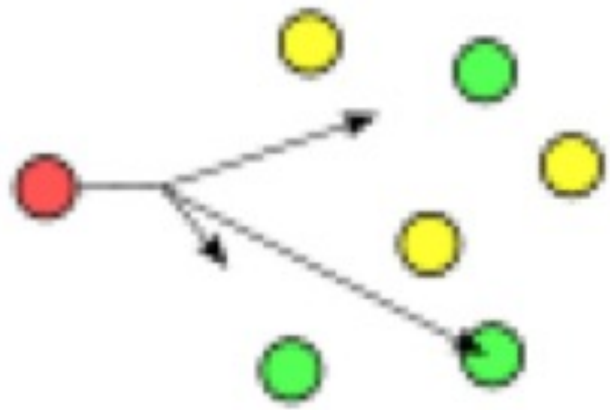


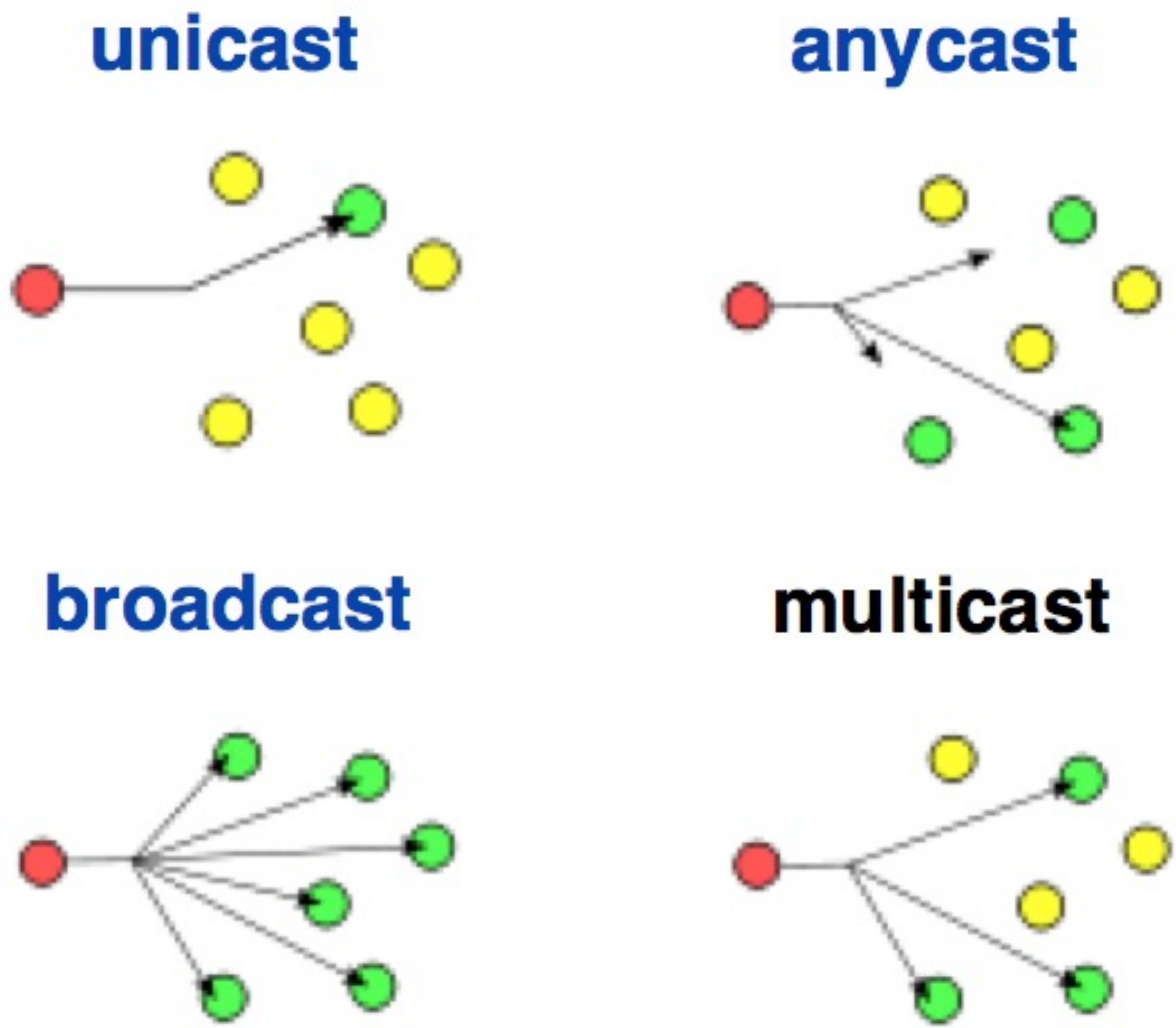
Multicast and Anycast

Note: The slides are adapted from the materials from Prof. Richard Han at CU Boulder and Profs. Jennifer Rexford and Mike Freedman at Princeton University, and the networking book (Computer Networking: A Top Down Approach) from Kurose and Ross.

Outline today

- IP Anycast
  - N destinations, 1 should receive the message
  - Providing a service from multiple network locations
  - Using routing protocols for automated failover

- Multicast protocols
  - N destinations, N should receive the message
  - Examples
    - IP Multicast and IGMP
    - PIM-DM
    - PIM-SM



http://en.wikipedia.org/wiki/Multicast

Limitations of DNS-based failover

- Failover/load balancing via multiple A records

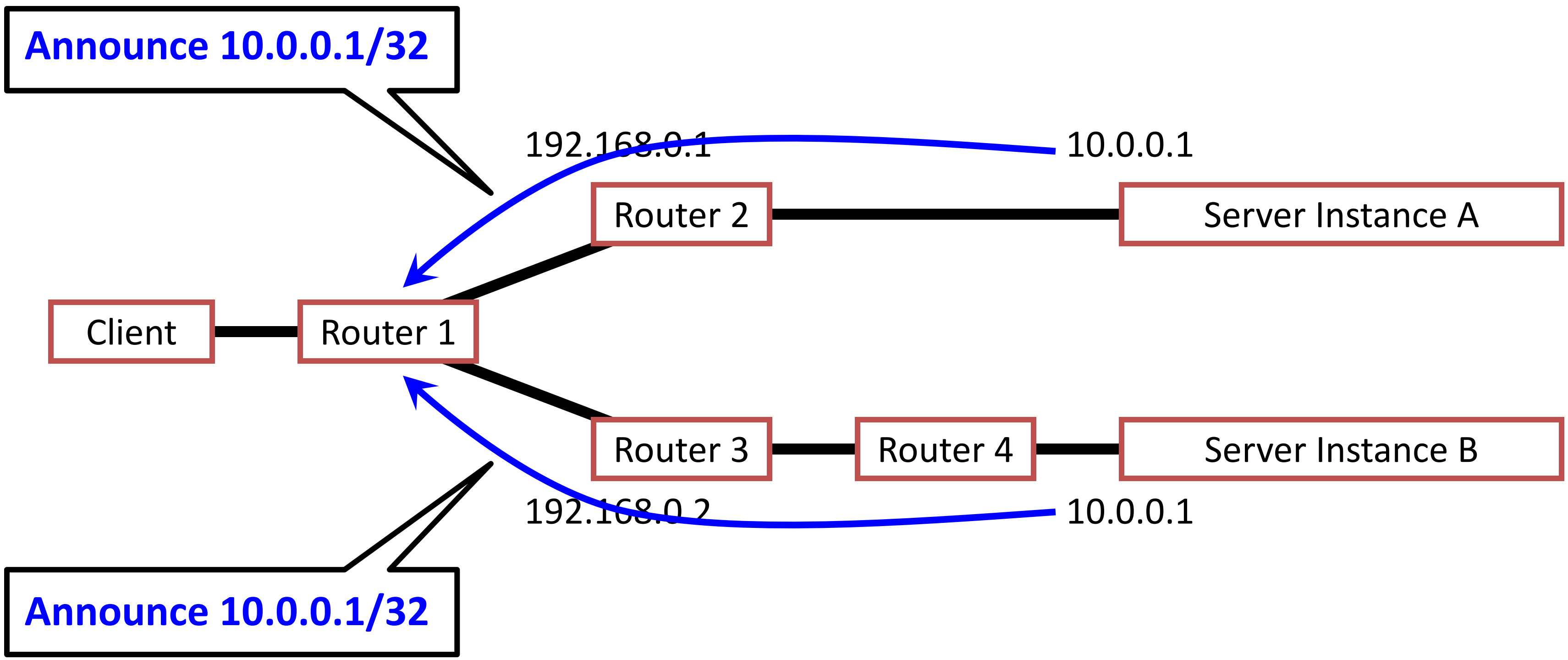
```
;; ANSWER SECTION:
www.cnn.com.     300    IN A   157.166.255.19
www.cnn.com.     300    IN A   157.166.224.25
www.cnn.com.     300    IN A   157.166.226.26
www.cnn.com.     300    IN A   157.166.255.18
```

- If server fails, service unavailable for TTL
  – Very low TTL: Extra load on DNS
  – Anyway, browsers cache DNS mappings ☹
- What if root NS fails? All DNS queries take > 3s?

