La Sapienza
Università degli Studi di Roma
Dipartimento di Informatica e Sistemistica

Computer Networks II
BGP - **B**order **G**ateway **P**rotocol

Luca Becchetti
Luca.Becchetti@dis.uniroma1.it

A.A. 2009/2010

Routing between Autonomous Systems -- BGP

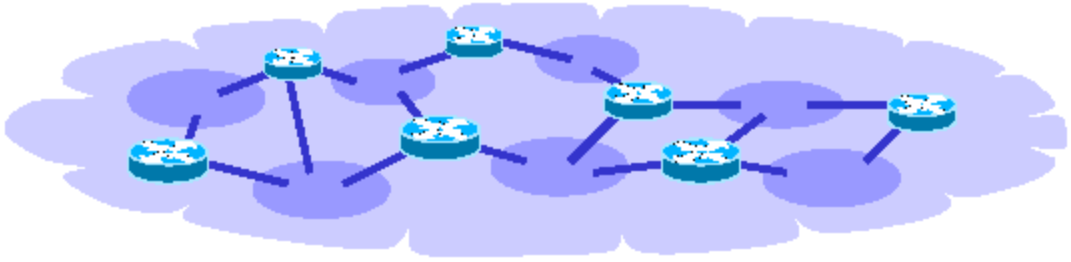
Thanks to:

Giuseppe Di Battista, Maurizio Patrignani, Maurizio Pizzonia:
Università di Roma Tre

Timothy G. Griffin
http://www.research.att.com/~griffin/interdomain.html

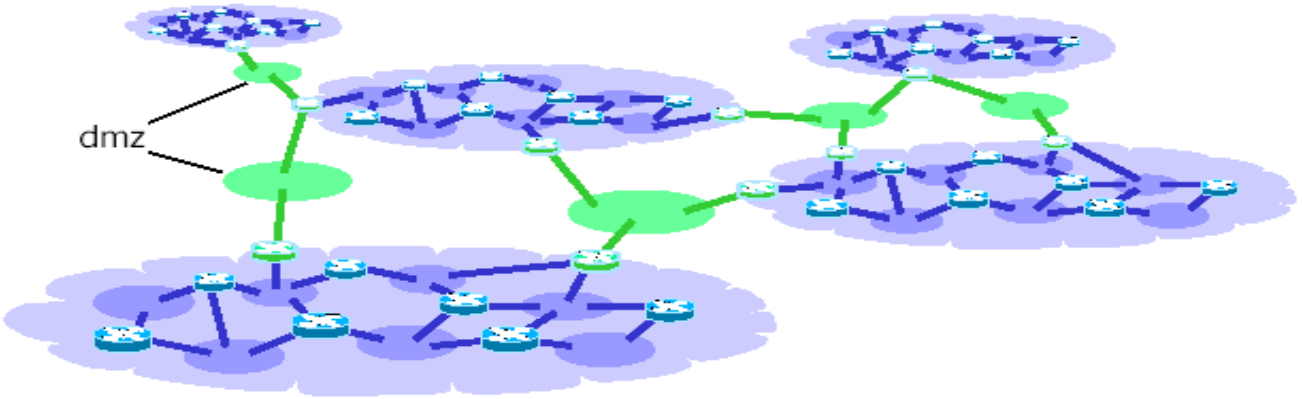
Autonomous systems

- Every organization's network consists of a set of routers under a single administration
- A routing algorithm is used to maintain routing tables within the AS



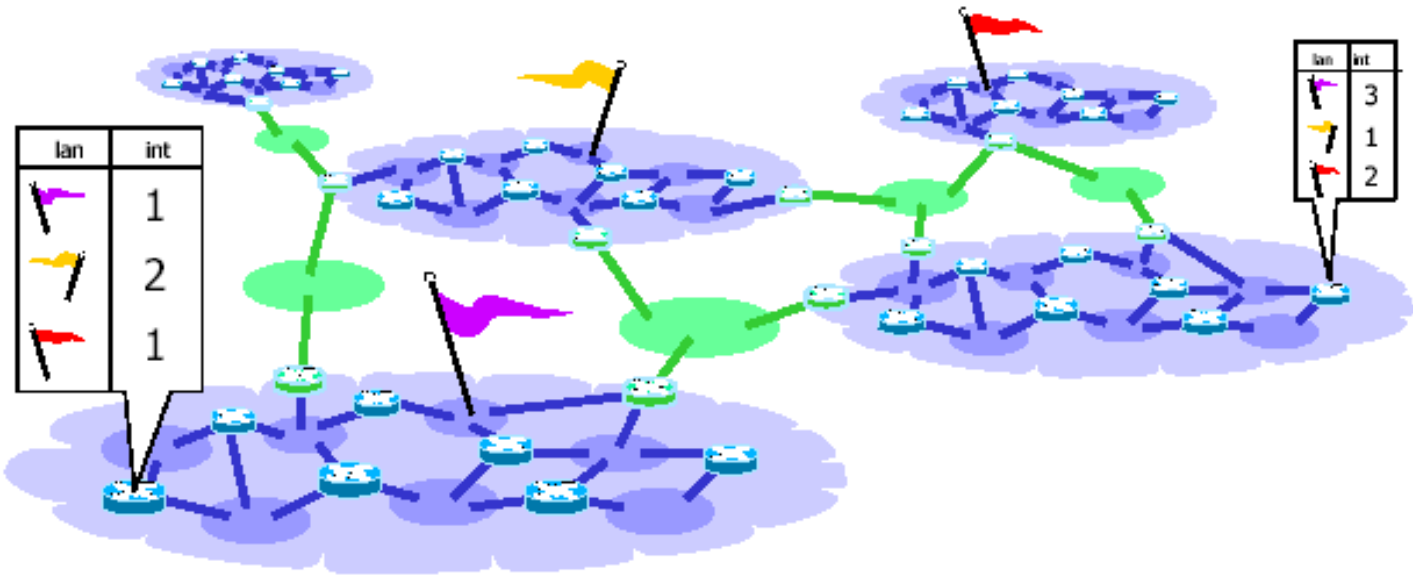
Ases interconnection

- When 2 or more organizations connect their networks into an inter-netwotks they need to establish *connection points*
- Added networks are said demarcation point



Routing between different ASes

- Every routing table must have an entry for *every possible destination*
- This has to hold for both *local* and *global* destinations



How to update routing tables in this case?

Three options in general:

- All organizations use the same routing algorithms
- Routing tables are updated manually, adding *static predefined routes*
- Combine an intra-domain with an inter-domain routing protocol: Exterior gateway protocol

1. Using one routing algorithm

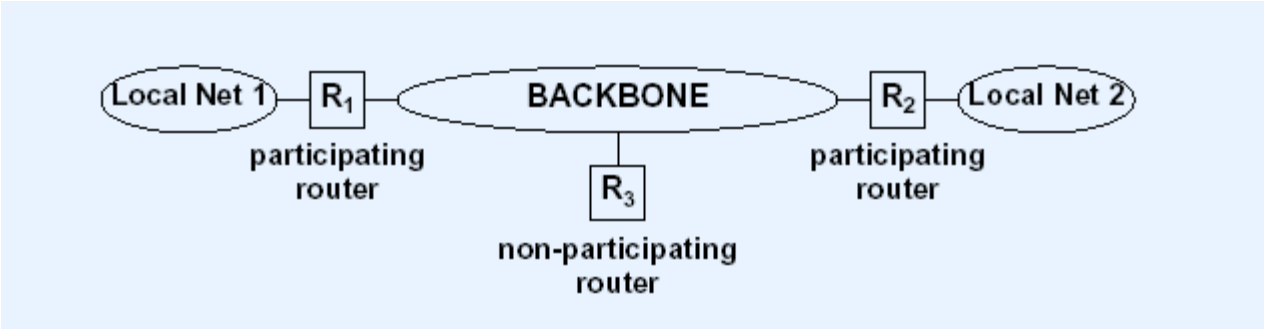
- Many drawbacks:
  - Propagation delay, ex: distance vector
  - Scalability
  - All organizations forced to use same algorithm
  - Difficult to adopt a new routing algorithm (everybody must change!)
  - “Political” and commercial relationships between ASes not considered

2. Static routes

- Hide internal part of AS
- For every external destination --> identify router at the border of destination AS
- Information about the path to reach the target
- Drawbacks:
  - Hard to maintain and fix
  - No (automatic) management of faults, no backup
  - No guarantees that all routers on source destination path available for traffic

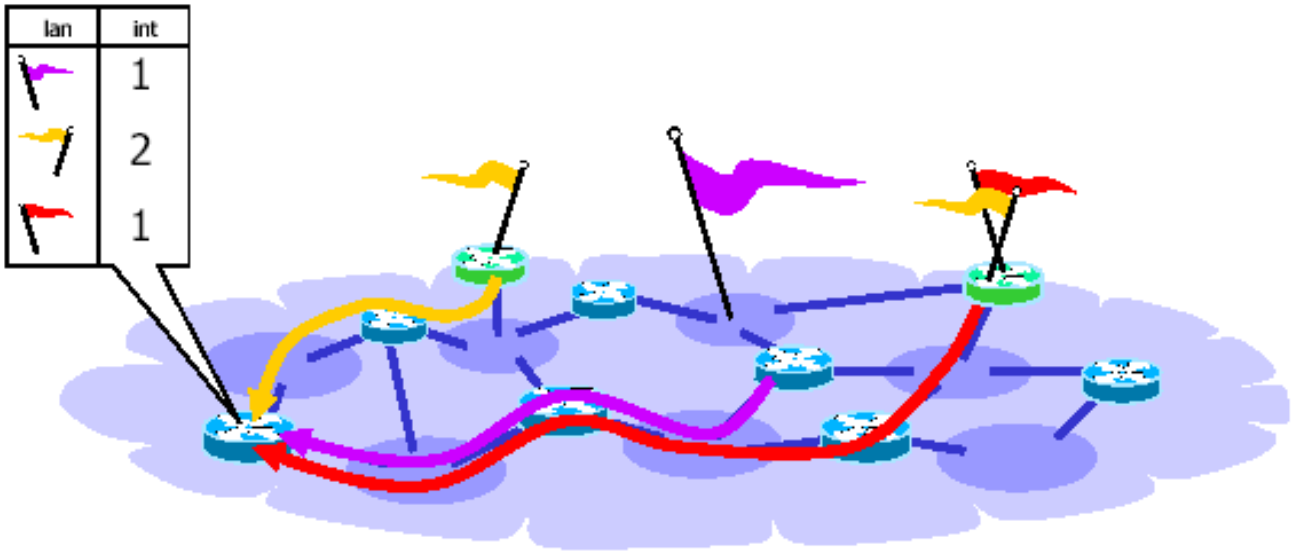
2. Static routes

- Routing may be inefficient
- In the example: R1 and R2 are part of the same AS. R3 forwards to R1 all traffic directed to the AS, including the one with destination LAN 2.
- Routing does not keep into account networks that are actually reachable



2. Static routes

- Routing algorithm diffuses within AS local traffic and traffic following static routes

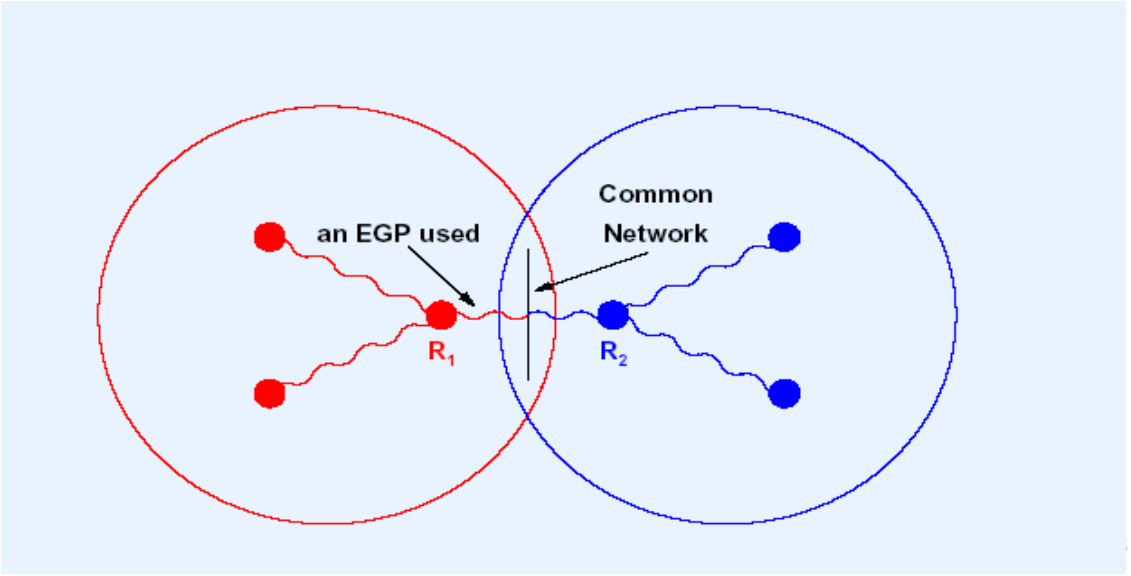


A different approach

- Occorre avere un flusso informativo in due direzioni, sia dall'interno verso l'esterno che dall'esterno verso l'interno
- L'AS si deve far carico di garantire la consistenza degli instradamenti interni
- Occorre annunciare all'esterno quali reti interne sono raggiungibili
- Occorre assegnare le responsabilità per la diffusione delle informazioni riguardo l'instradamento

3. Exterior gateway protocol

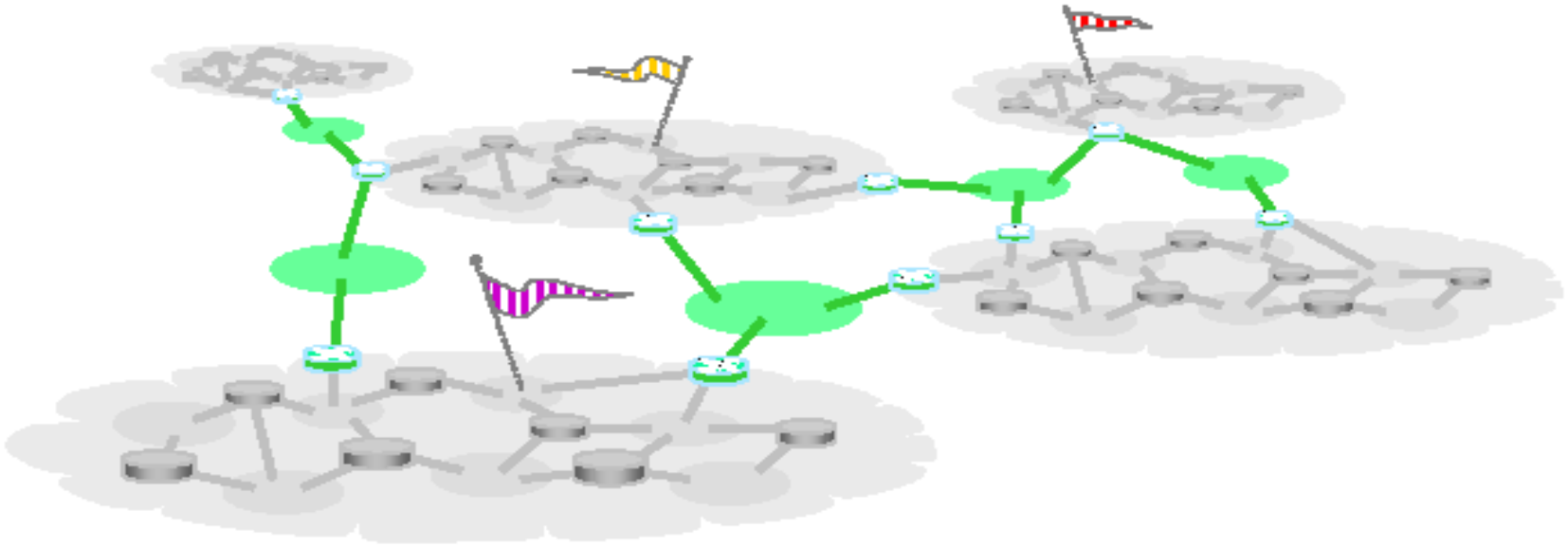
- Any protocol for the exchange of routing information between ASes
  - Also: a specific protocol, prior to BGP
- BGP – Border Gateway Protocol
- Two Ases exchanging routing information elect (at least) two routers to this purpose, which establish a peering session
- BGP routers are *border routers* or *gateways*



3. Exterior Gateway Protocol

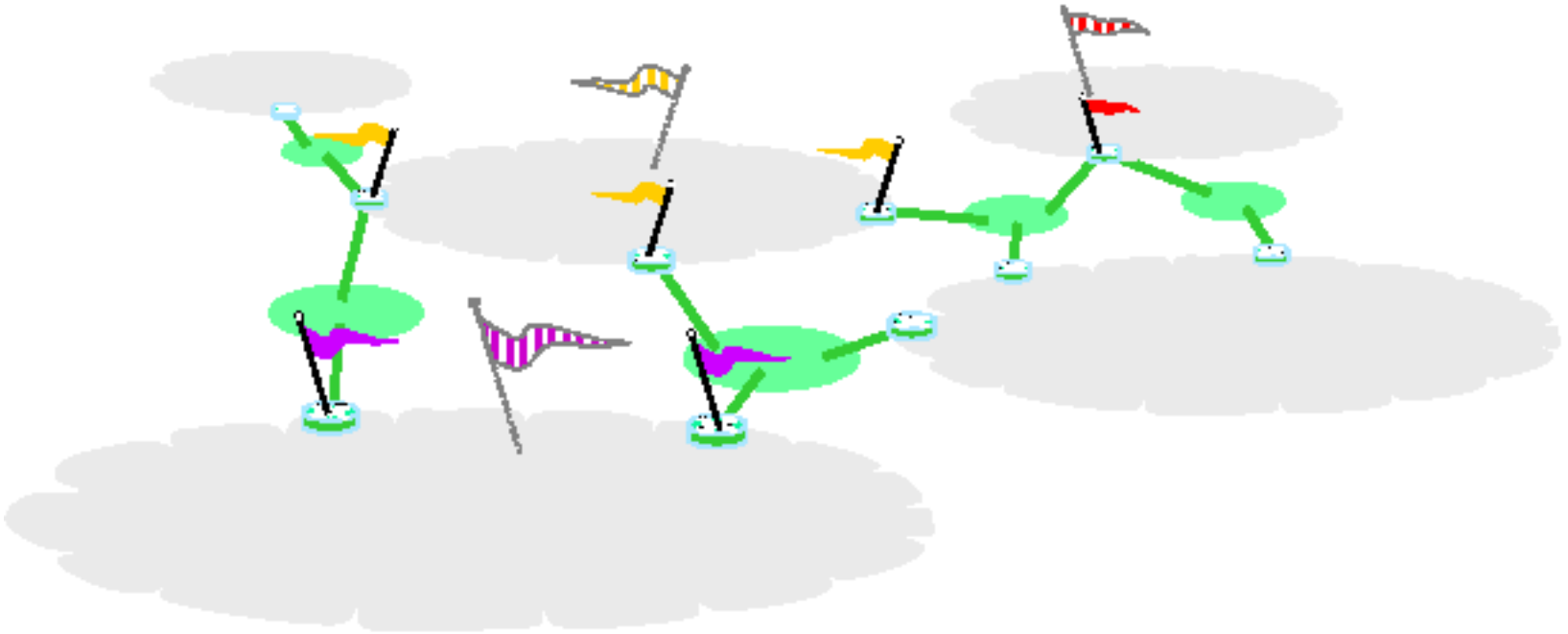
Approach:

- Inner part of Ases “hidden”
- Only summary information exchanged at AS borders by border routers



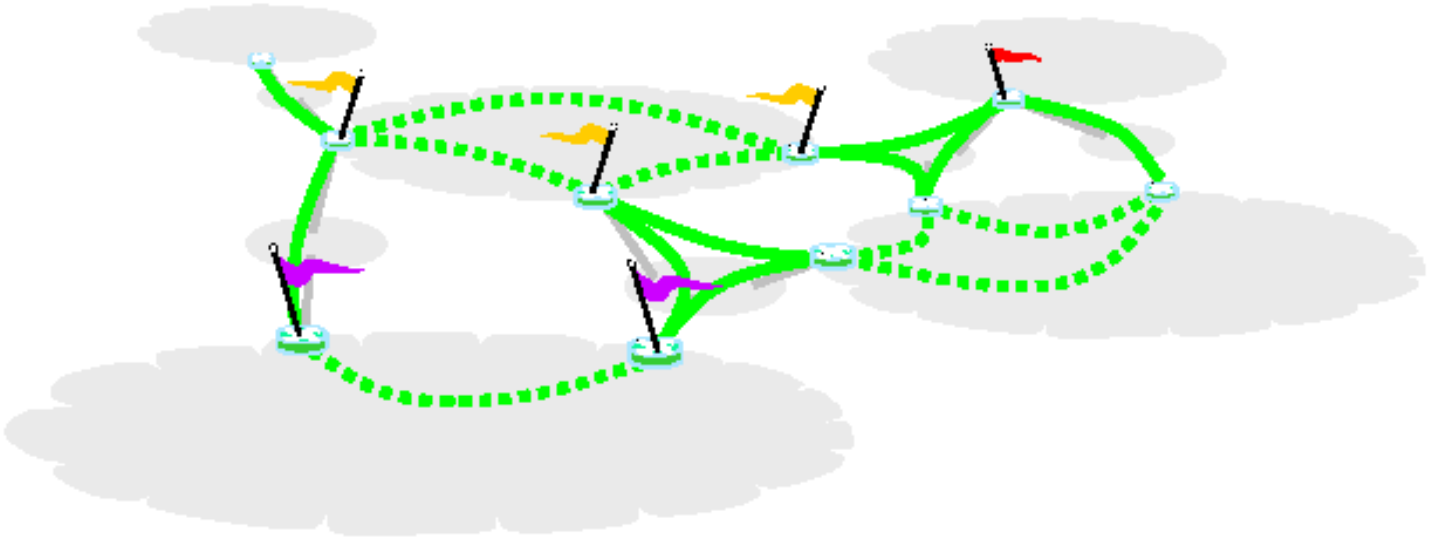
3. Exterior Gateway Protocol

- Every border router represents internal destinations *as if they were local*



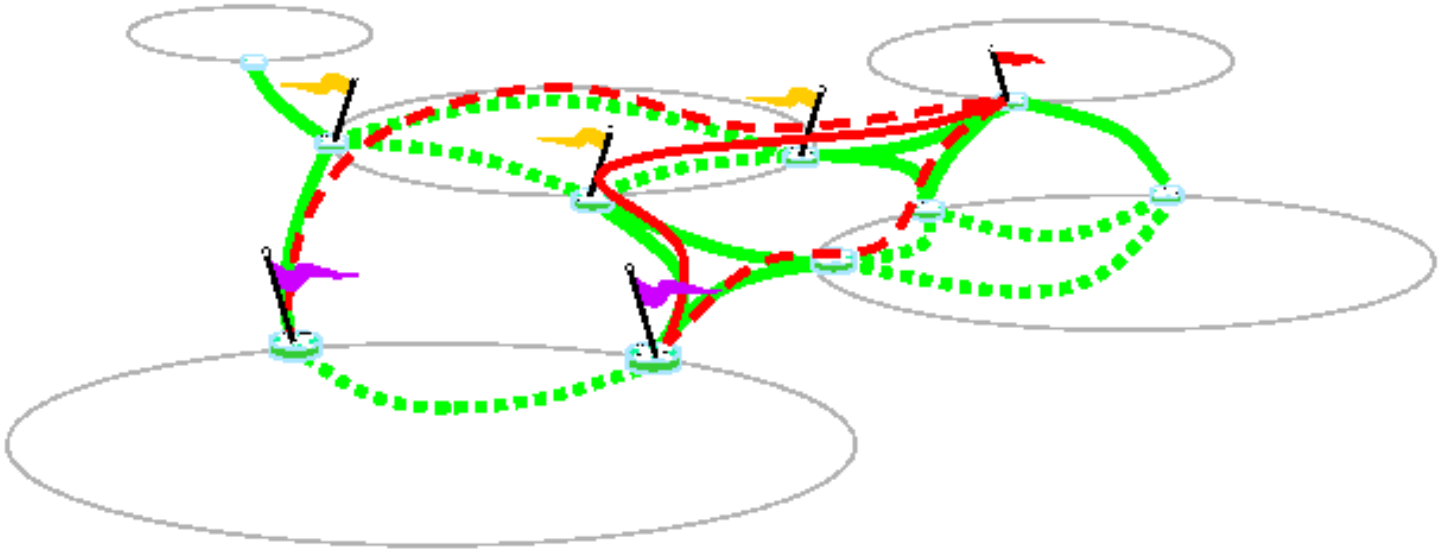
3. Exterior Gateway Protocol

3. Exterior Gateway Protocol



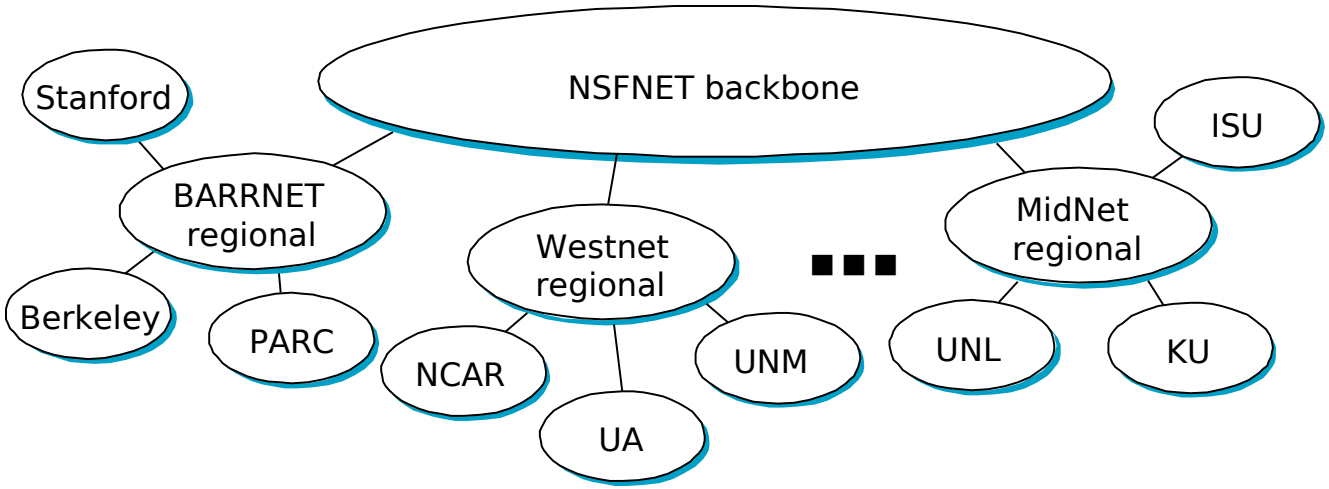
3. Exterior Gateway Protocol

- Possible to define pre-defined paths on the basis of policy considerations



3. Exterior Gateway Protocol

- Designed when the Internet was organized as described in the picture below
- AS graph had tree-like structure



