

Lecture 7

Multicasting

Internet Multicasting (1)



Fundamental issues in multicast

- **Joining and leaving a group**
 - Multicast sessions learning
 - Group members discovery
 - Dynamic group membership
- **Efficient transmission of multicast traffic**
 - Resource optimization
 - Delivery tree maintenance
- **Time-sensitive delivery of multicast traffic**
 - Data sequence maintenance
 - Synchronization

Internet Multicasting (2)

- Guaranteed arrival of multicast traffic
 - RTP (real-time transport protocol)
 - RMP (reliable multicast protocols)
- Scalability
 - Feedback implosions
 - The use of groups

Internet Multicasting (3)

System

- Group membership protocol
 - Internet Group Management Protocol (IGMP)
- Multicast routing mechanism
 - Distance vector multicast routing protocol (DVMRP)
 - Multicast open shortest path first (MOSPF)
 - Core-based tree (CBT)
 - Protocol independent multicast sparse mode (PIM-SM)

IGMP (1)

References

- [RFC-1112] "Host Extensions for IP Multicasting"
- [RFC-2236] "Internet Group Management Protocol, Version 2"
- [RFC-3376] "Internet Group Management Protocol, Version 3"

IGMP (2)

IGMP v1

- Multicast router: periodically sends a query message to the all-hosts address (224.0.0.1)
- A host sends a report in reply on a per group basis, thereby refreshing the tentative states
- IGMP v1 supports suppression for periodical refresh report messages
- IGMP v1 hosts send unsolicited reports upon joining a group, but leaves the group silently

IGMP (3)

IGMP v2

- IGMP v2 maintains two types of query and three types of report
 - Query
 - General query
 - Group-specific query
 - Report
 - Join
 - Leave
 - Refresh
- Periodical refresh report suppression is supported as well
- The approach is to lower leave latency

IGMP (4)

IGMP v3

- IGMP v3 maintains three types of query: general query, group-specific query, and group-and-source specific query
- IGMP v3 maintains four reports: join, leave, state change, and refresh
- No periodical refresh report suppression is supported
- The approach to support source filtering:
 - (group-id, filter mode, source list)
- General query
 - Multicast routers send General Queries periodically to request group membership information from an attached network
 - These queries are used to build and refresh the group membership state of systems on attached networks

IGMP (5)

- Group-specific query
 - A Group-specific query is sent to verify there are no systems that desire reception of the specified group or to "rebuild" the desired reception state for a particular group. Group-Specific Queries are sent when a router receives a State-Change record indicating a system is leaving a group
- Group-and-source-specific query
 - Used to verify there are no systems on a network which desire to receive traffic from a set of sources
- Source filtering is the ability for an individual host to specify the reception of packets sent to a multicast group only from a list of source addresses or to explicitly identify a list of the sources the host does not want to receive from a multicast group
- Eg., include{x,y,z}, exclude{x}, exclude{}, include{}
- IGMP v1/v2 → exclude{}

Introduction-Routing (1/4)



Requirements of routing protocols

- Routing table space minimization
- Control message minimization
- Robustness
- Optimal path construction



Two fundamental routing algorithms

- Distance-vector
- Link-state

Introduction-Routing (2/4)

Scalability

- Exterior gateway protocol (EGP)
 - Applied in the backbone (interconnecting multiple Autonomous Systems)
- Interior gateway protocol (IGP)
 - Applied in an autonomous system (AS)
 - Routers in an AS are controlled by a single administrative authority

Introduction-Routing (3/4)

- ✿ Routers exchange information to construct multicast delivery tree
- ✿ Pruning
 - Due to dynamic membership
 - Removing one branch from the delivery tree
 - Important to tree maintenance
- ✿ Dense mode vs. sparse mode multicast routing
 - Definition: number of networks with group members
 - Dense mode multicast routing: data-driven
 - Sparse mode multicast routing: receiver-initiated

Introduction-Routing (4/4)



Categories

- Multicast tree constructed
 - Shortest path tree
 - Constrained tree (for QoS purpose)
- Multicast tree rooted
 - Source-based
 - A tree rooted as a source node is constructed and connected to every member in the multicast group
 - E.g., DVMRP, MOSPF, PIM-DM
 - Centered-based (shared tree)
 - One node for each group is selected as the core (or termed a rendezvous point, RP) for the group. A tree rooted at the core is then constructed to span all the group members
 - E.g., CBT and PIM-SM

DVMRP (1/5)

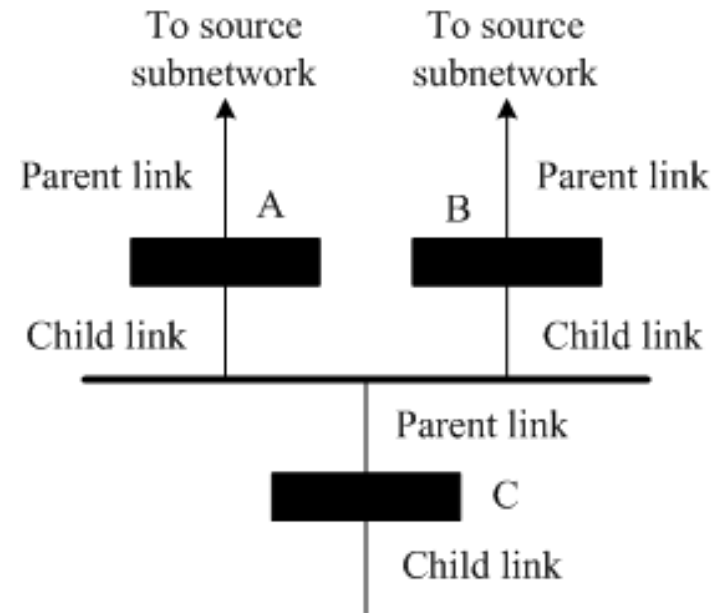
- ✿ Source-based multicast delivery tree
- ✿ Distance vector routing
 - Bellman-Ford algorithm
 - Multicast extension from DVRP (RIP)
 - Based on RPM
 - For the first multicast packet: send over the tree anyway
 - Send “prune message” if none wanna receive packets
 - Send “graft message” when a new group member joins in

DVMRP (2/5)



DVMRP router functions

- The problem of redundant links
- Dominant router
- Subordinate router
- How to determine?
 - With less metric or lower IP address
 - Poison-reverse routing updates
- It's possible that a different router may be the dominant router for each source



DVMRP (3/5)

- ✿ When needed, will a subordinate router send a prune message to the dominate router?
- ✿ Routing and forwarding
 - Routing table: periodically exchange routing table update messages with multicast capable neighbors
 - Forwarding table: multicast routing table + known groups + received prune messages

DVMRP (4/5)

Sample DVMRP Routing Table

Source Subnet	Subnet Mask	From-Gateway	Metric	Status	TTL	InPort	OutPorts
128.1.0.0	255.255.0.0	128.7.5.2	3	up	200	1	2,3
128.2.0.0	255.255.0.0	128.7.5.2	5	up	150	2	1
128.3.0.0	255.255.0.0	128.6.3.1	2	up	150	2	2,3
128.4.0.0	255.255.0.0	128.6.3.1	4	up	200	1	2

Source Subnet: a subnetwork which is a source of multicast datagrams

Subnet Mask: the subnet mask associated with the Source Prefix

From-Gateway: the previous hop router leading back toward a particular Source Prefix

Metric: cost associated with that port

Status: current availability of port (up or down)

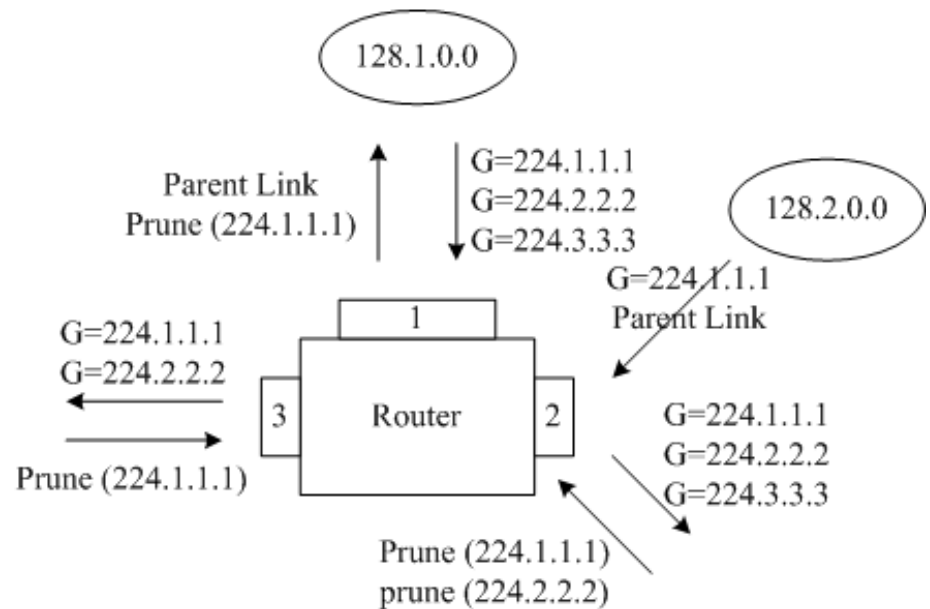
TTL: a time-to-live value is used for table management;
it indicates the number of seconds before an entry is removed from the routing table

InPort: router's interfaces for incoming traffic

OutPort: router's interfaces for outgoing traffic

DVMRP (5/5)

