Virtual Aggregation (VA)

Paul Francis,	MPI-SWS
Xiaohu Xu,	Huawei,
Hitesh Ballani,	Cornell
Dan Jen,	UCLA
Robert Raszuk,	Cisco
Lixia Zhang,	UCLA

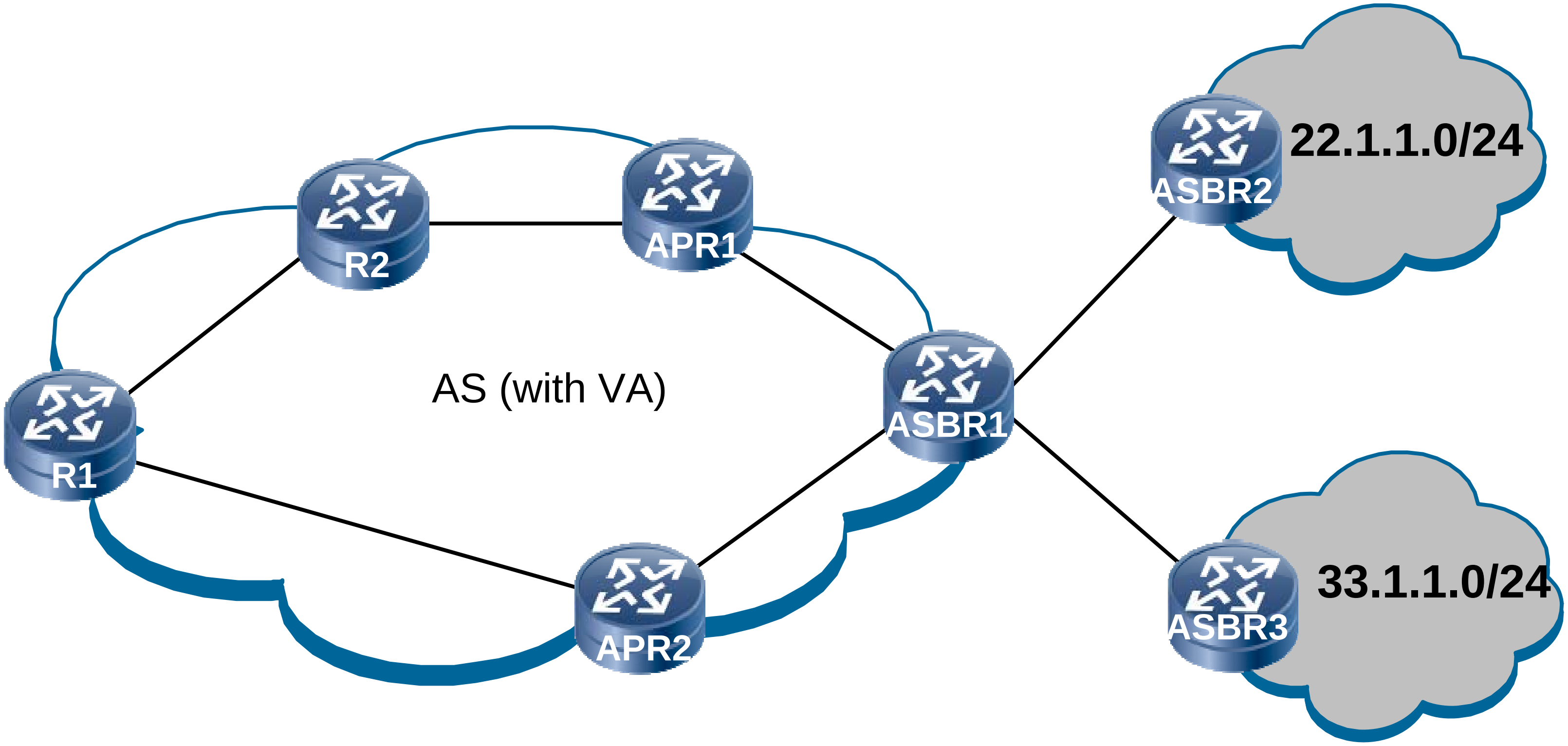
Current status

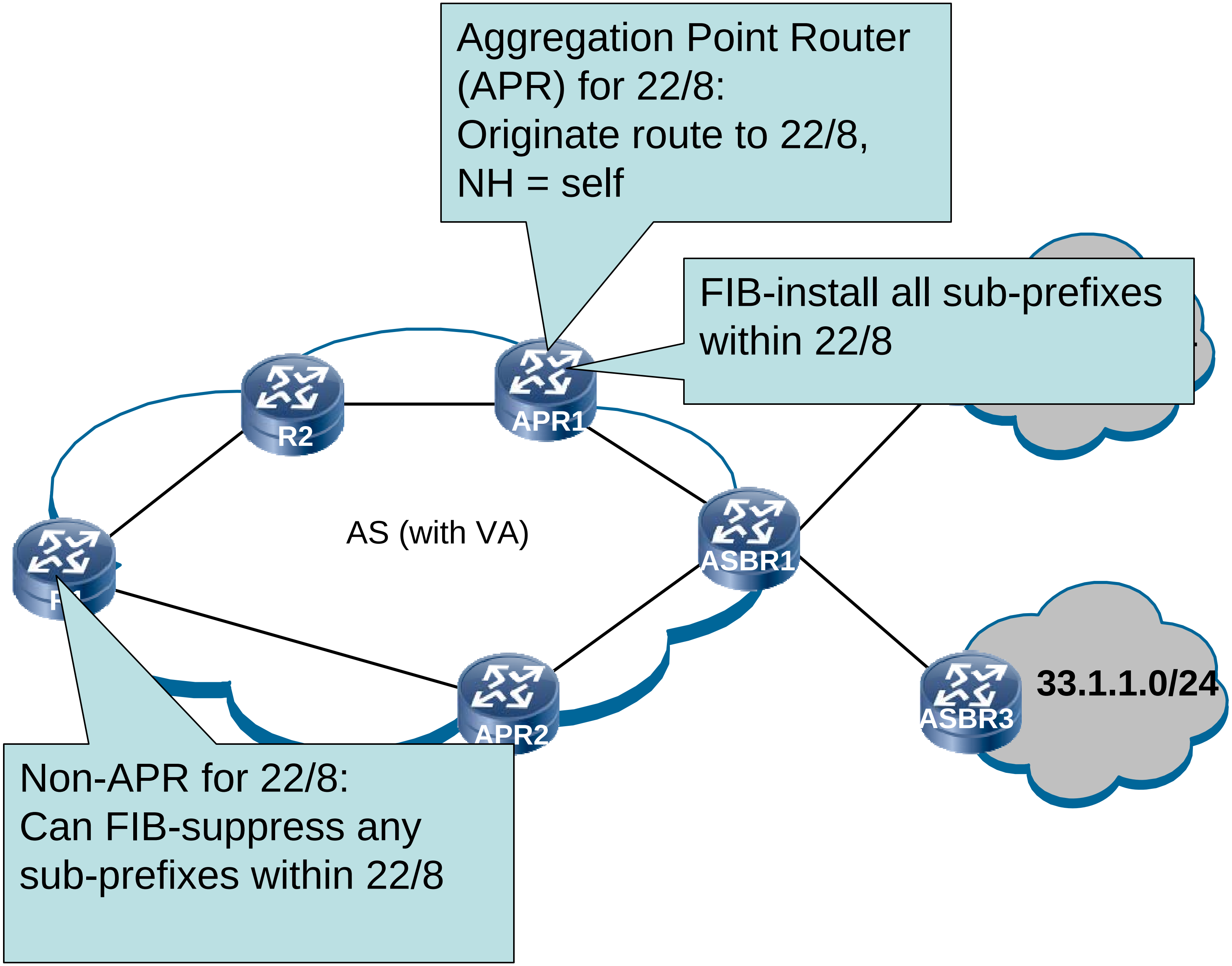
- New WG item in GROW
- Four informational drafts (six authors):
  - draft-ietf-grow-va-00
  - draft-ietf-grow-va-gre-00
  - draft-ietf-grow-va-mpls-00
  - draft-ietf-grow-va-perf-00
- Two partial implementations
  - Huawei, MPI-SWS (Quagga/Linux)

Virtual Aggregation (VA)

- A simple technique to shrink FIB size
  - Does not shrink RIB size
  - Can incur latency/load penalty, though this can be kept small
- Flexible: tight control over FIB size of *any* router in an ISP (core, edge, etc.)
  - Think up to 10X FIB reduction with negligible latency/load penalty

