# BGP 101

Jason Schiller
Google Network Engineering
jschiller@google.com

## Disclaimer

Any routing architecture depicted in these slides may or may not reflect reality.

They are intended as a reasonable approximation of what may exist in order to demonstrate a concept.

I freely trade between various vendor conventions for formatting convenience.  This is not intended to mean anything.

- E.g. Cisco router puck, with Juniper interface name, Cisco configuration snippets, and Juniper show output snippet

# BGP 101 Outline

- **What is BGP**
  - –Defining BGP
  - –BGP history
- **BGP at 30,000 Feet**
  - –Route propagation between ASes
  - –Distance-vector Routing and Link-state Routing
  - –IGP and BGP Interaction
  - –Routing and Forwarding
- **Path Selection**
  - –BGP Attributes

# BGP 102 Outline

- **BGP Policy**
  - –Common BGP Policy
- **BGP TE**
  - –Affecting Path Selection
- **iBGP Scaling**
  - –Route Reflection
  - –Confederations
- **Operational Considerations**
- **BGP Mechanics**
  - –Configuring BGP
  - –Troubleshooting BGP
  - –BGP Protocol

# What is BGP?

# What is BGP?

What is **BGP** ?

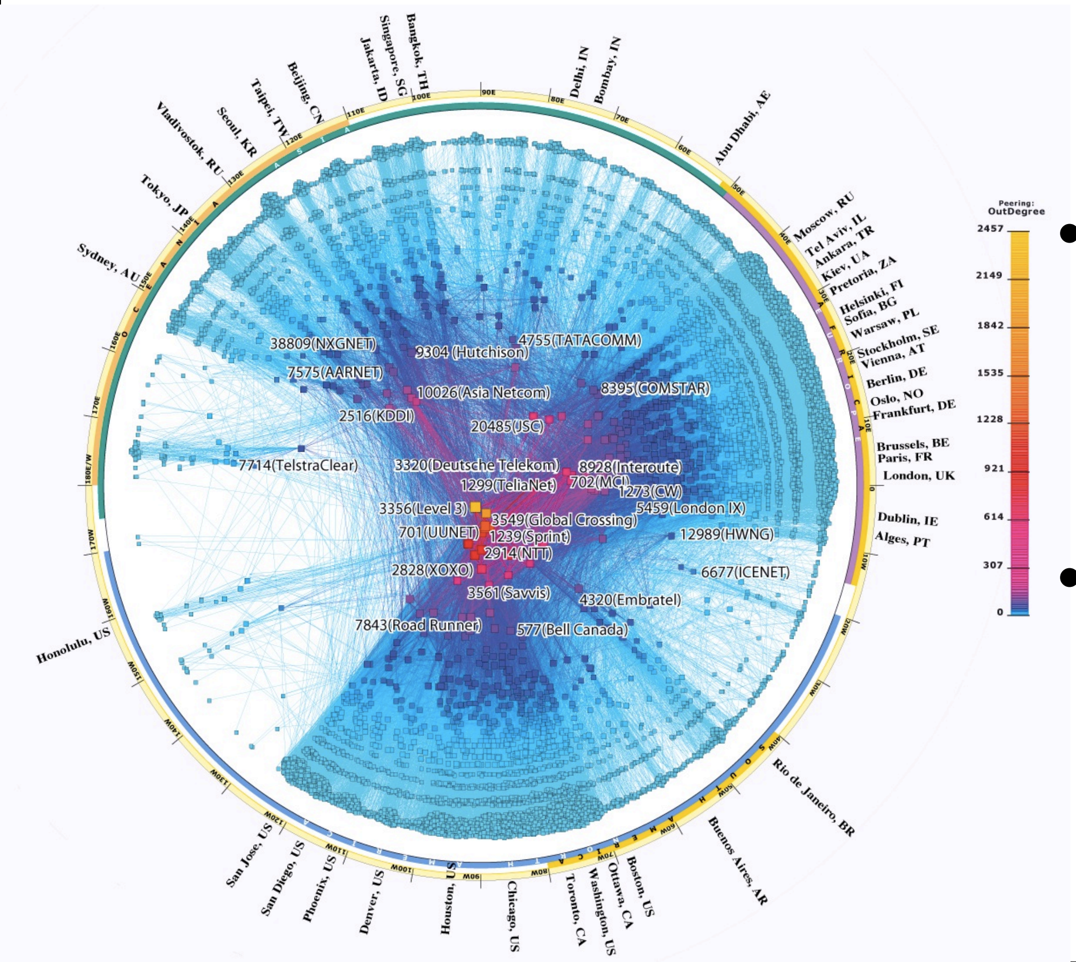
**B**order
**G**ateway
**P**rotocol

# What is BGP?

- BGP is an exterior gateway routing protocol used for inter-Autonomous System routing
- Defined in RFC-4271
  - –http://tools.ietf.org/html/rfc4271
- Autonomous System (AS)
  - –A single network that is under a single administration with a single routing policy
  - –This routing domain is assigned an Autonomous System Number (AS number or ASN)

# So What is BGP?

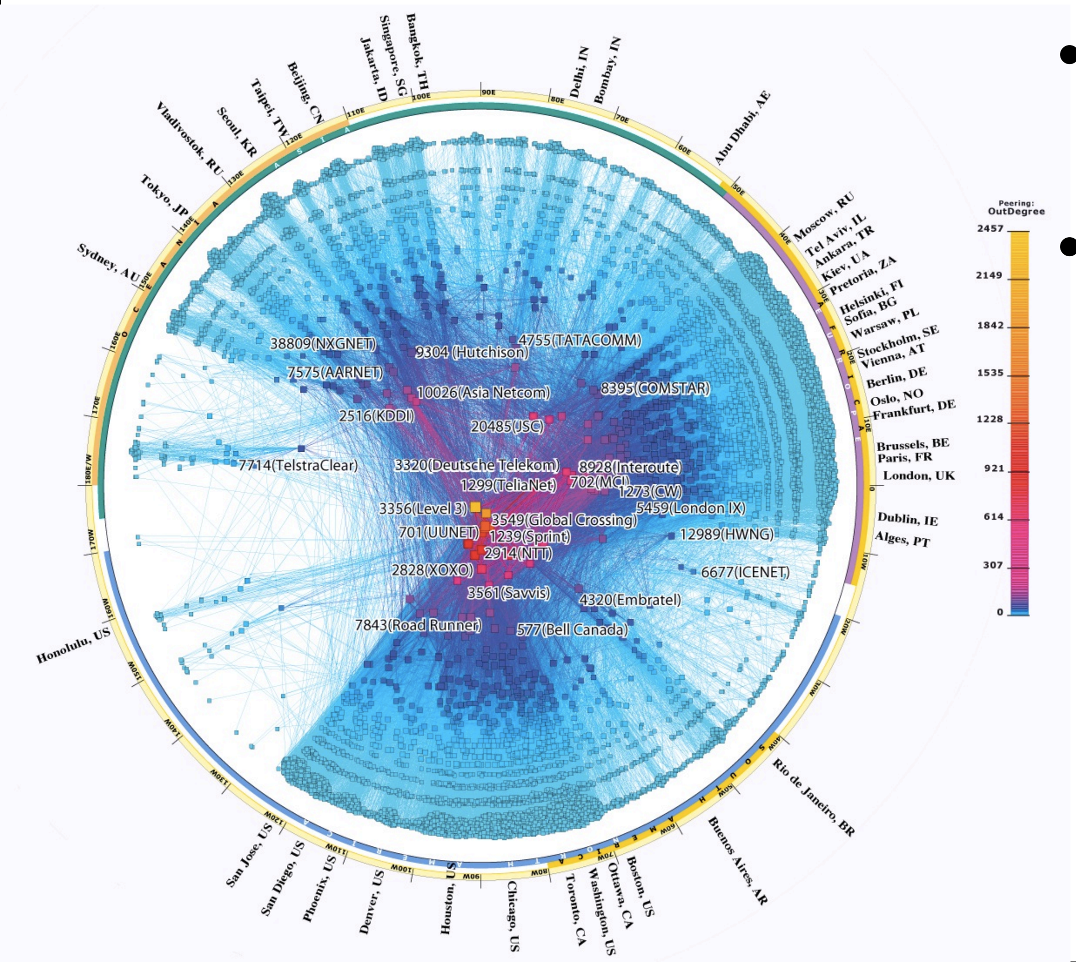
- BGP is the routing protocol that allows each network on the Internet to signal to other networks what destinations they can reach



- “Exterior” refers to the intention that it is to be used on the out side of the network
- “Inter-AS” refers to the communication between ASes (networks)

# So What is BGP?

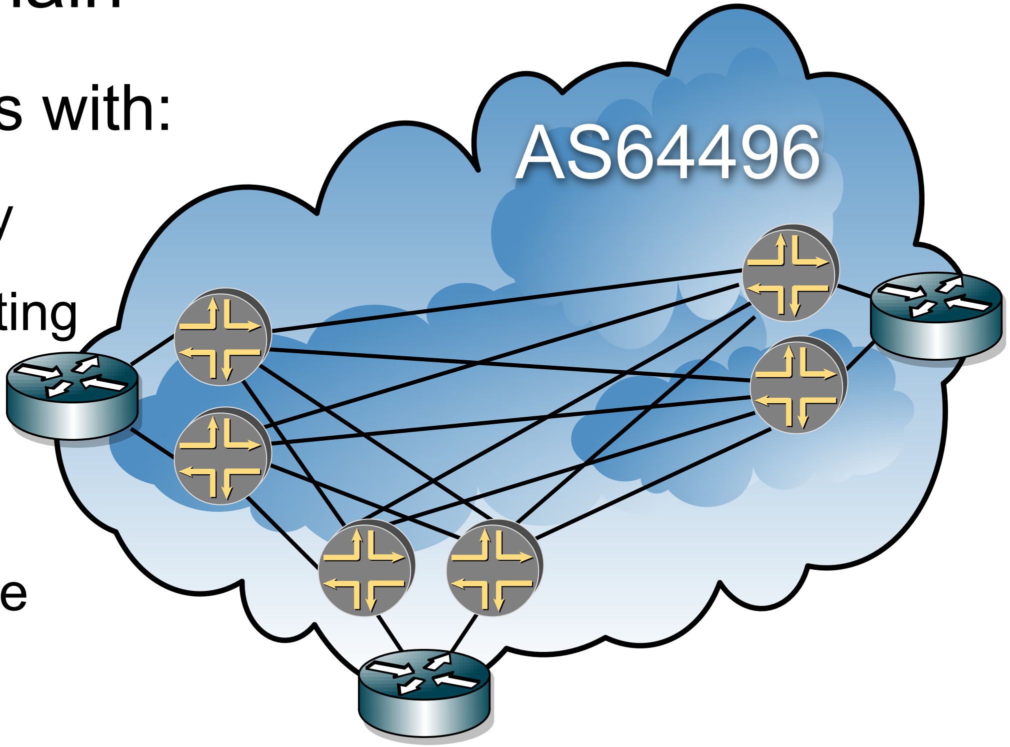
- BGP learns multiple paths to each route
- BGP selects the best path



- Best path is used to forward traffic
- Only the best path is sent along to other BGP neighbors
  - Each square represents an AS
  - Each line represents one or more peering relationship where paths are learned

# So What is an AS

- A single routing domain
  - –A collection of routers with:
    - Common routing policy
    - Consistent view of routing
      - –Usually a single IGP
    - Common ownership
    - Common administrative control
  - –A single globally unique AS number (ASN)



# Where Do ASNs Come From

- Globally Unique ASNs
  - Come from the Regional Internet Registries (RIRs)
  - AS 0 and AS65535 are reserved
- Private ASNs
  - Can be used for “internal” use
  - Range from 64512 – 65534
- Documentation ASNs
  - Can be used for documentation per RFC-5398
  - Range from 64496-64511 & 65536 - 65551

# ASN Format

- 2-byte ASNs
  - **Range from 0 to 65535**
- 4-byte ASNs
  - **Added another 16 bits**
  - **Additional addresses range from 65536 through 4294967296**
  - **AS plain**
    - Display the number as a decimal number
    - 10 digit number is hard to remember
    - Even telephone numbers have separators (NPA) NXX - xxxx
  - **AS dot**
    - Display the number as two 16 bit decimal numbers
    - Existing AS1234 becomes AS 0.1234
    - New 4-bye ASN range from 1.0 through 65535.65535
    - Issues with regex – “dot” means any character
      - Lots of debate about what character to use, and how many scripts, databases, tools, etc will break
- AS23456
  - **Used to represent a four-byte ASN to devices that are not 4-byte ASN compatible to enable backwards compatibility** (more on this later)

# AS Caveats

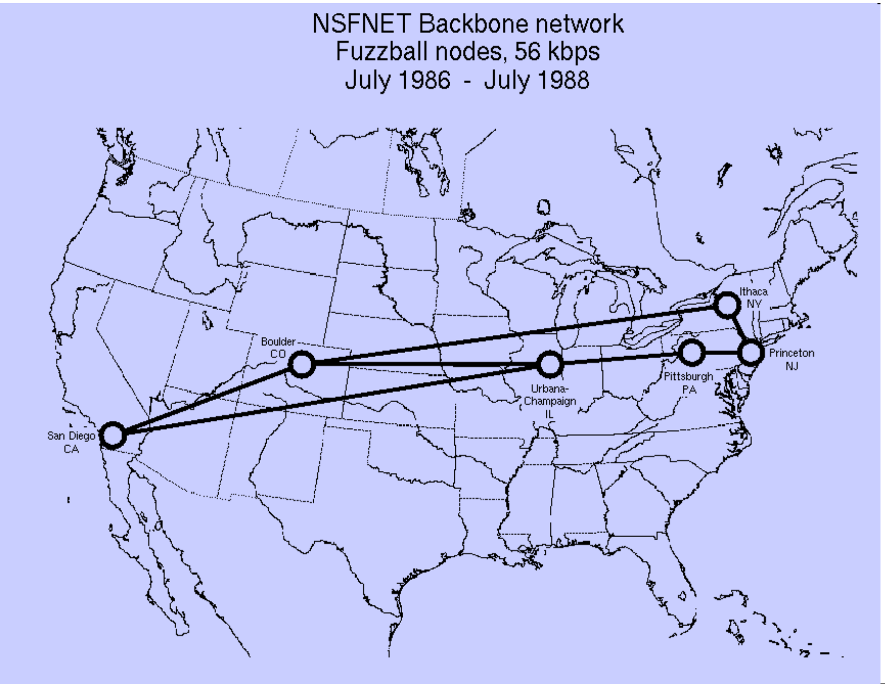
- Networks that are single homed, may use static routes to connect to the Internet
  - These networks “do not count” in the BGP world
  - These networks do not need an ASN
- Some transit providers allow customers to use a private ASN if they are multi-connected only to them
  - These private ASN are stripped from the routes when sent to the rest of the Internet
  - Creates an “ownership” problem because it looks like the transit provider is originating (and therefore responsible for) the route
- Some transit providers use RFC-2270 ASN for customers that are multi-connected only to them
  - Is a globally unique ASN registered to transit provider but designated for use by customers multi-connected to only that transit provider
  - Many different customer networks will use this same ASN

# BGP History

# NSFNET

- In 1985 the NSFNET was established to create an open network to provide academics access to 6 supercomputers
- In 1986 the NSFNET came online using 56K leased lines and PDP-11's as "gateways" running "fuzzball" software by David Mills
  - Ran EGP-2

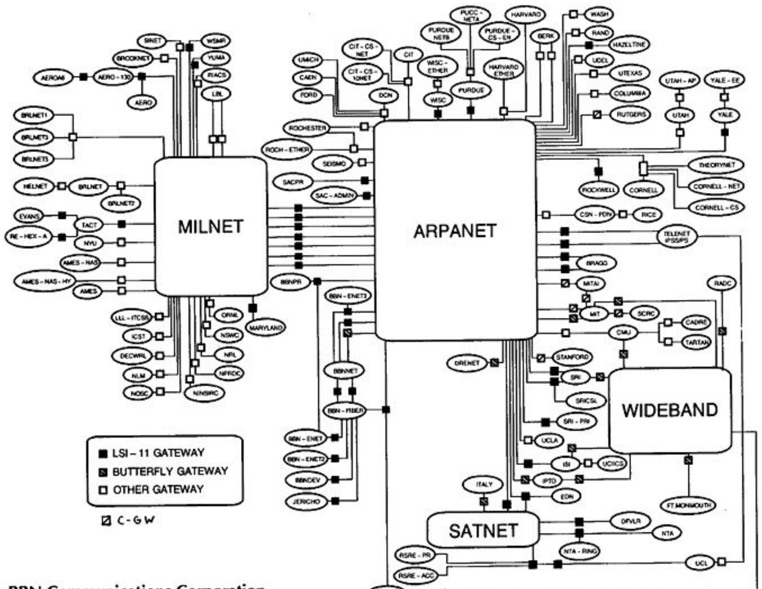
**http://www.eecis.udel.edu/~mills/database/papers/fuzz.pdf**
**http://www.eecis.udel.edu/~mills/database/papers/bone.pdf**



NSFNET Backbone network
Fuzzball nodes, 56 kbps
July 1986 - July 1988

# EGP – 2
# Exterior Gateway Protocol

**Internet in 1986**



BBN Communications Corporation

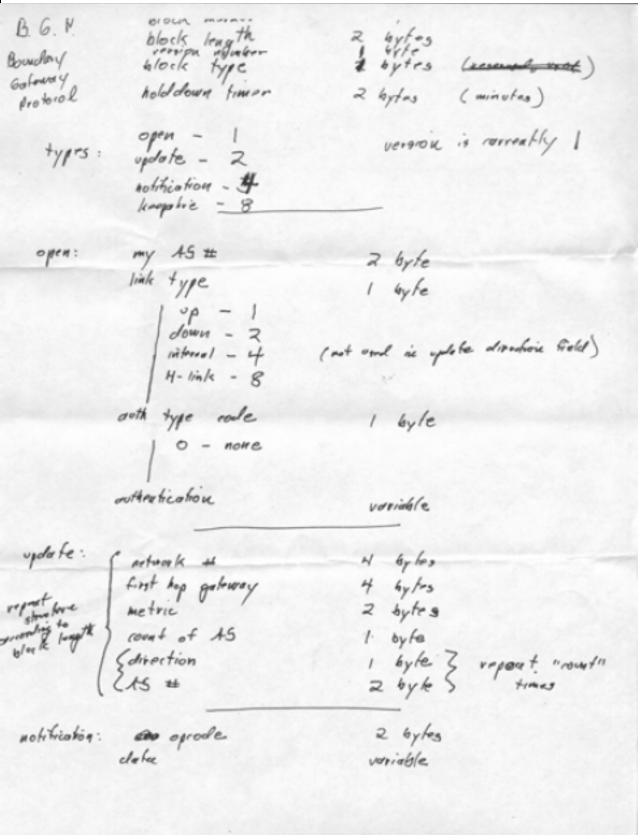
- Expected stub "gateways" would have a single connection to a single, centrally managed core of "backbone gateways"
  - Used spanning tree algorithm
  - Required everyone to single home to ARPANET
- Used unreliable IP transport causing routing instability
- Used periodic refresh of entire routing table

- The network was growing rapidly, and not a simple tree with sub-networks connecting to a single core
- EGP didn't scale well
- In 1987 NSF established an agreement with MCI / IBM / MERIT to upgrade and operate NSFNET Backbone Phase II
  - Deployed BGP

# BGP – A Three Napkin Protocol

## (Minus the Ketchup)

- The BGP protocol was designed in 1/1989 at IETF 12 over lunch
  - **Yakov Rekhter & Kurt Lougheed with help from Len Bosak**
  - **RFC-1105 formalized in June 1989**



B. G. P.
Boundary
Gateway
Protocol

block length 2 bytes
version number 1 byte
block type 1 bytes (~~[illegible]~~)
holddown timer 2 bytes (minutes)

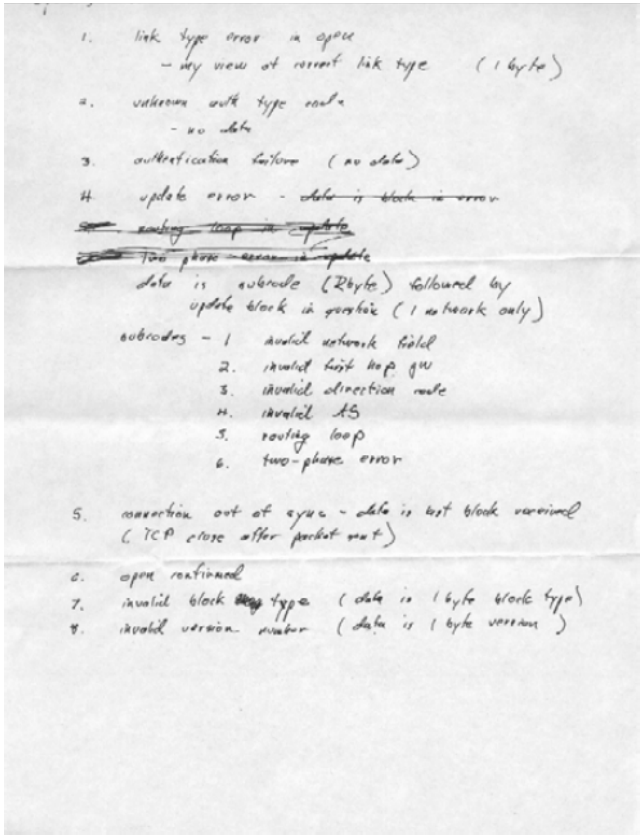
types: open - 1
update - 2
notification - 4
keepalive - 8

version is currently 1

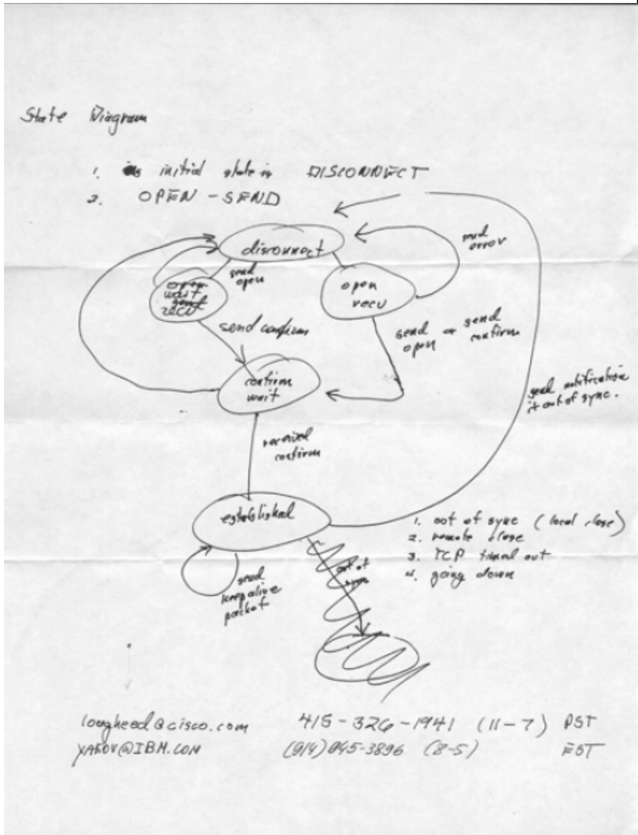
open: my AS # 2 byte
link type 1 byte
up - 1
down - 2
internal - 4
H-link - 8
(not used in update direction field)
auth type code 1 byte
0 - none
authentication variable

update: network # 4 bytes
first hop gateway 4 bytes
metric 2 bytes
count of AS 1 byte
direction 1 byte
AS # 2 byte
repeat "count" times
repeat structure according to block length

notification: opcode 2 bytes
data variable



1. link type error in open
   - my view of correct link type (1 byte)
2. unknown auth type code
   - no data
3. authentication failure (no data)
4. update error - ~~data is block in error~~
~~5. routing loop in update~~
~~6. two phase error in update~~
   data is subcode (2byte) followed by
   update block in question (1 network only)
   subcodes - 1 invalid network field
   2. invalid first hop gw
   3. invalid direction code
   4. invalid AS
   5. routing loop
   6. two-phase error
5. connection out of sync - data is last block received
   (TCP close after packet sent)
6. open confirmed
7. invalid block type (data is 1 byte block type)
8. invalid version number (data is 1 byte version)



State Diagram

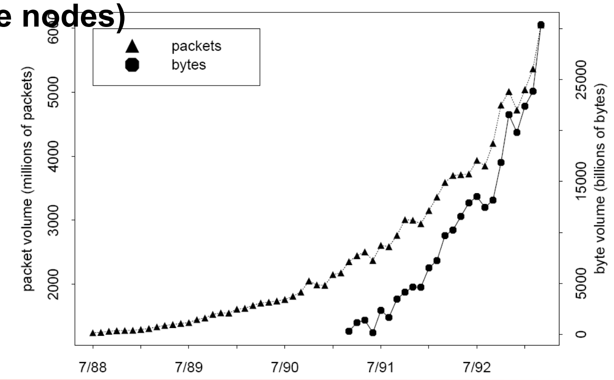
1. initial state is DISCONNECT
2. OPEN - SEND

1. out of sync (local close)
2. remote close
3. TCP timed out
4. going down

lougheed@cisco.com 415-326-1941 (11-7) PST
YAKOV@IBM.COM (914)945-3896 (8-5) EST

# The Rest is Internet History

Tracking NSFNET traffic aspects



**1988**

- NSFNET was upgraded to T-1
- Seven additional research networks were added (13 total Backbone nodes)
- NSFNET now connected over 170 TCP/IP networks
- Traffic doubled every seven months
- Operated by MERIT

**1989**

- BGP was deployed
- NSFNET upgraded to T-3
- Had 16 backbone nodes

**1990**

- ARPANET was dissolved.
  - Became BBN, GTE (1997), Genuity (2000) and merged with Level(3) (2003)
- Parallel commercial networks were built
  - UUNET, Performance Systems International network (PSI Net), CERFNet, and NEARNet
  - Merit spun out ANS and ANS CO+RE Systems to separate non-profit and profit entities

**1993**

- Opened a draft of its solicitation for a new architecture to the public
- MERIT awarded Route Arbiter Database (RADB)
- vBNS a Very High Speed Backbone Network Service awarded to MCI (later became MCI's commercial vBNS+, and Verizon Business Private IP)
- Network Access Point Manager awarded to Sprint for NY NAP, MFS Datanet for D.C. NAP, Bellcore for Chicago NAP & CA NAP

**1995**

- WWW traffic (21%) exceded FTP traffic (14%) for the first time in April 1995
- NSFNET was dissolved. At that time is supported 4,000 institutions and 50,000 networks across the US, Canada, and Europe
- NSF retained vBNS a Very High Speed Backbone Network Service as a core research network.

# BGP History

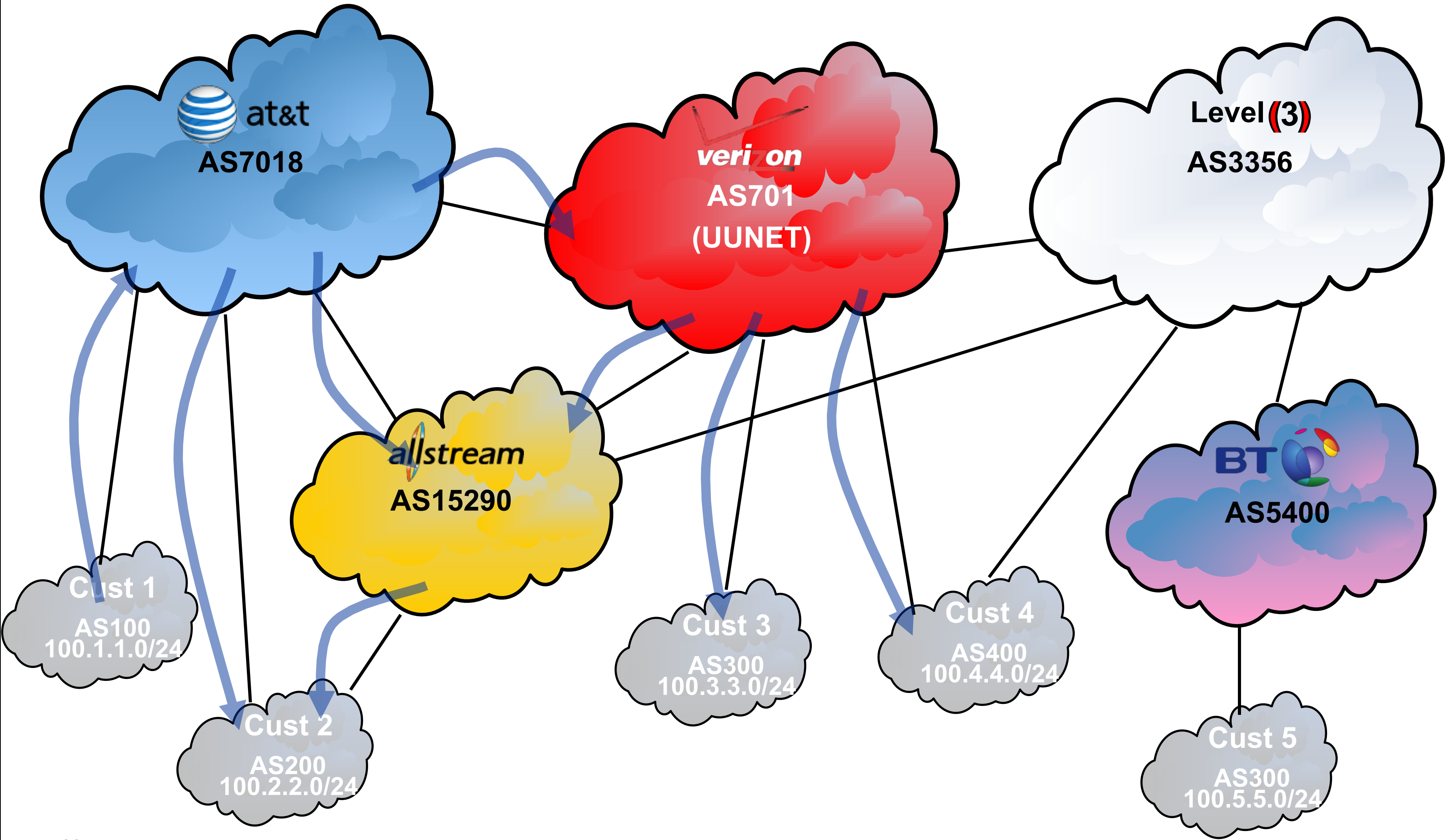
- RFC-4271 BGP v4
- RFC-1771 BGP v4
- RFC-1654 BGP v4
- RFC-1267 BGP v3
- RFC-1163 BGPv2
- RFC-1105 BGP
- RFC-904 EGPv2
- RFC-888 EGPv2
- RFC-827 EGPv1

# BGP At 30,000 Feet

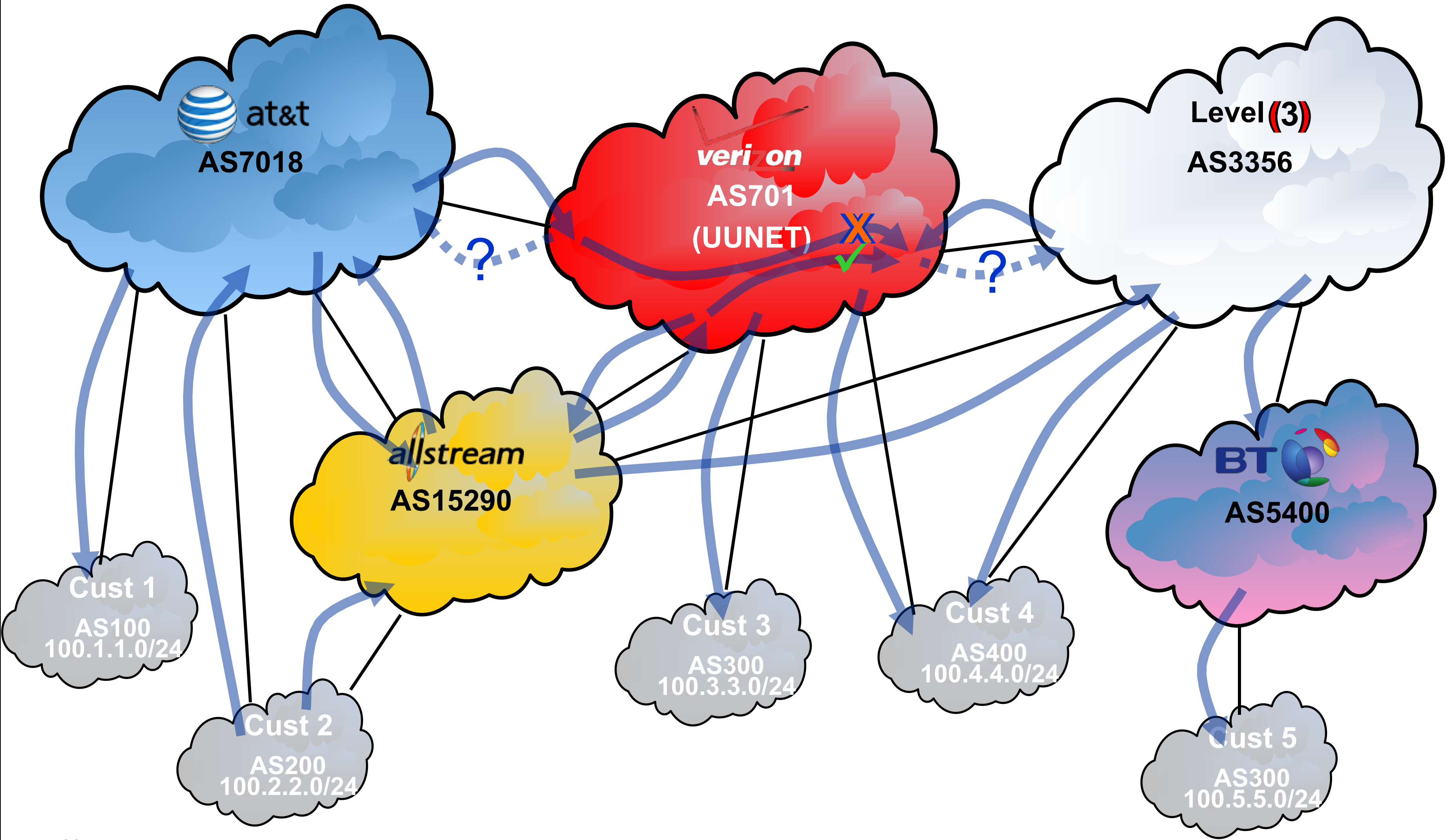
# Definitions

- Neighbors / peering – the relationship between two routers that exchange routing information over BGP
  - Types of peering relationships
    - Transit – The network is a “customer” of their upstream transit provider. The network usually pays to reach the whole Internet through their upstream
    - Peering – The network has a mutually beneficial settlement free interconnect with another network of roughly the same size and value. Peers generally only provide transit to their customers (and do not provide transit to customers of other Peers)
    - Customer – The network is a “transit provider” to their downstream “customer”. The network is usually paid to carry their customer’s traffic
- Default Routing
  - The “gateway of last resort”. It is the place where all traffic is sent if there is not more specific information about how to reach a destination.
- Default Free Zone
  - A collection of networks that have no default route. These networks must carry a “full Internet routing table” that carries reachability information for all destinations on the Internet

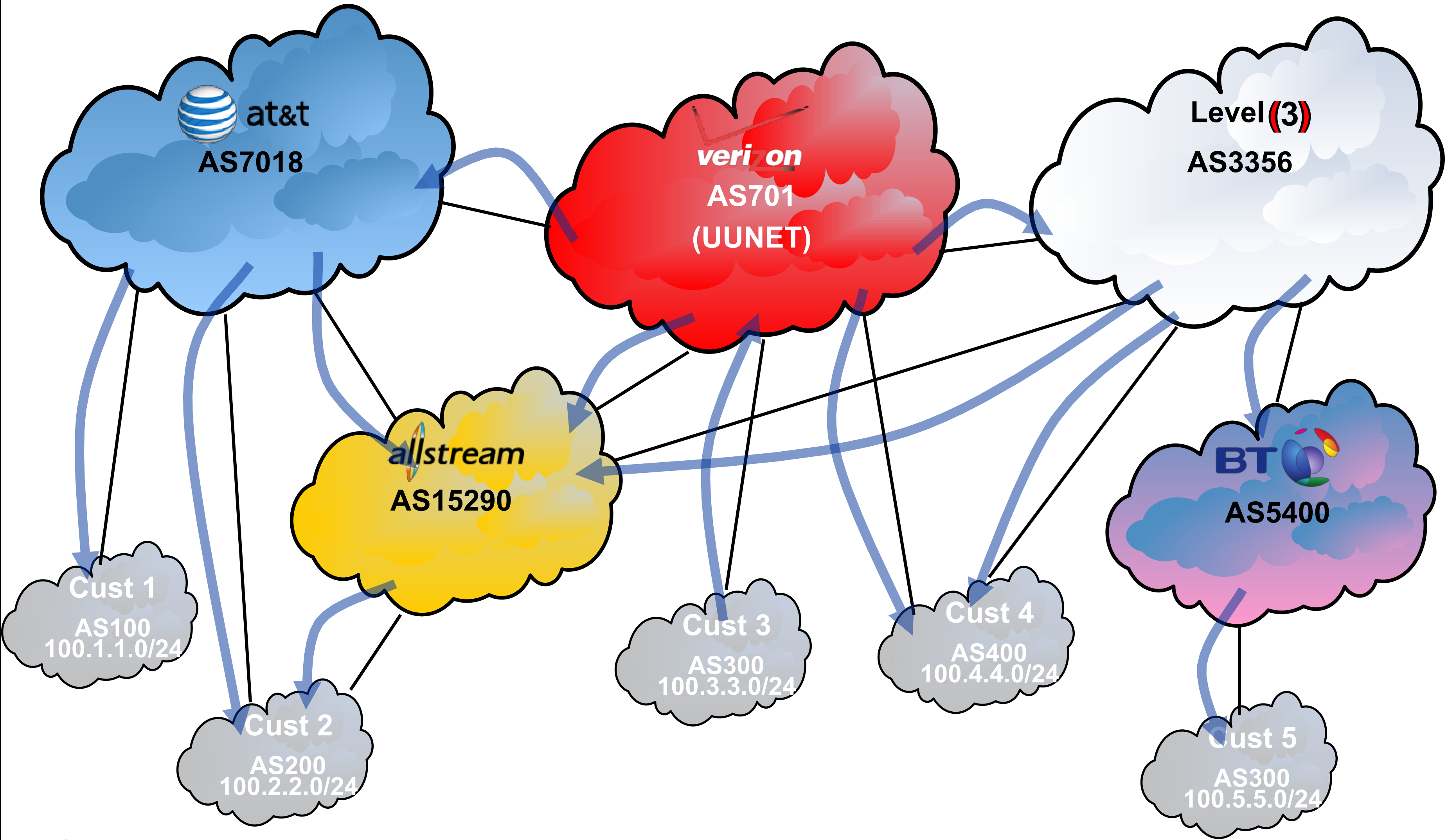
# BGP Routing at 30,000 Feet
# Routes Advertised From Customer 1



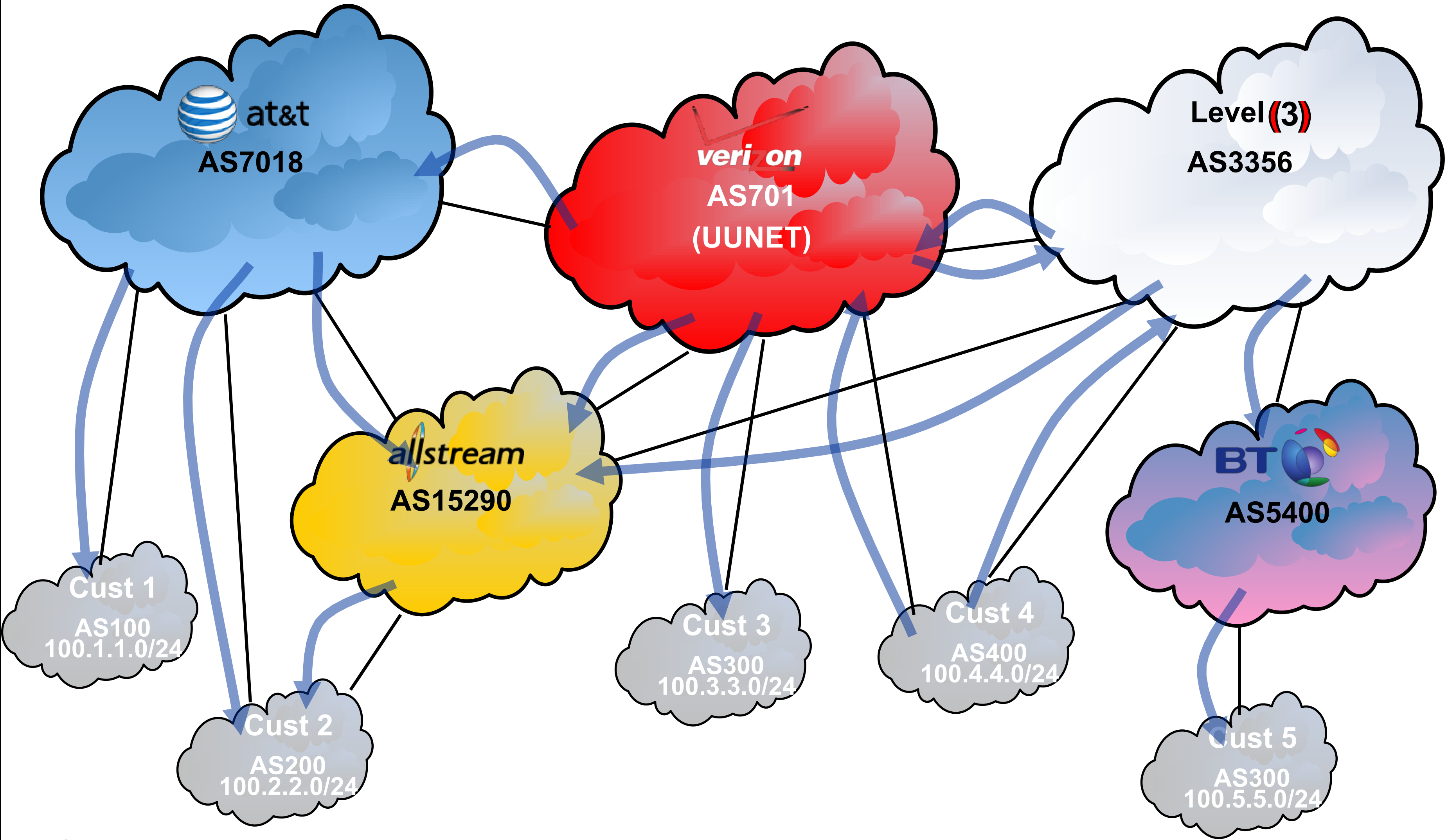
# BGP Routing at 30,000 Feet
## Routes Advertised From Customer 2



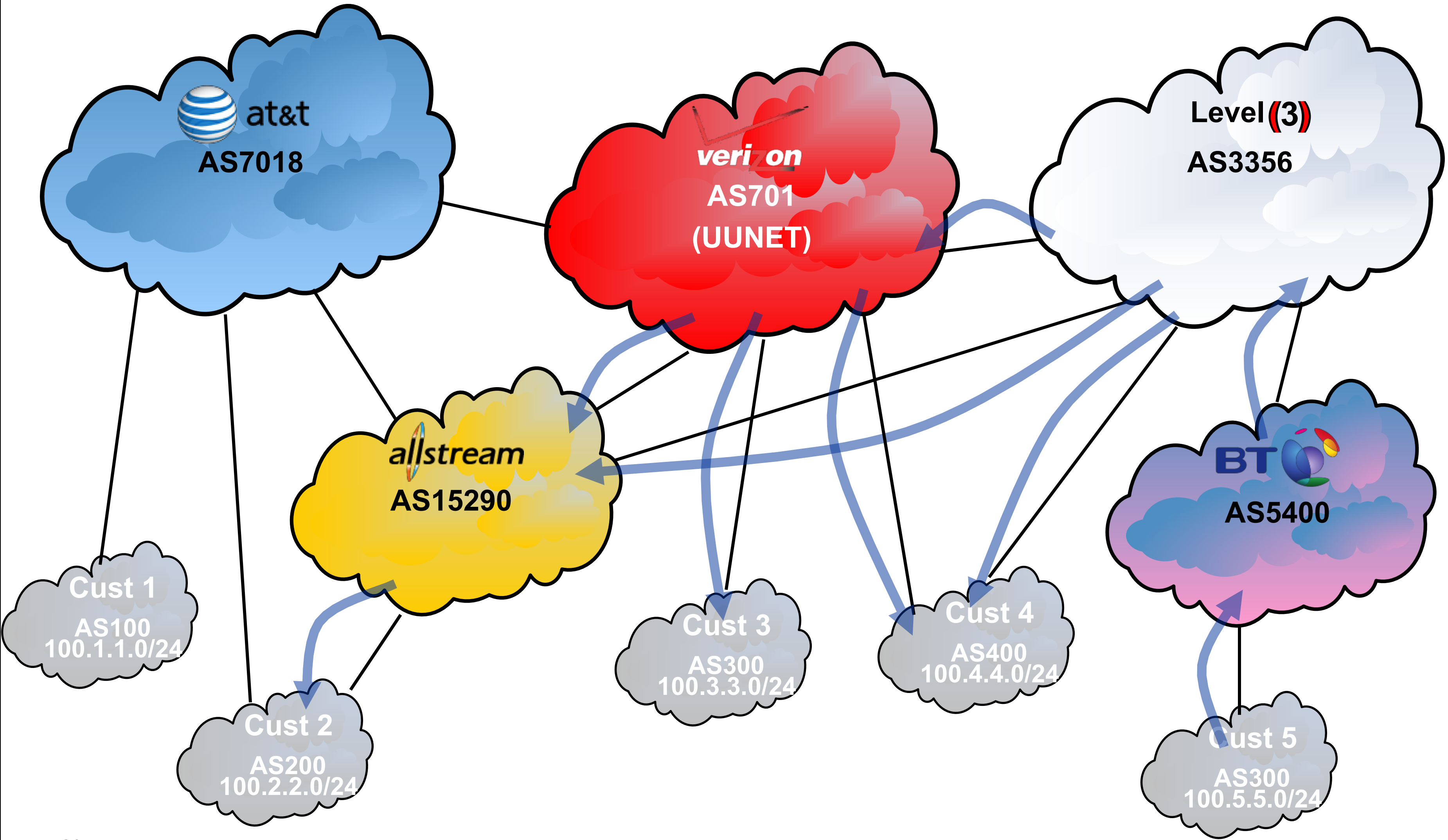
# BGP Routing at 30,000 Feet
# Routes Advertised From Customer 3



