Demystifying pros & cons of large scale BGP RR deployments

Rohit Bothra
Brocade Communications

Objectives

- Large scale iBGP deployments on ISP Core
  - Full-mesh iBGP
  - Route reflectors
  - Confederations

- Advantages with RR deployment

- The Problems with Route Reflectors
  - Routing anomalies caused by route reflectors
  - Understanding BGP convergence & its impact with RR

- Pros and cons of proposed solutions
- Solution available from different vendors.

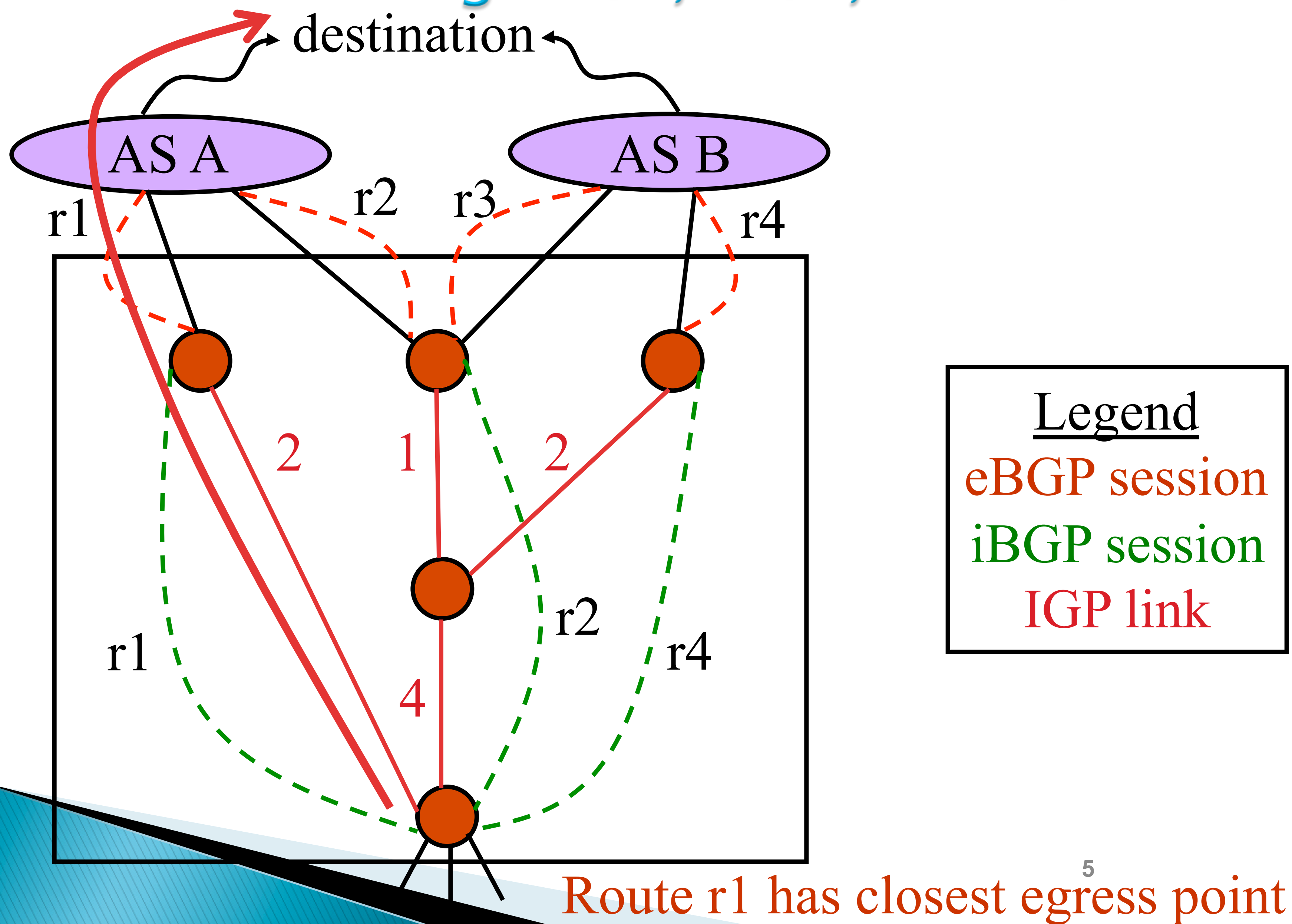
- Summary

Introduction

- BGP is the de-facto protocol of choice when it comes to Inter Domain Routing.
- Large ISPs BGP deployment involves many complexities at different levels.
- Route reflection was added to the routing architecture to solve the problem of scaling BGP.
- Despite the wide adoption of RR, a systematic evaluation and analysis on the impact of route reflection is not discussed widely, which will be helpful in:
  - Understanding of the protocol performance and enhancements
  - More realistic deployments.
  - New BGP solutions available
- We will discuss more on these lines today!

BGP Primer

Routers Running eBGP, iBGP, and IGP

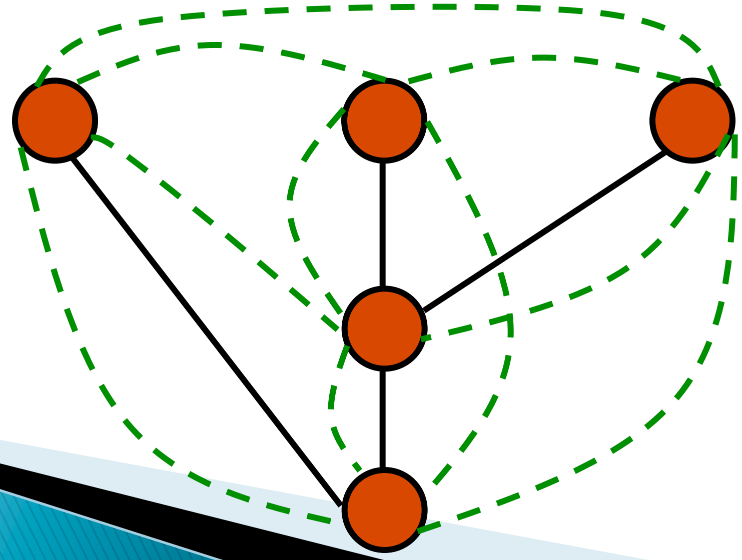


Roles of eBGP, iBGP, and IGP

- eBGP: External BGP
  - Learn routes from neighboring ASes
  - Advertise routes to neighboring ASes
- iBGP: Internal BGP
  - Disseminate BGP information within the AS
- IGP: Interior Gateway Protocol
  - Compute shortest paths between routers in AS
  - Identify closest egress point in BGP path selection

