Border Gateway Protocol (BGP)

Petr Grygárek

Role of Autonomous Systems on the Internet

Autonomous systems

- Not possible to maintain complete Internet topology information on all routers
  - big database, change processing overhead, instability
- Internet divided into Autonomous systems
  - ISP, big company
- Autonomous system = contiguous set of routers with common routing policy and under common administration
  - Routing policy: IGP, implemented route optimization, …
- Autonomous systems numbered with world-wide unique numbers (16 bit)

Hierarchical routing

- Intra-AS routing uses Interior Gateway Protocols (IGP)
  - knows only topology of it's own AS
  - outside of AS is reached using default
    - sometimes has summary information about networks behind individual external links
      - Limited by number of routes the protocol is capable to process efficiently
  - OPSF, RIP, IGRP, …
- Inter-AS routing uses Exterior Gateway Protocols
  - Operates on graph of AS interconnection
  - Does not know topology of other ASes, works only with information about networks contained in individual ASes
    - knows local next-hop border router to reach the destination
  - Currently only BGP (Border Gateway Protocol) is used

Inter-AS routing

- The purpose of EGP is to provide information to deliver packet to the boundary router of the destination AS
  - Boundary routers run both EGP and IGP
  - Boundary router delivers the packet to the final destination using IGP
- Every AS propagates networks contained within it into EGP
  - also networks reachable through it
  - It is useful to limit number of routes propagated using summarization (internal networks should have common prefix)

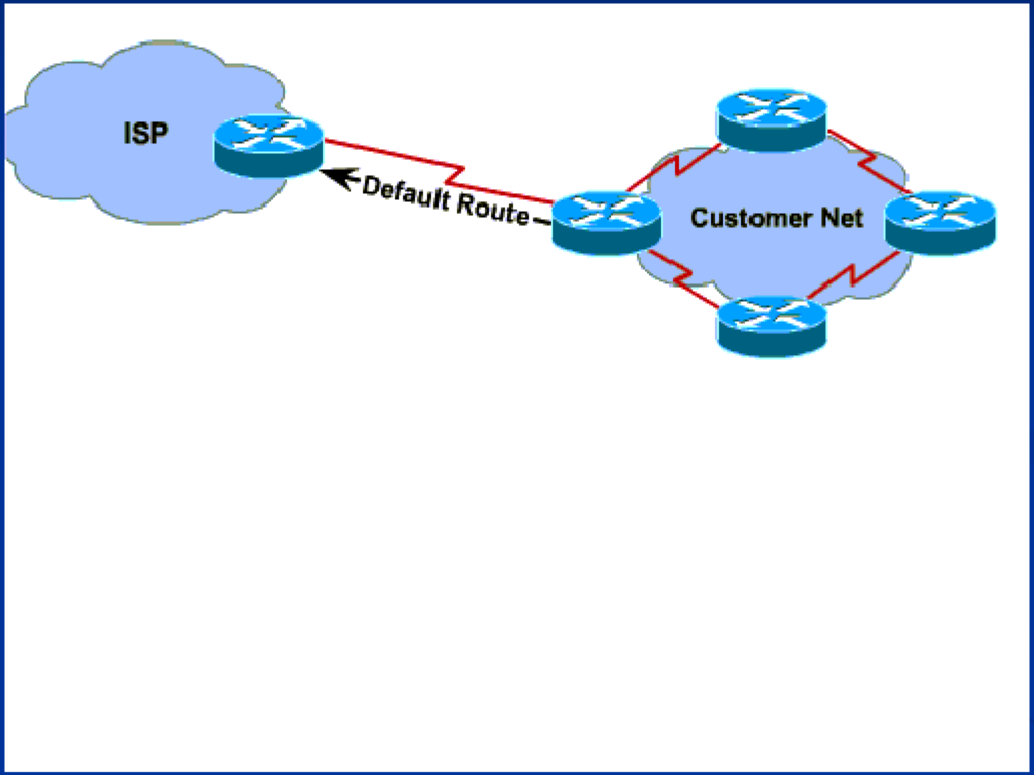
Note: For transit AS, packet has to be passed among border routers through AS internal routers

AS types

- Single-homed
- Multi-homed
  - More links to the same ISP or different ISPs

- Transit
  - Carries traffic not originated or destined to internal networks
  - multi-homed
- Non-transit
  - single-homed or multi-homed AS which doesn't allow transit

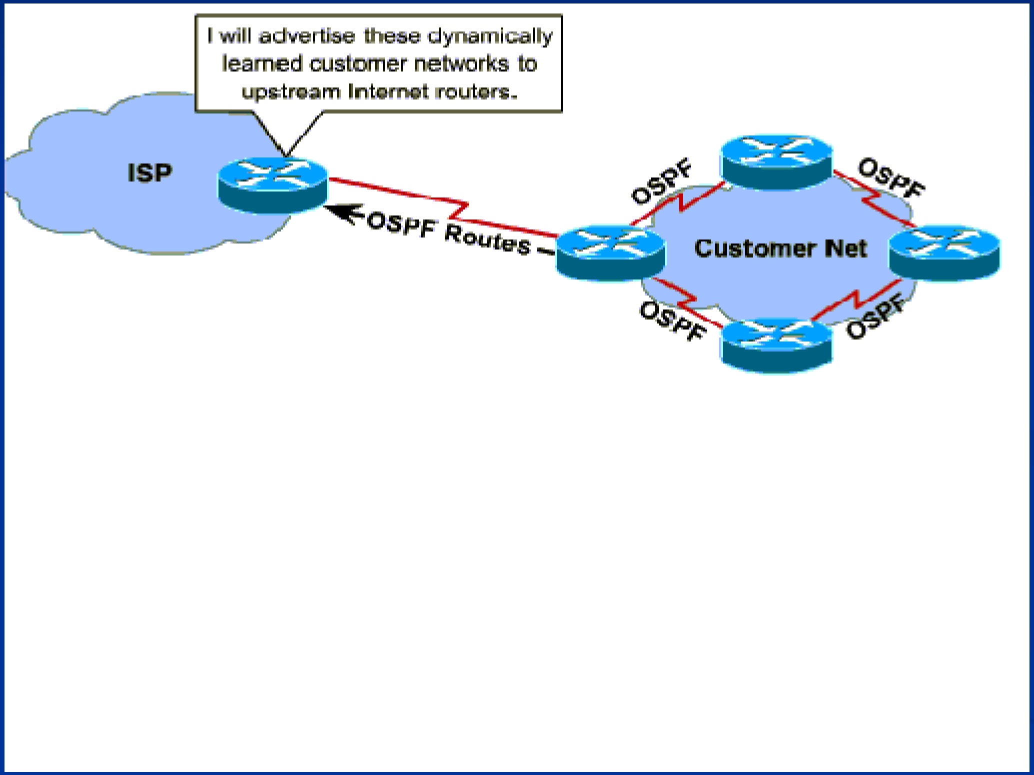
Single-homed AS



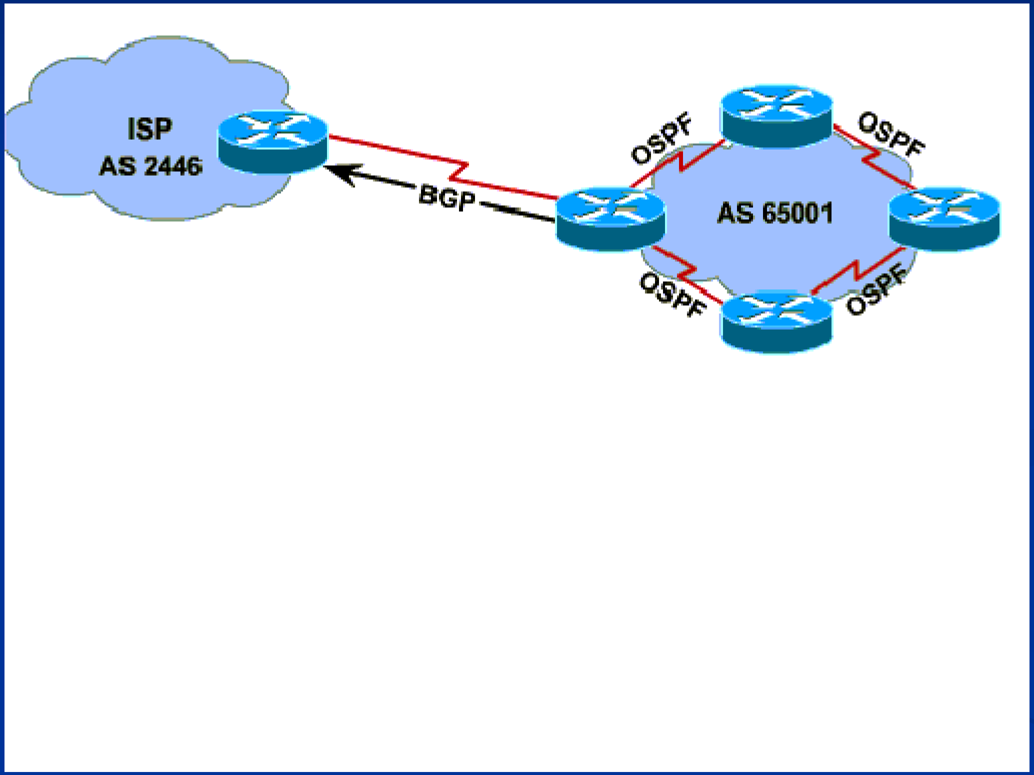
Single-homed AS: How to propagate internal networks into BGP ?

