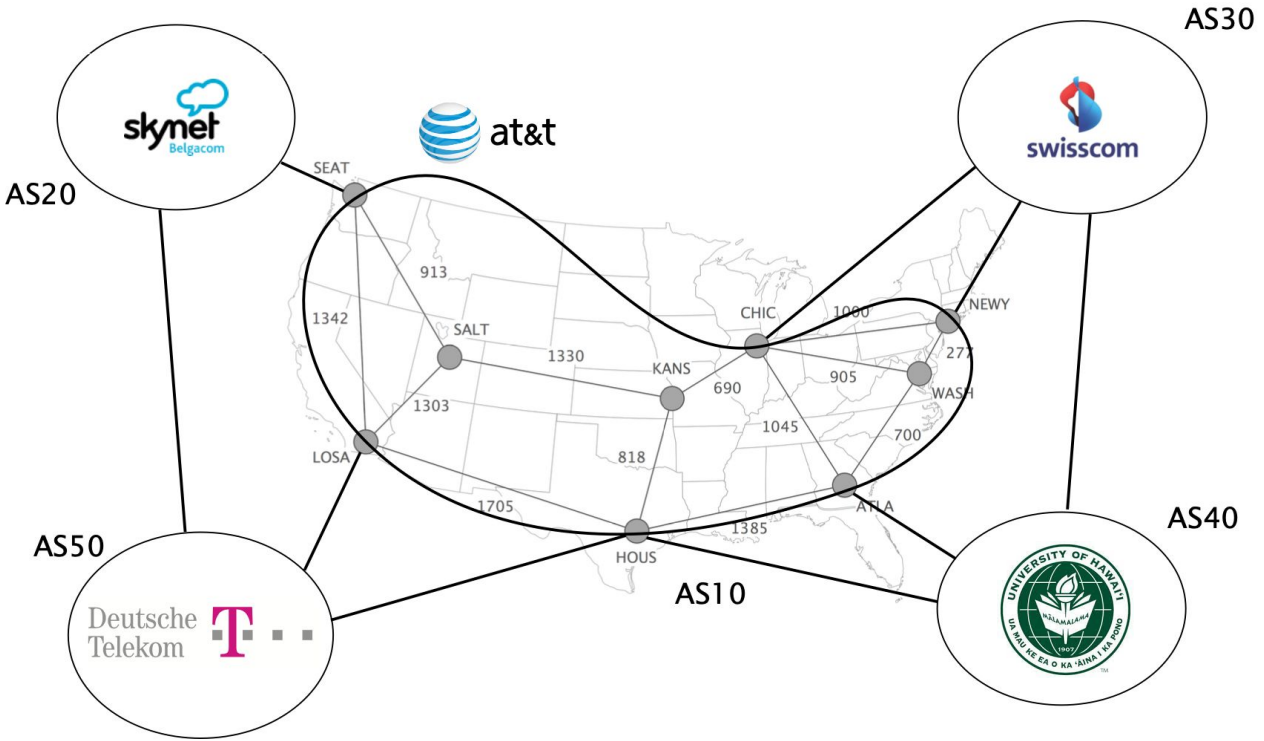
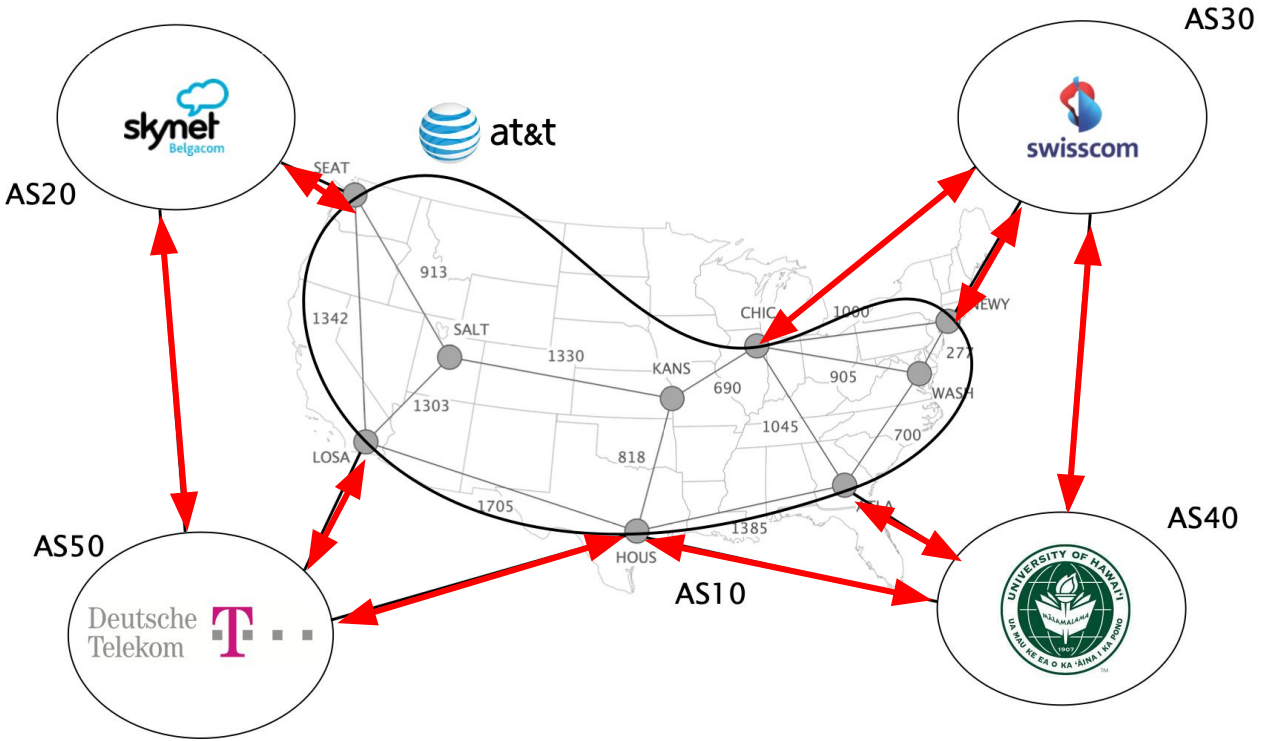
BGP Protocol

