DIPARTIMENTO
DI INFORMATICA
Via Salaria 113 - 00198, Roma

SAPIENZA
UNIVERSITÀ DI ROMA

Chapter 4
Network Layer

Reti di Elaboratori
Corso di Laurea in Informatica
Università degli Studi di Roma "La Sapienza"
Canale A-L
Prof.ssa Chiara Petrioli

Parte di queste slide sono state prese dal materiale associato al libro *Computer Networking: A Top Down Approach* , 5th edition.

Thanks also to Antonio Capone, Politecnico di Milano, Giuseppe Bianchi and Francesco LoPresti, Un. di Roma Tor Vergata

Chapter 4: Network Layer

- 4. 1 Introduction
- 4.2 Virtual circuit and datagram networks
- 4.3 What's inside a router
- 4.4 IP: Internet Protocol
  - Datagram format
  - IPv4 addressing
  - ICMP
  - IPv6
- 4.5 Routing algorithms
  - Link state
  - Distance Vector
  - Hierarchical routing
- 4.6 Routing in the Internet
  - RIP
  - OSPF
  - BGP
- 4.7 Broadcast and multicast routing

OSPF (Open Shortest Path First)

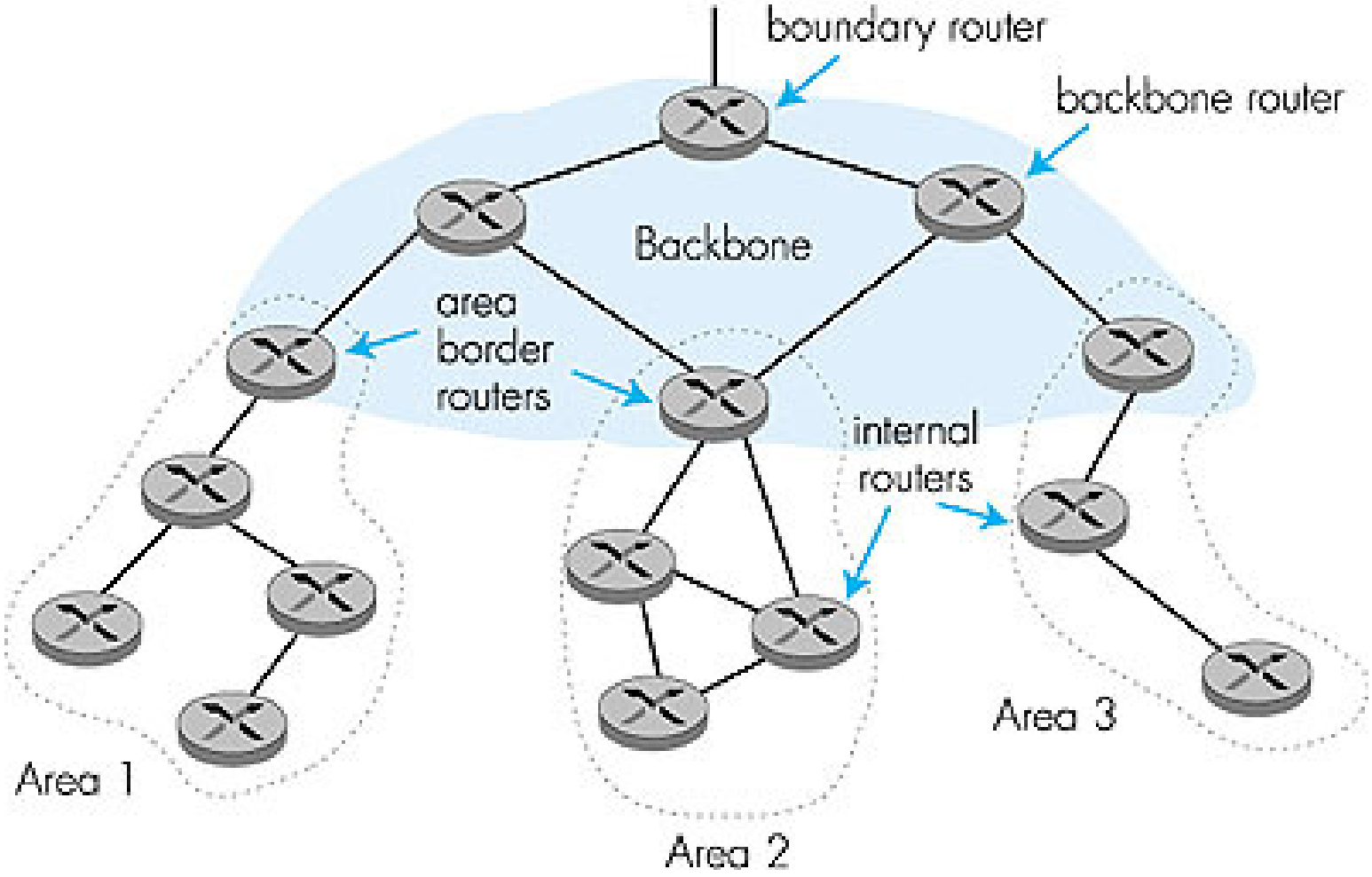
- "open": publicly available
- uses Link State algorithm
  - LS packet dissemination
  - topology map at each node
  - route computation using Dijkstra's algorithm
- OSPF advertisement carries one entry per neighbor router
  - Each node disseminates its local view of the topology
  - i.e., the router usable interfaces and reachable neighbors
- advertisements disseminated to entire AS (via flooding)
  - carried in OSPF messages directly over IP (using protocol nuner 89)

OSPF "advanced" features (not in RIP)

- security: all OSPF messages authenticated (to prevent malicious intrusion)
- multiple same-cost paths allowed (only one path in RIP)
- For each link, multiple cost metrics for different TOS (e.g., satellite link cost set "low" for best effort; high for real time)
  - Periodic updates (30 min) or event driven (link cost change)
- integrated uni- and multicast support:
  - Multicast OSPF (MOSPF) uses same topology data base as OSPF
- hierarchical OSPF in large domains.

Externally derived routing data is advertised throughout the Autonomous System unaltered

