Computer Networks II

Vecteur de distance à comptage de chemin sans cycles.

=> Peut osciller (jamais trouver de solution)

Ch.2 Interdomain Routing and BGP

(bruno.quoitin@umons.ac.be)

Important note: These slides are partly based on a course by Olivier Bonaventure (UCLouvain).

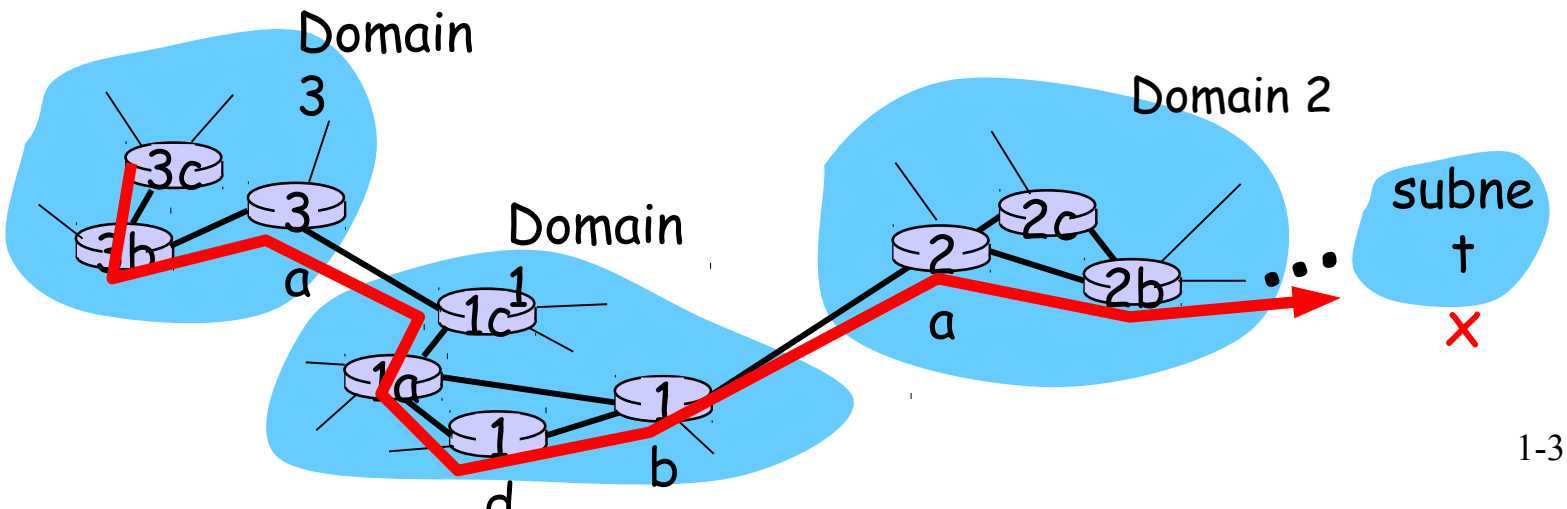
Chapter 2: roadmap

- 2.1 Inter-domain Routing
  - 2.1.1 Fundamental Objectives
  - 2.1.2 Definitions
  - 2.1.3 Routing Policies
- 2.2 The Border Gateway Protocol (BGP)
- 2.3 BGP-based Traffic Engineering
- 2.3 BGP Scalability
- 2.5 BGP Stability

Diffère des méthodes de Réseaux 1
(qui étaient intra domaine avec coûts fixés par opérateur réseau)

Inter-domain Routing

- **Fundamental Objectives**
  - Allows to transmit IP packets along the best path towards their destination through several transit domains Parfois, réseaux payants.
    - the notion of best-path often means "cheapest"
    - local objective functions (the best path of one domain might not be the best path of another domain)



Inter-domain Routing

- **Fundamental Objectives**
  - Take into account Routing Policies of each domain
  - Allow for domain Autonomy
    - Operation of the internals of each domain are performed independently
    - Each domain can run a different IGP, with its own IGP weights assignment
    - Each domain is free to specify its routing policy (domains for which it agrees to carry transit traffic, method used to select the best path, ...)

Plusieurs techniques :

1) coût unitaires: nbre de sauts. 2) Inverse de la bande pass. des liens

3) Meilleurs délais.

Inter-domain Routing

- **Fundamental Objectives**
  - Scalability
    - Hide the detailed topology of the transit domains
    - The inter-domain graph has millions of nodes and edges -> clearly not scalable
    - Rely on hierarchical routing : separate intra-domain routing from inter-domain routing

Les noeuds du graphe ne sont plus des routeurs mais des réseaux de routeurs

Inter-domain Routing

- **Definitions**
  - Autonomous System (AS)
    - Network under the administration of a single entity.
    - Each AS is identified with a unique AS number (ASN).
    - Examples : ISP network, university campus network, enterprise network, ...
  - Prefix
    - Set of IP addresses specified using the CIDR notation
    - Examples : 193.190.192/22 (UMONS), 138.48/16 (FUNDP), 130.104/16 (UCLouvain), ...

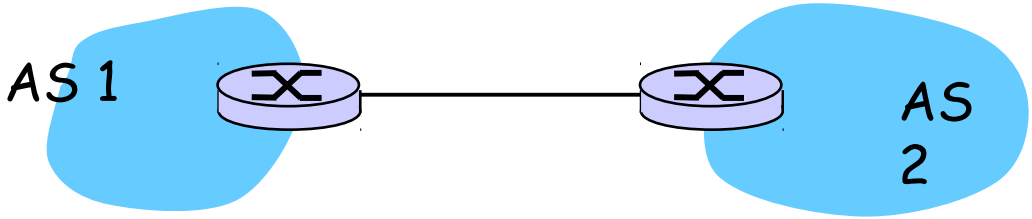
Whois this AS ? Whois this Prefix ?

```
bash-3.2$ whois -h whois.arin.net AS668

OrgName:     DoD Network Information Center
OrgID:       DNIC
Address:     3990 E. Broad Street
City:        Columbus
StateProv:   OH
PostalCode:  43218
Country:     US

ASNumber:    668
ASName:      ASN-DREN-NET
ASHandle:    AS668
Comment:
RegDate:     1990-04-24
Updated:     2009-04-10

OrgTechHandle: MIL-HSTMST-ARIN
OrgTechName:   Network DoD
OrgTechPhone:  +1-614-692-2708
OrgTechEmail:  HOSTMASTER@nic.mil

OrgTechHandle: REGIS10-ARIN
OrgTechName:   Registration
OrgTechPhone:  +1-800-365-3642
OrgTechEmail:  REGISTRA@nic.mil
...
bash-3.2$
```

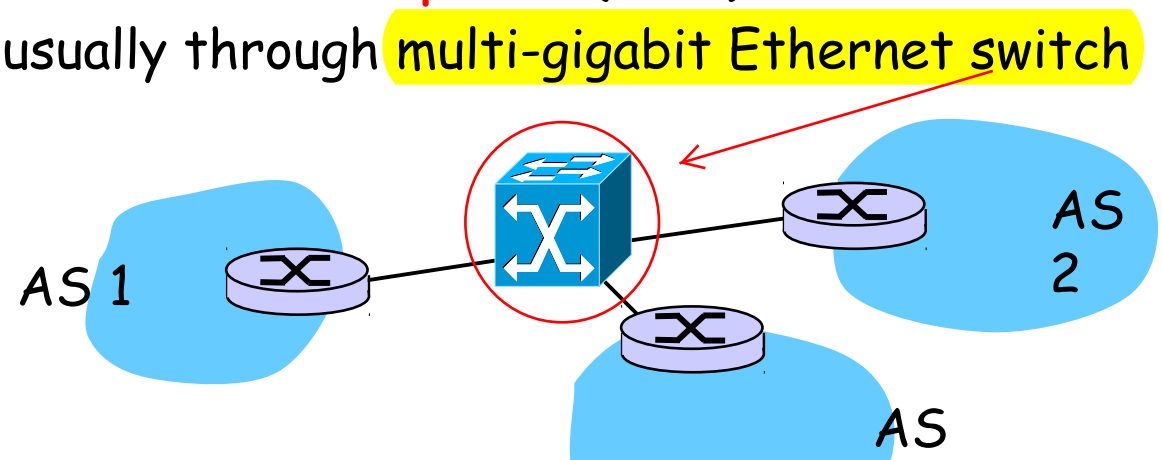
```
bash-3.2$ whois -h whois.arin.net 6.1.0.0

OrgName:     Headquarters, USAISC
OrgID:       HEADQU-3
Address:     NETC-ANC CONUS TNOSC
City:        Fort Huachuca
StateProv:   AZ
PostalCode:  85613-5000
Country:     US

NetRange:    6.0.0.0 - 6.255.255.255
CIDR:        6.0.0.0/8
NetName:     YUMA-NET
NetHandle:   NET-6-0-0-0-1
Parent:
NetType:     Direct Allocation
NameServer:  NS01.ARMY.MIL
NameServer:  NS02.ARMY.MIL
NameServer:  NS03.ARMY.MIL
Comment:
RegDate:     1994-02-01
Updated:     2009-06-19

OrgTechHandle: JIMAD-ARIN
OrgTechName:   DUNSCOMBE, JIM A
...
bash-3.2$
```

Inter-domain Routing

- Private link
  - usually a direct cable or leased line between two routers belonging to the connected domains

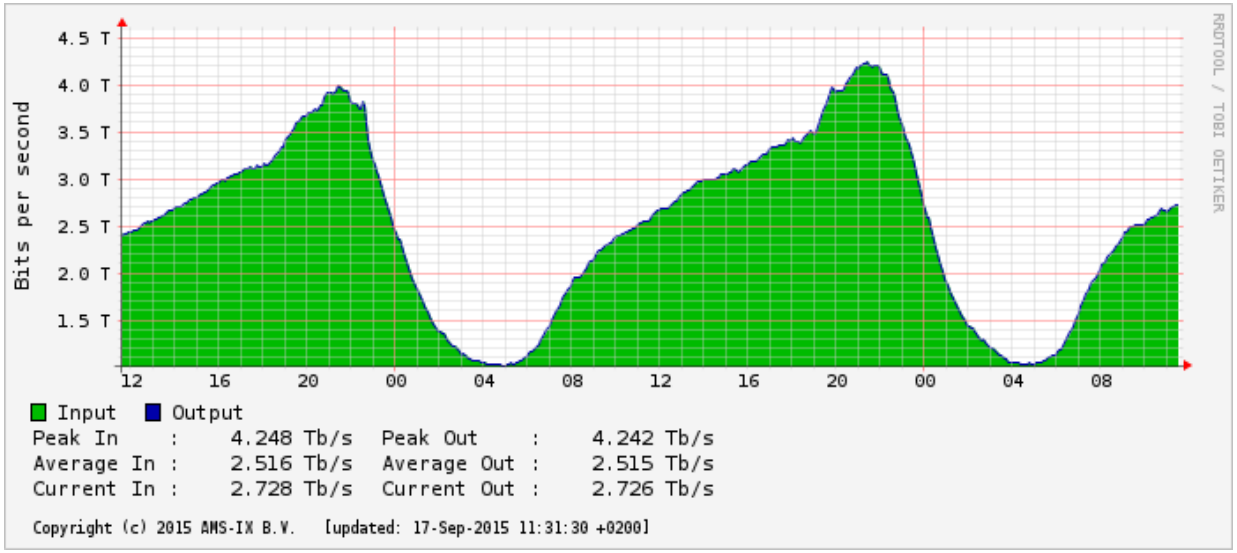


- Interconnection point (IXP)
  - usually through multi-gigabit Ethernet switch

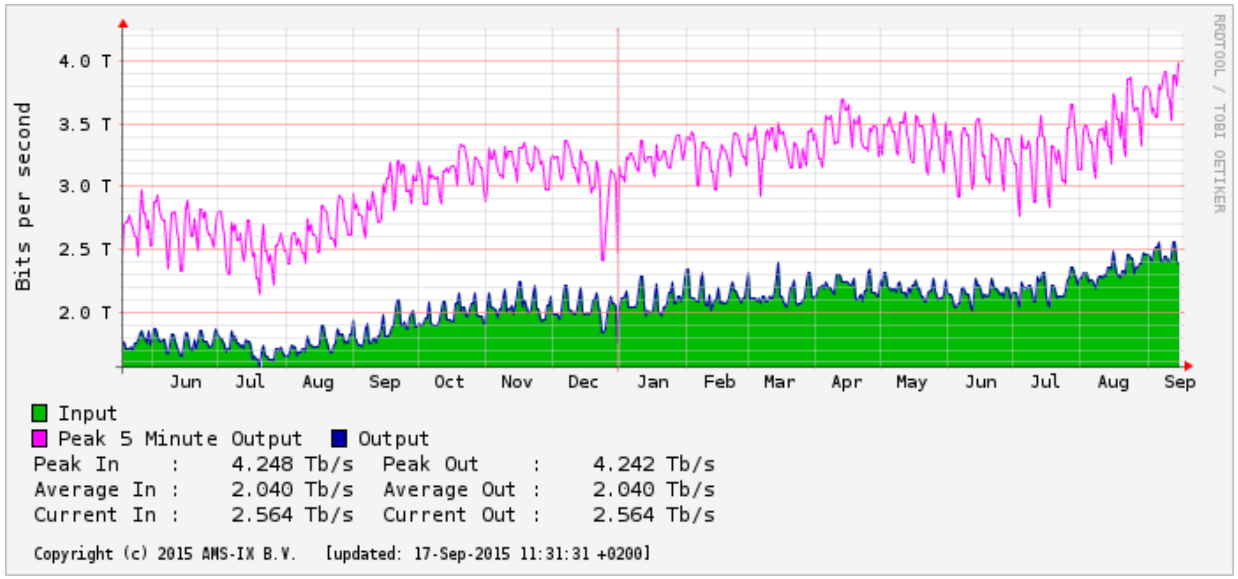


Example public IXP : AMS-IX

Sur deux jours :



Sur l'année :

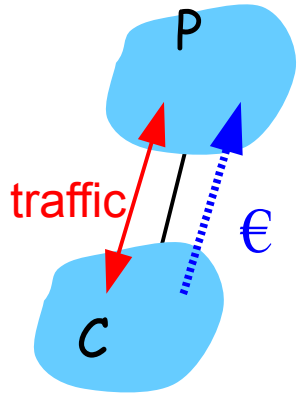


Source: AMS-IX, http://www.ams-ix.net/statistics/, 2014

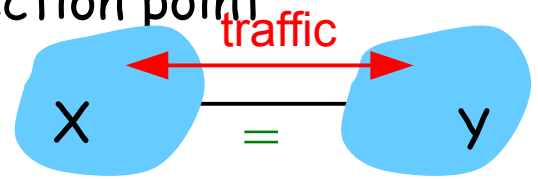
Routing Policies

Vertical fait exprès car relation client/ provider

- There are two main common policies
  - Customer-provider
    - **Customer C** buys Internet connectivity from a **provider P**
    - **P** propagates **C** routes
    - **P** announces to **C** the routes it knows



  - Shared-cost
    - **Peer domains X** and **Y** agree to exchange packets by using a direct link through an interconnection point
    - **X** and **Y** exchange their own routes and the routes of their customers

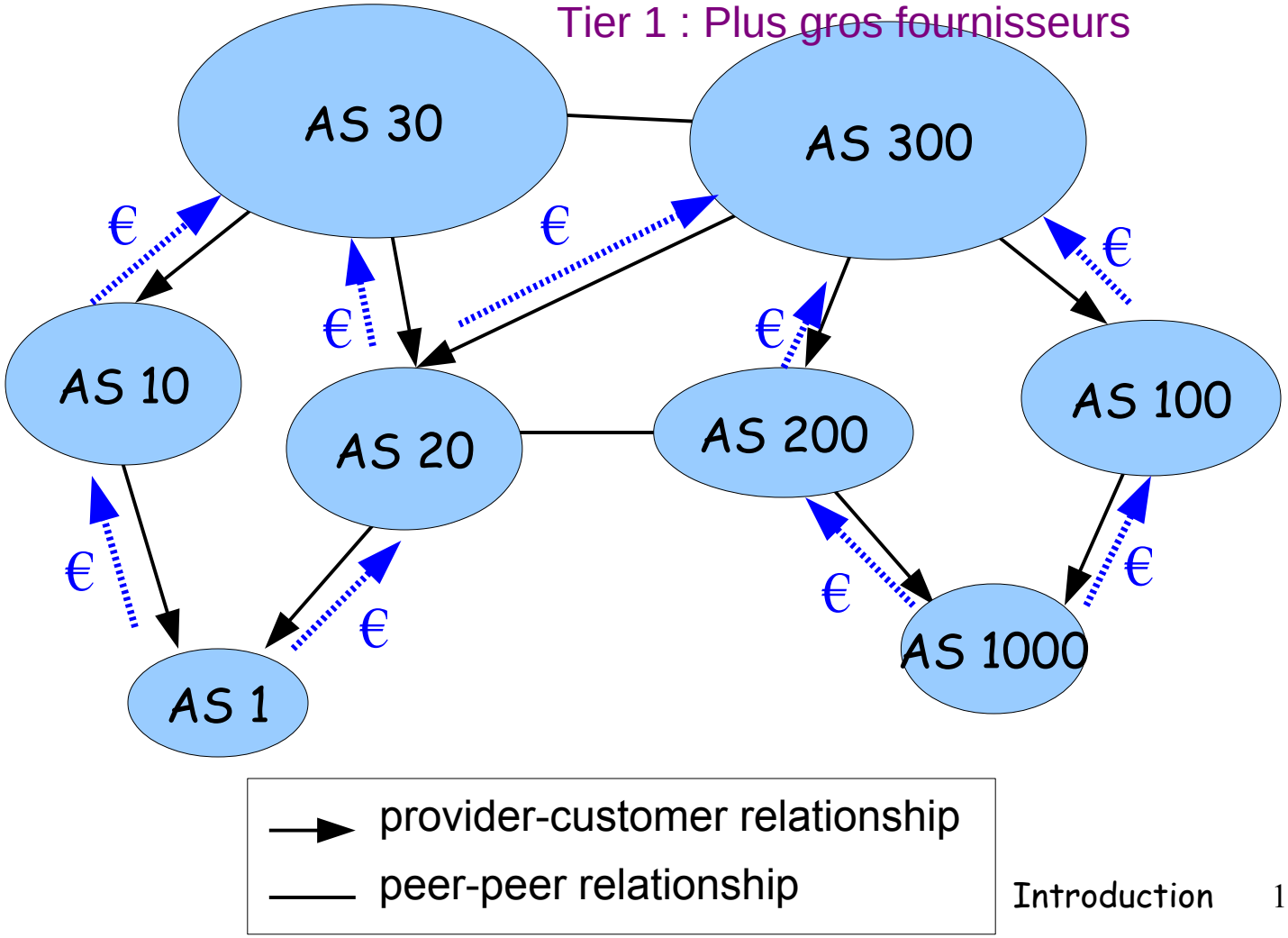


  - Other, more complex, policies are possible (and indeed exist)

Traffic de peer => horizontal

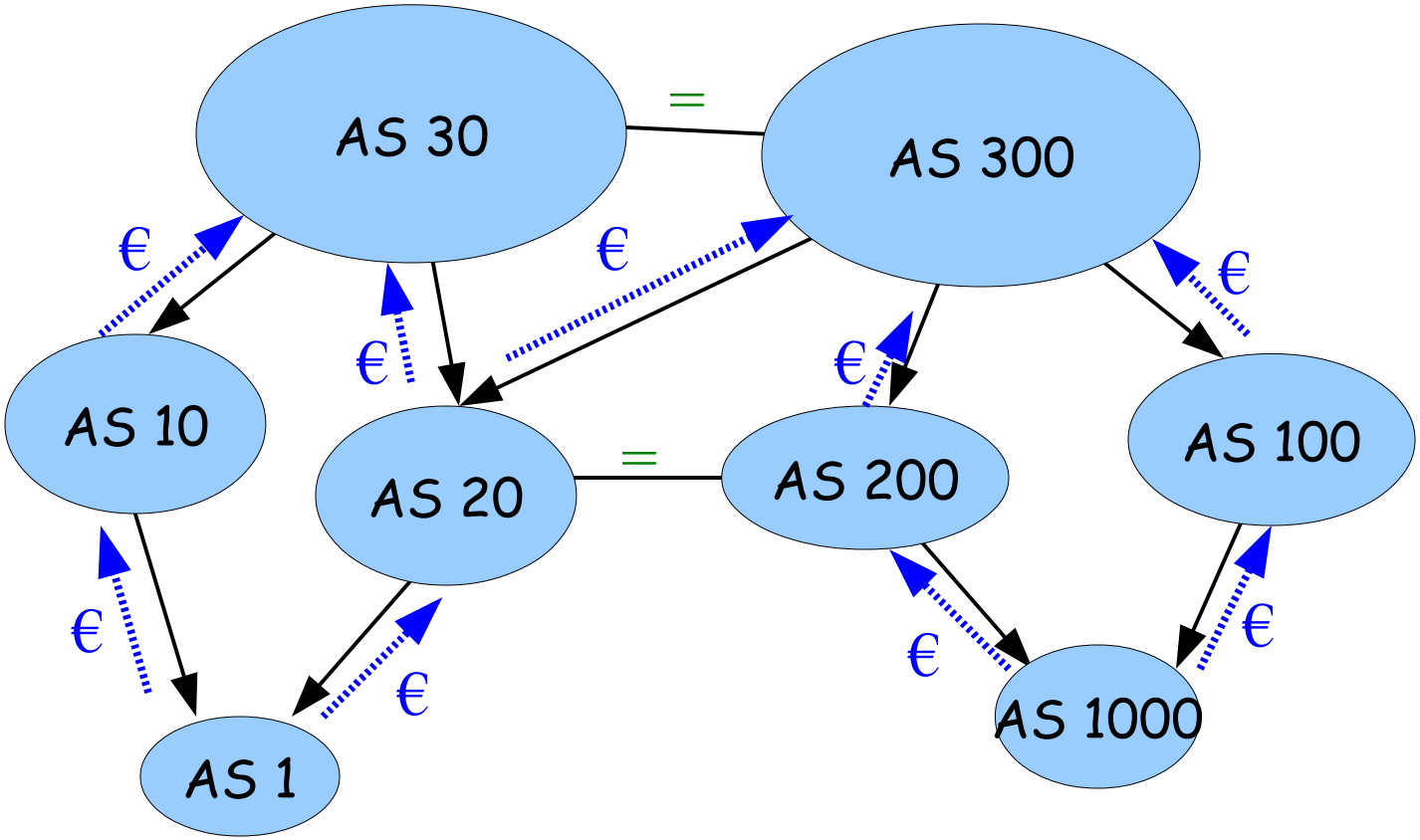
Customer-provider peering

Customer sends its internal routes and the routes of its own customers.
Provider sends to its customers all known routes.



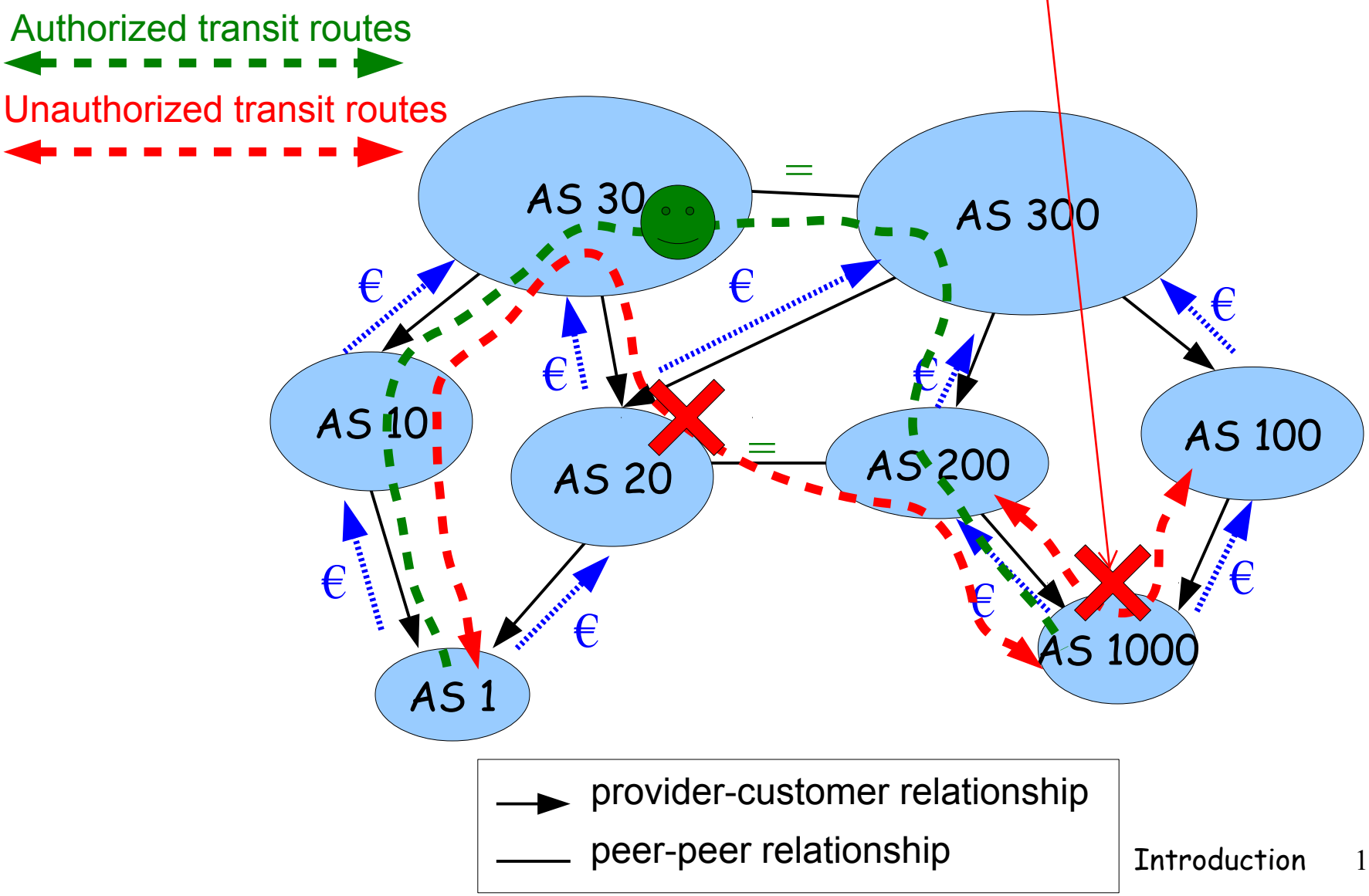
Shared-cost peering

Peers send to each others their internal routes and the routes of their own customers. A peer do not send the routes from its peer to its upstream provider !!!

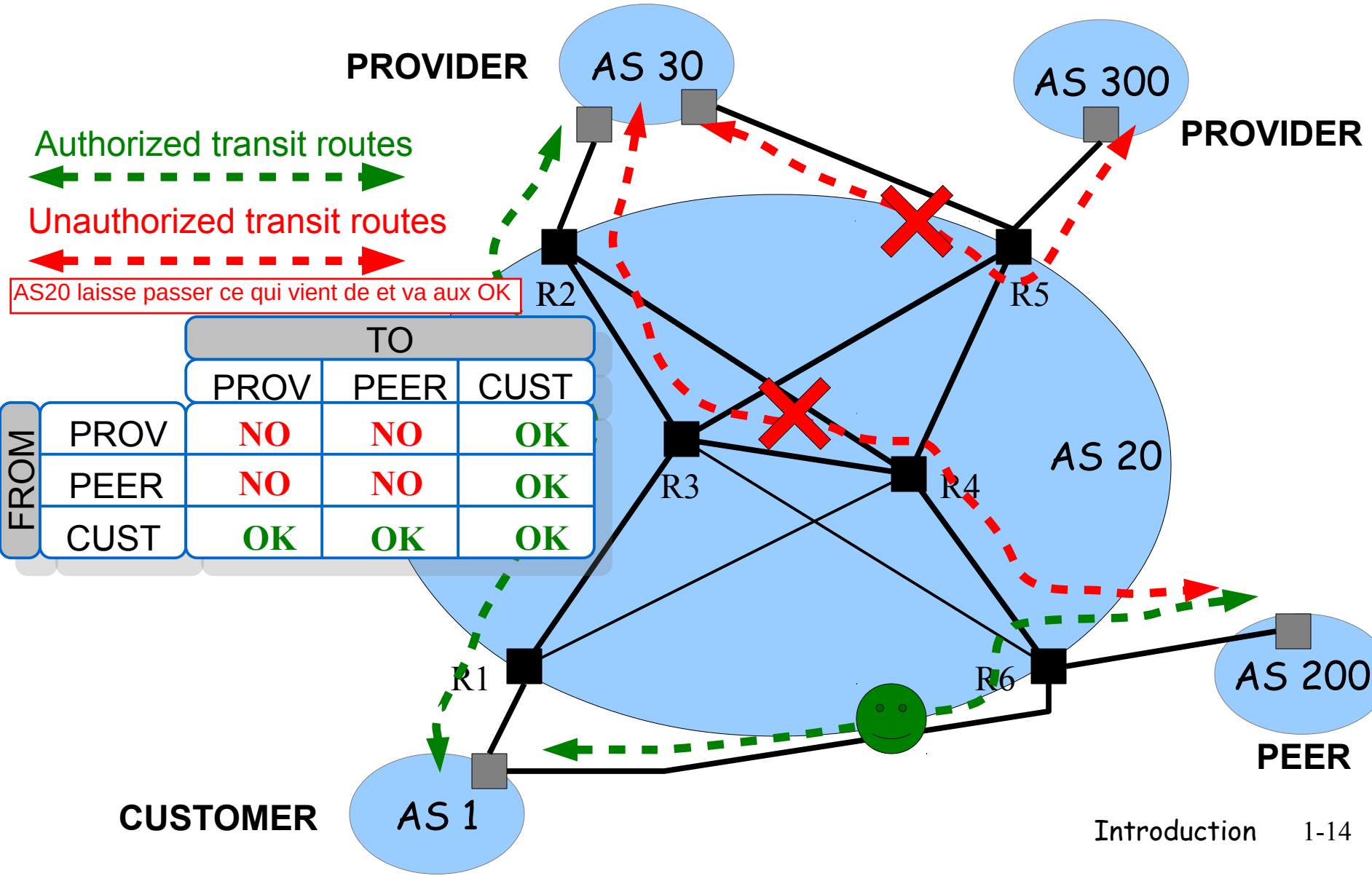


Authorized transit routes

Sinon, AS1000 pigeon car son provider l'utilise comme passage



Transit Policy Matrix



FROM \ TO	PROV	PEER	CUST
PROV	NO	NO	OK
PEER	NO	NO	OK
CUST	OK	OK	OK

Routing Policies

- **Principle**

Proche d'un protocole de routage,

A l'intérieur d'un AS, la meilleure route est un choix local

  - A domain specifies its routing policy by defining on each border router two sets of filters for each peer
    - Import filter : which route can be accepted by the router from a given peer
    - Export filter : which route can be advertised by the router to a given peer

  - Filters are usually expressed in a vendor-specific language
    - There is a standard, the **Routing Policy Specification Language** (RPSL, RFC2622 and RFC2650), but it is seldom used.
    - See also http://irrtoolset.isc.org/

RPSL example

- Simple import policies
  - Syntax

```
import: from AS# accept list_of_AS
```

  - Example

```
import: from BELNET accept UMONS
import: from LEVEL3 accept ANY
```

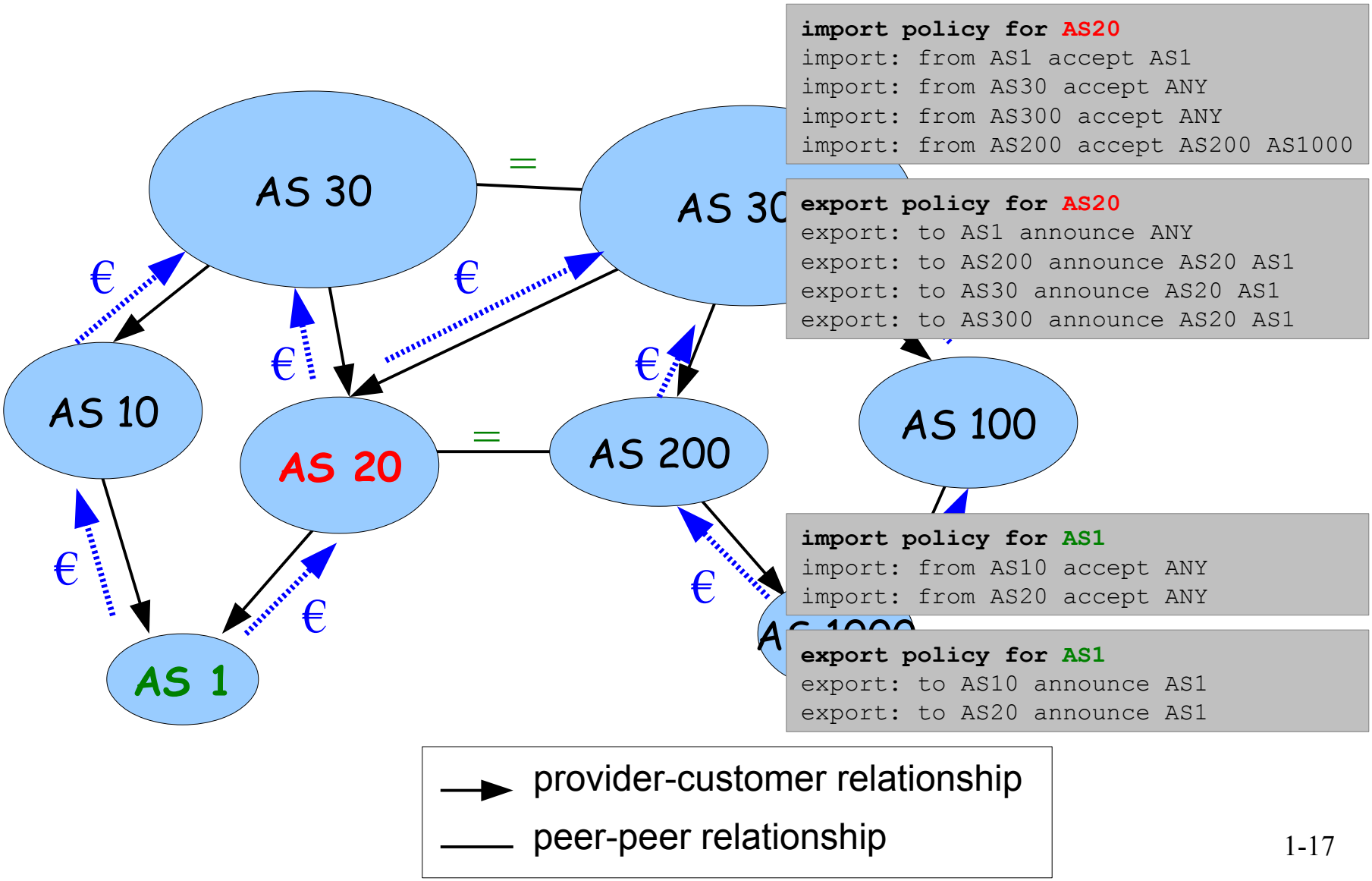
- Simple export policies
  - Syntax

```
export: to AS# announce list_of_AS
```

  - Example

```
export: to UMONS announce ANY
export: to LEVEL3 announce UMONS UCLOUVAIN ..
```

RPSL example



Chapter 2: roadmap

- 2.1 Inter-domain Routing
- 2.2 The Border Gateway Protocol (BGP)
  - 2.2.1 Principles
  - 2.2.2 Sessions
  - 2.2.3 Routes
  - 2.2.4 Path Attributes
  - 2.2.5 Messages
  - 2.2.6 Finite State Machine
  - 2.2.7 Decision Process
  - 2.2.8 Routing Filters
  - 2.2.9 Internal BGP (iBGP)
- ...

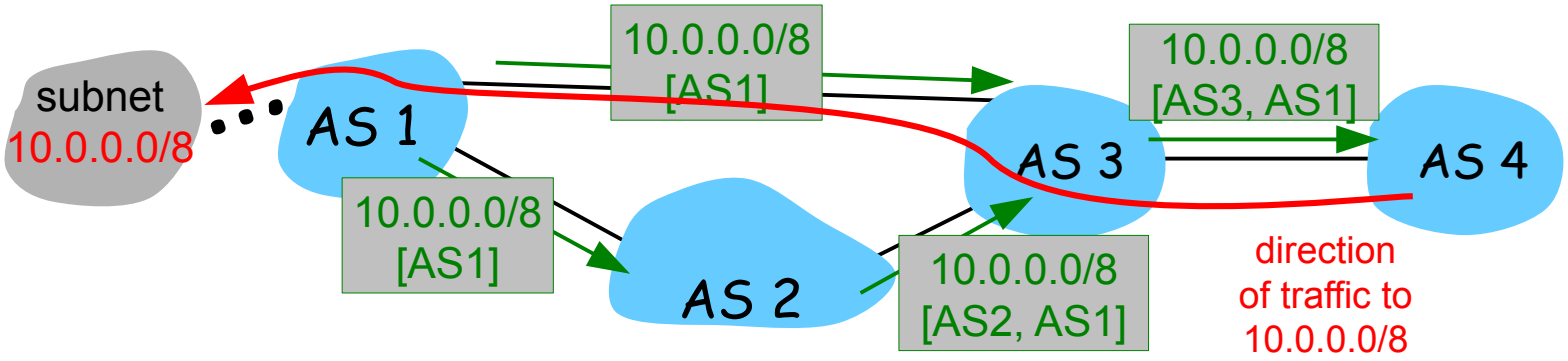
Border Gateway Protocol (BGP)

- **Principles**
  - BGP-4, RFC4271
  - de facto standard inter-domain routing protocol
  - provides each AS means to
    - obtain subnet (prefix) reachability information from neighboring ASs
    - propagate reachability information to all AS-internal routers
    - determine "good" routes to subnets based on reachability information and policy
  - allows subnet (prefix) to advertise its existence to rest of Internet : "I am here" !!!

