Reducing FIB Size with Virtual Aggregation (VA)

Paul Francis, MPI-SWS

Xiaohu Xu, Huawei,

Hitesh Ballani, Cornell

Dan Jen, UCLA

Robert Raszuk, Cisco

Lixia Zhang, UCLA

ISPs often want to extend the life of old routers

- Routers that have inadequate FIB, but otherwise are still useful
- A common approach: use old routers as customer PE, default to ISP core
  - Customer gets partial or no DFZ updates, but this often fine with customer
- But this is not always enough

Other FIB/RIB shrinking tricks

- Filter out more specific routes
  - Can lead to unreachability

- For lower-tier ISPs, default to transit ISPs
  - I.e. use 0.0.0.0/0 and load balance among transit ISPs
  - Filter out most or all routes from transit ISPs
- But:
  - Leads to non-optimal routes
  - Lots of configuration (peer routes, “important” routes like Google….)
  - Can’t be used by transit ISPs themselves

Mitigating non-optimal default routes

- Use more-specific “Semi-defaults”
  - I.e. AS3303 (Swisscom IP-Plus)
    - Andre Chapuis chapuis@ip-plus.net, SwiNOG 7 presentation, http://www.swinog.ch/
  - Various semi-defaults:
    - 62/8, 80/7, 212/7, 217/8 → EU transit ISP
    - ARIN/APNIC/LACNIC space → US transit
    - Class B: 128/3, 160/5 and 168/6 → US transit
- But still more configuration . . . . .
- Trade-off between RIB/FIB size and path quality
  - AS3303 gets very good paths for most traffic for 50% RIB/FIB reduction

IETF working on a more general solution: Virtual Aggregation

- GROW working group
  - People

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

-
  - Drafts:
    - draft-ietf-grow-va-00
      - draft-ietf-grow-va-gre-00
      - draft-ietf-grow-va-mpls-00
      - draft-ietf-grow-va-perf-00

What is Virtual Aggregation?

- A way to control FIB size in routers
  - DFZ FIB, not VPN tables
  - Does not shrink RIB size
- Tight control of FIB size for *any or all* routers
- No coordination between ISPs
- Works with legacy routers

Important today:
Perhaps critical tomorrow?

- Looking forward, BGP RIB growth rate could increase substantially
  - Because exhaustion of IPv4 erodes aggregation
  - Because of pressure to shrink default prefix size
  - Because of uptake of IPv6

- VA allows installed router base to absorb this growth

VA not perfect….

- Requires configuration of its own
- Entails a traffic load / FIB size trade-off
  - Which can be quite good
  - Academic study on Large Transit ISP:
    - 10X FIB reduction with negligible latency/load penalty
  - But in general we don’t know how easy to achieve this
    - Configuration……

Why this talk?

- You can help us define VA
  - Certain protocol or configuration details
  - Alternative ways to deploy
- Or, tell us that VA us useless….
- You can encourage your vendor to implement VA
  - Currently implementations from Huawei and MPI-SWS (Quagga/linux) in progress

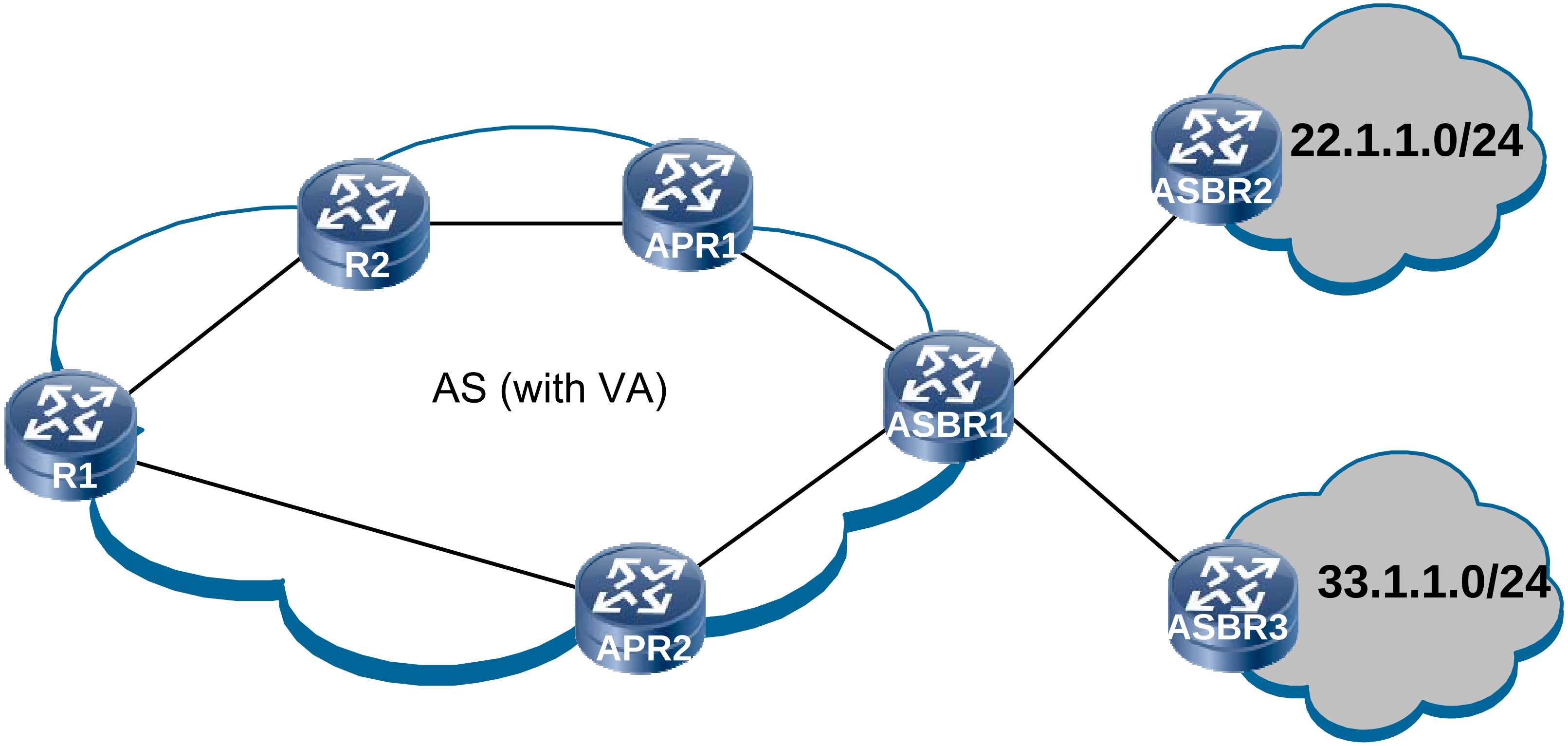
VA: Basic Idea

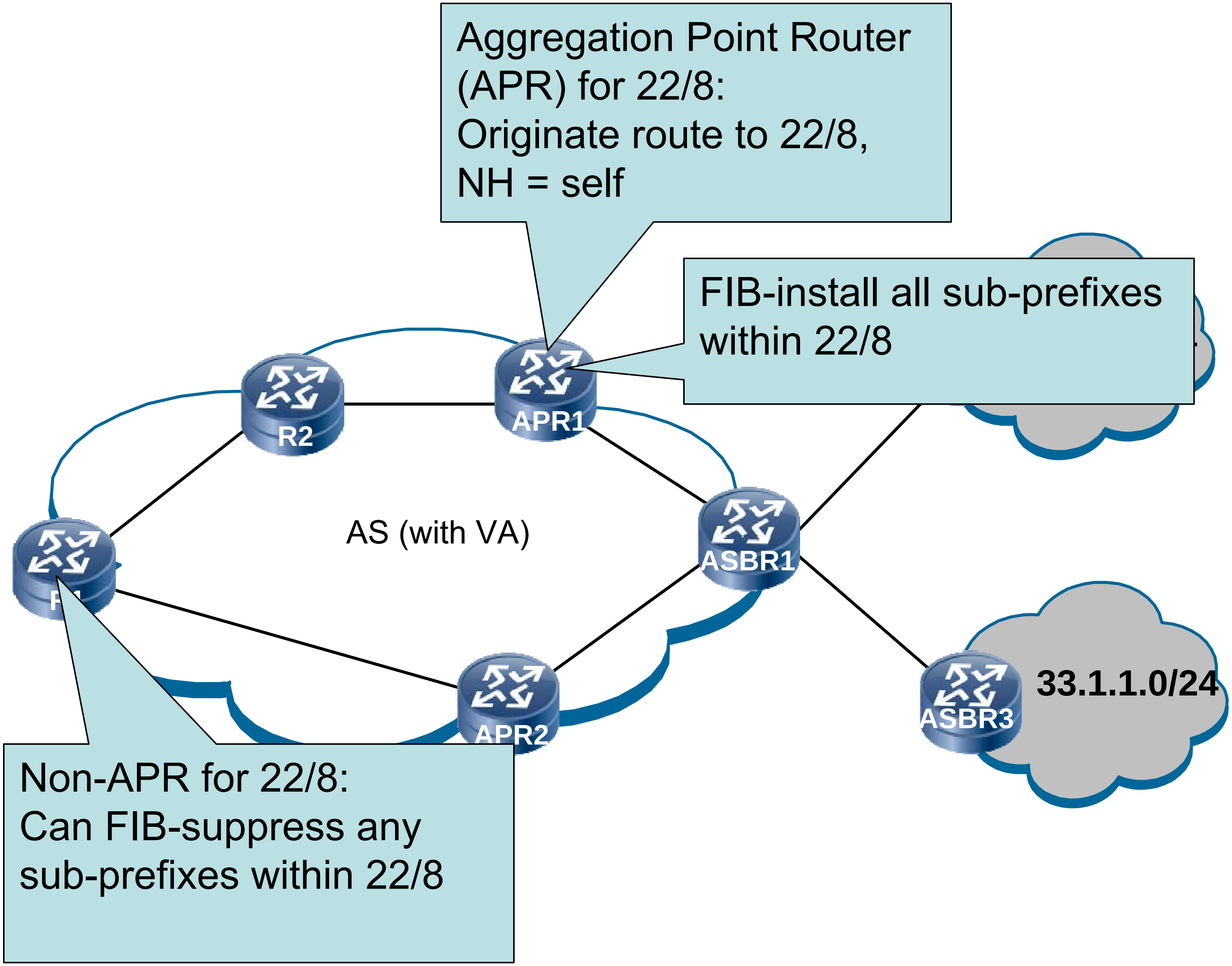
- Define “Virtual Prefixes” (VP)
  - These are shorter (bigger) than real prefixes
  - Thinks /6’s, /7’s, /8’s…..
- Assign different routers to be “responsible” for different Virtual Prefixes
  - I.e. they know how to route to everything in the VP
- Other routers don’t need to know how to route to everything
  - Rather, they can ***tunnel*** packets to the responsible routers

