UCL

INGI

Département d'ingénierie informatique

Interdomain routing with BGP4

Part 4/5

Olivier Bonaventure

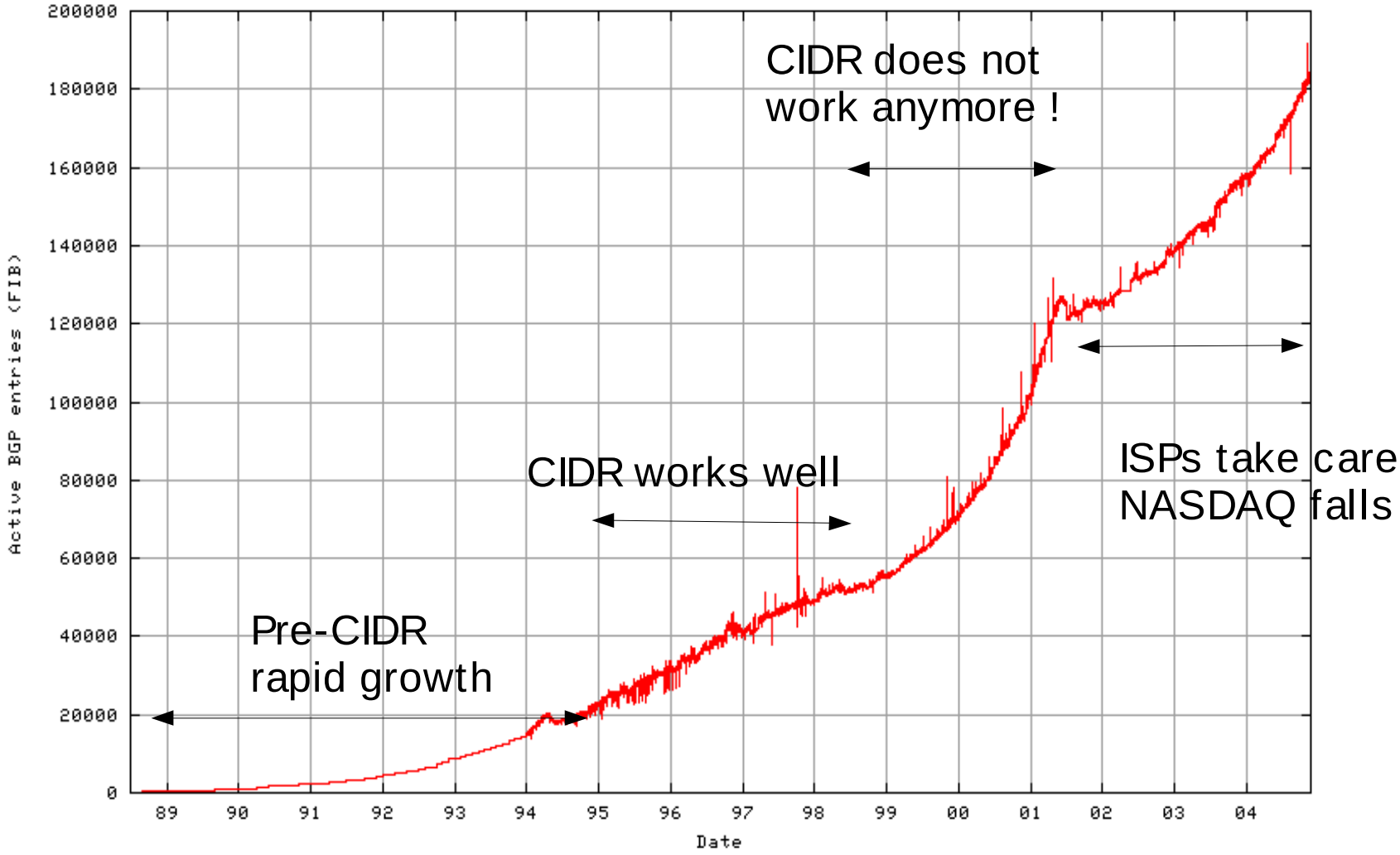
Department of Computing Science and Engineering
Université catholique de Louvain (UCL)
Place Sainte-Barbe, 2, B-1348, Louvain-la-Neuve (Belgium)

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

Outline

- Organization of the global Internet
- BGP basics
- BGP in large networks
- Interdomain traffic engineering with BGP
  - The growth of the BGP routing tables
  - The BGP decision process
  - Interdomain traffic engineering techniques
  - Case study
- BGP-based Virtual Private Networks

The growth of the BGP routing tables



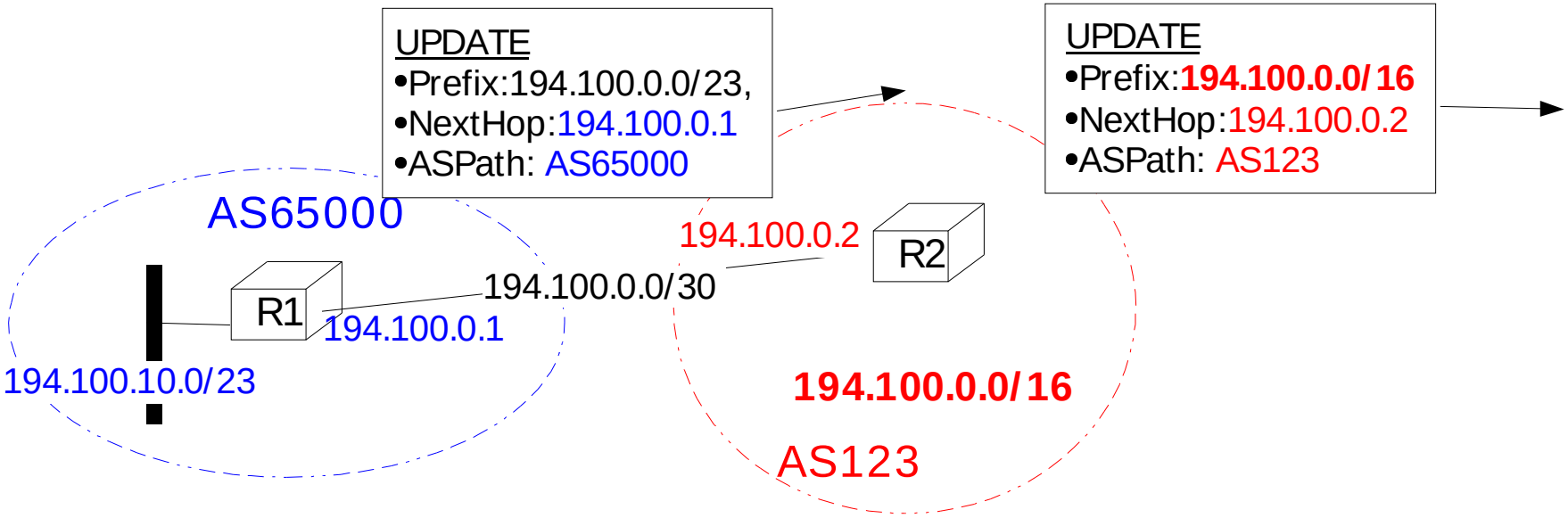
Source: http://bgp.potaroo.net , Nov. 2004

The reasons for the recent growth

- Fraction of IPv4 address space advertised
  - 24 % of total IPv4 space in 2000
  - 28 % of total IPv4 space in April 2003
  - 31% of total IPv4 space in Nov. 2004
- Increase in number of ASes
  - About 3000 ASes in early 1998
  - More than 18000 ASes in Nov 2004
  - Increase in multi-homing
    - Less than 1000 multi-homed stub ASes in early 1998
    - More than 6000 multi-homed stub ASes April 2003
- Increase in advertisement of small prefixes
  - Number of IPv4 addresses advertised per prefix
    - In late 1999, 16k IPv4 addr. per prefix in BGP tables
    - In April 2003, 8k IPv4 addr. per prefix in BGP tables

Evolution of typical stub AS

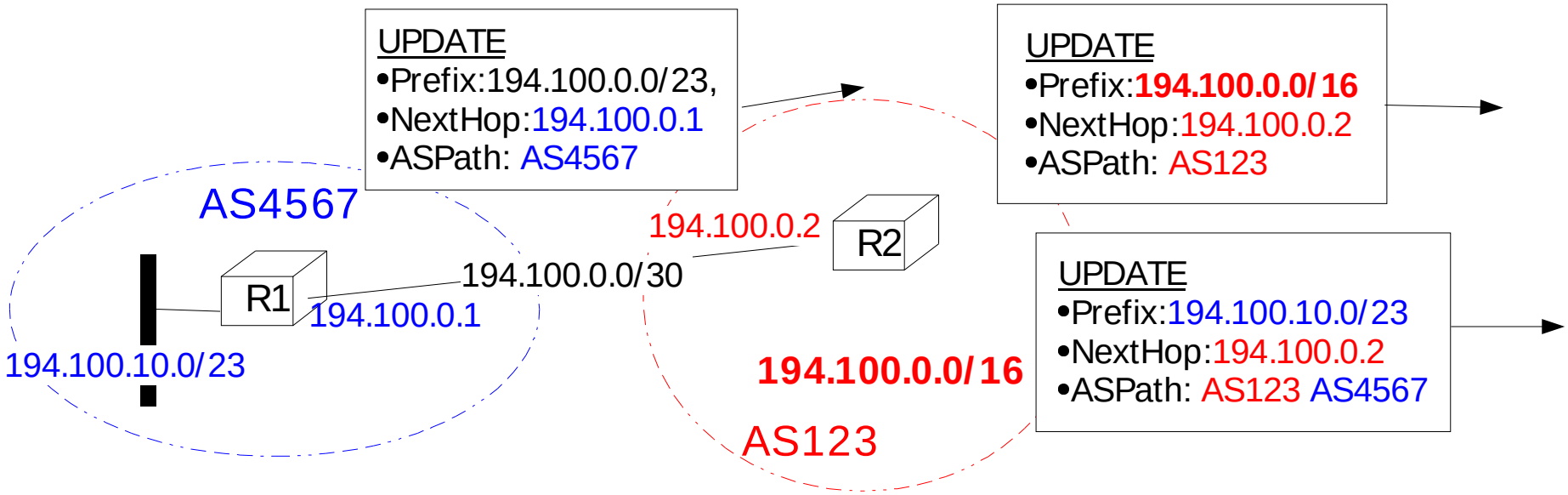
- Day one, first connection to upstream ISP
  - Stub receives address block from its ISP
  - Stub uses private AS number



  - Single homed-stub is completely hidden behind its provider
    - No impact on BGP routing table size

Evolution of typical stub AS (2)

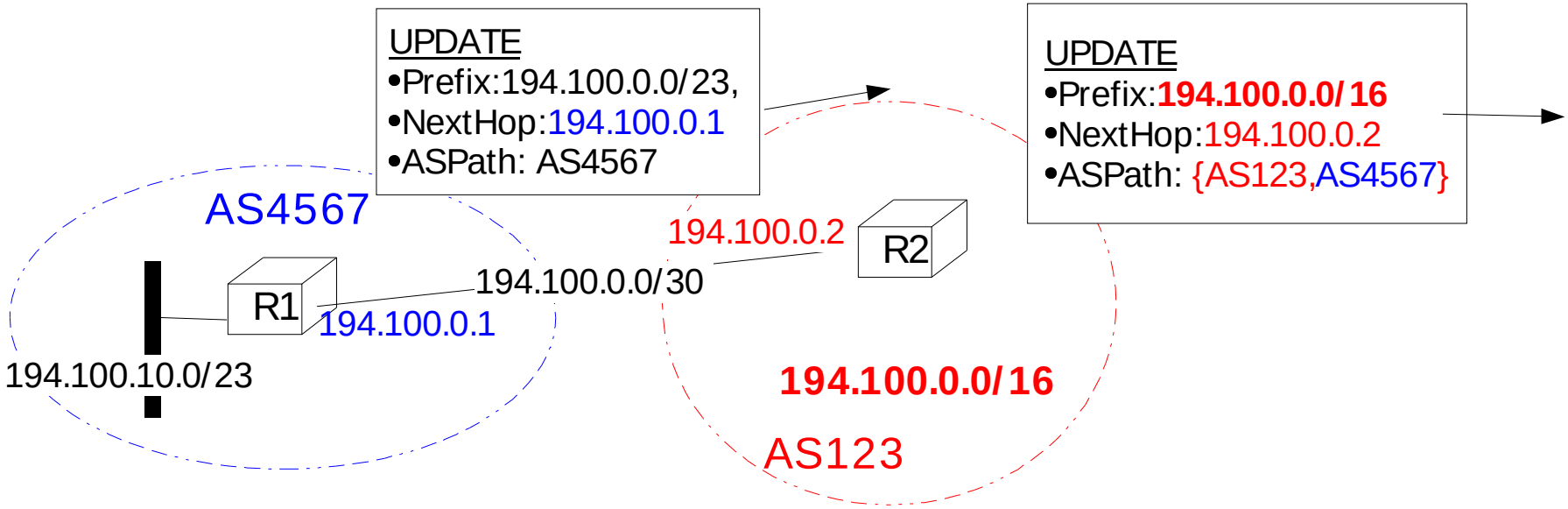
- Day two, stub AS expects to become multi-homed in near future and obtains official AS#



- Advantage
  - Simple to configure for AS123
- Drawback
  - Increases the size of all BGP routing tables