Full Mesh iBGP Configuration

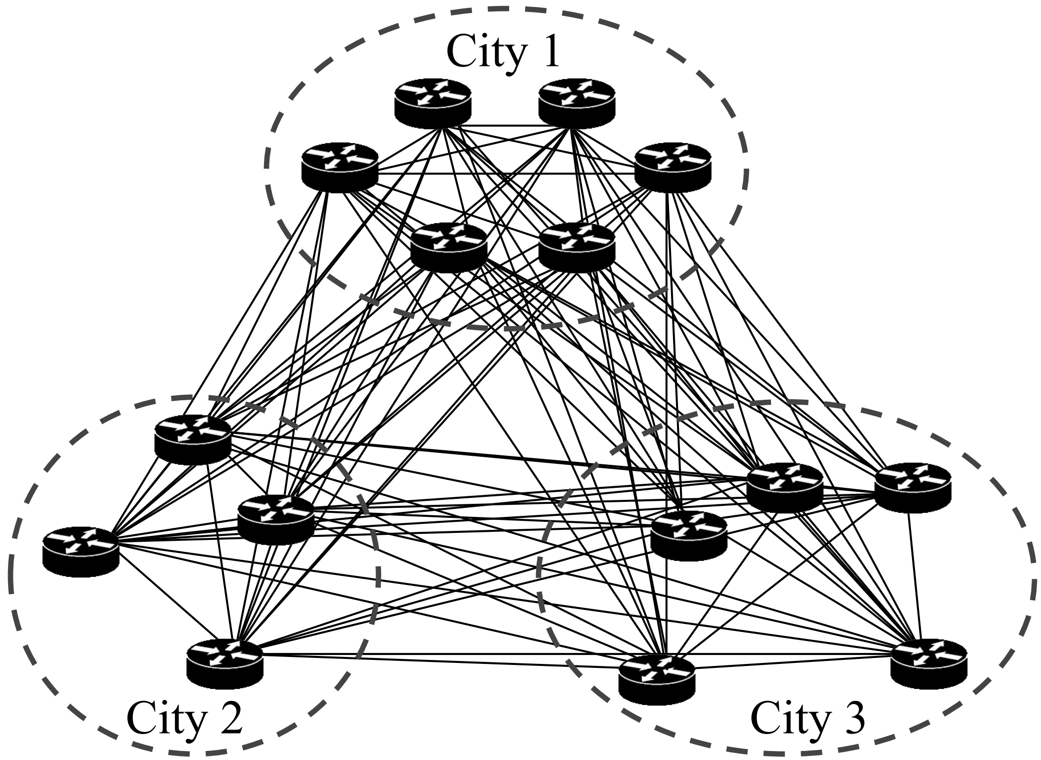
- Internal BGP session
  - Forward best BGP route to a neighbor
  - Do not send from one iBGP neighbor to another
- Full-mesh configuration
  - iBGP session between each pair of routers
  - Ensures complete visibility of BGP routes



Why Do Point-to-Point Internal BGP?

- Reusing the BGP protocol
  - iBGP is really just BGP
  - … except you don't add an AS to the AS path
  - … or export routes between iBGP neighbors
- No need to create a second protocol
  - Another protocol would add complexity
- And, full-mesh is workable for many networks
  - Well, until they get too big…

Full-mesh i-BGP does not scale

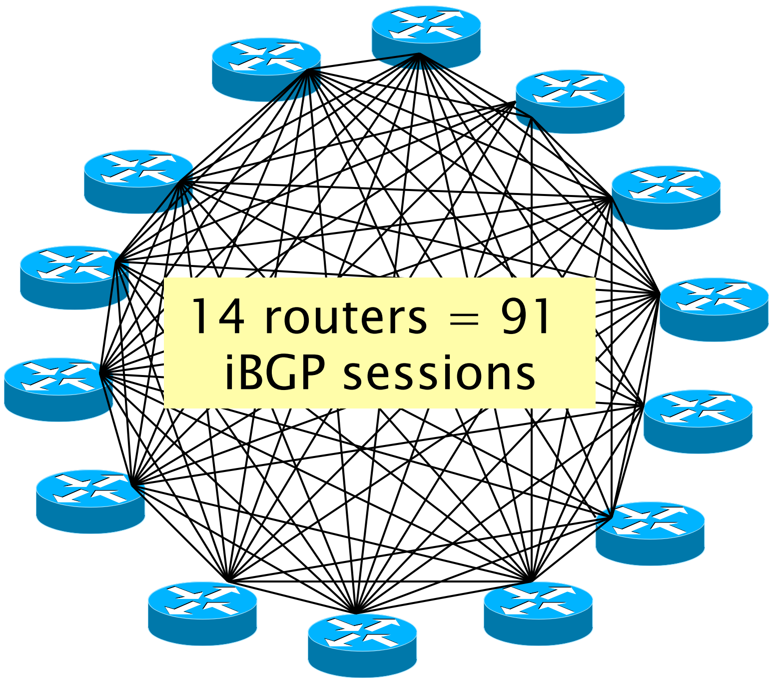


- Large ISPs have hundreds or even more than a thousand routers internally
- Full mesh leads to a high cost in provisioning
  - Adding or removing a router requires reconfigurations of all other routers

Scaling iBGP mesh

Avoid ½n(n−1) iBGP mesh

n=1000 ⇒ nearly half a million ibgp sessions!



- Two solutions
  - Route reflector – simpler to deploy and run
  - Confederation – more complex, has corner case advantages

Confederations: Benefits

- Solves iBGP mesh problem
- Packet forwarding not affected
- Can be used with route reflectors
- Policies could be applied to route traffic between sub-AS's

Scalability Limits of Full Mesh on the Routers

- Number of iBGP sessions
  - TCP connection to every other router
- Bandwidth for update messages
  - Every BGP update sent to every other router
- Storage for the BGP routing table
  - Storing many BGP routes per destination prefix
- Configuration changes when adding a router
  - Configuring iBGP session on every other router

