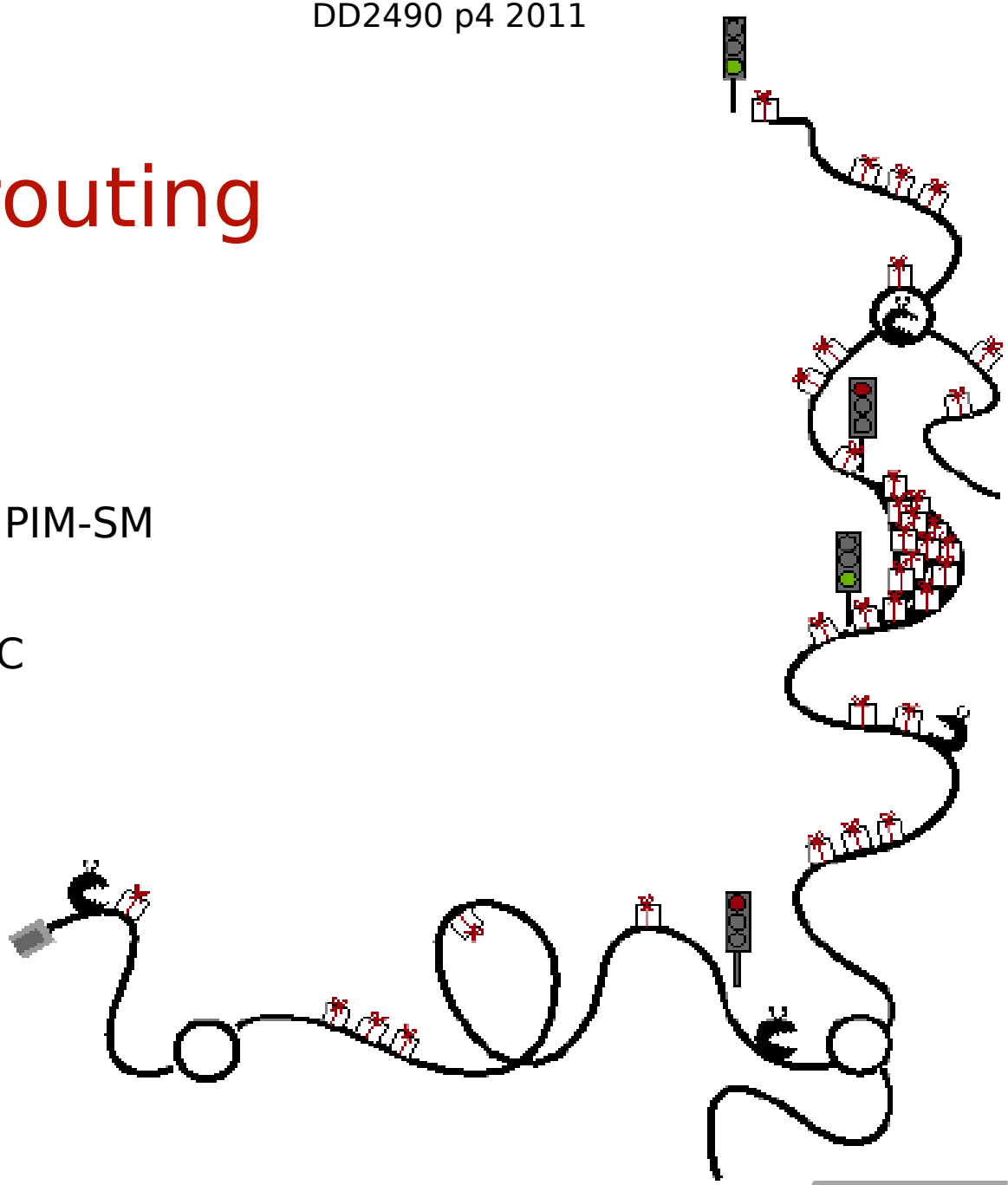
