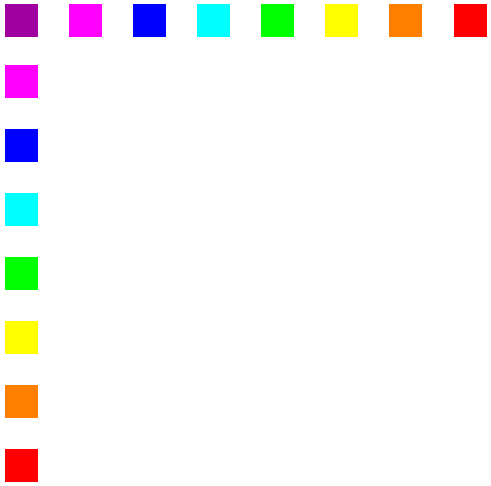


Border Gateway Protocol

Fulvio Risso

Politecnico di Torino

This set of slides is based on a previous version created by Mario Baldi and Giorgio Valent

Inter-domain routing (routing among AS)

- AS exchange routes through different protocols than the ones used inside
- Economic, scalability and management issues
  - A problem inside an AS should not affect the routing in the others
  - The external traffic crossing the AS should not damage internal traffic
  - Usually, an AS would like to use any path inside its domain; this may not be true among AS, and some paths (even if appear to be more convenient) should not be used for different reasons
  - For instance, an AS may not want that its traffic crosses a specific AS before reaching the destination (for security reason, for example)

Protocols for inter-domain routing (1)

- Static routing
  - Configuration of routers by hand
    - ⊕ This is the best "algorithm" to implement complex policies and to have the complete control over the network paths
    - ⊕ No control traffic needed
    - ⊖ Does not react to topological changes
    - ⊖ Easy to introduce inconsistencies
- Exterior Gateway Protocol (EGP)
  - Introduced by Internet Community in RFC 827
  - First protocol completely dedicated to routing among domains
  - It was a first try, but currently nobody used it

Protocols for inter-domain routing (2)

- Border Gateway Protocol (BGP)
  - Successor of EGP
  - Commonly used on Internet backbone
  - Defined in RFC 1105, 1163 and 1267, etc.
  - The current version is BGP4+ which includes extensions allowing to transport more protocol families (e.g., IPv4, IPv6)

Protocols for inter-domain routing (3)

- Inter-Domain Routing Protocol (IDRP)
  - Created as an evolution of BGP in order to support OSI addressing
  - Many improvements compared to the original BGP
    - BGP has evolved a lot since then, though
  - Rather complex parts
  - Currently nobody uses it
    - AS routing requires all the AS to used the same protocol
    - This is different from intra-domain routing, in which multiple protocols can coexist

Border Gateway Protocol (1)

- Path Vector protocol
  - It records the sequence of the Autonomous Systems (AS) for each destination
    - Each AS is identified through 2 (or 4) bytes
    - Actually, distance vector protocols store only the cost
    - No problems for count-to-infinity (a loop is detected by a duplicate AS in the path)
- Exchange of routing updates using reliable connections
  - Adjacent routers communicate through a TCP connection
- Support for routing policies
  - Policies are used to change the selection criteria for routes
    - Both internal routes and annoucements to the neighbors

Border Gateway Protocol (2)

- No automatic detection of neighboring routers
  - Peers must be configuration by hand
- Extensible
  - Data is transported as a set of “attributes”, formatted as Type-Length-Values (TLV)
  - A BGP instance may even transport attributes that are not understood by the current router
- Destinations are in form destination address/prefix length
- The routers can aggregate information of routing received before propagating them
  - Decreases the routing traffic
  - Decreases the routing table
  - Decreases the number of transients

Reliable transmission of updates

- Avoids periodic updates
  - Decreases bandwidth consumed to send routes
    - BGP routing tables may ba really big (hundres of thousands routes)
  - When a route is received, it means that something has changed
    - In classical Distance Vector protocols, the receipt of a DV does not necessarily imply that some routes changed
- Reliable route exchanges on a TCP connection
  - Other protocols (EIGRP, OSPF) implement reliable route exchange with a protocol-specific mechanism
    - E.g., protocol-specific timers, acknowledgement messages
  - BGP uses the TCP for setting up a reliable connection, and sends routes over it

Using TCP as transport protocol (1)

- All TCP messages (e.g., keepalives, route exchanges) are encapsulated in TCP
- It simplifies the specification of the BGP protocol
  - It reuses existing components instead of redefining yet another mechanism
  - BGP does not need to deal with retransmissions, lost messages, etc.
- It requires an explicit KeepAlive mechanism, managed by BGP itself
  - TCP does not provide the information if the remote peer is still reachable or not
  - Keepalive messages may be delayed by the TCP mechanisms

Using TCP as transport protocol (2)

- It may have some problems if a link is not really reliable
  - TCP will hide momentarily the failure of the link
- It may require some more configuration (e.g., QoS at the network level) so that the network will privilege BGP packets when required
  - In case of congestion, the TCP decreases the transmission bit rate, which may prevent the timely transmission of routing updates
  - The congestion may be the result of a routing problem, hence we would like to solve it as soon as possible

Path Vector

- Variant of DV algorithm, which records the "path" between two networks
  - "Path" is the list of the AS toward the destination
- BGP messages are much bigger than DVs
- PV is more stable, as it is easy to detect loops
  - If a router receives a PV that already contains its AS number, it discards the PV without propagating it, as we are running into a routing loop
  - If not, the router inserts its own AS number in the Path Vector and then it propagates it to its neighbors
    - Warning: this may not be valid in case of internal peering (more details later)

BGP costs

- BGP does not support explicit cost metric
  - Each AS can have diverse requirements, hence a different metrics (delay, or bandwidth, etc.)
  - No cost metric valid for all the Autonomous Systems
- The cost is the **number** of AS traversed
  - A path that traverses 2 AS is better than a path that crosses 4 AS, no matter how big the ASes are

Peers (1)

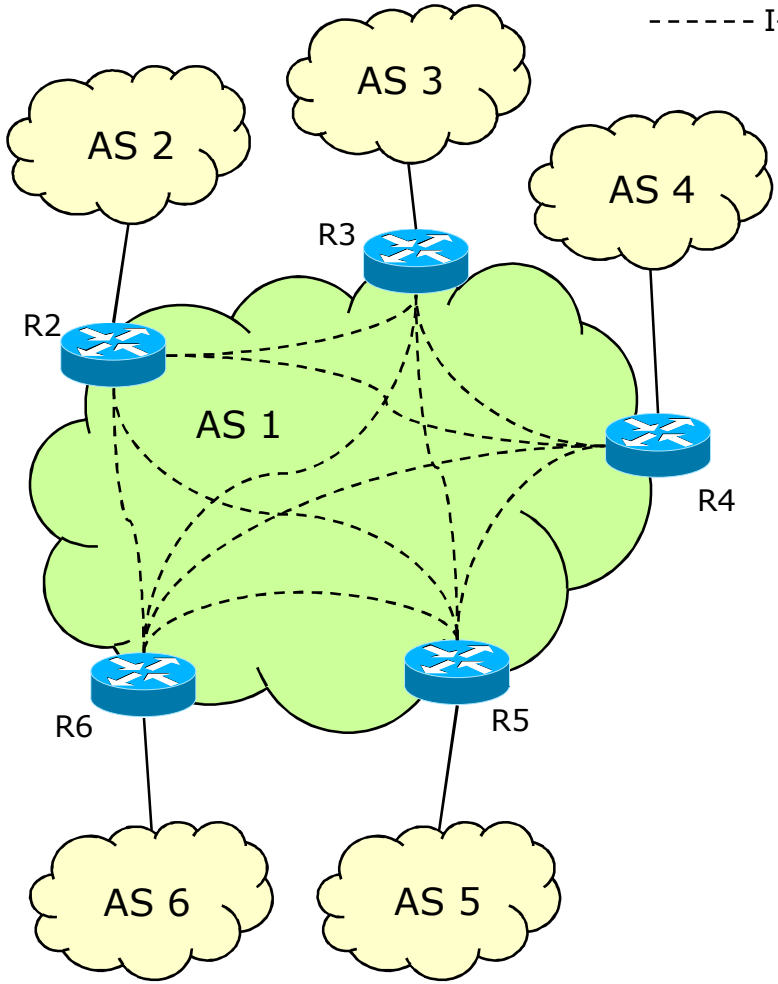
- Peers do not "recognize" (i.e., discover) each other automatically
  - Key difference compared to other routing protocols
  - The peering (internal as much as external) must be defined by the administrator
  - Peers may not be connected through a direct link
    - Other routers may exists between two peers
- Differentiation between Exterior and Interior peers
  - Exterior: two routers belonging to two ASes that "see" each other directly
  - Interior: two routers in the same AS that need to exchange BGP routes

Peers (2)

