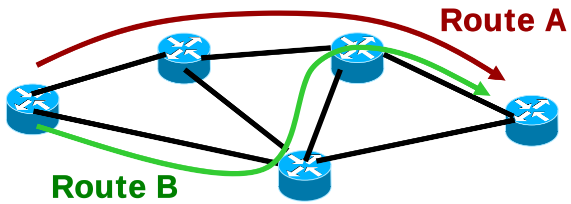
Some terminology

Terminology for packet switching networks

- ***Forwarding*: determining the output (port, “connection”) for a data packet**
  - **It may be done at many layers (link layer, IP layer, application layer)**
    - **The exchange/forwarding unit at the link-layer is the *frame***
    - **The exchange/forwarding unit at the network layer is the *packet***
    - **The exchange/forwarding unit at the application layer is the *message***
  - **It is done at every node through which the packet traverses**
    - **Eg: forwarding at Ethernet link-layer performed at node and *bridges***
    - **Forwarding at network layer performed at host and *routers***
    - **Application layer: local application entity, next application entity**
  - **Forwarding belongs to the *data-plane* (i.e. triggered by data packets)**
  - **Many forwarding strategies**
    - ***Flooding* (send through all available outputs)**
    - **Use a *spanning tree* for all the packets**
    - **Use information specific to the destination**
      - To indicate the outgoing port that correspond to a destination
        - Note that for this, packets must have destination
    - **Use detailed information carried in the data packet (*source routing*)**

Terminology for packet switching networks

- ***Routing*: control-plane function that determines the path to each destination and configures the forwarding function**
  - **Note that *forwarding* and *routing* are decoupled**
    - **i.e different routing mechanisms can be used to generate forwarding information**
  - **It works over an *identifier* for the destination**
    - **An identifier can be termed *name* or *address* regardless its dependency on location**
      - *Names* are independent of the location of the object identified
      - *Addresses* are dependent of the location
  - **The *route* is the path (or paths) that can be used to carry a data packet from a given point to its destination(s)**



  - **Routing can be**
    - ***Dynamic*: A *routing protocol* can be used to exchange routing information, and a *routing algorithm* is used to compute the routes**
    - ***Static*: forwarding information is configured manually by the router administrator, using network management…)**

Terminology for IP packet switching

- **IP layer is in charge of forwarding packets among different links**
  - **A *link* (from IP perspective) is the network region to which packets can be delivered with TTL=1, i.e. where just one IP forwarding operation is being performed**
- **Interfaces are identified at the IP layer by its *IP address***
  - **IP does not identify nodes but interfaces**
  - **These identifiers are termed addresses, because they depend on the location of the interface**
  - **Upper laters (transport, some times applications) use IP addresses as identifiers**
    - ✓ **This is a problem some times, since too much coupling exists among layers**
- **Networks are identified by *IP prefixes*, which are aggregations of contiguous IP addresses (eg. 163.117.139.0/24)**
  - **Aggregation reduces the amount of information to exchange by the routing protocol**
    - ✓ **It improves routing scalability**
      - Scalability: a system serves N users with R resources. The system scales if the function R=f(N) is linear or less than linear
  - **Unicast routing generates an *IP forwarding table* for each node, in which ONE output is determined for each destination prefix**
- **The forwarding algorithm is *Longest Prefix Match***

BGP
An introduction

Alberto García

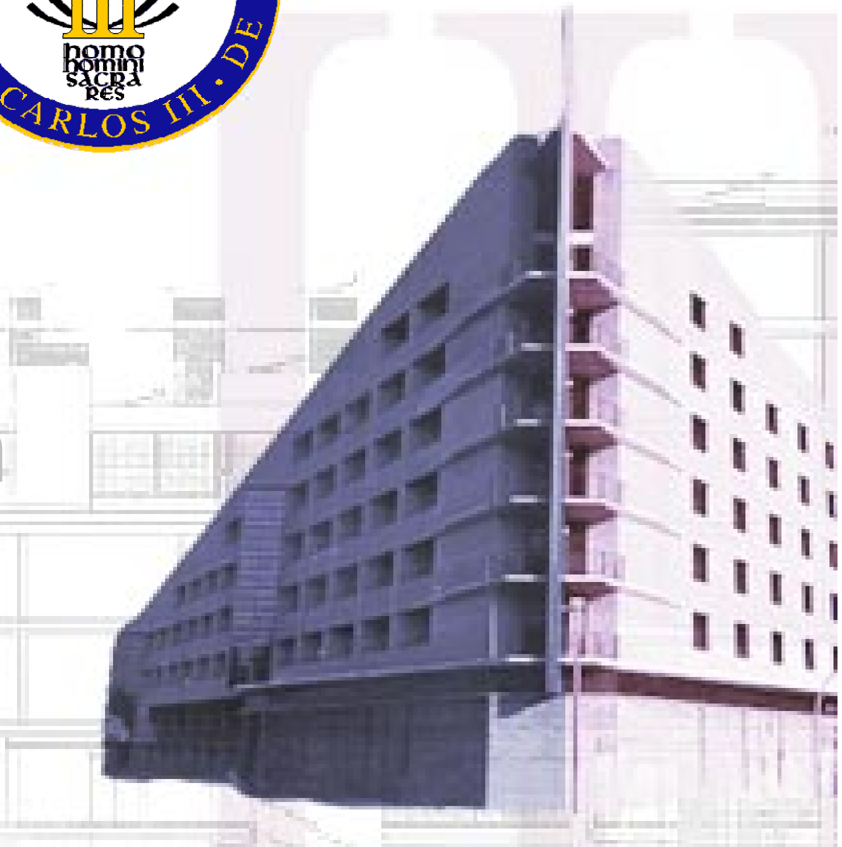
References

- **Books**
  - ***BGP. Building Reliable Networks with the Border Gateway Protocol.*** **Iljitsch van Beijnum. O' Reilly. 2002.**
    - ✓ **Available http://proquest.safaribooksonline.com/9780596002541 (when connecting from PCs located at UC3M or UPC)**
  - ***Internet Routing Architectures*. Sam Halal, Danny McPherson. Cisco Press. 2000.**
    - ✓ **http://proquest.safaribooksonline.com/1-57870-233-X**
  - ***Routing in the Internet*. Christian Huitema. Prentice Hall. 2000.**
- **RFCs**
  - ***RFC 4271. A Border Gateway Protocol 4 (BGP-4)*. Y. Rekhter, T. Li. S. Hares. Enero 06**
  - ***RFC 4274. BGP-4 Protocol Analysis.* D. Meyer, K. Patel. Enero 2006**
  - ***RFC 4276. BGP-4 Implementation Report.* S. Hares, A. Retana. Enero 2006.**
  - ***RFC 4277* Experience with the BGP-4 Protocol. D. McPherson, K. Patel. Enero 06.**
  - ***RFC 1930: Guidelines for creation, selection and registration of an Autonomous System (AS)***

References

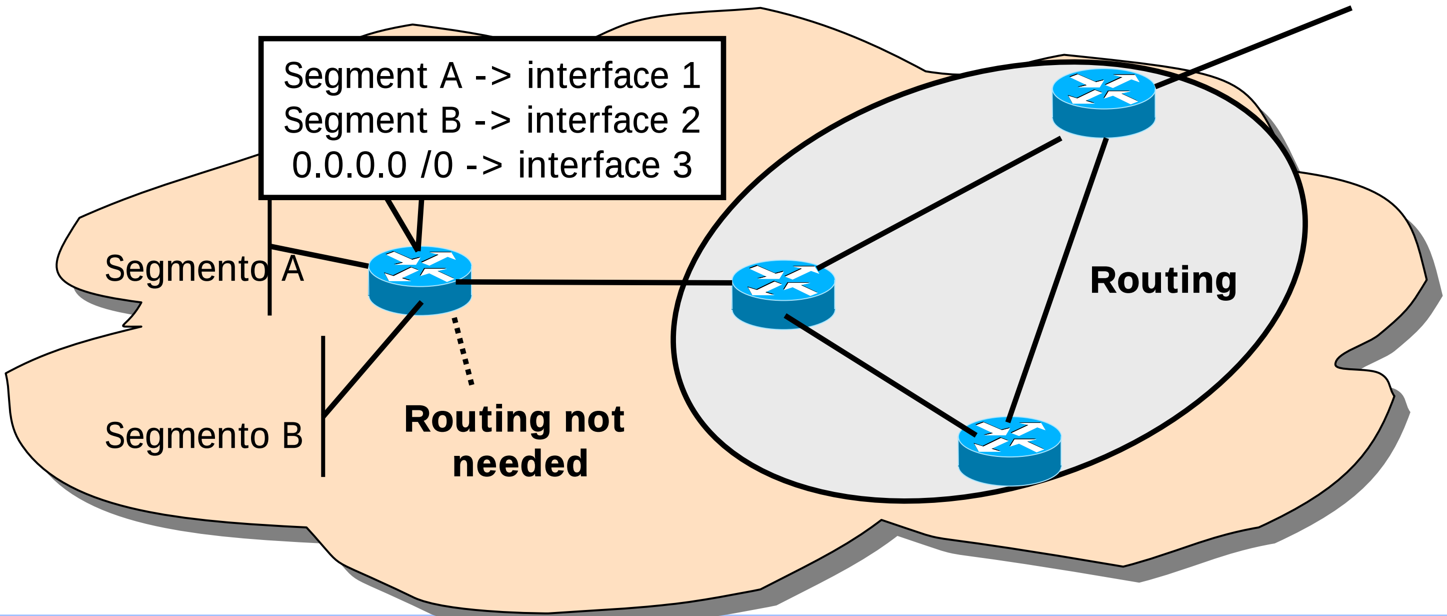
- *ISP column, Geoff Huston, www.potaroo.net*
  - Damping BGP, *June 2007*
  - 32-bit AS Numbers – The View from the old BGP World, *January 2007*
  - The BGP Report for 2005, *Jun 2006*
  - An Introduction to BGP - The Protocol. *May 2006*
  - *Exploring AS Numbers. Aug 2005*
  - The State of Inter-Domain Routing. *Mar 2004*
- *The Art of Peering - The Peering Playbook*. William Norton. arneill-py.sacramento.ca.us/ipv6mh/playbook.pdf
- *Internet Service Providers and Peering*. William Norton.
  http://arneill-py.sacramento.ca.us/ipv6mh/PeeringWP1_91.pdf
- *A business case for ISP peering*. William Norton. http://arneill-py.sacramento.ca.us/ipv6mh/ABusinessCaseforISPPeering1.2.pdf

Providing Interdomain Connectivity



Why routing?

- **Forwarding information can be obtained from**
  - **Dynamic routing, i.e. from routing protocols**
  - **Static routes (for simple topologies)**
  - **Mixed (static / dynamic)**
- **Why using (dynamic) routing**
  - **Configuration is easier by using routing protocols**
  - **If many paths exist, routing provides FAULT TOLERANCE, so there should be routing**

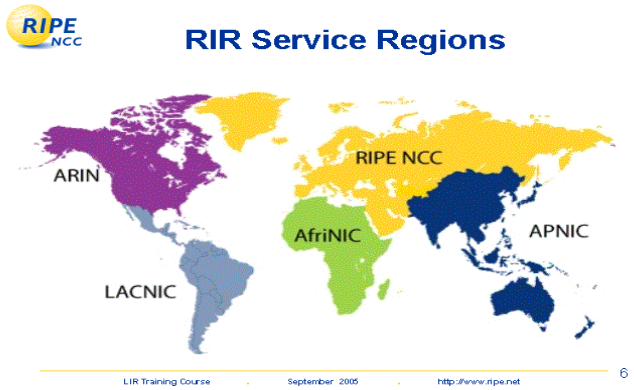


Routing in the Internet

- **There are too many network segments to be able to exchange detailed information between each node.**
  - **Current routing protocols are not designed to scale to such a size**
    - **Cost of computation of good routes**
    - **Large bandwidth requirements to exchange periodic routing information**
    - **Long convergence times; long recovery times after failure**
- **Solution: Abstract the routing information**
  - **Terminal equipment and routers belong to *administrative domains***
  - **Two independent levels of domains can be identified**
    - **Inter-domain Routing: The more "exterior" routing between distinct administrative domains**
      - Each administrative domain appears as a single network node
      - Every administrative domain shares the same type of information (speaks the same protocol)
    - **Intra-domain Routing: The more "interior" routing within an administrative domain**
      - The administrative domain does its own internal connectivity

Basic elements managed by the interdomain routing protocol

- THE PROTOCOL is BGP
- BGP transports reachability information for prefixes
  - Valid public prefixes: prefixes assigned by RIRs
    - ✓ or combinations of them: more specific aggregations
- "Domain" in BGP is an Autonomous System
  - 16 bit number
  - [0-64512] assigned by RIRs
    - ✓ [64513-65536] private AS numbers



READ BGP, Iljistch, pp. 62-71

Which flavor of protocol to choose?

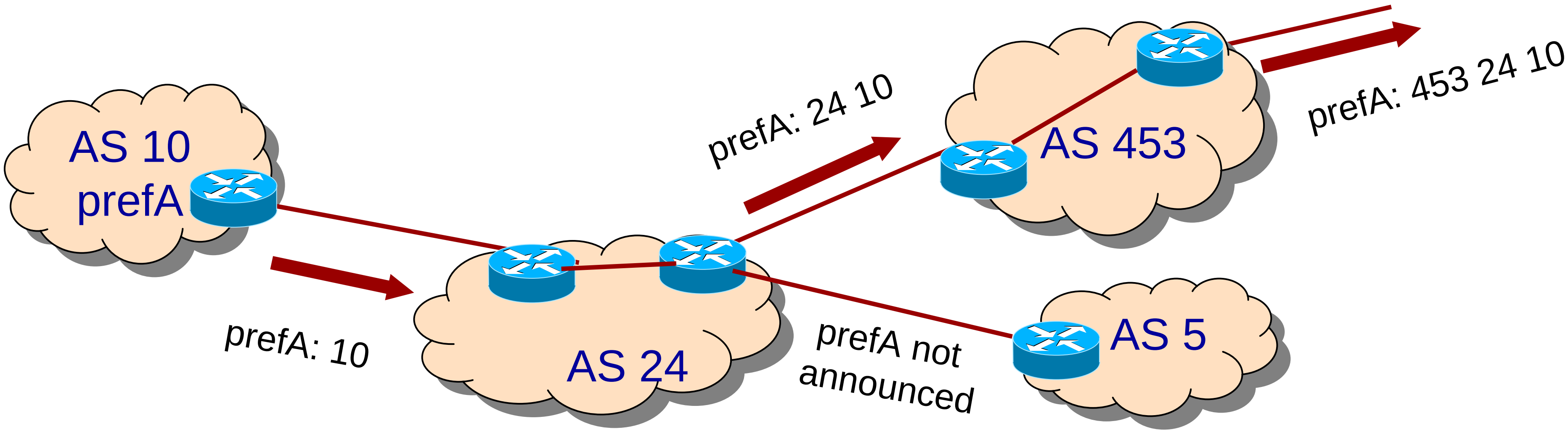
LINK STATE

- Router propagates to all routers the (correct) info of its neighbors
  - Routers send once the information to converge
- Carries weights (any distance can be used)
- Allow shortest path computation (Dijkstra algorithm)
  - Routers have complete topology
- Requires shared and uniform policies

VECTOR

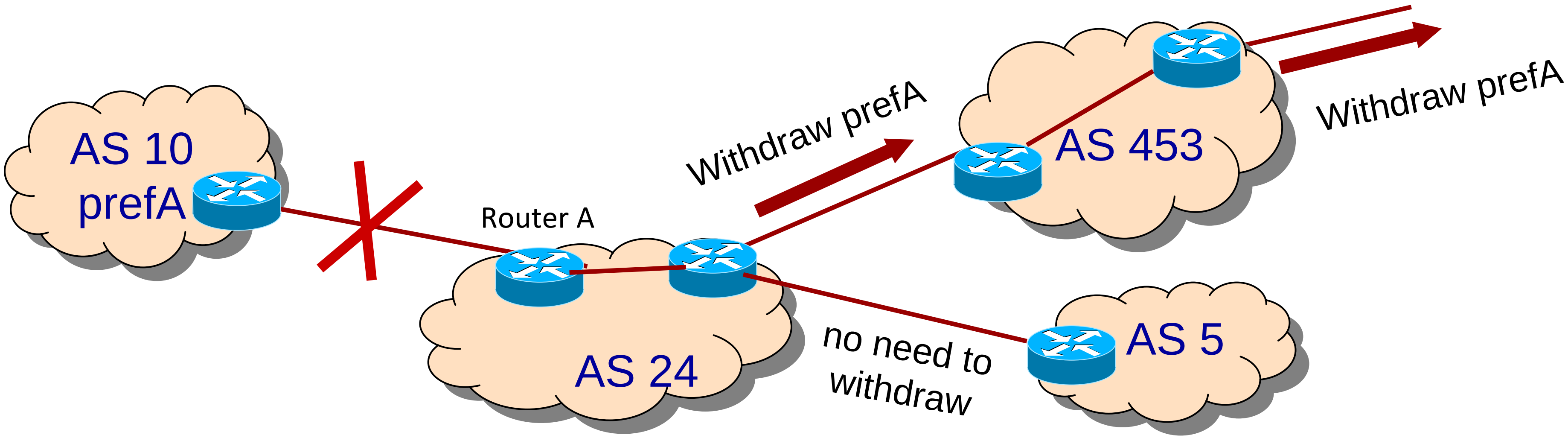
- Router propagates (only) to its neighbors (all) the info it has
  - Needs many exchanges to converge
- Does not carry weights (implicitly distance is number of hops)
- Allow shortest path computation (Bellman-Ford algorithm)
  - Routers only know next hop
- Well-known issues: count to infinite for detecting link loses…
- Each router can apply its own policy without any coordination

BGP is a path vector protocol



- **If an AS advertises a prefix to an adjacent AS => it indicates to its neighbour that**
  - **It knows how to reach any address with this prefix**
  - **It is willing to forward traffic for any address within this prefix**
- **Path Vector protocol: transmits a list of the AS numbers that were traversed to propagate info for the destination**
  - **Allows loop detection**
  - **The information can be changed or filtered during propagation**
    - ✓ **The announce itself may not be announced to any neighbor**
- **Data flows in the opposite direction of advertisements**