Hierarchical OSPF



Splitting the AD into areas

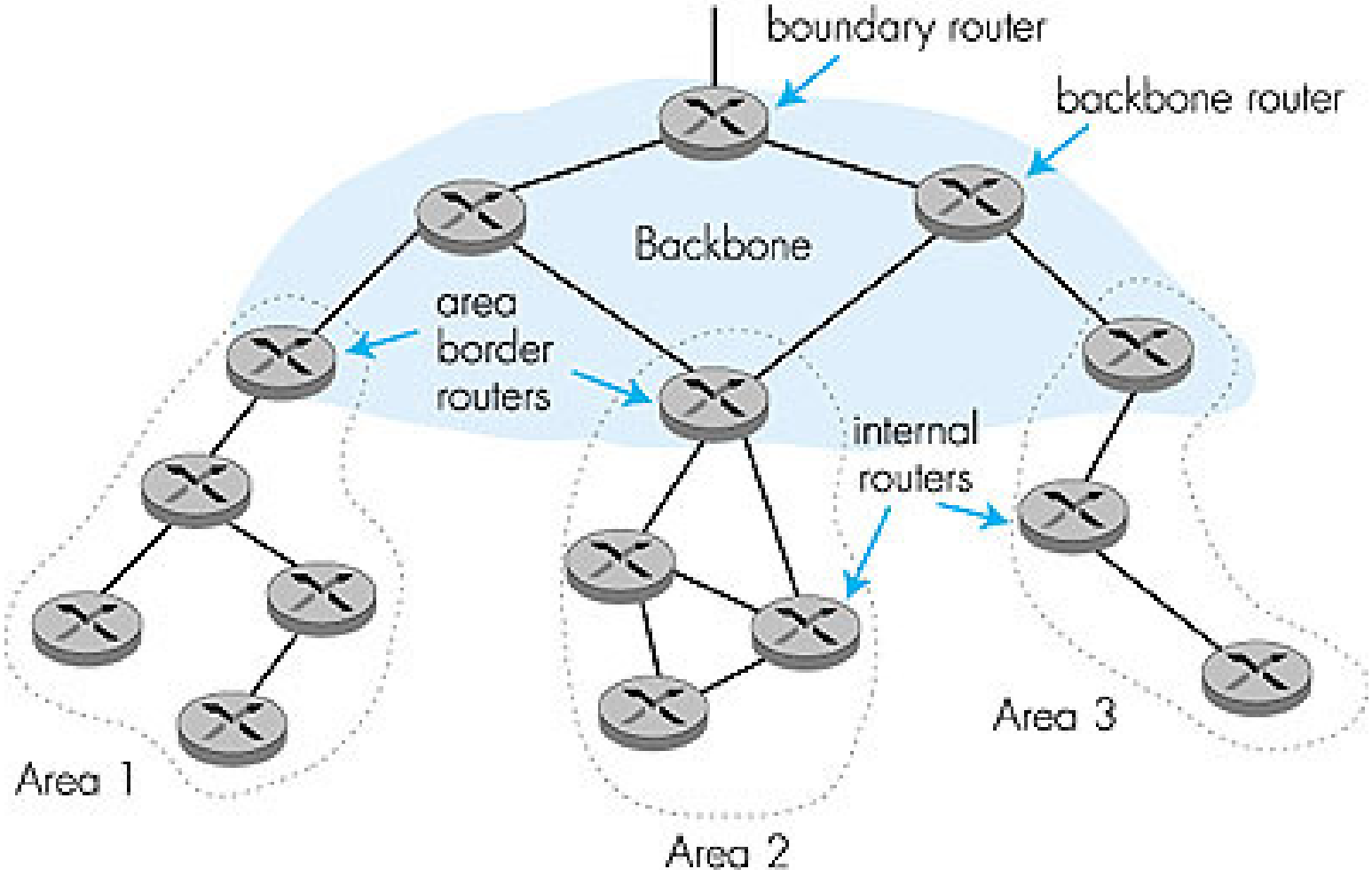
- OSPF allows collections of contigous networks and hosts to be grouped together
  - Such a group together with the routers with interfaces to any of the included networks is called an area
  - Each area runs a separate copy of the basic link-state routing algorirthm
    - has its own link state database
  - The topology of an area is invisble from the outside
  - Routers internal to a given area know nothing of the detailed topology external to the area

Hierarchical OSPF

Isolamento info di area funzionali a ridurre il traffico di routing

- two-level hierarchy: local area, backbone.
  - Link-state advertisements only in area
  - each node has detailed area topology; only known direction (shortest path) to nets in other areas.
- *area border routers:* "summarize" distances to nets in own area, advertise to other Area Border routers.

Hierarchical OSPF



Hierarchical OSPF

- The OSPF backbone is the special OSPF Area 0
- The OSPF backbone always contains backbone routers
- *Backbone routers:* run OSPF routing limited to backbone.
- The backbone is responsible for distributing routing information between non backbone areas
  - Every area border router hears the area summaries from all other area border routers
  - adding backbone distance+distance in summaries each router knows distance to different destinations
  - These distances are then advertised internally to their areas
- The backbone must be contiguous but not physically contiguous
  - Backbone connectivity can be established/maintained through the configuration of virtual links (part of the backbone with actual way to route between end piint of the virtual link based on intra_AS routing)

- *Boundary routers:* connect to other AS's.
- AS external LSAs are advertised in the AS WITH THE EXCEPTION OF stub areas
  - Stub areas use a default routing

Types of networks

- Transit networks are capable of carrying data traffic which is neither locally originated nor locally destined
- A stub network only carries traffic it either generates or addressed to it

LSA (Link State Advertisement)

- Periodic advertisement
- Link state is also advertised when a router state changes
  - Hello packets used to discover and maintain neighbor relationships
- Disseminated via flooding
- Flooding algorithm is reliable ensuring that all routers in the area have the same link state database

Chapter 4: Network Layer

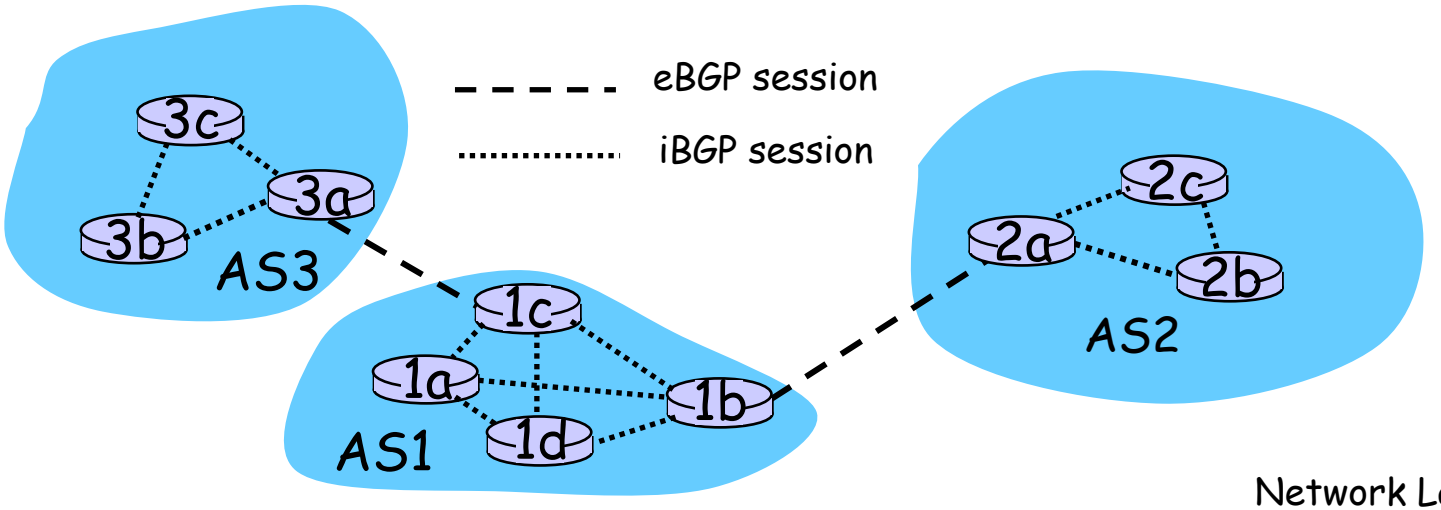
- 4. 1 Introduction
- 4.2 Virtual circuit and datagram networks
- 4.3 What's inside a router
- 4.4 IP: Internet Protocol
  - Datagram format
  - IPv4 addressing
  - ICMP
  - IPv6
- 4.5 Routing algorithms
  - Link state
  - Distance Vector
  - Hierarchical routing
- 4.6 Routing in the Internet
  - RIP
  - OSPF
  - BGP
- 4.7 Broadcast and multicast routing

Internet inter-AS routing: BGP

- BGP (Border Gateway Protocol): *the* de facto standard
- BGP provides each AS a means to:
  1. Obtain subnet reachability information from neighboring ASs.
  2. Propagate reachability information to all AS-internal routers.
  3. Determine "good" routes to subnets based on reachability information and policy.
- allows subnet to advertise its existence to rest of Internet: *"I am here"*