FIB-suppression

- BGP runs as normal
  - All routers have full RIB
  - Important not to muck with BGP operation per se
- VA simply doesn’t load certain prefixes into the FIB
  - i.e. those that the router is not responsible for

Basic VA mechanism

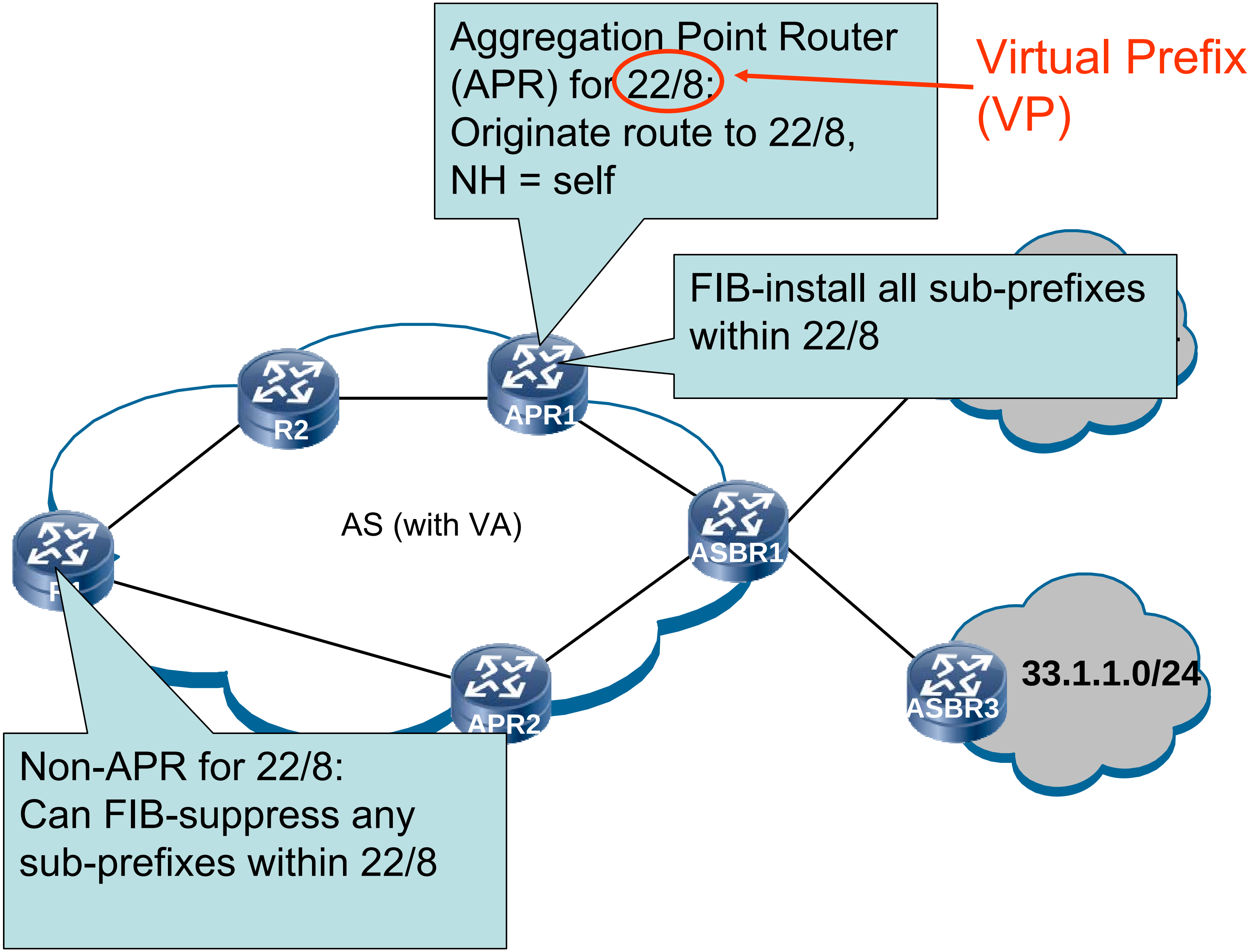




Aggregation Point Router (APR) for 22/8:
Originate route to 22/8,
NH = self

FIB-install all sub-prefixes within 22/8

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



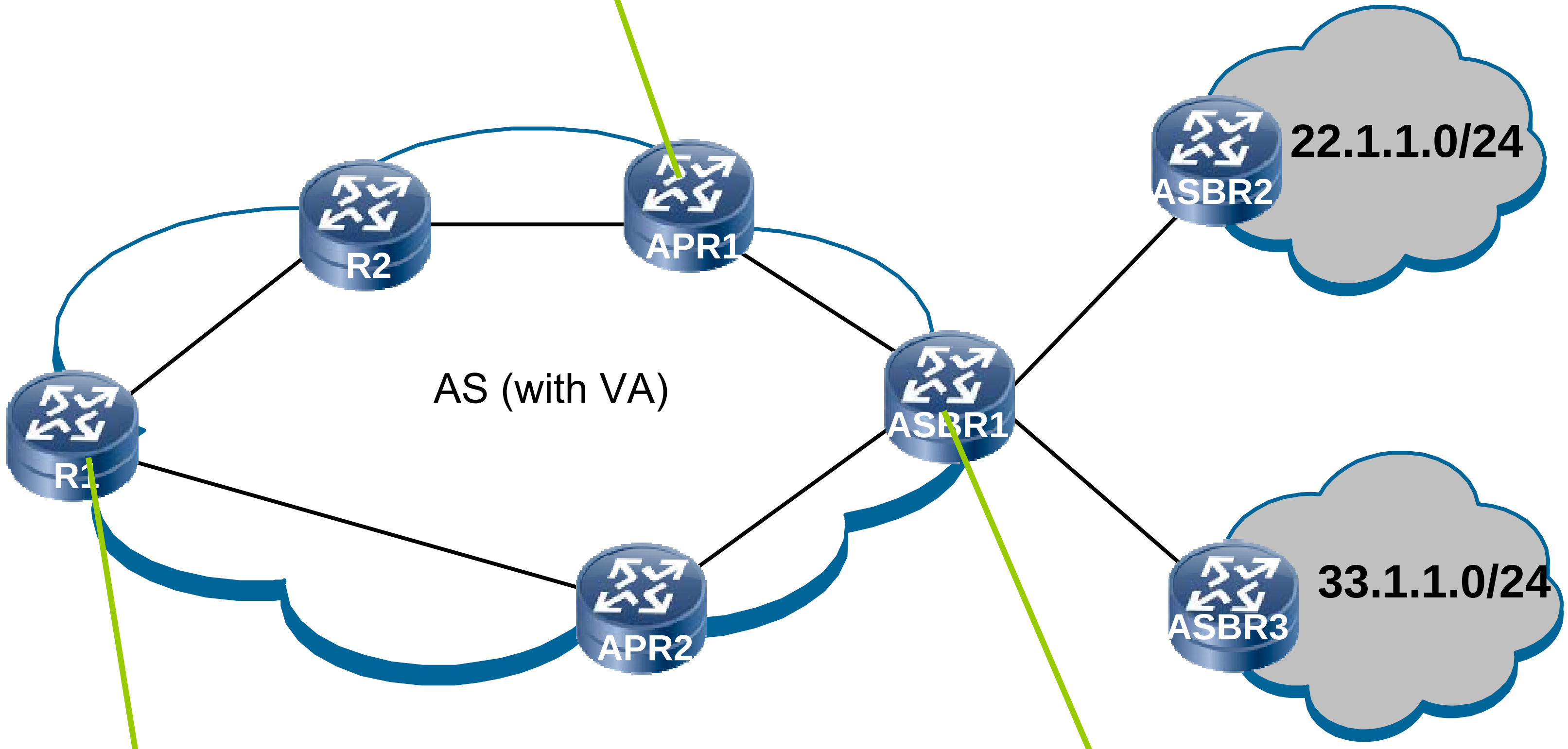
Aggregation Point Router (APR) for 22/8: Originate route to 22/8, NH = self