Aggregating routes

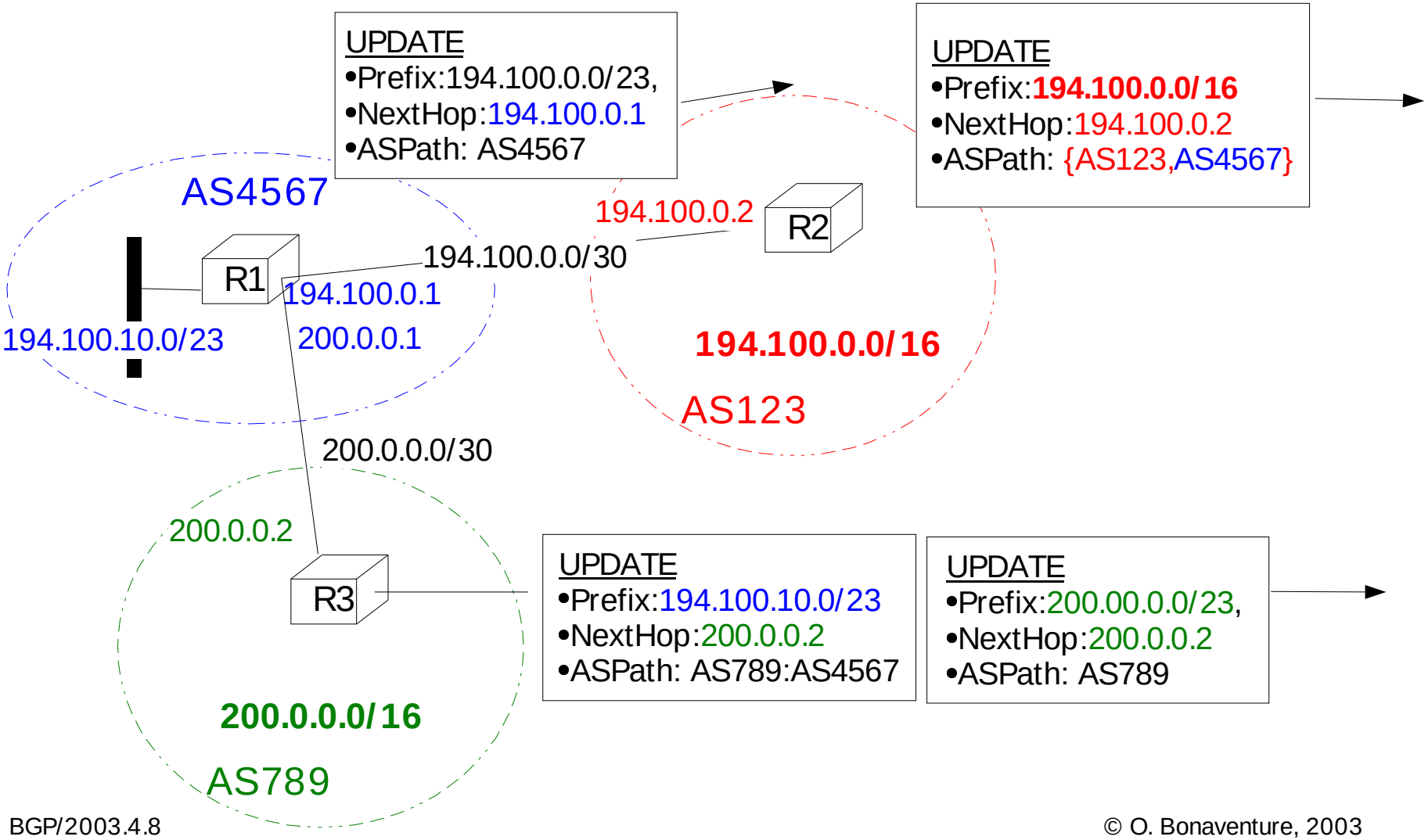
- BGP is able to aggregate received routes even if some ASPath information is lost



- One AS_SET contains several AS#
  - counts as one AS when measuring length of AS Path
  - used for loop detection, but ASPath may become very long when one provider has many clients to aggregate

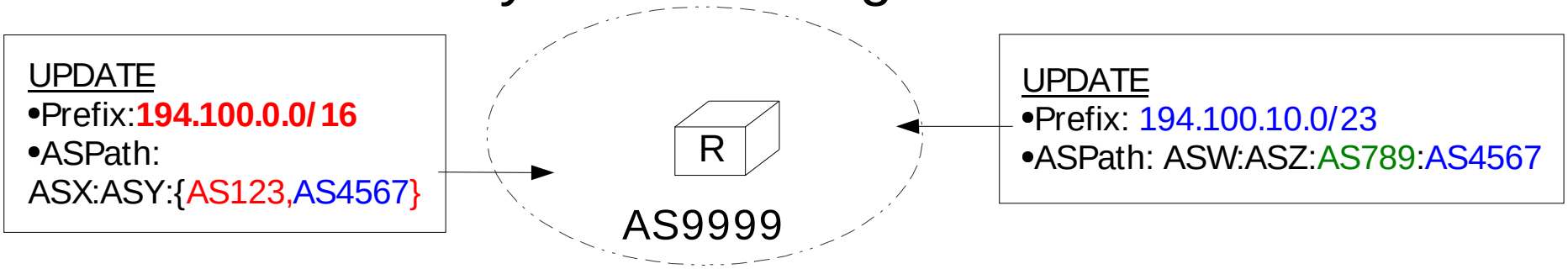
A dual-homed stub ISP

- Day three, stub AS is multi-homed



A dual-homed stub ISP (2)

- Drawback of this solution
  - Consider any AS receiving those routes



Routing table
194.100.10.0/23 Path:ASW:ASZ:AS789:AS4567
194.100.0.0/16 Path: ASX:ASY:{AS123,AS4567}

- Consequences
  - All traffic to 194.100.10.0/23 will be sent on non-aggregated path since it is the most specific !!!
  - AS123 might stop aggregating its customer prefixes, otherwise its customers will not receive packets
  - The global BGP routing tables are 50% larger than their optimal size if aggregation was perfectly used
  - Less than 7% of the BGP routes are aggregates

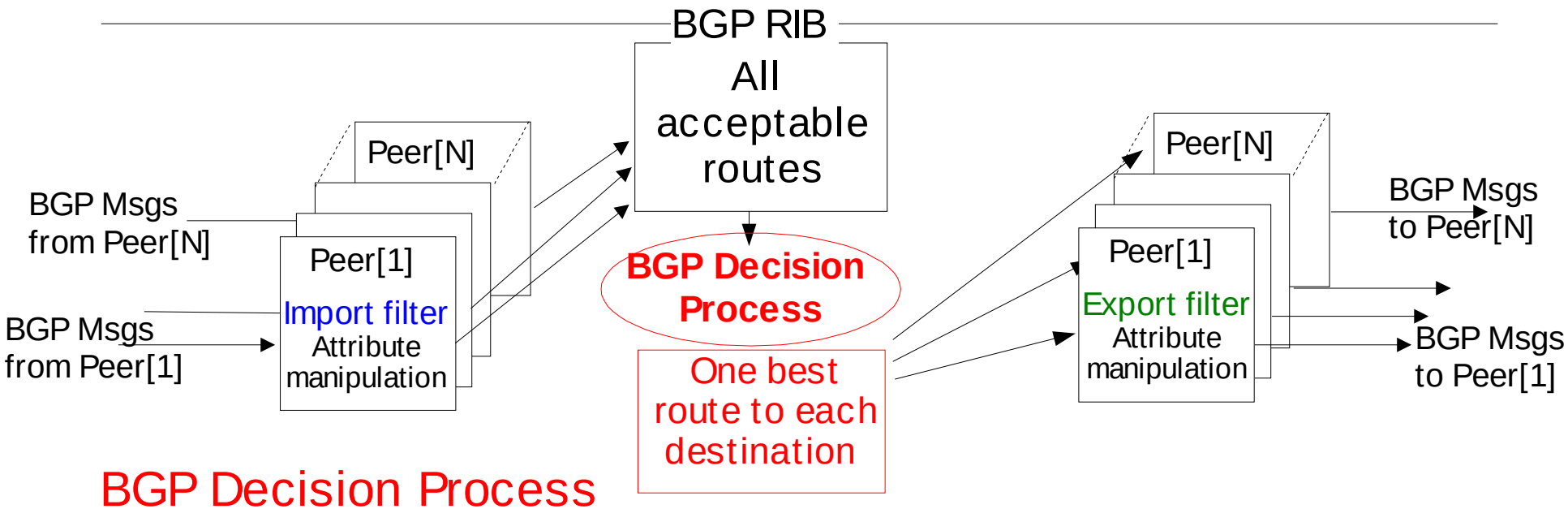
How to limit the growth of the BGP tables ?

- Long term solution
  - Define a better multihoming architecture
    - Will be difficult with IPv4
    - Work is ongoing to develop a better multihoming for IPv6
- Current «solution» (aka quick hack)
  - Some ISPs filter routes towards too long prefixes
  - Two methods are used today
    - Ignore routes with prefixes longer than *p* bits
      - Usual values range between 22 and 24
    - Ignore routes that are longer than the allocation rules used by the Internet registries (RIPE, ARIN, APNIC)
      - Ignore prefixes longer than /16 in class B space
      - Ignore RIPE prefixes longer than RIPE's minimum allocation (/20)
  - Consequence
    - Some routes are not distributed to the global Internet !

Outline

- Organization of the global Internet
- BGP basics
- BGP in large networks
- Interdomain traffic engineering with BGP
  - The growth of the BGP routing tables
  - The BGP decision process
  - Interdomain traffic engineering techniques
  - Case study
- BGP-based Virtual Private Networks

The BGP decision process

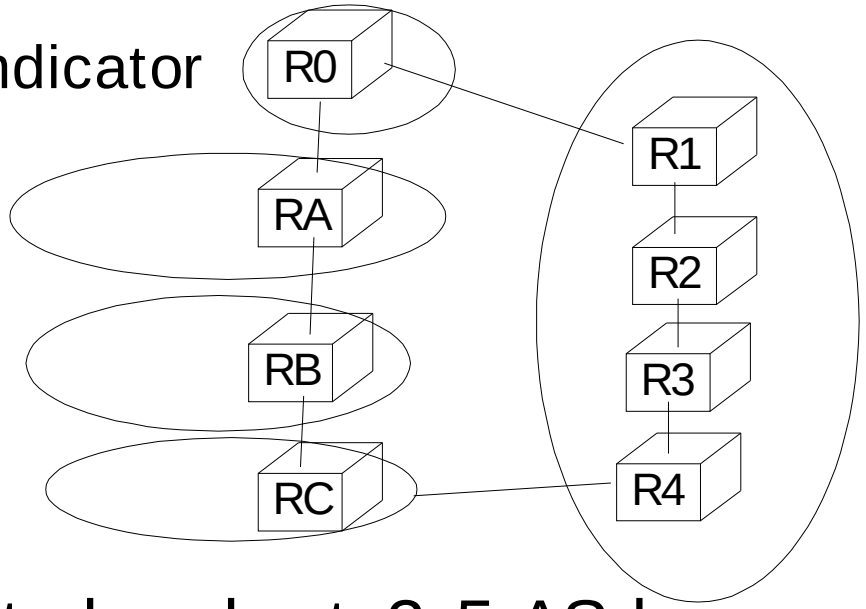


BGP Decision Process

- *Ignore routes with unreachable nexthop*
- Prefer routes with highest local-pref
- Prefer routes with shortest ASPath
- Prefer routes with smallest MED
- Prefer routes learned via eBGP over routes learned via iBGP
- Prefer routes with closest next-hop
- Tie breaking rules
  - Prefer Routes learned from router with lowest router id

The shortest AS-Path step in the BGP decision process

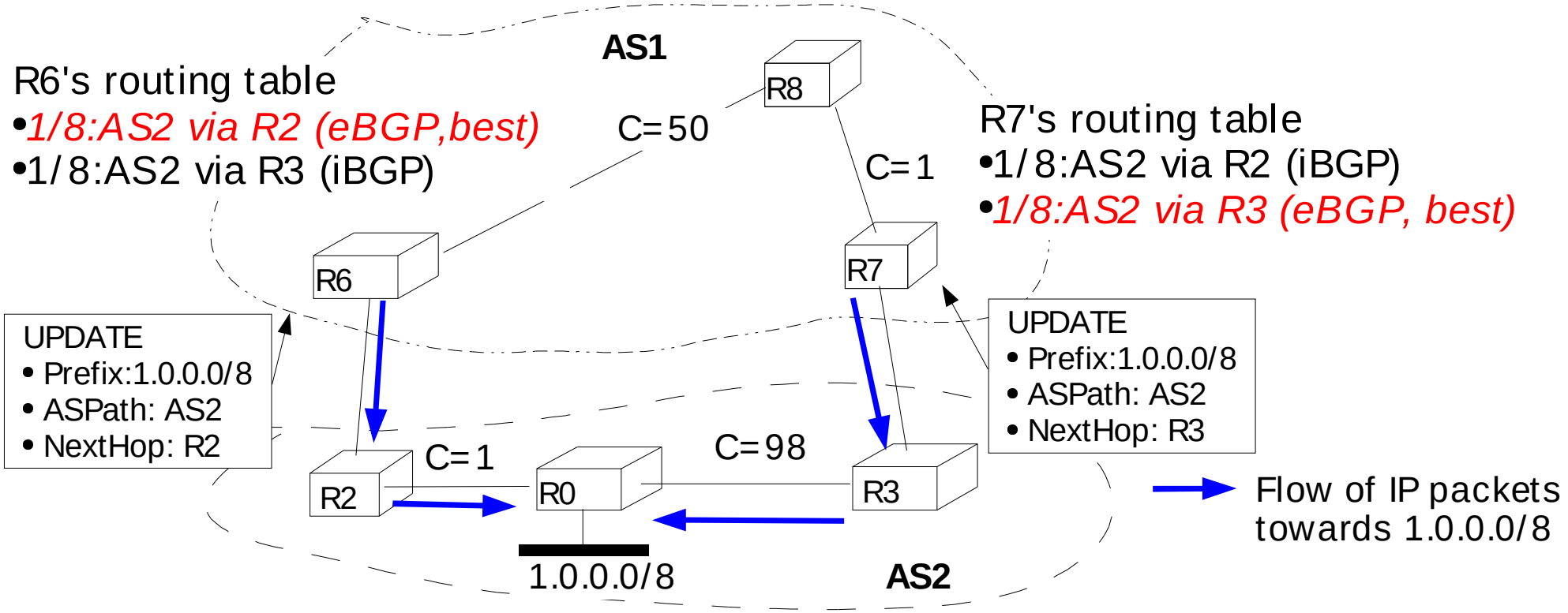
- Motivation
  - BGP does not contain a real " metric"
  - Use length of AS-Path as an indication of the quality of routes
    - Not always a good indicator



- Consequence
  - Internet paths tend to be short, 3-5 AS hops
  - Many paths converge at Tier-1 ISPs and those ISPs carry lots of traffic

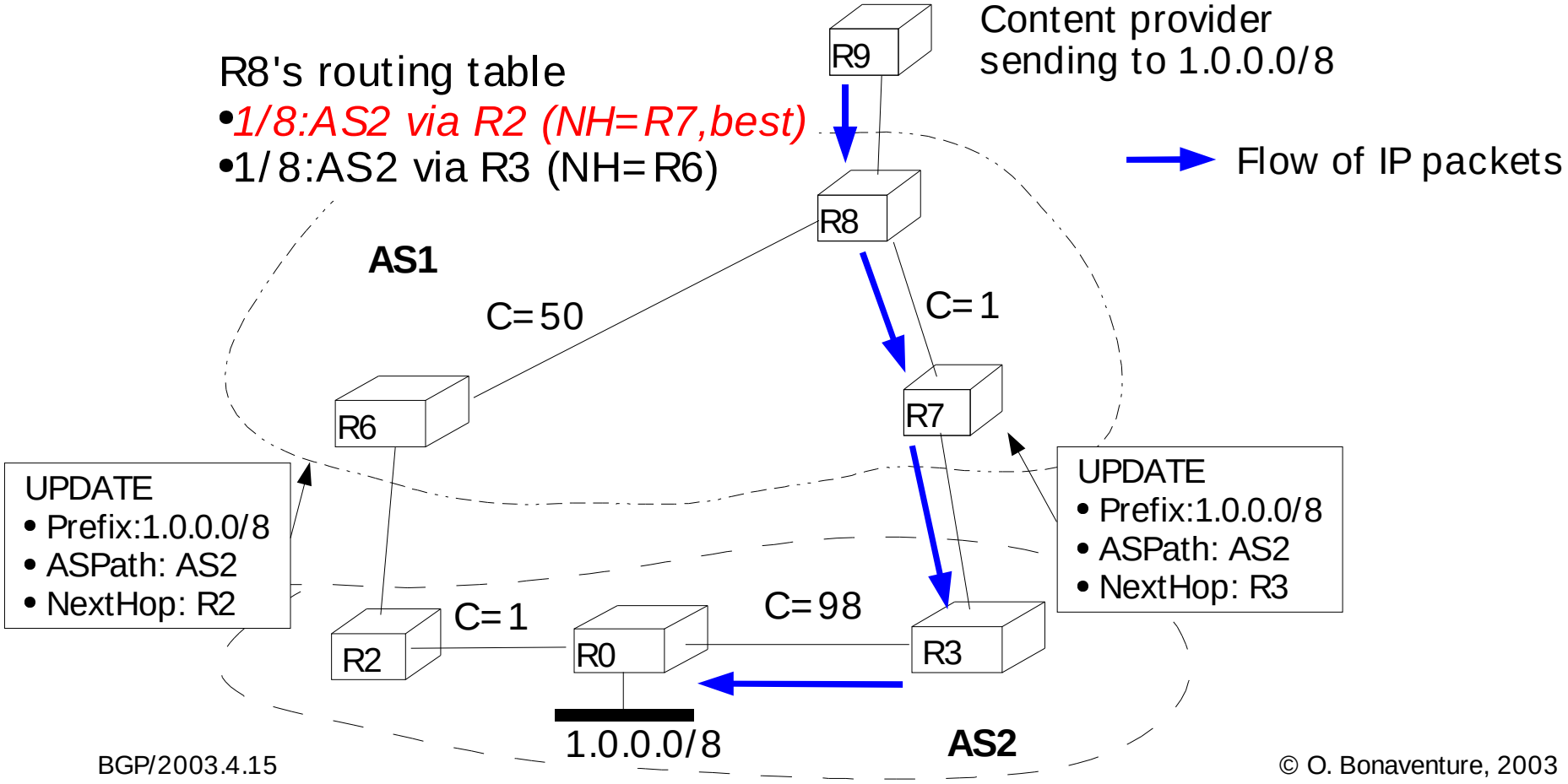
The prefer eBGP over iBGP step in the BGP decision process