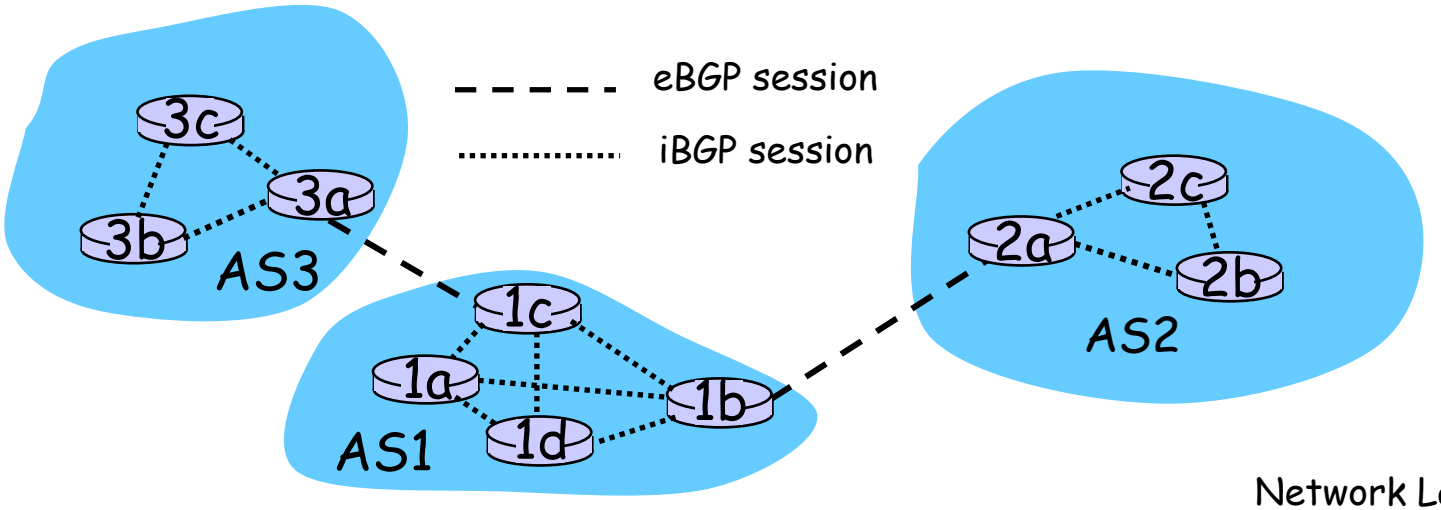
BGP basics

- pairs of routers (BGP peers) exchange routing info over semi-permanent TCP connections: BGP sessions
  - BGP sessions need not correspond to physical links.
- when AS2 advertises a prefix to AS1:
  - AS2 *promises* it will forward datagrams towards that prefix.
  - AS2 can aggregate prefixes in its advertisement



Distributing reachability info

- using eBGP session between 3a and 1c, AS3 sends prefix reachability info to AS1.
  - 1c can then use iBGP do distribute new prefix info to all routers in AS1
  - 1b can then re-advertise new reachability info to AS2 over 1b-to-2a eBGP session
- when router learns of new prefix, it creates entry for prefix in its forwarding table.



Path attributes & BGP routes

- advertised prefix includes BGP attributes.
  - prefix + attributes = "route"
- two important attributes:
  - AS-PATH: contains ASs through which prefix advertisement has passed: e.g, AS 67, AS 17
  - NEXT-HOP: indicates specific internal-AS router to next-hop AS. (may be multiple links from current AS to next-hop-AS)
- when gateway router receives route advertisement, uses import policy to accept/decline.

BGP route selection

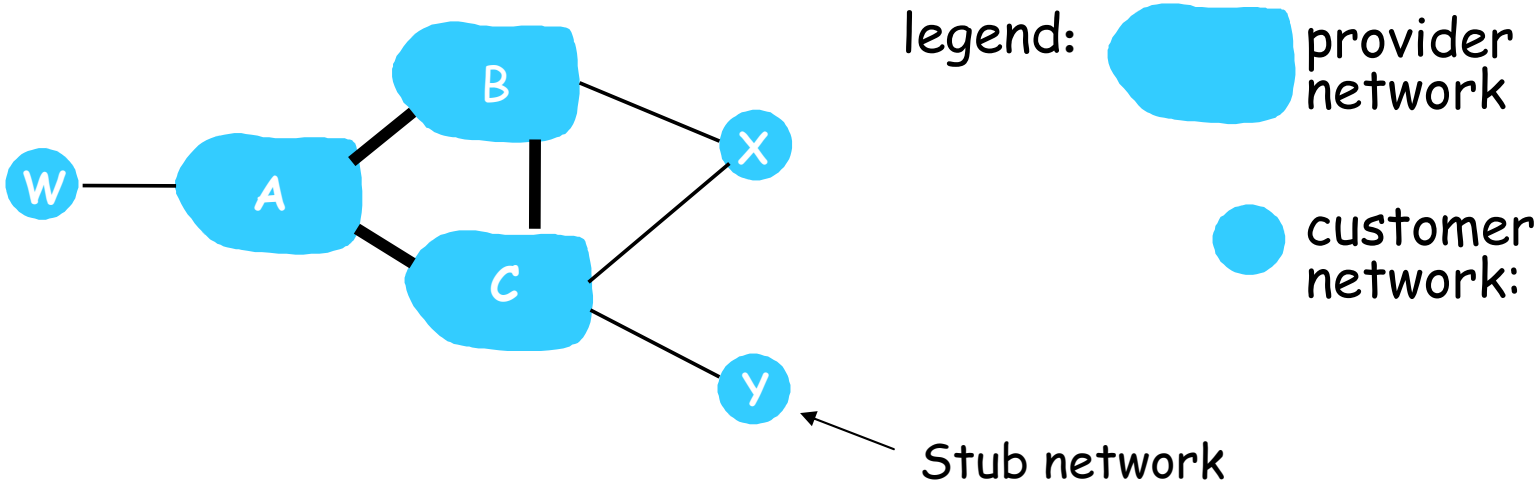
- router may learn about more than 1 route to some prefix. Router must select route.
- elimination rules (in priority order):
  1. local preference value attribute: policy decision
  2. (in case of same preference) shortest AS-PATH
  3. (in case of same preferehce and AS-PATH length) closest NEXT-HOP router: hot potato routing
  4. additional criteria to break the tie

BGP messages

- BGP messages exchanged using TCP.
- BGP messages:
  - OPEN: opens TCP connection to peer and authenticates sender
  - UPDATE: advertises new path (or withdraws old)
  - KEEPALIVE keeps connection alive in absence of UPDATES; also ACKs OPEN request
  - NOTIFICATION: reports errors in previous msg; also used to close connection

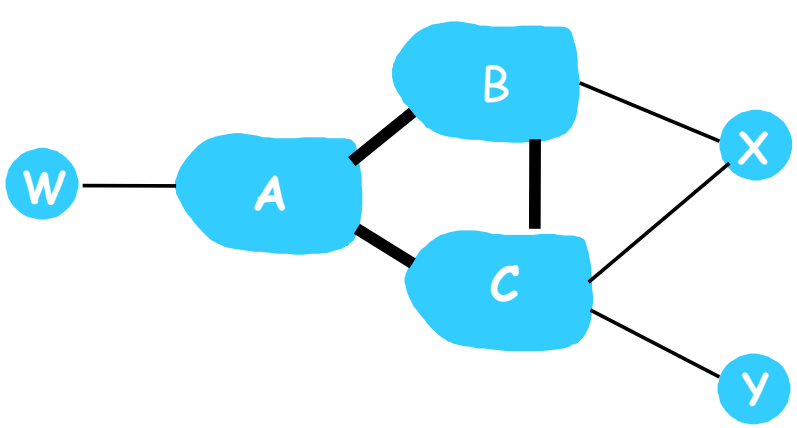
Sent periodically or in case of selected routes changes