BGP Confederations

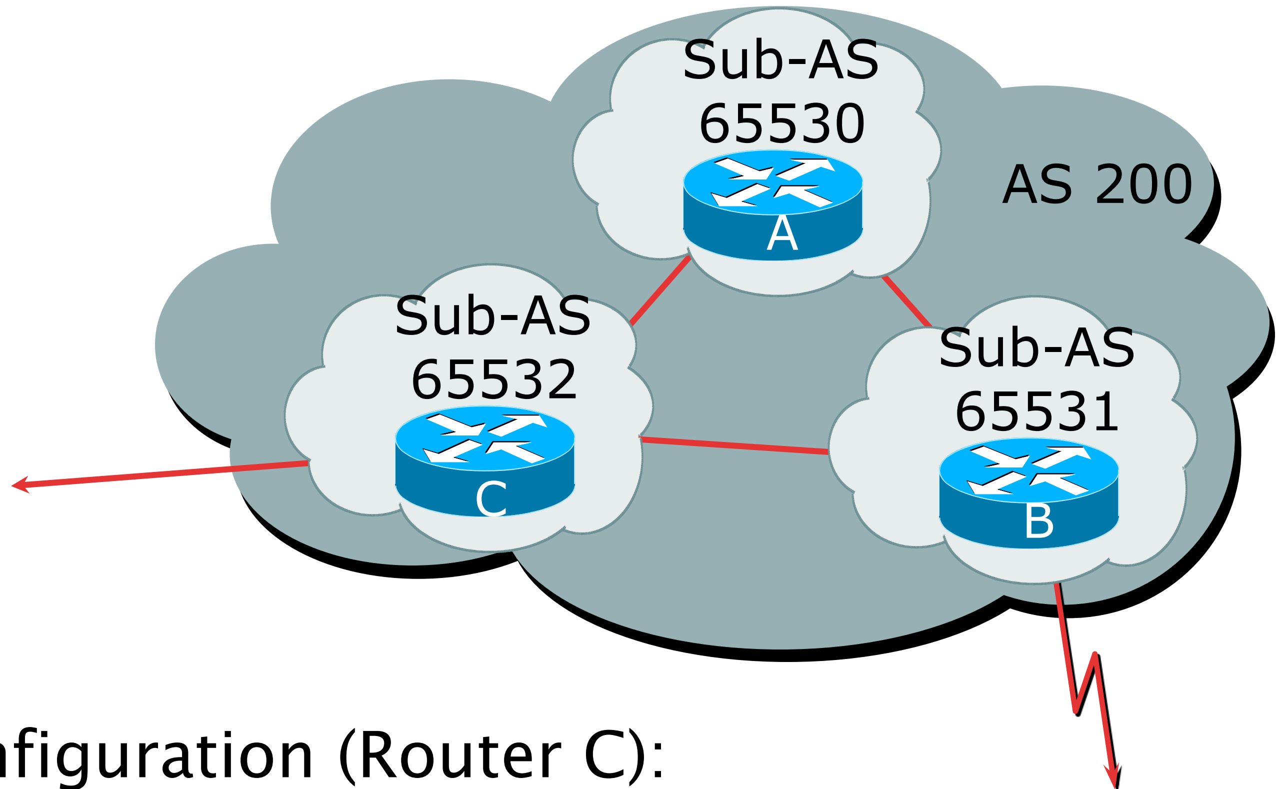
Confederations

- Divide the AS into sub-AS
  - eBGP between sub-AS, but some iBGP information is kept
    - Preserve NEXT_HOP across the sub-AS (IGP carries this information)
    - Preserve LOCAL_PREF and MED
- Usually a single IGP
- Described in RFC5065

Confederations

- Visible to outside world as single AS – "Confederation Identifier"
  - Each sub-AS uses a number from the private space (64512–65534)
- iBGP speakers in sub-AS are fully meshed
  - The total number of neighbors is reduced by limiting the full mesh requirement to only the peers in the sub-AS

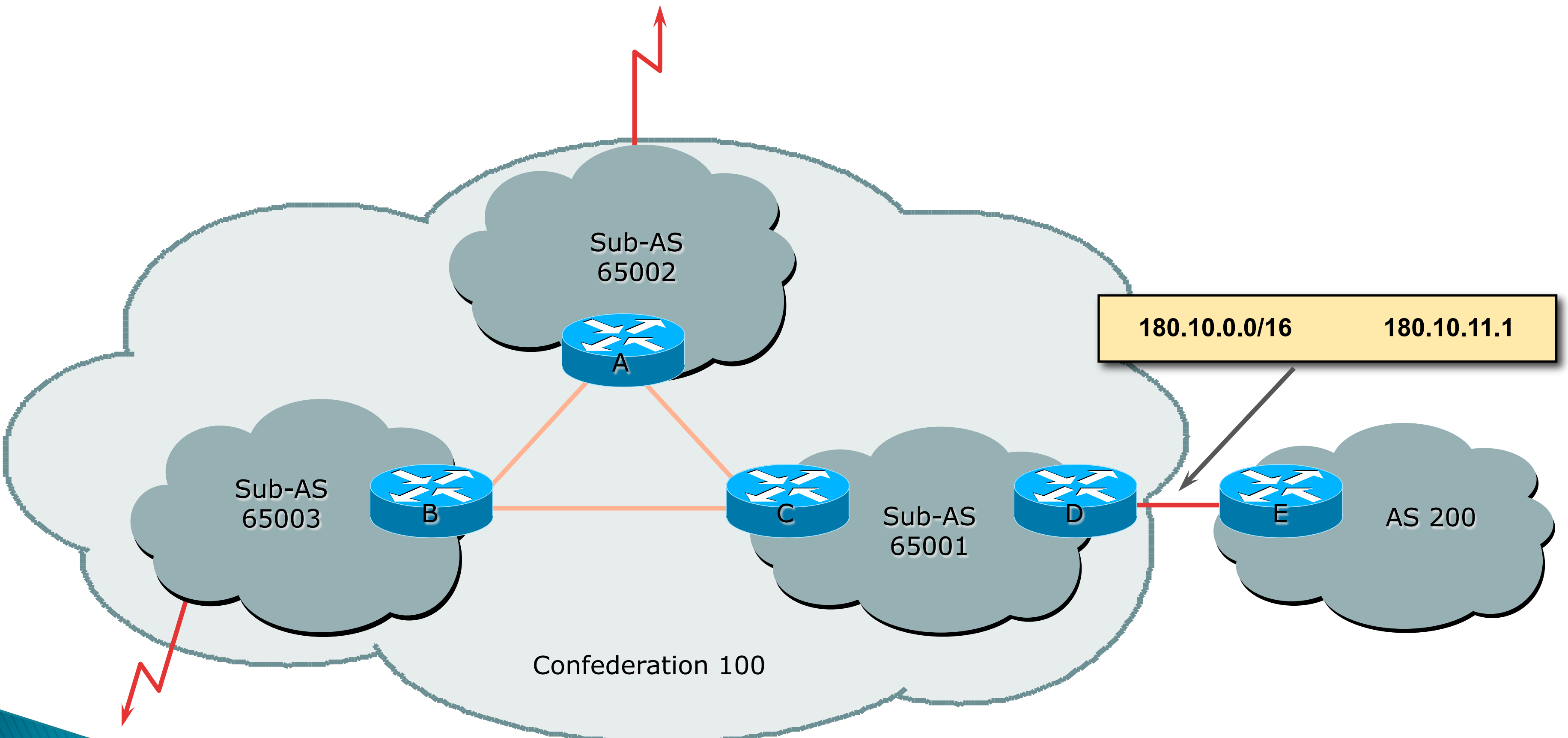
Confederations



- Configuration (Router C):

```
set protocols bgp 200 parameters confederation identifier 200
set protocols bgp 200 parameters confederation peers 65530 65531
set protocols bgp 200 neighbor 1.1.1.1 remote-as 65530
set protocols bgp 200 neighbor 2.2.2.2 remote-as 65531
```

