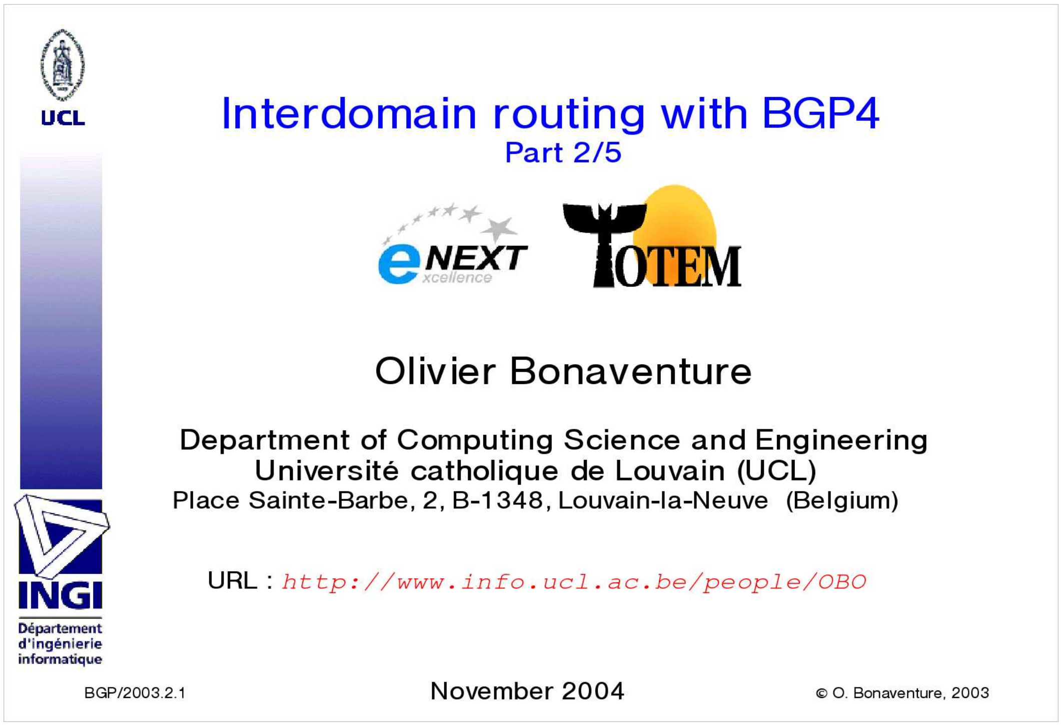


Interdomain routing with BGP4

Part 2/5

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

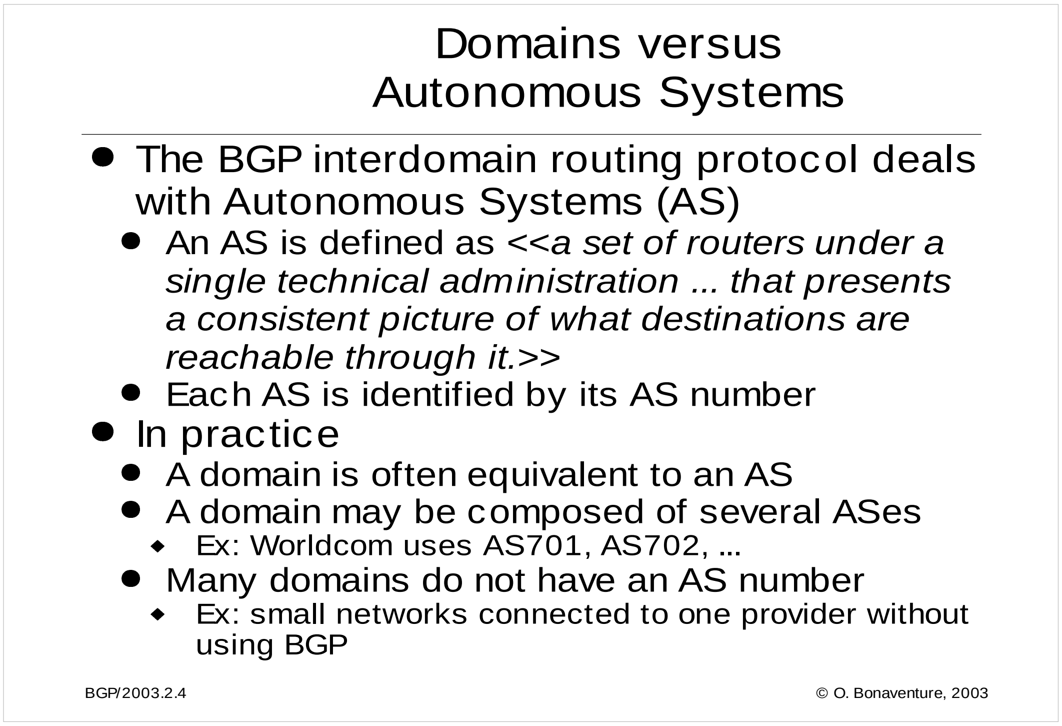
November 2004

Outline

- Organization of the global Internet
- BGP basics
  - Routing policies
  - The Border Gateway Protocol
  - How to prefer some routes over others
- BGP in large networks
- Interdomain traffic engineering with BGP
- BGP-based Virtual Private Networks

Interdomain routing

- Goals
  - Allow to transmit IP packets along the best path towards their destination through several transit domains while taking into account the routing policies of each domain without knowing the detailed topology of those domains
    - From an interdomain viewpoint, best path often means *cheapest* path
    - Each domain is free to specify inside its routing policy the domains for which it agrees to provide a transit service and the method it uses to select the best path to reach each destination



Domains versus Autonomous Systems

- The BGP interdomain routing protocol deals with Autonomous Systems (AS)
  - An AS is defined as <<*a set of routers under a single technical administration ... that presents a consistent picture of what destinations are reachable through it.*>>
  - Each AS is identified by its AS number
- In practice
  - A domain is often equivalent to an AS
  - A domain may be composed of several ASes
    - Ex: Worldcom uses AS701, AS702, ...
  - Many domains do not have an AS number
    - Ex: small networks connected to one provider without using BGP

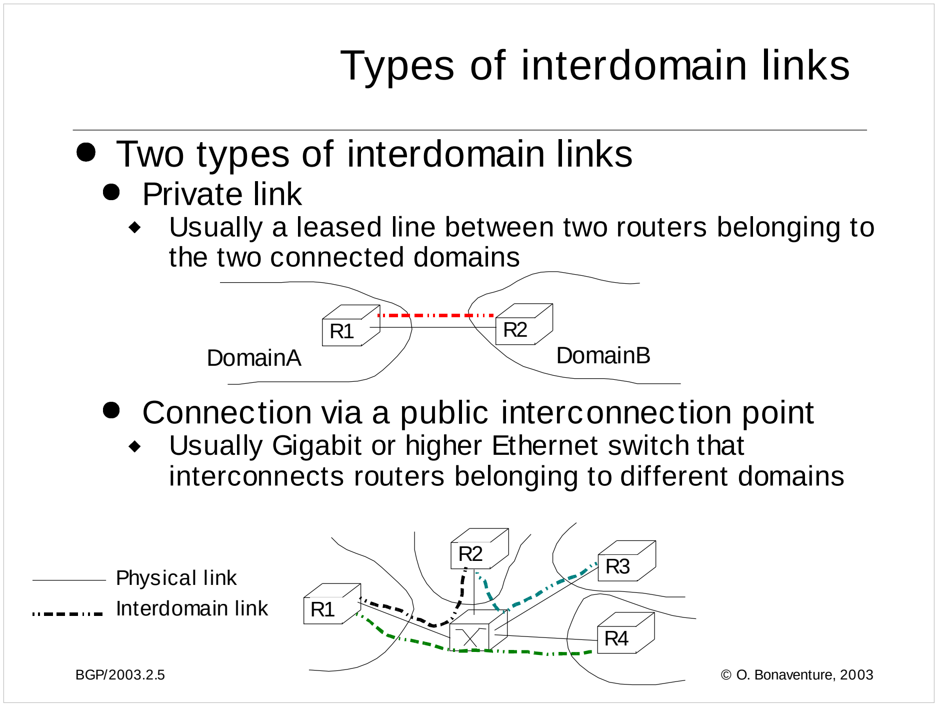
In the remainder of the tutorial, we will consider domains and Autonomous Systems as equivalent concepts.

Each AS on the Internet has been assigned a 16bits AS number by the Regional Internet Registries. For a current list of assigned AS numbers, see:

http://www.cidr-report.org/autnums.html

More information may be found in the whois databases :

http://whois.ripe.net
http://www.radb.net/



Types of interdomain links

- Two types of interdomain links
  - Private link
    - Usually a leased line between two routers belonging to the two connected domains

  - Connection via a public interconnection point
    - Usually Gigabit or higher Ethernet switch that interconnects routers belonging to different domains

For more information on the organization of the Internet, see :

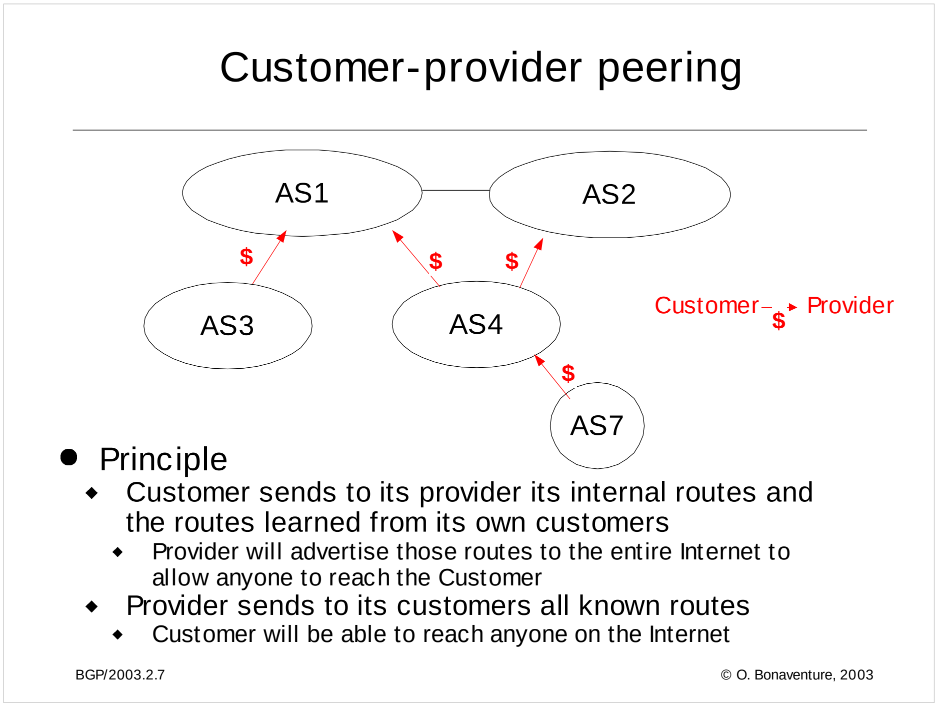
G. Huston, Peerings and settlements, Internet Protocol Journal, Vol. 2, N1 et 2, 1999,
http://www.cisco.com/warp/public/759/ipj_Volume2.html

For more information on interconnection points or Internet exchanges, see :

http://www.euro-ix.net/
http://www.ripe.net/ripe/wg/eix/index.html
http://www.ep.net/ep-main.html

Routing policies

- In theory BGP allows each domain to define its own routing policy…
- In practice there are two common policies
  - customer-provider peering
    - **Customer c** buys Internet connectivity from **provider P**
  - shared-cost peering
    - **Domains x** and **y** agree to exchange packets by using a direct link or through an interconnection point



