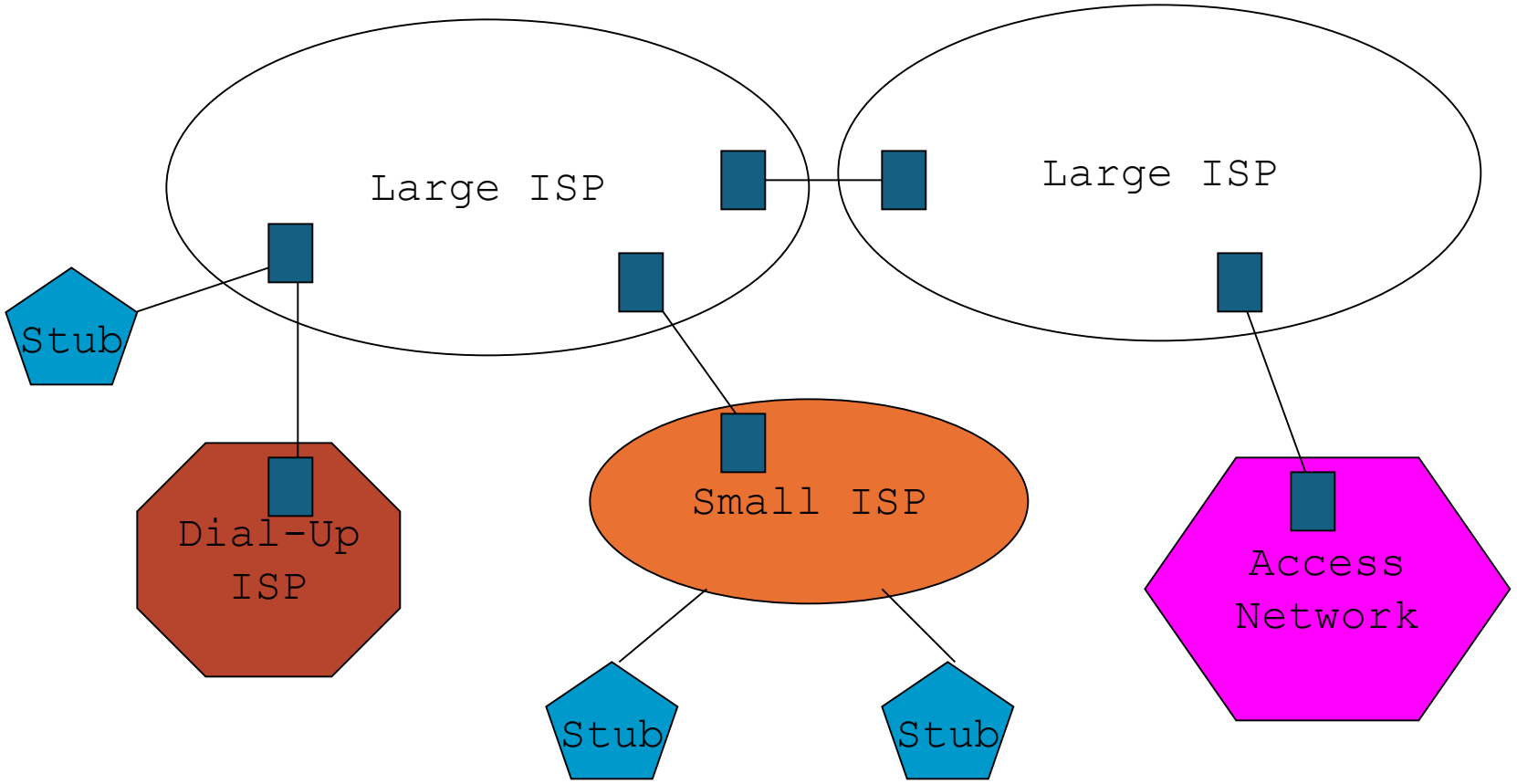
Chapter 5
Border Gateway Protocol(BGP)
RFC 1771

Rami Tawil

Outline

- Autonomous Systems and Gateways
- Inter-domain vs Intra-domain routing
- Classless Inter-domain routing
- BGP Messages, Attributes and Features
- AS Relationships and BGP Policies