Confederations: Next-hop



Confederation: Principle

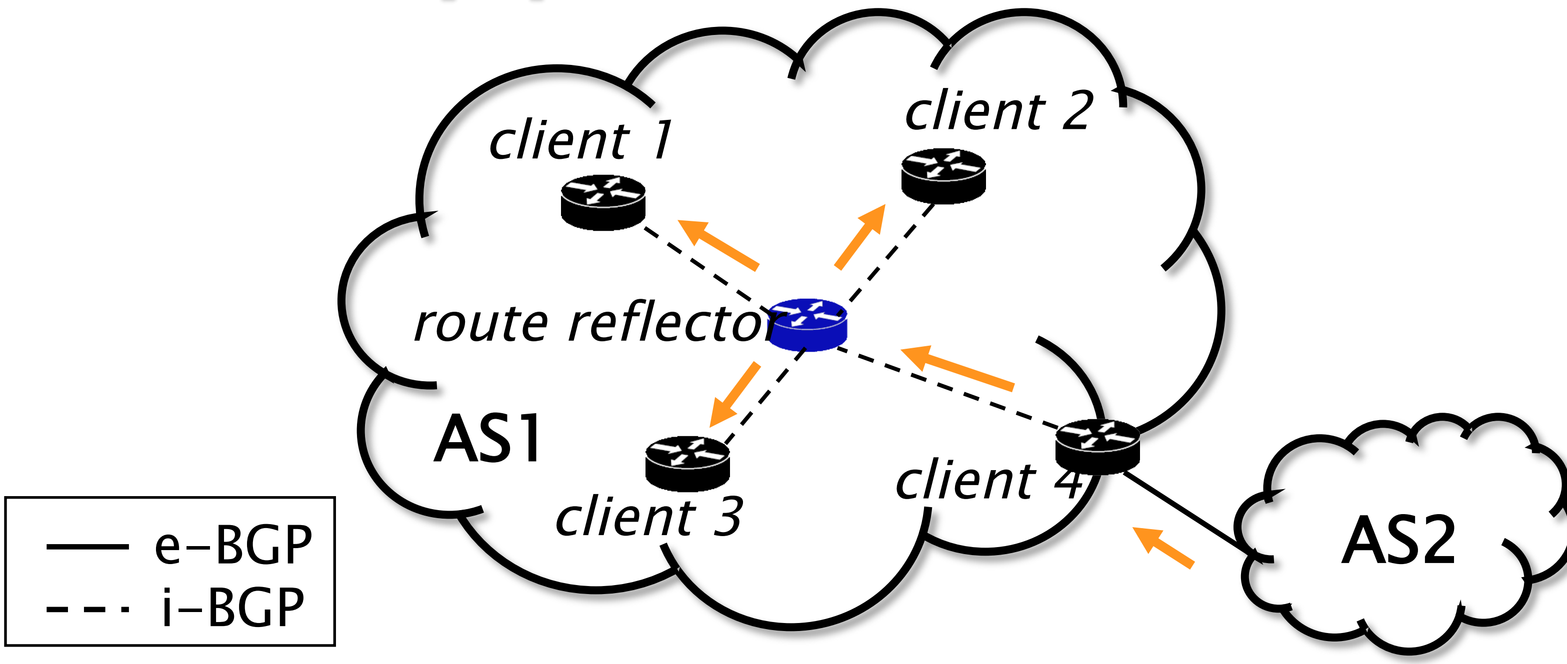
- Local preference and MED influence path selection
- Preserve local preference and MED across sub-AS boundary
- Sub-AS eBGP path administrative distance

Confederations: Caveats

- Minimal number of sub-AS
- Sub-AS hierarchy
- Minimal inter-connectivity between sub-AS's
- Path diversity
- Difficult migration
  - BGP reconfigured into sub-AS
  - must be applied across the network