Basic VA mechanism

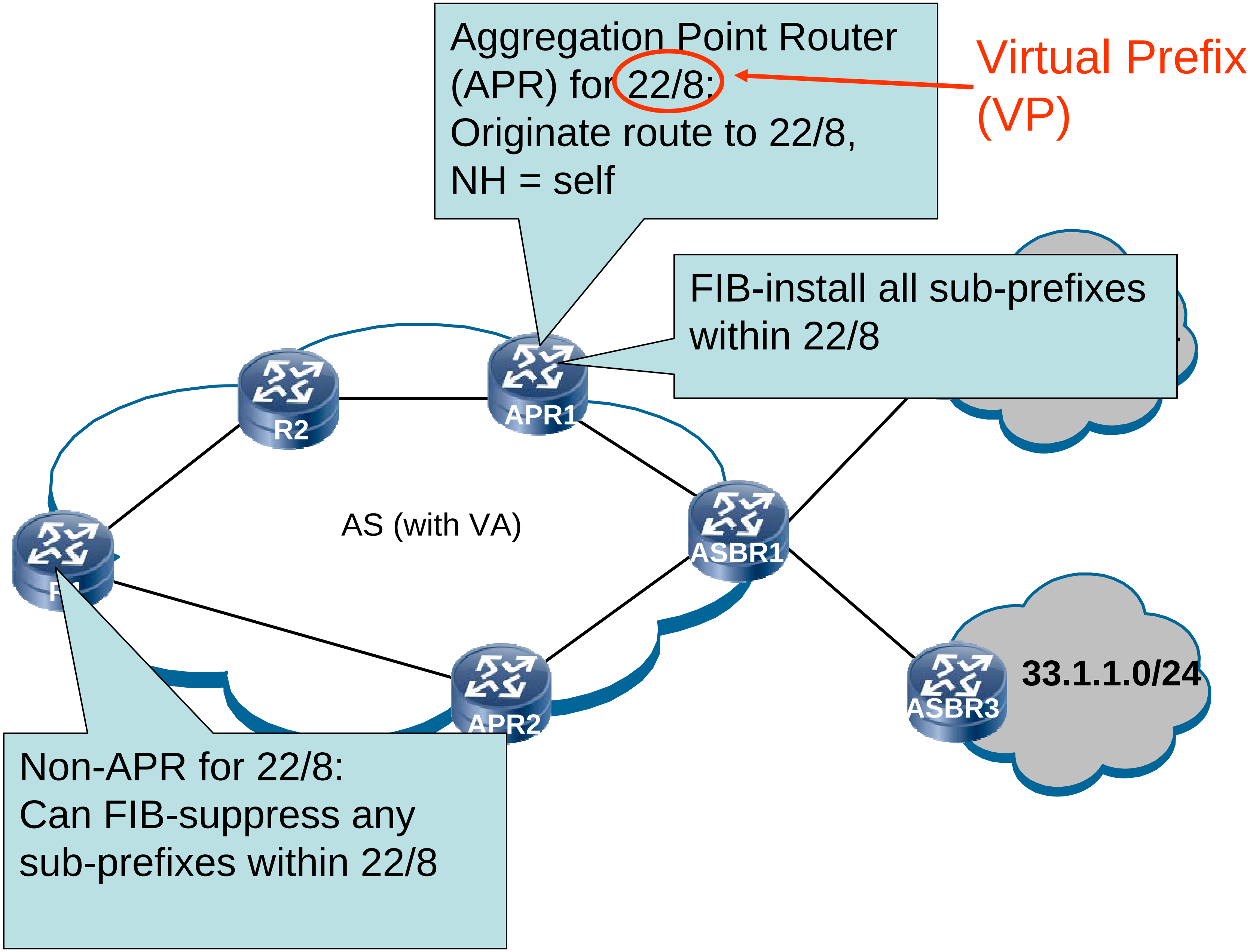




Aggregation Point Router (APR) for 22/8:
Originate route to 22/8,
NH = self

FIB-install all sub-prefixes within 22/8

Non-APR for 22/8:
Can FIB-suppress any sub-prefixes within 22/8



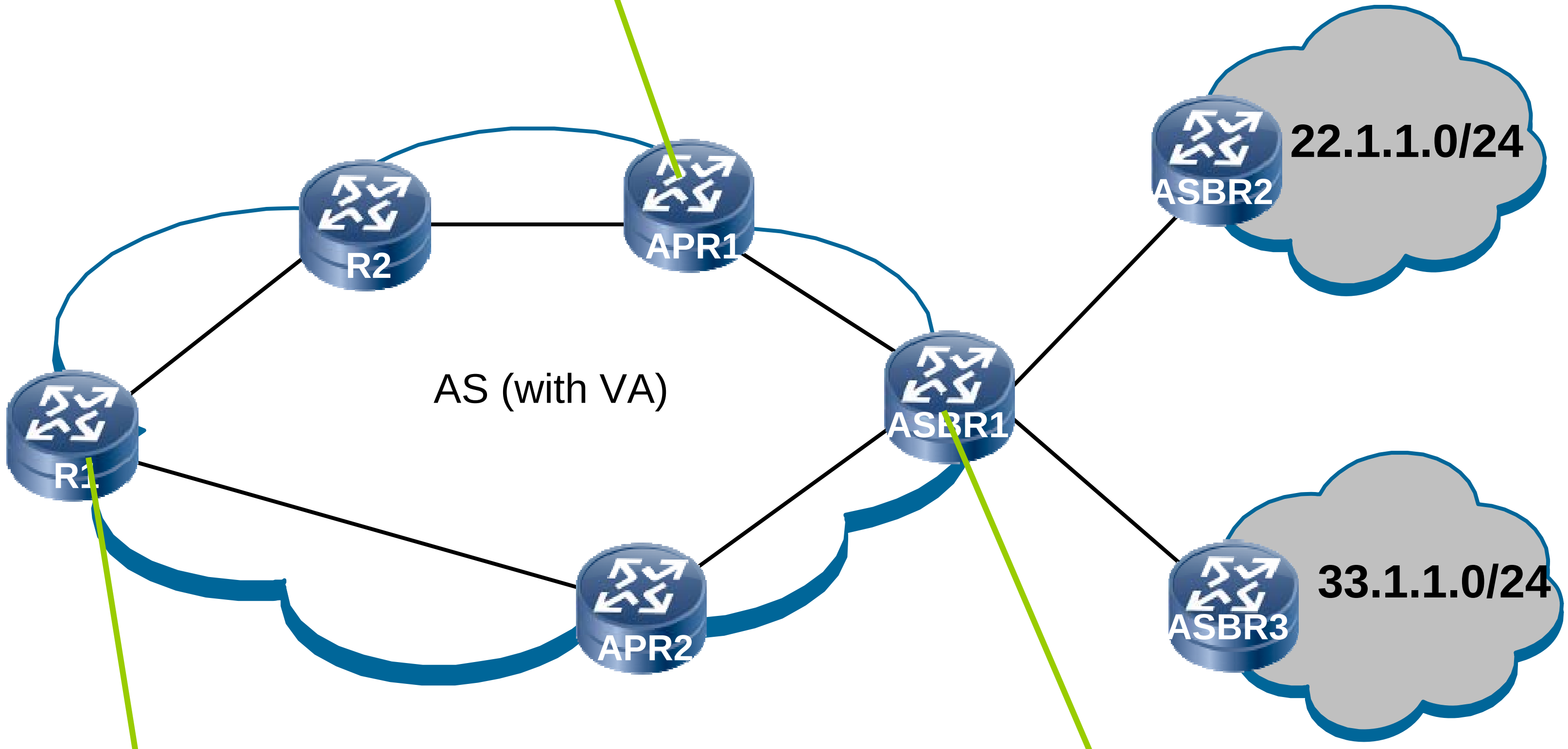
Aggregation Point Router (APR) for 22/8:
Originate route to 22/8,
NH = self

Virtual Prefix (VP)

