Interdomain routing with BGP4
Part 4/4

Olivier Bonaventure

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

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

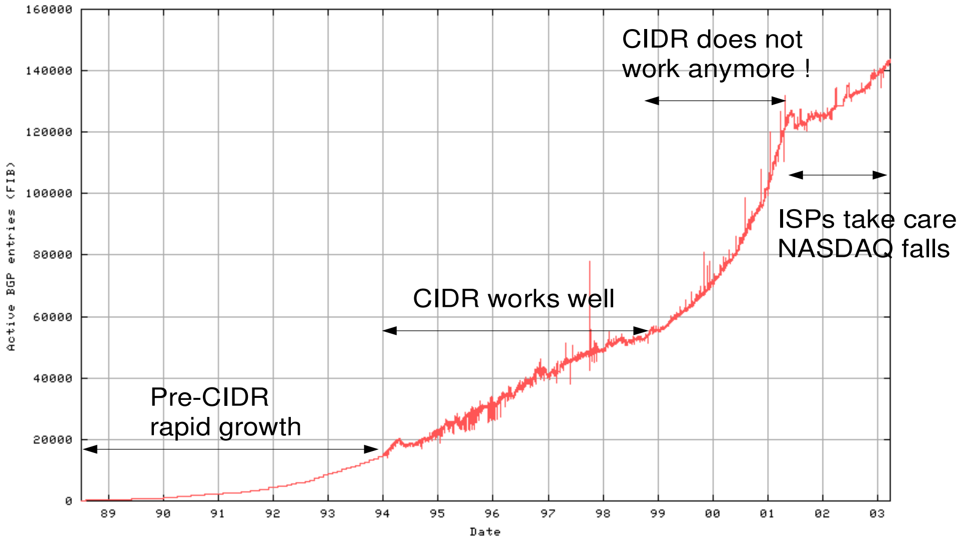
May 2003

Some of the note pages contain hypertext links to web pages. You can obtain an HTML or OpenOffice version of this tutorial with the hypertext links by sending an email to the author.

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 studies

The growth of the BGP routing tables



BGP/2003.4.3

Source: http://bgp.potaroo.net , April 2003

Source :

http://bgp.potaroo.net/as1221/bgp-active.html

For more information on the growth of the BGP tables, see :

http://bgp.potaroo.net
http://www.cidr-report.org

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
- Increase in number of ASes
  - About 3000 ASes in early 1998
  - More than 13000 ASes in April 2003
  - 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

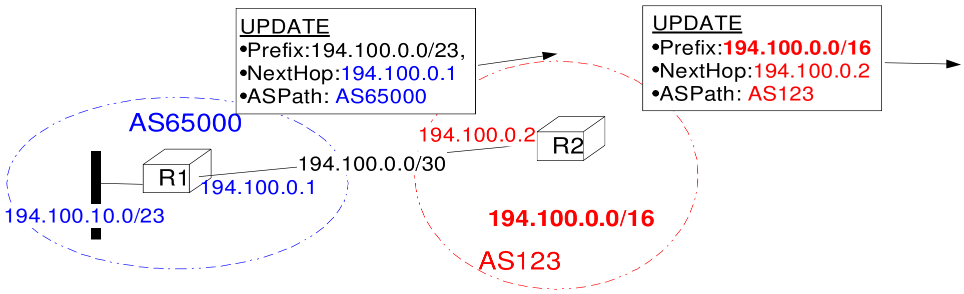
Source for this data :

http://bgp.potaroo.net

S. Agarwal, C. Chuah, R. Katz, OPCA : Robust interdomain policy routing and traffic control, IEEE OPENARCH 2003, April 2003

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

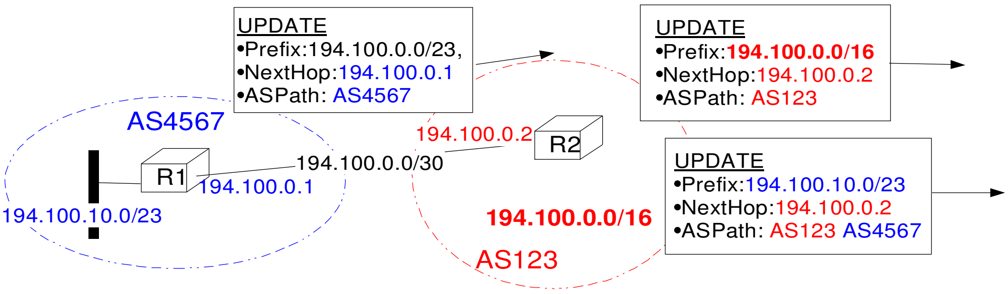
The private AS numbers (range 64512 through 65535) are reserved for private use and should not be advertised on the global Internet. See

J. Hawkinson, T. Bates, Guidelines for creation, selection, and registration of an Autonomous System (AS), RFC1930, March 1996

See also
J. Stewart, T. Bates, R. Chandra, E. Chen, Using a Dedicated AS for Sites Homed to a Single Provider, RFC2270, January 1998

Evolution of typical stub AS (2)

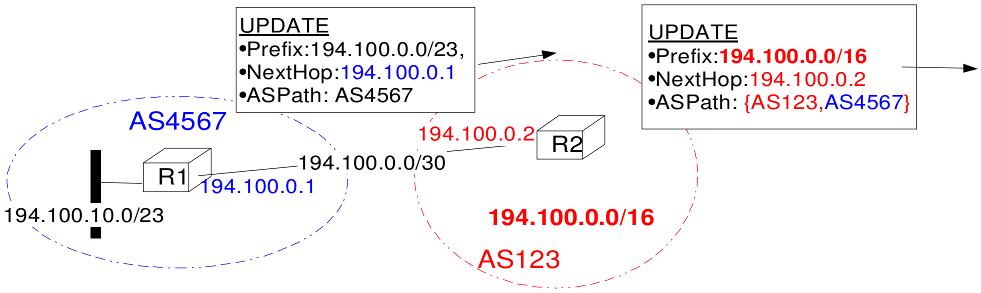
- Day two, stub AS expects to become multi-homed in near future and obtains official AS#



- 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

Another solution is to strip the AS# of the client network in the BGP advertisement. Removing this information may prohibit other domains from detecting loops. For this reason, two new attributes need to be added to the BGP advertisement :

- ATOMIC_AGGREGATE indicates that path information has been lost in the aggregation process

  Indicates also that the prefix should not be deaggregated further

AGGREGATOR contains info useful for debugging

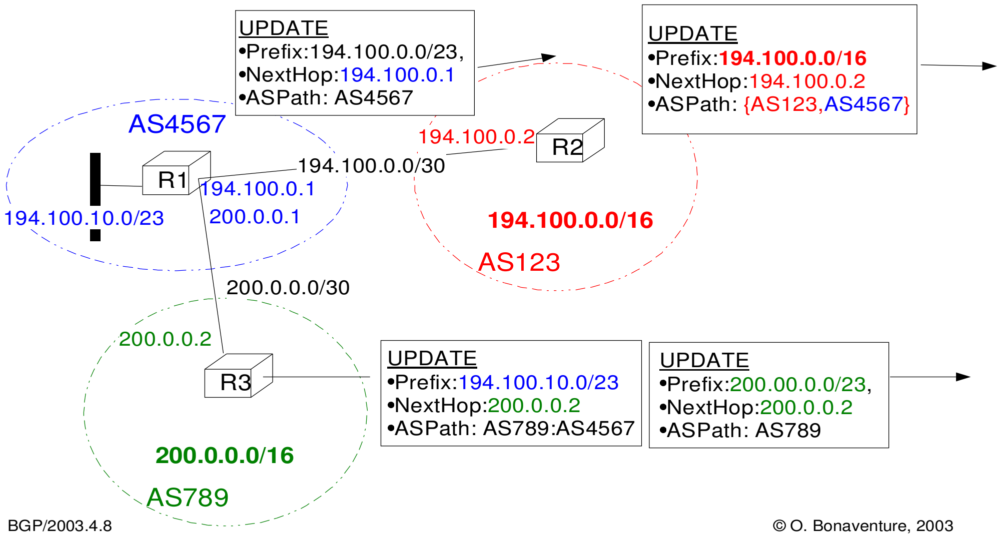
In this case, the BGP UPDATE message would be as follows :

UPDATE
•Prefix:**194.100.0.0/16**
•NextHop:194.100.0.2
•ASPath: AS123
•AGGREGATOR
AS123, 194.100.0.2
•ATOMIC_AGGREGATE

In April 2003, a BGP table collected by the RIPE RIS project contained about 7% of routes with the ATOMIC_AGGREGATE attribute

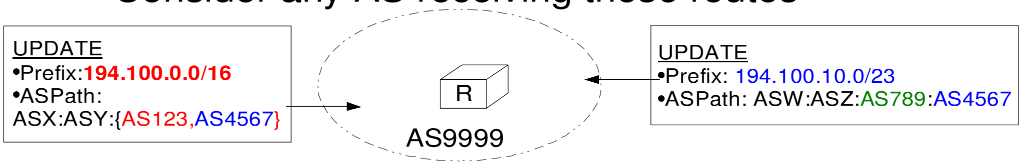
A dual-homed stub ISP

- 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 the non-aggregated path since it is the most specific !!!
  - AS123 might be forced to 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

See http://www.cidr-report.org for more information about the current status of the aggregation of BGP routes. This site computes regularly the optimum aggregates that should be announced by each AS based on BGP tables collected at various locations.

