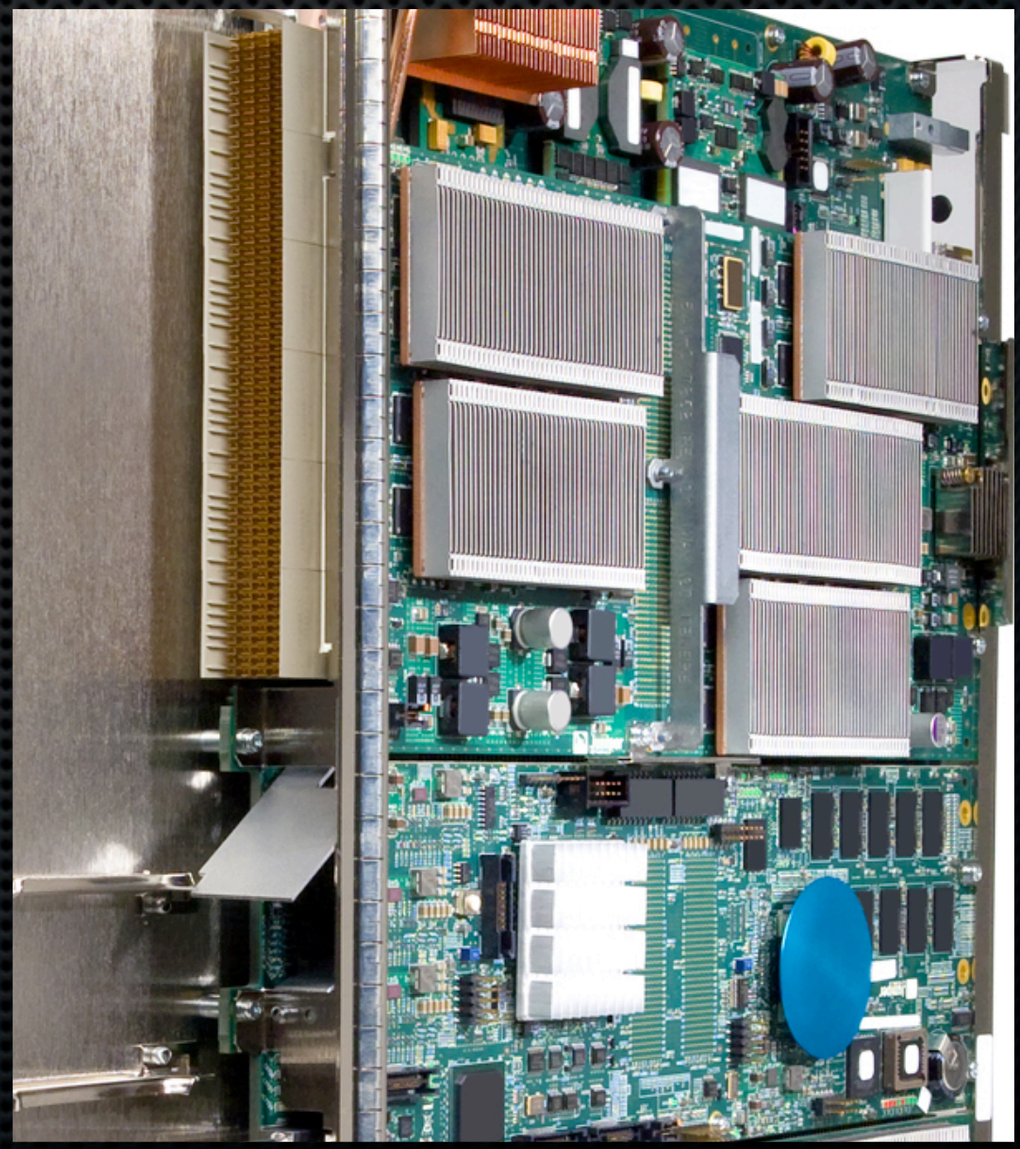
PIM-tunnels and MPLS P2MP as Multicast data plane in IPTV and MVPN

Lesson learned

Speaker

Rafał Szarecki

- JNCIE-M/T #136, JNCIP-E #106, JNCI
- rafal@juniper.net
- Curriculum
  - Juniper Networks; Professional Services Consultant - EMEA
    - Telenor, Telefonica, BT Media & Broadcast, etc.
  - Ericsson (Poland). Solutions EMEA
  - Polkomtel
  - NASK



Agenda

- What Experience
- Two customers profiles and overall solutions
- Plain IP multicast with PIM
- P2MP MPLS LSP

What Experience

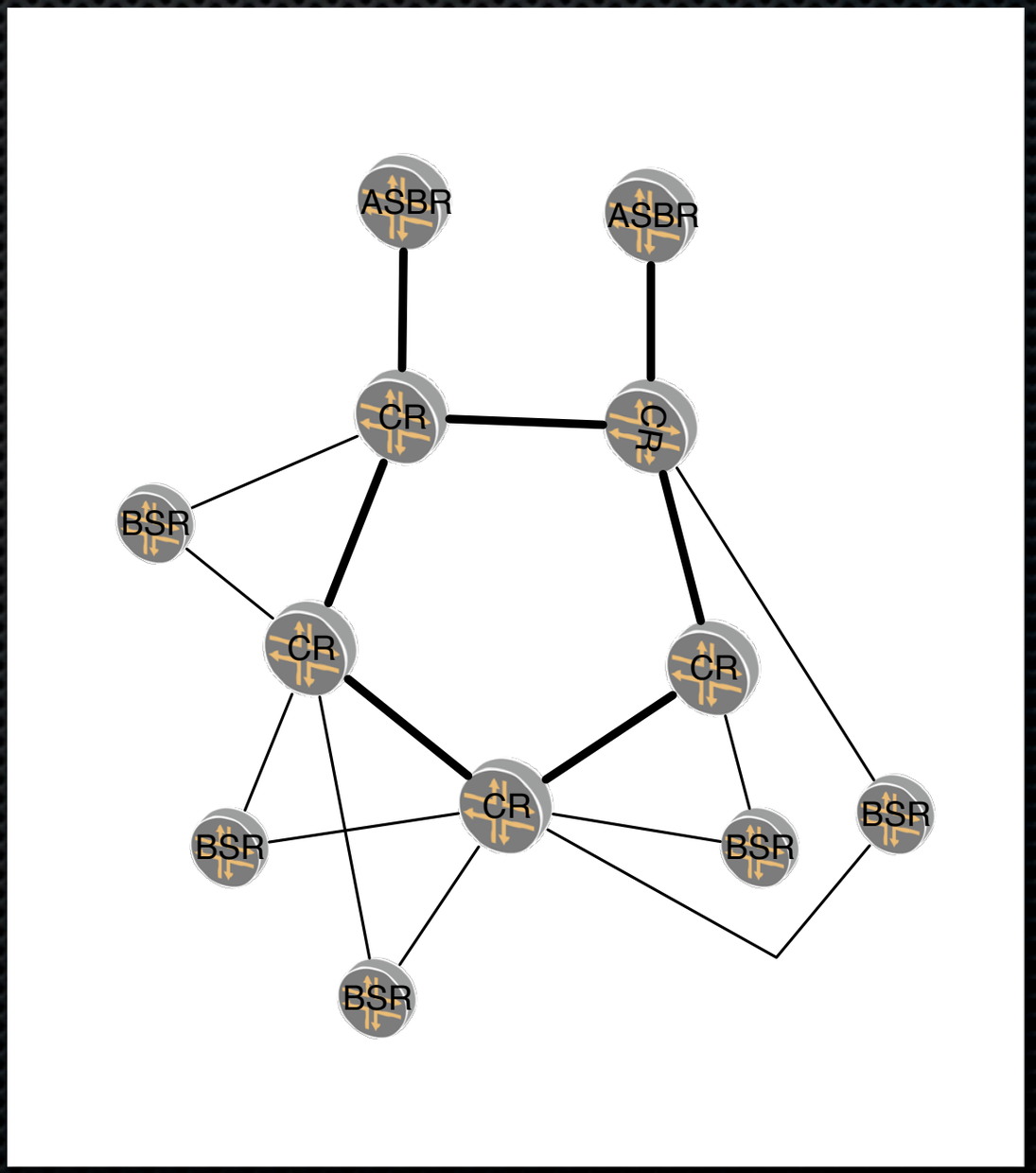
- I have a part in number of projects in 2007 and 2008
  - West Europe
    - Some of them are already in Commertial stage
- This talk will be base on two
  - technologically different
  - Both are commercially rolled out, and in use.

