Internet Routing Architecture

Stefano Secci

Outline

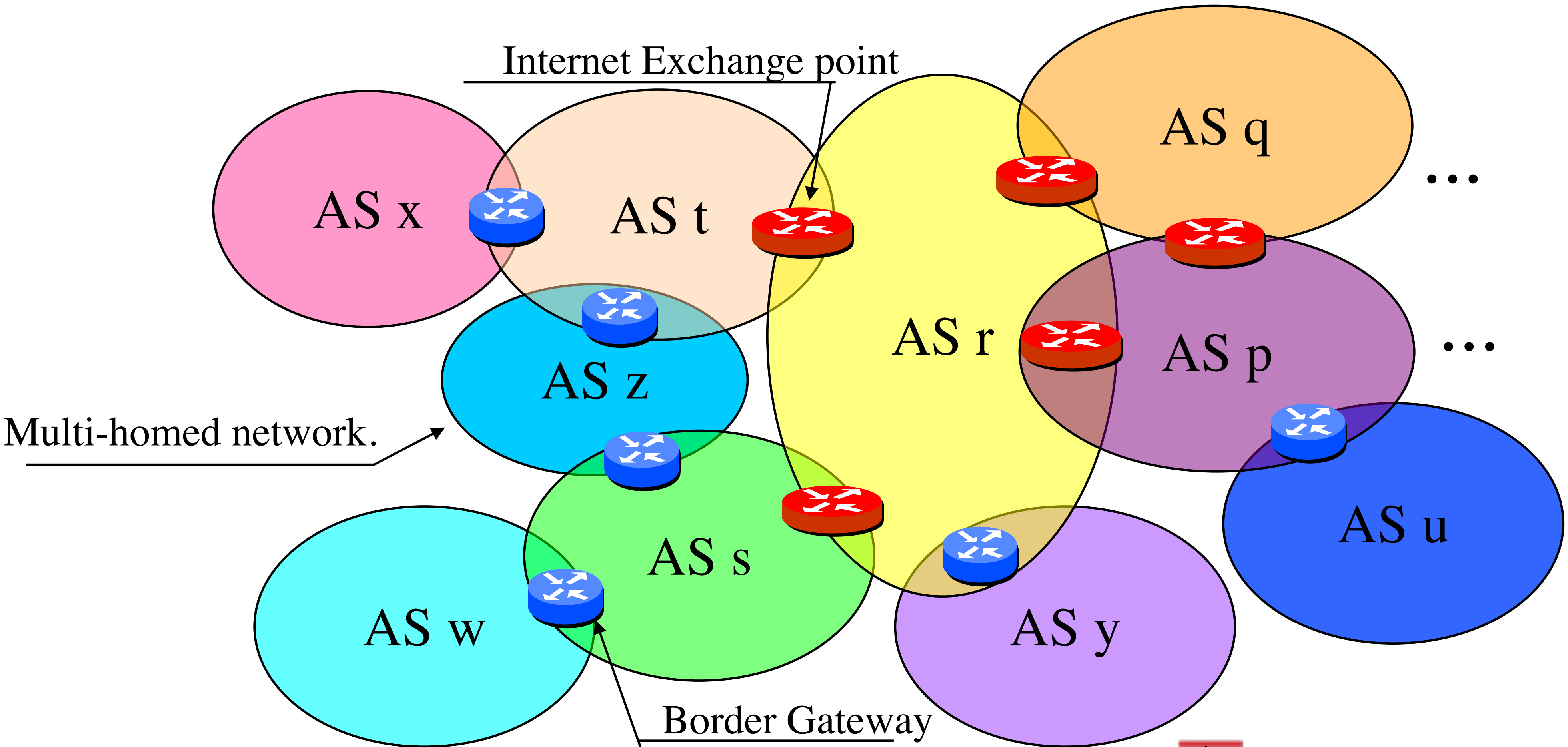
- The Internet Ecosystem
- The BGP protocol
- BGP scalability and management

The big picture

THE INTERNET ECOSYSTEM

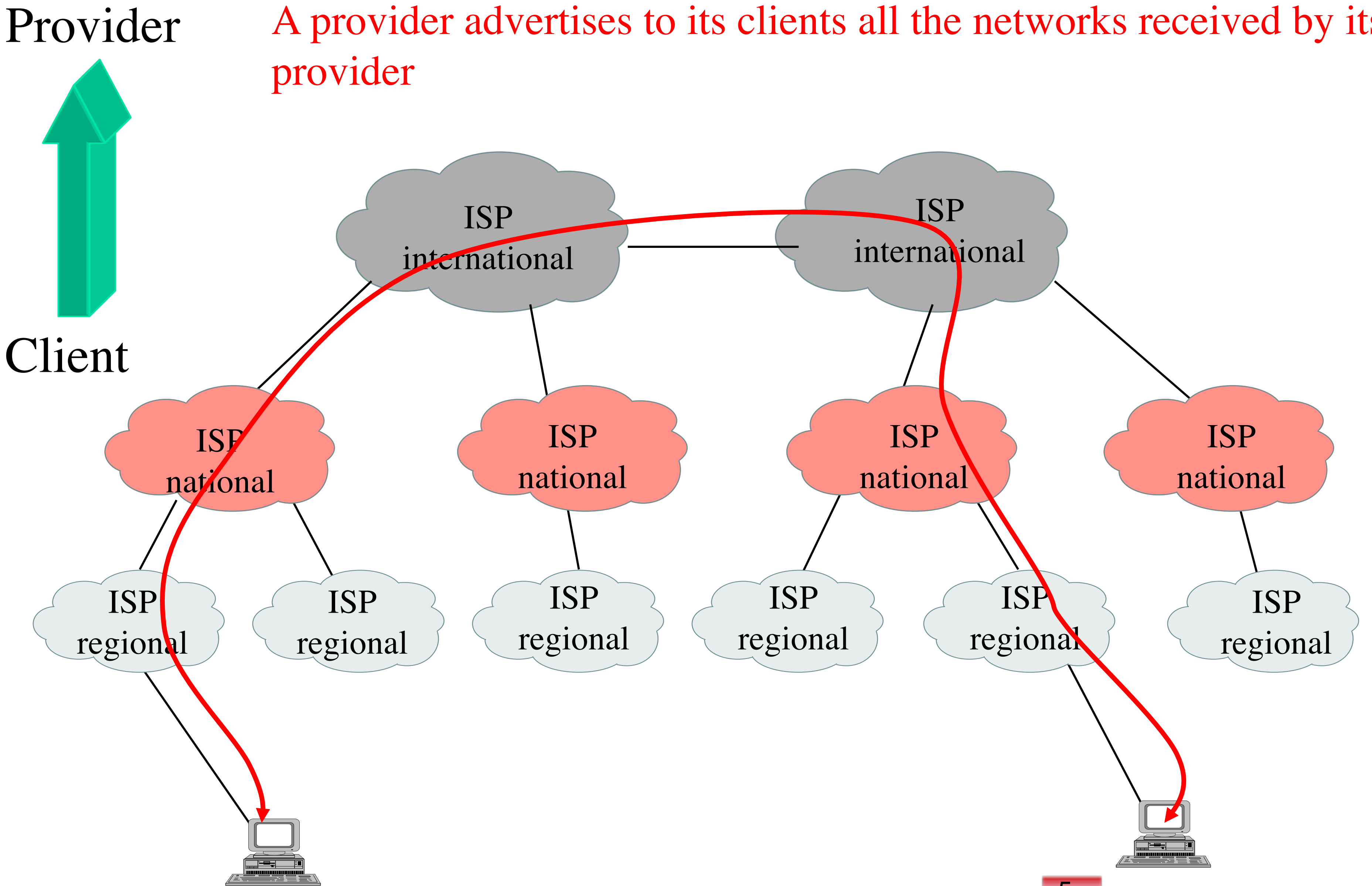
Internet routing: context

- 1992: Need to introduce a more efficient and robust external routing. Progressive introduction of BGP (Border Gateway Protocol).

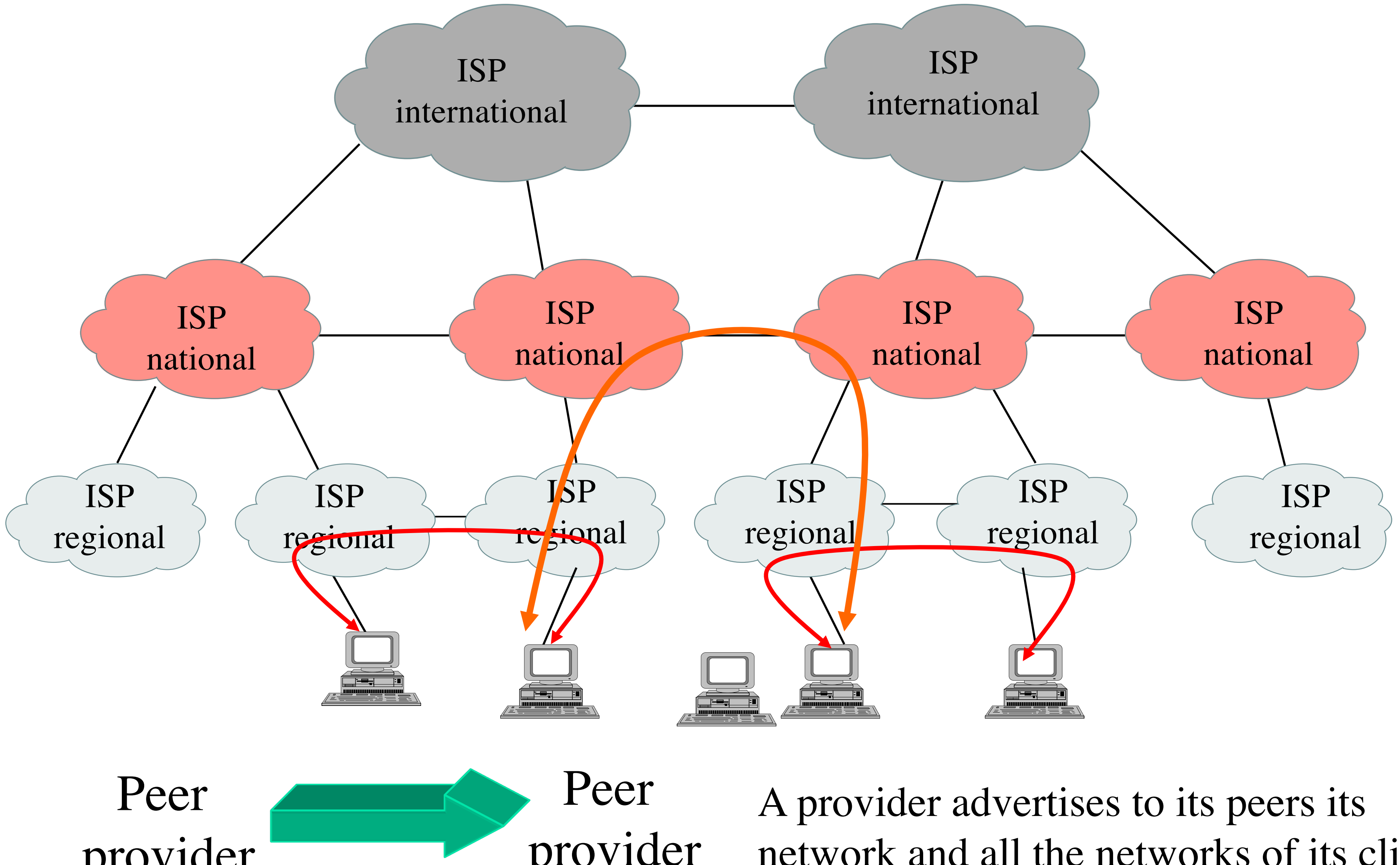


Transit business model

A provider advertises to its clients all the networks received by its provider

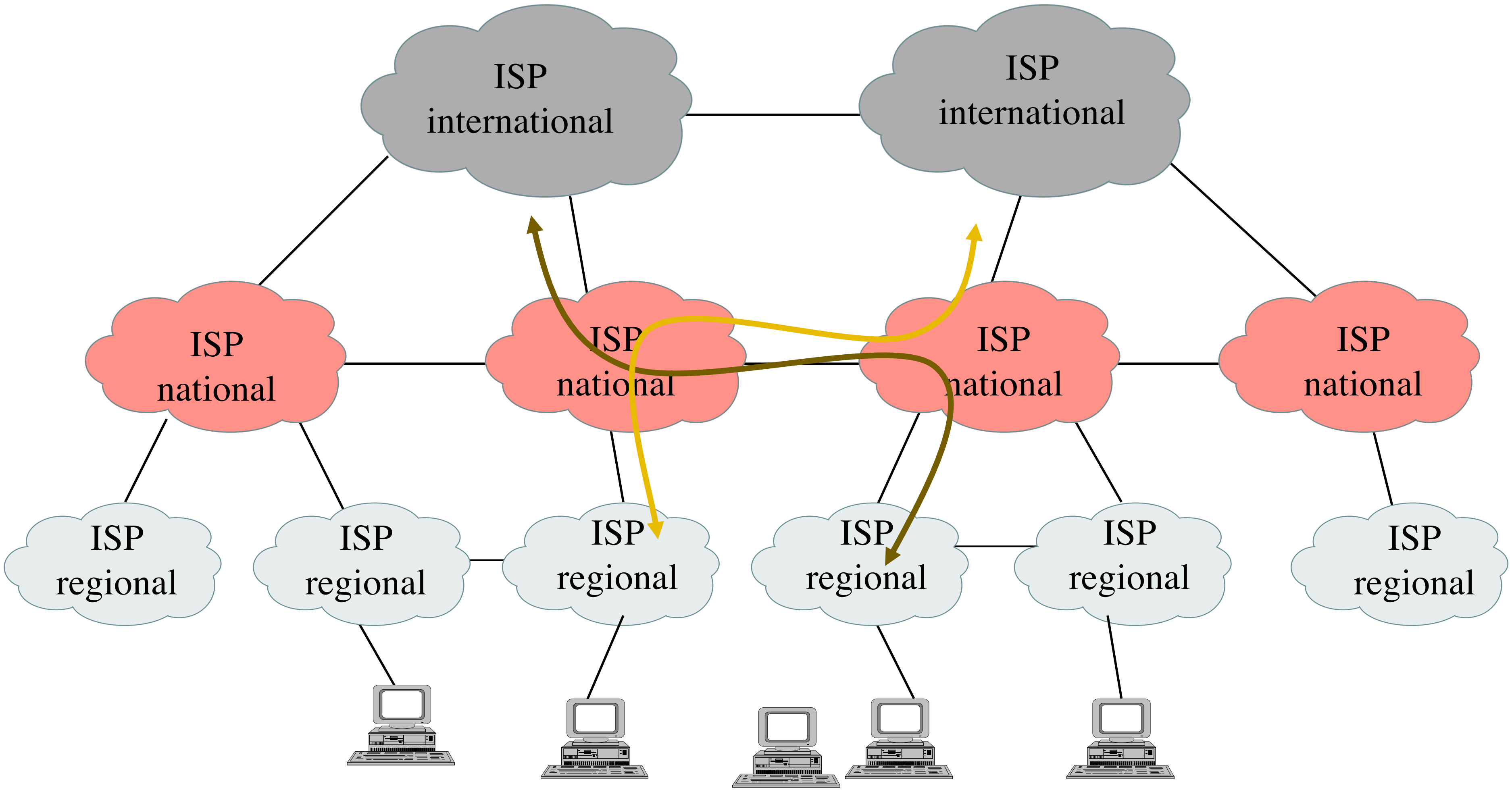


Peering business model



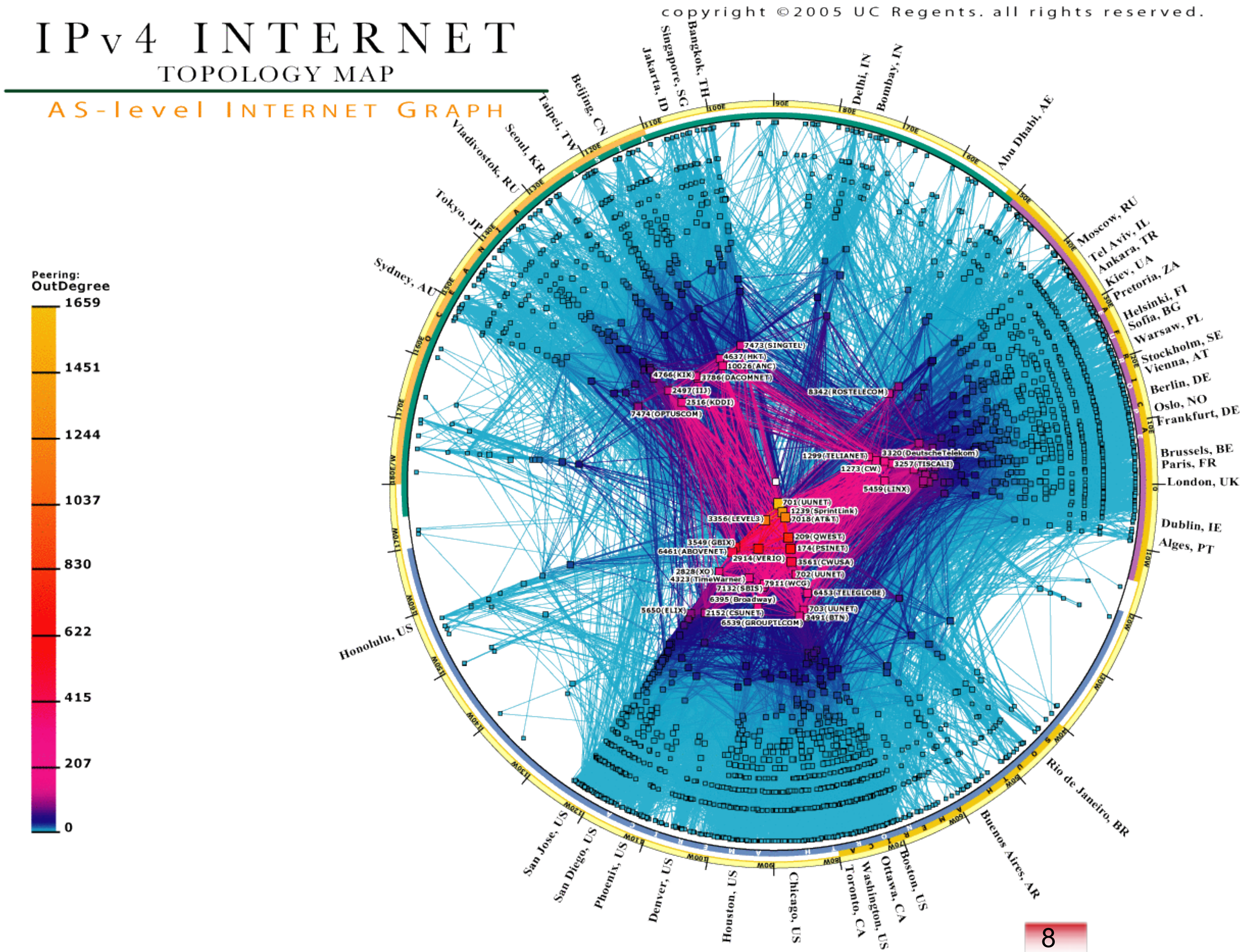
A provider advertises to its peers its network and all the networks of its clients

A new arising model: mutual-transit or paid-peering



A provider advertises to another provider part of its upstream provider cone.
Can be free of charge if opportunely balanced, or can be subject to payment (a sort of « paid peering), or can be activated only as backup agreement

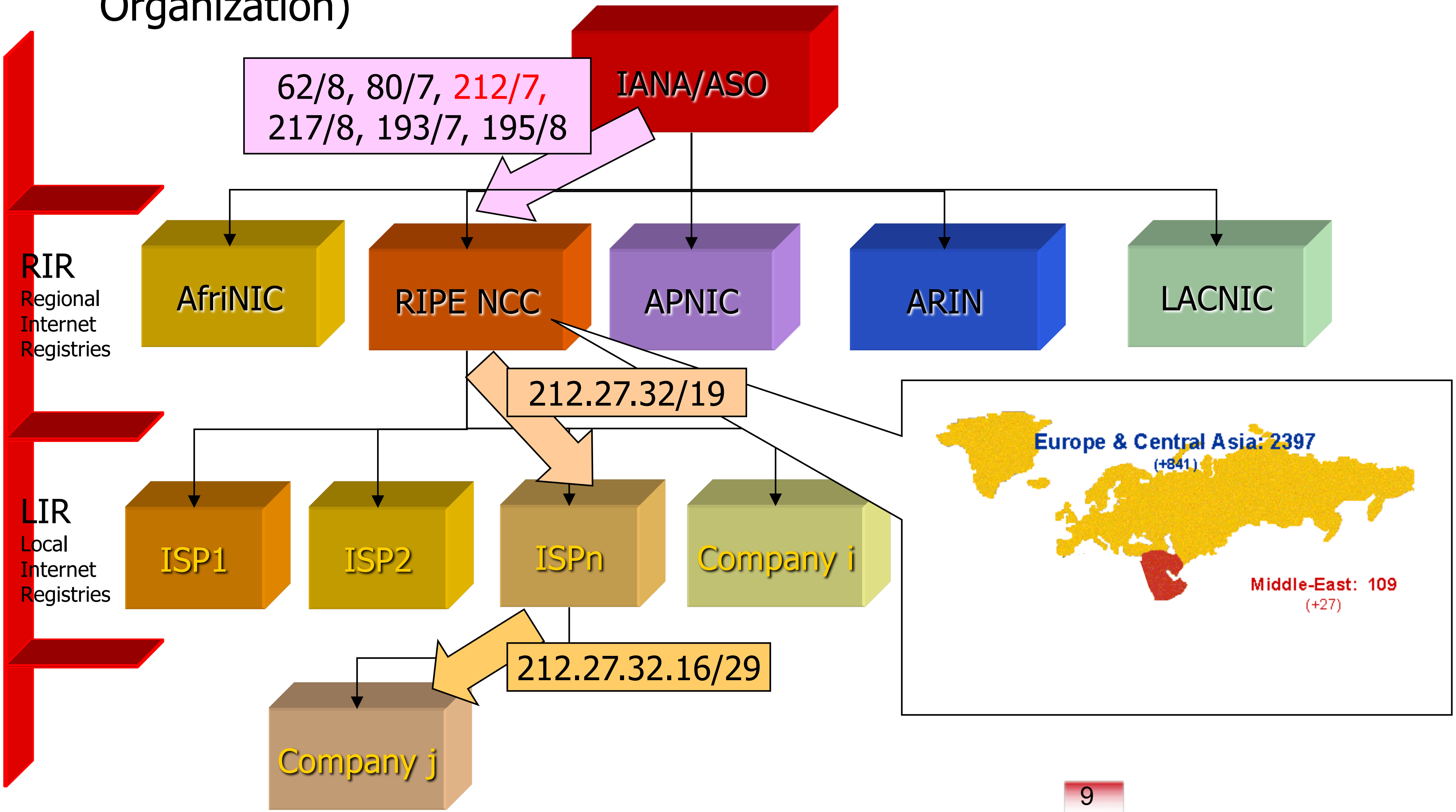
Internet Topology (AS map)