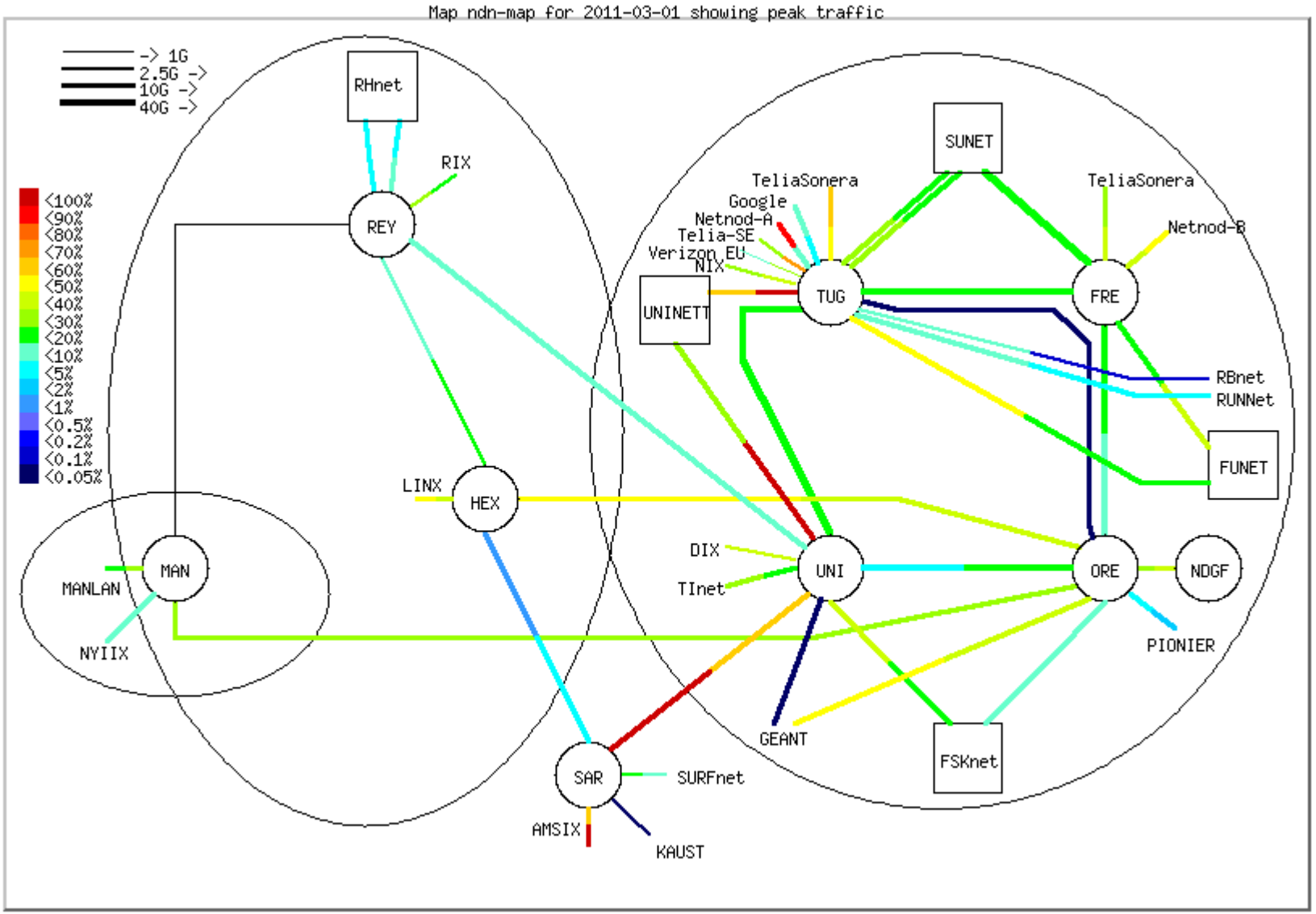
IP Multicast routing