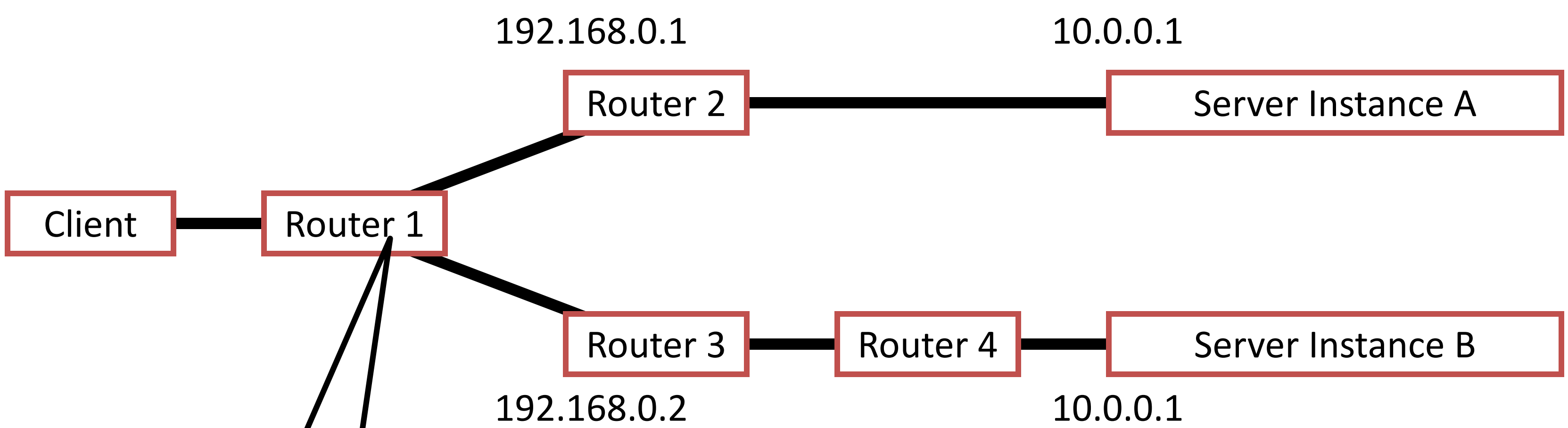
Motivation for IP anycast

- Failure problem: client has resolved IP address
  - What if IP address can represent many servers?
- Load-balancing/failover via IP addr, rather than DNS
- IP anycast is simple reuse of existing protocols
  - Multiple instances of a service share same IP address
  - Each instance announces IP address / prefix in BGP / IGP
  - Routing infrastructure directs packets to nearest instance of the service
    - Can use same selection criteria as installing routes in the FIB
  - No special capabilities in servers, clients, or network

IP anycast in action



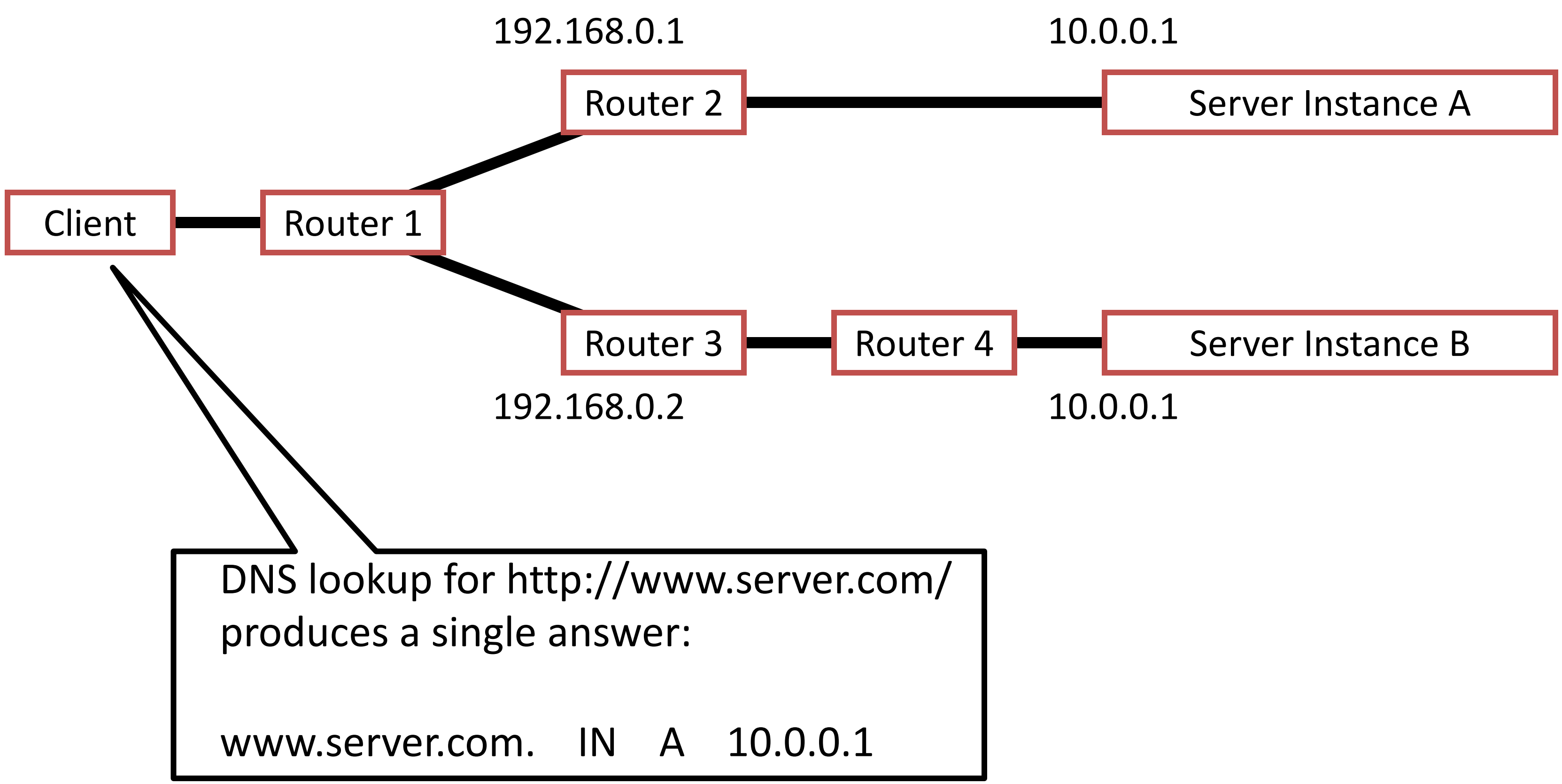
IP anycast in action



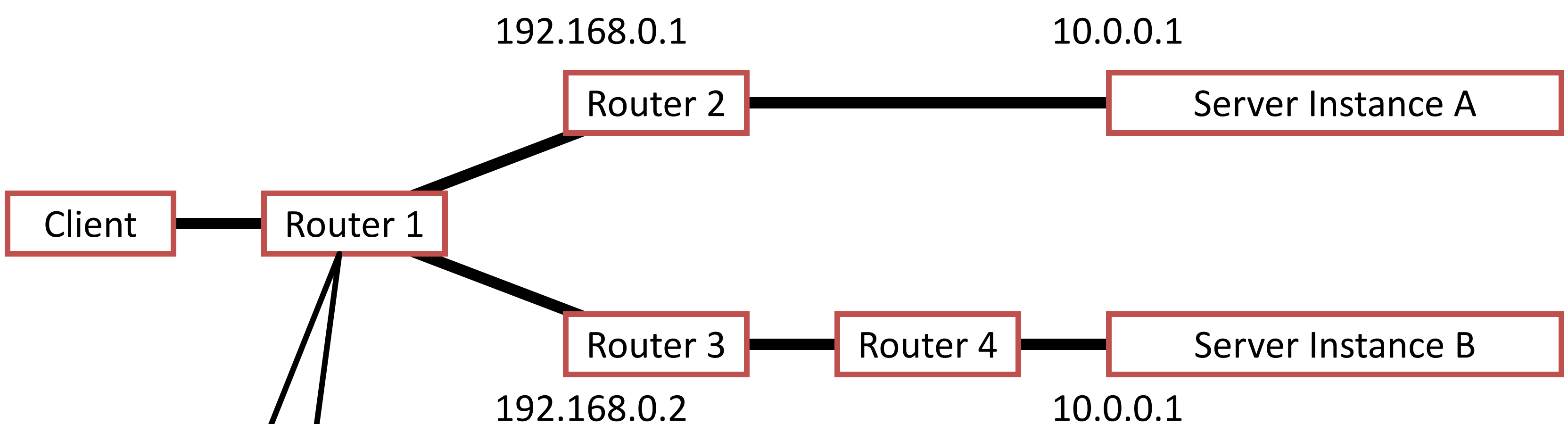
Routing Table from Router 1:

Destination	Mask	Next-Hop	Distance
192.168.0.0	/29	127.0.0.1	0
10.0.0.1	/32	192.168.0.1	1
10.0.0.1	/32	192.168.0.2	2

IP anycast in action



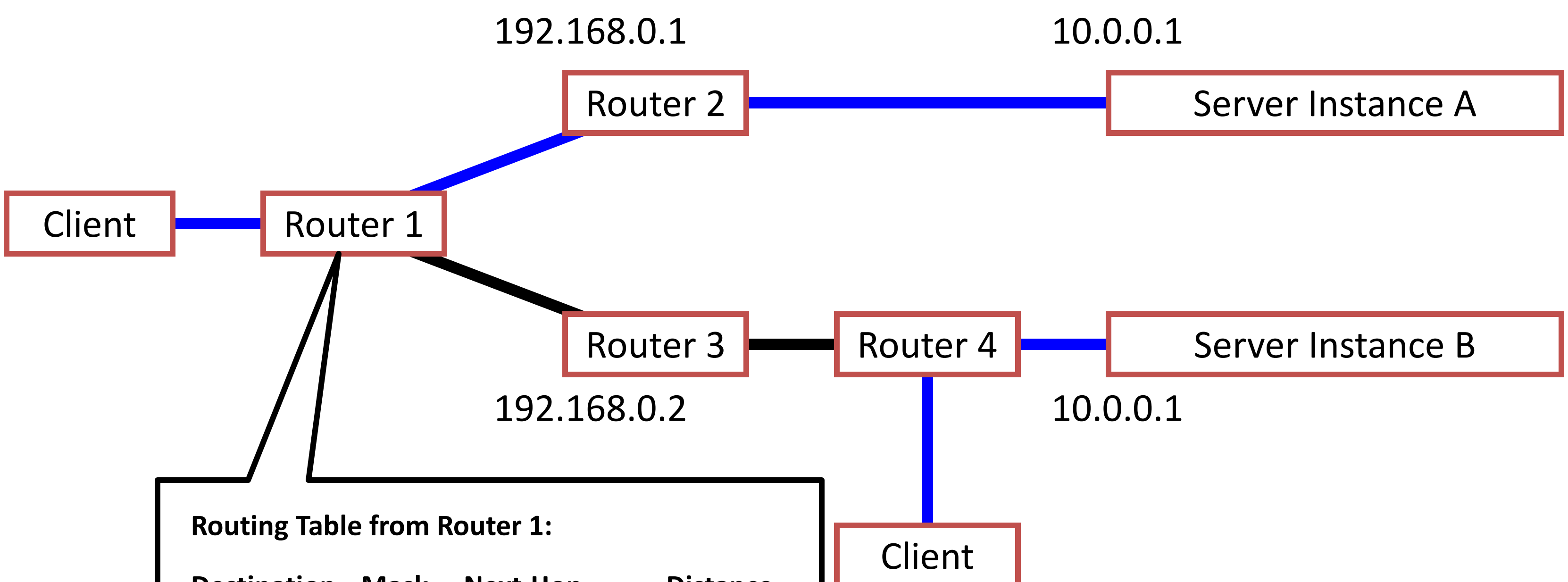
IP anycast in action



Routing Table from Router 1:

Destination	Mask	Next-Hop	Distance
192.168.0.0	/29	127.0.0.1	0
10.0.0.1	/32	192.168.0.1	1
10.0.0.1	/32	192.168.0.2	2

