Reti di Calcolatori 2019/20
Chapter 4 - Advanced Internetworking

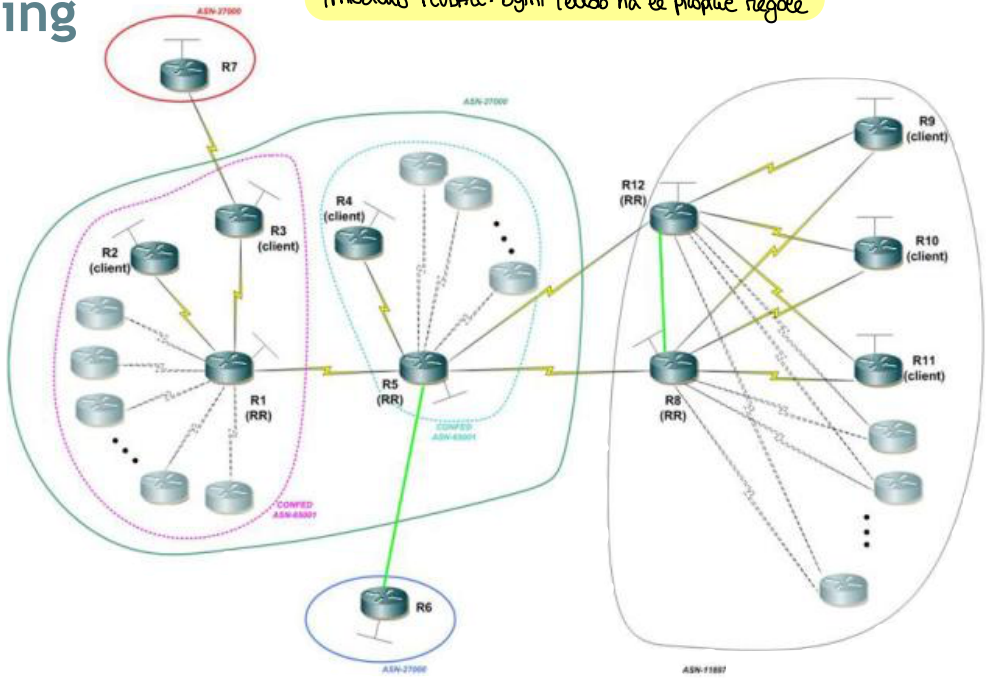
Prof. Marino Miculan

Interdomain Routing

- Internet is organized as **autonomous systems** (AS) each of which is under the control of a single administrative entity
  ↳ all'interno di essi si usa un solo algo di routing, es. RIP
- Autonomous System (AS)
  - corresponds to an administrative domain
  - examples: University, company, backbone network
- A corporation's internal network might be a single AS, as may the network of a single Internet service provider

Interdomain Routing

modello FEUDALE: ogni feudo ha le proprie regole



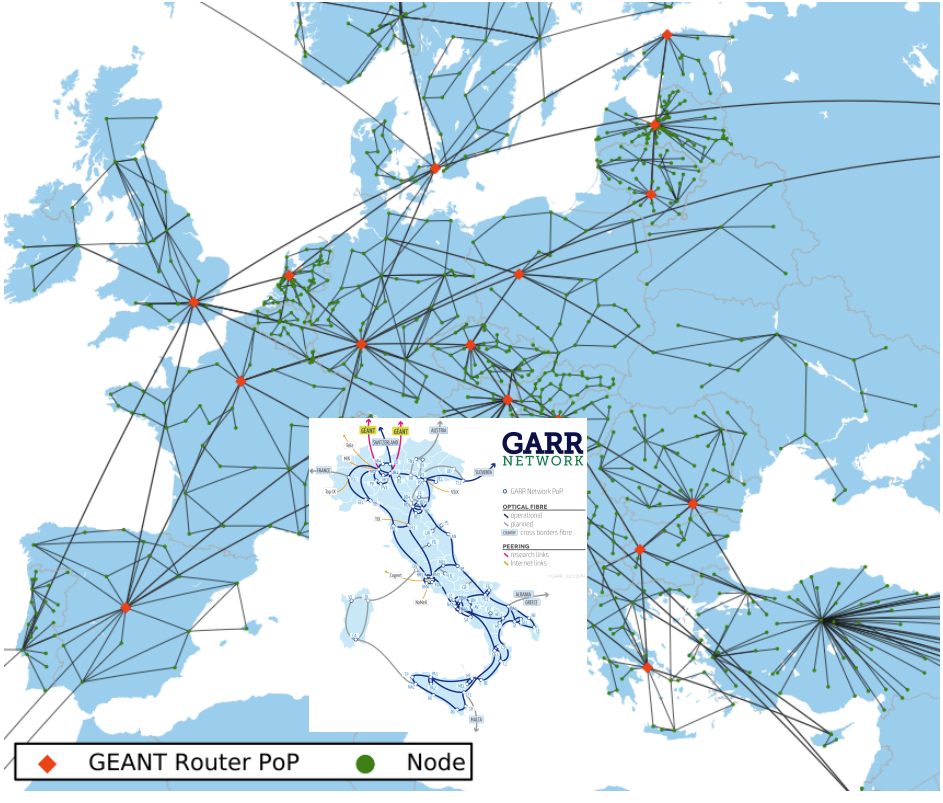
A network with some autonomous system

Some real networks...

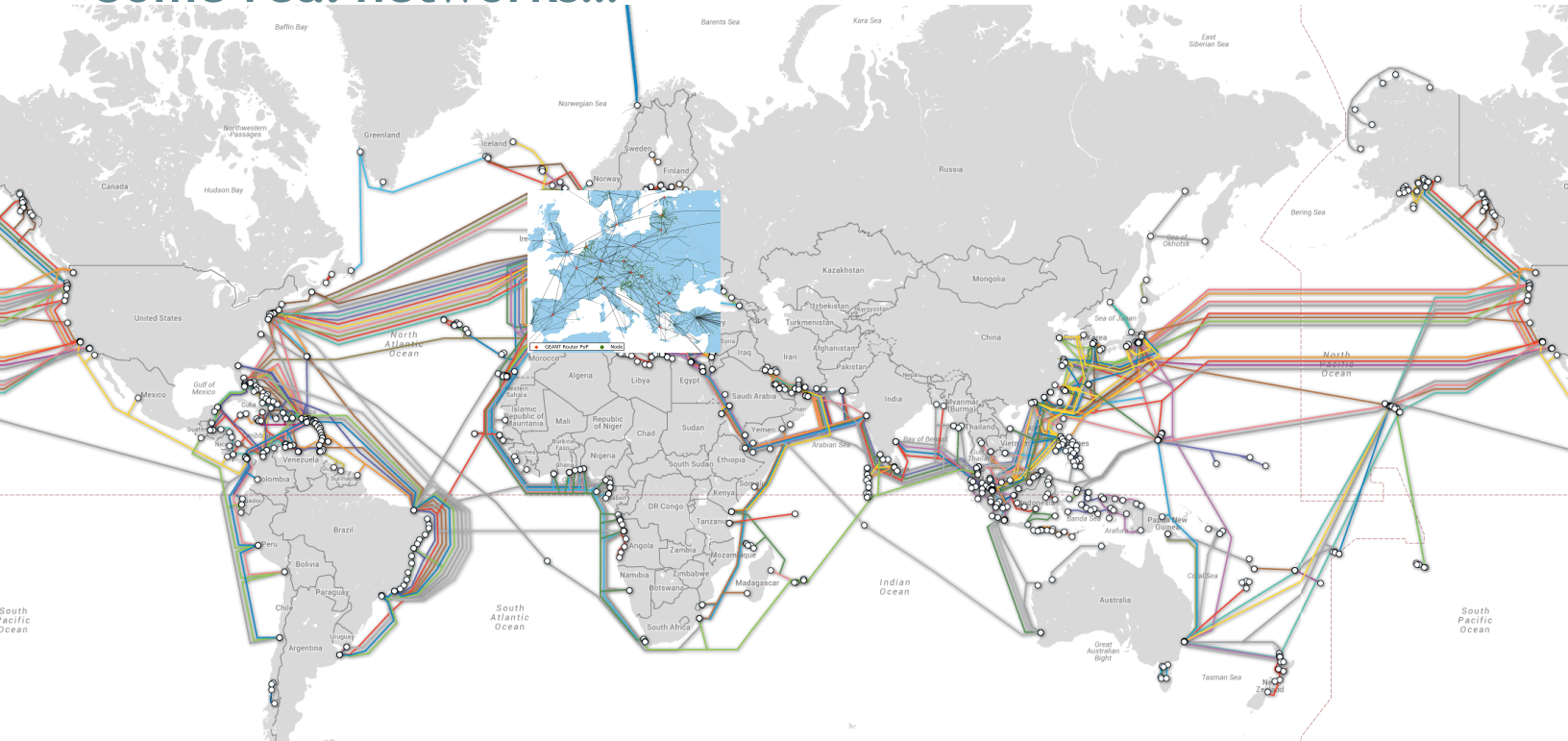
GARR
Se dall'uni mi voglio collegare al server del comune di Udine → TELECOM (provider commerciale)
NAMEX
il pacchetto viaggia sulla linea istituz. fino a Roma/Milano e lì cambia provider (passa a Telecom) e torna indietro fino al router del comune di Udine