Big Picture

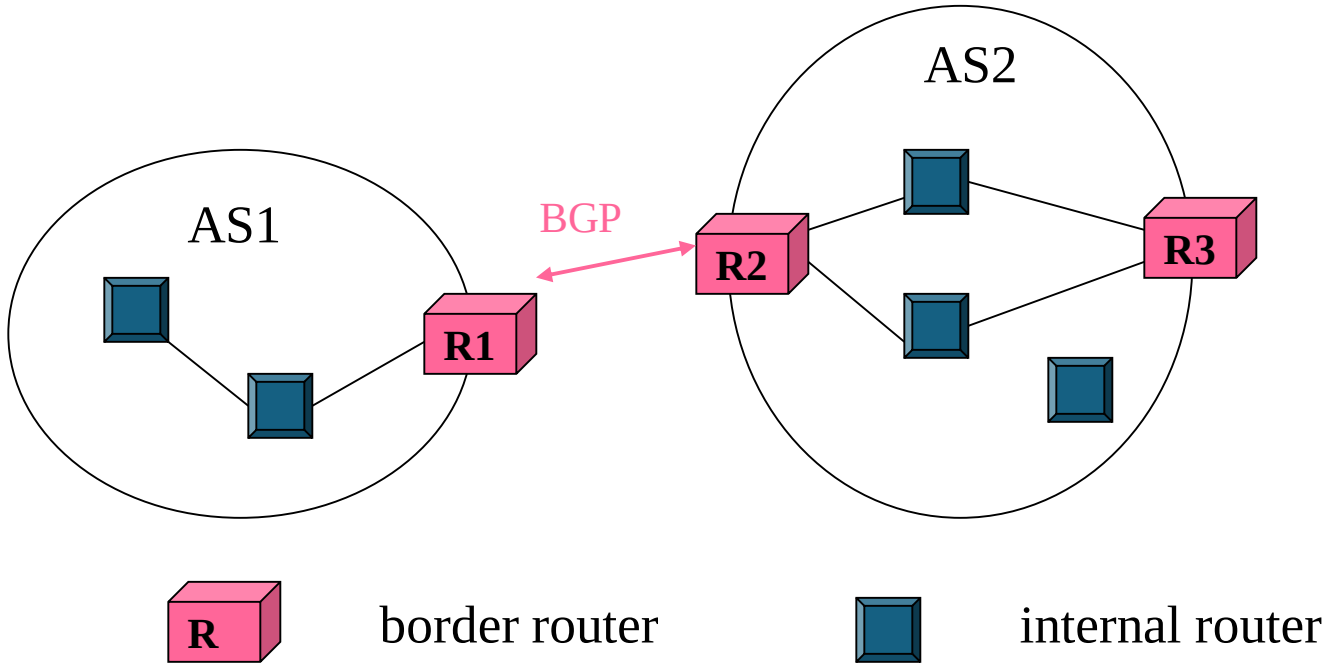


The Internet contains a large number of diverse networks

Autonomous System(AS)

- Internet is not a single network
  - Collection of networks controlled by different administrations
- An autonomous system is a network under a single administrative control
- An AS owns an IP prefix
- Every AS has a unique AS number
- ASes need to inter-network themselves to form a single virtual global network
  - Need a common protocol for communication

Who speaks BGP?



- Two types of routers
  - Border router(Edge), Internal router(Core)
- Two border routers of different ASes will have a BGP session

Purpose of BGP

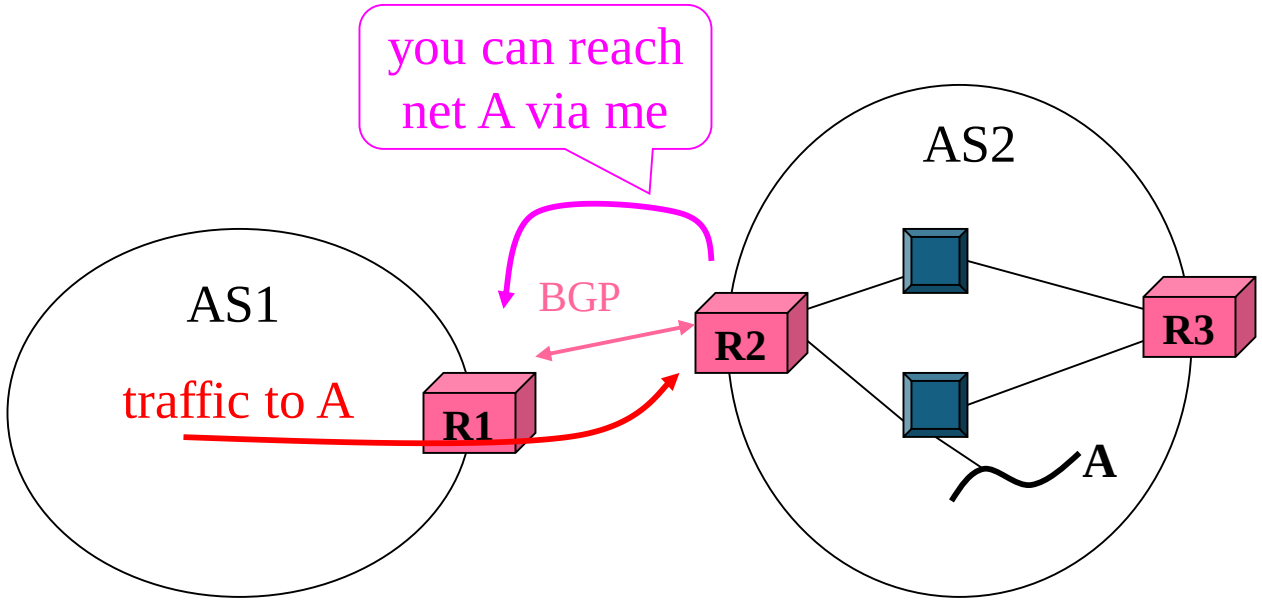
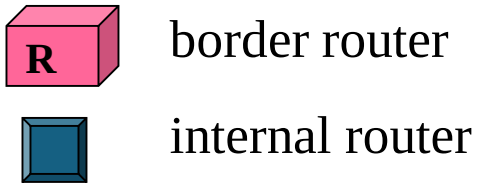