Border Gateway Protocol (BGP)

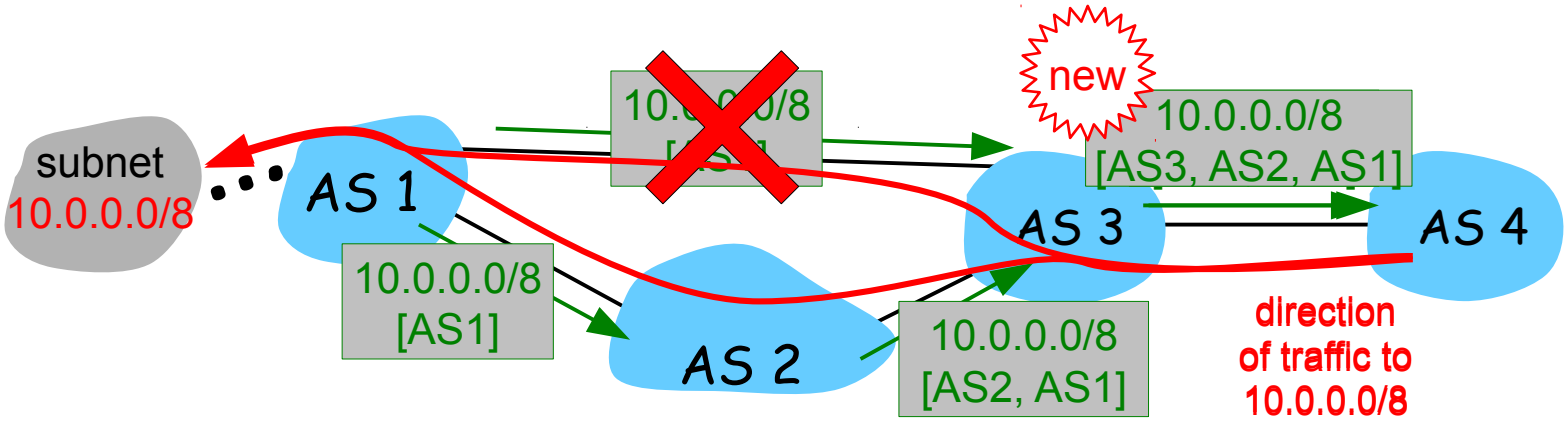
Propagation de l'information entre les AS en donnant linfo et le chemin de l'AS courant en premier, avant tout le reste du chemin (= prepending)

- **Principles**
  - Path Vector Protocol
    - Route = announces destination prefix (e.g. `10.0.0.0/8`)
    - Each router advertises its best route to each destination
    - Routers propagate received routes
    - ASN added in front of route's path (prepending)



Border Gateway Protocol (BGP)

- **Principles**
  - Incremental updates
    - Advertisements are only sent when routing changes
    - TCP is key : can't misunderstand or miss an update



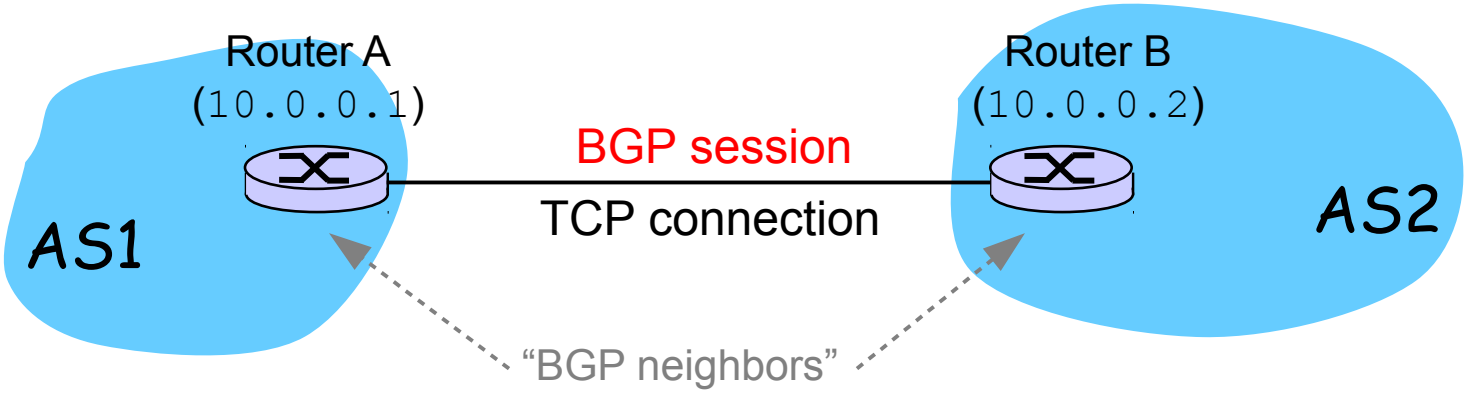
- In example : route from AS1 not available anymore to AS3 -> new route advertised from AS3 to AS4

Chapter 2: roadmap

- 2.1 Inter-domain Routing
- 2.2 The Border Gateway Protocol (BGP)
  - 2.2.1 Principles
  - 2.2.2 Sessions
  - 2.2.3 Routes
  - 2.2.4 Path Attributes
  - 2.2.5 Messages
  - 2.2.6 Finite State Machine
  - 2.2.7 Decision Process
  - 2.2.8 Routing Filters
  - 2.2.9 Internal BGP (iBGP)
- ...

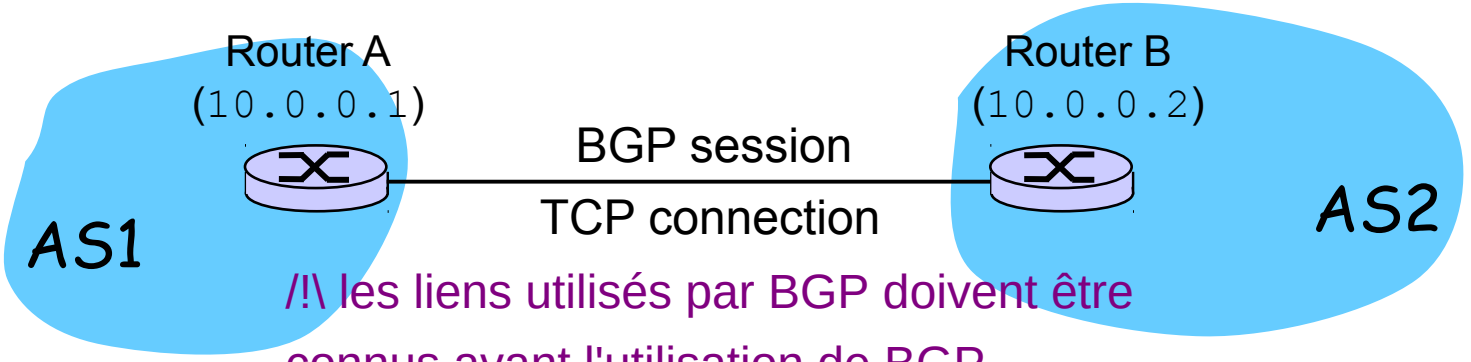
BGP Session

- **Principles**
  - Pairs of BGP routers exchange routes over semi-permanent TCP connections : BGP sessions
  - Routers that have a BGP session are called BGP peers or BGP neighbors
  - The default TCP port for BGP sessions is 179
  - Routers must be able to reach each other



BGP Session

- **Configuration of BGP session**
  - example using Cisco's config. language



/!\ les liens utilisés par BGP doivent être connus avant l'utilisation de BGP

ASN

```
# router bgp 1
# neighbor 10.0.0.2 remote-as 2
```

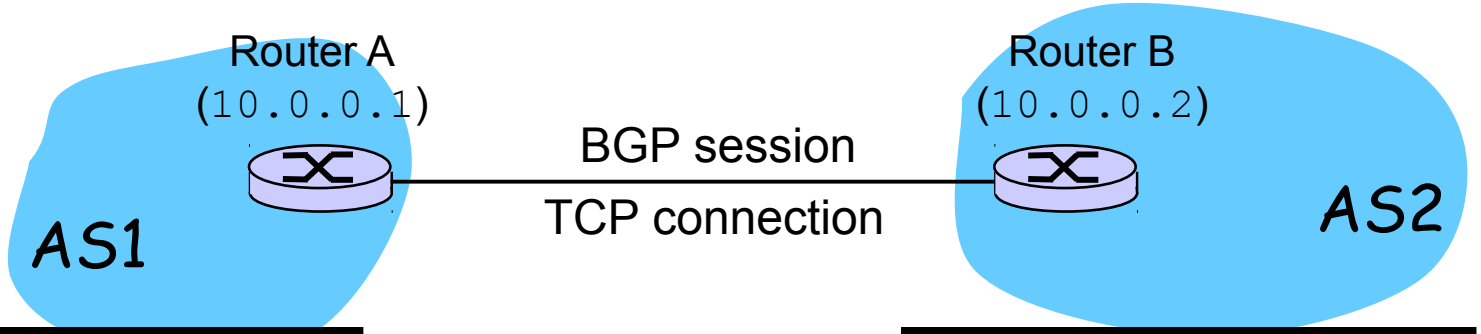
IP address and ASN of neighbor router

```
# router bgp 2
# neighbor 10.0.0.1 remote-as 1
```

BGP Session

Plusieurs config. possible

- **Configuration of BGP session**
  - using the JunOS config. language




```
routing-options {
  autonomous-system 1;
}
protocols {
  bgp {
    group mes_voisins {
      peer-as 2;
      type external;
      neighbor 10.0.0.2;
    }
  }
}
```

```
routing-options {
  autonomous-system 2;
}
protocols {
  bgp {
    group mes_voisins {
      peer-as 1;
      type external;
      neighbor 10.0.0.1;
    }
  }
}
```

Juniper NETWORKS

Chapter 2: roadmap

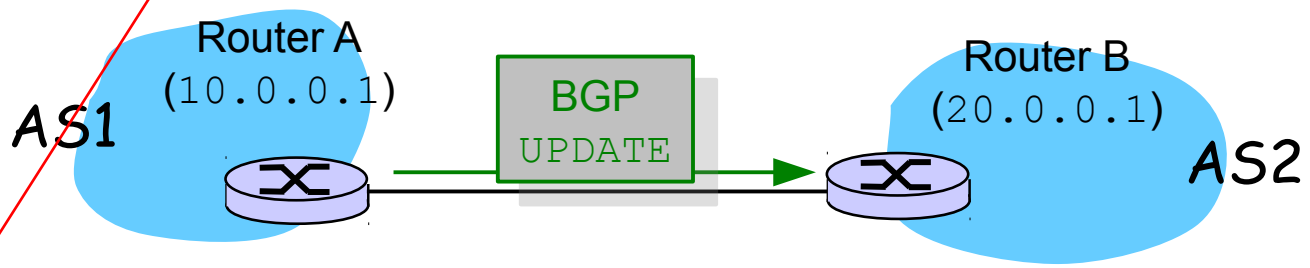
- 2.1 Inter-domain Routing
- 2.2 The Border Gateway Protocol (BGP)
  - 2.2.1 Principles
  - 2.2.2 Sessions
  - 2.2.3 Routes
  - 2.2.4 Path Attributes
  - 2.2.5 Messages
  - 2.2.6 Finite State Machine
  - 2.2.7 Decision Process
  - 2.2.8 Routing Filters
  - 2.2.9 Internal BGP (iBGP)
- ...

BGP Route

Tuyau : Une seule route possible par préfixe de destination. Si plusieurs pref. (ou routes) sont donnés, alors, il y a écrasement des autres info. => jamais 2 routes possibles pour un même préfixe

Definition

- A route is the combination of a destination prefix and path attributes
- A route is carried in an `UPDATE` message.



- Important constraint : A BGP router announces a **single route** per destination prefix

Routes Advertisements

- Where do the advertised routes come from ?
  - Static route
    - configured manually on the router

```
# router bgp 1
# neighbor 10.0.0.1 remote-as 2
# network 150.0.0.0 mask 255.255.255.0
```

  - Route redistribution
    - learned from an intra-domain routing protocol
    - bad practice: intra-domain routing instabilities propagated to BGP + scalability issues !

Problème de stabilité dans le réseau

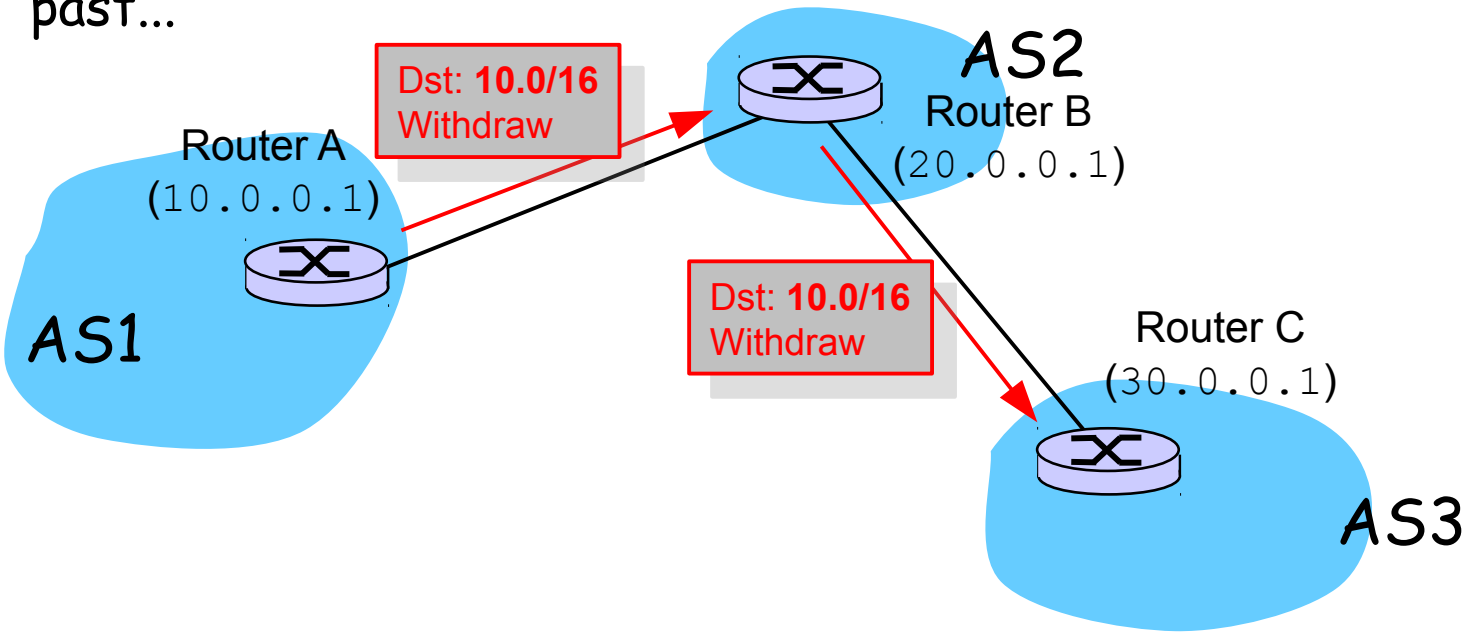
  - BGP, received from another router
    - each BGP router advertises its best routes to its neighbors

Routes Withdraw

- **Principle**
  - When a destination is not reachable anymore, a BGP router must also inform its neighbors
  - `UPDATE` messages are used to tell neighbors that a route is not reachable anymore.
    - This is called a `WITHDRAW` message (even if there is no `WITHDRAW` message in the BGP spec)
  - `WITHDRAW` should only be sent for previously announced routes.

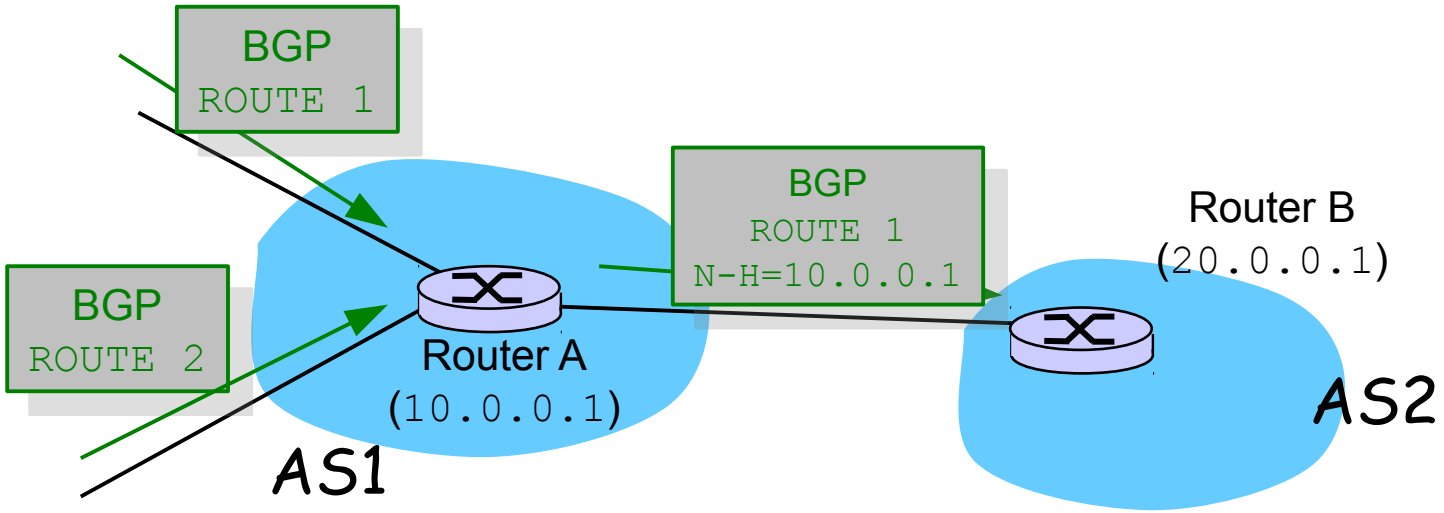
Route Withdraw

- **Principle**
  - When a prefix becomes unreachable, `WITHDRAW` messages are sent.
  - Suppose that a route towards `10.0/16` was announced in the past…



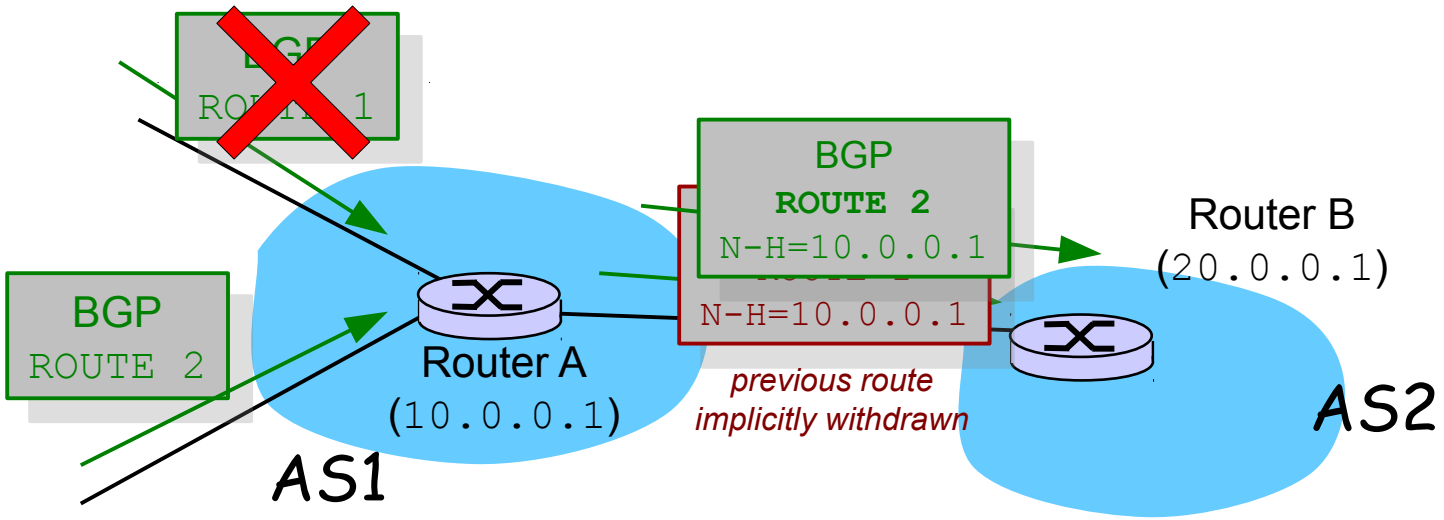
Route Withdraw

- **Principle**
  - A router can send an `UPDATE` message to notify a neighbor that the attributes of a previously announced route have changed.
  - This implicitly `WITHDRAWs` the previous route.



Route Withdraw

- **Principle**
  - A router can send an `UPDATE` message to notify a neighbor that the attributes of a previously announced route have changed.
  - This implicitly `WITHDRAWs` the previous route.



Chapter 2: roadmap

- 2.1 Inter-domain Routing
- 2.2 The Border Gateway Protocol (BGP)
  - 2.2.1 Principles
  - 2.2.2 Sessions
  - 2.2.3 Routes
  - 2.2.4 Path Attributes
  - 2.2.5 Messages
  - 2.2.6 Finite State Machine
  - 2.2.7 Decision Process
  - 2.2.8 Routing Filters
  - 2.2.9 Internal BGP (iBGP)
- ...

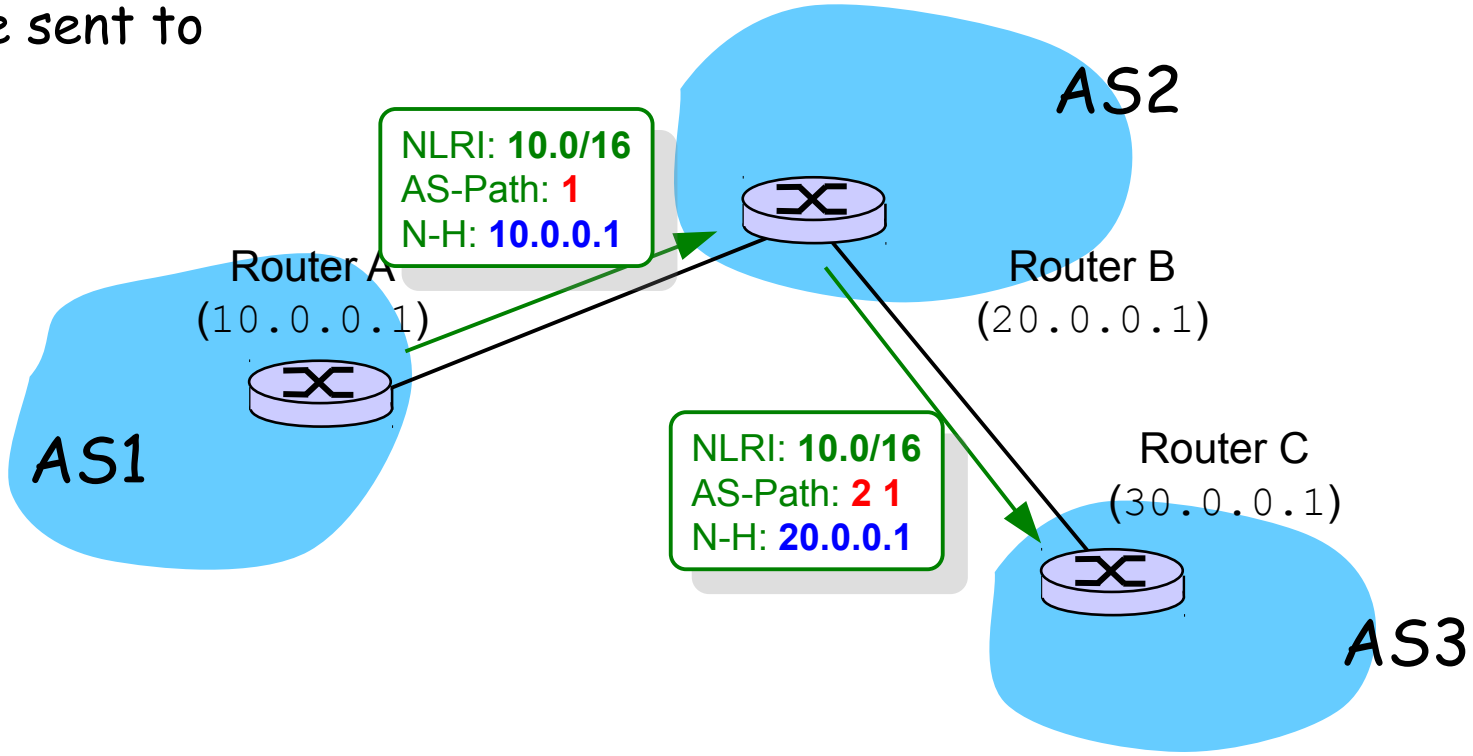
Path Attributes

- **Principles**
  - A route might contain several attributes.
  - The most important path attributes are
    - **NEXT-HOP**: specifies the IP address of the next-hop router on this route(1)
    - **AS-PATH**: contains the list of AS the BGP message has passed through. The list is specified in the form of a list of AS numbers (ASN).
    - **LOCAL-PREF**: integer value associated to the route. It tells how preferred the route should be.

(1) we'll see later that this is not necessarily the immediate next-hop : recursive lookups in the FIB can be used

Most important Path Attributes

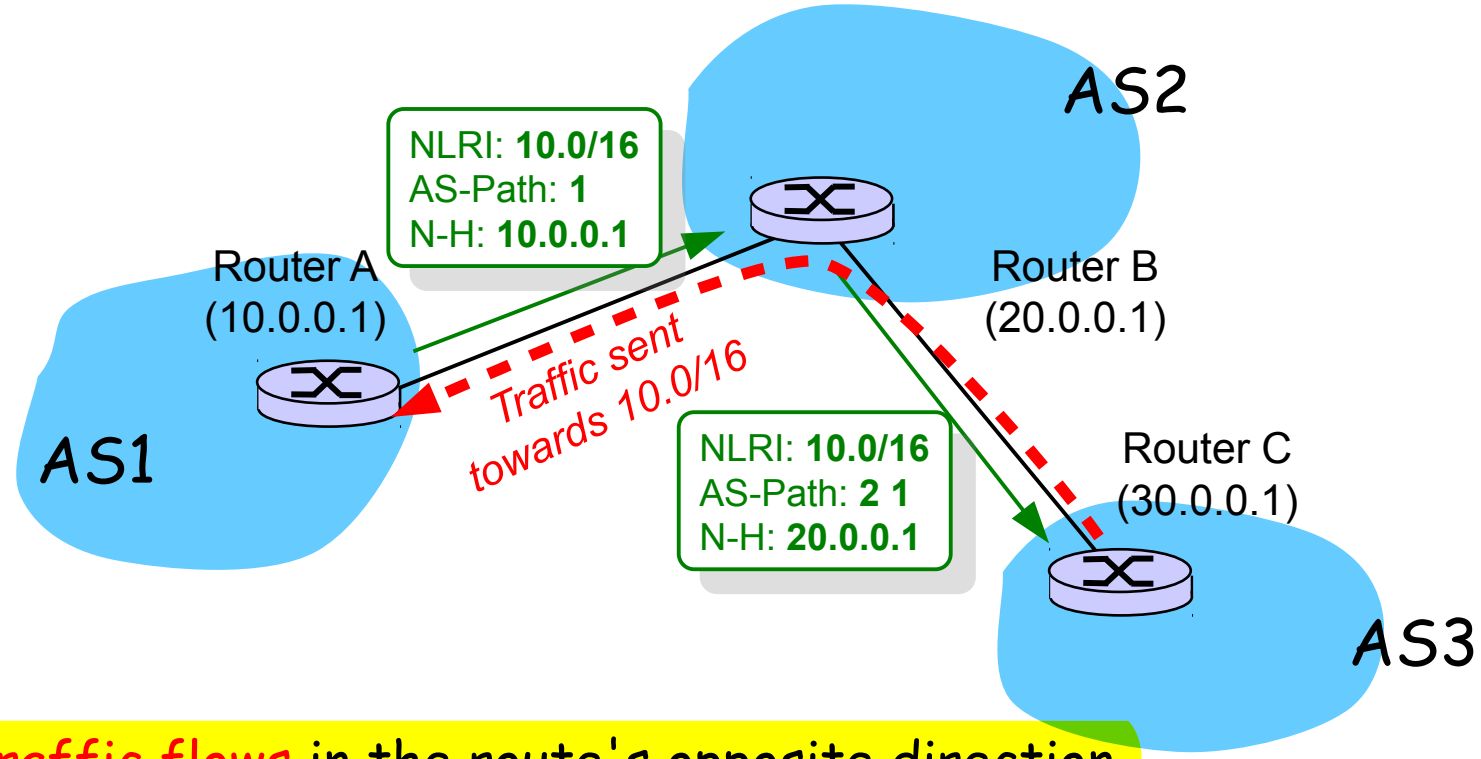
- The ASN of each traversed AS is prepended to the `AS-PATH` attribute.
- The `NEXT-HOP` is updated to specify which router traffic must be sent to



Le Next-Hop est mis à jour pour guider jusqu'au prochain routeur à parcourir.

Route versus Traffic

- A route advertised by a router is a *promise* that the router will carry datagrams towards the route's destination.



- Traffic flows in the route's opposite direction.

Roles of the AS-PATH attribute

- The AS-PATH attribute has a double role
  - It serves as a path metric
    - If multiple routes are available towards the same destination, the one with the shortest AS-PATH length is used.
  - It serves to avoid routing loops
    - A router that received a route with its own ASN in the AS-PATH will discard this route.
    - Most BGP implementations detect that a route, if announced to a neighbor, will cause a loop. This is called *sender-side loop detection*.

Path Attributes

- **Types of attributes**
  - well-known mandatory
    - must be supported by al implementations
    - must appear in every route
  - well-known discretionary
    - must be supported by all implementations
    - can optionally appear in a route
  - optional transitive
    - optional, remains in propagated routes
  - optional non-transitive
    - optional, removed when propagated

Standardized Path Attributes

Name	Type code	Well-known	Mandatory	Transitive	Extensions
ORIGIN	1	Y	Y		
AS_PATH	**2**	**Y**	**Y**		
NEXT_HOP	**3**	**Y**	**Y**		
MULTI_EXIT_DISCRIMINATOR (or MED)	4				
LOCAL_PREF	**5**	**Y**			
ATOMIC_AGGREGATE	6	Y			
AGGREGATOR	7			Y	
COMMUNITIES	8			Y	RFC1997
ORIGINATOR_ID	9				RFC2796
CLUSTER_LIST	10				RFC2796
EXTENDED_COMMUNITIES	16				RFC4360

Déterminer le type de route

Déterminer le nbre de préfixe

Spécifie les gr. pour les politique de routage

Plus précis que Communities

Example BGP Routes and Path Attributes

Prefix | AS-Path | Origin | Next-hop | Local-Pref | MED | Communities | ...

```
6.1.0.0/16|668|IGP|198.32.8.202|100|0|668:100 11537:3000|NAG||
6.2.0.0/22|668|IGP|198.32.8.202|100|0|668:100 11537:3000|NAG||
6.3.0.0/18|668|IGP|198.32.8.202|100|0|668:100 11537:3000|NAG||
6.4.0.0/16|668|IGP|198.32.8.202|100|0|668:100 11537:3000|NAG||
6.5.0.0/19|668|IGP|198.32.8.202|100|0|668:100 11537:3000|NAG||
6.6.0.0/16|668|IGP|198.32.8.202|100|0|668:100 11537:3000|NAG||
6.8.0.0/20|668|IGP|198.32.8.202|100|0|668:100 11537:3000|NAG||
6.9.0.0/20|668|IGP|198.32.8.202|100|0|668:100 11537:3000|NAG||
6.10.0.0/15|668|IGP|198.32.8.202|100|0|668:100 11537:3000|NAG||
6.14.0.0/15|668|IGP|198.32.8.202|100|0|668:100 11537:3000|NAG||
8.6.244.0/23|11096 6356|IGP|198.32.155.193|200|0|11096:307 11096:501 11537:950|NAG||
8.10.208.0/24|10466 32554|IGP|198.32.8.199|260|0|11537:260 11537:910 11537:950|NAG||
9.4.0.0/16|20965 559|IGP|198.32.8.202|100|0|11537:2501 20965:155|NAG||
12.0.48.0/20|10578 1742|IGP|198.32.8.199|200|0|10578:800 10578:840 11537:950|NAG||
12.6.208.0/20|10578 1742|IGP|198.32.8.199|200|0|10578:800 10578:840 11537:950|NAG||
12.107.208.0/23|81 22753|IGP|198.32.8.202|200|0|11537:950 11537:2000|NAG||
12.144.59.0/24|10466 13778|IGP|198.32.8.199|260|0|11537:260 11537:950 11537:2000|NAG||
12.151.0.0/24|10466 11558|IGP|198.32.8.199|260|0|11537:260 11537:902 11537:950|NAG||
12.151.1.0/24|10466 11558|IGP|198.32.8.199|260|0|11537:260 11537:902 11537:950|NAG||
12.161.8.0/21|10466 88|IGP|198.32.8.199|260|0|11537:260 11537:950|NAG||
12.174.210.0/23|5661 21712|IGP|131.247.47.246|200|0|11537:902 11537:950|NAG||
18.0.0.0/8|10578 3|IGP|198.32.8.199|200|0|10578:800 10578:840 11537:950|NAG||
18.3.4.0/24|10578 3|INCOMPLETE|198.32.8.199|200|0|10578:800 10578:840 11537:950|NAG||
...
```


Source: Abilene Observatory, http://www.internet2.edu/observatory/, http://ndb7.net.internet2.edu/bgp/RIBS/

Chapter 2: roadmap

- 2.1 Inter-domain Routing
- 2.2 The Border Gateway Protocol (BGP)
  - 2.2.1 Principles
  - 2.2.2 Sessions
  - 2.2.3 Routes
  - 2.2.4 Path Attributes
  - 2.2.5 Messages
  - 2.2.6 Finite State Machine
  - 2.2.7 Decision Process
  - 2.2.8 Routing Filters
  - 2.2.9 Internal BGP (iBGP)
- ...