# BGP Routing at 30,000 Feet
# Routes Advertised From Customer 4



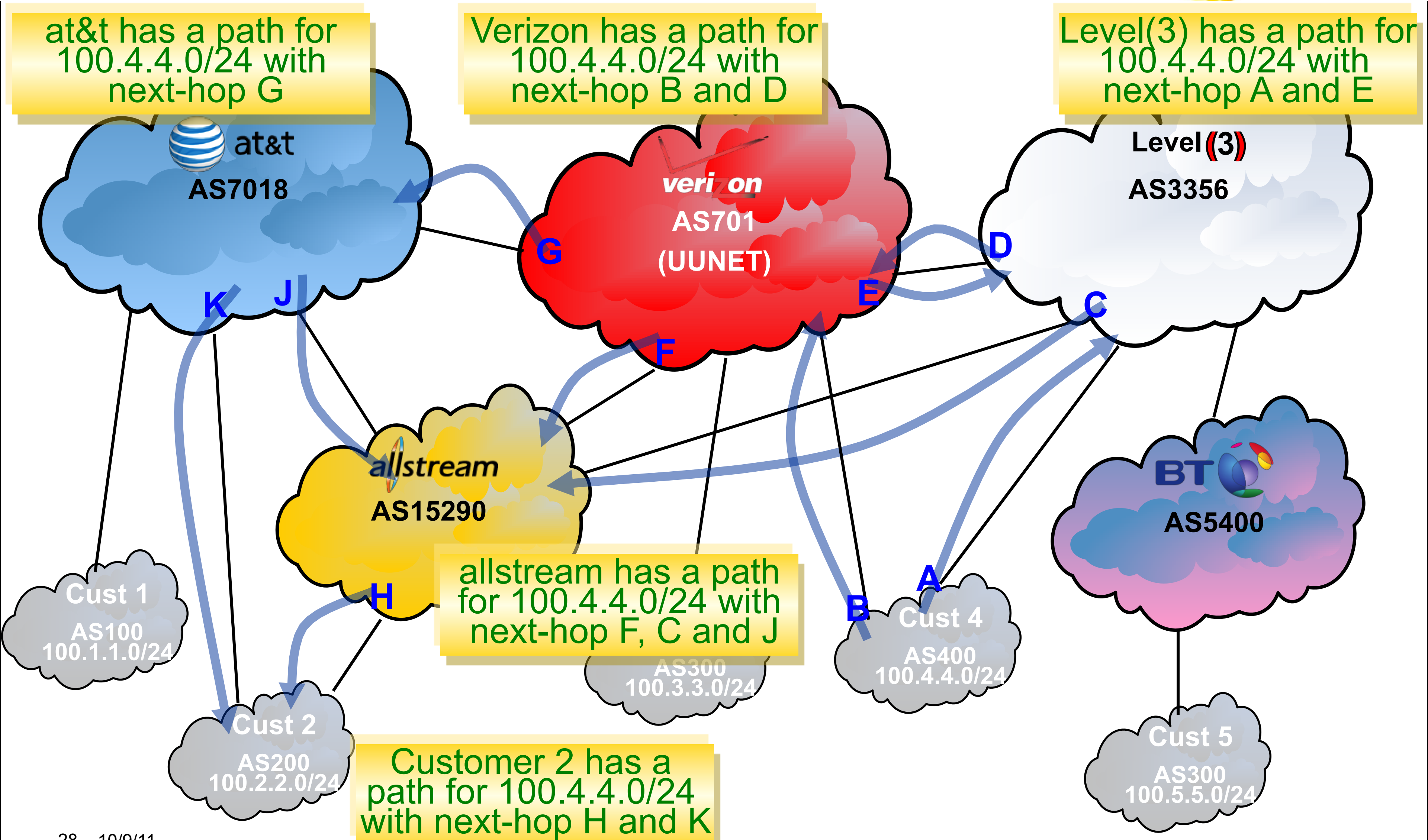
# BGP Routing at 30,000 Feet
# Routes Advertised From Customer 5



# Distance-Vector Routing

- The examples on the previous slides shows the distance vector nature of BGP routing
- In addition to announcing the reachability to the destination prefix BGP also announces the protocol next-hop
  - This next-hop is changed between each AS
- Each AS chooses the best path, and then forwards traffic towards the associated protocol next-hop (AS exit point)
  - This is the "distance" and "vector"
  - This is slightly more complicated as "distance" is not a simple link cost (more on path selection later)
- An AS in the middle of the topology does not know the entire topology
  - Only knows the topology to the edge of its network
  - Relies on cumulative down stream routers' path selection each with their limited view of their own topology
  - Relies on cumulative down stream routers' conveyance of "distance"

# BGP Protocol Next-Hop Changes From Customer 4 To Customer 2



at&t has a path for 100.4.4.0/24 with next-hop G

Verizon has a path for 100.4.4.0/24 with next-hop B and D

Level(3) has a path for 100.4.4.0/24 with next-hop A and E

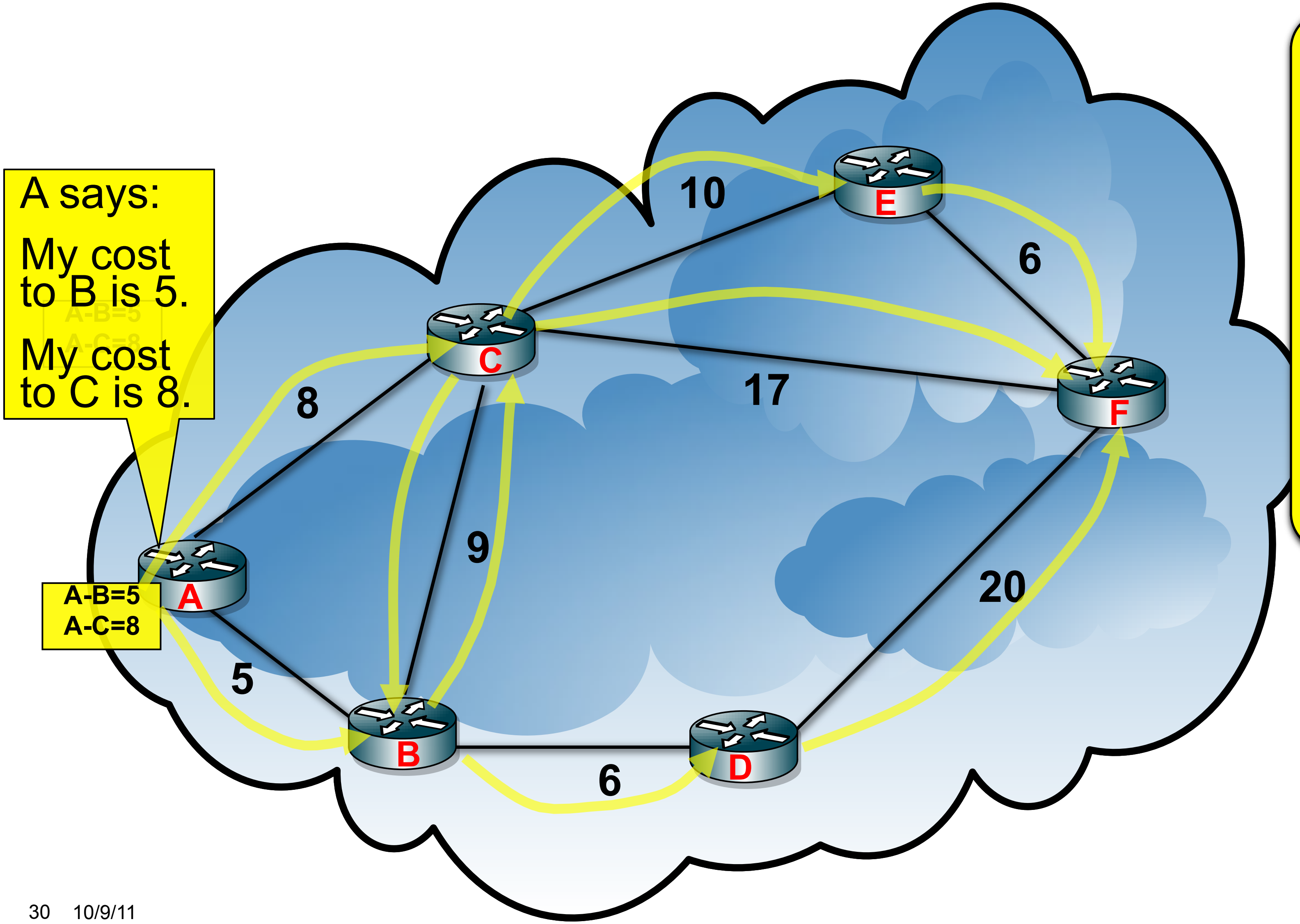
allstream has a path for 100.4.4.0/24 with next-hop F, C and J

Customer 2 has a path for 100.4.4.0/24 with next-hop H and K

# Link-State Routing Protocols

- Link-state routing protocols
  - OSPF and IS-IS are link state IGPs
  - RIP and  IGRP are distance vector IGPs
- Every node in the network announces all locally reachable destinations, and their link cost
  - This includes directly adjacent nodes
- This information is flooded to every node in the network
  - From this information, a graph is built showing the cost between each node, and each reachable destination
- Each node, from their unique location in the graph, then calculates the shortest path to every destination

# Link-State Routing Protocols



A says:

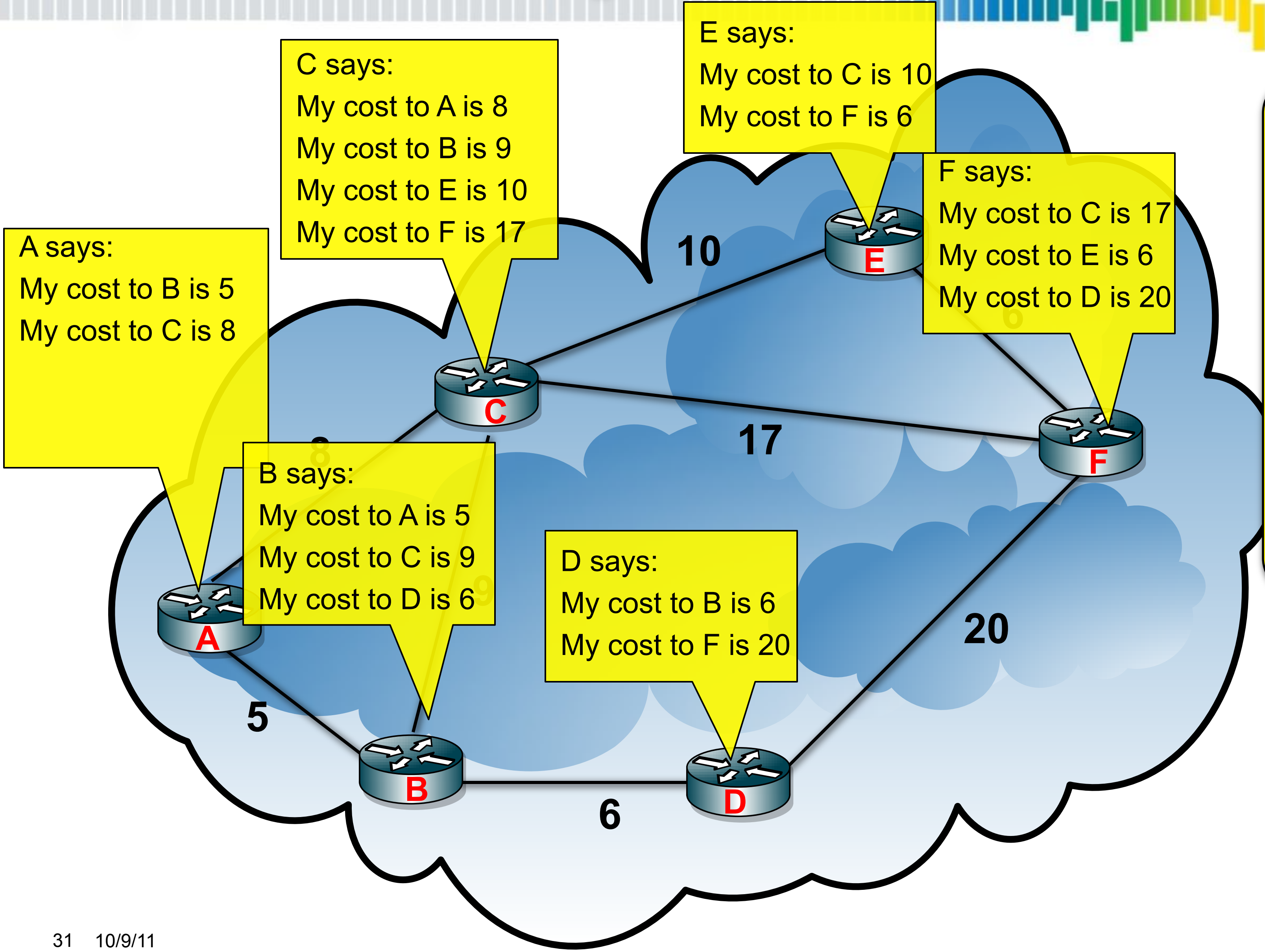
My cost to B is 5.

My cost to C is 8.

A – B = 5

A – C = 8

# Link-State Routing Protocols



A says:
My cost to B is 5
My cost to C is 8

C says:
My cost to A is 8
My cost to B is 9
My cost to E is 10
My cost to F is 17

E says:
My cost to C is 10
My cost to F is 6

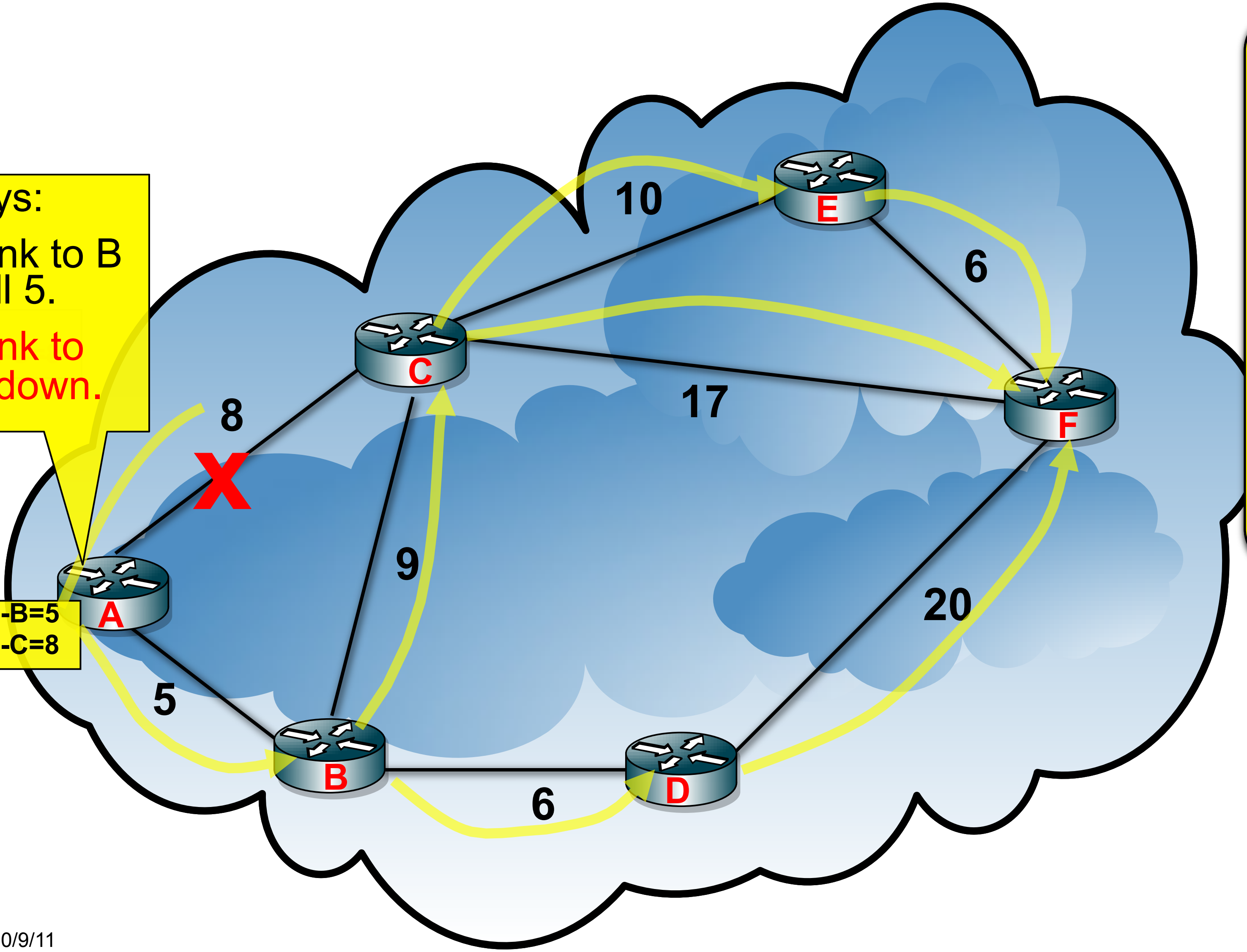
F says:
My cost to C is 17
My cost to E is 6
My cost to D is 20

B says:
My cost to A is 5
My cost to C is 9
My cost to D is 6

D says:
My cost to B is 6
My cost to F is 20

**A – B = 5**
**A – C = 8**
**B – C = 9**
**B – D = 6**
**C – E = 10**
**C – F = 17**
**D – F = 20**
**F – E = 6**

# Link-State Routing Protocols



A says:

My link to B is still 5.

My link to C is down.

A – B = 5

~~A – C = 8~~

B – C = 9

B – D = 6

C – E = 10

C – F = 17

D – F = 20

F – E = 6

# Link-state vs. Distance-Vector Routing

- Link-state protocols only need to flood information about what link or node has gone down
  - This information can be flooded independent of calculating a new best path
  - Each node them simply prunes off links or nodes that are no longer in service, and calculates a new best path
  - Link-state protocols converge very fast, as they do not depend on the change in topology to be recognized and then flooded
- Each node has a full picture of the topology
  - This does not scale well for large networks
    - Carries a lot of state
    - Each node (at least at a given hierarchy level) will see any instability in the network

# Link-state vs. Distance-Vector Routing

- Distance-vector protocols need only send updates when the best path changes
  - Distance-vector protocols need to first determine if a topology change causes a new best path to be selected
    - If so, a routing update must be crafted and sent
    - If not, the topology change is only locally significant, and no update is necessary
    - If there is no longer a best path, the previously best path must be withdrawn
- Each node does not have a full picture of the network
  - This scales better for large networks
    - Carries less state
    - Only nodes where the best path may change need to be sent an update / withdrawal
  - Takes longer for updates to cascade through the networks
    - Each node must process updates / withdrawals, then determine if there is a best path selection change, and only then build and send out a new update

# Why Use Distance-Vector for Internet Routing

- The large scale nature of the Internet routing table lends itself to distance-vector routing
  - Reduces the amount of state required
- Networks in the Internet are run by different organizations
  - No one organization has administrative control over the stability of the network
- Instability is a problem in large, disorganized, non-standardized networks
  - Reduces the amount of churn experienced network wide by localizing it
  - Provides a level of separation between administrative routing domains

# Why Use Link-State for Internal Routing

- Routing for the internal topology is much smaller than the global Internet
  - –Smaller amount of state required for internal topology can fit within limits of link-state protocols
- Internal topology generally run by a single organization
  - –Complete administrative control over the stability of the network
  - –Can set a single standard
- Converges very fast

# Bad 1990s Textbook Architecture

- Run BGP as your exterior gateway protocol (EGP) between routers at the edge of your AS, and the ASes you connect to
- Run your interior gateway protocol (IGP) usually OSPF or IS-IS between routers within your AS
- Redistribute between BGP and IGP on your edge routers facing another AS
- THIS DOES NOT SCALE
  - IGP becomes as large as the Internet table
  - Any instability in the Internet is translated into you IGP
  - Lose metrics when redistributing between protocols

# Good BGP / IGP design

- Run eBGP between your edge routers and other ASes to learn Internet routes
- Run iBGP within your AS to exchange Internet routes between your own routers
- Run an IGP within your AS to learn your internal topology
  - –Place a minimum of other routes into your IGP only if necessary
    - Loopbacks only (IS-IS) Loopbacks and backbone p-t-p (OSPF)
    - Anything that is a BGP next-hop
    - Any destination that requires IGP based load balancing
    - Any destination that requires fast convergence
    - All internal routes

# Good BGP / IGP design

- Best of both worlds approach
  - –BGP used for Internet routing
    - Converges slower but…
    - Has some level of isolation
    - Allows the Internet table to continue to scale
      - –Hundreds of thousands of routes
      - –Millions of paths
  - –IGP used for internal topology
    - Converges faster
    - Topology kept small so scaling is not an issue
    - Stability of topology is owned within the organization so stability is not a issue

# eBGP and iBGP

- eBGP is when your BGP neighbors belong to different ASes
  - Every best path is advertised across an eBGP boundary
  - eBGP neighbors are usually directly connected routers
  - Protocol next-hop is typically changed to advertising eBGP neighbor address
- iBGP is when your BGP neighbors belong to the same AS
  - If a path is learned via iBGP do not re-advertise to another iBGP neighbor
    - This requires a full mesh of iBGP peering (more on this later)
    - iBGP neighbors are generally not directly connected
  - Advertise every best eBGP learned path to iBGP neighbors
  - Protocol next-hop is typically unchanged

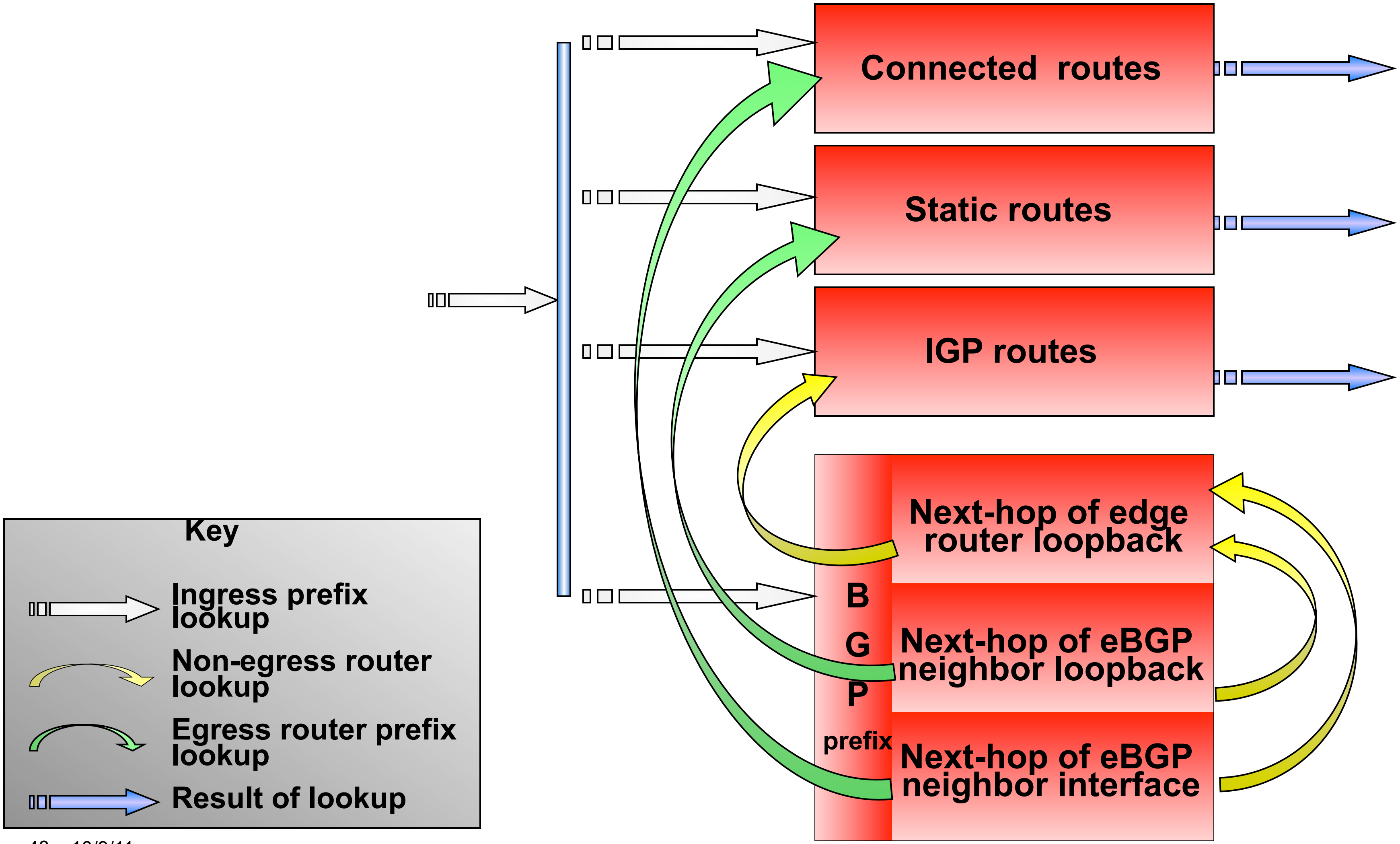
# BGP Protocol Next-hop

- A BGP route has a protocol next-hop that may be:
  - A connected interface
    - This is the case if your eBGP peer is directly connected to the local router and eBGP peering is between your interface IP address
  - A static route
    - This may be the case if your eBGP peer is directly connected to the local router and eBGP peering is between your loopbacks
  - A route reachable in your IGP
  - Another BGP route

# BGP Protocol Next-hop

- A BGP route has a protocol next-hop that may be:
  - A connected interface
  - A static route
  - A route reachable in your IGP
    - This may be the case if your eBGP peer is not peering with the local router
    - Edge routers that eBGP peer may set next-hop self so that routes advertised to the iBGP have a next-hop of the edge router’s loopback address which is in the IGP
    - Edge routers that eBGP peer may place directly connected interfaces passively into the IGP
    - Edge routers that eBGP peer may redistribute static routes into the IGP
  - Another BGP route
    - Edge routers that eBGP peer may redistribute static and connected routes into BGP

# BGP Protocol Next-hop Resolution



Connected routes

Static routes

IGP routes

BGP prefix

Next-hop of edge router loopback

Next-hop of eBGP neighbor loopback

Next-hop of eBGP neighbor interface

Key

Ingress prefix lookup

Non-egress router lookup

Egress router prefix lookup

Result of lookup

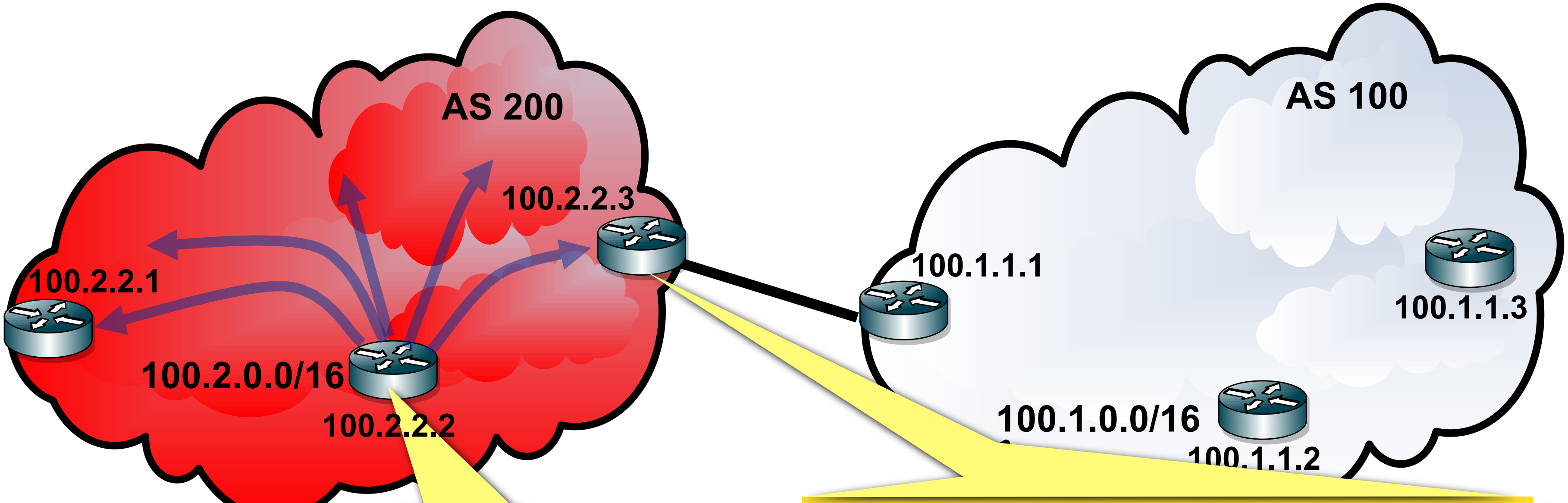
# BGP Protocol Next-hop Resolution

- Route regression is performed
  - Each route has a protocol next hop
    - If the protocol next hop is a directly connected interface, then the traffic can be forwarded
    - If the protocol next-hop is a static route, then the static route either points to a p-t-p interface which is reachable through a directly connected interface, or a next-hop on a broadcast medium with is reachable through some layer 2 protocol such as ARP
    - If the protocol next-hop is in the IGP, then the lowest cost IGP path and its egress interface can be determined by the IGP graph
    - If the next-hop is in BGP, then the BGP route to that next-hop is determined and the process is repeated

# Route Resolution for BGP Routes: Using Next-Hop Self

- In this first example AS100 will advertise a prefix to AS200  via eBGP
  - AS100 is using the AS100 side of the AS100 – AS200 interconnect as the BGP neighbor
  - When learned by AS200, this prefix will have the natural BGP next-hop of the of the AS100 neighbor address
- The AS200 edge router learning this prefix via eBGP will re-advertise the prefix to its iBGP and eBGP neighbors
  - It will set next-hop to "self" (its own loopback IP address) when advertising to the iBGP
  - Its loopback IP address is reachable in the IGP
- Non-egress routers learn the AS100 route via iBGP with a next-hop of the egress router' s loopback
  - That next-hop is reachable through the IGP

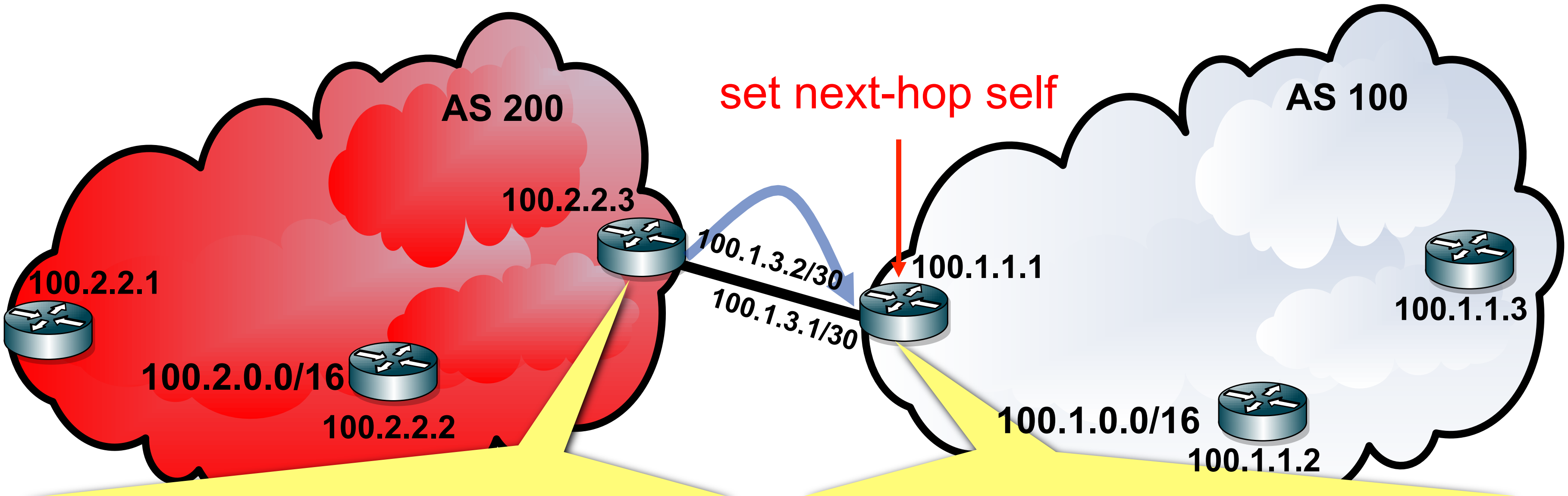
# AS 200 Routing Setup



```
100.2.0.0/16   *[Static/5] Discard
100.2.2.1/32   *[IS-IS/18] metric 1400
                  to 100.2.3.1 via ge-0/0/0
                > to 100.2.3.5 via ge-1/0/0
100.2.2.3/32   *[IS-IS/18] metric 1200
                > to 100.2.3.9 via ge-0/1/0
                  to 100.2.3.13 via ge-1/1/0
100.2.3.0/30   *[Direct/0] via ge-0/0/0
100.2.3.4/30   *[Direct/0] via ge-1/0/0
100.2.3.8/30   *[Direct/0] via ge-0/1/0
100.2.3.12/30 *[Direct/0] via ge-1/1/0
```

```
100.2.0.0/16 *[BGP/170] AS path: I
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
              Protocol next hop: 100.2.2.2
100.2.2.1/32 *[IS-IS/18] metric 2100
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
100.2.2.2/32 *[IS-IS/18] metric 1200
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
100.2.3.16/30  *[Direct/0] via ge-0/0/0
100.2.3.20/30  *[Direct/0] via ge-1/0/0
```

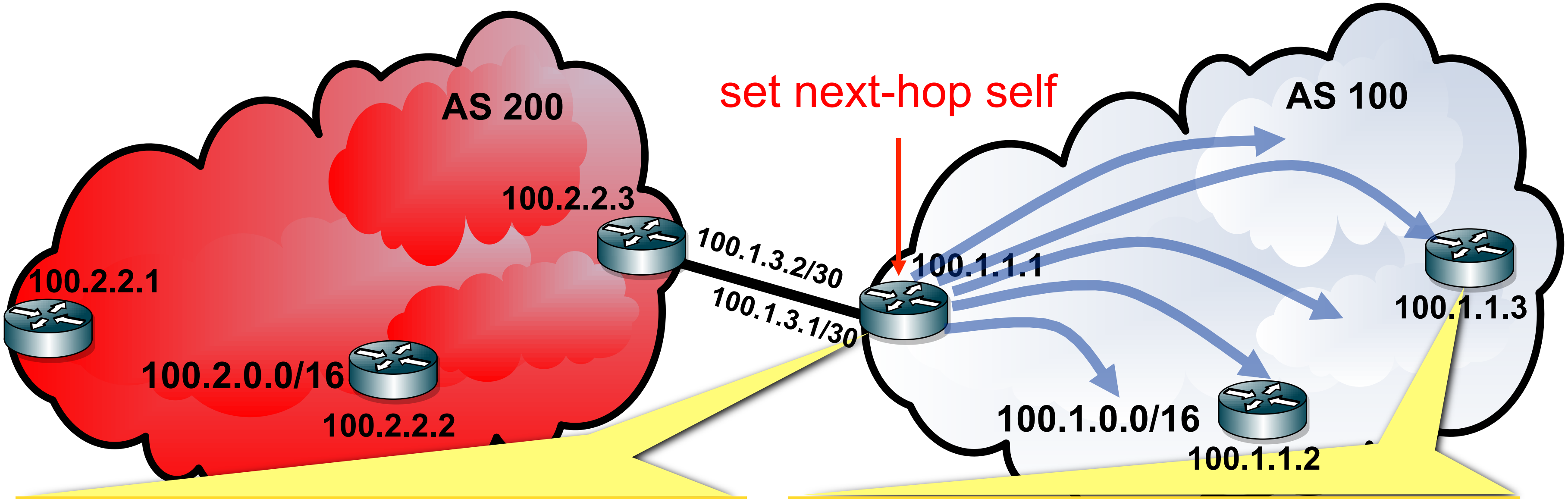
# AS 200 Advertises Route to AS100



```
100.1.3.0/30 *[Direct/0] via ge-1/1/0
100.2.0.0/16 *[BGP/170] AS path: I
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
              Protocol next hop: 100.2.2.2
100.2.2.2/32 *[IS-IS/18] metric 1200
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
100.2.3.16/30  *[Direct/0] via ge-0/0/0
100.2.3.20/30  *[Direct/0] via ge-1/0/0
```

```
100.2.0.0/16 *[BGP/170] AS path: 200 I
              > to 100.1.3.2 via ge-1/1/0
              Protocol next hop: 100.1.3.2
100.1.1.2/32 *[IS-IS/18] metric 2200
                to 100.1.3.5 via ge-0/0/0
              > to 100.1.3.9 via ge-1/0/0
100.1.1.3/32 *[IS-IS/18] metric 10010
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.3.0/30  *[Direct/0] via ge-1/1/0
100.1.2.4/30  *[Direct/0] via ge-0/0/0
100.1.2.8/30  *[Direct/0] via ge-1/0/0
```

# AS 100 Setting Next-hop self



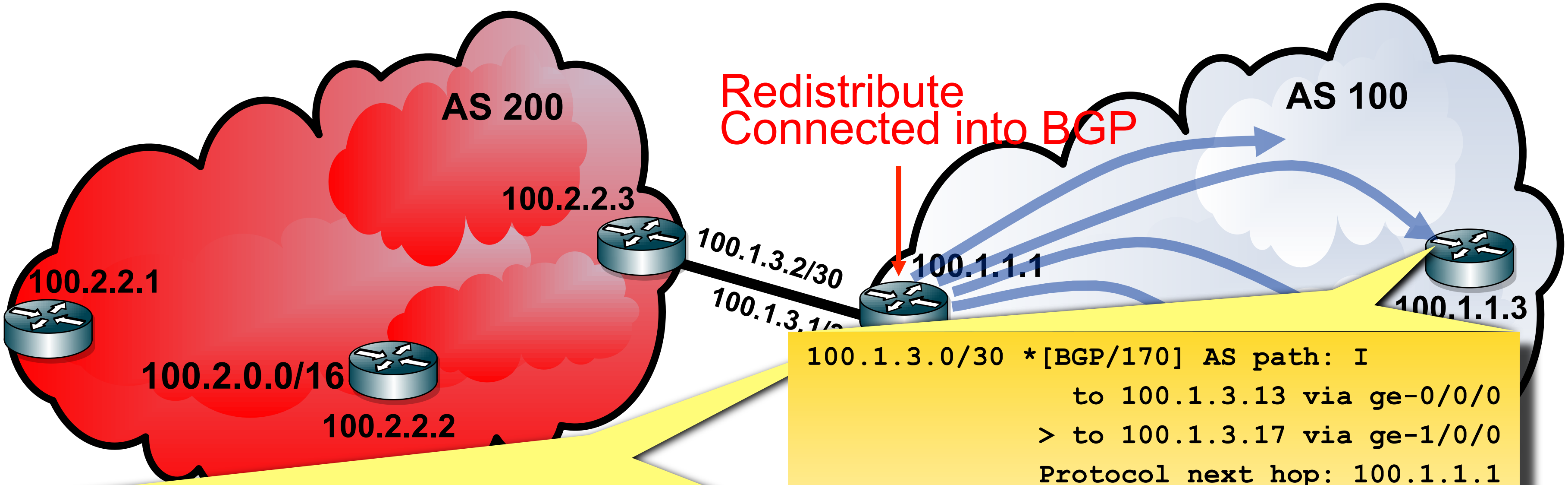
```
100.2.0.0/16 *[BGP/170] AS path: 200 I
              > to 100.1.3.2 via ge-1/1/0
              Protocol next hop: 100.1.3.2
100.1.1.2/32 *[IS-IS/18] metric 2200
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.1.3/32 *[IS-IS/18] metric 10010
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.3.0/30  *[Direct/0] via ge-1/1/0
100.1.2.4/30  *[Direct/0] via ge-0/0/0
100.1.2.8/30  *[Direct/0] via ge-1/0/0
```

```
100.2.0.0/16 *[BGP/170] AS path: 200 I
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
              Protocol next hop: 100.1.1.1
100.1.1.1/32 *[IS-IS/18] metric 10010
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
100.1.1.2/32 *[IS-IS/18] metric 2000
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
100.1.2.12/30 *[Direct/0] via ge-0/0/0
100.1.2.16/30 *[Direct/0] via ge-1/0/0
```

# Route Resolution for BGP Routes: Redistributing Connected into BGP

- In this next example AS100 will advertise a prefix to AS200 via eBGP
  - AS100 is using the AS100 side of the AS100 – AS200 interconnect as the BGP neighbor
  - When learned by AS200, this prefix will have the natural BGP next-hop of the of the AS100 neighbor address
- The AS200 edge router learning this prefix via eBGP will re-advertise the prefix to its iBGP and eBGP neighbors
  - This router will also redistribute the route for the AS100 – AS200 interconnect to BGP and advertise it to its iBGP and eBGP neighbors
    - The next-hop for this prefix will be the redistributing router's loopback IP address
  - Its loopback IP address is reachable in the IGP
- Non-egress routers learn the AS100 route via iBGP with a next-hop of the eBGP neighbor's IP address
  - That next-hop is reachable through the BGP prefix for the AS100 – AS200 interconnect
    - The route for the AS100 – AS200 interconnect has a next-hop of the egress router's loopback
      - That next-hop is reachable through the IGP

# AS 100
# Routing Setup



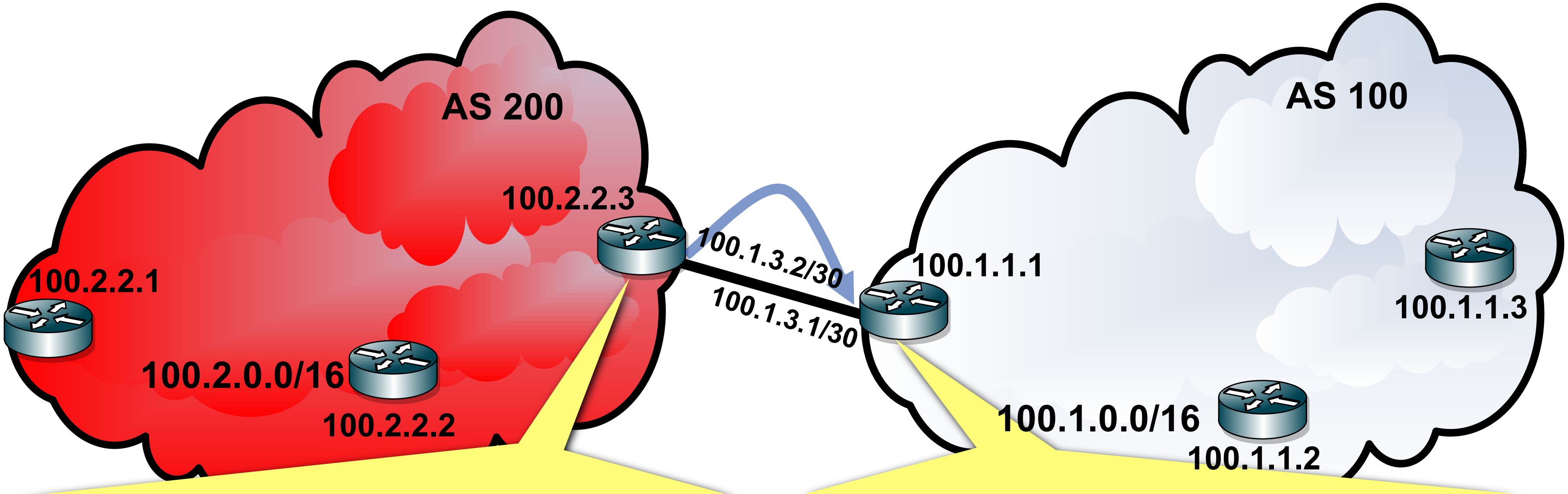
Redistribute
Connected into BGP

```
100.1.1.2/32 *[IS-IS/18] metric 2200
                  to 100.1.3.5 via ge-0/0/0
                > to 100.1.3.9 via ge-1/0/0
100.1.1.3/32 *[IS-IS/18] metric 10010
                  to 100.1.3.5 via ge-0/0/0
                > to 100.1.3.9 via ge-1/0/0
100.1.3.0/30  *[Direct/0] via ge-1/1/0
100.1.2.4/30  *[Direct/0] via ge-0/0/0
100.1.2.8/30  *[Direct/0] via ge-1/0/0
```

```
100.1.3.0/30 *[BGP/170] AS path: I
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
              Protocol next hop: 100.1.1.1

100.1.1.1/32 *[IS-IS/18] metric 10010
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
100.1.1.2/32 *[IS-IS/18] metric 2000
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
100.1.2.12/30 *[Direct/0] via ge-0/0/0
100.1.2.16/30 *[Direct/0] via ge-1/0/0
```