Customer-provider peering

- Principle
  - Customer sends to its provider its internal routes and the routes learned from its own customers
    - Provider will advertise those routes to the entire Internet to allow anyone to reach the Customer
  - Provider sends to its customers all known routes
    - Customer will be able to reach anyone on the Internet

On link AS7-AS4

AS7 advertises its own routes to AS4

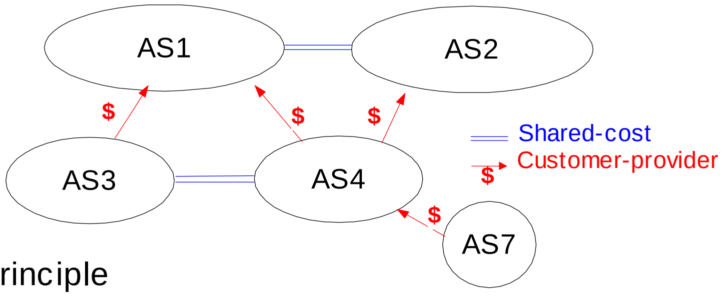
AS4 advertises to AS7 the routes that allow to reach the entire Internet

On link AS4-AS2

AS4 advertises its own routes and the routes belonging to AS7

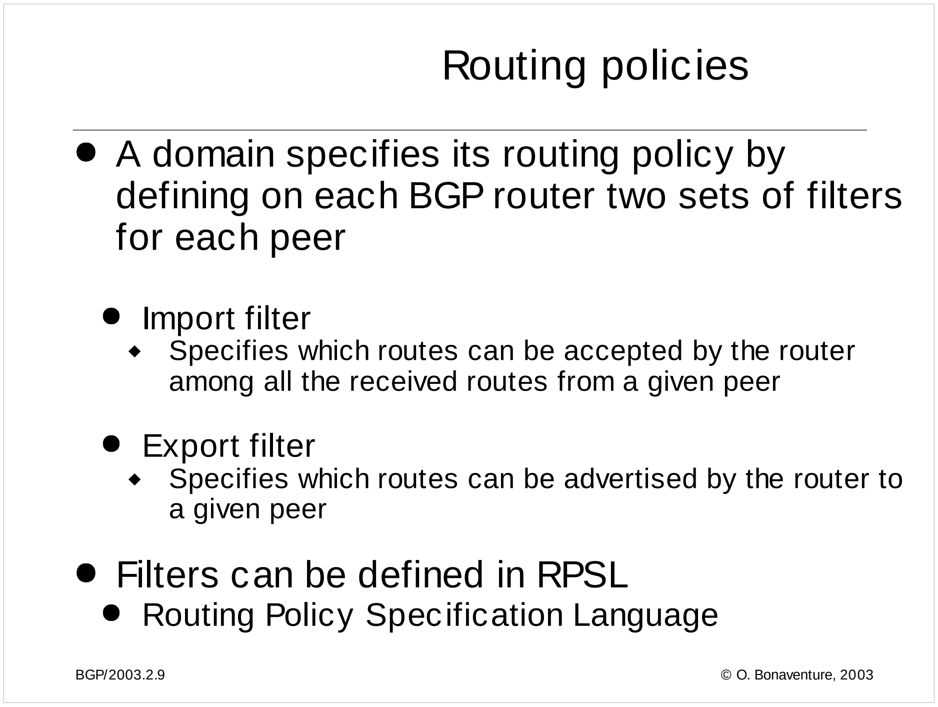
AS2 advertises the routes that allow to reach the entire Internet

Shared-cost peering



- Principle
  - PeerX sends to PeerY its internal routes and the routes learned from its own customers
    - PeerY will use shared link to reach PeerX and PeerX's customers
    - PeerX's providers are not reachable via the shared link
  - PeerY sends to PeerX its internal routes and the routes learned from its own customers
    - PeerX will use shared link to reach PeerY and PeerY's customers
    - PeerY's providers are not reachable via the shared link

On link AS3-AS4
    AS3 advertises its internal routes
    AS4 advertises its internal routes and the routes learned
        from AS7 (its customer)
On link AS1-AS2
    AS1 advertises its internal routes and the routes received
        from AS3 and AS4 (its customers)
    AS2 advertises its internal routes and the routes learned
        from AS74(its customer)



Routing policies

- A domain specifies its routing policy by defining on each BGP router two sets of filters for each peer
  - Import filter
    - Specifies which routes can be accepted by the router among all the received routes from a given peer
  - Export filter
    - Specifies which routes can be advertised by the router to a given peer
- Filters can be defined in RPSL
  - Routing Policy Specification Language

RFC 2622 Routing Policy Specification Language (RPSL). C. Alaettinoglu, C. Villamizar, E. Gerich, D. Kessens, D. Meyer, T. Bates, D. Karrenberg, M. Terpstra. June 1999.

RFC 2650 Using RPSL in Practice. D. Meyer, J. Schmitz, C. Orange, M. Prior, C. Alaettinoglu. August 1999.

Internet Routing Registries contain the routing policies of various ISPs, see :

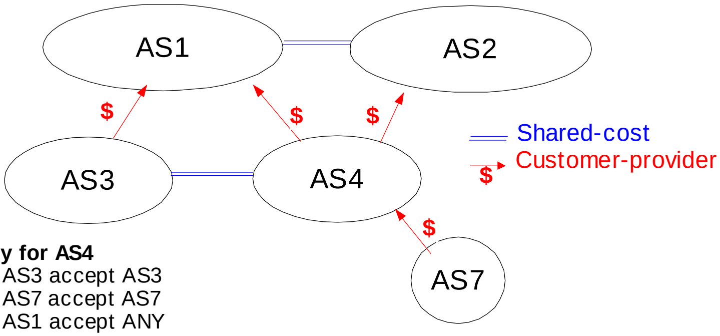
http://www.ripe.net/ripencc/pub-services/whois.html
http://www.arin.net/whois/index.html
http://www.apnic.net/apnic-bin/whois.pl

RPSL

- Simple import policies
  - Syntax
    - `import: from AS# accept list_of_AS`
  - Examples
    - `Import: from Belgacom accept Belgacom WIN`
    - `Import: from Provider accept ANY`

- Simple export policies
  - Syntax
    - `Export: to AS# announce list_of_AS`
  - Example
    - `Export: to Customer announce ANY`
    - `Export: to Peer announce Customer1 Customer2`

Routing policies
Simple example with RPSL



Import policy for AS4
Import: from AS3 accept AS3
import: from AS7 accept AS7
import: from AS1 accept ANY
import: from AS2 accept ANY

Export policy for AS4
export: to AS3 announce AS4 AS7
export: to AS7 announce ANY
export: to AS1 announce AS4 AS7
export: to AS2 announce AS4 AS7

Import policy for AS7
Import: from AS4 accept ANY

Export policy for AS4
export: to AS4 announce AS7

Scalable routing policies with RPSL

- How to specify policies of large domains ?
  - Define one `route` object for each advertised prefix
    - `route:` prefix
    - `descr:` human-readable description
    - `origin:` AS# advertising the prefix
  - Define one as-set for all the clients of a given AS
    - `as-set:` macro name
    - `descr:` human-readable description
    - `members:` list of clients AS#
  - Specify the routing policies by using `as-sets` instead of AS numbers whenever possible

Scalable routing policies with RPSL (2)

- Example

```
aut-num:    AS20965
as-name:     GEANT
descr:      The GEANT IP Service
...
import:     from AS2611 action pref=100;accept AS-BELNET
...
export:     to AS2611 announce AS-GEANTNRN ...
```

```
as-set:      AS-BELNET
descr:       BELNET AS Macro
members:       AS2611, AS15383, AS9208, AS2111
```

```
route:       130.104.0.0/16
descr:        NET-UCLOUVAIN
origin:      AS2611
...
route:       138.48.0.0/16
descr:        FUNDP-AC-BE
origin:      AS2611
```

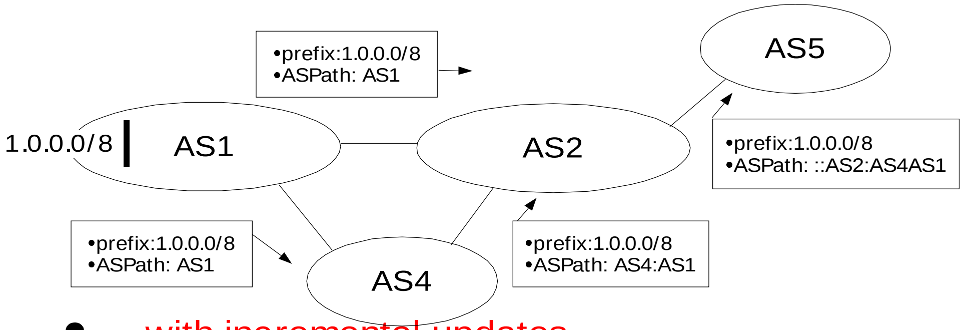
```
route:       81.19.48.0/20
descr:        IST-ATRIUM-EXP-20030212
origin:      AS2111
```

Outline

- Organization of the global Internet
- BGP basics
  - Routing policies
  - The Border Gateway Protocol
  - How to prefer some routes over others
- BGP in large networks
- Interdomain traffic engineering with BGP
- BGP-based Virtual Private Networks

The Border Gateway Protocol

- Principle
  - Path vector protocol
    - BGP router advertises its best route to each destination



•prefix:1.0.0.0/8
•ASPath: AS1

•prefix:1.0.0.0/8
•ASPath: ::AS2:AS4AS1

•prefix:1.0.0.0/8
•ASPath: AS1

•prefix:1.0.0.0/8
•ASPath: AS4:AS1

  - ... with incremental updates
    - Advertisements are only sent when their content changes

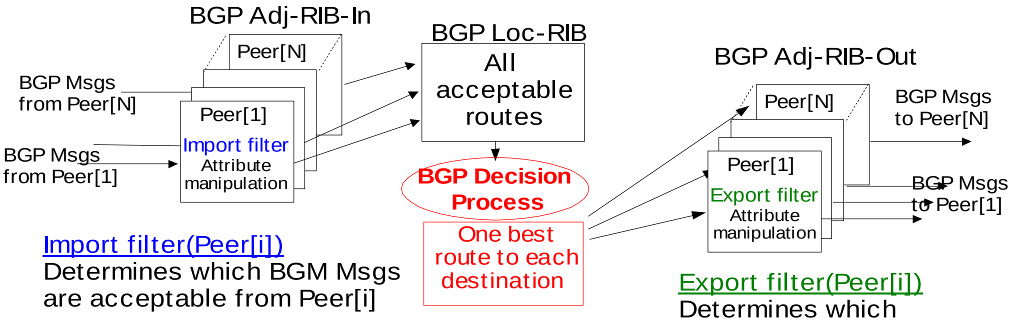
"Origin" of the routes announced by BGP

- Where do the routes announced by a BGP router come from ?
  - Learned from other BGP routers
    - BGP router only propagates the received routes
  - Static configuration
    - BGP router is configured to advertise some prefixes
    - Drawback : requires manual configuration
    - Advantage : Stable set of advertised prefixes
  - Learned from an Interior Gateway Protocol
    - The prefixes received from the IGP are advertised by the BGP router usually as an aggregate
    - Advantage
      - BGP advertisements follow network state, prefix is automatically withdrawn by BGP it is not reachable via IGP
    - Drawback
      - BGP announcements will be unstable if IGP is unstable...