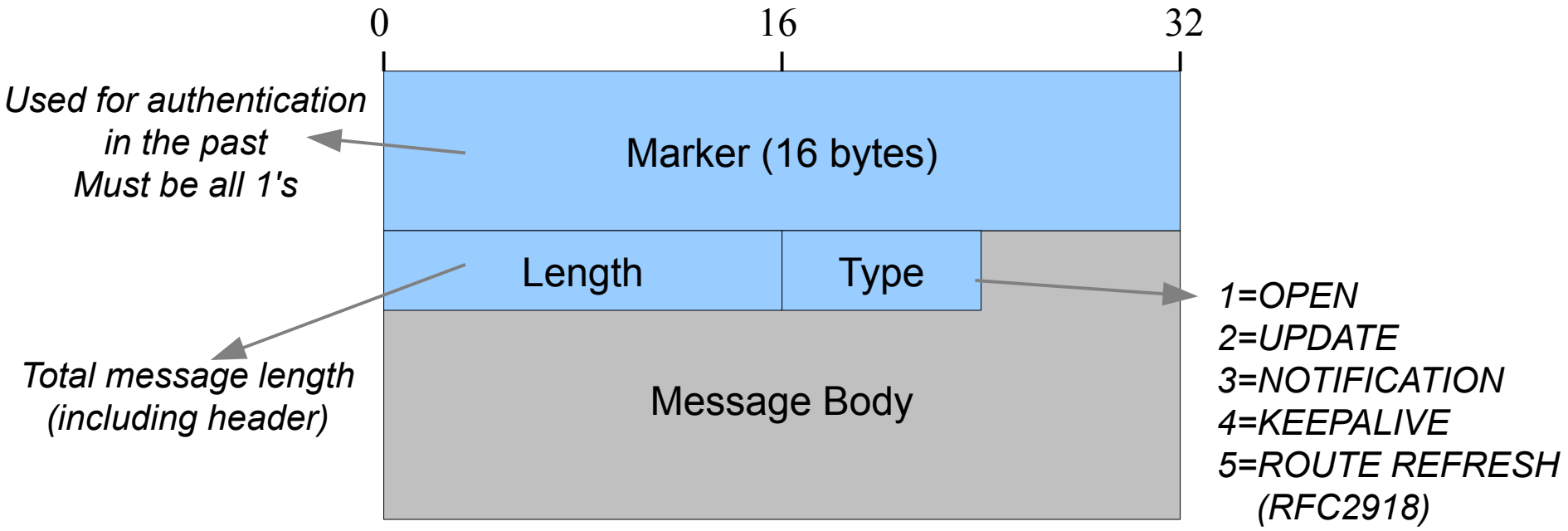
BGP Messages

→ Minimum de sécurité => authentification, puis on fait ce qu'on veut

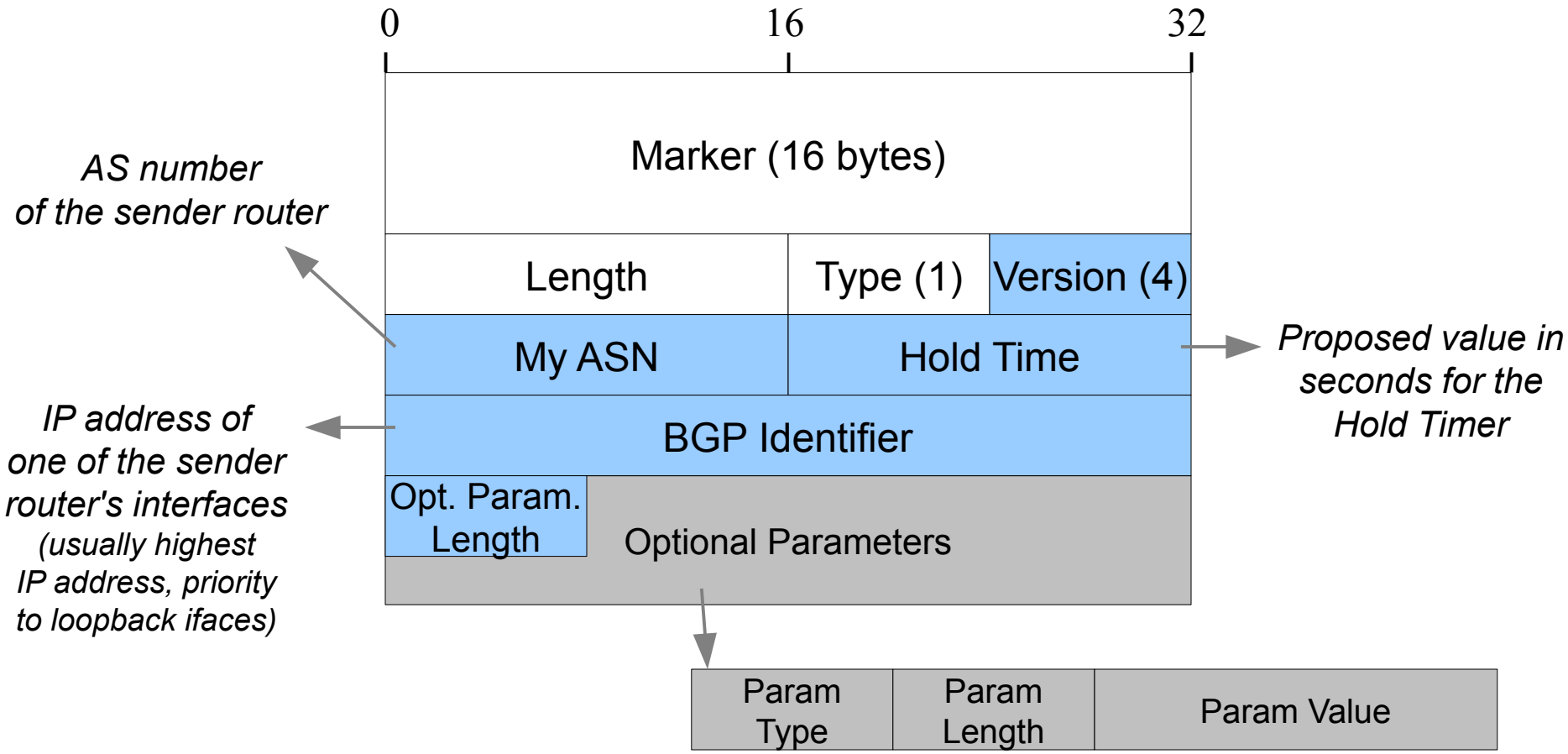
- **BGP Message types**
  - **OPEN**
    - opens TCP connection to peer and optionally authenticates sender
  - **UPDATE**
    - advertises new path (and/or withdraws existing)
  - **KEEPALIVE**
    - keeps connection alive in the absence of UPDATE messages; also acks OPEN request
  - **NOTIFICATION**
    - reports errors in previous message; also used to close connection (Cease)

BGP Messages

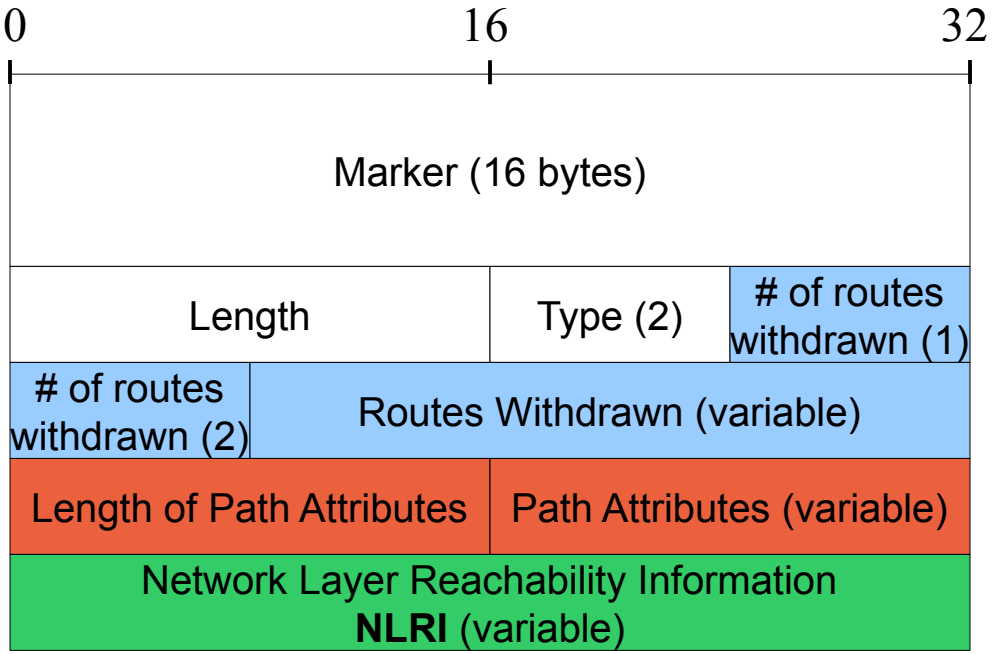
- **Header Format**
  - Each message has a fixed size header
  - The maximum size for a BGP message is 4096 bytes

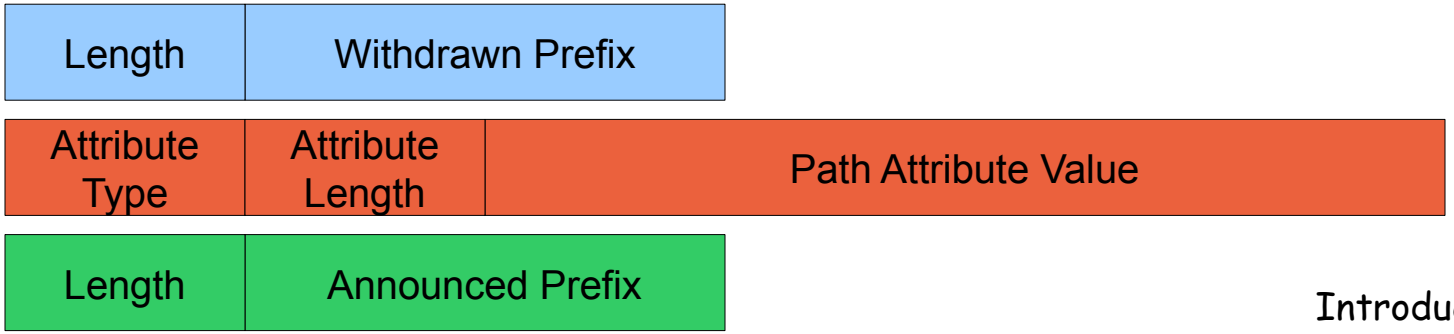


OPEN message



UPDATE message





Chapter 2: roadmap

- 2.1 Inter-domain Routing
- 2.2 The Border Gateway Protocol (BGP)
  - 2.2.1 Principles
  - 2.2.2 Sessions
  - 2.2.3 Messages
  - 2.2.4 Routes
  - 2.2.5 Path Attributes
  - 2.2.6 Finite State Machine
  - 2.2.7 Decision Process
  - 2.2.8 Routing Filters
  - 2.2.9 Internal BGP (iBGP)
- ...

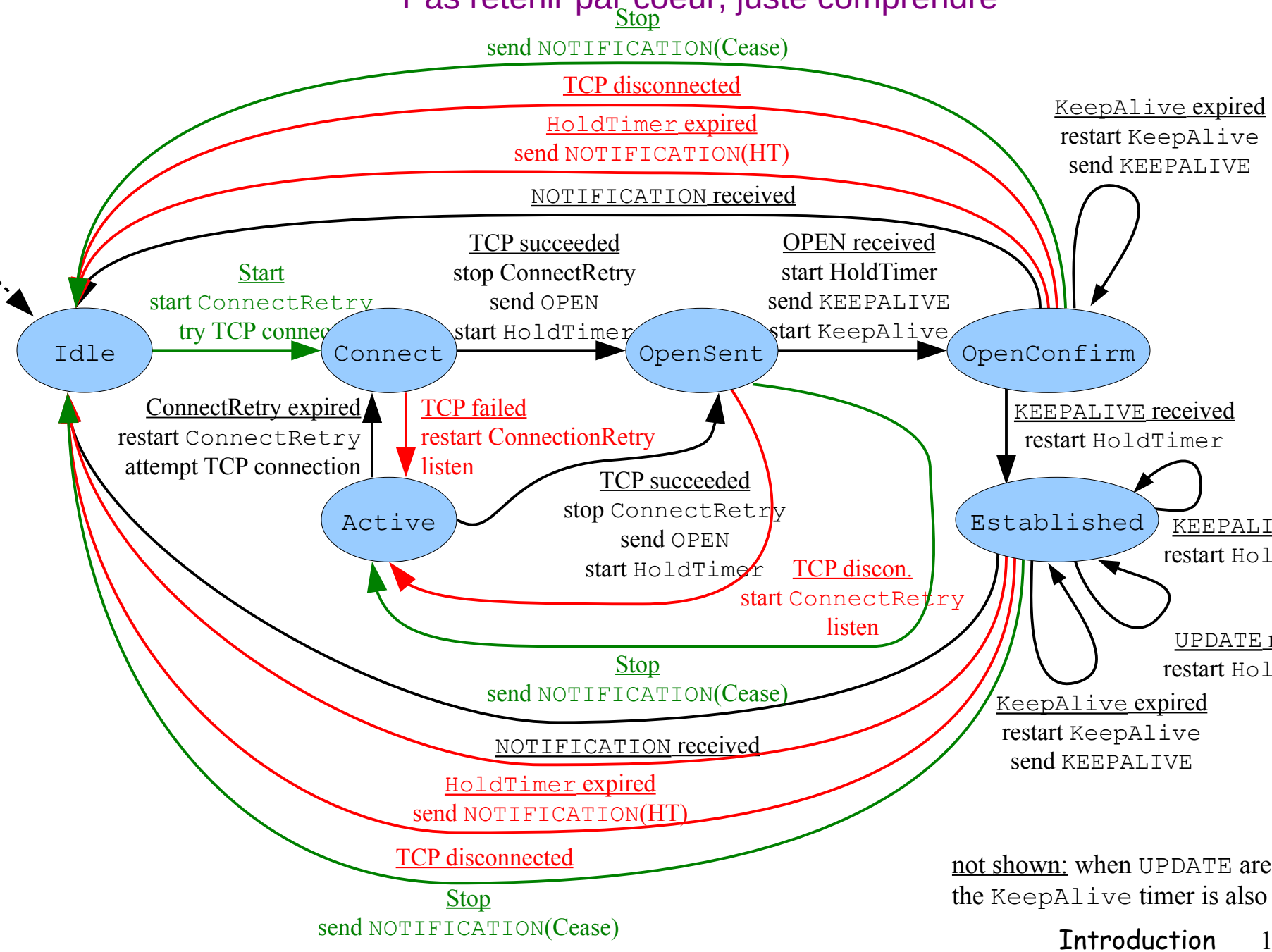
Finite State Machine (FSM)

- **Objectives**
  - The aim of the BGP Finite State Machine is to ensure the proper handling of BGP messages as well as to ensure the robustness of BGP sessions.
  - The FSM is composed of 6 different states.
  - 3 timers are also used to manage BGP sessions.

Finite State Machine (FSM)

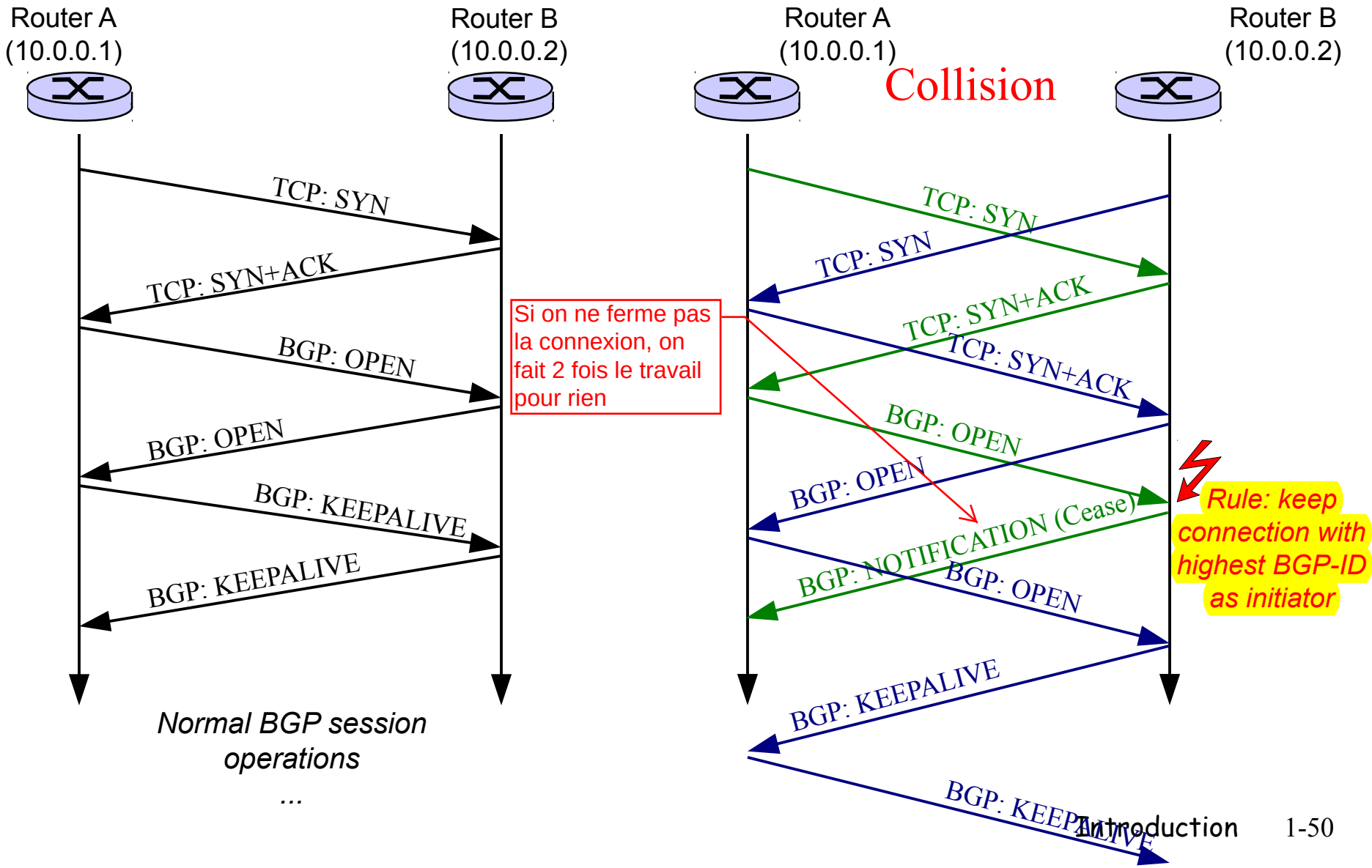
- 6 States
  - **Idle** - initial state
  - **Connect** - try to initiate TCP connection
  - **Active** - listen to incoming TCP connections
  - **OpenSent** - OPEN sent, waiting for incoming OPEN
  - **OpenConfirm** - OPEN received, KEEPALIVE sent, waiting for incoming KEEPALIVE
  - **Established** - up and running, exchanging KEEPALIVE and UPDATE
- 3 Timers
  - **ConnRetry** - spaces TCP connection attempts
  - **HoldTimer** - used to check BGP session activity
  - **KeepAlive** - sends KEEPALIVE on a regular basis

Pas retenir par coeur, juste comprendre



not shown: when UPDATE are sent
the KeepAlive timer is also restarte

BGP connection collision

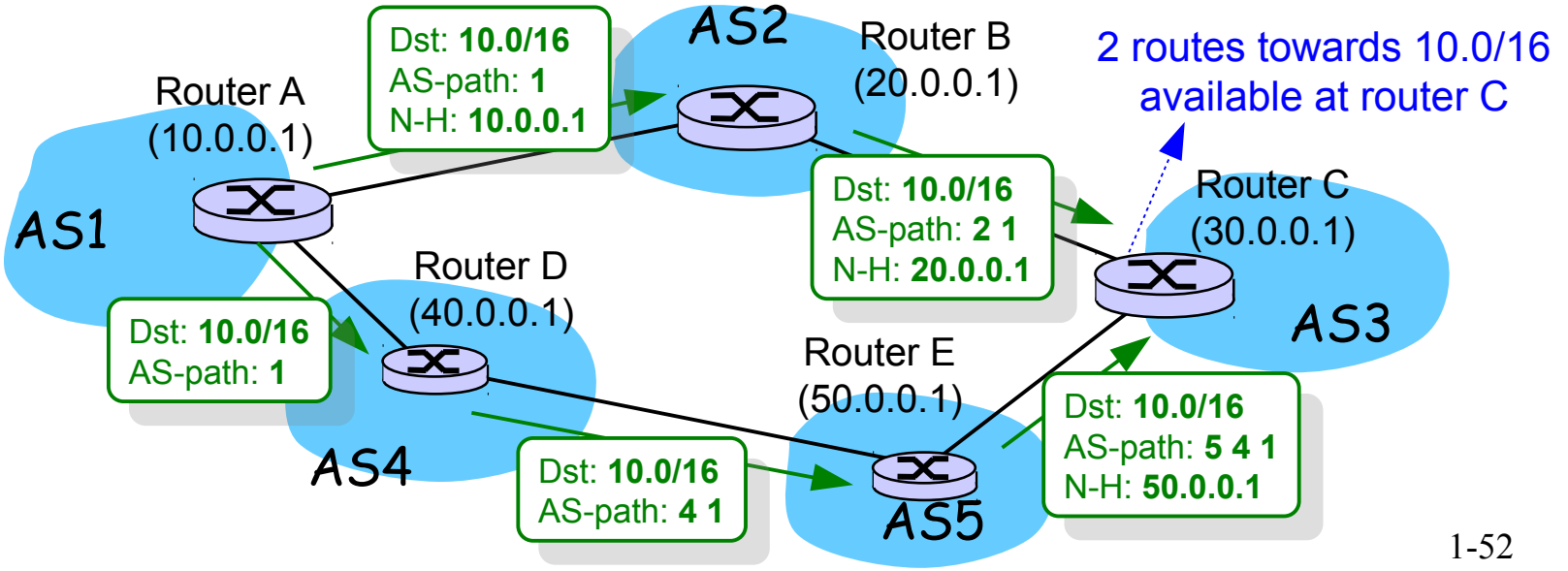


Chapter 2: roadmap

- 2.1 Inter-domain Routing
- 2.2 The Border Gateway Protocol (BGP)
  - 2.2.1 Principles
  - 2.2.2 Sessions
  - 2.2.3 Routes
  - 2.2.4 Path Attributes
  - 2.2.5 Messages
  - 2.2.6 Finite State Machine
  - 2.2.7 Decision Process
  - 2.2.8 Routing Filters
  - 2.2.9 Internal BGP (iBGP)
- ...

Routes Selection

- **Objectives**
  - A router will often receive multiple routes to reach the same destination prefix. When a router has multiple alternative routes it needs to pick a single best route for forwarding.

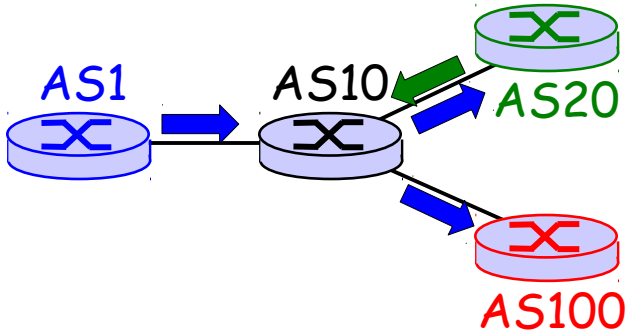


Routes Selection

- **BGP Route Information Bases**
  - A BGP router uses several data structures for storing routes
  - Adj-RIB-in[1]
    - stores received routes
    - one per neighbor
  - Loc-RIB
    - stores best routes (selected by decision process)
    - unique
  - Adj-RIB-out[1]
    - stores sent routes
    - used to avoid sending duplicate messages
    - one per neighbor

(1) "Adj" stands for adjacent.

Decision Process



Inside AS10's BGP Router

AS 1

Dst: **1.0/16**
AS-path: **1**

AS 20

Dst: **1.0/16**
AS-path: **20 1**

AS 100

no route

RIB-in

RIB-in

RIB-in

BGP decision process

Loc-RIB

Dst: **1.0/16**
AS-path: **1**

RIB-out

RIB-out

RIB-out

nothing sent (same source)

AS 1

Dst: **1.0/16**
AS-path: **10 1**

AS 20

Dst: **1.0/16**
AS-path: **10 1**

AS 100

control plane

data plane

Single best route towards 1.0/16

used for forwarding

FIB

Route Selection

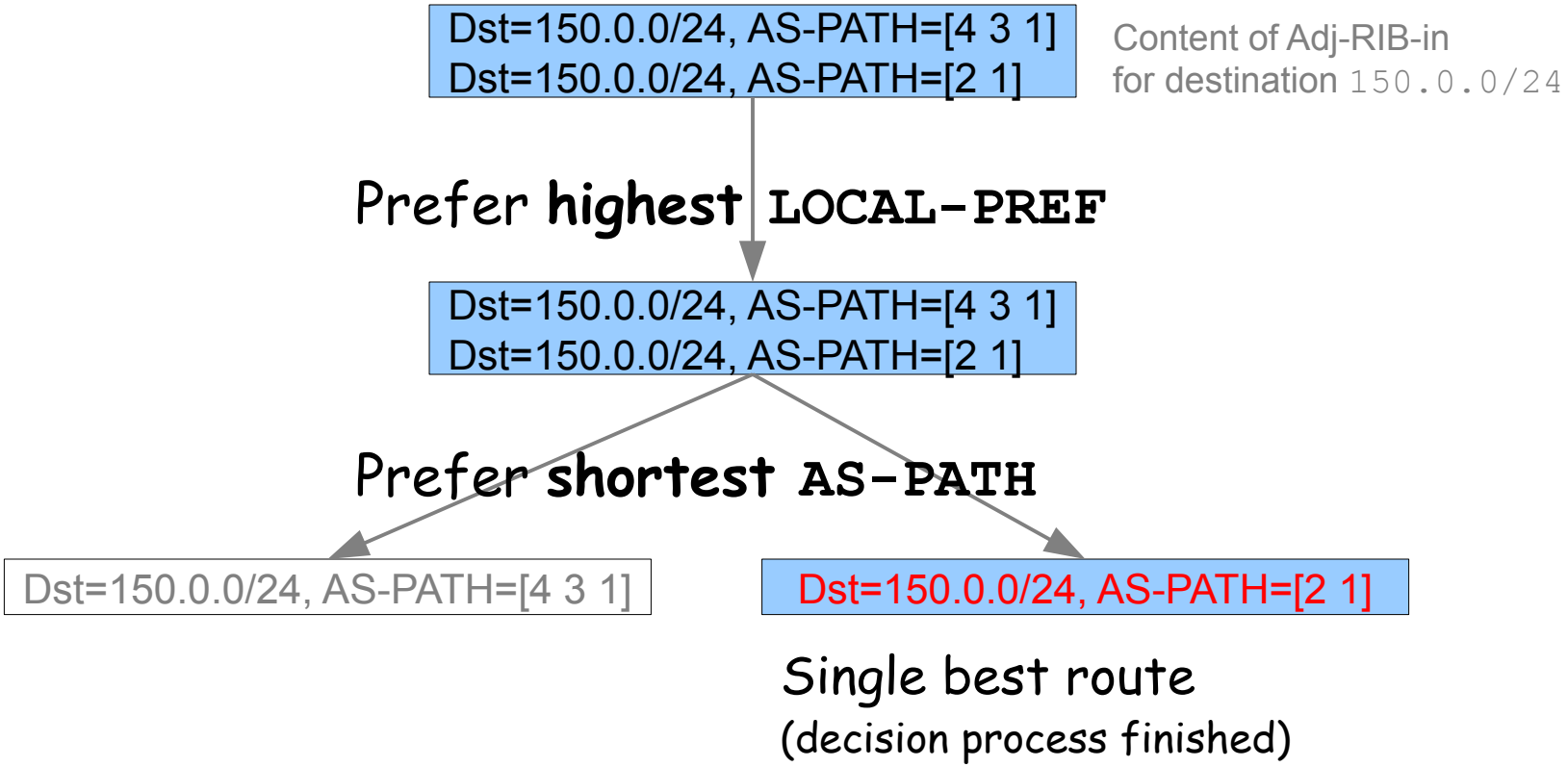
- The BGP Decision Process is used to select a single best route towards each destination prefix
  - It is composed of several rules applied in sequence on the set of available routes.
  - After each rule, the non-dominant routes are removed from the set.
  - The decision process stops when a single route remains (the so-called best route)
  - The best route is used for forwarding and it is propagated to all neighbor routers except the one that advertised the route.

Decision Process (simplified)

1. Ignore if next-hop unreachable
2. Prefer **highest** **`LOCAL-PREF`**
3. Prefer **shortest** **`AS-PATH`**
4. Tie-break

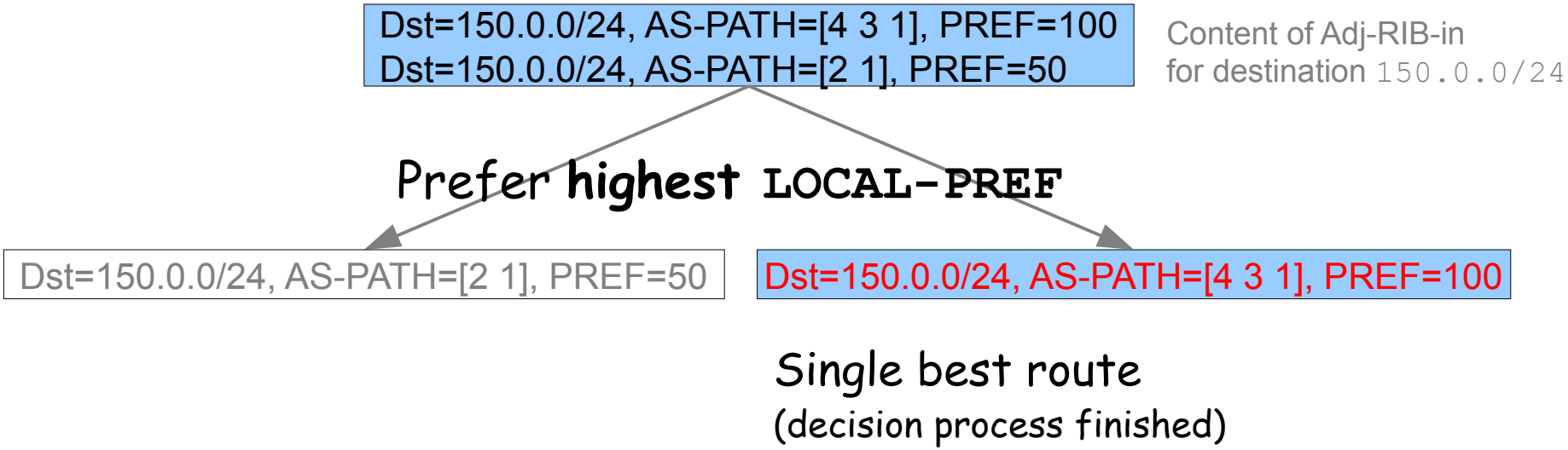
Decision Process

- Example



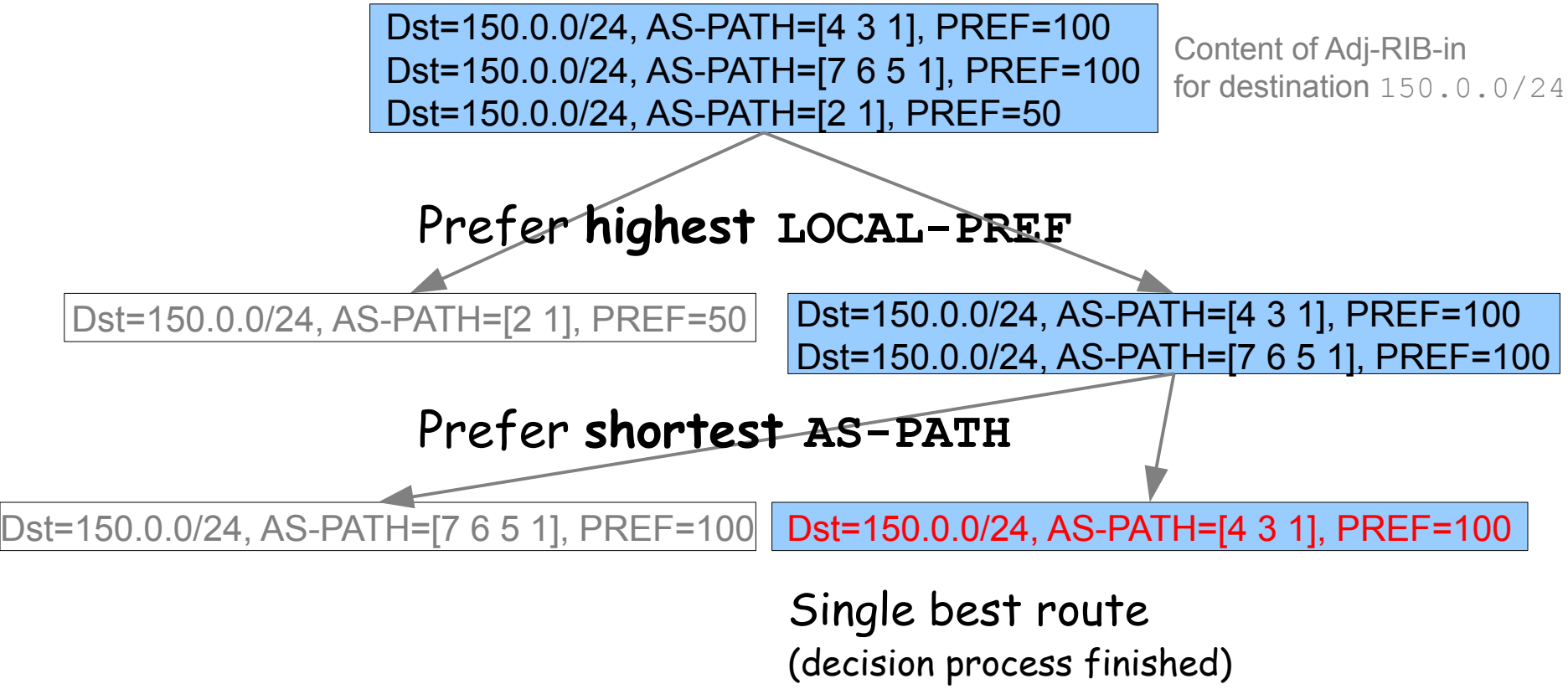
Decision Process

- Example



Decision Process

- Example



Decision Process (almost complete)

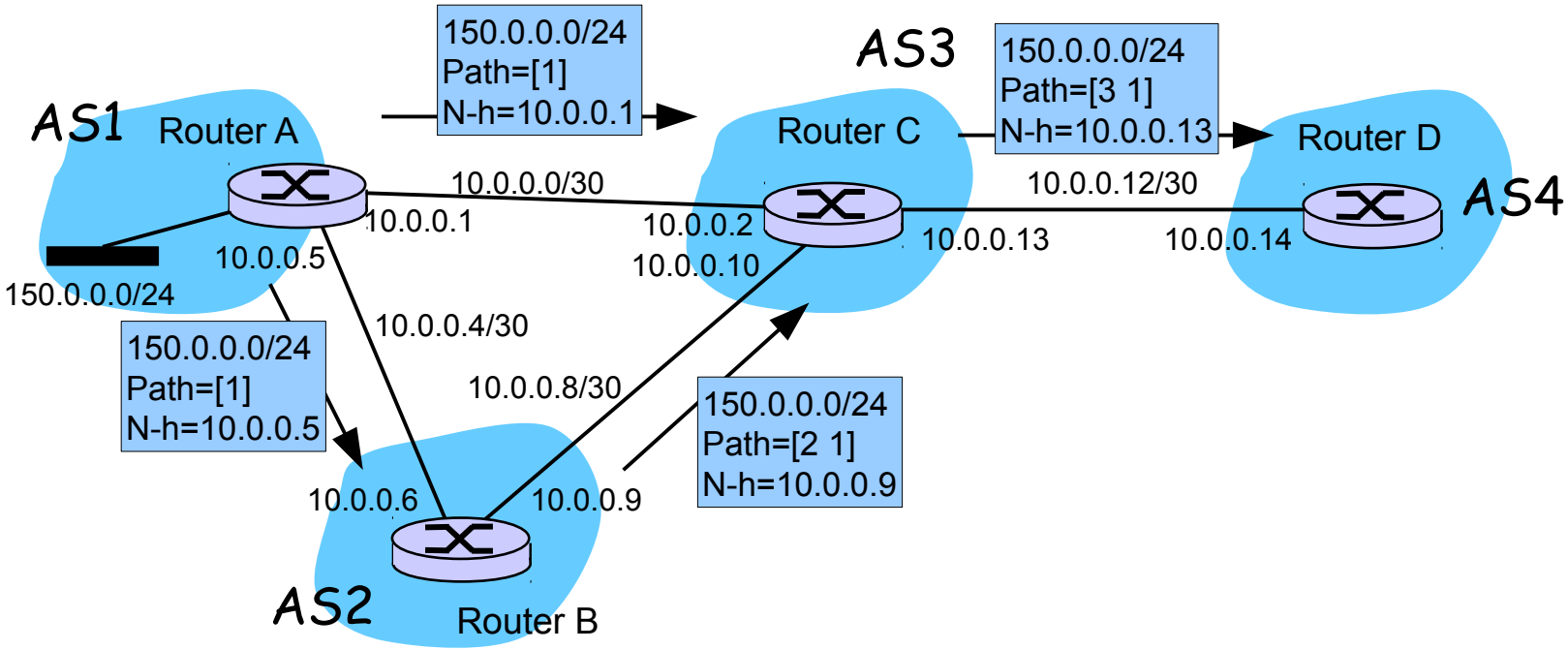
1. Ignore if next-hop unreachable
2. Prefer locally originated networks
3. Prefer **highest** `LOCAL-PREF`
4. Prefer **shortest** `AS-PATH`
5. Prefer **lowest** `ORIGIN`
6. Prefer **lowest** `MED`
7. Prefer **eBGP over iBGP**
8. Prefer **nearest next-hop**

tie-breaks

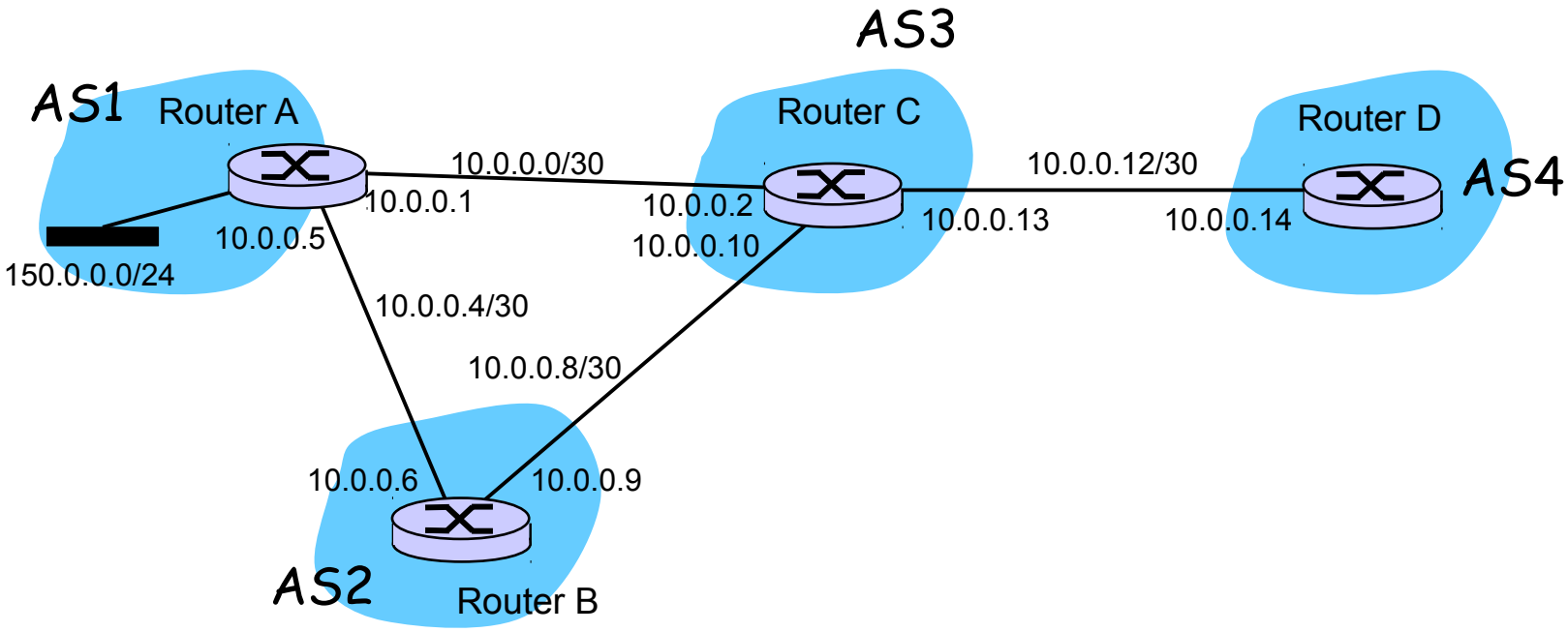
9. Prefer **lowest Router-ID** / `ORIGINATOR-ID`
10. Prefer **shortest** `CLUSTER-LIST`
11. Prefer **lowest neighbor address**

Note : the exact decision process may slightly vary from vendor to vendor. See for example http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080094431.shtml

Time for some more examples



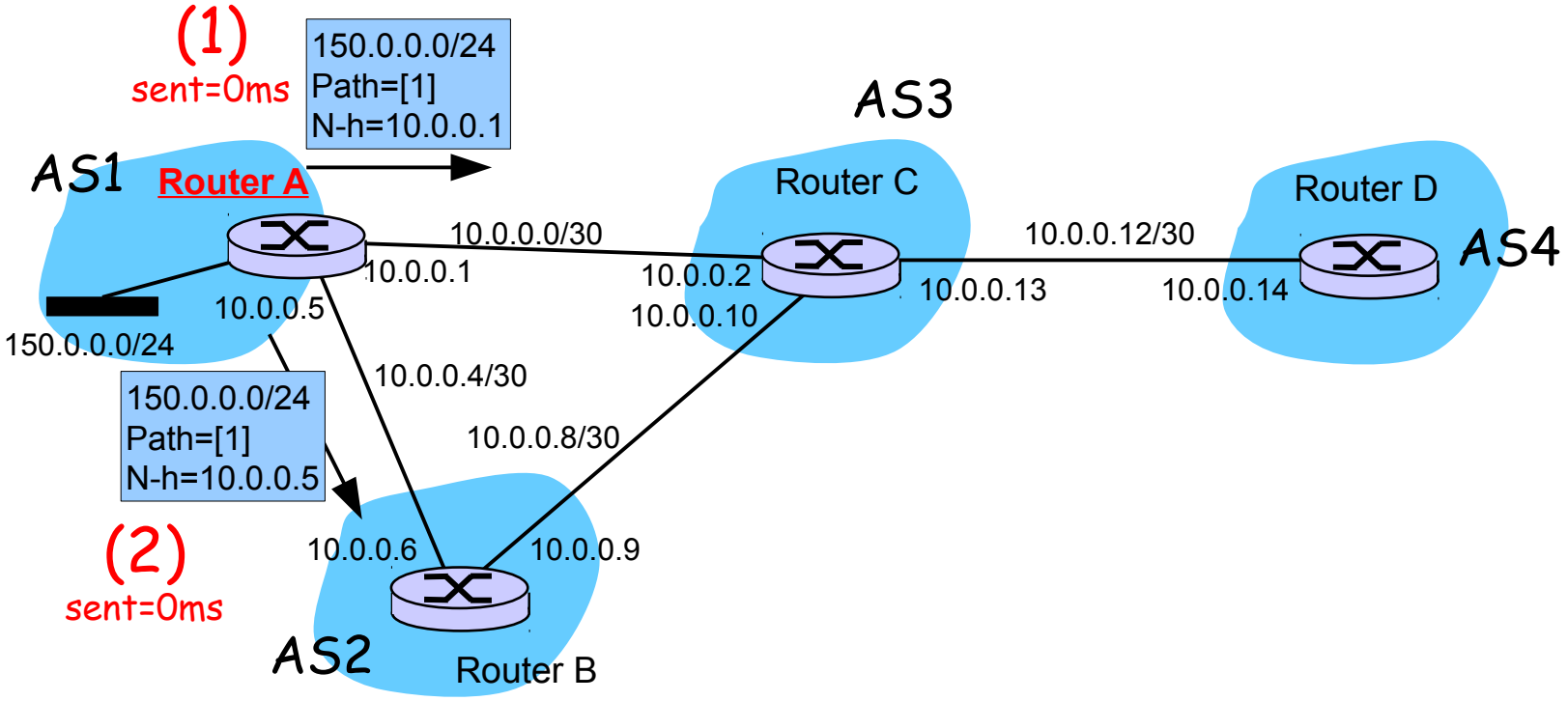
Slow motion…



Let's restart with an additional hypothesis. UPDATE messages take some time to fly from one router to another. Flight times are given below:

A → B: 10ms
B → C: 10ms
A → C: 100ms
C → D: 10ms

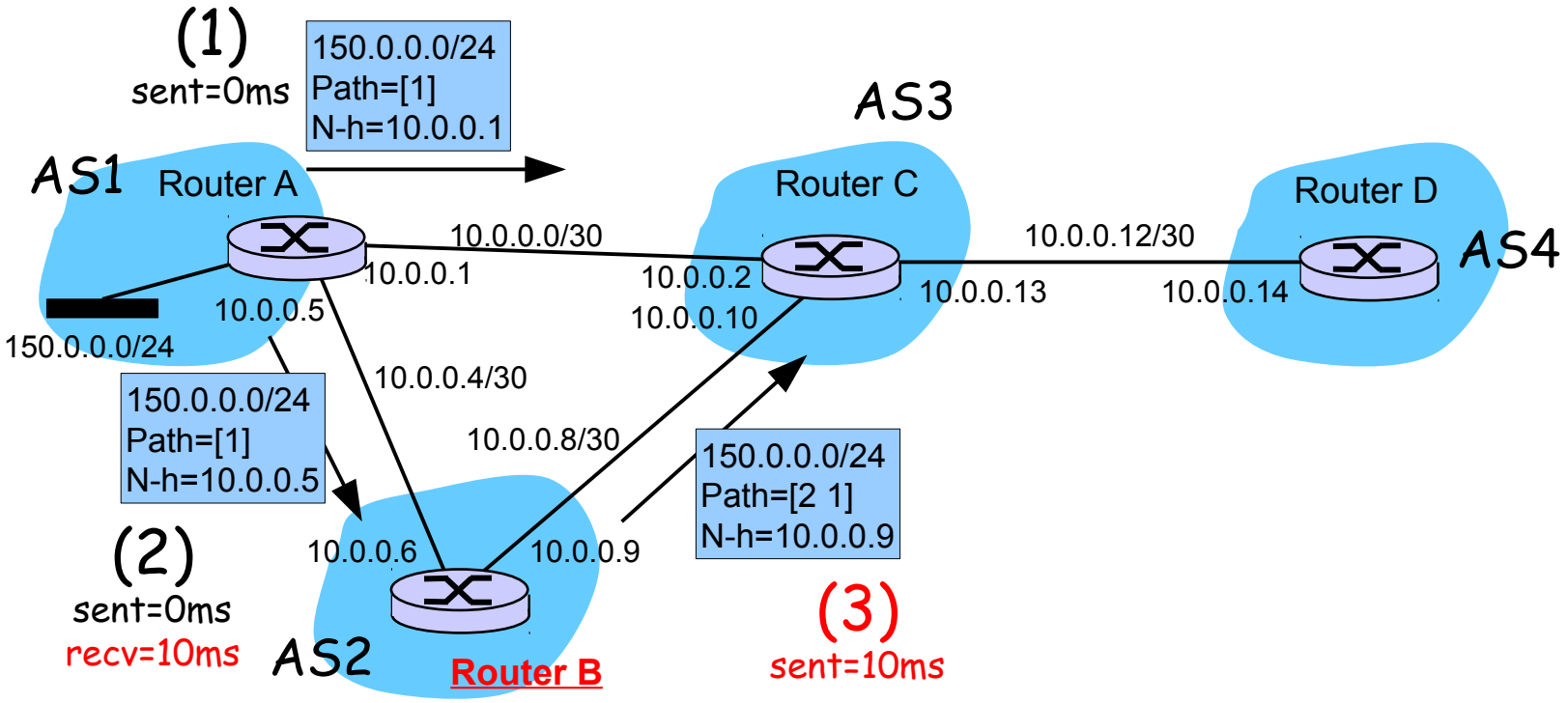
Slow motion...



