Ongoing stories in Inter-domain routing

(Some of them)
Pierre.Francois@imdea.org

Recommendation for a smooth afternoon

- Talks given to ISPs, router vendors, and one CDN
  - No rocket science
  - maybe not in your field
  - **Interrupt** me for definition/clarification questions
- I don't need to get to the end of my talk (2h)

BGP

- Path vector protocol (Autonomous System Paths) Paths are stuffed with attributes on which routing decisions are based
- Policy driven (1st rule is to not offer free lunch)
- 400k+ independent routing decisions to keep up with (+ some v6 prefixes...)
- Flexible and hence complex to operate (BGP can fail)
- eBGP and iBGP

Agenda

- iBGP, eBGP, (mis-)behaving on the IP transit market with BGP
- Giving ISPs a convergence vs. scaling tradeoff for **i**BGP
  - We got good at scaling by making convergence worse
  - Add-Paths, Why? How?
  - I should have KISS'ed that project
- **e**BGP Policy violations in the data-plane
  - No free lunch on the Internet, really?
- (What you should not do about your transit bill)

BGP Add-Paths

Hundreds of proposals hiding behind one...

Motivation for Add-paths

- Initial "motivation" was MED oscillation avoidance
  - A major change to BGP to solve a problem which has gone public... once.
- Emergence of new IDR requirements a few years ago
  - Fast recovery upon peering link / ASBR failure
  - Load balancing among multiple primary BGP NHs
  - Hitless planned maintenance
  - "Optimal" hot-potato routing
  - (Churn reduction / convergence concealment)

Add-Paths at the IETF

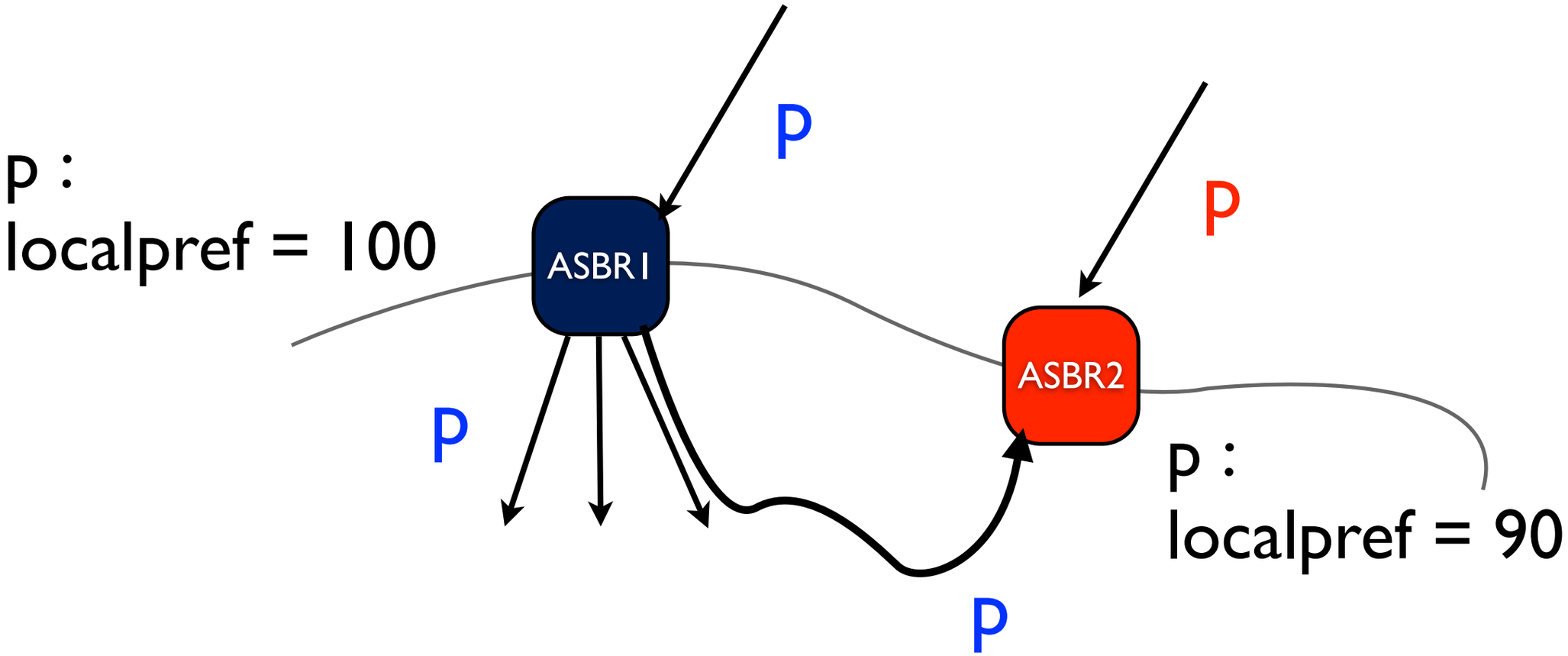
- draft-ietf-idr-add-paths
  How to announce multiple paths for the same destination

- draft-ietf-idr-add-paths-guidelines
  Why only a small subset of proposals will be supported

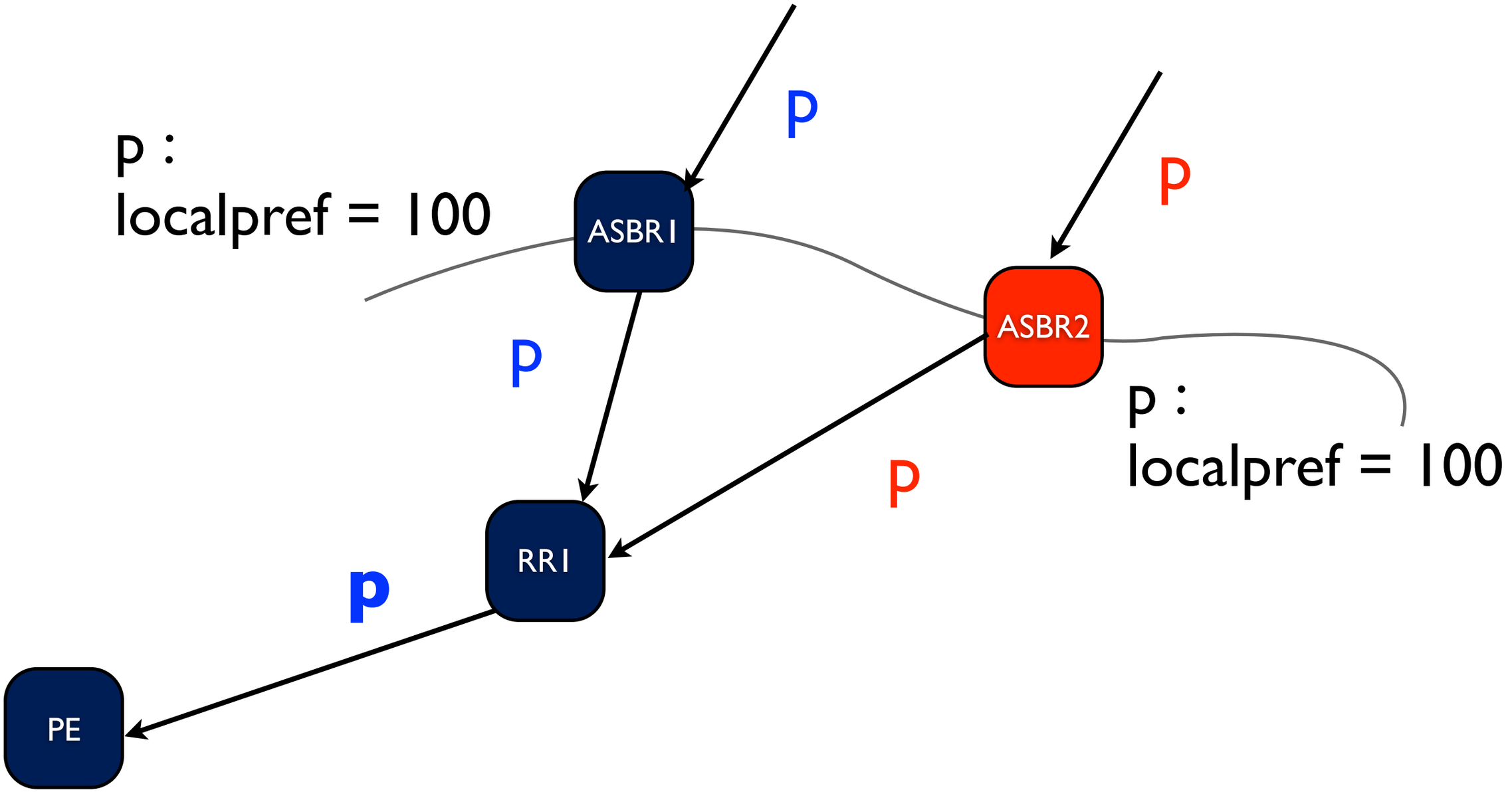
iBGP Path hiding

- Lack of path diversity in iBGP deployments
  - Policies
  - Route Reflection

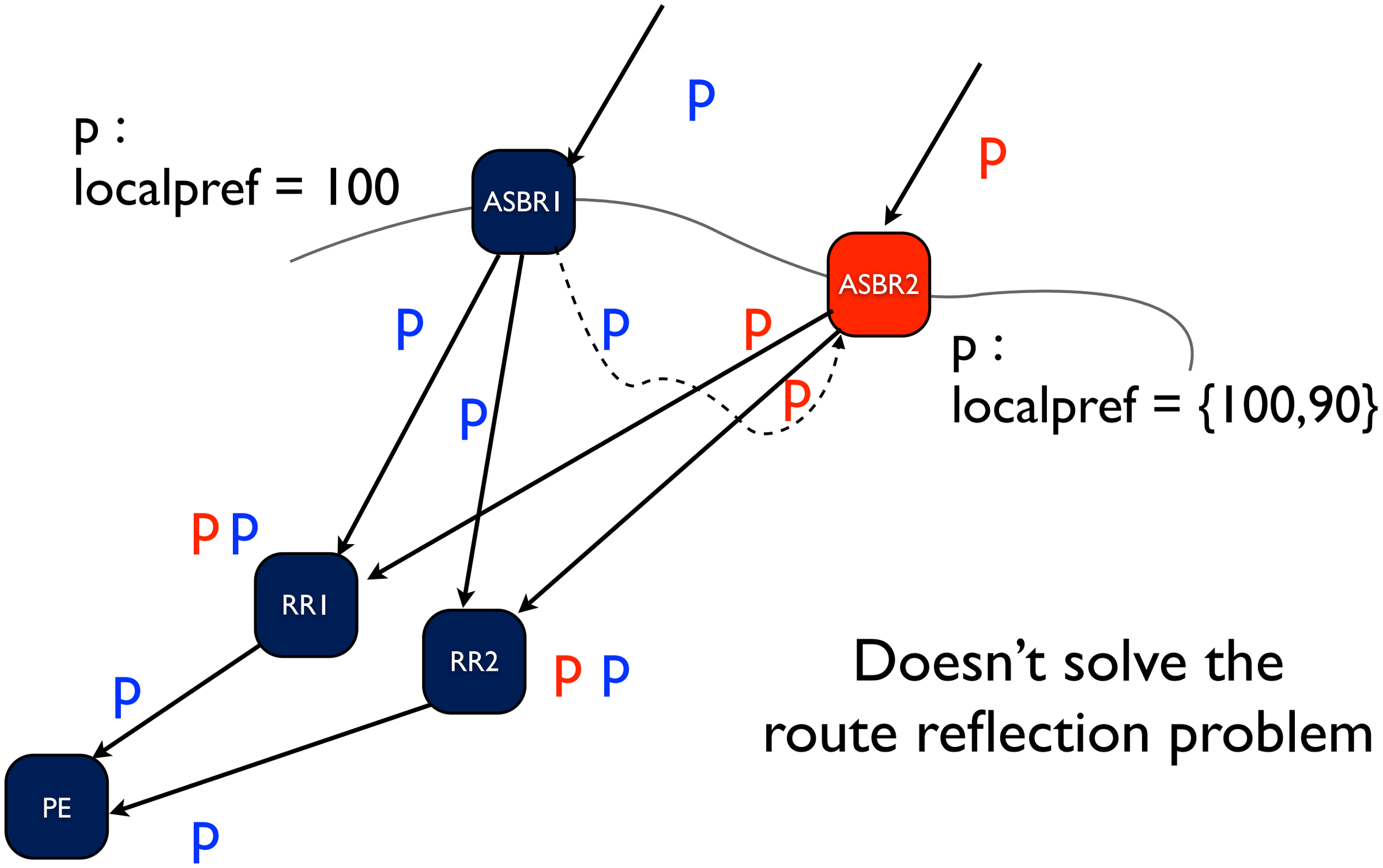
Policies let paths be hidden



Route Reflection hides paths

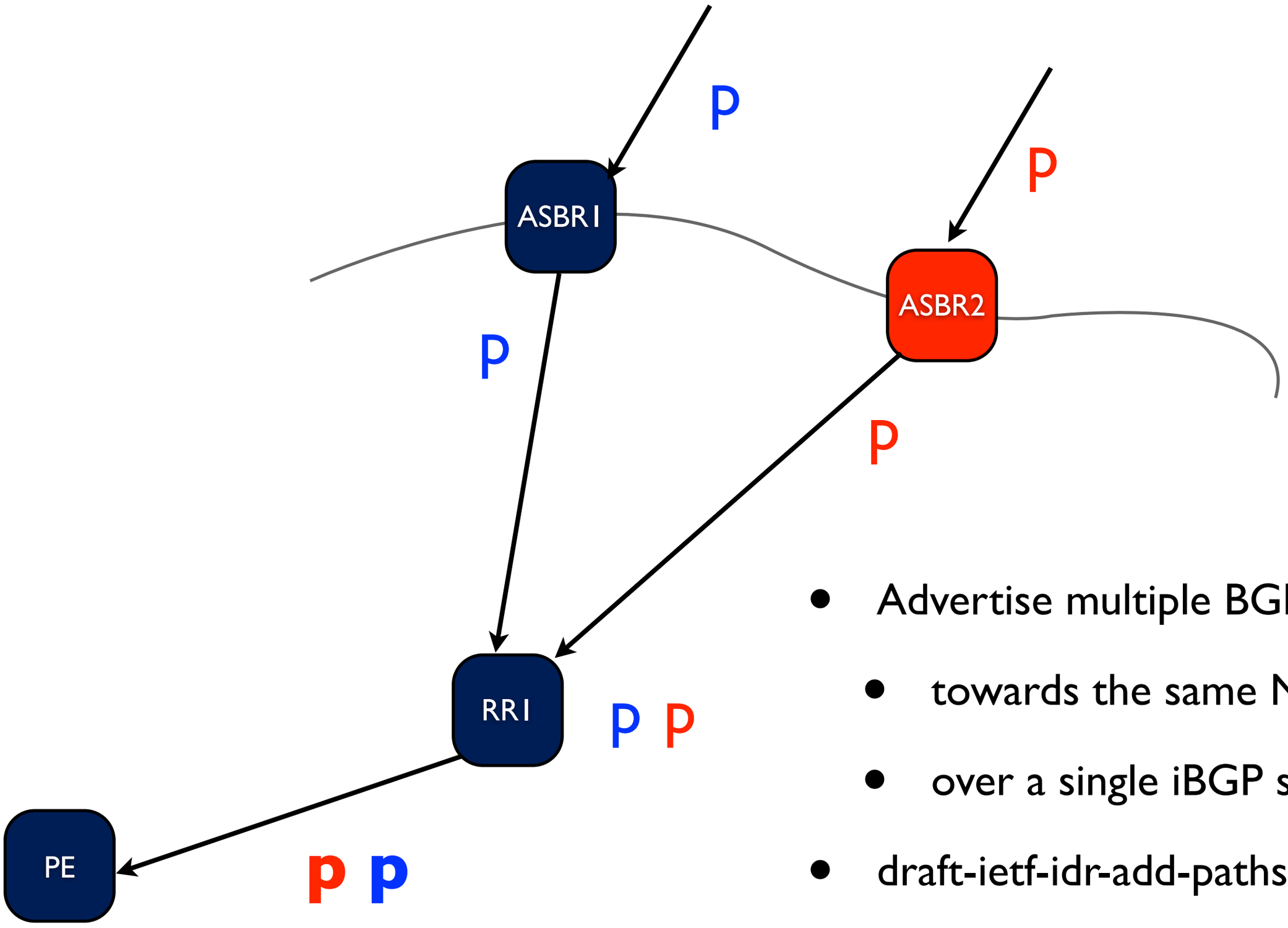


Can’t we just turn adv-best-external on ?



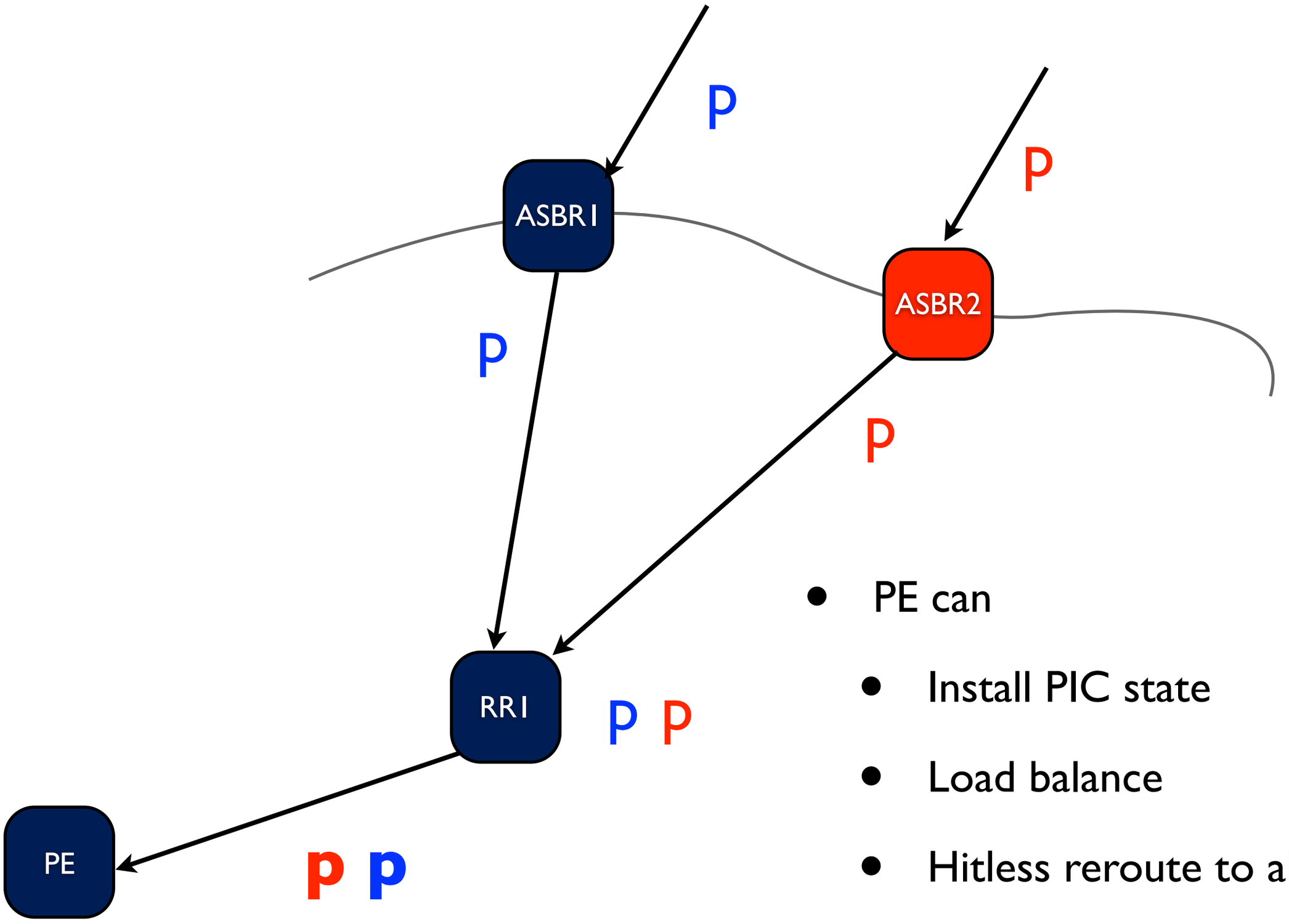
Doesn’t solve the
route reflection problem

BGP Add paths



- Advertise multiple BGP paths
  - towards the same NLRI
  - over a single iBGP session
- draft-ietf-idr-add-paths

