Improving internal BGP routing

V. Van den Schrieck, ICTEAM

Public defense, Dec. 7th 2010

UCL
Université catholique de Louvain

Preliminary observations

New Internet services

=> new requirements for performance

Internet = {competing Internet Service Providers}

Global improvement requires cooperation

What about local improvement?

Objective of the thesis

Improve Internet Service Providers internally to increase the global efficiency of Internet routing

Outline

I. Background on Internet routing

II. Problem Statement

III. Improving iBGP with single path advertisement

IV. Improving iBGP with multiple paths advertisement : BGP Add-Paths

Outline

I. **Background on Internet routing**

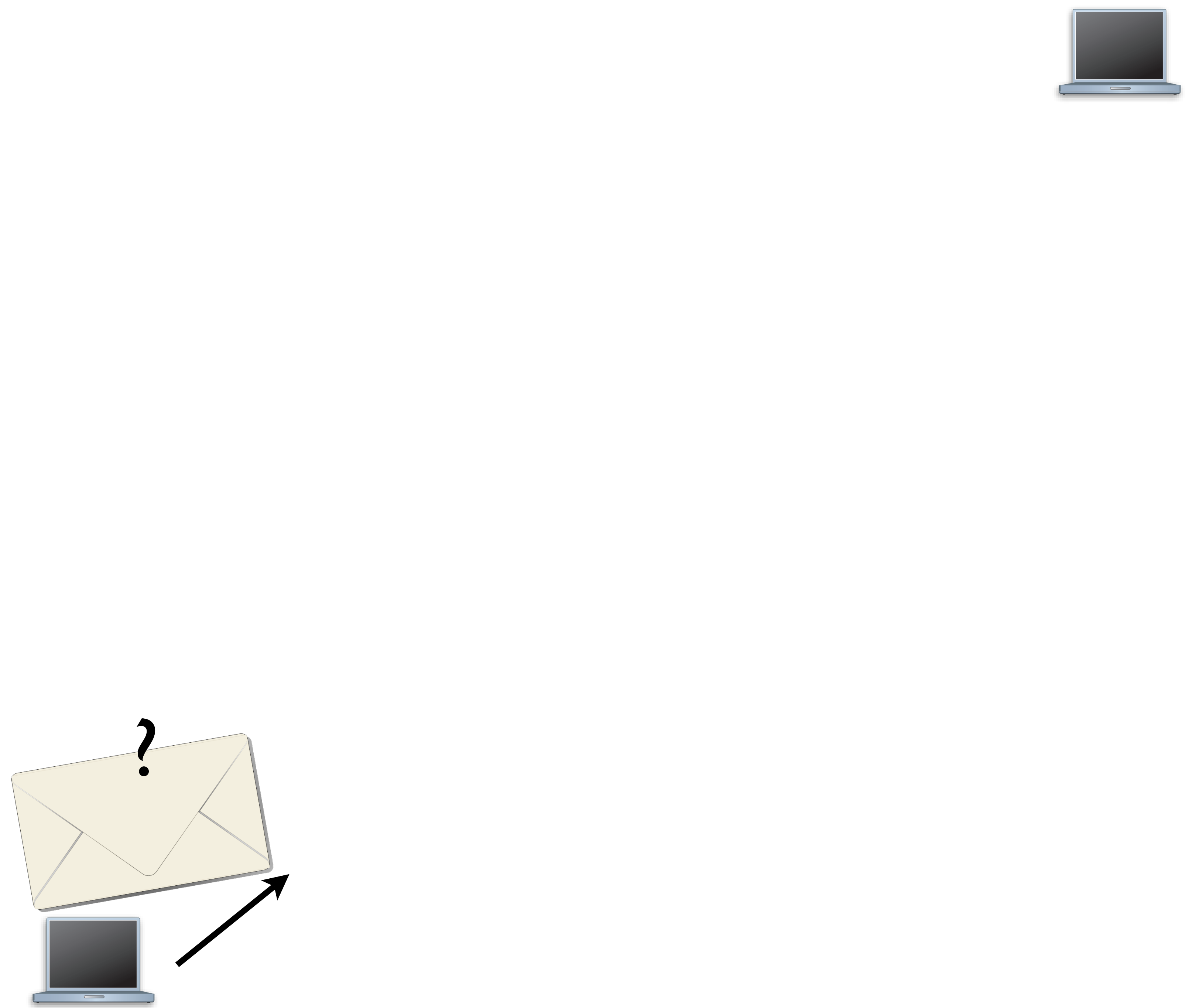
II. Problem Statement

III. Improving iBGP with single path advertisement

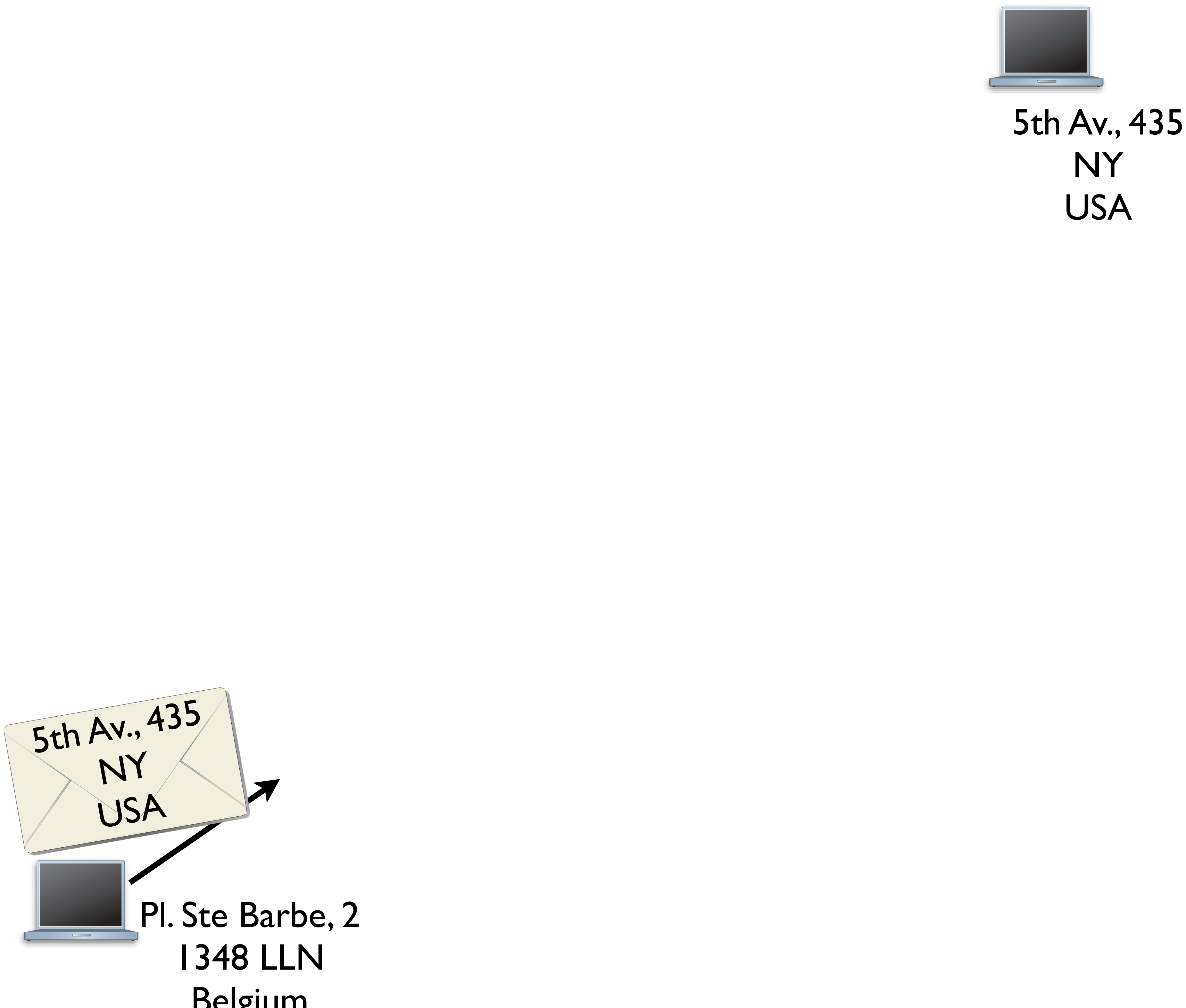
IV. Improving iBGP with multiple paths advertisement : BGP Add-Paths

How does Internet work?

How to send data to a distant end-host?



How does Internet work?



End-Host must be identified

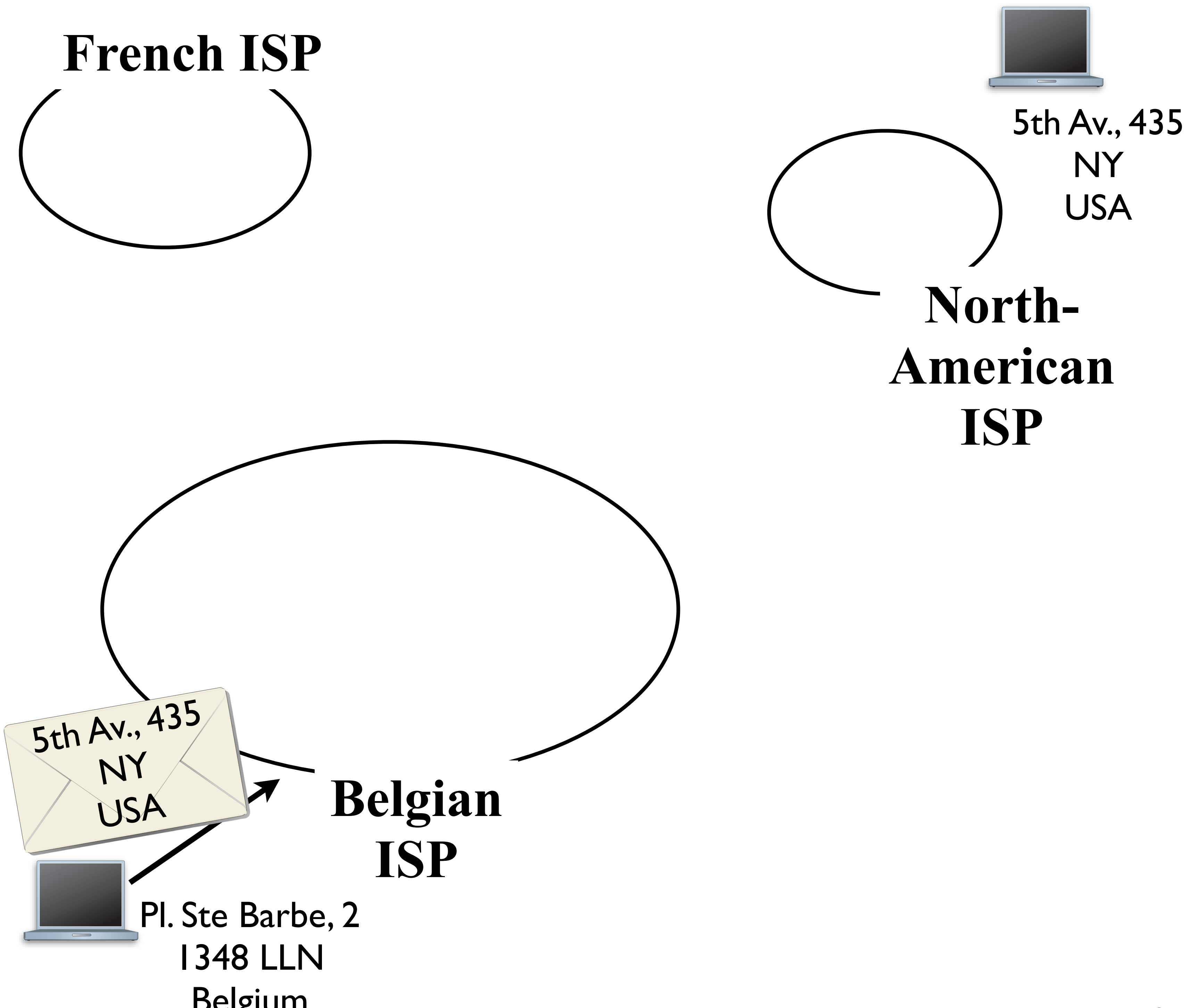
- Address assignment
- In practice : IP addresses (ex : 130.104.128.231)

Destination address is inserted in data exchanged (Packets)

How does Internet work?

End-Hosts send their packets to their Internet Service Providers (ISPs)

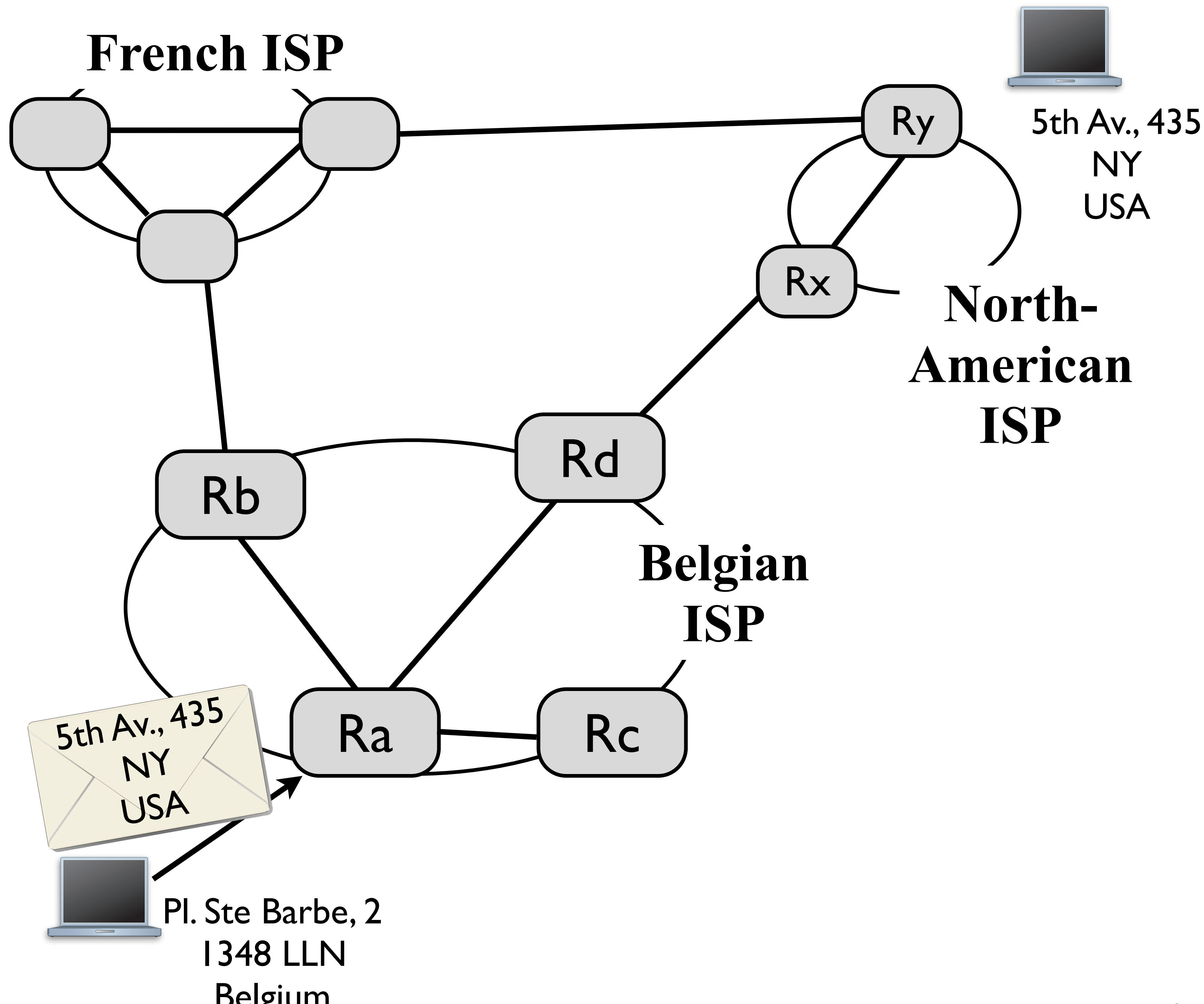
- ISPs handle packet forwarding to the destination
- ISPs can be regional or international
- They exchange data packets with each other



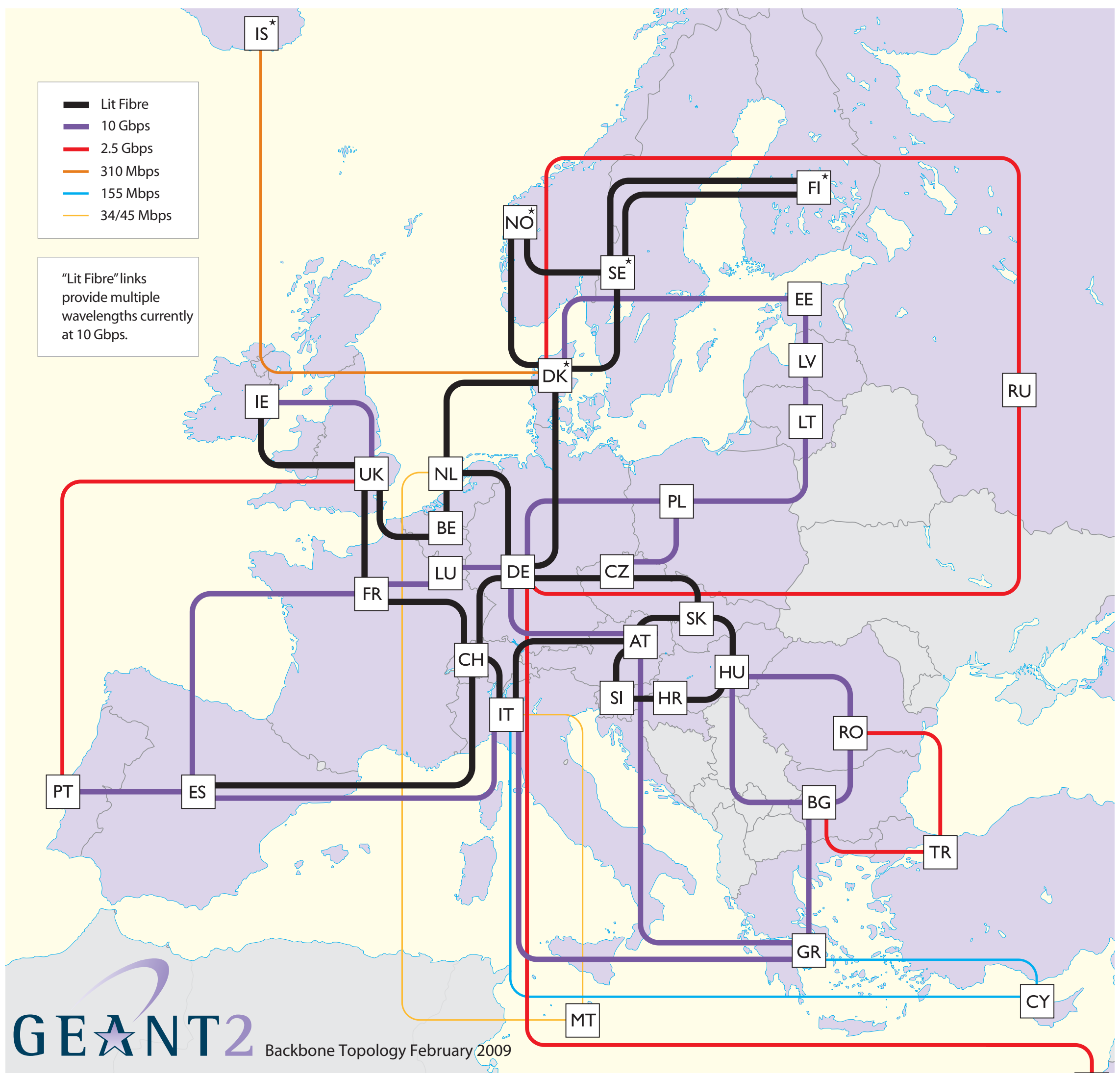
How does Internet work?

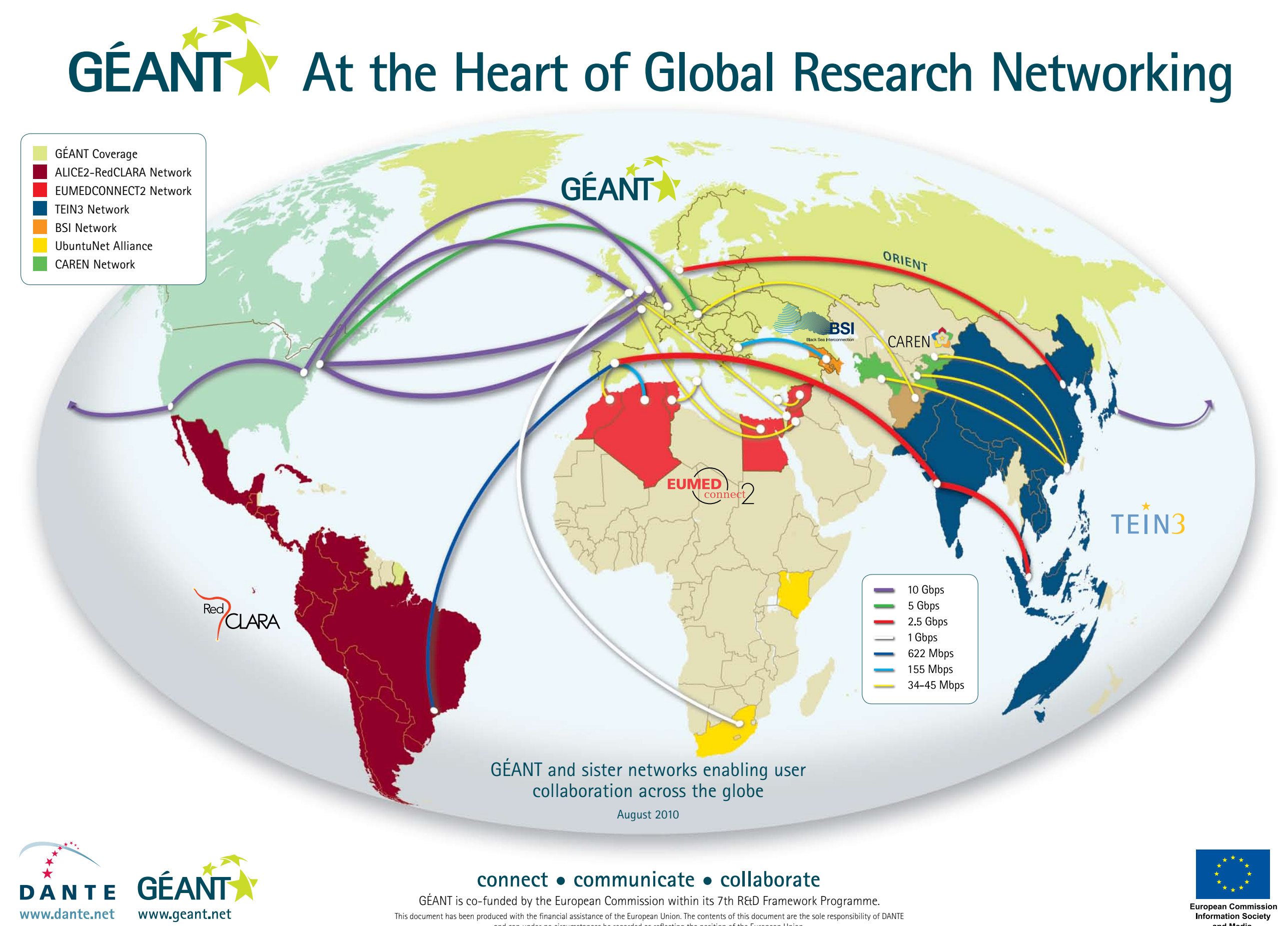
ISPs use devices called "Routers" to forward packets to their destinations

- Routers are connected to each other inside ISPs
- Routers at the borders of ISPs can be connected to routers of other ISPs

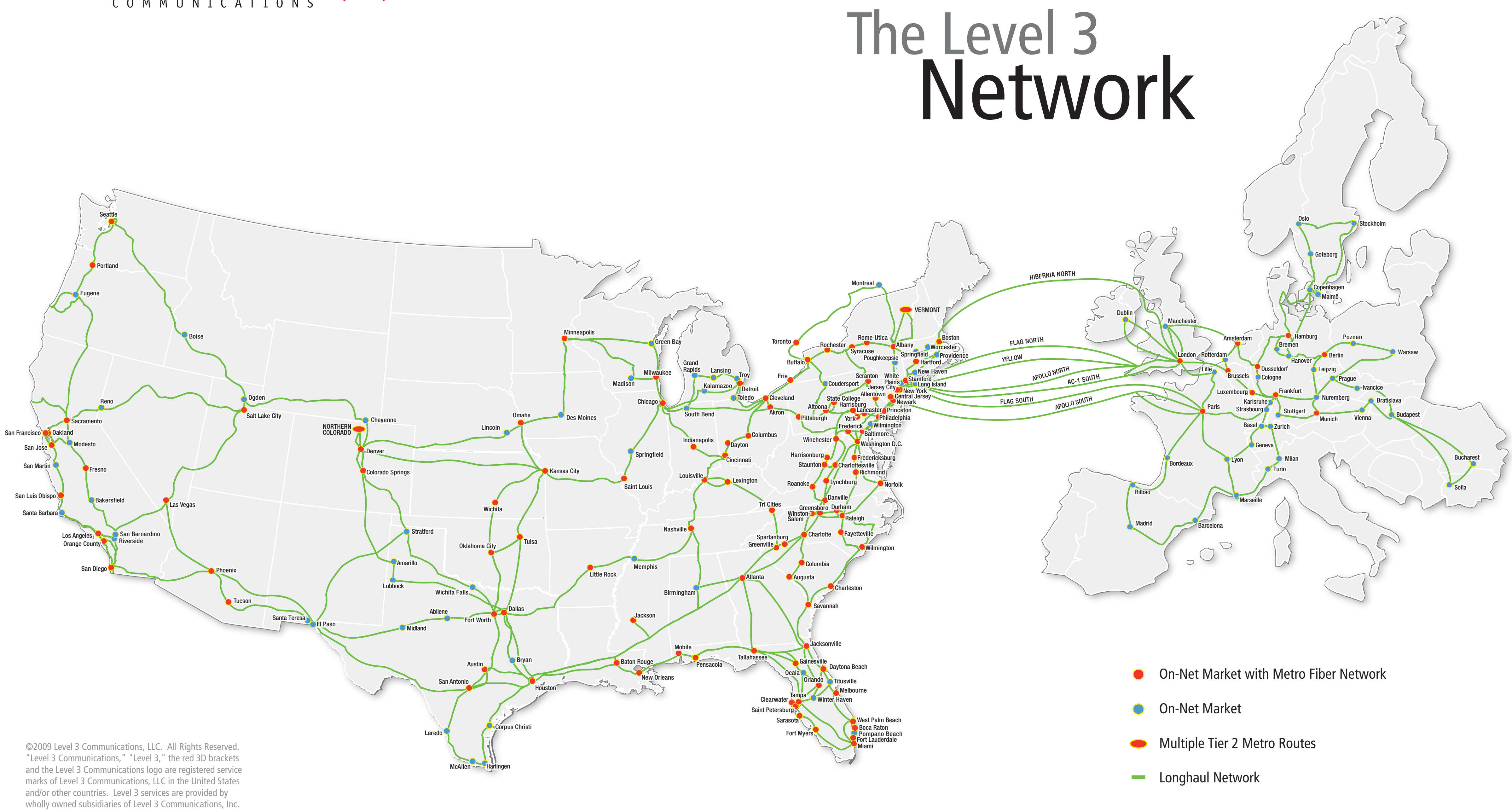


Geant : European ISP





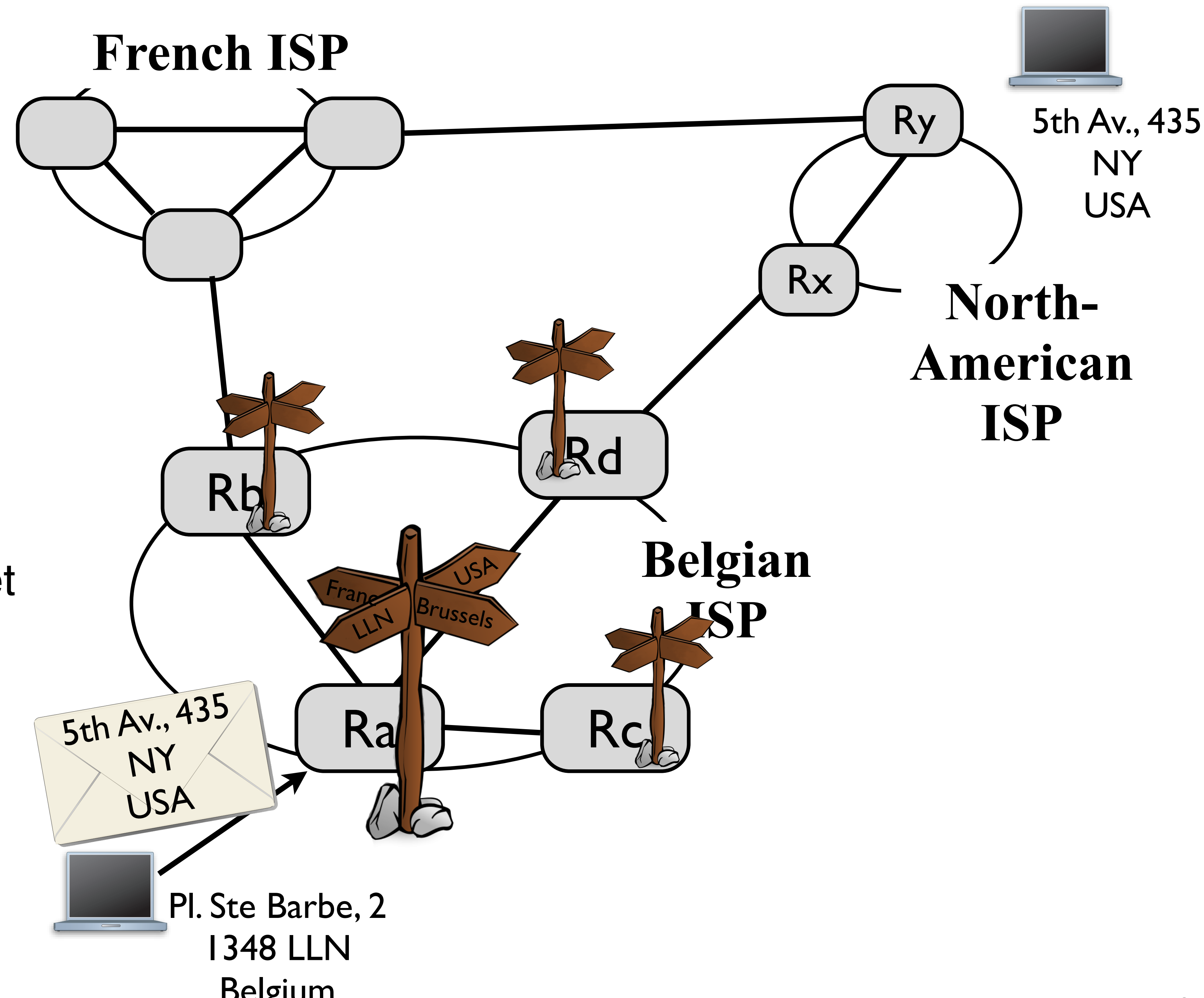
Level 3 : International ISP



How does Internet work?