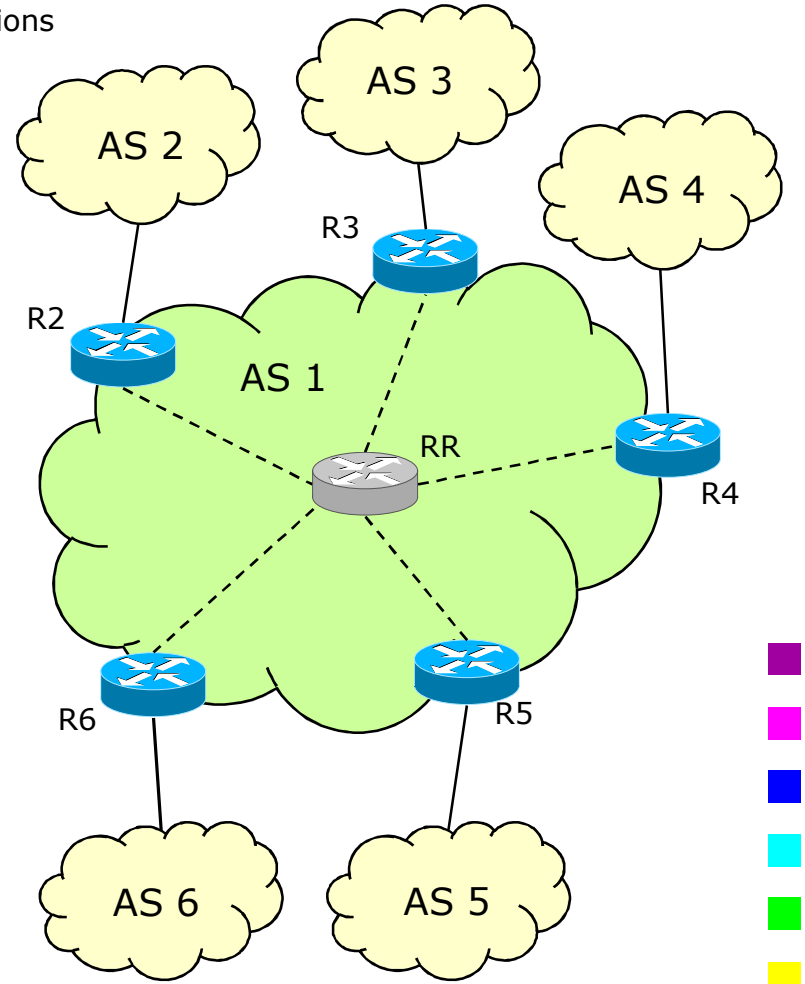
- I-BGP and E-BGP
  - Connection to internal (I-BGP) and external (E-BGP) peers
  - I-BGP: usually peers are not directly connected
  - E-BGP: usually peers are directly connected
  - The processing of BGP messages and the routes announced on those connections may be different

- Each BGP router of an AS must have an I-BGP session with all other internal routers (full I-BGP mesh)
  - The routes that stay inside an AS are not exchanged in the I-BGP session
  - Route Reflector: eliminates the full mesh requirement
  - Confederated AS

Route Reflector



I-BGP full mesh



I-BGP full mesh with route reflector

AS Confederation

- AS is divided in mini-Ases
  - Each one has a private AS number
- Peering between routers in different mini-ASes is based on E-BGP
- Advertisements are handled by inter-mini-AS routers are I-BGP does:
  - No changes to some of the attributes: next-hop, multi-exit-disc and local pref
- Routers in the same mini-AS have a complete I-BGP mesh

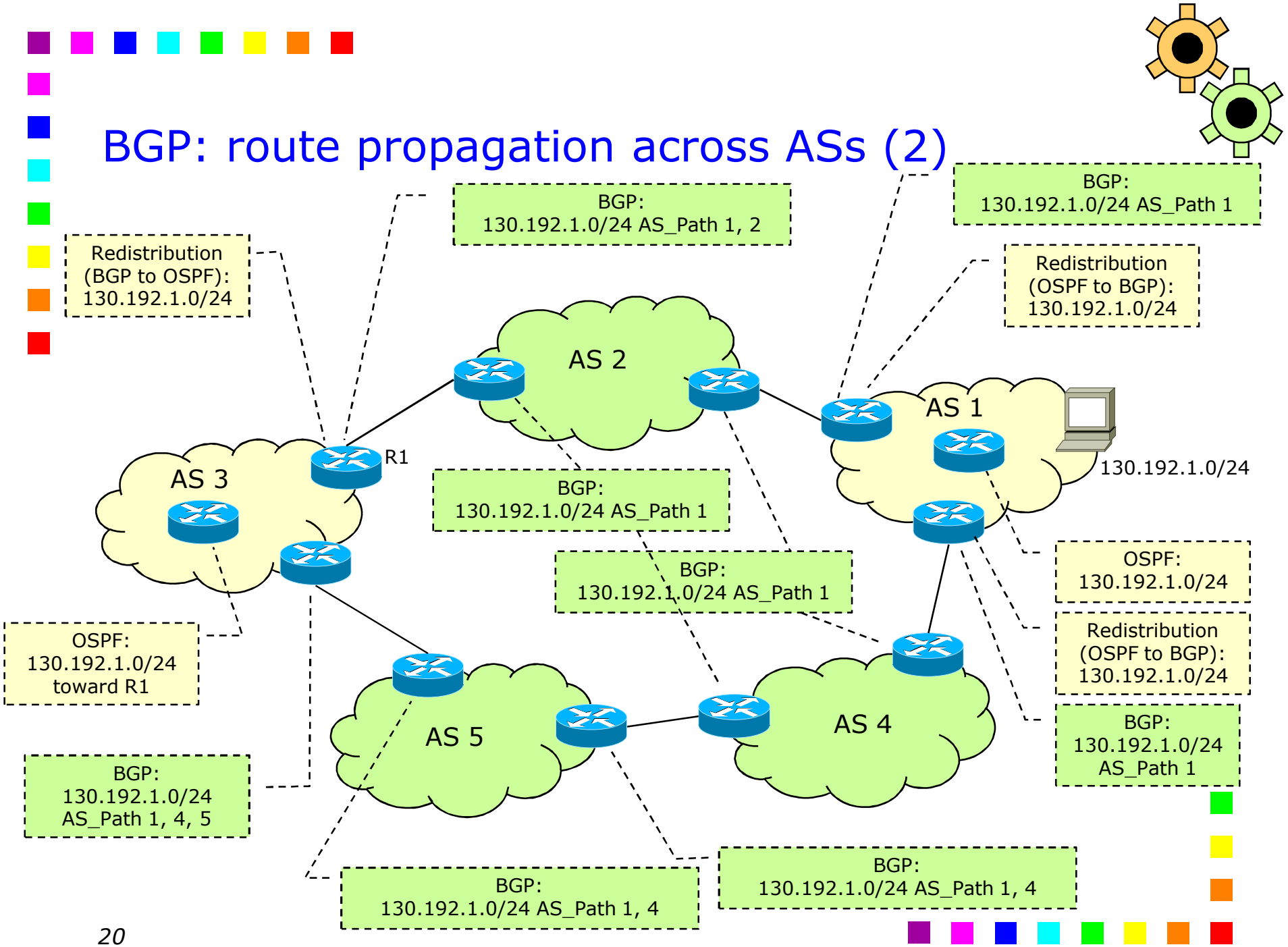
Why are I-BGP sessions needed?

- Used to exchange the external routes independently from the **routes** exchanged by the interior protocol
  - A modification of a BGP route does not necessarily have any impact on the internal routes
    - A variation of the AS external routes does not necessarily require the re-computation of the internal routes (no transient, less processing)
    - Smaller number of routes handled by the internal protocol
  - The BGP updates are normal packets of data for routers inside the AS
- Used to exchange the external routes independently from the **interior protocol**
  - The network can use any protocol for the internal routing
  - Internal protocols cannot transport information specific for BGP (e.g., AS_Path)