Policies and BGP

- Two mechanisms to support policies in BGP
  - Each domain defines itself which is the best route to reach each destination based on the routes learned from its peers
    - The chosen best route is not necessarily the "shortest" route as with IGPs
    - Only the best route towards each destination can be announced to external peers
  - Each domain determines, on its own, which routes can be advertised to each peer
    - An AS does not necessarily advertise to all its neighbors all the routes that it knows

Conceptual model of a BGP router



Import filter(Peer[i])
Determines which BGM Msgs
are acceptable from Peer[i]

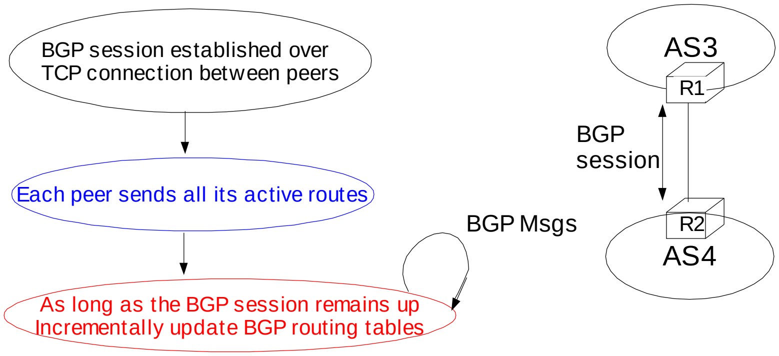
Export filter(Peer[i])
Determines which
routes can be sent to Peer[i]

BGP Routing Information Base
Contains all the acceptable routes
learned from all Peers + internal routes

- BGP decision process selects *the* best route towards each destination

BGP : Principles of operation

- Principles
  - BGP relies on the incremental exchange of path vectors

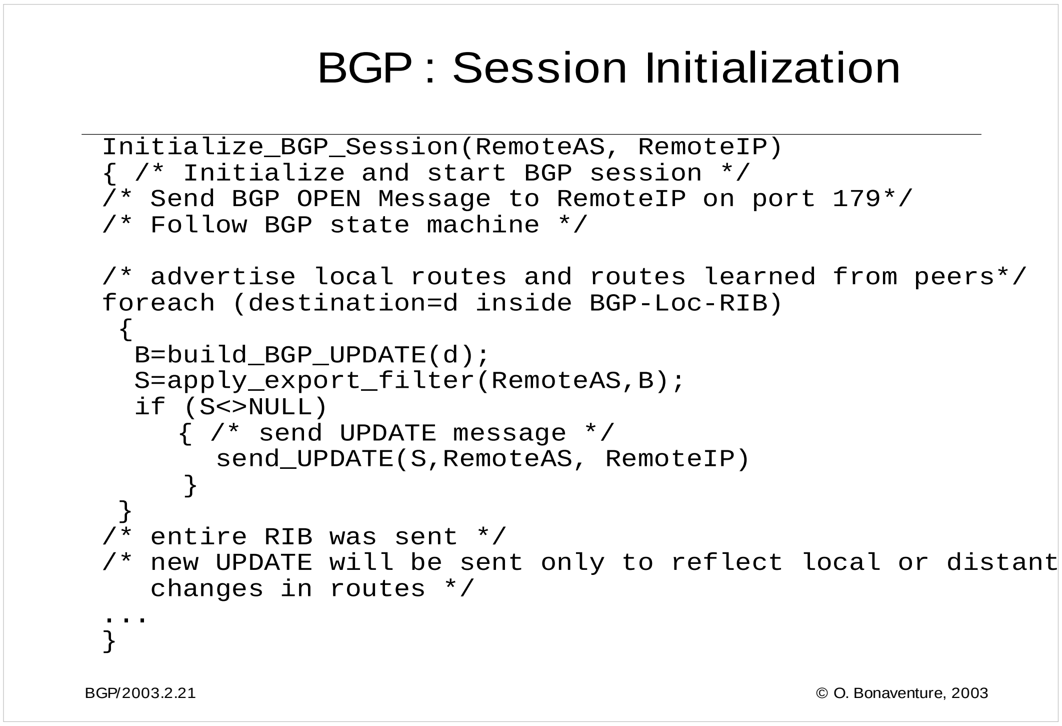

BGP session established over TCP connection between peers
Each peer sends all its active routes
As long as the BGP session remains up Incrementally update BGP routing tables
BGP Msgs
AS3
R1
BGP session
R2
AS4

BGP : Principles of operation (2)

- Simplified model of BGP
  - 2 types of BGP path vectors

  - UPDATE
    - Used to announce a route towards one prefix
    - Content of UPDATE
      - Destination address/prefix
      - Interdomain path used to reach destination (AS-Path)
      - Nexthop (address of the router advertising the route)

  - WITHDRAW
    - Used to indicate that a previously announced route is not reachable anymore
    - Content of WITHDRAW
      - Unreachable destination address/prefix



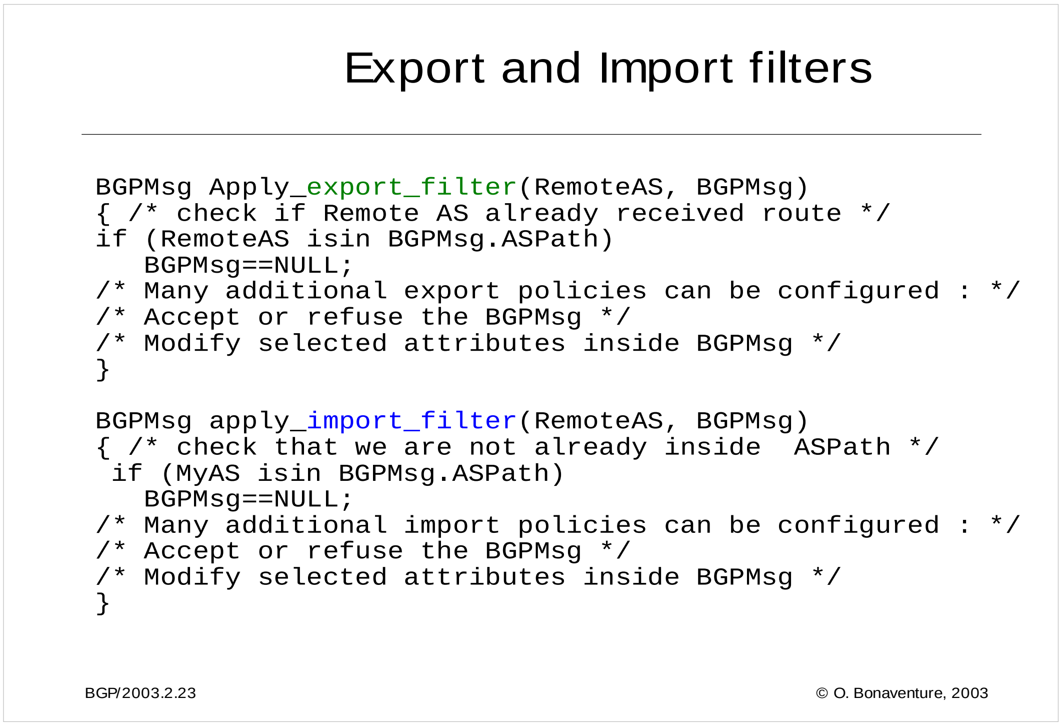
BGP : Session Initialization

```
Initialize_BGP_Session(RemoteAS, RemoteIP)
{ /* Initialize and start BGP session */
/* Send BGP OPEN Message to RemoteIP on port 179*/
/* Follow BGP state machine */

/* advertise local routes and routes learned from peers*/
foreach (destination=d inside BGP-Loc-RIB)
 {
  B=build_BGP_UPDATE(d);
  S=apply_export_filter(RemoteAS,B);
  if (S<>NULL)
     { /* send UPDATE message */
       send_UPDATE(S,RemoteAS, RemoteIP)
      }
 }
/* entire RIB was sent */
/* new UPDATE will be sent only to reflect local or distant
   changes in routes */
...
}
```

Events during a BGP session

1. Addition of a new route to RIB
   - A new internal route was added on local router
     - static route added by configuration
     - Dynamic route learned from IGP
   - Reception of UPDATE message announcing a new or modified route
2. Removal of a route from RIB
   - Removal of an internal route
     - Static route is removed from router configuration
     - Intradomain route declared unreachable by IGP
   - Reception of WITHDRAW message
3. Loss of BGP session
   - All routes learned from this peer removed from RIB



Export and Import filters

```
BGPMsg Apply_export_filter(RemoteAS, BGPMsg)
{ /* check if Remote AS already received route */
if (RemoteAS isin BGPMsg.ASPath)
   BGPMsg==NULL;
/* Many additional export policies can be configured : */
/* Accept or refuse the BGPMsg */
/* Modify selected attributes inside BGPMsg */
}

BGPMsg apply_import_filter(RemoteAS, BGPMsg)
{ /* check that we are not already inside  ASPath */
 if (MyAS isin BGPMsg.ASPath)
   BGPMsg==NULL;
/* Many additional import policies can be configured : */
/* Accept or refuse the BGPMsg */
/* Modify selected attributes inside BGPMsg */
}
```

In the above export filter, we assume that the BGP sender does not send to PeerX the routes learned from this peer. This behavior is not required by the BGP specification, but is a common optimization, often called sender-side loop detection.

The check for the presence of the localAS number in the routes learned is specified in the BGP RFC.

BGP : Processing of UPDATES

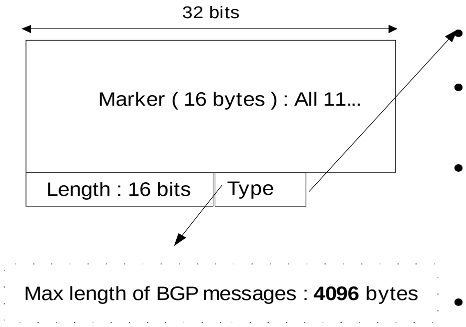
```
Recvd_BGPMsg(Msg, RemoteAS)
{
 B=apply_import_filer(Msg,RemoteAS);
 if (B==NULL) /* Msg not acceptable */
     exit();
 if IsUPDATE(Msg)
 {
  Old_Route=BestRoute(Msg.prefix);
  Insert_in_RIB(Msg);
  Run_Decision_Process(RIB);
  if (BestRoute(Msg.prefix)<>Old_Route)
  { /* best route changed */
    B=build_BGP_Message(Msg.prefix);
    S=apply_export_filter(RemoteAS,B);
    if (S<>NULL) /* announce best route */
     send_UPDATE(S,RemoteAS);
    else if (Old_Route<>NULL)
     send_WITHDRAW(Msg.prefix);
 } ...
```

BGP : Processing of WITHDRAW

```
Recvd_Msg(Msg, RemoteAS)
...
if IsWITHDRAW(Msg)
 {
  Old_Route=BestRoute(Msg.prefix);
  Remove_from_RIB(Msg);
  Run_Decision_Process(RIB);
  if (Best_Route(Msg.prefix)<>Old_Route)
  { /* best route changed */
    B=build_BGP_Message(d);
    S=apply_export_filter(RemoteAS,B);
    if (S<>NULL) /* still one best route */
       send_UPDATE(S,RemoteAS, RemoteIP);
    else if(Old_Route<>NULL)/* no best route anymore */
       send_WITHDRAW(Msg.prefix,RemoteAS,RemoteIP);
  }
 }
}
```

