Interdomain routing with BGP4

Part 3/5

Olivier Bonaventure

Department of Computing Science and Engineering
Université catholique de Louvain (UCL)
Place Sainte-Barbe, 2, B-1348, Louvain-la-Neuve (Belgium)

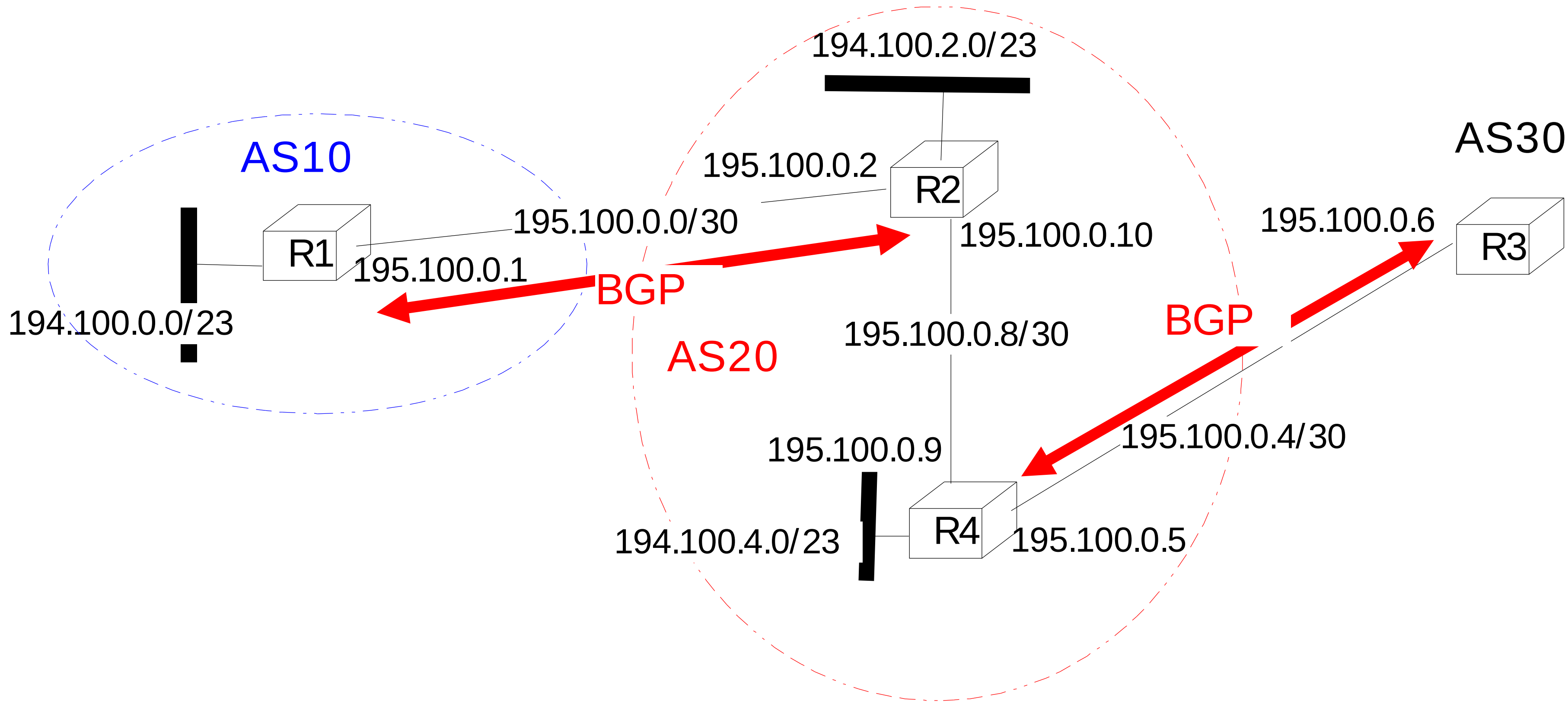
URL : *http://www.info.ucl.ac.be/people/OBO*

May 2003

Outline

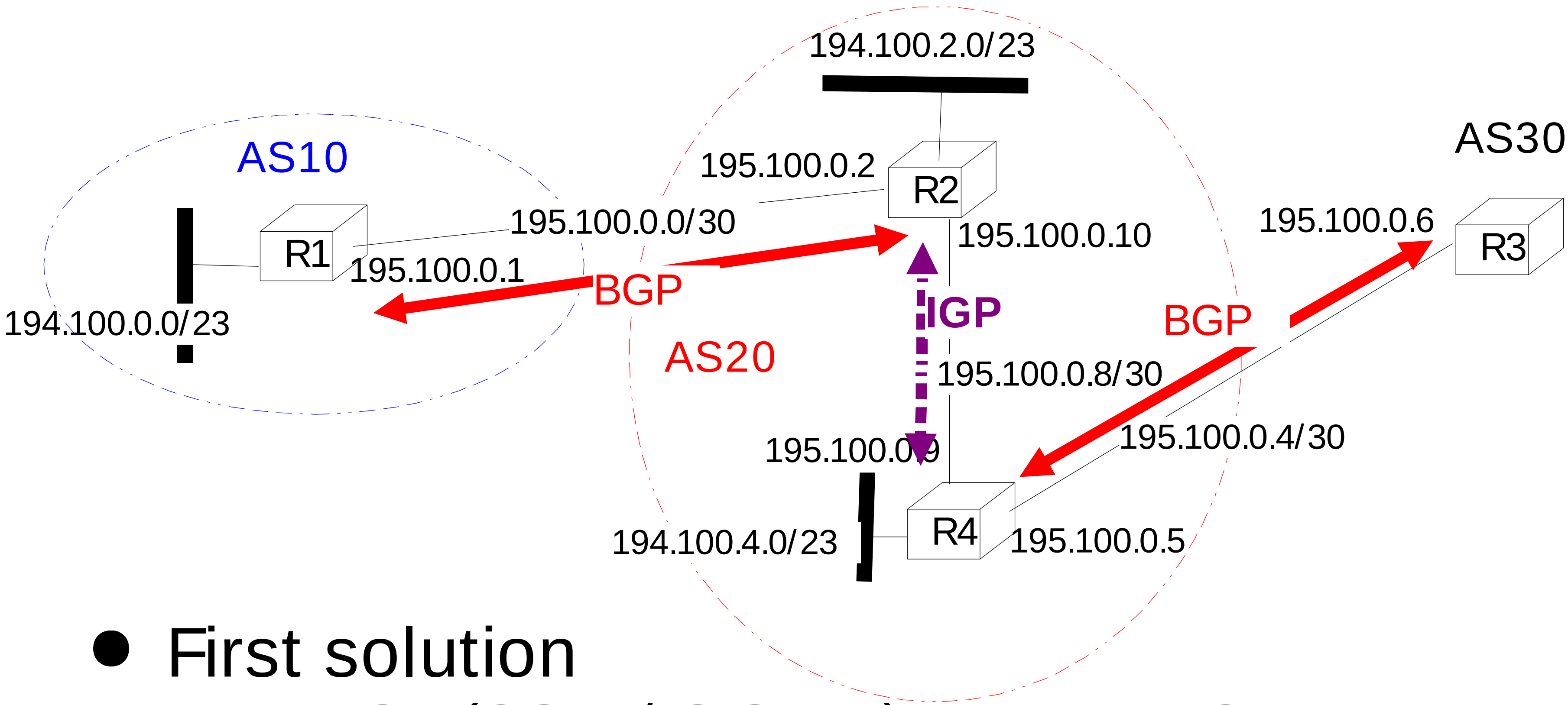
- Organization of the global Internet
- BGP basics
- BGP in large networks
  - The needs for iBGP
  - Confederations and Route Reflectors
  - Scalable routing policies
  - The dynamics of BGP
- Interdomain traffic engineering with BGP
- BGP-based Virtual Private Networks

BGP and IP
Second example



- Problem
  - How can R2 (resp. R4) advertise to R4 (resp. R2) the routes learned from AS10 (resp. AS30) ?

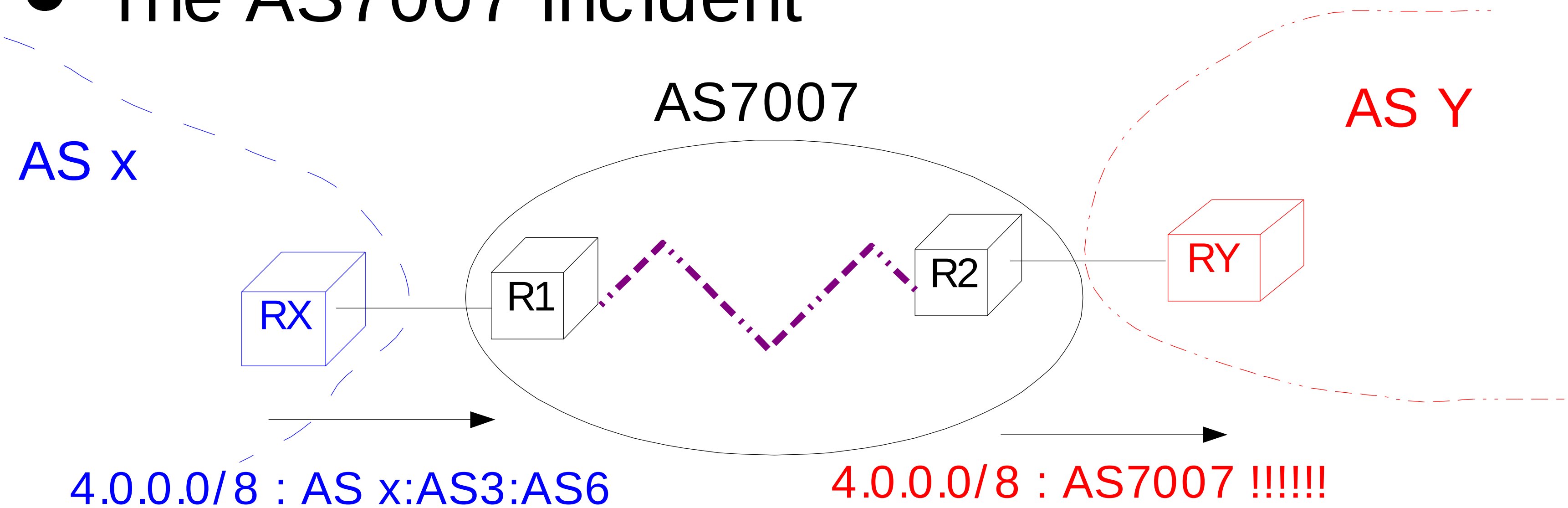
BGP and IP
Second example (2)



- First solution
  - Use IGP (OSPF/ ISIS,RIP) to carry BGP routes
- Drawbacks
  - IGP may not be able to support so many routes
  - IGP does not carry BGP attributes like ASPath !

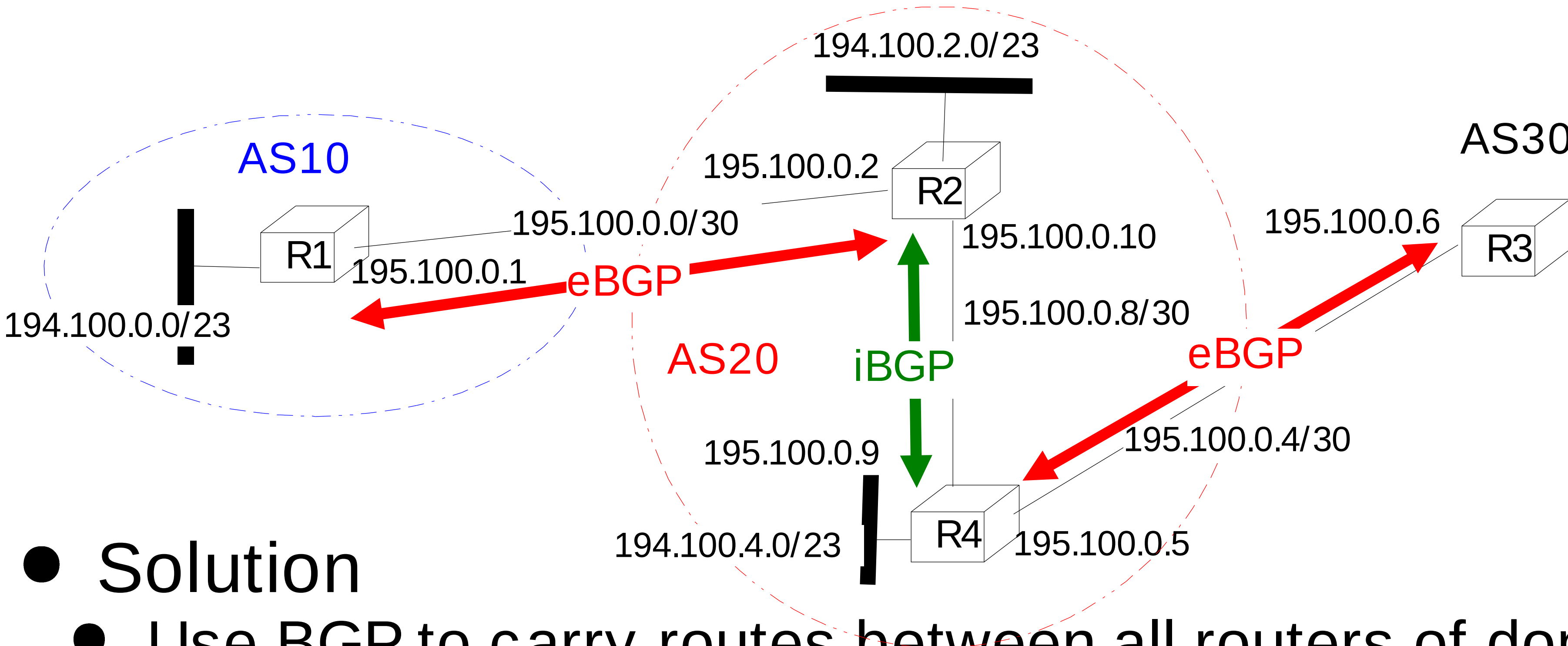
The AS7007 incident

- The AS7007 incident



- A single configuration error in two routers
  - All routes learned from ASX on R1 were redistributed to R2 via IGP and R2 announced them to ASY
  - Consequence
    - AS7007 advertised routes that almost all IP addresses were belonging to AS7007
    - These routes were shorter than the real routes ...
  - Two hours of disruption for large parts of the Internet !

iBGP and eBGP

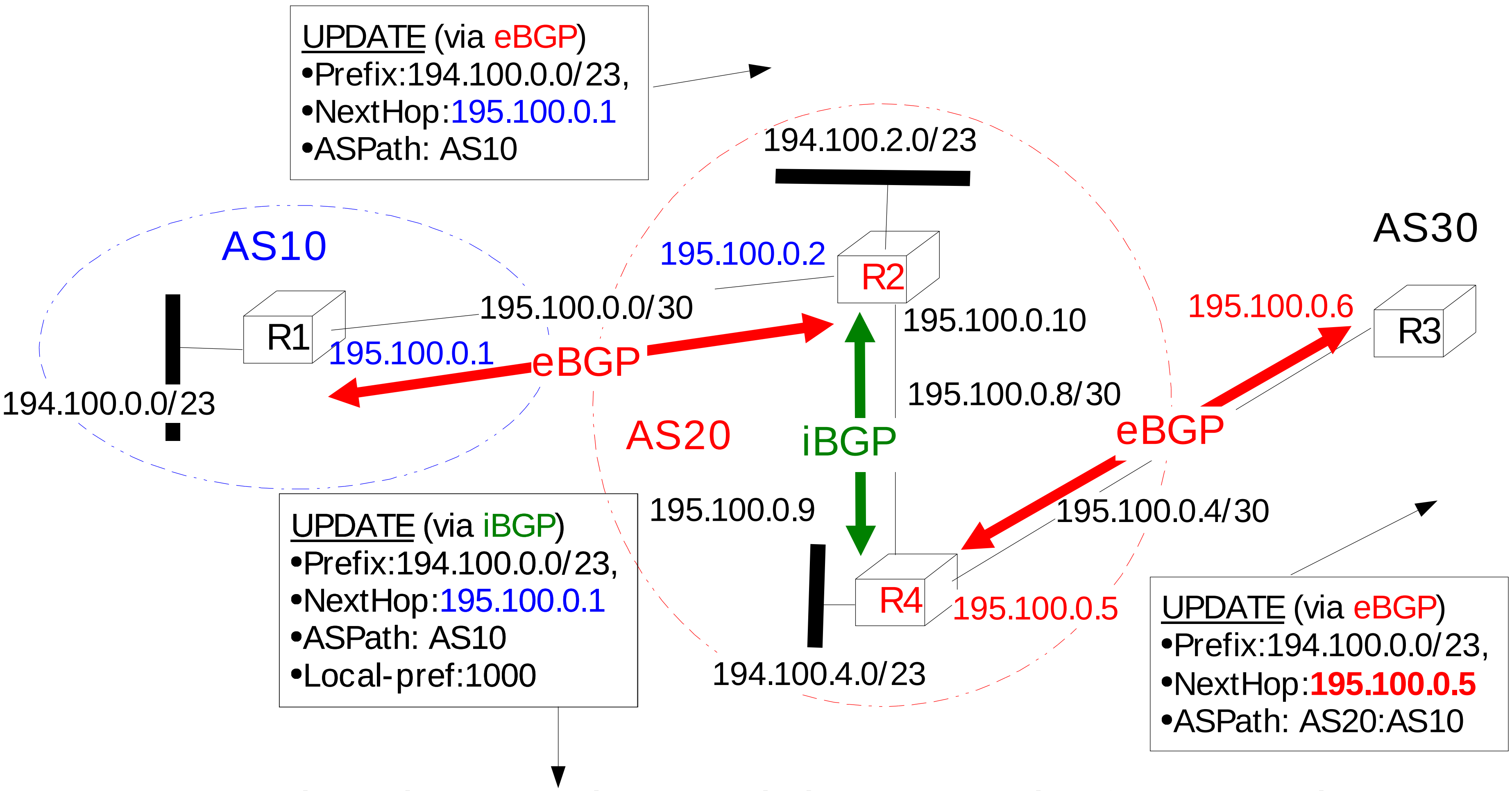


- Solution
  - Use BGP to carry routes between all routers of domain
    - Two different types of BGP sessions
    - eBGP between routers belonging to different ASes
    - iBGP between each pair of routers belonging to the same AS
      - Each BGP router inside ASx maintains an iBGP session with all other BGP routers of ASx (full iBGP mesh)
      - Note that the iBGP sessions do not necessarily follow physical topology