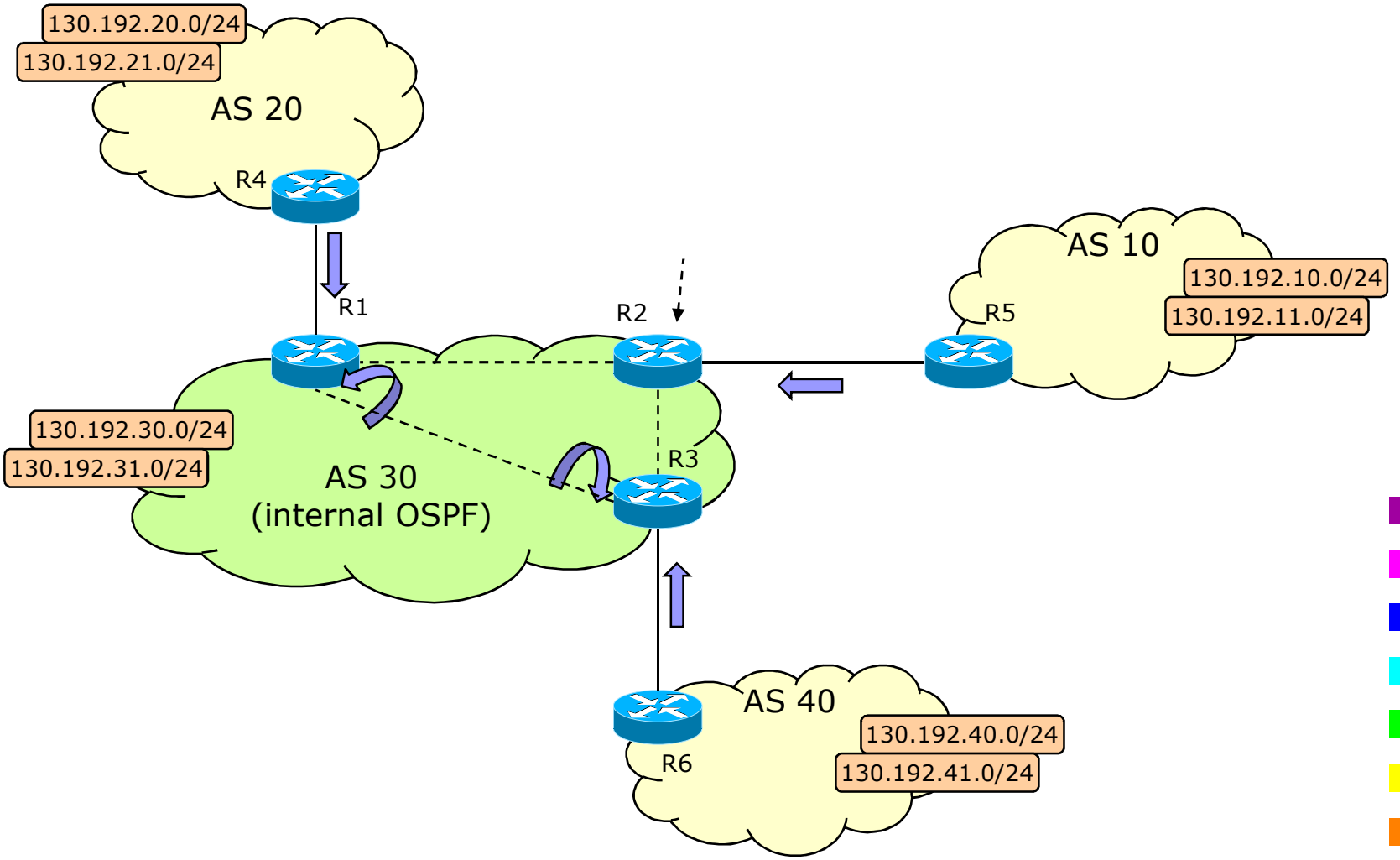
BGP: route propagation across ASs (1)

- The actual algorithm is more complicated but… let's see what happens when leaving some of the details out of the picture (for now)
- Each BGP router propagates the list of reachable networks toward other ASs
  - Reachable networks:
    - Internal destinations (i.e., networks in its AS)
    - External destinations (i.e., networks in other AS)
- E-BGP sessions: each destination is **prepended** with the current AS number
  - The receiver knows exactly the list of ASes that need to be traversed in order to reach that network
    - Warning: in some cases, the list of ASes is not complete; more details later
- I-BGP sessions: AS list is sent as is

BGP: route propagation across ASs (2)



BGP: route propagation



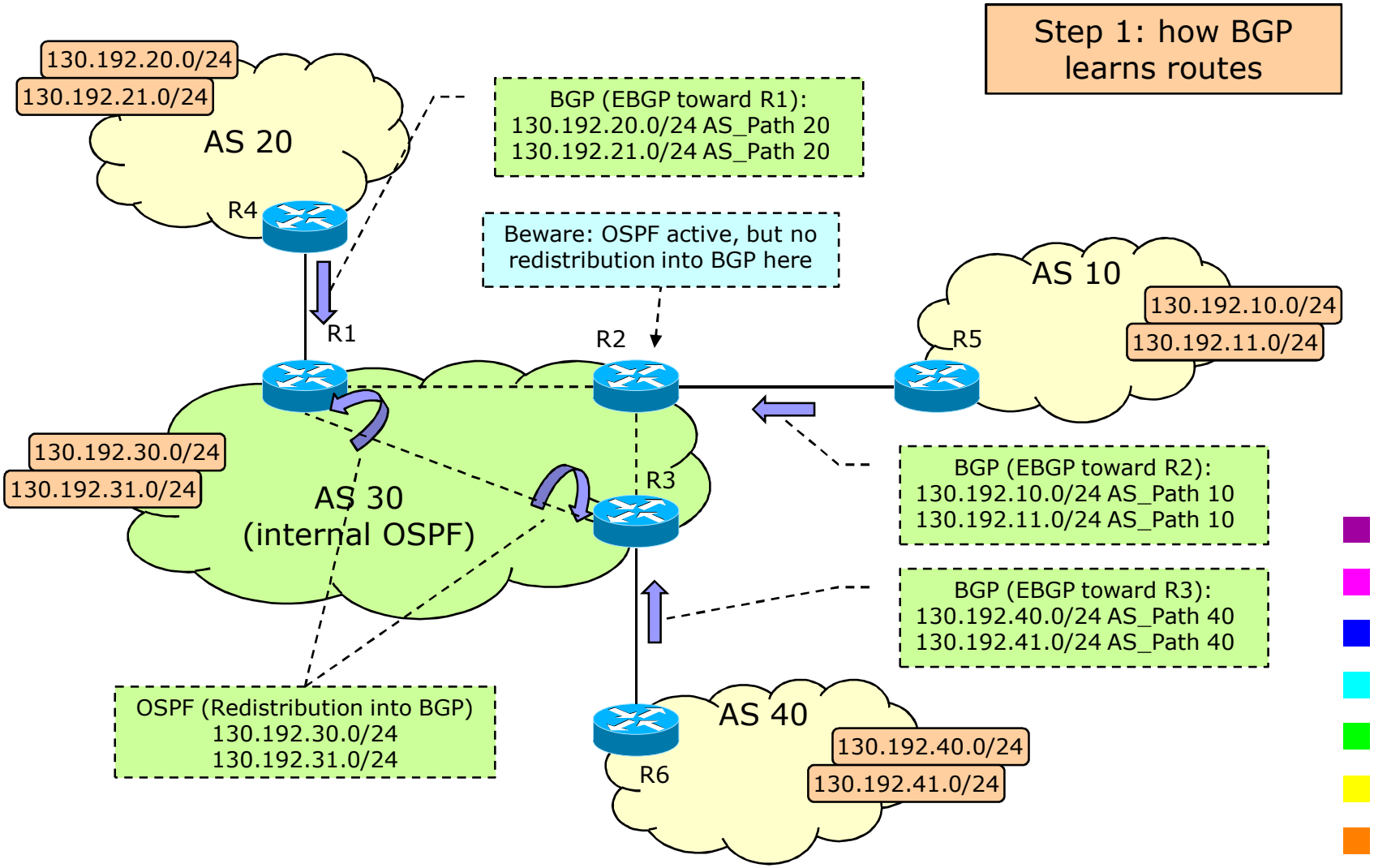
BGP: route propagation (1)

- The actual algorithm for route propagation is, in fact, more complicated
  - We have to distinguish between I-BGP and E-BGP connections
  - Let’s look at the next slides with an example
- E-BGP connections
  - Redistributed internal routes are propagated toward other E-BGP peers
    - R1 propagates AS 30 destinations to R4
  - External routes are propagated toward other E-BGP peers if their AS is not on the best path toward those destinations (split horizon-like mechanism)
    - R1 propagates AS 40 destinations to R4
    - R1 does not propagate AS 20 destinations back to R4