Customer A

- 2nd ISP in a country
- project goal - deliver IPTV to xDSL subscribers
- Use of plain IP multicast was mandatory - political reasons (the board decision)
- 5 core routers in 10GE ring
- About 15 BSRs. Dual-homed.
- Multicast streams received from other AS:
  - over same connection as Internet.
  - two connections for redundancy.

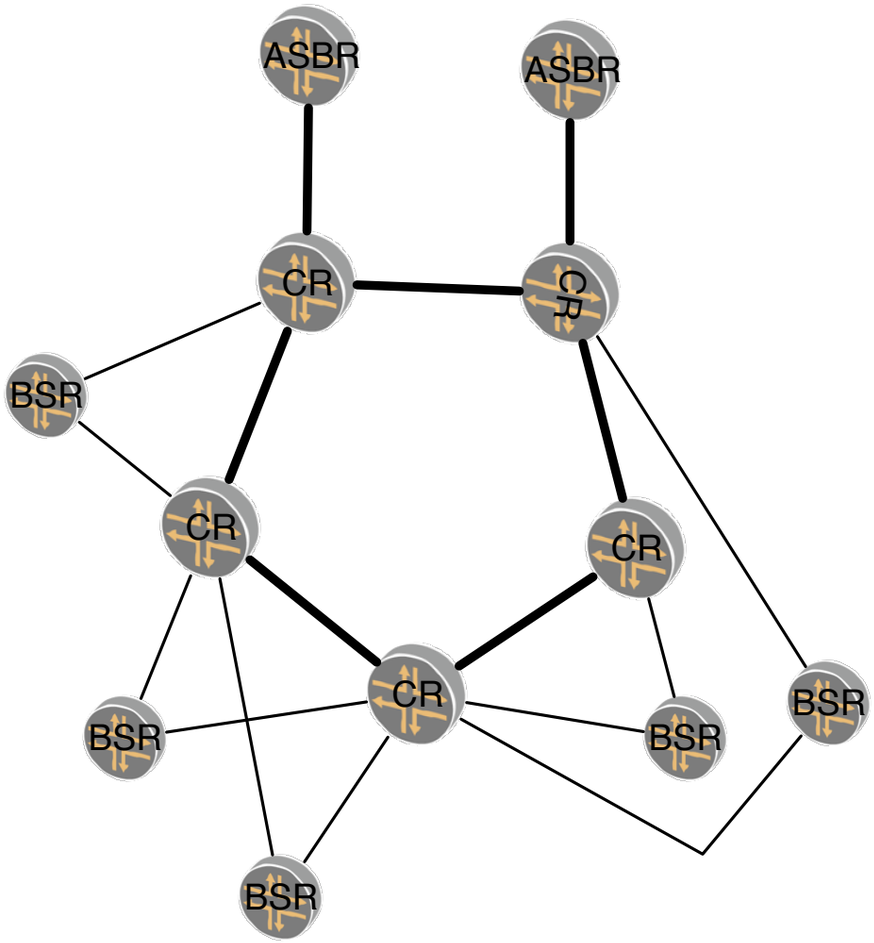
Customer A
How it works (1)

- RPF routes distributed as MBGP (AFI1, SAFI2)
  - same instance as Internet
  - CRs are a route reflectors for connected BSRs
  - CRs and ASBR are in full mesh
- OSPF is IGP.
  (0/0 Ext. LSA in IGP)
- MPLS LDP on all CRs, and BSRs.



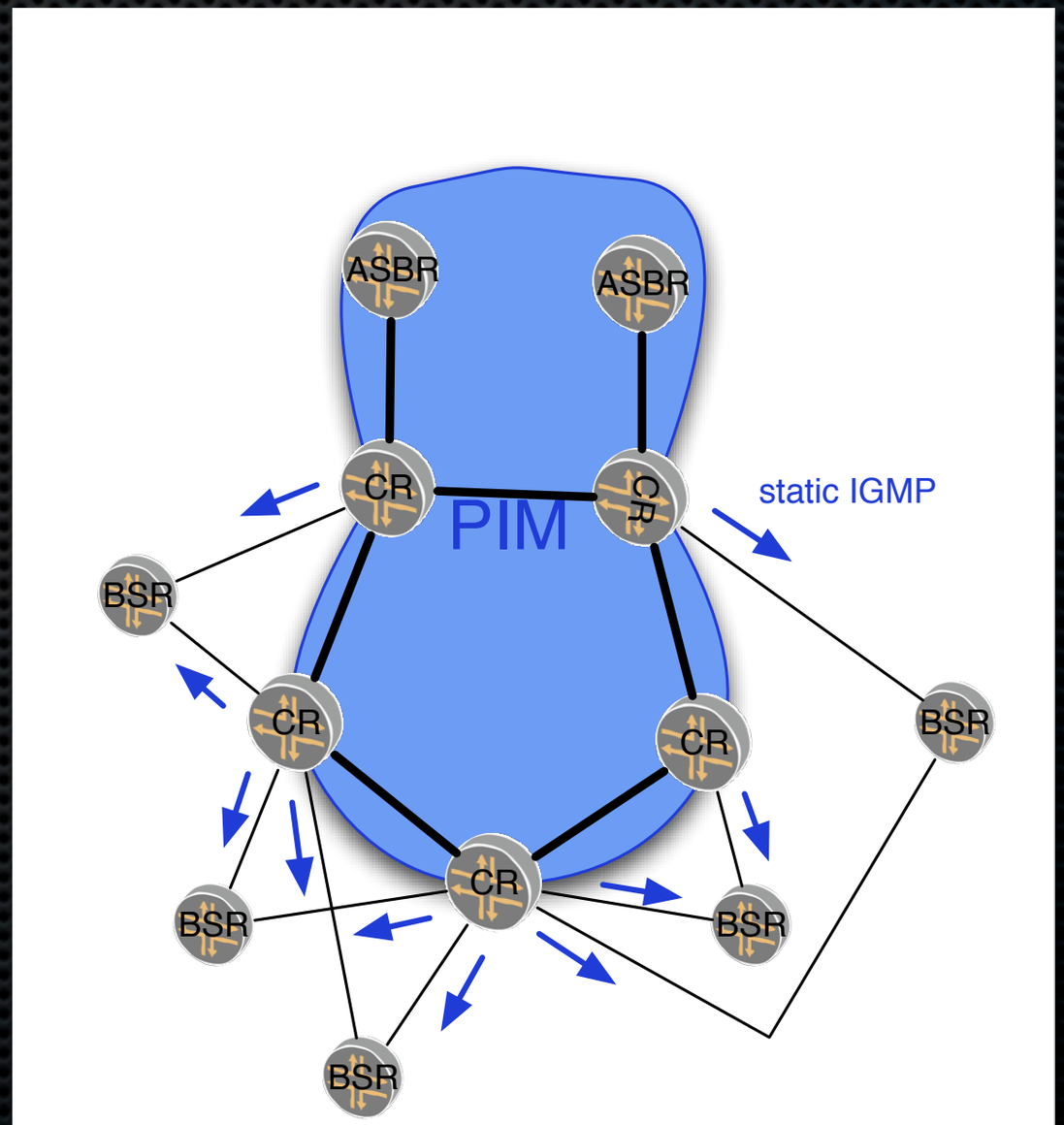
Customer A
How it works (2)

- Design challenges
  - IPTV is very sensitive for losses.
  - BSR <==> CR links are not so stable
  - BSR <==> CR can be over L2 infrastructure. Some failure in transmission not visible to BSR and CR.
  - BSRs are significantly less stable then CR.
  - If ASBR fail, BGP RPF routes learned by CR from this ASBR remains active until BGP keepalive expires (90 sec).