BGP Add paths



- PE can
  - Install PIC state
  - Load balance
  - Hitless reroute to alternate

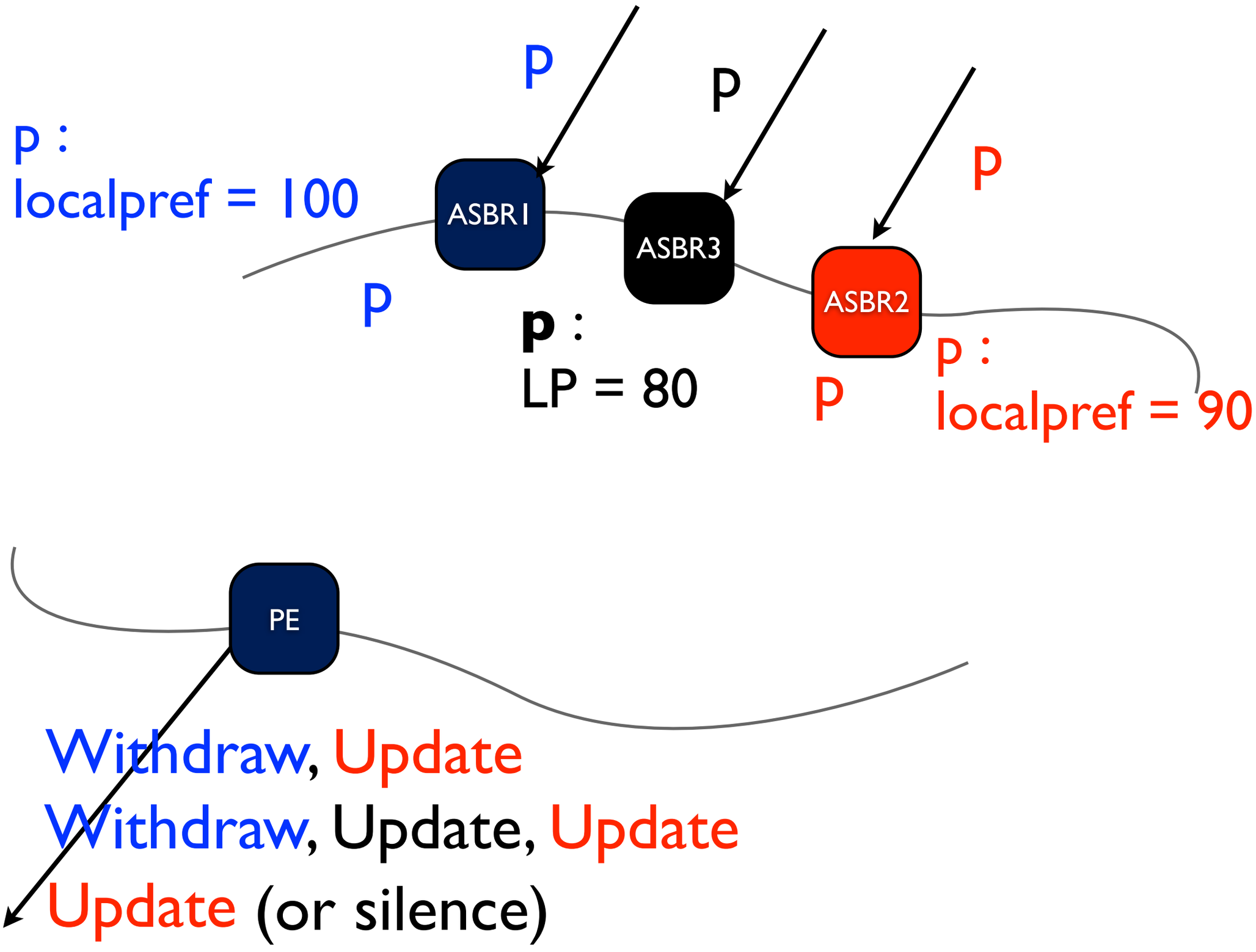
Optimal Hot Potato

- RRs may perform different IGP tie-breaking
  - Clients don't get the path that they would pick
  - Add-paths enabled RRs let the IGP tie-break to clients
    - Depending on which paths it advertises

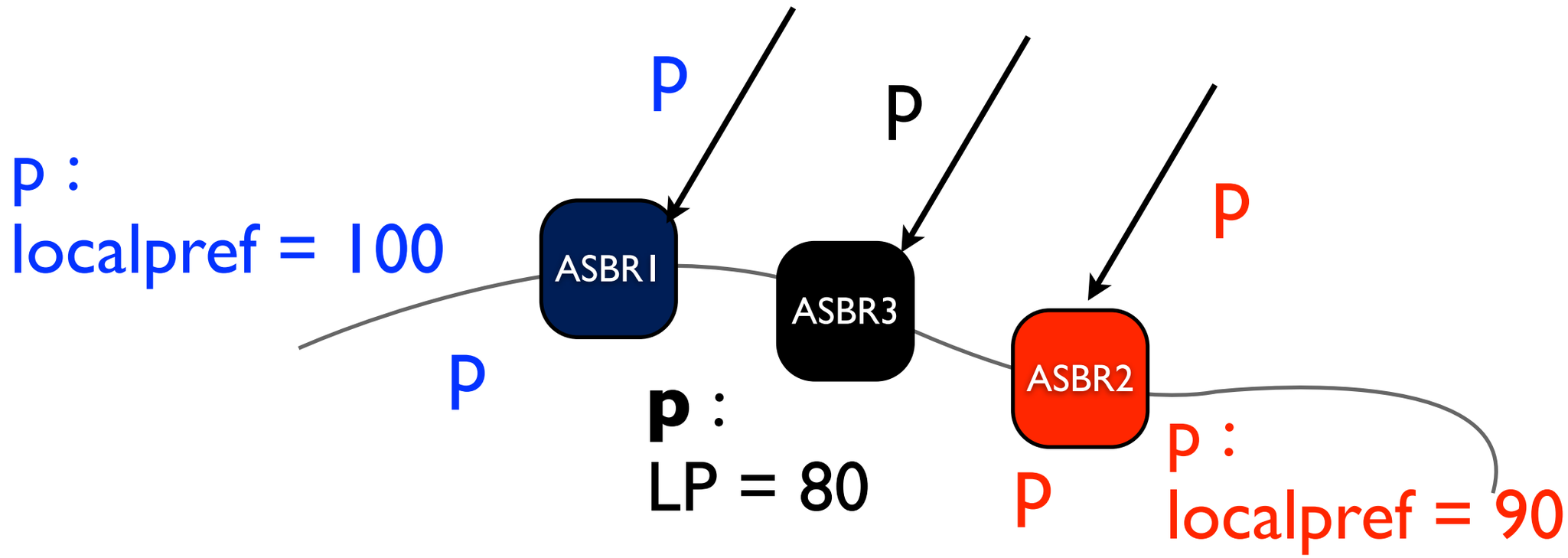
Churn reduction ???

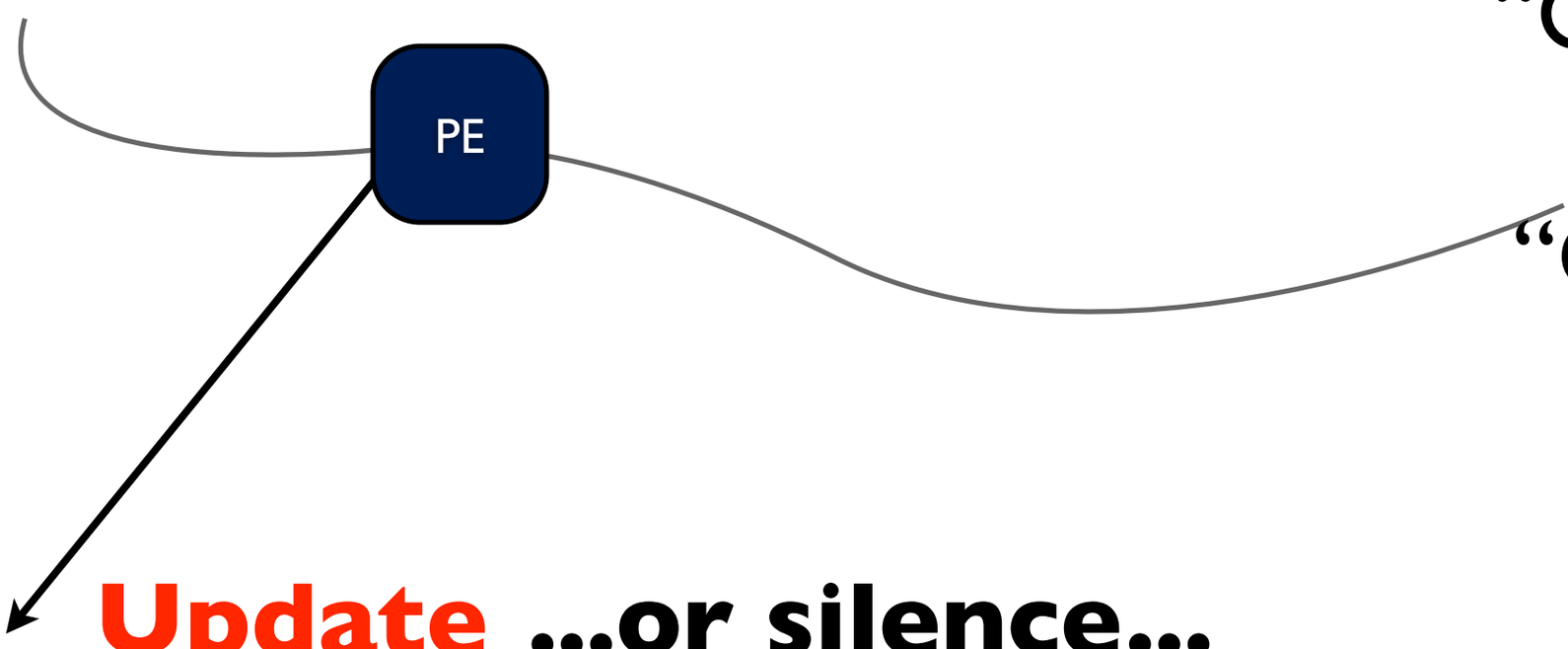
- Churn reduction for primary paths...
- ...with internal churn increase for non-primary ones

Churn Reduction



Churn Reduction





"Churny" internal
for
"Quiet" external

Update ...or silence...

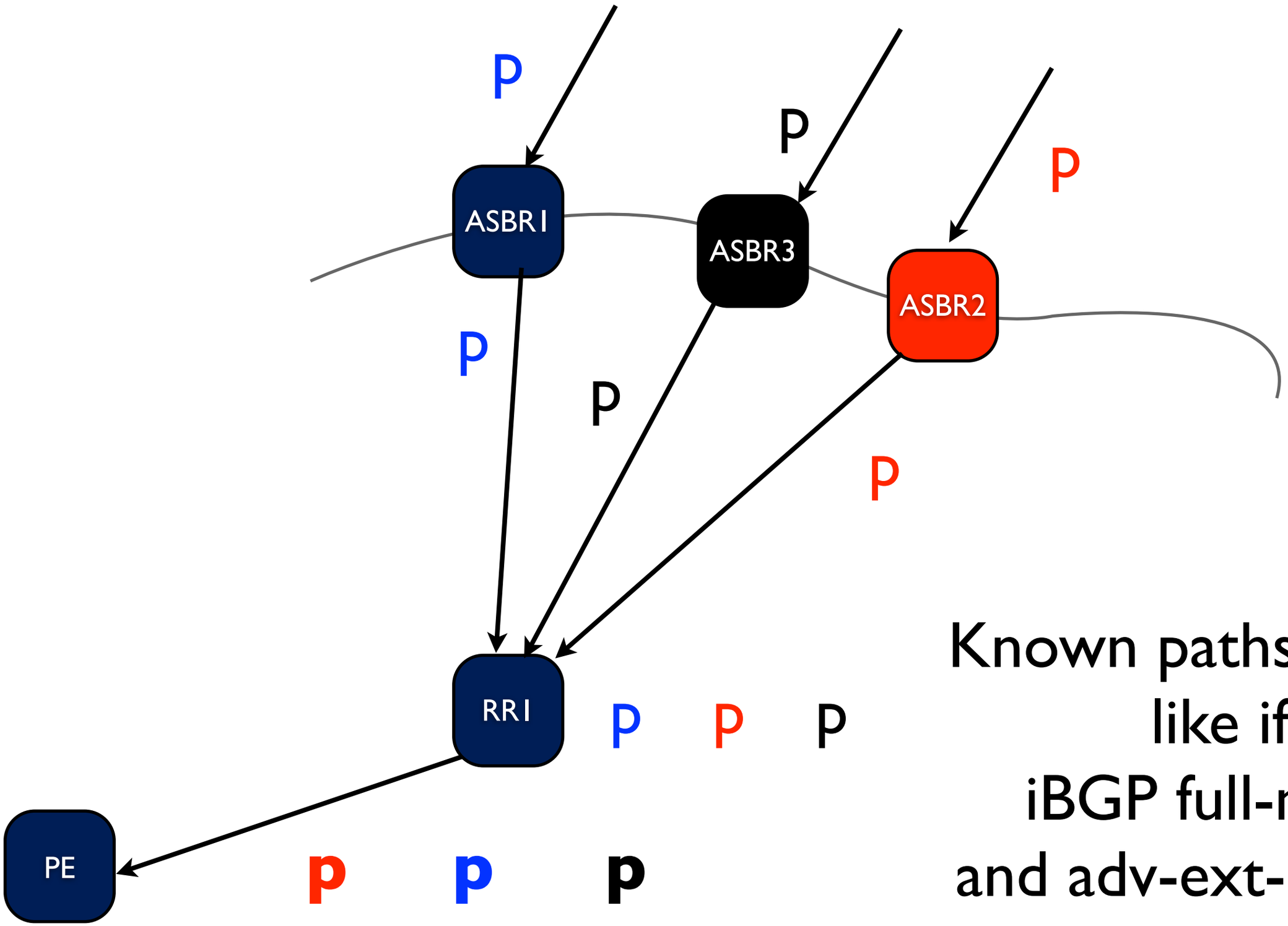
draft-ietf-idr-add-paths-guidelines

- draft-ietf-idr-add-paths doesn't tell which paths to select
  - Multiple motivations lead to different "selection modes"
    - Evaluate them (what they give, at which cost)
      - analytical
      - *"numbers"*

Modes

- All paths
- N paths
- AS-Wide best paths (and variants)
- Best Loc Pref / Second best Loc Pref paths
- Decisive step -1 paths
- Neighbor-AS group best paths

All Paths

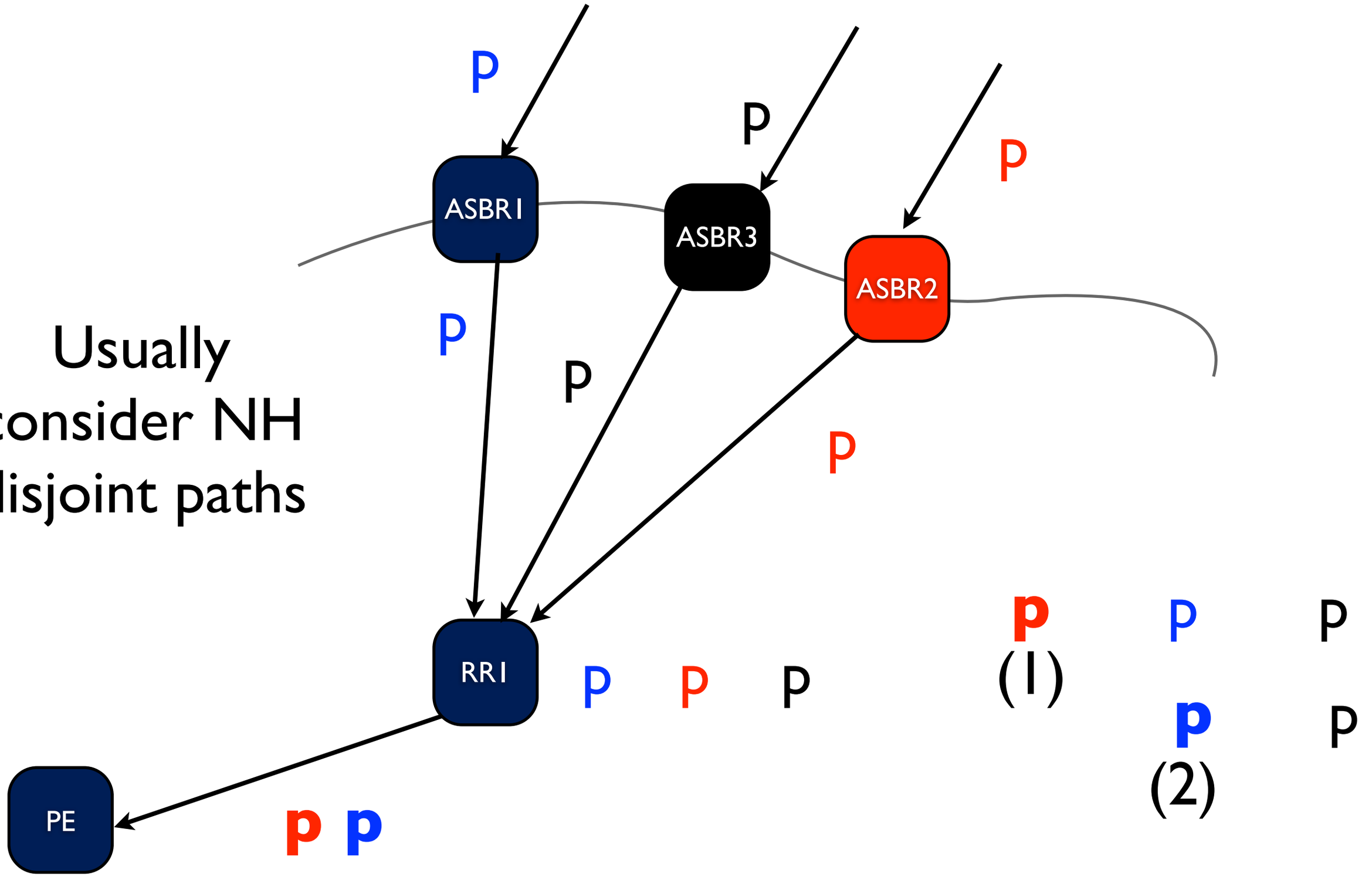


Known paths almost
like if
iBGP full-mesh
and adv-ext-best on

Add-All

- Easiest Decision Process algorithm
- Nice mode to turn on towards a BGP monitor
- Memory/internal update churn monster
  - Depending on how many paths for each p

N paths (N is configured)



Usually consider NH disjoint paths

Add-N-Paths

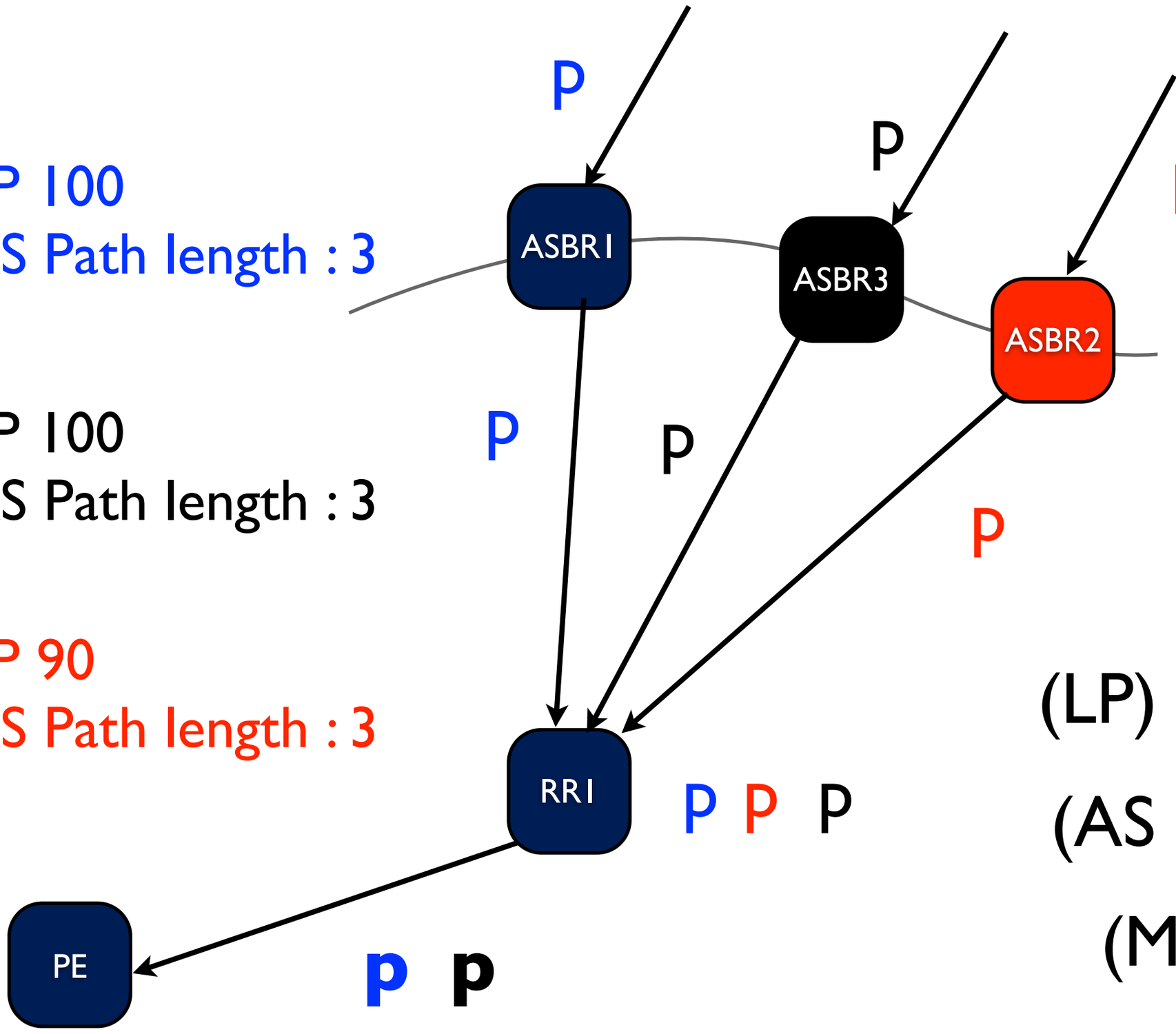
- Most practical use cases
  - Set N to 2 for basic PIC support
  - Set N to desired number of NHs for LB
- Memory hit kept under control through configuration of N
- Doesn't solve MED oscillations
- Developers tend to implement it as N*DP