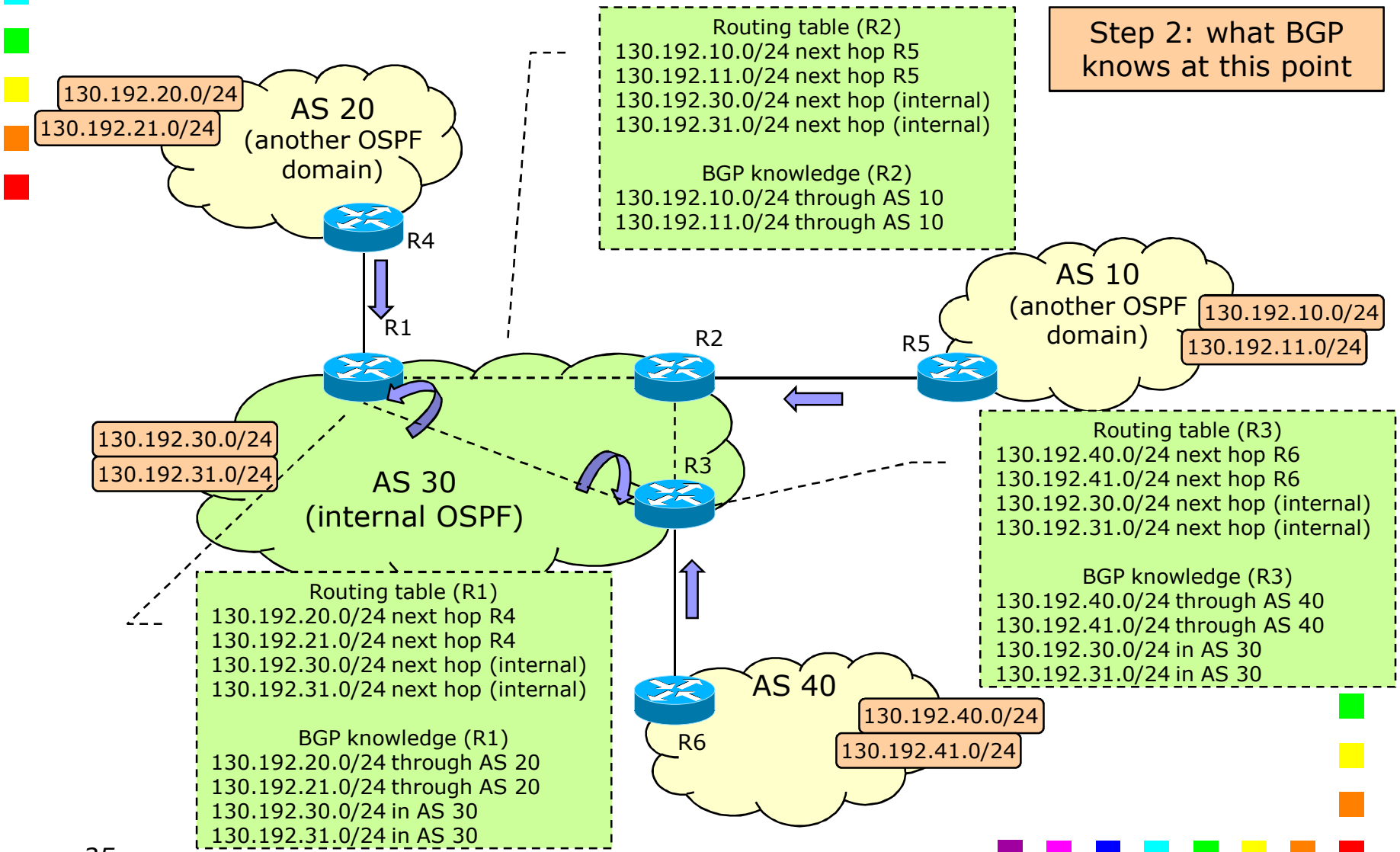
BGP: route propagation (2)

- I-BGP connections
  - All routes learned from E-BGP are propagated toward the other I-BGP peers
    - R1 propagates AS 20 destinations to R2 and R3
  - Internal routes are NOT propagated toward other I-BGP peers
    - R1 does not propagate destinations in AS 30 to R2 and R3
  - All routes received from I-BGP are NOT propagated to other I-BGP peers
    - R1 receives AS 40 destinations in I-BGP, so it does not send those destinations to R2 and R3
    - AS Path is not updated, so there would be no way to detect routing loops

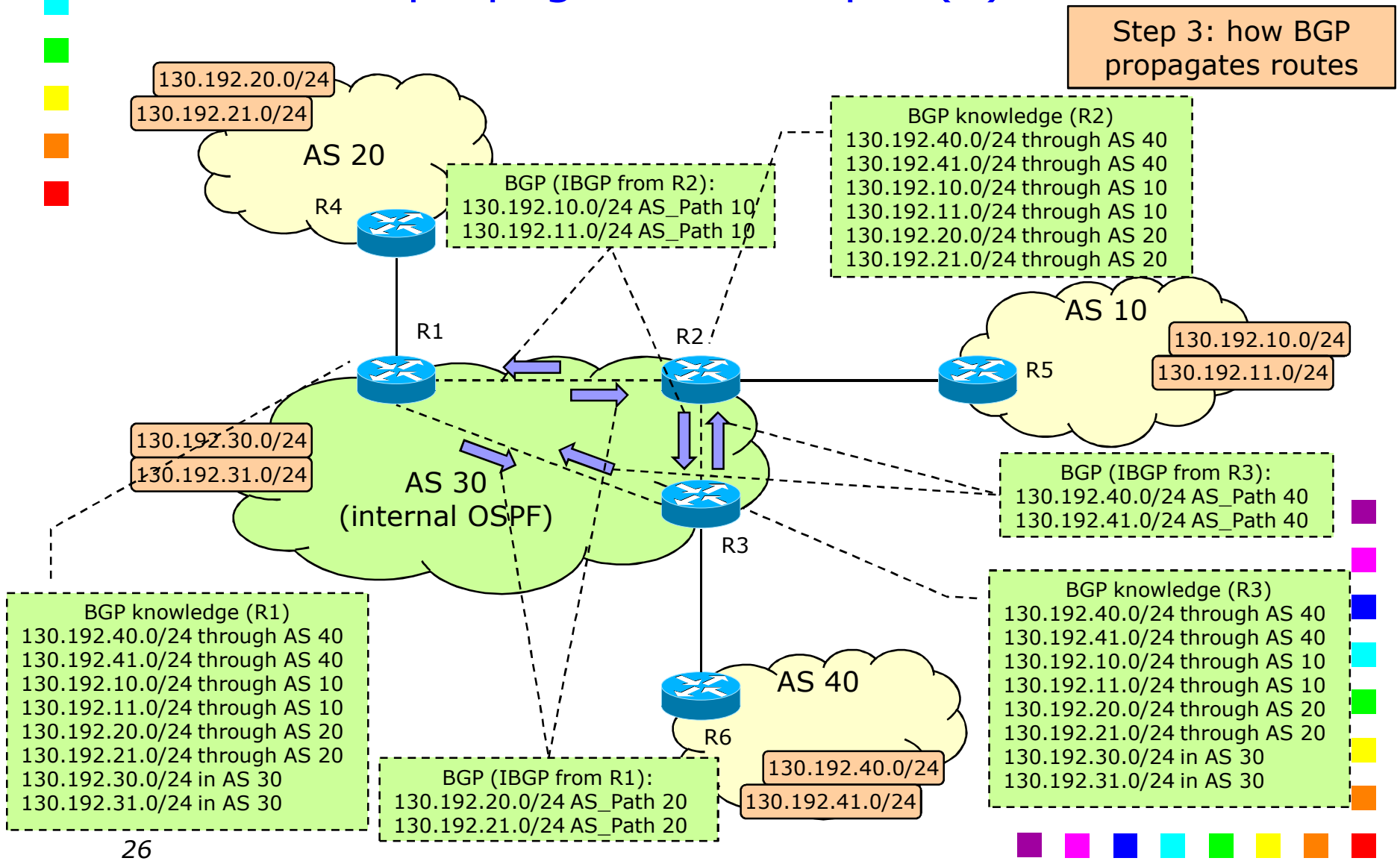
BGP: route propagation example (1)