Routers know how to forward packets to their destination

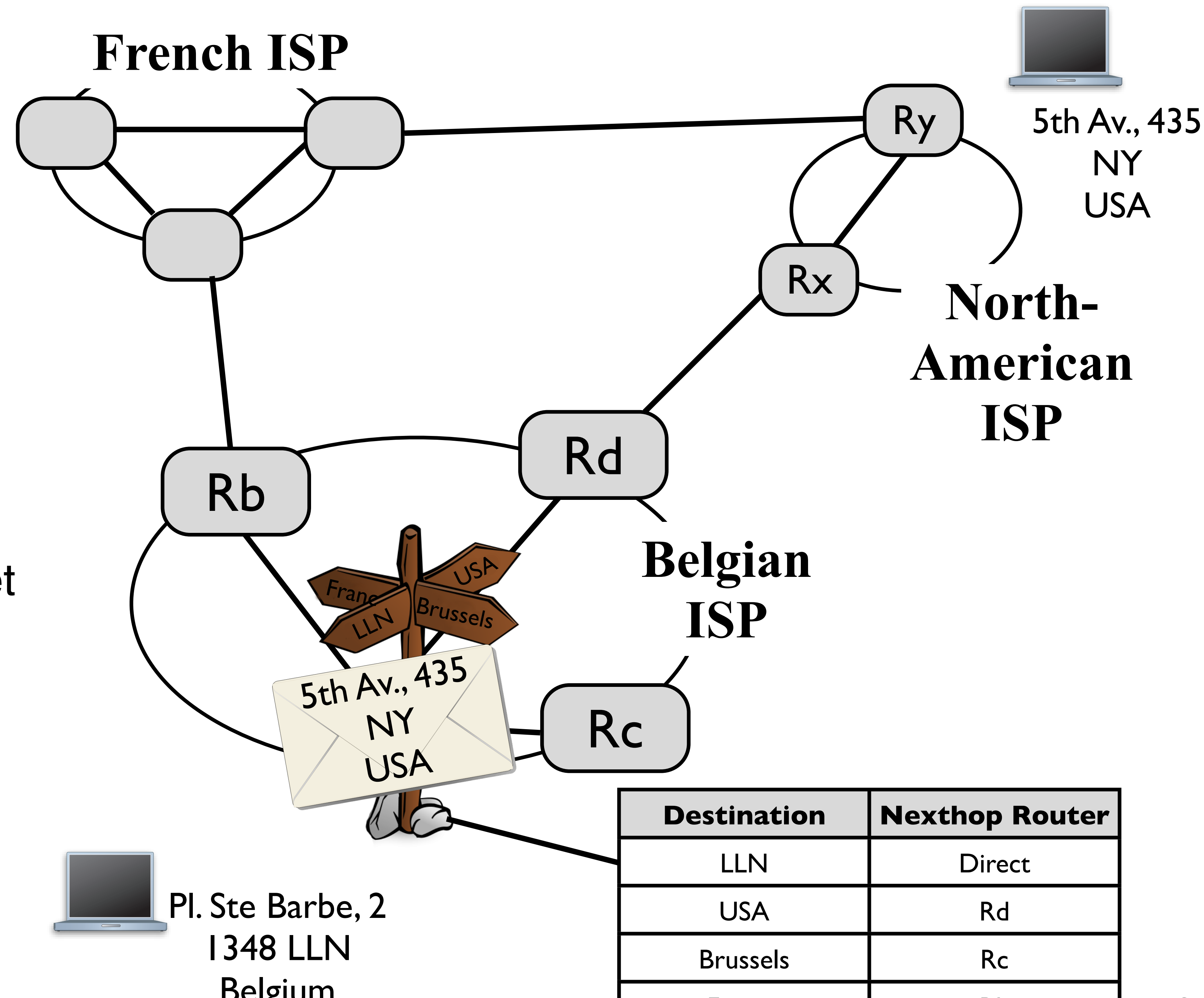
- They maintain information about local and global directions inside a “**Routing Table**”
- A router sends each packet to the neighbor corresponding to the destination address



How does Internet work?

Routers know how to forward packets to their destination

- They maintain information about local and global directions inside a “**Routing Table**”
- A router sends each packet to the neighbor corresponding to the destination address

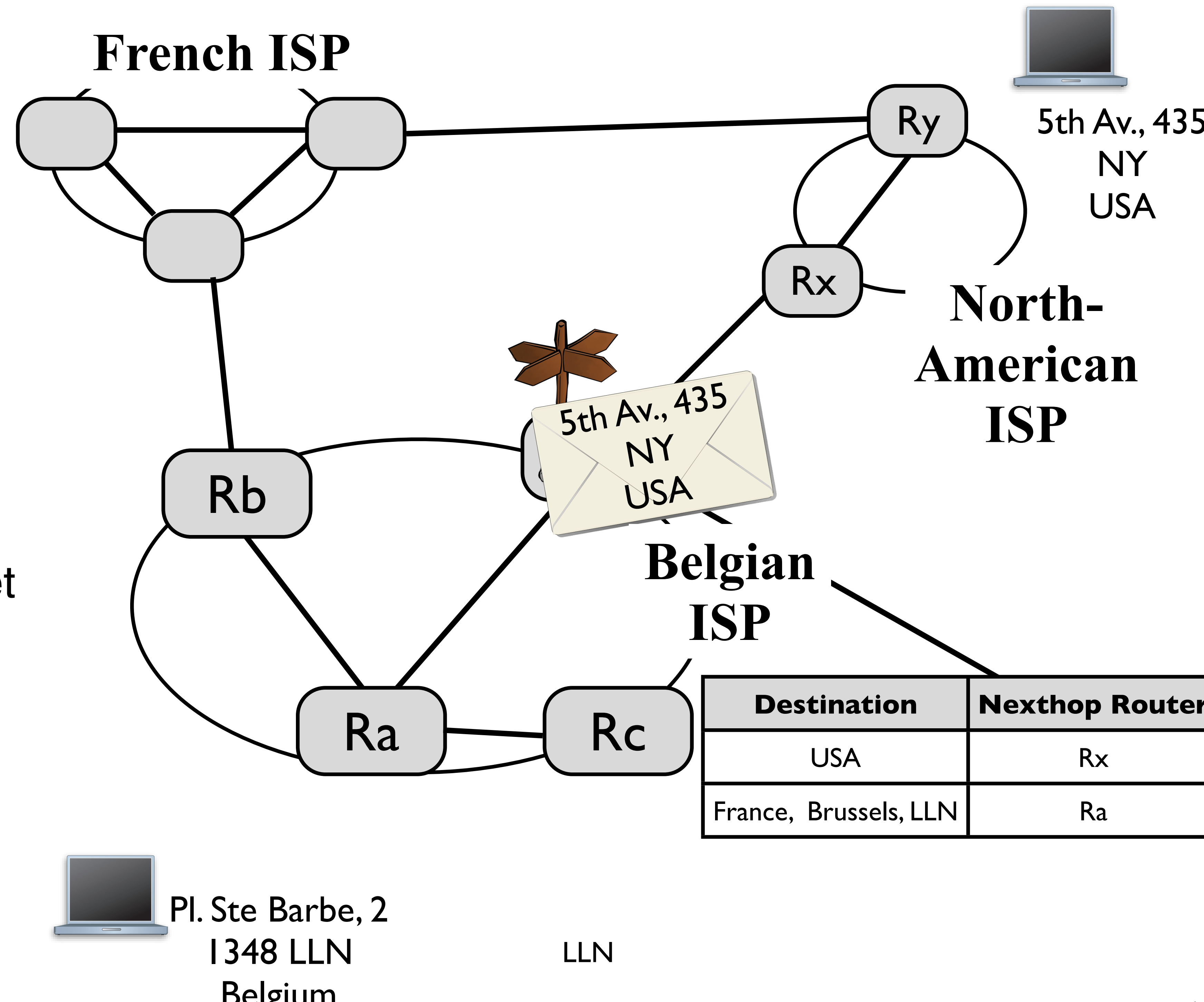


Destination	Nexthop Router
LLN	Direct
USA	Rd
Brussels	Rc
France	Rb

How does Internet work?

Routers know how to forward packets to their destination

- They maintain information about local and global directions inside a “**Routing Table**”
- A router sends each packet to the neighbor corresponding to the destination address



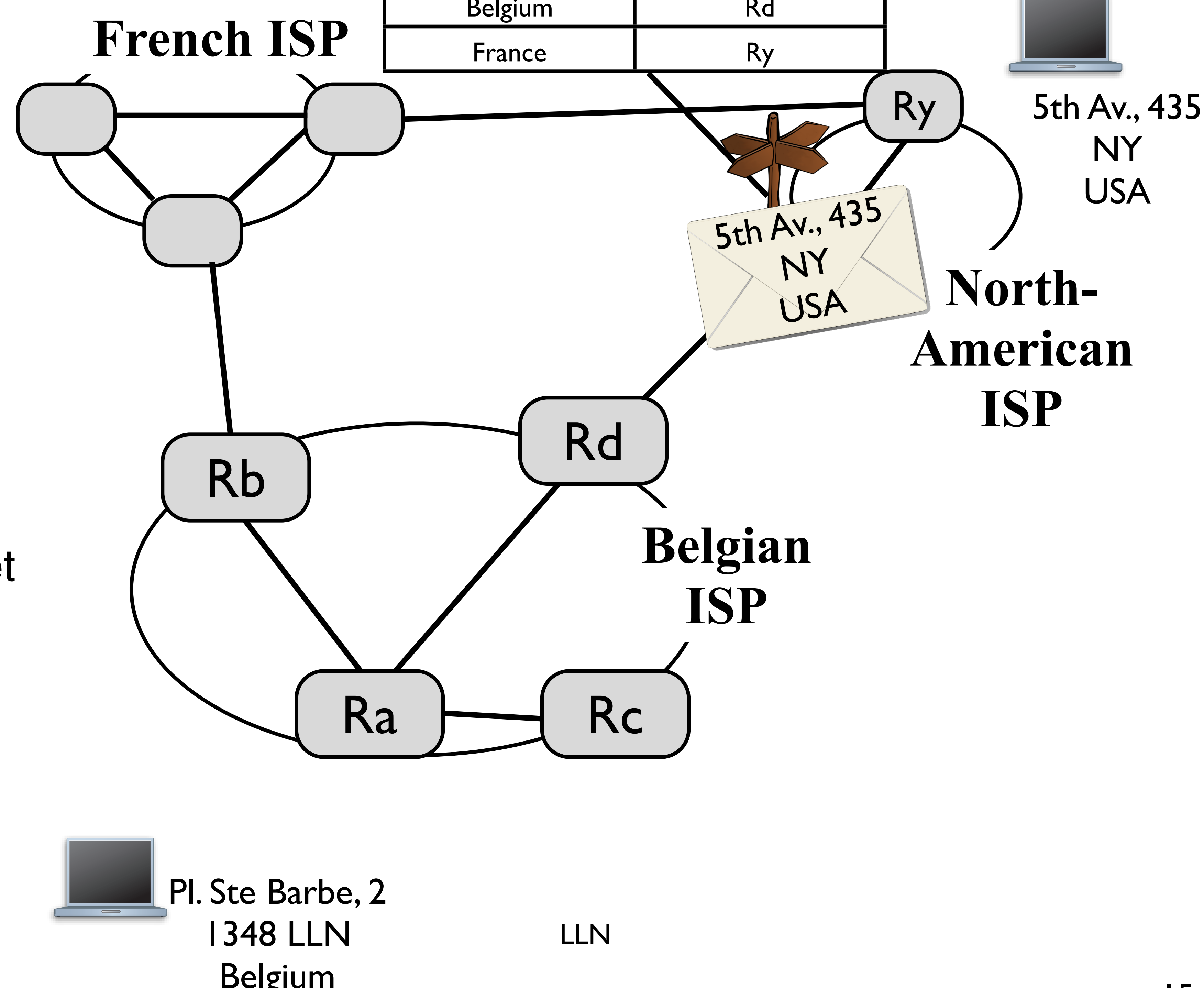
Destination	Nexthop Router
USA	Rx
France, Brussels, LLN	Ra

How does Internet work?

Destination	Nexthop Router
New York	Ry
Belgium	Rd
France	Ry

Routers know how to forward packets to their destination

- They maintain information about local and global directions inside a "**Routing Table**"
- A router sends each packet to the neighbor corresponding to the destination address

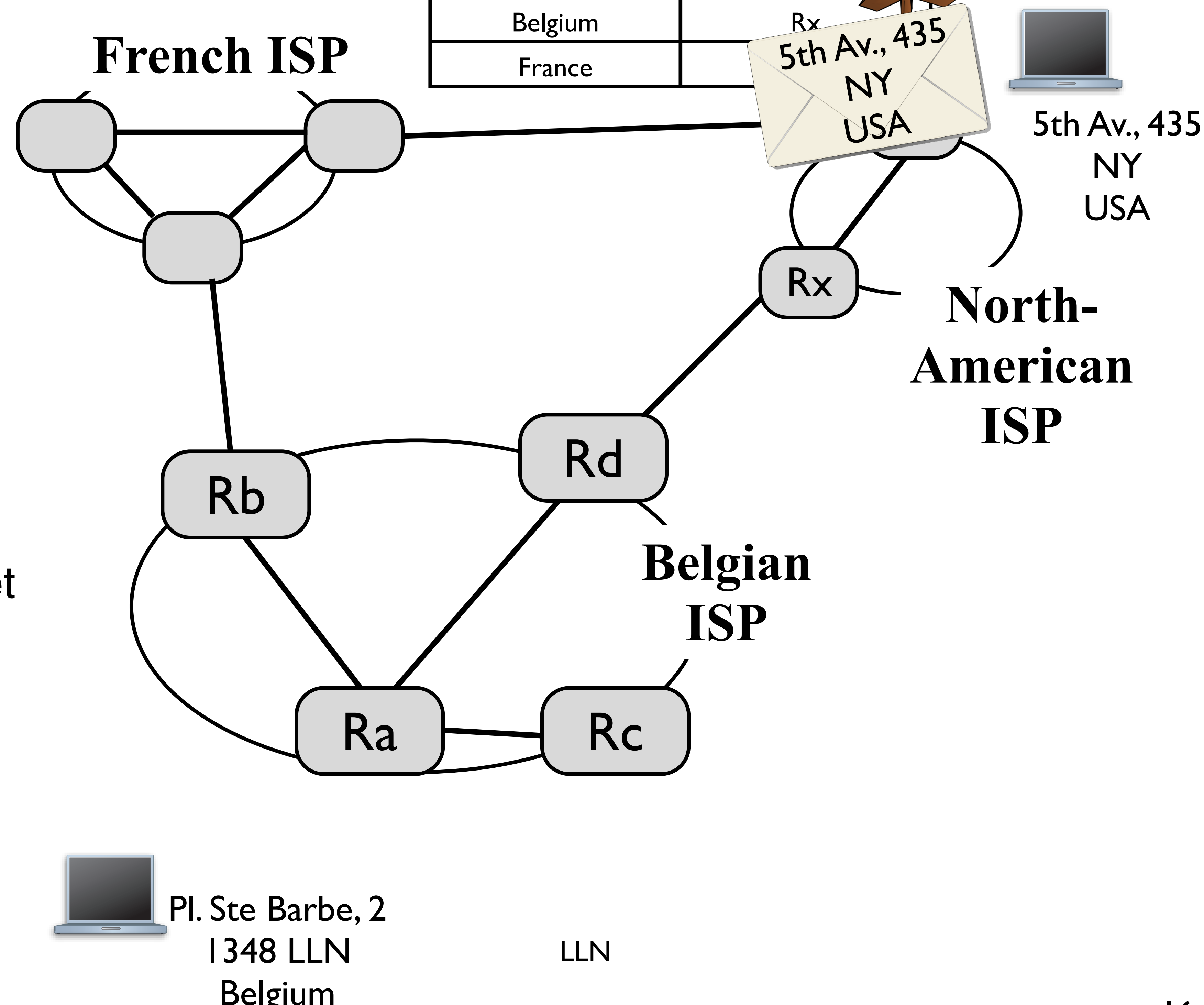


How does Internet work?

Routers know how to forward packets to their destination

- They maintain information about local and global directions inside a "**Routing Table**"
- A router sends each packet to the neighbor corresponding to the destination address

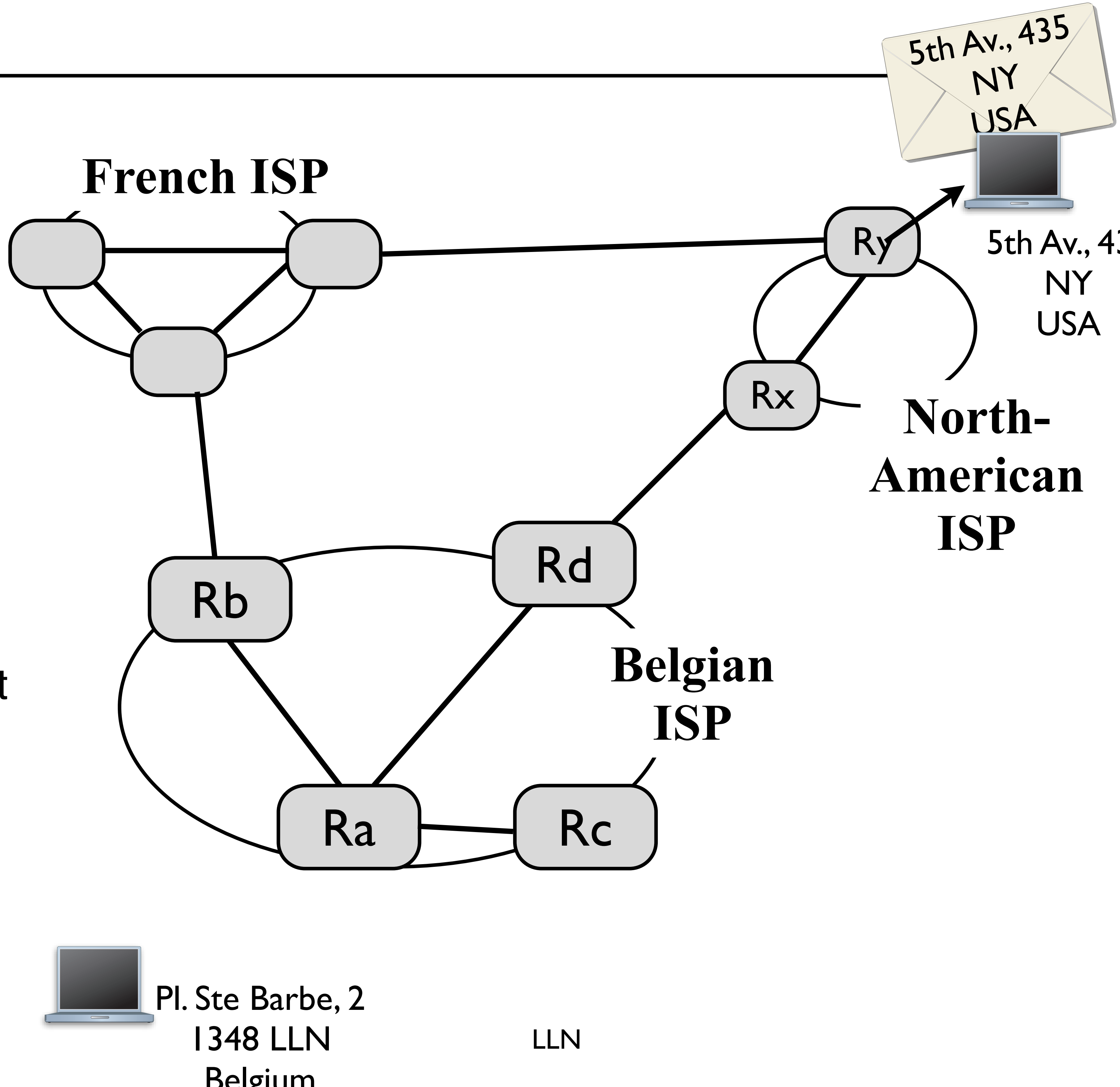
Destination	Nexthop Router
New York	Direct
Belgium	Rx
France	



How does Internet work?

Routers know how to forward packets to their destination

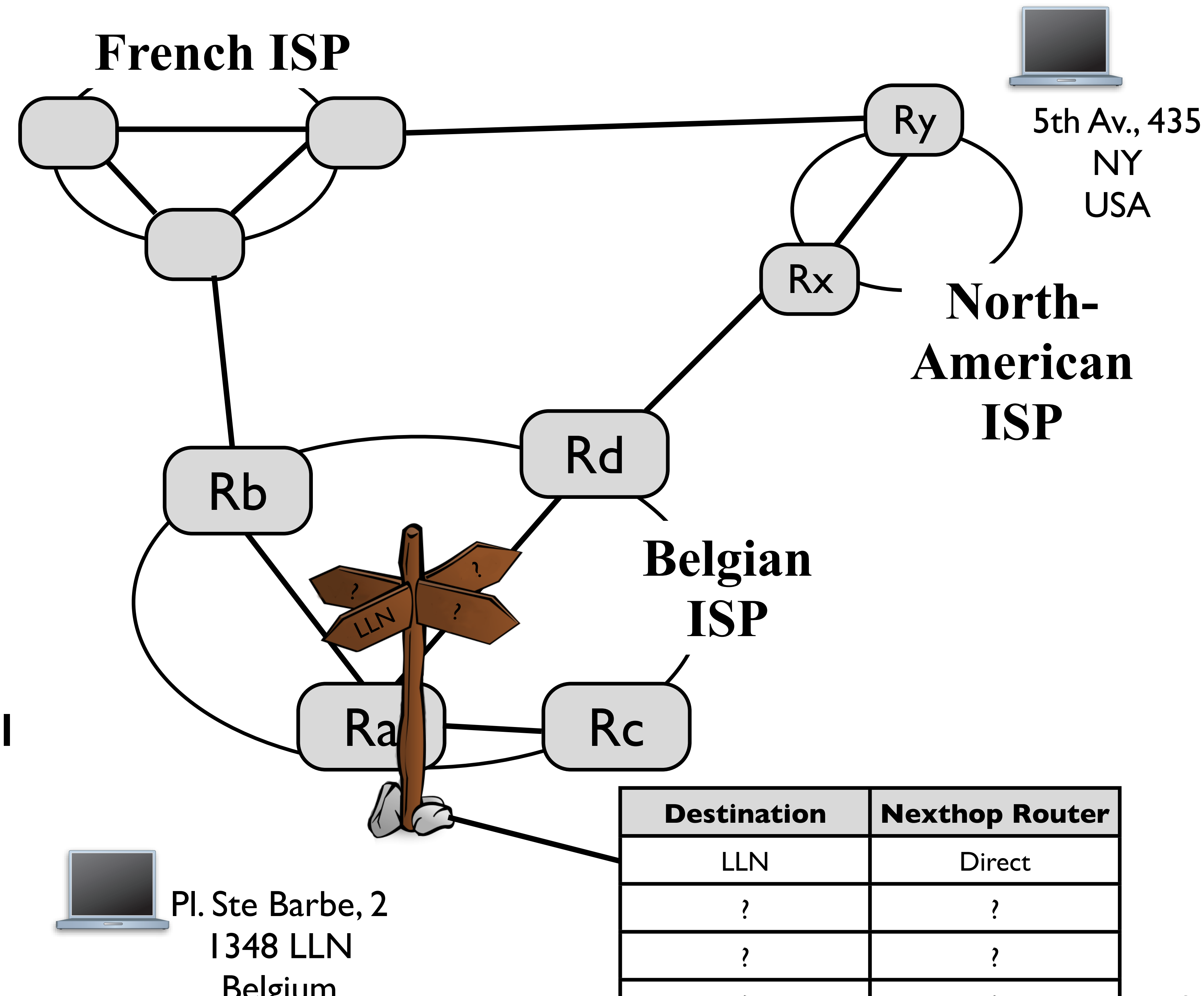
- They maintain information about local and global directions inside a "**Routing Table**"
- A router sends each packet to the neighbor corresponding to the destination address



How to build Routing Tables?

Routers exchange information about reachable destinations

- They use **Routing Protocols**
- In the Internet, the Interdomain Routing protocol used is the **Border Gateway Protocol (BGP)**



Destination	Nexthop Router
LLN	Direct
?	?
?	?
?	?

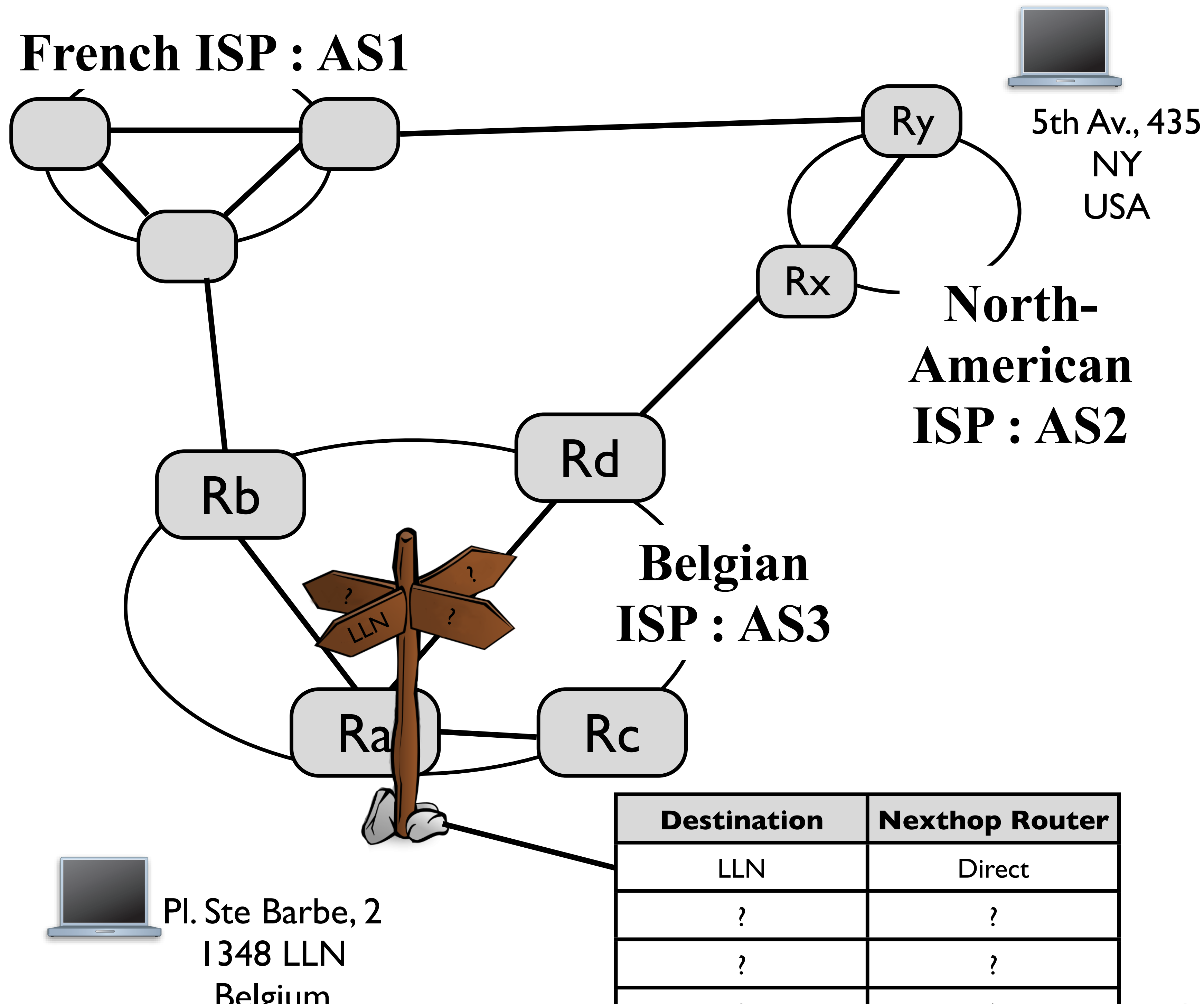
The Border Gateway Protocol

ISPs are called **Autonomous Systems** (AS)

- Each AS is identified by an **AS number**

Two components to BGP :

- Between ASes : external BGP (**eBGP**)
- Inside an AS : internal BGP (**iBGP**)

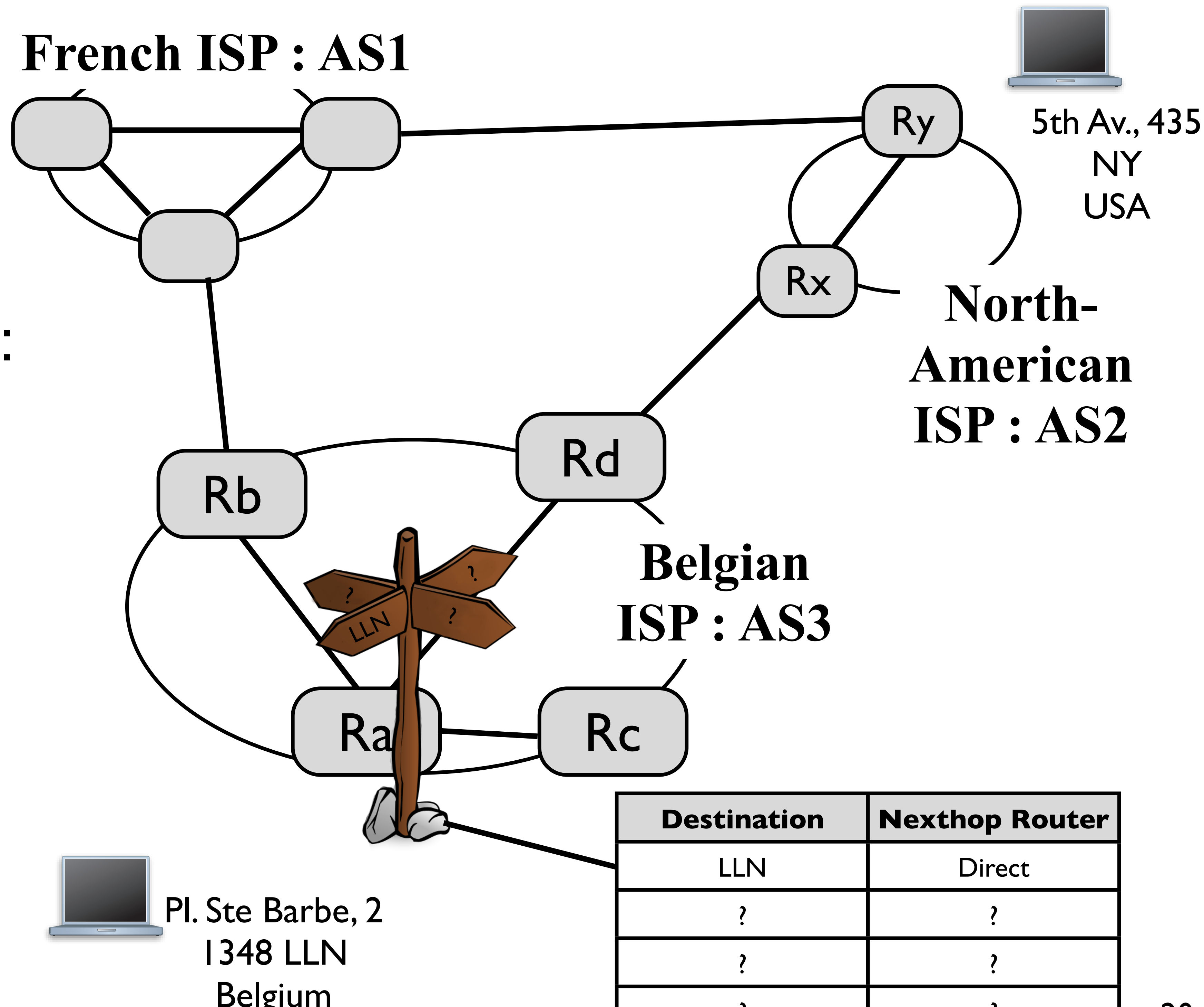


Destination	Nexthop Router
LLN	Direct
?	?
?	?
?	?

The Border Gateway Protocol

Each AS advertises reachable destinations in BGP messages containing :

- An **AS Path** : List of ASes on the path to the destination
- A **Nexthop** : The router to which packets for that destination must be forwarded
- Miscellaneous **BGP attributes** (ex : Local Preference)

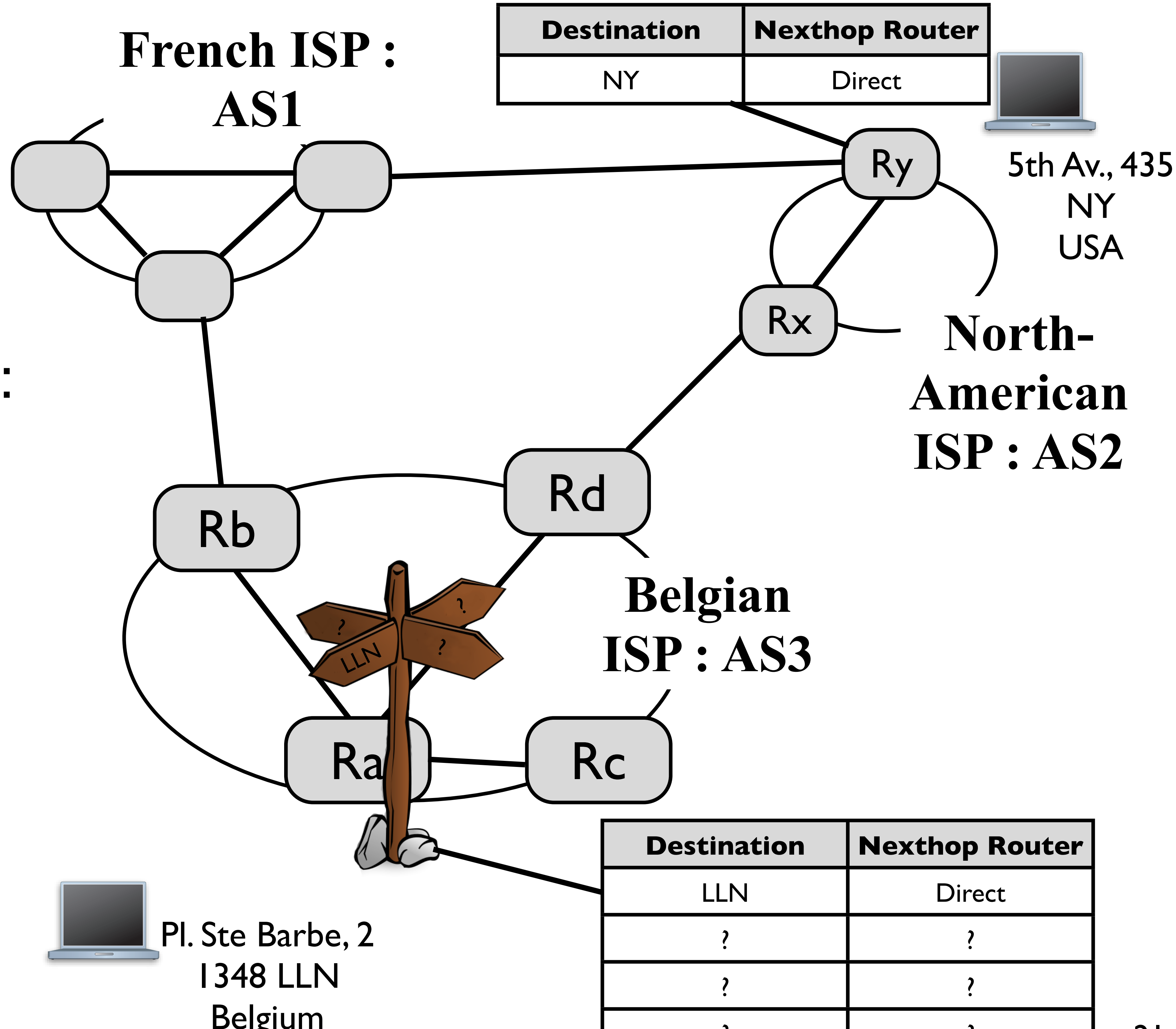


Destination	Nexthop Router
LLN	Direct
?	?
?	?
?	?