Customer A
How it works (2)

- Challenges Addressed
  - Feed all multicast data to BSR on each of 2 uplinks permanently
    - BSR restore IPTV as soon as new RPF IIF is elected.
    - No need to wait for PIM signaling.
  - BFD for OSPF between CR and BSR. Detect loss of connectivity fast (in some cases).
  - Static IGMP SSM reports on CR.
    - Multicast states in core do not depend on BSRs state (fails).
    - Very long config. Exposed for errors.
      (JNPR CLI apply-groups helps)
  - iBGP session between CR and ASBR, backed by BFD. (to not wait for BGD 3xKeepaLive

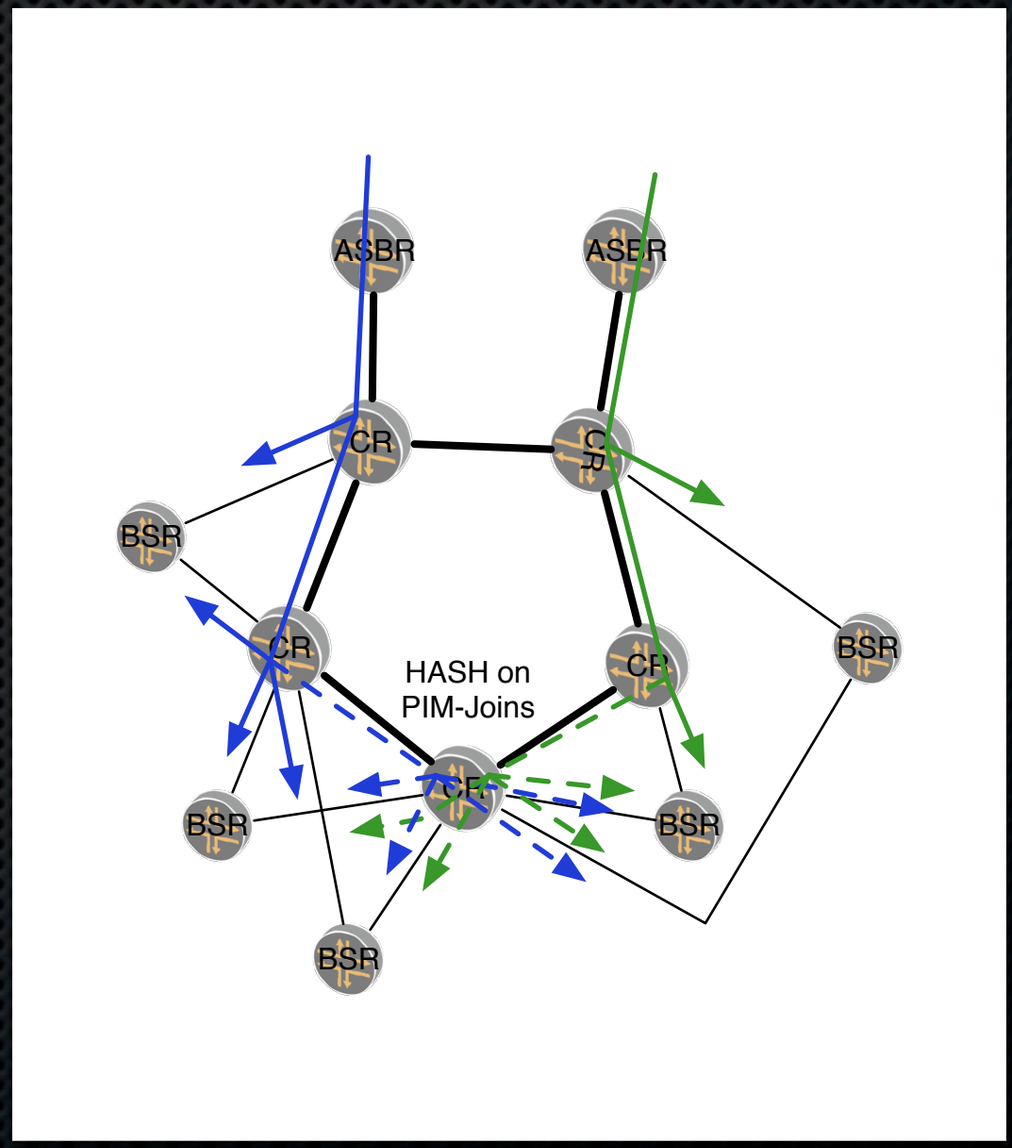


Customer A BSR feeding

- Redundancy in ACTIVE-ACTIVE fashion
- Switchover do not affect core of network. No PIM-joins to CR and between CRs.
- Resources consumed on both uplinks.
  - Uplinks are GE
  - IPTV is projected to be ~ 600Mbps
  - Only 2 x 400 Mbps remains for unicast. But LB can’t be equal - topology.
- Fail of uplink (measured):
  - If BSR sees LoL on his RPF uplink, switchover is fast - 170 ms
  - If BFD (3x100 ms interim) detects fail, switchover is about 380 ms
- Uplink restoration (measured):
  - RPF interfaces is updated (reverted) in about 60 ms losses.

Customer A Multicast in core

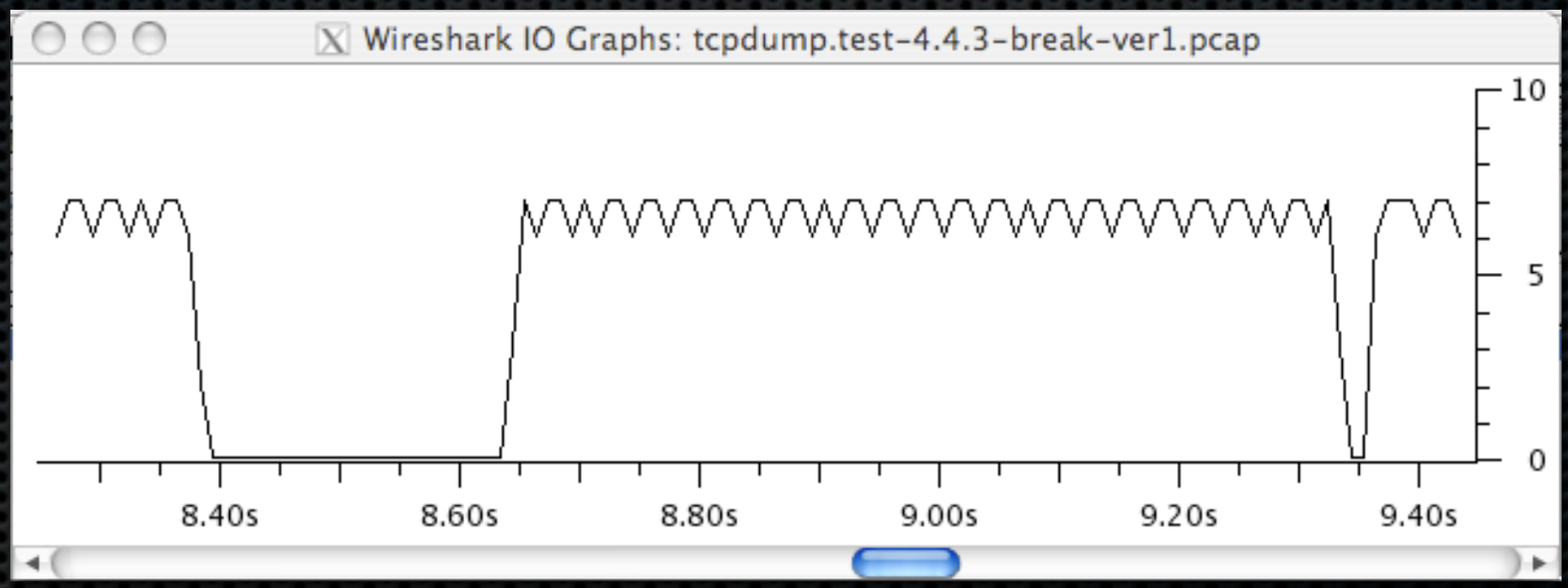
- ACTIVE-ACTIVE not possible with IP MULTICAST in ring topology.
  - Same (S,G)/(*,G) states for both feeds - distribution tree can't cross same link/node.
- Regular PIM-SSM used.
  - On one of CR RPF interfaces for given (S,G) is elected base on hash. (Operationally complex)
  - Traffic restoration after link/node failure requires convergency of OSPF, and then PIM
  - Traffic restoration after link/node REPAIR requires convergency of OSPF, and then PIM.



PIM convergency fingerprint

Link fail:

260ms+30ms traffic loss
950 ms instability
Few second of IPTV loss



Link repaired:

30ms+280ms+290ms
traffic loss
50 (3) sec instability
Two artifacts on IPTV.