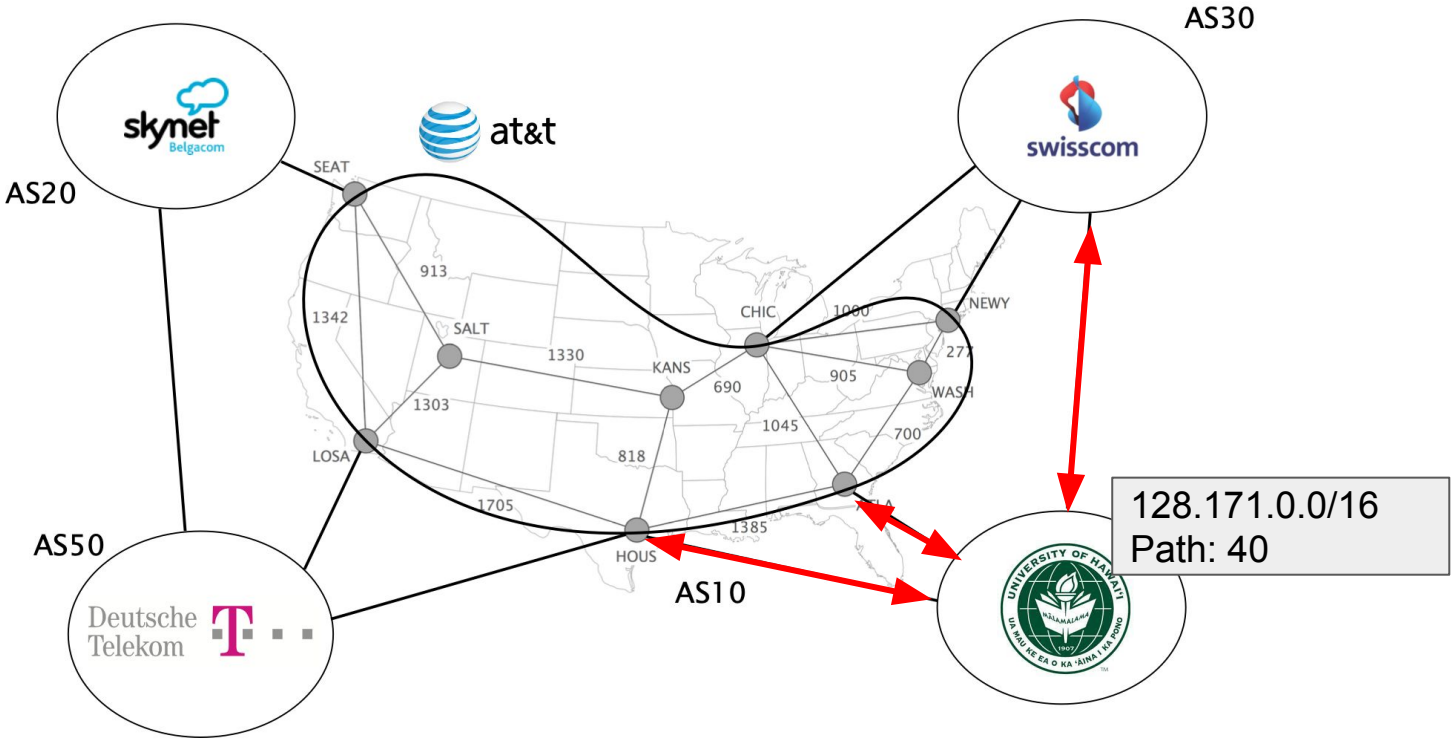
BGP Comes in Two Flavors



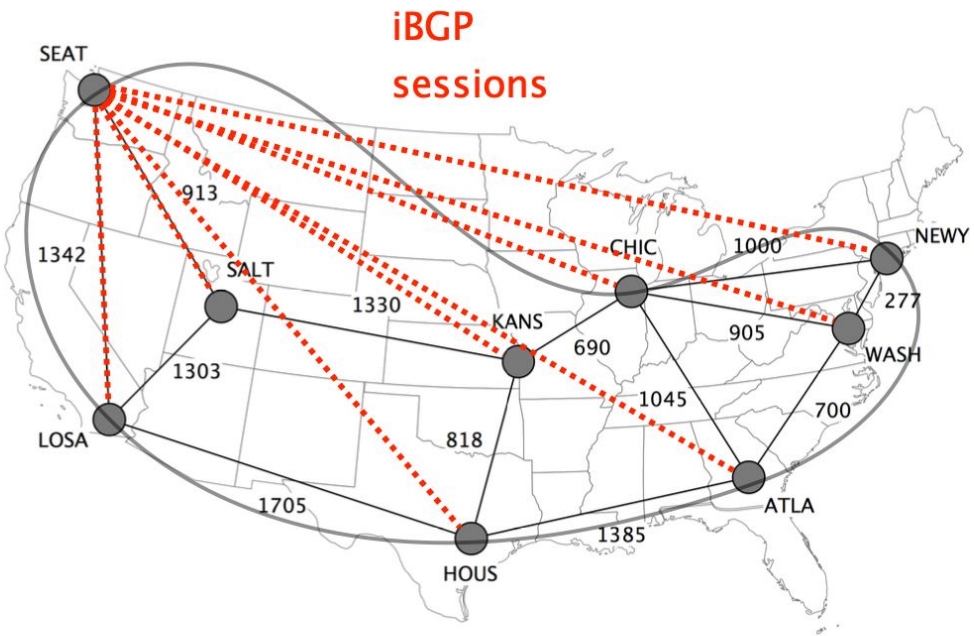
External BGP (eBGP) Sessions Connect Border Routers in Different ASes



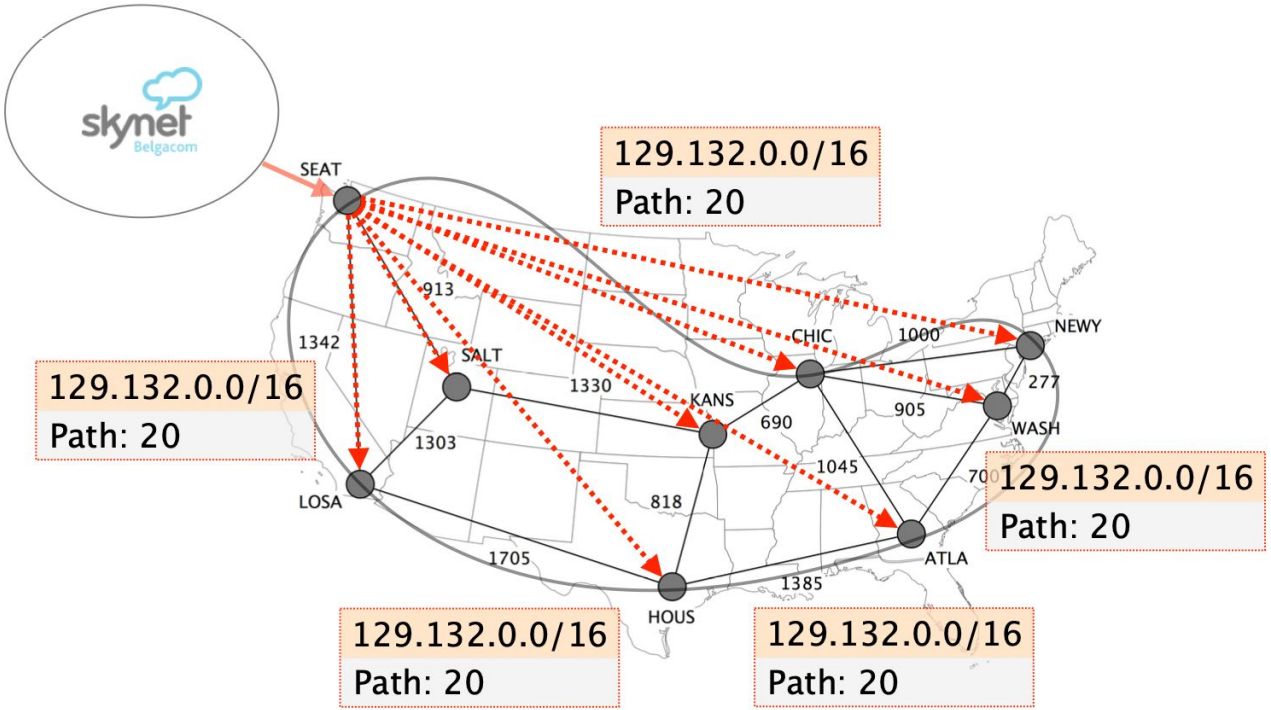
eBGP Sessions are used to Learn Routes to External Destinations



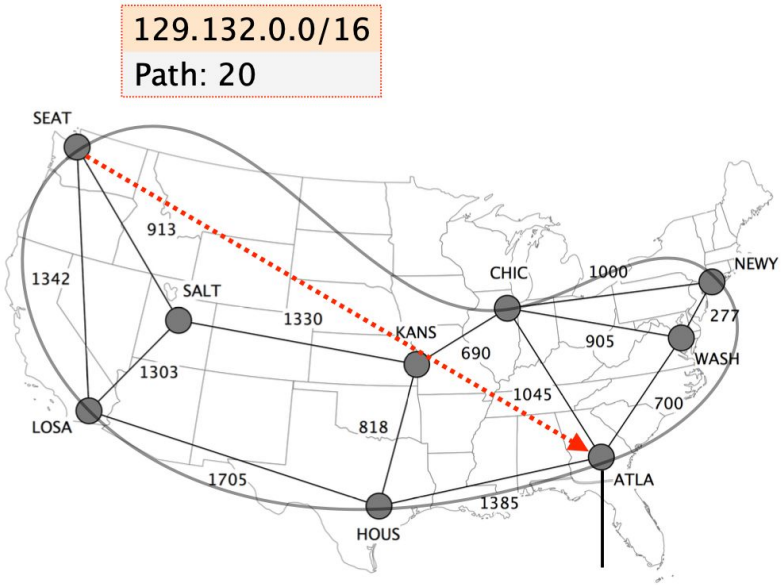
Internal BGP (iBGP) Sessions Connect Routers in the Same AS



iBGP Sessions are used to Disseminate Externally Learn Routes Internally



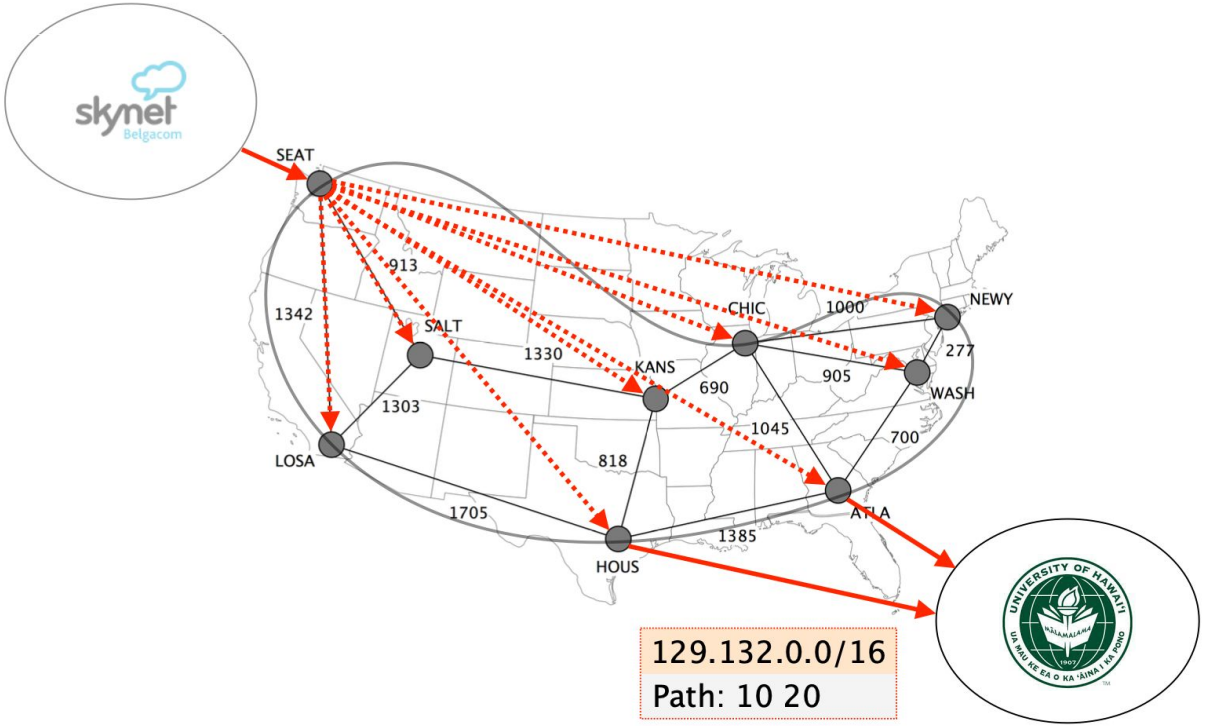
iBGP Sessions are used to Disseminate Externally Learn Routes Internally



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

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

Routes Learned via iBGP are then Announced Externally, using eBGP



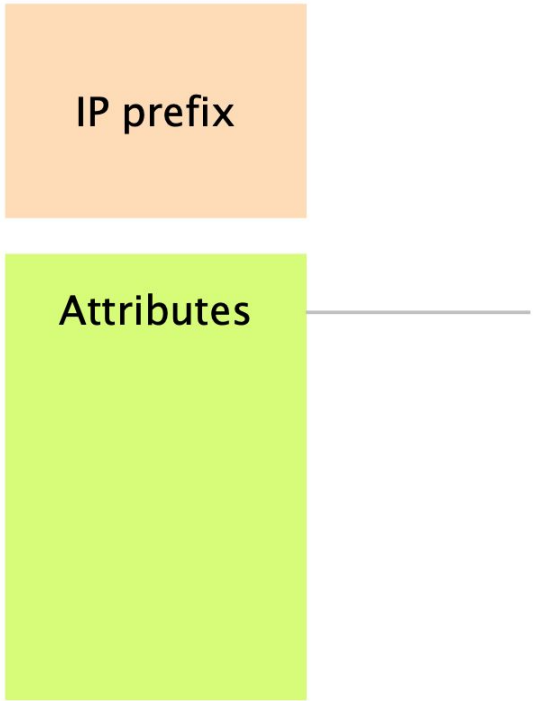
BGP is Simple, Composed of Four Basic Messages