Exterior Gateway Protocol (EGP)

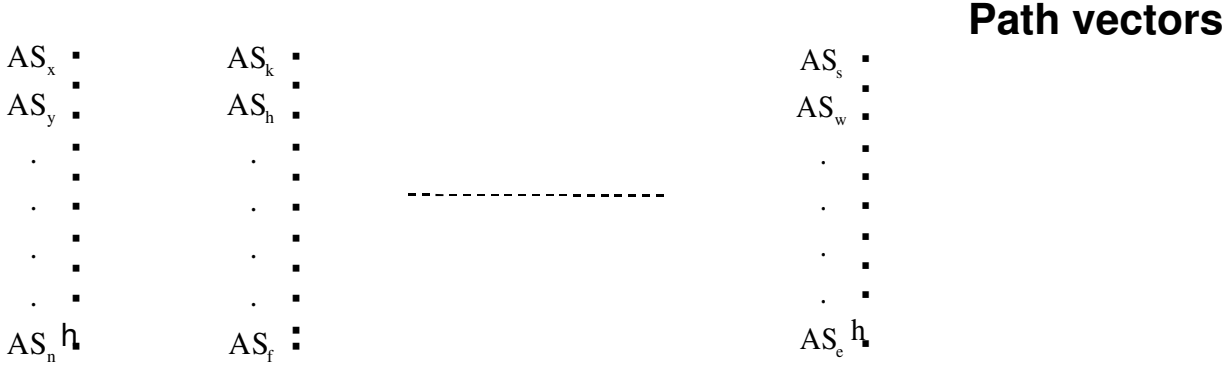
- Distance-vector routing protocols (e.g., RIP) unsuited as EGPs
  - Assume same metric for all routers; cannot be assumed for different ASes
  - No indication of intermediate routers between source and destination; in inter-AS scenario, prefferred or forbidden paths are possible
- Link-state routing protocols (e.g., OSPF) unsuited as EGPs
  - Different ASes may use different metrics
  - Flooding not applicable among different ASes
- Inter-AS routing does not only depend on efficiency (shortest path); it often depends on different factors [e.g., commercial agreements]
  - For this reason: EGP protocols **don't use metrics to define paths but only reachability info**

Exterior Gateway Protocol (EGP)

- **Path vector routing** technique used in EGP protocols
- Only information about the following aspects is used:
  - Which networks are reachable over a given router
  - Which routers are traversed by a path
- No notions of distance or cost (almost)
- Determine list of ASes that have to be traversed along the path to reach a specific destination network
  - Routing will keep preference for certain ASes with respect to others (trade agreements, performance, etc.) into account

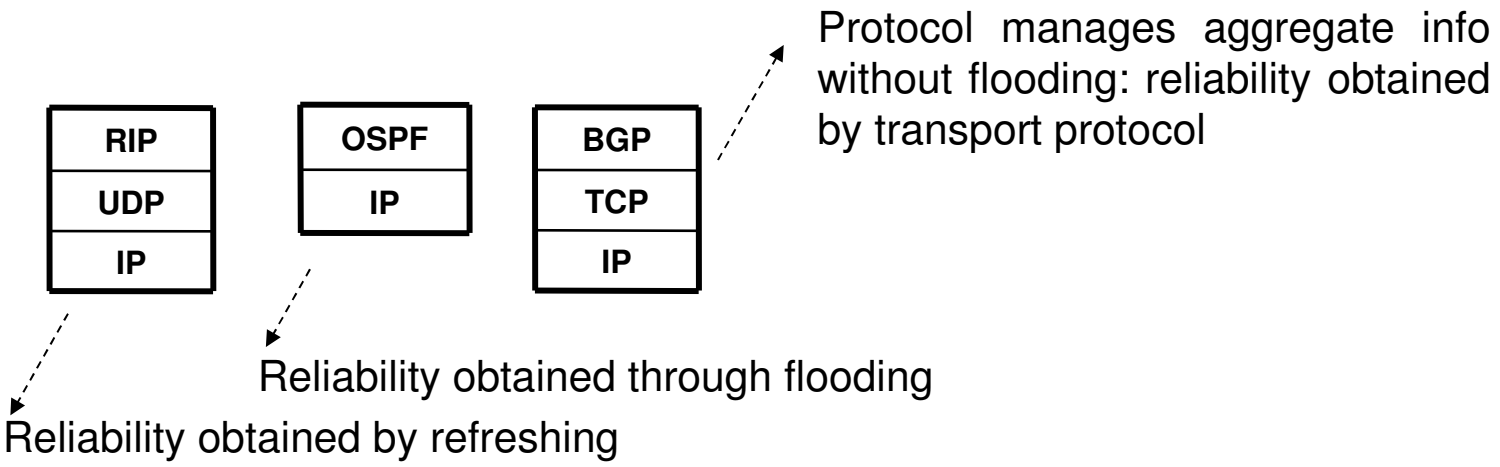
Path Vector Routing Protocols

- Routing based on “path vector”
- Idea: for every < origin subnet, destination subnet>, collect a list of alternatives, where each alternative is a sequence of ASes to traverse in order to reach destination subnet from the origin one
  - One of the possible alternatives selected on the basis of specific criteria [agreements, preferences,…] from AS sending a packet



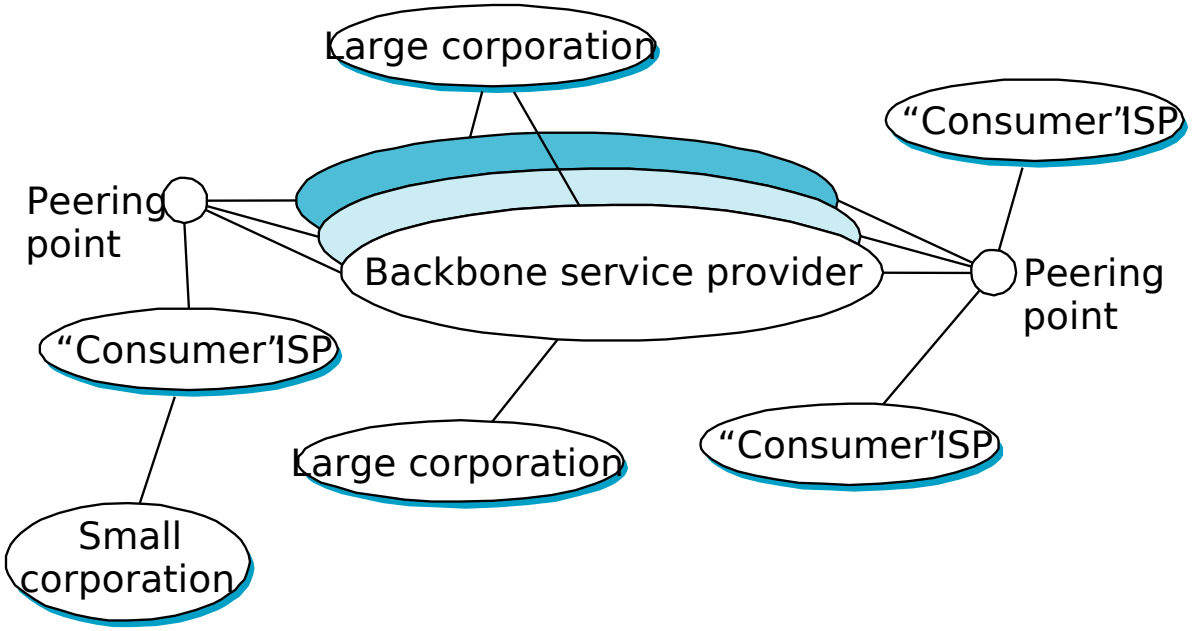
Border Gateway Protocol (BGP)

- BGP:
  - Allows routers belonging to different ASes to exchange reachability information
  - Supports CIDR [in particular, also variable-length subnetting]
  - BGP message exchange relies on TCP
  - Most recent version: BGPv5
  - Comparison:



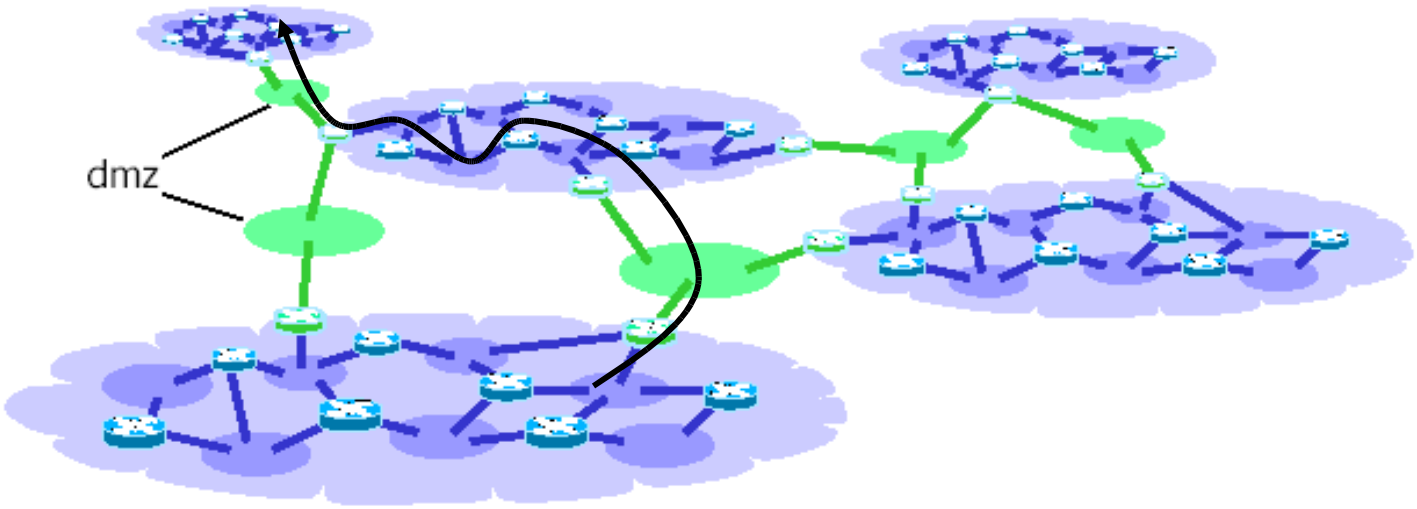
BGP v4 – Border Gateway Protocol

- No assumption on AS graph
- More interconnected backbone networks
  - Service provider networks
- Many Service Providers exist



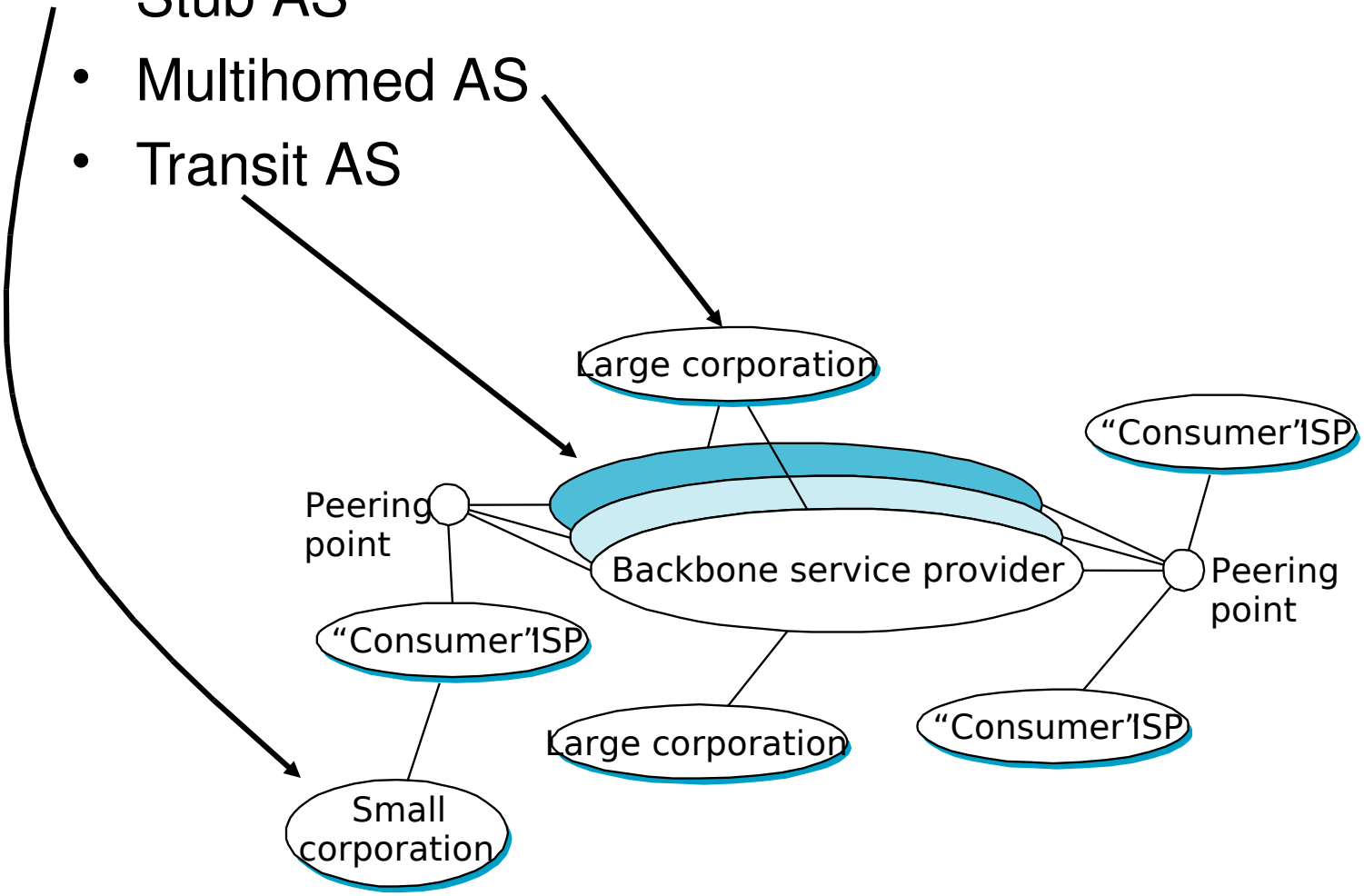
BGP v4 – Border Gateway Protocol

- *Local traffic*
  - begins **or** ends at internal nodes
- *Transit traffic*
  - Crosses borders between different ASes



BGP v4 – Border Gateway Protocol

- Stub AS
- Multihomed AS
- Transit AS



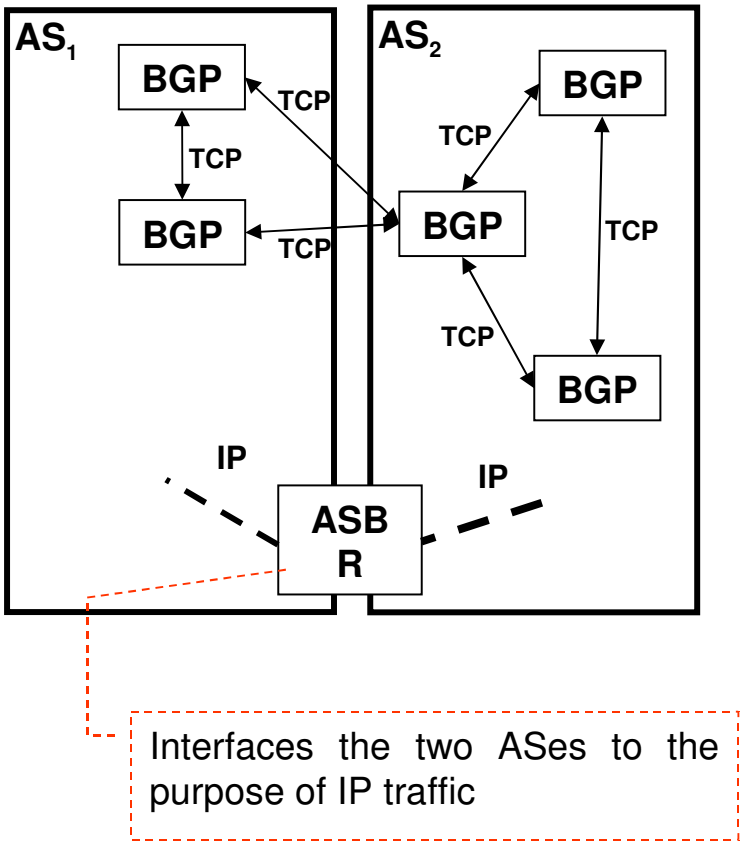
BGP v4 – Border Gateway Protocol

Each AS has:

- One or more border routers
- One BGP *speaker* advertises:
  - local networks
  - other reachable networks (transit AS only)
    - There is always a default route
  - gives *path* information

BGP Terminology

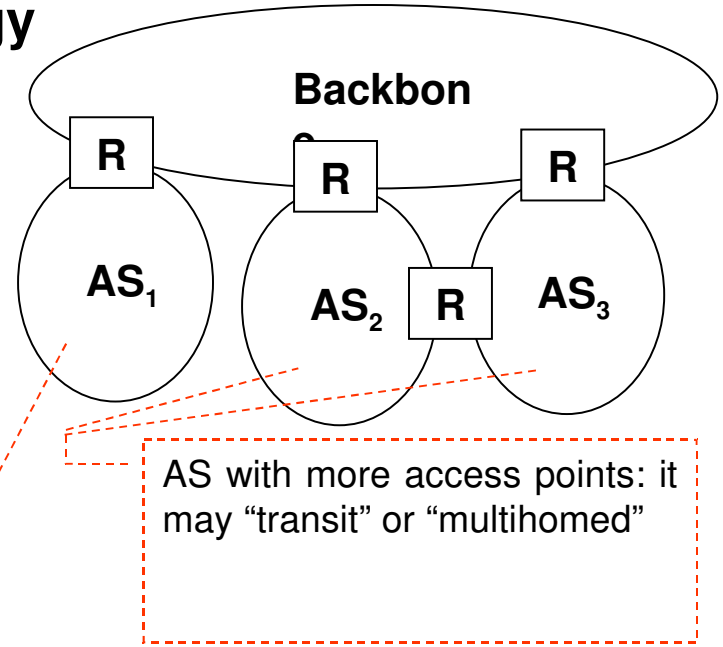
- **BGP speaker**
  - A router supporting BGP
  - A BGP router is not necessarily a border router (ASBR)
- **BGP Neighbors**
  - Pair of BGP speakers exchanging *inter-AS* routing information
  - Two types possible:
    - *internal*: if they belong to same AS
    - *external*: if they belong to different ASes
- **BGP session**
  - The TCP connection supporting a BGP session between two BGP speakers



BGP Terminology

- **AS Border Router (ASBR)**
  - A router connected to other ASes
  - *Internal*
    - An ASBR belonging to the same AS as a BGP speaker under consideration
  - *External*
    - An ASBR that is in a different AS than a BGP speaker under consideration
- **AS connection**
  - Physical connection
    - Two Ases share the physical subnet
  - BGP connection
    - There is BGP session between a pair of BGP speakers belonging to different ASes

BGP Terminology



- **Traffic**
  - *Local*
    - Originating or destined to *this* AS
  - *Transit*
    - Non local traffic
- **AS type**
  - Stub
    - It has just one single inter-AS connection, it only carries local traffic
  - Multihomed
    - It has multiple connections to different ABes but it does not carry transit traffic
  - Transit
    - It has multiple connections to different Ases and it carries transit traffic

BGP Terminology

- **AS number**
  - *Unique* 16-bit identifier for an AS
- **AS path**
  - List of ASes traversed by a path
- **Routing policies**
  - No fixed rules to select inter-AS paths; every AS adminstrator can define own rules
    - A multi-homed AS may refuse to forward transit traffic
    - A multi-homed AS may allow transit traffic, but only for some ASes
    - An AS may select ASes to which it forwards transit traffic
  - Among possible choices, a BGP speaker selects the most suitable to meet the routing policy requirements decided by the AS administrator
  - In the presence of multiple paths to the same destination, BGP speaker keeps all, *but only one* is communicated to other ASes

BGP

- Il protocollo BGP impone che un AS presenti la stessa visione a tutti gli AS che usano i suoi servizi
  - questa condizione è garantita dal protocollo IGP (es. OSPF)
- Il protocollo BGP di un AS comunica ad altri AS solo cammini che lo usano come next-hop
  - conforme al classico schema di routing in IP
- Due BGP speaker, dopo aver instaurato una sessione, si scambiano i path completi verso ogni altro AS di destinazione
  - un path è indicato sotto forma di lista di AS
  - la disponibiltà dell'intera lista di AS **evita l'insorgere di loop**

BGP functionalities

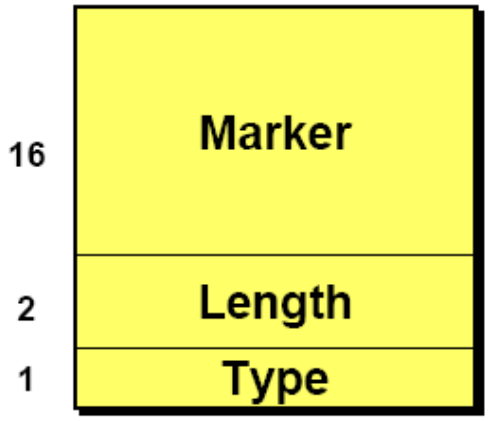
- **Neighbor Acquisition Procedure**
  - Used when two AS routers sharing the same subnet want to start the exchange of reachability information
  - Both have to agree in order to avoid overload
  - Procedure consists in sending one request (Open message) and one reply (Keepalive message)
  - Procedure can be started by network manager

BGP functionalities

- **Neighbor Reachability Procedure**:
  - Used to keep sessions between BGP routers active
  - Every router ensure that peer is alive and maintains session
  - How: routers exchange keepalive message periodically
- **Network Reachability Procedure**:
  - Every router maintains a database of the networks that are reachable and the preferred path to reach them
  - When database changes router sends Update message to BGP peers to inform them about change

BGP Messages

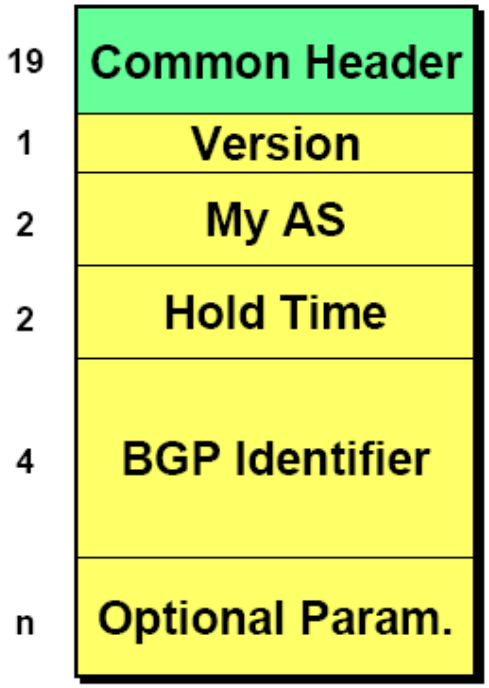
- **Header (19 bytes)**
  - Common to all BGP messages
  - Marker (16 bytes)
    - Used to allow destination of message to authenticate and identify sender
  - Length (2 bytes)
    - Message length in bytes
  - Type (1 byte)
    - Type of the message (Open, Update, Keepalive,Notification)



BGP Header: 19 bytes

BGP Messages

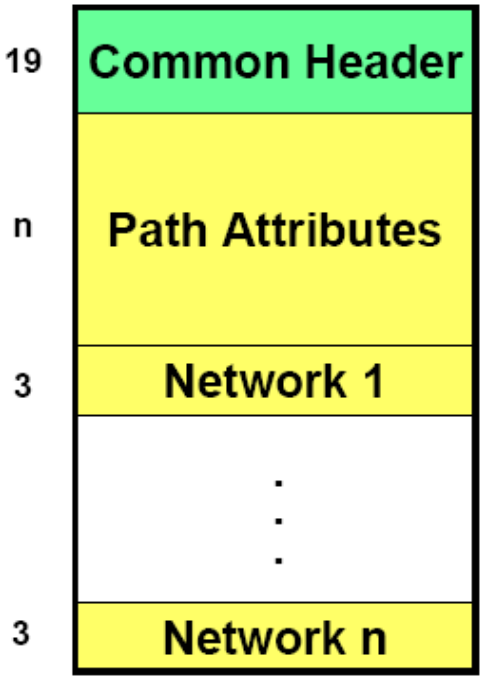
- **Open Message**
  - Used in Neighbor Acquisition procedure
  - My AS
    - Identifier of AS to which router belongs
  - Hold time
    - Proposed duration for timer used in keepalive procedure
  - BGP identifier
    - Router IP address



BGP Messages (Update)

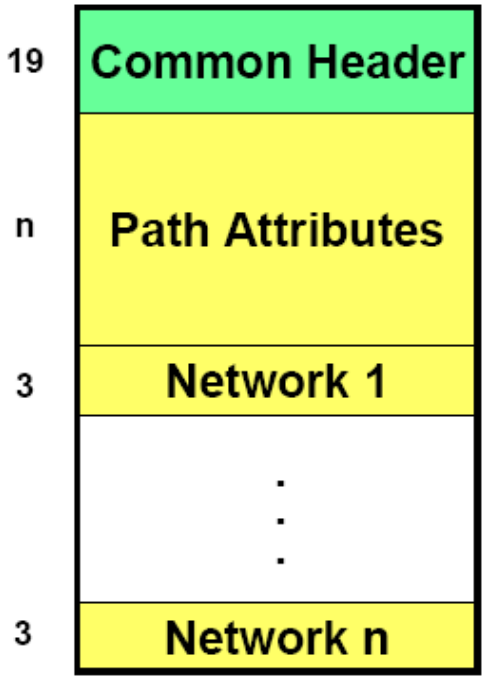
"key" message for BGP: which networks are reachable over the path, with the path described as a sequence of ASes to traverse

- **Update Message**
  - Used to send to peer BGP routers reachability information about a *single path*
  - Path Attributes
    - Describes characteristics of proposed path
  - Network 1,…, Network n
    - List of network addresses for networks reachable over proposed path
    - Can be specified in CIDR notation



BGP Messages (Update cont.)