- ISP router has static routes to customer's networks
  - and redistributes them into BGP
- IGP between ISP router and customer router
  - ISP redistributes IGP into BGP
- BGP between ISP router and customer router
  - If customer has it's own AS number
    - or uses private AS number

Propagation via IGP



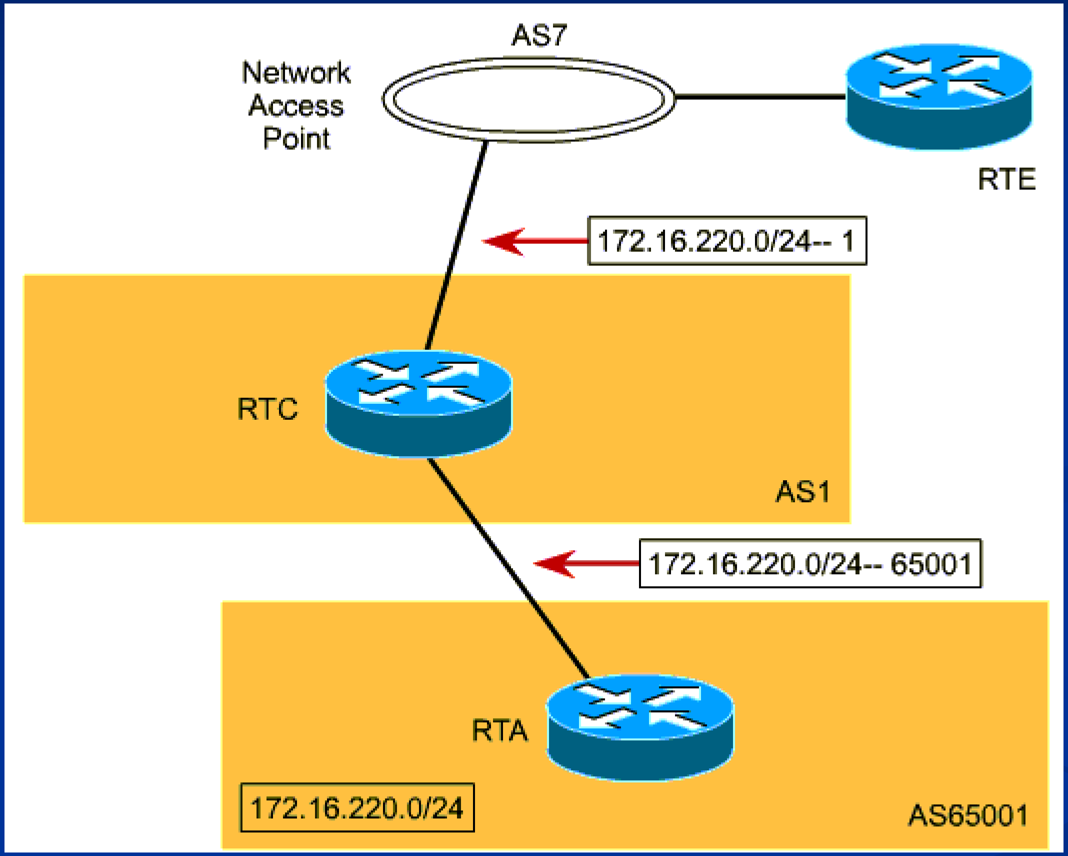
Propagation via BGP



Private AS-es

- 64512 – 65535
- Used and known only in context of single provider’s AS
  - Can be used only for AS connected to single provider (by one or more links)
- Outside of provider AS, private AS-es presents themselves like part of that’s provider AS

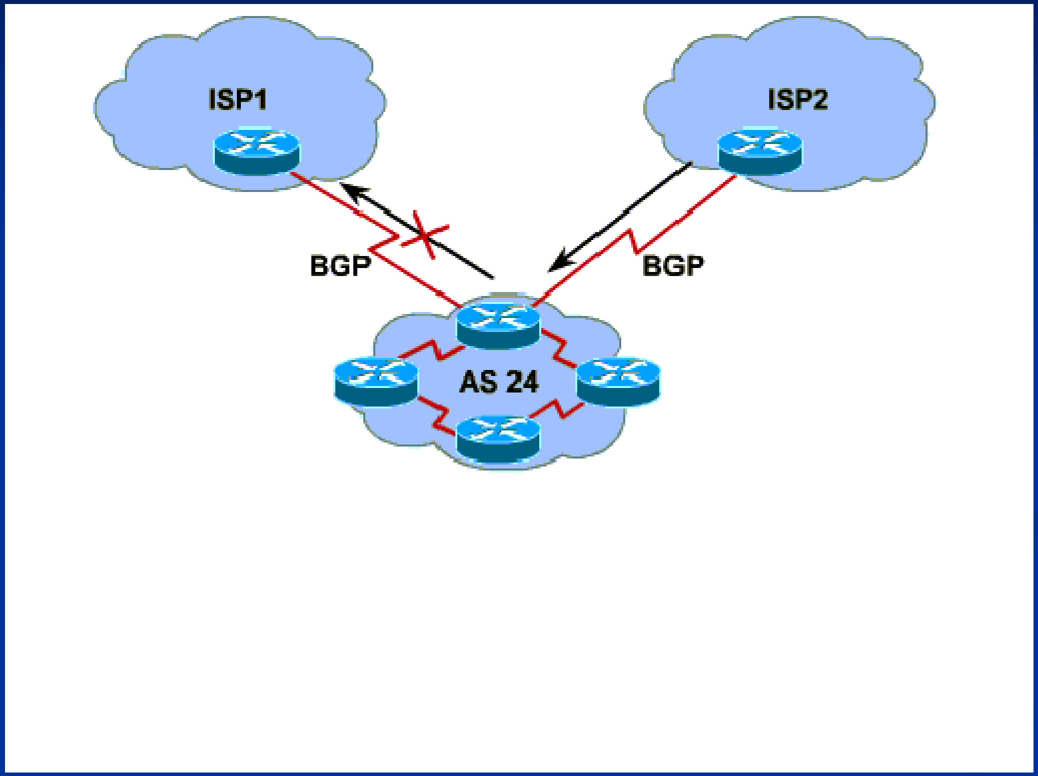
Private AS-es



Who has it's own AS ?