Message flows for sample DVMRP forwarding table



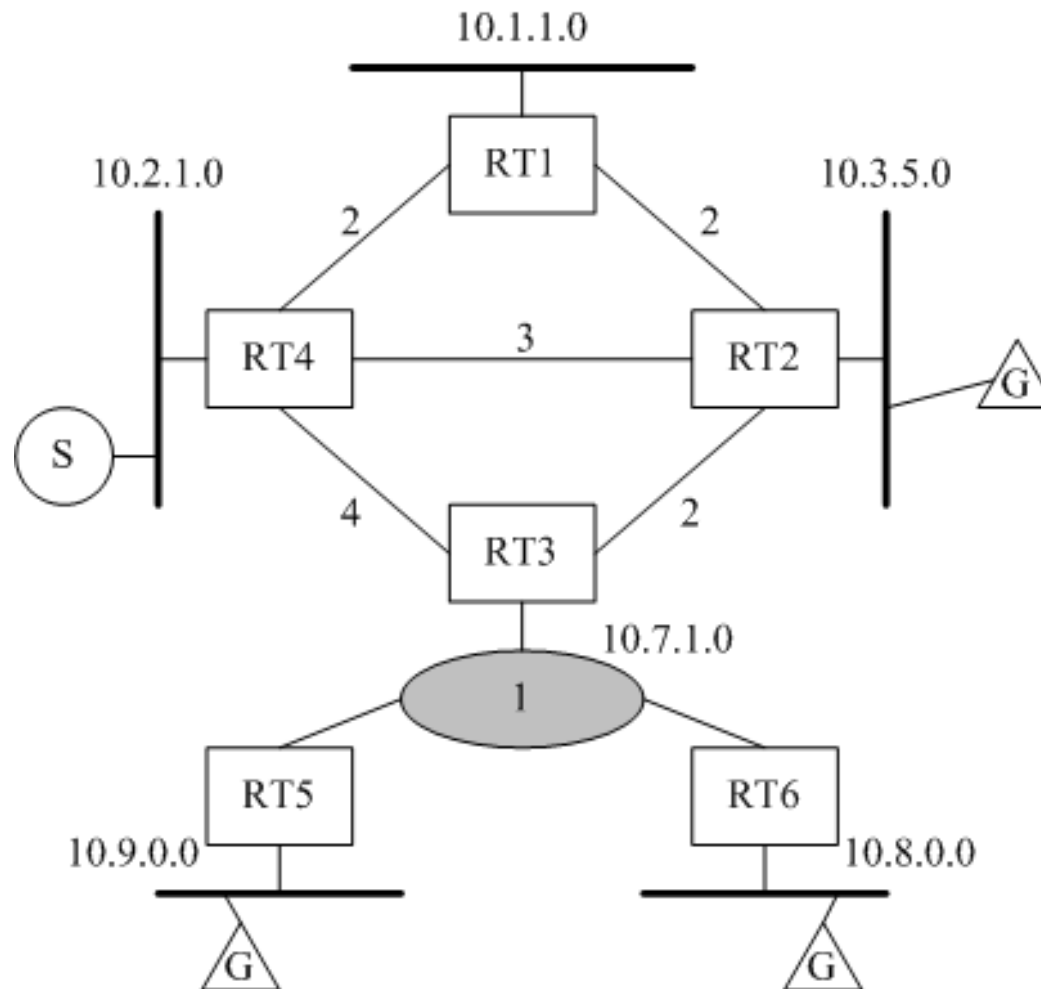
Sample DVMRP Forwarding Table

Source Subnet	Multicast Group	TTL	InPort	OutPorts
128.1.0.0	224.1.1.1	200	1 Pr	2p 3p
	224.2.2.2	100	1	2p 3
	224.3.3.3	250	1	2
128.2.0.0	224.1.1.1	150	2	2p 3p

MOSPF (1/7)

- ✿ Source-based multicast delivery tree
- ✿ Link state routing
 - Dijkstra algorithm
 - Multicast extension to LSA (OSPF)
 - Group-membership-LSA (link-state advertisement)
 - Trees are built on-demand to conserve CPU and memory resources

MOSPF (2/7)



MOSPF (3/7)

Basic operations

- MOSPF routers maintain a local group database
- For a given multicast datagram, all routers within an OSPF area calculate the same source-based shortest-path delivery tree
- No need to flood the first multicast datagram to all routers in an area
- “On-demand” construction spreads calculations over time

MOSPF (4/7)



Protocol mechanisms

- Link state database
- Local group database: labeling the serving MOSPF router
 - Reduce the size of the link state database
 - Streamlining the MOSPF routing calculation
- Forwarding cache
 - (source, group)
 - Upstream interface, downstream interface
 - Only the routers that are parts of a particular tree maintain such info
 - This cache is not aged or periodically refreshed until
 - Topology changes
 - Group member has changed

MOSPF (5/7)

Sample MOSPF Forwarding Cache

Destination	Source	Upstream	Downstream	TTL
224.1.1.1	128.1.0.2	!1	!2 !3	5
224.1.1.1	128.4.1.2	!1	!2 !3	2
224.1.1.1	128.5.2.2	!1	!2 !3	3
224.2.2.2	128.2.0.3	!1	!1	7

TTL: the minimum number of hops a datagram must cross to reach any of the Dest. Group's members

MOSPF (6/7)



Protocol operation

- **Joining a multicast group**
 - Host: IGMP → Designated router (DR) + backup DR
 - DR: group-membership-LSA → flooding
 - Other routers: update local group database, and may update forwarding cache
- **Leaving the multicast group**
 - Host: IGMP → DR → LSA → update
- **Inter-area multicast forwarder: OSPF area border router (ABR)**
- **Wild-card multicast receiver**

MOSPF (7/7)

- Data forwarding
 - Global topology
 - Link state database + group membership LSA
 - Shortest path tree computation
 - Computation on demand
 - Based on source network and destination multicast group
 - Forwarding cache construction

MOSPF – Complementary (1)

Area v.s. AS

- In OSPF, a single autonomous system (AS) can be divided into smaller groups called *areas*
- (+) Reduce the number of link-state advertisements (LSAs) and other OSPF overhead traffic sent on the network
- (+) Reduce the size of the topology database that each router must maintain

Area border router (ABR)

- Routing devices that belong to more than one area and connect one or more OSPF areas to the backbone area are called *area border routers* (ABRs)
- At least one interface is within the backbone while another interface is in another area
- ABRs also maintain a separate topological database for each area to which they are connected

Autonomous system boundary router (ASBR)

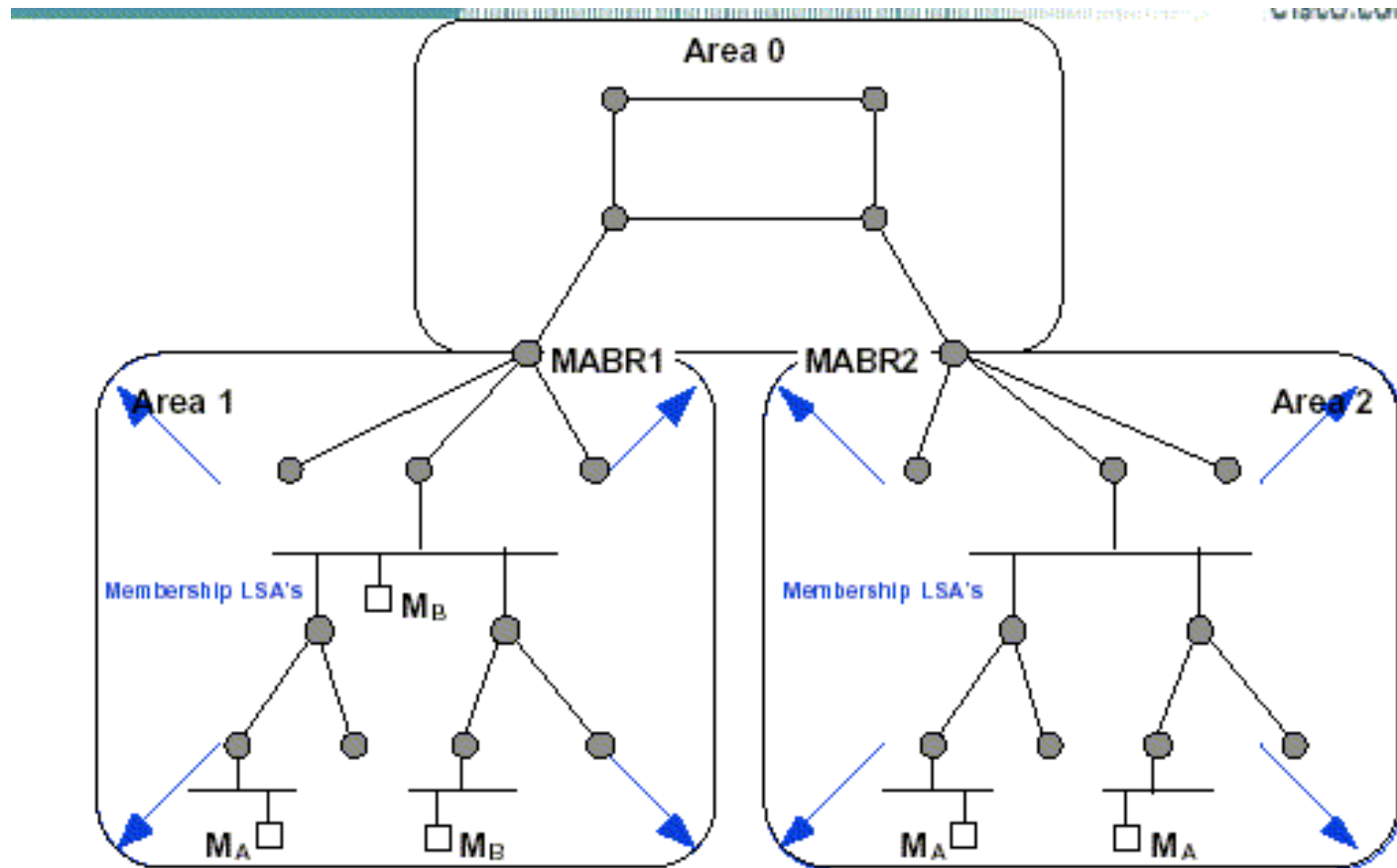
- Routing devices that exchange routing information with routing devices in non-OSPF networks are called *AS boundary routers*