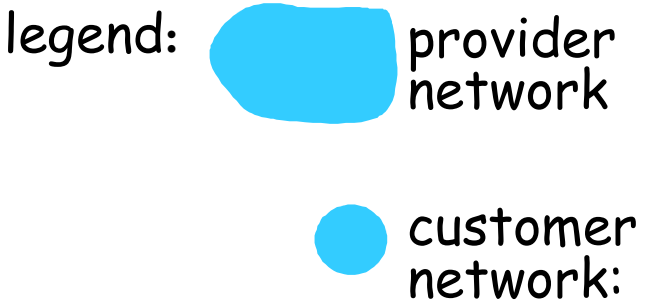
BGP routing policy



- A,B,C are provider networks
- X,W,Y are customer (of provider networks)
- X is dual-homed: attached to two networks
  - X does not want to route from B via X to C
  - .. so X will not advertise to B a route to C

BGP routing policy (2)





- A advertises path AW to B
- B advertises path BAW to X (who is its client)
- Should B advertise path BAW to C?
  - No way! B gets no "revenue" for routing CBAW since neither W nor C are B's customers
  - B wants to force C to route to w via A
  - B wants to route *only* to/from its customers!
  - Peering agreements amongs pairs of ISP possible to solve this problem

Decision process

- The decision process selects routes for subsequent advertisement applying the policies in the local Policy Information Base (PIB) to the routes stored in its Adj-RIBs_In (Incoming Routing Information Base)
- A function takes as argument the attributes of a give route and returns a) either a non negative integer identifying the degree of preference for the route or b) a value indicating the route is inelegible

Why different Intra- and Inter-AS routing ?

Policy:

- Inter-AS: admin wants control over how its traffic routed, who routes through its net.
- Intra-AS: single admin, so no policy decisions needed

Scale:

- hierarchical routing saves table size, reduced update traffic

Performance:

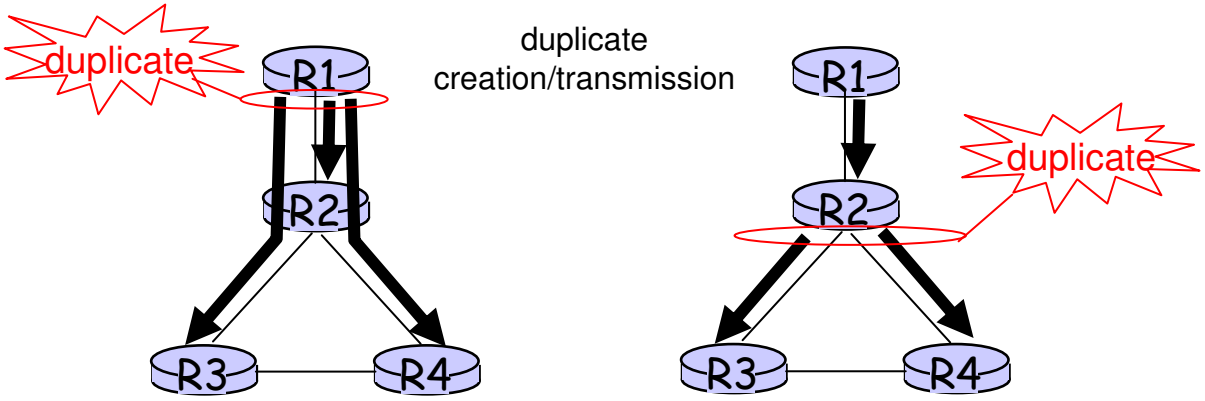
- Intra-AS: can focus on performance
- Inter-AS: policy may dominate over performance

Chapter 4: Network Layer

- 4. 1 Introduction
- 4.2 Virtual circuit and datagram networks
- 4.3 What's inside a router
- 4.4 IP: Internet Protocol
  - Datagram format
  - IPv4 addressing
  - ICMP
  - IPv6
- 4.5 Routing algorithms
  - Link state
  - Distance Vector
  - Hierarchical routing
- 4.6 Routing in the Internet
  - RIP
  - OSPF
  - BGP
- 4.7 Broadcast and multicast routing

Broadcast Routing

- deliver packets from source to all other nodes
- source duplication is inefficient:

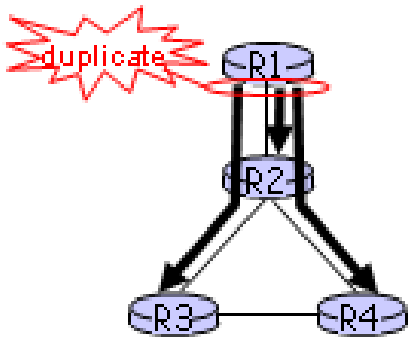


source duplication

in-network duplication

- source duplication: how does source determine recipient addresses?

Unicast ad N vie

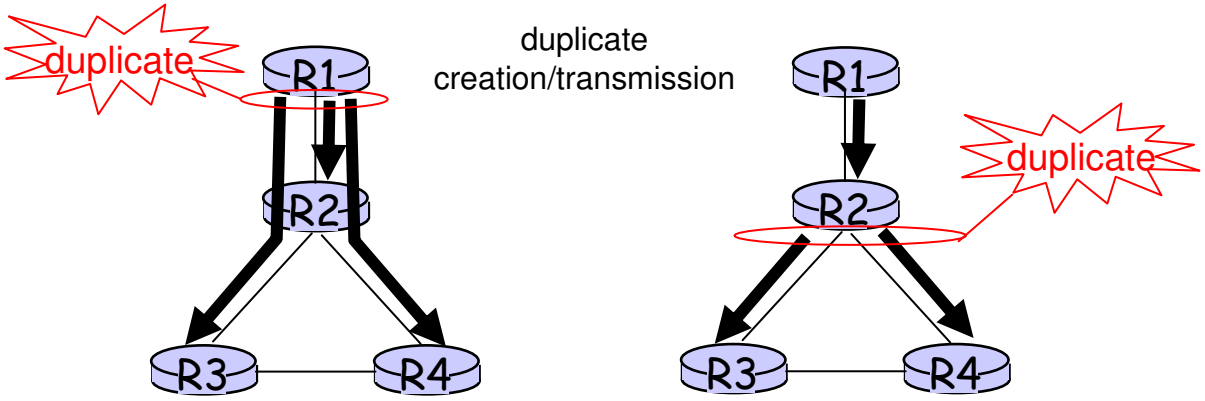


source duplication

- Inefficiente
  - Un singolo collegamento attraversato da N copie del messaggio se il nodo origine è connesso al resto della rete tramite un unico collegamento
- Indirizzi di tutte le destinazioni devono essere noti al mittente
  - altri meccanismi protocollari sono richiesti
- Broadcast può essere usato per inoltrare informazioni di topologia in una situazione in cui le rotte non sono ancora note
  - es. OSPF

Broadcast Routing

- deliver packets from source to all other nodes
- source duplication is inefficient:



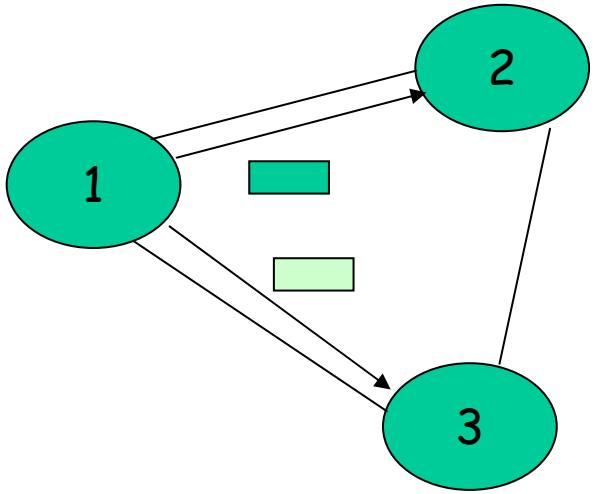
source duplication

in-network duplication

- source duplication: how does source determine recipient addresses?