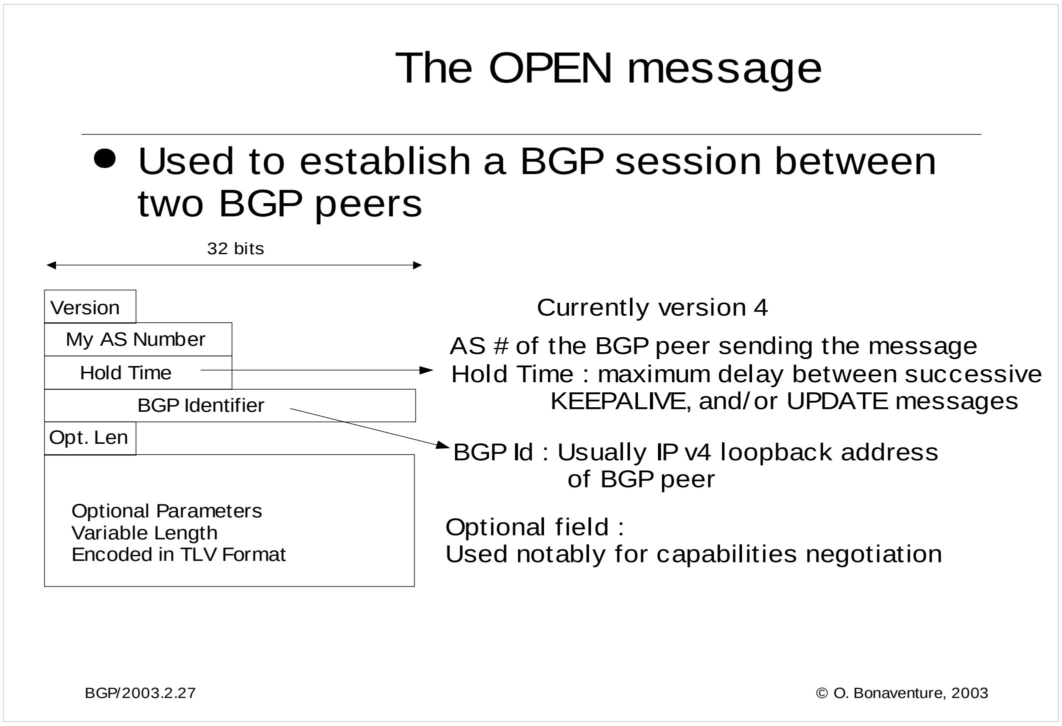
The BGP messages

- Variable length messages
  - With fixed size header



Max length of BGP messages : **4096** bytes

- OPEN
  used to establish BGP session
- UPDATE
  used to send new routes and to remove unusable routes
- NOTIFICATION
  used to inform the remote peer of an error
  BGP session is closed upon transmission or reception of NOTIFICATION message
- KEEPALIVE
  one message must be sent at least every 30 seconds on each BGP session
- ROUTE_REFRESH
  used to support graceful restart



The OPEN message

- Used to establish a BGP session between two BGP peers

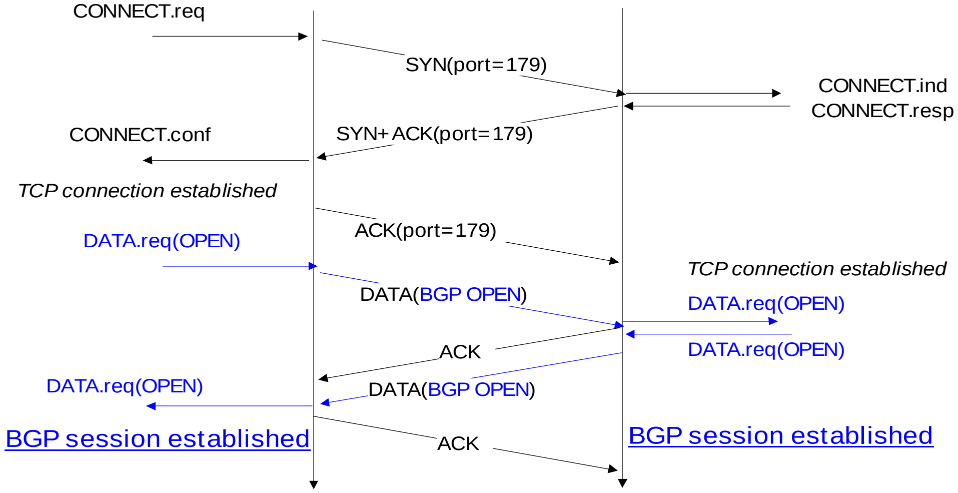
32 bits

Version	Currently version 4
My AS Number	AS # of the BGP peer sending the message
Hold Time	Hold Time : maximum delay between successive KEEPALIVE, and/or UPDATE messages
BGP Identifier	BGP Id : Usually IP v4 loopback address of BGP peer
Opt. Len	
Optional Parameters Variable Length Encoded in TLV Format	Optional field : Used notably for capabilities negotiation

Inside the OPEN message, and also in the Path attributes of the UPDATE message, the AS number is encoded as a 16 bits field. This limits the number of Ases in the global Internet. Given the rapid growth in the number of AS present on the Internet, the AS space could become completely full within a few years.

Work in under way to allow BGP to support 32 bits wide AS numbers. See Q. Vohra, E. Chen, "BGP support for four-octet AS number space", Work in Progress, <draft-ietf-idr-as4bytes-04.txt>, September 2001.

Establishment of a BGP session



Usually, a BGP session can only be established between two manually configured peers. Each peer needs to be configured with the IP address and the AS number of the remote peer.

For a security point of view, several solutions have been proposed to ensure that a BGP session will not be hijacked :

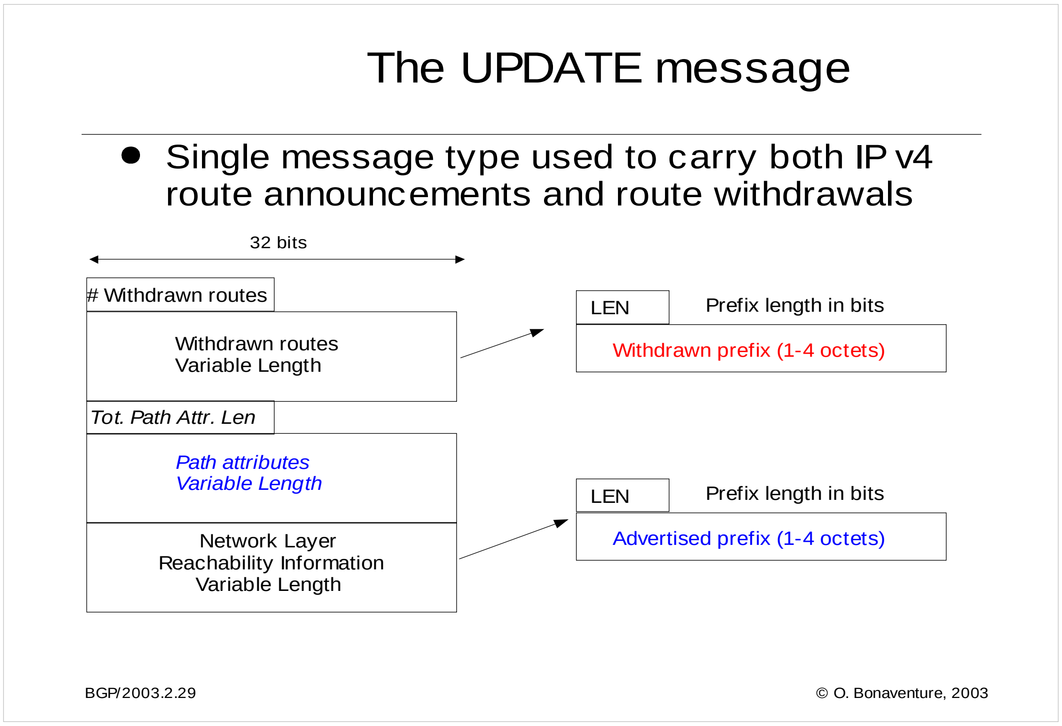
- One solution is to protect the TCP connection with MD5 digests. See
, A. Heffernan, Protection of BGP Sessions via the TCP MD5 Signature Option
, RFC2385, August 1998
- Another solution is to utilize IP packets with a TTL value of 255 on single-hop eBGP sessions :
V. Gill, J. Heasley, D. Meyer, The BGP TTL Security Hack (BTSH), Internet draft, draft-gill-btsh-00.txt , October 2002, Work in progress
- Another solution is to send the BGP session over an IPSec association

For a discussion of BGP security issues, see :

- Sandra Murphy, BGP Security Analysis, Internet draft, draft-murphy-bgp-secr-04.txt , work in progress, November 2001
- S. Murphy, BGP Security Vulnerabilities Analysis, Internet draft, draft-murphy-bgp-vuln-01.txt , work in progress, Oct. 2003

See also the RPSEC IETF working group

- http://www.ietf.org/html.charters/rpsec-charter.html



The UPDATE message

- Single message type used to carry both IP v4 route announcements and route withdrawals

32 bits

# Withdrawn routes

Withdrawn routes
Variable Length

Tot. Path Attr. Len

Path attributes
Variable Length

Network Layer
Reachability Information
Variable Length

LEN

Prefix length in bits

Withdrawn prefix (1-4 octets)

LEN

Prefix length in bits

Advertised prefix (1-4 octets)

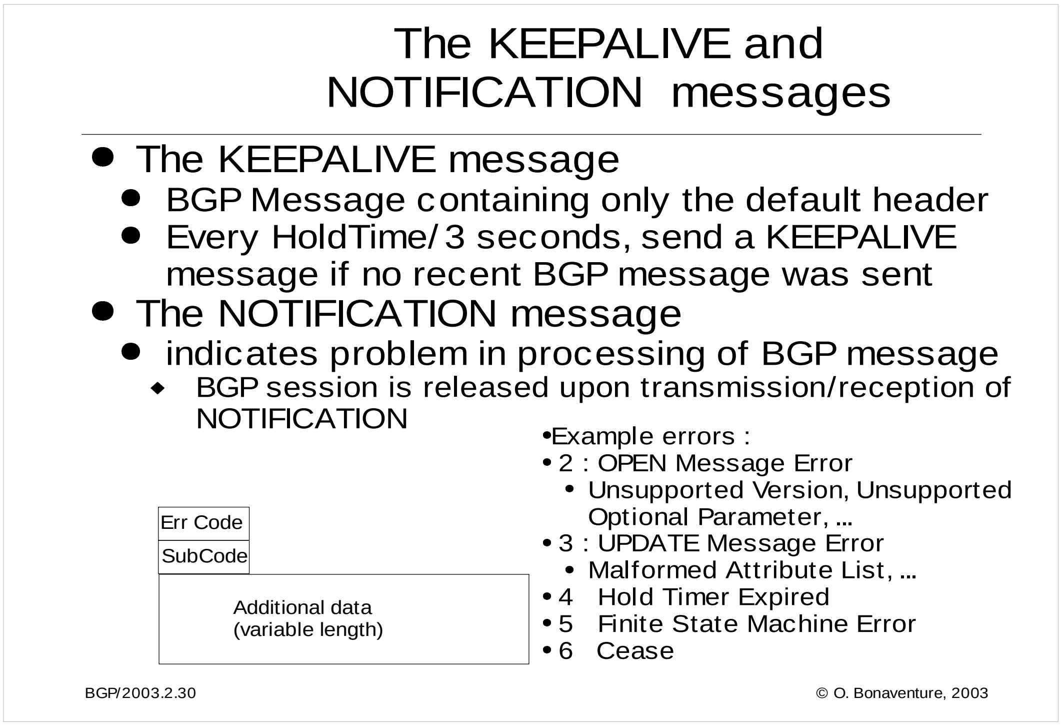
This format is used when BGP carries IP v4 routing information. With the MultiProtocol extensions, BGP can be used to carry different types of addresses instead the same BGP session (e.g. IP v6, RFC2547 VPNs, MPLS labels, or IP Multicast routing information). See e.g. :

P. Marques, F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, March 1999.

In this case, the capabilities optional parameter is used inside the OPEN message to negotiate the utilization of other addresses formats. Those non-IPv4 addresses are carried inside optional path attributes (MP_REACH_NLRI and MP_UNREACH_NLRI). Those attributes are encoded as described in :

T. Bates, R. Chandra, D. Katz, Y. Rekhter, Multiprotocol Extensions for BGP-4, Internet draft, draft-ietf-idr-rfc2858bis-02.txt, October 2002, work in progress

Being able to pack multiple route announcements and withdrawals in the same BGP message is very important for performance reasons, since a good packing of the BGP messages can significantly reduce the number of BGP messages exchanged. In this tutorial, for simplicity, we will only utilize BGP messages carrying an advertisement or a withdrawal for a single IP prefix. We will utilize the word “UPDATE” for a BGP UPDATE message containing a single advertised prefix and the word “WITDRAW” for a BGP UPDATE message containing a single withdrawn prefix.