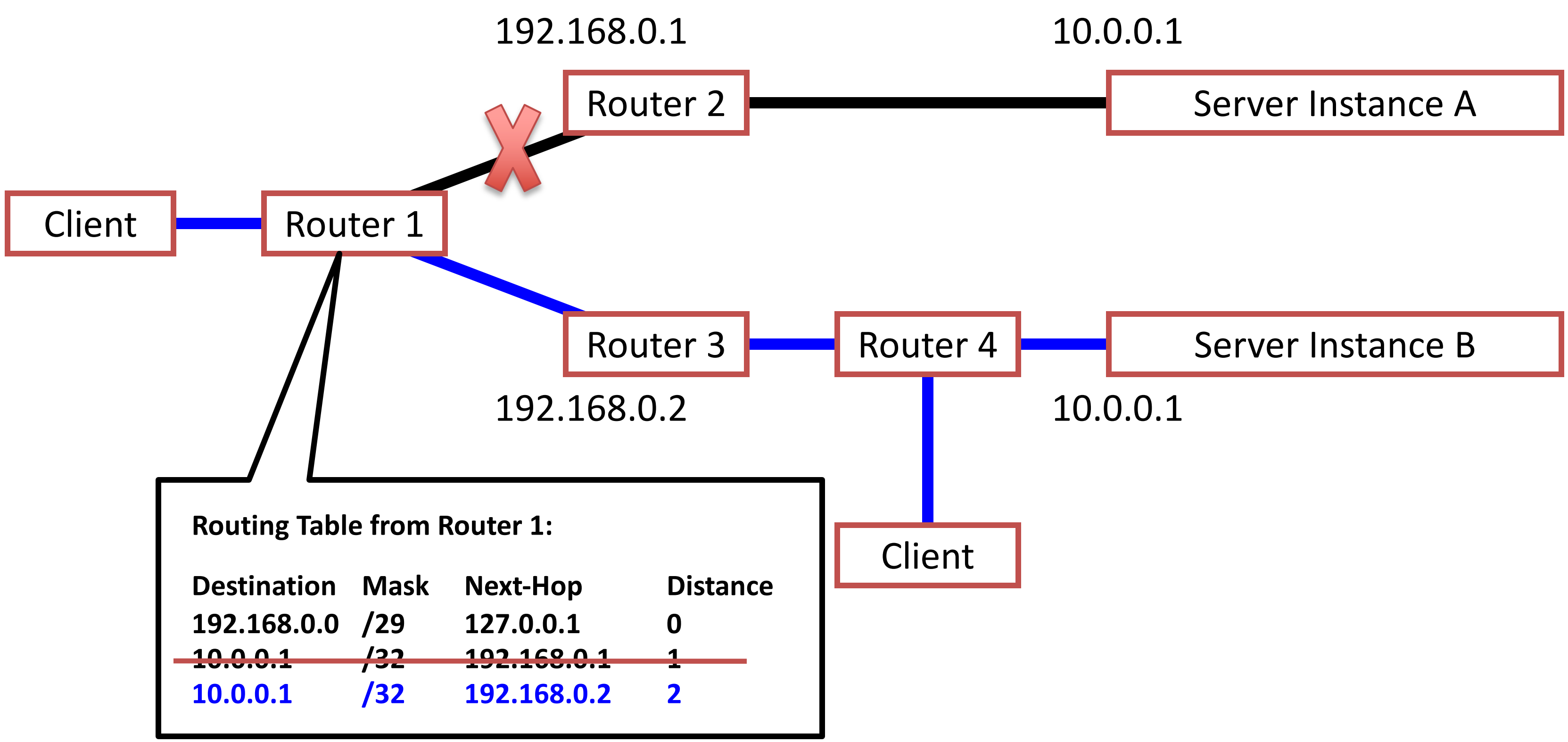
IP anycast in action



Routing Table from Router 1:

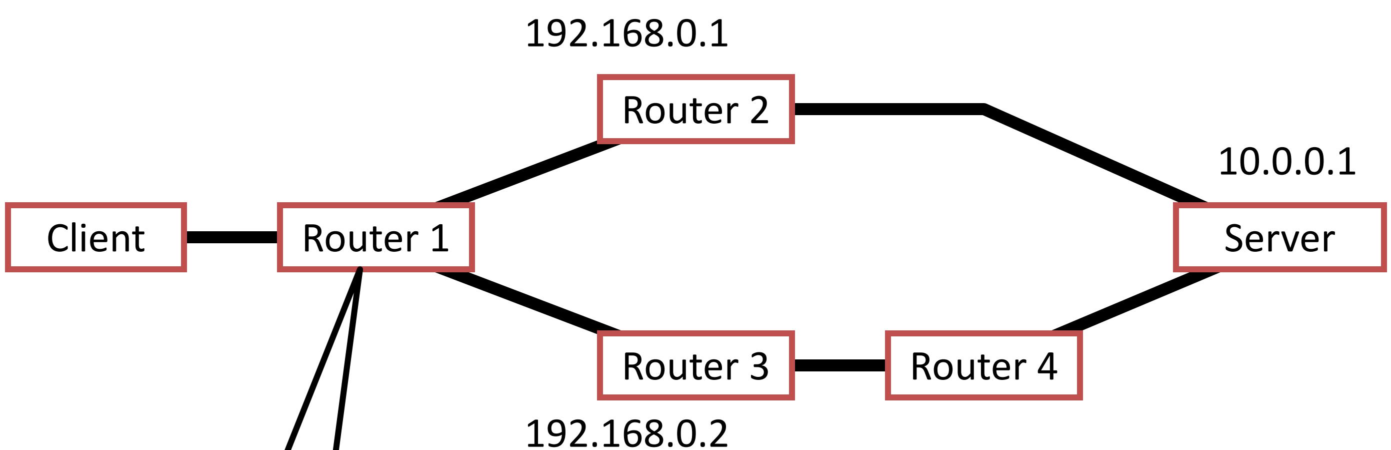
Destination	Mask	Next-Hop	Distance
192.168.0.0	/29	127.0.0.1	0
10.0.0.1	/32	192.168.0.1	1
10.0.0.1	/32	192.168.0.2	2

IP anycast in action



IP anycast in action

From client/router perspective, topology could as well be:



Routing Table from Router 1:

Destination	Mask	Next-Hop	Distance
192.168.0.0	/29	127.0.0.1	0
10.0.0.1	/32	192.168.0.1	1
10.0.0.1	/32	192.168.0.2	2

Downsides of IP anycast

- Many Tier-1 ISPs ingress filter prefixes > /24
  - Publish a /24 to get a “single” anycasted address: Poor utilization
- Scales poorly with the # anycast groups
  - Each group needs entry in global routing table
- Not trivial to deploy
  - Obtain an IP prefix and AS number; speak BGP

Downsides of IP anycast

- Subject to the limitations of IP routing
  - No notion of load or other application-layer metrics
  - Convergence time can be slow (as BGP or IGP converge)
- Failover doesn't really work with TCP
  - TCP is stateful: if switch destination replicas, other server instances will just respond with RSTs
  - May react to network changes, even if server online
- Root nameservers (UDP) are anycasted, little else

Multicast protocols

What Is Multicast?

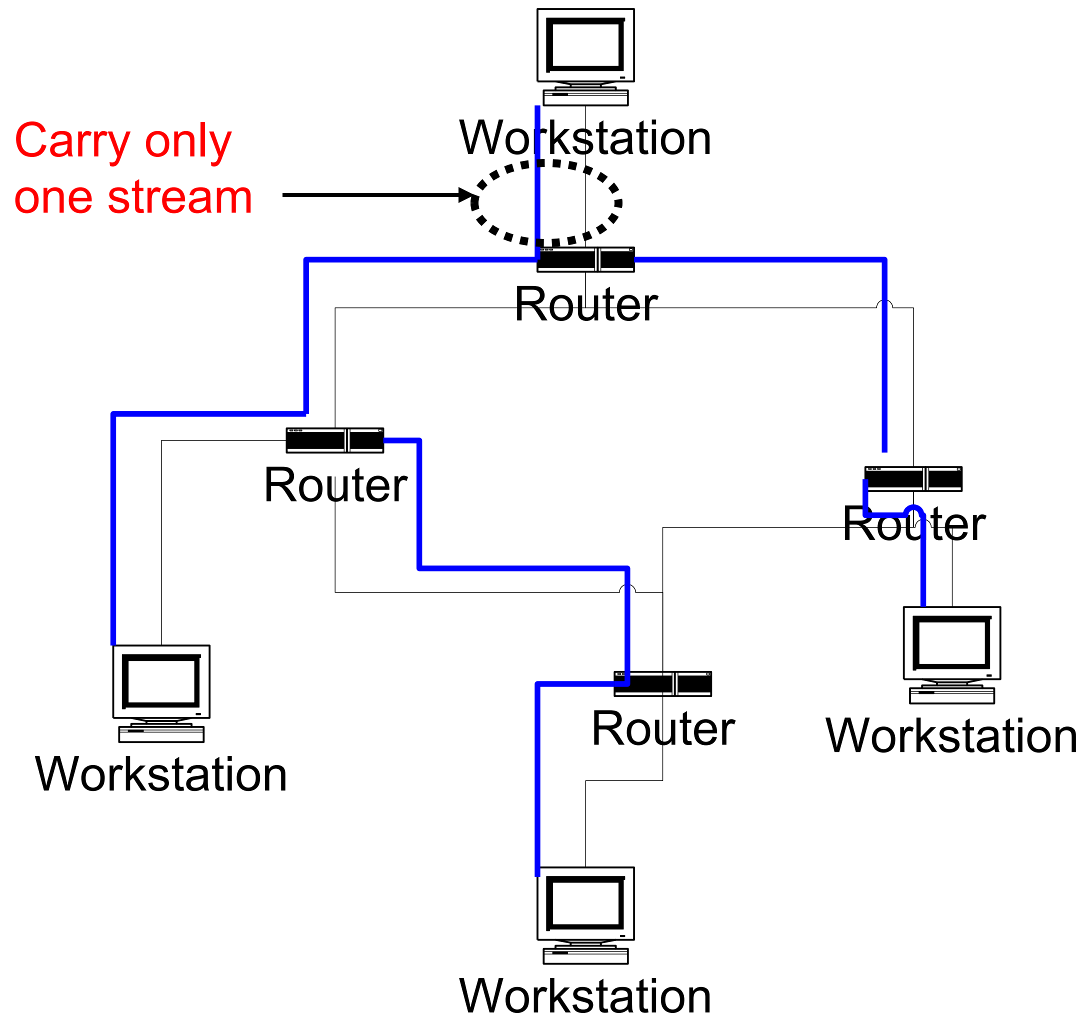
- Transmit the same content to a set of hosts (multicast group)
- Differs from broadcast where it involves all hosts either at a local or at a remote network