→ provider di ricerca (per UNIVERSITÀ, ECC...)

Some real



Some real networks…

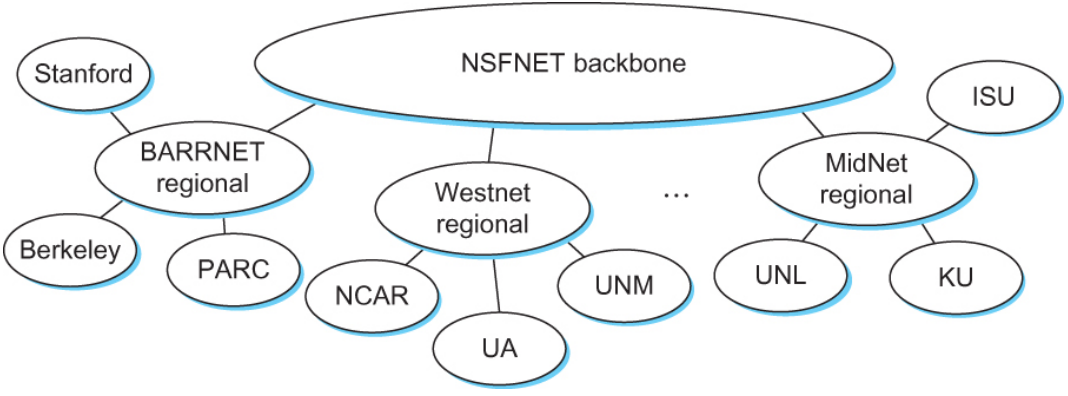


Route Propagation

- Idea: Provide an additional way to hierarchically aggregate routing information is a large internet.
  - Improves scalability
- Divide the routing problem in two parts:
  - Routing **within** a single autonomous system
  - Routing **between** autonomous systems
- Another name for autonomous systems in the Internet is *routing domains*
- Two-level route propagation hierarchy
  - Inter-domain routing protocol (Internet-wide standard)
  - Intra-domain routing protocol (each AS selects its own)

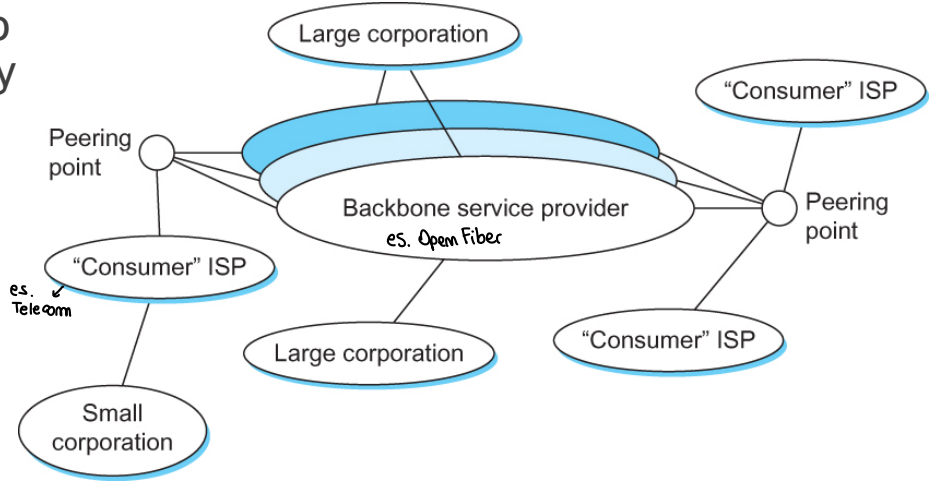
EGP and BGP

- Inter-domain Routing Protocols (ovvero tra SISTEMI AUTONOMI (AS) diversi) → EGP, BGP
  - Exterior Gateway Protocol (EGP)
    - Forced a tree-like topology onto the Internet
    - Did not allow for the topology to become general
      - Tree like structure: there is a single (SPINA DORSALE) backbone and autonomous systems are connected only as parents and children and not as peers - used to be like this in 1990:



Border Gateway Protocol (BGP)

- Assumes that the Internet is an arbitrarily interconnected set of ASs.
- Today's Internet consists of an interconnection of multiple backbone networks (they are usually called service provider networks, and they are operated by private companies rather than the government)
- Sites are connected to each other in arbitrary ways



BGP

- Some large corporations connect directly to one or more of the backbone, while others connect to smaller, non-backbone service providers.
- Many service providers exist mainly to provide service to "consumers" (individuals with PCs in their homes), and these providers must connect to the backbone providers
- Often many providers arrange to interconnect with each other at a single "peering point"

BGP-4: Border Gateway Protocol

→ più importante trovare dei percorsi (vista la complessità) piuttosto che trovare percorsi efficienti (a differenza di RIP, OSPF)

- Assumes the Internet is an arbitrarily interconnected set of AS's.
- Define **local traffic** as traffic that originates at or terminates on nodes within an AS, and **transit traffic** as traffic that passes through an AS.
- We can classify AS's into three types:
  - es. SMALL CORPORATION
    Stub AS: an AS that has only a single connection to one other AS; such an AS will only carry local traffic (small corporation in the figure of the previous page).
  - es. LARGE CORPORATION → ha es. più connessioni con più provider
    Multihomed AS: an AS that has connections to more than one other AS, but refuses to carry transit traffic (large corporation at the top in the figure of the previous page).
  - es. PROVIDER
    Transit AS: an AS that has connections to more than one other AS, and is designed to carry both transit and local traffic (backbone providers).

Se si sbaglia una configurazione BGP si rischia di fare danni → es. down di Facebook

I transit li posso attraversare, gli altri due no

BGP

- The goal of Inter-domain routing is to find any path to the intended destination that is **loop free**
  - We are concerned with reachability than optimality
  - Finding path anywhere close to optimal is considered to be a great achievement

- Why?