iBGP versus eBGP

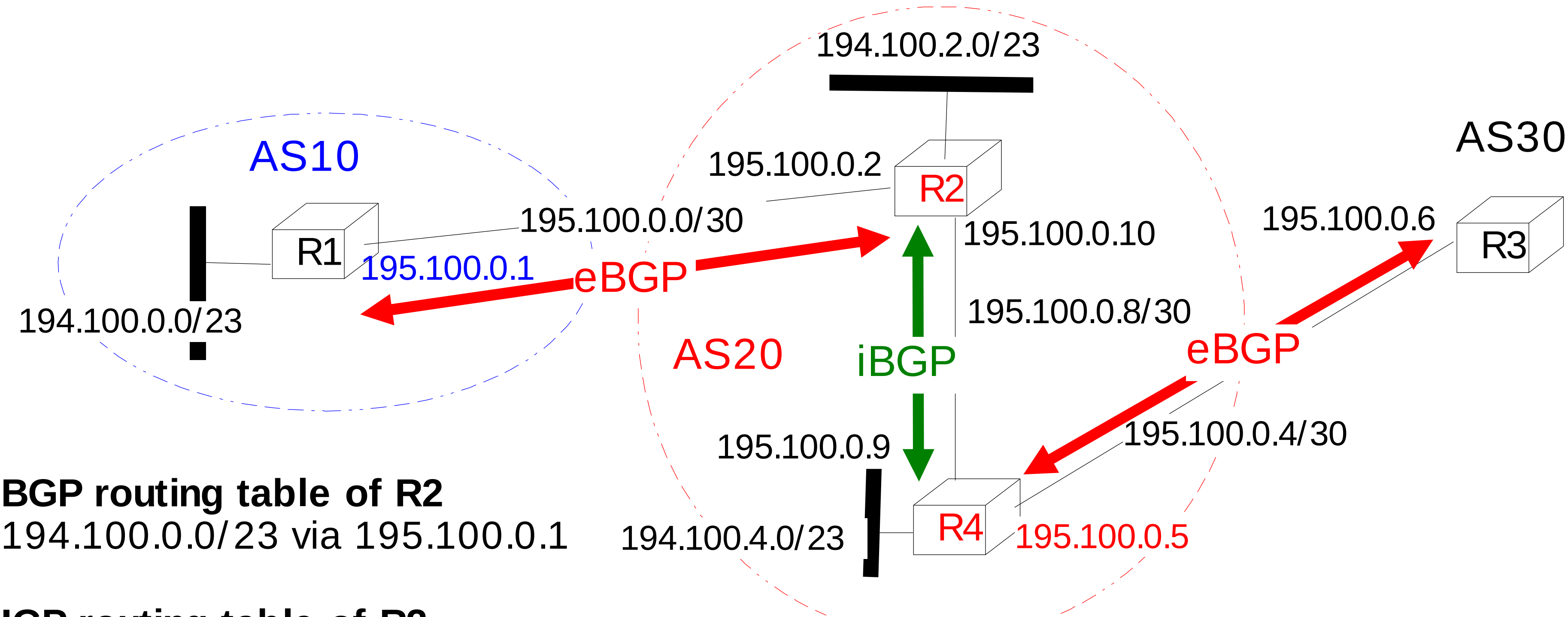
- Differences between iBGP and eBGP
  - `local-pref` attribute is only carried inside messages sent over iBGP session
  - Over an eBGP session, a router only advertises its best route towards each destination
    - Usually, import and export filters are defined for each eBGP session
  - Over an iBGP session, a router advertises only its best routes learned over eBGP sessions
    - A route learned over an iBGP session is *never* advertised over another iBGP session
    - Usually, no filter is applied on iBGP sessions

iBGP and eBGP : Example



- Note that the next-hop and the AS-Path of BGP update messages are only updated when sent over an eBGP session

iBGP and eBGP Packet Forwarding



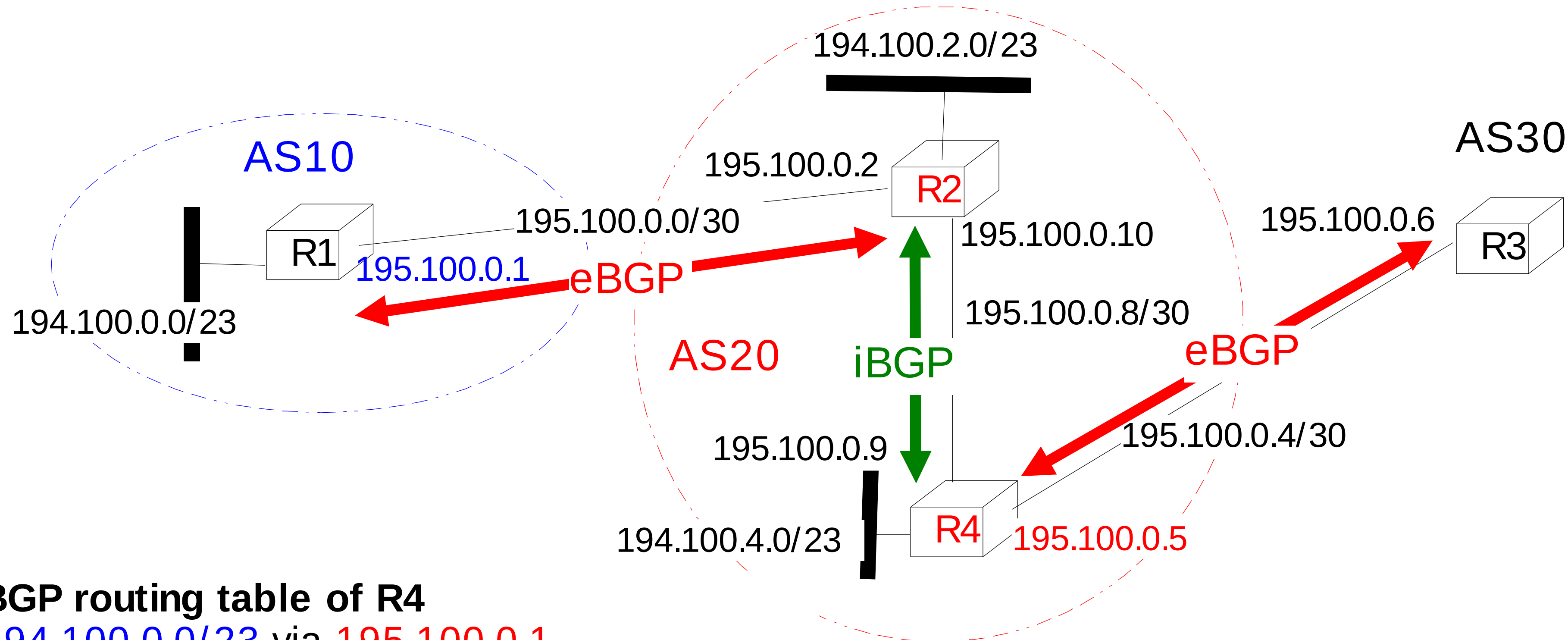
BGP routing table of R2
194.100.0.0/23 via 195.100.0.1

IGP routing table of R2
195.100.0.0/30 West
195.100.0.4/30 via 195.100.0.9
195.100.0.8/30 South
194.100.0.4/23 via 195.100.0.9
194.100.2.0/23 North

BGP routing table of R4
194.100.0.0/23 via 195.100.0.1

IGP routing table of R4
195.100.0.0/30 via 195.100.0.10
195.100.0.4/30 East
195.100.0.8/30 North
194.100.2.0/23 via 195.100.0.10
194.100.0.4/23 West

iBGP and eBGP Packet Forwarding (2)



BGP routing table of R4

194.100.0.0/23 via 195.100.0.1

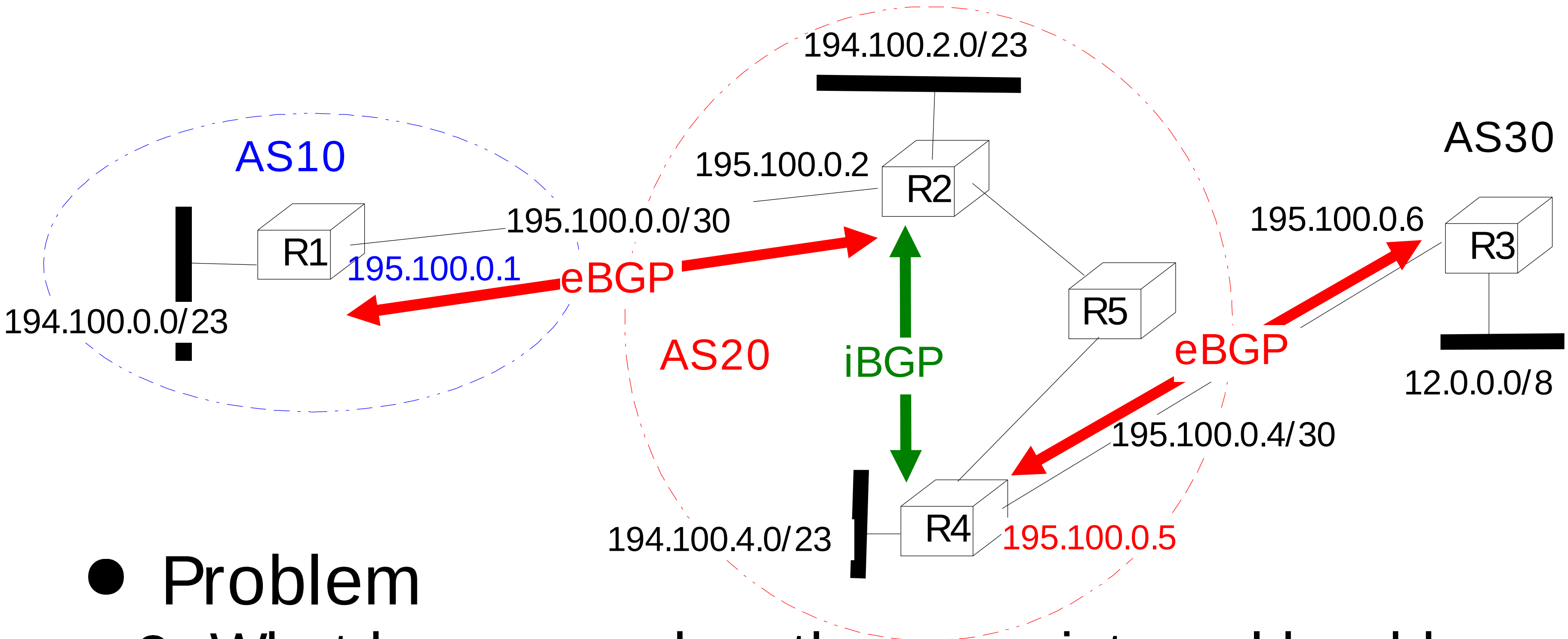
IGP routing table of R4

195.100.0.0/30 via **195.100.0.10**
195.100.0.4/30 East
195.100.0.8/30 North
194.100.2.0/23 via 195.100.0.10
194.100.4.0/23 West

Forwarding of R4

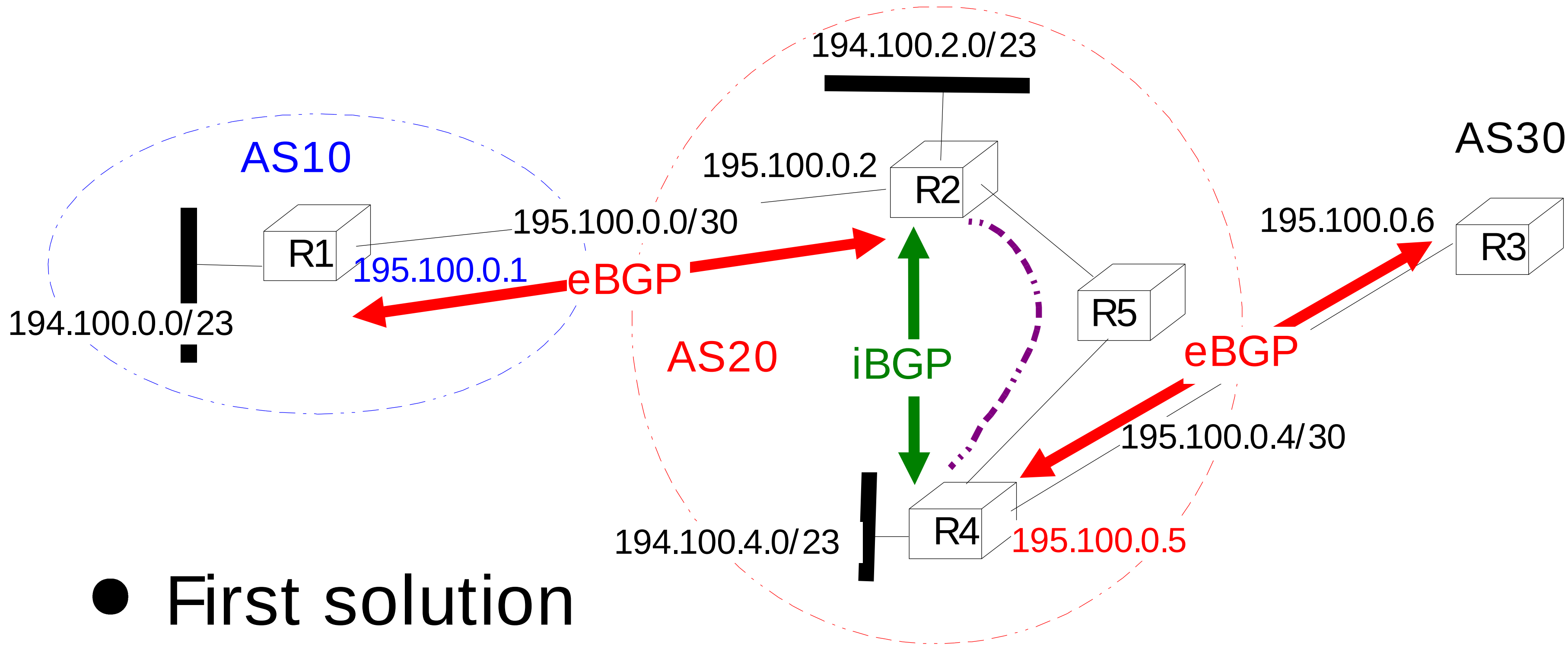
194.100.0.0/23 via **195.100.0.10**
195.100.0.0/30 via 195.100.0.10
195.100.0.4/30 East
195.100.0.8/30 North
194.100.2.0/23 via 195.100.0.10
194.100.4.0/23 West

Using non-BGP routers



- Problem
  - What happens when there are internal backbone routers between BGP routers inside an AS ?
    - iBGP session between BGP routers is easily established when IGP is running since iBGP runs over TCP connection
    - How to populate the routing table of the backbone routers to ensure that they will be able to route any IP packet ?