Applications

- Audio-video distribution (1 to many)
  - Video-on-demand
- File distribution
  - Stock quotes, new software
  - Content caching
  - News
- Distributed simulation (war gaming, multi-player games)

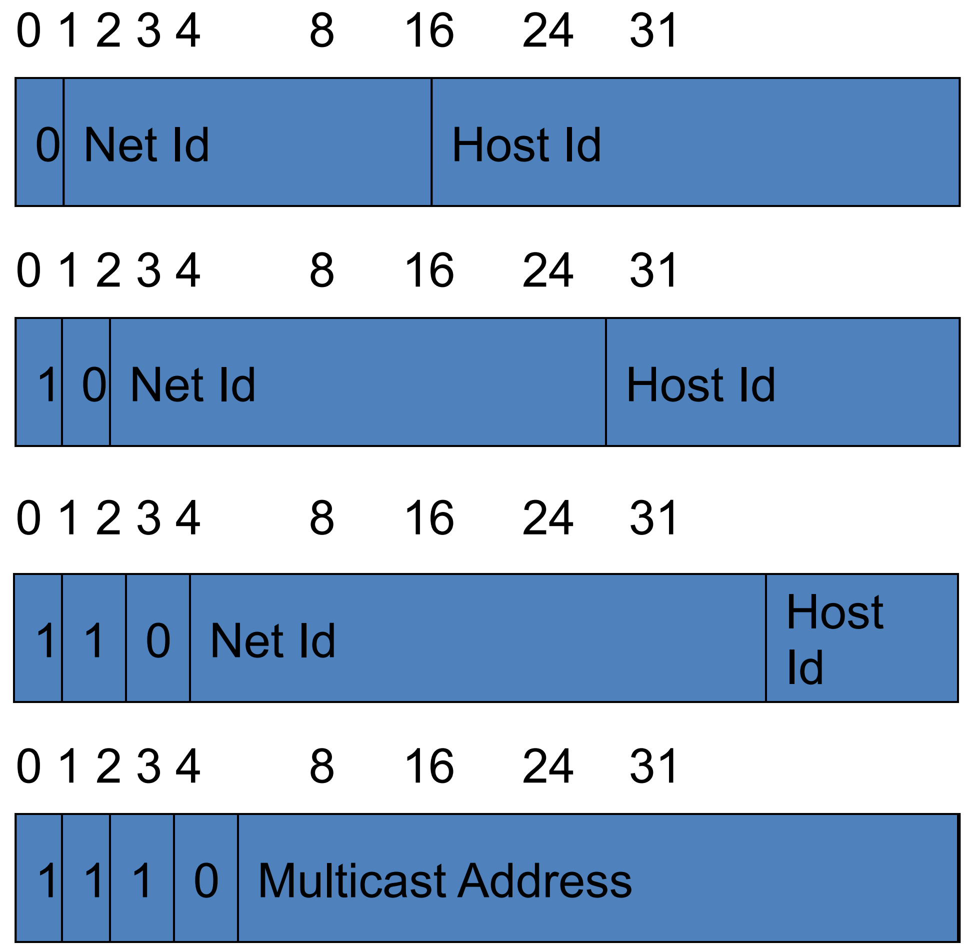
Multicast Advantage

- Bandwidth saving
  - Common links to a group of receivers will carry the same packet only once.



Multicast Addressing

- The 4 major classes of IP addresses
  - Class A address
    - 128 networks
    - > 65,536 hosts
  - Class B address
    - 16,384 networks
    - 256-65,536 hosts
  - Class C address
    - 2^21 networks
    - Less than 256 hosts
  - Class D address
    - 28 bit multicast addr.



224.0.0.0 through 239.255.255.255

Multicast Addresses

- Can only be used as a destination address
- Corresponds to a multicast group
- Special group address
  - 224.0.0.0 – reserved
  - 224.0.0.1 – all hosts group, all multicast capable hosts & routers (with TTL = 1)
  - 224.0.0.2 – all routers group, multicast capable routers only

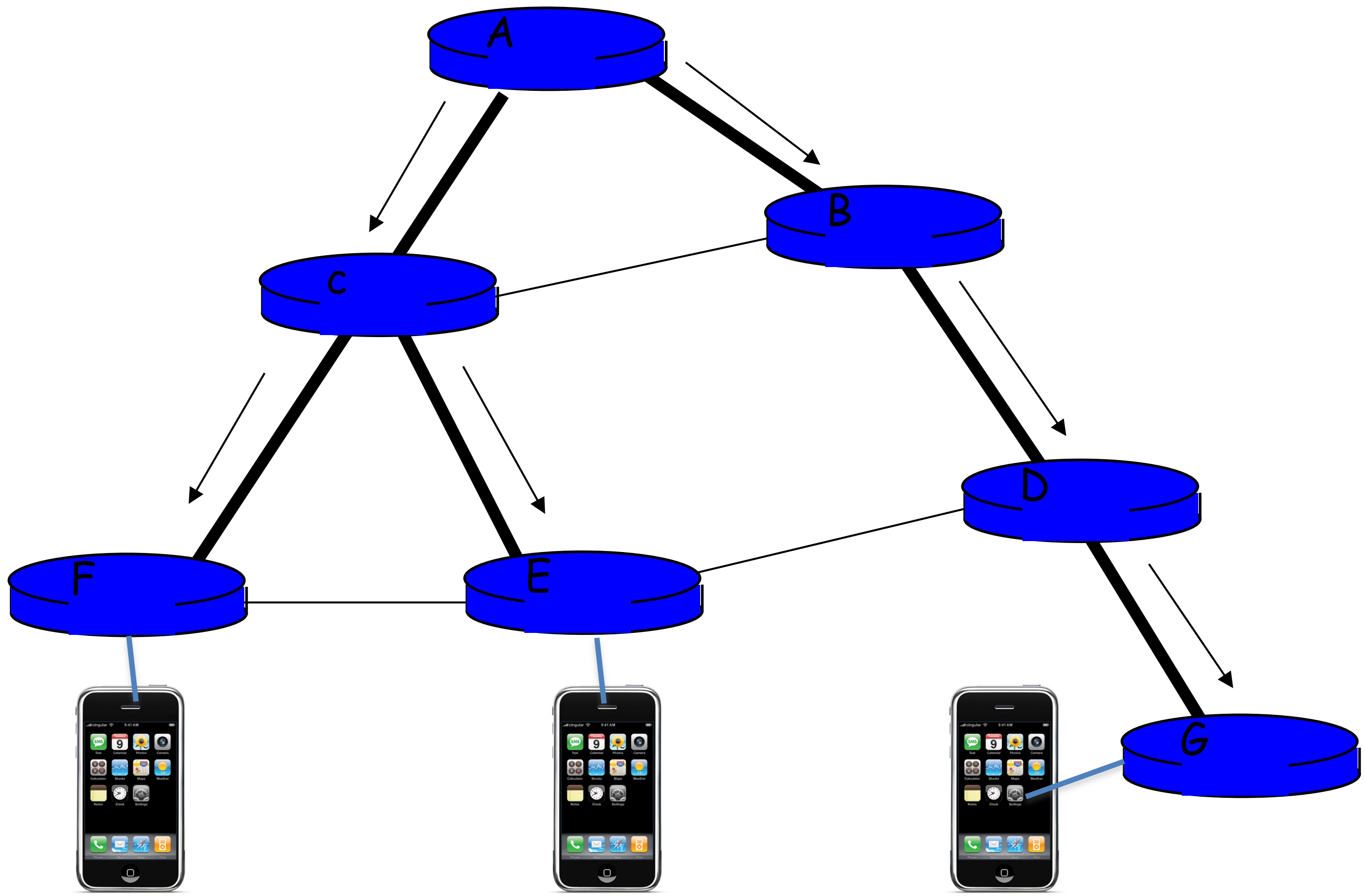
Multicast Groups

- Groups may be of any size
- Group members may be located anywhere in the Internet
- Hosts can join and leave groups at will
- There is no “list” of group members
- Senders need not be members of the group
- Receiver-driven membership

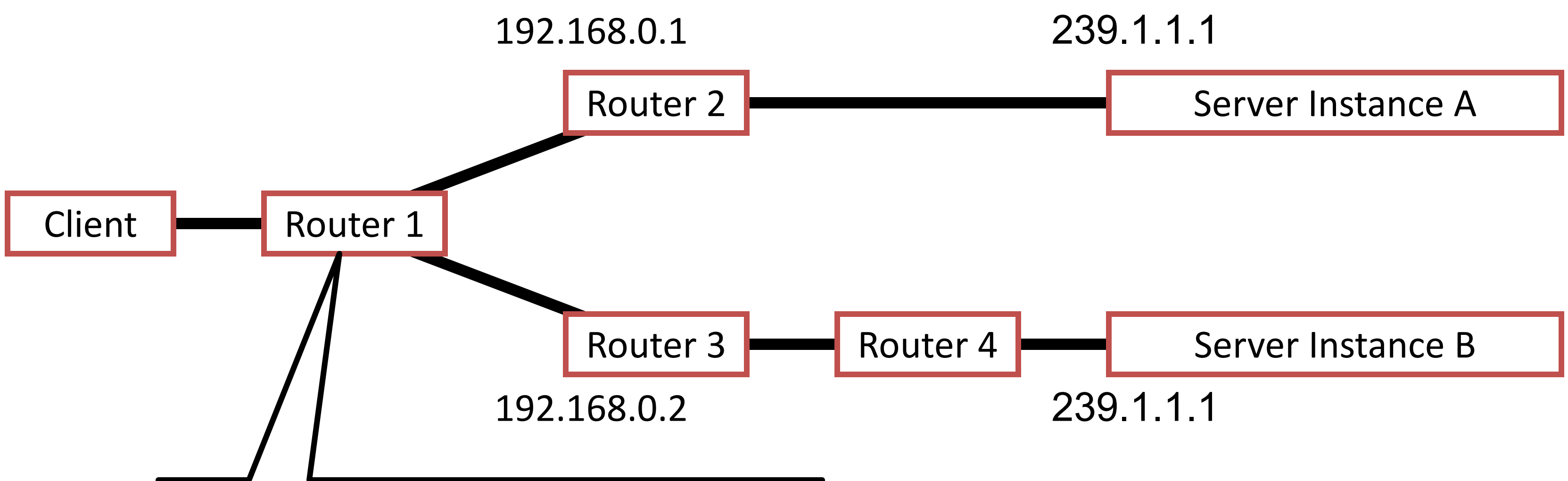
Multicasting messages

- Simple application multicast: Iterated unicast
  - Client simply unicasts message to every recipient
  - Pros: simple to implement, no network modifications
  - Cons: O(n) work on sender, network
- Advanced overlay multicast ("peer-to-peer")
  - Build receiver-driven tree
  - Pros: Scalable, no network modifications
  - Cons: O(log n) work on sender, network; complex to implement
- IP multicast
  - Embed receiver-driven tree in network layer
  - Pros: O(1) work on client, O(# receivers) on network
  - Cons: requires network modifications; scalability concerns?

