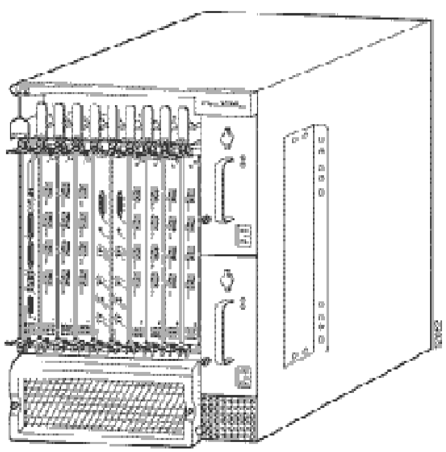
CS551
Inter-domain Routing

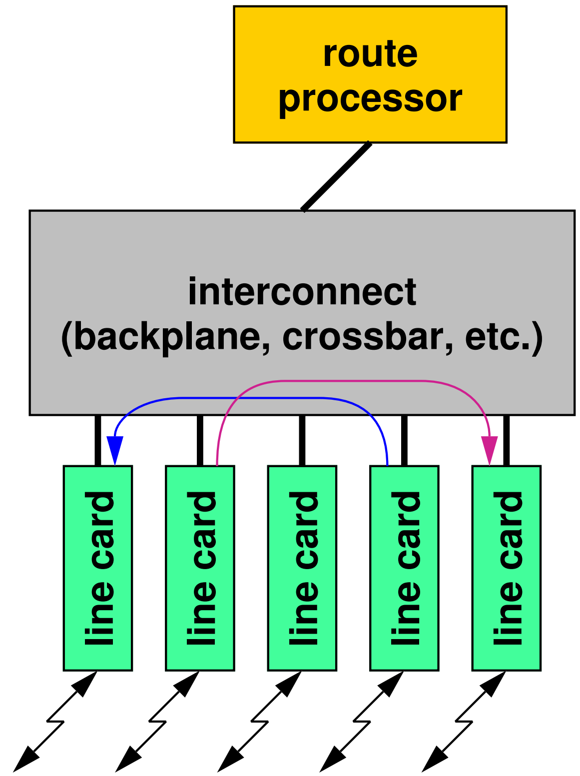
Bill Cheng

http://merlot.usc.edu/cs551-f12

Inter-domain Routing

Cisco 7xxx Router





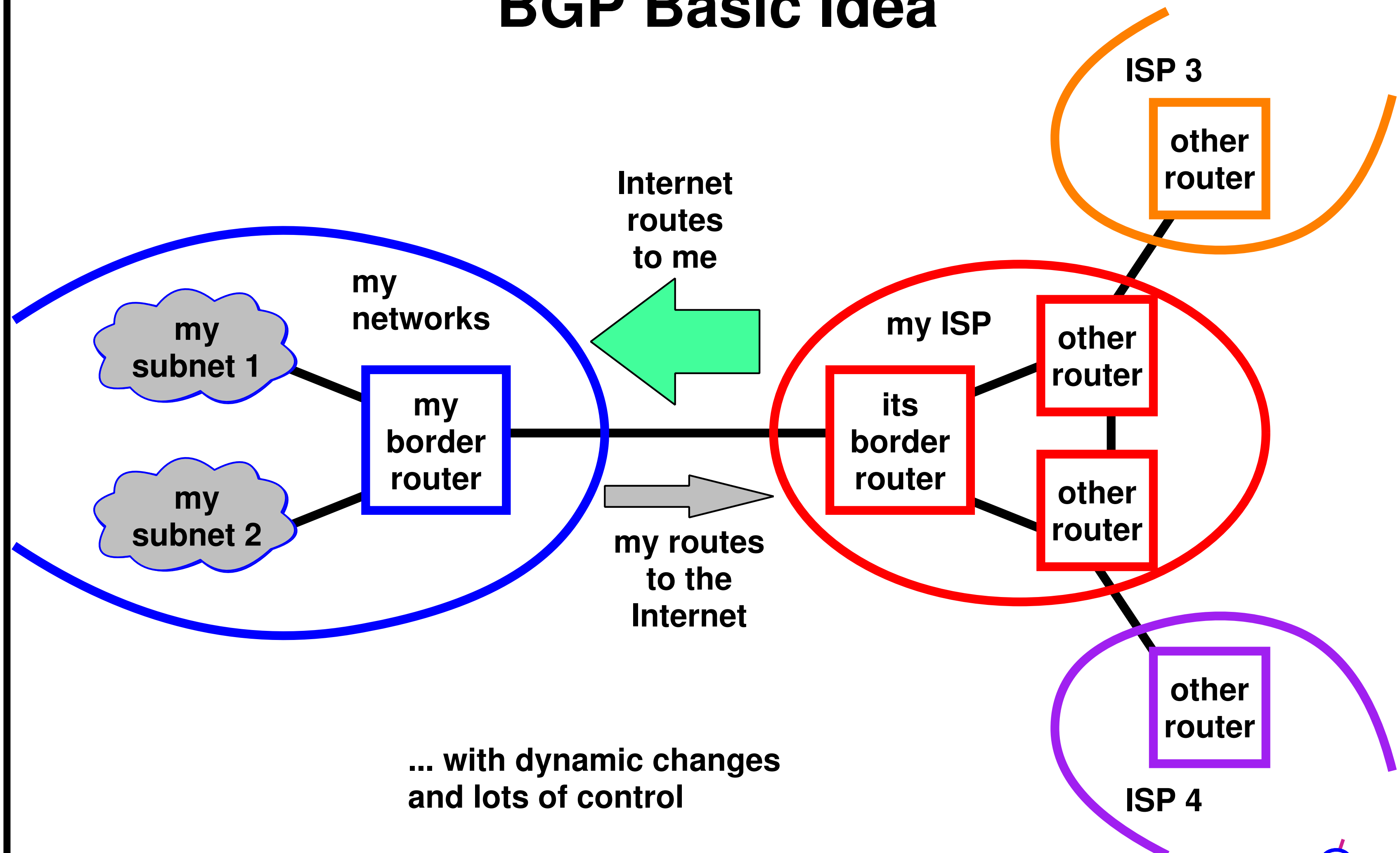
Sources

- John Stewart III: *"BGP4 - Inter-domain routing in the Internet"*
- RFC1771 [Rekhter95a]: main BGP RFC
- RFC1772-3-4: application, experiences, and analysis of BGP
- RFC1965: AS confederations for BGP
- Christian Huitema: *"Routing in the Internet"*, chapters 8 and 9
- Cisco tutorial online
- [Gao00b] sections 2.1 and 3.1
  - excellent terse overview of BGP

BGP History

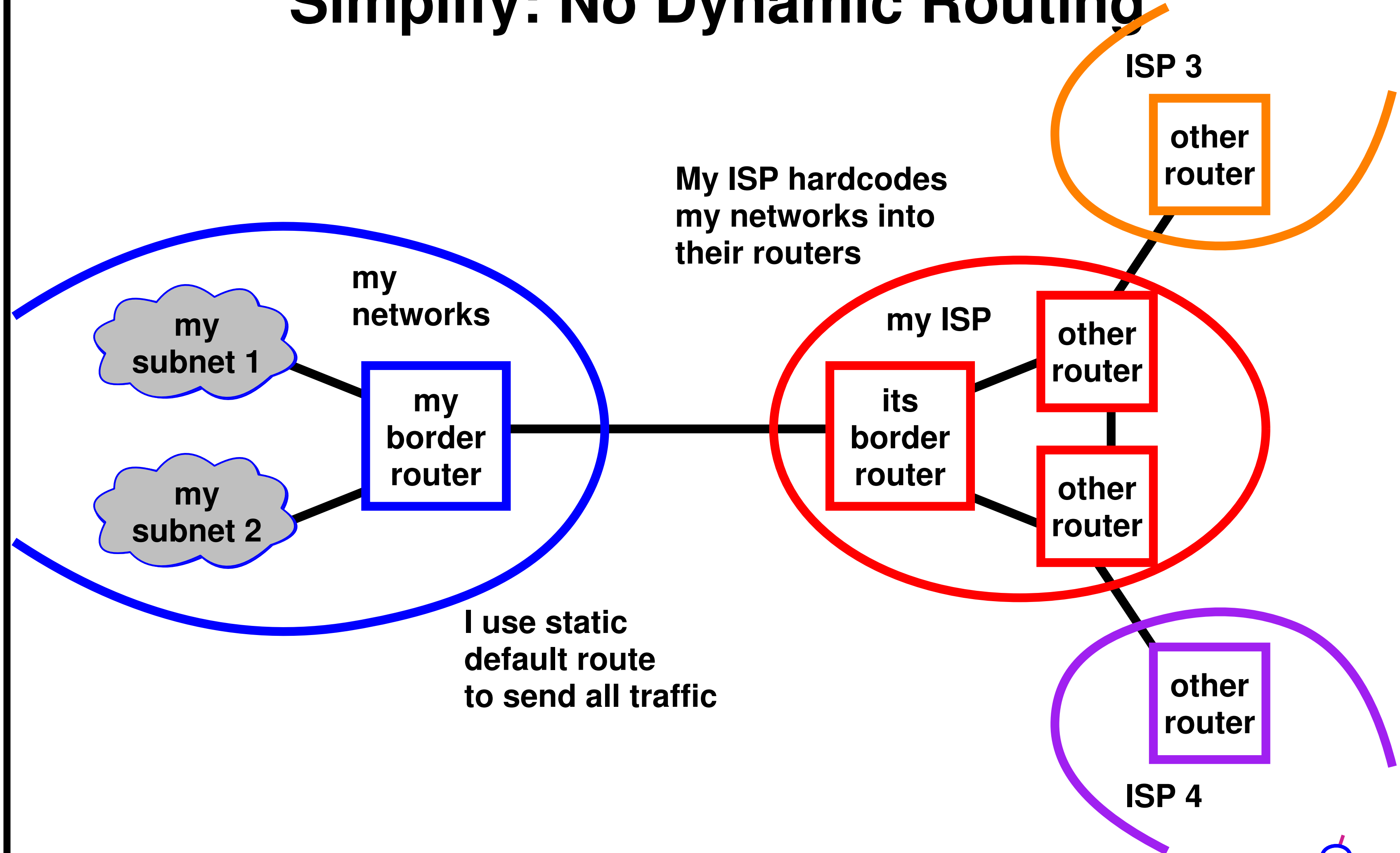
- Mid-80s: EGP
  - reachability protocol (no shortest path)
  - did not accommodate cycles (tree topology)
  - evolved when all networks connected to ARPANET
- Limited size network topology
- Result: BGP introduced as routing protocol
- Today: BGP-4 is the standard, IETF working on BGP-NG