- Motivation : hot potato routing
  - A router should try to get rid of packets sent to external domains as soon as possible



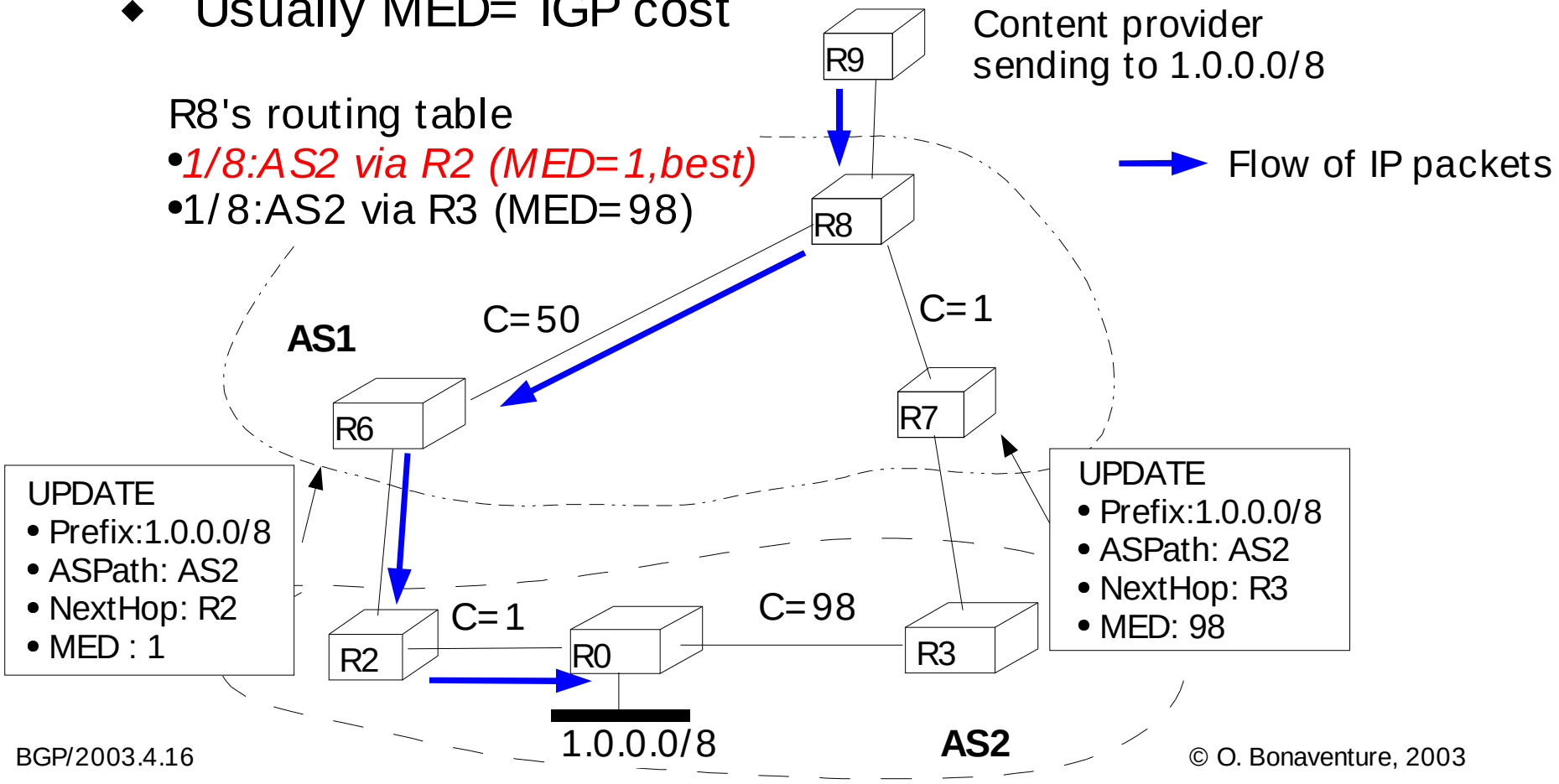
The closest nexthop step in the BGP decision process

- Motivation : hot potato routing
  - A router should try to get rid of packets sent to external domains as soon as possible



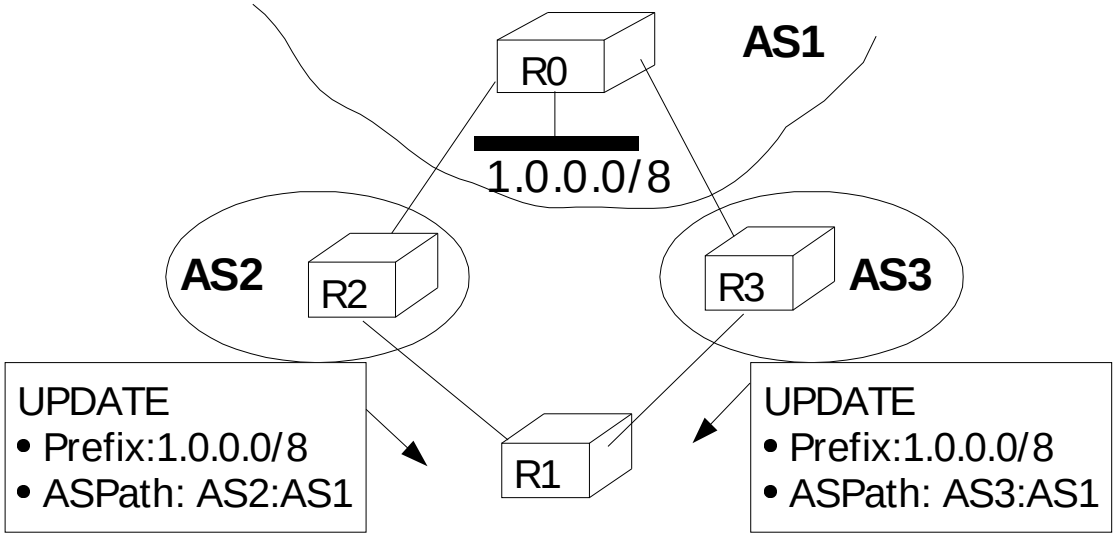
The lowest MED step in the BGP decision process

- Motivation : cold potato routing
  - In a multi-connected AS, indicate which entry border router is closest to the advertised prefix
    - Usually MED= IGP cost



The lowest router id step in the BGP decision process

- Motivation
  - A router must be able to determine *one* best route towards each destination prefix
    - A router may receive several routes with comparable attributes towards one destination



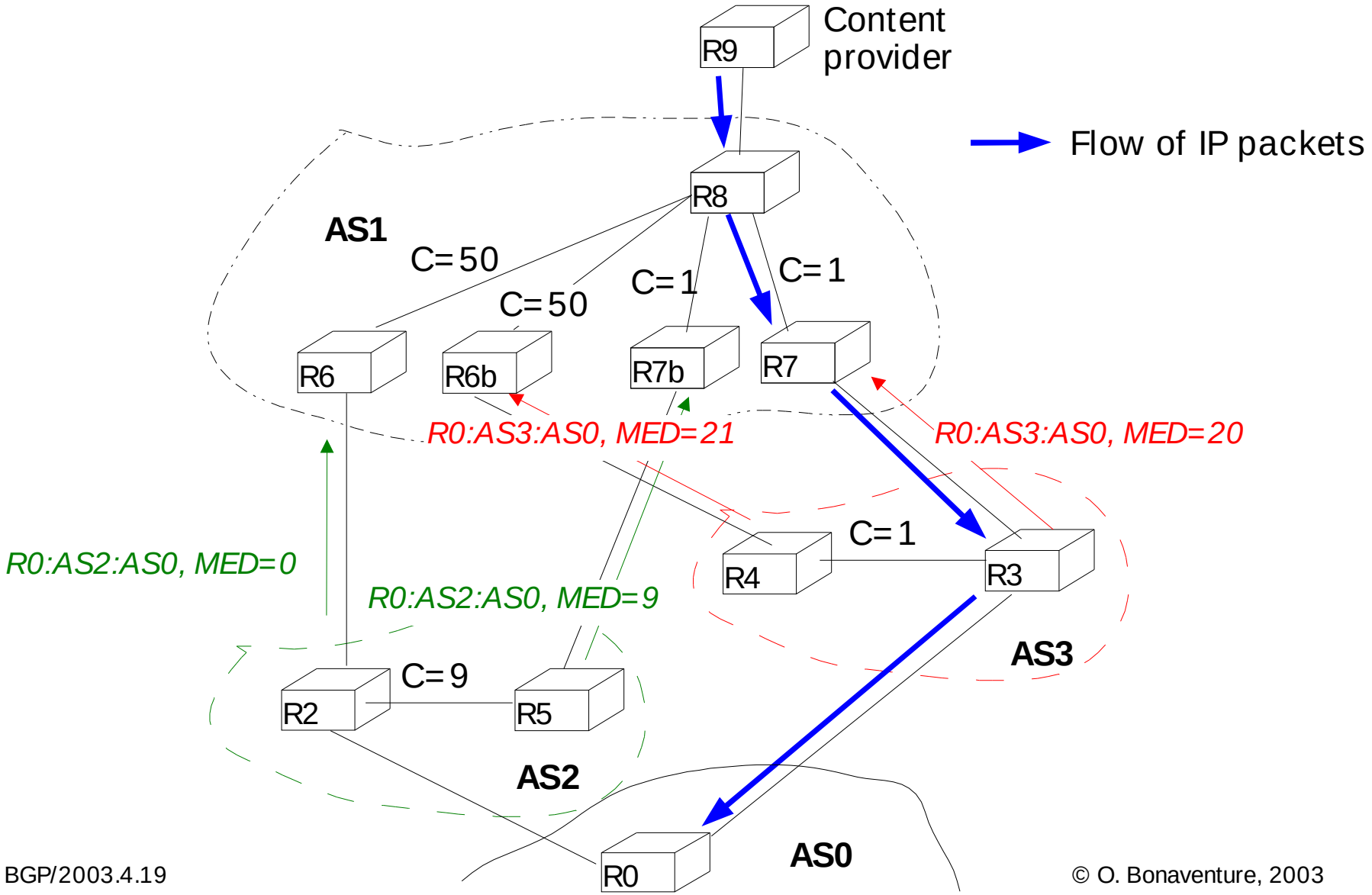
- Consequence
  - A router with a low IP address will be preferred

More on the MED step in the BGP decision process

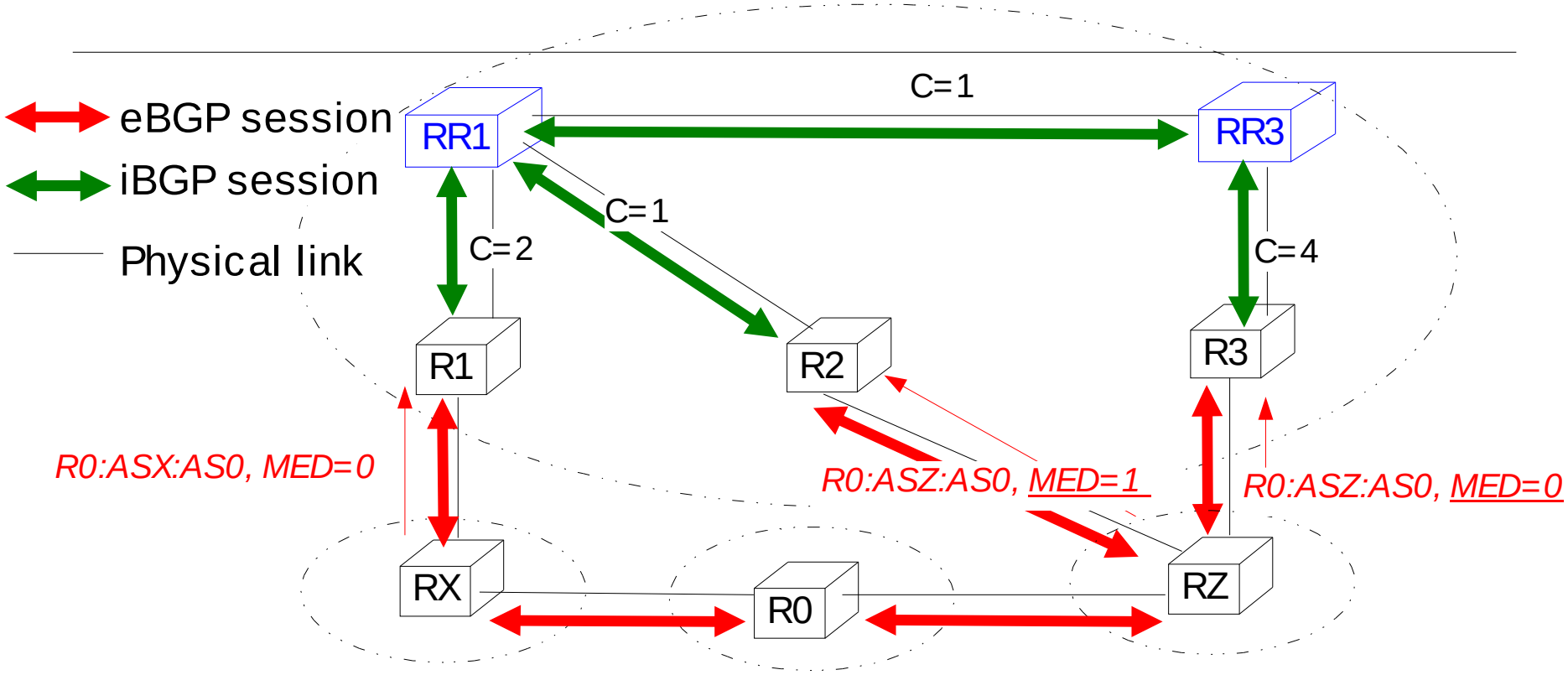
- Unfortunately, the processing of the MED is more complex than described earlier
- Correct processing of the MED
  - MED values can only be compared between routes receiving from the SAME neighboring AS
    - Routes which do not have the MED attribute are considered to have the lowest possible MED value.
  - Selection of the routes containing MED values

```
for m = all routes still under consideration
 for n = all routes still under consideration
   if (neighborAS(m) == neighborAS(n)) and
        (MED(n) < MED(m))
        {
          remove route m from consideration
        }
```

Why such a complex MED step ?