Virtual Prefix (VP)

FIB-install all sub-prefixes within 22/8

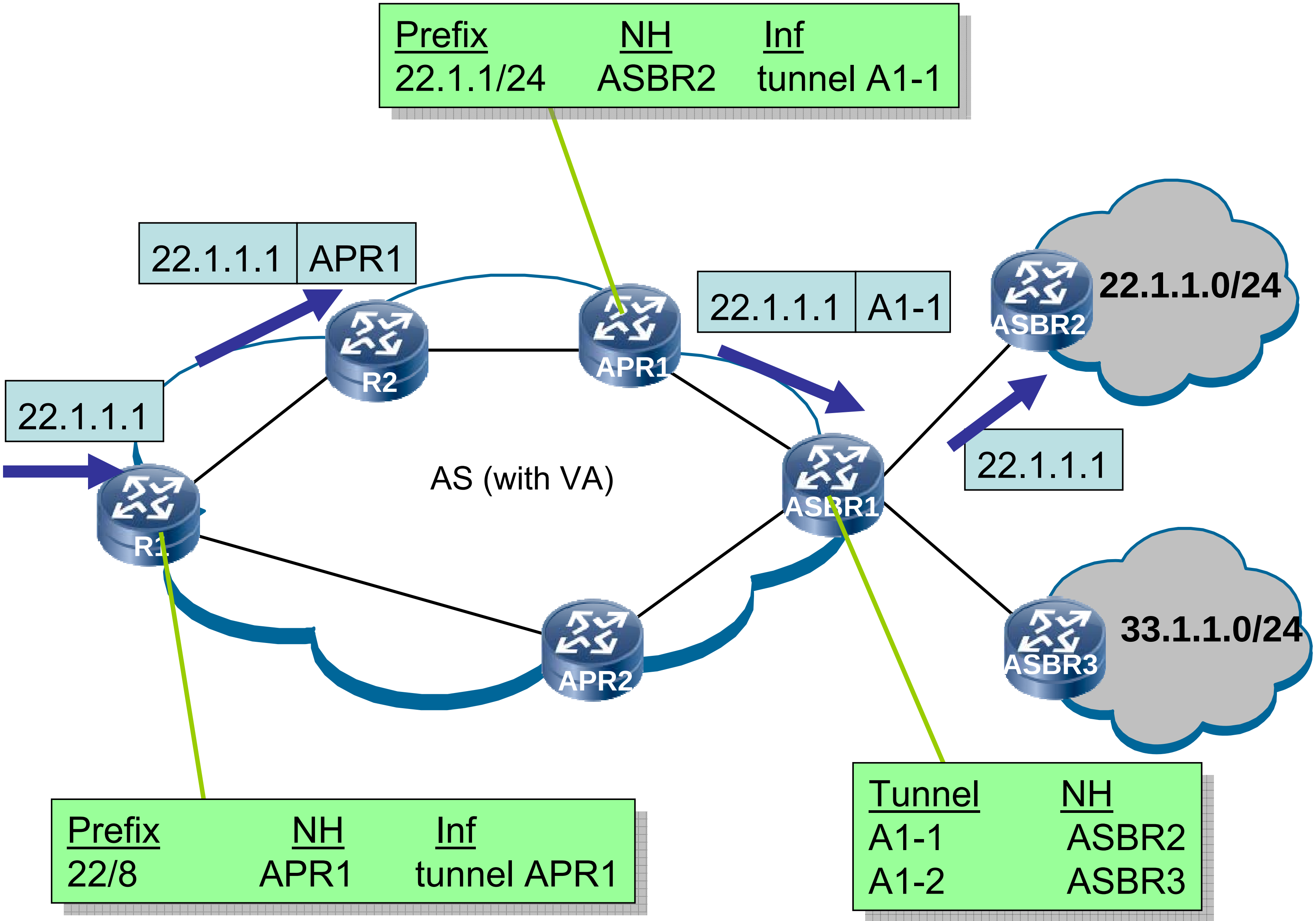
Non-APR for 22/8: Can FIB-suppress any sub-prefixes within 22/8