Let's restart with an additional hypothesis. UPDATE messages take some time to fly from one router to another. Flight times are given below:

A → B: 10ms
B → C: 10ms
A → C: 100ms
C → D: 10ms

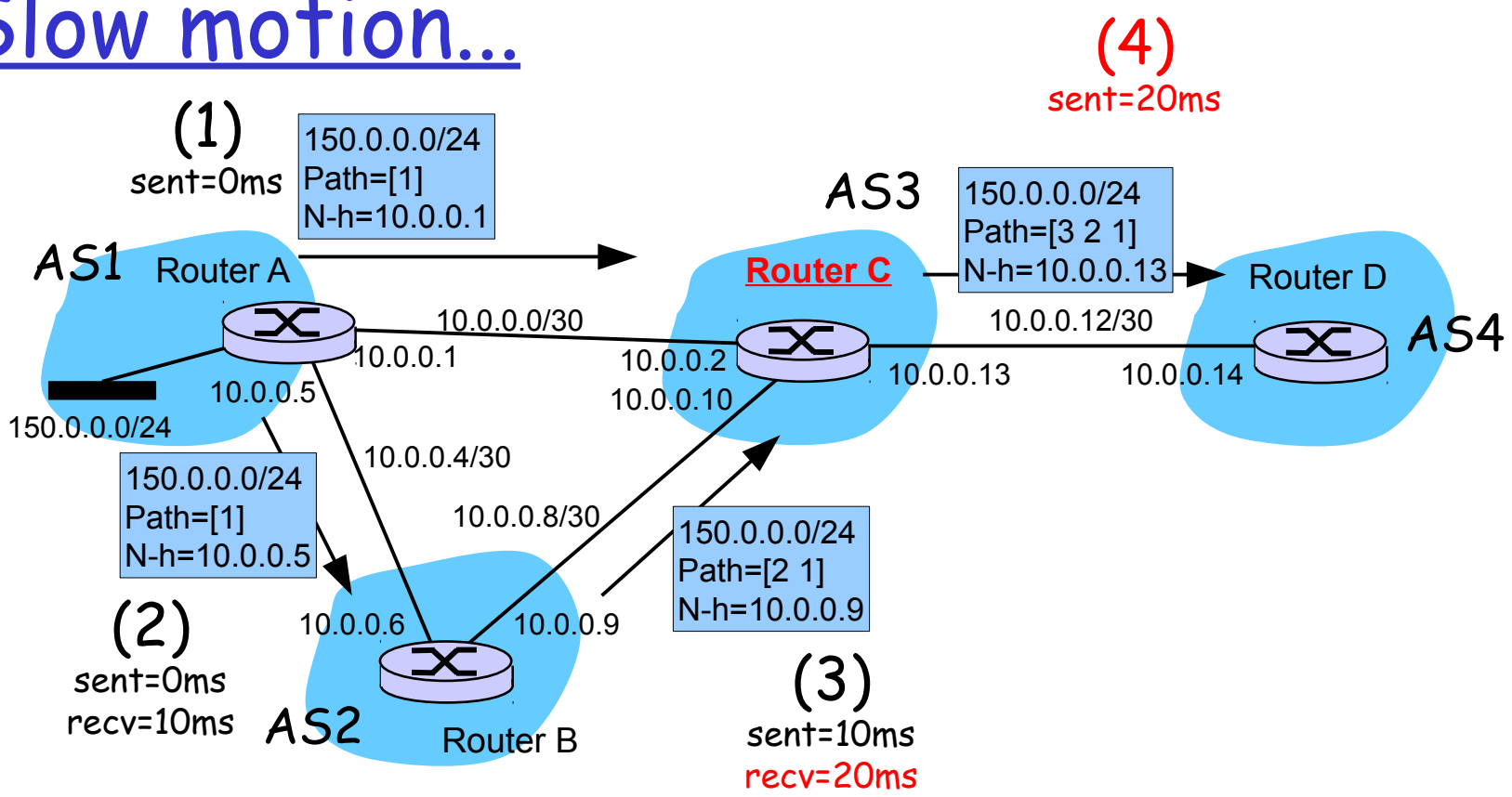
Slow motion…



Let's restart with an additional hypothesis. UPDATE messages take some time to fly from one router to another. Flight times are given below:

A → B: 10ms
B → C: 10ms
A → C: 100ms
C → D: 10ms

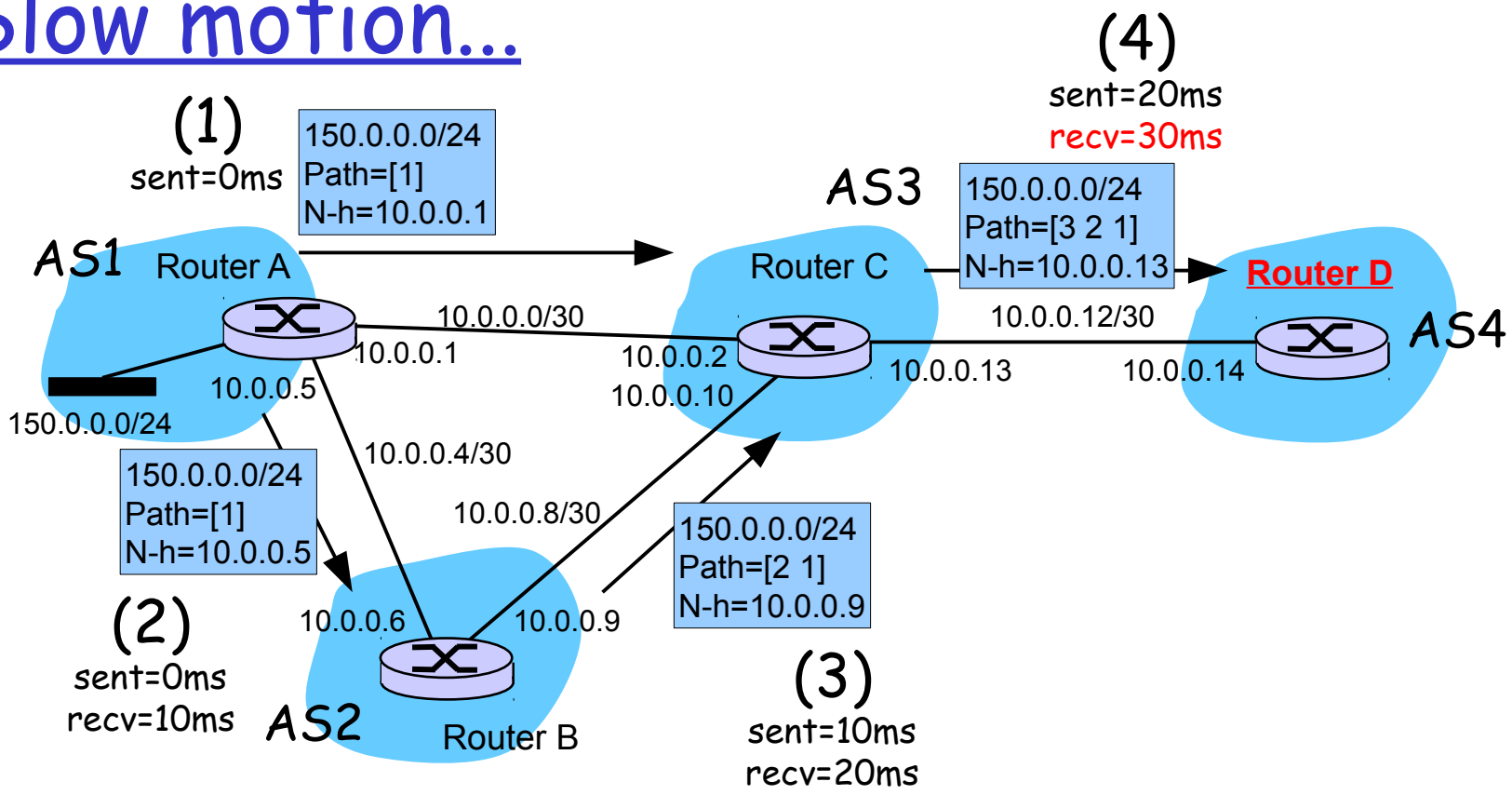
Slow motion...



Let's restart with an additional hypothesis. UPDATE messages take some time to fly from one router to another. Flight times are given below:

A → B: 10ms
B → C: 10ms
A → C: 100ms
C → D: 10ms

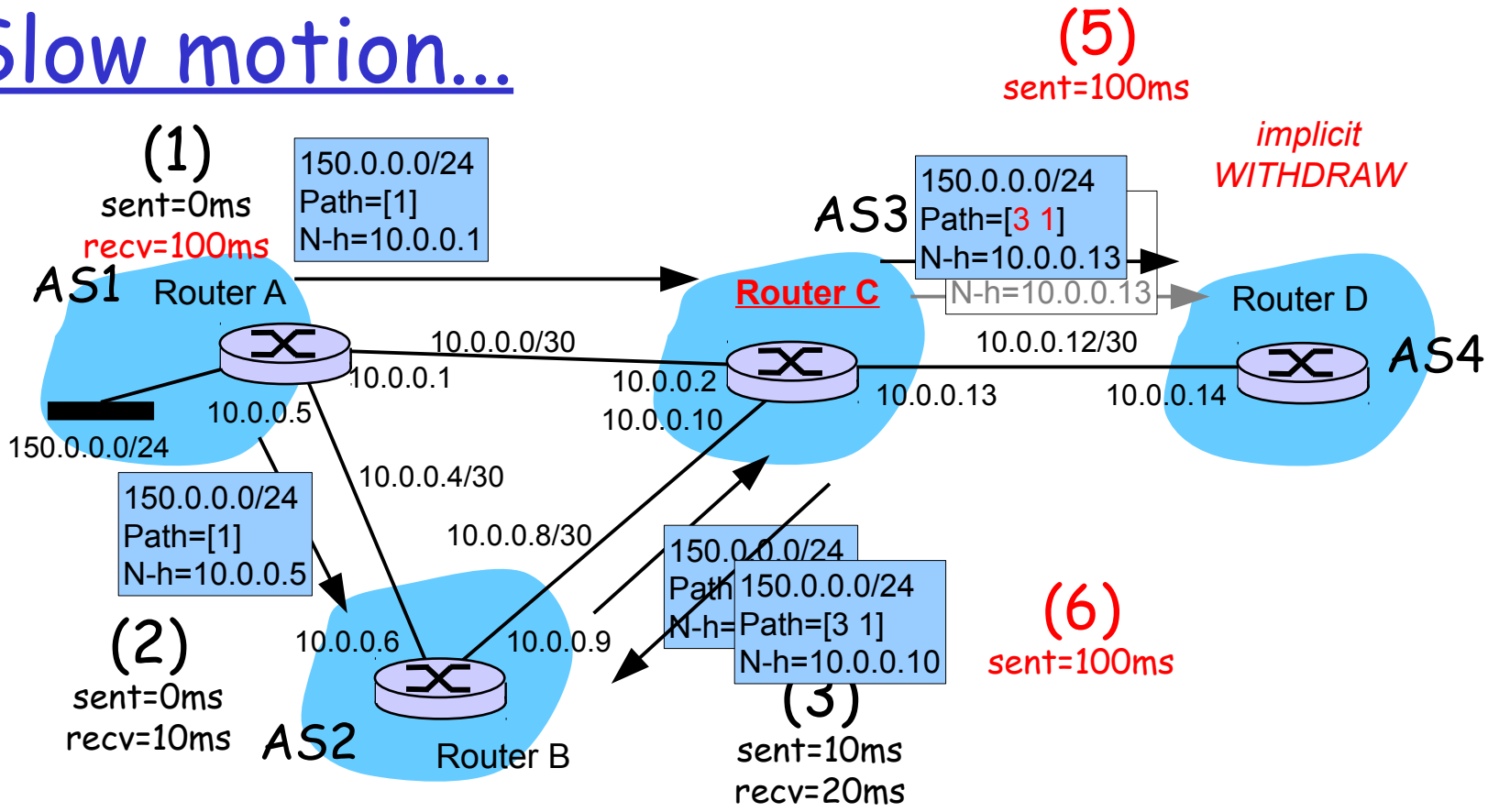
Slow motion...



Let's restart with an additional hypothesis. UPDATE messages take some time to fly from one router to another. Flight times are given below:

A → B: 10ms
B → C: 10ms
A → C: 100ms
C → D: 10ms

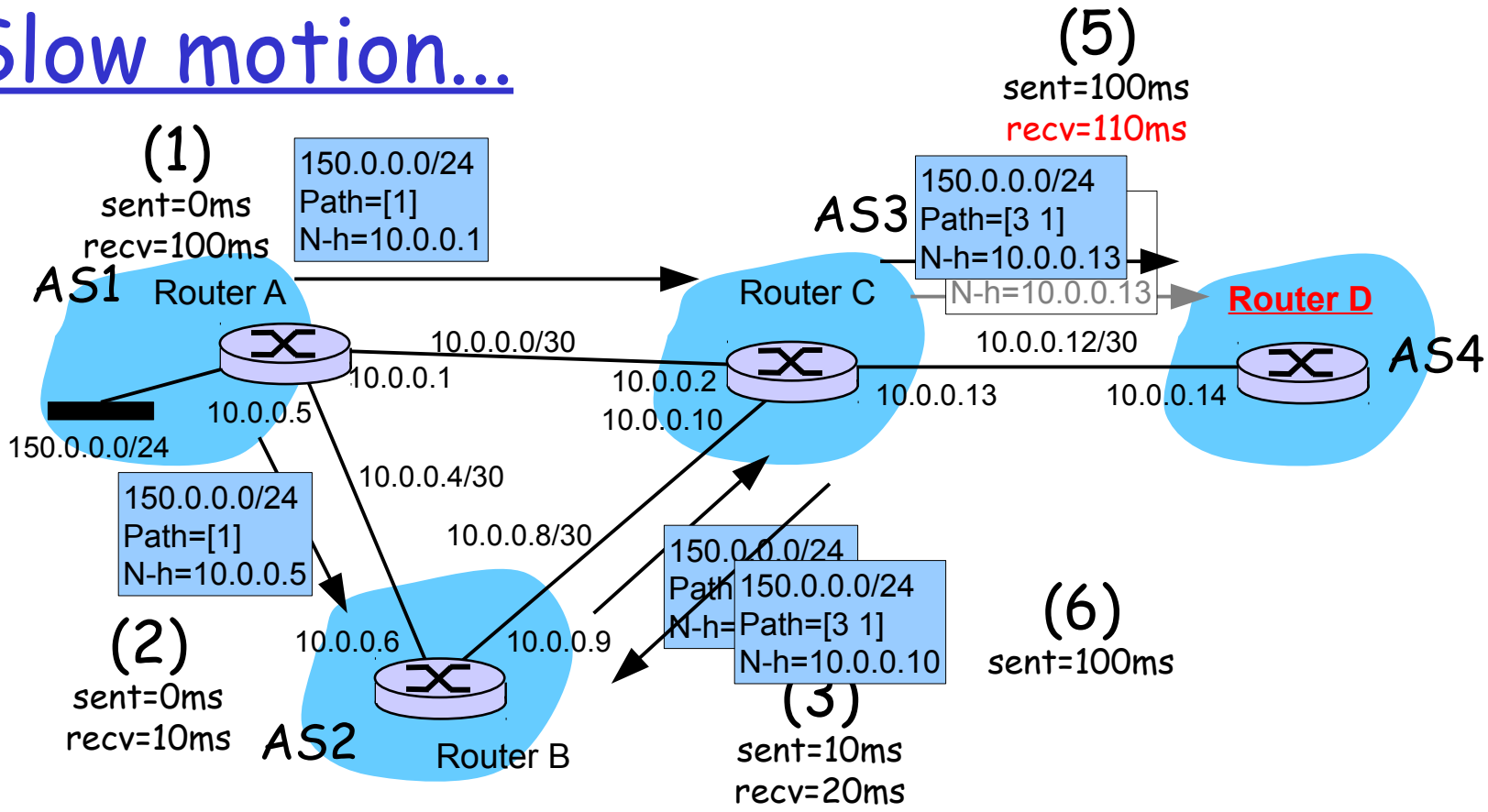
Slow motion...



Let's restart with an additional hypothesis. UPDATE messages take some time to fly from one router to another. Flight times are given below:

A → B: 10ms
B → C: 10ms
A → C: 100ms
C → D: 10ms

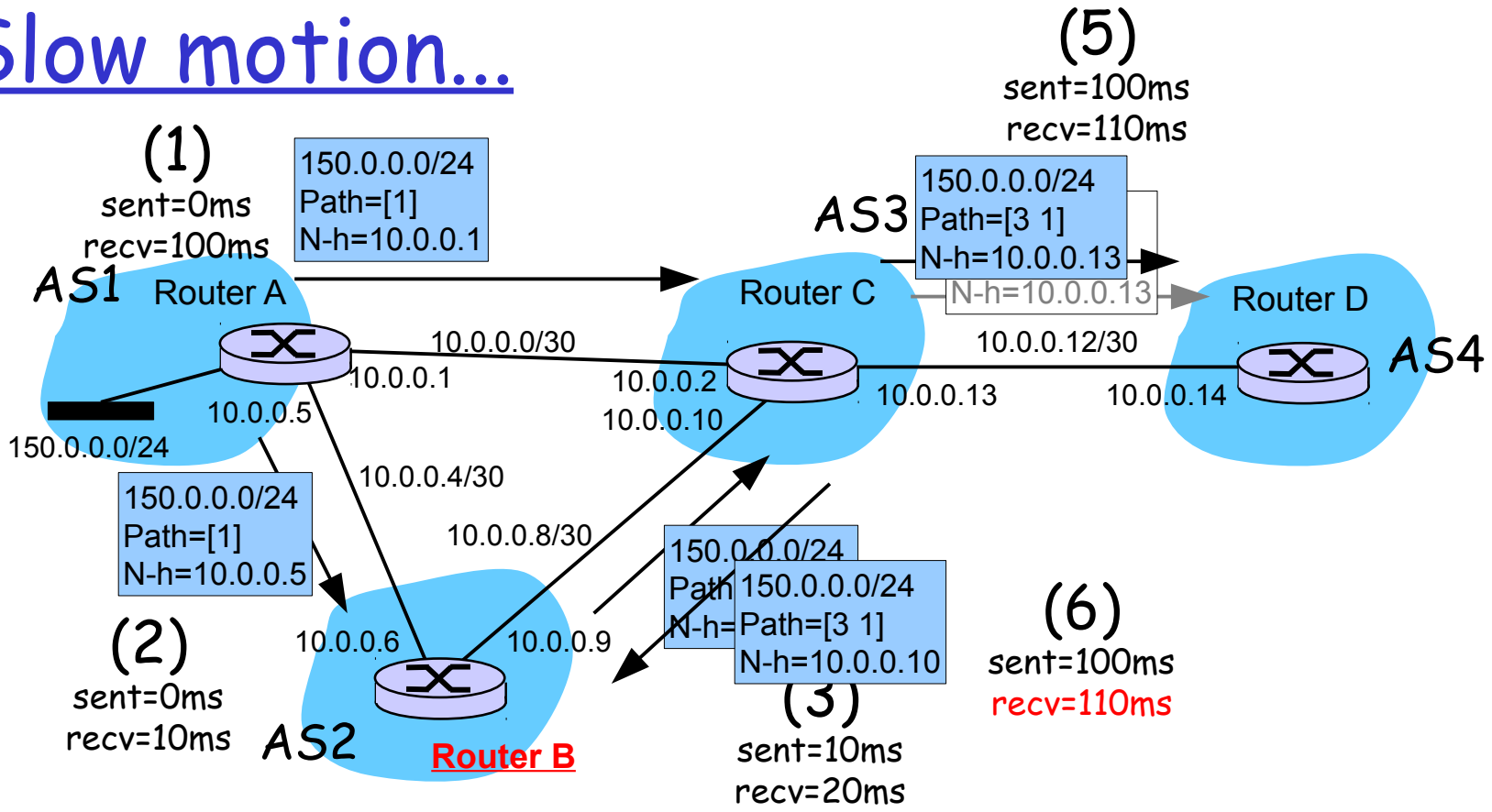
Slow motion...



Let's restart with an additional hypothesis. UPDATE messages take some time to fly from one router to another. Flight times are given below:

A → B: 10ms
B → C: 10ms
A → C: 100ms
C → D: 10ms

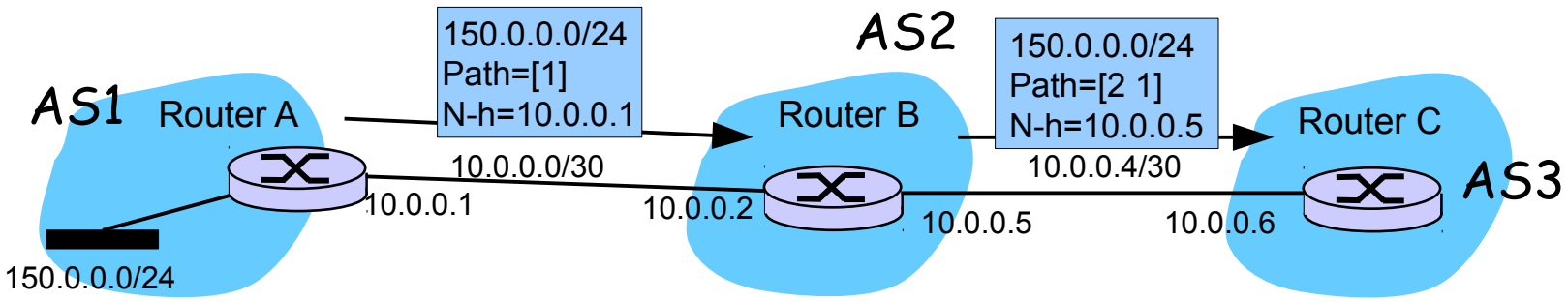
Slow motion...



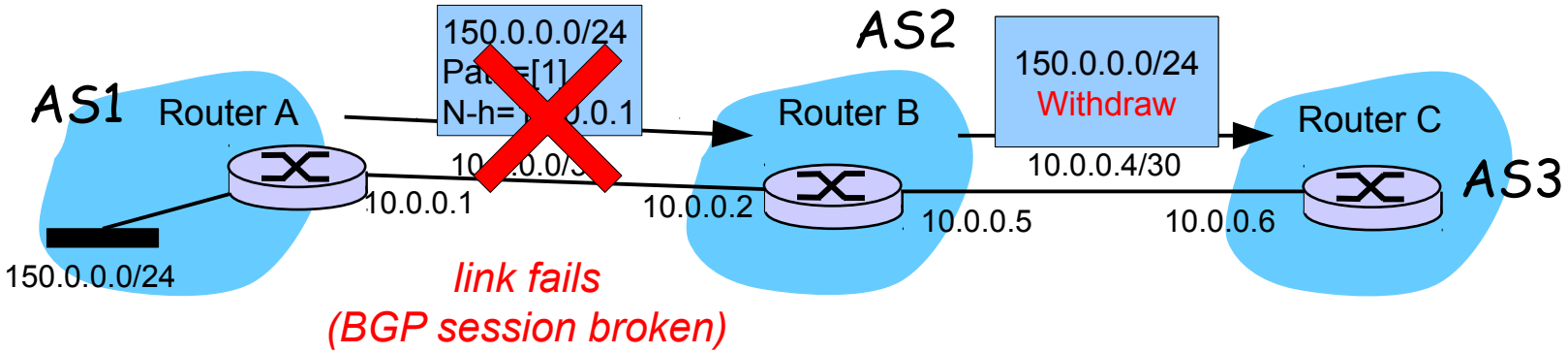
Let's restart with an additional hypothesis. UPDATE messages take some time to fly from one router to another. Flight times are given below:

A → B: 10ms
B → C: 10ms
A → C: 100ms
C → D: 10ms

Another example



Another example



Chapter 2: roadmap

- 2.1 Inter-domain Routing
- 2.2 The Border Gateway Protocol (BGP)
  - 2.2.1 Principles
  - 2.2.2 Sessions
  - 2.2.3 Routes
  - 2.2.4 Path Attributes
  - 2.2.5 Messages
  - 2.2.6 Finite State Machine
  - 2.2.7 Decision Process
  - 2.2.8 Routing Filters
  - 2.2.9 Internal BGP (iBGP)
- ...

Routing Filters

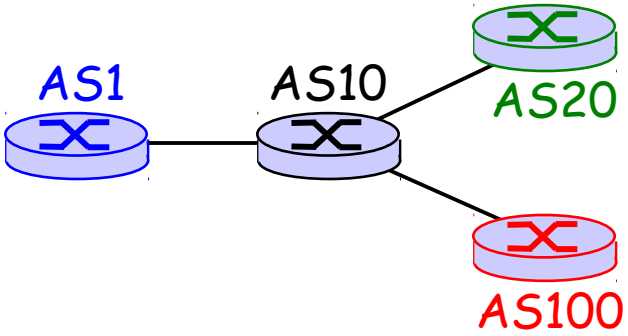
- **Objectives**
  - How to influence the ranking of routes ?
    - example : prefer cheap link over expensive link
    - example : send traffic over primary link instead of backup link
  - How to prevent routes from being redistributed to a specific neighbor
    - example : a route from a provider must not be sent to another provider
  - How to reject a route from a specific neighbor ?
    - example : avoid to send traffic through that neighbor

Routing Filters

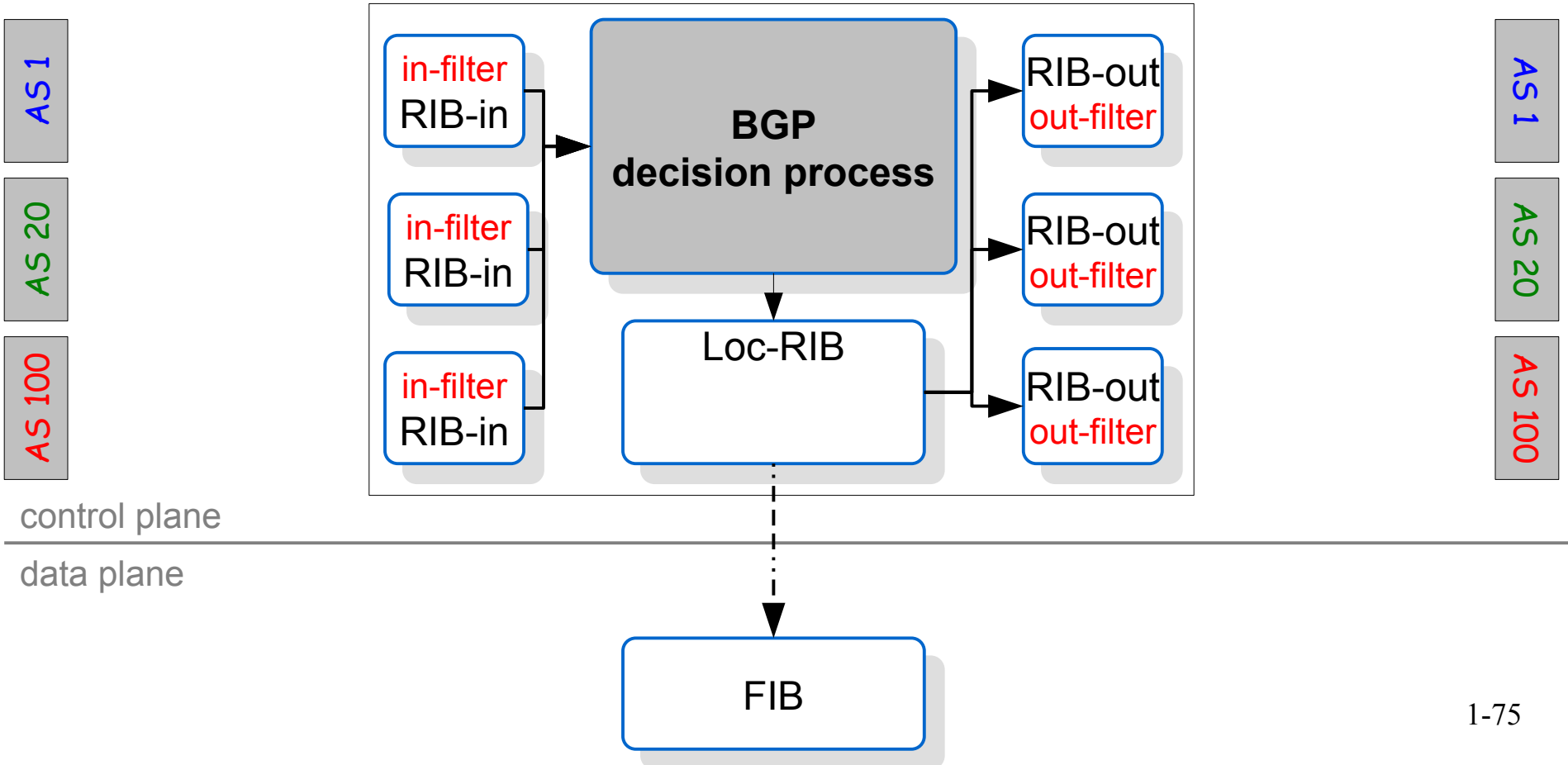
- **Principles**
  - Add import and export filters to our BGP router model
  - Can be done separately for each neighbor
  - Filters are expressed in a vendor-specific policy definition language

  - Import filter (in-filter)
    - select acceptable routes
    - change route attributes
  - Export filter (out-filter)
    - select re-distributable routes
    - change route attributes

Routing Filters

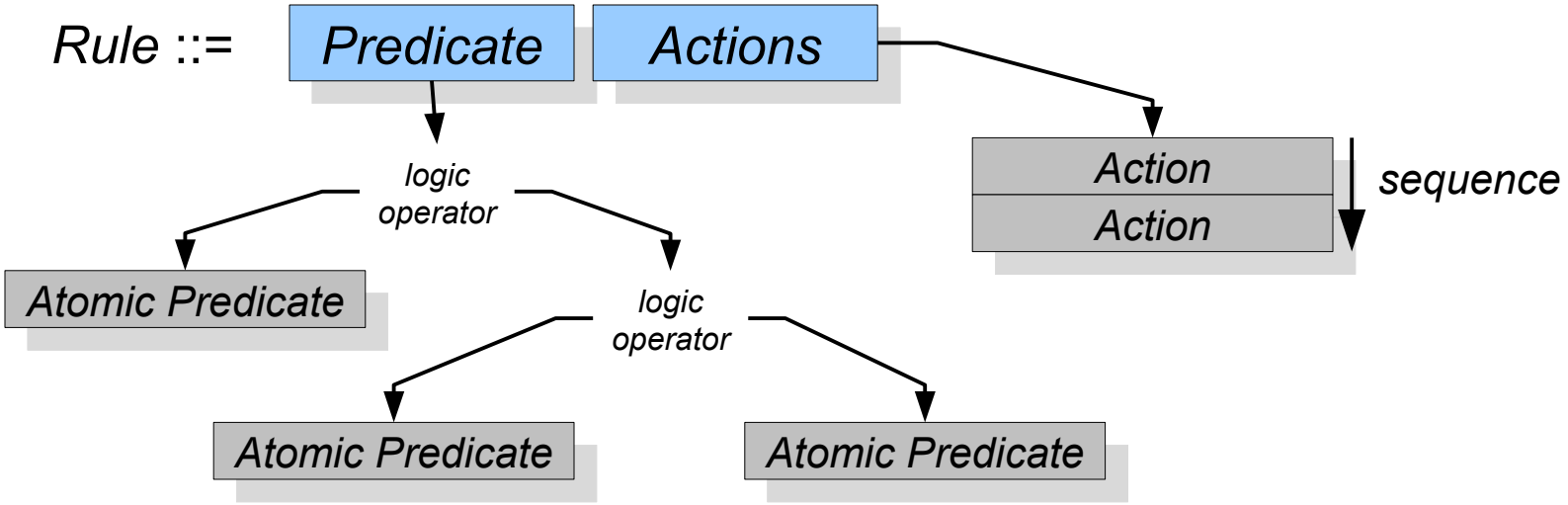


Inside AS10's BGP Router



Routing Filters

- **Structure of filters**
  - Expressed as a sequence of rules
  - Each rule is composed of a predicate and a sequence of actions. If the predicate is true, the actions are applied.