Multicast Tree



IP multicast in action

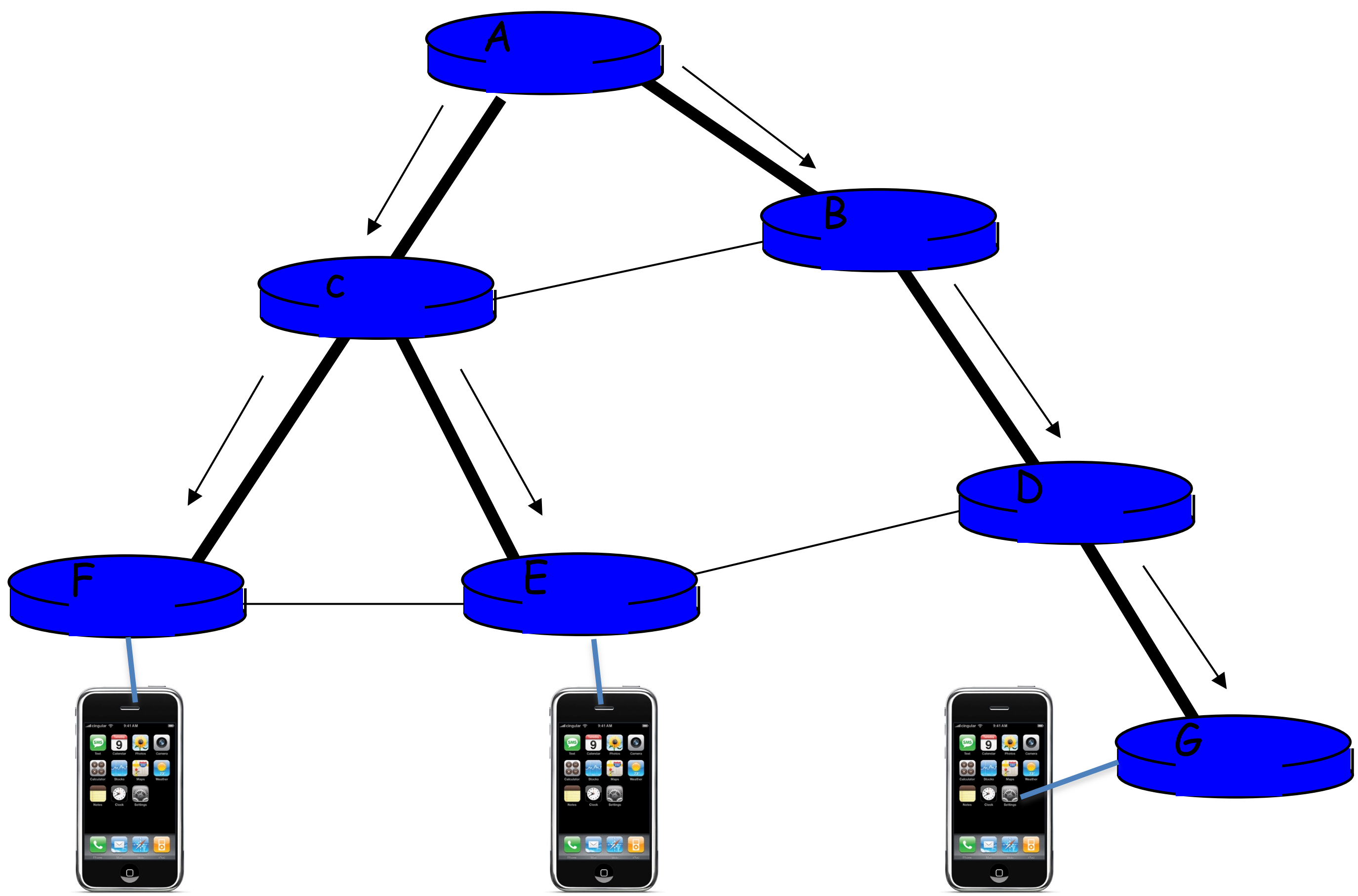


Routing Table from Router 1:

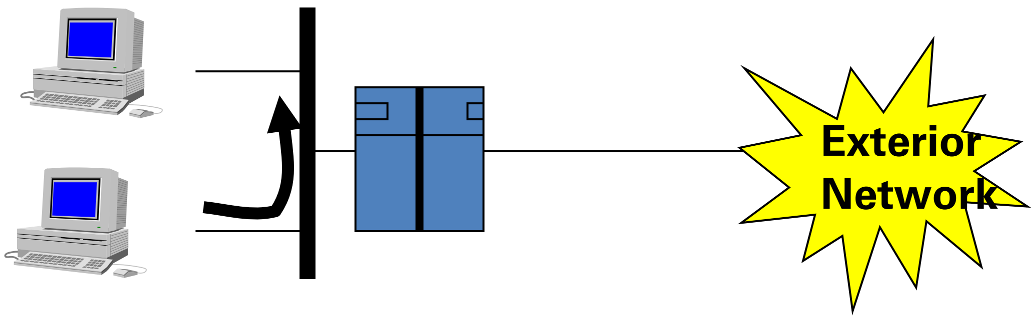
Destination	Mask	Next-Hop	Distance
192.168.0.0	/29	127.0.0.1	0
239.1.1.1	/32	192.168.0.1	1
239.1.1.1	/32	192.168.0.2	2

Example Multicast Protocol

- Receiver sends a "join" messages to the sender
  - And grafts to the tree at the nearest point

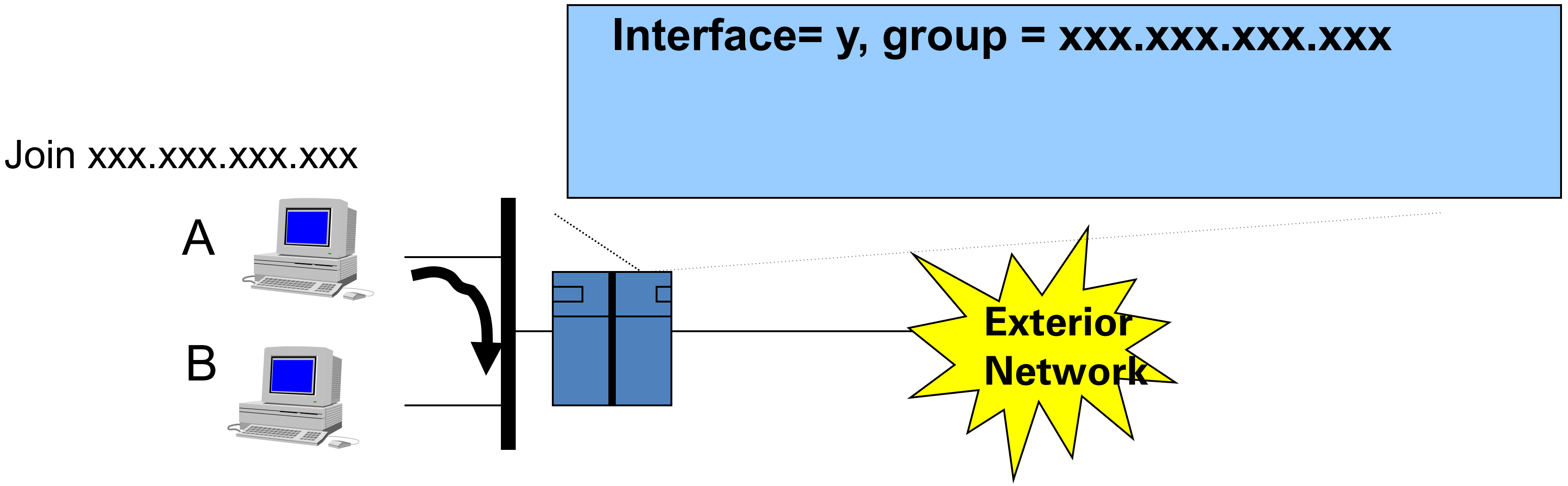


Multicast Datagram Delivery



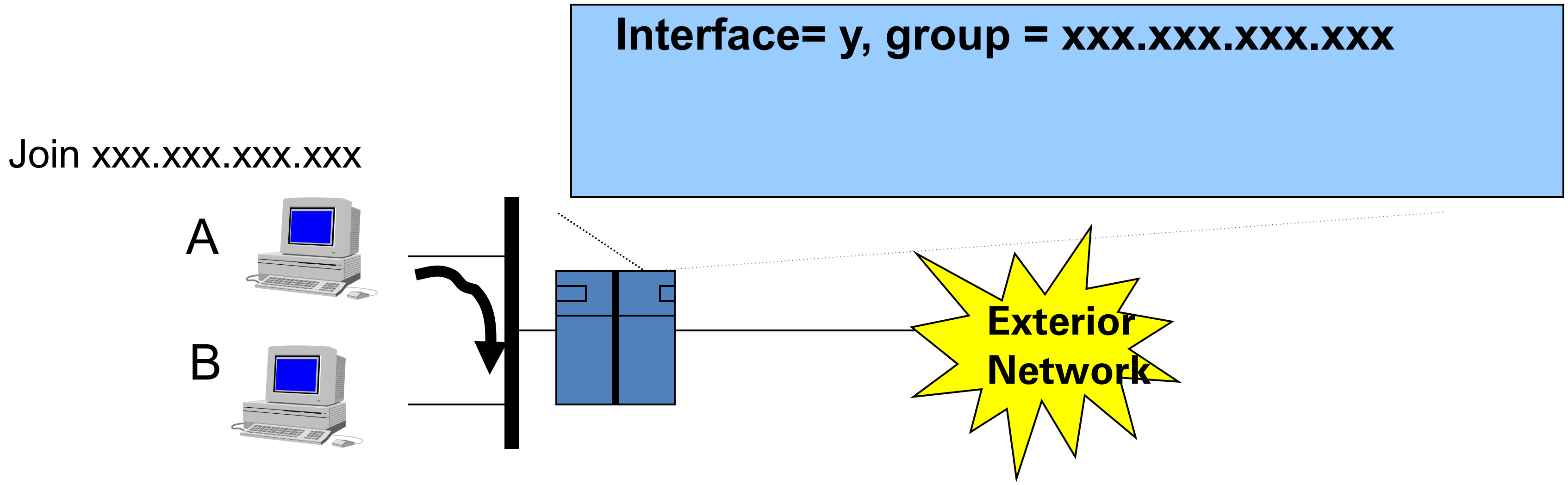
- Same network: hosts can always send and receive locally generated multicast packets by themselves. (Sender switches the IP addr to Ethernet (MAC) addr, receiver sets its network interface to receive pcks.).
- But to outside, multicast router must be present.