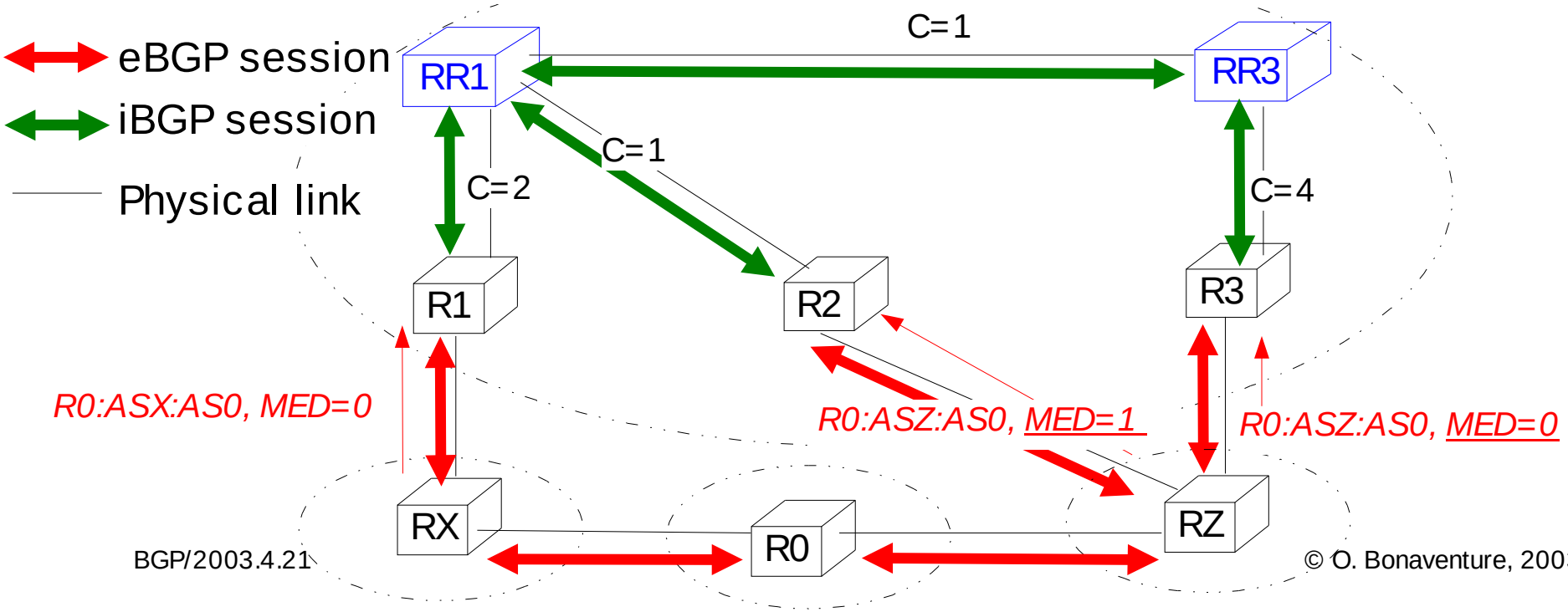
Route oscillations with MED



- Consider a single prefix advertised by R0 in AS0
  - R1, R2 and R3 always prefer their direct eBGP path
  - Due to the utilization of route reflectors, RR1 and RR3 only know a subset of the three possible paths
    - This limited knowledge is the cause of the oscillations

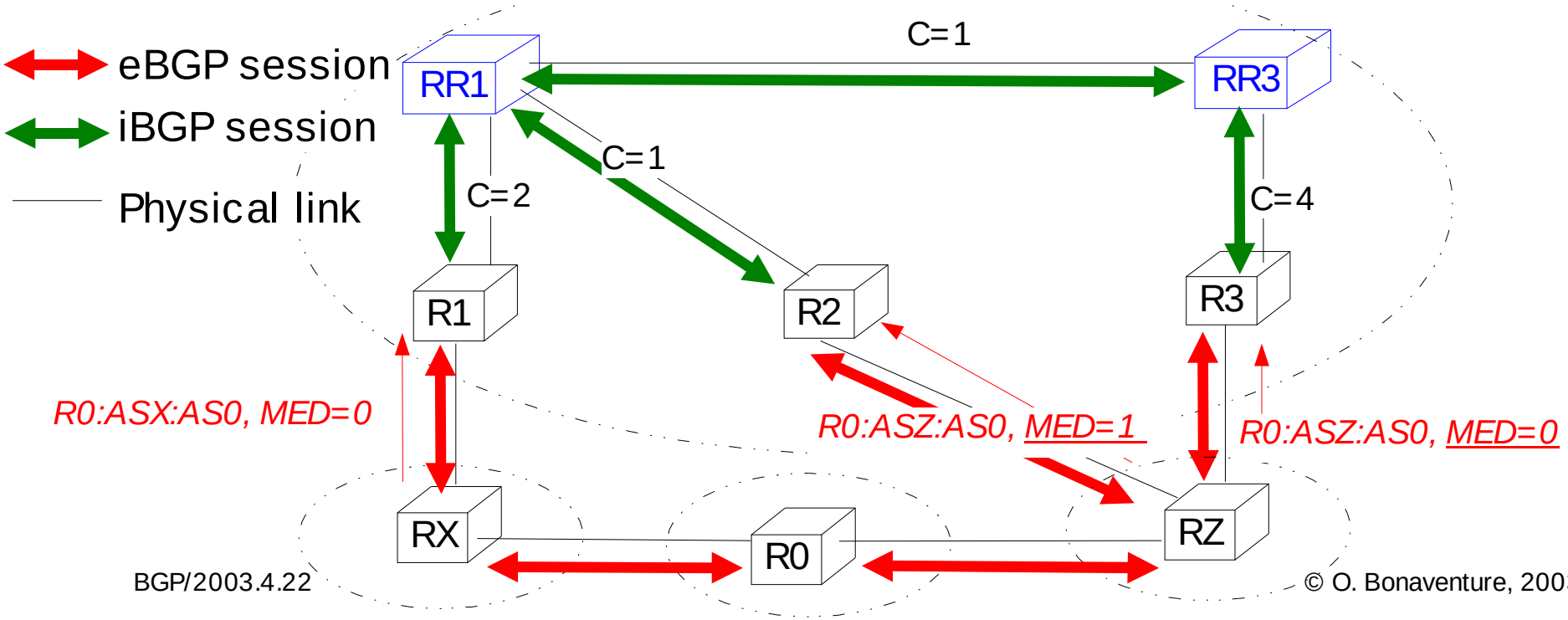
Route oscillations with MED (2)

- RR3's best path selection
  - If RR3 only knows the R3-RZ path, this path is preferred and advertised to RR1
  - RR3 knows the R1-RX and R3-RZ paths, R1-RX is best (IGP cost) and RR3 doesn't advertise a path to RR1
  - If RR3 knows the R2-RZ and R3-RZ paths, RR3 prefers the R3-RZ path (MED) and R3-RZ is advertised to RR1

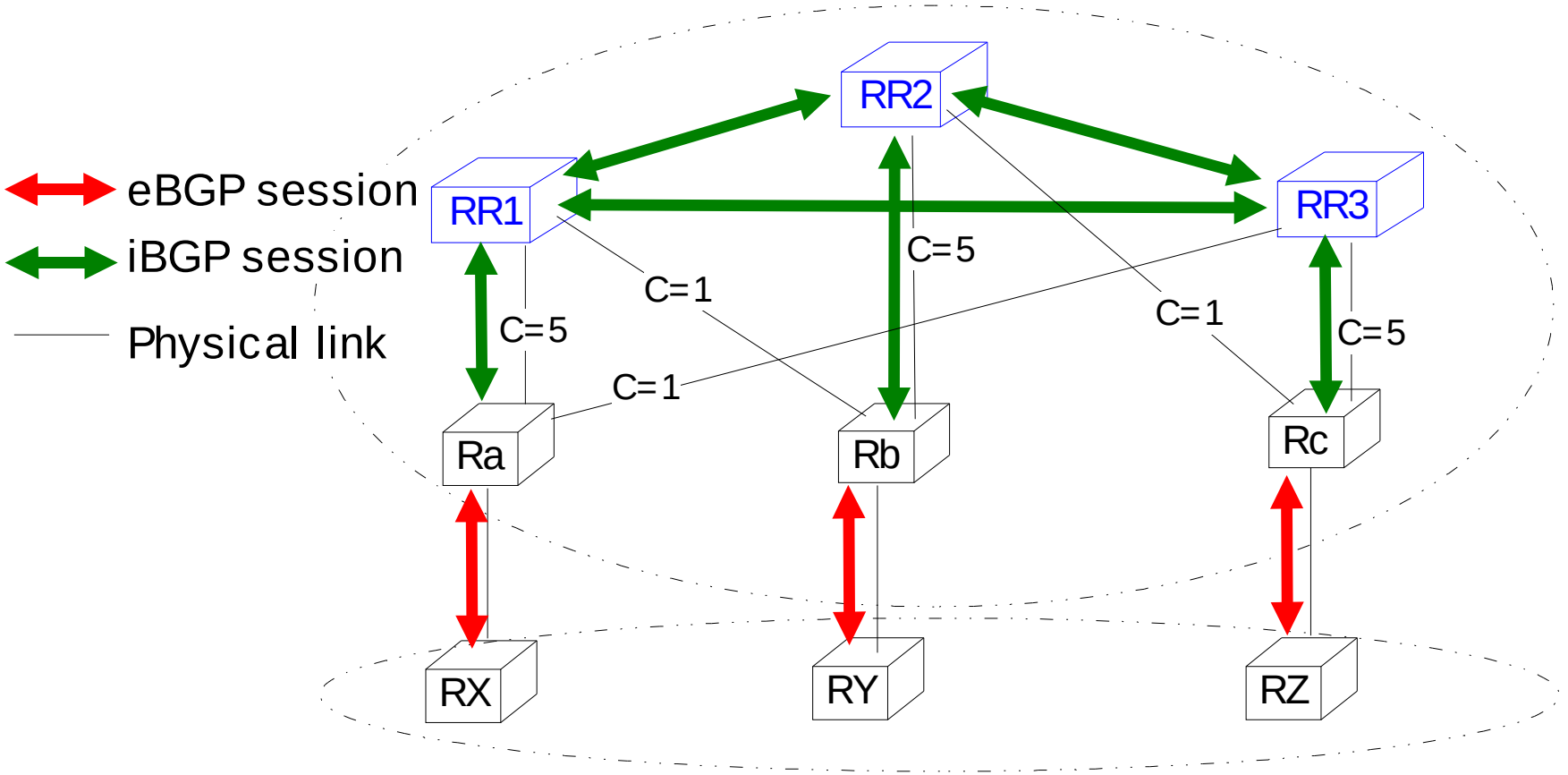


Route oscillations with MED (3)

- RR1's best path selection
  - If RR1 knows the R1-RX, R2-RZ and R3-RZ paths, R1-RX is preferred and RR1 advertises this path to RR3
    - But if RR1 advertises R1-RX, RR3 does not advertise any path !
  - If RR1 knows the R1-RX and R2-RZ paths, RR1 prefers the R2-RZ path and advertises this path to RR3
    - But if RR1 advertises R2-RZ, RR3 prefers and advertises R3-RZ !

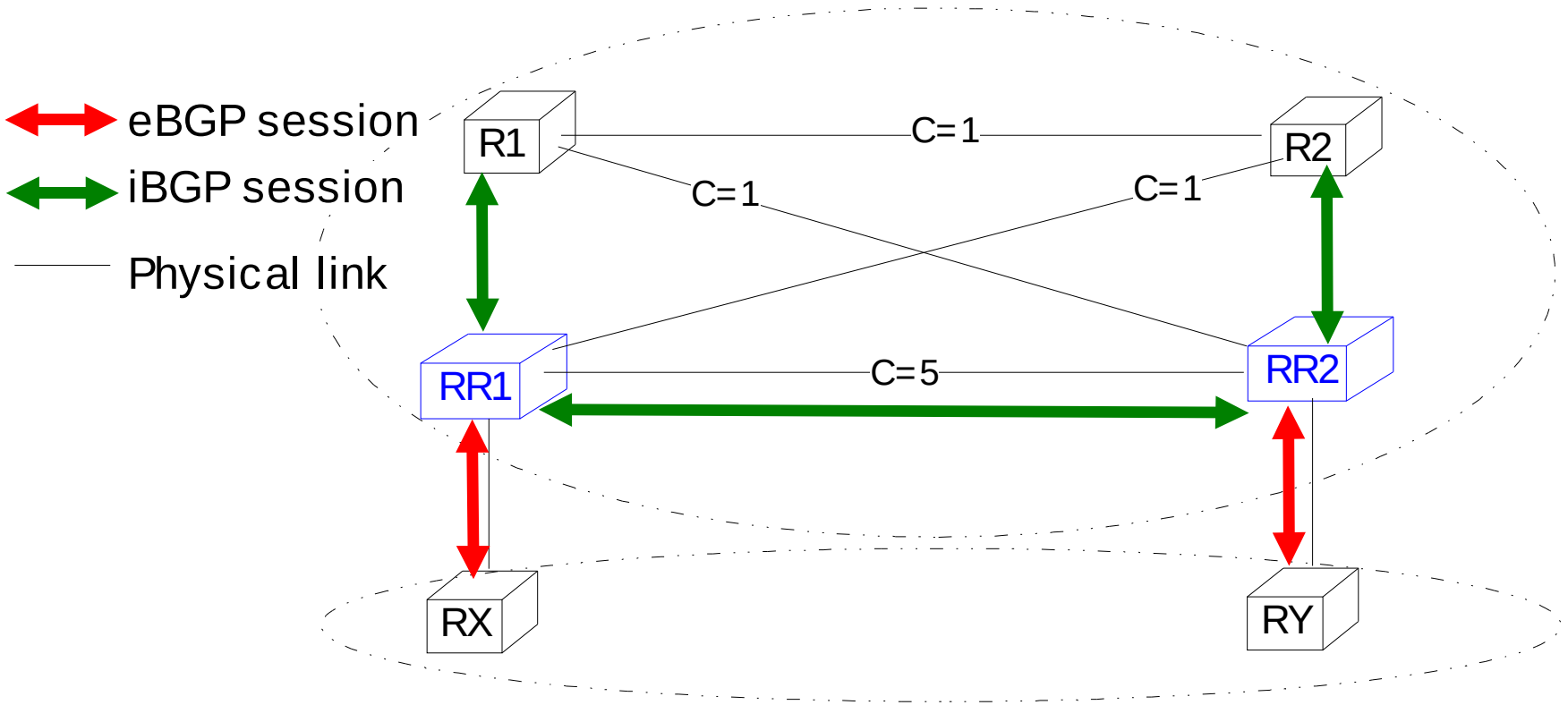


Other problems with Route Reflectors



- Consider one prefix advertised by RX,RY,RZ
  - Ra, Rb, and Rc will all prefer their direct eBGP path
  - RR1, RR2 and RR3 will never reach an agreement