How 0.5 sec traffic loss is manifested on IPTV

~500 ms loss

~250 ms loss

How 0.5 sec traffic loss is manifested on IPTV

~500 ms loss

~250 ms loss

How 0.5 sec traffic loss is manifested on IPTV

~500 ms loss

~250 ms loss

PIM tunnels

- Used in MVPN as per draft-rosen
- Also an option in NG-MVPN
- Signaling in Core - PIM. ==> Same behavior as plane IP Multicast w/ PIM
- Data-plane in core - IP (C-mcast) over GRE over IP (SP-mcast). ==> Same behavior as plane IP Multicast.

PIM-base Operation and Maintenance

- PIM states are effect of PIM-join/prune. Independent from traffic.
- Forwarding states are a cache entries - triggered by traffic and subject to time-out.
  - Make troubleshooting vary hard in practice - you need traffic to see something on data-plane.
  - Needs refreshment
  - No such thing as steady state
- Many inter-protocol dependencies - hard to manage:
  - PIM-JOIN is send out of interfaces selected by RPF
  - RPF interfaces depends on MP-BGP
  - MP-BGP depends on IGP

Customer B

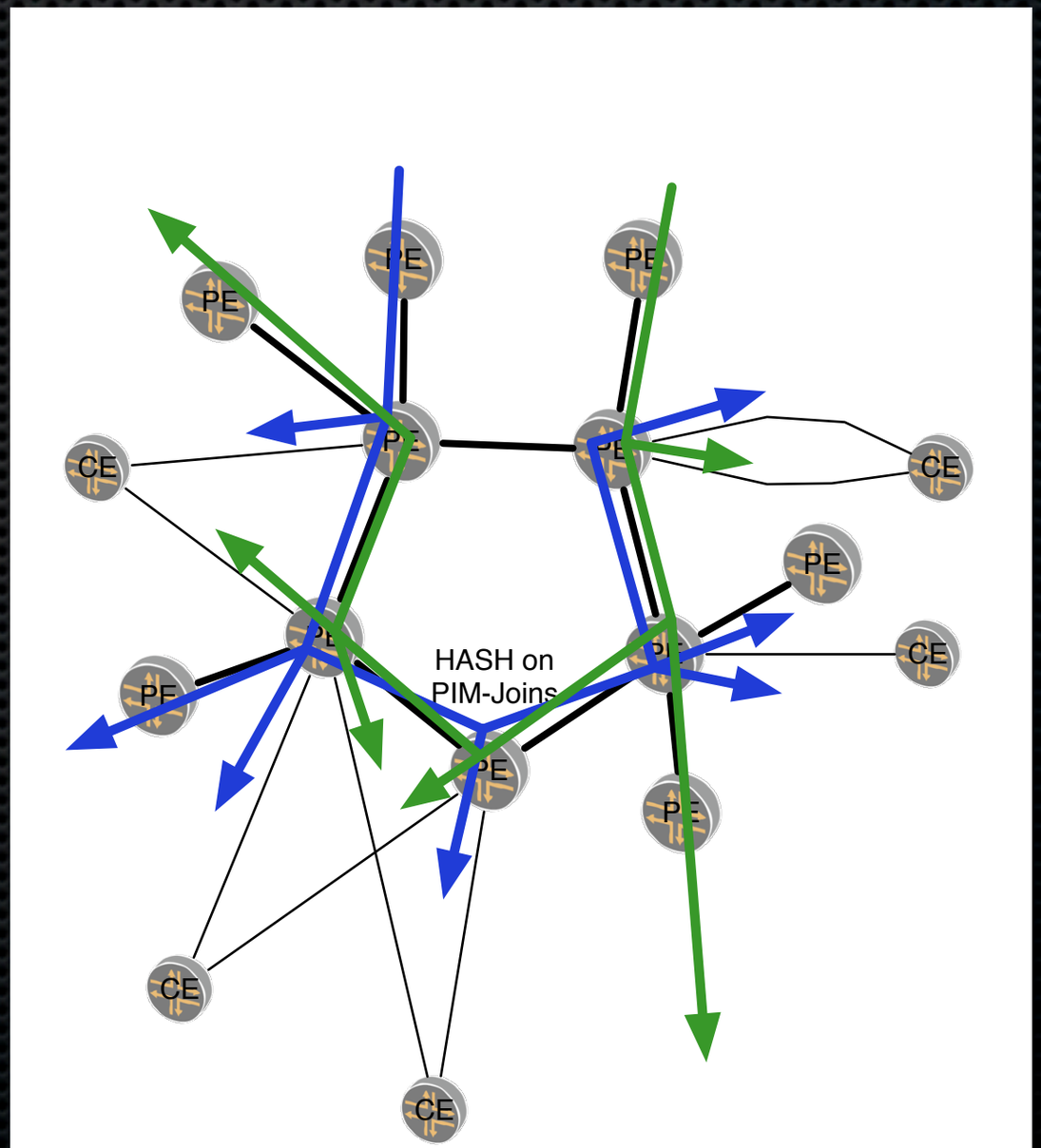
- Biggest in country. Country bigger then PL.
- Business - distribution of digital TV signals.
  - In country, and
  - around the Globe
  - Customer has used IP/MPLS network for this purposes for years.
- Project driver: DVP-T switchover.
  - Customer want to deliver DVB-T signal to every Broadcast antenna in country.
  - For multiple Broadcasters.
- Network:
  - More then 500 routers in a network.
  - Single-plain core

Customer B
Design requirements

- ACTIVE-ACTIVE distribution for some premium channels
  - Single plane topology - no “natural” demarcation
  - Scale to several thousands “leaf” routers.
  - Separation of broadcasters
  - Complex, reach topology, but… some connection are expected to be too narrow.
    - directing channels individually over topology.

Customer B Design basics (SP core)

- Next-Generation MVPN
  - **draft-ietf**-l3vpn-2547bis-mcast-07, **draft-ietf**-l3vpn-2547bis-mcast-bgp-05.txt
  - MP-BGP for multicast signaling in SP-domain
  - MPLS P2MP instate S–PMSI
  - Set of dedicated RR.
- MPLS P2MP
  - explicit staticERO calculated off-line
    - diverse path for ACTIVE-ACTIVE
    - link BW usage control
  - BW constrain (backed by CSPF) - double-protection
  - Link-Protection by facility backup - non ACTIVE-ACTIVE streams sub-second restoration