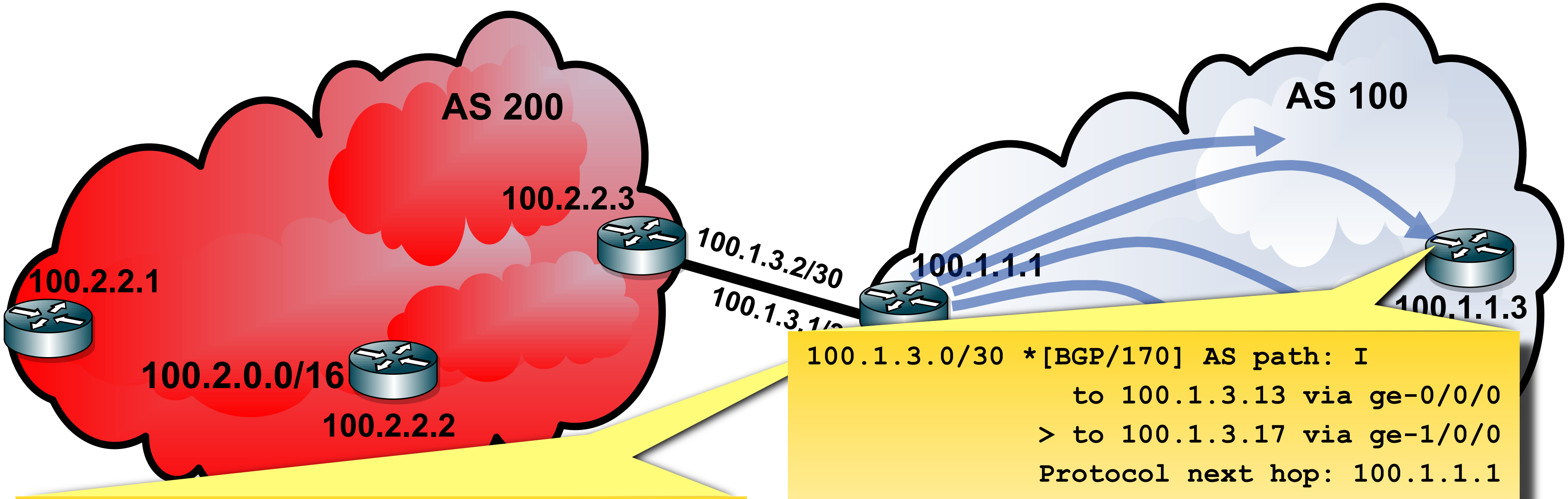
# AS 200 Advertises Route to AS100



```
100.1.3.0/30 *[Direct/0] via ge-1/1/0
100.2.0.0/16 *[BGP/170] AS path: I
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
              Protocol next hop: 100.2.2.2
100.2.2.2/32 *[IS-IS/18] metric 1200
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
100.2.3.16/30  *[Direct/0] via ge-0/0/0
100.2.3.20/30  *[Direct/0] via ge-1/0/0
```

```
100.2.0.0/16 *[BGP/170] AS path: 200 I
              > to 100.1.3.2 via ge-1/1/0
              Protocol next hop: 100.1.3.2
100.1.1.2/32 *[IS-IS/18] metric 2200
                to 100.1.3.5 via ge-0/0/0
              > to 100.1.3.9 via ge-1/0/0
100.1.1.3/32 *[IS-IS/18] metric 10010
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.3.0/30  *[Direct/0] via ge-1/1/0
100.1.2.4/30  *[Direct/0] via ge-0/0/0
100.1.2.8/30  *[Direct/0] via ge-1/0/0
```

# AS 100
# Redistributes Connected Routes into BGP



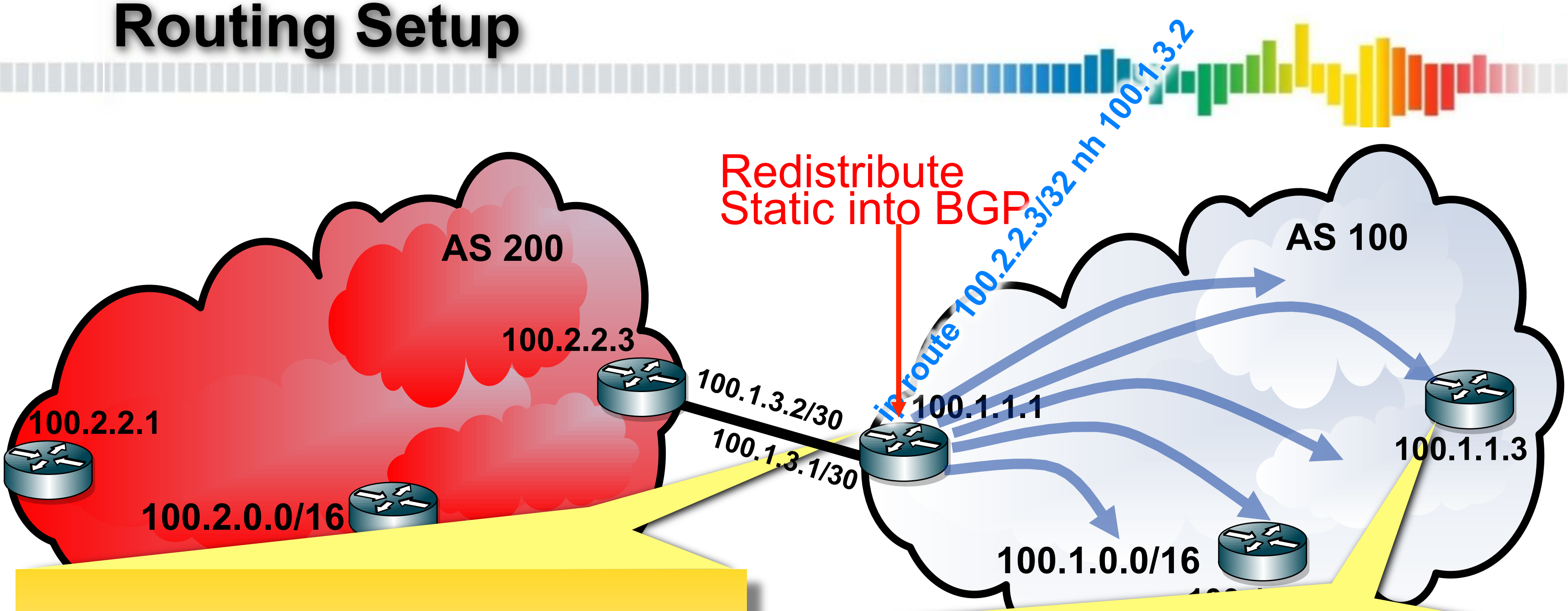
```
100.2.0.0/16 *[BGP/170] AS path: 200 I
              > to 100.1.3.2 via ge-1/1/0
              Protocol next hop: 100.1.3.2
100.1.1.2/32 *[IS-IS/18] metric 2200
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.1.3/32 *[IS-IS/18] metric 10010
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.3.0/30  *[Direct/0] via ge-1/1/0
100.1.2.4/30  *[Direct/0] via ge-0/0/0
100.1.2.8/30  *[Direct/0] via ge-1/0/0
```

```
100.1.3.0/30 *[BGP/170] AS path: I
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
              Protocol next hop: 100.1.1.1
100.2.0.0/16 *[BGP/170] AS path: 200 I
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
              Protocol next hop: 100.1.3.2
100.1.1.1/32 *[IS-IS/18] metric 10010
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
100.1.1.2/32 *[IS-IS/18] metric 2000
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
100.1.2.12/30 *[Direct/0] via ge-0/0/0
100.1.2.16/30 *[Direct/0] via ge-1/0/0
```

# Route Resolution for BGP Routes: Redistributing Static Routes into BGP

- In the last example AS100 will advertise a prefix to AS200 via eBGP
  - AS100 is its loopback IP address as the BGP neighbor
  - When learned by AS200, this prefix will have the natural BGP next-hop of the of the AS100 neighbor address
- The AS200 edge router learning this prefix via eBGP will re-advertise the prefix to its iBGP and eBGP neighbors
  - The AS200 router requires a static route for the AS100 neighbor’s loopback address as it is not directly connected
  - This router will also redistribute this static route to BGP and advertise it to its iBGP and eBGP neighbors
    - The next-hop for this prefix will be the redistributing router’s loopback IP address
  - Its loopback IP address is reachable in the IGP
- Non-egress routers learn the AS100 route via iBGP with a next-hop of the eBGP neighbor’s IP address
  - That next-hop is reachable through the BGP prefix for the AS100 neighbor’s loopback Address
    - That route has a next-hop of the egress router’s loopback
      - That next-hop is reachable through the IGP

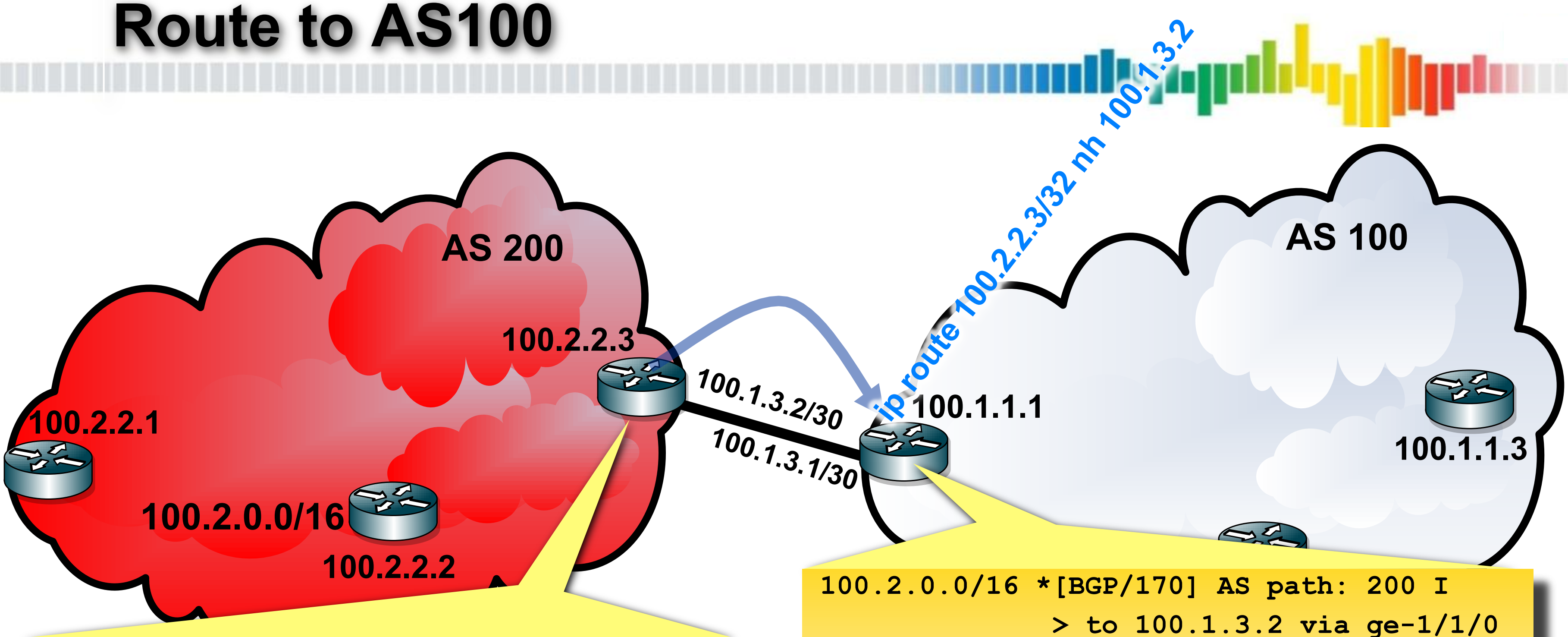
# AS 100
# Routing Setup



```
100.1.1.2/32 *[IS-IS/18] metric 2200
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.1.3/32 *[IS-IS/18] metric 10010
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.3.0/30  *[Direct/0] via ge-1/1/0
100.1.2.4/30  *[Direct/0] via ge-0/0/0
100.1.2.8/30  *[Direct/0] via ge-1/0/0
100.2.2.3/32  *[Static/5] 100.1.3.2
               via ge-1/1/0
```

```
100.2.2.3/32 *[BGP/170] AS path: I
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
              Protocol next hop: 100.1.1.1
100.1.1.1/32 *[IS-IS/18] metric 10010
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
100.1.1.2/32 *[IS-IS/18] metric 2000
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
100.1.2.12/30 *[Direct/0] via ge-0/0/0
100.1.2.16/30 *[Direct/0] via ge-1/0/0
```

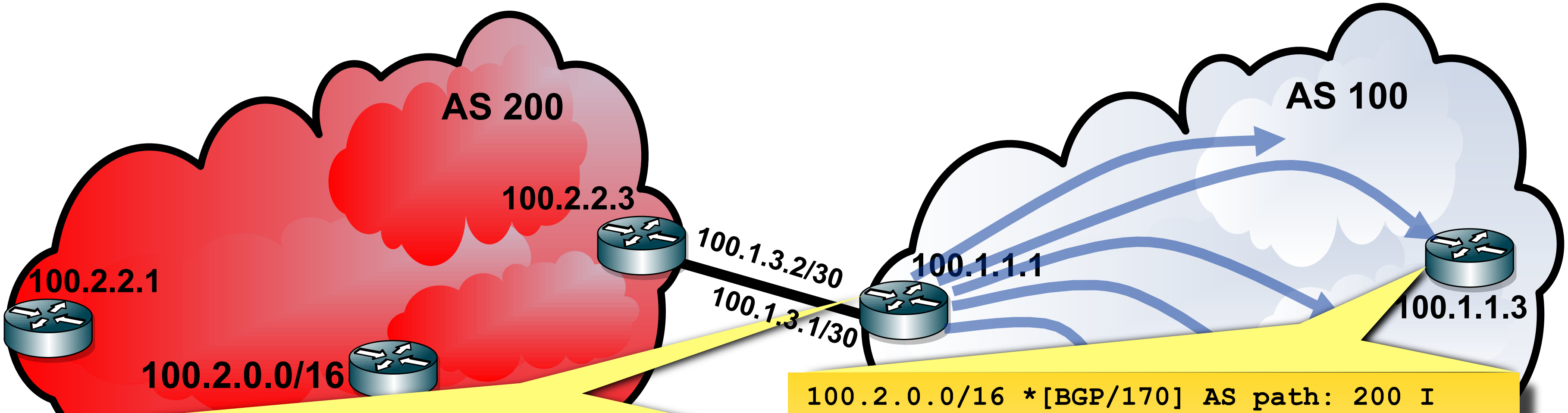
# AS 200 Advertises Route to AS100



```
100.1.3.0/30 *[Direct/0] via ge-1/1/0
100.2.0.0/16 *[BGP/170] AS path: I
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
              Protocol next hop: 100.2.2.2
100.2.2.2/32 *[IS-IS/18] metric 1200
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
100.2.3.16/30  *[Direct/0] via ge-0/0/0
100.2.3.20/30  *[Direct/0] via ge-1/0/0
```

```
100.2.0.0/16 *[BGP/170] AS path: 200 I
              > to 100.1.3.2 via ge-1/1/0
              Protocol next hop: 100.2.2.3
100.1.1.2/32 *[IS-IS/18] metric 2200
                to 100.1.3.5 via ge-0/0/0
              > to 100.1.3.9 via ge-1/0/0
100.1.1.3/32 *[IS-IS/18] metric 10010
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.3.0/30  *[Direct/0] via ge-1/1/0
100.1.2.4/30  *[Direct/0] via ge-0/0/0
100.1.2.8/30  *[Direct/0] via ge-1/0/0
100.2.2.3/32  *[Static/5] 100.1.3.2
               via ge-1/1/0
```

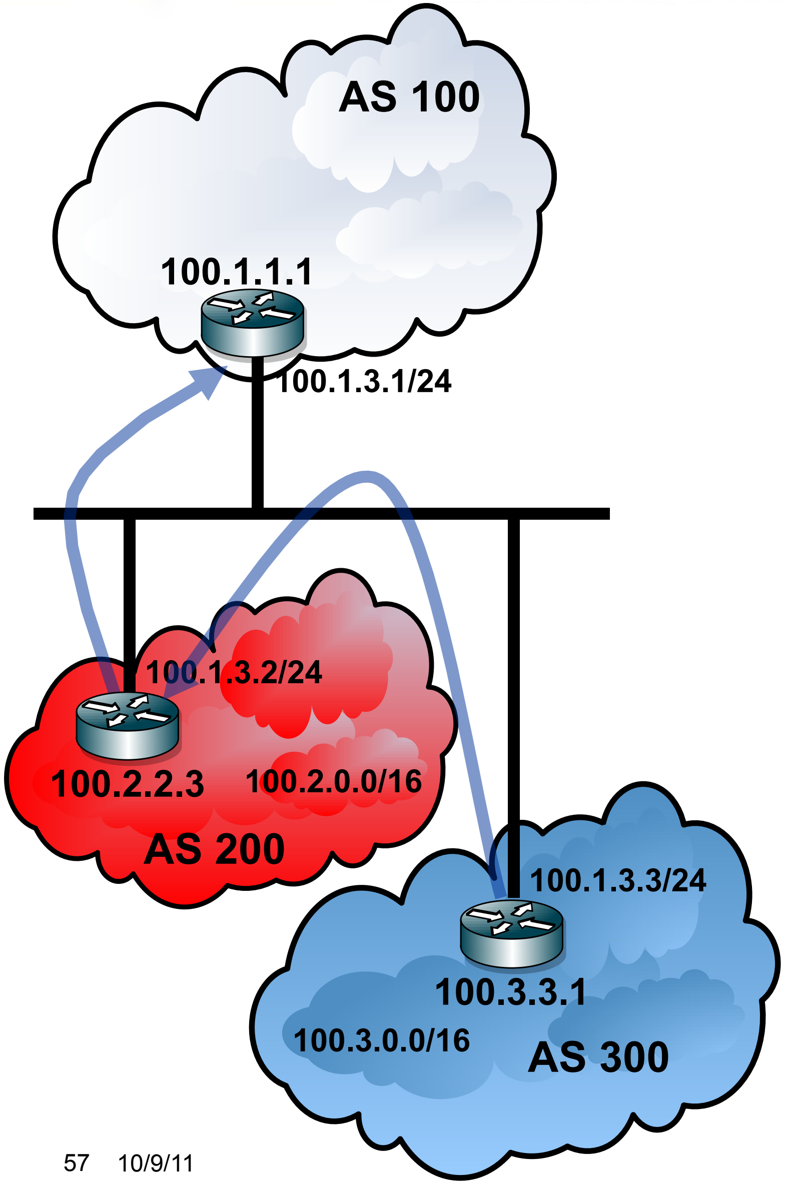
# AS 100
# Redistributes Static Routes into BGP



```
100.2.0.0/16 *[BGP/170] AS path: 200 I
              > to 100.1.3.2 via ge-1/1/0
              Protocol next hop: 100.2.2.3
 100.1.1.2/32 *[IS-IS/18] metric 2200
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.1.3/32 *[IS-IS/18] metric 10010
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.3.0/30  *[Direct/0] via ge-1/1/0
100.1.2.4/30  *[Direct/0] via ge-0/0/0
100.1.2.8/30  *[Direct/0] via ge-1/0/0
100.2.2.3/32  *[Static/5] 100.1.3.2
               via ge-1/1/0
```

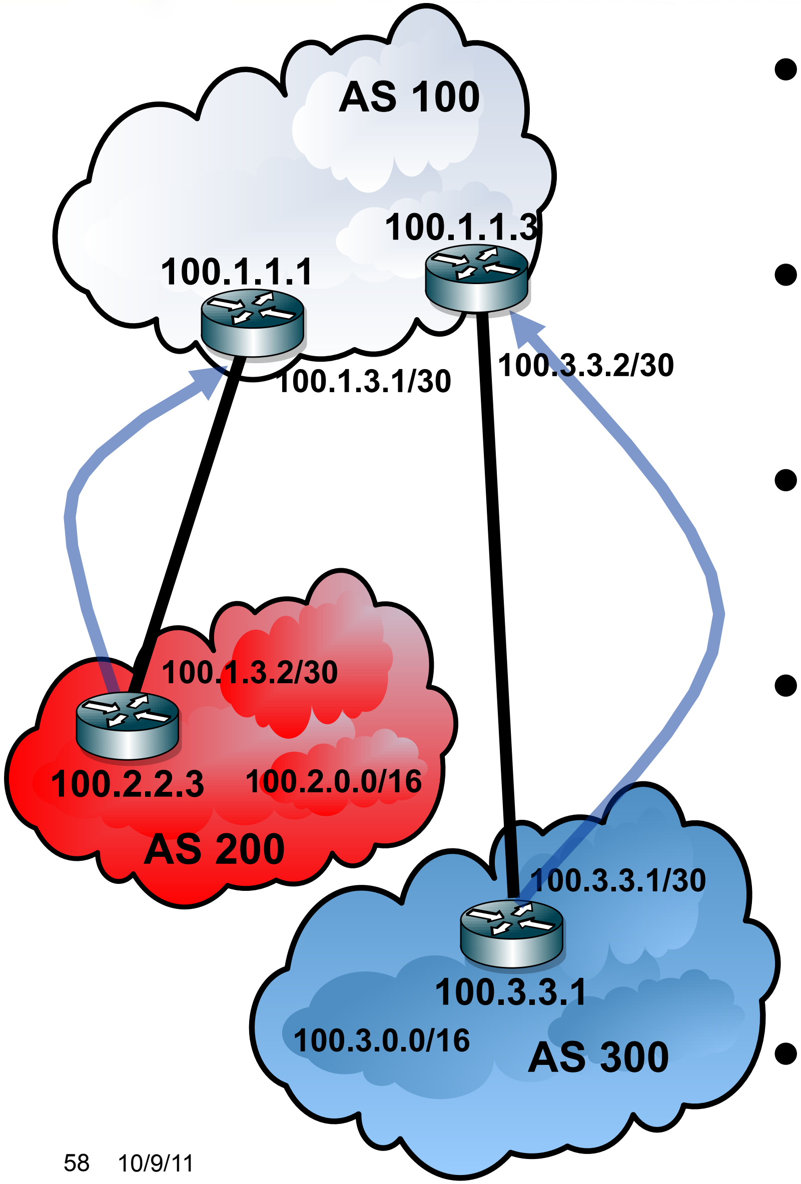
```
100.2.0.0/16 *[BGP/170] AS path: 200 I
              > to 100.1.3.2 via ge-1/1/0
              Protocol next hop: 100.2.2.3
100.2.2.3/32 *[BGP/170] AS path: I
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
              Protocol next hop: 100.1.1.1
100.1.1.1/32 *[IS-IS/18] metric 10010
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
100.1.1.2/32 *[IS-IS/18] metric 2000
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
100.1.2.12/30 *[Direct/0] via ge-0/0/0
100.1.2.16/30 *[Direct/0] via ge-1/0/0
```

# Third Party Next-hops



- AS300 eBGP peers with AS200
  - Advertises 100.3.0.0/16 NH 100.1.3.3
- AS200 eBGP peers with AS100
  - Advertises 100.2.0.0/16 NH 100.1.3.2
  - Advertises 100.3.0.0/16 NH 100.1.3.3
- AS100 will forward directly to AS300
  - More efficient
  - AS100 does not need extra configs to eBGP peer with AS300

# Third Party Next-hops Danger



- AS100 peers with AS200 and AS300
- AS200 and AS300 are competitors and do not peer
- One of AS200’s services is under a DoS attack
- AS200 send a route for the destination under attack with a third party next-hop of the AS300 router
- AS100 now forwards the DoS traffic to AS300

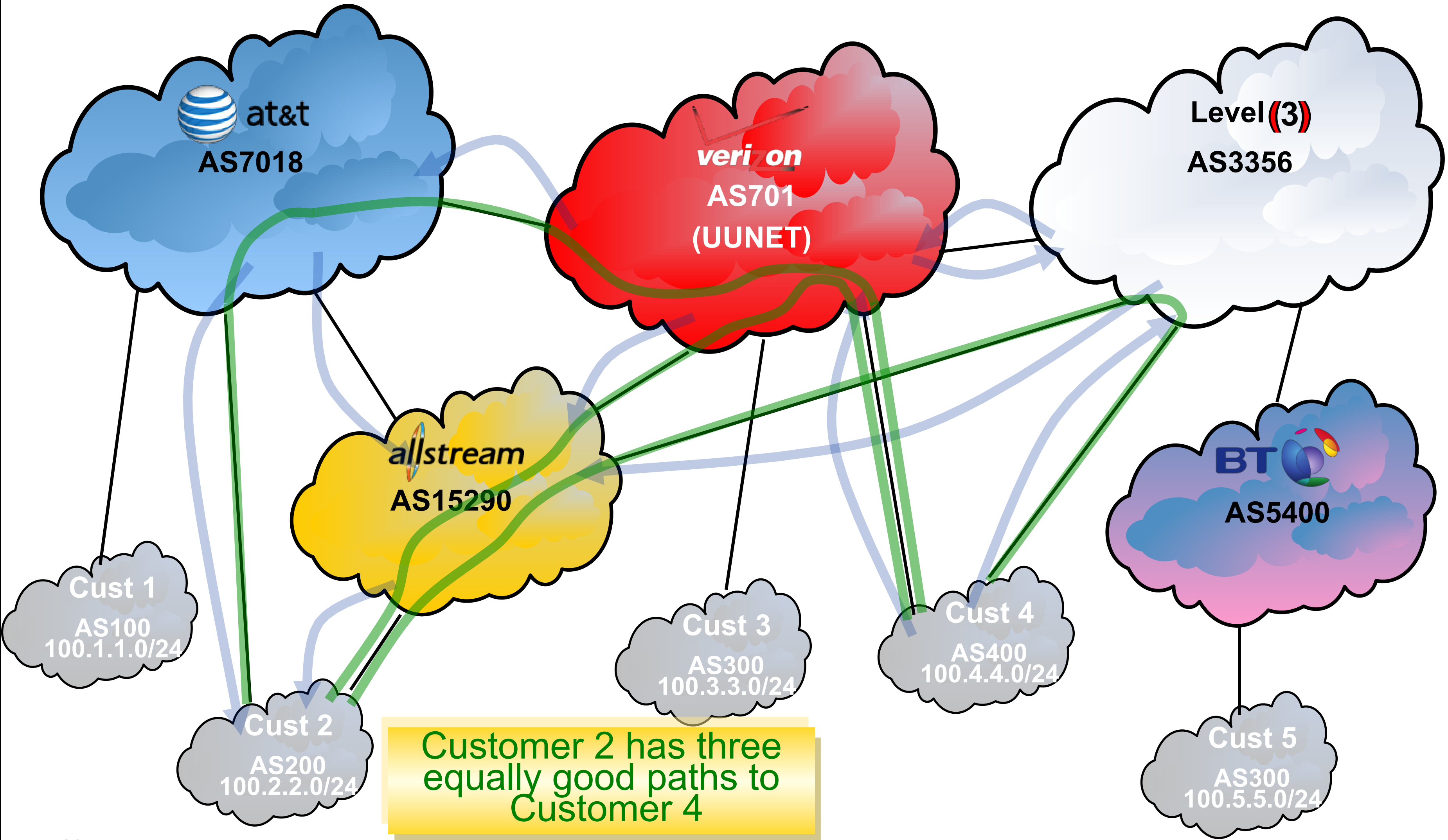
# Routing vs Forwarding

- Routing is the process where information about reachability of destinations is shared
  - –A router may learn multiple paths for a given route
    - These paths are stored by the routing engine (RE) or route processor (RP) in the Routing Information Base (RIB)
    - Path selection is performed by the RE or RP on this information
    - The best paths are chosen
  - –Occurs in the control plane

# Routing vs Forwarding

- Forwarding is the process of receiving a packet on an ingress interface, determining the egress interface, and sending the packet out of the router
  - The best paths for each route are stored by the forwarding engine (FE) in the Forwarding Information Base (FIB)
  - Occurs in the forwarding plane

# Forwarding Traffic at 30,000 Feet From Customer 2 To Customer 4



Customer 2 has three equally good paths to Customer 4

# BGP Path Selection

# BGP Path Selection Overview

- Validity
- Reachability
  - Is there a valid route to the BGP protocol Next-Hop
- Specificity
- Lowest protocol preference / Highest Cisco weight
- Higher Local Preference
- Locally Originated (Cisco)
- Shortest AS path length
- Locally Originated (Juniper)
- Lowest Origin code
  - IGP > EGP > Incomplete (Unknown)
- Lowest MED (null MED = 0)
  - Always-compare-med /cisco-non-deterministic-med / deterministic-med
- Prefer eBGP learned routes over iBGP learned routes
- Lowest IGP distance
  - Routes whose next-hop is static or connected have no IGP distance
    - Cisco makes no IGP distance comparison between a route with an IGP distance and a route without an IGP distance
    - Juniper compares a route with no IGP distance as a value of zero
  - Juniper: multiple routes for the next-hop, use lowest protocol preference, prefer inet.3 over inet.0, choose route with highest equal cost paths
- Shortest route-reflection cluster length (Juniper)
- Oldest external
  - Juniper on eBGP prefer oldest, unless always compare router-ID is enabled
- Lowest Router ID
- Shortest route-reflection cluster length (Cisco)
- Lowest Neighbor Address

# What are BGP attributes?

- A BGP message consists of one or more NLRI
  - –NLRI – Network Layer Reachability Information
    - A network prefix and its mask length
  - –All of the prefixes in the message must share the same set of attributes
  - –Attributes describe properties associated with the route(s)
  - –Attributes are encoded in TLVs (Type Length Value) making them easily extensible

# BGP Attribute Types

- Mandatory
  - Must be included in the update
  - Without it, the update is malformed
- Optional
  - Not all implementations need to understand this attribute
- Discretionary
  - Not required to be in the update
- Transitive
  - Can be shared between networks
- Non-transitive
  - Must be kept within the AS

# Validity / Reachability / Specificity

- Malformed
  - Invalid attribute value
  - Missing required attributes
  - Variable length attributes don’t match length
- Looped AS-path
- Looped cluster-ID list
- Enforce first AS and neighbor AS is not the first ASN
- iBGP protocol next-hop is not synchronized (Cisco)

Validity

- Is there a valid route to the protocol next-hop
  - Next hop is unreachable
  - Route reachable through itself
  - No route to next-hop

Reachability

- Prefer routes that are more specific
  - Prefer 1.2.3.0/24 over 1.2.0.0/16
- This is how prefix hi-jacking creates problems

Specificity

# Protocol Preference / Cisco Weight

- Lowest protocol preference
- Cisco Proprietary weight
  - Highest weight preferred
  - Locally configured value between 0 and 65,535
    - Per neighbor
    - Per prefix (filter-list)
    - Via route map policy
  - Default value for locally originated paths 32,767
  - Default value for BGP learned routes 0
  - Weight not propagated to BGP neighbors
- Juniper lacks weight, but can create similar behavior by manipulating locally configured protocol preference on specific BGP prefixes

# Protocol Preference

| Protocol | Juniper | Cisco |
|---|---|---|
| Direct / connected | 0 | 0 |
| System routes | 4 | -- |
| Static | 5 | 1 |
| EIGRP summary | -- | 5 |
| MPLS | 7 | -- |
| LDF | 8 | -- |
| LDP | 9 | -- |
| eBGP | -- | 20 |
| Internal EIGRP | -- | 90 |
| IGRP | -- | 100 |
| OSPF | -- | 110 |
| OSPF internal | 10 | -- |
| IS-IS | -- | 115 |
| IS-IS Level 1 internal | 15 | -- |
| IS-IS Level 2 internal | 18 | -- |
| Default | 20 | -- |
| Redirects | 30 | -- |
| Kernel | 40 | -- |
| SNMP | 50 | -- |

| Protocol | Juniper | Cisco |
|---|---|---|
| Router Discovery | 55 | -- |
| RIP | 100 | 120 |
| RIPng | 100 | -- |
| PIM | 105 | -- |
| DVMRP | 110 | -- |
| Routes to down interfaces | 120 | -- |
| Aggregate | 130 | -- |
| OSPF AS-external | 150 | -- |
| IS-IS Level 1 external | 160 | -- |
| IS-IS Level 2 external | 165 | -- |
| BGP | 170 | -- |
| EGP | -- | 140 |
| ODR | -- | 160 |
| External EIGRP | -- | 170 |
| iBGP | -- | 200 |
| MSDP | 175 | |
| Unknown | -- | 255 |

# Local Preference

- Highest Local Preference is preferred
  - –Default Local-pref is 100
  - –Local-pref can be set by policy
    - Used to modify outbound traffic
  - –Local-pref is sent to iBGP neighbors
  - –Local-pref is not transitive, and is lost between ASes
    - Transit providers typically support communities for customers to convey a desired local-pref value for their transit provider to set
    - Used to set inbound traffic

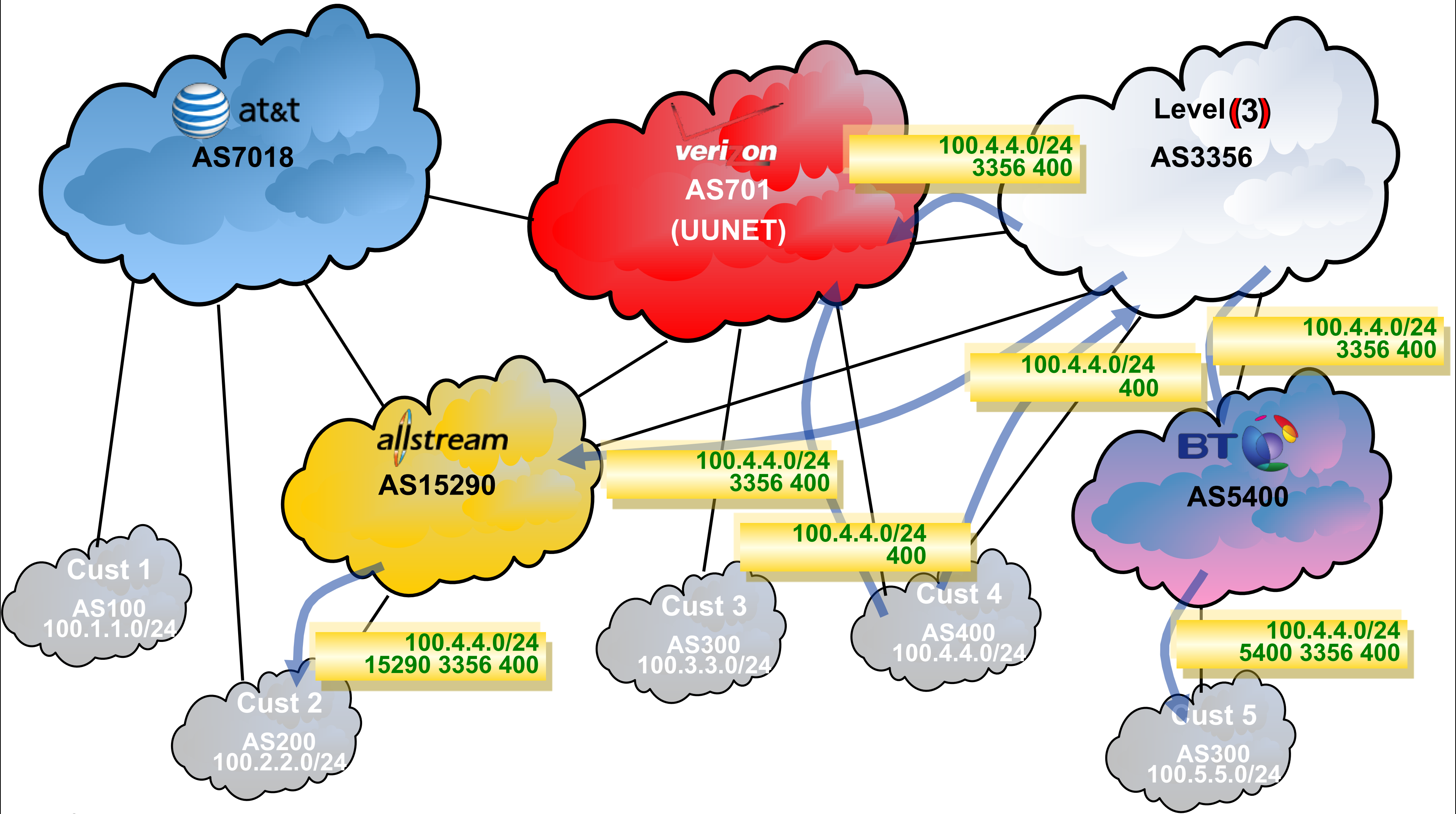
# Locally Originated (Cisco)

- Prefer routes that the local router redistributed into BGP
  - Prefer paths redistributed into BGP by network statements
  - Prefer paths that are redistributed from another protocol via the redistribute command
  - Prefer paths that are redistributed into BGP from the aggregate command

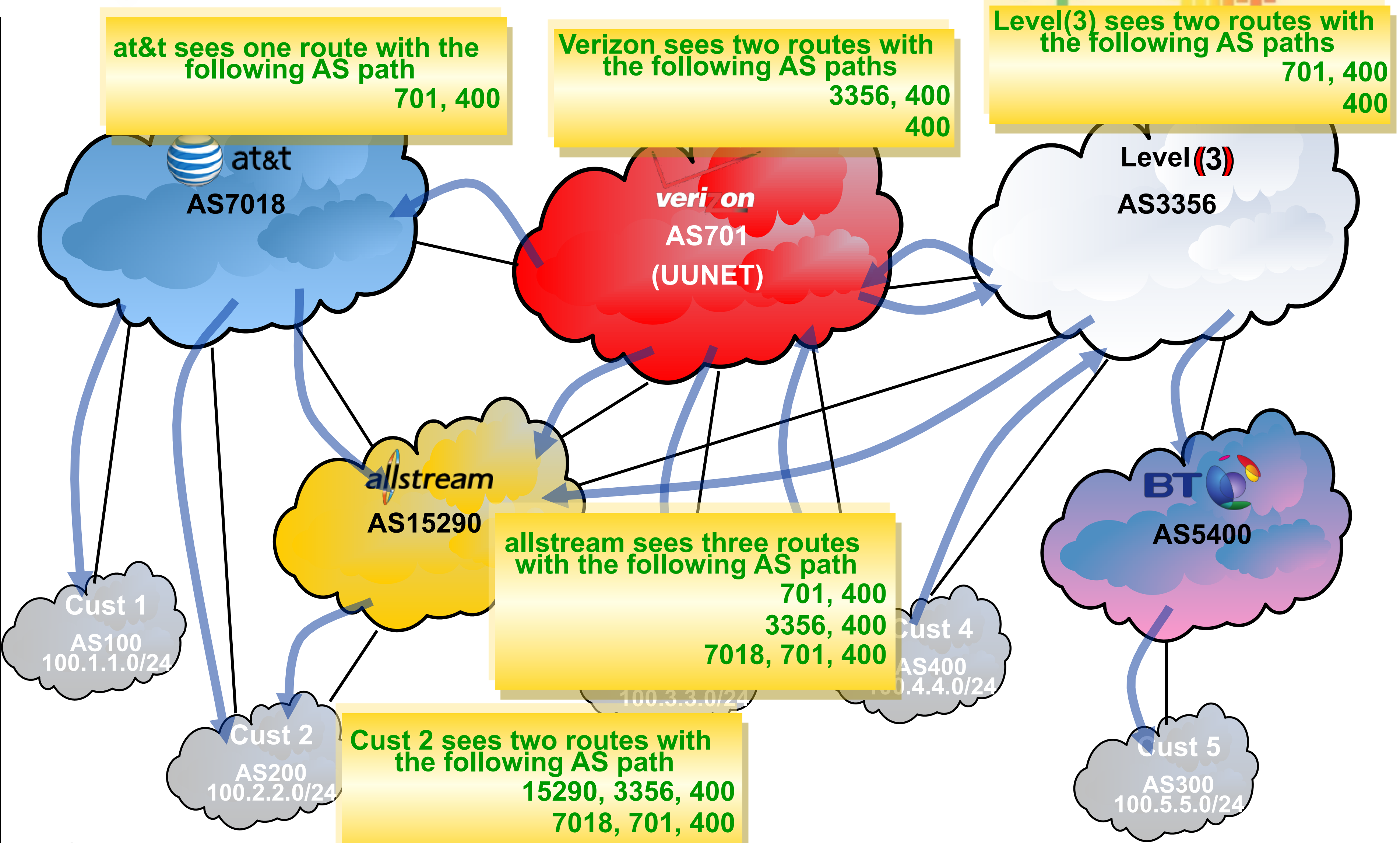
# Shortest AS-Path

- As a route propagates from network to network the AS is added to the AS path
  - Origin is to the right
  - Neighboring AS is to the left
  - AS is added to the AS Path when exiting an AS
  - AS sets count as one AS
  - Confed ASes do not count
- AS Path is used to prevent routing loops
- AS Path is also used a a metric of distance
  - A route with fewer number of ASes in the AS path is better than a route with more ASes in the AS path
    - A route passing through many networks is likely to be longer than a path passing through few networks
    - AS path length is a poor measure of distance
      - Is possible the distance across two small ASes is shorter than one very large AS
      - Unfortunately it is the only measure of end to end distance for the Internet

# AS Paths


at&t
AS7018
verizon
AS701
(UUNET)
Level(3)
AS3356
allstream
AS15290
BT
AS5400
Cust 1
AS100
100.1.1.0/24
Cust 2
AS200
100.2.2.0/24
Cust 3
AS300
100.3.3.0/24
Cust 4
AS400
100.4.4.0/24
Cust 5
AS300
100.5.5.0/24
100.4.4.0/24
3356 400
100.4.4.0/24
400
100.4.4.0/24
3356 400
100.4.4.0/24
3356 400
100.4.4.0/24
400
100.4.4.0/24
15290 3356 400
100.4.4.0/24
5400 3356 400

# AS Path



at&t sees one route with the following AS path
701, 400

Verizon sees two routes with the following AS paths
3356, 400
400

Level(3) sees two routes with the following AS paths
701, 400
400

allstream sees three routes with the following AS path
701, 400
3356, 400
7018, 701, 400

Cust 2 sees two routes with the following AS path
15290, 3356, 400
7018, 701, 400

# Locally Originated (Juniper)

- Prefer strictly internal paths
  - –IGP routes
  - –Static, Direct, Local
- Juniper does not redistribute routes into BGP
  - –Well, not easily… you can export/import between RIBs

# Origin

- Lowest Origin is preferred
- IGP > EGP > incomplete / unknown
- Used by the router that places the route into BGP to indicate what protocol the route came from
  - **Supposed to indicate the reliability of the routing information**
  - **IGP has the most up to date information and is most reliable thus most preferred**
  - **EGP indicates the route was redistributed from EGP into BGP**
    - **Is less reliable than the IGP and thus less preferred than IGP**
    - **Was useful during the transition from EGP to BGP**
    - **No networks on the Internet use EGP**
    - **Some now view this to mean any EGP such as BGP**
    - **Some now view this to mean any other dynamic routing protocol that is not your IGP**
  - **Incomplete / unknown is the least reliable type of route, where no dynamic routing information is available such as a static route**

# Origin
# Default Behavior

- Inconsistent default behavior
  - –By default when redistributing into BGP Juniper sets origin as IGP and Cisco as incomplete
  - –Has lead some transit providers to conclude this attribute is not used, allowing them to freely stomp on the origin value to influence path selection

# Mutli-Exit Discriminator (MED)

- Inter-AS, non-transitive, optional attribute
  - Can send a MED value to a neighboring AS, but it will not propagate beyond that AS
- Lowest MED is preferred
- MED is intended to be used when there are multiple exits to a given destination
  - The downstream AS can indicate which entrance points are better
  - This is typically accomplished by conveying the IGP distance in the MED value
  - The down stream AS can also set an arbitrary value such as 10 and 20 to create a primary / backup configuration with a single upstream AS
- MED values are only compared when the neighbor AS is the same
- If MED is not set, the assumed value is 0

# Default MED Value

- If MED is not set, the assumed value is 0
  - –This turns out to not be helpful
  - –Cannot set a MED value to make one link better than default
  - –Must set all the other links to be worse
- Discussion about making the missing MED value be ∞ or the highest value
  - –Some implementations treat missing MED differently
  - –∞, (2^32)-1, (2^32)-2

# Med Options

- Always Compare MED
  - –Compares the MED value of two paths even if the neighbor AS is different
- Cisco Non-deterministic MED
  - –Cisco only does a pair-wise MED comparison
    - The first route becomes active
    - The next route is compared to the active route
    - And so on
  - –Order the routes are learned in could change the best path selection
    - More on MED Oscillation later

# Prefer eBGP Learned Paths Over iBGP Learned Paths

- Prefer paths learned from an eBGP neighbor over an iBGP neighbor
  - This maximizes the number of paths sent to the iBGP

# Lowest IGP Distance

- Prefer paths with the lowest IGP distance to the protocol next-hop
  - If the protocol next-hop is a BGP route, recurse until you resolve a route in the IGP
- Some vendors treat static and direct routes as an IGP cost of 0
- Some vendors treat static and connected routes as a null IGP cost and make no comparison

# Shortest Route-Reflection Cluster List (Juniper)

- Prefer routes with the shortest route-reflection cluster list

# Oldest eBGP Path

- For eBGP learned routes, prefer the older route
  - –At this point the tie breaker is nearly arbitrary
  - –Preferring the old path reduces unnecessary churn
  - –Preferring the oldest path is non-deterministic
    - A flap in a long lived BGP session could cause a lot of traffic to slosh
- Can over-ride with always compare router-ID

# Lowest Router-ID

- Prefer the path with the neighbor that has the lowest router-ID
  - –4.4.4.4 is lower than 8.1.1.1
  - –Each router has a globally unique router-ID

# Shortest Route-Reflection Cluster List (Cisco)

- Prefer routes with the shortest route-reflection cluster list

# Lowest Neighbor Address

- If a router has multiple eBGP sessions to the same router, then the router-ID for those sessions will be the same
- In that case choose the lowest neighbor address
  - A router cannot have two BGP sessions to the same IP address

# Other Attributes

# BGP Communities

- The ability to put one or more labels on a route
- Communities are transitive and optional
- Defined in RFC-1997
- 32 bit number
  - Usually represented as two 16-bit numbers separated by a colon
  - Typical format is ASN:<action-value>
    - 701:80 – AS701 should set local-pref of 80 on this route when learned from a customer
    - 7018:80 – AS7018 should set local-pref of 80 on this route when learned from a customer
- Useful to classify a type of route, and then write policy to manipulate that type of route
- Useful for informational purposes, such as determining the location the route was learned

