VOCUS communications

MPLS + Traffic Engineering

Tim Warnock – Vocus Communications

Who am I?

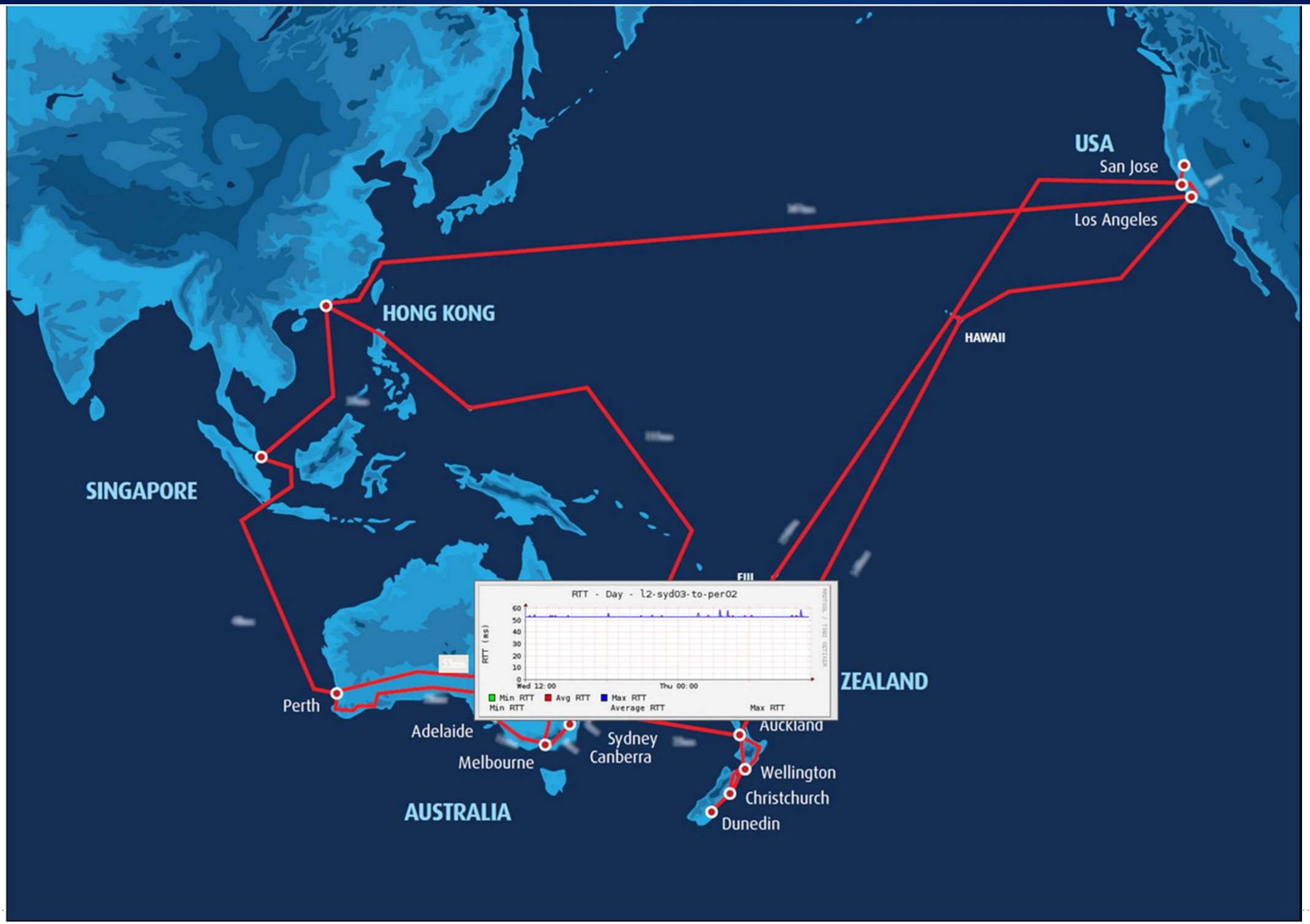
- I’m a senior network engineer at Vocus Communications based in Brisbane.
- I primarily work with Brocade and Cisco equipment within the Vocus network.
- I have previously built WISP networks using the MikroTik platform combined with other vendors.