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 !

For more information on filtering based on the RIR allocation guidelines, see Steve Bellovin, Randy Bush, Timothy G. Griffin, and Jennifer Rexford, "Slowing routing table growth by filtering based on address allocation policies," June 2001, available from http://www.research.att.com/~jrex

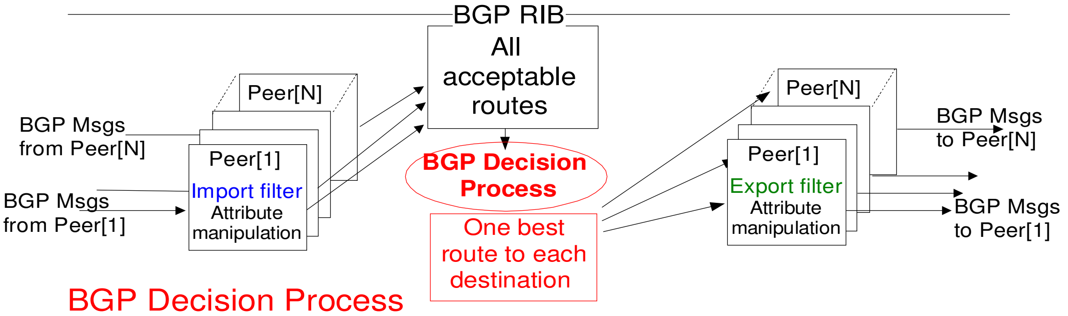
The RIPE allocation guidelines may be found at :
http://www.ripe.net/ripe/docs/ir-policies-procedures.html

For a discussion of the Ipv6 multi-homing solutions being developped, see the site multi-homing with Ipv6 working group of the IETF
http://www.ietf.org/html.charters/multi6-charter.html

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 studies

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 BGP decision process also contains a additional step after the ASPath step where the routes with the lowest ORIGIN attribute are preferred. We ignore this step and this attribute in this tutorial.

The BGP decision process used by router vendors may change compared to this theoretical description. For real BGP decision processes, see :

http://www.cisco.com/en/US/tech/tk826/tk365/technologies_tech_note09186a0

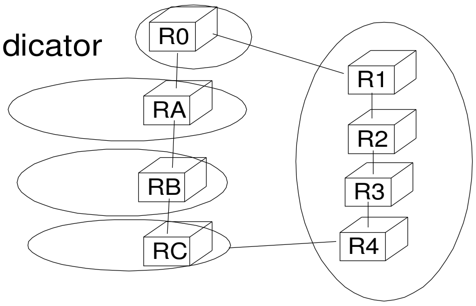
http://www.riverstonenet.com/support/bgp/routing-model/index.htm#_Route_Se

http://www.juniper.net/techpubs/software/junos53/swconfig53-ipv6/html/routing-

http://www.foundrynet.com/services/documentation/ecmg/BGP4.html

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



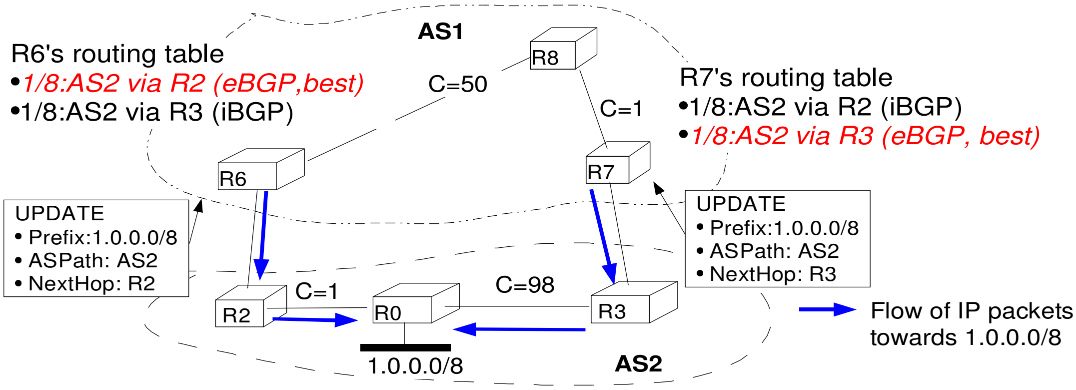
- Consequence
  - Internet paths tend to be short, 3-5 AS hops
  - Many paths converge at Tier-1 ISPs and those ISPs carry lots of traffic

A recent study of the quality of the AS Path as a performance indicator compared the round trip time with the length of the AS Path and has shown that the length of the AS Path was only a good indicator for 50% of the considered paths. See :

Bradley Huffaker, Marina Fomenkov, Daniel J. Plummer, David Moore and k claffy, Distance Metrics in the Internet, Presented at the IEEE International Telecommunications Symposium (ITS) in 2002. http://www.caida.org/outreach/papers/2002/Distance/

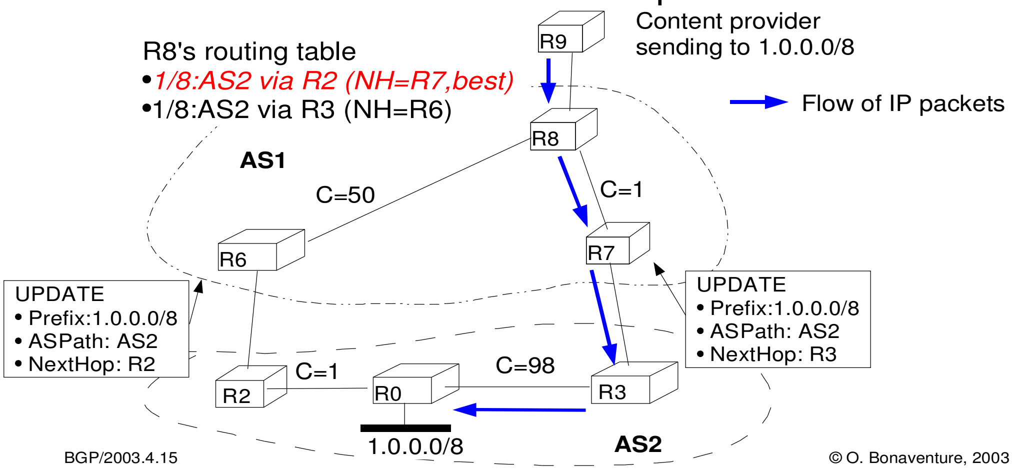
The prefer eBGP over iBGP step in the BGP decision process

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



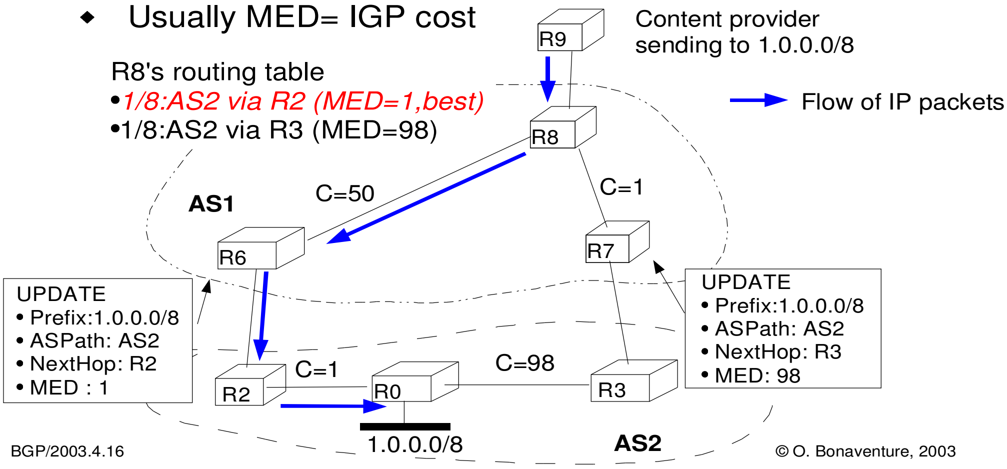
The closest `nexthop` step in the BGP decision process

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



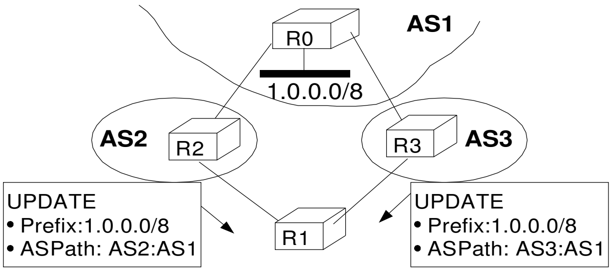
The lowest MED step in the BGP decision process