The Border Gateway Protocol

Each AS advertises reachable destinations in BGP messages containing :

- An **AS Path** : List of ASes on the path to the destination
- A **Nexthop** : The router to which packets for that destination must be forwarded
- Miscellaneous **BGP attributes** (ex : Local Preference)



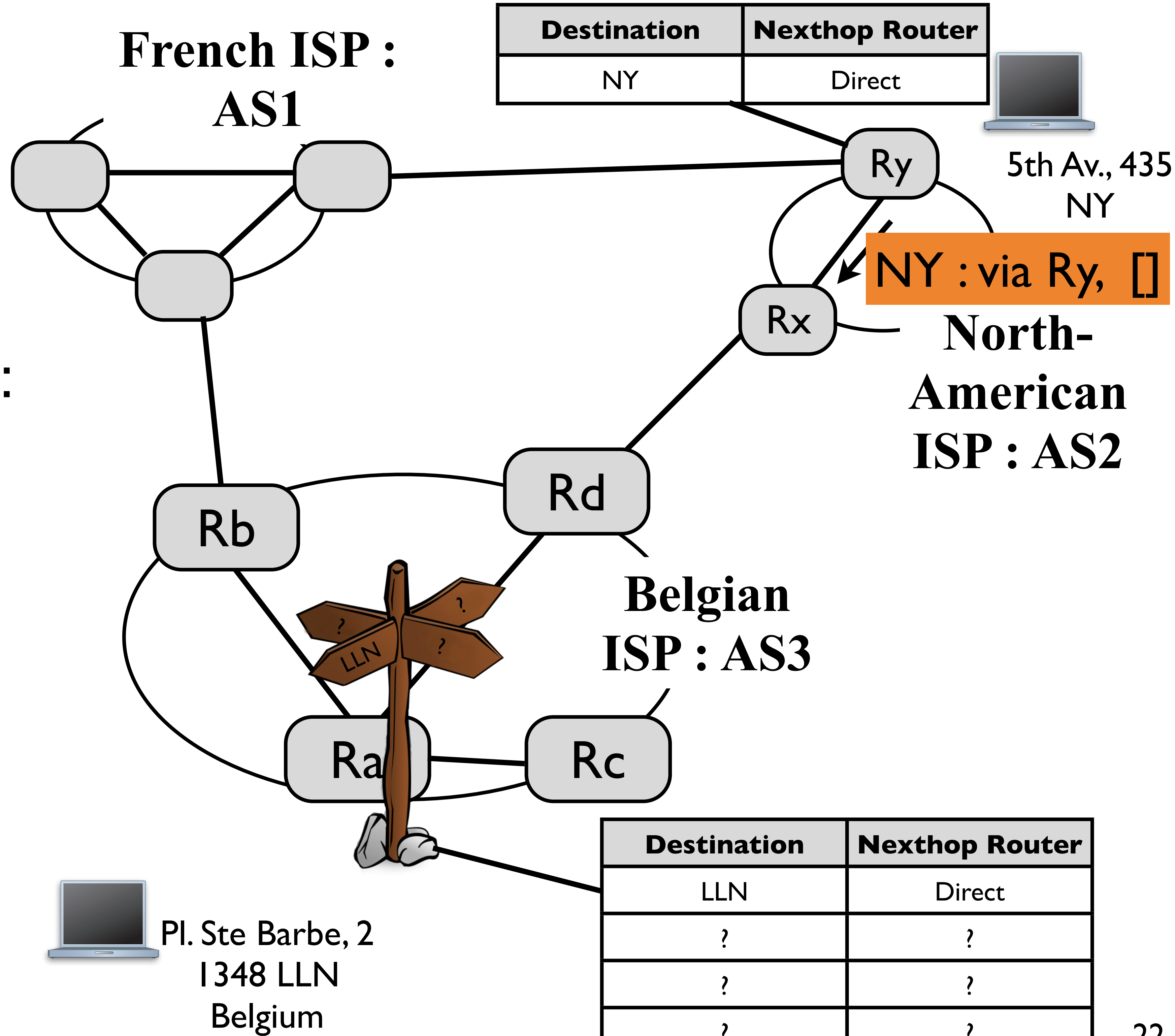
Destination	Nexthop Router
NY	Direct

Destination	Nexthop Router
LLN	Direct
?	?
?	?
?	?

The Border Gateway Protocol

Each AS advertises reachable destinations in BGP messages containing :

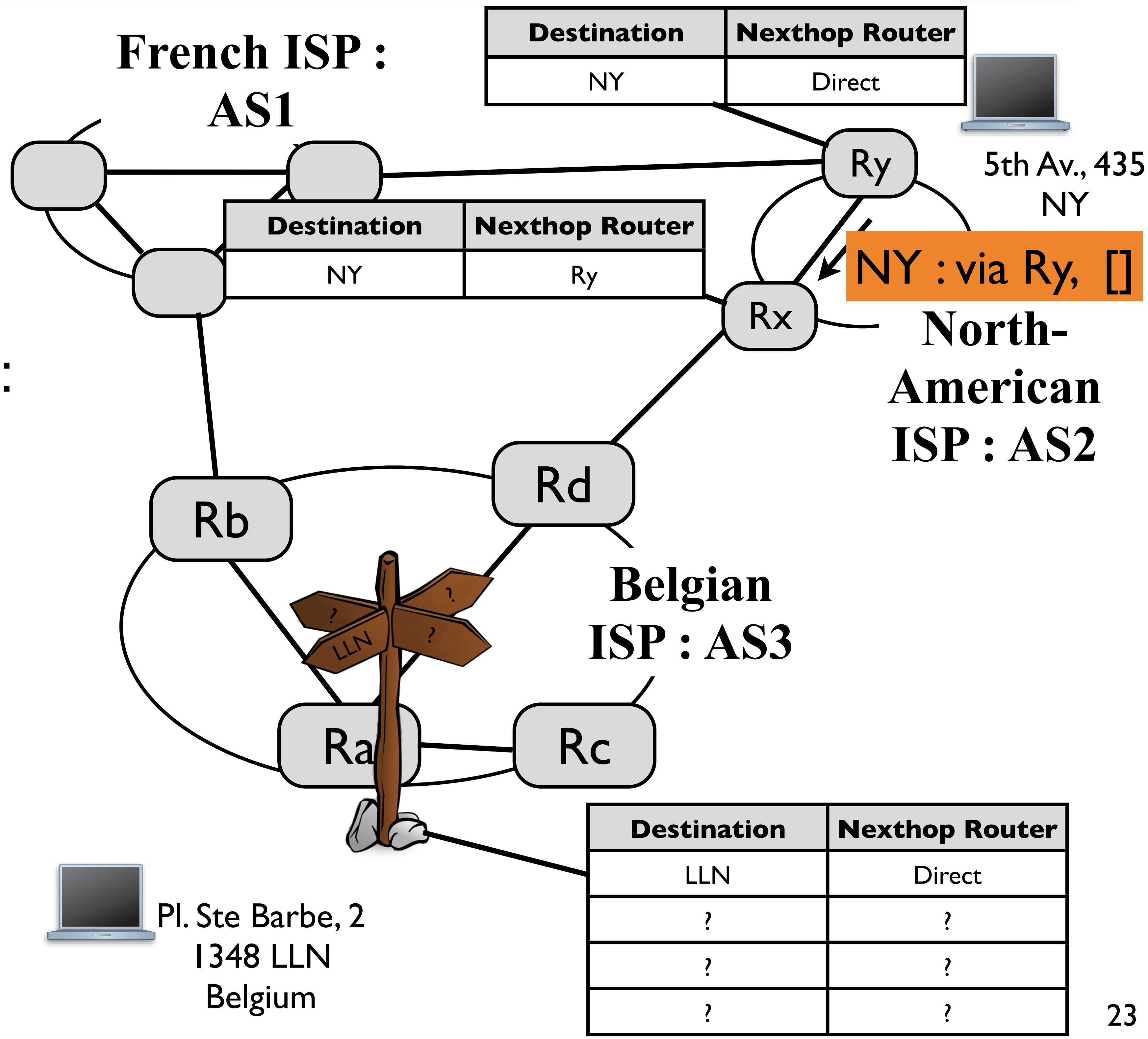
- An **AS Path** : List of ASes on the path to the destination
- A **Nexthop** : The router to which packets for that destination must be forwarded
- Miscellaneous **BGP attributes** (ex : Local Preference)



The Border Gateway Protocol

Each AS advertises reachable destinations in BGP messages containing :

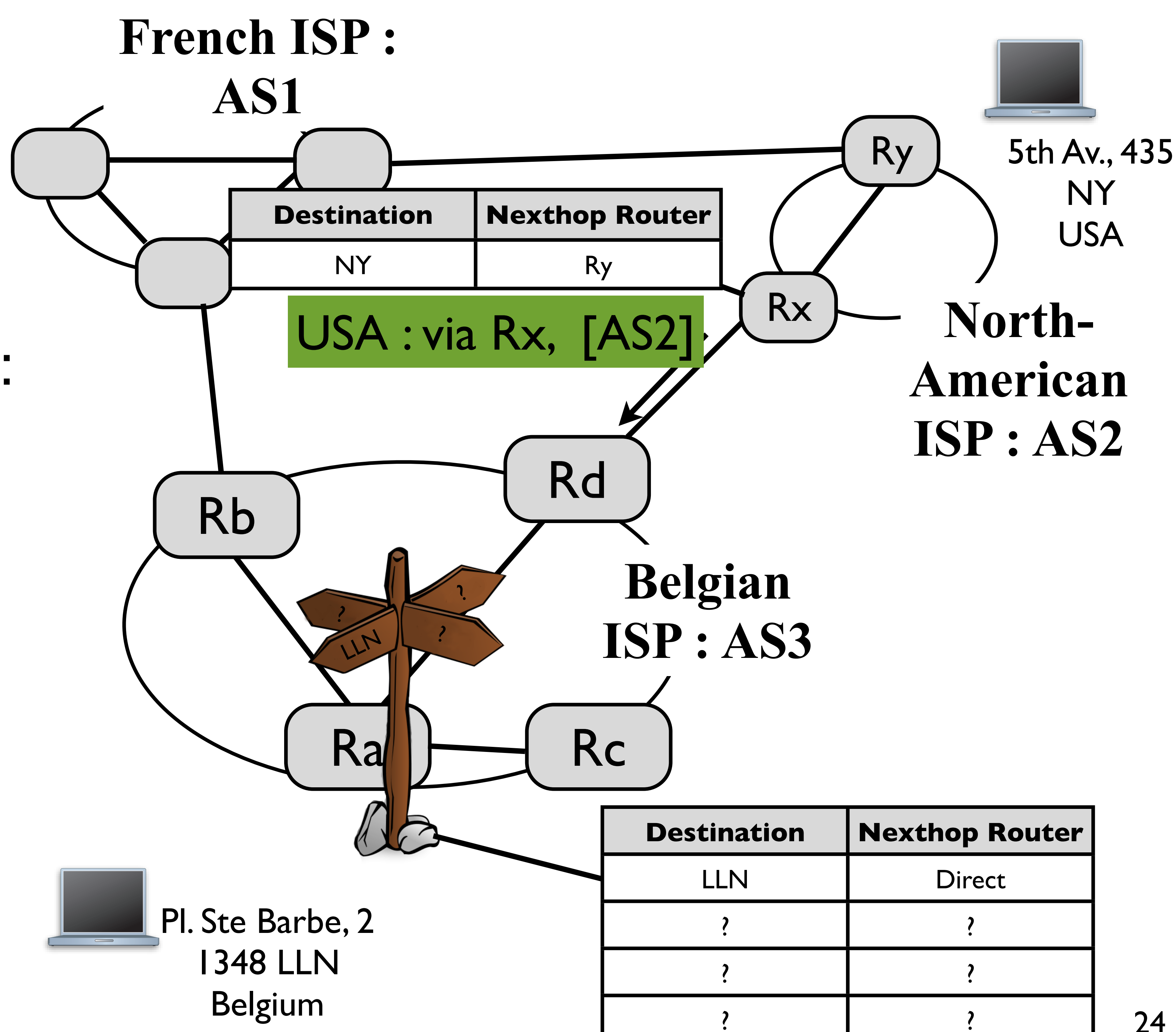
- An **AS Path** : List of ASes on the path to the destination
- A **Nexthop** : The router to which packets for that destination must be forwarded
- Miscellaneous **BGP attributes** (ex : Local Preference)



The Border Gateway Protocol

Each AS advertises reachable destinations in BGP messages containing :

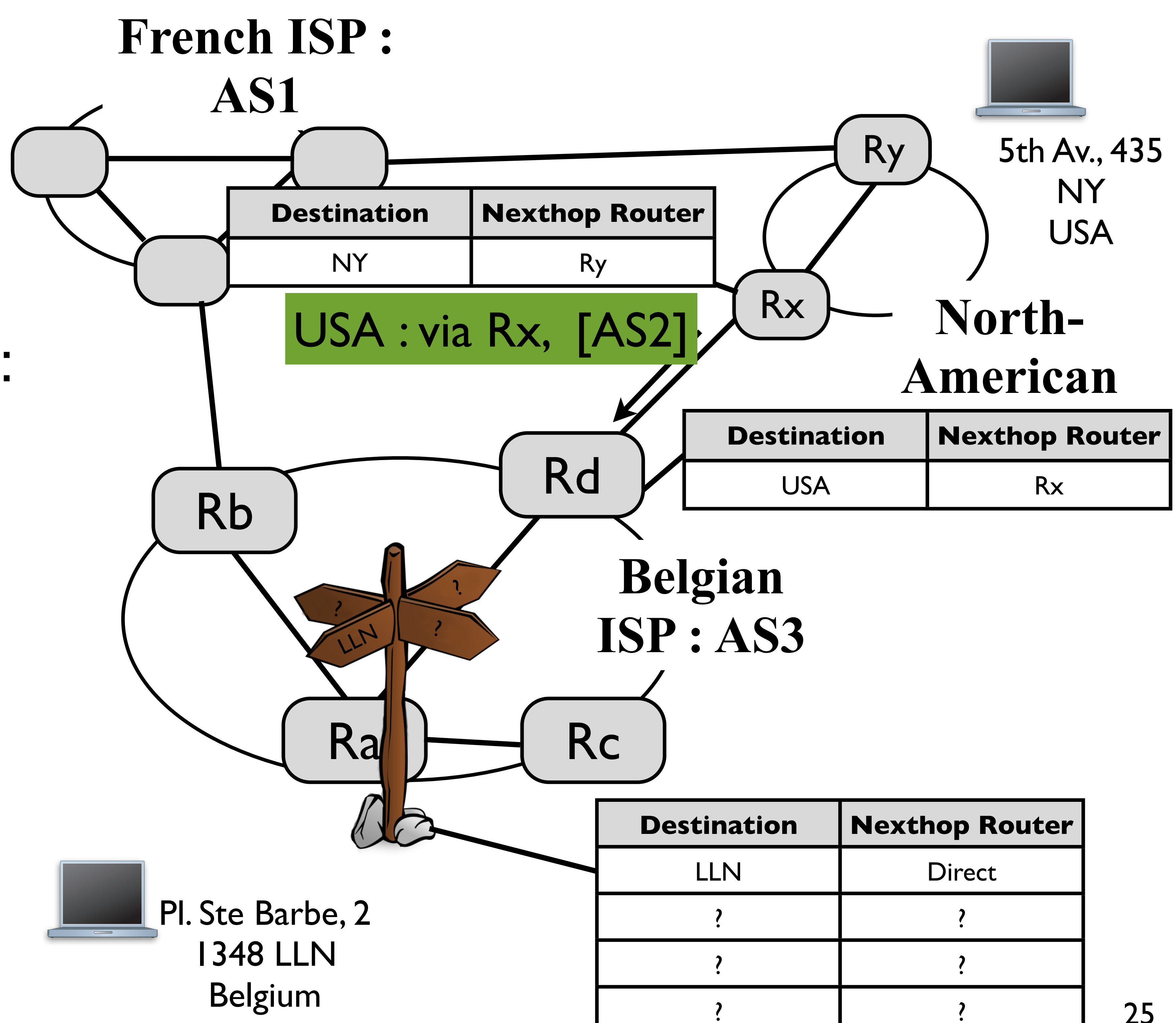
- An **AS Path** : List of ASes on the path to the destination
- A **Nexthop** : The router to which packets for that destination must be forwarded
- Miscellaneous **BGP attributes** (ex : Local Preference)



The Border Gateway Protocol

Each AS advertises reachable destinations in BGP messages containing :

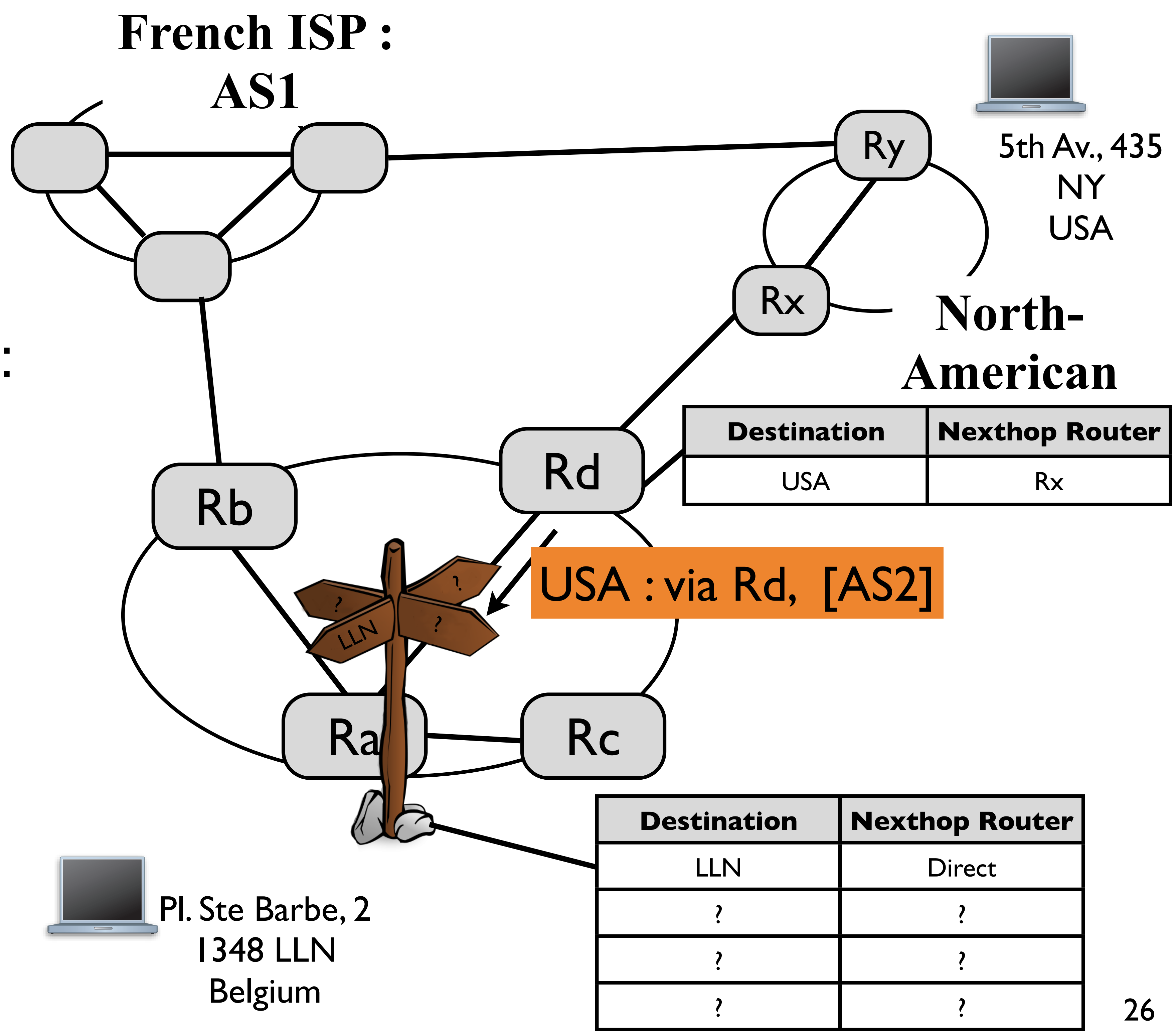
- An **AS Path** : List of ASes on the path to the destination
- A **Nexthop** : The router to which packets for that destination must be forwarded
- Miscellaneous **BGP attributes** (ex : Local Preference)



The Border Gateway Protocol

Each AS advertises reachable destinations in BGP messages containing :

- An **AS Path** : List of ASes on the path to the destination
- A **Nexthop** : The router to which packets for that destination must be forwarded
- Miscellaneous **BGP attributes** (ex : Local Preference)



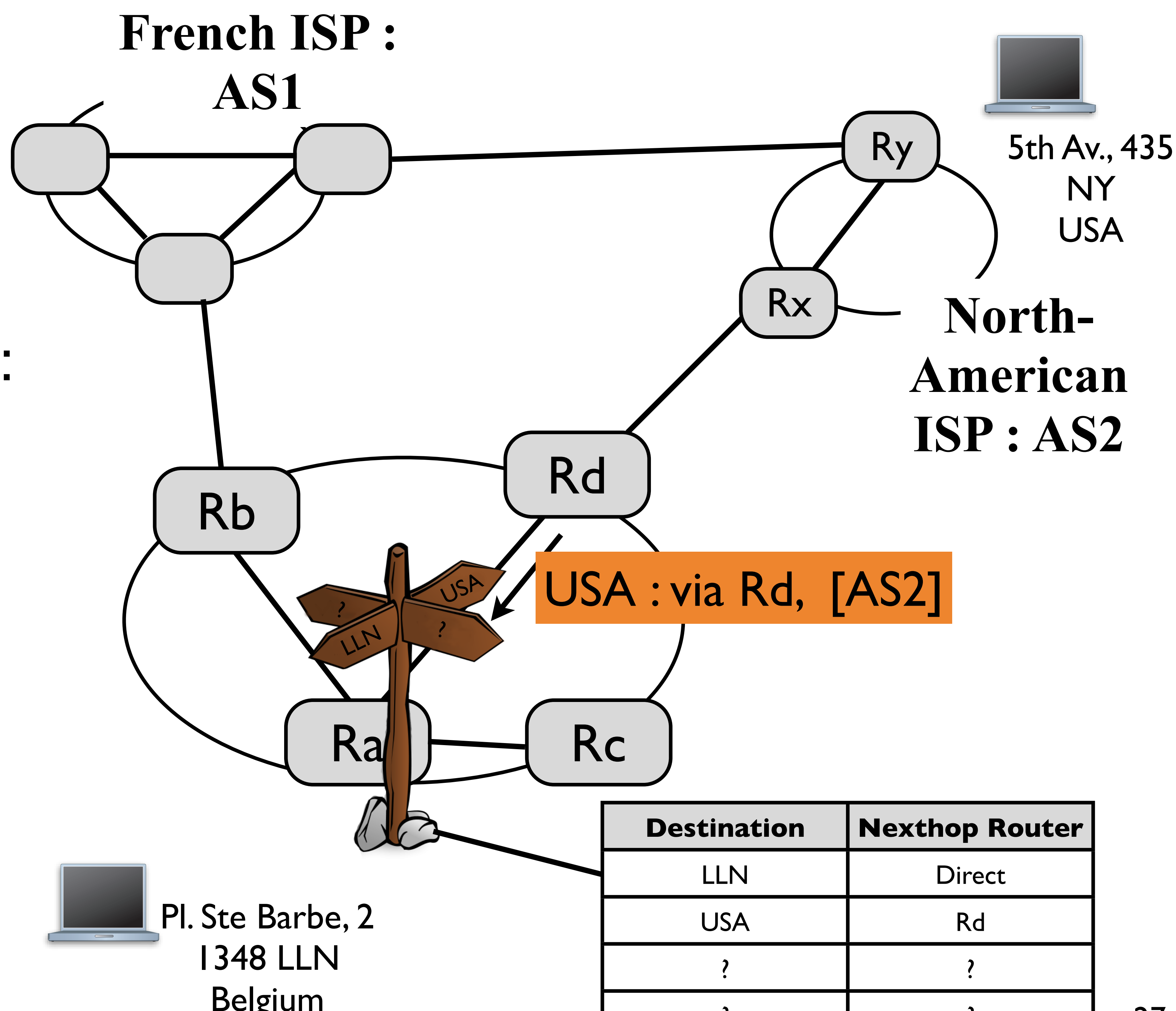
Destination	Nexthop Router
USA	Rx

Destination	Nexthop Router
LLN	Direct
?	?
?	?
?	?

The Border Gateway Protocol

Each AS advertises reachable destinations in BGP messages containing :

- An **AS Path** : List of ASes on the path to the destination
- A **Nexthop** : The router to which packets for that destination must be forwarded
- Miscellaneous **BGP attributes** (ex : Local Preference)

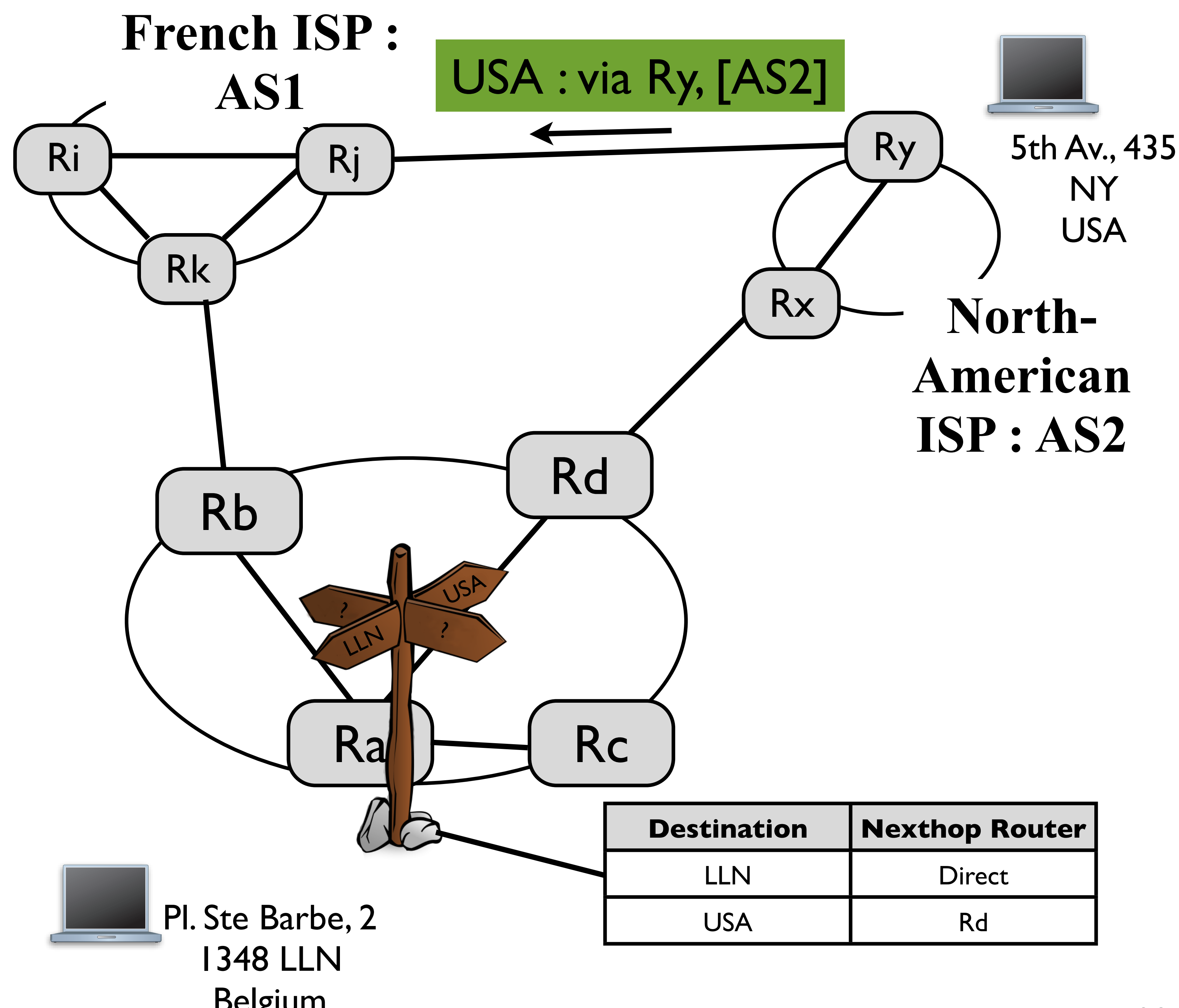


Destination	Nexthop Router
LLN	Direct
USA	Rd
?	?
?	?