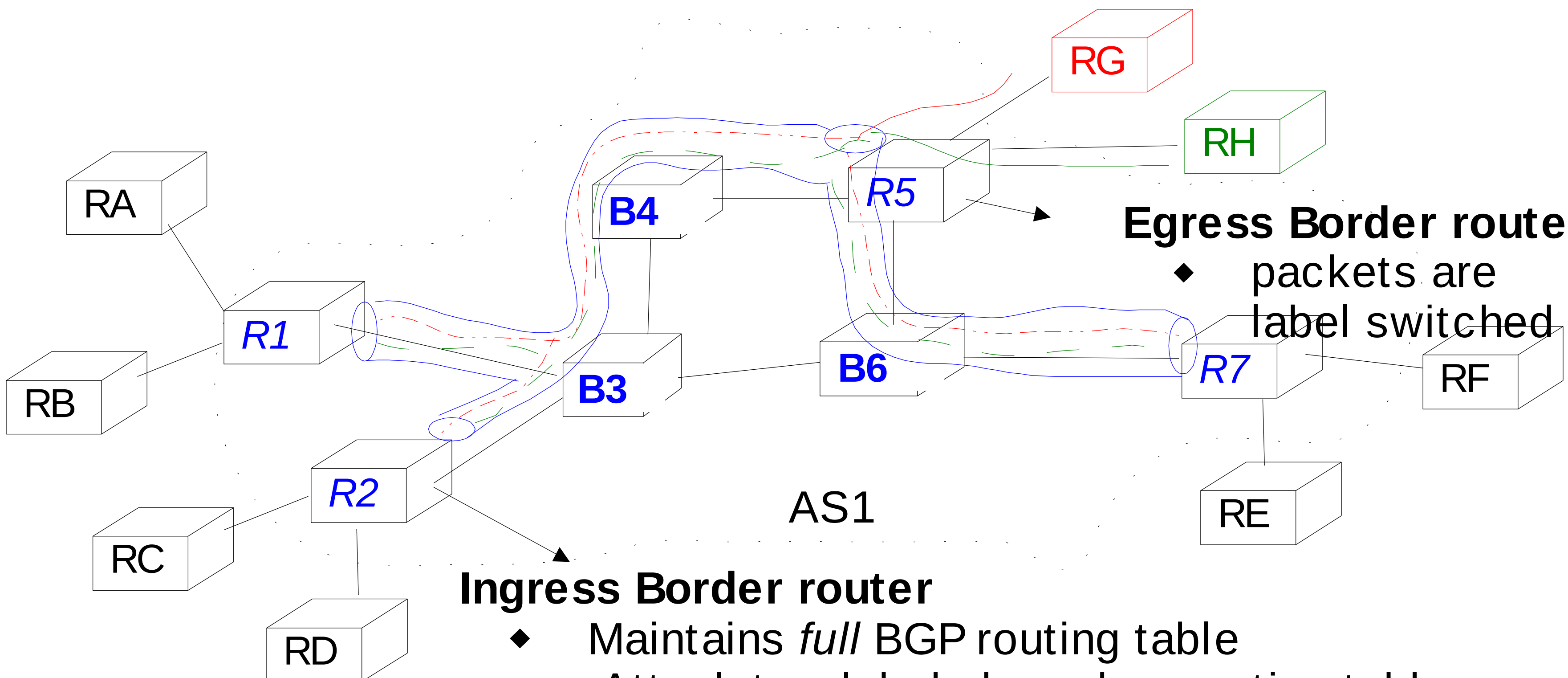
Using non-BGP routers (2)



- First solution
  - Use tunnels between BGP routers to encapsulate interdomain packets
    - GRE tunnel
      - Needs static configuration and be careful with MTU issues
    - MPLS tunnel
      - Can be dynamically established in MPLS enabled backbone

MPLS in large ISP networks

- Only one BGP table lookup inside the AS
  - Use a hierarchy of labels
    - top label is used to reach egress router
    - second label is used to reach eBGP peer



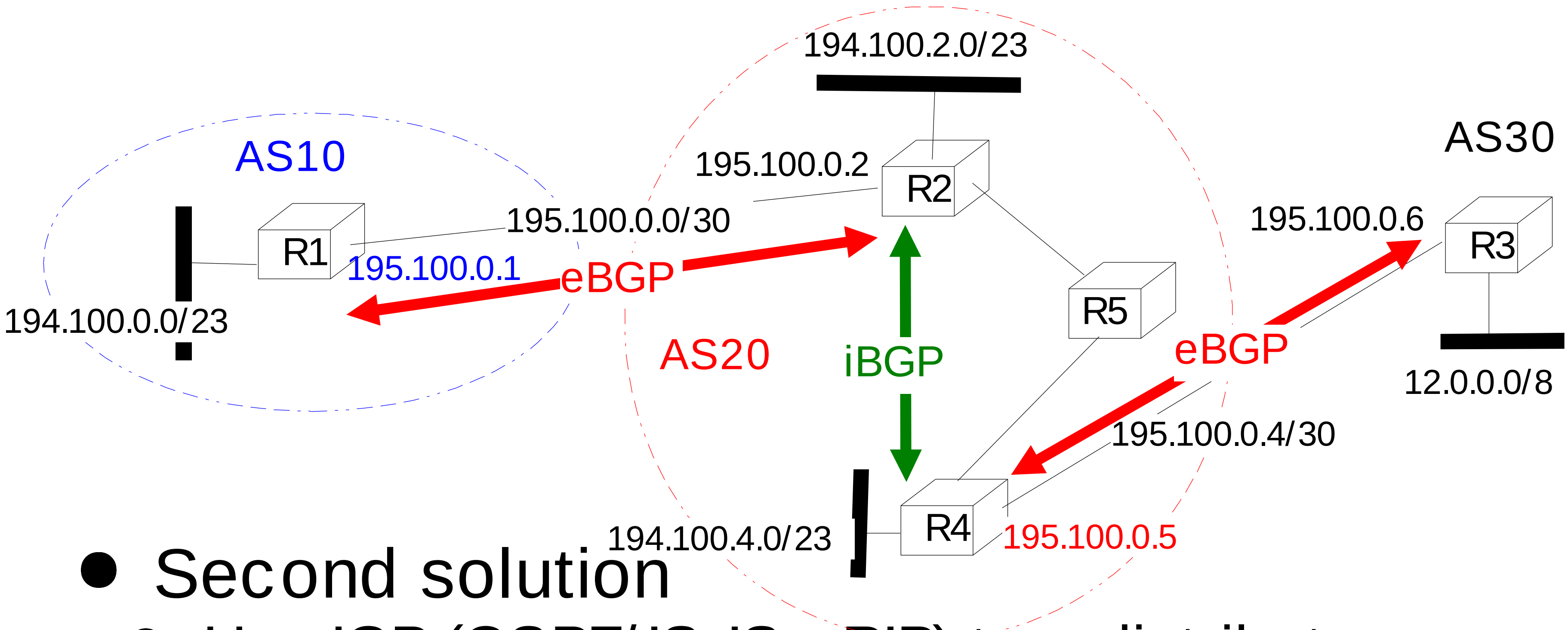
Egress Border router

- packets are label switched

Ingress Border router

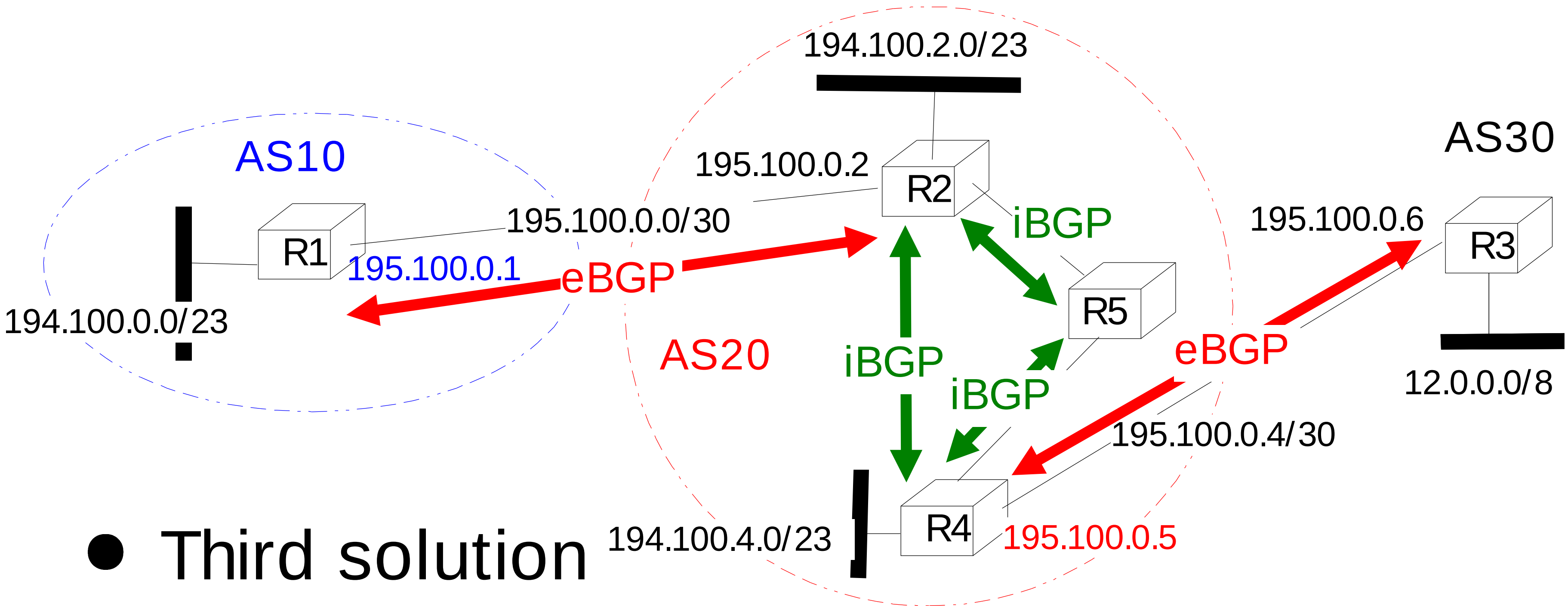
- Maintains *full* BGP routing table
- Attach two labels based on routing table

Using non-BGP routers (3)



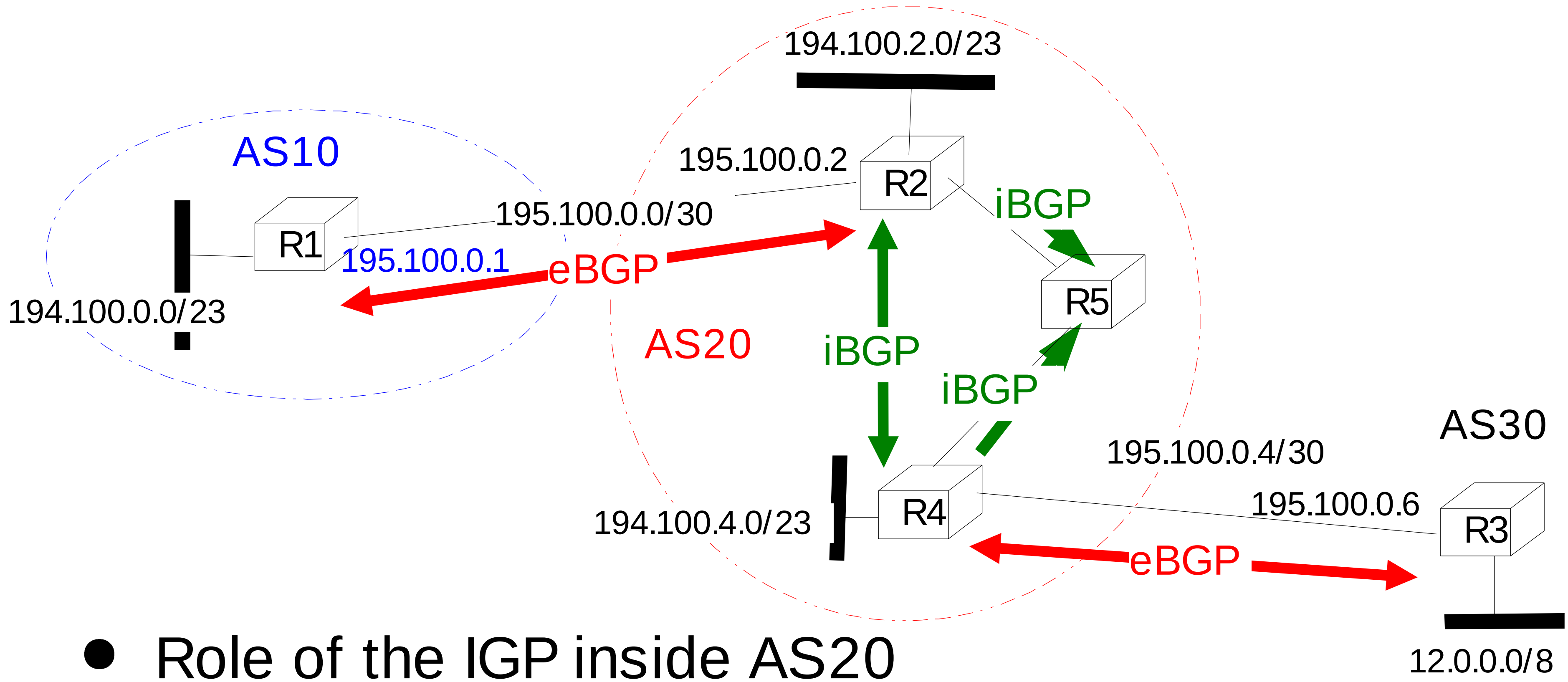
- Second solution
  - Use IGP (OSPF/ IS-IS - RIP) to redistribute interdomain routes to internal backbone routers
  - Drawbacks
    - Size of BGP tables may completely overload the IGP
    - Make sure that BGP routes learned by R2 and injected inside IGP will not be re-injected inside BGP by R4 !

Using non-BGP routers (4)



- Third solution
  - Run BGP on internal backbone routers
  - Internal backbone routers need to participate in iBGP full mesh
    - Internal backbone routers receive BGP routes via iBGP but never advertise any routes
      - Remember : a route learned over an iBGP session is never advertised over another iBGP session

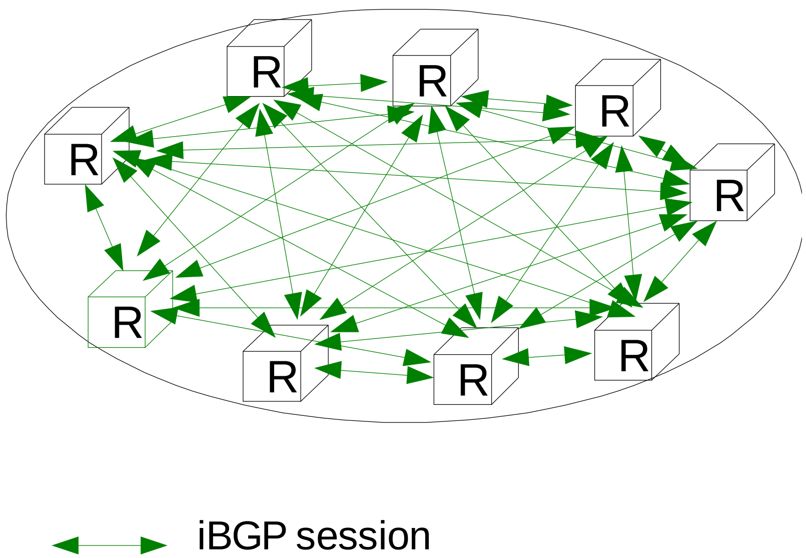
The roles of IGP and BGP

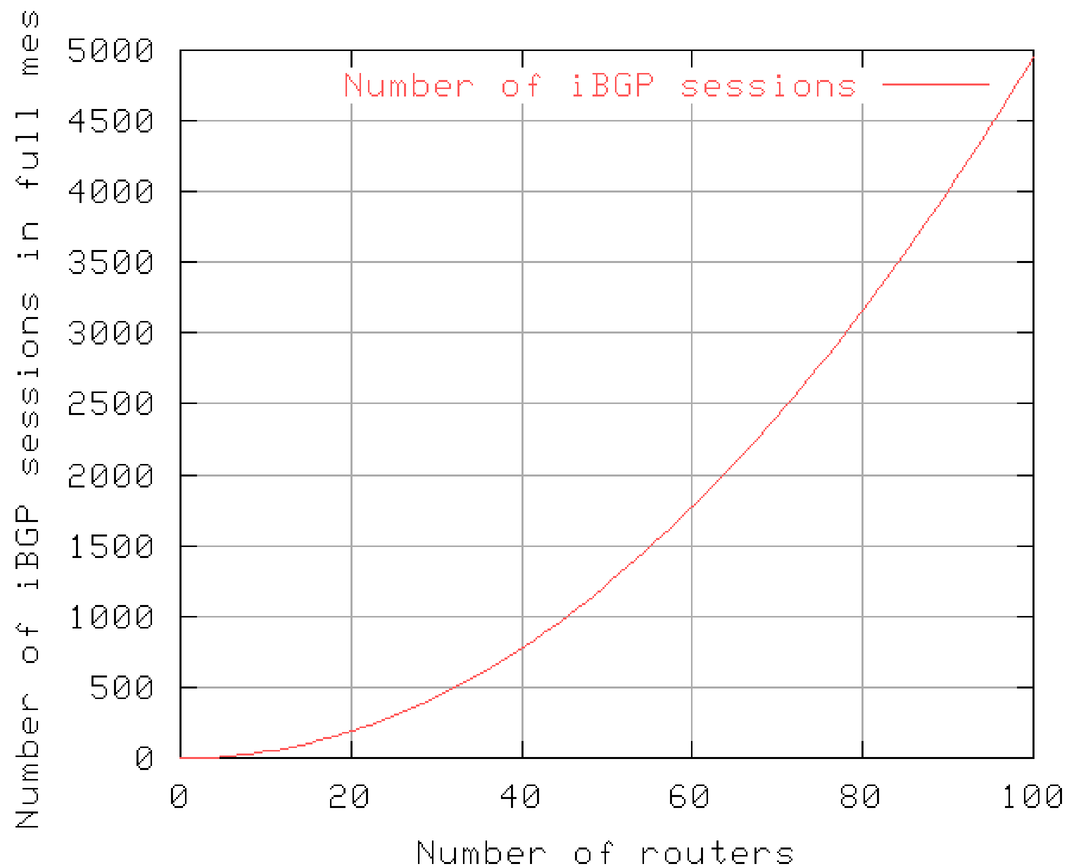


- Role of the IGP inside AS20
  - Distribute internal topology and internal addresses R2-R4-R5)
- Role of BGP inside AS20
  - Distribute the routes towards external destinations
  - IGP must run to allow BGP routers to establish iBGP sessions