BGP Basic Idea

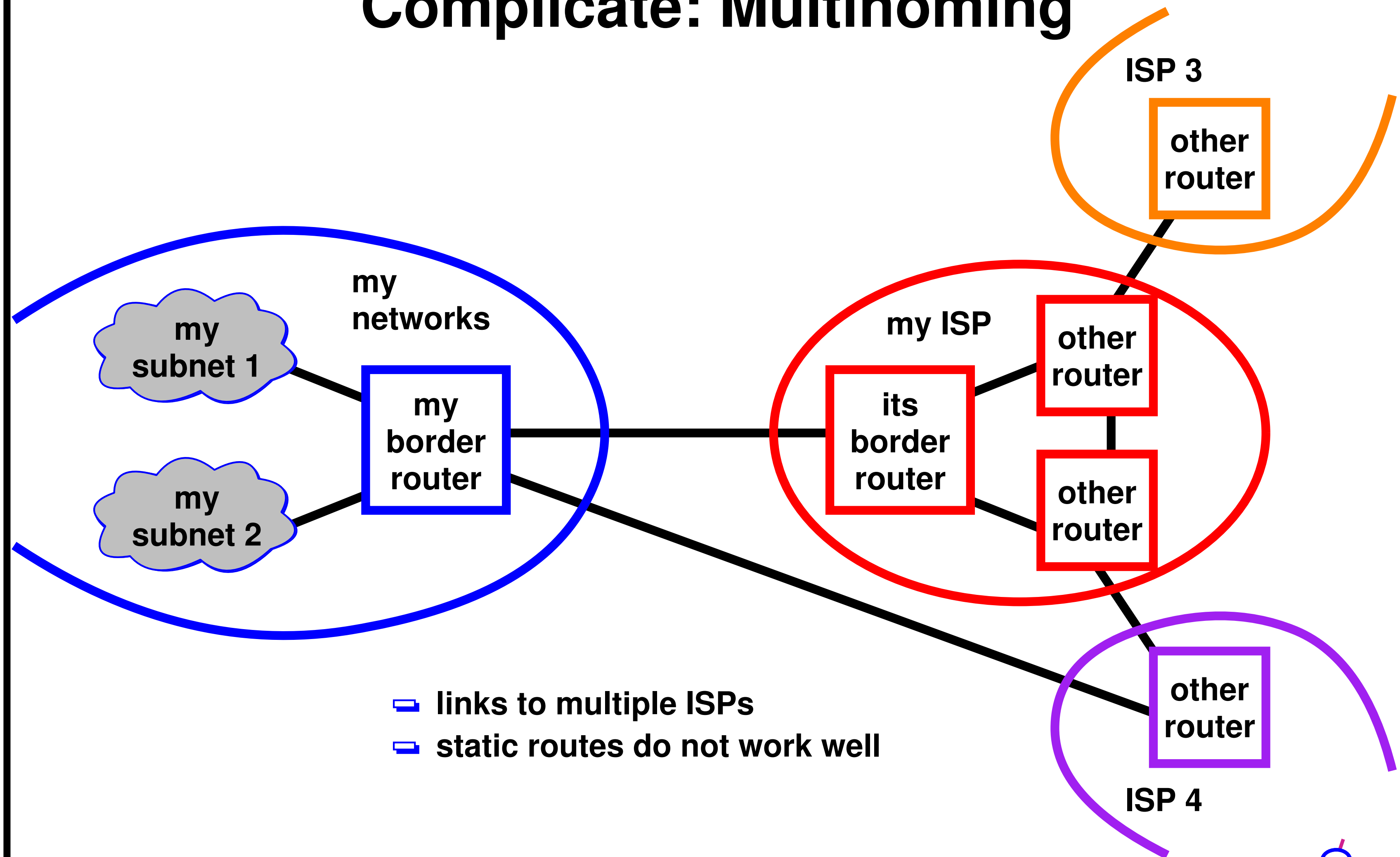


... with dynamic changes and lots of control

Simplify: No Dynamic Routing



Complicate: Multihoming



- links to multiple ISPs
- static routes do not work well

Where And Why BGP?

- Where?
  - multihomed hosts
  - E-BGP for inter-domain routing (between AS's)
  - I-BGP for intra-domain routing (within an AS)
- Why?
  - to deal with dynamics (link failure/recovery)
  - configurable policies on routes

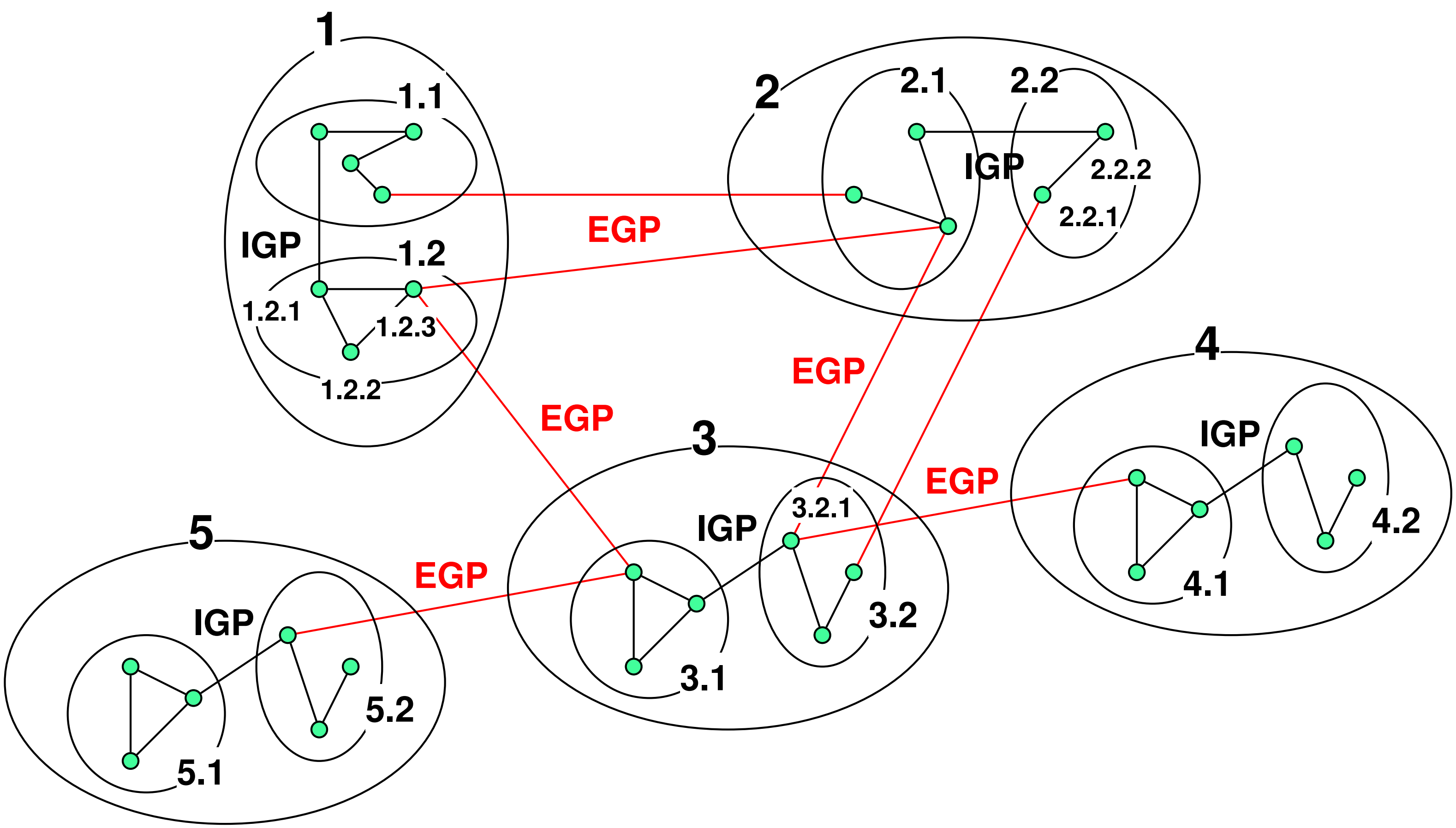
BGP Terminology

- ***AS:*** **autonomous system**
- ***Peer:*** **an adjacent router (Note: not the same meaning as ISP peering)**
- ***Exchange point:*** **place where many ISPs have routers and connections**
- ***RIB:*** **routing information base**
- ***Adj-RIB-In:*** **incoming routing information**
- ***Loc-RIB:*** **local routing information**
- ***Adj-RIB-Out:*** **outgoing routing information**

Autonomous Systems

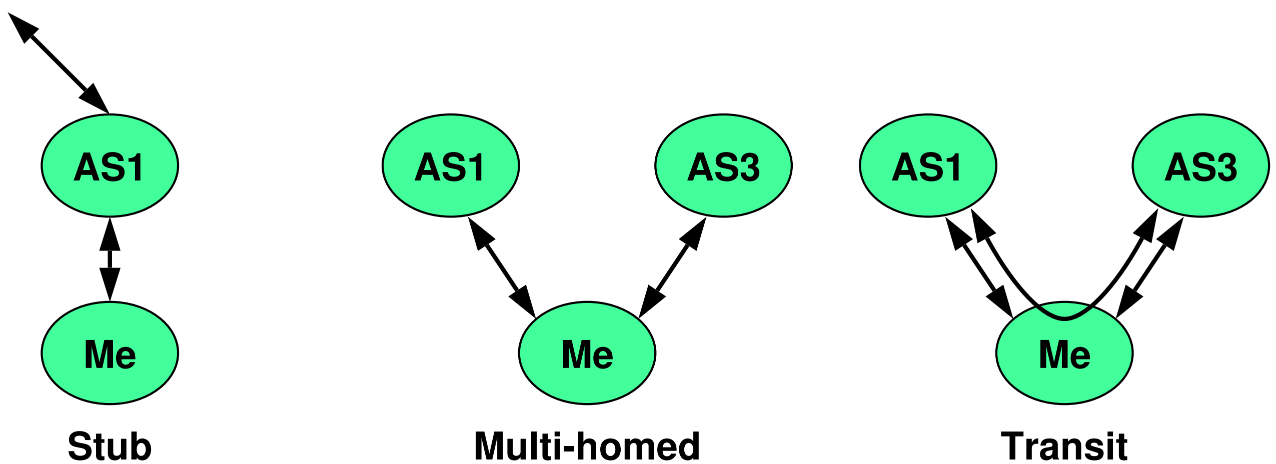
- What is an AS?
  - a set of routers under a single technical administration
  - uses an *interior gateway protocol (IGP)* and common metrics to route packets within the AS
  - uses an *exterior gateway protocol (EGP)* to route packets to other AS's
- AS may use multiple IGPs and metrics, but appears as single AS to other AS's
- Why have both EGP and IGP?
  - know different levels of detail
  - different levels of trust
  - policy issues are much more important in EGP

Example


1
1.1
IGP
1.2
1.2.1
1.2.3
1.2.2
2
2.1
2.2
IGP
2.2.2
2.2.1
EGP
EGP
EGP
3
3.2.1
IGP
3.1
3.2
EGP
4
IGP
4.1
4.2
5
IGP
EGP
5.1
5.2

AS Categories

- ***Stub:*** **an AS that has only a single connection to one other AS - carries only local traffic**
- ***Multi-homed:*** **an AS that has connections to more than one AS, but does not carry transit traffic**
- ***Transit:*** **an AS that has connections to more than one AS, and carries both transit and local traffic (under certain policy restrictions)**



EGP Protocol Choices

- Link state or distance vector?
  - no universal metric - policy decisions
- Problems with distance-vector:
  - Bellman-Ford algorithm slow to converge (counting to infinity problem)
- Problems with link state:
  - metric used by routers in different AS's is not the same - may create loops
  - link state database too large - entire Internet
  - may expose policies to other AS's

Solution: DV with Path Vectors

- Each routing update carries the entire path (AS's appear in the path)
- Loops are detected as follows:
  - when AS gets route check if AS already in path
- if yes, reject route
- if no, add itself and (possibly) advertise route further
- Advantage:
  - metrics are local - AS chooses path, protocol ensures no loops

Interconnecting BGP Peers

- BGP uses TCP to connect peers (port 179)
- Advantages:
  - simplicity from reliability and ordering
  - I-BGP communicate over multiple hops
  - no need for periodic refresh - routes are valid until with-drawn, or the connection is lost (hard state or soft state?)
  - incremental updates
- Disadvantages
  - congestion control on a routing protocol?
- TCP has keepalive option, why BGP keepalive also?
  - *end-to-end argument:* TCP connection can be alive but routing mechanism can be hung, must have BGP keepalive

Hop-by-hop Model

- BGP advertises to neighbors only those routes that it uses
  - consistent with the hop-by-hop Internet paradigm
  - e.g., AS1 cannot tell AS2 to route to other AS's in a manner different than what AS2 has chosen (need source routing for that)

Protocol Observations

- How does BGP know when a link is down/out?
  - timeout (hold time)
  - see [Shaihk00a]
- How does BGP avoid looping paths?
  - path vector via AS_PATHS
  - loop detection on route receipt (or transmission [Labovitz00a])

BGP Messages

- ***OPEN:*** **sets up timeout, AS, id, etc.**
- ***UPDATE:*** **update (inject, withdraw) routes with attributes**
- ***NOTIFICATION:*** **error reporting**
- ***KEEPALIVE:*** **no change, but link is up**
  - **TCP has keepalive option, why BGP keepalive also?**
    - **end-to-end argument: TCP connection can be alive but routing mechanism can be hung, must have BGP keepalive**

BGP Attributes

- ORIGIN: where prefix orginates
- AS_PATH: path for routing
- NEXT_HOP: where to send data
- MULTI_EXIT_DISCRIMINATOR: used to influence multi-homing

- Why BGP Attributes?
  - want to do policy routing
  - want some way to prevent looping in DV routing (AS_PATH)
  - flexibilities (allows extensibility)