- **Path Attributes**
  - Origin
    - Indicates origin of information: IGP, EGP or "incomplete"
  - AS_Path
    - List of ASes traversed by *this* path
  - Next_hop
    - IP addres of next BGP router on this path
  - Multi_Exit_Disc
    - Info on internal routing within an AS
  - Local_Pref
    - Degree if preference for *this* path
  - Atomic_Aggregate, Aggregator
    - Aggregation of network addresses, useful for hierarchical routing

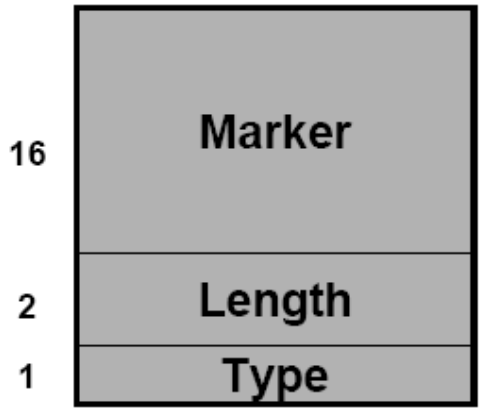


BGP Messages

- Update messages sent at start of peering realtionship between two BGP routers and then when path changes occur
- A router receiving an Update message compares path received with the one currently used for the given destination
  - If new path is “better” --> old path substituted and this is notified to peer routers
  - If new path “worse” than current one --> no modifications

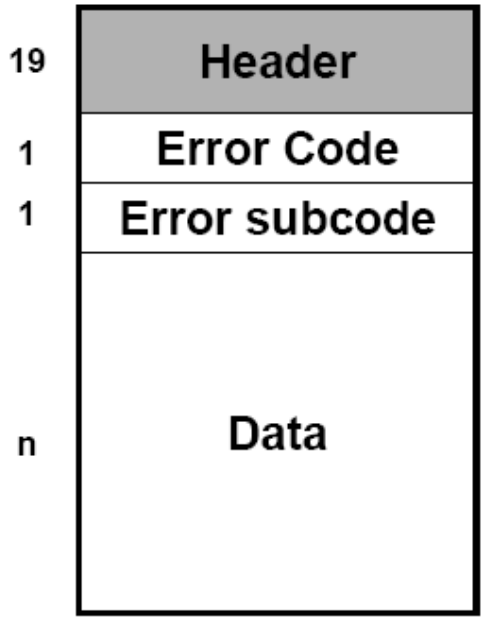
BGP Messages

- **Keep-alive Message**
  - Used to notify activity of a BGP router --> resets Hold Timer
  - Hold timer decremented between receipt of two consecutive BGP messages (Keep-alive or Update) from peer
  - If Hold timer expires before next message arrives --> BGP peer session is reset
  - Ensures reachability of sending router
  - Contains only header bytes

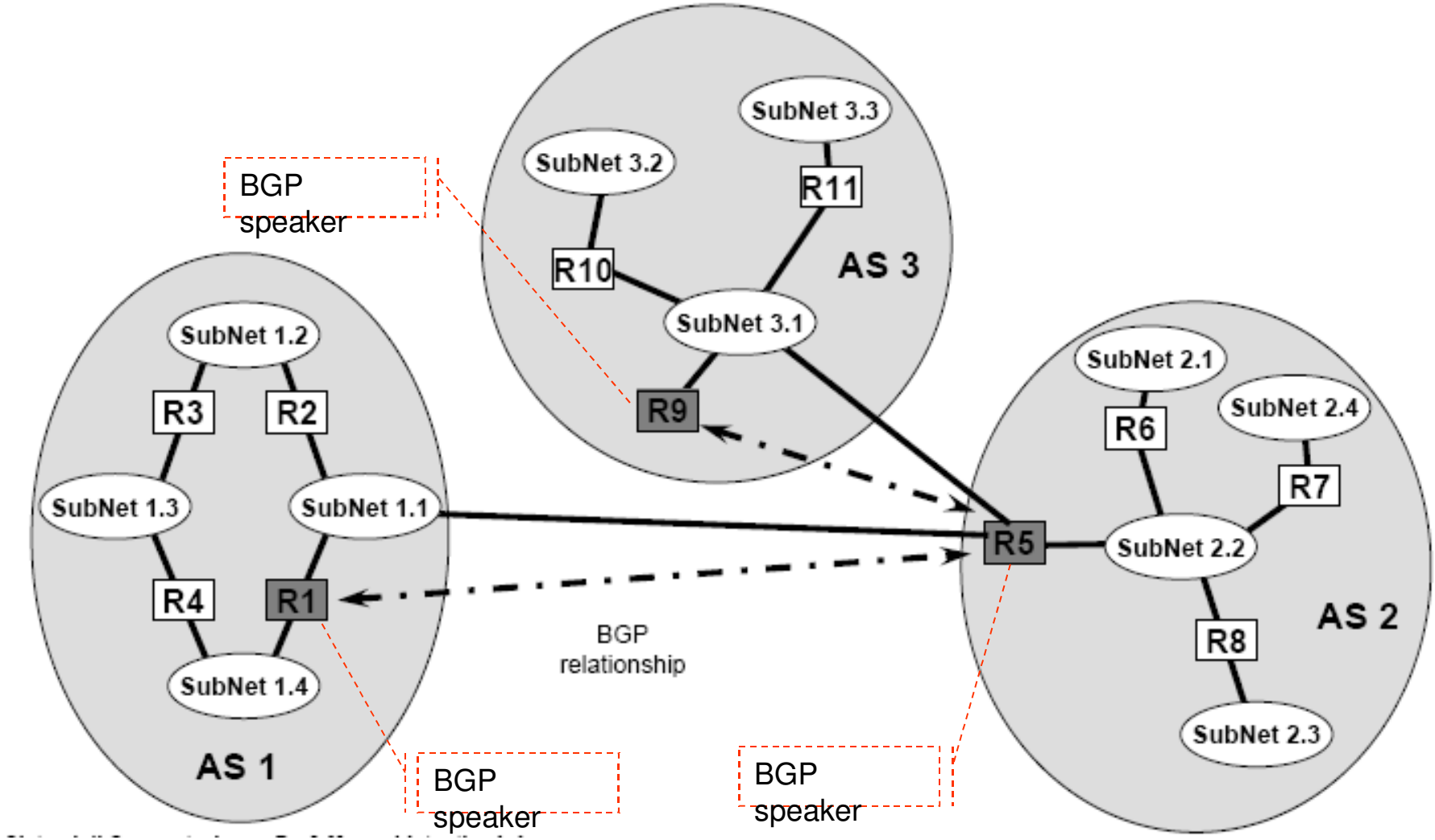


BGP Messages

- **Error Notification Message**
  - Used to notify error to peer routers
    - Hold Timer expiry
    - Procedure errors or wrong messages
    - Address errors
    - …



Example BGP 1/3



Example BGP 2/3

- Routers R1 and R5 implement both BGP and an IGP (e.g., OSPF); as a consequence, R1 knows the structure of AS1
- R1 sends Update message to R5 containing:
  - AS1's identifier
  - R1's IP address (highest IP address among all interfaces)
  - List of subnets belonging to AS1
- R5 stores reachability of AS1's subnets over R1
- R5 sends Update message to R9 containing:
  - Identifiers of AS1 and AS2
  - R5's IP address
  - AS1's subnet list

Example BGP 3/3

- Update message from R5 notifies R9 that AS1's subnets are reachable over router R5 and that path traverses both AS2 and AS1 in this sequence
- Finally, R9 sends Update message to all its peers containing:
  - Identifiers for AS1, AS2 and AS3
  - R9's IP address
  - AS1's subnet list
- In this way, reachability informations spreads over the network

BGPv4 – more details

- Numerazione, peering e scambio di messaggi
  - Messaggi BGP
  - EBGP e IBGP
- Annunci BGP - Route advertisement
  - Messaggi di UPDATE
  - Attributi
- Selezione dei cammini
- Interazione con IGPs
- Limitazioni BGP e soluzioni
- Architetture BGP e bilanciamento del carico

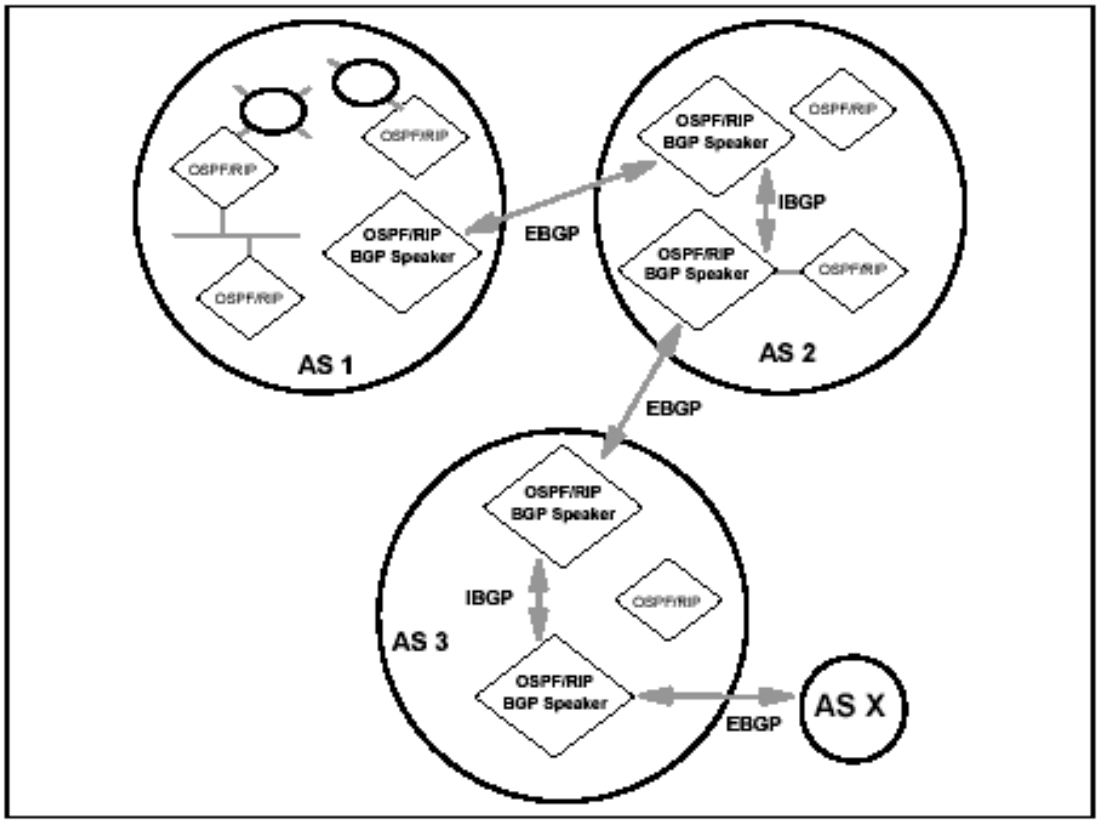
BGP v4

Peering e scambio di messaggi

BGP v4 – Border Gateway Protocol

- Peer: coppia di router BGP che si scambiano informazione di instradamento
  - IBGP peer: stesso AS
  - EBGP peer: AS diversi

Comunicazione tra peer avviene mediante connessioni TCP

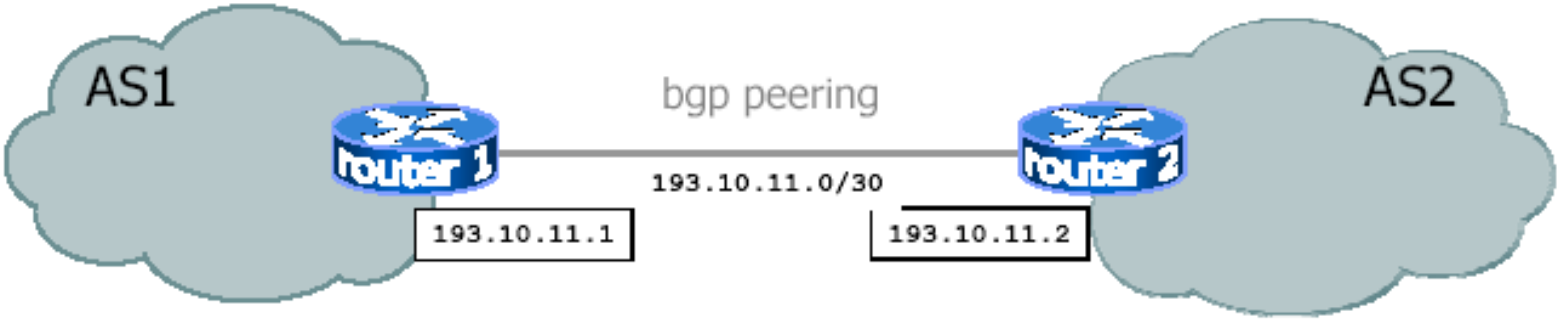


AS numbering

- BGP requires an identifier for every AS (Autonomous System Number, asn) between 1 and 65,535
- An asn may be
  - Global asn– obtained from regional internet authority: ripe, arin, apnic
  - Private asn– obtained from ISP

Peering tra due AS

- Le informazioni possono essere scambiate tra due AS solo se una sessione peering è attiva
- La sessione peering è una connessione TCP tra i due AS



Funzionalità BGP

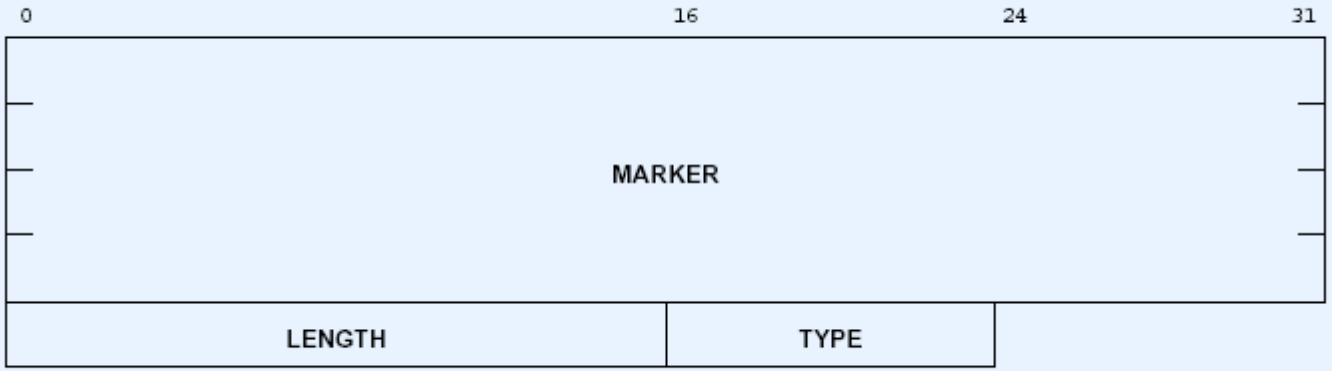
1. Apertura connessione tra peer
2. Annuncio informazioni sulla raggiungibilità
3. Verifica corretto funzionamento

Quattro tipi di messaggio BGP

Type Code	Message Type	Description
1	OPEN	Initialize communication
2	UPDATE	Advertise or withdraw routes
3	NOTIFICATION	Response to an incorrect message
4	KEEPALIVE	Actively test peer connectivity

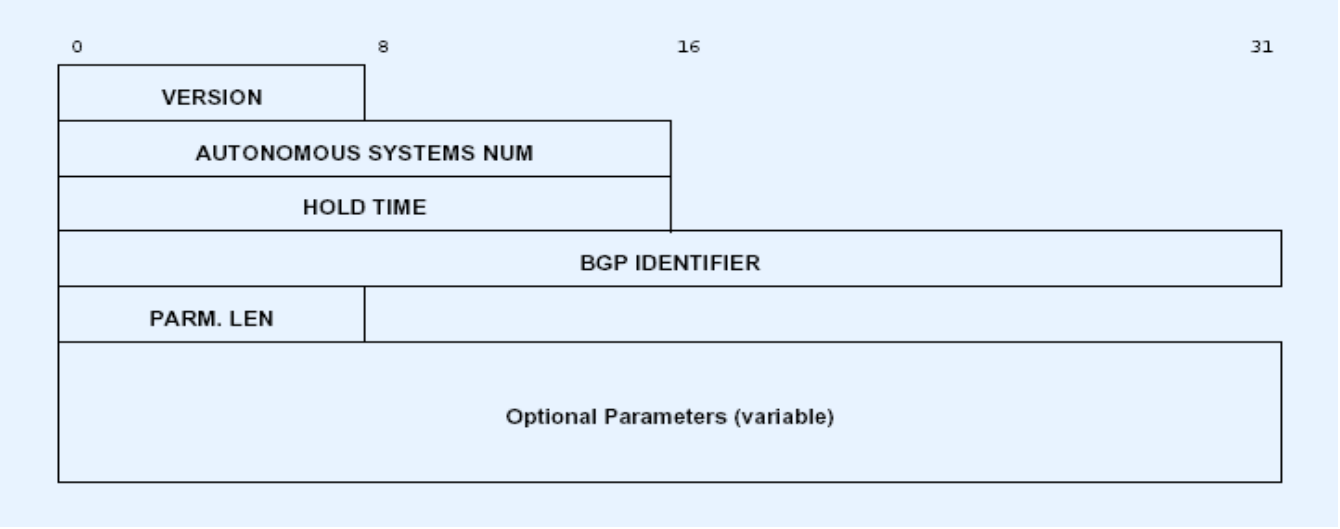
Intestazione messaggi BGP

- Precede ogni messaggio BGP ed identifica il tipo di messaggio
- Marker (16 byte): autenticazione e sincronizzazione tra i peer
- Length (2 byte): lunghezza del messaggio tra 19 e 4096 byte
- Type: tipo di messaggio BGP



Peering/apertura connessione

- OPEN: usato per aprire una connessione peer
- Il campo Hold specifica il massimo numero di secondi tra due messaggi successivi
- Un router bgp è caratterizzato dall'asn e da un indentificatore unico a 32 bit che deve usare per tutte le connessioni peering
- Parametri opzionali: ad esempio per l'autenticazione



Messaggi/OPEN

- Il router destinatario di un messaggio OPEN risponde con un KEEPALIVE
- Connessione aperta quando entrambi i router hanno inviato un messaggio OPEN e ricevuto un messaggio KEEPALIVE

Messaggi/KEEPALIVE

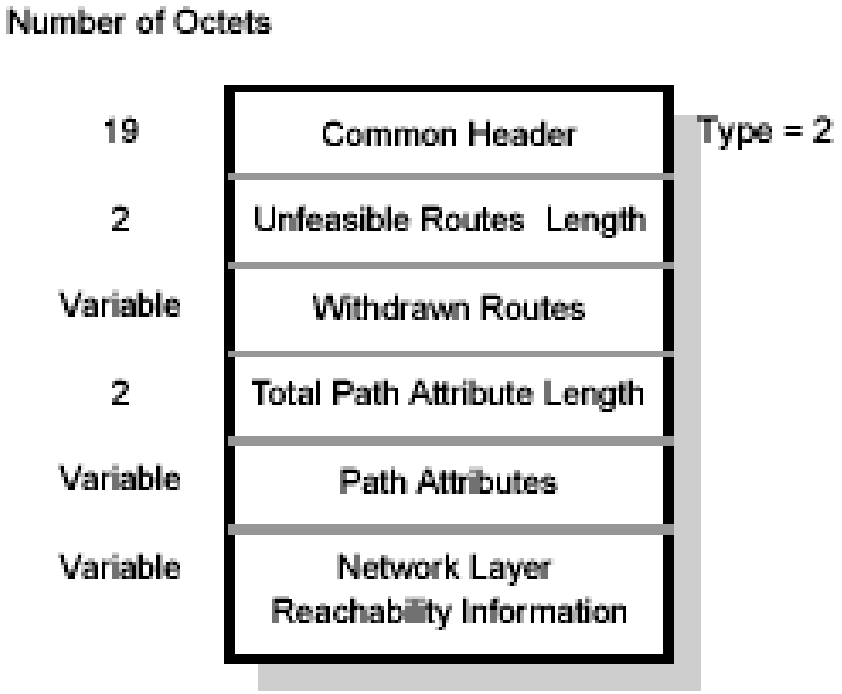
- Verifica periodicamente la connessione TCP tra entità peer
- Più efficiente rispetto ad inviare periodicamente messaggi di instradamento
- Intervallo KEEPALIVE ogni 1/3 di HOLD time, mai inferiore a 1 sec.

Messaggi/NOTIFICATION

- Controllo o segnalazione errori
- BGP invia un messaggio di notifica e chiude la connessione TCP
- Errori:
  1. Errore nell'intestazione del messaggio
  2. Errore nel messaggio OPEN
  3. Errore nel messaggio UPDATE
  4. Timer di attesa scaduto
  5. Errore nella macchina a stati finiti
  6. Fine (connessione terminata)

Messaggi/UPDATE

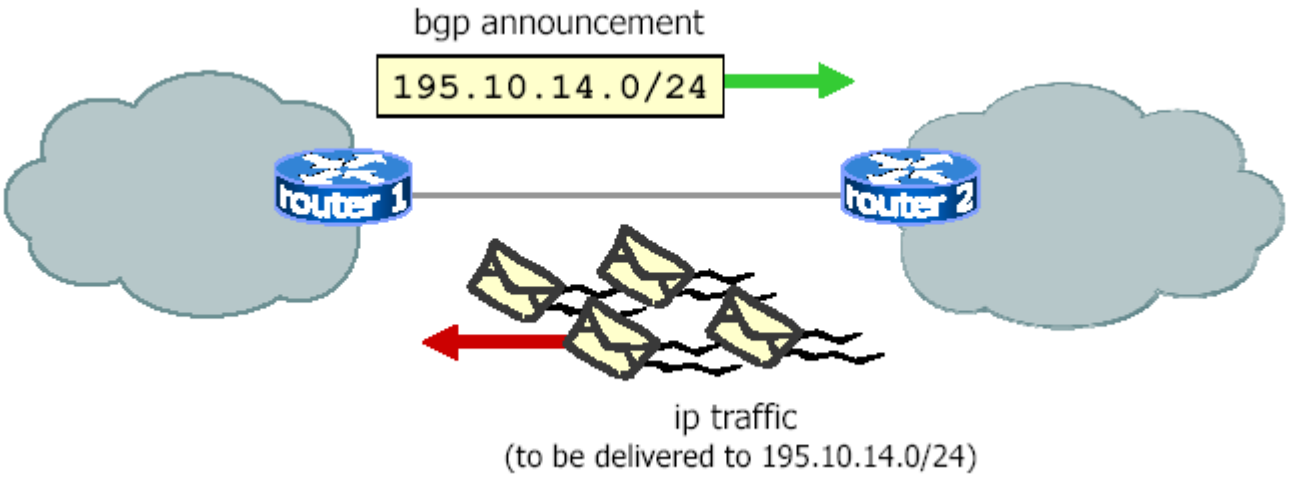
- Announcement = prefix + attributes values
- Annuncia nuove reti raggiungibili ed eventualmente l'instradamento
- Annuncia reti precedentemente annunciate non più raggiungibili



BGP updates

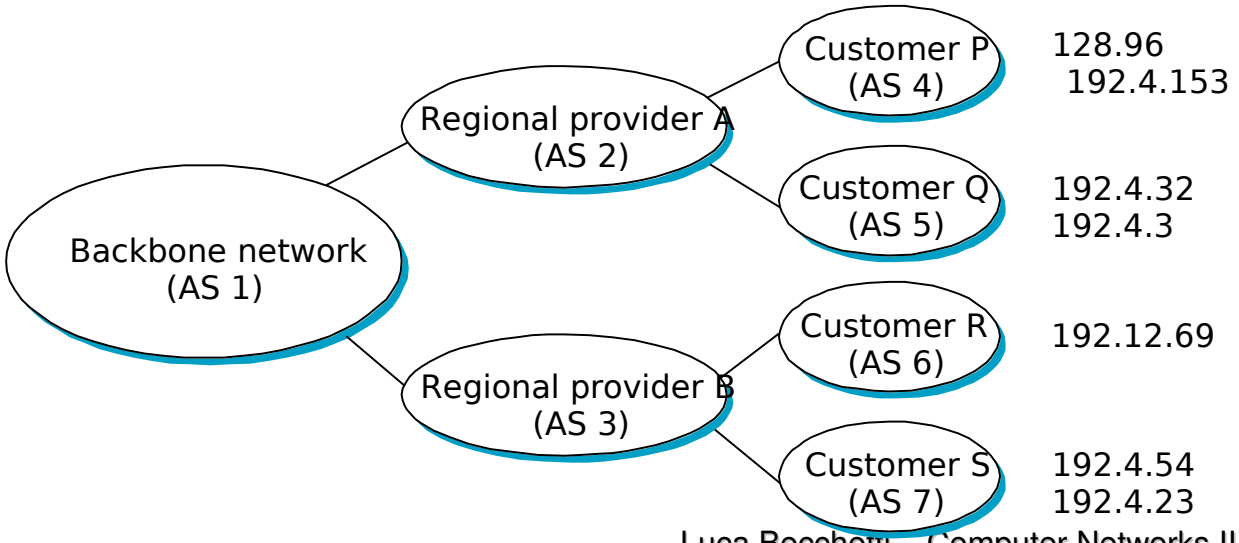
Annunci BGP

- BGP permette ad un AS di offrire connettività ad un altro AS
- Offrire connettività significa promettere il recapito ad una specifica destinazione
- Destinazione specificata da (Netmask, Prefix)
  - Si adotta convenzione CIDR
- **Annunci BGP in messaggi UPDATE**