Source: www.caida.org

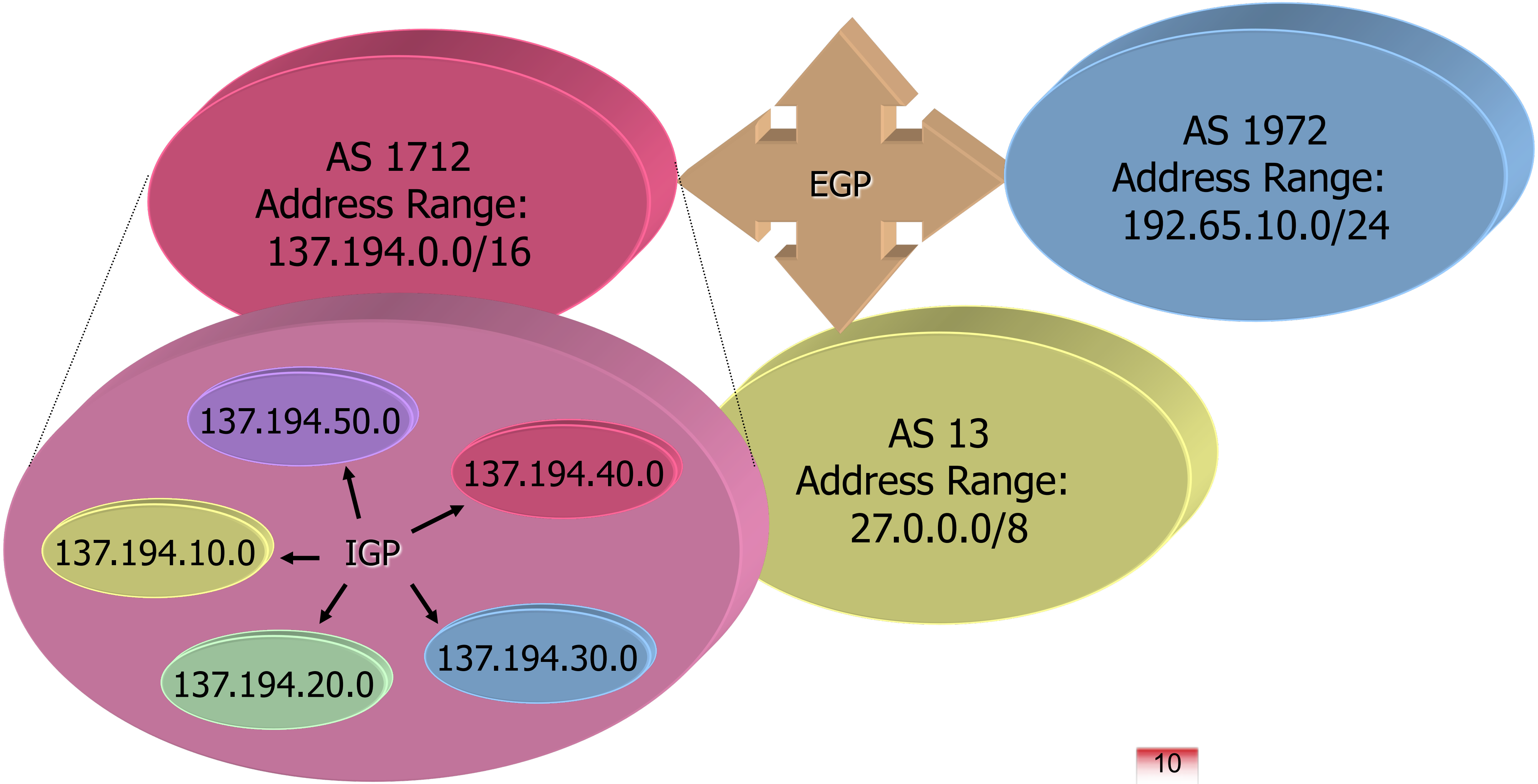
Internet Routing Architecture: Address Assignment

- Internet Hierarchical Political Organization (Address Supporting Organization)



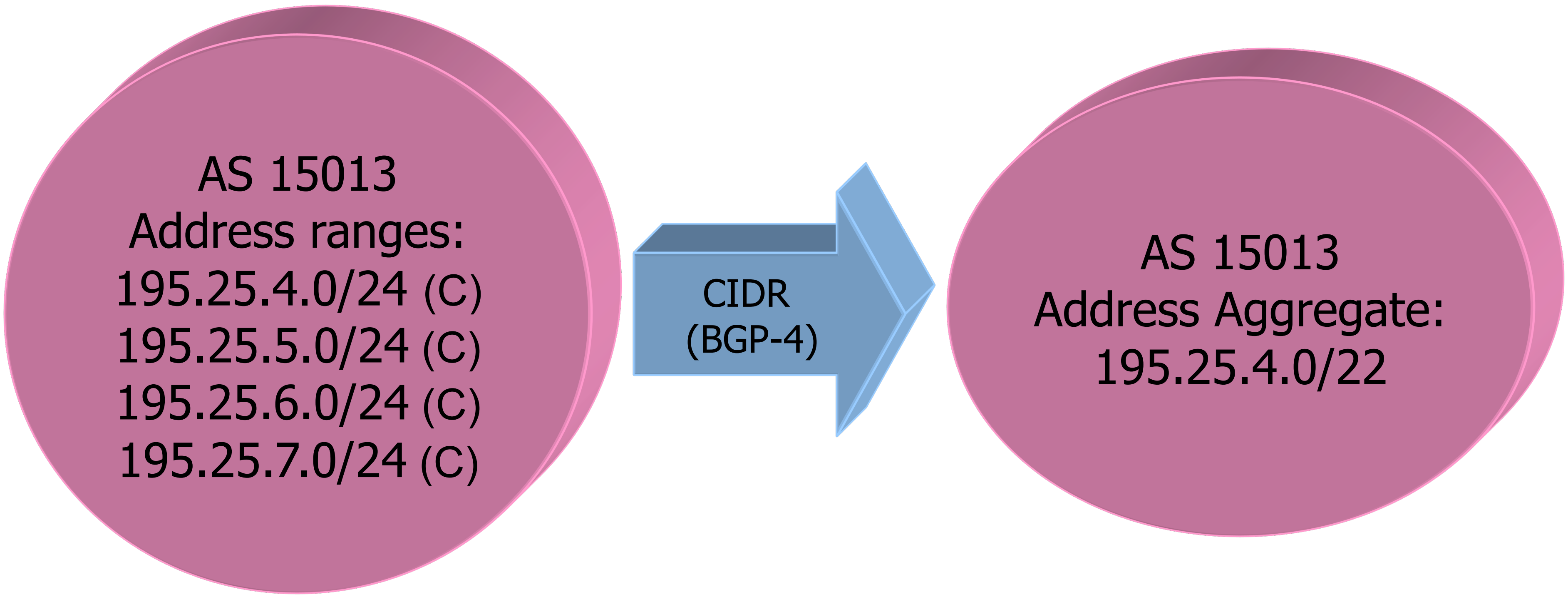
Internet Hierarchical Architecture: IGP/EGP

- Address Assignment:
  - Organization based (no hierarchical structure)
  - Subnetting within an AS (manipulated by IGP)

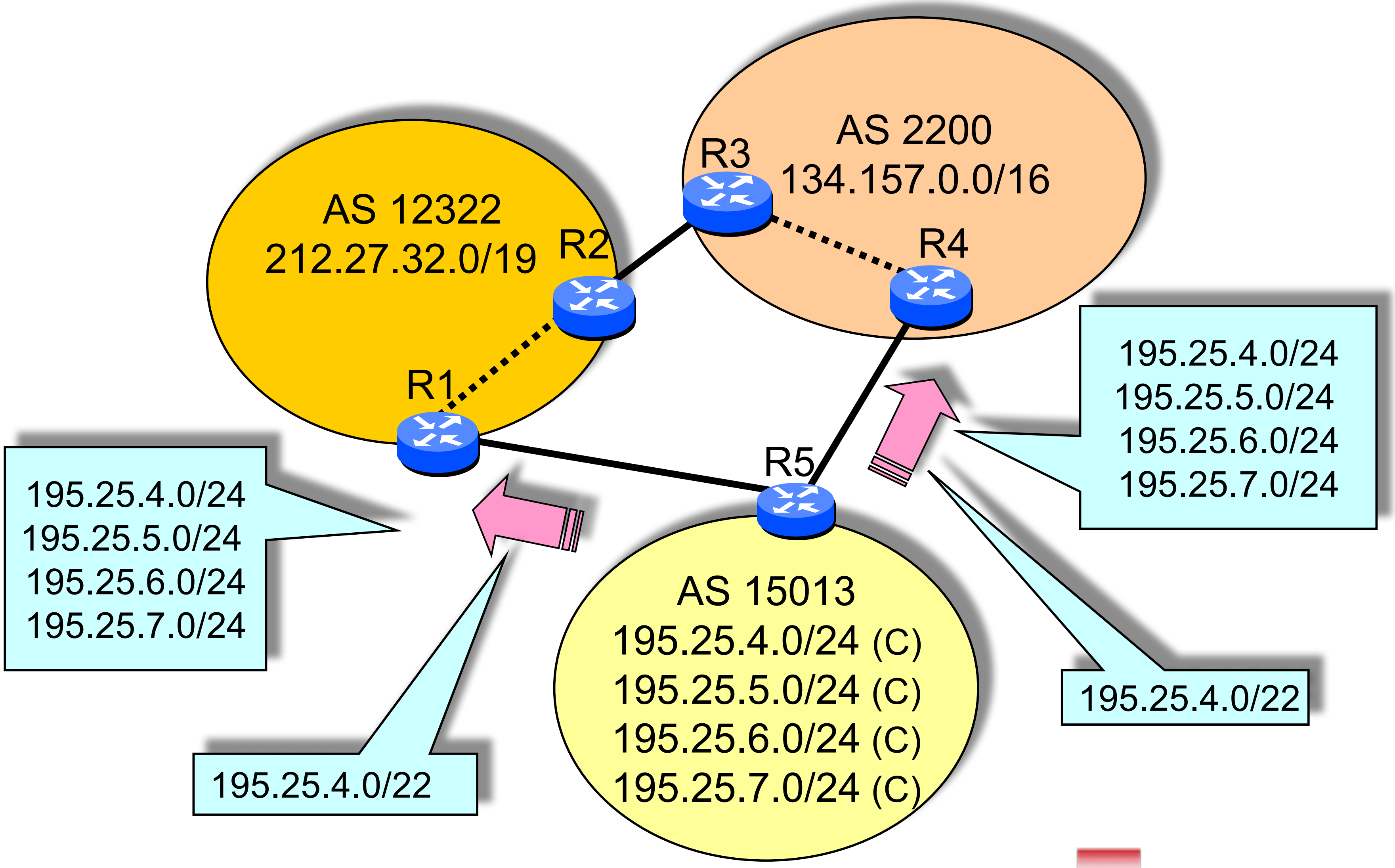


Internet Routing Architecture: CIDR

- CIDR (Class-Less Inter-Domain Routing)

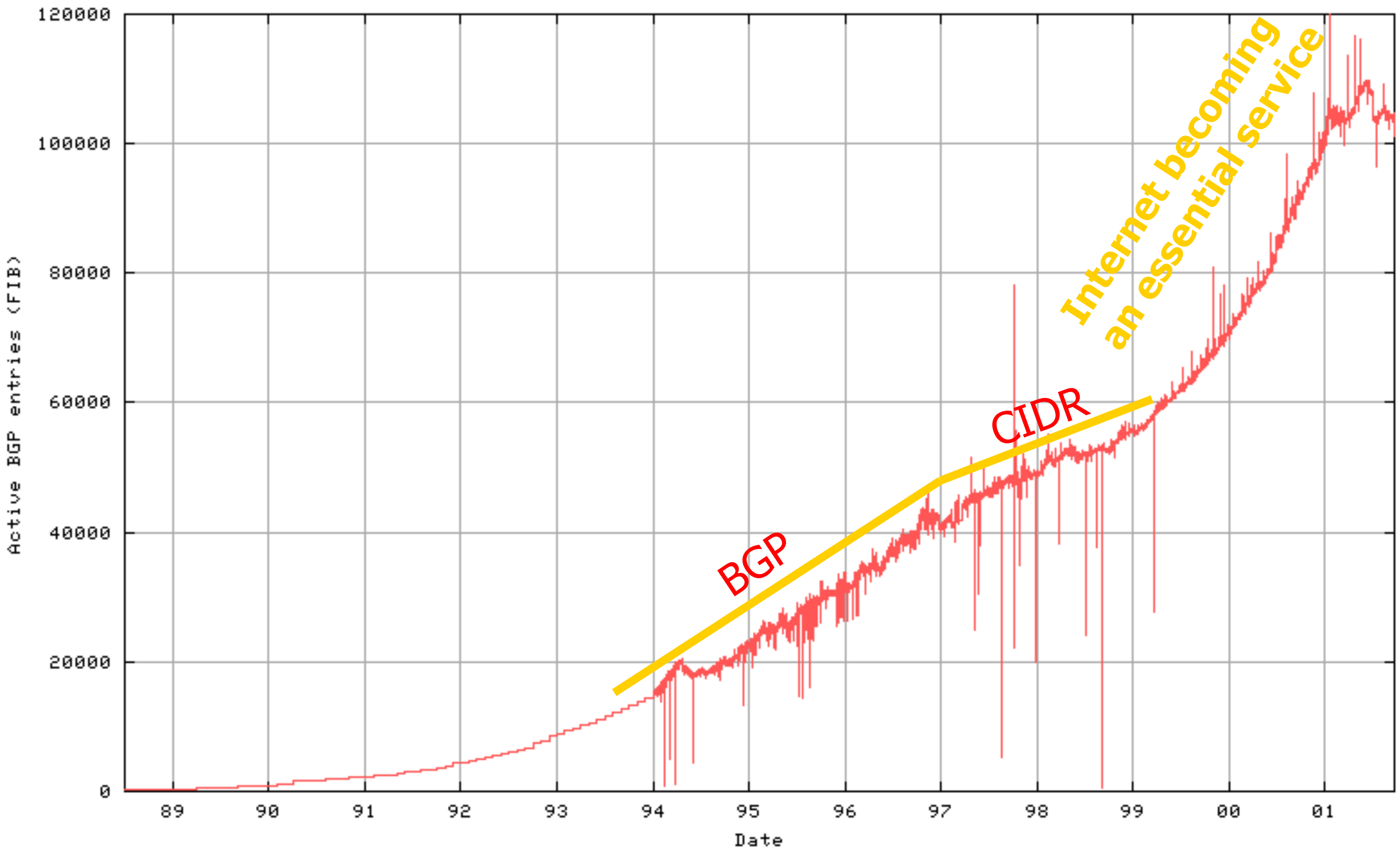


CIDR aggregation



The power(less) of the hierarchy

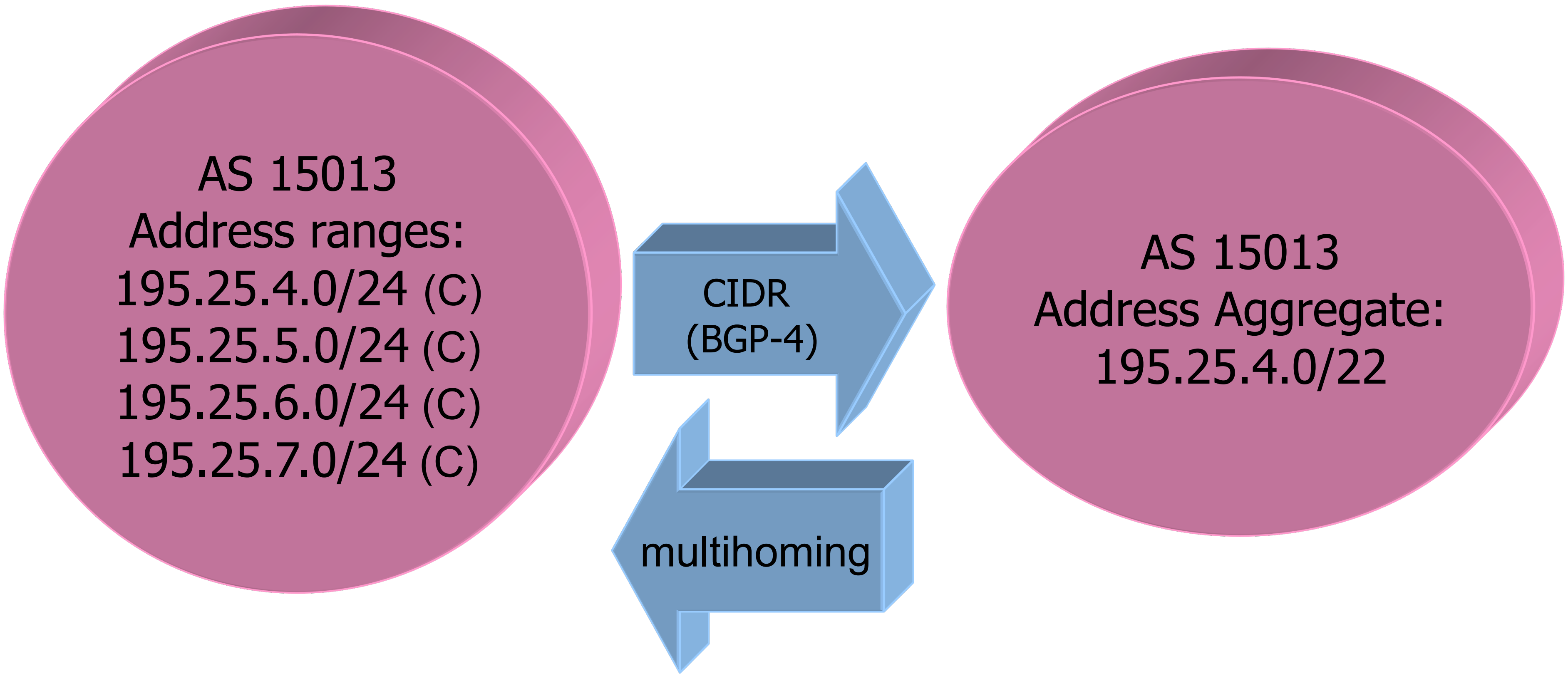
- Important increase of routing table size since mid 99.
  - Main reason in 99-01: many new Ases



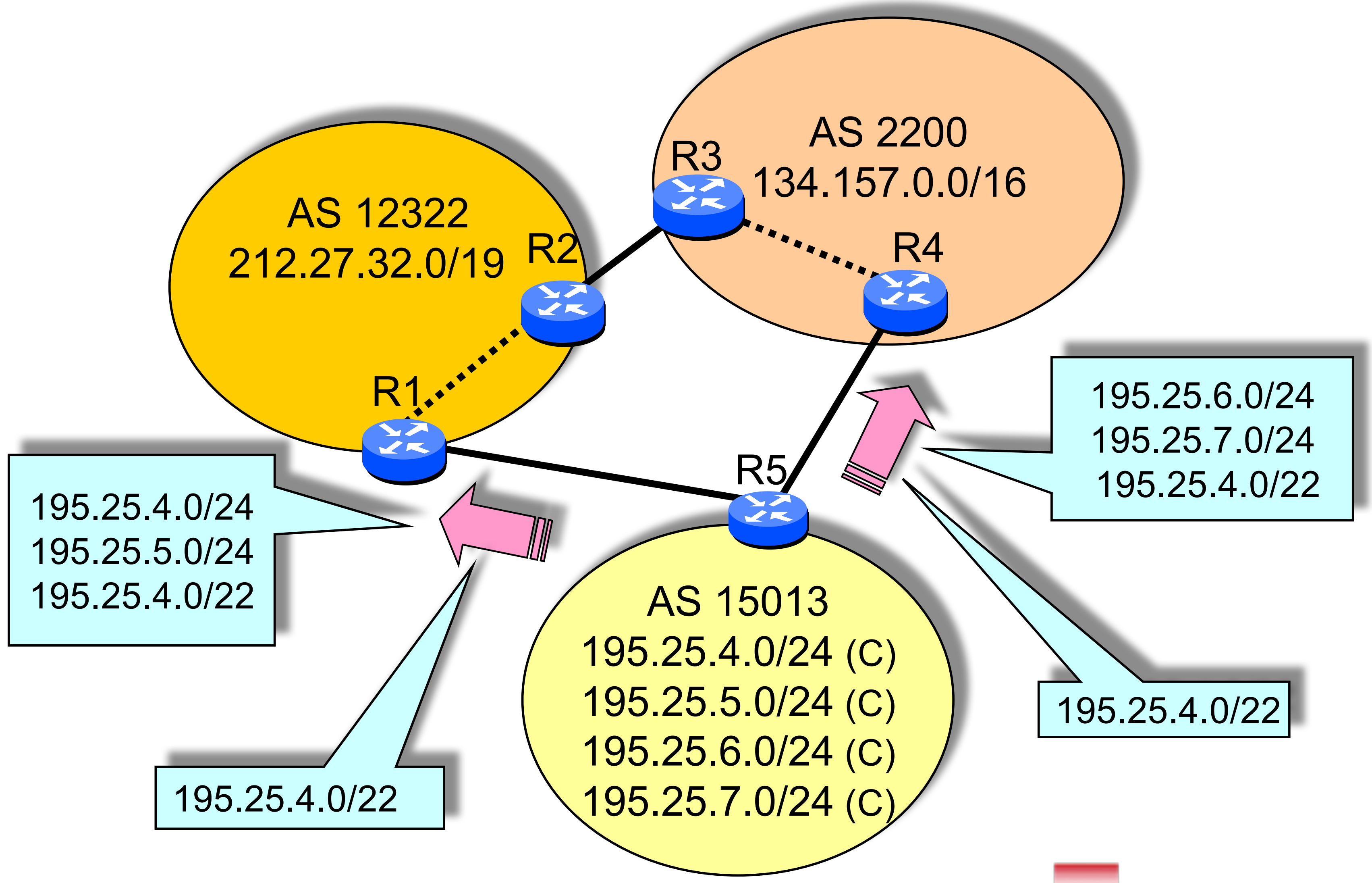
(source: bgp.potaroo.net)