Policy With BGP

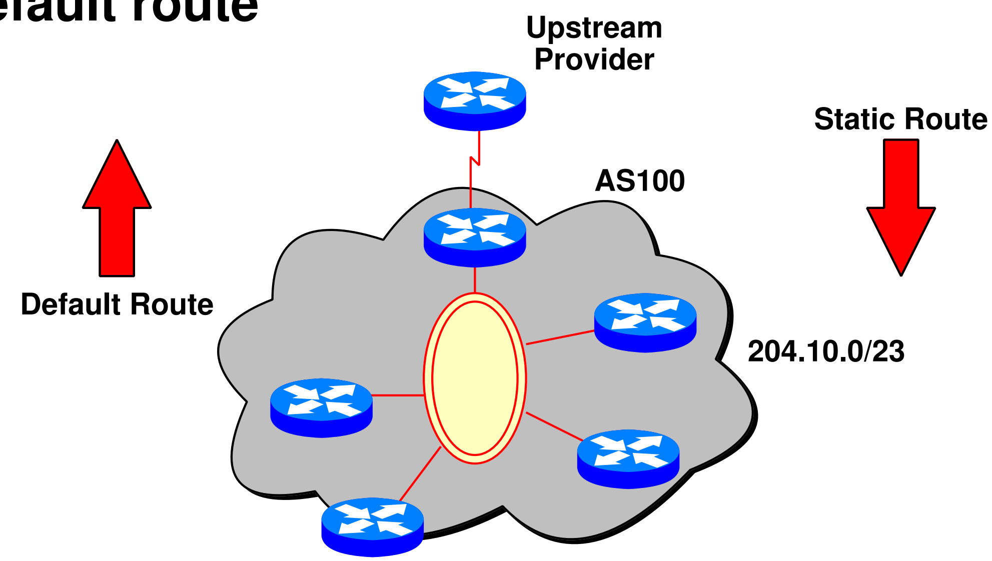
- **BGP provides capability for enforcing various policies**
- **Policies are not part of BGP (no policy messages)**
  - **they are provided to BGP as configuration information**
- **BGP enforces policies by *choosing paths from multiple alternatives* and *controlling advertisement to other AS's***

Examples of BGP Policies

- A multi-homed AS refuses to act as transit
  - limit path advertisement
- A multi-homed AS can become transit for some AS's
  - only advertise paths to some AS's
- An AS can favor or disfavor certain AS's for traffic transit from itself

BGP Is NOT Needed If:

- Single homed network (stub)
- AS does not provide downstream routing
- AS uses a default route



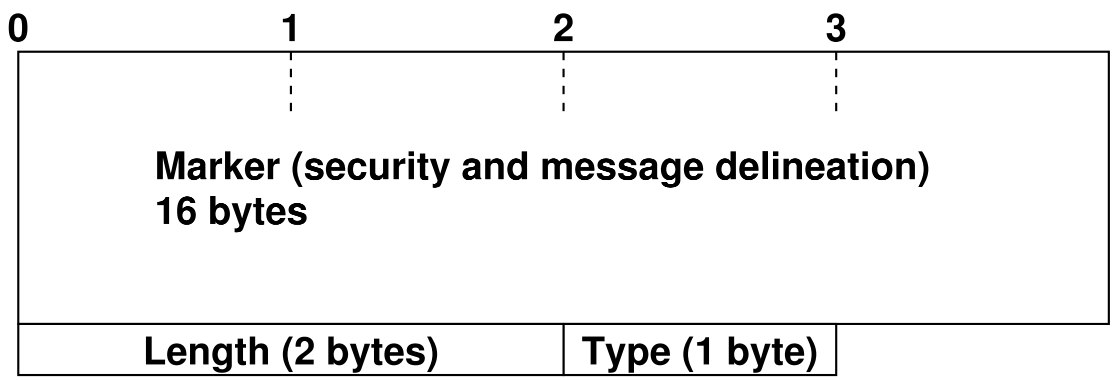
BGP-4

- Latest version of BGP
- BGP-4 supports CIDR

Routing Information Bases (RIB)

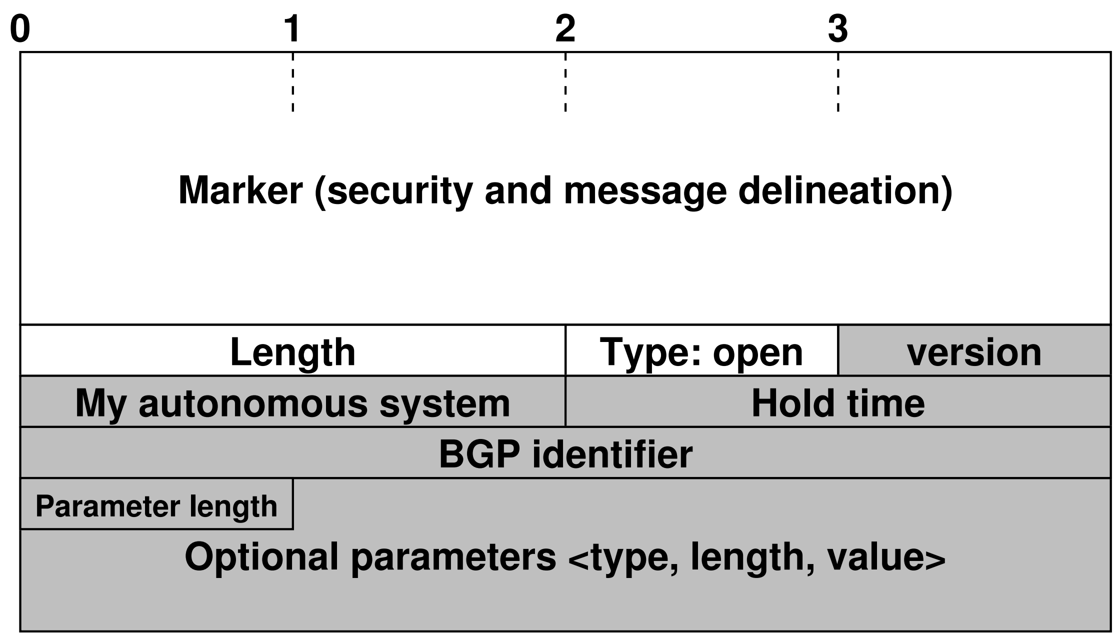
- **Routes are stored in RIBs**
- ***Adj-RIBs-In:*** **routing info that has been learned from other routers (unprocessed routing info)**
- ***Loc-RIB:*** **local routing information selected from Adj-RIBs-In (routes selected locally)**
- ***Adj-RIBs-Out:*** **info to be advertised to peers (routes to be advertised)**

BGP Common Header



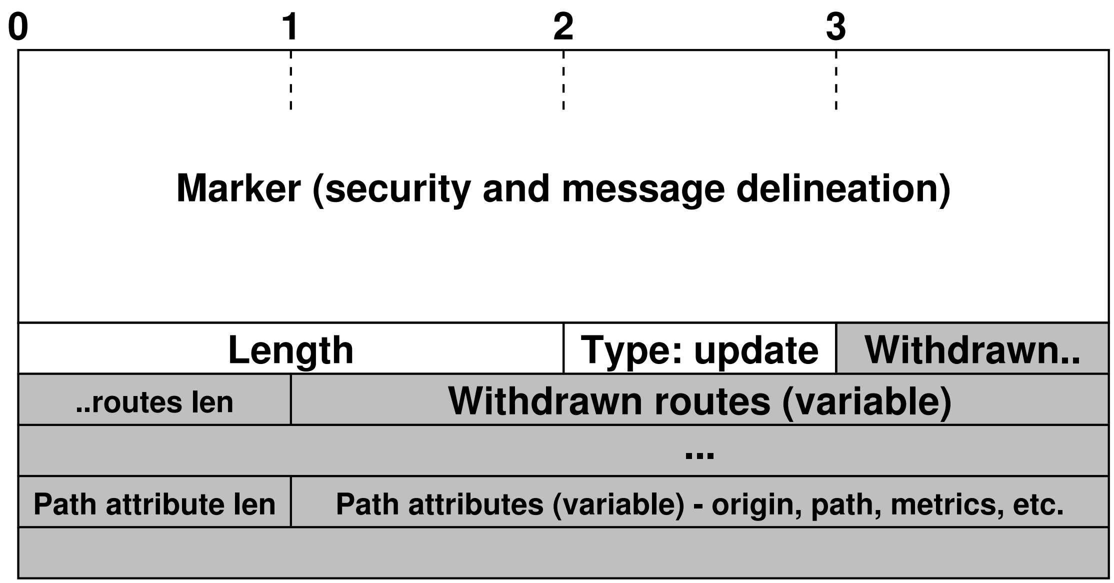
Types: OPEN, UPDATE, NOTIFICATION, KEEPALIVE

BGP OPEN Message



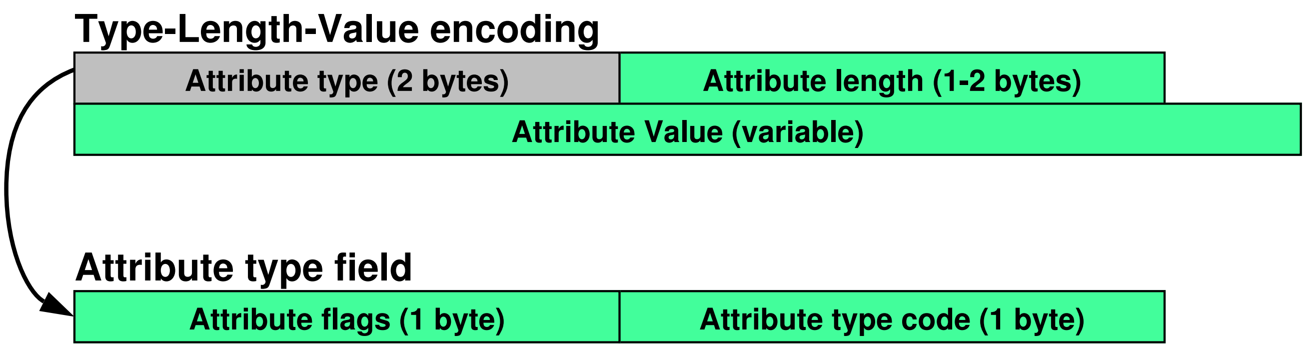
- My autonomous system: ID assigned to that AS
- Hold timer: max interval between KEEPALIVE or UPDATE messages
- BGP ID: address of one (typically virtual) interface and is same for all messages

BGP UPDATE Message



- UPDATE message may report multiple withdrawn routes.
- Many prefixes may be included in UPDATE, but must share same attributes.

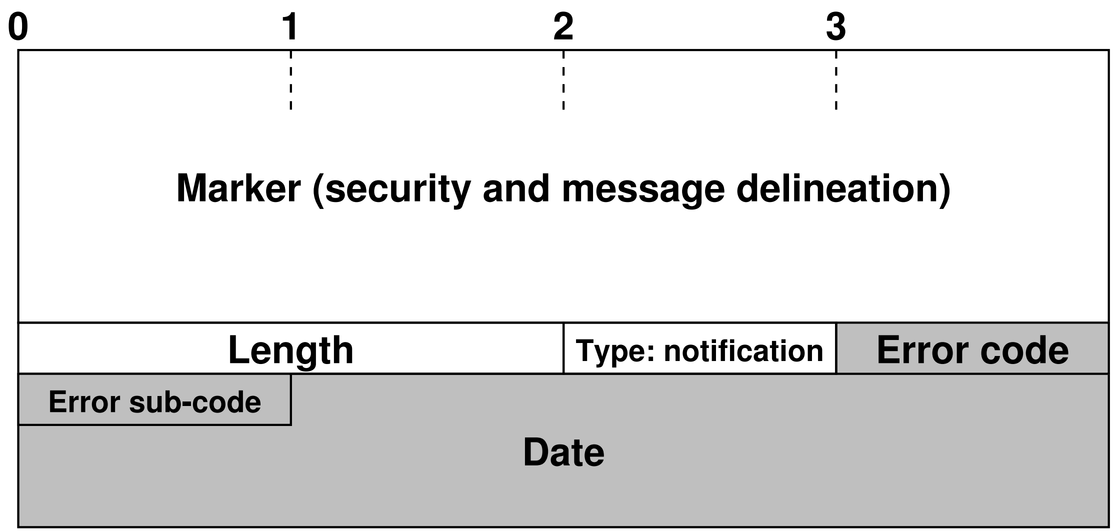
Path Attributes



Flags: optional v.s. well-known
transitive v.s. non-transitive (passed on)
partial (someone in path did not understand this attribute)
extended length (2 bytes instead of 1)

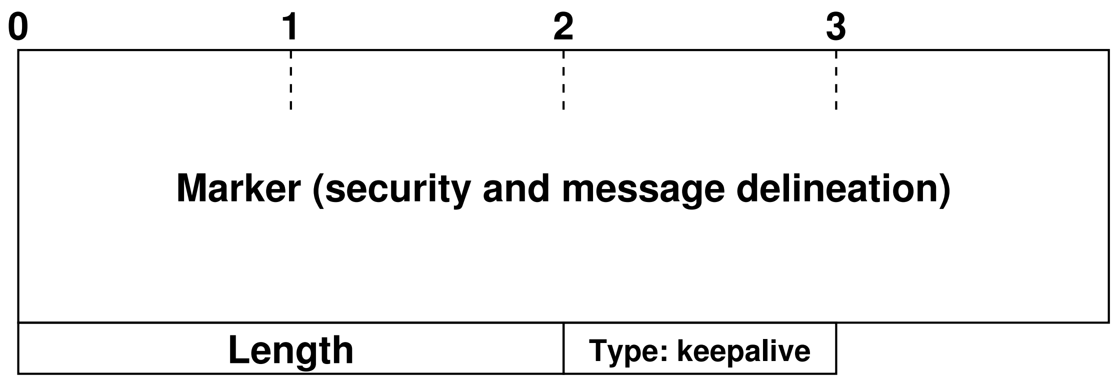
Attribute types: Origin, AS_PATH, Next_Hop (more later..)