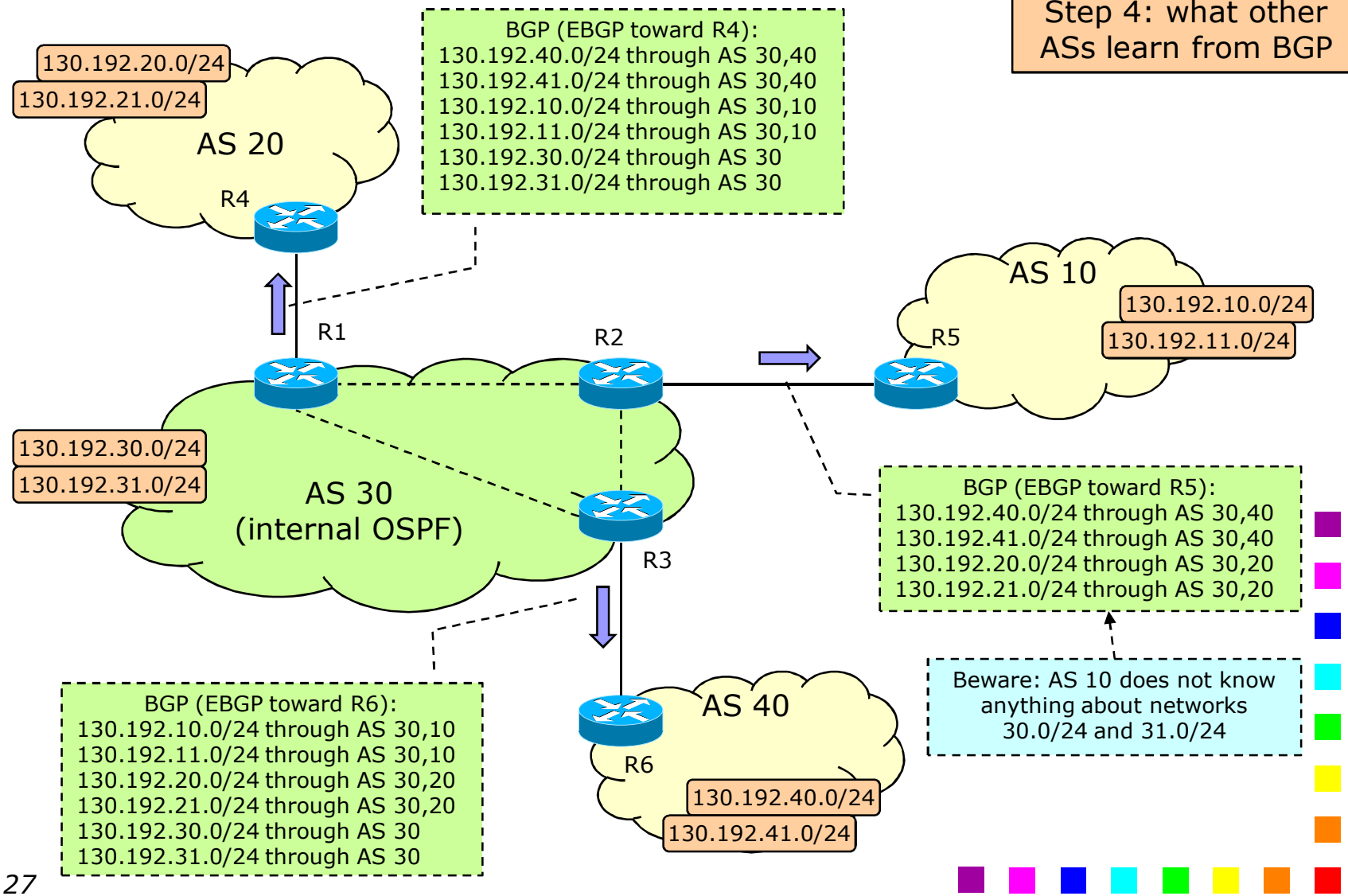
BGP: route propagation example (2)



BGP: route propagation example (3)



BGP: route propagation example (4)



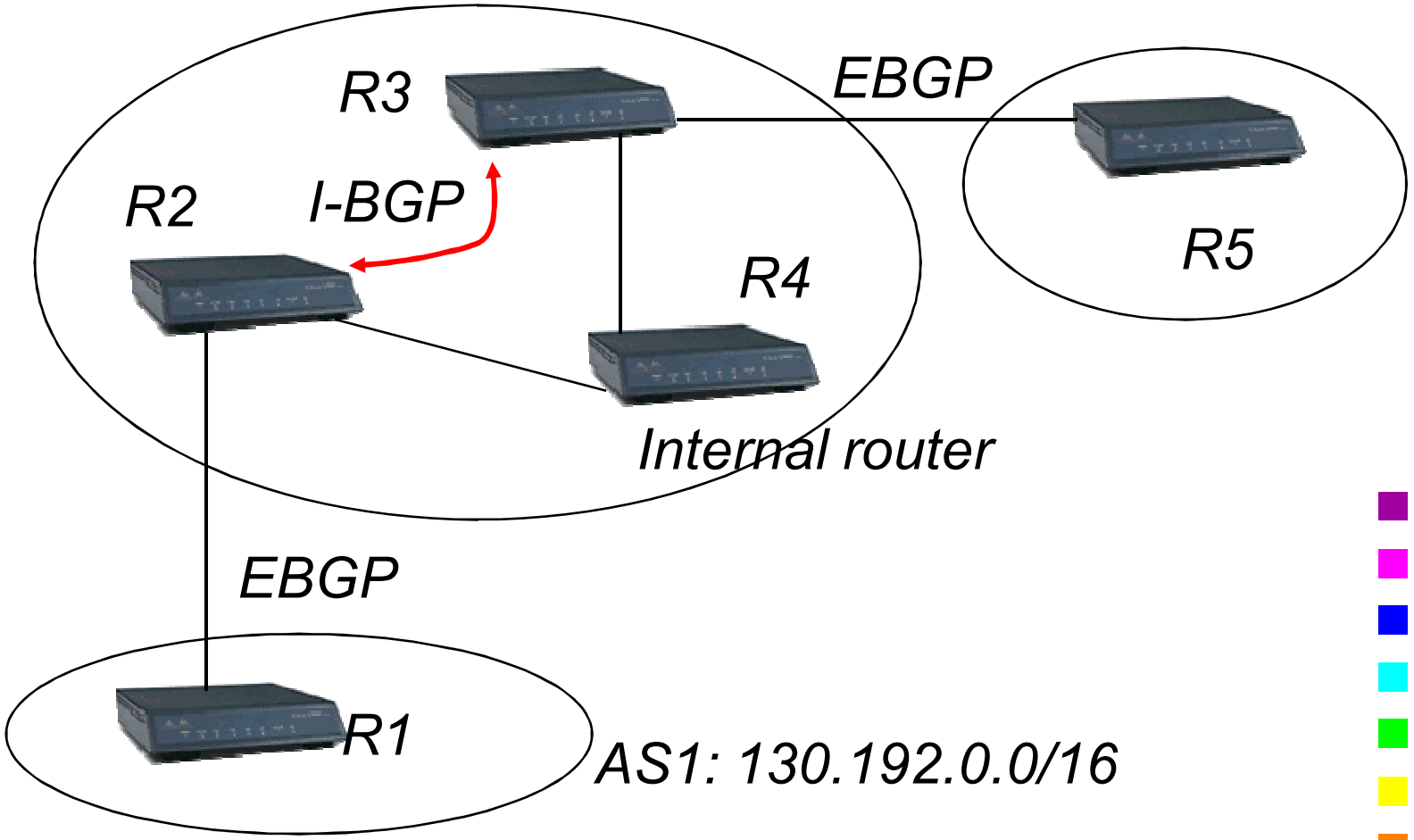
Some considerations about internal routes

- BGP does not propagate internal routes in I-BGP
- Some consequences
  - All I-BGP routers must be aware of the networks internal to the AS
    - I.e., something else MUST tell internal routes to BGP
    - E.g. redistribution from IGP (e.g., OSPF) to BGP
  - This "something else" must be active on all the BGP routers of the AS and must inject the same info into all the BGP routers
    - E.g., troubles for R2 in the previous example
      - Please note that R2 has the complete routing table
      - Do not confuse the routing table (which is the output of all the routing processes active on the router) with the BGP routing process

IGP-BGP Synchronization

- BGP routers in a transit AS learn external destinations from other BGP routers through I-BGP, but routing of packets through the AS (towards egress BGP router) relies on internal routers
- Routing tables of internal routers must contain entries to extra-AS destinations
  - Must have been advertised by IGP
  - There is no point in advertising with BGP destinations not advertised by IGP

IGP-BGP Synchronization



Without IGP-BGP Synchronization

- R1 advertises 130.192.0.0/16 through E-BGP
- Advertisements reach R2 and propagates to R3 through I-BGP
- R3 distributes through E-BGP to R5
- R4 can learn of 130.192.0.0/16 only through IGP
- When R5 sends to R3 traffic directed to 130.192.0.0/16 R3 forwards it to R4 to reach the exit point R2
- If R4 has not reaceived any advertisement for 130.192.0.0/16 through an IGP, it will discard the packet

With IGP-BGP Synchronization

- R3 advertises 130.192.0.0/16 only when that
- It might be good to disable synchronization when:
  - AS is not a transit one
  - All AS routers use BGP

