E6998-02: Internet Routing

Lecture 13
Border Gateway Protocol, Part II

John Ioannidis

AT&T Labs – Research

ji+ir@cs.columbia.edu

Announcements

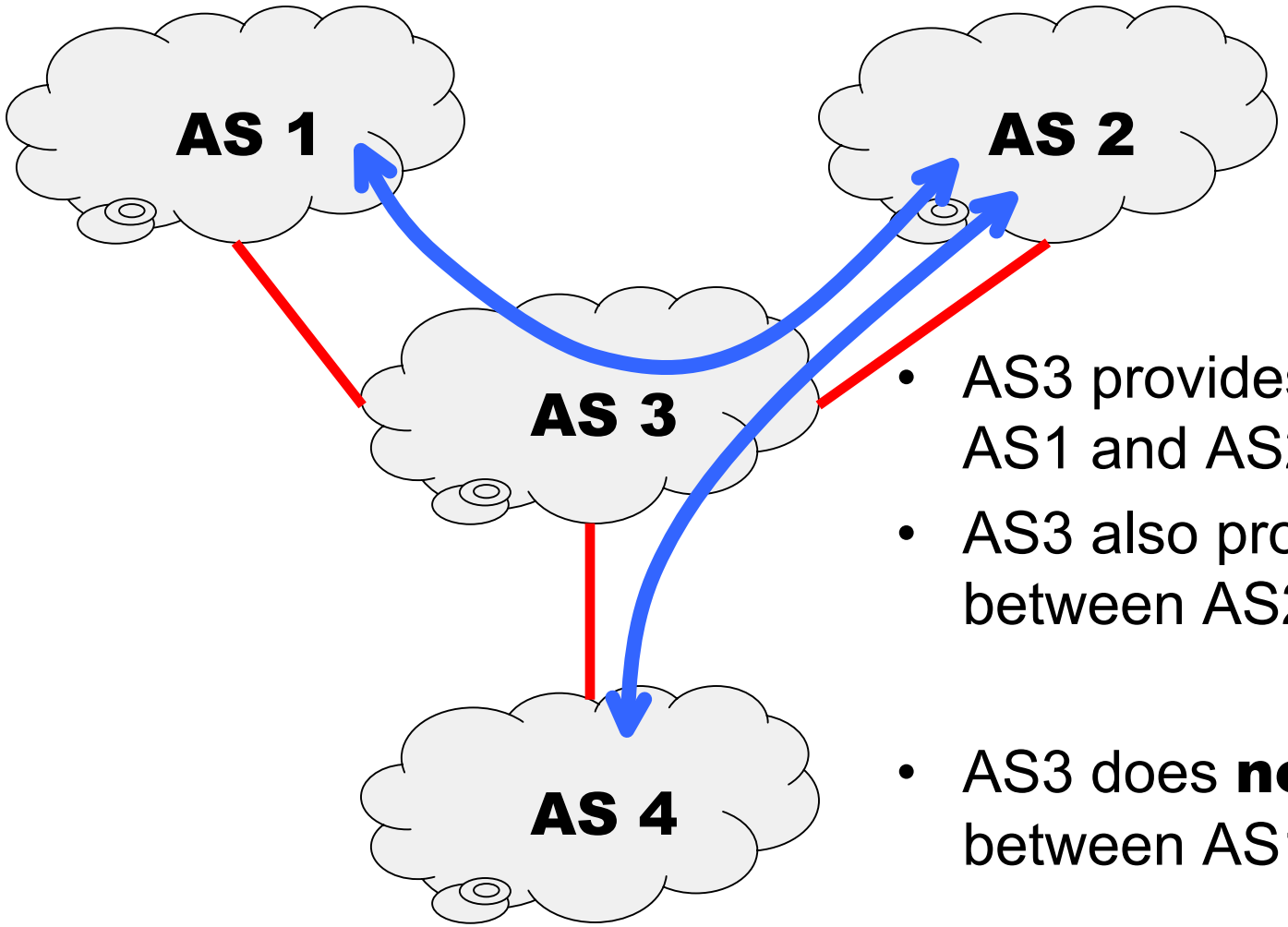
Lectures 1-13 are available.

Have you been working on your project proposal?

Still looking for a TA.

Acknowledgement: some of the slides for this lecture have been “inspired” by Tim Griffin’s BGP Tutorial.

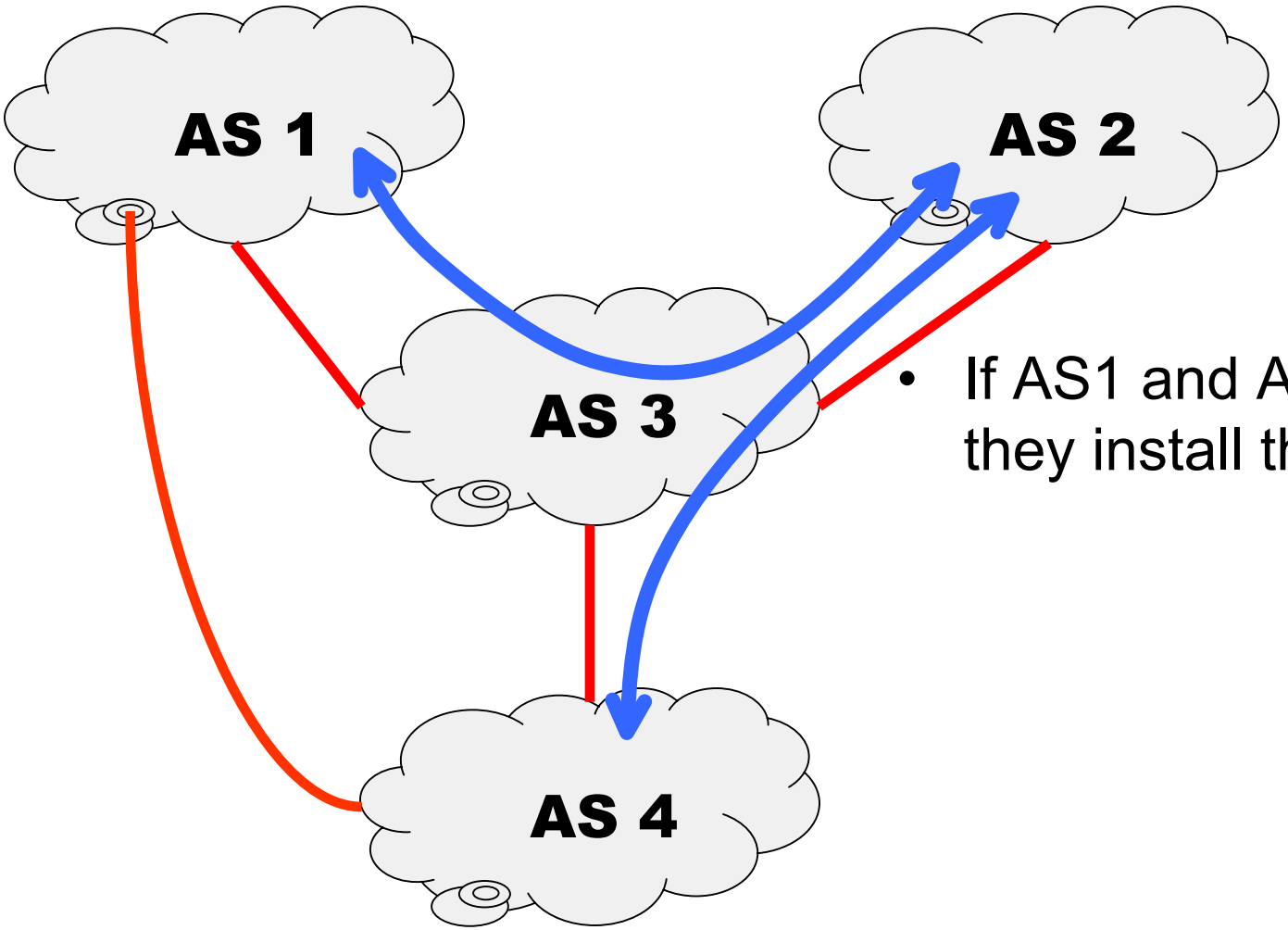
Transit *vs.* Non-transit Networks (review)



- AS3 provides transit between AS1 and AS2.
- AS3 also provides transit between AS2 and AS4.

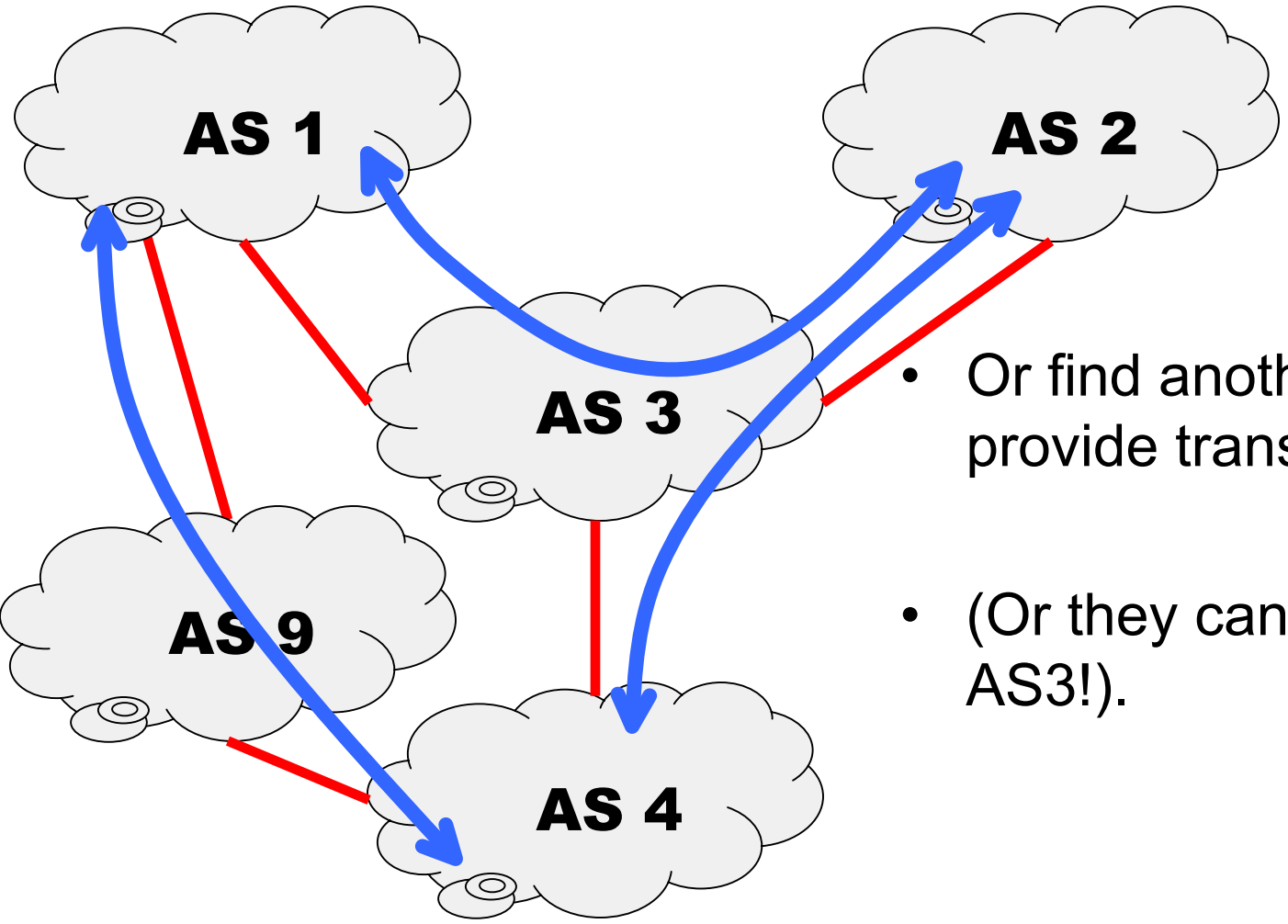
- AS3 does **not** provide transit between AS1 and AS4.