Vocus Communications

- AS4826
- ASX Listed Wholesale + Direct supplier of Dark Fibre, Metro Ethernet, IP Transit and Voice.
- Services Australia, New Zealand, South East Asia and the United States.
- Uses MikroTik RB2011 “probes” in 15 key POP sites for performance analysis. Data is collected via API and stored in RRD.

Vocus MikroTik Probes

Vocus MikroTik Probes



Back on topic

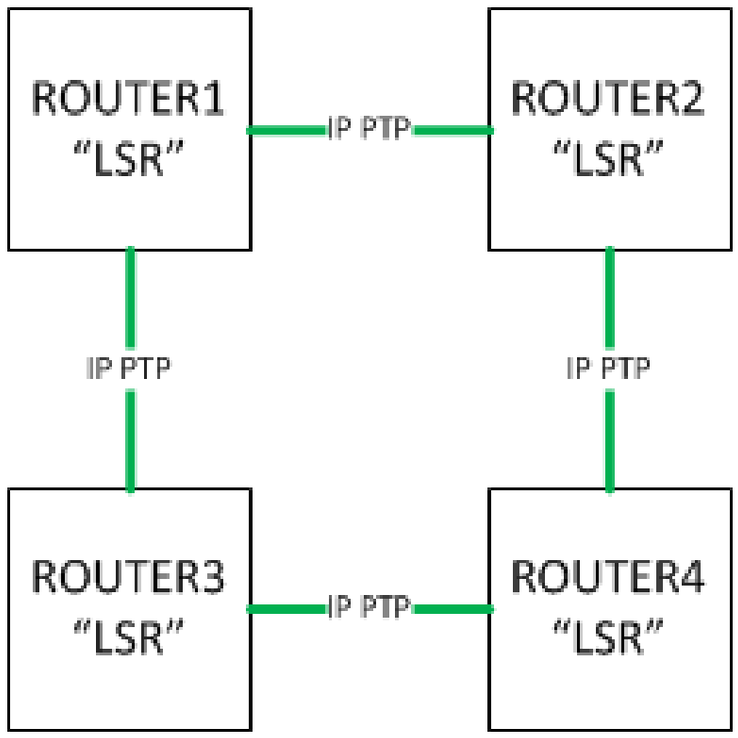
VOCUS communications

What is MPLS?

- Stands for Multi Protocol Label Switching.
- Requires both OSI Layer 2 and OSI Layer 3 to be functioning.