BGP Route Reflectors

Route reflection solves scalability problem



Total number of i-BGP routers = **5 = N**

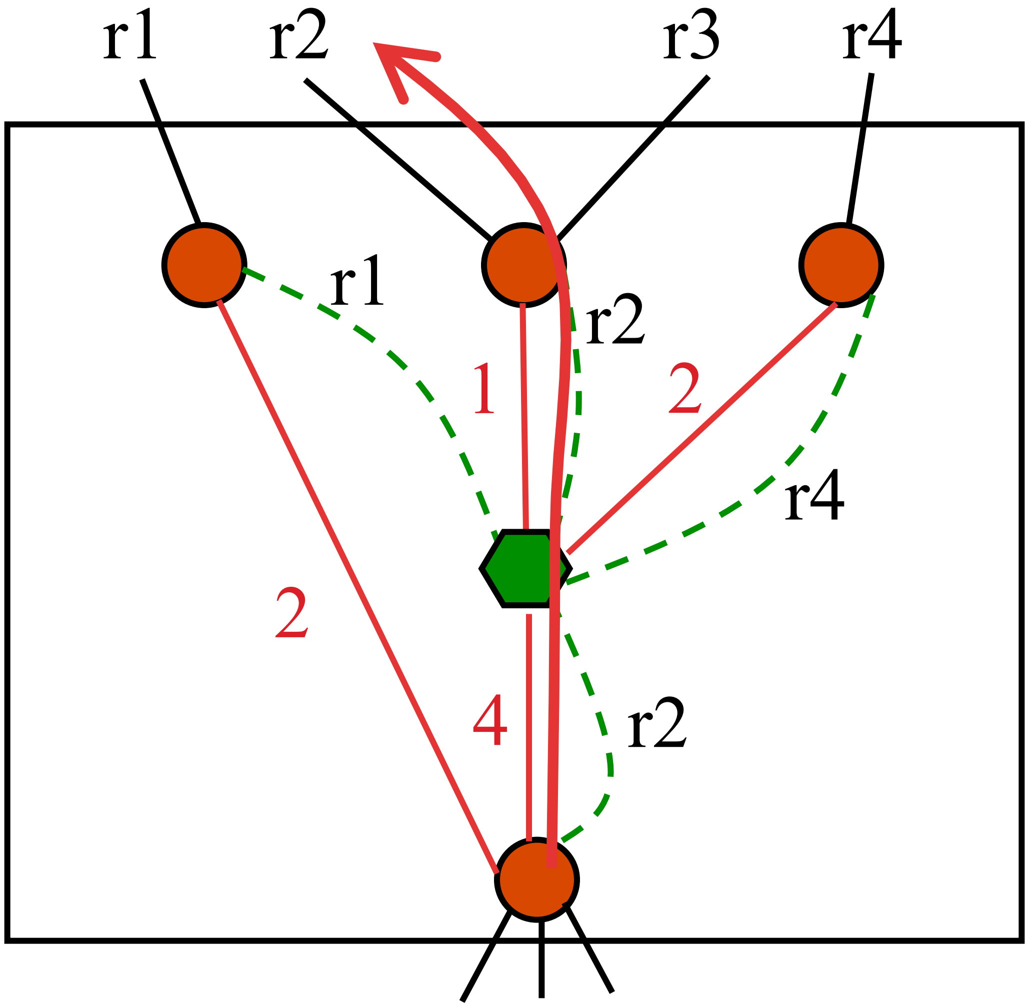
Total number of sessions = **4**

Number of *additional* sessions for an additional i-BGP = 1

Route Reflectors

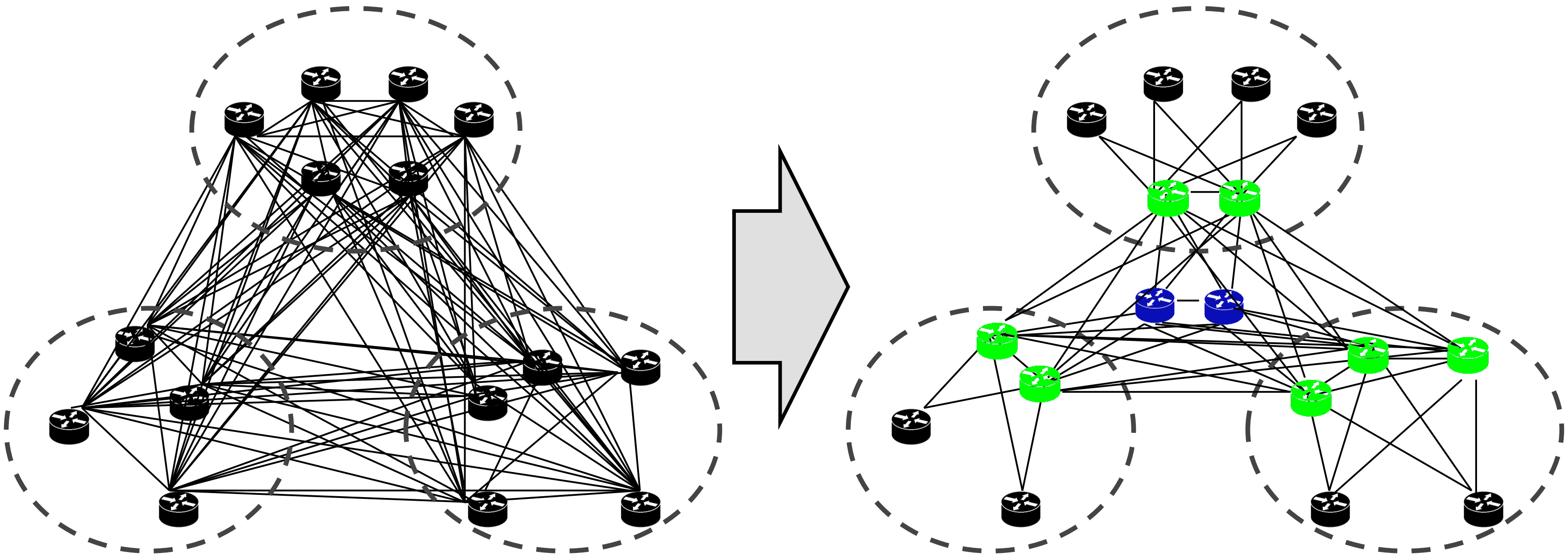
- Relax the iBGP propagation rule
  - Allow sending updates between iBGP neighbors
- Route reflector
  - Receives iBGP updates from neighbors
  - Send a single BGP route to the clients
- Very much like provider, peer, and customer
  - To client: send all BGP routes
  - To peer route reflector: send client-learned routes
  - To route reflector: send all client-learned routes

Example: Single Route Reflector



Router only learns about r2

Large ISP revisited with hierarchical RR



- Route reflection substantially reduces the total number of sessions
- Route reflection can be deployed hierarchically to reduce even more

The Advantages with Route Reflectors

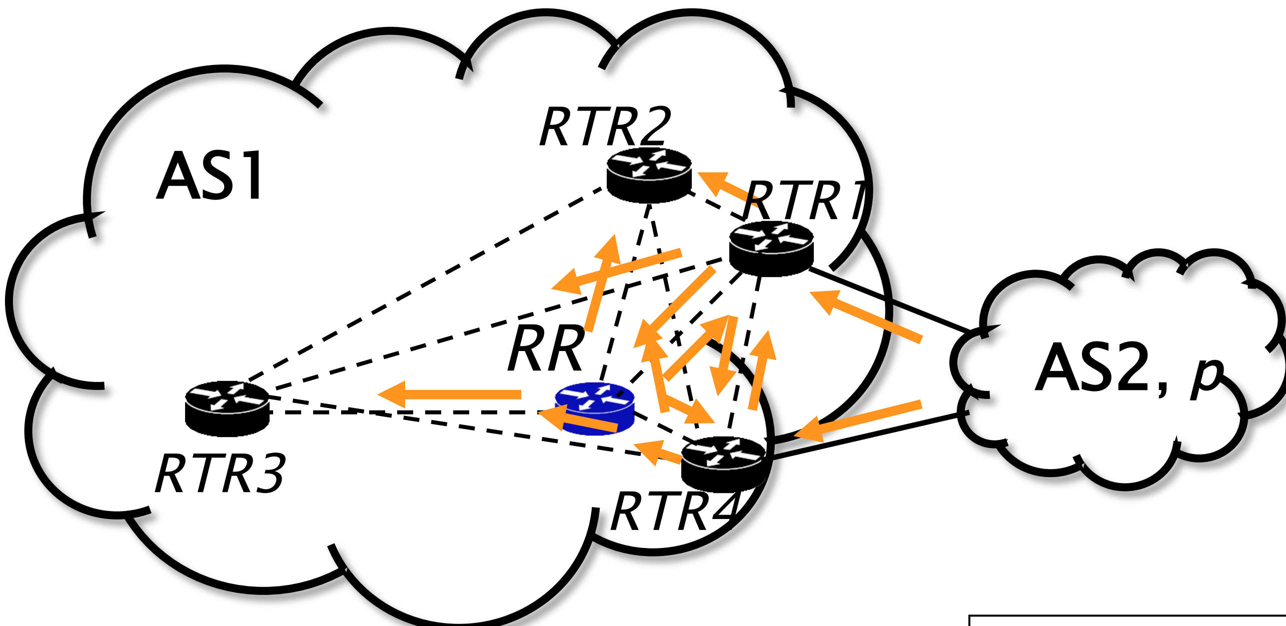
- Advantage: scalability
  - Fewer iBGP sessions
  - Lower bandwidth for update messages
  - Smaller BGP routing tables
  - Lower configuration overhead
  - Lower cost
  - Lower number of deployment nodes

BGP Route Reflector Disadvantages

The disadvantages with RRs

- The story is going to take a U turn
  - Routing performance
    - **Path diversity**
    - **Convergence**
    - **Others**
      - Robustness to failures
      - Internal update explosion
      - Optimal route selection
  - Routing correctness
    - Data forwarding loop
    - Route oscillations

Path diversity reduction due to route reflection



ALL
p: NH = RTR1, ASPATH = AS2
p: NH = RTR4, ASPATH = AS2

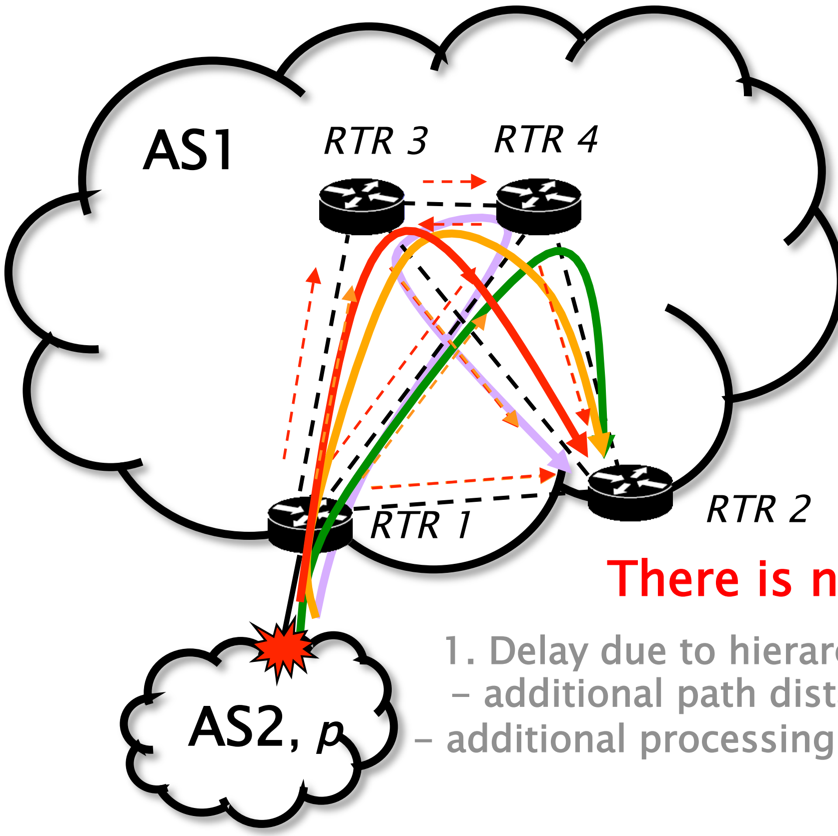
RTR1, RR
p: NH = RTR1, ASPATH = AS2
p: NH = RTR4, ASPATH = AS2

OTHERS
p: NH = RTR4, ASPATH = AS2

Paths can be hidden due to path preference

- BGP path attribute values used by a BGP router in BGP best path selection
  - First 4 are independent from the i-BGP topological location of the given router
    - LOCAL_PREF
    - AS_PATH length
    - ORIGIN
    - MED
  - The rest 3 attribute values change depending on the i-BGP topological location of the given router
    - Prefer e-BGP over i-BGP
    - IGP cost
    - Router ID

Increased convergence delay in i-BGP RR



Update path

1. ~~RR2->RTR1~~
2. ~~RR1->RTR1~~
3. ~~RR2->RR1->RTR1~~
4. ~~RR1->RR2->RTR1~~
5. Not reachable

There is no path to prefix *p*!