The iBGP full mesh

- Drawback
  - N*(N-1)/2 iBGP sessions for N routers



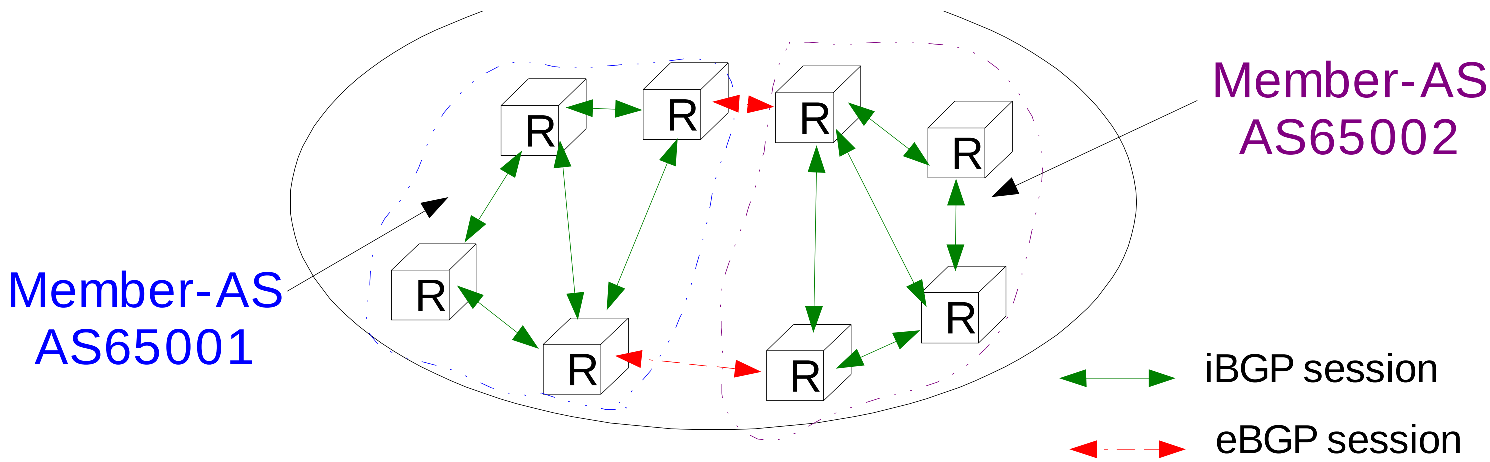


Outline

- Organization of the global Internet
- BGP basics
- BGP in large networks
  - The needs for iBGP
  - Confederations and Route Reflectors
  - Scalable routing policies
  - The dynamics of BGP
- Interdomain traffic engineering with BGP
- BGP-based Virtual Private Networks

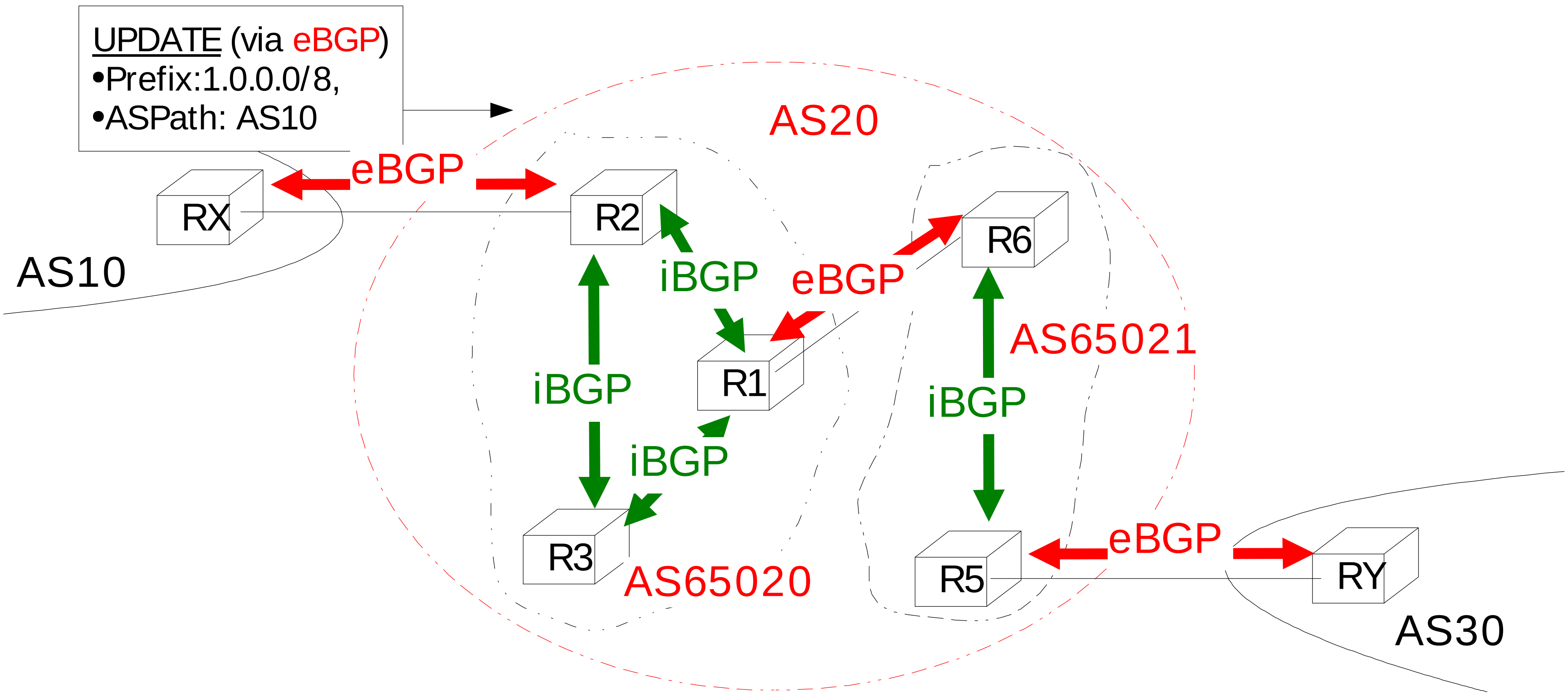
How to scale iBGP in large domains ?

- Confederations
  - Divide the large domain in smaller sub-domains
    - Use iBGP full mesh inside each sub-domain
    - Use eBGP between sub-domains



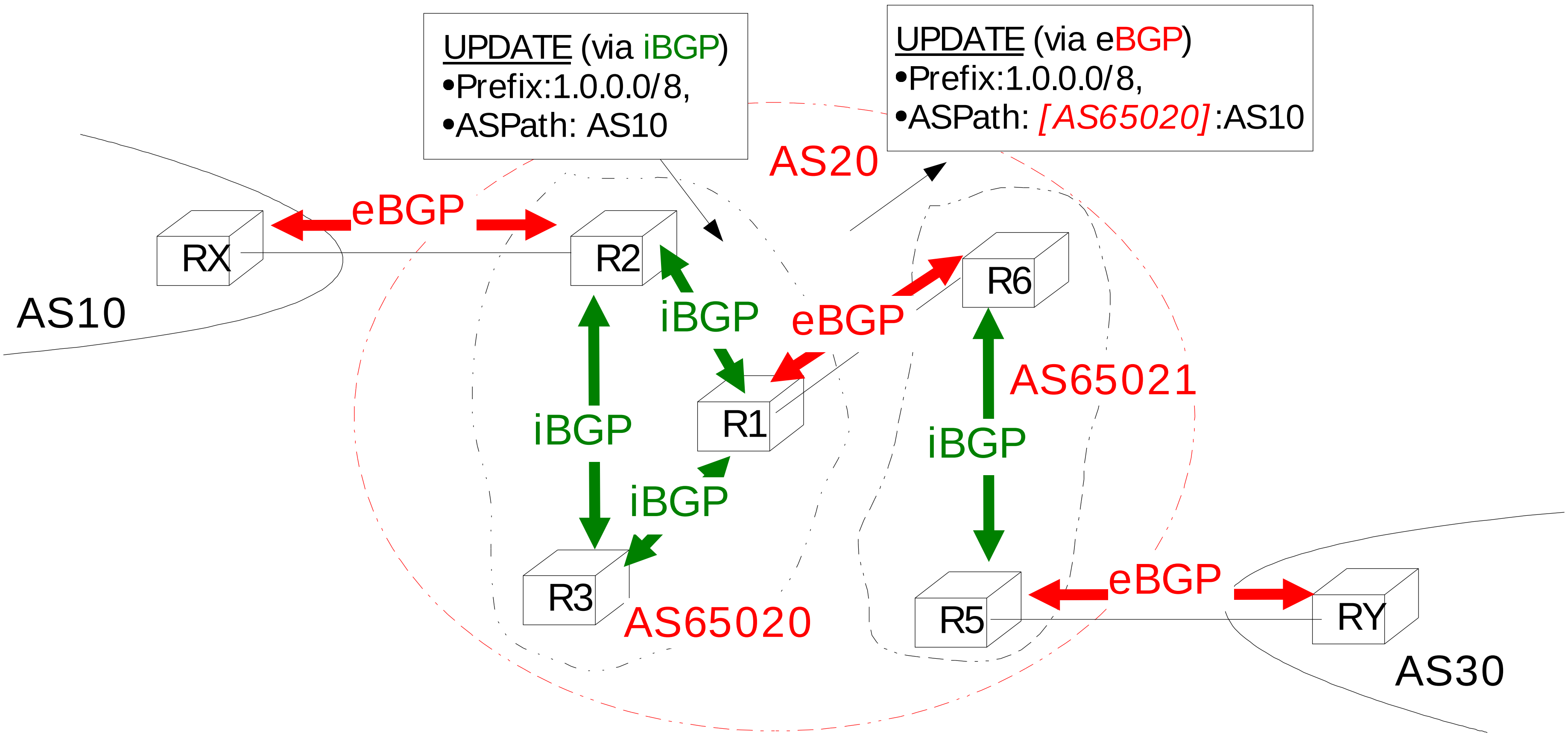
  - Each router is configured with two AS numbers
    - Its confederation AS number
    - Its Member-AS AS number
  - Usually, a single IGP covers the whole domain

Confederations : example



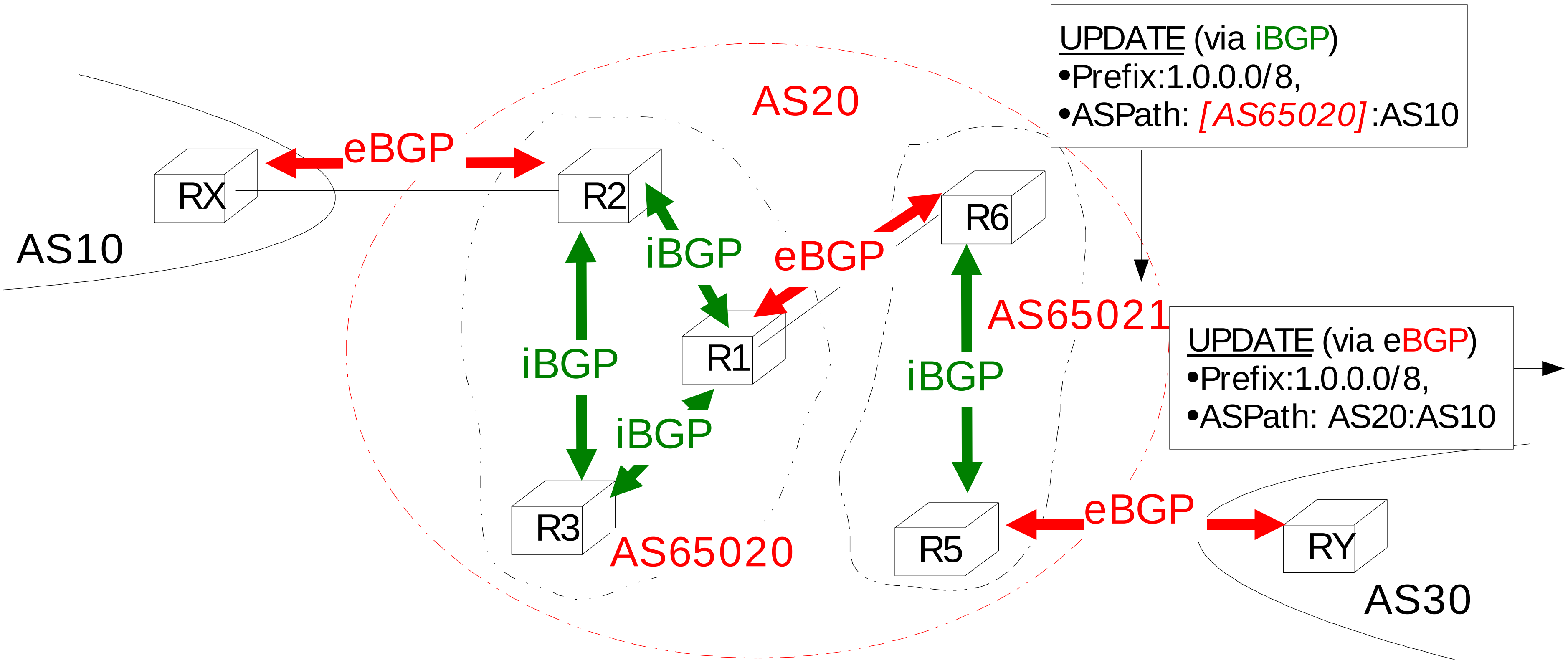
- On the eBGP session between R2 and RX, R2 belongs to AS20
- On the eBGP session between R5 and RY, R5 belongs to AS20
- On the eBGP session between R1 and R6, R1 belongs to AS65020 and R6 belongs to AS65021

Confederations : example (2)



- When propagating an UPDATE via eBGP to another router of the same confederation, R1 inserts its Member-AS number in the AS_PATH

Confederations : example (3)

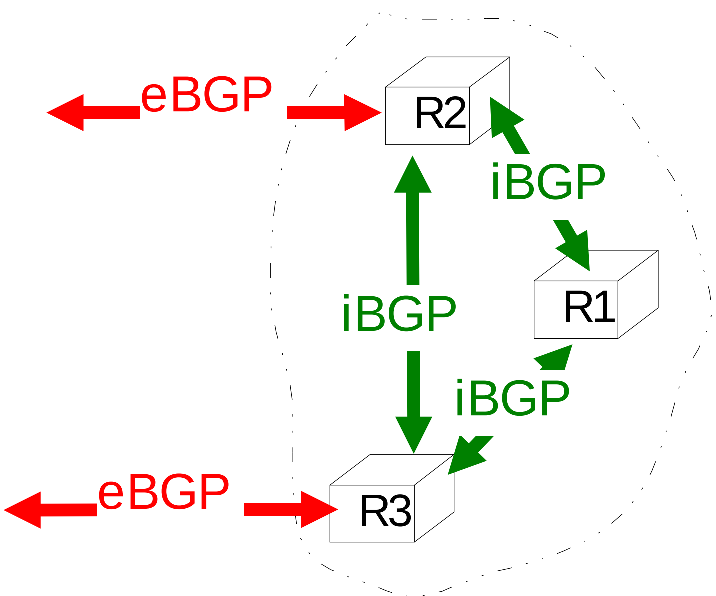


- When propagating an UPDATE via eBGP to a router outside its confederation, R5 removes the internal path from the AS_Path and inserts its Confederation AS number in the AS_PATH

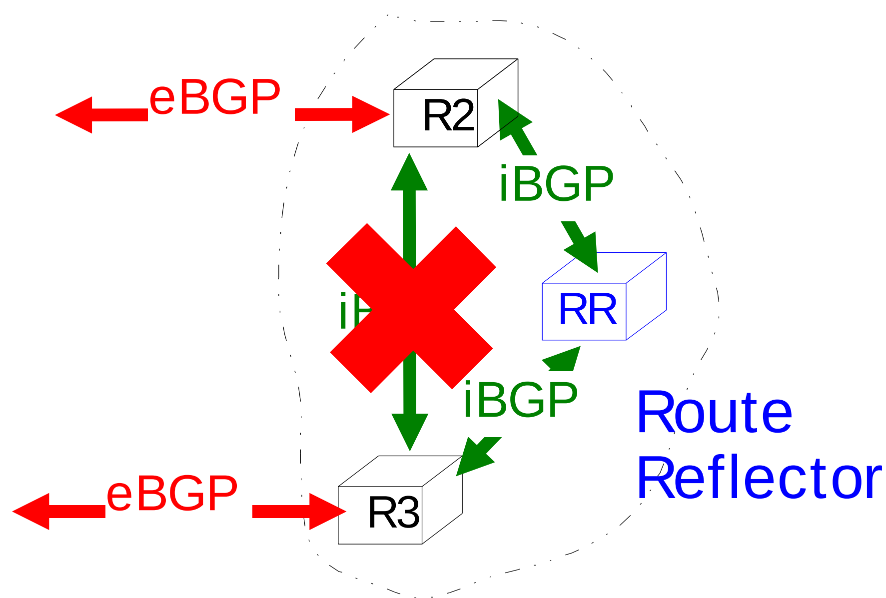
Route reflectors
An alternative to confederations