Availability of multiple paths

A destination can be reachable via several Nexthops

- Ex : USA also reachable via AS 1 (French ISP)

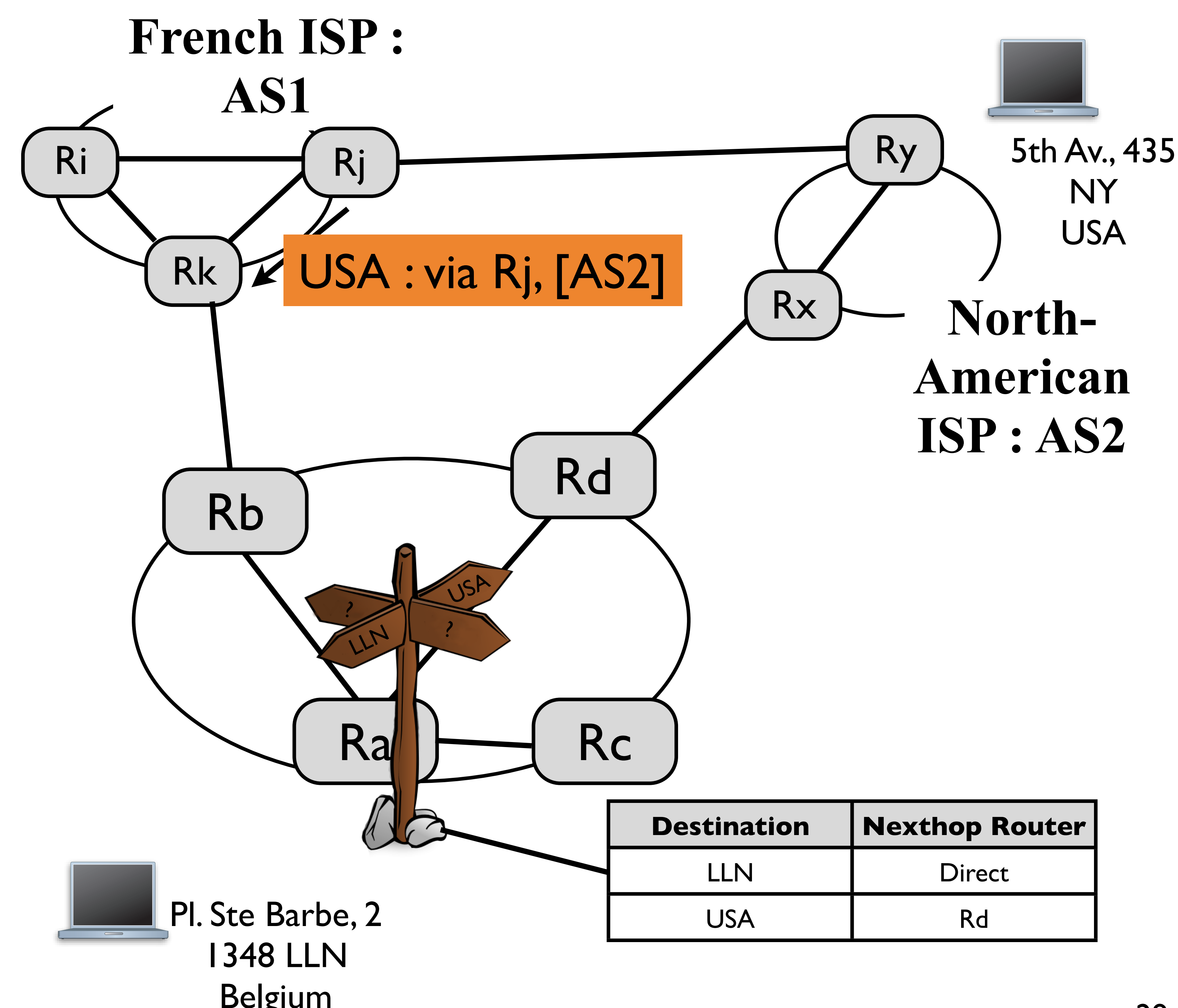


Destination	Nexthop Router
LLN	Direct
USA	Rd

Availability of multiple paths

A destination can be reachable via several Nexthops

- Ex : USA also reachable via AS 1 (French ISP)

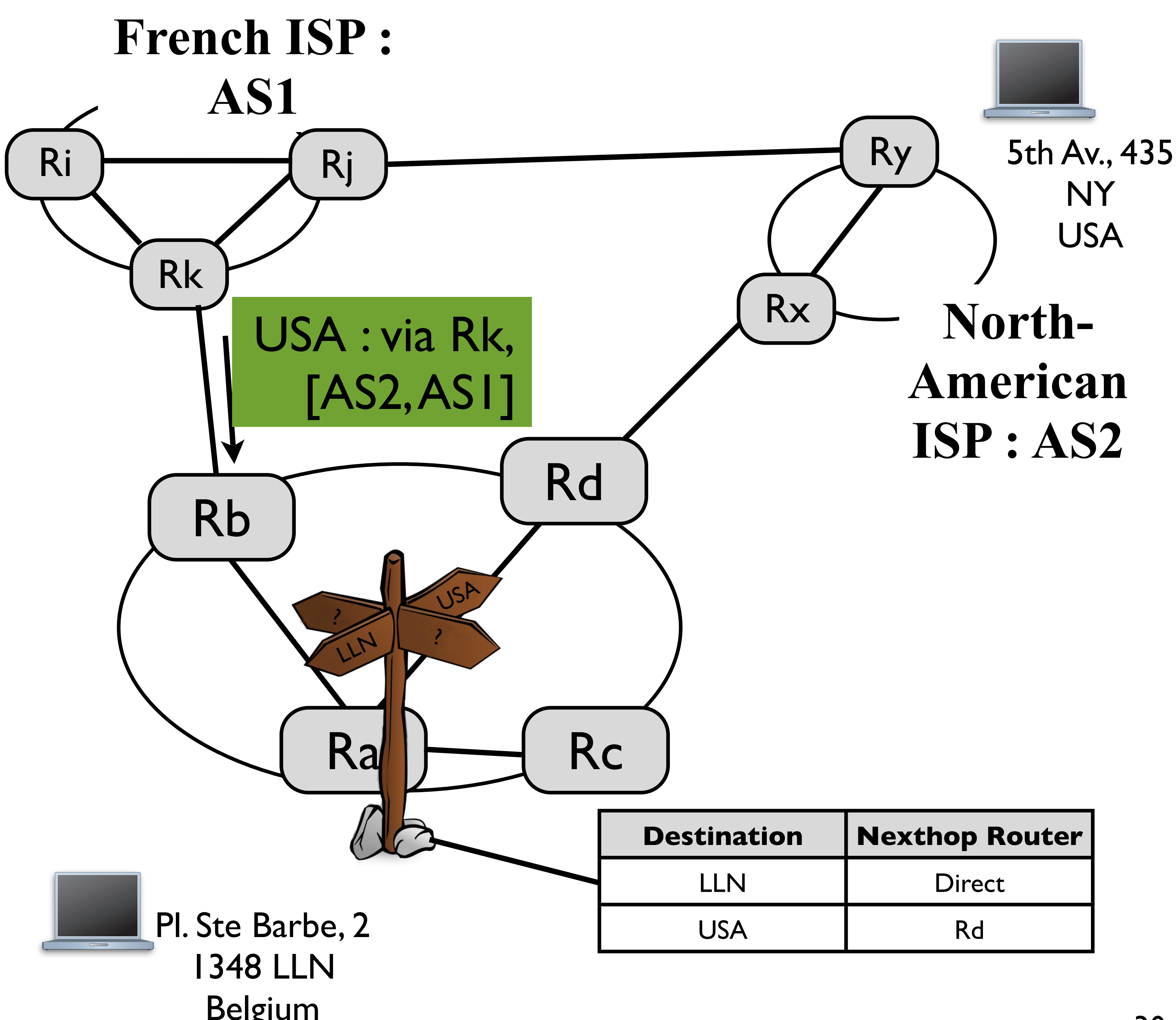


Destination	Nexthop Router
LLN	Direct
USA	Rd

Availability of multiple paths

A destination can be reachable via several Nexthops

- Ex : USA also reachable via AS 1 (French ISP)

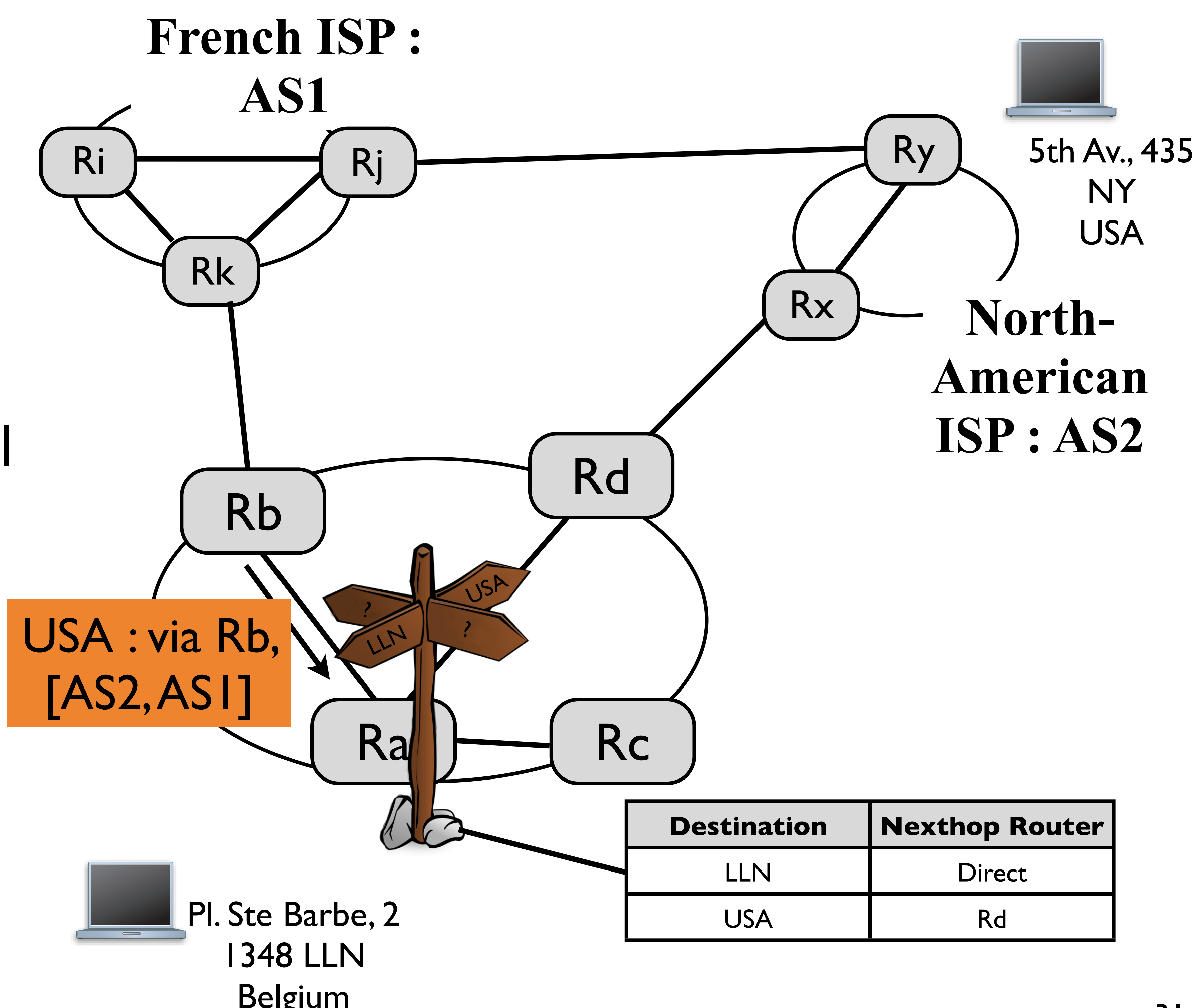


Destination	Nexthop Router
LLN	Direct
USA	Rd

Availability of multiple paths

A destination can be reachable via several Nexthops

- Ex : USA also reachable via AS 1 (French ISP)

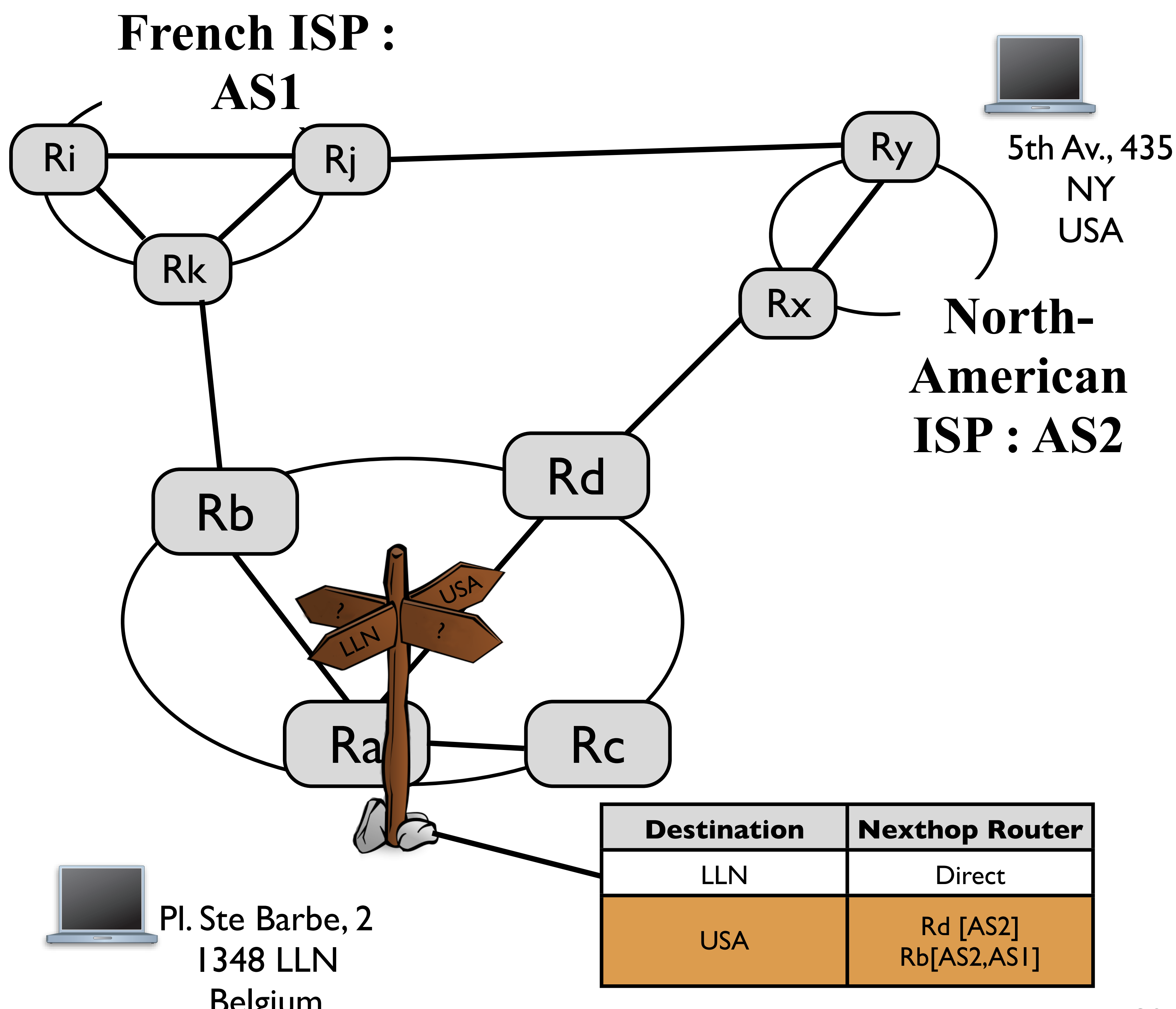


Destination	Nexthop Router
LLN	Direct
USA	Rd

Availability of multiple paths

A destination can be reachable via several Nexthops

- Ex : USA also reachable via AS 1 (French ISP)

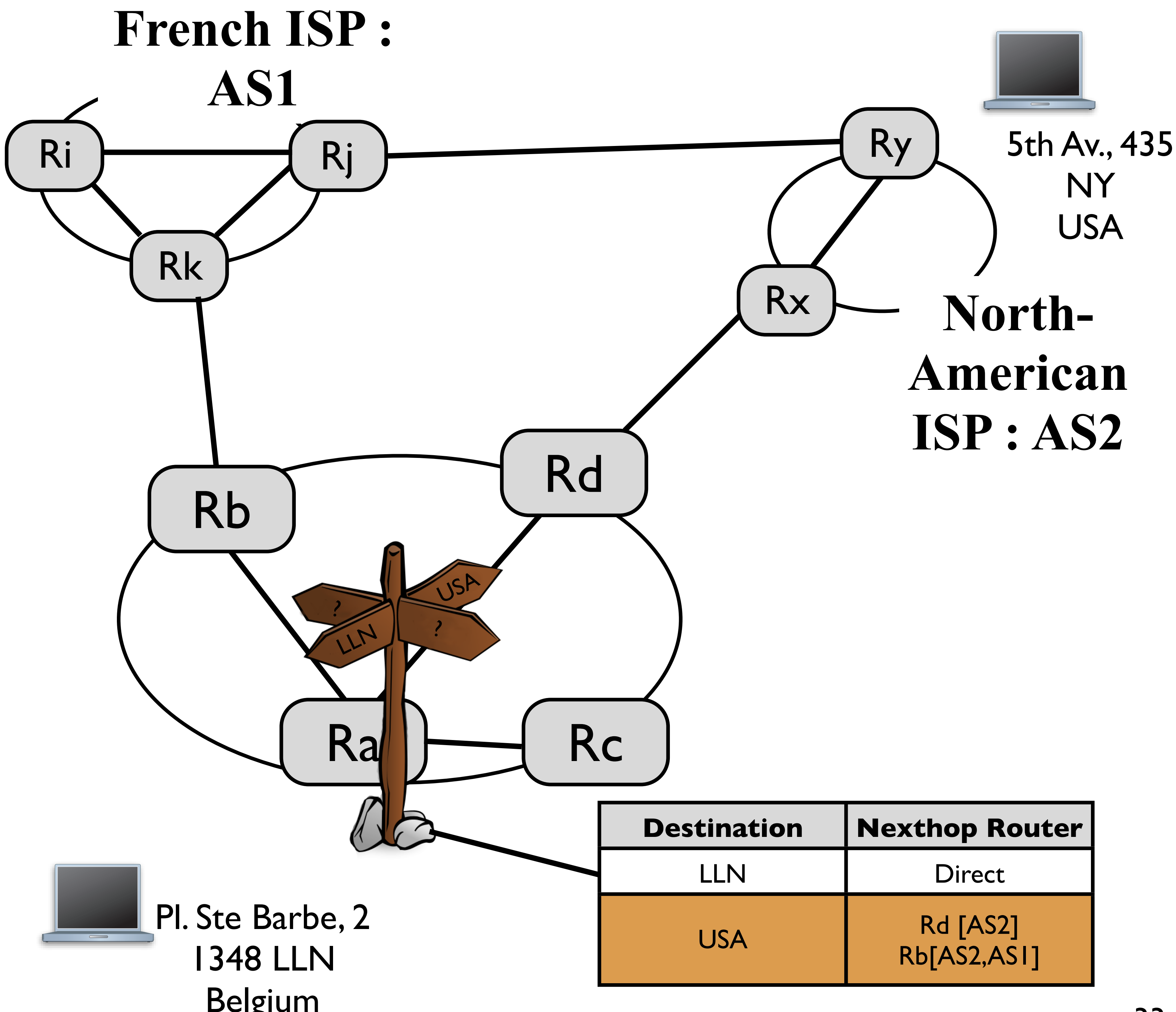


Destination	Nexthop Router
LLN	Direct
USA	Rd [AS2] Rb[AS2,AS1]

Best Path selection

Each router selects a Best Path for each destination : **Decision Process**

- Higher Local-Pref
- Shortest AS-Path
- ...
- Closest Nexthop
- ...



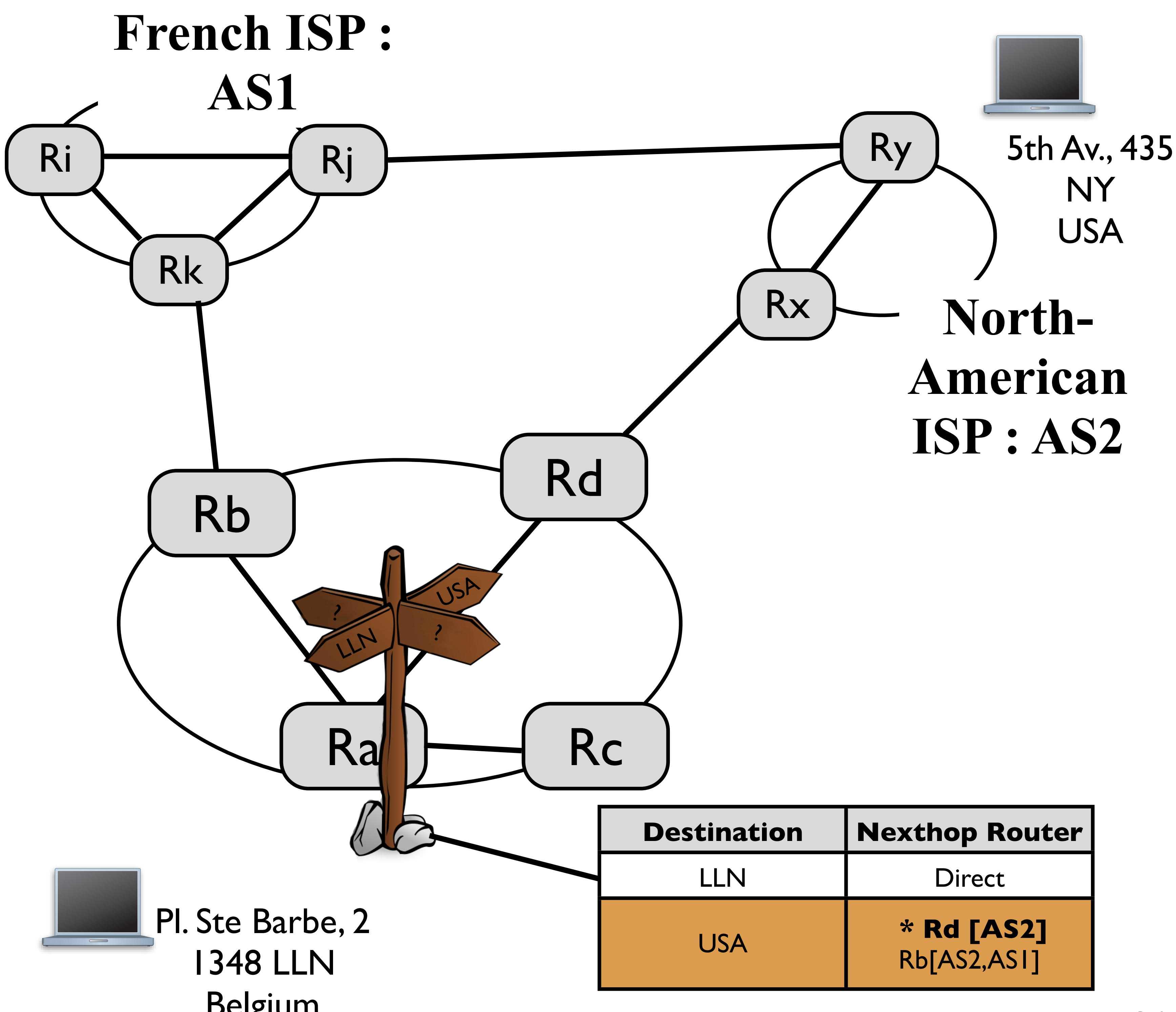
Destination	Nexthop Router
LLN	Direct
USA	Rd [AS2] Rb[AS2,AS1]

Best Path selection

Each router selects a Best Path for each destination :

Decision Process

- Higher Local-Pref
- **Shortest AS-Path**
- ...
- Closest Nexthop
- ...



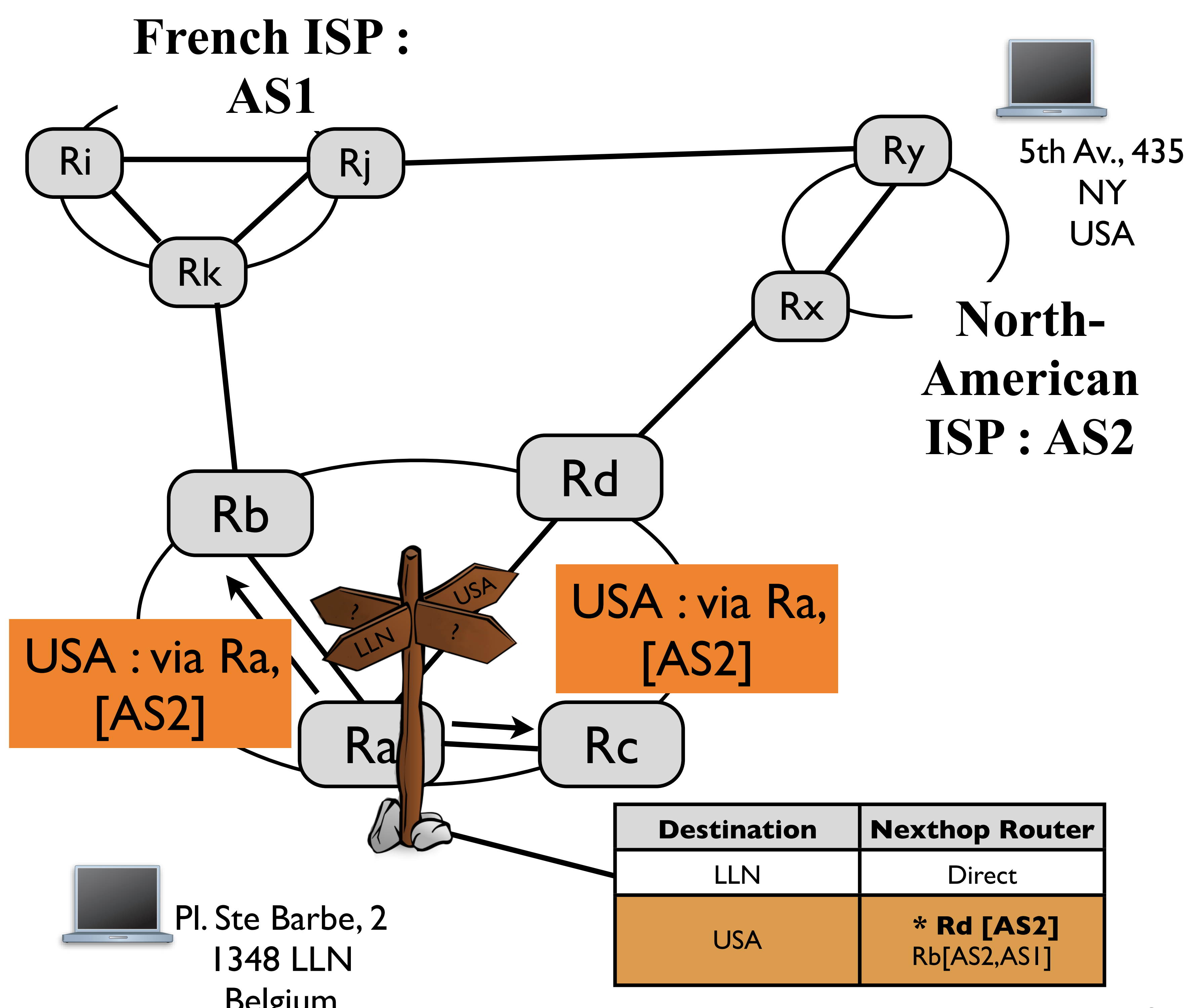
Destination	Nexthop Router
LLN	Direct
USA	*** Rd [AS2]** Rb[AS2,AS1]

Best Path selection

Each router selects a Best Path for each destination :

Decision Process

Only the **best path** is advertised to BGP neighbors



Destination	Nexthop Router
LLN	Direct
USA	*** Rd [AS2]** Rb[AS2,AS1]

Outline

I. Background on Internet routing

II. Problem Statement

III. Improving iBGP with single path advertisement

IV. Improving iBGP with multiple paths advertisement : BGP Add-Paths

BGP issues

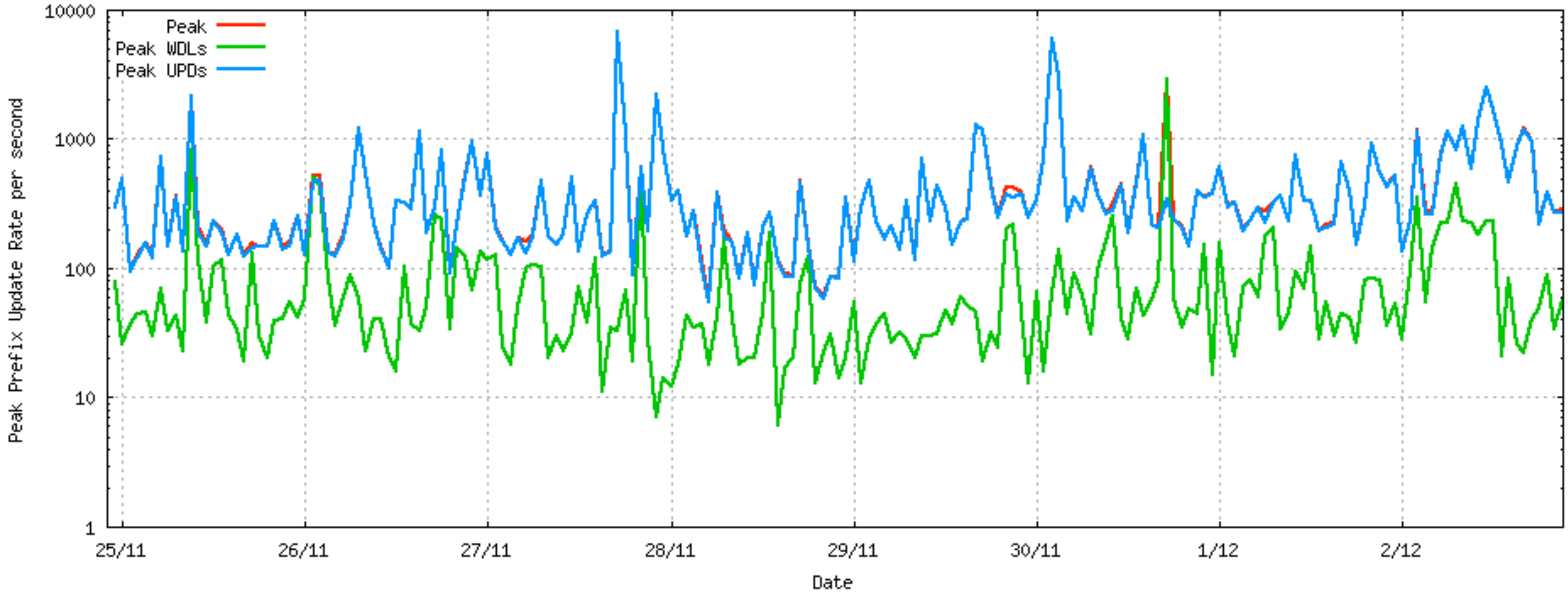
BGP is a huge distributed system

- > 35,000 ASes, each with between 1 and several thousands of routers
- 300,000 destinations
- And still growing...

The system suffers from scalability issues

- Ex : High rate of BGP messages => pressure on routers' CPU

BGP Churn



Hourly peak of per-second updated and withdrawn prefix rate on a single eBGP session

Source : www.potaroo.net

BGP issues

Churn : High number BGP messages exchanged

Slow and “noisy” failure recovery

Convergence anomalies

Misconfigurations

The weaknesses of current iBGP

Usually, several paths are available for a given destination at the borders of an ISP

- Via several neighbor ASes, as shown in the previous example
- Via several links with a given neighbor AS [Merindol et al., 2009]

But with current iBGP, lots of routers inside an ISP **only** receive **one path per destination**

The weaknesses of current iBGP

Consequences of a **lack of diversity** in iBGP

- Transient iBGP failures and slow convergence [Wang05]
- Lack of failure isolation : Local event seen outside the AS
- Convergence anomalies in iBGP as in eBGP [Griffin02][Griffin02a]

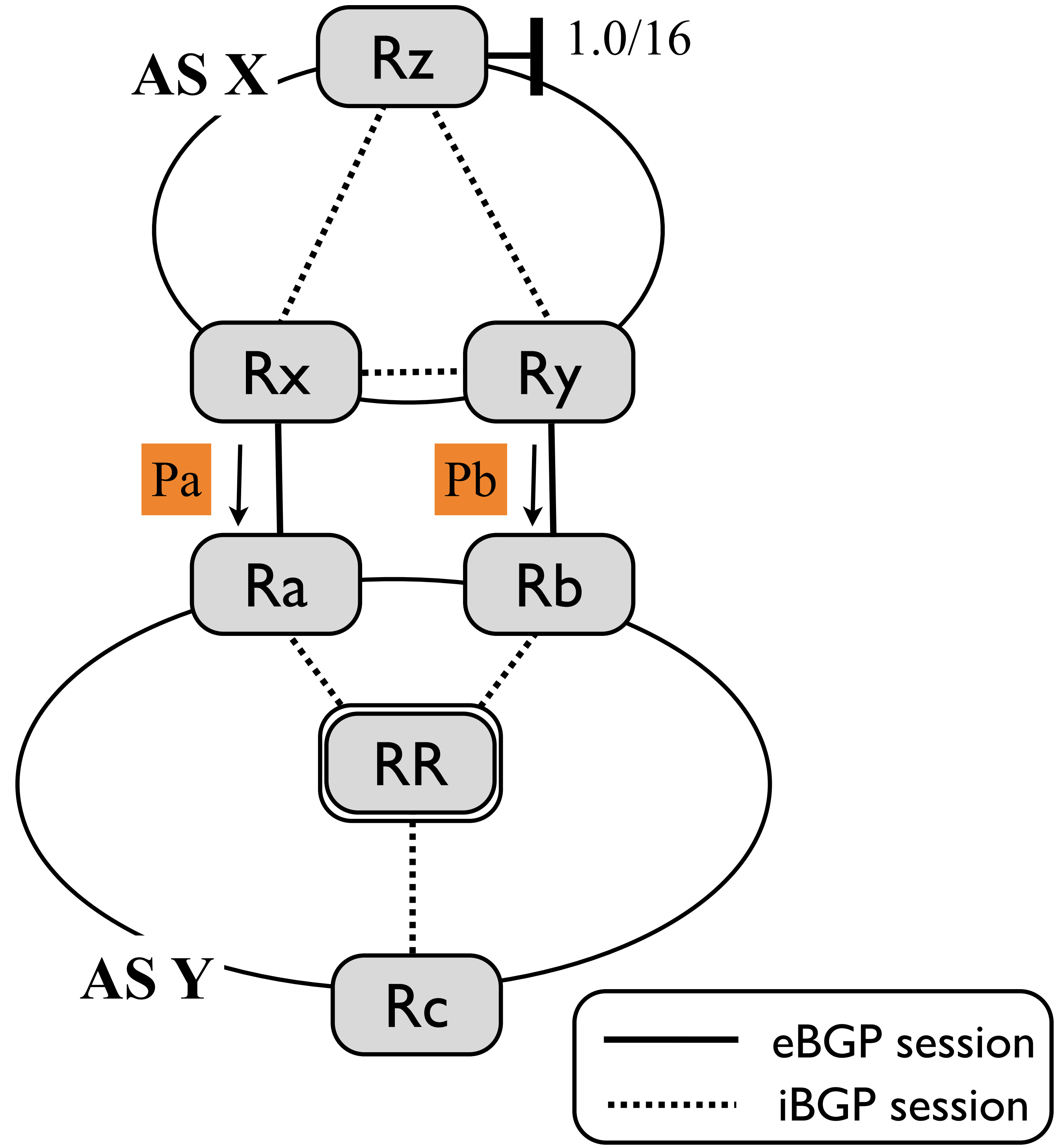
Misconfigurations

=> iBGP contributes to BGP instabilities

Path Diversity in ISPs

In this example :