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

BGP is Simple, 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

BGP Updates Carry an IP Prefix and Some Attributes



Describe route properties

used in route selection/exportation decisions

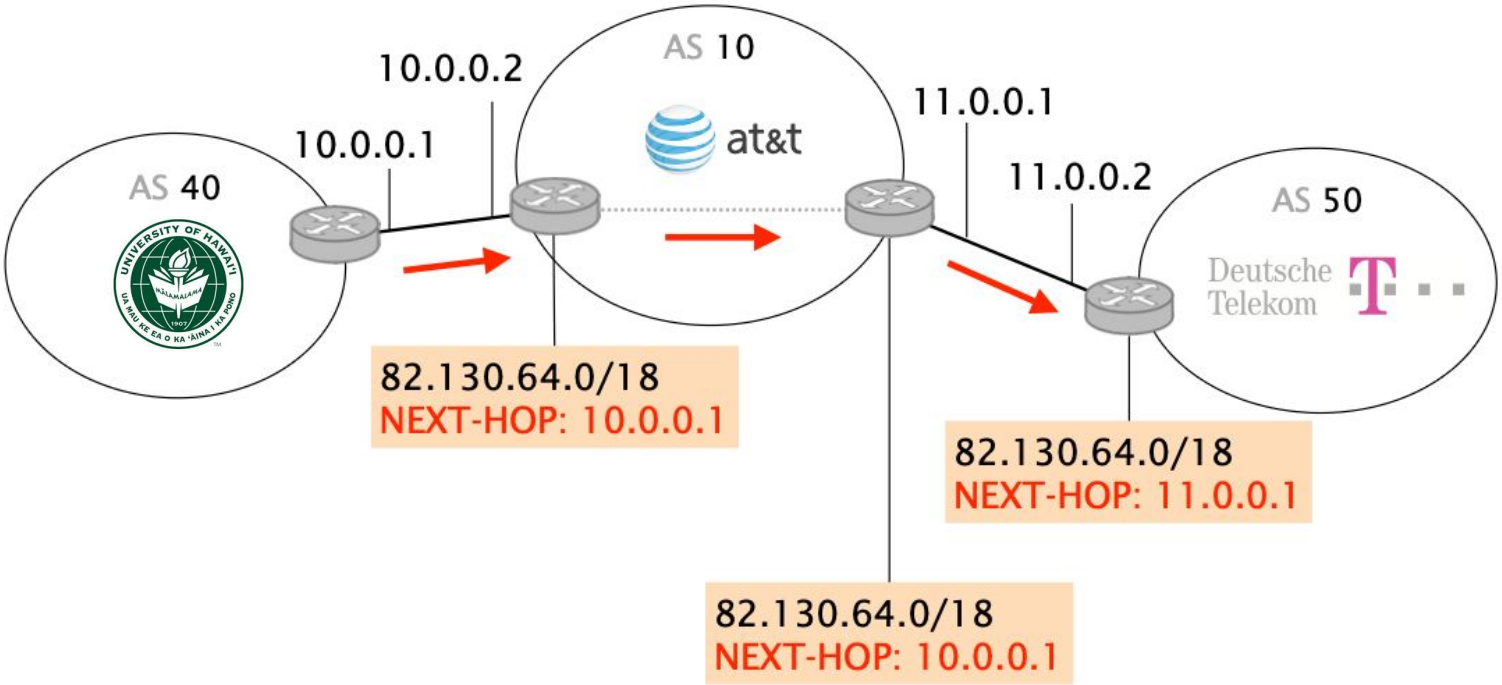
are either local (*only* seen on iBGP)

or global (seen on iBGP *and* eBGP)

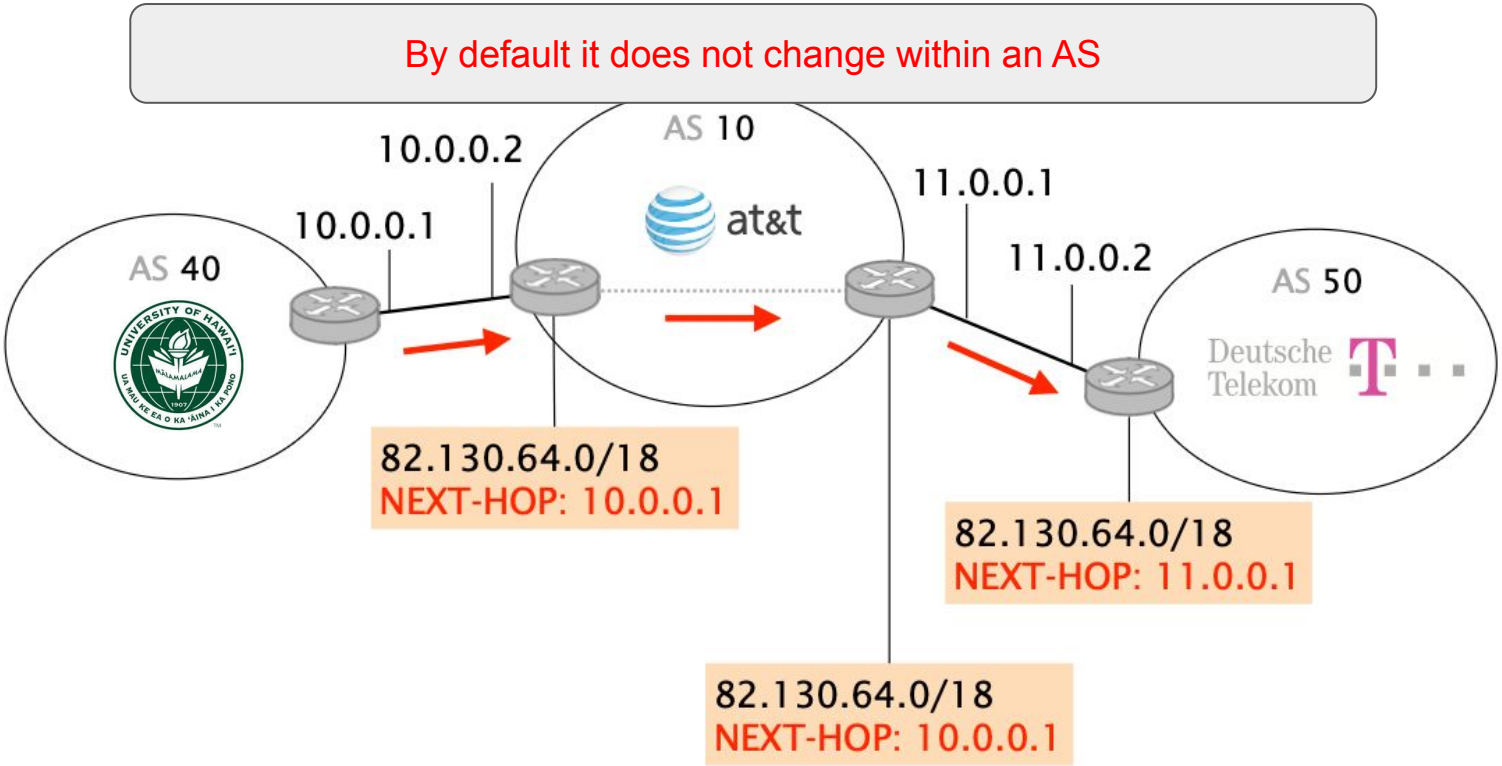
BGP Updates Carry an IP Prefix and Some Attributes

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

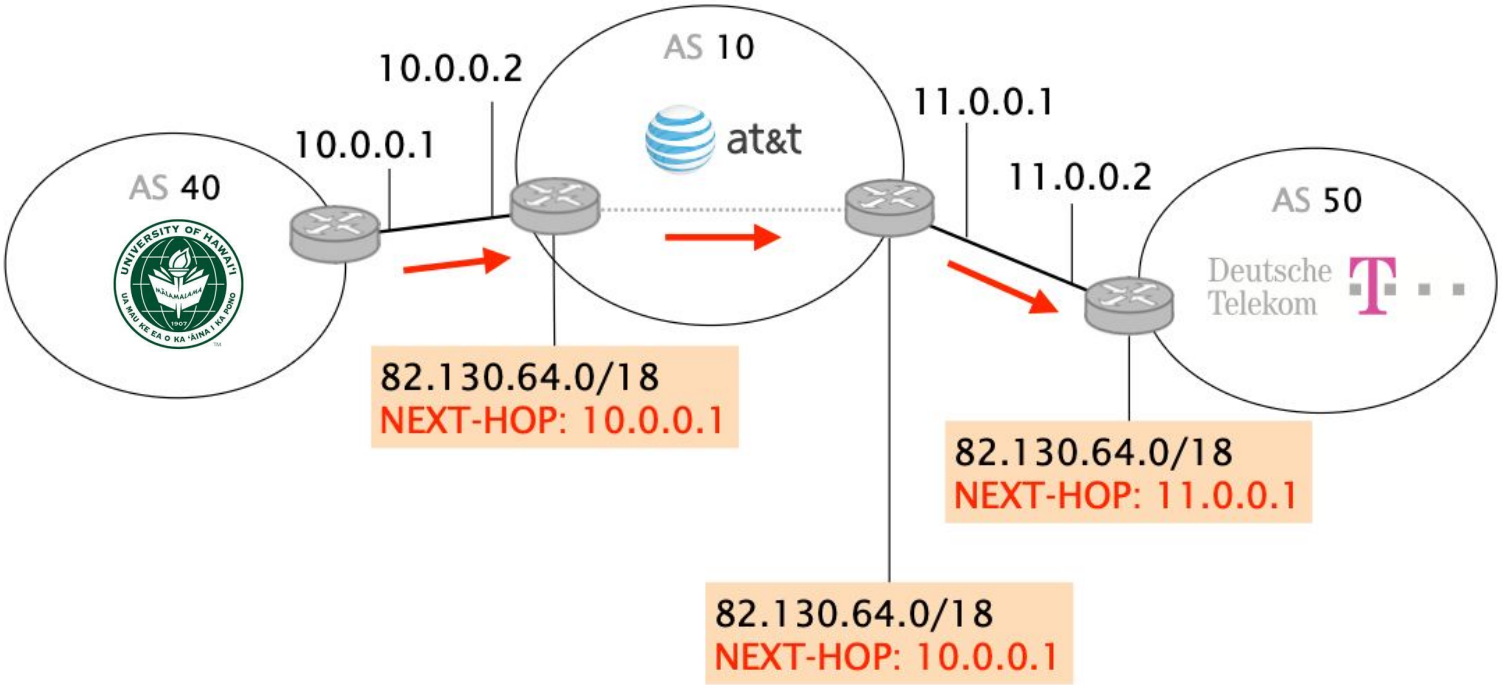
NEXT-HOP: Indicates Where to Send Traffic Next