FIB-install all sub-prefixes within 22/8

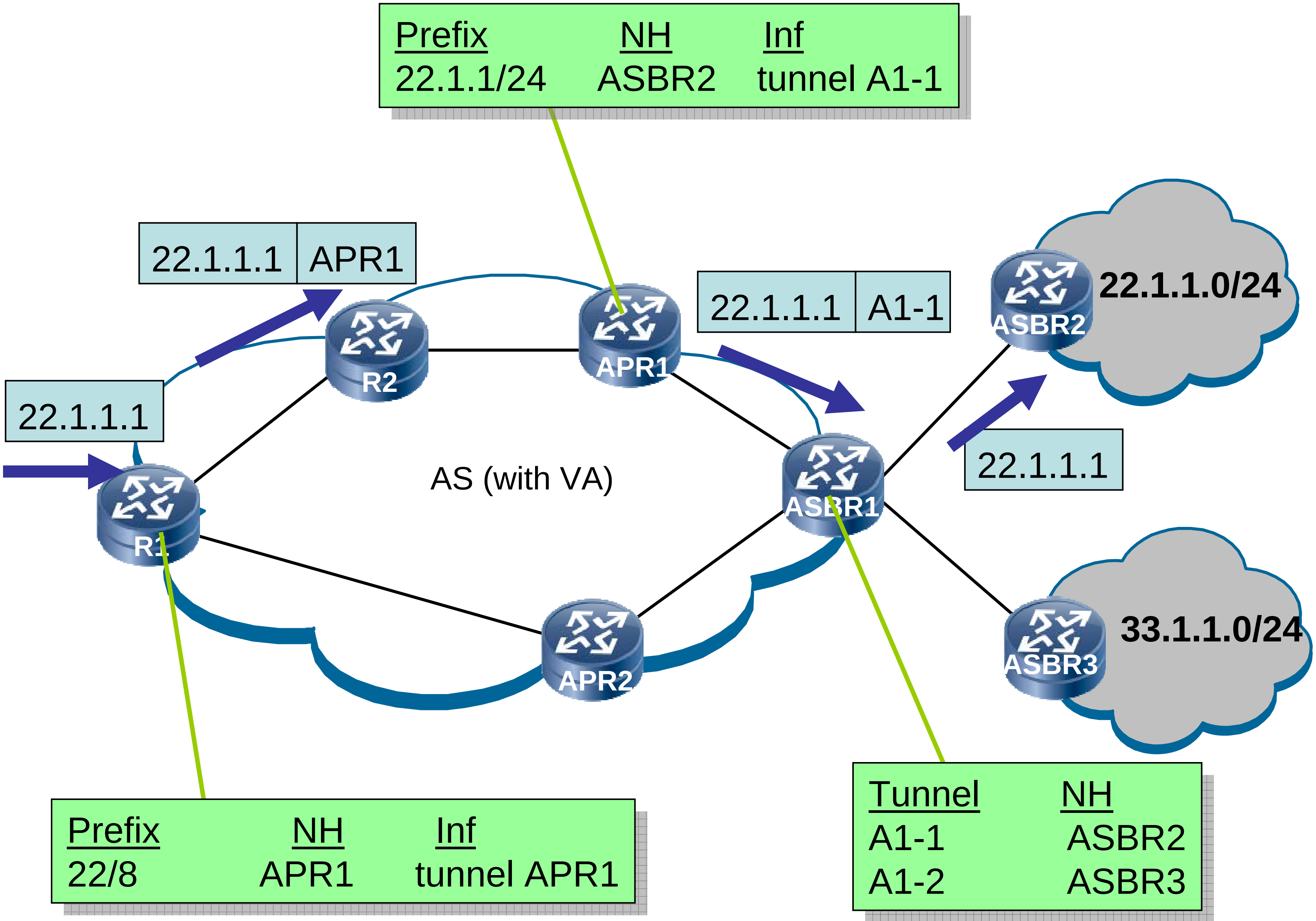
Non-APR for 22/8:
Can FIB-suppress any sub-prefixes within 22/8

Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1



Prefix	NH	Inf
22/8	APR1	tunnel APR1

Tunnel	NH
A1-1	ASBR2
A1-2	ASBR3

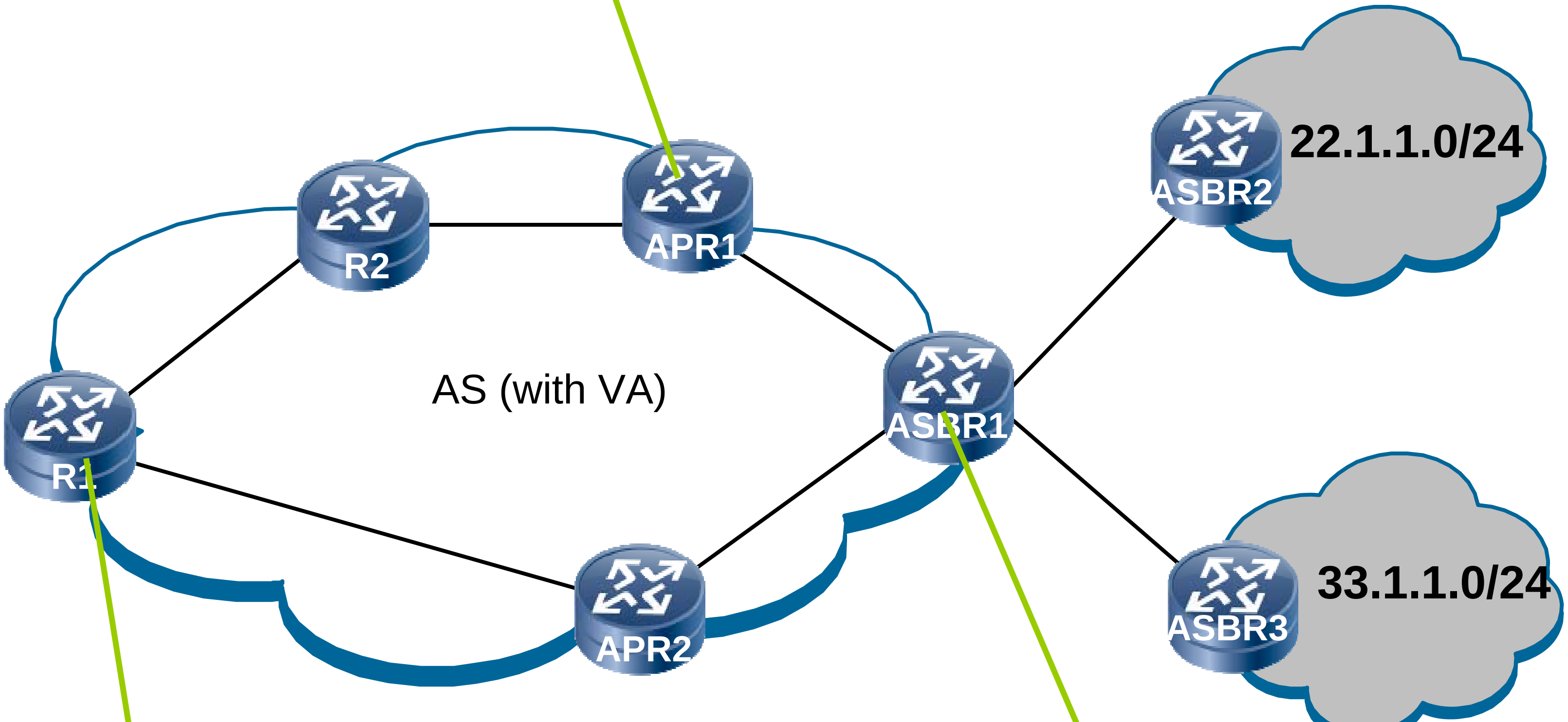


Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1

Prefix	NH	Inf
22/8	APR1	tunnel APR1

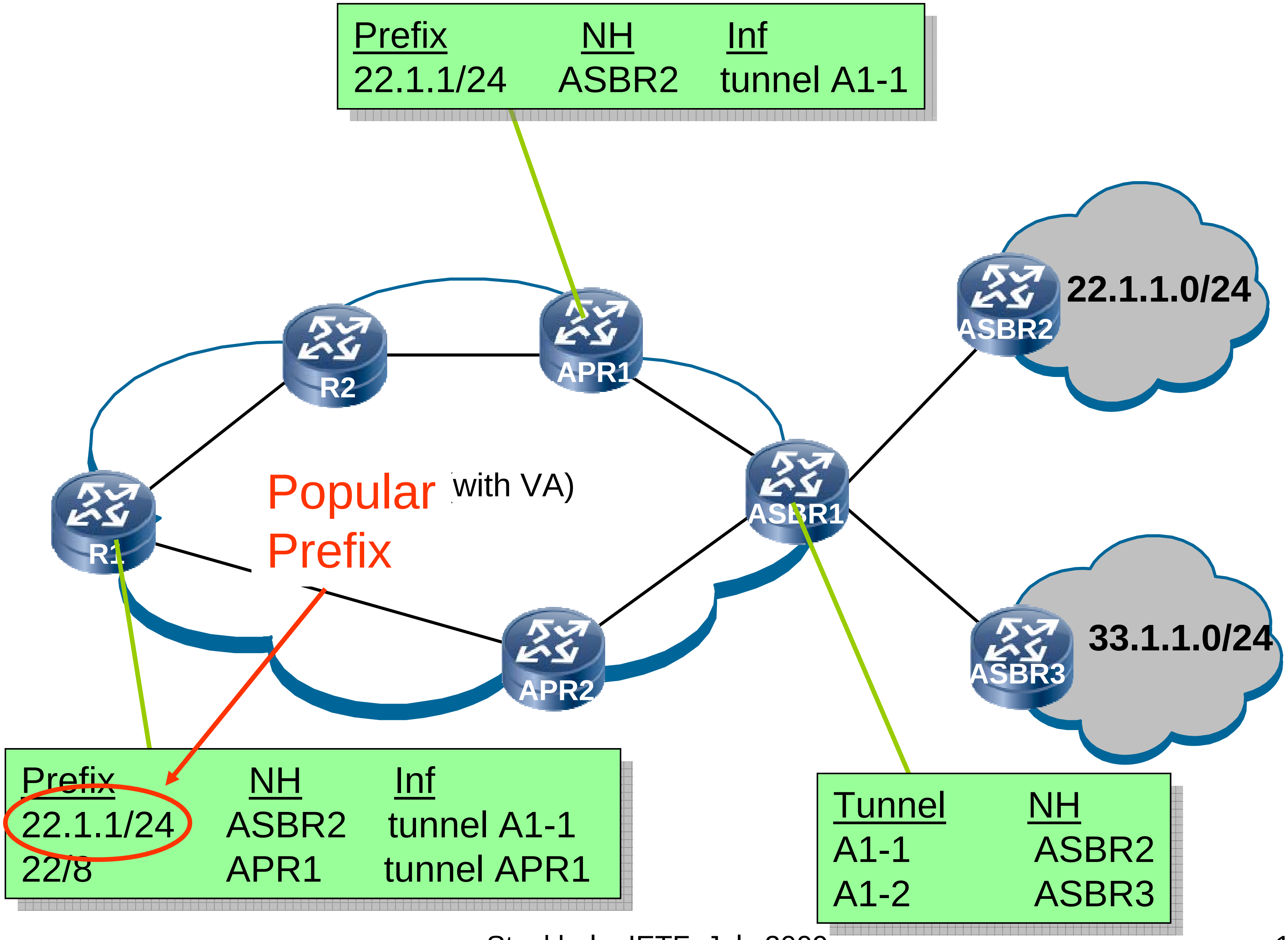
Tunnel	NH
A1-1	ASBR2
A1-2	ASBR3

Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1



Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1
22/8	APR1	tunnel APR1

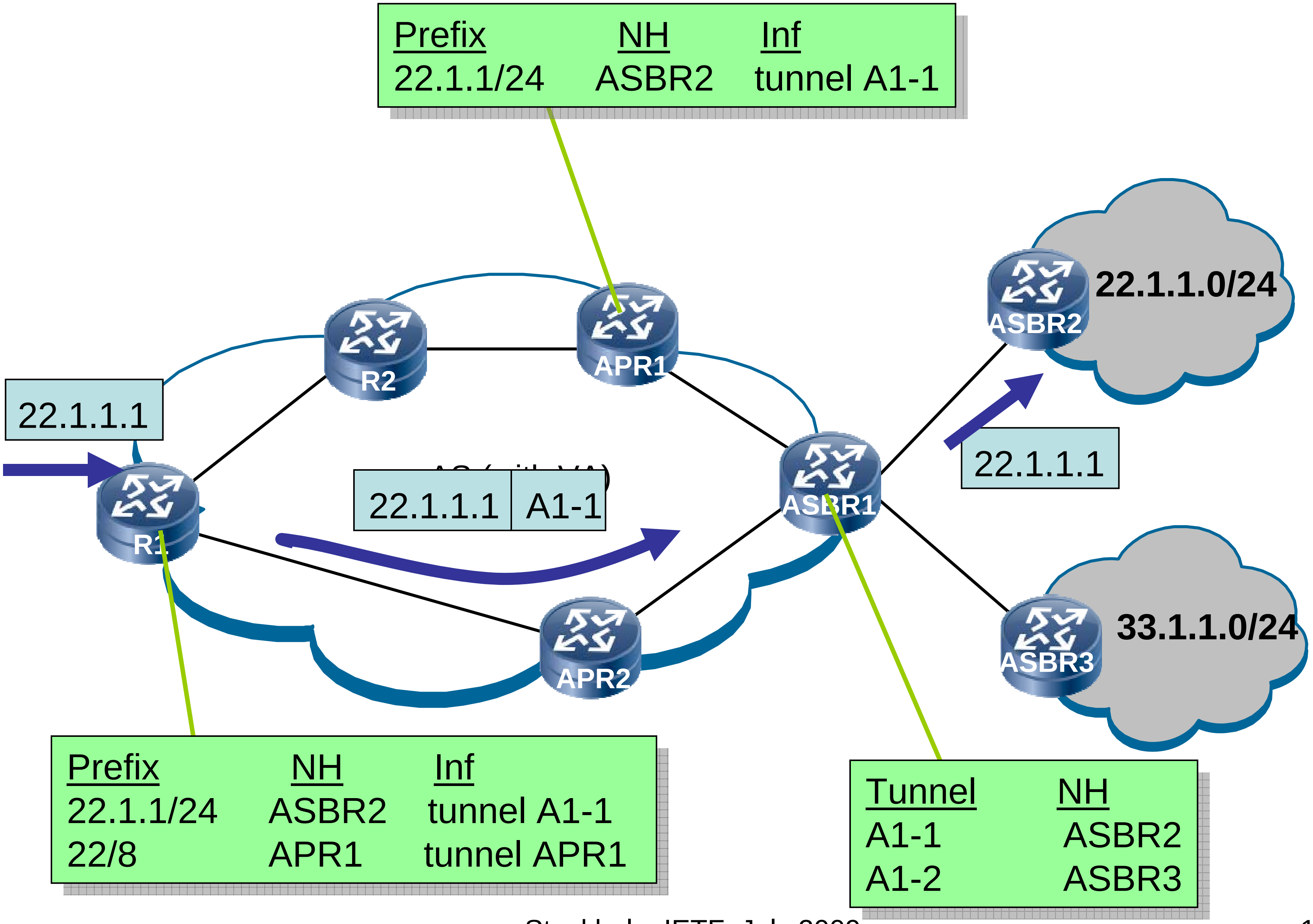
Tunnel	NH
A1-1	ASBR2
A1-2	ASBR3



Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1

Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1
22/8	APR1	tunnel APR1

Tunnel	NH
A1-1	ASBR2
A1-2	ASBR3

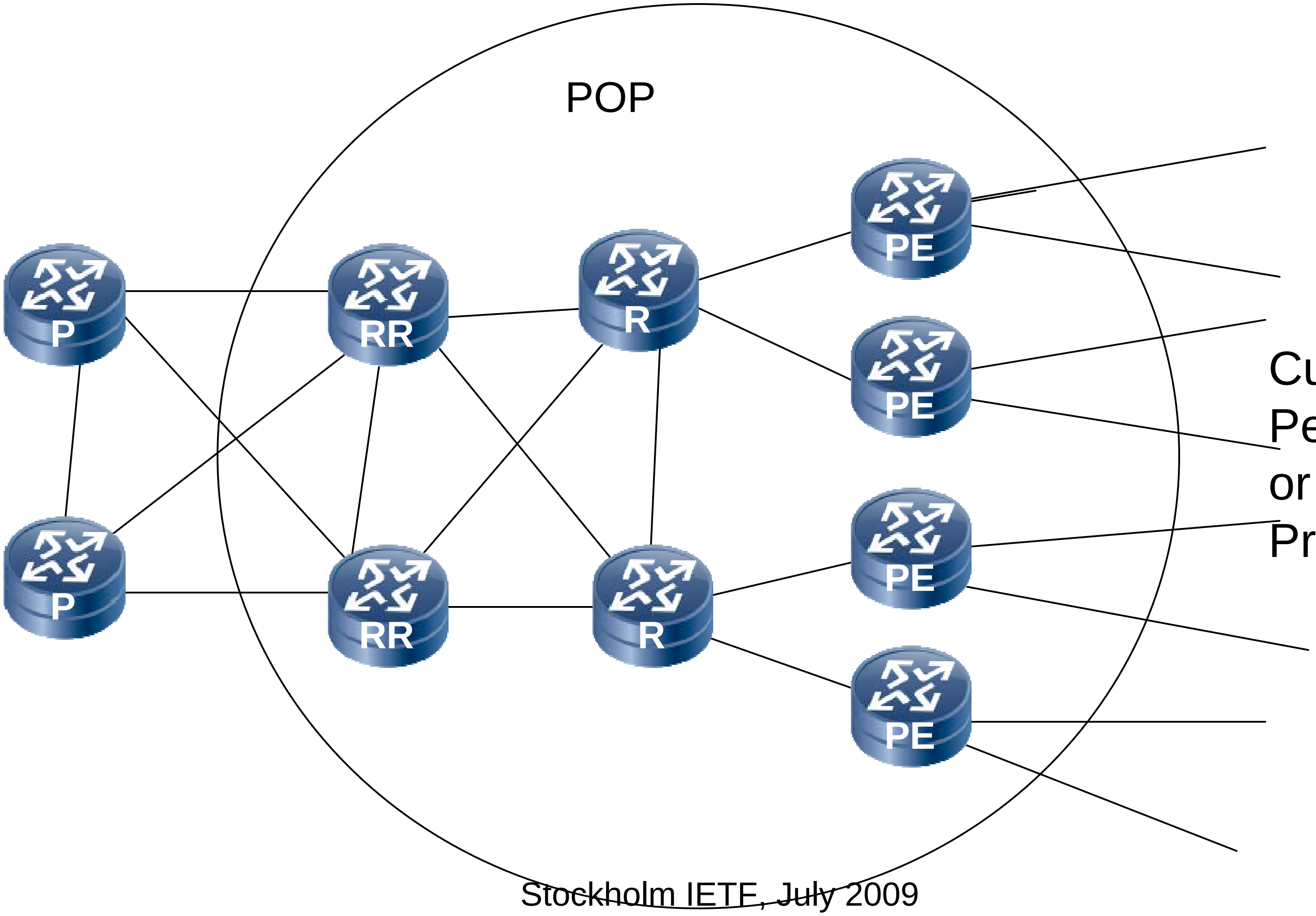


Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1

Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1
22/8	APR1	tunnel APR1

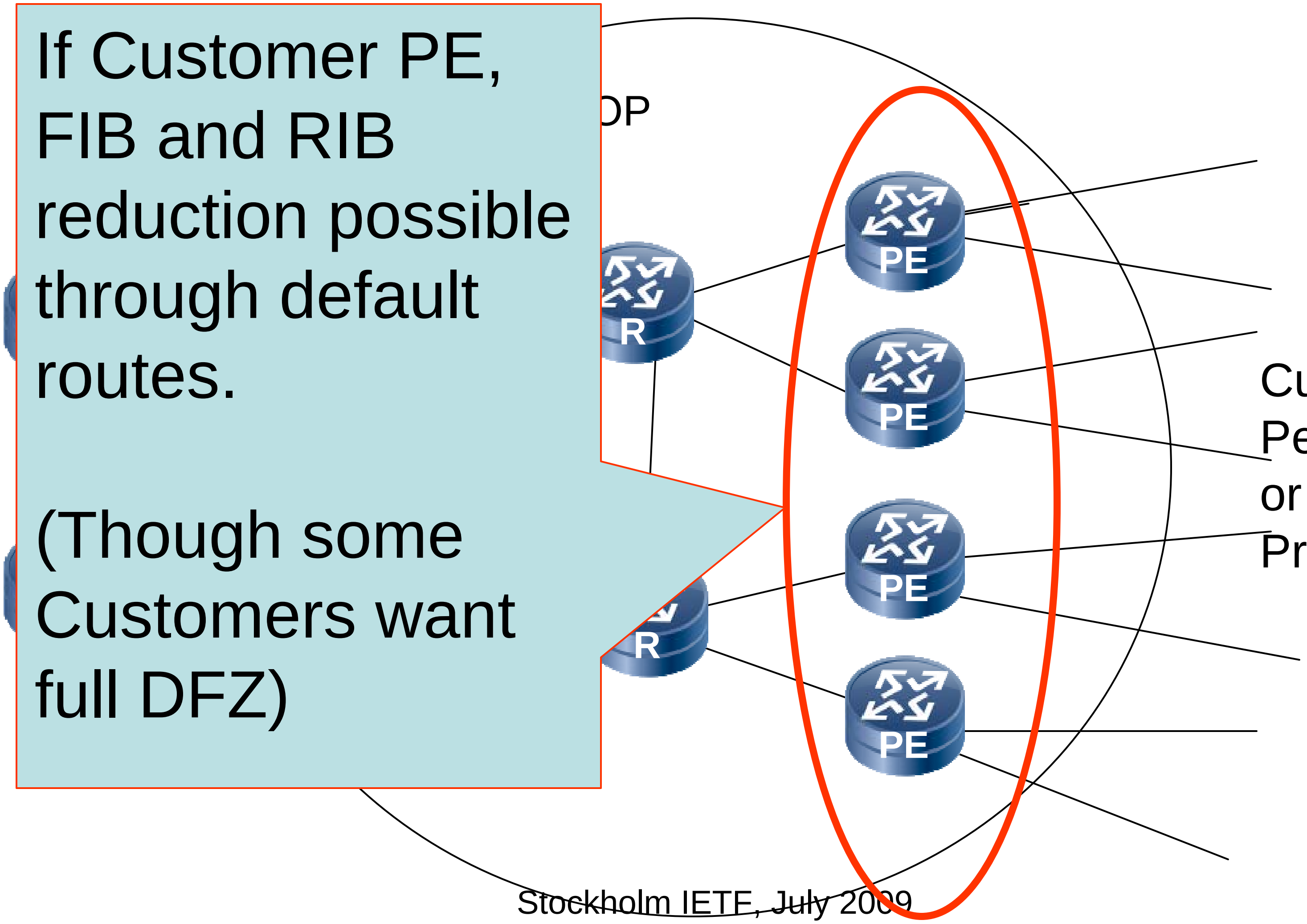
Tunnel	NH
A1-1	ASBR2
A1-2	ASBR3

A typical POP structure



Customers,
Peers,
or
Providers

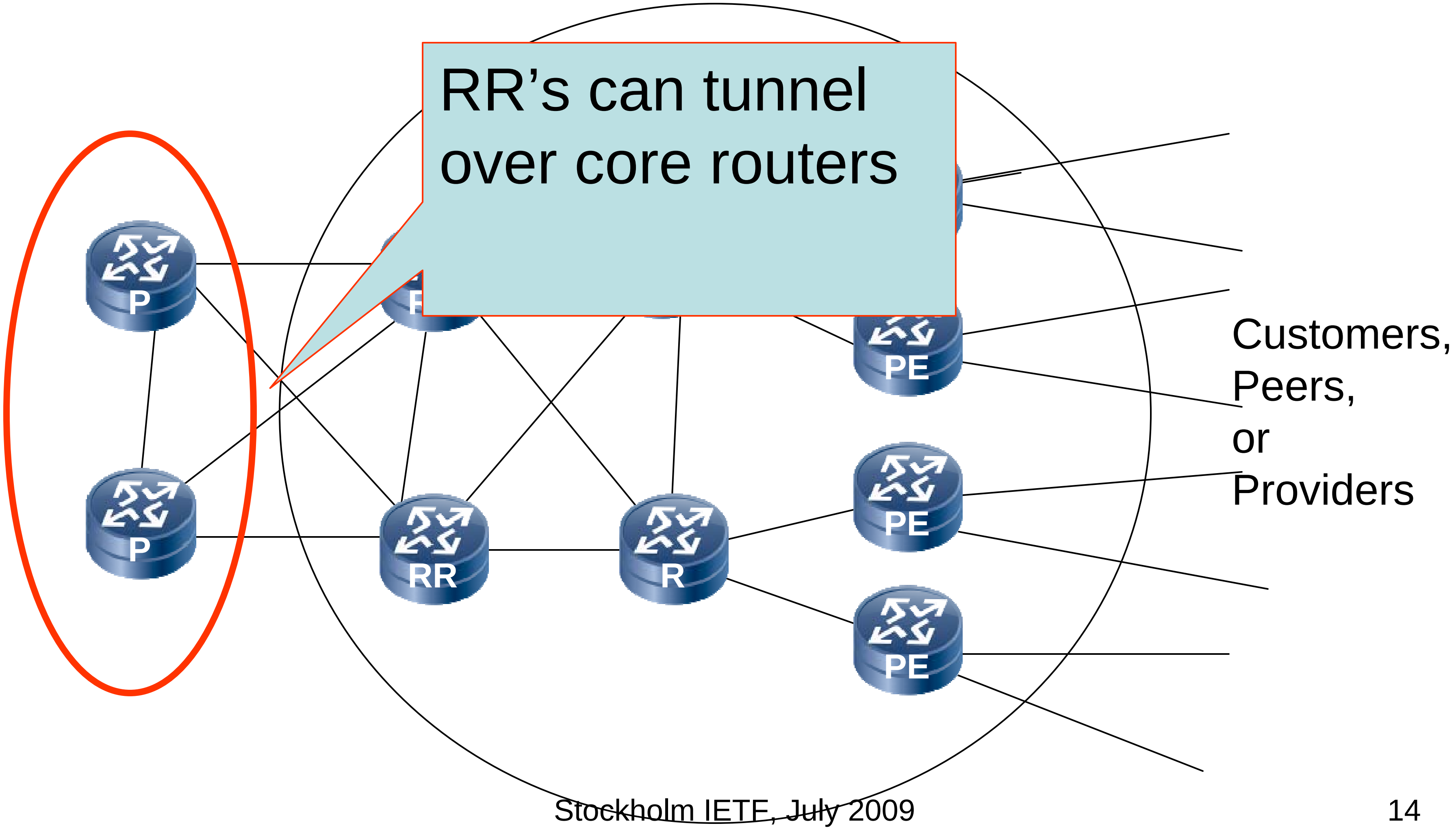
FIB reduction today



If Customer PE, FIB and RIB reduction possible through default routes.