MOSPF – Complementary (2)

- 🌾 Wild-card multicast receiver
- 🌾 Inter-area multicast forwarder
- 🌾 Inter-AS multicast forwarder

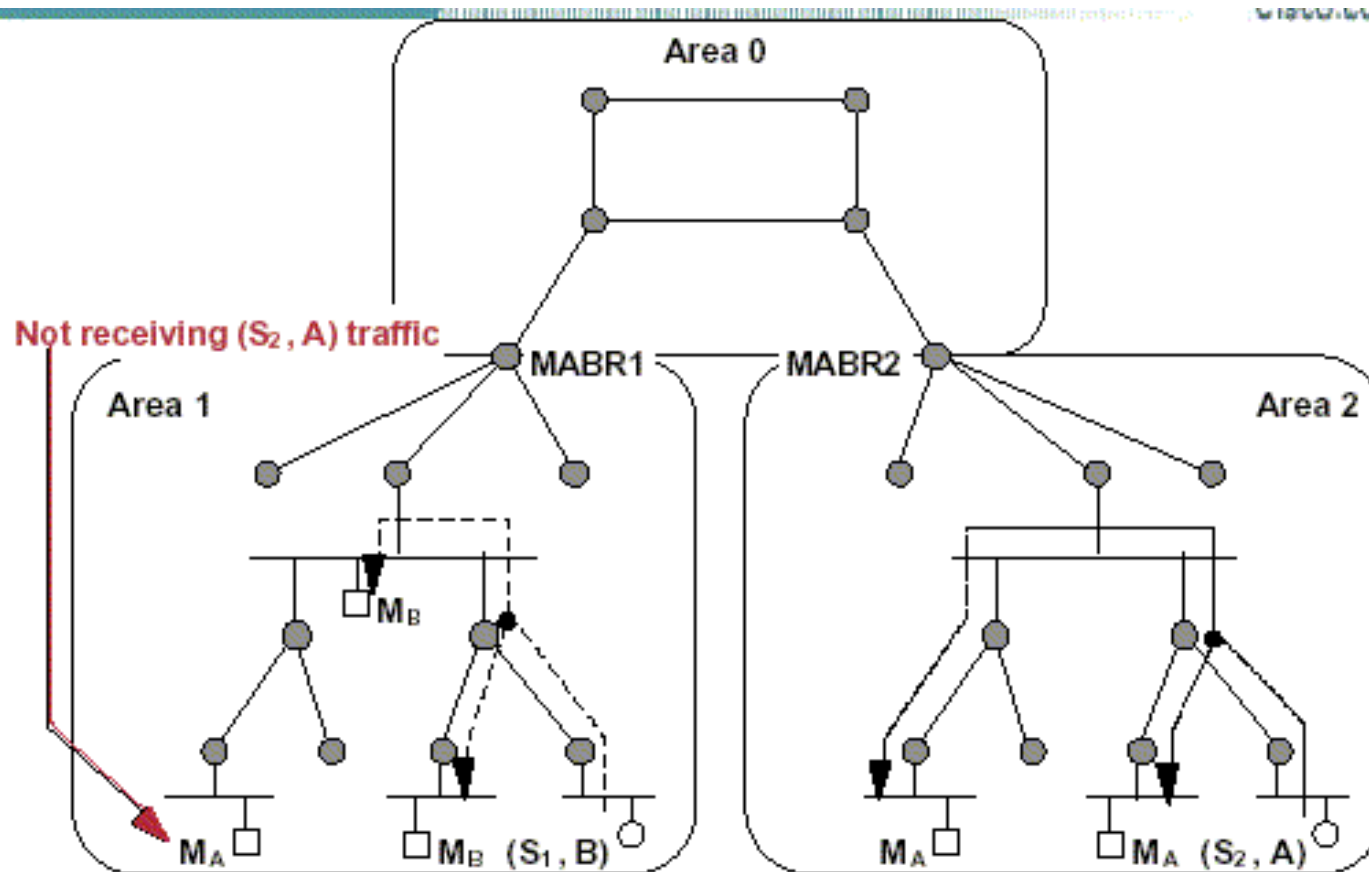
MOSPF – Complementary (3)

- ✿ One AS consists of several areas
- ✿ Within an area, routers flood Group Membership-LSA



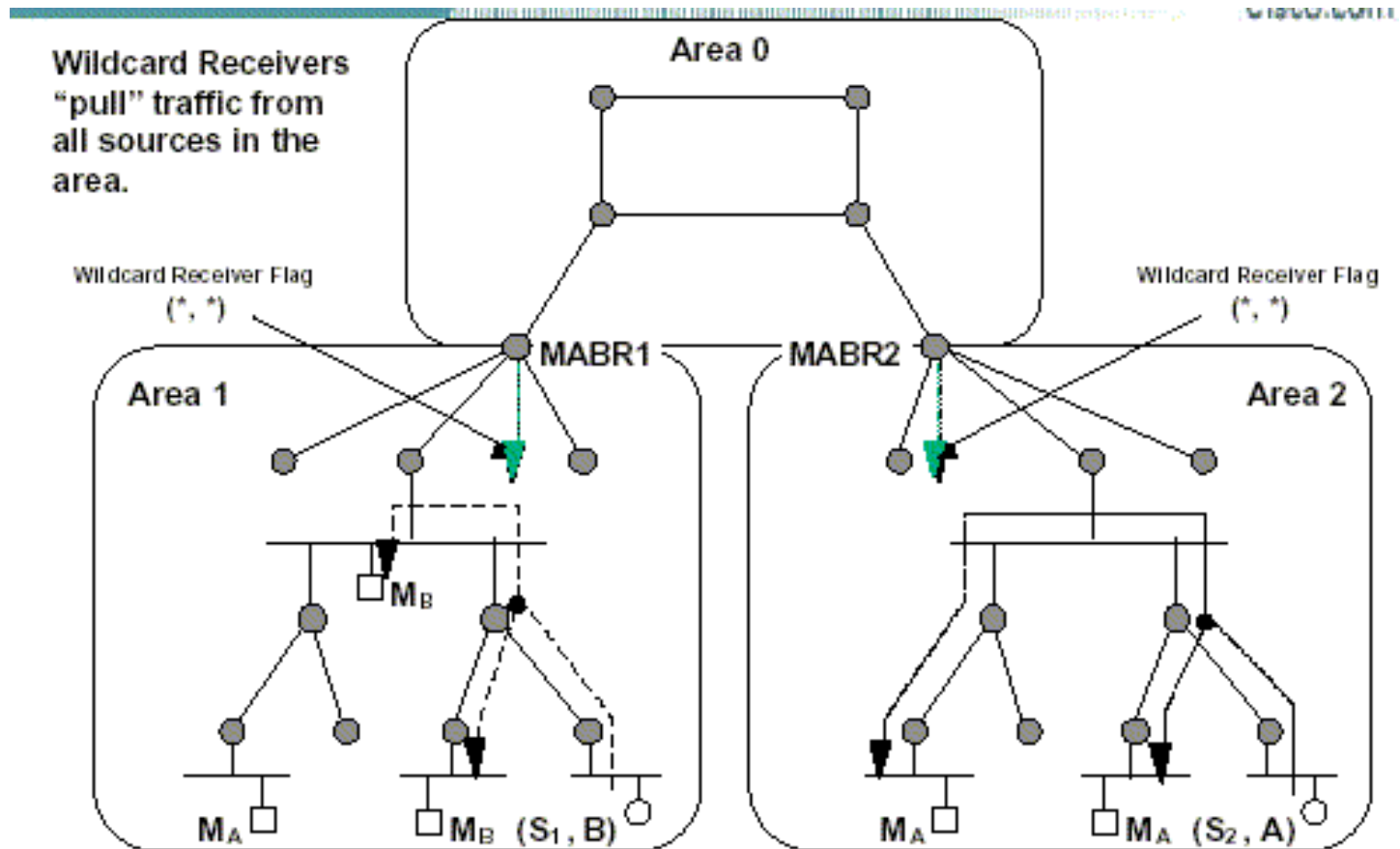
MOSPF – Complementary (4)

- ✿ Routers within an area construct source-network trees for the multicast traffic forwarding



