Communication Networks

Spring 2022

Laurent Vanbever

nsg.ee.ethz.ch

ETH Zürich (D-ITET)

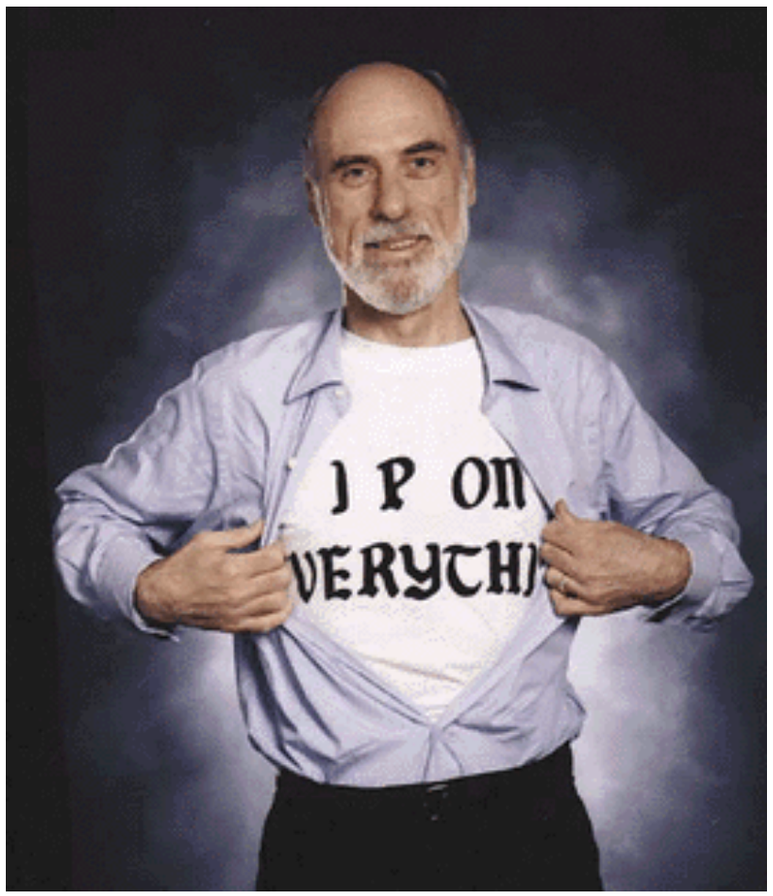
April 4 2022

Materials inspired from Scott Shenker & Jennifer Rexford

Last week on

Communication Networks

Internet Protocol and Forwarding



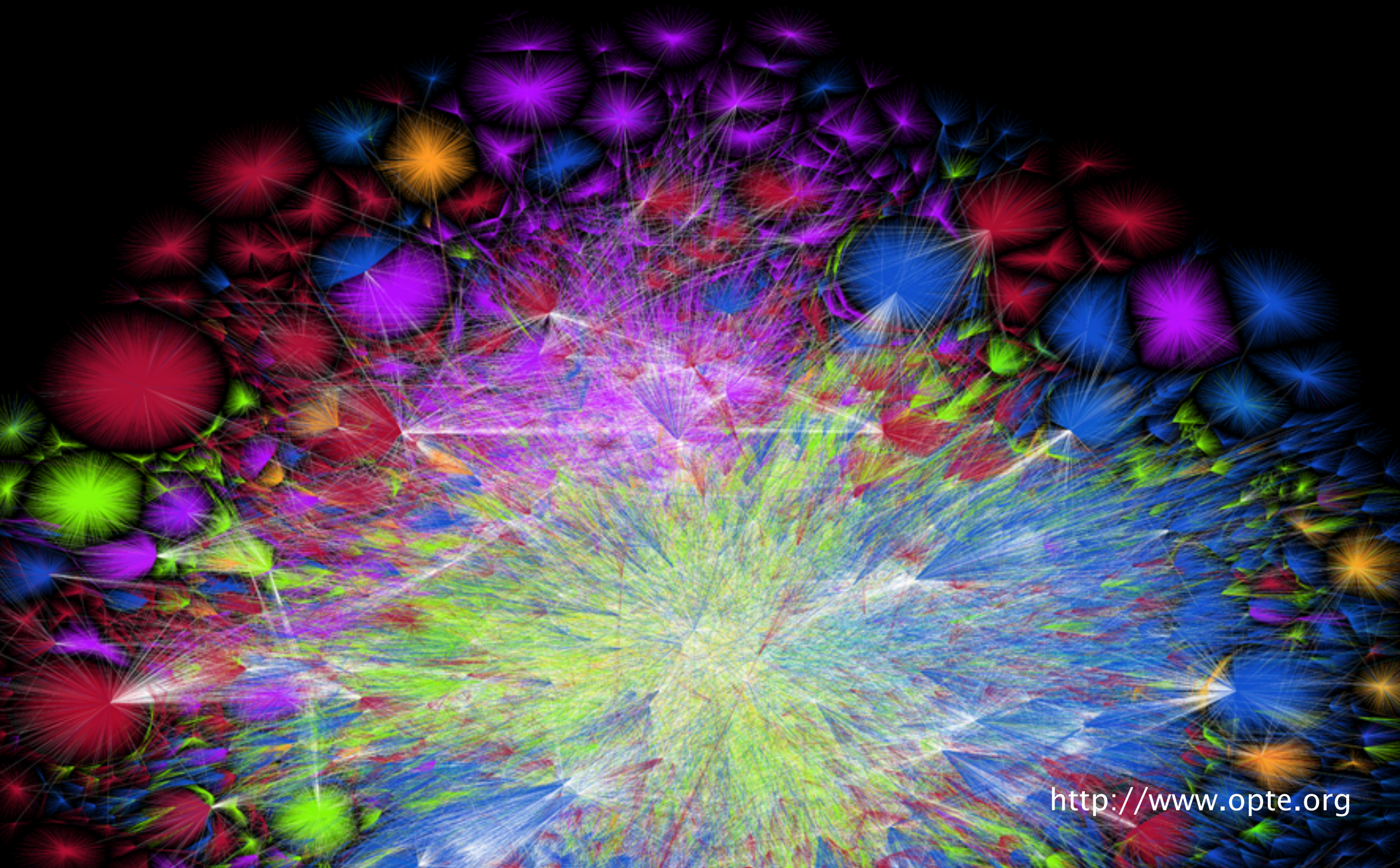
source: Boardwatch Magazine

1 **IP addresses**
use, structure, allocation

2 **IP forwarding**
longest prefix match rule

3 **IP header**
IPv4 and IPv6, wire format

Internet routing



Internet routing

from here to there, and back

1 Intra-domain routing

Link-state protocols

Distance-vector protocols

2 Inter-domain routing

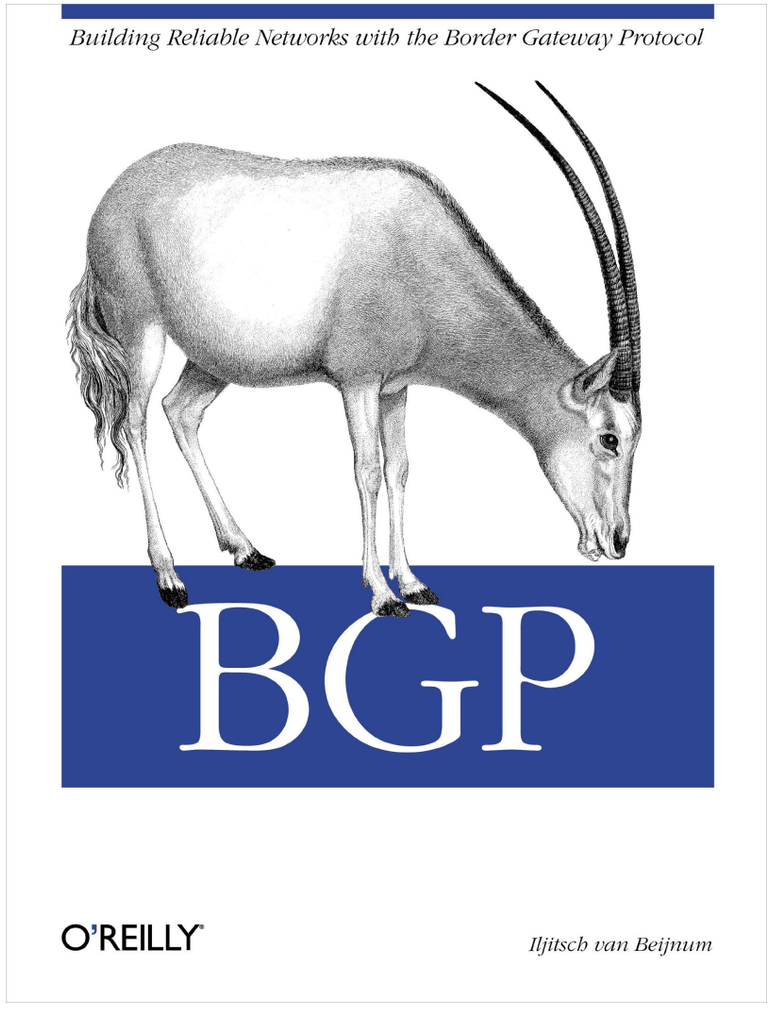
Path-vector protocols

This week on

Communication Networks

Border Gateway Protocol

policies and more



1 **Protocol**

How it works

2 **Policies**

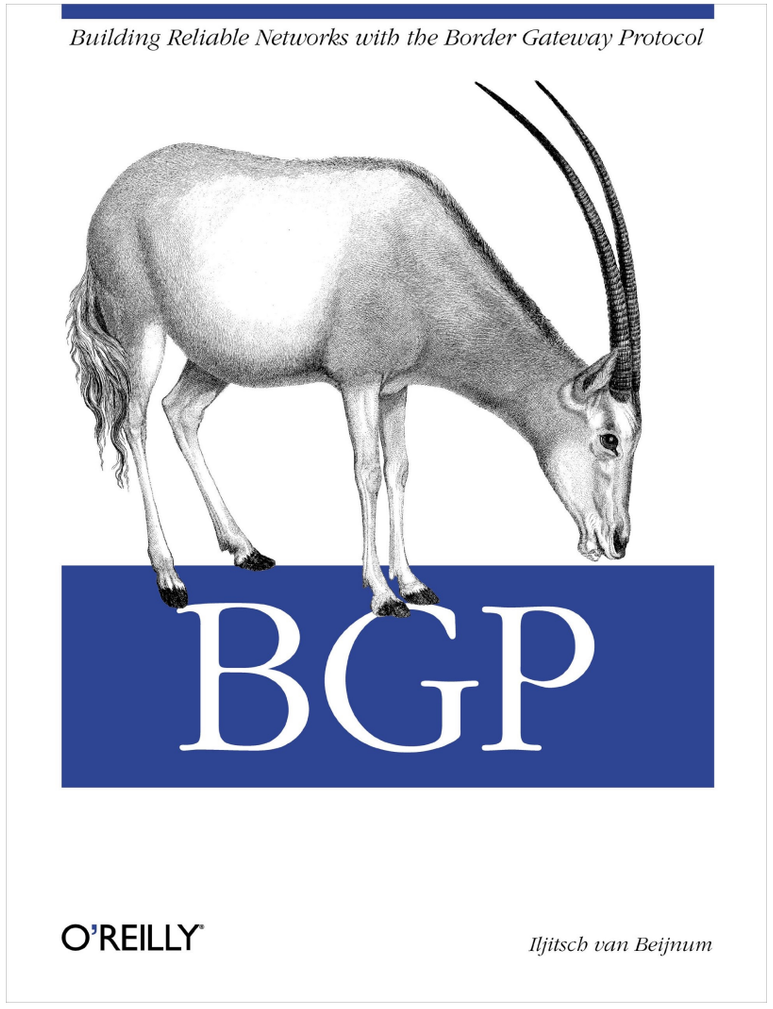
"Follow the money"

3 **Problems**

Security, performance, …

Border Gateway Protocol

policies and more



1 Protocol

How it works

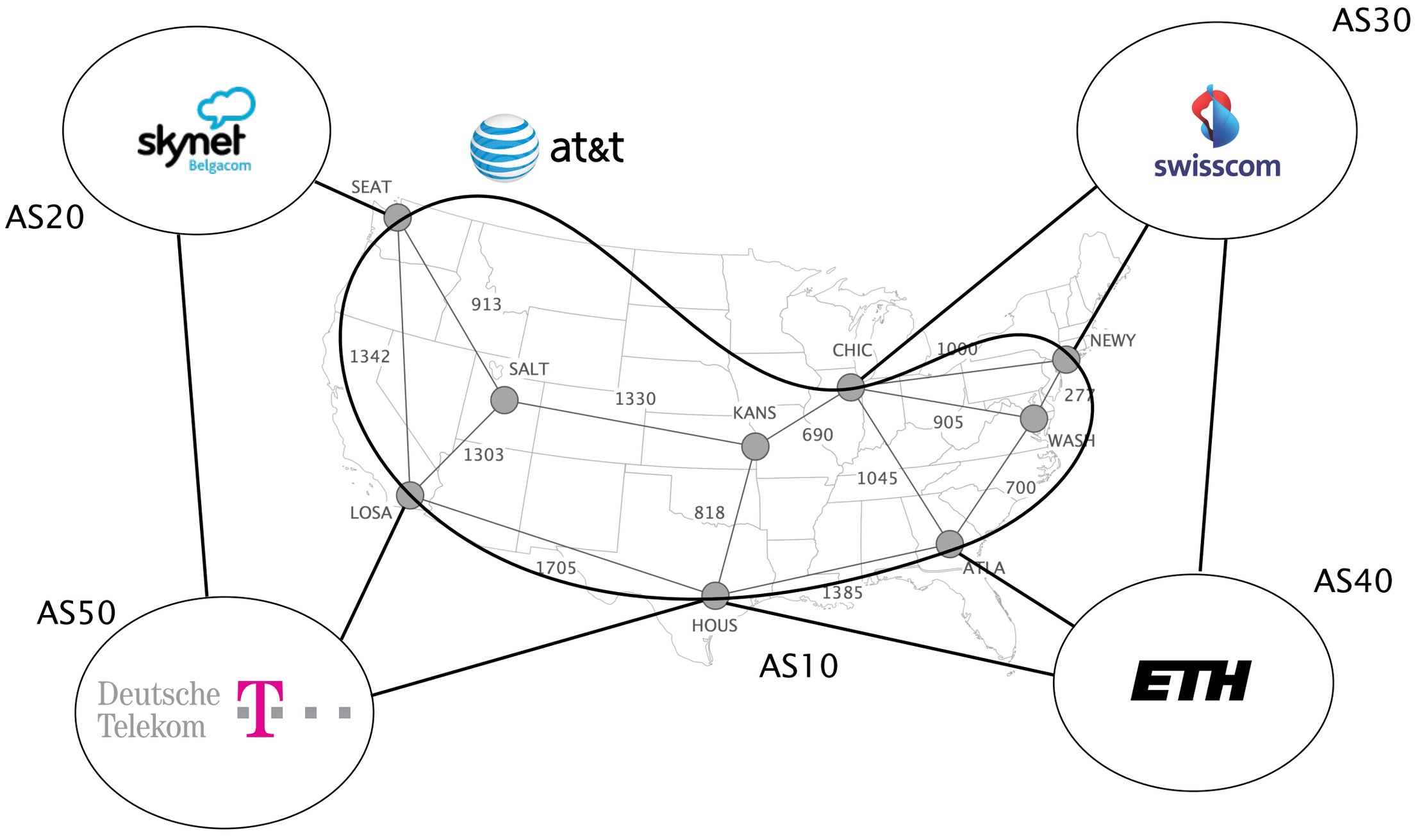
Policies

"Follow the money"

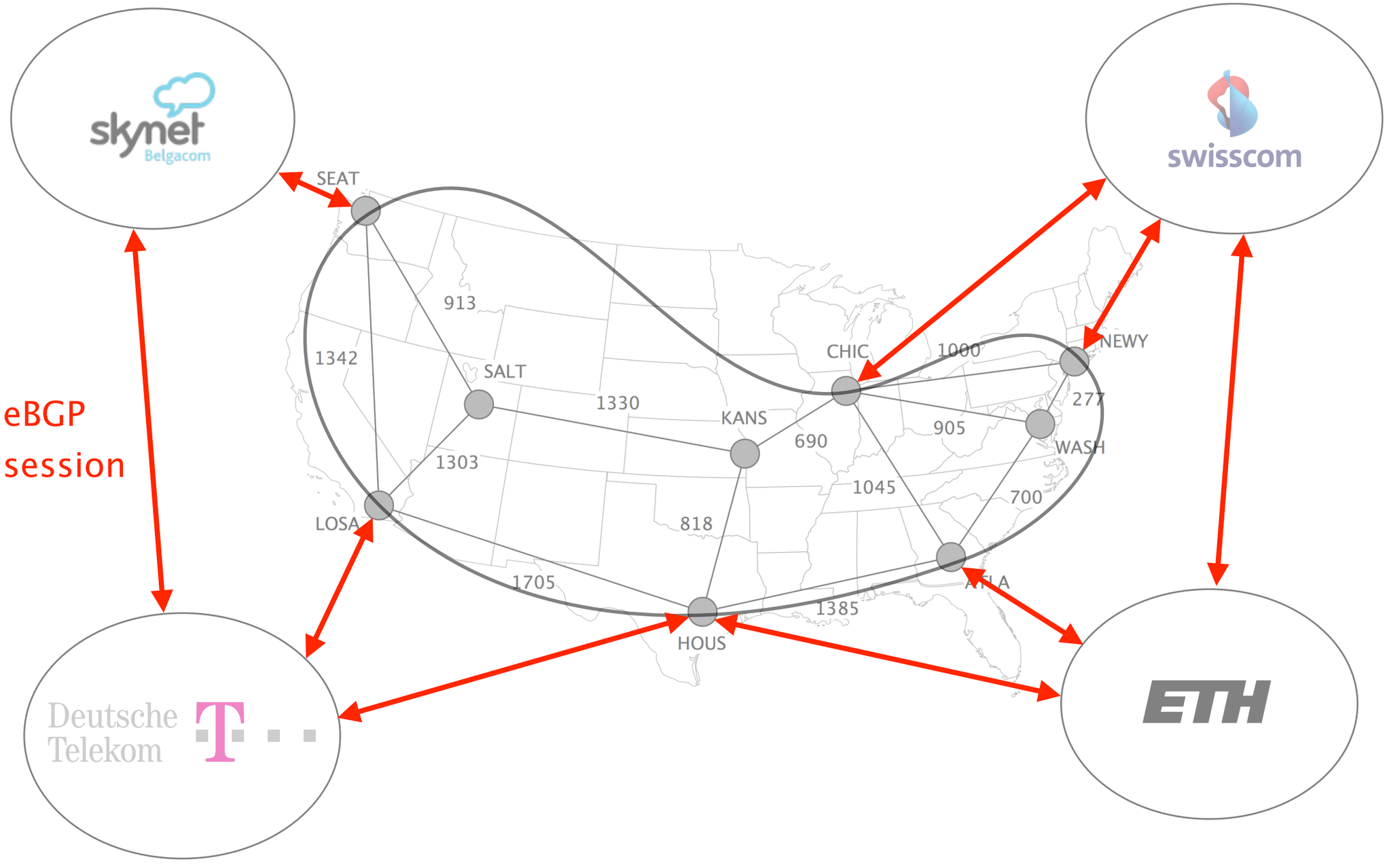
Problems

Security, performance, …

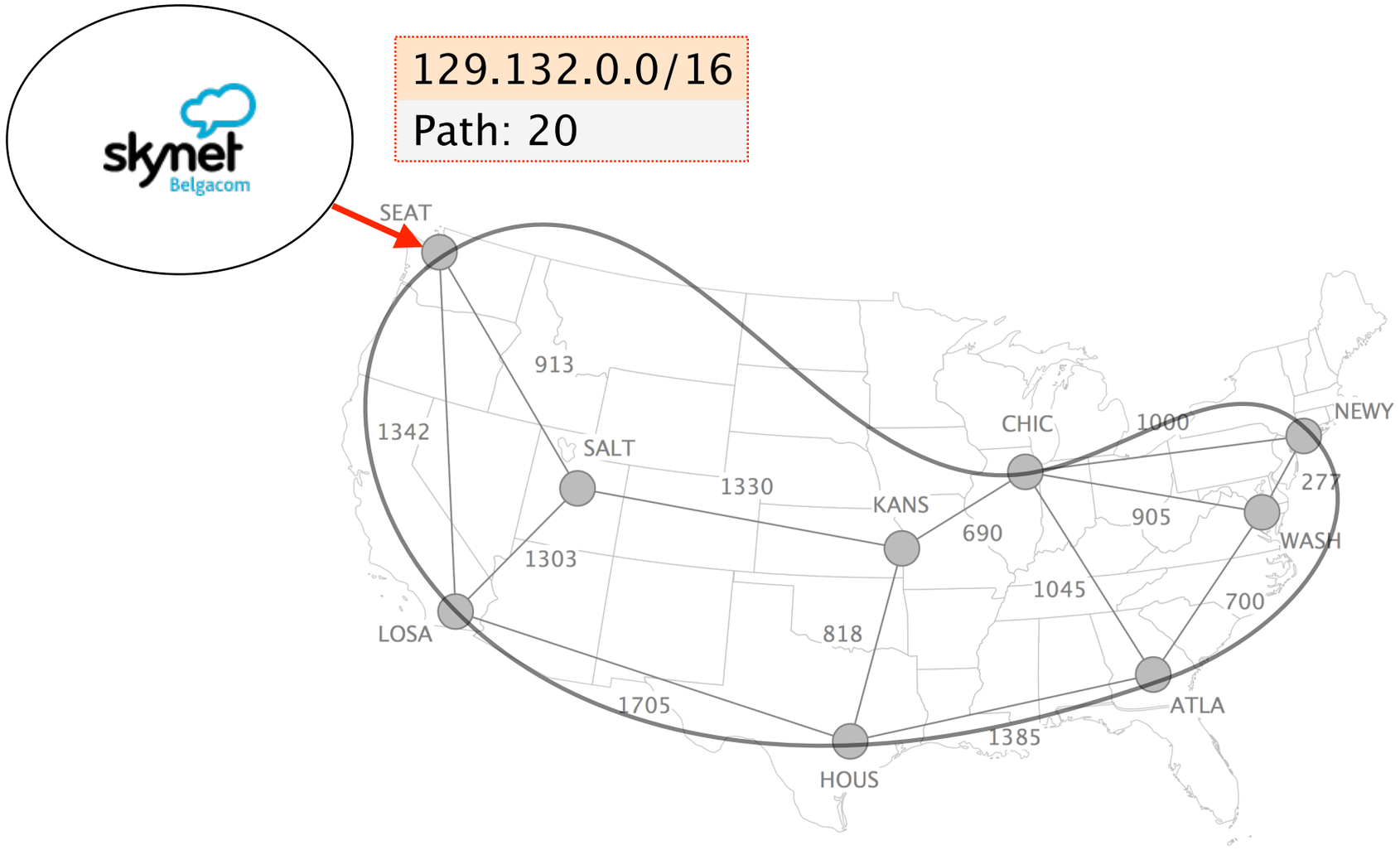
BGP sessions come in two flavors



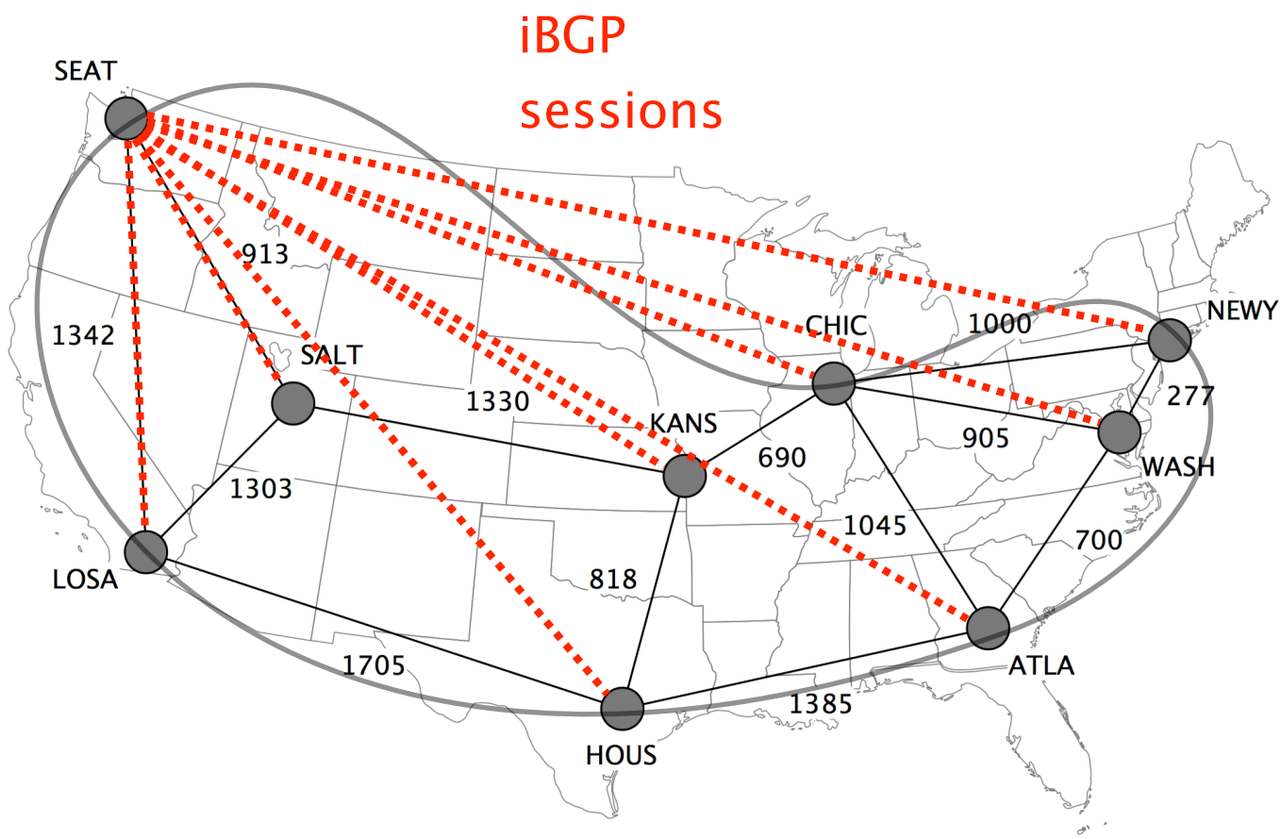
external BGP (eBGP) sessions
connect border routers in different ASes



