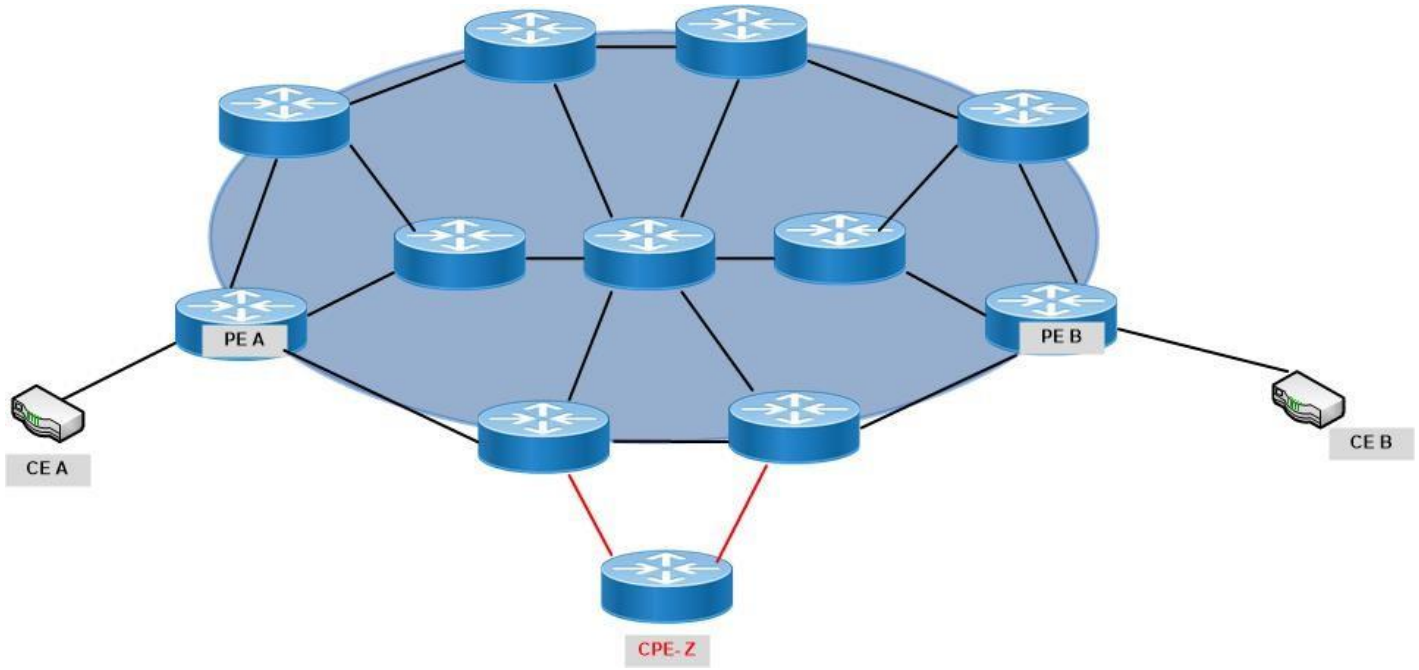
- If nothing is changed on a router interface and configuration kept as default (default MTU is 1500 bytes), labelled packets will get fragmented.
- Fragmentation can create a lot of problems and it is not recommended.
- One of the reasons is, for successfully reassembling, no fragment can be missed or corrupted. There is no mechanism in place yet to let the sender know that a fragment is missing.

Why is higher MTU than default 1500 bytes is necessary?

Checkpoint 2 :LSP Breaking- MPLS not running on a new link

- May be after the VPN was first configured a new link has been added in the backbone but MPLS was not enabled in the interfaces
- Again, it was not creating problems so far because VPN traffic was not going through this particular path as it is a new link.
- Check if LDP is running and neighborship has formed.