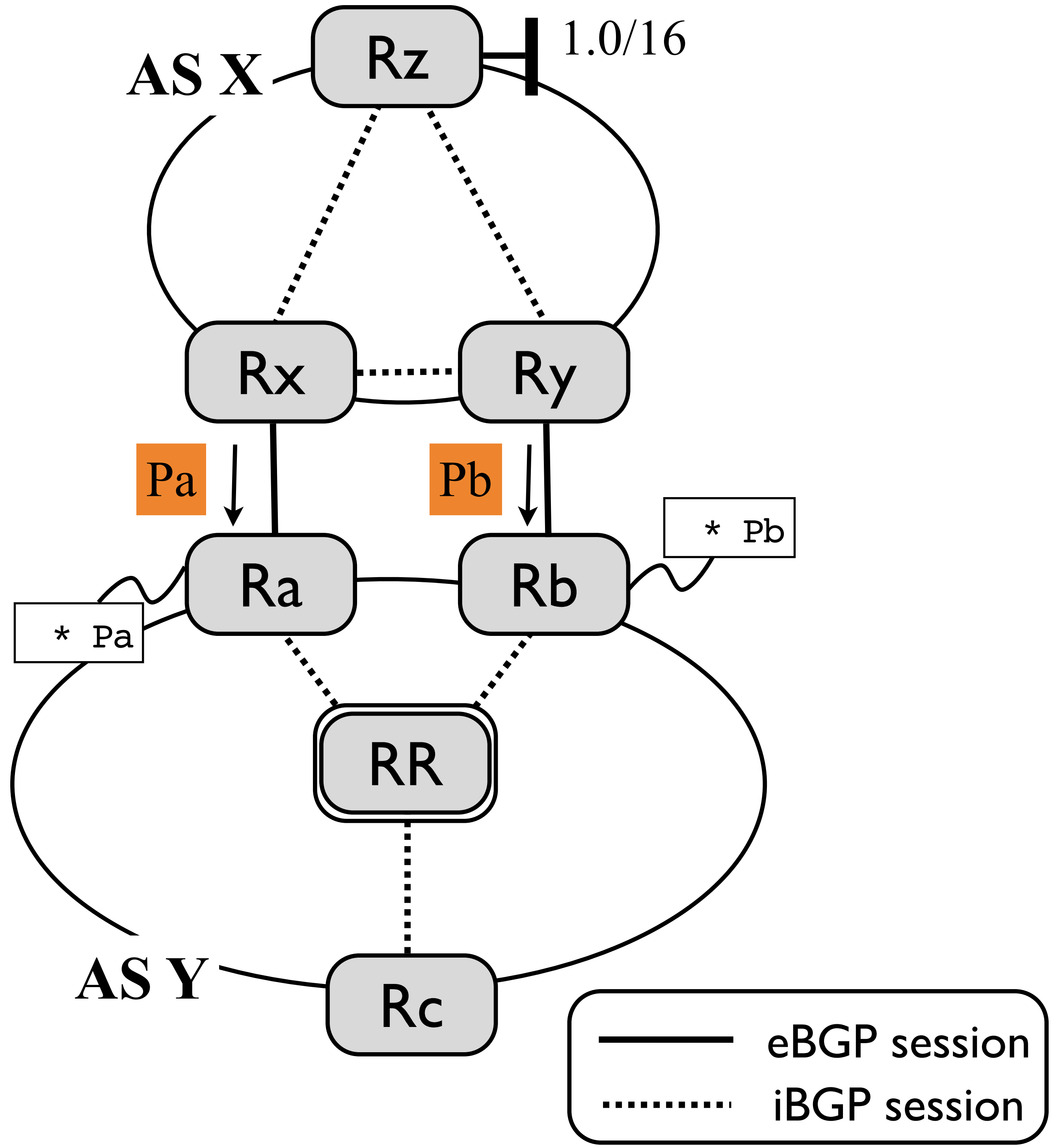
- Two ISPs : ASX and ASY
- One destination advertised by AS X
- Two links between AS X and AS Y
- AS Y receives a path to the destination on both links : Two paths **Pa** and **Pb**



Path Diversity in ISPs

In this example :

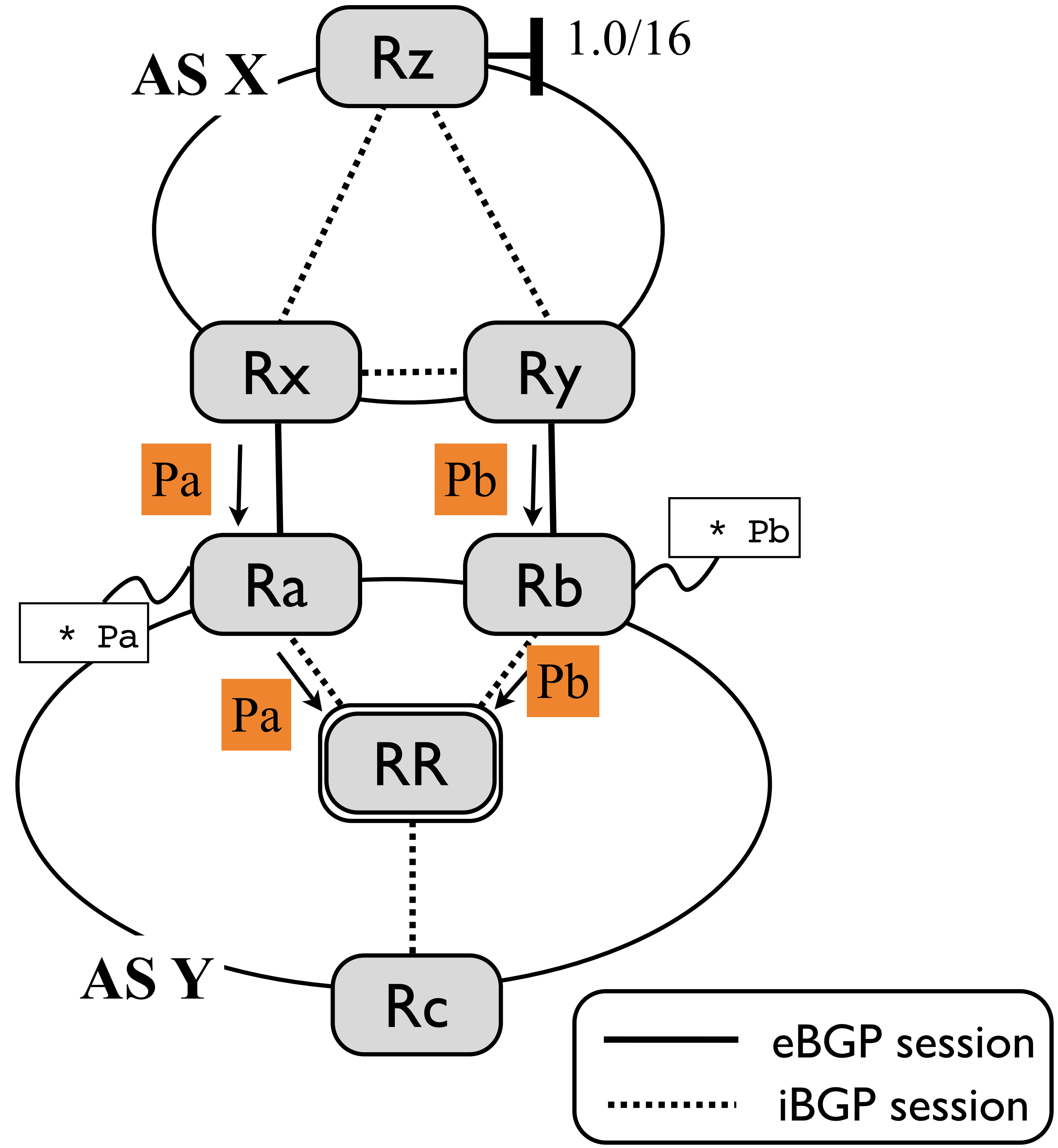
- Two ISPs : ASX and ASY
- One destination advertised by AS X
- Two links between AS X and AS Y
- AS Y receives a path to the destination on both links : Two paths **Pa** and **Pb**



Path Diversity in ISPs

Ra and Rb advertise the destination to their neighbors

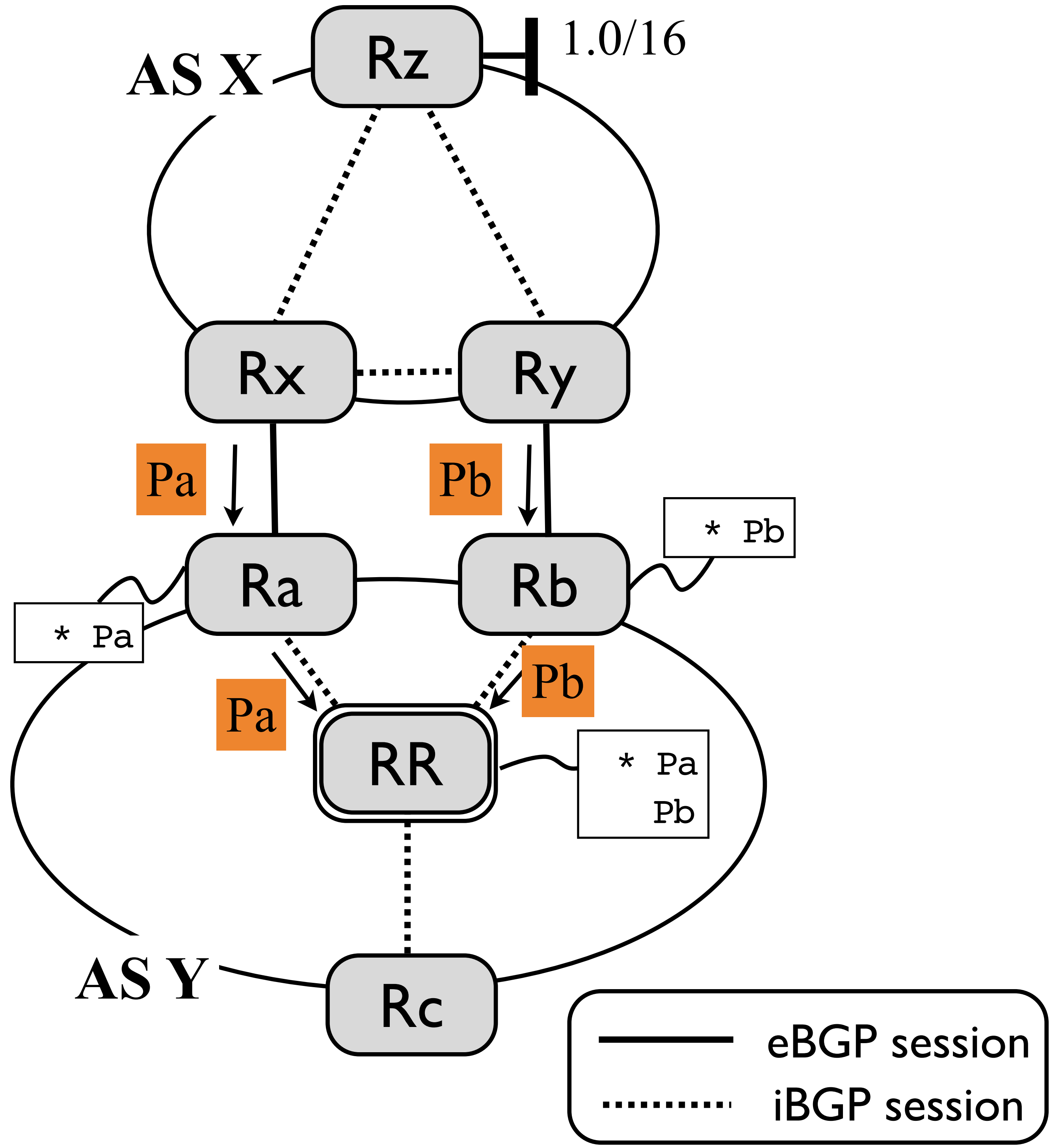
- eBGP neighbors (not shown on the figure)
- iBGP neighbors : In this case, RR



Path Diversity in ISPs

RR has two paths to the destination in its routing table

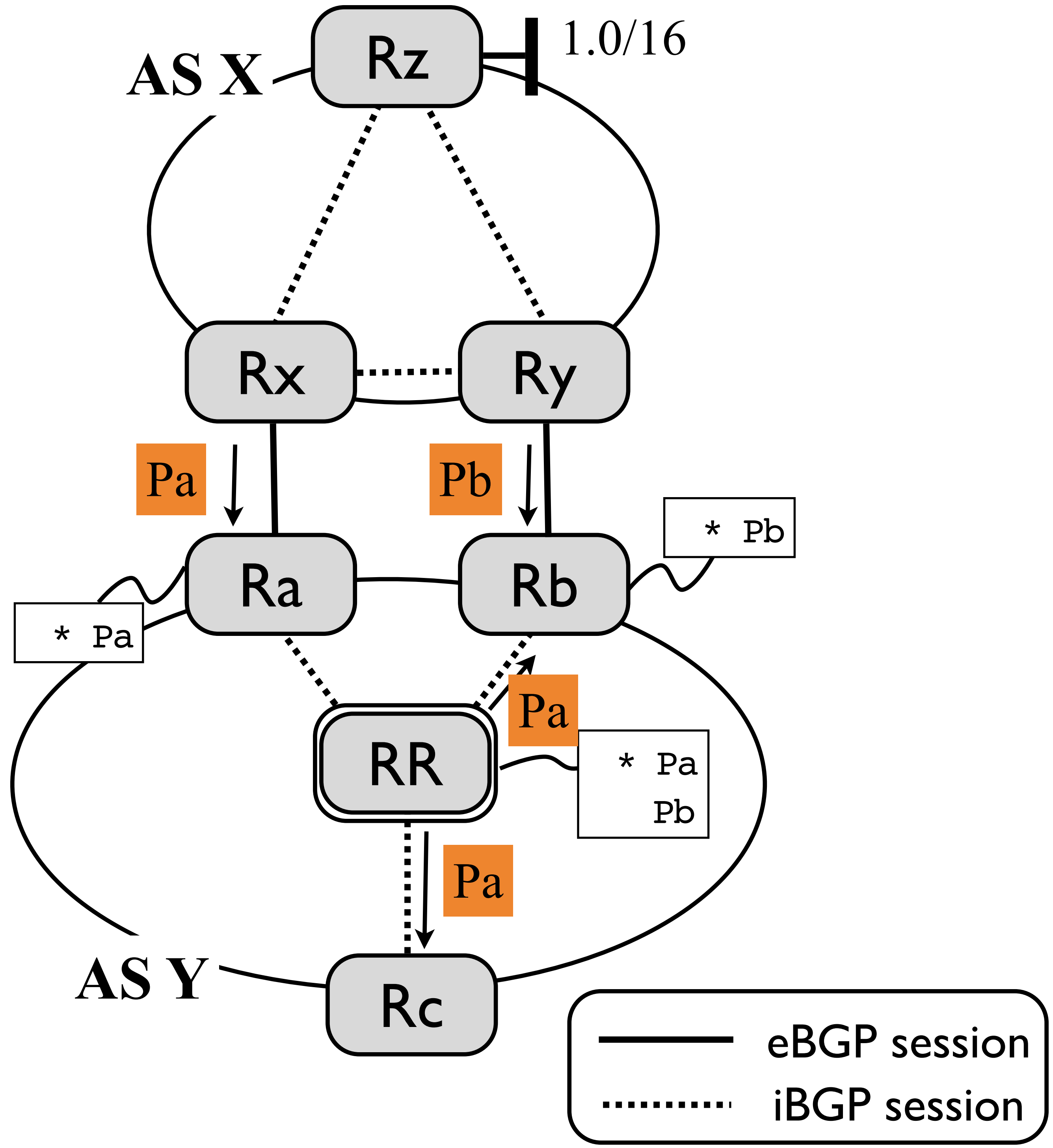
RR selects Pa as best (closest Nexthop)



Path Diversity in ISPs

Ra and Rb advertise the destination to RR

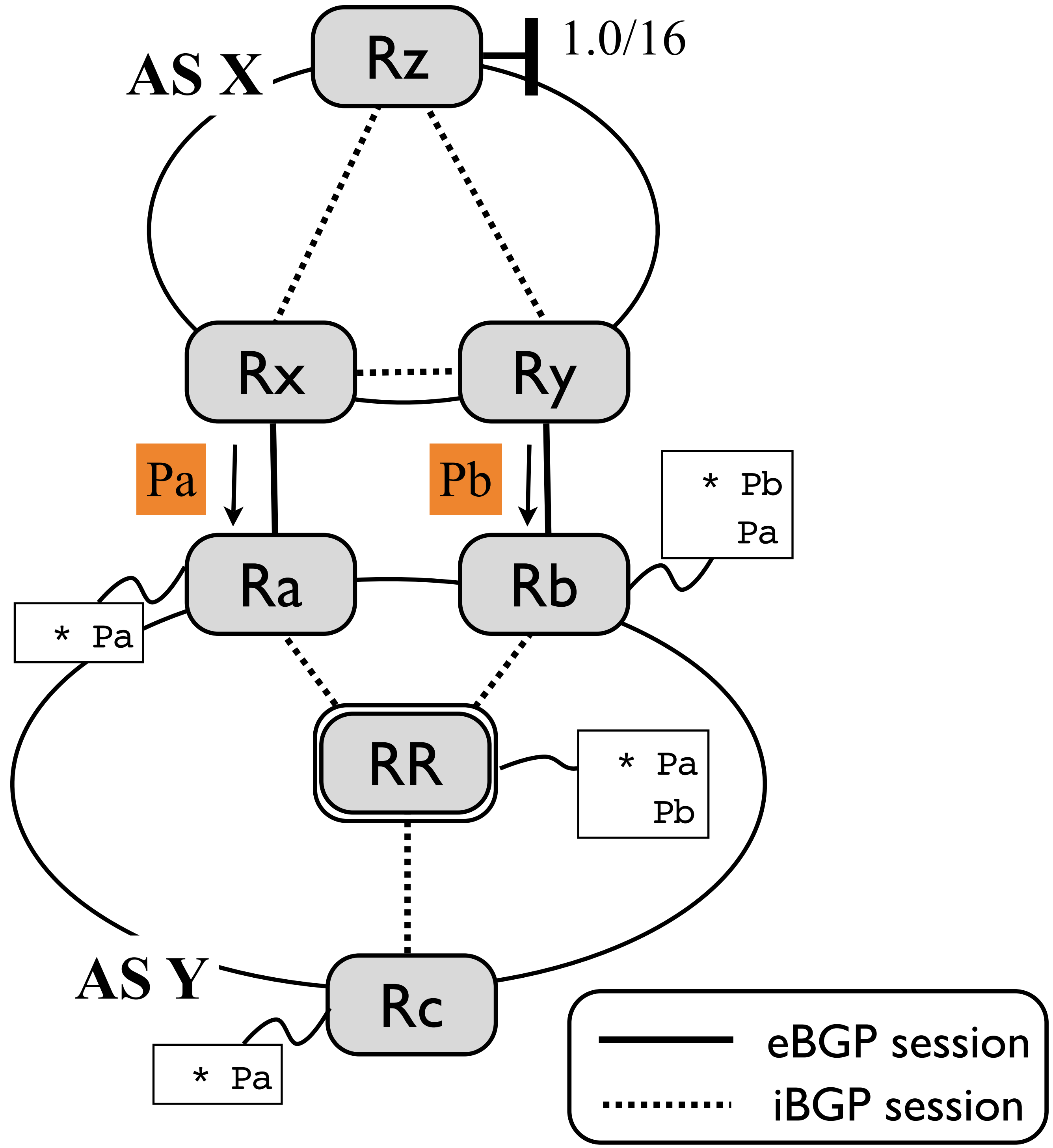
- RR has two paths in its routing table
- RR selects Pa as best (closest nexthop)
- RR advertises its best path Pa to its neighbors



Path Diversity in ISPs

Ra and Rb advertise the destination to RR

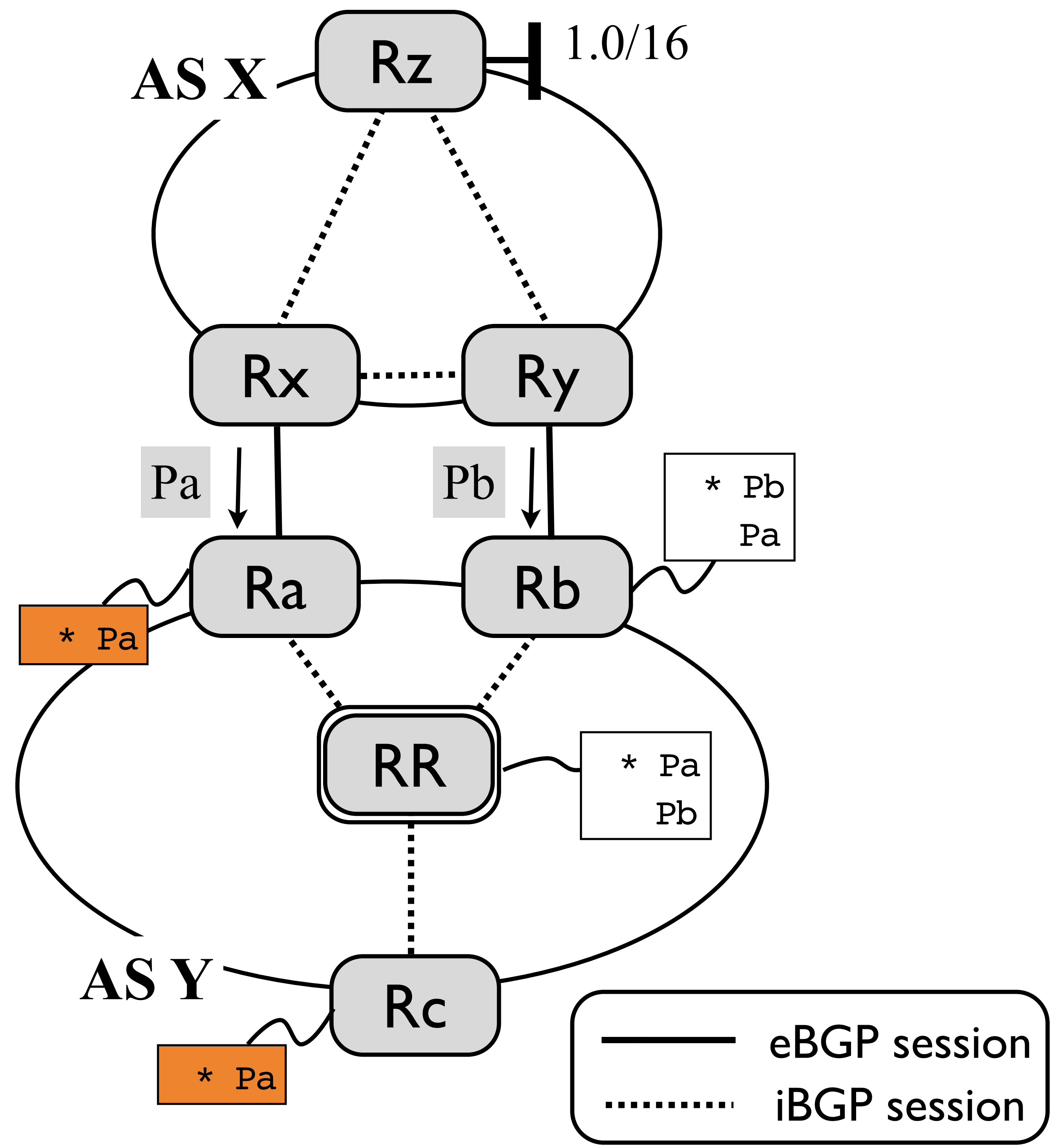
- RR has two paths in its routing table
- RR selects Pa as best (closest nexthop)
- RR advertises its best path Pa to its neighbors



Lack of Path Diversity in iBGP

Path diversity is not well propagated in iBGP

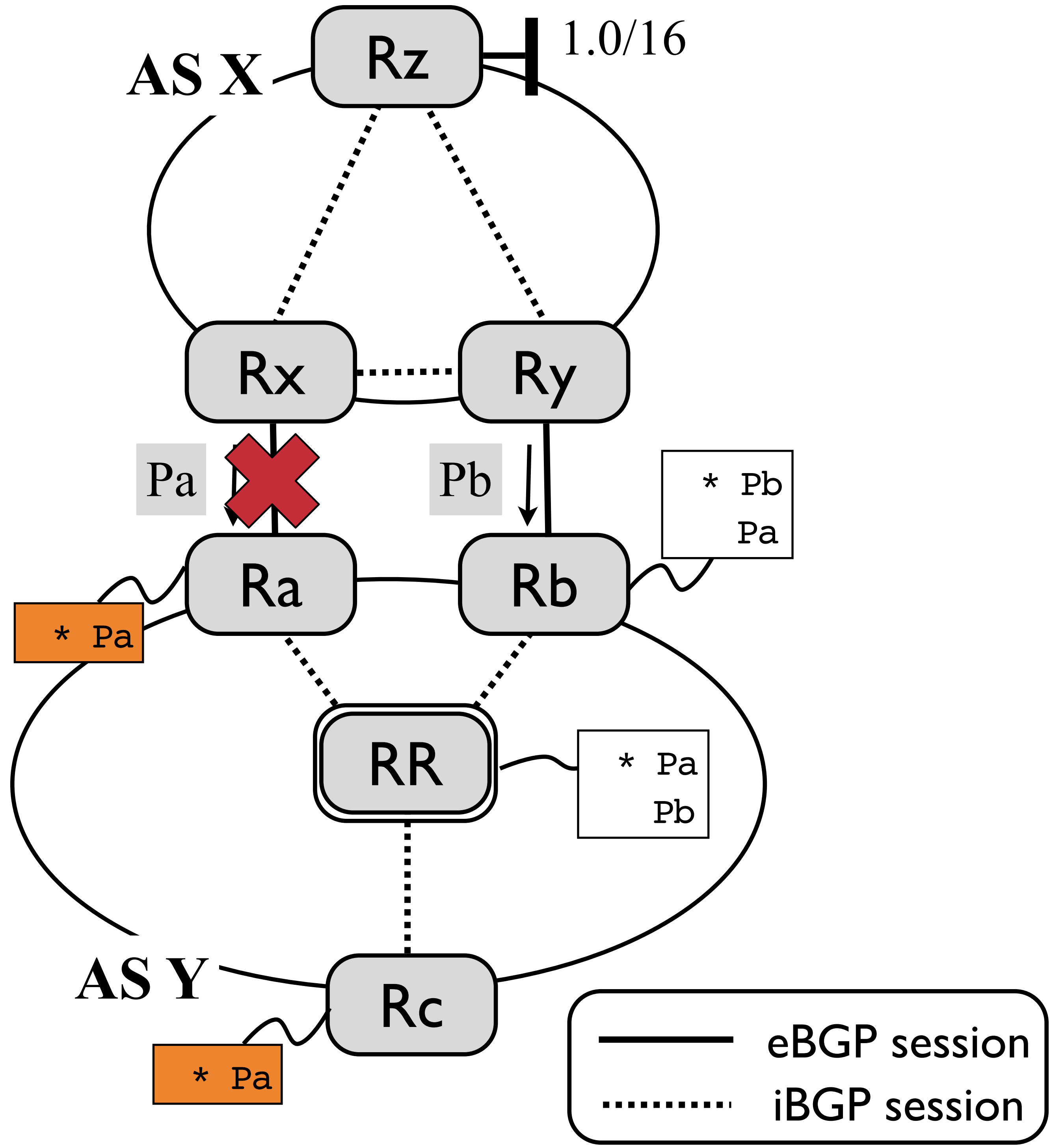
- Ra and Rc only know Pa
- But alternate path Pb is known in the AS!



Lack of Path Diversity in iBGP

Path diversity is not well propagated in iBGP

- Ra and Rc only know Pa
- But alternate path Pb is known in the AS!
- This is problematic when there is a link failure (ex : Link Rx-Ra)



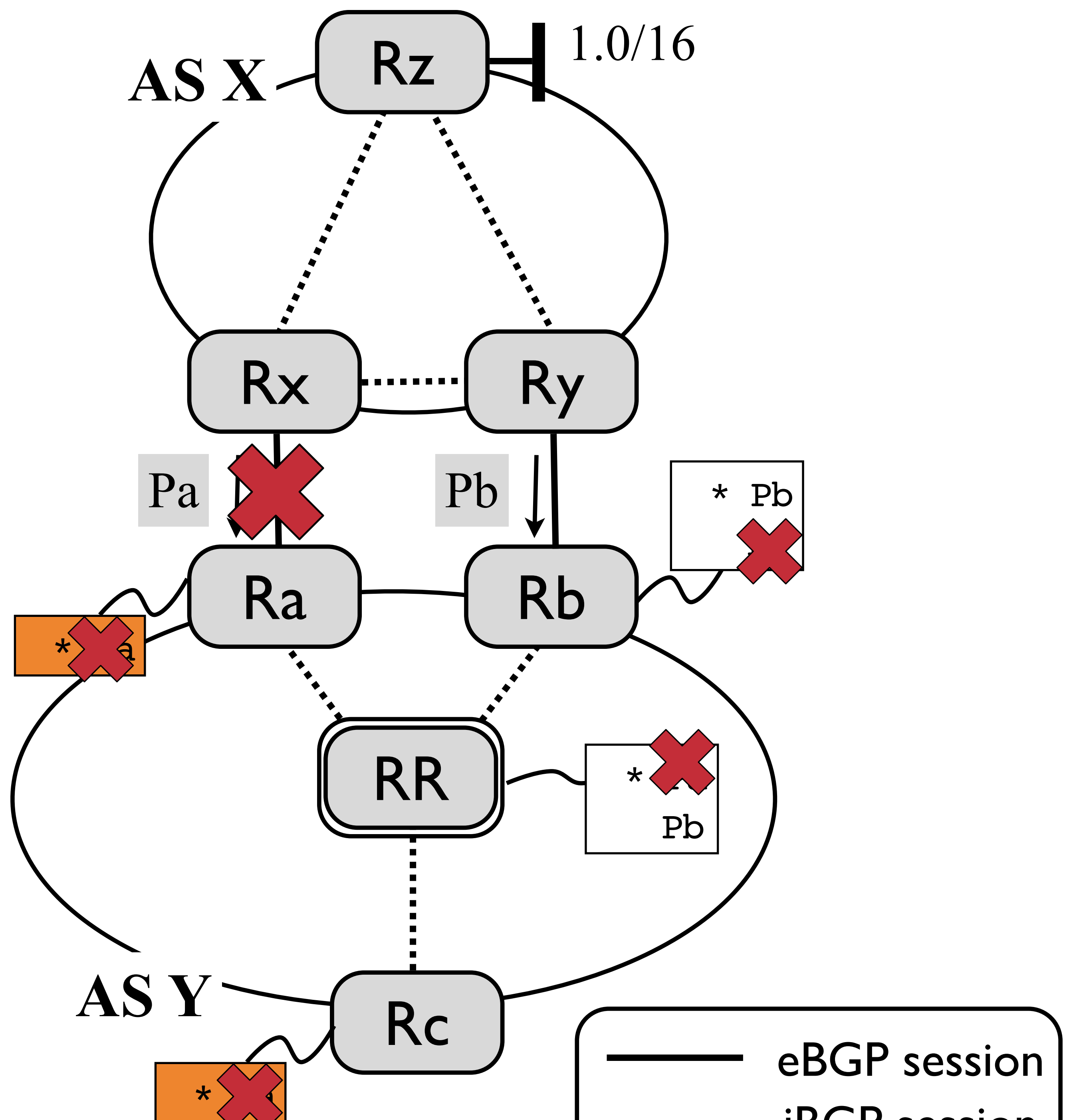
Path Diversity and Fast Recovery

Rb and RR have an alternate path

- They select the path Pb as best immediately, and packets are forwarded to ASX via this path

Ra and Rc loose their only path

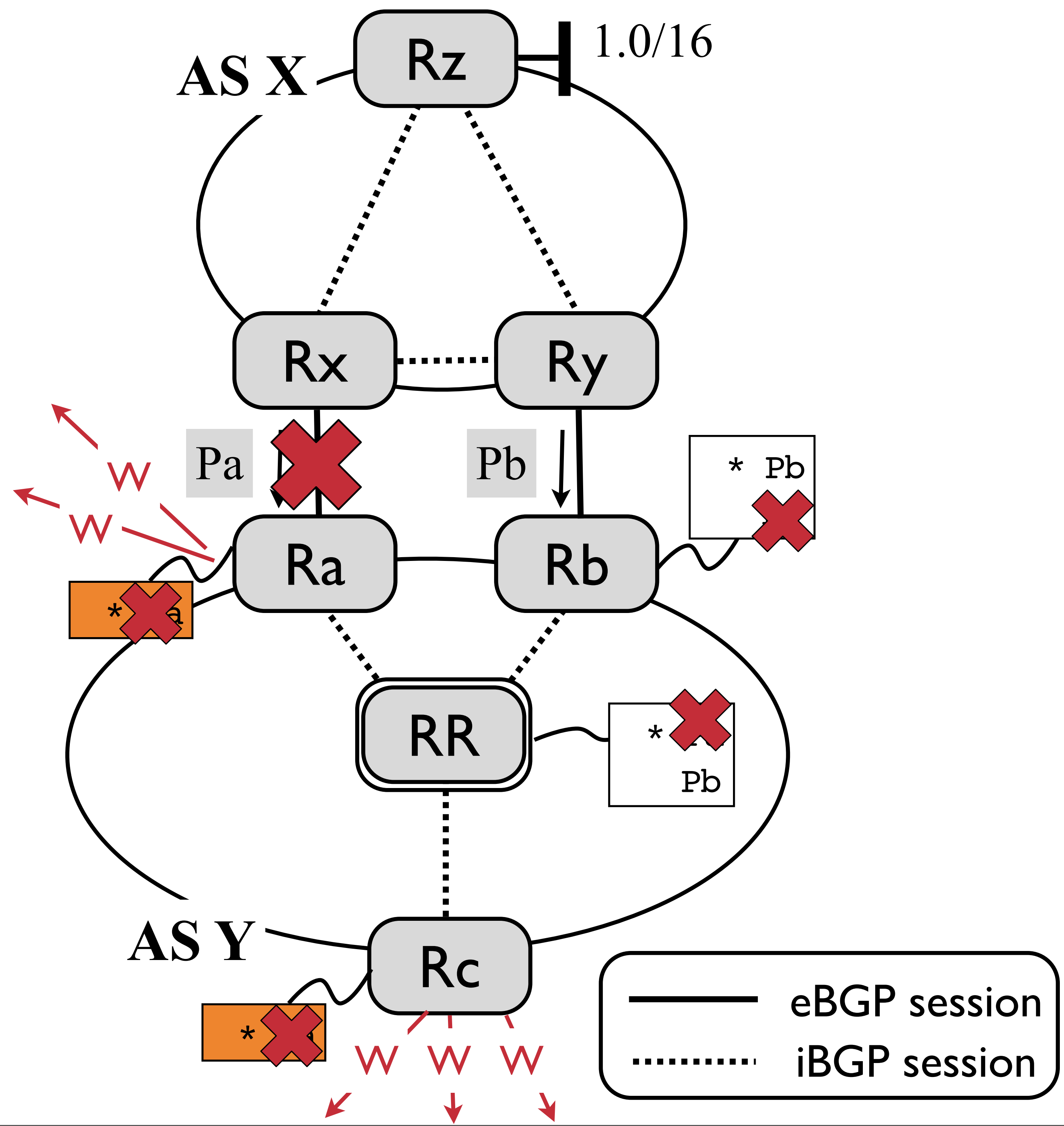
- They cannot forward the packets to the destination, and drop them
- **Traffic is lost** until RR advertises the alternate path Pb to Ra and Rc