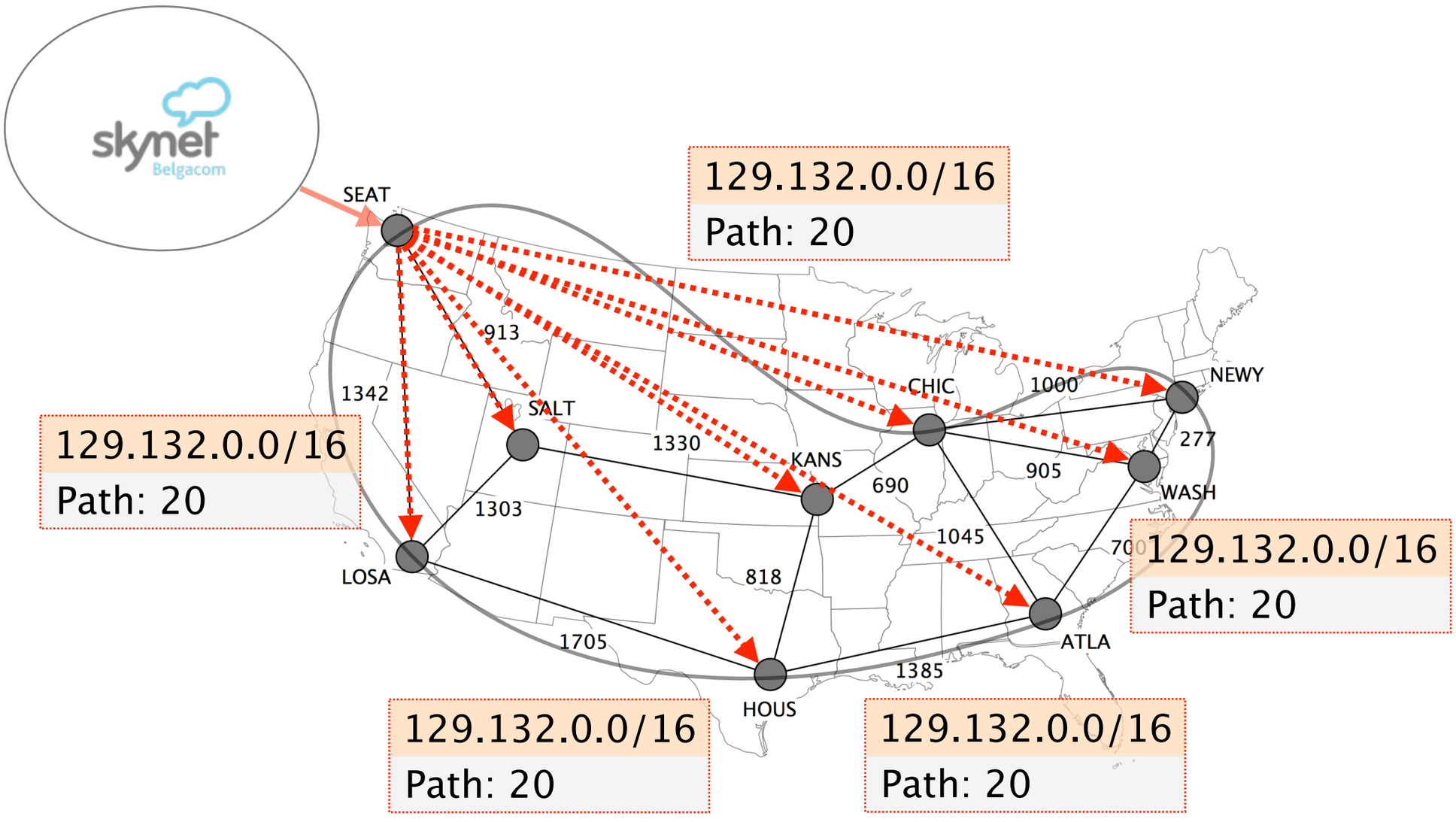
eBGP sessions are used to learn routes to external destinations

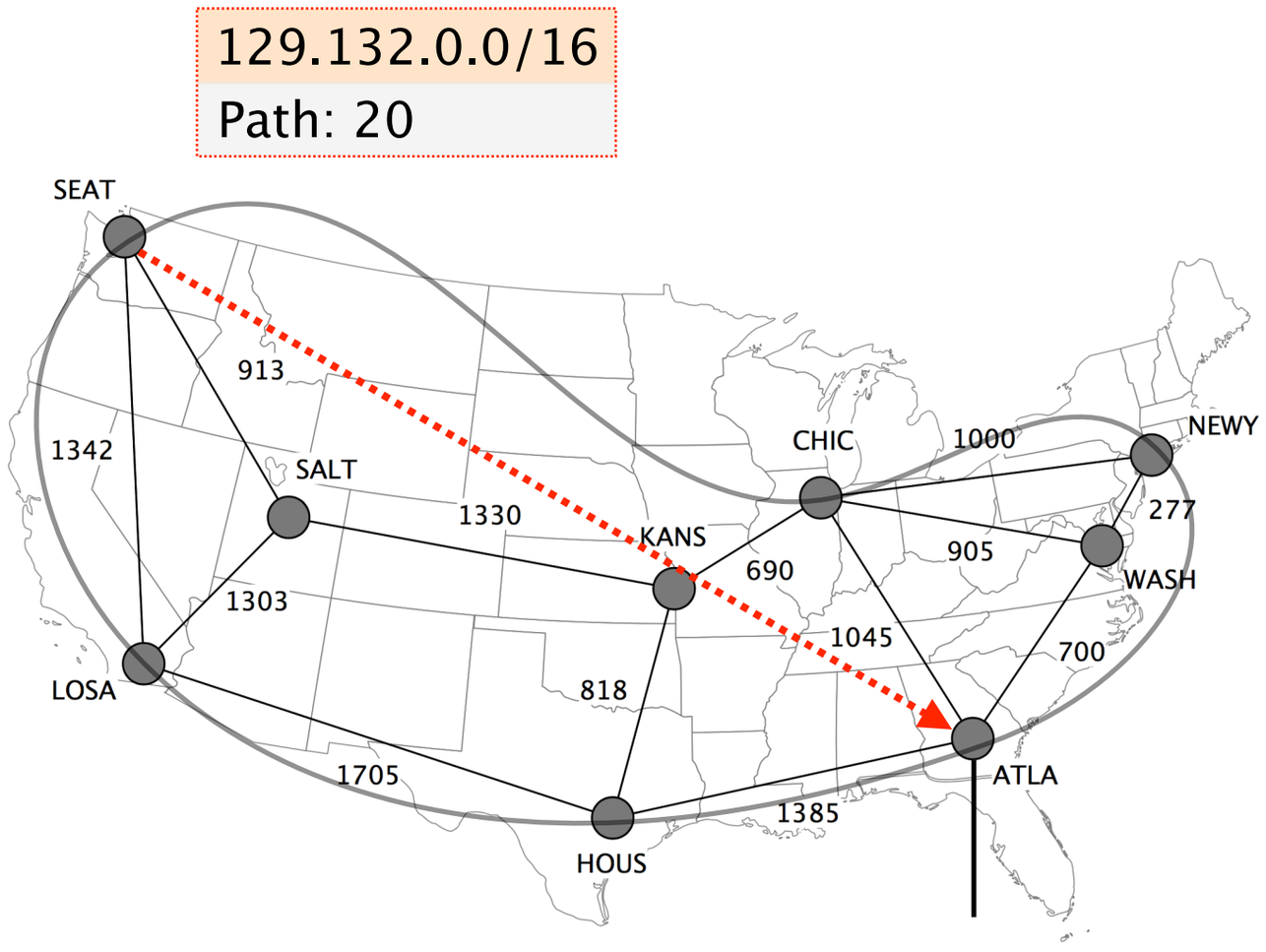


internal BGP (iBGP) sessions connect the routers in the same AS



iBGP sessions are used to disseminate externally-learned routes internally

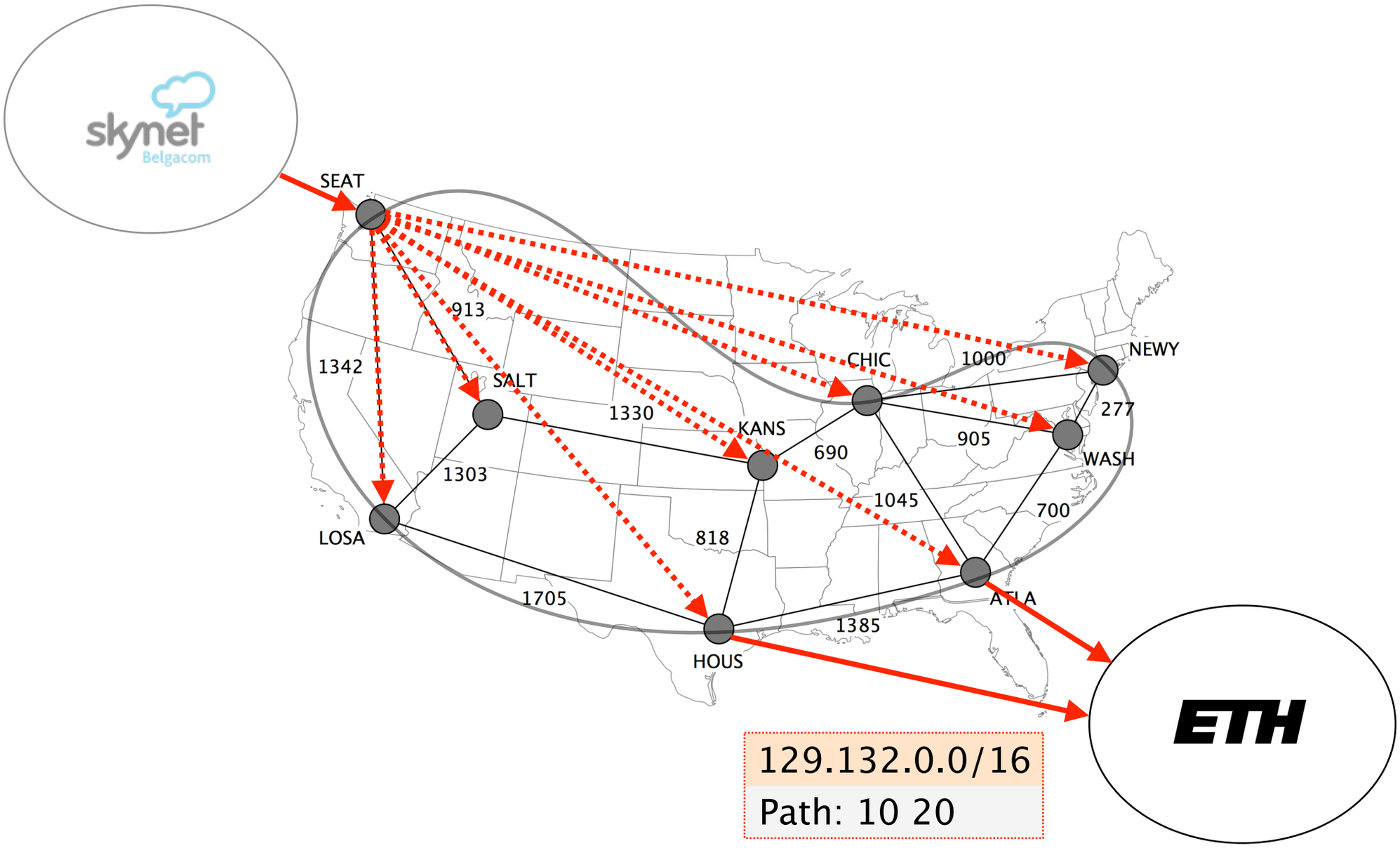




I can reach “129.132/16” via SEAT,
internal NH is CHIC

learned via IGP (*e.g.*, OSPF)

Routes disseminated internally are then announced externally again, using eBGP sessions



On the wire, BGP is a rather simple protocol composed of four basic messages

type	used to…
OPEN	establish TCP-based BGP sessions
NOTIFICATION	report unusual conditions
UPDATE	inform neighbor of a new best route a change in the best route the removal of the best route
KEEPALIVE	inform neighbor that the connection is alive

UPDATE

inform neighbor of a new best route

a change in the best route

the removal of the best route

BGP UPDATEs carry an IP prefix
together with a set of attributes

IP prefix

Attributes

BGP UPDATEs carry an IP prefix
together with a set of attributes

IP prefix

Attributes

Describe route properties

used in route selection/exportation decisions

are either local (*only* seen on iBGP)

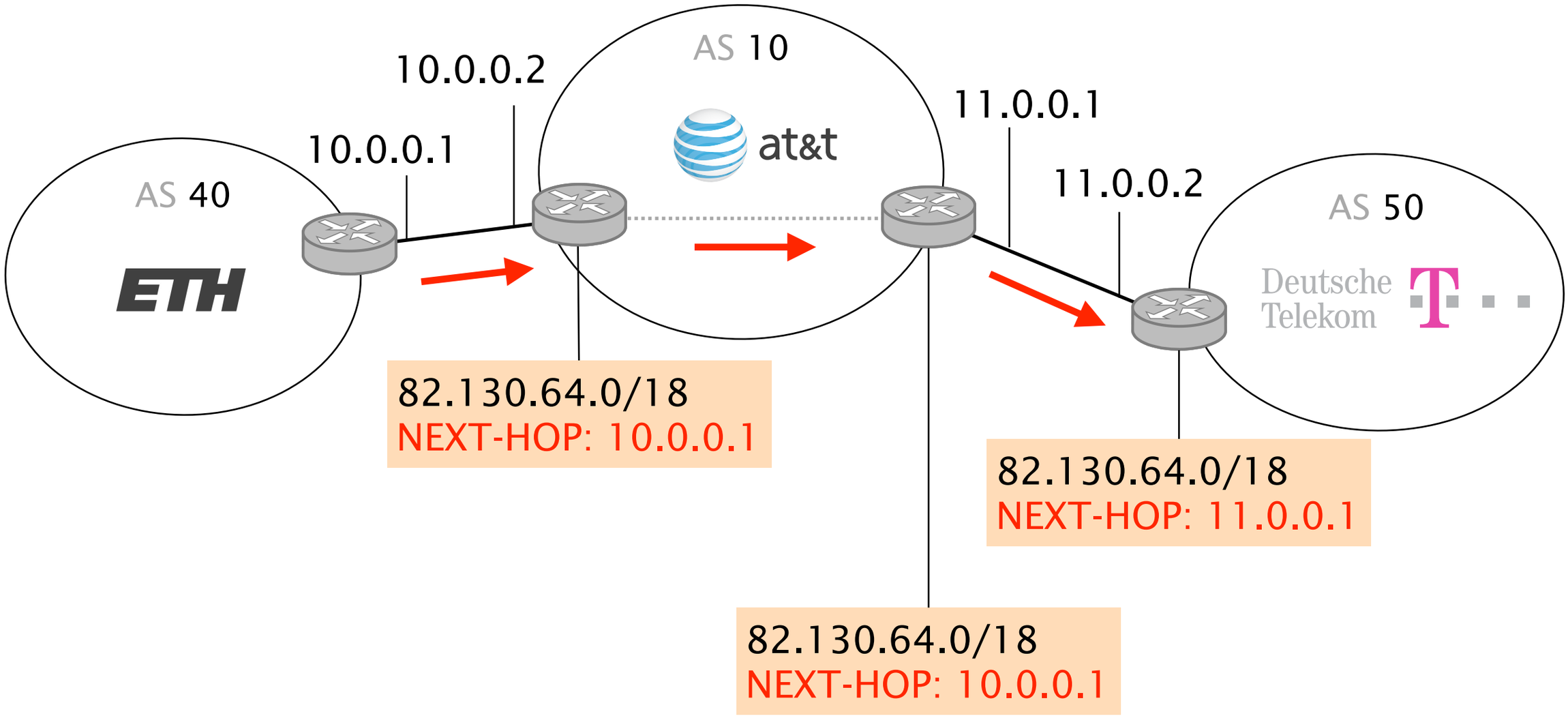
or global (seen on iBGP *and* eBGP)

Attributes	Usage
NEXT-HOP	egress point identification
AS-PATH	loop avoidance
	outbound traffic control
	inbound traffic control
LOCAL-PREF	outbound traffic control
MED	inbound traffic control

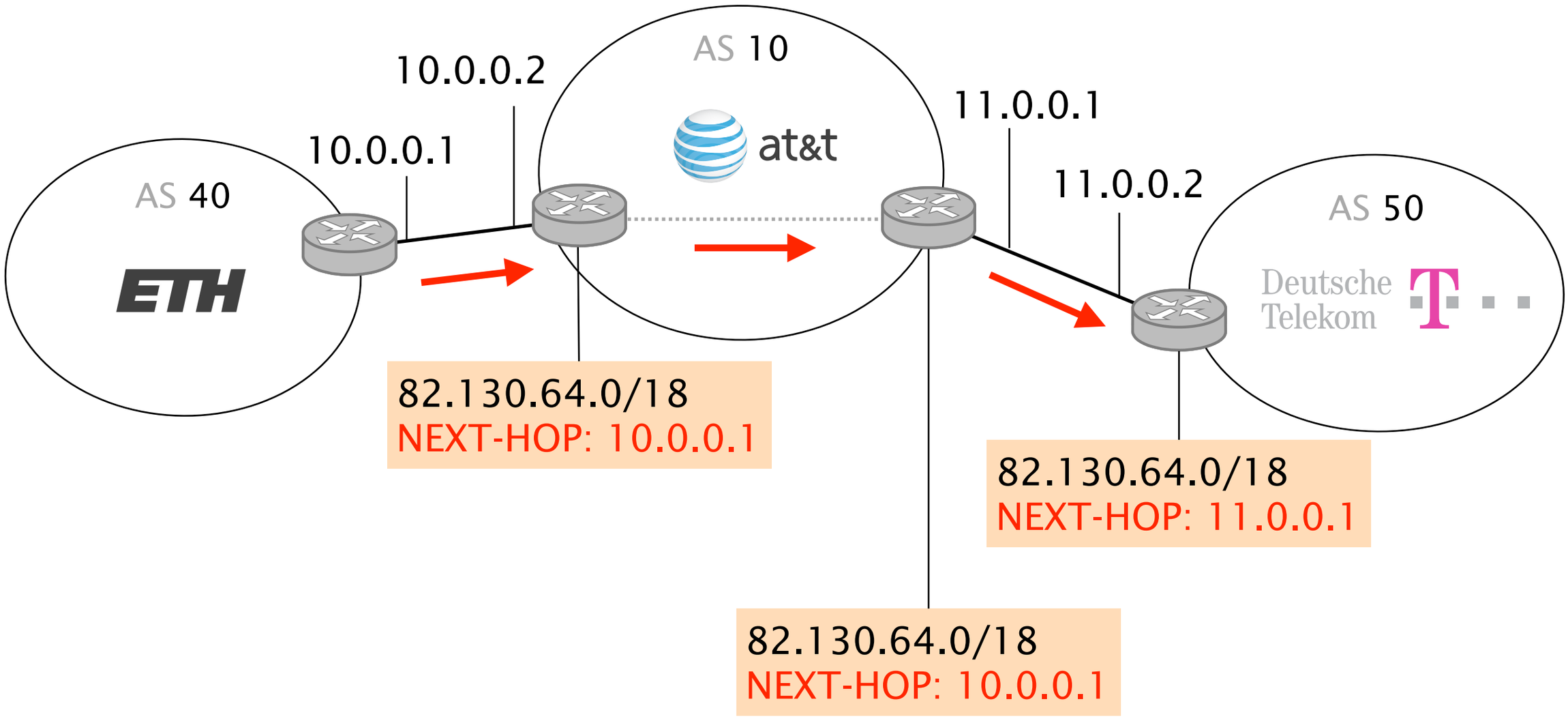
The NEXT-HOP is a global attribute which indicates where to send the traffic next

The NEXT-HOP is set when the route enters an AS,

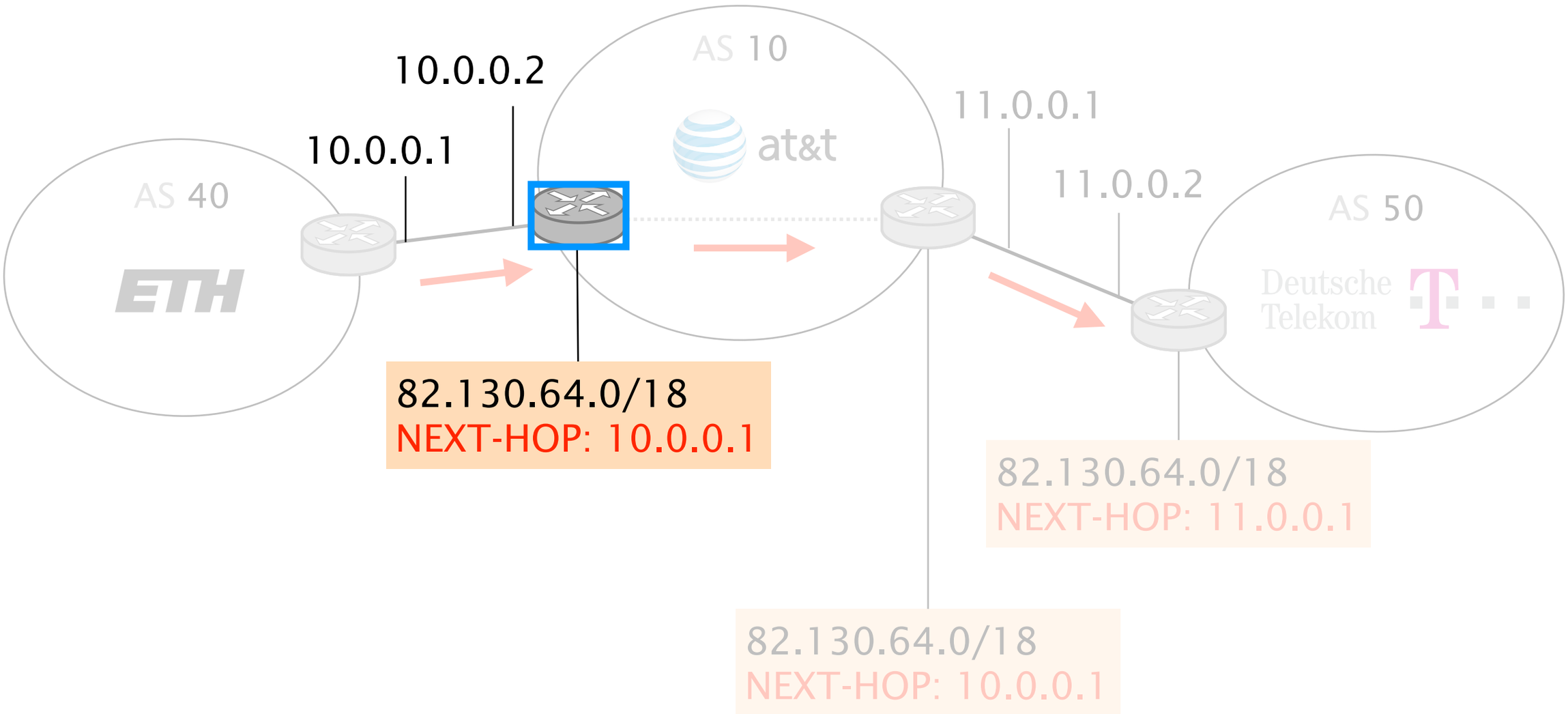
by default, it does not change *within* the AS