Overlapped routes

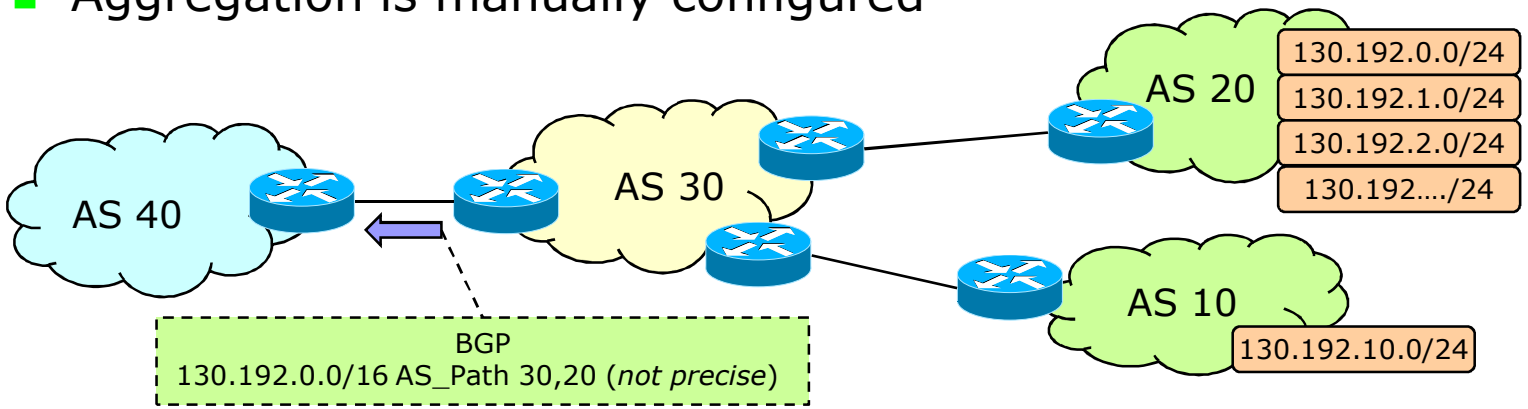
- Overlapped routes are a feature defined in IP in order to increase the capability to aggregate routes
  - E.g., destinations 30.0.0.0/16 and 30.0.2.0/24 represent two different routes
    - More specific path valid for a smaller set of destinations
      - Longer address prefix
    - Less specific path valid for a larger set of destinations
      - Shorter address prefix
  - When forwarding traffic, routers use the most specific route
- BGP can transport overlapped routes in its messages

```
Router# show ip route

S    30.0.0.0/16 [1/0] via 20.0.0.2
S    30.0.3.0/24 [1/0] via 20.0.0.3
```

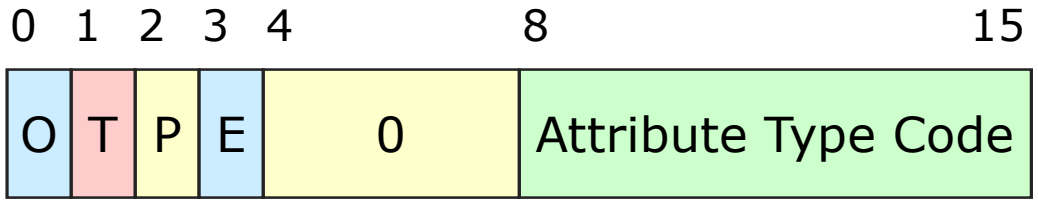
Aggregate routes

- BGP can *aggregate* routes if those are reachable through the same AS
- In this case, the route is propagated in such a way that the receiver knows that the announced AS_PATH may not be valid for all the destinations
  - The receiver knows that its packets may not follow the advertised path
  - The route cannot be disaggregated
- Aggregation is manually configured



Attributes

- Attributes are encoded in the TLV format and are used to transport information in BGP
- Many attributes are defined
  - E.g., the list of AS traversed to reach the destination
- Two bytes are used to encode the “type”
- Some additional flags are used to discriminate the “scope” of the attributes



BGP attributes: types

- BGP has four types of attributes as listed:
  - Well-known mandatory
    - Must be understood in all the implementations, and it must be present in all the messages
  - Well-known discretionary
    - Must be understood in all the implementations, but it may not be present in all the messages
  - Optional transitive
    - May not be undestood by all the implementation, and it has to be propagated along the entire path
  - Optional non-transitive
    - May not be undestood by all the implementation, and it does not have to be propagated along the entire path

Attributes and Flags (1)

- Two flags (O/T) are used to discriminate those categories
  - A flag for discriminating between mandatory/discretionary is not needed (the attribute is well-known, so everybody knows)
- O (well-known/optional): equal to "1" if this attribute must be understood by all the implementations (i.e., all the BGP routers in the path)
- T (transitive/local)
  - transitive: if the attribute can be updated or understood by other routers (for example, the bandwidth of the smallest link in the path) and thus should be transmitted to the following routers
  - local: if the attribute is meaningful only inside an precise AS; it does not have to be propagated to other AS
- Well-known attributes are always transitive and therefore their "transitive" bit is always set to one

Attributes and Flags (2)

- P (partial): set if a router on the path is not able to understand the meaning of this attribute (for example an optional one); it means that at least one router on the path did not understand that attribute
- E (length): determines if the LENGTH field is coded with one byte (E = 0) or two

Well-known attributes: Origin (Type = 1)

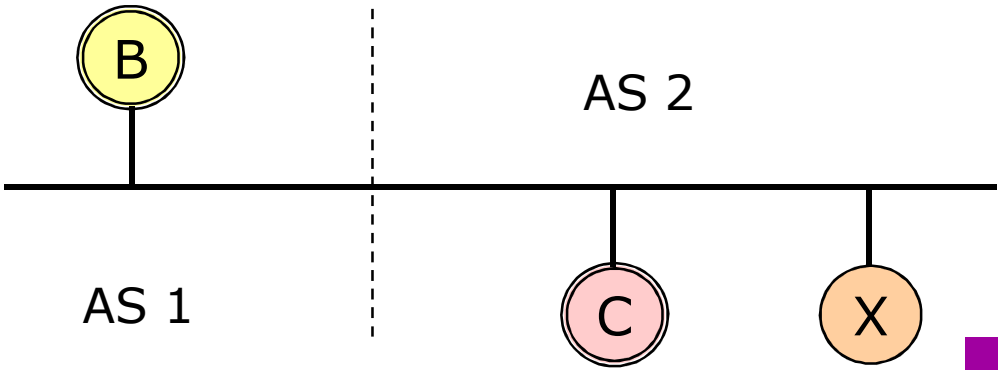
- Origin of the route; three possible values are allowed:
  - IGP: the route originated on a BGP router
    - This route type includes any route that originated from the BGP process on a BGP-speaking router
    - In practice, it can be learned by a router that is configured to advertise that prefix (e.g., "network" command in Cisco routers)
  - EGP: learned from an "EGP" session
    - Historical, since nobody is still using EGP
  - Incomplete: the route originated from a routing process other than BGP, and entered BGP by means of manual redistribution, such as redistribution from an IGP protocol, static route, or connected route
    - The destination is not reachable through BGP only

Well-known attributes: AS_PATH (Type = 2)

- List of the AS traversed to reach the destination
  - Ordered components (AS_SEQUENCE)
  - Non-ordered components (AS_SET)
- The AS_PATH is modified only when propagated through E-BGP sessions
  - If the first component of the current AS_PATH is ordered, the AS number is added at the beginning of the sequence
  - If the first component of the current AS_PATH is not ordered, an ordered component containing the identifier of the AS is added

Well-known attributes: NEXT_HOP (Type = 3)

- Used when a network contains multiple routers belonging to two different AS
- Example
  - B and C have an established BGP session
  - However, the best path to the destination goes through X
  - C uses the NEXT_HOP attribute to tell B to use X as a next hop to reach the destination
- This attribute is “non transitive”
  - It is useless to transmit it to other BGP nodes in the path

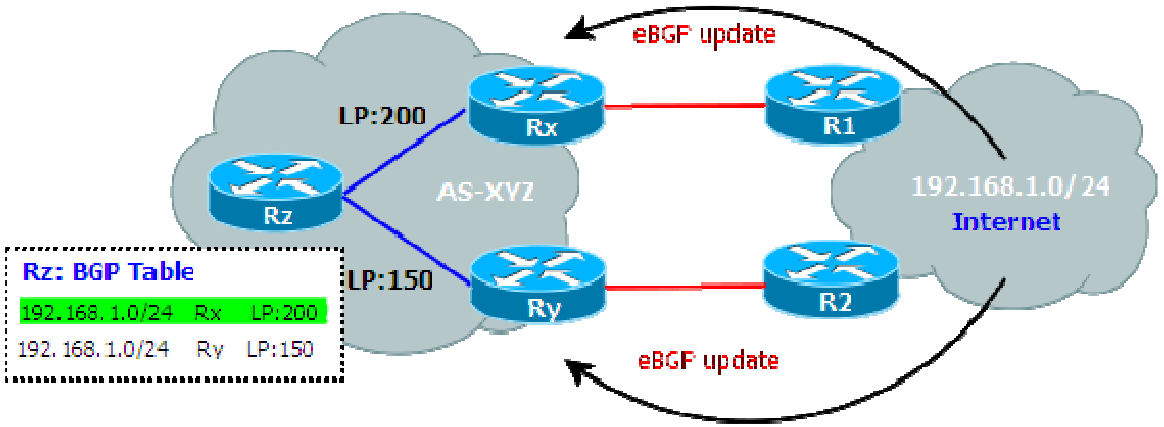


Other Well-known attributes

- Unreachable (Type = 4)
  - Set when a path previously announced becomes unreachable
  - Since advertisements are not periodic, routes do not expire, they need to be withdrawn
  - This update message also has the function of the "poisoned reverse" in DV protocols
- Inter-AS Metric (Type = 5)
  - Used when two ASes are connected through multiple links
  - This attribute allows to discriminate between those links (e.g., a "standard" connection and a "backup" link)
  - Non transitive and local attribute