BGP

I principali 3 motivi che rendono arduo l'instradamento:

- **Scalability**: An Internet backbone router must be able to forward any packet destined anywhere in the Internet
  - Having a routing table that will provide a match for any valid IP address
- **Autonomous** nature of the domains (al mondo ci sono circa 60.000 Autonomous Systems)
  - It is impossible to calculate meaningful path costs for a path that crosses multiple ASs (FACILMENTE) → proprio perché i vari AS usano algoritmi di ROUTING che possono essere diversi
  - A cost of 1000 across one provider might imply a great path but it might mean an unacceptable bad one from another provider
- Issues of **trust** → tra AZIENDE CONCORRENTI
  - Provider A might be unwilling to believe certain advertisements from provider B

BGP

- Each AS has one (or more) **BGP speaker** that advertises to its peers:
  - local networks
  - other reachable networks (transit AS only)
  - gives path information
- In addition to the BGP speakers, the AS has one or more **border gateways** which need not be the same as the speakers
- The border gateways are the routers through which packets enter and leave the AS

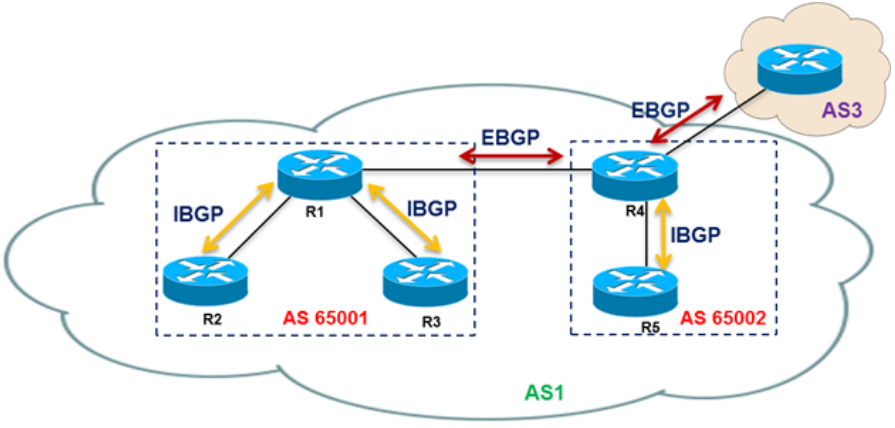
PORTAVOCE che comunica quali sono le sue reti locali e le reti non locali per le quale è transit

"potete mandarmi il traffico per queste reti..."

comunica con altri speaker BGP di altri AS per scambiare informazioni in merito alla raggiungibilità

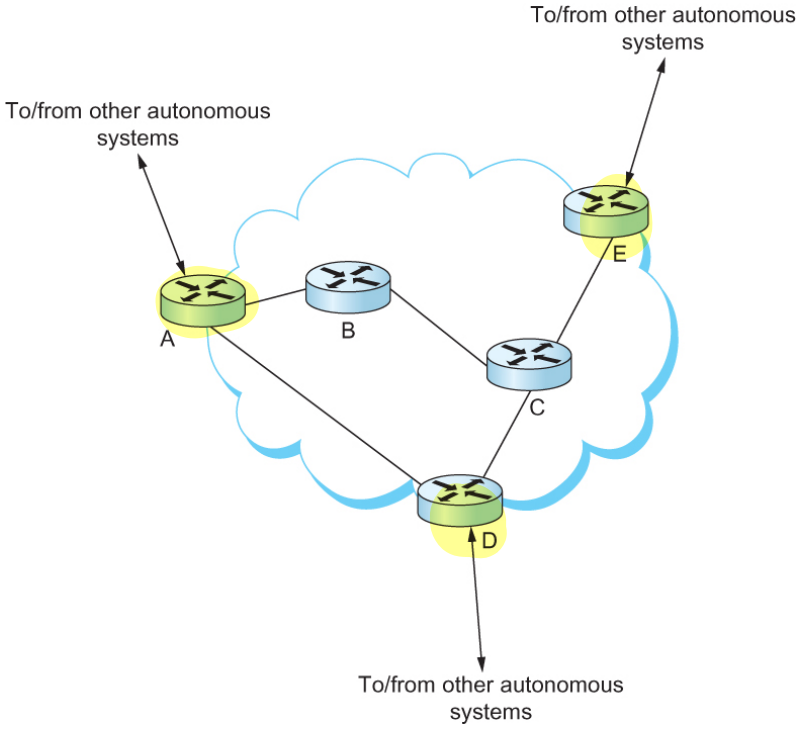
eBGP and iBGP

- Two versions of protocol:
  - **External BGP** (eBGP): used between routers belonging to different ASs
  - **Internal BGP** (iBGP): used between routers inside the same ASs



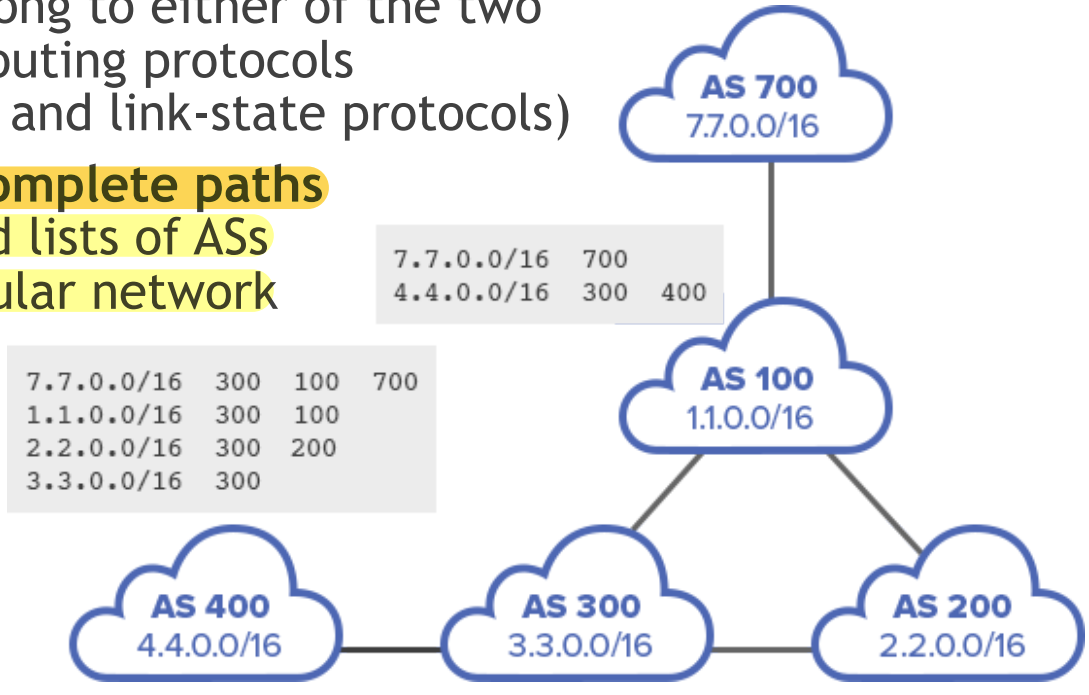
Integrating Interdomain and Intradomain Routing

- All routers run iBGP and an intradomain routing protocol (e.g. OSPF).
- Border routers (A, D, E) also run eBGP to other ASs



BGP

- BGP does not belong to either of the two main classes of routing protocols (distance vectors and link-state protocols)
- BGP advertises **complete paths** as an enumerated lists of ASs to reach a particular network



BGP loop prevention and policy routing

- A BGP router ignores a path if
  - its AS appears in it (this avoids loops between ASs)
  - Or if it violates some policy of the AS

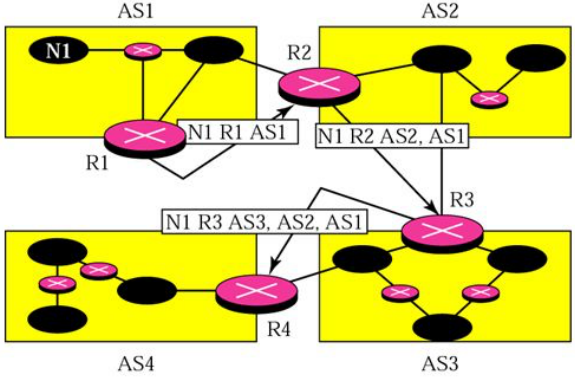
↳ ovvero se ci sono AS "indesiderati" o poco affidabili nel cammino