Scalability: Routing tables size

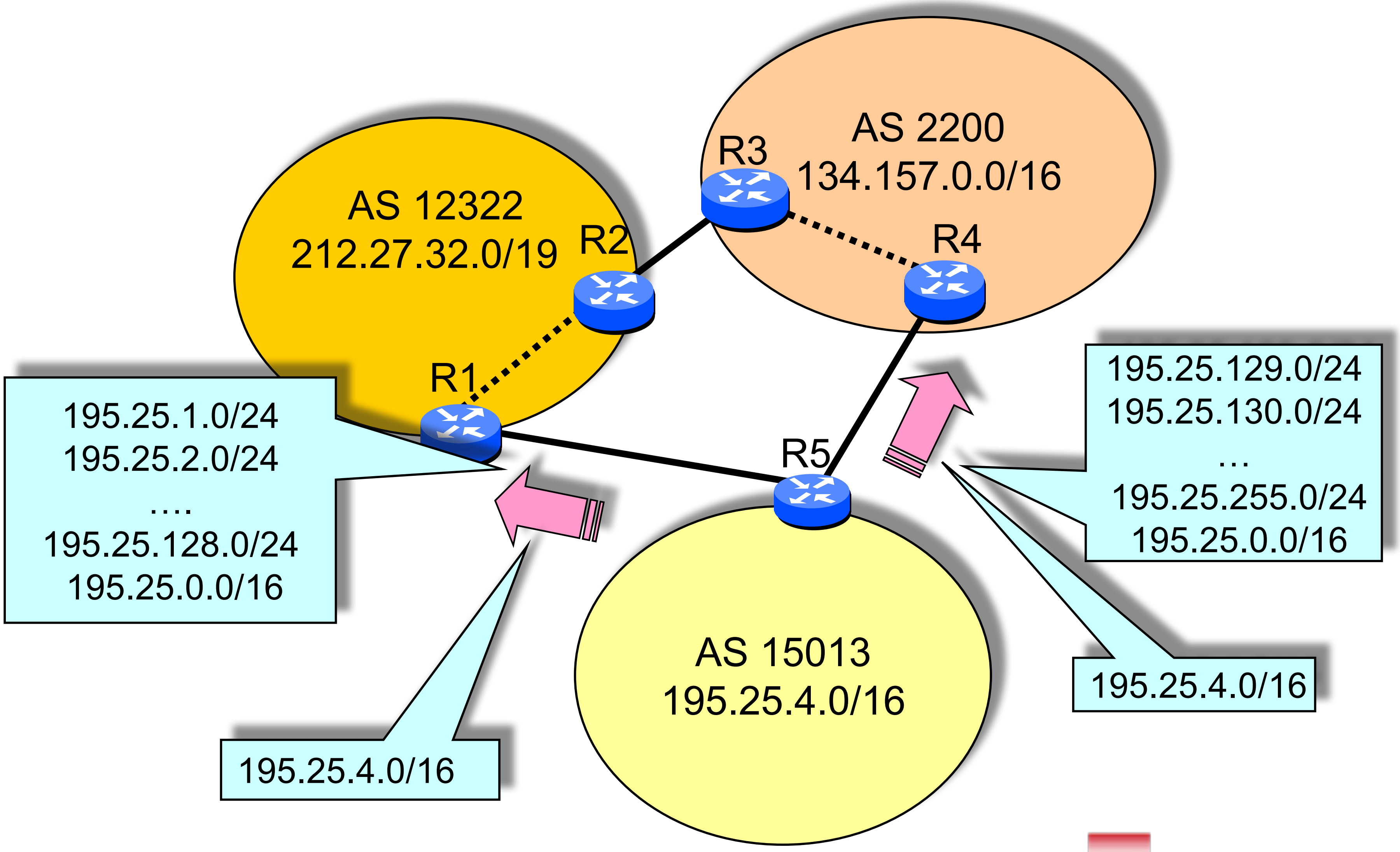
- CIDR (Class-Less Inter-Domain Routing)



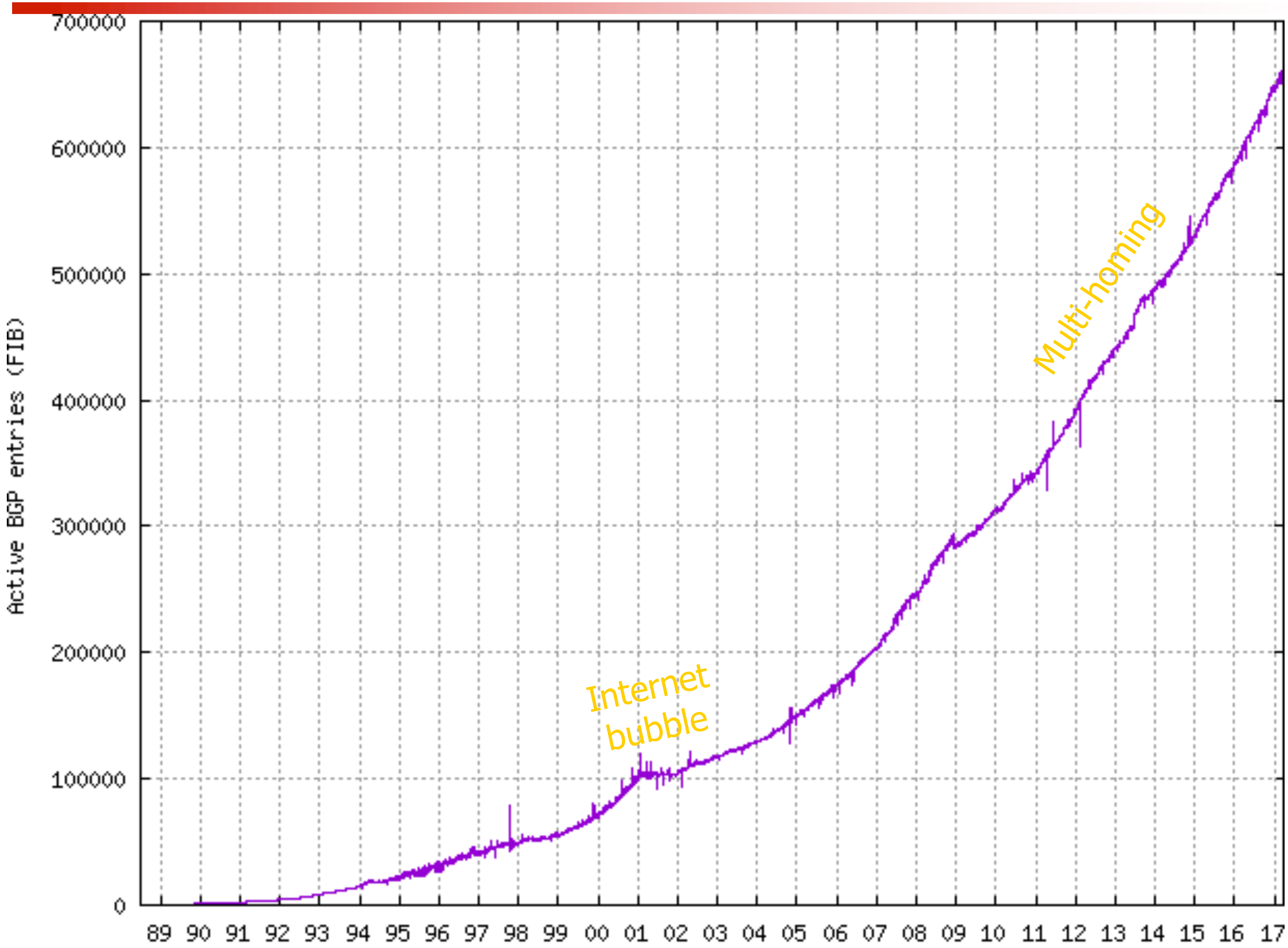
Multihoming: prefix desaggregation



Multihoming: even worst than without CIDR



Current situation: fast growth!!



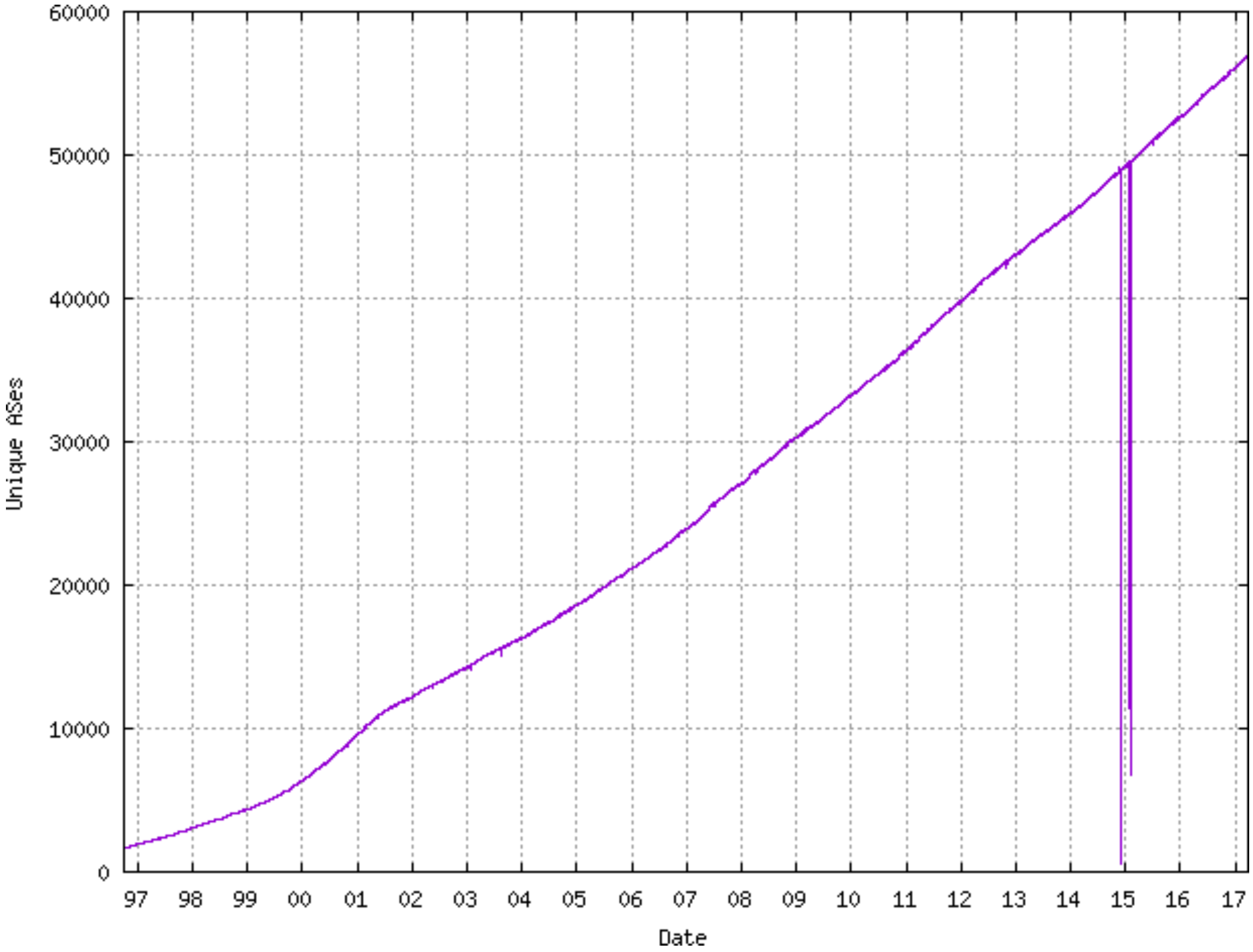
(source: bgp.potaroo.net)

How much memory for so many entries?

** 18 April 2012

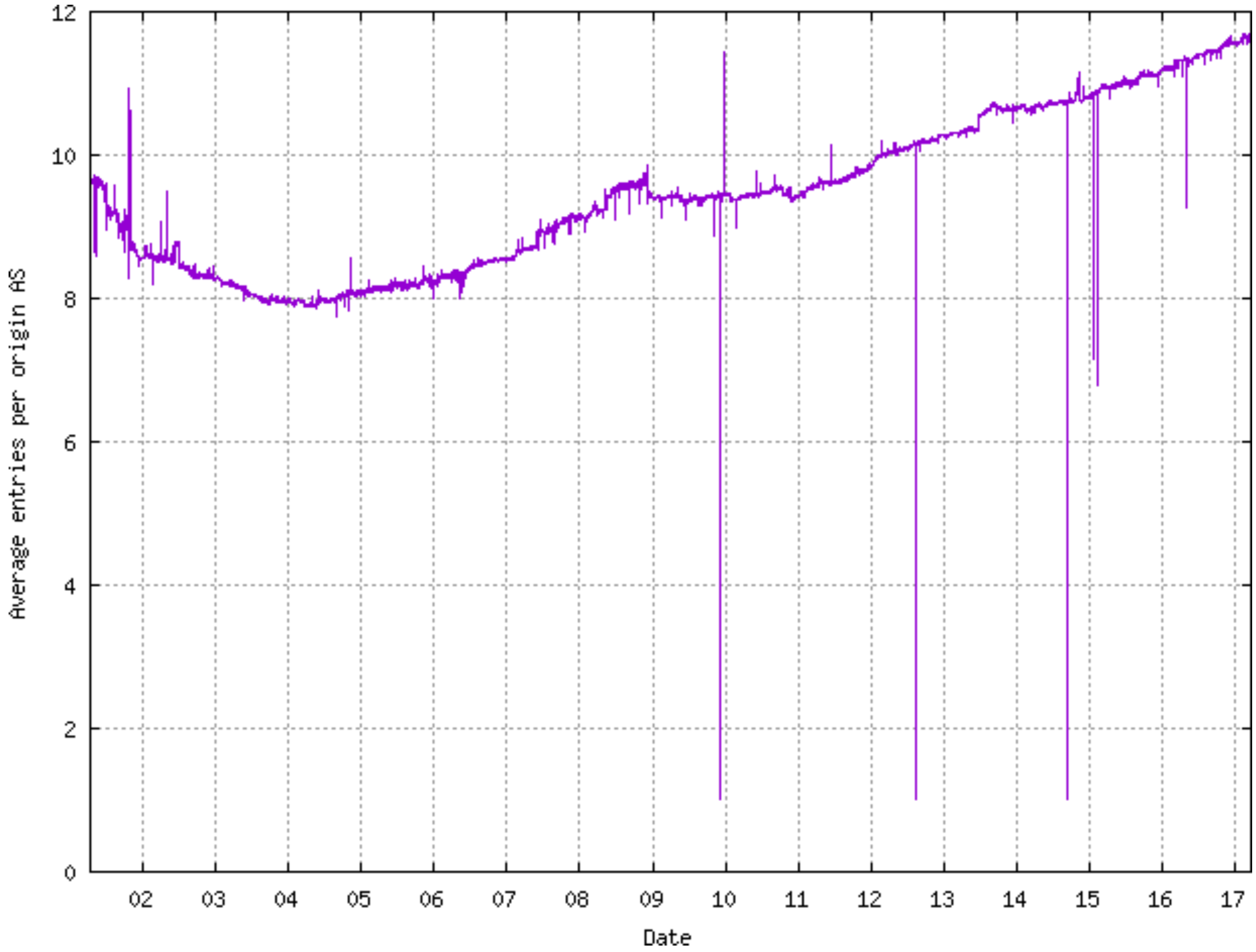
```
$ bgpctl show rib mem
RDE memory statistics
   404372 IPv4 unicast network entries using 15.4M of memory
    63223 IPv6 unicast network entries using 3.4M of memory
   935183 rib entries using 57.1M of memory
  4138172 prefix entries using 253M of memory
   559992 BGP path attribute entries using 64.1M of memory
   488476 BGP AS-PATH attribute entries using 29.7M of memory,
          and holding 559992 references
     7811 BGP attributes entries using 305K of memory
          and holding 399505 references
     7810 BGP attributes using 61.1K of memory
RIB using 423M of memory
```

AS number... a fast increase !



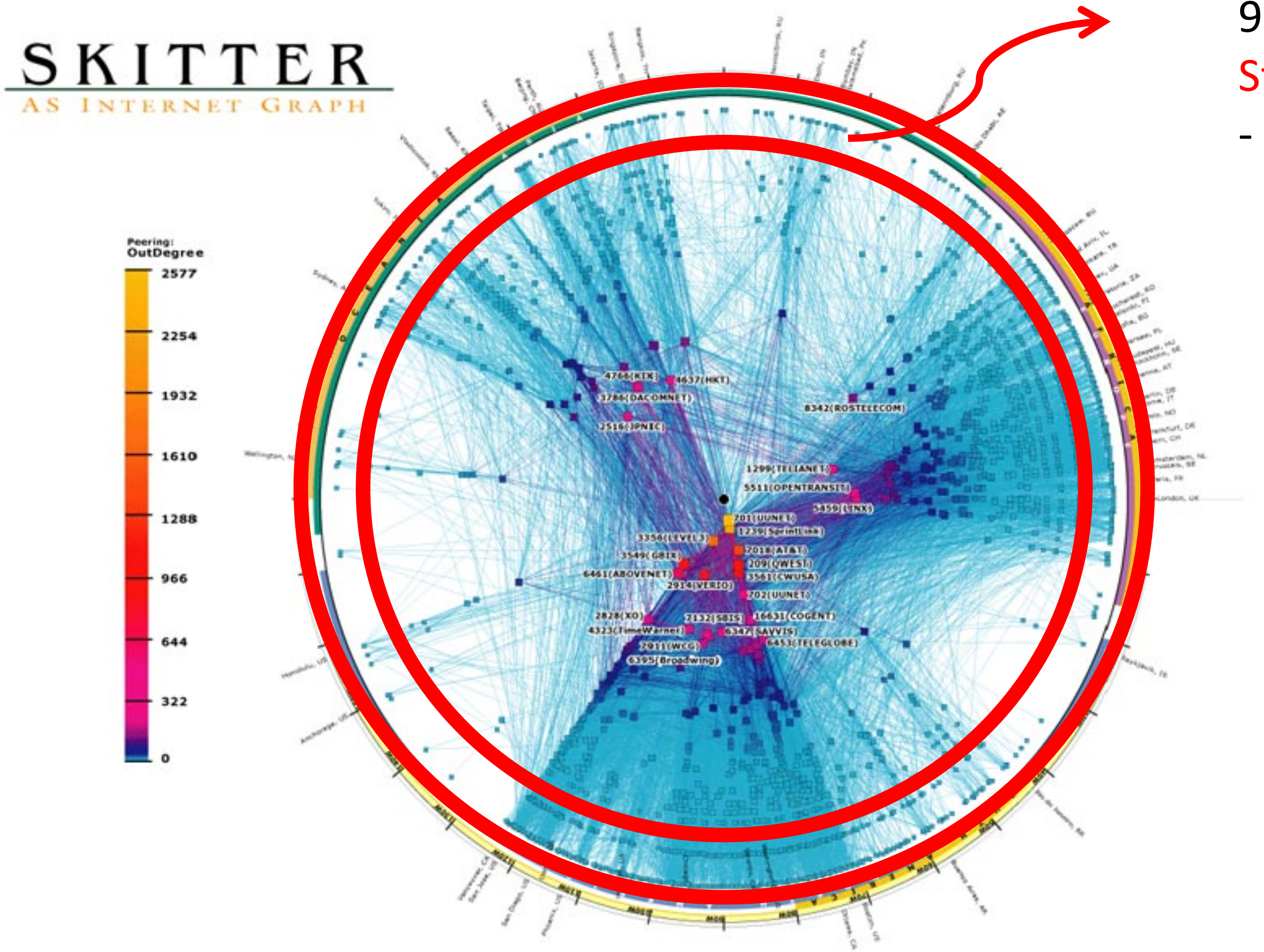
(source: bgp.potaroo.net)

An insight on the growth



(source: bgp.potaroo.net)

Where are most of the ASs?

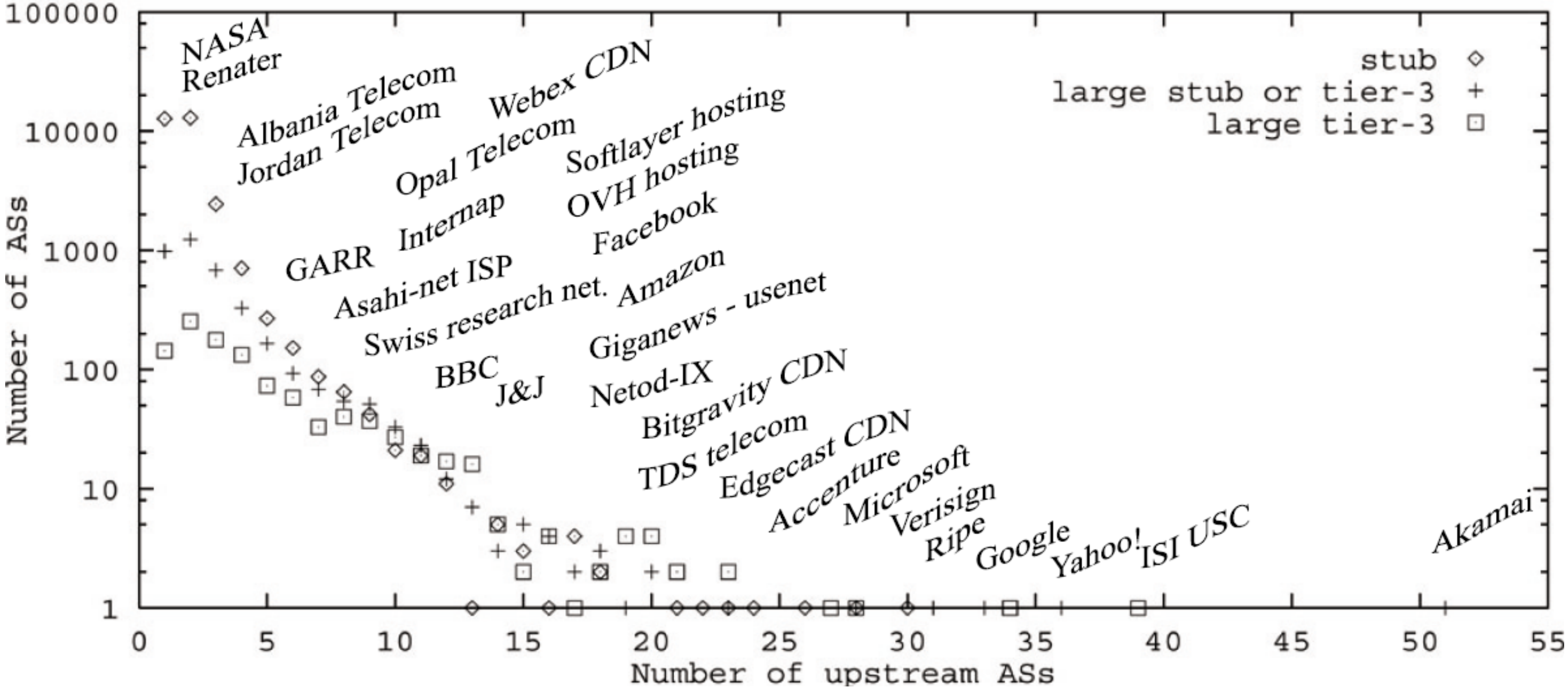


94% of the ASs are here!