Transit *vs.* Non-transit Networks (review)



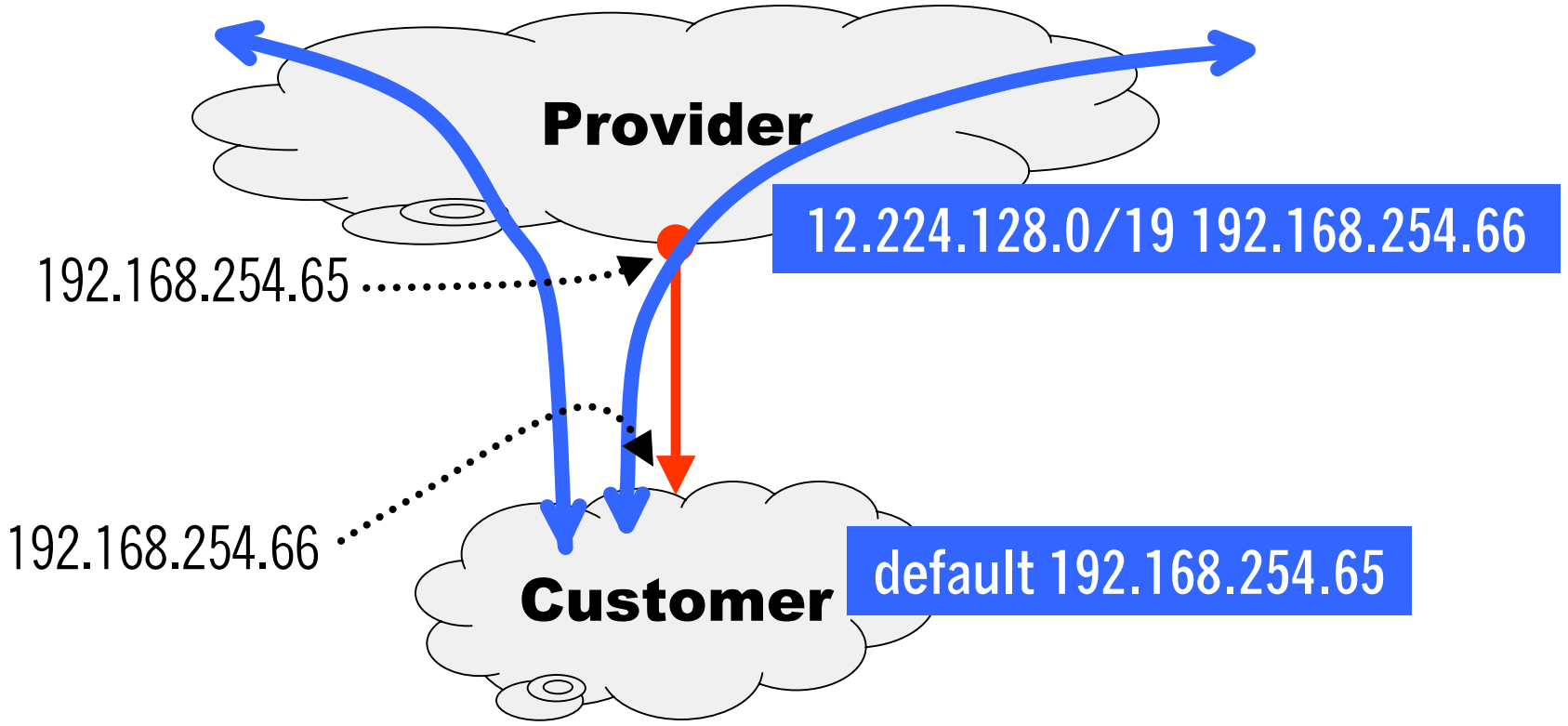
- If AS1 and AS4 need to talk, they install their own link.

Transit *vs.* Non-transit Networks (review)



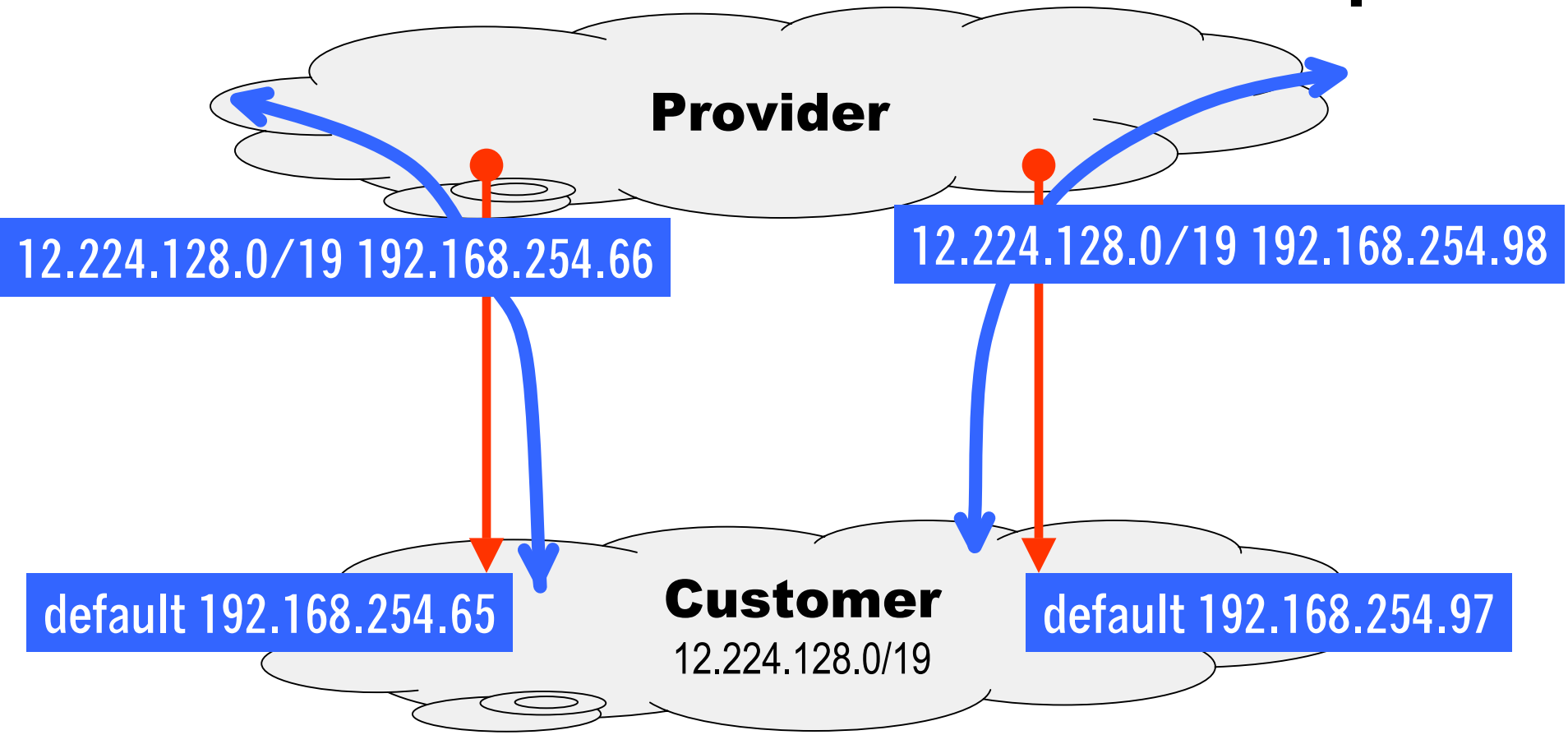
- Or find another network to provide transit traffic.
- (Or they can negotiate with AS3!).

Customer-Provider Relationship



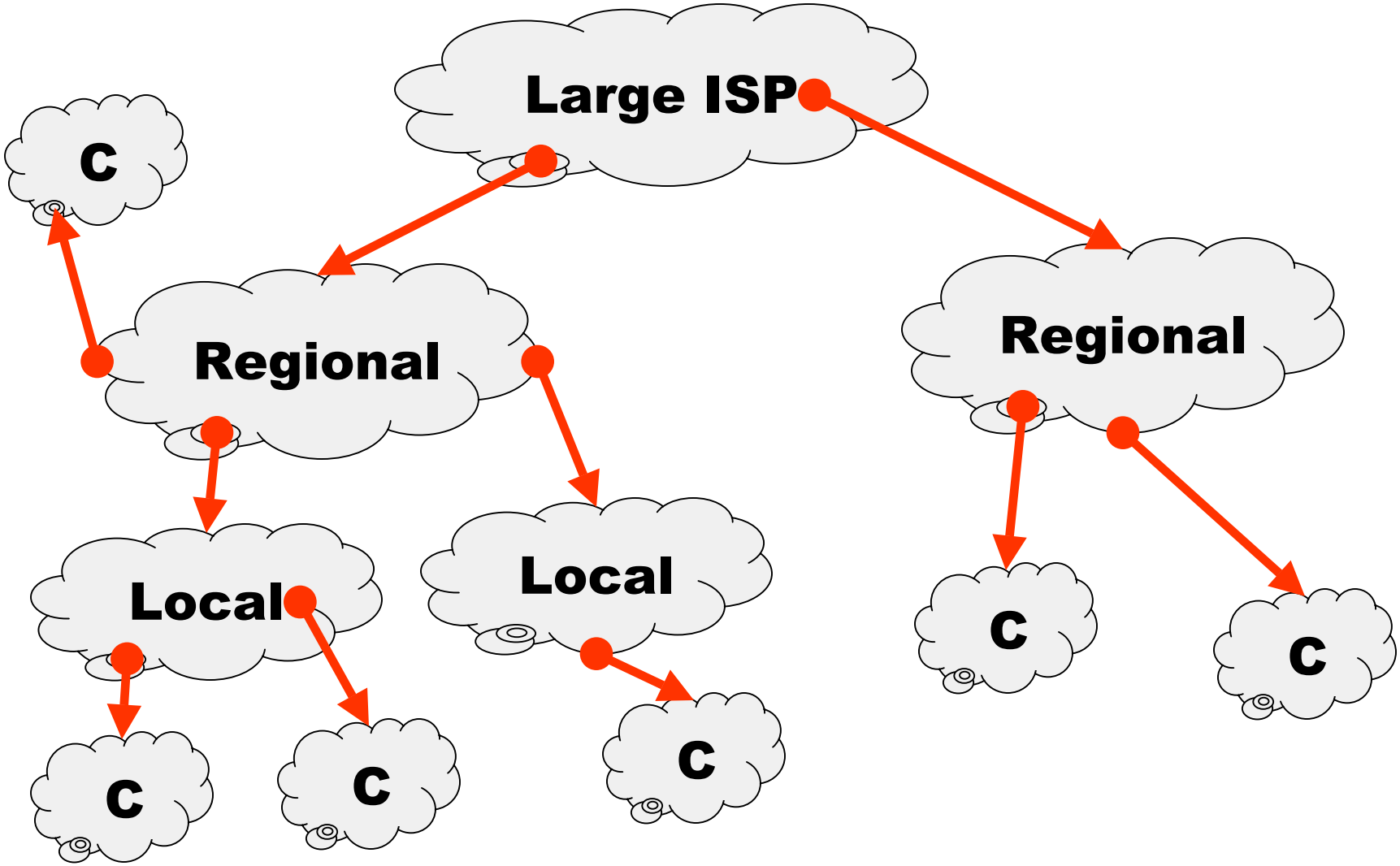
- Customer pays provider for access.
- Customer just has default route pointing to provider.
- Provider has static route pointing to customer.
- Customer does not need BGP.