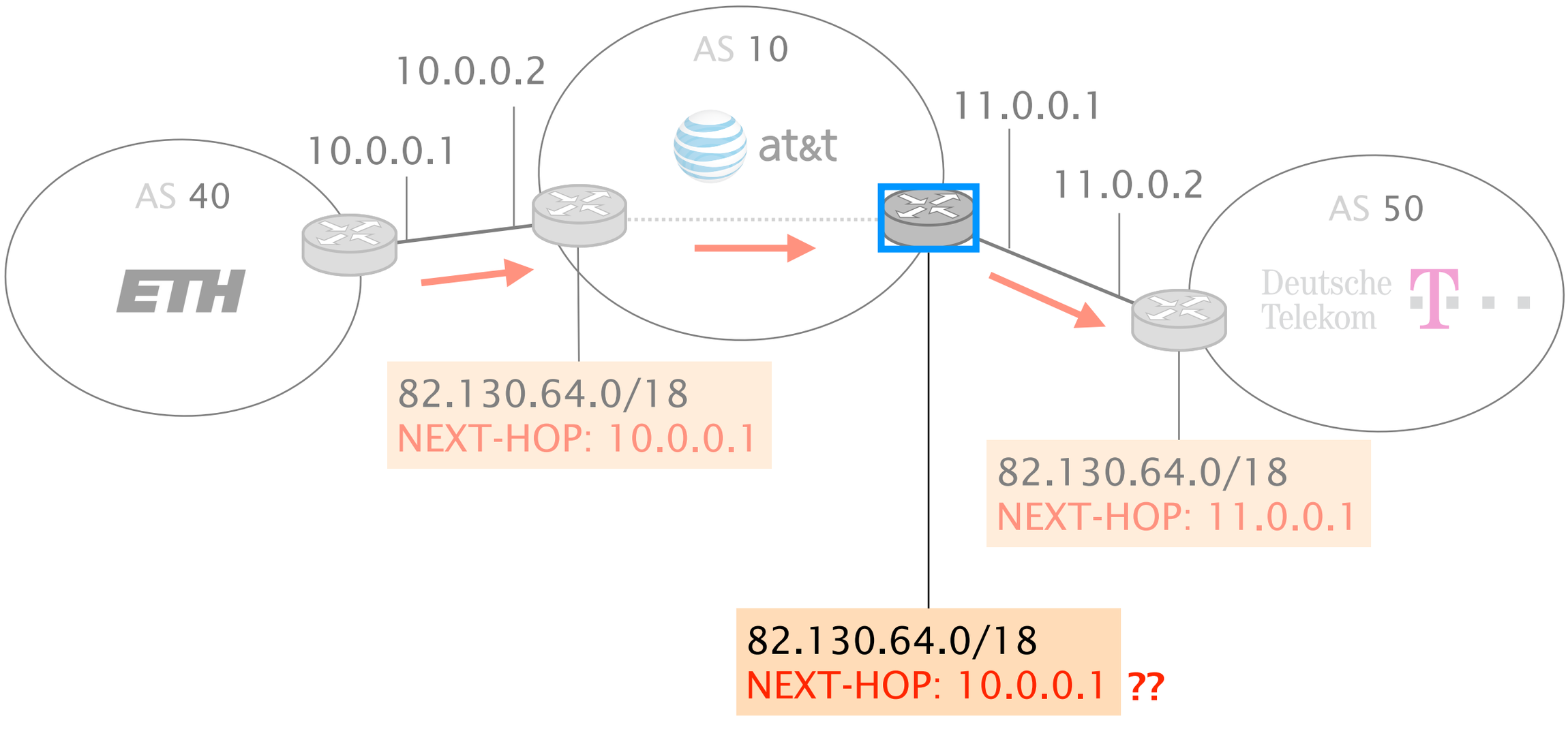
For externally-learned routes, this means that the NEXT-HOP is the IP address of the neighbor's eBGP router, here **10.0.0.1** for at&t



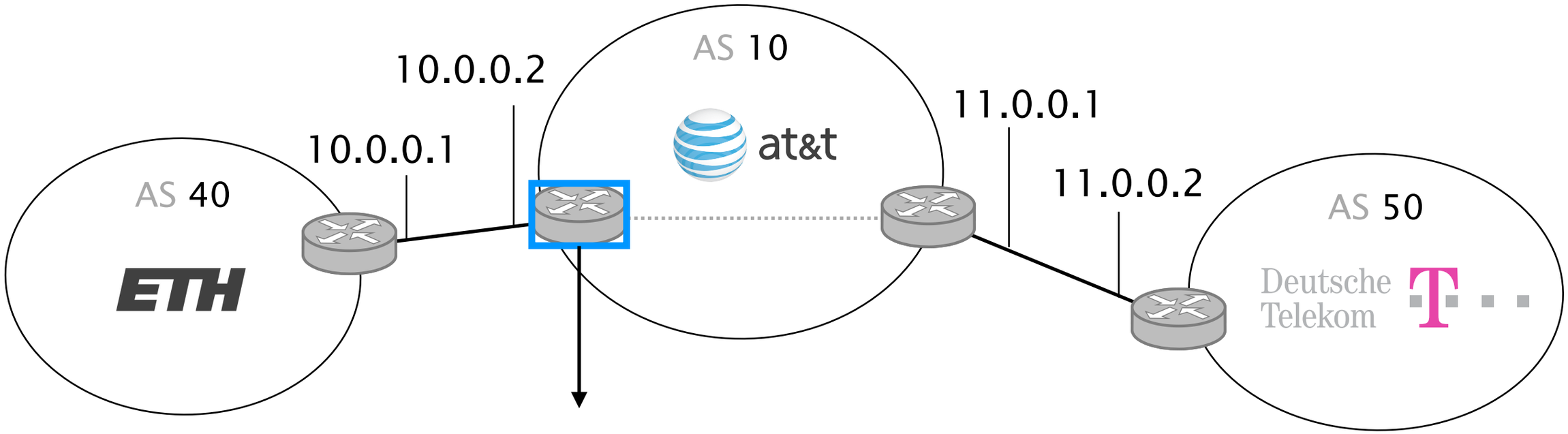
For this router, reaching 10.0.0.1 is not a problem as it is directly connected to the corresponding subnet (10.0.0.0/30)



That router is *not* directly to the NEXT-HOP's subnet (10.0.0.0/30)
and does not know how to reach it, it will therefore drop the BGP route…

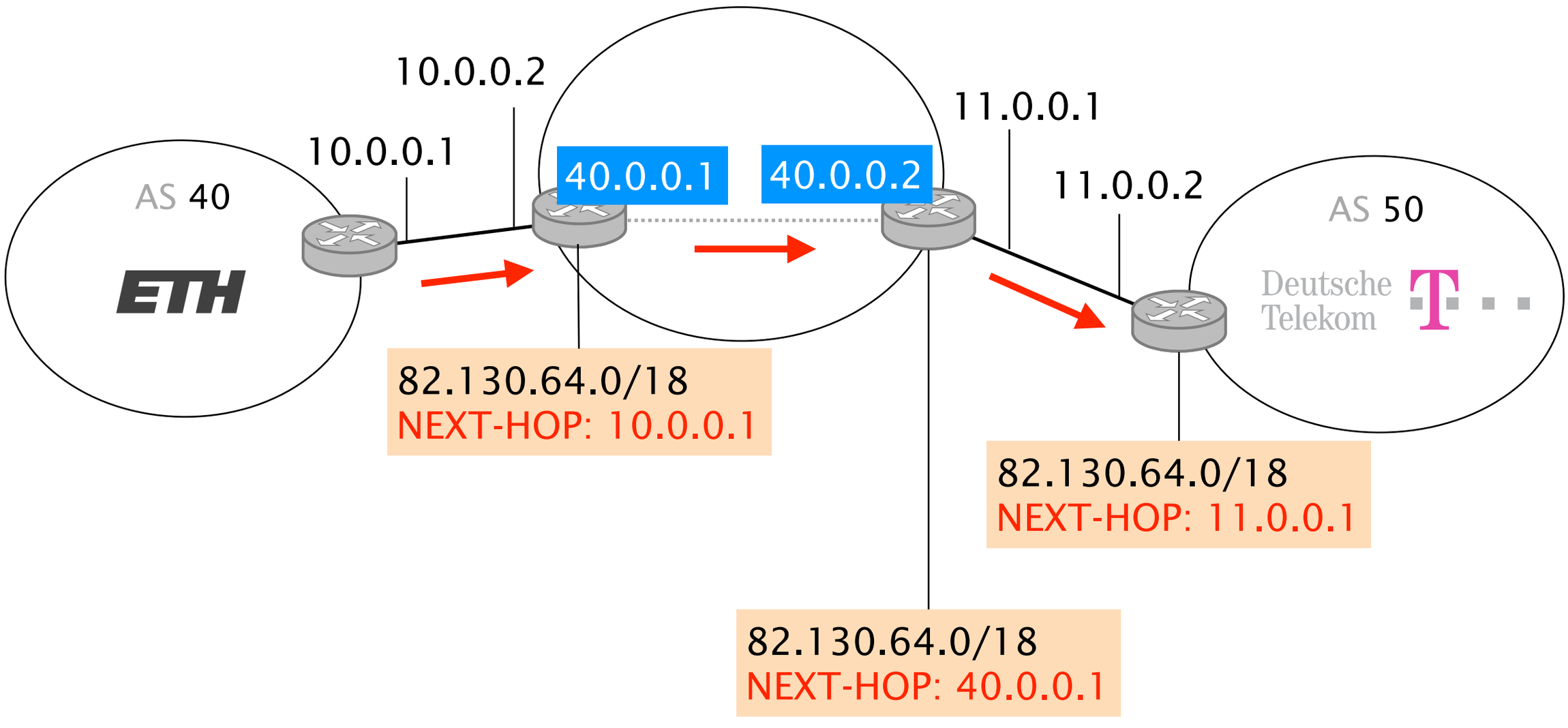


One solution is for the external router to redistribute
the prefixes attached to the external interfaces into the IGP

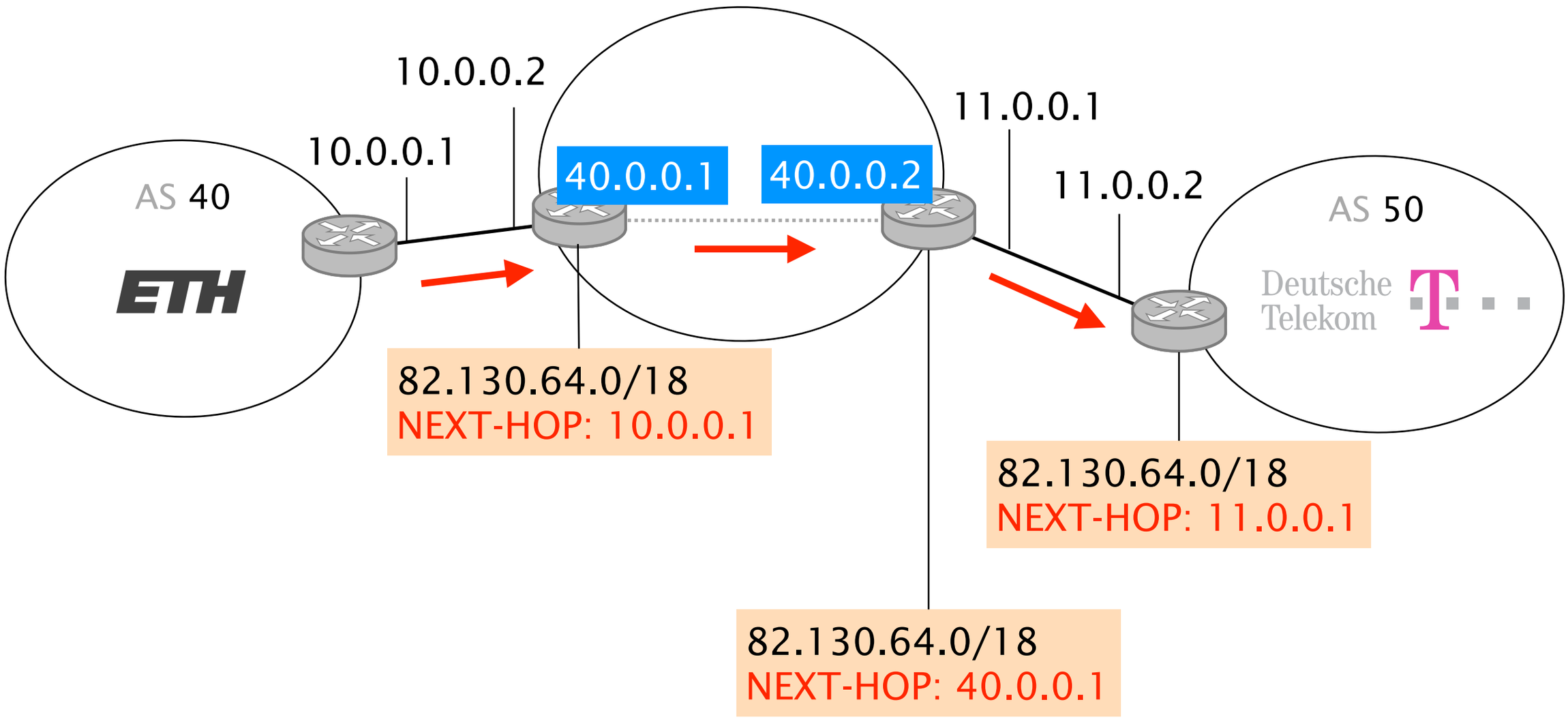


announce 10.0.0.0/30
in OSPF

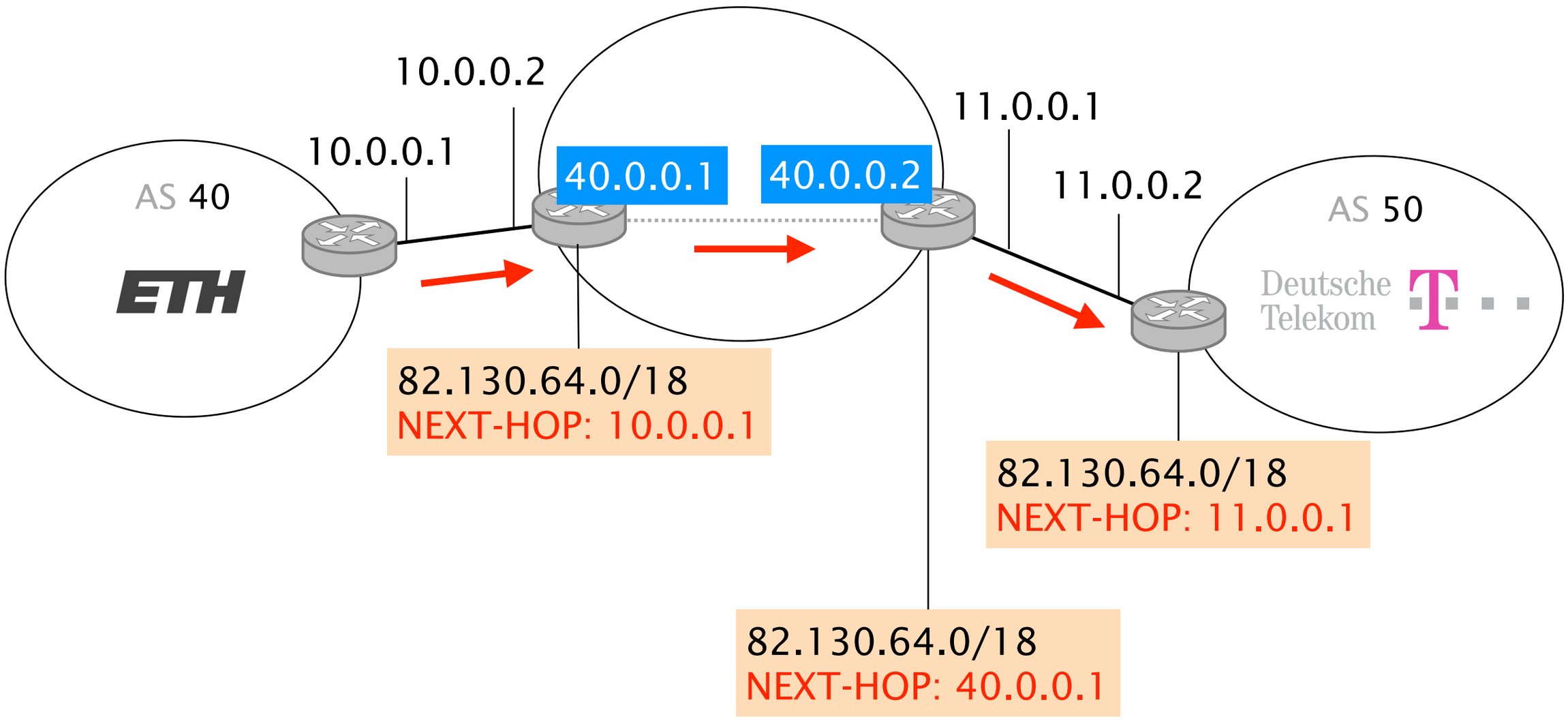
Another solution is for the border router to rewrite the NEXT-HOP before sending it over iBGP, usually to its loopback address



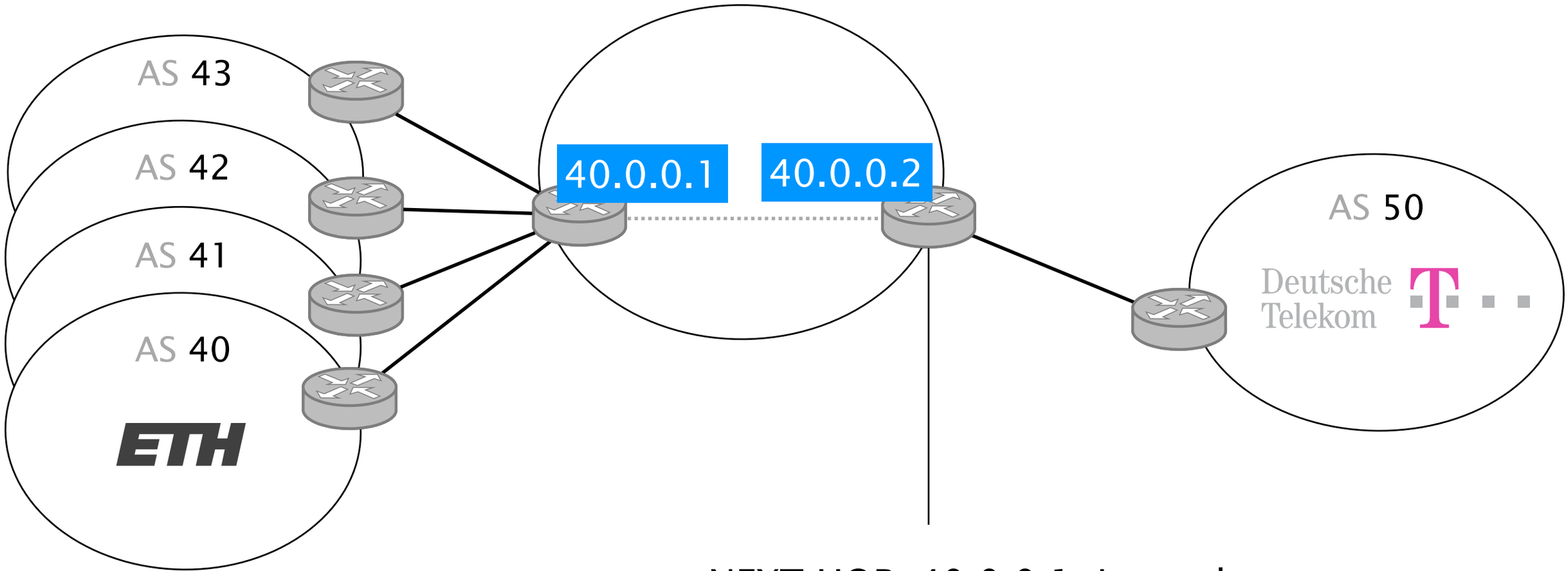
Of course, loopback addresses need to be reachable network-wide.
Typically, each router advertise its loopback (as a /32) in the IGP



Rewriting the next-hop to the eBGP router's loopback is known as "**next-hop-self**"

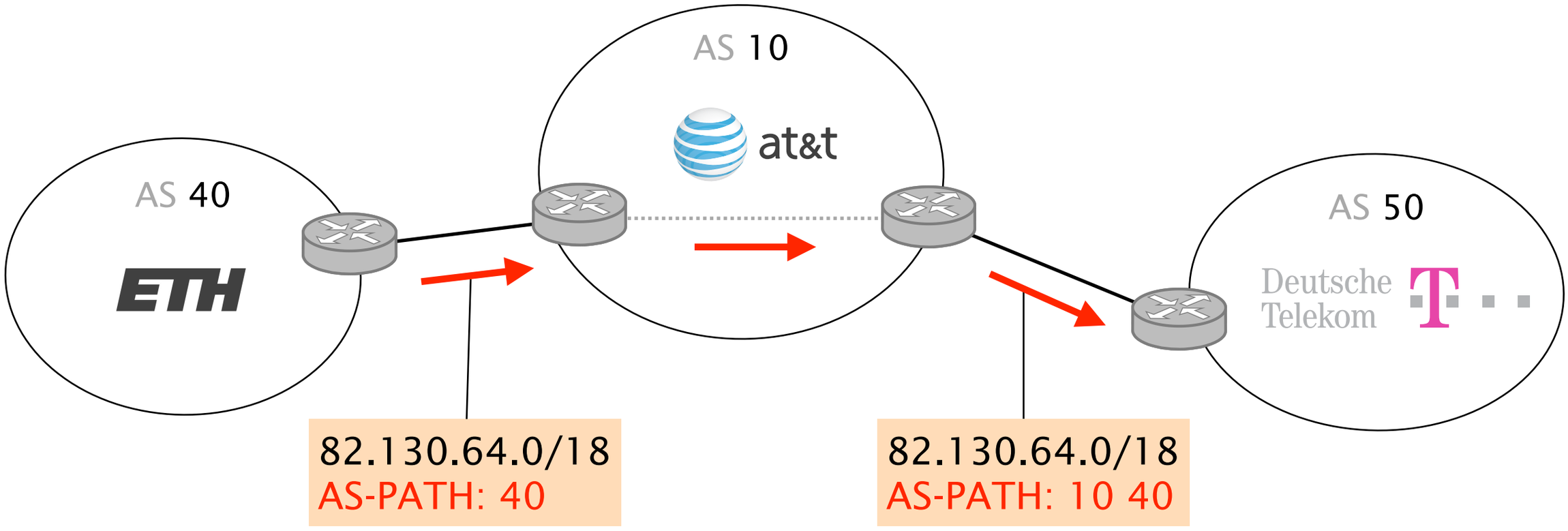


The advantage of next-hop-self is to spare the need to advertise *each* prefix attached to an external link in the IGP