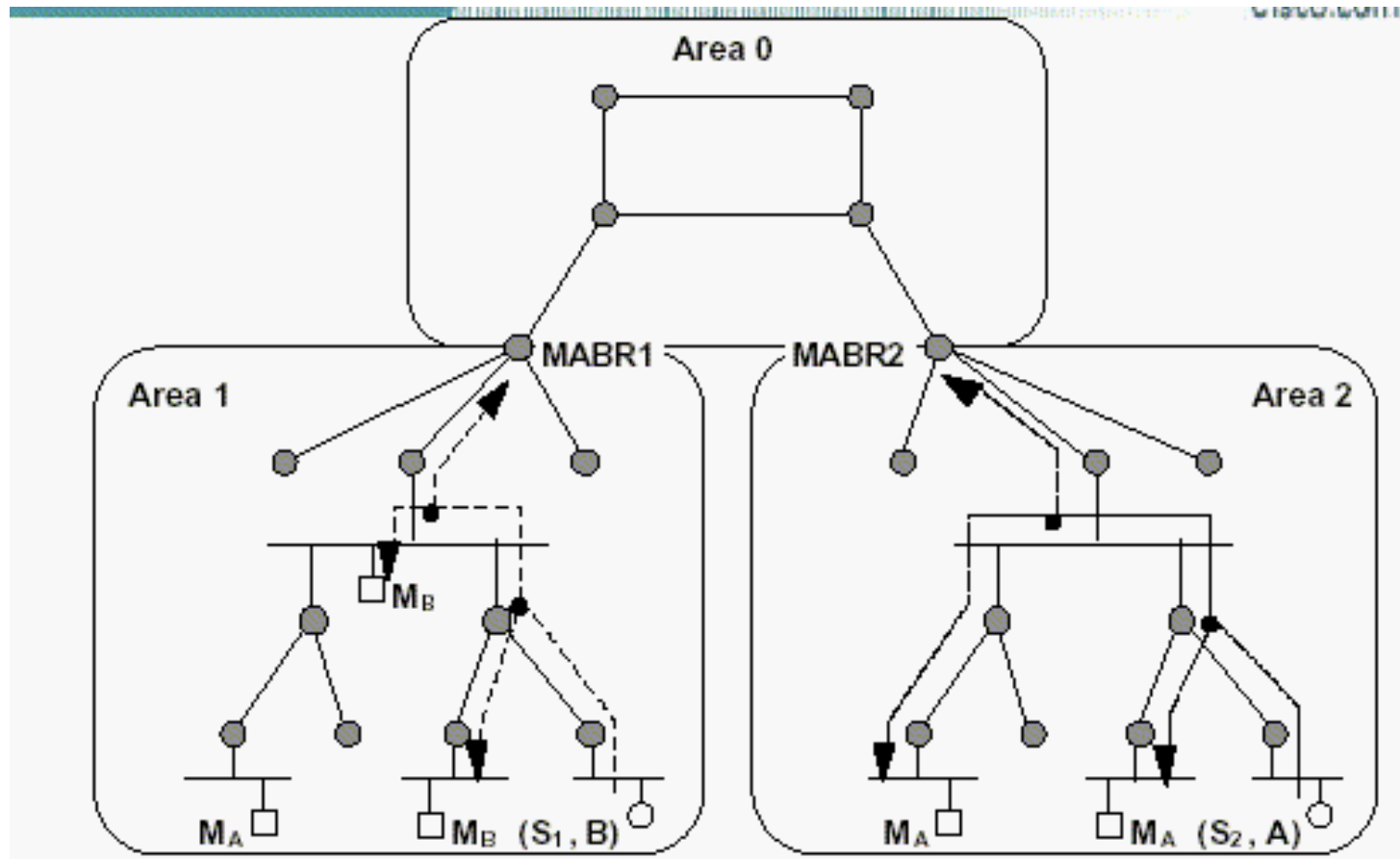
MOSPF – Complementary (5)

- Inter-area traffic deliveries through MABRs (wildcard multicast receiver)



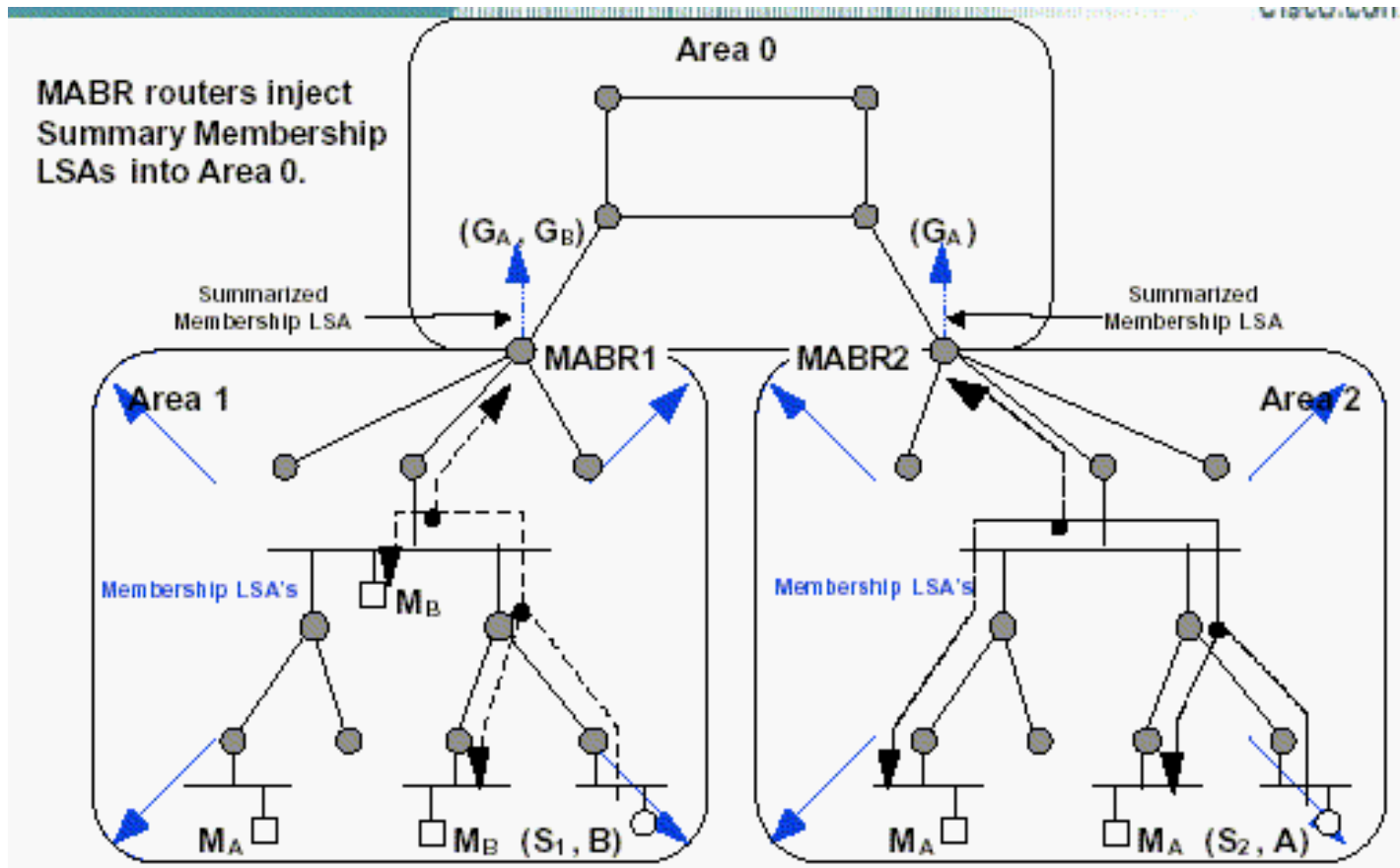
MOSPF – Complementary (6)

- 🌿 Multicast data delivered to MABRs



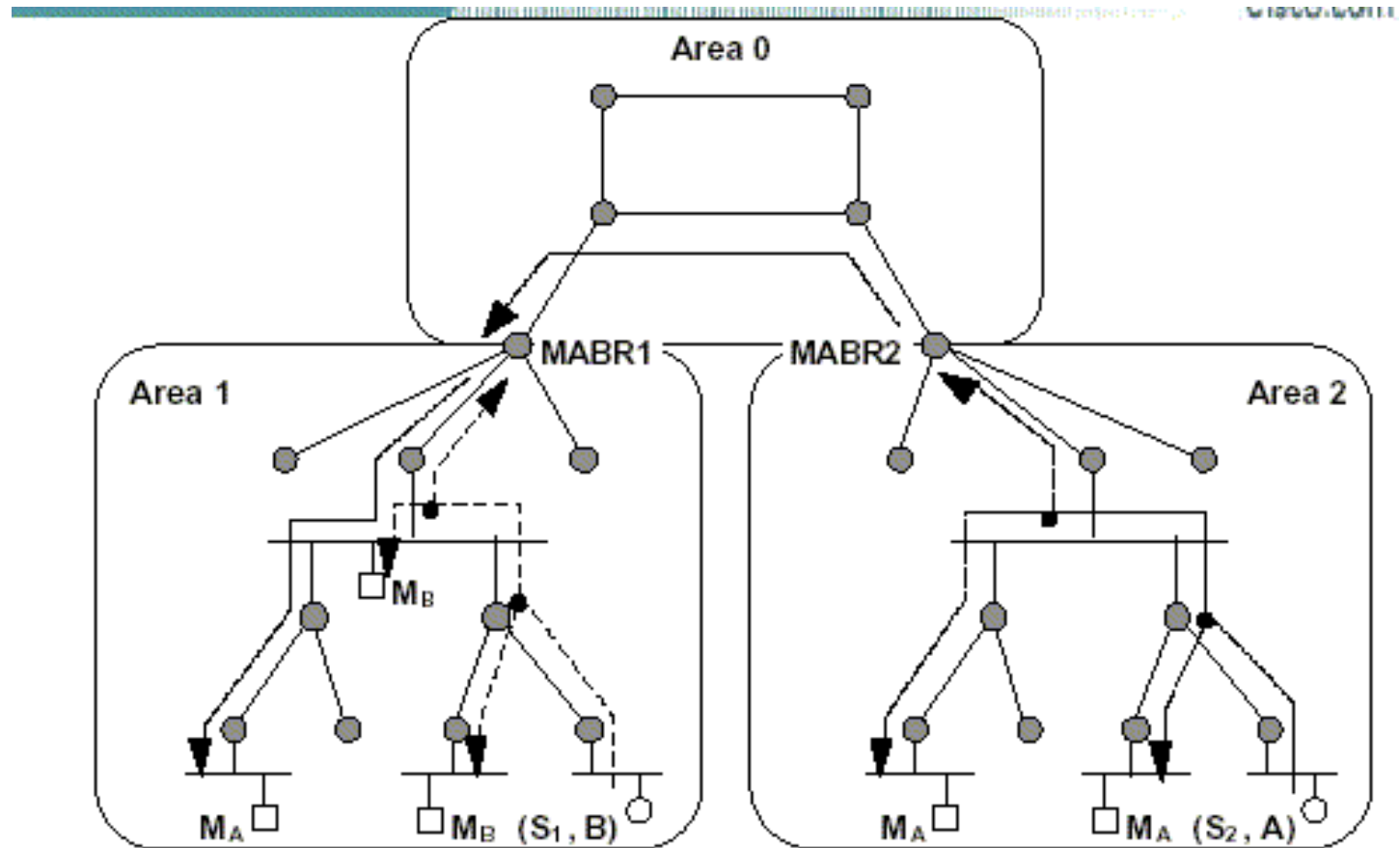
MOSPF – Complementary (7)