Example of Network Reachability

Network	Next router	Path
N1	R1	AS14,AS23,AS67
N2	R5	AS22,AS67,AS5,AS89
N3	R6	AS67,AS89,AS9,AS34
N4	R12	AS62,AS2,AS9

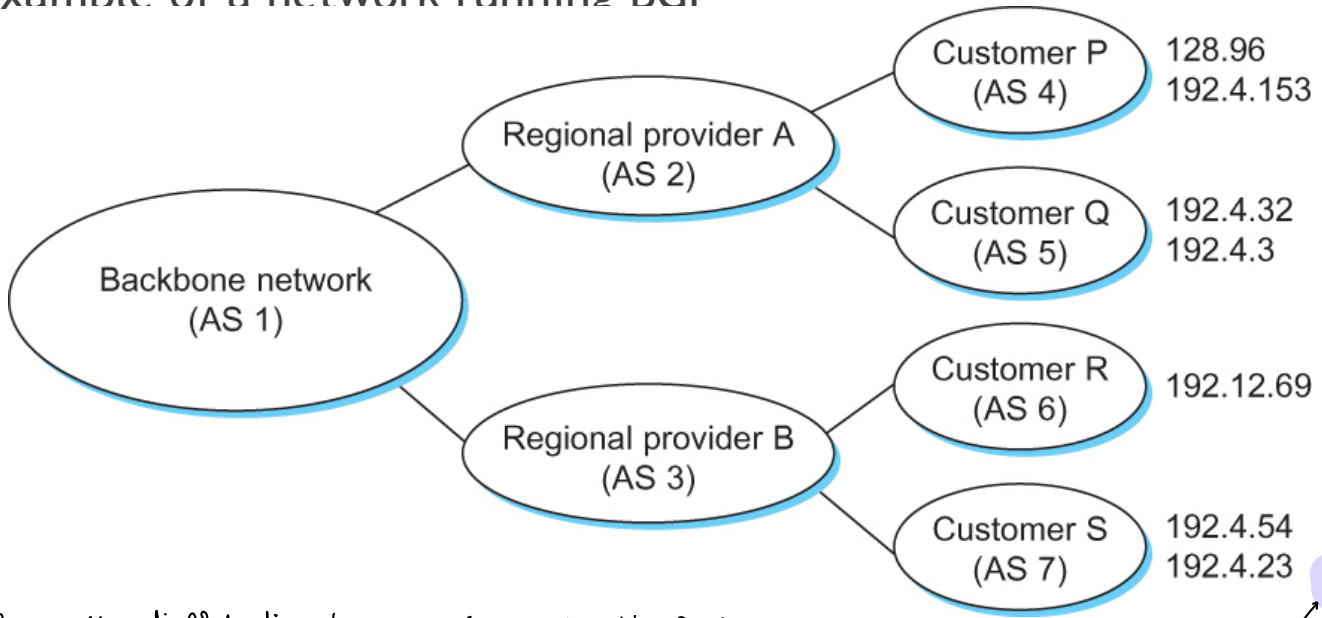
- Loop Prevention in BGP:
  - Checks the Path before updating its database. (If its AS is in the path ignore the message)
- Policy Routing:
  - If a path consist of an AS against the policy of the current AS, message discarded.

Example of Message adverstisements



BGP Example

- Example of a network running BGP



ogni AS è identificato UNIVOCAMENTE

es. lo speaker di AS 1 dice che per arrivare alle reti 128.96, 192.4.153, 192.4.32, 192.4.3 si passa per <AS1, AS2>

BGP Example

- Speaker for AS 2 advertises reachability to P and Q
  - Network 128.96, 192.4.153, 192.4.32, and 192.4.3, can be reached directly from AS 2.
- Speaker for backbone network then advertises
  - Networks 128.96, 192.4.153, 192.4.32, and 192.4.3 can be reached along the path <AS 1, AS 2>.
- Speaker can also cancel previously advertised paths

BGP Issues

- AS numbers carried in BGP need to be unique
  - For example, AS 2 can only recognize itself in the AS path in the example if no other AS identifies itself in the same way
- AS numbers (ASN) are 32-bit numbers assigned by a central authority (IANA and regional internet registries) (RFC 4893)
- About 60.000 ASs, at the moment
  - these are the nodes of the network running eBGP

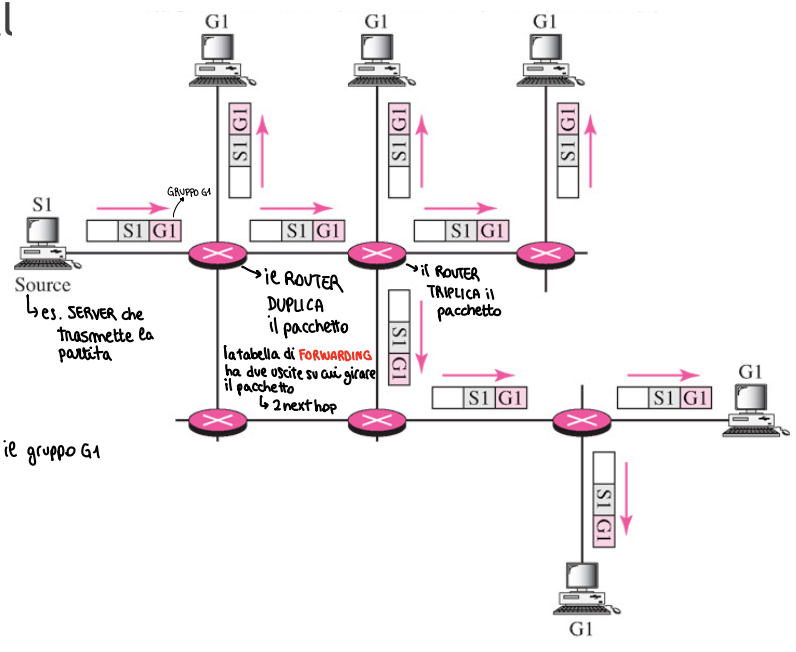
Internet Multicast

SINGOLO PACCHETTO IP ricevuto da un gruppo di HOST

- Multicast = the possibility of sending a message to a group of receiving hosts, without knowing or specifying them
- One-to-many: Source specific multicast (SSM) → es. DAZN (uno condivide a tutti)
  - A receiving host specifies a multicast group where a specific host is sending. Examples:
    - Radio station broadcast
    - Transmitting news, stock-price
    - Software updates to multiple hosts
- Many-to-many: Any source multicast (ASM) → es. chat gruppo whatsapp
  ↳ tutti condividono a tutti
    - Multimedia teleconferencing
    - Online multi-player games
    - Distributed simulations