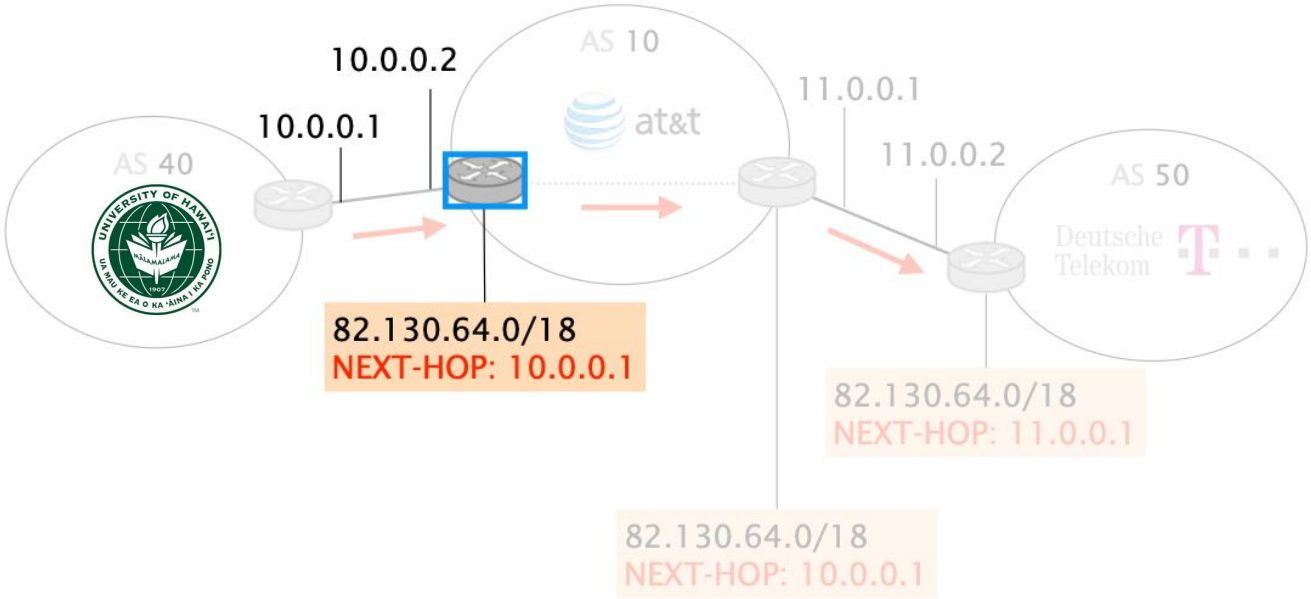
NEXT-HOP: Indicates Where to Send Traffic Next



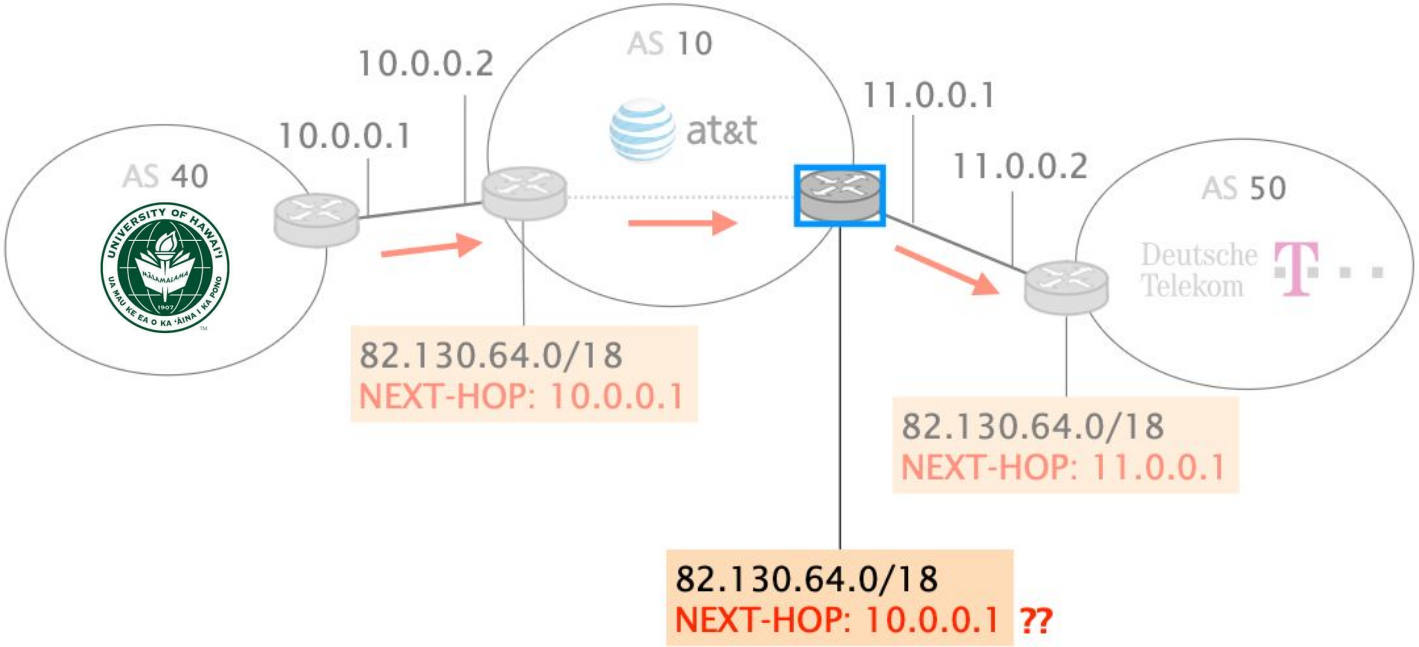
For externally-learned routes, this means that the NEXT-HOP is the IP address of the neighbor's eBGP router, here 10.0.0.1 for at&t



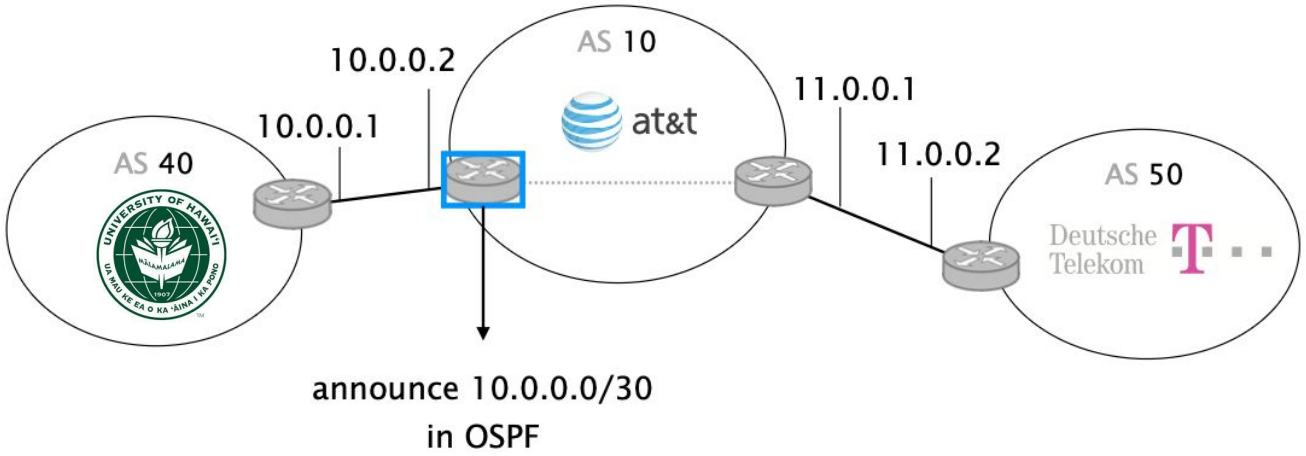
For this router, reaching 10.0.0.1 is not a problem as it is directly connected to the corresponding subnet (10.0.0.0/30)