Routing Filters

- **Filter algorithm**

```
def filter_apply(filter, route):
    for rule in filter.rules:
        if predicate_matches(rule.predicate, route):
            for action in rule.actions:
                result= action_apply(action, route)
                if result in (ACCEPT, DENY):
                    return result
    return ACCEPT
```

Routing Filters

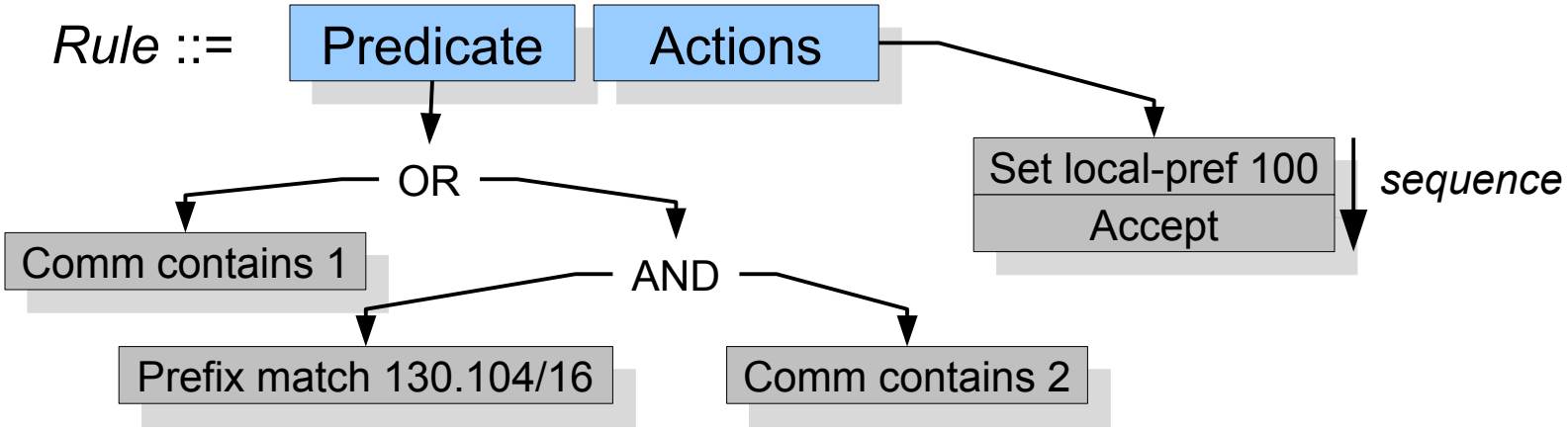
- **Atomic Predicates**
  - match destination prefix against IP prefix
  - match next-hop against IP address / prefix
  - test existence of an ASN in AS-Path
  - match AS-Path against regular expression
  - …
    - (depends on creativity of router vendor)

Routing Filters

- **Actions**
  - Accept / reject the route
  - Set / Increase / Decrease Local-Pref
  - Prepend AS-Path
  - Set Multi-Exit-Discriminator
  - Add / Remove Community
  - Remove private ASN from AS-Path
  - …

Routing Filters

- **Example**
  - The filter below sets the `Local-Pref` attribute of the route to 100 and accept the route iff the `Communities` attribute contains 1 or if the destination prefix is 130.104/16 and the `Communities` attribute contains 2



Routing Filter Example

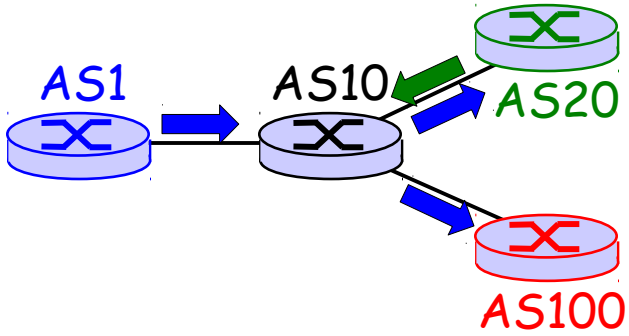
- with CISCO IOS ("route-maps")

```
# ip as-path access-list 1 permit ^1 2$
#
# route-map IN_SET_LOCAL_PREF 10 permit
#   match as-path 1
#   set local-preference 100
#
# router bgp 1
#   neighbor 10.0.0.1 remote-as 2
#   neighbor 10.0.0.1 route-map IN_SET_LOCAL_PREF in
#   network 150.0.0.0 mask 255.255.255.0
```

This route-map will only allow routes with an AS-PATH equal to [1 2] and set their LOCAL_PREF to 100

  - note : the semantics of the CISCO route-maps slightly differs from the one used in the lecture. If a route is not matched by any match statement, the route is rejected

No Routing Filter



Inside AS10's BGP Router

AS 1

Dst: **1.0/16**
AS-path: **1**

AS 20

Dst: **1.0/16**
AS-path: **20 1**

AS 100

no route

RIB-in

RIB-in

RIB-in

BGP decision process

Loc-RIB

Dst: **1.0/16**
AS-path: **1**

RIB-out

RIB-out

RIB-out

nothing sent (same source)

Dst: **1.0/16**
AS-path: **10 1**

Dst: **1.0/16**
AS-path: **10 1**

AS 1

AS 20

AS 100

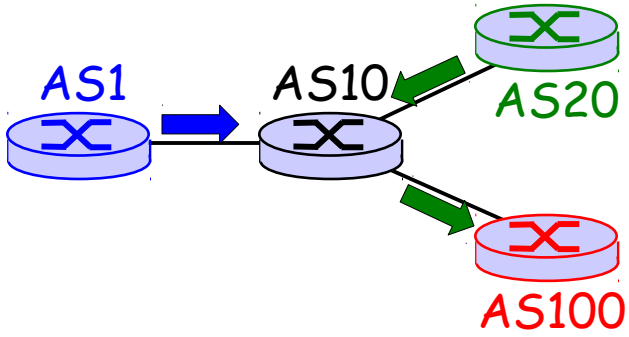
control plane

data plane

used for forwarding

FIB

Inbound Routing Filter



set local-preference 200
(default is 100)

Inside AS10's BGP Router

AS 1

Dst: **1.0/16**
AS-path: **1**

AS 20

Dst: **1.0/16**
AS-path: **20 1**

AS 100

no route

RIB-in

in-filter
RIB-in

RIB-in

BGP decision process

Loc-RIB

Dst: **1.0/16**
AS-path: **20 1**
Local-Pref: **200**

RIB-out

RIB-out

RIB-out

nothing sent (loop)

nothing sent (same source)

Dst: **1.0/16**
AS-path: **10 20 1**

AS 1

AS 20

AS 100

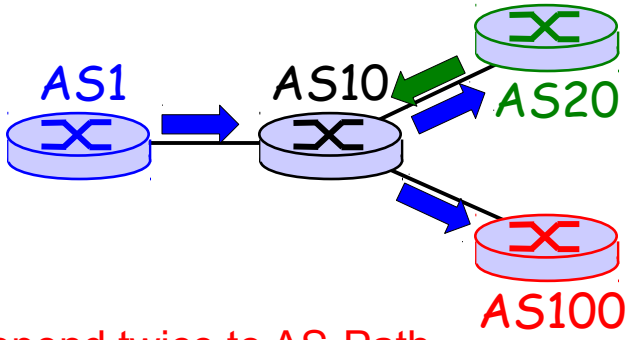
control plane

data plane

used for forwarding

FIB

Outbound Routing Filter



prepend twice to AS-Path

Inside AS10's BGP Router

AS 1

Dst: **1.0/16**
AS-path: **1**

AS 20

Dst: **1.0/16**
AS-path: **20 1**

AS 100

no route

RIB-in

RIB-in

RIB-in

BGP decision process

Loc-RIB

Dst: **1.0/16**
AS-path: **1**

RIB-out

RIB-out
out-filter

RIB-out

Dst: **1.0/16**
AS-path: **10 10 10 1**

Dst: **1.0/16**
AS-path: **10 1**

AS 1

AS 100

control plane

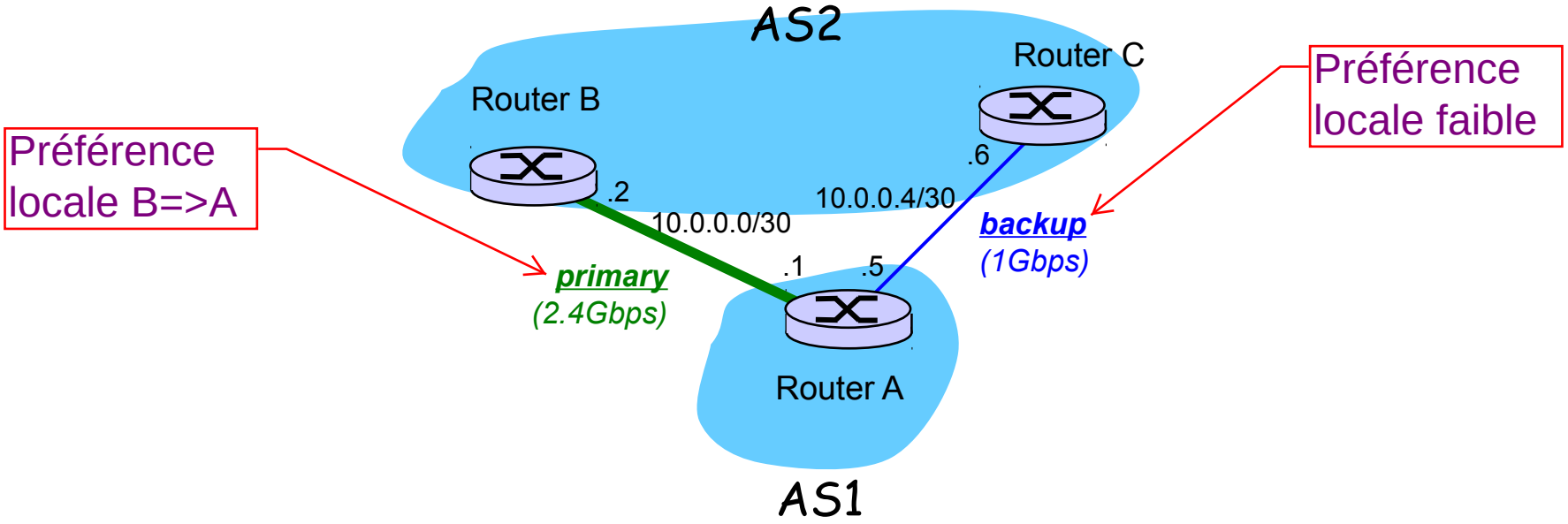
data plane

used for forwarding

FIB

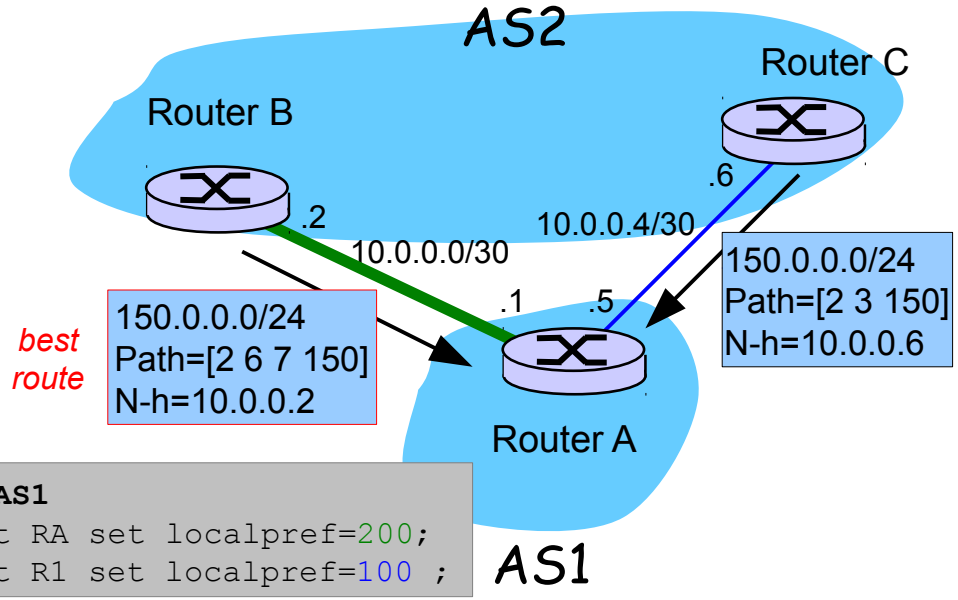
Application

- **Backup route**
  - It is frequent that an AS is connected through multiple links to its upstream provider
  - In this case, it is frequent to have a primary peering link and a backup peering link



Application

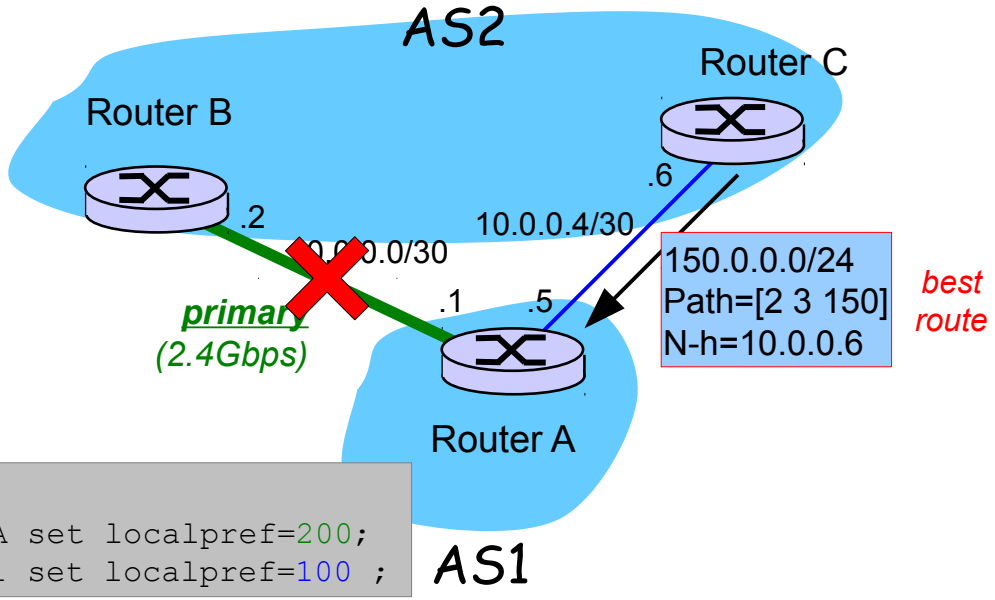
- **Backup route**
  - When primary link is up, traffic should go out through AS2. An import filter is configured on router A to set a higher preference on routes received from router B




```
RPSL-like policy for AS1
import: from AS2 RB at RA set localpref=200;
        from AS2 RC at R1 set localpref=100 ;
```

Application

- **Backup route**
  - When primary link fails, the route from AS2 is removed from the RIB-in of router A and the decision process falls back on router C's routes.




```
RPSL-like policy for AS1
import: from AS2 RB at RA set localpref=200;
        from AS2 RC at R1 set localpref=100 ;
```

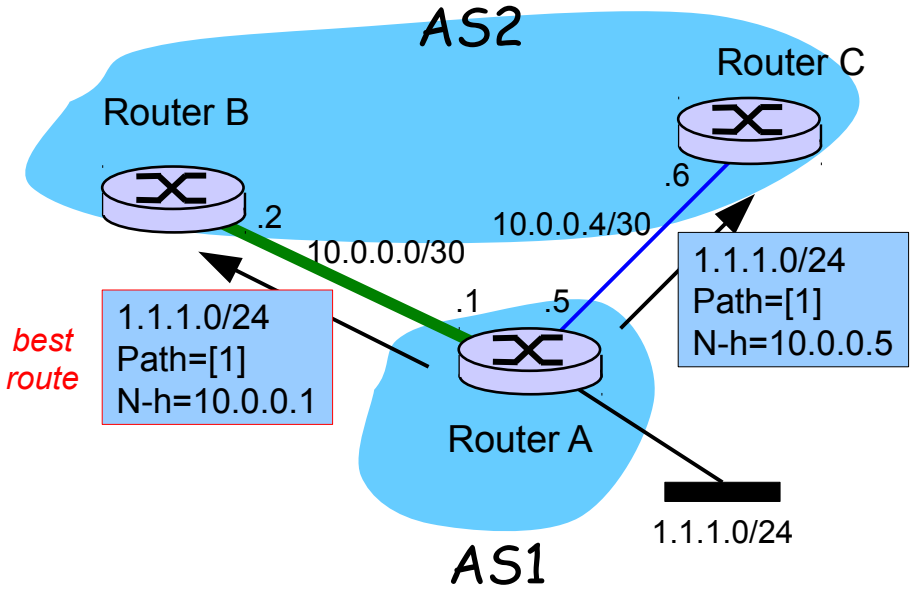
Application

- **Backup route**
  - Note that there might be multiple reasons in the previous example for a route to be removed from the RIB-in
    - BGP session could be broken (remote router is down or unreachable).
    - BGP next-hop could become unreachable. Remember the decision process first checks that the next-hop is reachable before considering a route is eligible.
    - An upstream router might fail, causing all the routes it had announced to be withdrawn downstream

Application

- **Backup route**
  - The provider will also configure filters to prefer routes received from the customer through the primary link

```
RPSL-like policy for AS2
import: from AS1 RA at RB set localpref=200;
        from AS1 RA at RC set localpref=100 ;
```



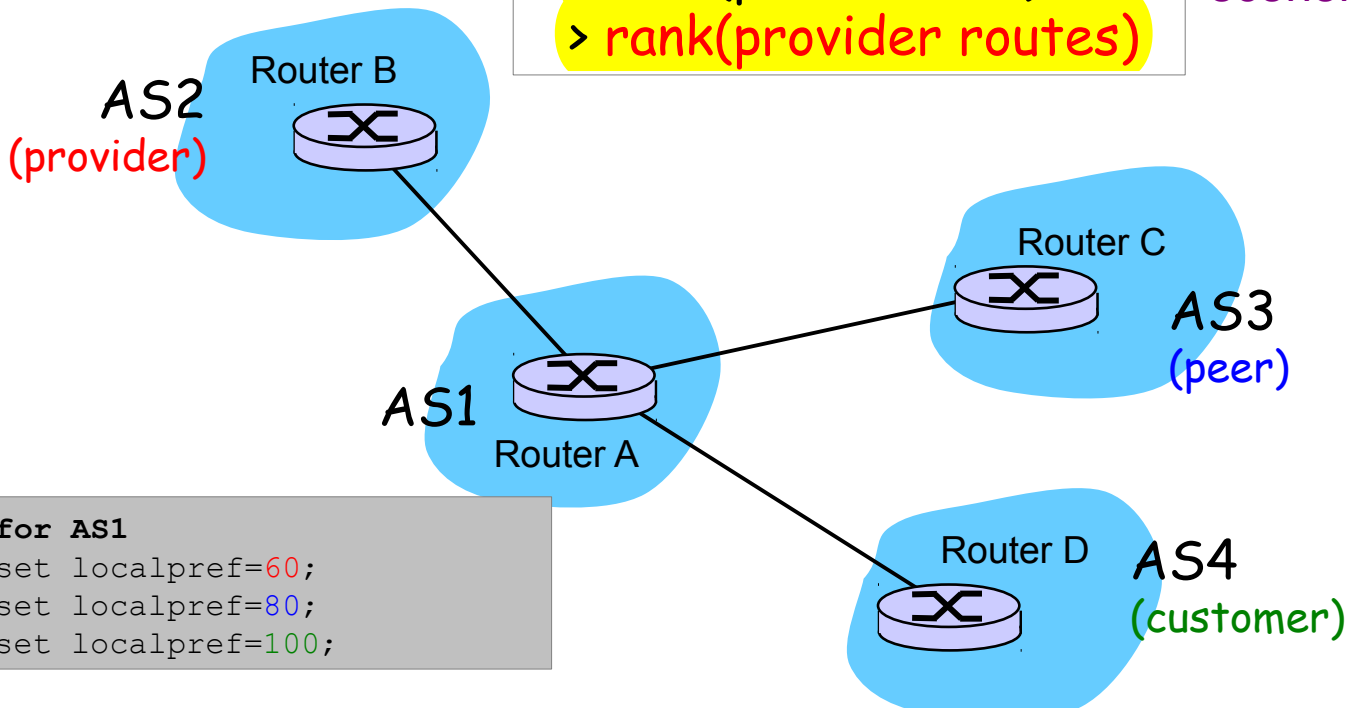
Application

Business Relationships

- The `Local-Pref` attribute is often set based on the economical relationships among ASs. `Local-Pref` values are assigned such that

rank(customer routes)
> rank(peer routes)
> rank(provider routes)

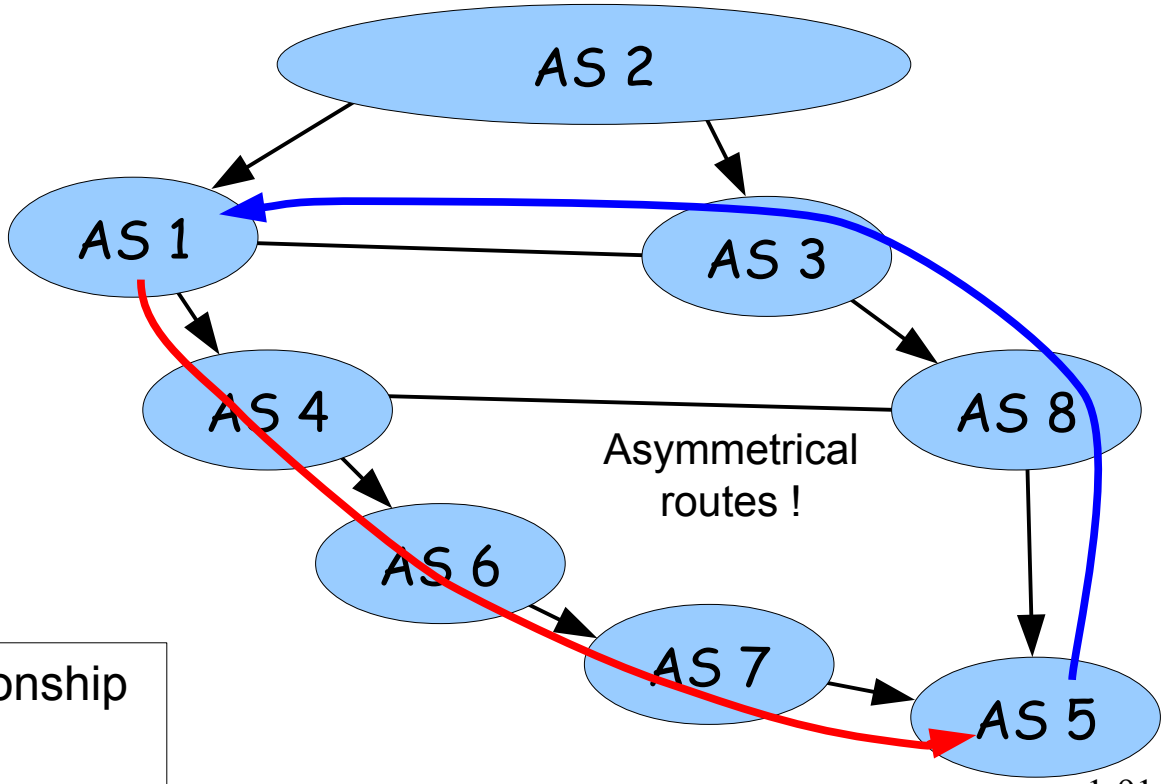
Contraintes économiques

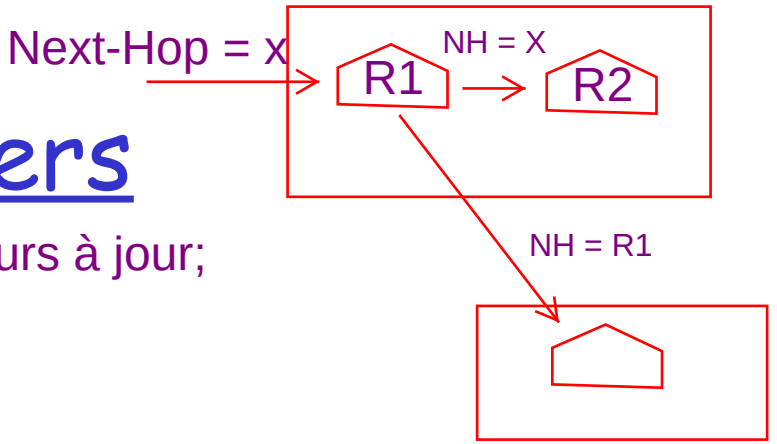



```
RPSL-like policy for AS1
import: from AS2 set localpref=60;
        from AS3 set localpref=80;
        from AS4 set localpref=100;
```

Application

- **Business Relationships**
  - Which route will be used by AS1 to reach AS5 ? and in the reverse direction ?





Protocol Routing Filters

En interne, le Next-Hop ne se met pas toujours à jour;

En externe oui.

Principle

- Apart from the configurable import and export filters, a BGP router will automatically filter routes as follows
  - **`Local-Pref`** : removed from a route before it is advertised to a BGP router in another AS.
  - **`AS-Path`** : prepended with the local ASN before the route is sent to a BGP router in another AS.
  - **`Next-Hop`** : updated before a route is sent to another BGP router.
  - a route with an `AS-Path` that contains the ASN of a neighbor will not be advertised to that neighbor (*sender-side loop detection*).

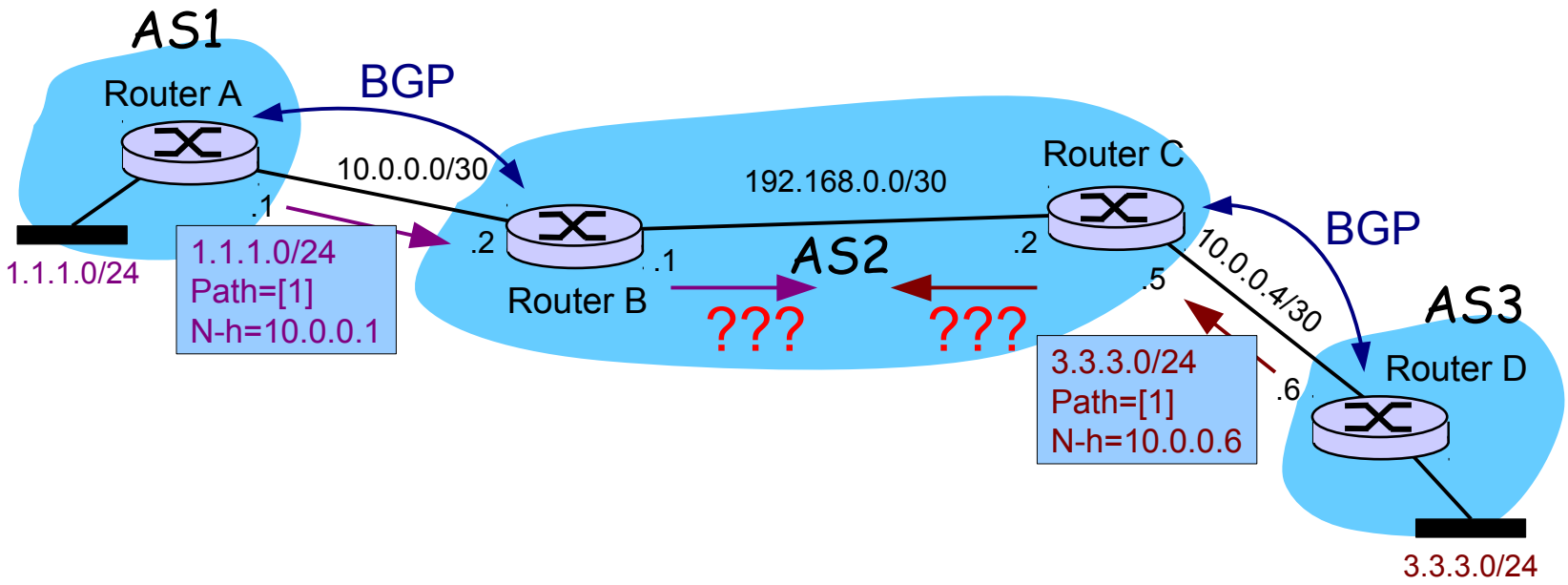
Chapter 2: roadmap

- 2.1 Inter-domain Routing
- 2.2 The Border Gateway Protocol (BGP)
  - 2.2.1 Principles
  - 2.2.2 Sessions
  - 2.2.3 Routes
  - 2.2.4 Path Attributes
  - 2.2.5 Messages
  - 2.2.6 Finite State Machine
  - 2.2.7 Decision Process
  - 2.2.8 Routing Filters
  - 2.2.9 Internal BGP (iBGP)
- ...

BGP inside an AS

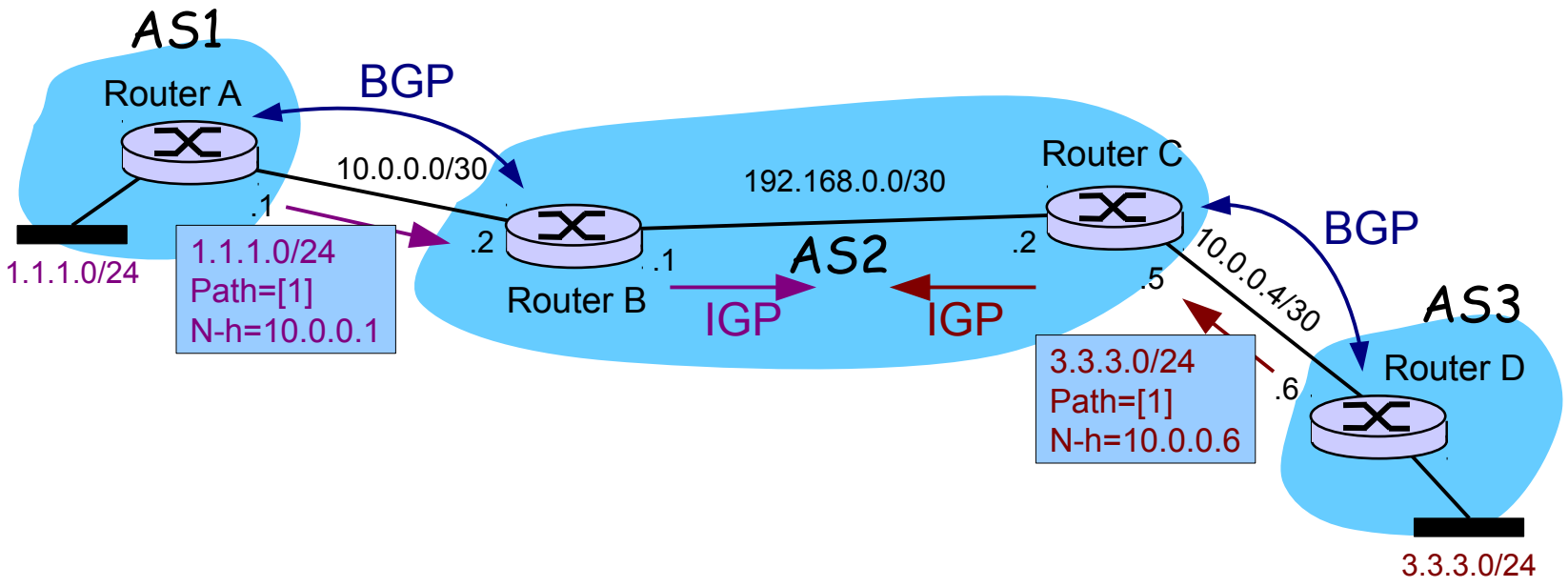
Comment propager le
flux à l'intérieur d'un AS ?

- **Problem**
  - How to advertise BGP routes from one router to another router within the same AS ?



BGP inside an AS

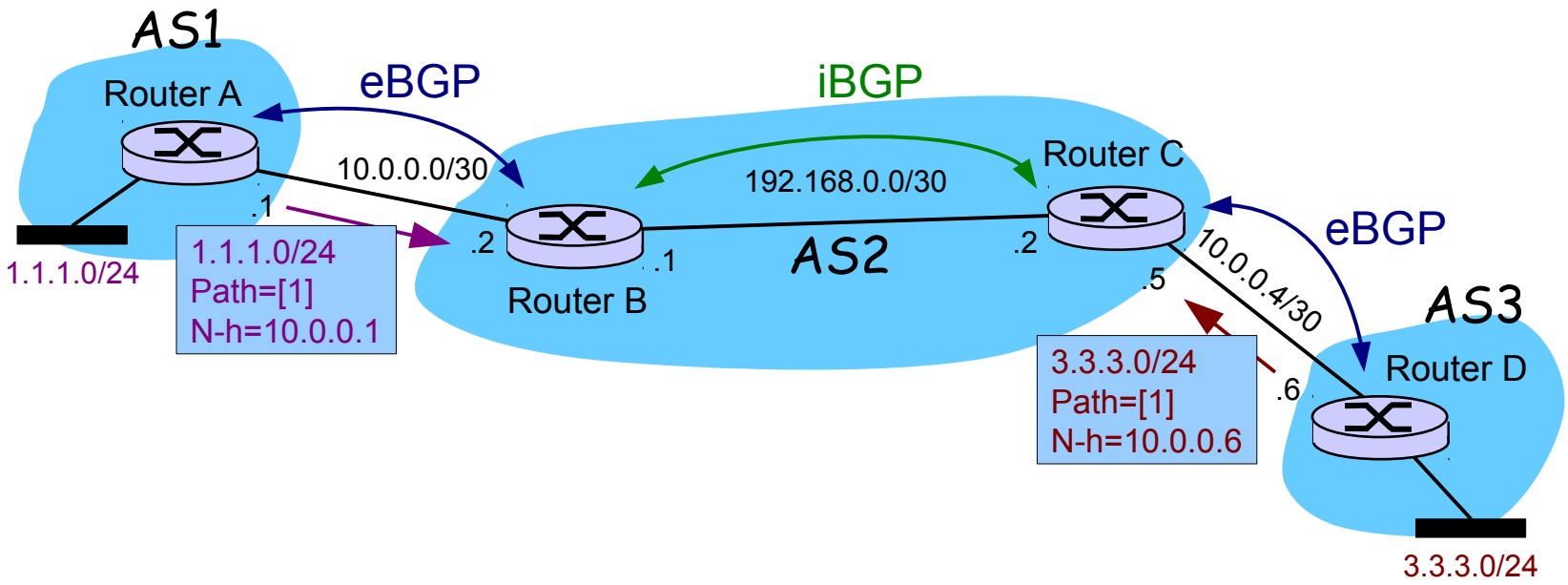
- **First Solution: use IGP**(1)
  - Drawback: IGP might not be able to support so many routes
  - Drawback: IGP does not carry BGP attributes (`AS-Path`, ...)



(1) IGP : *Interior Gateway Protocol*, routing protocol used within a domain, e.g. OSPF, RIP, IS-IS

BGP inside an AS

- **Solution: use BGP**
  - Two different types of BGP sessions: internal (iBGP) and external (eBGP)



iBGP versus eBGP

Règles de routage

Differences between iBGP and eBGP

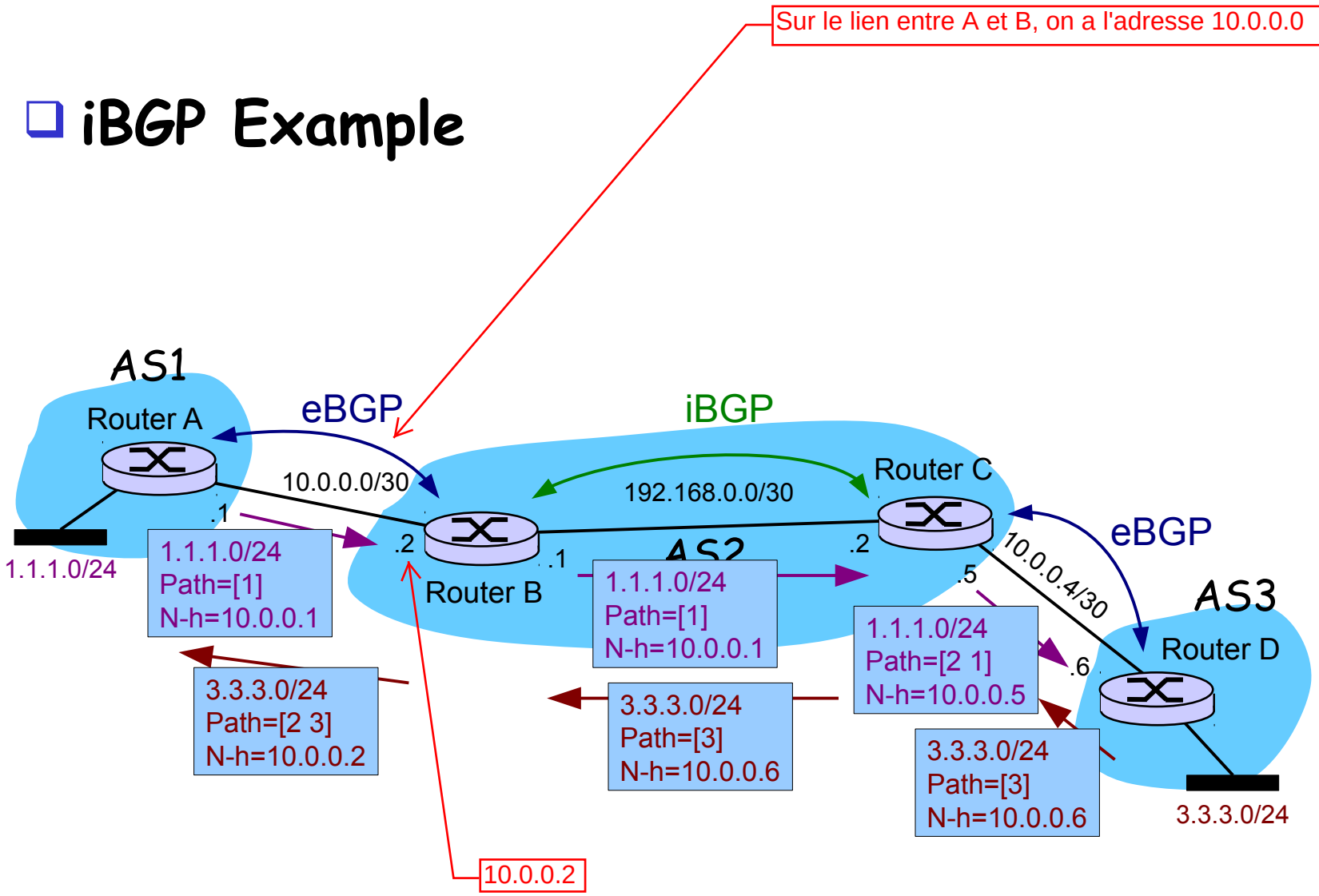
- The `Local-Pref` attribute is only carried over iBGP sessions
- Usually, import and export filters are only defined for eBGP sessions.
- /!\ Routes learned over iBGP sessions cannot be redistributed over another iBGP session.
  - Otherwise, routes could loop forever over the iBGP sessions.
- The `Next-Hop` attribute of a route is usually only updated when the route is being sent over an eBGP session (see also `next-hop-self`).
- The `AS-Path` attribute of a route is only prepended with the local ASN before the route is sent over an eBGP session.

Exemple : session de la veille

Problème de convergence de routage !
Vu que l'AS-Path n'et pas mis à jour avec la propagation de BGP...
Comment déterminer le chemin à utiliser ?