AS-Wide Best paths

P
LP 100
AS Path length : 3

p
LP 100
AS Path length : 3

P
LP 90
AS Path length : 3



Not hiding paths that another node would have preferred

(LP) P p p
(AS Path) p p
(MED) p p

AS-Wide Best paths

- "The router doesn't make local decisions"
- DP complexity < not running add-paths
- Provides routing optimality and max LB potential
- Provides MED oscillation avoidance

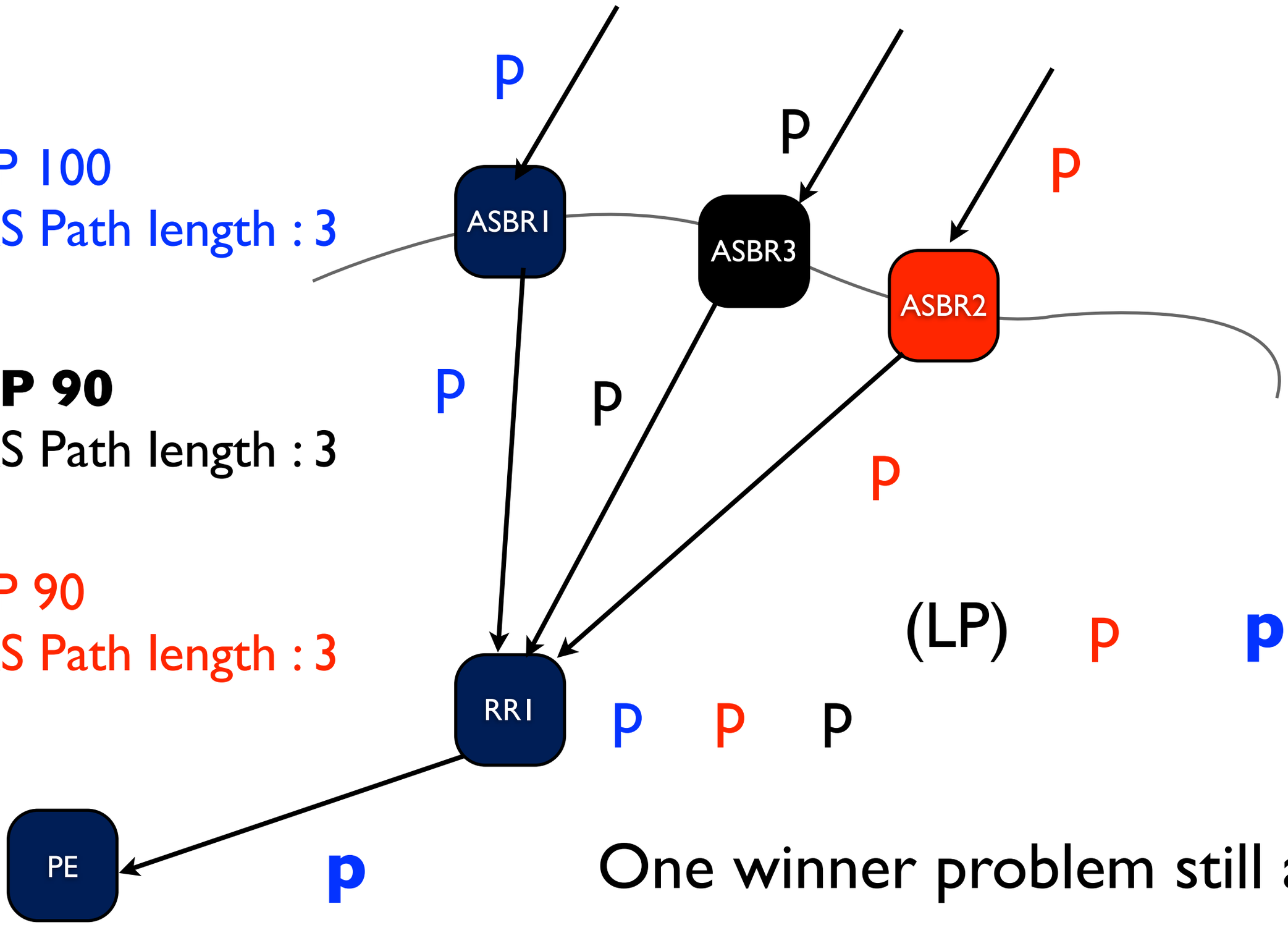
- !!! Doesn't feed PIC !!!

AS-Wide Best paths

P
LP 100
AS Path length : 3

P
LP 90
AS Path length : 3

P
LP 90
AS Path length : 3



(LP) P p P

P P P

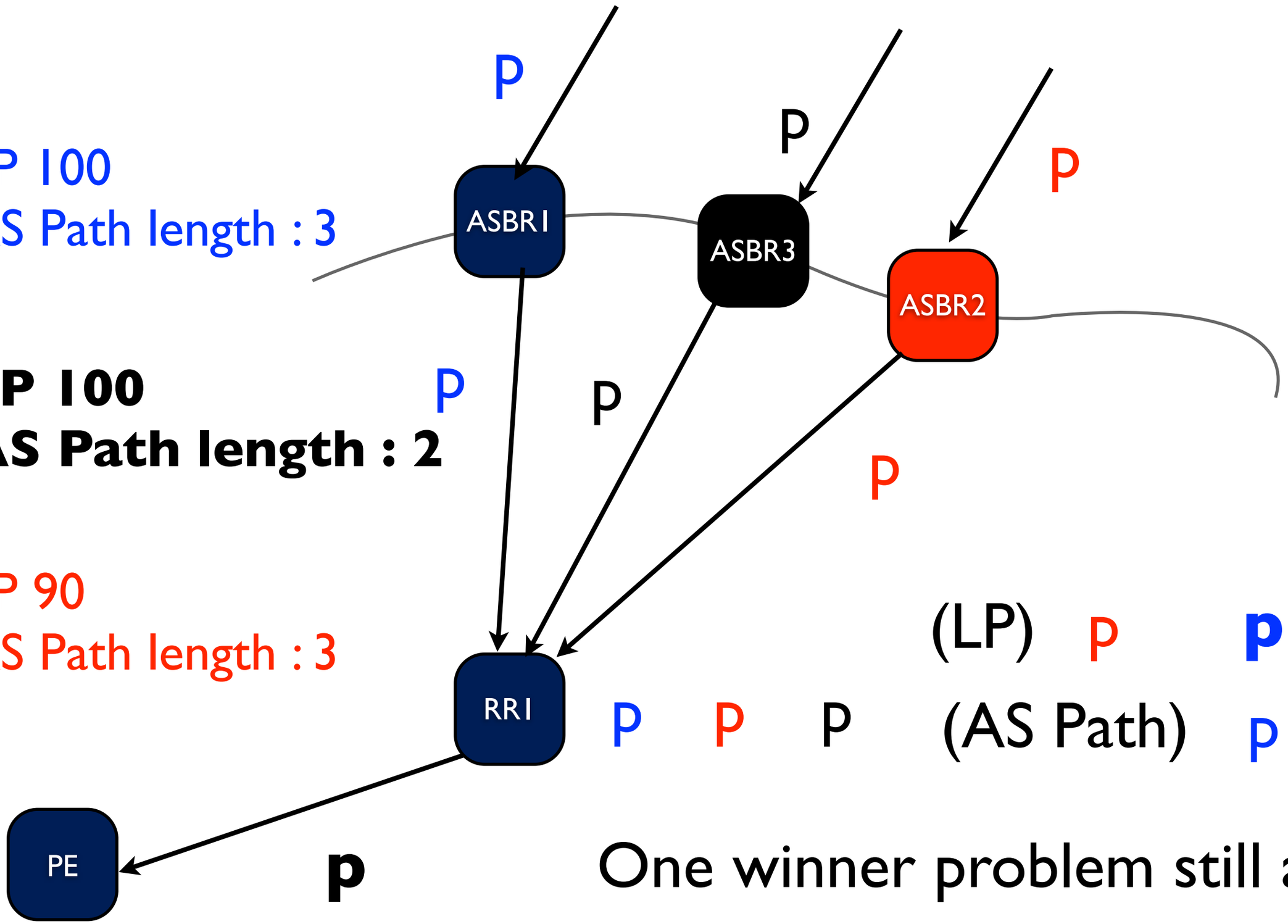
One winner problem still applies

AS-Wide Best paths

P
LP 100
AS Path length : 3

P
LP 100
AS Path length : 2

P
LP 90
AS Path length : 3



(LP) P **p** **p**

P P P (AS Path) p **p**

p

One winner problem still applies

Best LP/Second Best LP

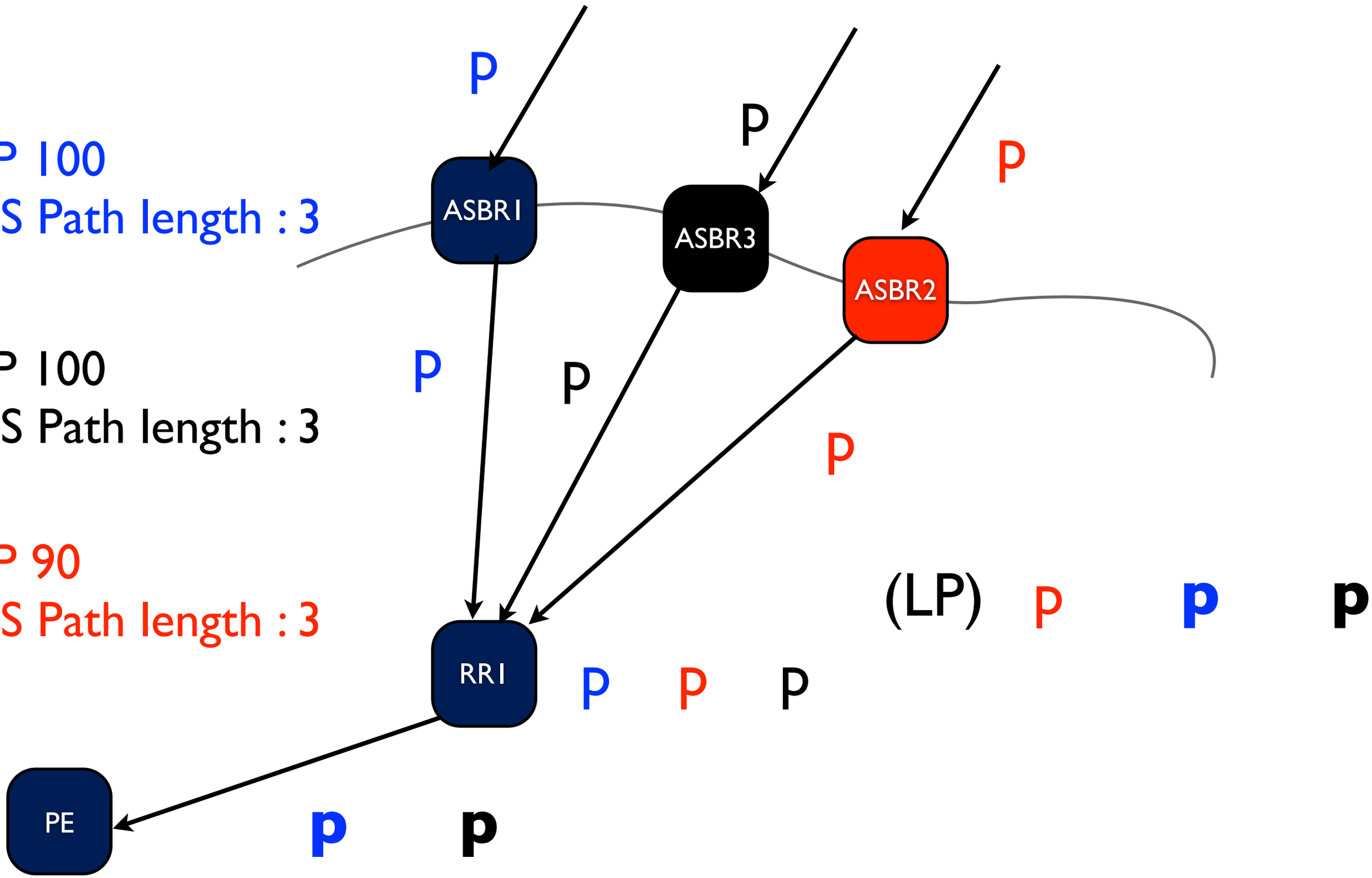
- If #(paths with highest LP) > 1
  - advertise paths with highest LP
- else
  - advertise the path with highest LP
  - advertise the paths with second highest LP

Best LP/Second Best LP

P
LP 100
AS Path length : 3

P
LP 100
AS Path length : 3

P
LP 90
AS Path length : 3

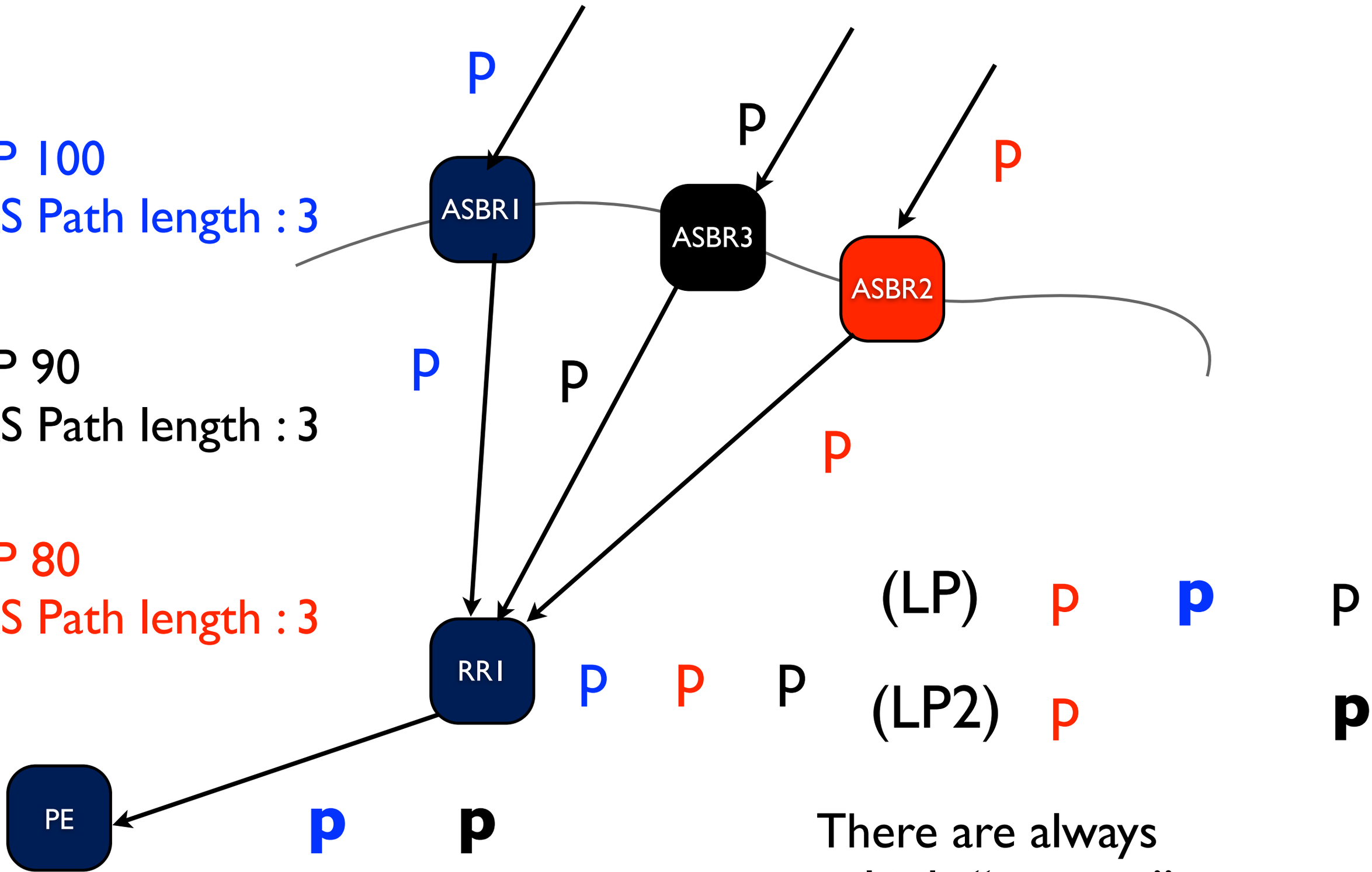


Best LP/Second Best LP

P
LP 100
AS Path length : 3

P
LP 90
AS Path length : 3

P
LP 80
AS Path length : 3



(LP) P p P

(LP2) P p

There are always multiple "winners"

Best LP / Second Best LP

- Adj-Rib-In optimized for this mode contains two or three sets of paths per NLRI
  - Best bin
  - Second best bin if required
  - Others
- Decision Process :
  Select what's in first and second bin

Decisive step - 1

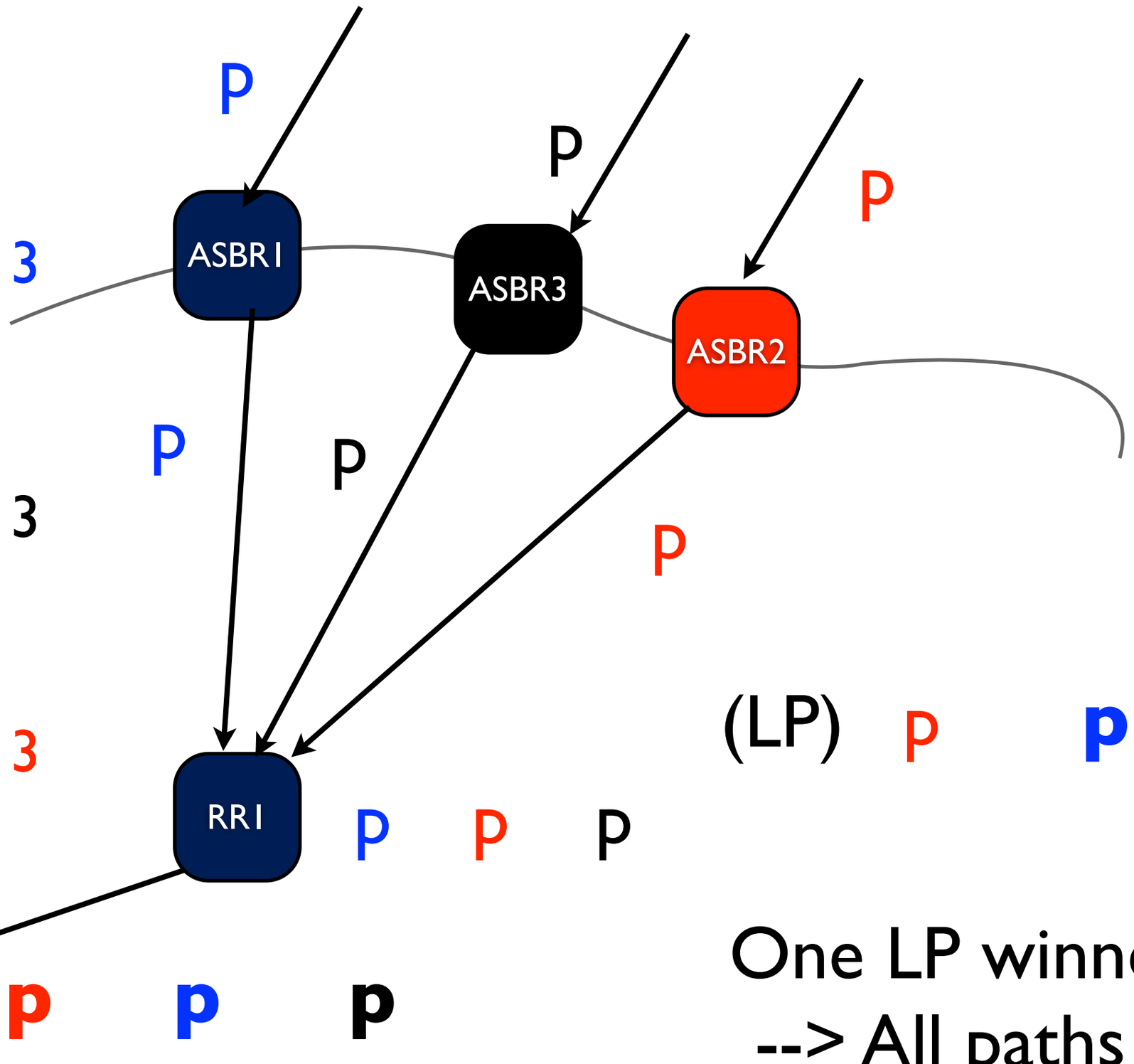
- Apply normal BGP selection process, but
  - If IGP tie-break rule is reached, advertise what remains
    - If best path is found at a preceding rule i, advertise what remained when applying rule i-1
- Tries to obtain diversity while advertising as few paths as possible