- 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

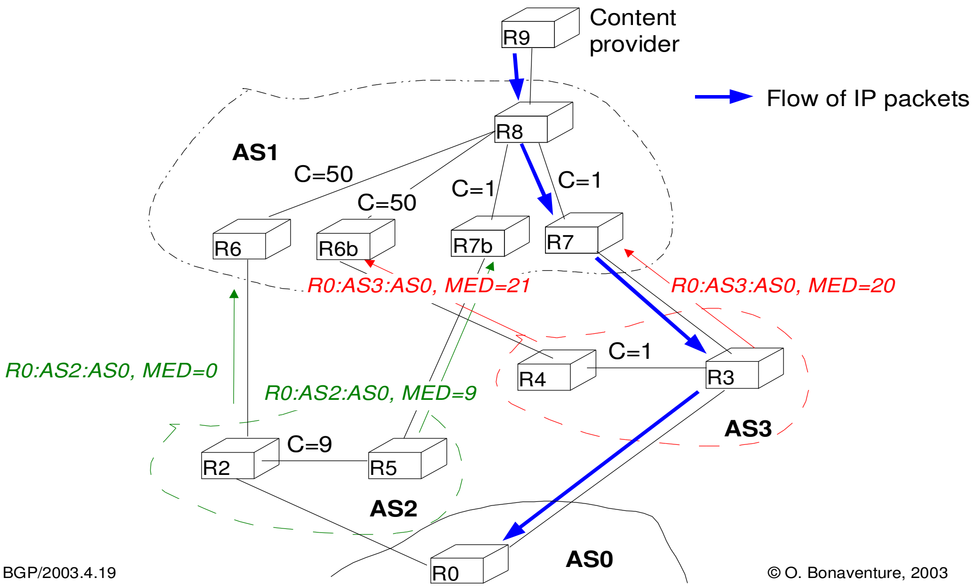
Note that on some router implementations, the lowest router id step in the BGP decision process is replaced by the selection of the oldest route. See e.g. : http://www.cisco.com/warp/public/459/25.shtml
Preferring the oldest route when breaking times is used to prefer stable paths over unstable paths, however, a drawback of this approach is that the selection of the BGP routes will depend on the arrival times of the corresponding messages. This makes the BGP selection process non-deterministic and can lead to problems that are difficult to debug.

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 ?



•In the example above, assuming a full iBGP mesh inside AS1 and that all routes have the same local-pref value, router R8 will receive four paths to reach router R0 :

- One path going via R5 in AS2 and received with MED=9
- One path going via R3 in AS3 and received with MED=20
- One path going via R2 in AS2 and received with MED=0
- One path going via R4 in AS3 and received with MED=21

The local-pref and AS-Path steps of the decision process will not remove any path from consideration.

The MED step of the BGP decision process will select, from each neighboring AS, the paths with the smallest MED, namely :

- One path going via R2 in AS2 and received with MED=0
- One path going via R3 in AS3 and received with MED=20

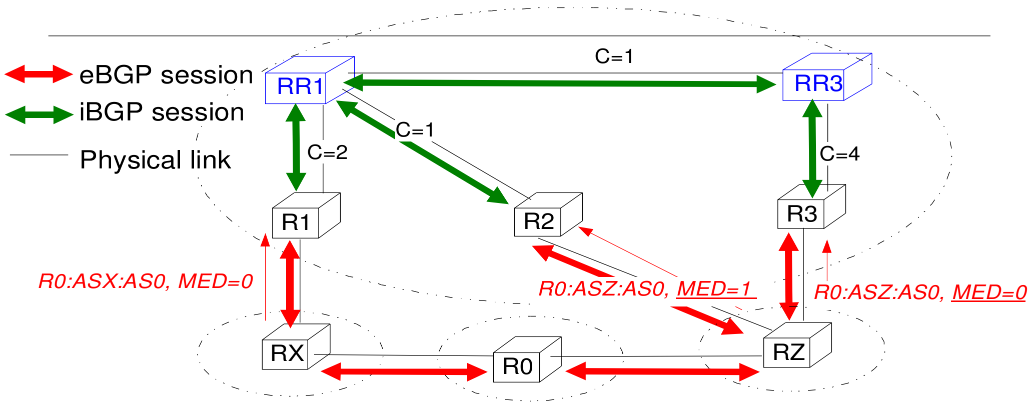
Then, the closest nexthop step of the BGP decision process will select as best path the path that leaves AS1 router R7, i.e. :

- One path going via R3 in AS3 and received with MED=20

This is the standardized processing of the MED attribute in BGP4. As always with BGP4 implementations, some implementations allow operators to :

- Ignore the MED values from a given peer
- Process all MED values without considering the AS from which the MED value was learned
  - in this case, the path via R5 would be selected
- ...

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

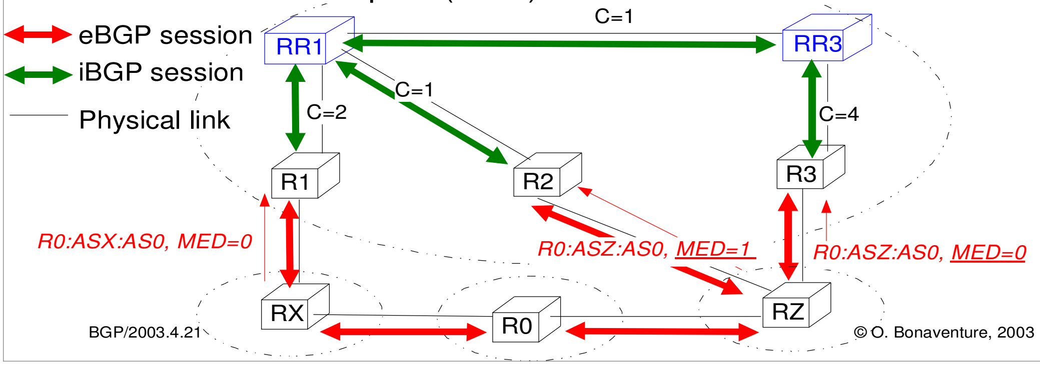
This route oscillation problem is described in :

D. McPherson, V. Gill, D. Walton, A. Retana, BGP Persistent Route Oscillation Condition, Internet draft, draft-ietf-idr-route-oscillation-01.txt, work in progress, Feb 2002

A better description and analysis may be found in :
Analysis of the MED Oscillation Problem in BGP. Timothy G. Griffin and Gordon Wilfong. ICNP 2002

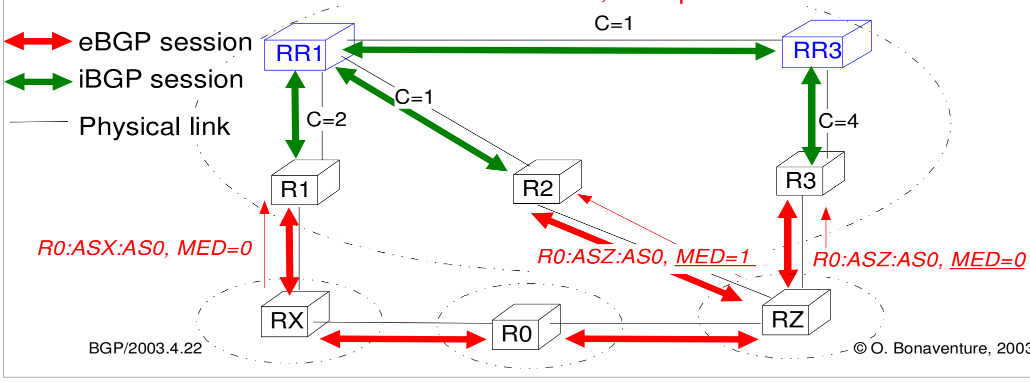
Route oscillations with MED (2)

- 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

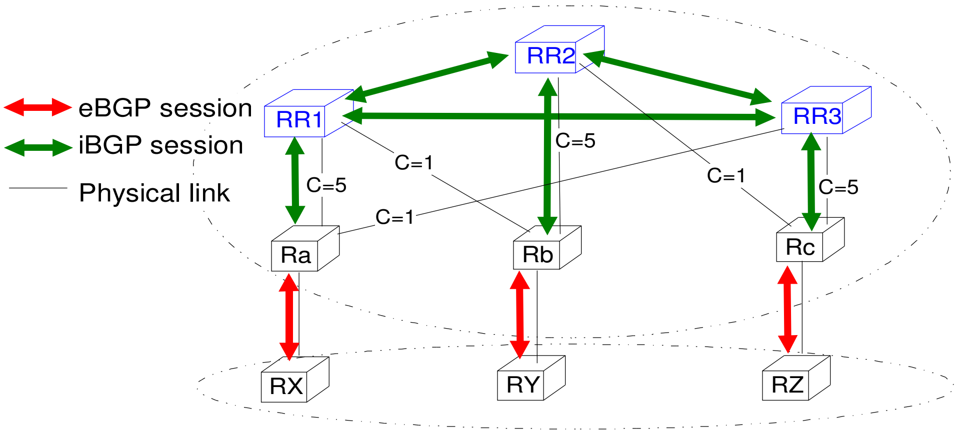


Route oscillations with MED (3)

- 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

With an iBGP full mesh, all BGP routers would received the three possible paths and RR1 would prefer the path via R2, RR2 would prefer the path via R3 and RR3 would prefer the path via R1.

With Route Reflectors, the situation is more complex because each RR only knows some of the routes since each RR only advertises its best path on the iBGP full mesh with the other Rrs.