(Though some Customers want full DFZ)

FIB reduction today

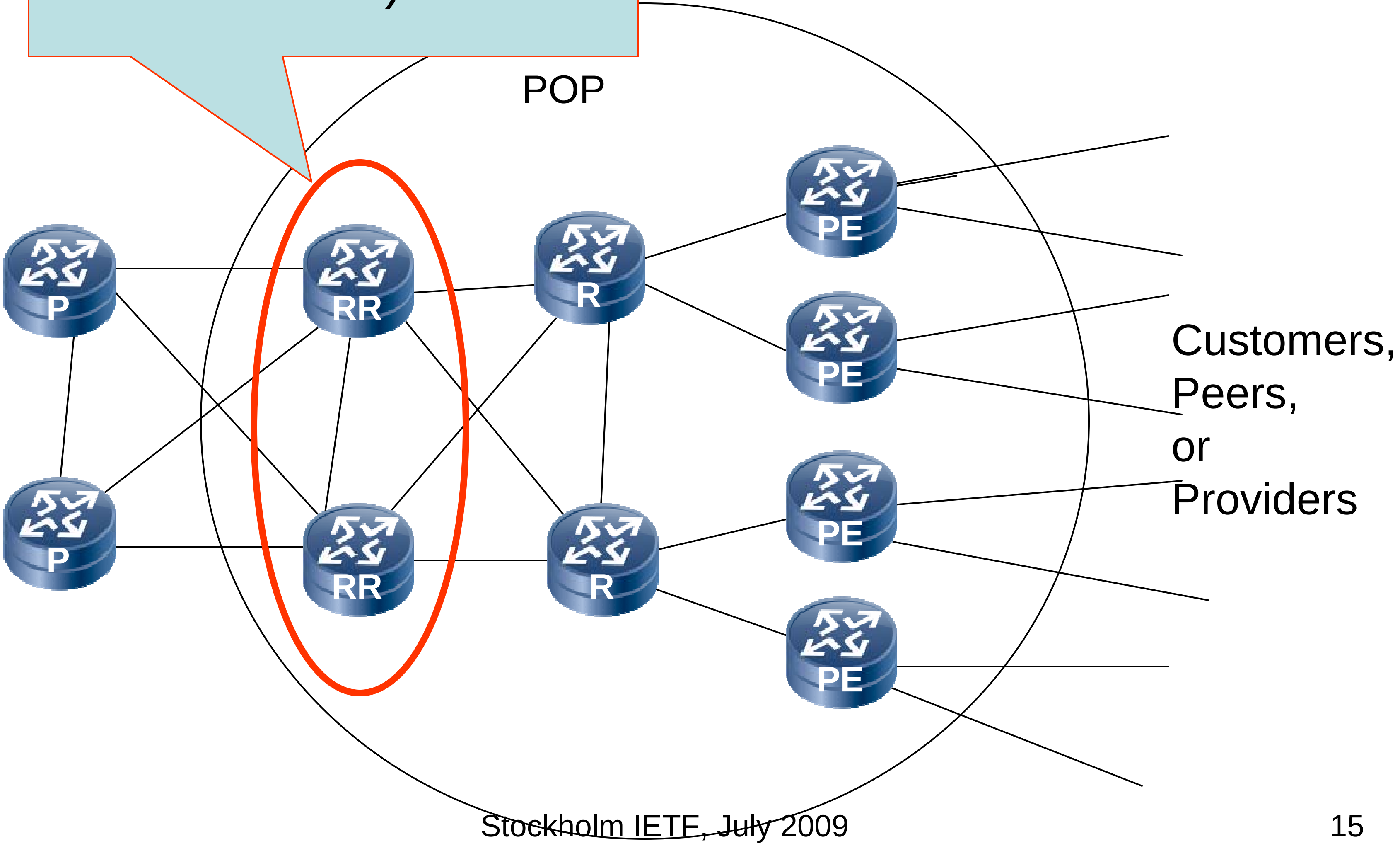


RR’s can tunnel over core routers

Customers, Peers, or Providers

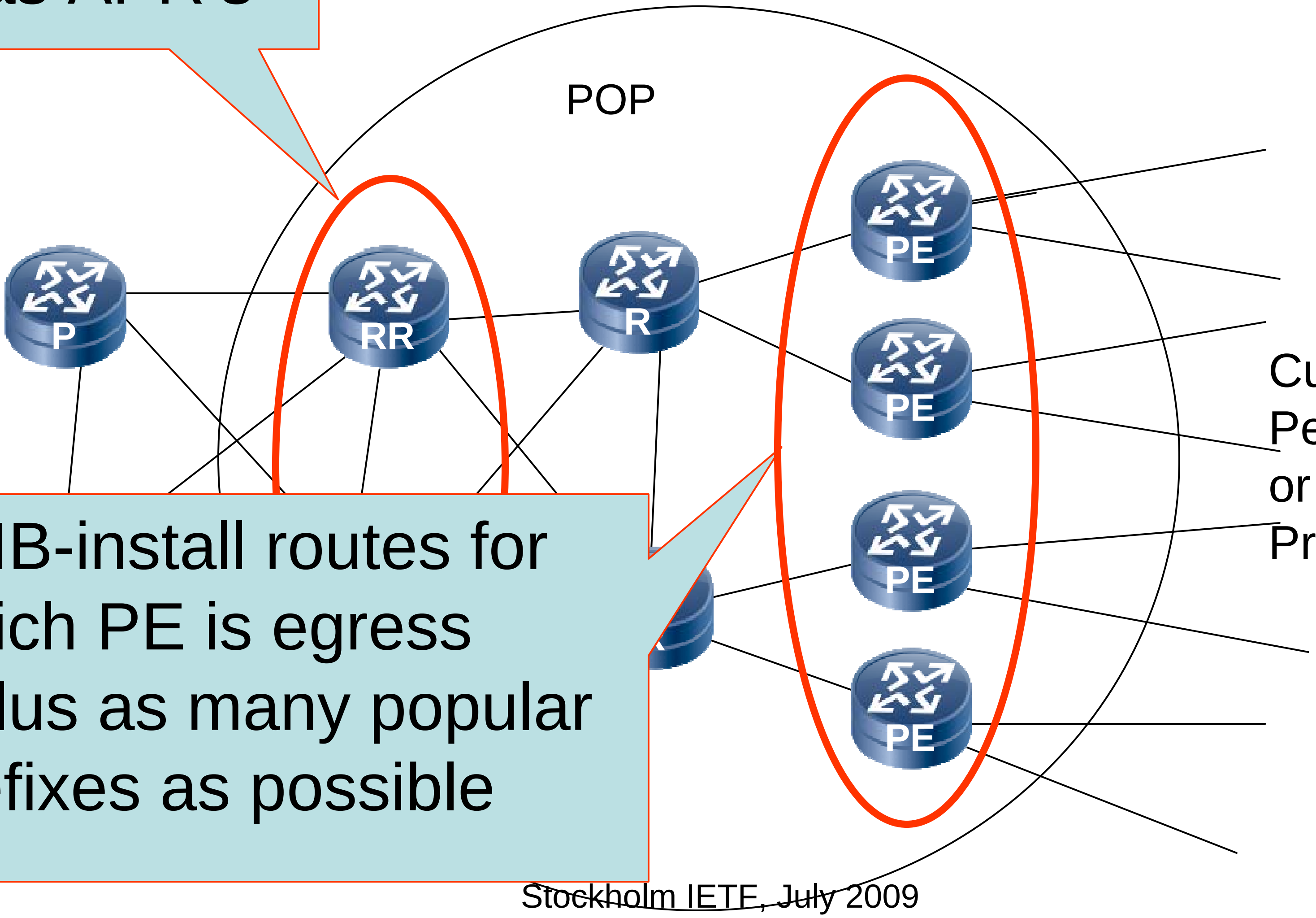
Use RR’s as APR’s

(Can optionally do FIB reduction here)