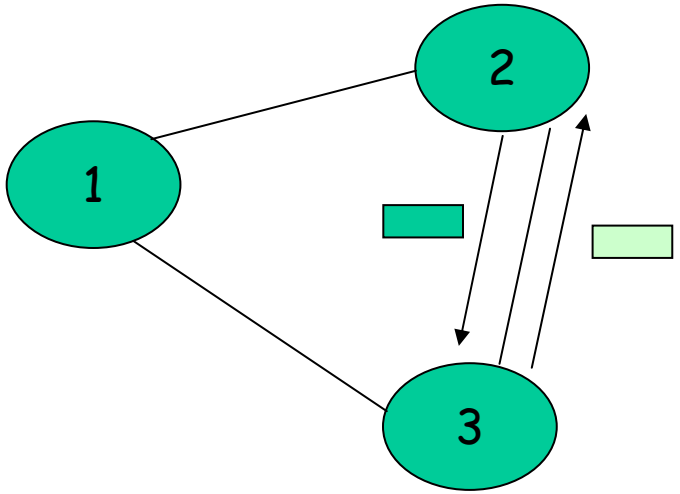
In-network duplication

- flooding: when node receives brdcst pckt, sends copy to all neighbors EXCEPT the one from which the pckt was received
  - Problems: cycles & broadcast storm



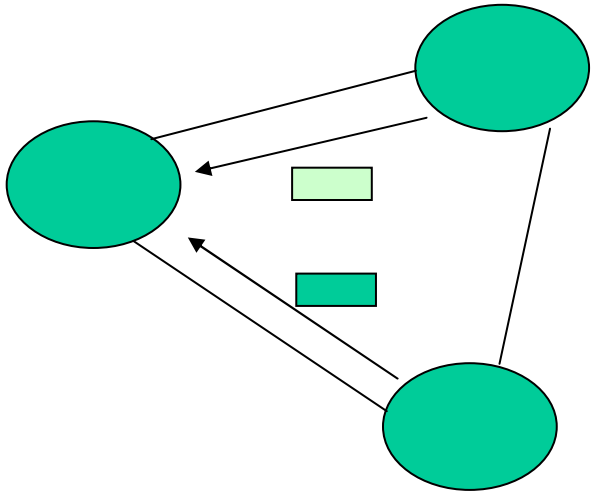
In-network duplication

- flooding: when node receives brdcst pckt, sends copy to all neighbors
  - Problems: cycles & broadcast storm



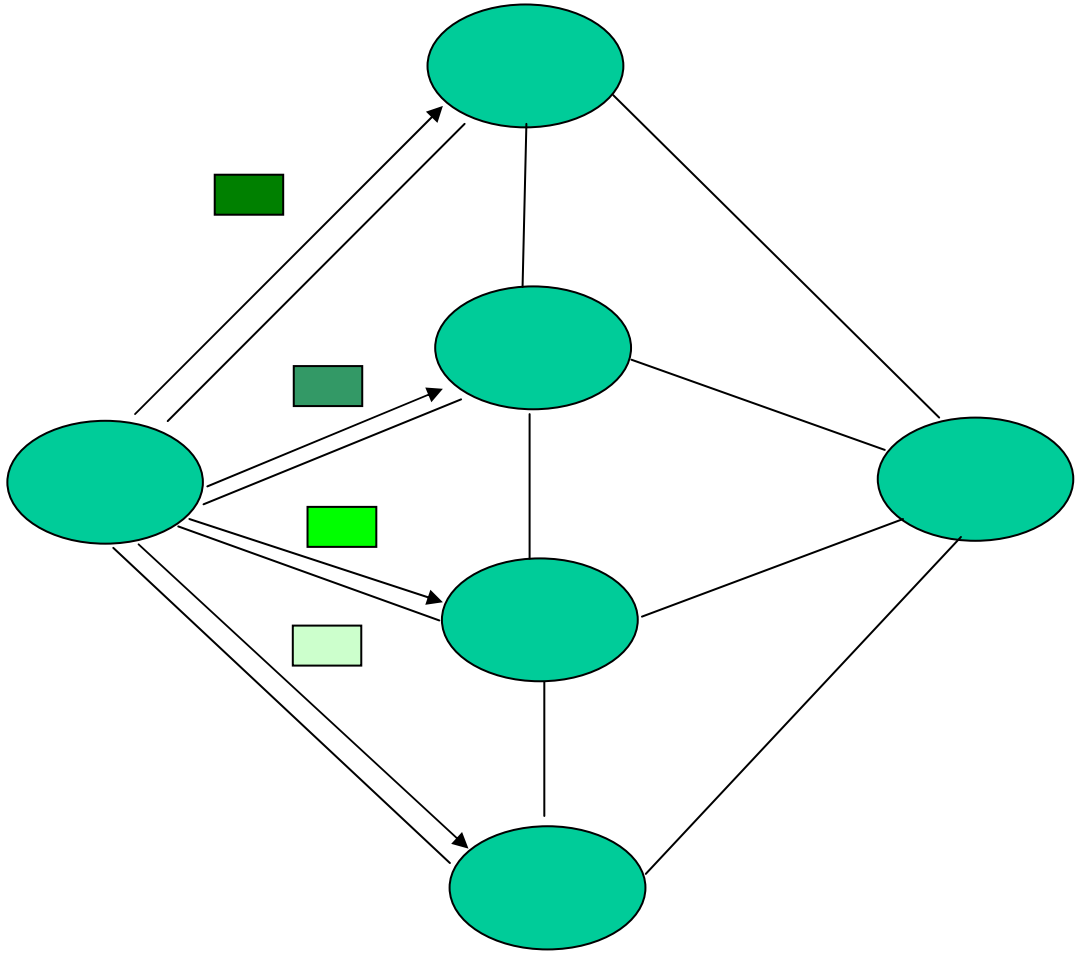
In-network duplication

- flooding: when node receives brdcst pckt, sends copy to all neighbors
  - Problems: cycles & broadcast storm