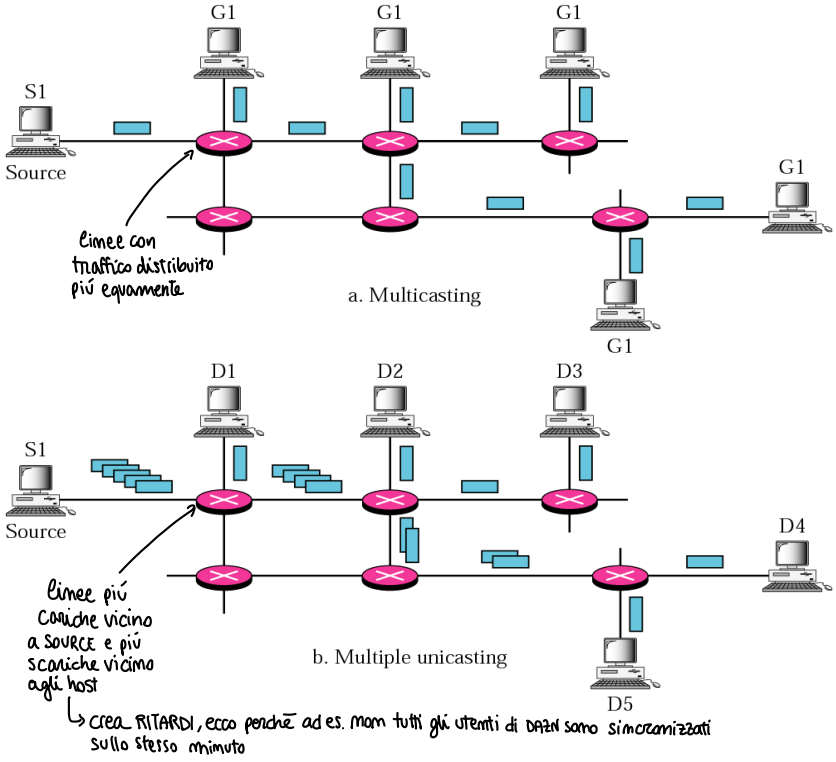
Multicast in IP

- Using IP multicast to send the identical packet to each member of the group
  - A host sends a *single* copy of the packet, addressed to the group's multicast address
  - Each host members of that group receives a copy of the packet
  - The source host does not need to know the individual unicast IP address of each member
  - Packets are duplicated by routers along the way, when needed



Multicast without multicast?

- Without support for multicast, a source needs to send a separate packet with the identical data to each member of the group
  - This redundancy consumes more bandwidth
  - Redundant traffic is not evenly distributed, concentrated near the sending host
  - Source needs to keep track of the IP address of each member in the group
    - Group may be dynamic



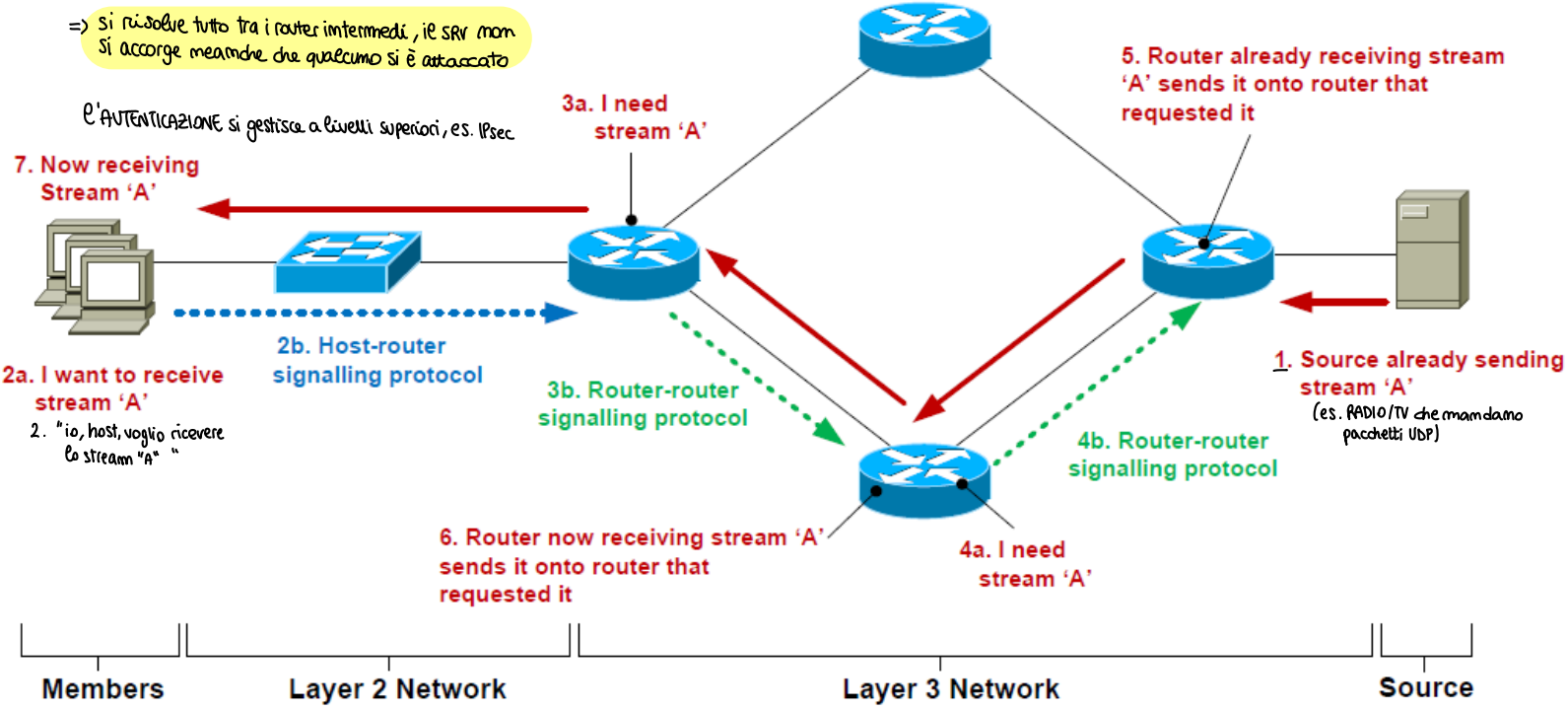
Multicast in IP

- IP provides an IP-level many-to-many multicast model based on **multicast groups**
  - Each group has its own IP multicast address in the D class (224.0.0.0 – 239.255.255.255)
  - Group addresses are assigned in many ways:
    - 224.0.0.0/24: Use is defined by IANA, but restricted to local network
    - 224.0.1.0/24: Global groups, statically assigned by IANA; e.g. 224.0.1.1 is NTP
    - Dynamically leased for the time of a *session*, using SAP/SDP protocol
    - …

Multicast group management

- A host can join and leave groups
- A host can be in multiple groups
- A host signals its desire to join or leave a multicast group by communicating with its *local router* using a special protocol
  - In IPv4: Internet Group Management Protocol (**IGMP**)
  - In IPv6: Multicast Listener Discovery (**MLD**)
- The router has the responsibility for making multicast behave correctly with regard to the host

Multicast Service Model Overview



Multicast service

→ broadcast e multicast su WiFi funziamano male (con CSMA-CA)
perchè bisogna aspettare CTS e RTS da chi precisamente? → per il MULTICAST devi usare il cavo