BGP NOTIFICATION Message



- Used for error notification (update error, expired timer, FSM, cease)
- TCP connection is closed immediately after notification.

BGP KEEPALIVE Message



- Sent periodically (but before hold timer expires) to peers to ensure connectivity.
- Sent in place of an UPDATE message.

Note: hold_time = zero means no keepalives will be sent

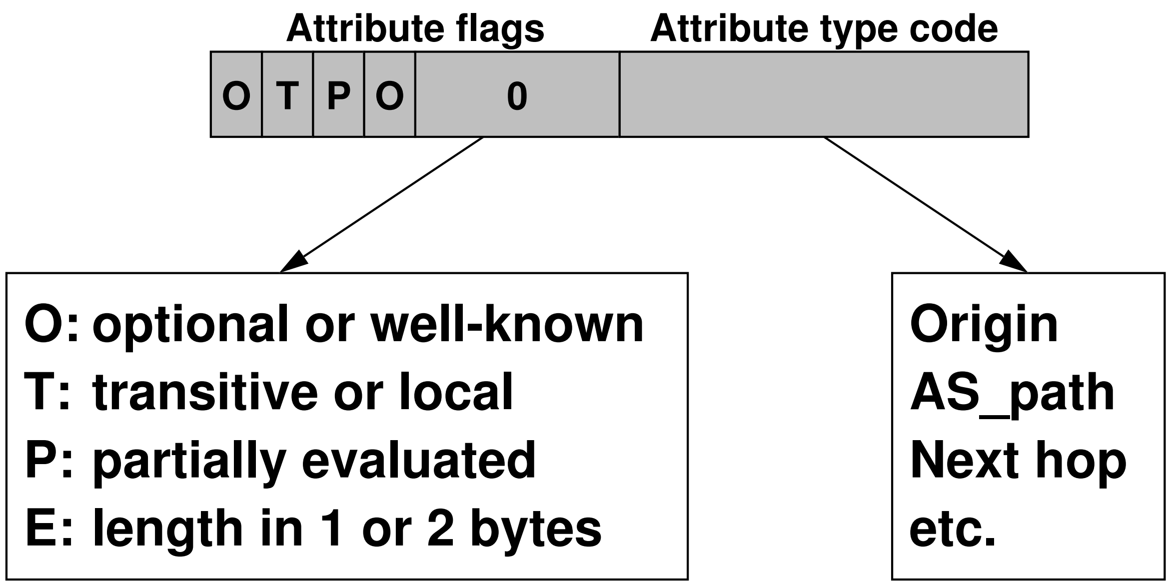
Path Selection Criteria

- Information based on path attributes
- Attributes + external (policy) information
- Examples:
  - hop count
  - policy considerations
  - presence or absence of certain AS
  - path origin (EGP, IGP)
  - link dynamics (flapping, stable)

Path Attributes

- Categories (recall flags):
  - well-known mandatory (passed on)
  - well-known discretionary (passed on)
  - optional transitive (passed on)
  - optional non-transitive (if unrecognized, not passed on)
- Optional attributes allow for BGP extensions

Path Attribute Message Format (Repeated)



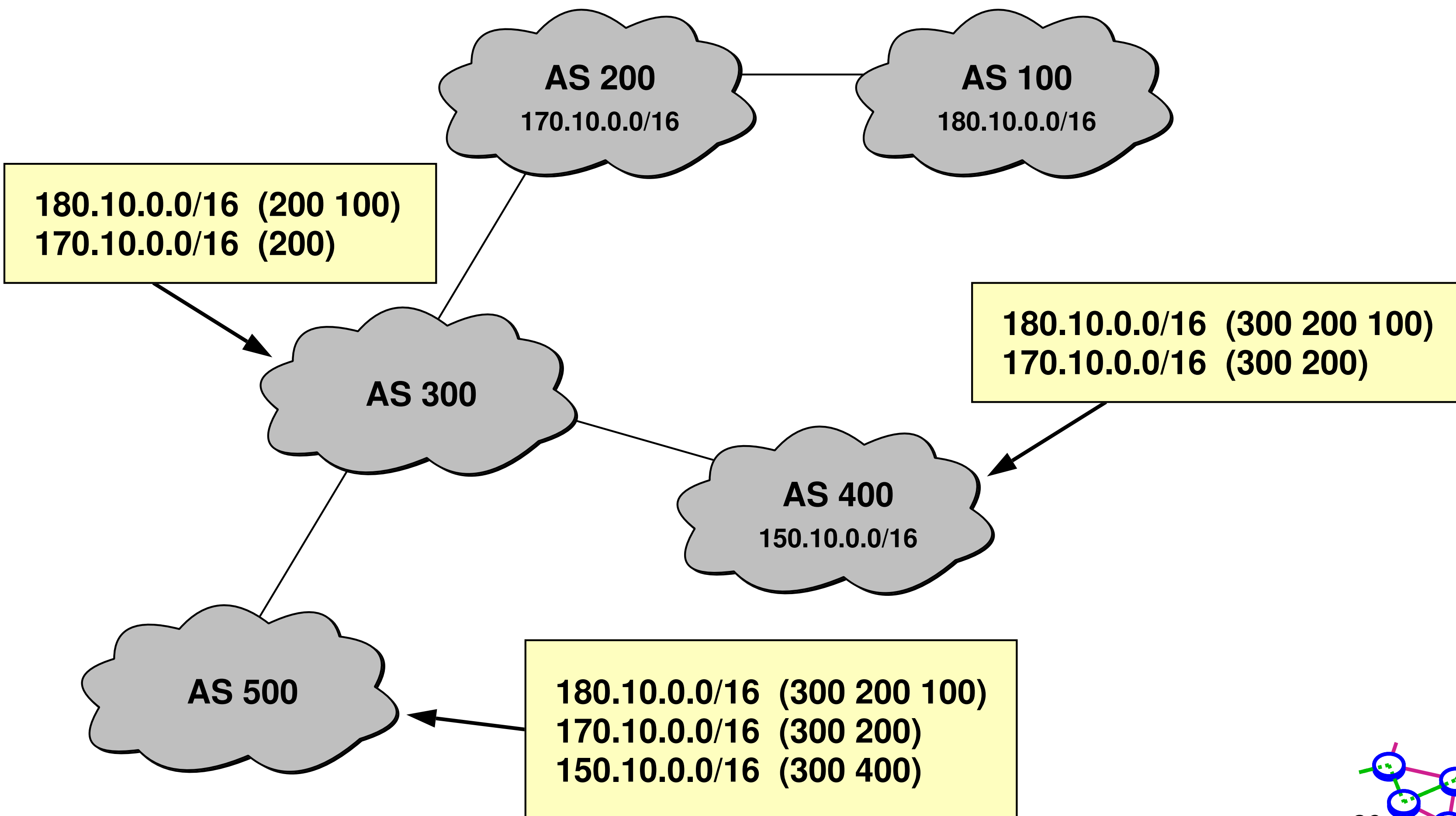
ORIGIN Path Attribute

- Well-known, mandatory attribute
- Describes how a prefix was generated at the origin AS. Possible values:
  - *IGP:* prefix learned from IGP
  - *EGP:* prefix learned through EGP
  - *INCOMPLETE:* none of the above (often seen for static routes)

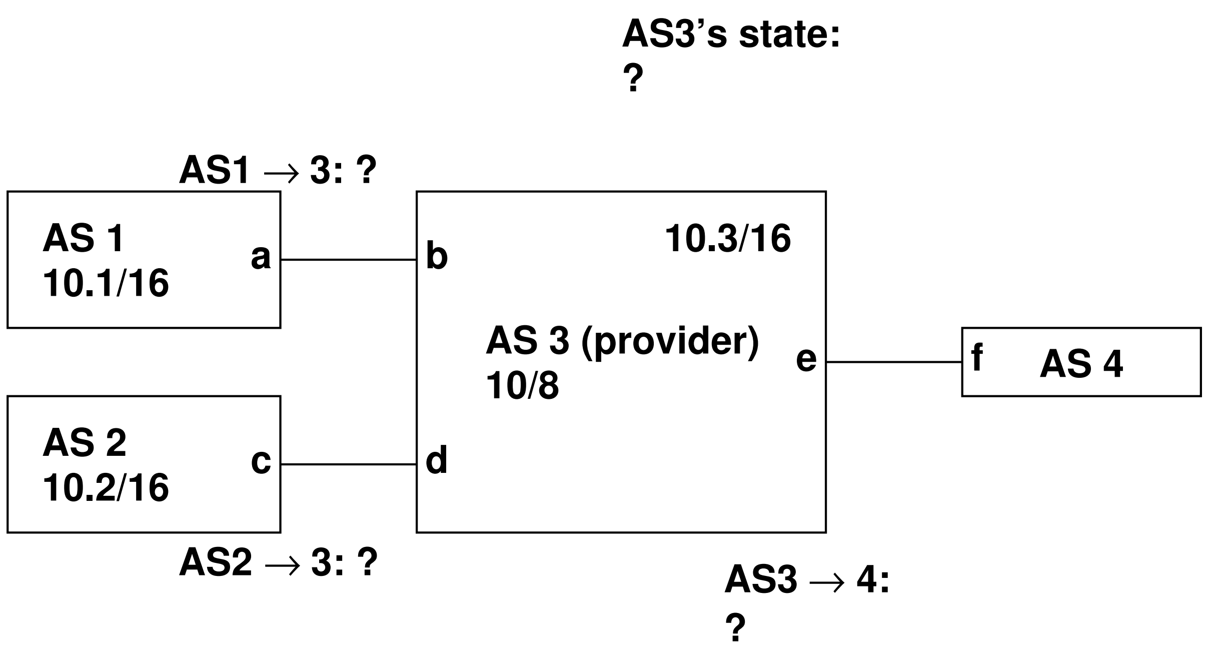
AS_PATH Attribute

- Well-known, mandatory attribute
- Important components:
  - list of traversed AS's
- If forwarding to internal peer:
  - do not modify AS_PATH attribute
- If forwarding to external peer:
  - prepend self into the path

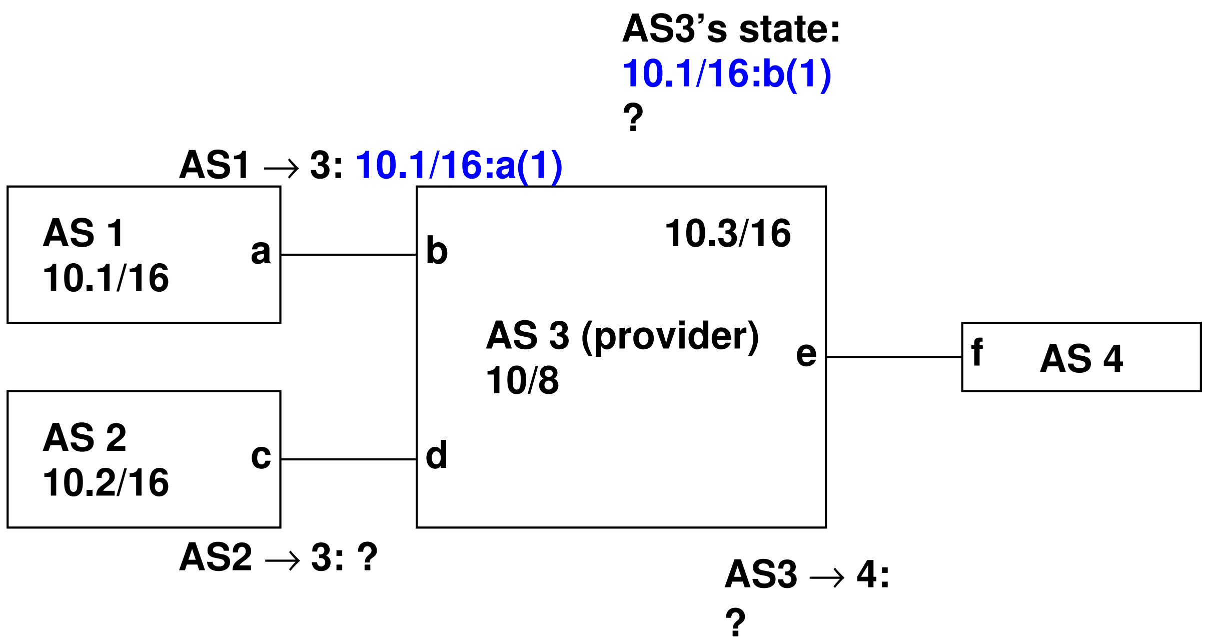
AS_PATH Attribute



CIDR and BGP


AS3's state:
?
AS1 → 3: ?
AS 1
10.1/16
a
b
10.3/16
AS 3 (provider)
10/8
e
f
AS 4
AS 2
10.2/16
c
d
AS2 → 3: ?
AS3 → 4:
?