Route BGP/Path vector

- Speaker for AS2 advertises reachability to P and Q
  - network 128.96, 192.4.153, 192.4.32, and 192.4.3, can be reached directly from AS2
- Speaker for backbone advertises
  - networks 128.96, 192.4.153, 192.4.32, and 192.4.3 can be reached along the path (AS1, AS2).
- Speaker can cancel previously advertised paths

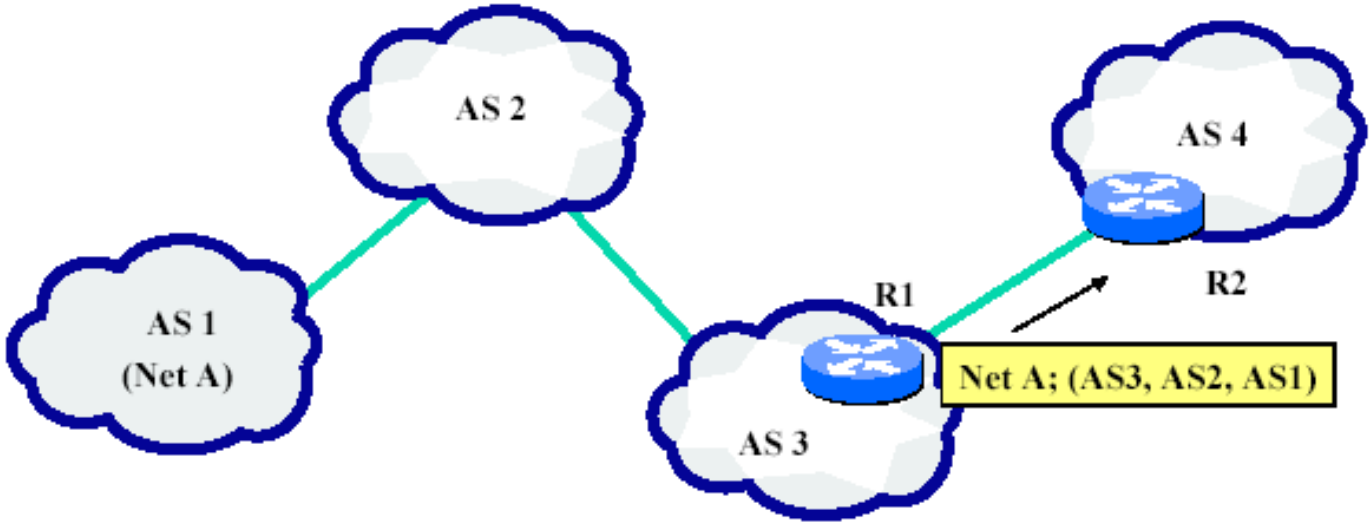


Path vector/cont.

- Structure of information in the updates:

  DestNet:(<lista di AS>)

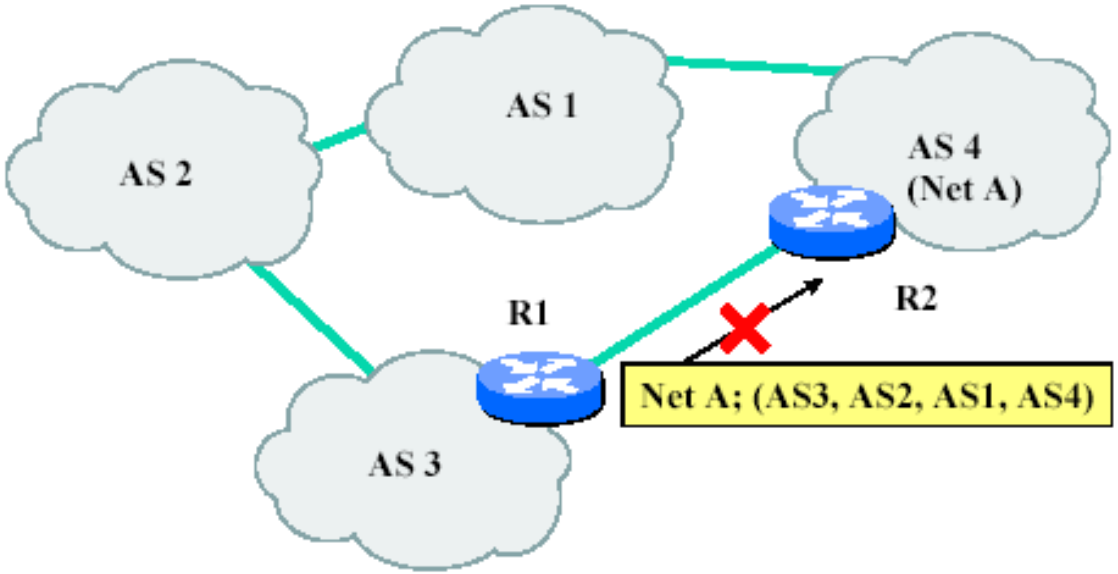
- No shortest paths in general



R1 says to R2: to reach subnet A go over AS3, AS2 and AS1 in this sequence

Managing loops

- Paths contain complete paths
- Cycles can be detected



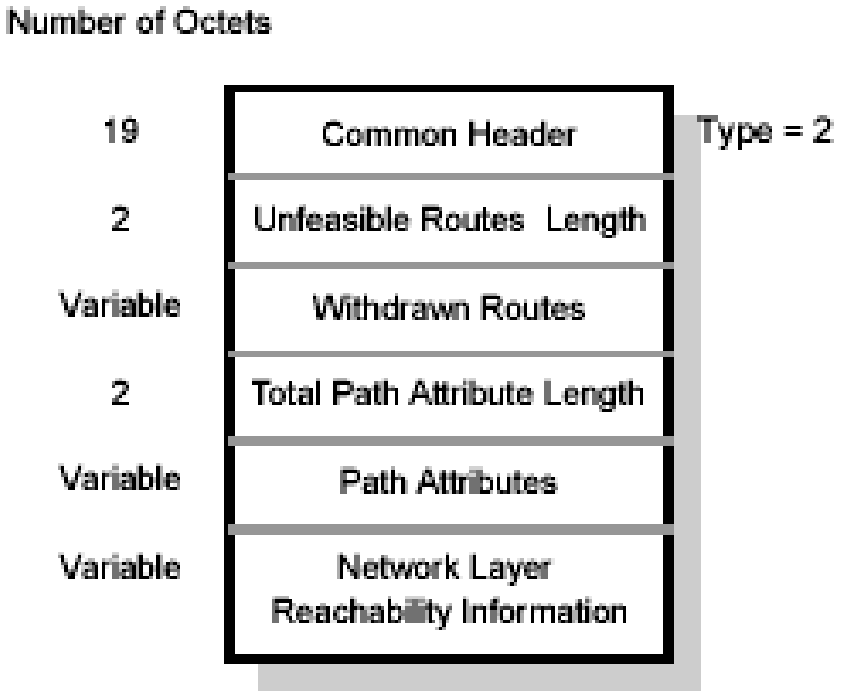
R2 always refuses Net A is associated AS path includes AS to which R2 belongs

Filtro degli annunci

- Gli annunci sono inviati e/o accettati solo se alcune condizioni sono verificate
- Gli annunci possono essere filtrati sulla base di:
  - Una lista di prefissi validi
  - Una lista di numeri di AS

Update with withdrawn routes

- May also contain list of networks that are no longer reachable over *this* path
- Withdrawn routes fields not present if no network to "withdraw"



Update/cont.

- Withdrawn routes: list of pairs <length, IP prefix> for unreachable destinations
  - Length: network prefix length in bits
- Network Layer Reachability Information (NLRI): list of pairs <length, IP prefix> for announced destination networks
  - Length: network prefix length in bits
- NLRI example:

  /25, 204.149.16.128

  /23, 206.134.32

  /8, 11

- Value of AS_PATH: sequence of AS identifiers --> sequence of ASes to traverse in order to reach networks announced in NLRI

Attributes

- Variable field in UPDATE message
- Attribute values common to *all announced destinations*
- Destinations with different attributes *must be announced with different messages*
- Attribute categories
  - **Well-known**: must be supported in every BGP implementation - if also **mandatory:** must be forwarded, possibly after modification
  - **Optional**: need not be implmented, may be forwarded or not

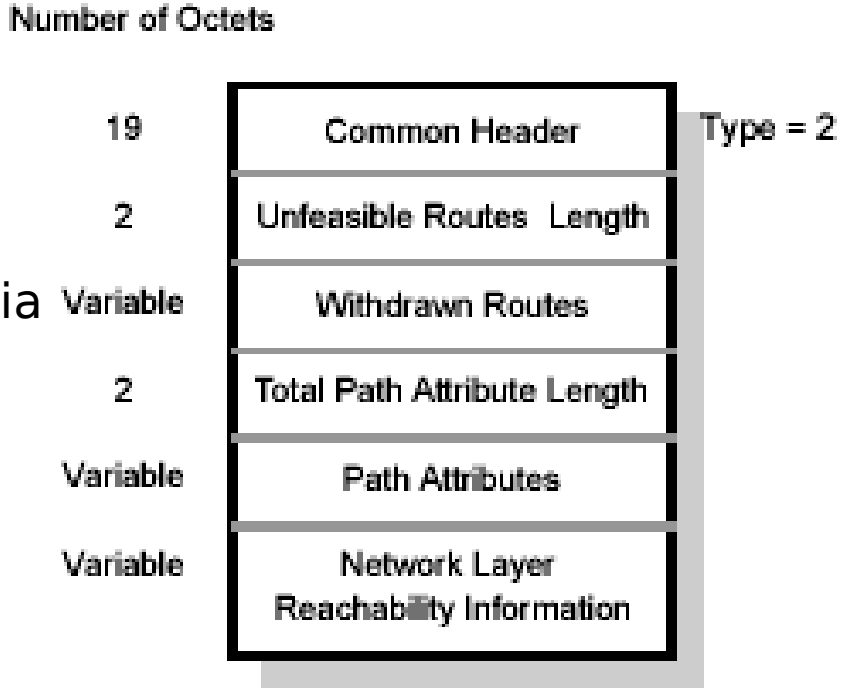
Path attributes - every route

- **AS_PATH**
  - AS list - *well-known*
- **ORIGIN**
  - Origin of routing info - *well-known*
- **NEXT_HOP**
  - IP address of next BGP speaker on route to destination - *well-known*
- **MED**: discriminate between different exits from AS
  - MED: MULTI_EXIT_DISCRIMINATOR - *optional*
  - Useful in path selection (see further)
- **LOCAL_PREF**: preference within AS
  - Useful in path selection (see further) - *well-known*
- Aggregator
  - ID of AS that aggregated routes

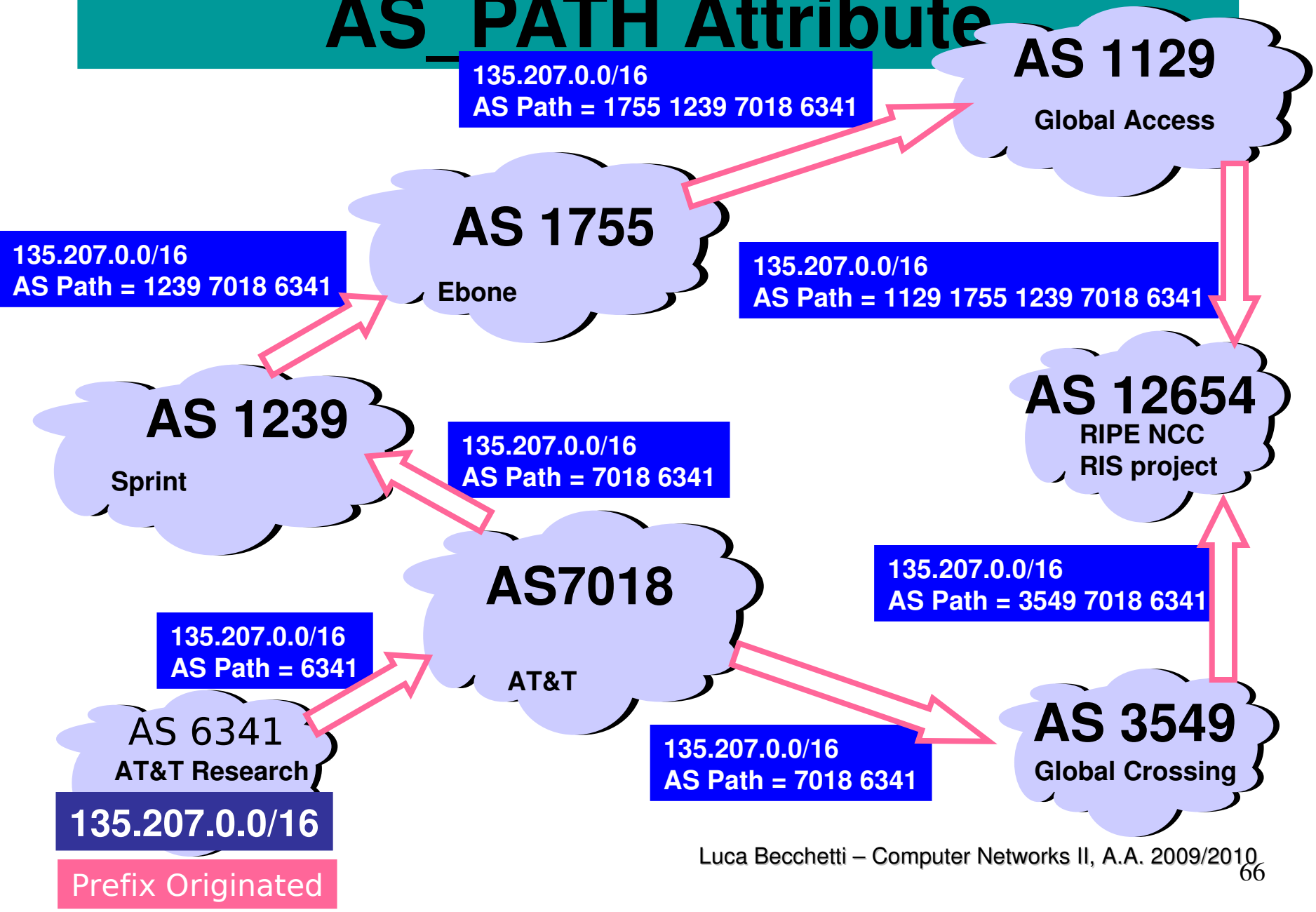
ORIGIN

- Definisce l'origine dell'informazione annunciata
- Puo' essere IGP, EGP o INCOMPLETE

INCOMPLETE: si ha nel caso in cui le reti annunciate siano state inserite come route statiche nello speaker che invia l'annuncio

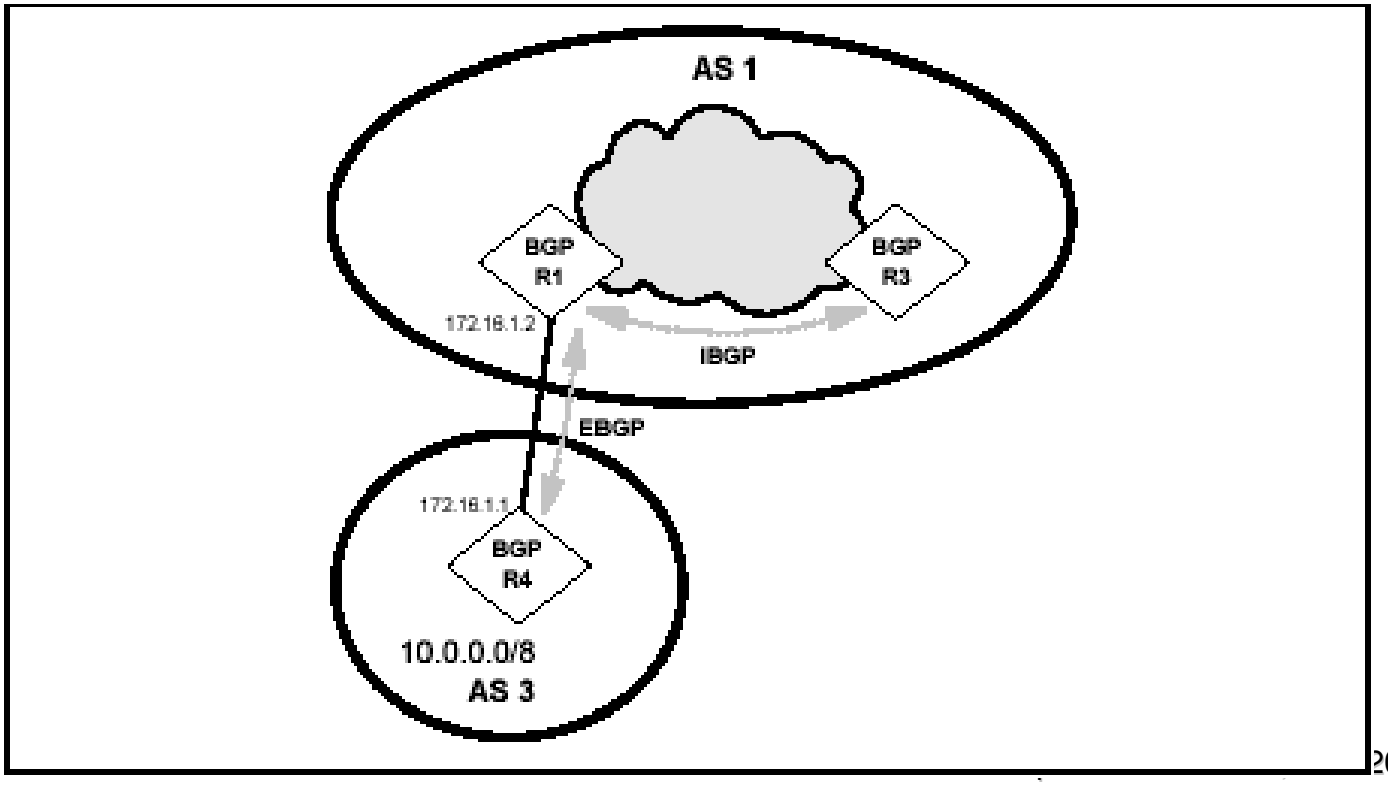


AS_PATH Attribute

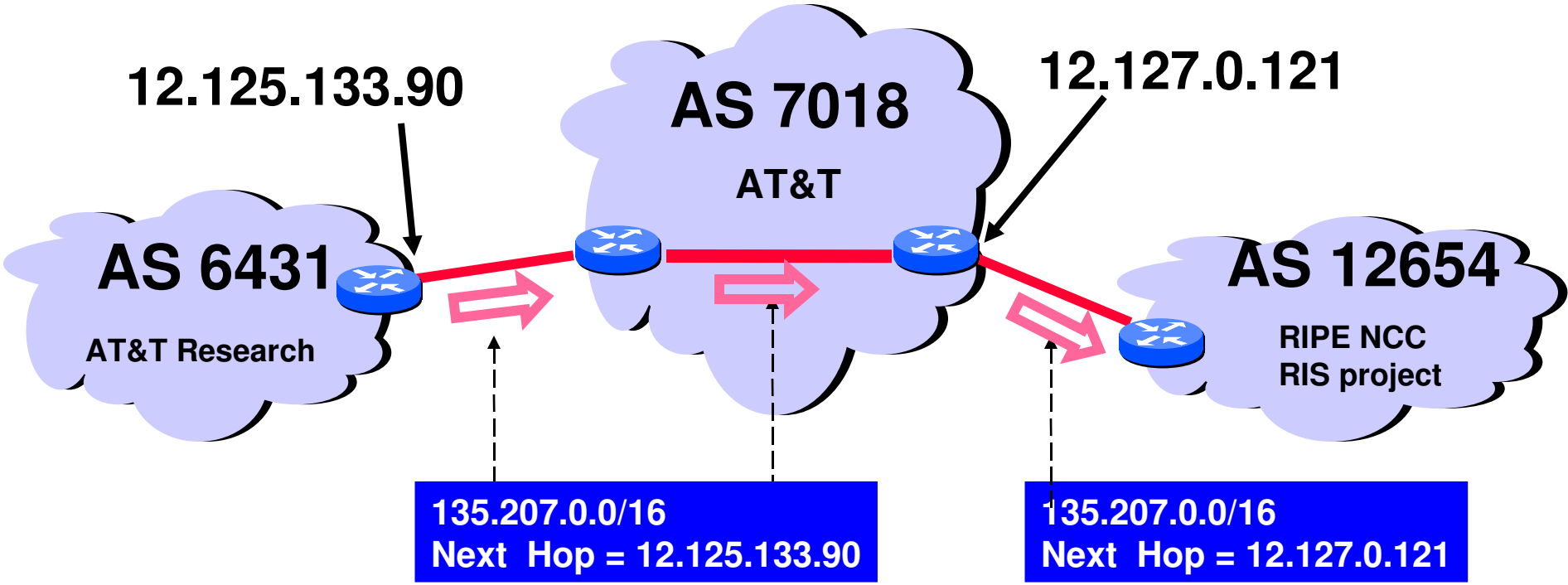


NEXT_HOP

- IP address of BGP speaker of next-hop AS on path destination
- For network 10.0.0.0/8: R1 sends 172.16.1.2 as next hop to R4



BGP Next Hop Attribute

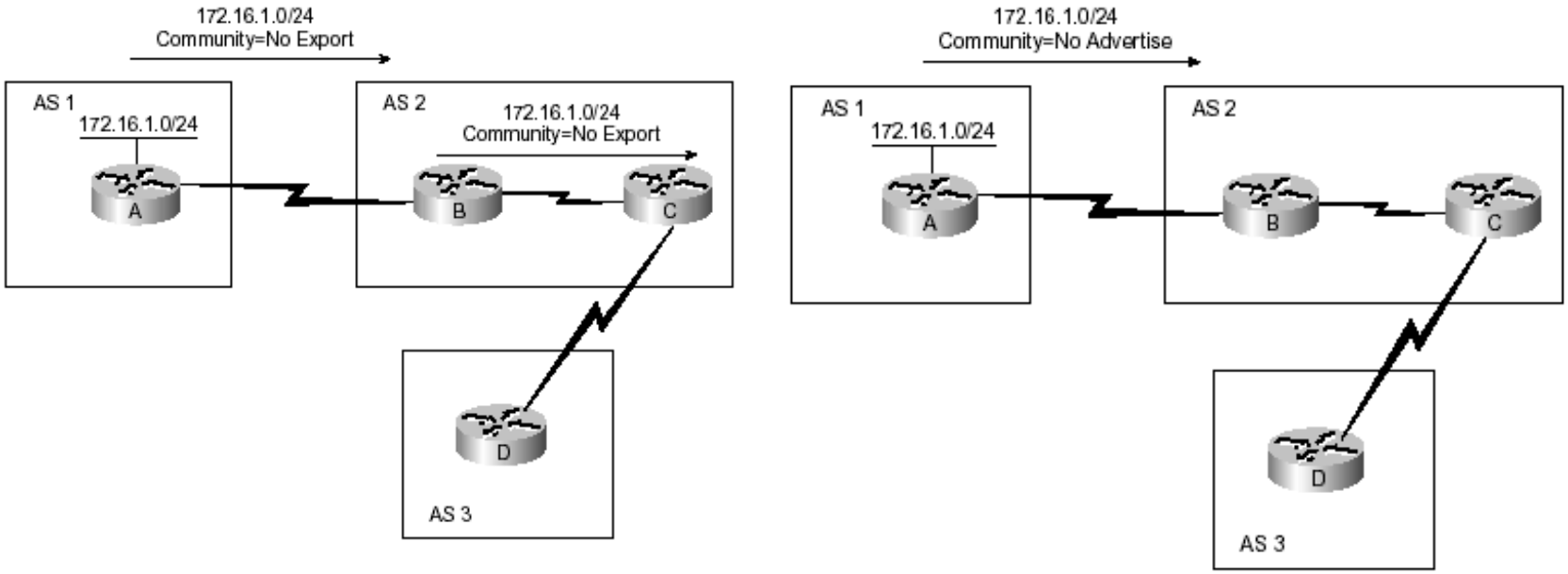


Every time a route announcement crosses an AS boundary, the Next Hop attribute is changed to the IP address of the border router that announced the route.