# Extended BGP Communities

- Extended by RFC-4360
- 64 bit number
  - First 8-bits indicate type “regular” or “extended”
    - Can indicate transitive or non-transitive
  - For extended the second 8-bits indicate sub-type
    - 16-bit ASN : 32-bit value
    - 32-bit ASN : 16-bit value
- Useful to encode 32-bit ASN : action
- Useful to extend the number space for more action values for 16-bit ASNs
- Useful for network based VPN to signal route target
  - Which routing instance the address belongs to

# Well Know Communities

- Several well known communities
  - www.iana.org/assignments/bgp-well-known-communities
  - No-export 65535:65281
    - Do not advertise out side the local AS
  - No-advertise 65535:65282
    - Do not advertise to any BGP neighbor
  - No-export-subconfed 65535:65283
    - Do not advertise outside of the local sub-AS (used only with confederations)
  - No-peer 65535:65284
    - Do not advertise to bi-lateral Peers (RFC3765)

# Common Transit Provider Communities

- Set local-pref
- Prepend transit provider’s AS towards Peers
- Don’t leak prefix
  - –To Peers
  - –Out of the continent
  - –Out of the country
  - –Out of the state
- Some providers offer this behavior targeted to a short list of ASes

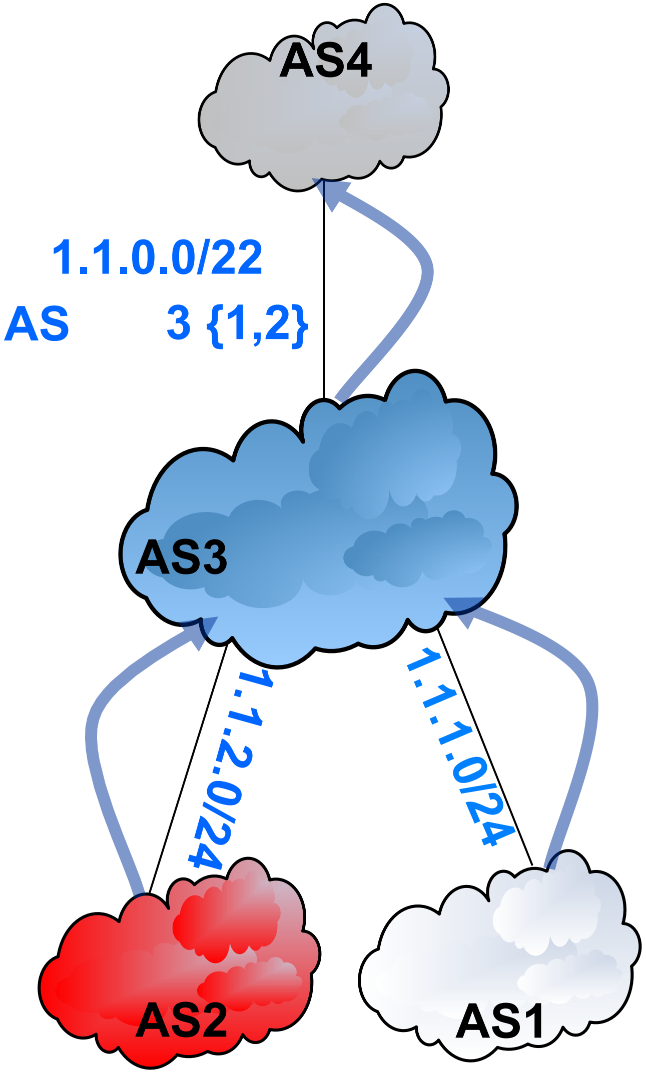
# Community Stripping

- Default behavior
  - –Juniper sends communities by default
    - Communities must be removed via policy if this behavior is not desired
  - –Cisco does not send communities by default
    - Send-community has to be enabled if this behavior is not desired
- Standard Peering behavior
  - –Settlement free Peers do not typically allow their Peers to control traffic engineering on their network
  - –Peers typically strip all communities and flatten MEDs

# Aggregator

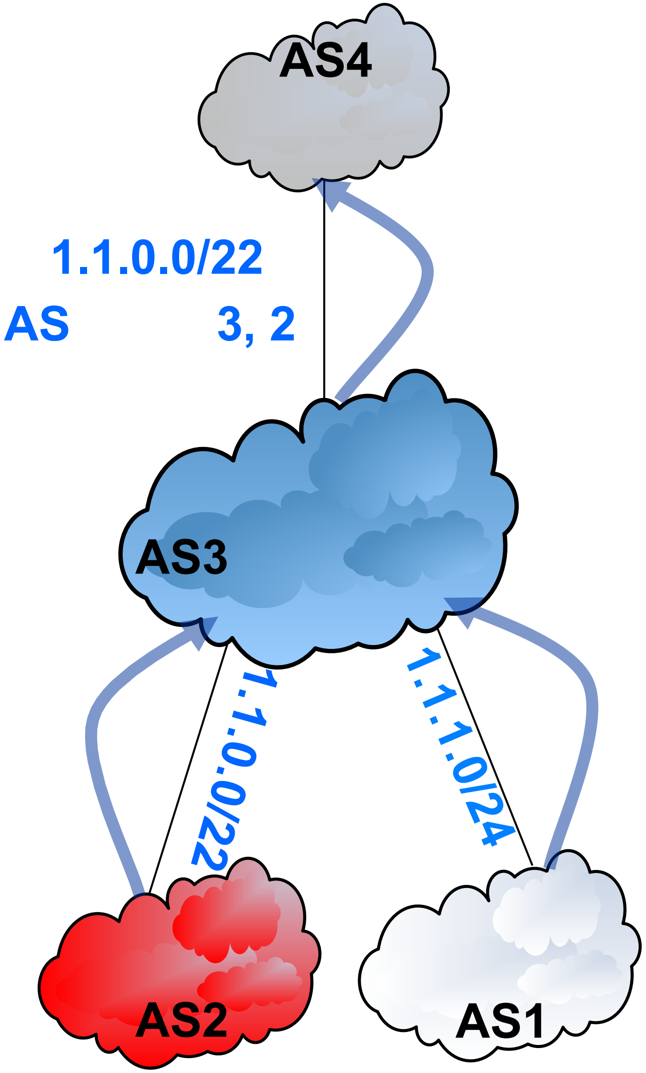
- Optional, transitive attribute
- Used when a router receives one or more routes that it aggregates into a supernet
- The Aggregator is the RID of router performing the aggregation
  - –Useful for debugging purposes
- The aggregate route can inherit the AS paths of all contributing routes and store them in an AS set
  - –Policy can be crafted to limit the routes contributing to the aggregate, and thus the ASes in the AS set

# Aggregate Route Example



- AS1 advertises 1.1.1.0/24 to AS3
- AS2 advertises 1.1.2.0/24 to AS3
- AS3 aggregates 1.1.1.0/24 and 1.1.2.0/24 into a less specific route say 1.1.0.0/22
- AS3 advertises the aggregate 1.1.0.0/22 to AS4
  - –AS3 can suppress the announcement of 1.1.1.0/24 and 1.1.2.0/24
  - –AS3 can originate the aggregate 1.1.0.0/22
  - –Or can use an as-set to inherit the AS path of the contributing routes

# Atomic Aggregate Route Example



- AS1 advertises 1.1.1.0/24 to AS3
- AS2 advertises 1.1.0.0/22 to AS3
- AS3 aggregates 1.1.1.0/24 and 1.1.0.0/22 into a less specific route say 1.1.0.0/22
- AS3 advertises the aggregate 1.1.0.0/22 to AS4
  - –AS3 can suppress the announcement of 1.1.1.0/24 and advertise the atomic aggregate of 1.1.0.0/22
  - –AS3 sets the atomic aggregate flag to indicate some routing information of more specific routes are being suppressed

# AS4path

- AS path is not capable of carrying 32-bit ASNs
- Changing AS path to carry 32-bit ASNs would make it non-backward compatible
- AS23456 is used as a “place holder” to represent a 32-bit ASN in the AS path
- AS4path contains the actual 32-bit ASN
- Non capable 32-bit ASN routers use the AS path, and pass the AS4path information along
- Capable 32-bit ASN routes construct the AS path from the (16-bit) AS path and the (32-bit) AS4path

# BGP 102

Jason Schiller
Google Network Engineering
jschiller@google.com

# BGP 102 Outline

- **BGP Policy**
  - –Common BGP Policy
- **BGP TE**
  - –Affecting Path Selection
- **iBGP Scaling**
  - –Route Reflection
  - –Confederations
- **Operational Considerations**
- **BGP Mechanics**
  - –Configuring BGP
  - –Troubleshooting BGP
  - –BGP Protocol

# BGP Policy

# ISP Customer Route Filtering

- Best ISP practice is to filter customers by prefix
  - Customers can only advertise a route or a more specific route that they are authorized to use
    - Direct allocation from an RIR to the customer
    - Assigned to a customer by an ISP and properly SWIP'ed
  - Often difficult in practice
    - Some customers have tens of thousands of entries
      - Some may be customers of customers with direct allocations (hard to verify)
    - Some customers update thousands of lines per week
    - Near misses on company names
    - Branch office of an organization asserting ability to use corporate office space
  - Peering through an IRR can help
    - If everyone uses it and it is kept accurate
- Some ISPs will do AS path filtering and max-AS

# ISP Peer Route Filtering Policy

- Often difficult know what a Peer should advertise to you
  - Similar to customers with down stream customers
  - Much larger scale
  - Peers often do not want to reveal business relationships
- Restrict routes Peers should not advertise
  - RFC-1918 (not globally routable)
  - Bogons (unallocated space)
    - This list is constantly changing, and if not kept current creates issues
  - Internal trusted space
- Peer AS filtering
  - Prevent Peers from advertising routes that have an AS in the AS path that is likely to be one of their Peers

# ISP Route Aggregation

- Static customers using a more specific of their provider’s address space do not need to be advertised to the Internet
- BGP customers connecting to only one upstream AS and using a more specific of their upstream do not need to be advertised to the Internet
- The provider needs to carry all of the more specific routes to deliver traffic to these customers, but the Internet only needs to know about the aggregate to reach the provider

# ISP Route Aggregation Goals

- Offer a stable route to the ISP aggregate
- Keep the Internet routing table small
  - –Suppress more specific routes of static and singly homed BGP customers that use provider space
- Do not suppress routes of static customers that use their own space
- Do not suppress routes of multi-homed BGP customers
  - –Suppressing routes of multi-homed BGP customer will impact their traffic engineering

# Typical ISP
# Route Aggregation Implementation

- Create a pull-up / tie-down
  - –Statically route ISP aggregate to null0 / discard
  - –Redistribute route into BGP
  - –Do not place pull-up routes on edge routers
    (do not use aggregate or network on edge routers)
    - If an edge router become isolated from the core it will not advertise the ISP aggregates
  - –Place route on all core routes
    - If an edge router has reachability to any portion of the core, the the ISP aggregates will be announced
    - Route will persist even if many core routers go down

# Typical ISP
# Route Aggregation Implementation

- Create a pull-up / tie-down
- Maintain a prefix-list of ISP aggregates
  - –Match on more specific of ISP aggregate when re-distributing from static and tag with a BGP community
  - –Don’t add this community when redistributing the aggregate pull-up routes!
  - –Match on more specific of ISP aggregate when learning routes from multi-connected eBGP customers
    - Match on neighbor RFC-2270 ASN or private ASN
- On eBGP match on this community and deny announcing this prefix
  - –Not needed if the well known no-export community is used

# BGP TE

# BGP TE Policy

- There are 3 basic flavors of BGP multi-homing
  - –Best path
  - –Primary / backup
  - –Load sharing across all links
- Dialing traffic
  - –Additional requirement to shift traffic between links with any of the three options, such as pushing traffic away from over utilized links
- Complex combinations of the three cases

# Best Path

- Requires the ability to use a non-random “best” path (For some definition of best).
- Current best path is based on routing information based on BGP path selection algorithm
- Best path approximates “shortest” path to destination
- This is the default behavior

# Primary / Backup

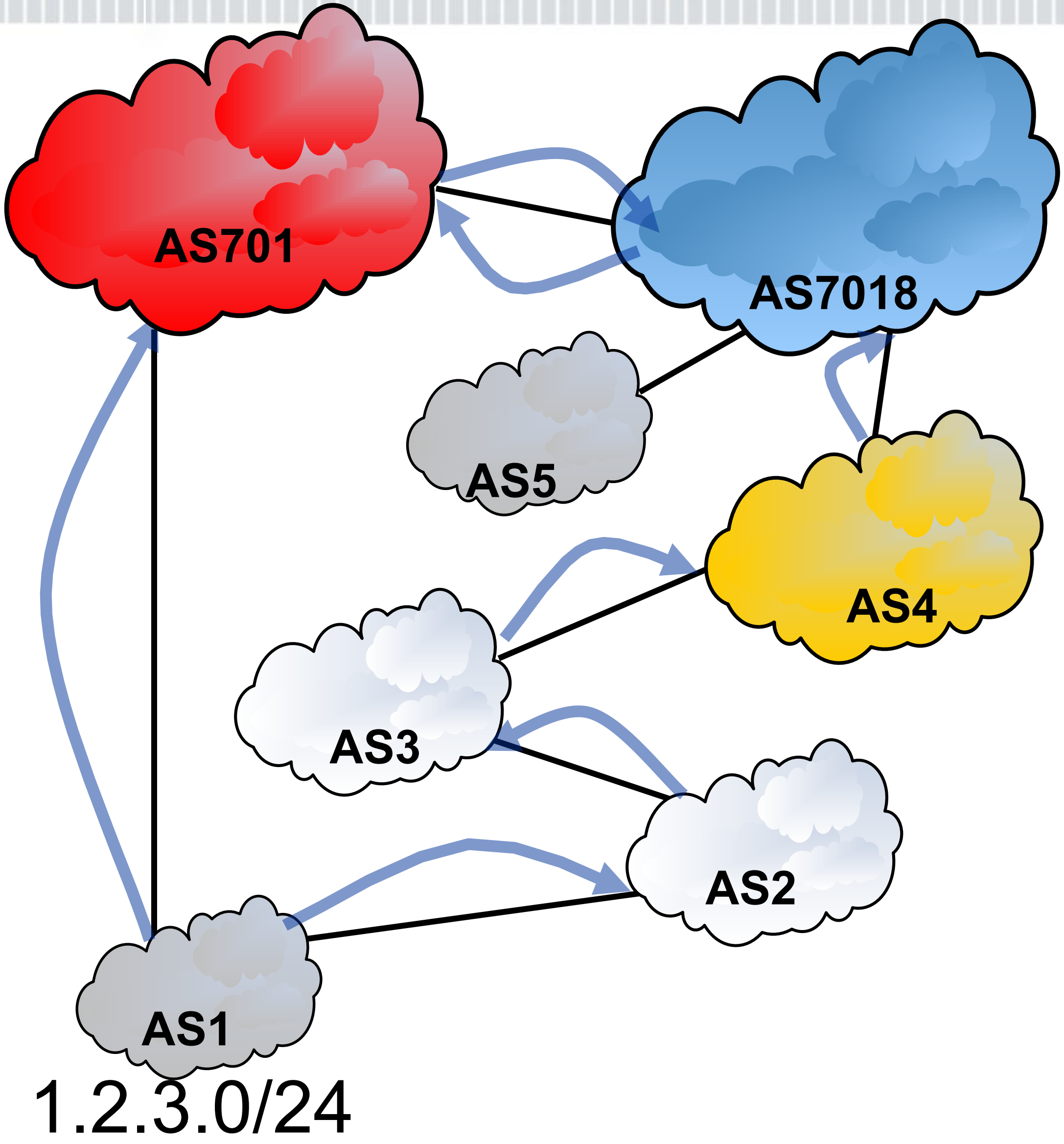
- Requires the ability to designate a link or set of links as the primary link to use for all traffic to one or more destination prefixes.
  - Primary link should carry all traffic for the designated prefixes.
  - The backup link should only carry traffic if the primary goes down.
- Common reason for primary / backup configuration is one link is more expensive than the other
  - Cost
  - Latency
  - Performance
  - Bandwidth
- Usually configured with local Preference for different upstream ASes, or MED with a single upstream AS

# Load Sharing

- Requires loading traffic on all links
  - The goal is to load the links as evenly as possible without negative impact on traffic flows
- Common reason for load sharing is to squeeze as much bandwidth out of multiple links as possible
  - This is often the case where larger links are cost prohibitive such as for small companies or locations where circuit cost is high
- Typically accomplished by slicing up prefixes and advertising different prefixes on different links

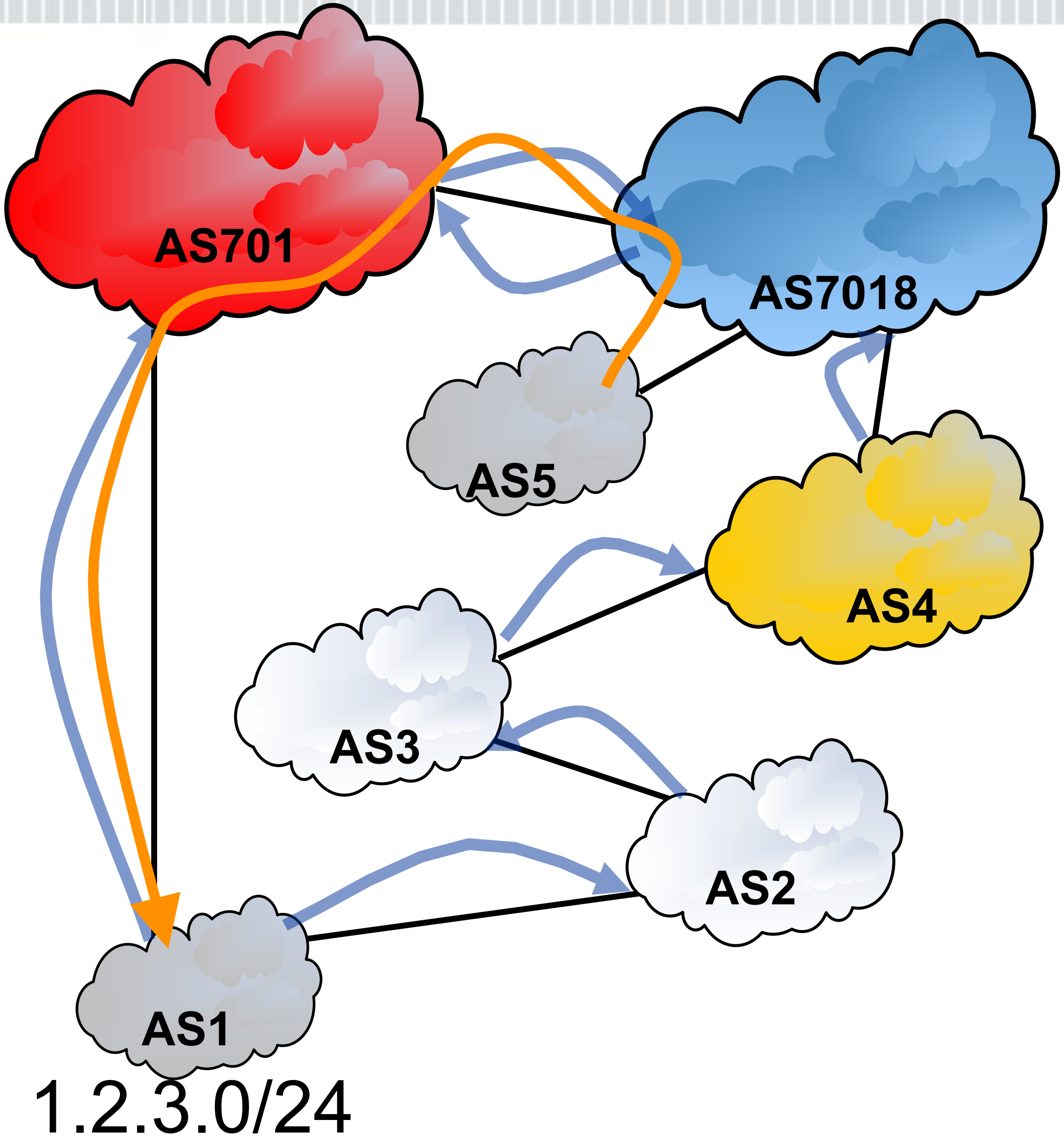
# ISP Best Path TE

# Default Outbound Tier 1 Provider Paths (Best Path)



- AS1 gets transit from AS701 and AS2
- AS2 gets transit from AS3
- AS3 gets transit from AS4
- AS4 gets transit from AS7018
- AS701 and AS7018 are Peers

# Default Outbound Tier 1 Provider Traffic (Best Path)



- AS7018 prefers the path to 1.2.3.0/24 via its Peer AS701
  - –Shortest AS path
    - 2 AS hops vs. 4 AS hops
- AS5 will forward traffic to AS1 via AS7018 and AS701
- Is it better for AS7018 to prefer delivering traffic to its Peer or its customer?

# Tier 1 Provider
# Outbound Traffic Engineering

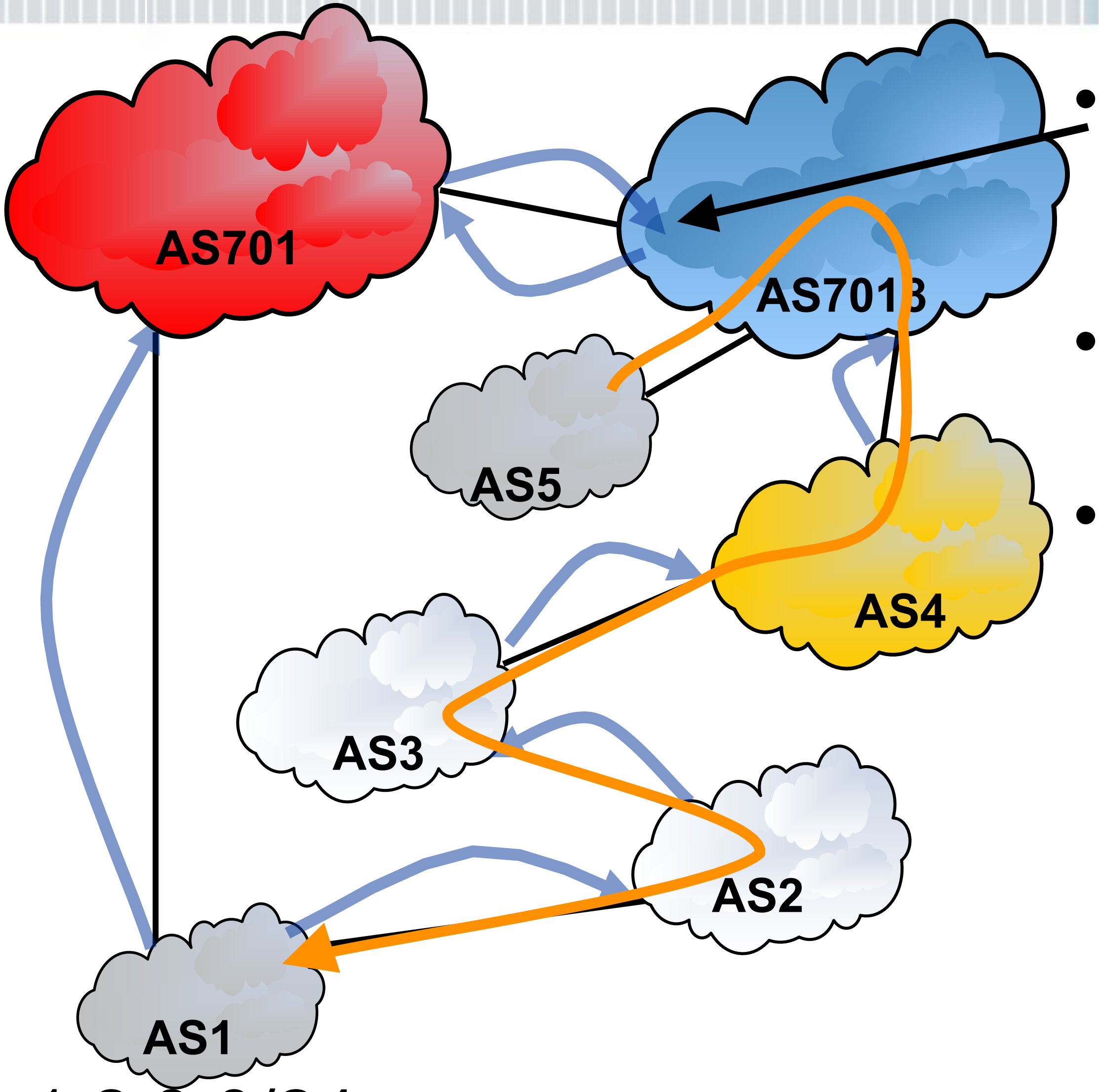
Is it better for AS7018 to prefer delivering traffic to its Peer or its customer?

- Via its Peer is a shorter AS path an might provide better service
- Via its Peer is a short AS path, and AS7018 is more likely to be chosen by multi-homed customers
- Peer links are generally larger than customer links, and Peering contracts usually require bandwidth upgrades when they reach capacity
- Customers pay the transit provider for traffic, Peers do not
  - Preferring customers generates more revenue
- Peering points often become congested and are not easily upgraded, thus may provide poorer performance
- Most Tier 1 ISPs prefer customers over Peers
  - This is a primary / backup configuration

# ISP Local Preference TE

Primary / Backup

# Local Preference
# Outbound Tier 1 Provider Traffic



1.2.3.0/24

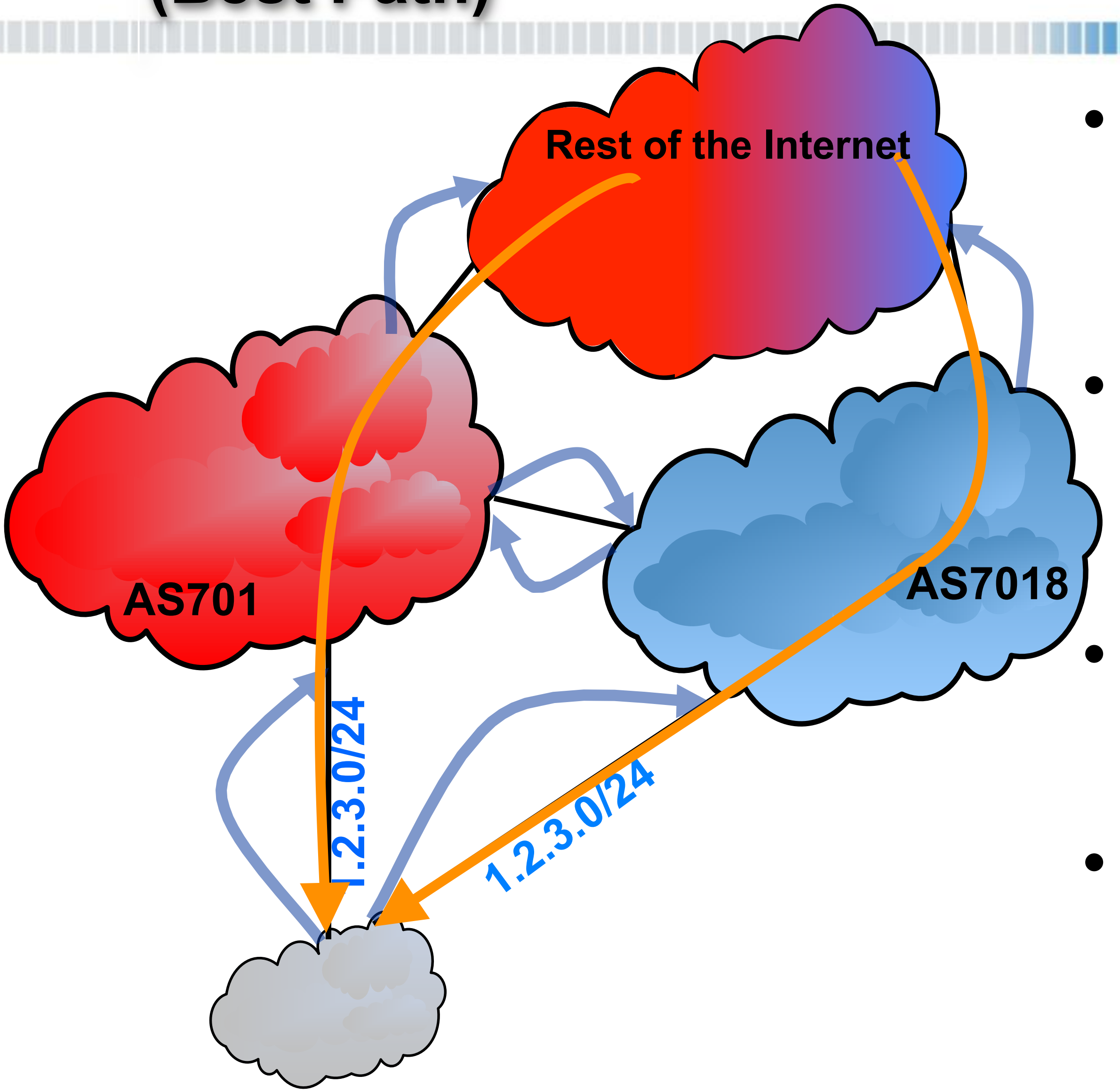
- AS7018 sets policy to lower the local-pref of Peers to 80
- Customers have a default local-pref of 100
- AS7018 prefers customer routes over Peer routes
  - AS5 forwards traffic to AS1 via 7018, 4, 3, 2
  - Customers offered communities to modify their local-pref
    - Can set local-pref equal to Peer value

# Local Preference
# Outbound Tier 2 Provider Traffic

- Tier 2 ISPs have a similar policy to tier 1s but with one added level
  - –Customer are more preferred over Peers
  - –Peers are more preferred over Transit Providers
    - Tier 2 ISPs have to pay for transit to the “rest of the Internet” which is not a customer or a customer of a Peer

# End-site Best Path TE

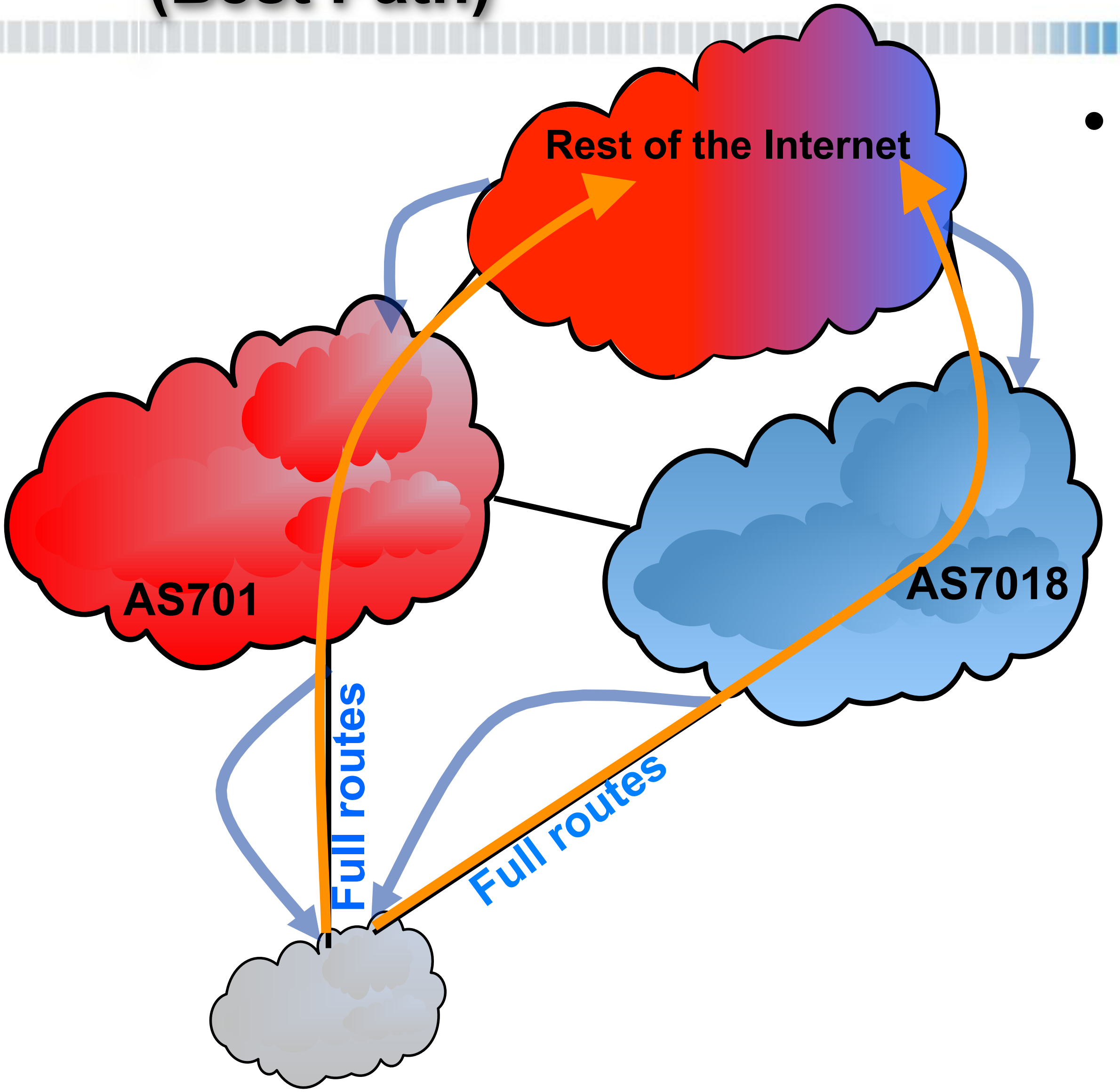
# Default Inbound Customer Traffic (Best Path)



- Customer advertises its route to both upstream providers, at&t and UUNET
- Each provider announces the route to each other and the rest of the Internet
- The portion of the Internet closer to at&t will prefer at&t
- The portion of the Internet closer to UUNET will prefer UUNET

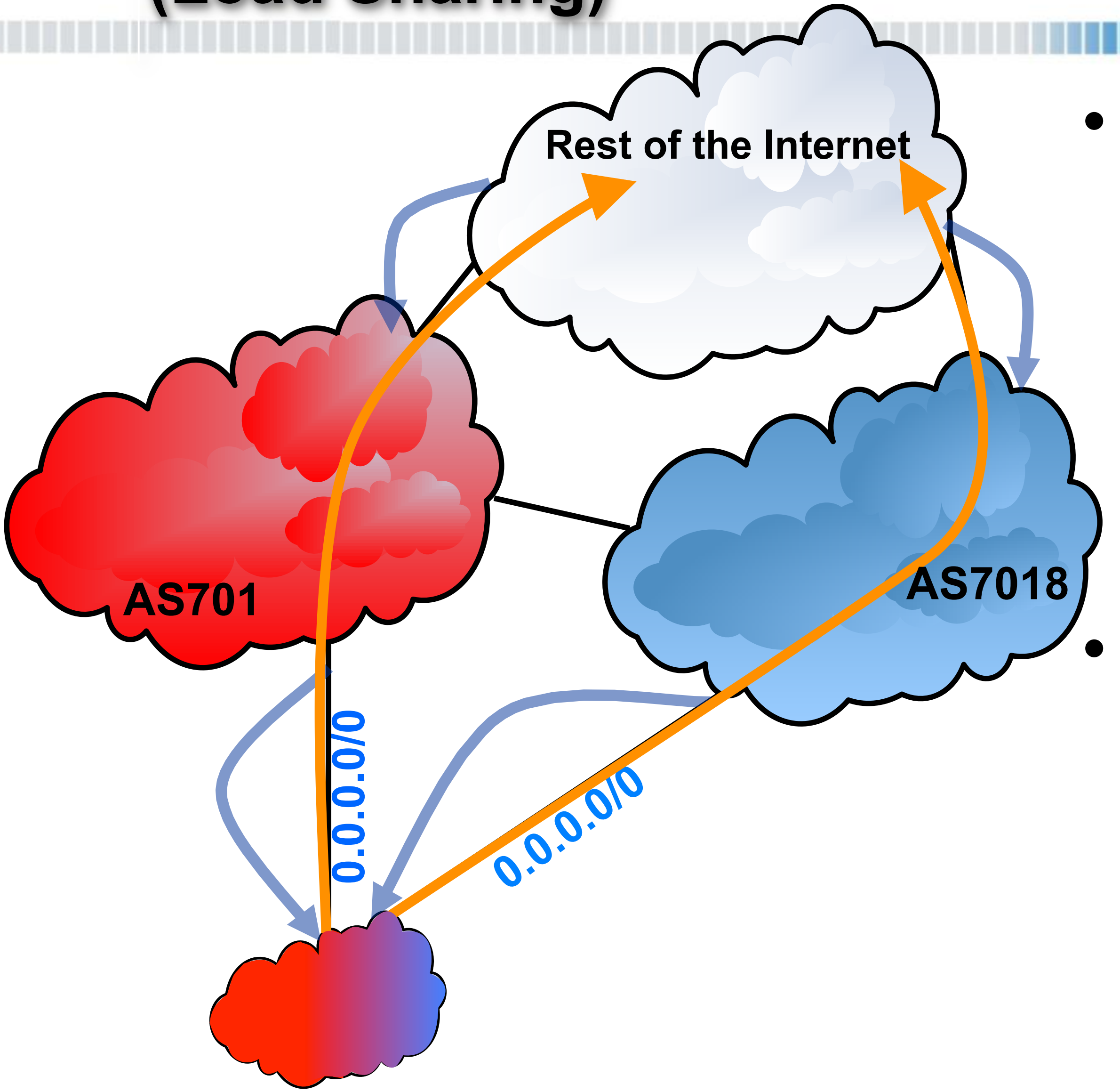
1.2.3.0/24

# Default Outbound Customer Traffic (Best Path)



- The customer can learn full routes from both ISPs
  - Will use the upstream ISP that the destination is closer too
  - Will round robin equally good paths

# Default Outbound Customer Traffic (Load Sharing)

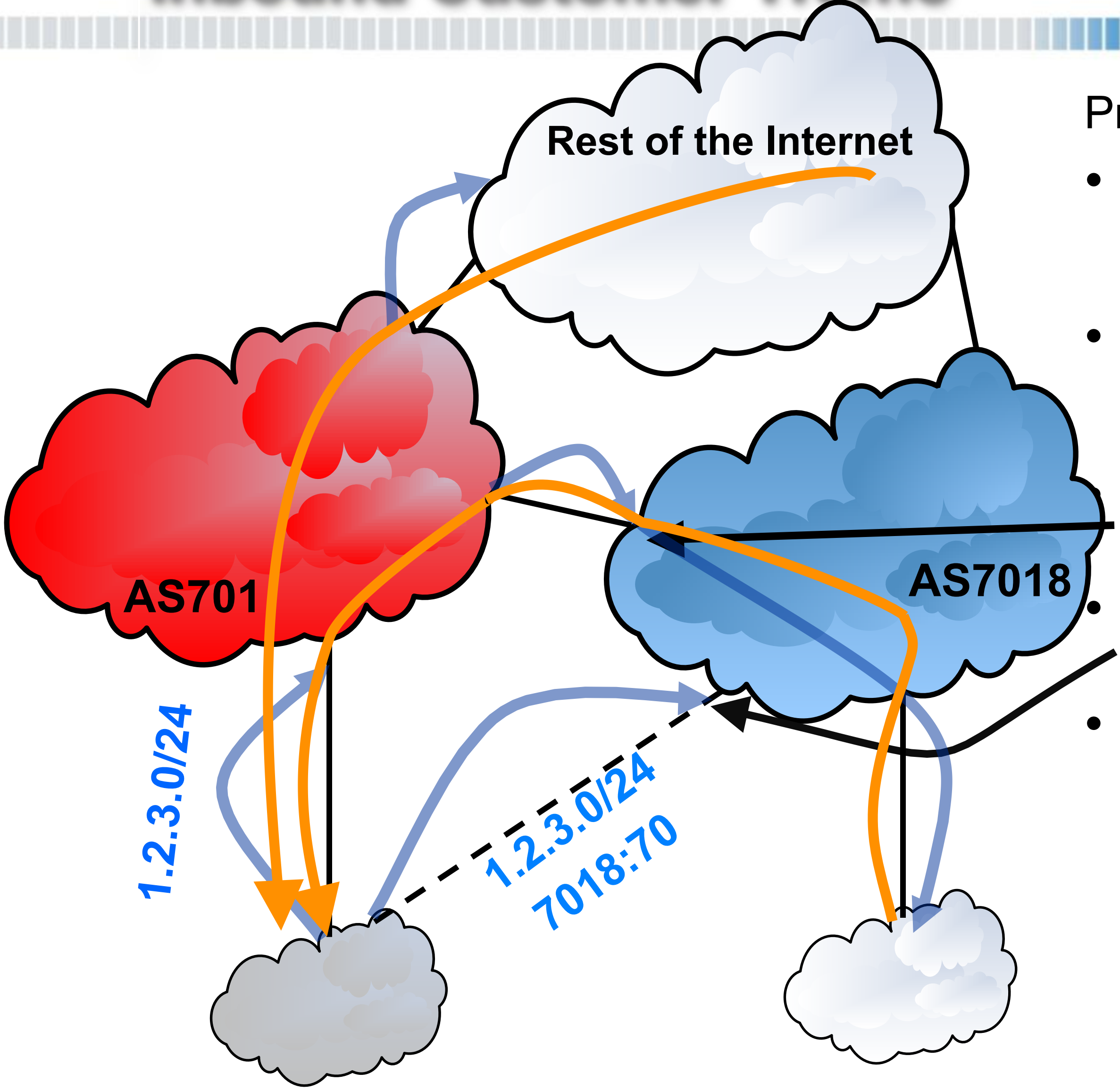


- The customer can learn full routes from both ISPs
  - Will use the upstream ISP that the destination is closer too
  - Will round robin equally good paths
- Each provider can announce a default route
  - Will round robin outbound traffic
  - Or prefer the exit that is closer depending on local IGP metrics

# End-site Local Preference TE

## Primary / Backup

# Local Preference
# Inbound Customer Traffic

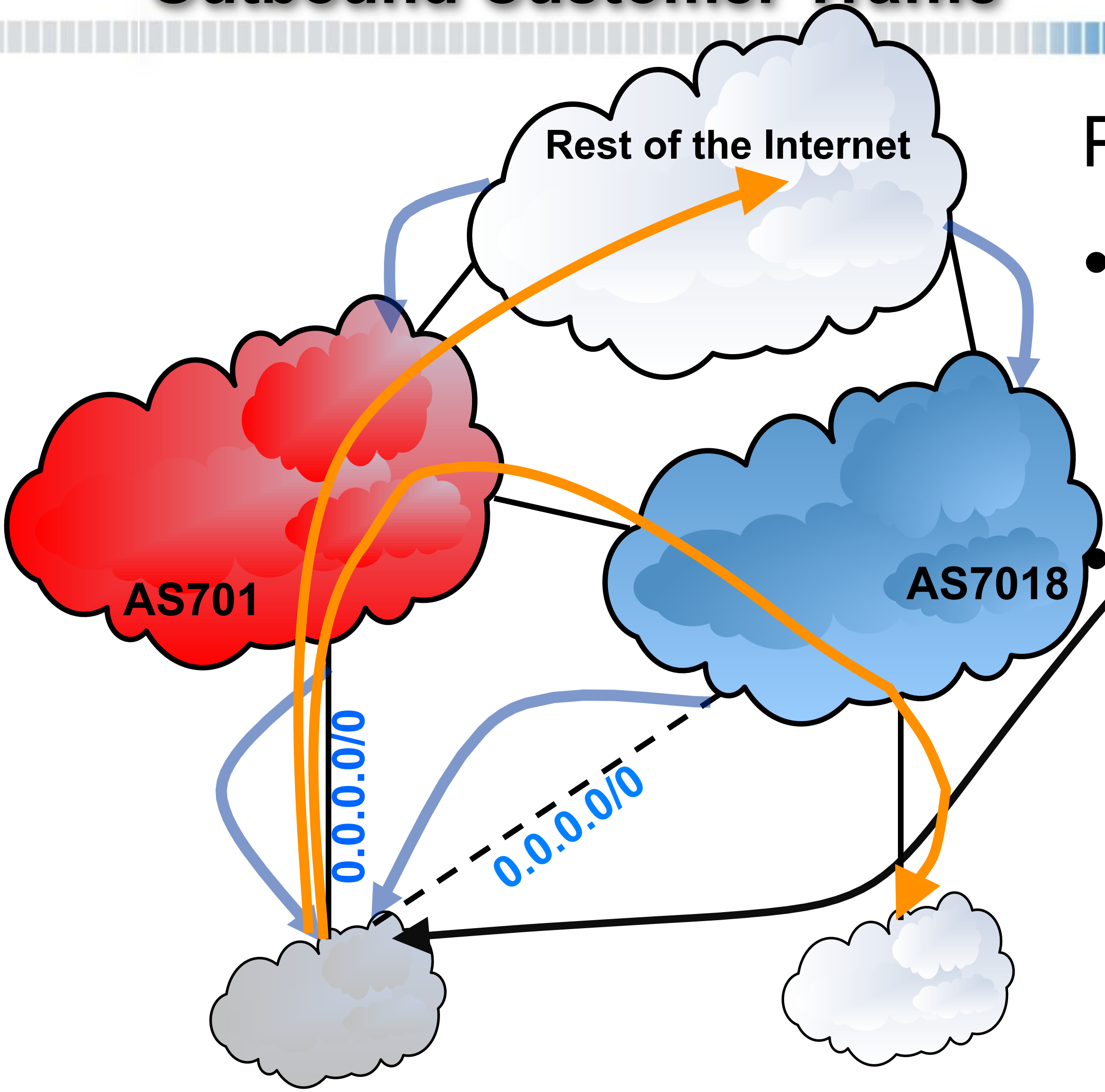


Primary / backup ISPs

- Customer wants to only use the at&t link if UUNET goes down
- Customer sets community 7018:70 on announcements to at&t
- at&t sets local-pref of 80 on all Peer routes
- at&t matches on 7018:70 and sets local-pref of 70 for cust
- at&t prefers the route learned from its Peers
  - at&t announces this route to customers
  - at&t does not announce this route to Peers unless UUNET withdraws or times out