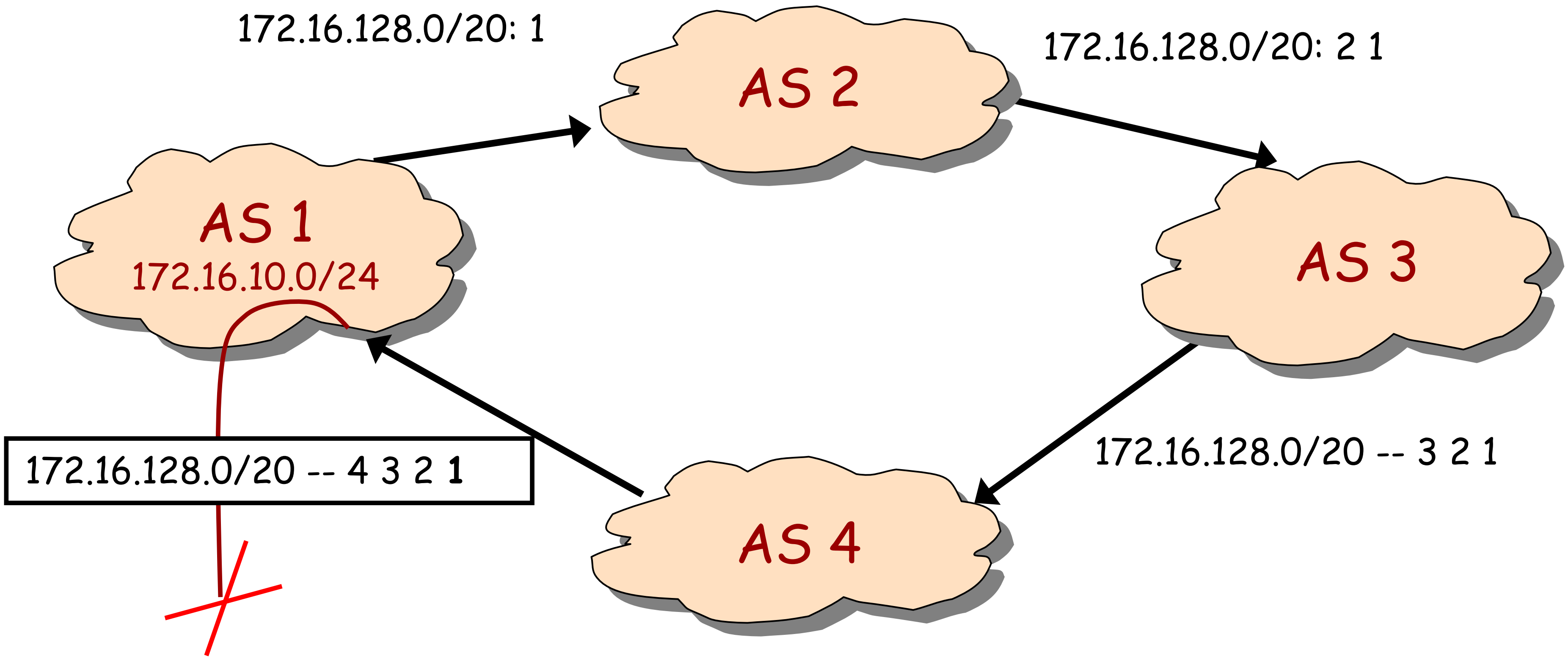
Route withdrawal in BGP



- **Router A detects that a BGP session fails, it deletes all the routing info received through this session**
  - **Tries to find an alternative route to prefA (it does not exist)**
  - **Finally realizes that prefA is no longer reachable**
- **Propagates a withdrawal request to all routers to which it has previously announced the prefix**

- **If a change occurs in the route, it generates a new advertisement with the list of Autonomous Systems traversed in the new route**

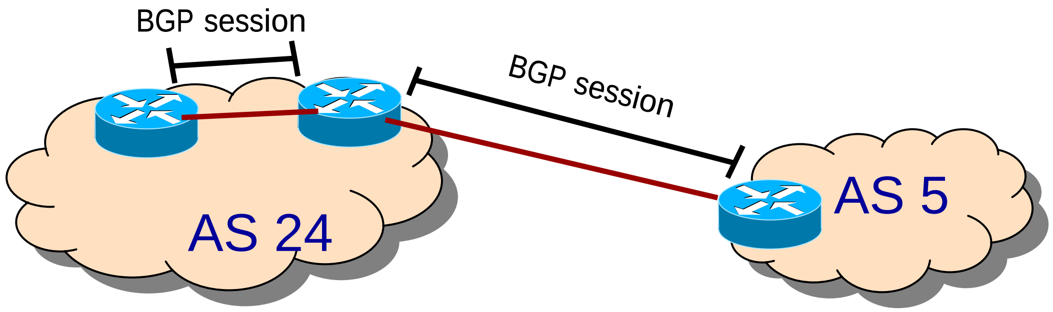
Loop Detection



- **Loops: Each AS checks if its own AS number is on the path list**

BGP: Basic Functionality

- **Two BGP neighbor routers establish BGP session by starting a connection using TCP (port 179)**
  - **Only "neighbor" routers establish connections**



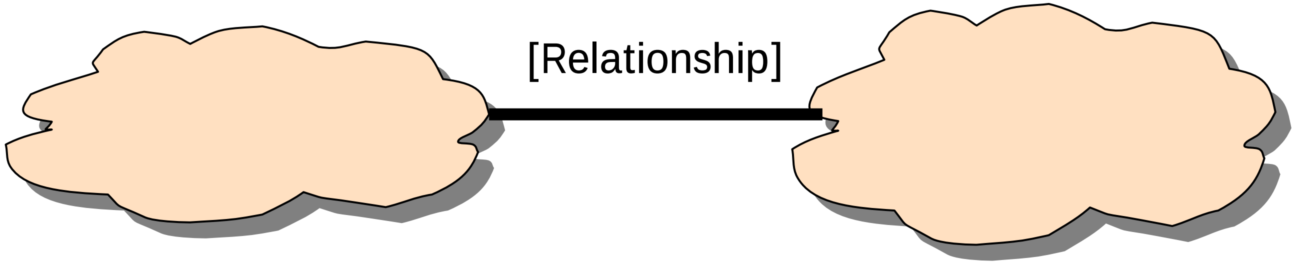
- **At first, each peer sends ALL its routing information**
  - **Can contain local information or from other previously established BGP sessions**
- **Afterwards, only routing changes (incremental protocol)**
  - New or modified routes
  - Withdrawals of previously transmitted routes

  ➔ **There are no refreshes, only changes!**
  - Different to the predecessor of BGP
  - **There is a mechanism to detect that the neighbor is alive (exchange of KEEPALIVE messages)**

Business model

Alberto García

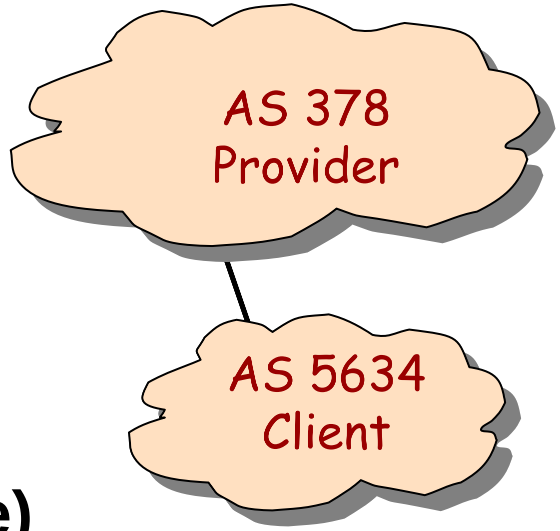
Relationships among ASs



- **Business models promotes THREE main relationship types among two ASs**
  - **One provides connectivity to the other -> TRANSIT (client – provider)**
    - ✓ **Transit is always PAID**
  - **Both want to use the link just to communicate with each other -> PEERING (peer - peer)**
    - ✓ **Peering may be FREE, or PAID**
  - **“Mutual transit” relationship or SIBLING (sibling – sibling). It is established among ASs with close relationship; the link is used in general as peering, but can be used as backup**
    - ✓ **Different ASs for the same company, mergers, acquisitions, …**
- **The first two options account roughly for 99.___ % of the relationships in the Internet**
  - **[Paper in ACM CCR ene 07]: sampling among many link relationships, peering: 77.6%, transit: 19.2%, siblings: 2.7%**

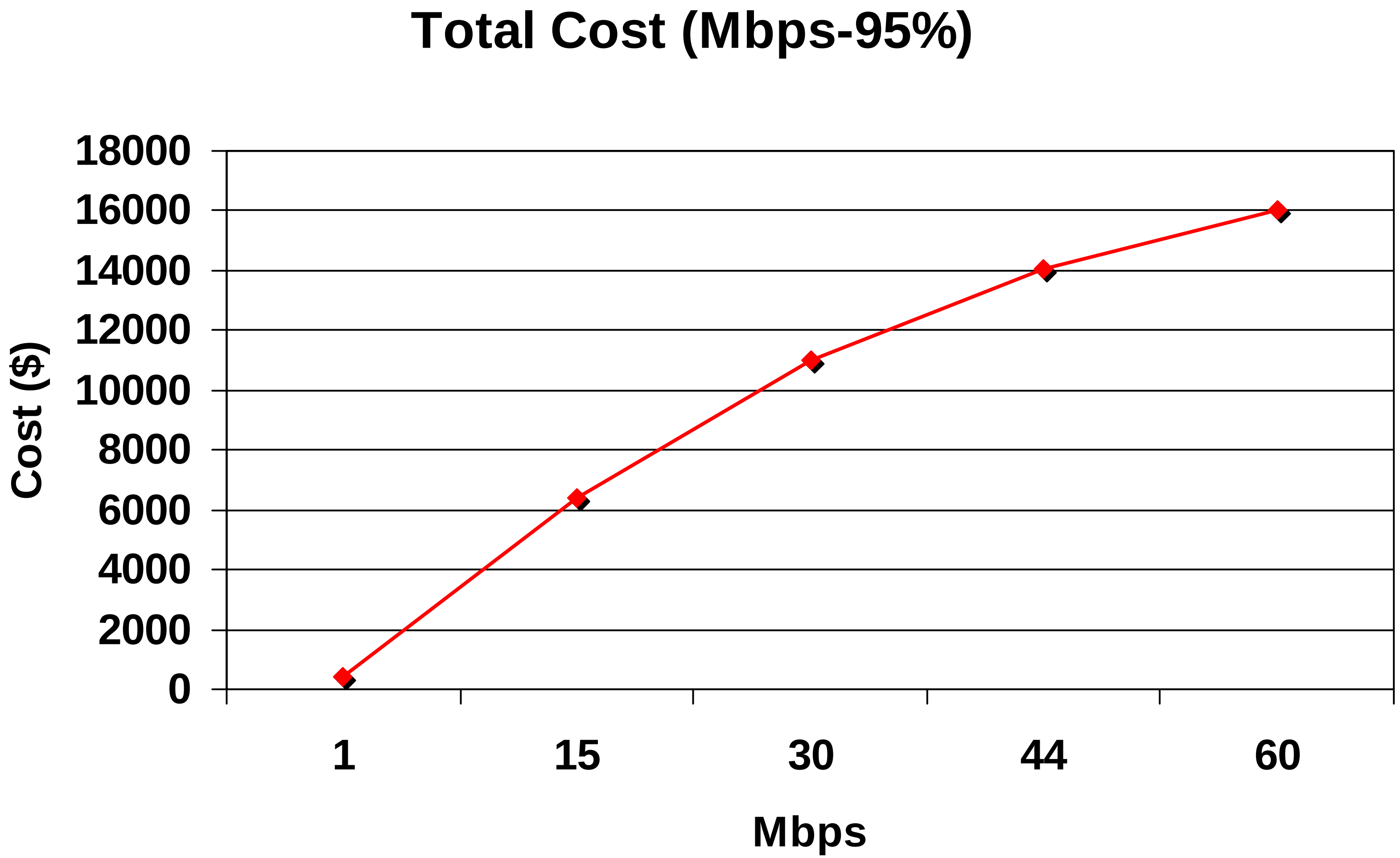
READ: *A business case for ISP peering*. William Norton. http://arneill-py.sacramento.ca.us/ipv6mh/ABusinessCaseforISPPeering1.2.pdf

Transit Costs

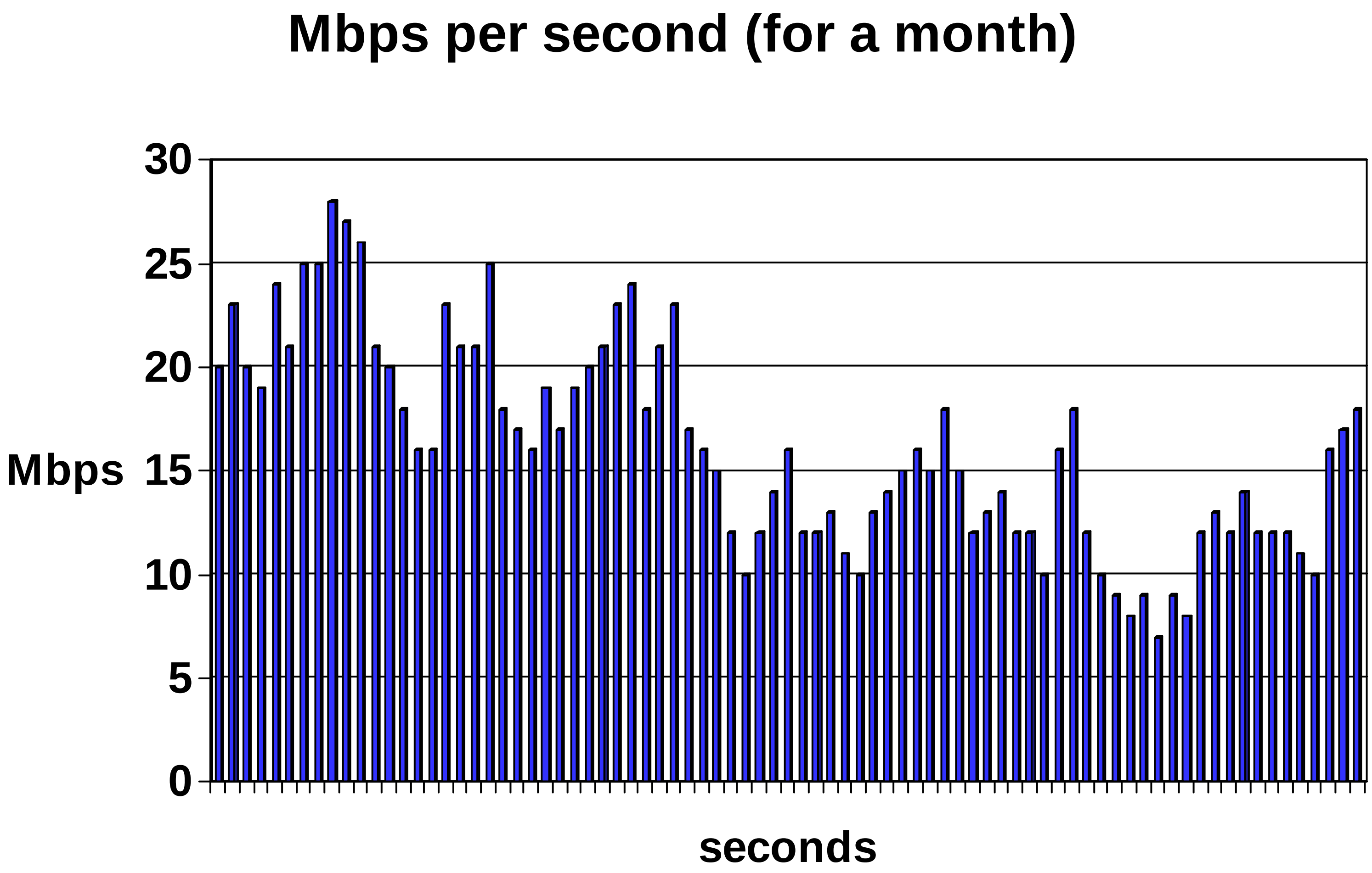


- **Cost of establishing the service**
- **Cost:**
  - **Line +**
  - **Traffic crossing (some times flat rate)**
    - **Measurement process:**
      - Take the average in 5 minute intervals, and see the 95 percentile
    - **The cost per megabit falls with size of consumption**
    - **The cost per megabit falls around 30% annually**

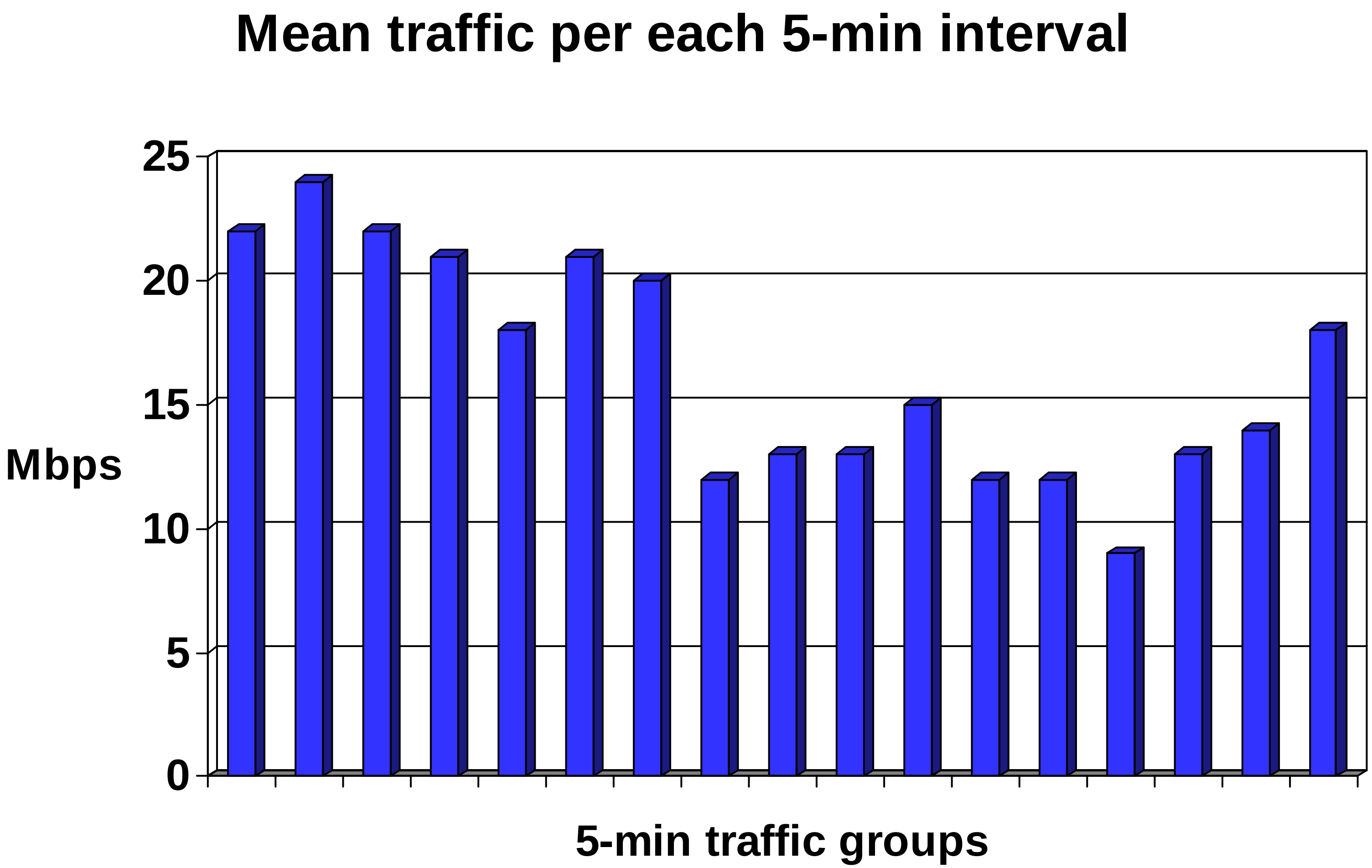
Transit cost: ISP transit rate



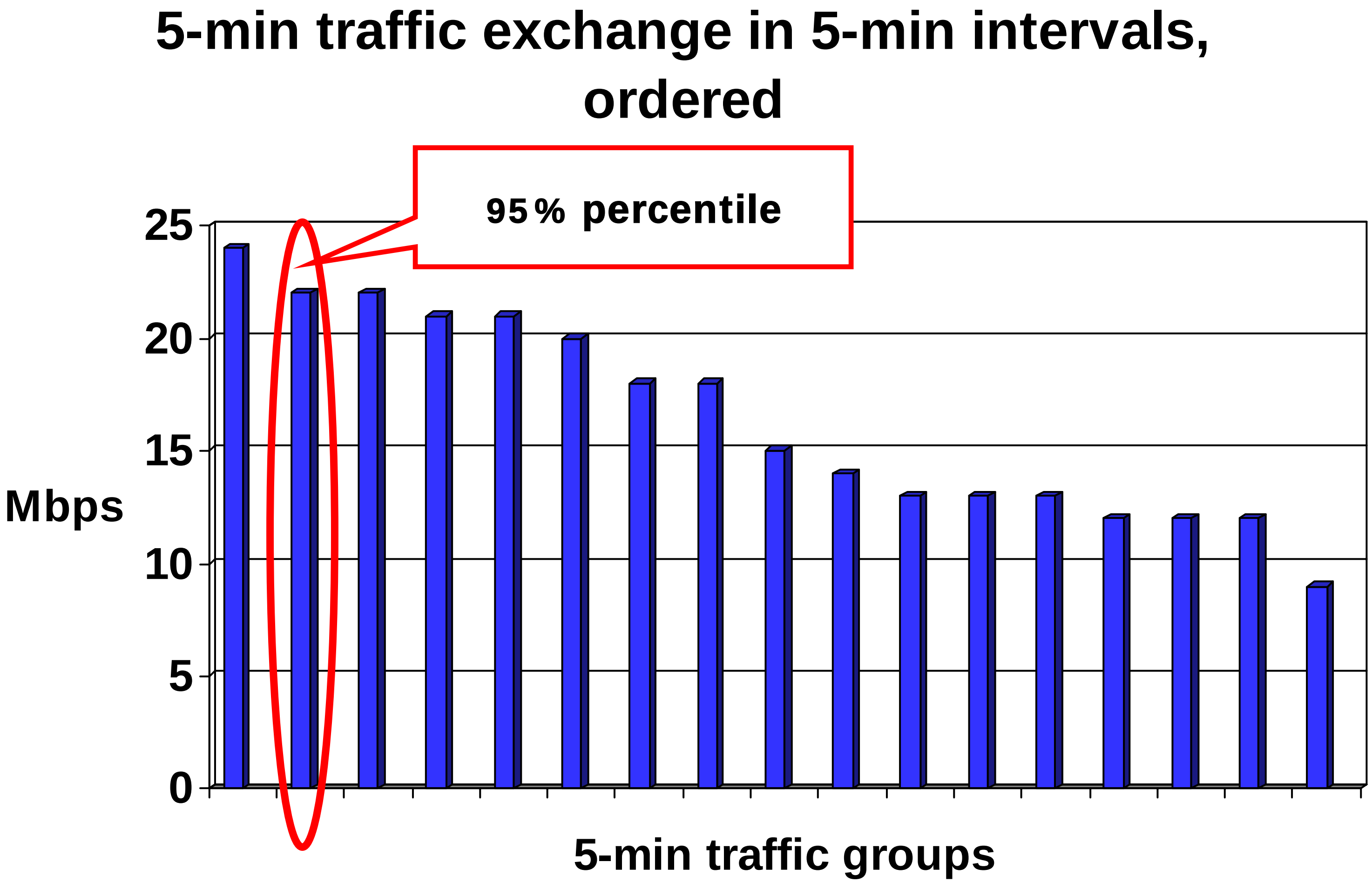
Transit cost: Exchanged traffic per second