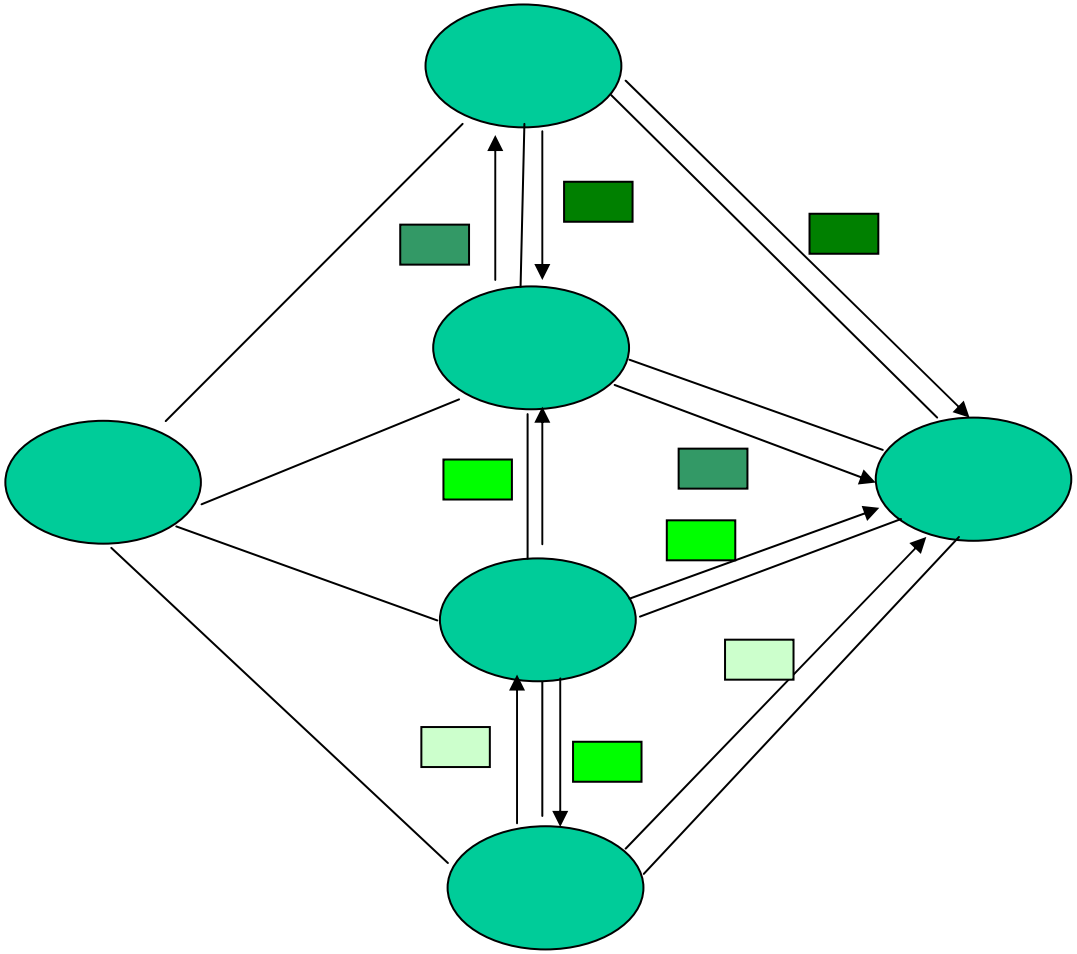


E ricominciamo come nella prima situazione
Bisogna saper distinguere tra quando mandiamo un nuovo messaggio e quando stiamo ritrasmettendo qualcosa che abbiamo già visto
→ Sequence numbers!

Broacast storm

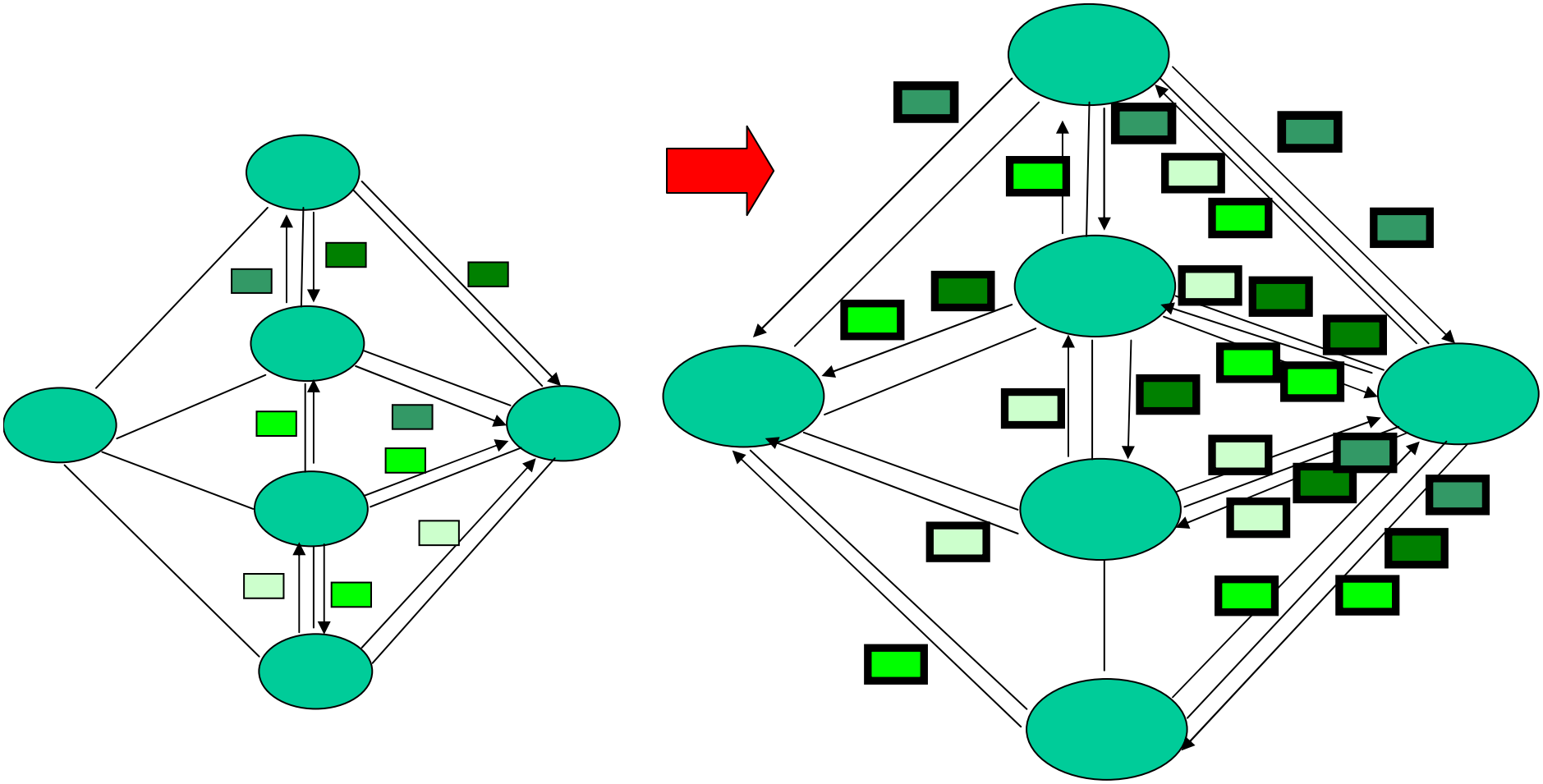


Broacast storm



Broacast storm

Il numero di pacchetti in rete cresce significativamente!!