1.2.3.0/24

# Local Preference
# Outbound Customer Traffic

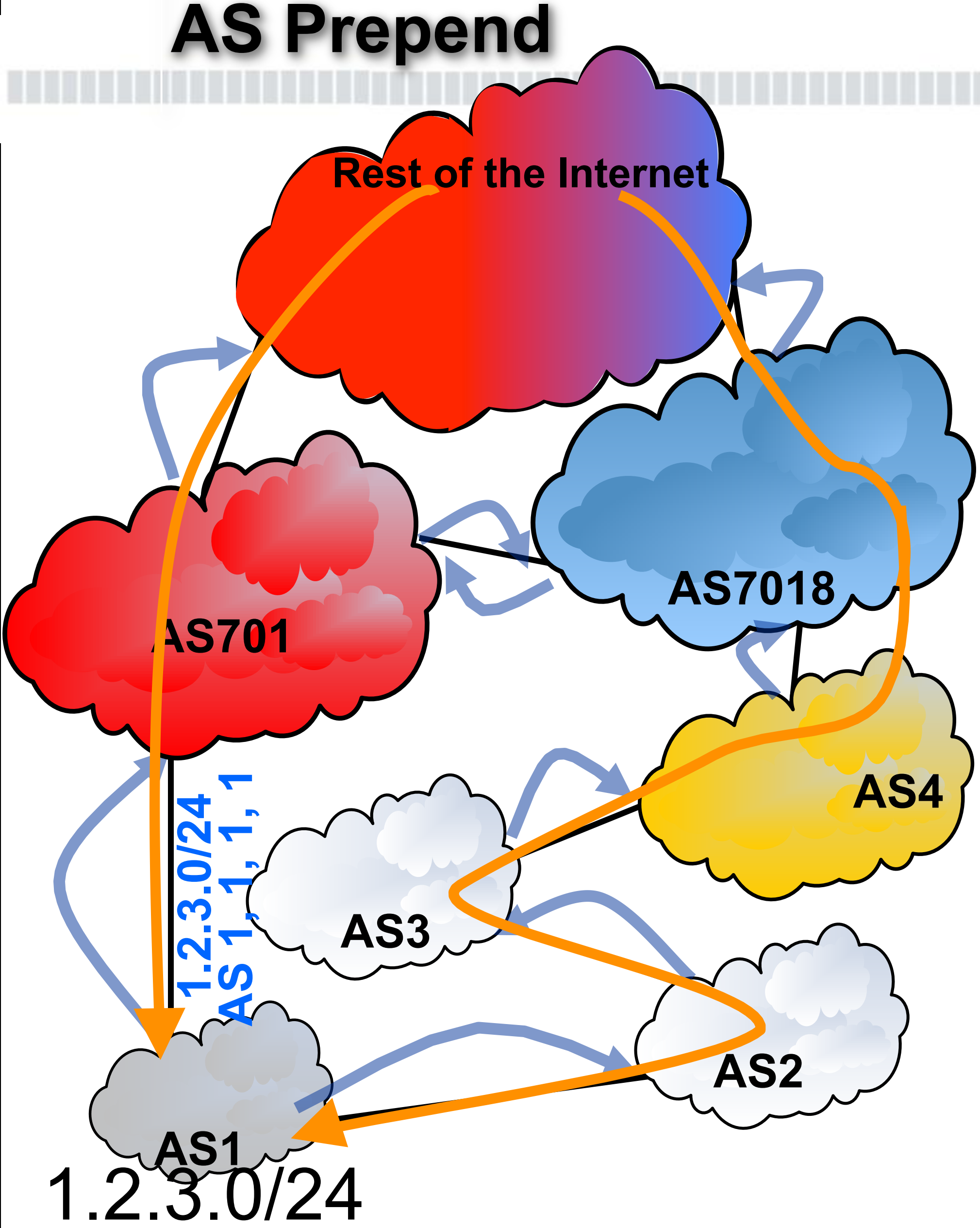


Primary / backup

- Each provider announces a default route
- Customer sets local-pref of 90 on routes from at&t
  - –Will prefer UUNET
  - –Will only use at&t if UUNET is down

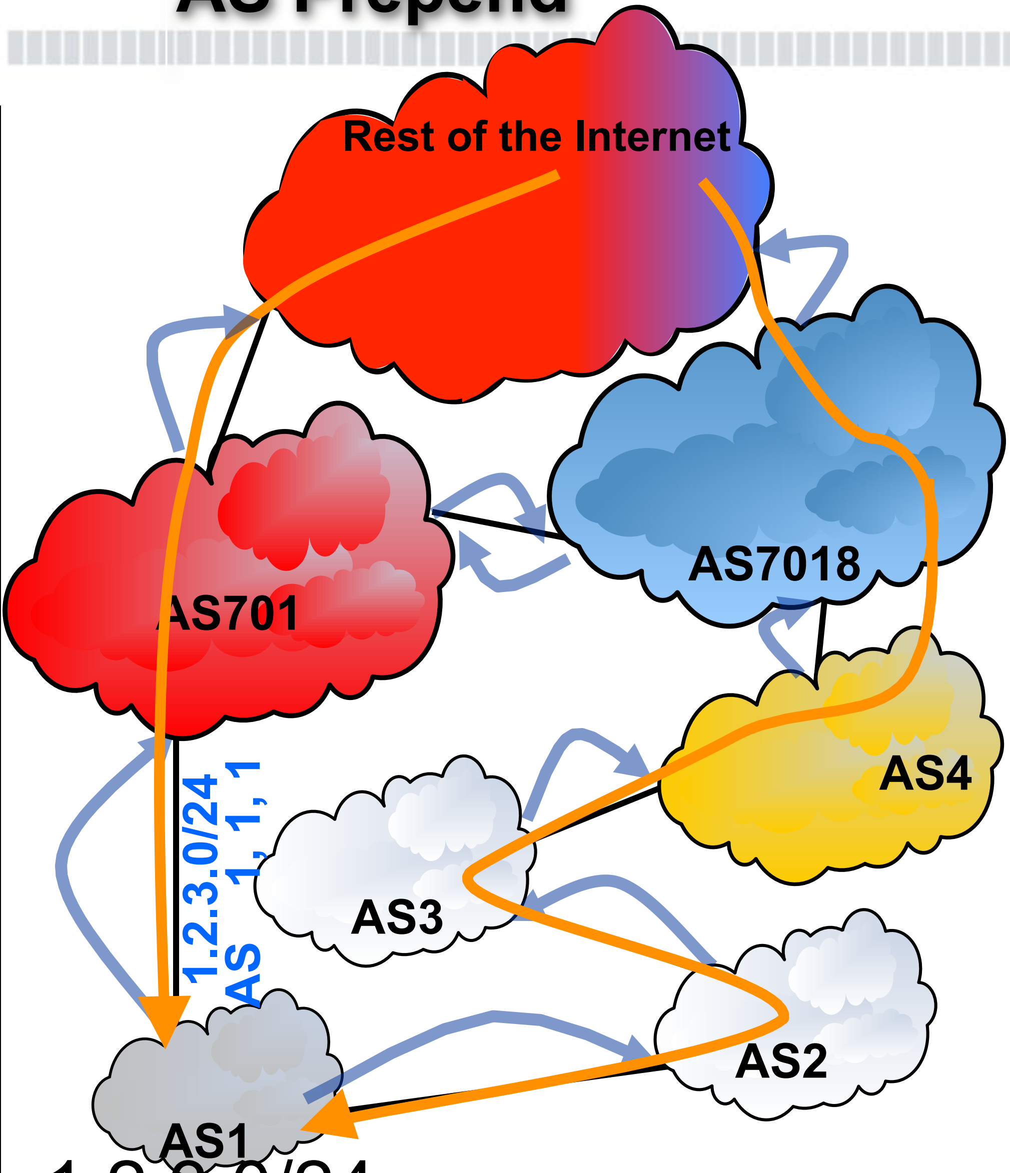
# AS Path TE

# AS Prepend



- AS1 prepends its AS three times to AS701
- AS1 advertises to AS2
- AS2 advertises to AS3
- AS3 advertises to AS4
- AS4 advertises to AS7018
- AS701 & AS7018 announce to each other and the rest of the Internet
- Portion of the Internet closer to at&t will use at&t
- Portion of the Internet closer to UUNET will use UUNET

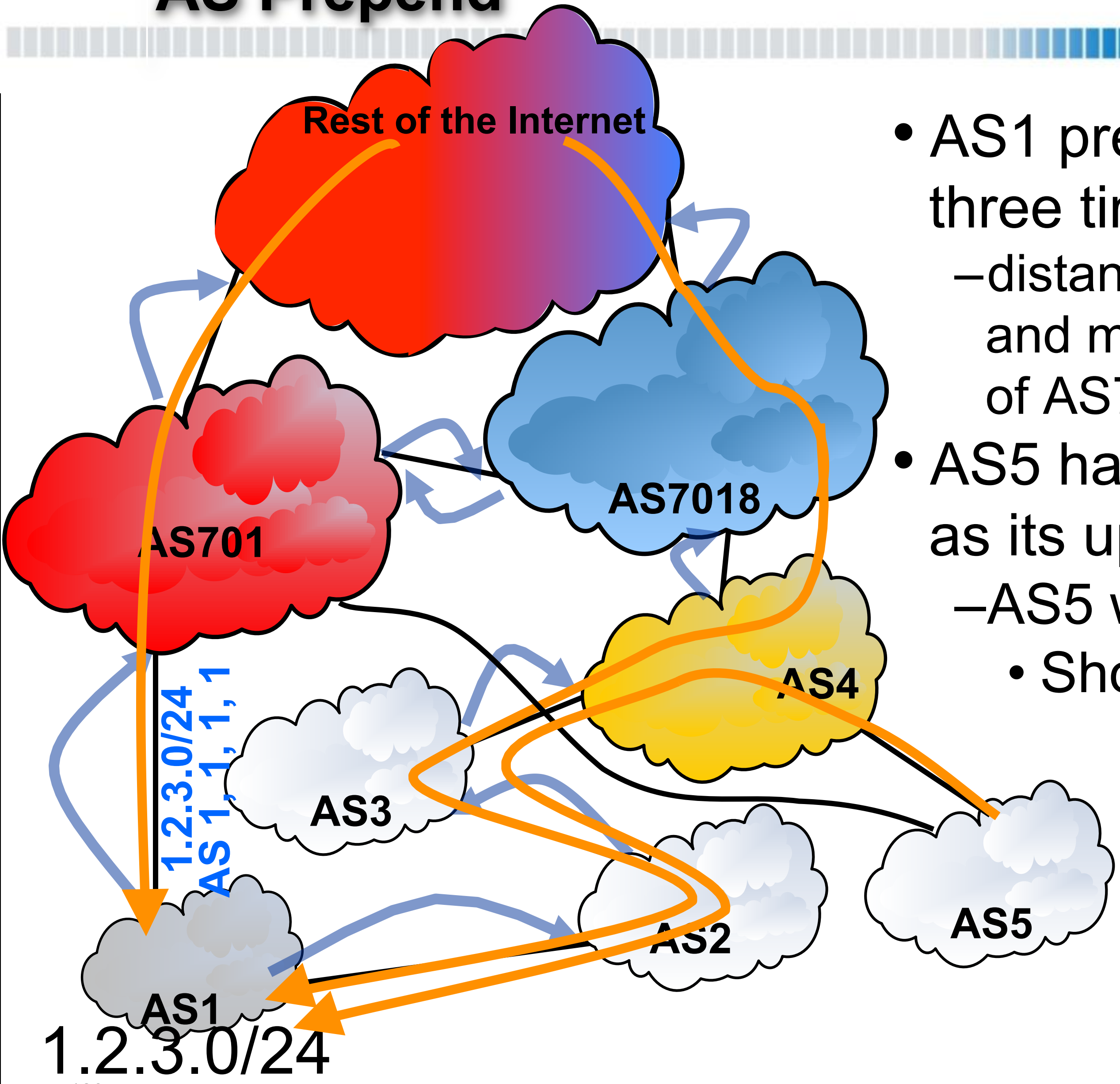
# AS Prepend



- AS1 prepends its AS only two times to AS701
- Same behavior as before, but now more of the rest of the Internet is closer to AS701

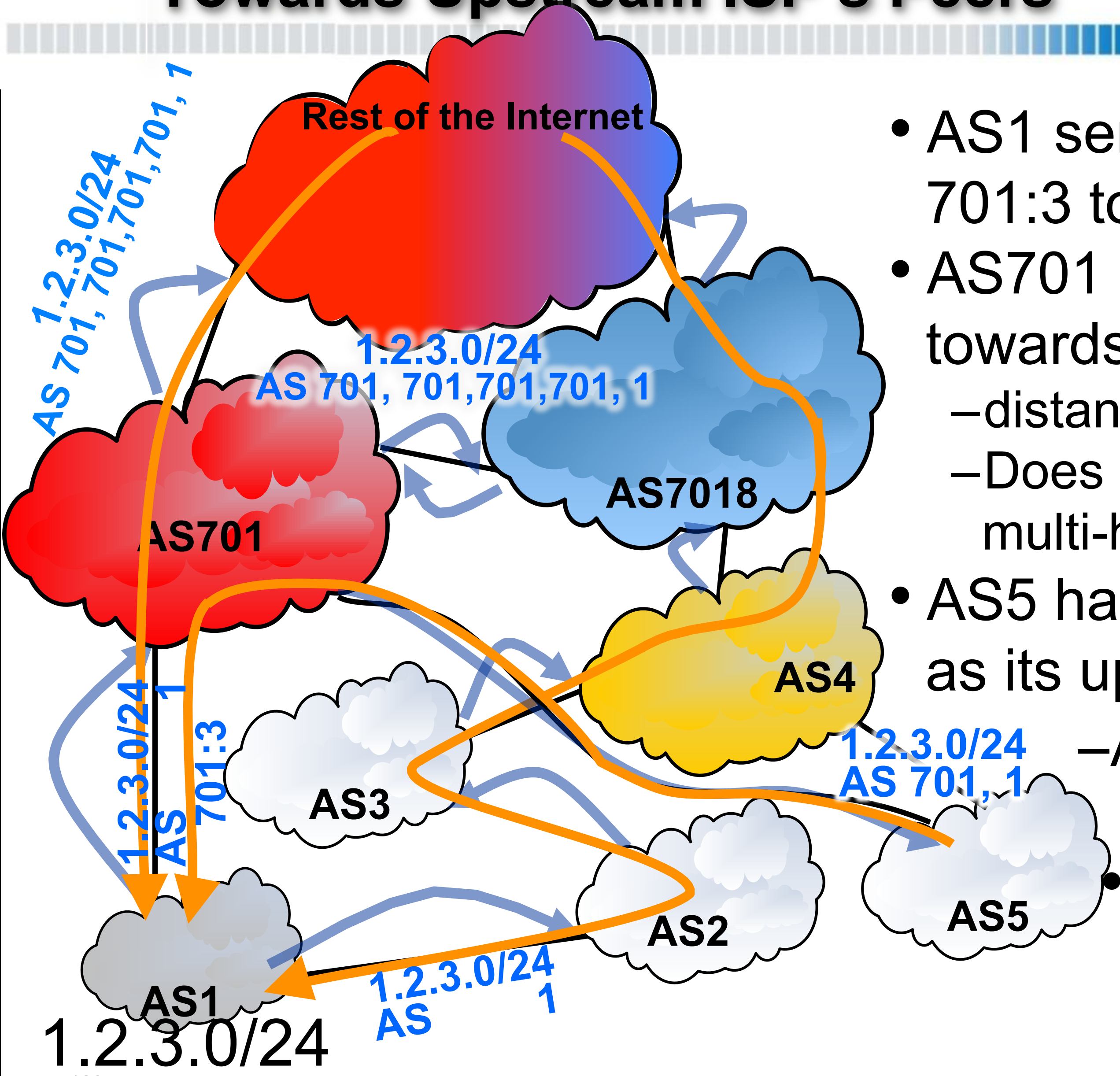
1.2.3.0/24

# AS Prepend



- AS1 prepending its AS three times to AS701
  - –distances AS701 Peers, and multi-homed customers of AS701
- AS5 has AS701 and AS4 as its upstream providers
  - –AS5 will prefer AS4
    - Shorter AS path

# AS Prepend
# Towards Upstream ISP's Peers

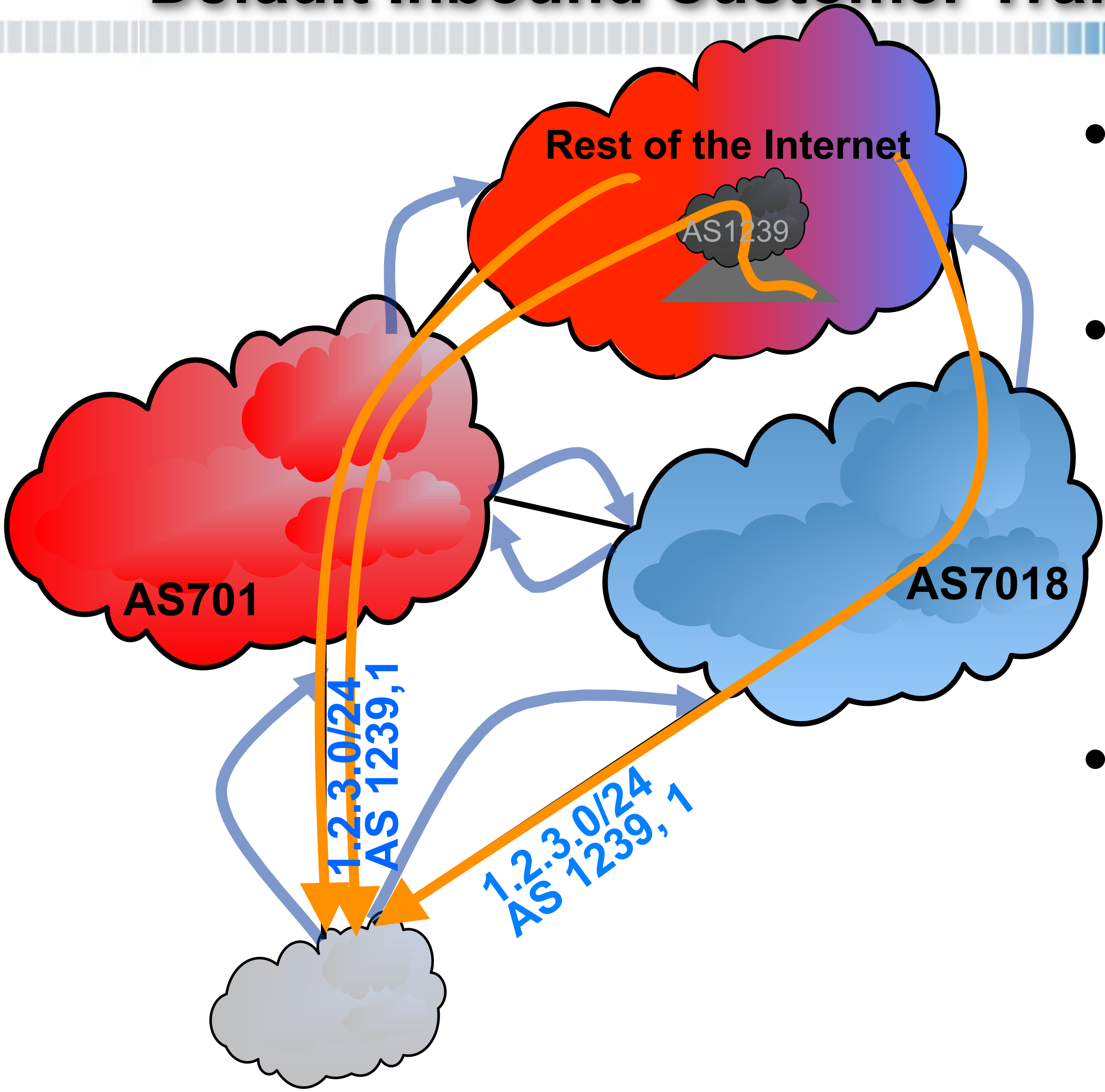


- AS1 sends community 701:3 to AS701
- AS701 prepend 3 times towards its Peers
  - –distances AS701 Peers
  - –Does not distance AS701 multi-homed customers
- AS5 has AS701 and AS4 as its upstream providers
  - –AS5 will prefer AS701
    - Shorter AS path
- Rest of the Internet is unchanged

# Prepending Someone Else’s AS

- Customer may advertise a more specific route to one provider, to draw extra traffic to links with that provider
- Customer may limit propagation of that route by pre-pending another network’s ASN
  - –This will prevent that network from learning that prefix
  - –This will prevent networks down stream from that AS from learning that prefix

# Default Inbound Customer Traffic

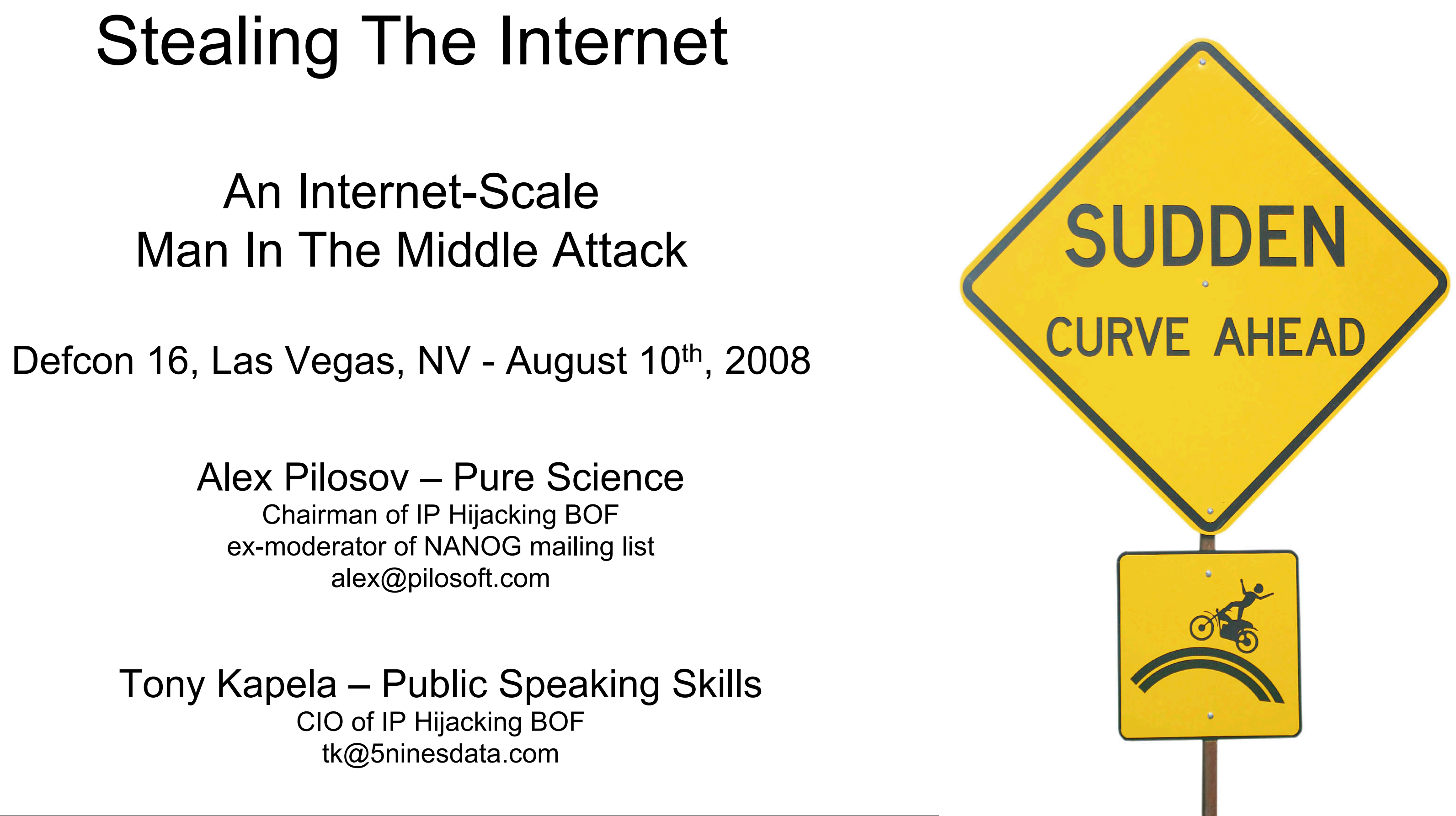


1.2.3.0/24

- AS 1239 will not learn 1.2.3.0/24
- The portion of the Internet that is only down stream from AS 1239 will not learn 1.2.3.0/24
- This may be useful if 1.2.3.0/24 is part of an aggregate
  - –E.g. AS701 announces 1.2.0.0/16

# Prepending Someone Else’s AS

- Can insert another’s ASN in the path to create a Kapela style man in the middle attack



Stealing The Internet

An Internet-Scale
Man In The Middle Attack

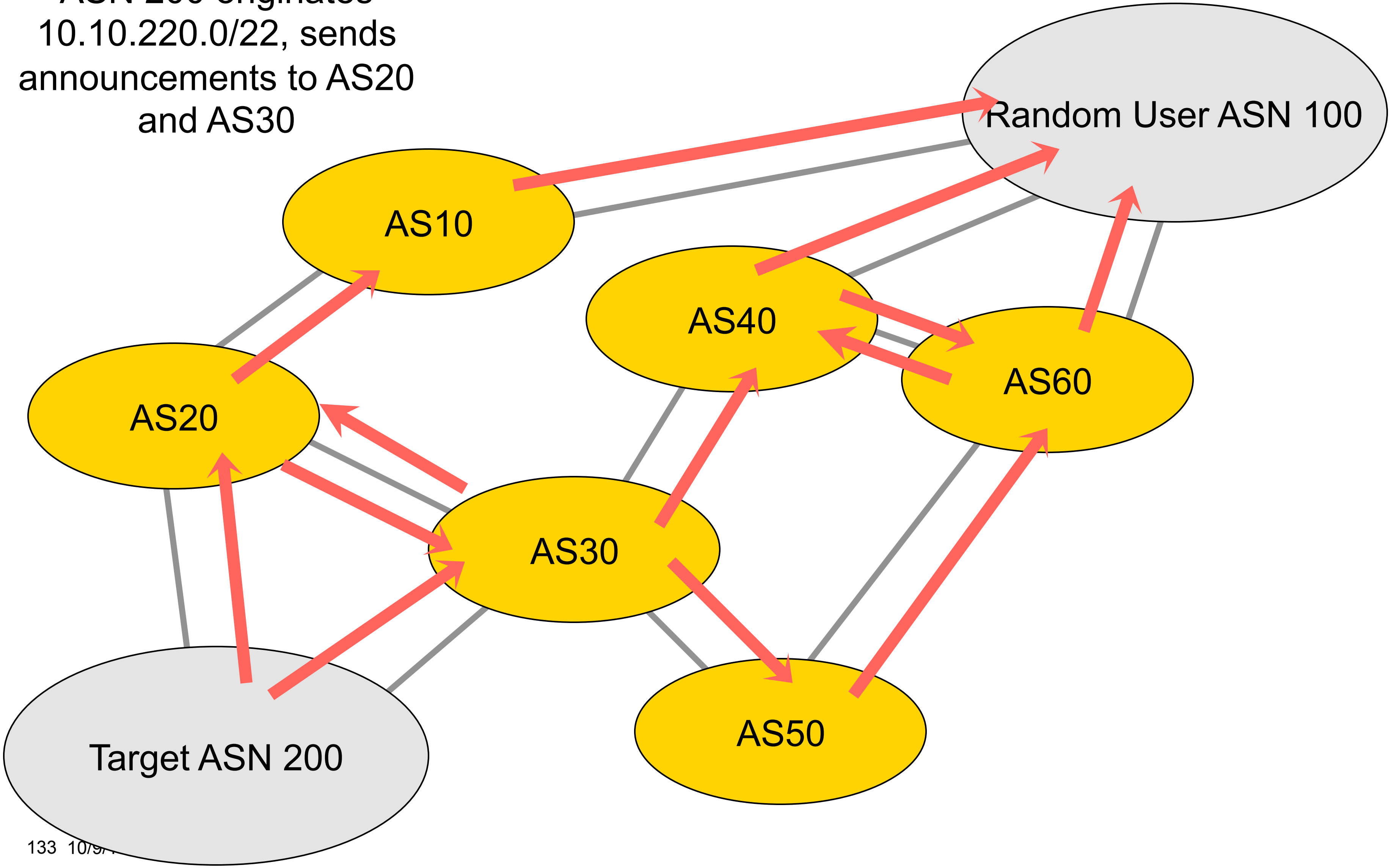
Defcon 16, Las Vegas, NV - August 10th, 2008

Alex Pilosov – Pure Science
Chairman of IP Hijacking BOF
ex-moderator of NANOG mailing list
alex@pilosoft.com

Tony Kapela – Public Speaking Skills
CIO of IP Hijacking BOF
tk@5ninesdata.com

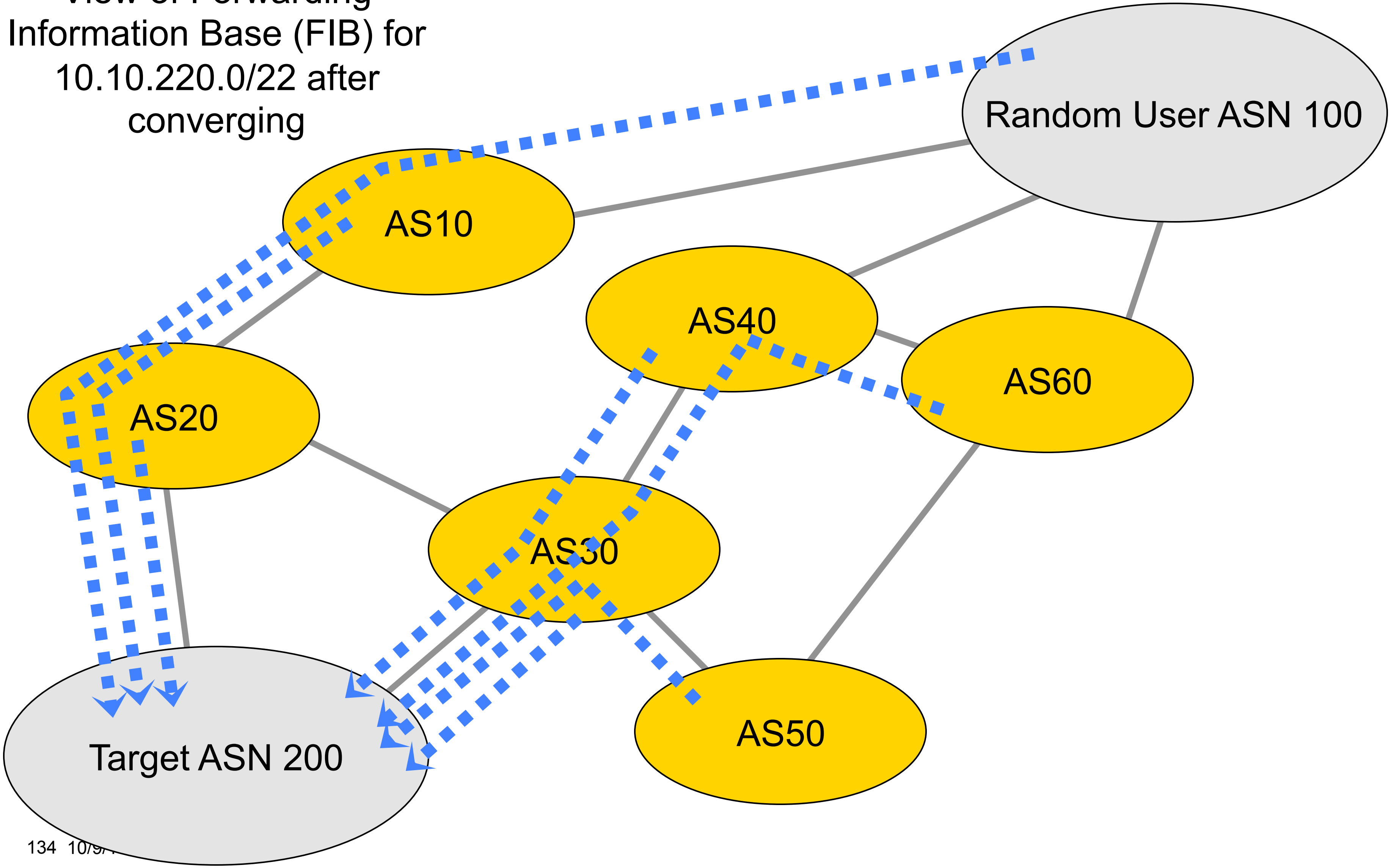
# BGP MITM – First Observe

ASN 200 originates 10.10.220.0/22, sends announcements to AS20 and AS30



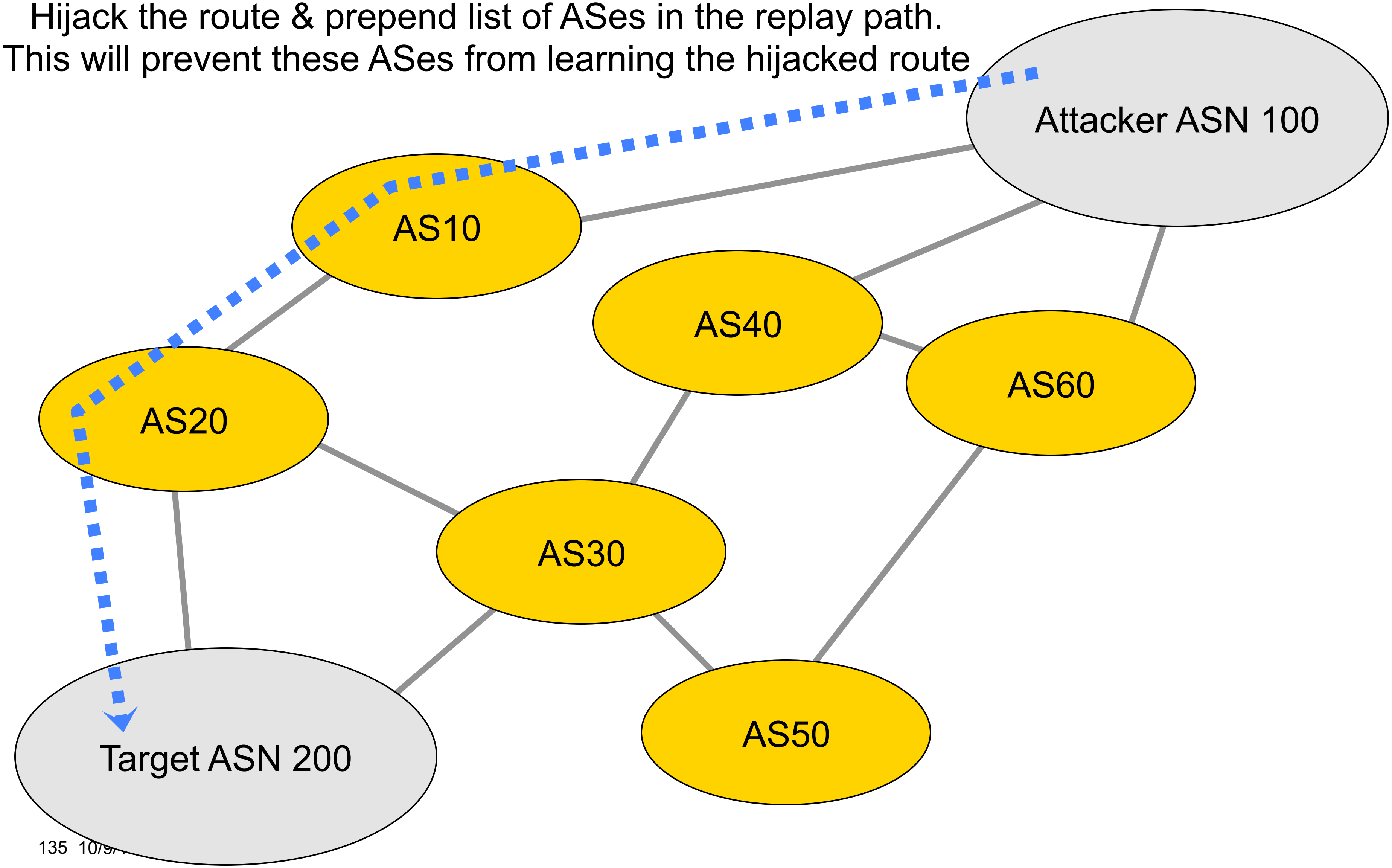
# BGP MITM – First Observe

View of Forwarding Information Base (FIB) for 10.10.220.0/22 after converging



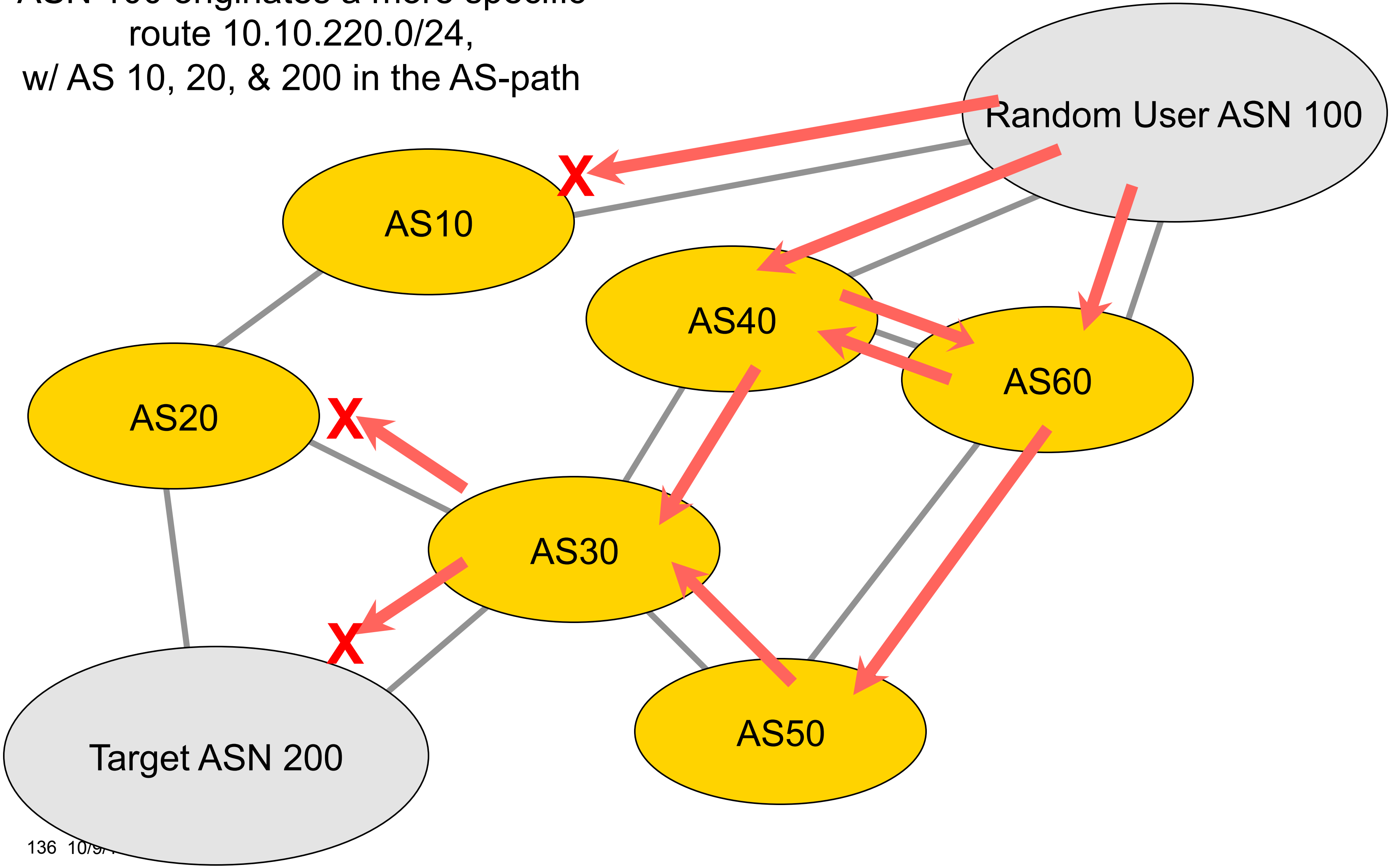
# BGP MITM – Plan Reply Path

Hijack the route & prepend list of ASes in the replay path.
This will prevent these ASes from learning the hijacked route



# BGP MITM – Hijack Routes

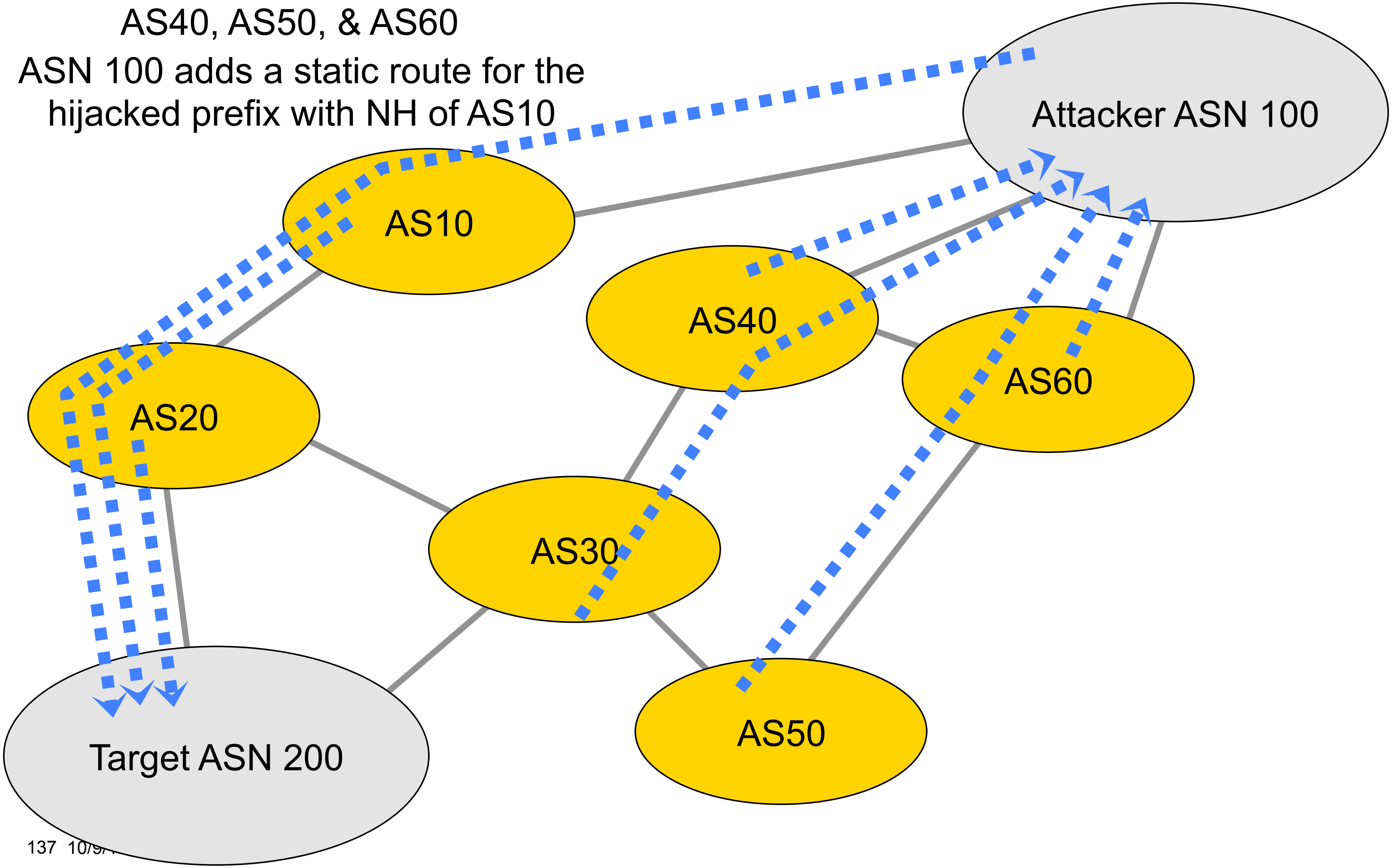
ASN 100 originates a more specific route 10.10.220.0/24, w/ AS 10, 20, & 200 in the AS-path



# BGP MITM – Return Hijacked Traffic

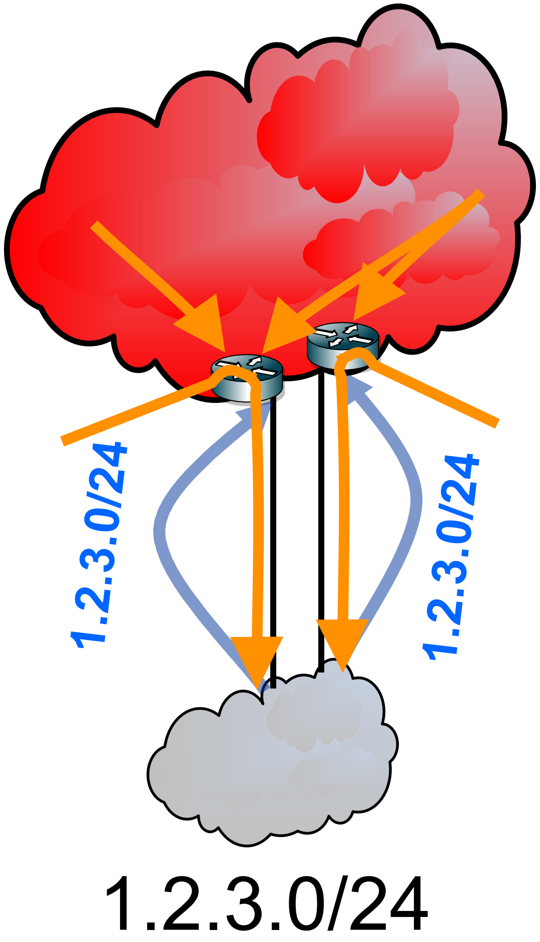
ASN 100 hijacks traffic from AS30, AS40, AS50, & AS60