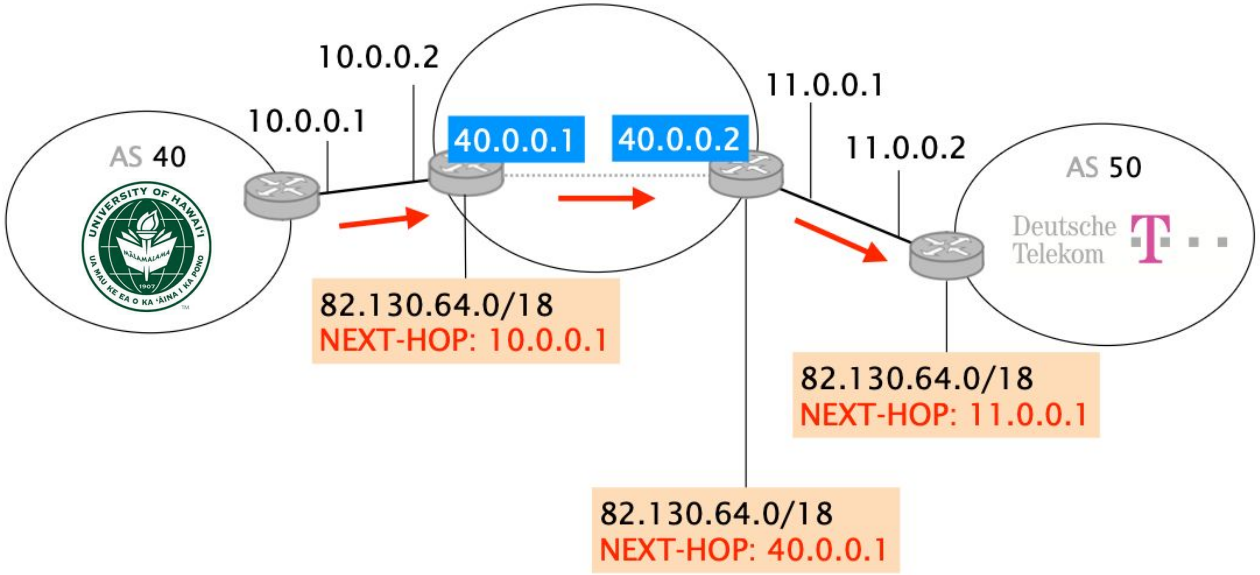
That router is not directly to the NEXT-HOP's subnet (10.0.0.0/30)
and does not know how to reach it, it will therefore drop the BGP route...



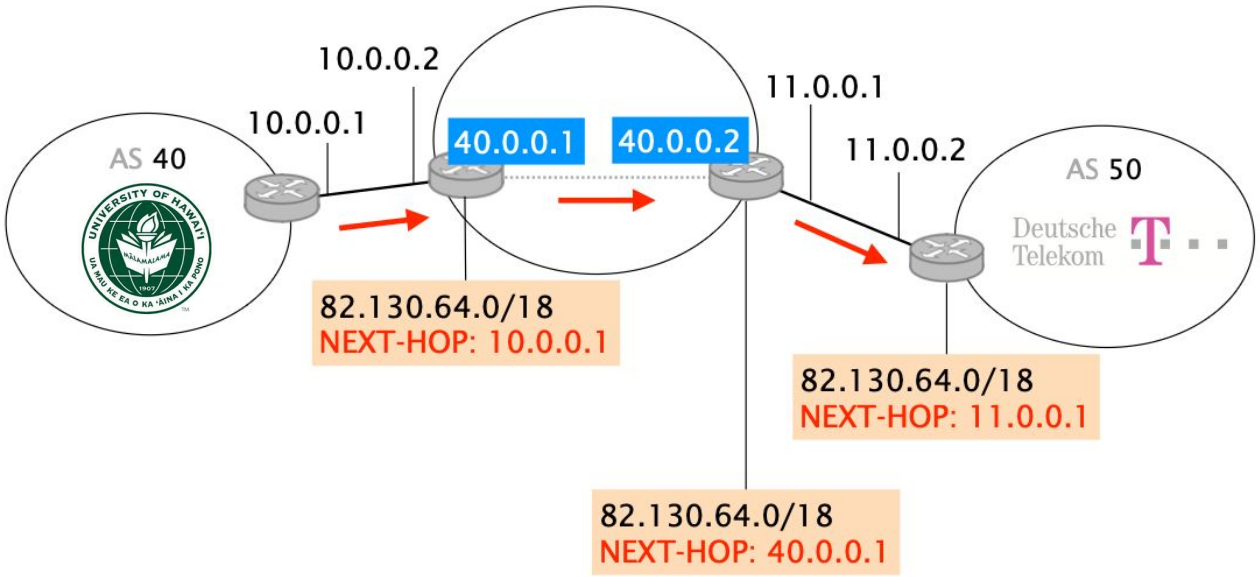
One solution is for the external router to redistribute
the prefixes attached to the external interfaces into the IGP



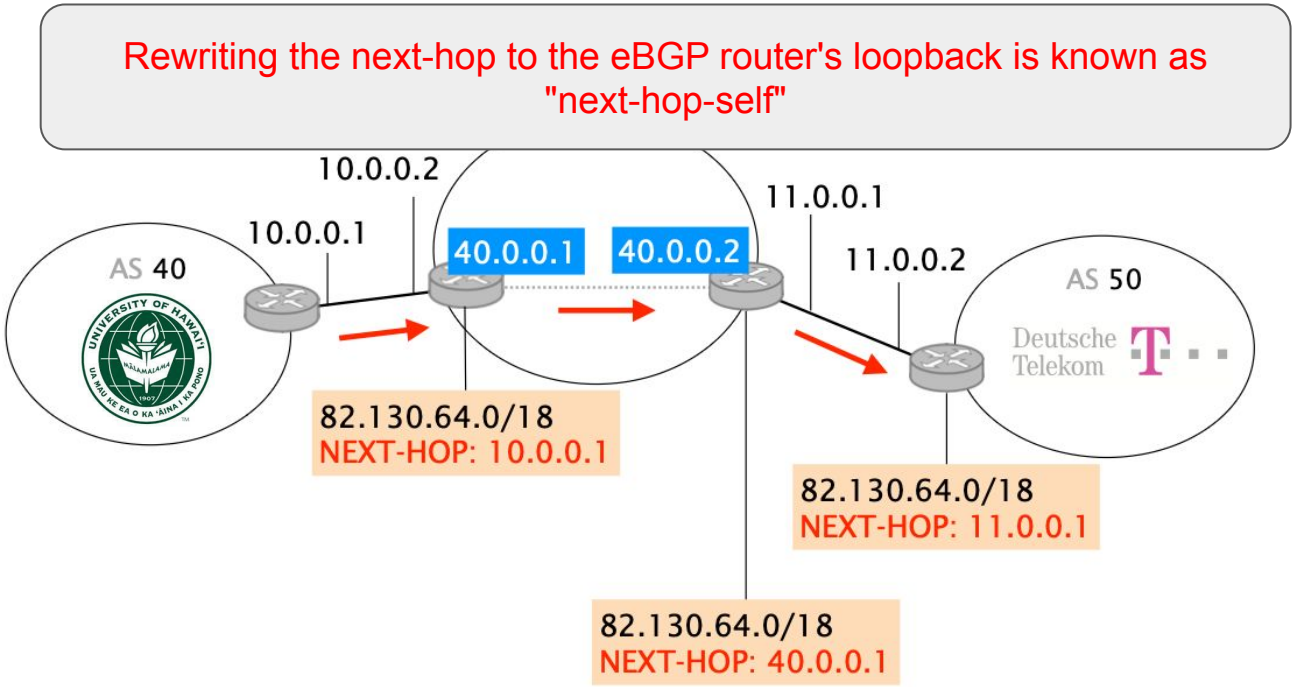
Another solution is for the border router to rewrite the NEXT-HOP before sending it over iBGP, usually to its loopback address



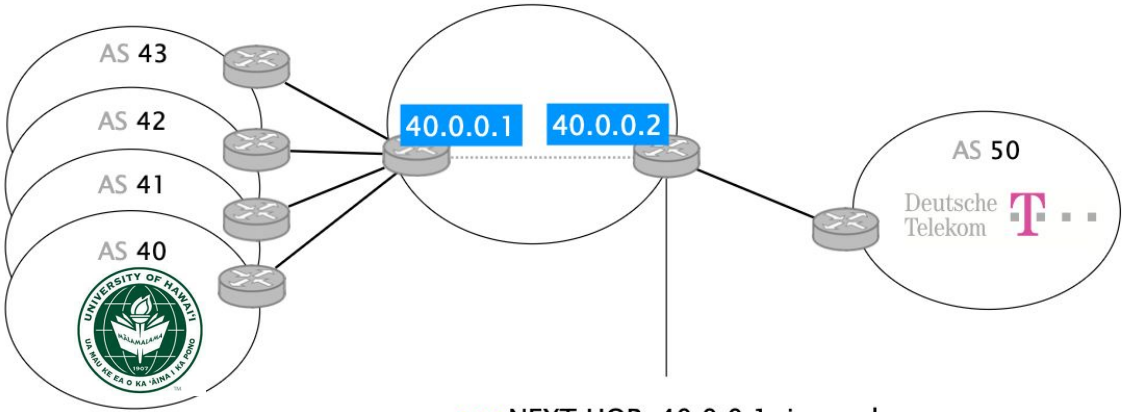
Loopback addresses need to be reachable network-wide. Typically, each router advertise its loopback (as a /32) in the IGP



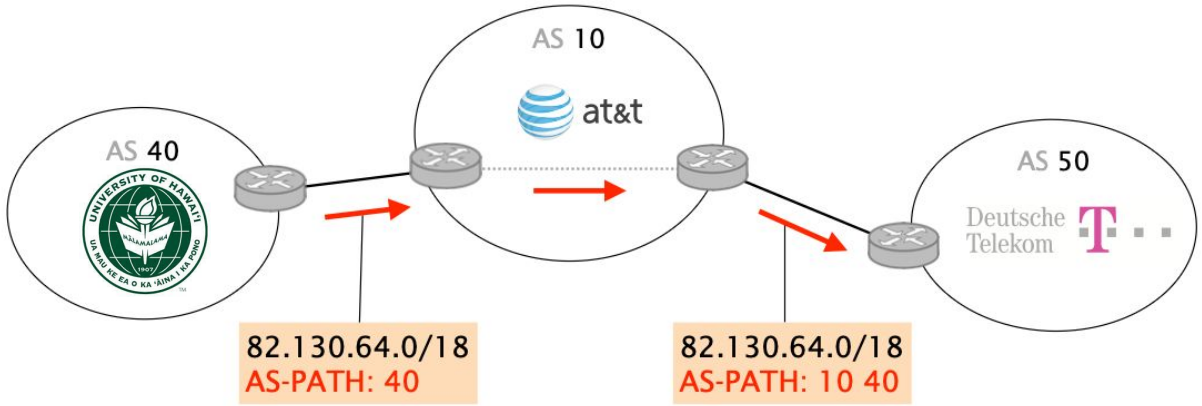
Loopback addresses need to be reachable network-wide. Typically, each router advertise its loopback (as a /32) in the IGP