Controlled flooding

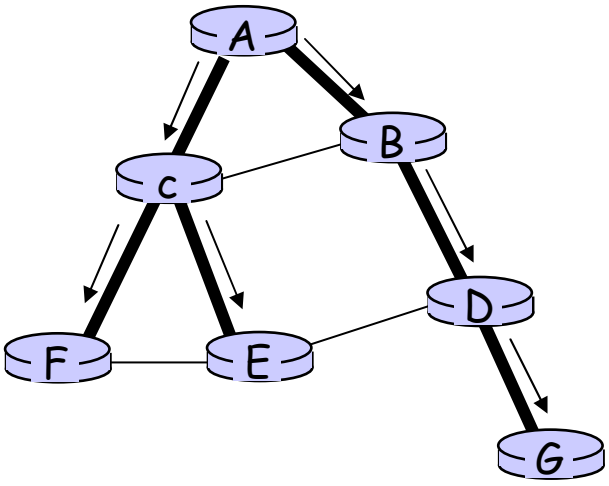
- Il nodo origine pone il proprio indirizzo ed il numero di sequenza nei pacchetti che invia in broadcast
- Ciascun nodo mantiene una lista di ID origine, SEQN per i broadcast ricevuti, trasmesso o inoltrato
- Se riceve un pacchetto broadcast per prima cosa verifica se <ID, SEQN> compare nella lista dei pacchetti *già gestiti*
  - Se si scarta
  - Altrimenti riinvia su tutte le interfacce tranne quella da cui ha ricevuto

Controlled flooding, altre opzioni

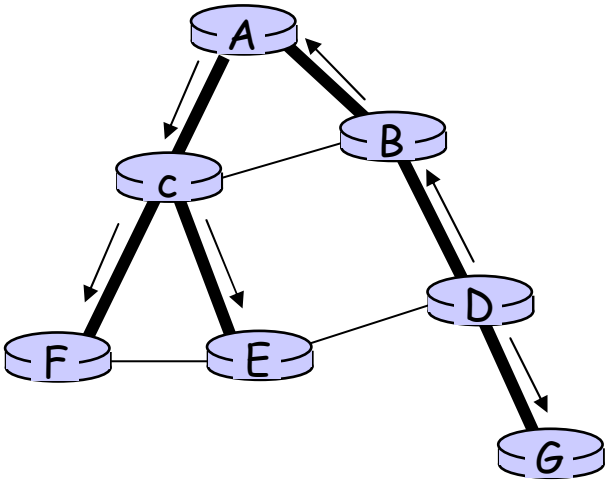
- Reverse path forwarding (RPF): only forward pckt (on all links but the one from which the packet was received) if it arrived on shortest path between node and source

Spanning Tree

- First construct a spanning tree
- Nodes forward copies only along spanning tree



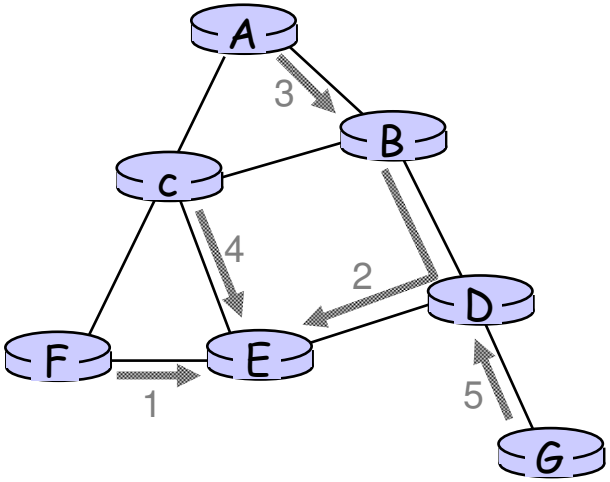
(a) Broadcast initiated at A



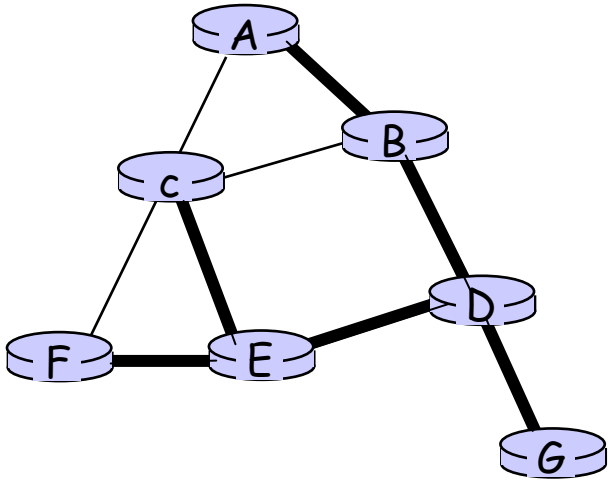
(b) Broadcast initiated at D

Spanning Tree: Creation

- Center node
- Each node sends unicast join message to center node
  - Message forwarded until it arrives at a node already belonging to spanning tree



(a) Stepwise construction of spanning tree



(b) Constructed spanning tree

Multicasting

- Molte applicazioni richiedono il trasferimento di pacchetti da uno o più mittenti ad un gruppo di destinatari
  - trasferimento di un aggiornamento SW su un gruppo di macchine
  - streaming (audio/video) ad un gruppo di utenti o studenti
  - applicazioni con dati condivisi (lavagna elettronica condivisa da più utenti)
  - aggiornamento di dati (adnamento di borsa)
  - giochi multi-player interattivi
  - ...

Indirizzamento Multicast

- L'identificatore che rappresenta un gruppo multicast è un indirizzo IP multicast di classe D
- Come ci si affilia ad un indirizzo multicast? Come vengono gestiti i cambiamenti dinamici (join/remove) nel gruppo?
  - Gestione dinamica del gruppo OLTRE a
  - Algoritmi per la consegna delle informazioni ad un gruppo multicast

IGMP Internet Group Management Protocol

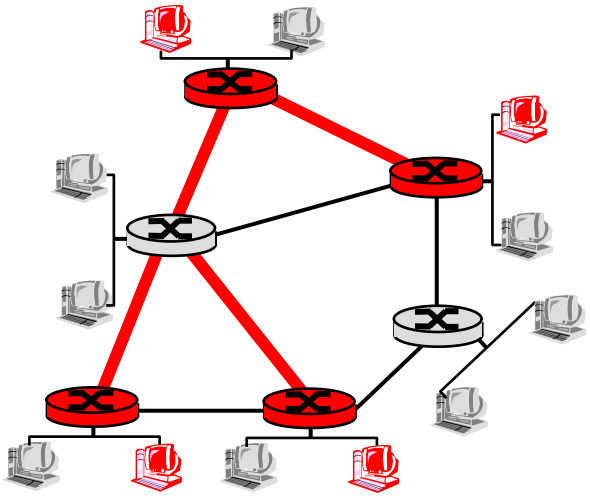
- Messaggi incapsulati in datagrammi IP, con IP protocol number 2
- Mandati con TTL a 1
- Messaggi IGMP
  - Type (8bit) Query (richiesta dal router)/ Membership Report (risposta dagli host)/ Leave group (ma anche possible timeout + mancata risposta alla richiesta del router→ soft state)
- Max Response Time (per rispondere ad una query)
- Checksum
- Group Address (0 se si manda una general query, indirizzo IP del gruppo nel caso di una group specific query con cui si richiede chi sia affiliato a quel gruppo)

IGMP Internet Group Management Protocol

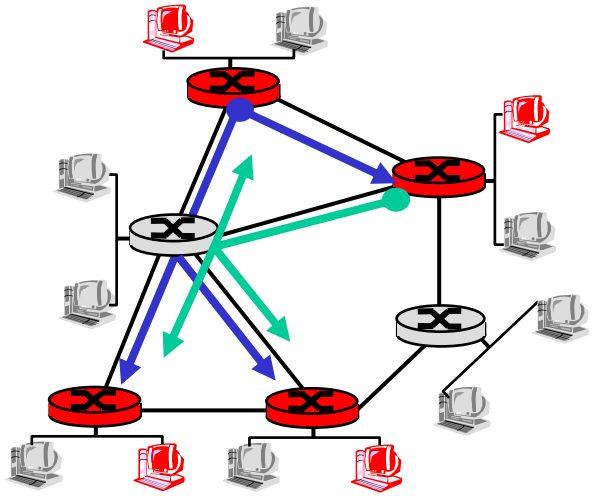
- IGMP consente ad un router di imparare quali gruppi multicast hanno affiliati sulle sottoreti connesse a ciascuna delle loro interfacce
- Un router multicast tiene una lista per ciascuna sottorete dei multicast group (multicast group membership→ almeno un elemento del gruppo fa parte della sottorete) con un timer per membership
  - la membership deve essere aggiornata da report inviati prima della scadenza del timer
  - pò essere anche aggiornata tramite messaggi di leave espliciti