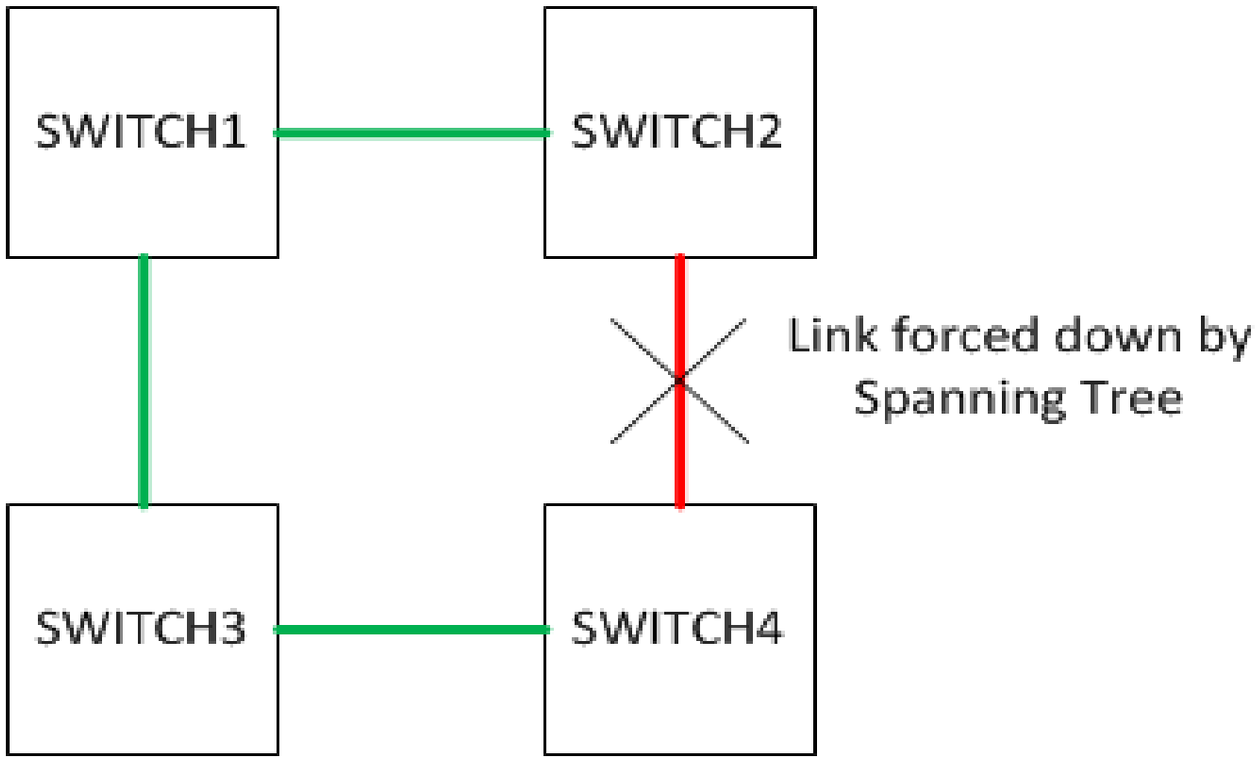
MPLS is layer 2.5

- Requires Layer 3 (IP) routing information to reach destination.
- Frames are “label switched” through “label switch routers“ over point-to-point links at Layer 2.