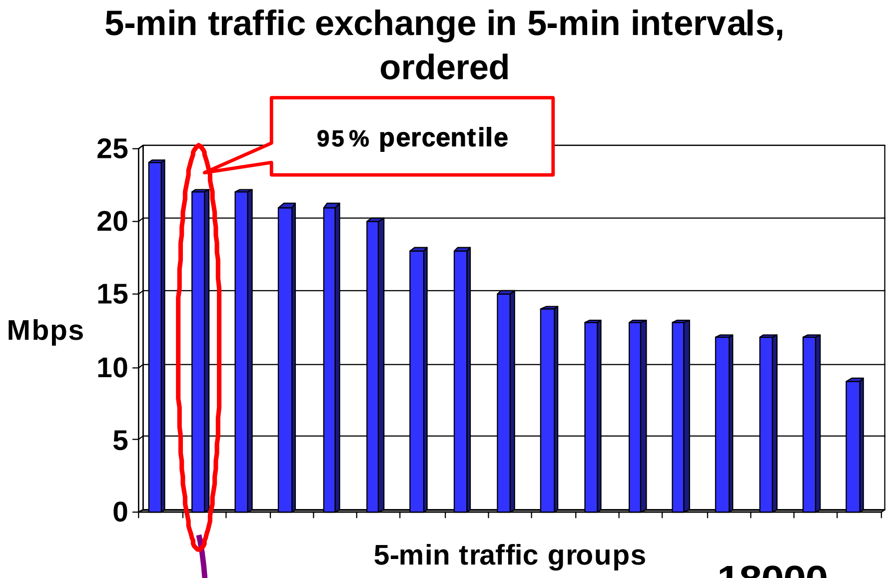
Transit cost: mean of exchanged traffic per 5 minutes

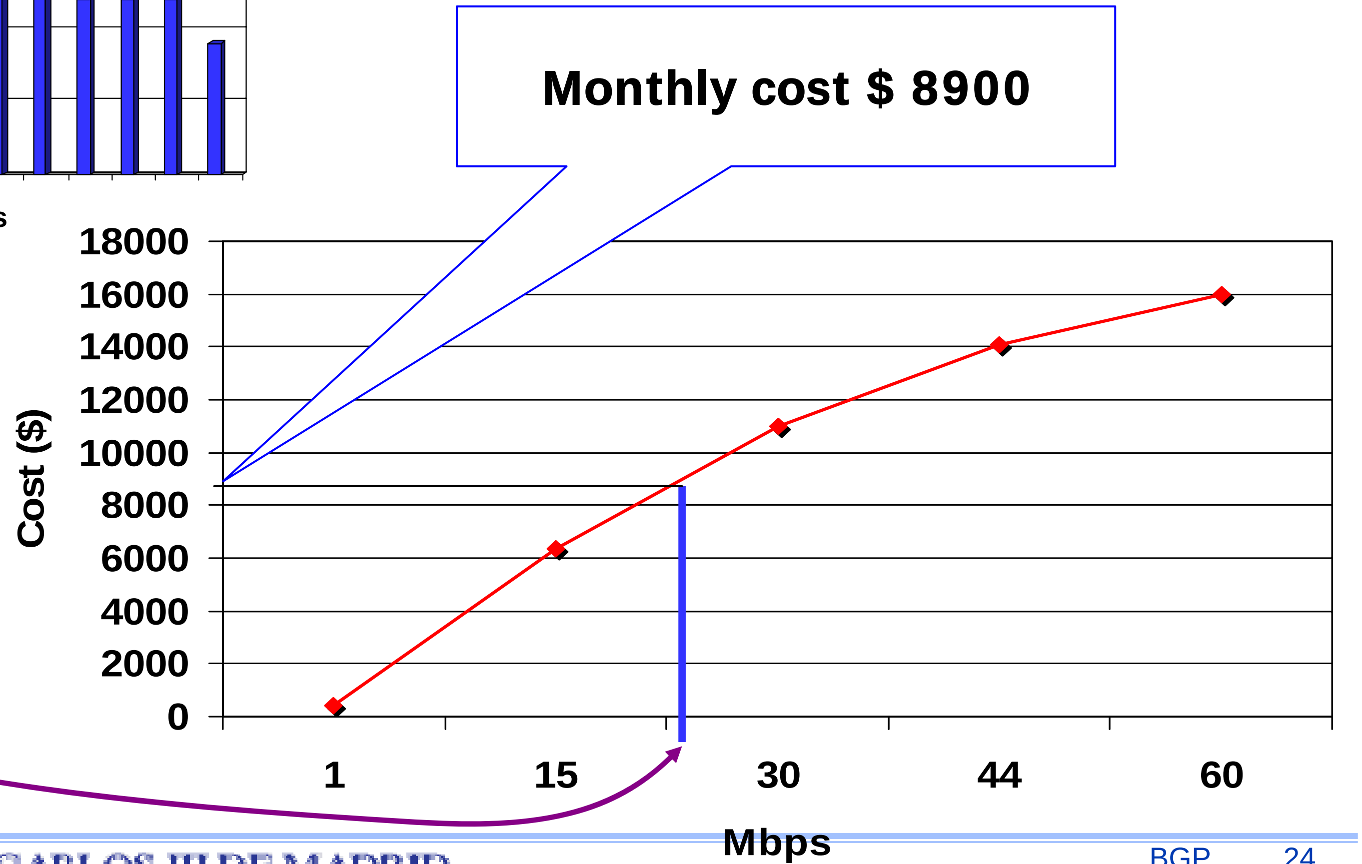


Transit cost: mean of exchanged traffic per 5 minutes

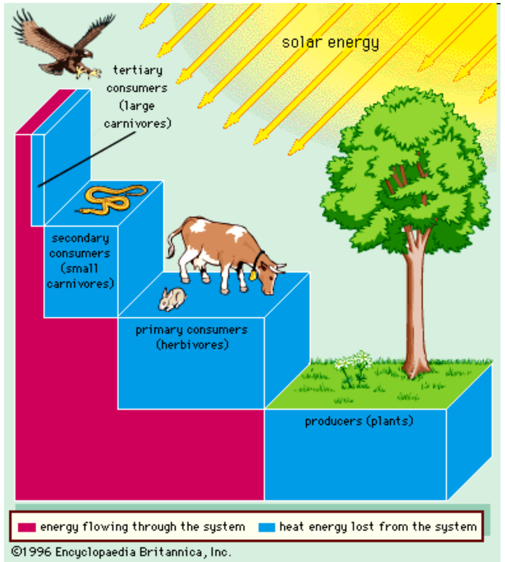


Transit cost

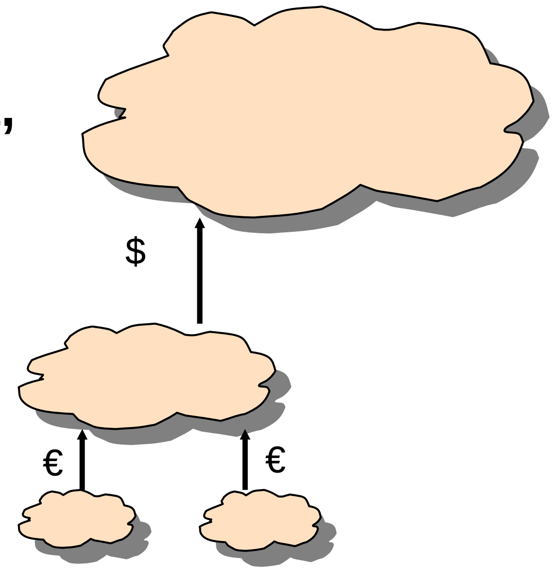




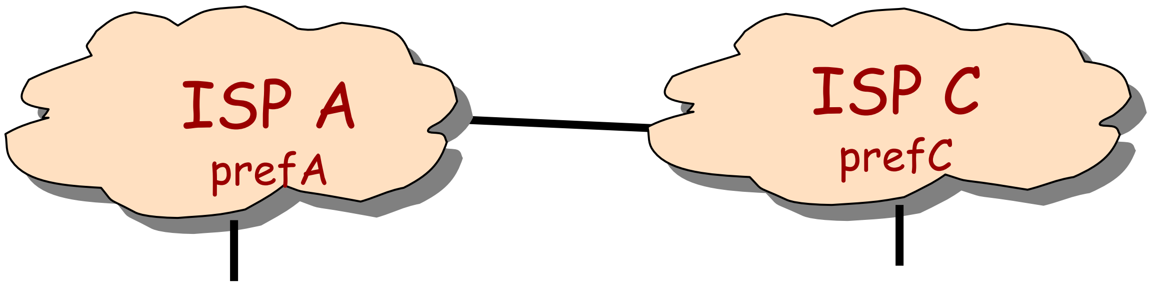
Internet Funding



- **Cost chain**
  - **End user pays its provider for ADSL**
  - **Provider**
    - ✓ **Pays costs of its own infrastructure**
    - ✓ **Gets some benefits… and**
    - ✓ **Pays upper layer provider**
    - **(n times)**
  - **Tier-1 receives payment from its direct client, pays its own infrastructure, and obtains benefits**
    - ✓ **Don’t pay anyone for IP data transit**
- **The internet infrastructure as a whole is paid commercially!**



Reasons for Peering

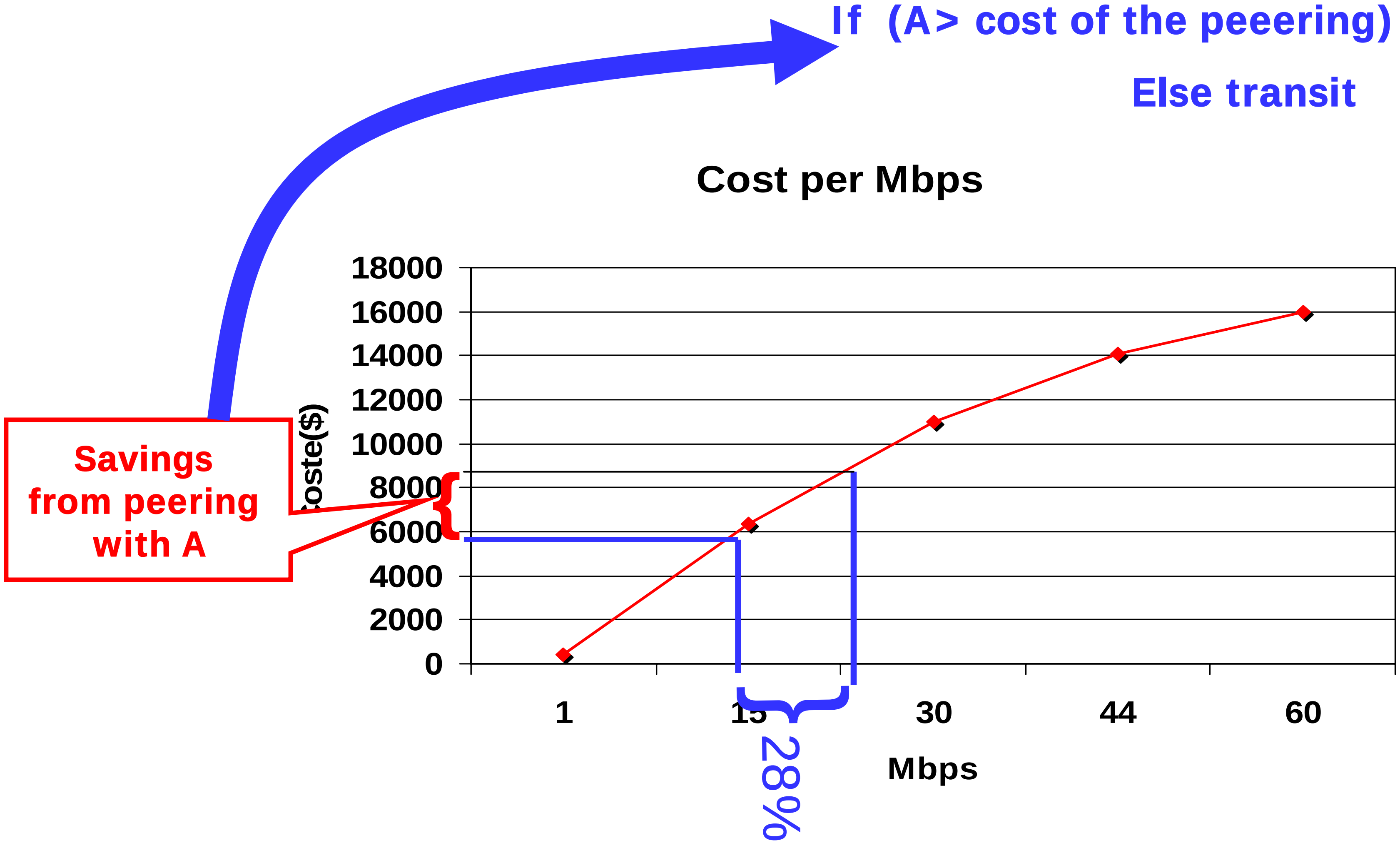


- **(free peering) To reduce costs**
  - **Sending less transit traffic ⇨ paying less to the transit provider**
    - **Peering always has a cost. You don't pay another ISP but:**
      - Need a line between the organisations that peer
        - *Usually pay half each*
      - Cost of management

Peering or transit?

If (A> cost of the peeering) then peering

Else transit



BGP commercial strategy

- Try to establish peering relationships with as many ASs
  - To which you exchange large amounts of traffic
  - AND with low cost to connect to with a fiber
    - Don’t try to peer with an AS in New Zealand
  - AND will never be your clients
    - Let them pay you, … or pay any other
    - Try with those that could be your providers, just in case, although…
- For the rest of the communications, use providers

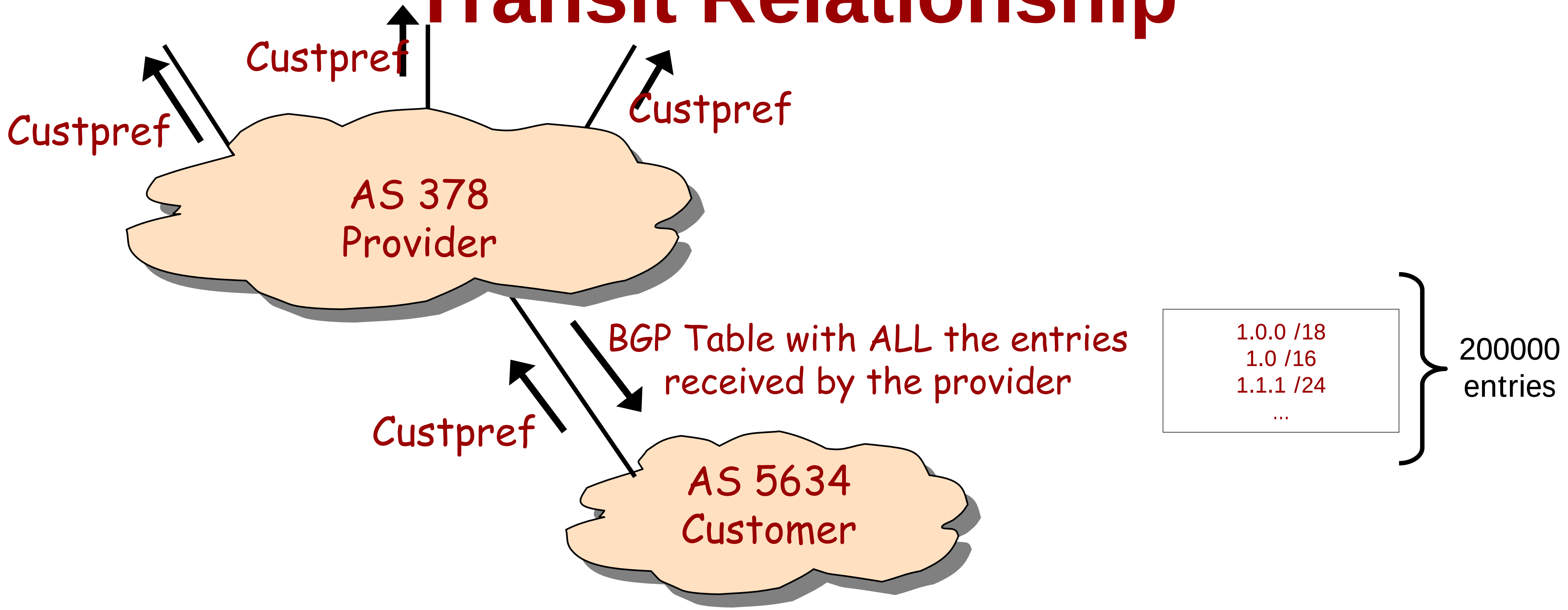
Defining peering and transit

- **Peering and transit are defined by two behaviors:**
  - **Which routes are preferred (depending on the roles of the neighbors generating the routes)**
  - **Which routes are propagated to a neighbor (depending on the roles of the neighbors, and the neighbors generating the routes)**
- **For both considerations, the objective is to take the decision that REDUCE COSTS most**