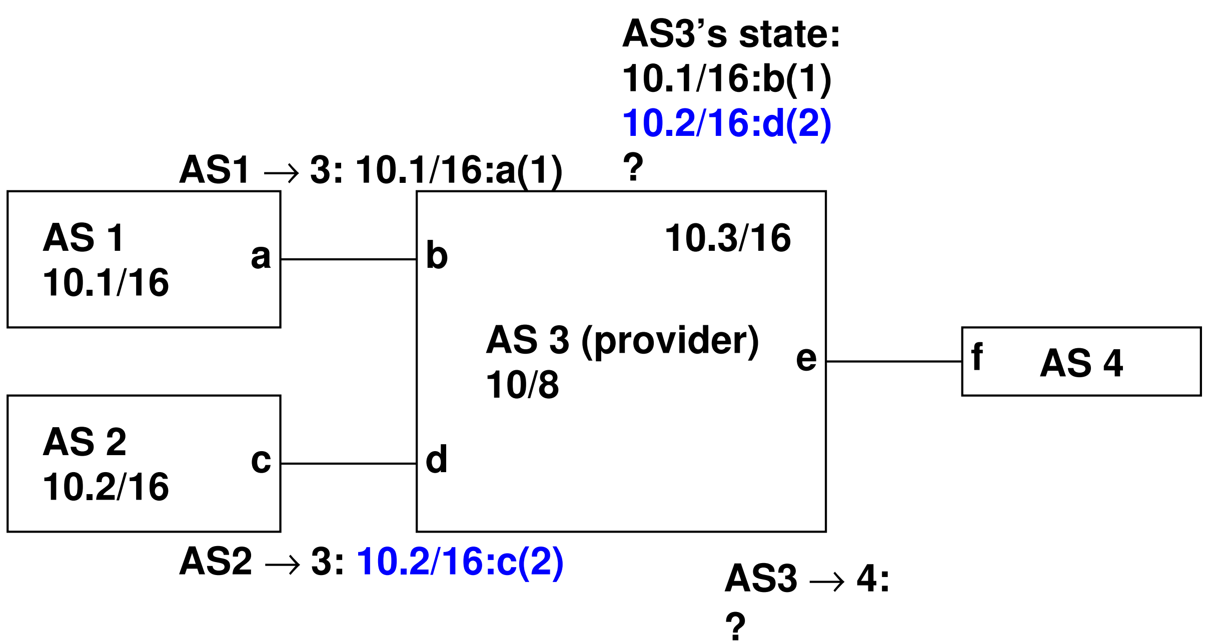
CIDR and BGP



Notation:

- export: destination network:exit router(AS_PATH)
- state: destination network:entry router(AS_PATH)

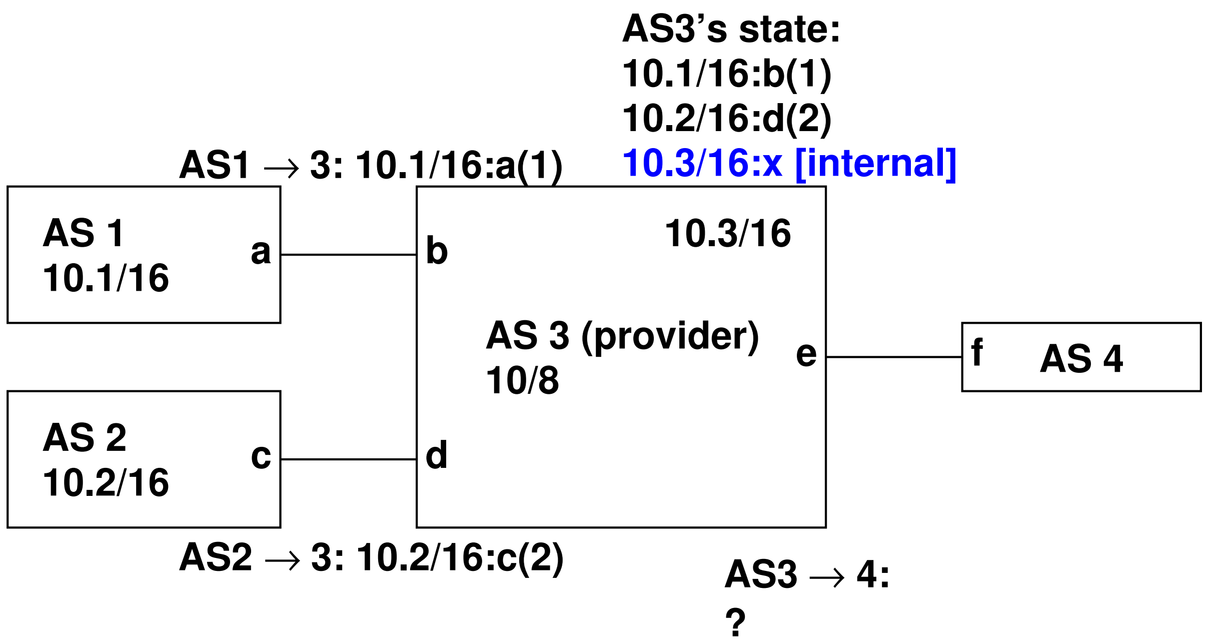
CIDR and BGP



Notation:

- export: destination network:exit router(AS_PATH)
- state: destination network:entry router(AS_PATH)

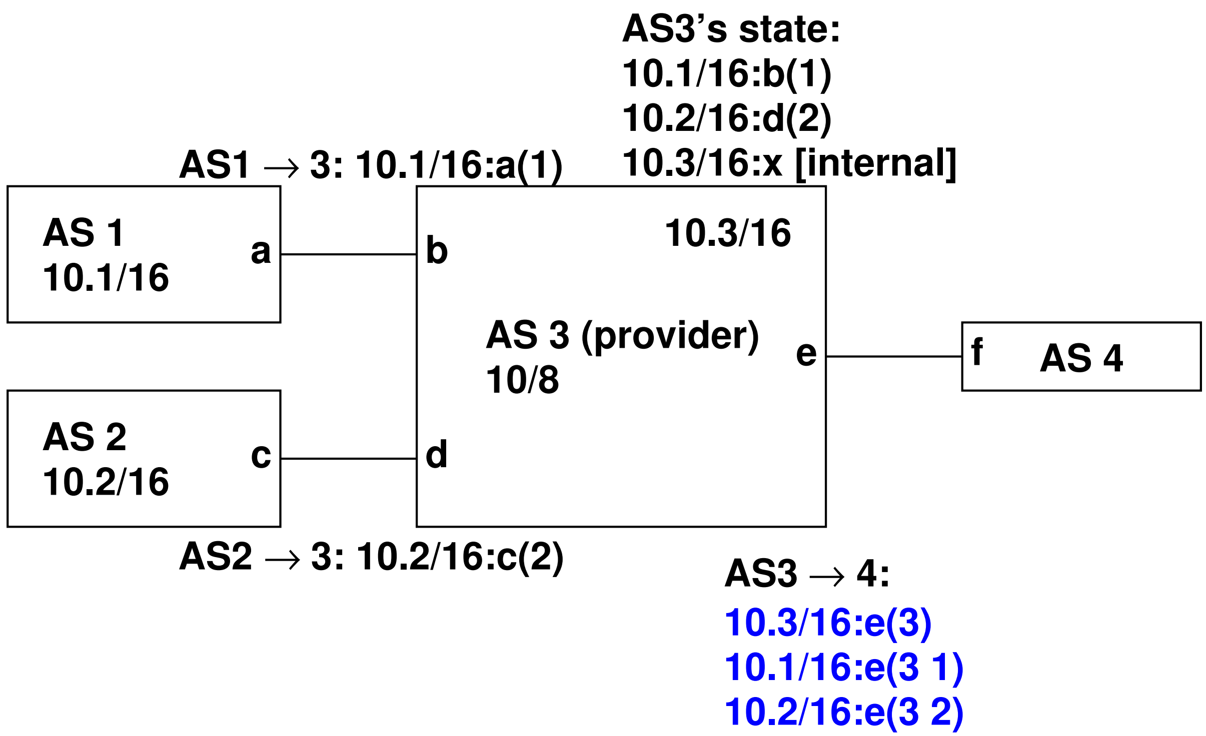
CIDR and BGP



Notation:

- export: destination network:exit router(AS_PATH)
- state: destination network:entry router(AS_PATH)

CIDR and BGP

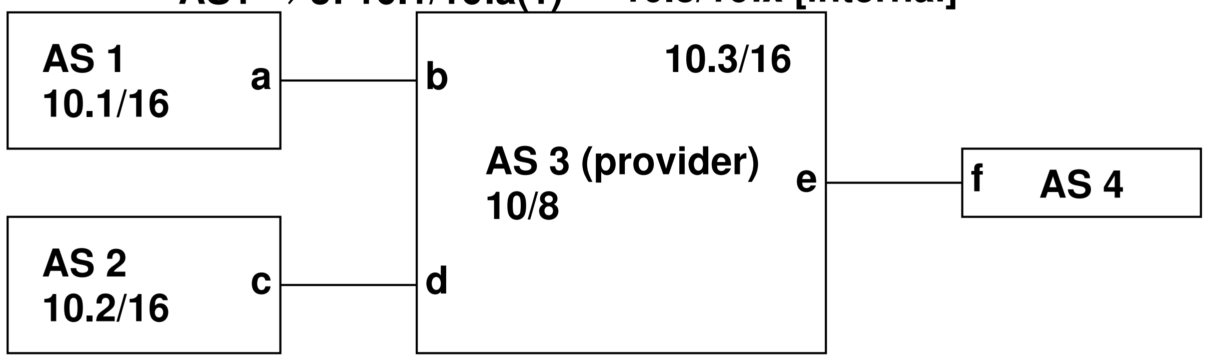


- too many entries in routing tables
- need aggregation for scalability
- also need enough information in AS-PATH for loop detection
- how?