Decisive step - 1

P
LP 100
AS Path length : 3

P
LP 90
AS Path length : 3

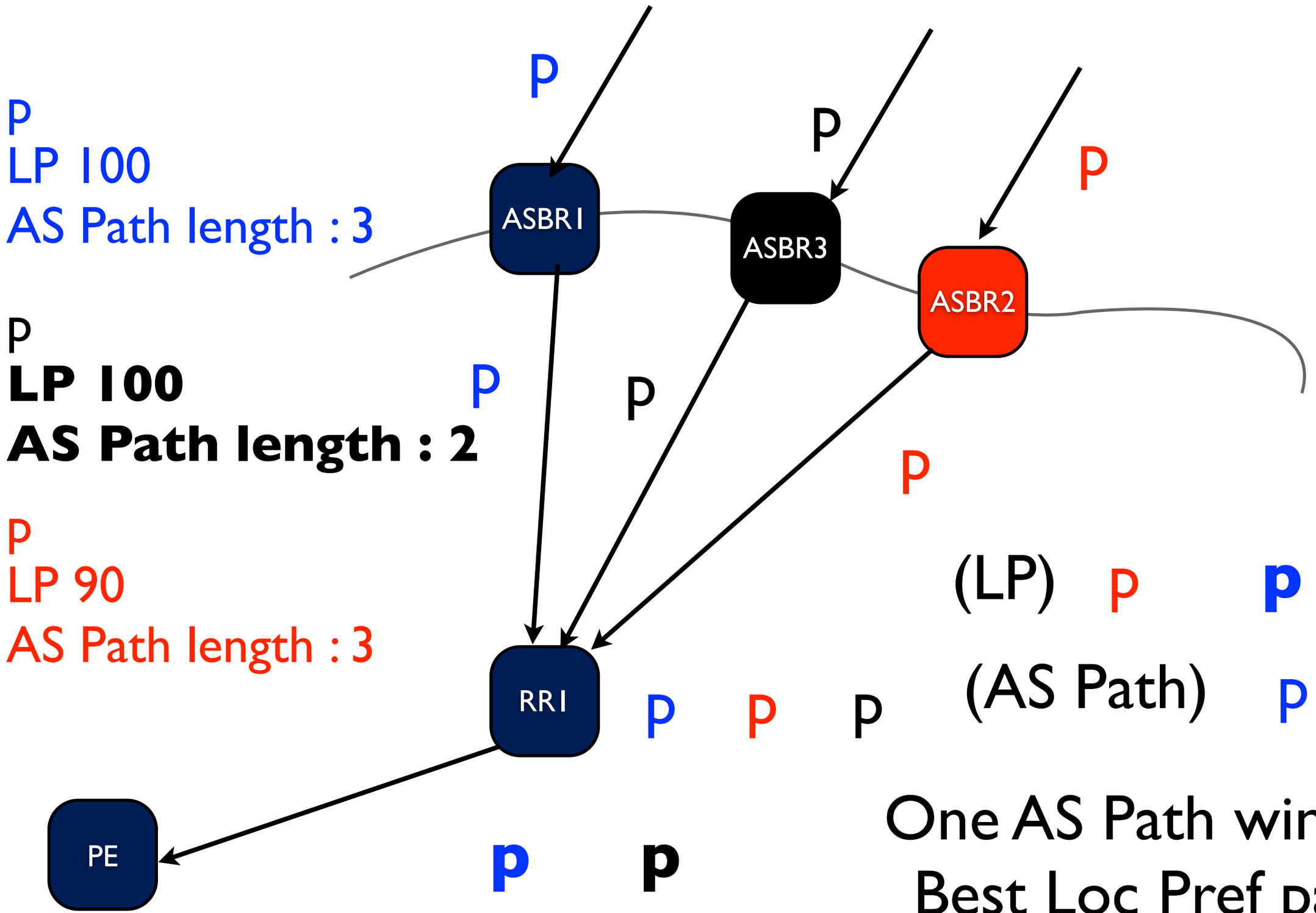
P
LP 90
AS Path length : 3



(LP) P **p** P

One LP winner
--> All paths !

Decisive step - 1



Neighbor-AS group best

- Avoids MED oscillations
  - draft-walton-bgp-route-oscillation-stop
- Advertise the best path from each neighboring AS
  - No ASBR picks as best a non-lowest MED path

Neighbor-AS group best

- Provides paths from different neighboring ASes, but
  - their existence is not guaranteed
  - nothing to deal with post-convergence paths

Summary

	Path optimality	Backup availability / optimality	Control plane load and stress	DP Complexity	MED osc. avoidance
All	OK	OK	Max	EASIEST	OK
N	?	OK / ?	Bounded	Depends on N can be optimized	?
AS-Wide	OK	KO / ~OK	~MAX	EASY	OK
LP1/LP2	OK	OK	~MAX	EASIER	OK
Decisive-1	OK	OK	~MAX	Easy but "spaghetti"	OK
Group best	KO ...	KO	~MAX	?	OK

Summary

	Path optimality	Backup availability / optimality	Control plane load and stress	DP Complexity	MED osc. avoidance
All	OK	OK	**Max**	EASIEST	OK
N	?	OK / ?	**Bounded**	Depends on N can be optimized	?
AS-Wide	OK	KO / ~OK	**~MAX**	EASY	OK
LP1/LP2	OK	OK	**~MAX**	EASIER	OK
Decisive-1	OK	OK	**~MAX**	Easy but "spaghetti"	OK
Group best	KO ...	KO	**~MAX**	?	OK

Summary

	Path optimality	Backup availability / optimality	Control plane load and stress	DP Complexity	MED osc. avoidance
All	OK	OK	Max	EASIEST	**OK**
N	?	OK / ?	Bounded	Depends on N can be optimized	**?**
AS-Wide	OK	KO / ~OK	~MAX	EASY	**OK**
LP1/LP2	OK	OK	~MAX	EASIER	**OK**
Decisive-1	OK	OK	~MAX	Easy but "spaghetti"	**OK**
Group best	KO ...	KO	~MAX	?	**OK**

Summary

	Path optimality	Backup availability / optimality	Control plane load and stress	DP Complexity	MED osc. avoidance
All	OK	OK	Max	EASIEST	**OK**
N	?	OK / ?	Bounded	Depends on N can be optimized	**?**
AS-Wide	**OK**	KO / ~OK	~MAX	EASY	**OK**
LP1/LP2	OK	OK	~MAX	EASIER	**OK**
Decisive-1	OK	OK	~MAX	Easy but "spaghetti"	**OK**
Group best	**KO ...**	KO	~MAX	?	**OK**

Summary

	Path optimality	Backup availability / optimality	Control plane load and stress	DP Complexity	MED osc. avoidance
All	OK	OK	Max	EASIEST	OK
N	?	OK / ?	Bounded	Depends on N can be optimized	?
AS-Wide	OK	KO / ~OK	~MAX	EASY	OK
LP1/LP2	OK	OK	~MAX	EASIER	OK
Decisive-1	OK	OK	~MAX	Easy but "spaghetti"	OK
Group best	KO ...	KO	~MAX	?	OK

Summary

	Path optimality	Backup availability / optimality	Control plane load and stress	DP Complexity	MED osc. avoidance
All	OK	OK	Max	EASIEST	OK
N	?	OK / ?	Bounded	Depends on N can be optimized	?
AS-Wide	OK	KO / ~OK	~MAX	EASY	OK
LP1/LP2	OK	OK	~MAX	EASIER	OK
Decisive-1	OK	OK	~MAX	**Easy but "spaghetti"**	OK
Group best	KO ...	KO	~MAX	?	OK

Summary

	Path optimality	Backup availability / optimality	Control plane load and stress	DP Complexity	MED osc. avoidance
All	OK	OK	Max	EASIEST	OK
N	?	OK / ?	Bounded	Depends on N can be optimized	?
AS-Wide	OK	KO / ~OK	~MAX	EASY	OK
LP1/LP2	OK	OK	~MAX	EASIER	OK
Decisive-1	OK	OK	~MAX	Easy but "spaghetti"	OK
Group best	KO ...	KO	~MAX	?	OK

Summary

	Path optimality	Backup availability / optimality	Control plane load and stress	DP Complexity	MED osc. avoidance
All	OK	OK	Max	EASIEST	OK
N	?	OK / ?	Bounded	Depends on N can be optimized	?
AS-Wide	OK	KO / ~OK	~MAX	EASY	OK
LP1/LP2	OK	OK	~MAX	EASIER	OK
Decisive-1	OK	OK	~MAX	Easy but "spaghetti"	OK
Group best	KO ...	KO	~MAX	?	OK

Summary

	Path optimality	Backup availability / optimality	Control plane load and stress	DP Complexity	MED osc. avoidance
All	OK	OK	Max	EASIEST	OK
N	?	OK / ?	Bounded	Depends on N can be optimized	?
AS-Wide	OK	KO / ~OK	~MAX	EASY	OK
LP1/LP2	OK	OK	~MAX	EASIER	OK
Decisive-1	OK	OK	~MAX	Easy but "spaghetti"	OK
Group best	KO ...	KO	~MAX	?	OK

Current Recommendations

- MUST: Add-N
  - Default MUST be 2
    - N MUST be configurable
    - Option to not limit N (Add-All)
- OPTIONAL: AS-Wide best variants
- OPTIONAL-: All others

Deployment

- Session wide upgrade required
- Ongoing works on migration
- Forget about deployment w/o Ingress-Egress encap

Next Steps

- Add-path for eBGP
  - Route Server implementation
  - draft-jasinska-ix-bgp-route-server
    +Add-All
    +Filtering
    +Pick one for clients not supporting add-paths
- Migrating to Add-paths (ISPs freak out)
- Adj-Rib-out management optimization

BGP Policy violations in the data-plane

With Paolo Lucente, Cariden/PMACCT
paolo@pmacct.net

- Two well-known facts about routing...
- leading to policy violations...

Observation 1

- Policy-constrained path selection in BGP...
  Flexible, per-prefix granularity

- *"A BGP-router's* ***route processor*** *will pick a path towards a given destination* ***prefix*** *by applying the following rules"*

  Weight
  Local-pref
  As Path Length
  IGP/Med
  ...

Observation 1

- ... dominated in the data-plane
- A **FIB** will pick a path towards a given destination **address** by applying the following rules

Longest prefix match to get the prefix

(
Best path towards that prefix was picked based on
Weight
Local-pref
As Path Length
IGP/Med
...)

Observation 11

- Common to provide a lot of routing flexibility
- Route propagation control offered by Sprint
  - Have to be a customer of Sprint
  - 65000:XXX : Do not advertise to ASXXX
    can be AOL, NTT, BT, Level3, GBLX, Verizon, AT&T, ...

Powerful complementary means to limit path knowledge

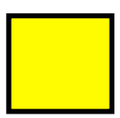
- Selective advertisement, performed locally
- Selective advertisement, triggered remotely

Control-plane/Data-plane can mismatch

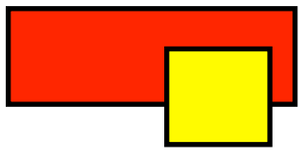
- Paths for **overlapping** prefixes are controlled independently
  - By yourself
  - By your BGP neighborhood
- Forwarding plane dominated by the longest prefix match rule
- What if your policy differs for overlapping prefixes ?

Toy case study

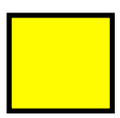
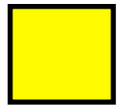
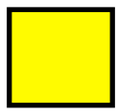
A BGP advertisement for NLRI P/p



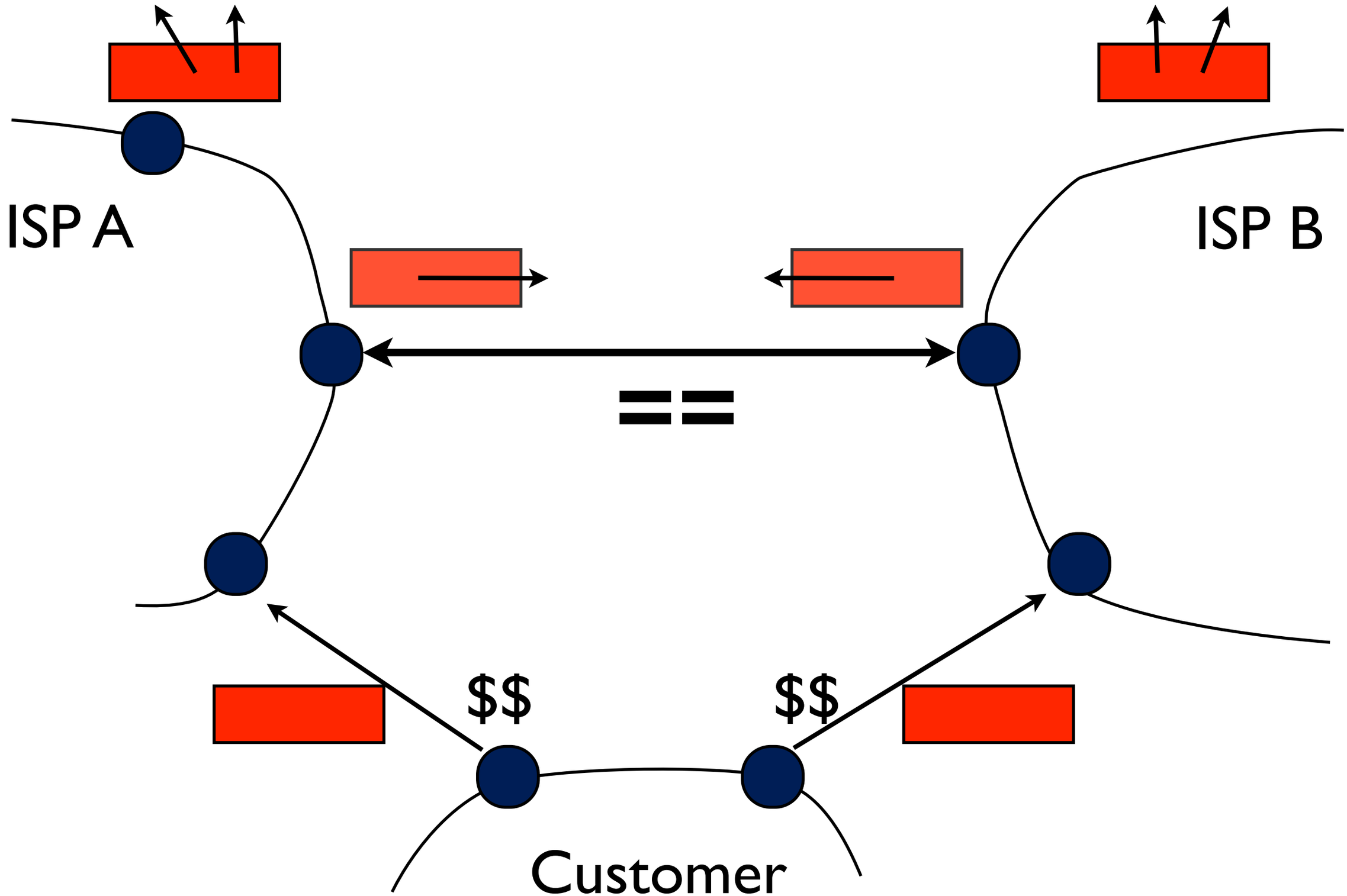
A BGP advertisement of a prefix more specific than P/p, say P/p+1



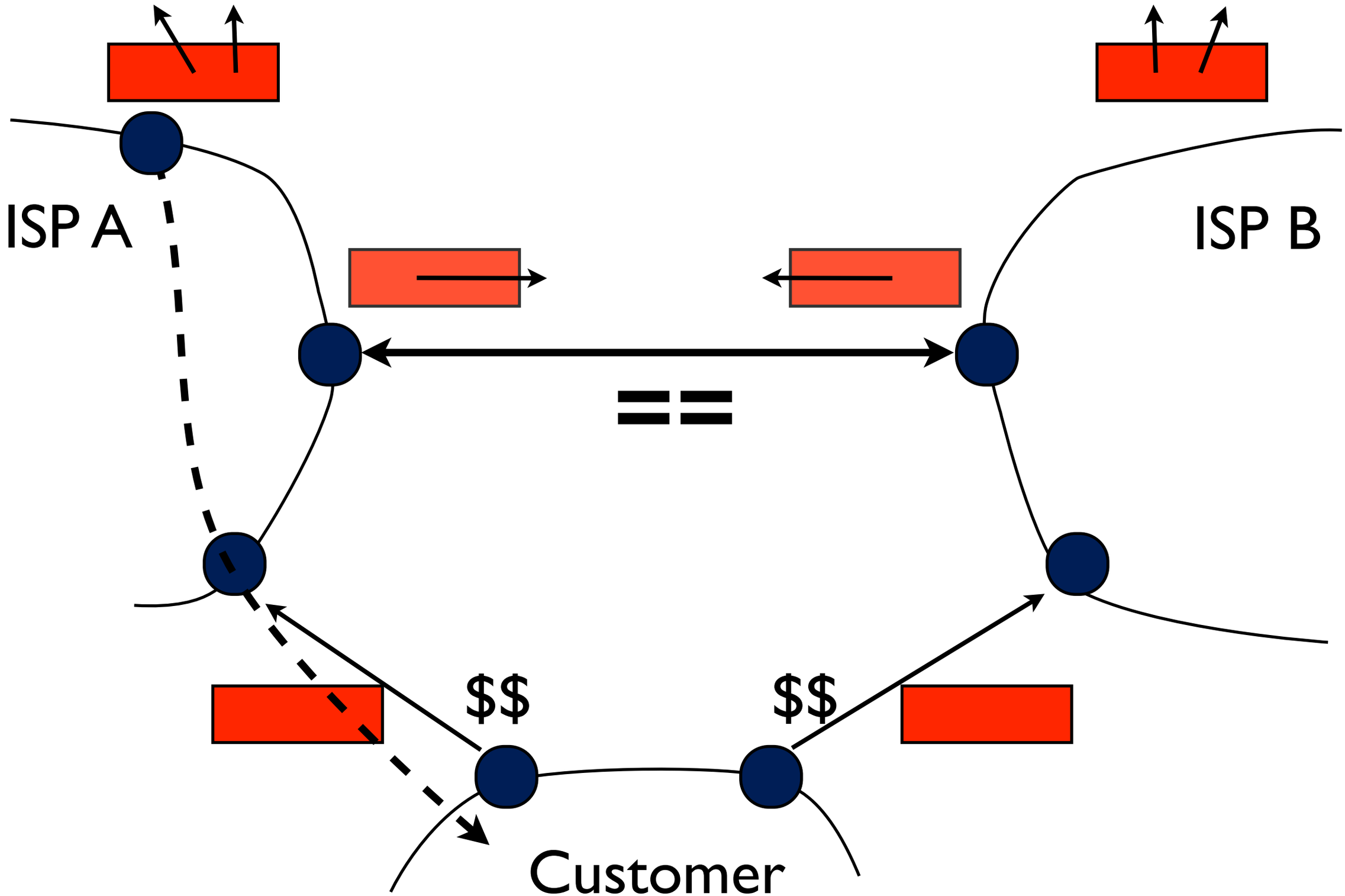
The BGP policy violation trick

- Play with  and communities
- Make  reach only a subset of the ASes
  - Some ASes forward  according to 
  - Until packet reaches an AS knowing 
  - Resulting data-plane not necessarily fitting everyone's policy...

