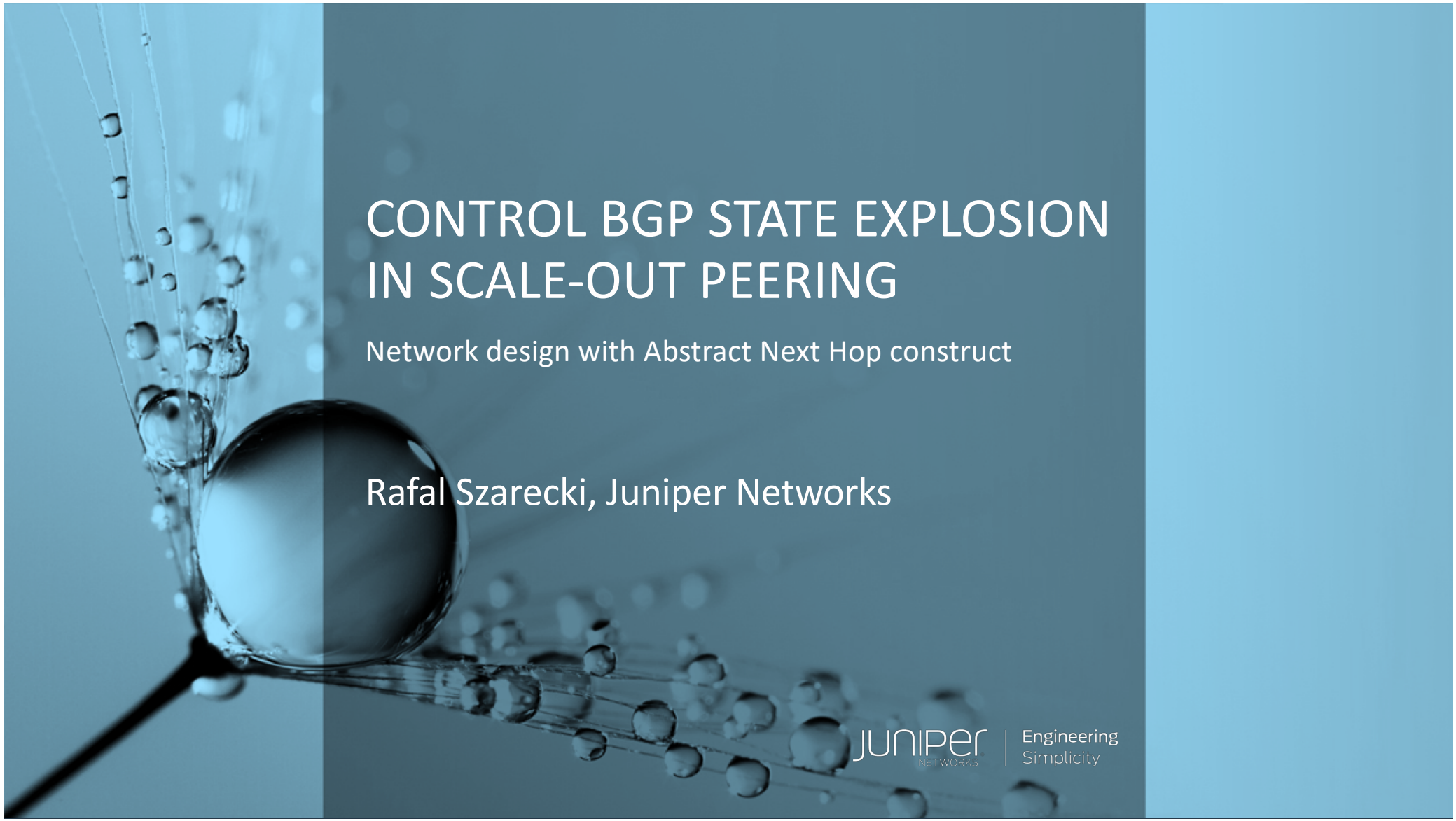


# CONTROL BGP STATE EXPLOSION IN SCALE-OUT PEERING

Network design with Abstract Next Hop construct

Rafal Szarecki, Juniper Networks

JUNIPER NETWORKS | Engineering Simplicity

# AGENDA

Scale-out peering:

- Why
- What it is and what it is not
- Principles of operation

Control-Plane challenges

- BGP Scale
- Common practice limitations

Solution

- Abstract Next-Hop (ANH)
- iBGP architecture w/ ANH
- Configuration examples

# AGENDA

Scale-out peering:

- Why
- What it is and what it is not
- Principles of operation

Control-Plane challenges

- BGP Scale
- Common practice limitations

Solution

- Abstract Next-Hop (ANH)
- iBGP architecture w/ ANH
- Configuration examples

# WHY SCALE OUT PEERING

## BUSINESS AND INDUSTRY

- Screen Time & interactive applications
- Bandwidth (B/W) + RoundTripTime (RTT) = experience
- Data consumer/producer in distinct networks.

## SCALE-OUT ATTRIBUTES

- resiliency
- capacity
- competition

## DENSE PEERING SURFACE

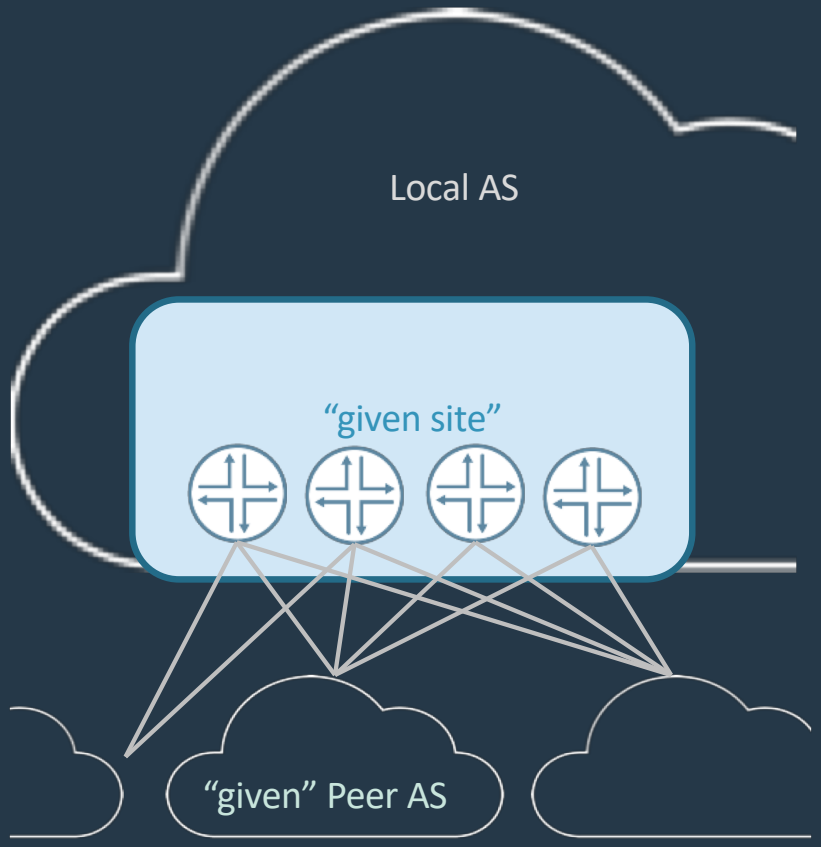
# SCALE-OUT PEERING

Given Peer AS is connected to multiple ASBRs at given site (of local AS)