Stub ASs:

- do not transit traffic

Edge ASs: multi-homing behavior



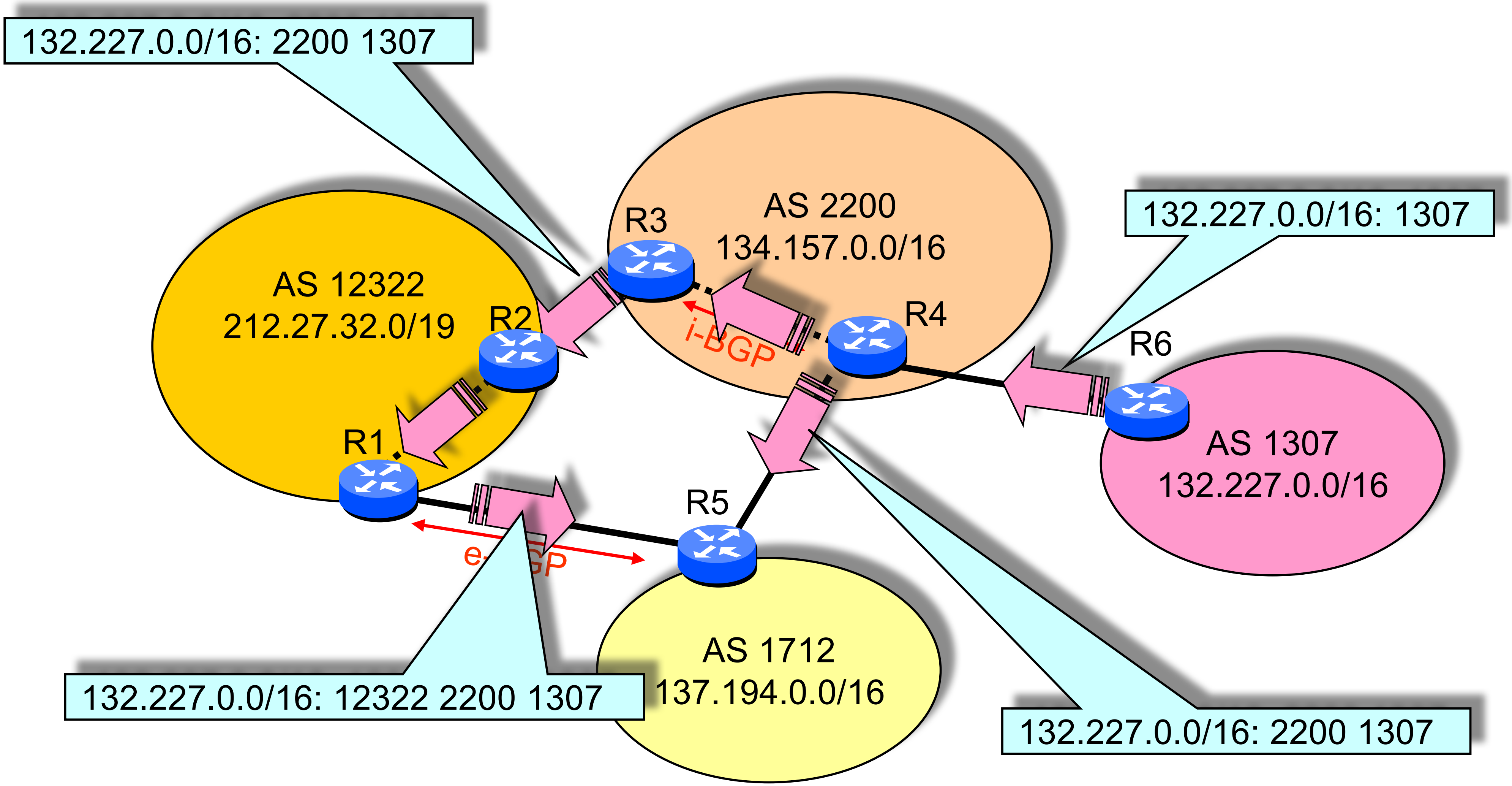
EGP routing decision process

THE BGP PROTOCOL

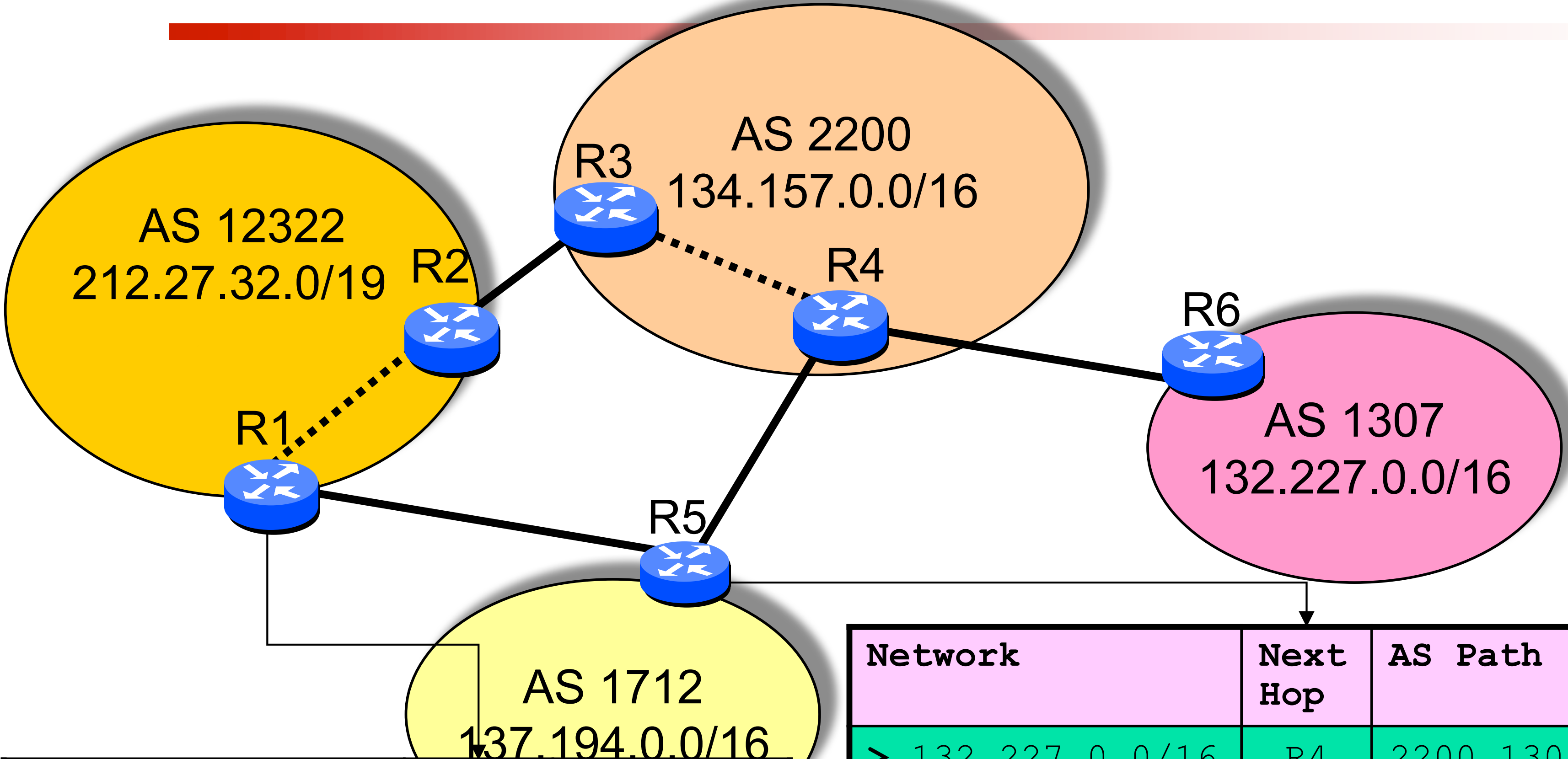
BGP: Rationales

- Choice of the routing technology:
  - Link State can't be used for internet wide graph
    - Use of hierarchy seems difficult to implement, in particular for political reasons
  - Distance Vector scales…
    - But has robustness issues that need to be addressed.
    - Proposed extensions for robustness:
      - Path Vector (see below) for loop avoidance
      - Incremental updates (scalability, limitation of overhead)
      - Runs on top of TCP (robustness)

BGP: Path Vector



BGP: Path Vector (2)



Network	Next Hop	AS Path
> 132.227.0.0/16	R2	2200 1307
> 134.157.0.0/16	R2	2200
> 137.194.0.0/16	R5	1712
> 212.27.32.0/19	--	12322

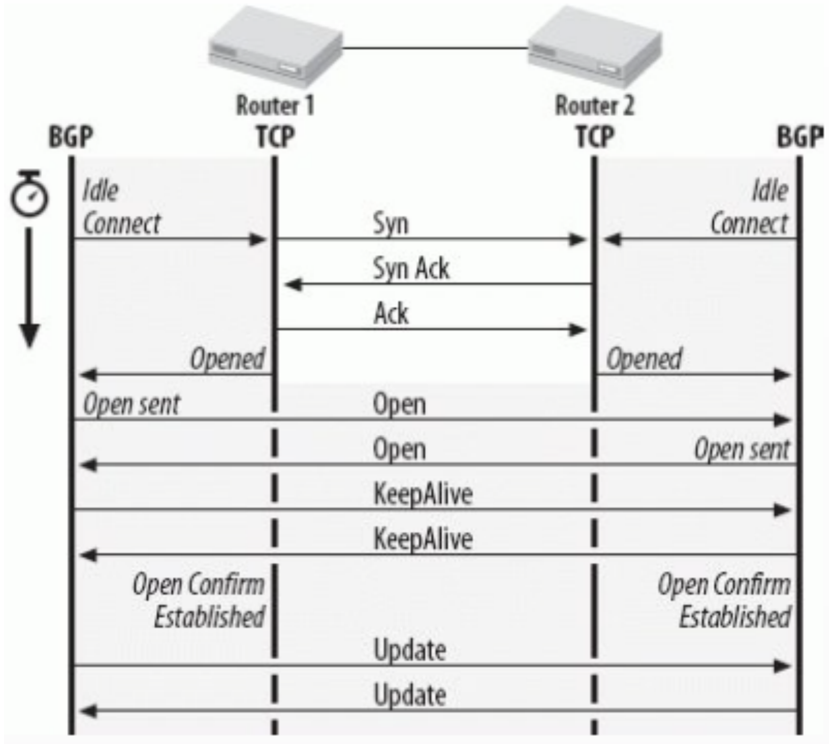
Network	Next Hop	AS Path
> 132.227.0.0/16	R4	2200 1307
132.227.0.0/16	R1	12322 2200 1307
> 134.157.0.0/16	R4	2200
134.157.0.0/16	R1	12322 2200
> 137.194.0.0/16	--	1712
> 212.27.32.0/19	R1	12322
212.27.32.0/19	R4	2200 12322

Incremental Updates

- Advertisements are only sent when their content changes
  - Avoid to limit volume of routing overhead exchanged.
  - But requires caching of neighbor advertisements (Adjacency RIB)

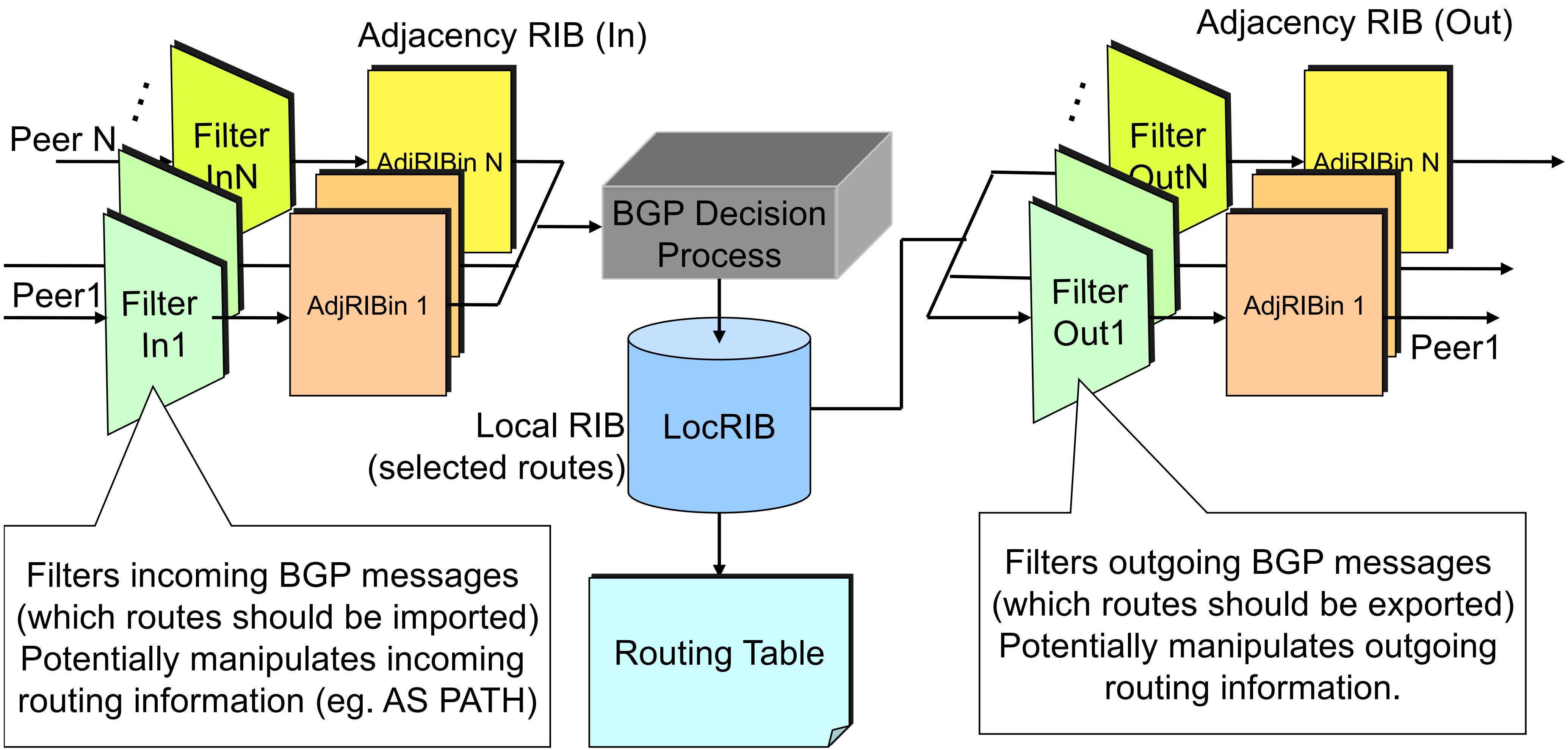
- Incremental update insured by means of two main messages:
  - Route Update
  - Route Withdraw

- Notes:
  - Import and Export filters used to control routes exchanged with neighbors
    - only acceptable routes are cached
  - Refreshes are insured by simple keep alive messages.



BGP Router functional architecture

RIB: Routing Information Base

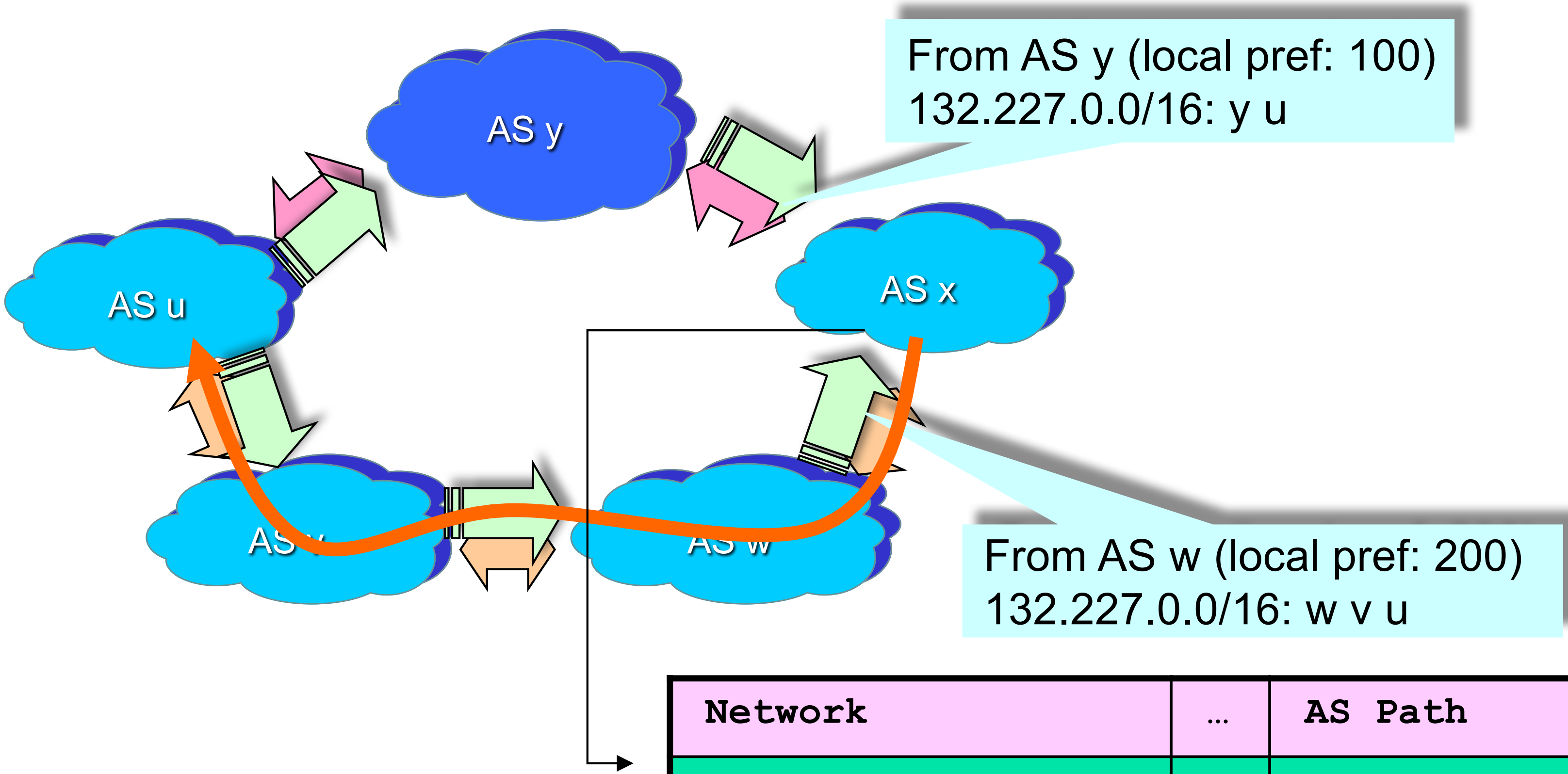


BGP: Policy Routing

- Why "policy routing" ?
  - Of paramount importance in order to segregate peers w.r.t transit AS, for instance.
- BGP-4 « Policy-routing » support.
  - Through filtering and BGP information manipulations (eg. AS Path)
  - Different criteria are used for path selection thanks to the complex BGP Decision process.
    - Local preference attribution.
    - AS Hop count (from the AS Path Vector)
    - MED (Multi Exit Discriminator)
    - "Hot Potato" (eBGP versus iBGP, closest Next Hop)
    - Tie Breaking

BGP Policies: Local preference

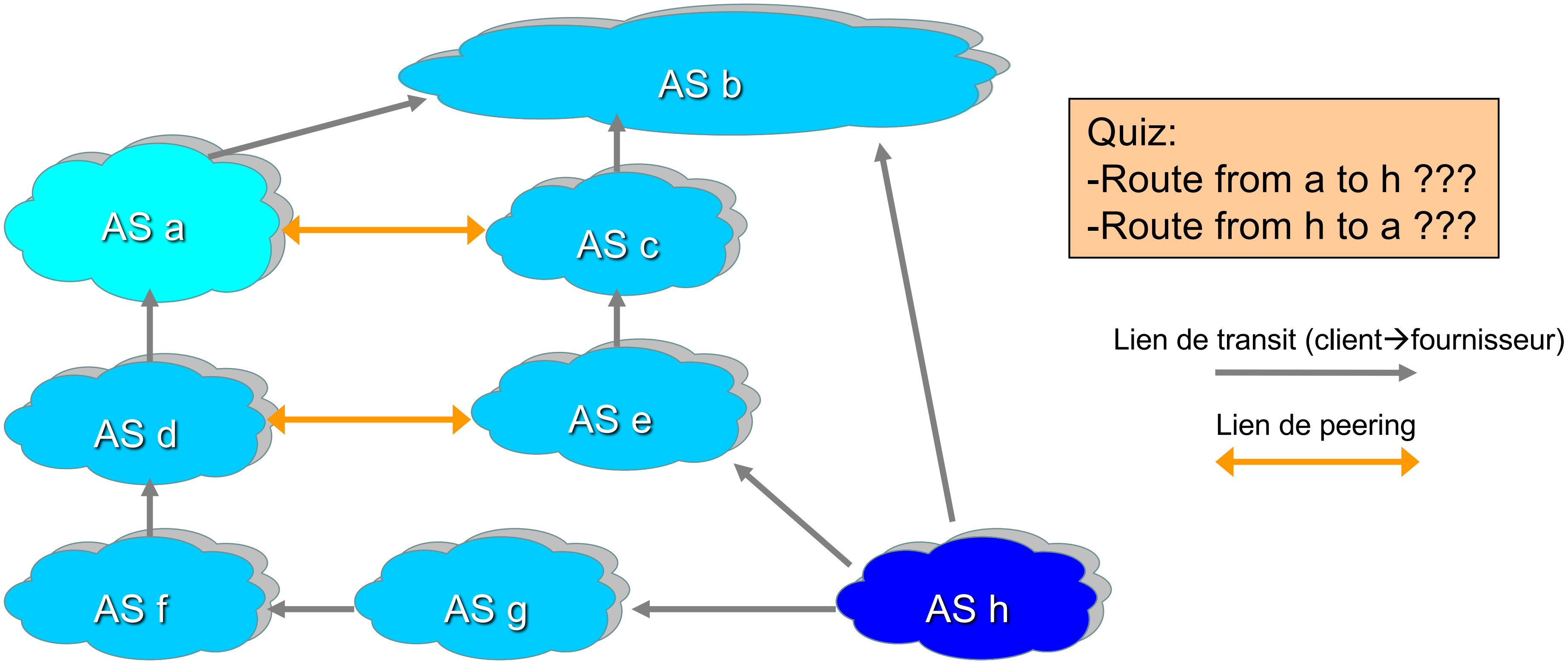
- Peers are given a local preference. BGP Updates with the highest preference are chosen.
- Limitations:
  - For Out-bound traffic control only (i.e. incoming routes)
  - Local policies only (peer-to-peer and not end-to-end policy scope)



Network	...	AS Path
132.227.0.0/16		Y,U
> 132.227.0.0/16		W,V,U

BGP Policies: Local preference (2)

- Possible use of LocalPref for ISPs (transit AS):
  - High Local-Pref (100) for routes received from customers
  - Medium Local-Pref (50) for routes received from peers
  - Low Local-Pref (0) for routes received from a provider
- Consequences:
  - Asymmetry of Internet routes.