- Route reflectors
  - A route reflector is a special router that is allowed to propagate the routes learned over iBGP sessions on other iBGP sessions

Normal iBGP full mesh

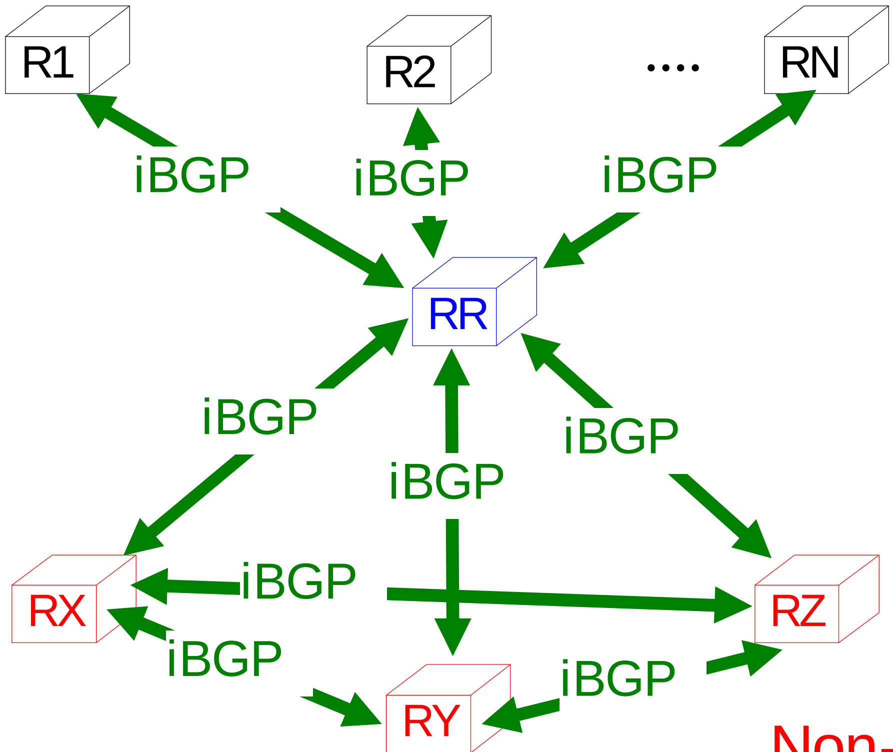


iBGP with one route reflector



Behavior of a Route Reflector

- Two types of iBGP peers of a route reflector

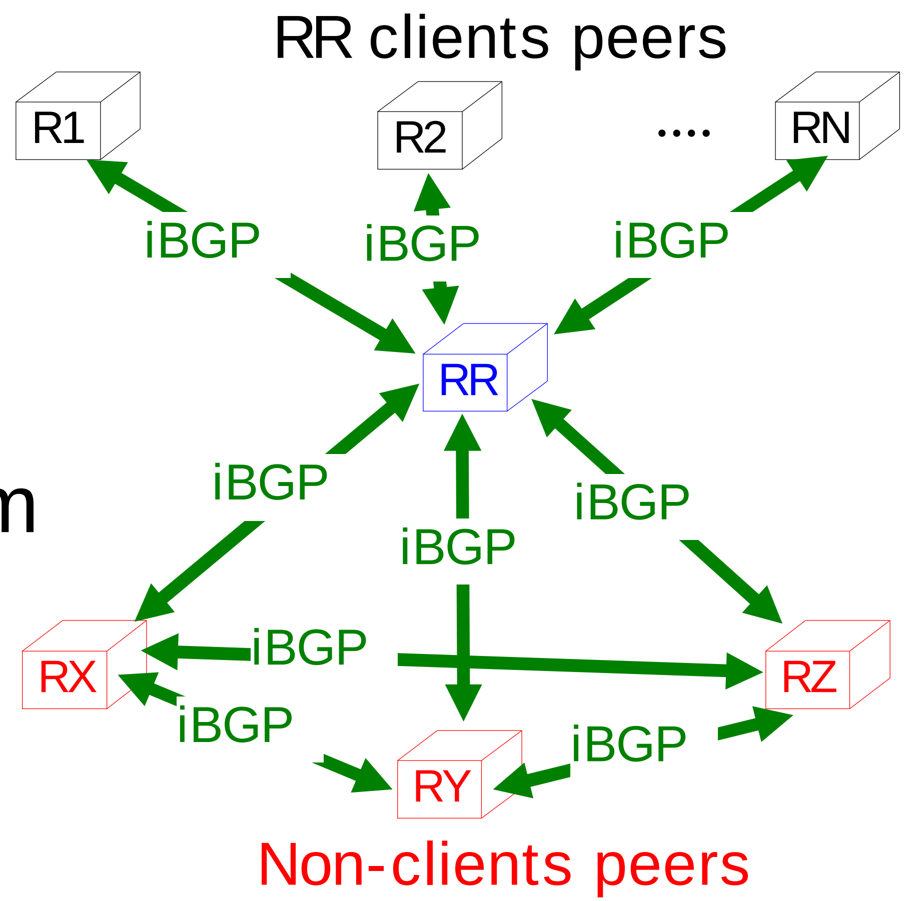


RR clients peers
(do not participate in
iBGP full mesh)

Non-clients peers
(participate in iBGP full mesh)

Behavior of a Route Reflector

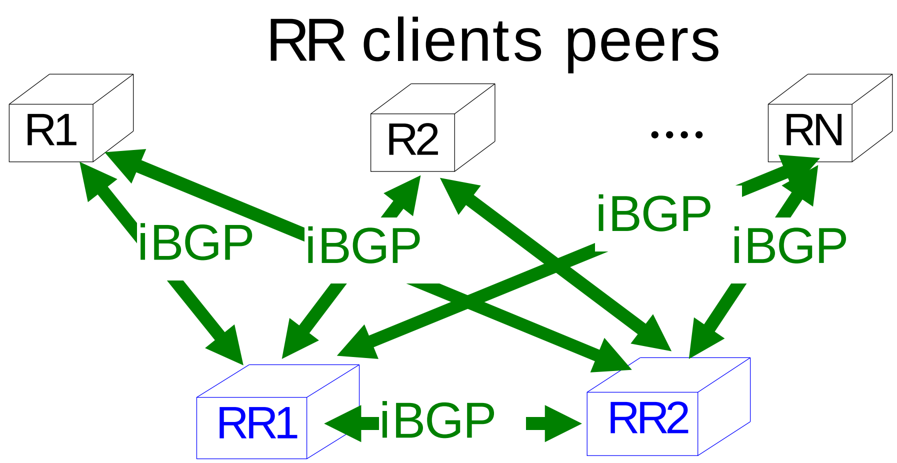
- Route received from an eBGP session or a client peer
  - Select best path
  - Advertise to
    - All client peers
    - All non-client peers



- Route received from non-client peer
  - Select best path
  - Advertise to :
    - All client peers

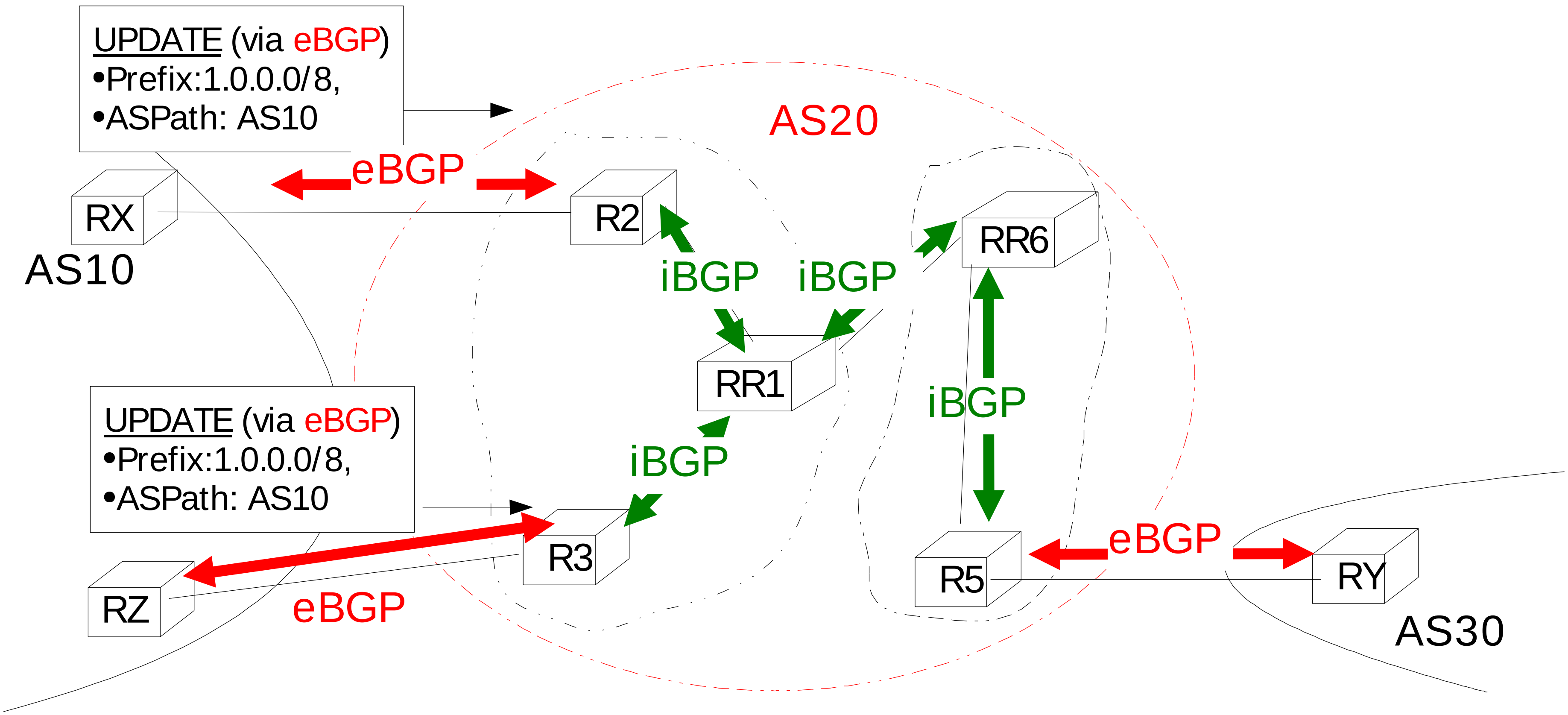
Fault tolerance of route reflectors

- How to avoid having the RR as a single point of failure ?
  - Solution
    - Allow each client peer to be connected at 2 RRs



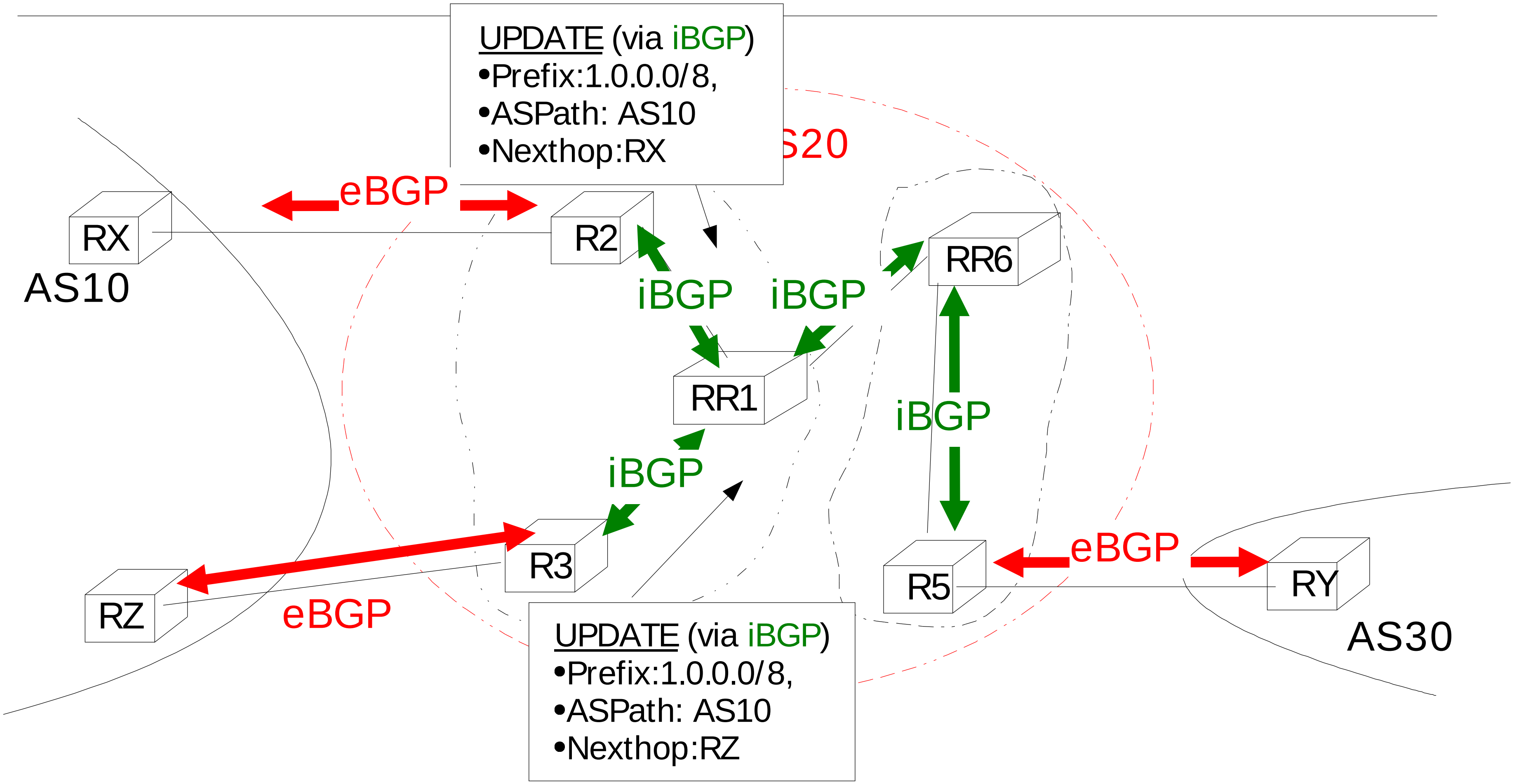
  - Issue
    - Configuration errors may cause redistribution loops
      - ORIGINATOR_ID used to carry router ID of originator of route
      - CLUSTER_LIST contains the list of RR that sent the UPDATE message inside the current AS

Route reflectors : an example



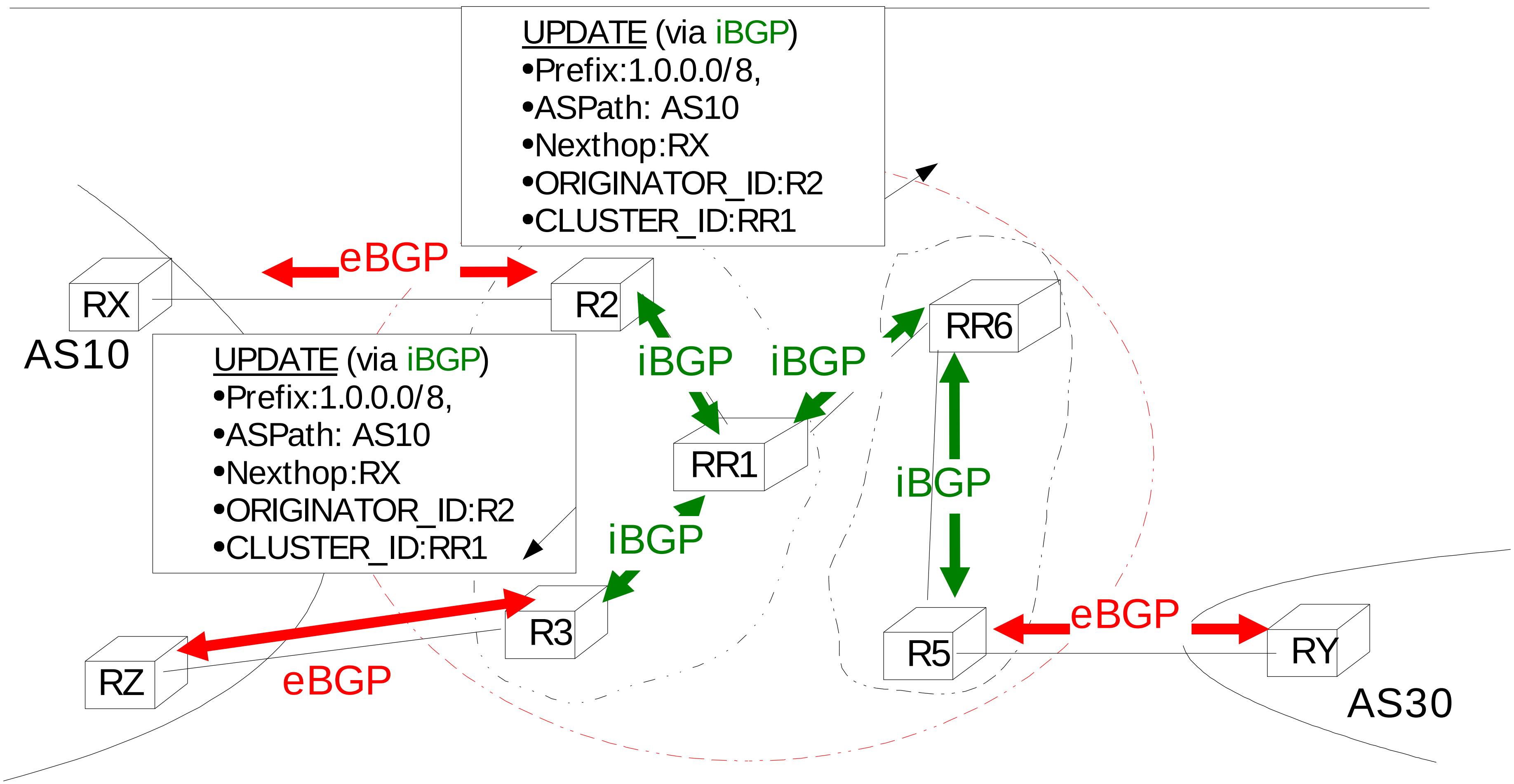
- R2 and R3 are clients of Route Reflector RR1
- RR1 and RR6 are in iBGP full mesh
- R5 is client of Route Reflector RR6