Forwarding problems with Route Reflectors



- Consider a prefix advertised by RX and RY
  - BGP routing will converge
    - RR1 (and R1) prefer path via RX, RR2 (and R2) prefer path via RY
  - But forwarding of IP packets will cause loop !
    - R1 sends packets towards prefix via R2 (to reach RX, its best path)
    - R2 sends packets towards prefix via R1 (to reach RY, its best path)

Outline

- Organization of the global Internet
- BGP basics
- BGP in large networks
- Interdomain traffic engineering with BGP
  - The growth of the BGP routing tables
  - The BGP decision process
  - Interdomain traffic engineering techniques
  - Case study
- BGP-based Virtual Private Networks

Interdomain traffic engineering

- Objectives of interdomain traffic engineering
  - Minimize the interdomain cost of your network
  - Optimize performance
    - prefer to send/receive packets over low delay paths for VoIP
    - prefer to send/receive packets over high bandwidth paths
  - Balance the traffic between several providers

- How to engineer your interdomain traffic ?
  - Carefully select your main provider(s)
  - Negotiate peering agreements with other domains at public interconnection points
  - Tune the BGP decision process on your routers
  - Tune your BGP advertisements

Traffic engineering prerequisite

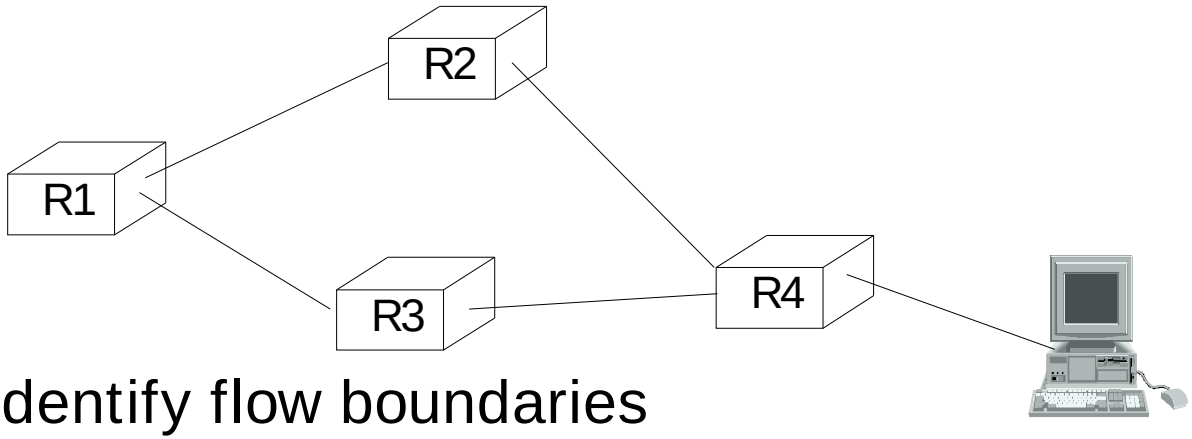
- To engineer the packet flow in your network... you first need to know :
  - amount of packets entering your network
    - preferably with some information about their source (and destination if you provide a transit service)
  - amount of packets leaving your network
    - preferable with some information about their destination (and source if you provide a transit service)

- How to obtain this information in an accurate and cost effective manner ?

Link-level traffic monitoring

- Principle
  - rely on SNMP statistics maintained by each router for each link
  - management station polls each router frequently
- Advantages
  - Simple to use and to deploy
  - Tools can automate data collection/presentation
  - Rough information about network load
- Drawbacks
  - No addressing information
  - Not always easy to find the cause of congestion

Flow-level traffic capture



- Principle
  - routers identify flow boundaries
    - does not cause huge problems on cache-based routers
  - Layer-3 flows
    - IP packets with same source (resp. destination) prefix
    - IP packets with same source (resp. destination) AS
    - IP packets with same IGP (resp. BGP) next hop
  - Layer-4 flows
    - one TCP connection corresponds to one flow
    - UDP flows
  - routers forwards this information inside special packets to monitoring workstation

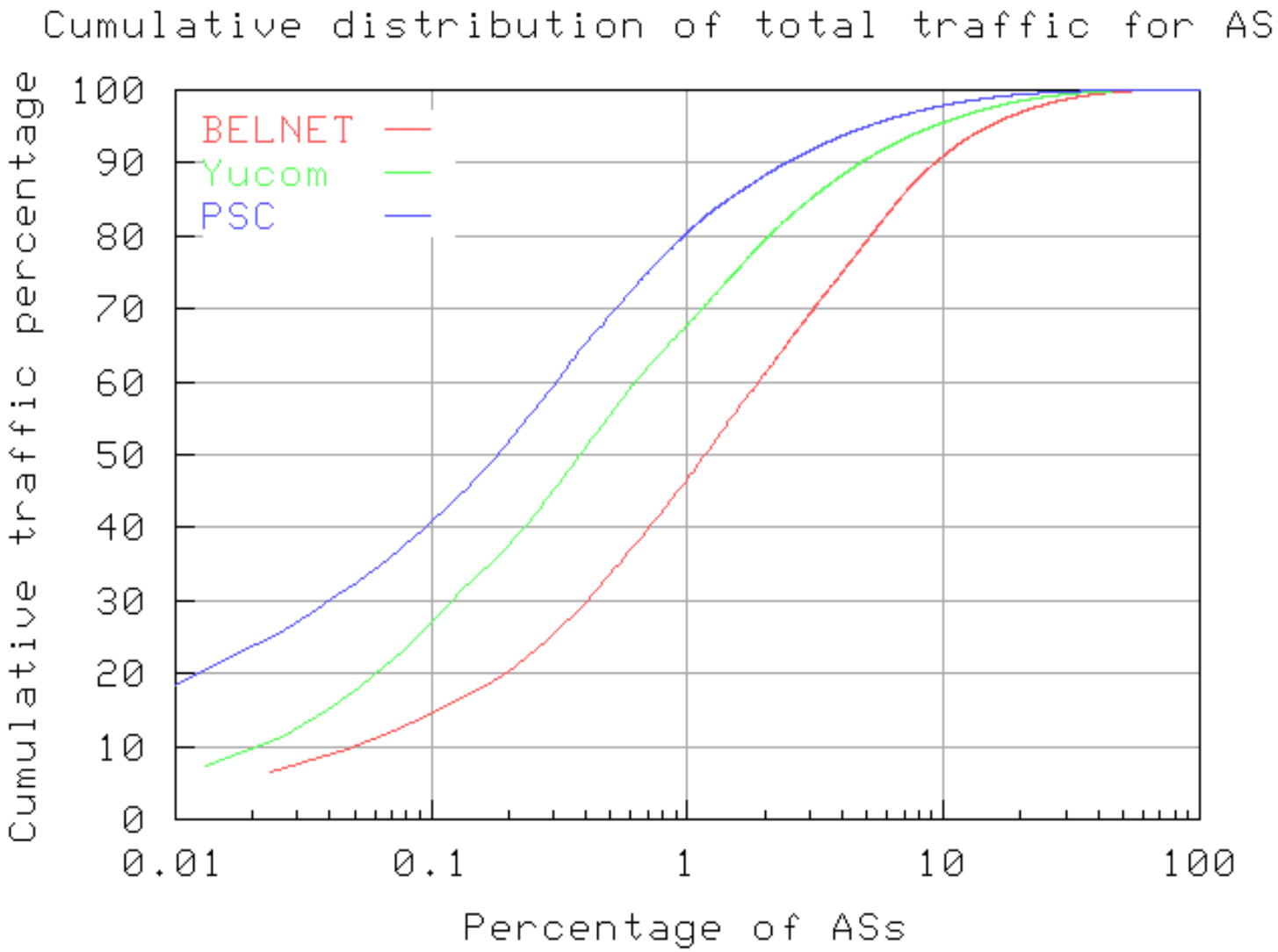
Flow level traffic capture (3)

- Advantages
  - provides detailed information on the traffic carried out on some links
- Drawbacks
  - flow information needs to be exported to monitoring station
    - information about one flow is 30 - 50 bytes
    - average size of HTTP flow is 15 TCP packets
  - CPU load on high speed on routers
    - not available on some router platforms
  - Disk and processing requirements on monitoring workstation

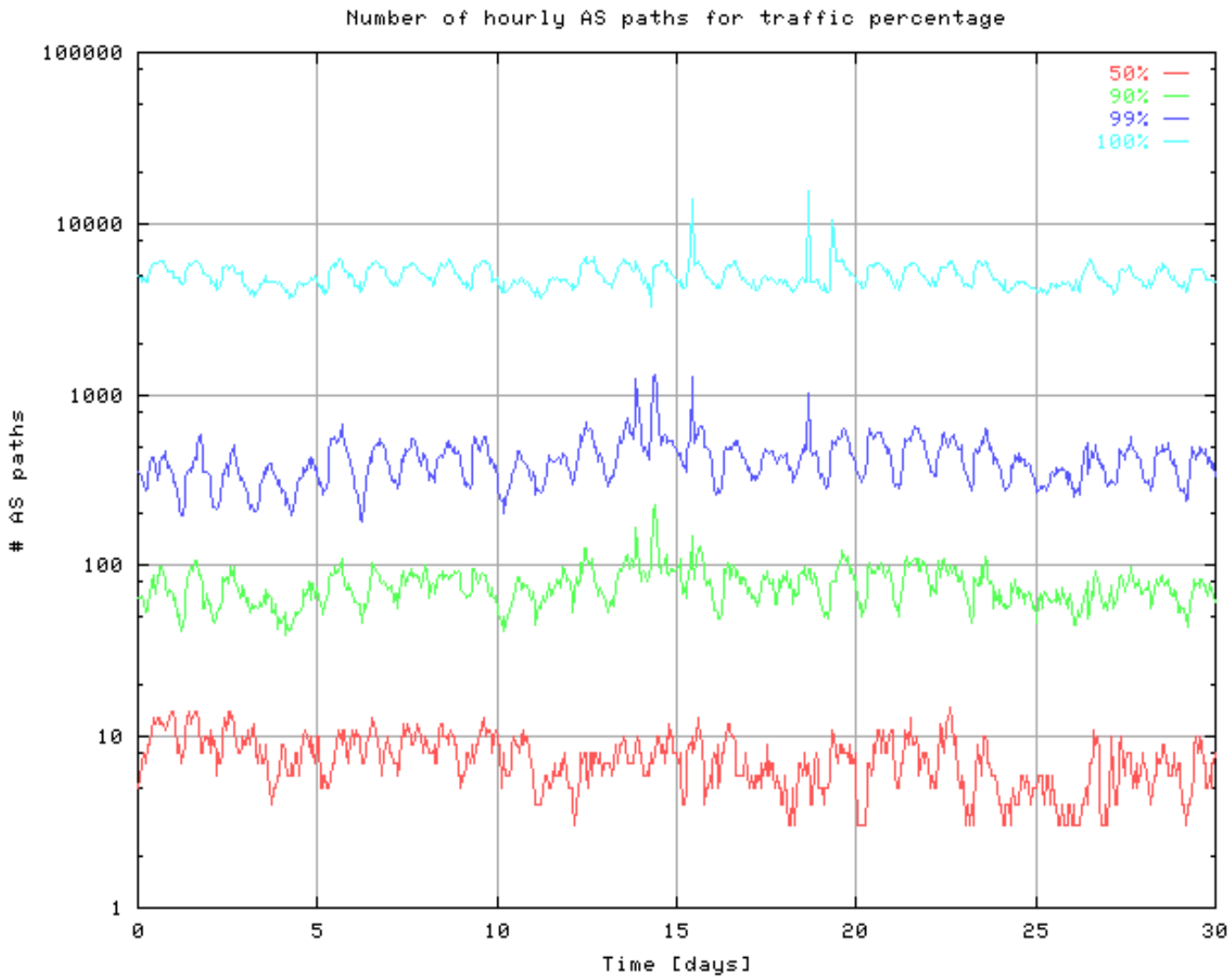
Netflow

- Industry-standard flow monitoring solution
  - Netflow v5
    - Router exports per layer-4 flow summary
      - Timestamp of flow start and finish
      - Source and destination IP addresses
      - Number of bytes/packets, IP Protocol, TOS
      - Input and output interface
      - Source and destination ports, TCP flags
      - Source and destination AS and netmasks
  - Netflow v8
    - Router performs aggregation and exports summaries
      - AS Matrix
        - interesting to identify interesting peers
      - Prefix Matrix
        - SourcePrefixMatrix, DestinationPrefixMatrix, PrefixMatrix
        - provides more detailed information than ASMatrix