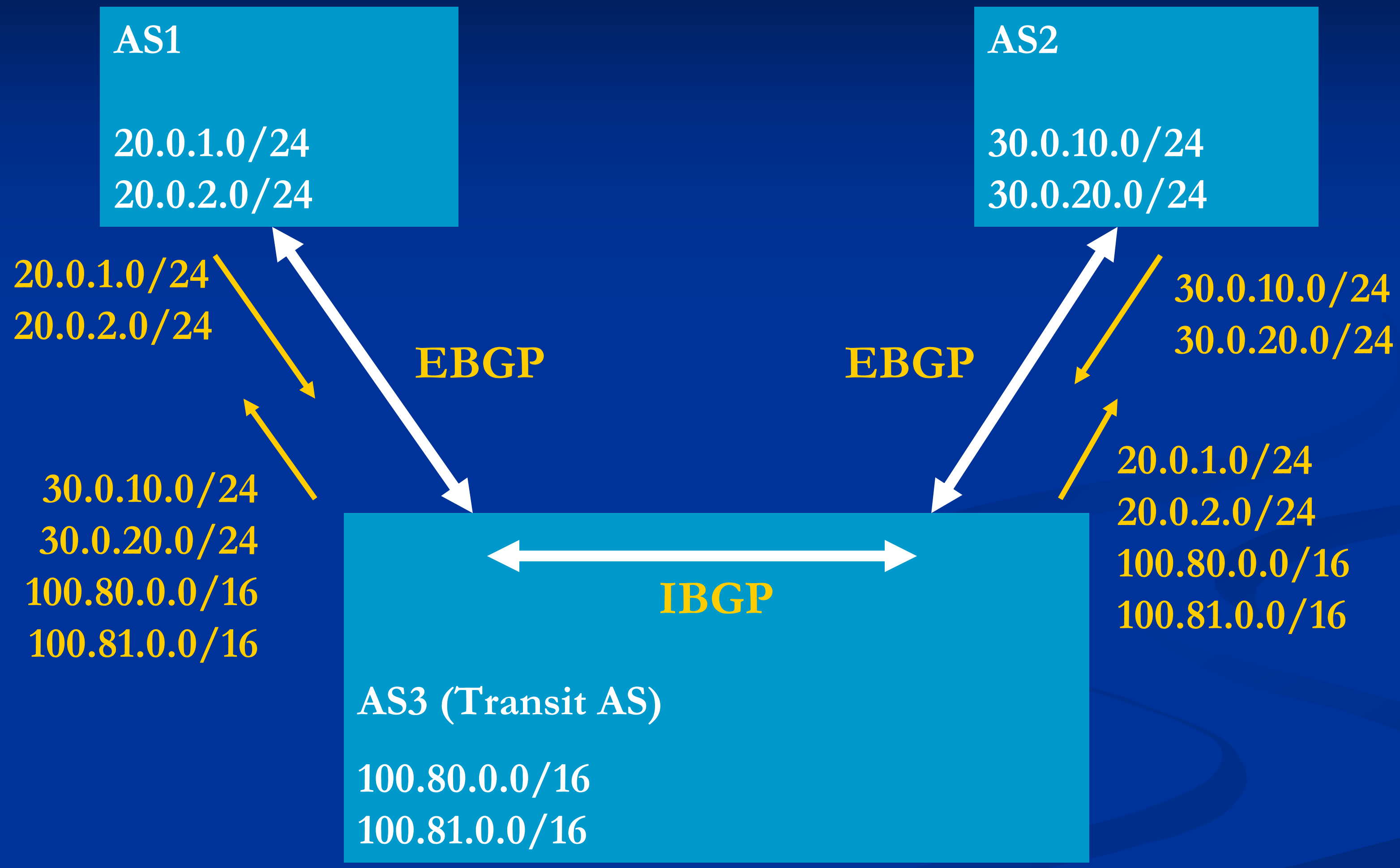
- Normally, customer's networks are part of provider AS
  - Sometimes private AS-es used
- Customer has to have it's own AS number if he indents to connect to multiple providers
- Customer commonly needs it's own AS number if it requires provider-independent addresses

Nontransit Multi-homed AS



- Packet filters can be used on ingress links to protect against injection of unwanted traffic
  - ISP1 could use static route to route to ISP2 networks

Transit multi-homed AS



Routing symmetry, load balancing

- Symmetry - the link used for outgoing traffic for some network is also used for returning traffic
- Load balancing – some destinations reached by one link, others by another

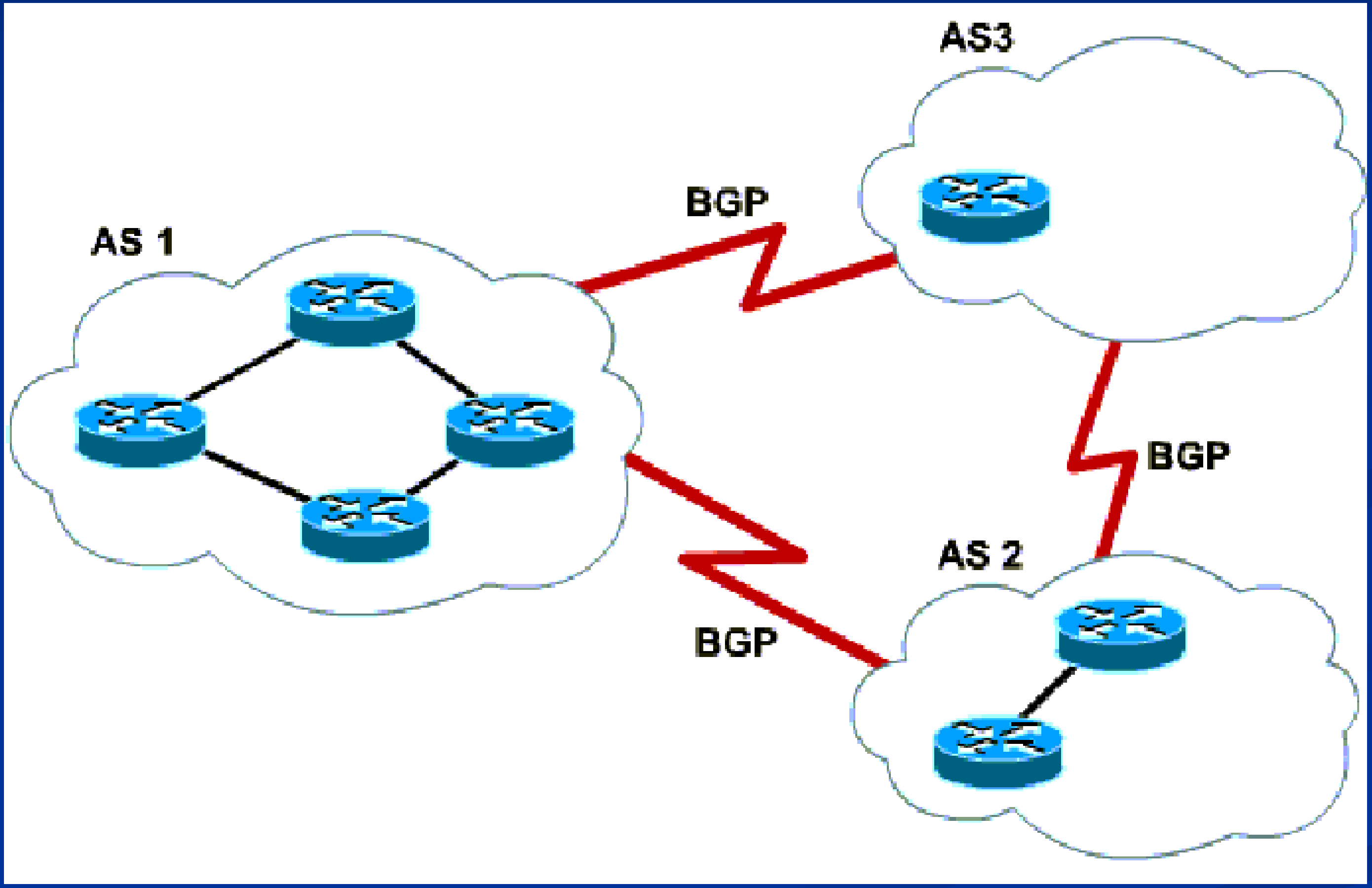
Often not possible to reach both

Border Gateway Protocol

Border Gateway Protocol

- Exchanges information between AS border routers
  - What networks are in each AS
  - List of AS-es to transit when reaching particular network
- Today, BGP v.4 is used
  - Sometimes BGPv4+: multiprotocol extension
    - Other address families, multicasting, VPNs, …
  - Supports classless addressing
    - Propagates subnet masks with every prefix
    - Allows for address range aggregation

BGP operation on graph of AS-es



Path selection, routing policies

- BGP operates on AS interconnection graph
- Path = sequence of AS numbers to transit to get to particular network
- BGP does not have simple concept of metric to select best path
  - Path has to be chosen with regard to business policy of individual AS operators
    - BGP configuration has to reflect appointed routing policy
  - Details of routing policy have to be configured manually
    - Peer routers, prefix filtering and route preferences, …
    - Configuration more complicated than configuration of IGP's

Examples of routing policies

- Which destination we allow to transit packets to through our AS ?
- From which source address we allow to transit traffic through our AS ?
- Which external link will we use to reach particular external network ?
- Which ingress link we want other ASes use for traffic destined for particular network inside our AS ?