Route reflectors : an example (2)



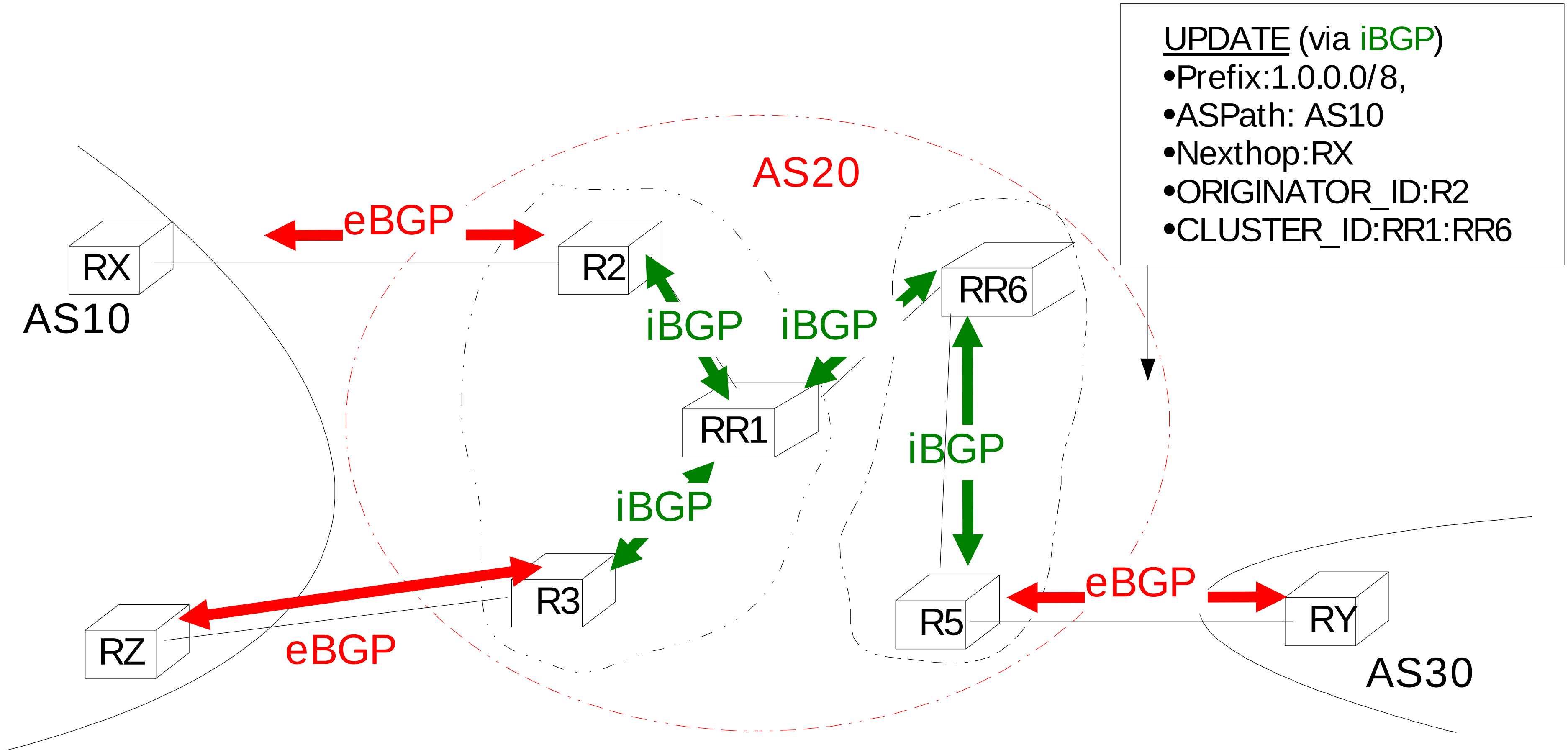
- RR1 will select its best path towards 1.0.0.0/8 and will re-advertise it by adding the ORIGINATOR_ID and the CLUSTERID

Route reflectors : an example (3)



- RR1 prefers the path to 1.0.0.0/8 via RX-R2
  - RR1 advertises this path to its client peer (R3)
    - the path is not advertised to R2 since R2 already received it
  - RR1 advertises this path to its non-client peer (RR6)

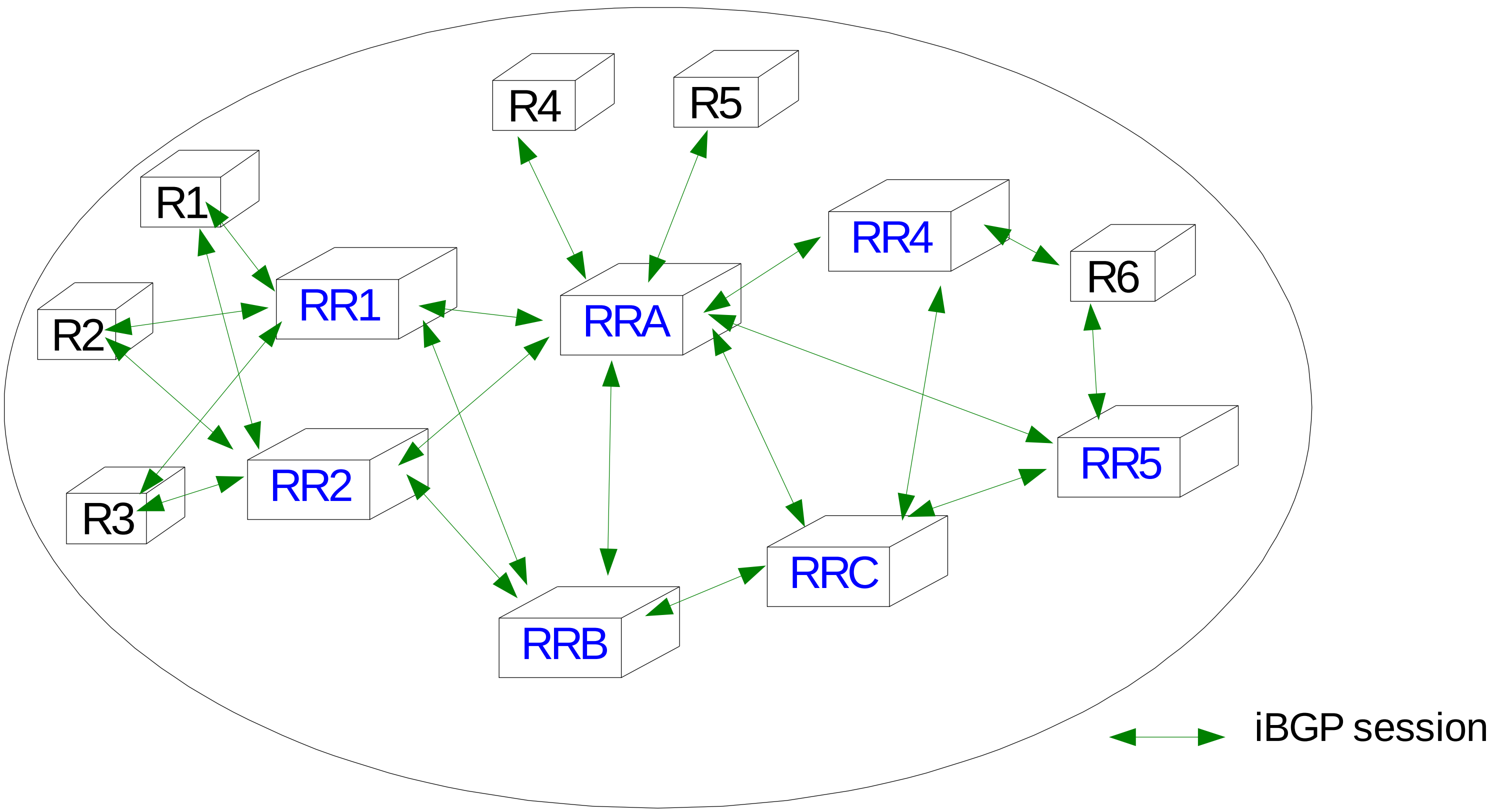
Route reflectors : an example (4)



- RR6 advertises the path to 1.0.0.0/8 via RX-R2
  - to its client peer R5
- R5 will remove ORIGINATOR_ID and CLUSTER_ID before advertising the path to RY via eBGP

Hierarchy of route reflectors

- In large domains, a hierarchy of route reflectors can be built



Confederations versus Route reflectors

- Confederations
  - Solves iBGP scaling
  - Redundancy with iBGP full-mesh inside each MemberAS
  - Possible to run one IGP per Member AS
  - Requires manual router configuration
  - Can be used when merging domains
  - Can lead to some routing oscillations
- Route reflectors
  - Solves iBGP scaling
  - Redundancy by using Redundant RRs
  - Usually a single IGP for the whole AS
  - Requires manual router configuration
  - Can lead to some routing oscillations

Outline

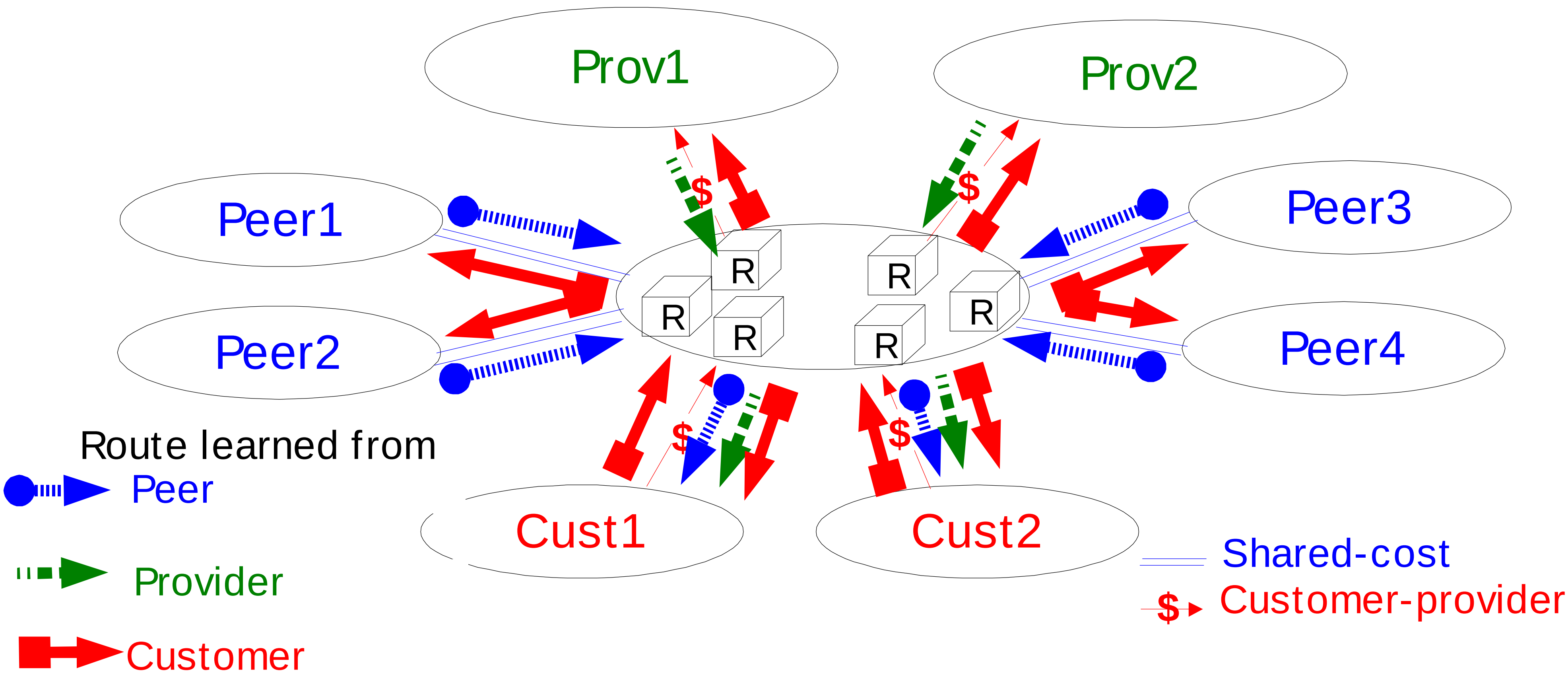
- Organization of the global Internet
- BGP basics
- BGP in large networks
  - The needs for iBGP
  - Confederations and Route Reflectors
  - Scalable routing policies
  - The dynamics of BGP
- Interdomain traffic engineering with BGP
- BGP-based Virtual Private Networks

The Community attribute

- Principle
  - Optional transitive attribute containing a set of communities
  - each community acts as a marker
    - one community is represented as a 32 bits value
    - usually routes with same marker are treated same manner
  - Standardized communities
    - NO_EXPORT (0xFFFFFF01)
    - NO_ADVERTISE (0xFFFFFF02)
  - Delegated communities
    - 65536 communities have been delegated to each AS
      - ASX65536 ASX:0 through ASX:65535

Scalable routing policies with communities

- Principle
  - attach same community value to all routes that need to receive the same treatment



More complex routing policies with communities

- Other utilizations of communities
  - Research ISP providing two types of services
    - Access to research networks for universities
    - Access to the commercial Internet for universities and government institutions
    - Solution
      - Tag routes learned from research network and commercial Internet
      - Only announce the universities to research network
      - Only advertise research network to universities
  - Commercial ISP providing several transit services
    - Full transit service
      - Announce all known routes to all customers
      - Advertise customer routes to all peers, customers, providers
    - Client routes only
      - Only advertise to those customers the routes learned from customers, but not the routes learned from peers
      - Advertise the routes learned from those customers only to customers

Other utilizations of communities

- Communities used for tagging
  - Community attached by router that receives route to indicate country where route was received
    - Example (Eunet, AS286)
      - 286:1000 + countrycode for Public peer routes
      - 286:2000 + countrycode for Private peer routes
      - 286:3000 + countrycode for customer routes
    - Another example (C&W, AS3561)
      - 3561:SRCC
        - S : Peer or Customer
        - R : Regional Code
        - CC : ISO3166 country code
  - Community to indicate IX where route was learned
    - Example : AS12369 (Global Access Telecommunications)
      - 13129:2110 : route leared at DE-CIX
      - 13129:2120 : route learned at INXS
      - 13129:2130 : route learned at SFINX