The KEEPALIVE and NOTIFICATION messages

- The KEEPALIVE message
  - BGP Message containing only the default header
  - Every HoldTime/3 seconds, send a KEEPALIVE message if no recent BGP message was sent
- The NOTIFICATION message
  - indicates problem in processing of BGP message
    - BGP session is released upon transmission/reception of NOTIFICATION

Err Code

SubCode

Additional data
(variable length)

•Example errors :
- 2 : OPEN Message Error
  - Unsupported Version, Unsupported Optional Parameter, ...
- 3 : UPDATE Message Error
  - Malformed Attribute List, ...
- 4 Hold Timer Expired
- 5 Finite State Machine Error
- 6 Cease

The error codes and subcodes

•1: Message Header Error
- 1 : Connection not synchronized
- 2: : Bad message length
- 3 : Bad message type

• 2 : OPEN Message Error
- 1 - Unsupported Version Number.
- 2 - Bad Peer AS.
- 3 - Bad BGP Identifier.
- 4 - Unsupported Optional Parameter.
- 6 - Unacceptable Hold Time.

• 3 : UPDATE Message Error
- 1 - Malformed Attribute List.
- 2 - Unrecognized Well-known Attribute.
- 3 - Missing Well-known Attribute.
- 4 - Attribute Flags Error.
- 5 - Attribute Length Error.
- 6 - Invalid ORIGIN Attribute.
- 8 - Invalid NEXT_HOP Attribute.
- 9 - Optional Attribute Error.
- 10 - Invalid Network Field.
- 11 – Malformed AS_PATH

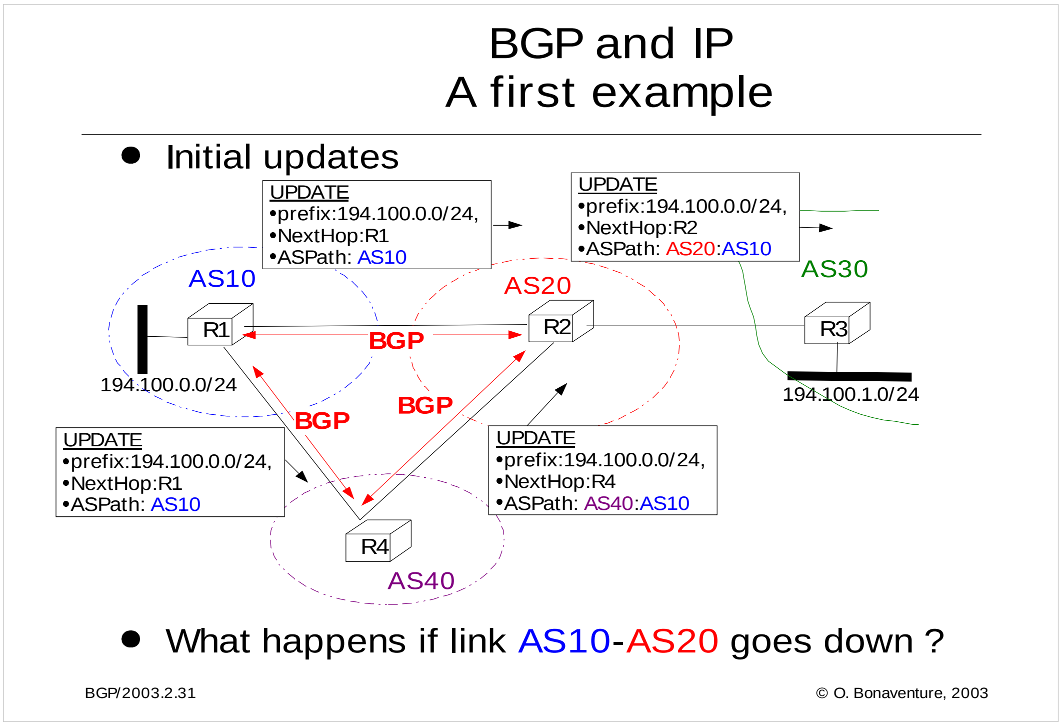
• 4 Hold Timer Expired

• 5 Finite State Machine Error

• 6 Cease

Besides the NOTIFICATION messages, there have been recent proposals within IETF to use a new BGP message to indicate not too severe errors without releasing the BGP session :

G. Nalawade, J. Scudder, D. Ward, BGPv4 INFORM Message, Internet draft, draft-nalawade-bgp-inform-01.txt, Work in progress, Dec. 2002



BGP and IP
A first example

- Initial updates

UPDATE
•prefix:194.100.0.0/24,
•NextHop:R1
•ASPath: AS10

UPDATE
•prefix:194.100.0.0/24,
•NextHop:R2
•ASPath: AS20:AS10

UPDATE
•prefix:194.100.0.0/24,
•NextHop:R1
•ASPath: AS10

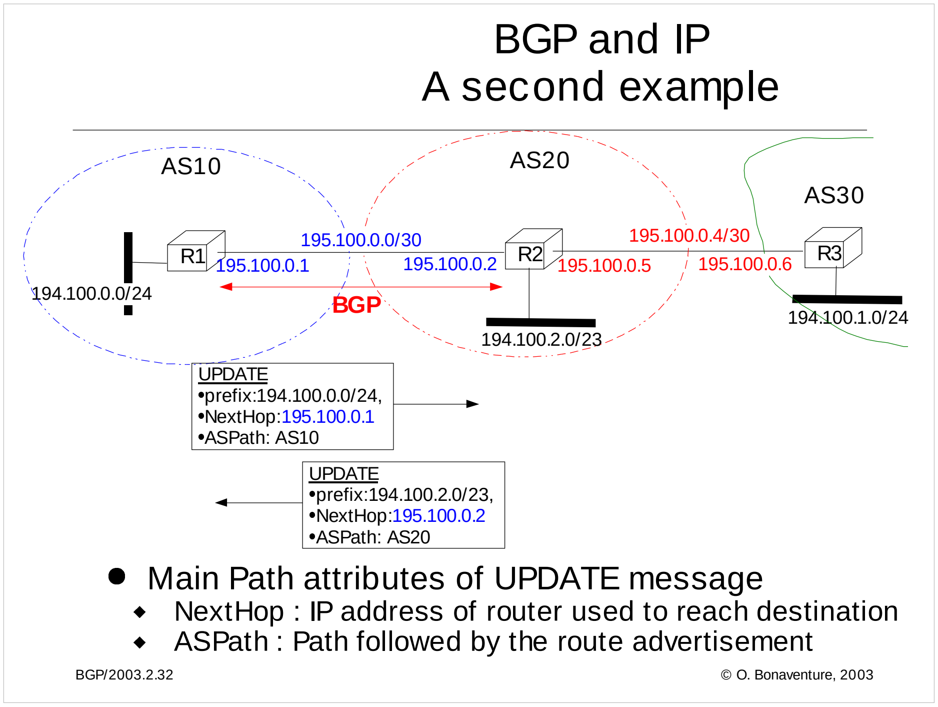
UPDATE
•prefix:194.100.0.0/24,
•NextHop:R4
•ASPath: AS40:AS10

- What happens if link AS10-AS20 goes down ?

BGP/2003.2.31

If link AS10-AS20 goes down, AS20 will not consider anymore the path learned from AS10. It will thus remove this path from its routing table and will instead select the path learned from AS40. This will force AS20 to send the following UPDATE to AS30 :

UPDATE
•prefix:194.100.0.0/24,
•NextHop:R2
•ASPath: AS20:AS40: AS10



BGP and IP
A second example

UPDATE
- prefix:194.100.0.0/24,
- NextHop:195.100.0.1
- ASPath: AS10

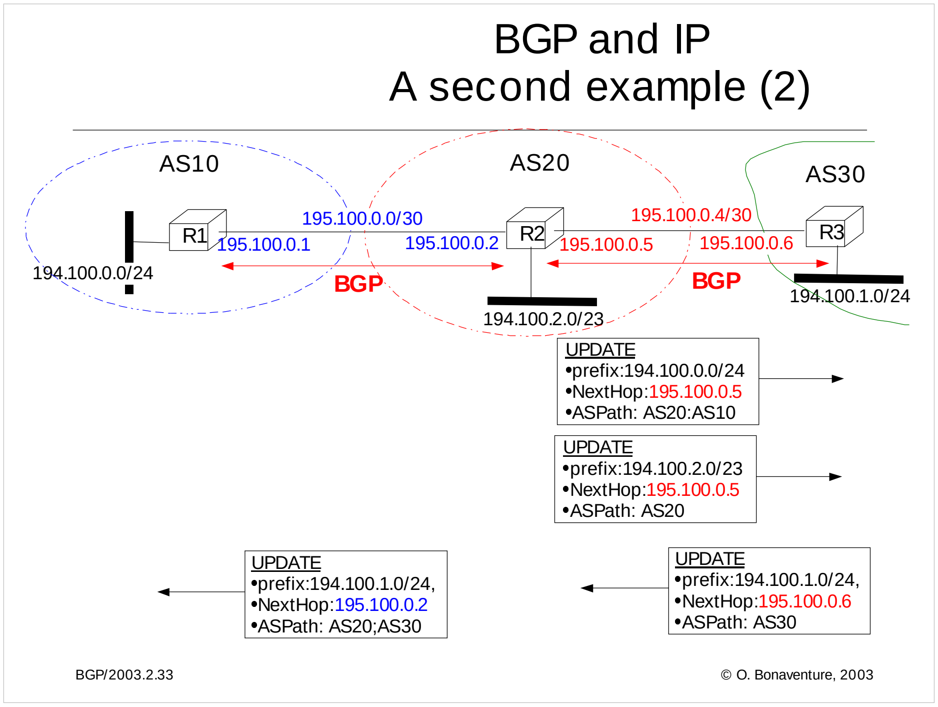
UPDATE
- prefix:194.100.2.0/23,
- NextHop:195.100.0.2
- ASPath: AS20

- Main Path attributes of UPDATE message
  - NextHop : IP address of router used to reach destination
  - ASPath : Path followed by the route advertisement

In this example, we only consider the BGP messages concerning the following IP networks :194.100.0.0/24, 194.100.1.0.0/24 and 194.100.2.0/23. Routes concerning networks 195.100.* also need to be distributed in practice, but they are not considered in the example.

The UPDATE message carries the ASPath in order to be able to detect routing loops.

The nexthop information in the UPDATE is often equal to the IP address of the router advertising the route, but it can be sometimes useful to advertise as a next hop another IP address than the address of the router producing the BGP UPDATE message. For example, a router supporting BGP could advertise a route on behalf of another router who cannot run the BGP protocol.