BGP inside an AS

- iBGP Example



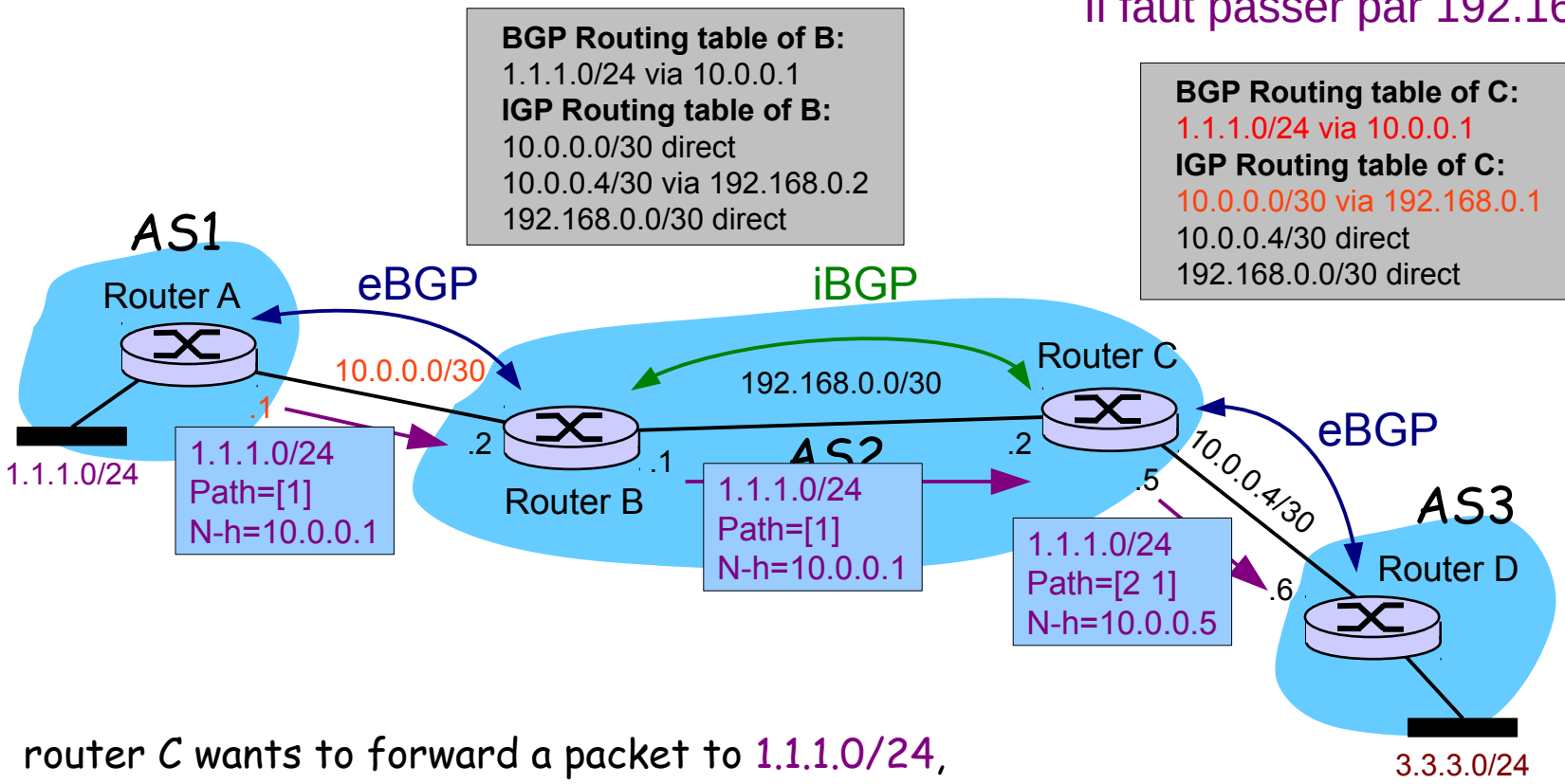
BGP inside an AS

Grâce à la table de routage

- Packet **forwarding**

Pour joindre 10.0.0.0/30,

Il faut passer par 192.168.0.1



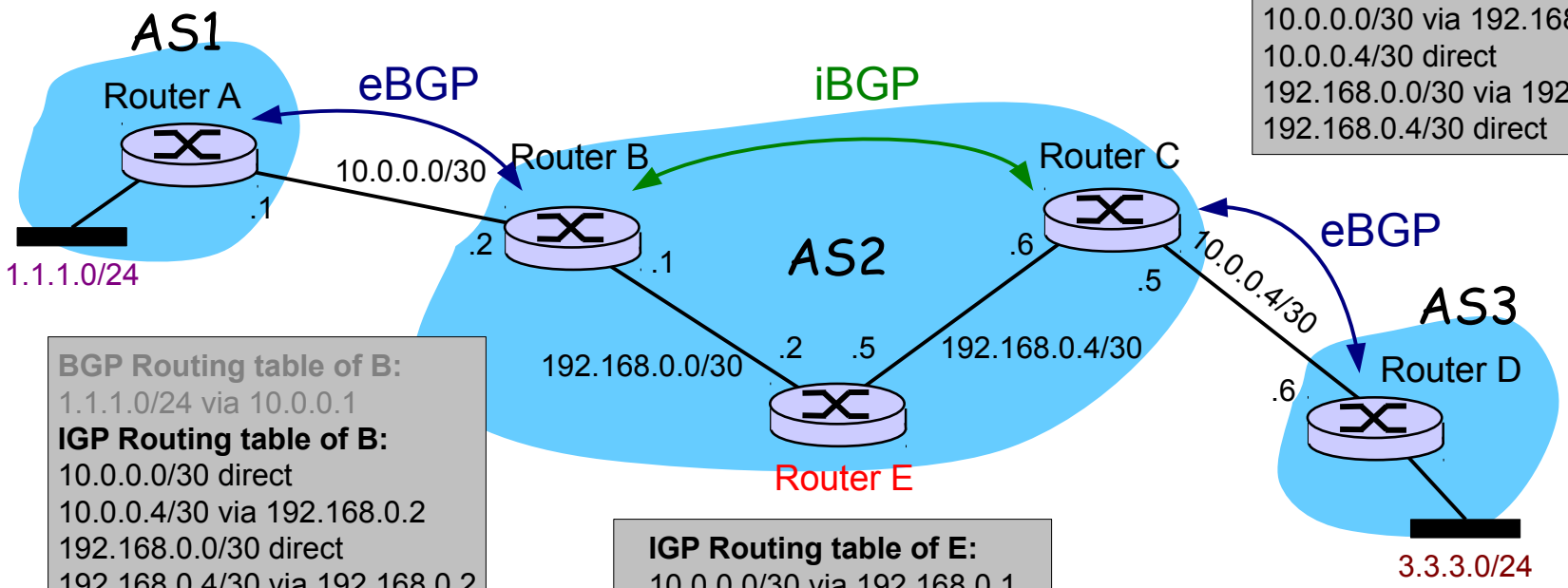
If router C wants to forward a packet to 1.1.1.0/24,
it needs to rely on the BGP route towards 1.1.1.0/24
AND on the IGP route towards 10.0.0.1, the BGP route's next-hop

BGP inside an AS

Est-ce que la session BGP entre B et C est-elle possible ? Oui, un protocole IGP interdomaine utilisant IP

Dealing with non-BGP routers

- What happens if there are internal, backbone routers between BGP routers inside an AS ?
  - iBGP session (TCP connection) : OK



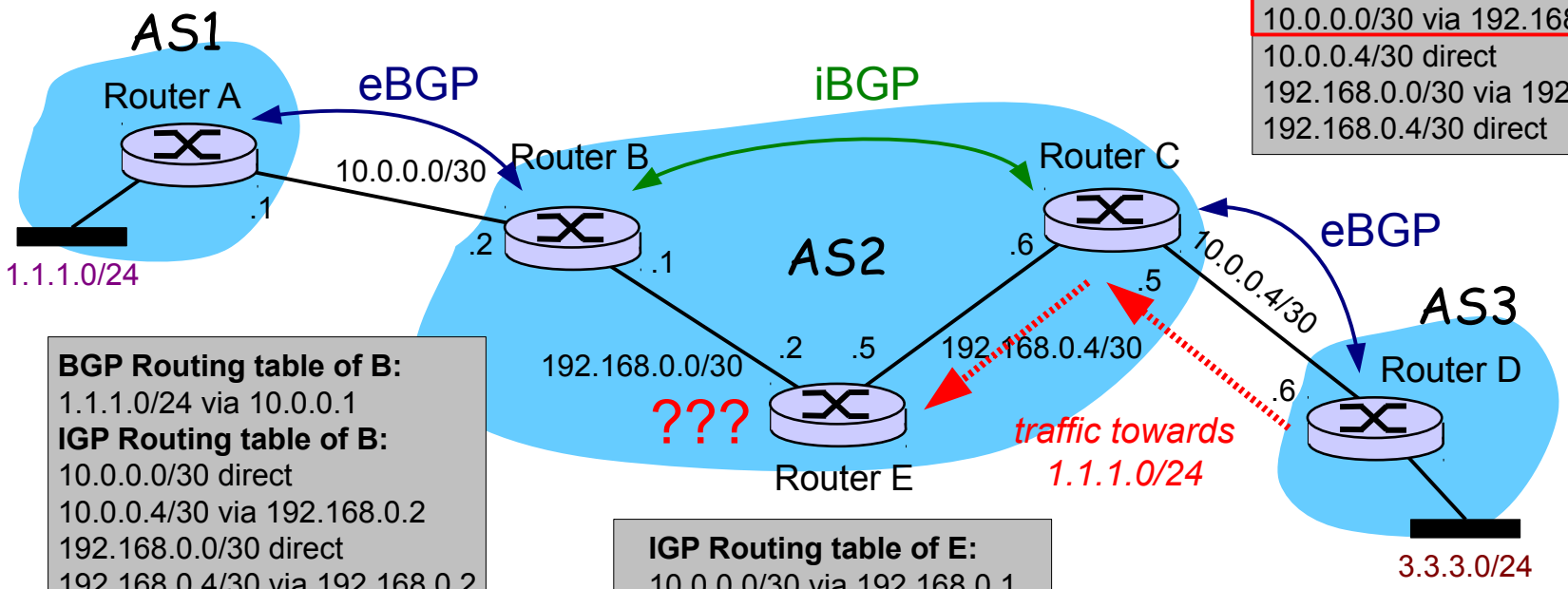
BGP Routing table of C:
1.1.1.0/24 via 10.0.0.1
IGP Routing table of C:
10.0.0.0/30 via 192.168.0.5
10.0.0.4/30 direct
192.168.0.0/30 via 192.168.0.5
192.168.0.4/30 direct

BGP Routing table of B:
1.1.1.0/24 via 10.0.0.1
IGP Routing table of B:
10.0.0.0/30 direct
10.0.0.4/30 via 192.168.0.2
192.168.0.0/30 direct
192.168.0.4/30 via 192.168.0.2

IGP Routing table of E:
10.0.0.0/30 via 192.168.0.1
10.0.0.4/30 via 192.168.0.6
192.168.0.0/30 direct
192.168.0.4/30 direct

BGP inside an AS

- **Dealing with non-BGP routers**
  - What happens if there are internal, backbone routers between BGP routers inside an AS ?
    - What about external routes ?



BGP Routing table of C:
1.1.1.0/24 via 10.0.0.1
IGP Routing table of C:
10.0.0.0/30 via 192.168.0.5
10.0.0.4/30 direct
192.168.0.0/30 via 192.168.0.5
192.168.0.4/30 direct

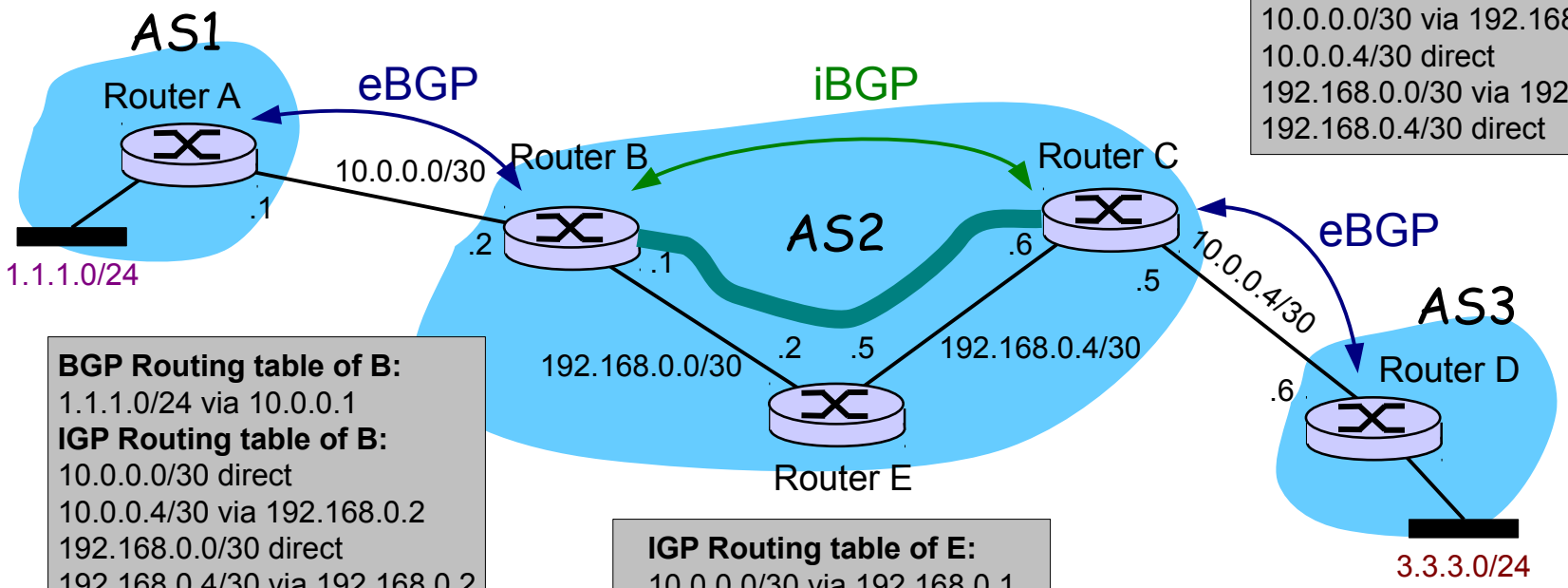
BGP Routing table of B:
1.1.1.0/24 via 10.0.0.1
IGP Routing table of B:
10.0.0.0/30 direct
10.0.0.4/30 via 192.168.0.2
192.168.0.0/30 direct
192.168.0.4/30 via 192.168.0.2

IGP Routing table of E:
10.0.0.0/30 via 192.168.0.1
10.0.0.4/30 via 192.168.0.6
192.168.0.0/30 direct
192.168.0.4/30 direct

BGP inside an AS

Envisageable, mais en pratique, ce n'est pas faisable (ici, il n'y a que 3 routeurs...)

- **Dealing with non-BGP routers**
  - First Solution: use tunnels. Traffic towards a BGP next-hop is encapsulated in a tunnel until egress (exit) router is reached.



BGP Routing table of C:
1.1.1.0/24 via 10.0.0.1
IGP Routing table of C:
10.0.0.0/30 via 192.168.0.5
10.0.0.4/30 direct
192.168.0.0/30 via 192.168.0.5
192.168.0.4/30 direct

BGP Routing table of B:
1.1.1.0/24 via 10.0.0.1
IGP Routing table of B:
10.0.0.0/30 direct
10.0.0.4/30 via 192.168.0.2
192.168.0.0/30 direct
192.168.0.4/30 via 192.168.0.2

IGP Routing table of E:
10.0.0.0/30 via 192.168.0.1
10.0.0.4/30 via 192.168.0.6
192.168.0.0/30 direct
192.168.0.4/30 direct

BGP Inside an AS

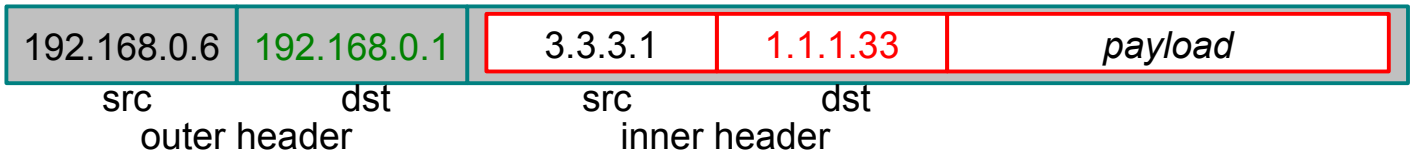
Si n est le nbre de routeurs, la config est de l'odre de n^2 (trop important)

Tunnel configuration example

- To setup a tunnel, a virtual network interface is created.
- The tunnel end-point must be specified: address of the router at the end of the tunnel.
- A static route is added to forward "some traffic" through the tunnel interface.
- with CISCO IOS, a GRE tunnel config. would look like

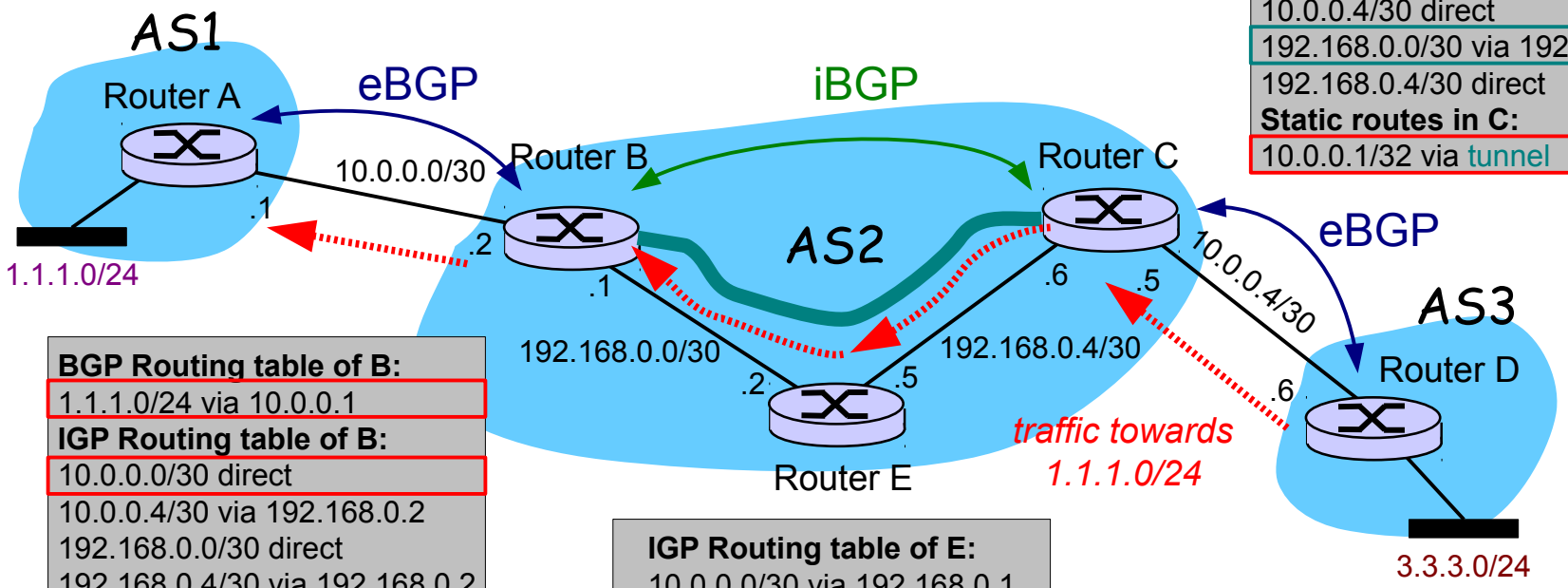
```
# interface Tunnel0
#   tunnel source Ethernet0
#   tunnel destination 192.168.0.1
#   tunnel mode gre ip
#
# ip route 10.0.0.1 255.255.255.255 Tunnel0
```

- an encapsulated packet would look like



BGP inside an AS

- **Dealing with non-BGP routers**
  - First Solution : use tunnels. Traffic towards a BGP next-hop is encapsulated in a tunnel until egress (exit) router is reached.



BGP Routing table of C:
1.1.1.0/24 via 10.0.0.1
IGP Routing table of C:
10.0.0.0/30 via 192.168.0.5
10.0.0.4/30 direct
192.168.0.0/30 via 192.168.0.5
192.168.0.4/30 direct
Static routes in C:
10.0.0.1/32 via tunnel

BGP Routing table of B:
1.1.1.0/24 via 10.0.0.1
IGP Routing table of B:
10.0.0.0/30 direct
10.0.0.4/30 via 192.168.0.2
192.168.0.0/30 direct
192.168.0.4/30 via 192.168.0.2

IGP Routing table of E:
10.0.0.0/30 via 192.168.0.1
10.0.0.4/30 via 192.168.0.6
192.168.0.0/30 direct
192.168.0.4/30 direct

Tunnels = OK

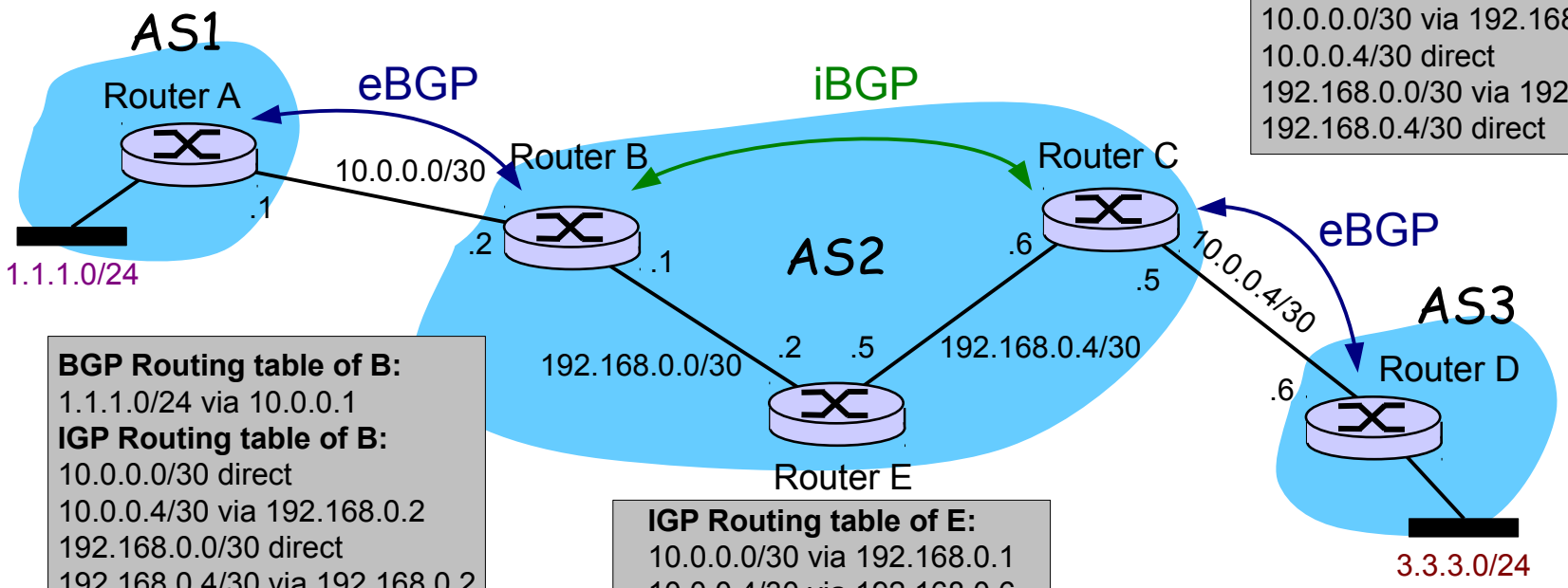
BGP inside an AS

/!\ Utiliser les tunnels diminuent le MTU => augmentation de fragmentation => overhead réseau

- **Issues with tunnels**
  - MTU is reduced due to outer header
  - Might increase processing load on tunnel head-end and tail-end due to encap-/decapsulation
    - modern routers are able to do this at line speed
  - Requires static route configuration (manual)
  - MPLS tunnels can be established dynamically but require an MPLS-enabled core
    - more on this in another chapter…

BGP inside an AS

- **Dealing with non-BGP routers**
  - Second solution: redistribute BGP route into IGP
  - Bad: scalability + danger to re-inject in BGP



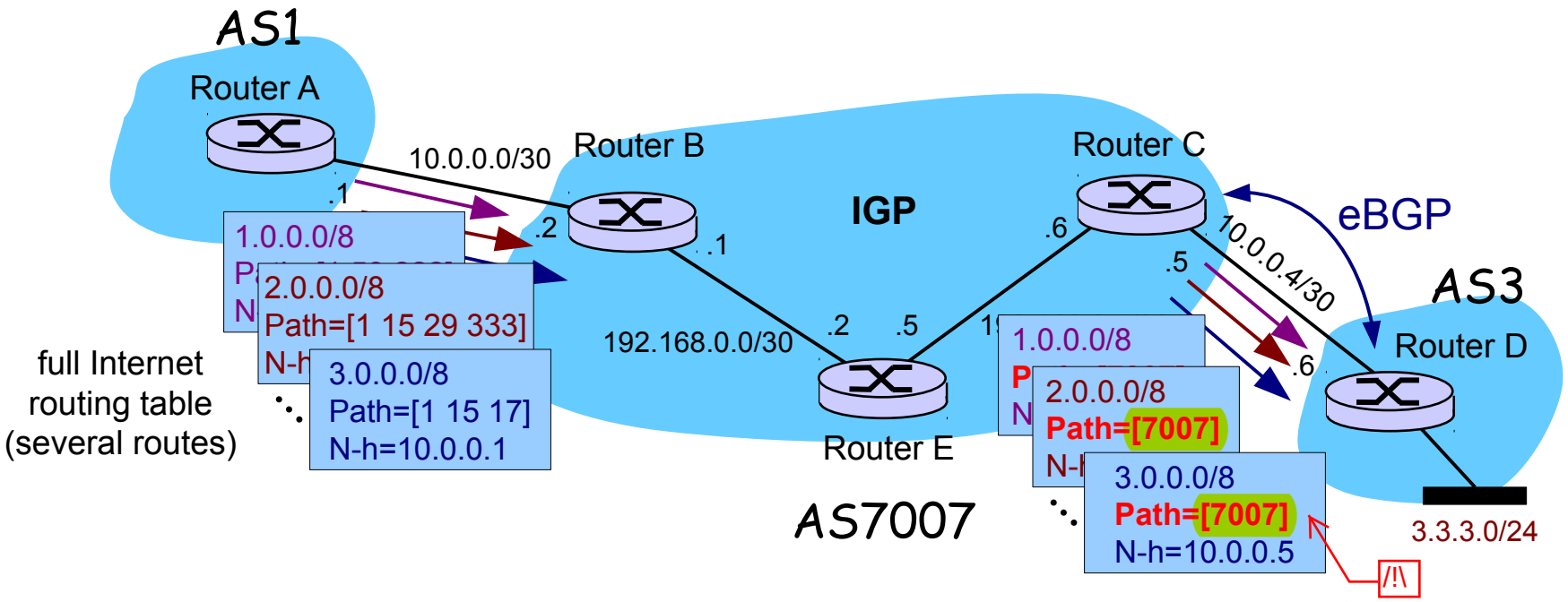
BGP Routing table of C:
1.1.1.0/24 via 10.0.0.1
IGP Routing table of C:
10.0.0.0/30 via 192.168.0.5
10.0.0.4/30 direct
192.168.0.0/30 via 192.168.0.5
192.168.0.4/30 direct

BGP Routing table of B:
1.1.1.0/24 via 10.0.0.1
IGP Routing table of B:
10.0.0.0/30 direct
10.0.0.4/30 via 192.168.0.2
192.168.0.0/30 direct
192.168.0.4/30 via 192.168.0.2

IGP Routing table of E:
10.0.0.0/30 via 192.168.0.1
10.0.0.4/30 via 192.168.0.6
192.168.0.0/30 direct
192.168.0.4/30 direct
1.1.1.0/24 via 192.168.0.1

BGP inside an AS

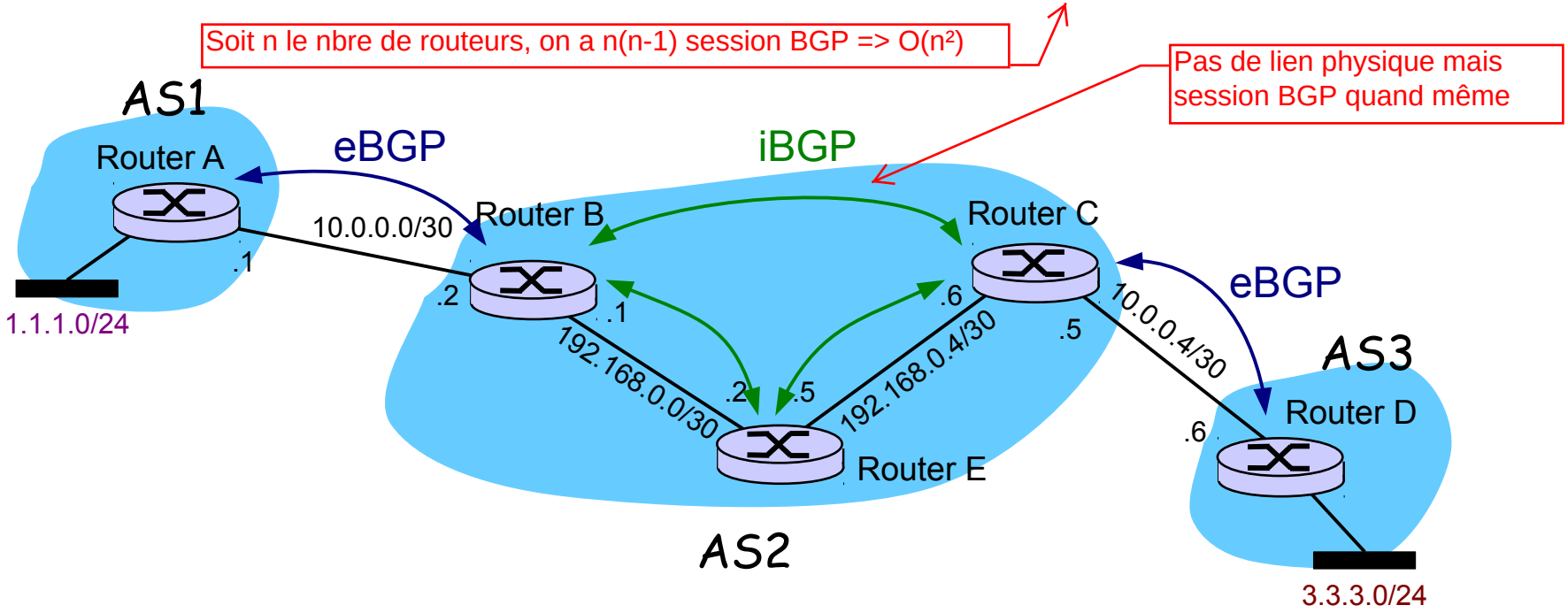
- **THe AS7007 incident**
  - On April 25, 1997, all routes received from an eBGP session are redistributed to the IGP and then back to BGP at the other side of the network... with a fresh AS-Path of length 1



BGP inside an AS

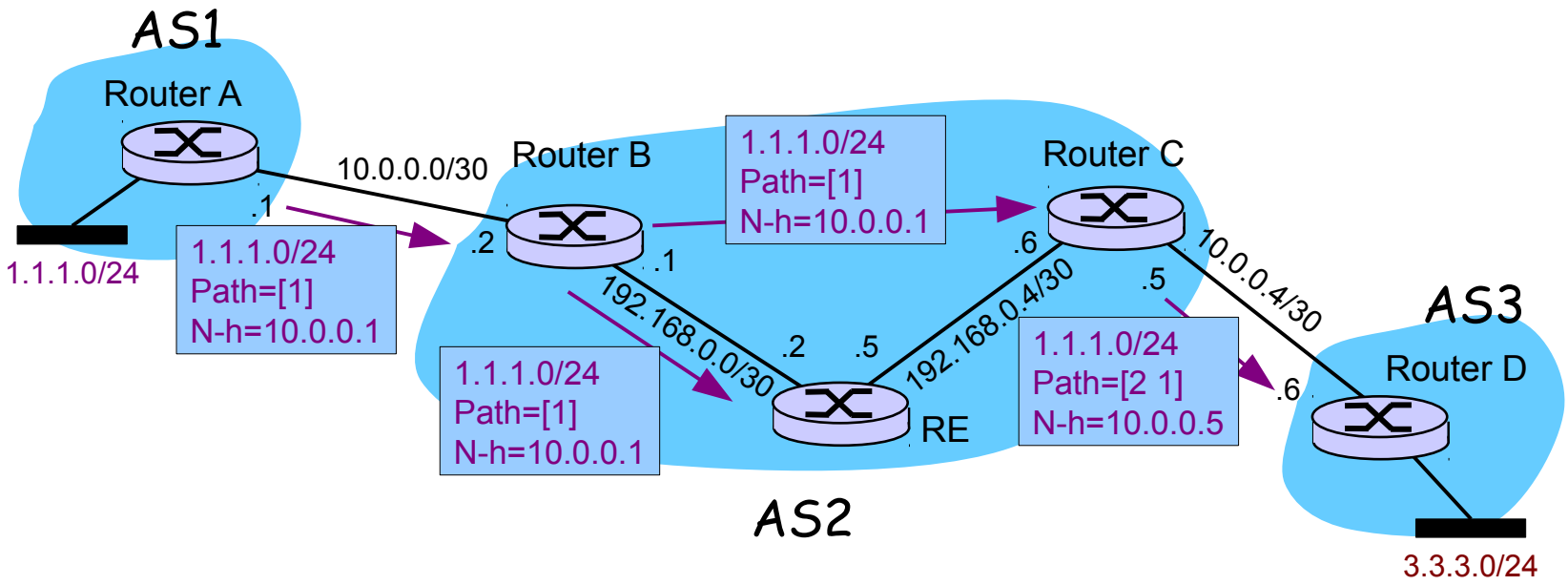
Avoir une session BGP sur l'ensemble des routeurs

- **Third solution: run BGP on all routers**
  - There must be an iBGP session between each pair of BGP routers inside the AS. This is called a *full-mesh* iBGP topology.



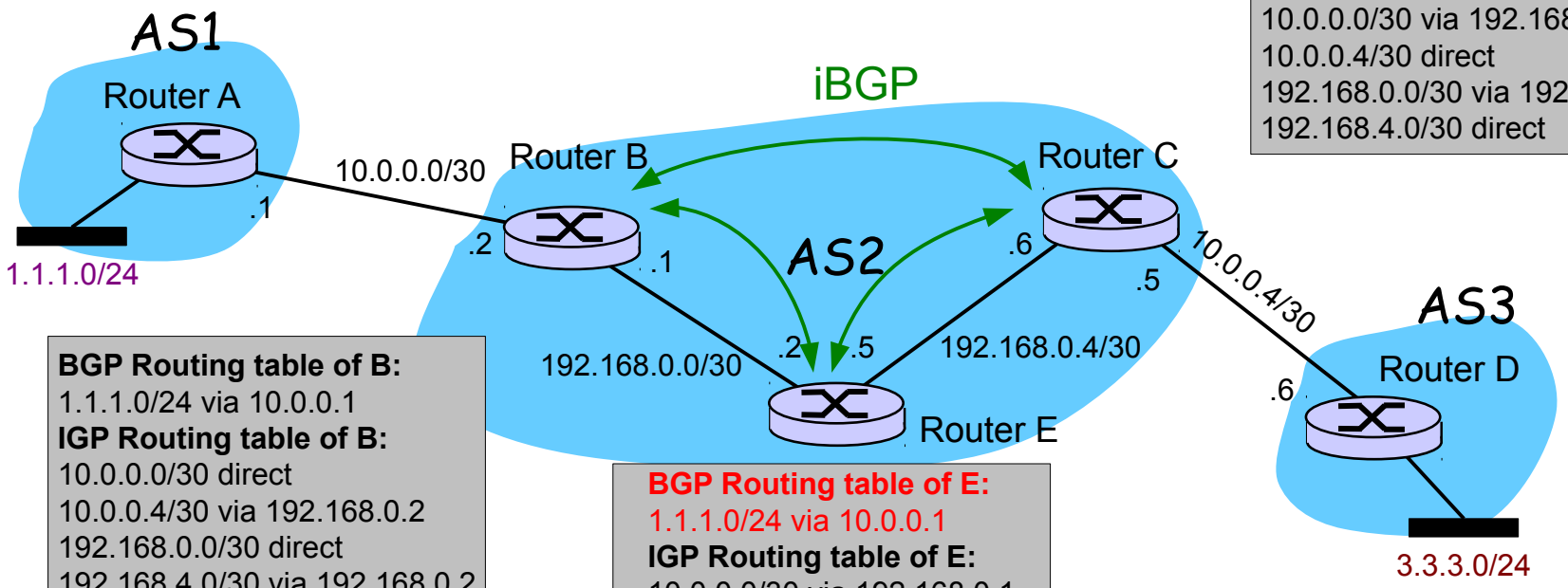
BGP inside an AS

- **Third solution: run BGP on all routers**
  - Rule to avoid iBGP routing loops: a route received over an iBGP session cannot be sent over another iBGP session !