Other attributes (1)

- local pref
  - Well-known discretionary attribute
  - Included only in I-BGP and never propagated outside
  - It represents one of the ways to alter the path taken by one AS to reach a given destination in another AS
  - It is valid for all the routers in the local AS
  - It influences the outgoing traffic
- Local Pref influences *routes*, Inter-AS Metric influences *links*



Other attributes (2)

- Atomic aggregate
  - Well-known discretionary attribute
  - The purpose of the attribute is to alert BGP speakers along the path that some information have been lost due to the route aggregation process and that the aggregate path might not be the best path to the destination
- Aggregator
  - Added by the router that has generated the Atomic Aggregate
  - Contains the identifier of the AS and the IP address of the router

Policies

- Configured manually
- Allow the router to assign a “preference” to the possible paths and to choose the best existing path that satisfies the policies
- Very complex policies can be imposed

Routing Information Base (RIB)

- Adj-RIBs-In
  - Information learned from the received announcements
- Loc-RIB
  - Information used by the forwarding process
  - Selected through the decision process
- Adj-RIBs-Out
  - Information that will be propagated to other domains
  - Selected through the decision process

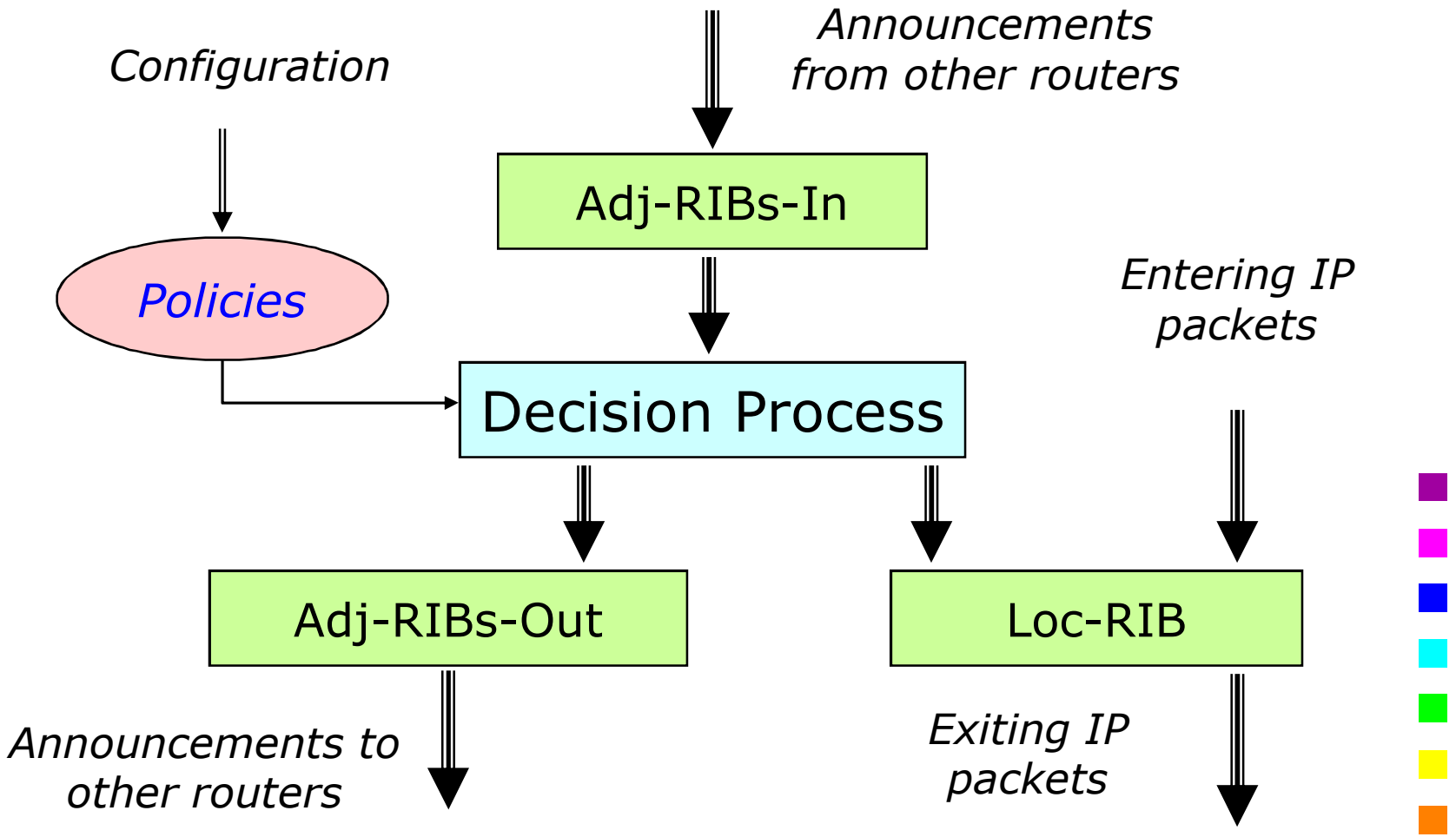
Reception and propagation (1)

- When a routing update is received
  - If the destination is in Adj-RIBs-In the new route replaces the old one
    - The decision process is executed
  - If the new route is more specific than the existing one
    - Different attributes the decision process is executed
    - Same attributes the new route is ignored

Reception and propagation (2)

- The destination is not in Adj-RIBs-In the new route is added
  - The decision process is executed
- La new route is less specific than an existing one the decision process is executed
  - Only on the destinations described by the new route

Policies, table, announcements



Decision Process (1)

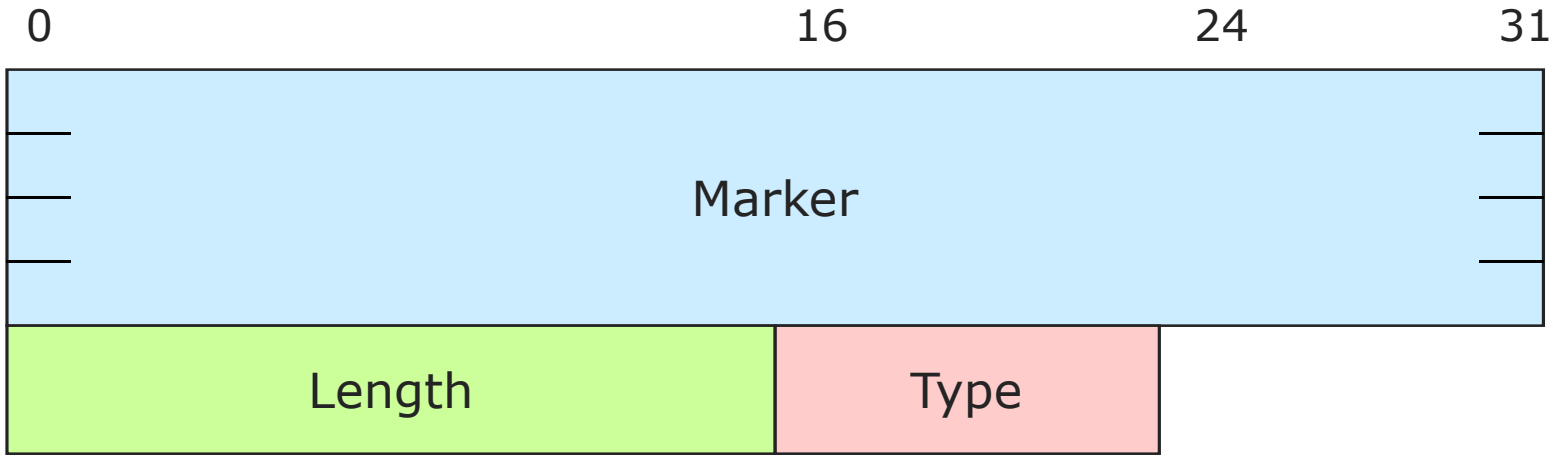
- Applies the policies defined in the Policy Information Base (PIB) to select the routes that need to be propagated
- This is a function that given the attributes of a route returns an integer (preference degree)
- The decision process does not take into account
  - The existence of other routes
  - The non-existence of other routes
  - The attributes of other routes
- Once the function applied to each route to a destination, the route with the largest preference degree is chosen

Decision Process (2)

- Works on all routes in Adj-RIB-In
- Select the routes propagated inside the Autonomous System
- Select the routes propagated outside Autonomous System
- Aggregate the routes and reduces the information to be transmitted