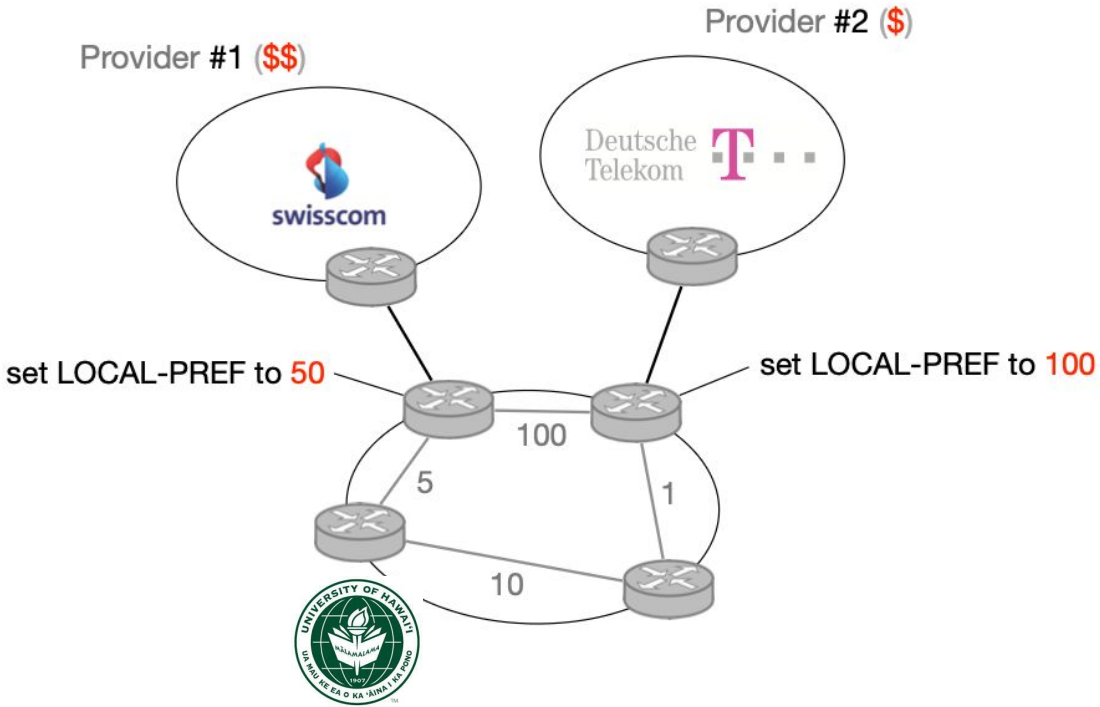
The advantage of next-hop-self is to spare the need to advertise each prefix attached to an external link in the IGP



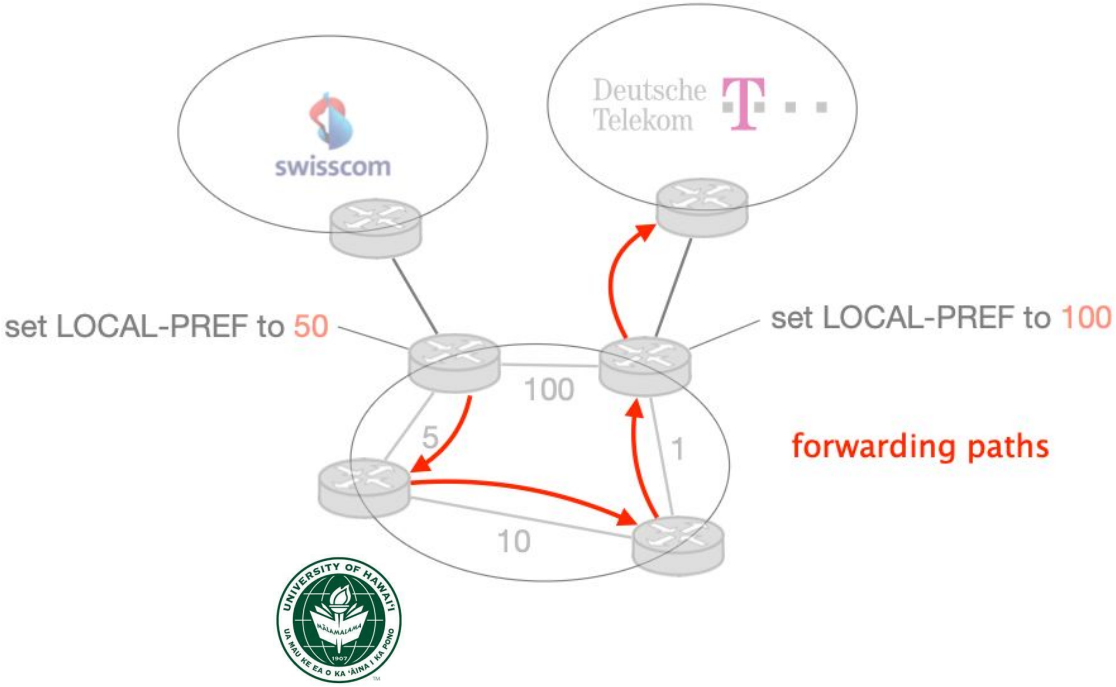
The AS-PATH is a global attribute that lists all the ASes a route has traversed (in reverse order)



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

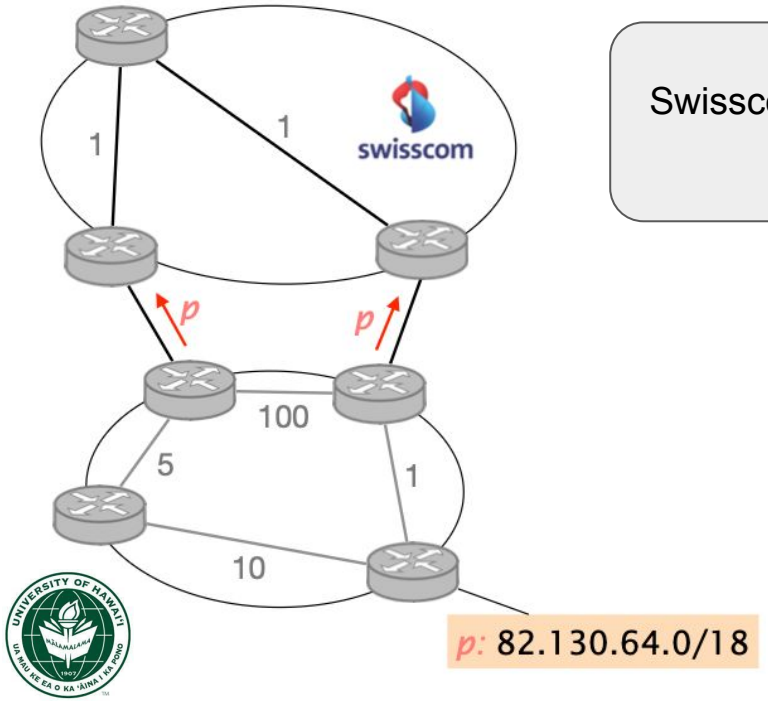


By setting a higher LOCAL-PREF, all routers end up using DT to reach any external prefixes, even if they are closer (IGP-wise) to the Swisscom egress



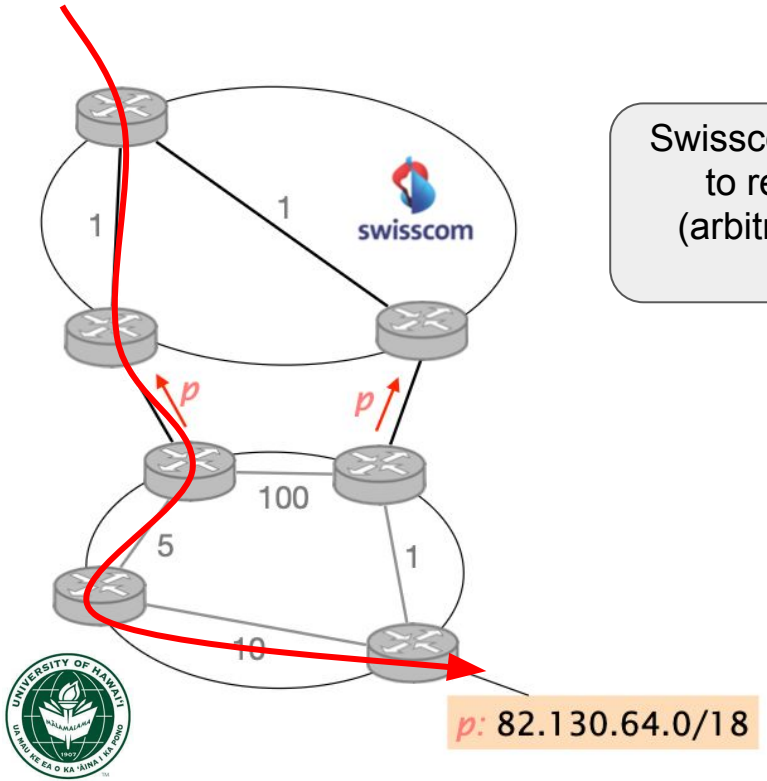
The Multi-Exit Discriminator (MED) is a global attribute which encodes the relative “proximity” of a prefix w.r.t. to the announcer

The Multi-Exit Discriminator (MED) is a global attribute which encodes the relative “proximity” of a prefix w.r.t. to the announcer