table at R1:

dest	next hop
A	R2

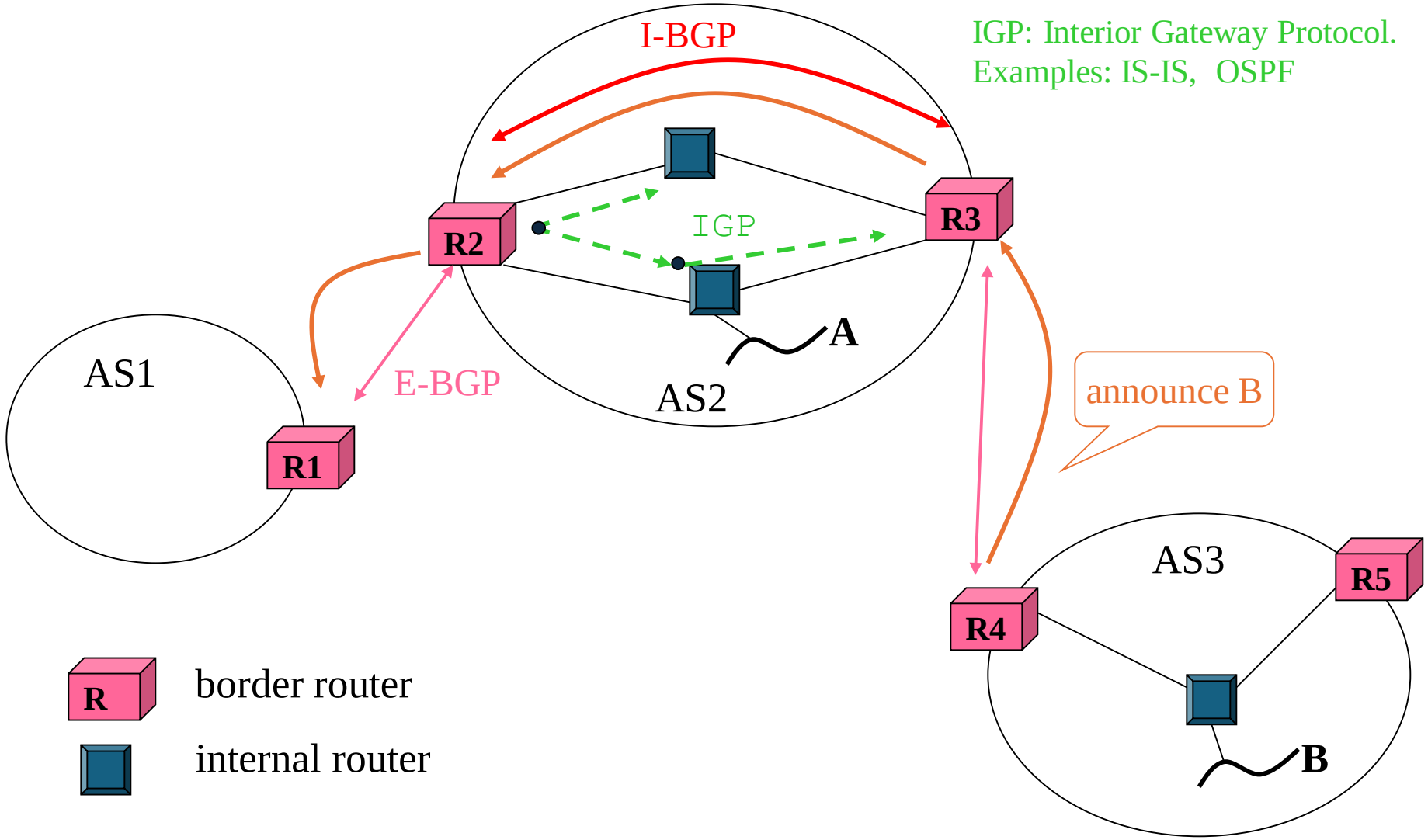


Share connectivity information across ASes

Intra-domain vs Inter-domain

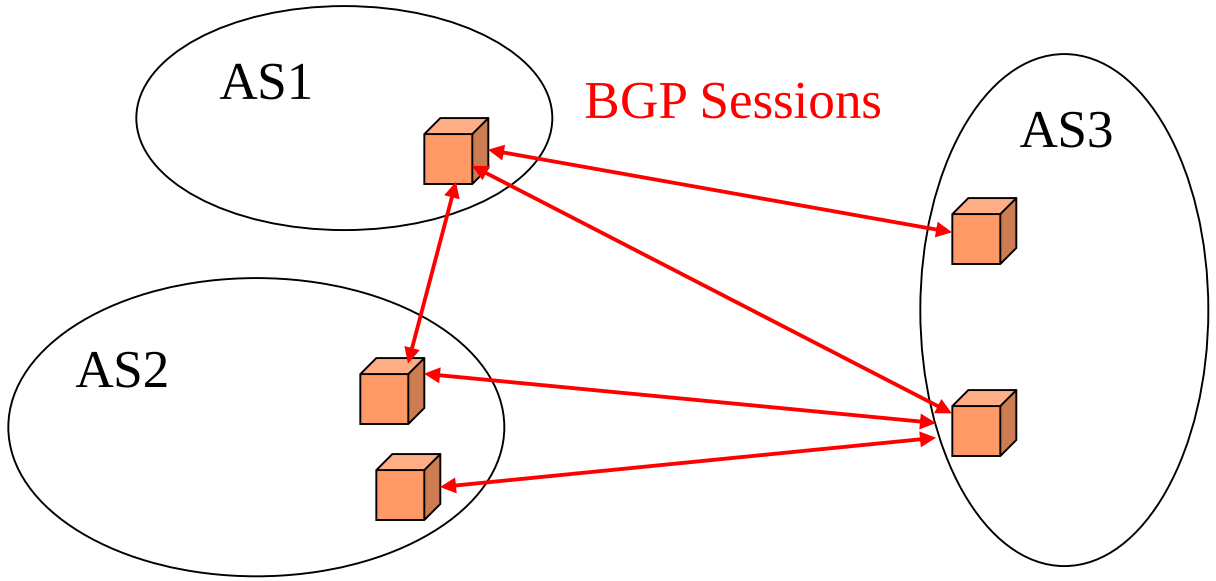
- An AS is a routing domain
- Within an AS:
  - Can run a link-state routing protocol
  - Trust other routers
  - Scale of network is relatively small
- Between ASes:
  - Lack of information about other AS's network (Link-state not possible)
  - Crossing trust boundaries
  - Link-state protocol will not scale
  - Routing protocol based on route propogation

I-BGP and E-BGP



Sharing routes

- One router can participate in many BGP sessions.
- *Initially* … node advertises ALL routes it wants neighbor to know (could be >50K routes)
- *Ongoing* … only inform neighbor of changes

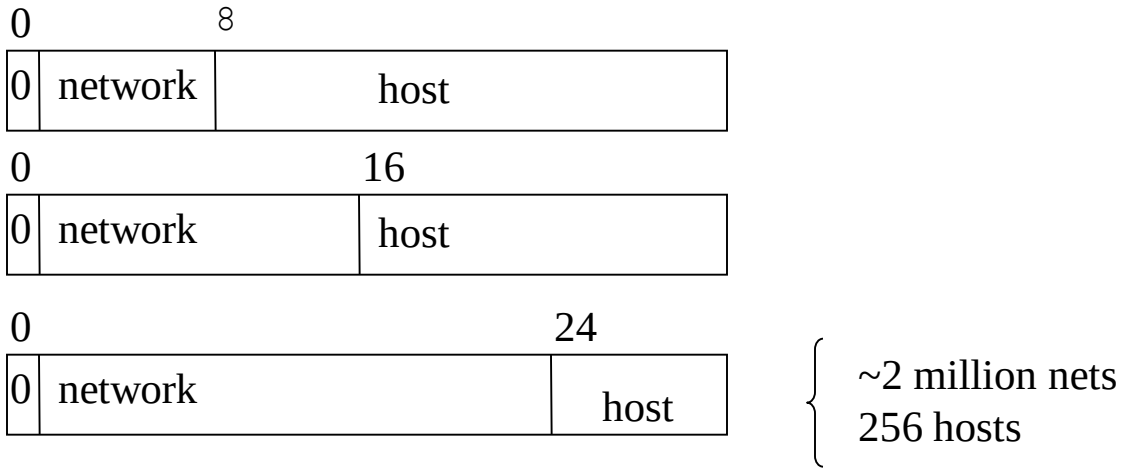