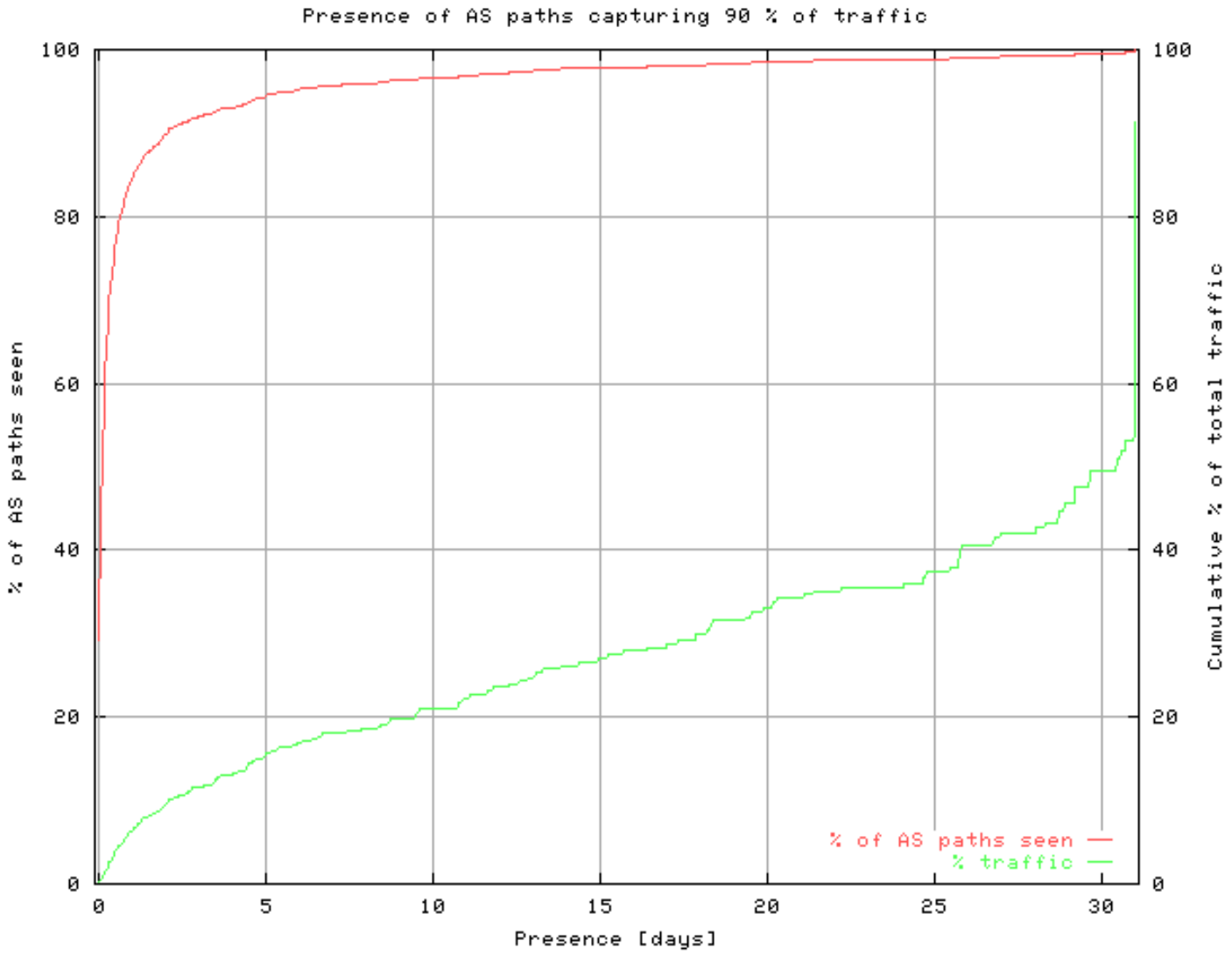
Characteristics of interdomain traffic



Topological distribution of the traffic sent by a stub during one month



Topological dynamics of the traffic sent by a stub during one month

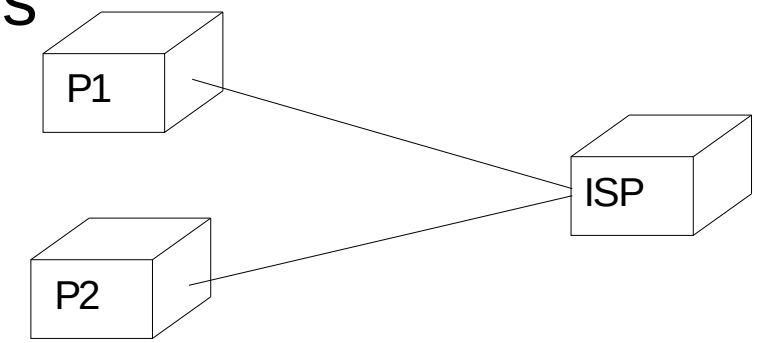


The provider selection problem

- How does an ISP select a provider ?
  - Economical criteria
    - Cost of link
    - Cost of traffic
  - Quality of the BGP routes announced by provider
    - Number of routes announced by provider
    - Length of the routes announced by provider
  - Often, ISPs have two upstream providers for technical and economical redundancy reasons

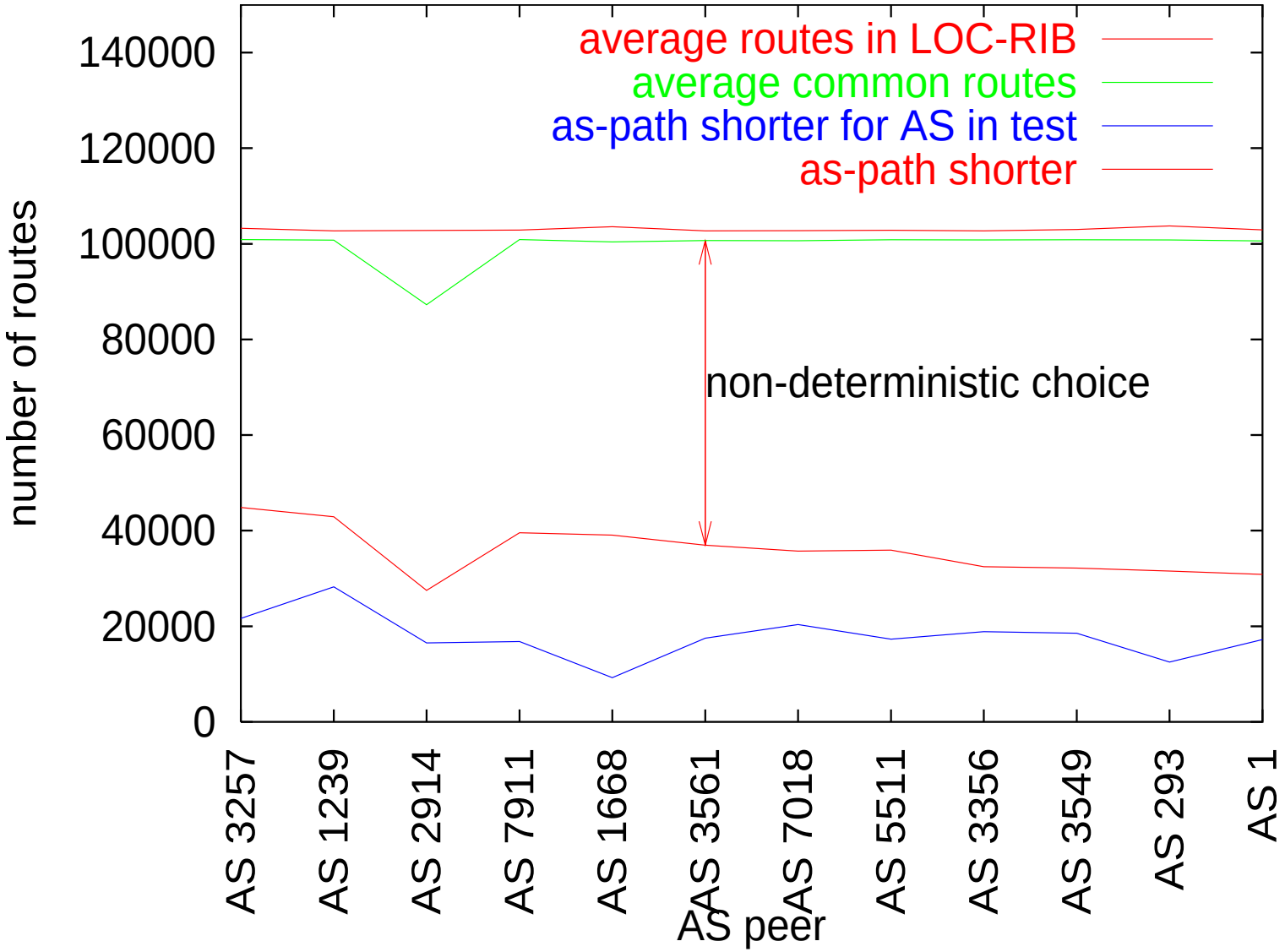
An experiment in provider selection

- Principle
  - Obtain BGP routing tables from several providers
    - 12 large providers peering with routeviews
  - Simulate the connection of an ISP to 2 of those providers



  - Rank providers based on the routes selected by the BGP decision process of the simulated ISP

Selection among the 12 largest providers

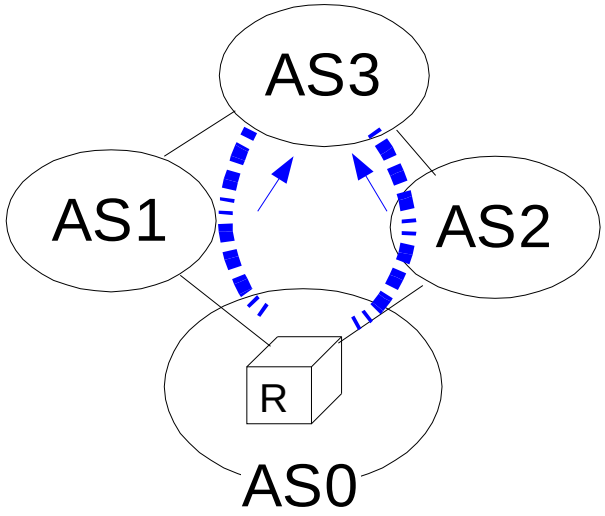


Tuning BGP to ...
control the outgoing traffic

- Principle
  - To control its outgoing traffic, a domain must tune the BGP decision process on its own routers
- How to tune the BGP decision process ?
  - Filter some routes learned from some peers
  - `local-pref`
    - usual method of enforcing economical relationships
  - MED
    - usually, MED value is set when sending a route
    - but some routers allow to insert a MED in a received route
      - allows to prefer routes over others with same AS Path length
  - IGP cost to nexthop
    - setting of IGP cost for intradomain traffic engineering
  - Several routes in fowarding table instead of one

BGP Equal Cost MultiPath

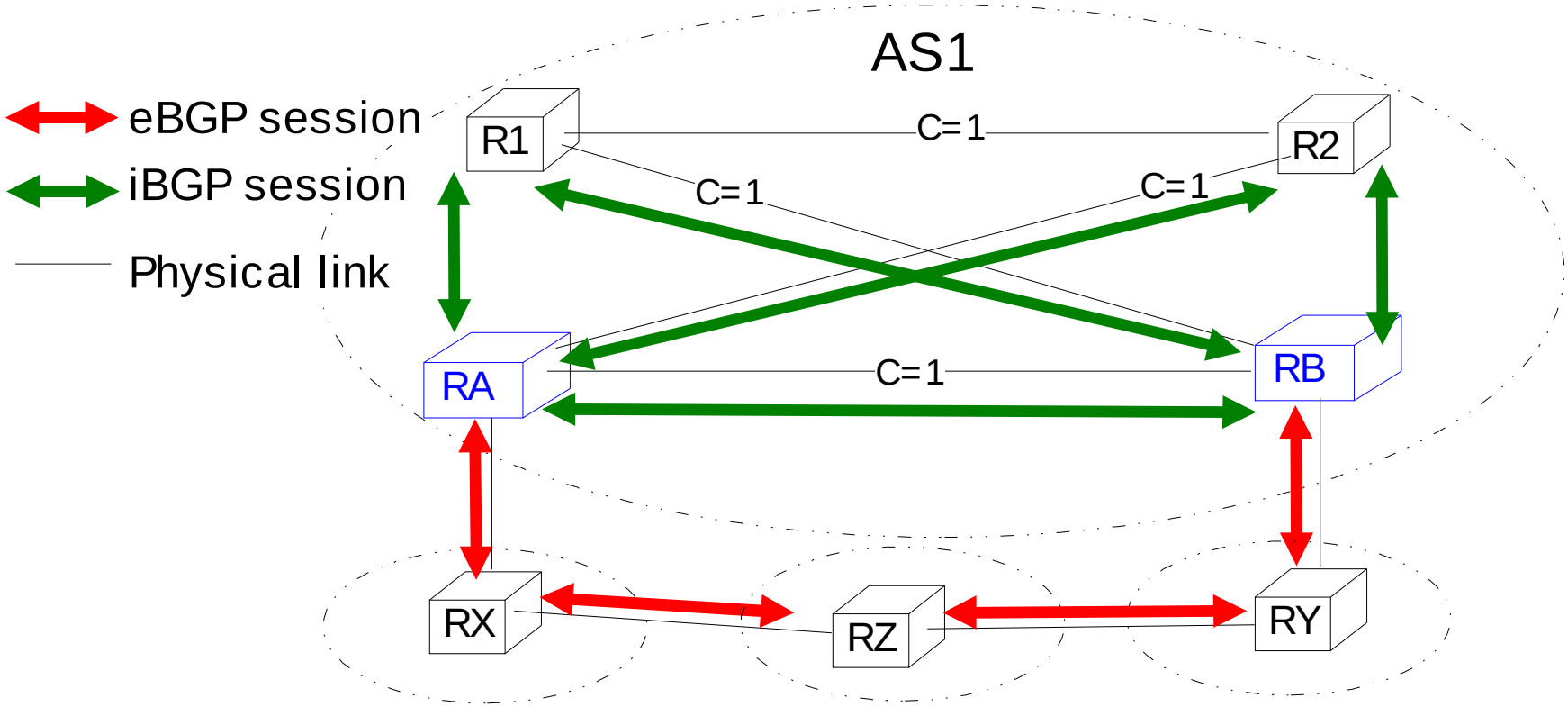
- Principle
  - Allow a BGP router to install several paths towards each destination in its forwarding table
  - Load-balance the traffic over available paths



- Issues
  - Which AS Path will be advertised by AS0
    - BGP only allows to advertise one path
    - Downstream routers will not be aware of the path
      - Beware of routing loops !

BGP equal cost multipath (2)

- How to use BGP equal cost multipath here ?



- RB could send the packets to RZ via RY and RA
- R1 could also try to send the packets to RZ via RA and RB since R1 knows those two paths

BGP Equal Cost Multipath (3)

- Which paths can be used for load balancing ?
  - Run the BGP decision process and perform load balancing with the leftover paths at RouterId step
- Consequences
  - Border router receiving only eBGP routes
    - Perform load balancing with routes learned from same AS
    - Otherwise, iBGP and eBGP advertisements will not reflect the real path followed by the packets
  - Internal router receiving routes via iBGP
    - Only consider for load balancing routes with same attributes (AS-Path, local-pref, MED) and same IGP cost
    - Otherwise loops may occur