Customer-Provider Relationship



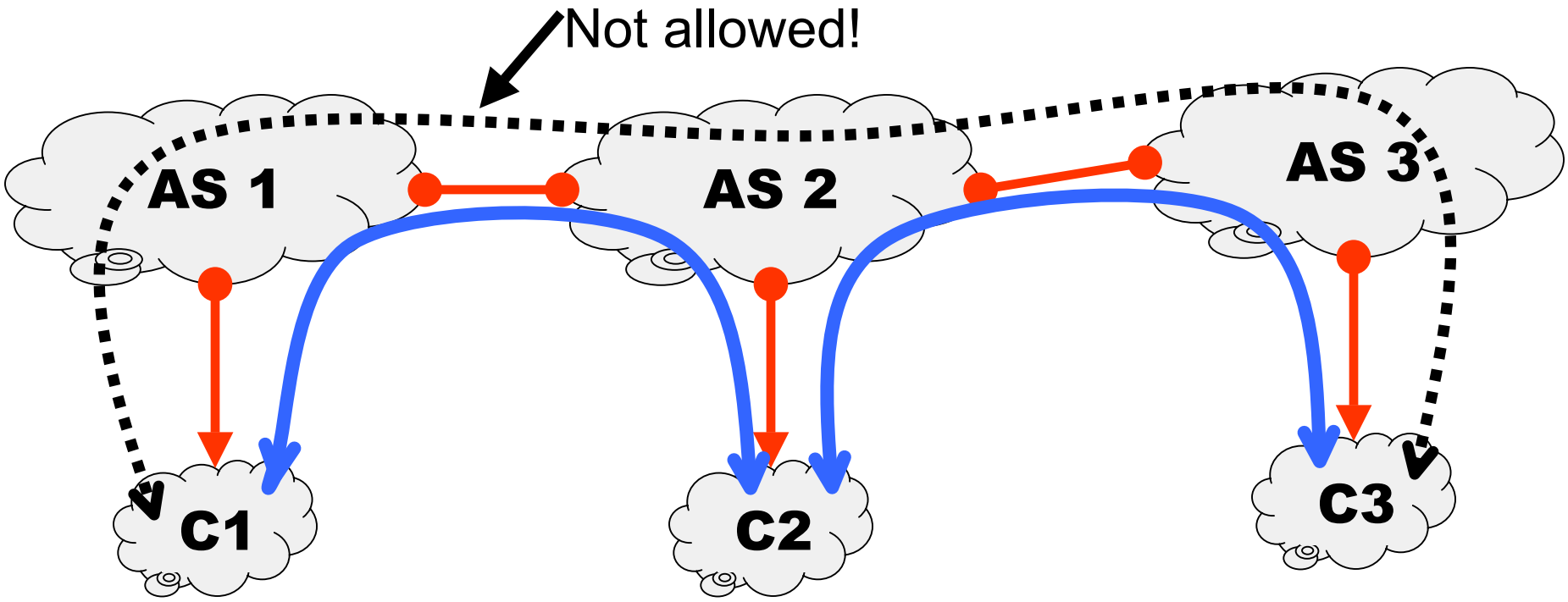
- This also works with multiple connections between Customer and Provider.
- IGP actually takes care of using closest link (how?).

Customer-Provider Hierarchy



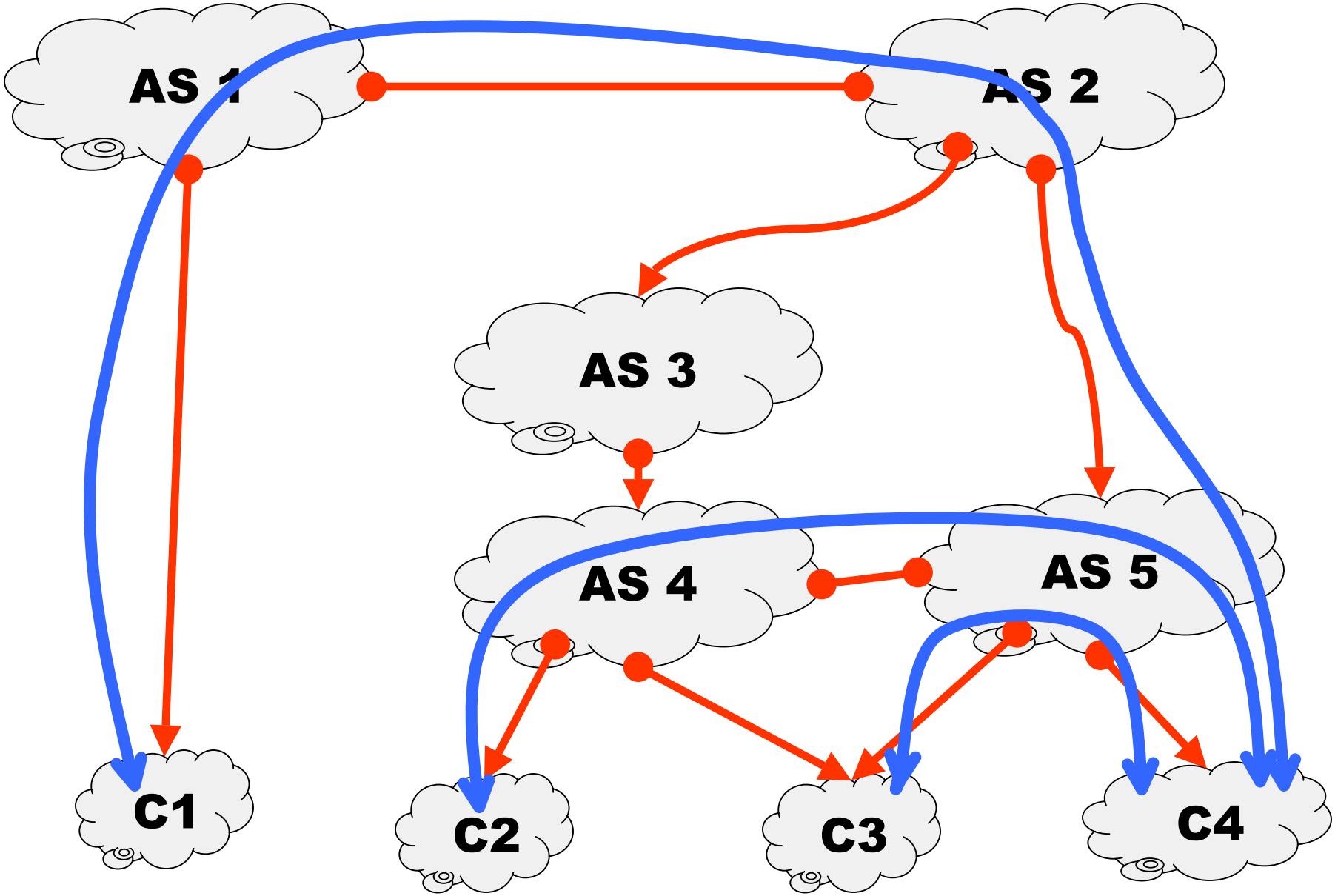
- Customer-Provider relationships can be hierarchical.
- Each network pays their *upstream* provider.

Peering



- Peers provide transit between their respective customers.
- Peers DO NOT provide transit for other peers.
  - They do if they have a customer relationship!
  - How is this enforced?

Peering is About Shortcuts



Peer or Customer?

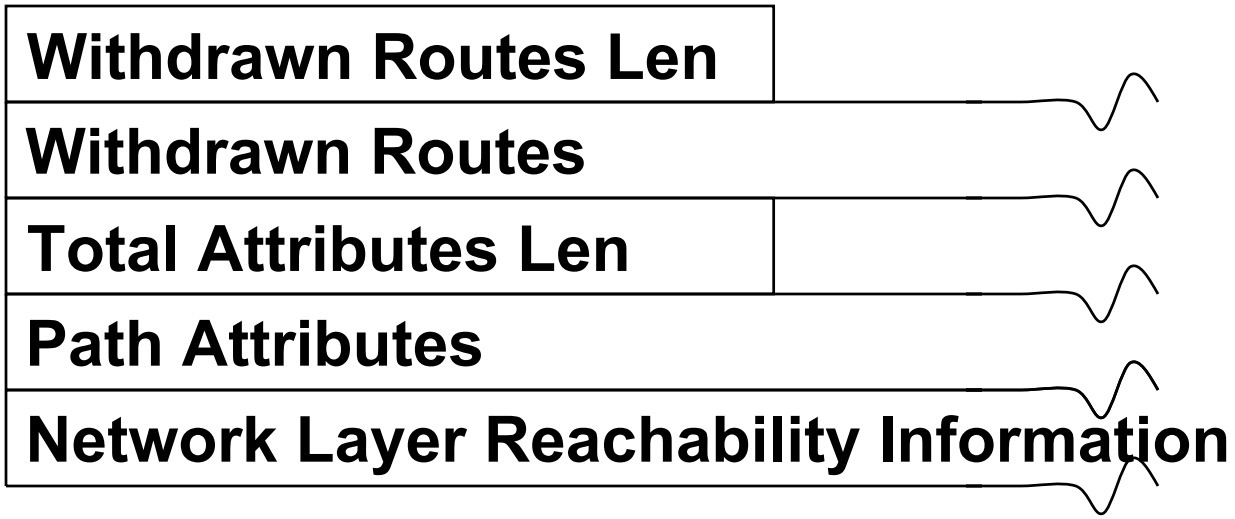
- Each provider's customers:
  - Want to "connect" to customers of other providers.
  - Provide services that others may want/need.
- Providers, in response:
  - Should pay to provide upstream service to their customers.
  - Should get paid to make their customers available.
- Peering agreements result from this contention.
  - Peering implies no exchange of money.
  - Your peers are your competitors!
  - Peering agreements are often confidential.
    - And subject to periodic negotiation.

Peer or Customer? Cont'd

- Similar-size providers peer.
  - Tier-1, Tier-2, etc. providers.
- Customers who exchange a lot of traffic may also peer!
- A customer may have multiple upstream providers.
  - Multihoming.
- "Back-doors" may be installed for special customers.
  - Columbia is not Verizon's customer.
  - But lots of Verizon DSL customers want to connect to Columbia.
  - Verizon may install a private link to Columbia just for their DSL customers.

Back to BGP

- Path Attributes in particular.



Path Attributes

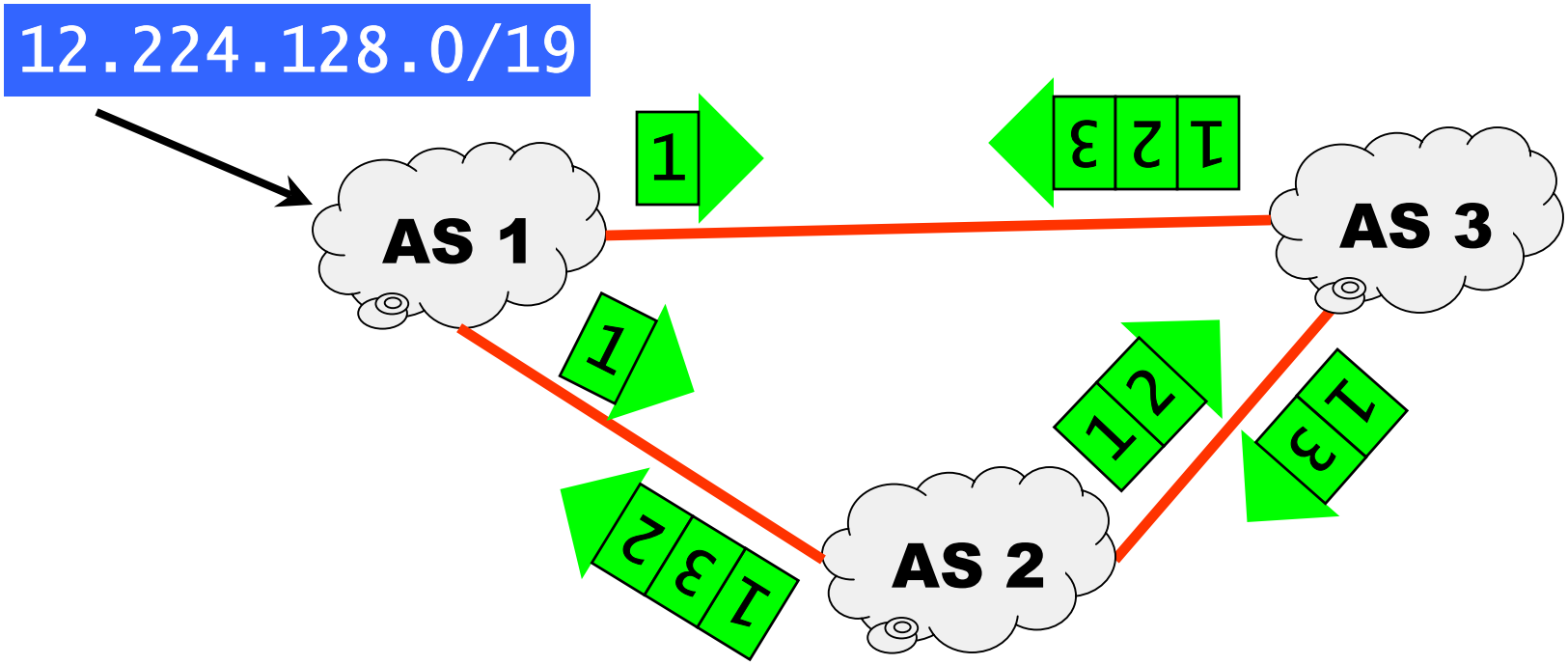
1	ORIGIN	RFC 1771
2	AS_PATH	RFC 1771
3	NEXT_HOP	RFC 1771
4	MULTI_EXIT_DISCRIMINATOR	RFC 1771
5	LOCAL_PREF	RFC 1771
6	ATOMIC_AGGREGATE	RFC 1771
7	AGGREGATOR	RFC 1771
8	COMMUNITY	RFC 1997
9	ORIGINATOR_ID	RFC 2796
10	CLUSTER_LIST	RFC 2796
11	DPA	deprecated
12	ADVERTISER	RFC 1863
13	RCID_PATH/CLUSTER_ID	RFC 1863
14	MP_REACH_NLRI	RFC 2283
15	MP_UNREACH_NLRI	RFC 2283
16	EXTENDED COMMUNITIES	draft-ietf-idr-bgp-ext-communities-05.txt
17	NEW_AS_PATH	draft-ietf-idr-as4bytes-05.txt
18	NEW_AGGREGATOR	draft-ietf-idr-as4bytes-05.txt
...	...	
255	Reserved for development	