Assigning IP address and AS numbers (Ideally)

- A host gets its IP address from the IP address block of its organization
- An organization gets an IP address block from its ISP's address block
- An ISP gets its address block from its own provider OR from one of the 3 routing registries:
  - ARIN: American Registry for Internet Numbers
  - RIPE: Reseaux IP Europeens
  - APNIC: Asia Pacific Network Information Center
- Each AS is assigned a 16-bit number (65536 total)
  - Currently 10,000 AS's in use

Addressing Schemes

- Original addressing schemes (class-based):
  - 32 bits divided into 2 parts:
  - Class A
  - Class B
  - Class C



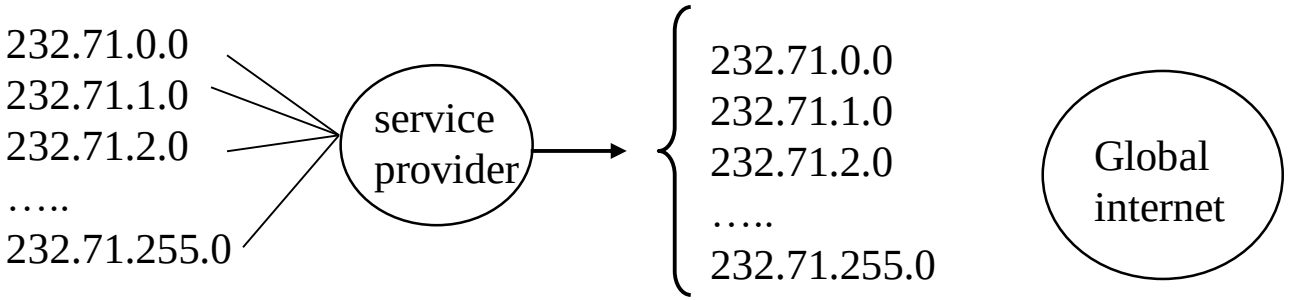
- CIDR introduced to solve 2 problems:
  - exhaustion of IP address space
  - size and growth rate of routing table