COMMUNITY attribute

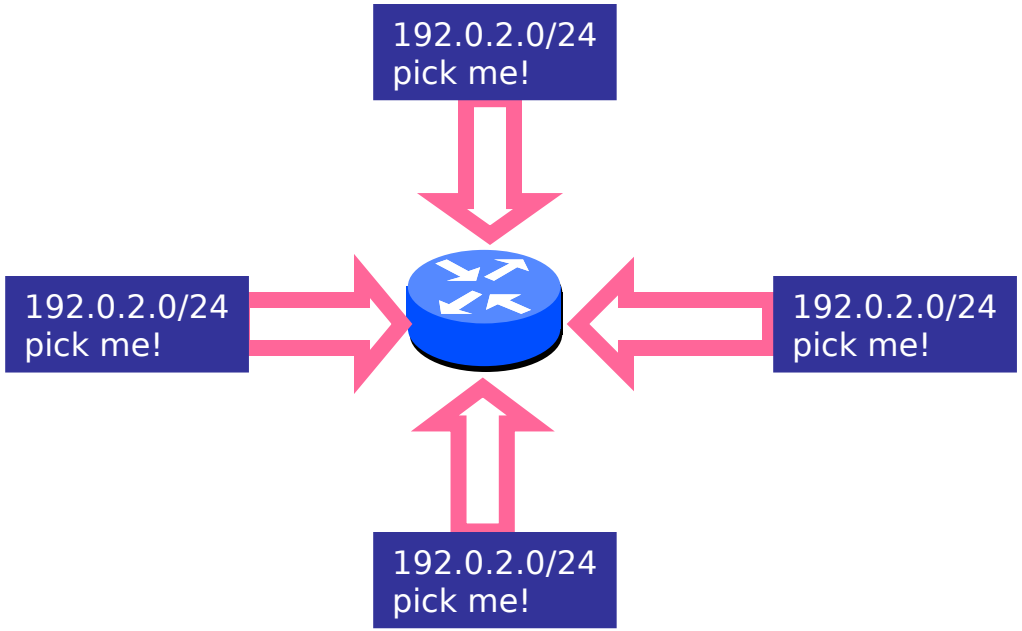
- Possible to define groups of destinations to which same forward policy should be applied
- Policy defined by value of COMMUNITY attribute
- Optional attribute
- Pre-defined values for COMMUNITY
  - **No-export**: don't announce route to EBGP peers
  - **No-advertise**: don't announce to any peer
  - **Internet**: announce route to every peer

COMMUNITY


172.16.1.0/24
Community=No Export
AS 1
172.16.1.0/24
A
AS 2
172.16.1.0/24
Community=No Export
B
C
D
AS 3
172.16.1.0/24
Community=No Advertise
AS 1
172.16.1.0/24
A
AS 2
B
C
D
AS 3

Path selection

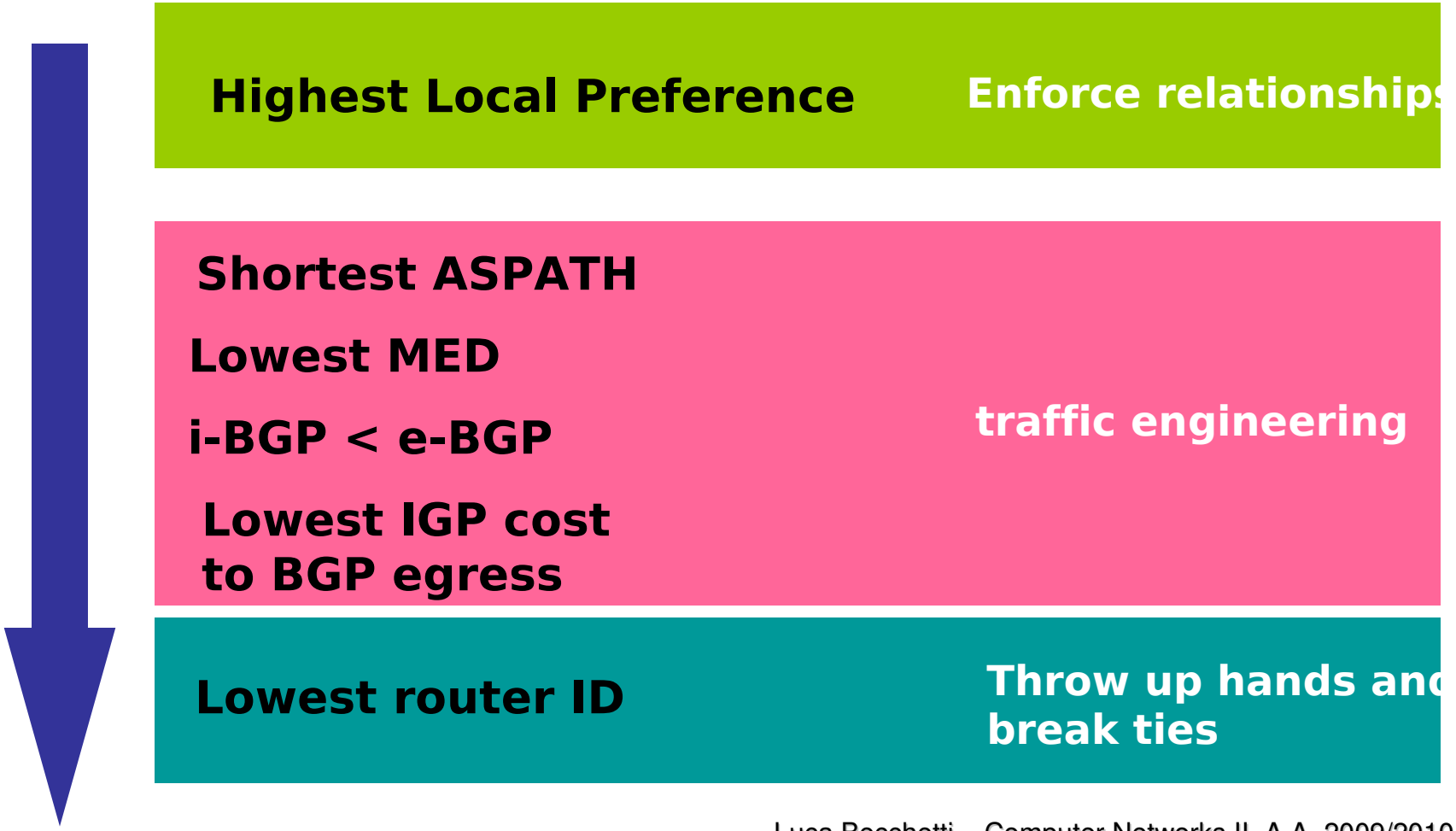
Attributes are Used to Select Best Routes



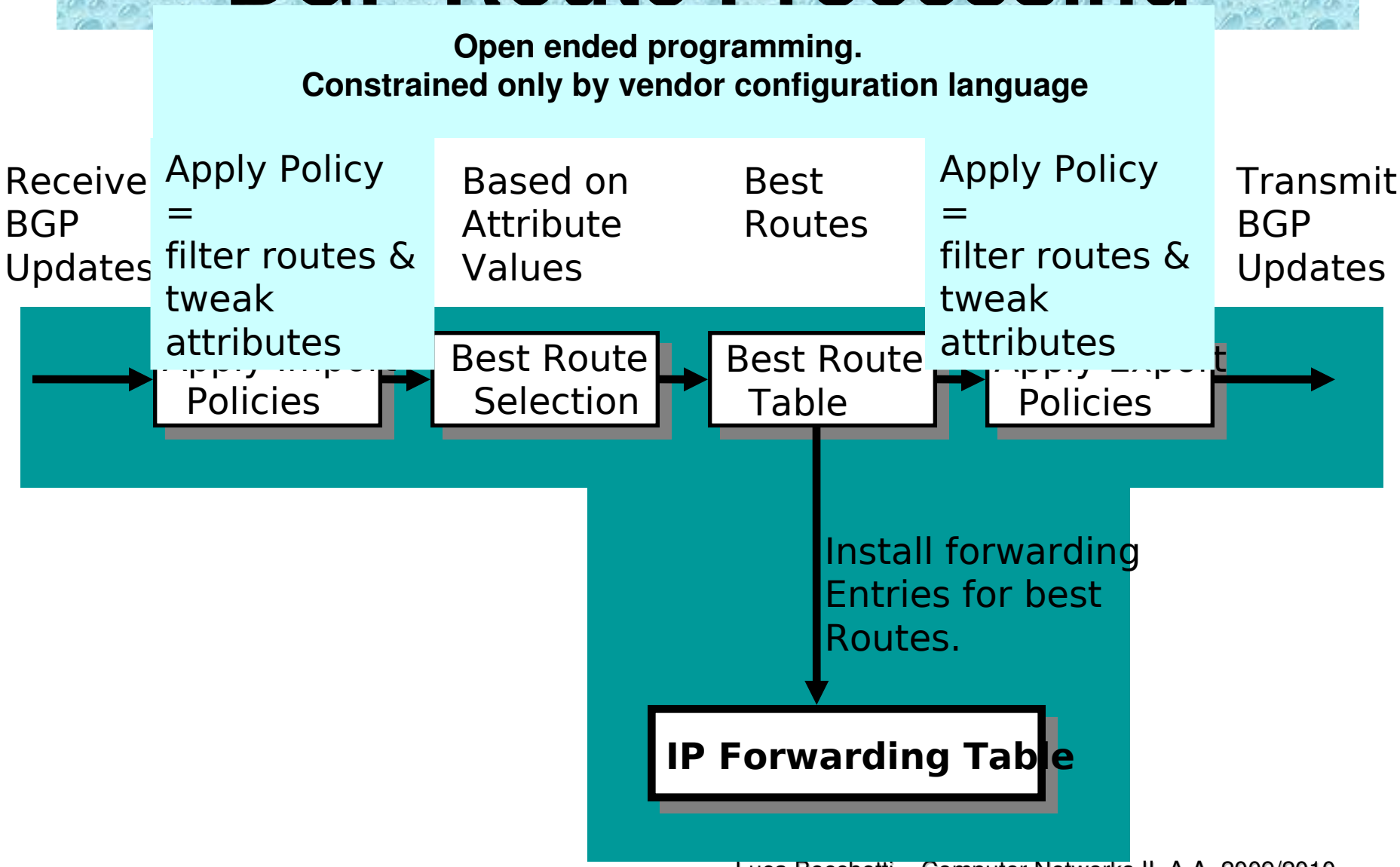
Given multiple routes to the same prefix, a BGP speaker must pick at most one best route

(Note: it could reject them all!)

Route Selection Summary

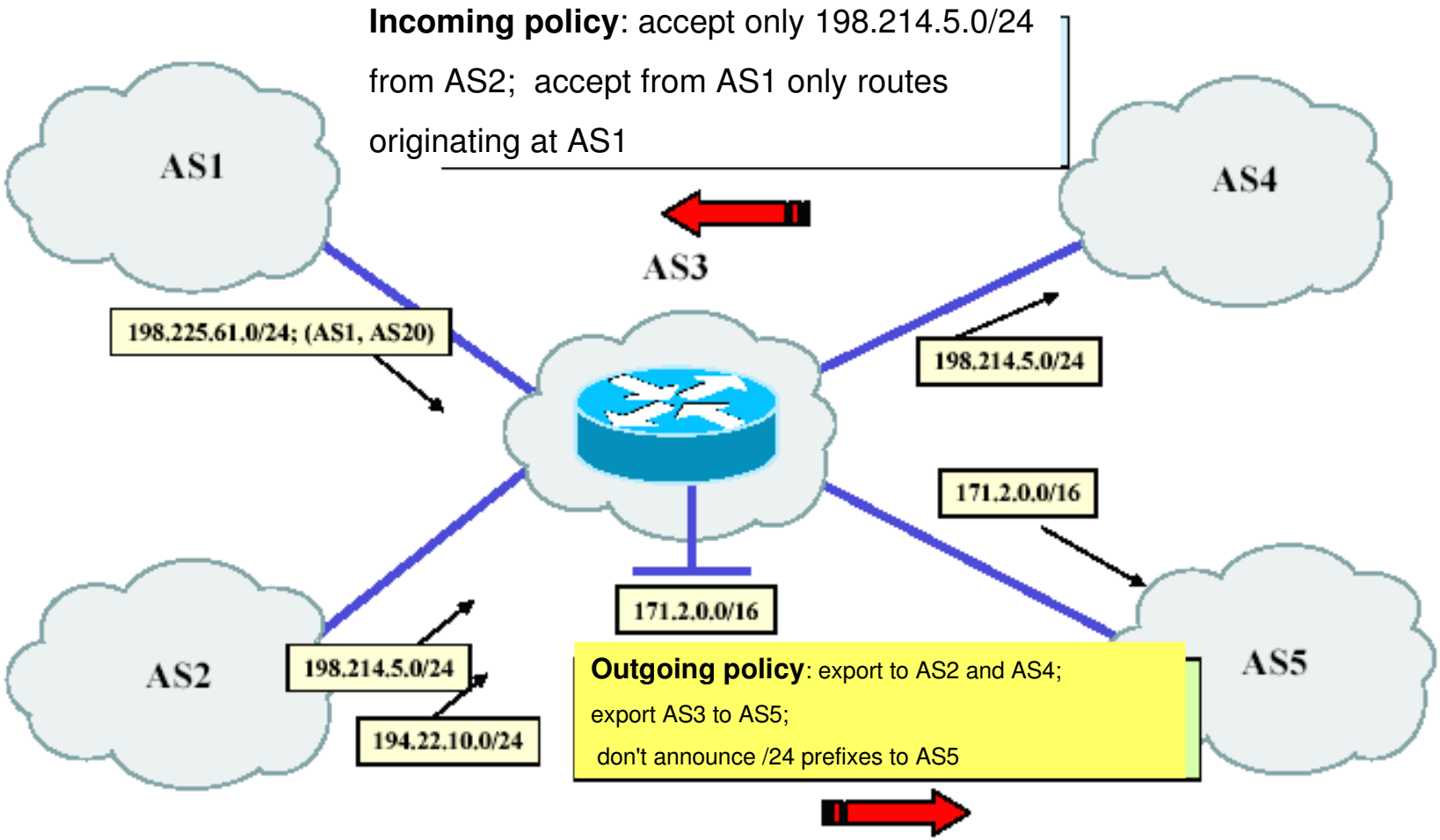


BGP Route Processing



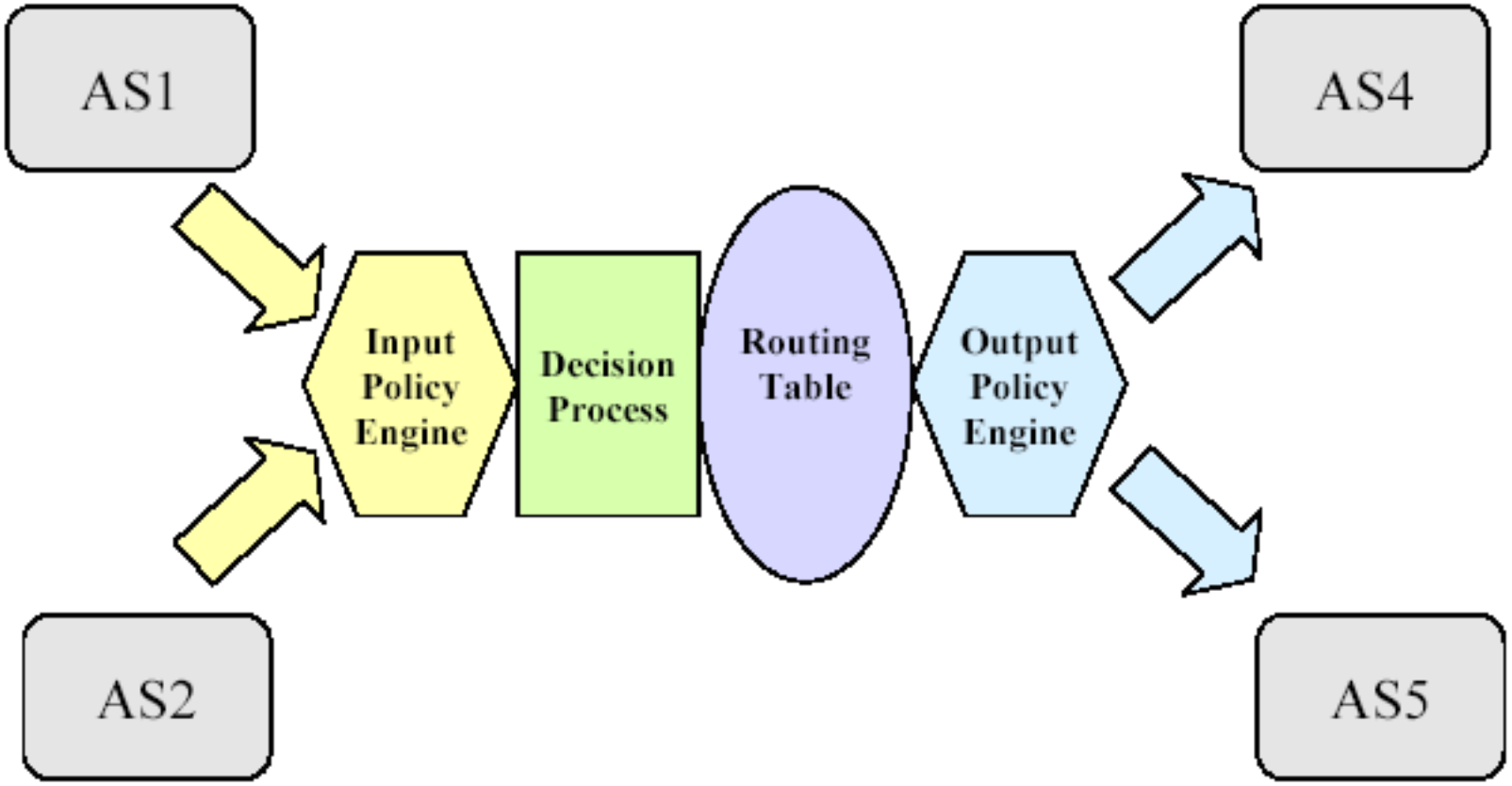
BGP - routing policies

- Administrator decides incoming/outgoing traffic policies



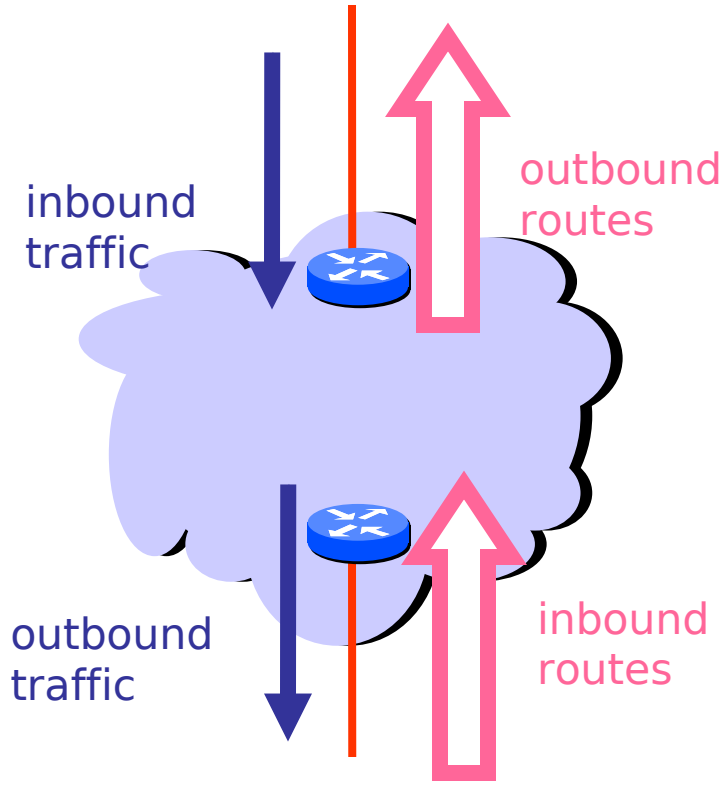
BGP - routing policies

- Architectural scheme



Tweak Tweak Tweak

- For inbound traffic
  - Filter outbound routes
  - Tweak attributes on outbound routes in the hope of influencing your neighbor's best route selection
- For outbound traffic
  - Filter inbound routes
  - Tweak attributes on inbound routes to influence best route selection

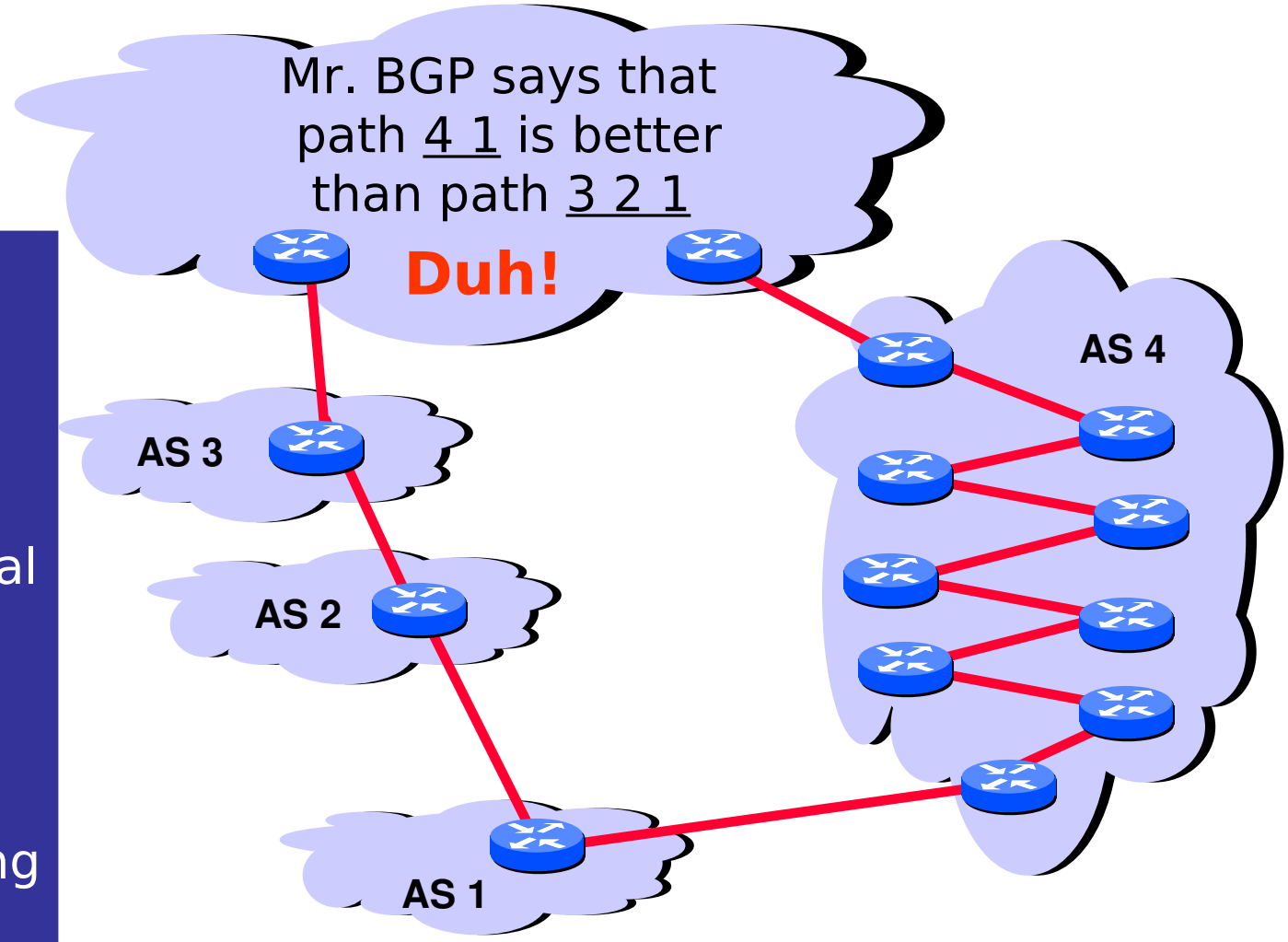


In general, an AS has more control over outbound traffic

Shorter Doesn't Always Mean Shorter

In fairness: could you do this "right" and still scale?