ORIGIN

- Well-known, Mandatory. Type=1
- Shows how a prefix was learned.
  - Prefixes are *injected* into BGP
- Length=1
- Value:
  - IGP (=1): Prefix was learned from an IGP.
  - EGP (=2): Prefix was learned from the EGP (BGP).
  - INCOMPLETE (=3): Prefix was learned some other way.
    - Static routes/directly connected networks.

AS_PATH

- ASNs through which the announcement for these prefixes has passed.
- First ASN in the AS_PATH: Origin AS.
- Each AS appends its own ASN before passing on the update.

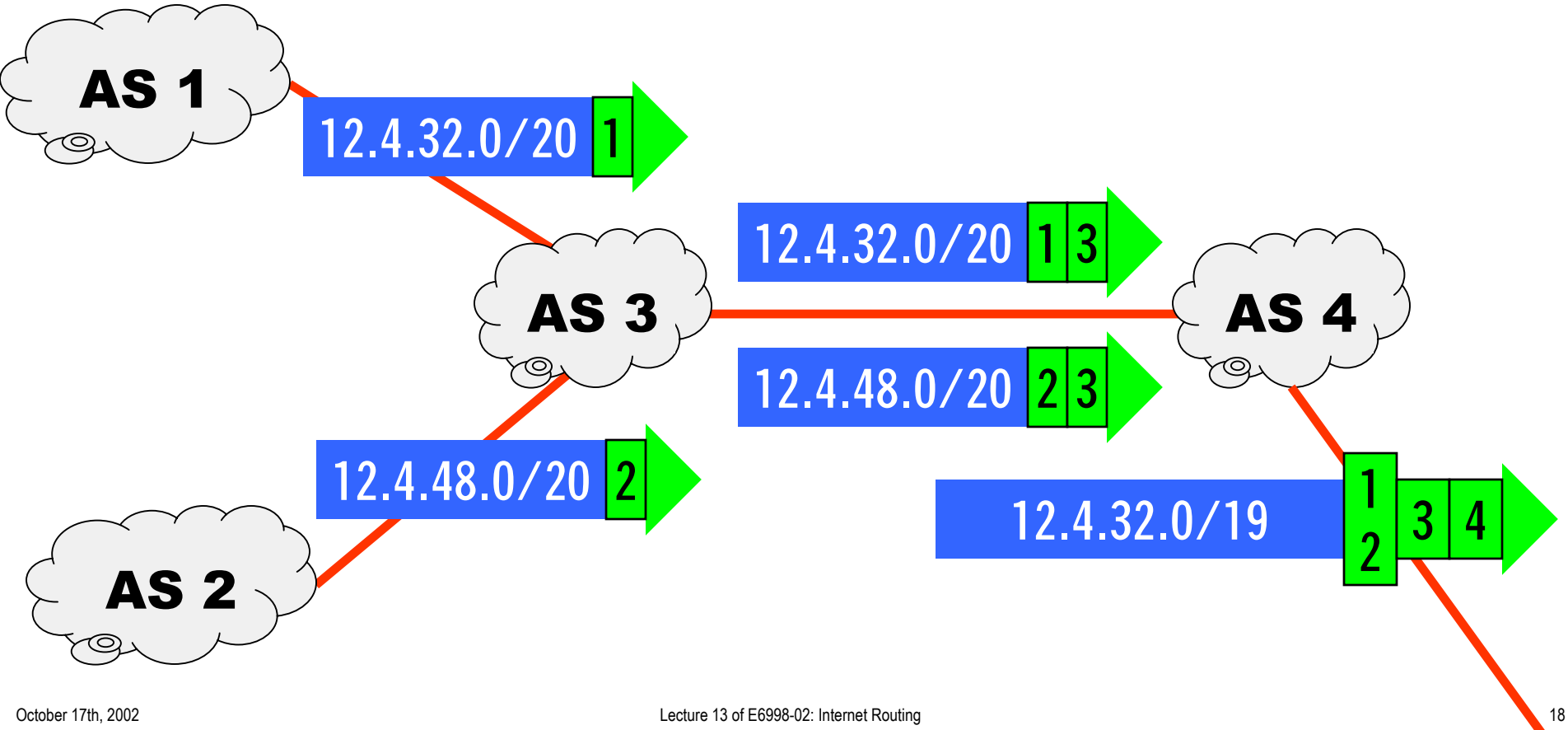


AS_PATH Cont'd

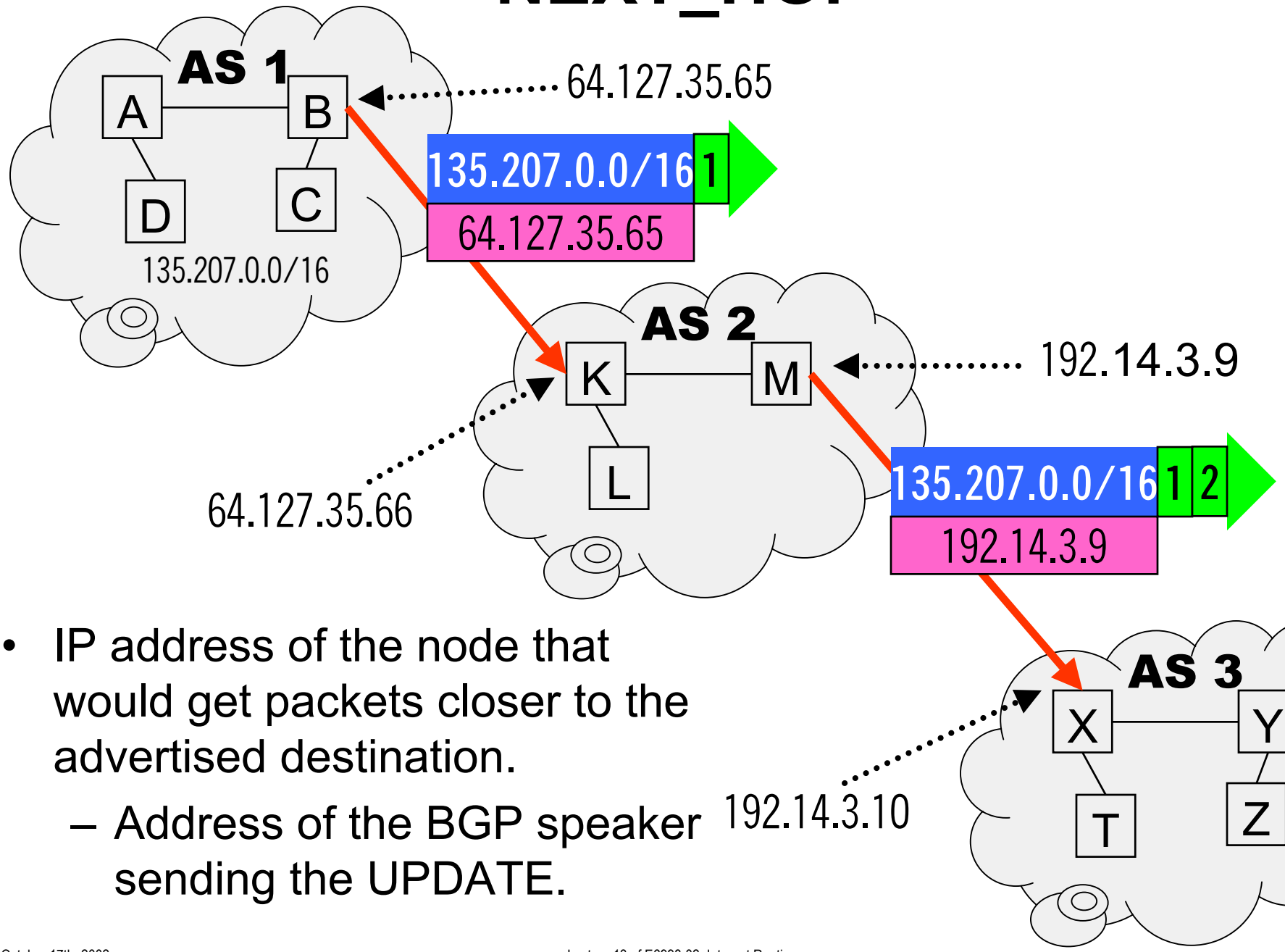
- Well-known, Mandatory. Type=2
- Encoded as sequence of AS_PATH segments.
- Each segment is encoded as:
  - Path Segment Type:
    - AS_SET (1): unordered set of ASNs.
    - AS_SEQUENCE (2): ordered set of ASNs.
  - Path Segment Length: 1 octet, #of ASNs in segment.
  - Path Segment Value: 2*PSL octets, list of ASNs.
- New ASNs are actually **prepended** in the packet.
- If leading segment is AS_SET, a new AS_SEQUENCE is prepended with the ASN as its sole member.
- If leading segment is AS_SEQUENCE, the ASN is just prepended to the sequence.

AS_PATH Cont'd

- Most AS_PATHs are encoded as a single AS_SEQUENCE.
- If a router needs to aggregate, it has to use AS_SET.
- Not common, since most routers aggregate prefixes from their own AS.