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

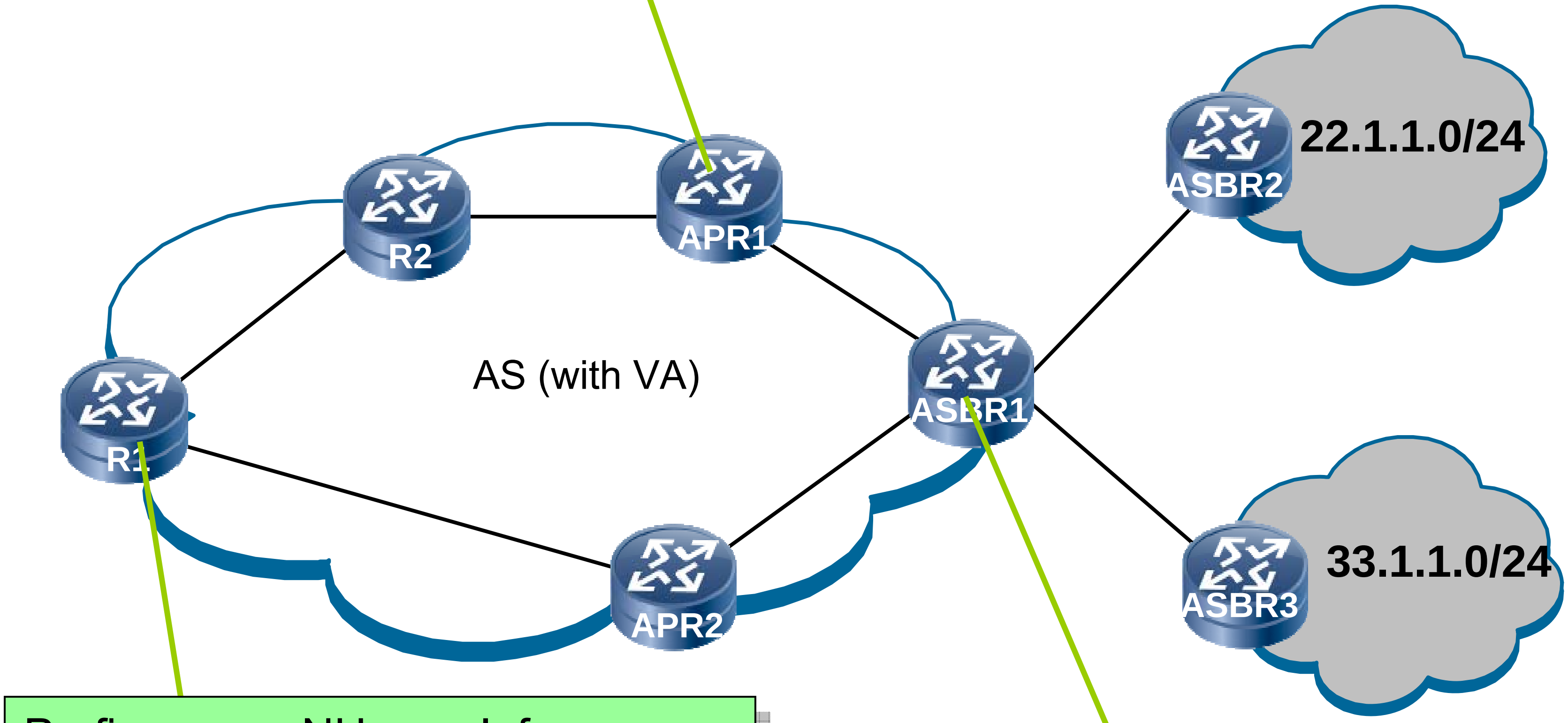


Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1

Prefix	NH	Inf
22/8	APR1	tunnel APR1

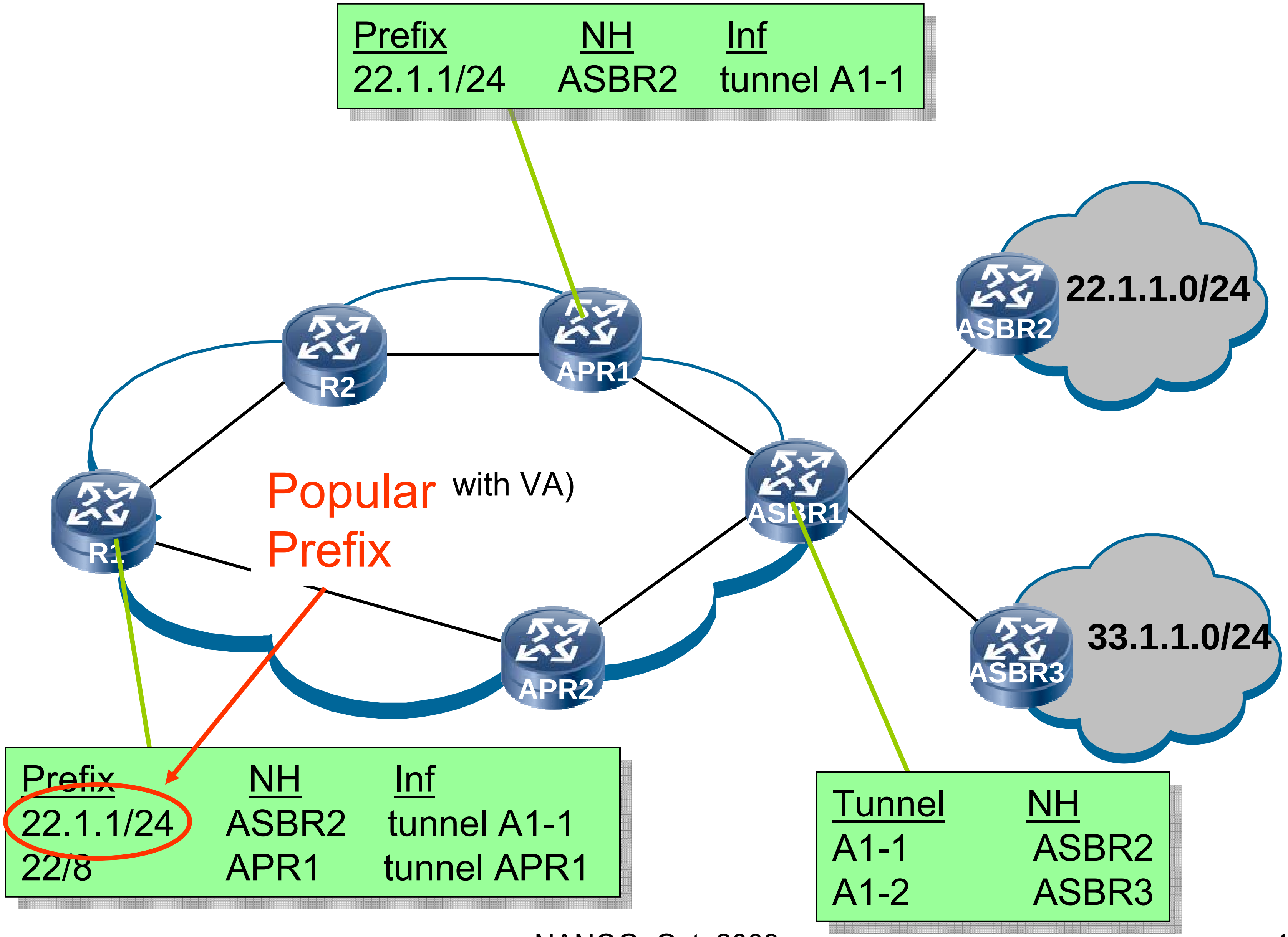
Tunnel	NH
A1-1	ASBR2
A1-2	ASBR3

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

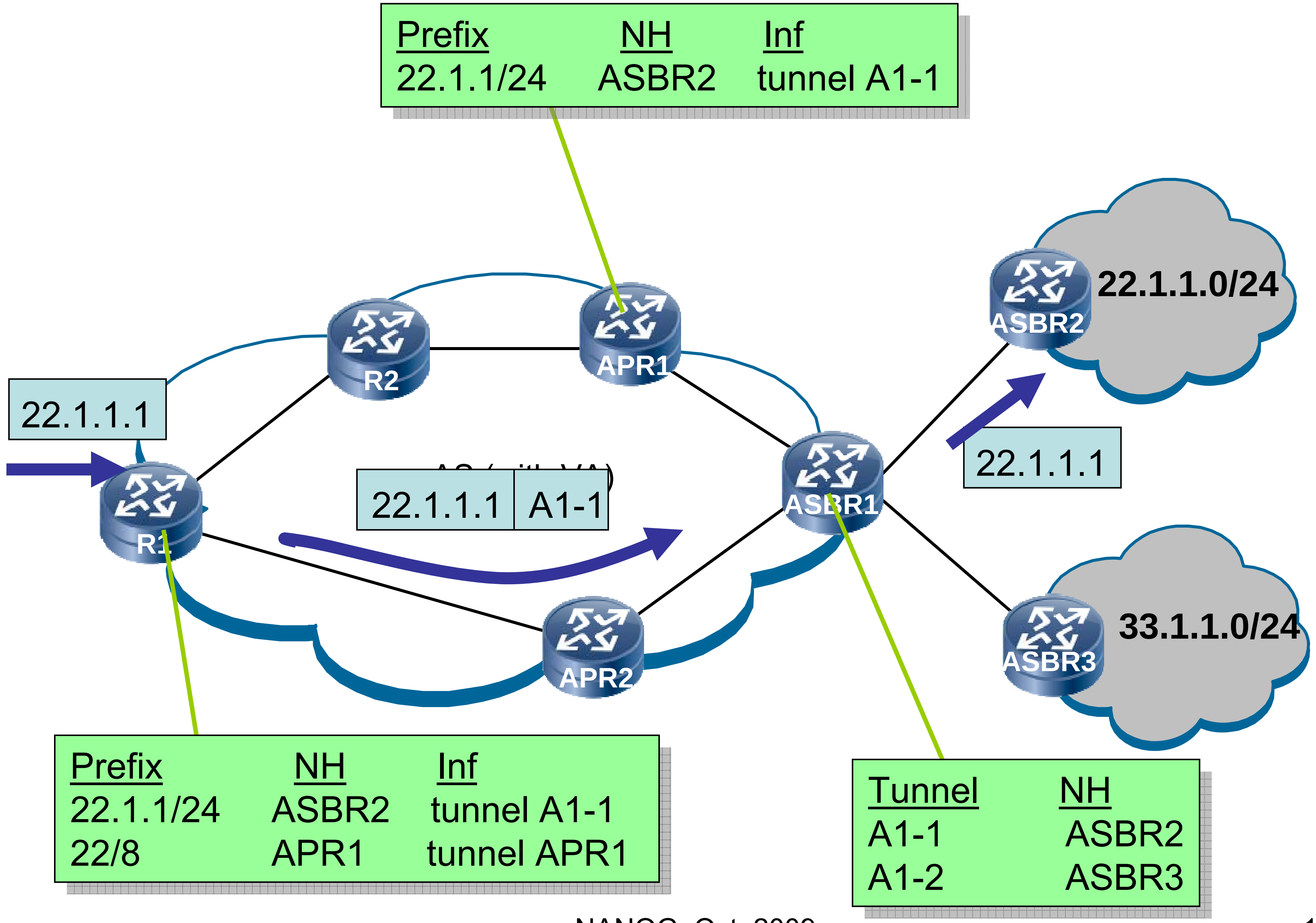
Tunnel	NH