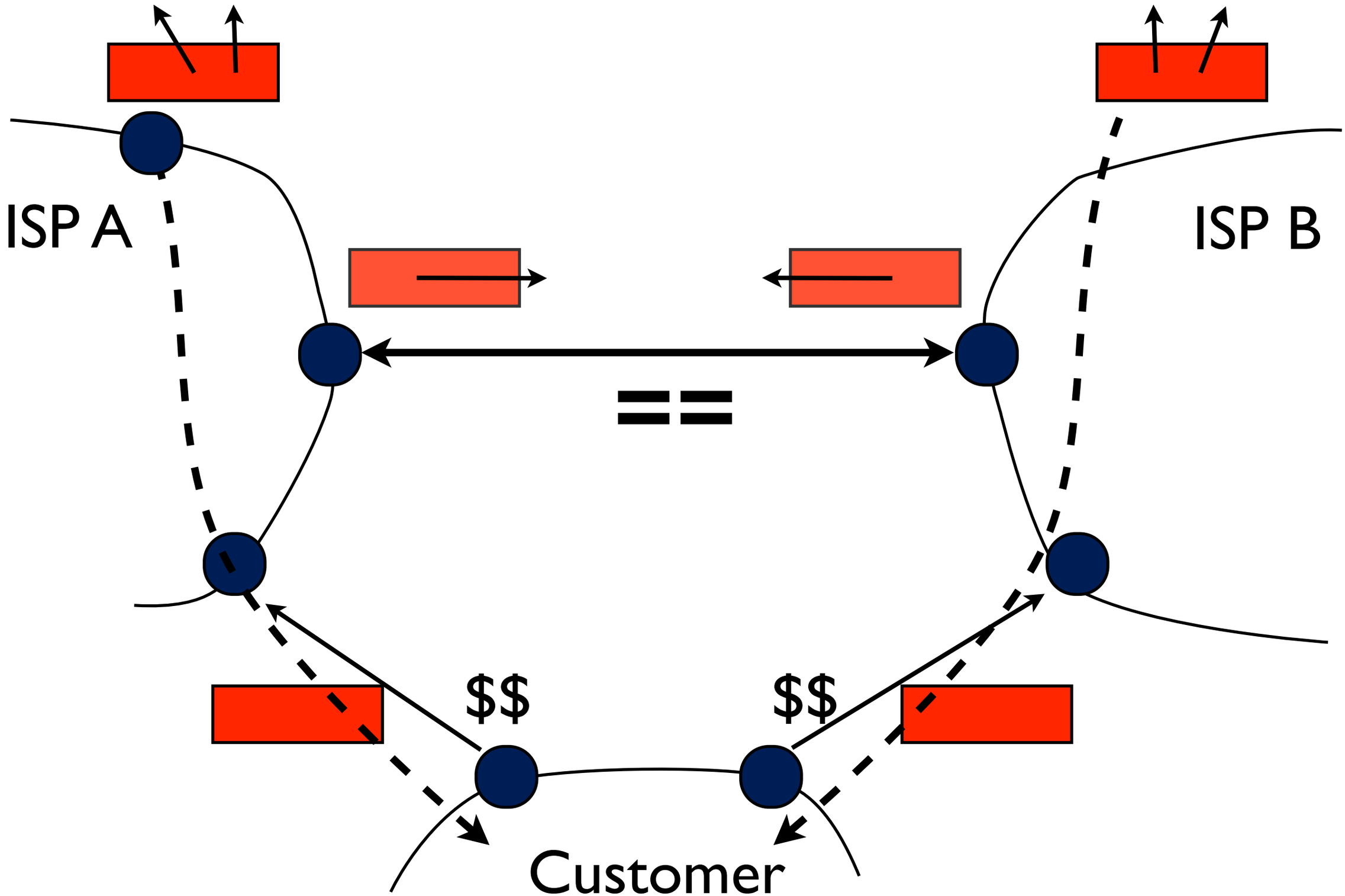
Initial routing status



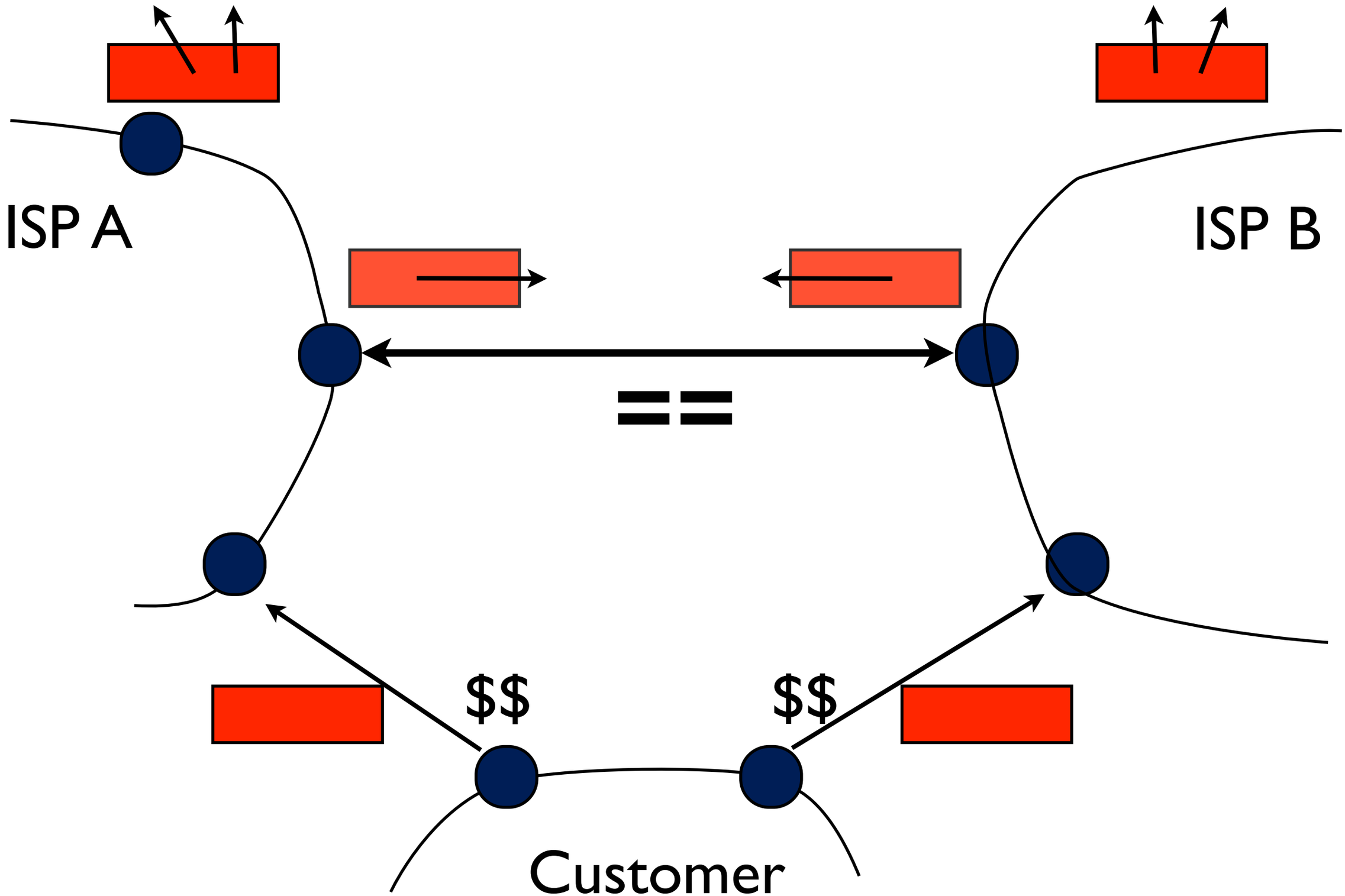
Initial routing status



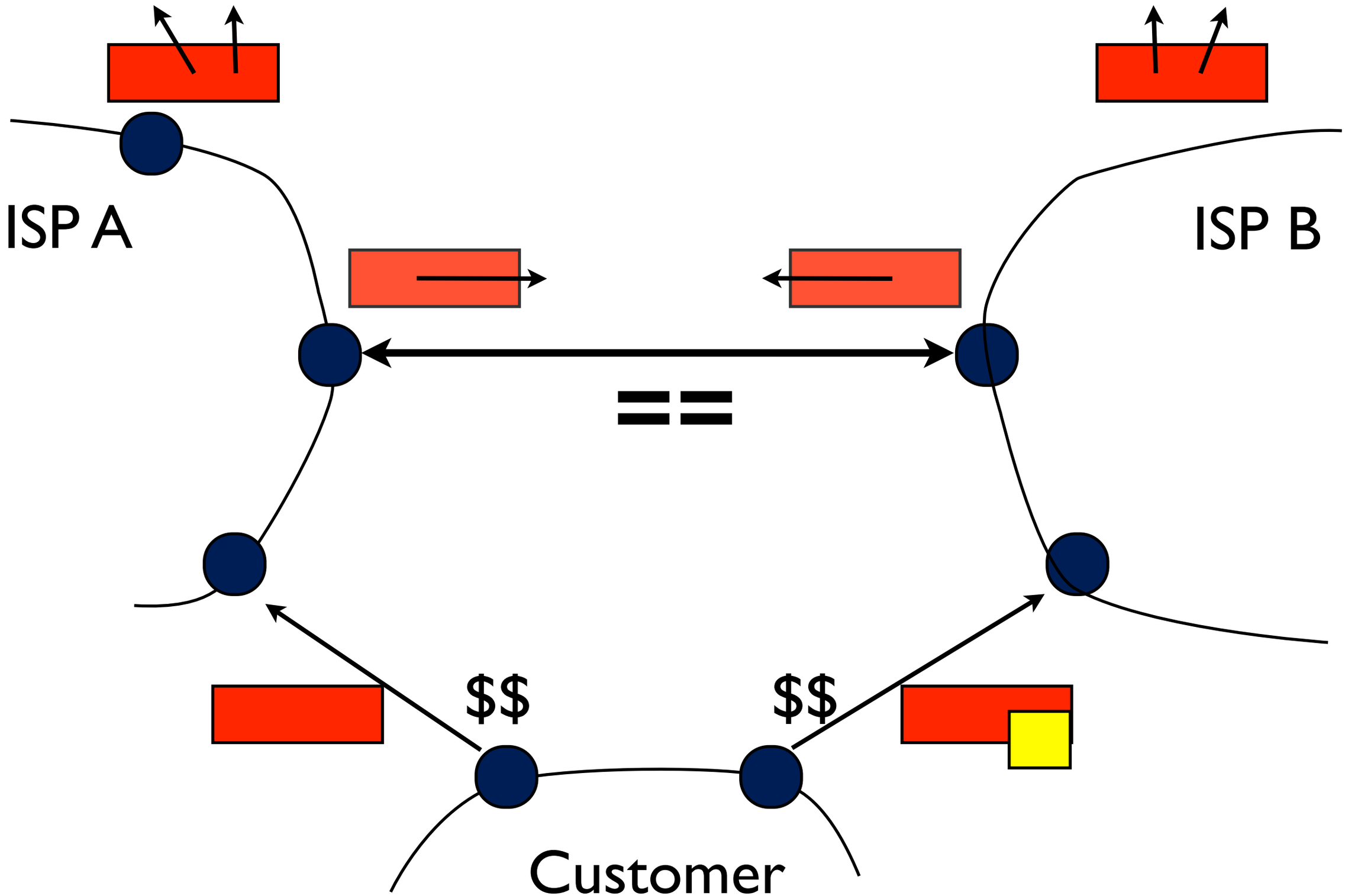
Initial routing status



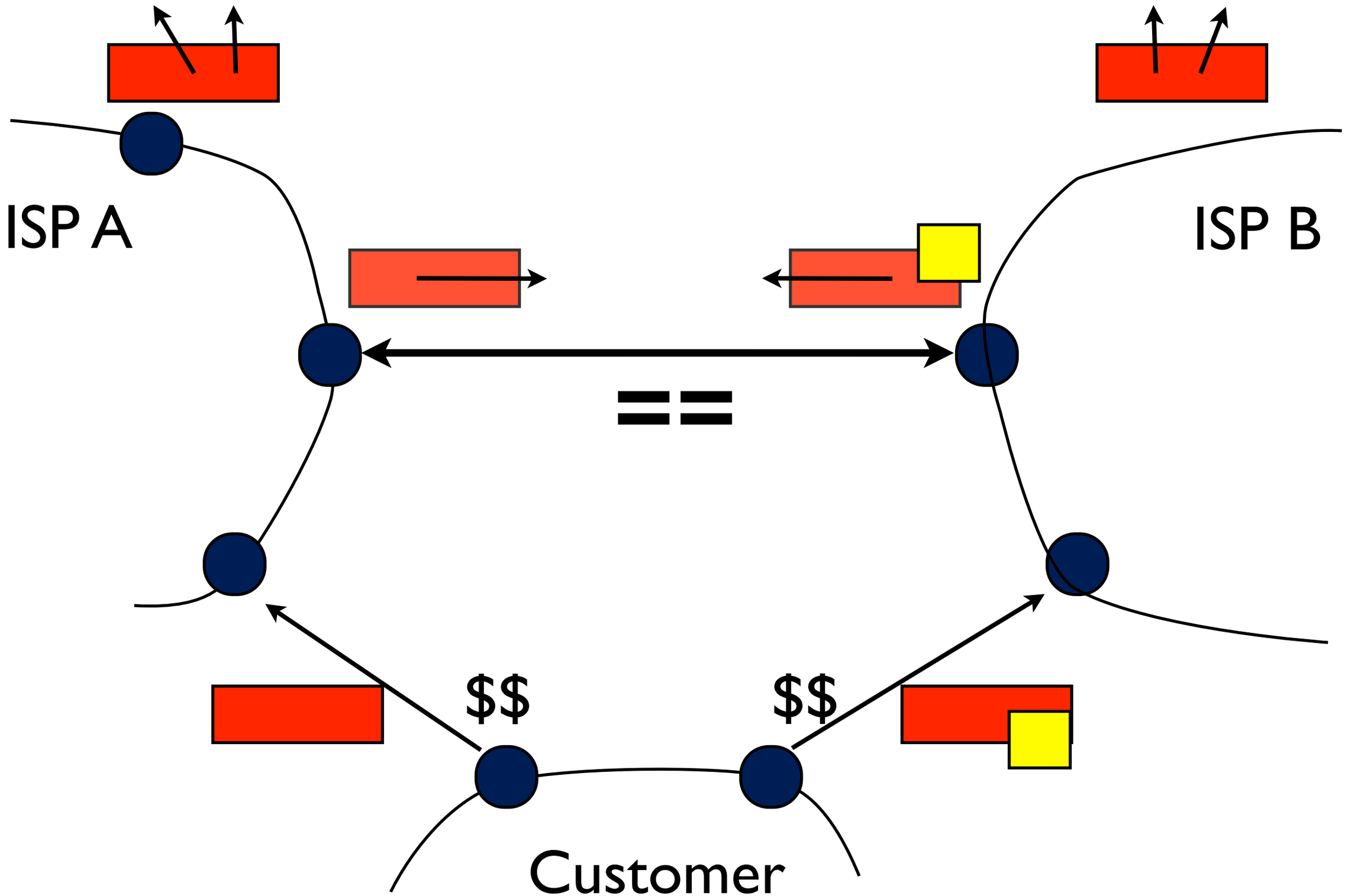
Inbound TE, selective advertisement of a more specific prefix



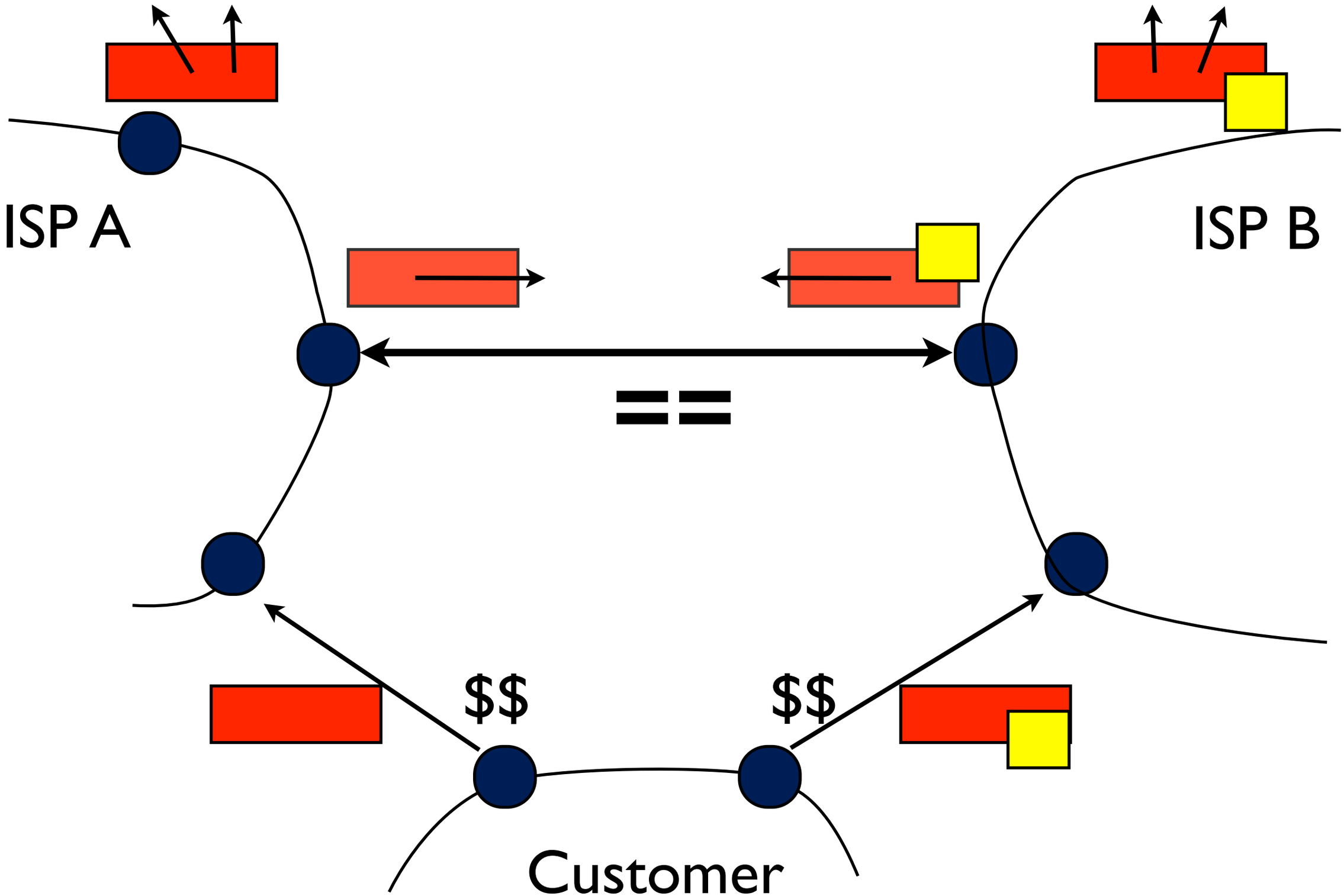
Inbound TE, selective advertisement of a more specific prefix



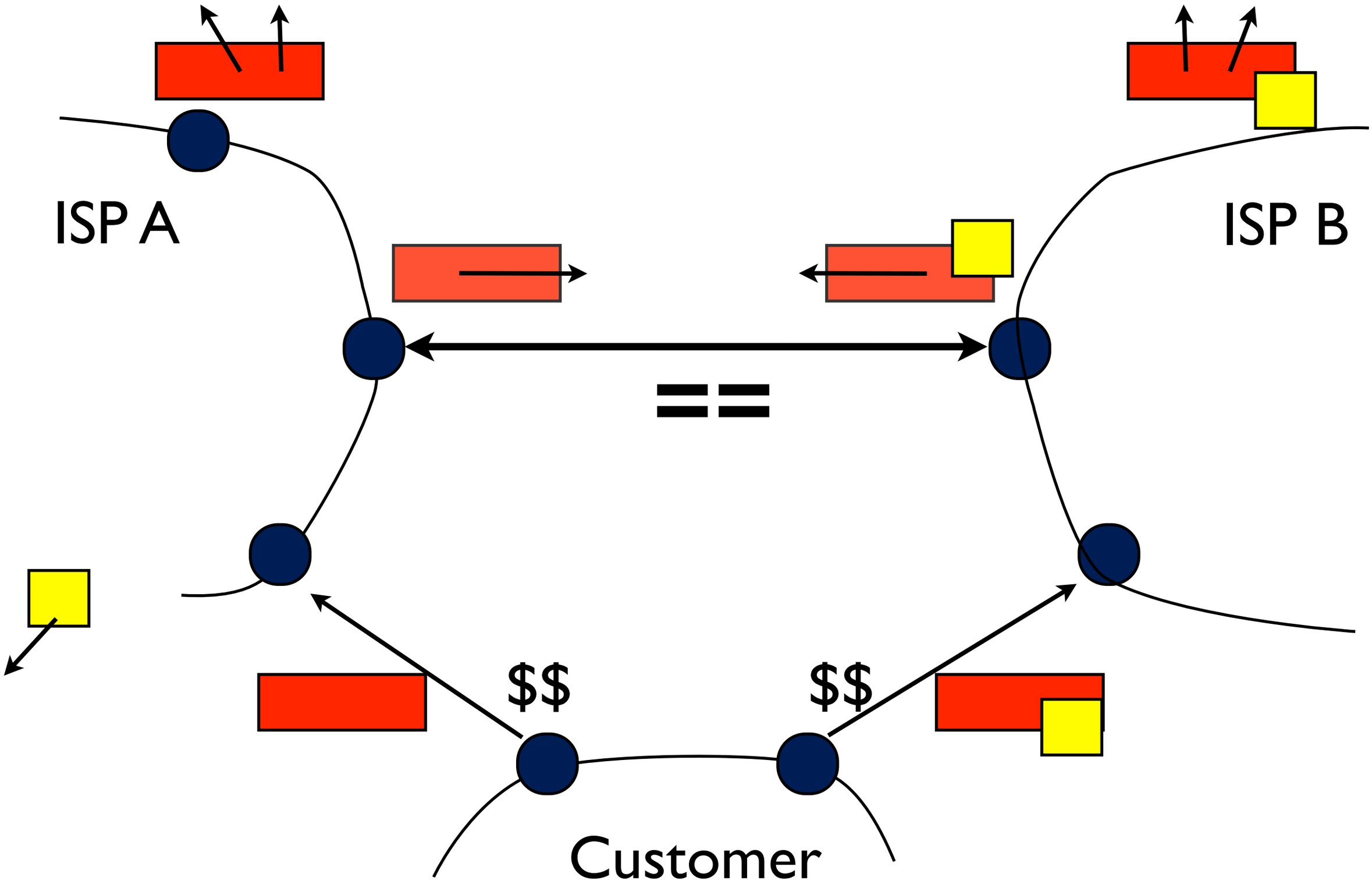
Inbound TE, selective advertisement of a more specific prefix