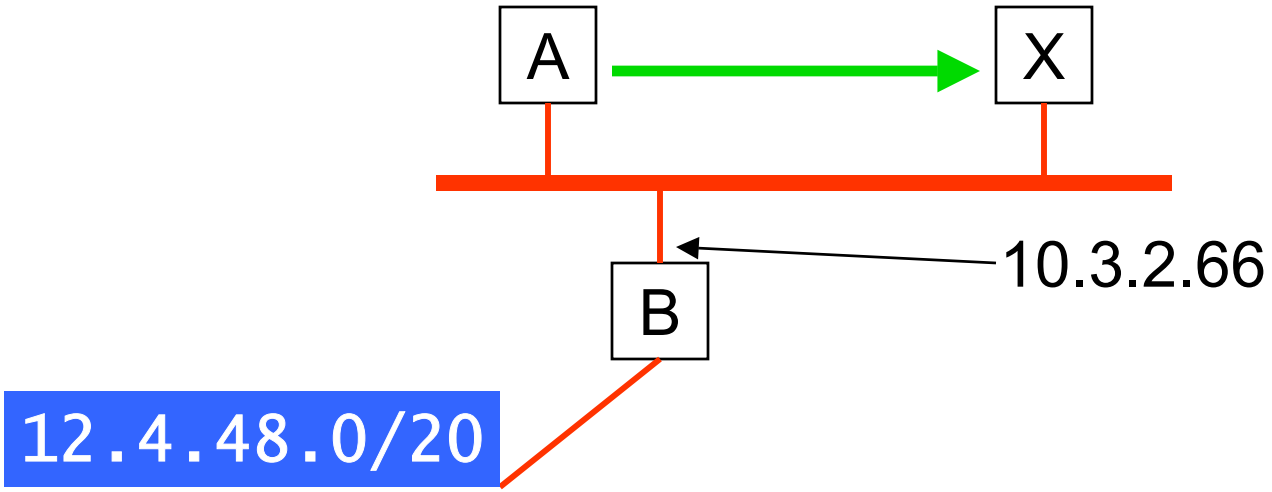
NEXT_HOP



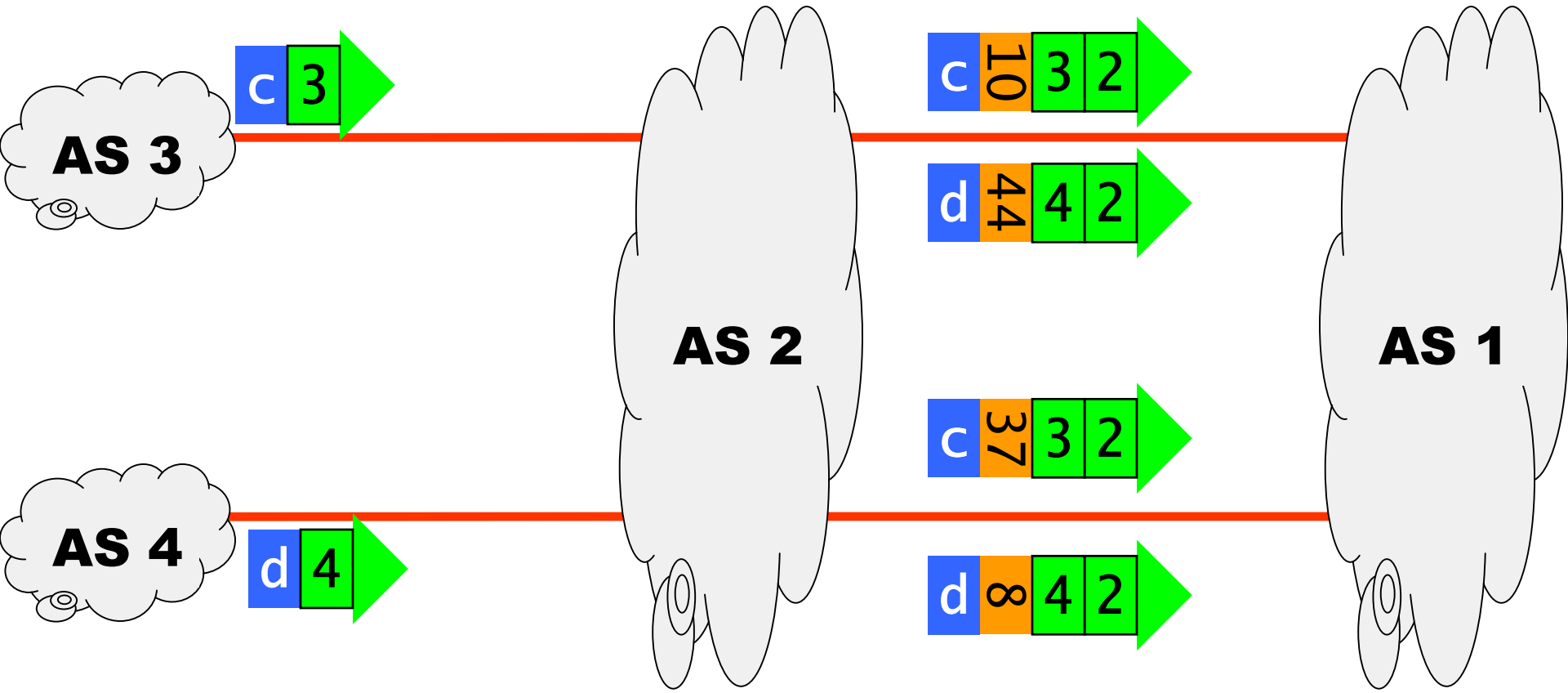
- IP address of the node that would get packets closer to the advertised destination.
  - Address of the BGP speaker sending the UPDATE.

NEXT_HOP cont'd

- Well-known, Mandatory. Type=3
- Encoded as the 4-octet address right after the Type Code.
- IP address of the node that would get packets closer to the advertised destination.
  - Address of the BGP speaker sending the UPDATE.
- Exception: A (BGP speaker) sends X (BGP speaker) an UPDATE indicating B (10.3.2.66 interface) (not a BGP speaker) is the router for 12.4.48.0/20.



MULTI_EXIT_DISCRIMINATOR (MED)



- AS2 includes MED to the updates it sends to AS1.
- AS3 and AS4 are advertised over both links, of course.
- AS1 can now make a better choice about sending packets to AS3 and AS4.

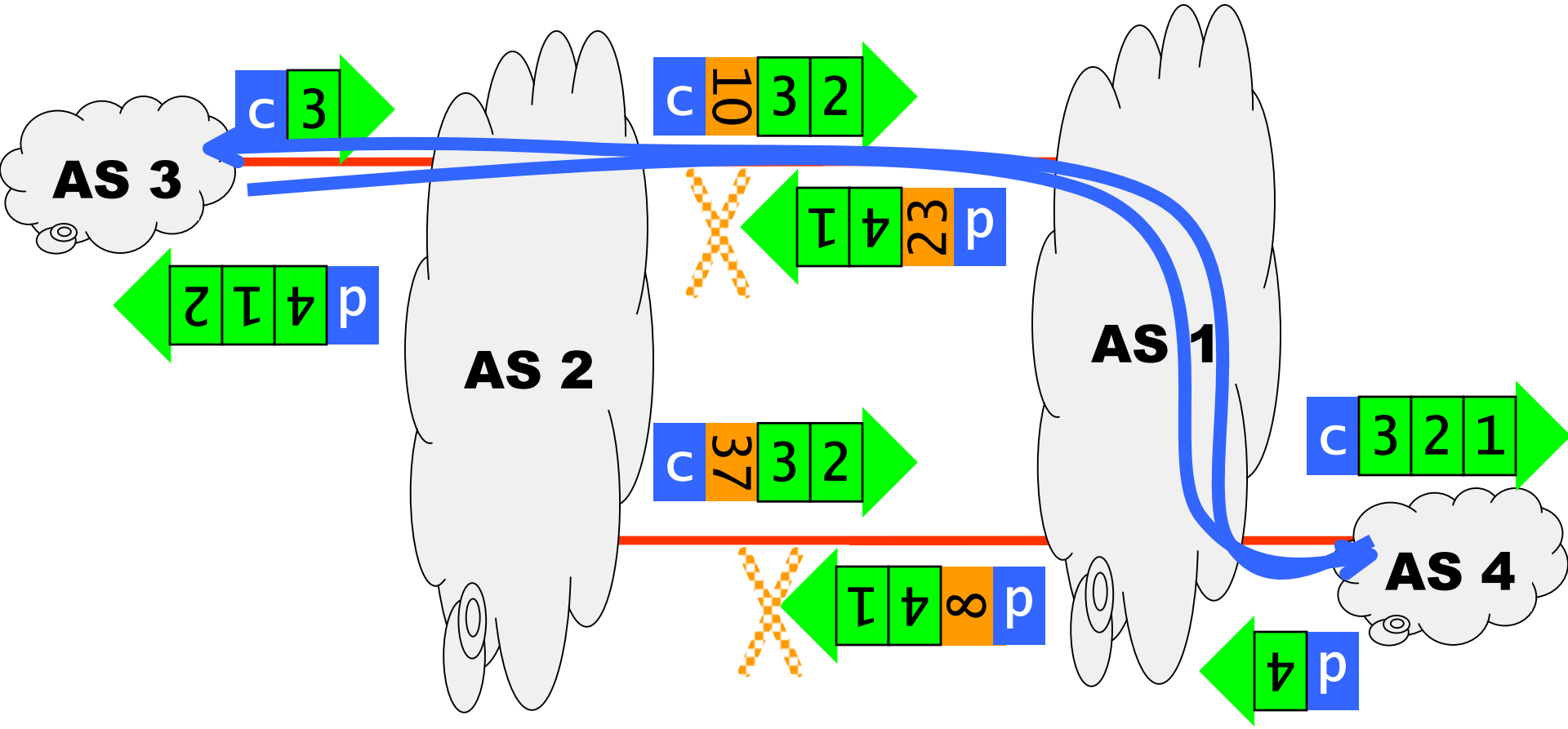
MED Cont'd

- One AS sets MED, but another uses it.
  - MED only used in Customer/Provider relationships (why?).
- Peers usually ignore received MEDs (why?).

- Optional, Non-transitive (why?). Type=4
- Length is always 4, encoding is unsigned integer.

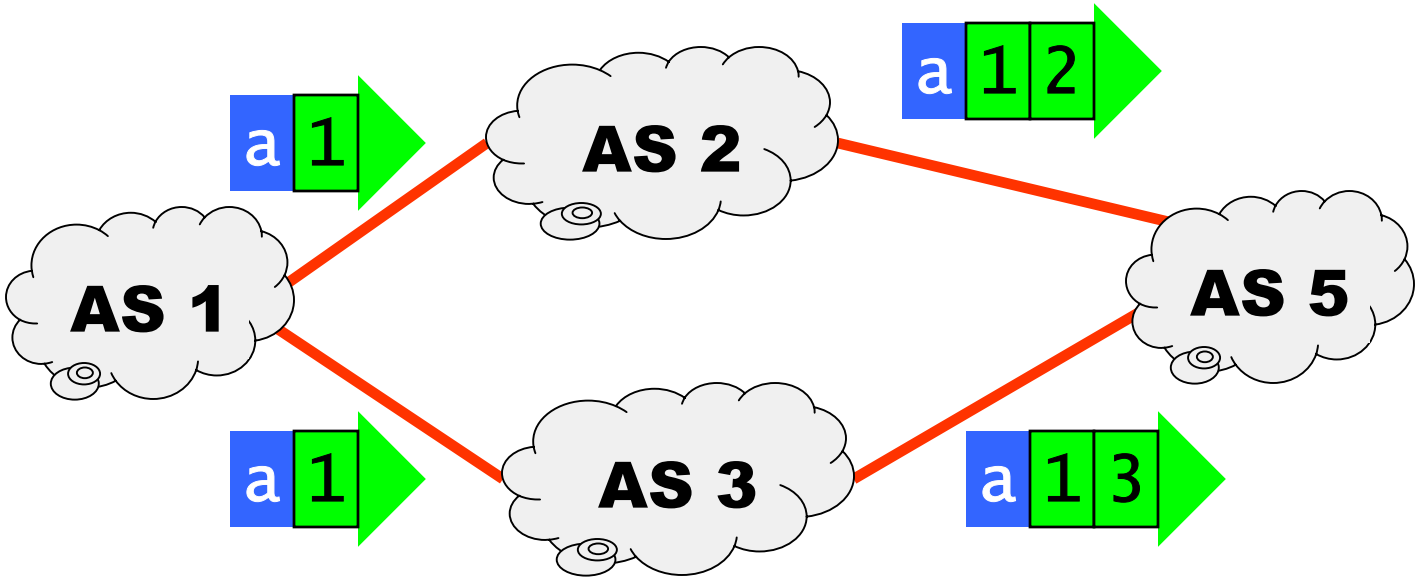
- MED is usually the IGP metric for the advertised prefix.
- MED comparison only makes sense when received from the same AS.