A1-1	ASBR2
A1-2	ASBR3



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

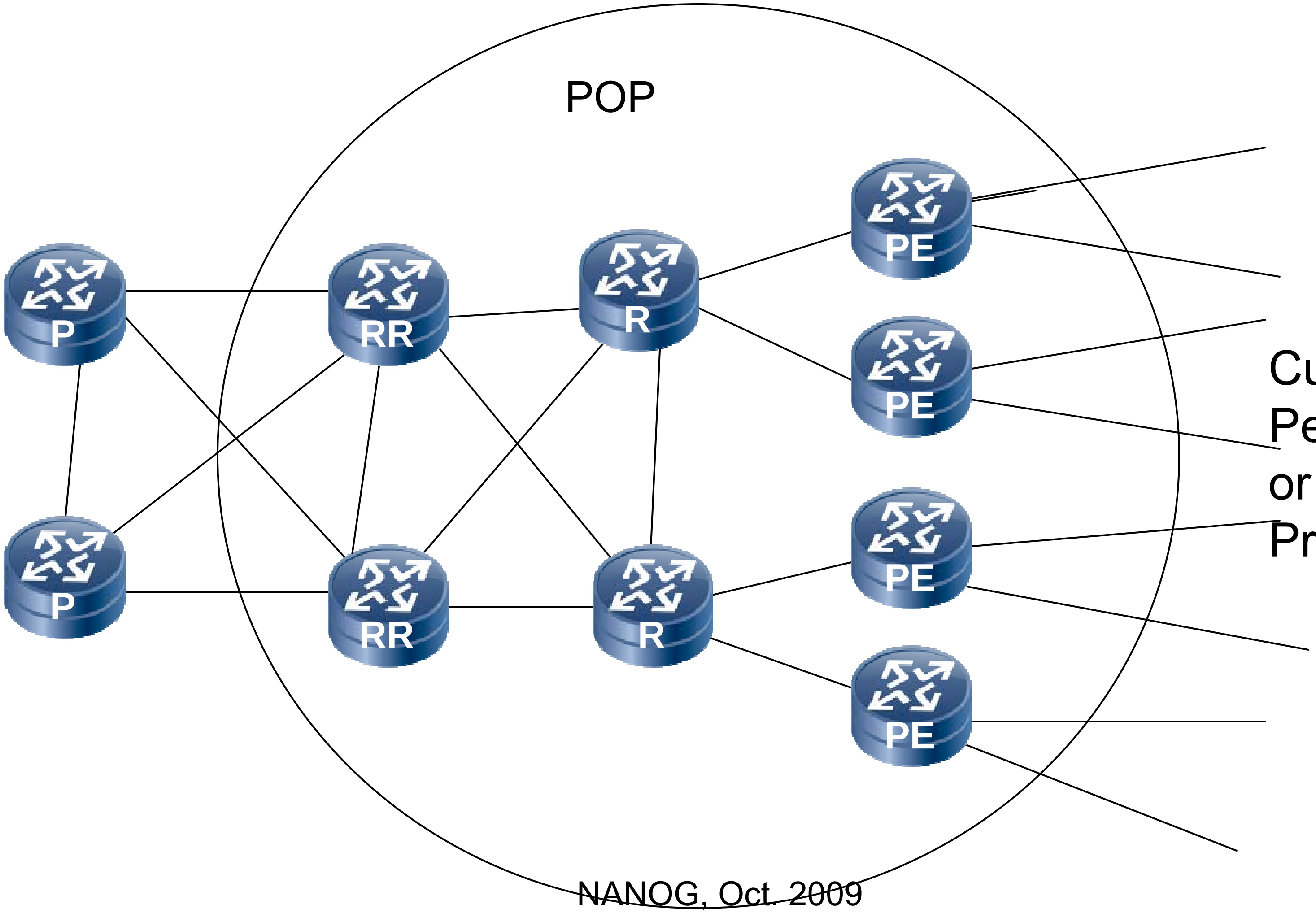
Types of tunnels defined

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

A deployment example

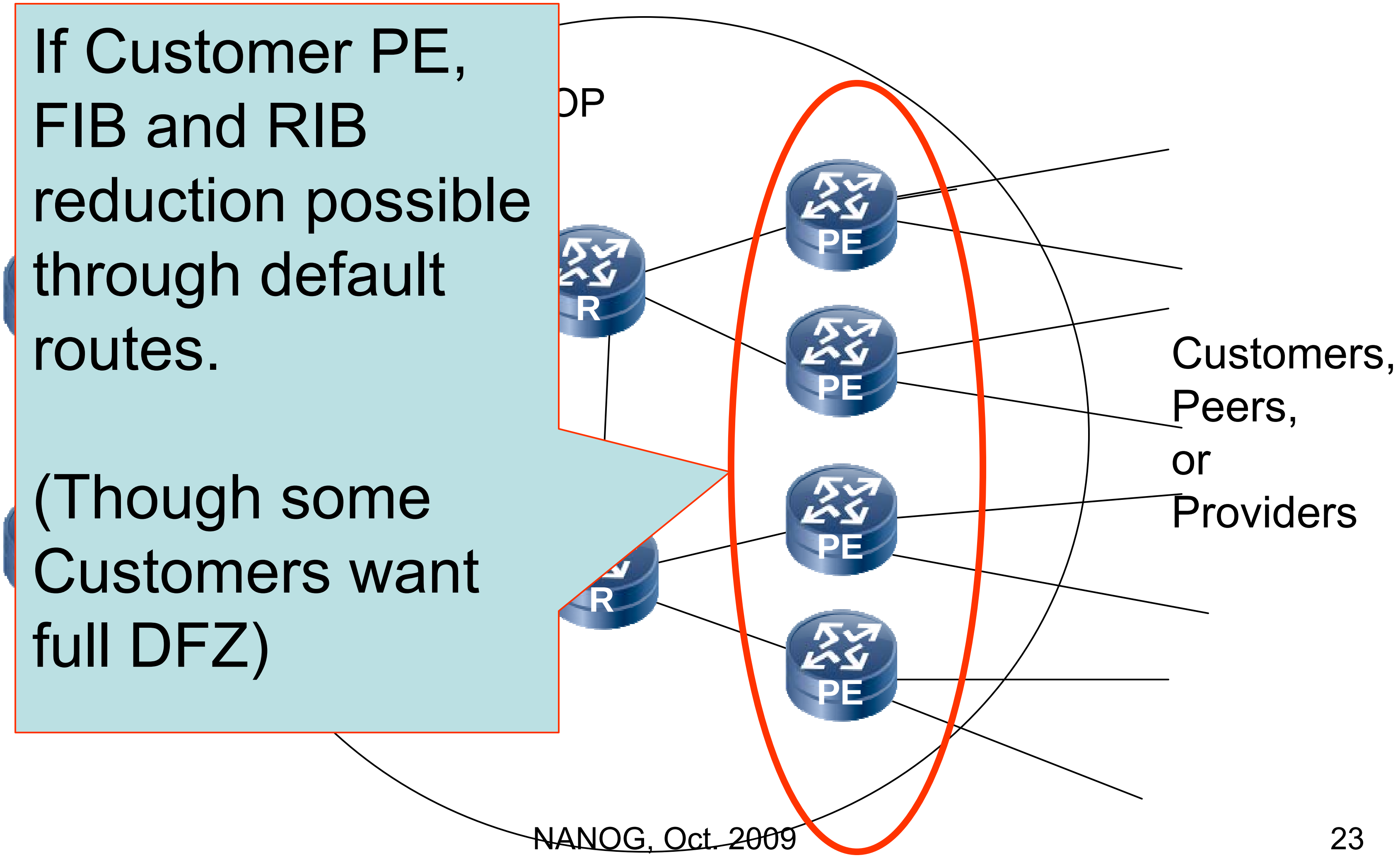
- Courtesy of Robert Raszuk, Cisco

A typical POP structure



Customers,
Peers,
or
Providers