Issues with communities

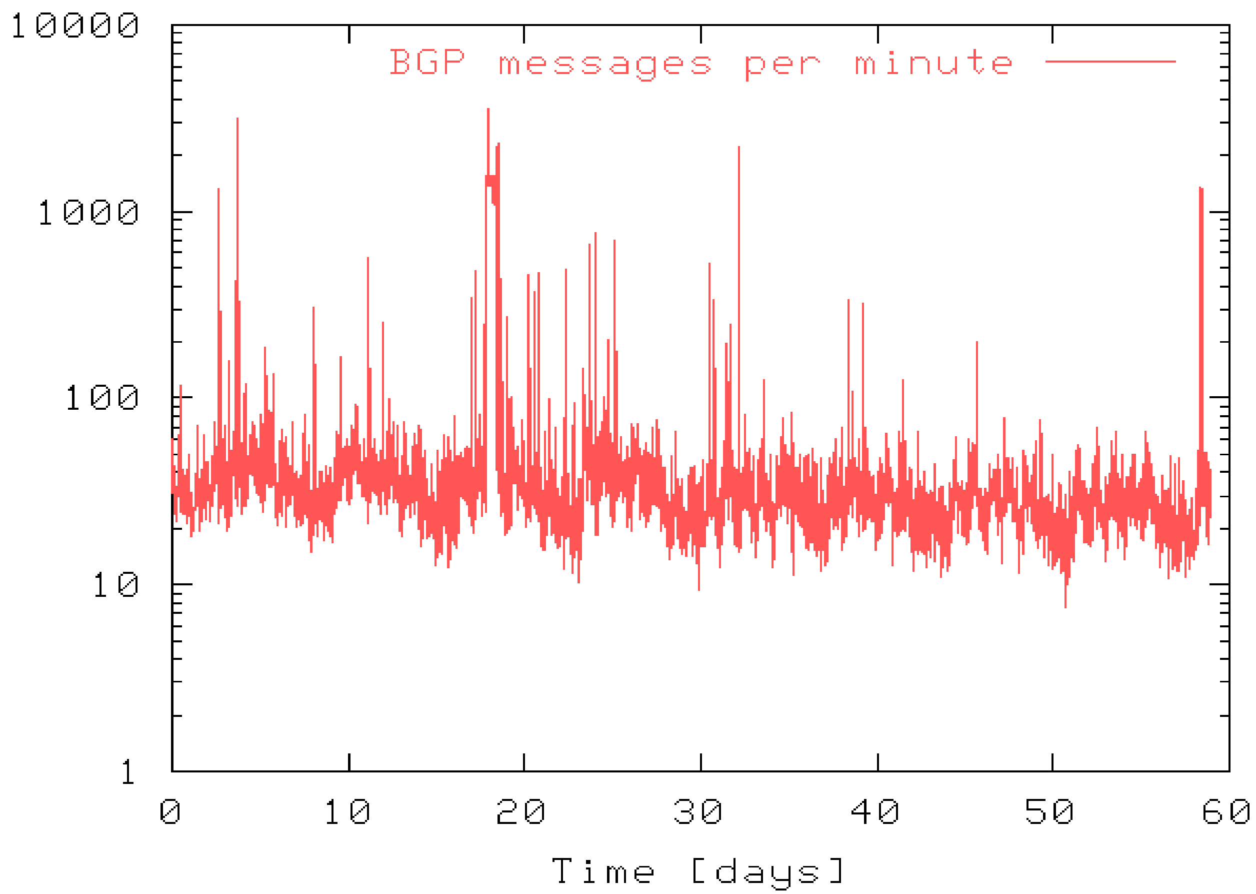
- Issues
  - A router may easily add community values
  - The community attribute is optional and transitive
    - A community value added by one router could be propagated to the global Internet
      - In Jan 2003, 50% of the BGP routes contained communities
      - Some routes may contain several tens of communities
  - The semantics of communities is defined locally
    - Some ASes advertise the semantics of their communities by using RPSL
    - Most of the community values that a router receives are useless, but they consume memory and some CPU and may cause BGP UPDATEs to be widely distributed
- Best Current Practice
  - If you use communities, make sure that they are not advertised uselessly to the entire Internet...

Outline

- Organization of the global Internet
- BGP basics
- BGP in large networks
  - The needs for iBGP
  - Confederations and Route Reflectors
  - Scalable routing policies
  - The dynamics of BGP
- Interdomain traffic engineering with BGP
- BGP-based Virtual Private Networks

The dynamics of BGP

- Ideally, BGP routes should be stable and a BGP router should seldom receive messages
  - On the global Internet, things are less simple

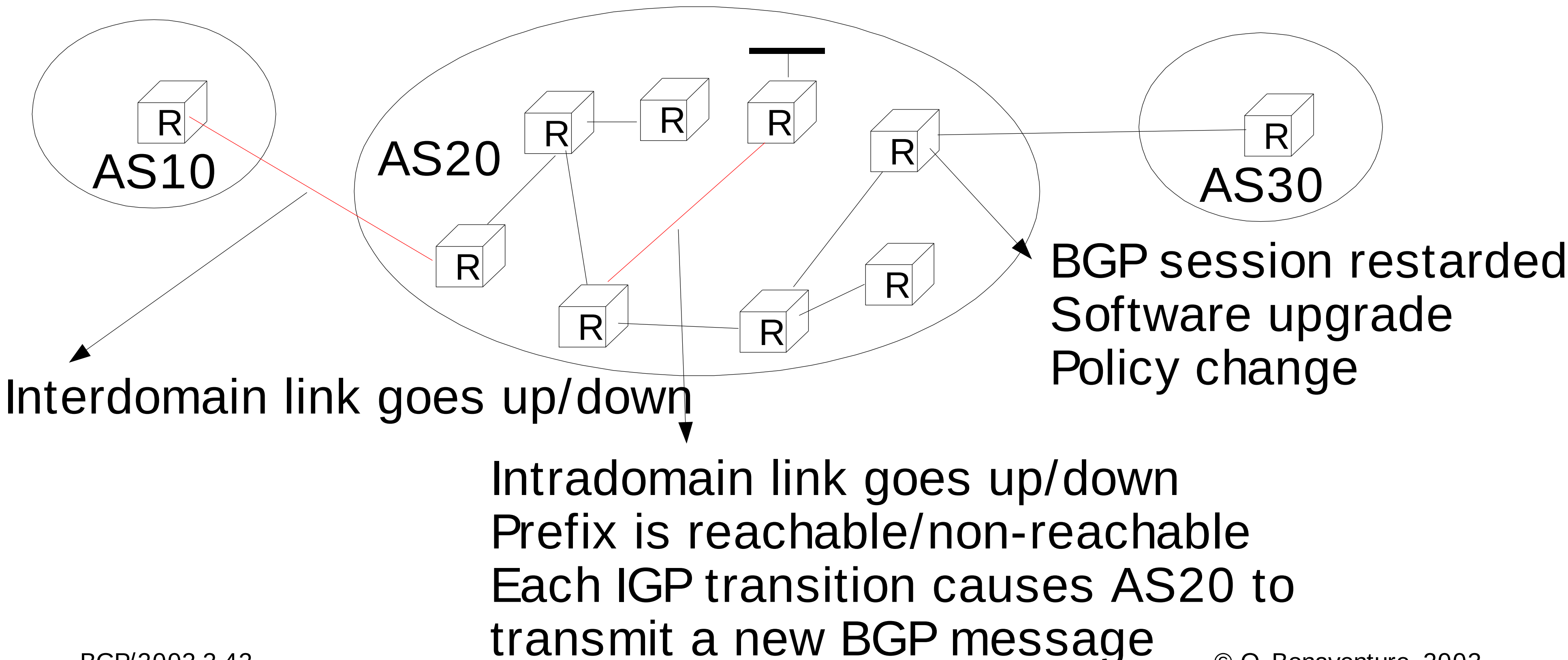


A closer look at the BGP messages

- One month study of a client of AS2611
  - Captured all outgoing traffic sent to AS2611
  - Captured all BGP messages received from AS2611
- Some findings
  - Received advertisements for 103,853 # AS Paths
  - But
    - 50% of those AS Paths appeared in our BGP routing table for less than 9 minutes
      - Other studies have shown that a small number of prefixes were responsible for most BGP messages
    - Only 31,151 AS Paths were actually used to send packets
    - 95% of all the traffic sent by the stub AS was transmitted over 13,000 AS Paths that were stable for more than 99% of time

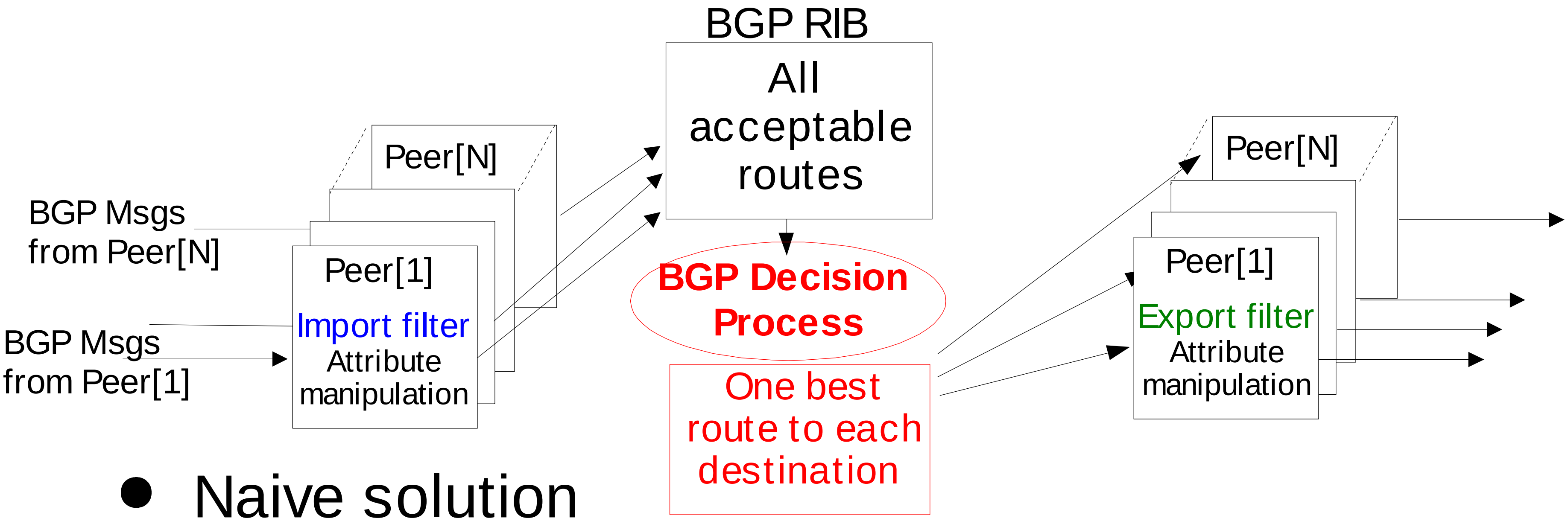
Why so many BGP messages ?

- The Internet is large and complex
- A small remote event may result in sending BGP messages to all BGP routers



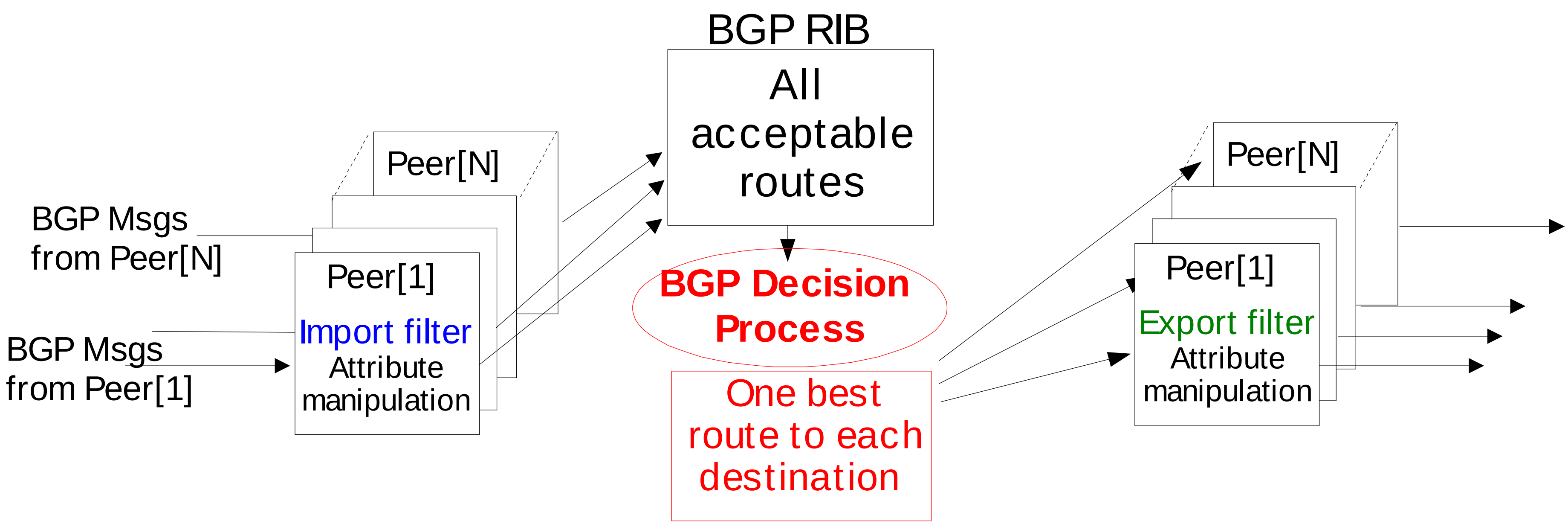
Changes in BGP policies

- How to change the import/export policies used by one BGP router ?



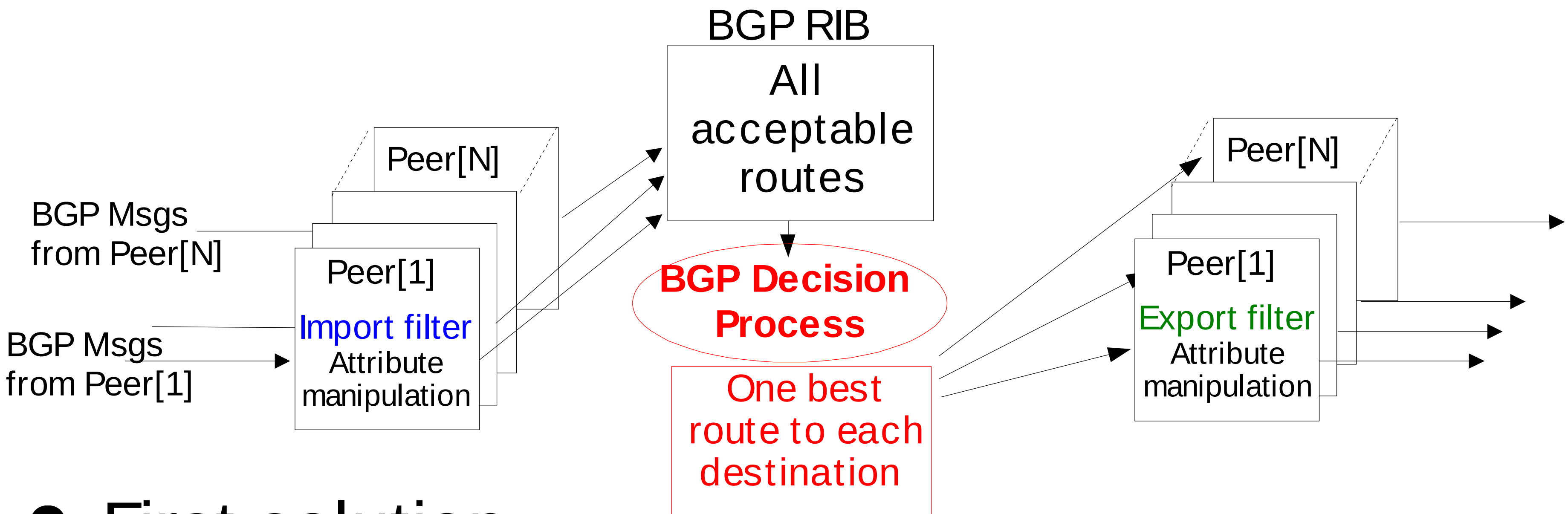
- Naive solution
  - Change import/export filters
  - Stop BGP sessions
    - Peers may need to send lots of Withdraw messages !
  - Reestablish BGP sessions
    - BGP router will receive and process lots of Update messages !

How to smoothly change export filters ?



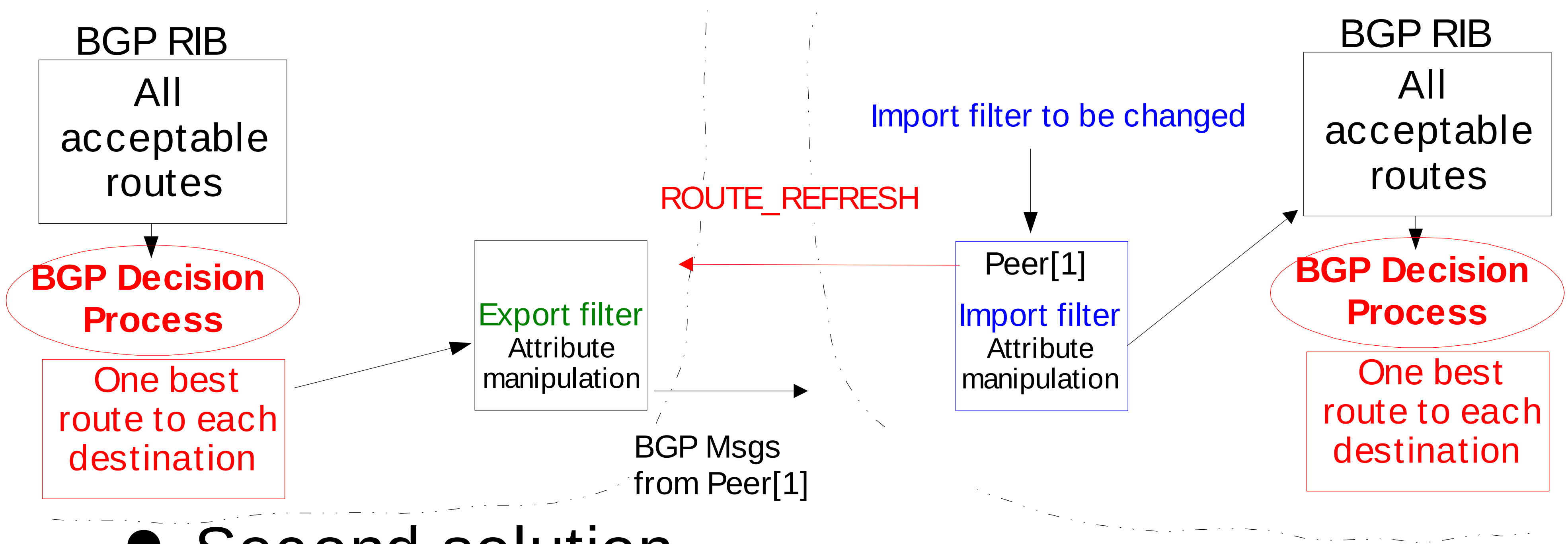
- Principle
  - Update export filters that need to be changed
  - For each BGP session using a modified filter
    - Scan BGP routing tables to determine the BGP messages to be sent according to the new filter
    - Send the required BGP messages

How to smoothly change import filters ?