MED Cont'd



- MED can be (ab)used to get one ISP to carry more traffic.
- Traffic from AS3 to AS4 goes to closest link.
- Traffic from AS4 to AS3 obeys MED.

LOCAL_PREF



- How does AS5 decide how to send traffic to prefix a?
- MED doesn't help here.
  - Only one link between AS pairs.
  - AS5 may want to set its own policy about this.
- AS5 uses the LOCAL_PREF attribute on routes it receives.
- LOCAL_PREF is the first attribute used in route selection.

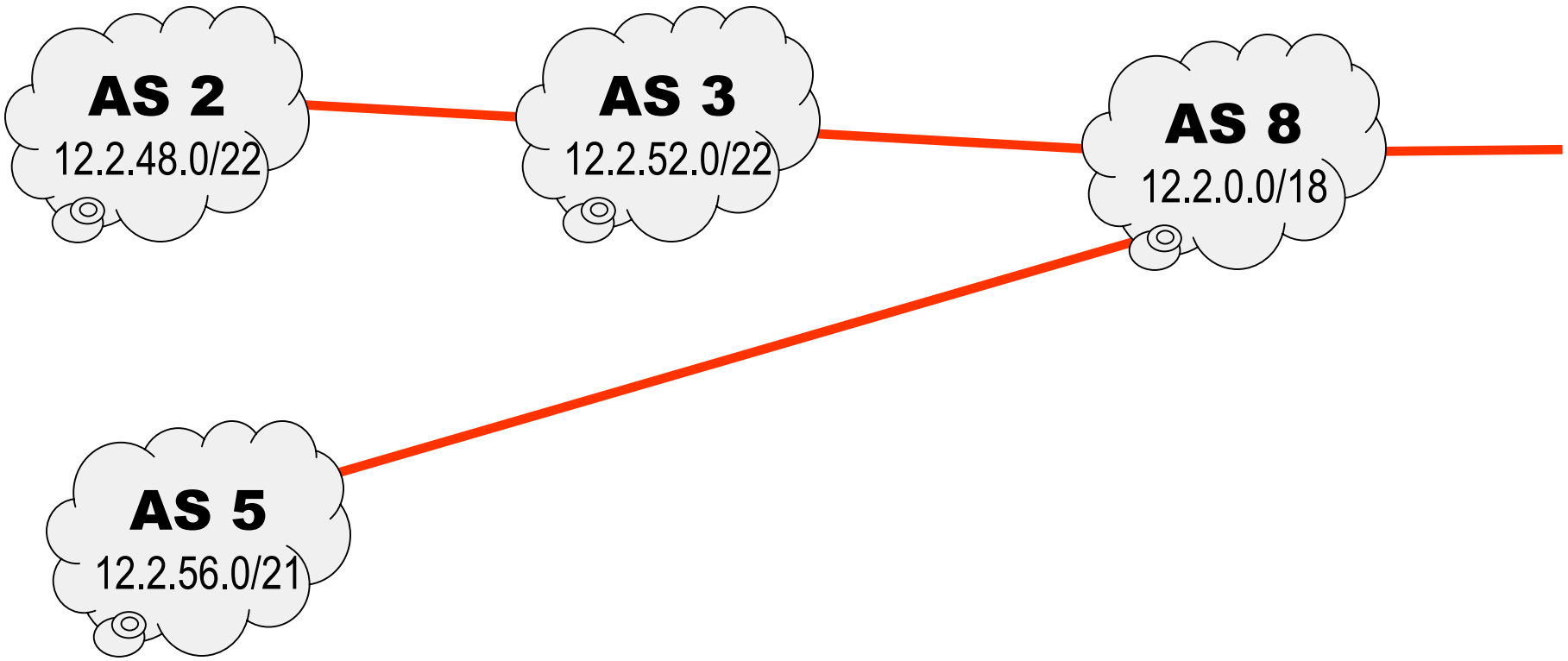
LOCAL_PREF Cont'd

- LOCAL_PREF is computed locally when route received from E-BGP, IGP, or statically assigned.
  - Part of the interface configuration.
  - Stored in the Adj-RIB-In.
- LOCAL_PREF is carried in I-BGP.
  - Don't worry about this right now!

- Well-known, Discretionary. Type=5
- Length is always 4.
- Encoding is unsigned integer.

Route Aggregation

- AS2 and AS3 can be aggregated into 12.2.48.0/21.
- AS8's space covers that of AS2, AS3, and AS5.
- What should AS8 advertise upstream?

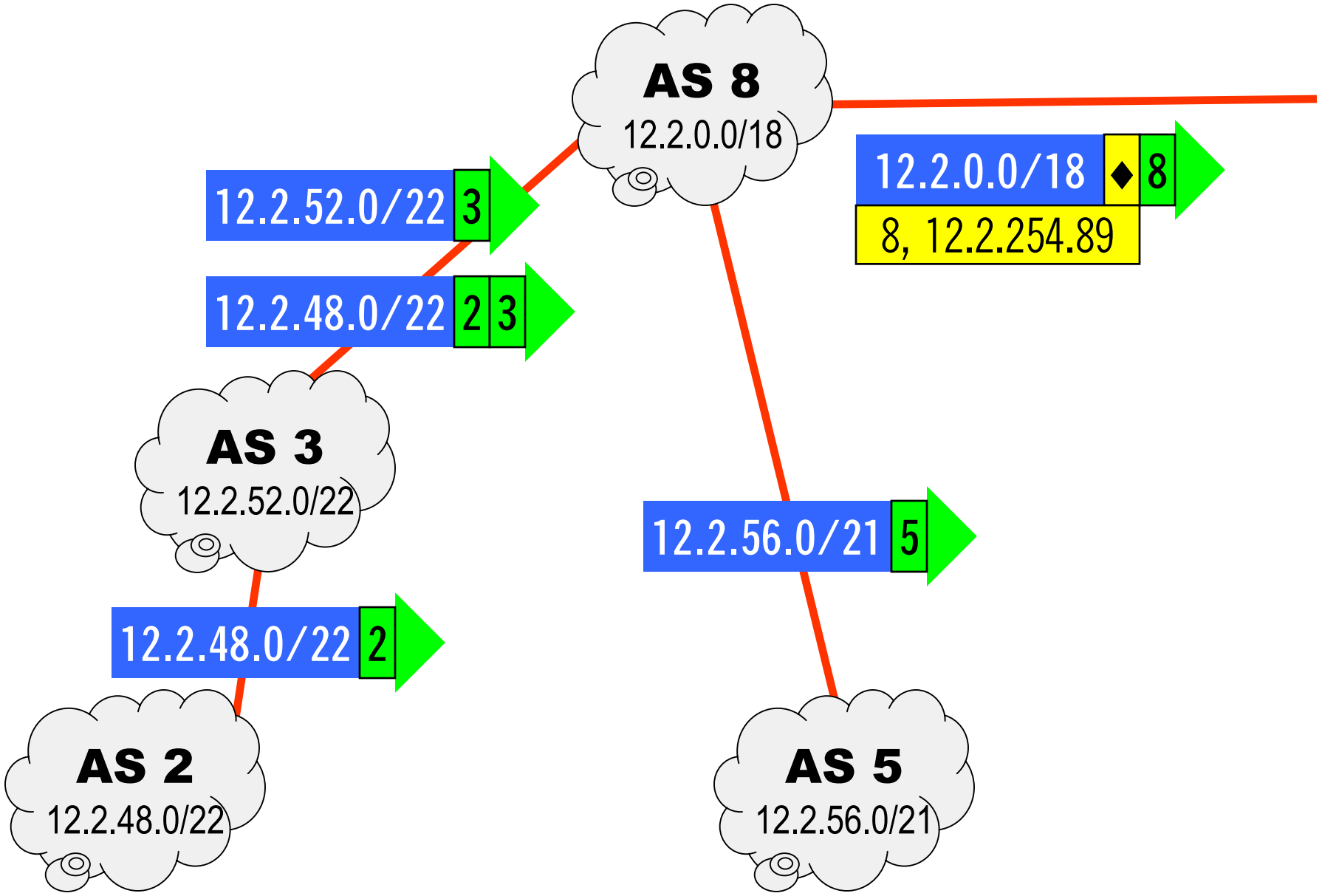


Route Aggregation, Cont'd

- AS8 could advertise:
  - Nothing, or some subset of the routes (subj. to policy).
  - All four routes.
  - Advertise just its own (less-specific) route.
    - 12.2.0.0/18 (AS8)
  - De-aggregate its own prefix and advertise more-specifics:
    - 12.2.0.0/19 (AS8)
    - 12.2.32.0/20 (AS8)
    - 12.2.48.0/22 (AS2, AS3, AS8)
    - 12.2.52.0/22 (AS3, AS8)
    - 12.2.56.0/21 (AS5, AS8)
- Aggregation saves space but destroys information.

ATOMIC_AGGREGATE & AGGREGATOR

- If a BGP speaker aggregates routes.
  - AS_PATH information is lost.
- Following routers must be alerted.
  - So they don't de-aggregate the advertised prefix.
- The ATOMIC_AGGREGATE attribute provides that feature.
  - Well-known, Discretionary. Type=6.
  - Zero length (just a flag).
  - Must remain attached.
- AGGREGATOR attribute:
  - Indicates which AS and router performed the aggregation.
  - Optional, transitive. Type=7.
  - Length is always 6.
  - 2-byte ASN, 4-byte IP address of aggregator.


AS 8
12.2.0.0/18
12.2.0.0/18 ◆ 8
8, 12.2.254.89
12.2.52.0/22 3
12.2.48.0/22 2 3
AS 3
12.2.52.0/22
12.2.48.0/22 2
12.2.56.0/21 5
AS 2
12.2.48.0/22
AS 5
12.2.56.0/21

COMMUNITY

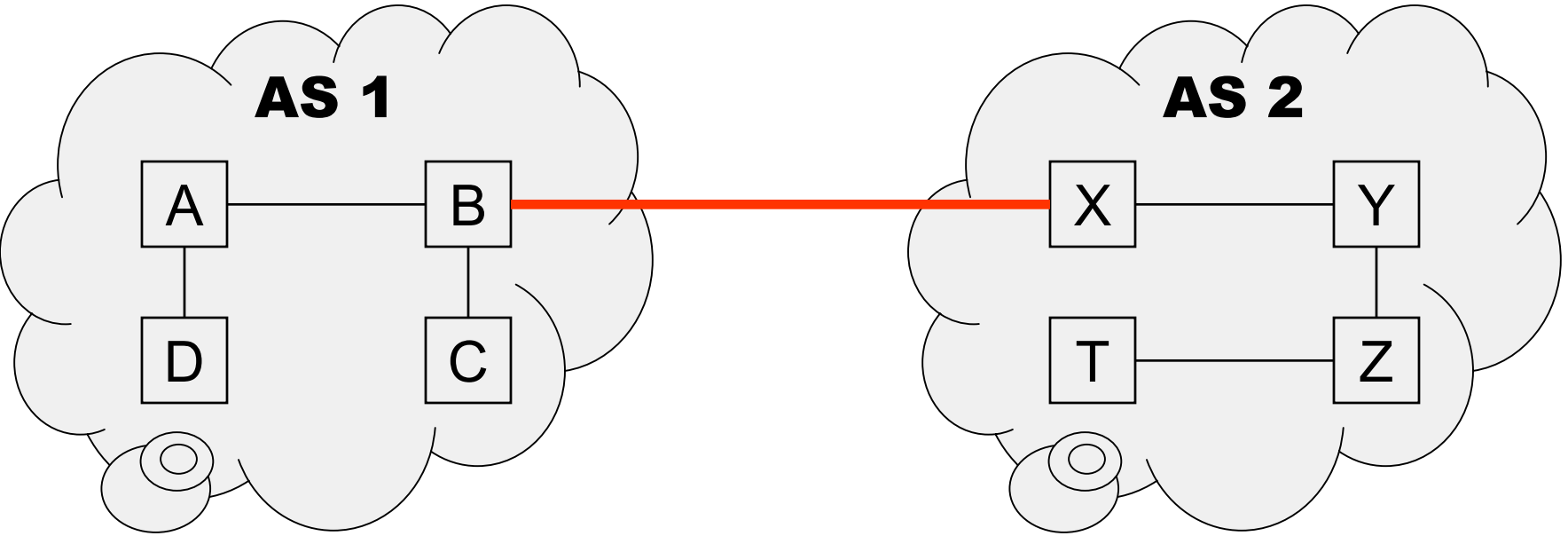
- Specified in RFC 1997.
- Encodes arbitrary properties.
  - E.g., all of customer's routes get a specific COMMUNITY.
- Much of the policy is specified using communities.