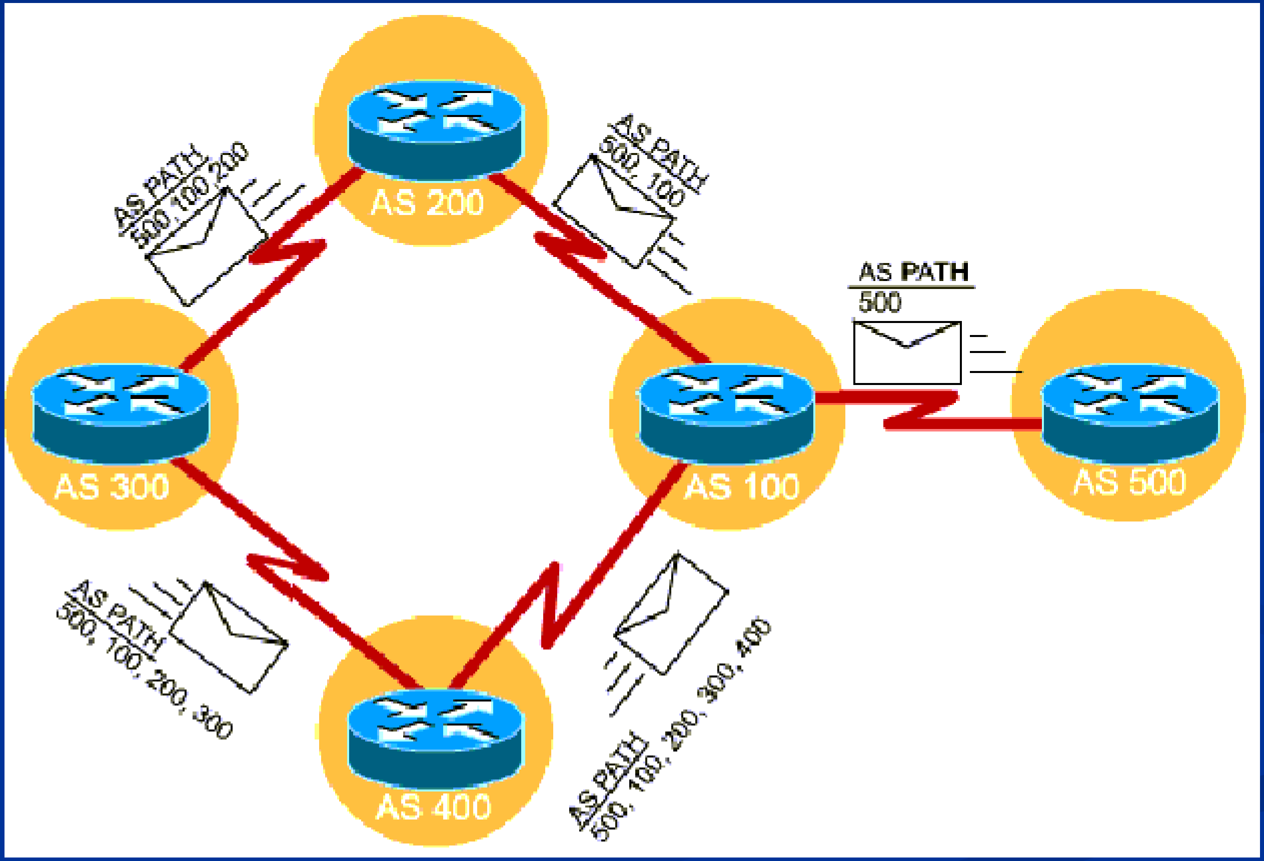
Suboptimal routing on the Internet

- Internet routing is not optimal from point of view of any metric
  - There is no common metric, various IGPs use different metrics
- Optimality not reachable neither desired
  - Hierarchical routing is suboptimal
    - but limits the number of routes in routing table
  - Need to respect routing policies

BGP Principle

- Path-vector routing algorithm
  - from point of view of topology knowledge, BGP stands between distance-vector and link-state protocols
- Path vector = sequence of AS numbers to transit before getting to particular network
- Every route is propagated together with it's path vector
  - Path vector collects number of AS-es the route was passed through
  - If AS receives route with path vector containing it's own AS number, route is discarded (loop avoidance)
- Path vector serves as metric
  - route with shorter path vector is preferred

Passing of BGP routes



Spreading of routing information

- Routing information exchanged between AS boundary routers
  - Peer routers to exchange routing information with are configured manually
  - Reliable exchange (TCP, port 179)
- When BGP session is established among peers, complete routing information is exchanged
- After initial exchange, only changes are sent

Peer reachability testing

- BGP router periodically checks reachability of every peer
  - Keepalive message sent once per minute
- If some peer fails, the router has to remove all routes through that peer and inform other peers

BGP messages

Exchanged between peer routers

(TCP/179, support for authentication)

- OPEN – session establishment
  - Negotiation of protocol version, hold time for keepalives, AS numbers, …
- UPDATE
  - Advertised prefixes (+ route attributes), withdrawn routes
- KEEPALIVE – peer reachability testing
- NOTIFICATION – operation error, close session

BGP database

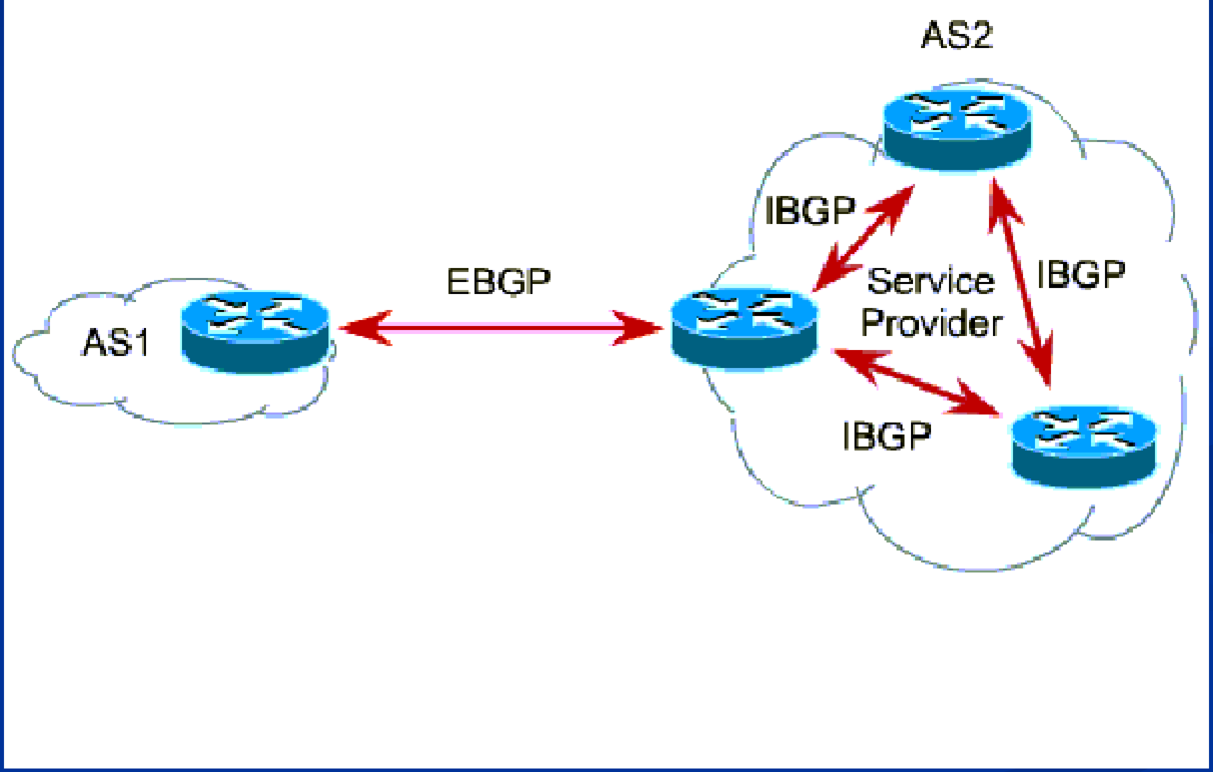
- BGP database contains all routes learned from peers
- For every destination, one route is chosen based on routing policy criteria
  - No support for load balancing
- Chosen routes are placed into routing table
- Only routes used by router itself (i.e. those chosen into routing table) are propagated to other neighbors

External and Internal BGP

External and Internal BGP

- If there is more than one boundary router in some AS, BGP information has to be passed between them
  - Special case, exchange between routers in the same AS
- Boundary routers can possibly be separated by internal structure of routers (running IGP)
- Solution: there exists two types of BGP session
  - External BGP (EBGP)
  - Internal BGP (IBGP)
    - Peers do not have to be physically connected