Multicast routing and PIM-SM

Olof Hagsand KTH CSC



Example: Ski world cup 2011



Literature

- RFC 4601 Section 3 (you may need some definitions from Section 2). See reading instructions on web.

Deployment

- Multicast *routing* is in general not deployed in current networks
- However multicast for *control* is widely used
  - –Routing and other protocols use multicast locally
- Some sites, eg ”stadsnät” and ISPs have deployed local multicast delivery. IPTV
  - –IPTV distribution is commonly using multicast
- IP Multicast is not gaining the attention it deserves

IPTV - today

- A multicast driving force today is IPTV - a component of ”triple-play”
  - You can also distribute video via unicast: http (eg youtube), p2p (eg bit-torrent), streaming(eg rstp),...
- An operator has a *head-end* from where it distributes video content
  - Terestial, Satelite, hard-disk,....
- The operator distributes the video using IP multicast from its head-end via its distribution network using IP multicast routing. Typically based on MPLS.
- From the routers, an L2 network, via xDSL modems, distributes the video the last step.
- Video is encoded using MPEG-2 (ca 3-4Mb/s) or MPEG-4 (HD 12-16Mb/s)
- Encapsulation is either MPEG directly over UDP (DVB) or MPEG over RTP over UDP
- One video channel per multicast group
  - When you change channel, you change group

IP Multicast Overview

- *Receiver-based* multicast:
  - –Senders send to any group
  - –Receivers join groups
- *Dynamic group membership*
  - –Hosts leave and join groups dynamically
- Group addresses (class D)
- Uses multicast in hardware if available
- Best-effort delivery semantics (unreliable)
- Notation:
  - –(S, G) - specific sender S to group G
  - –(*, G) - all senders to group G
- Prime architect: Steve Deering

IP Multicast Addresses

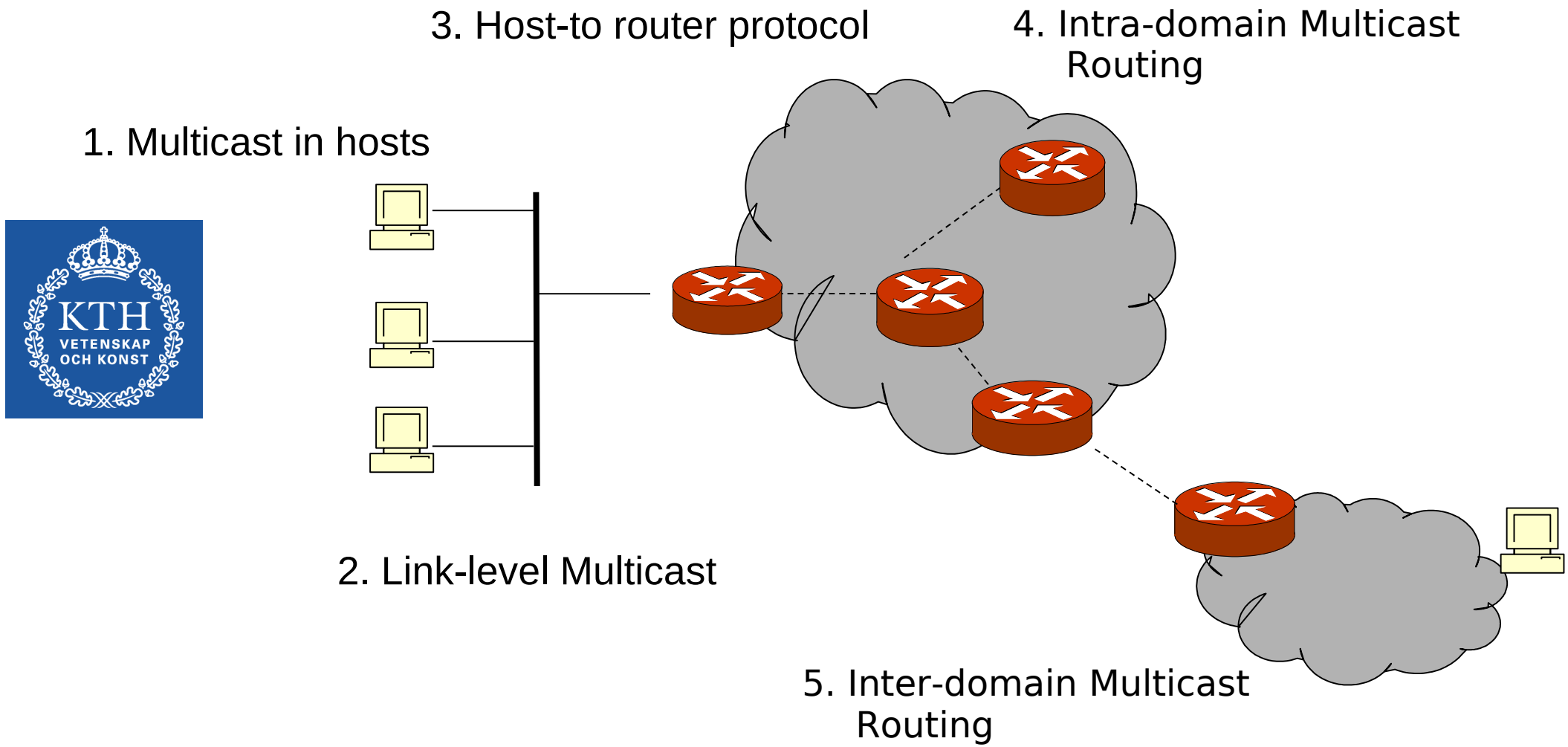
- IP-multicast addresses, class D addresses (binary prefix: 1110/4)