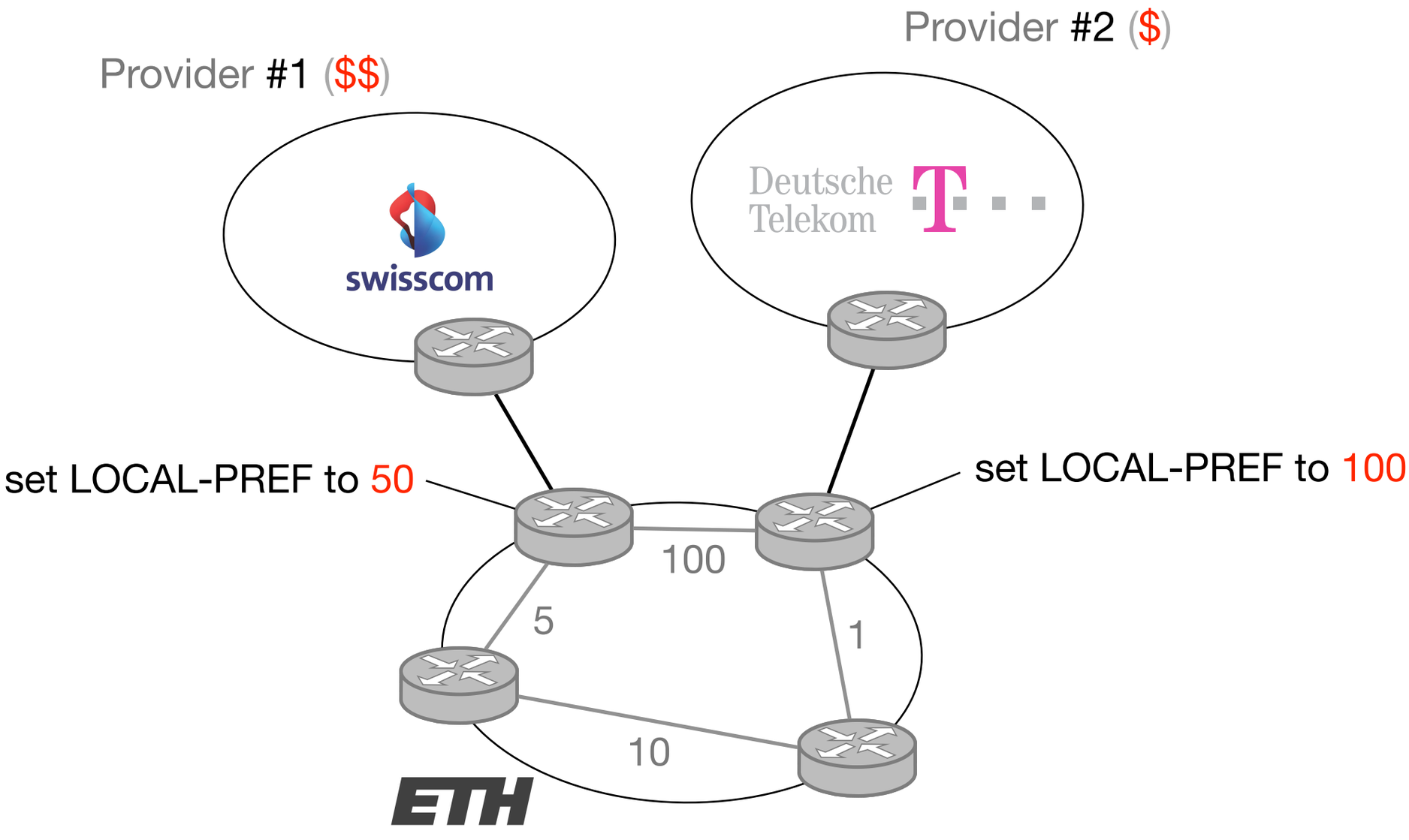


one NEXT-HOP, 40.0.0.1, is used
to reach routes announced by AS 40, 41, 42, 43…

The AS-PATH is a global attribute that lists
all the ASes a route has traversed (in reverse order)


AS 10
at&t
AS 40
ETH
AS 50
Deutsche Telekom
82.130.64.0/18
AS-PATH: 40
82.130.64.0/18
AS-PATH: 10 40

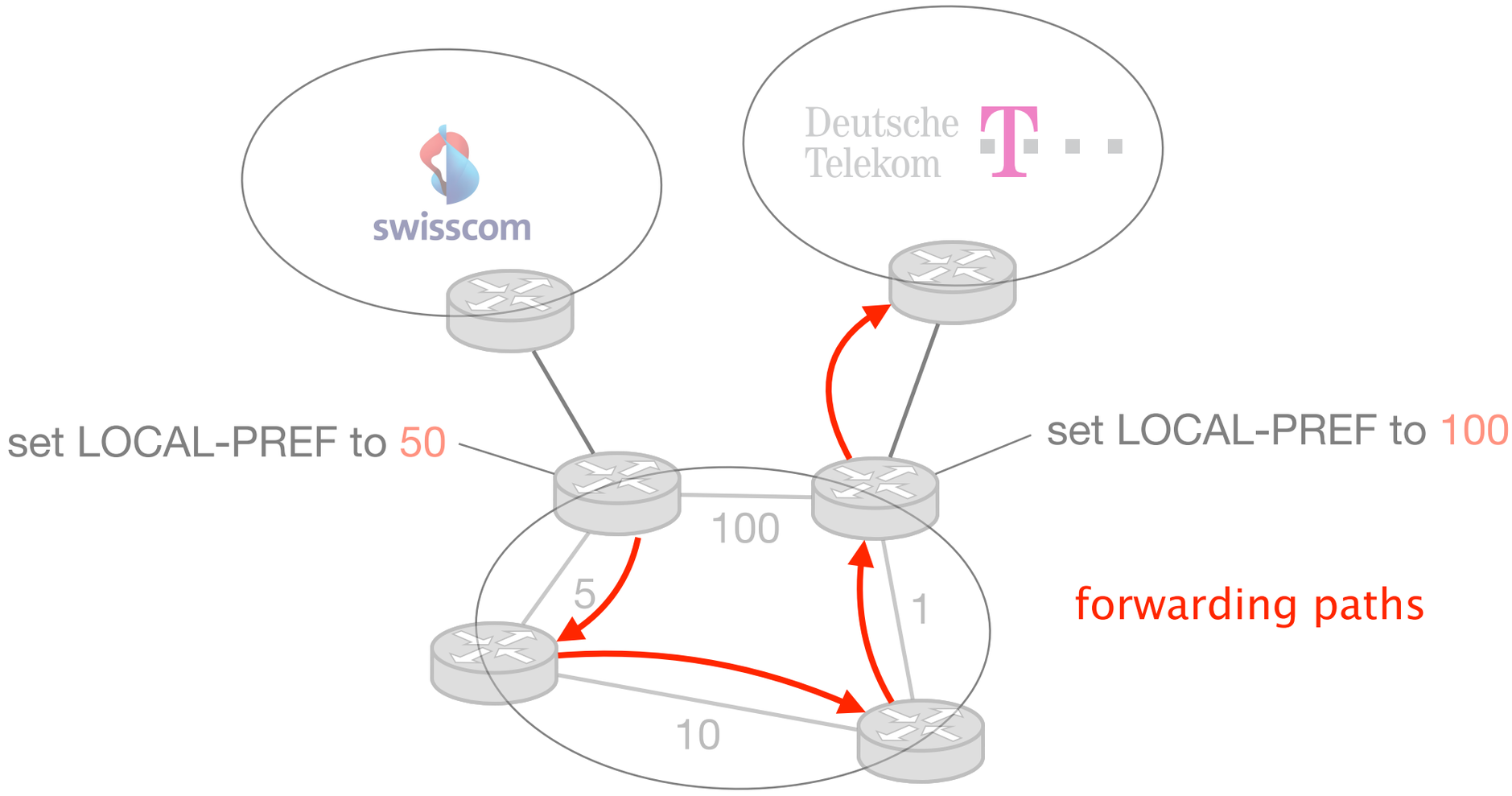
The LOCAL-PREF is a *local* attribute set at the border, it represents how “preferred” a route is


Provider #1 ($$)
Provider #2 ($)
Deutsche Telekom
swisscom
set LOCAL-PREF to 50
set LOCAL-PREF to 100
100
5
1
10
ETH

By setting a higher LOCAL-PREF,

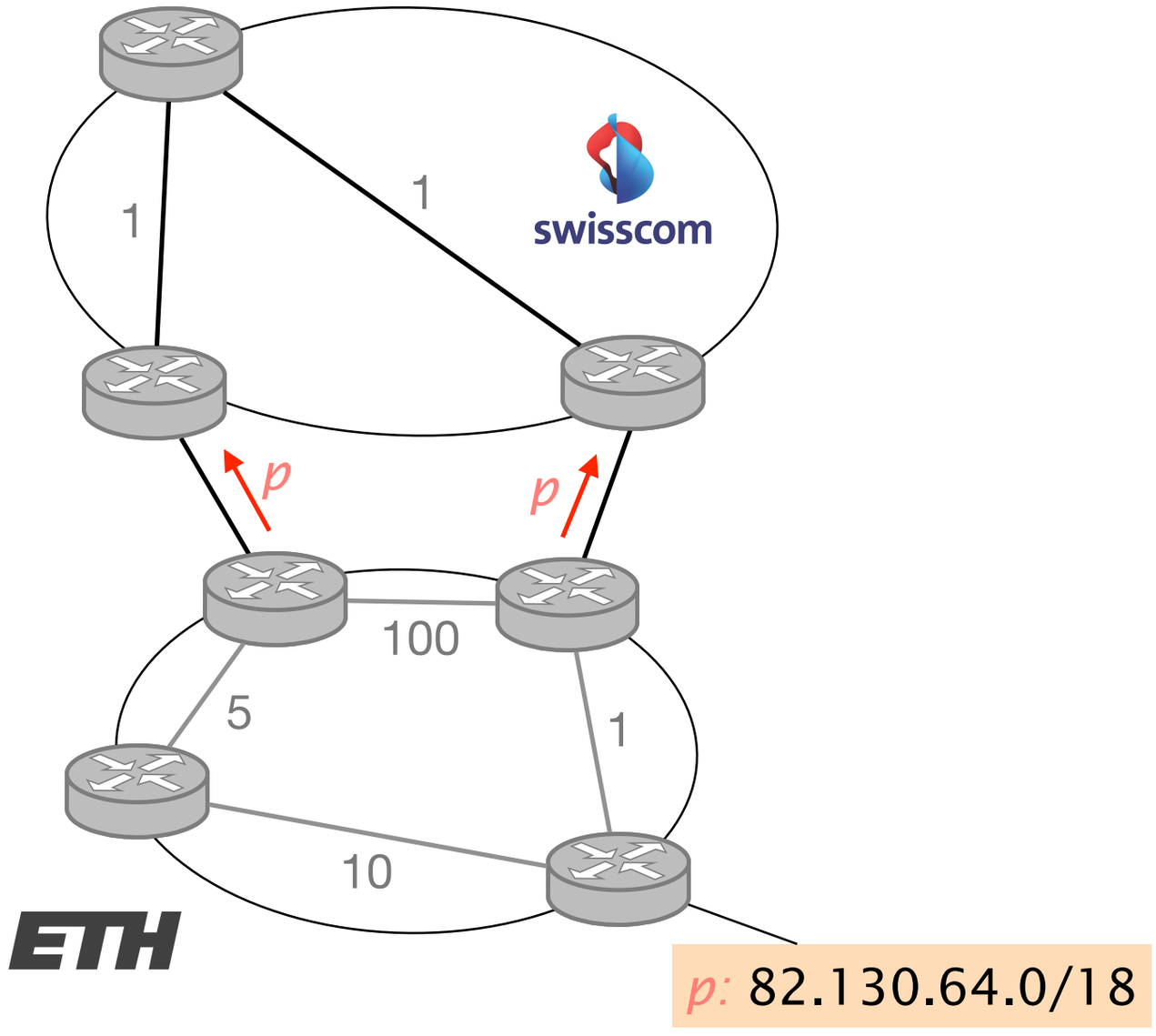
all routers end up using DT to reach any external prefixes,

even if they are closer (IGP-wise) to the Swisscom egress



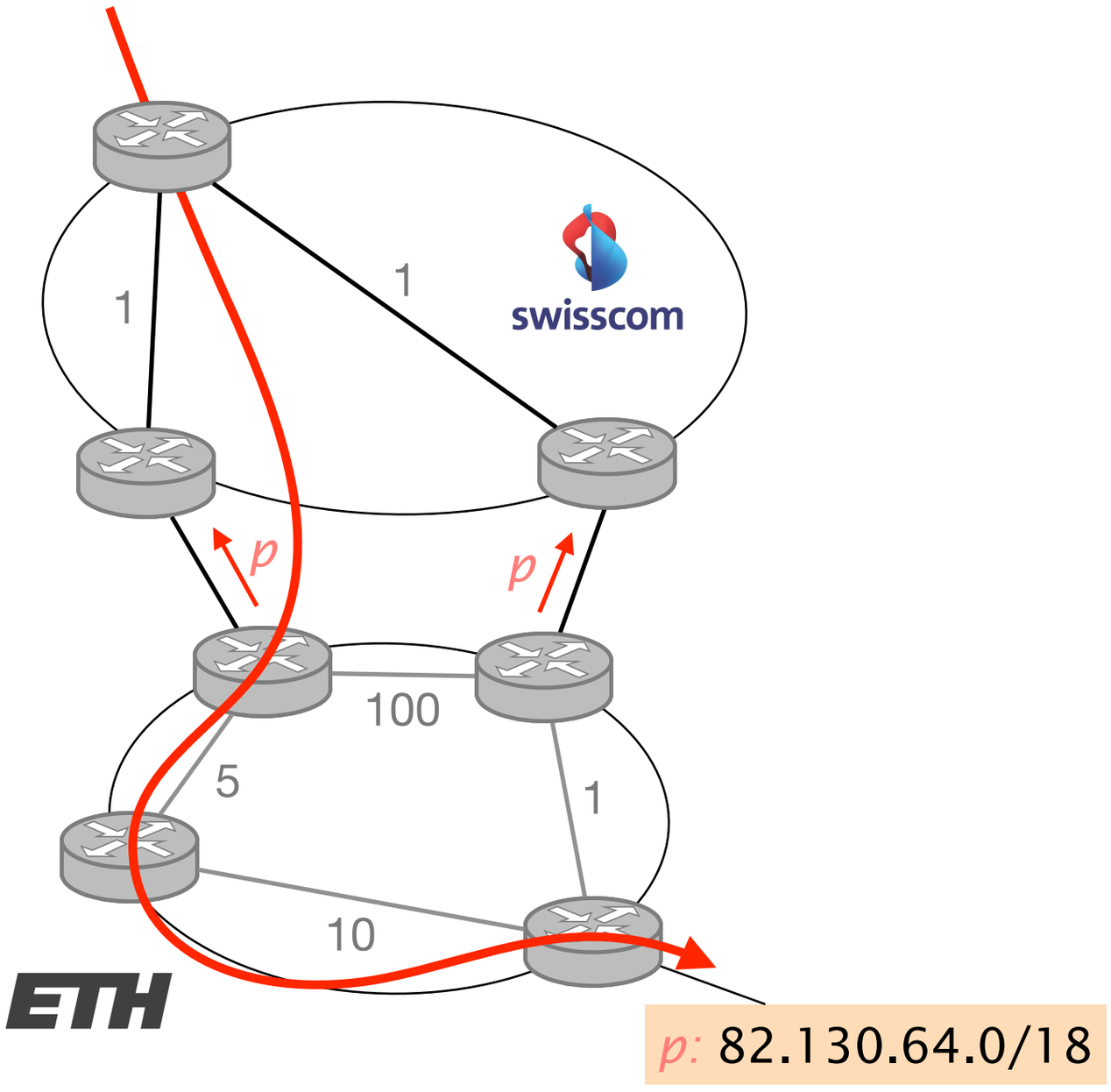
The MED is a *global* attribute which encodes
the relative “proximity” of a prefix wrt to the announcer

Swisscom receives two routes to reach *p*



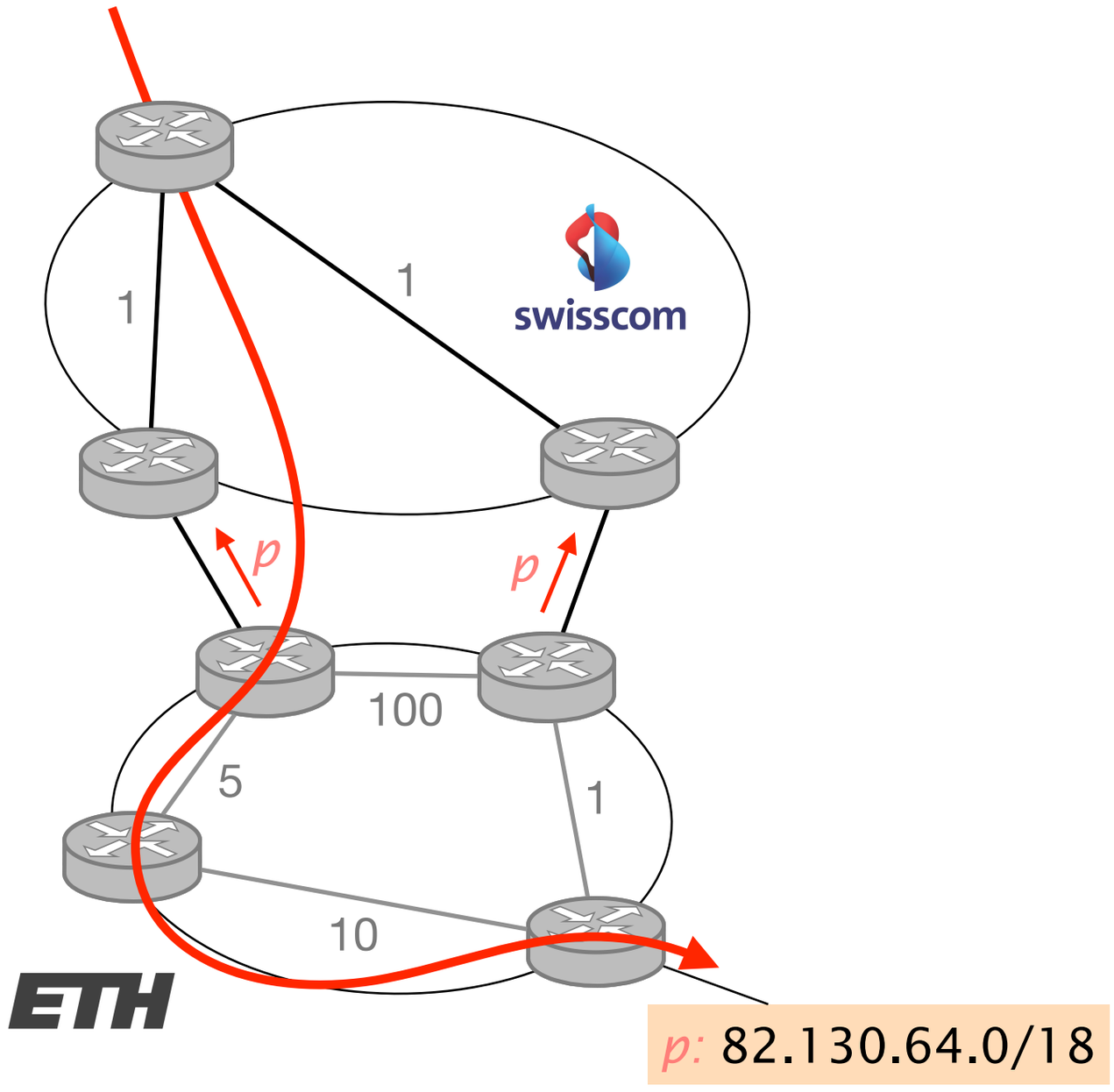
Swisscom receives two routes to reach *p*

and chooses (arbitrarily) its left router as egress

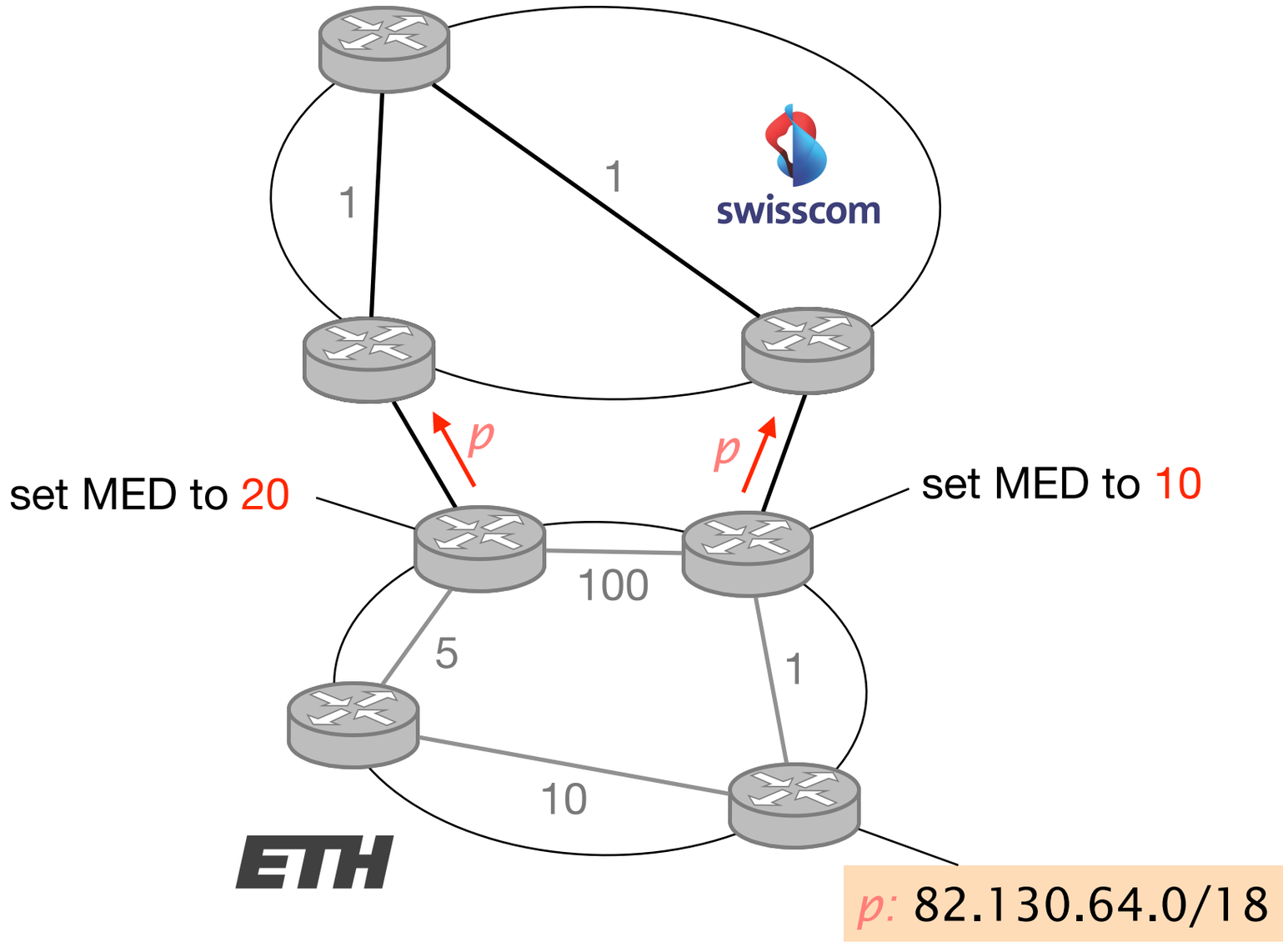


Yet, ETH would prefer to receive traffic for *p*

on its right border router which is closer to the actual destination



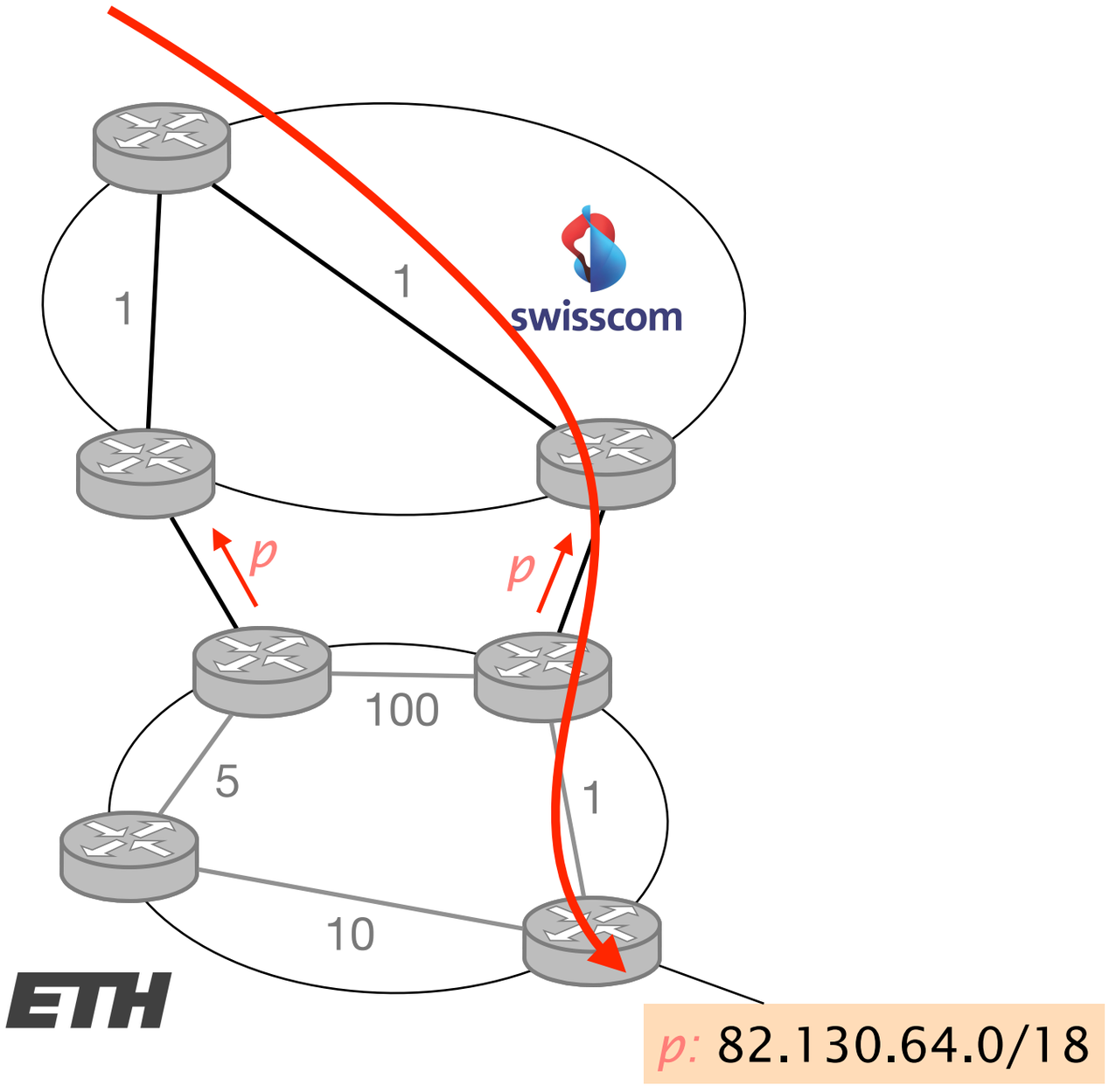
ETH can communicate that preferences to Swisscom
by setting a higher MED on *p* when announced from the left