BGP inside an AS

- **Third solution: run BGP on all routers**
  - Rule to avoid iBGP routing loops: a route received over an iBGP session cannot be sent over another iBGP

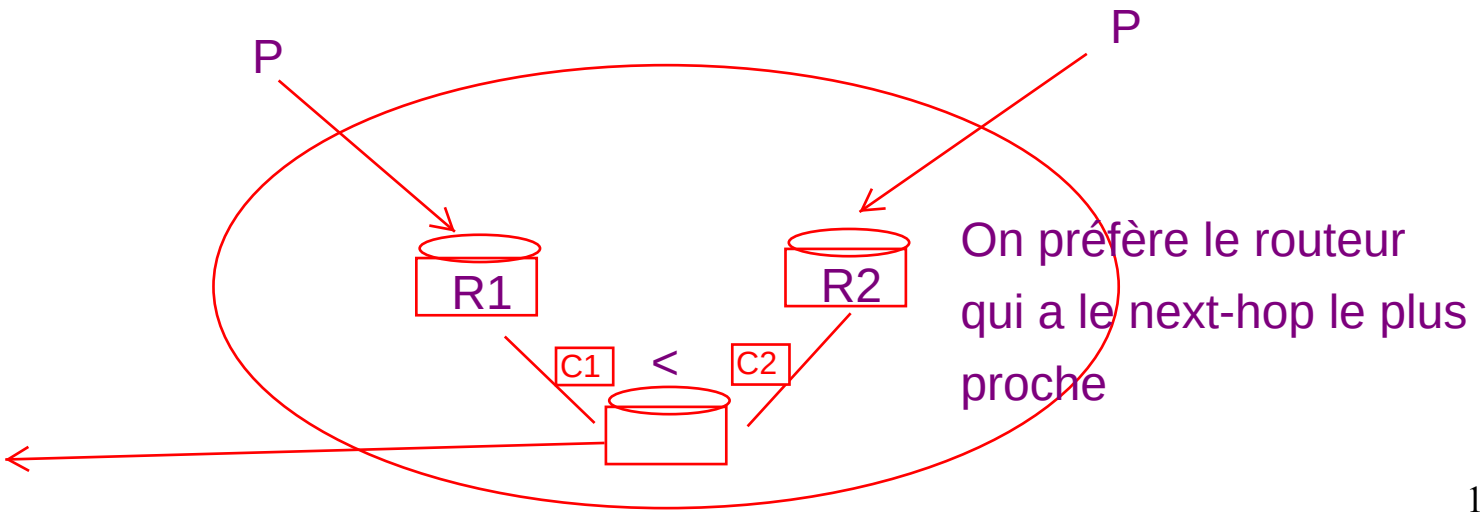


BGP inside an AS

- **Summary: IGP versus iBGP**
  - Role of IGP ← Propage que des routes internes et dépend du type de protocole de routage
    - distribute the internal topology and internal addresses within the AS
    - in previous example : 10.0.0.0/30, 10.0.0.4/30, 192.168.0.0/24 and 192.168.0.4/14
  - Role of iBGP ← Propage que des routes extérieures
    - distribute routes towards external destinations, learned via eBGP to all internal routers
    - in previous example : 1.1.1.0/24 and 3.3.3.0/24
    - iBGP sessions can be established thanks to IGP routes

BGP inside an AS

- **iBGP and decision process**
  - Once iBGP has been introduced, the decision process needs to be updated with additional rules.
    - A router can receive equally ranked routes from 2+ iBGP neighbors
    - A router can receive equally ranked routes from iBGP and eBGP neighbors



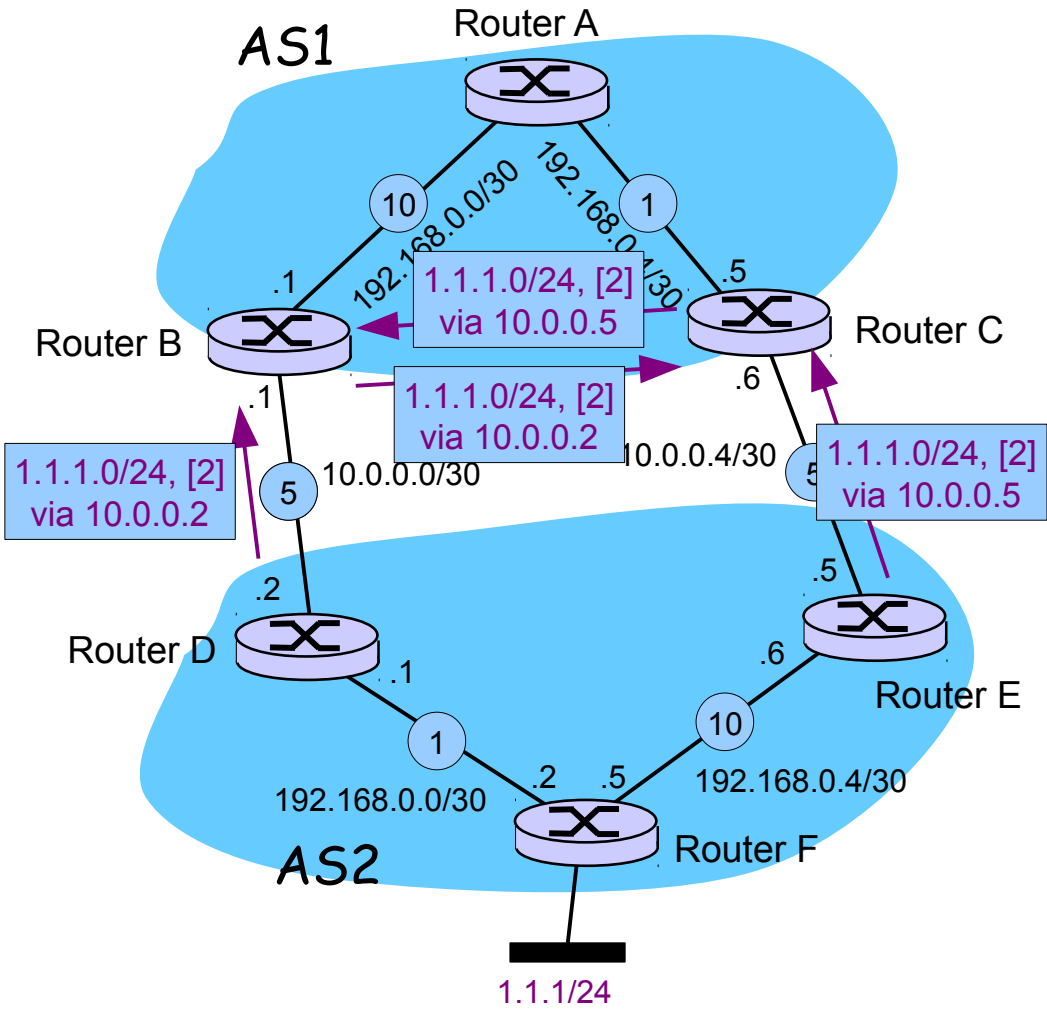
BGP inside an AS

1. Ignore if next-hop unreachable
2. Prefer locally originated networks
3. Prefer **highest** `LOCAL-PREF`
4. Prefer **shortest** `AS-PATH`
5. Prefer **lowest** `ORIGIN`
6. Prefer **lowest** `MED`
7. Prefer **eBGP over iBGP**
8. Prefer **nearest next-hop**
9. Prefer **lowest Router-ID** / `ORIGINATOR-ID`
10. Prefer **shortest** `CLUSTER-LIST`
11. Prefer **lowest neighbor address**

Decision Process

RIB-in of A:
1.1.1.0/24, [2] via 10.0.0.5 (eBGP)
1.1.1.0/24, [2] via 10.0.0.2 (iBGP)
Loc-RIB of A:
1.1.1.0/24 via 10.0.0.5

- **eBGP over iBGP**
  - Hot-potato routing
  - A router should try to get rid of packets sent to external domains as soon as possible



RIB-in of B:
1.1.1.0/24, [2] via 10.0.0.2 (eBGP)
1.1.1.0/24, [2] via 10.0.0.5 (iBGP)
Loc-RIB of B:
1.1.1.0/24 via 10.0.0.2

Decision Process

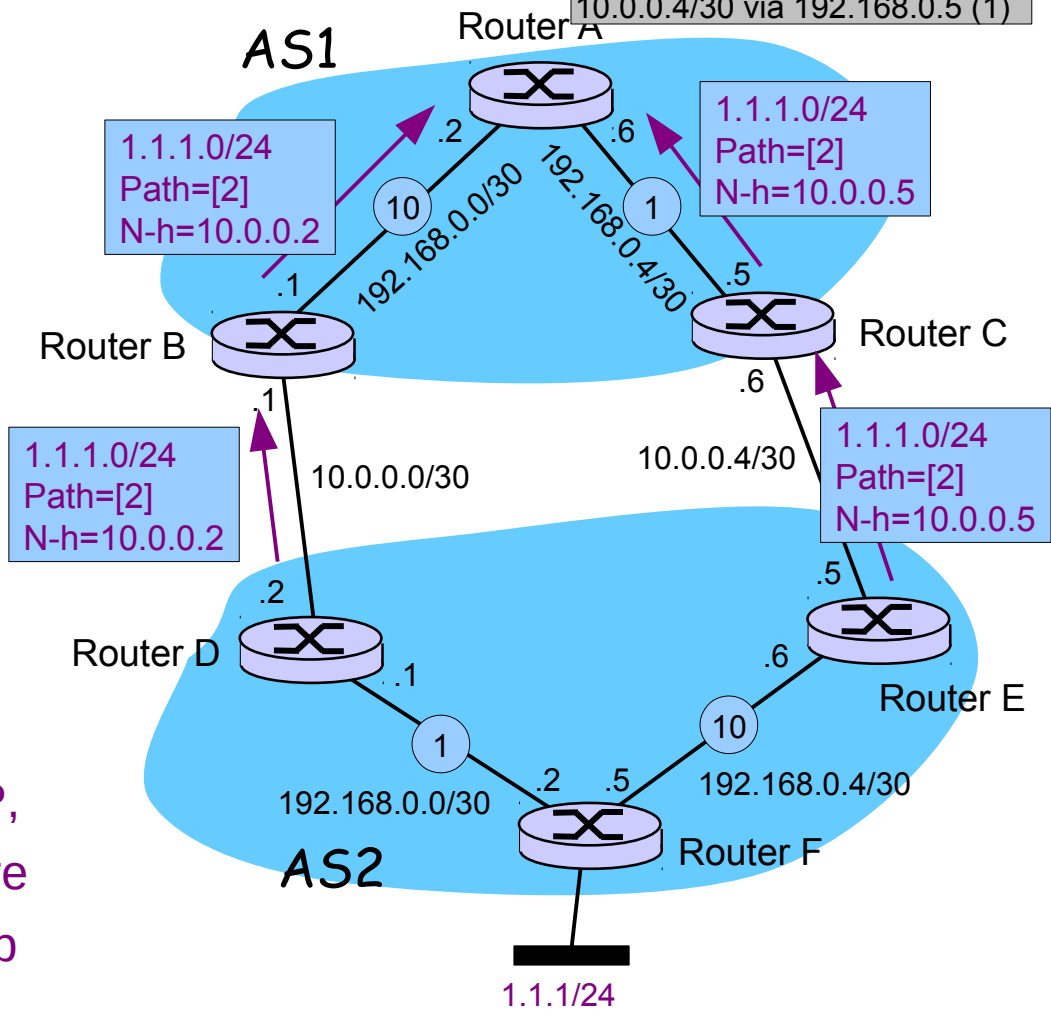
RIB-in of A:
1.1.1.0/24, Path=[2] via 10.0.0.5 (E)
1.1.1.0/24, Path=[2] via 10.0.0.2 (D)
Loc-RIB of A:
???

IGP Routing table of A:
192.168.0.0/30 direct (0)
192.168.0.4/30 direct (0)
10.0.0.0/30 via 192.168.0.1 (10)
10.0.0.4/30 via 192.168.0.5 (1)

- **nearest next-hop**
  - Router A receives two routes with the same `Local-Pref` value and the same `AS-Path` length. Need to break the ties !

    => Next-Hop

=> Comme les sessions sont eBGP,
On ne peut pas appliquer la 1ère solution => utilisons le Next-Hop



Decision Process

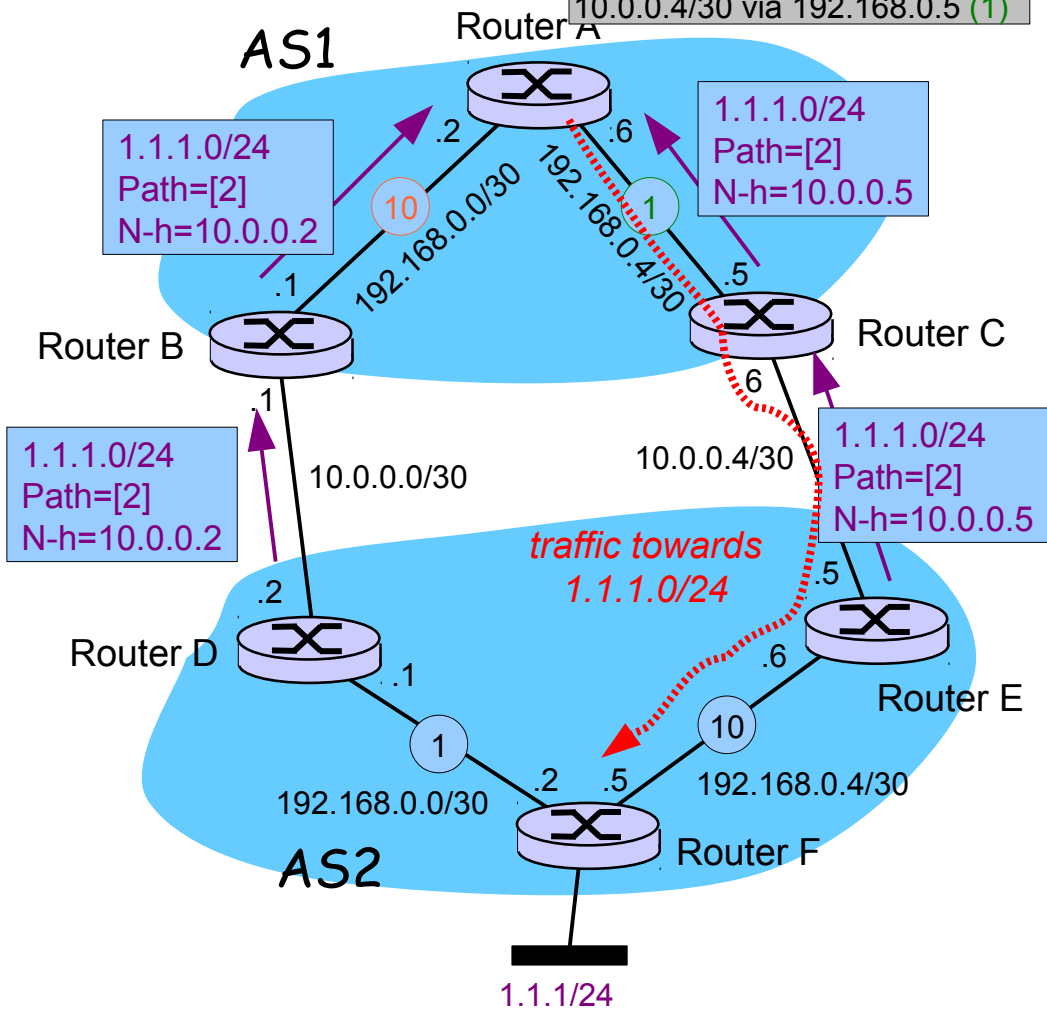
RIB-in of A:
1.1.1.0/24, Path=[2] via 10.0.0.5 (E)
1.1.1.0/24, Path=[2] via 10.0.0.2 (D)
Loc-RIB of A:
1.1.1.0/24 via 10.0.0.5

IGP Routing table of A:
192.168.0.0/30 direct (0)
192.168.0.4/30 direct (0)
10.0.0.0/30 via 192.168.0.1 (10)
10.0.0.4/30 via 192.168.0.5 (1)

- **nearest next-hop**
  - Hot-potato routing
  - Rely on the IGP cost to the BGP next-hops

cost(A → D) > cost(A → E)

  - Therefore prefer BGP route through E



BGP inside an AS

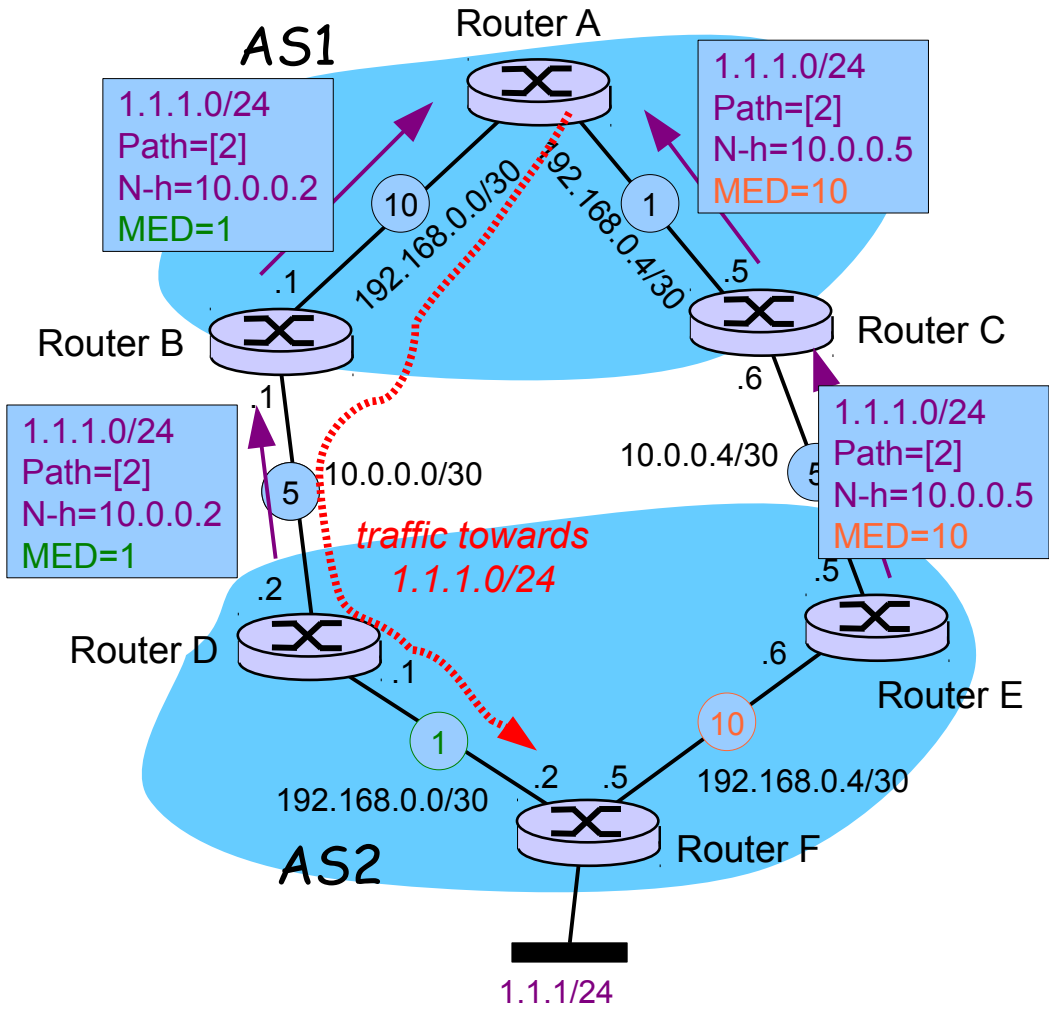
1. Ignore if next-hop unreachable
2. Prefer locally originated networks
3. Prefer **highest** `LOCAL-PREF`
4. Prefer **shortest** `AS-PATH`
5. Prefer **lowest** `ORIGIN`
6. Prefer **lowest** `MED`
7. Prefer **eBGP over iBGP**
8. Prefer **nearest next-hop**
9. Prefer **lowest Router-ID** / `ORIGINATOR-ID`
10. Prefer **shortest** `CLUSTER-LIST`
11. Prefer **lowest neighbor address**

Influence le processus de décision (multi-exit discrimination)
=> Permet de choisir le lien à prendre

Decision Process

RIB-in of A:
1.1.1.0/24, [2], MED=1 via 10.0.0.2
1.1.1.0/24, [2], MED=10 via 10.0.0.5
Loc-RIB of A:
1.1.1.0/24 via 10.0.0.2

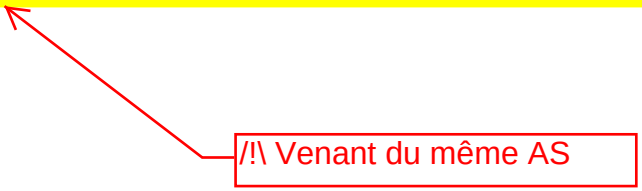
- **lowest MED**
  - Cold-potato routing
  - Tell your neighbor which ingress router to use…
  - Add MED attribute to routes announced
  - Neighbor AS will select routes with lowest MED value
  - An AS can ignore MED values from selected neighbors



The MED step

Inconvénient : On ne peut pas comparer les routes venant d'AS différents

- **Details**
  - The MED attribute is often set based on the IGP costs in the local AS
  - BUT the way IGP costs are assigned differs from one AS to another → the IGP costs of an AS are usually not comparable to that of another AS
  - As a consequence, MED values can only be compared between routes received from the same neighboring AS !!!



The MED step

- **Algorithm**

```
for R1 in all routes still under consideration
    for R2 in all routes still under consideration
        if (neighborAS(R1) == neighborAS(R2)) and
           (MED(R1) < MED(R2))
            remove route R2 from consideration
```

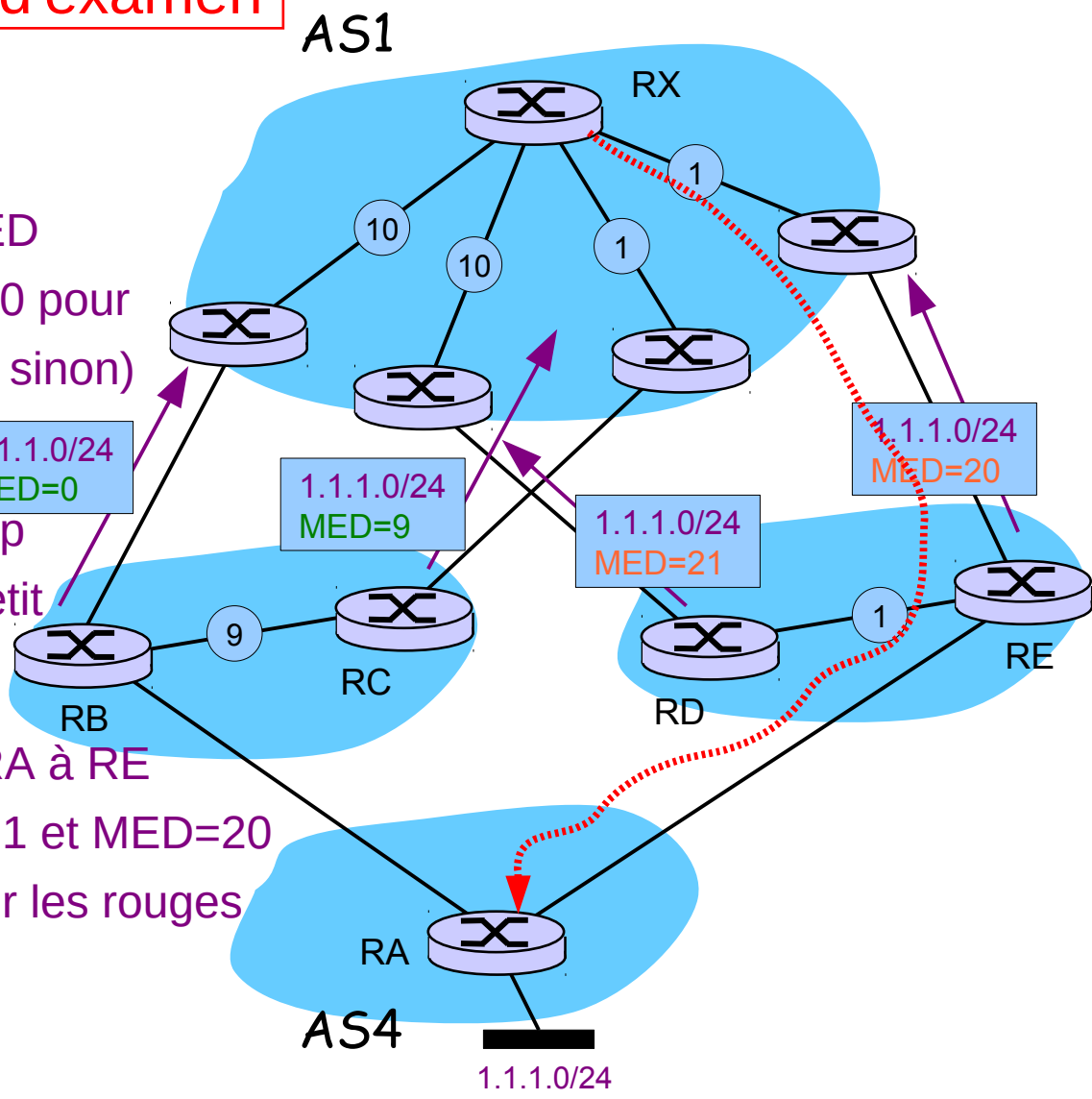
Question d'examen

The MED step

1ère étape : Comparaison des MED
(il va garder le MED=0 pour les verts et MED=20 sinon)

2ème étape : Comparer le next-hop et prendre le plus petit

Conclusion : Prendre la route de RA à RE
Car le Next-Hop vaut 1 et MED=20
A été sélectionné pour les rouges



BGP decision process

...
6. Prefer lowest MED
7. Prefer eBGP over iBGP
8. Prefer nearest next-hop
...

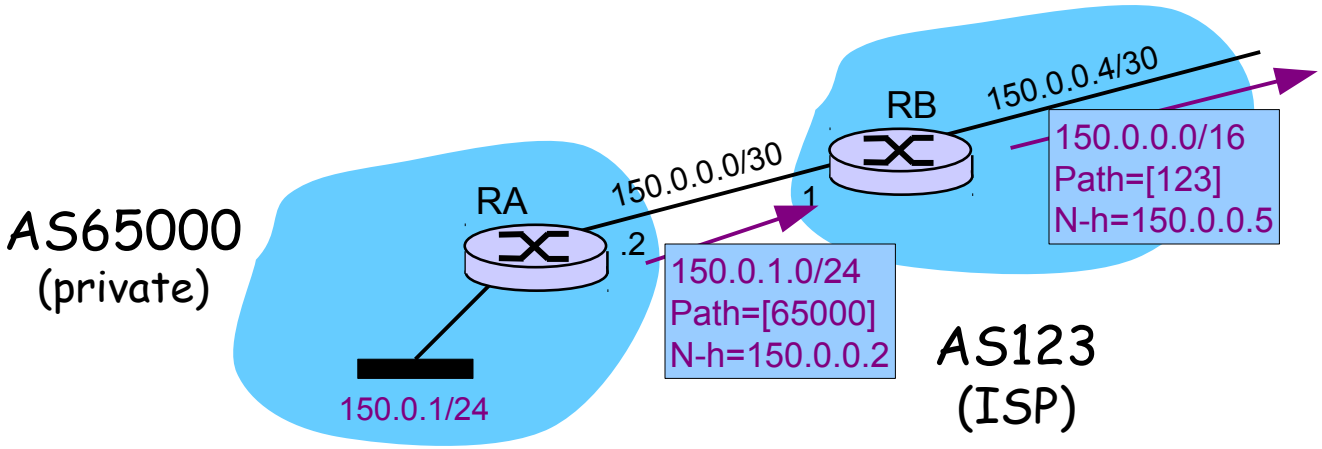
Chapter 2: roadmap

- 2.1 Inter-domain Routing
- 2.2 The Border Gateway Protocol (BGP)
- 2.3 BGP-based Traffic Engineering
  - Multi-homing
  - Selective Announcements
  - More Specific Prefixes
  - AS-Path Prepending
- 2.3 BGP Scalability
- 2.5 BGP Stability

Single-homed stub AS

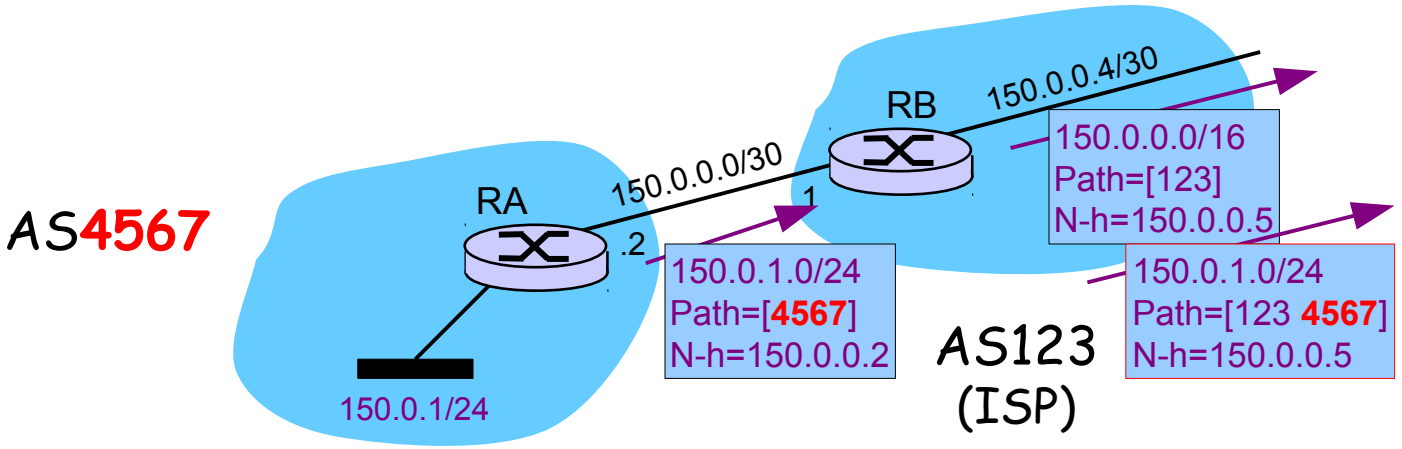
Le parent n'est pas obligé de propager la route 150.0.1.0/24 qui est plus spécifique.
Il va propager l'agrégat 150.0.0.0/16
Qui va englober l'adresse spécifique

- First days on the Internet…
  - Connection to a single provider
  - Uses private ASN (in range 64512-65535)
  - Stub AS completely hidden behind ISP



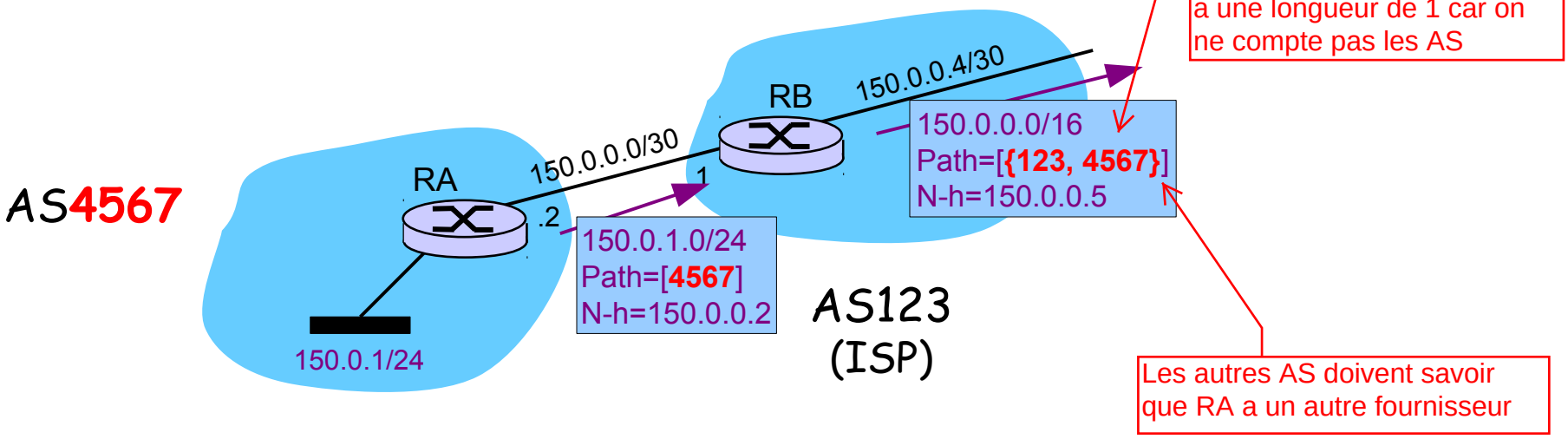
Single-homed stub AS

- Later...
  - Stub AS expects to **become** **multi-homed** and obtains public ASN.
  - ISP needs to advertise customer's prefix with real AS-Path → increases size of global routing tables.



Route Aggregation

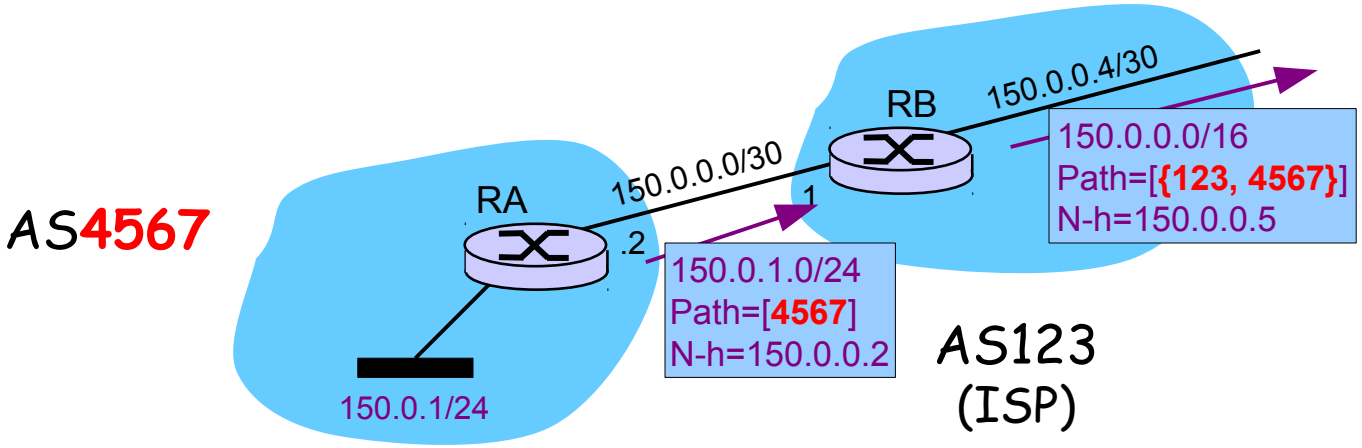
- **Principle**
  - BGP is able to **aggregate** received routes.
  - The AS-Path attribute is made of `AS_SEQUENCE` and `AS_SET` elements.
    - An `AS_SEQUENCE` is a sequence (ordered list) of ASNs.
    - An `AS_SET` is a set (unordered list) of ASNs(1).



(1) Note : an `AS_SET` is usually shown between curly braces { }

Route Aggregation

- **Details**
  - When considered by the decision process, the length of an `AS_SET` is 1 while the length of an `AS_SEQUENCE` is equal to the number of ASNs it contains.
  - Information about the real AS-Path is lost when `AS_SET` elements are used.



Si c'était agrégé correctement, il y aurait ce nbre de préfixe annoncé

AS try to aggregate as much as possible their prefixes. But it is not always possible or desirable...

Nbre de préfixe différents annoncés

Différence entre les 2 premiers => gain de performance

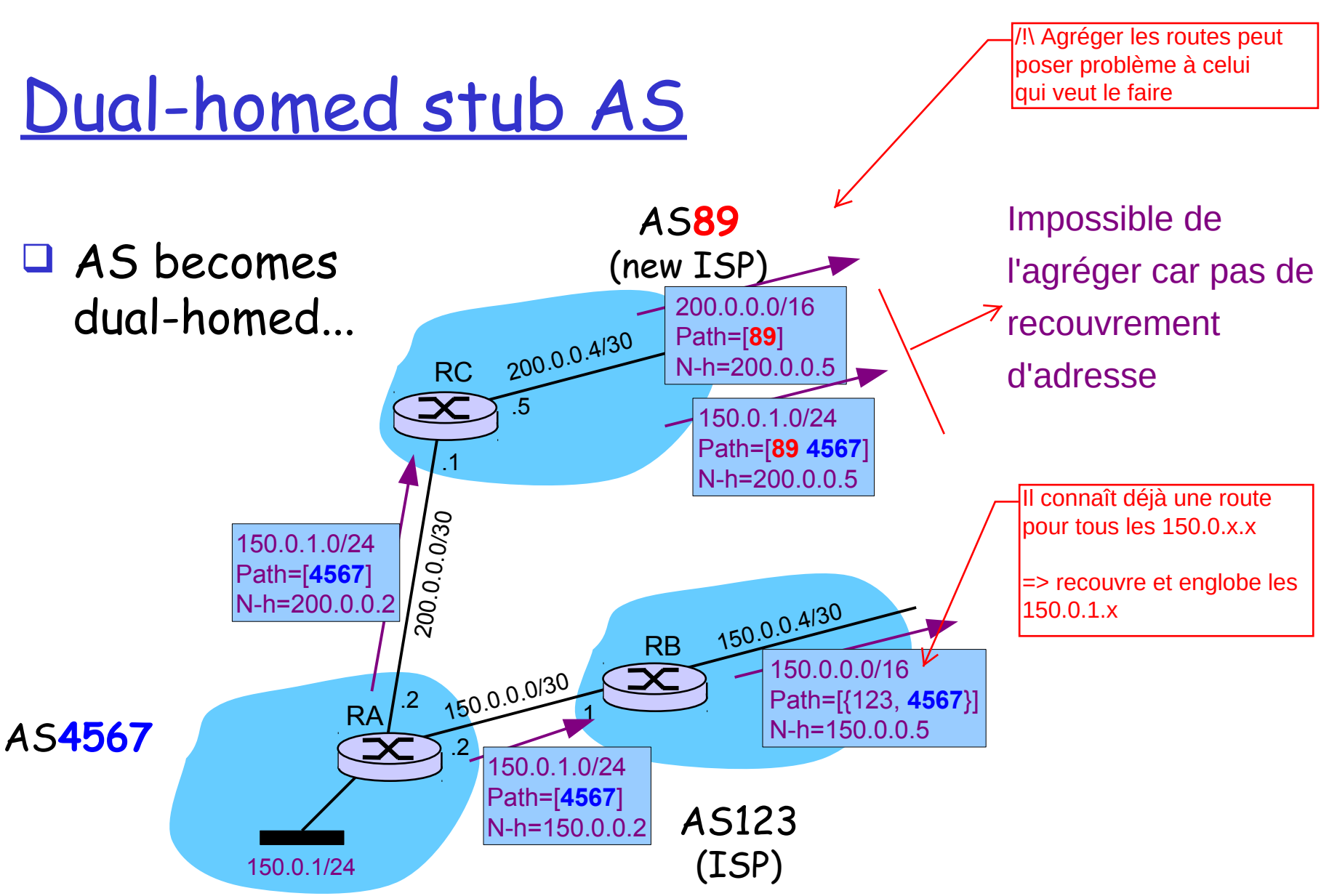
```
--- 08Oct10 ---
ASnum   NetsNow  NetsAggr  NetGain  % Gain  Description

Table   338240    209651    128589   38.0%  All ASes
AS6389    3774       282      3492   92.5%  BELLSOUTH-NET-BLK - BellSouth.net Inc.
AS4323    4485      1985      2500   55.7%  TWTC - tw telecom holdings, inc.
AS19262   1779       279      1500   84.3%  VZGNI-TRANSIT - Verizon Online LLC
AS4766    1863       523      1340   71.9%  KIXS-AS-KR Korea Telecom
AS22773   1204        66      1138   94.5%  ASN-CXA-ALL-CCI-22773-RDC - Cox Communications Inc.
AS4755    1365       297      1068   78.2%  TATACOMM-AS TATA Communications formerly VSNL is Leading ISP
AS17488   1360       308      1052   77.4%  HATHWAY-NET-AP Hathway IP Over Cable Internet
AS5668    1063        93       970   91.3%  AS-5668 - CenturyTel Internet Holdings, Inc.
AS10620   1333       376       957   71.8%  Telmex Colombia S.A.
AS6478    1368       427       941   68.8%  ATT-INTERNET3 - AT&T Services, Inc.
AS18566   1058       175       883   83.5%  COVAD - Covad Communications Co.
AS1785    1795      1013       782   43.6%  AS-PAETEC-NET - PaeTec Communications, Inc.
AS7545    1418       696       722   50.9%  TPG-INTERNET-AP TPG Internet Pty Ltd
AS7303     805       101       704   87.5%  Telecom Argentina S.A.
AS8452    1046       371       675   64.5%  TE-AS TE-AS
AS8151    1340       690       650   48.5%  Uninet S.A. de C.V.
AS33363   1368       734       634   46.3%  BHN-TAMPA - BRIGHT HOUSE NETWORKS, LLC
AS4808     936       303       633   67.6%  CHINA169-BJ CNCGROUP IP network China169 Beijing Province Network
AS18101    884       251       633   71.6%  RELIANCE-COMMUNICATIONS-IN Reliance Communications Ltd.DAKC MUMBAI
AS28573   1172       604       568   48.5%  NET Servicos de Comunicao S.A.
AS7552     650       121       529   81.4%  VIETEL-AS-AP Vietel Corporation
AS4780     707       182       525   74.3%  SEEDNET Digital United Inc.
AS17676    604        81       523   86.6%  GIGAINFRA Softbank BB Corp.
AS7018    1469       947       522   35.5%  ATT-INTERNET4 - AT&T Services, Inc.
AS24560   1045       523       522   50.0%  AIRTELBROADBAND-AS-AP Bharti Airtel Ltd., Telemedia Services
AS9443     575        75       500   87.0%  INTERNETPRIMUS-AS-AP Primus Telecommunications
AS7011    1156       668       488   42.2%  FRONTIER-AND-CITIZENS - Frontier Communications of America, Inc.
AS22047    558        82       476   85.3%  VTR BANDA ANCHA S.A.
AS4804     665       205       460   69.2%  MPX-AS Microplex PTY LTD
AS36992    651       196       455   69.9%  ETISALAT-MISR

Total    39496     12654     26842   68.0%  Top 30 total
```

Source: http://www.cidr-report.org (8/10/2010)

NetsNow: number of prefixes currently advertised
NetsAggr: number of prefixes advertised if fully aggregated

Dual-homed stub AS

- AS becomes dual-homed...