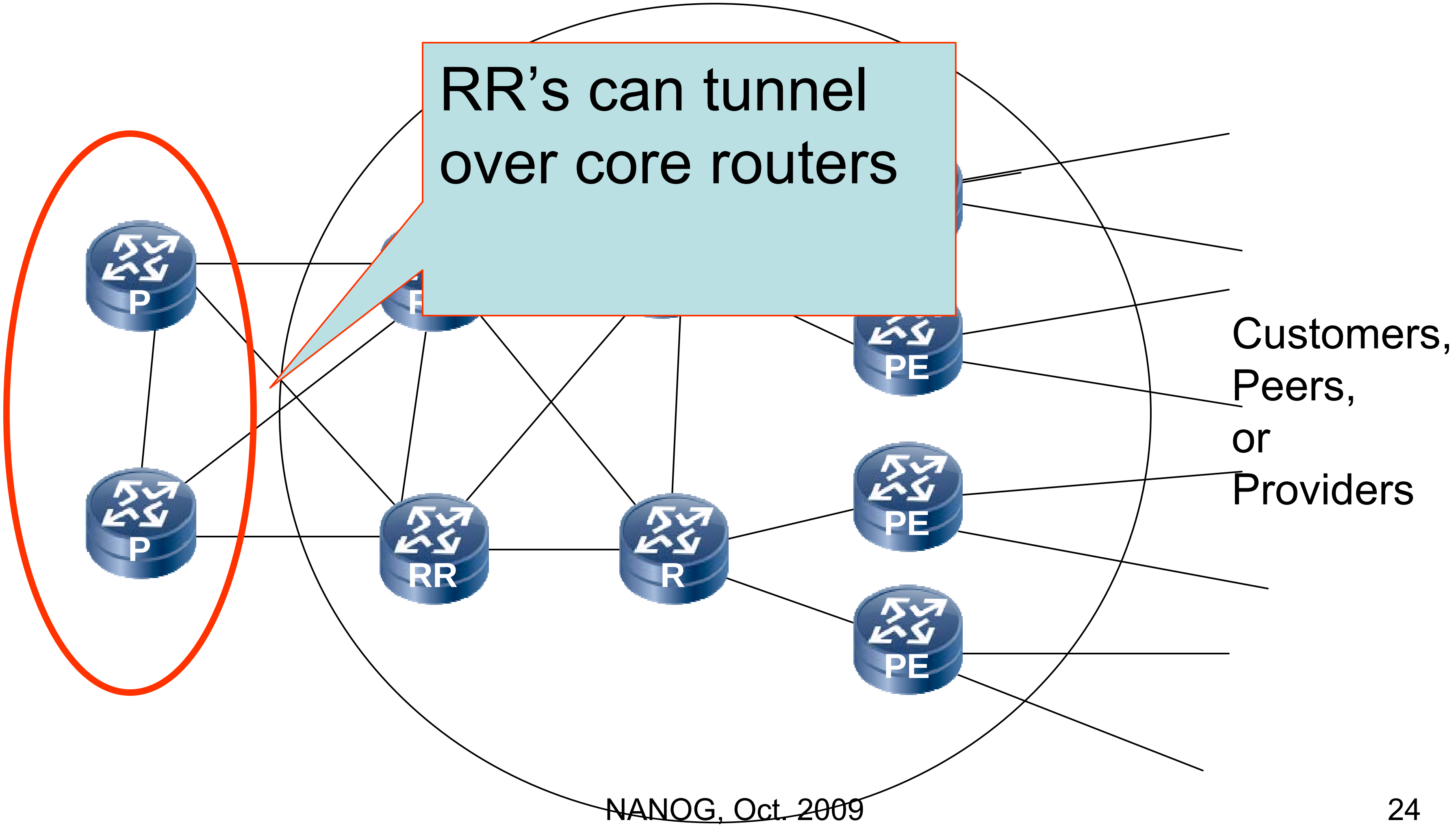
FIB reduction today



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

(Though some Customers want full DFZ)

FIB reduction today

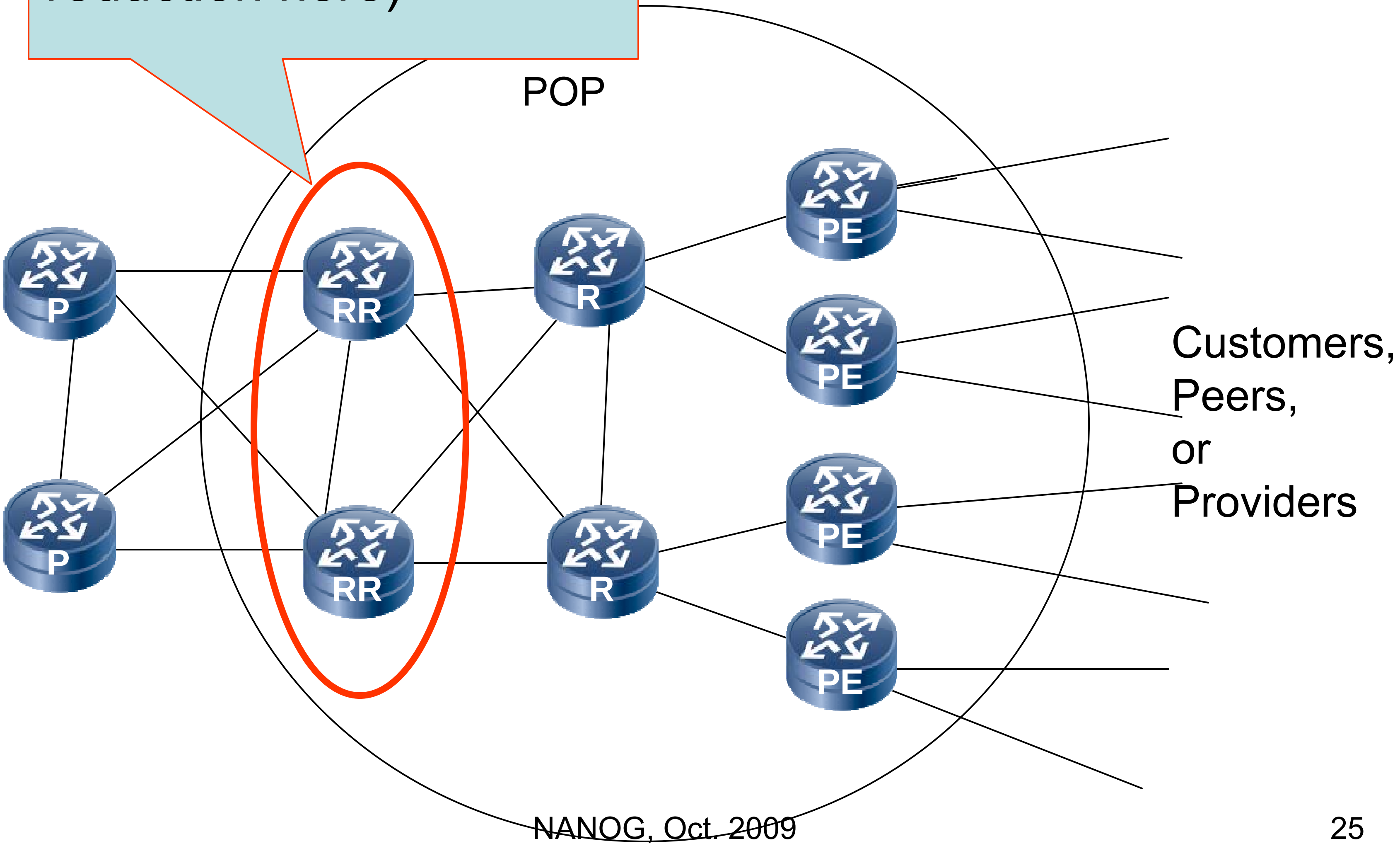


RR's can tunnel over core routers

Customers,
Peers,
or
Providers

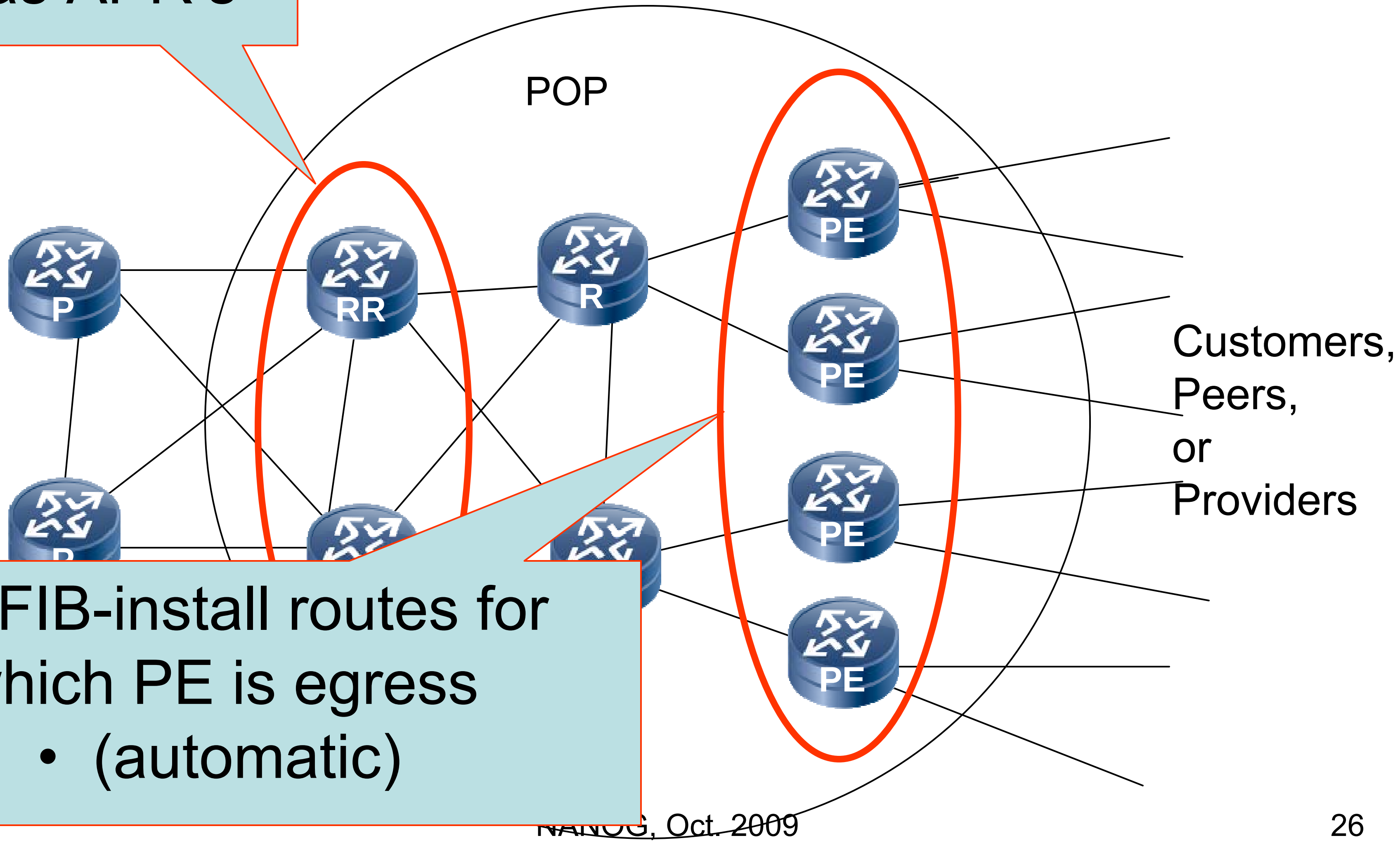
Use RR’s as APR’s

(Can optionally do FIB reduction here)