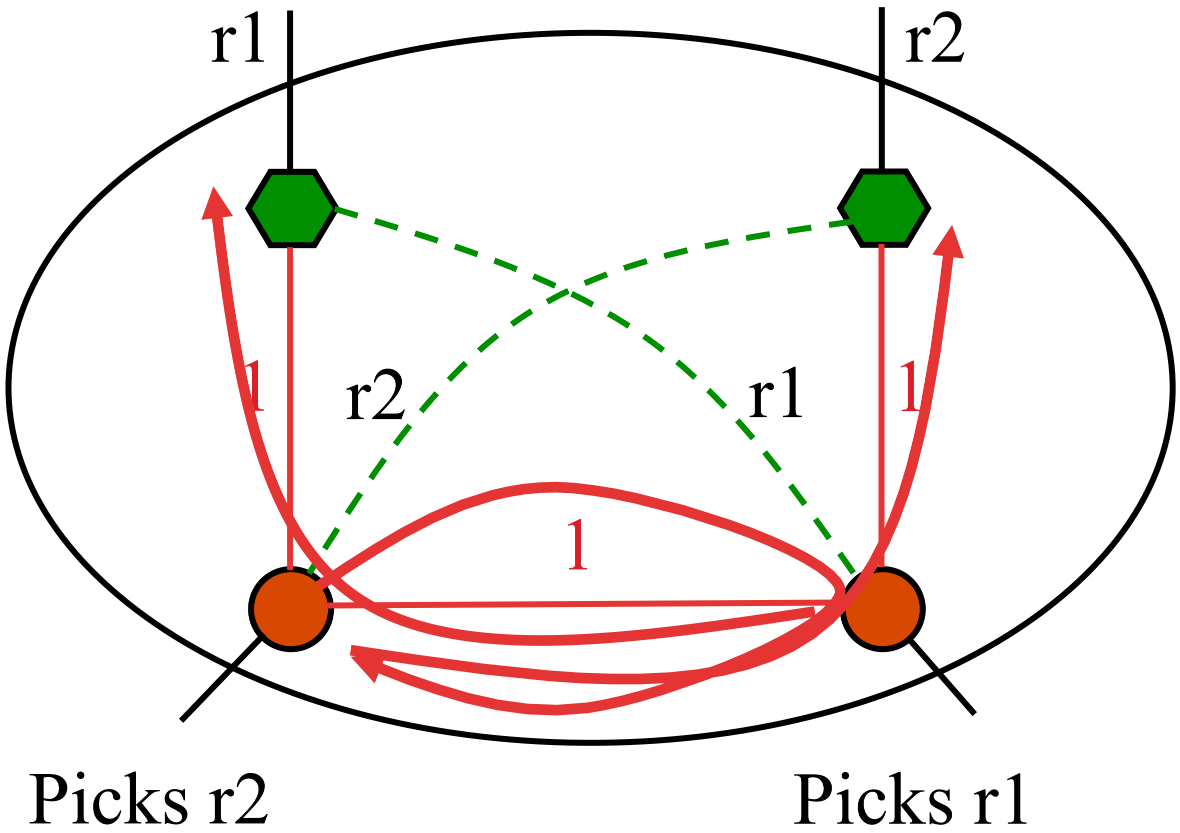
1. Delay due to hierarchy
   - additional path distance
   - additional processing delays
2. Delay due to route reflector redundancy
   - increased # of control paths

Delay caused by RRs
Estimating the additional delay caused by route reflection

- Additional delays due to **route reflector redundancy**
  - Identify the *superfluous updates* generated purely due to route reflector redundancy
  - What is the additional convergence time solely contributed by these updates?

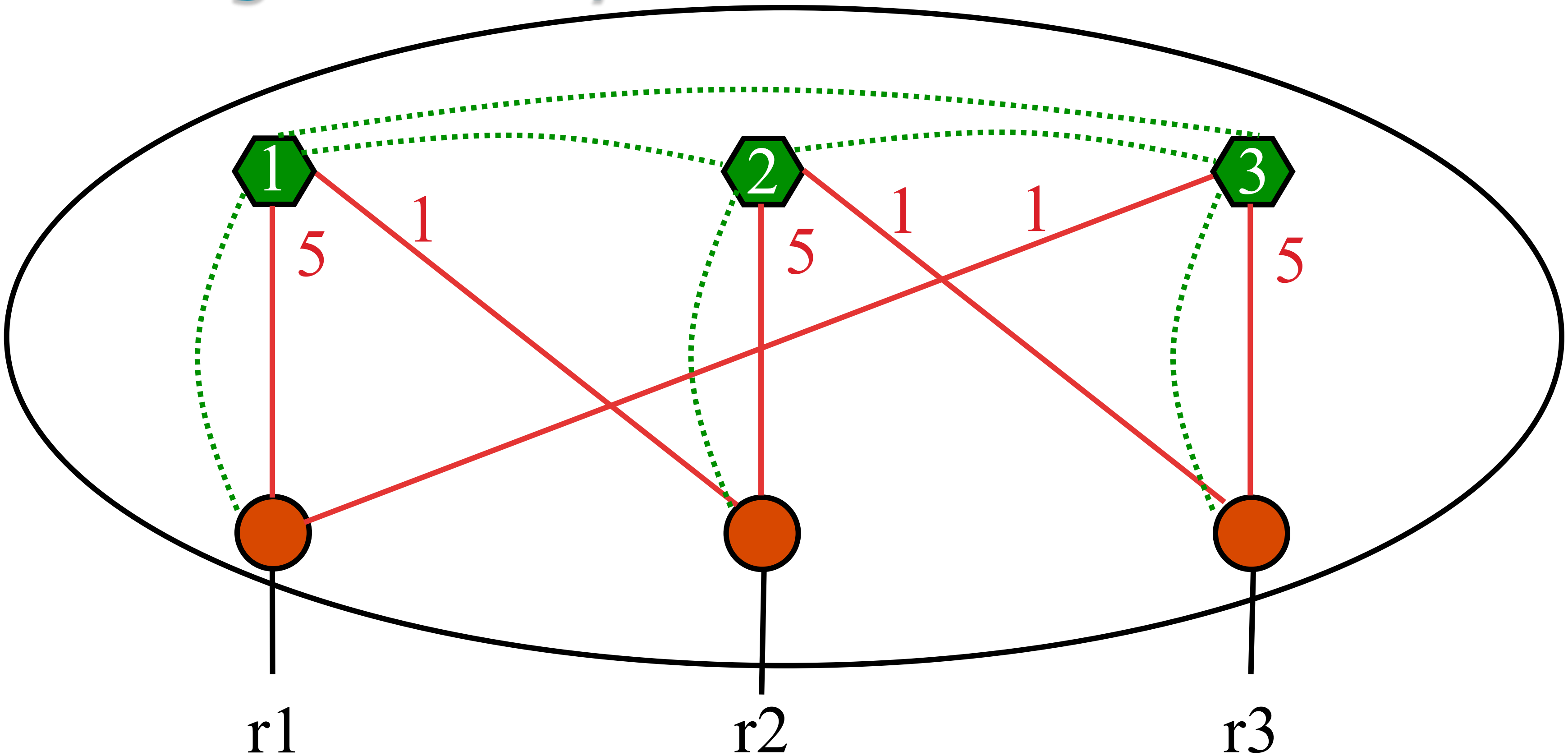
- Additional delays due to **hierarchy**
  - Compare the direct and RR paths between all monitors in the backbone routing infrastructure inside ISP_{RR}