- All ASBRs are set to construct ECMP toward given Peer AS (egress ECMP)
- All ASBRs sends equal routing information to given Peer AS (desired ingress ECMP)

Principles

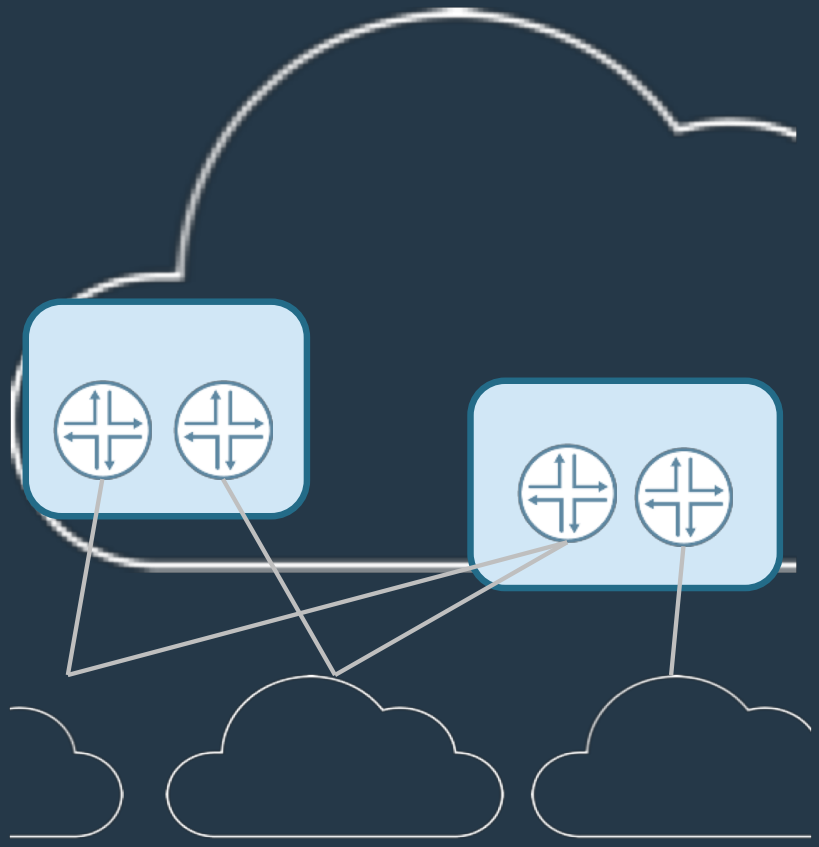
- Provides required B/W with ECMP
- Provides site-local N+1 redundancy



# SCALE-OUT PEERING ⊂ DISTRIBUTED PEERING

If given Peer AS is connected to ASBRs, each at different site (of local AS) → distributed but not scale-out.

Multiple ASBRs at given site, each serving disjoined set of  Peer ASes → distributed not scale-out.

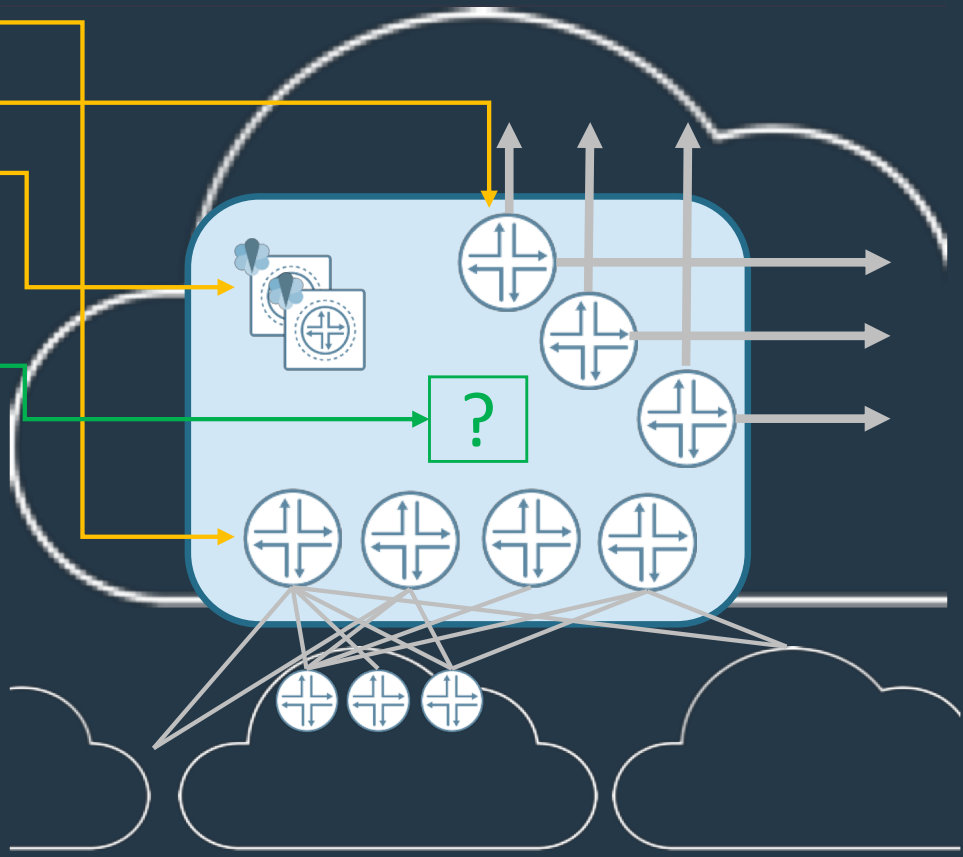


# SCALE-OUT PEERING SITE

Multiple (N) ASBRs; 1-**2**-M Core Routers(CR); Route Reflectors(RR)

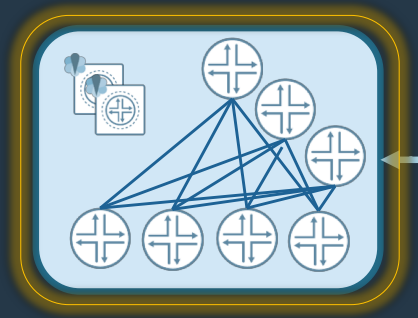
Something between

- Passive network
- Active nodes
  - L2 (vintage)
  - L3 underlay
  - L3
  - Many combination but some commonalities…