CIDR and BGP

AS3's state:
10.1/16:b(1)
10.2/16:d(2)
10.3/16:x [internal]

AS1 → 3: 10.1/16:a(1)



AS2 → 3: 10.2/16:c(2)

AS3 → 4:
10.3/16:e(3)
10.1/16:e(3 1)
10.2/16:e(3 2)

⇩ aggregation with *set notation*

AS3 → 4:
10/8:e(3 [1 2])

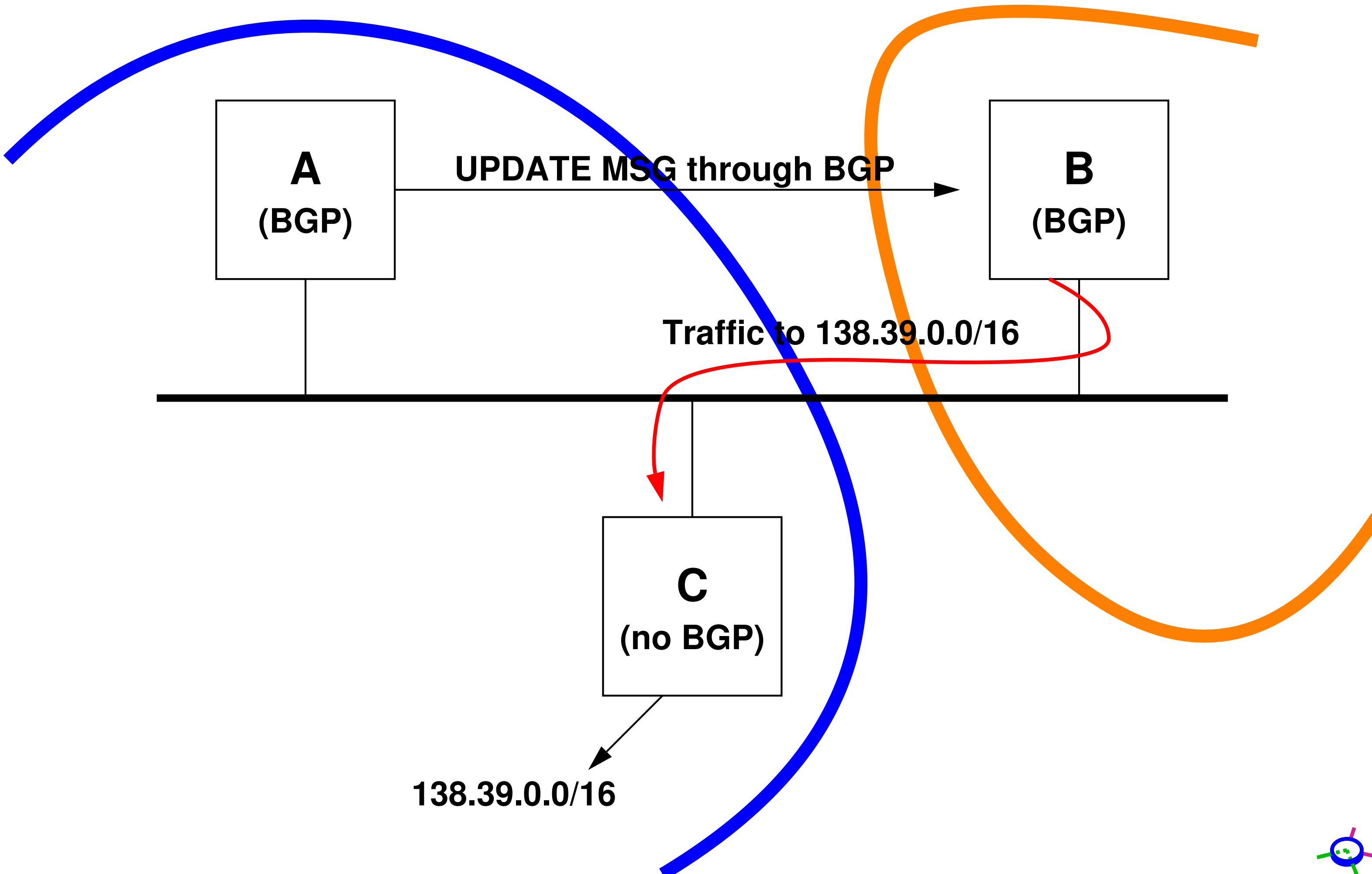
Sets and Sequences

- Solution: restructure AS-PATH attribute as:
  - path: (Sequence (3), Set (1, 2))
- if AS4 wants to advertise path:
  - Path: (Sequence (4, 3), Set (1, 2))
- In practice used only if paths in set have same attributes

NEXT-HOP Path Attribute

- Well-known, mandatory attribute
- NEXT-HOP: IP address of border router to be used as next hop
- Usually, next hop is the router sending the UPDATE message
- Useful when some routers do not speak BGP

Example of NEXT-HOP



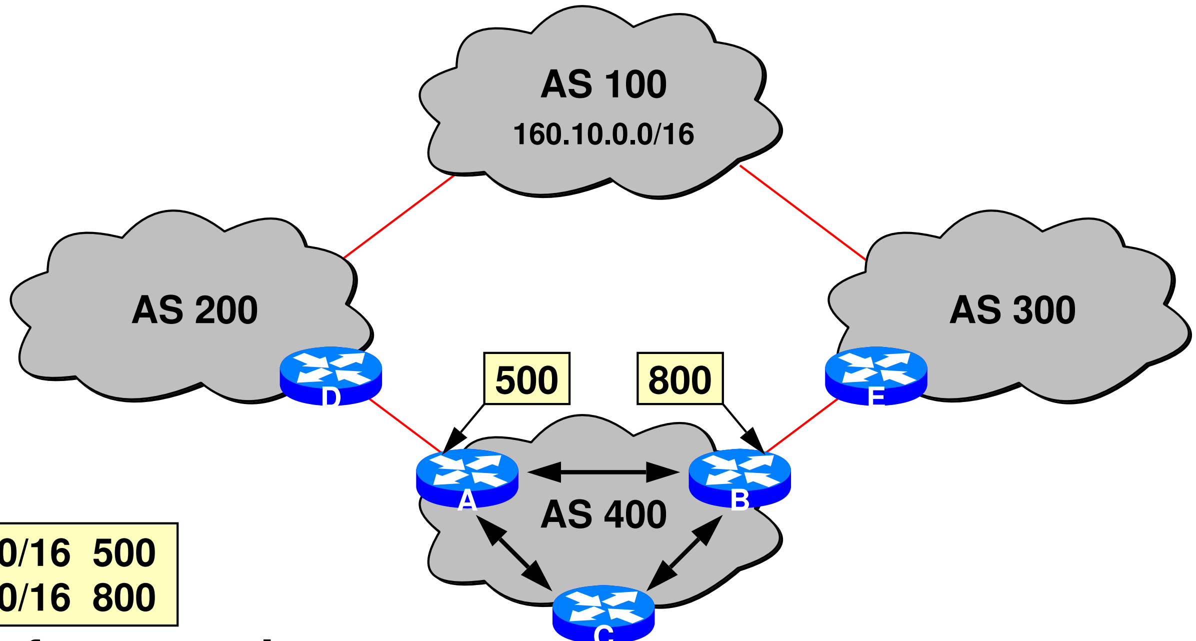
BGP Attributes For Policy Control

- So, why policy routing?
  - business relationships
  - control (optimize) routes
  - multi-homing: control traffic over multiple links

Policy 1: LOCAL-PREF Path Attribute

- Well-known, discretionary
- Provided by a BGP router to all other internal BGP routers
  - denotes degree of preference for each destination
- From local configuration
  - affects *your* AS only
  - (does not propagate to others)
  - can influence any prefixes
- Pick with path to prefer for a prefix
- Rule: *BGP prefers paths with higher LOCAL-PREF*

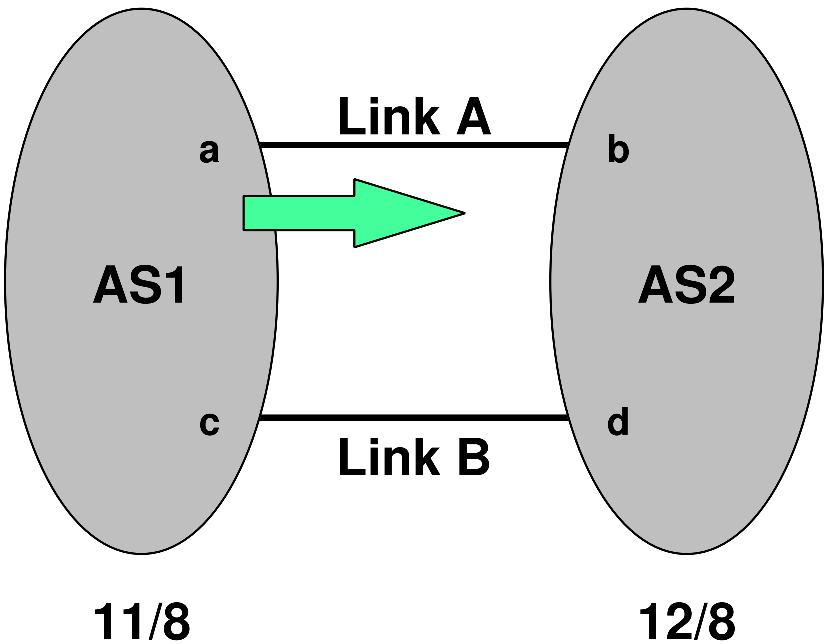
LOCAL-PREF



	160.10.0.0/16	500
>	160.10.0.0/16	800

Higher preference wins

LOCAL-PREF Example 1

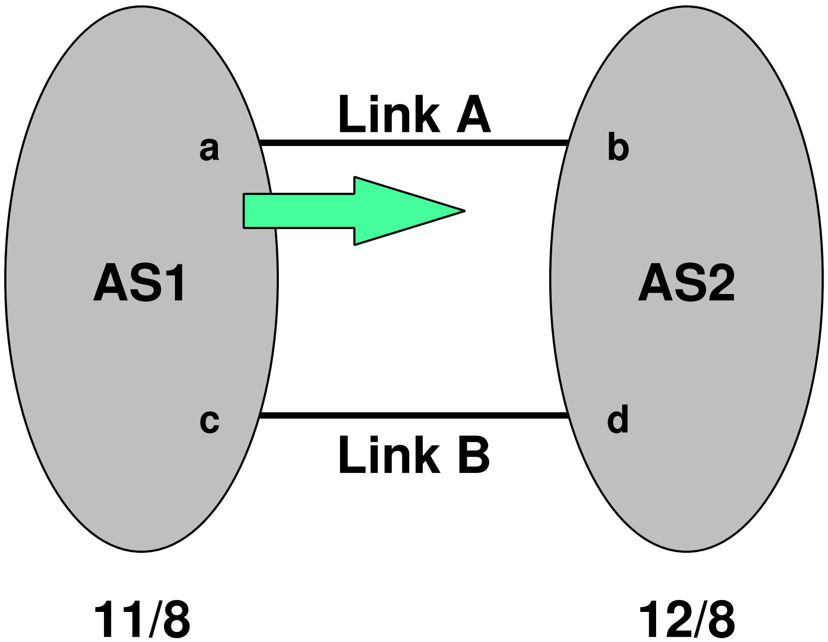


You are AS1 with two links A & B to AS2. How to force all traffic to AS2's prefix 12/8 through link A?

AS1's routing table:

?

LOCAL-PREF Example 1



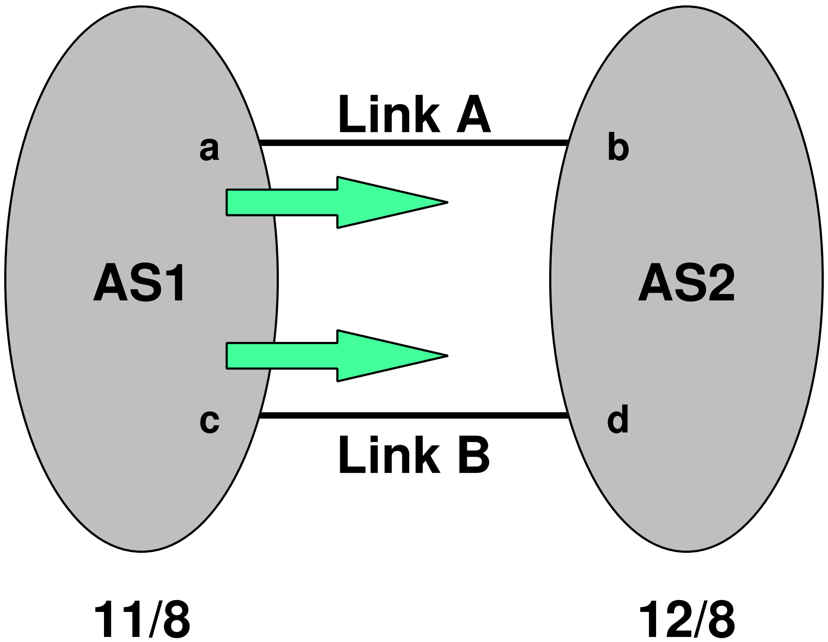
You are AS1 with two links A & B to AS2. How to force all traffic to AS2's prefix 12/8 through link A?