Routing Anomaly: Forwarding Loop



Packet deflected toward other egress point, causing a loop

Routing Anomaly: Protocol Oscillation



RR1 prefers r2 over r1
RR2 prefers r3 over r2
RR3 prefers r1 over r3

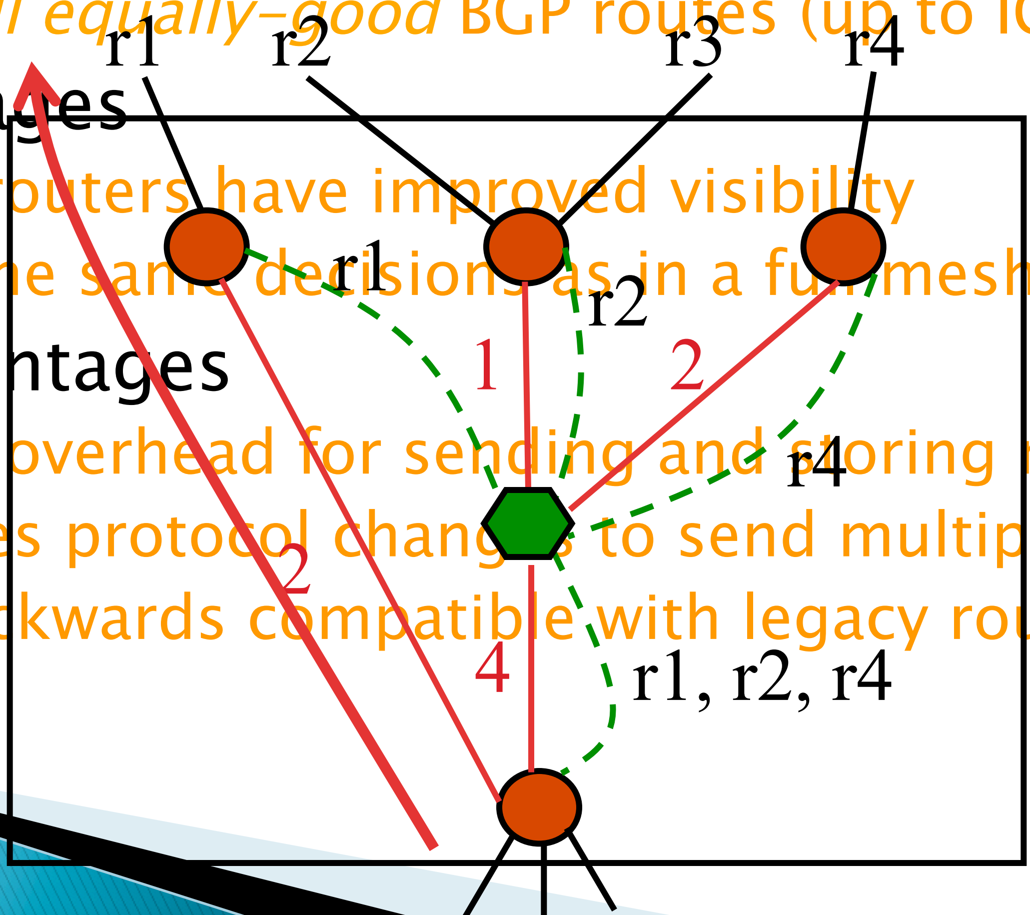
Solutions

Avoiding Routing Anomalies

- Reduce impact of route reflectors
  - Ensure route reflector is close to its clients
  - … so the RR makes consistent decisions
- Sufficient conditions for ensuring consistency
  - RR preferring routes through clients over "peers"
  - BGP messages should traverse same path as data
- Forces a high degree of replication
  - Many route reflectors in the network
  - E.g., a route reflector per PoP for *correctness*
  - E.g. have a second RR per PoP for *reliability*

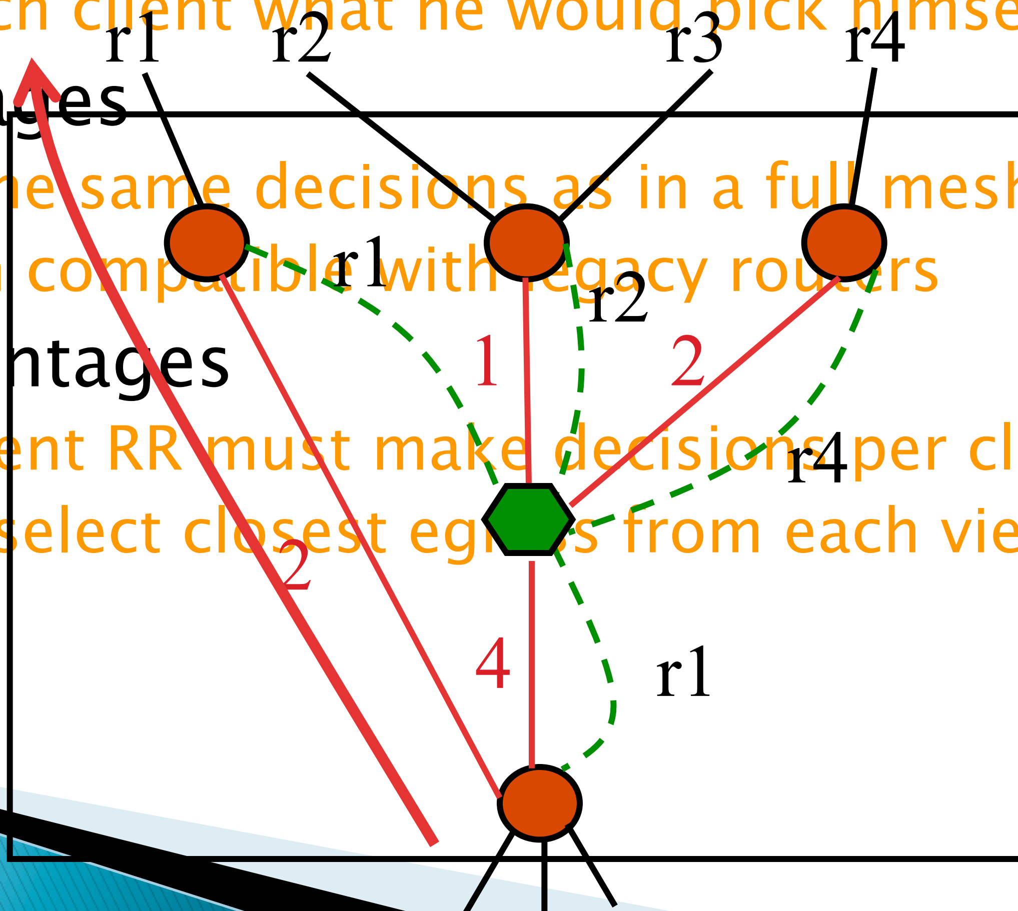
Possible Solution: Disseminating More Routes