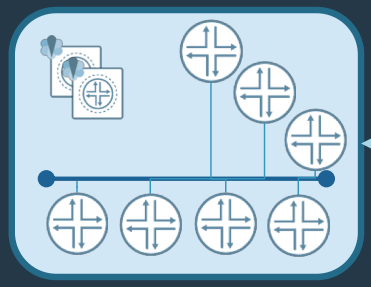


# SITE NETWORK - SOMETHING BETWEEN

passive connections -> Leaf/Spine-Spine/Leaf

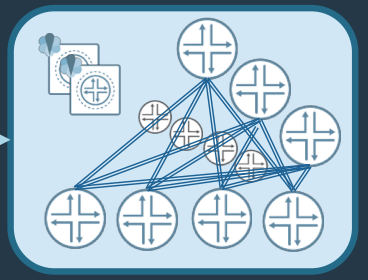


L2 switches in the middle

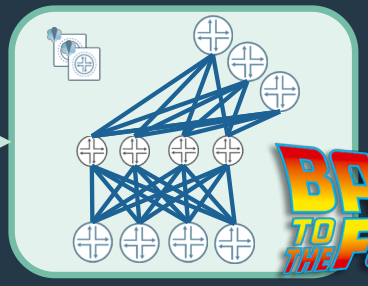


- CRs and ASBRs are L3 neighbors
- Low requirements on device in the middle

Leaf-Spine w/overlay



L3 SPINEs w/o overlay



- SPINEs participate in AS-wide routing
- Requirements could be high RIB/FIB, MPLS-TE, etc

CORE-DISTRIBUTION-ACCESS

# COMMON PRINCIPLES

Equal LB among ASBR – BGP multipath – RIB scale-up

N+1 redundancy model for:

- Inter-AS Interface failures
- Multiple eBGP session failure
- ASBR failure
- Site failure

Scale-out ASBRs w/o impacting other sites

- 2 x Route Reflector (RR) per site
- keep # session under control

Restore optimal traffic in seconds, not 10's minutes. Regardless of scale.

# AGENDA

Scale-out peering:

- Why
- What it is and what it is not
- Principles of operation

Control-Plane challenges

- BGP Scale
- Common practice limitations

Solution

- Abstract Next-Hop (ANH)
- iBGP architecture w/ ANH
- Configuration examples

# RIB SCALE-UP EXAMPLE

Network

- 40 sites
- 4 ASBR's per site
- 2 RR per site
- 500k external prefixes (pfx) per ASBR; 250K active
- 1M unique prefixes

80M path's overall @ peering surface;

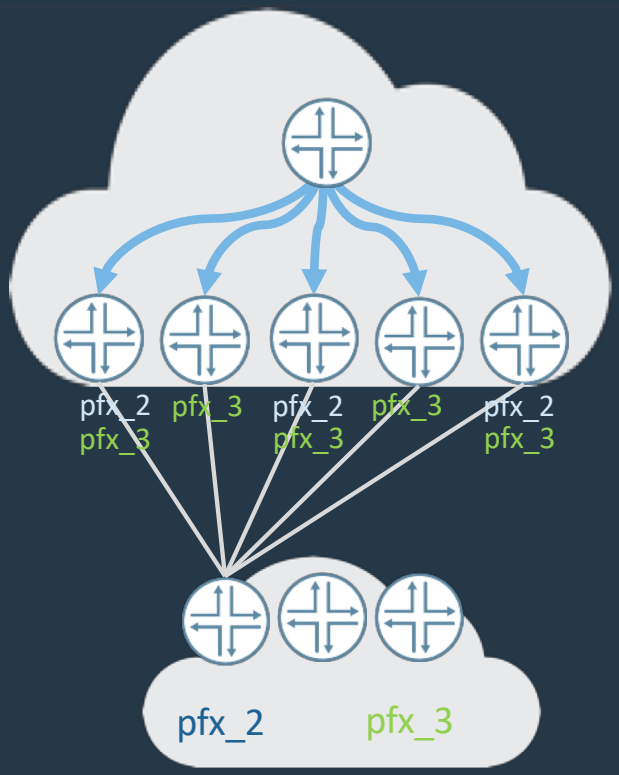
- **80 paths per prefix**
- 160M @ RR.
- 4 best (or more if same IGP cost to multiple sites; 8? 12?)
- 1... 4... (8?) backup
- 80%+ really not needed.

# THE PEER NETWORK

Don’t assume too much on Peer AS– non-consistent prefix-set

- Some prefix received on sub-set of ASBR but not on all.
- Tradeoff:
  - **Full (Internet) FIB on CR or SPINE – OR –**
  - Sub-optimal routing (redirection form one ASBR to other) -OR –
  - exposure of all path from peering surface to any ASBR in AS (scale explosion)

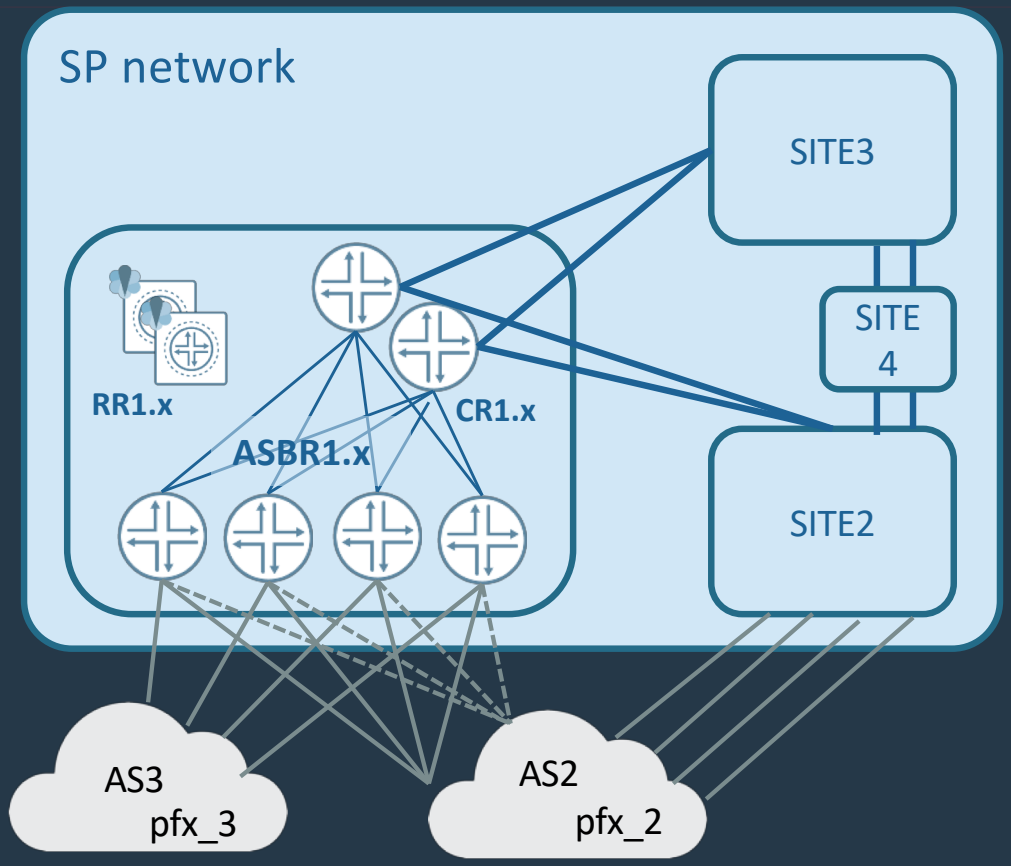
Can be also Scale-out – a lot of eBGP sessions.