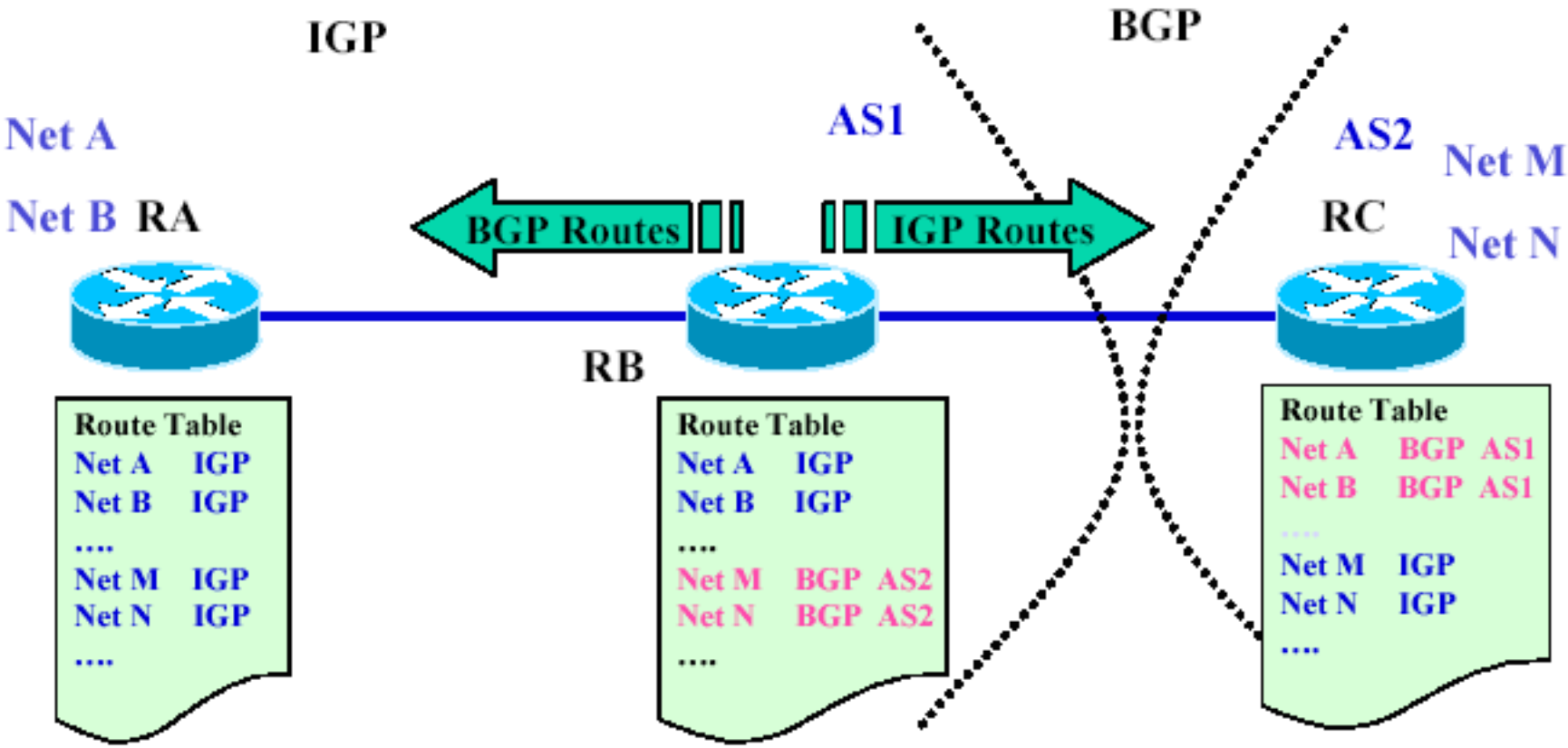
Exporting internal state would dramatically increase global instability and amount of routing state



Interaction with IGP

Interaction with IGP

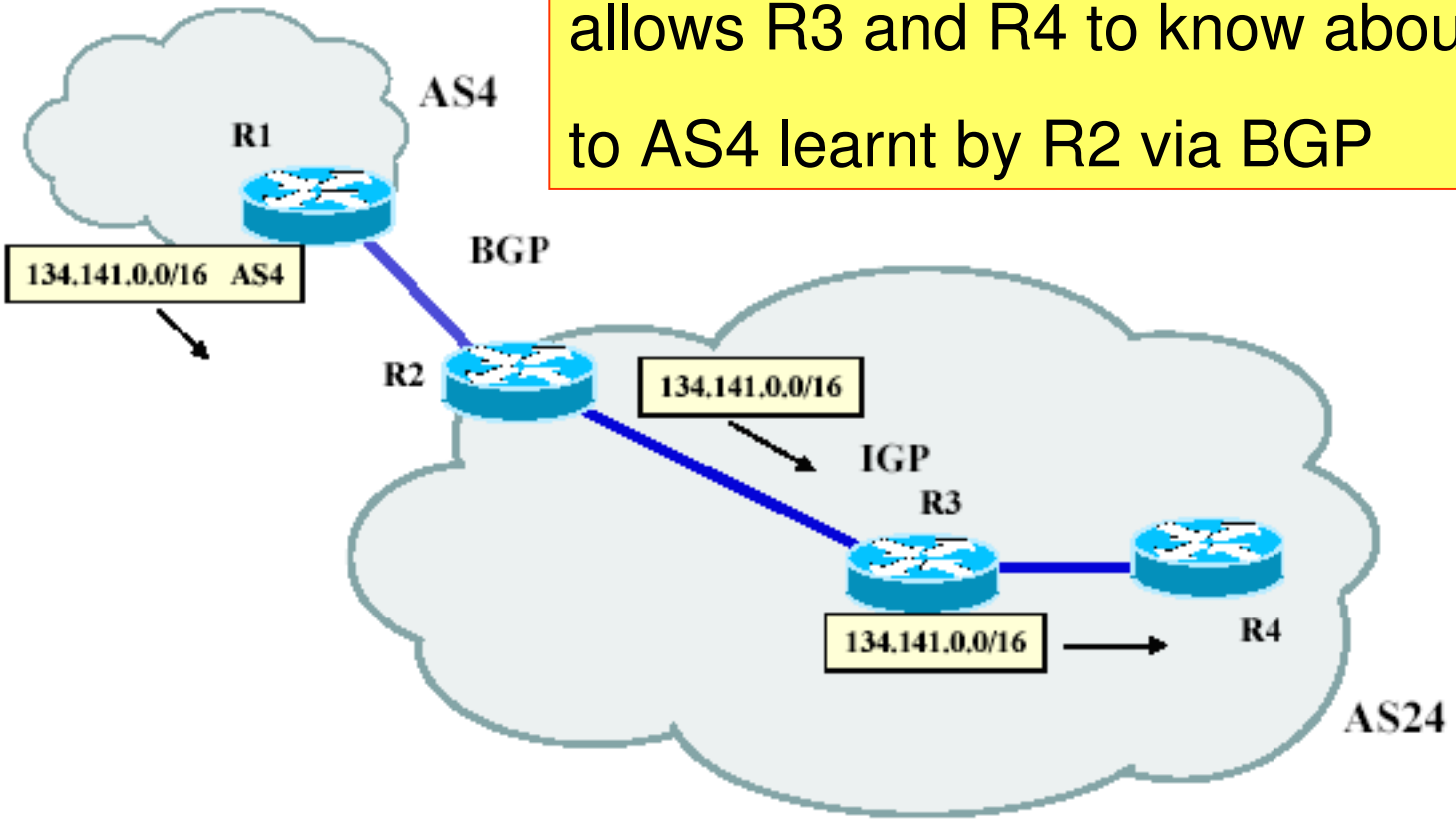
- Border router implements BGP *and* IGP



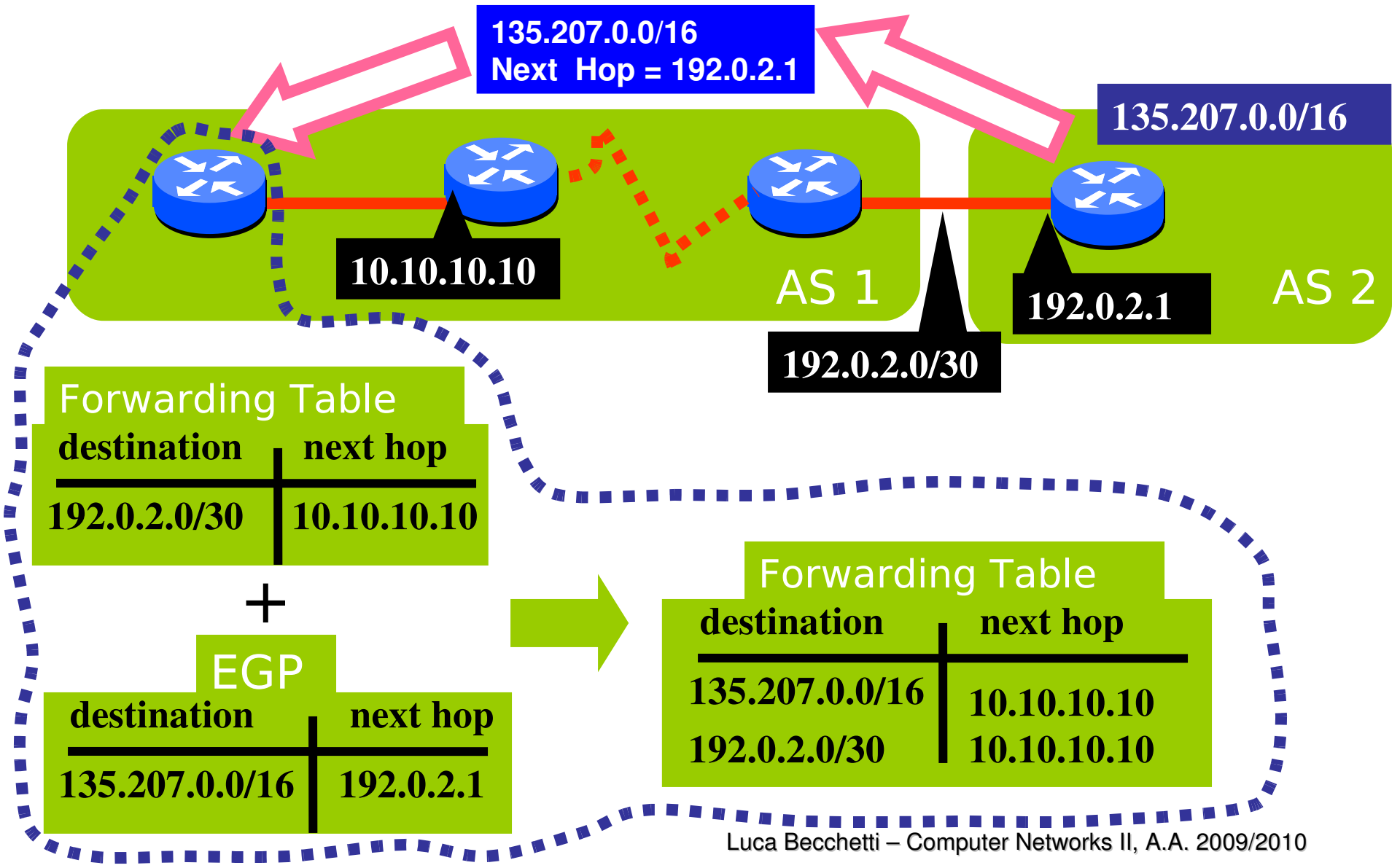
Interaction with IGP

- Border router implements BGP *and* IGP

BGP --> IGP redistribution mechanism allows R3 and R4 to know about routes to AS4 learnt by R2 via BGP



Join EGP with IGP For Connectivity



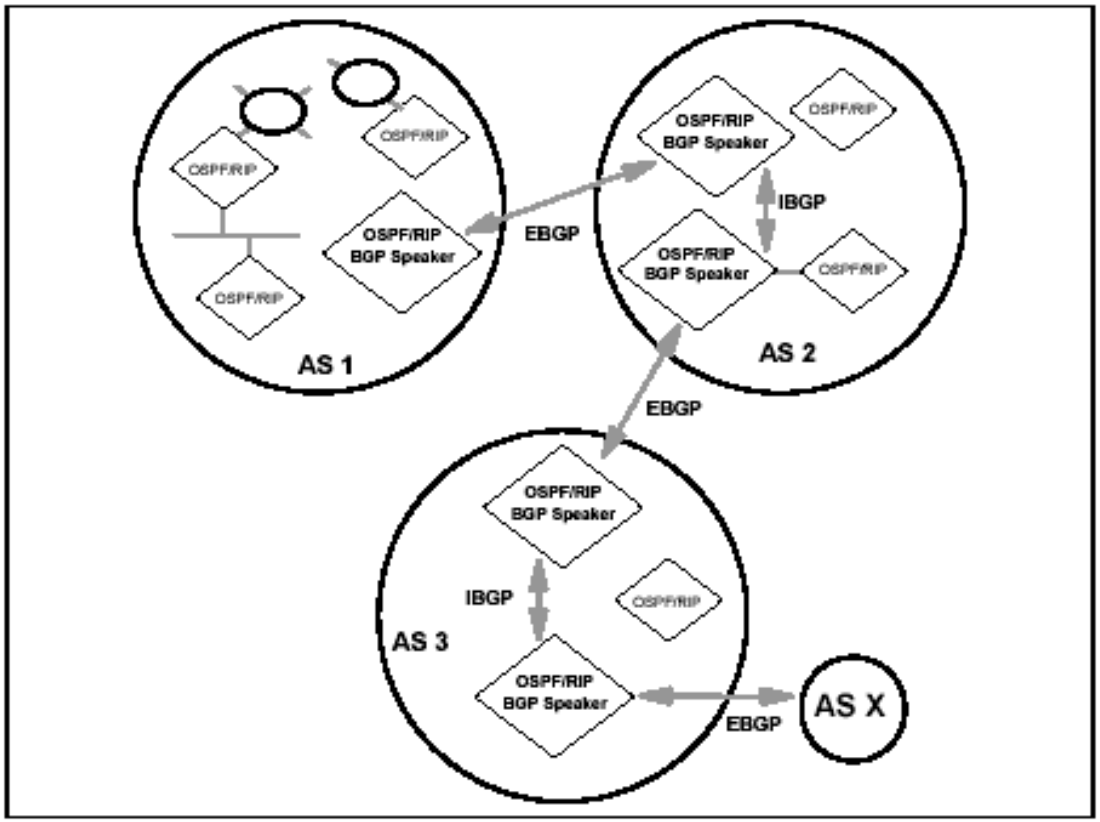
BGP limits and solutions

BGP limits

- BGP cannot choose between two paths based on cost or congestion
- BGP allows to distribute traffic among more links but not dynamically
- Necessary to manually configure which networks are announced by which border routers
- All autonomous sytems have to agree on consistent scheme to announce reachability

BGP limits/2

- If AS2 does not forward AS3's updates to AS1 --> AS3 and ASX unreachable from AS1



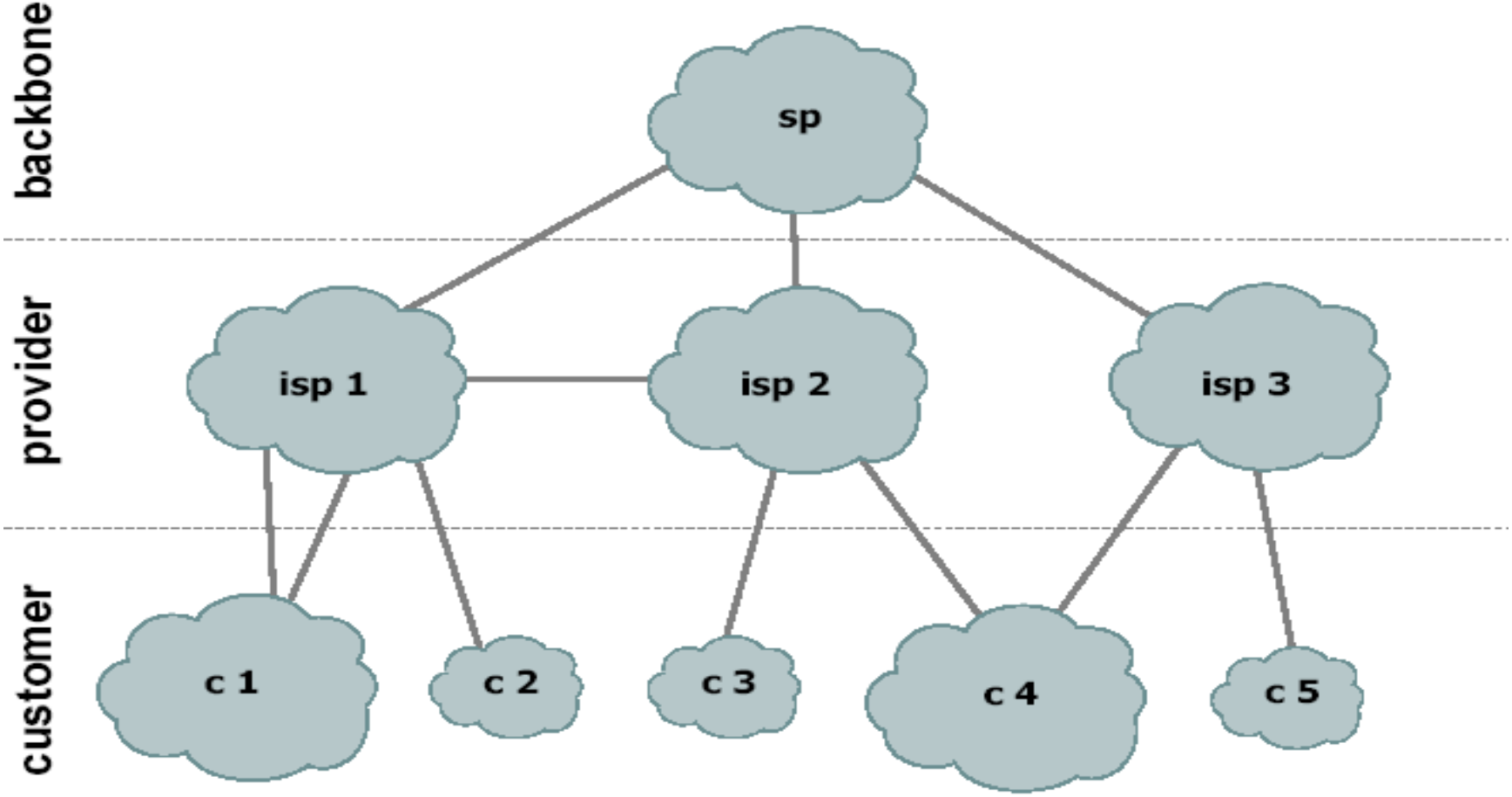
Instradamento con arbitraggio

- Occorre un sistema per garantire la coerenza sulle informazioni di instradamento
- Database autenticato e replicato che contiene le informazioni sulla raggiungibilità
- Autenticazione: solo AS autenticati possono annunciare la raggiungibilità di una rete
- NAP sono i router di interconnessione tra ISP
- I NAP hanno un Router Server che mantiene il data base BGP ma non sono necessariamente speaker BGP
- Gli speaker BGP mantengono aperto un collegamento verso il Router Server

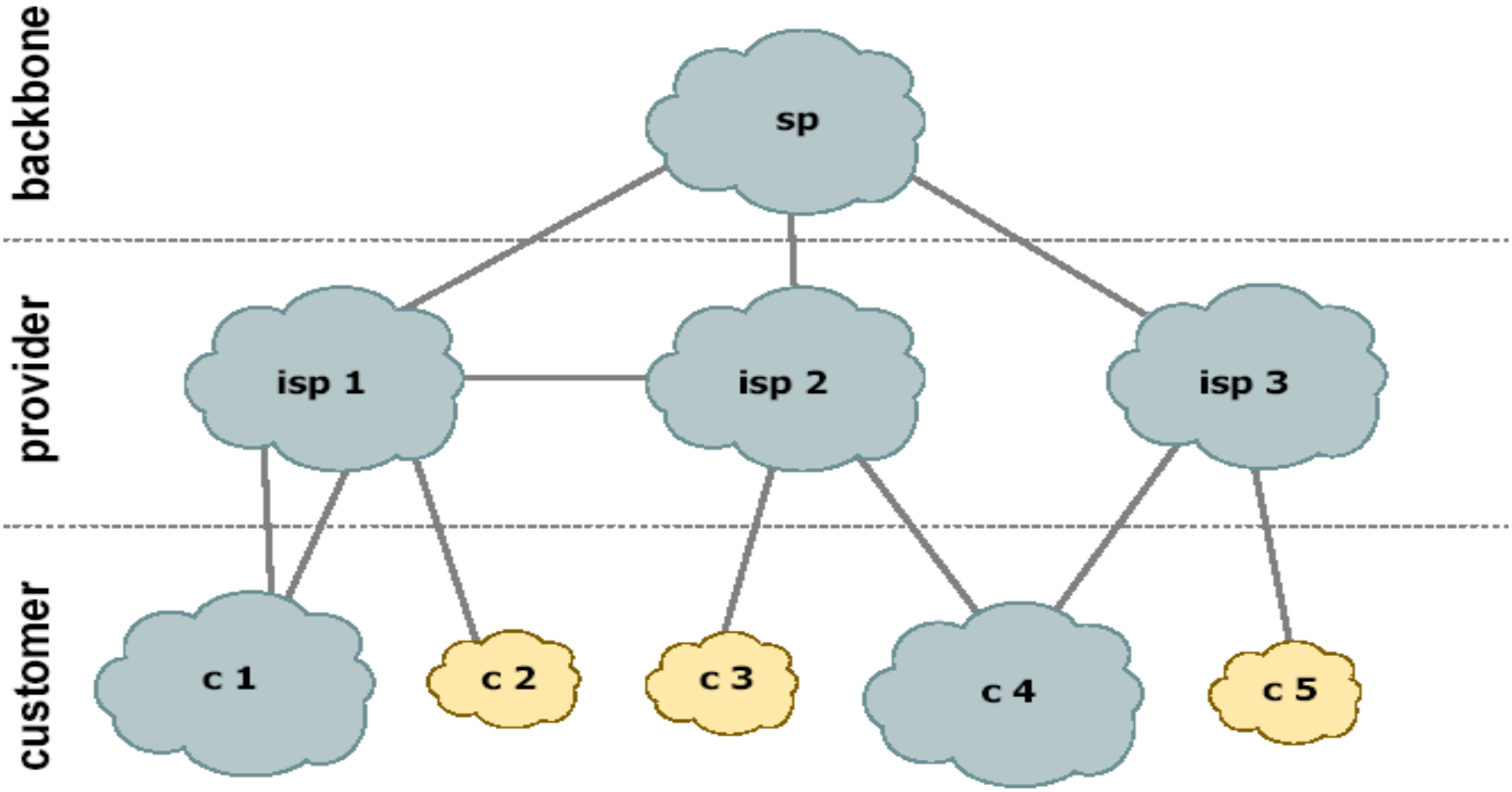
Examples of BGP architectures

AS STUB and Multi-Homed

Complex BGP scenario



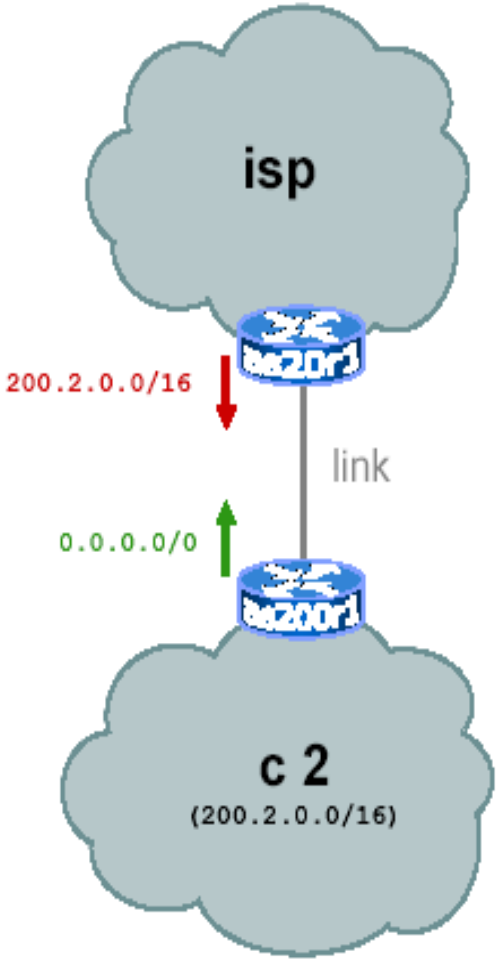
Stub network



Stub network, architecture

- A router in the network is default gateway and is connected to *a single* ISP's router
- Single BGP peering sessions over which as200 provides its reachability information and accepts routing over default router

Static routing for a stub network

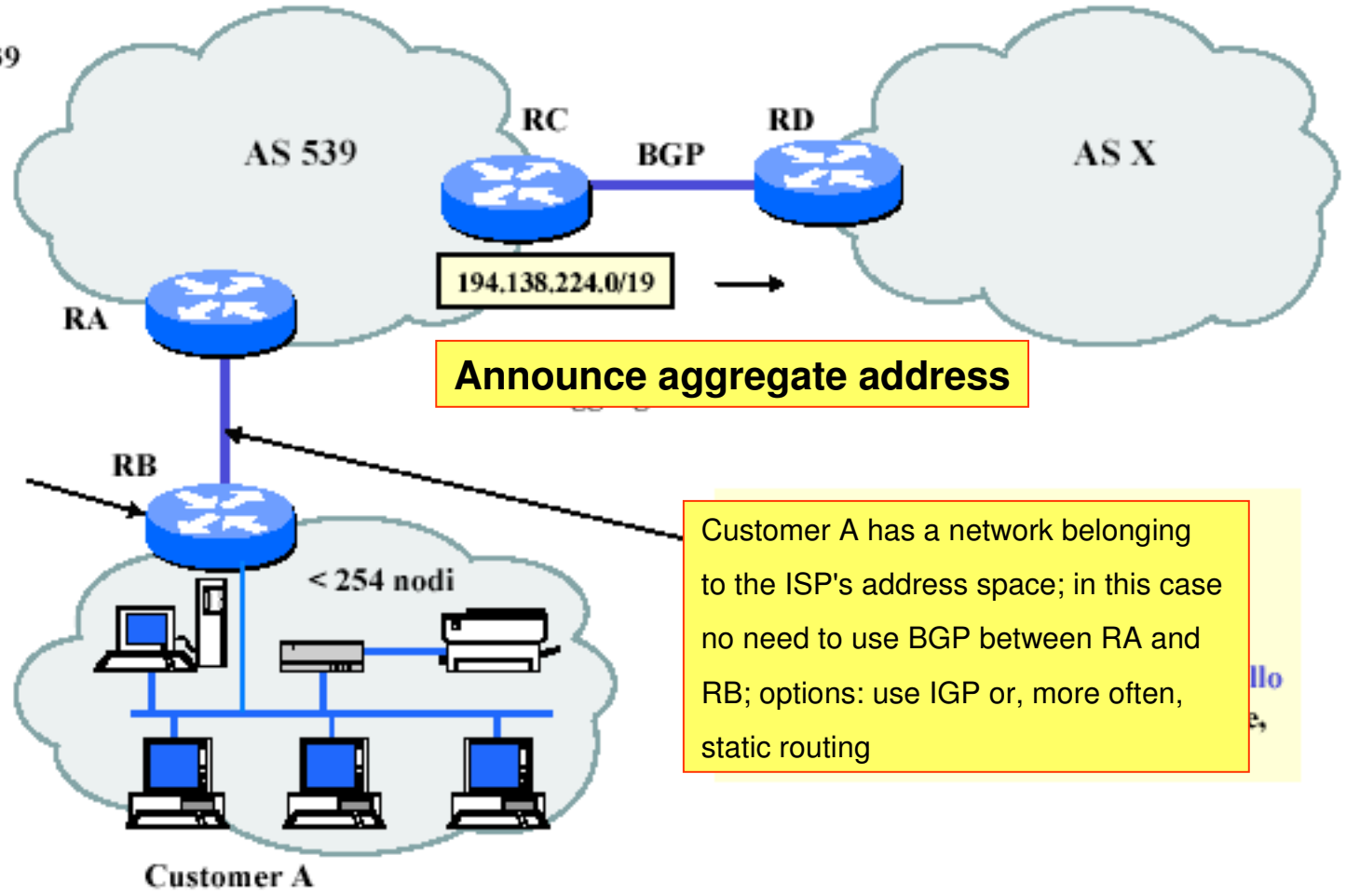


- A default static route enough for IP packets to be sent to the Internet over ISP connection
- Also enougn for incoming packets to reach the stub network after traversing the ISP's network
- No need for BGP