Route selection criteria

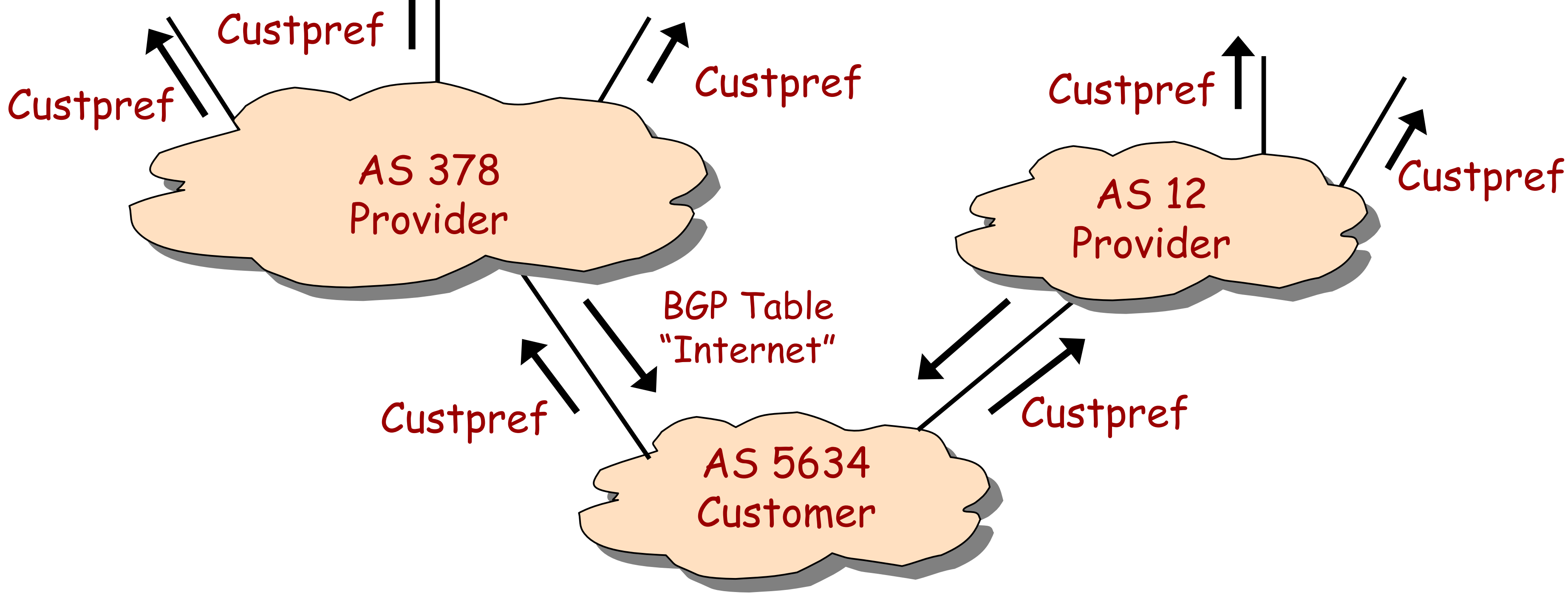
- **This is the preference order for selecting a given route**
  - **Prefer always routes to clients (i.e. send data preferably to clients)**
    - ✓ **When you send traffic through that link, you obtain PROFIT**
  - **If not, prefer routes to peers**
    - ✓ **When you send traffic through that link, you do at LOW COST, through short path to destination…**
  - **Else, transit**
    - ✓ **HIGH COST**
- **Among the same ‘class’, do whatever you want**
  - **The same preference for all, different preference levels…**

Route propagation behavior for the Transit Relationship



- **Transit Connection (Provider / customer)**
  - **Customer propagates its prefixes (and the prefixes of its customer). The provider propagates his customer's prefixes to the outside**
  - **Provider sends to the customer all the prefixes that he knows**
- **The customer pays the provider, according to the quantity of traffic crossing between them**

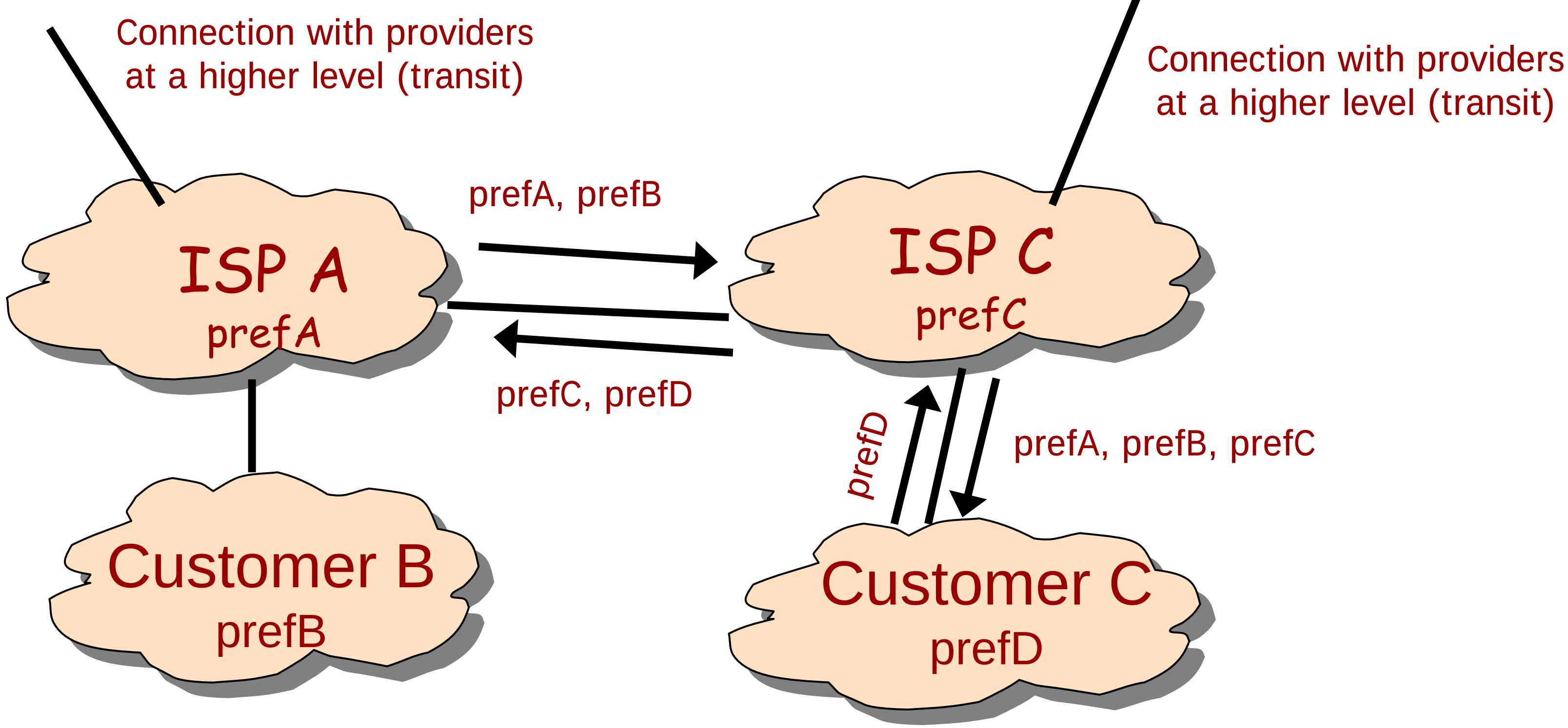
Route propagation for a Multihoming Scenario



- **Multihoming: the customer is contracting more than one provider to obtain fault tolerance, load sharing etc.**
- **The customer doesn't announce to the new provider any prefix external to him**
  - **If he did, the provider AS12 could get into reachability problems!**
- **There are commercial reasons for not sending all routes to everyone**

Route propagation behavior for the Peering relationship

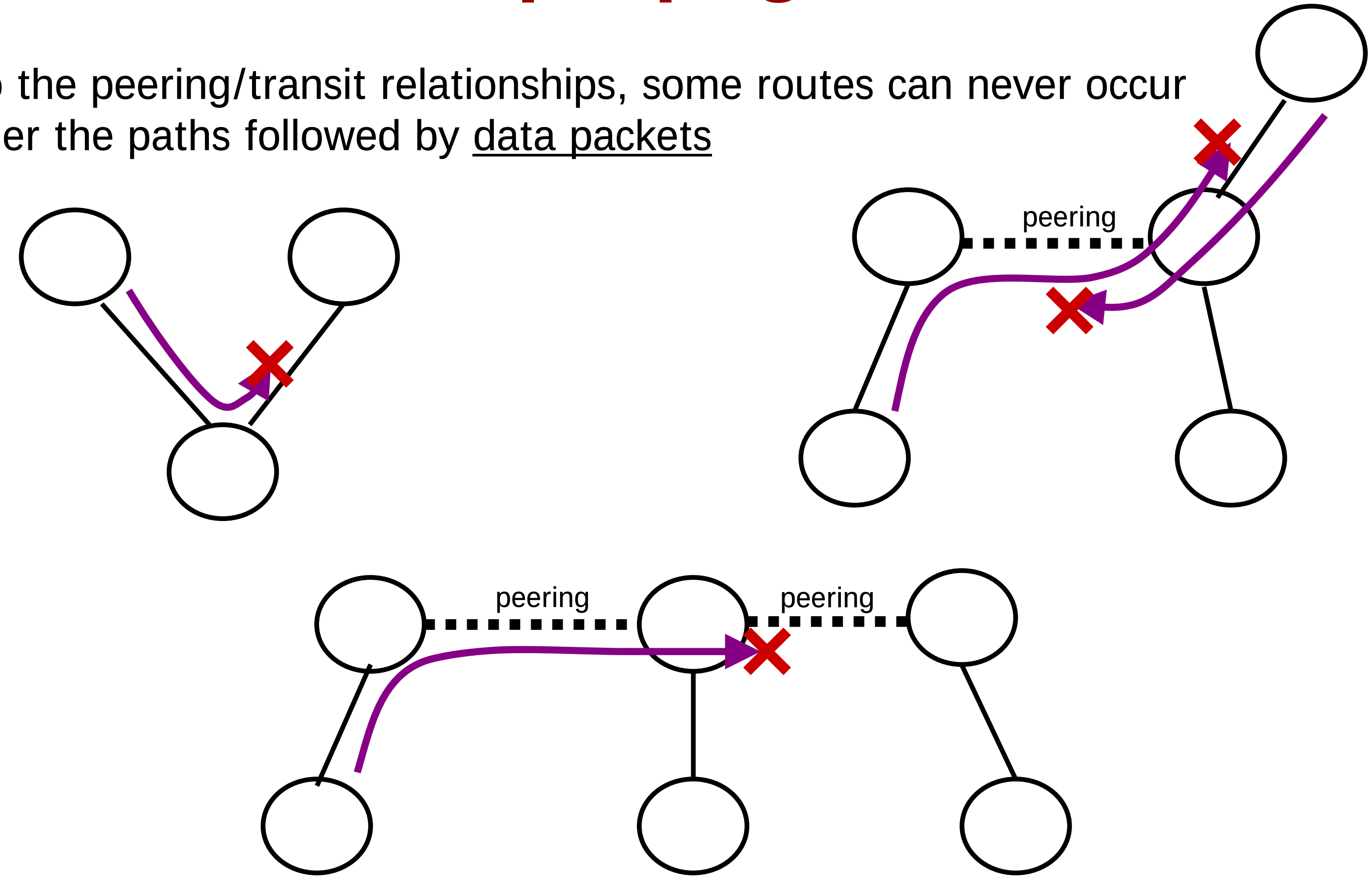
- ***Peering*: Relationship by which a provider offers to another provider's customers connectivity to his own customers**
  - **He doesn't offer general connectivity to the Internet**



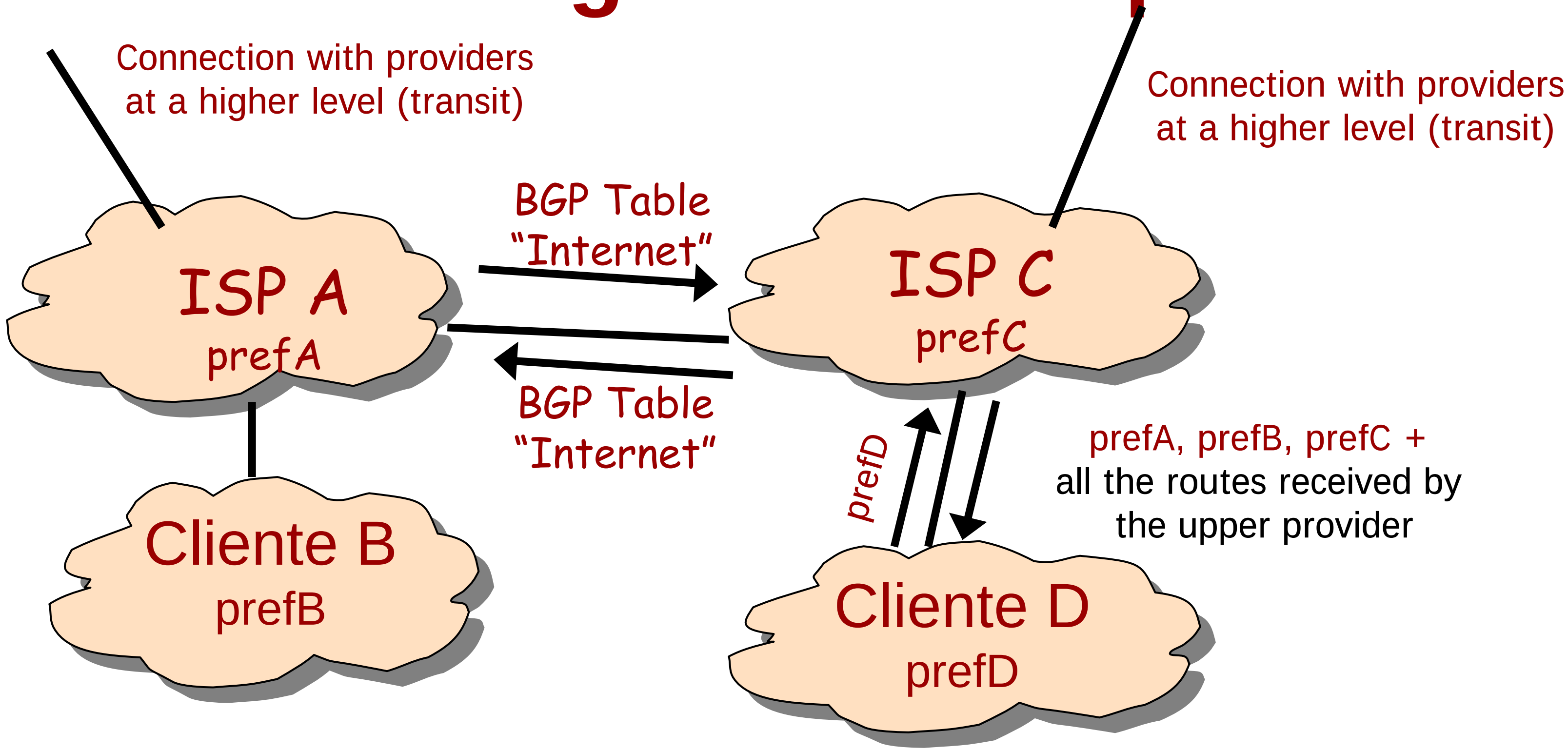
  - **Note: the term "peer" is overloaded – two routers that exchange BGP routes are "peers BGP"**

Paths resulting from restrictions in route propagation

Due to the peering/transit relationships, some routes can never occur
Consider the paths followed by <u>data packets</u>



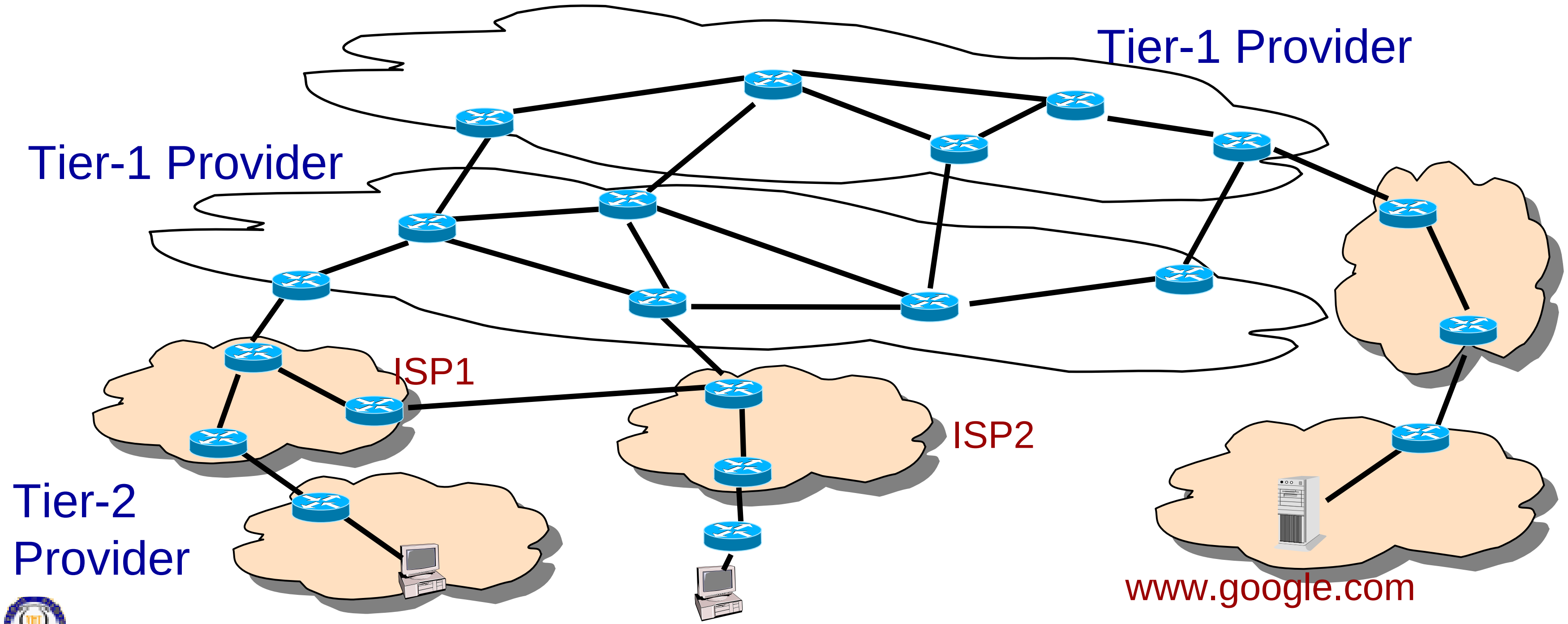
Route propagation behavior for the Sibling relationship



- **When become SIBLINGs: ISPA and ISPC belong to the same organization (although may have different policies that justify different ASs, or coordinated networks (such as Research & Academic European Networks)…**
- **For the prefixes received as transit backup through the link, assign lower preference for the sibling link**