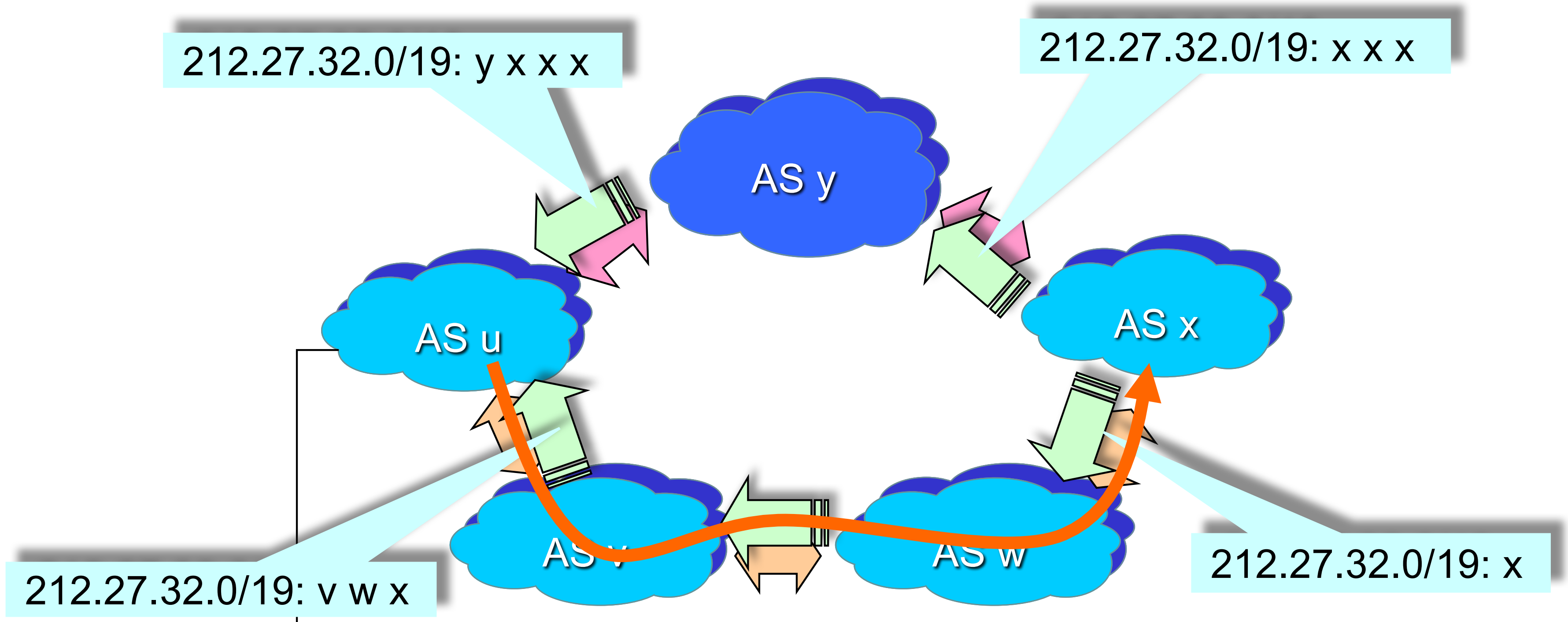


BGP: Policy Routing

- Why “policy routing” ?
  - Of paramount importance in order to segregate peers w.r.t transit AS, for instance.
- BGP-4 « Policy-routing » support.
  - Through filtering and BGP information manipulations (eg. AS Path)
  - Different criteria are used for path selection thanks to the complex BGP Decision process.
    - Local preference attribution.
    - AS Hop count (from the AS Path Vector)
    - MED (Multi Exit Discriminator)
    - “Hot Potato” (eBGP versus iBGP, closest Next Hop)
    - Tie Breaking

BGP Policies: Path pre-pending

- AS-Path is manipulated in order to control incoming traffic path



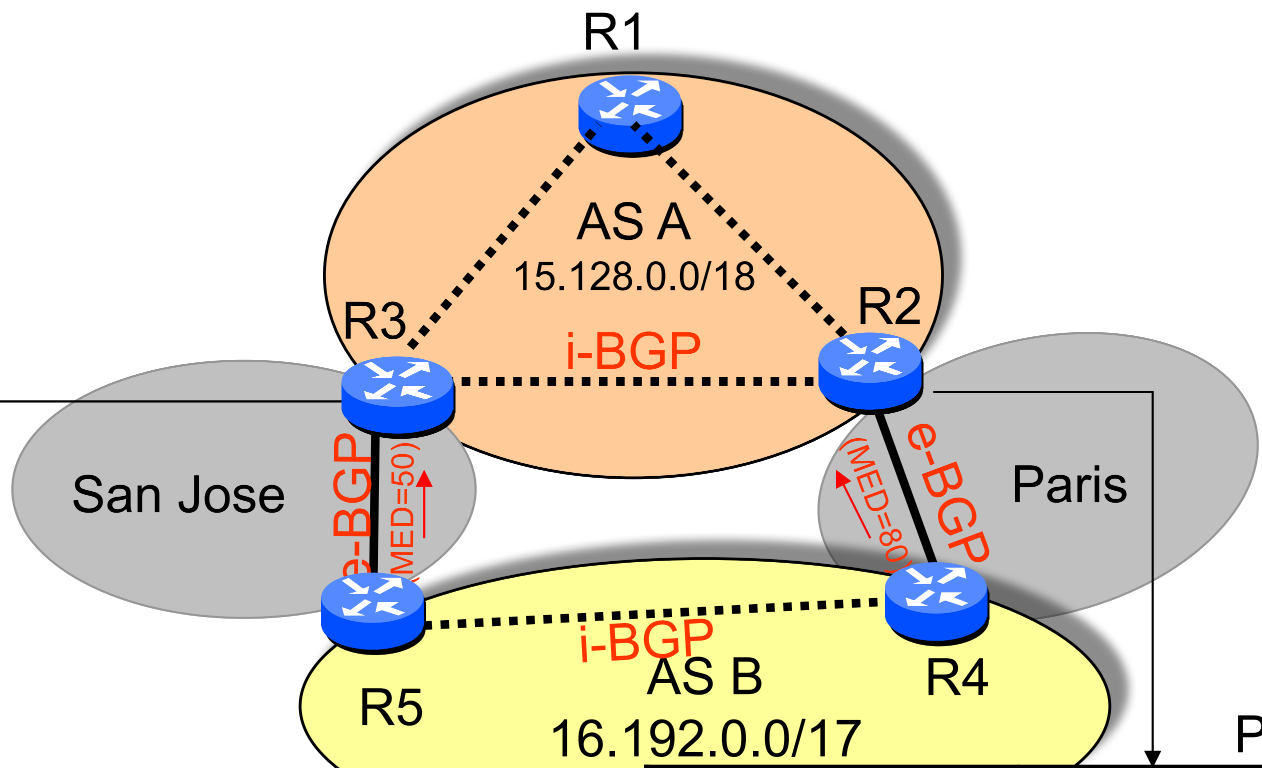
Network	...	AS Path
212.27.32.0/19		Y X X X
> 212.27.32.0/19		V W X

BGP: Policy Routing

- Why "policy routing" ?
  - Of paramount importance in order to segregate peers w.r.t transit AS, for instance.
- BGP-4 « Policy-routing » support.
  - Through filtering and BGP information manipulations (eg. AS Path)
  - Different criteria are used for path selection thanks to the complex BGP Decision process.
    - Local preference attribution.
    - AS Hop count (from the AS Path Vector)
    - MED (Multi Exit Discriminator)
    - "Hot Potato" (eBGP versus iBGP, closest Next Hop)
    - Tie Breaking

BGP: Multiple Exit Discriminator (MED)

Cold potato MED routing: only for routes acquired from the same AS border



Paris Router :

Network	Next HOP	AS Path
16.192.0.0/17	R4	B (from eBGP, MED=80)
> 16.192.0.0/17	**R3**	**B i** (from iBGP, **MED=50)**

San Jose Router :

Network	Next HOP	AS Path
16.192.0.0/17	R2	B i (from iBGP, MED=80)
> 16.192.0.0/17	**R5**	**B** (from eBGP, **MED=50)**

BGP: Multiple Exit Discriminator (MED) (2)

- A lower MED is preferred over a higher MED
  - The lower MED rule is also called "cold potato" rule
  - A MED attribute that is received by an AS does not leave the AS
    - i.e., it has a per-AS scope

- In practice,
  - Often disabled because it may lead to oscillations, e.g. with route reflectors (see after)
  - If used, only for transit agreements (customers pay for) and not for peering agreements (free transit)

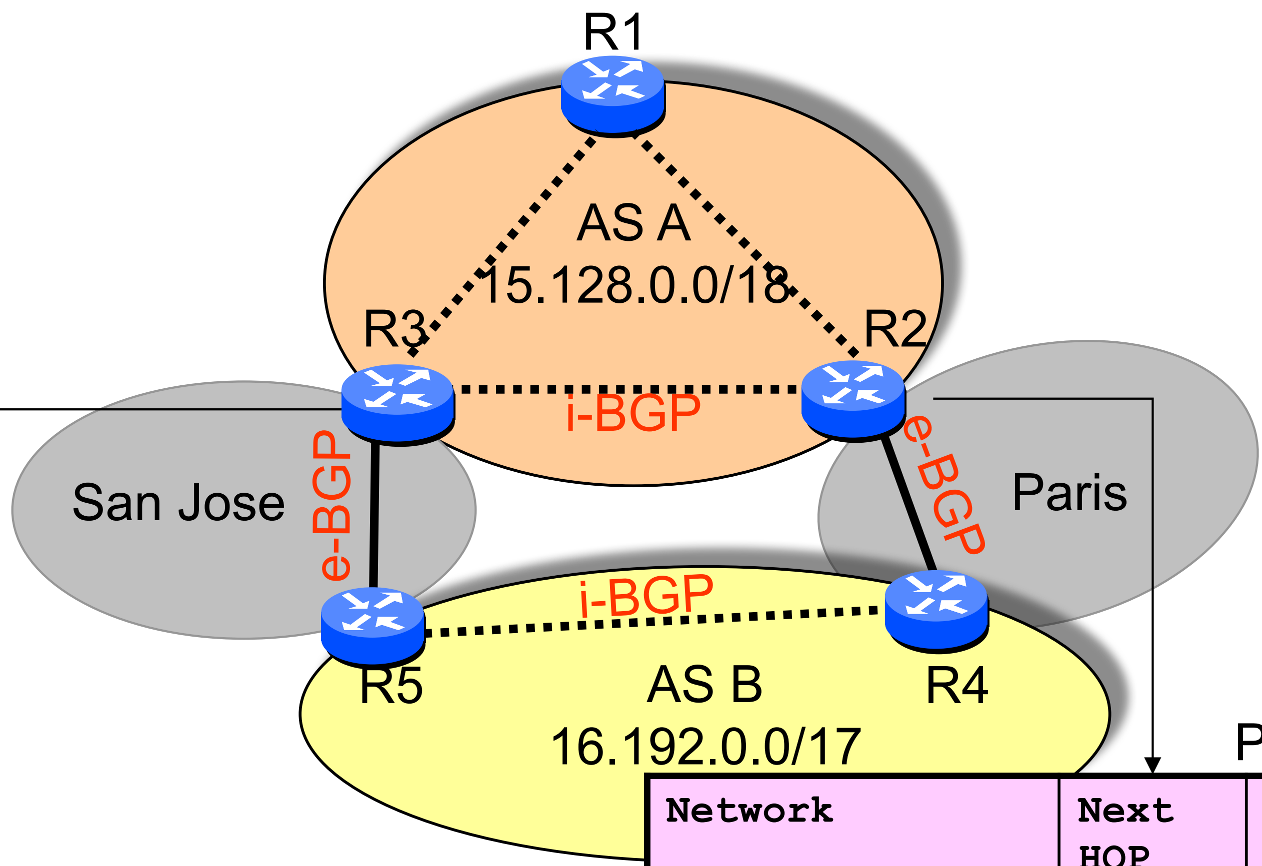
- Note:
  - Its scope may be extended to multiple ASs
  - Its use for peering agreements would need to be coordinated (see the Appendix)

BGP: Policy Routing

- Why “policy routing” ?
  - Of paramount importance in order to segregate peers w.r.t transit AS, for instance.
- BGP-4 « Policy-routing » support.
  - Through filtering and BGP information manipulations (eg. AS Path)
  - Different criteria are used for path selection thanks to the complex BGP Decision process.
    - Local preference attribution.
    - AS Hop count (from the AS Path Vector)
    - MED (Multi Exit Discriminator)
    - “Hot Potato” (eBGP versus iBGP, closest Next Hop)
    - Tie Breaking

BGP: Hot Potato Routing

Routes learned from E-BGP are preferred over routes from i-BGP



Paris Router :

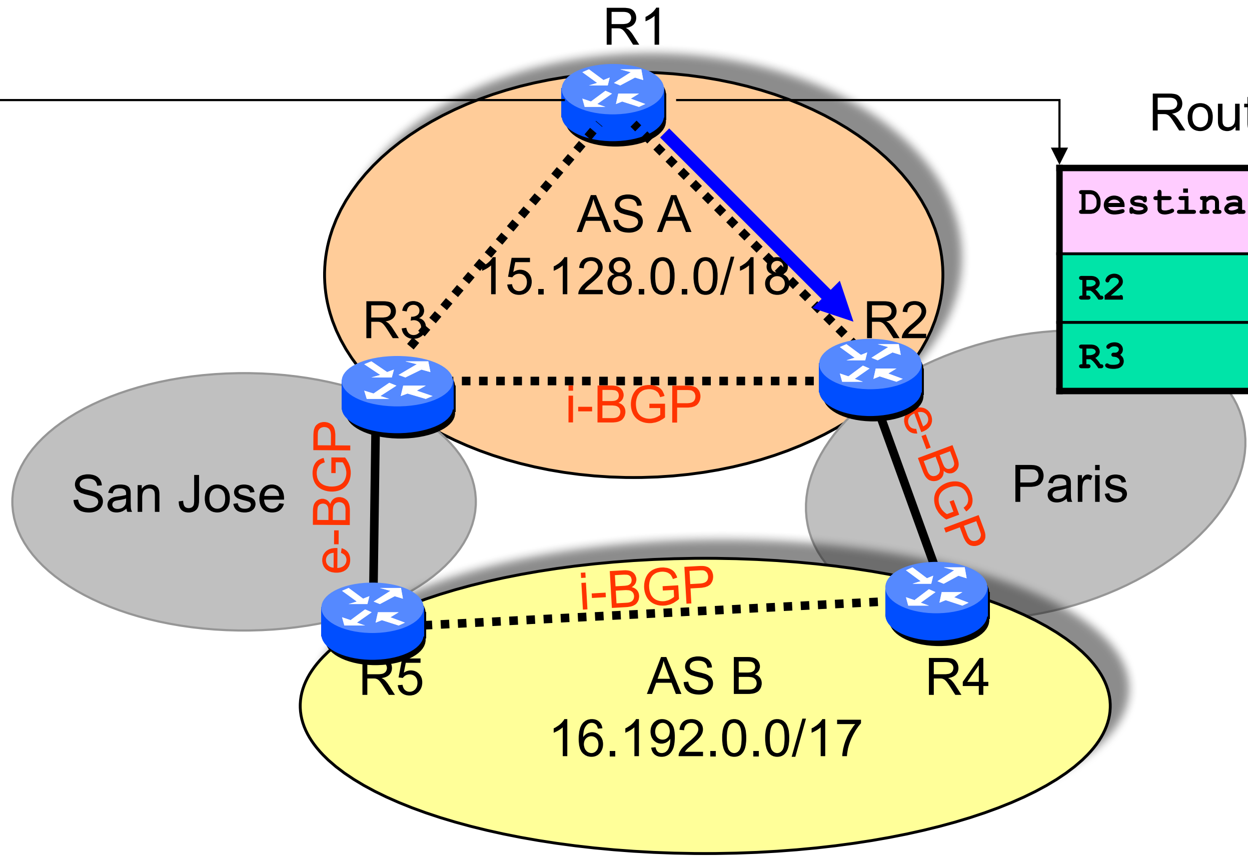
Network	Next HOP	AS Path
> 16.192.0.0/17	**R4**	**B (from eBGP)**
16.192.0.0/17	R3	B i(from iBGP)

San Jose Router :

Network	Next HOP	AS Path
16.192.0.0/17	R2	B i(from iBGP)
> 16.192.0.0/17	**R5**	**B (from eBGP)**

BGP: Hot Potato Routing (2)

IGP distance is used to choose between iBGP routes



Router R1 (IGP table):

Destination	…	Cost
R2		60
R3		100

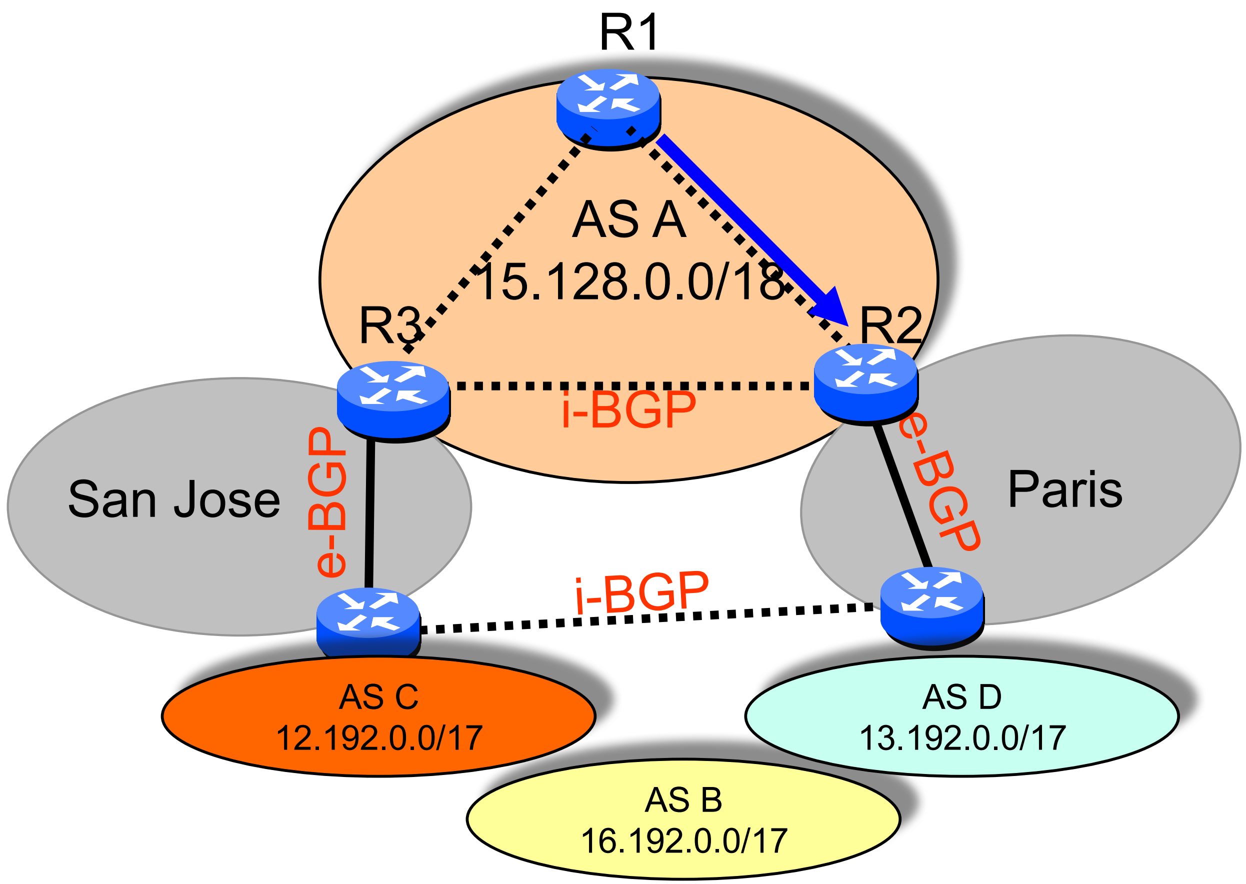
RFC 4721 (2006) spécifie de ne plus implémenter ce comportement pour des routes externes. Pas toujours implémenté

Router R1 (BGP table):

Network	Next HOP	AS Path
16.192.0.0/17	R3	B i(from iBGP) - IGP cost: 100
> 16.192.0.0/17	**R2**	**B i(from iBGP) - IGP cost: 60**

BGP: Hot Potato Routing (3)

Applied also when there are multiple downstream ASs (differently than the MED)

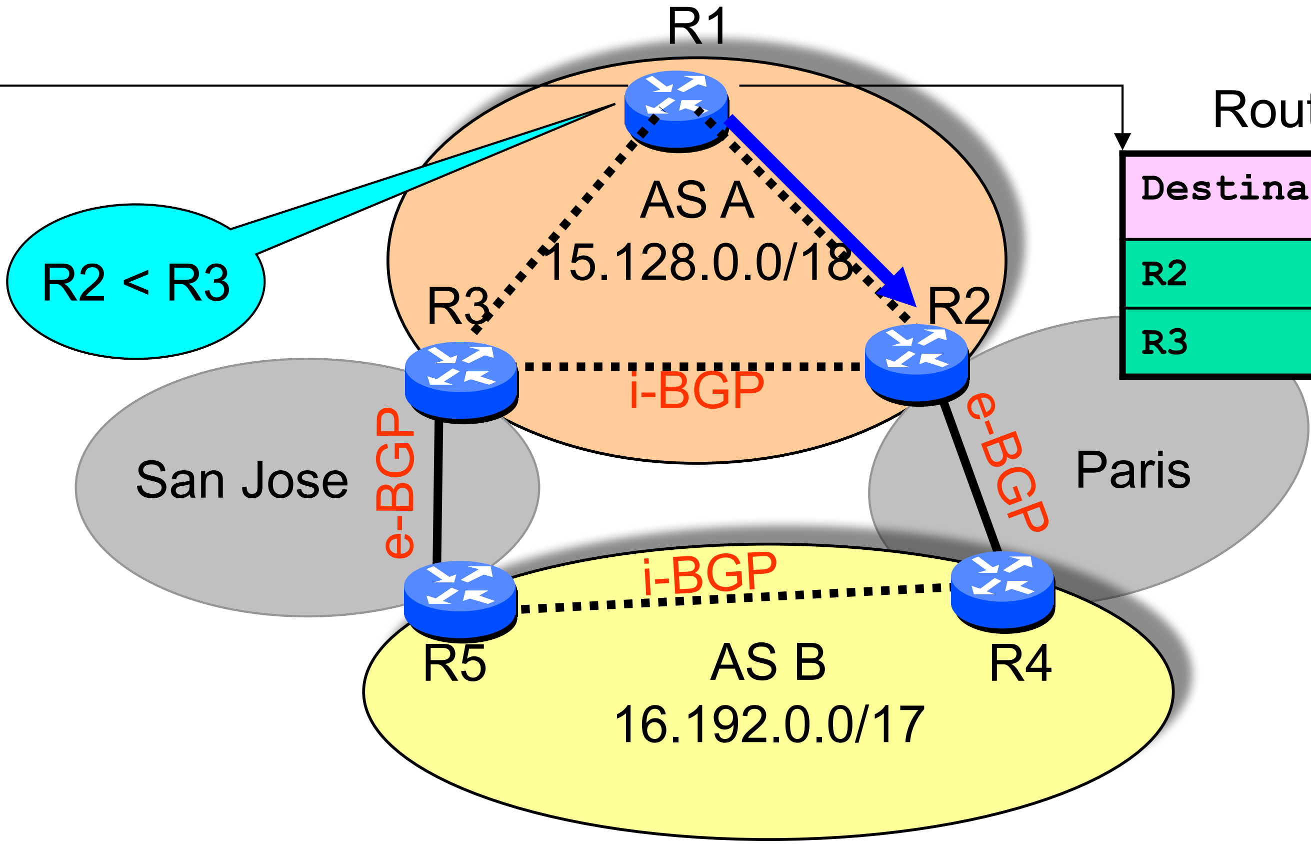


BGP: Policy Routing

- Why "policy routing" ?
  - Of paramount importance in order to segregate peers w.r.t transit AS, for instance.
- BGP-4 « Policy-routing » support.
  - Through filtering and BGP information manipulations (eg. AS Path)
  - Different criteria are used for path selection thanks to the complex BGP Decision process.
    - Local preference attribution.
    - AS Hop count (from the AS Path Vector)
    - MED (Multi Exit Discriminator)
    - "Hot Potato" (eBGP versus iBGP, closest Next Hop)
    - Tie Breaking

BGP: Tie Breaking

If nothing else is different, use next hop with lowest IP address !!!