Problem #1: Lifetime of Address Space

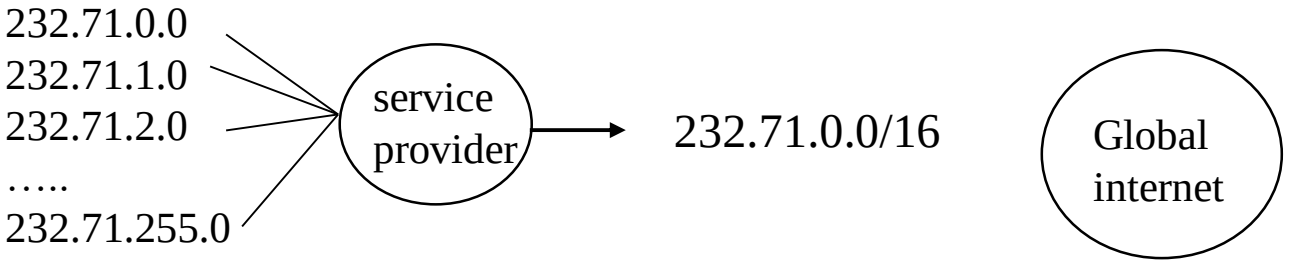
- Example: an organization needs 500 addresses. A single class C address not enough (256 hosts).
  Instead a class B address is allocated. (~64K hosts) That's overkill -a huge waste.
- CIDR allows networks to be assigned on arbitrary bit boundaries.
  - permits arbitrary sized masks: 178.24.14.0/23 is valid
  - requires explicit masks to be passed in routing protocols
- CIDR solution for example above: organization

Problem #2: Routing Table Size

Without CIDR:



With CIDR:



CIDR: Classless Inter-Domain Routing

- Address format <IP address/prefix P>.
  - The prefix denotes the upper P bits of the IP address.
- Idea - *use aggregation* - provide routing for a large number of customers by advertising one common prefix.
  - This is possible because nature of addressing is hierarchical
- Summarization reduces the size of routing tables, but maintains connectivity.
- Aggregation

BGP Details

- Classless Inter-domain Routing
- Path-vector protocol
- BGP Messages, Attributes
- Preference-based routing
- Export and Import Policies

BGP: A Path-vector protocol

```
ner-routes>show ip bgp
BGP table version is 6128791, local router ID is 4.2.34.165
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal
Origin codes: i - IGP, e - EGP, ? - incomplete
   Network          Next Hop            Metric LocPrf Weight Path
* i3.0.0.0         4.0.6.142             1000     50      0 701 80 i
* i4.0.0.0         4.24.1.35                0    100      0 i
• i12.3.21.0/23    192.205.32.153        1000     50      0 7018 4264 6468 I
• e128.32.0.0/16   192.205.32.153        1000     50      0 7018 4264 6468 25 e
```

- Every route advertisement contains the entire AS path
  - Generalization of distance vector
- Can implement policies for choosing best route

Example:

AS_PATH: 65001 65010 65055 65077

Route Attributes

- NEXT-HOP: tells the router which next router to send the packet to.
- LOCAL PREF: used inside the AS (higher => more preferred path)
- MED (Multiple-exit discriminator): used between two different ASes (lower => better path) – ***if you are sending traffic to me use this exit.***
- ORIGIN: shows the route originally came from: I -> IGP, e -> EGP or ? -> Incomplete (static path - router prefer: i->e->?)
- AS-PATH: a list of AS's through which the announcement for a prefix has passed

Basic Messages in BGP

OPEN Message

- **Purpose**

To **start a BGP connection** between two routers.

- **What it contains**
  - **Version** (usually 4)
  - **My AS number**
  - **Hold time** **(If I don't hear anything (a KEEPALIVE or UPDATE) from you for n seconds, I will think our BGP connection is dead)**
  - **BGP Identifier** (router ID)
  - **Optional parameters** (like capabilities)
- **Example:**

Router A sends:

```
BGP OPEN:
  Version: 4
  My AS: 65010
  Hold Time: 180
  BGP Identifier: 10.1.1.1
  Capabilities: IPv4 Unicast, IPv6 Unicast
```

Router B replies with its own OPEN message.

Basic Messages in BGP

Update Message

- **Purpose**

To **advertise** (send) or **withdraw** (remove) routes.