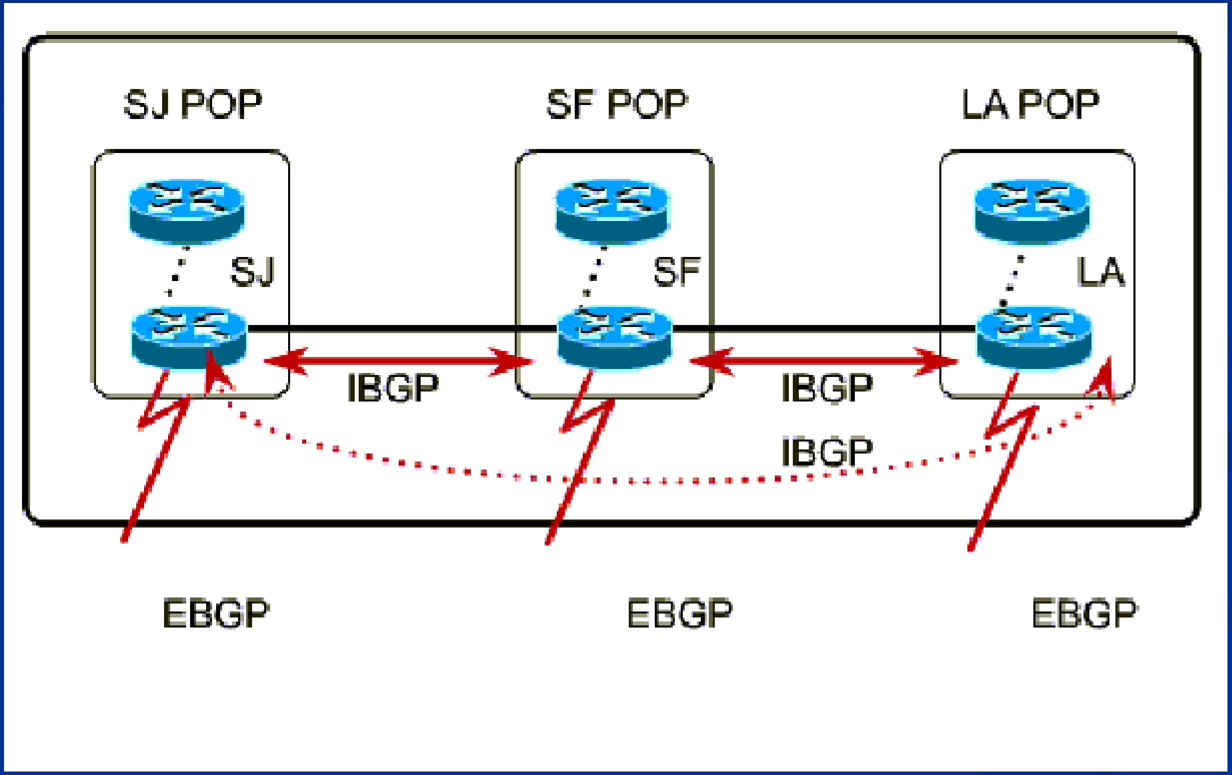
EBGP and IBGP



Passing of routes in IBGP sessions

- Need to avoid loops when passing routes through IBGP
  - Test for presence of receiving peer's AS number in path vector doesn't work
- Special rules defined for passing of routes in IBGP session
  - Information from IBGP is passed to EBGP peers, but not to other IBGP peers.
  - Information from EBGP is passed to other EBGP peers and all IBGP peers

Full mesh of IBGP sessions



Definition of BGP Routing Policy

BGP Attributes

- Mechanism of implementation of routing policies
- Every route passed between peers can be assigned one or more attributes
- Routes are processed and selected based on values of attributes they carry

Attribute Types

- Well-known
  - understood by every BGP implementation
  - Mandatory – must be appended to each route
  - Discretionary – may be appended to route
- Optional
  - not every BGP implementation must understand it
  - Transitive
    – if implementation doesn't understand the attribute, it passes it next unchanged
  - Nontransitive
    – if implementation doesn't understand the attribute, it doesn't pass it next

Most commonly used Attributes

BGP Attribute Codes and Their Respective Types

Attribute Code	Type
1 -- ORIGIN	Well-known mandatory
2 -- AS_PATH	Well-known mandatory
3 -- NEXT_HOP	Well-known mandatory
4 -- MULTI_EXIT_DISC	Optional nontransitive
5 -- LOCAL_PREF	Well-known discretionary
6 -- ATOMIC_AGGREGATE	Well-known discretionary
7 -- AGGREGATOR	Well-known discretionary
8 -- COMMUNITY	Optional transitive (Cisco)
9 -- ORIGINATOR_ID	Optional nontransitive (Cisco)
10 -- Cluster List	Optional nontransitive (Cisco)
11 -- Destination Preference	(MCI)
12 -- Advertiser	(Baynet)
13 -- rcid_path	(Baynet)
255 -- Reserved	--

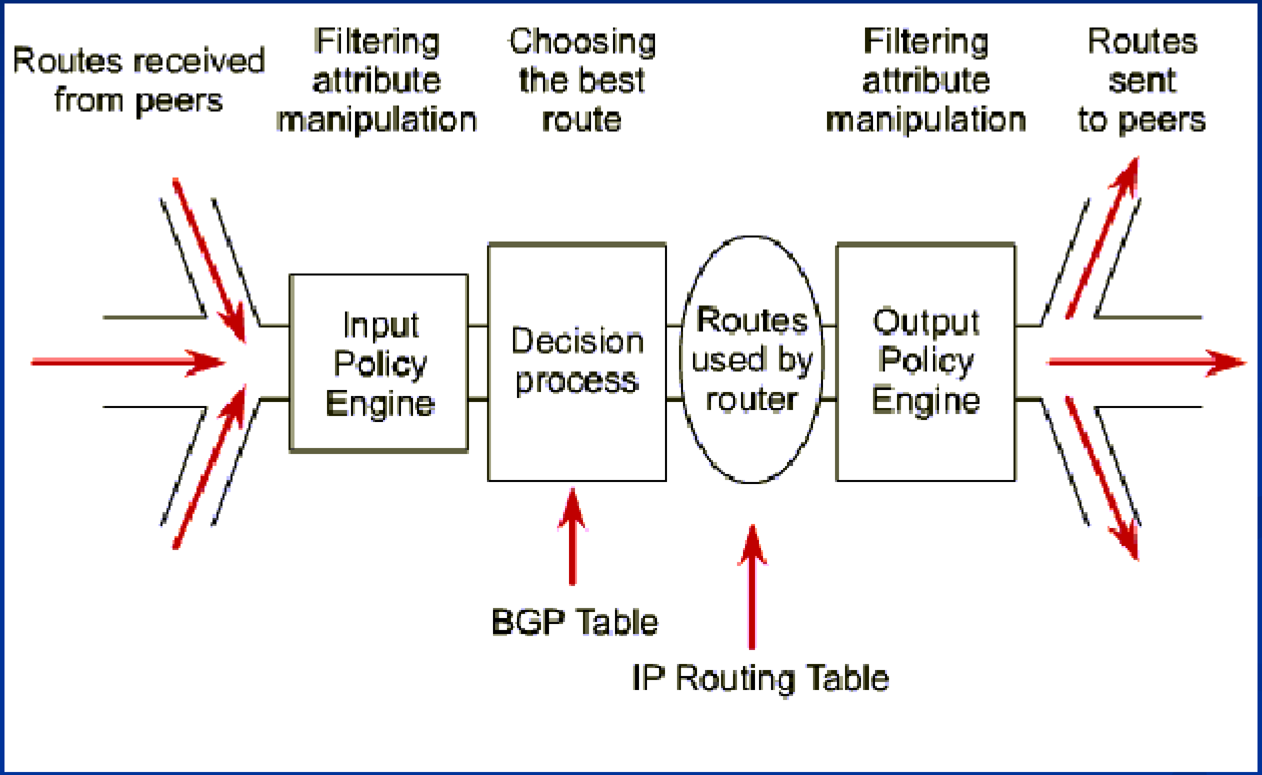
How to influence routing policy using attributes ?