Router R1 (IGP table):

Destination	…	Cost
R2		100
R3		100

Router R1 (BGP table):

Network	Next HOP	AS Path
16.192.0.0/17	R3	B i(from iBGP) - IGP cost: 100
> 16.192.0.0/17	**R2**	**B i(from iBGP) - IGP cost: 100**

Importance of BGP Traffic Engineering

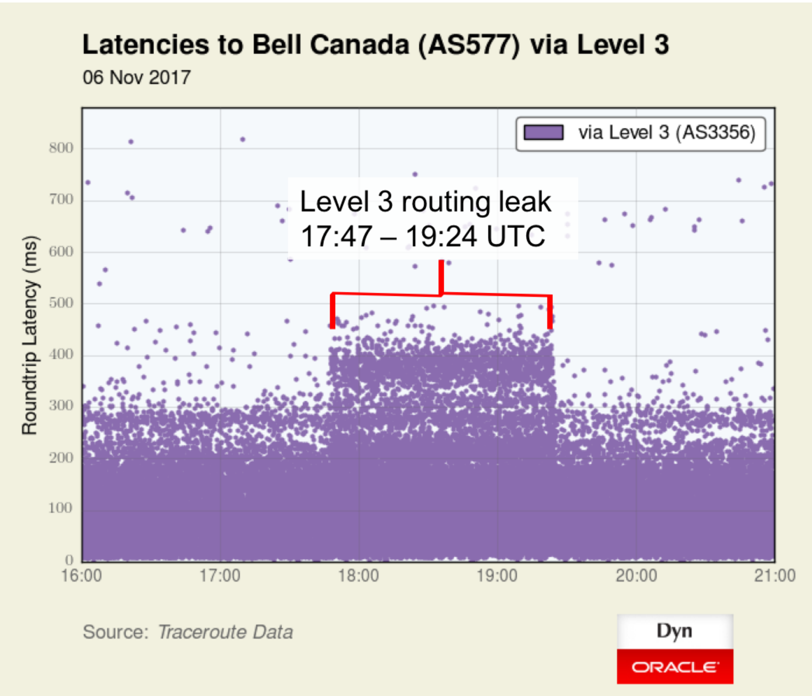
an example

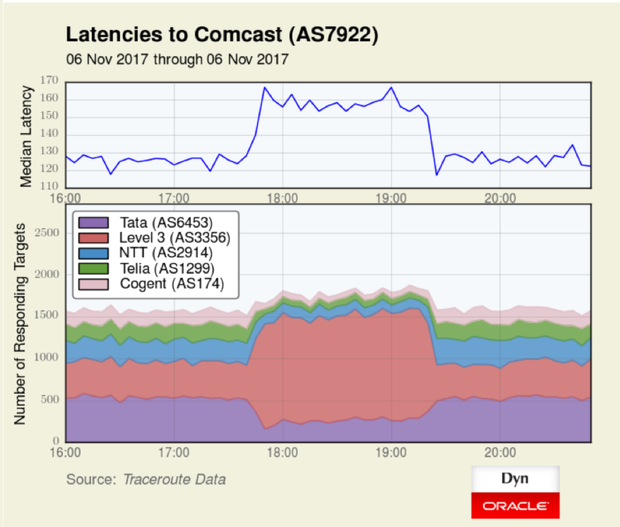
Widespread impact caused by Level 3 BGP route leak

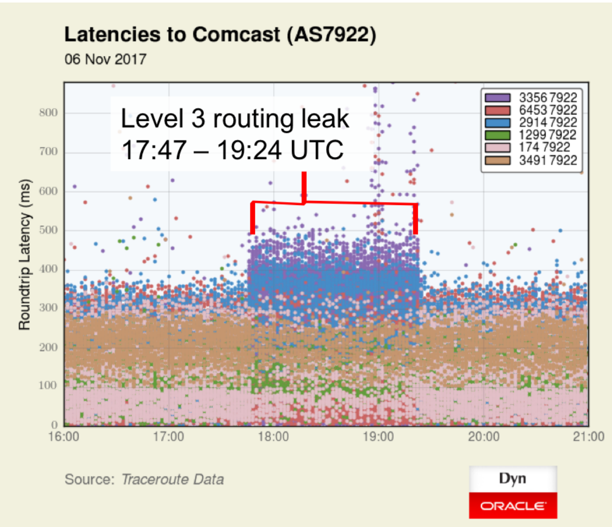
Research // Nov 7, 2017 // Doug Madory

What happened?

At 17:47:05 UTC yesterday (6 November 2017), Level 3 (AS3356) began globally announcing thousands of BGP routes that had been learned from customers and peers and that were intended to stay internal to Level 3. By doing so, internet traffic to large eyeball networks like Comcast and Bell Canada, as well as major content providers like Netflix, was mistakenly sent through Level 3's misconfigured routers. Traffic engineering is a delicate process, so sending a large amount of traffic down an unexpected path is a recipe for service degradation. Unfortunately, many of these leaked routes stayed in circulation until 19:24 UTC leading to over 90 minutes of problems on the internet.





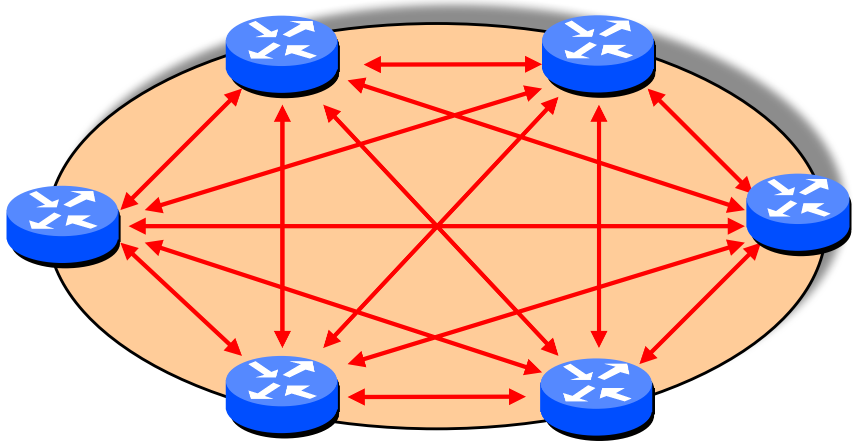


BGP SCALABILITY AND MANAGEMENT

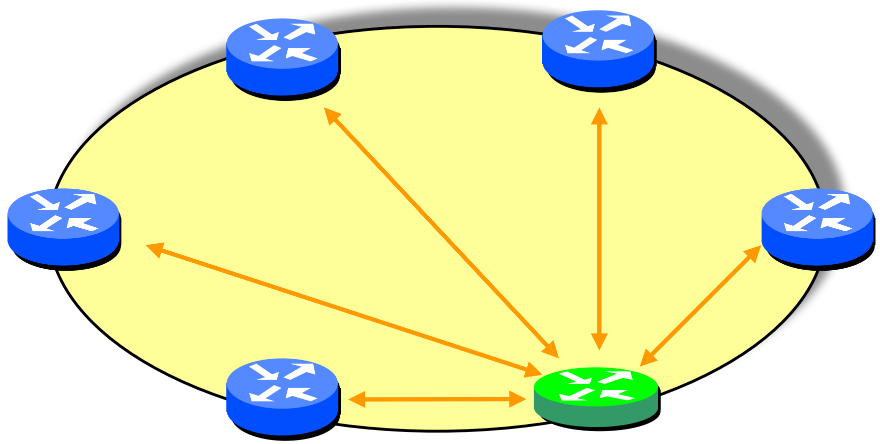
BGP Peers and Route Reflectors

i-BGP peers need to be full meshed (no intermediate BGP hops)

- Scalability and routing convergence issues…

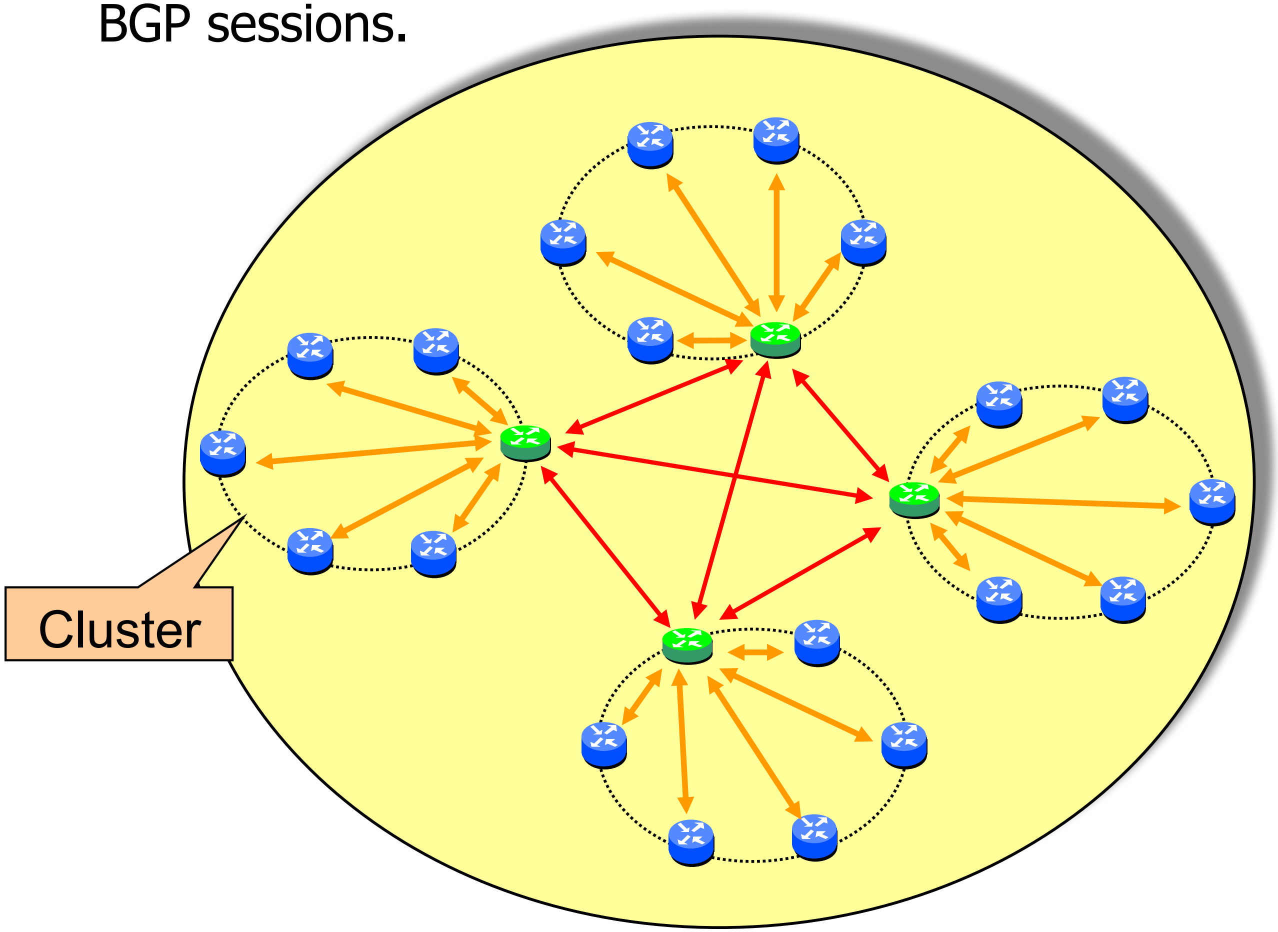


… except if a Route Reflector is configured



BGP Clusters

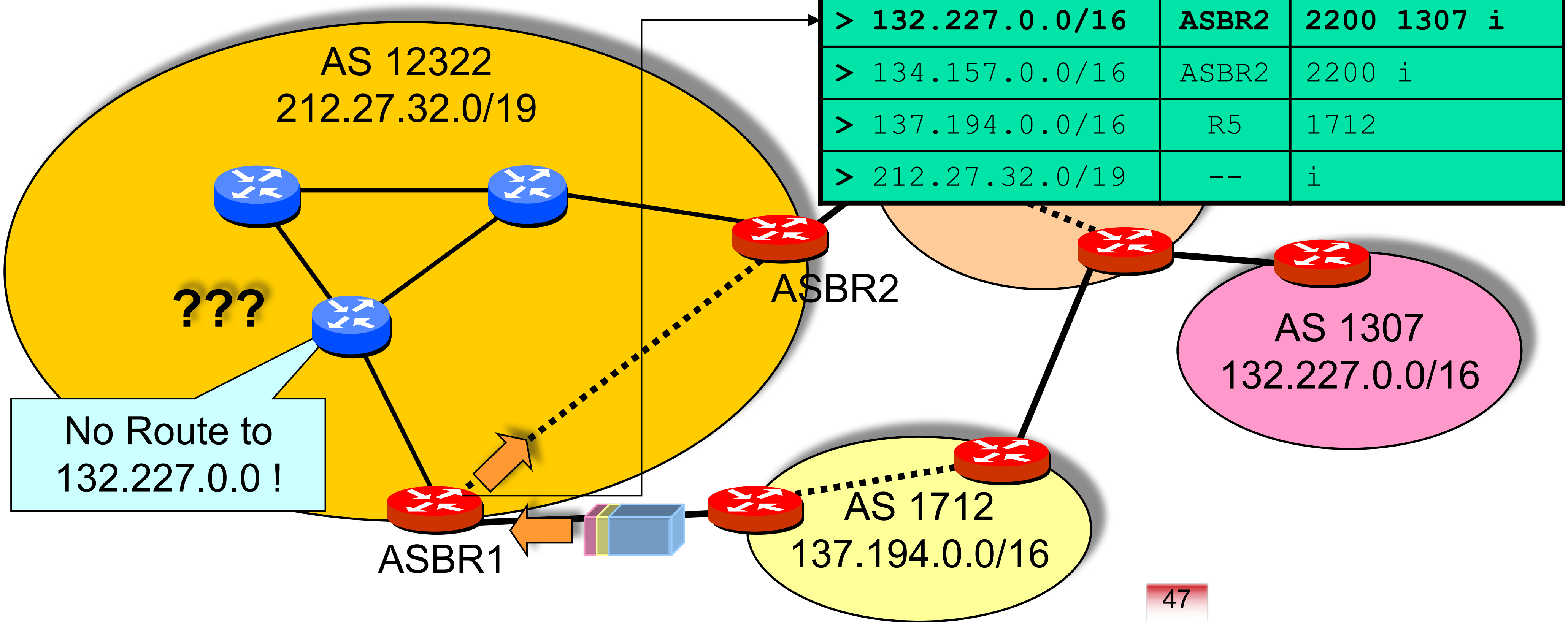
- For large networks, a set of Route Reflectors can be configured, defining “clusters”.
  - Route Reflectors are usually connected together using full meshed i-BGP sessions.



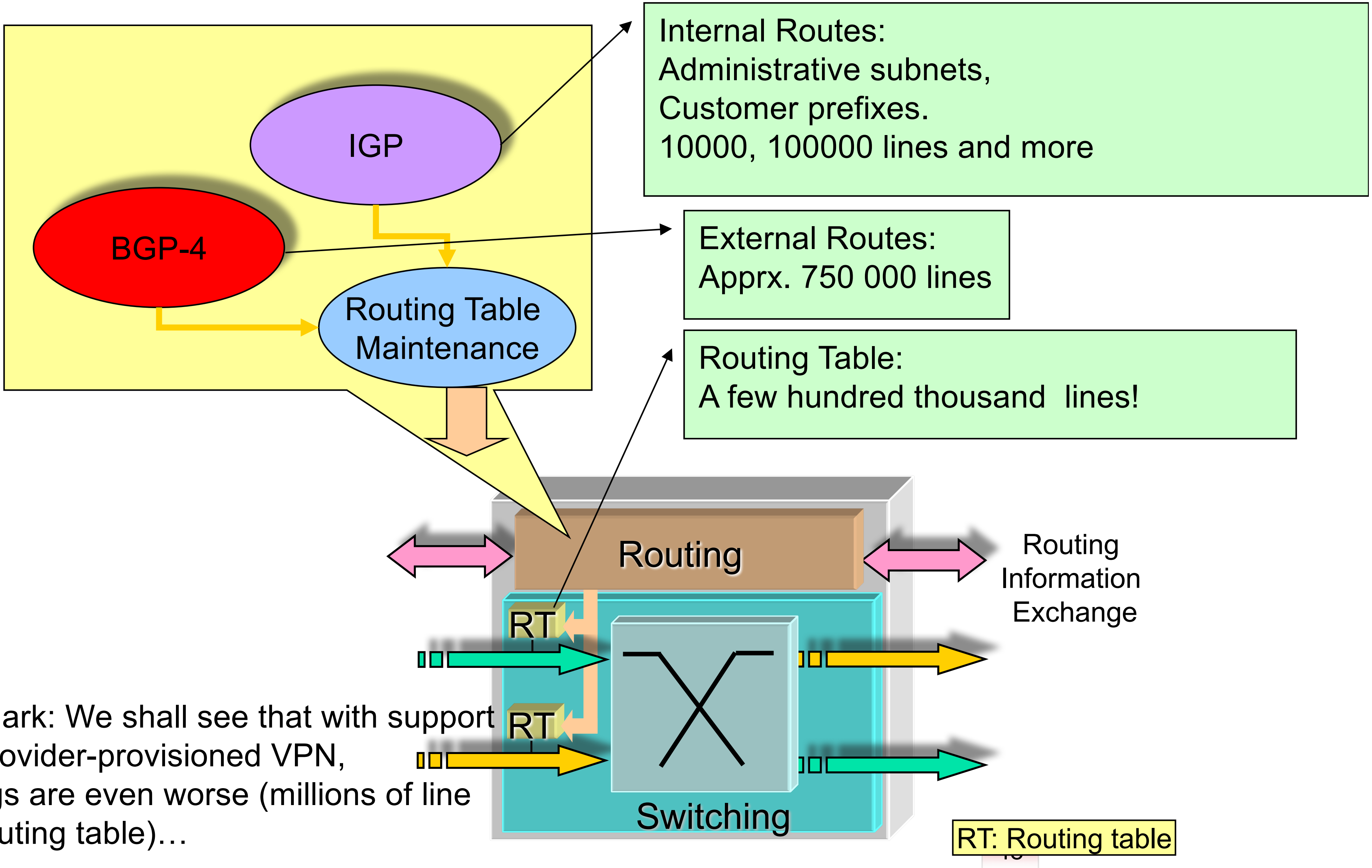
Transit Networks: Interactions between IGP and EGP

- Internal routers don't know external address prefixes - Three solutions:
  - **BGP in all routers** (using Route Reflectors, clusters etc. to avoid overhead represented by large number of i-BGP sessions)
  - **Leaking** (EGP routes injected in IGP, dangerous !!!)
  - Tunneling (GRE, **MPLS**)

Network	Next Hop	AS Path
> 132.227.0.0/16	**ASBR2**	**2200 1307 i**
> 134.157.0.0/16	ASBR2	2200 i
> 137.194.0.0/16	R5	1712
> 212.27.32.0/19	--	i