## SOLUTION GOALS

1. Reduce number of BGP path across network
2. Keep ECMP
3. Enable sub-second restoration after failures for 500k+ impacted active BGP paths.
   - Interface
   - Sessions
   - ASBR
   - Site

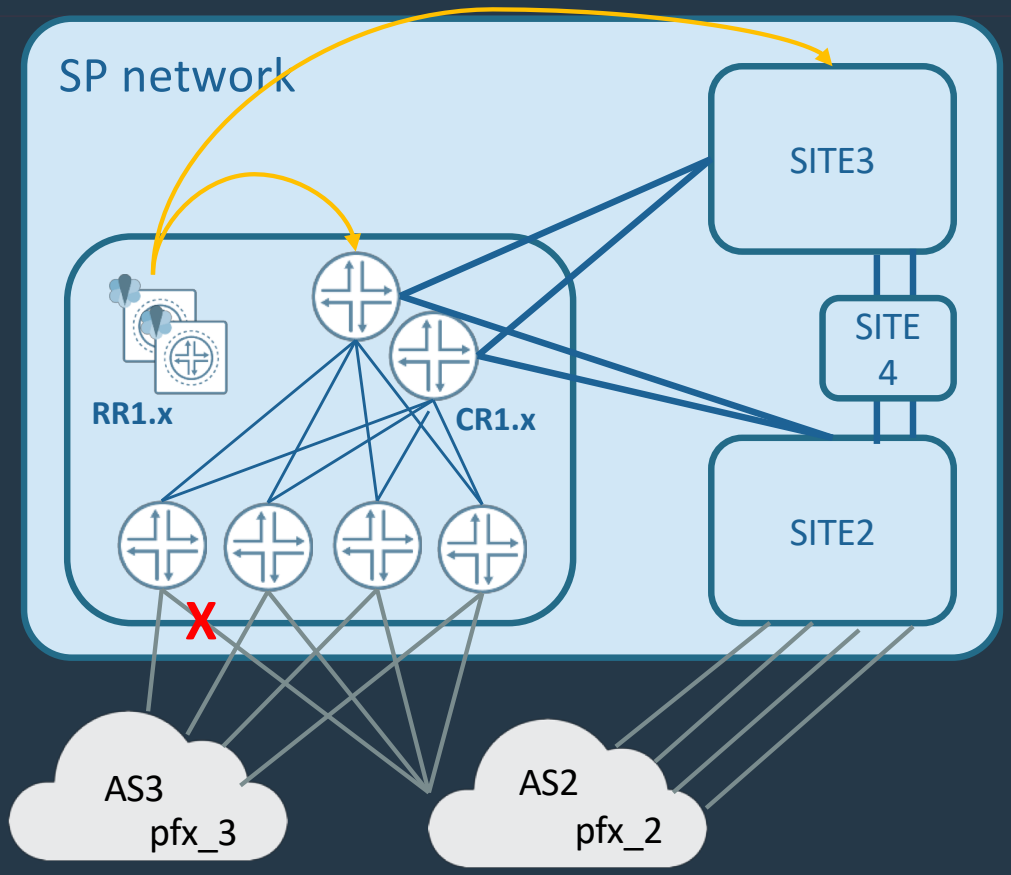
# EXAMPLE NETWORK




Native IP network.
BGP:

- RRx.1 full mesh; RRx.2 full mesh
  - one active path advertised.
- CR: multipath, ADD-PATH from RR's
- ASBR:
  - multipath, ADD-PATH from RR's
  - only active path to RR's
  - multipath from eBGP (LB)

# SHORTAGE OF NEXT HOP SELF



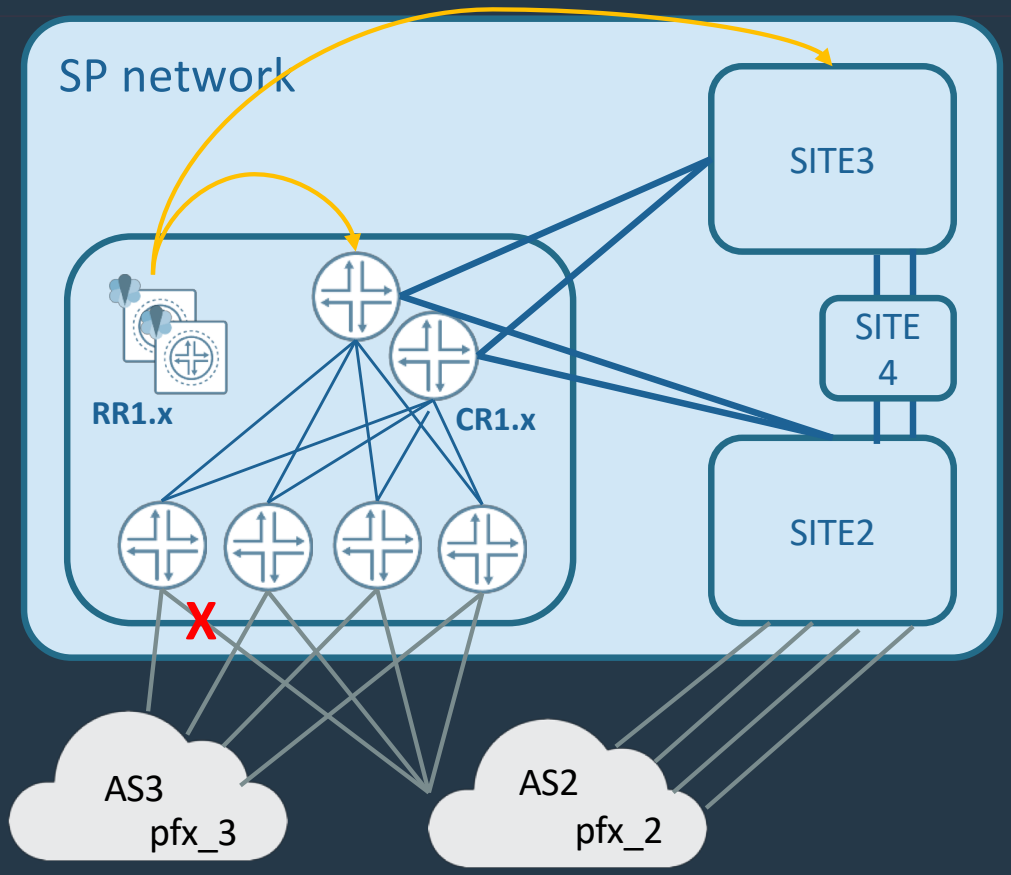

BEFORE:

- CR route to pfx_2 w/ 4 BGP Next-Hops (B_NH) ASBR1.1-ASBR1.4 loopbacks)
- all 4 loopbacks in IGP

AFTER:

- all 4 loopbacks in IGP
- ASBR1.1 send (500k+) withdraws (minutes?)
- RR reflects
- CR removes path and update FIB
- Routes at other sites need to do the same.