- Area's summary group membership LSA
- Routers in backbone area (area 0) calculate the shortest path accordingly



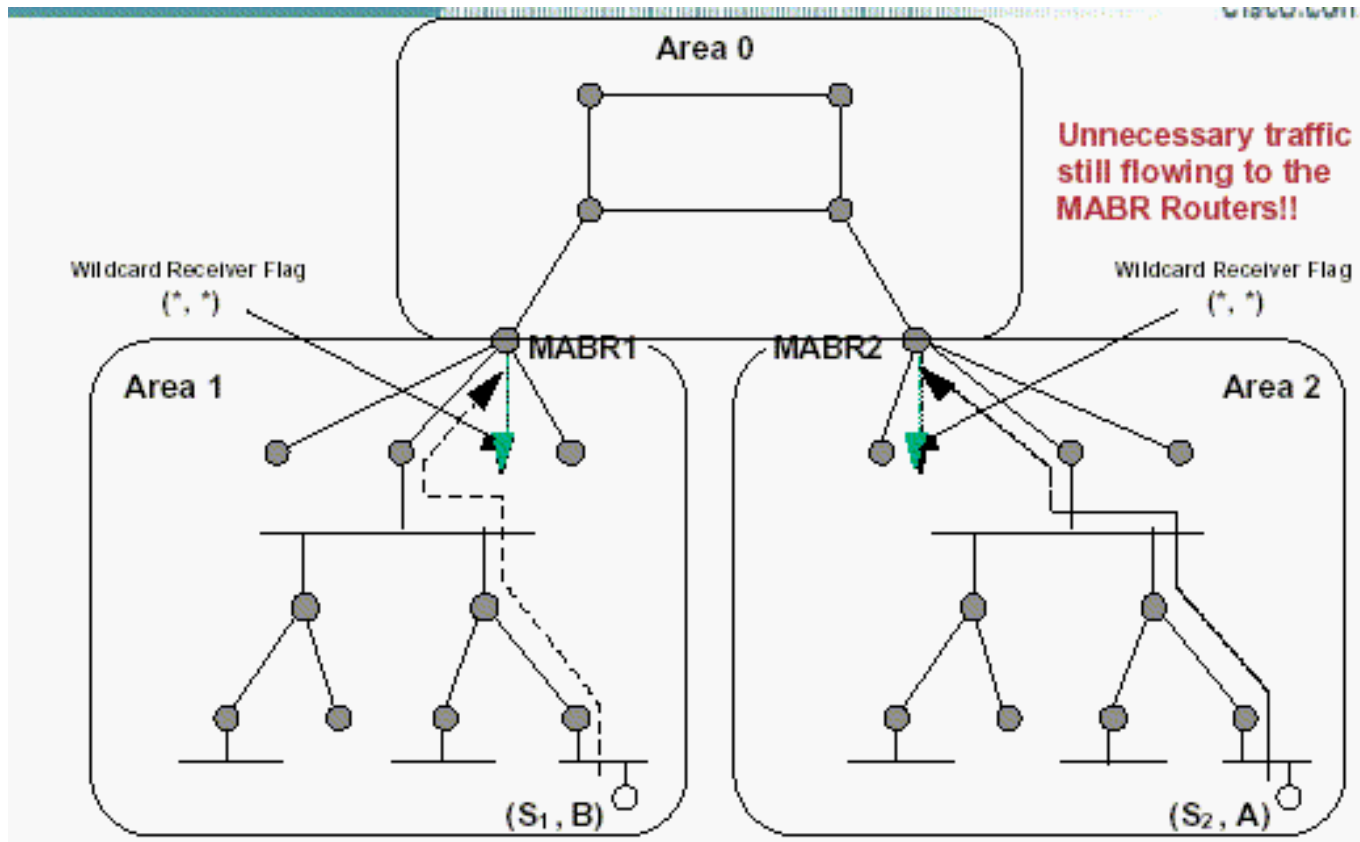
MOSPF – Complementary (8)

🌾 Inter-area multicast data delivery



MOSPF – Complementary (9)

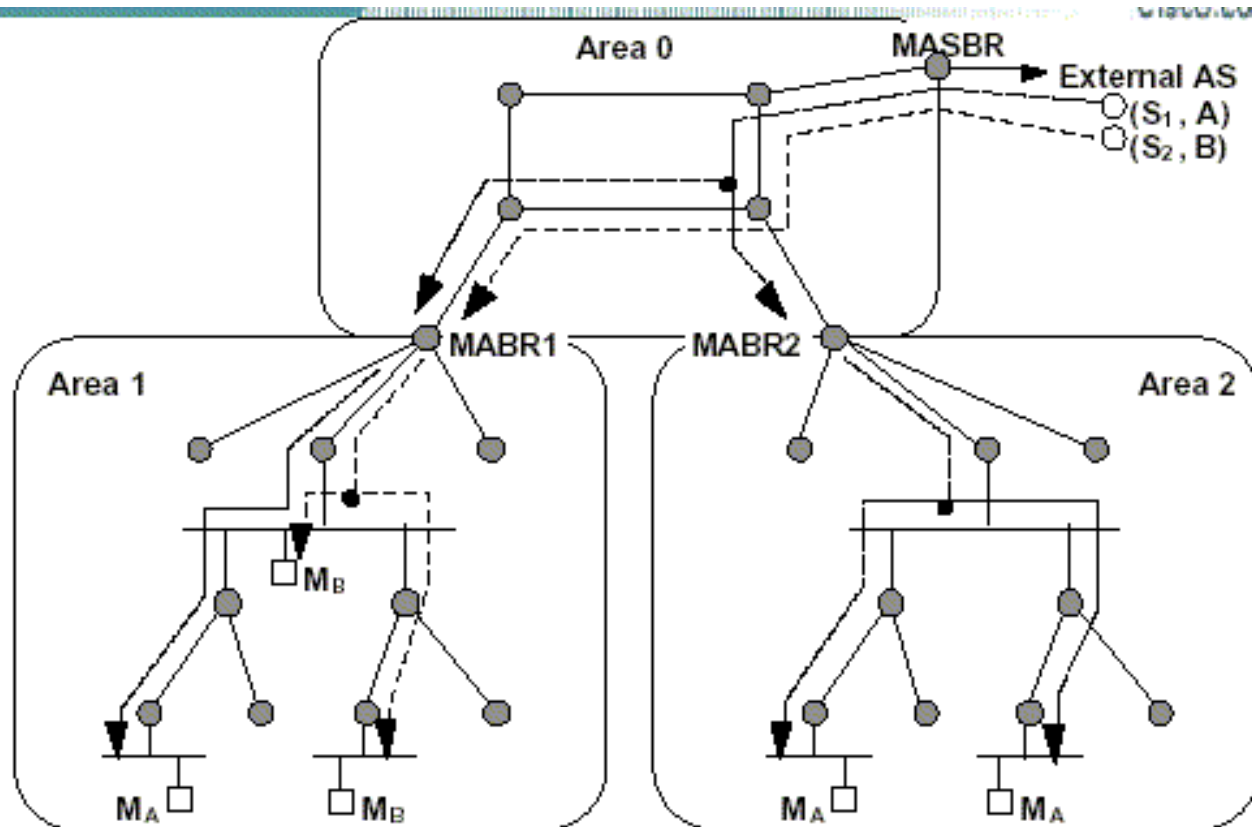
- When there are no members for a multicast group, traffic is still pulled to the MABRs as a result of the Wildcard Receiver mechanism
- This can result in unnecessary bandwidth consumption



MOSPF – Complementary (10)