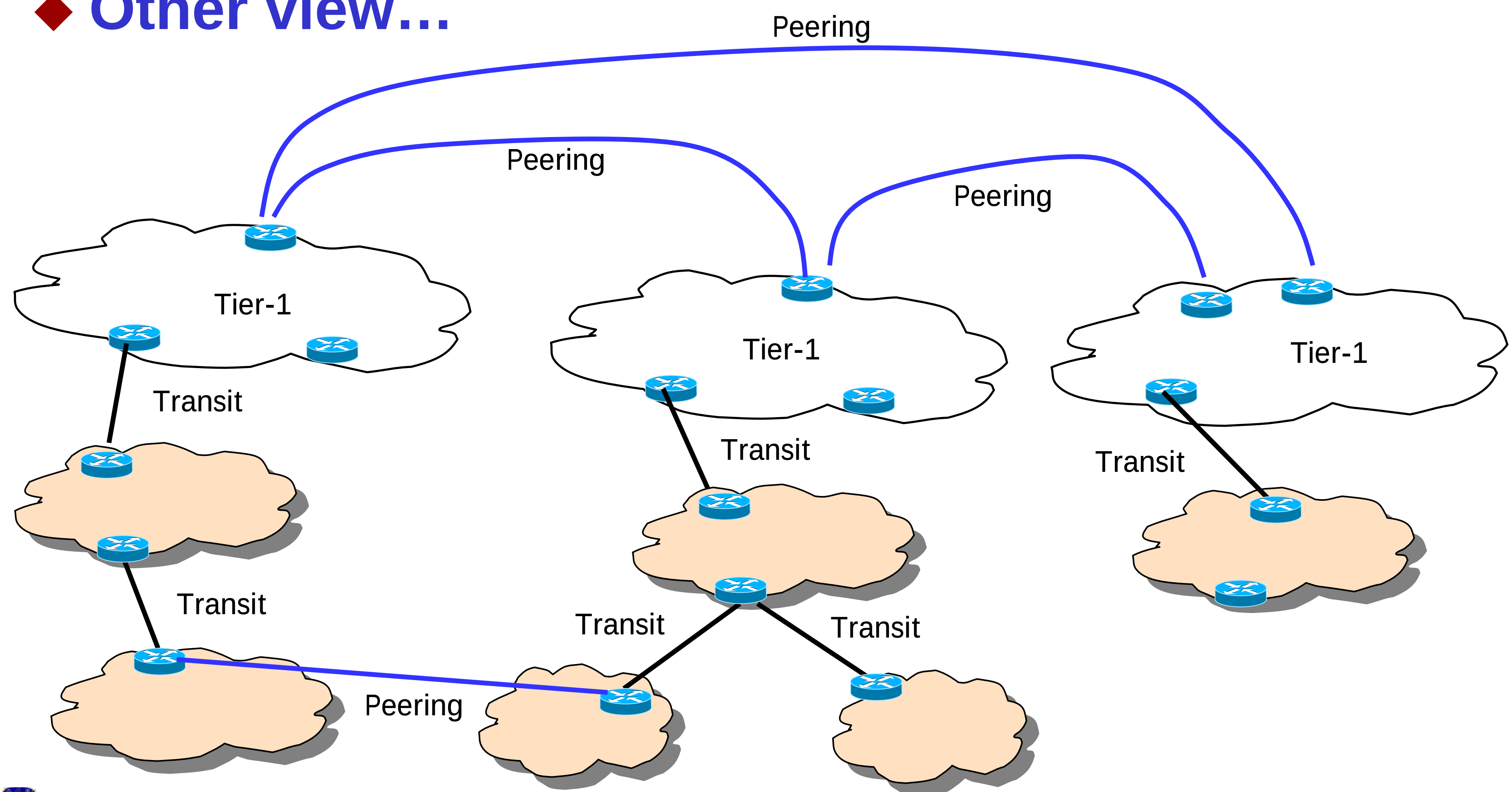
Transit levels

- *Tier 1 Providers* (layer-1)
  - By definition, don't pay
    - They don't have transit providers and they don't pay for peering
    - Definition is based on a commercial relationship: difficult to know from the outside
  - Exchange traffic using free peerings with other Tier 1
- *Tier 2* Providers (layer-2): they pay (either the have transit providers, or pay for peering)
- *Leaf AS* (or *stubs*): AS that are not providers for anyone (just customers)



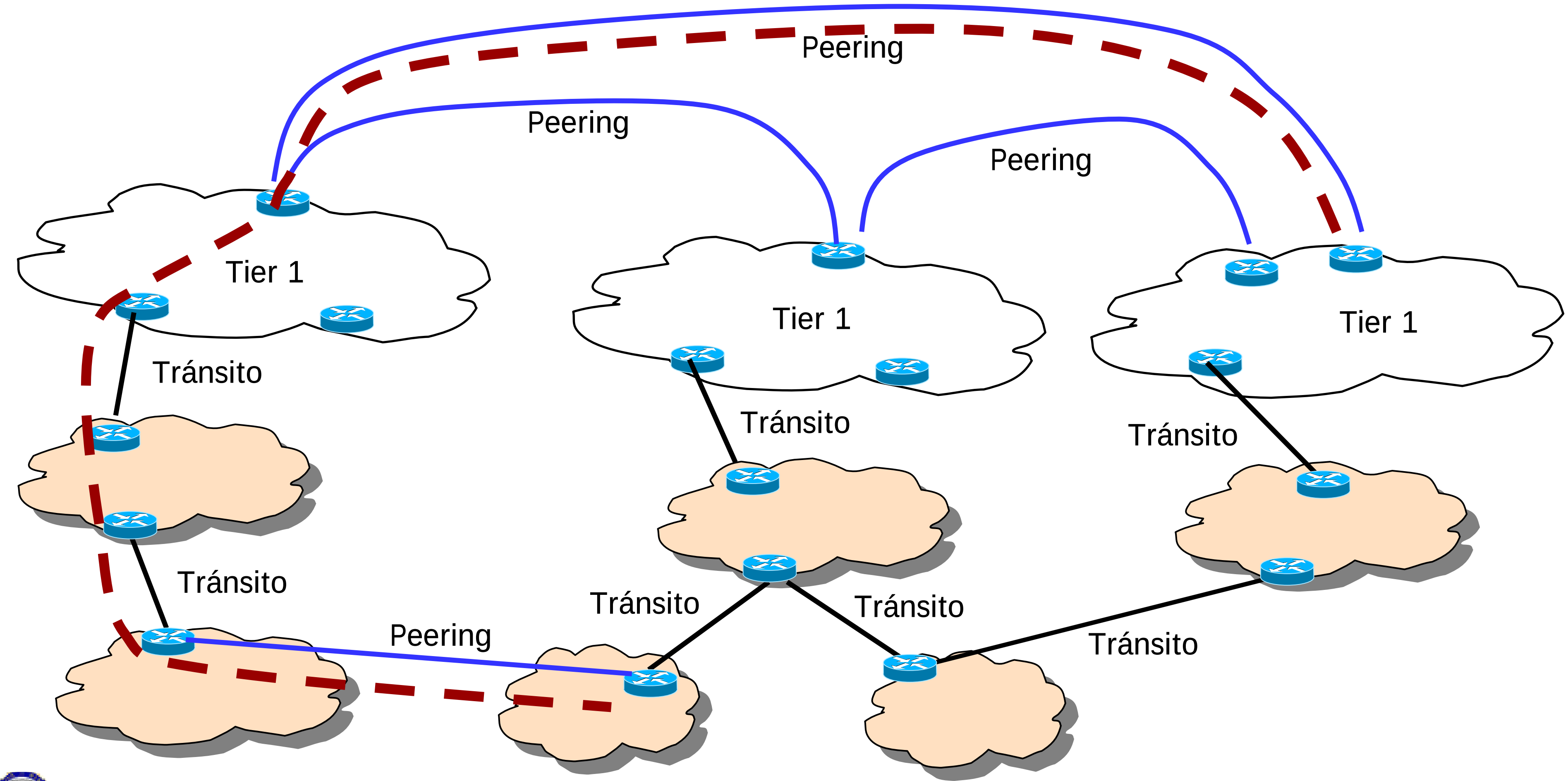
Transit Levels

- Other view…



Transit levels

- Remember that not all paths are possible!



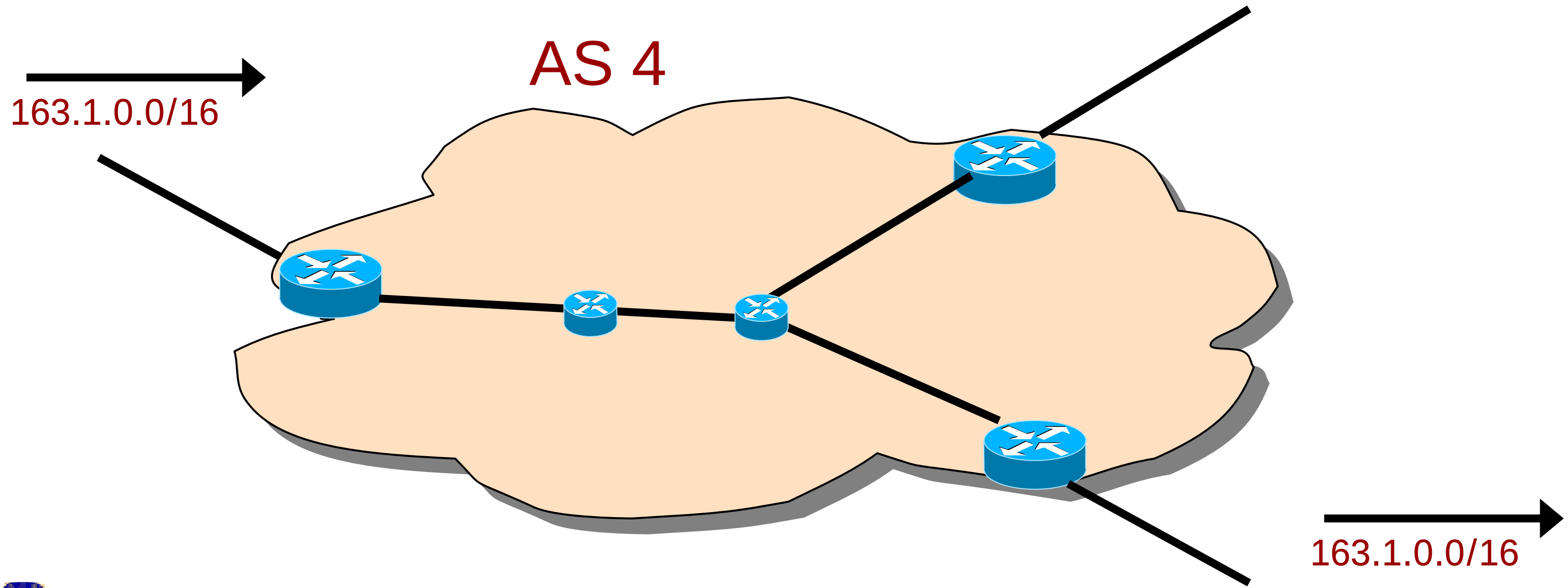
Tier-1 Providers

- **Tier-1 providers**
  - **AT&T**
  - **Global Crossing,**
  - **Level 3 Communications**
  - **NTT Verio**
  - **QWEST**
  - **Sprint**
  - **Verizon**
- **Networks that do not have transit providers, but it is believed that they pay for peering**
  - **Above.Net, Cogent, TeliaSonera, Teleglobe, XO Communications**

READ: **Internet Inter-Domain Traffic. Craig Labovitz, Scott Iekel-Johnson, Danny McPherson, Jon Oberheide, Farnam Jahanian. SIGCOMM 2010.** **(available in Reading list of the course)**

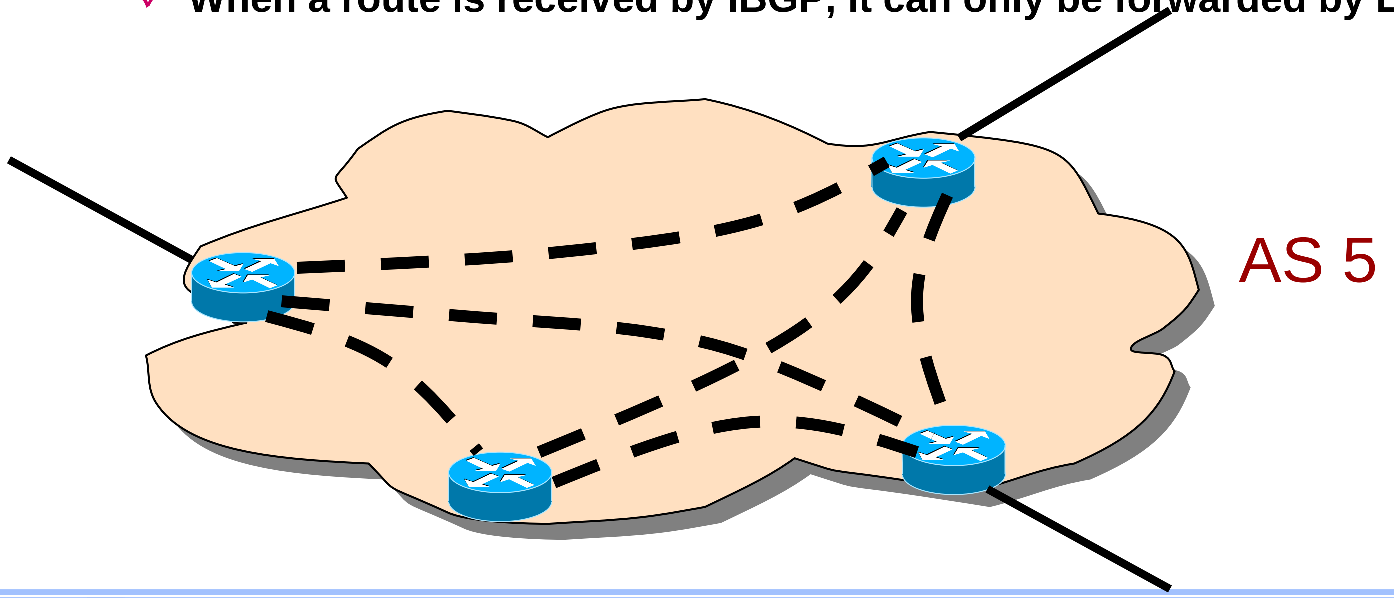
EBGP, IBGP

- **We have assumed that an AS has no internal structure… but they do**
  - **MANY routers that talk BGP to the outside**
  - **Perhaps routers that perform internal forwarding**



EBGP, IBGP

- **To pass BGP information across a domain use IBGP**
- **Build a virtual topology in the interior**
- **In order to avoid loops when forwarding information, the list of ASs does not help (all are inside the same AS)**
  - **Solution:**
    - ✓ **Ensure that all BGP routers belonging to the same domain must be have a BGP session with each other (EVERYONE with EVERYONE).**
    - ✓ **When a router receives a route by EBGP, it sends it to ALL the internal routers by IBGP (as well as to EBGP neighbors)**
    - ✓ **When a route is received by IBGP, it can only be forwarded by E-BGP**



Traffic Engineering

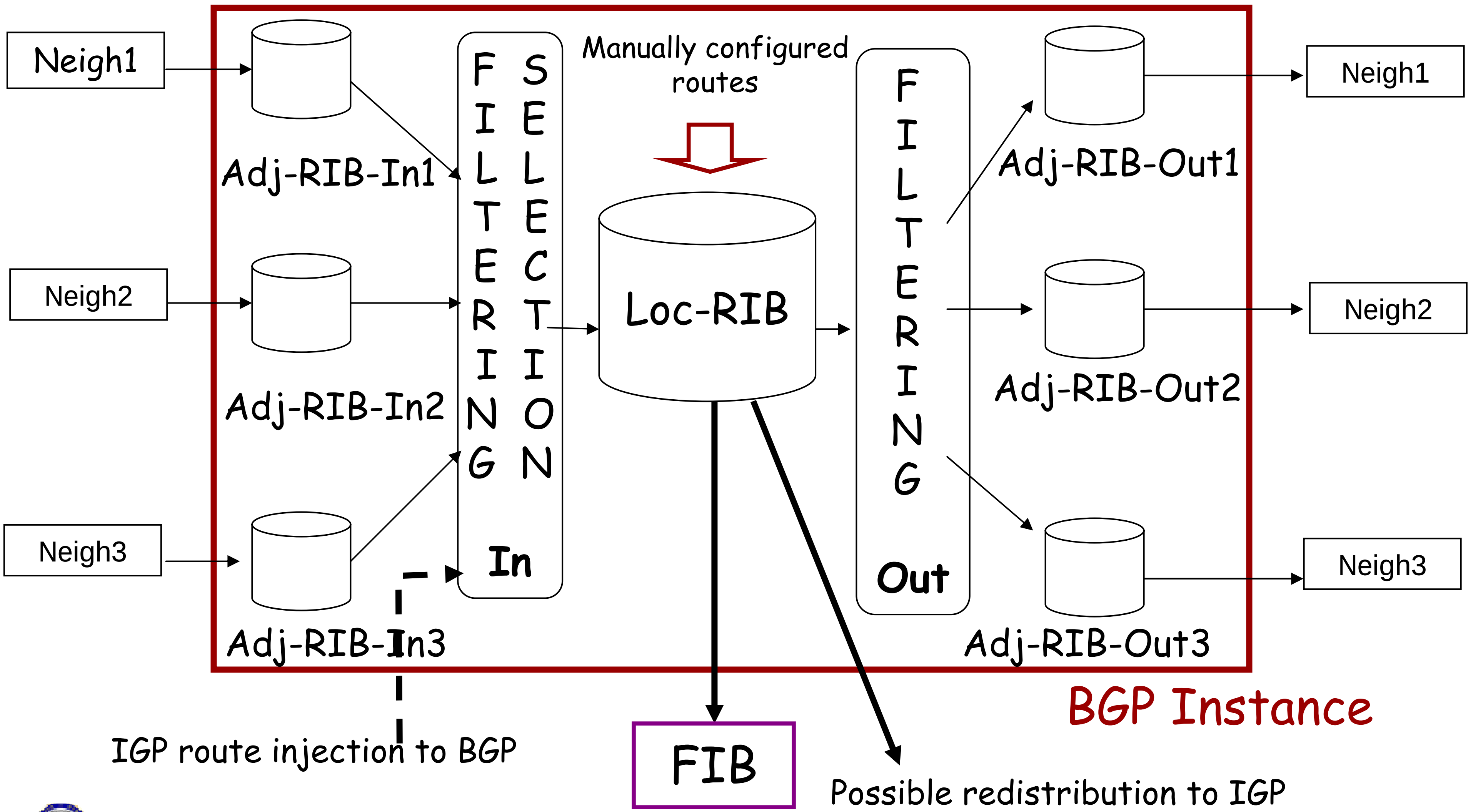
Introduction to Traffic Engineering

- Routing protocols usually provide connectivity and some kind of optimization according to stable metrics (shortest path, minimum cost…)
- *Traffic engineering (TE)*: tailoring of routing to achieve additional goals such as
  - Commercial goals (pay less to providers)
  - Circumvent congested links
  - Achieve minimum end-to-end delays
    - Note that the metrics for the last two are DYNAMIC (i.e. change with time)
- Tools to apply TE (in intradomain)
  - Use routing protocols adapted for TE
    - Integrated with routing, but may generate instabilities (loops either/or oscillations)
    - Example: include *queuing delay* (or other congestion indication) to link metrics and then use Bellman Ford (or link state)
  - Use *management plane* (a central element obtains information from the network, process all the information and configures the routing in the elements)
    - Configuration may change routes directly, or weights (and then routing can still react when failures occur)
    - It is an optimization problem, that can be solved with different heuristics
      - Model of network that allows “what if” estimations
    - Allows network-wide coordination (no loops, no oscillations)

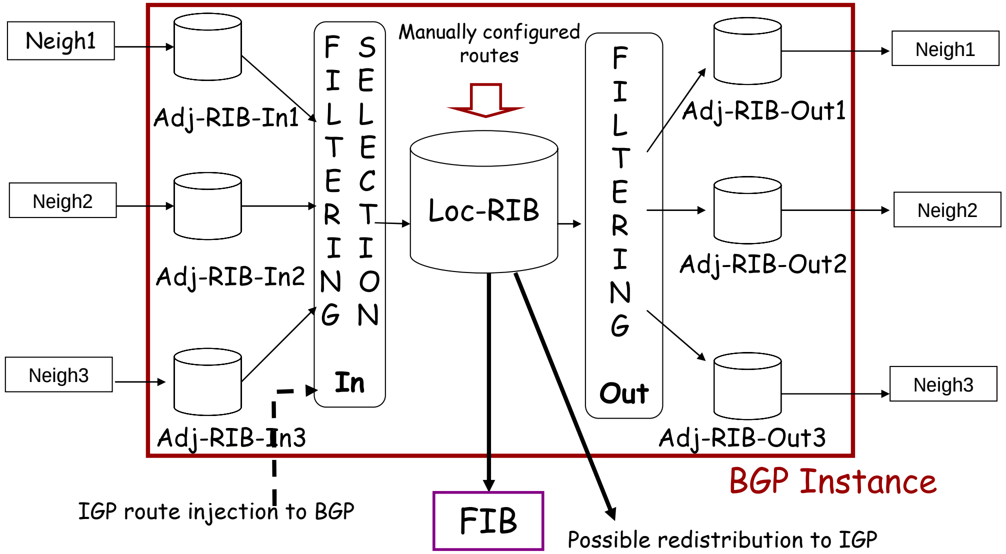
Traffic Engineering in BGP

- **The path that data packets will follow to a given destination depends on**
  - **How BGP announcements are propagated**
  - **How each router selects the path for a given prefix**
    - ✓ **I.e., given several routes, which is selected**