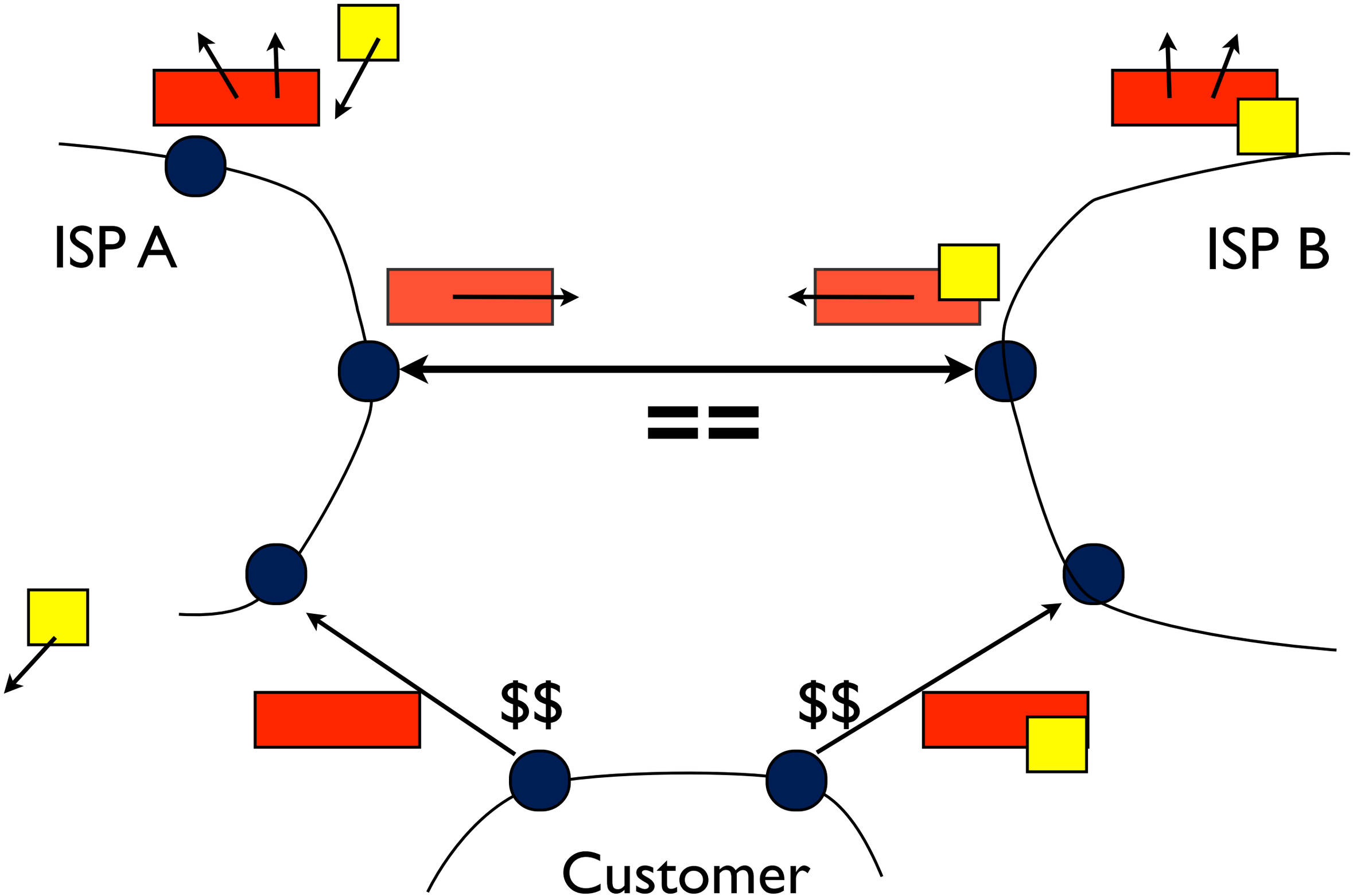
Inbound TE, selective advertisement of a more specific prefix



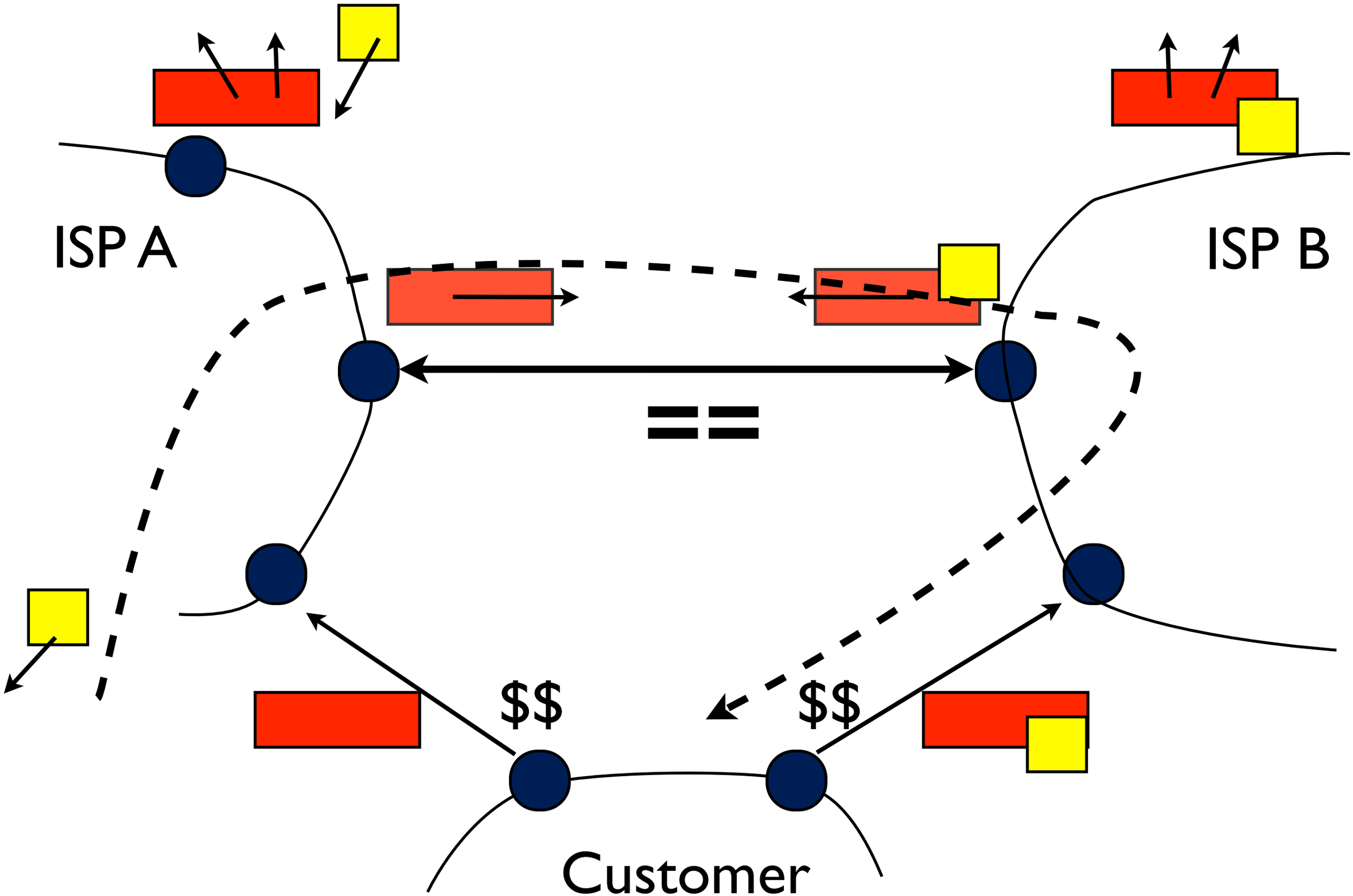
Inbound TE, selective advertisement of a more specific prefix



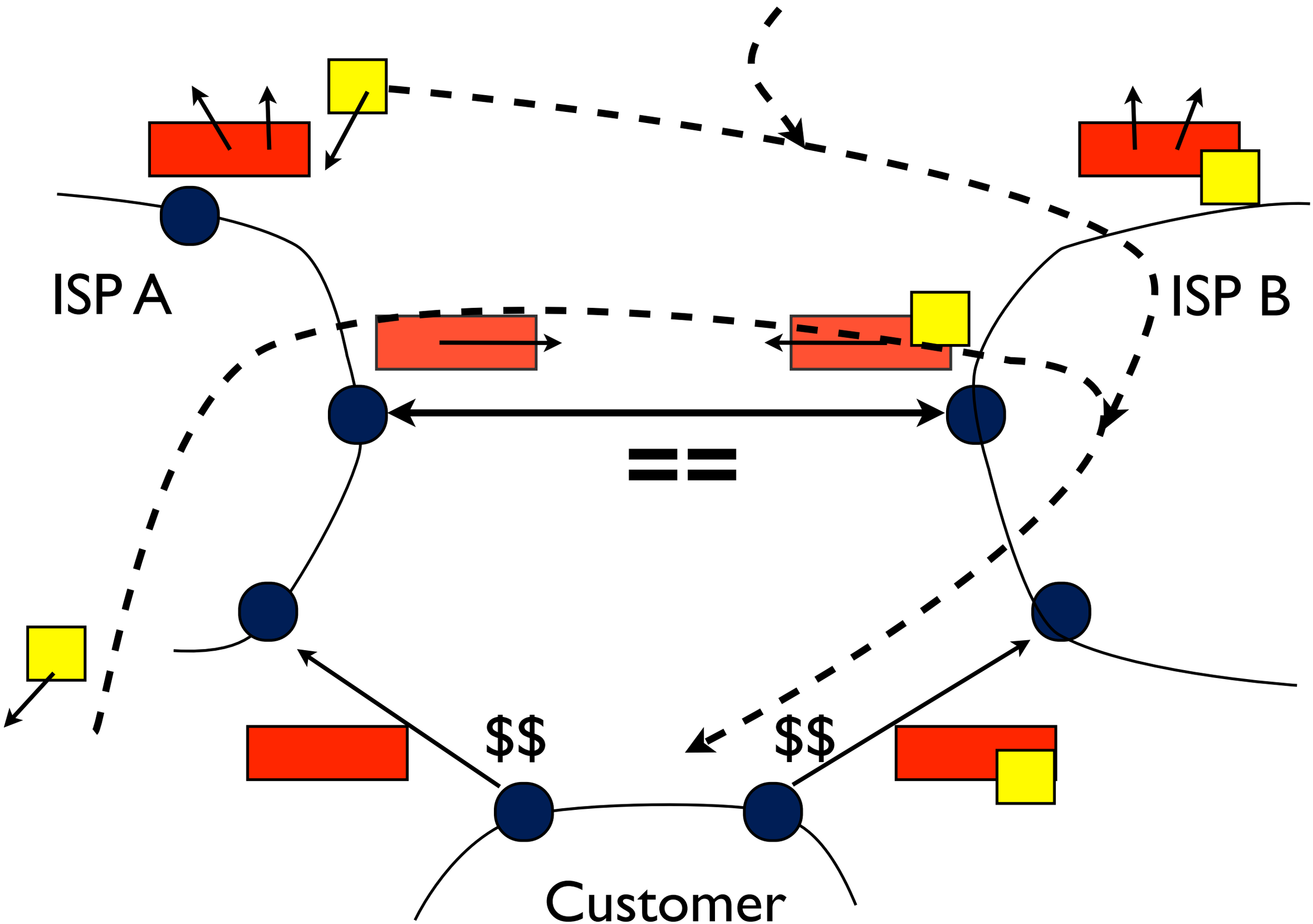
Inbound TE, selective advertisement of a more specific prefix



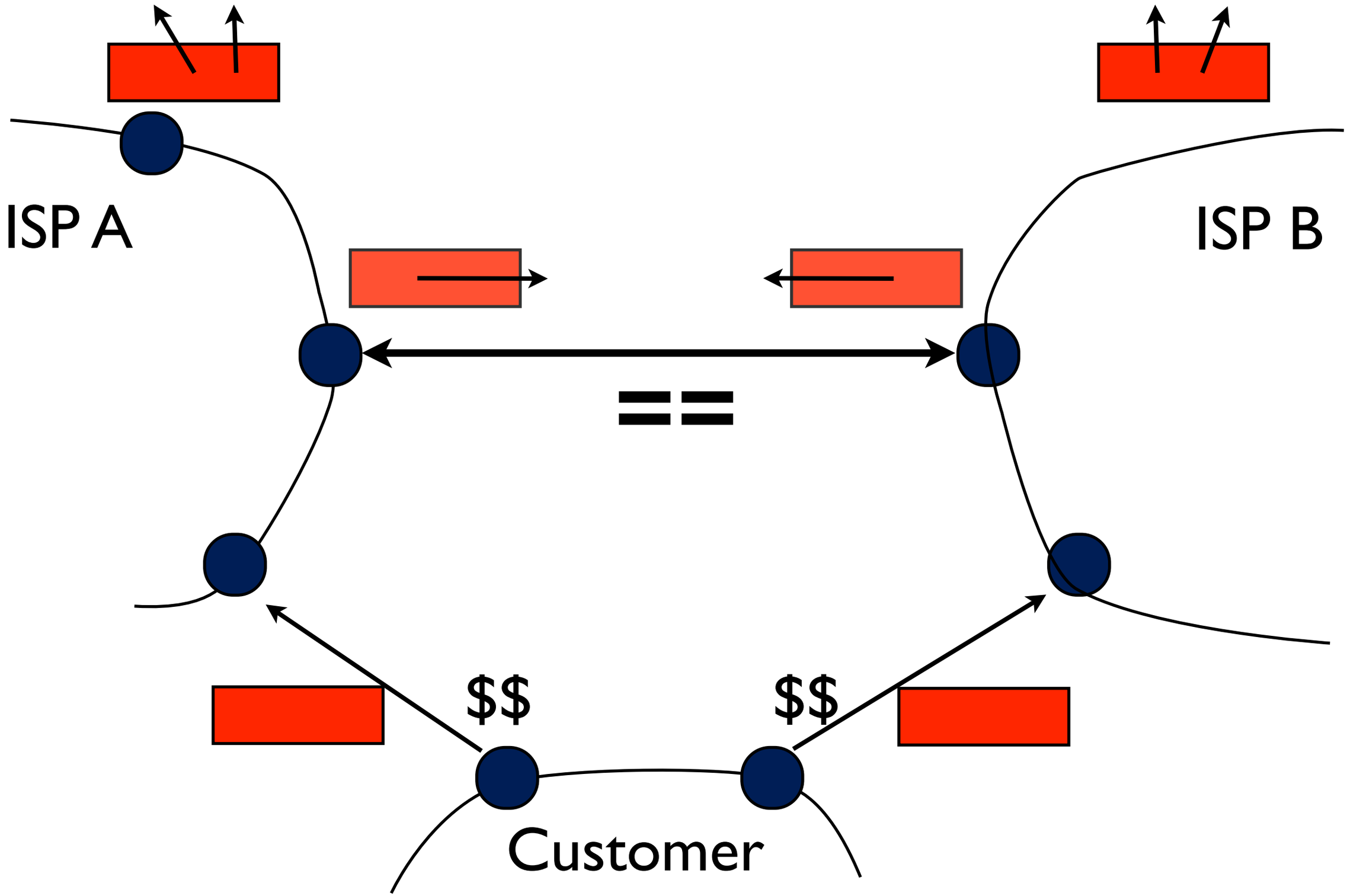
Inbound TE, selective advertisement of a more specific prefix



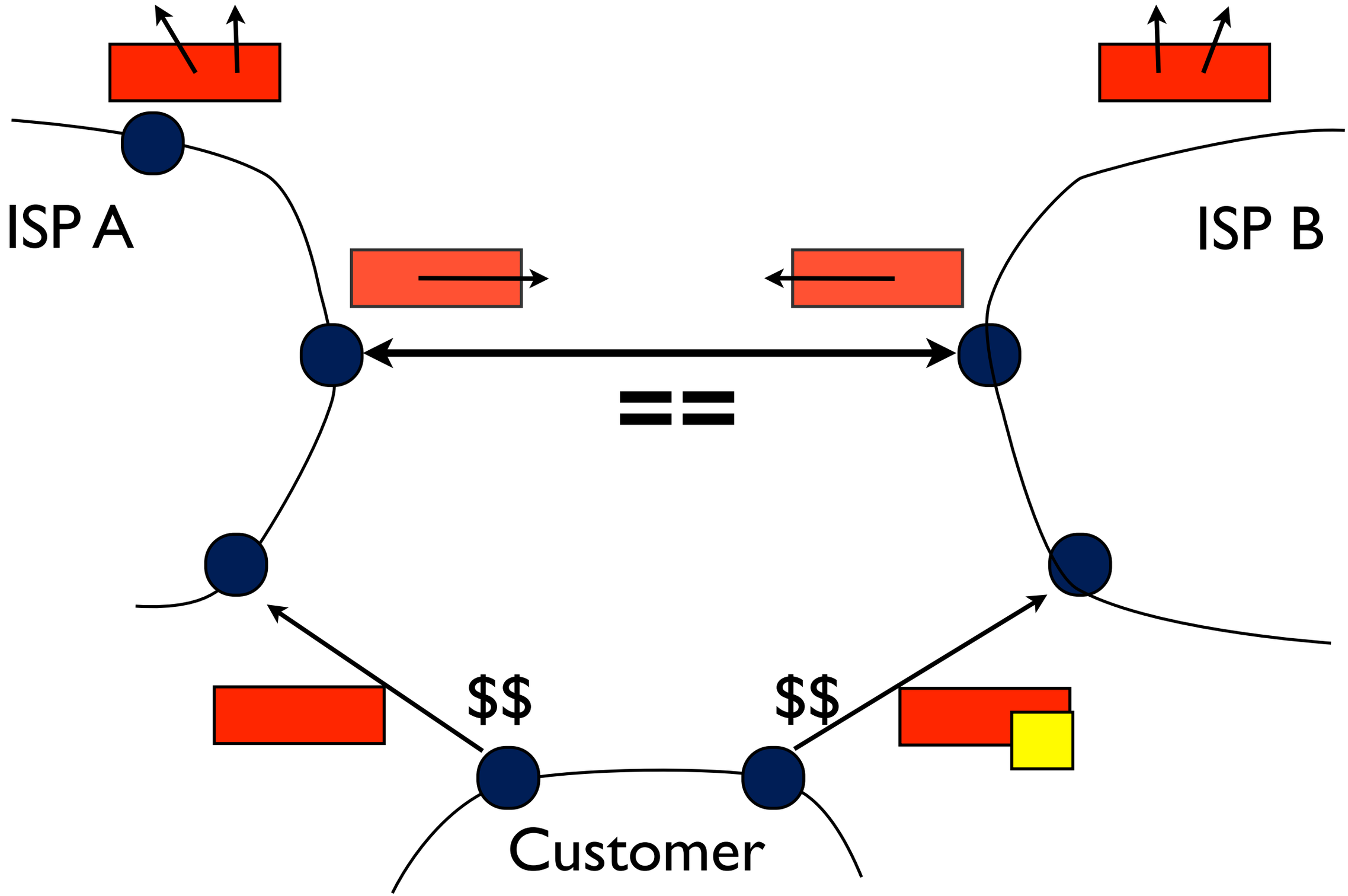
Inbound TE, selective advertisement of a more specific prefix