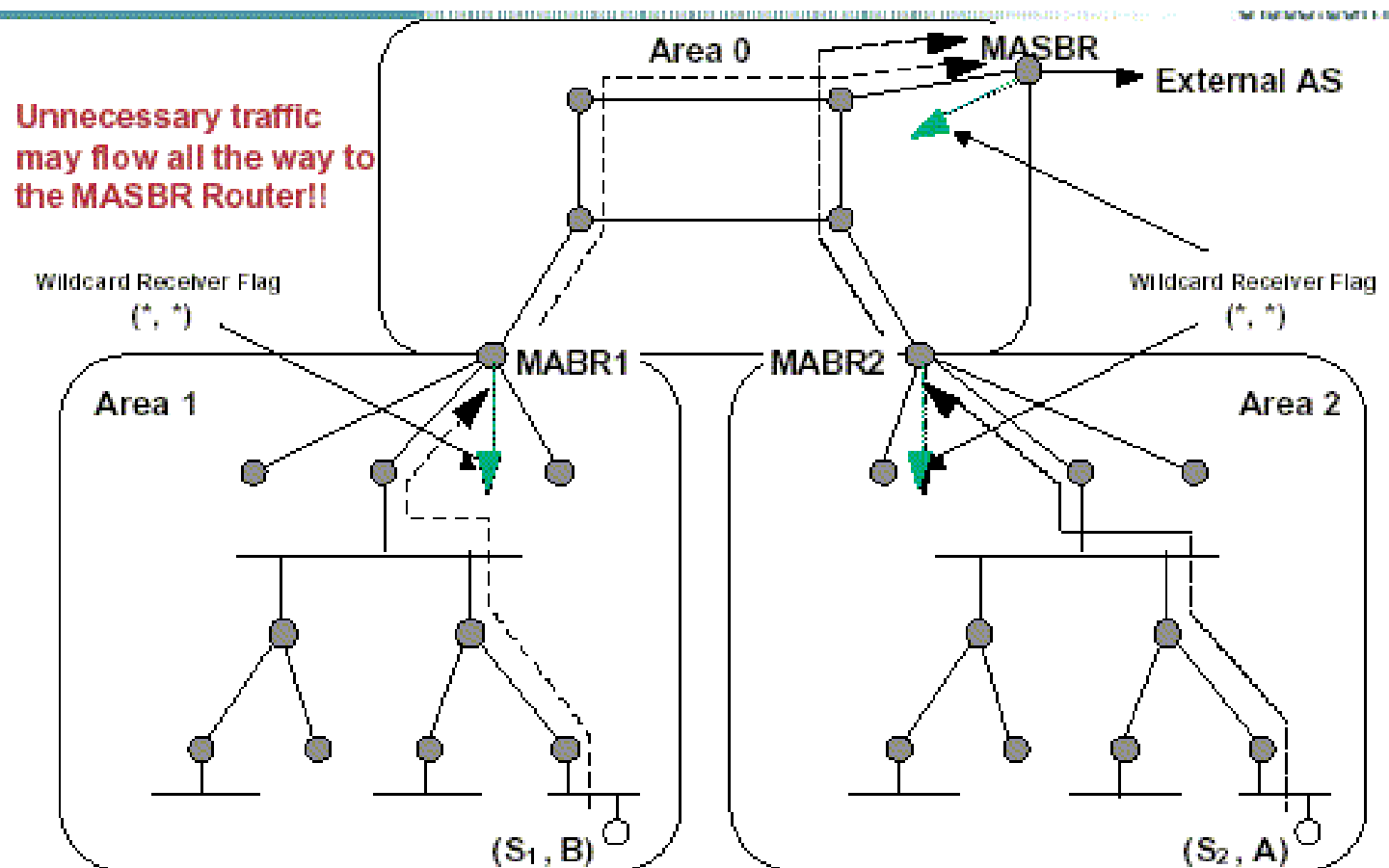
🌾 Inter-AS multicast data delivery

- When traffic arrives from outside the domain via the Multicast AS Border Router (MASBR), this traffic is forwarded across the backbone to the MABRs as necessary based on the Summary Membership LSAs



MOSPF – Complementary (11)

- Still have the problem of unnecessary bandwidth waste



CBT (1/7)

- ✿ Core Based Trees
- ✿ Center-based multicasting
- ✿ Design goals
 - Scalability: $S \times N$ vs. N
 - Tree creation
 - Receiver-based formation
 - Significant benefit to all routers on the SP non-receiver senders and the multicast tree
 - Unicast routing separation

CBT (2/7)

Components

- Core
 - Primary Core
 - Secondary Core
- Designated Router (DR)
 - The same as IGMP querier for the subnet

Function

- Every member of the group sends an explicit Join towards the primary core resulting in either creation of a branch ending in the primary core or ending in an existing branch of the tree

Weakness of CBT

- Core placement and shortest path
- Longer latency compared to SPT
- Traffic concentration
- The core as a point of failure

CBT (3/7)

- ✿ Cores and core placement
 - Core list
 - Priority/ranking
 - Explicit ordered list
 - Primary core, secondary core, ...
 - Placement schemes
 - Statically configured
 - Any router can become a core when a host on one of its attached subnets wished to initiate a group

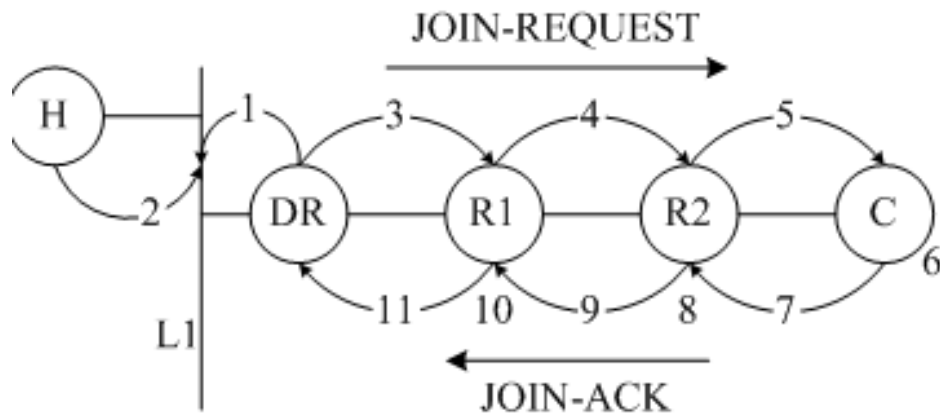
CBT (4/7)

Protocol operations

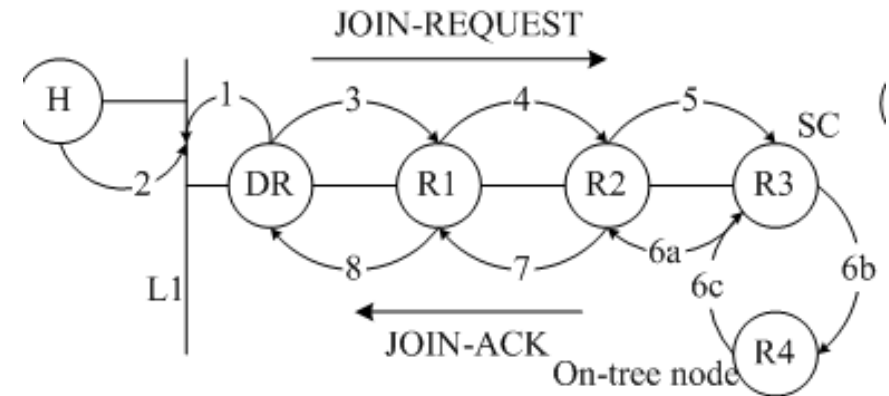
- Tree creation
 - Joining a primary core
 - Unicast Join-Request (with active bit=1) and Join-ACK hop-by-hop + create forwarding entry backwards
 - “join pending” state
 - “CBT-non-core” state
 - In pending state and get a better path,...
 - One parent link and multiple child links
 - Joining a secondary core
 - As joining a primary core but split to two parts: DR joins a secondary core → the secondary core joins a primary core → may hit an on-tree router before reaching the primary core

CBT (5/7)

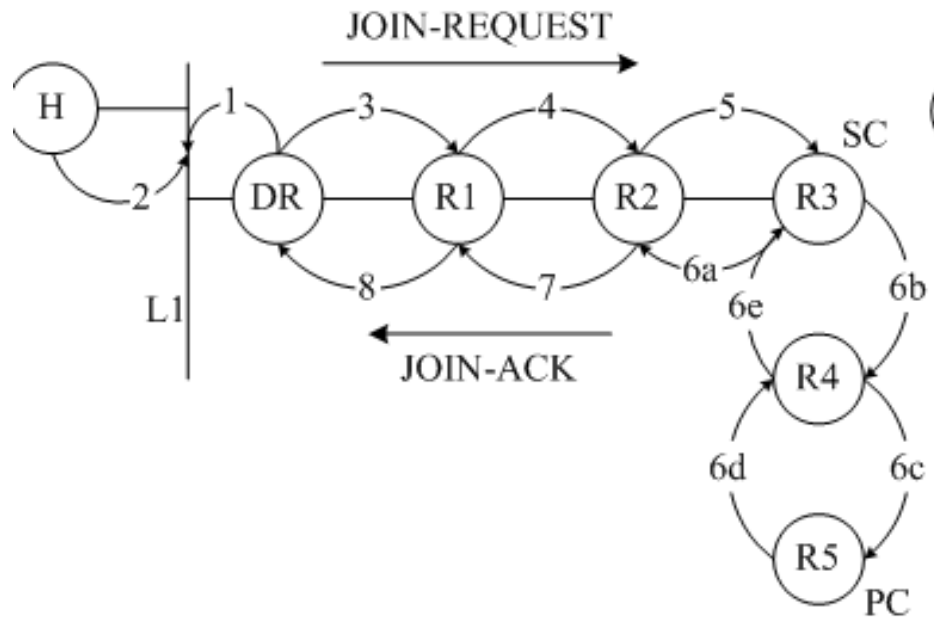
CBT Join



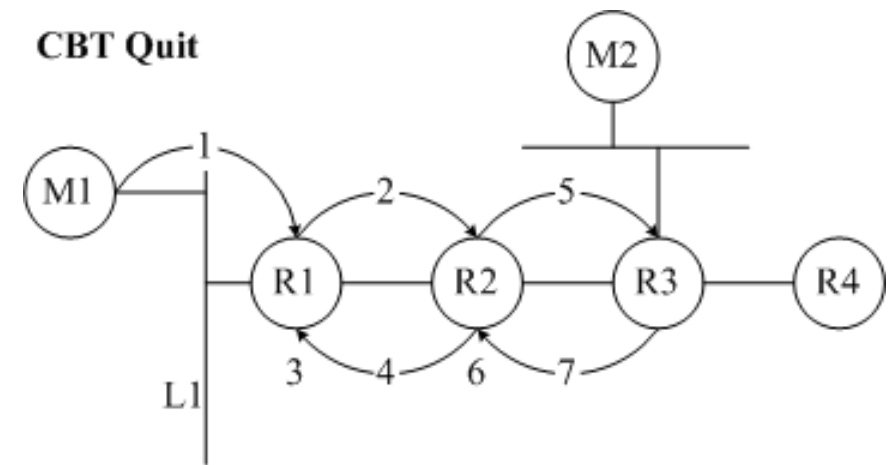
CBT Join



CBT Join



CBT Quit



CBT (6/7)