Internet Group Management Protocol (IGMP)



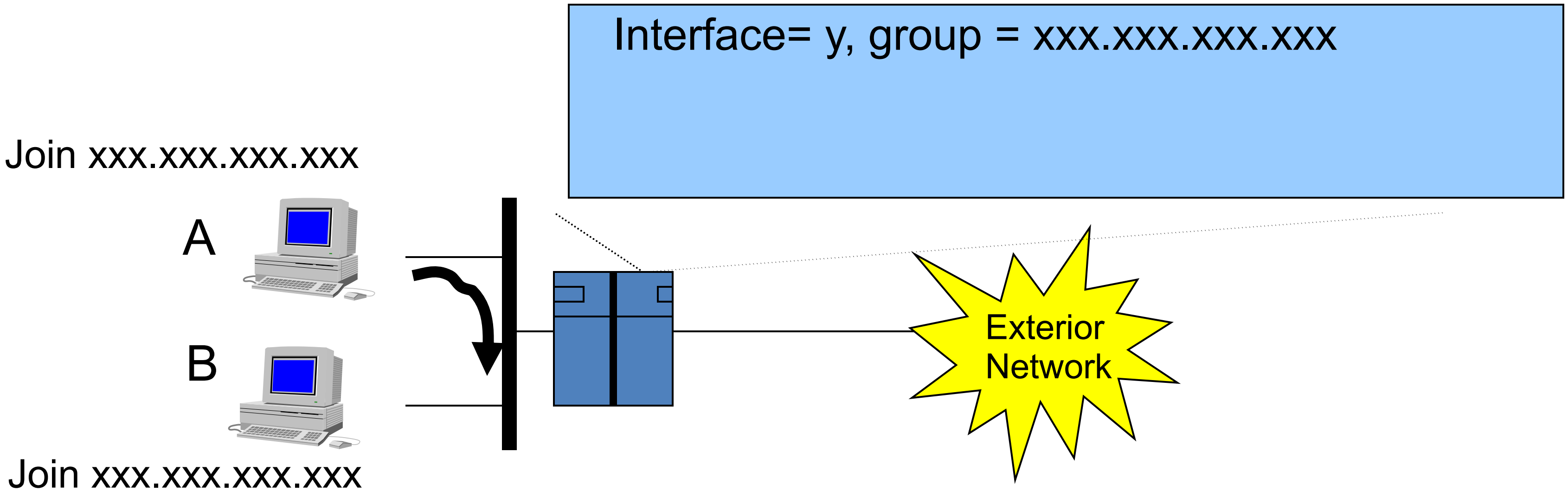
- Hosts subscribe to a multicast group by issuing IGMP host membership report messages.
  - Multicast to all hosts group with TTL 1.

Internet Group Management Protocol (IGMP)



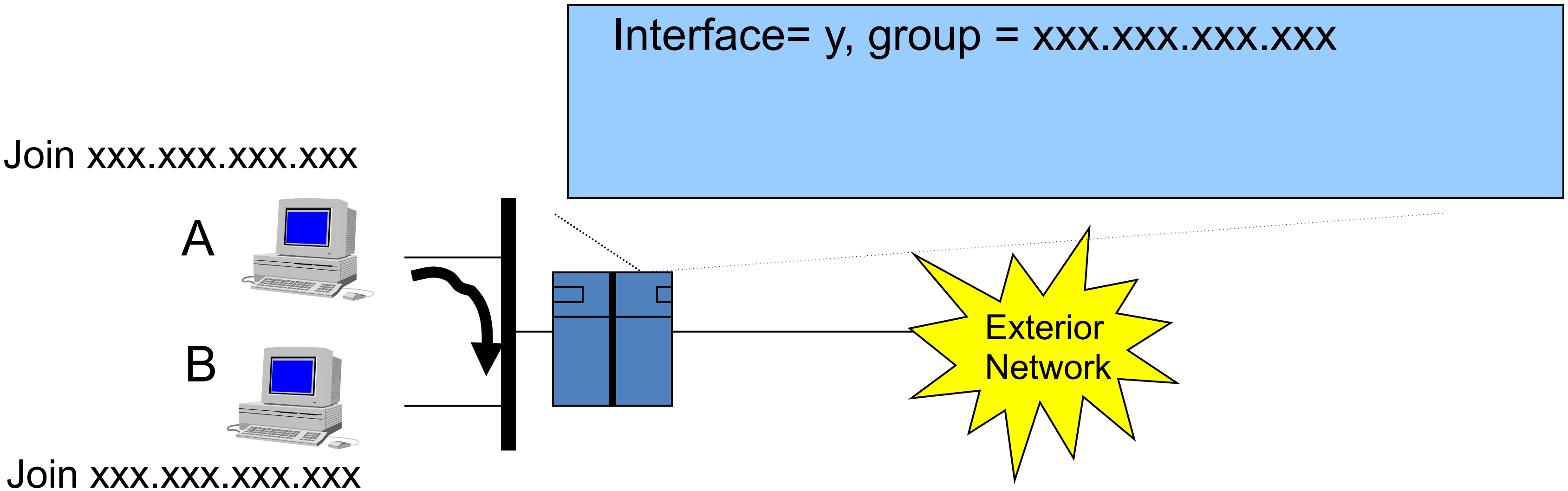
- A local multicast router….
  - Records the group address.
  - Begins forwarding packets destined for that group into its own network when receiving multicast packets.

IGMP Operation



- Multicast routers need not keep track of host addresses.
  - Configure to receive all multicast messages.

IGMP Operation



- Routers advertises group memberships to other routers.
- Hosts can join and leave groups dynamically.

IGMP v1

- Two types of IGMP msgs (both have IP TTL of 1)
  - Host membership query: Routers query local networks to discover which groups have members
  - Host membership report: Hosts report each group (e.g., multicast addr) to which belong, by broadcast on net interface from which query was received
- Routers maintain group membership
  - Host senders an IGMP "report" to join a group
  - Multicast routers periodically issue host membership query to determine liveness of group members
  - Note: No explicit "leave" message from clients

IGMP: Improvements

- IGMP v2 added:
  - If multiple routers, one with lowest IP elected querier
  - Explicit leave messages for faster pruning
  - Group-specific query messages

- IGMP v3 added:
  - Source filtering: Join specifies multicast "only from" or "all but from" specific source addresses

IGMP: Parameters and Design

- Parameters
  - Maximum report delay: 10 sec
  - Membership query internal default: 125 sec
  - Time-out interval: 270 sec = 2 * (query interval + max delay)
- Router tracks each attached network, not each peer
- Should clients respond immediately to queries?
  - Random delay (from 0..D) to minimize responses to queries
  - Only one response from single broadcast domain needed
- What if local networks are layer-2 switched?
  - L2 switches typically broadcast multicast traffic out all ports
  - Or, IGMP snooping (sneak peek into layer-3 contents), Cisco's proprietary protocols, or static forwarding tables

IP multicast often best effort

- Application protocols on top of UDP
  - Within enterprises
  - Commercial stock exchanges
  - Multimedia content delivery
    - Streaming audio, video, etc.
    - Everybody in group listening/watching same content
    - IPTV
  - Many applications insensitive to loss, and networks managed/provisioned so little/no loss

Multicast Routing

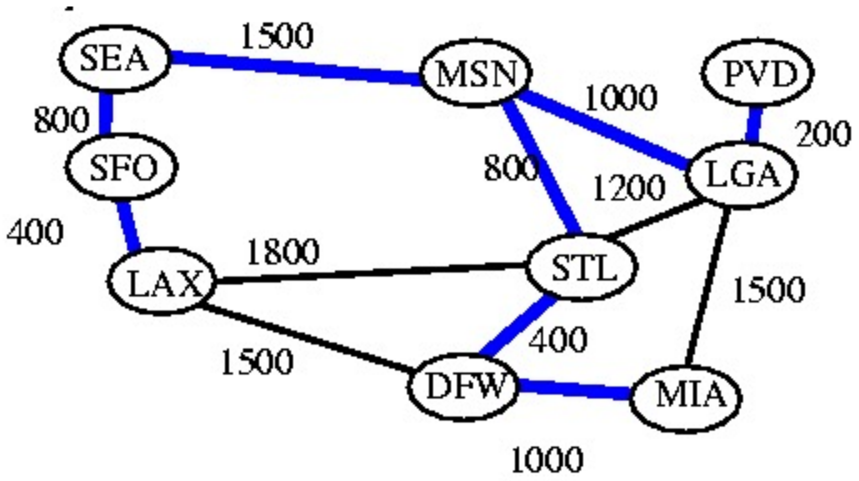
- Reverse path multicast
- Core-based Trees
- PIM
  - Dense mode
  - Sparse mode

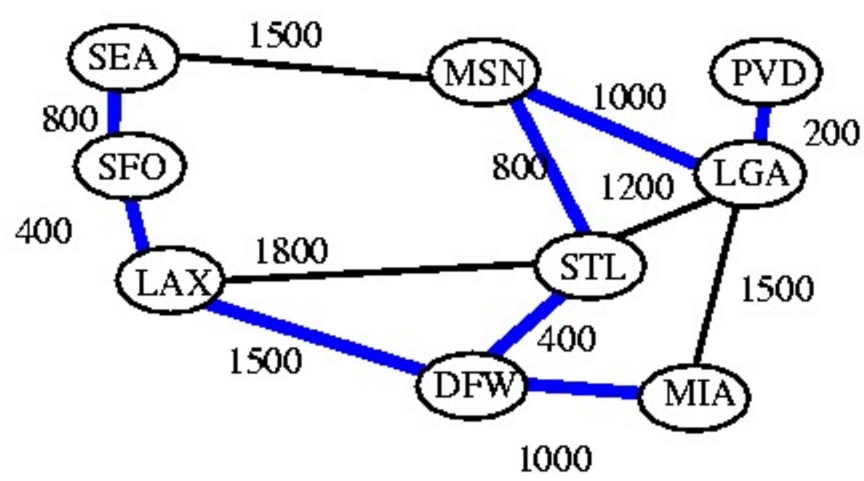
Multicast Tree

- Shared tree.
  - Single tree for all sources.
  - Spanning tree: tree that connects all the vertices (hosts and routers).
    - Minimum spanning tree (MST).
    - More...
- Per-source tree.