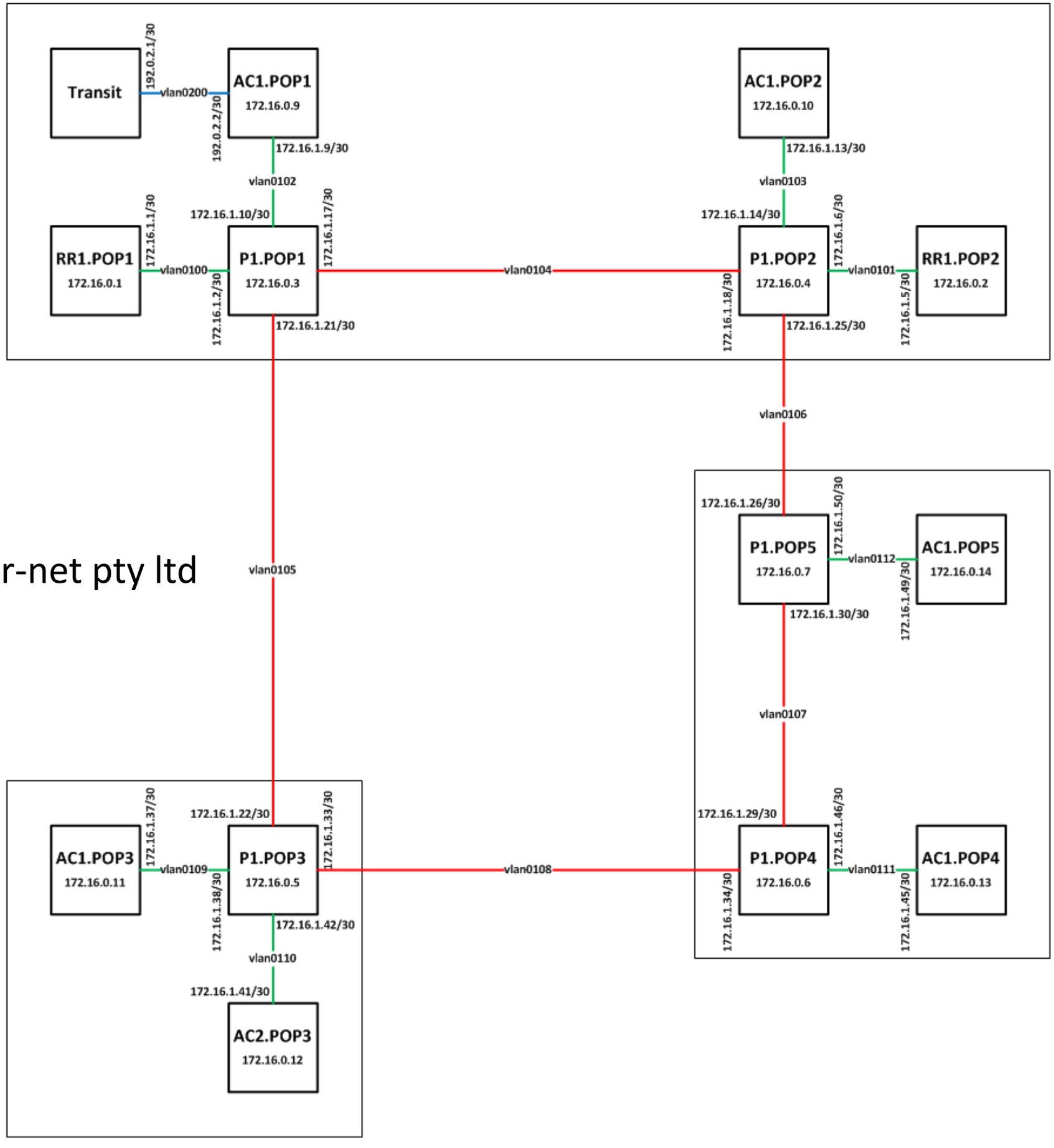


MPLS vs STP

- Why use MPLS over traditional Layer 2 + Spanning Tree Protocol?
- Its more efficient
- You can safely build loops



compu-global-mega-hyper-net pty ltd



OPSF

- This network uses Open Shortest Path First (OSPF) for its interior gateway protocol (IGP).
- We will only be using area 0.
- OSPF will be used only to distribute routing information on connected interfaces.

BFD

- This network is configured to use Bidirectional Forwarding Detection helper for OSPF.
- BFD (on MikroTik) defaults to sending a hello every 200ms and will consider the partner down after missing 5 hellos in a row (1 second).
- This triggers a down event for OSPF.
- It is useful for detecting faults when link state may not be lost or a link may go unidirectional.

BGP

- Border Gateway Protocol is an exterior gateway protocol used on the wider internet for communication route information between autonomous systems (AS).
- This network will also use BGP as an IGP (along side OSPF).
- This network will employ BGP Route Reflectors rather than maintaining a full mesh of BGP neighbours.

LDP

- MPLS employs the Label Distribution Protocol for maintaining a database of labels.
- In this network we will use LDP for signalling the setup of MPLS services.

RSVP - TE

- Resource reSerVation Protocol is an MPLS transport protocol.
- RSVP-Traffic Engineering is an extension to RSVP.
- RSVP-TE Maintains a database (Traffic Engineering Database) of labels and paths.
- RSVP-TE Database distributed by extension of existing IGP – OSPF in our case.

Putting it all together

VOCUS communications

Roles

- RR router
  - OSPF + RSVP-TE / LDP / RSVP
  - BGP Route Reflector role comprising
    - BGP instance with client-to-client-reflection=yes
    - BGP peers configured with route-reflect=yes

Roles

- P router
  - MPLS LSR – “Core router”
  - OSPF + RSVP-TE / LDP / RSVP
  - Configured as a BGP free core.

Roles

- Access router
  - OSPF + RSVP-TE / LDP / RSVP
  - BGP Route Reflector client

Interfaces

- Faux “Loopback Interface”

```
/interface bridge
add name=Loop0 protocol-mode=none
```

- I have used VLANs to segregate traffic.
- I have used EtherType=0x88a8 so that virtual-box interface doesn’t strip VLAN tag

```
/interface vlan
add interface=ether2 name=vlan0102 use-service-tag=yes vlan-id=102
add interface=ether2 name=vlan0200 use-service-tag=yes vlan-id=200
```