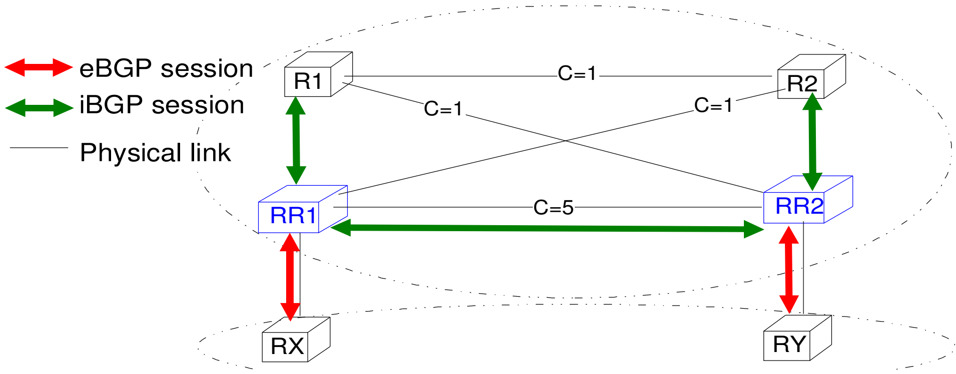
RR1 will learn the path via RX from its client R1. RR2 learns the path via RY from its client R2 and RR3 learns the path via RZ from its client R3.

Assume RR1is the first to select its path. It selects the RX path since it only knows this path and advertises it to RR2 and RR3. Upon reception of this advertisement, RR3 compares the path via RZ and the path via RX and prefers the path via RX. RR3 advertises its best path to R3, but R3 still prefers its direct path to RZ.. Note that RR3 does not advertise the path via RZ to the other RRs since this is not its best path.

Now, assume that RR2 selects its best path. It knows the paths via RX (learned from RR1) and RY (learned via R2). The current best path is clearly the path via RY and RR2 advertises this path to RR1 and RR3. Upon reception of this advertisement, RR1 will select again its best path. Now, RR1's best path is clearly the path via RY. Unfortunately, the selection of this path forces RR1 to withdraw the path via RX that it initially advertised. Upon reception of the withdraw message, RR3 will need to select its best path... The RRs will exchange BGP messages forever without reaching a consensus.

For more information about this problem and others, see :
T. Griffin, G. Wilfong, On the correctness of iBGP configuration, Proc. ACM SIGCOMM2002, August 2002
Route Oscillations in I-BGP with Route Reflection. Anindya Basu, Chih-Hao Luke Ong, April Rasala, F.Bruce Shepherd, and Gordon Wilfong. SIGCOMM 2002

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)

Note that this forwarding problem does not occur if R1 and R2 use some tunneling mechanism (e.g. MPLS) to send packets towards RX and RY via RR1 and RR2

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 studies

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

For a vendor-oriented discussion of interdomain traffic engineering, see :

T. Monk, Inter-domain Traffic Engineering: Principles and case examples, Proc. INET 2002, http://inet2002.org/CD-ROM/lu65rw2n/papers/t06-c.pdf

In you intend to negotiate peering agreements, you should probably read :
W. Norton, The Art of Peering: The Peering Playbook , available from < wbn@equinix.com> or http://www.xchangepoint.net/white_papers/wp20020625.pdf

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 ?

For a discussion on the types of monitoring or measurements suitable for traffic engineering purposes, see :

Wai Sum Lai et al., A framework for internet traffic engineering measurement, Internet draft, draft-ietf-tewg-measure-02.txt, March 2002

Other references include

Anja Feldmann, Albert Greenberg, Carsten Lund, Nick Reingold, Jennifer Rexford, and Fred True. Deriving traffic demands for operational ip networks: methodology and experience. In *Proc. ACM SIGCOMM2000*, September 2000.
An extended version appeared in IEEE/ACM Transactions on Networking

Matthias Grossglauser and Jennifer Rexford, "Passive traffic measurement for IP operations," to appear as a chapter in The Internet as a Large-Scale Complex System, Oxford University Press, 2002 (INFORMS slides).

Traffic Matrix Estimation: Existing Techniques and New Directions. A. Medina (Sprint Labs, Boston University) , N. Taft (Sprint Labs), K. Salamatian (University of Paris VI), S. Bhattacharyya, C. Diot (Sprint Labs)

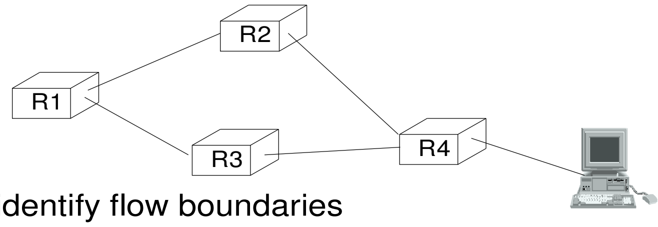
See also the papers presented at the ACM SIGCOMM Internet Measurement Workshops and at PAM

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

A very popular tool for link-level monitoring is MRTG, see
http://people.ee.ethz.ch/~oetiker/webtools/mrtg/

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 monitoring tools started with the development of Netflow on Cisco routes (http://www.cisco.com/warp/public/732/Tech/nmp/netflow/). Netflow is available in various formats (V1, V5, V7, V8), depending on the router platform and the desired monitoring information.
Since then, several third-party software have been developed to collect Netflow data. A good list of pointers for such tools is maintained by Simon Leinen at SWITCH (http://www.switch.ch/tf-tant/floma/software.html).

Several vendors have also adopted the Netflow format (http://www.juniper.net/techpubs/software/junos53/swconfig53-policy/html/samp)

Within IETF, the IPFIX working group is expected to develop a standard alternative to Netflow. See http://www.ietf.org/html.charters/ipfix-charter.html

Open source tools can also be used to capture traffic in Netflow format, see e.g. http://www.ntop.org

Flow level traffic capture (3)

- Advantages
  - provides detailed information on the traffic carried out on some links

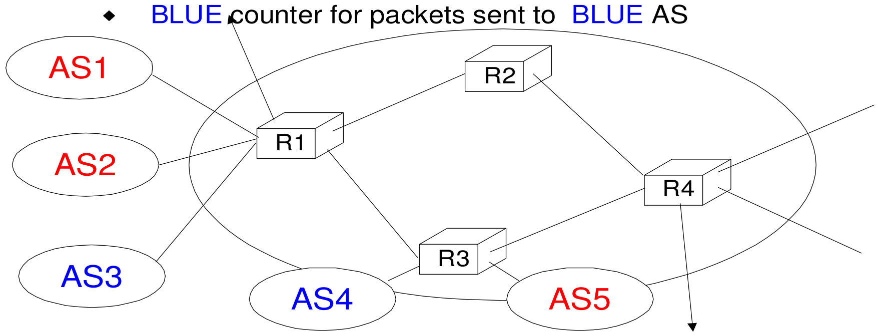
- Drawbacks
  - flow information needs to be exported to monitoring station
    - information about one flow is 30 - 50 bytes
    - average size of HTTP flow is 15 TCP packets
  - CPU load on high speed on routers
    - not available on some router platforms
  - Disk and processing requirements on monitoring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

BGP policy accounting

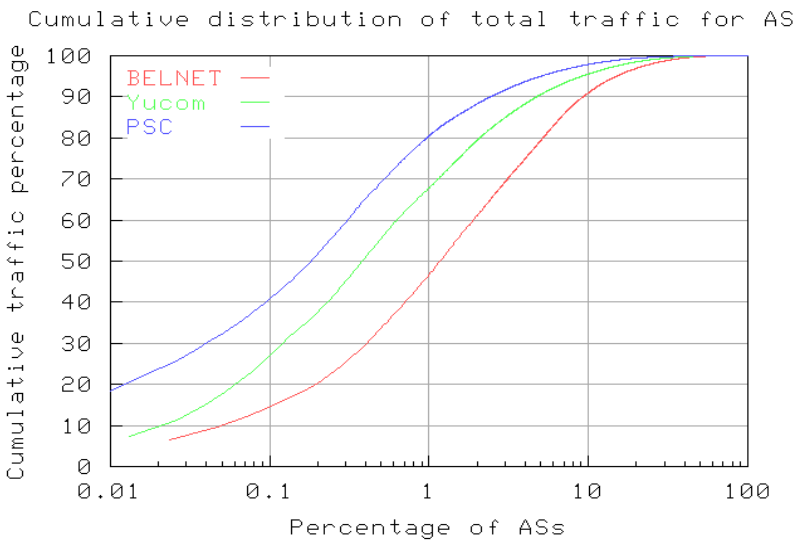
- Color some routes in BGP tables
  - RED counter for packets sent to RED AS
  - BLUE counter for packets sent to BLUE AS



- Border router maintains per-color statistics
- when packets are forwarded, statistic associated to the color of the route used to route the packet is updated
- Drawback
  - currently restricted to a limited number of colors

For more information on this feature, see
http://www.switch.ch/misc/leinen/snmp/monitoring/bucket-accounting.html
http://www.riverstonenet.com/technology/bgp_policy.shtml
http://www.cisco.com/warp/public/459/38.html