BGP and IP
A second example (2)

UPDATE
- prefix:194.100.0.0/24
- NextHop:195.100.0.5
- ASPath: AS20:AS10

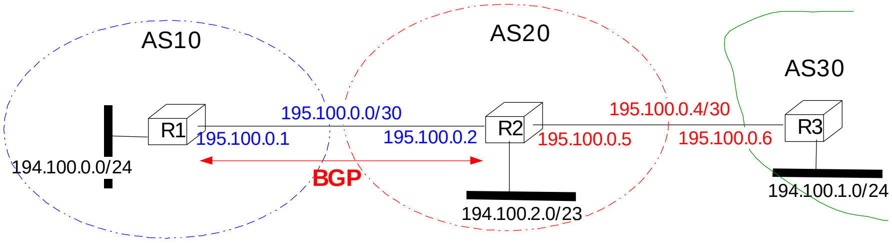
UPDATE
- prefix:194.100.2.0/23
- NextHop:195.100.0.5
- ASPath: AS20

UPDATE
- prefix:194.100.1.0/24,
- NextHop:195.100.0.2
- ASPath: AS20;AS30

UPDATE
- prefix:194.100.1.0/24,
- NextHop:195.100.0.6
- ASPath: AS30

In this example, we only consider the BGP messages concerning the following IP networks :194.100.0.0/24, 194.100.1.0.0/24 and 194.100.2.0/23. Routes concerning networks 195.100.* also need to be distributed, but they are not considered in the example.

BGP and IP
A second example (3)



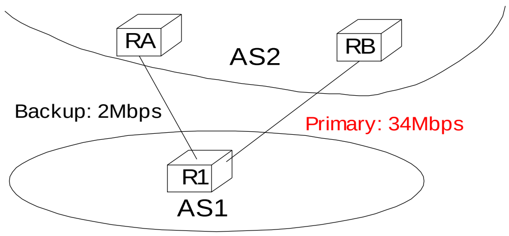
<u>WITHDRAW</u>
•prefix:194.100.1.0/24

In this example, we only consider the BGP messages concerning the following IP networks :194.100.0.0/24, 194.100.1.0.0/24 and 194.100.2.0/23. Routes concerning networks 195.100.* also need to be distributed, but they are not considered in the example.

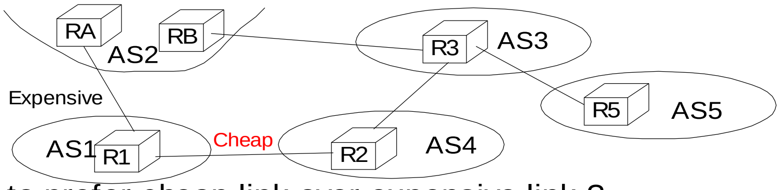
Outline

- Organization of the global Internet
- BGP basics
  - Routing policies
  - The Border Gateway Protocol
  - How to prefer some routes over others
- BGP in large networks
- Interdomain traffic engineering with BGP
- BGP-based Virtual Private Networks

How to prefer some routes over others ?

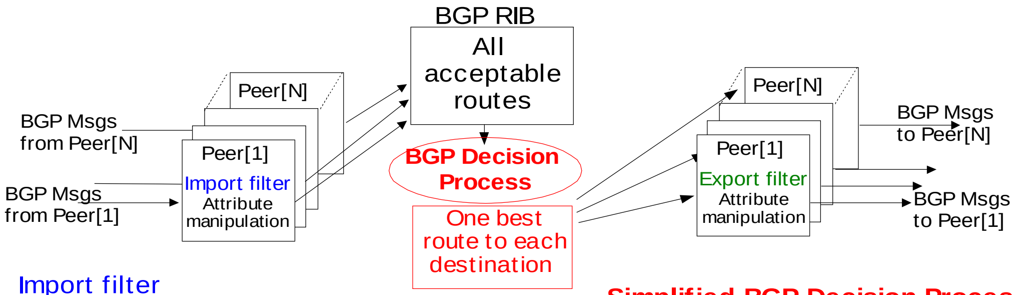


- How to ensure that packets will flow on primary link ?



- How to prefer cheap link over expensive link ?

How to prefer some routes over others (2) ?



BGP RIB

All acceptable routes

BGP Decision Process

One best route to each destination

BGP Msgs from Peer[N]

BGP Msgs from Peer[1]

Peer[N]

Peer[1]

Import filter

Attribute manipulation

Peer[N]

Peer[1]

Export filter

Attribute manipulation

BGP Msgs to Peer[N]

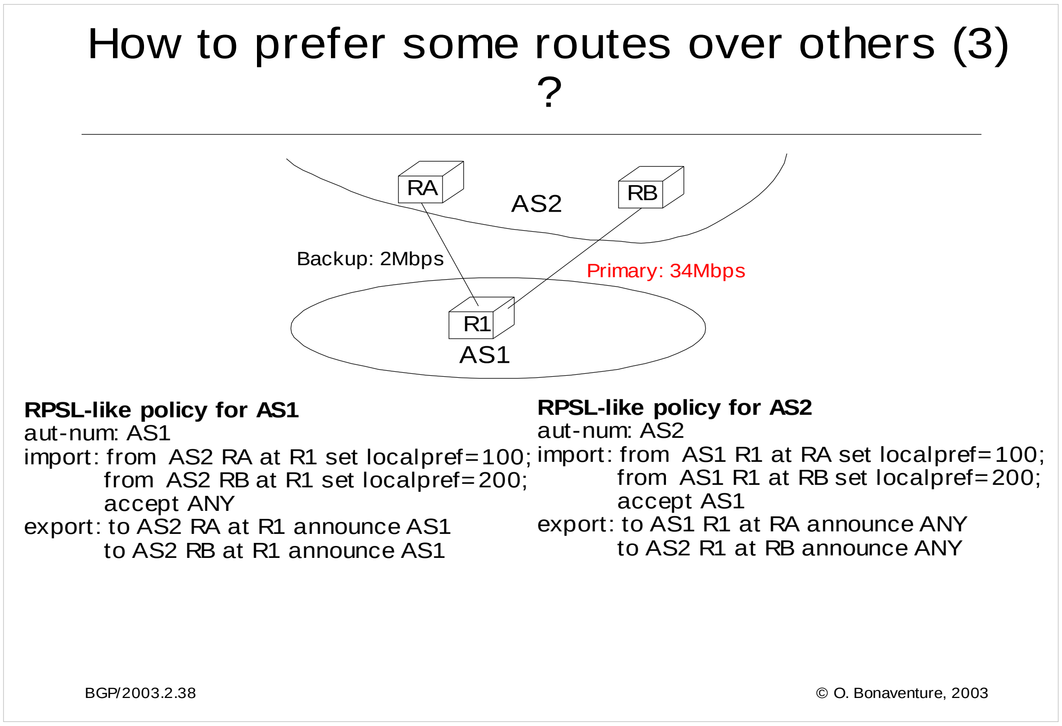
BGP Msgs to Peer[1]

Import filter

- Selection of acceptable routes
- Addition of `local-pref` attribute inside received BGP Msg
  - Normal quality route : `local-pref=100`
  - Better than normal route :`local-pref=200`
  - Worse than normal route :`local-pref=50`

Simplified BGP Decision Process

- Select routes with highest `local-pref`
- If there are several routes, choose routes with the shortest ASPath
- If there are still several routes tie-breaking rule



How to prefer some routes over others (3) ?

RA
RB
AS2
Backup: 2Mbps
Primary: 34Mbps
R1
AS1

RPSL-like policy for AS1

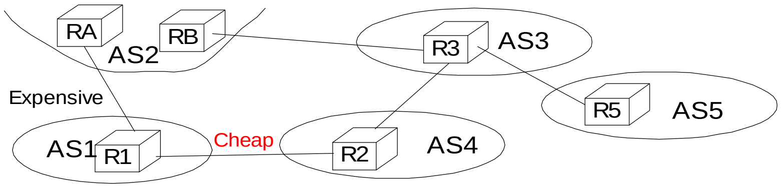
```
aut-num: AS1
import: from  AS2 RA at R1 set localpref=100;
        from  AS2 RB at R1 set localpref=200;
        accept ANY
export: to AS2 RA at R1 announce AS1
        to AS2 RB at R1 announce AS1
```

RPSL-like policy for AS2

```
aut-num: AS2
import: from  AS1 R1 at RA set localpref=100;
        from  AS1 R1 at RB set localpref=200;
        accept AS1
export: to AS1 R1 at RA announce ANY
        to AS2 R1 at RB announce ANY
```

Note that in RPSL, the set localpref construct does not exist. It is replaced with action preference=x. Unfortunately, in RPSL the routes with the lowest preference are preferred. RPSL uses thus the opposite of local-pref....

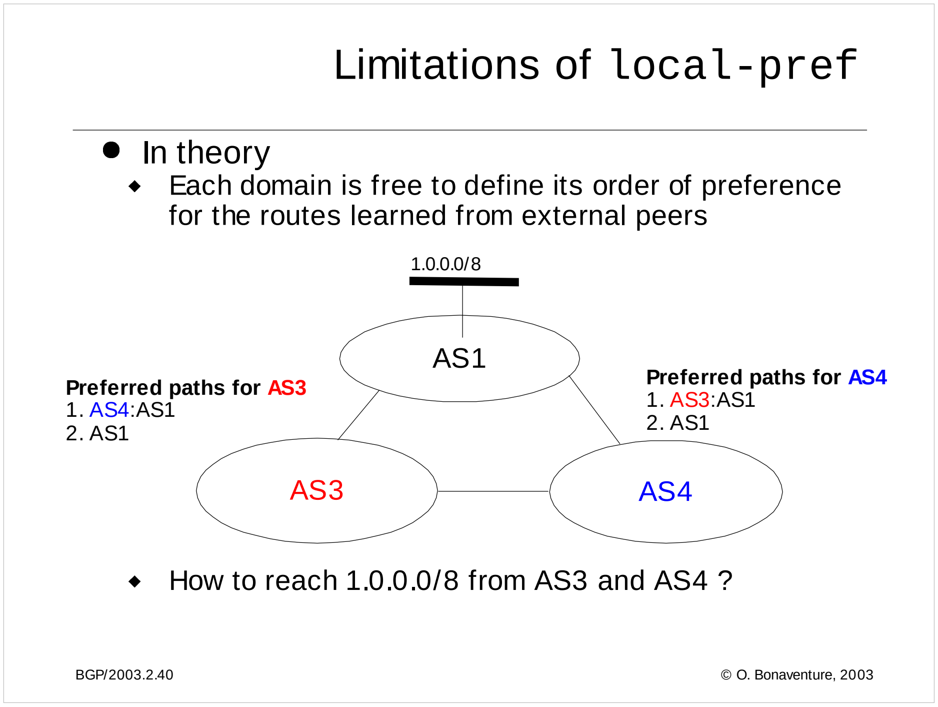
How to prefer some routes over others (4) ?



RPSL policy for AS1

```
aut-num: AS1
import: from  AS2 RA at R1 set localpref=100;
          from  AS4 R2 at R1 set localpref=200;
          accept ANY
export: to AS2 RA at R1 announce AS1
          to AS4 R2 at R1 announce AS1
```

- AS1 will prefer to send packets over the cheap link
- But the flow of the packets destined to AS1 will depend on the routing policy of the other domains



Limitations of `local-pref`

- In theory
  - Each domain is free to define its order of preference for the routes learned from external peers

1.0.0.0/8
AS1
AS3
AS4

Preferred paths for AS3
1. AS4:AS1
2. AS1

Preferred paths for AS4
1. AS3:AS1
2. AS1

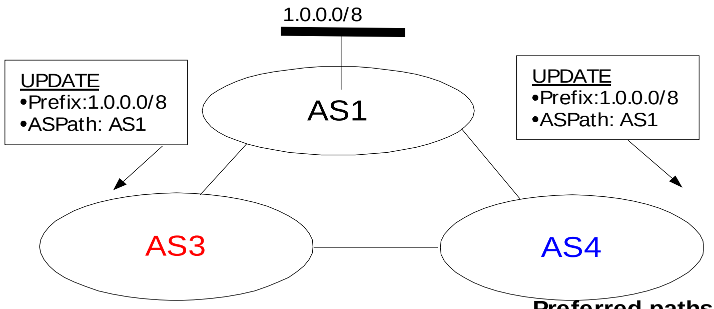
  - How to reach 1.0.0.0/8 from AS3 and AS4 ?