MST

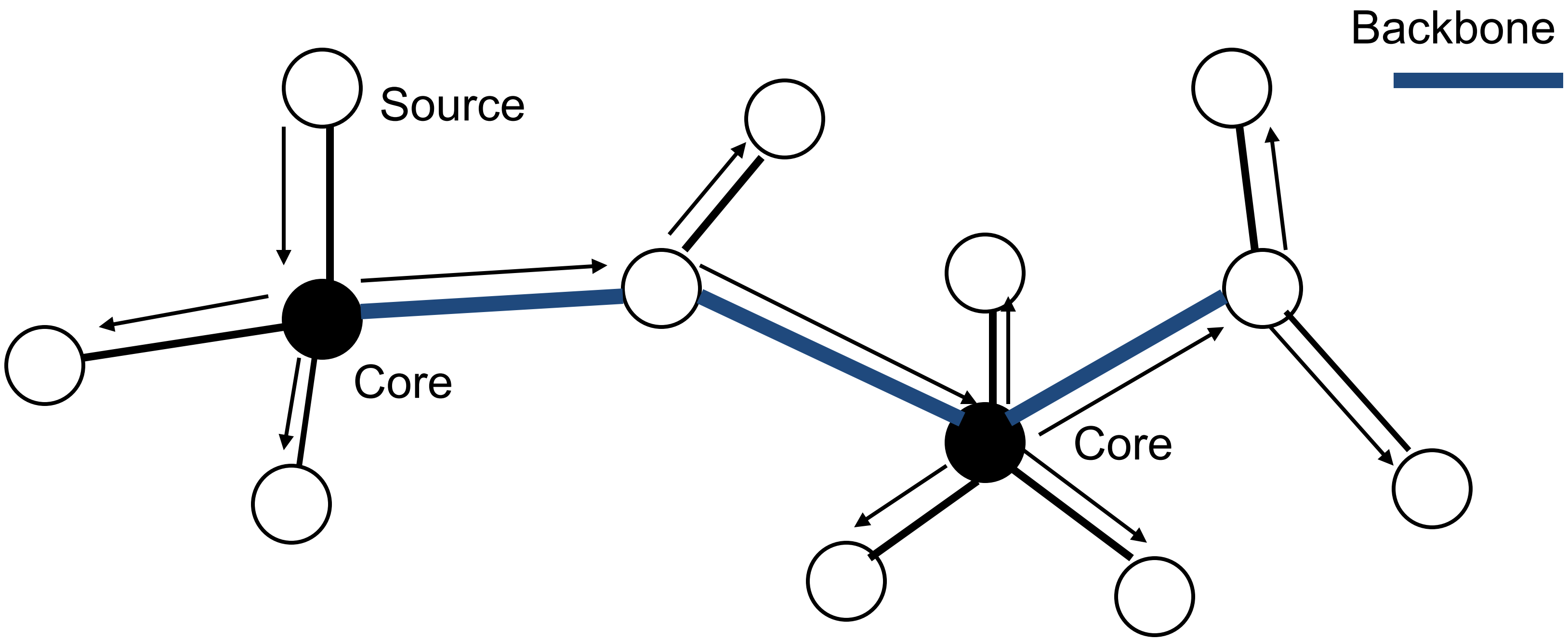
- Prim's and Kruskal's MST algorithms.
- Minimize costs involving all nodes.





Core-based Tree (CBT)

- One or more routers form the core tree of a multicast backbone.

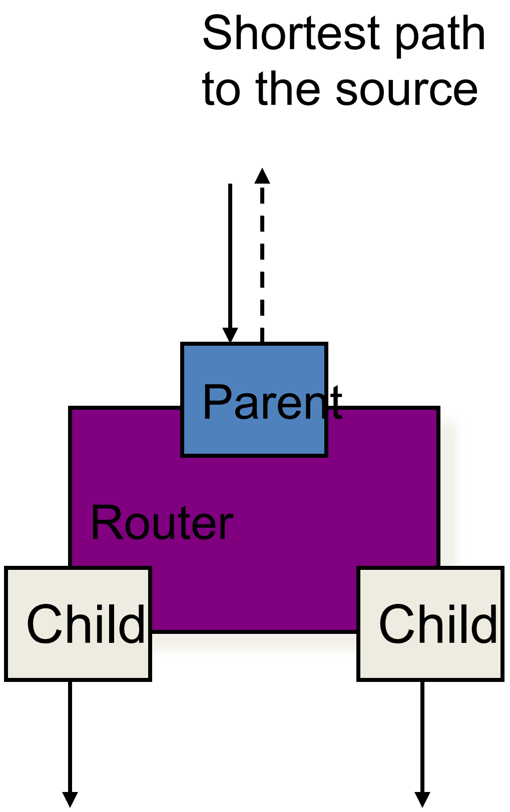


CBT

- Non-core routers send join requests toward core routers.
- Routers know the locations of nearest core routers.
- Forward packets to all oifs (outgoing interface) on the CBT (except the one the packet came from).

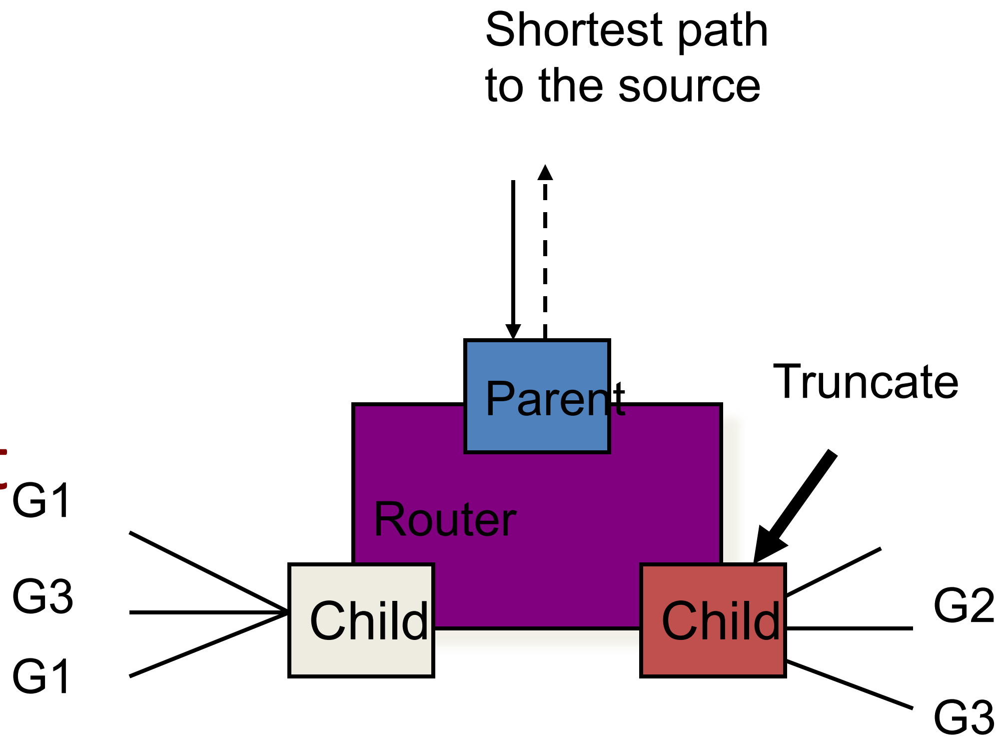
Reverse Path Broadcast

- iif: incoming interface; oif: outgoing interface
- If iif is on the shortest path to source S,
  - forward to all other oifs.
  - avoid forwarding duplicates.
- Can be enhanced:
  - Forward to oif if that next hop router considers “me” as a parent.



Truncated Reverse Path Broadcasting (TRPB)

- Using IGMP, if no member exists in a leaf network, its router does not forward.

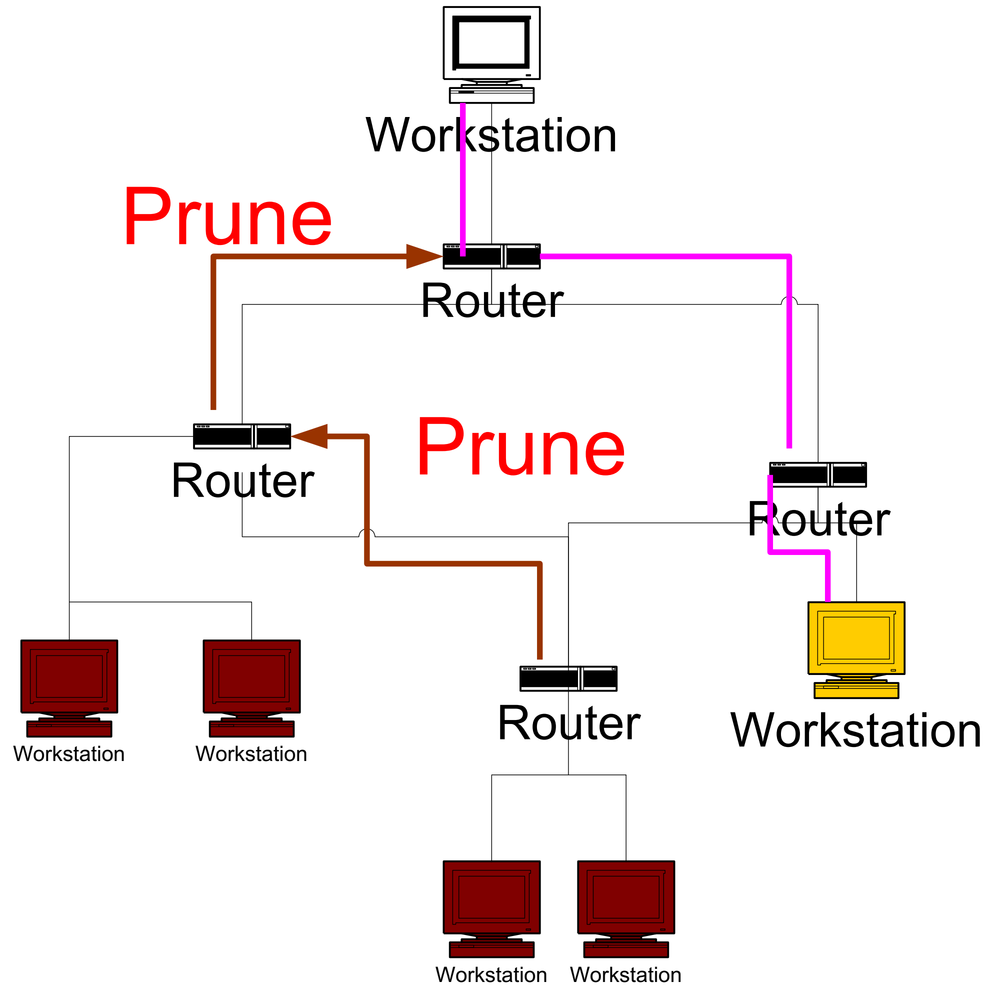


Reverse Path Multicasting (RPM)

- Creates a delivery tree that spans only
  - Subnetworks with group members, and
  - Routers and subnetworks along the shortest path to subnetworks with group members.
- Initially TRPB.
- Each router that does not have any subnetworks with a group member sends "prune" message to its parent and its parent does the same thing.