- Optional, Transitive. Type=8
- Four bytes: (*e.g.,* 7018:100)
  - 2 bytes ASN (by convention).
  - 2 bytes administratively defined (no predefined meaning).

- We'll talk about this in the next lecture.

Learning External Prefixes

- So far, BGP has been presented as a pure EGP.
  - A protocol that runs between ASs.



- How do A, C and D learn about AS2's routes?
  - Ditto for Y, Z, T about AS1's routes?
- I.E., how are prefixes learned by an ASBR distributed inside the AS?

Learning External Prefixes, cont'd

- Inject into the IGP (using AS-External LSAs).
- Small networks can do this.
  - Default route + a few external routes.
- Does not work for large ISPs.
  - They carry a full routing table (100K-400K routes!).
- Would lose policy information.
  - No way to carry attributes.
- IGPs don't scale well.
  - Computational complexity.
  - Memory requirements.
  - Additional traffic.
    - Fragmented LSAs.
- Clearly need a different way!

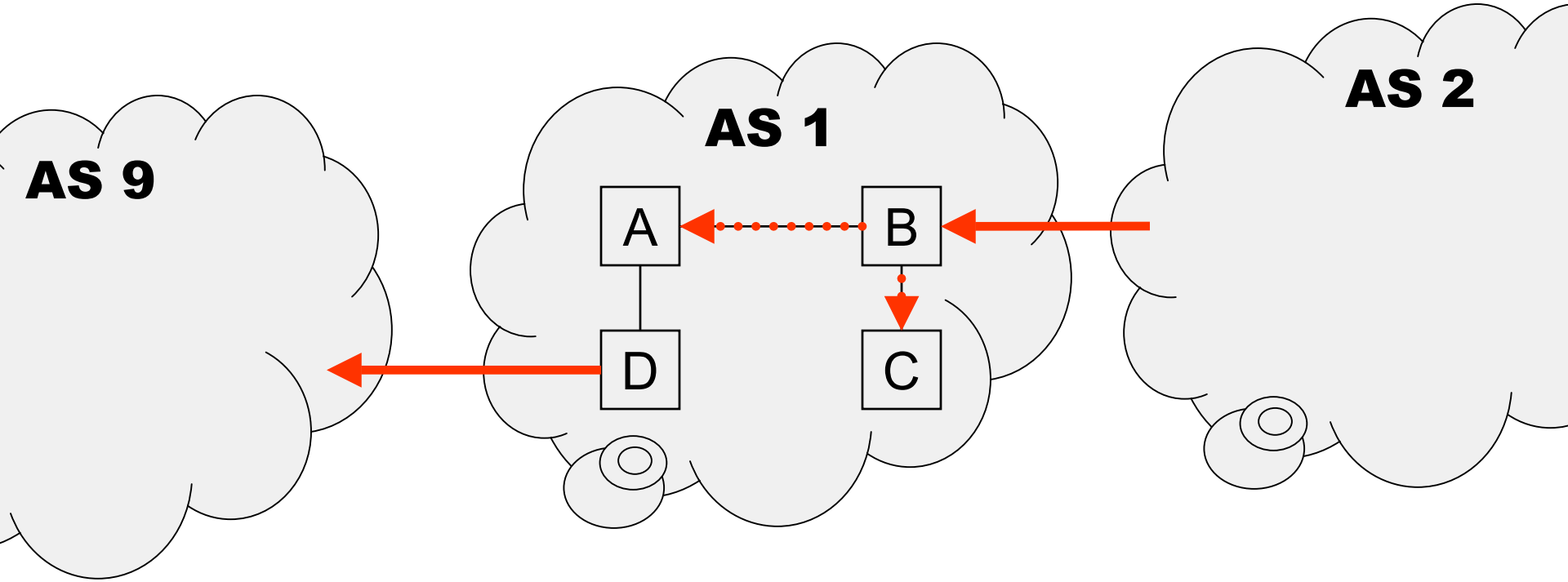
E-BGP and I-BGP

- The solution is called *Internal-BGP (I-BGP)*.
  - As opposed to *External-BGP (E-BGP)*.
- E-BGP is used between ASs.
- I-BGP is used **within** an AS.
  - Is used to distribute routes learned with E-BGP.
- E-BGP and I-BGP are the same protocol.
  - Same messages, attributes, state machine, etc.
- But: different rules about route redistribution:

		Redistribute to	
		I-BGP	E-BGP
Learned from	I-BGP	no	yes
	E-BGP	yes	(yes)

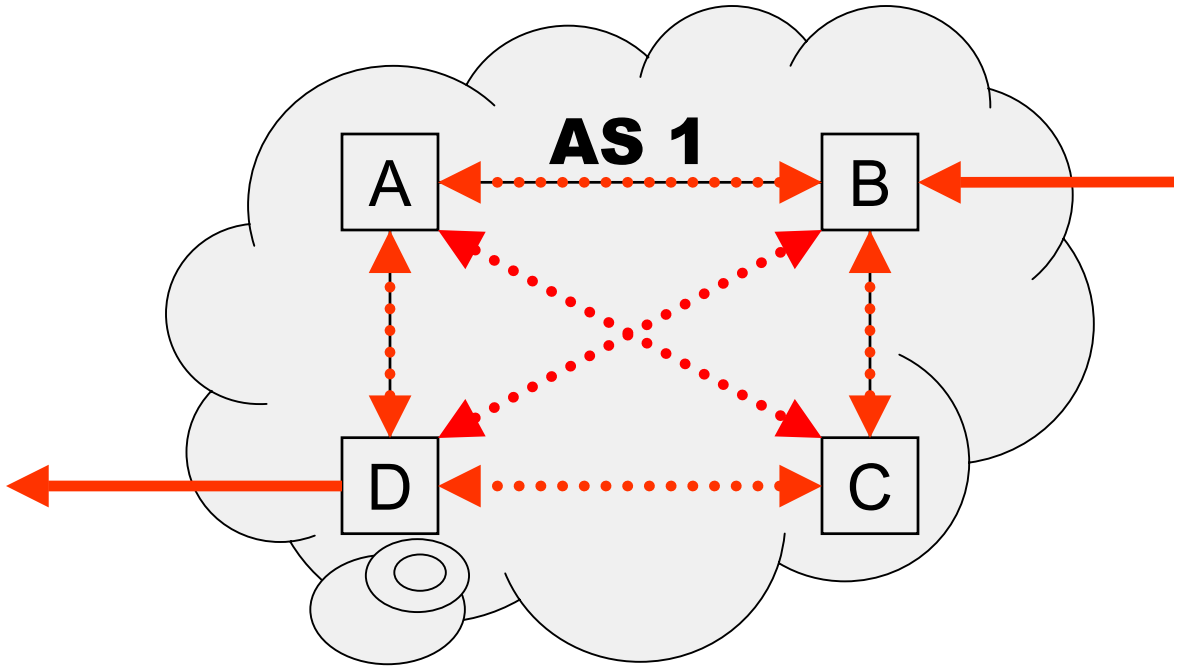
I-BGP Route Redistribution

- How does D learn routes acquired by B?
  - Since A can't redistribute routes learned over I-BGP?
- If D also had an external connection, how would it redistribute routes learned from other ASs?



I-BGP Route Redistribution, cont'd

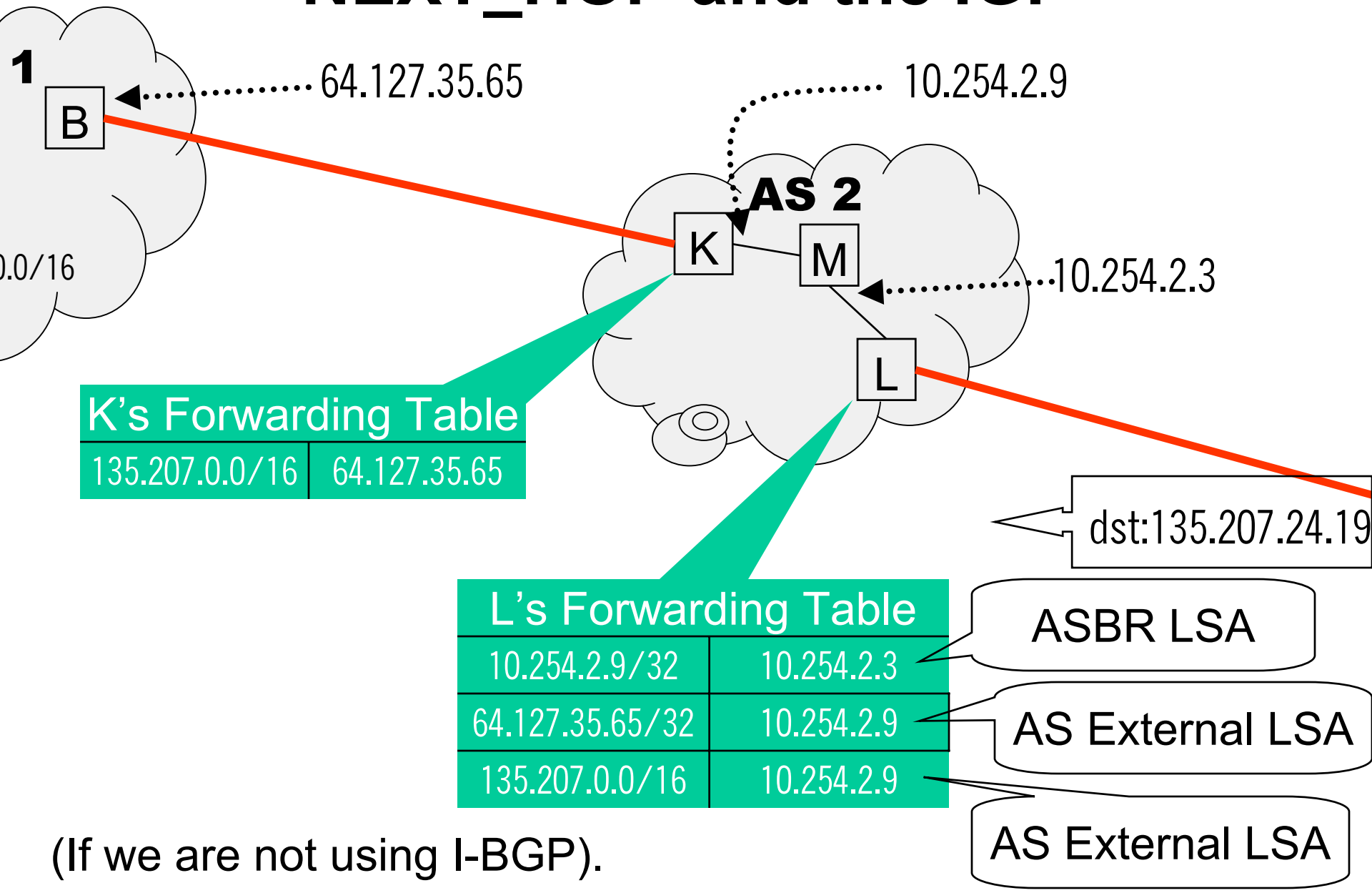
- Remember: BGP is a **routed** protocol.
- Routes between routers already exist.
  - Carried by the IGP.
- I-BGP sessions can be formed between non-adjacent routers.
- I-BGP sessions must form a full mesh:



I-BGP, cont'd

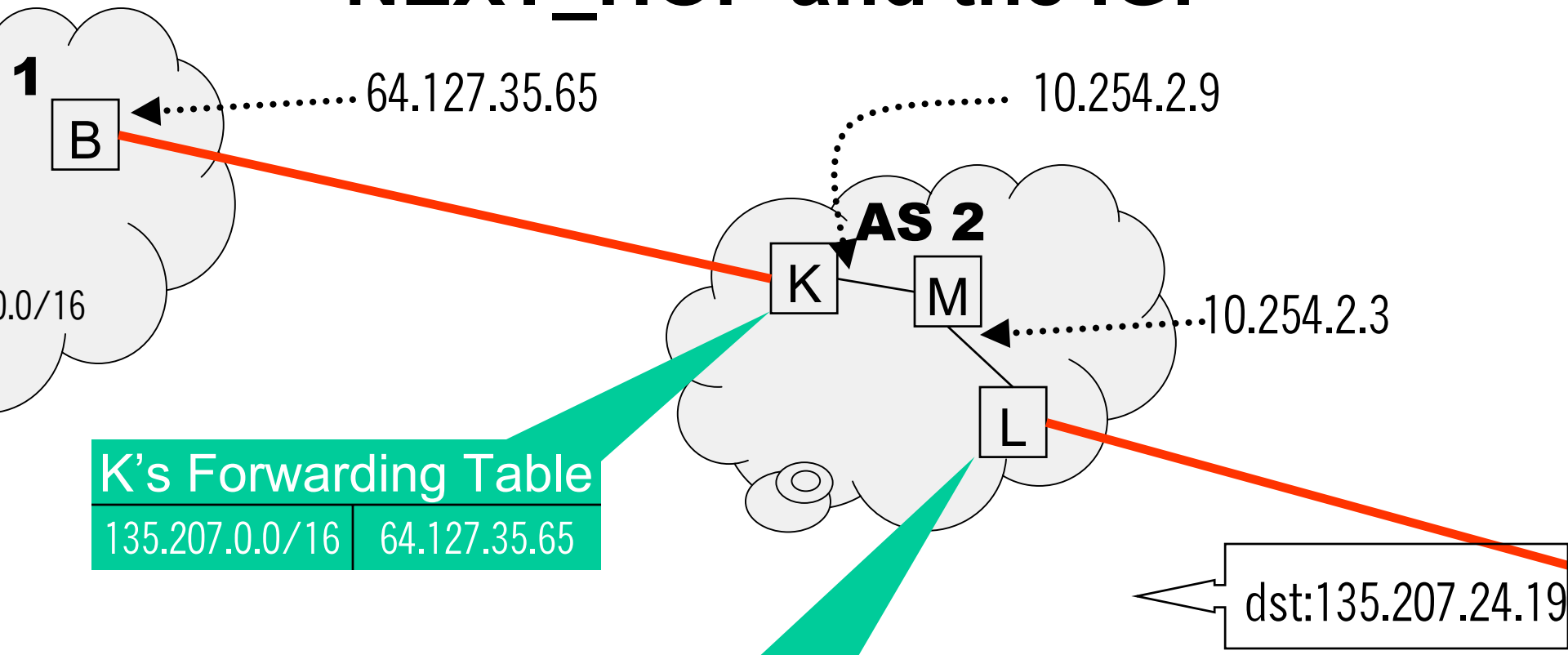
- Full mesh.
- Independent of actual links between (internal) routers.
- TCP src/dst of I-BGP session must be a loopback address.
  - Routing to the router must be independent of interfaces going up/down.
- Full mesh is necessary to prevent loops.
  - AS_PATH is used to detect loops in E-BGP.
  - ASN appended to AS_PATH only when route is advertised to E-BGP peer.
- I-BGP is **NOT** an IGP.
  - Nor can be used as one.

NEXT_HOP and the IGP



(If we are not using I-BGP).

NEXT_HOP and the IGP



K's Forwarding Table	
135.207.0.0/16	64.127.35.65

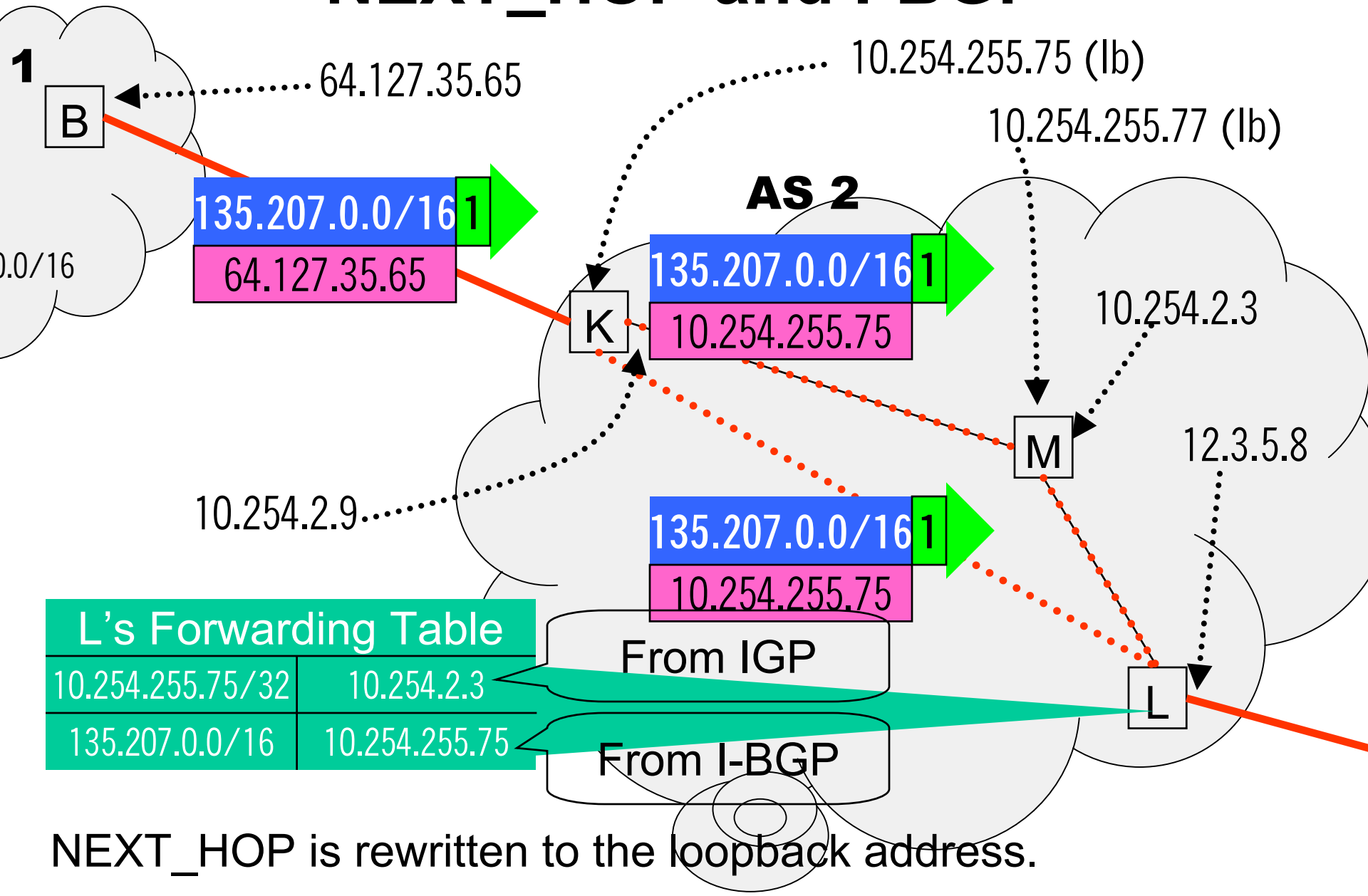
L's Forwarding Table	
10.254.2.9/32	10.254.2.3
64.127.35.65/32	10.254.2.3
135.207.0.0/16	10.254.2.3

Rewritten

Rewritten

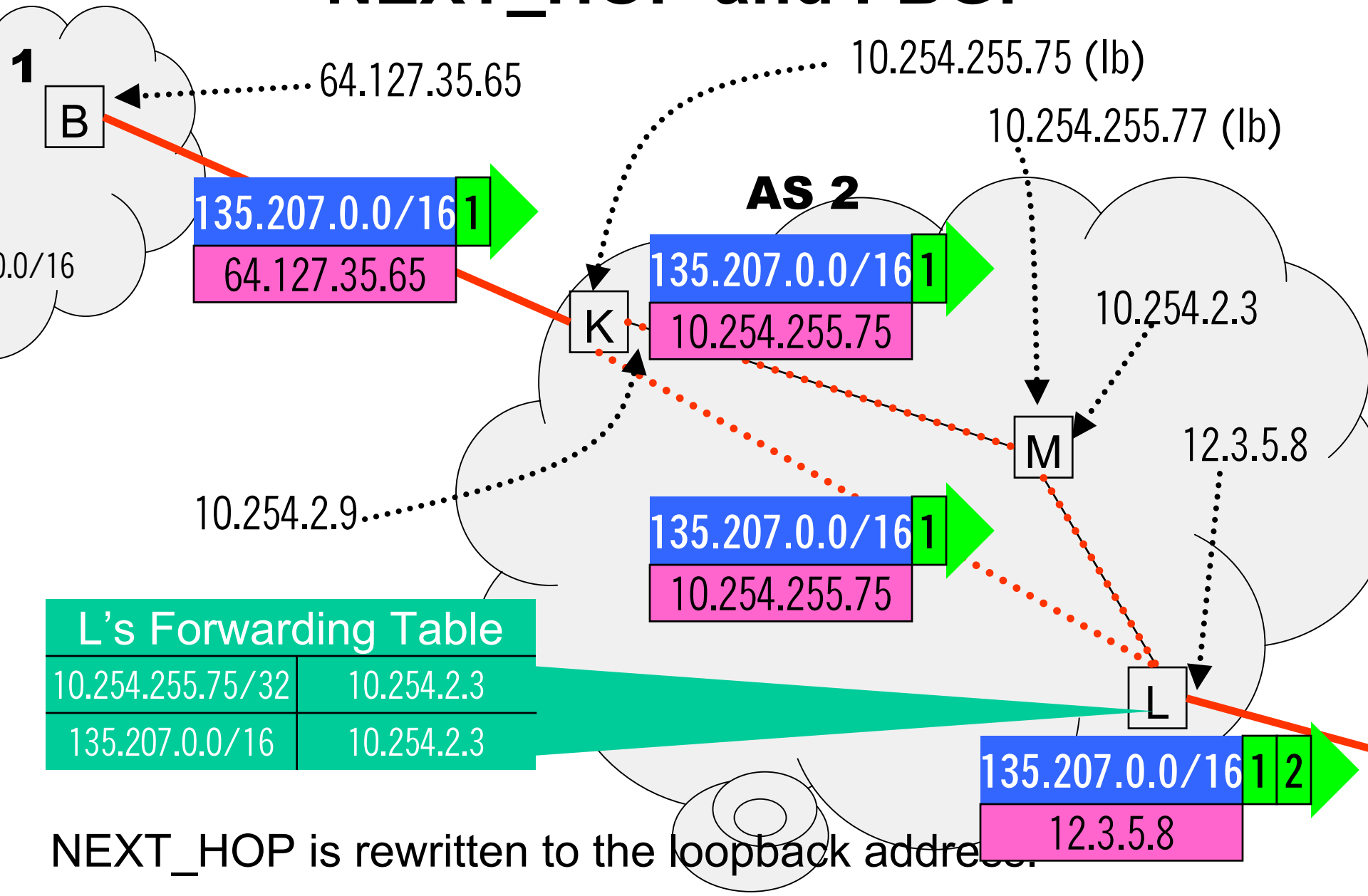
(If we are not using I-BGP).

NEXT_HOP and I-BGP



NEXT_HOP is rewritten to the loopback address.

NEXT_HOP and I-BGP



L's Forwarding Table	
10.254.255.75/32	10.254.2.3
135.207.0.0/16	10.254.2.3

NEXT_HOP is rewritten to the loopback address