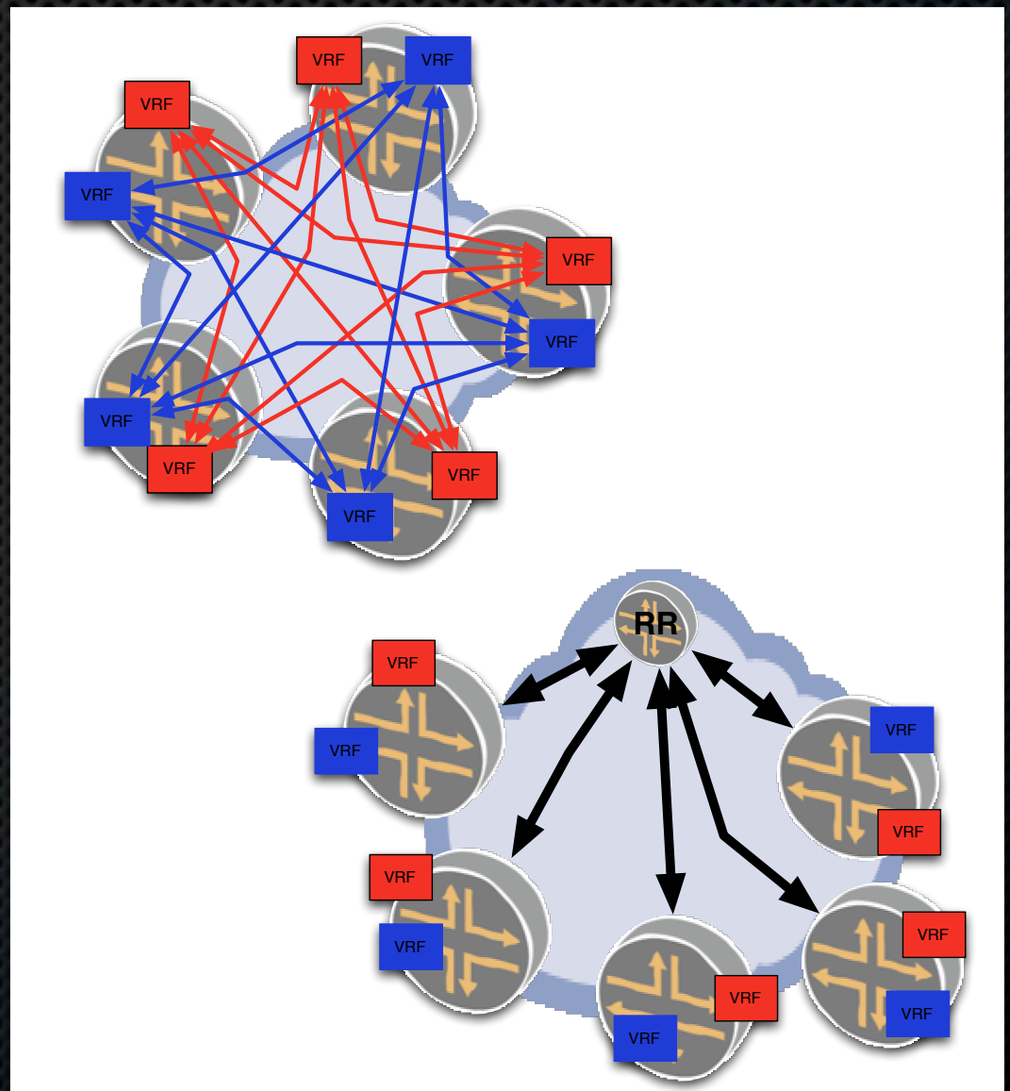


Customer B
Design basics (PE - CE)

- CE managed by SP
- On PE access interface static IGMP join (SSM).
  - Scaling
  - Core stability
- eBGP toward VRF for RPF routes.

Customer B
Why MP-BGP?

- IPTV signaling not need to be ultra-fast - once channels are signaled, they remains for ever.
- Control-plane scaling - only 2 BGP session needed on “leaf” PE and “head-end” PE
- NOTE:draft-rosen requires PIM session beween each two VRF in VPN - VR model rather then 2547bis.
  - 500 PE & 5 VPNs ==> 2.495 PIM sessions on each PE toward other PE + as required by PE-CE routing



Common understanding among SPs

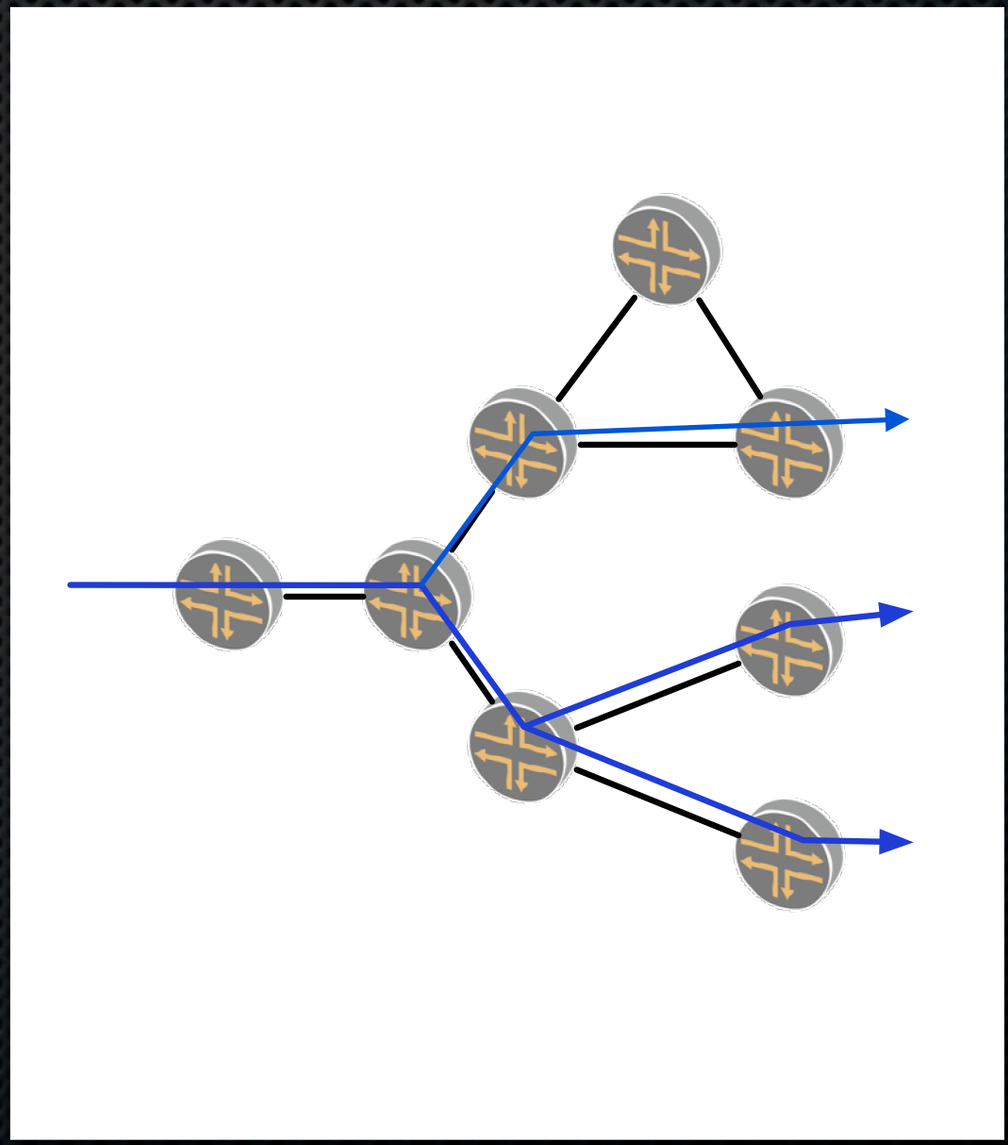
- **draft-ietf-l3vpn-mvpn-considerations-01**
- Authors: BT, FT, NTT, DT, Verizon. No vendors involved.
- Signalling:
  - Auto-discovery - MP-**BGP**
  - S-PMSI signaling - **BGP** preferred
  - PE-PE C-multicast routing - **BGP** or modified PIM (implementation do not exist)
- Data-plane:
  - tradeoff between resource usage optimization and simplicity
  - P2MP **RSVP-TE** gives **better** resource **utilization**, and **FRR** capability.
  - mLDP and GRE/IP-multicast are easy to provision.

Customer B
Data-plane behavior

- Link-loss covered by MPLS FRR and then make-before-break branch-LSP re-signaling.
  - Platform dependent, ~10ms/~25ms
- some node's control plane issues covered by GPR
  - Loss-less on transit and egress nodes
  - GPR of RSVP do not cover head-end (no helper) by IETF RFC.
    - IGP do not know about issue - GPR
    - P2MP LSP re-establishing base on RSVP timers.
  - Is not an issue for streams protected in ACTIVE-ACTIVE fashion
  - For others NG-MVPN designated forwarder election allows to connect active and backup sources to different ingress PEs. Traffic form only one will be forwarded.
- Light-out of the transit node covered by IGP convergency and branch-LSP re-signaling.