Multicast Routing: Problem Statement

- ***Goal:*** find a tree (or trees) connecting routers having local mcast group members
  - *tree:* not all paths between routers used
  - *source-based:* different tree from each sender to rcvrs
  - *shared-tree:* same tree used by all group members



Shared tree



Source-based trees

Approaches for building mcast trees

Approaches:

- source-based tree: one tree per source
  - shortest path trees
  - reverse path forwarding
- group-shared tree: group uses one tree
  - minimal spanning (Steiner)
  - center-based trees

…we first look at basic approaches, then specific protocols adopting these approaches

Instradamento multicast con albero condiviso dal gruppo

- Stesso approccio visto per calcolare un albero di ricoprimento per il flooding
- **I router multicast connessi a sottoreti con elementi del gruppo si affiliano tramite un unicast message mandato all'iniziatore dell'albero**
- **Tutti i nodi attraversati prima di unirsi ad un elemento già parte dell'albero di multicast fanno parte dell'albero**

R3

Reverse Path Forwarding

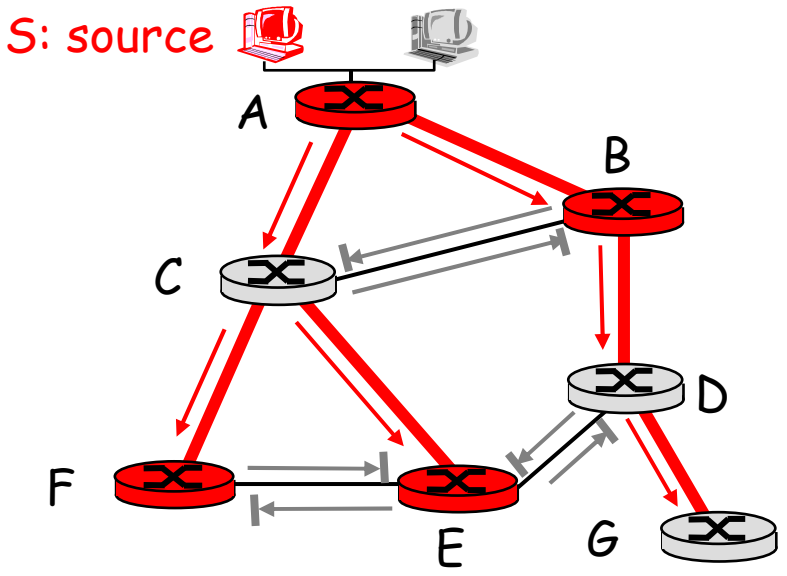
- rely on router's knowledge of unicast shortest path from it to sender
- each router has simple forwarding behavior:

if (mcast datagram received on incoming link on shortest path back to center)

then flood datagram onto all outgoing links

else ignore datagram

Reverse Path Forwarding for source based multicasting: example



LEGEND

router with attached group member

router with no attached group member

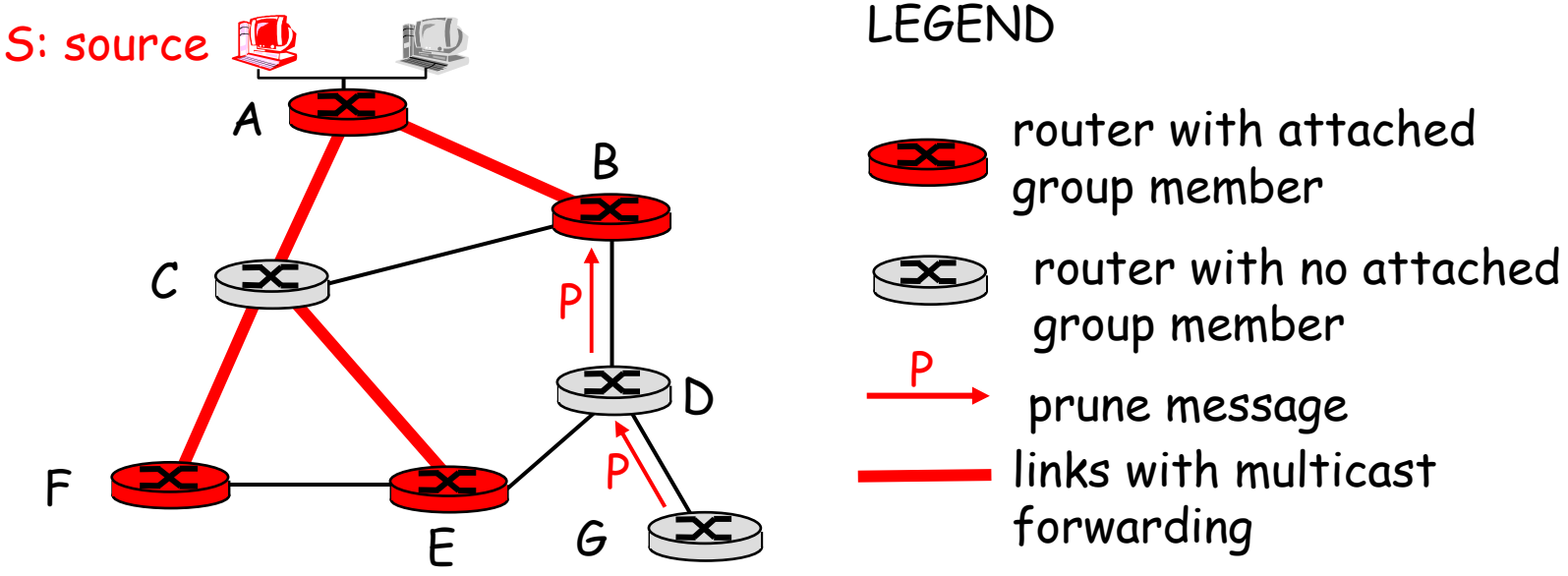
datagram will be forwarded

datagram will not be forwarded

- result is a source-specific *reverse* SPT
  - may be a bad choice with asymmetric links
  - unneeded forwarding e.g., from D to G

Reverse Path Forwarding: pruning

- forwarding tree contains subtrees with no mcast group members
  - no need to forward datagrams down subtree
  - "prune" msgs sent upstream by router with no downstream group members



Shared-Tree: Steiner Tree

- Steiner Tree: minimum cost tree connecting all routers with attached group members
- problem is NP-complete
- excellent heuristics exists
- not used in practice:
  - computational complexity
  - information about entire network needed
  - monolithic: rerun whenever a router needs to join/leave

Center-based trees

- single delivery tree shared by all
- one router identified as *"center"* of tree
- to join:
  - edge router sends unicast *join-msg* addressed to center router
  - *join-msg* "processed" by intermediate routers and forwarded towards center
  - *join-msg* either hits existing tree branch for this center, or arrives at center
  - path taken by *join-msg* becomes new branch of tree for this router

Chapter 4: summary

- 4. 1 Introduction
- 4.2 Virtual circuit and datagram networks
- 4.3 What's inside a router
- 4.4 IP: Internet Protocol
  - Datagram format
  - IPv4 addressing
  - ICMP
  - IPv6
- 4.5 Routing algorithms
  - Link state
  - Distance Vector
  - Hierarchical routing
- 4.6 Routing in the Internet
  - RIP
  - OSPF
  - BGP
- 4.7 Broadcast and multicast routing