  `224.0.0.0 - 239.255.255.255`

- 28 bit multicast group id
- Selected addresses reserved by IANA for special purposes:

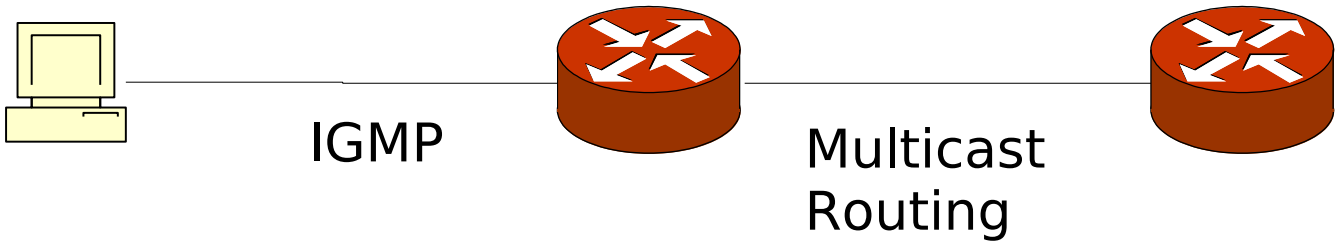
Address	Description
`224.0.0.0 – 224.0.0.255`	Local Network Control Block (dont forward)
`224.0.0.1`	All Systems on this subnet
`224.0.0.2`	All Routers on this subnet
`224.0.0.4`	DVMRP Routers
`224.0.0.9`	RIP Routers
`233/8`	GLOP addressing
`232/8`	Source-specific multicast