ASN 100 adds a static route for the hijacked prefix with NH of AS10



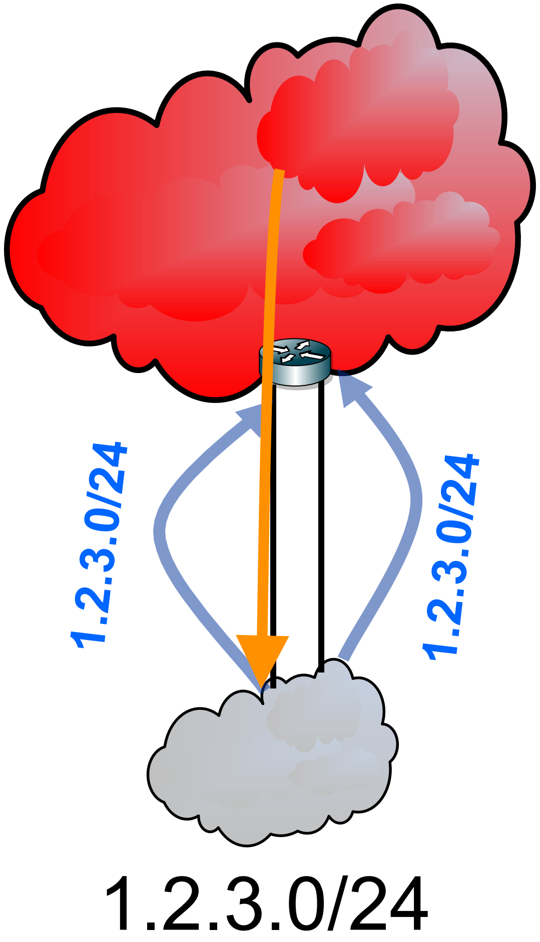
# End-site MED TE

# Default Inbound Customer Traffic



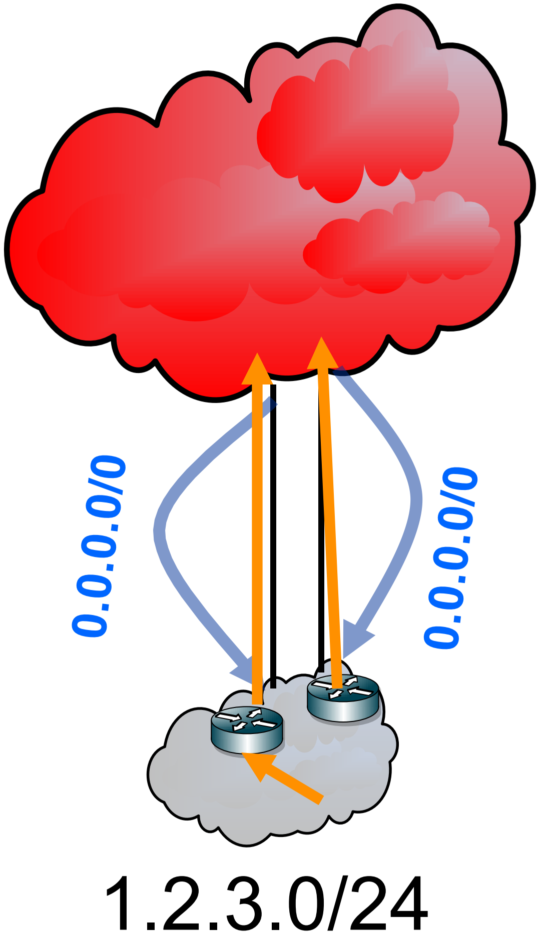
- Customer advertises the same prefixes across two links to a provider
- If terminated on different ISP routers
  - Each ISP router (and their down stream customers) will prefer the local eBGP session
  - The rest of the network will choose the IGP closest eBGP router
  - If IGP distance is equal, the rest of the network will choose the lowest router ID

# Default Inbound Customer Traffic



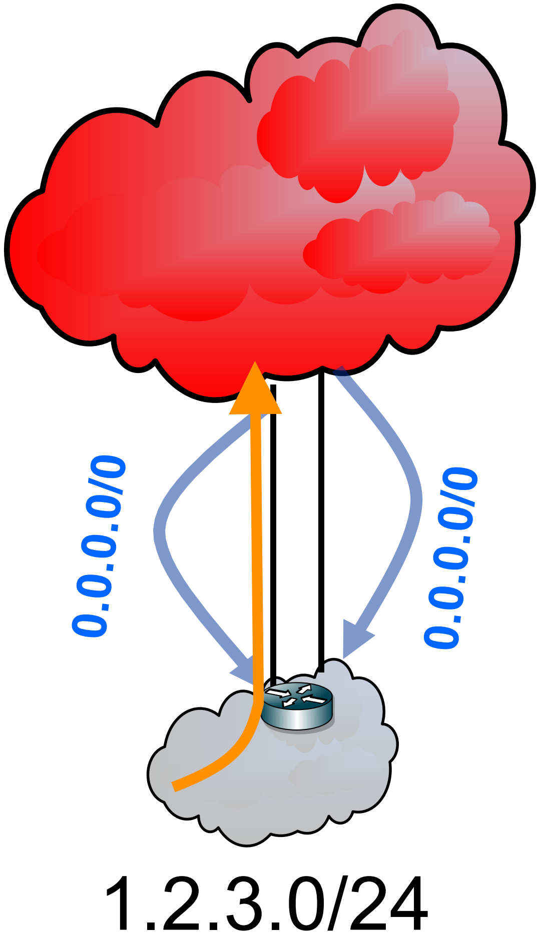
- Customer advertises the same prefixes across two links to a provider
- If terminated on a single ISP router
  - –The network will choose one eBGP path
    - Oldest path
    - Lowest Router ID
    - Lowest neighbor if terminated on a single customer router

# Default Outbound Customer Traffic



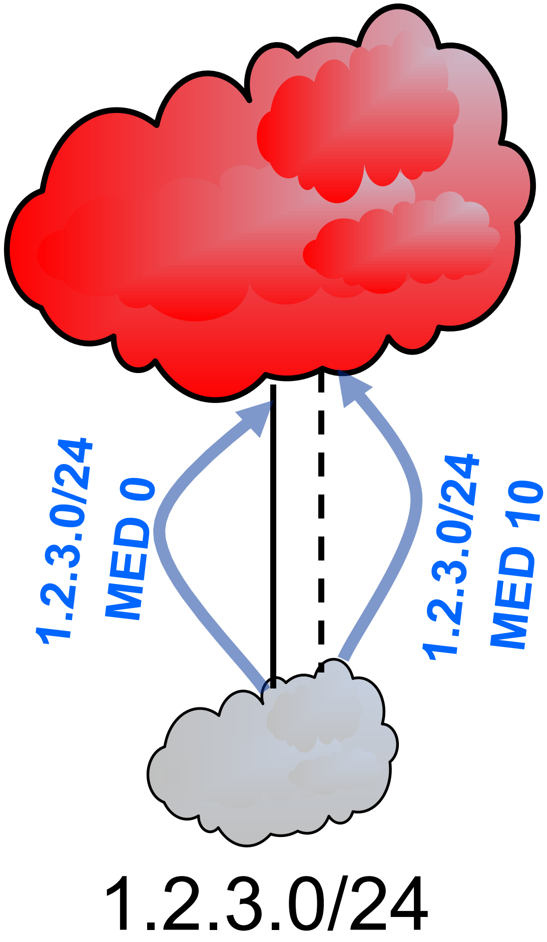
- ISP advertises the same prefixes across two links to a customer
- If terminated on different customer routers
  - –Each customer router will prefer the local eBGP session
  - –The rest of the network will choose the IGP closest eBGP router
  - –If IGP distance is equal, the rest of the network will choose the lowest router ID

# Default Outbound Customer Traffic



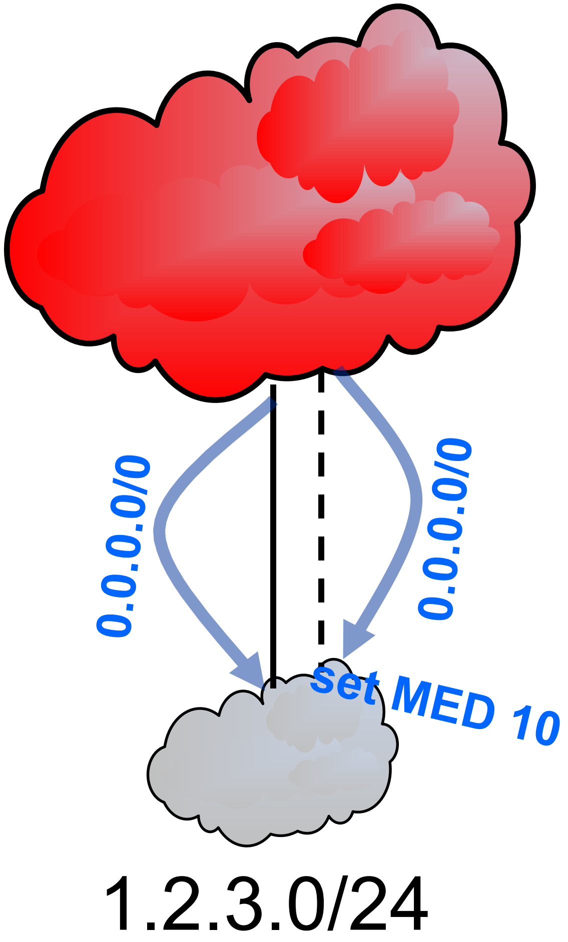
- ISP advertises the same prefixes across two links to a customer
- If terminated on a single customer router
  - –The network will choose one eBGP path
    - Oldest path
    - Lowest Router ID
    - Lowest neighbor if terminated on a single ISP router

# Lowest Mutli-Exit Discriminator (MED)



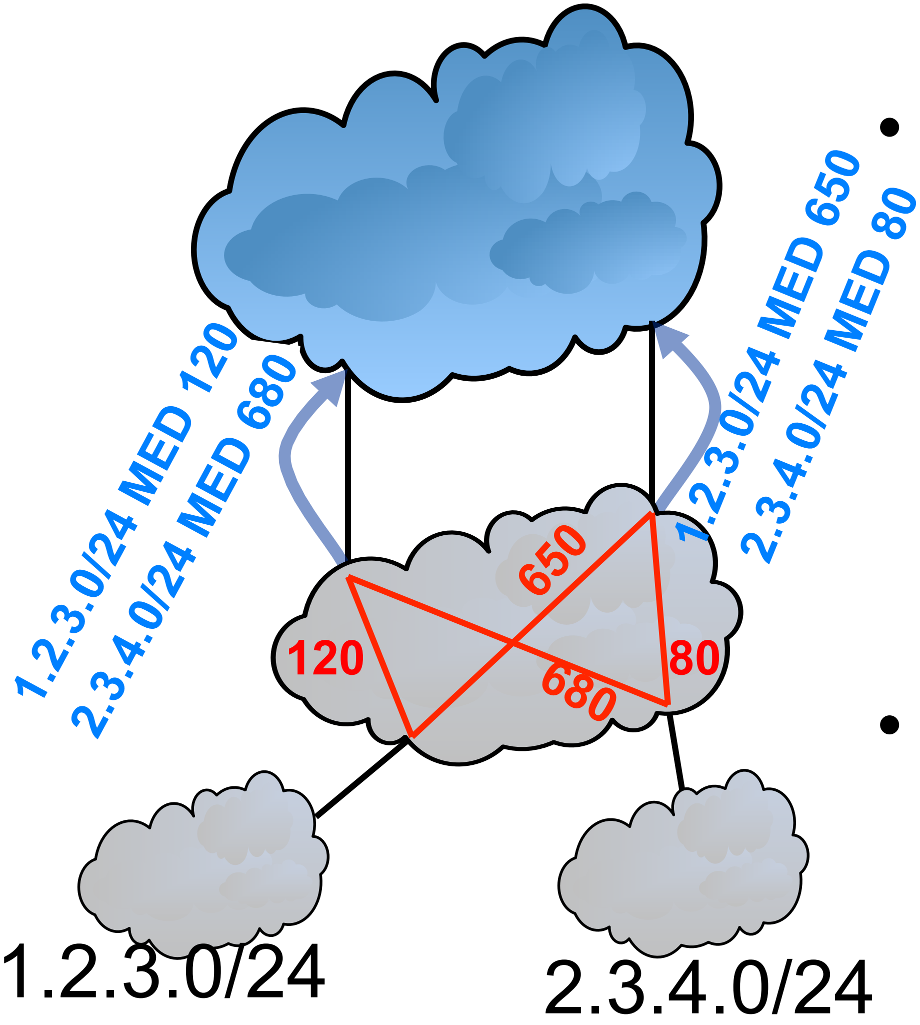
- Customer may prefer a primary / backup configuration
- Customer can advertise the same prefix with a higher MED on the back-up link

# Lowest Mutli-Exit Discriminator (MED)



- Customer may prefer a primary / backup configuration
- Customer can advertise the same prefix with a higher MED on the back-up link
- Customer can set MED inbound on routes learned from the provider in the same manor

# Lowest Mutli-Exit Discriminator (MED)



- MED can also be used to convey local IGP metrics
- This will cause the upstream network to do cold potato routing
  - Upstream delivers traffic to your network at the entry point closest to the destination
  - ISP will often offer this to their multi-connected customers
- Be careful to avoid *always compare MEDs* if you have more than one upstream
  - Their metrics may be incompatible

# ISP MED TE

# ISP MED TE

- ISPs do not manipulate the MEDs of customer routes
  - –Customers pay their transit providers to honor their traffic engineering preferences
  - –That includes the MED values customers set
- ISPs flatten MEDs from their Peers
  - –Settlement Free Peers do not pay to exchange traffic
  - –They are generally not permitted to specify traffic engineering preferences on their Peer’s network
  - –Some larger ISPs will force smaller ISPs to accept their MEDs in order to qualify for Peering
    - The smaller Peer must perform cold potato routing, and deliver traffic to the larger Peer on the entry point closest to the destination

# ISP IGP Distancing

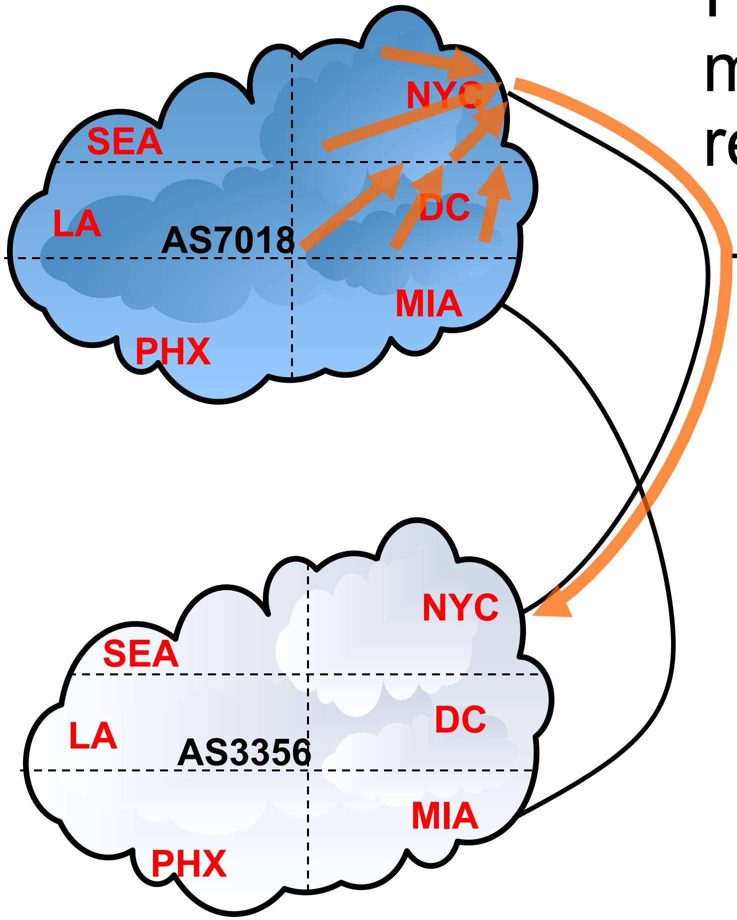
# ISPs IGP Distance

- ISPs typically design their IGP metrics based on network topology and distance (latency / route miles)
- All things being equal, ISPs prefer to deliver traffic to destinations that are local to a given region over destinations in a remote region
- If a destination is not connected to the local region it will pick the closest region where the destination is connected

# ISP – Peer Interconnects

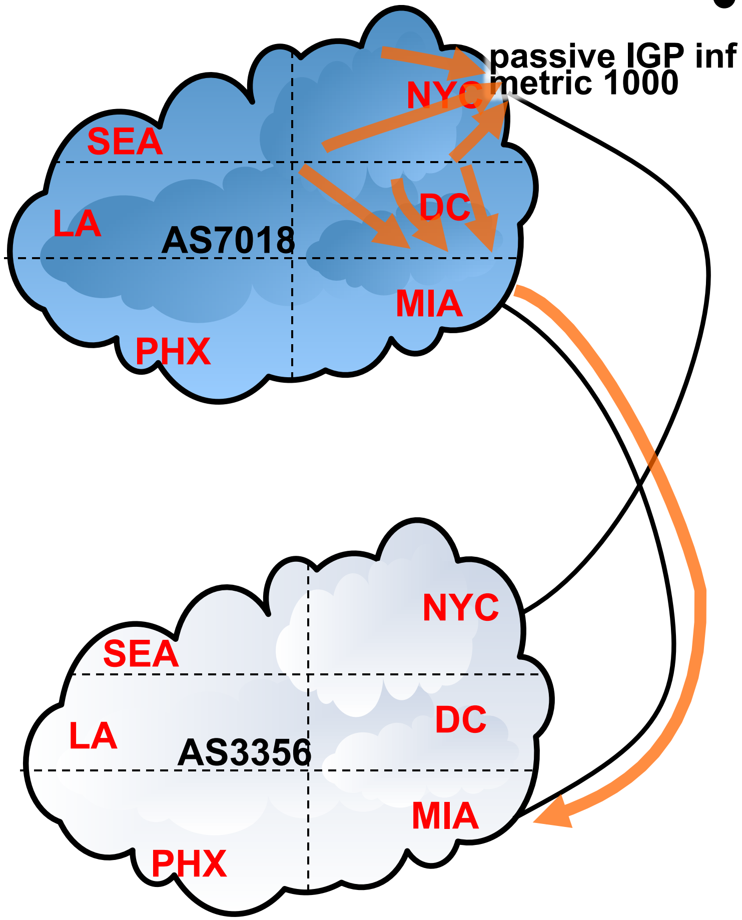
- Peers typically connect in multiple regions
- Peers are typically required to send the same set or routes at all interconnect points, and are typically not permitted to do traffic engineering
- ISPs can configure the interface facing a Peer for passive IGP, and place a small IGP metric to push traffic away from a hot interface
  - Peer biasing

# Peer Biasing (Remote Region)



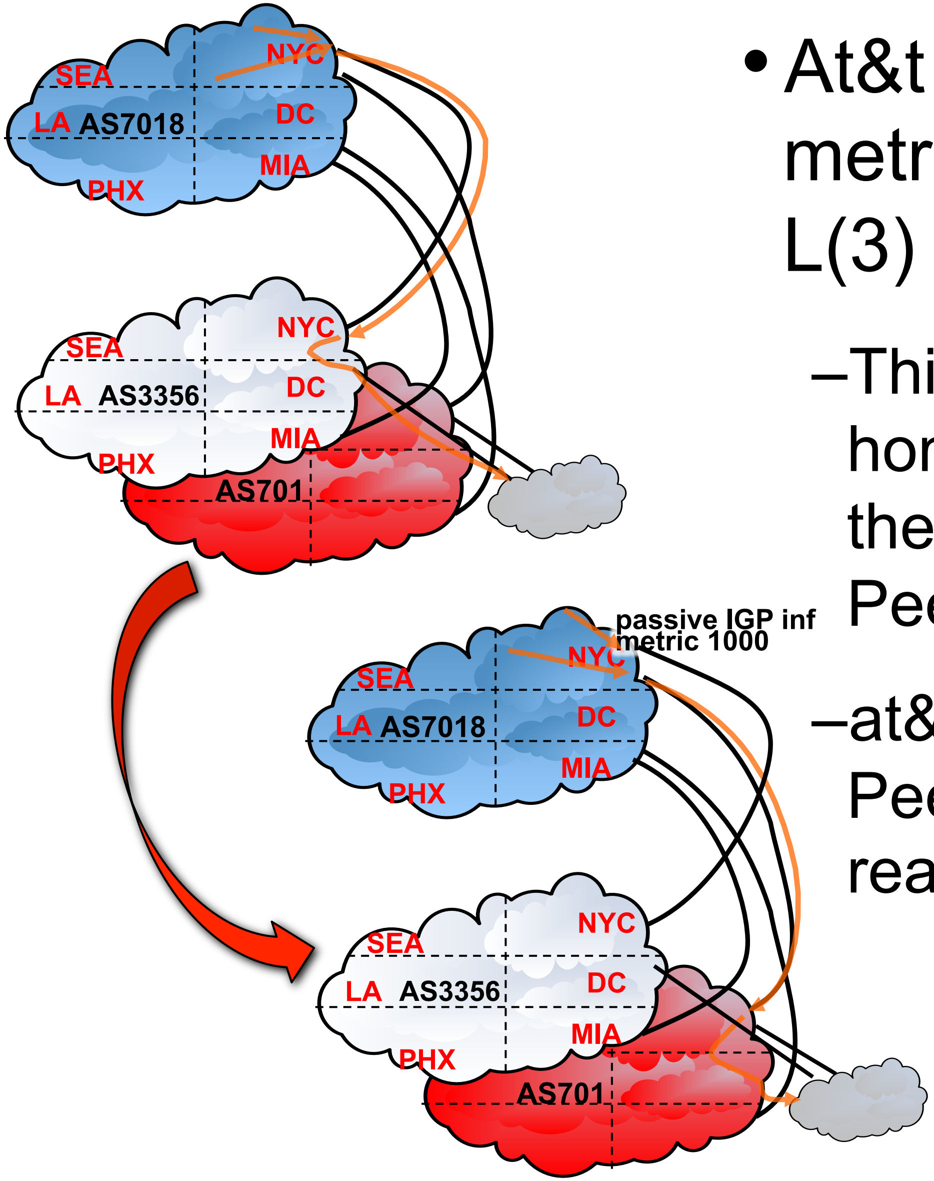
- Peers will often interconnect in many regions, but possibly not all regions
  - Ex. L(3) and at&t Peer on the east coast in New York and Miami, but not DC
    - at&t sources in the New York region will use the New York link to L(3) destinations
    - Likewise at&t sources in the Miami region will use the Miami link to L(3) destinations
    - at&t sources in the DC region will use the New York link to L(3) because the DC region is IGP closer to New York than Miami

## Peer Biasing (Remote Region)



- At&t can add a passive IGP metric on the NYC interface to L(3)
  - –This will make the DC region IGP closer to the Peering point w L(3) in Miami
  - –at&t DC region will still prefer NYC Peering point for other Peers connected to both NYC and MIA

# Peer Biasing



- At&t can add a passive IGP metric on the NYC interface to L(3)
  - –This will make destinations multi-homed to two Peers connected in the same region to prefer the other Peer
  - –at&t DC region will still prefer MIA Peering point for other destinations reachable by Level(3)

# End-site IGP Distancing

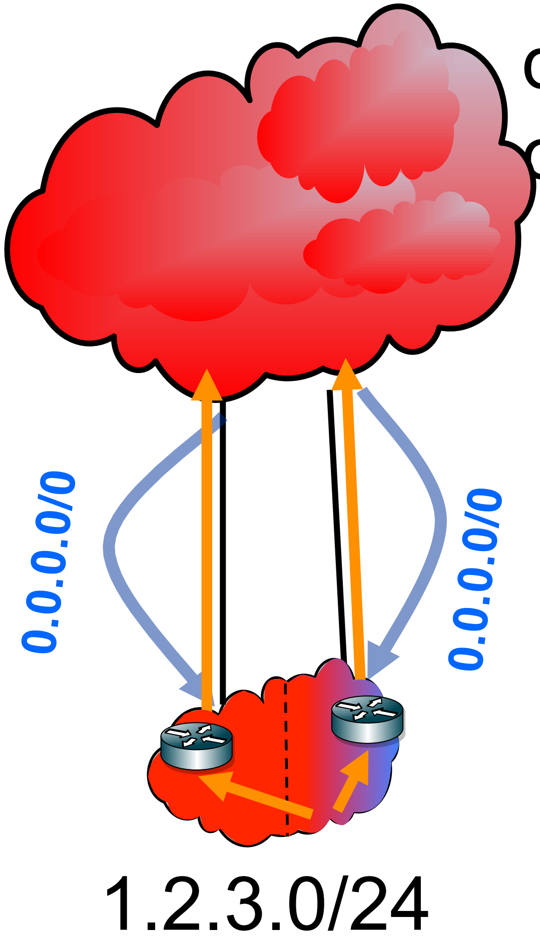
# End-site IGP distancing

- End-site has equally good reachability through two upstream ASes to a BGP prefix
  - –Both providers sending a default route
  - –Both providers sending full routes
    - Applies only to the set of destinations that are equally good through both providers (same AS path length from both providers)

# End-site IGP distancing

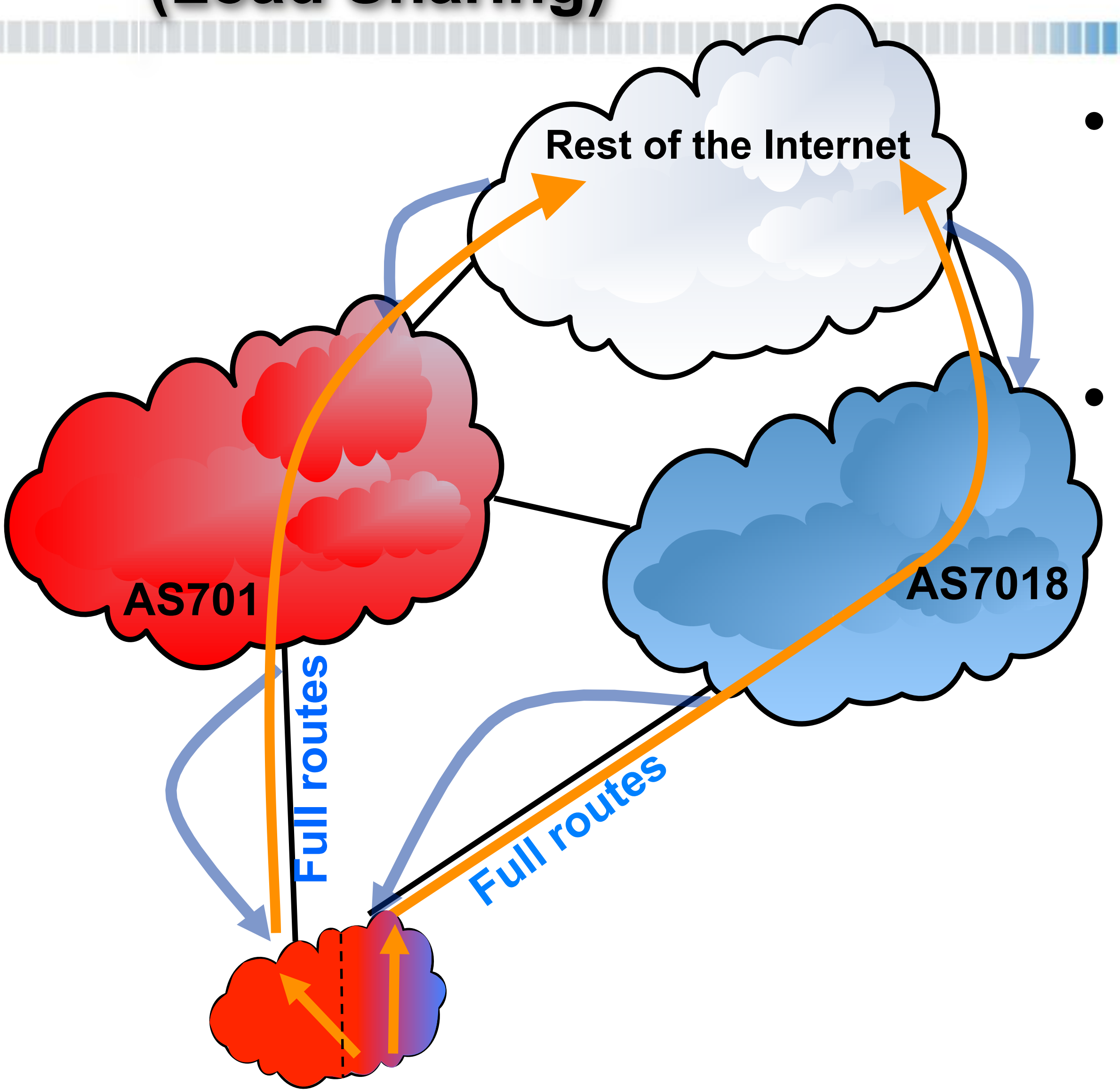
- End-site has equally good reachability through two upstream ASes to a BGP prefix
  - –If it terminates on a single customer router
    - BGP path selection algorithm breaks on oldest path or lowest router ID
  - –If it terminates on two different customer routers
    - Each router will choose the eBGP learned path as best
    - Each router will advertise its path to the iBGP
    - Other iBGP routers will break on IGP distance to the BGP protocol next-hop or router ID

# End-site IGP distancing



- Other iBGP routers will break on IGP distance to the BGP protocol next-hop or router ID
  - –If half of the network is IGP closer to one egress router, then traffic from that half will prefer that egress point
  - –Likewise for the other half
  - –Making a larger portion of the network IGP closer to one egress point will increase the traffic load for that egress

# Default Outbound Customer Traffic (Load Sharing)



- The same behavior can be done for two providers sending default routes
- The same can be done with two providers sending full route
  - Will only shift traffic for destinations that are equally good from both providers

# ISP Advertising Different Routes

# ISP Prefix Advertisement Options

- ISPs often offer different sets of prefixes
  - Customer equipment may not have enough memory for multiple full Internet feeds
  - Customer may prefer a sub-set of routes for traffic engineering
    - Want to make one transit provider preferred for only that ISP’s customers, and us the other provider for everything else
      - Take only customer prefixes from one provider and a default from the other
    - Customer is well Peered outside of the Asia / Pacific, and wants global customer routes and only Peer routes from the Asia / Pacific region

# ISP Prefix Advertisement Options
# Full Routes

- Full Internet routes
  - Everything but the internal more specifics
  - All provider aggregates, all customer routes, all global Peer routes, all regional Peer routes for the local region, all national Peer routes for the local country, and routes from transit providers
- Full Internet routes, no transit
  - Same as Full Internet routes, but no routes from transit providers
- Full regional routes
  - Routes for all local regional aggregates, all in region customers all Regional Peers
- Full continental routes
  - Routes for all local continental aggregates, all customer of the continent, continental Peers
- Full national routes
  - Routes for all local national aggregates, all customers of the country, national Peers

# ISP Prefix Advertisement Options
# No Peer routes

- Global aggregates, and customers
  - Everything but the internal more specifics, and Peer routes
  - All provider aggregates, all customer routes
- Regional aggregates, and customers
  - Routes for all local regional aggregates, all in region customers
- Continental aggregates, and customers
  - Routes for all local continental aggregates, all customer of the continent
- National aggregates, and customers
  - Routes for all local national aggregates, all customers of the country

# ISP Prefix Advertisement Options
# Aggregates Only

- Global aggregates
  - –All provider aggregates
- Regional aggregates
  - –Routes for all regional aggregates
- Continental aggregates
  - –Routes for all continental aggregates
- National aggregates
  - –Routes for all national aggregates

# ISP Prefix Advertisement Options
# Default Route

- Any of the previous offering with a default route
- Only a default route
  - –0.0.0.0/0
- No routes

# End-site Advertising Different Routes

# Advertising More Specifics

- The most granular for of inbound TE is advertising more specific routes
  - –It is often used in conjunction with the other inbound TE mechanisms to further dial traffic around in a more granular fashion
  - –The larger the number of more specific routes, the more granular the TE
    - Routes more specific than a /24 will likely not be accepted outside your direct transit provider, and will likely only impact how your transit provider send you traffic

# Primary / Backup ISPs
# Dialing Up Inbound Traffic on the Backup

- Customer wants a majority of the traffic to use the primary link
- Customer wants some small amount of traffic to "spill over" to the back-up link
  - –Primary link is saturated
- Customer will advertise
  - –Aggregate to both providers with a community for the back-up provider to set low local preference
  - –A more specific route to the back-up provider for some small slice of prefixes that draws the appropriate amount of traffic

# Best Path
# Dialing Up Inbound Traffic on the Backup

- Customer wants a majority of the traffic to use the best path
- Customer wants to draw some small amount of traffic to an underutilized upstream provider
  - Customer wants this traffic to be from sources that do not have a huge disparity in distance from the preferred provider and under utilized provider
  - Customer will AS prepend to the over utilized provider making sources with equally good connectivity to both providers prefer the under utilized provider
    - Customer may send a more specific route (up to /24) to both providers and only AS prepend the more specific to the over utilized provider

# Load Sharing
# Dialing Around Inbound Traffic

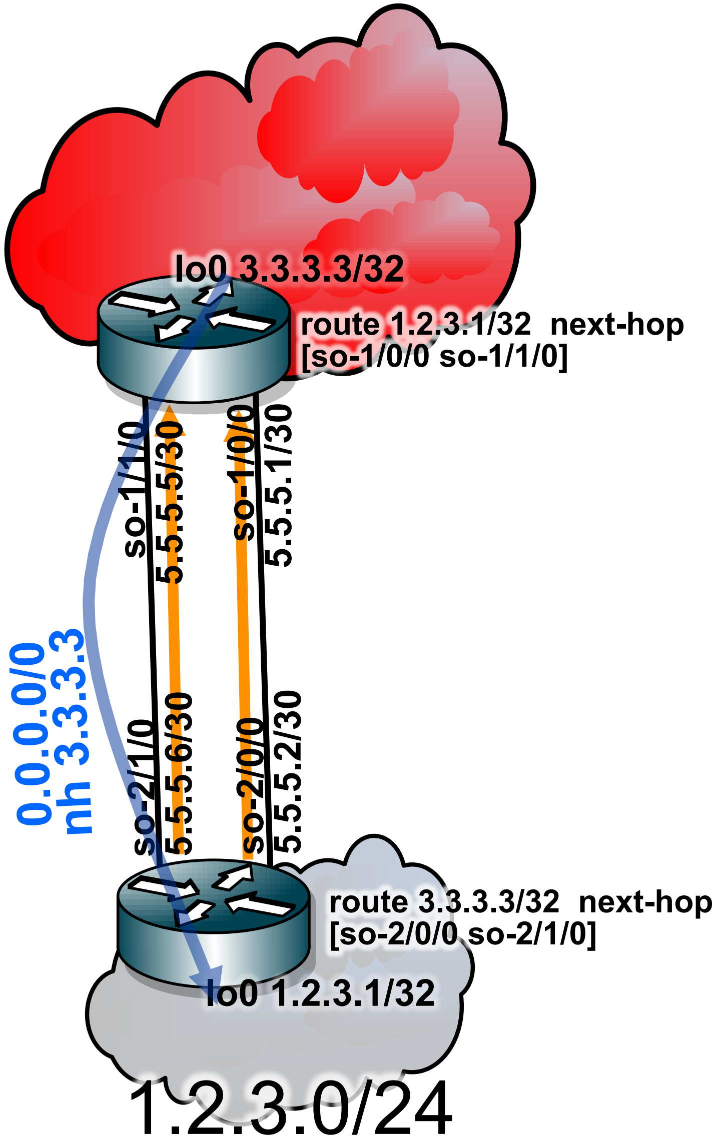
- Customer wants to balance traffic across multiple links (possible of different sizes)
- Customer will advertise the aggregate to all providers over all links
- Customer will advertise various more specific slices across each link
  - Customer will vary the size or specific prefix to draw more or less traffic
- Customer may use local preference communities to indicate where the traffic for the more specific slice should go if the preferred link fails

# Other TE Considerations

# eBGP Multi-hop

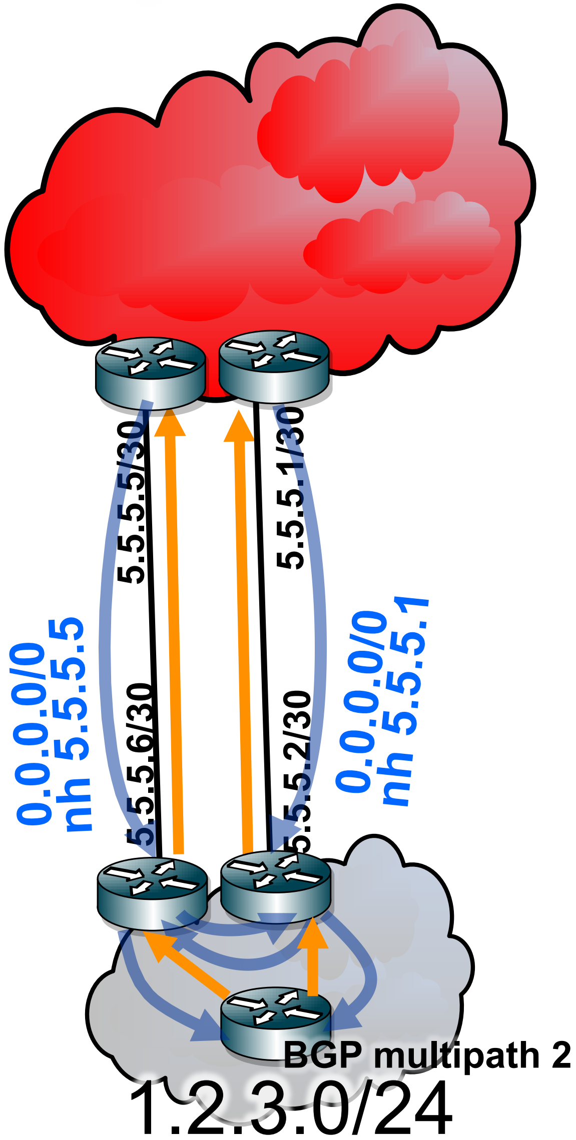
- Allows the BGP neighbor to not be directly connected
  - –Useful when the router terminating the physical circuit cannot support BGP, but a router further in the customer network can
  - –Also useful in peering loopback to loopback and load sharing across multiple parallel links
    - Only work when all the links are between the same pair of routers

# eBGP Multi-hop



- Allows the BGP neighbor to not be directly connected
  - Useful when the router terminating the physical circuit cannot support BGP, but a router further in the network can
  - Also useful in peering loopback to loopback and load sharing across multiple parallel links
    - Only work when all the links are between the same pair of routers

# BGP Multi-path

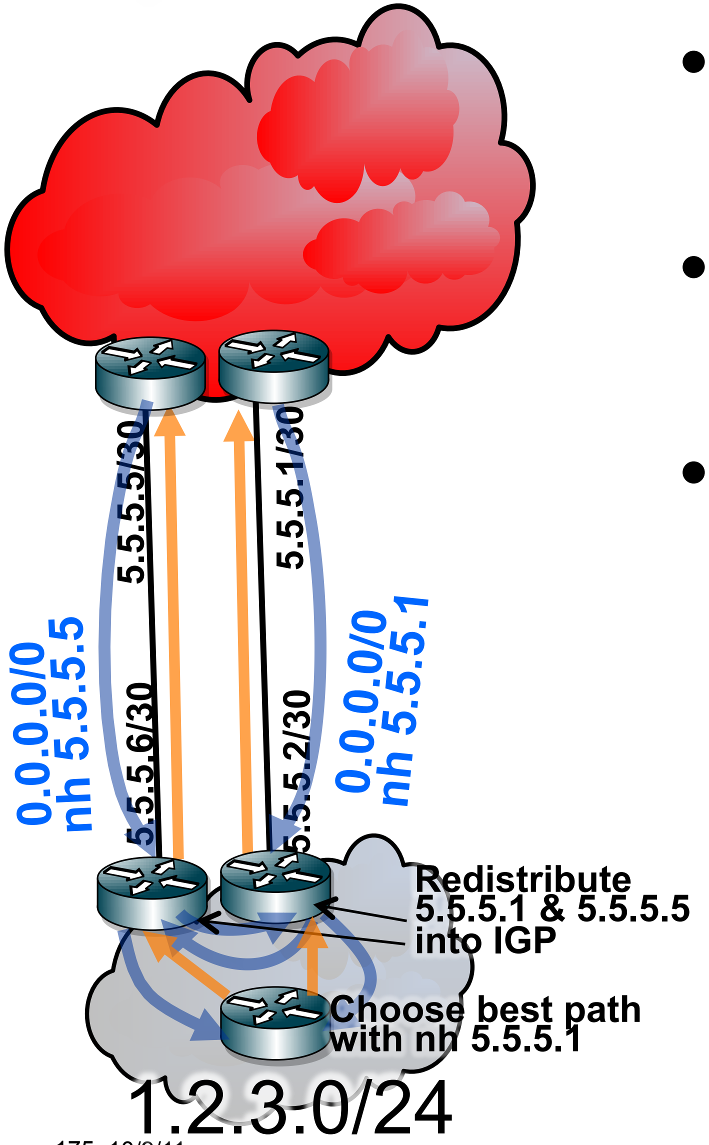


- Router selects one best BGP path
- Router only advertises one best BGP path
- Router will install multiple BGP paths in the FIB if they are equally good up to IGP distance

# Multi-hop vs Multi-path

- Multi-hop
  - Only works when all links are between a single pair of routers
  - Single eBGP session between loopbacks
  - Single eBGP path with a single protocol next-hop
  - Multiple equal cost static routes to each far end link for the BGP protocol next-hop
- Multi-path
  - Only works when each link is to a different eBGP speaking router in the local network
  - eBGP session per link
  - BGP multi-path on all routers speaking iBGP to each eBGP speaking router where one of the paths is terminated
  - As many FIB entries as there are discrete eBGP speaking routers that learn a path (up to the number of BGP multi-path)

# IGP ECMP to the BGP Protocol Next-hop



Redistribute 5.5.5.1 & 5.5.5.5 into IGP

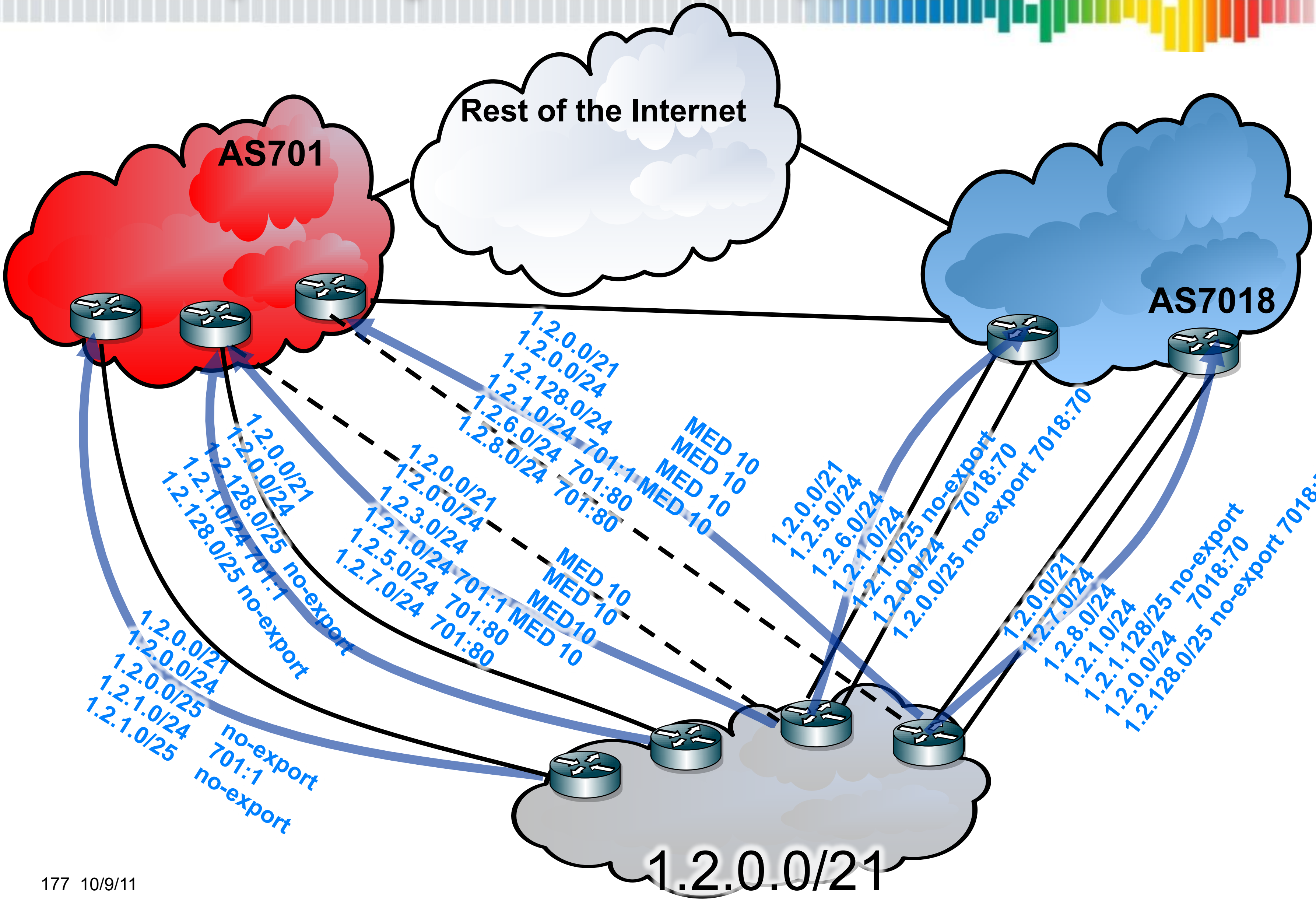
Choose best path with nh 5.5.5.1

- Router selects one best BGP path
- Router only advertises one best BGP path
- Router will install one BGP path in the FIB
  - If both eBGP speaking routers redistribute all protocol next-hops into the IGP then routers that are equidistant to both eBGP routers, will pick one best path, but load both routers ther IGP ECMP