OSPF

- Set up OSPF instance enabling RSVP-TE extensions

```
/routing ospf instance
set [ find default=yes ] mpls-te-area=backbone mpls-te-router-id=Loop0 redistribute-
connected=as-type-1 router-id=172.16.0.9
```

- Set each Point-to-Point IP range as area=backbone (aka area=0 or area=0.0.0.0)

```
/routing ospf network
add area=backbone network=172.16.1.8/30
```

- Default passive and then enable PtP interface

```
/routing ospf interface
add network-type=broadcast passive=yes
add interface=vlan0102 network-type=point-to-point use-bfd=yes
```

BGP (Client)

- Configure BGP Instance

```
/routing bgp instance
set default redistribute-connected=yes redistribute-static=yes router-id=172.16.0.9
```

- Configure BGP Filtering

```
/routing filter
add action=jump chain=To-RouteReflectors jump-target=ConnectedToBGP protocol=connect
add action=jump chain=To-RouteReflectors jump-target=StaticToBGP protocol=static
add action=discard chain=To-RouteReflectors
add action=return chain=ConnectedToBGP
add action=return chain=StaticToBGP
```

- Configure BGP Peers

```
/routing bgp peer
add name=RR1.POP1-IPv4 nexthop-choice=propagate out-filter=To-RouteReflectors remote-
address=172.16.0.1 remote-as=65530 ttl=default update-source=Loop0
add name=RR1.POP2-IPv4 nexthop-choice=propagate out-filter=To-RouteReflectors remote-
address=172.16.0.2 remote-as=65530 ttl=default update-source=Loop0
```

BGP (Route Reflector)

- Configure BGP Instance

```
/routing bgp instance
set default redistribute-connected=yes redistribute-static=yes router-
id=172.16.0.1
```

- Configure BGP Filtering

```
/routing filter
add action=jump chain=RR-IN-IPv4 jump-target=NO-DEFAULT-IPv4
add action=accept chain=RR-IN-IPv4
add action=jump chain=RR-IN-IPv6 jump-target=NO-DEFAULT-IPv6
add action=accept chain=RR-IN-IPv6
add action=discard chain=NO-DEFAULT-IPv4 prefix=0.0.0.0/0
add action=return chain=NO-DEFAULT-IPv6
add action=discard chain=NO-DEFAULT-IPv6 prefix=::/0
add action=return chain=NO-DEFAULT-IPv4
```

BGP (Route Reflector)

- Configure BGP Peers

```
/routing bgp peer
add in-filter=RR-IN-IPv4 name=AC1.POP1-IPv4 nexthop-choice=propagate remote-address=172.16.0.9
remote-as=65530 route-reflect=yes ttl=default update-source=Loop0

add in-filter=RR-IN-IPv4 name=AC1.POP2-IPv4 nexthop-choice=propagate remote-
address=172.16.0.10 remote-as=65530 route-reflect=yes ttl=default update-source=Loop0

add in-filter=RR-IN-IPv4 name=AC1.POP3-IPv4 nexthop-choice=propagate remote-
address=172.16.0.11 remote-as=65530 route-reflect=yes ttl=default update-source=Loop0

add in-filter=RR-IN-IPv4 name=AC2.POP3-IPv4 nexthop-choice=propagate remote-
address=172.16.0.12 remote-as=65530 route-reflect=yes ttl=default update-source=Loop0

add in-filter=RR-IN-IPv4 name=AC1.POP4-IPv4 nexthop-choice=propagate remote-
address=172.16.0.13 remote-as=65530 route-reflect=yes ttl=default update-source=Loop0

add in-filter=RR-IN-IPv4 name=AC1.POP5-IPv4 nexthop-choice=propagate remote-
address=172.16.0.14 remote-as=65530 route-reflect=yes ttl=default update-source=Loop0
```

MPLS Protocols (LDP + RSVP)

- Configure LDP instance

```
/mpls ldp
set enabled=yes lsr-id=172.16.0.9 transport-address=172.16.0.9
```

- Configure LDP interface

```
/mpls ldp interface
add interface=vlan0102
```

- Configure RSVP interface

```
/mpls traffic-eng interface
add bandwidth=1Gbps interface=vlan0102
```

MPLS and MTU

- Each packet that is label switched has a stack of MPLS labels on them.
- This increases the size of the packet.
- MikroTik has a special interface setting called L2MTU for Non IP traffic.
- To carry 1500 bytes of Ethernet Traffic you will need to account for the MPLS label stack count (4 bytes per label).
- Bare minimum recommended is 1508 (2 labels).
- Best practise is to build to your network maximum capability.