Remark: Routing Processes



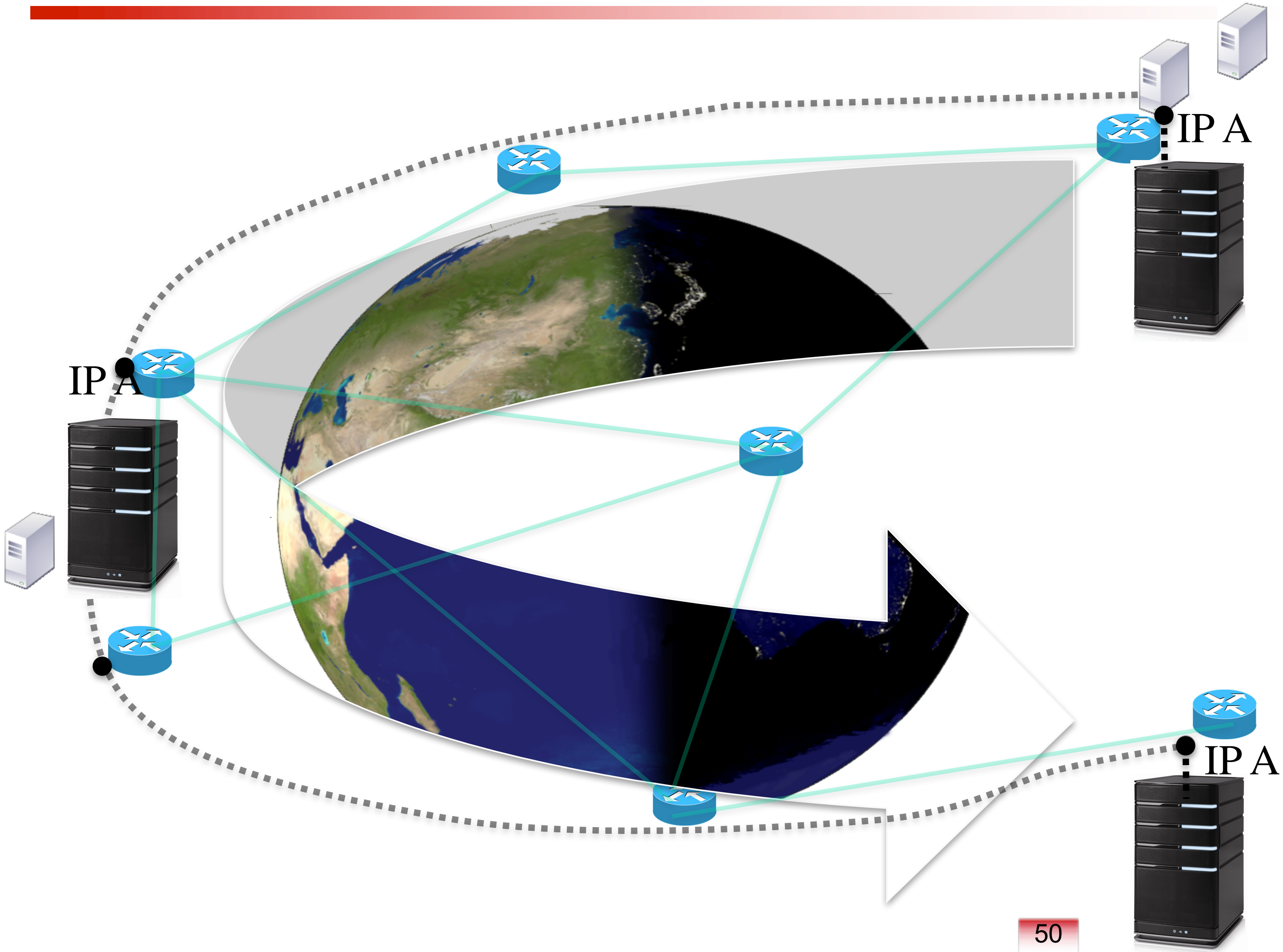
Remark: We shall see that with support of provider-provisioned VPN, things are even worse (millions of line in routing table)…

LISP

LOCATOR/IDENTIFIER SEPARATION PROTOCOL ROUTING

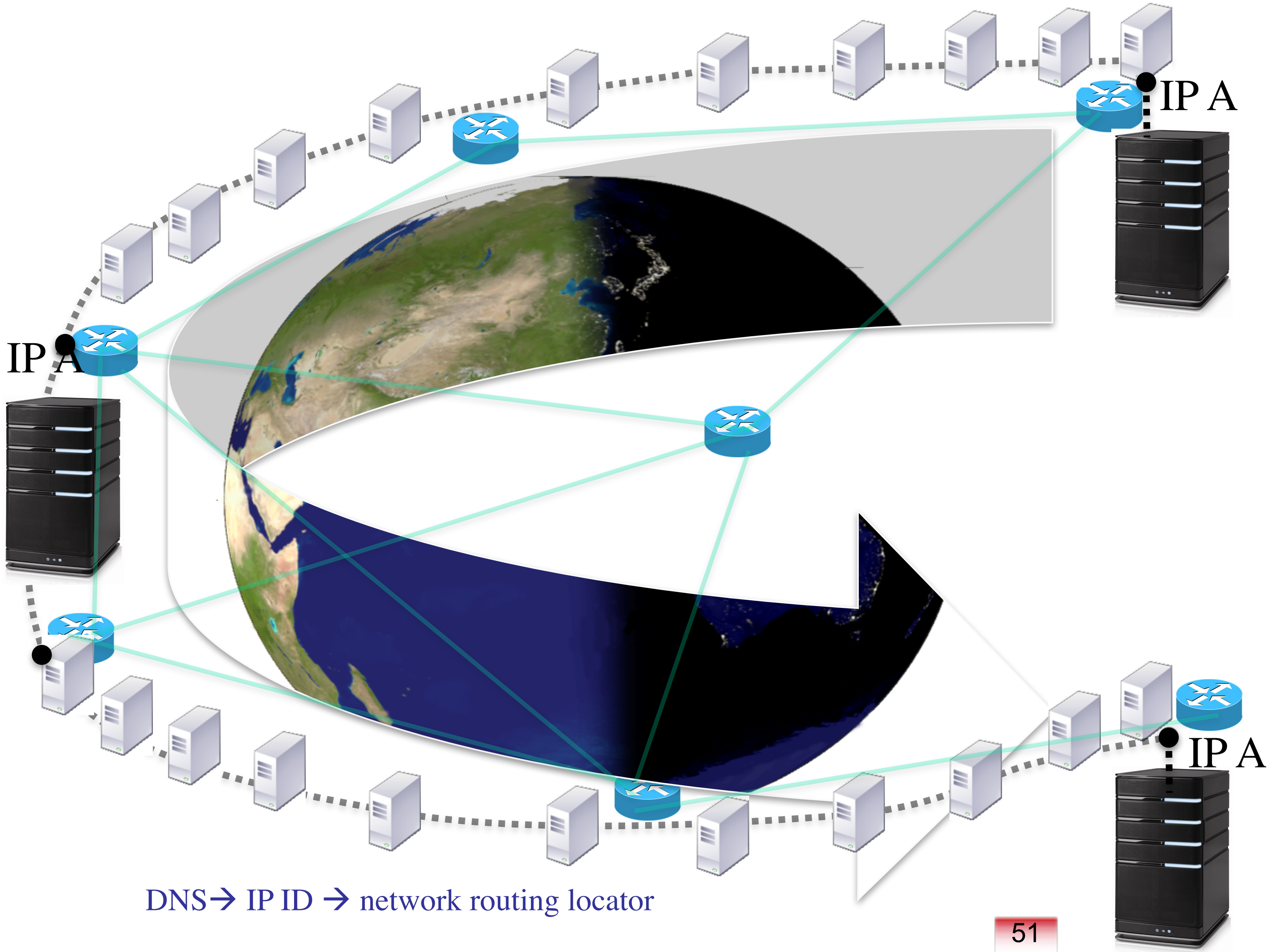
Cloud-centric Internet

comment supporter une migration de machines virtuelles à l'échelle Internet?



Cloud-centric Internet

comment supporter une migration de machines virtuelles à l'échelle Internet?



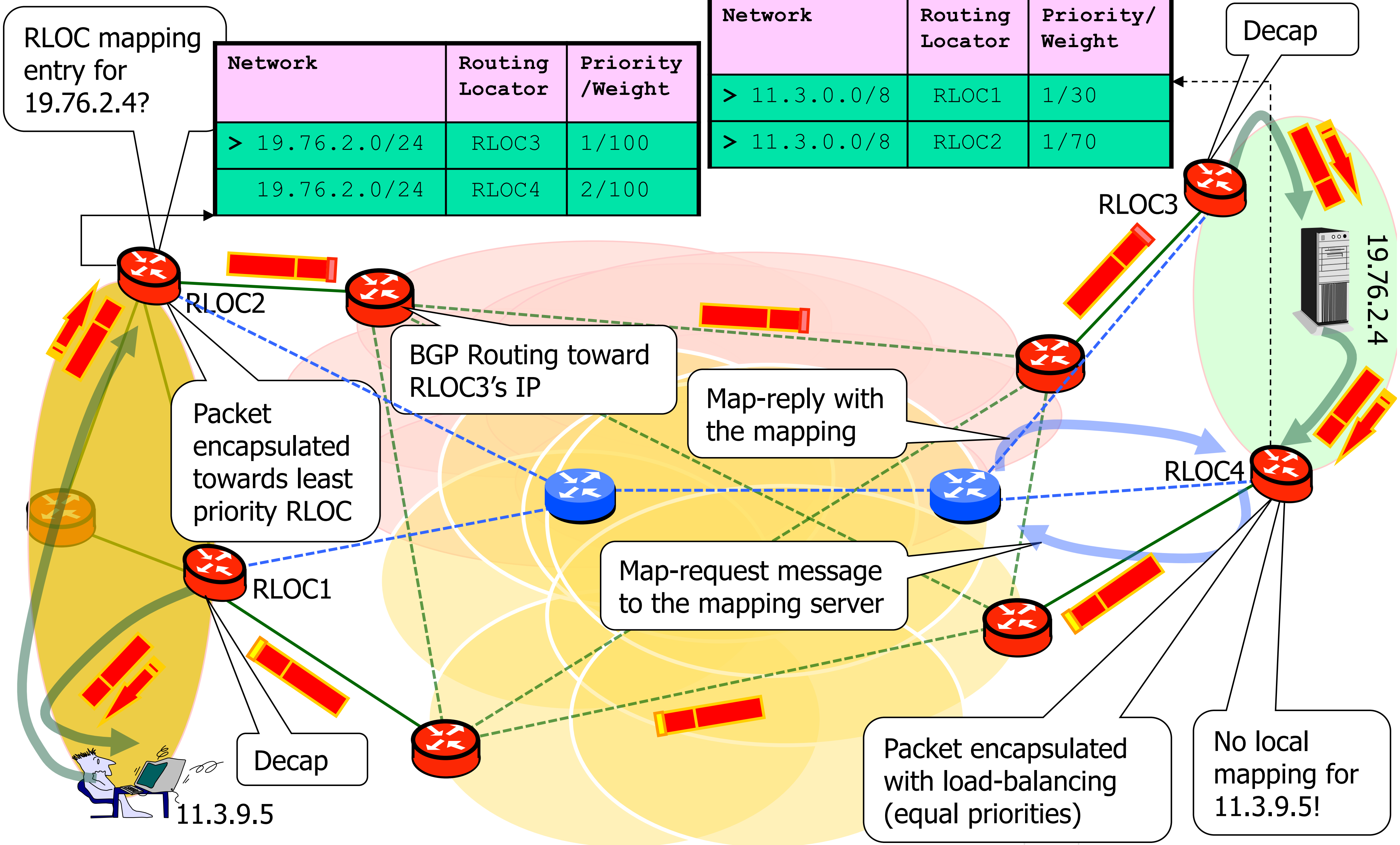
DNS→ IP ID → network routing locator

Locator/Identifier Separation Protocol (LISP)

(IP-UDP-LISP-IP data plane, BGP+LISP control plane)

Network	Routing Locator	Priority /Weight
> 19.76.2.0/24	RLOC3	1/100
19.76.2.0/24	RLOC4	2/100

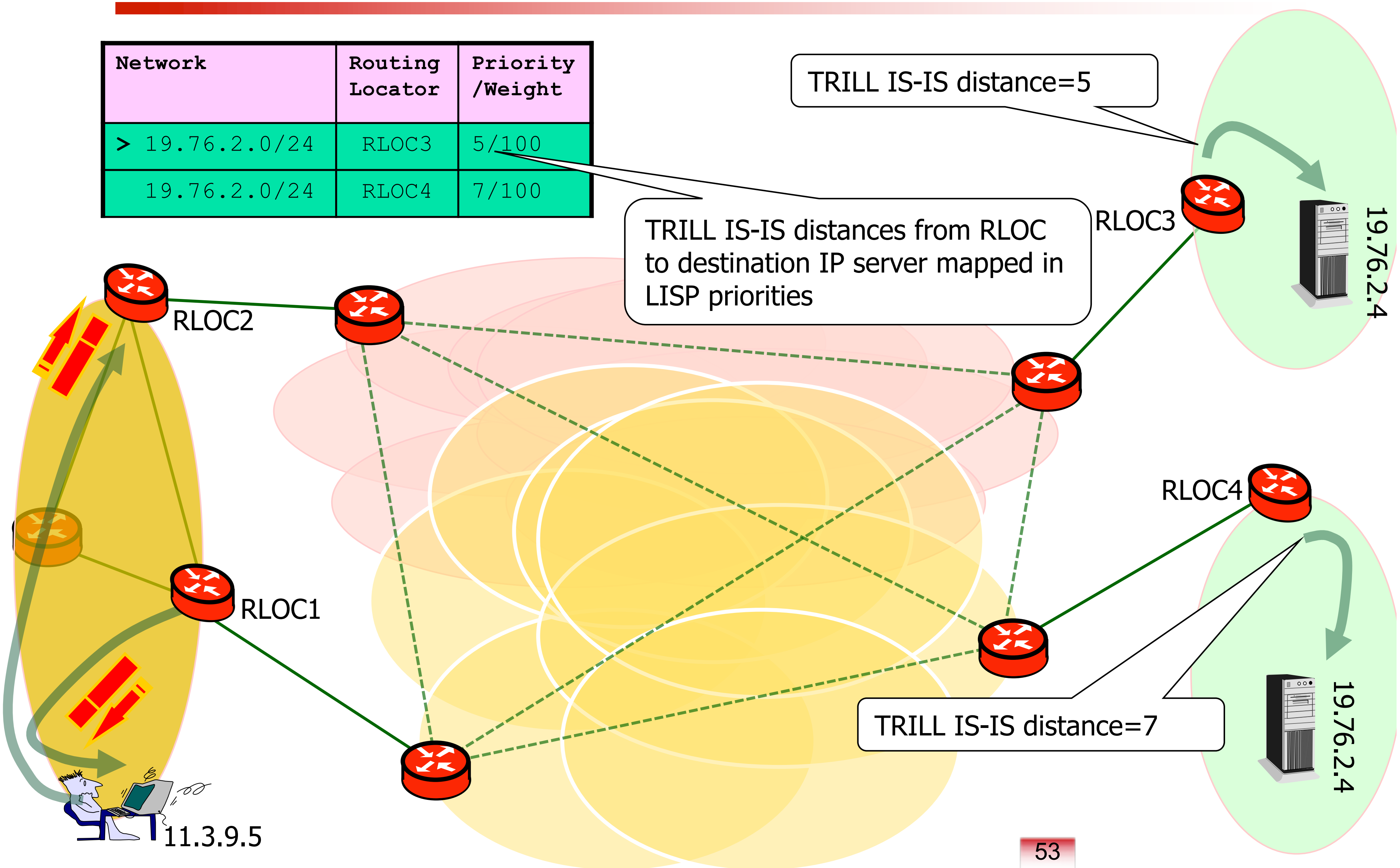
Network	Routing Locator	Priority/ Weight
> 11.3.0.0/8	RLOC1	1/30
> 11.3.0.0/8	RLOC2	1/70



Pilotage de Datacenters distribués

(IP data plane, IGP+LISP control plane)

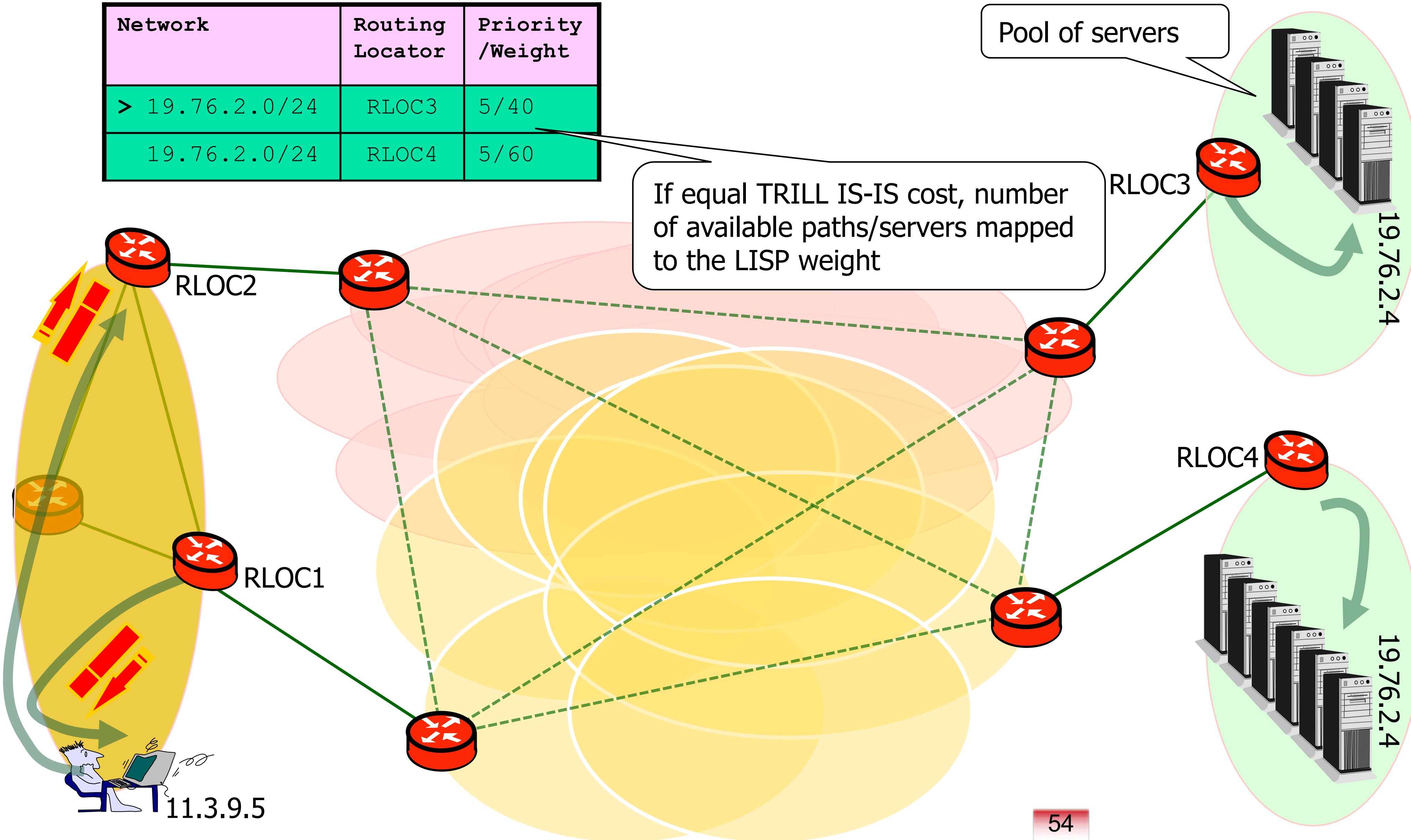
Network	Routing Locator	Priority /Weight
> 19.76.2.0/24	RLOC3	5/100
19.76.2.0/24	RLOC4	7/100



Pilotage de Datacenters distribués (2)

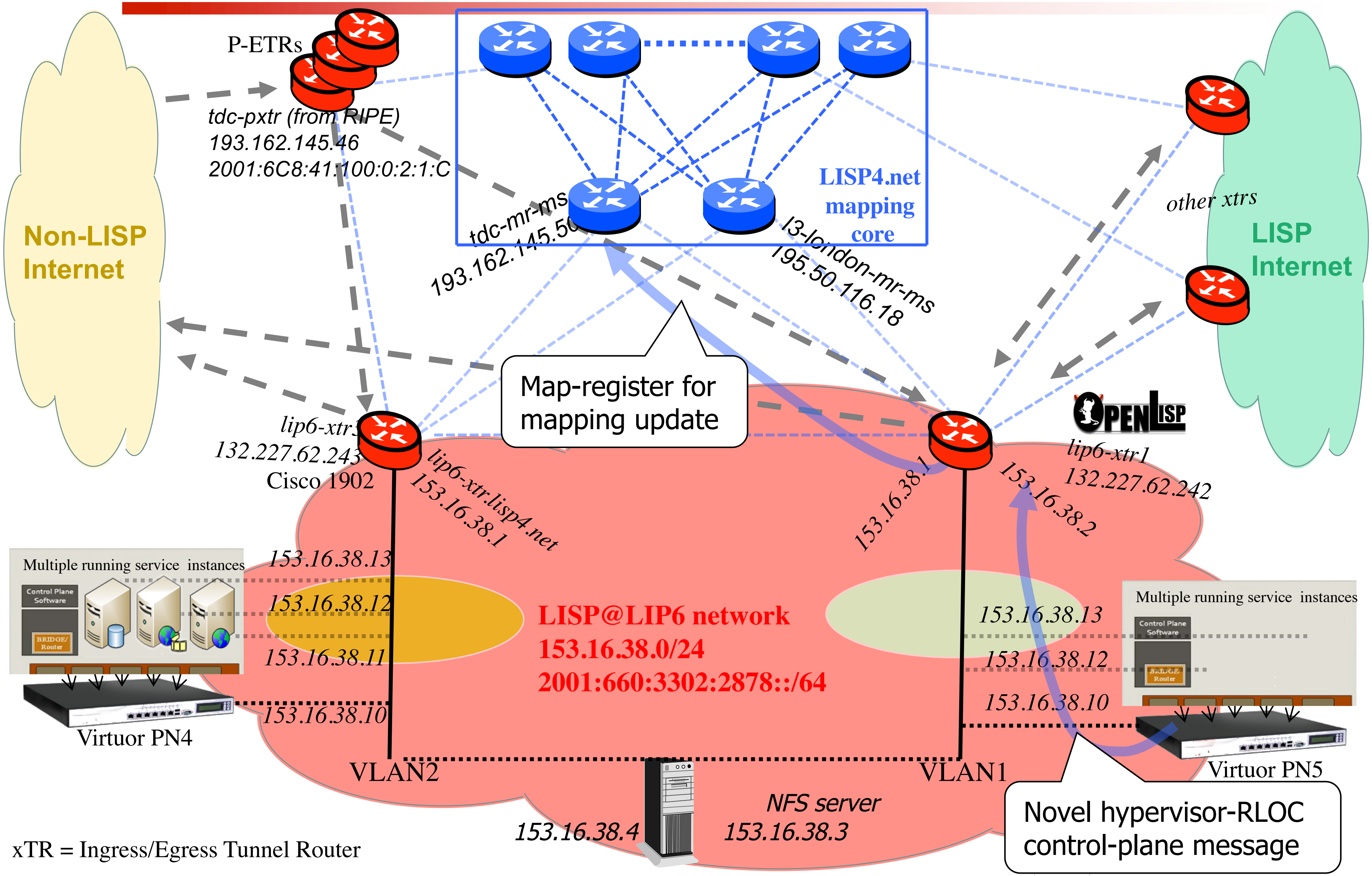
(IP data plane, IGP+LISP control plane)

Network	Routing Locator	Priority /Weight
> 19.76.2.0/24	RLOC3	5/40
19.76.2.0/24	RLOC4	5/60



Interconnexion au testbed LISP4.Net et migration de MVs

LISP@LIP6: http://www.lisp.ipv6.lip6.fr (behind LISP!)