Example

Address block assigned to AS539

194.138.224.0/19



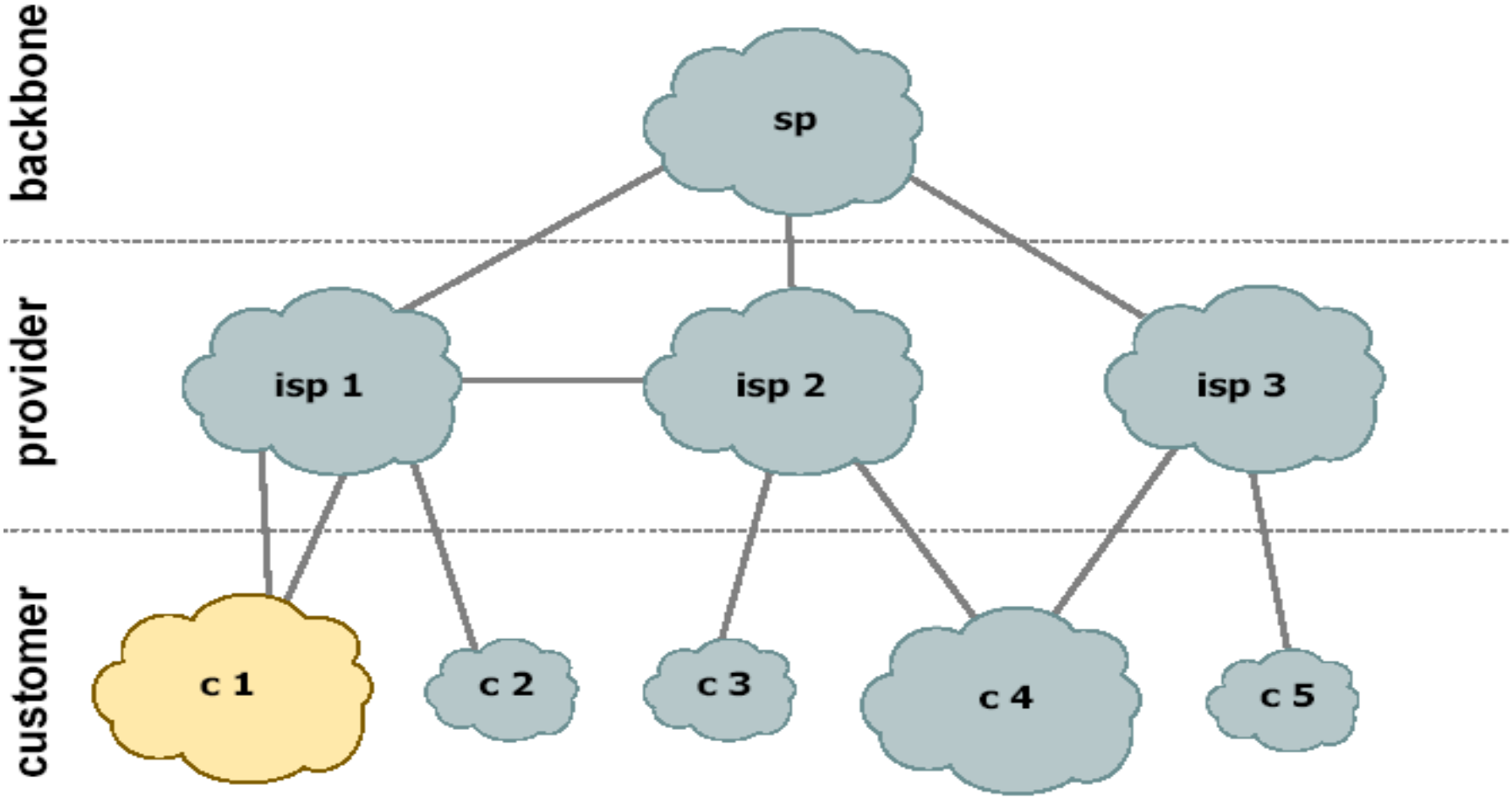
Announce aggregate address

Customer A has a network belonging to the ISP's address space; in this case no need to use BGP between RA and RB; options: use IGP or, more often, static routing

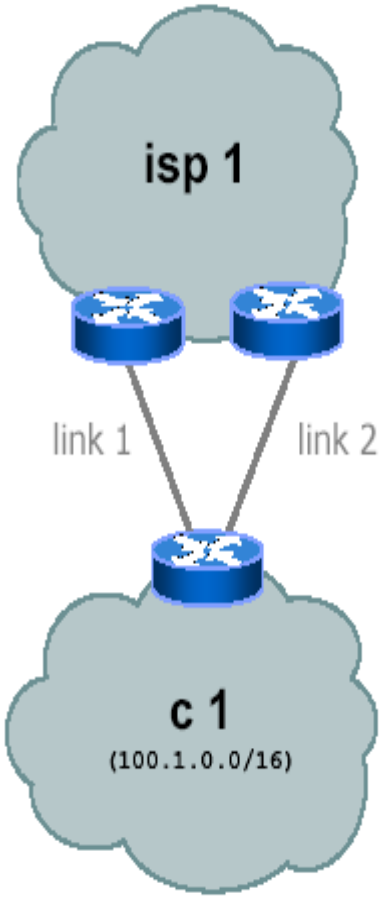
Address block assigned from ISP

194.138.230.0/24

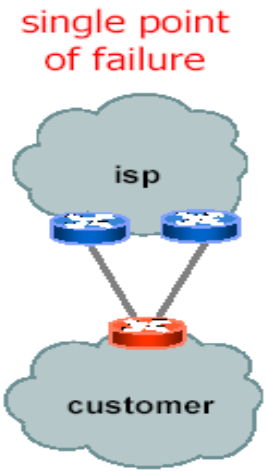
Multi-homed stub networks



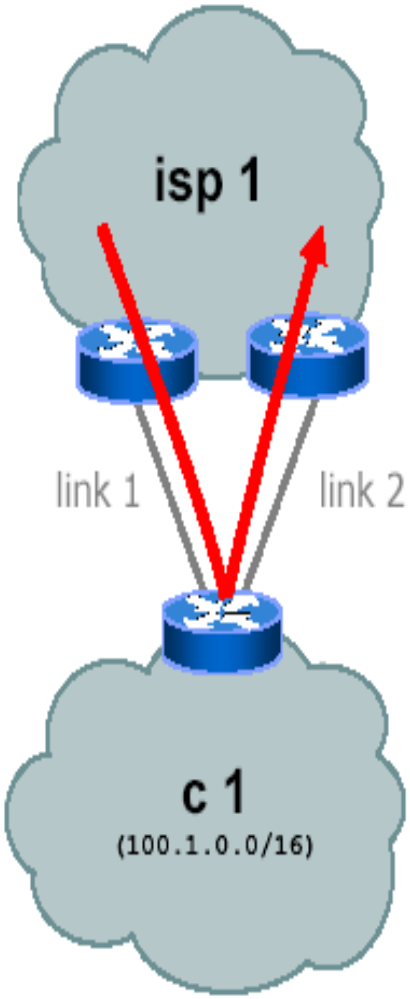
Multi-homed stub networks



- Two connections to ISP
- Usually two routers of customer network involved

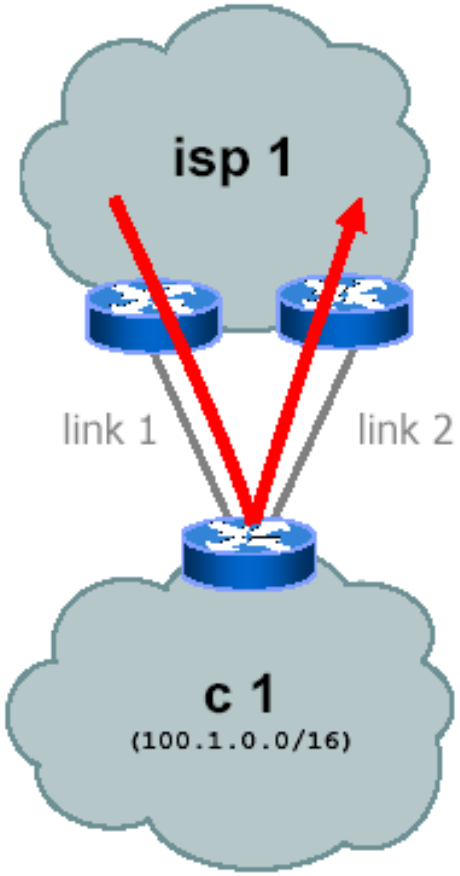


Routing



- A packet for the internet may traverse one of the links
- A transit packet may traverse both links
- Should not happen in a stub network

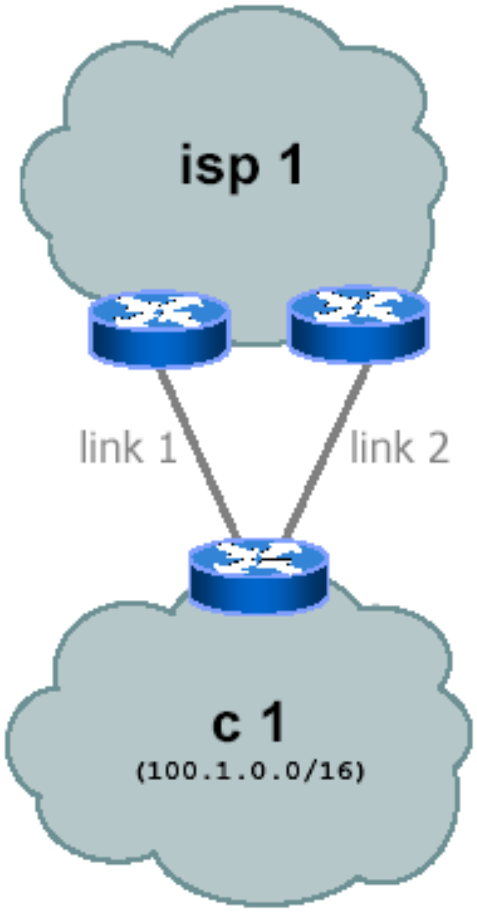
Desired policies - Backup



Example:

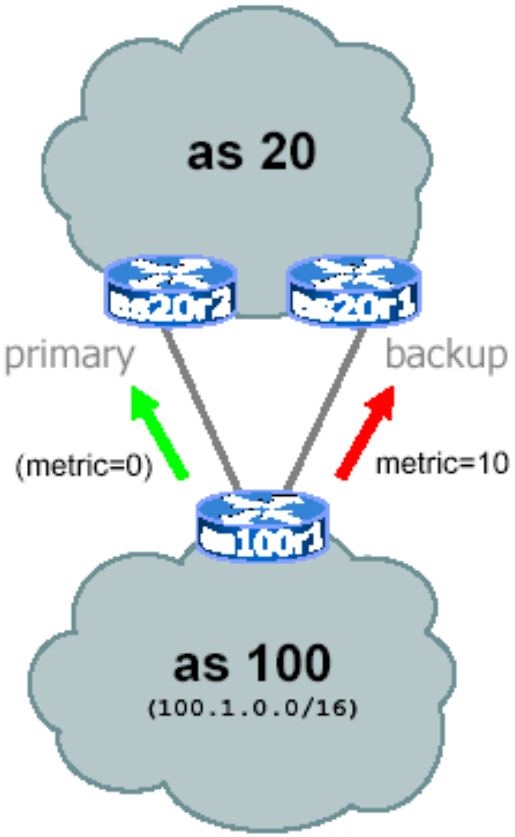
- Eliminate transit traffic
- Incoming traffic:
  - Use link 1
  - Use link 2 if link 1 faults
- Outbound traffic:
  - Use link 1
  - Use link2 if link 1 faults

Alternatives to BGP



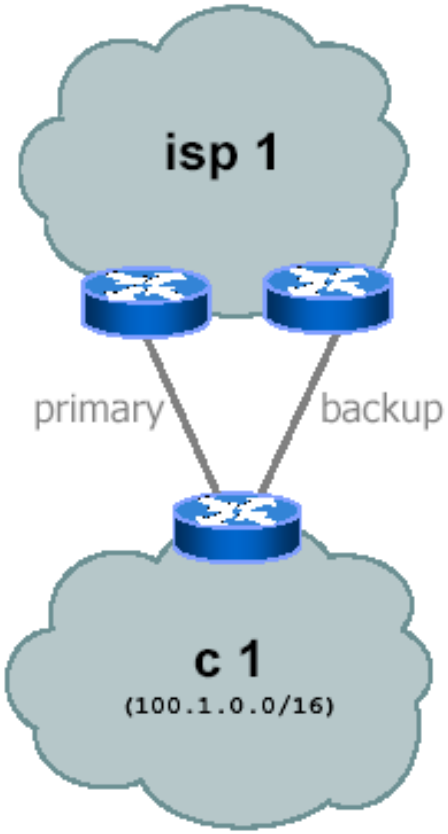
- Use IGP:
  - Packets traverse link 1 or link 2 according to shortest path to c1
  - not possible to exclude transit path if link1 and link2 on SP to destination
- Use static routes:
  - ISP's and c1's routers must be manually and consistently configured.
  - No automatic backup possible

BGP/MED



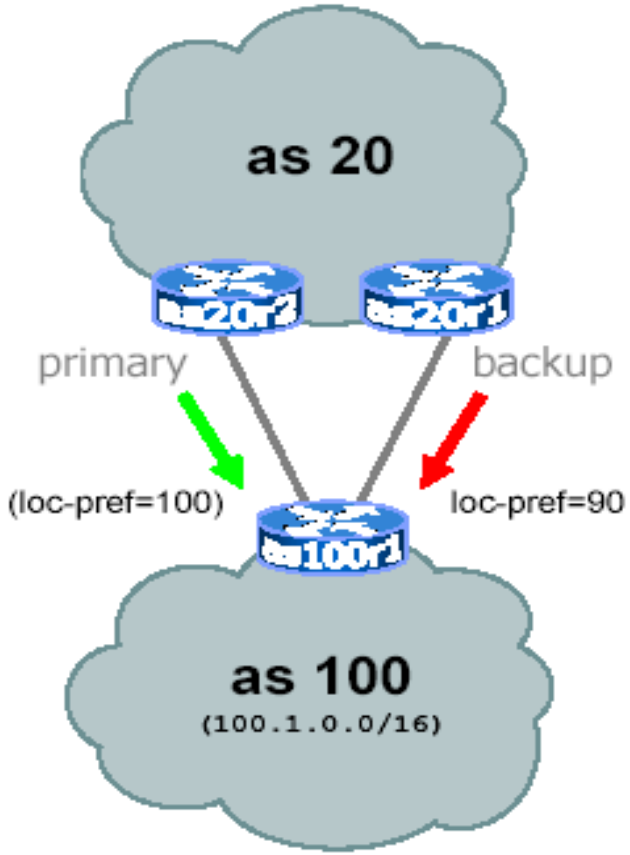
- the attribute called "metric" by the sender as, is called "multi-exit-discriminator" by the receiver as
- apon reception of the same announcement with two different meds, the provider will (hopefully) adopt the one with the smaller one
- default value is zero
- metric is set on outgoing announcements and manages inbound traffic flows

BGP strategy



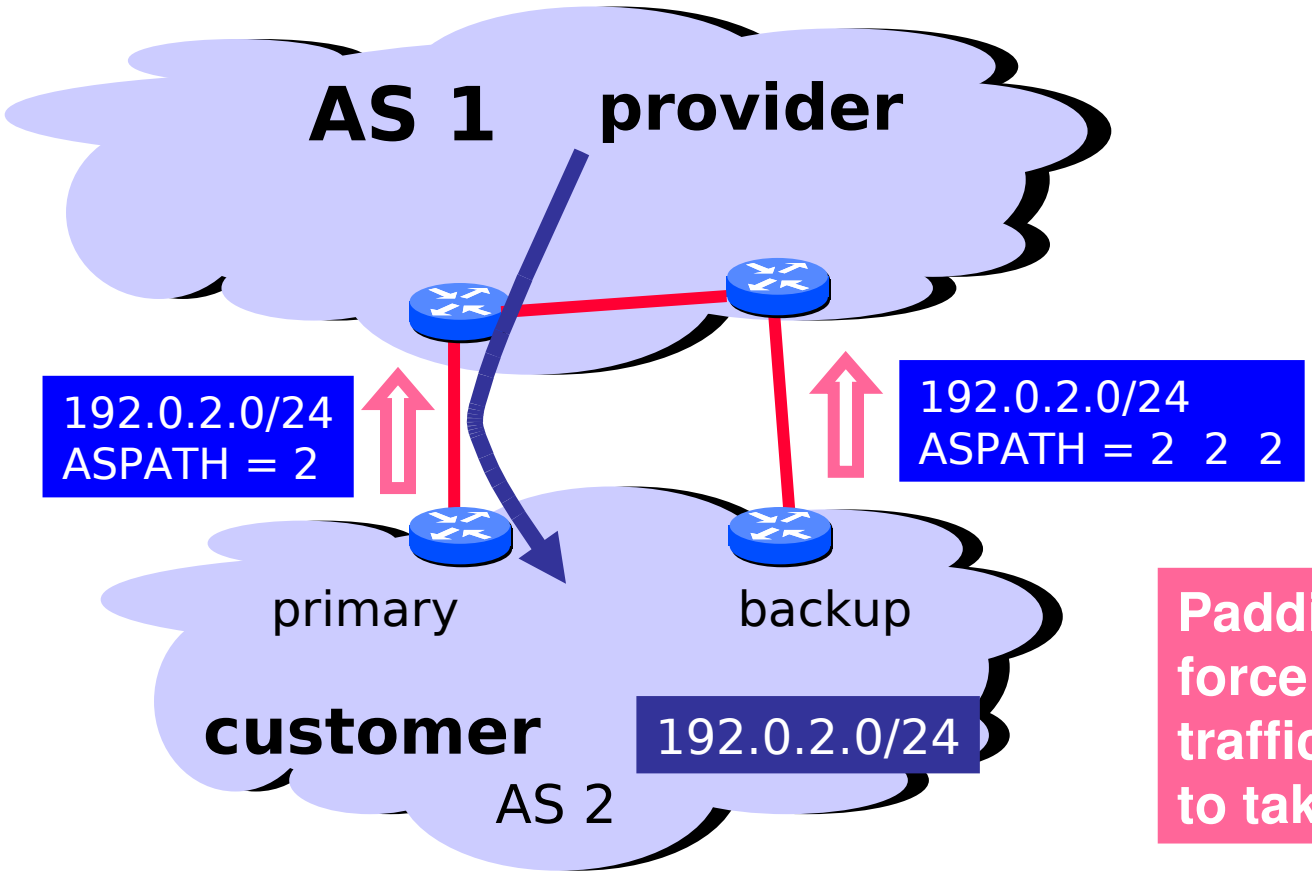
- announce 100.1.0.0/16 on both links:
  - Primary link sends standard announcement
  - Backup link increases MED on exit announcements and decreases LOCAL_PREF on inbound announcements
  - MED: MULTI_EXIT_DISCRIMINATOR
- When one link faults, /16 announcement on backup link ensures connectivity

BGP/Local Preference



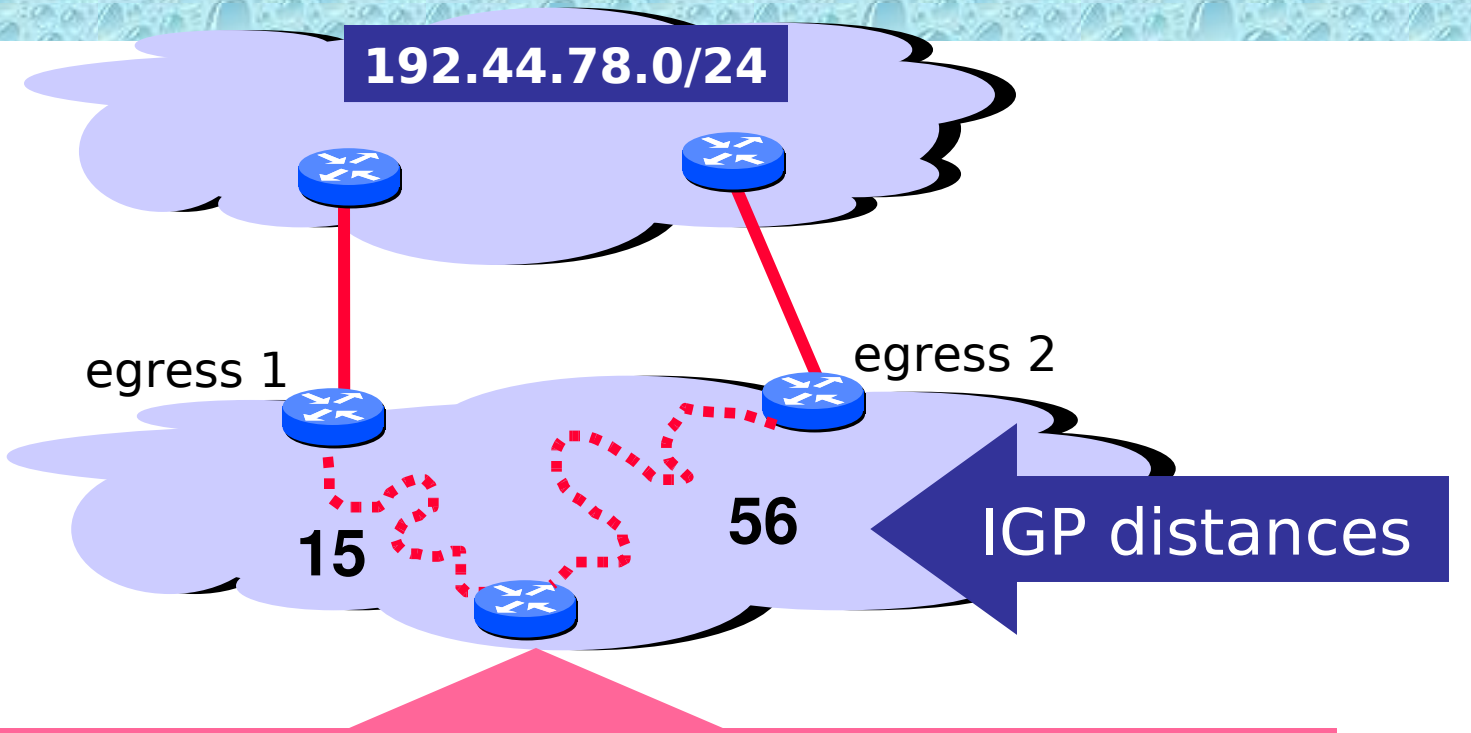
- the customer assigns a lower local-preference to the announcement coming from the backup peer
- local-preference attribute is checked before as-path length in the route selection process
- default value is 100
- local-preference applies to incoming announcements and manages outbound traffic flows

Shedding Inbound Traffic with ASPATH Padding Hack



Padding will (usually) force inbound traffic from AS 1 to take primary link

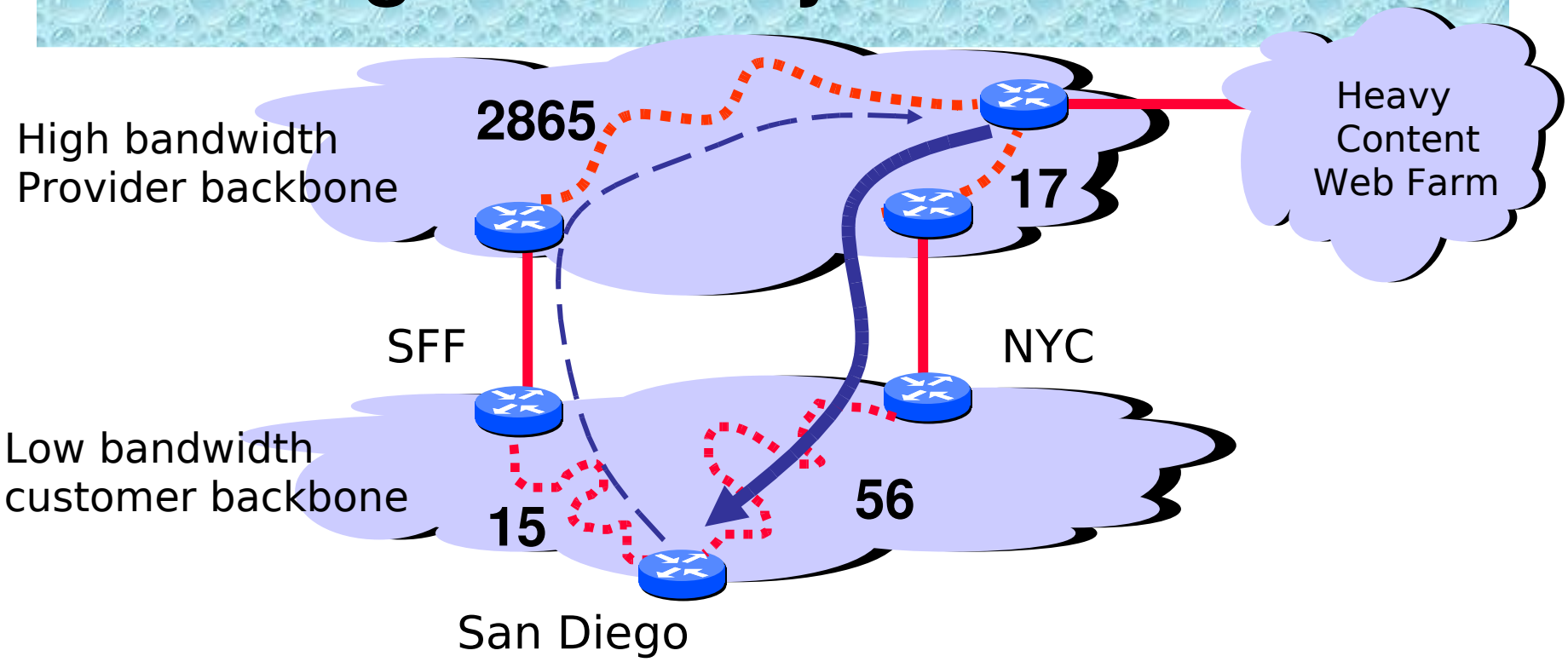
Hot Potato Routing: Go for the Closest Egress Point



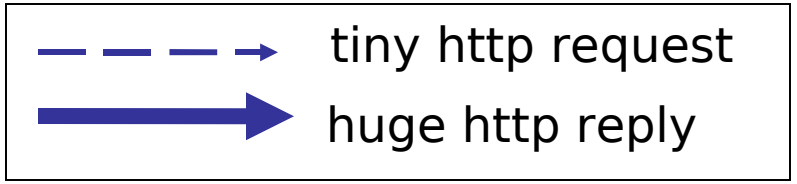
This Router has two BGP routes to 192.44.78.0/24.