RPM

- Drawbacks
  - Periodic flooding
  - Each router needs to maintain state information for all groups and each source.



Protocol Independent Multicast (PIM) - DM

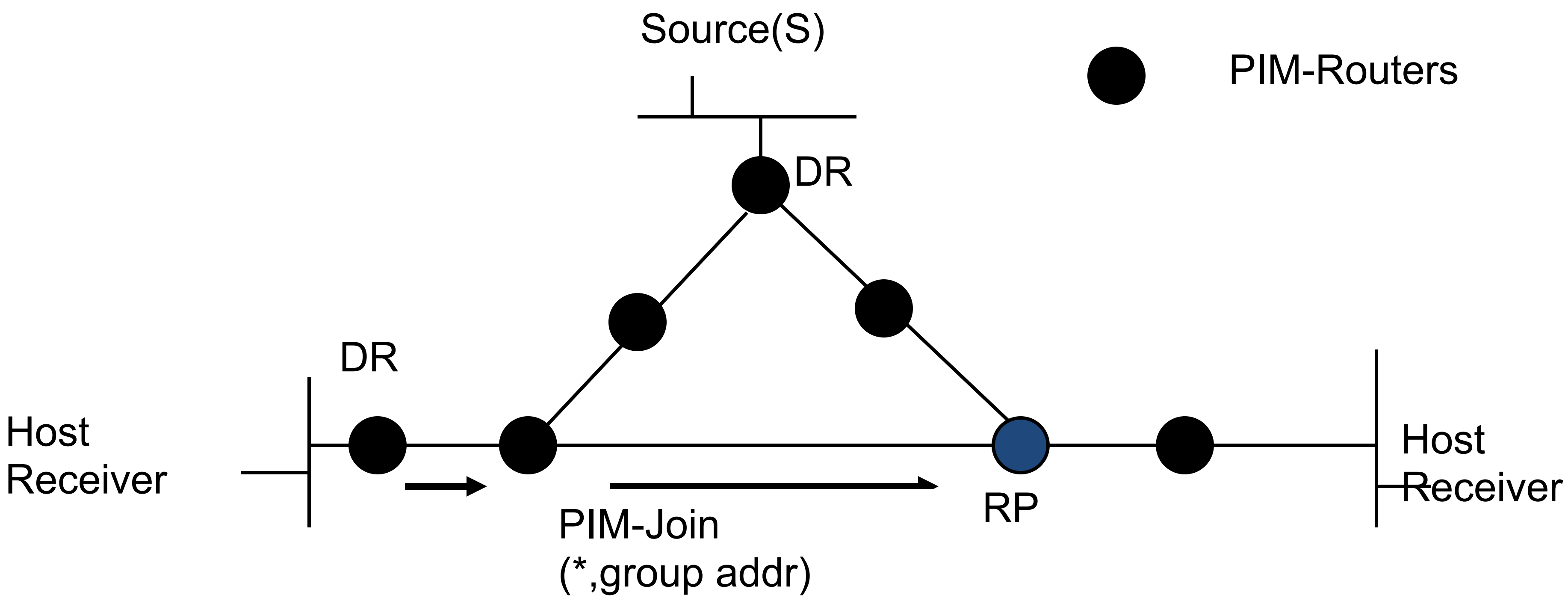
- Forward to all outgoing interfaces (unless pruned).
- Allow packet duplication and avoid any overhead in building trees.

PIM-SM

- Use CBT (source based).
- Default setting: don't forward unless explicitly told to do so.
  - Contrast with Dense mode: the default setting is "forward unless explicitly told otherwise".
- Core routers are called rendezvous point(RP) – a group specific.
  - one primary RP selected or elected for a group.

PIM-SM: Member Join

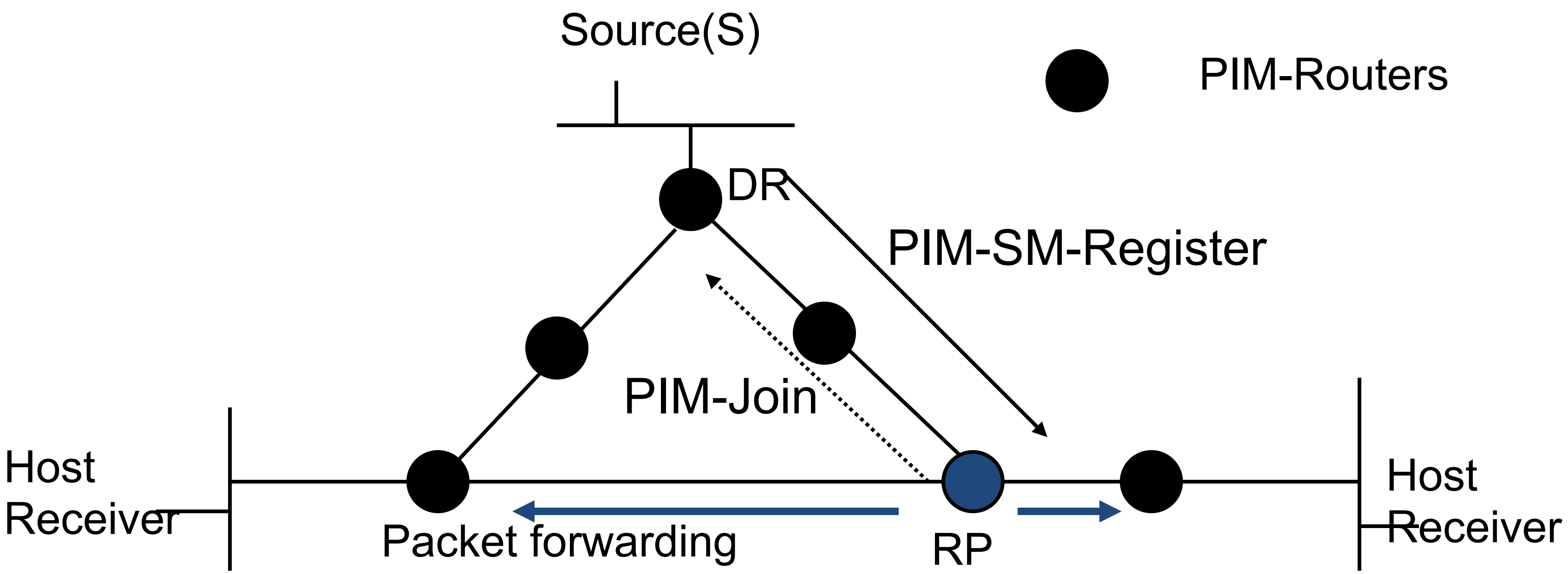
- Each receiver contacts its directly attached router (DR), which in turn joins the multicast tree by sending a PIM-Join message to RP (*,group addr) -- * indicates any source. This turns on the interfaces on the path for multicast forwarding for that address.



PIM-SM: Sending Multicast Packets.

- When a sender first transmits a multicast packet to a group,
  - The DR of the source must forward the first data packet encapsulated in a unicast PIM-SM-Register packet to the primary RP of the group.
  - RP learns of a new source, and then RP transmits PIM-Join message back to the source's DR.
  - Routers between DR and RP set their interfaces for forwarding toward the RP from the PIM-JOIN messages

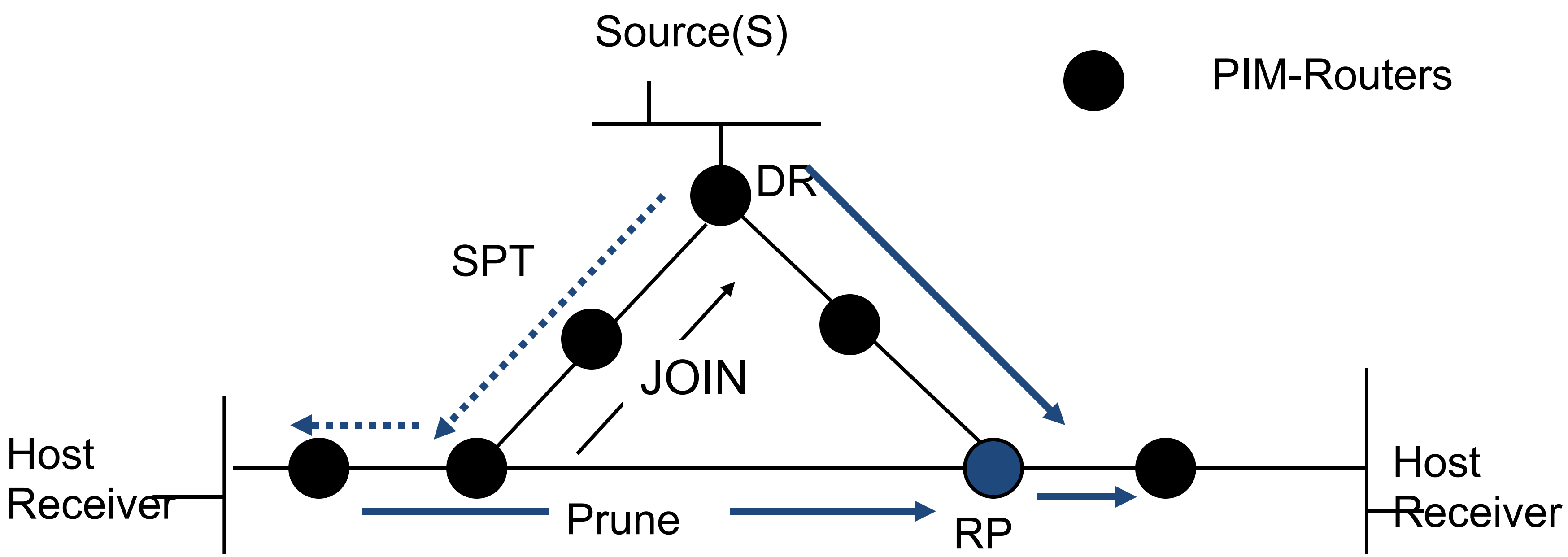
PIM-SM: sending multicast



- Next packets are multicast without encapsulation.

RP-Shared Tree or Shortest Path Tree (SPT)

- PIM-SM gives an option for receivers to switch to a source-rooted SPT.



PIM-SM Problems

- Single point of failure
- Hot spot
- non-optimal path
- Complexity