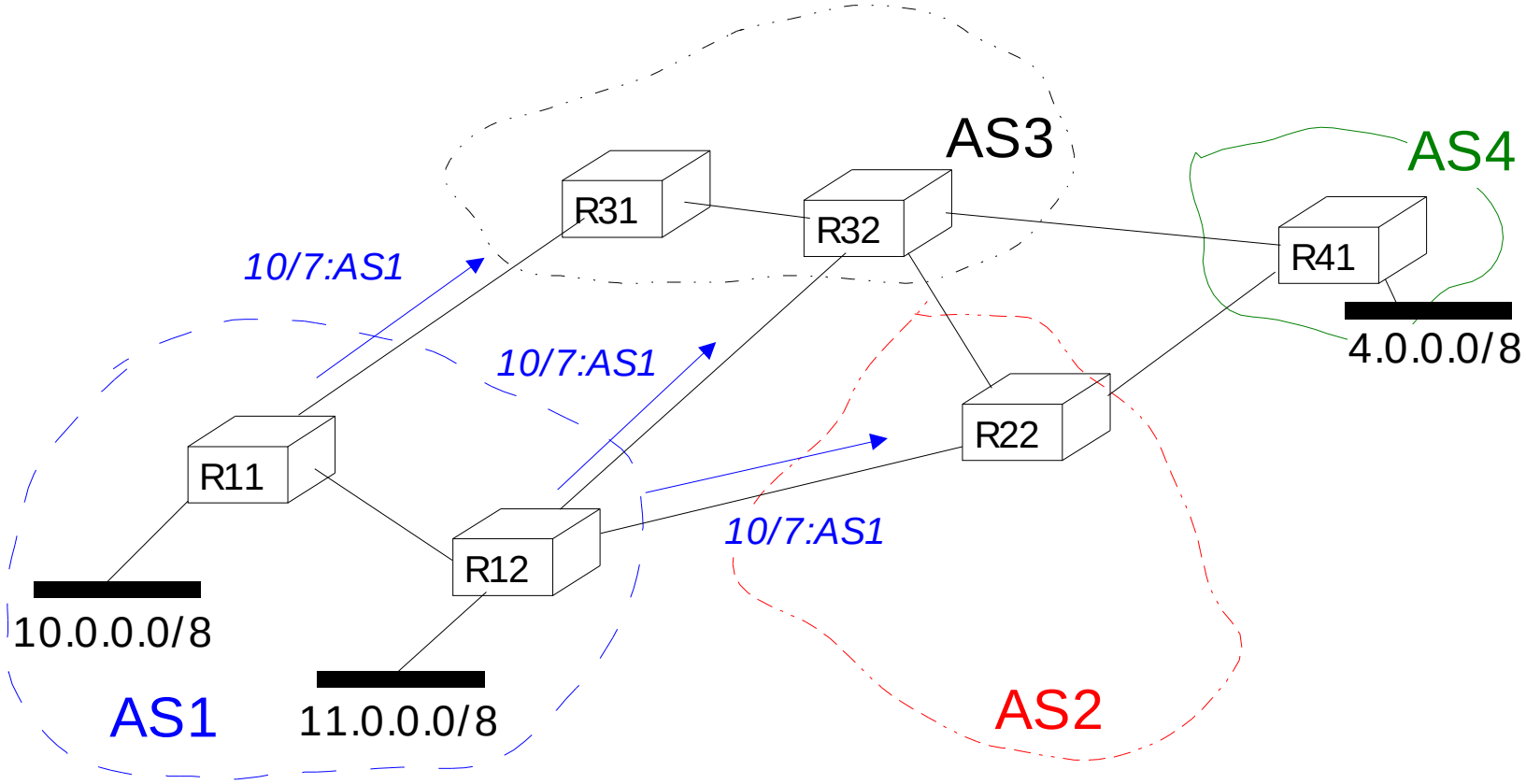
Tuning BGP to ...
control the incoming traffic

- Principle
  - To control its incoming traffic, a domain must tune the BGP advertisements sent by its own routers

- How to tune the BGP advertisements ?
  - Do not announce some routes to from some peers
    - advertise some prefixes only to some peers
  - MED
    - insert MED=IGP cost, usually requires bilateral agreement
  - AS-Path
    - artificially increase the length of AS-Path
  - Communities
    - Insert special communities in the advertised routes to indicate how the peer should run its BGP decision process on this route

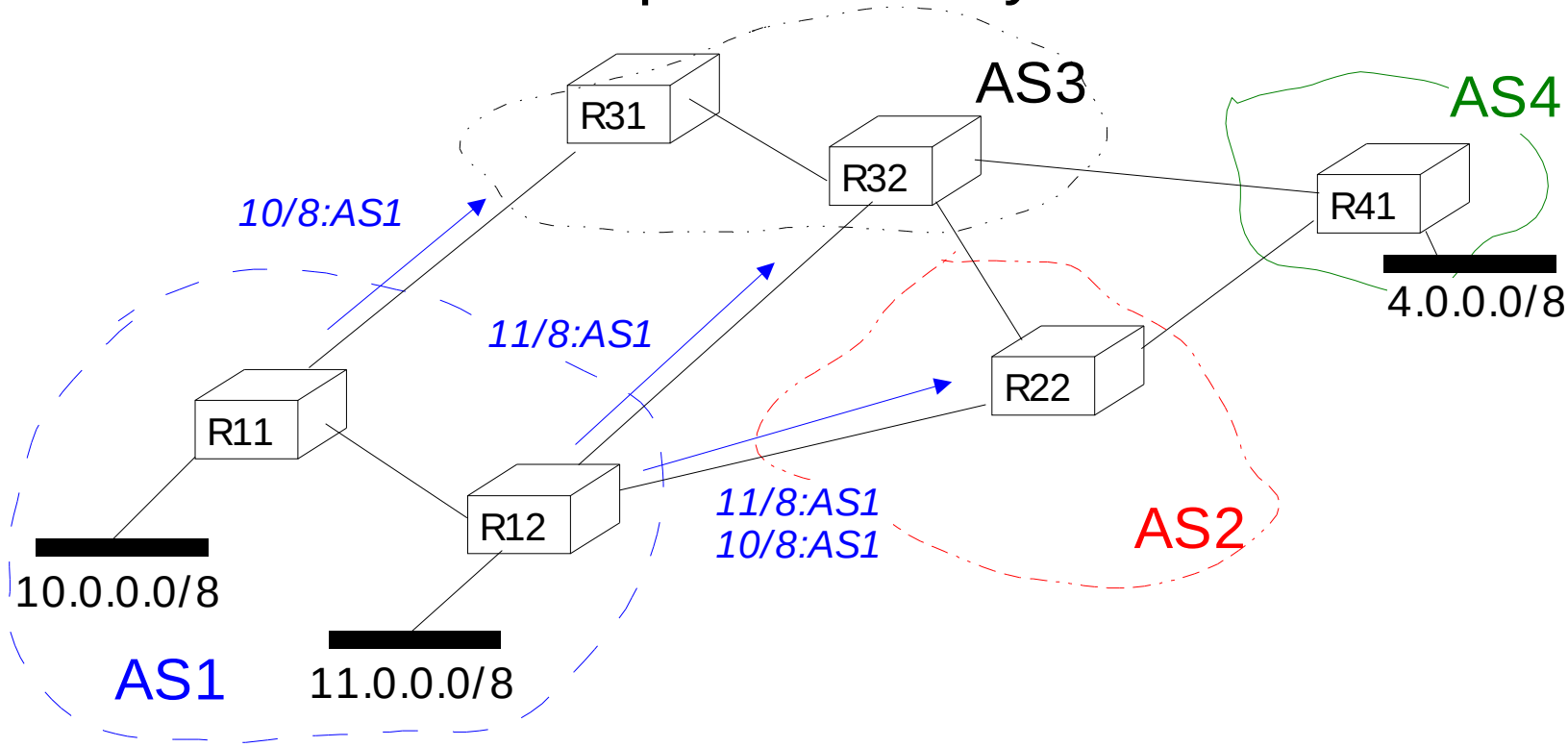
Control of the incoming traffic Sample network

- Routing without tuning the announcements
  - packet flow towards AS1 will depend on the tuning of the decision process of AS2, AS3 and AS4



Control of the incoming traffic Selective announcements

- Principle
  - Advertise some prefixes only on some links



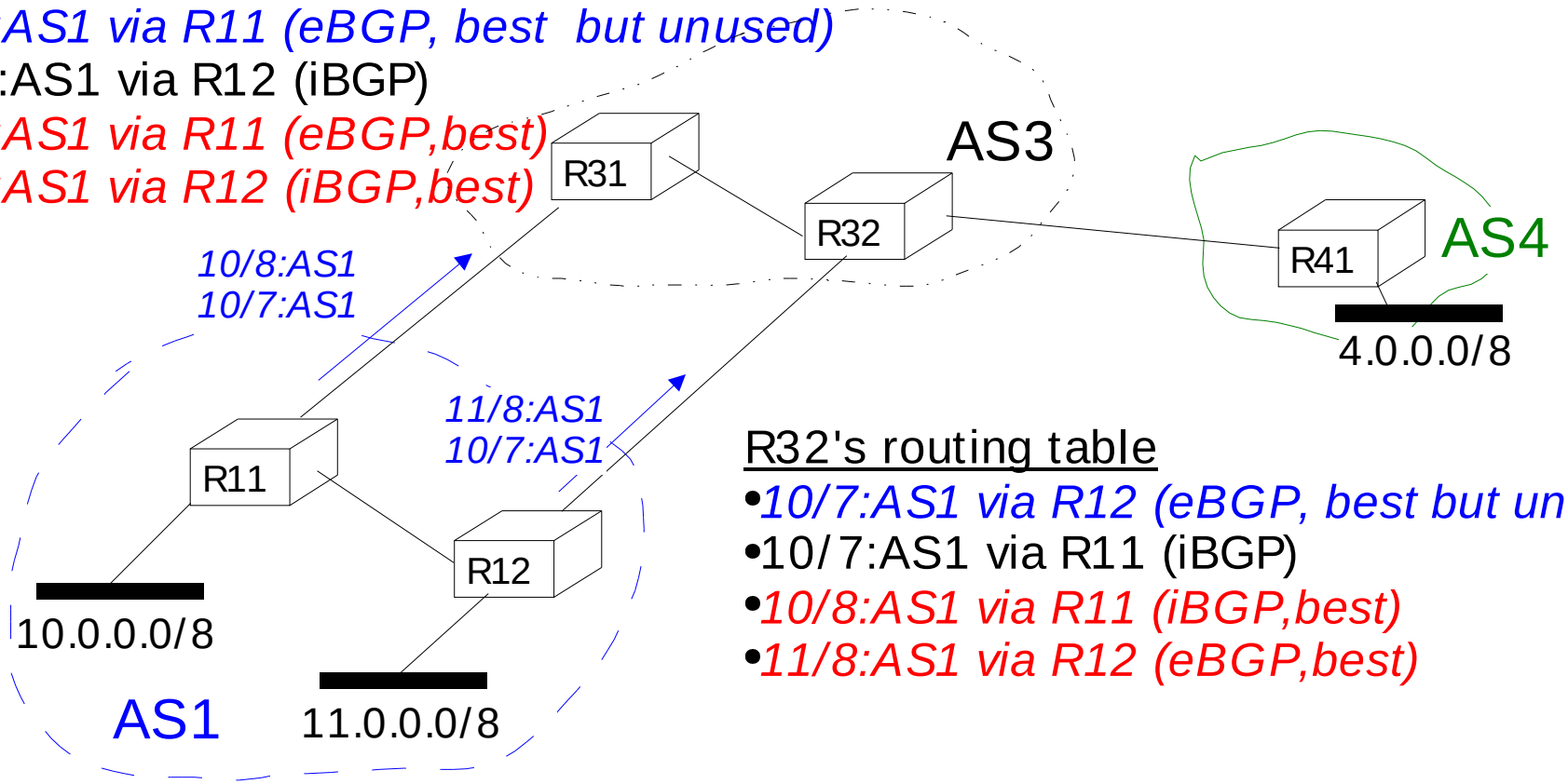
  - Drawbacks
    - splitting a prefix increases size of all BGP routing tables
    - Limited redundancy in case of link failure

Control of the incoming traffic
More specific prefixes

- Objective
  - Announce a large prefix on all links for redundancy but prefer some links for parts of this prefix
- Remember
  - When forwarding an IP packet, a router will always select the *longest match* in its routing table
- Principle
  - advertise different overlapping routes on all links
    - The entire IP prefix is advertised on all links
    - subnet1 from this IP prefix is also advertised on link1
    - subnet2 from this IP prefix is also advertised on link2
    - ...

Control of the incoming traffic
More specific prefixes (2)

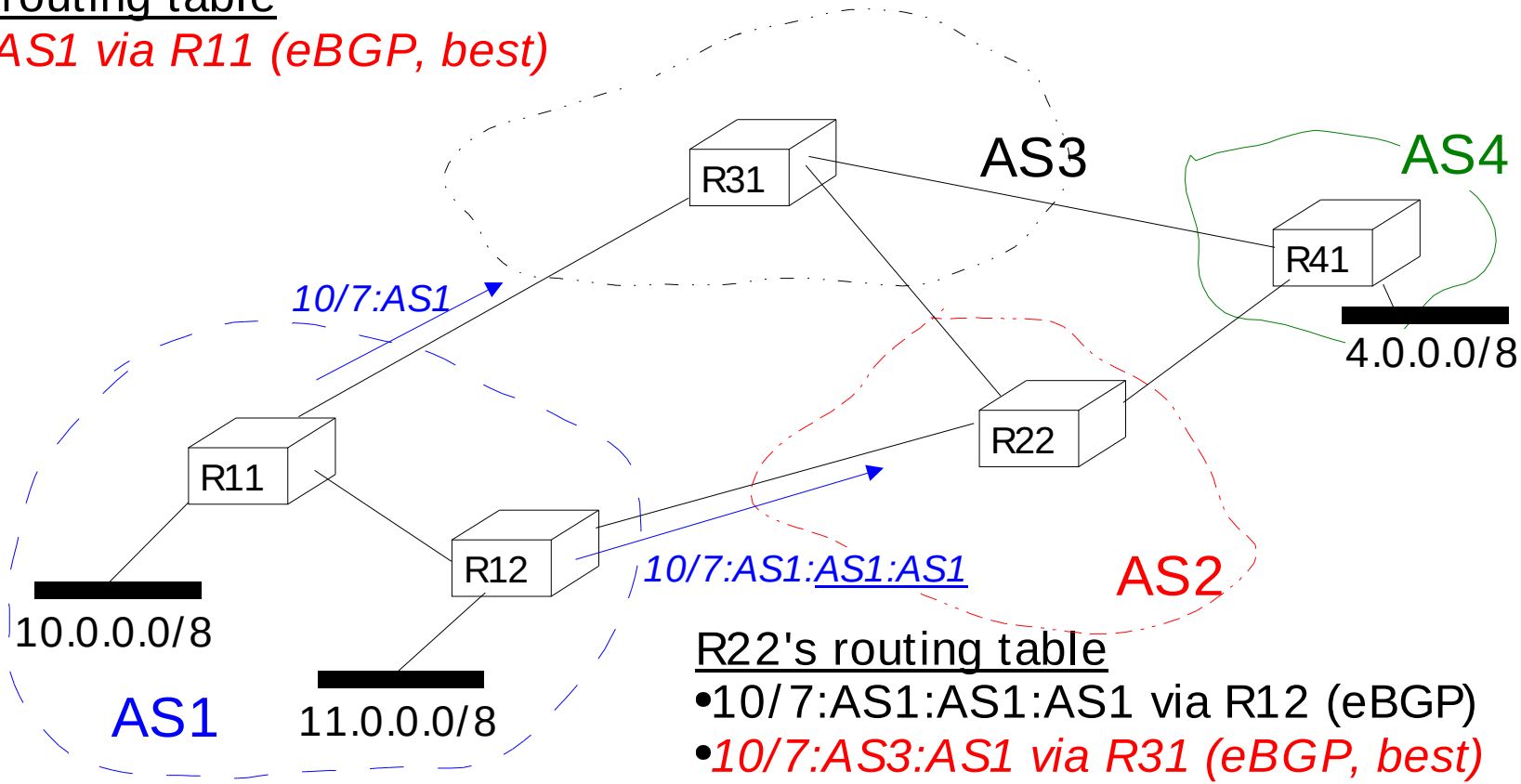
- Principle
  - Advertise partially overlapping prefixes



Control of the incoming traffic AS-Path prepending

- Principle
  - Artificially prepend own AS number on some routes

R31's routing table
- *10/7:AS1 via R11 (eBGP, best)*



R22's routing table
- 10/7:AS1:AS1:AS1 via R12 (eBGP)
- *10/7:AS3:AS1 via R31 (eBGP, best)*
- 10/7:AS4:AS3:AS1 via R41 (eBGP)