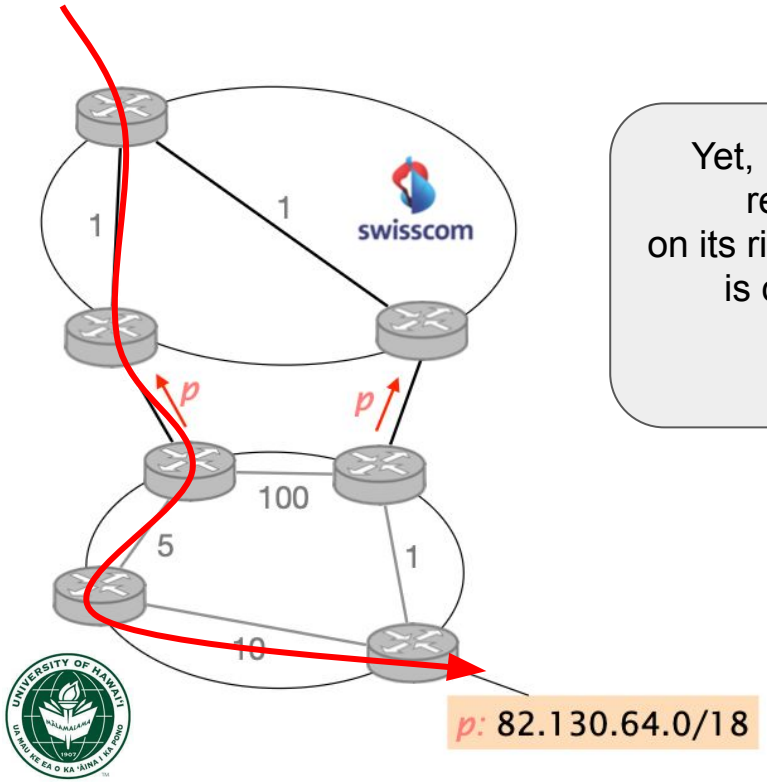
Swisscom receives two routes to reach p

The MED is a global attribute which encodes the relative “proximity” of a prefix w.r.t. to the announcer



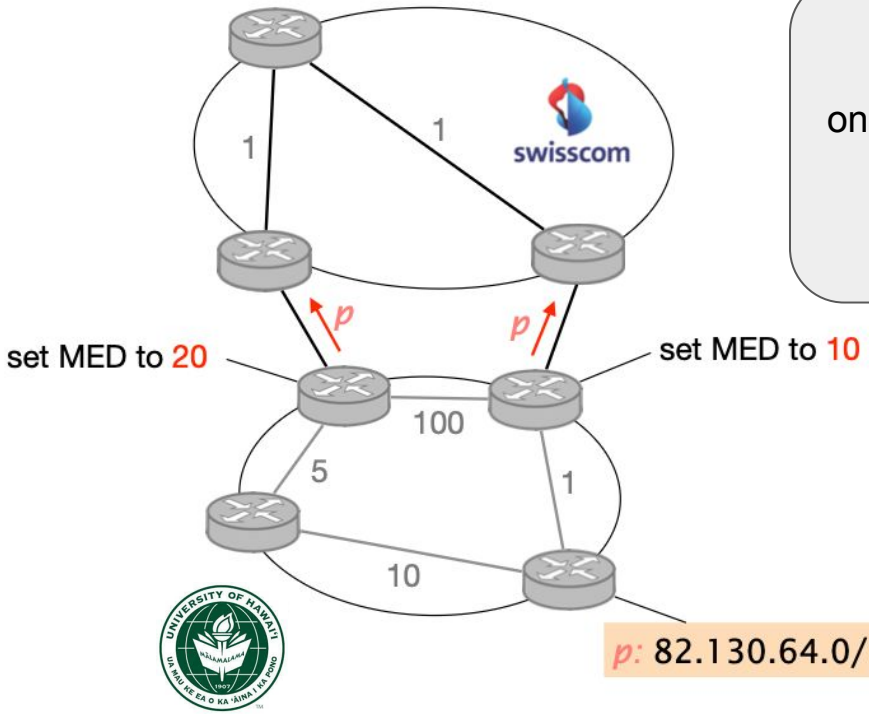
Swisscom receives two routes to reach p and chooses (arbitrarily) its left router as egress

The MED is a global attribute which encodes the relative “proximity” of a prefix w.r.t. to the announcer



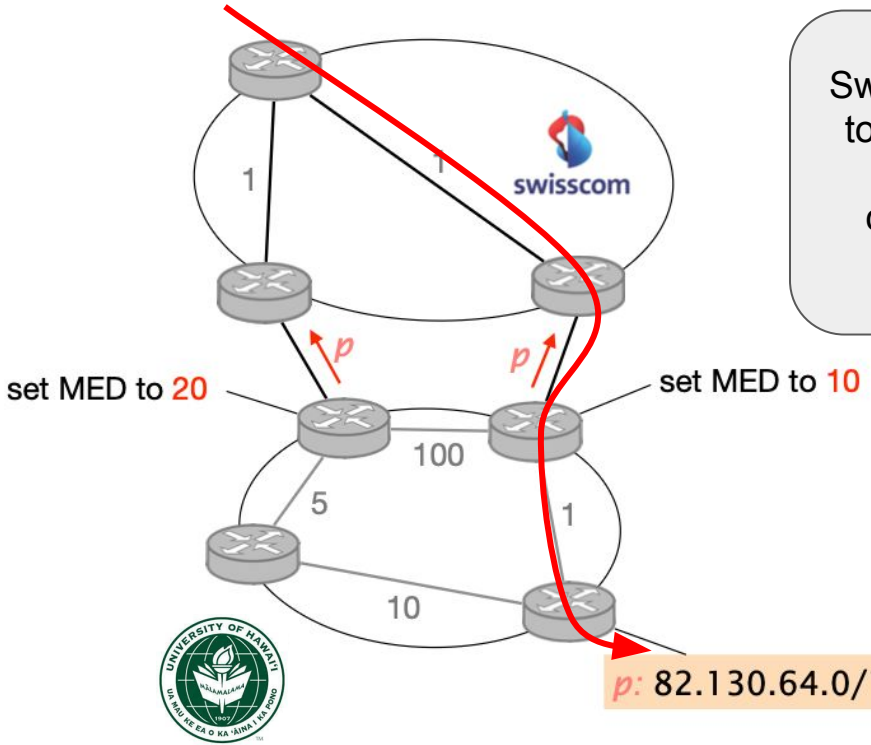
Yet, UHM would prefer to receive traffic for p on its right border router which is closer to the actual destination

UHM can communicate that preference to Swisscom by setting a higher MED on p when announced from the left



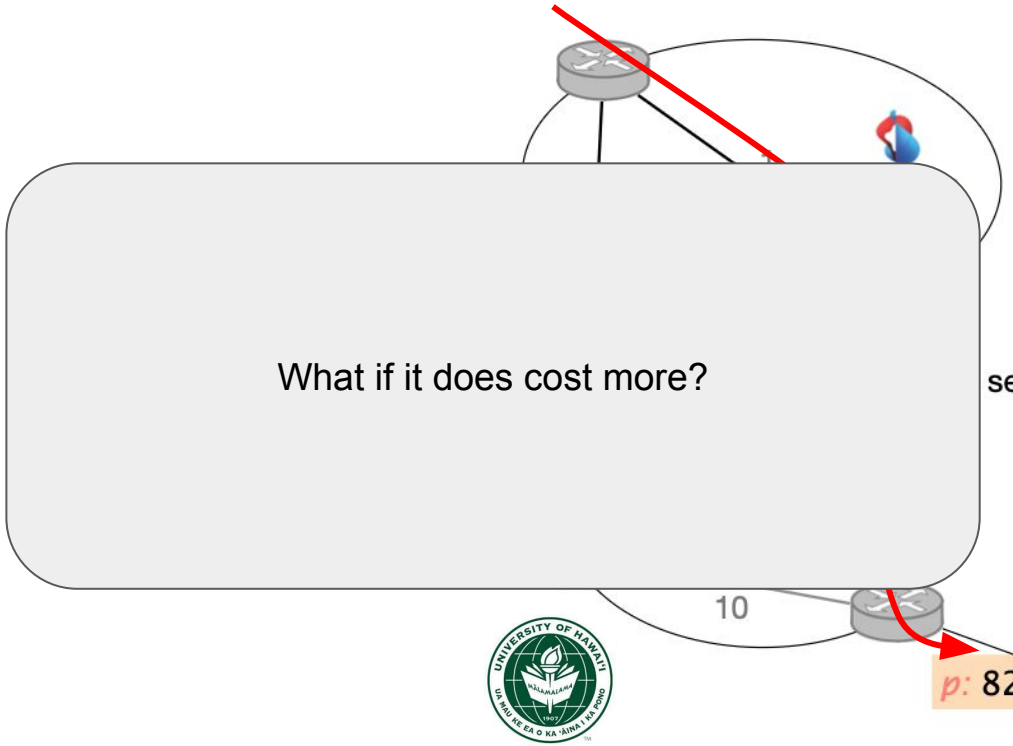
Yet, UHM would prefer to receive traffic for p on its right border router which is closer to the actual destination

UHM can communicate that preference to Swisscom by setting a higher MED on p when announced from the left