Hot potato: get traffic off of your network as Soon as possible. Go for egress 1!

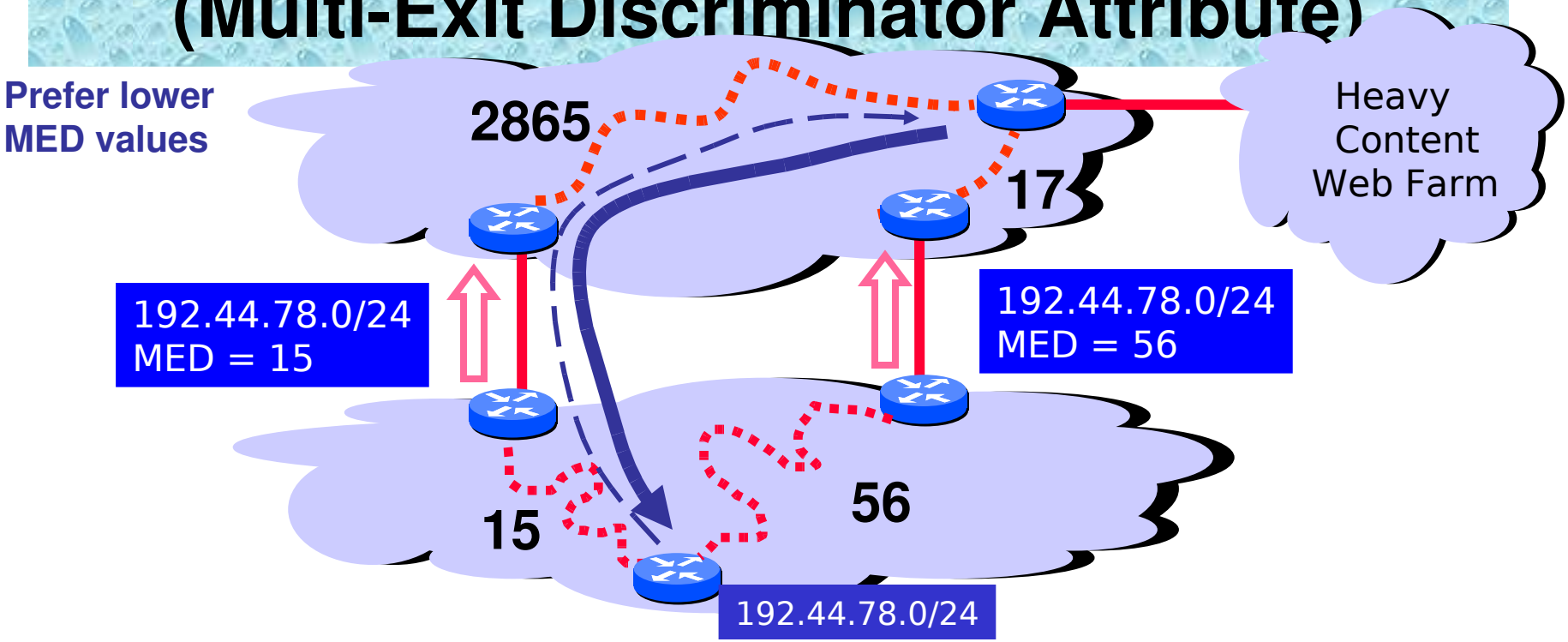
Getting Burned by the Hot Potato



Many customers want their provider to carry the bits!



Cold Potato Routing with MEDs (Multi-Exit Discriminator Attribute)

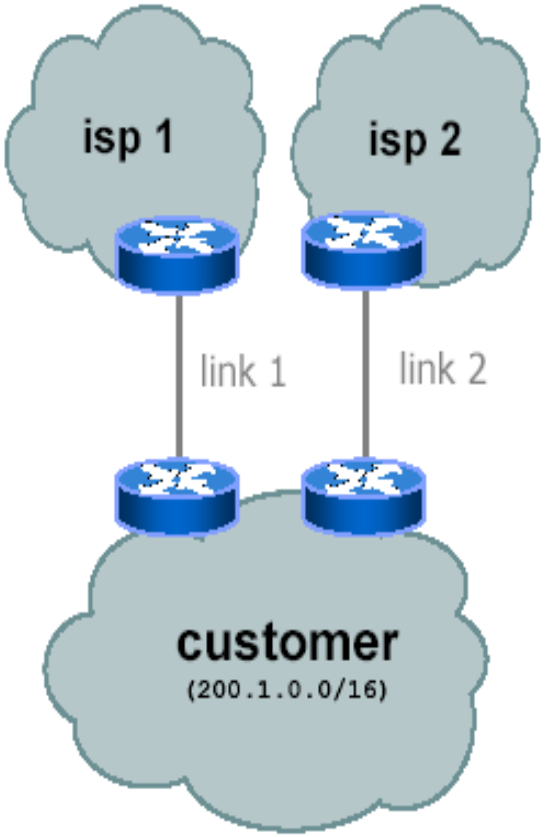


This means that MEDs must be considered BEFORE IGP distance!

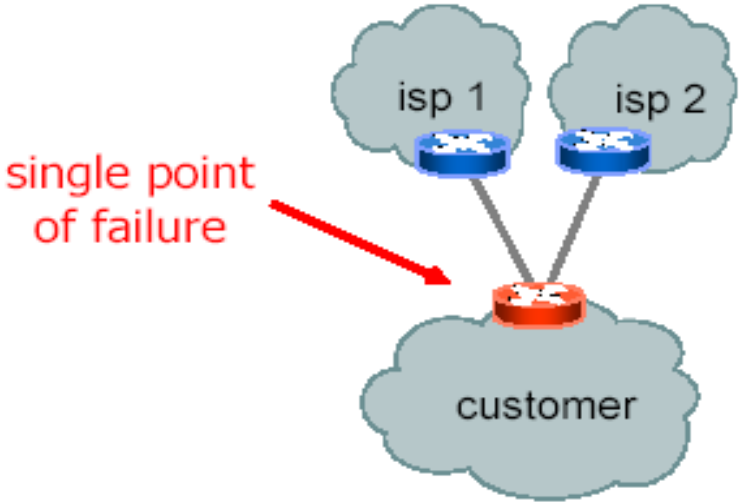
Note1 : some providers will not listen to MEDs

Note2 : MEDs need not be tied to IGP distance

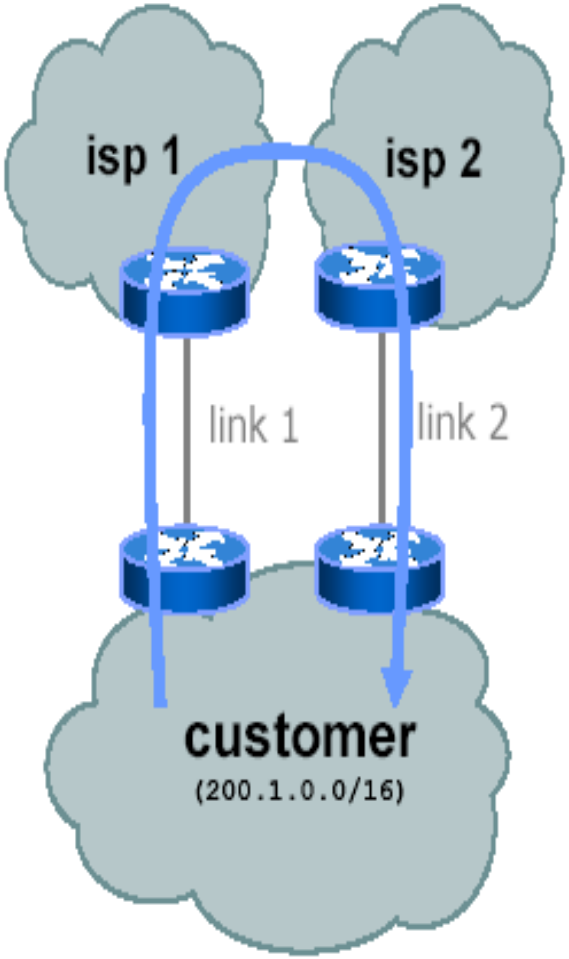
Multi-homed network



- Two (or more) links to different provider
- Typically two routers involved for fault tolerance

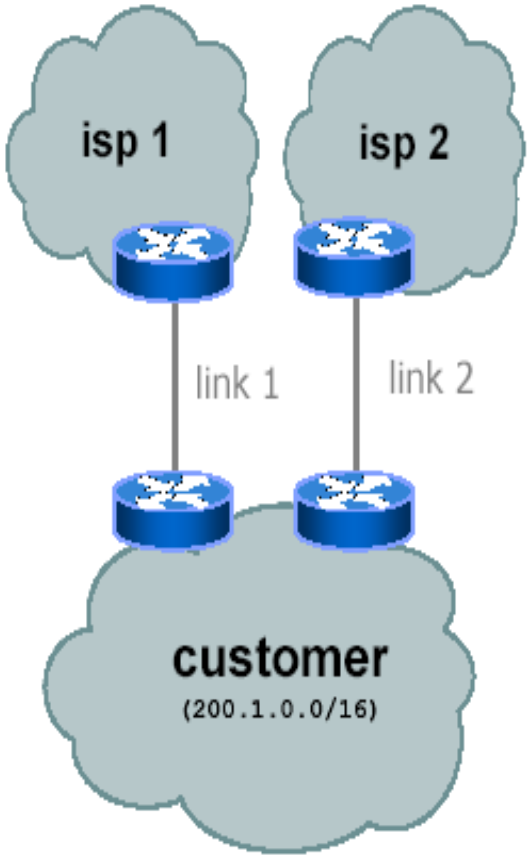


Routing



- Outbound packet may use any of the links to reach the internet
- Inbound pcacket may use any of the two links to reach AS
- Internet packet may cross both link 1 and link 2
- Local packet may cross both links

Load balancing



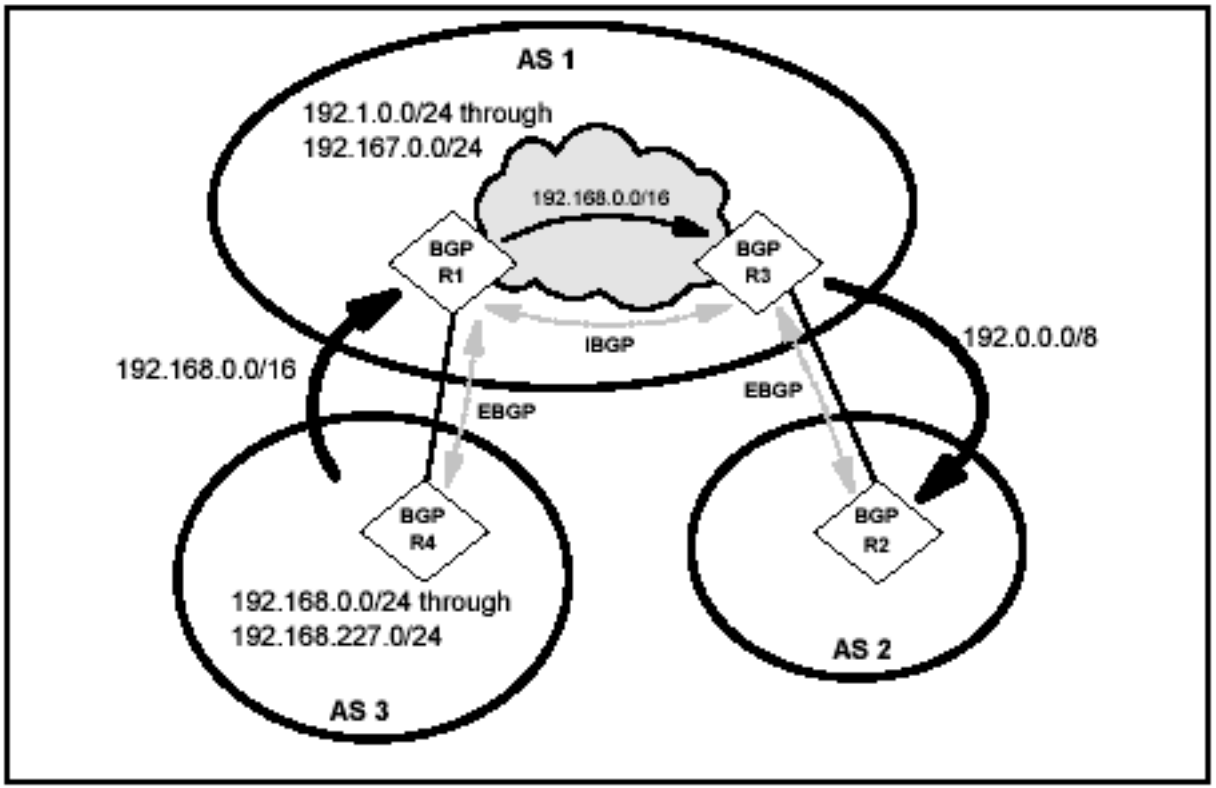
- Outbound traffic:
  - Half of the hosts use link link 1,
  - Other half uses link 2
- Inbound traffic:
  - Use link 1 to reach half of the hosts
  - Use link 2 to reach other half

Load balancing with BGP

- Inbound traffic: split /16 and send two /17 announcements, one per link
  - E.g.: 200.1.0.0/17 on link1 and 200.1.128.0/17 on link2
  - Approximate partitioning of inbound traffic
  - Assumes same link capacity and uniform distribution of traffic with respect to AS'
  - In practice: modify split until load is suitably balanced
- Outbound traffic: accept default upstream routing
  - Use IGP to partition traffic (hot potato)

Route aggregation

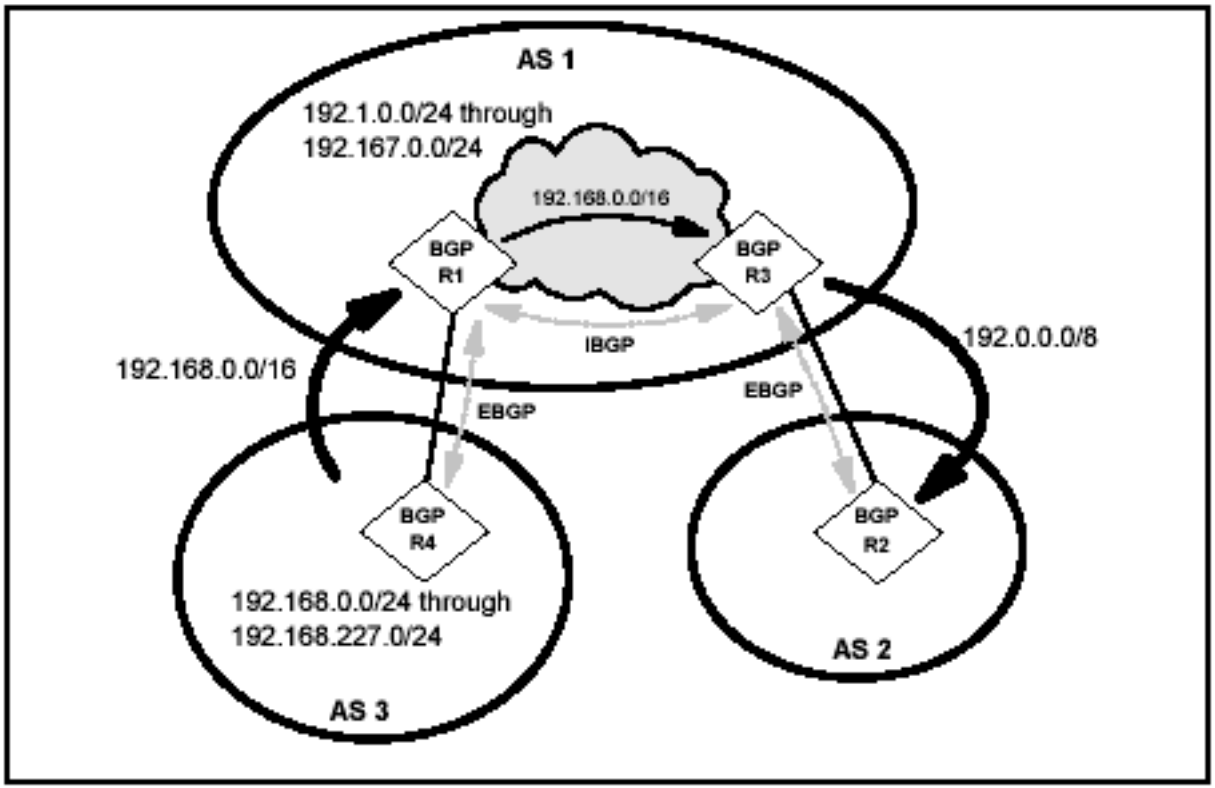
- BGP v4 uses CIDR for route aggregation
- Increases scalability



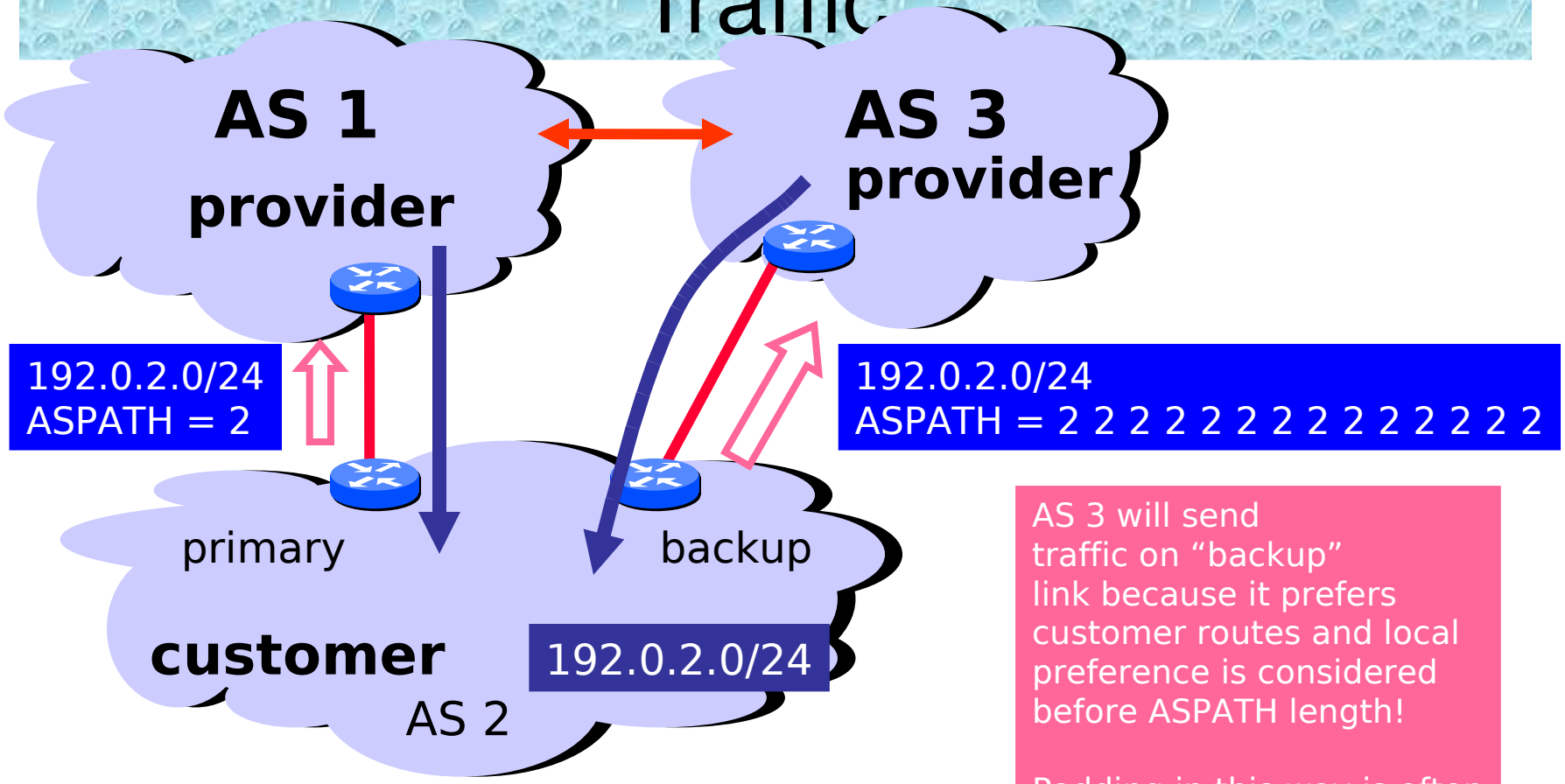
Attenzione: 182….

Route aggregation/cont.

- R4 announces 182.168.0.0/16 <3> to R1
- R3 announces 182.0.0.0/8 <1 3> to R2



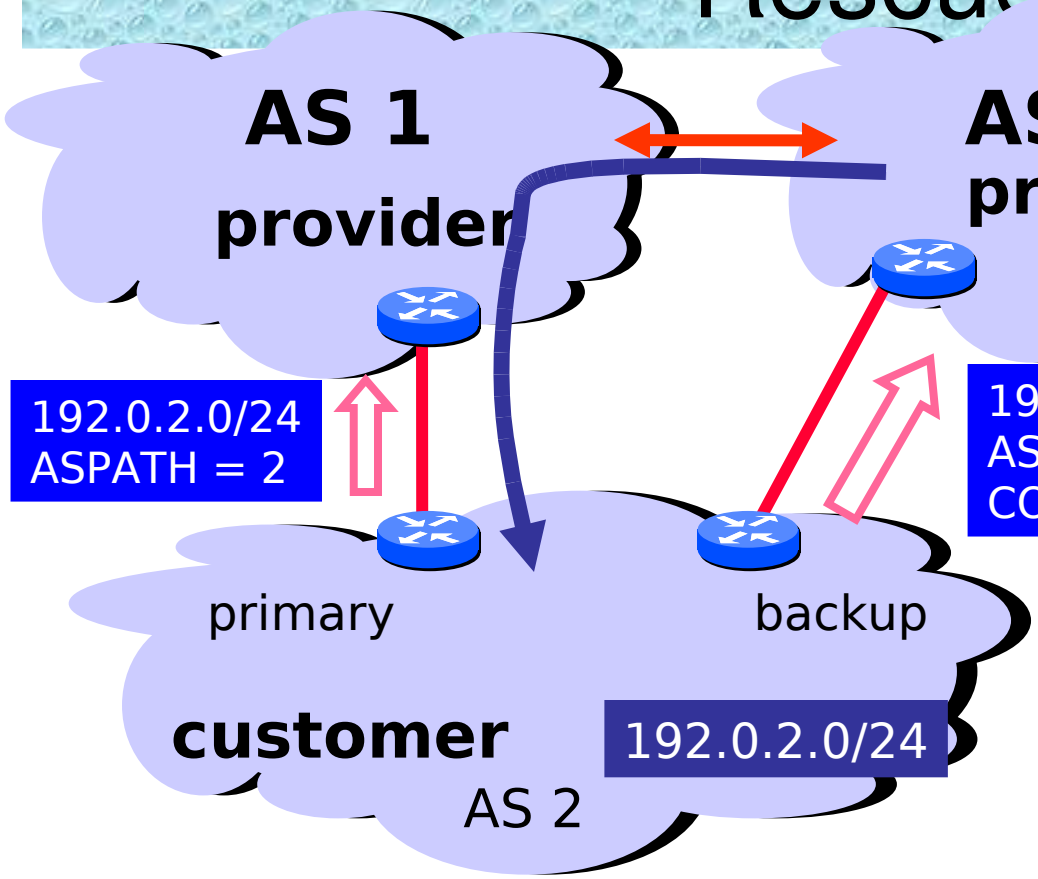
Padding May Not Shut Off All Traffic



AS 3 will send traffic on "backup" link because it prefers customer routes and local preference is considered before ASPATH length!

Padding in this way is often used as a form of load balancing

COMMUNITY Attribute to the Rescue!



AS 3: normal customer local pref is 100, peer local pref is 90

Customer import policy at AS 3:
If 3:90 in COMMUNITY then
set local preference to 90
If 3:80 in COMMUNITY then
set local preference to 80
If 3:70 in COMMUNITY then
set local preference to 70

References

- General
  - TCP/IP guide: http://www.tcpipguide.com/free/t_BGPFundamentalsandGeneralOperation.htm
  - White paper CISCO: http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a00800c95bb.shtml
- Attributes: http://www.cisco.com/en/US/docs/internetworking/technology/handbook/bgp.html