# SHORTAGES OF NEXT HOP UNCHANGED (1)




BEFORE:

- CR route to pfx_2 w/ 4 B_NH (one Peer_IP per ASBR)
- all 4 NNI subnets in IGP

FAILURE:

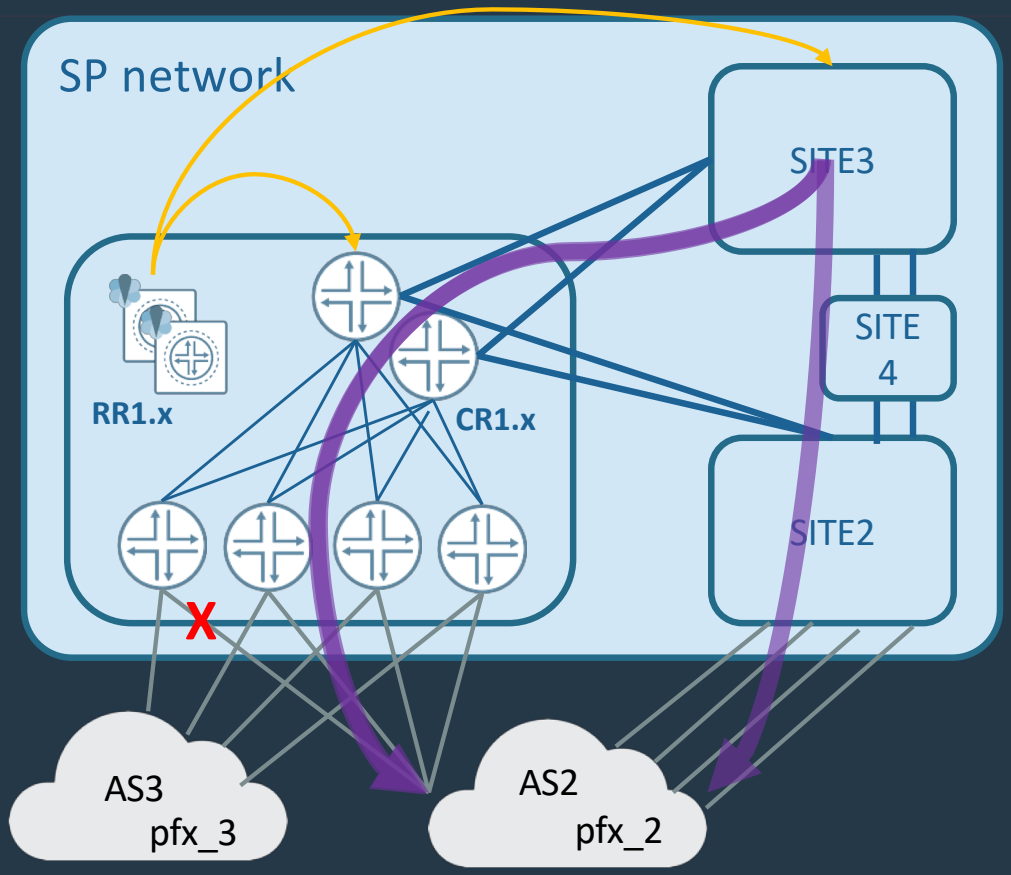
- BGP session down, Interface UP

AFTER:

- all 4 Peer_IPs in IGP
- ASBR1.1 send (500k+) withdraws (minutes?)
- …

Overkill ?

# SHORTAGES OF NEXT HOP UNCHANGED (2)




Interfaced goes down…

- NNI subnet removed from IGP. Peer_IP unreachable

CR

RR

ADD-PATH (4) among sites will solve this problem but:

- 4 x scale-up control plane
- 4 x scale-up in Next-Hop table (ASIC constrain)
- “cold-start” convergence time increase
- Churn exposure.

Other sites

- original B_NH unreachable
- switch to alternate ASBR (SITE2)
- then switch back to SITE1 after update

Long-Haul link usage - congestion; RTT;

# AGENDA

Scale-out peering:

- Why
- What it is and what it is not
- Principles of operation

Control-Plane challenges

- BGP Scale
- Common practice limitations

Solution

- Abstract Next-Hop (ANH)
- iBGP architecture w/ ANH
- Configuration examples

# ABSTRACT NEXT-HOP (ANH)

No protocol changes. Same old good BGP!

Arbitrary IP/32 address:

- set as B_NH when path form (member of) sub-set of eBGP advertised to iBGP.
- CONDITIONALY inserted into IGP
  - When at least one eBGP session form sub-set is ESTABLISHED/Converged

Sub-set of eBGP sessions – configuration, up to operator's decision: E.g.

- all sessions on given ASBR with same peer AS
- all sessions on given ASBR with same Transit providers
- all sessions on given SITE with same peer AS

Generic concept, not only for scale-out peering.

# ASBR OPERATION W/ ANH – INTRA-SITE

## Setup

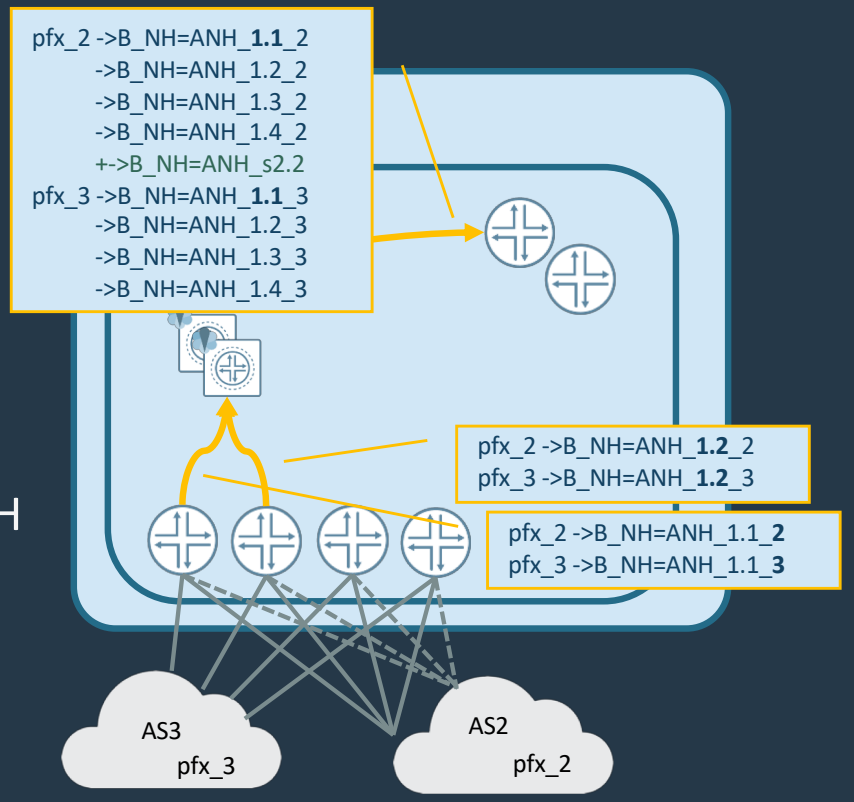
eBGP sessions sub-set - all session w/ AS2 from ASBR.