Import policy for AS3
Import: from AS1 accept ANY; set localpref=10
import: from AS4 accept ANY; set localpref=200

Import policy for AS4
Import: from AS1 accept ANY; set localpref=10
import: from AS3 accept ANY; set localpref=200

Limitations of `local-pref` (2)

- AS1 sends its UPDATE messages ...



1.0.0.0/8

UPDATE
•Prefix:1.0.0.0/8
•ASPath: AS1

AS1

UPDATE
•Prefix:1.0.0.0/8
•ASPath: AS1

AS3

AS4

Preferred paths for AS3
1. AS4:AS1
2. AS1

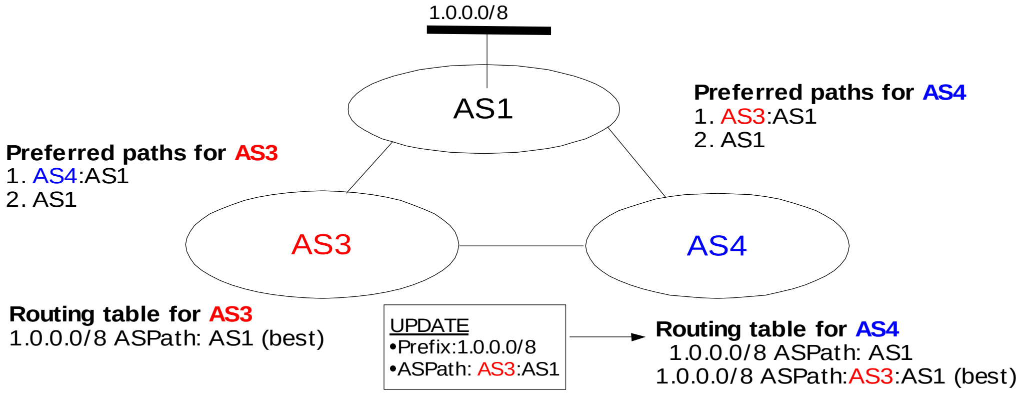
Preferred paths for AS4
1. AS3:AS1
2. AS1

Routing table for AS3
1.0.0.0/8 ASPath: AS1 (best)

Routing table for AS4
1.0.0.0/8 ASPath: AS1 (best)

Limitations of `local-pref` (3)

- First possibility
  - AS3 sends its UPDATE first...



Preferred paths for AS3
1. AS4:AS1
2. AS1

Preferred paths for AS4
1. AS3:AS1
2. AS1

Routing table for AS3
1.0.0.0/8 ASPath: AS1 (best)

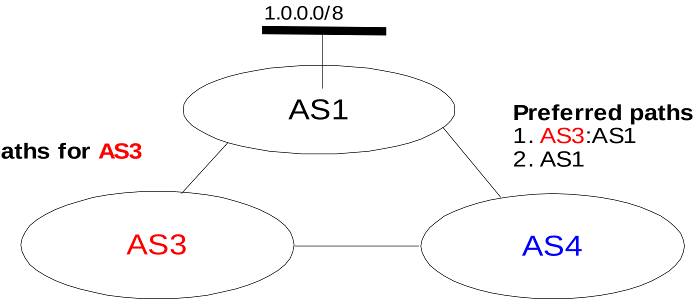
UPDATE
•Prefix:1.0.0.0/8
•ASPath: AS3:AS1

Routing table for AS4
1.0.0.0/8 ASPath: AS1
1.0.0.0/8 ASPath:AS3:AS1 (best)

- Stable route assignment

Limitations of `local-pref` (4)

- Second possibility
  - AS4 sends its UPDATE first...



Preferred paths for AS3
1. AS4:AS1
2. AS1

Preferred paths for AS4
1. AS3:AS1
2. AS1

Routing table for AS3
1.0.0.0/8 ASPath: AS1
1.0.0.0/8 ASPath: AS4:AS1 (best)

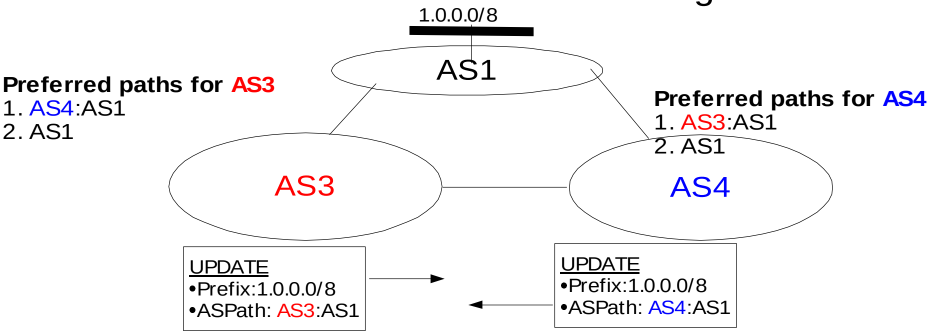
UPDATE
•Prefix:1.0.0.0/8
•ASPath: AS4:AS1

Routing table for AS4
1.0.0.0/8 ASPath: AS1 (best)

- ◆ Another (but different) stable route assignment

Limitations of `local-pref` (5)

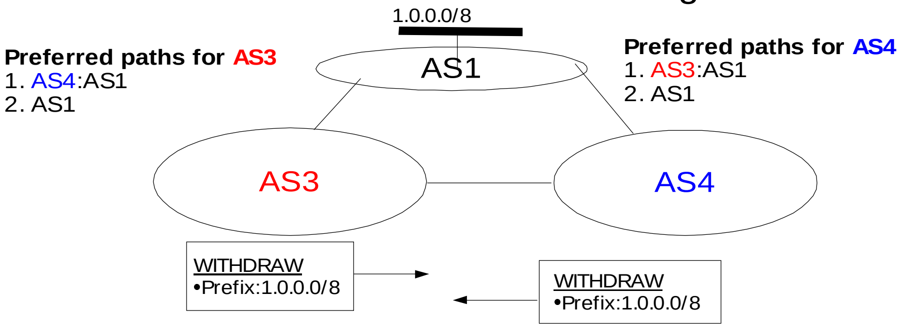
- Third possibility
  - AS3 and AS4 send their UPDATE together...



- ◆ AS3 prefers the indirect path and will thus send withdraw since the chosen best path is via AS4
- ◆ AS4 prefers the indirect path and will thus send withdraw since the chosen best path is via AS3

Limitations of `local-pref` (6)

- Third possibility (cont.)
  - AS3 and AS4 send their UPDATE together…



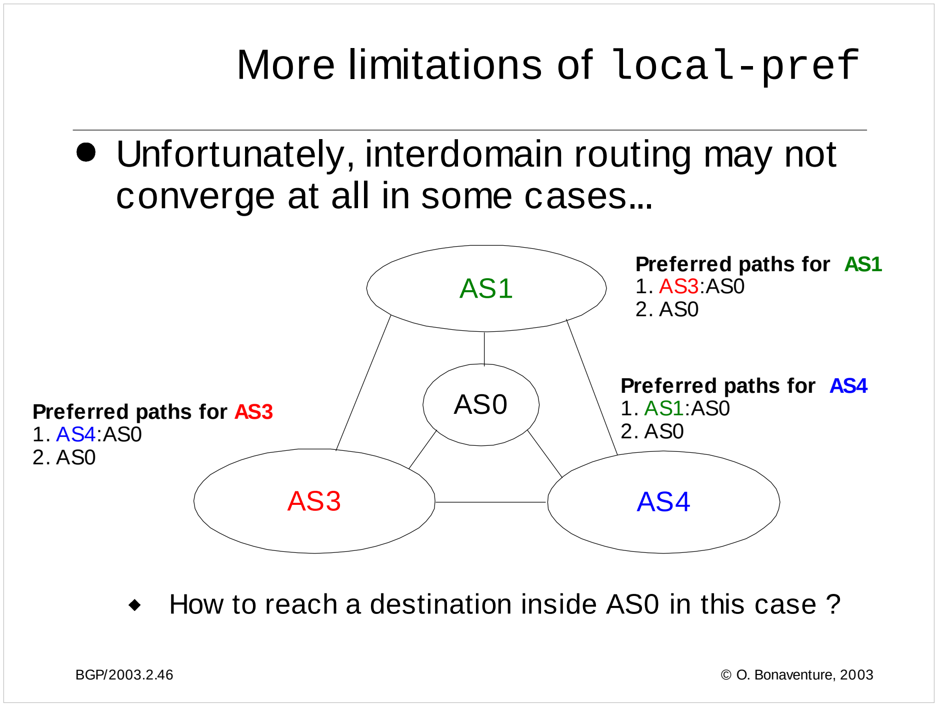
Preferred paths for AS3
1. AS4:AS1
2. AS1

Preferred paths for AS4
1. AS3:AS1
2. AS1

<u>WITHDRAW</u>
•Prefix:1.0.0.0/8

<u>WITHDRAW</u>
•Prefix:1.0.0.0/8

- AS3 learns that the indirect route is not available anymore
  - AS3 will reannounce its direct route…
- AS4 learns that the indirect route is not available anymore
  - AS4 will reannounce its direct route…



More limitations of `local-pref`

- Unfortunately, interdomain routing may not converge at all in some cases...

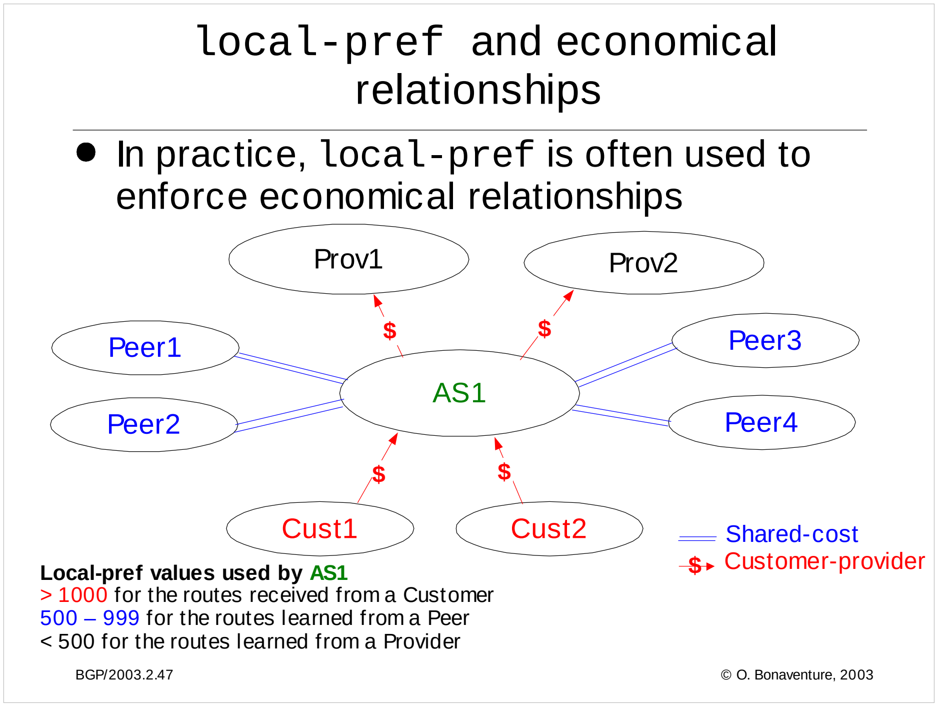
Preferred paths for AS1
1. AS3:AS0
2. AS0

Preferred paths for AS4
1. AS1:AS0
2. AS0

Preferred paths for AS3
1. AS4:AS0
2. AS0

- How to reach a destination inside AS0 in this case ?