BGP Operation Model



BGP operation model explained

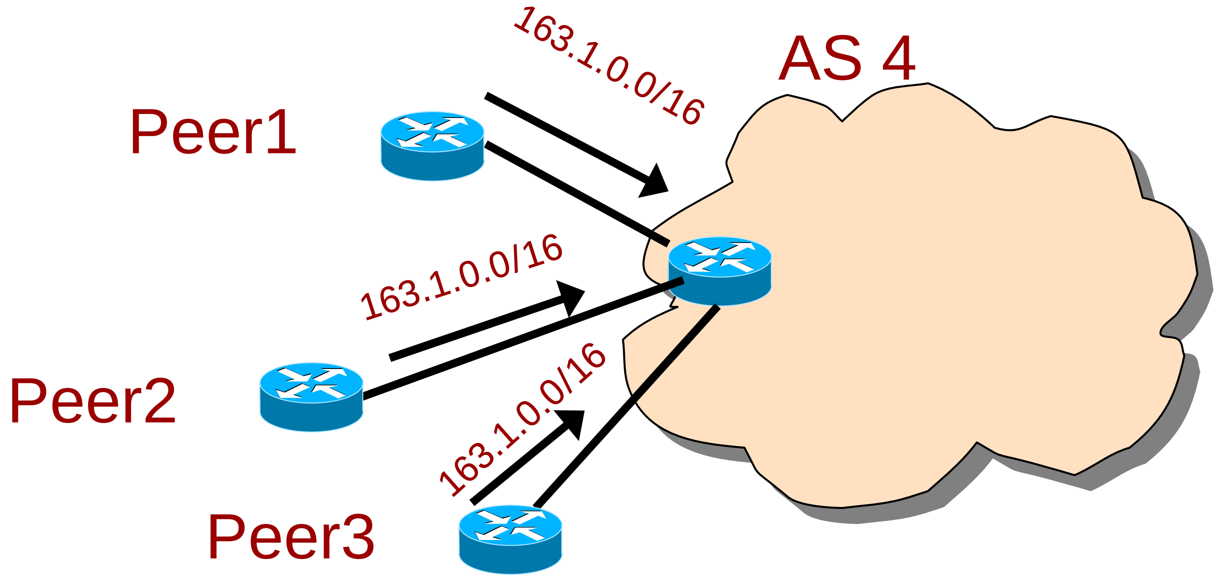


- **Consider a prefix advertised by peer1**
  - The prefix is stored in the "database" Adj-RIB-In1 (adjacent Routing Information Base input from 1)
  - Then, the prefix is filtered (suppose not)
  - When a new prefix arrives, selection process is started again. For that, info for the same prefix in other Adj-RIB-In's is considered
  - Suppose this prefix is preferred. Then, information is stored in the Local Routing Information Base, and then installed in the Forwarding Information Base (=IP forwarding table)
    - ✓ Note that maybe some translation is required – consider NEXT_HOP example
  - Then, the outgoing filter decides to which neighbors it must be advertised
    - ✓ The fact that it has been advertised is stored, to know to which neighbors send future withdraws or changes in the route

Basic Processing of BGP Routes

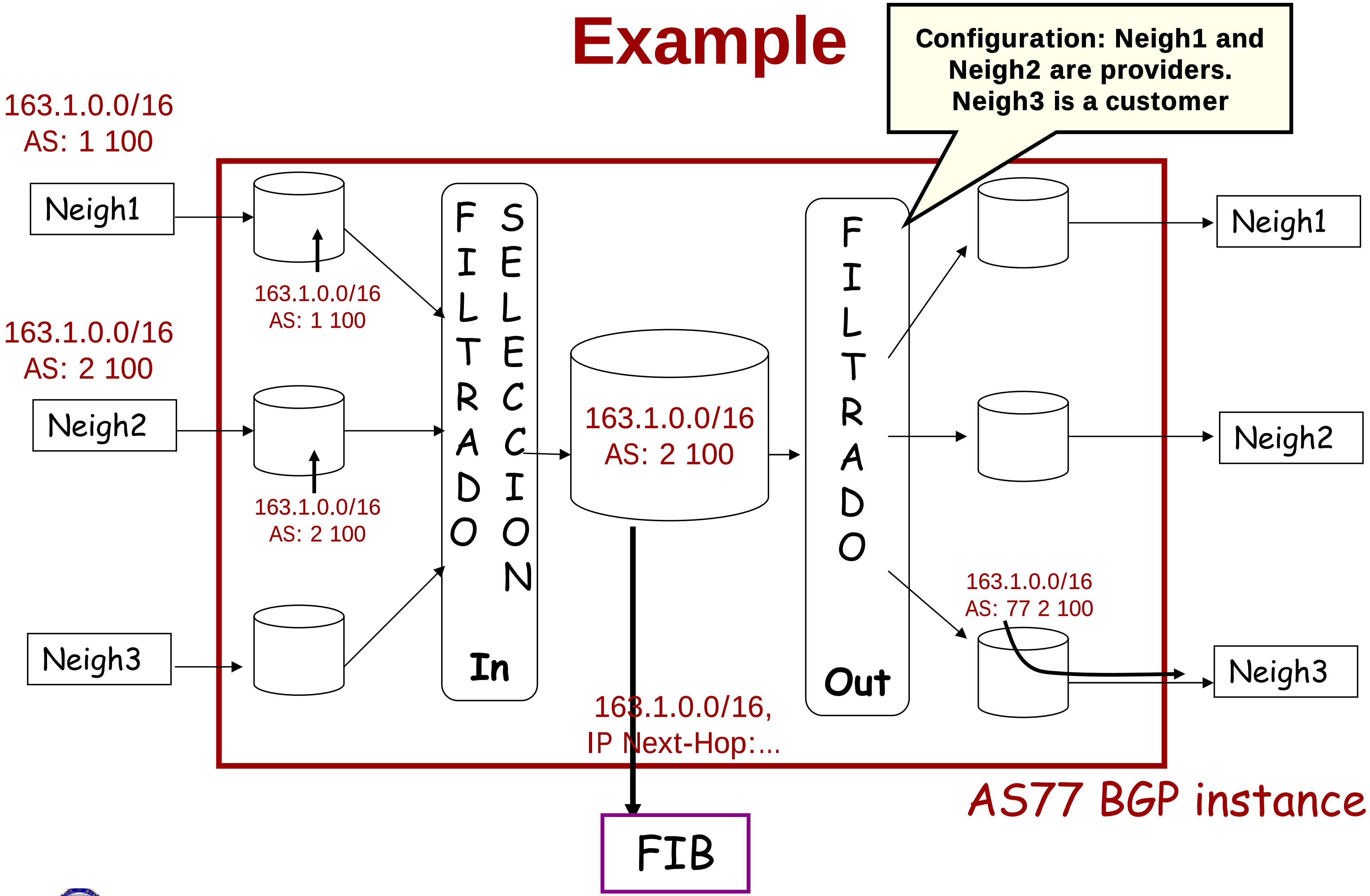
1. **Input selection:**
   - **Filter received routes, delete non-acceptable routes**
     - **Routes with loops (loop detection)**
     - **Unacceptable routes (private addresses, non-allocated addresses)**
     - **Route filtered due to a policy (policy-based filtering)**
       - Prefix
       - AS_PATH
       - COMMUNITY
     - **Unstable routes**
2. **Route-selection algorithm:**
   - **Select the best route to the different routes**
     - **Applying policy**
3. **Output selection:**
   - **Decide which routes to propagate to the peers**
     - **Applying policy**

Selection of routes



- **If a router receives announcements for the same prefix from different neighbours, it must choose one of them as best path**
  - **BGP chooses only one path to reach the destination**
  - **BGP propagates the best path to its neighbours**
  - **BGP stores the non-selected routes to be able to recover them if needed**
- **Criteria for choosing**
  - **The administrator can select any criteria**
    - ✓ **Always choose a particular exit link, always prefer routes that go through a particular AS, …**

Example



Configuration: Neigh1 and Neigh2 are providers. Neigh3 is a customer
163.1.0.0/16
AS: 1 100
Neigh1
163.1.0.0/16
AS: 1 100
163.1.0.0/16
AS: 2 100
Neigh2
163.1.0.0/16
AS: 2 100
Neigh3
FILTRADO
SELECCION
In
163.1.0.0/16
AS: 2 100
163.1.0.0/16,
IP Next-Hop:...
FILTRADO
Out
Neigh1
Neigh2
163.1.0.0/16
AS: 77 2 100
Neigh3
AS77 BGP instance
FIB

Outgoing route filtering and business model

- **We have said before (AS relationships) that business model suggest**
  - **Never carry traffic between two of your providers**
    - ✓ **To do this, don't advertise (filter out) to providers routes received from providers**
  - **Never carry traffic between a peer and a provider (and vice versa)**
    - ✓ **To do this, don't advertise (filter out) to providers routes received from peers, and filter to peers route received from providers.**
  - **Send as much traffic as you can to your clients**
    - ✓ **Don't put any filter (out) involving clients**

BGP Path Attributes

- **Attributes: BGP specific information that travels with a prefix, and can be used to make decisions on filtering or route selection**
  - **Some attributes are: AS_PATH, ORIGIN, NEXT_HOP…**
  - **Example of route selection criteria: Choose the path that traverses fewest ASs (i.e. less number of components in the AS_PATH)**
- **Attributes may change in transit**
  - **It depends on the attribute type**

Route selection in BGP

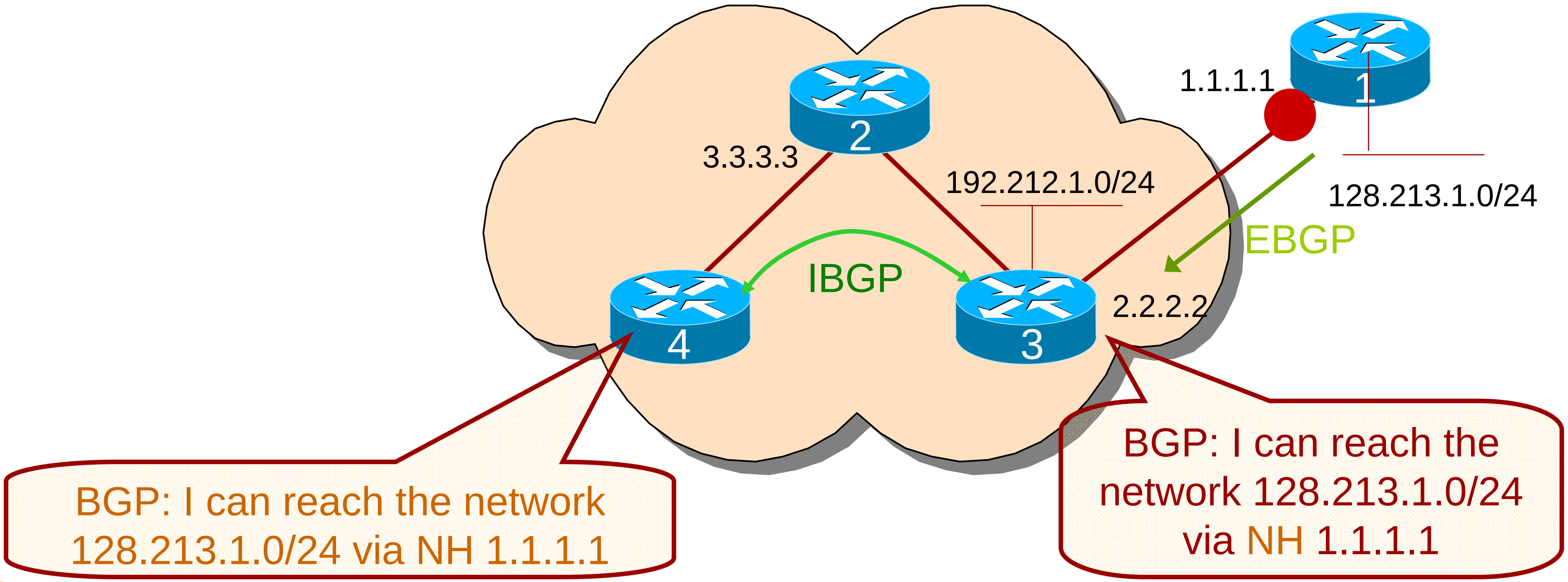
- What do we want with BGP route selection?
- An administrator could find interesting
  - To decide explicitly the route that he wants
    - Maybe due to economic reasons, performance…
  - Select more “robust” routes (if information available)
  - Path traversing the minimum number of ASs
    - This is a reasonable criteria, if we do not have more detailed data related with performance
  - Allow a network informing other about preferences, if there are common links
  - Select routes according to “hot potato” routing
    - Hot potato: send out of a network a packet as fast as possible, because in this way we spend less network resources
  - If two routes are almost equivalent, we should have a criteria to select one

BGP Path Attributes: AS_PATH

- Contains the AS numbers passed on the announced route
  - In so-called path segments (one per AS)
- For each UPDATE message passed along to another AS (EBGP):
  - The AS prepends (=inserts at the beginning) its AS number to the list of path segments
    - ✓ List must remain unchanged if UPDATE passed to a router within the AS (IBGP)
- Sequence of path segments
  - A path segment is described with:
    - ✓ Type (AS_SET or AS_SEQUENCE)
    - ✓ Length of the path segment (# of AS in the path segment)
    - ✓ Values (one or more AS numbers)
- This allows you to:
  - Apply routing policies based on the transit AS
  - Detect loops: if the receiving AS is already contained in the path

BGP Path Attributes: NEXT_HOP

- **NEXT_HOP shows the IP address of the border router that provides access to the announced routes**
  - **In the example, a route generated in the 1 is propagated to 3 (EBGP) and to 4 (IBGP)**
    - ✓ **NH both in 3 and 4 is 1.1.1.1**
  - **If 4 propagates the route outside, it should put the IP of its outgoing interface as NEXT_HOP**



BGP Path Attributes: NEXT_HOP

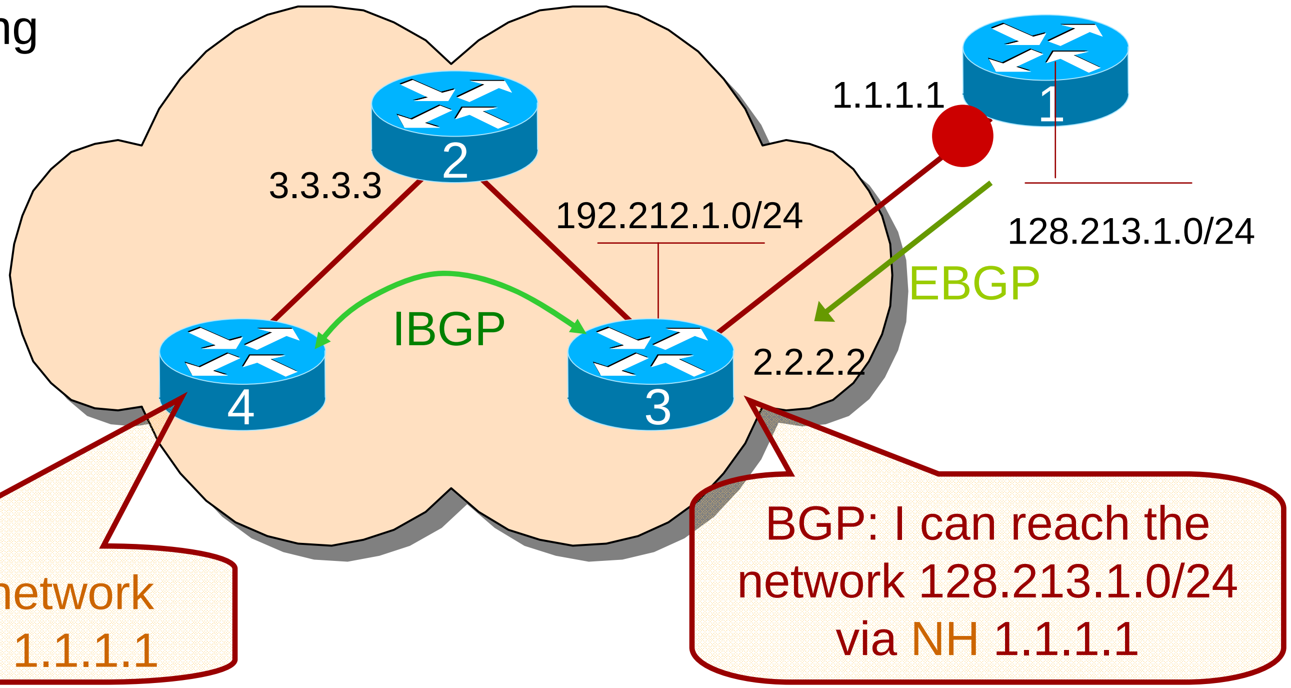
- **The NEXT_HOP info along with the IP routing table is processed to generate a new entry in the IP routing table**
- **An entry for the NEXT_HOP *must* exist in the IP routing table either through IGP or statically)**
  - ✓ **For example: R4 must know (through IGP or static route) how to route to 1.1.1.1**