ASBR-PeerAS-ANH (AP-ANH) - unique per ASBR, per PeerAS – (ANH_1.1_2)

RR1.x

- gets **1** path from each ASBR w/ B_NH==AP-ANH (regardless of # eBGP session w/ AS2)
- advertise ADD-PATH (5) to CR1.x and ASBR

CR1.x load-balance among 4 B_NHs



# ASBR OPERATION W/ ANH – INTRA-SITE

## Failures

### One session with AS2

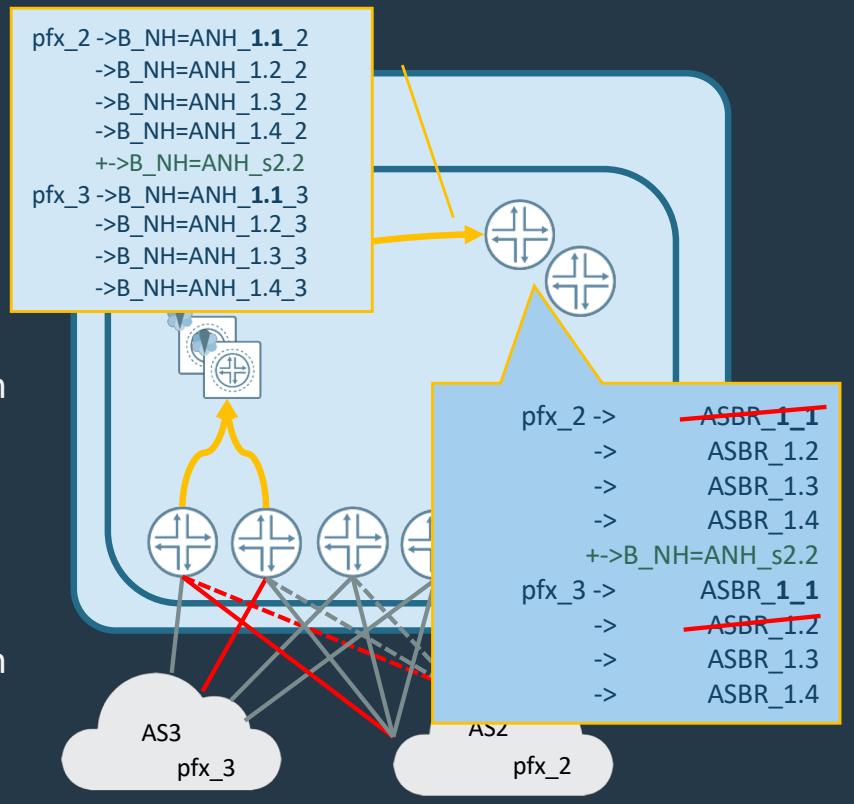
- B_NH on iBGP not changed. Other attributes unchanged.
- No Update send to RR.
- like NHS

### One session with AS3 (@ ASBR1.2)

- B_NH unreachable in IGP. Path invalid. CR remove path form ECMP group. CR sent to un-affected ASBR only.
- ASBR1.1 (slowly) withdraws paths
- like NH unchanged (peer IP)

### All session with AS2 (@ ASBR1.1)

- B_NH unreachable in IGP. Path invalid. CR remove path form ECMP group. CR sent to un-affected ASBR only.
- ASBR1.1 (slowly) withdraws paths
- like nothing else

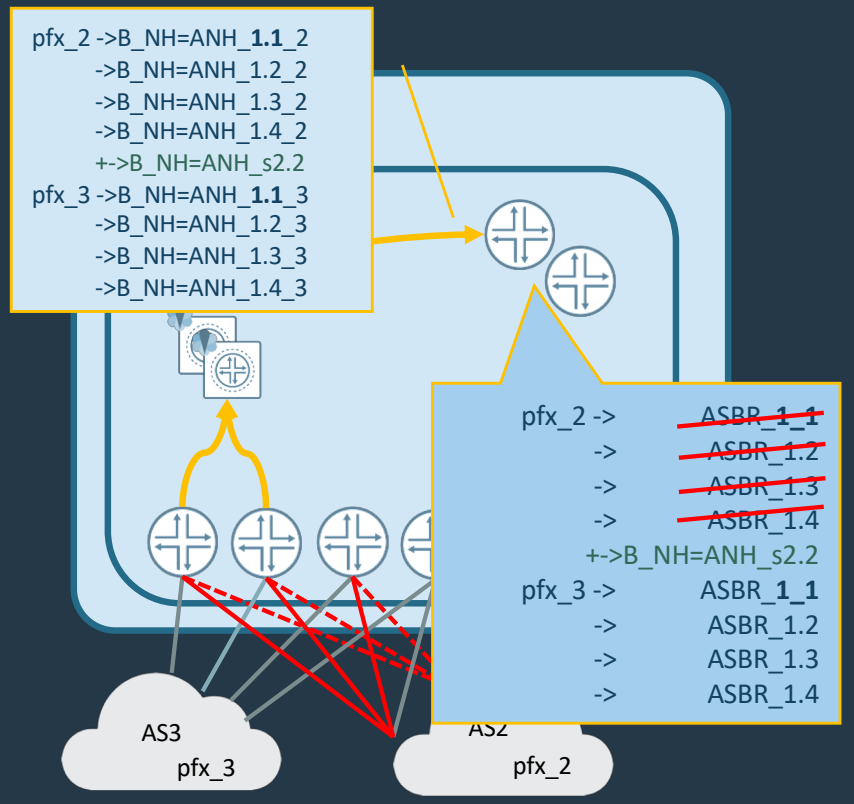


# ASBR OPERATION W/ ANH – INTRA-SITE

## Failures

All session with AS2 on all ASBR_1.x

- B_NHs unreachable in IGP.
  - Path invalid. CR remove  4 path form ECMP group.
  - CR sent traffic to other site.
- ASBR1.x (slowly) withdraws paths
- like nothing else



# ASBR OPERATION W/ ANH – AS-WIDE

## Setup

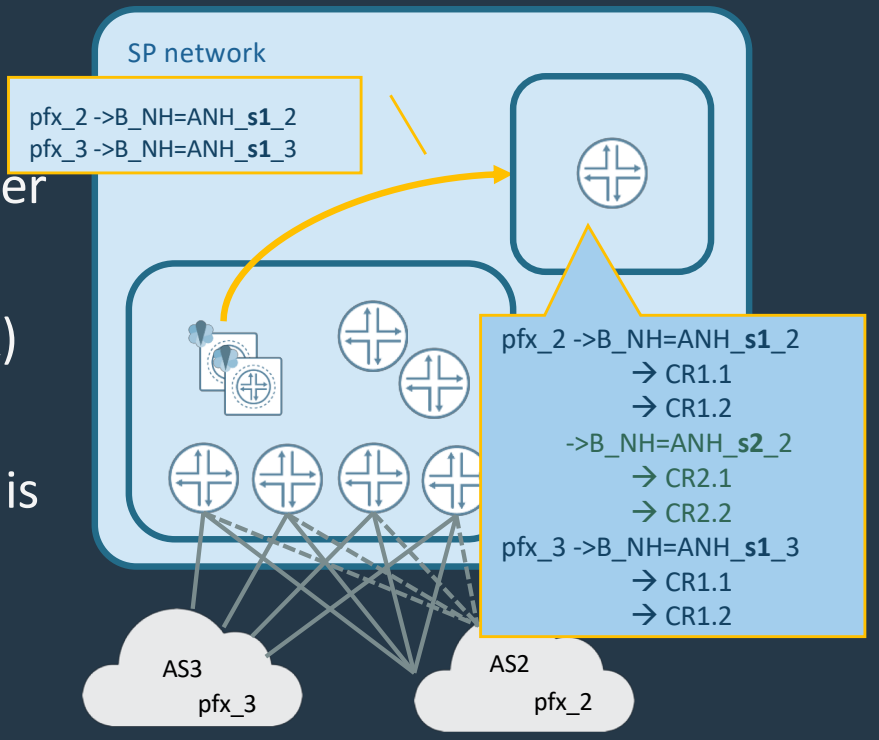
eBGP sessions set - all session w/ AS2 from all ASBR at site.