In practice, the exchange of BGP UPDATE messages will cease due to the utilization of timers by BGP routers and the routing will stabilize on one of the two stable route assignments.



`local-pref` and economical relationships

- In practice, `local-pref` is often used to enforce economical relationships

Local-pref values used by AS1
> 1000 for the routes received from a Customer
500 – 999 for the routes learned from a Peer
< 500 for the routes learned from a Provider

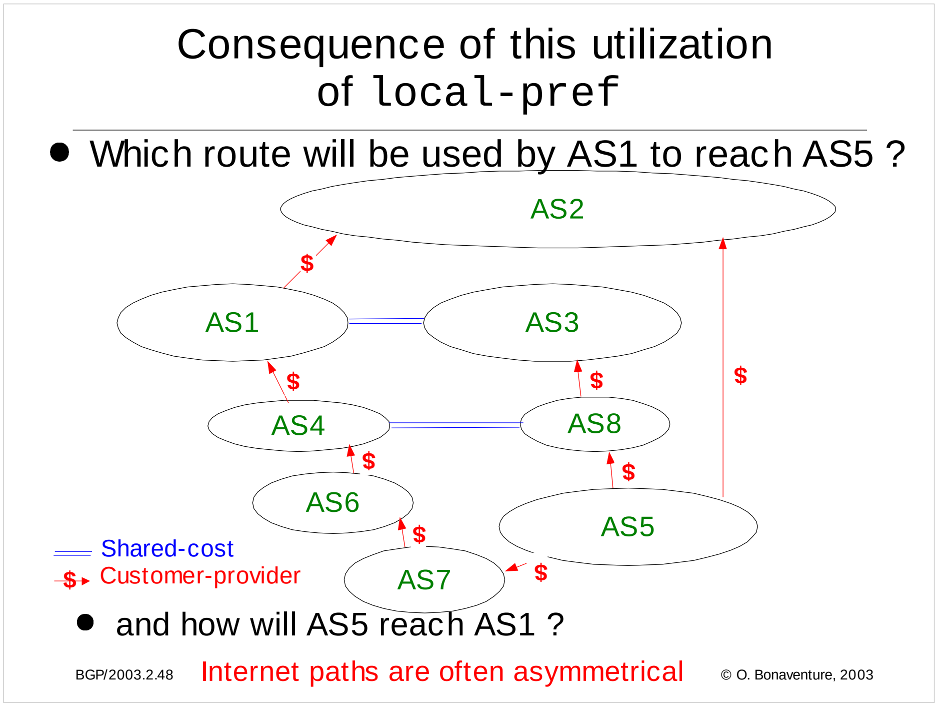
This local-pref settings corresponds to the economical relationships between the various ASes.
Since AS1 is paid to carry packets towards Cust1 and Cust2, it will select a route towards those networks whenever possible.
Since AS1 does not need to pay to carry packets towards Peer1-4, AS1 will select a route towards those networks whenever possible.
AS1 will only utilize the routes receive from its providers when there is no other choice.

It is shown in the following papers that this way of utilizing the local-pref attribute leads to stable BGP routes :
Lixin Gao, Timothy G. Griffin, and Jennifer Rexford, "Inherently safe backup routing with BGP," Proc. IEEE INFOCOM, April 2001
Lixin Gao and Jennifer Rexford, "Stable Internet routing without global coordination," IEEE/ACM Transactions on Networking, December 2001, pp. 681-692

The RPSL policy of AS1 could be as follows :

RPSL policy for AS1

```
aut-num: AS1
import:    from  Cust1 action set localpref=200; accept Cust1
           from  Cust2 action set localpref=200; accept Cust2
           from  Peer1 action set localpref=150; accept Peer1
           from  Peer2 action set localpref=160; accept Peer2
           from  Peer3 action set localpref=170; accept Peer3
                      from  Peer4 action set localpref=180; accept Peer4
           from  Prov1 action set localpref=100; accept ANY
           from  Prov2 action set localpref=100; accept ANY
```



Consequence of this utilization of `local-pref`

- Which route will be used by AS1 to reach AS5 ?

- and how will AS5 reach AS1 ?

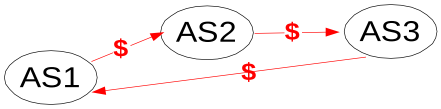
Internet paths are often asymmetrical

Due to the utilization of the local-pref attribute, some paths on the Internet are longer than their optimum length, see :

Lixin Gao and Feng Wang , The Extent of AS Path Inflation by Routing Policies, GlobalInternet 2002

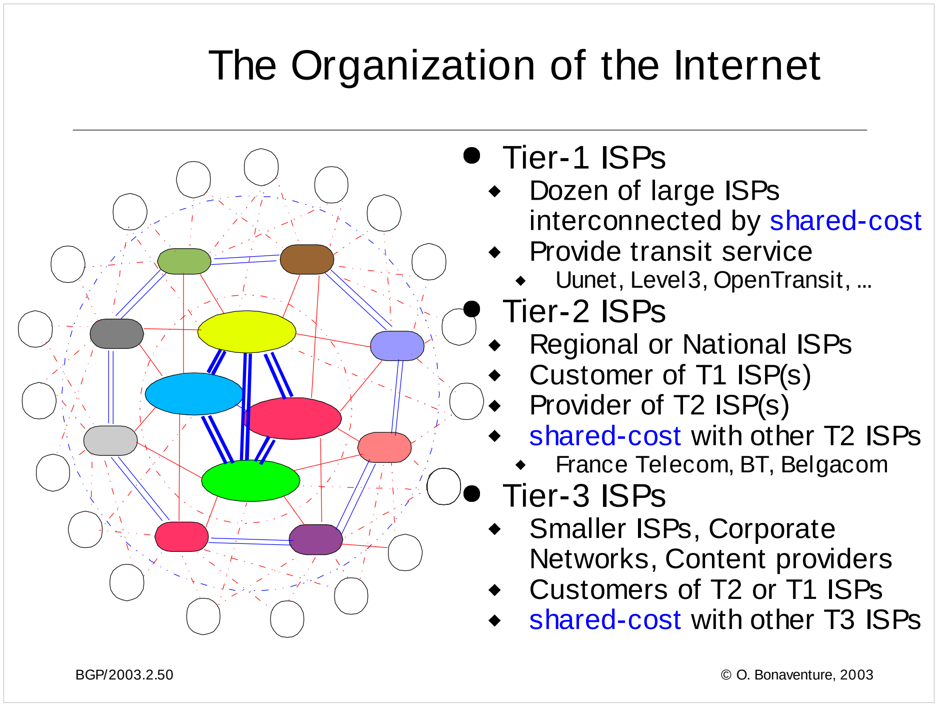
Guidelines for a safe utilization of `local-pref`

- The directed graph composed of the customer->provider links is loop-free
  - An AS cannot be a customer of a provider of its providers



  - An AS always prefer a route via a customer over a route via a provider or a peer
    - With some restrictions on the graph composed of peer-to-peer relationships, it is also possible to allow an AS to give the same preference to a route via a customer or via a peer

Lixin Gao and Jennifer Rexford, "Stable Internet routing without global coordination," IEEE/ACM Transactions on Networking, December 2001, pp. 681-692

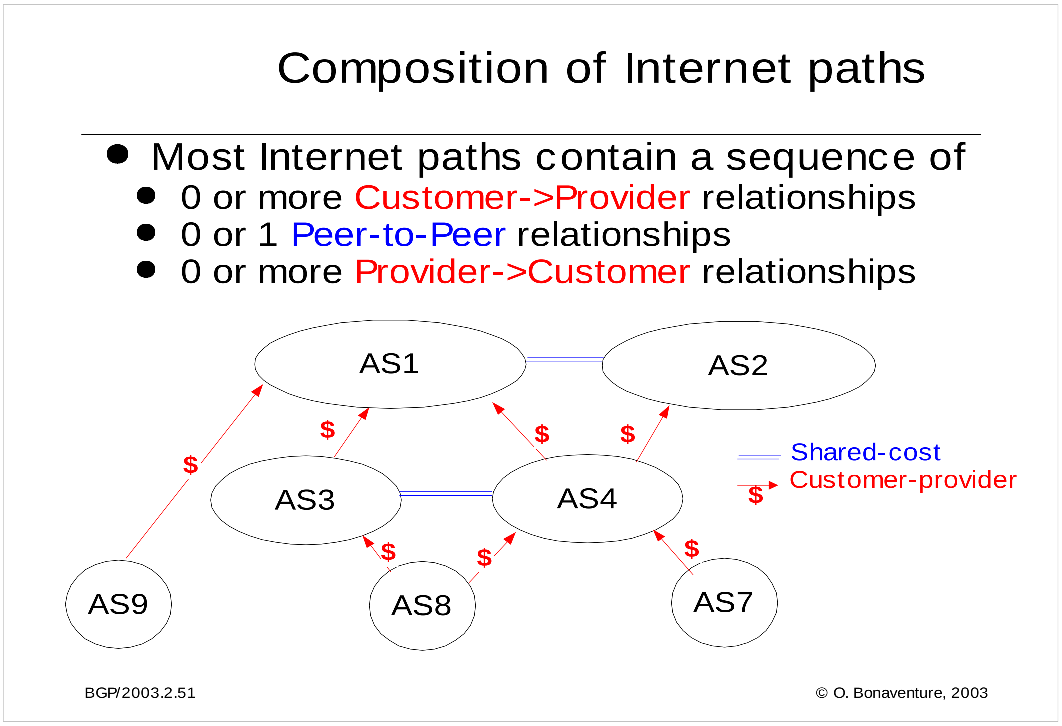


The Organization of the Internet

- Tier-1 ISPs
  - Dozen of large ISPs interconnected by shared-cost
  - Provide transit service
    - Uunet, Level3, OpenTransit, ...
- Tier-2 ISPs
  - Regional or National ISPs
  - Customer of T1 ISP(s)
  - Provider of T2 ISP(s)
  - shared-cost with other T2 ISPs
    - France Telecom, BT, Belgacom
- Tier-3 ISPs
  - Smaller ISPs, Corporate Networks, Content providers
  - Customers of T2 or T1 ISPs
  - shared-cost with other T3 ISPs

See :

L. Subramanian, S. Agarwal, J. Rexford, and RH Katz. Characterizing the Internet hierarchy from multiple vantage points. In IEEE INFOCOM, 2002



Composition of Internet paths

- Most Internet paths contain a sequence of
  - 0 or more Customer->Provider relationships
  - 0 or 1 Peer-to-Peer relationships
  - 0 or more Provider->Customer relationships

For a discussion of this and its implication on the organization of the global Internet, see e.g. :

Lakshminarayanan Subramanian, Sharad Agarwal, Jennifer Rexford, and Randy H. Katz, "Characterizing the Internet hierarchy from multiple vantage points," in Proc. IEEE INFOCOM, June 2002

Summary

- Routing policies
  - Two main routing policies
    - Customer-Provider relationship
    - Peer-to-Peer relationship
- The Border Gateway Protocol
  - Path vector protocol with incremental updates
  - Import and export filters to implement routing policies
  - Utilization of local-pref
    - Influence BGP decision process
    - Prefer some routes over others
    - Be careful with possible oscillations due to bad setting