Characteristics of interdomain traffic



This figure is based on a study of all the interdomain traffic of three distinct ISPs at different periods of time. The trace was collected during one week for BELNET, the Belgian Research ISP, five days for YUCOM, a dialup ISP based in Belgium and one day for PSC, a gigapop in the US. This figure is analyzed in :
B. Quoitin, S. Uhlig, C. Pelsser, L. Swinnen and O. Bonaventure, Interdomain traffic engineering with BGP, IEEE Communications Magazine, May 2003, http://www.info.ucl.ac.be/people/OBO/biblio.html

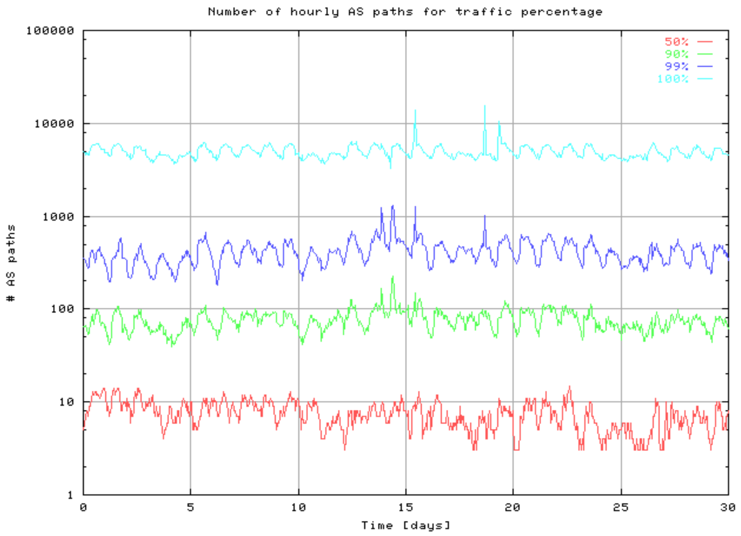
A detailed analysis of the characteristics of interdomain traffic based on a stub ISP may be found in :

S. Uhlig and O. Bonaventure, Implications of interdomain traffic characteristics on traffic engineering, European Transactions on Telecommunications, Jan. 2002, http://www.info.ucl.ac.be/people/OBO/biblio.html

A similar result concerning the traffic distribution was obtained by studying the traffic of a tier-1 ISP, see

N. Feamster, J. Borkenhagen, J. Rexford, Controlling the impact of BGP policy changes on IP traffic, AT&T Technical Memorandum, 2001

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

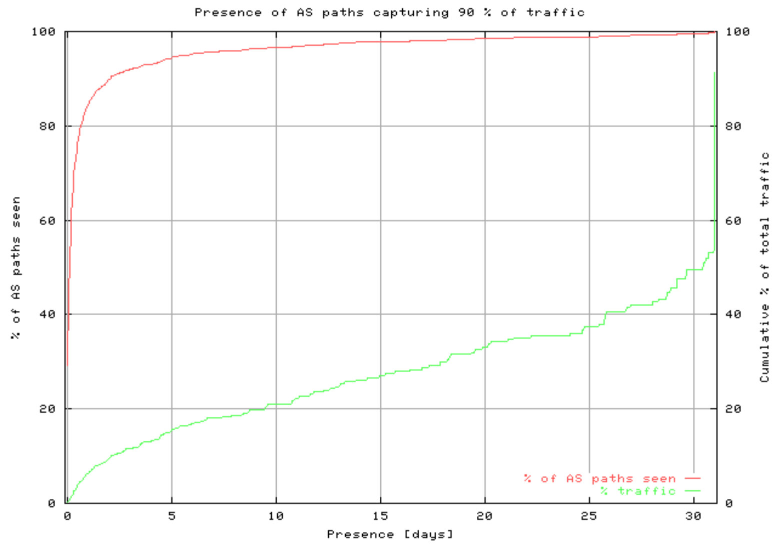


This figure is taken from :

S. Uhlig, V. Magnin, O. Bonaventure, C. Rapier, L. Deri, On the Topological Stability of Interdomain Traffic, unpublished manuscript, May2003

This paper analyses the stability of the traffic sent by the UCL network to the Internet during one month. The figure above was drawn by computing during each hour, the sorted list of active AS Paths during this period and then counting how many of those top AS-Paths were required to capture a given amount of traffic.

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



This figure is taken from :

S. Uhlig, V. Magnin, O. Bonaventure, C. Rapier, L. Deri, On the Topological Stability of Interdomain Traffic, unpublished manuscript, May2003

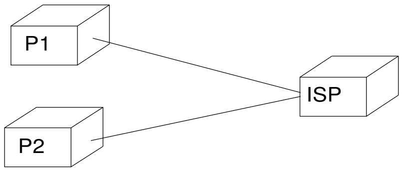
The figure above was drawn by counting the number of times each AS Path that appeared in thehourly top 90% figure and comparing this information with the amount of traffic sent on those AS Paths. It shows that a small number of AS Paths are always present, but that most AS Paths only appear during small periods of time.

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

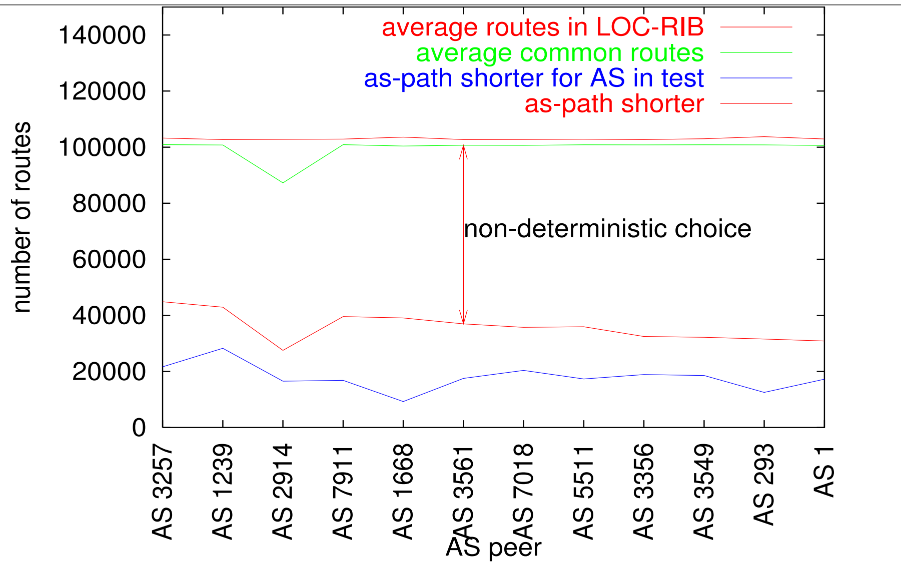


  - Rank providers based on the routes selected by the BGP decision process of the simulated ISP

This study was conducted by Sébastien Tandel in November 2002 based on the BGP routing tables stored by Routeviews. Additional information may be found in :

L. Swinnen, S. Tandel, S. Uhlig, B. Quoitin and O. Bonaventure, An Evaluation of BGP-based Traffic Engineering Techniques, under submission, Dec. 2002
http://www.info.ucl.ac.be/people/OBO/papers/cost263-chapter.pdf

Selection among the 12 largest providers



The twelve considered providers are large T1 ISPs :

AS2914	: Verio
AS3257	: TISCALI
AS1239	: Sprint
AS7911	: Williams
AS3561	: C&W USA
AS1668	: AOL
AS7018	: ATT
AS5511	: FT Backbone
AS3549	: GLBIX
AS3356	: Level3
AS1	: Genuity
AS293	: ESnet

For these ISPs that are in majority tier 1, the figure shows that the number of common routes is very high varying between 96.9 and 98.1% of the full BGP table except for AS2914 having on average 85% of the routes in common with the 11 other peers. The figure also shows that between 56033 and 69735 routes are selected in a non-deterministic manner by the BGP decision process of our stub AS. A closer look at those routes reveals that 80% of them have an AS-Path length of 3 to 4 AS-hops. On average, for all considered pairs, almost 62% of the routes are chosen in a non deterministic manner. This result implies that the length of AS-Path is not always a sufficient condition to select BGP routes and that ISPs could easily influence their outgoing traffic by defining additional criteria to prefer one provider over the other.