BGP routing table in 4

DESTINATION	NH
128.213.1.0/24	1.1.1.1

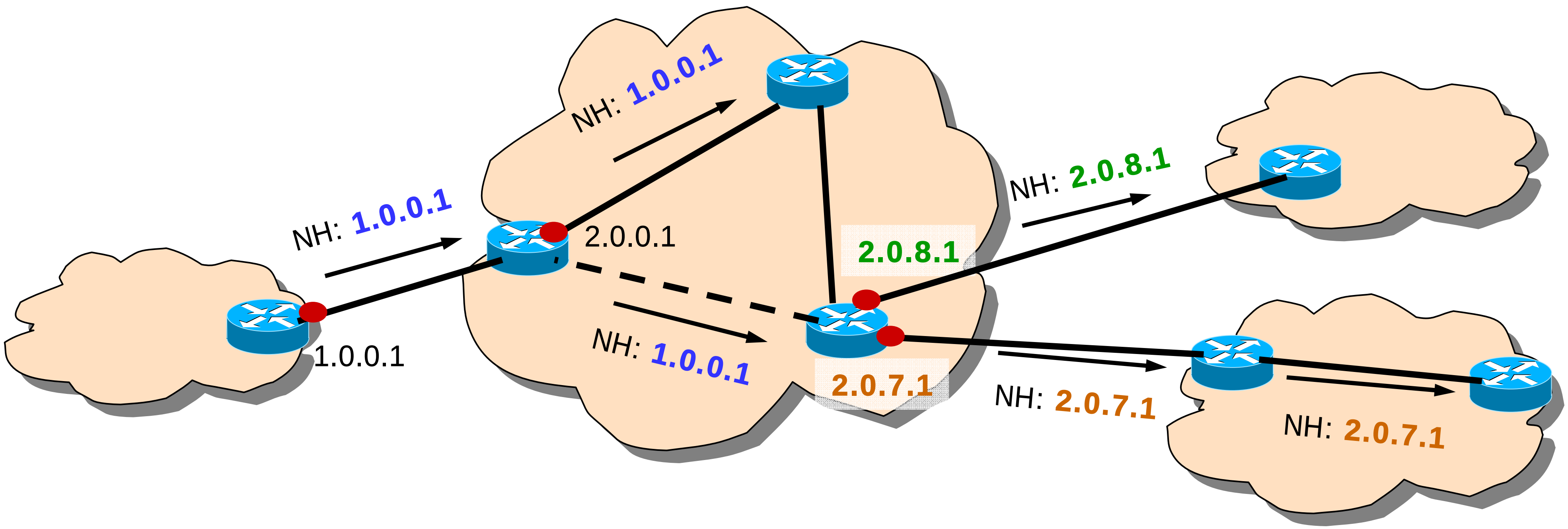
IP routing table in 4

DESTINATION	NH
192.212.1.0/24	3.3.3.3
128.213.1.0/24	3.3.3.3
2.2.2.0/24	3.3.3.3
3.3.3.0/24	-
1.1.1.0/24	3.3.3.3



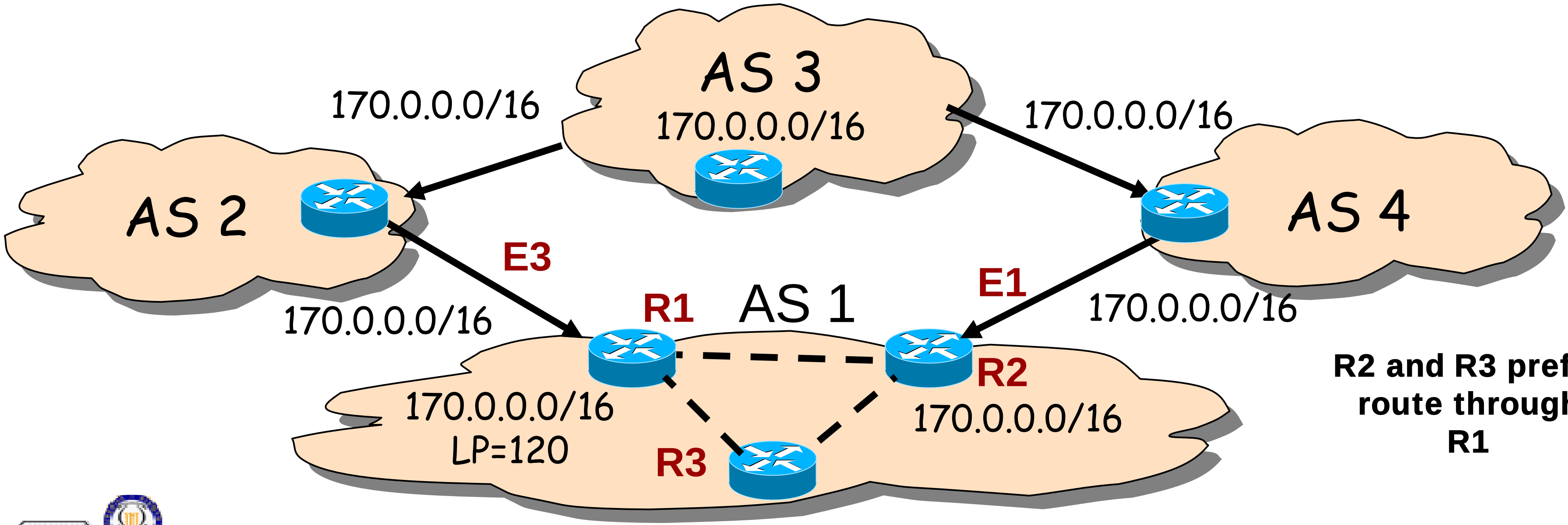
NEXT_HOP

- **Example of NEXT_HOP use**
  - **Route originated in left-most router**



BGP Path Attributes: LOCAL_PREF

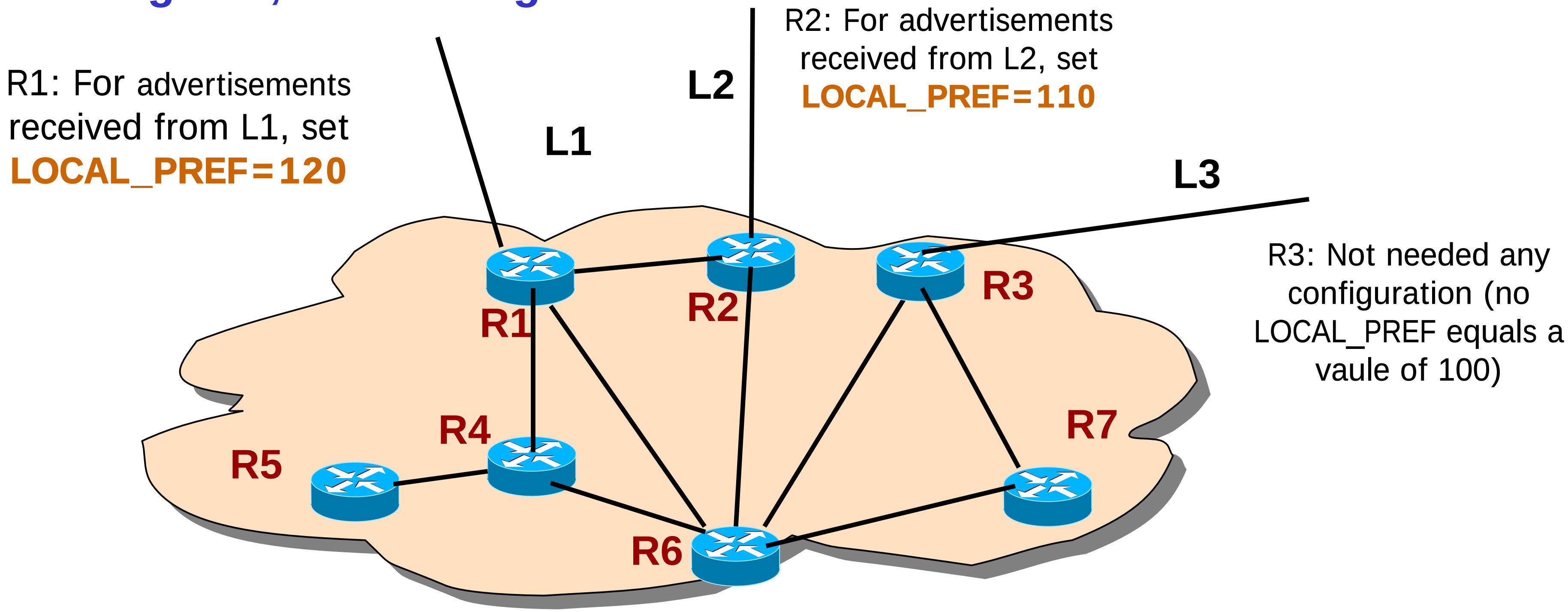
- **Aim: allow the propagation of link preference for some external prefix inside an AS**
  - **It is configured in a single router, and it is propagated through IBGP to all internal peers**
  - **Prefer routes with the highest local preference**
  - **Default value of 100 (i.e. if it is not explicitly set, it is equivalent to 100)**
  - **Note: it is only used inside a given AS (it is only transmitted by IBGP)**



R2 and R3 prefer route through R1

Example of use of LOCAL_PREF

- **Desired policy: Prefer always outgoing path through L1, if not through L2, else through L3**



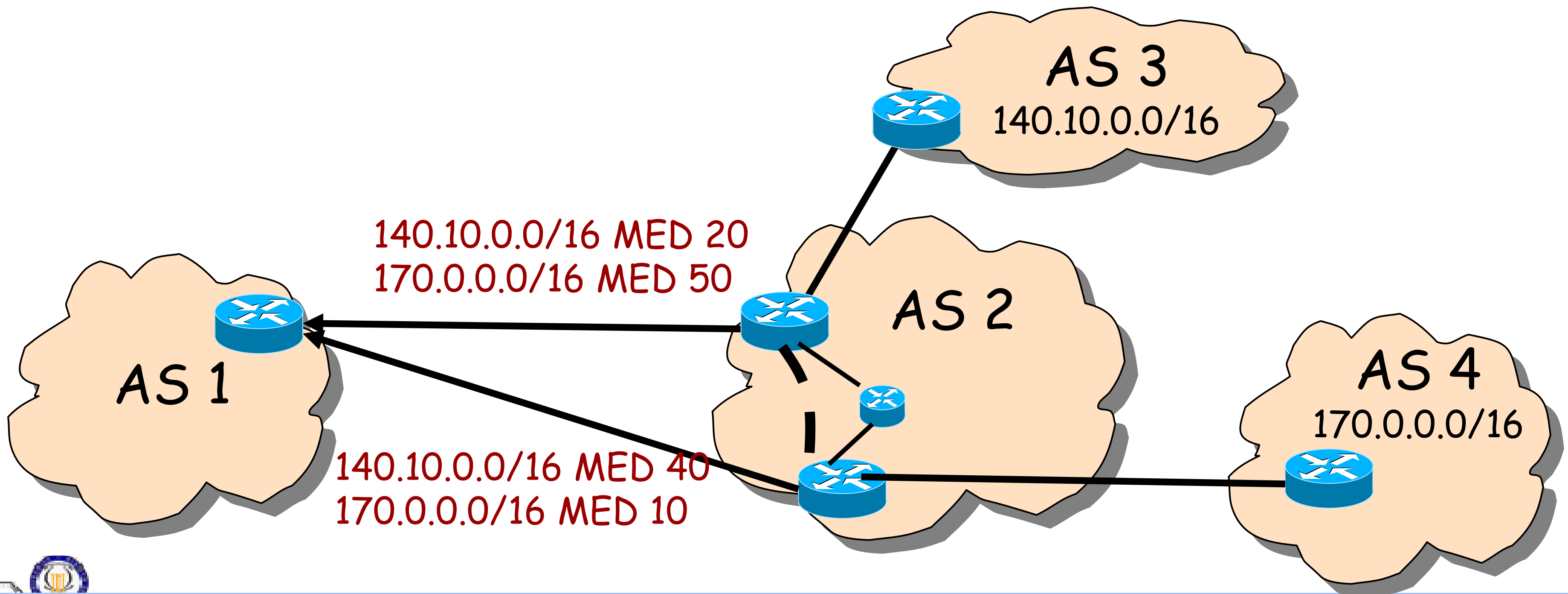
- **Think: if we do not use LOCAL_PREF, how many configurations should be required to apply the same policy?**

LOCAL_PREF to enforce business model

- It has been stated before (AS relationships) that the usual business model suggest the following route preference
  - Prefer always routes to clients
    - ✓ When you send traffic through that link, you obtain PROFIT
  - If not, prefer routes to peers
    - ✓ When you send traffic through that link, you do at LOW COST, through short path to destination…
  - Else, transit
    - ✓ HIGH COST
- This behavior is enforced by proper configuration of LOCAL_PREF

MULTI-EXIT DISCRIMINATOR (MED)

- Allows an AS to suggest to its neighbours a preferred connection (when multiple exist) for a given route
  - Distance metric: Always prefer the lower value
  - In principle it discriminates between routes with equal AS_PATH values
  - The metric is local between the two ASs, it is not propagated further

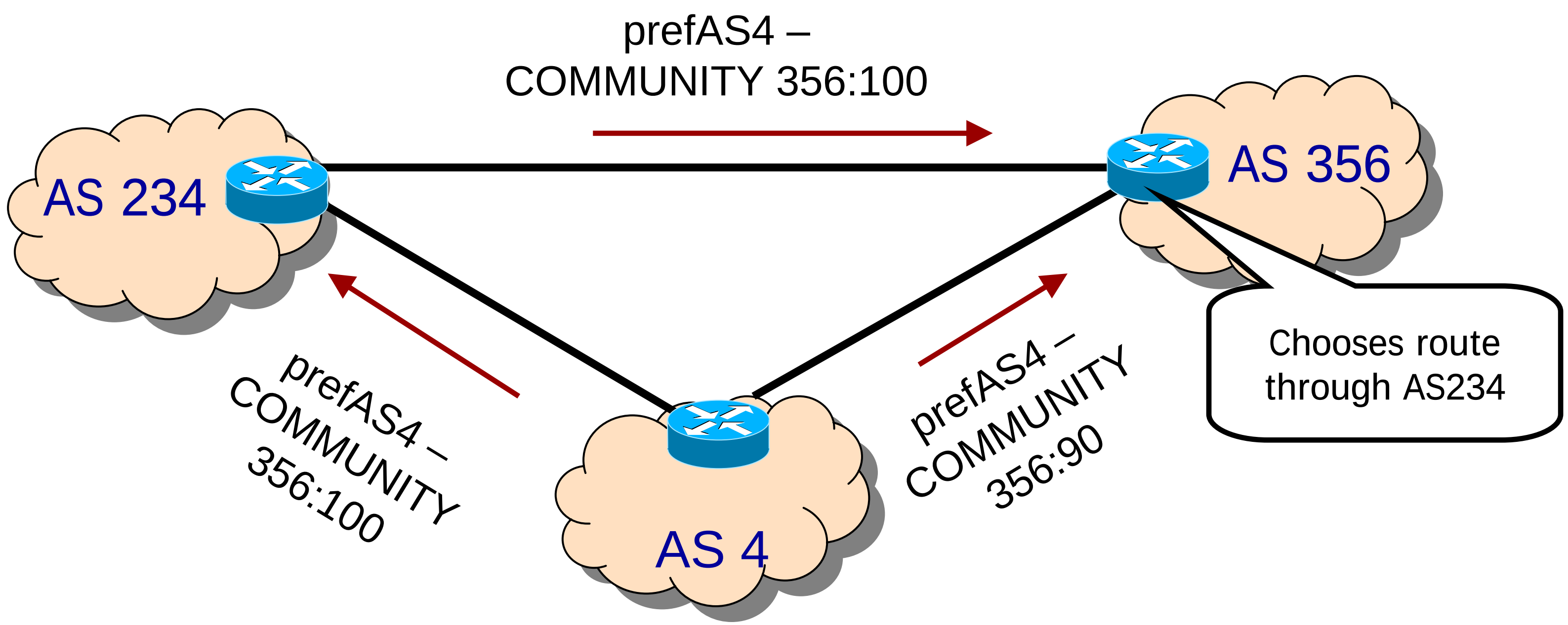


BGP Path Attributes: COMMUNITY

- **COMMUNITY value:**
  - **Group of destinations sharing common properties**
  - **32 bit number acting as a tag to qualify a route**
- **Alleviates managing route distribution**
  - **See following example**
- **The COMMUNITY attribute is a list of COMMUNITY values**
  - **An advertisement can be associated to multiple COMMUNITY values**

Example of COMMUNITY

- **AS4 wants to be able to configure how does it receive traffic from AS356 (sometime directly, other through AS234)**
  - **AS356 is willing to cooperate**
- **AS4 and AS356 agree in a particular COMMUNITY use: if AS4 generates COMMUNITY 356:lp, AS356 configures LOCAL_PREF lp for the received route**
- **Example below: AS4 prefers using route through AS234**

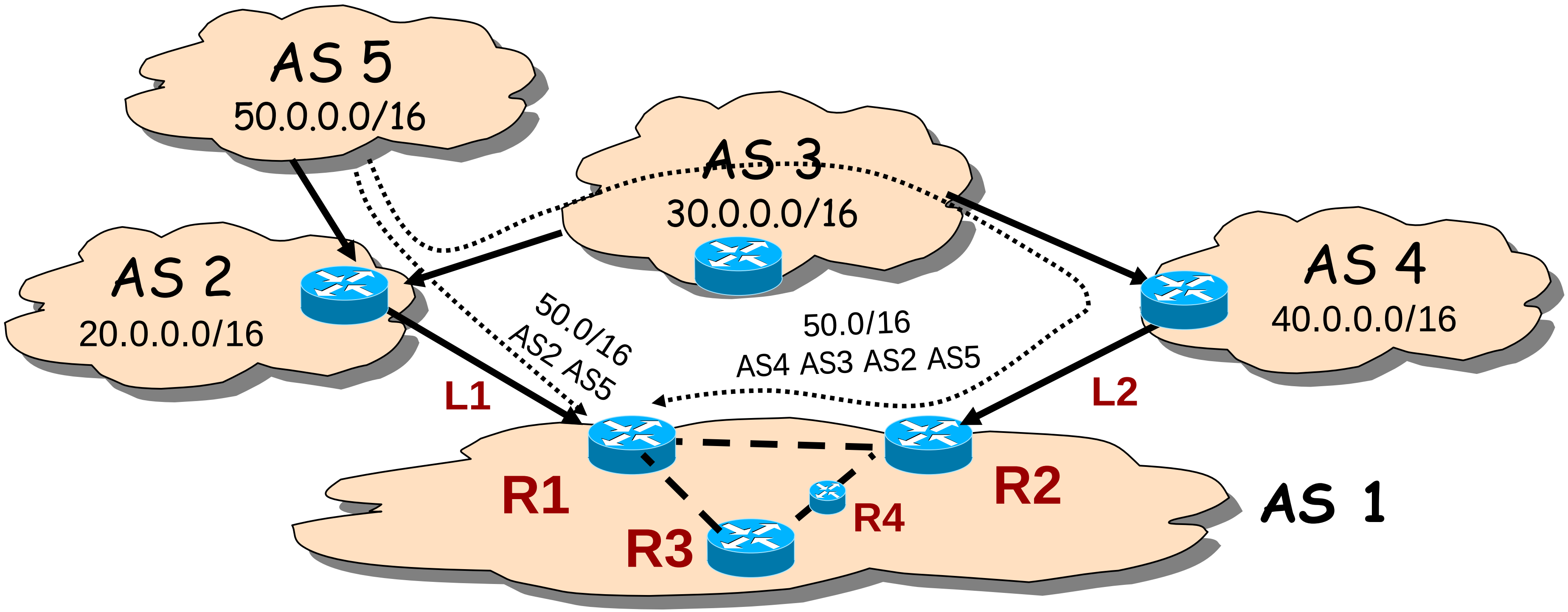


BGP: Route Selection Rules, Tie breaking

- **SAME PREFIX:** As the list is browsed routes that do not tie in the best value in each of the criteria are deleted:

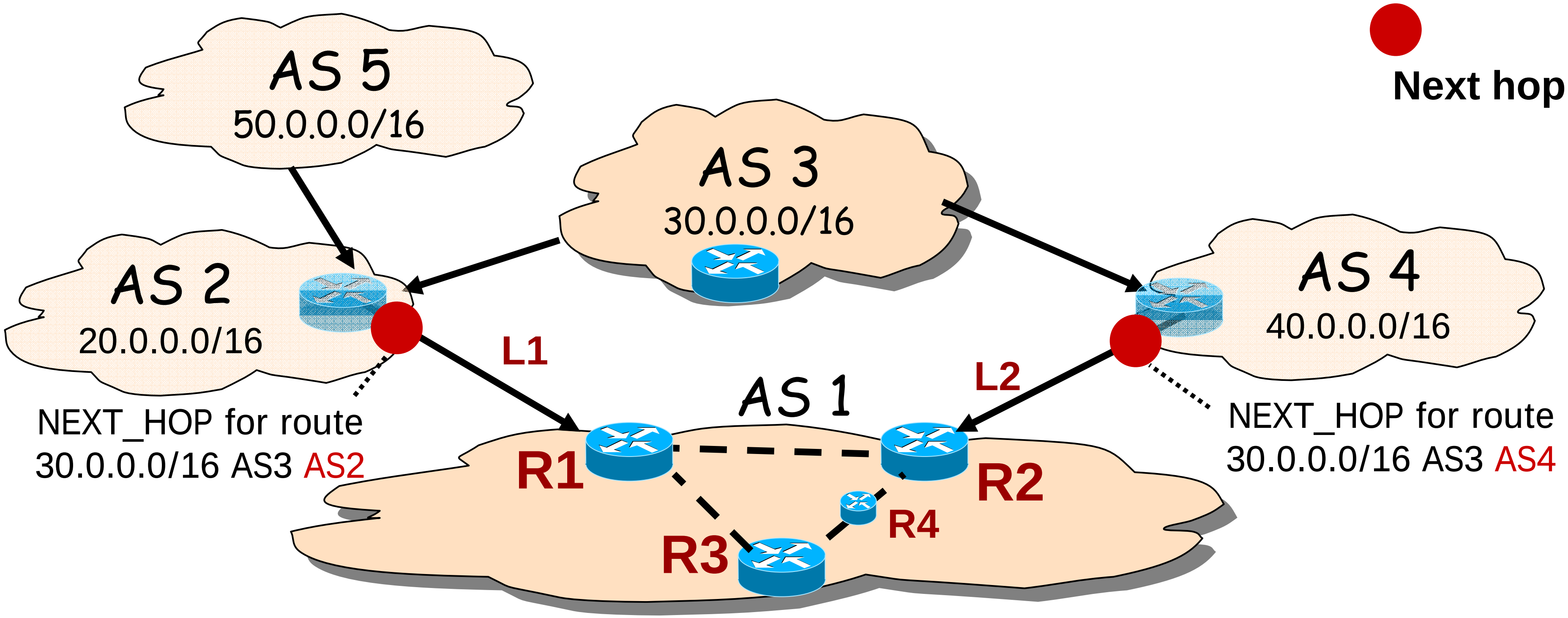
1. If NEXT_HOP is not available (there is no route in the IP forwarding table), ignore the route.
2. Delete routes with lower LOCAL_PREF.
   - Specific rules are used, which correspond with internal politics (prefer route that crosses through AS_X, by a link...) in order to generate LOCAL_PREF.
   - It was generated by the administrator, therefore it is very trustworthy.
3. Delete routes with longest AS_PATH (larger amount of AS to transit)
   - Very much applied.
4. Delete routes with higher ORIGIN. (Freshness)
5. Delete routes (coming from the same AS) with higher MED
   - If two routes come from the same AS, it is probable that they will have the same AS_PATH, on the contrary rule 3 would not have been applied.
6. Delete routes that were learnt by IBGP, if there are routes learnt by EBGP.
   - Hot potato: sending traffic to the exterior if it is possible.
7. Delete routes to NEXT_HOP with higher costs.
   - Note that only considers AS own metric
   - Hot potato: send traffic to the faster way to exterior.
8. Prefer routes announced by router with lower BGP identifier.
   - BGP identifier is in the OPEN message of the protocol.
9. Prefer route received from the interface with lower address to the neighbor.