Customers,
Peers,
or
Providers

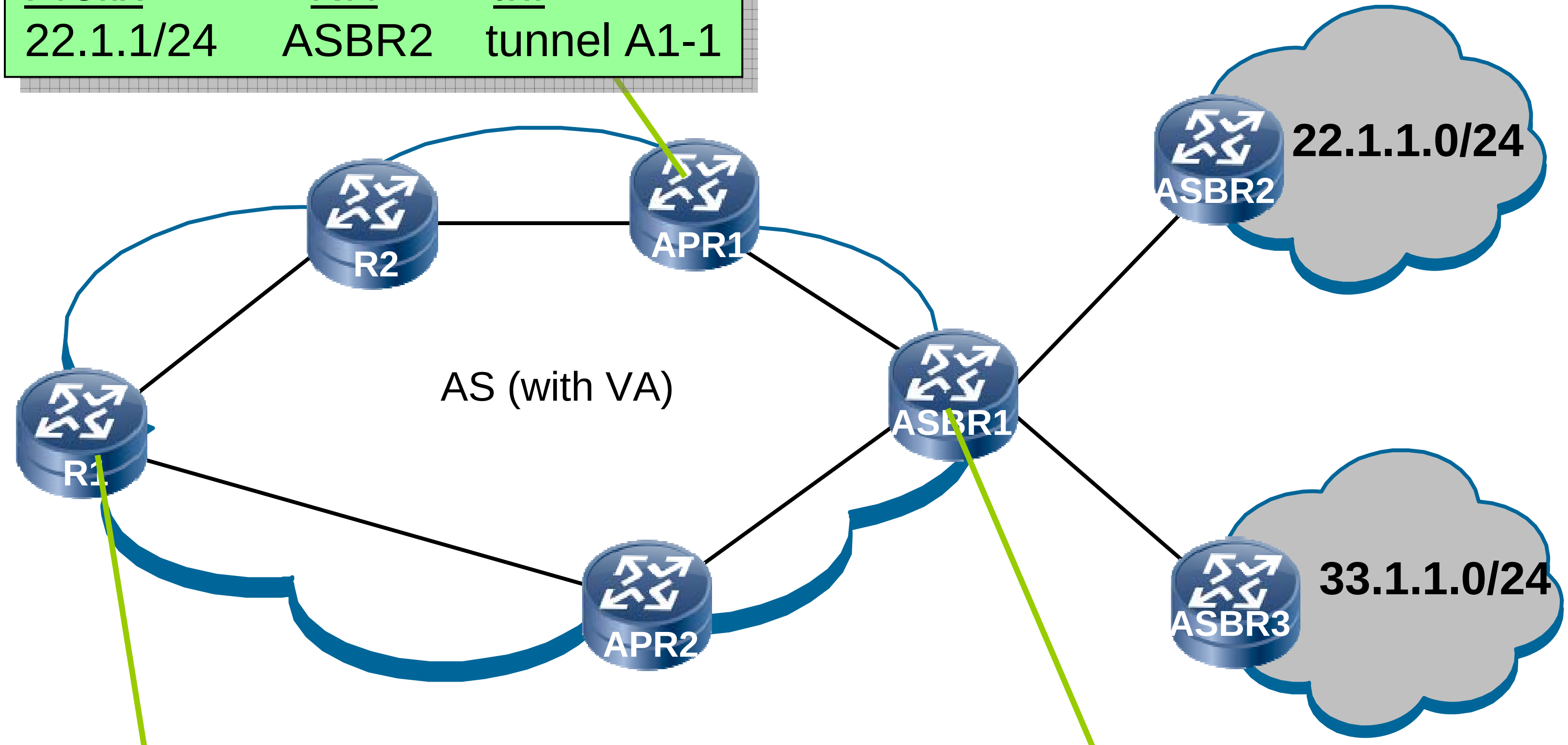
Use RR’s as APR’s



- FIB-install routes for which PE is egress
- Plus as many popular prefixes as possible

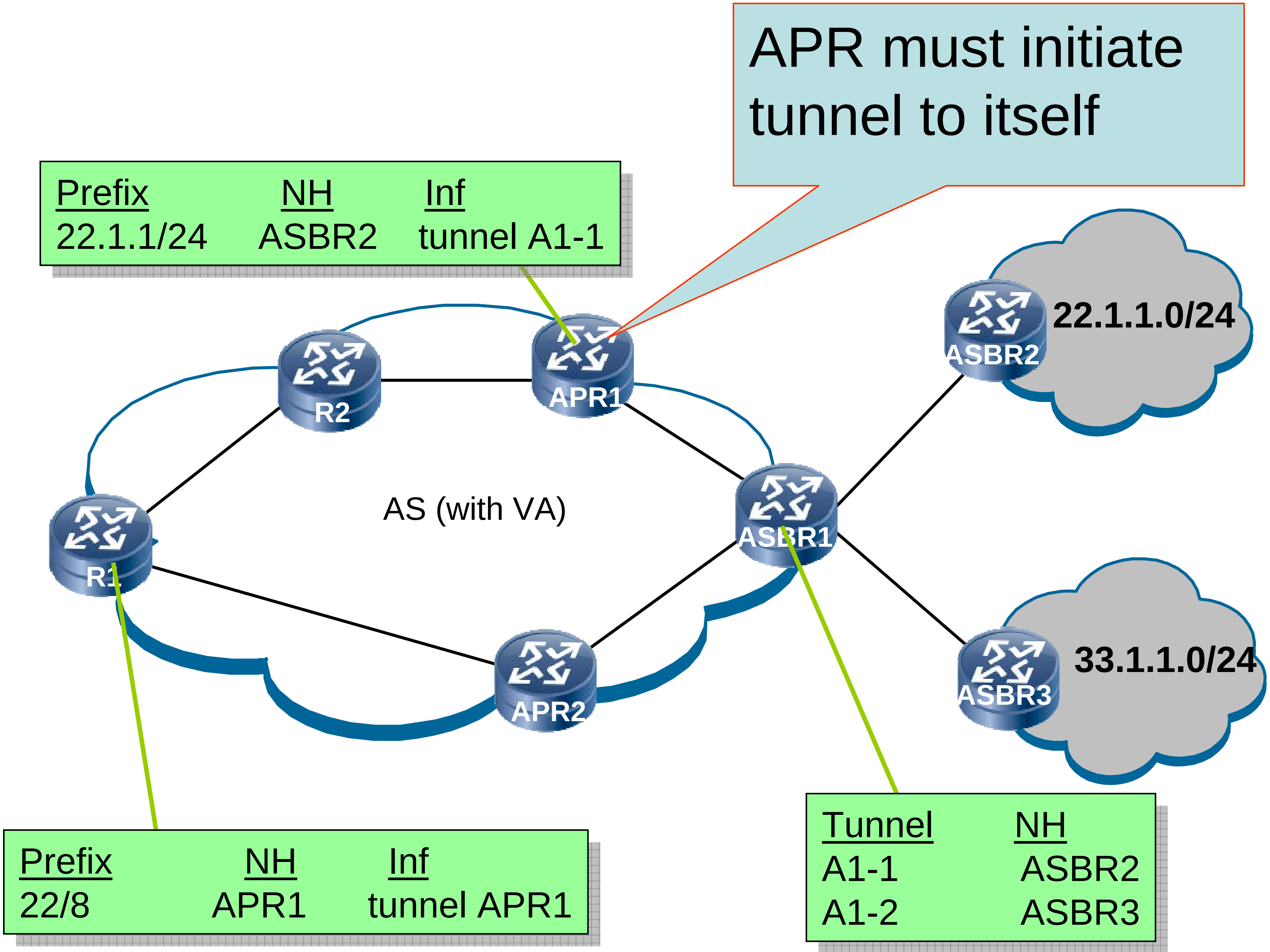
How are tunnels configured?

Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1



Prefix	NH	Inf
22/8	APR1	tunnel APR1

Tunnel	NH
A1-1	ASBR2
A1-2	ASBR3



APR must initiate tunnel to itself

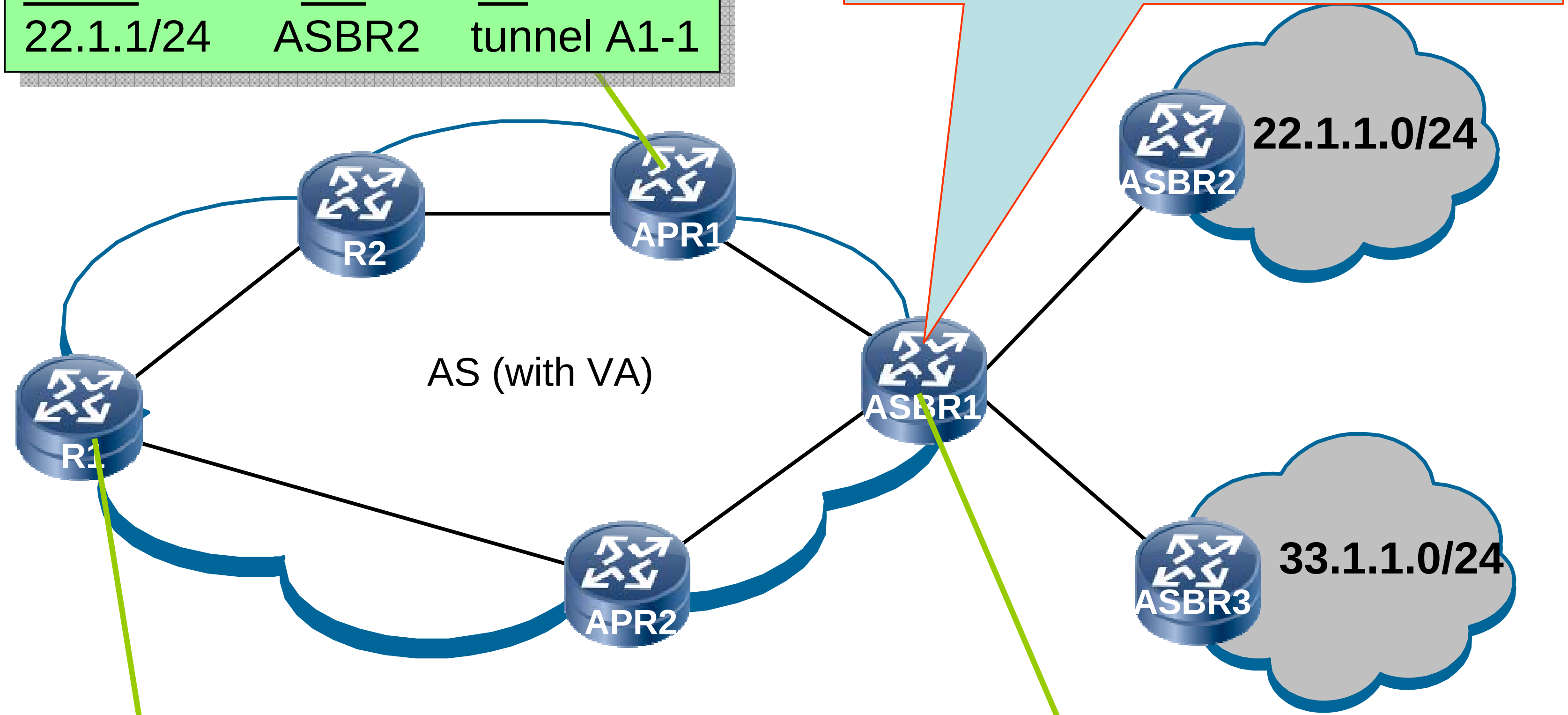
Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1

Prefix	NH	Inf
22/8	APR1	tunnel APR1

Tunnel	NH
A1-1	ASBR2
A1-2	ASBR3

ASBR must initiate a tunnel per neighbor remote ASBR

Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1



Prefix	NH	Inf
22/8	APR1	tunnel APR1

Tunnel	NH
A1-1	ASBR2
A1-2	ASBR3

Tunnels in VA drafts

- MPLS (using LDP)
- IP-in-IP (using RFC5512)
- GRE (using RFC5512)

Tunnel to APR

- Advertise loopback address as Next_Hop (NH) in BGP update for VP route
- If MPLS
  - Use LDP to establish tunnels to its loopback address (/32)
- If IP-in-IP
  - Use RFC5512 BGP Encapsulation Extended Attribute in VP route
- If GRE with Key
  - Use RFC5512 Tunnel Encapsulation Attribute in VP route

Tunnels to ASBR

- If MPLS
  - – Use LDP to establish tunnel to every remote neighbor ASBR
    - Remote ASBR address is tunnel target
  - – Use remote ASBR address as NH in BGP updates
  - – Use PHP mechanism to strip MPLS header before delivering to remote ASBR

Tunnels to ASBR

- If GRE with Key
  - Assign a unique GRE Key to every remote neighbor ASBR
  - In BGP update:
    - Use remote ASBR address as NH
    - Advertise Key value in RFC5512 Tunnel Encapsulation Attribute

Tunnels to ASBR

- If IP-in-IP or GRE without Key
  - – Assign a unique loopback address to every remote neighbor ASBR
    - i.e. remote ASBR1 = 10.1.1.1, remote ASBR2 = 10.1.1.2, etc.
  - – In BGP update:
    - Use unique loopback address as NH
    - Use RFC5512 BGP Encapsulation Extended Attribute to indicate that tunneling should be used

Scalability of tunnels

- MPLS signals one tunnel per remote ASBR
  - – Roughly 20K tunnels in transit ISP we studied
  - ☹ Each tunnel requires LDP signaling, and a /32 in OSPF
  - ☺ Can reduce to one tunnel per local ASBR
    - By using stacked MPLS tags

☺ IP-in-IP advertises one prefix per local ASBR

☺ Keyed GRE has one tunnel per remote ASBR

FIB-install rules

- APRs must FIB-install all sub-prefixes within VP
- All routers must FIB-install all Virtual Prefixes (VP)
- All other prefixes may be FIB-suppressed

This requires that:

- APRs must know their own VPs
- All routers must know complete VP-list

All routers must know complete VP-list

- Current spec proposes a static table configured in all routers
  - Same table for all routers
- Current spec describes how to modify list (add, remove, merge, split)
  - Must be done in such a way that:
    - Forwarding is not disrupted
    - The FIB doesn't temporarily grow beyond its "before" and "after" sizes

Adding and removing VPs

- Adding a VP:
  - First configure VP in APR
    - FIB-install sub-prefixes
  - Then add VP to all VP-lists
    - FIB-suppress sub-prefixes
- Removing a VP:
  - First remove VP from all VP-lists
    - FIB-install sub-prefixes
  - Then remove VP from APR
    - FIB-suppress sub-prefixes

Splitting and Merging VP

- Splitting a VP
  - First do an add on both nested child VPs
  - Then do a remove on the parent VP
- Merging VPs
  - First do an add on the parent
  - Then do a remove on the child VP

Configuring Popular Prefixes

- The current spec mostly punts on this
  - Or, more politically correctly, leaves it to vendors as a competitive feature
- Some simple things can be done:
  - FIB-install all customer sub-prefixes
  - FIB-install all sub-prefixes for which the router is the egress
- But FIB-installing high-volume sub-prefixes is less easy

Automatic configuration?

- Do we need automatic config of the VP-list and high-volume sub-prefixes?
- And if so, how do we do it?

Automating config of high-volume sub-prefixes

- Note that it is the ingress router that needs to FIB-install to obtain shortest-path benefit

Two cases:

1. ASBR sees high volume incoming
   - Independently FIB-install high-volume sub-prefixes
2. ASBR sees high volume outgoing
   - Can be from many ingress routers, few of which see high-volume
   - Must somehow inform the ingress routers

Tagging high-volume sub-prefixes

- ASBR (or data-plane RR) identifies high-volume outgoing sub-prefixes
- ASBR/RR attaches a “should FIB-install” tag (attribute) to BGP updates for the sub-prefix
- Other routers use this as a hint in their FIB installing decision process
  - i.e. don’t need to FIB-install if there isn’t room

Auto-config of VP-list:
Tag VP approach

- Original VA spec had auto-config of VP-list:
  - APR would tag VP routes with “this is a VP” attribute
    - ☺ No new config required, since APRs must know their VPs in any event
  - Routers install sub-prefixes unless within a VP
  - ☹ Problem was that a booting router may not see tagged VP route until *after* installing many sub-prefixes and possibly over-flowing the FIB

Auto-config of VP-list: Tag VP approach

- One solution:
  - – Keep “this is a VP” attribute as originally envisioned
  - – Rather than “FIB-install by default”
    - Unless shown to be within a VP
  - – Do: “FIB-suppress by default”
    - Unless shown NOT to be within a VP

  - – Downside is that many entries not FIB-installed until BGP done initializing
  - – But this mitigated by GR (graceful restart)

Auto-config of VP-list: “May suppress” tag approach

- Another solution:
  - – Install “VP ranges” in some fraction of routers
    - Only RRs
    - Only edge routers
  - – Routers with “VP ranges” tag updates for sub-prefixes within VPs with a “may FIB-suppress” attribute
    - Routers know they can FIB-suppress the sub-prefix as soon as they learn the route

☹ This solution requires static configuration of “VP ranges” in some routers

Next steps

- Discuss various auto-config approaches on mailing list
  - May lead to standards track rather than informational
- Discuss stacked MPLS tags on mailing list
- Write deployment/scenarios draft
- Continue working on implementations

Auto-config of VP-list: Tag VP approach

- One solution:
  - – Keep “this is a VP” attribute as originally envisioned
    - This gives routers the VP-list in steady state
  - – Routers remember the VP-list between boots
  - – Routers assume “old” VP-list when start booting, modify VP-list during boot as new attributes received
    - Normally no or few changes between boots…

Current status

- WG item in GROW
- Four drafts:
  - draft-ietf-grow-va-00
    - Francis, Xu, Ballani, Jen, Raszuk, Zheng
  - draft-ietf-grow-va-gre-00
    - Xu, Francis, Raszuk
  - draft-ietf-grow-va-mpls-00
    - Francis, Xu
  - draft-ietf-grow-va-perf-00
    - Ballani, Francis, Jen, Xu, Zhang