IP Multicast Architecture



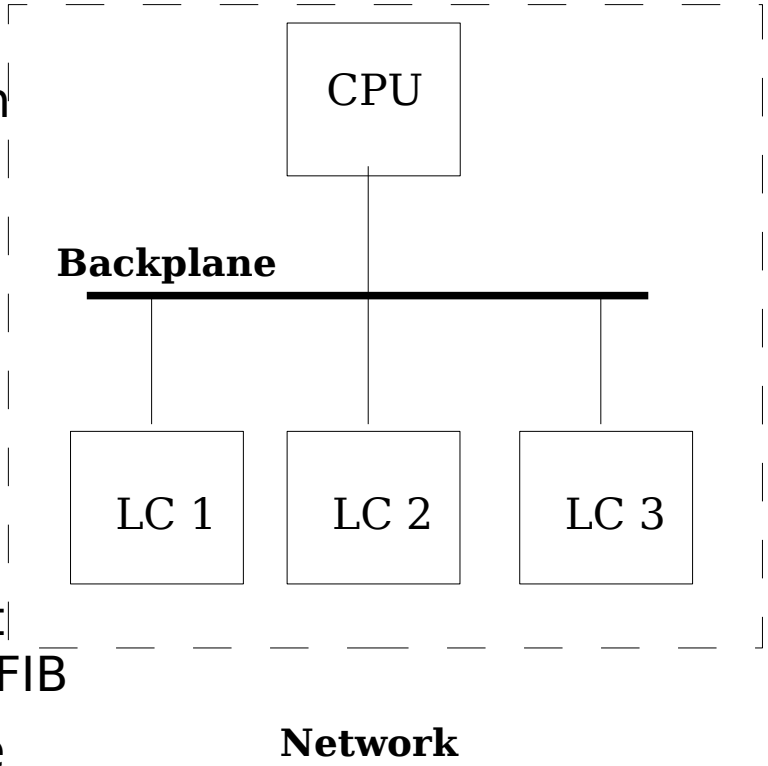
Multicast router

- Listens to all multicast traffic and forwards if necessary.
- Multicast router listens to *all* multicast addresses.
  - Ethernet: 2^{23} link layer multicast addresses
  - Listens *promiscuously* to all LAN multicast traffic
- Communicates with directly connected hosts via *IGMP*
- Communicates with other multicast routers with *multicast routing protocols*