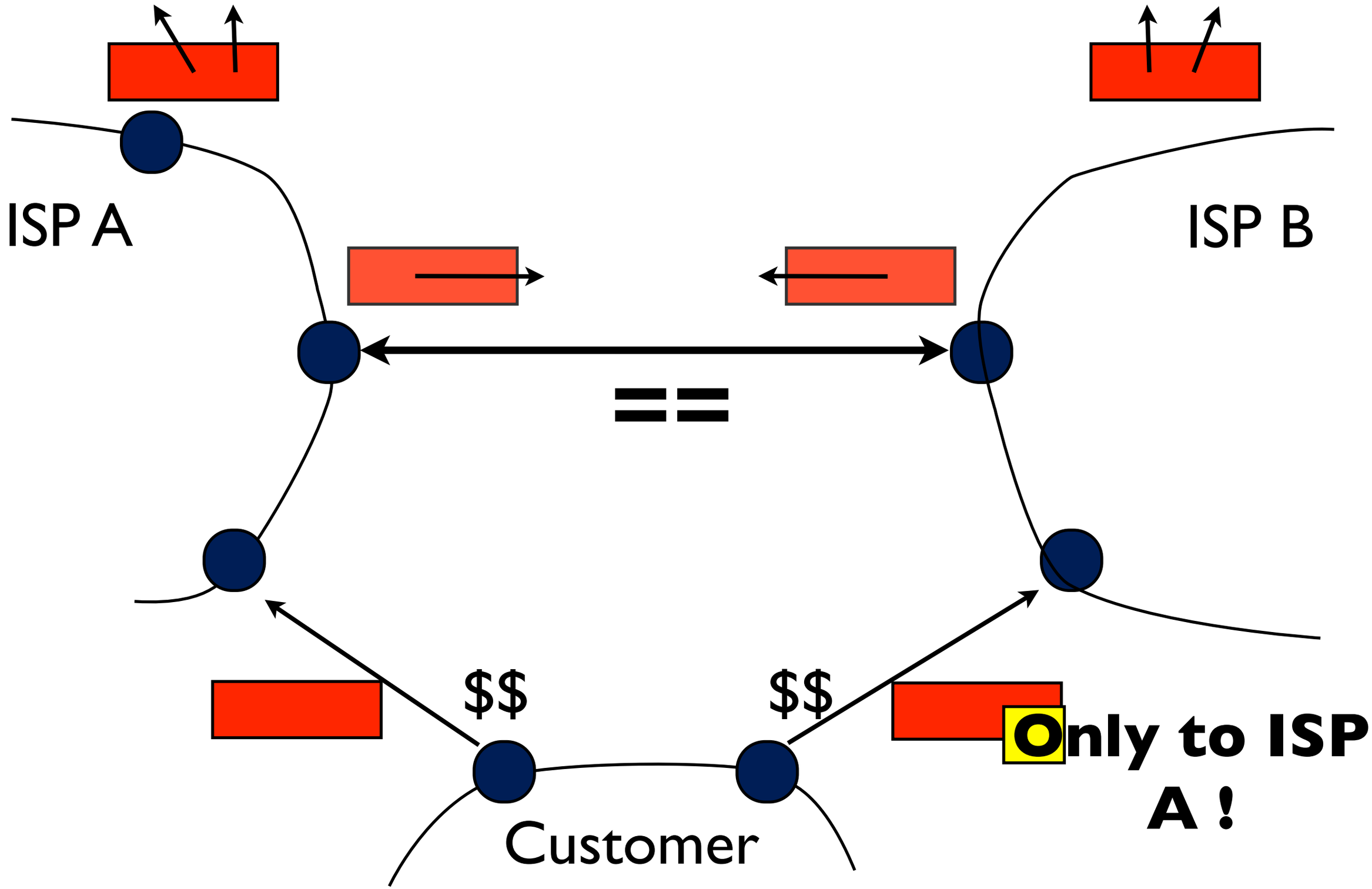
Scope the advertisement of the more specific



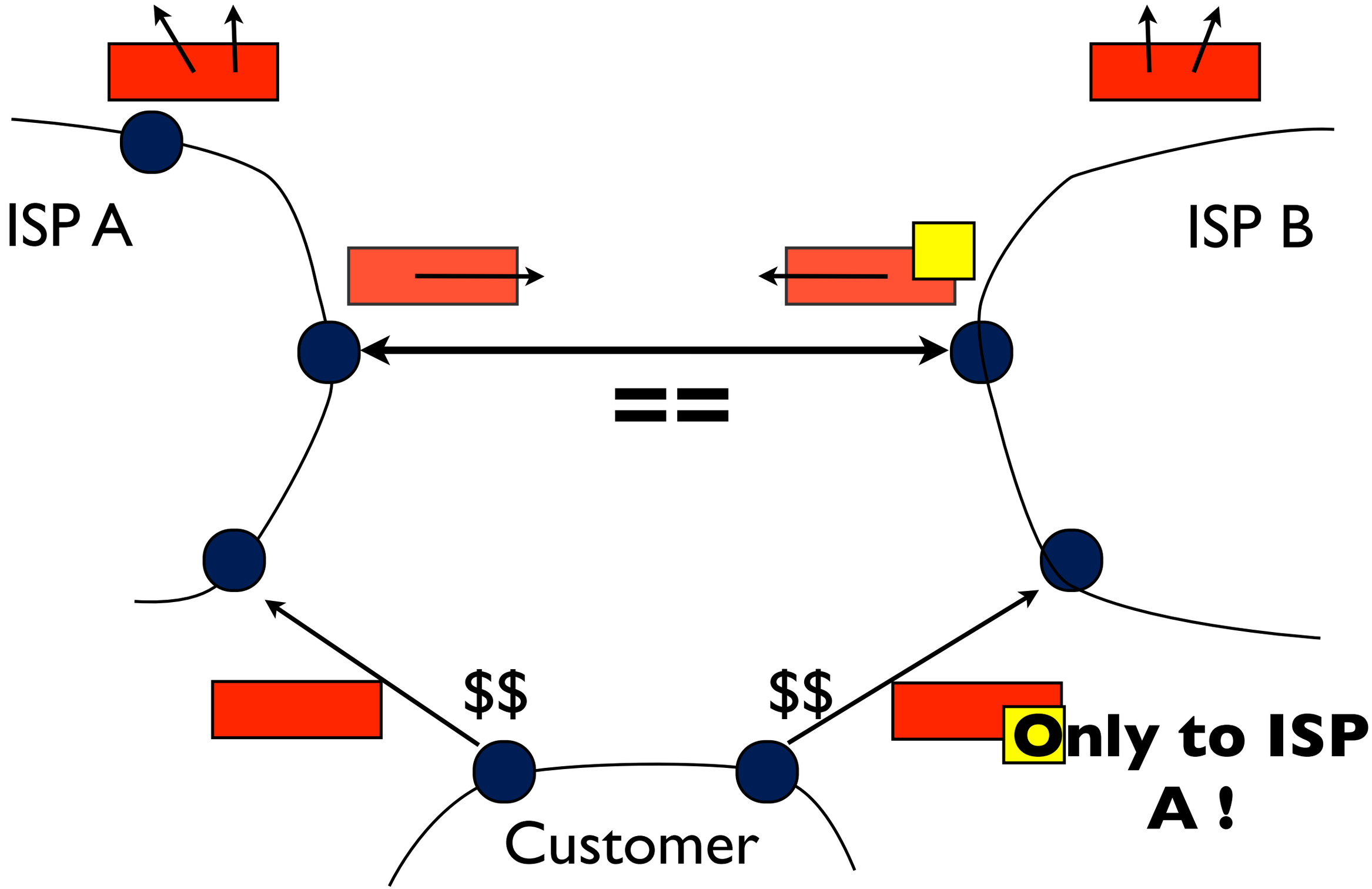
Scope the advertisement of the more specific



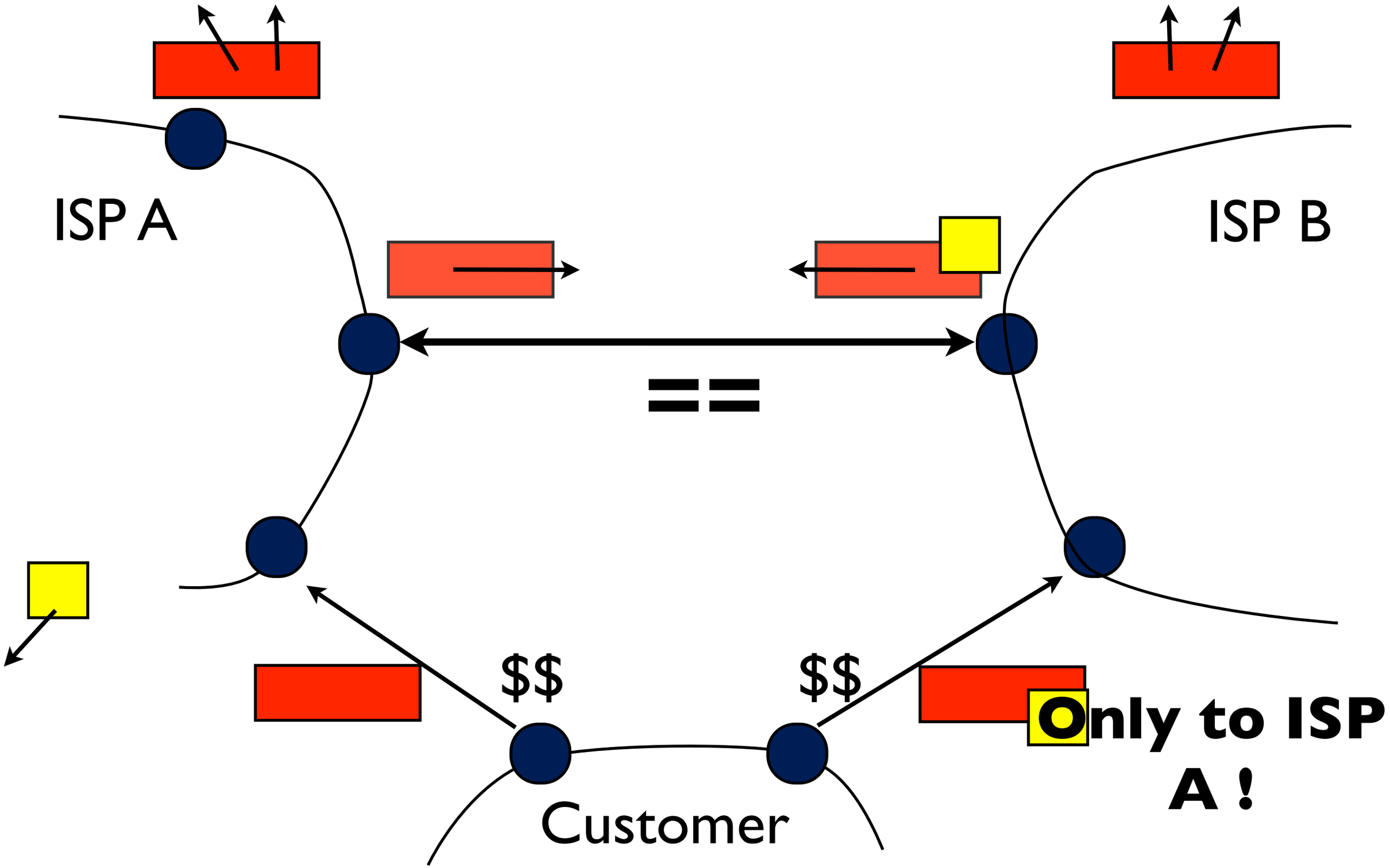
Scope the advertisement of the more specific



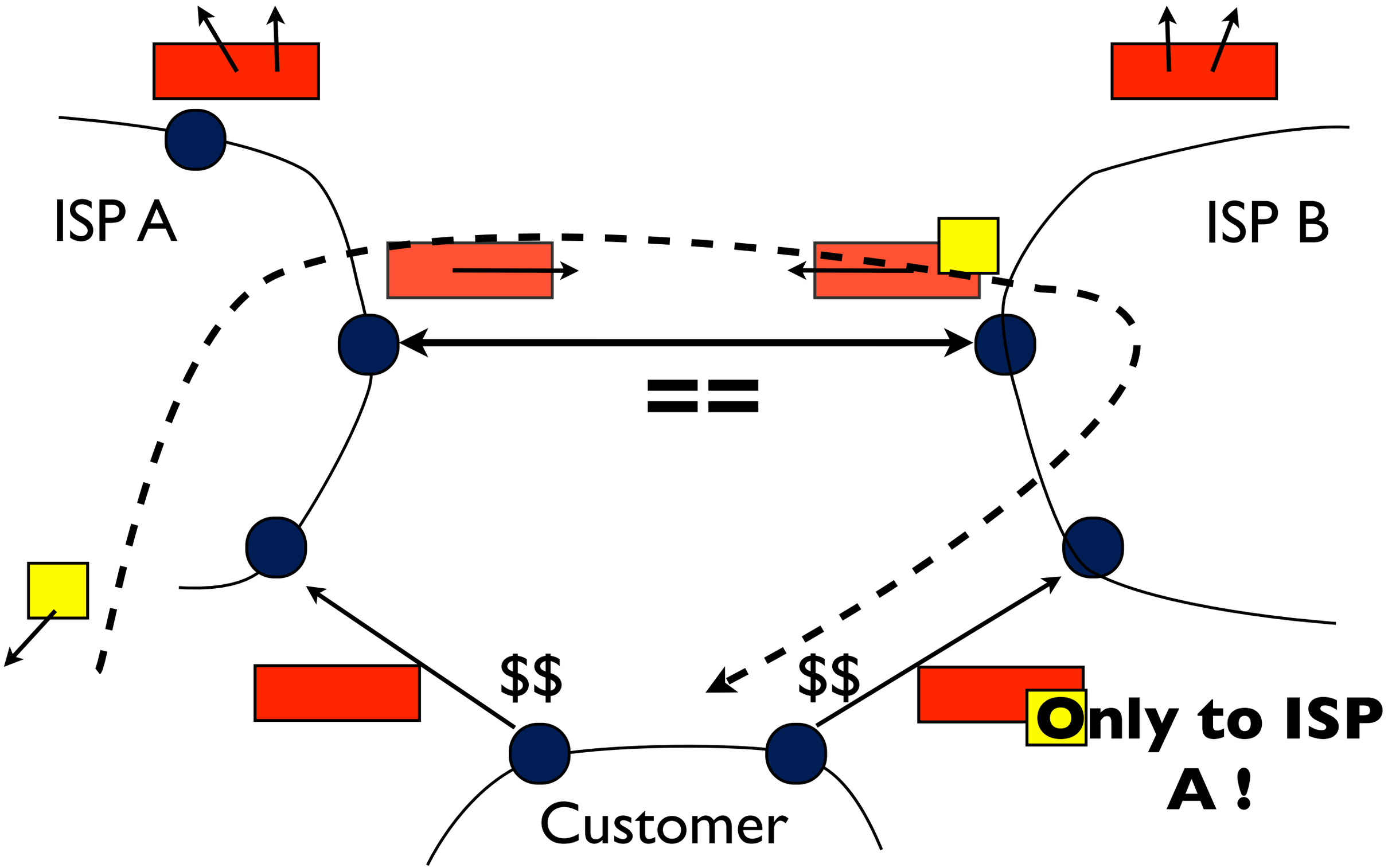
Scope the advertisement of the more specific



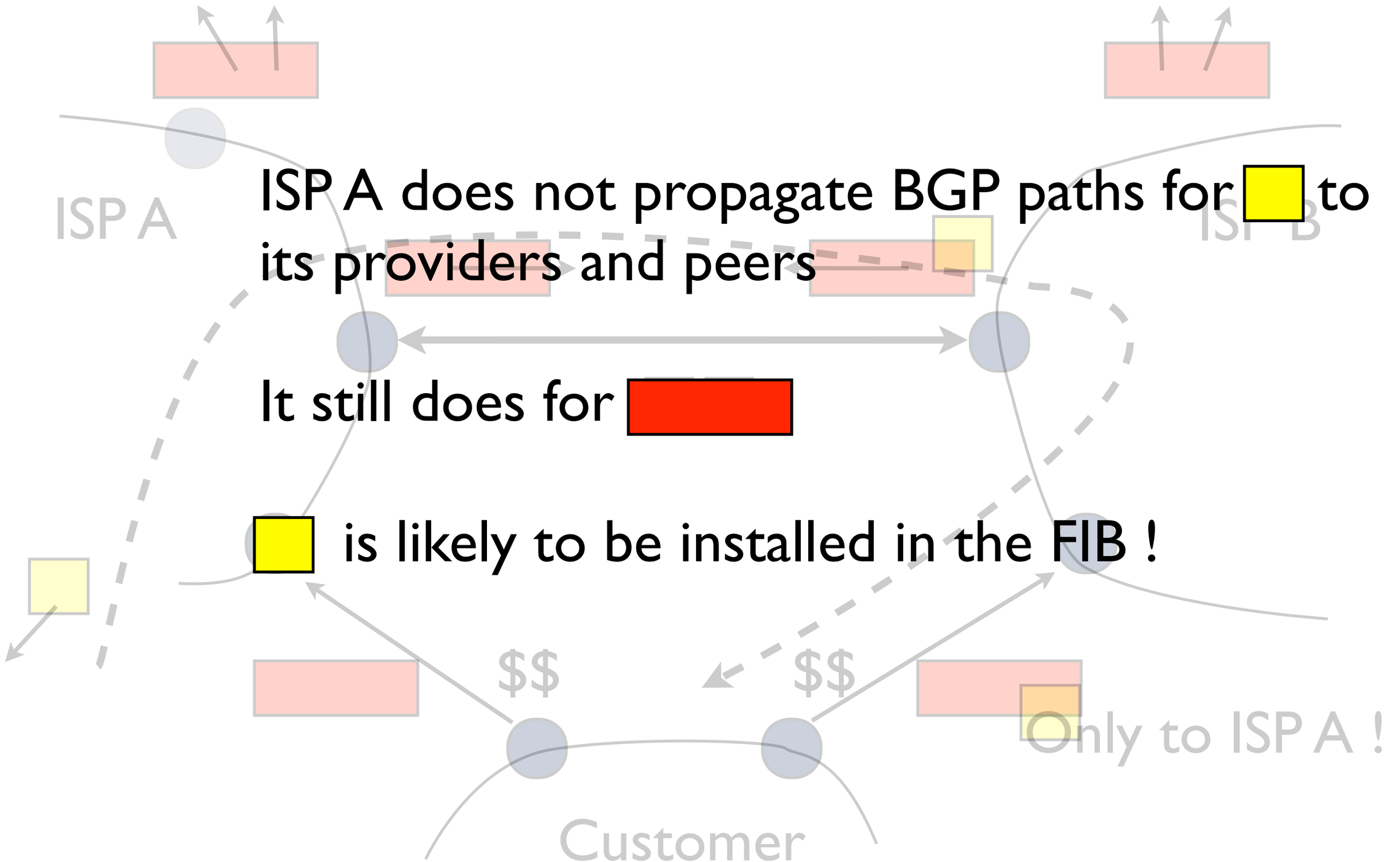
Scope the advertisement of the more specific



Scope the advertisement of the more specific



Let's start playing : Scope advertisement of the more specific

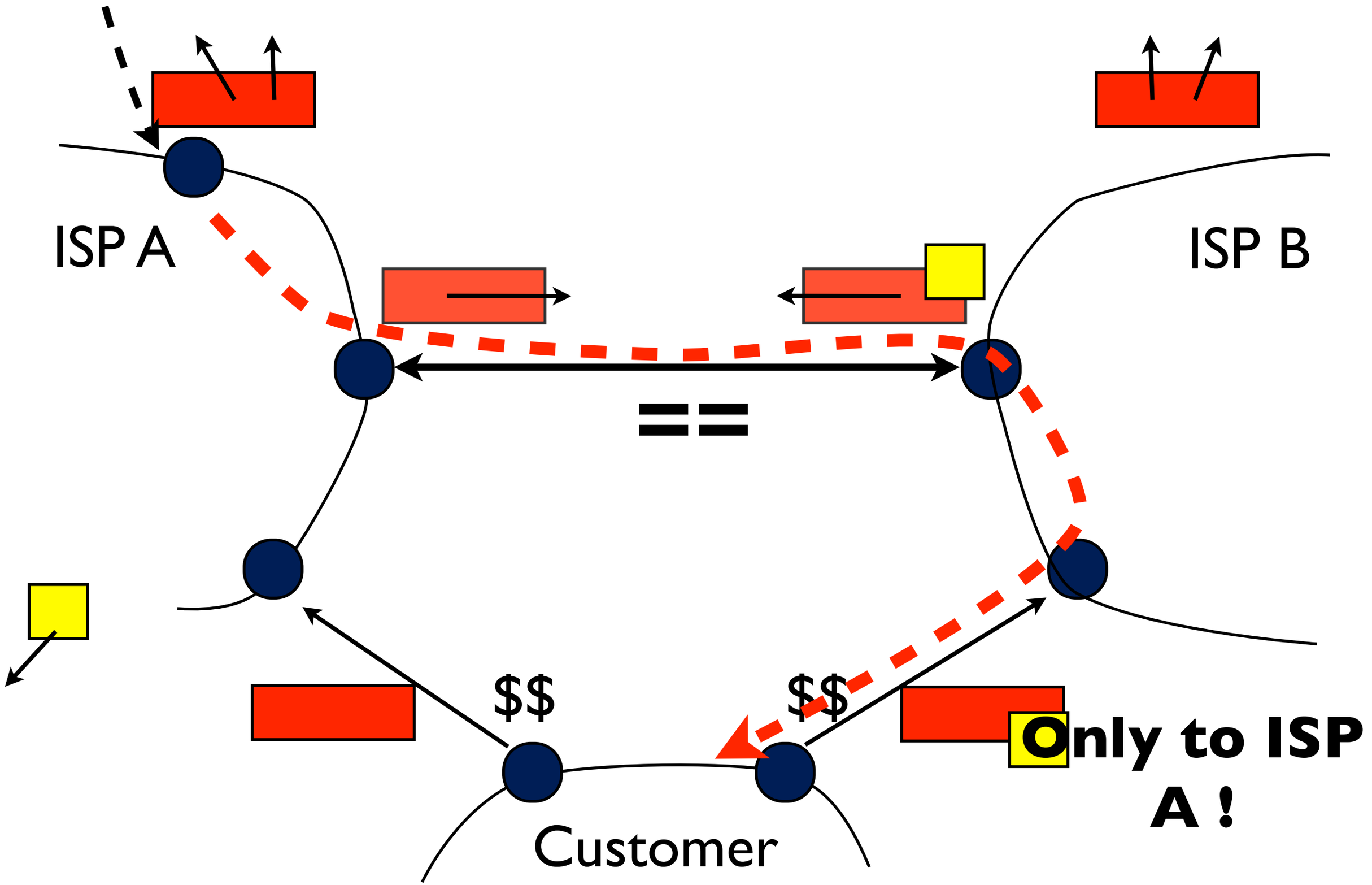


ISP A does not propagate BGP paths for ■ to its providers and peers

It still does for ■

■ is likely to be installed in the FIB !