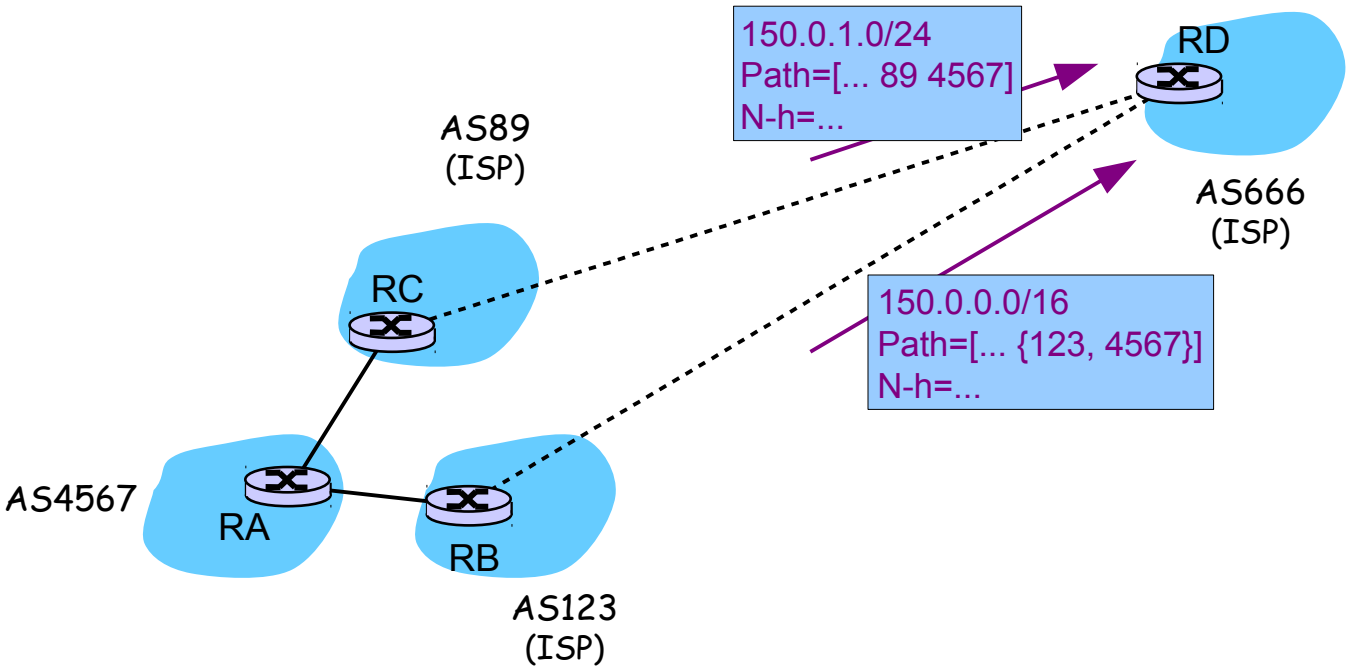


Dual-homed stub AS

BGP ne compare jamais de préfixe.
=> regarder le forwarding => prendre le préfixe le plus long

Mais alors : Quelle route prendre ??

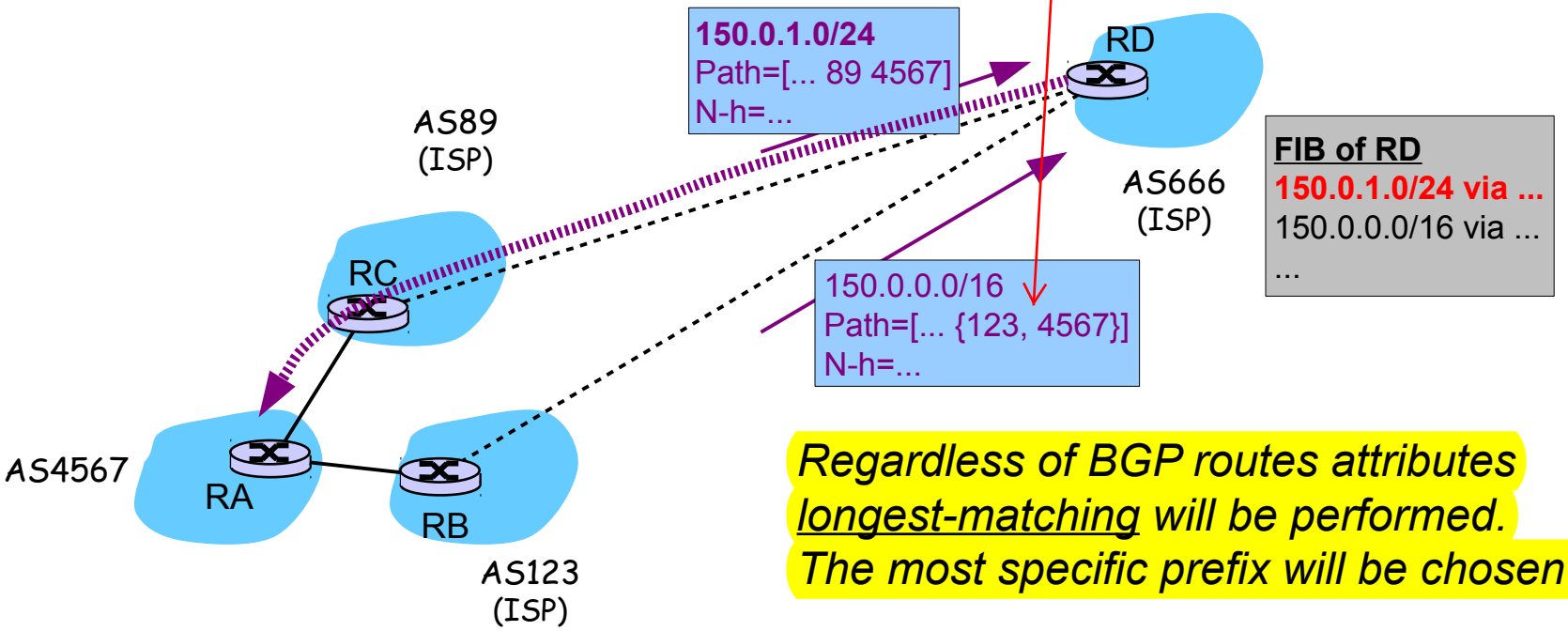
- **Issue**
  - Consider a remote AS receiving the routes advertised by our dual-homed stub's ISPs



/!\ BGP ne choisit les routes que pour le même préfixe

Dual-homed stub AS

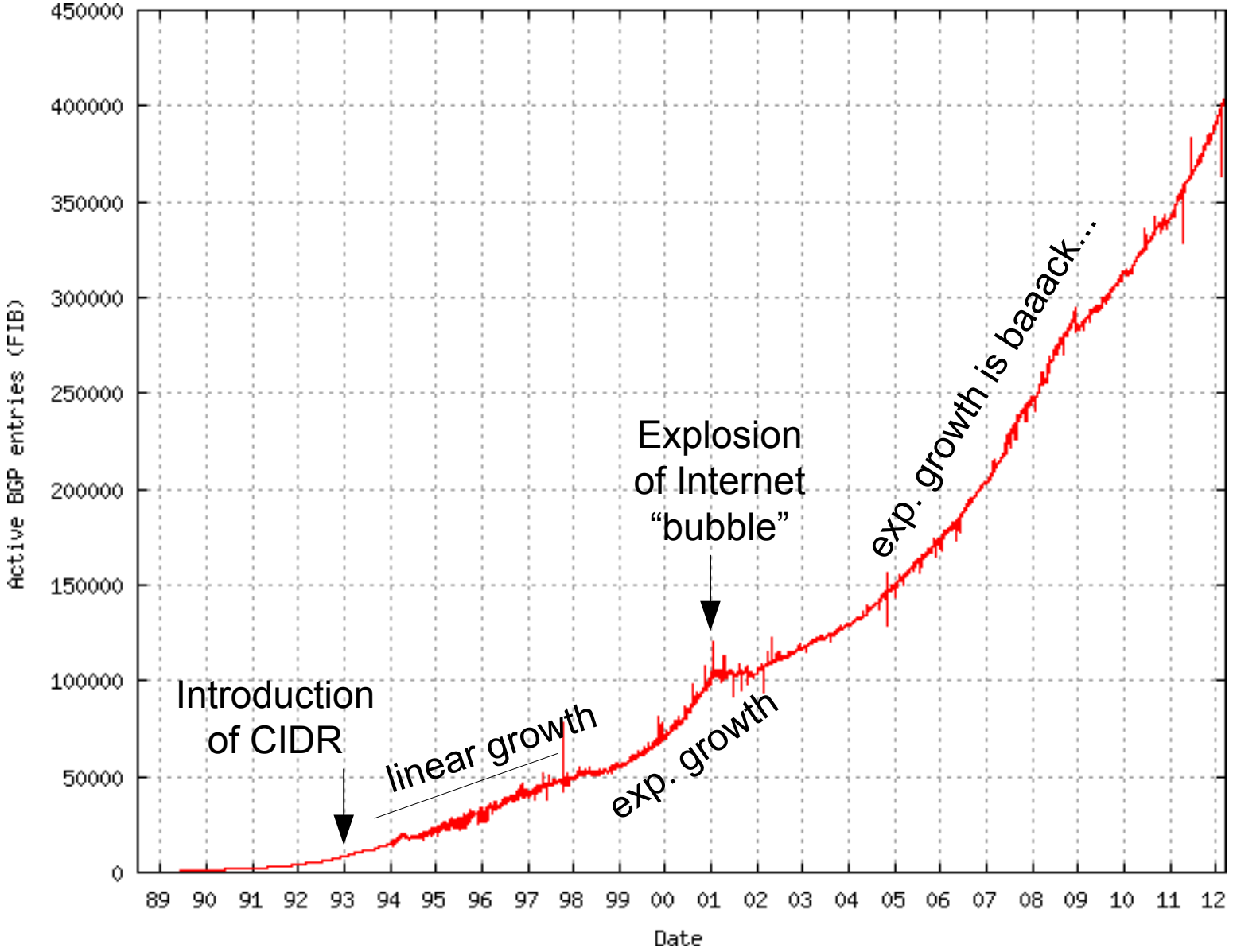
- **Issue**
  - Consider a remote AS receiving the routes advertised by our dual-homed stub's ISPs



Dual-homed stub AS

- **Issues**
  - All traffic to 150.0.1.0/24 will be sent on the non-aggregated path (most specific prefix)
  - AS123 ISP might stop aggregating its customer prefixes. Otherwise, its customers will not receive packets through its links.
  - Hence, the global BGP routing tables are 50% larger than their optimal size (if aggregation was perfectly used)

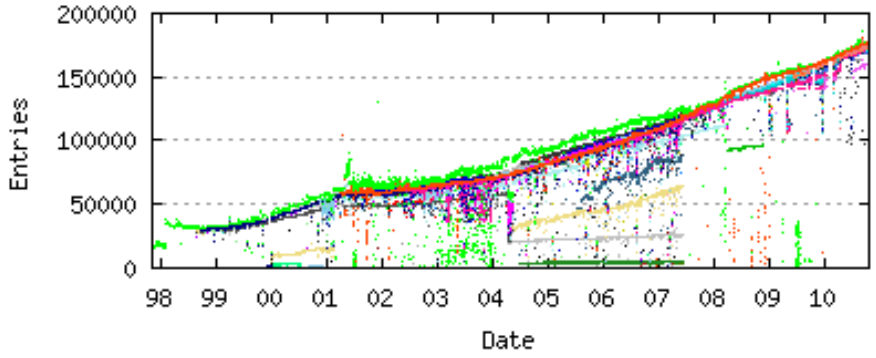
Internet Forwarding Table Growth



Source: The CIDR Report, 22/10/2012
http://www.cidr-report.org

Dual-homed stub AS

- **How to limit growth of BGP tables ?**
  - Long term solution
    - Define a better multi-homing solution
    - Difficult with IPv4, feasible with IPv6 (more later)
  - Short term solution
    - Some ISPs filter routes towards too long prefixes !



Number of /24 in RouteViews Routing tables

Source: The CIDR Report, 8/10/2010
http://www.cidr-report.org

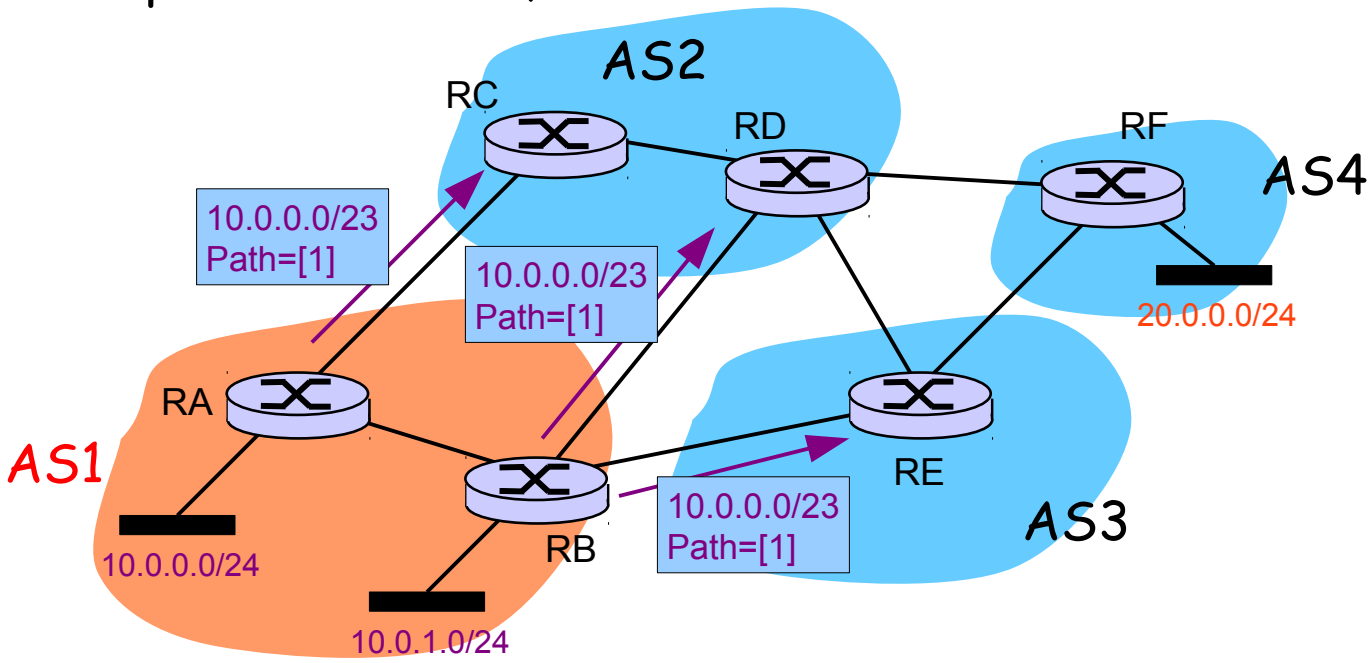
- See for example talk by P. Smith (CISCO) at RIPE 2006 : no reason to see prefixes longer than /22 on the Internet. Excepté les /32 qui sont les racines DNS

Dual-homed stub AS

- **Second Issue to solve**
  - How to better control
    - how traffic reaches a multi-homed stub AS ?
    - how traffic reaches a remote destination ?
  - How ? Tuning BGP attributes to control incoming and outgoing traffic -> Interdomain Traffic Engineering (TE)

Traffic Engineering

- **Sample Network**
  - Routing without tuning the announcements.
  - Packet flows towards AS1 depend on the decision process and policies of AS2, AS3 and AS4.

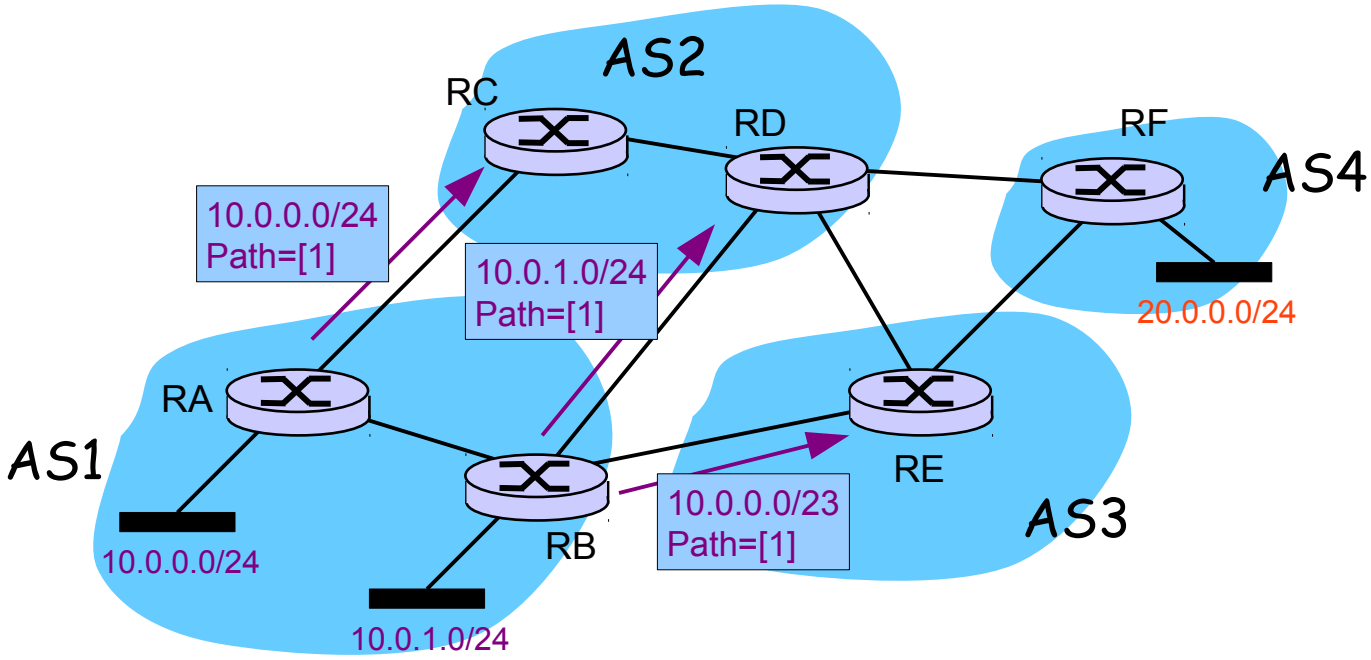


Traffic Engineering

P1 10.0.0.0/24
P2 10.0.1.0/24
P3 10.0.0.0/23

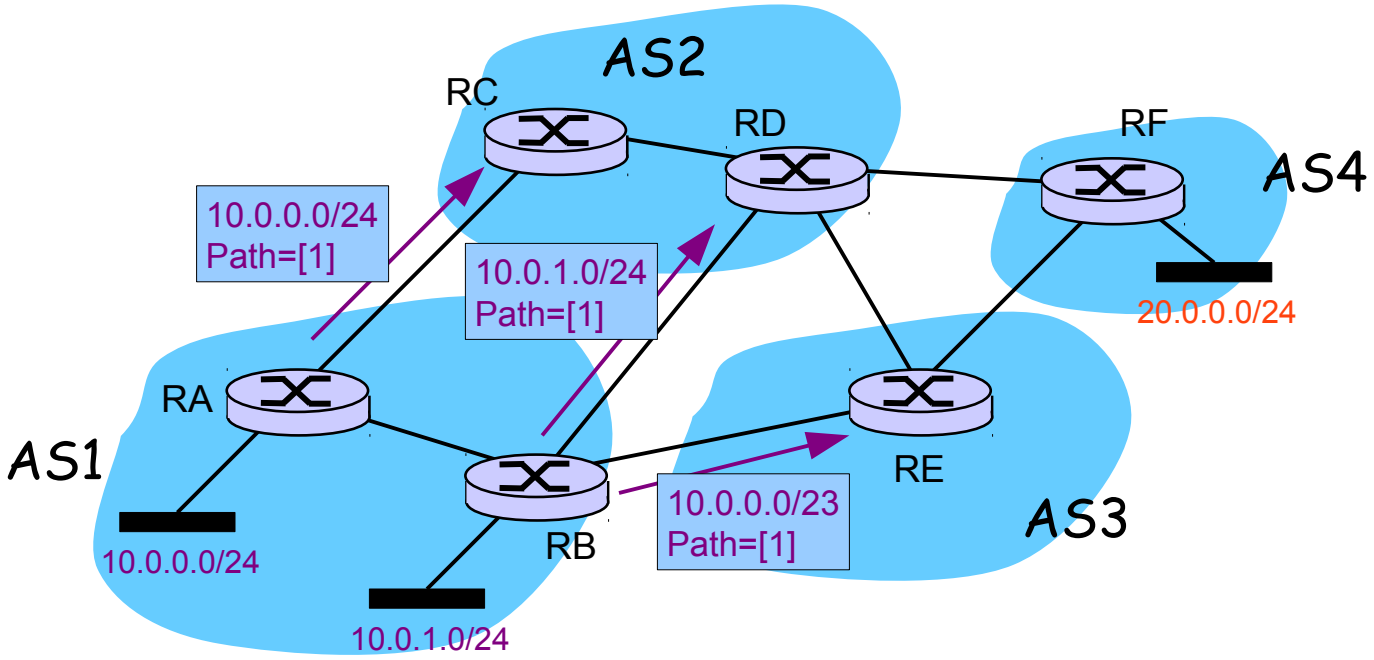
On regarde les 23 premiers bits
=> P3 englobe P1 et P2

- **1st Technique: Selective announcements**
  - Advertise some prefixes only on some links
  - Drawbacks ?



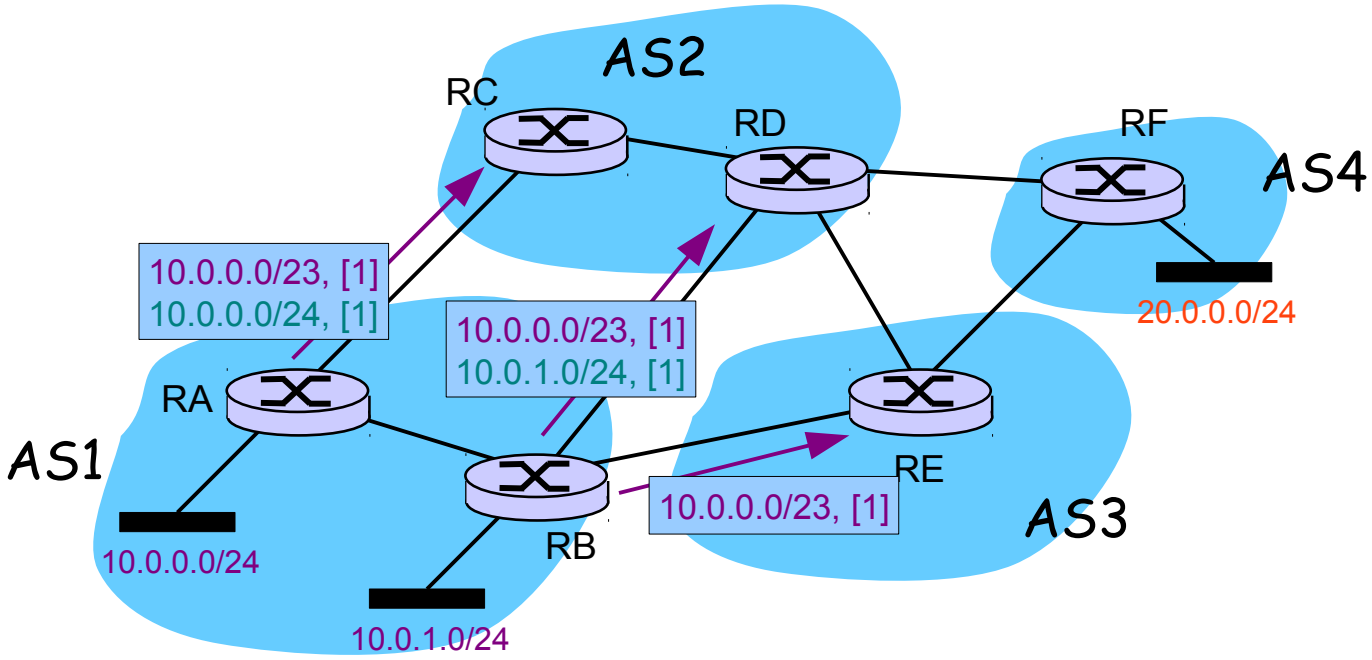
Traffic Engineering

- **1st Technique: Selective announcements**
  - Splitting a prefix increases the size of all BGP routing tables
  - What if a link fails ?



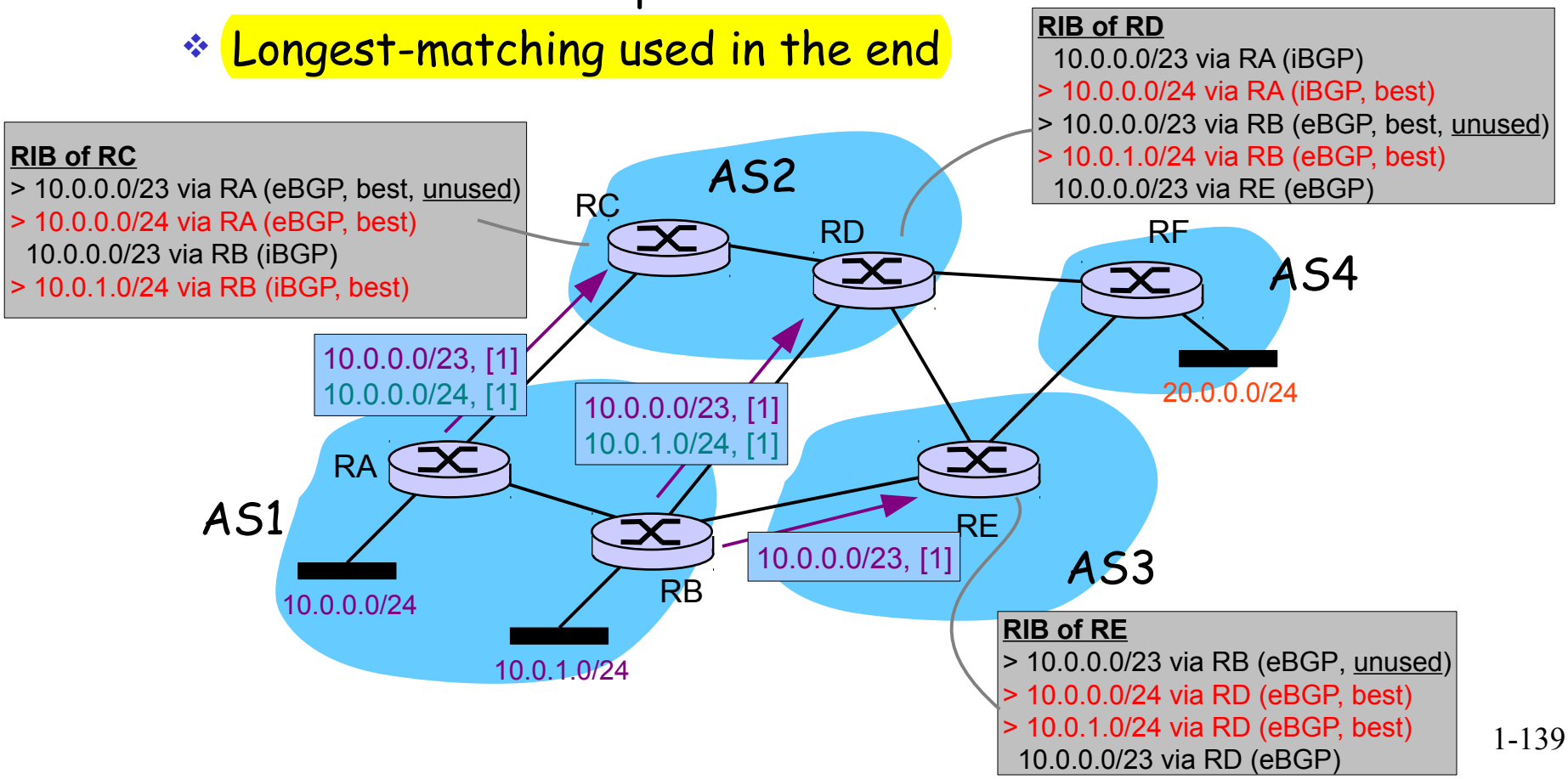
Traffic Engineering

- **2nd Technique: More specific prefixes**
  - Announce a large prefix (aggregate) on all links for redundancy, and selectively send more specific prefixes on some links.



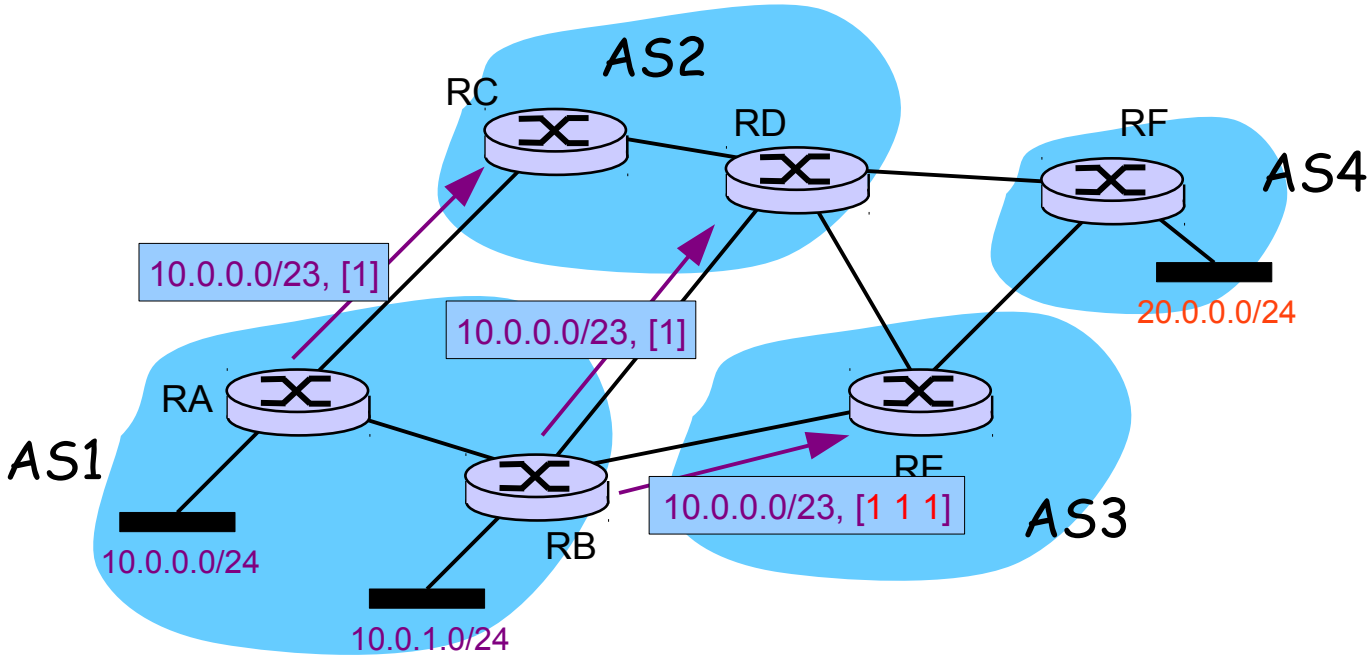
Traffic Engineering

- 2nd Technique: More specific prefixes
  - Routes towards all prefixes selected
  - Longest-matching used in the end



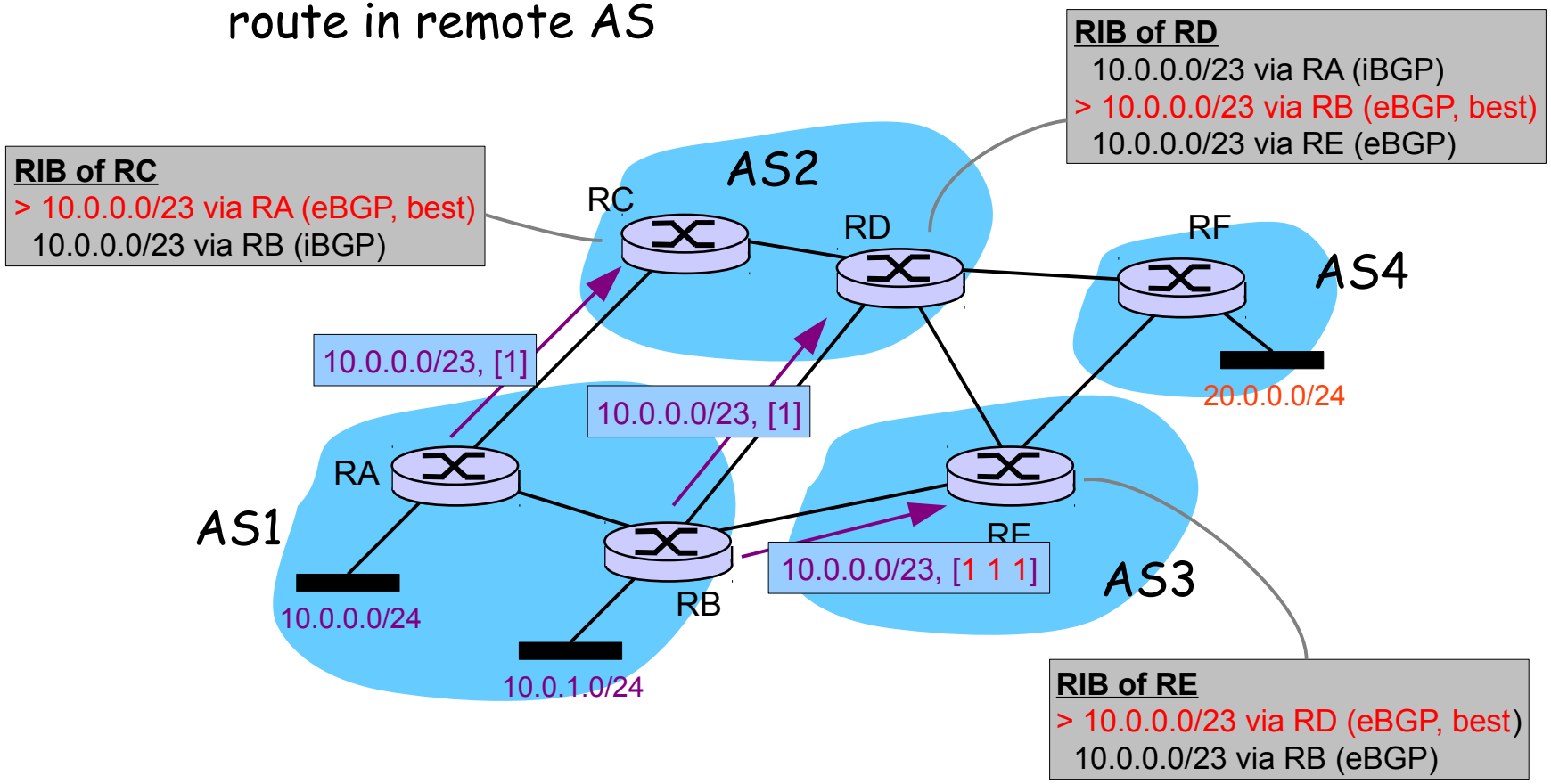
Traffic Engineering

- **3rd Technique: AS-Path prepending**
  - Make AS-Path artificially longer to decrease ranking of route in remote AS



Traffic Engineering

- **3rd Technique: AS-Path prepending**
  - Make AS-Path artificially longer to decrease ranking of route in remote AS



Traffic Engineering

- 3rd Technique: AS-Path prepending
  - Drawback ?