AS1's routing table:

12/8:a(2) w/LP=10

12/8:c(2) w/LP=5

LOCAL-PREF Example 2

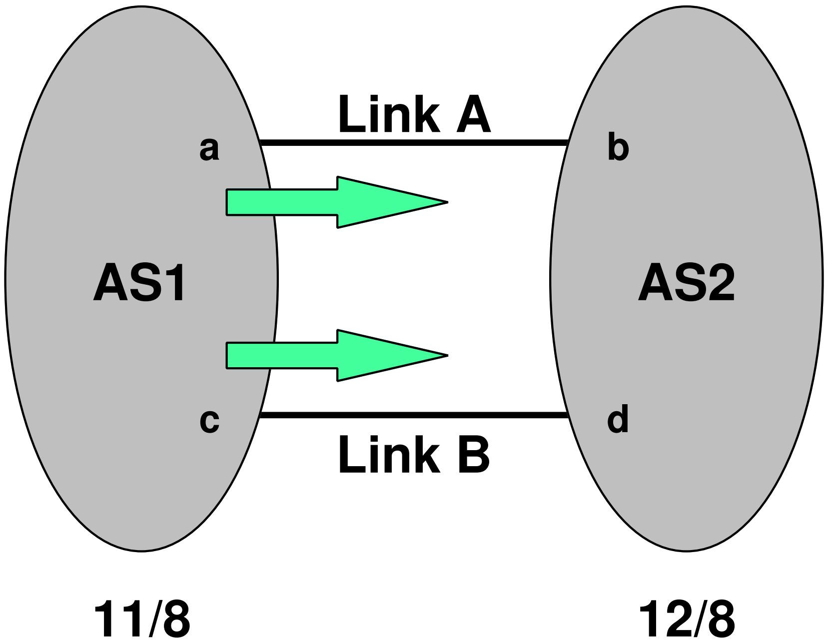


You are AS1 with two links A & B to AS2. How to load-share AS2-bound traffic between links A & B?

AS1's routing table:

?

LOCAL-PREF Example 2



You are AS1 with two links A & B to AS2. How to load-share AS2-bound traffic between links A & B?

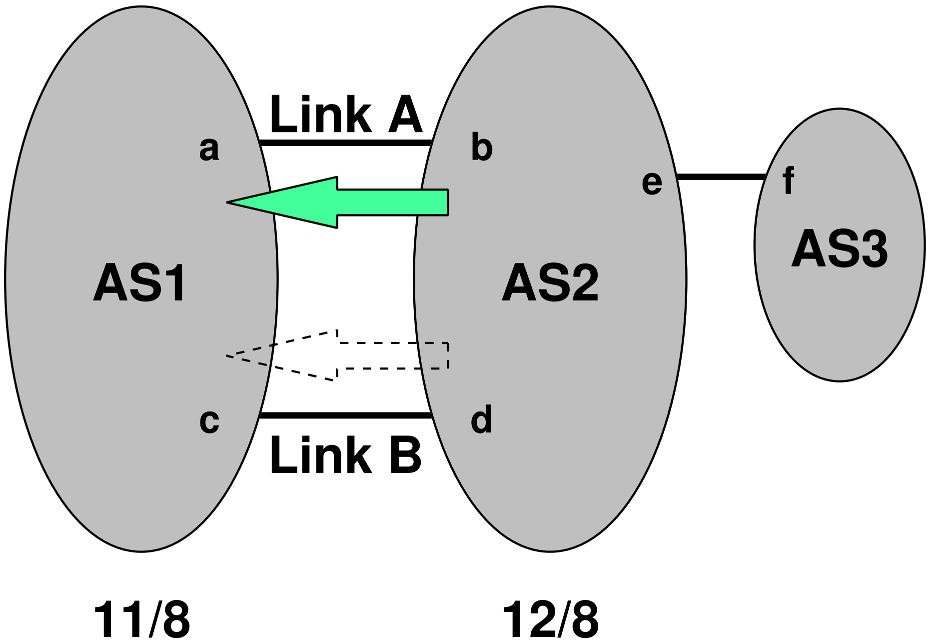
AS1's routing table:

12.0/9:a(2) w/LP 10
12.0/9:c(2) w/LP 5
12.128/9:a(2) w/LP 5
12.128/9:c(2) w/LP 10

Policy 2: AS-PATH Inflation

- From local configuration
  - affects *all* AS's in the Internet
  - affects only your prefixes
- Make a path look worse than it is
- Rule: *BGP prefers shorter AS-PATHs*

AS-PATH Inflation Example



You are AS1 with two links A & B to AS2. How to make link A primary and B backup for incoming traffic?

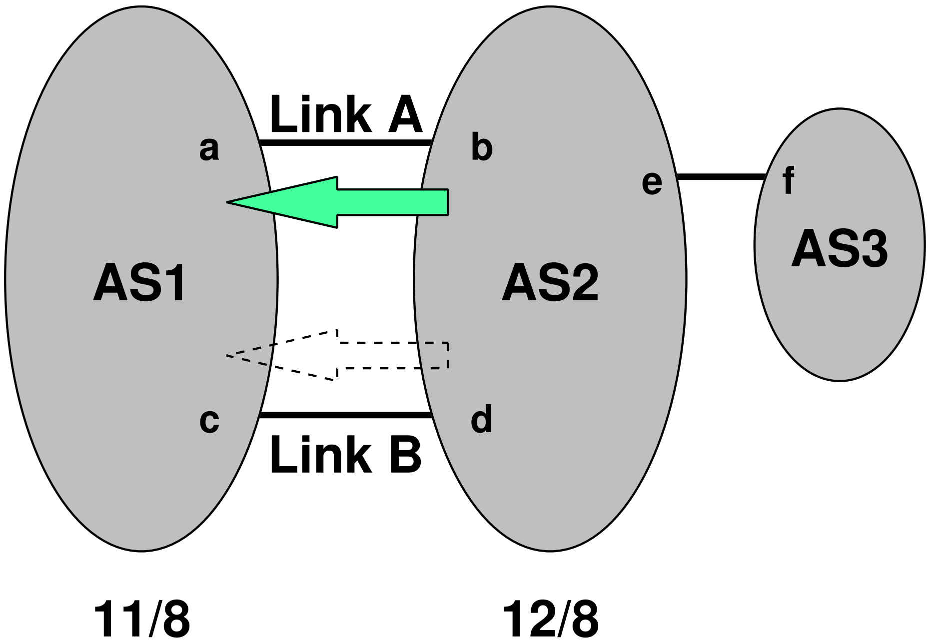
AS1 exports:

?

AS2's state:

?

AS-PATH Inflation Example



You are AS1 with two links A & B to AS2. How to make link A primary and B backup for incoming traffic?

AS1 exports:

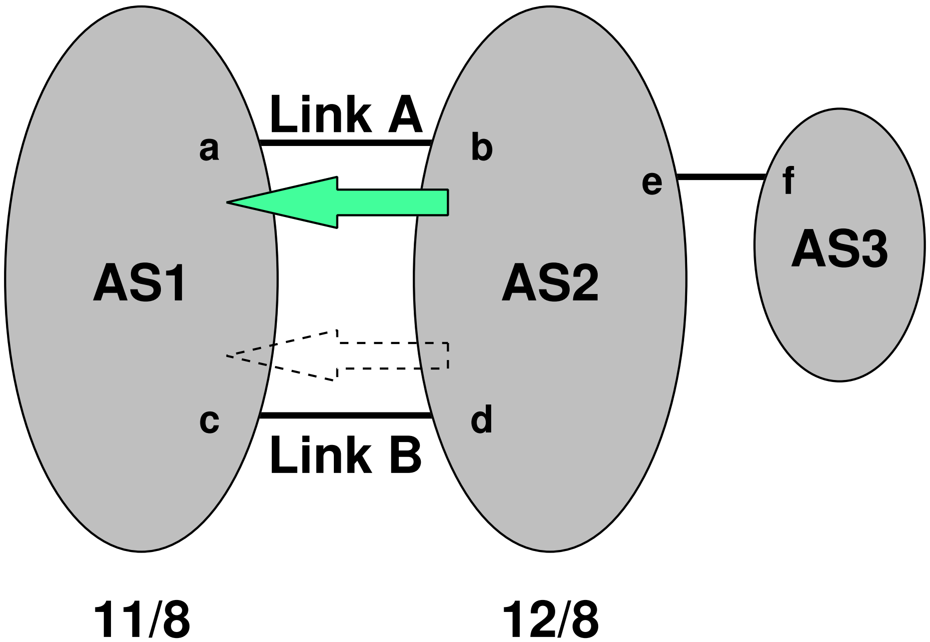
11/8:a(1)

11/8:c(1 1)

AS2's state:

?

AS-PATH Inflation Example



You are AS1 with two links A & B to AS2. How to make link A primary and B backup for incoming traffic?

AS1 exports:

11/8:a(1)

11/8:c(1 1)

AS2's state:

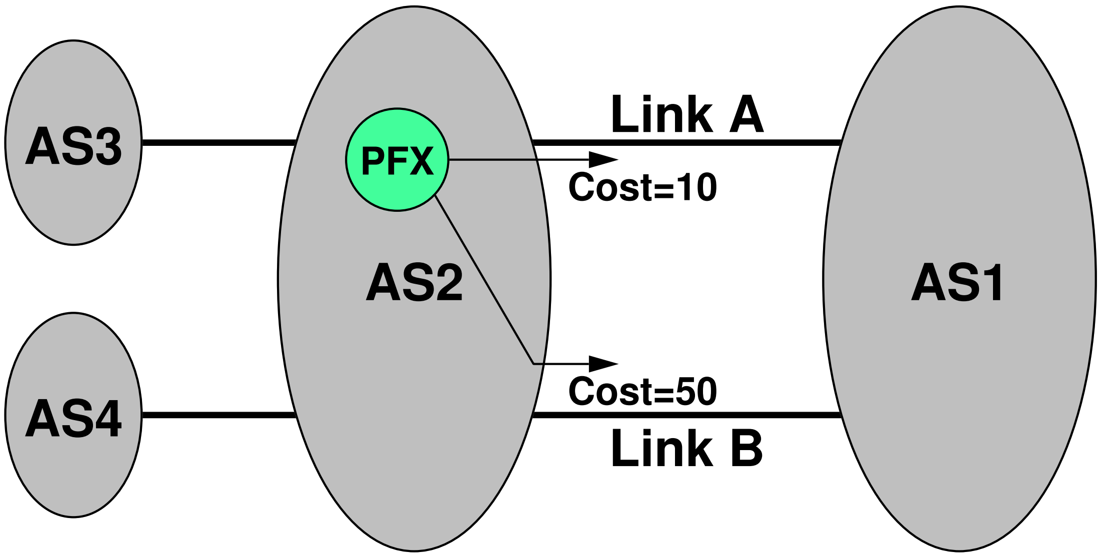
11/8:b(1) ← **prefers this**

11/8:d(1 1)

Policy 3: Multi-exit Discriminator (MED) Path Attribute

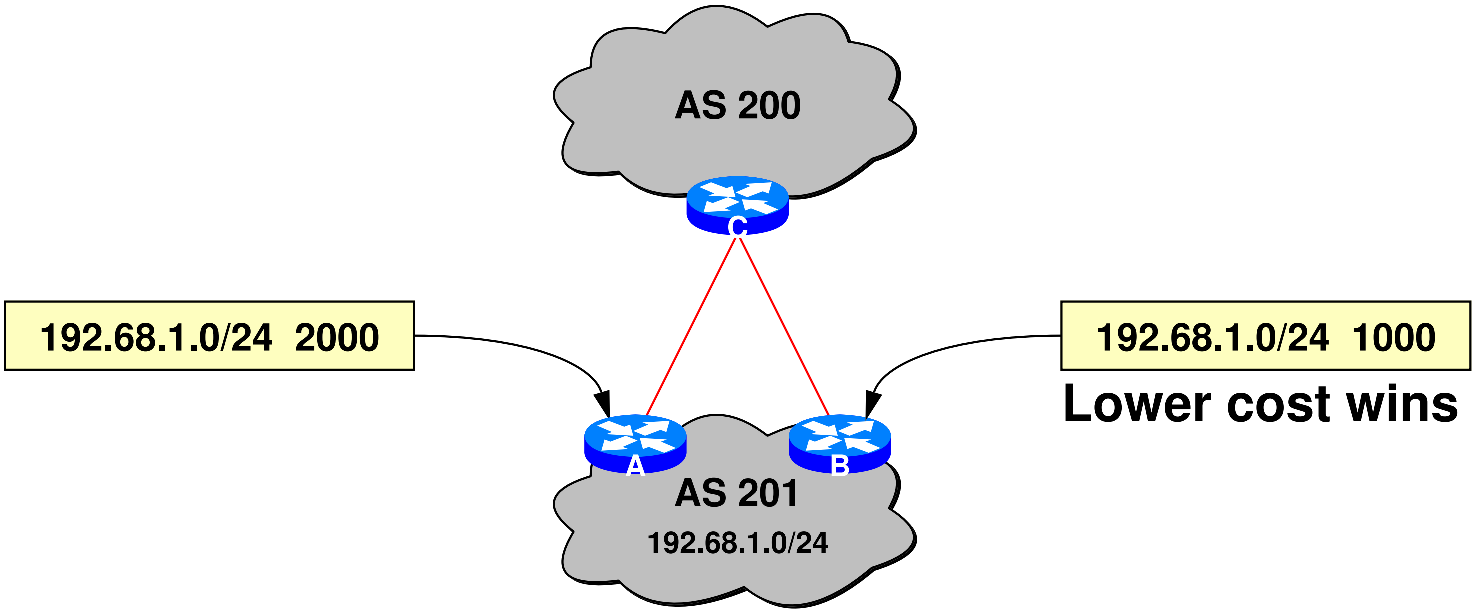
- Optional, non-transitive attribute
- Used when two AS's connect to each other in more than one place
- Carries a metric expressing degree of preference
- From local configuration
  - affects prefixes you propagate
  - affects *adjacent* AS's
- Used to help others pick the right exit point
  - therefore they probably trust you (e.g., client/provider relationship)
- Rule: *BGP prefers the lowest MED*