Swisscom receives two routes to reach *p*

and, *given it does not cost it anything more*,

chooses its right router as egress



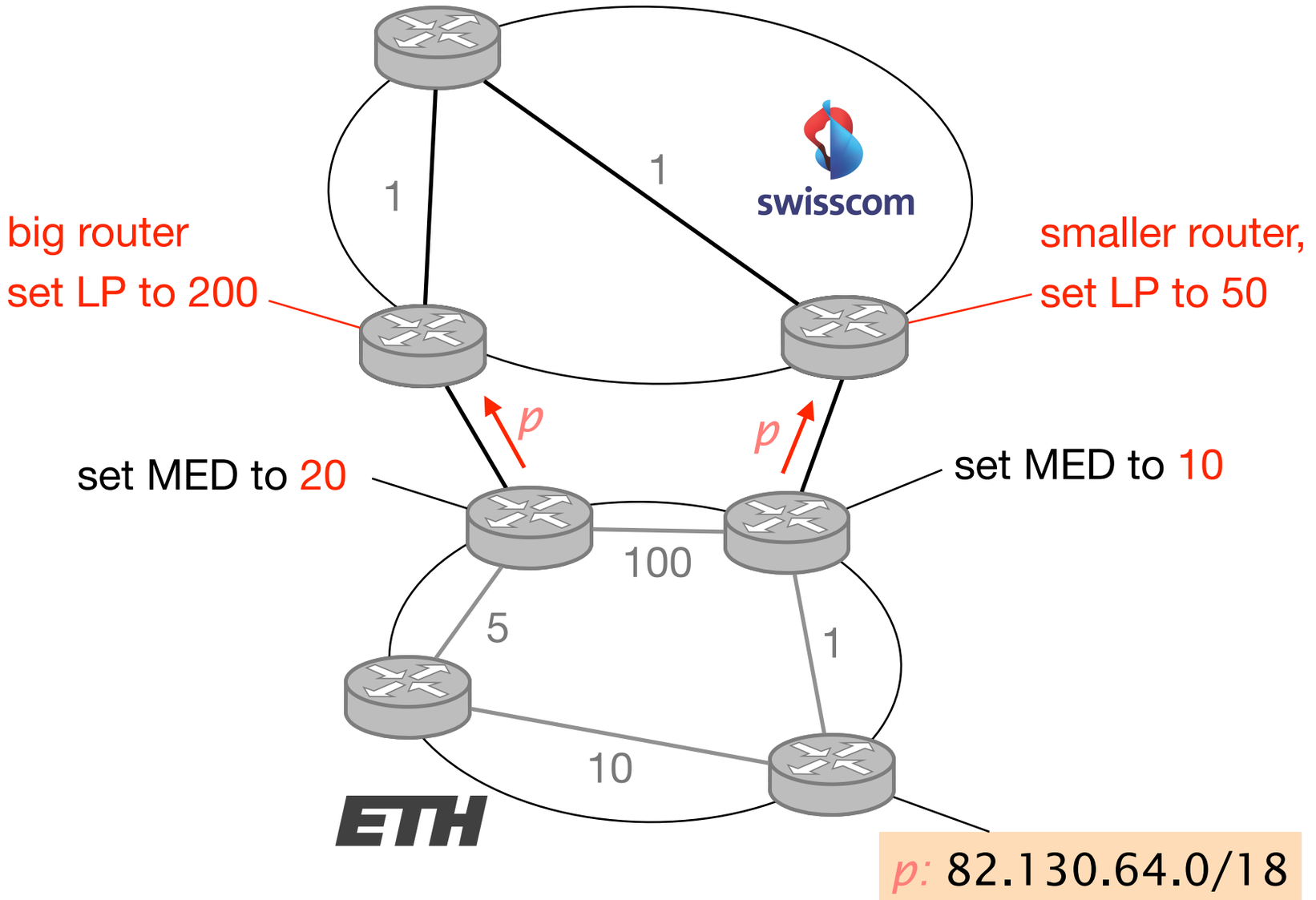
Swisscom receives two routes to reach *p*

and, *given it does not cost it anything more*,

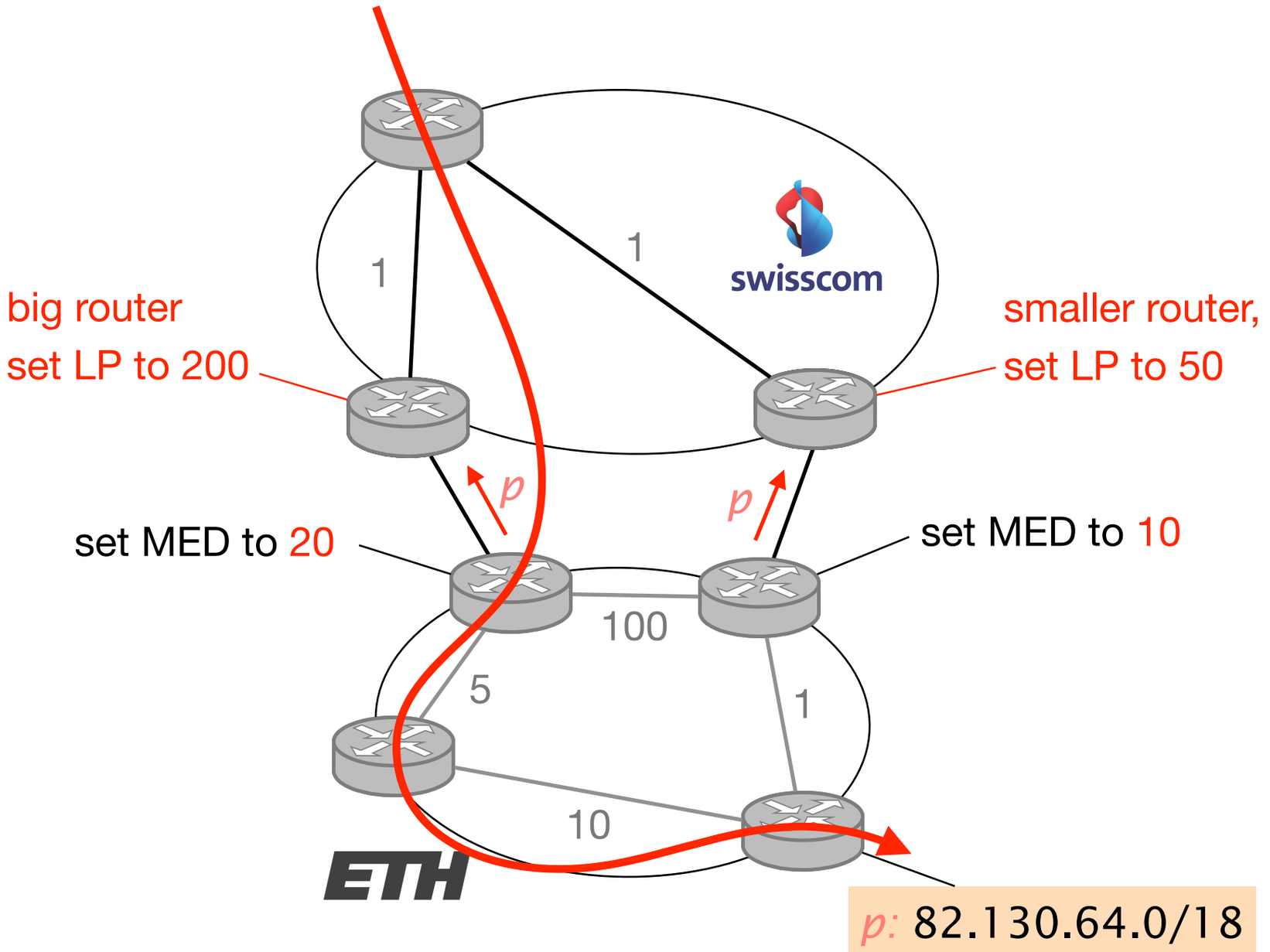
chooses its right router as egress

But what if it does?

Consider that Swisscom always prefer to send traffic via its left egress point (bigger router, less costly)



In this case, Swisscom will not care about the MED value and still push the traffic via its left router



Lesson

The network which is sending the traffic always has the final word when it comes to deciding where to forward

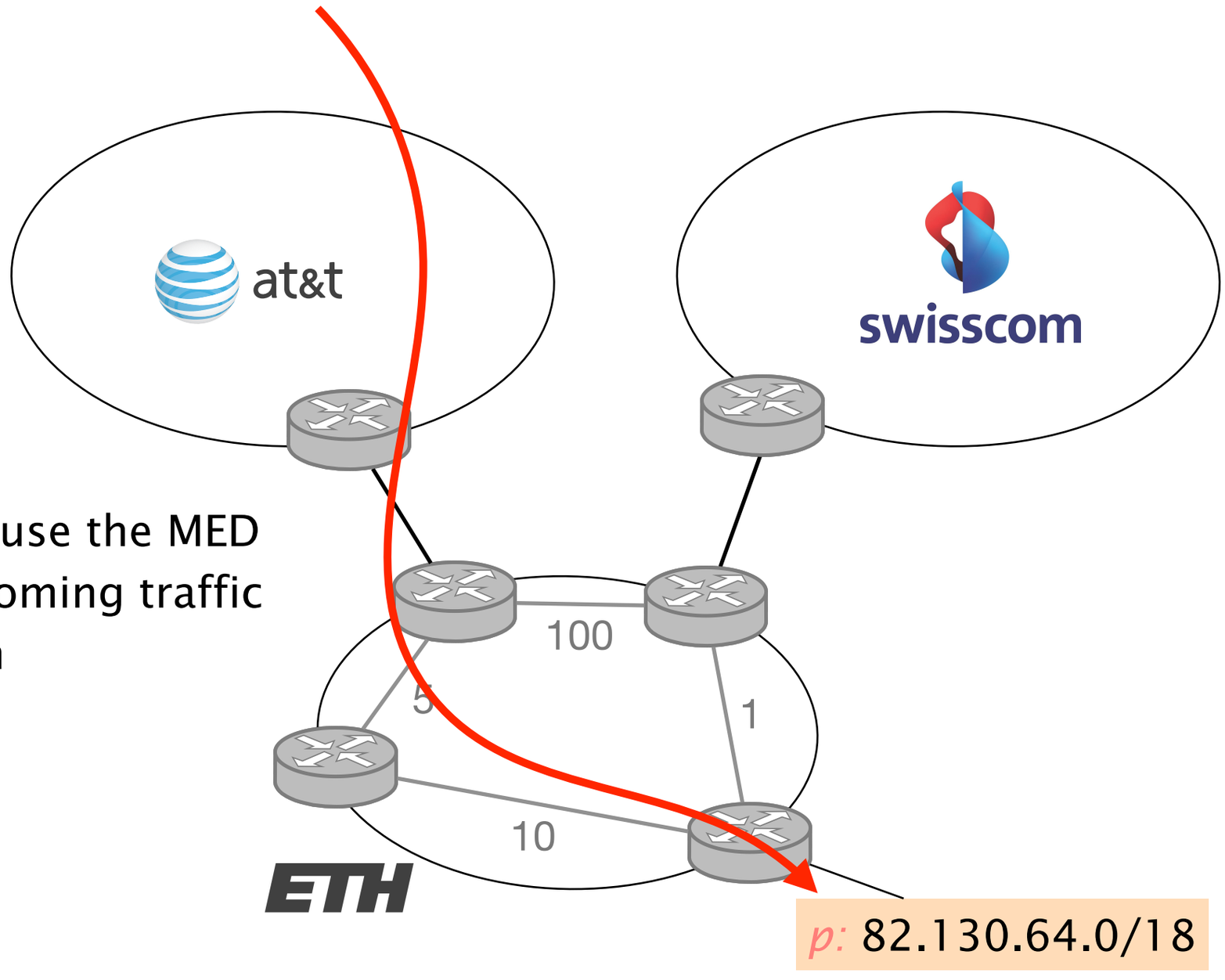
Corollary

The network which is receiving the traffic can just influence remote decision, not control them

With the MED, an AS can influence its inbound traffic

between multiple connection towards the same AS

ETH cannot use the MED
to move incoming traffic
to Swisscom



BGP UPDATEs carry an IP prefix
together with a set of attributes

IP prefix

Attributes

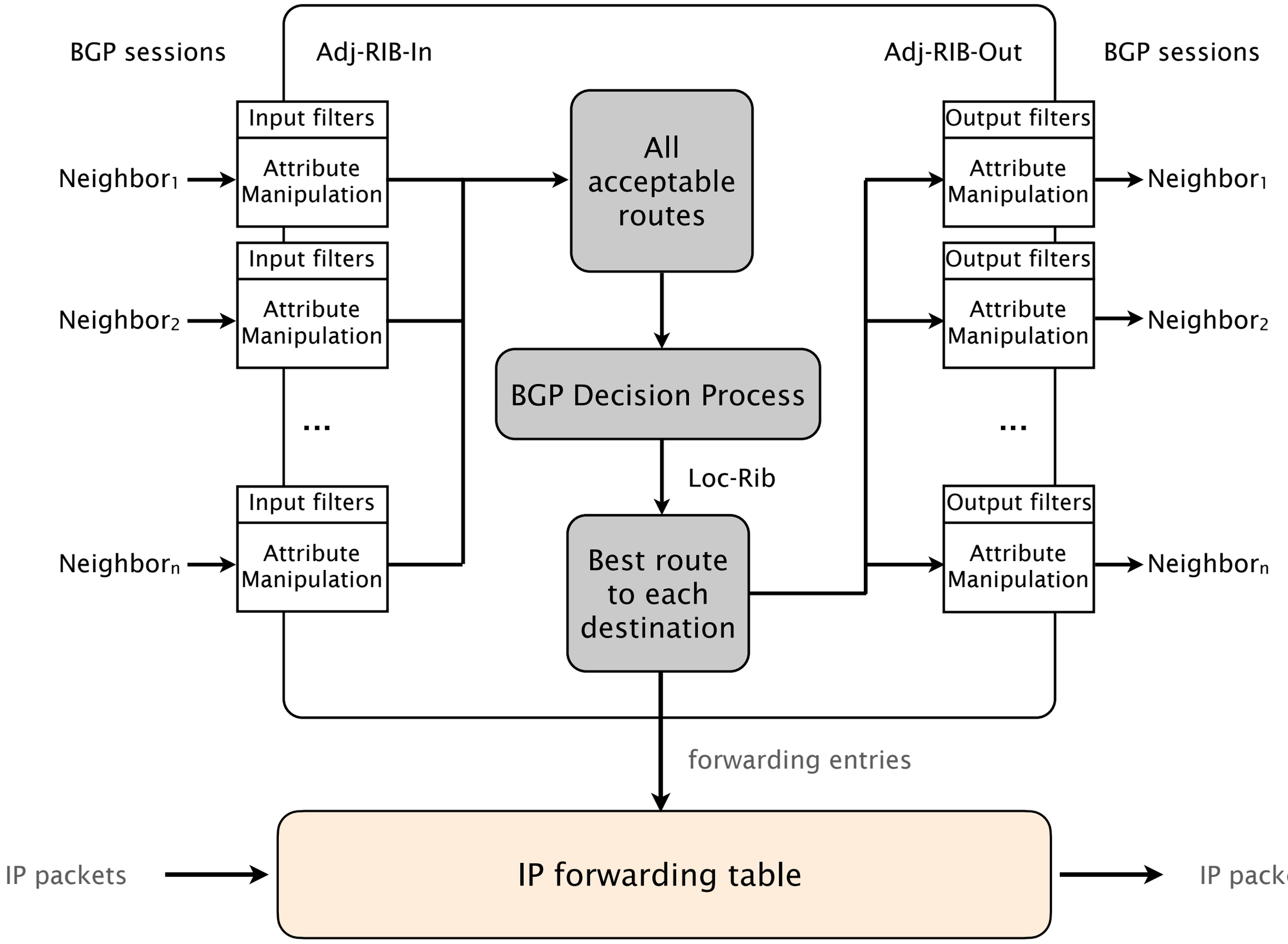
Describe route properties

used in route selection/exportation decisions

are either local (*only* seen on iBGP)

or global (seen on iBGP *and* eBGP)

Each BGP router processes UPDATEs according to a precise pipeline


BGP sessions
Adj-RIB-In
Adj-RIB-Out
BGP sessions
Input filters
Attribute Manipulation
Neighbor1
Input filters
Attribute Manipulation
Neighbor2
...
Input filters
Attribute Manipulation
Neighborn
All acceptable routes
BGP Decision Process
Loc-Rib
Best route to each destination
Output filters
Attribute Manipulation
Neighbor1
Output filters
Attribute Manipulation
Neighbor2
...
Output filters
Attribute Manipulation
Neighborn
forwarding entries
IP packets
IP forwarding table
IP packets

Given the set of all acceptable routes for each prefix, the BGP Decision process elects a single route

BGP is often referred to as a single path protocol

Prefer routes…

with higher LOCAL-PREF

with shorter AS-PATH length

with lower MED

learned via eBGP instead of iBGP

with lower IGP metric to the next-hop

with smaller egress IP address (tie-break)

learned via eBGP instead of iBGP

with lower IGP metric to the next-hop

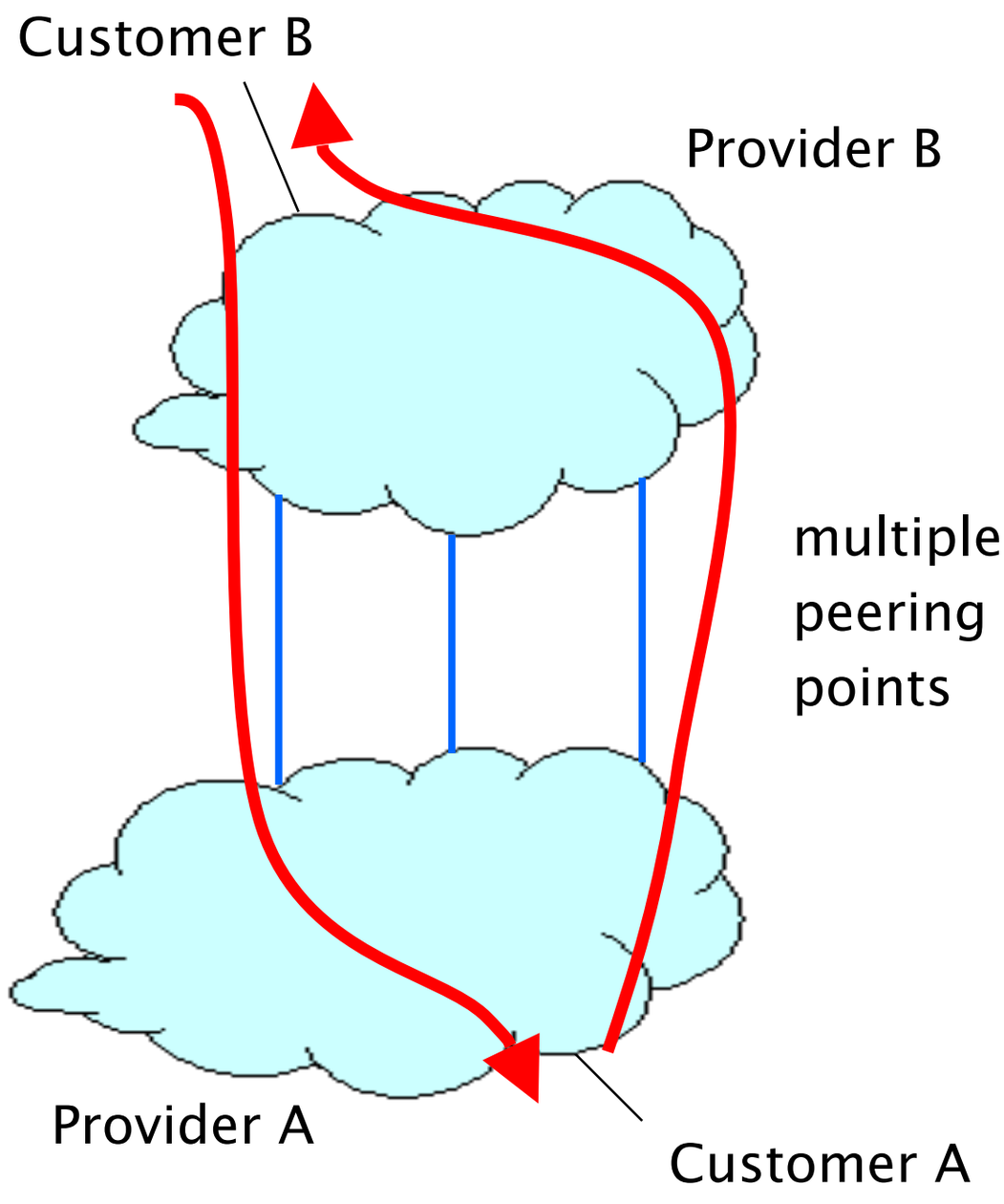
These two steps aim at directing traffic
as quickly as possible out of the AS (early exit routing)