Site-PeerAS-ANH (SP-ANH) - unique per SITE, per PeerAS – (ANH_s1_2)

RR1.x - advertise one path to other sites (RRy.x) w/ B_NH:=SP-ANH

ASBR1.x – insert SP-ANH into IGP if its AP-ANH is active.

CR/CS at other sites – resolves SP-ANH via IGP-load balance among all CR1.x/CRy.x



# ASBR OPERATION W/ ANH – AS-WIDE

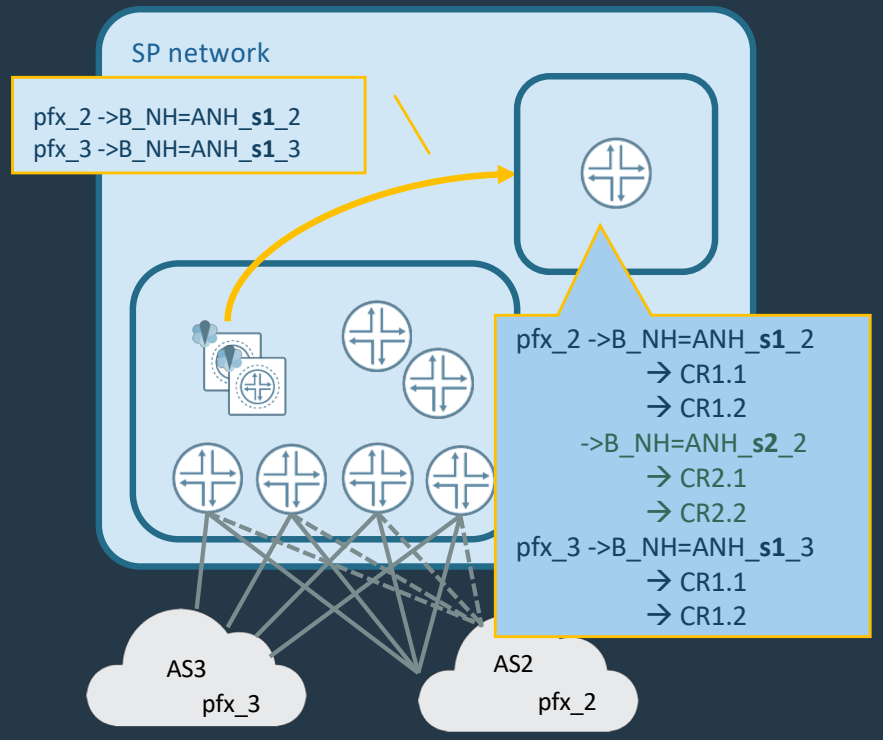
## Failures

### One session with AS2

- No Update/Withdraw send to among RR.
- like NHS

### One session with AS3 (@ ASBR1.2)

- SP-ANH reachable in IGP.
- No FIB changes on other sites (same B_NH).
- ORIGINATOR ID changed - Update send to among RR.



# ASBR OPERATION W/ ANH – AS-WIDE

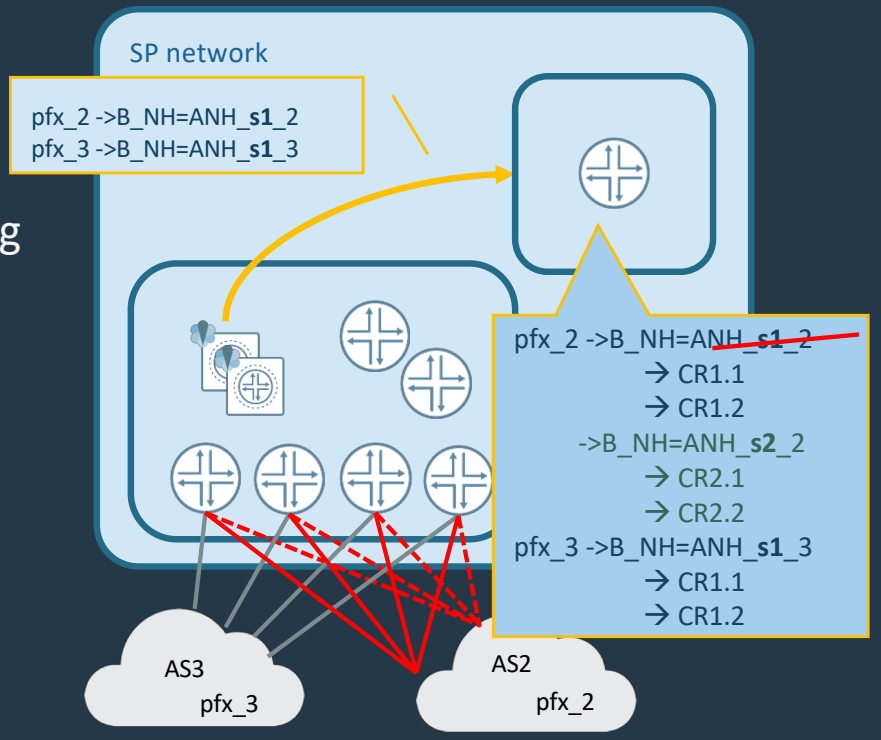
## Failures

All session with AS2 (@ ASBR1.1)

- SP-ANH reachable in IGP.
- No FIB changes on other sites (same B_NH).
- ORIGINATOR ID changed - Update send to among RR.

All session with AS2 on all ASBR_1.x

- B_NHs unreachable in IGP. Other site
  - Path invalid. CR/CS remove 4 path form ECMP group.
  - At other site CR/CS sent traffic to other site.
- ASBR1.x (slowly) withdraws paths
- like nothing else



# RESULTS – CONTROL PLANE STATE MITIGATION

## Network

- 40 sites
- 4 ASBR's per site
- 2 RR per site
- 500k external pfx per ASBR; 250k best.
- 1M prefixes

Peering surface – 80M path, 40M active.