RSVP-TE influencing IP routing

- We can influence the IP Routing table with RSVP-TE tunnels.
- This allows for better utilisation of links that might otherwise be idle or provide a better (or worse) experience for customers.

Building a TE tunnel

- First we start with the path
  - I will define 3 paths.
  - “IGP” will take the shortest path as calculated via Constrained Shortest Path First (CSPF). This will check every 30 minutes that it’s the best available path and if its not – self adjust.
  - “TO-AC1.POP5-SHORT” will only be the Loopback IP of P1.POP4
  - “TO-AC1.POP5-LONG” will be required to take the scenic route via the Loopbacks of P1.POP4 -> P1.POP3 -> P1.POP1 -> P1.POP2.

```
/mpls traffic-eng tunnel-path
add name=IGP reoptimize-interval=30m
add hops=172.16.0.6:loose, 172.16.0.5:loose, 172.16.0.3:loose, 172.16.0.4:loose name=TO-AC1.POP5-LONG use-cspf=no
add hops=172.16.0.6:loose name=TO-AC1.POP5-SHORT use-cspf=no
```

Loose vs Strict

- When building up paths, you have the option of loose or strict. I’ve found it’s best not to mix them.
- Loose paths don’t have to be completely defined - only enough to get you close and let the network handle the rest. Traditionally the hops are the Loopback IPs of LSRs in the path.
- Strict mode is useful when you have a specific path you want traffic to take. Traditionally the hops are the IPs of each PTP link you want it to take.

Final steps

- Define the paths (We'll make it take the long path, fail to the short path with a backup of IGP)
- Add the TE Interface to the destination Loopback IP.

```
/interface traffic-eng
add disabled=no name=traffic-eng1 primary-path=TO-AC1.POP5-LONG record-route=no secondary-
paths=TO-AC1.POP5-SHORT, IGP to-address=172.16.0.14

/ip route
add distance=1 dst-address=172.16.0.14/32 gateway=traffic-eng1
```

IGP vs RSVP-TE

IGP

```
[admin@AC1.POP4] > /tool traceroute use-dns=yes 172.16.0.14 count=10
# ADDRESS                          LOSS SENT    LAST      AVG     BEST   WORST
 1 vlan0111.p1.pop4.lab             0%   10    2.1ms      2.1      0.7     3.6
 2 vlan0107.p1.pop5.lab             0%   10    2.1ms      1.7      0.6     2.4
 3 loop0.ac1.pop5.lab               0%   10    1.7ms      1.8      0.5     3.8
```

RSVP-TE

```
[admin@AC1.POP4] > /tool traceroute use-dns=yes 172.16.0.14 count=10
 # ADDRESS                          LOSS SENT    LAST      AVG     BEST   WORST
 1 vlan0111.p1.pop4.lab             0%   10    3.2ms        3      1.2     4.2
 2 vlan0108.p1.pop3.lab             0%   10    2.8ms      2.5      0.9     3.3
 3 vlan0105.p1.pop1.lab             0%   10    2.6ms      2.4      0.6     3.6
 4 vlan0104.p1.pop2.lab             0%   10    2.6ms      2.4      0.5     3.2
 5 vlan0106.p1.pop5.lab             0%   10    2.4ms      2.3      0.5     3.6
 6 loop0.ac1.pop5.lab               0%   10    2.7ms      2.4      0.6     3.6
```

More detail:

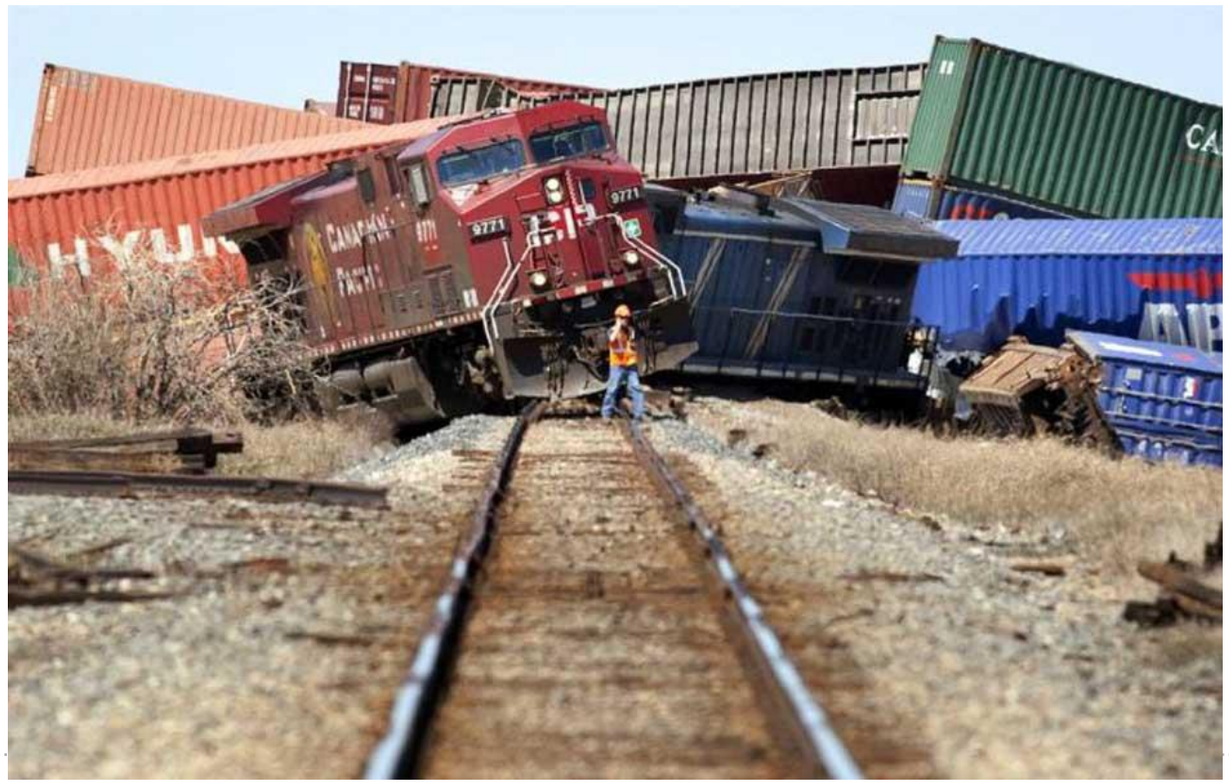
```
[admin@AC1.POP4] /ip route> print where dst-address=172.16.0.14/32

Flags: X - disabled, A - active, D - dynamic,
C - connect, S - static, r - rip, b - bgp, o - ospf, m - mme,
B - blackhole, U - unreachable, P - prohibit
 #      DST-ADDRESS        PREF-SRC        GATEWAY            DISTANCE
 0   S  172.16.0.14/32                     traffic-eng1              1
 1 ADo  172.16.0.14/32                     172.16.1.46             110
```

- In this example I changed the path the traffic takes to the Loopback of AC1.POP5.
- Its important to highlight the consequences of this action. As we are using BGP with source of Loopbacks – Any prefix advertised from AC1.POP5 into BGP will take the TE path. Unfortunately in my testing I found this broke things.

Limitations

- Influencing the IP Routing table means all traffic for that destination prefix (be it a loopback or a subnet) will use the traffic engineering tunnel.
- Be careful to not accidentally advertise TE prefixes to your IGP. Things could get, uh, interesting…



RSVP-TE Influencing VPLS

- VPLS will automatically use an RSVP-TE tunnel if its present.
- It will still signal via LDP (targeted mode)
- This allows for better utilisation of links that might otherwise be idle or provide a better (or worse) experience for customers.