- **What it contains**
  - Withdrawn routes (if removing)
  - Path attributes (AS-PATH, NEXT-HOP, LOCAL PREF…)
  - Network Layer Reachability Information (NLRI) → the actual routes being advertised
- **Example (Advertising a route):**

Router A wants to announce network **192.168.10.0/24**:

```
BGP UPDATE:
  Withdrawn Routes:
None
  Path Attributes:
      NEXT-HOP:
10.0.0.1
      AS-PATH: [65010]
      LOCAL_PREF: 100
  NLRI: 192.168.10.0/24
```

- **Example (withdrawing a route)**

Router A removes a network:

```
BGP UPDATE:
  Withdrawn Routes:
      192.168.50.0/24
```

Basic Messages in BGP

KEEPALIVE Message

- **Purpose**

To make sure the **connection is still alive**.

- **What it contains**
  - No payload → it is empty
  - Only a header
- **Example:**

Router A sends every 60 seconds:

BGP KEEPALIVE
(Just a heartbeat — no data)

Basic Messages in BGP

NOTIFICATION Message

- **Purpose**

To report **errors** and **close** the connection.

What it contains

  - Error code
  - Error subcode
  - Optional error data

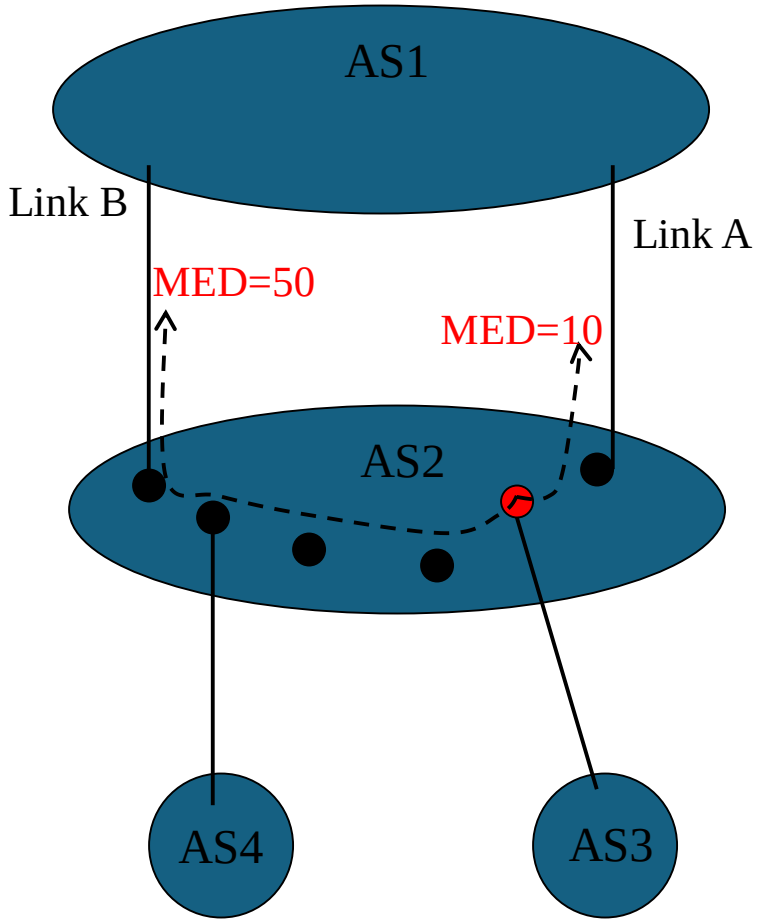
- **Example:**

Router B detects a bad AS number:

```
BGP NOTIFICATION:
  Error Code: 2 (OPEN Message
Error)
  Subcode: 2 (Bad Peer AS)
  Data: AS number mismatch
```

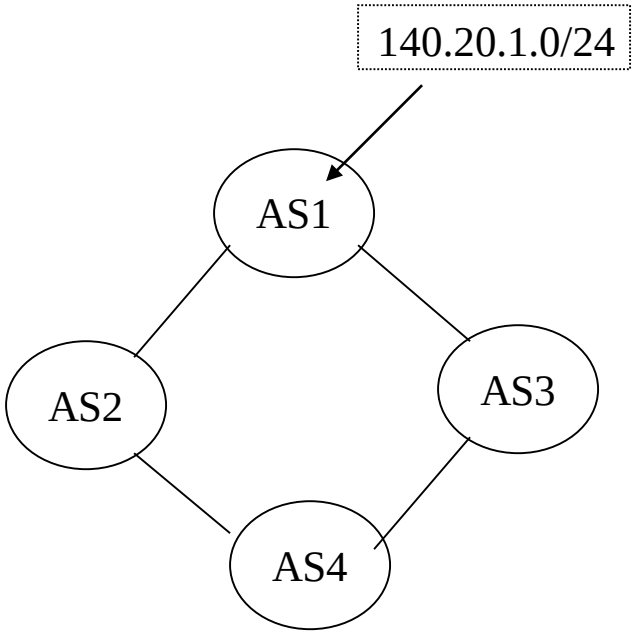
Attribute: Multi-Exit Discriminator (MED)

- when AS's interconnected via 2 or more links
- AS announcing prefix sets MED
- enables AS2 to indicate its preference
- AS receiving prefix uses MED to select link
- a way to specify how close a prefix is to the link it is announced on



Attribute: Local Preference

- Used to indicate preference among multiple paths for the same prefix *anywhere* in the internet.
- The higher the value the more preferred
- Exchanged between IBGP peers only. Local to the AS.
- Often used to select a specific exit point for a particular destination