- Leave
 - Conditions
 - » No members for that group on any directly attached subnets
 - » It has received a QUIT-REQUEST on each of its child interfaces for that group
 - Operations
 - » Unicast Quit-Request (to its parent router)
 - » Quit-Ack (reply by the parent router) hop-by-hop
 - » Delete forwarding entry
- Data forwarding
 - Two phase routing
 - Unicast routing to send multicast packets to a multicast tree, allowing multicast groups and multicast packets to remain invisible to routers on the tree
 - Once on the multicast tree, multicast data packets to all group members based on group id
 - » FIB (Forwarding Information Base)

CBT (7/7)

Core router discovery

– Bootstrap

- For intra-domain
- Some domain's routers are configured to be CBT candidate core routers, and periodically advertises themselves to the domain's Bootstrap Router (BSR), using "Core Advertisement" messages
- BSR is elected dynamically from all domain's routers
- BSR collects these "Core Advertisement" messages and periodically advertises a candidate core set (CC-set) by hop-by-hop unicast forwarding, by using "Bootstrap Messages"

– Manual configuration

PIM

- 🌿 Protocol Independent Multicast
- 🌿 Function: DVMRP + CBT

PIM-DM

- 🌿 Source-rooted tree
- 🌿 When a source starts sending, all downstream hosts want to receive multicast datagrams
- 🌿 Explicit prune message with a timer
- 🌿 Graft message
- 🌿 Similar to DVMRP

PIM-SM (1/7)

Design goals

- Efficient sparse group support
 - Sparse means
 - The ratio of the number of networks/domains with members is significant smaller than the total number of networks/domains in a region
 - Group members are widely distributed
 - Flooding and pruning overhead of non-member is significant high
 - Sparse-mode does not imply that the group has a few members

PIM-SM (2/7)

- High quality data distribution
 - Supporting low-delay data distribution if needed
 - Avoiding imposing a single shared tree
 - Multiple sources send data simultaneously
 - Path lengths between sources and destinations in SPT's are significantly shorter than in shared tree
- Unicast routing independence
- Robustness
 - Soft state mechanism
 - Avoiding a single point of failure
- Interoperability

PIM-SM (3/7)

Components

- Rendezvous Point (RP)
- Designated Router (DR)
 - Senders: PIM Register
 - Piggybacked
 - Receivers: PIM Join-Request
 - PIM-Register-Stop
- Bootstrap Router (BSR)

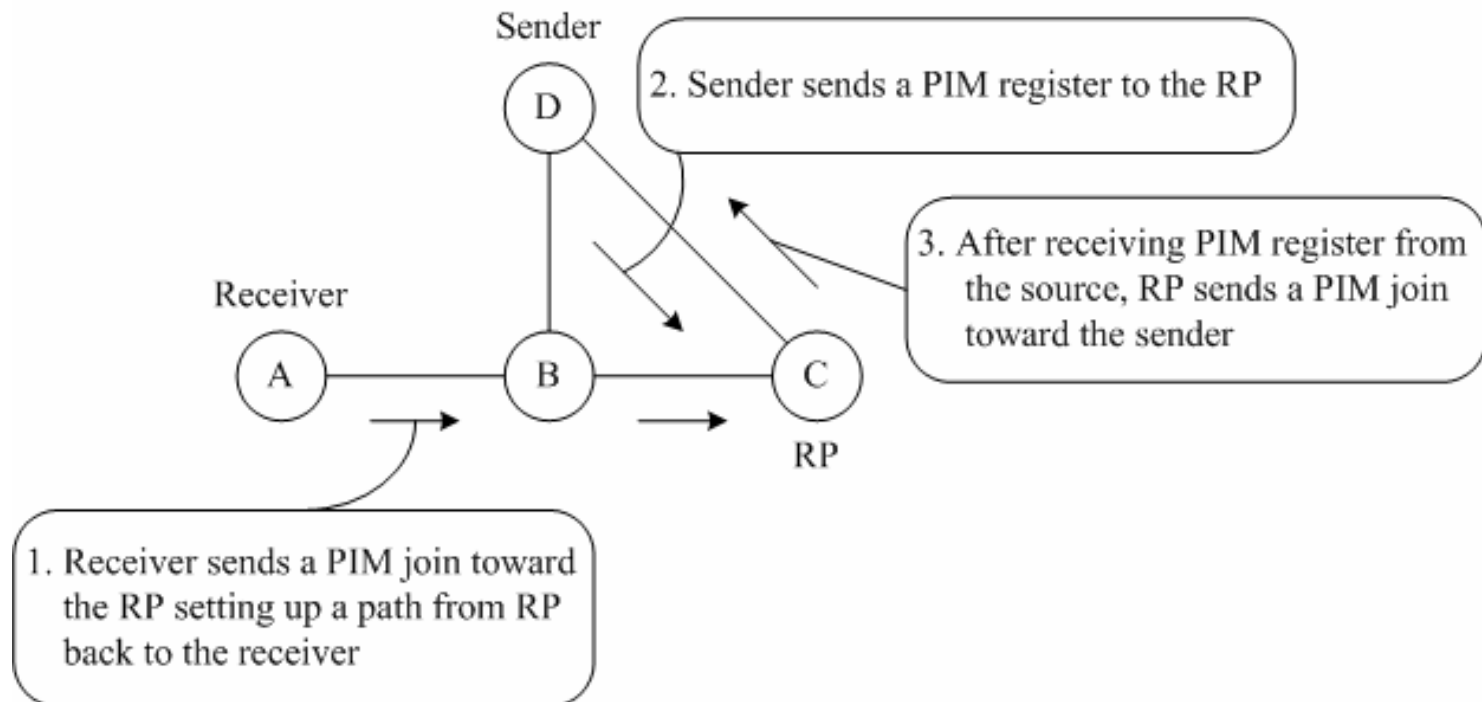
Creating the PIM framework

- An RP sends C-RP-Advs to BSR → BSR announces the RP-set using bootstrap message (BSM) to all routers → every router can uniquely identify the RP for the group

PIM-SM (4/7)

🌿 Creating a RP-rooted PIM tree

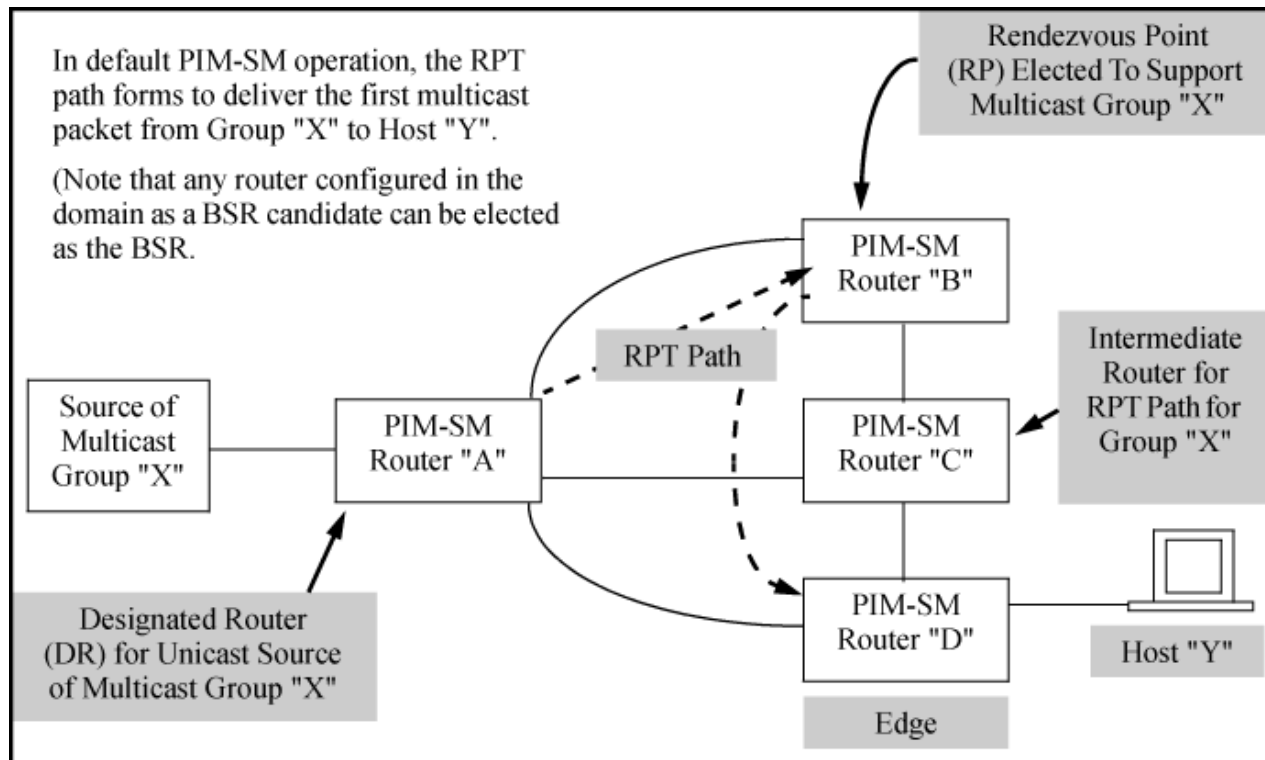
- Receiver join
- Source join (join request piggybacked in multicast data packet)
 - If none potential receivers, the RP router simply drop received data packets or inform the source stopping transmissions