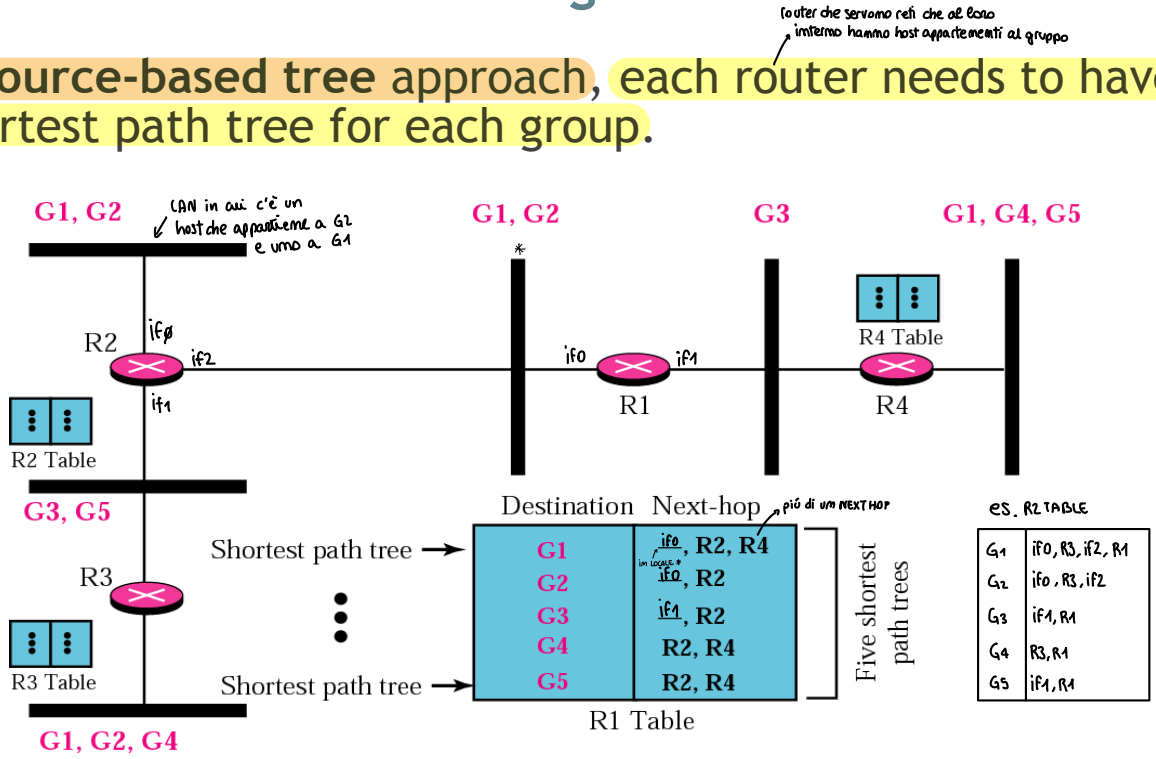
- Once a host has joined a group, it receives all messages sent to that group, and can send any message to that group as well
- Scope of messages can be defined by choosing a suitable TTL
- Security issues (secrecy, integrity, authentication, access control…) are not handled at this level (can be handled at session/application level)

Multicast Routing → le tabelle di FORWARDING non cambia molto, per un gruppo multicast invia a uno o più next hop

- A router's unicast forwarding tables indicate for any IP address, which link to use to forward the unicast packet
- To support multicast, a router must additionally have **multicast forwarding tables** that indicate, based on multicast destination address, which links to use to forward the multicast packet
- Multicast forwarding tables collectively specify a set of trees, one for each group: **Multicast distribution trees**
- **Multicast routing** is the process by which multicast distribution trees are determined
- Many routing protocols have been proposed: DVMRP, RPB, PIM…
- Two families: **source-based** and **group-shared** trees

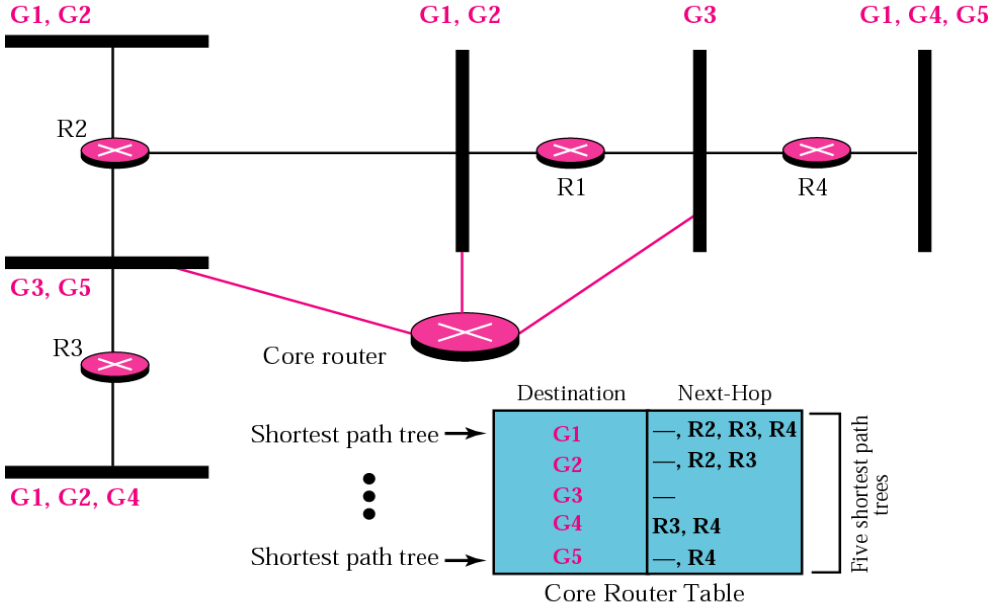
Source-based multicast routing

- In the **source-based tree** approach, each router needs to have one shortest path tree for each group.



Group-shared tree multicast routing

- In the group-shared tree approach, only a router (called **core** or **rendezvous** router) has a shortest path tree for each group, and is involved in multicasting.



Distance-Vector Multicast Routing Protocol (DVMRP)

↳ simile al ROUTING DISTANCE VECTOR UNICAST

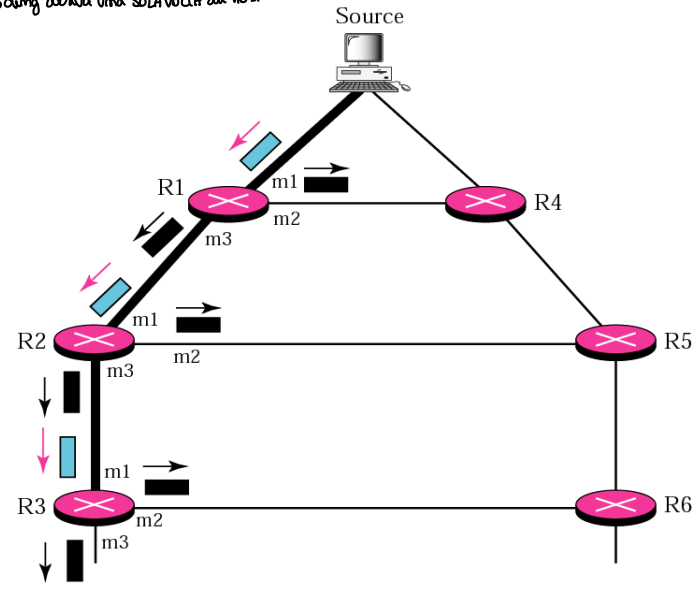
- First multicast routing protocol implemented (RFC 1075, Nov. 1988)
- Extends distance vector routing protocols for unicast routing
- "Flood and prune" strategy: flood the network with multicast traffic, and in the meanwhile prune the branches not interested in the traffic

RIMUOVI

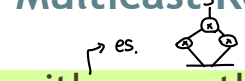
↳ provo a mandare il traffico a tutti

Distance-Vector Multicast Routing Protocol (DVMRP)

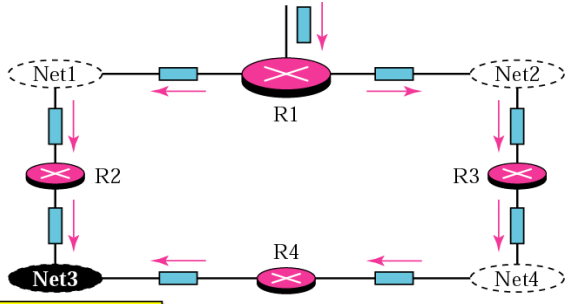
- Flooding part: **Reverse Path Flooding** → non è banale flooding, se ho 10 vie per arrivare a destinazione, il pacchetto in flooding arriva una SOLA VOLTA all'host
  - From the unicast routing table, each router already knows the next hop which the shortest path to a source S goes through.
  - When receives a multicast packet, a router looks at its source address S, and then forwards it on all outgoing links (except the one on which the packet arrived), **if and only if** packet arrived from the next hop towards S (i.e., in "reverse path" direction). → se il pacchetto gli è arrivato da quello che per lui è il percorso migliore (verso S), quindi se coincide con il suo next Hop per andare verso S, allora lo forwarda agli altri
  - In this way, there are no loops in the flooding process