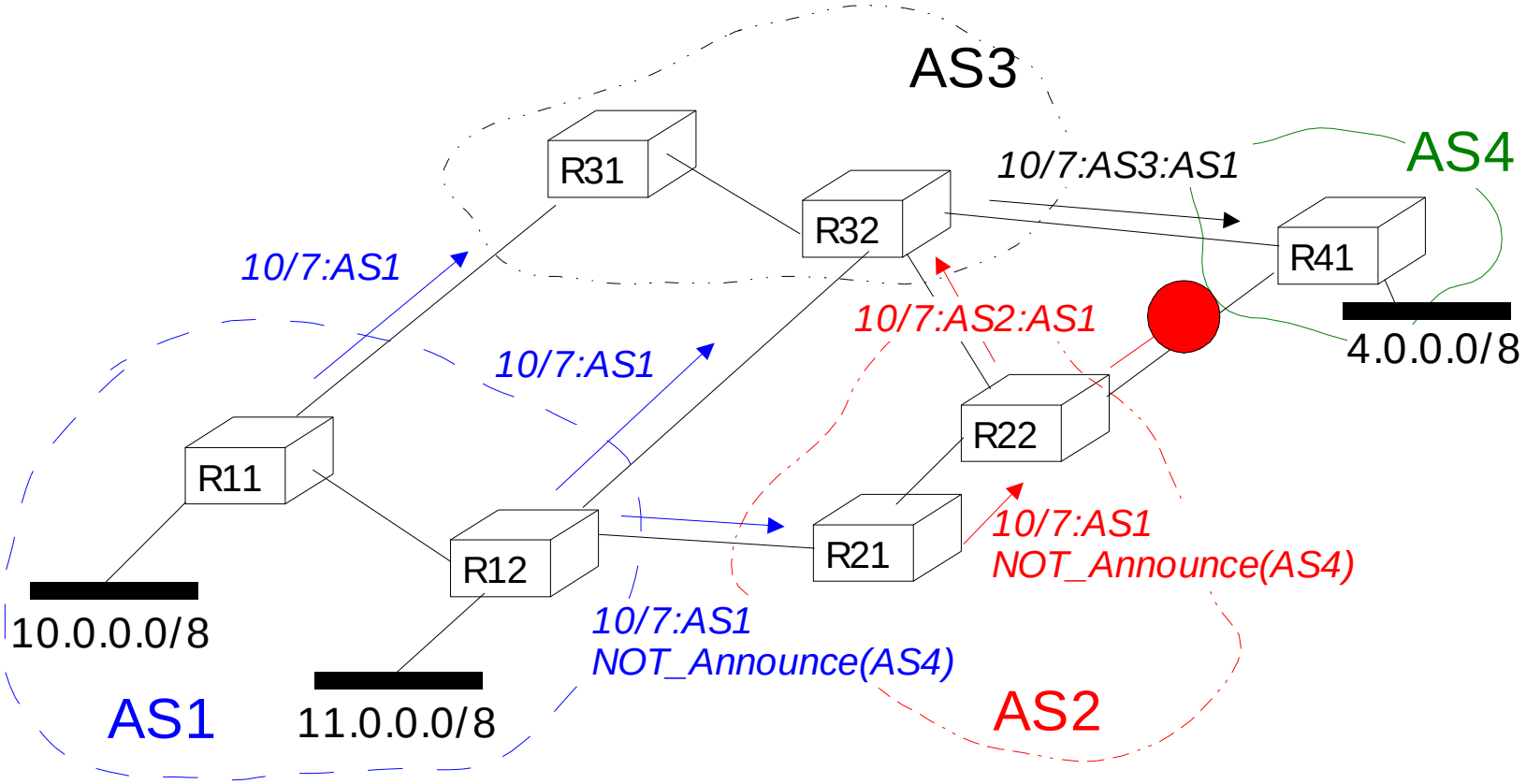
Traffic engineering with BGP communities

- Principle
  - Attach special community value to request downstream router to perform a special action
- Possible actions
  - Set local-pref in downstream AS
    - Example from UUnet (AS702)
      - 702:80 : Set Local Pref 80 within AS702
      - 702:120 : Set Local Pref 120 within AS702
  - Do not announce the route to ASx
    - Example from OpenTransit (AS1755)
      - 1755:1000 : Do not announce to US
      - 1755:1101: Do no announce to Sprintlink(US)
  - Prepend AS-Path when announcing to ASx
    - Example from BT Ignite (AS5400)
      - 5400:2000 prepend when announcing to European peers
      - 5400:2001 prepend when announcing to Sprint (AS1239)

The BGP redistribution communities

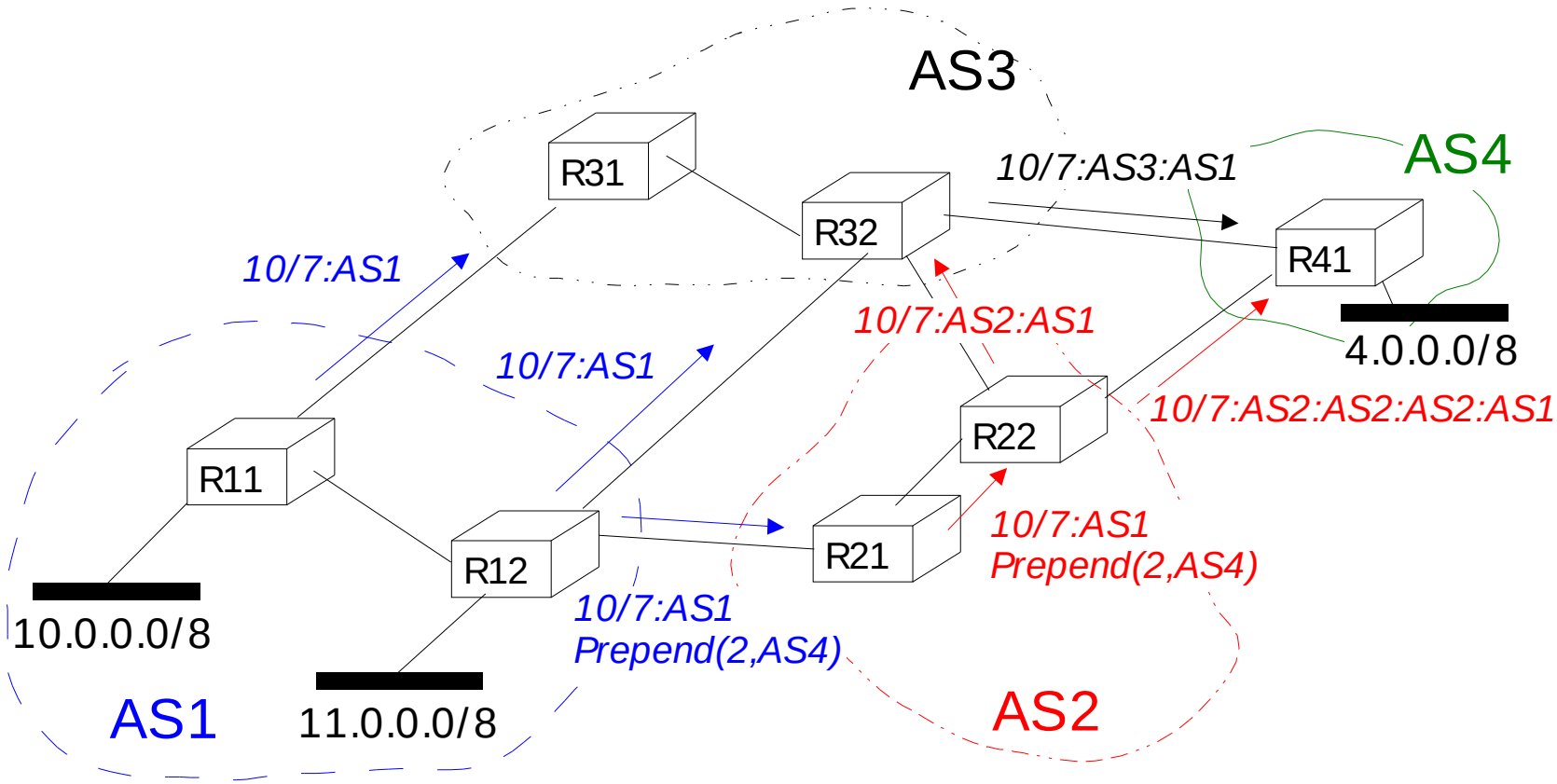
- Drawbacks of community-based TE
  - Requires error-prone manual configurations
  - BGP communities are transitive and thus pollute BGP routing tables
- Proposed solution
  - Utilize extended communities to encode TE actions in a structured and standardized way
  - actions
    - do not announce attached route to specified peer(s)
    - attach NO_EXPORT when announcing route to specified peer(s)
    - prepend N times when announcing attached route to specified peer(s)

Community-based selective announcements



- R22 does not announce 10/7 to R41
- R41 will only know one path towards 10/7

Community-based AS-Path prepending



- R22 announces 10/7 differently to R32 and R21
- R41 will prefer path via R32 to reach 10/7

Control of the incoming traffic Summary

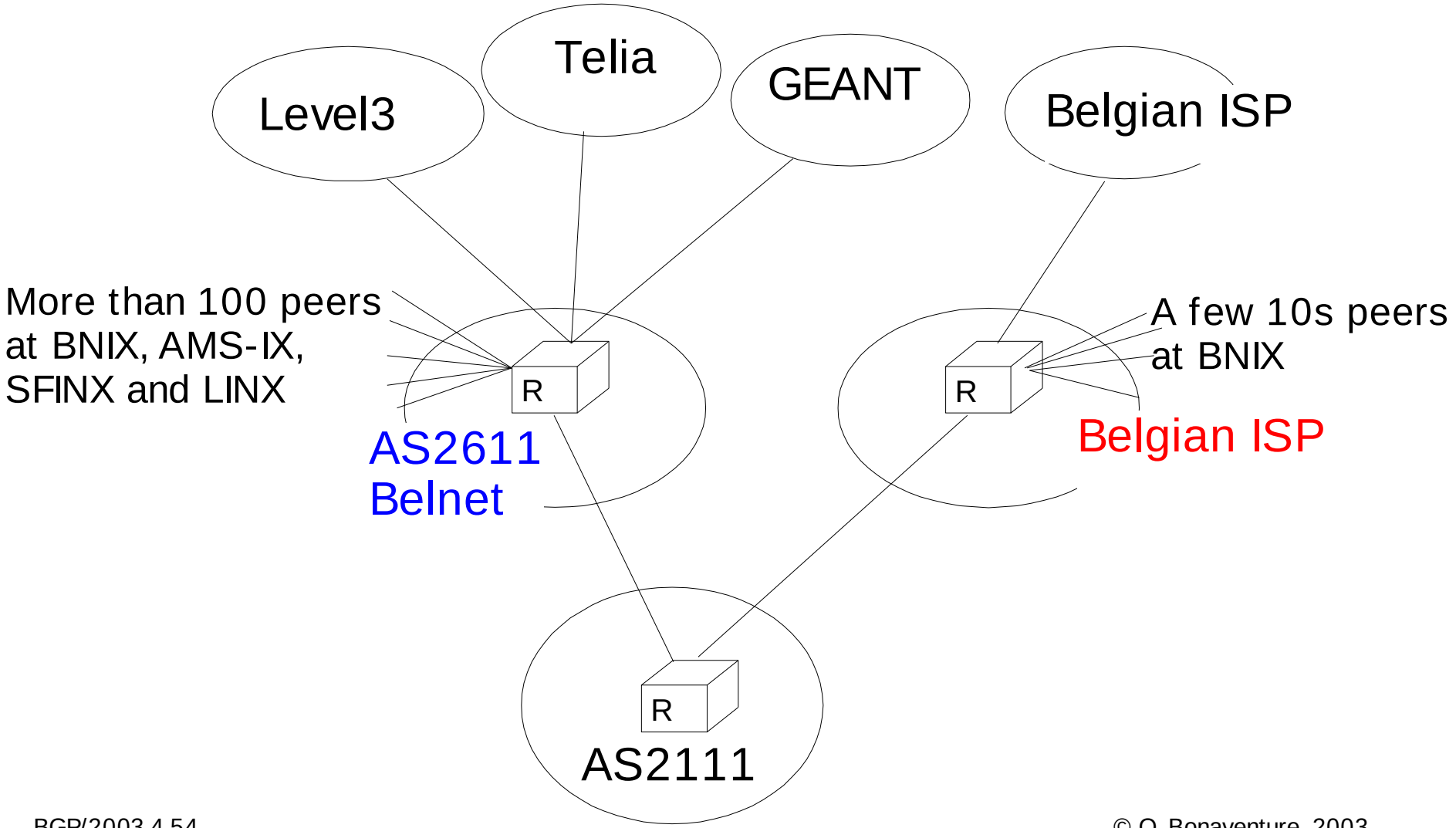
- Advantages and drawbacks
  - Selective announcements
    - always work, but if one prefix is advertised on a single link, it may become unreachable in case of failure
  - More specific prefixes
    - better than selective announcements in case of failure
    - but increases significantly the size of all BGP tables
    - some ISPs filter announcements for long prefixes
  - AS-Path prepending
    - Useful for backup link, but besides that, the only method to find the amount of prepending is trial and error...
  - Communities/redistribution communities
    - more flexible than AS-Path prepending
    - Increases the complexity of the router configurations and thus the risk of errors...

Outline

- Organization of the global Internet
- BGP basics
- BGP in large networks
- Interdomain traffic engineering with BGP
  - The growth of the BGP routing tables
  - The BGP decision process
  - Interdomain traffic engineering techniques
  - Case study
- BGP-based Virtual Private Networks

AS-Path prepending and communities in practice

- An experiment in the global Internet



Measurements with AS-Path prepending

- Study with 56k prefix from global Internet
  - For each prefix, sent TCP SYN on port 80 and measure from which upstream reply came back
- Without prepending
  - 68 % received via Belnet, 32% received via BISP
- With prepending once on Belnet link
  - 22% received via Belnet, 78% received via BISP
- With prepending twice on Belnet link
  - 15% received via Belnet, 84% received via BISP

How to better balance the incoming traffic ?

- AS Path prepending is clearly not sufficient
- Can we do better with the communities ?
  - Need to move some traffic from one upstream to another

- Level3 Communities
  - 65000:0
    - announce to customers but not to peers
  - 65000:XXX
    - do not announce to peer ASXXX
  - 65001:0
    - prepend once to all peers
  - 65001:XXX
    - prepend once to peer ASXXX

- Telia Communities
  - 1299:2009
    - Do not annouce EU peers
  - 1299:5009
    - Do not annouce US peers
  - 1299:2609
    - Do not anounce to Concert
  - 1299:2601
    - Prepend once to Concert

Before you start tuning your BGP routers...

'' My top three challenges for the Internet are scalability,
scalability, and
scalability''

Mike O'Dell, Chief scientist, UUNet

'' BGP is running on more than 100K routers (my estimate), making it one of the world's largest and most visible distributed system Global dynamics and scaling principles are still not well understood...''

Tim Griffin, AT&T Research