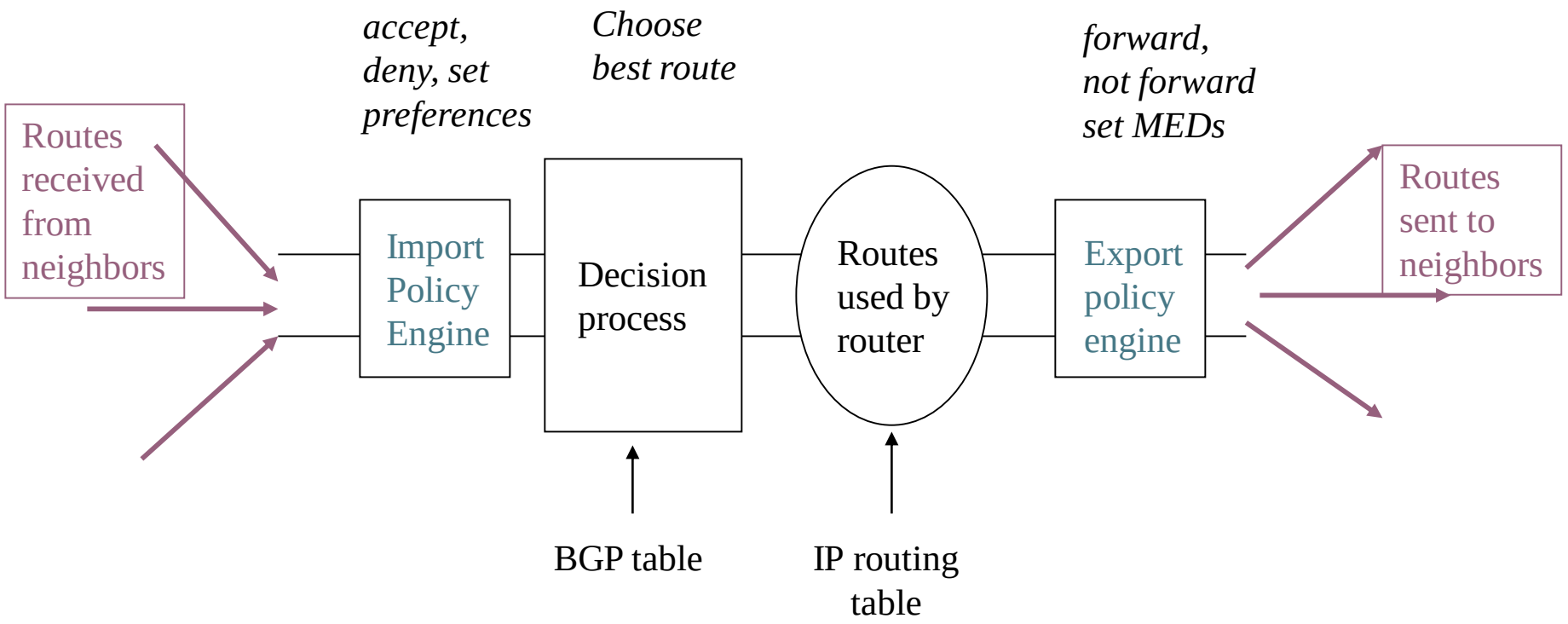
BGP table at AS4:

Destination	AS Path	Local Pref
140.20.1.0/24	**AS3 AS1**	**300**
140.20.1.0/24	**AS2 AS1**	**100**

Choosing best route

- Choose route with highest **LOCAL_PREF**
  - Preference-based routing
- If multiple choices, select route with shortest **hop-count**
- If multiple choices for same neighboring AS, choose path with max MED value
- Choose route based on lowest origin type
  - IGP < EGP < INCOMPLETE
- Among IGP paths, choose one with lowest cost
- Finally use router ID to break the tie.

Routing Process Overview



Import and Export Policies

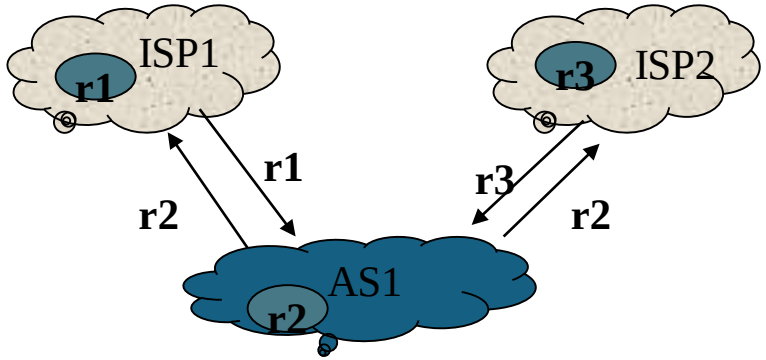
- Inbound filtering controls outbound traffic
  - filters route updates received from other peers
  - filtering based on IP prefixes, AS_PATH, community
- Outbound Filtering controls inbound traffic
  - forwarding a route means others may choose to reach the prefix through you
  - not forwarding a route means others must use another router to reach the prefix
- Attribute Manipulation
  - Import: LOCAL_PREF (manipulate trust)
  - Export: AS_PATH and MEDs