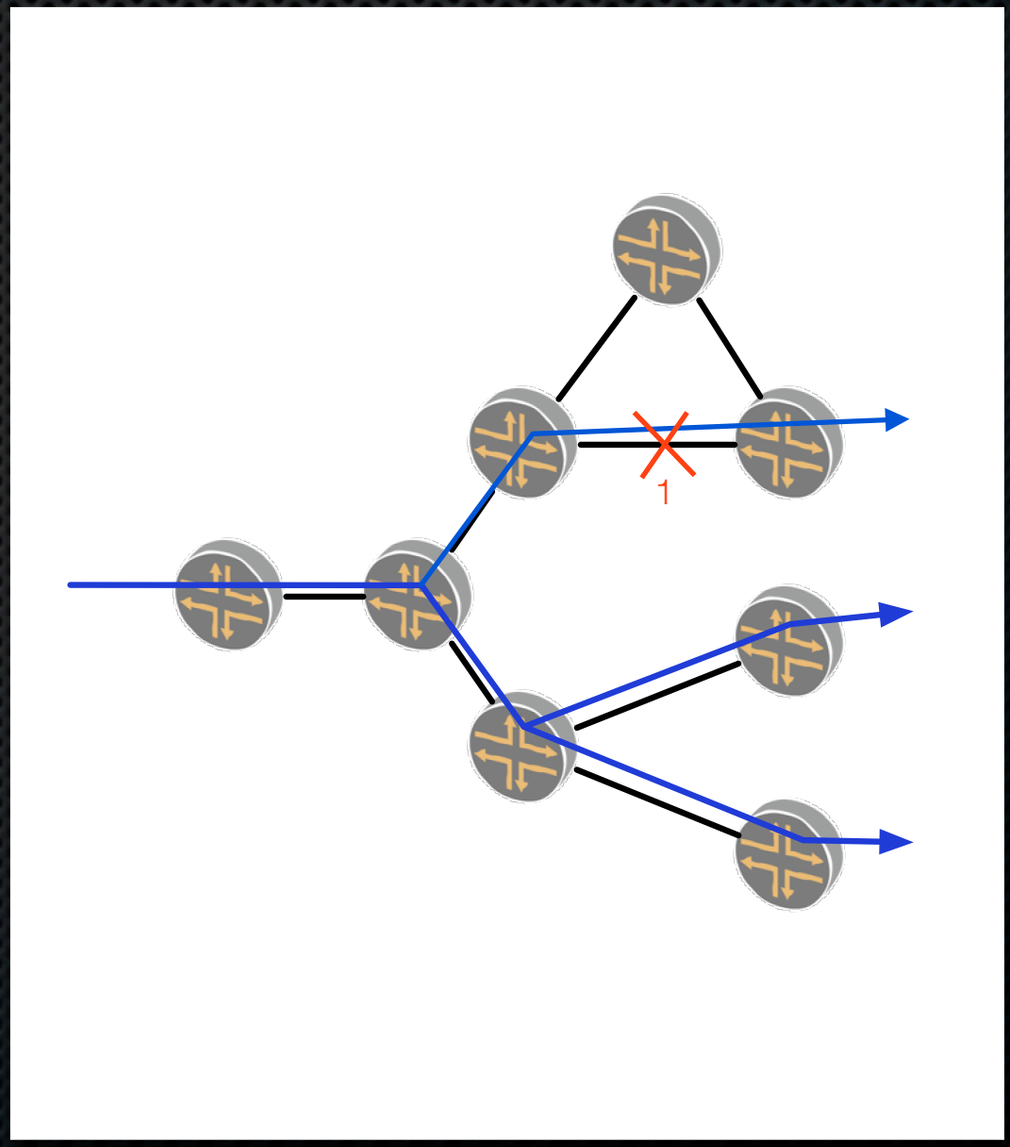
P2MP branch re-signaling

- No FRR in this example
- Southern branches are not affected at all.
- This is fail scenario, but can be others:
  - new staticERO
  - new BW request
  - etc.



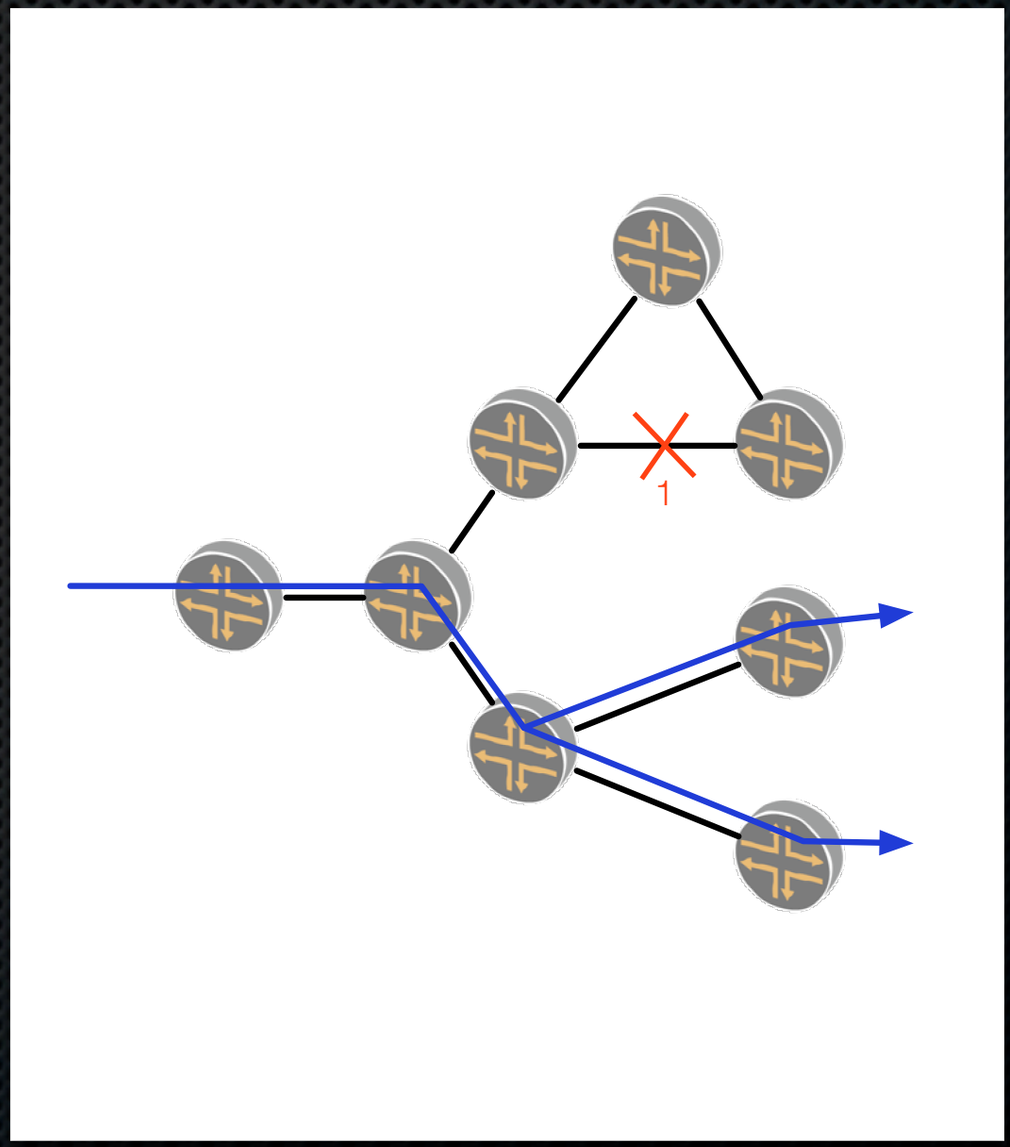
P2MP branch re-signaling

- No FRR in this example
- Southern branches are not affected at all.
- This is fail scenario, but can be others:
  - new staticERO
  - new BW request
  - etc.