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 (ignore) 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 some routes over others with same AS Path length
  - IGP cost to nexthop
    - setting of IGP cost for intradomain traffic engineering
  - Several routes in the fowarding table instead of one

Usually, the control of the outgoing traffic is based on a manual configuration of the routers. However, recently some vendors have proposed tools to automate the control of the outgoing traffic based on measurements. See e.g. :

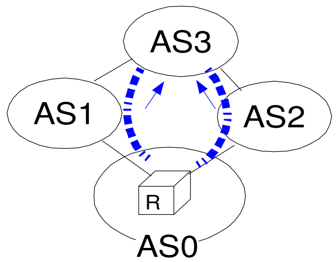
J. Bartlett, Optimizing multi-homed connections,Business Communications Review, January 2002

D. Allen, NPN: Multihoming and Route Optimization: Finding the Best Way Home, Network Magazine, Feb. 2002, http://www.networkmagazine.com/article/NMG20020206S0004

S. Borthick, Will route control change the Internet, Business Communications Review, September 2002

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



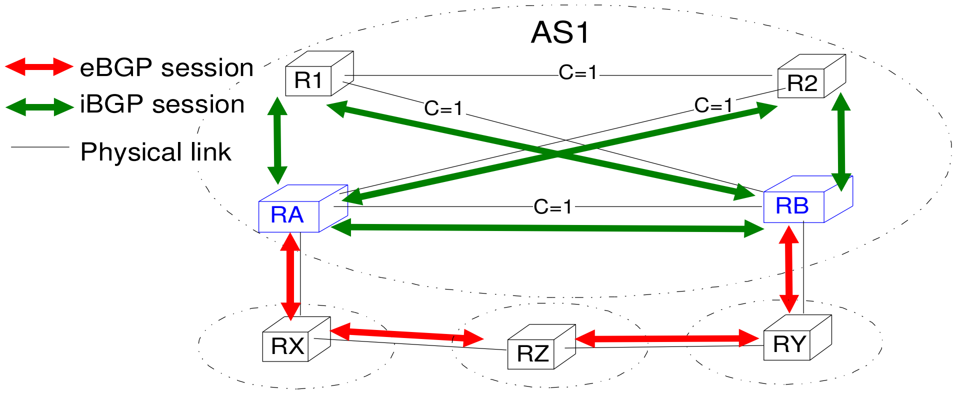
- Issues
  - Which AS Path will be advertised by AS0
    - BGP only allows to advertise one path
    - Downstream routers will not be aware of the path followed
      - Beware of routing loops !

Those multipath extensions are supported by several vendors, see:
http://www.cisco.com/univercd/cc/td/doc/product/software/ios122/122newft/122

http://www.juniper.net/techpubs/software/junos53/swconfig53-ipv6/html/ipv6-bg

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

Besides considering equal cost paths for load balancing, some vendors also support unequal load balancing by relying on the link bandwidth extended community that allows routers to determine the bandwidth of external links.
See :
S. Sangli, D. Tappan, Y. Rekhter, BGP Extended Communities Attribute, Internet draft, work in progress, Nov. 2002
http://www.ietf.org/internet-drafts/draft-ietf-idr-bgp-ext-communities-05.txt

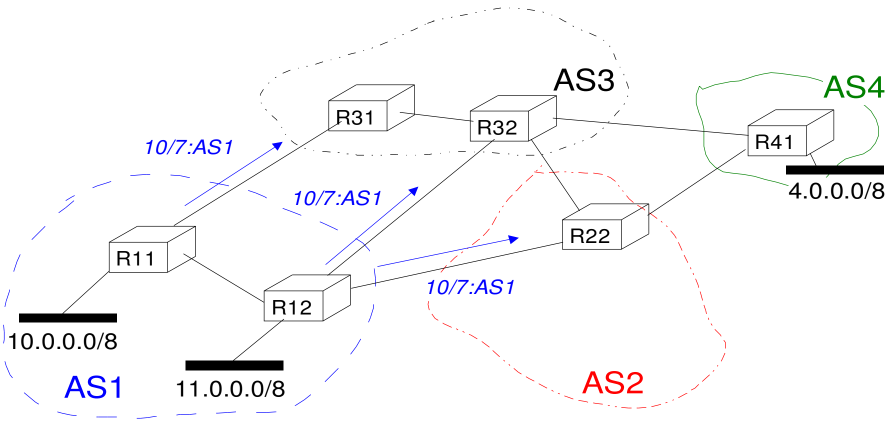
For a vendor usage of this community, see :
http://www.cisco.com/en/US/products/sw/iosswrel/ps1839/products_feature_gu

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

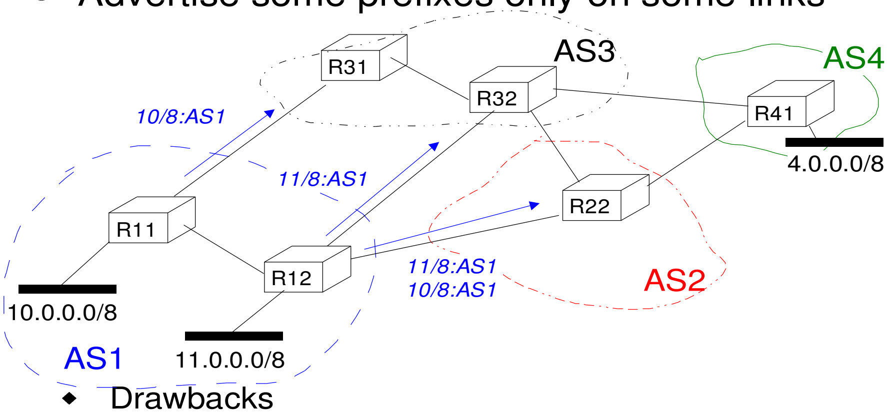
- Routing without tuning the announcements
  - packet flow towards AS1 will depend on the tuning of the decision process of AS2, AS3 and AS4



In this example, we assume that no filters are applied by AS2, AS3 and AS4 on the routes received from AS1.

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
  - No redundancy in case of link failure

In this example, AS1 forces AS3 to send the packets towards 10.0.0.0/8 on the R31-R11 link and the packets towards 11.0.0.0/8 on the R32-R12 link. This is a common method used to balance traffic over external links, but an important drawback is that if the R11-R31 link fails, AS3 would not be able to utilize the R12-R32 link to reach 10.0.0.0/8 and would be forced to used the path through AS2.

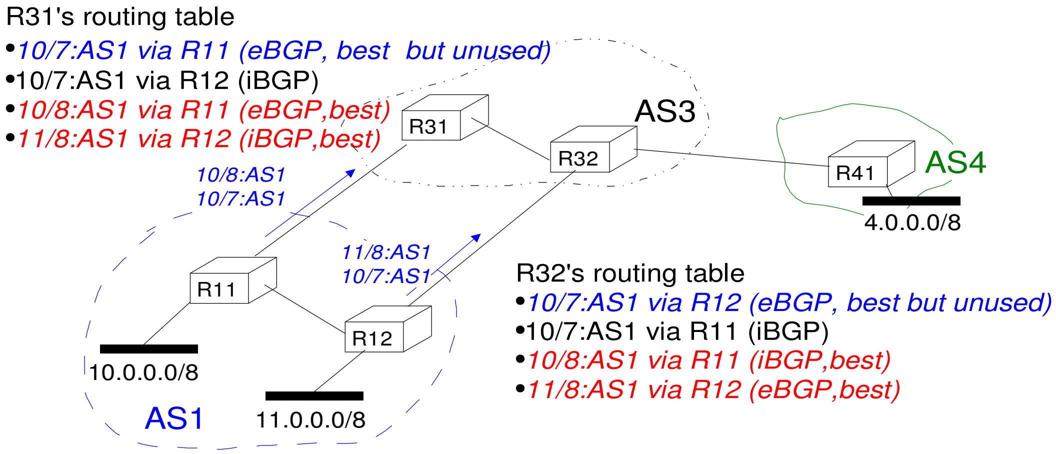
Note that if R12 advertised 10.0.0.0/7 instead of advertising both 10.0.0.0/8 and 11.0.0.0/8, then, most of the traffic could be received via AS3 since AS3 is advertising a more specific prefix (see later).

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



Compared with the utilization of the selective announcements, the main advantage of using more specific prefixes is that if link R11-R31 fails, then the packets towards 10.0.0.0/8 will still be sent by AS3 through the R32-R12 link since they are part of the 10.0.0.0/7 router learned from R12.

An important drawback of this solution is that it unnecessarily increases the size of the BGP routing tables of all routers on the Internet. For this reason, several ISPs block prefixes that are too long. For example, some ISPs do not accept prefixes longer than /22, and other try to filter prefixes based on the allocation rules of the regional IP address registries.

For more information on this filtering, see :

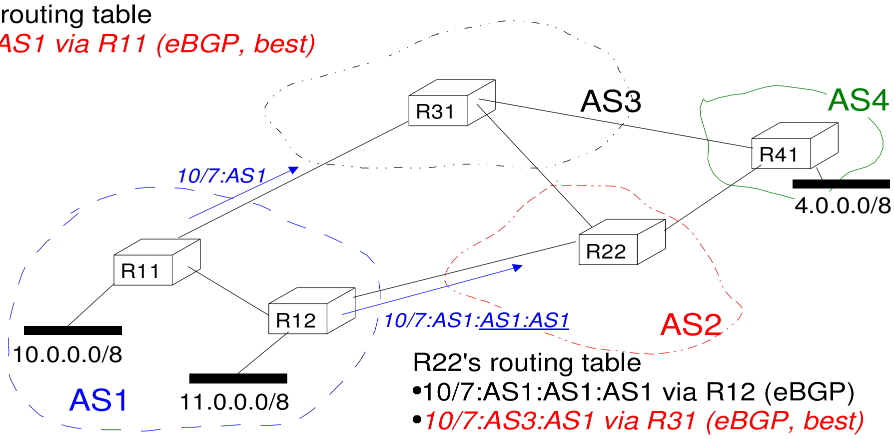
S. Bellovin et al., Slowing routing table growth by filtering on address allocation policies, preprint available from http://www.research.att.com/~jrex , June 2001

Note that if AS1 wants to use the more selective prefixes only to control the traffic on its links with AS3 and not beyond, then, the more specific prefixes should be advertised with the NO_EXPORT community while 10.0.0.0/7 would be advertised without community values. With this community value, the two more specific prefixes will not be advertised by AS3 and thus will not contribute to the growth of the global BGP routing table.