# Complex Layered BGP TE Approach

- 1.2.0.0/24 should use best path out of both of AS701's primary paths and both of AS7018's bundles
- 1.2.1.0/24 should mostly use best path out of both of AS701's primary paths and both of AS7018's bundles, but AS7018 should draw slightly more traffic
- 1.2.2.0/23 should be preferred over AS701 primary links
- 1.2.5.0/22 should be preferred over AS7018's bundles
- Any AS701 traffic that fails should use the AS701 backup links
- If any one AS7018 bundle fails it should use the AS701 backup links

# Complex Layered BGP TE Approach



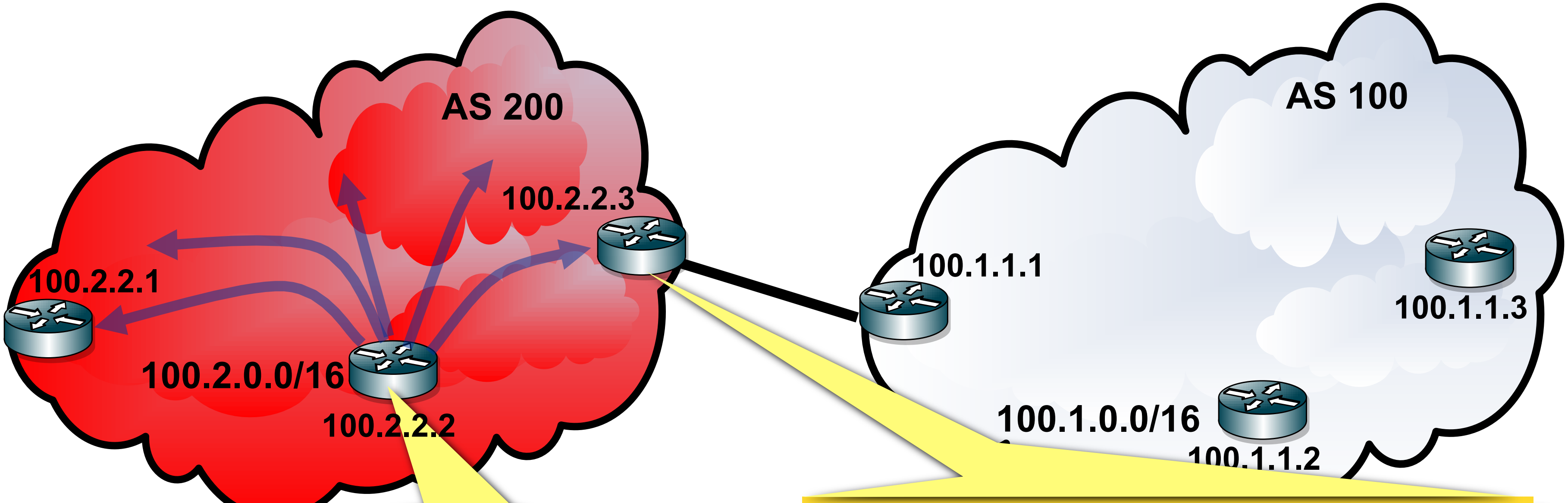
# BGP Protocol Scaling

## Routes and Paths and Sessions.. Oh my!

# eBGP and iBGP

- eBGP is when your BGP neighbors belong to different ASes
  - Every best path is advertised across an eBGP boundary
  - eBGP neighbors are usually directly connected routers
  - Protocol next-hop is typically changed to advertising eBGP neighbor address
- iBGP is when your BGP neighbors belong to the same AS
  - If a path is learned via iBGP do not re-advertise to another iBGP neighbor
    - This requires a full mesh of iBGP peering (more on this later)
    - iBGP neighbors are generally not directly connected
  - Advertise every best eBGP learned path to iBGP neighbors
  - Protocol next-hop is typically unchanged

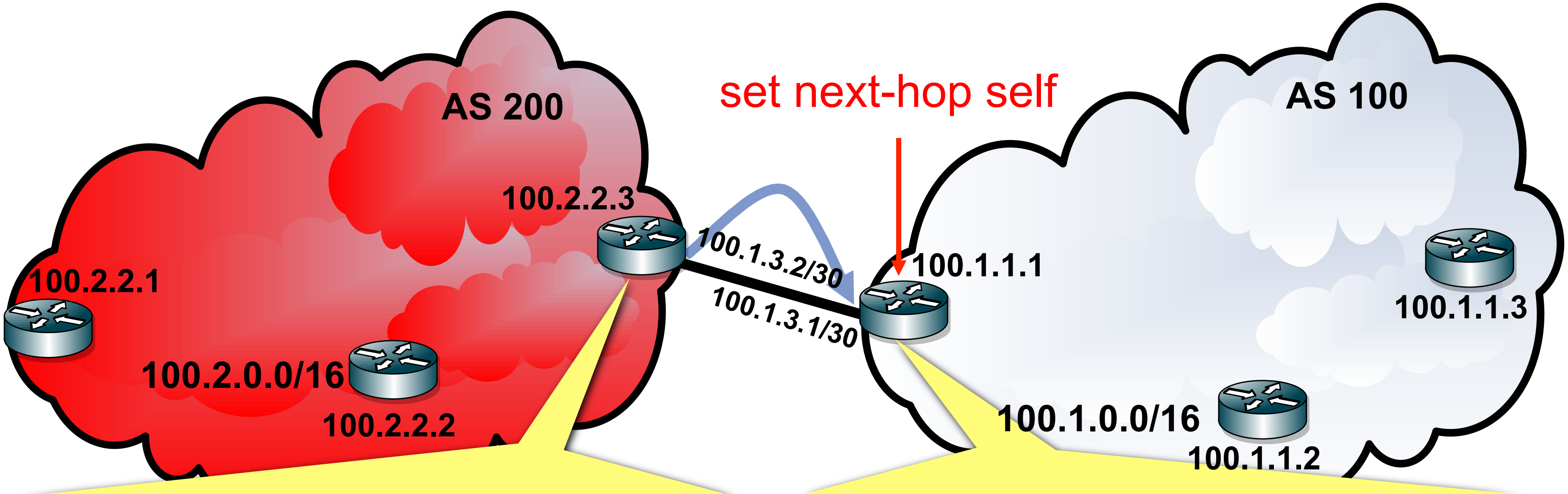
# AS 200 Redistributing Static to BGP Advertising Best Path to iBGP



```
100.2.0.0/16   *[Static/5] Discard
100.2.2.1/32   *[IS-IS/18] metric 1400
                  to 100.2.3.1 via ge-0/0/0
                > to 100.2.3.5 via ge-1/0/0
100.2.2.3/32   *[IS-IS/18] metric 1200
                > to 100.2.3.9 via ge-0/1/0
                  to 100.2.3.13 via ge-1/1/0
100.2.3.0/30   *[Direct/0] via ge-0/0/0
100.2.3.4/30   *[Direct/0] via ge-1/0/0
100.2.3.8/30   *[Direct/0] via ge-0/1/0
100.2.3.12/30 *[Direct/0] via ge-1/1/0
```

```
100.2.0.0/16 *[BGP/170] AS path: I
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
              Protocol next hop: 100.2.2.2
100.2.2.1/32 *[IS-IS/18] metric 2100
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
100.2.2.2/32 *[IS-IS/18] metric 1200
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
100.2.3.16/30  *[Direct/0] via ge-0/0/0
100.2.3.20/30  *[Direct/0] via ge-1/0/0
```

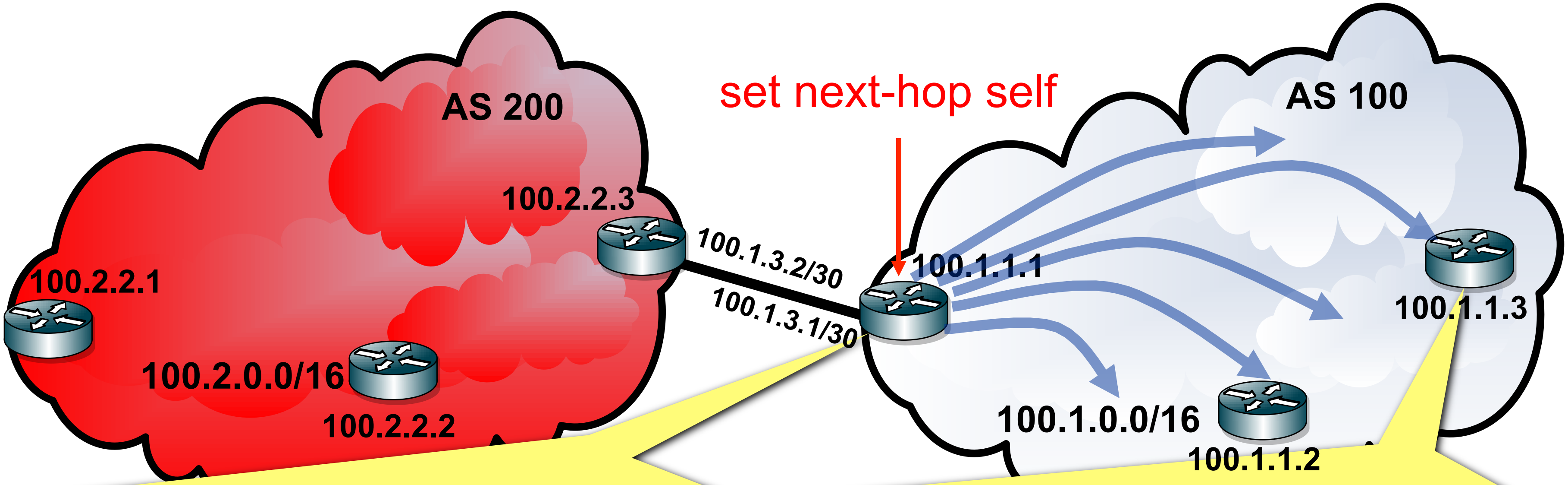
# AS 200 Advertises Best Path to eBGP



```
100.1.3.0/30 *[Direct/0] via ge-1/1/0
100.2.0.0/16 *[BGP/170] AS path: I
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
              Protocol next hop: 100.2.2.2
100.2.2.2/32 *[IS-IS/18] metric 1200
              > to 100.2.3.17 via ge-0/0/0
                to 100.2.3.21 via ge-1/0/0
100.2.3.16/30  *[Direct/0] via ge-0/0/0
100.2.3.20/30  *[Direct/0] via ge-1/0/0
```

```
100.2.0.0/16 *[BGP/170] AS path: 200 I
              > to 100.1.3.2 via ge-1/1/0
              Protocol next hop: 100.1.3.2
100.1.1.2/32 *[IS-IS/18] metric 2200
                to 100.1.3.5 via ge-0/0/0
              > to 100.1.3.9 via ge-1/0/0
100.1.1.3/32 *[IS-IS/18] metric 10010
                to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.3.0/30  *[Direct/0] via ge-1/1/0
100.1.2.4/30  *[Direct/0] via ge-0/0/0
100.1.2.8/30  *[Direct/0] via ge-1/0/0
```

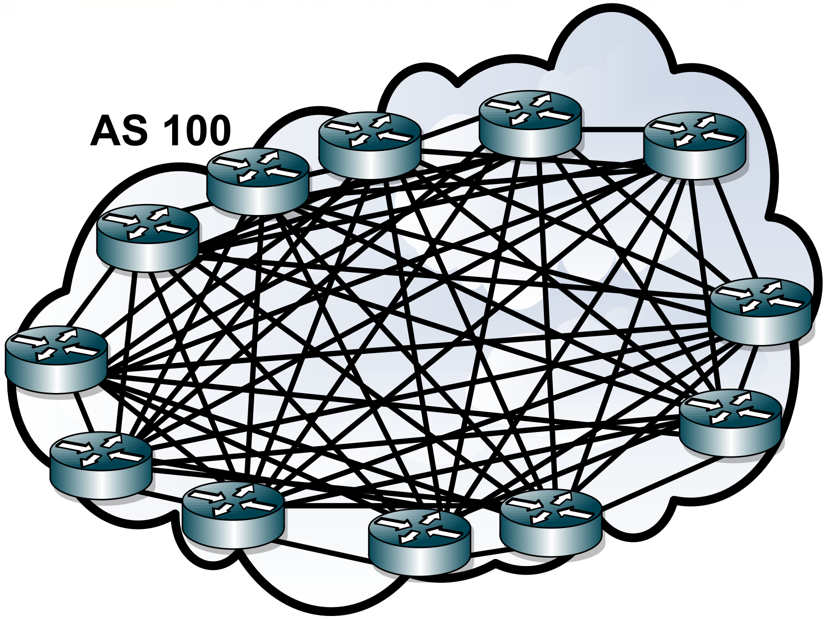
# AS 100 Adbvertises Best eBGP path to iBGP



```
100.2.2.0/16 *[BGP/170] AS path: 200 I
              > to 100.1.3.2 via ge-1/1/0
              Protocol next hop: 100.1.3.2
100.1.1.2/32 *[IS-IS/18] metric 2200
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.1.3/32 *[IS-IS/18] metric 10010
                 to 100.1.3.5 via ge-0/0/0
               > to 100.1.3.9 via ge-1/0/0
100.1.3.0/30  *[Direct/0] via ge-1/1/0
100.1.2.4/30  *[Direct/0] via ge-0/0/0
100.1.2.8/30  *[Direct/0] via ge-1/0/0
```

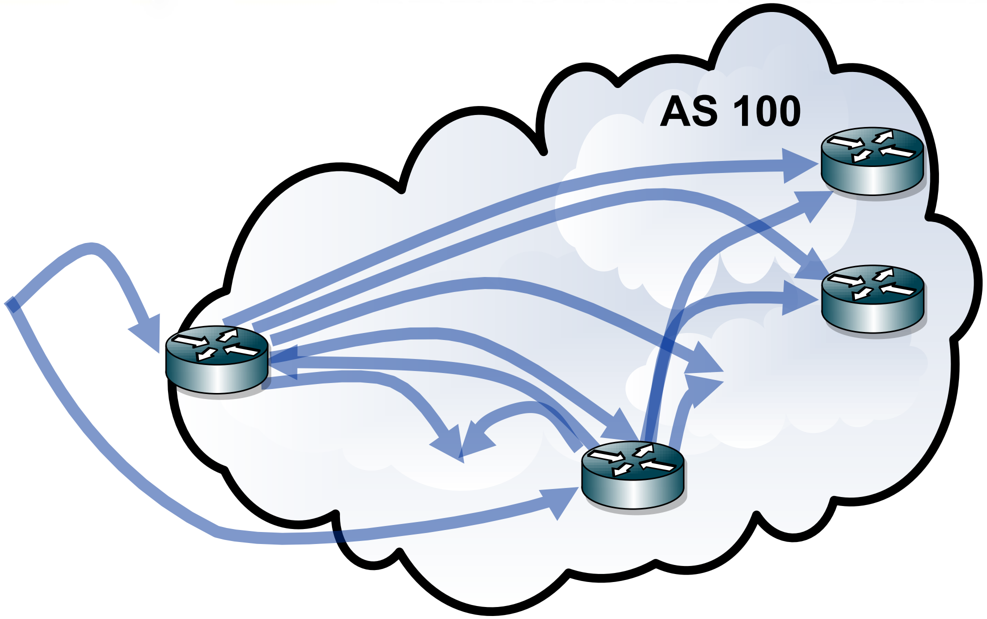
```
100.2.2.0/16 *[BGP/170] AS path: 200 I
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
              Protocol next hop: 100.1.1.1
100.1.1.1/32 *[IS-IS/18] metric 10010
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
100.1.1.2/32 *[IS-IS/18] metric 2000
                to 100.1.3.13 via ge-0/0/0
              > to 100.1.3.17 via ge-1/0/0
100.1.2.12/30 *[Direct/0] via ge-0/0/0
100.1.2.16/30 *[Direct/0] via ge-1/0/0
```

# iBGP Full Mesh



- When a router learns a path from one iBGP neighbor, it does not advertise it to another iBGP neighbor
- This requires that all routers in the AS be configured in a full iBGP mesh
- Full mesh can be difficult to manage
  - Each router needs to have an iBGP session configured to every other router
  - [n*(n-1)/2] iBGP sessions
  - Activating a new router requires the addition of hundreds of neighbor statements
  - Every other router in the AS must be configured with a neighbor statement for the newly activated router
  - A single missing neighbor can result in routing loops or blackholes
- Beyond a few hundred neighbors doesn't scale

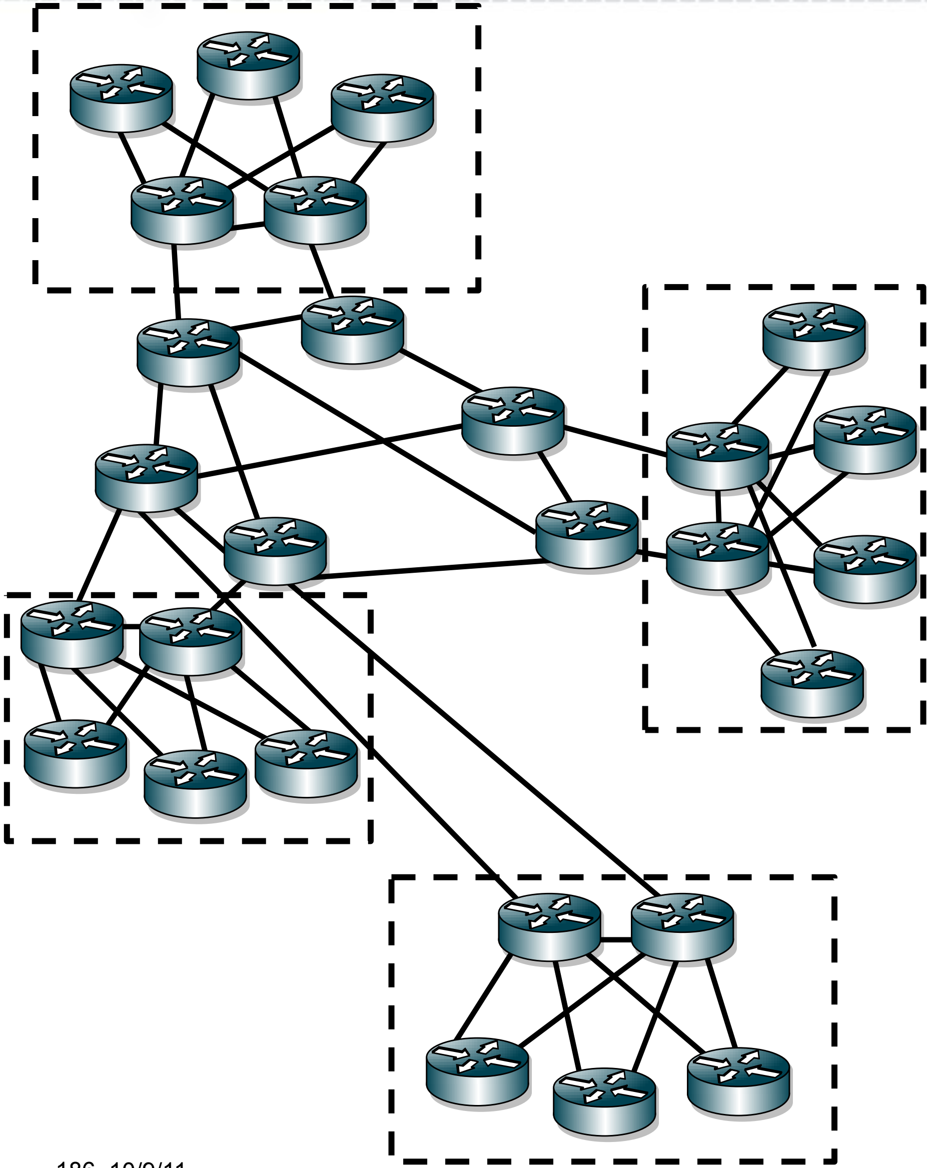
# iBGP Full Mesh
# Path Implications



- When a router learns a path from an eBGP neighbor, if it is selected as best, then it will advertise to all iBGP neighbors
  - All routers will see one path for each singly homed destination
    - If the edge router fails, the path will get withdrawn (path is unreachable anyway)
  - Multi-homed destinations sending the same prefix to multiple routers in the AS will result in as many paths as the number of routers the destination neighbors with

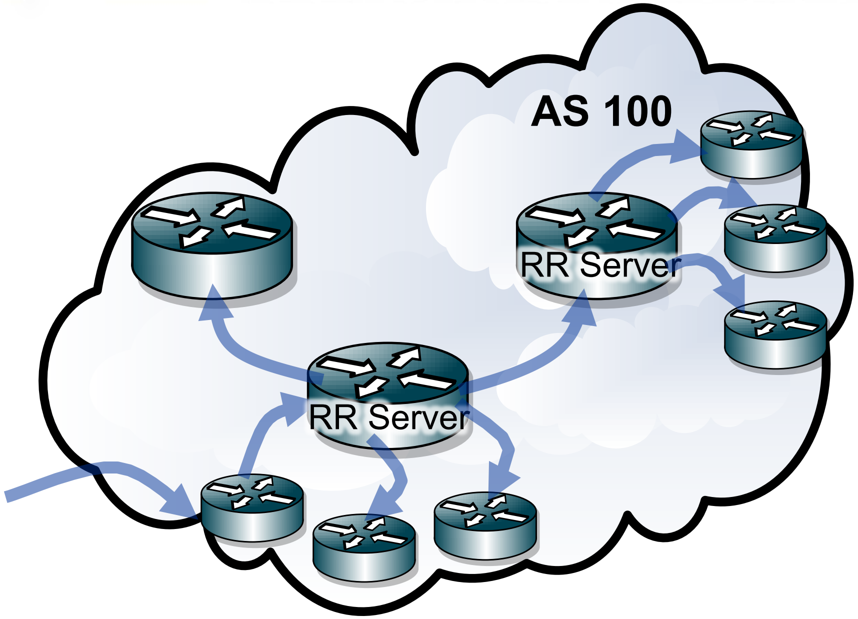
# Route Reflection

# Route Reflection



- Network topology often lends itself to hierarchy
  - Edge Routers
  - Hub aggregators
  - Core routers
- Rather than iBGP full mesh, each router can learn routes for the router one level up who is "reflecting" routes
  - Edge routers are route reflection clients of the hub aggregators
  - Hub aggregators are clients of the core routers
  - Core routers have a full iBGP mesh

# Route Reflection Announcement Rules

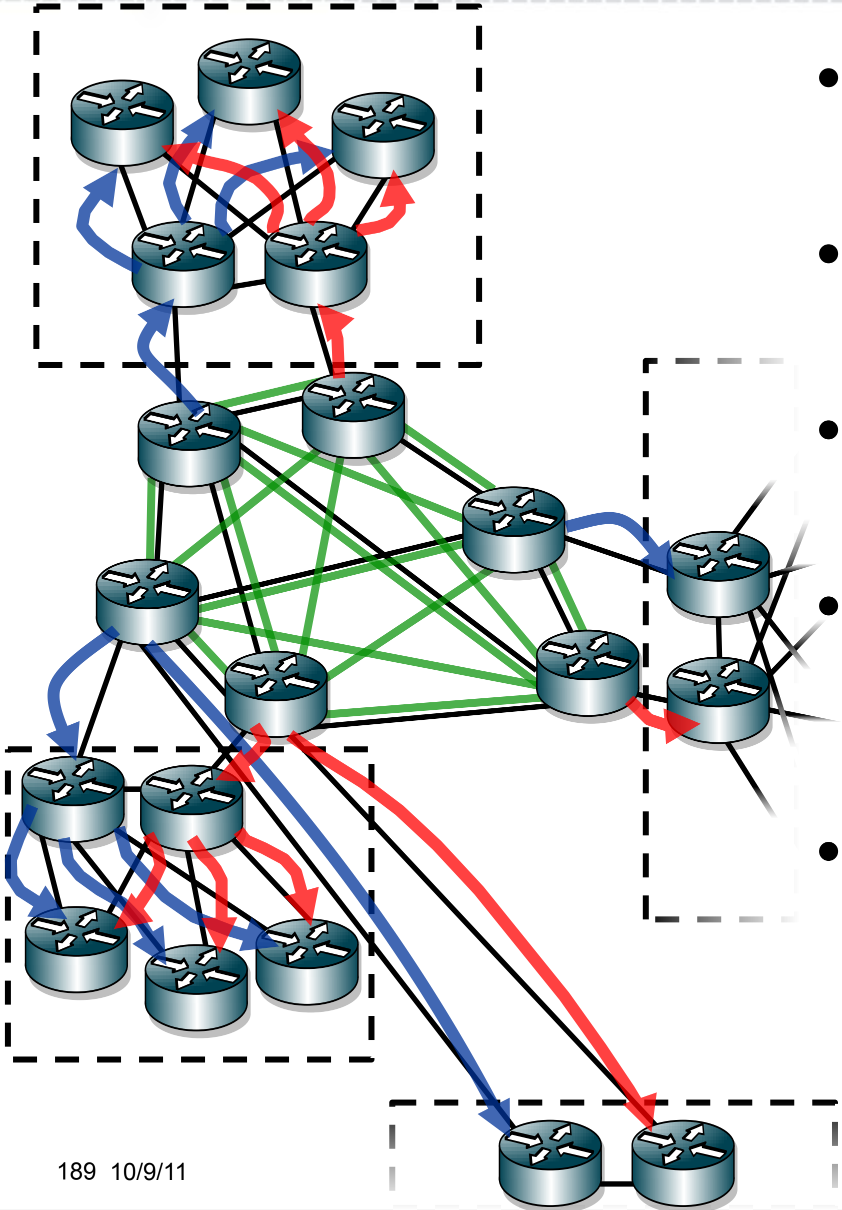


- Rather then fully meshing iBGP, a router may be a client of a route reflection server
- Route reflector servers will reflect best paths from one of its clients to all other clients and non-clients
- Route reflector servers will reflect best paths from non-clients to clients
- Route reflectors servers will NOT reflect best paths from non-clients to other non-clients
  - Standard iBGP behavior
  - Route reflector servers at the highest level of the hierarchy must have a full iBGP mesh
- Route reflector servers can be deployed in hierarchy
  - Router reflector server may have clients and also be a client of another route reflection server
- RFC-4456

# Route Reflection Clusters

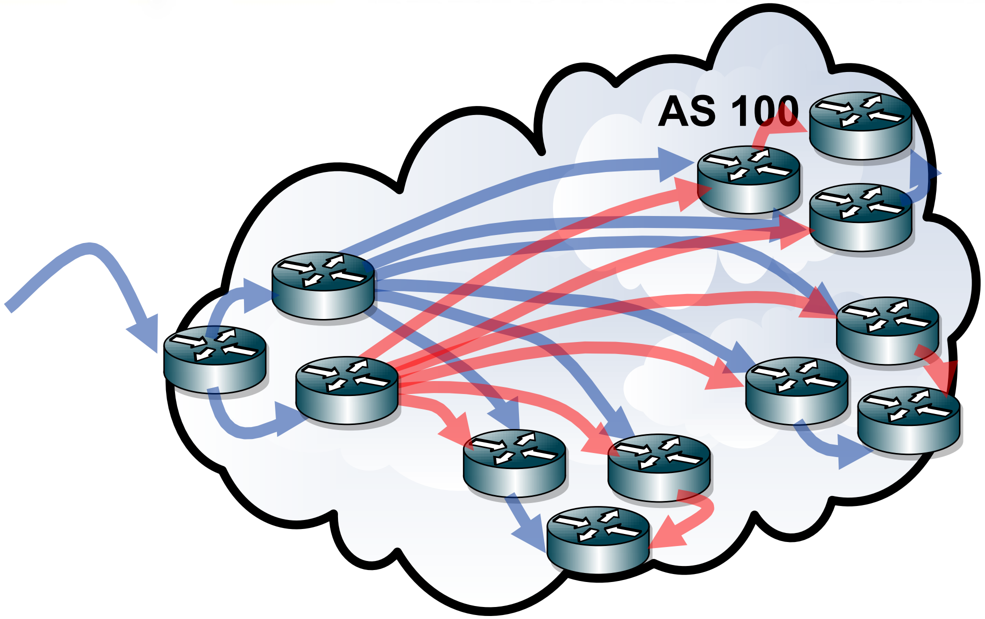
- Each route reflection server and its down stream clients from a grouping called a cluster
  - –Each cluster has a unique number
  - –If each route reflection server is in its own cluster, can use Router ID
    - This is the most common implementation
- Route reflector servers will add their cluster ID to the cluster list when advertising a path
- Route reflector servers will not learn paths with their own cluster ID in the cluster list
  - –Loop prevention

# Route Reflection Construction



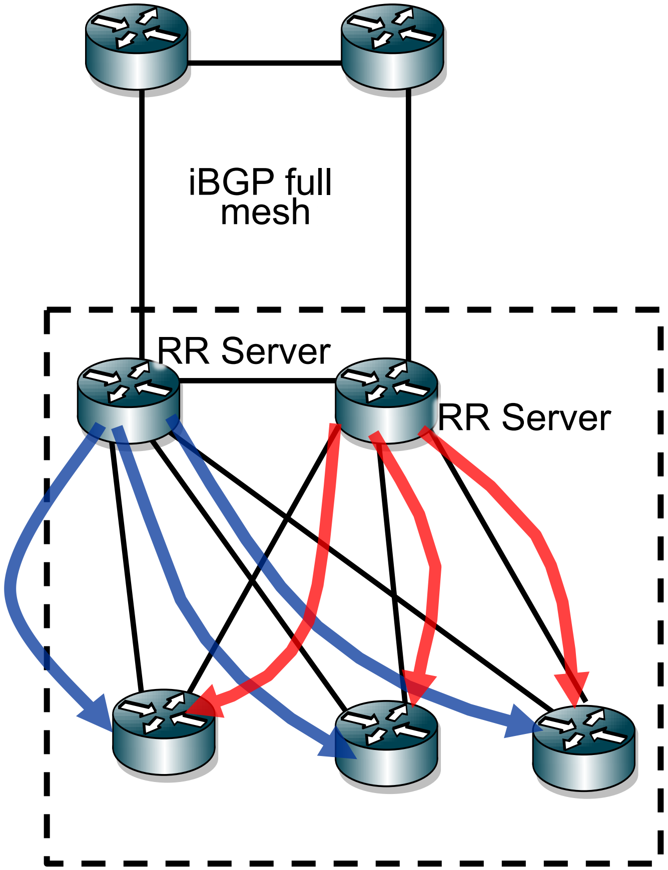
- Each edge device will be a client of both upstream hub aggregators
- Hub aggregators will be route reflector servers for edge routers
- Hub aggregators will be clients of the core routes
- Core routers will be route reflection servers for hub aggregators
- Core routers will be the highest level of route reflection hierarchy
  - –Core routers will have a full iBGP mesh

# Route Reflection Paths



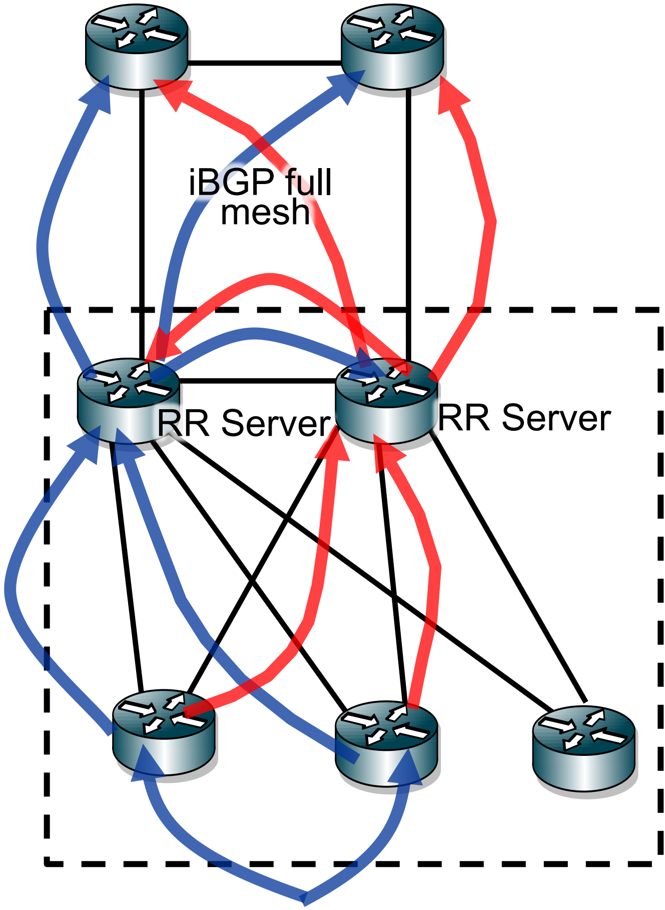
- Route reflection will increase the number of paths for single homed destinations
- In an iBGP full mesh, all routers will have one path for each router that has an eBGP session that learns the route
- In route reflection there will be as many paths as there are route reflection servers serving the cluster(s) where the route is learned
- Path reduction can occur when a destination is multi-homed to edge routers that use the exact same set of route reflector servers
- Path reduction can occur is multiple route reflection servers iBGP peer with each other and they share the same cluster ID

# Route Reflection Clusters and Paths



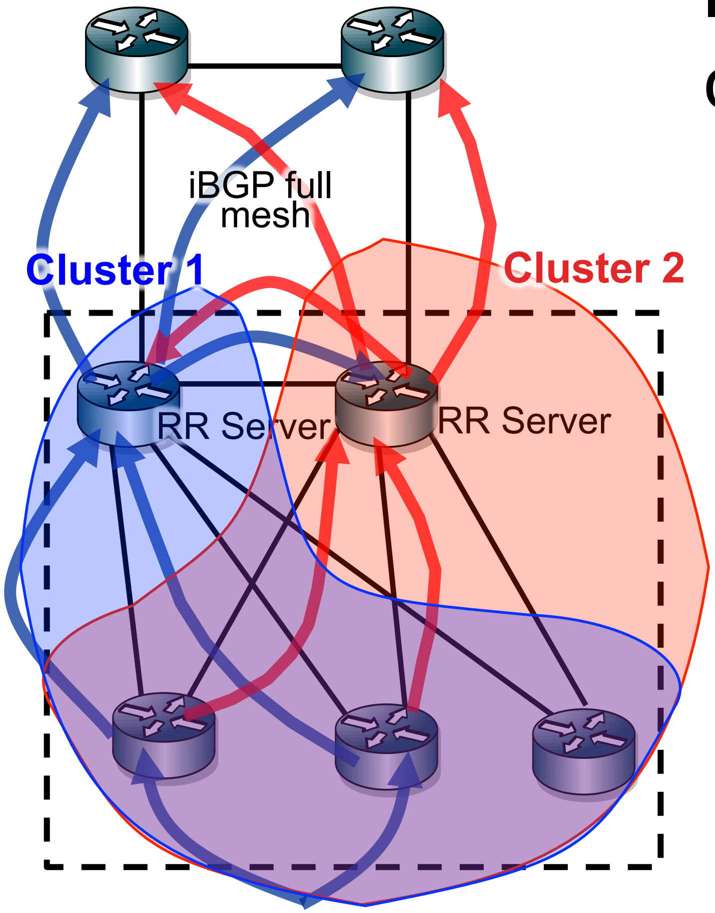
- Each route reflection server will choose one best path and reflect that to its clients
  - Each route reflection client will learn exactly one path for each route in the network from each of its route reflection servers

# Route Reflection Clusters and Paths



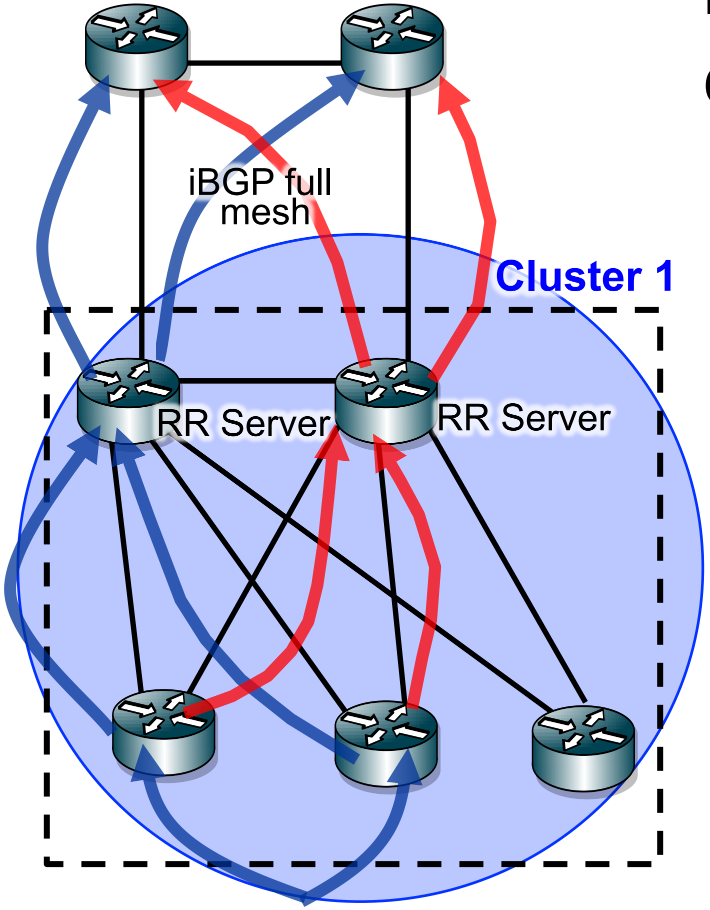
- A route reflection client is not aware that it is a client
  - Follows standard iBGP rules. Best eBGP learned paths are advertised to iBGP (but only its RR servers)
- Each route reflection server will learn one path from each client in the cluster
- Each route reflection server will choose one best path and reflect that to clients (previous slide)
- Each route reflection server will choose one best client path and advertise that to non-clients

# Route Reflection Clusters and Paths



- Route reflection servers will not learn paths that have their own cluster in the cluster list
  - –If each route reflection server is in its own cluster, then a router reflection server will learn extra paths
    - One extra path for each additional route reflection servers that serve the same set of clients

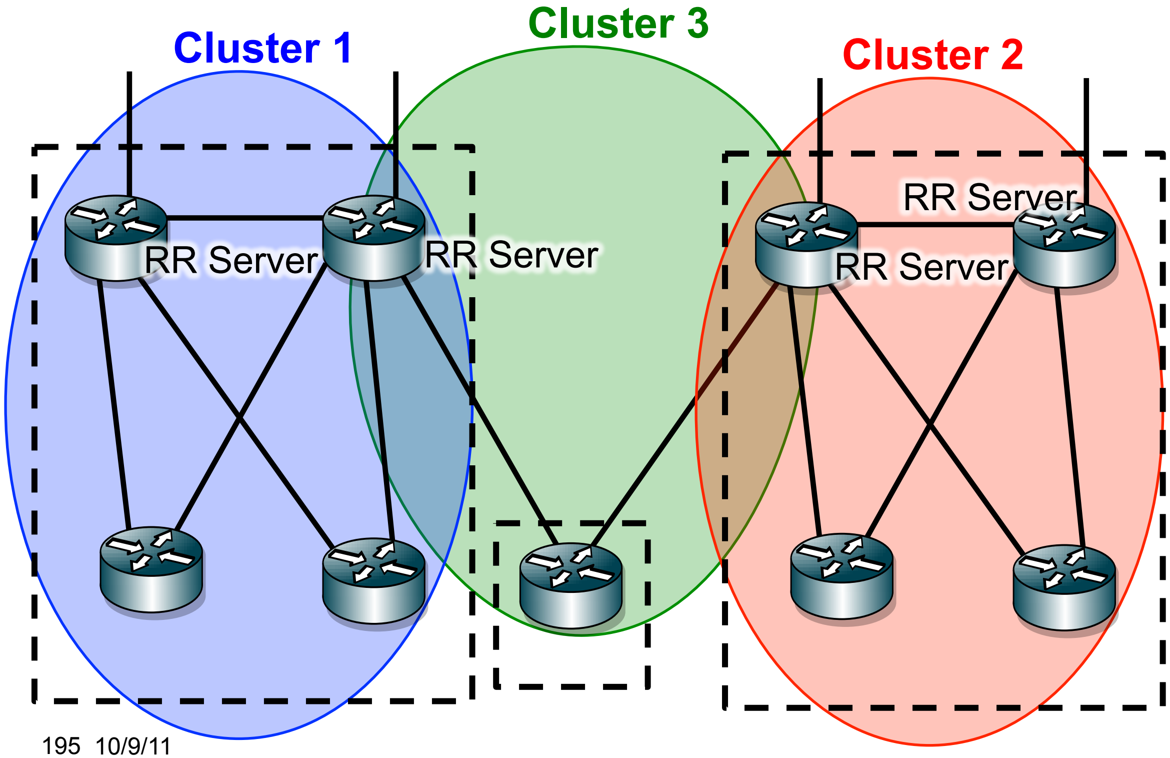
# Route Reflection Clusters and Paths



- Route reflection servers will not learn paths that have their own cluster in the cluster list
  - –If each all of the route reflection servers for a set of clients share a single cluster, then a router reflection server will not learn extra paths from other route reflection servers that serve the same set of clients

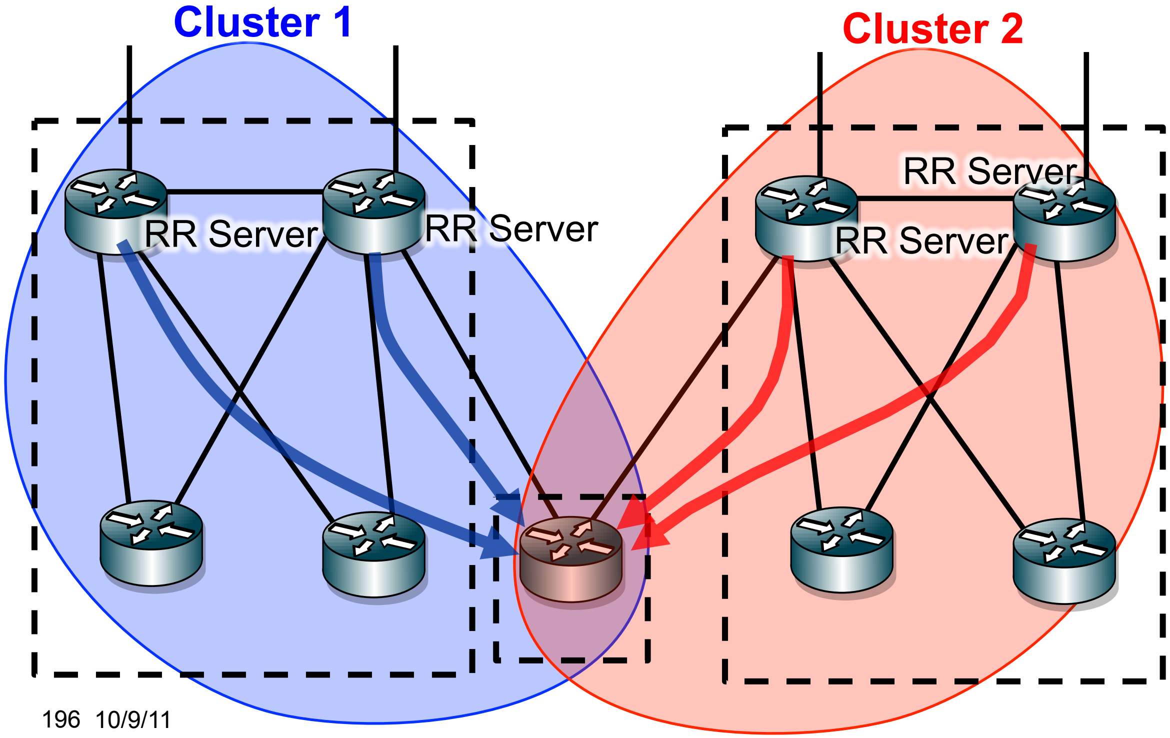
# Route Reflection Clusters and Paths

- It is not always easy to configure all route reflection servers for a set of clients to be a single cluster
  - Clients may have partial over lap with a set of route reflection servers
  - Remote edge router in cluster 3
  - Requires router reflectors to support multiple clusters



# Route Reflection Clusters and Paths

- All route reflection slides so far have demonstrated the route reflection topology following the physical
  - –Following physical is recommended!