New path in the network



This is annoying

- Your policies can be violated
- Your flexible routing service can turn **you** into a transit thief when misused by **your** customers
- "Nothing breaks" when the violation takes place
- Ex. : Just consider the Tier-I clique...

So what can you do ?

- Forward differently
- Filter-out / Drop
- Monitor !

Forwarding differently

- Deploy BGP so as to have forwarding at an incoming interface solely based on policy fitting paths
  - Put the Internet in VRFs
  - Careful configuration of import rules
- Complex, Costly

Filtering out / Drop

- Drop packets, at ingress, for routes that are not supposed to be served there
  - Assume malicious behavior by default
  - Interrupts service from/to customers
- Filter out, at egress
  - Range served as if the msp did not exist

Monitor

- You got the means to monitor ingress-egress traffic demand to run your business, right ?
- “Just” check if counters for non-policy compliant transit
  - Pick the phone when counters are not at 0
  - Filter-out if the issue is not getting fixed early enough
- Seems like few operators run the check

PMACCT

- Tool developed by Paolo Lucente
  (See talk at RIPE 61 plenary)
- Policy violation check is a matter of a couple of lines

  http://wiki.pmacct.net/DetectingRoutingViolations
- Tools integrating with pmacct can benefit from this work
  (ie. Cariden)

Ignoring more specifics In the blind ?

Ignoring more specifics ?

- It is frustrating to forward traffic to a transit provider...
- ...when a less specific covering the destination is known via a peer
- See talk by Fredy Künzler (INIT7) RIPE 63

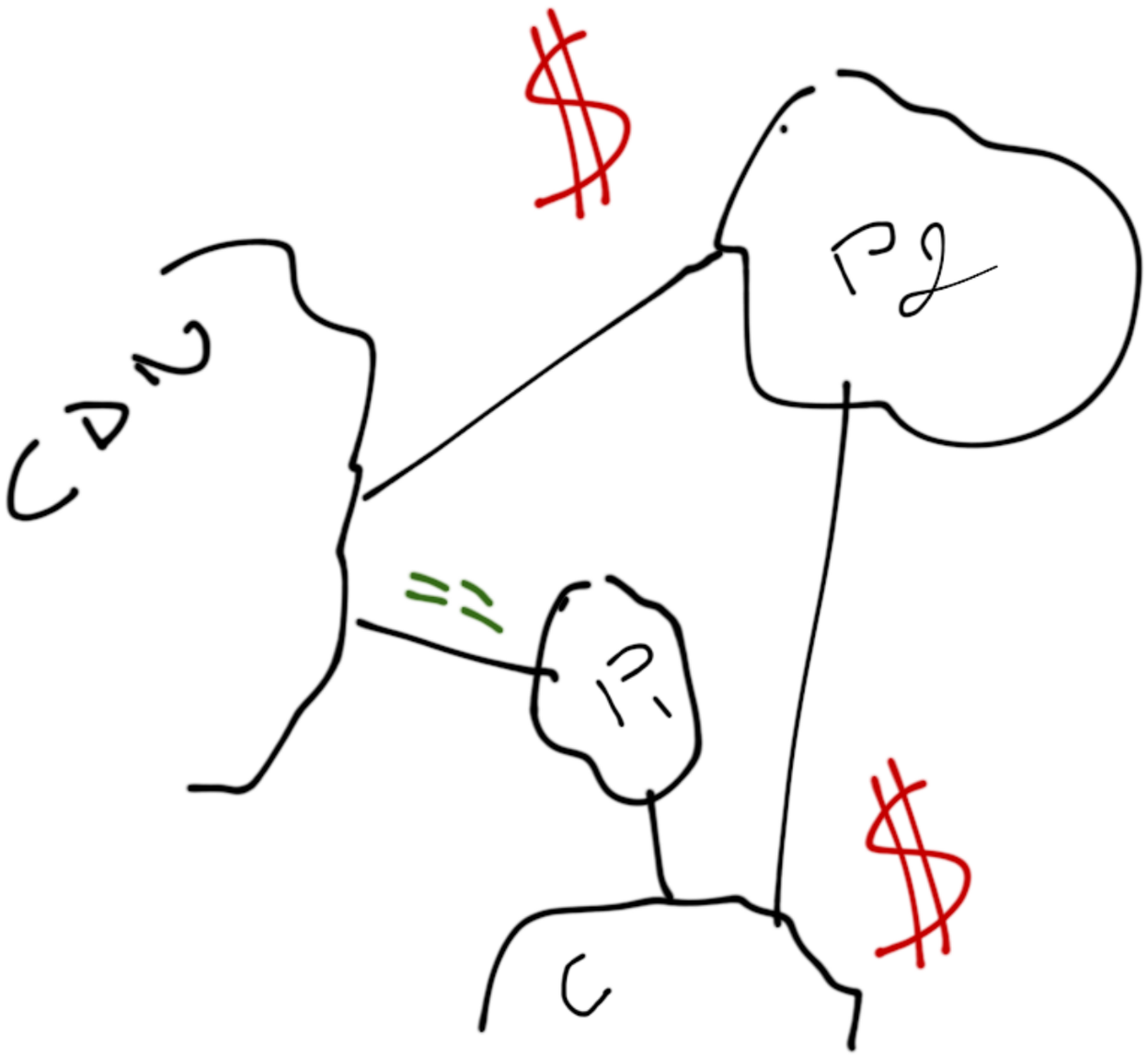
Ignore them

- Filter out the announcement of the more specific prefix received over the eBGP session with the transit provider

Context

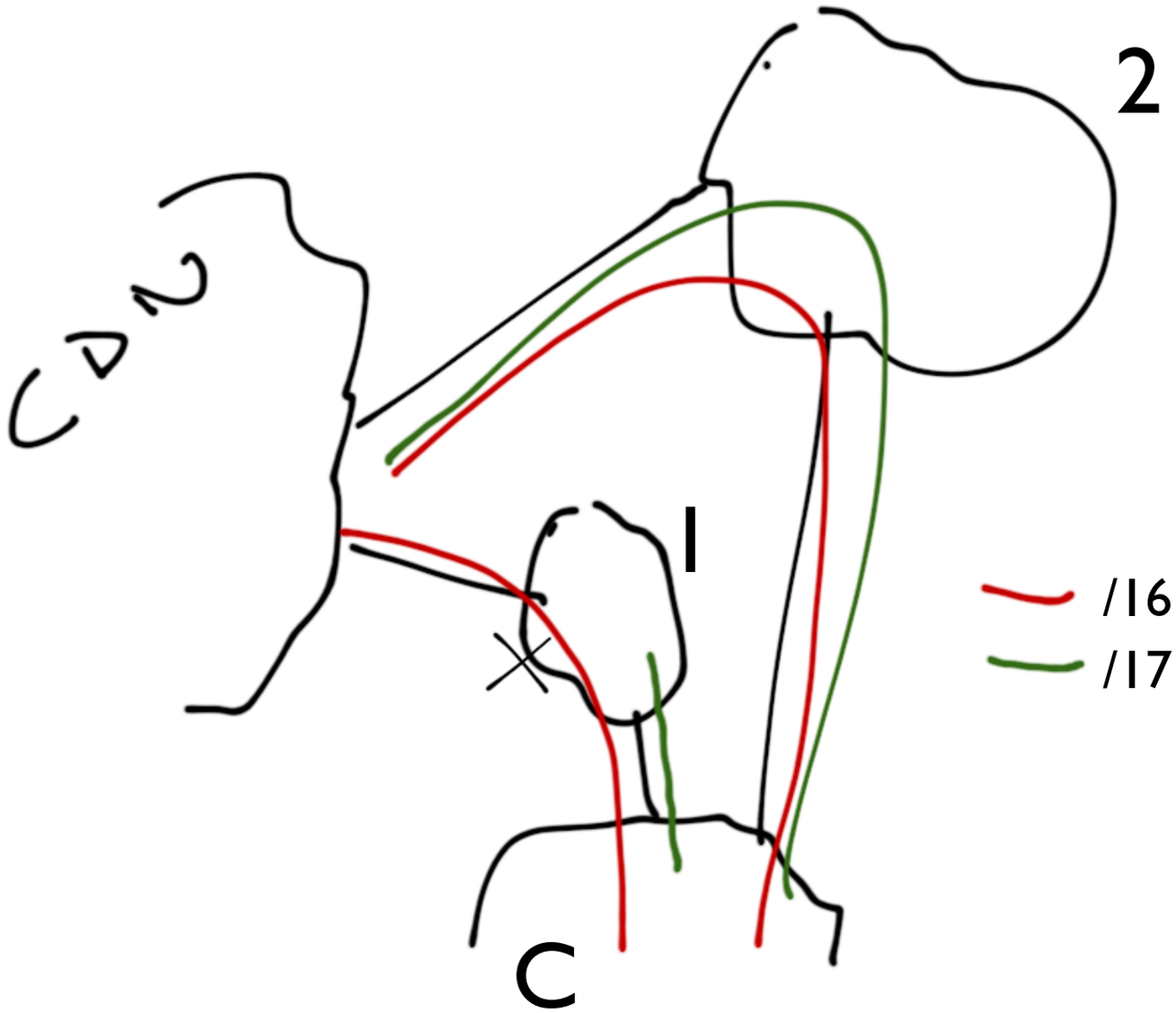
- What are the reasons for a CDN to receive more specific prefixes from providers, only ?

Reference context 1



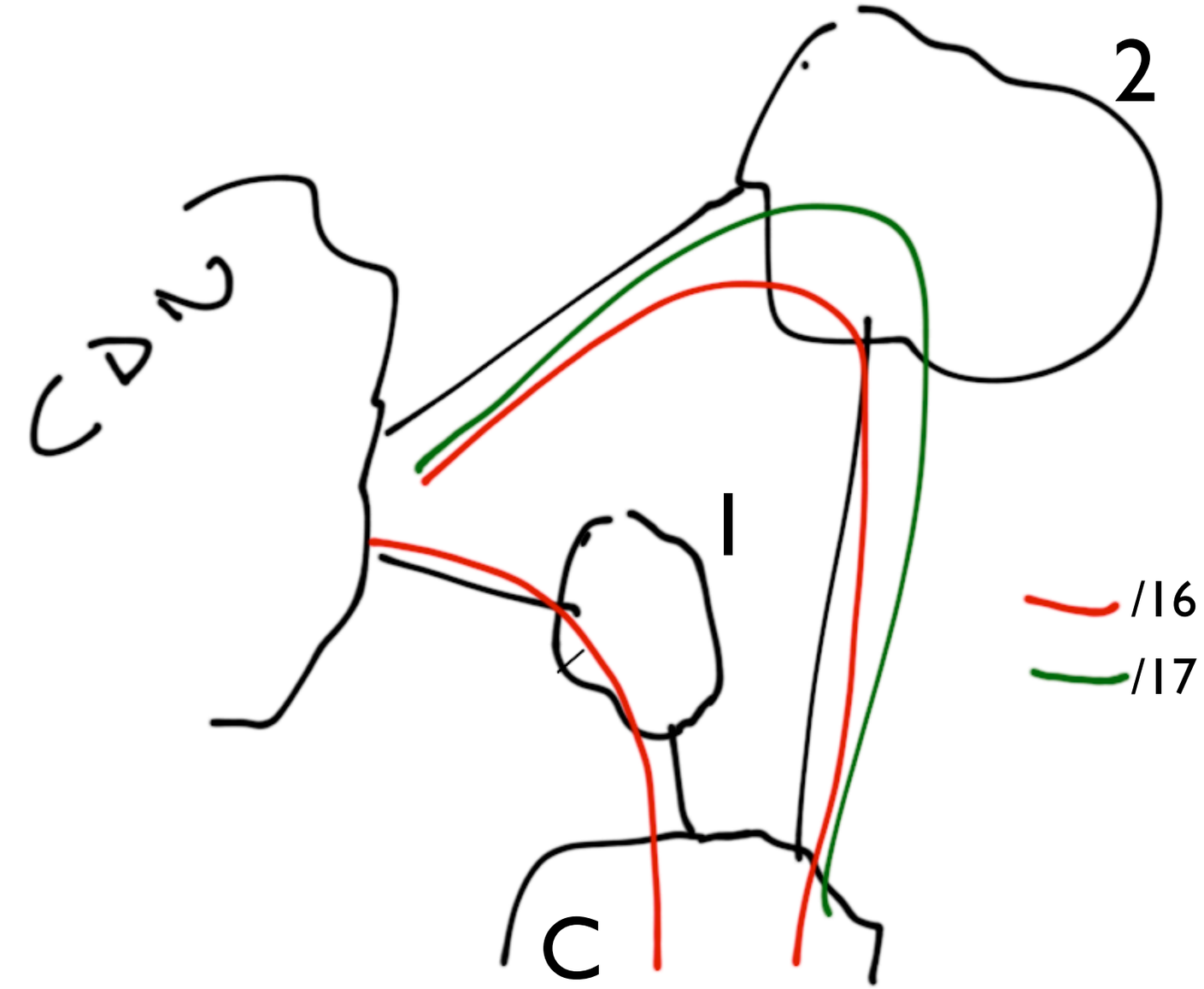
- Destination Eyeball ISP C
- C in customer base of Peer P1
- C in customer base of Provider P2

Case 1
No export



- C tags NO_EXPORT when advertising the more specific to peer P1
  - C does not want the entire incoming traffic shares for the /17 to be delivered by P1
  - C gives traffic shares to P1 only for the single homed customers of P1. **C Expects to receive the rest from P2**
- Do you want to bypass C ?

Case II
Selective advertisement

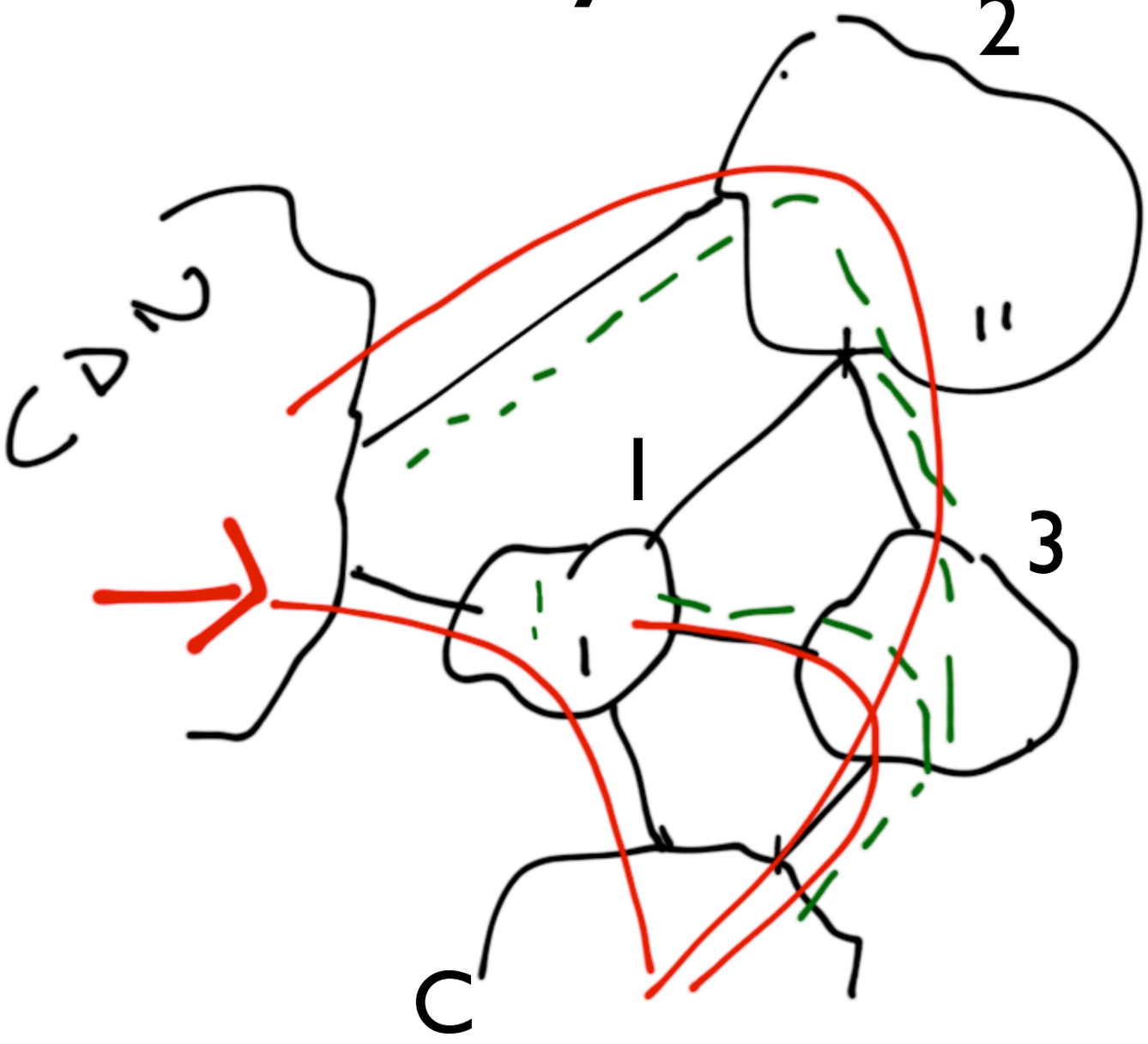


- C does not advertise the /17 to P1
  - C does not want to allow the incoming traffic shares for the /17 to be delivered by P1
  - P1 is only allowed to deliver its own customer traffic for C through its own transit/peering towards C
- Do you want to bypass C ?

Impact of bypassing more specifics

- Disrespect of your peers' customers traffic engineering requirements/needs
- Policy violations ?

Case III
Policy violation at peer



- C does not advertise the /17 to P1
- Only to P3
- P3 and P1 are peers
- If you ignore the transit path, you violate P1's policy doing CDN-P1-P3

Marketing

- You act against your neighborhood
- What is the cost of a public announcement

  "*These CDN guys are the ones making the money, they force me to peer instead of paying me, and now, on top of that, they make moves to get free transit through my network ???*"

Take away

- Ignoring more specifics can do you good
  - With a cost for your peers, and peers customers
    - Reduction of their traffic engineering capabilities
      - Disrespect of their explicit requirements expressed with selective advertisement of more specific prefixes
  - With a risk of policy violation at your peers
- Should not be done blindly

Ignoring MSP
Making it an habit

- Requires external knowledge
  - BGP state deep down the hierarchy
  - Business
- Do that
  - with careful considerations about your BGP neighborhood,
  - on prefixes tracking relevant amounts of traffic