- Make route reflectors more verbose
  - Send *all* BGP routes to clients, not just best route
  - Send *all equally-good* BGP routes (up to IGP cost)
- Advantages
  - Client routers have improved visibility
  - Make the same decisions as in a full mesh
- Disadvantages
  - Higher overhead for sending and storing routes
  - Requires protocol changes to send multiple routes
  - Not backwards compatible with legacy routers



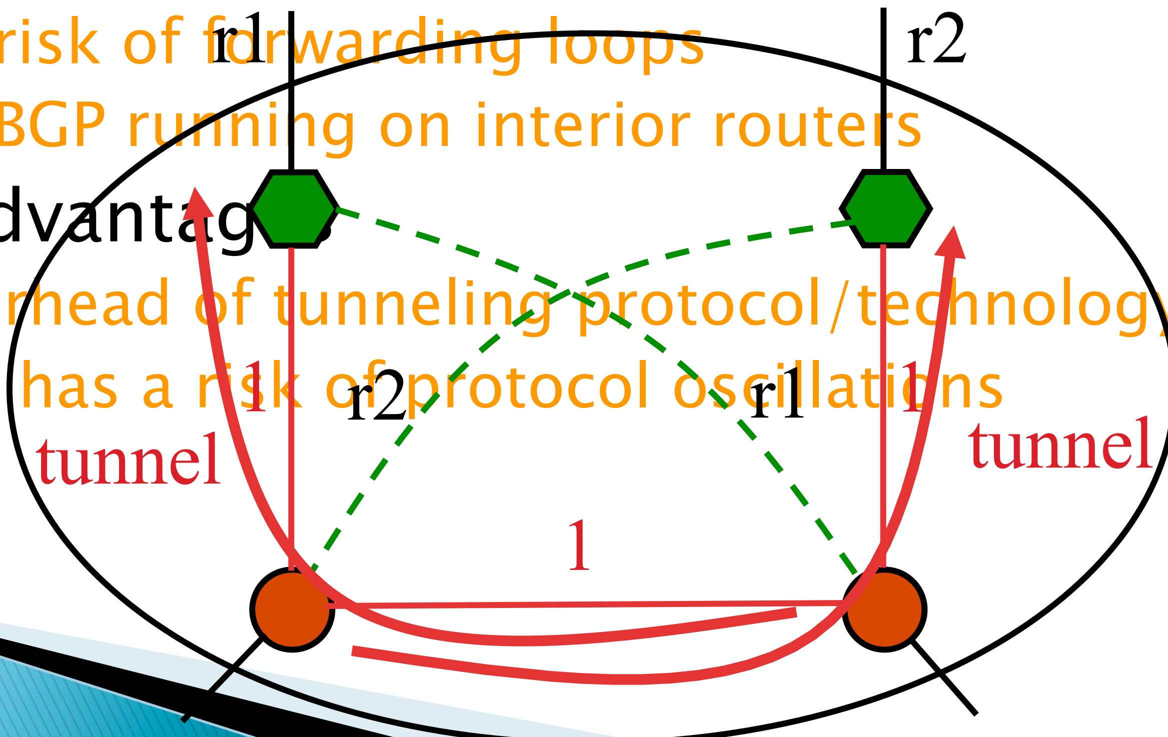
Possible Solution: Customized Dissemination

- Make route reflector more intelligent
  - Send customized BGP route to each client
  - Tell each client what he would pick himself
- Advantages
  - Make the same decisions as in a full mesh
  - Remain compatible with legacy routers
- Disadvantages
  - Intelligent RR must make decisions per client
  - … and select closest egress from each viewpoint



Possible Solution: Tunnel Between Edge Routers

- Tunneling through the core
  - Ingress router selects ingress point
  - Other routers blindly forward to the egress
- Advantages
  - No risk of forwarding loops
  - No BGP running on interior routers
- Disadvantages
  - Overhead of tunneling protocol/technology
  - Still has a risk of protocol oscillations



State-of-the-Art of BGP Distribution in an AS

- When full-mesh doesn't scale
  - Hierarchical route-reflector configuration
    - One or two route reflectors per PoP
  - Some networks use "confederations" (mini ASes)
- Recent ideas
  - Sufficient conditions to avoid anomalies
  - Enhanced RRs sending multiple or custom routes
  - Flooding/multicast of BGP updates
  - Tunneling to avoid packet deflections
- Open questions
  - Are the sufficient conditions too restrictive?
  - Good comparison of the various approaches

Vendor solution considerations

Vendor Solutions

Solutions	Description	Advantages
BGP PIC	Prefix independent convergence for CORE link failures as well as Edge node failures	Fast Convergence
BGP Add path	Multiple paths ready to use in dataplane	Fast Convergence, ECMP
BGP virtual RR	optimize/virtualize BGP route-reflector functions due to integration of more BGP services	Scalability & Performance
BGP multipath	Helps in BGP diversity	Avoid Route Oscillation, ECMP
BGP Best External	Provides support for advertisement of Best-External path to the iBGP/RR peers when a locally selected bestpath is from an internal peer	Back up sends its own external path
VPN unique RD	PE can reflect same prefix with unique RDs	Recommended method for MPLS VPN
BGP optimal route reflection	An RR selects best path based on IGP metric	Solves Hot potato routing for VRR
BGP multiple cluster IDs	allows an iBGP neighbor (usually a route reflector) to have multiple cluster IDs: a global cluster ID and additional cluster IDs that are assigned to clients.	Solves Route oscillation

Summary

- Networks are getting bigger, so plan your iBGP scaling with all pros & cons in mind.
- Techniques for scaling the routing design needs to be considered very carefully.
- Define, quantify, and analyze i-BGP convergence before deployment.
- RR topology design may mitigate expected convergence numbers.
- There are many optimized solutions available from different vendors around RR
  - Choose as per your network requirements.

THANK YOU

BACKUP SLIDE

Brocade Vyatta NFV Use Case: vRouter

vRouter

- Deployment model:
  - Virtualize typical SP PE router for business VPN services
- Brocade Vyatta vRouter benefits:
  - High performance and scale, designed for virtualization
  - Advanced routing – BGP, OSPF, Multicast, etc
  - Stateful Firewall with NAT
  - MPLS/VPN, VRFs, QoS, etc
- Other NFV benefits:
  - Agility – Click of button provisioning of new customers
  - Flexibility – easy to scale out or repurpose
  - Lower cost – Lower CAPEX running VNF on COTS versus dedicated HW PEs;

    lower OPEX from automated provisioning and typically pay as you use