Distance-Vector Multicast Routing Protocol (DVMRP)



- Still, networks with more than one router can get duplicate broadcast packets
- Eliminate duplicate packets by fixing a "parent" (relative to S) router for each LAN and letting only parents forward packets
- Parent routers are those with the shortest path to S (which they know via distance vector) - in the case above, R2 is parent for Net3
- Break ties by choosing router with smallest IP address
  - (Compare with the Spanning tree protocol in switched LANs)
- This strategy, called **Reverse Path Broadcast**, guarantees that each LAN receives exactly one copy of each packet (less loss of packets), along a **source-based shortest path tree**



Distance-Vector Multicast Routing Protocol (DVMRP)

- **Pruning part**: Prune networks that have no hosts interested in group G. Done in two steps.
- Step 1: Determine if LAN is a leaf (FOGLIA) of the distribution tree with no members in G
  - A LAN is a leaf if its parent is the only router on the LAN
  - Hosts in the LAN willing to participate to G must notify the router using IGMP (periodically - they are removed after a timeout)
  - Thus, the router knows if LAN has no members of G

Distance-Vector Multicast Routing Protocol (DVMRP)

- Step 2: Propagate “no members of G here” information
  - augment each entry <Destination, Cost> in distance vectors sent to neighbours with the list of groups for which this network is interested in receiving multicast packets
  - Thus, each neighbour knows whether they have to consider this router in the reverse path flooding/broadcast, by looking at the groups it is interested into
  - Including always the list of interested groups in distance vectors may be heavy and useless (e.g. in the case a LAN is interested in several groups, but no host is transmitting on these groups)
  - Instead, the router adds to distance vectors the list of groups which it is NOT interested into, and only when multicast address becomes active (someone is transmitting)

Distance-Vector Multicast Routing Protocol (DVMRP)

- So, when a multicast transmission starts on group G:
  - at first it floods all the network, via RPF (reverse path flooding) / RPB (reverse path broadcast), building a distribution tree which covers basically all the network (because no one is saying that it is not interested in G, yet)
  - and then the tree is pruned: the routers not interested in G start notifying back to stop transmitting to them
  - If a host of a pruned net joins the group G later on, it notifies its local router using IGMP, and the router will start accepting the traffic from his neighbours (*grafting*)
- Overall, DVMRP works well on small scale (↳ intra-domain), less on large scale, due to this continuous flooding-and-pruning

Multicast routing: PIM

→ un'evoluzione di DVMRP: è impensabile in grandi reti inviare il traffico in modalità broadcast finché un router dichiara di non essere interessato
↳ bisogna trovare una modalità per reti di scala grande
↳ simile a RIP

- Protocol Independent Multicast
  - independent from the construction of the network topology (to be solved with a different protocol)
- Has two different modes
- PIM-DM (dense mode) (RFC 3973, 2005) → dentro un AS
  - Used when many hosts are origin of data with respect to the routers
    - e.g. inside a LAN, or an autonomous system
- PIM-SM (sparse mode) (RFC 2362, 1998) → gruppo sparpagliato nel globo (quindi in mezzo tra loro ci sono molti router)
  - Used when multicast involves few nodes in the net (with respect to the number of routers)
    - E.g. few hosts across the Internet

PIM Dense Mode

- PIM-DM (dense mode)
  - Similar to DVMRP, but independent from the underlying unicast routing protocol (differently from DVMRP, which is tied to distance vector protocols)
  - **Source-based trees**: each node (router) keeps its copy of multicast distribution trees
  - Uses flood-and-prune strategy: RPF to distribute traffic, and pruning/grafting to handle trees.
  - Works well when the group is not too large, and not too many routers have to be traversed (e.g. inside a LAN or AS)

simile a DVMRP però non è legato all'algoritmo di UNICAST usato dal router, invece DVMRP è legato al DISTANCE VECTOR

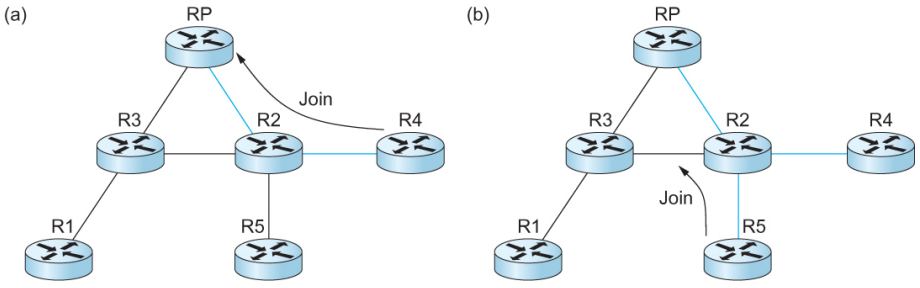
PIM Sparse Mode

ho un src che trasmette (il suo router) al router rendez-vous e poi questo in multicast trasmette a tutti i router interessati. Questi router si collegano al router RP quando ricevono un pacchetto IGMP all'interno della loro rete

- PIM-SM (sparse mode)
  - Can scale to large groups, sparse on the Internet
    - (e.g. it is used for IPTV)
  - For each group, a **group-shared tree** is built and in each AS a "rendez-vous" router is elected
    ↳ router RADICE
  - Routers (in consequence of an IGMP request from some local host) can join the shared trees by sending a (unicast) "join" request to their rendez-vous point → il ROUTER RP è un "punto di ritrovo" per tutti i router che vogliono entrare in qualche gruppo multicast
  - As it traverses the routers, the "join" message to group G creates a multicast distribution tree, rooted at the rendez-vous point RP for G

PIM Sparse Mode

Shared Tree



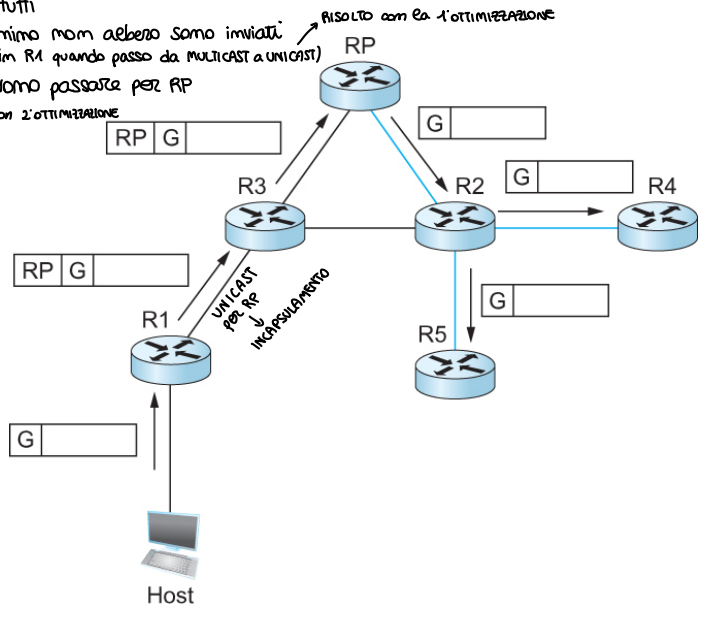
- Each router analyses the message and adds to its table the rule to forward downwards the traffic from group G along the interface the "join" message came from
- If the router was not participating to the tree already (a), then it forwards the "join" request towards RP, and marks the corresponding interface as the only one where the traffic can come from; otherwise (b) does nothing.