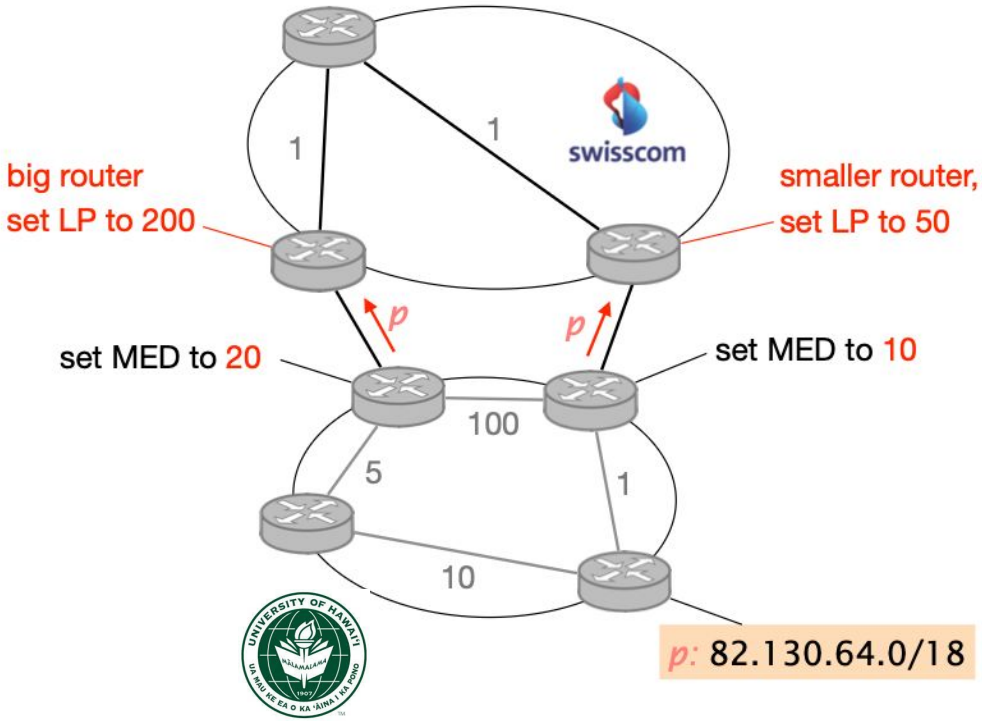
Swisscom receives two routes to reach p and, given it does not cost it anything more, chooses its right router as egress

UHM can communicate that preference to Swisscom by setting a higher MED on p when announced from the left

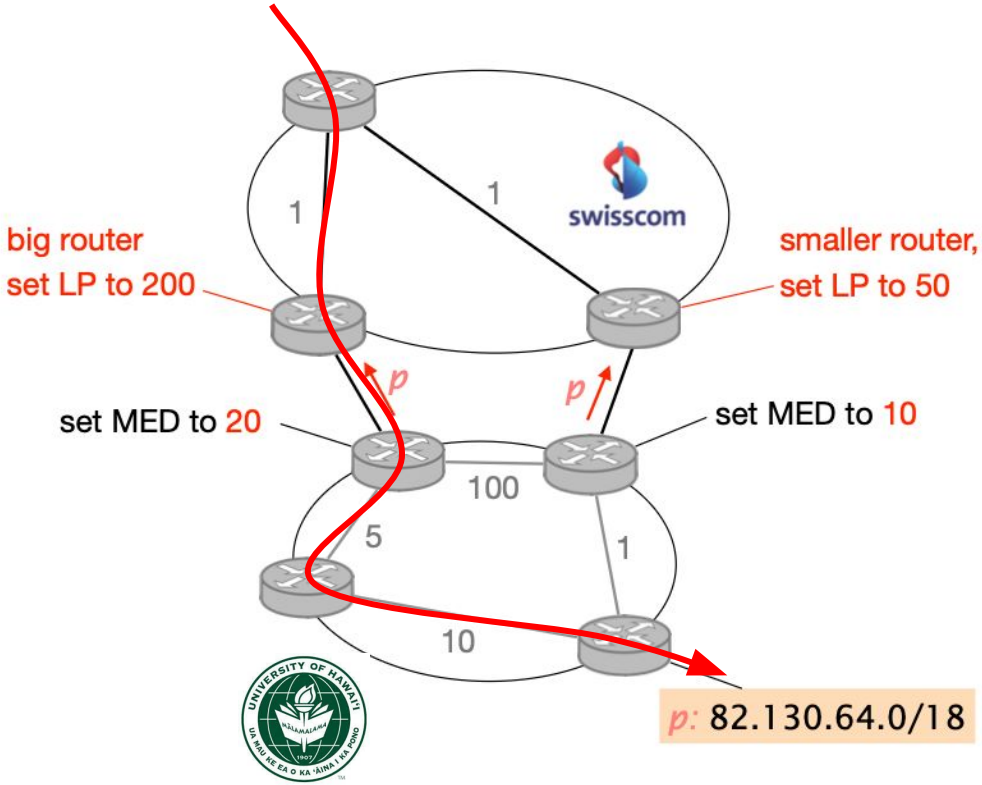


Swisscom receives two routes to reach p and, given it does not cost it anything more, chooses its right router as egress

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



Bottom Line

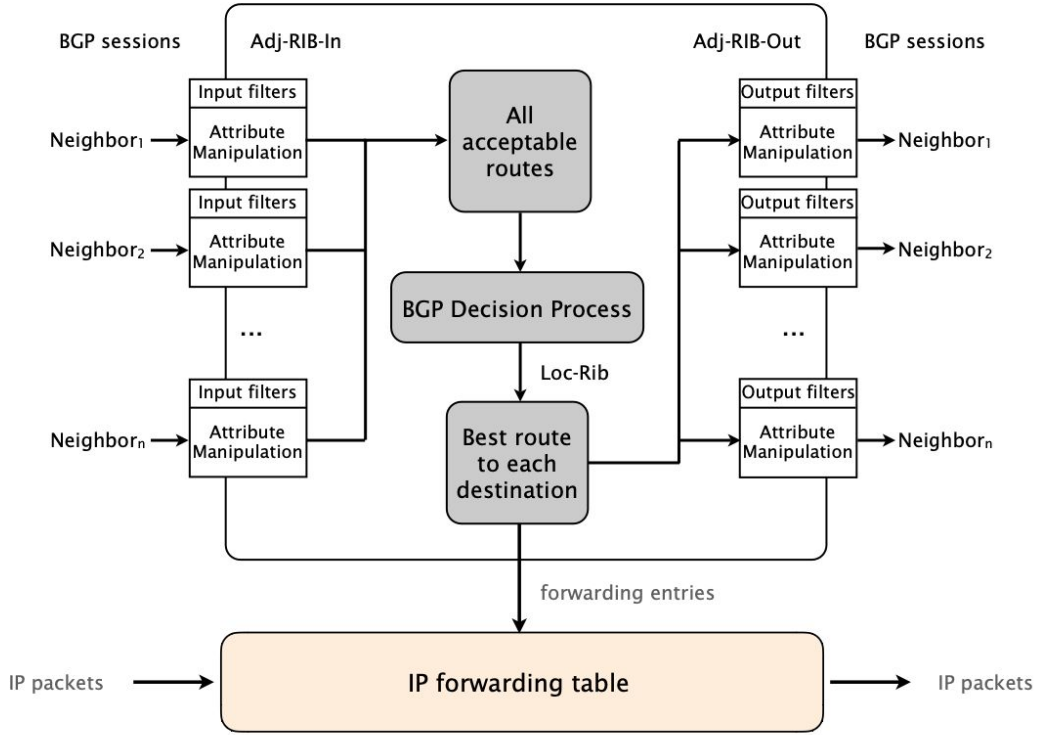
Lesson

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

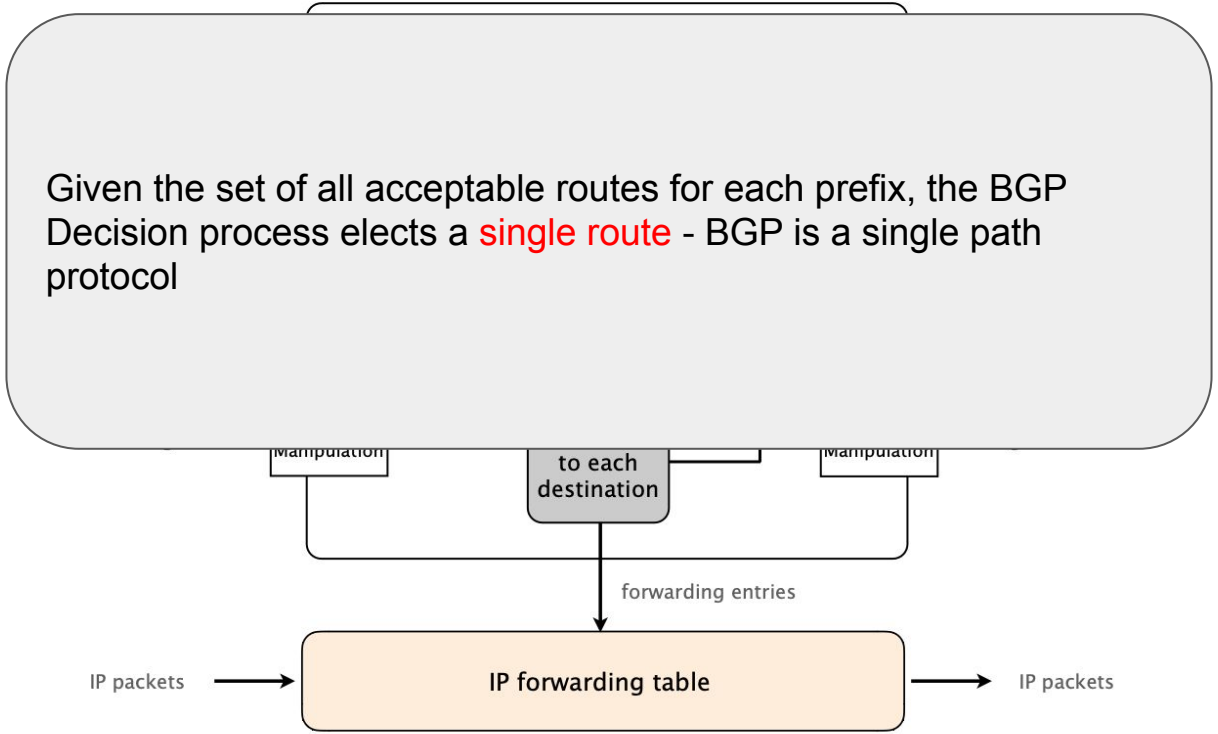
Corollary

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

Each BGP router processes UPDATEs according to a precise pipeline



Each BGP router processes UPDATEs according to a precise pipeline



Each BGP router processes UPDATEs according to a precise pipeline

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)

Each BGP router processes UPDATEs according to a precise pipeline

Prefer routes…

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

learned via eBGP instead of iBGP

with lower IGP metric to the next-hop

with smaller egress IP address (tie-break)

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