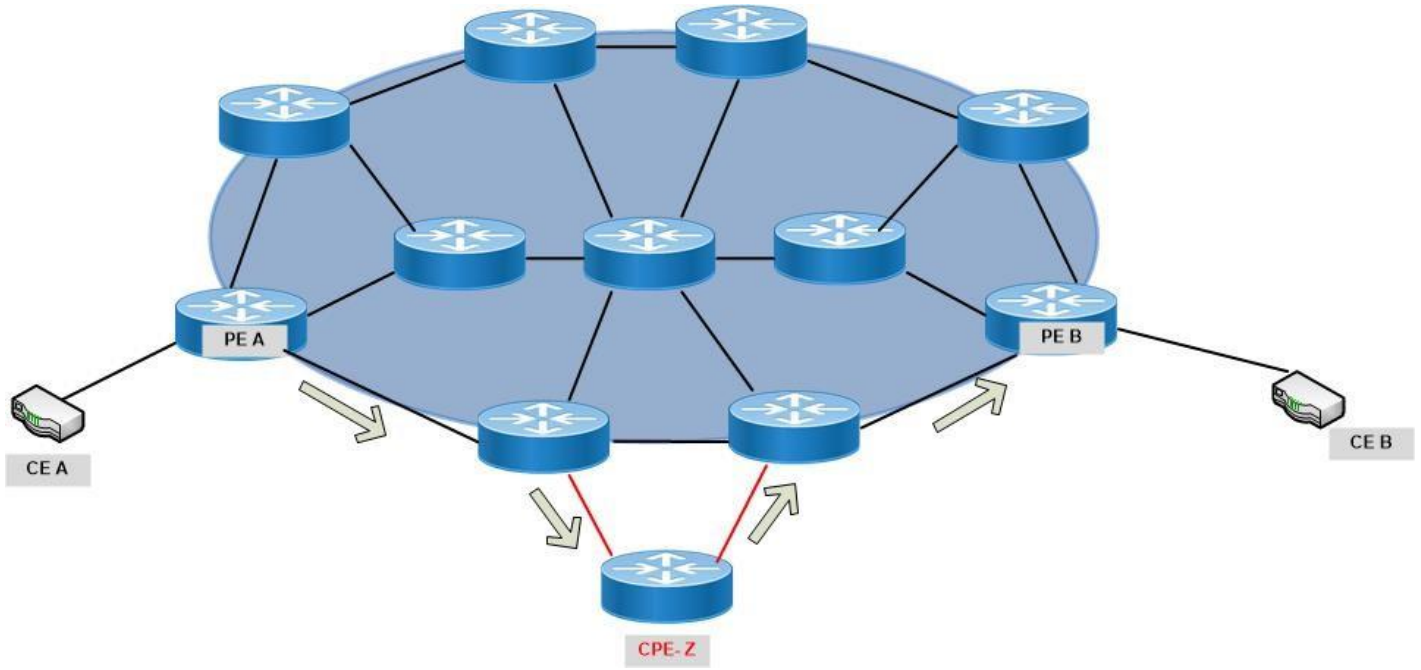
CHECKPOINT 3 :LSP BREAKING- CPE (CLIENT-PREMISE-EQUIPMENT) BECOMES TRANSIT, CPE DOES NOT RUN MPLS



Sometimes there are routers in the backbone which are not supposed to become transit.

It could be that such router does not have the required MPLS configs, may be just basic Ipv4 routing.

In our backbone, often to manage last mile protection for an internet client we place a router at client's premise.



CHECKPOINT 3 :CPE (CLIENT-PREMISE-EQUIPMENT) BECOMES TRANSIT, CPE DOES NOT RUN MPLS

If MPLS traffic choses the CPE as transit , MPLS VPN traffic will fail because LDP is not enabled in this router.

Tweaking in the IGP routing can be done so that traffic does not take this path.

CHECKPOINT 4 : USE OF SAME RT,RD FOR DIFFERENT CLIENTS :

- This issue happens only if proper and updated databases of all VPN IDs, RDs, RTs are not maintained.
- Once when a client of ours was not getting traffic, the route table of the particular VRF showed that CE's advertised prefixes were present at the PE of the other end.
- However a closer look at this VRF routing table revealed that the next hop for this VPNv4 prefix was a different PE than the expected.

CHECKPOINT 4 : USE OF SAME RT,RD FOR DIFFERENT CLIENTS :

- What actually happened was, there was a mistake in the database consisting of VPN IDs, RDs, RTs .
- So wrong RDs and RTs were being used.In the VRF, we were receiving prefixes of a different client because of using the wrong RTs.
- Both of the clients were using the same private IP blocks to make things more confusing.