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)

AS-Path prepending is a popular technique since in the BGP decision process, the selection of the shortest AS-Path is one of the most important criteria. In theory, the length of the AS-Path is not necessarily an indication of the quality of a path, but some studies have shown that, on average, short AS-Paths offered a better performance that longer paths.

More information on these studies may be found in :

A. Broido et al., Internet expansion : refinement and churn, European Transactions on Telecommunications, special issue on traffic engineering, January 2002

Due to the importance of the "shortest AS-Path" criteria in the BGP decision process, most interdomain routes used in the Internet are relatively short (up to 3-4 transit AS between source and destination for most routes).

See
http://ipmon.sprintlabs.com/paccess/routestat/trends.php?type=addrReachabili

for some information on the addresses that are reachable at N AS hops from a large ISP like Sprint.

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)

E. Chen, and T. Bates, "An Application of the BGP Community Attribute in Multi-home Routing", RFC 1998, August 1996.

A detailed survey of the utilization of the community attribute today may be found in :

B. Quoitin and O. Bonaventure, A survey of the utilization of the BGP community attribute, Technical Report Infonet-TR-2002-02, Feb 2002, available from http://www.infonet.fundp.ac.be/doc/tr/

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)

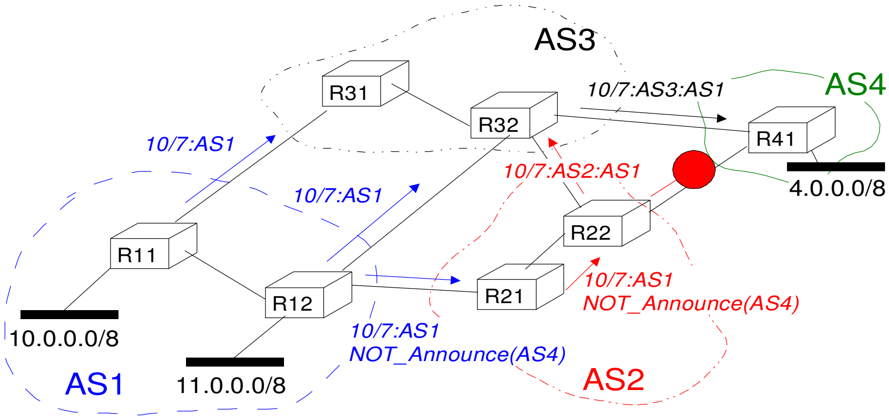
The BGP redistribution communities are described in :

O. Bonaventure et al., Controlling the redistribution of BGP routes
Internet draft, draft-ietf-ptomaine-redistribution-01.txt, work in progress, August 2002

An implementation of these communities in zebra is described in :

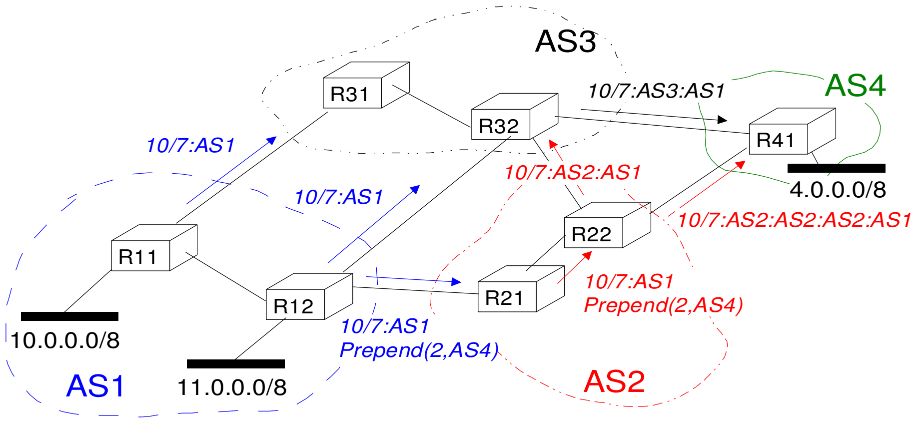
B. Quoitin, An implementation of the BGP redistribution communities in Zebra, Technical report Infonet-TR-2002-03, Feb 2002
http://www.infonet.fundp.ac.be/doc/tr/Infonet-TR-2002-03.html

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

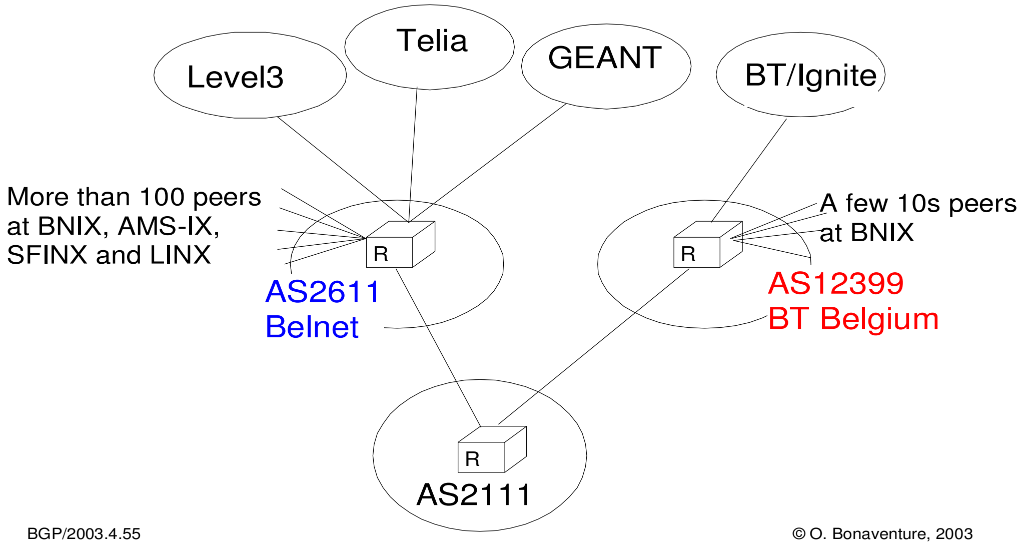
- Advantages and drawbacks
  - Selective announcements
    - always work, but if one prefix is advertised on a single link, it may become unreachable in case of failure
  - More specific prefixes
    - better than selective announcements in case of failure
    - but increases significantly the size of 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 studies

AS-Path prepending and communities in practice

- An experiment in the global Internet



This evaluation was carried out by Cristel Pelsser in March-April 2003. The links with the two upstream providers were GRE tunnels. Those measurements could not have been done without the help of Jan Torrele (Belnet), Benoît Piret (BT) and Patrice Devemy (Skynet). This evaluation should be considered as an experiment and not as a "comparison" between Belnet and BT Belgium.

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 BT
- With prepending once on Belnet link
  - 22% received via Belnet, 78% received via BT
- With prepending twice on Belnet link
  - 15% received via Belnet, 84% received via BT

When prepending was used on the BT link, the following results were obtained :

- With prepending once on BT link
  - 80% received via Belnet, 20% received via BT
- With prepending twice on BT link
  - 80% received via Belnet, 20% received via BT
- With prepending three times on BT link
  - All traffic was received via Belnet

How to better balance the incoming traffic ?

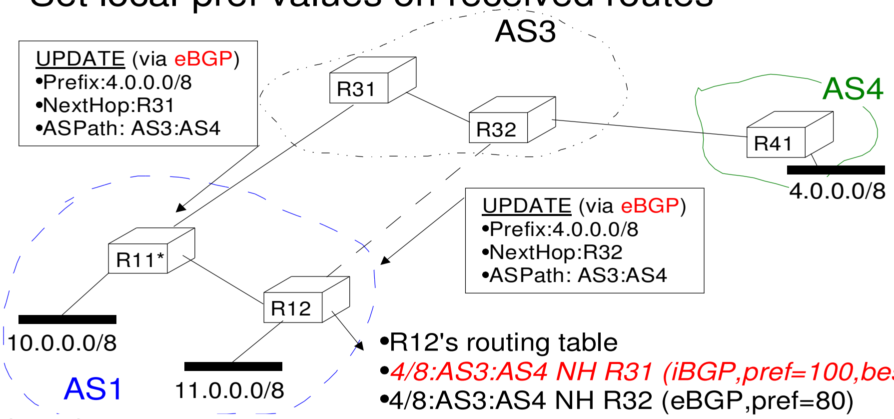
- AS Path prepending is clearly not sufficient
- Can we do better with the communities ?
  - Need to move some traffic from one upstream to another
- Level3 Communities
  - 65000:0
    - announce to customers but not to peers
  - 65000:XXX
    - do not announce to peer ASXXX
  - 65001:0
    - prepend 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

Community-based traffic engineering

- Study with 56k prefix from global Internet
  - For each prefix, sent TCP SYN on port 80 and measure from which upstream reply came back
- Results
  - Without communities
    - 68 % received via Belnet, 32% received via BT
  - With community 65000:0
    - Level3 does not announce to peers
    - 45% received via Belnet, 55% received via BT
  - With communities 1299:2009 and 1299:5009
    - Telia does not announce to US and EU peers
    - 63% received via Belnet, 36% received via BT