MED



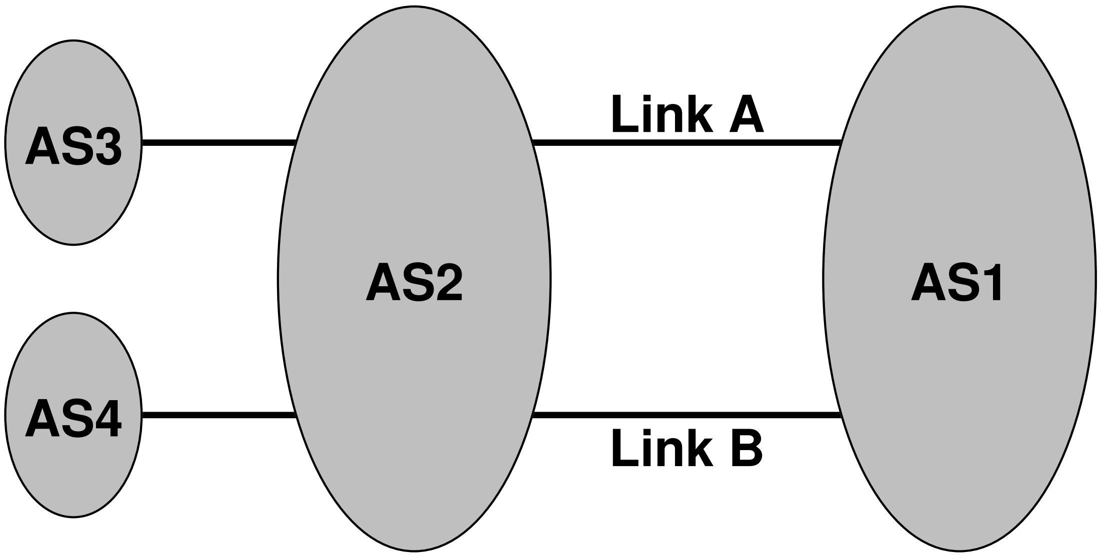
- AS2 includes MEDs with prefixes sent to AS1 over Links A and B.
- AS1 uses these to select appropriate link when sending to PFX.

...MED



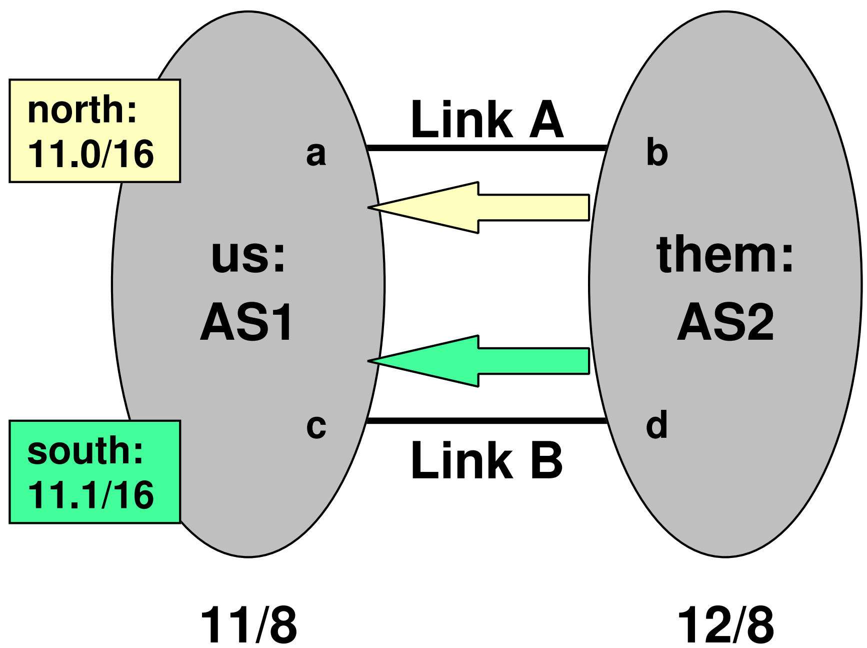
Lower cost wins

MED (Cont...)



- AS2 can use MED to instruct AS1 to use link A for traffic to AS3, and link B for traffic to AS4.
- How is this done?

MED Example

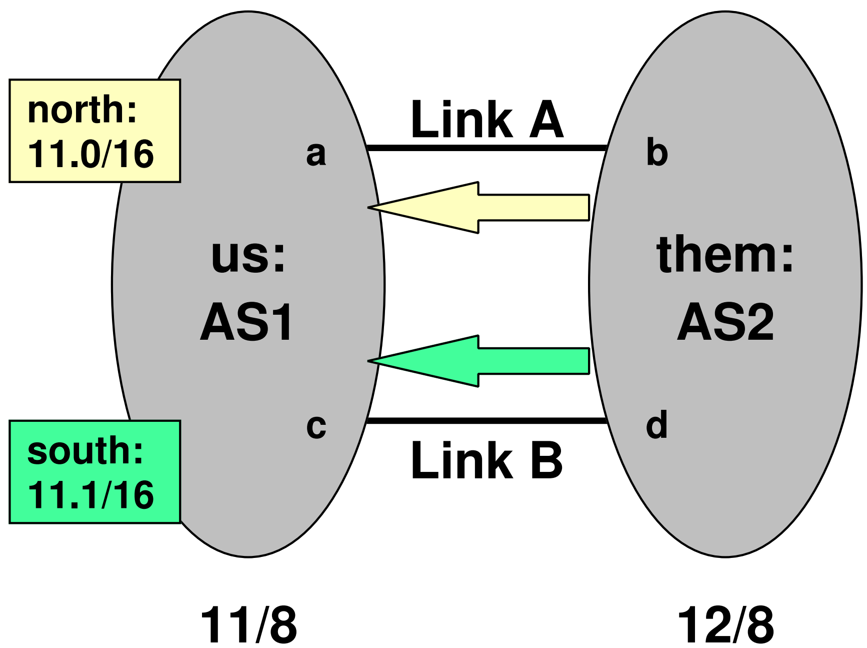


You are AS1 with two links A & B to AS2. How can you make AS2 send north traffic to link A and south traffic to link B?

AS1 exports:

?

MED Example



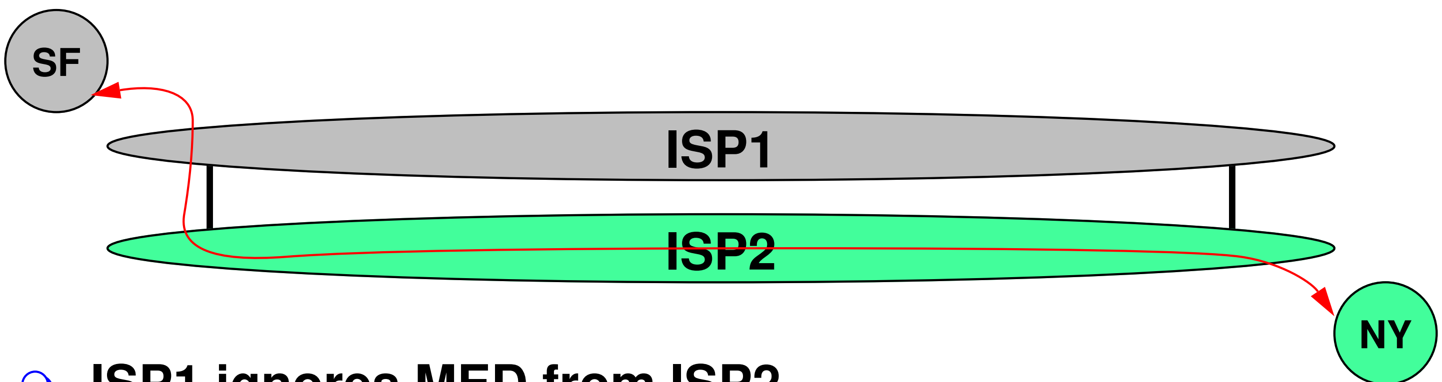
You are AS1 with two links A & B to AS2. How can you make AS2 send north traffic to link A and south traffic to link B?

AS1 exports:

11.0/16:a(1) w/ MED=10
11.0/16:c(1) w/ MED=20
11.1/16:a(1) w/ MED=20
11.1/16:c(1) w/ MED=10

MED (Cont...)

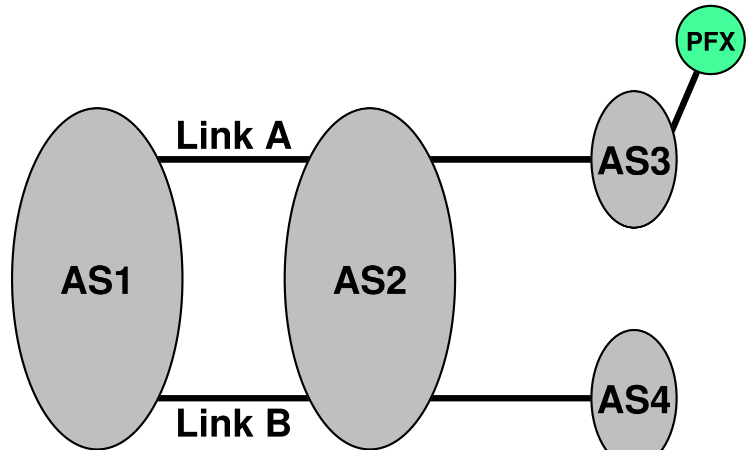
- MED is typically used in provider/subscriber scenarios.
- It can lead to unfairness if used between ISPs because it may force one ISP to carry more traffic:



- ISP1 ignores MED from ISP2
- ISP2 obeys MED from ISP1
- ISP2 ends up carrying traffic most of the way
- *"hot potato routing"*
- Results: MED ignored by ISP's that don't trust each other

MED Is Non-transitive

- AS1 sends MEDs to AS2, AS2 will *not* pass these MEDs to AS3 and AS4
  - MEDs are relative to links A and B *only*
- Cannot combine or compare MEDs from different AS's
  - AS1 learns two ways to reach PFX, one from AS2 and one from AS3, cannot compare MEDs



Route Selection

- Question: which routes should be installed in the forwarding table?
- Input: All routes that have been learned and accepted by a router
  - If only one route, then select it
  - If multiple routes (with same length prefix) then we have a decision to make

UPDATE Message Handling

- Unrecognized, optional, non-transitive attributes are ignored. Unrecognized, optional, transitive attributes cause the Partial bit to be set.
- WITHDRAWN routes are processed first.
- Feasible routes are placed in Adj-RIB-In, replacing old ones, if any.

Decision Process

- **Calculate degree of preference for each route in Adj-RIB-In as follows (apply following steps until one route is left):**
  1) **Select route with *highest LOCAL-PREF***
  2) **Select route with *shortest AS-PATH***
  3) **Apply MED (if routes learned from same neighbor), choose *lowest MED***
  4) **Select route with smallest NEXT-HOP cost (from IBGP, cost to edge router)**
  5) **Select route learned from E-BGP peer with lowest BGP ID**
  6) **Select route from I-BGP neighbor with lowest BGP ID**
- **Install selected route in Loc-RIB**
- **Disseminate routes to peers, update Adj-RIB-Out**
- **Done**

BGP's Importance

- BGP is a very powerful protocol
  - support for ***policy*** is unique among deployed routing protocols
- The key to global connectivity of the Internet
- Yet, it is so complex that many pathologies are being discovered even now, nearly a decade after initial deployment
  - delayed convergence [Labovitz00]
  - persistent oscillation (Varadhan 1996 and Griffin 2000)
  - router-reflector pathologies (Basu 2002)