Path Diversity and Failure Isolation

Ra and Rc must **withdraw** the destination advertised to eBGP neighbors

- Cascading effet : Reachability losses at other ASes
- Unnecessary BGP messages
- **The failure is propagated to the global Internet**



Evaluation of path diversity in a large ISP

Network : Large ISP, > 100 routers, worldwide connectivity

Data :

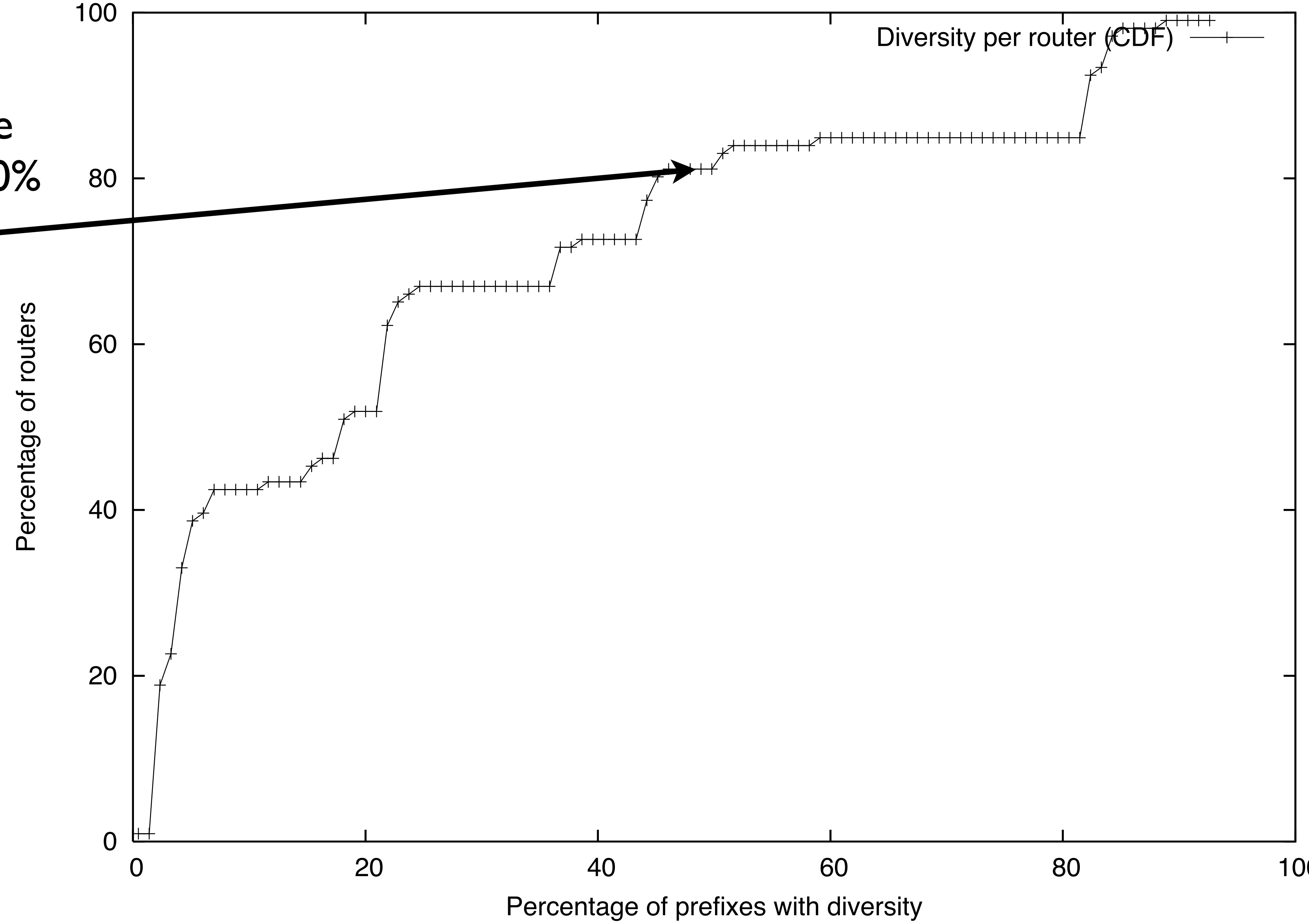
- Model of the topology
- Paths advertised in eBGP

Methodology :

- Simulate the propagation of BGP messages in the ISP (C-BGP simulator)
- Count the number of routers with/without path diversity for each destination

Evaluation of path diversity in a large ISP

80% of the routers have path diversity for max. 50% of the destinations



Goal of the thesis

Goal of the thesis : Improve iBGP to provide path diversity to all routers of an ISP

Outline

I. Background on Internet routing

II. Problem Statement

III. Improving iBGP with single path advertisement

IV. Improving iBGP with multiple paths advertisement : BGP Add-Paths

Outline

I. Background on Internet routing

II. Problem Statement

III. Improving iBGP with single path advertisement

- Overview of the three proposed improvements
- Focus on liBGP

IV. Improving iBGP with multiple paths advertisement : BGP Add-Paths

Overview of proposed improvements

IiBGP

- Providing on-demand diversity with lightweight iBGP sessions
- Path diversity, Auto-Config, Failure isolation

AiBGP [V.Van den Schrieck et al., CoNEXT Student workshop'07]

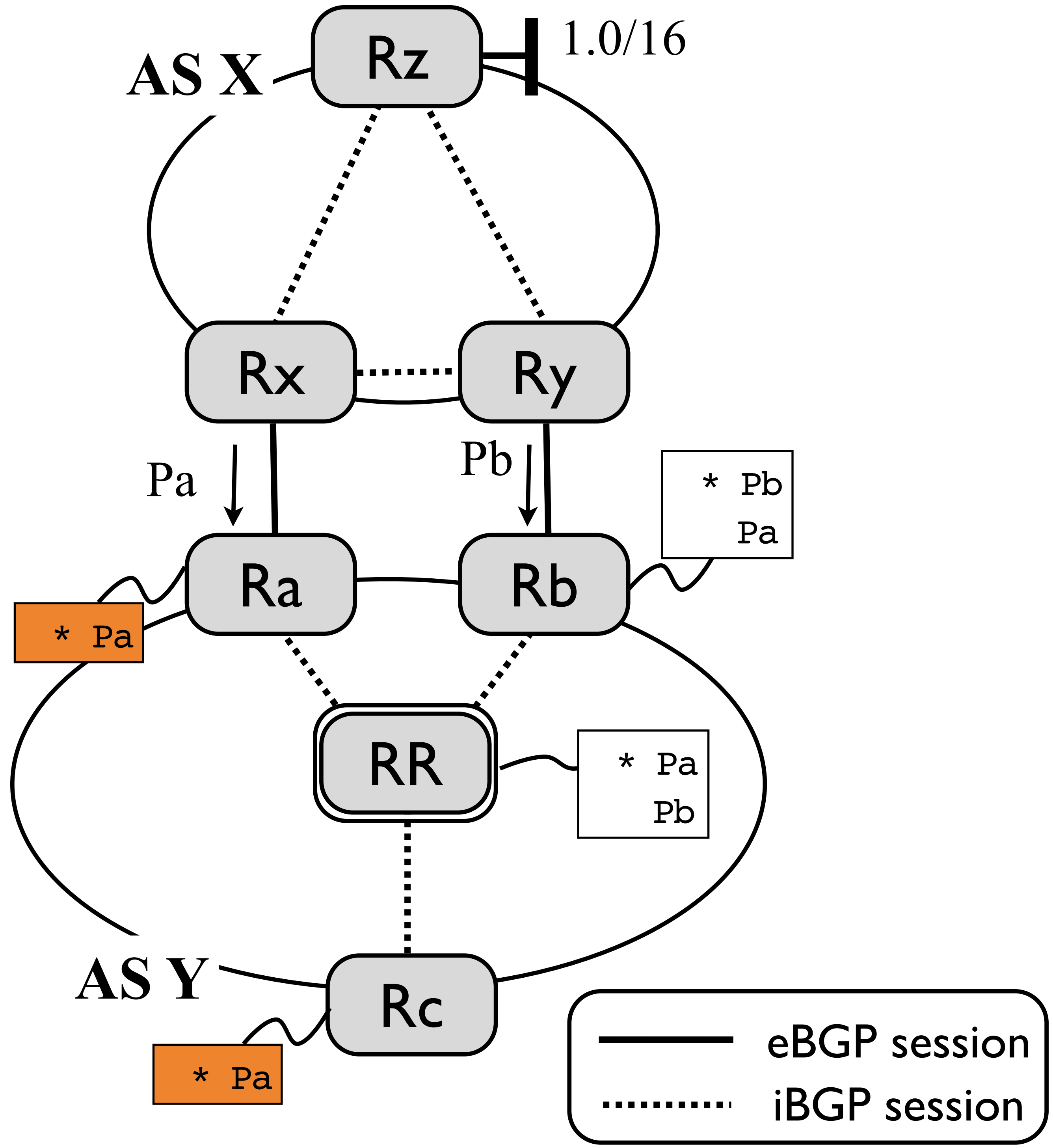
- Auto-configurable iBGP organization based on IiBGP sessions
- Auto-Config, Routing Correctness, (Path diversity)

Path diversity community [V.Van den Schrieck et al., Networking'09]

- Preventing unnecessary BGP withdraws
- Failure Isolation and churn reduction

IiBGP : Motivations

- Alternate paths are important for fast recovery
- Neighboring ASes are often multi-connected [Mérindol]
- Bad diversity propagation in iBGP



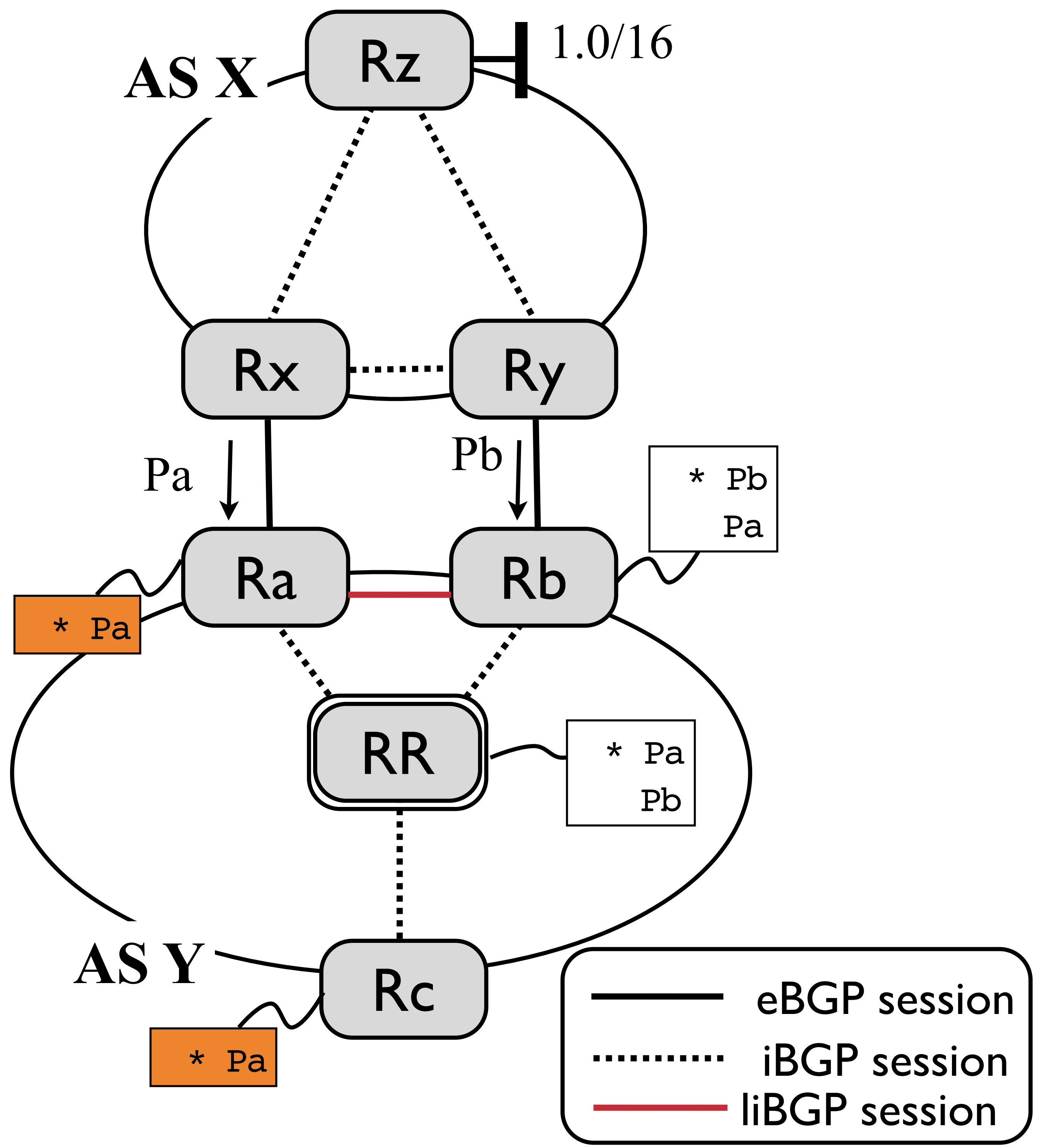
liBGP : Additional sessions

Obtain diversity from where it can be found

- ASBR connected to same neighboring AS

Add a lightweight iBGP session with that ASBR

- Only the paths received from the common neighboring AS



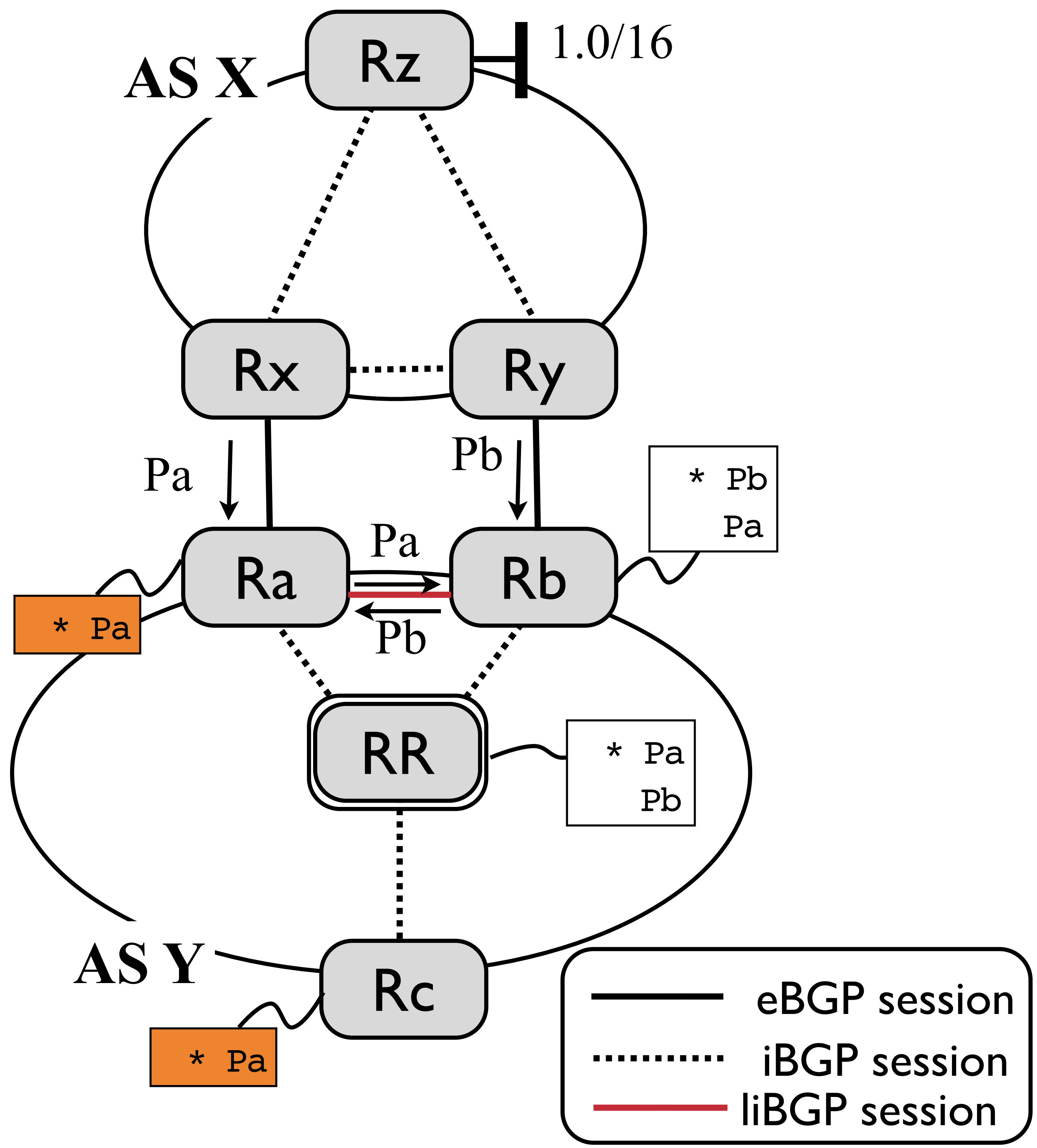
liBGP : Additional sessions

Obtain diversity from where it can be found

- ASBR connected to same neighboring AS

Add a lightweight iBGP session with that ASBR

- Only the paths received from the common neighboring AS



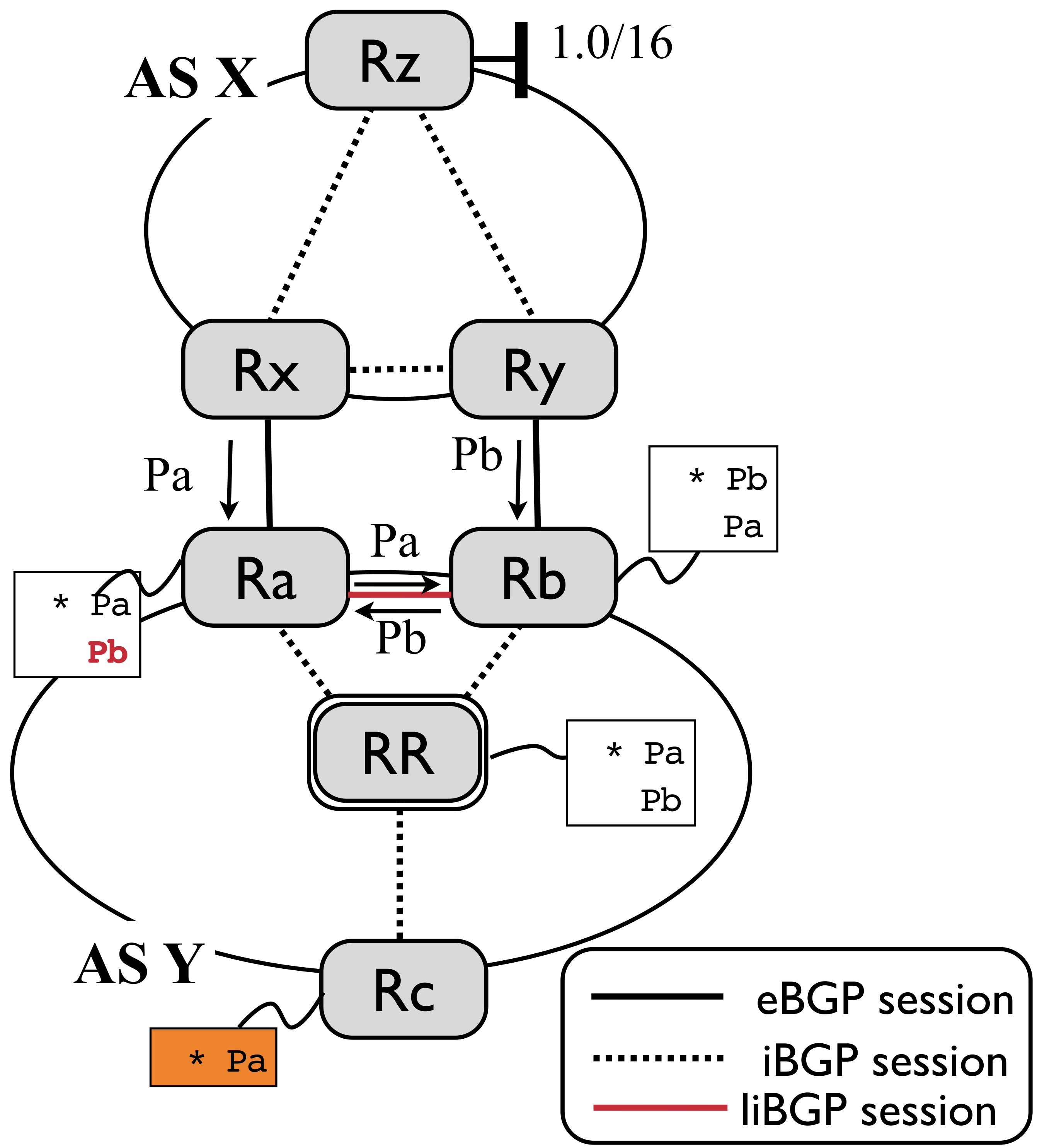
IiBGP : Additional sessions

Obtain diversity from where it can be found

- ASBR connected to same neighboring AS

Add a lightweight iBGP session with that ASBR

- Only the paths received from the common neighboring AS

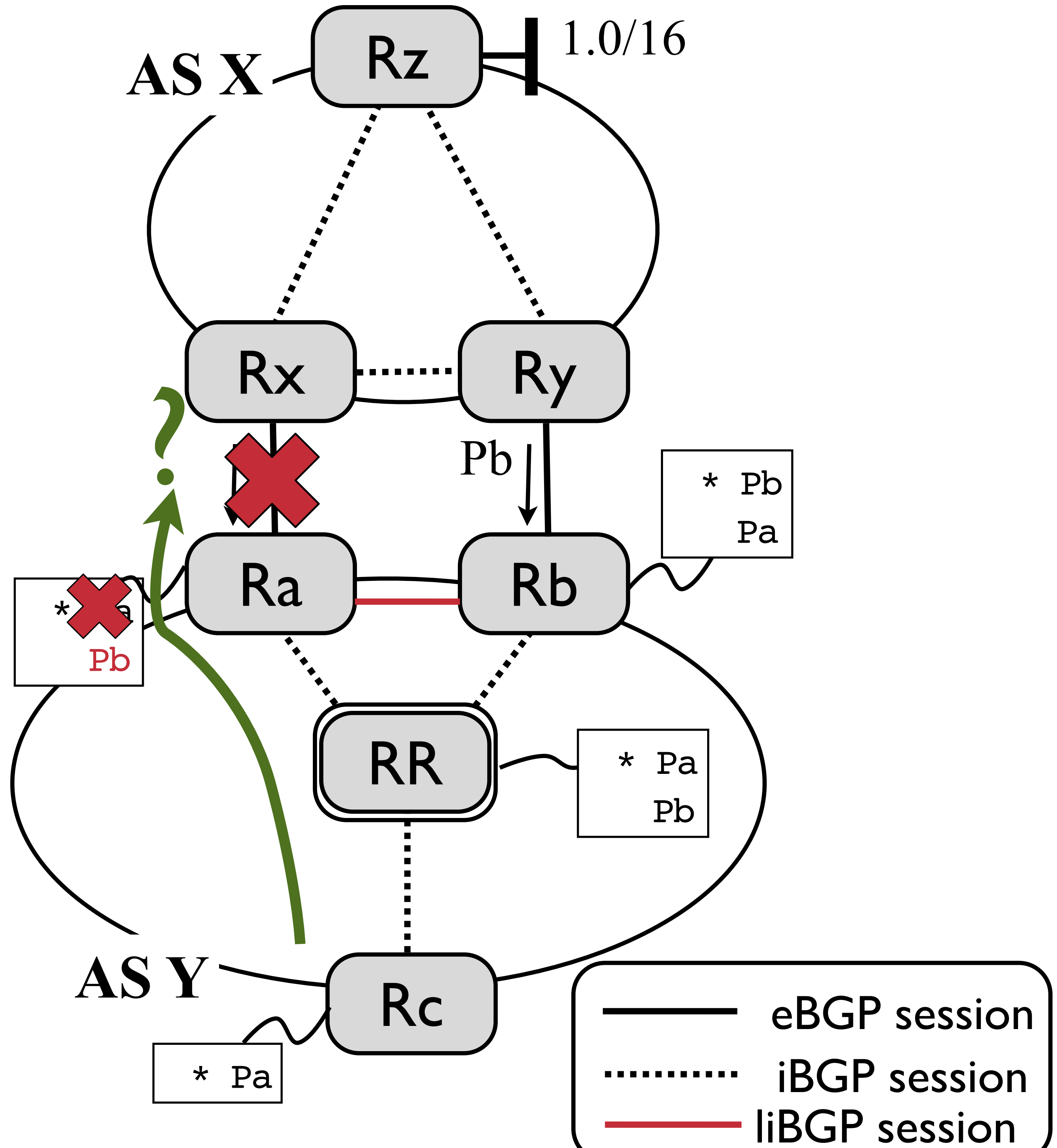


liBGP : Failure Recovery

1st step :

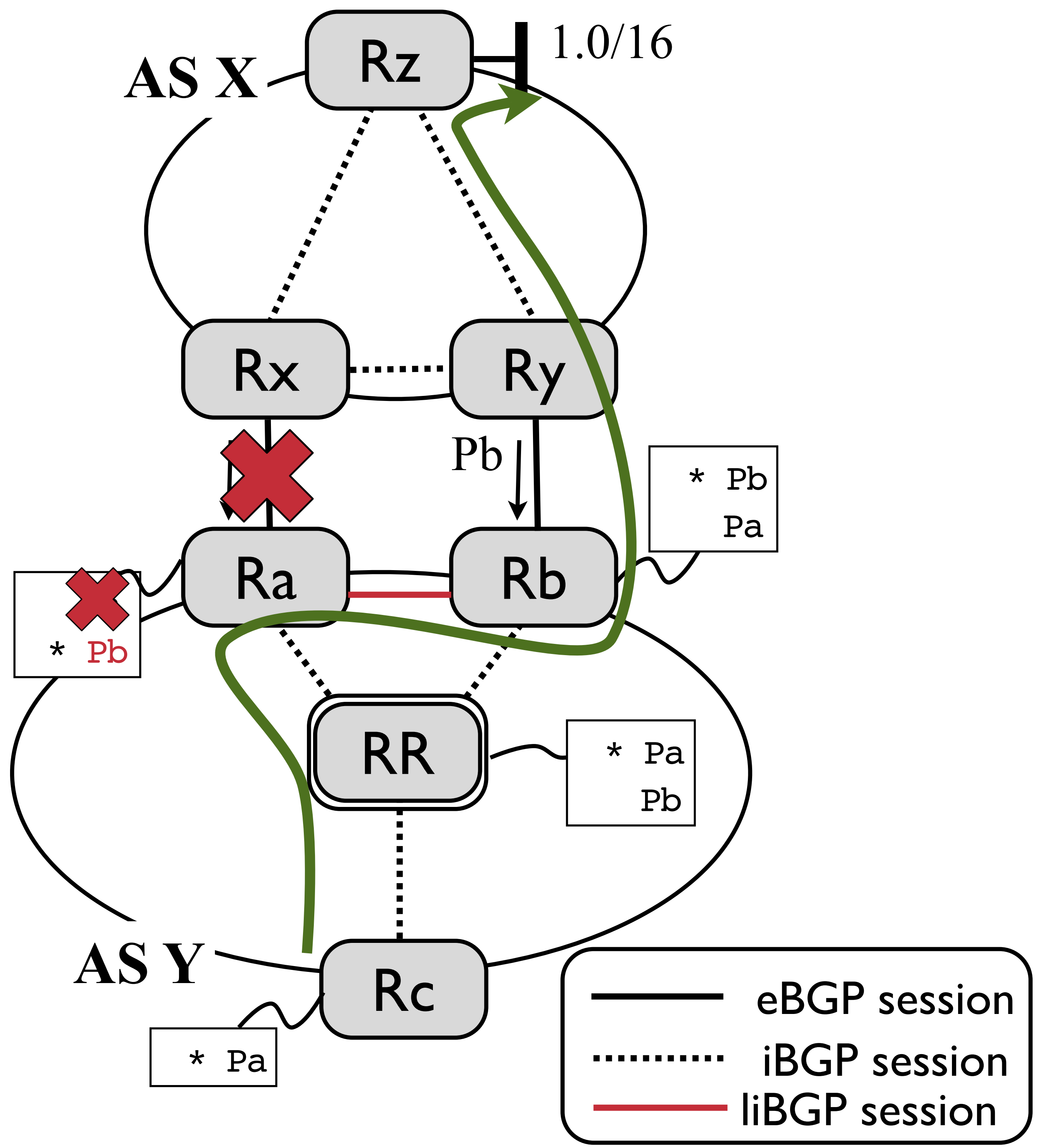
Ra detects the failure of Pa

- In this example, failure detection mechanism is such that Rc does not know about the link failure (but other case supported as well)
- Ra cannot forward traffic to Rx anymore
- But Ra knows alternate path Pb



liBGP :
Egress Failure Recovery

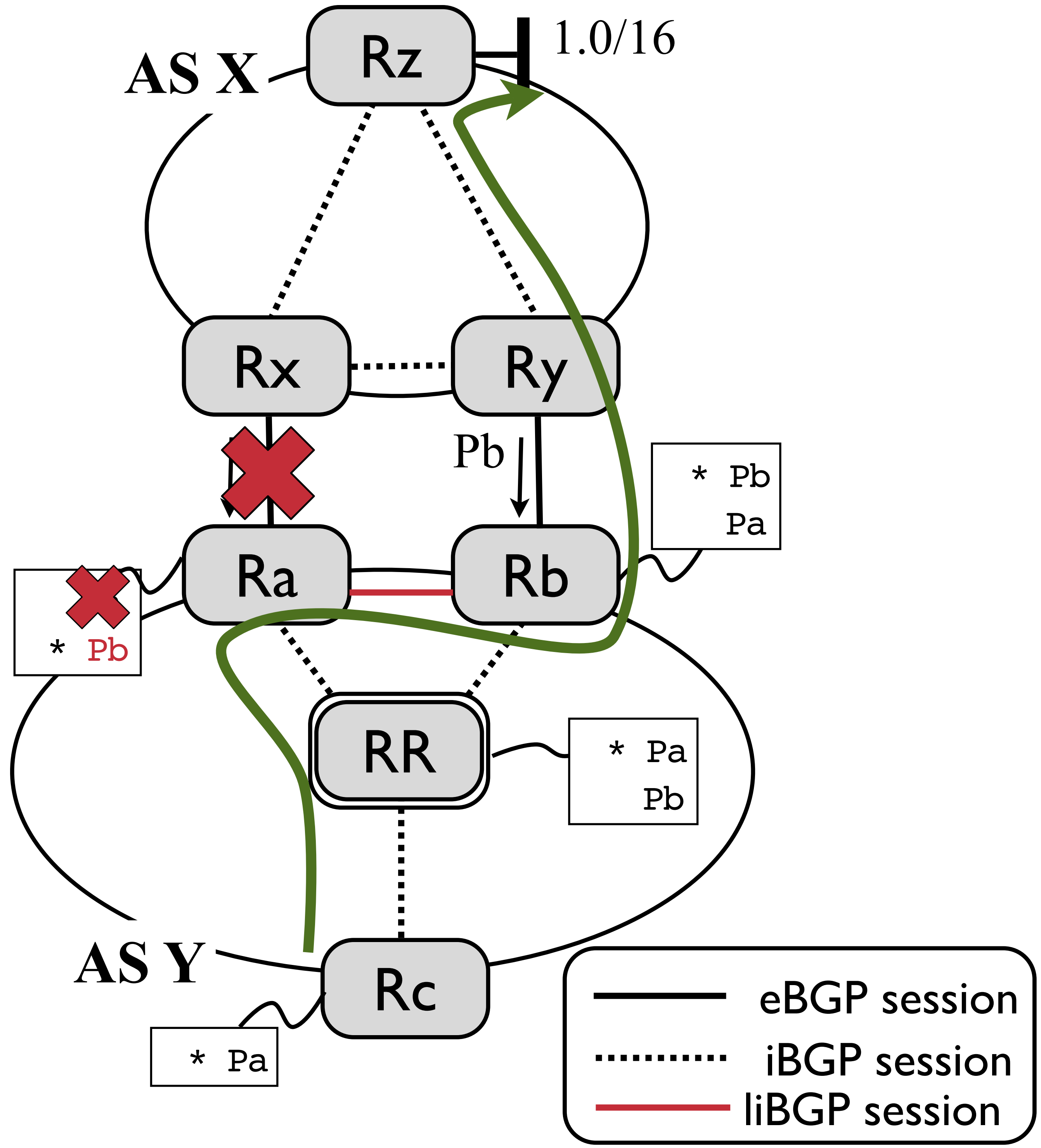
Ra redirects traffic on Pb via a tunnel to Rb



liBGP : Failure Recovery

Traffic loss at Ra is prevented, but routing is not optimal anymore!