- First solution
  - Store all UPDATE messages (unmodified) received from each peer before applying the import filter
  - When an import filter changes
    - Apply the new filter to the stored UPDATE messages
- Drawback
  - Memory consumption

How to smoothly change import filters (2) ?

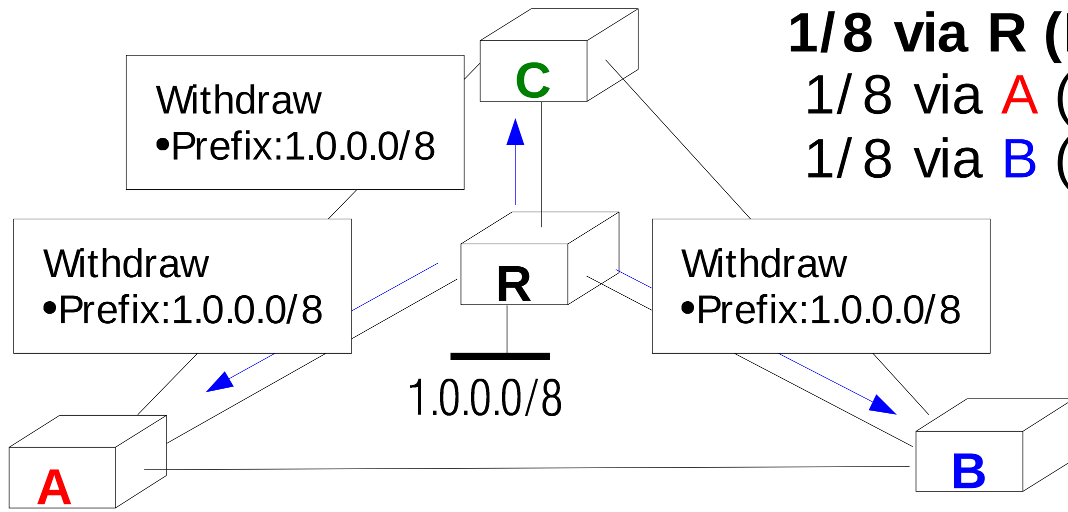


- Second solution
  - Do not store received UPDATE messages
  - When an import filter changes
    - Send the ROUTE_REFRESH BGP message to request the concerned peer to send again <u>all his messages</u>
    - Apply the new filter to BGP messages received after the transmission of the ROUTE_REFRESH

Another reason for the BGP messages

- In some cases, BGP may try several paths

Routing table of C
1/8 via R (Path: R) (best)
1/8 via A (Path: A-R)
1/8 via B (Path: C-R)



Routing table of A
1/8 via R (Path: R) (best)
1/8 via B (Path: B-R)
1/8 via C (Path: C-R)

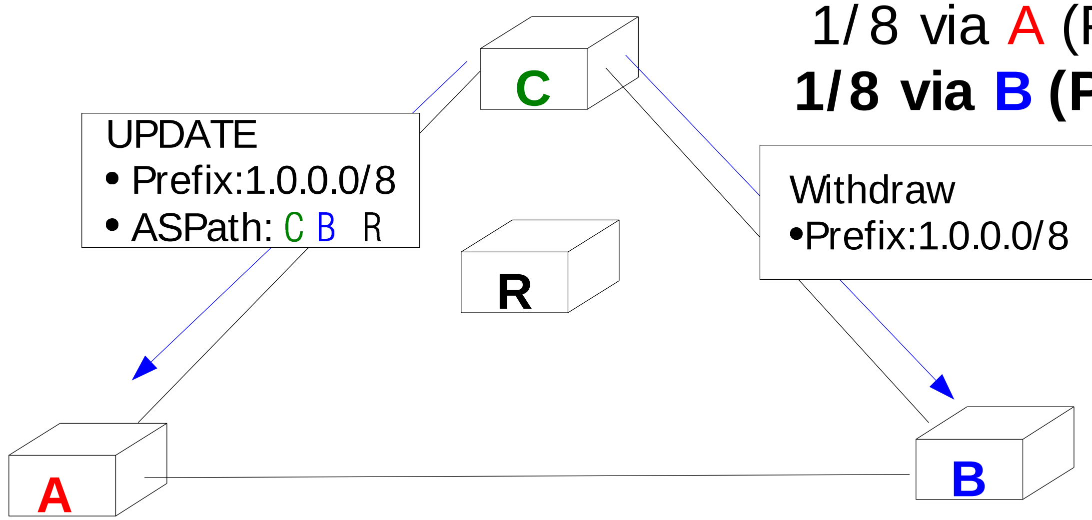
Routing table of B
1/8 via R (Path: R) (best)
1/8 via A (Path: A-R)
1/8 via C (Path: C-R)

  - Routers will process the withdraw message and ... advertise alternate routes to their peers

Another reason for the BGP messages (2)

- C processes first the withdraw

<u>Routing table of C</u>
1/8 via A (Path: A-R)
1/8 via B (Path: B-R) (best)



<u>Routing table of A</u>
1/8 via B (Path: B-R) (best)
1/8 via C (Path: C-R)

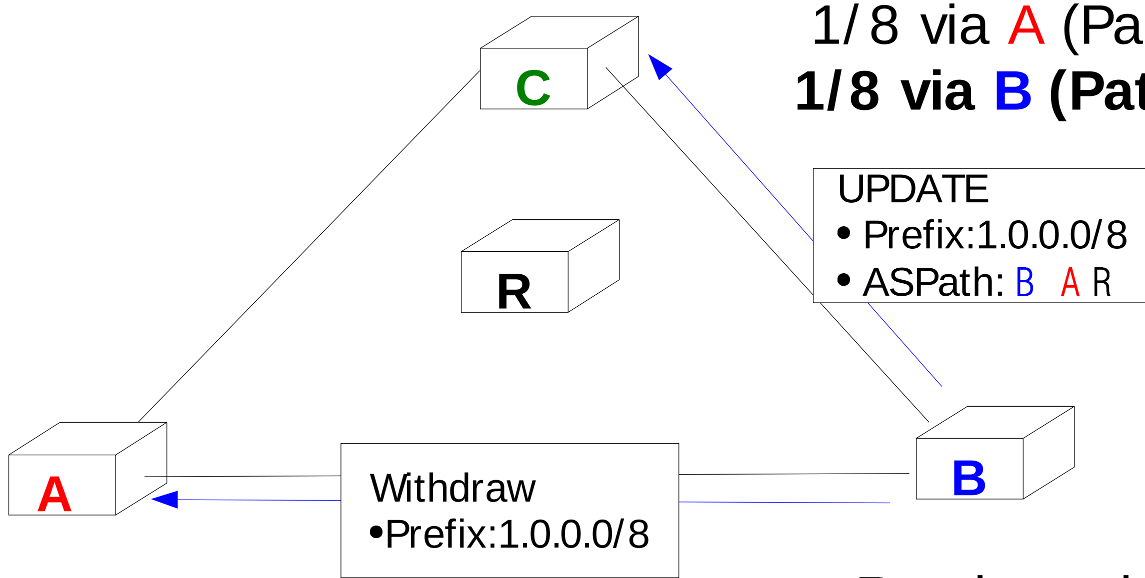
<u>Routing table of B</u>
1/8 via A (Path: A-R)
R via C (Path: C-R) (best)

  - A learns a worse (but valid) route towards 1/8
  - C sends withdraw to B since previous advertised path (C-R) is not available anymore and C has chosen route via B

Another reason for the BGP messages (3)

- B sends announcements

<u>Routing table of C</u>
1/8 via A (Path: A-R)
1/8 via B (Path: B-R) (best)



<u>Routing table of A</u>
1/8 via B (Path: B-R) (best)
1/8 via C (Path: C-B-R)

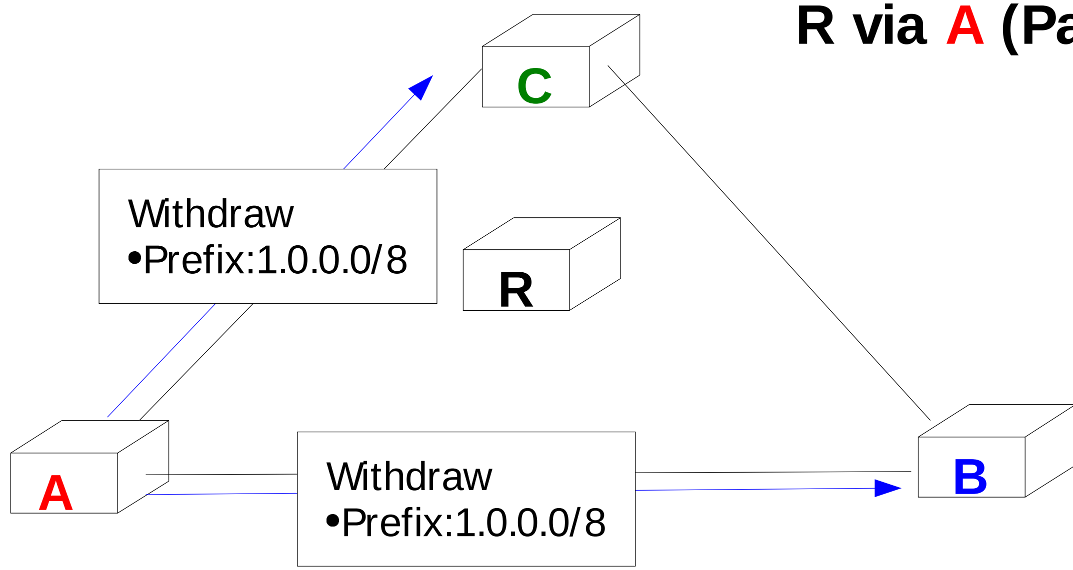
<u>Routing table of B</u>
1/8 via A (Path: A-R)

- C learns a longer path towards 1/8
- B sends a withdraw to A since its only route is via A

Another reason for the BGP messages (4)

- A sends announcements

Routing table of C
R via A (Path: A-R)



Routing table of A
1/8 via C (Path: C-B-R)

Routing table of B
1/8 via A (Path: A-R)

- A can only send a withdraw to C and B since they both appear in the ASPath of their route to reach 1/8
- B and C learn that their route via A is invalid

How to reduce the number of unnecessary BGP messages ?

- Avoid transmitting messages too frequently

  - Two UPDATE messages sent by the same BGP peer and advertising the same route should be separated by at least *MinRouteAdvertisementInterval* (MRAI) seconds
    - Default value for MRAI : 30 seconds

  - Advantage
    - Reduces the number of unnecessary BGP messages
  - Drawback
    - May delay the propagation of BGP messages and thus decrease the convergence time
      - For this reason, MRAI is usually disabled on iBGP sessions

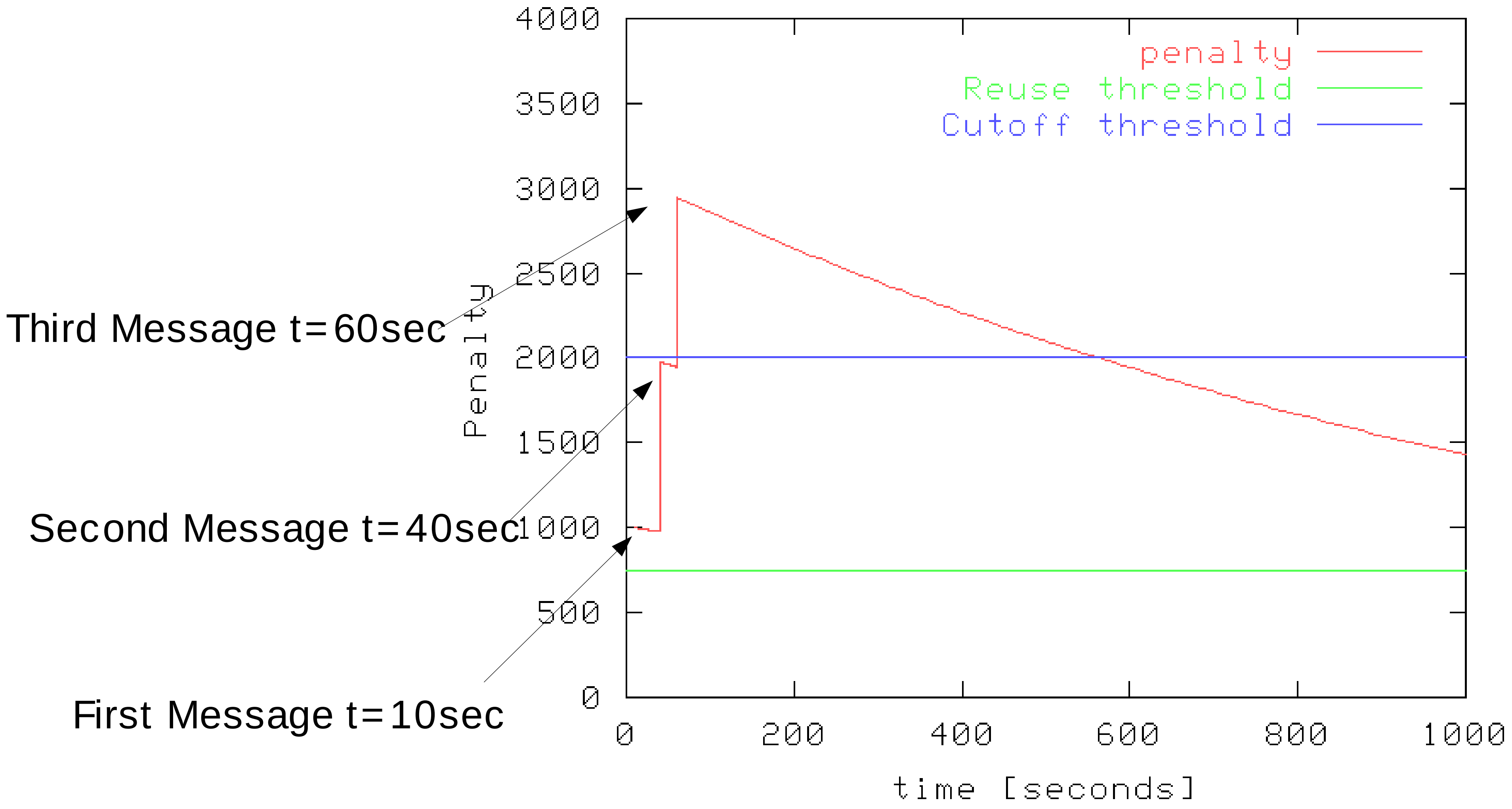
BGP dampening

- Observation
  - Most routes do not change frequently
  - A small fraction of the routes are responsible for most of the BGP messages exchanged
    - Can we penalize those unstable routes to preserve the more stable routes ?
- Principle
  - Associate a penalty counter to each route
    - Increase penalty counter each time route changes
    - Use exponential decay to slowly decrease penalty counter with time
  - Routes with a too large penalty are suppressed

BGP Dampening parameters

- Main parameters of BGP dampening
  - Penalty per BGP message
    - Penalty per withdraw message
    - Penalty per attribute change in Update message
    - Penalty per Update message
  - Cutoff threshold
    - Penalty value above which route is suppressed
  - Reuse threshold
    - Minimum penalty value required to reuse a route
  - Halftime
    - For the exponential decay
  - Maximum suppress time
    - A route cannot be suppressed longer than this time

BGP Dampening : example



Evaluation of BGP Dampening

- Advantages
  - Only penalizes unstable routes without affecting usually stable routes
- Issues
  - What are the best configurations values to use ?
    - No definite scientific answer today
  - ISPs often don't apply dampening on all sessions
    - No dampening on iBGP sessions
    - No dampening on eBGP sessions with customers
    - No dampening for the root/GTLD DNS prefixes
    - Some propose to use more aggressive dampening parameters for longer prefixes

Summary

- iBGP versus eBGP
  - EBGP distributes routes between domains
  - IBGP distributes interdomain routes inside a domain
- iBGP sessions inside a domain
  - Full mesh (unscalable)
  - Route reflectors (change iBGP processing rule)
  - Confederations (useful when merging domains)
- Scalable routing policies with communities
- The dynamics of BGP
  - A few sources produce most BGP UPDATES
  - How to reduce the churn
    - MRAI timer
    - Dampening
    - Route refresh capability