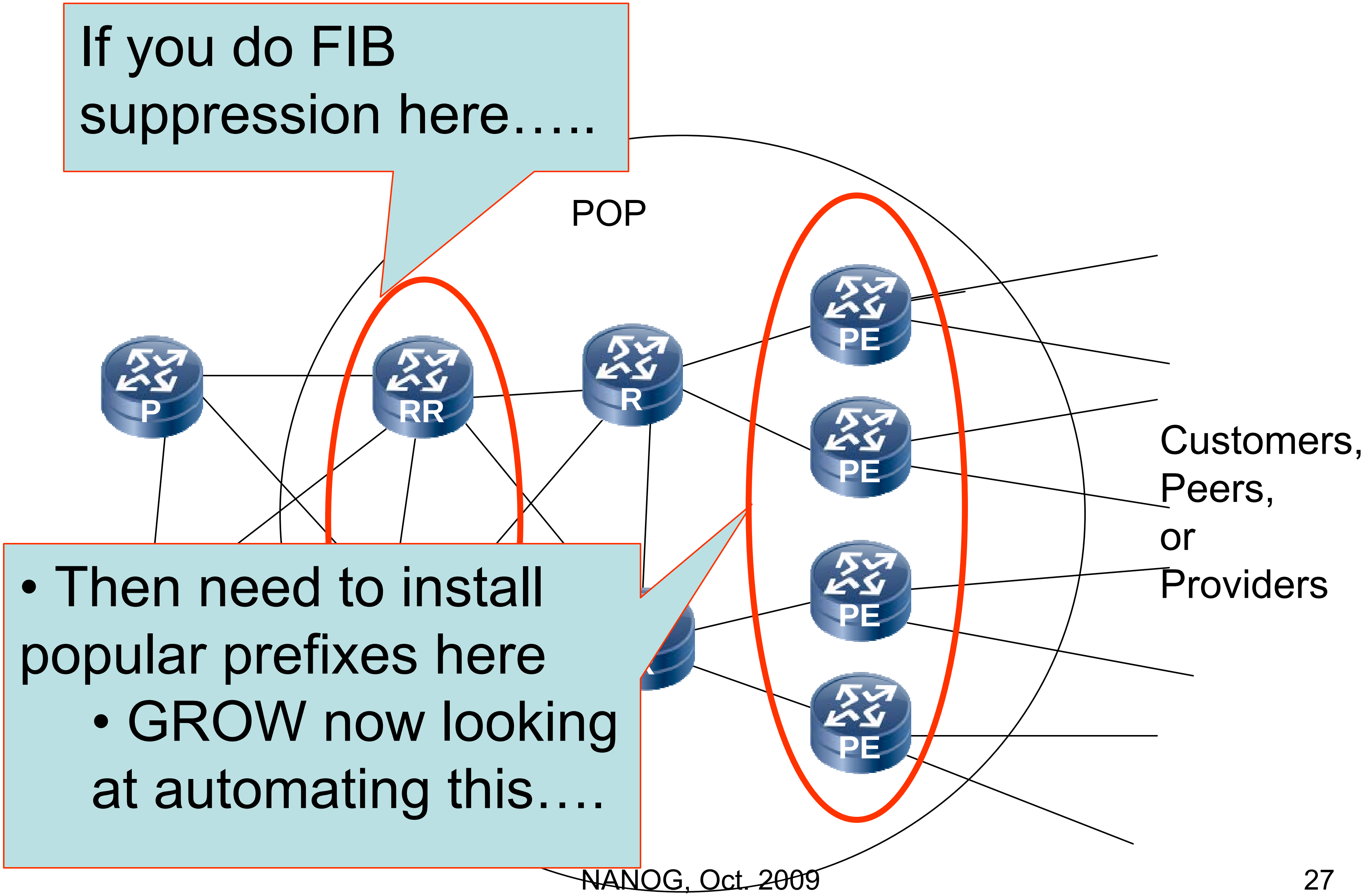


Customers,
Peers,
or
Providers

Use RR’s as APR’s



- FIB-install routes for which PE is egress
    - (automatic)



If you do FIB suppression here…..

- Then need to install popular prefixes here
  - GROW now looking at automating this….

VA from your point of view

- Figure out where you need FIB reduction
- Based on this, design your deployment
  - Select VPs
    - Just one /0 if all RRs keep full FIB
    - Otherwise, probably just all /7’s or something….
  - Assign routers as APRs, configure
  - Configure “VP-list” in all routers
    - (Though we are looking at how to eliminate this requirement)
- If you have FIB reduction everywhere (RRs included), then need to configure popular prefixes
  - (Though we are looking at how to automate this)

To summarize

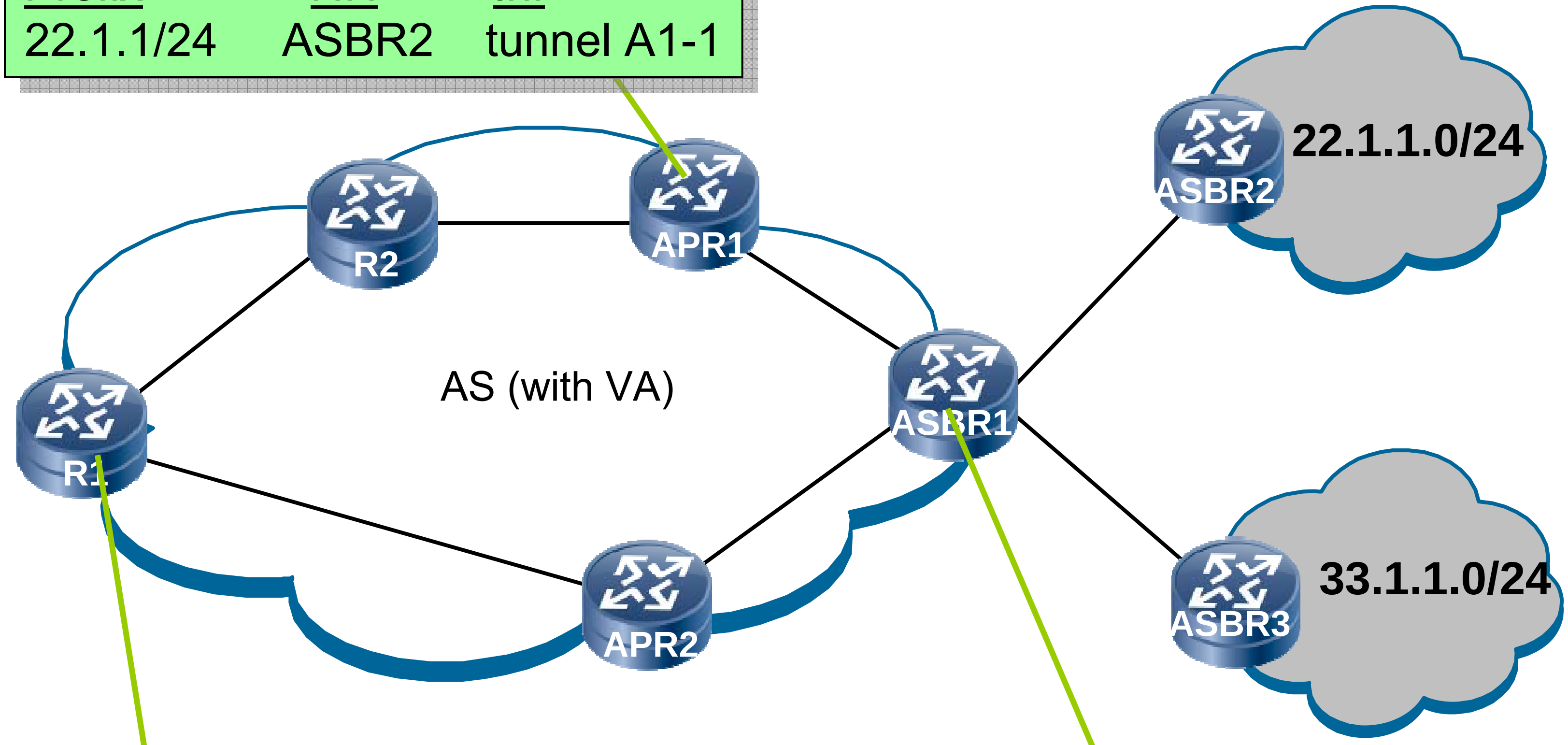
- New IETF GROW WG work item for FIB suppression
  - Allows you to extend the lifetime of older routers indefinitely
- Still early in the standards process
  - You can influence the design
- If this sounds useful, please talk to your favorite vendor

Thanks!