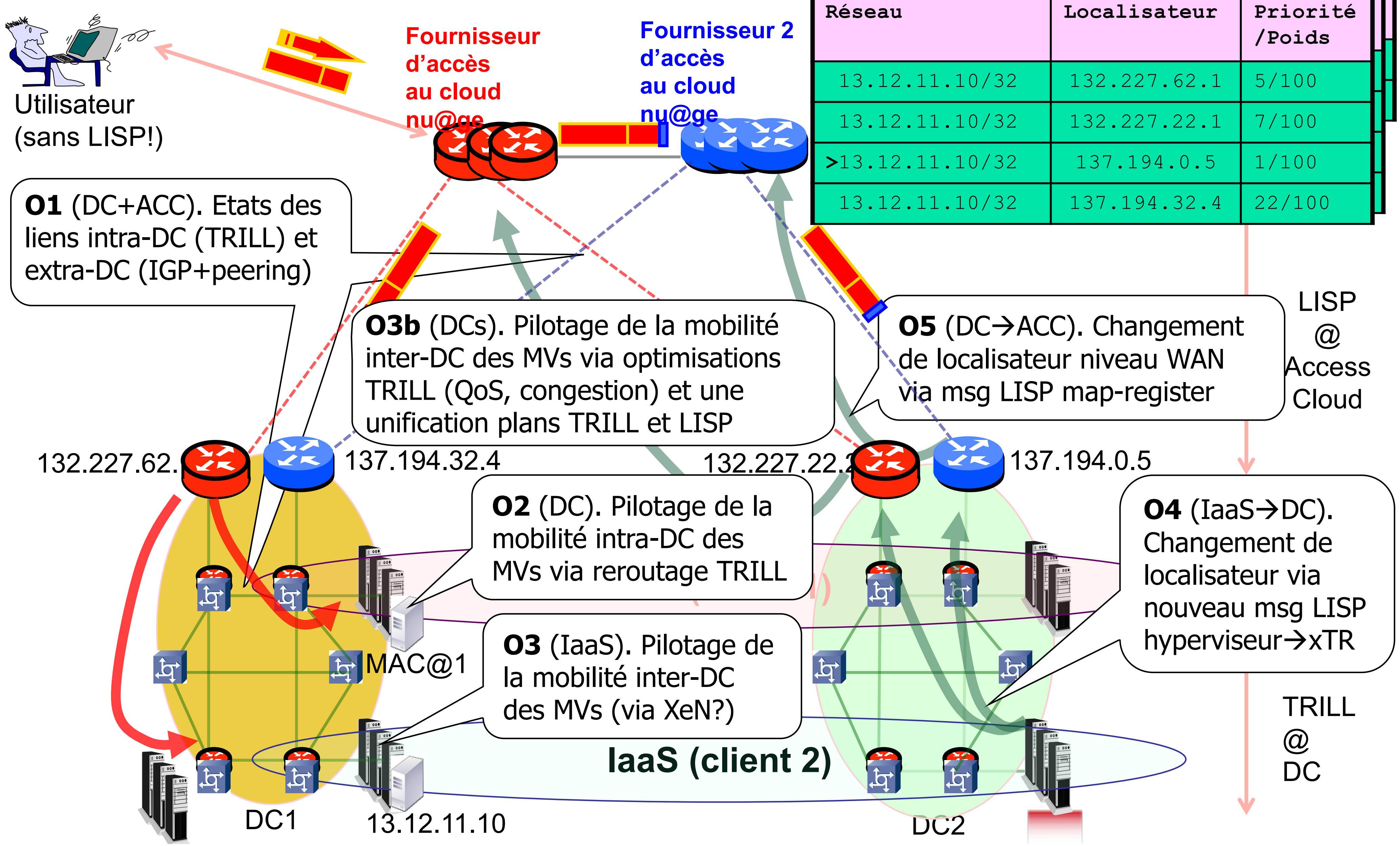


xTR = Ingress/Egress Tunnel Router

Vers un hyperviseur WAN

(plan de données IP, plan de contrôle TRILL+LISP)

Réseau	Localisateur	Priorité /Poids
13.12.11.10/32	132.227.62.1	5/100
13.12.11.10/32	132.227.22.1	7/100
>13.12.11.10/32	137.194.0.5	1/100
13.12.11.10/32	137.194.32.4	22/100



S. Secci et al., “ClubMED: Coordinated Multiple Exit Discriminator Strategies for Peering Carriers”, in Proc. of NGI 2009

APPENDIX: BGP MED-BASED COORDINATION

A 2-link peering game example

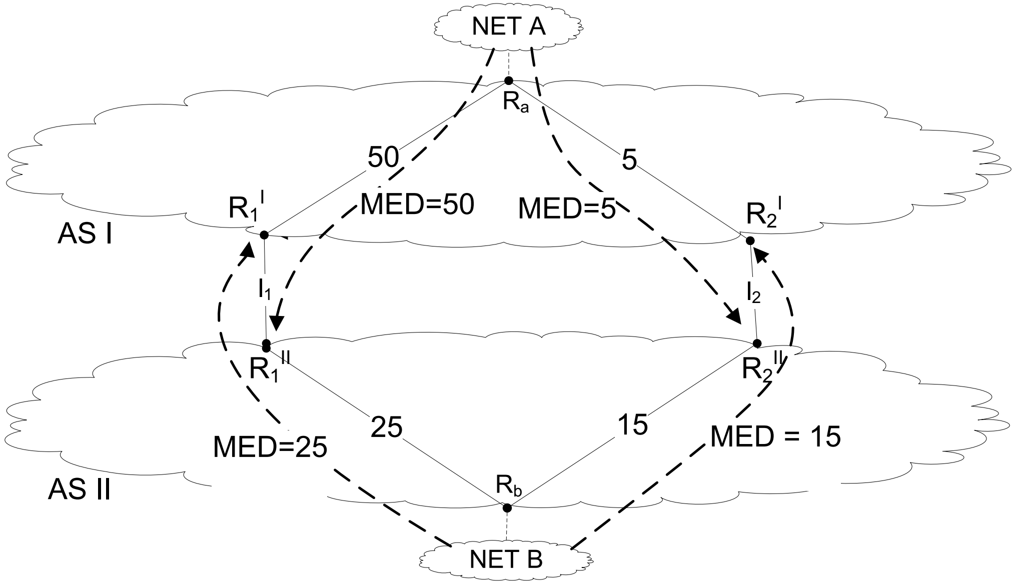


TABLE I
A DUMMY GAME.

I \ II	l_1	l_2
l_1	**(50,25)**	**(5,25)**
l_2	**(50,15)**	**(5,15)**

TABLE II
A CLUBMED GAME.

I \ II	l_1	l_2
l_1	(100,50)	(55,40)
l_2	(55,40)	**(10,30)**

- AS I and AS II exchange their internal routing cost via the MED
  - MED-icated BGP announcements for NET A and NET B (resp.)
- The strategies available for each peer are the possible egress links
  - l_1 and l_2
- Table I: form including the MEDs affecting the peer routing decision
  - → dummy game (unilateral choices l_1,l_2 are equivalent): 4 Nash equilibria
- Table II: sum its own IGP routing costs
  - → **ClubMED (Coordinated MED)** game: 1 Nash equilibrium (hot potato on both sides)

The ClubMED game: proven properties

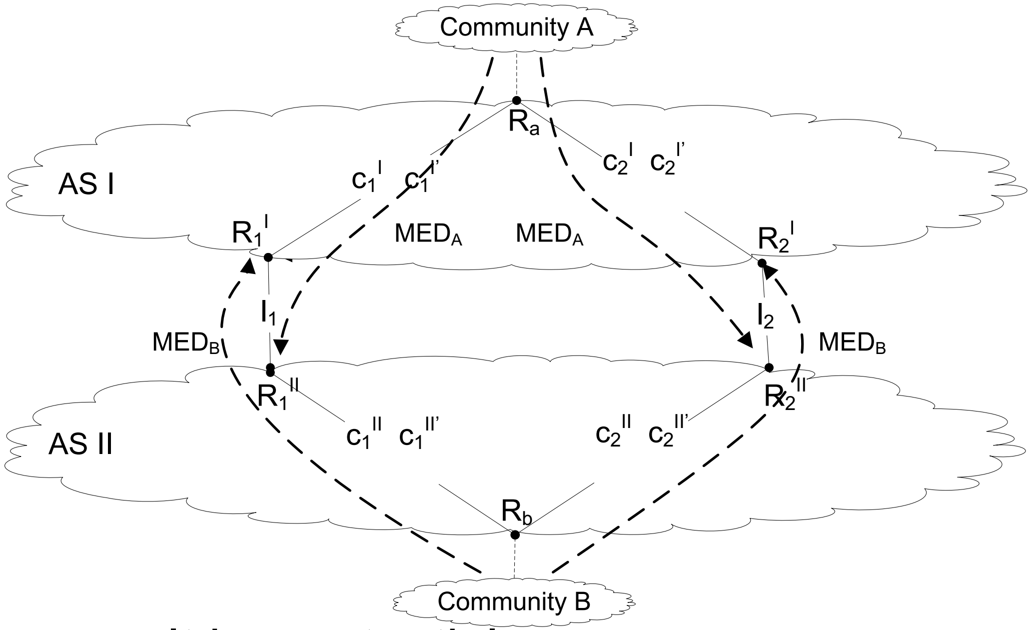


TABLE I
2-LINK CLUBMED GAME, SUM OF TWO GAMES WITH POTENTIAL.

I \ II	l_1	l_2
l_1	(c_1^I, c_1^{II})	(c_1^I, c_2^{II})
l_2	(c_2^I, c_1^{II})	(c_2^I, c_2^{II})

+

I \ II	l_1	l_2
l_1	$(c_1^{I\prime}, c_1^{II\prime})$	$(c_2^{I\prime}, c_1^{II\prime})$
l_2	$(c_1^{I\prime}, c_2^{II\prime})$	$(c_2^{I\prime}, c_2^{II\prime})$

$$\begin{pmatrix} 0 & c_1^{II} - c_2^{II} \\ c_1^I - c_2^I & c_1^{II} - c_2^{II} + c_1^I - c_2^I \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

It is a potential game

- The incentive to change expressed in one global potential function;
- The cost difference by an individual strategy move Is equal to the potential difference

Nash equilibrium ←→ Potential minimum → Low complexity for routers

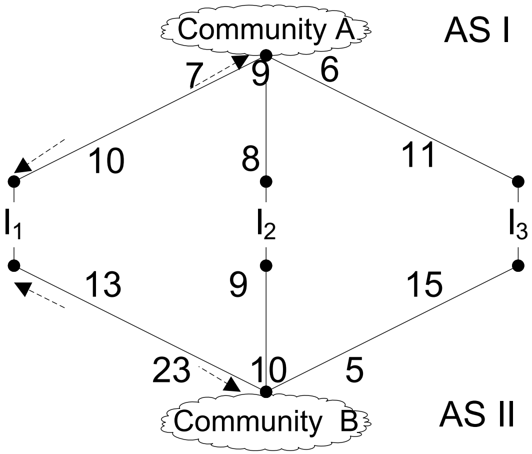
- And a Nash equilibrium always exists

If multiple equilibria, there are equal egress costs at, at least, one side

A ClubMED Nash equilibrium is not necessarily a Pareto-efficient profile

- G_d guides the Pareto-efficiency, G_s guides the equilibrium

3-link ClubMED game examples



I\II	l_1	l_2	l_3
l_1	(17,36)	(19,32)	(16,38)
l_2	(15,23)	(**17,19**)	(14,25)
l_3	(18,18)	(20,14)	(17,20)

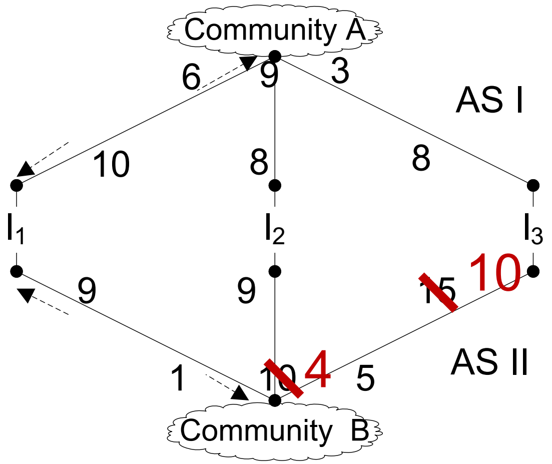
The Nash equlibrium is unique and Pareto-efficient

Pareto efficiency:

•A strategy profile s is **Pareto-superior** to another strategy profile s' if a player's cost can be decreased from s to s' without increasing the other player's cost. And s' is Pareto-inferior to s.

•A strategy profile is **Pareto-efficient** if it is not Pareto-inferior to any other strategy profile.

•The set of Pareto-efficient profiles is the Pareto-frontier of the game.



I\II	l_1	l_2	l_3
l_1	(16,10)	(19,10)	(13,~~18~~) 13
l_2	(**14,~~19~~**) 13	(**17,~~19~~**) 13	(11,~~25~~) 14
l_3	(***14,14***)	(**17,14**)	(11,~~20~~) 15

The Pareto-superior Nash equilibrium is not Pareto-efficient any longer!

ClubMED-based coordination strategies

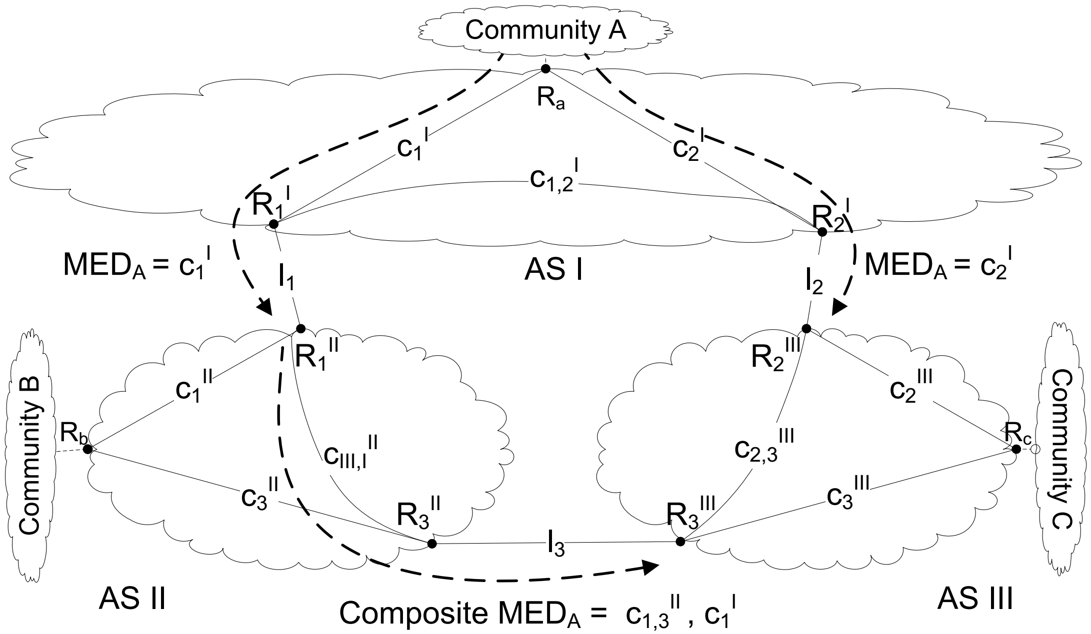
1. Implicit coordination *(one-shot)*
   a) Choose the Nash equilibrium if it is unique; if many, balance the load on the equilibria. →*Nash Equilibrium Multi Path (NEMP) policy*
   b) Choose the Pareto-superior Nash equilibrium if it is unique; NEMP on the Pareto-superior equilibria, if many (equal).

2. Explicit spot agreement: *(binding agreements)*
   Choose the Pareto-efficient strategy profiles; if many (equal), balance the load on them

3. Repeated coordination: *(repeated)*
   - After shrinking the Nash set w.r.t. the Pareto-efficiency, the ASs might agree to make both a further step toward another choice;
   - The loss that one surely has moving is compensated by the improvement upon the other AS. I.e if, e.g. for AS I moving from I_{i_0} toward I_i:

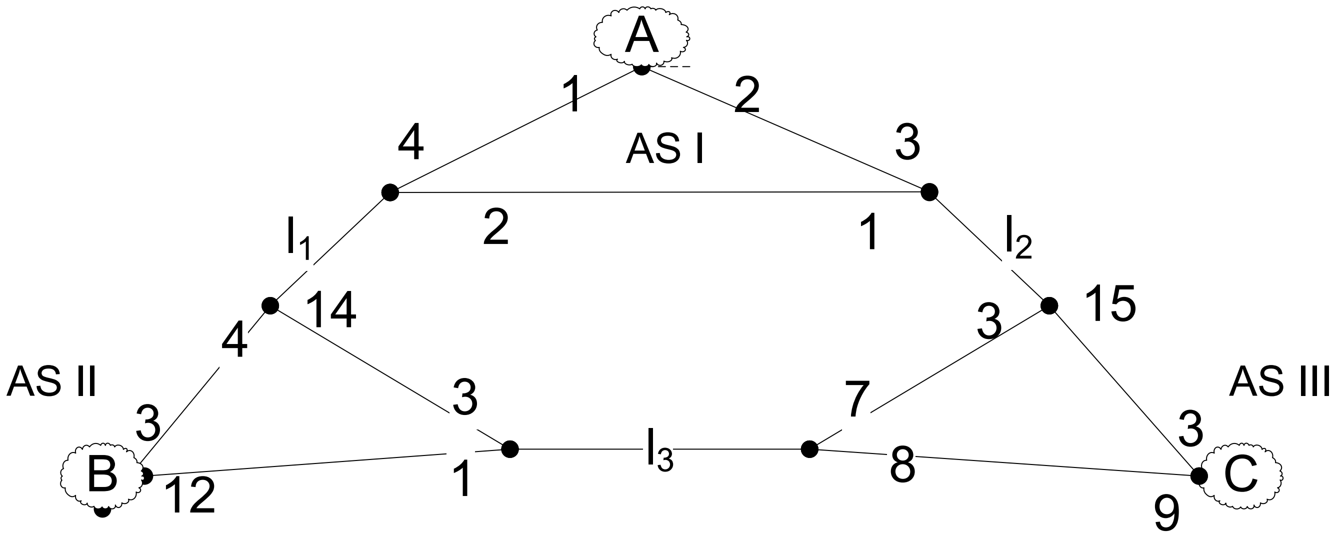
$$\psi_d(I_i) - \psi_d(I_{i_0}) + \varphi_d(I_i) - \varphi_d(I_{i_0}) < 0$$

The extended peering game



- Each peer sees the extended peering frontier as a unique frontier such as for the classical peering
  - Routing decision: where to route the egress aggregate flow from its community toward the other communities of the other peers
  - The receiving peer deaggregates the flow: one toward its destination, one transiting toward another peer folllowing BGP
- The dummy game is characterized by ingress costs and transit costs

An extended peering game example



III

III	I \ II	l_1	l_3
l_2	l_1	(12,13,27)	(14,10,36)
	l_2	(11,19,28)	(13,16,37)
l_3	l_1	(7,36,20)	(9,33,29)
	l_2	(6,42,21)	**(8,39,30)**

III	I \ II	l_1	l_3
l_2	l_1	(4,4,15)	(4,1,15)
	l_2	(3,4,15)	(3,1,15)
l_3	l_1	(4,4,8)	(4,1,8)
	l_2	(3,4,8)	(3,1,8)

+

For: $x_{III} = l_2, \quad P = \begin{pmatrix} 4 & 1 \\ 3 & 0 \end{pmatrix} \quad x_I$

Fast IGP Convergence, Global Internet Stability, Router Redundancy

APPENDIX: IMPROVING ROUTING AVAILABILITY

IGP Convergence

- Facts:
  - Standard Link-state IGP protocols converge within seconds (1s in average in a single domain, 1 min in a complex multi-area network).
  - Broadcast Link State advertisement method can be pretty inefficient and load routers uselessly.
  - Race conditions met due to loss of hello packets (e.g. ATT world-wide network collapse in 2001…)

- On-going Developments
  - (done) sub-second IGP convergence (millisecond IGP convergence):
    - Decrease of Hello refresh period,
    - code optimization,
    - implementation of fast routing re-computation (iSPF), …
  - Developments for better IGP stability:
    - Priority given to hello packet handling
    - Flow control, Graceful restart
  - Research on "Fast IP Re-Routing" (50ms convergence)

Global Internet Stability

- Facts:
  - Routing loops or incorrect routes detected during supervision campaigns. Main reasons: Incongruent policies between ASes, misconfiguration of BGP4.
  - A problem in one router… is spread all over the Internet !
  - Threat !!! More and more ASes… More and more policies… And BGP-4 is now manipulated by customers (multihoming etc.) !

- On-going Developments
  - Better control on Policies (RIPE Databases, etc…): but has usually little effect has these databases are not used for router configuration (manual configuration)
  - Understanding policy routing, incongruent policy effects (models, simulations, etc.).
  - Develop routing configuration debugging tools, routing table consistency analysis tools, etc. (mostly proprietary scripts form carriers and ISPs).
  - No Solution available to date…
  - … but at least more and more supervision tools available and implemented.

Increasing Router Availability

- Possible solution: Router Redundancy
  - IETF solution: VRRP (Virtual Router Redundancy Protocol)
  - Some proprietary solutions: HSPR (Host Standby Router Protocol) from Cisco or IPSTB (IP STandBy) from DEC, …
- Goal: Passive redundancy
  - Several routers are seen as a single virtual router. One a active (elected master) and others are in stand-by mode.
  - Works on LAN (broadcast networks).