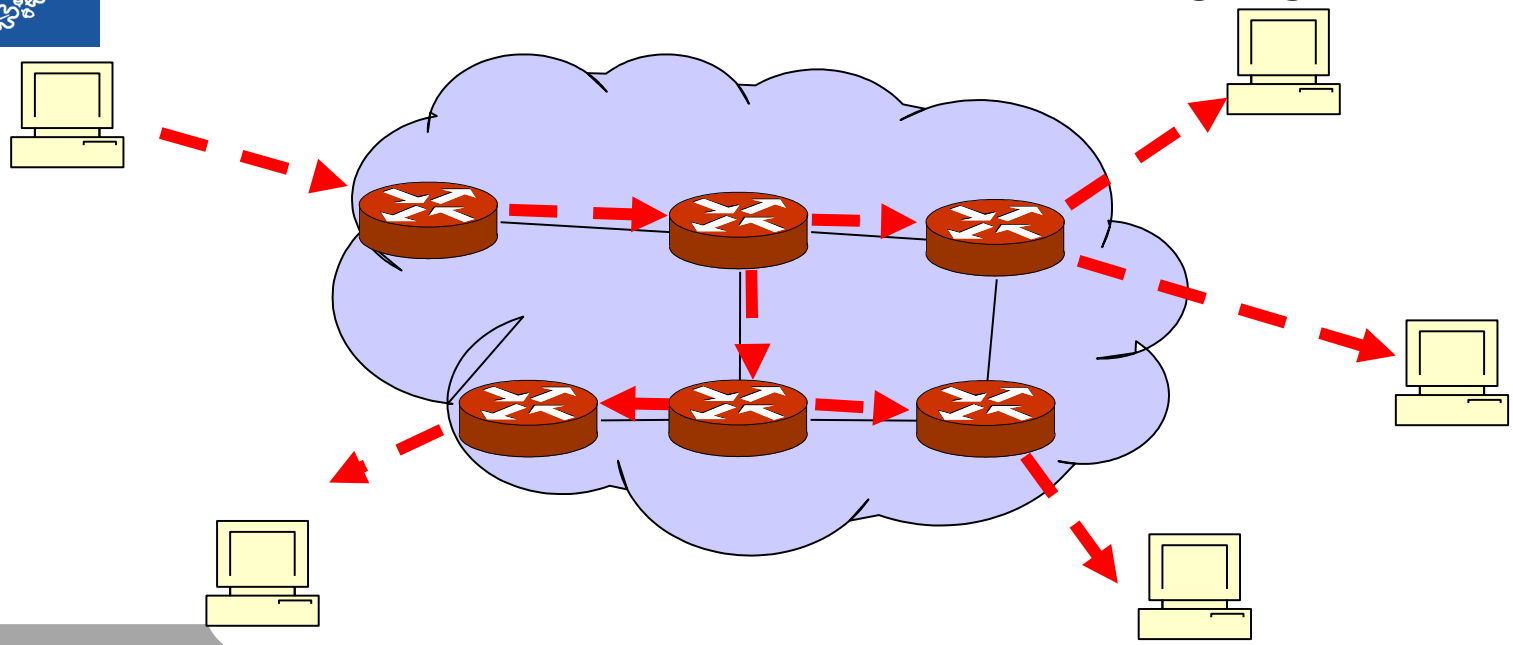
Multicast in routers

- Packets are replicated to many output ports
- Forwarding used to be done in slowpath – in the main CPU
- Modern routers can do forwarding in the linecards or in hardware
- Replication can be made in:
  - –incoming linecards,
  - –backplane
  - –outgoing linecards
- Routes computed by main CPU by multicast routing protocol are installed in the linecard FIB
- In multicast forwarding you look at both the destination and source IP address



Multicast Routing

- A packet received on a router is forwarded on many interfaces
- The network replicates the packets – not the hosts
- All routing protocols aim at building delivery trees through a network
- Mostly multicast routing is more concerned where packets *come from* instead of where they are going
- This is why *unicast routing algorithms* are used to find the shortest path to the single source.
  - You do not need a multicast *routing* algorithm !



Delivery Trees

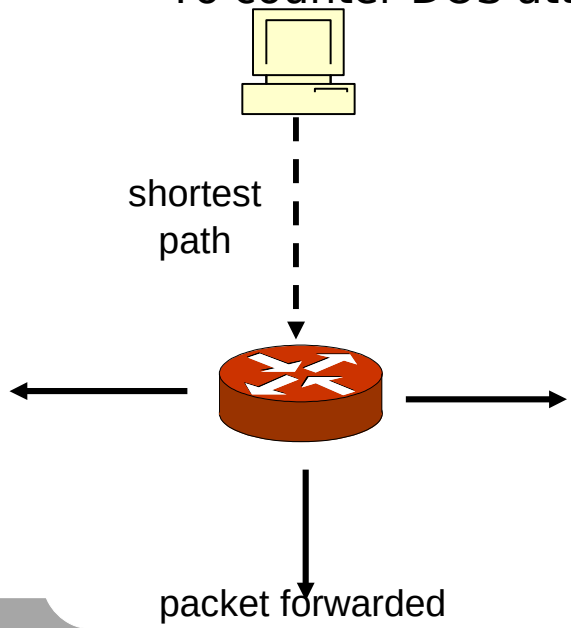
- Group Shared Trees: (*, G)
  - –Same delivery tree for all senders
  - –Router state: O(G)
  - –But suboptimal paths and delays
  - –Unidirectional or Bidirectional
  - –Tree's root is in Rendez-vous point (RP)
  - –Also called RP Trees.
- Source Based Trees: (S, G)
  - –Different delivery tree for each sender
  - –Router state: O(S*G),
  - –Optimal paths and delay.
- Data-driven / Push
  - –Trees built when data packets are sent
- Demand-driven / Pull
  - –Build when members join