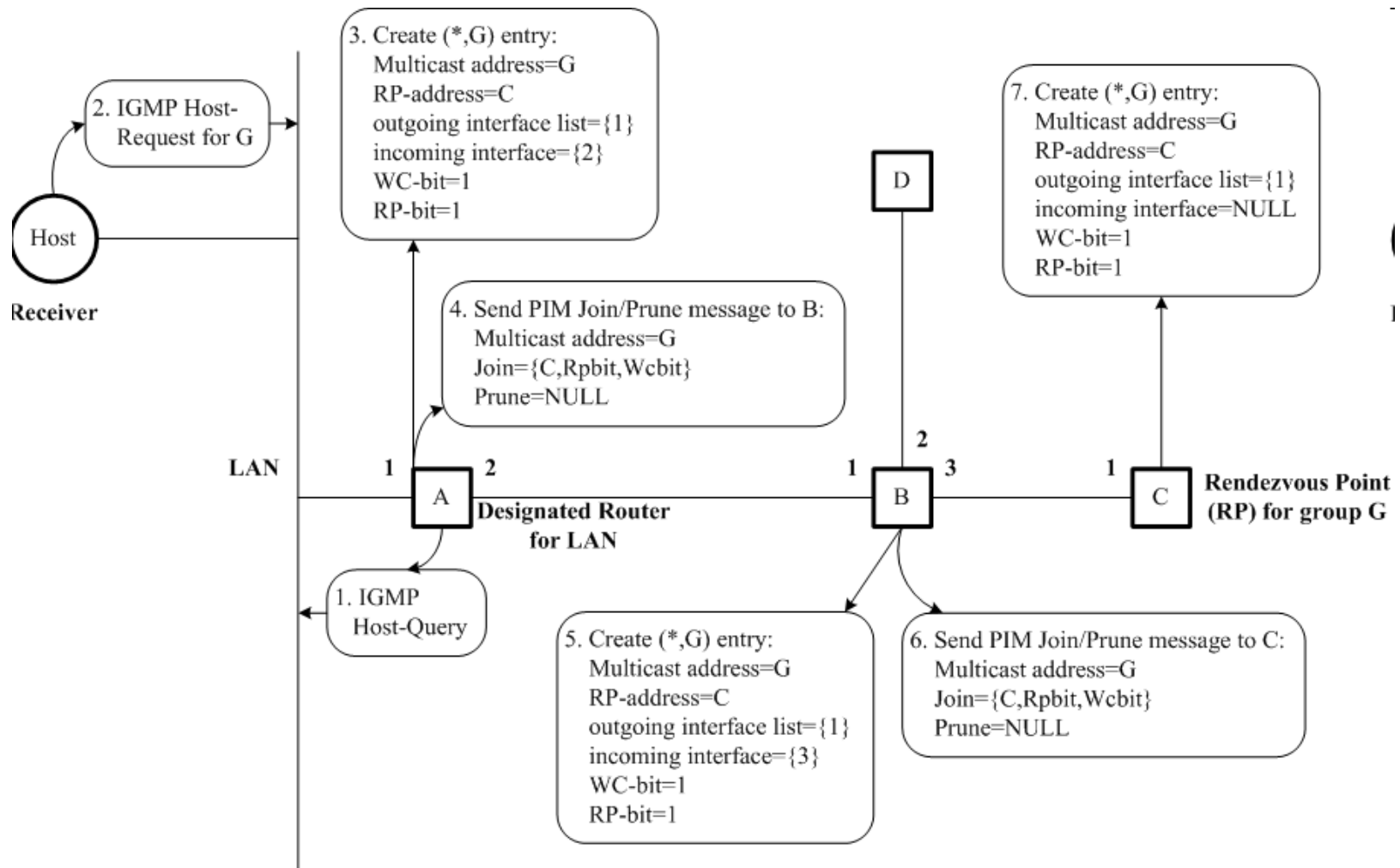
PIM-SM (5/7)

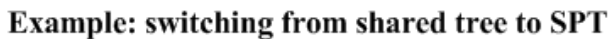
Protocol operations

- Local host joining a group
- Establishing the RP-rooted shared tree (a.k.a. RP tree, RPT)
- Switching from shared tree to SPT (shortest path tree)



PIM-SM (6/7)

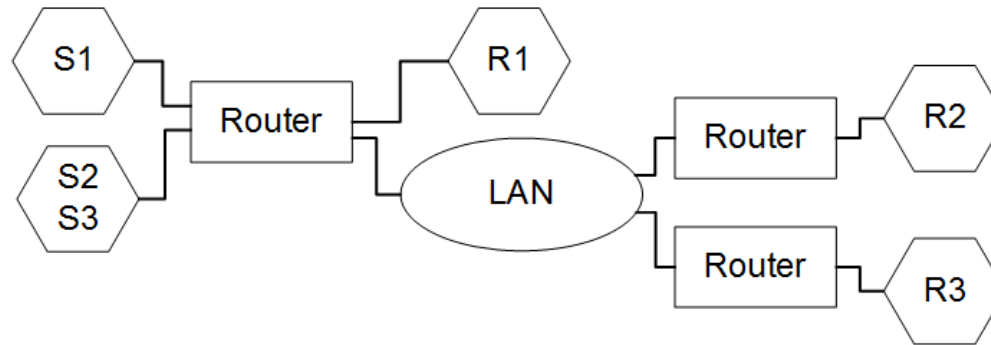




Comparison of DVMRP and CBT Router States

No. groups	10	10	10	100	100	100	1000	1000	1000
Group size	20	20	20	40	40	40	60	60	60
Sources per group	10%	50%	100%	10%	50%	100%	10%	50%	100%
# DVMRP router entries	20	100	200	400	2000	4000	6000	30000	60000
# CBT router entries	10	10	10	100	100	100	1000	1000	1000





Problem Set 2



	FF	SE	WF
R ₁	FF(S ₁ {6B}, S ₂ {6B})	SE((S ₁ , S ₂) {5B})	WF(* {3B})
R ₂	FF(S ₁ {4B}, S ₃ {3B})	SE((S ₁ , S ₃) {4B})	WF(* {5B})
R ₃	FF(S ₁ {7B})	SE((S ₂) {3B})	WF(* {4B})
S ₁	7B	5B	5B
S ₂	6B	5B	5B
S ₃	3B	5B	5B

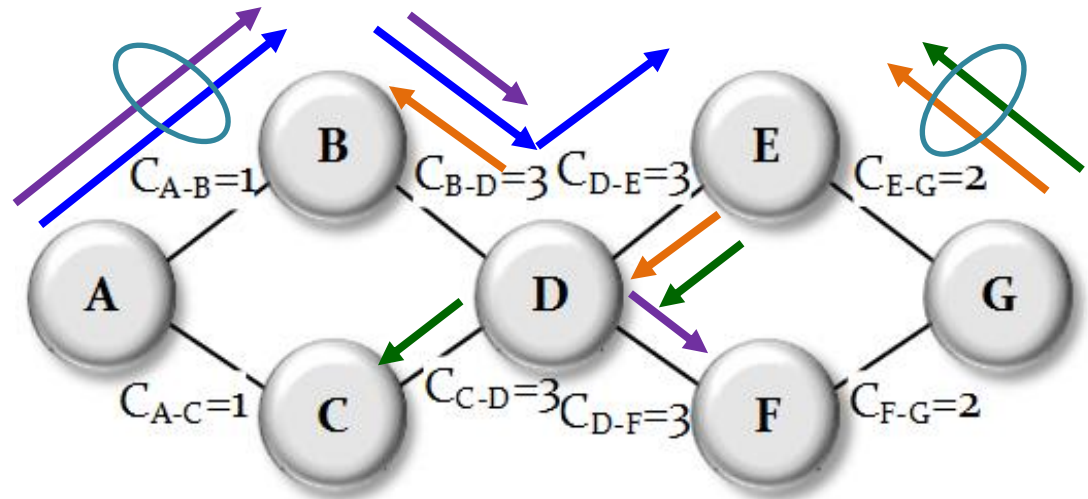
Problem Set 5 (1)

✿ Considering the following network topology

- Flow 1 is from A to E (path is $A \rightarrow B \rightarrow D \rightarrow E$) 
- Flow 2 is from A to F (path is $A \rightarrow B \rightarrow D \rightarrow F$) 
- Flow 3 is from G to B (path is $G \rightarrow E \rightarrow D \rightarrow B$) 
- Flow 4 is from G to C (path is $G \rightarrow E \rightarrow D \rightarrow C$) 

✿ Q1: What are the rate allocations for these flows, respectively, when utilizing max-min resource allocation?

- Flow 1: 0.5
- Flow 2: 0.5
- Flow 3: 1
- Flow 4: 1



Problem Set 5 (2)

Q2: We further assume that the rate demands of flows 1-4 are 1, 1, 2, and 2. How will you manage these flows through traffic engineering to achieve the least total delay? [hint: assign different routes to flows]

– The maximum of link utilization is minimized, the total delay the packets experience is also minimized

– Flow 1 is from A to E →

– Flow 2 is from A to F →

– Flow 3 is from G to B →

– Flow 4 is from G to C →