–Remote edge router in clusters 1 and 2

–Requires remote edge router to be a client of all four route reflectors

- Increases paths

# Confederations

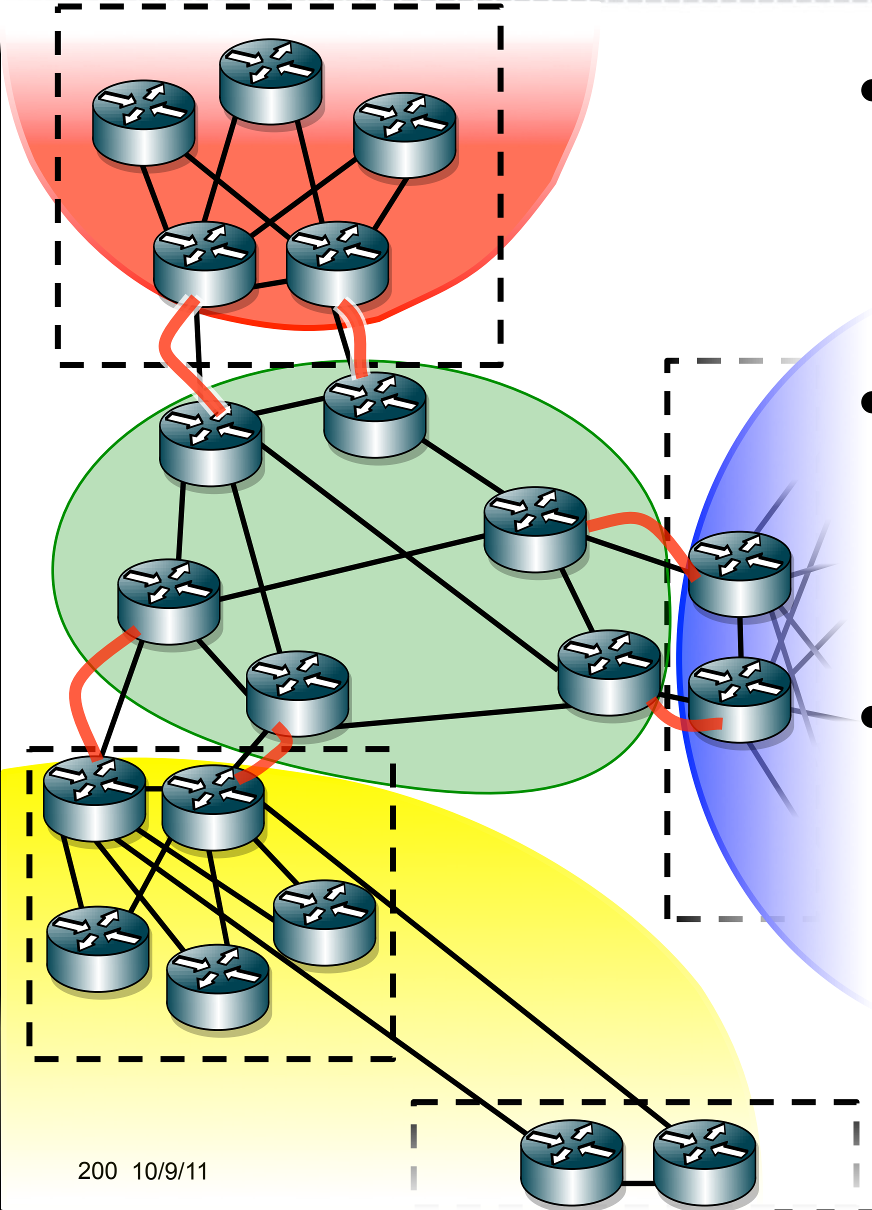
# Confederations

- Rather than building a full iBGP mesh, the network can be divided into sub-ASes
- Each sub-AS will have a full iBGP mesh
- The sub-ASes are connected by eiBGP sessions
  - –Enhanced Interior Border Gateway Protocol
  - –Sometimes called cBGP (confederation BGP)
  - –Has some properties of iBGP and eBGP
- Appears like a single AS to the rest of the Internet
- RFC-5065

# eiBGP Announcement Rules

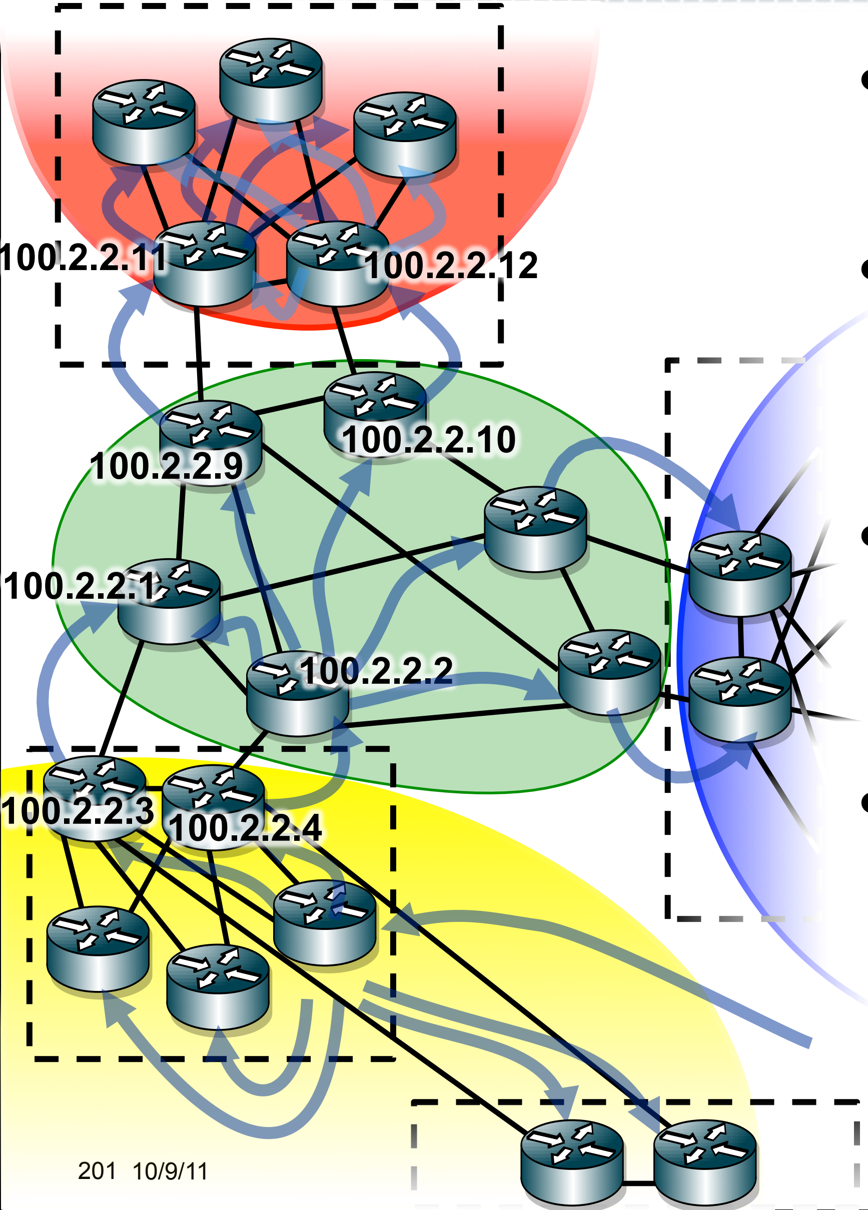
- Send all best paths to eiBGP neighbors
- If best path is learned from eiBGP announce to iBGP
- If best path is learned from iBGP do not announce to iBGP neighbors
- Append sub-AS when advertising to eiBGP
  - Sub-ASes are used for loop detection, but are not counted in the AS hop count
- Strip all sub-ASes when advertising to eBGP
- Do not change next-hop, local preference, or MED values

# Confederation Construction



- Typical implementation is to place the core into its own sub-AS
- Each region that connects to the core becomes its own sub-AS
- Run eiBGP at each point where a regional sub-AS connects to the core sub-AS

# Confederation Paths



- An edge router learns a destination via eBGP
- That edge router announces that destination to all other routers in the sub-AS
- Each eiBGP speaking router will advertise one best path to ieBGP
- Core routers will announce to the core iBGP mesh if their best path was learned from eiBGP

# Route Reflection and Confederations

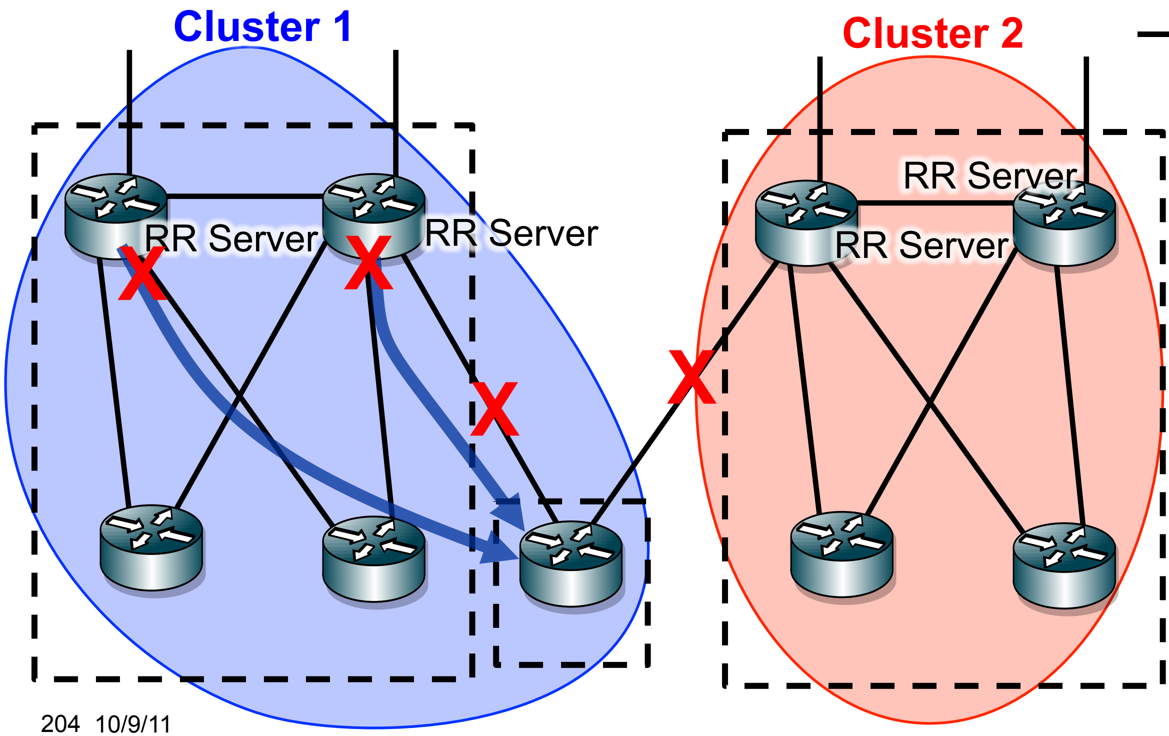
## Comparison & Consideration

# Hazards of the Control Plan Not Following the Physical

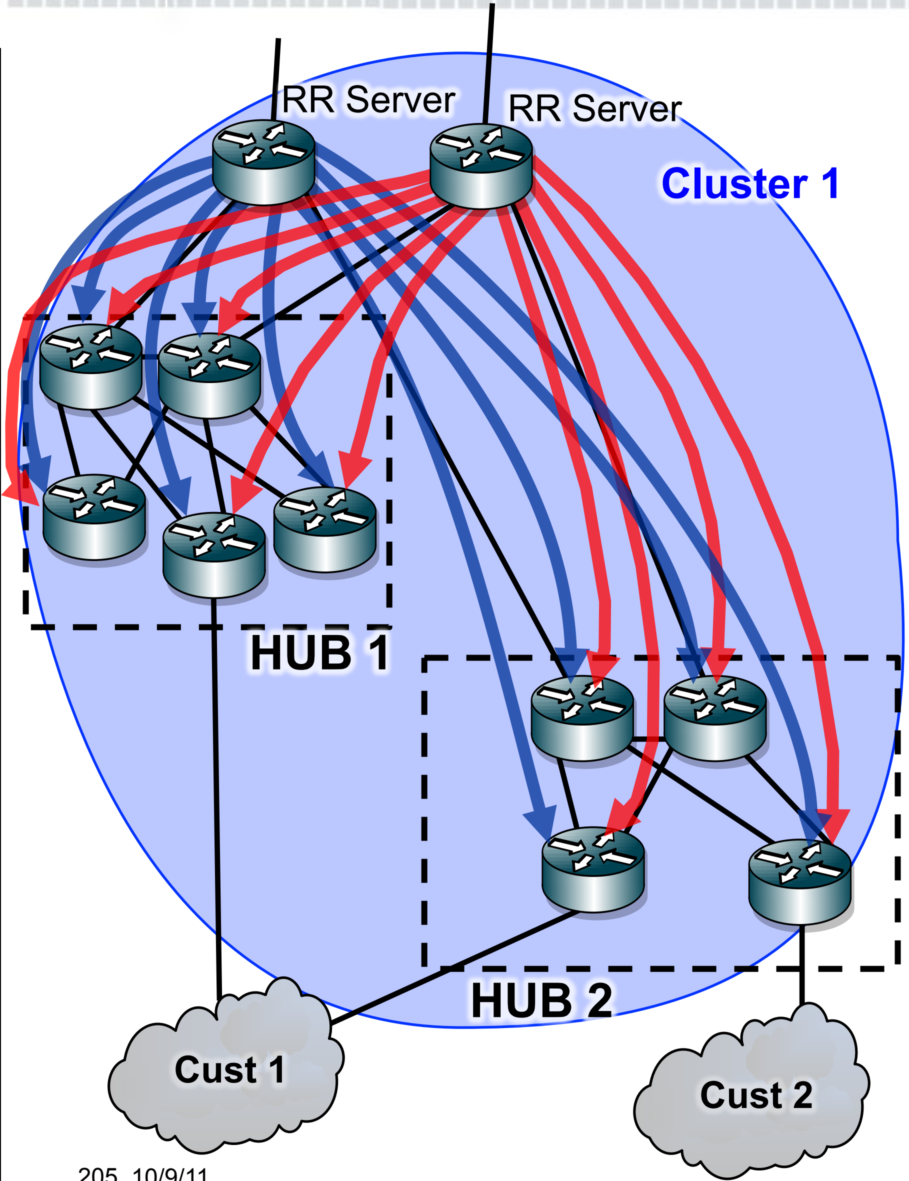
- Always follow the physical
  - –Have route reflection servers be routers higher up in the physical architecture
  - –Configure eiBGP session between all routers that have links to remote sub-ASes
- Inconsistent control plane and forwarding plane:
  - – can be difficult to troubleshoot
  - –increases the number failure scenarios
  - –Causes path reduction that may reduce some paths that are better for some part of the network

# Hazards of Route Reflection Not Following the Physical

- Divorces routing failures from forwarding failures
  - Increases the total number of possible failure conditions
  - Can result in counter intuitive failure conditions
- Can result in path reduction and loss of better paths
  - Two possible failure conditions
    - Both edge router uplinks fail or upstream routers fail isolating forwarding plane
    - Both RRs in cluster 1 fail isolating control plane

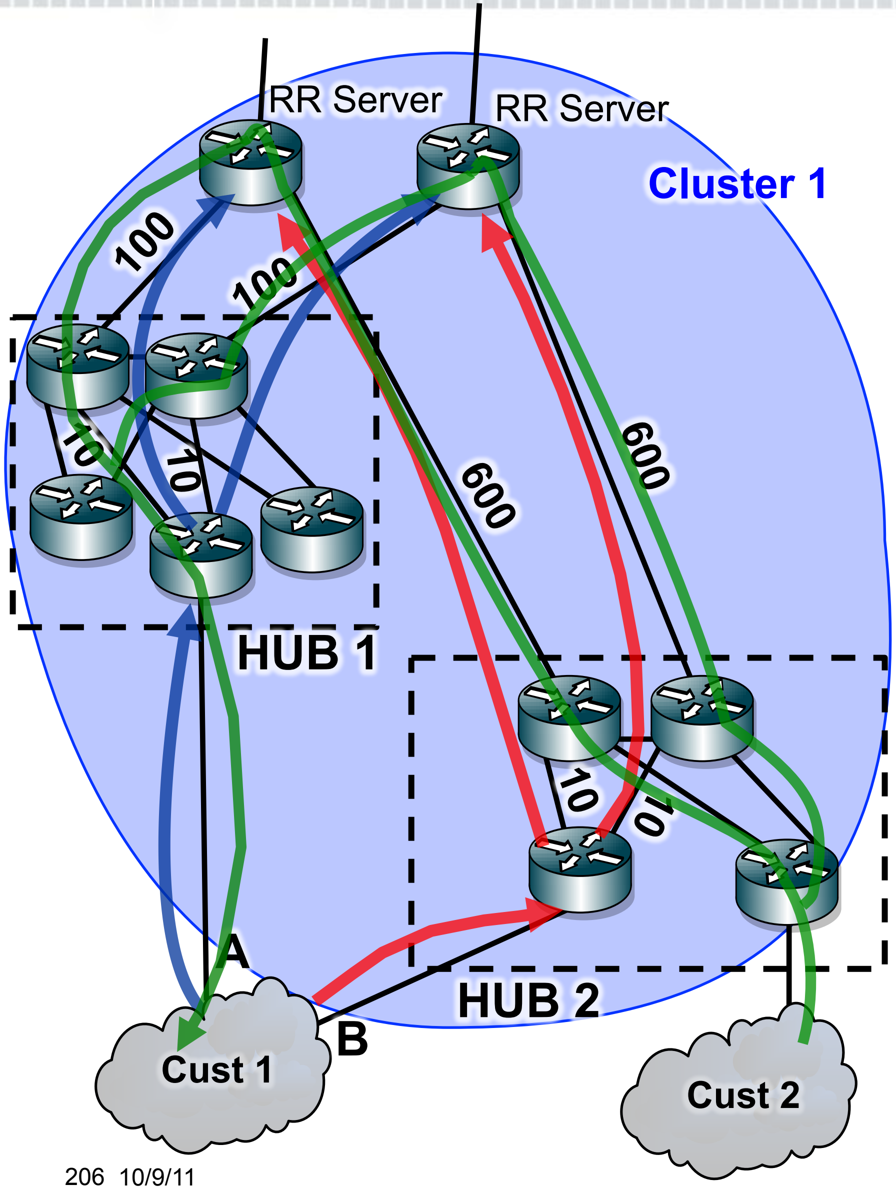


# Hazards of Route Reflection Not Following the Physical



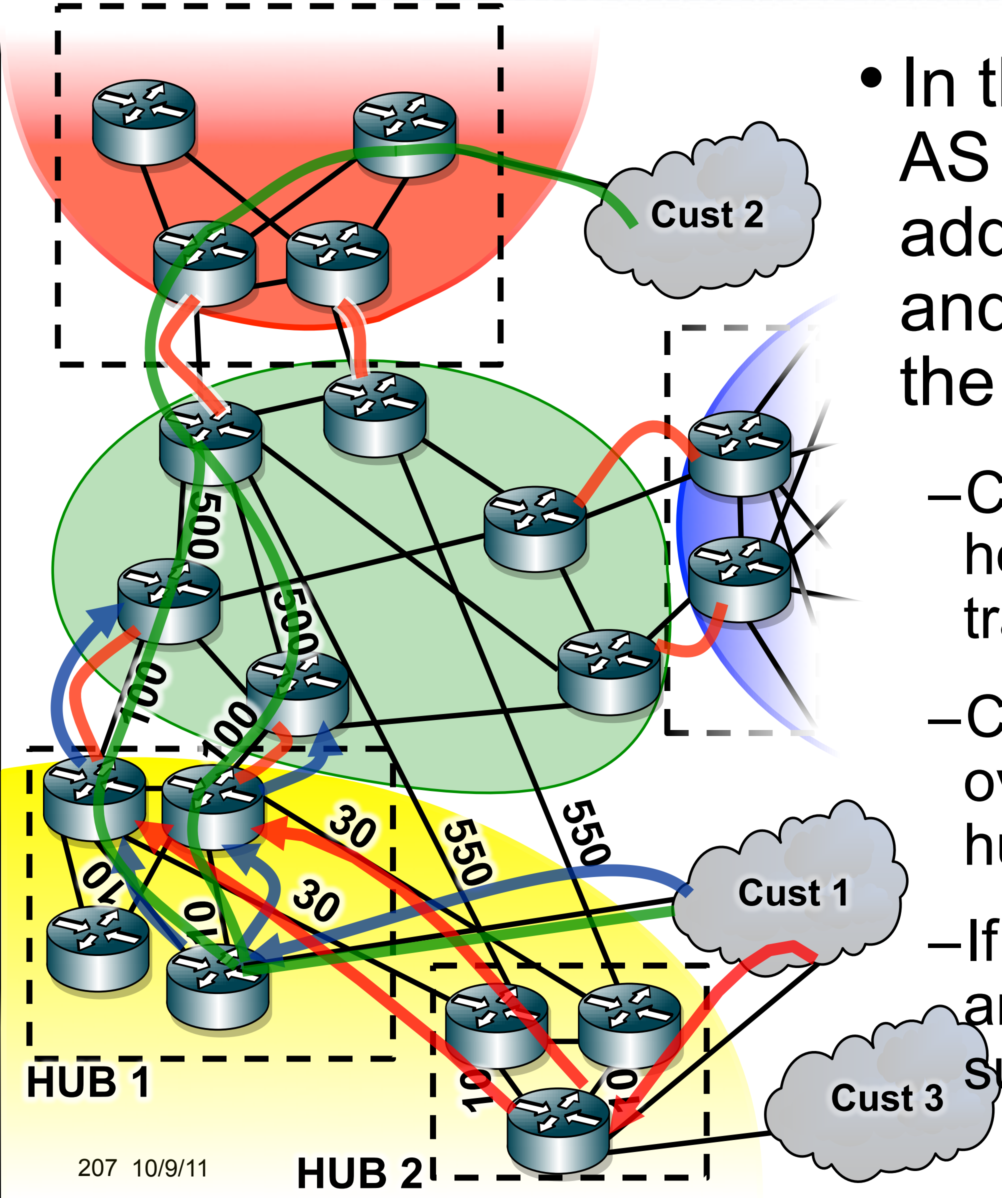
- Core routers are route reflection servers for the entire region
  - –Hub aggregators are clients of the core routers
  - –Edge routers are clients of the core aggregators
- IGP cost from core routers to hub 1 edge routers is much lower than to hub 2 edge routers

# Hazards of Route Reflection Not Following the Physical – Path Reduction



- Customer 1 is multi-homed to HUB 1 and HUB 2
  - Customer 1 send the same route to both routers
- HUB 1 edge router advertises the route to both core routers (blue NH A)
- HUB 2 edge router advertises the route to both core routers (red NH B)
- Core router choose best path
  - Blue with a NH A is best due to its lower IGP cost (110 vs 610)
- Customer 2 will reach Customer 1 via HUB 1

# Hazards of Confederations Not Following the Physical



- In this case, the yellow sub-AS has been provisioned additional links between hub 2 and the transit core, but lacks the associated eiBGP session
  - Customer 2 will reach the multi-homed customer 1 over the transit links to hub 1
  - Customer 2 will reach customer 3 over the short-cut transit link to hub 2
  - If the connectivity between the and hub 1 fails, then the yellow sub-AS will be isolated

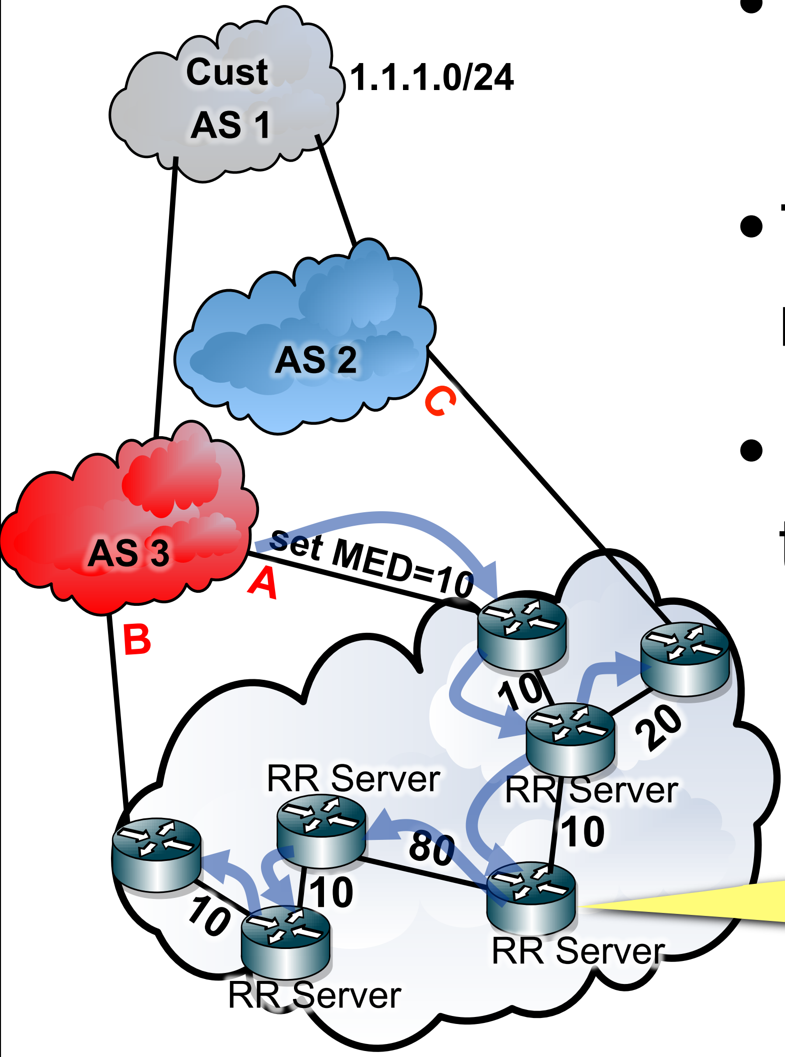
# Confeds vs Route Reflection

- Route reflection is more commonly used
- Migration to route reflection is easier
- Both provide a convenient place to configure routing policy to support geographical division
- Both reduce the iBGP mesh
- Route reflection necessarily increase paths as a function of the number of route reflection servers
- Increase of number of paths can be managed with confederations and careful IP numbering
- Confederations can be used for migration where one network is swallowing another
  - One network is made a sub-AS of the other as a transition step
  - If both networks have non-overlapping sub-ASes, eBGP can simply be replaced with eiBGP once the IGP is unified

# MED Oscillation

- MED oscillation can occur in any hierarchical network
- Have three paths where
  - –Route B is better than route A
  - –Route C is better then route B
  - –Route A is better than C
- Can occur in a flat network where vendors do pair wise path comparison
  - –Is non-deterministic, but not persistent
  - –Cisco IOS
  - –JunOS with “cisco-non-deterministic” MEDs

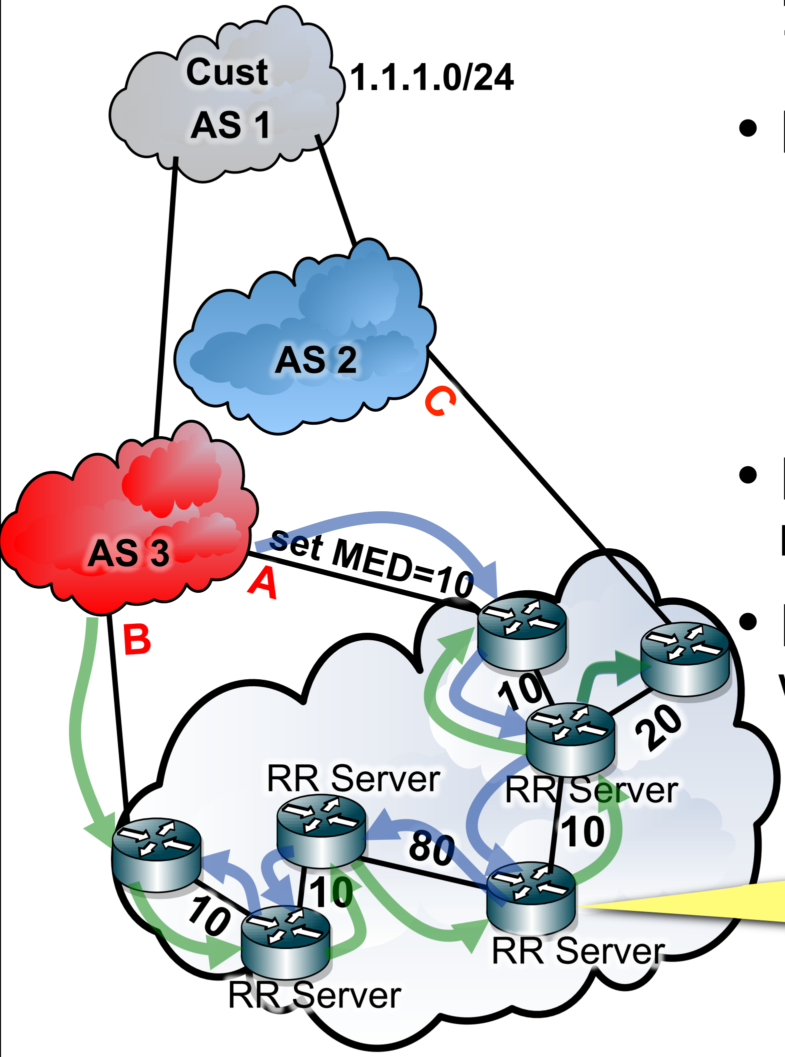
# MED Oscillation



- The network learns a path for 1.1.1.0/24 with next-hop A
- This is the only path for this route and is therefore best
- Path A is advertised throughout the network

| | Route | NH | AS Path | MED | IGP |
|---|---|---|---|---|---|
| B | 1.1.1.0/24 | A | 3 1 | 10 | 20 |

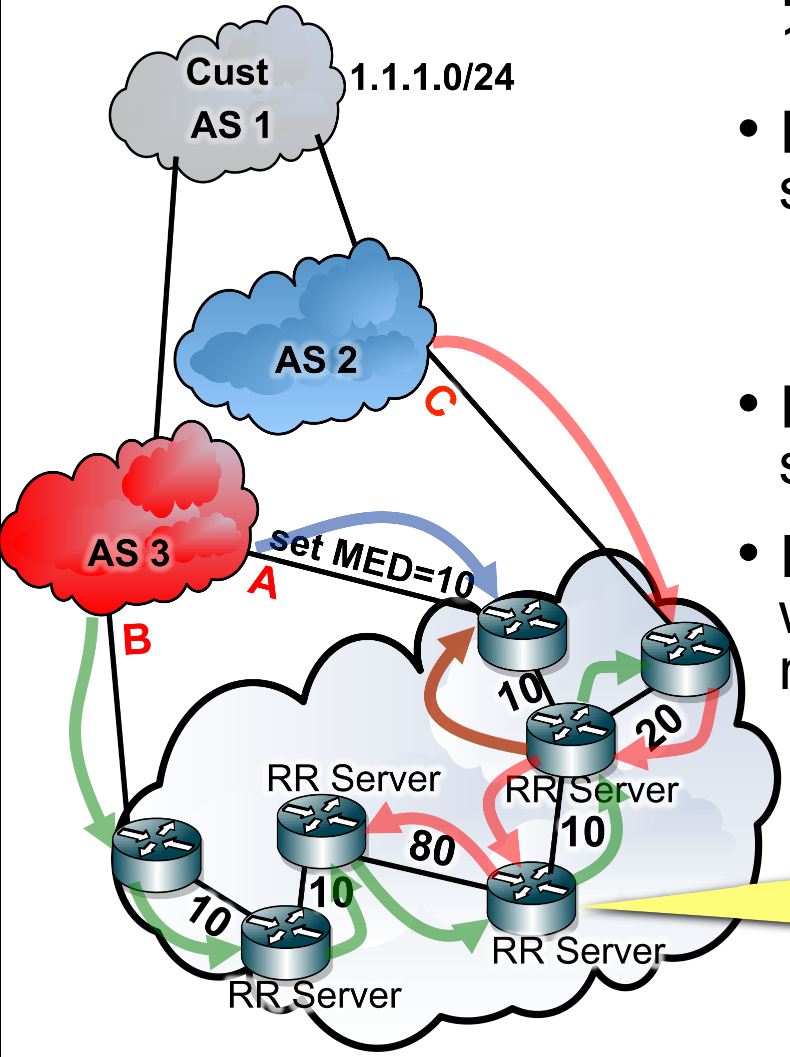
# MED Oscillation



- Next the network learns a path for 1.1.1.0/24 with next-hop B
- Path B is better than path A
  - Path B has a lower MED (0) than path A (MED 10)
  - The neighbor AS is 3 in both cases so MED is compared
- Path B is advertised throughout the network
- Path A is no longer best and is withdrawn

| Route | NH | AS Path | MED | IGP |
|---|---|---|---|---|
| ~~B 1.1.1.0/24~~ | ~~A~~ | ~~3 1~~ | ~~10~~ | ~~20~~ |
| B 1.1.1.0/24 | B | 3 1 | 0 | 100 |

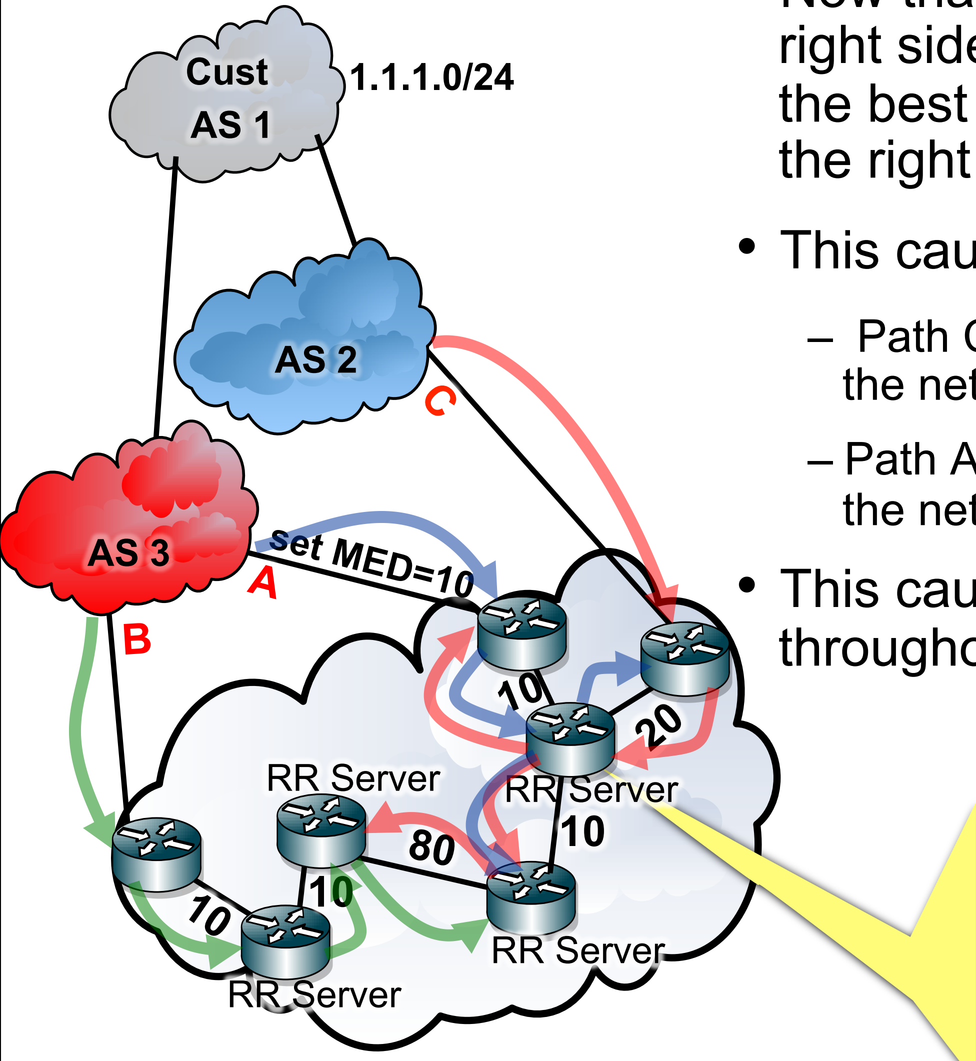
# MED Oscillation



- Finally the network learns a path for 1.1.1.0/24 with next-hop C
- Path C is better than path B on the right side of the network
  - Path C has a lower IGP cost (20/30) than path B (110/100)
- Path C is advertised throughout the right side of the network
- Path B is no longer best and is withdrawn from the right side of the network

| | Route | NH | AS Path | MED | IGP |
|---|---|---|---|---|---|
| | ~~1.1.1.0/24~~ | ~~A~~ | ~~3 1~~ | ~~10~~ | ~~20~~ |
| B | 1.1.1.0/24 | B | 3 1 | 0 | 100 |
| B | 1.1.1.0/24 | C | 2 1 | 0 | 30 |

# MED Oscillation



- Now that B has been withdrawn from the right side of the network, path A becomes the best path and is advertised throughout the right side of the network
- This causes path C to no longer be best
  - Path C is withdrawn throughout the right side of the network
  - Path A is advertised throughout the right side of the network
- This causes path B to be best and advertise throughout the right side of the network …

| | Route | NH | AS Path | MED | IGP |
|---|---|---|---|---|---|
| | ~~1.1.1.0/24~~ | ~~A~~ | ~~3 1~~ | ~~10~~ | ~~20~~ |
| | ~~1.1.1.0/24~~ | ~~B~~ | ~~3 1~~ | ~~0~~ | ~~100~~ |
| B | 1.1.1.0/24 | C | 2 1 | 0 | 20 |
| B | 1.1.1.0/24 | A | 3 1 | 10 | 10 |

# BGP Mechanics

# BGP Mechanics

- BGP uses TCP port 179
  - TCP is reliable transmission
    - 3 way handshake
    - Session numbers
    - Flow control
    - Retransmits
    - TCP MD5 signature

# BGP Messaging

- Open
  - BGP capabilities
    - BGP version/IPv4/IPv6/unicast/multicast/labeled/timer values/authentication type
  - Determined at time of BGP open
    - Each neighbor presents a set of capabilities
    - Common capabilities are agreed upon
- Update
  - Prefix withdrawn / announced / modified
    - Incremental updates
  - Attributes
    - Origin, AS Path, Next Hop, MED, local-pref, Atomic Aggregate, Aggregator …
- Notification
  - Explain an unexpected behavior (various codes and sub-codes)
  - Tear down the session
- Keepalive

# BGP Capabilities

| Value | Description | Reference |
|---|---|---|
| 0 | Reserved | RFC-5492 |
| 1 | Multiprotocol Extensions for BGP-v4 | RFC-2858 |
| 2 | Route Refresh Capability for BGP-4 | RFC-2918 |
| 3 | Outbound Route Filtering Capability | RFC-5291 |
| 4 | Multiple routes to a destination capability | RFC-3107 |
| 5 | Extended Next Hop Encoding | RFC-5549 |
| 6-63 | Unassigned | |
| 64 | Graceful Restart Capability | RFC-4724 |
| 65 | Support for 4-octet AS number capability | RFC-4893 |
| 66 | Deprecated (2003-03-06) | |
| 67 | Support for Dynamic Capability | Enke_Chen |
| 68 | Multisession BGP Capability | Chandra_Appanna |
| 69 | ADD-PATH Capability | draft-ietf-idr-add-paths |
| 70 | Enhanced Route Refresh Capability | draft-keyur-bgp-enhanced-router-refresh |
| 71-127 | Unassigned | |
| 128-255 | Reserved for Private Use | RFC-5429 |

http://www.iana.org/assignments/capability-codes/capability-codes.xml

## Multiprotocol Extensions

- Allows BGP to support more than IPv4 Unicast

  IPv4 multicast, IPv6, IPv6 multicast,IPv4/IPv6 layer 3 VPN, IPv4/IPv6 multicast VPN, …

- Communicated through AFI / SAFI
  - AFI – Address Family Identifier
  - SAFI – Subsequent Address Family Identifier

| AFI | Meaning |
|---|---|
| 1 | IPv4 |
| 2 | IPv6 |

| SAFI | Meaning |
|---|---|
| 1 | Unicast |
| 2 | Multicast |
| 3 | Unicast and Multicast |
| 4 | MPLS Label |
| 128 | MPLS-Labeled VPN |

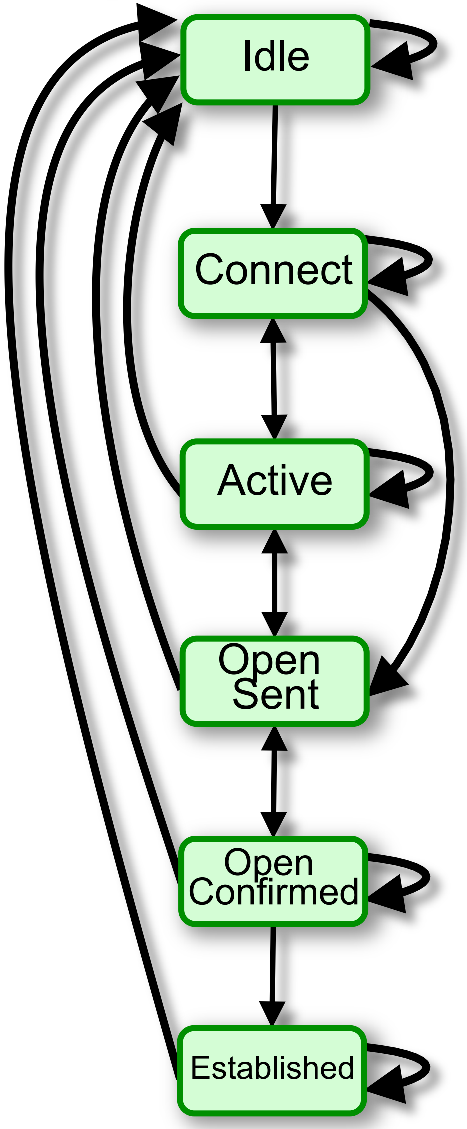
# BGP Capabilities
# Route Refresh

- In bound policy changes on an already established session take affect for only new updates
- To see the impact of a policy change the inbound router needs to keep a copy of all routes prior to the application of policy
  - RIB-IN
  - Consumes memory
- Route refresh allows the inbound router to ask the outbound router to resend its updates

# BGP Capabilities
# Graceful Restart

- Router indicates that it can continue to forward packets when routing capabilities stops
  - NSF (non-stop forwarding)
- Useful when the BGP process restarts or the router switches from primary to backup RE/RP
- When the BGP process restarts
  - It notifies its neighbors of a restart
  - Neighbors mark the state as stale, and continue to forward on it for a brief period of time
  - It requests a new set of updates while forwarding on the old state which is marked as stale and made less prefered
  - Once all the BGP routes are relearned, and the IGP converges, the new routing state is installed in the FIB
  - Stale routing state in the FIB ages out

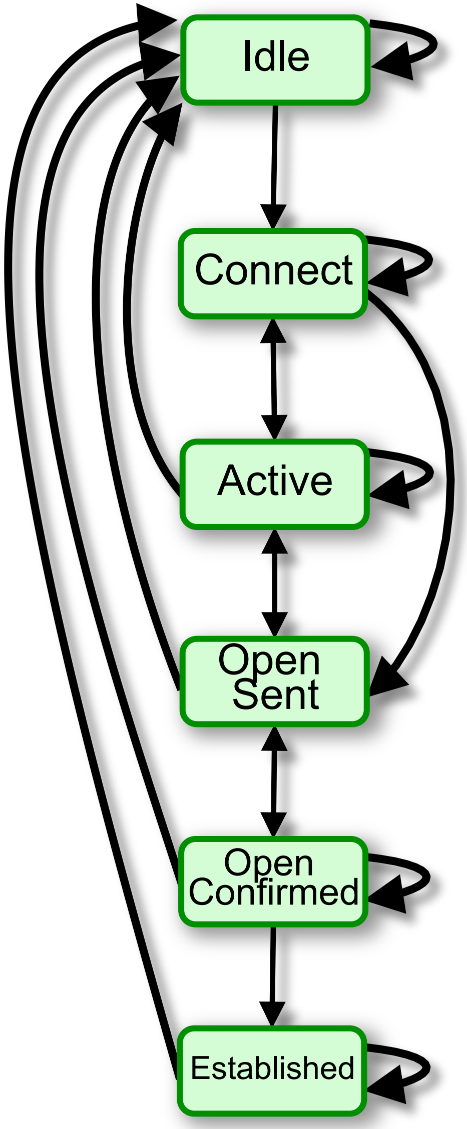
# BGP State Machine



- Idle state
  - Housekeeping
  - Open a TCP session
  - Listen for TCP session
- Connect State
  - Wait for TCP session establishment
  - Send BGP open message post TCP establishment – move to open sent state
  - Move to active state on errors
- Open sent state
  - Listen for BGP open message
  - Check if received open message is valid and there is not compatibility mismatch
    - (wrong AS expected, incompatible family, incompatable version, wrong MD5)
  - If there is a mismatch, send a notification
  - If there is not a mismatch send keepalives – move to open confirmed state

RFC-4271 section 8. BGP Finite State Machine

# BGP State Machine



- Open confirmed state
  - –Listen for keepalives – when received move to established state
  - –If timers time out move to idle state
- Established State
  - –Send BGP updates
  - –If the hold time expires, or there is an error send a notification and move to idle state
- Active State
  - –Restart the TCP session – if successful send an open message and move to open sent sate
  - –If unsuccessful move to Idle state

# Use Templates

- Have a standard set of policies that can be applied
- Apply policy even if it preforms no function
  - –Allows you reserve the policy name
  - –Familiarize operational staff with policies
  - –Allow application of policy later without resetting the session

# Use Peer Groups

- Implemented to reduce router processing by building a single update (single RIB-OUT) for peer group
  - Operationally not so significant
- Good for organizing and managing BGP neighbors
  - Easy to apply the same policy to a class of neighbors
  - Consider separate peer groups for each class of neighbors even if the policy is currently the same
    - Greater flexibility to change policy later without moving neighbors to a different peer group which requires a BGP reset

# Other BGP Considerations

- TTL hack
  - –Default TTL on eBGP is 1
  - –Need to set mutli-hop TTL for peering loopback to loopback, or Peering with routers not directly connected
  - –An attacker four hops away can set TTL to 5 and attack the TCP session
  - –Having neighbor set TTL to 255, and discarding all packets with TTL lower than 254 prevents attack for devices that are not directly connected

# Other BGP Considerations

Route Damping

- Intended to suppress unstable routes
- Don’t use it with your customers, they pay you to carry their routes
- Don’t use it with Peers or transit providers
  - Causes more harm than good
  - A single flap can cause down stream ASes to experience multiple flaps as each network calculates new best path, and floods updates
  - A simple update can look like many updates if a race condition occurs as updates are advertised differently through different equipment, and different network configurations
  - Route Damping reduces stability, increases CPU load, and makes destinations unreachable

?