- Other routers should stop using Ra as Nexthop, but use Rb instead
- Without loosing traffic themselves, i.e. they should keep using Pa as long as they don't have Pb in their routing tables

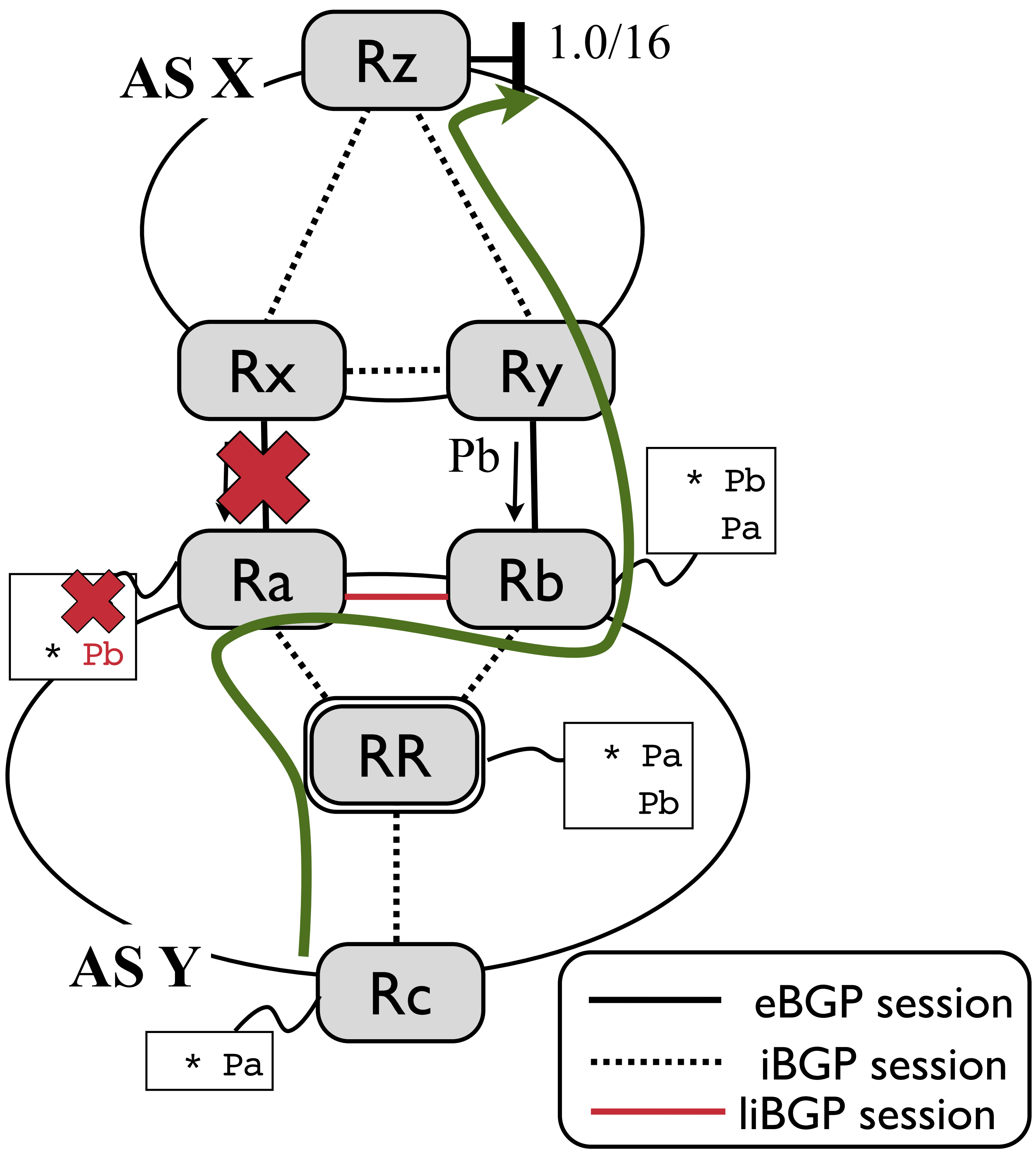


liBGP : Failure Recovery

2nd step :

Graceful Shutdown of the failed link

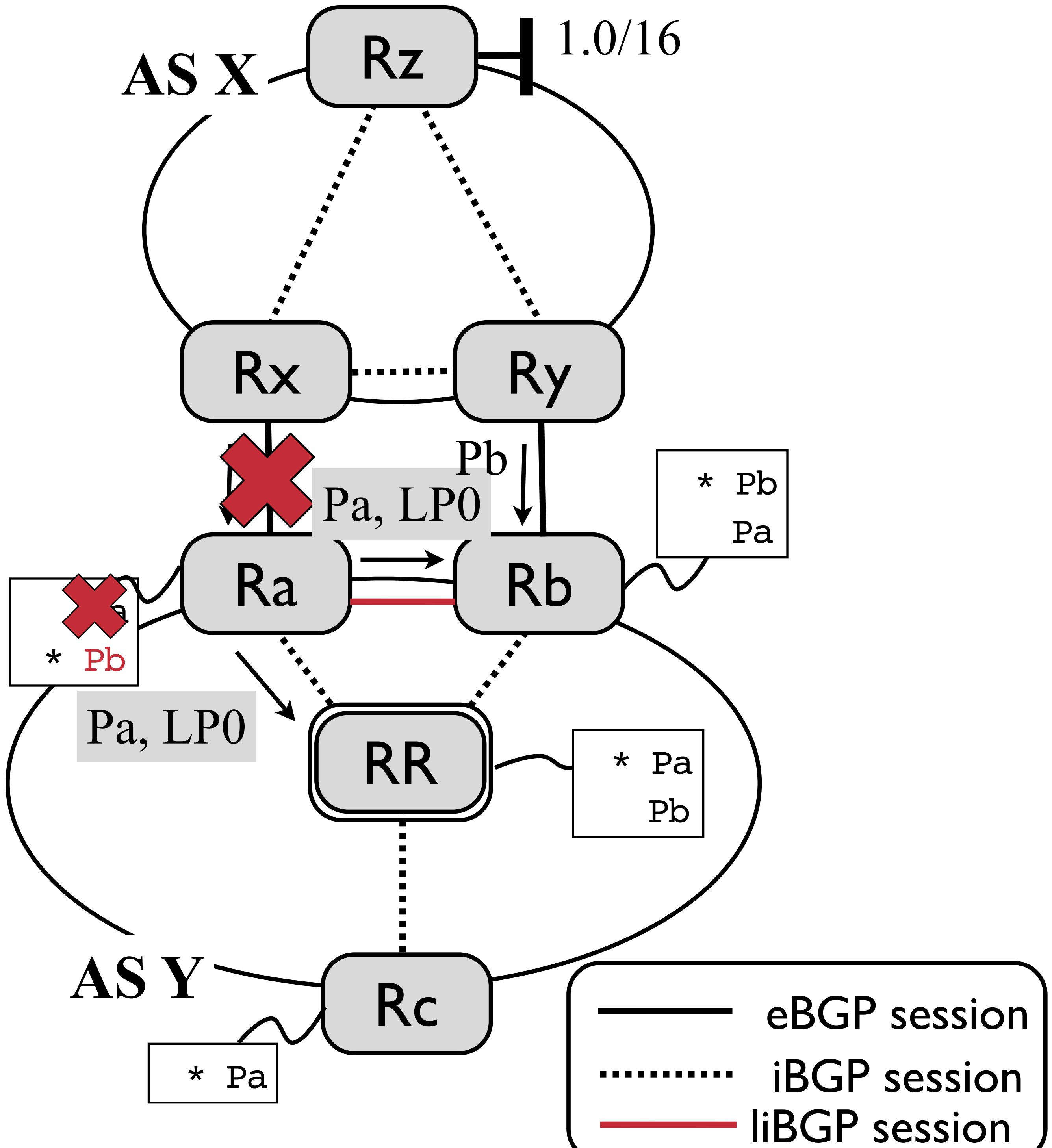
- Don't tell immediately to other routers that Pa is invalid
- But try to convince them to use Pb instead of Pa when they have the choice



liBGP : Failure Recovery

Graceful Shutdown of failed link

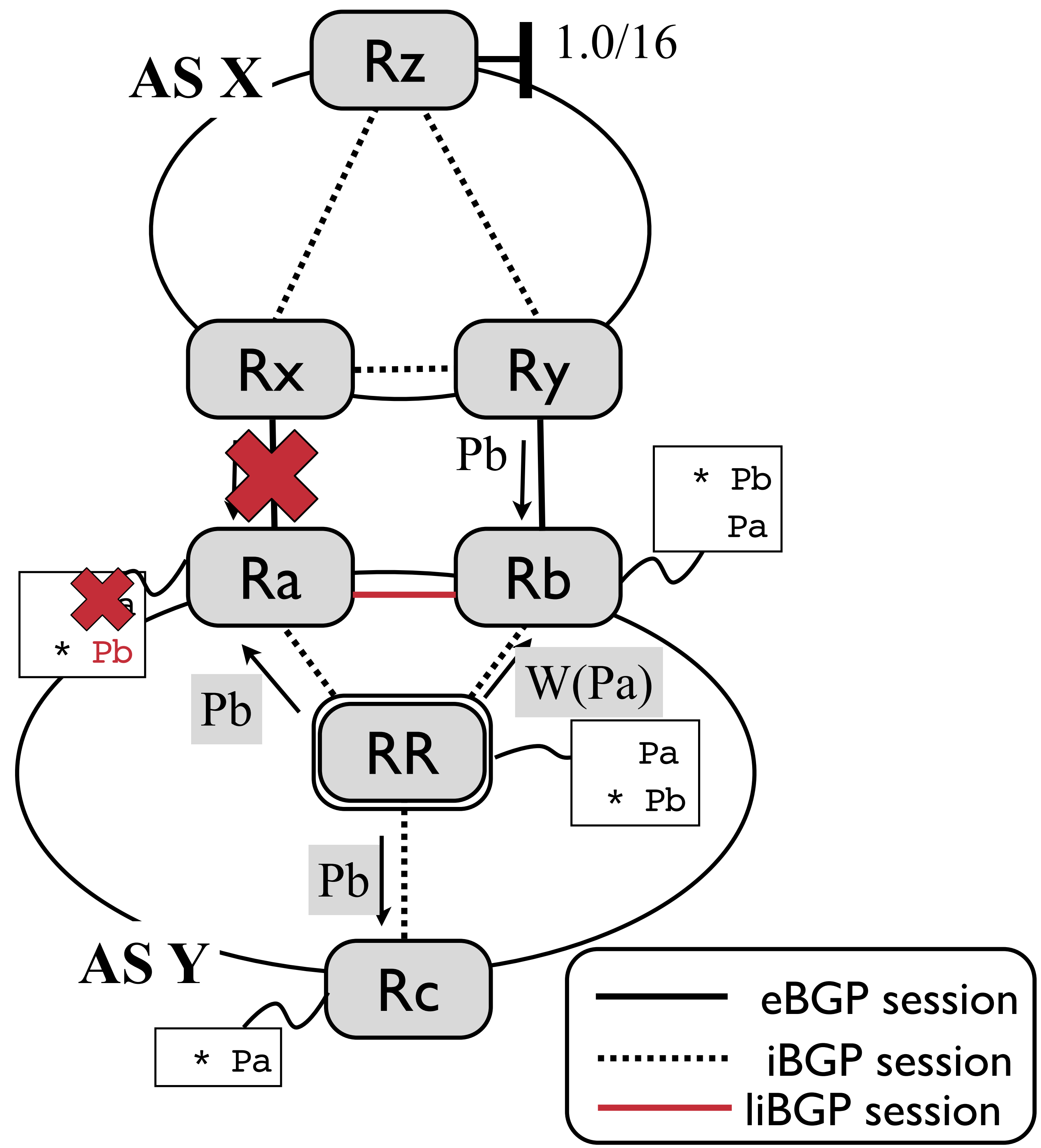
- Ra re-advertises primary path Pa with Local-Pref 0



liBGP : Failure Recovery

Graceful Shutdown of failed link

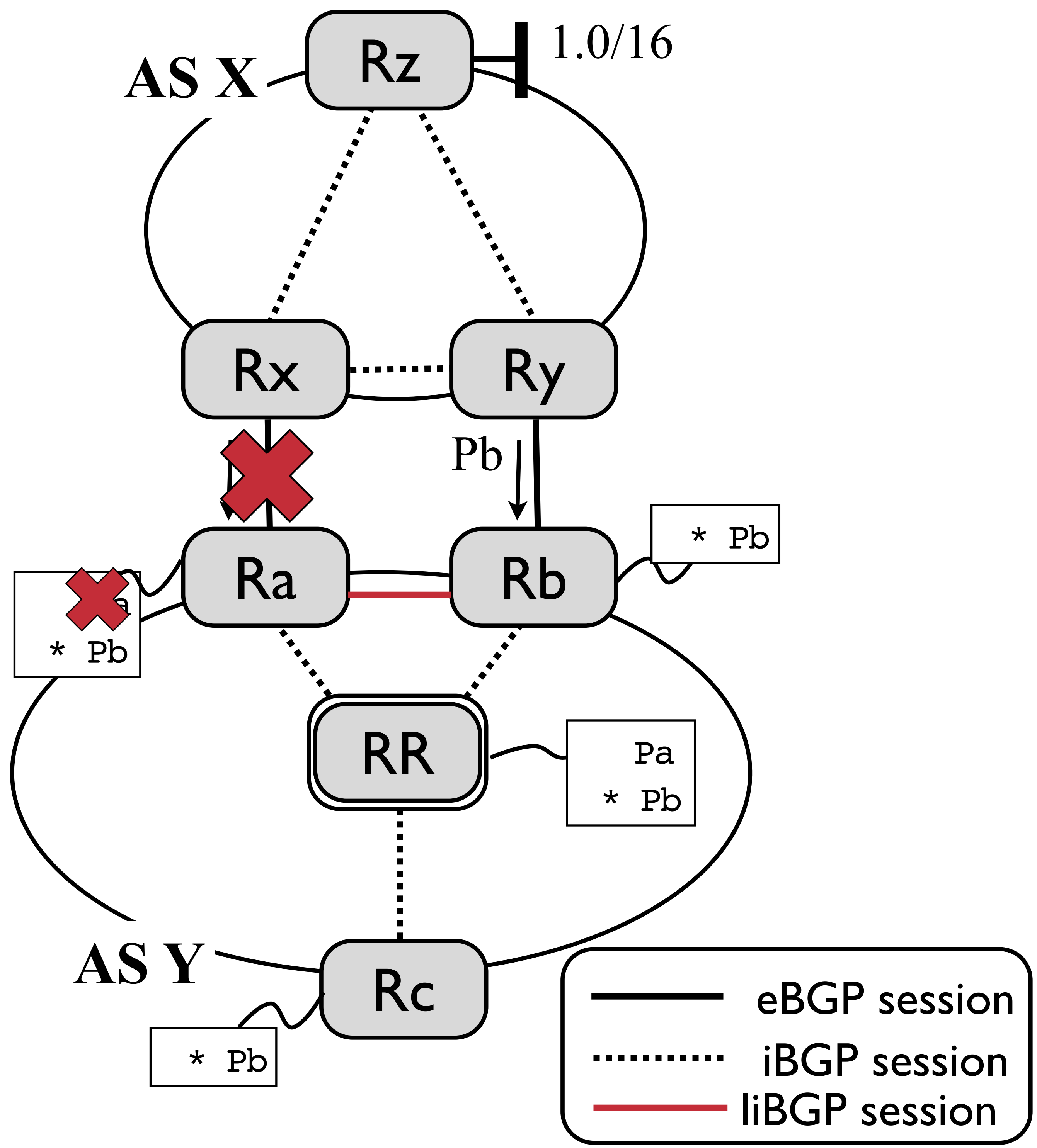
- Ra re-advertises primary path Pa with Local-Pref 0
- RR now prefers Pb, and advertises it to its neighbors



liBGP : Failure Recovery

Graceful Shutdown of failed link

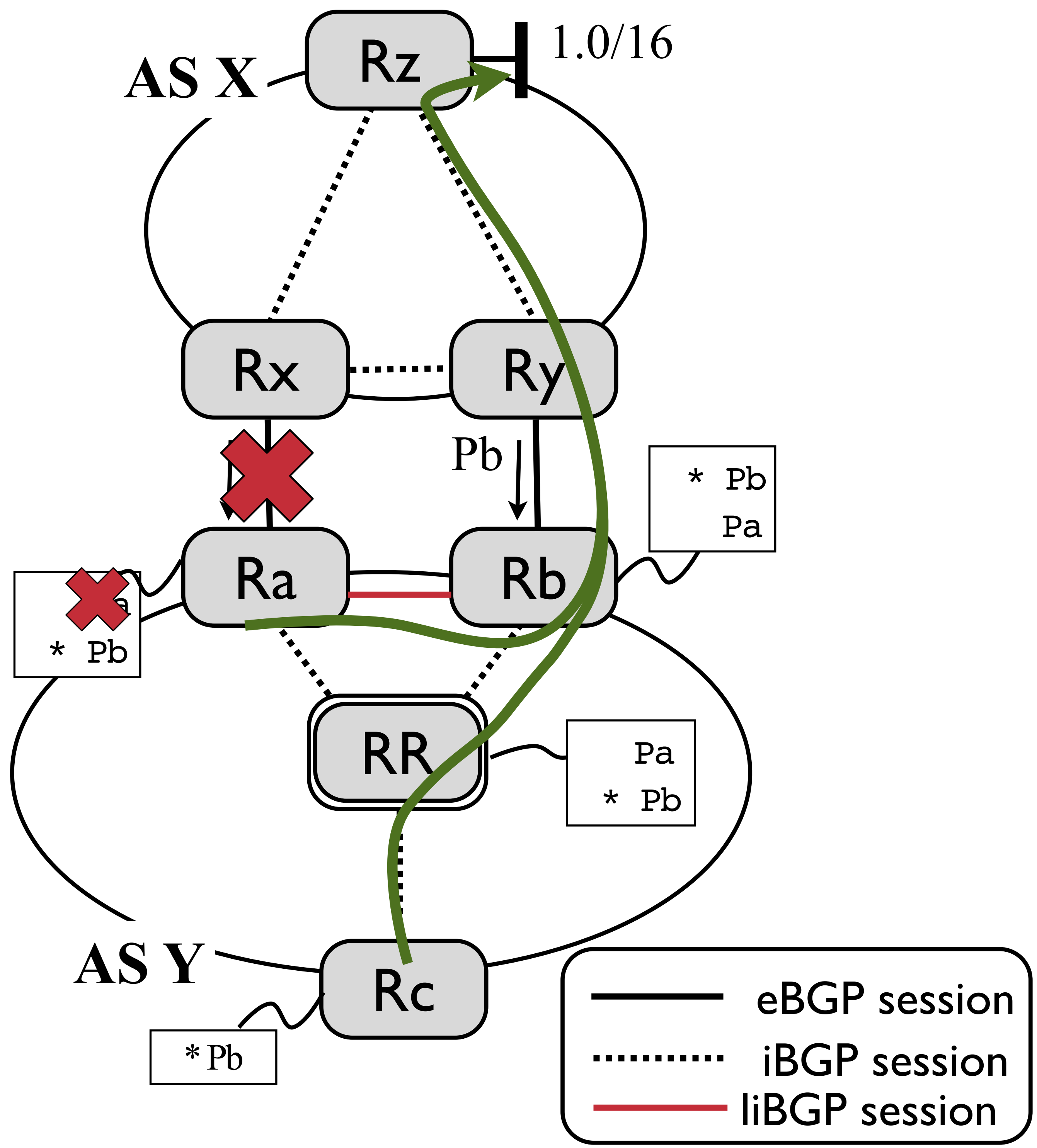
- Ra re-advertises primary path Pa with Local-Pref 0
- RR now prefers Pb, and advertises it to its neighbors
- Pb replaces Pa in the routing table of Rc



liBGP :
Failure Recovery

Graceful Shutdown of failed link

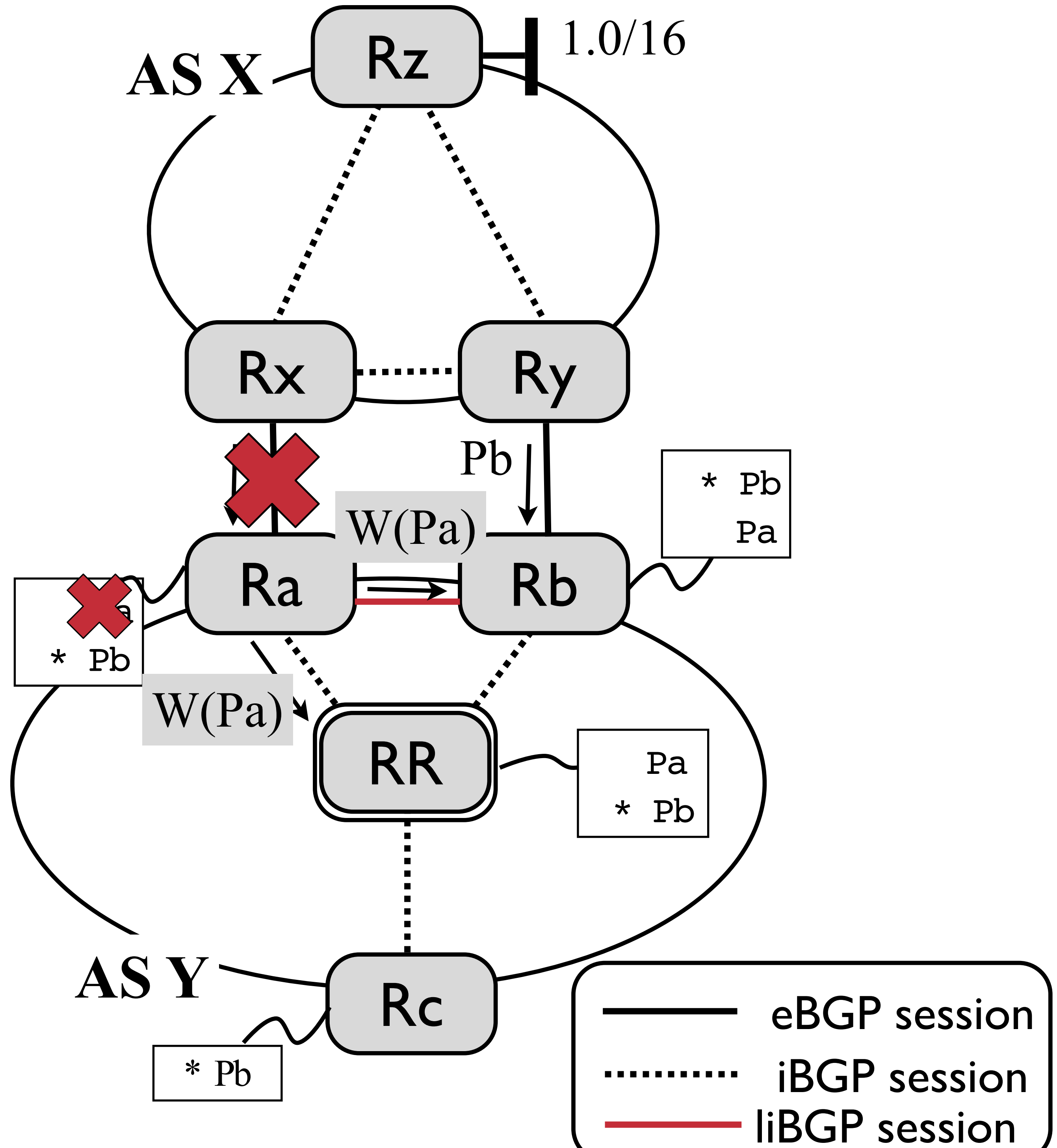
- Ra re-advertises primary path Pa with Local-Pref 0
- RR now prefers Pb, and advertises it to its neighbors
- Pb replaces Pa in the routing table of Rc
- Routing is now optimal



IiBGP : Failure Recovery

Withdrawal of the failed path

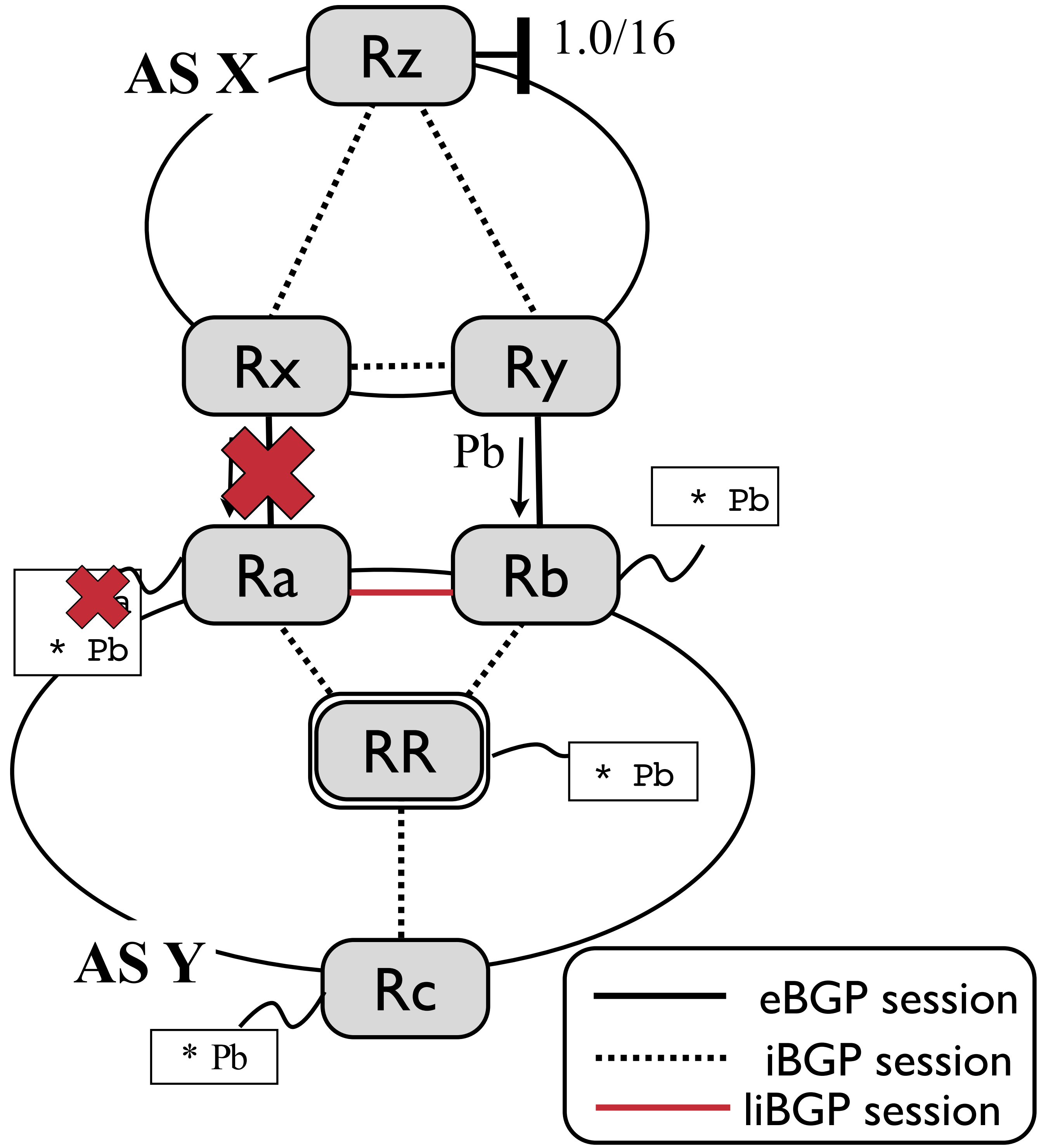
- Now that Pa isn't used anymore, it can be definitely withdrawn
- Ra advertises a BGP messages saying that it cannot reach the destination anymore



liBGP : Failure Recovery

Withdrawal of the failed path

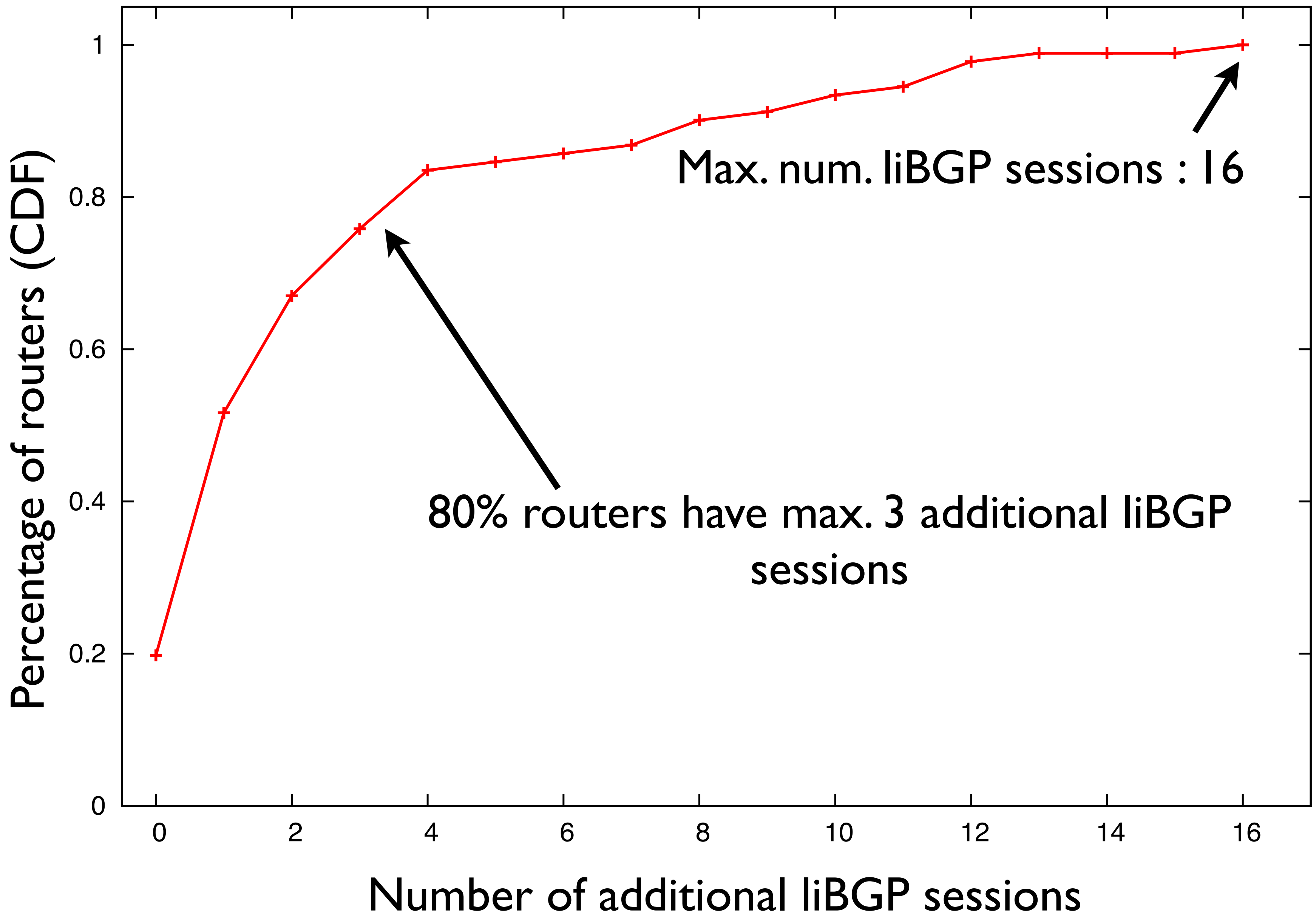
- Now that Pa isn't used anymore, it can be definitely withdrawn
- Ra advertises a BGP messages saying that it cannot reach the destination anymore
- Pa doesn't appear in the routing tables anymore



liBGP : Cost of additional sessions

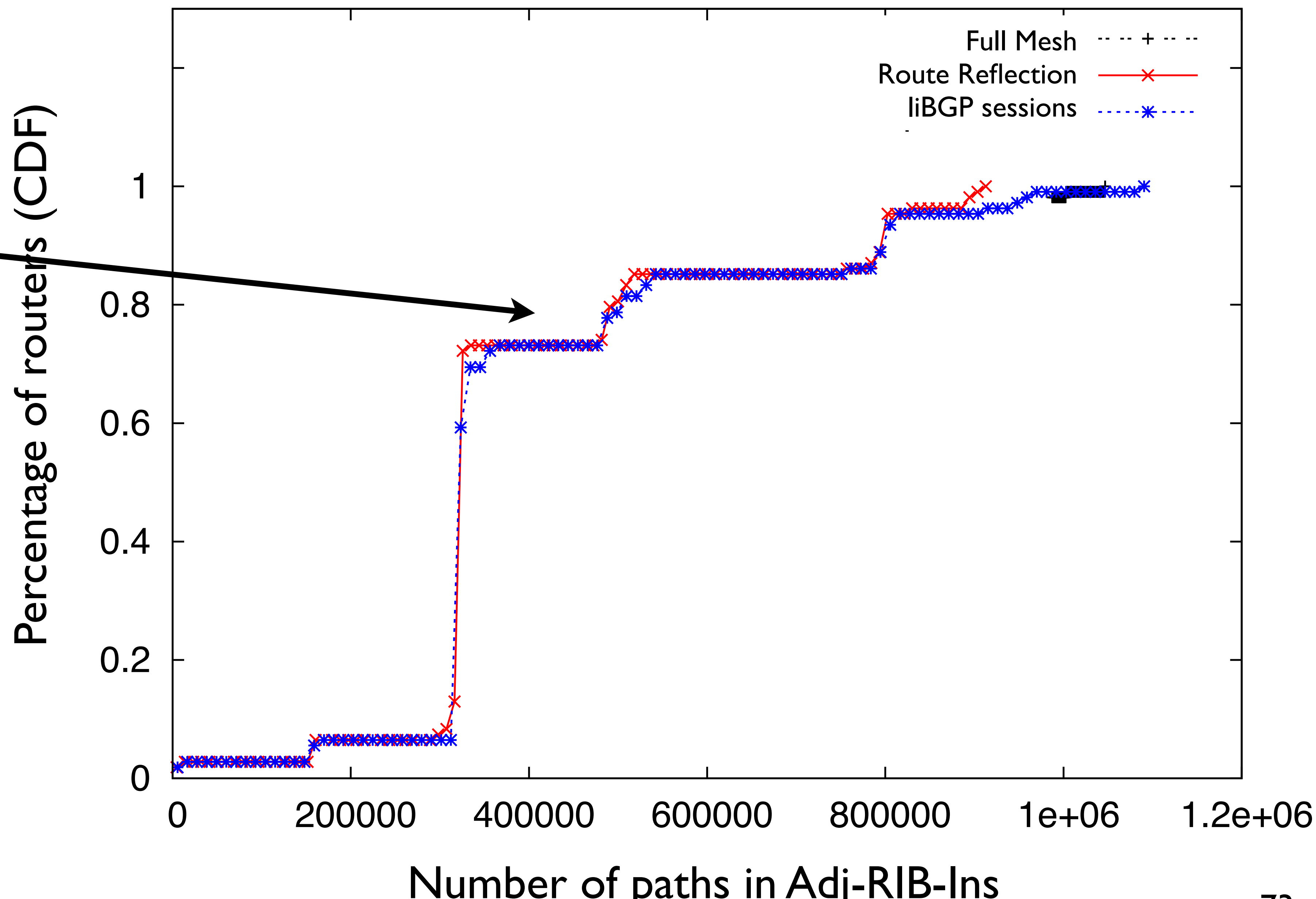
On T1 ISP :