ASes are selfish

They dump traffic
as soon as possible
to someone else

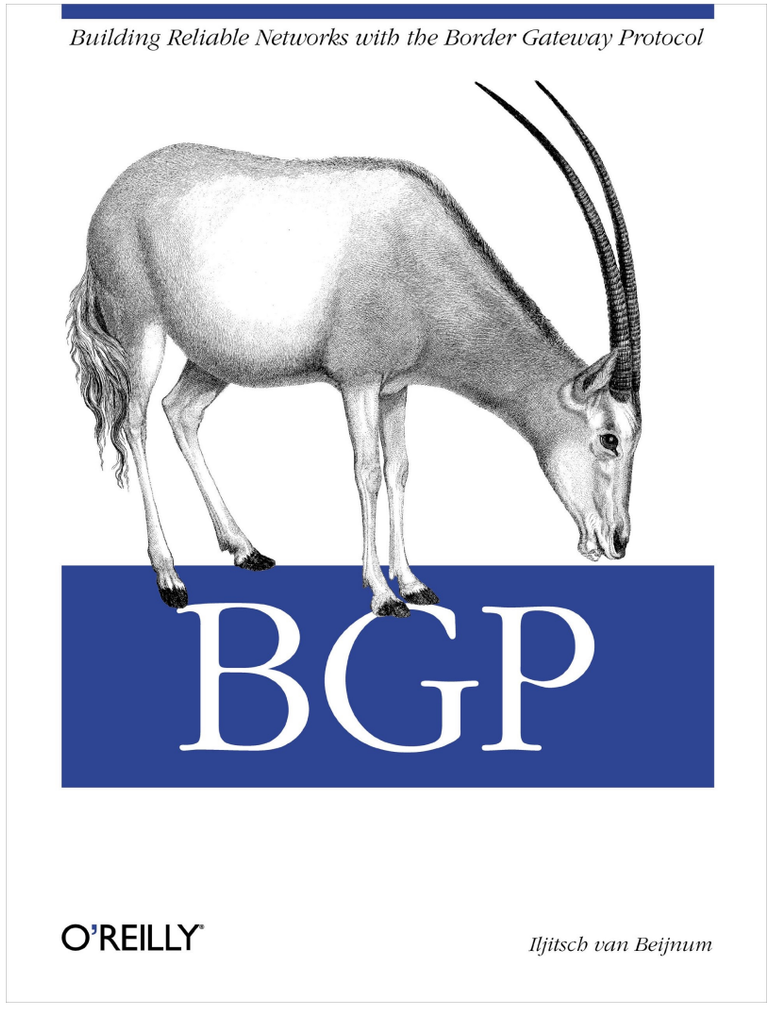
This leads to asymmetric routing

Traffic does not flow on
the same path
in both directions



Border Gateway Protocol

policies and more



Protocol

How it works

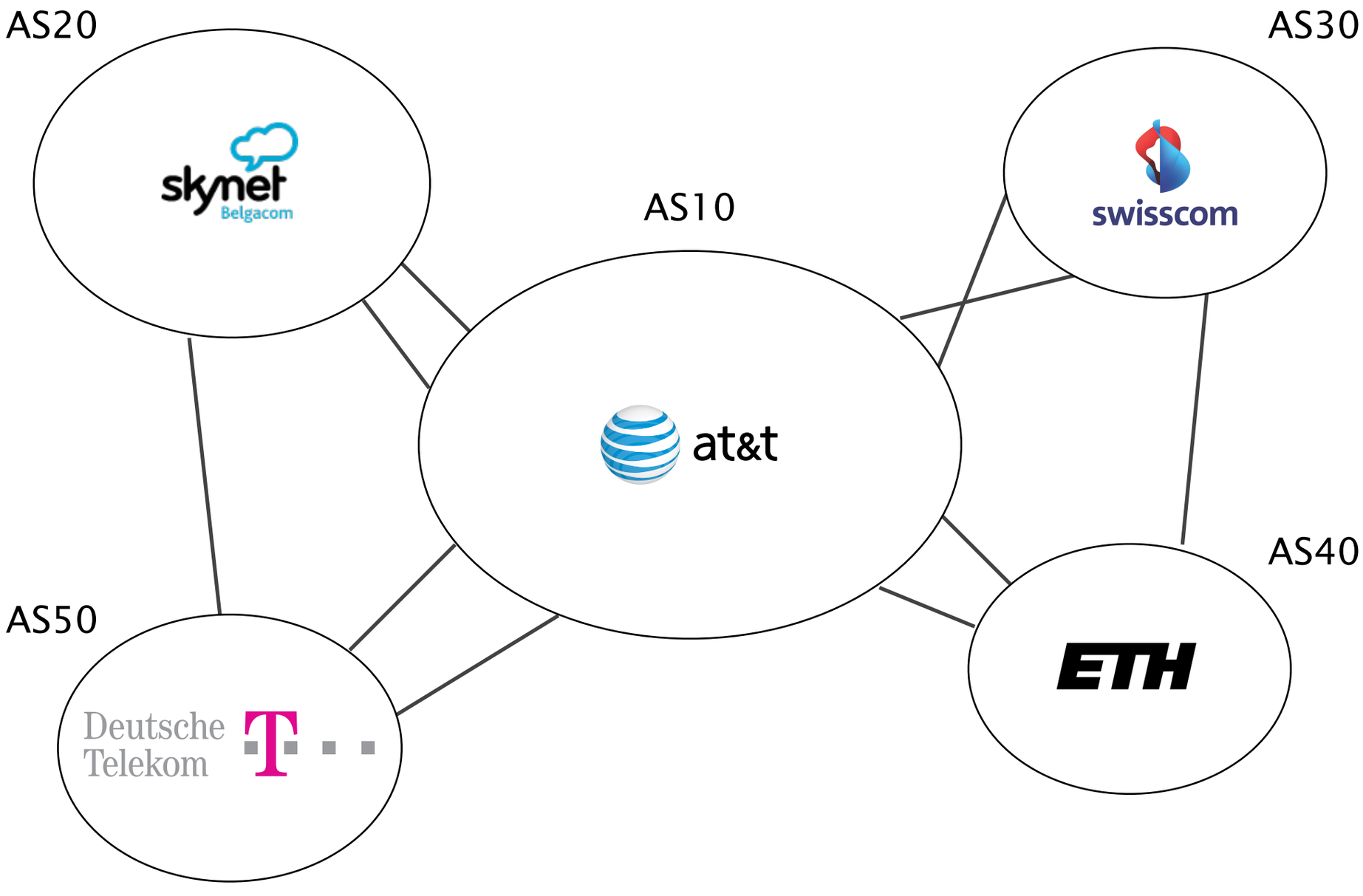
2 Policies

"Follow the money"

Problems

Security, performance, …

The Internet topology is shaped according to business relationships



Intuition

2 ASes connect only if they have a business relationship

BGP is a “follow the money” protocol

There are 2 main business relationships today:

- customer/provider
- peer/peer

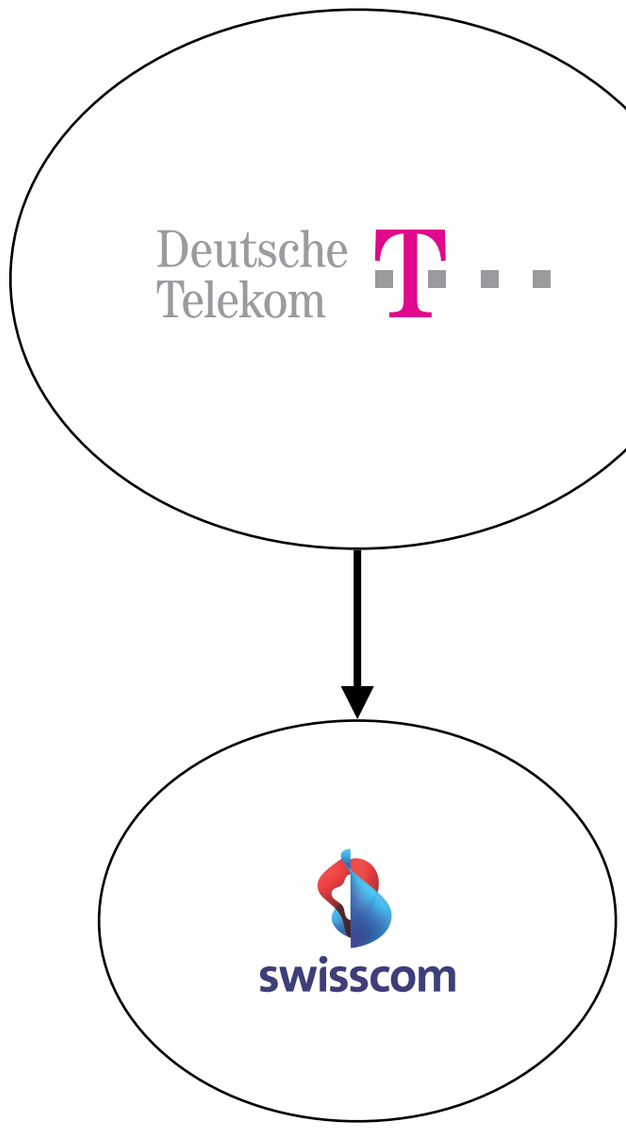
many less important ones (siblings, backups,…)

There are 2 main business relationships today:

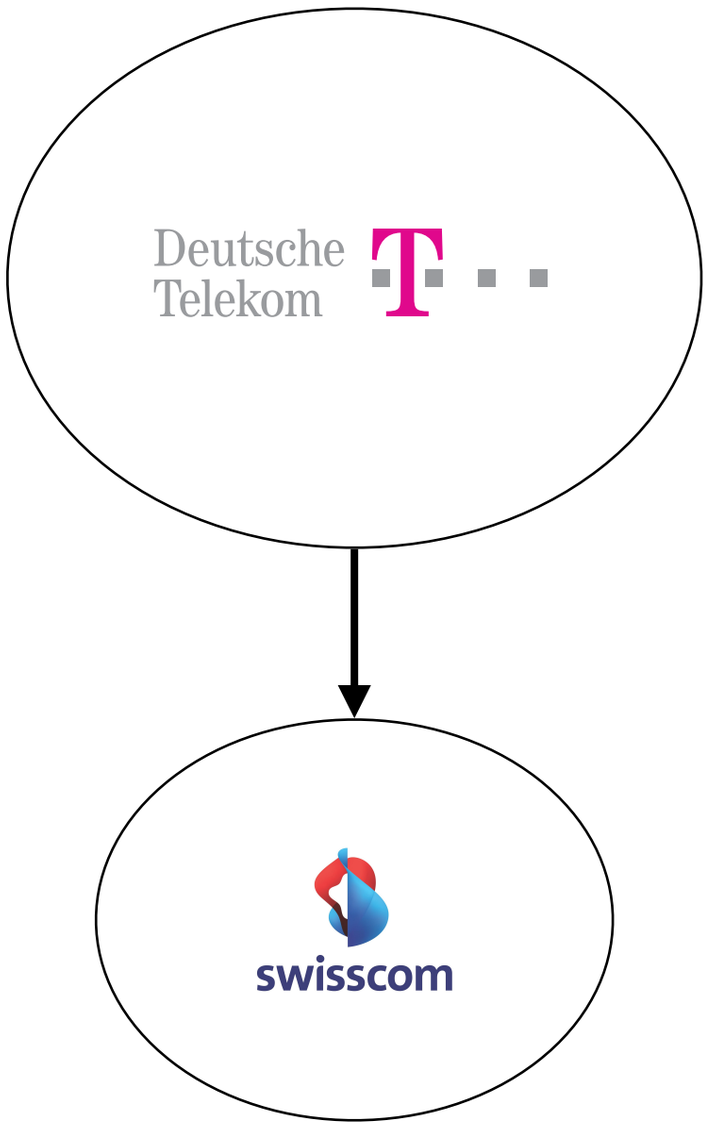
- customer/provider
- peer/peer

Customers pay providers

to get Internet connectivity



The amount paid is based on peak usage, usually according to the 95th percentile rule



Every 5 minutes, DT

records the # of bytes sent/received

At the end of the month, DT

- sorts all values in decreasing order
- removes the top 5% values
- bills wrt highest remaining value

Most ISPs discounts traffic unit price when pre-committing to certain volume

commit		unit price ($)	Minimum monthly bill ($/month)
10	Mbps	12	120
100	Mbps	5	500
1	Gbps	3.50	3,500
10	Gbps	1.20	12,000
100	Gbps	0.70	70,000

Examples taken from The 2014 Internet Peering Playbook

Internet Transit Prices have been continuously declining during the last 20 years

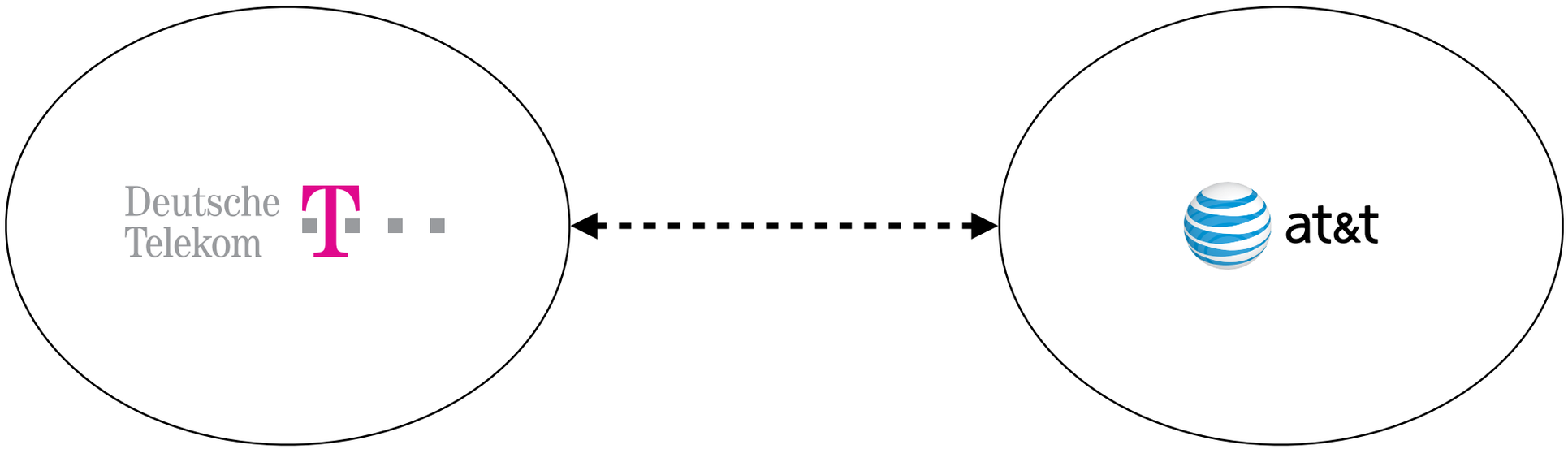
Internet Transit Pricing (1998-2015)			
Source: http://DrPeering.net			
Year	**Internet Transit Price**		**% decline**
1998	$1,200.00	per Mbps	
1999	$800.00	per Mbps	33%
2000	$675.00	per Mbps	16%
2001	$400.00	per Mbps	41%
2002	$200.00	per Mbps	50%
2003	$120.00	per Mbps	40%
2004	$90.00	per Mbps	25%
2005	$75.00	per Mbps	17%
2006	$50.00	per Mbps	33%
2007	$25.00	per Mbps	50%
2008	$12.00	per Mbps	52%
2009	$9.00	per Mbps	25%
2010	$5.00	per Mbps	44%
2011	$3.25	per Mbps	35%
2012	$2.34	per Mbps	28%
2013	$1.57	per Mbps	33%
2014	$0.94	per Mbps	40%
2015	$0.63	per Mbps	33%

The reason? Internet commoditization & competition

There are 2 main business relationships today:

- customer/provider
- peer/peer

Peers don’t pay each other for connectivity,
they do it *out of common interest*

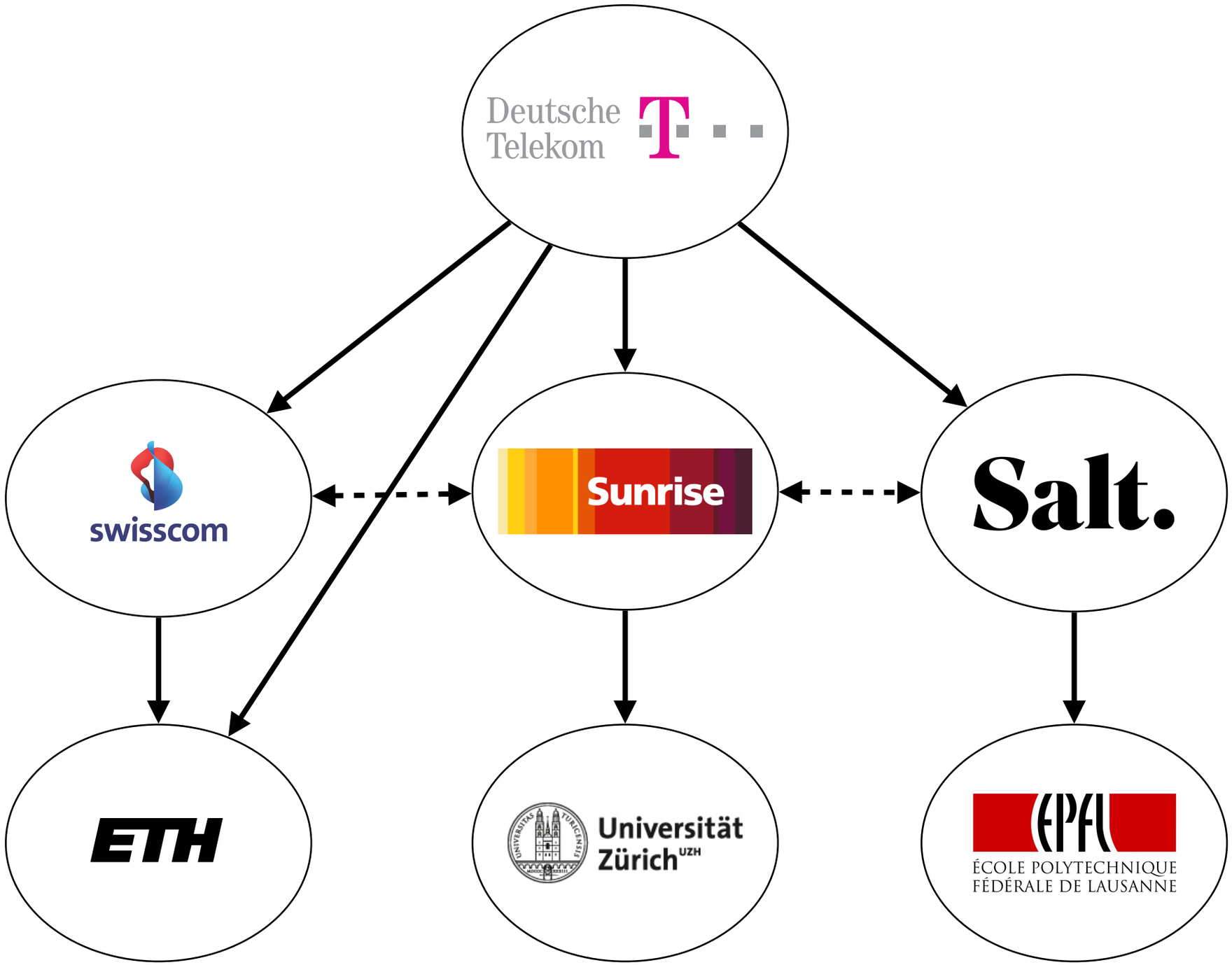


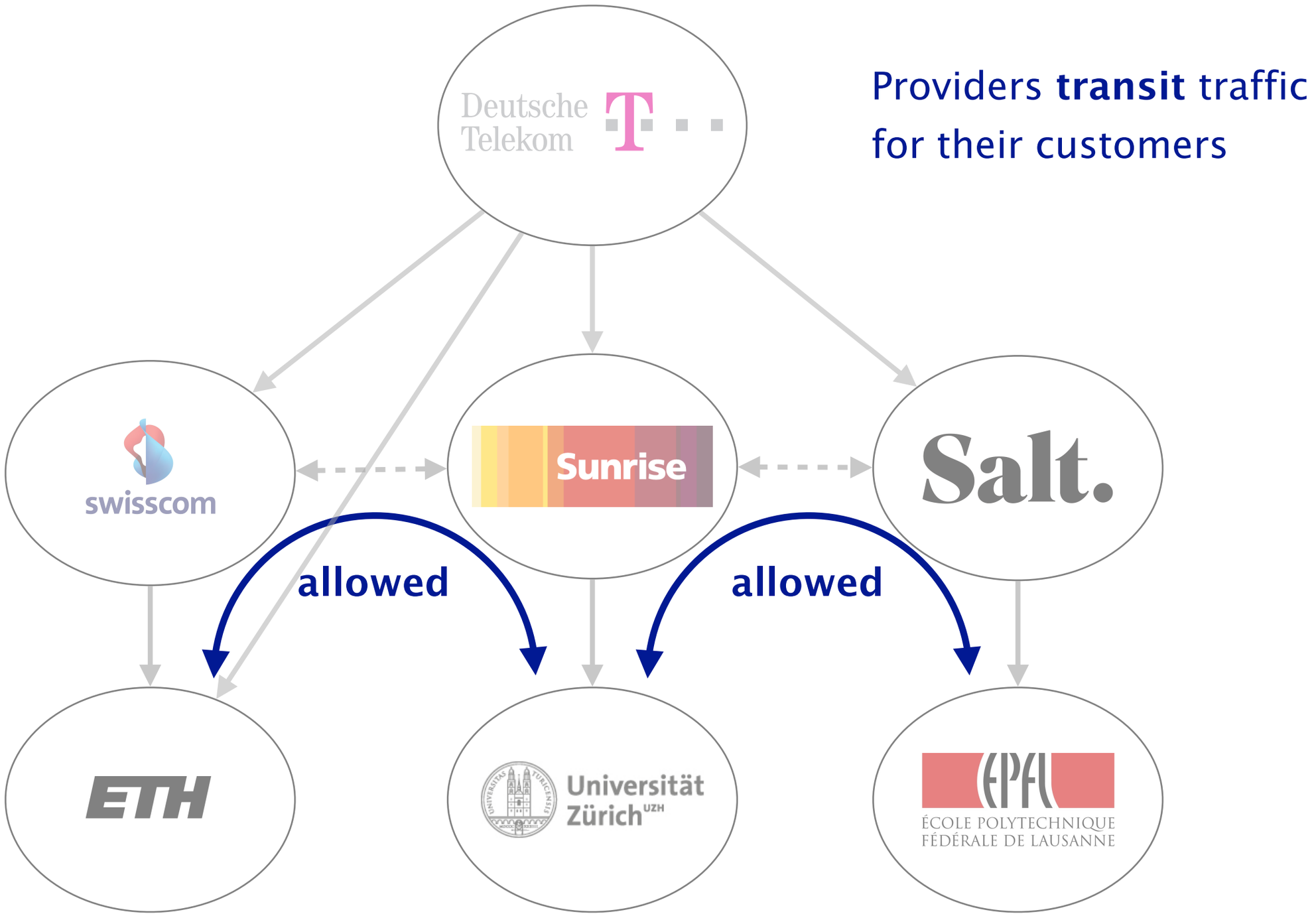
peer peer

DT and ATT exchange *tons* of traffic.
they save money by directly connecting to each other

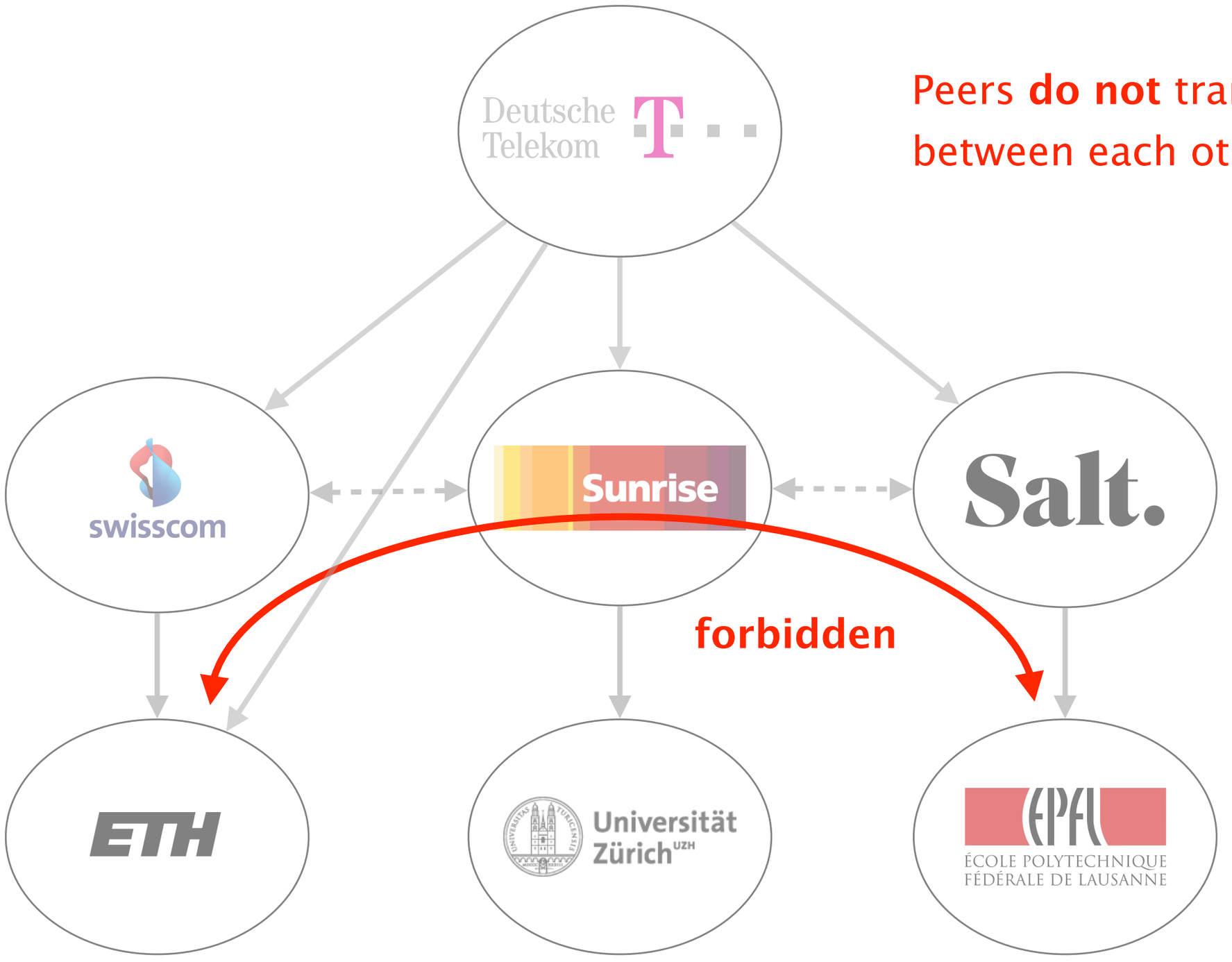
To understand Internet routing,

follow the money

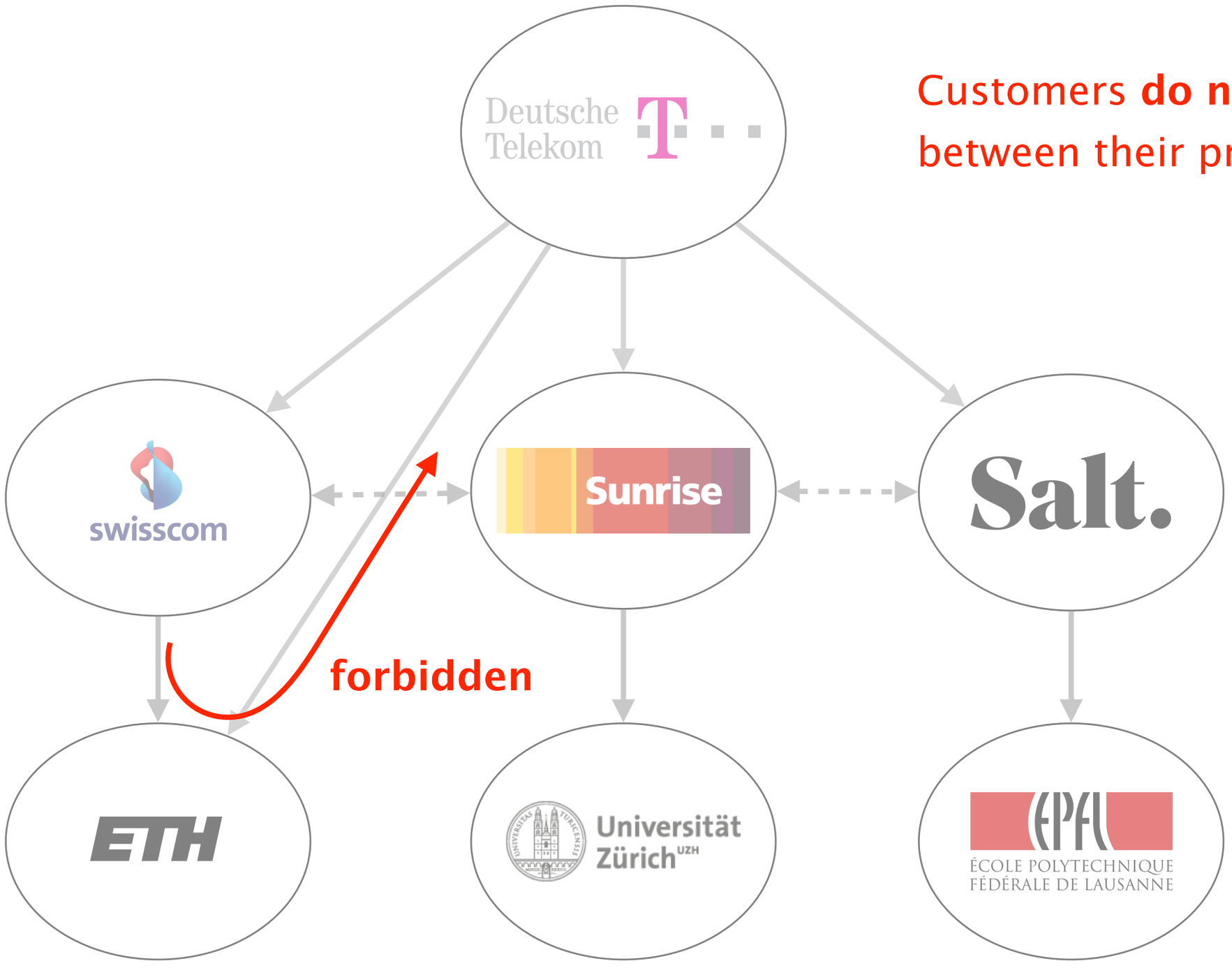

Deutsche Telekom
swisscom
Sunrise
Salt.
ETH
Universität Zürich[UZH]
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE



Providers **transit** traffic
for their customers



Peers **do not** transit traffic between each other



Customers **do not** transit traffic between their providers

These policies are defined by constraining which BGP routes are *selected* and *exported*

Selection	Export
which path to use?	which path to advertise?