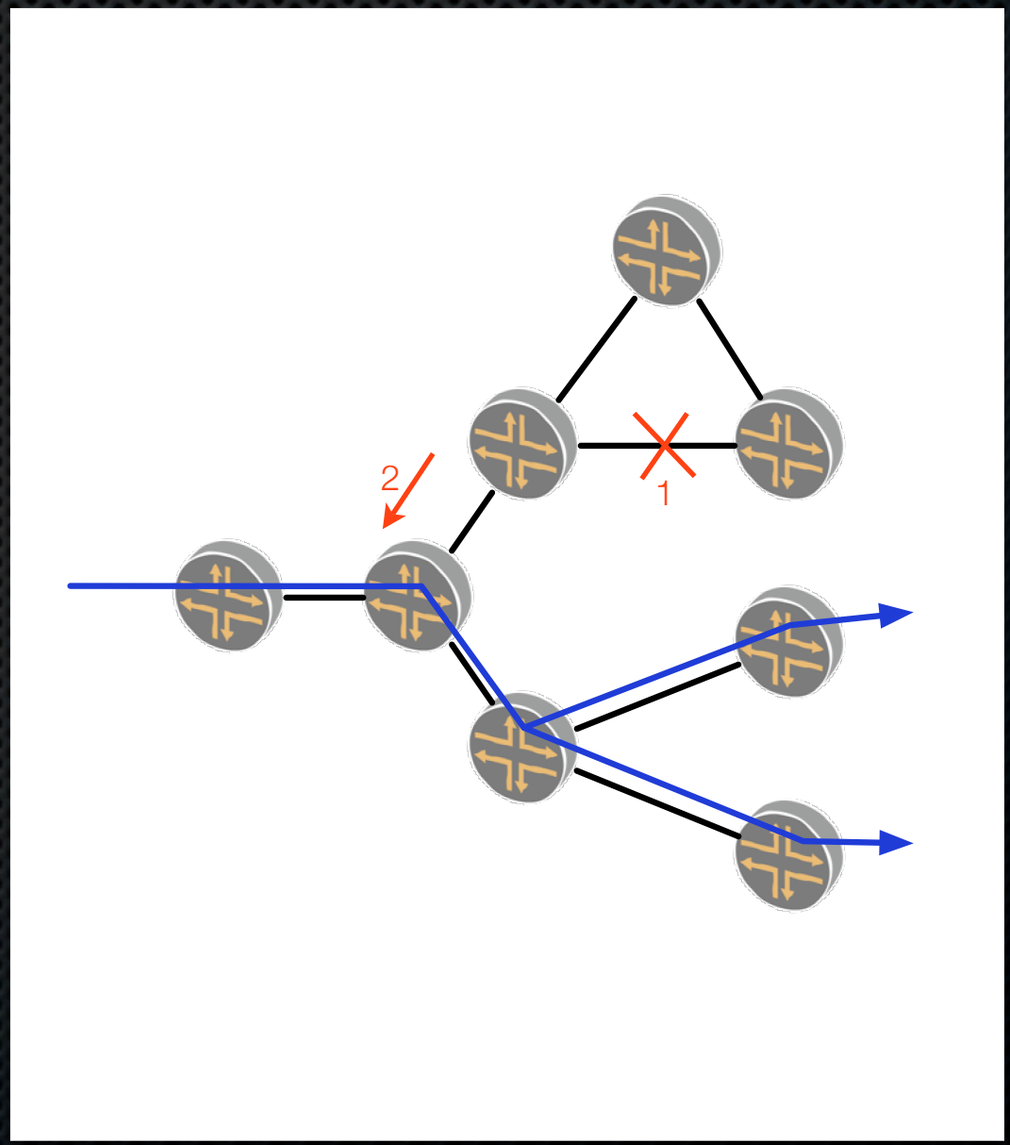
P2MP branch re-signaling

- No FRR in this example
- Southern branches are not affected at all.
- This is fail scenario, but can be others:
  - new staticERO
  - new BW request
  - etc.



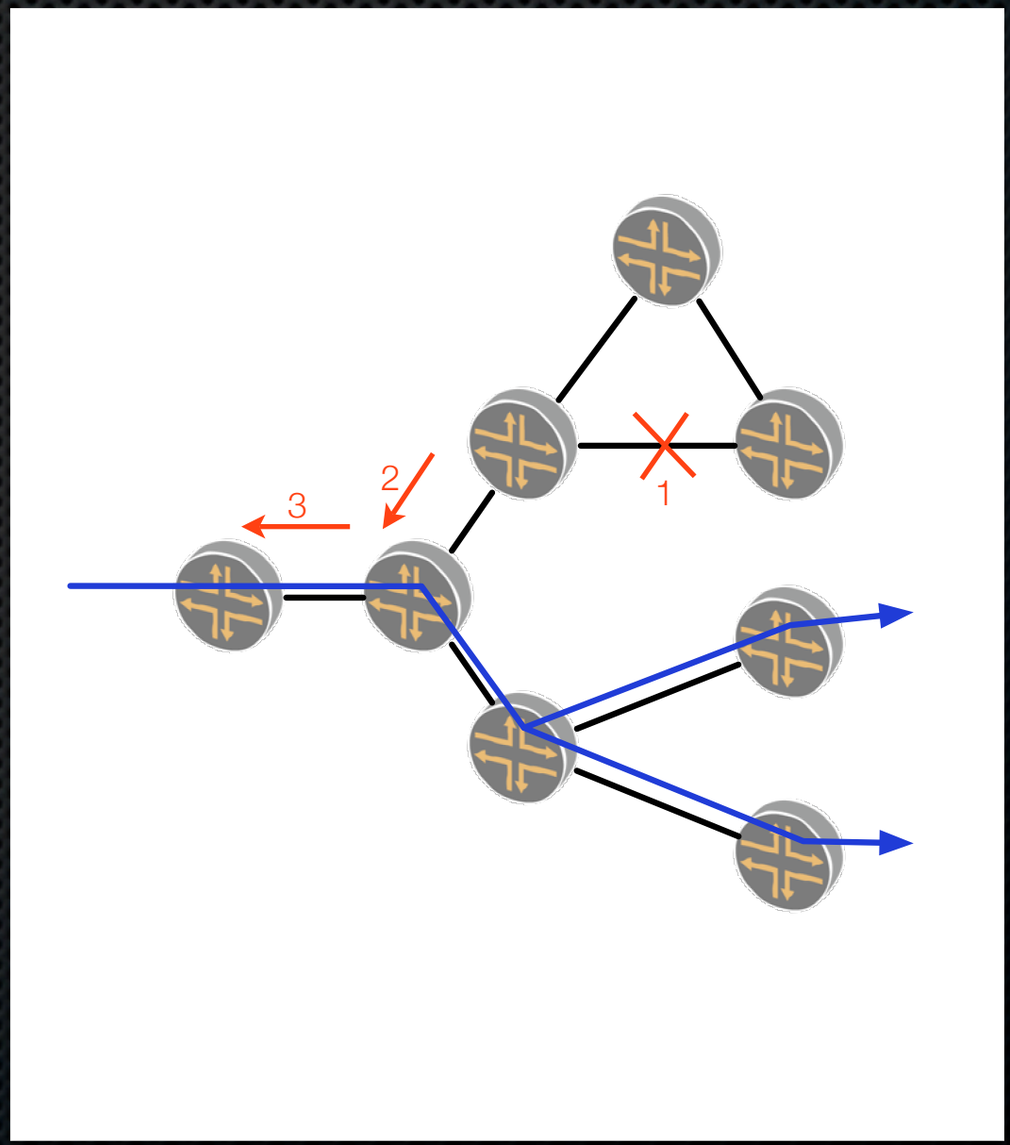
P2MP branch re-signaling

- No FRR in this example
- Southern branches are not affected at all.
- This is fail scenario, but can be others:
  - new staticERO
  - new BW request
  - etc.



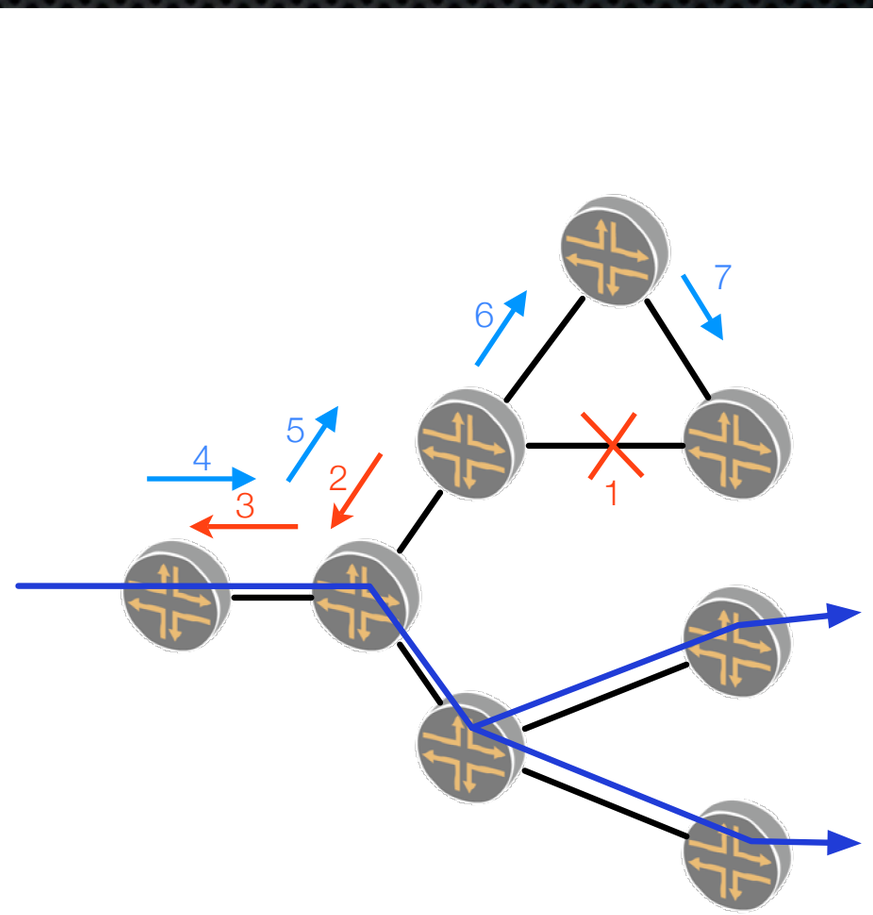
P2MP branch re-signaling

- No FRR in this example
- Southern branches are not affected at all.
- This is fail scenario, but can be others:
  - new staticERO
  - new BW request
  - etc.