- Manipulation with attributes received from individual peers
  - Input Policy Engine
  - Includes filtering of routes received from individual peers
- Manipulation with attributes of routes propagated to individual peers
  - Output Policy Engine
  - Includes filtering of routes propagated to individual peers
- Route used (and propagated next) by BGP router is determined by candidate route’s attribute values

Function of policy engines

- Test for attribute values
- Test for prefixes (including prefix length)
- Setting of attribute value when predefined criteria met
- Filtering of route when predefined criteria met

Processing of BGP routes



Definition of Routing Policies

- Separately for each peer
- Separately for incoming and outgoing routes

BGP Table (BGP database)

- Contains routes passed through (and possibly manipulated by) input policy engine
  - Routes from every peer
- For every destination (prefix), one best route is chosen
  - Selection is based on attribute values
    - Standardized algorithm (will be discussed next)
  - Best route placed into routing table
  - Best route passed next to Output Policy Engine

Well-known Mandatory Attributes

AS-PATH

- Necessary for path-vector algorithm function
- AS which gets the route prepends it's number to the beginning
- AS doesn't accept route if AS-PATH already contains it's own AS number
- Route with shorter AS-PATH is preferred

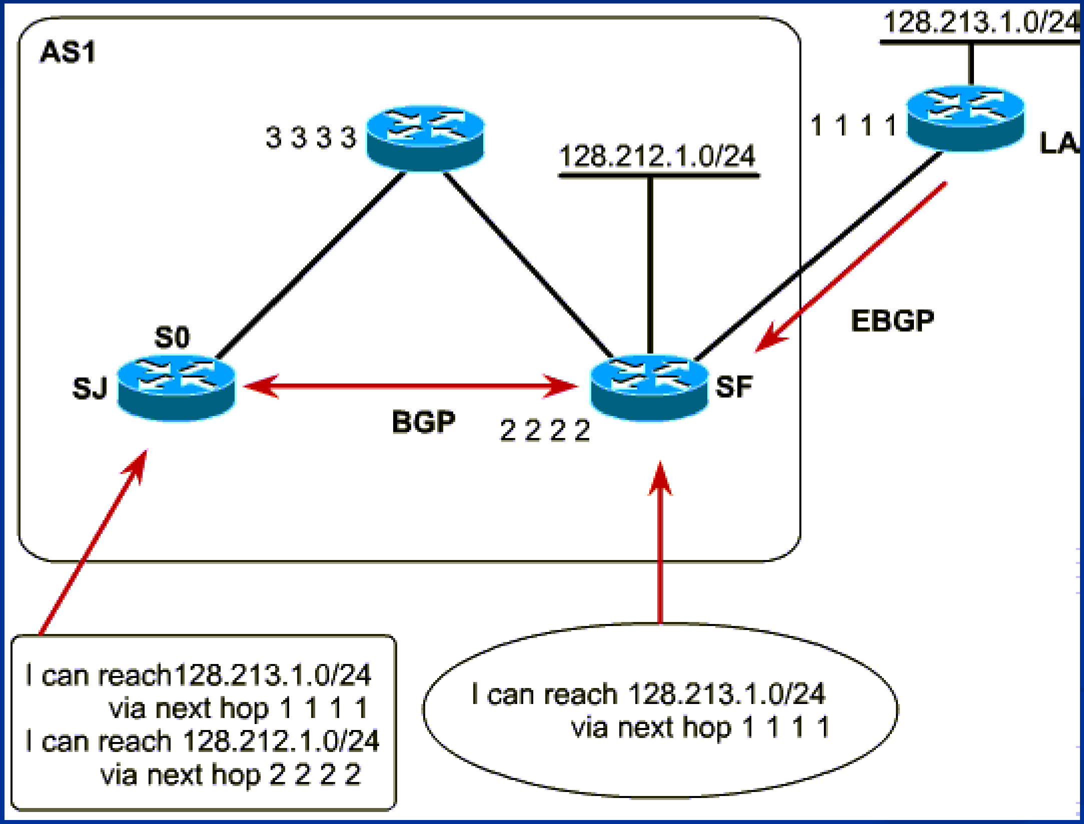
AS-PATH manipulation

- AS-PATH handled as string
  - (AS numbers separated by spaces)
- Regular expression used to test presence of some pattern (AS sequence)
  - Originating AS, AS in path, …
- Inserting AS number multiple times makes AS-PATH longer and route less preferred
  - Router can insert only it’s own AS number (possibly multiple times)

NEXT-HOP

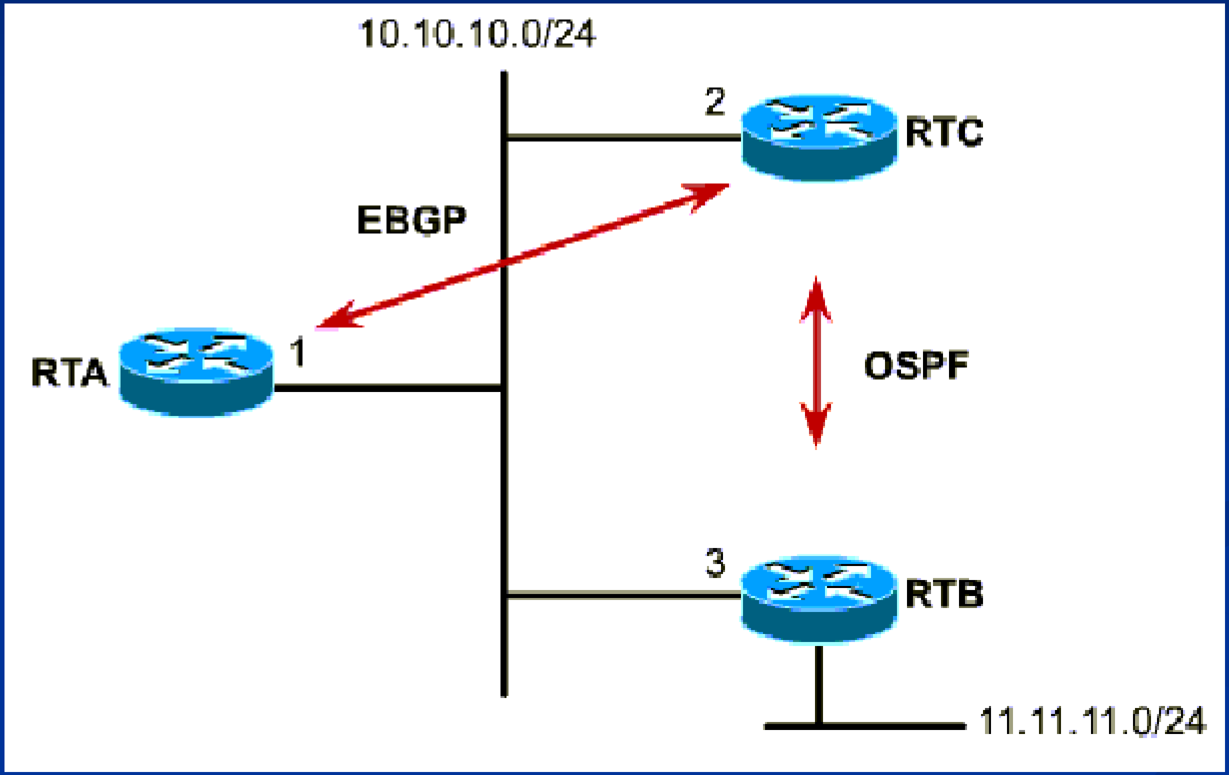
- Next hop of BGP route is boundary router which propagated that route into AS
  - Difference from IGP – not neighbor on the same link
- Router has to know route to next-hop address from IGP (or IBGP)
  - Otherwise, BGP route is not accepted
- Recursive routing table lookup when routing packets