PIM Sparse Mode

- nella modalità normale tutti i pacchetti inviati lungo un cammino nom albero sono inviati im UNICAST (quindi INCAPSULAMENTO im R1 quando passo da MULTICAST a UNICAST) → RISOLTO con la 1' OTTIMIZZAZIONE
- tutti i pacchetti per le gruppo devono passare per RP
  ↳ risolto con 2' OTTIMIZZAZIONE

- Once the shared tree is in place:

1. Host sends packet to group G on its LAN (→ è il SAV che trasmette)
2. It is received by the Designed Router (here R1), besides any other local host
3. The DR tunnels it to the RP, encapsulated in a normal *unicast* IP packet
4. RP receives the packet, opens it, and forwards it down the shared tree (here to R2, then R4 and R5)



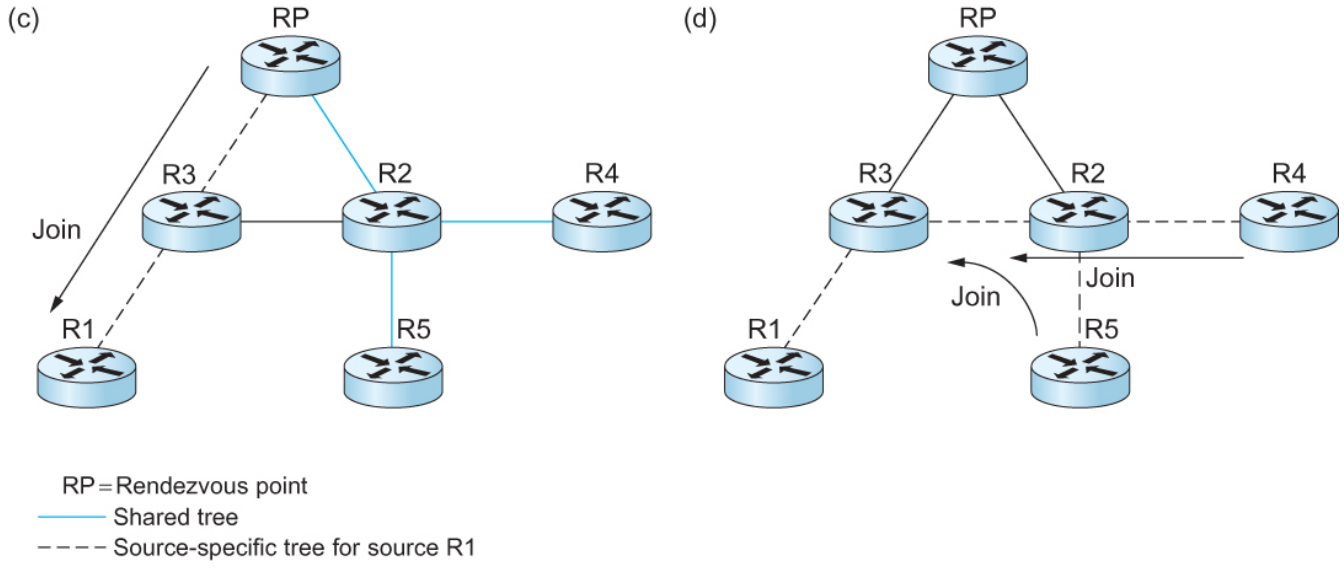
- notice that we can assume that routers between DR and RP are not participating to the tree (they may be not implementing multicast at all, as often happens on the Internet); only DR needs to know which is the RP

PIM Sparse Mode

- If the source router generates a lot of traffic, encapsulation may be costly
- Optimisation: build a **source-specific tree** on-the-fly, whose root is the source router, and the RP is a leaf
  - RP sends a "join" message to the DR (here R1) (c)
  - Intermediate routers (here R3) become aware of the new tree
  - Now R1 can send native multicast traffic (without encapsulation): these go down (root is R1) along its specific tree, reach RP, which forward them down along the shared tree on the other side
- Further optimisation: clients can decide to participate to the source specific tree instead of the shared tree, by sending "join" to R1

PIM Sparse Mode

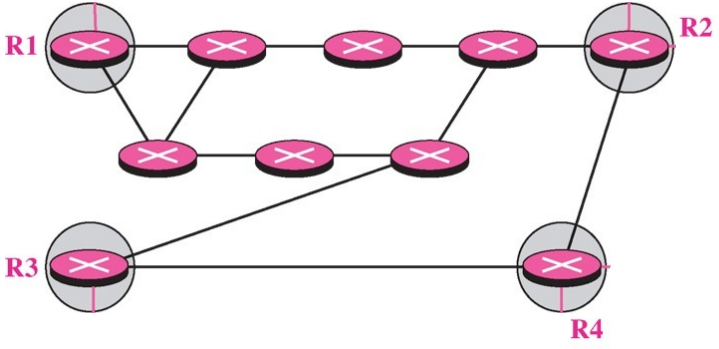
1' OTTIMIZZAZIONE



Source specific tree

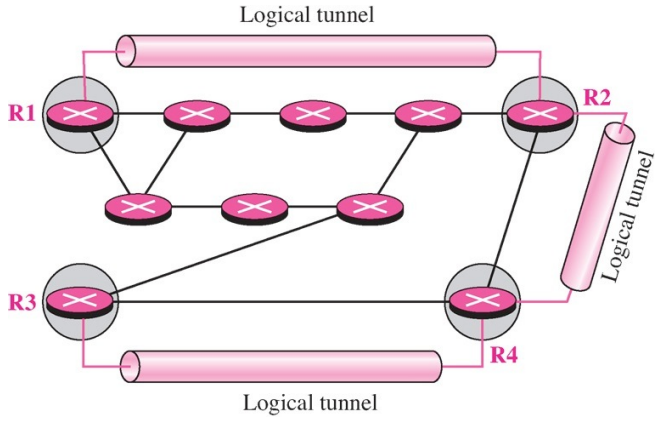
Multicast Backbone (MBONE)

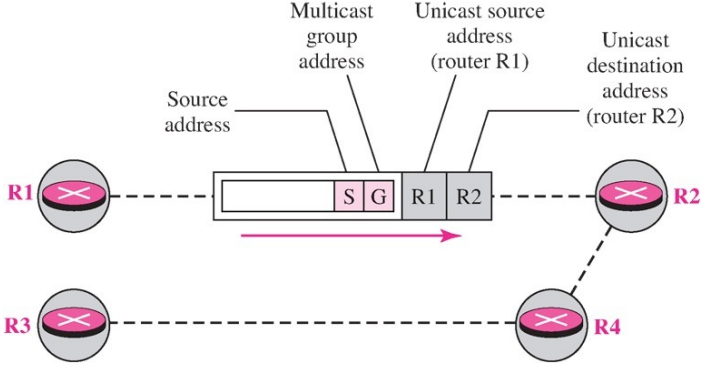
- As per today, only a small fraction of routers implement multicast routing
  - Administrators are not keen to admit multicast traffic due to their cost in routers
- Often these routers are not contiguous, i.e., are connected with regions which are not multicast-aware
  - aside: R1, R2, R3, R4 are multicast-aware routers, but all the others are not (they just drop all multicast traffic)



Multicast Backbone (MBONE)

- "Temporary" solution: create some unicast **tunnel** between the multicast router, crossing through non-multicast router
- Tunnel are logical links, and act as a *backbone* for multicast, called **MBONE** = **Multicast backbone**
- Multicast packets are encapsulated inside normal unicast IP packet, and can move through the multicast-unaware routers as usual.
- Multicast routing protocols can run over the MBONE





Multicast Backbone (MBONE)

- MBONE (and multicast in general) are not very widespread
  - accounting issues, traffic management, traffic on the providers…
  - More load on routers
  - largest event ever transmitted live: Rolling Stones concert in Dallas,1994 (50.000 clients)
- MBONE is obsoleted, and it is going to be replaced by IPv6, with PIMv6
- Still, multicast is adopted within single organisations (at levels of LANs and AS), e.g. for carrying IPTV content within an hotel/ hospital/provider domain…

Summary

- We have looked at the issues of scalability in routing in the Internet
- We have discussed IPv6 (in chapter 3, actually)
- We have discussed Multicasting
- We have discussed Mobile IP