Route Selection: Example 1



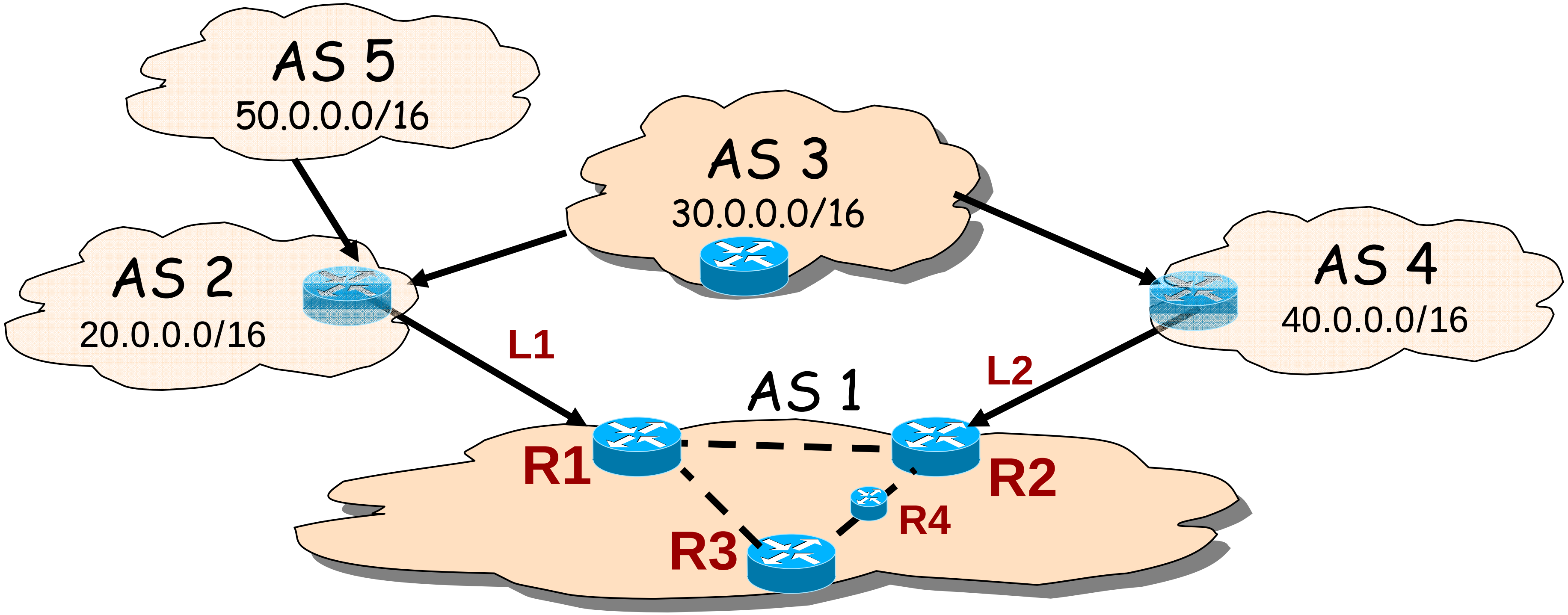
- **Question:** Which routes select AS1 routers?
- **No LOCAL_PREF configuration. R4 is not a BGP router**
- **R1, R2 and R3 choose the rule of lowest AS_PATH**
  - **R1, R2 and R3 choose link L1 for routes received from AS2 y AS5**
    - ✓ **R2 and R3 will send packets to prefixes inside those ASs through R1**
  - **R1, R2 and R3 choose link L2 for routes received from AS4**
  - **This rule does not decide the path to AS3**

Route Selection: Example 1



- Route to AS3 is not decided yet. There is neither ORIGIN nor MED.
- Suppose there are intradomain metrics in AS1 (for example, because RIP is used – number of hops)
- Then, apply rule of preferring EBGP over IBGP
  - R1 chooses {AS2 AS3}, R2 chooses {AS4 AS3}
- R3 receives both routes by IBGP. Applies rule of less distance to NEXT_HOP
  - R3 chooses {AS2 AS3} (2 hops away NEXT_HOP)
- Note: this is *Hot-Potato* behaviour

Route Selection: Example 1

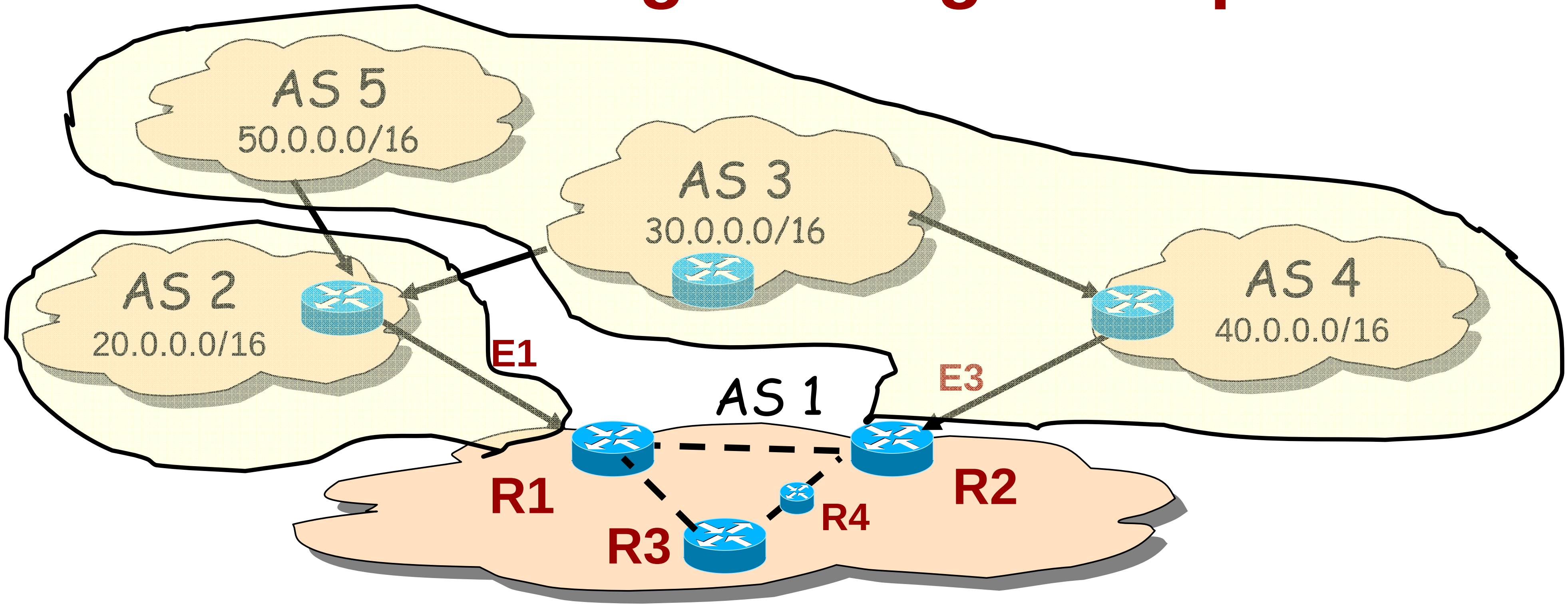


- Route to AS3 is not decided yet. There is neither ORIGIN nor MED.
- Suppose there are NO intra-domain metrics in AS1
- Apply rule of preferring EBGP over IBGP
  - **R1 chooses route through {AS2 AS3}**
  - **R2 chooses route through {AS4 AS3}. Useless rule for R3 (both routes through IBGP)**
- R3 can not apply rule of less distance to NEXT_HOP
- Note: R1 y R2 behave as *Hot-Potato*

What can a site express about route selection

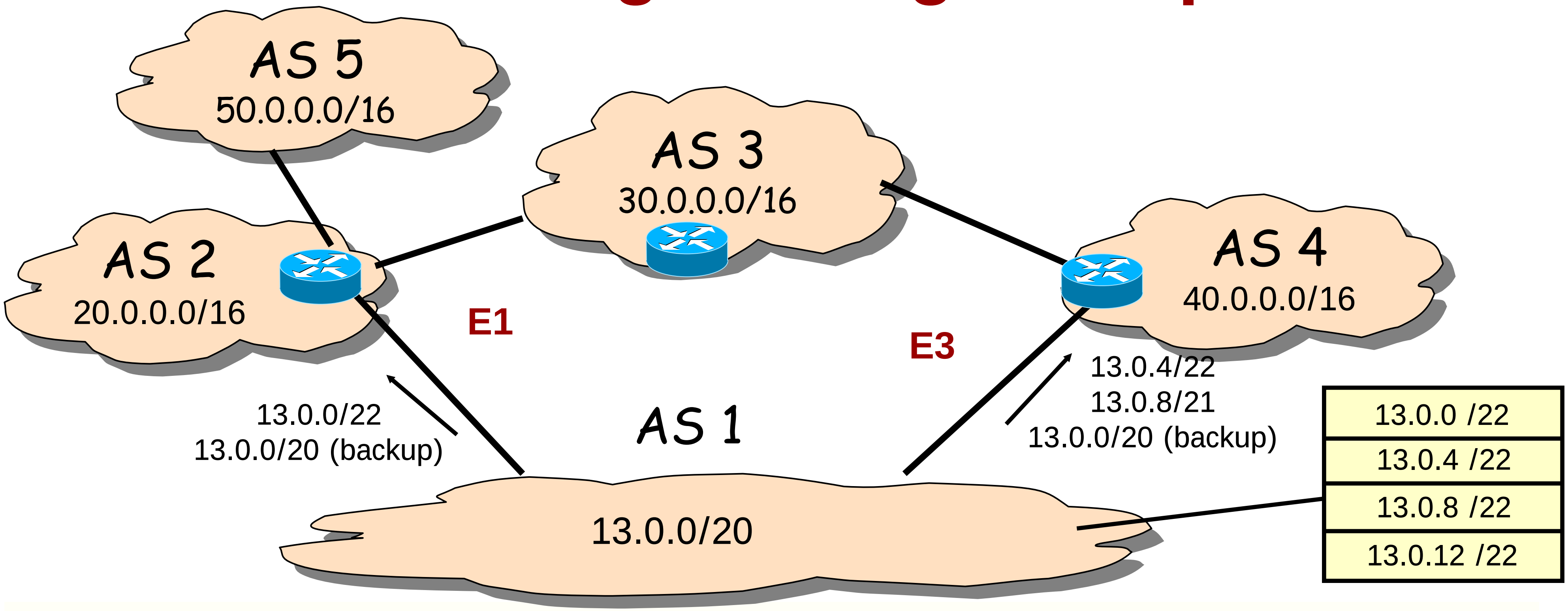
- **Send traffic**
  - **A site can always decide where to send its packets**
    - Except for such cases more specific prefixes are in use
      - Although more specific prefixes could be filtered
    - **LOCAL_PREF, generated from a prefix, an AS…**
    - **All following possibilities are dependent from this one**
- **Receive traffic**
  - **A site can force a path by means of more specific prefixes**
  - **A site can express a preference from which of two distinct sites it wants to receive traffic, and can extend this to sites farther away (than its immediate neighbors)**
    - **Fictitious increment of AS_PATH: "AS Prepending"**
      - E.g.: 3352 15630 15630 15630 15630
    - **Use of previously agreed communities**
      - Remote sites must know and use them
  - **A site can suggest to his immediate neighbor where it wishes to receive traffic if they have two or more common connections**
    - **MED, accompanied by a prefix**

Traffic Engineering Example



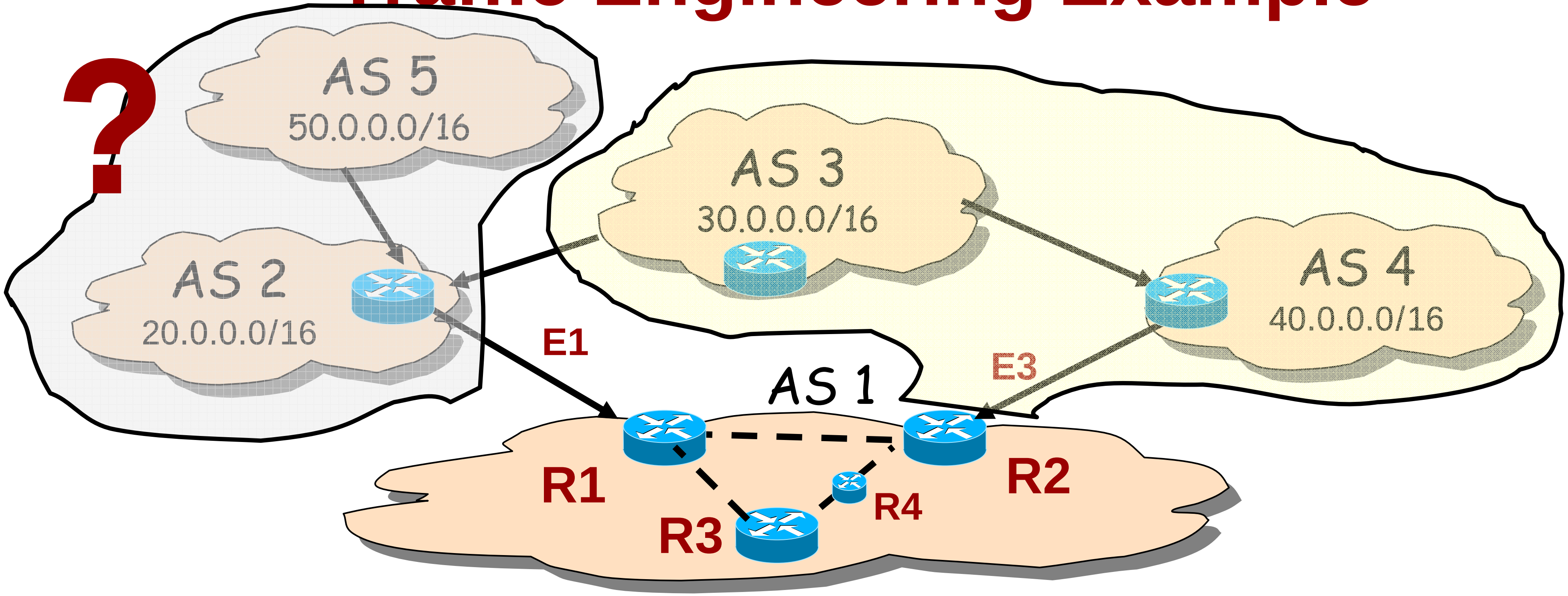
- Link E3 provides 4 times more bandwidth than E1.
- Aim: try to suit the traffic **sent** to the infrastructure, to <u>send 4 times more traffic through E3 than through E1</u>
  - Suppose that the amount of traffic exchanged by each one of the remotes ASs is similar
- Solution: configure R2 with LOCAL_PREF 120 for routes 30.0/16, 40.0/16, 50.0/16
  - Configure R2 with LOCAL_PREF 120 for route 20.0/16

Traffic Engineering Example



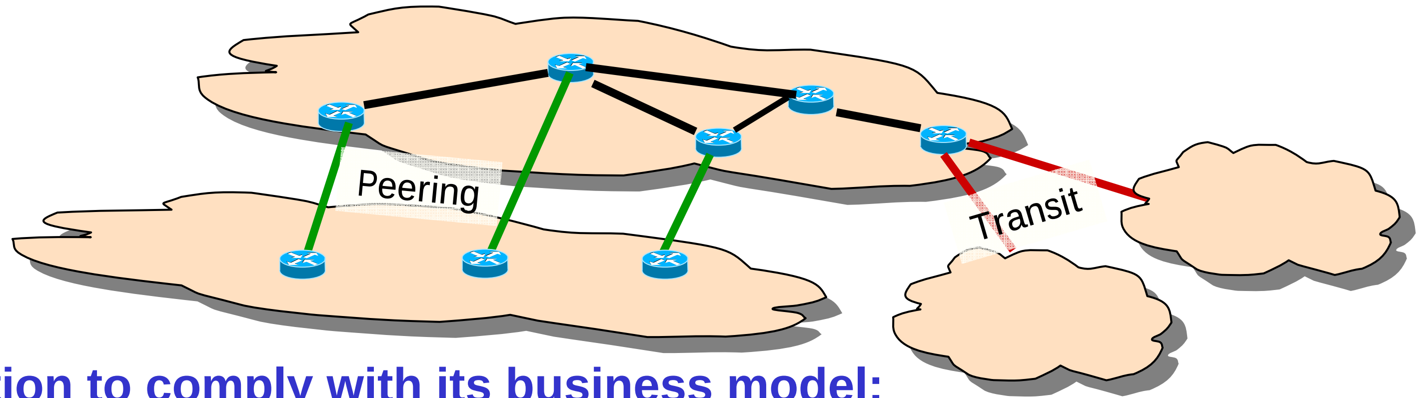
- Context: E3 link has 4 times more capacity than E1. Objective: Try to suit the traffic **received** by AS1, in order to receive about 4 times more traffic by E3 than by E1.
  - Supose that the exchanged traffic volume with every AS is similar.
- Propagate more specific routes of the addressing assigned to AS1, in order that every link attracts, preferently, different traffic.
  - ✓ Adjust the announcements based on the measures taken.
- In the example, enters 3 times more traffic by E3 than by E1, if the traffic is homogenously distributed in the address space.
- Note: actually, half the forwarding table entries are more specific routes of other also propagated.

Traffic Engineering Example



- Solution: suppose that remote ASs apply rule of lowest AS_PATH, so perform AS prepend
  - Through E1, to make this link less preferred, send local prefixes with AS1 AS1 AS1
  - Through E3, send route with just AS1
- Result: AS3 and AS4 prefer route send through E3
  - AS2… just don't know, since distance is the same though both options ({AS1, AS1, AS1} and {AS1 AS3 AS4}). The decision will depend on other factors
  - AS5 receives routes from AS2, so its traffic will follow the AS2 decision
- Another option: if four AS1s are propagated through E1=> all the traffic will enter through E3
- Conclusion: AS prepending allows traffic engineering configuration for incoming traffic, but the configurations are not very precise

Traffic Engineering Example - 2



- **Tier 1 configuration to comply with its business model:**
  - **Send to customers as most traffic you can (to earn more money)**
  - **When communicating with peers, try to spend the lower amount of your own resources**
- **Configuration:**
  - **To Customers: always send traffic by transit links**
    - ✓ **Never by peers, even if my customers are also customers of my peers**
    - ⇒ **Configure LOCAL_PREFERENCE in the links with customers**
  - **Peers: want to put hot potato in practice**
    - ✓ **Independently of AS_PATH, MED…**
    - ⇒ **Disable rule which prefers lower AS_PATH**

      `cisco% bgp bestpath as-path ignore`

      Applied rules:

      ⇒Lower value of ORIGIN (marginal impact)

      ⇒Prefer EBGP to IBGP, prefer smaller metric to NEXT_HOP