NEXT-HOP



- Line between ASes propagated into IGP

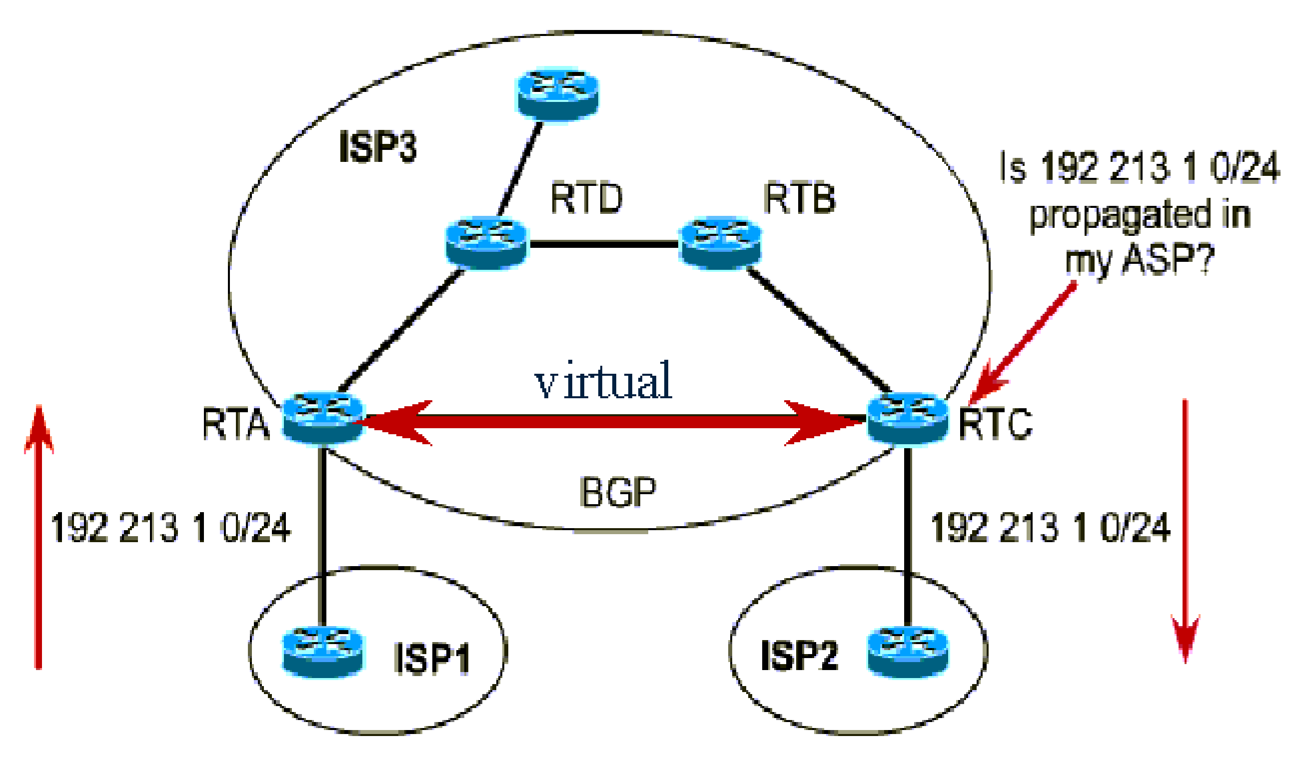
NEXT-HOP on broadcast network



ORIGIN

- Informs where BGP learnt the route from
  - IGP – redistributed from IGP
  - EGP – unused (from outdated protocol EGP)
  - INCOMPLETE – unknown origin

BGP and IGP synchronization problem



Route synchronization

- Route is synchronized, if router can see it both from BGP and IGP
- Only synchronized routes are propagated out of AS
  - Otherwise, traffic would have to be discarded by internal routers
- When IBGP is ran on every router, switch off the synchronization test

Transit system routing implementation choices

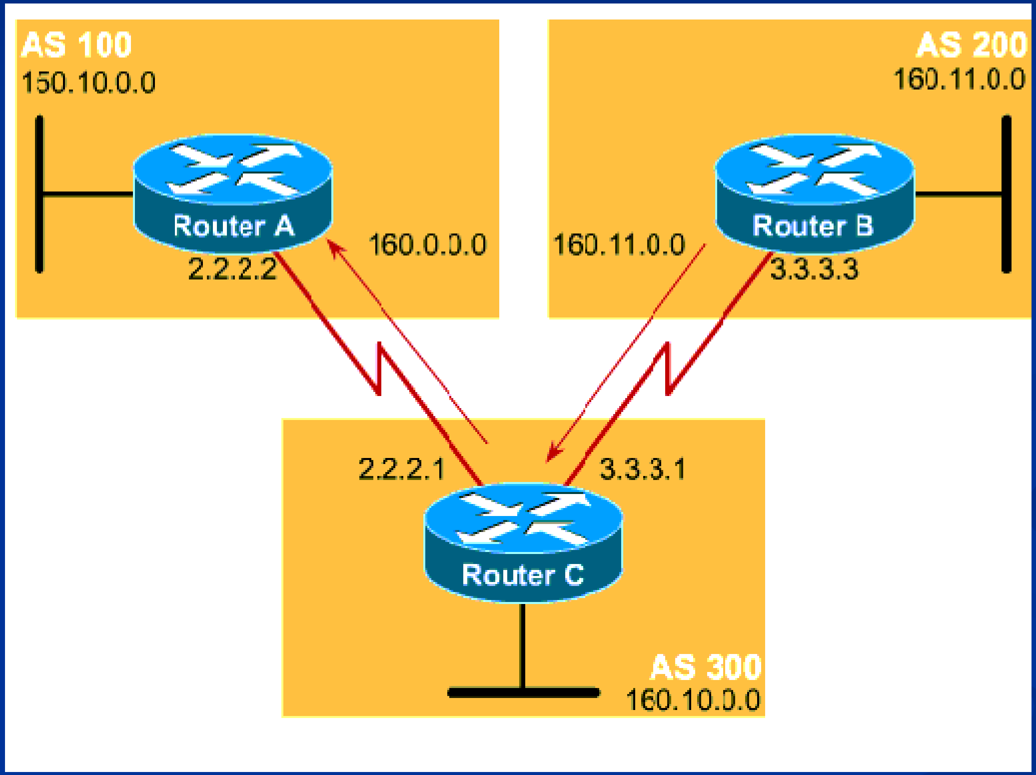
- BGP on every router (IBGP)
  - At least on every transit router
  - Common solution of ISPs
- Redistribution of BGP routes into IGP
  - But IGPs are not capable to handle so many routes

Route aggregation in BGP

Aggregation Attributes

- Router can aggregate more routes into one with shorter prefix
  - Only when aggregator "owns" whole address range
- ATOMIC-AGGREGATE=True
- AGGREGATOR: ID of aggregating router
- AS-SET= AS-PATH_1+AS-PATH_2
- AS-PATH: set as if route originated from AS of aggregating router

Aggregation Example

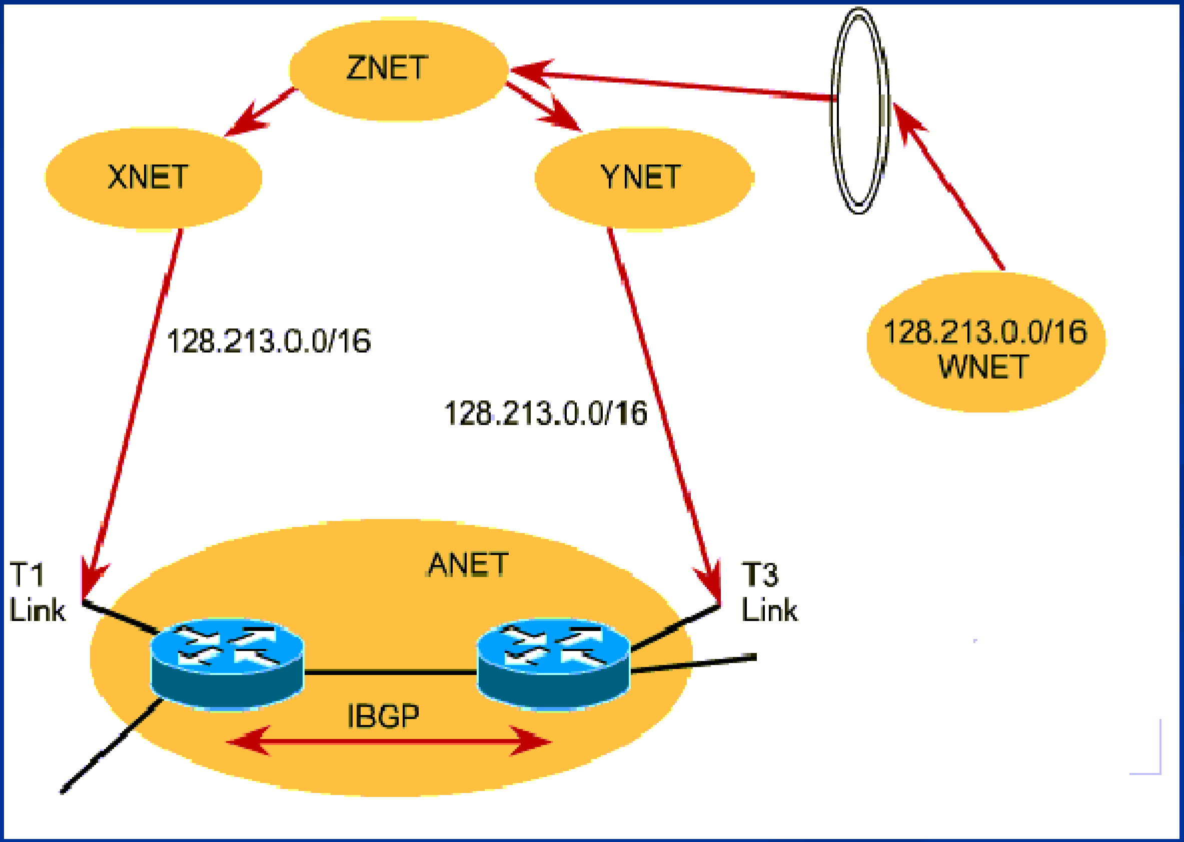


How to influence route selection using attributes

LOCAL_PREFERENCE

- Well-known discretionary
- Allows routers of one AS to unify exit link they will use to reach some particular external network
  - Route with higher LOCAL_PREFERENCE is preferred
- Never passed behind AS boundary