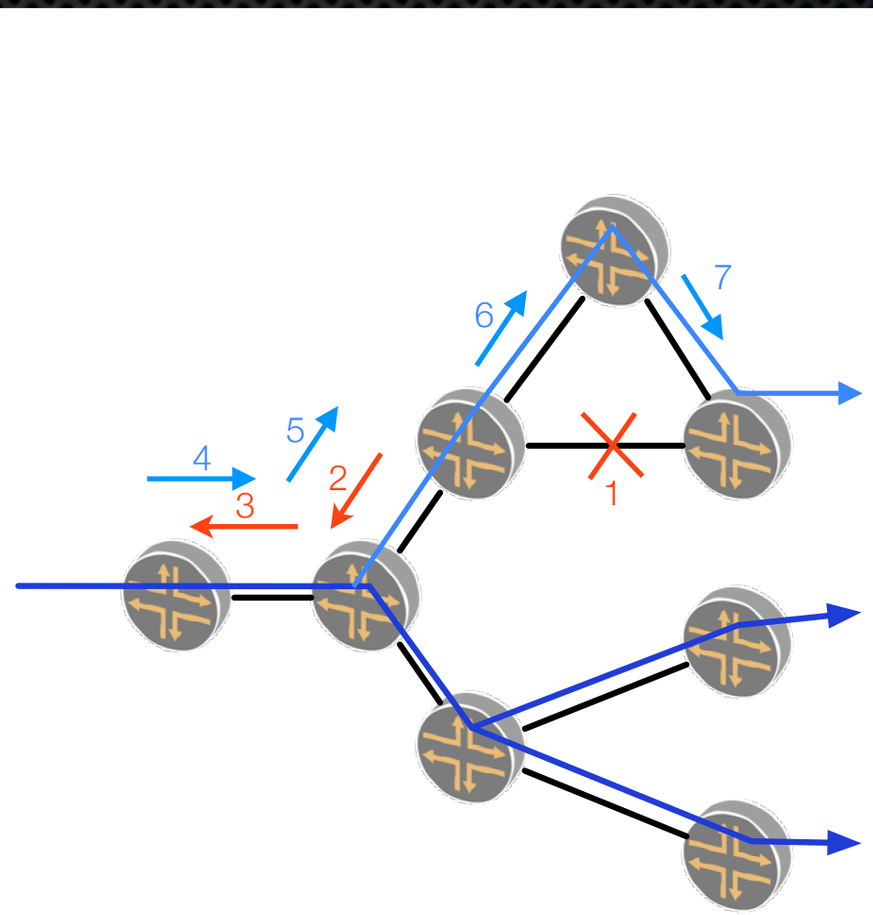
P2MP branch re-signaling

- No FRR in this example
- Southern branches are not affected at all.
- This is fail scenario, but can be others:
  - new staticERO
  - new BW request
  - etc.

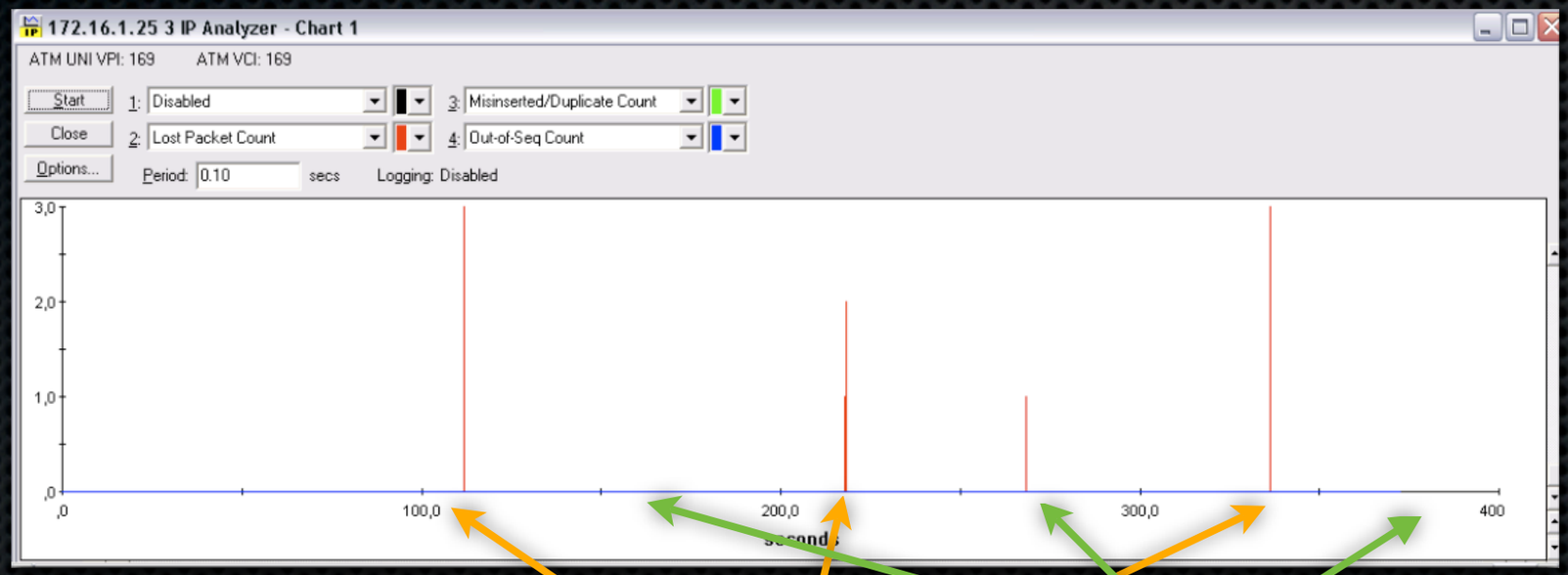


P2MP branch re-signaling

- No FRR in this example
- Southern branches are not affected at all.
- This is fail scenario, but can be others:
  - new staticERO
  - new BW request
  - etc.



P2MP FRR restoration fingerprint



- Three runs of link-fail/ link-repair test.
- The single channel streaming at 267pps

- Link fail
  - ~10ms loss == 10ms instability
- Link repair
  - Mostly loss-less
  - 1 packet loss can happens
  - thanks to make-before-break

How 0.01 sec traffic loss is manifested on IPTV

- Impact on IPTV watchers depend on type of lost MPEG frame
  - I-frame lost.
    - Worst case
    - Resynchronization needed.
    - Can take few seconds
    - probability is ~20%
  - non-I-frames are lost
    - small artifact
    - typical audio degraded but still understandable.
- 11 ms loss clip.

How 0.01 sec traffic loss is manifested on IPTV

- Impact on IPTV watchers depend on type of lost MPEG frame
  - I-frame lost.
    - Worst case
    - Resynchronization needed.
    - Can take few seconds
    - probability is ~20%
  - non-I-frames are lost
    - small artifact
    - typical audio degraded but still understandable.
- 11 ms loss clip.

How 0.01 sec traffic loss is manifested on IPTV

- Impact on IPTV watchers depend on type of lost MPEG frame
  - I-frame lost.
    - Worst case
    - Resynchronization needed.
    - Can take few seconds
    - probability is ~20%
  - non-I-frames are lost
    - small artifact
    - typical audio degraded but still understandable.
- 11 ms loss clip.

Summary

PIM base IP multicast

- ACTIVE-ACTIVE protection possible only on very specific network topologies - dual-plane
- Traffic driven forwarding states. This complicates troubleshooting
- Basic requirements for transit nodes - just IP Multicast
- Traffic restoration depend on full IGP convergency and PIM-Join re-signaling

P2MP MPLS LSP

- ACTIVE-ACTIVE protection possible on virtually any topology. At cost of TE tools and work.
- Pre-signaled forwarding states. You can verify corectness of all states befor traffic arrive.
- All routers on path needs to understand MPLS P2MP signaling
- MPLS TE capabilities - BW reservation, ERO, colors, etc.
- ~50 ms traffic restoration possible. very good effects in 80% of cases.

Qestions and Answers (eventually)