Defining the TE paths

- I’ve built paths on P1.POP4 and P1.POP5

```
[admin@P1.POP4] > /mpls traffic-eng tunnel-path
add name=IGP reoptimize-interval=30m
add hops=172.16.0.7:loose name=TO-P1.POP5-SHORT use-cspf=no
add hops=172.16.0.5:loose, 172.16.0.3:loose, 172.16.0.4:loose,
172.16.0.7:loose name=TO-P1.POP5-LONG use-cspf=no
```

```
[admin@P1.POP5] > /mpls traffic-eng tunnel-path
add name=IGP reoptimize-interval=30m
add hops=172.16.0.6:loose name=TO-P1.POP4-SHORT use-cspf=no
add hops=172.16.0.4:loose, 172.16.0.3:loose, 172.16.0.5:loose,
172.16.0.6:loose name=TO-P1.POP4-LONG use-cspf=no
```

Building the TE tunnels

- I've built the traffic engineering tunnels on P1.POP4 and P1.POP5

```
[admin@P1.POP4] > /interface traffic-eng
add disabled=no name=traffic-eng1 primary-path=TO-P1.POP5-LONG record-route=no
secondary-paths=TO-P1.POP5-SHORT, IGP to-address=172.16.0.7

[admin@P1.POP4] > /tool traceroute use-dns=yes 172.16.0.7 count=10
 # ADDRESS                          LOSS SENT    LAST     AVG    BEST   WORST
 1 loop0.p1.pop5.lab                  0%   10    0.3ms     0.5     0.3     0.9
```

```
[admin@P1.POP5] > /interface traffic-eng
add disabled=no name=traffic-eng1 primary-path=TO-P1.POP4-LONG record-route=no
secondary-paths=TO-P1.POP4-SHORT, IGP to-address=172.16.0.6

[admin@P1.POP5] > /tool traceroute use-dns=yes 172.16.0.6 count=10
 # ADDRESS                          LOSS SENT    LAST     AVG    BEST   WORST
 1 loop0.p1.pop4.lab                  0%   10    0.4ms     0.5     0.3     0.6
```

Building the VPLS on P1.POP4

```
[admin@P1.POP4] > /interface vpls
add cisco-style=yes cisco-style-id=10000 disabled=no l2mtu=1500 name=vpls1
remote-peer=172.16.0.7

[admin@P1.POP4] > /ip address
add address=10.0.0.1/30 interface=vpls1

[admin@P1.POP4] /interface vpls> monitor 0
       remote-label: 73
        local-label: 100
      remote-status:
          transport: traffic-eng1
transport-nexthop: 172.16.1.33
   imposed-labels: 58,73
```

Building the VPLS on P1.POP5

```
[admin@P1.POP5] > /interface vpls
add cisco-style=yes cisco-style-id=10000 disabled=no l2mtu=1500 name=vpls1
remote-peer=172.16.0.6

[admin@P1.POP5] > /ip address
add address=10.0.0.2/30 interface=vpls1

[admin@P1.POP5] /interface vpls> monitor 0
       remote-label: 100
        local-label: 73
      remote-status:
          transport: traffic-eng1
 transport-nexthop: 172.16.1.25
    imposed-labels: 58,100
```

Testing the VPLS

- Run a duplex bandwidth test over the VPLS between P1.POP4 and P1.POP5.
- I've set MTU to 1400 because I can't do more than 1500 L2MTU in the lab.

```
[admin@P1.POP5] > /tool bandwidth-test 10.0.0.1 local-tx-speed=5M
remote-tx-speed=5M duration=10s user=admin protocol=udp local-udp-tx-
size=1400 remote-udp-tx-size=1400 direction=both

tx-current: 4.9Mbps
rx-current: 4.9Mbps
lost-packets: 0
direction: both
tx-size: 1400
rx-size: 1400
```

- Here is the 5M/5M being label switched by P1.POP1:

```
[admin@P1.POP1] > /tool torch ether2
        TX          RX TX-PACKETS RX-PACKETS
  10.2Mbps   10.2Mbps         922         925
  10.2Mbps   10.2Mbps         922         925
```

Live Demo/Questions?

Thanks ☺