EXAMPLE

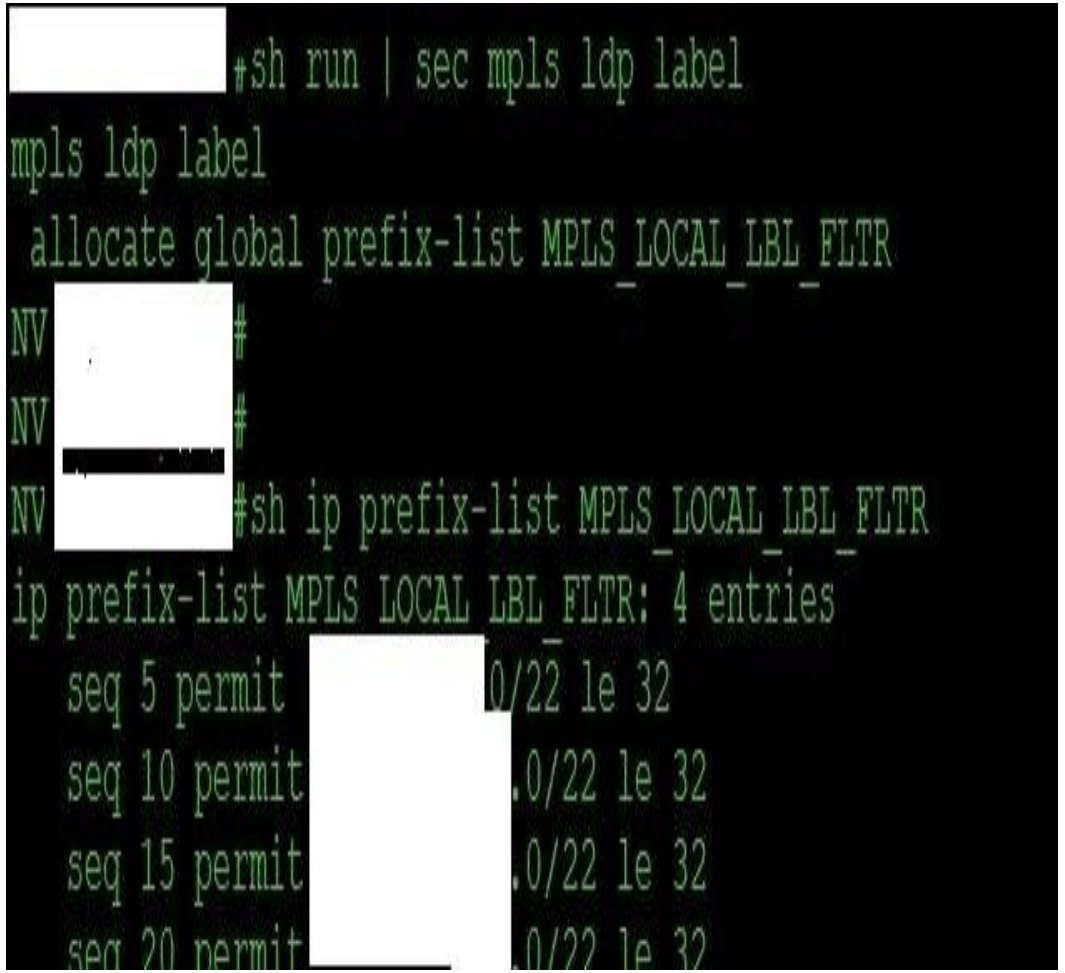
- PE-RTR1#sh bgp vpnv4 unicast vrf VRF-TEST

```
     Network          Next Hop            Metric LocPrf Weight Path
Route Distinguisher: 10.11.12.13:8 (default for vrf VRF-TEST)
 *>  10.1.1.201/32    10.73.15.81                          0 65534 ?
 *>  10.51.1.224/30   10.73.15.81                          0 65534 ?
 *>  10.73.9.80/30    172.26.0.54              80          0 58923 65420 ?
 *>  10.79.9.80/30    10.9.13.3                0    100    0 58923 65420 ?
 *>  10.79.11.84/32   172.26.0.54              80          0 58923 65420 ?
 *>  10.74.9.80/30    10.9.13.3                0    100    0 58923 65420 ?
```

CHECKPOINT 5 :LABEL FILTERING GONE WRONG:

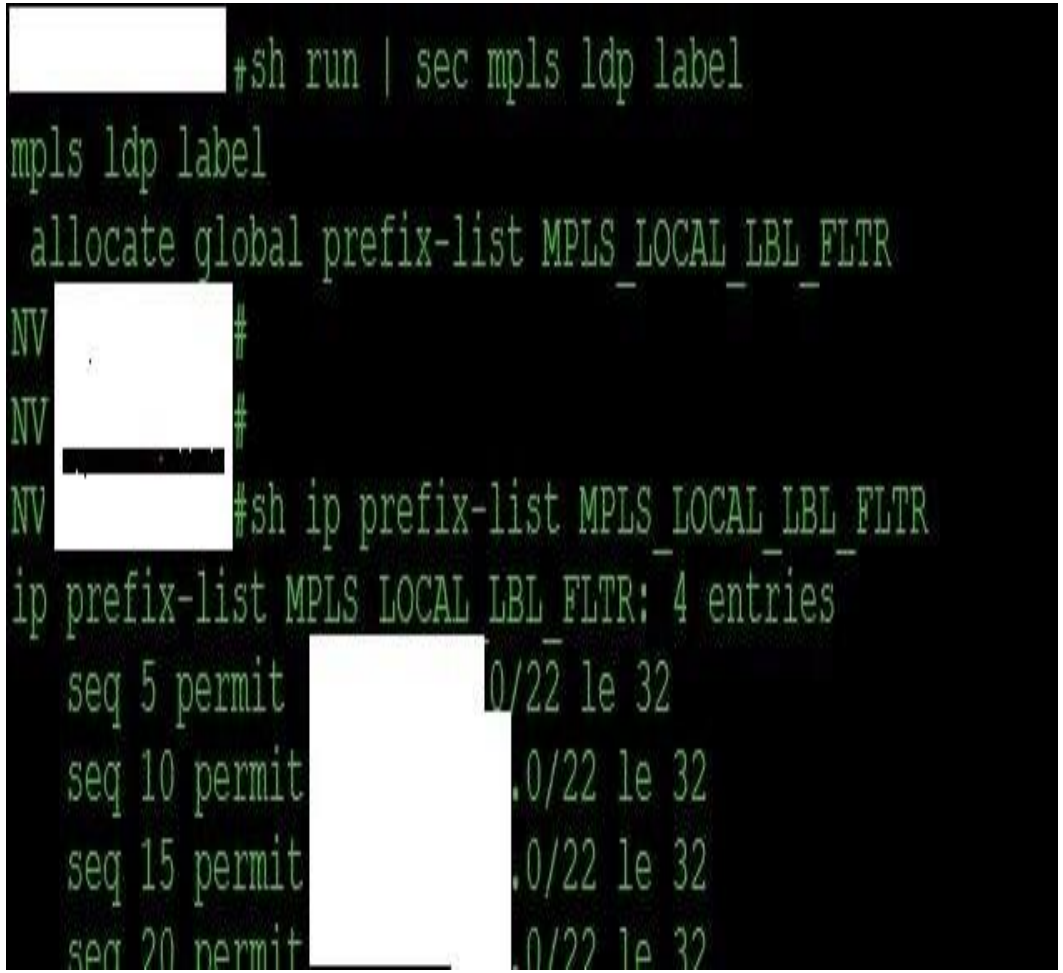
Why Label Filtering?

- We have a practice to use label filters in our backbone which enables a P router to create labels for only certain prefixes and advertise those labels to other P routers.
- Because we just need to create labels of prefixes of only the desired Label switched paths.
- It will be a waste of router's processor and memory to create labels for all routes.



CHECKPOINT 5 :LABEL FILTERING GONE WRONG:

- Once we deployed a fresh and new IP block in our MPLS backbone, but forgot to allow them in the label filter.
- This caused L2 and L3 MPLS VPN established problems for some clients.
- All these were solved in the end when labels were being created for that new IP block by allowing in label filter.
- But it took quite some time to figure out what was actually causing the issue.



EXAMPLE

****Checking whether labels are being created appropriately is always a great troubleshooting tool**

```
RTR-Z#sh mpls ldp bindings 192.168.136.0 30
  lib entry: 192.168.136.0/30, rev 2
    local binding:  label: 405
        remote binding: lsr: 10.10.10.3:0, label: 29
```

```
RTR-Y#sh mpls ldp bindings 192.168.136.0 30
  lib entry: 192.168.136.0/30, rev 1661
        local binding:  label: 29
        remote binding: lsr: 10.10.10.4:0, label: 79
        remote binding: lsr: 10.10.10.1:0, label: 72
```

CHECKPOINT 5 :LABEL FILTERING GONE WRONG:

- In another incident another small mistake was made, but the outcome was not so tiny.

- In case of BGP's prefix advertisement filters, usually we allow prefixes in this manner-

```
ip prefix-list EXAMPLE_PRFX: 2 entries
   seq 15 permit 10.10.216.0/22 le 24
    seq 25 permit 10.10.11.0/22 le 24
```

- This is because in case of eBGP we usually discard prefixes those are smaller in size than /24.

CHECKLIST 5 :LABEL FILTERING GONE WRONG:

- Because of this practice, an engineer configured the MPLS label filter in the same manner:

seq 5 permit 192.168.11.0/22 le **24**

- So when we divided this IP block and used it in subnets of /30 or any subnet smaller than /24 in different parts of the MPLS backbone, labels were not being created for them.
- The problem was resolved after the prefix list was corrected to the following form-

seq 5 permit 192.168.11.0/22 le **32**

THANK YOU