- We protect all multi-connected neighbors
- Few additional sessions per router



IiBGP : Routing tables load

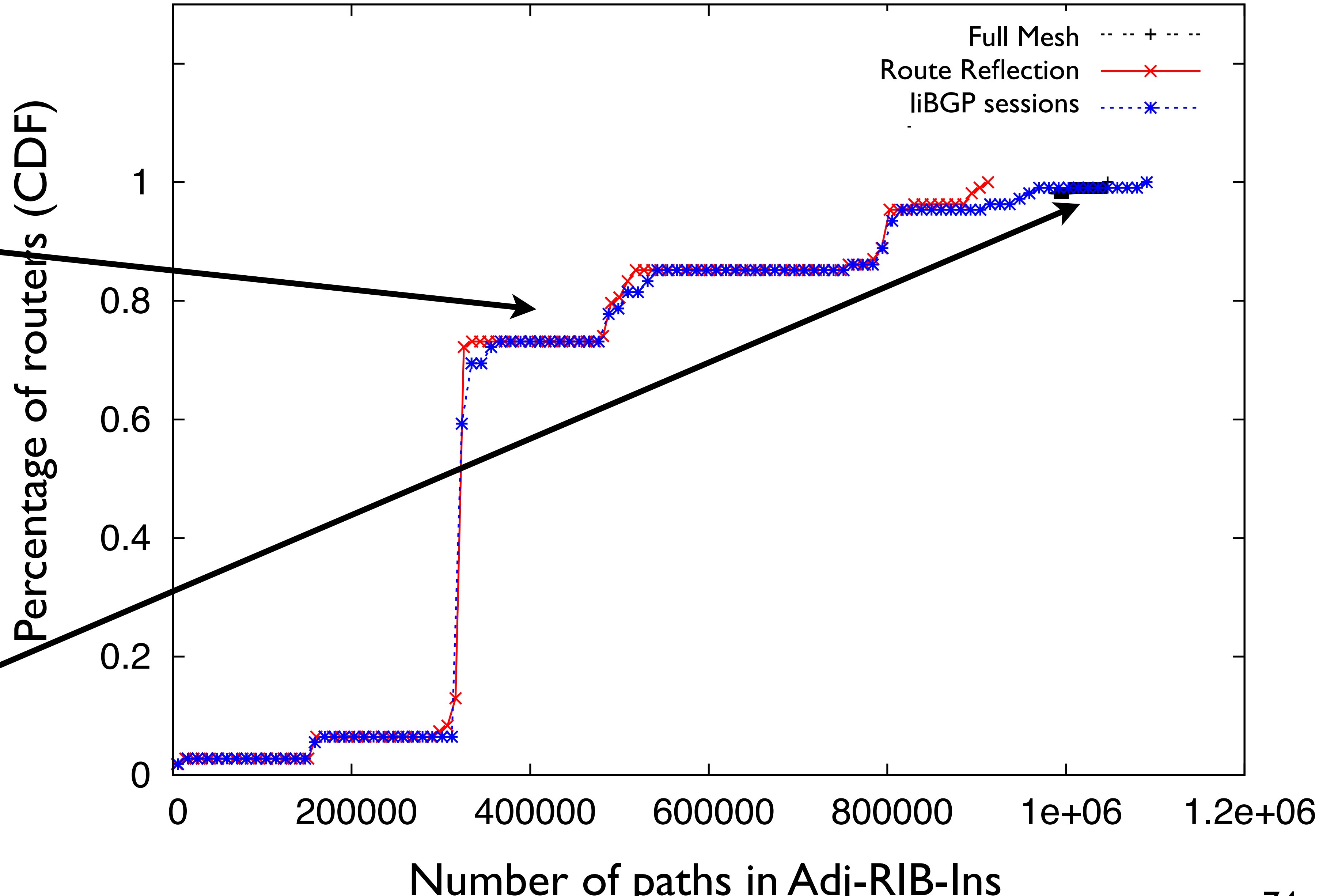
Few additional paths for most routers



IiBGP : Routing tables load

Few additional paths for most routers

A few routers have to support a higher load



liBGP : Summary

On demand path diversity for multi-connected neighbors

Cost is moderate, few modifications to BGP

Auto-configuration of liBGP sessions

Does not work for single-connected neighbors

Outline

I. Background on Internet routing

II. Problem Statement

III. Improving iBGP with single path advertisement

IV. Improving iBGP with multiple paths advertisement : BGP Add-Paths

- BGP Add-Paths : Motivations and Selection modes
- Qualitative and quantitative analysis

Motivations

Origin of iBGP issues : Single path advertisement

- Lack of path diversity for fast recovery
- Routing anomalies

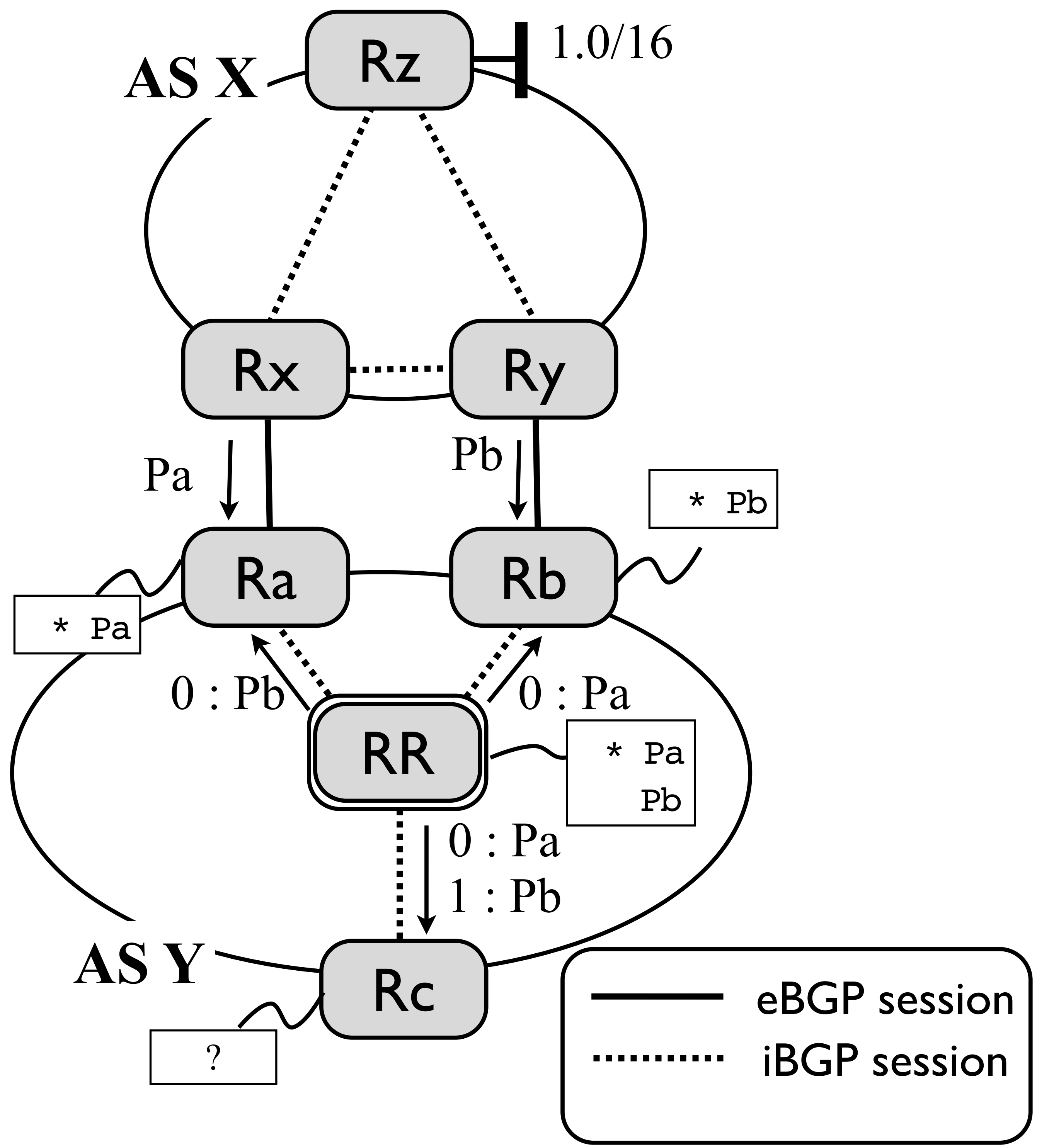
Why not remove this limitation?

- Scalability of routing tables might be frightening
- But iBGP is under control!

Add-Paths : Example

RR can advertise more than the best path

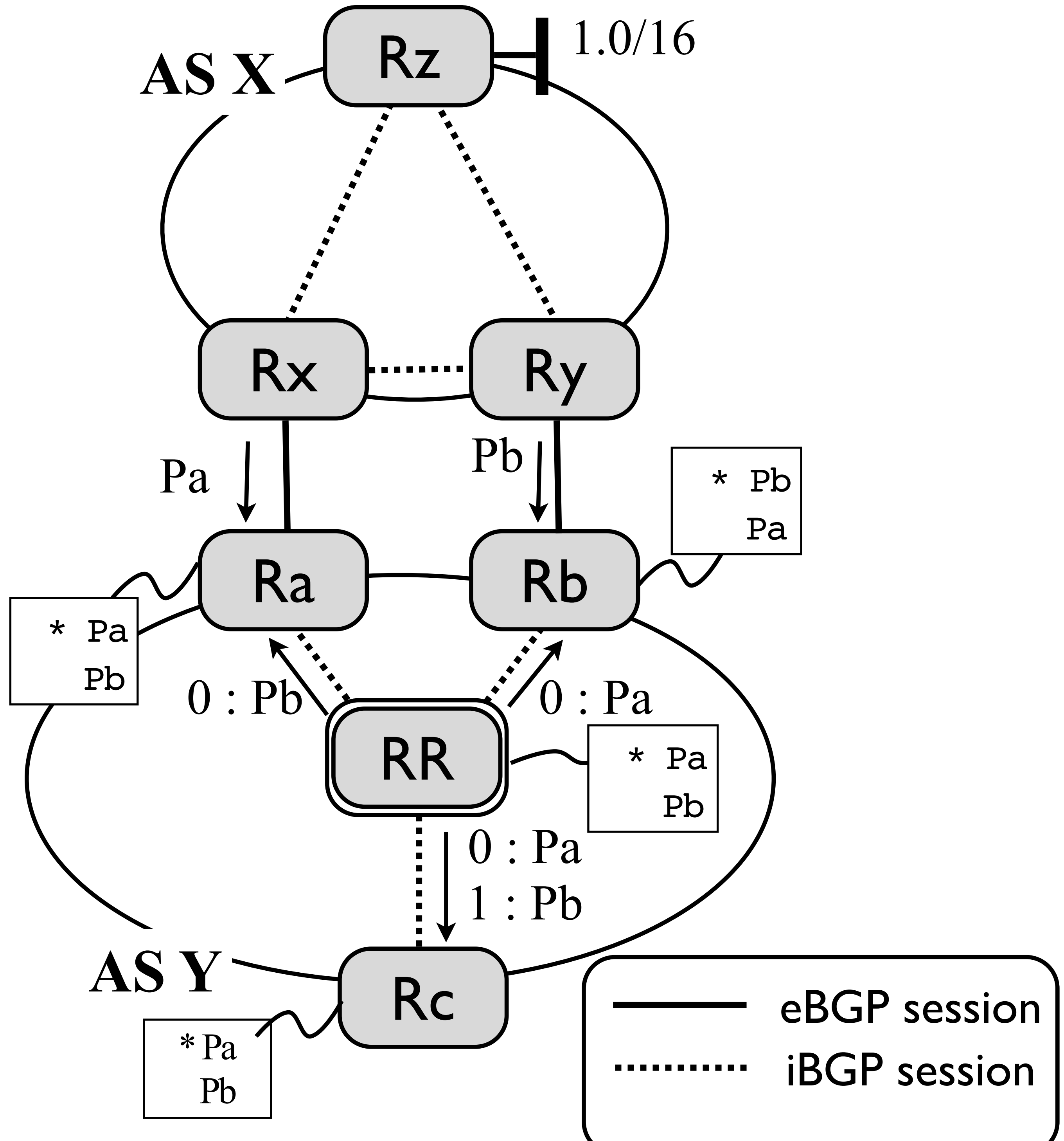
- Non-Best path to Ra
- Two paths to Rc



Add-Paths : Example

RR can advertise more than the best path

- Non-Best path to Ra
- Two paths to Rc



Add-Paths : Which paths?

Internet draft defines the encoding of multiple paths per destination

But which paths should be advertised to neighbors?

- All paths?
- Best path + Backup paths?
- ...

Contribution of this thesis : Analysis of the tradeoffs of the sets of paths advertised by Add-Paths

Selection modes

A **selection mode** is an algorithm selecting the paths to be advertised with Add-Paths

Properties of selections modes :

- Routing oscillations prevention
- Provisioning of N alternate paths
- ...

Cost of selection modes :

- Routing table load
- Churn
- ...

Selection modes

Some selection modes existing in literature :

- **Add-All-Paths** [Basu]
- **Add-N-Paths** [Mohapatra]
- **Add-Group-Best-Paths** [Walton]

New selection modes proposed :

- **Advertise-LP1-LP2-Paths**

Add-Paths Analysis

What are the trade-offs of each selection mode

Qualitative analysis : Properties of each mode?

- **[draft vvds-add-paths-analysis]**
- **[V. Van den Schrieck et al., JSAC 2010]**

Quantitative analysis : Impact of a mode for a given ISP?

- **[V. Van den Schrieck et al., JSAC 2010]**

Qualitative analysis

	Backup path	Routing oscillation avoidance	Routing tables load + Churn	...
Add-All-Paths	OK	OK	Maximum, unbounded	...
Add-N-Paths	OK	KO [draft-vvds-add-paths-osc.], but improved	Bounded : Max. N per dest.	...
Add-LP1-LP2-Paths	OK	OK	Large, unbounded	...
Add-Group-Best-Paths	**KO**	OK	Unbounded : # neighbors advertising a destination	...

Unbounded cost : Depends on the ISP considered

=> Has to be evaluated case by case

Quantitative analysis

Analyser : SimBGP, extended with Add-Paths

Metrics extracted from simulation output

- Convergence times
- Routing table load, path diversity
- Churn
- ...

Quantitative analysis : Example

Simulations on generated Internet topologies

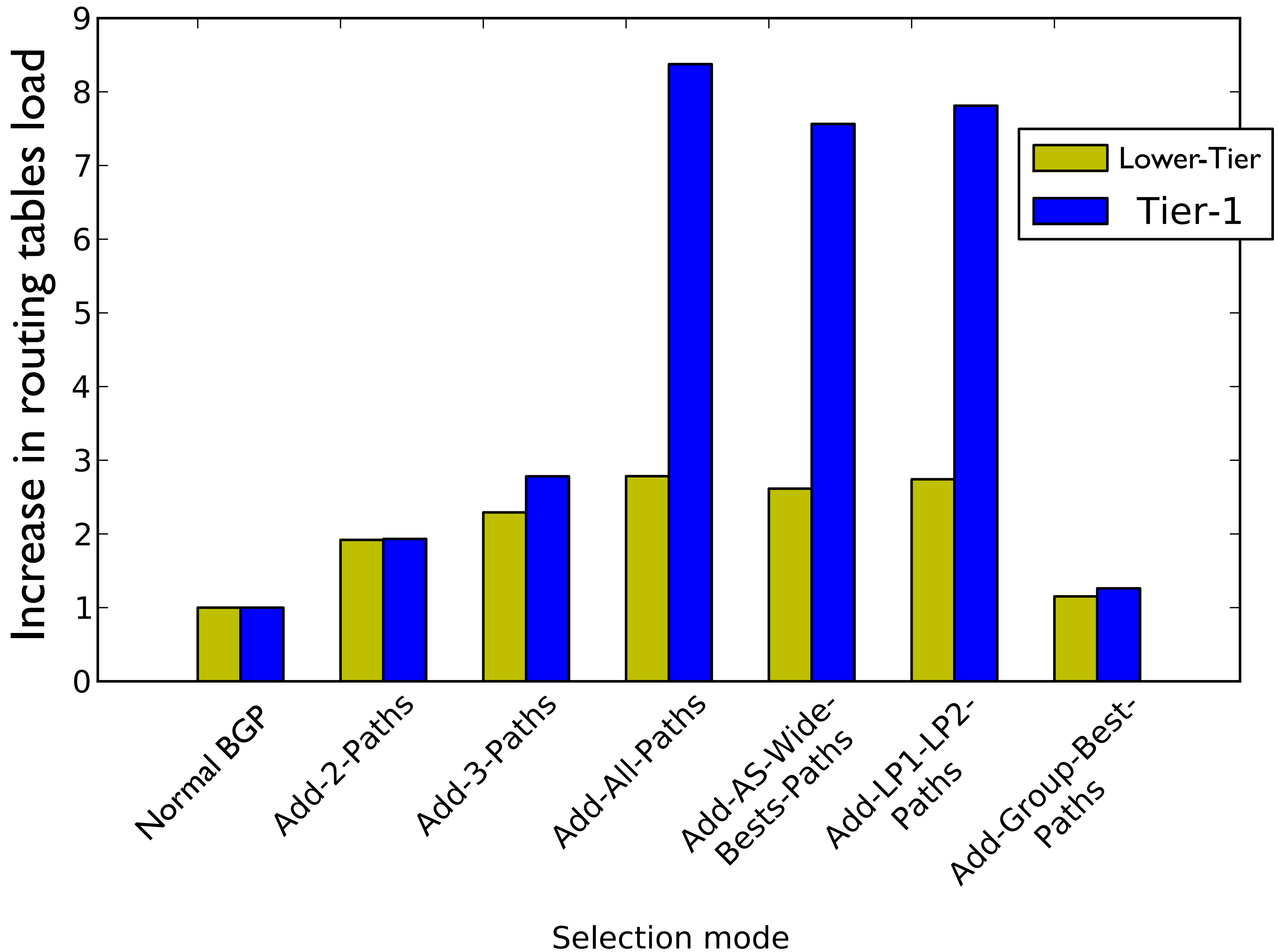
- Ghitle (AS-level graph) + iGen (router-level topo)
- Destinations advertised at random locations

One provider running Add-Paths at each run

Metric considered : Average **number of paths per destination** in the routing tables of the provider running Add-Paths

Routing Tables Load

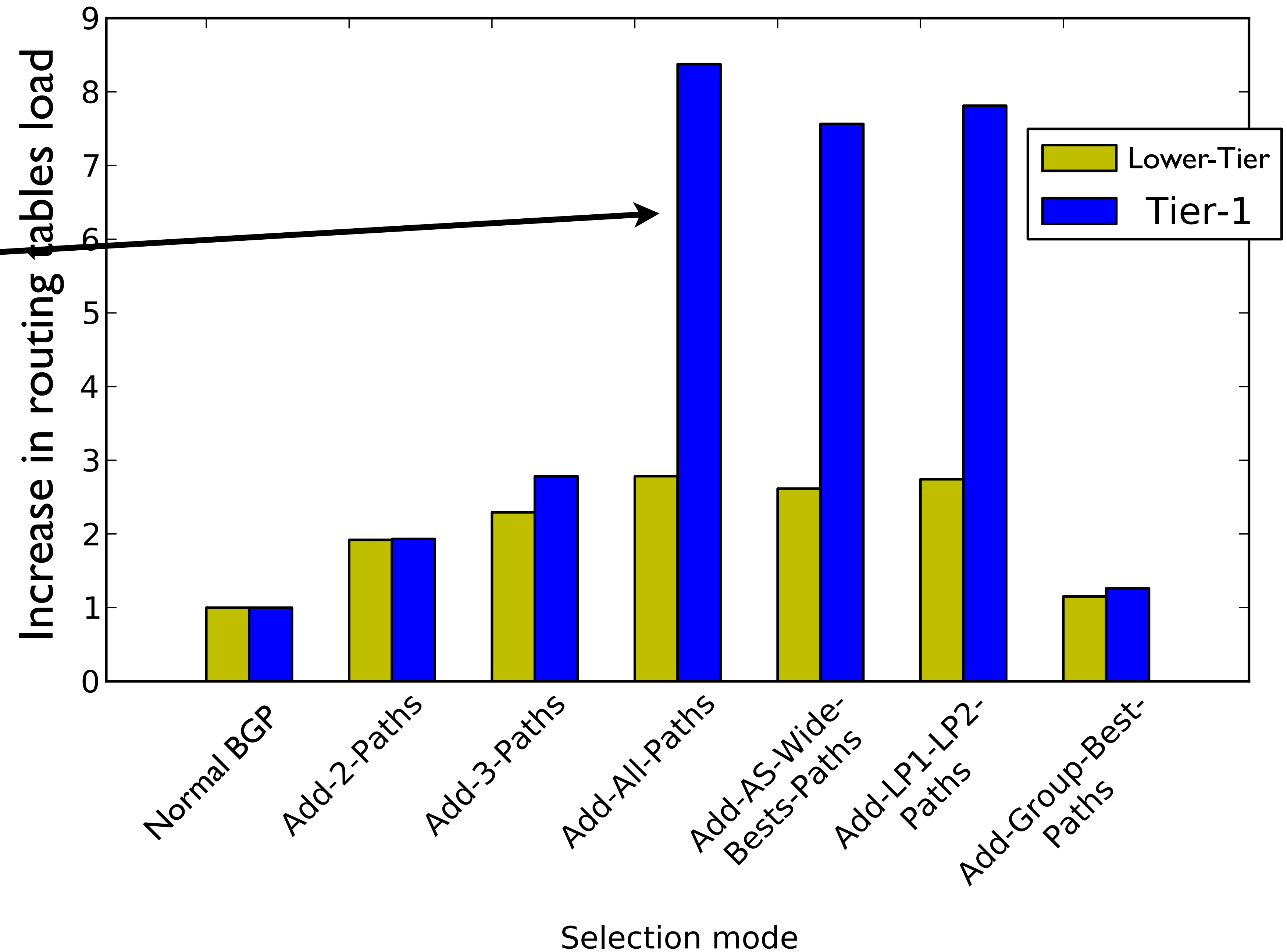
Increase in number of paths per router



Routing Tables Load

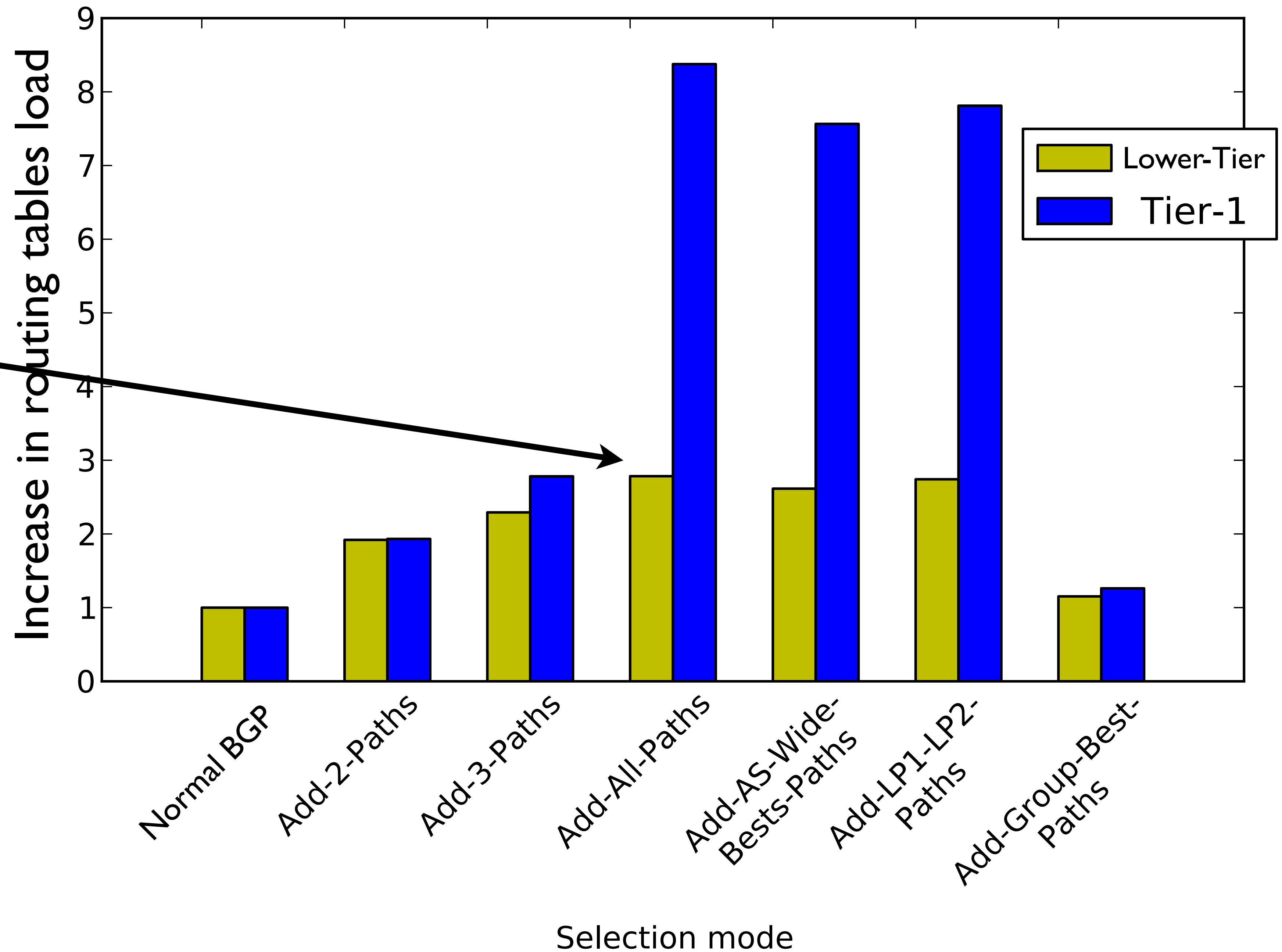
Increase in number of paths per router

Lots of additional paths with non-bounded modes for Tier-1 ISP



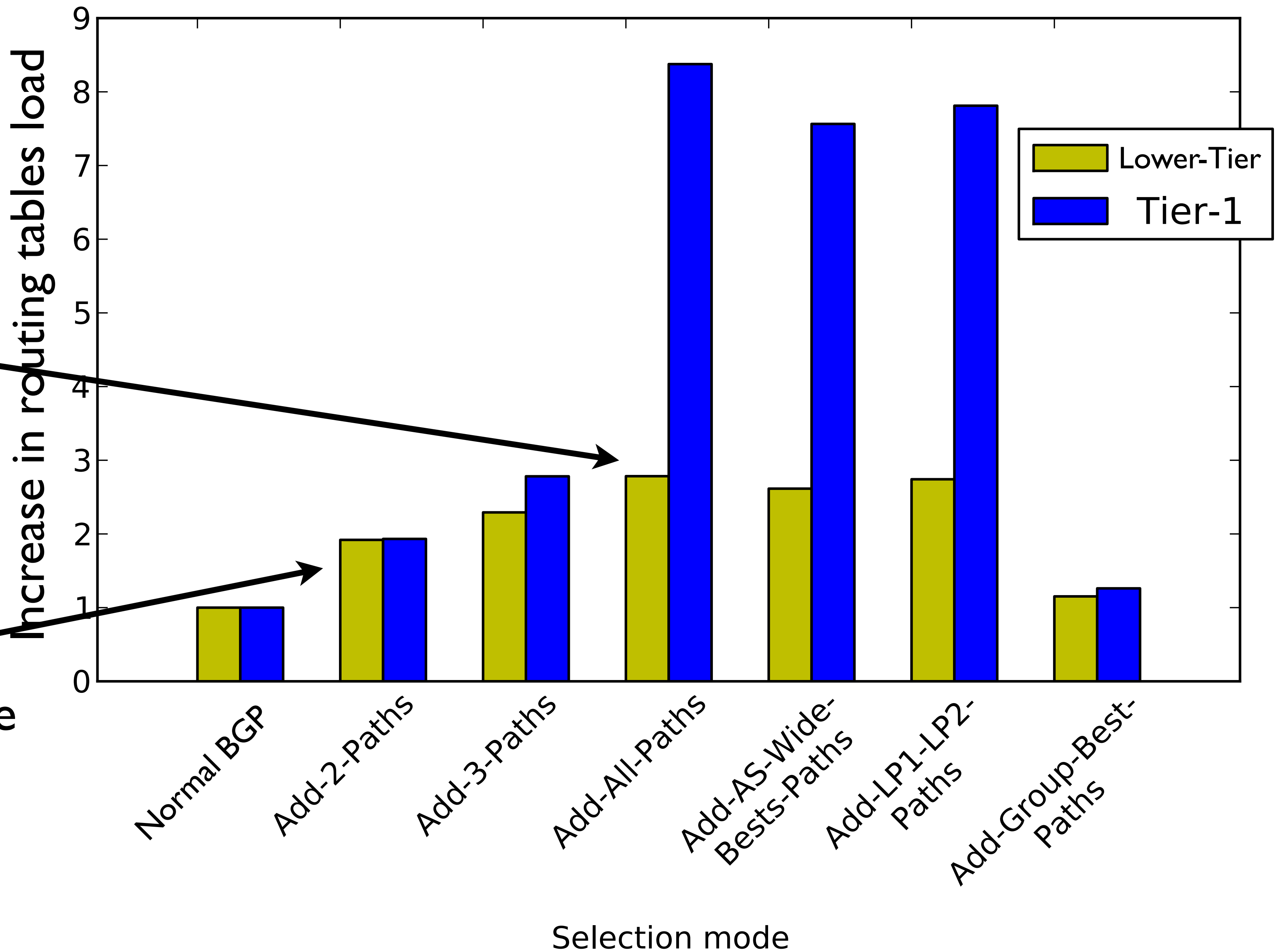
Routing Tables Load
Increase in number of paths per router

Load is more reasonable for lower-Tier ISPs



Routing Tables Load

Increase in number of paths per router



Load is more reasonable for lower-Tier ISPs

Add-2-Paths has an increase factor of 2

Add-Paths : Conclusion

Add-ALL-Paths

- All nice properties
- **But** load is maximal and non bounded

Add-LP1-LP2-Paths

- Same properties than Add-All-Paths
- Also non-bounded load, slightly lighter

Add-N-Paths

- Fast recovery, correctness improved but not guaranteed
- Controlled load

Add-Group-Best-Paths

- Routing correctness
- No Fast Recovery

Conclusion of the thesis

We improved single path iBGP

- Three multi-purpose proposals
- But not “THE” perfect solution

We analysed the trade-offs of Add-Paths

- Complete solution with different “flavours” (selection modes)
- “Hot Topic” at the IETF, under implementation by router vendors
- **Add-N-Paths** has an interesting cost/performance trade-off

Conclusion of the thesis

Goal : **improve global routing by changing iBGP routing**

iBGP improved by :

- Increasing fast recovery and failure isolation upon failure
- Preventing routing oscillations
- Preventing misconfigurations with Auto-Configuration

Impact on global routing :

- Less connectivity losses
- Shorter convergence times
- Reduced interdomain churn

Questions?