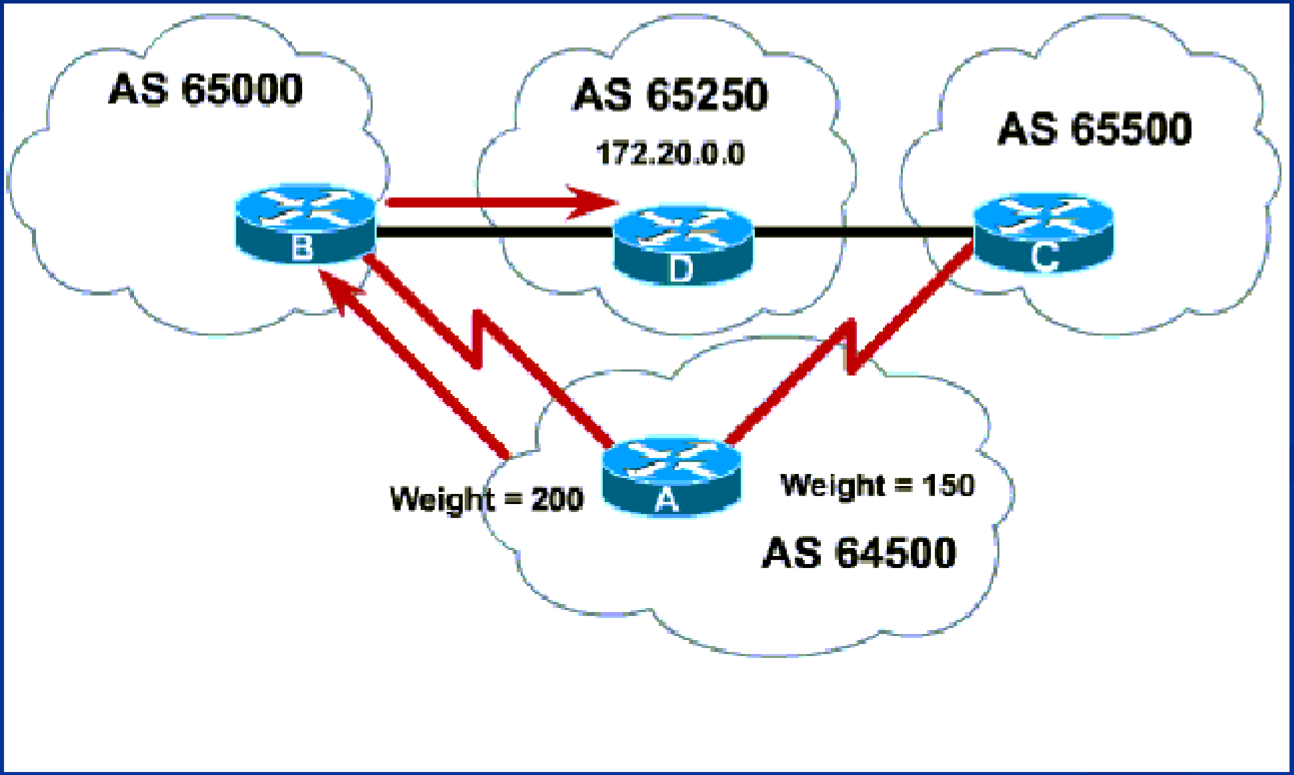
LOCAL_PREFERENCE Example



WEIGHT

- Proprietary (Cisco, …)
- Used to increase/decrease preference of some route in Input Policy Engine
  - Higher Weight is preferred
- Only local significance, does not passed outside of single router
  - In fact, not a standard-defined attribute

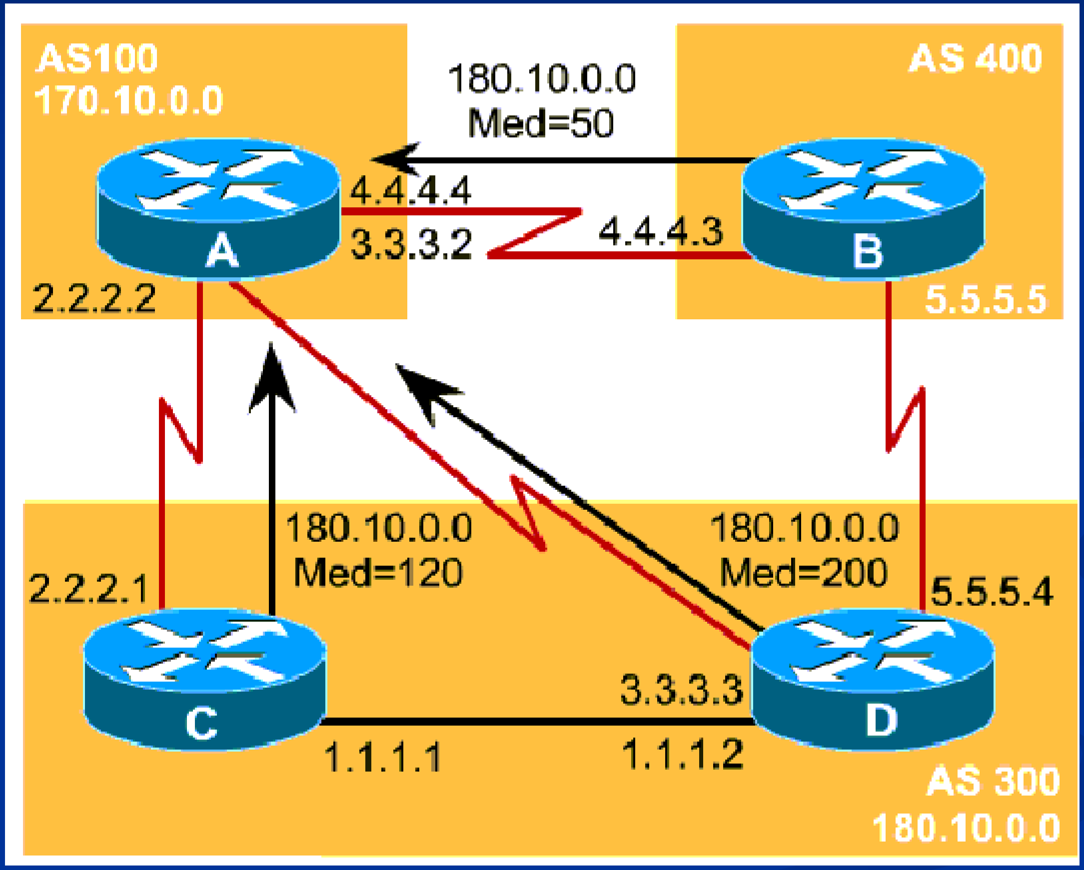
WEIGHT example



Multi-Exit Discriminator (MED)

- Influences other AS's decision which link to use when routing packets into networks inside "our" AS
- Lower MED is preferred
  - treated similary like IGP metric
  - MED value can be set manually or taken from IGP metric
- Normally, only MEDs from the same AS may be compared

MED example



AS100
170.10.0.0
180.10.0.0
Med=50
AS 400
4.4.4.4
3.3.3.2
4.4.4.3
A
B
2.2.2.2
5.5.5.5
180.10.0.0
Med=120
180.10.0.0
Med=200
2.2.2.1
5.5.5.4
C
D
3.3.3.3
1.1.1.1
1.1.1.2
AS 300
180.10.0.0

Route Selection Algorithm

1. Higher WEIGHT
2. Higher LOCAL_PREFERENCE
3. Route generated by router itself
4. Shorter AS_PATH
5. More preferred ORIGIN
   - (IGP best, INCOMPLETE worst)
6. Lower MED
7. EBGP preferred over IBGP
8. Better IGP metric to NEXT-HOP
9. Lower peer Router_ID (tiebreaker)