- IETF Drafts
  - draft-ietf-grow-va-00
  - draft-ietf-grow-va-gre-00
  - draft-ietf-grow-va-mpls-00
  - draft-ietf-grow-va-perf-00
- Other
  - "Making Routers Last Longer with ViAggre", NSDI 2009

How are tunnels configured?

Prefix	NH	Inf
22.1.1/24	ASBR2	tunnel A1-1

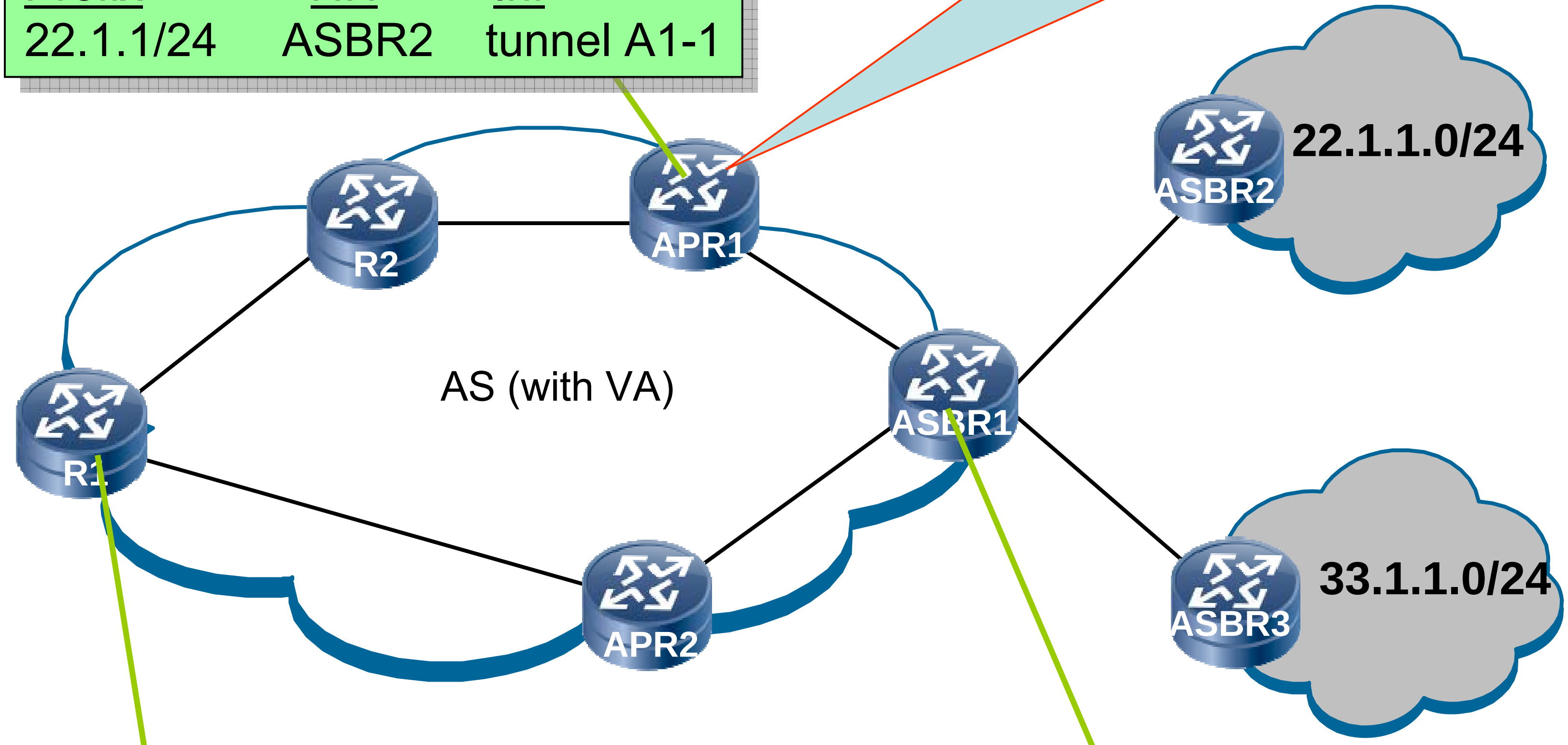


Prefix	NH	Inf
22/8	APR1	tunnel APR1

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

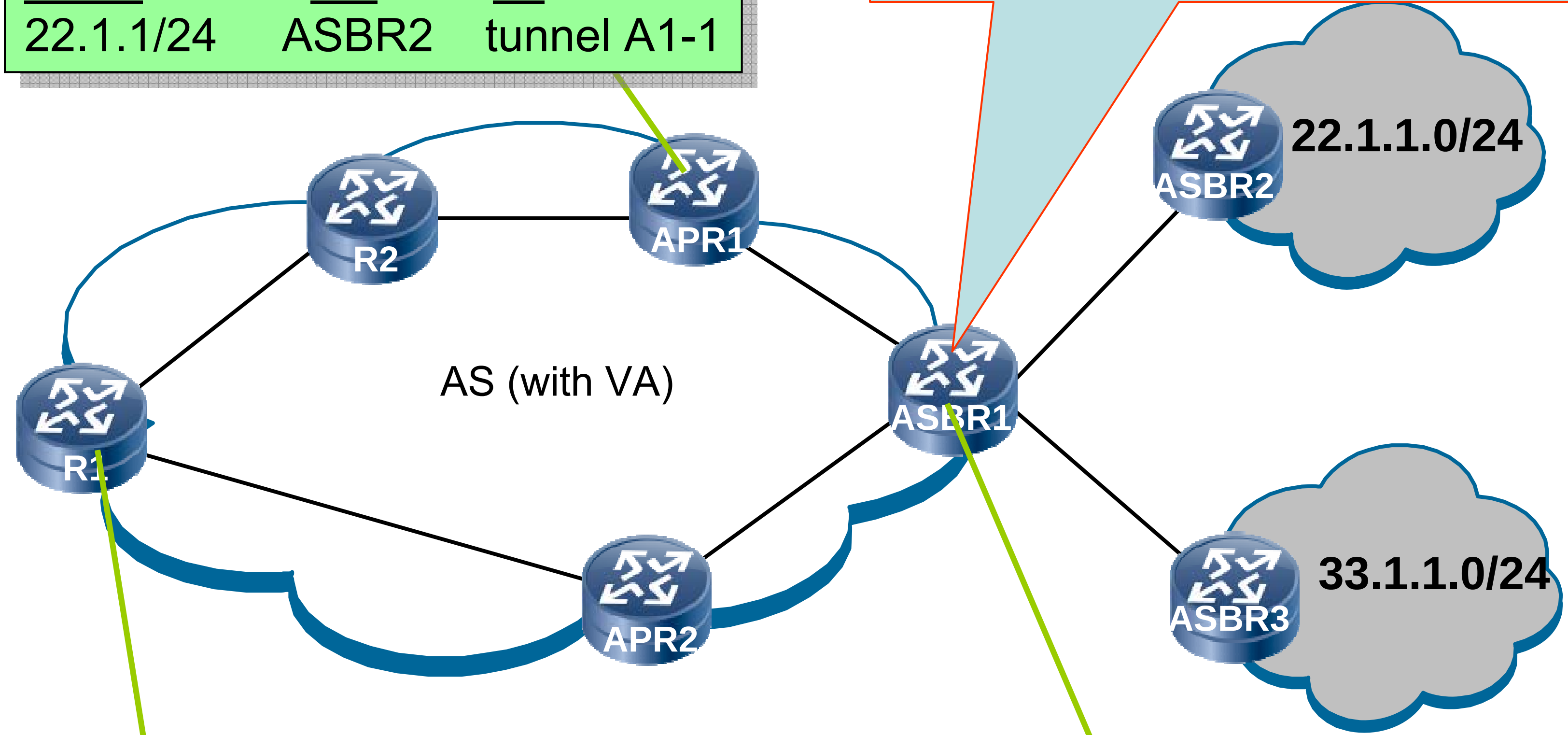


Prefix	NH	Inf
22/8	APR1	tunnel APR1

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

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

FIB-install rules

- APRs must FIB-install all sub-prefixes within VP
- All routers must FIB-install all Virtual Prefixes (VP)
- All other prefixes <u>may</u> 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?

- WG is considering automatic config of the VP-list and high-volume sub-prefixes
- Involves tagging routes with appropriate community attribute
- Stay tuned….

To summarize

- New IETF GROW WG work item for FIB suppression
  - Allows you to extend the lifetime of older routers indefinitely
- Still early in the standards process
  - You can influence the design
- If this sounds useful, please talk to your favorite vendor

Thanks!

- IETF Drafts
  - – draft-ietf-grow-va-00
  - – draft-ietf-grow-va-gre-00
  - – draft-ietf-grow-va-mpls-00
  - – draft-ietf-grow-va-perf-00
- Other
  - – “Making Routers Last Longer with ViAggre”, NSDI 2009

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