Transit vs. Nontransit AS

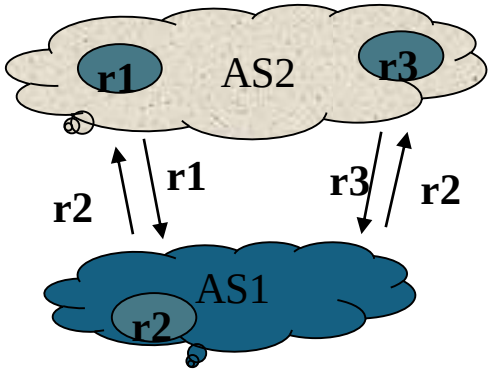
Transit traffic = traffic whose source and destination are outside the AS

Nontransit AS: does not carry transit traffic

- Advertise own routes only
- Do not propagate routes learned from other AS's
- case 1:

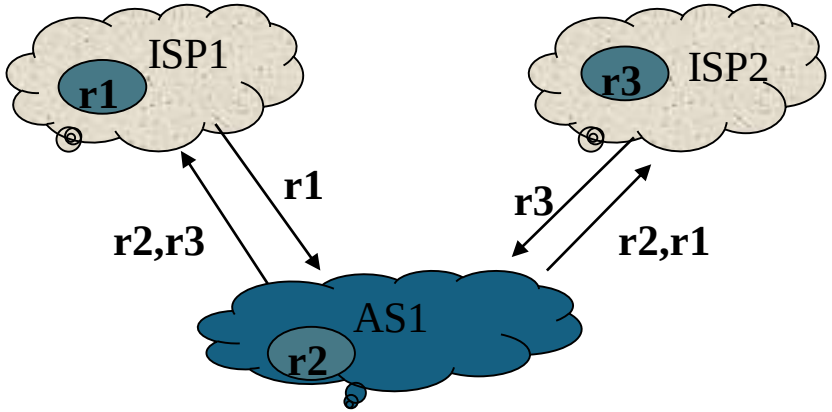


- case 2:



Transit AS: does carry transit traffic

- Advertises its own routes PLUS routes learned from other AS's

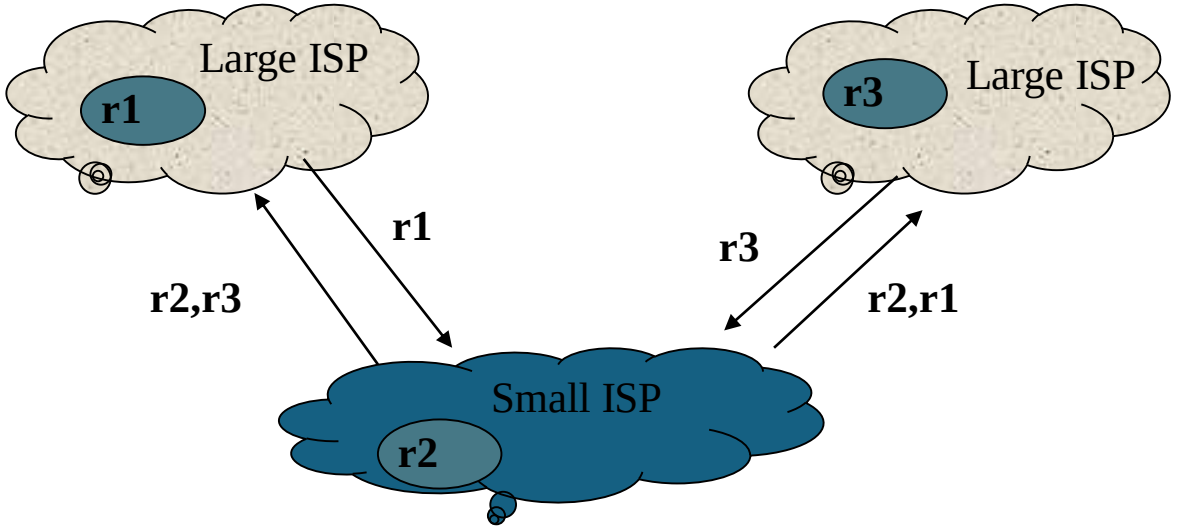


AS relationships, Export rules

- AS has customers, providers and peers
- Relationships between AS pairs:
  - customer-provider
  - peer-to-peer
- Type of relationship influences policies

- Exporting to provider:
  AS exports its routes & its customer's routes, but not routes learned from other providers or peers
- Exporting to peer:
  (same as above)
- Exporting to customer:
  AS exports its routes plus routes learned from its providers and peers

Customer-Transit problem



- Assume that the small ISP is a customer of two large ISPs
- If customer ISP does not obey export rules
  - forwards advertisements from one large ISP to another

Take-aways

- Internet is composed of various ASes which use BGP to inter-network themselves
- Internet switched to classless addressing
- BGP as a routing protocol
  - Path-vector based
  - Supports route-aggregation
  - Supports preferential routing
  - Uses Import and Export policies
- BGP is the protocol that “holds” the Internet intact