Selection

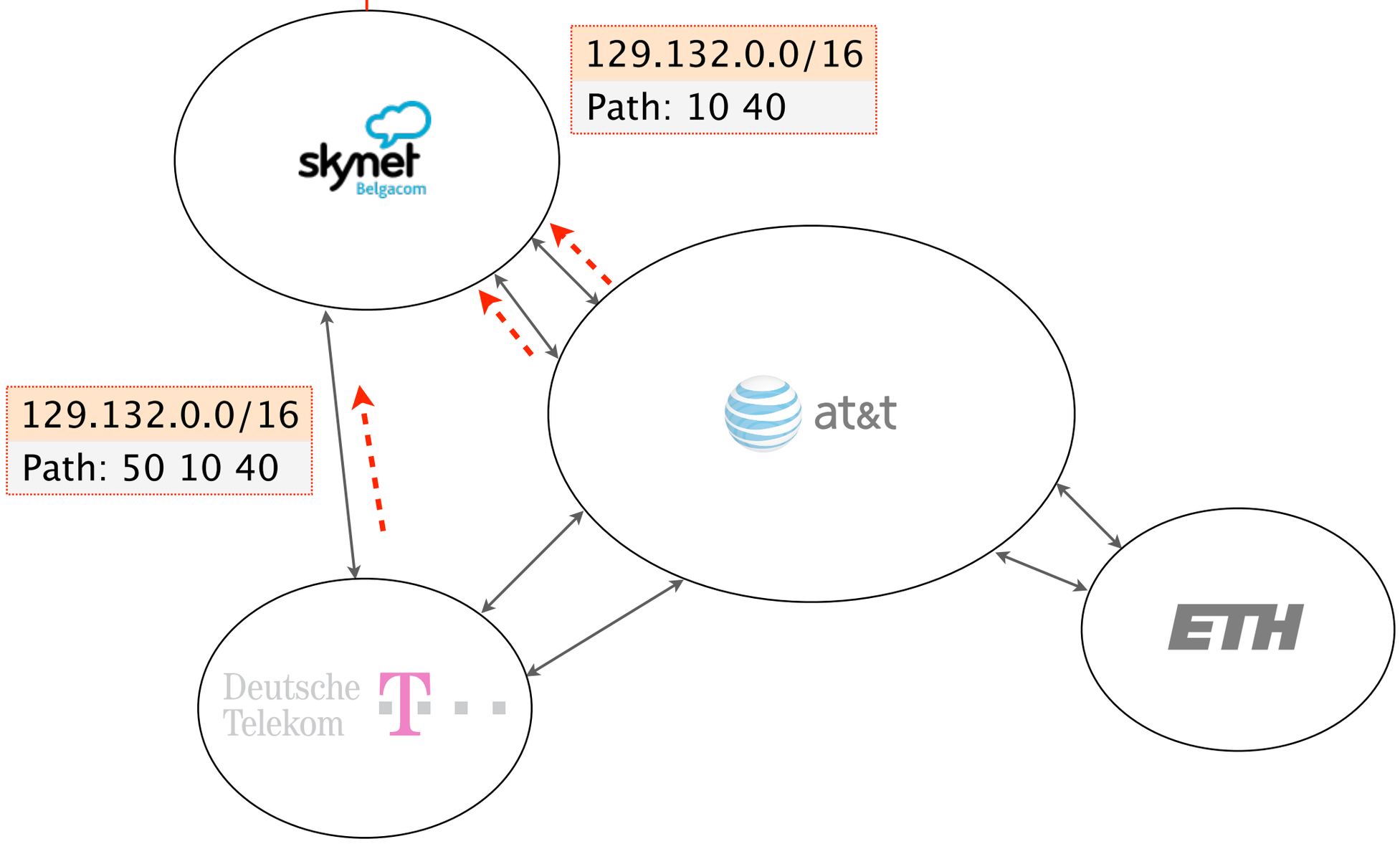
Export

which path to use?

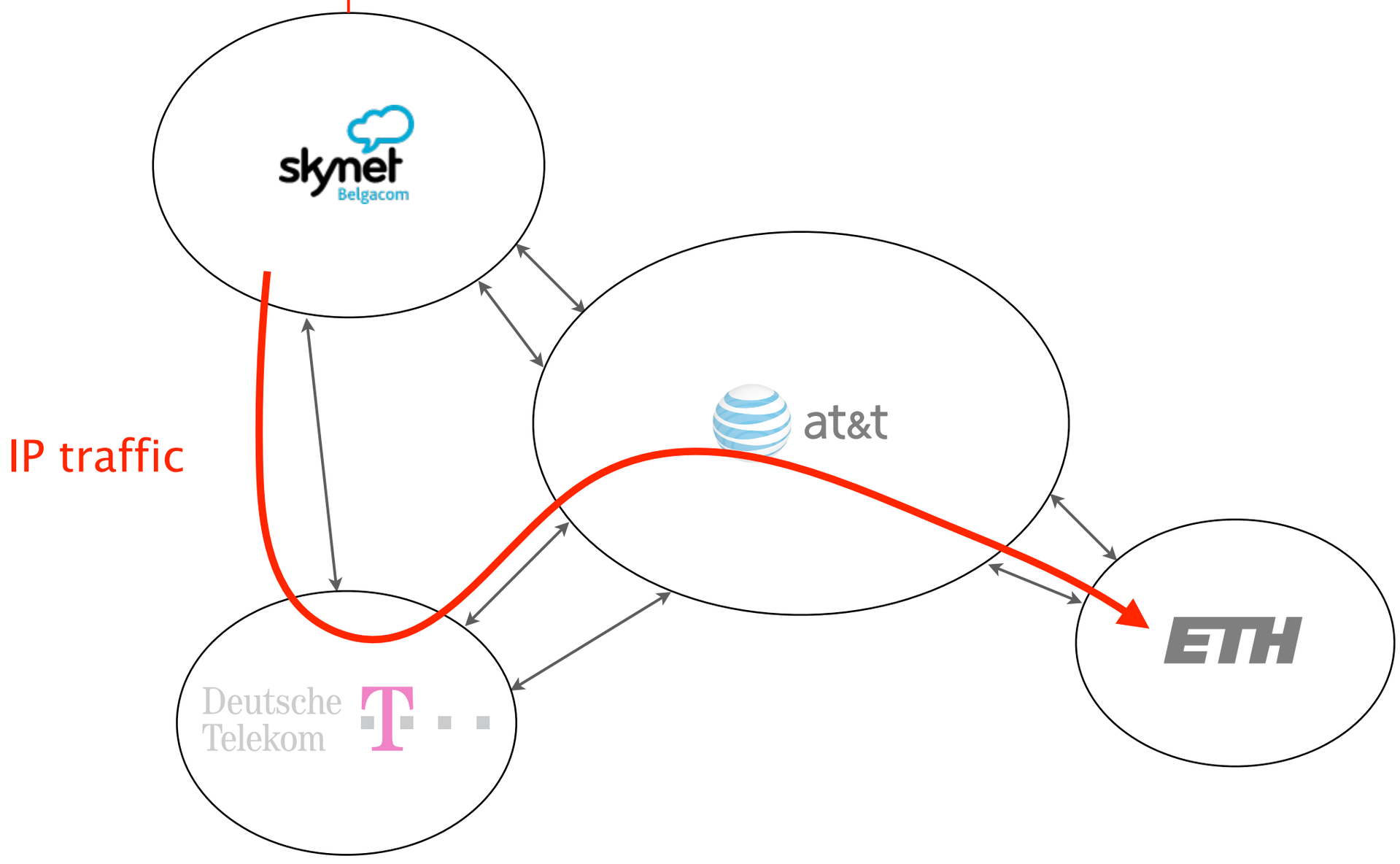
control outbound traffic

which path to advertise?

always prefer Deutsche Telekom routes over AT&T



always prefer Deutsche Telekom routes over AT&T



IP traffic

Business relationships conditions

route selection

For a destination *p*, prefer routes coming from

- customers over
- peers over
- providers

route type

Selection

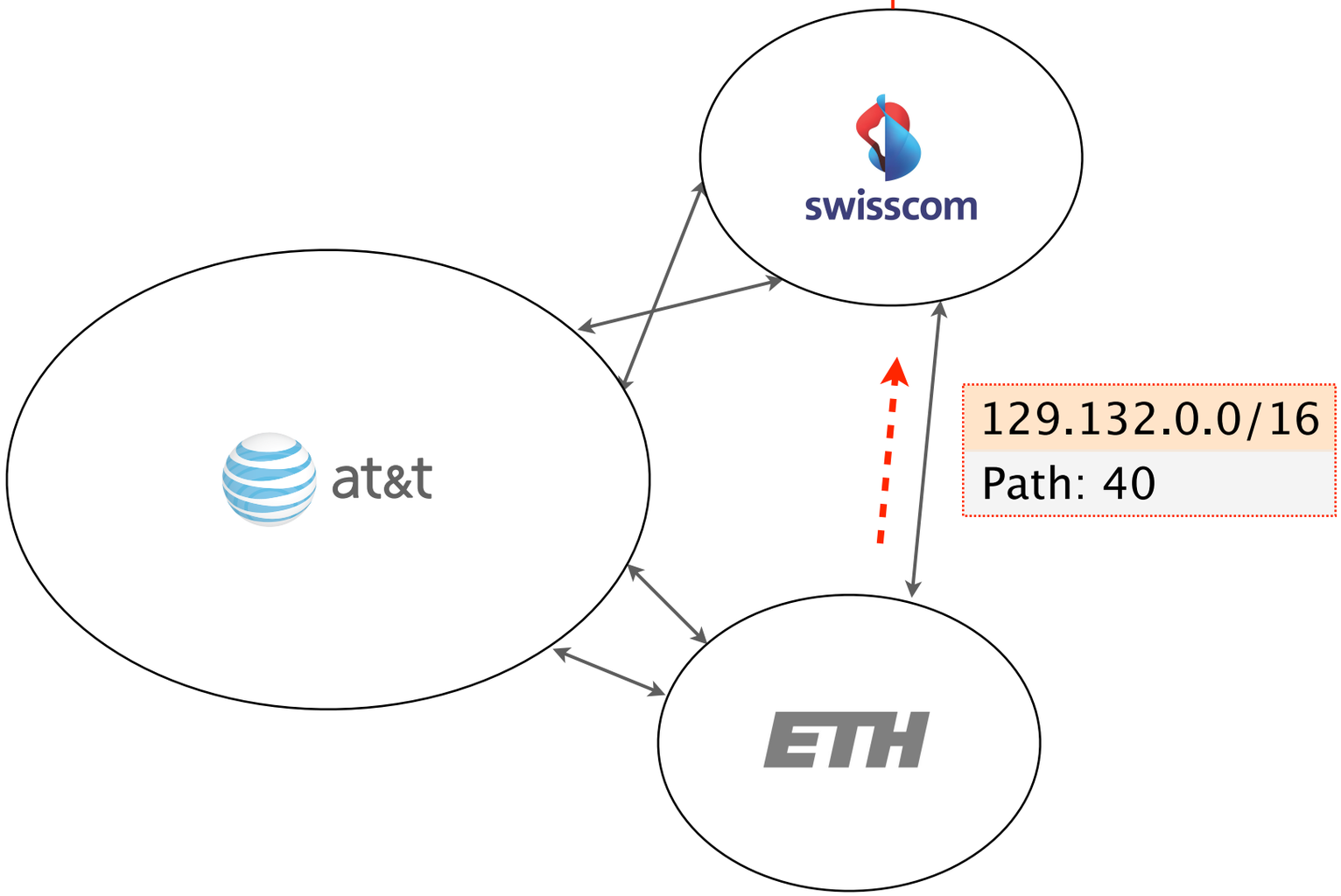
Export

which path to use?

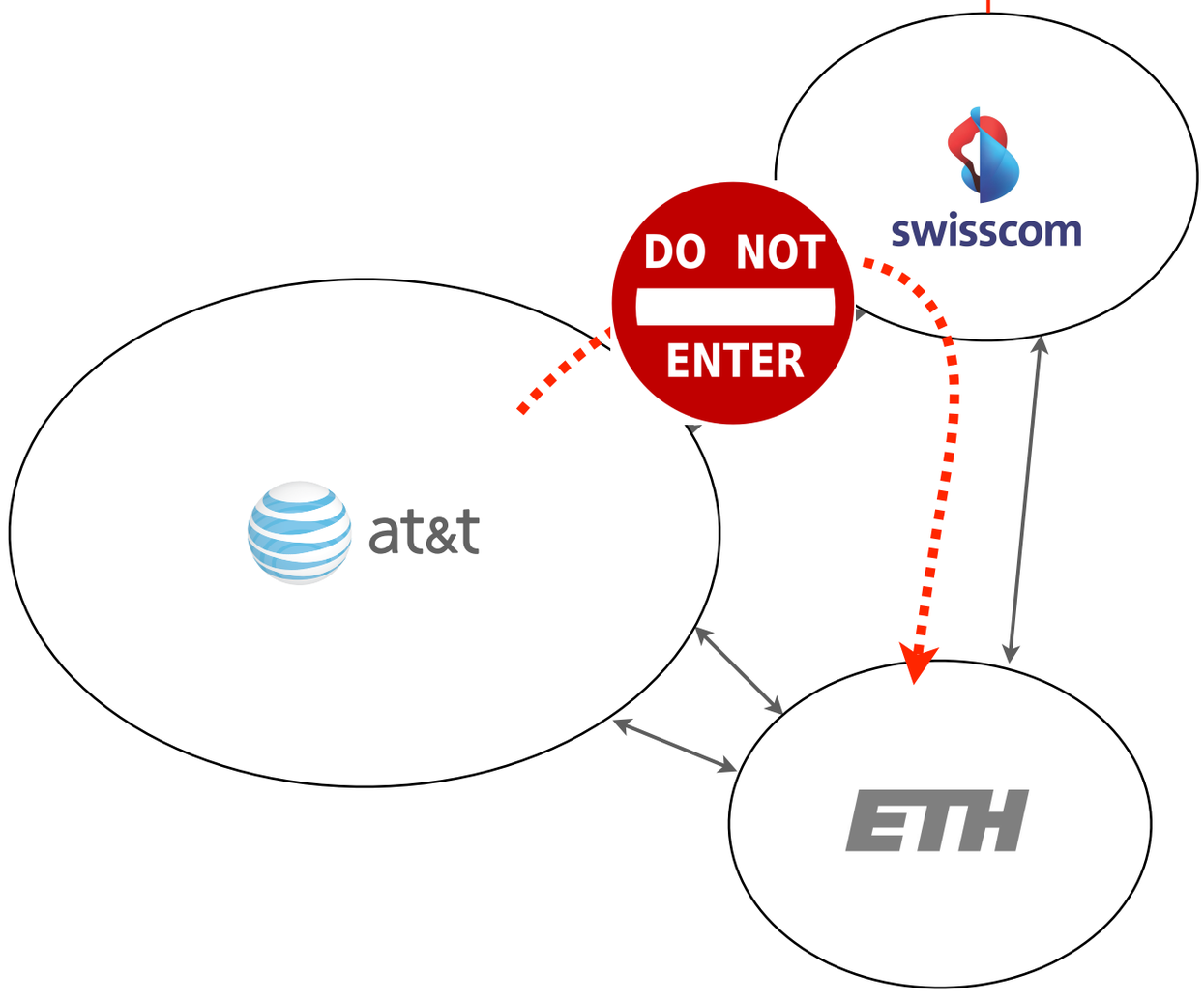
which path to advertise?

control inbound traffic

do not export ETH routes to AT&T



do not export ETH routes to AT&T



These policies are defined by constraining which BGP routes are *selected* and *exported*

Selection	Export
which path to use?	which path to advertise?

Selection

Export

which path to use?

control outbound traffic

which path to advertise?

Business relationships conditions

route selection

For a destination *p*, prefer routes coming from

- customers over
- peers over
- providers

route type

Selection

Export

which path to use?

which path to advertise?

control inbound traffic

Business relationships conditions

route exportation

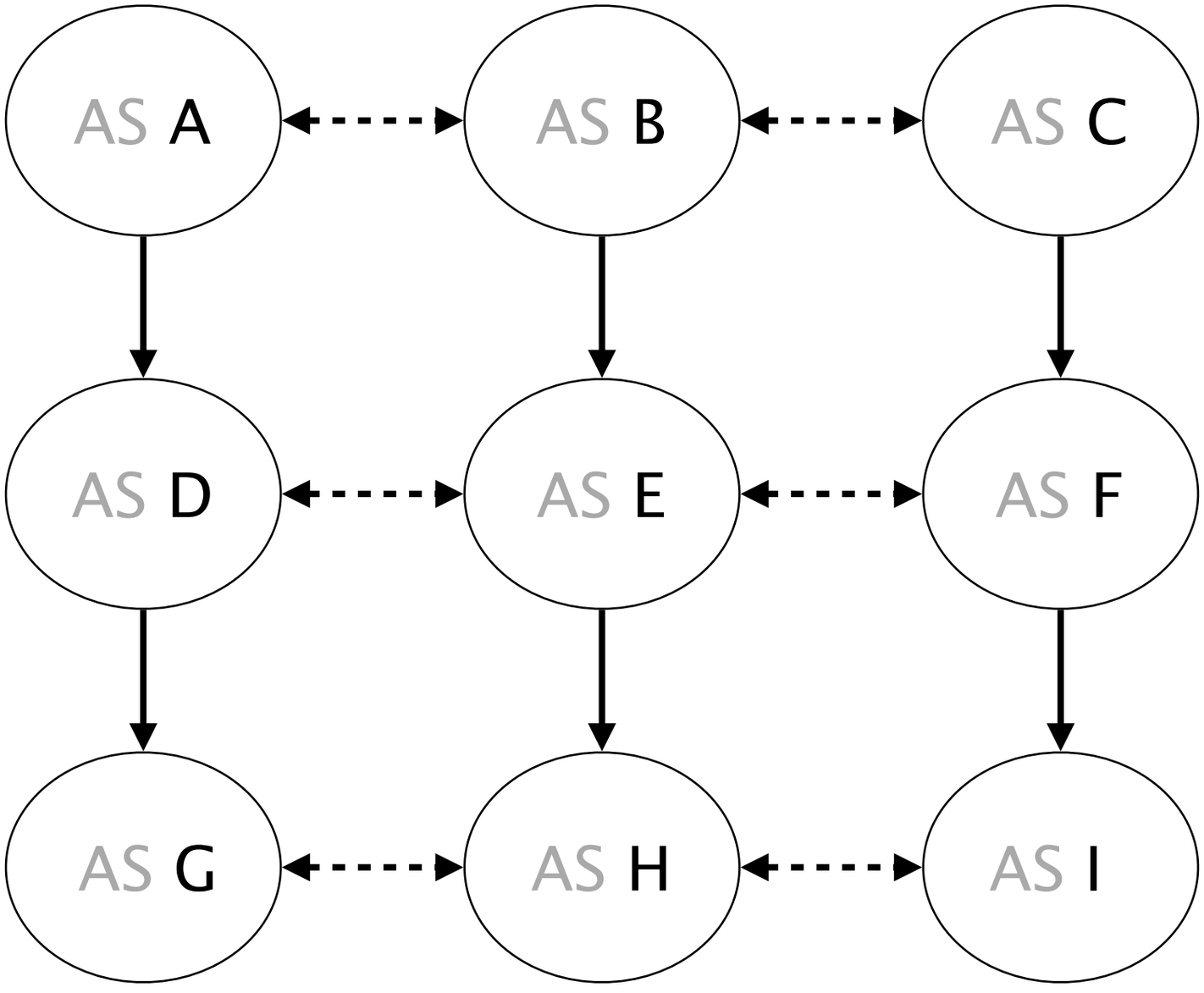
		send to		
		customer	peer	provider
	customer			
from	peer			
	provider			

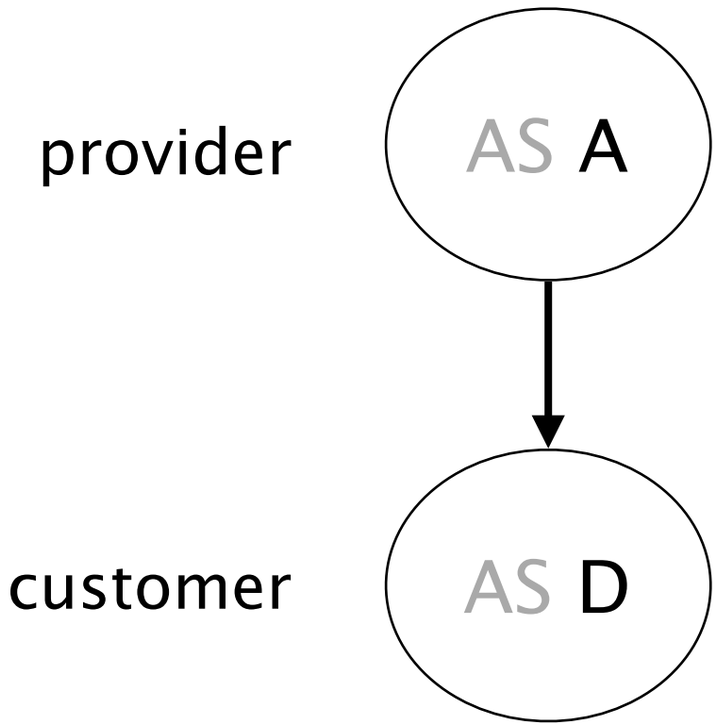
Routes coming from customers
are propagated to everyone else