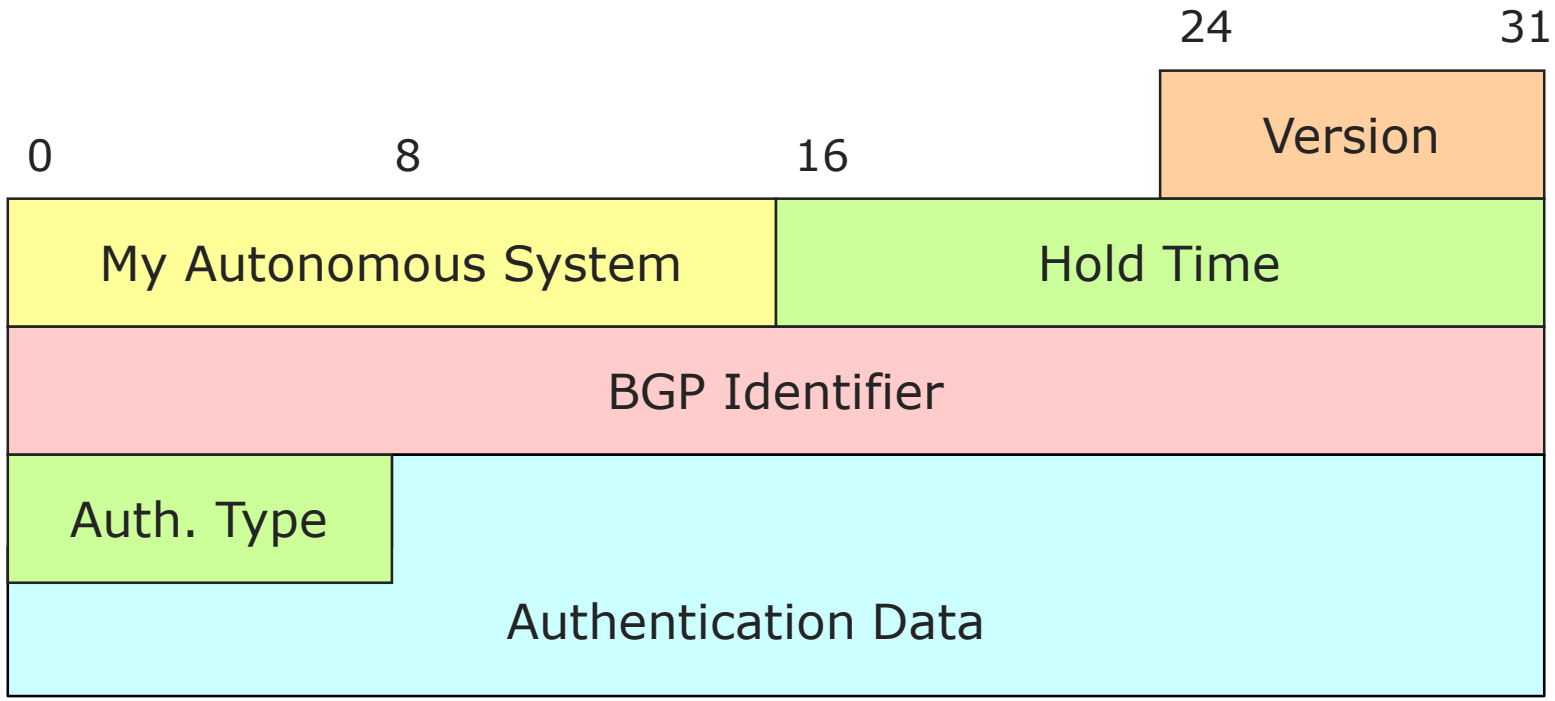
BGP: Messages format

- Common header
  - Marker: inserted for security reasons: a special algorithm (negotiated during the Initial Exchange phase) decides what it contains depending on the message. It represent a special type of checksum
  - Length: length of the TCP message
  - Type: Type of the BGP message (Open = 1, Update = 2, Notification = 3, Keepalive = 4)



OPEN Message (1)

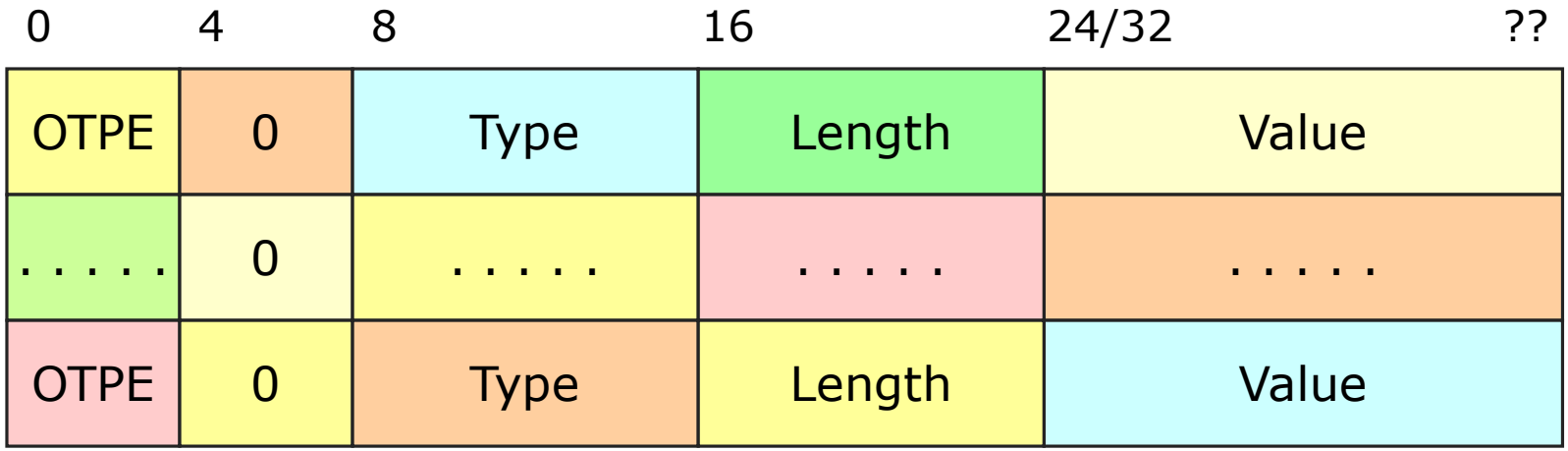
- First message transmitted when a TCP peering session is established
- Used to negotiate the BGP version
  - The connection is accepted only if both use the same BGP version

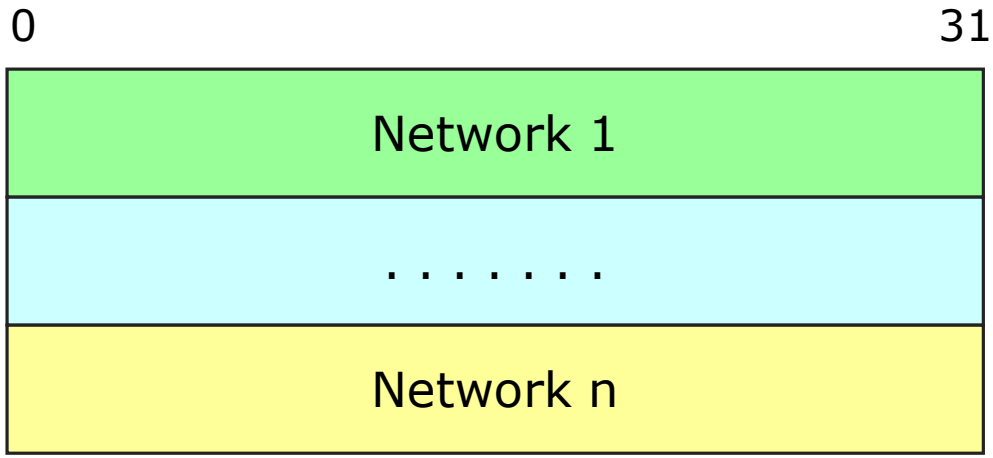


OPEN Message (2)

- My AS: number (ID) of the AS (declared by the IANA)
- Hold Time: number of seconds used by the Keep Alive procedure
- BGP Identifier: the IP address of one of the router interfaces
  - It is a parameter of the router set by management, independently of which interface is actually used to send packets
- Auth. Type: type of the authentication
- Authentication Data: authentication; does not exist if Auth_Type = 0
  - The authentication length is drawn from the field Length of the common header

UPDATE Message (1)





UPDATE Message (2)

- Update
  - Contains routing information
  - Distributes only one route
  - May withdraw multiple routes
  - May include many destinations (for which the route is valid)
- It is composed of 2 parts
  - A list of Path Attributes (in the format Flag, Type, Length of current attribute, Value)
  - A list of networks for which this announcement is valid
  - The number of Path Attributes that can be drawn from the field Length of the BGP header that indicates the length of the field Path Attributes (not of the whole packet)

KEEPALIVE Message

- Keepalive
  - Indicate to the neighboring router that the sender is still active
  - Used when there is no routing information to transmit
  - It is an empty packet composed only of Marker, Length and Type

NOTIFICATION Message

- Sent if an error or an abnormal situation occurs
  - Sent with the last message before the cutting of a connection with an other router (for example because a shutdown is being done)
  - Error Code: indicates the error situation
  - Subcode: better specifies the error
  - Data: to better identify the error