Case study 1
Stub with one provider and a backup link

- Control of the outgoing traffic
  - Set local-pref values on received routes



•R12's routing table
•*4/8:AS3:AS4 NH R31 (iBGP,pref=100,best)*
•4/8:AS3:AS4 NH R32 (eBGP,pref=80)

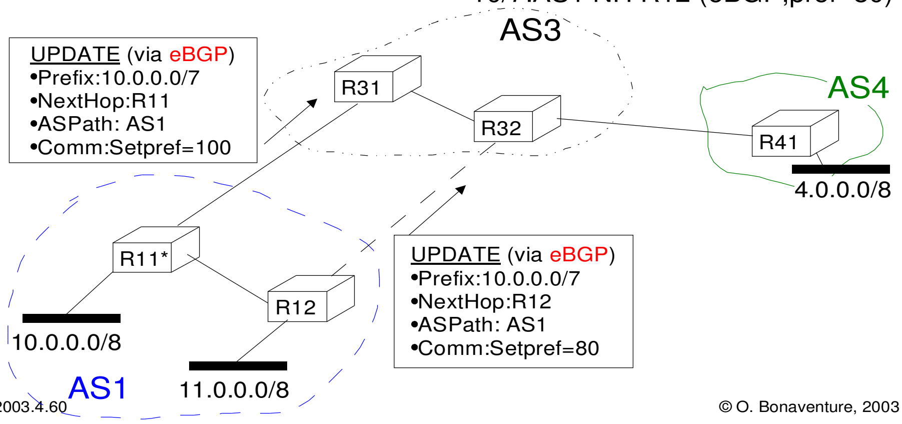
R11's routing table
•*4/8:AS3:AS4 NH R31 (eBGP,pref=100,best)*
•4/8:AS3:AS4 NH R32 (iBGP,pref=80)

Case study 1
Stub with one provider and a backup link

- Control of the incoming traffic
  - utilize communities to set local-pref in provider

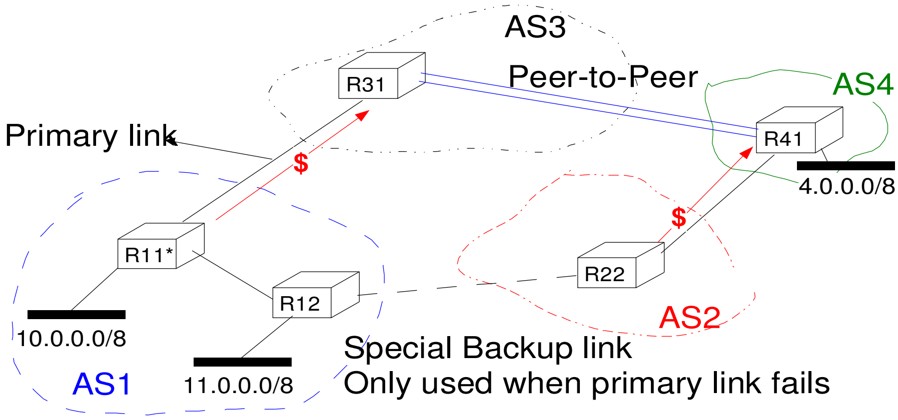
R31's routing table
- *10/7:AS1 NH R11 (eBGP,pref=100,best)*
- 10/7:AS1 NH R12 (iBGP,pref=80)

R32's routing table
- *10/7:AS1 NH R11 (iBGP,pref=100,best)*
- 10/7:AS1 NH R12 (eBGP,pref=80)



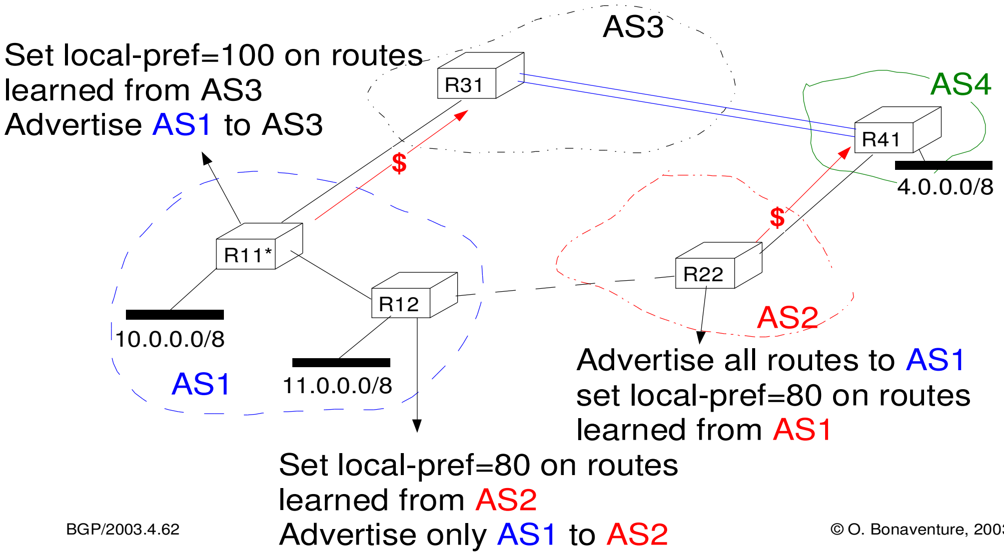
Case study 2
Stub with provider and external backup

- How to allow link AS1-AS2 to serve as a backup in case of failure of AS1-AS3 ?



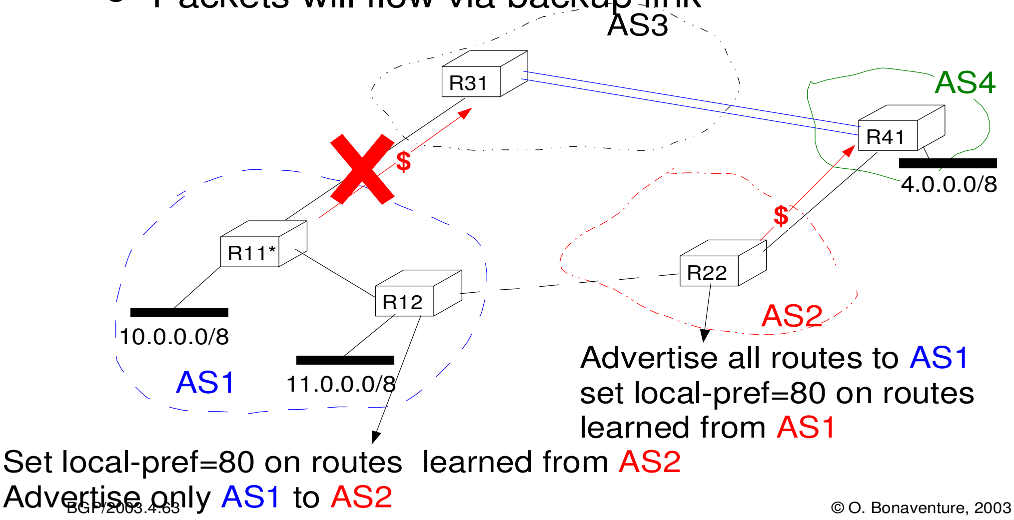
Case study 2
Stub with provider and external backup (2)

- Control of the outgoing traffic



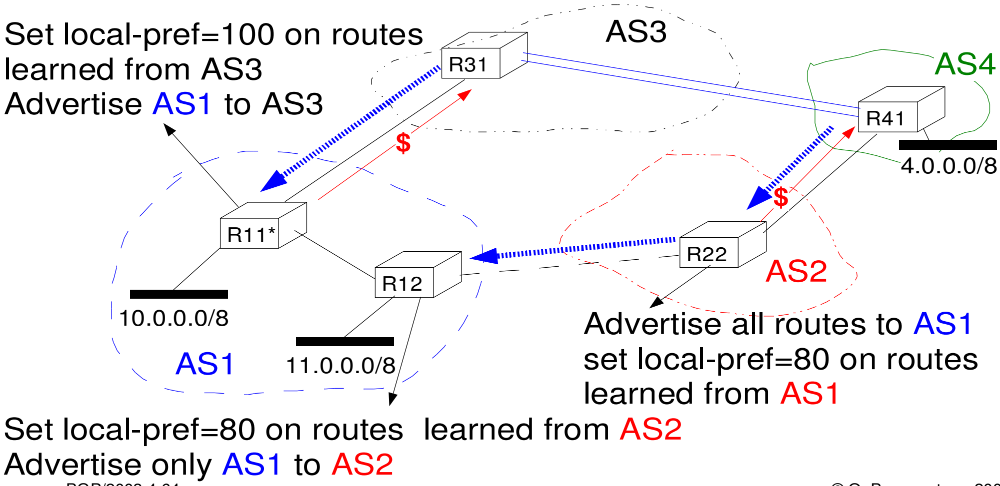
Case study 2
Stub with provider and external backup (3)

- What happens when primary link fails ?
  - Packets will flow via backup link



Case study 2
Stub with provider and external backup (4)

- What happens when primary link is back ?
  - Some packets will still flow via backup link...



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

Thank you

Questions and comments can be sent to

Olivier Bonaventure

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

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

To be informed about updates to this tutorial, send an email to Bonaventure@info.ucl.ac.be .