## Control Plane Scale

- RR : 11M path
  - 40 x 250k = 10M path from other sites
  - 4 x 250k =1M path from local ASBRs
- CR : 2.5-10M path
  - 2 x (4-5) x 250k = 2-2.5M path (best on local ASBR)
  - 2 x (1-5*) x 750k = 1.5-7.5M path
- ASBR : 3-10M path
  - 2 x (3-5) x 250k = 1.5-2.5M path (best on local ASBR)
  - 2 x (1-5*) x 750k = 1.5-7.5M path

upto **80% less**

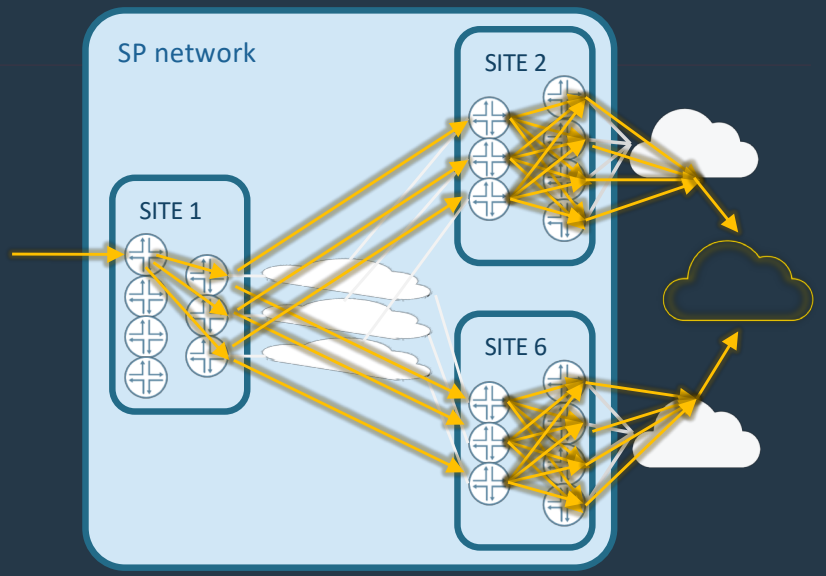
* if given prefix, when learned from eBGP is best at 5 remote sites

# RESULTS - ECMP

Network

- 40 sites
- 4 ASBR's per site
- 2 RR per site
- 3 CR per site (3-ple plane core)

let SITES 2-6 are in same IGP distance form SITE 1.




ECMP – 24 way

- across up-to 5 remote sites, each reachable over 3 core planes, and then over 4 ASBRs there.
- If best exit on local ASBR – 4 way.

# CONFIGURATION EXAMPLE – AP-ANH (JUNOS)

```
[edit protocols bgp group PeerAS2]
type external;
egress-te {
    install-address 1.1.1.2;
    rib {
        inet.0;
    }
}
peer-as 2;
neighbor 11.1.1.1;
neighbor 11.1.1.5;

[protocols isis export]
export AP-ANH-TO-ISIS;

[policy-options policy-statement AP-NH-TO-ISIS term IPv4]
from {
    protocol arp;
    rib inet.0;
}
    then accept;

[protocols bgp group to-RR]
export [ ...  SET-AP-ANH ]
neighbor 11.0.0.6

[policy-options policy-statement SET-AP-ANH term 10]
from {
    as-path fromAS2;
}
then {
    next-hop 1.1.1.2;
}
```

ANH route is UP if at leas one session @ ASBR is converged.
(ESTABLISHED state and received EoR marker)

Simple IGP export policy

```
ASBR1.1> show route 1.1.1.2 terse
A V Destination        P Prf    Metric 1   Metric 2  Next hop
AS path
* ? 1.1.1.2/32         A 170                         >11.1.1.1
  ?                    A 170                         >11.1.1.5

ASBR1.1> show isis database ASBR1.1.00-00 detail

IS-IS level 2 link-state database:

ASBR1.1.00-00 Sequence: 0x894, Checksum: 0x8b43, Lifetime: 706 secs
   IS neighbor: CR1.1.00                       Metric:       45
   IS neighbor: CR1.2.00                       Metric:       50
   IP prefix:  1.1.1.2/32                      Metric:       10
Internal Up

ASBR1.1> show route advertising-protocol bgp 11.0.0.6 150.0.0.0

inet.0: 2103071 destinations, 5958606 routes (2103070 active, 5
holddown, 1 hidden)
  Prefix                 Nexthop          MED     Lclpref    AS path
* 150.0.0.0/24           1.1.1.2          90      199        2 1234 I
```

# CONFIGURATION EXAMPLE – AP-ANH (IOS-XR)

```
router static
    address-family ipv4 unicast
        1.0.1.2 255.255.255.255 GigabitEthernet0/0/0/1 11.1.1.1
tag 2
        1.0.1.2 255.255.255.255 GigabitEthernet0/0/1/3 11.1.1.5
tag 2
    !
!
route-policy SP-NH-TO-ISIS
    if tag is 2 and rib-has-route in
    (57.57.1.1/32, 57.57.1.5/32) then pass
    endif
  end-policy


router isis 1
   address-family ipv4 unicast
      redistribute static SP-NH-TO-ISIS
     !
  !

router bgp 1
  neighbor 192.168.40.24
    remote-as 1
    address-family ipv4 unicast
    route-policy SET-AP-ANH out
  !
!
```

```
route-policy SET-AP-ANH
    if as-path neighbor-is '2' then
      set next-hop 1.0.1.2
    endif
    […]
  end-policy
```

ANH route is UP if at least one peer is UP.
(but all session could be down…)

Conditional IGP export policy – “if BGP session UP”
Known, super-stable ”leading” external prefix needed.

”leading” external prefix:

- Known, agreed upfront
- Advertised by peer on given session only
- Always advertised (stable)
- Exists in RIB == given session is ESTABLISHED

# CONFIGURATION EXAMPLE – SP-ANH (JUNOS)

```
@ASBR:

[edit routing options static route]
1.0.1.2/32 {
    next-hop discard;
    tag 2;
}
[protocols isis export]
export SP-ANH-TO-ISIS;
[policy-options policy-statement SP-NH-TO-ISIS term IPv4]
from {
    condition AP-ANH-AS2
    protocol static;
    tag 2;
}
    then accept;
[policy-options condition AP-ANH-1.1-2]
if-route-exists{
    1.1.1.2/32;
    table inet.0;
}

@RR:
[protocols bgp group RR-mesh]
export [ ...  SET-SP-ANH ]
neighbor 11.0.0.17
[policy-options policy-statement SET-SP-ANH term 10]
from {
    as-path fromAS2;
}
then {
    next-hop 1.0.1.2;
}
```

SP-ANH route is in IGP if at least one session @ SITE is converged.
(so at least one AP-ANH is in RIB)

```
ASBR1.1> show route 1.1.1.2 terse
S V Destination        P Prf   Metric 1   Metric 2  Next hop        AS path
* ? 1.0.1.2/32         A 170                        > dsc


ASBR1.1> show isis database ASBR1.1.00-00 detail

IS-IS level 2 link-state database:

ASBR1.1.00-00 Sequence: 0x894, Checksum: 0x8b43, Lifetime: 706 secs
   IS neighbor: CR1.1.00                       Metric:       45
   IS neighbor: CR1.2.00                       Metric:       50
   IP prefix:  1.1.1.2/32                      Metric:       10 Internal Up
   IP prefix:  1.0.1.2/32                      Metric:       10 Internal Up

RR1.1> show route advertising-protocol bgp 11.0.0.17 150.0.0.0

inet.0: 2103071 destinations, 5958606 routes (2103070 active, 5 holddown, 1
hidden)
  Prefix                  Nexthop           MED     Lclpref    AS path
* 150.0.0.0/24            1.0.1.2           90      199        2 1234 I
```

# SUMMARY

Scale-out solves challenges with Bandwidth, redundancy, RTT, etc.

At cost of scale-up Control Plane states.

New construct and practices in protocol configuration/network design needed.

- Abstract NH is useful construct.
- Use it to control BGP scale in scale-out peering is example of practices.
- draft-szarecki-grow-abstract-NH-scaleout-peering

# THANK YOU!