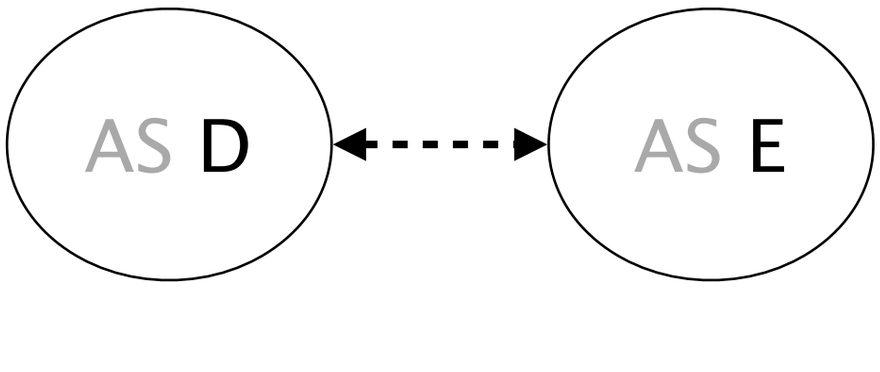
		send to		
		customer	peer	provider
from	customer	✓	✓	✓
	peer			
	provider			

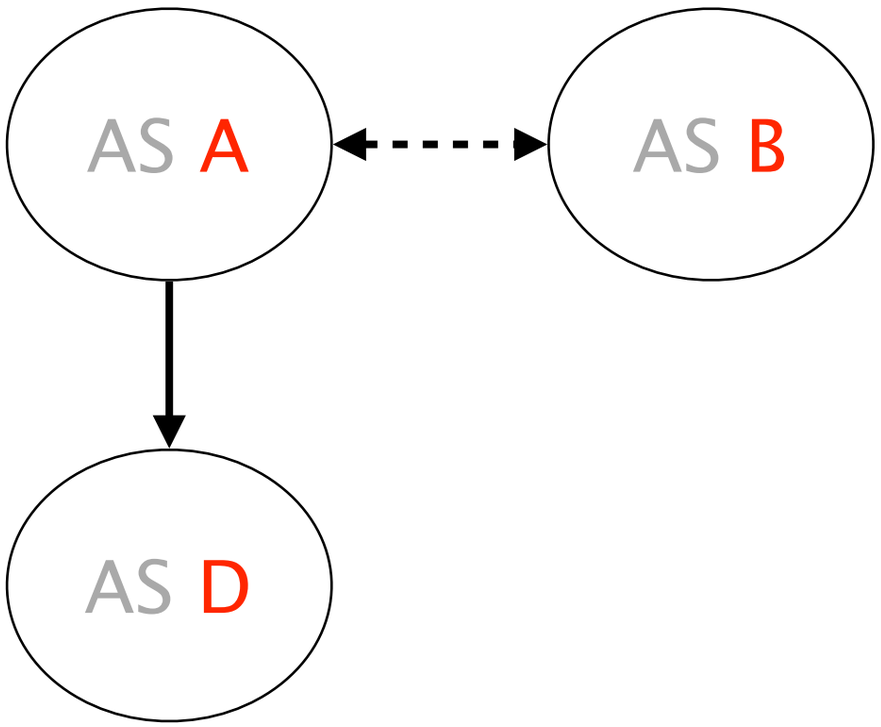
Routes coming from peers and providers are only propagated to customers

		send to		
		customer	peer	provider
	customer	✓	✓	✓
from	peer	✓	-	-
	provider	✓	-	-

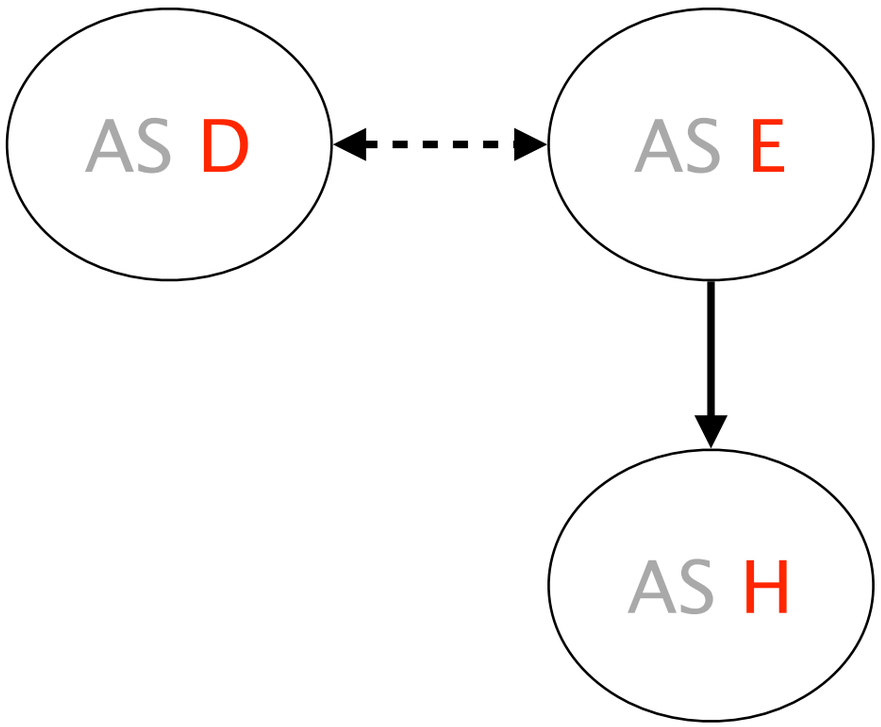

AS A
AS B
AS C
AS D
AS E
AS F
AS G
AS H
AS I


provider
AS A
customer
AS D

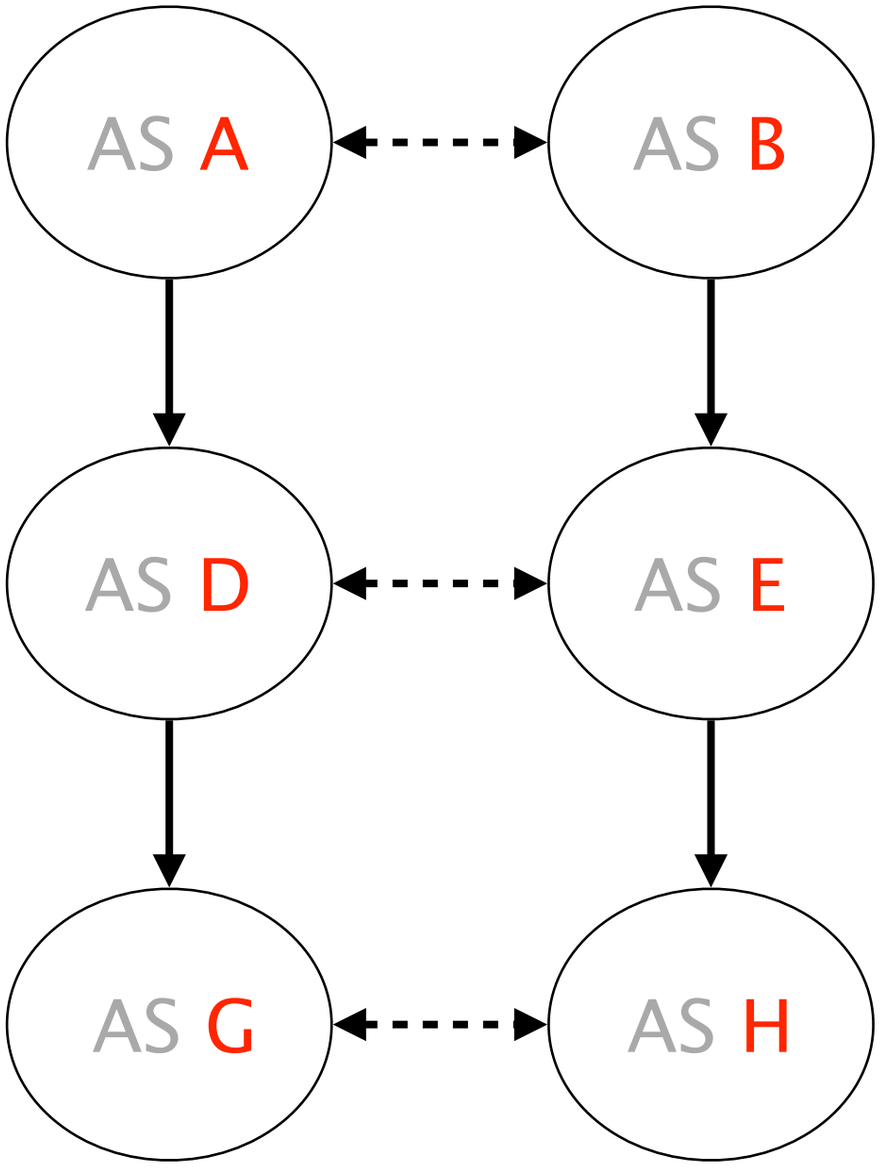

AS D
AS E
peer
peer



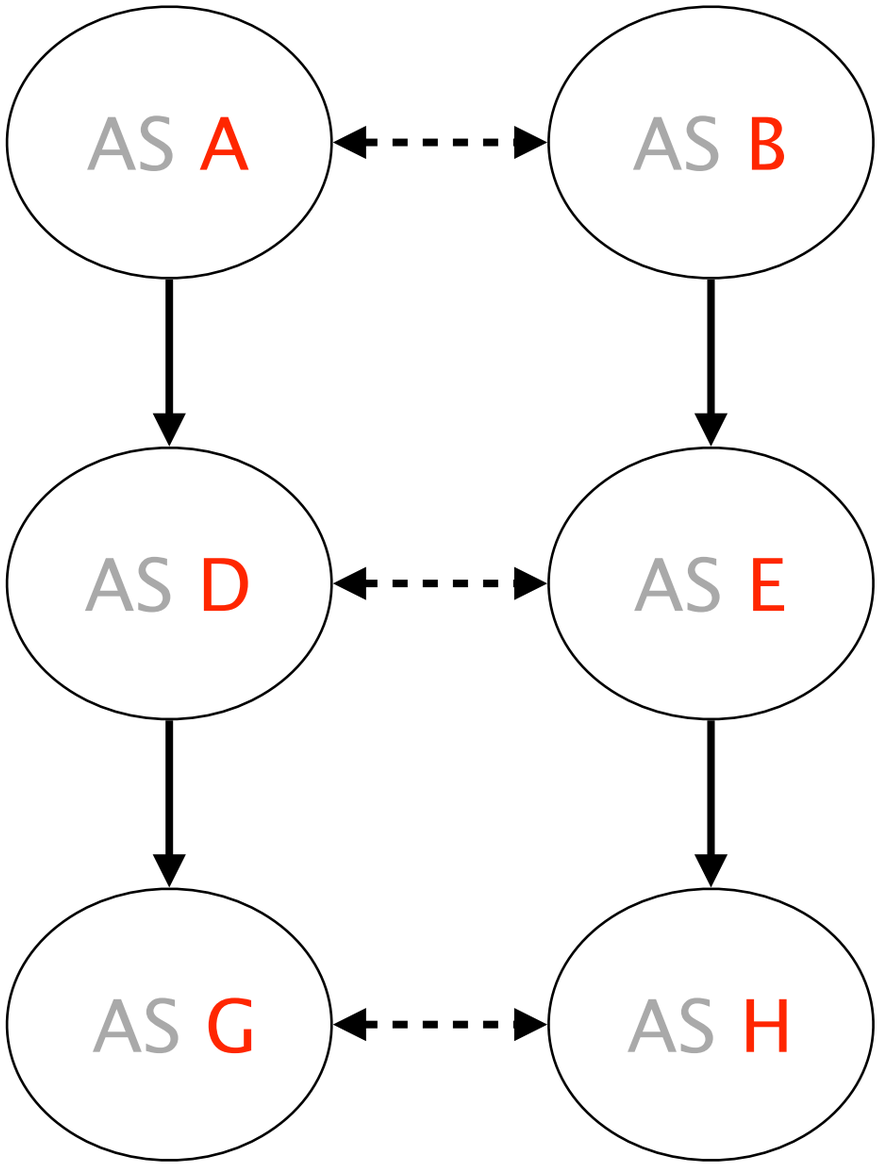
Is (B, A, D) a valid path? Yes/No



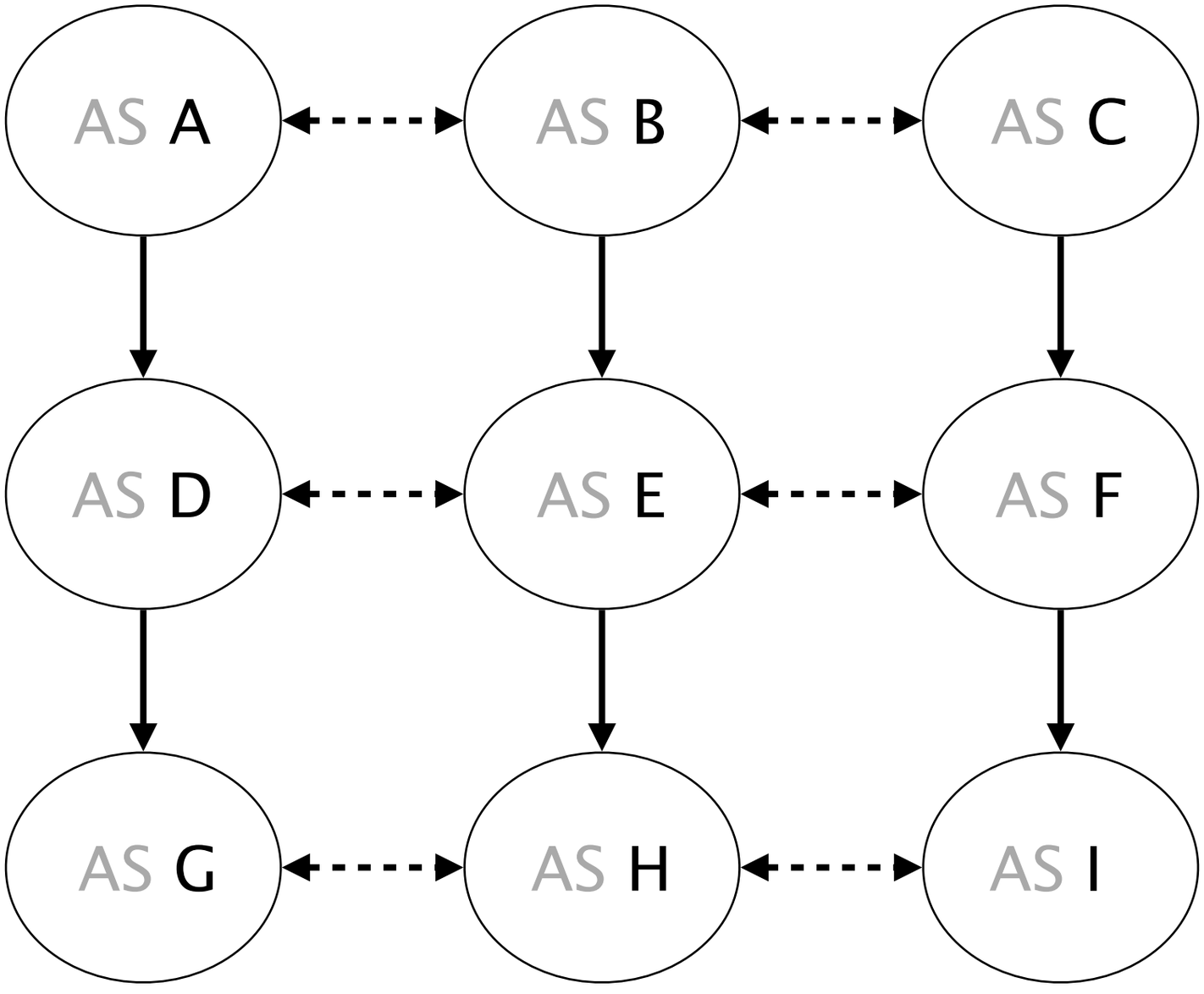
Is (H, E, D) a valid path? Yes/No



Is (G,D,A,B,E,H) a valid path? Yes/No



Will (G,D,A,B,E,H) actually see packets? Yes/No



What's a valid path between G and I?

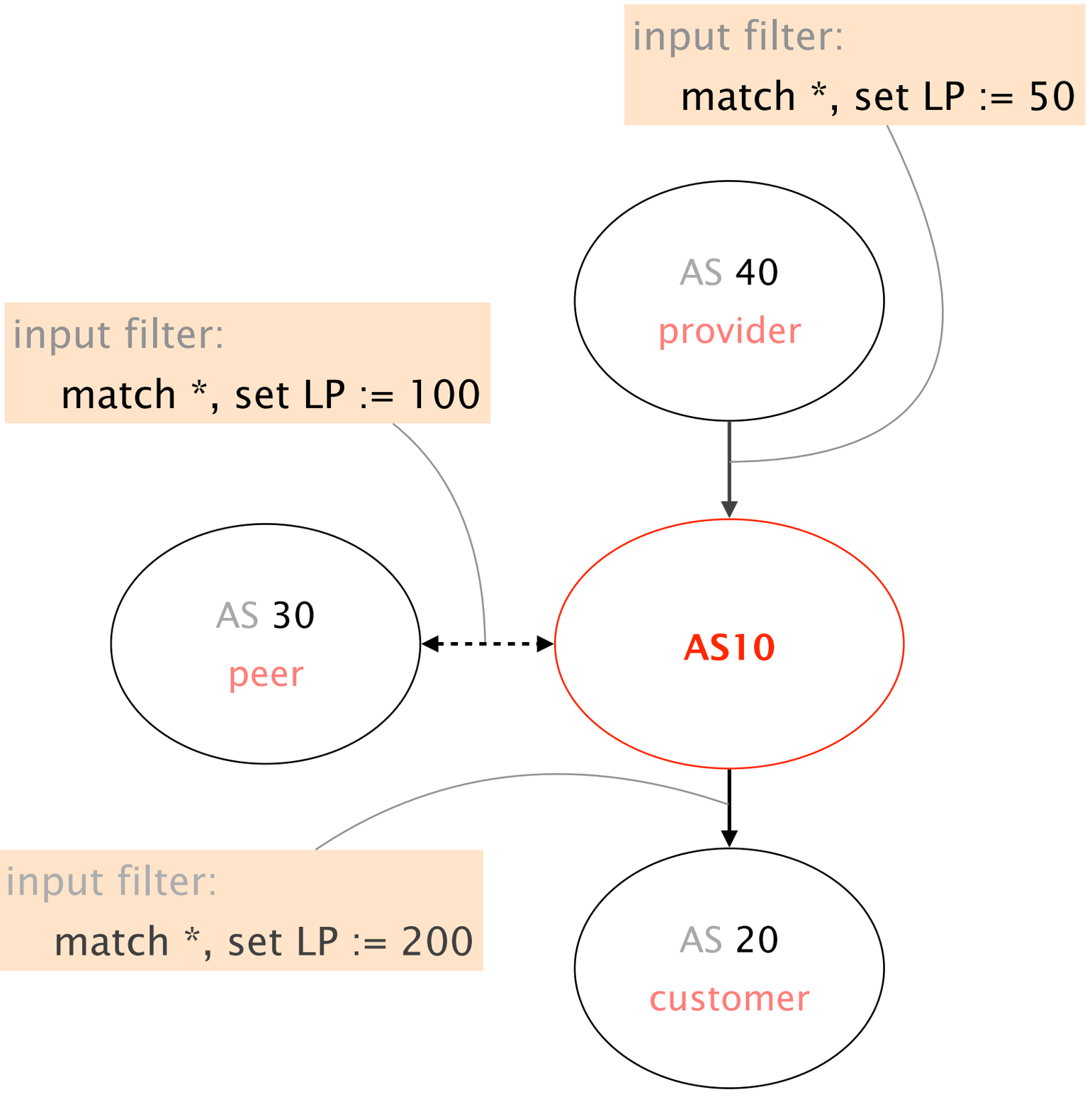
Let’s look at how operators implement customer/provider and peer policies in practice

To implement their selection policy, operators define input filters which manipulates the LOCAL-PREF

For a destination *p*, prefer routes coming from

- customers over
- peers over
- providers

route type

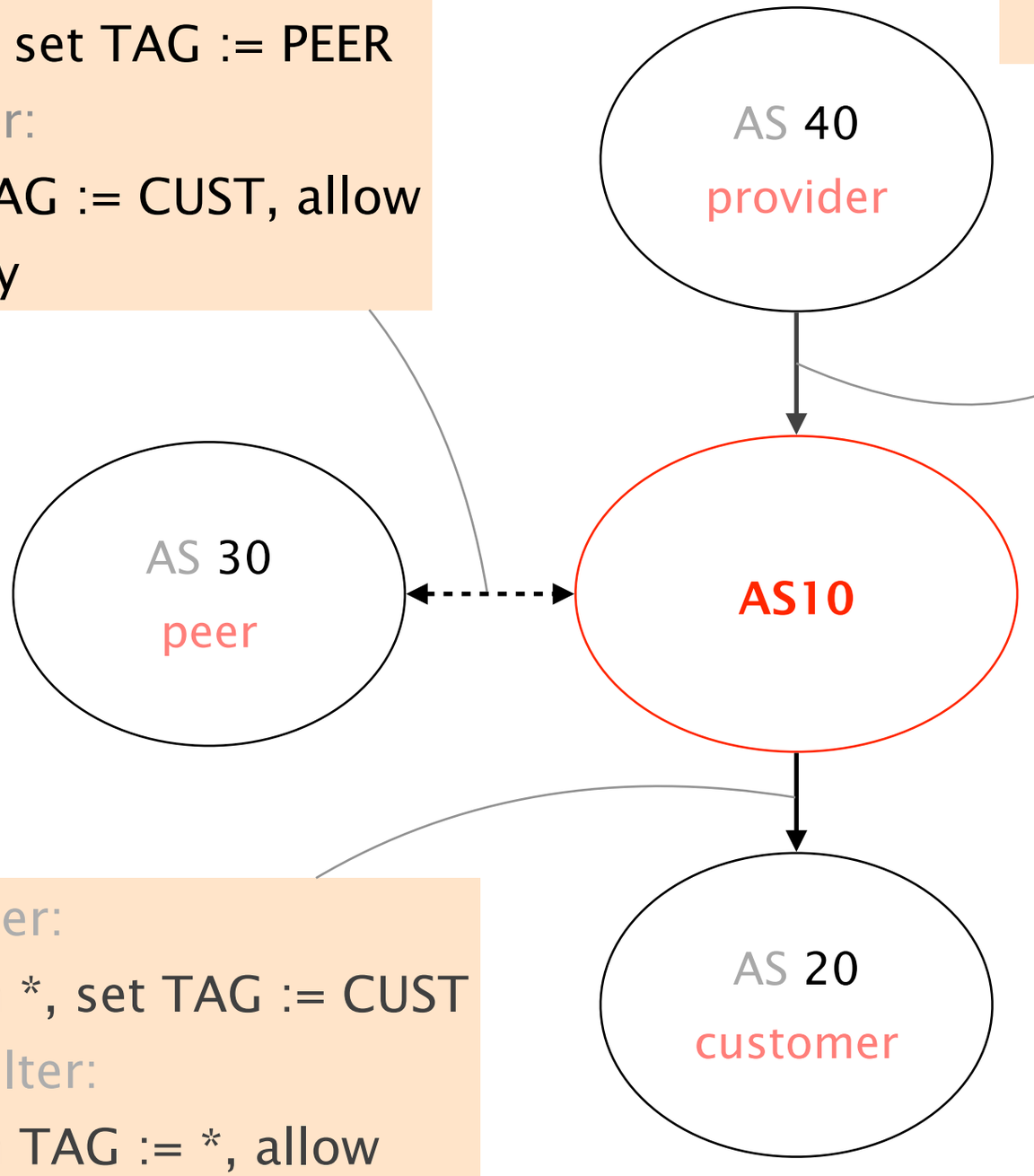

input filter:
match *, set LP := 50
AS 40
provider
input filter:
match *, set LP := 100
AS 30
peer
AS10
input filter:
match *, set LP := 200
AS 20
customer

To implement their exportation rules,
operators use a mix of import and export filters

		send to		
		customer	peer	provider
from	customer	✓	✓	✓
	peer	✓	-	-
	provider	✓	-	-

input filter:
match *, set TAG := PEER
output filter:
match TAG := CUST, allow
else deny

input filter:
match *, set TAG := PROV
output filter:
match TAG := CUST, allow
else deny



input filter:
match *, set TAG := CUST
output filter:
match TAG := *, allow

Communication Networks

Spring 2022

Laurent Vanbever

nsg.ee.ethz.ch

ETH Zürich (D-ITET)

April 4 2022