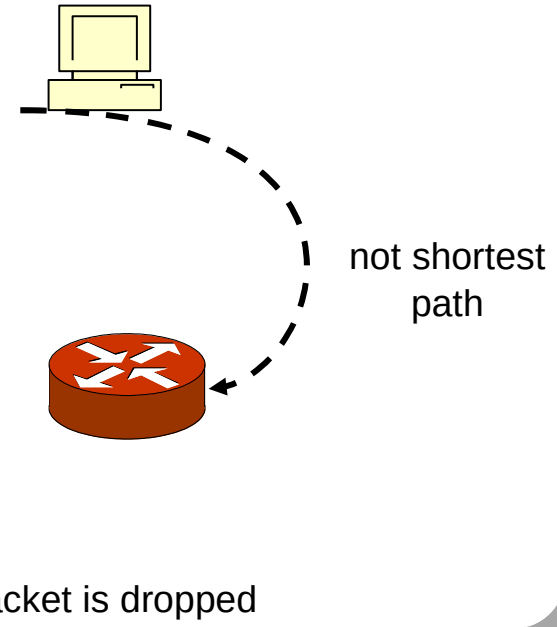
Protocol classification

- Dense-mode protocols
  - –Intended for domains with high receiver density
  - –Push, Source trees
  - –Examples: PIM-DM, DVMRP
- Sparse-mode protocols
  - –Intended for domains with small receiver density
  - –Pull, shared trees (not always,..)
  - –Examples: PIM-SM, CBT
- Link-state protocols
  - –Source trees
  - –Examples: MOSPF

Multicast forwarding step 1: RPF

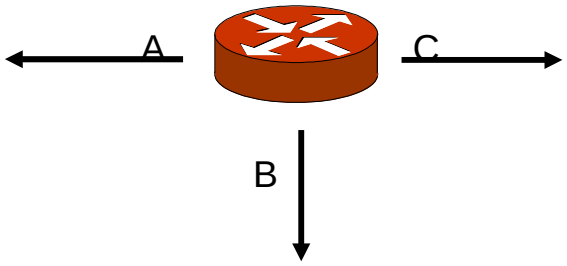
- Reverse Path Forwarding
- Basic forwarding principle in multicast
- Forward datagram only if it arrives on interface used to send unicast to source
  - –or rendez-vous-point (RP) if (*, G) entry
  - –Send out on all other interfaces: Flooding!
- You can re-use Unicast routing tables for multicast
  - –but used in reverse
- RPF can also be used in unicast to detect source address spoofing
  - –To counter DOS attacks





Multicast forwarding step 2

- Keep a table with (source, group) entries and an interface forwarding list
- This is the delivery tree built by the multicast routing protocol
- Data driven (push): Entries are created when data is sent
- Demand-driven (pull): Entries appear when receivers join.



Sender, group	interface list
*, 231.2.3.4	A, B, C
10.0.0.1, 231.2.3.4	A
10.0.2.1, 231.2.3.4	C
*, 229.5.6.7	B, C

Exercise1: Multicast forwarding

Prefix	Outgoing IF
180.70.65.0/24	A
201.4.22.0/24	C
120.14.96.0/24	B
153.18.16.0/24	D
0.0.0.0/0	A

- A router has interfaces A, B, C, D
- The router has a unicast forwarding table and a multicast forwarding table
- How are the following IP packets forwarded by the router?
  - –A packet with src = 180.70.65.12 and dst = 228.3.2.2, arriving on interface B
  - –A packet with src = 201.4.22.4 and dst = 228.3.2.2, arriving on interface C
  - –A packet with src 120.14.96.42 and dst = 241.16.53.2, arriving on interface B
  - –A packet with src = 153.18.16.3 and dst = 234.12.32.5. arriving on interface D
- Solution on web after the lecture

Sender, Group	Outgoing IF list
*, 228.3.2.2	A, B, C
201.4.22.4, 228.3.2.2	A, D
*, 234.12.32.5	A, C, D

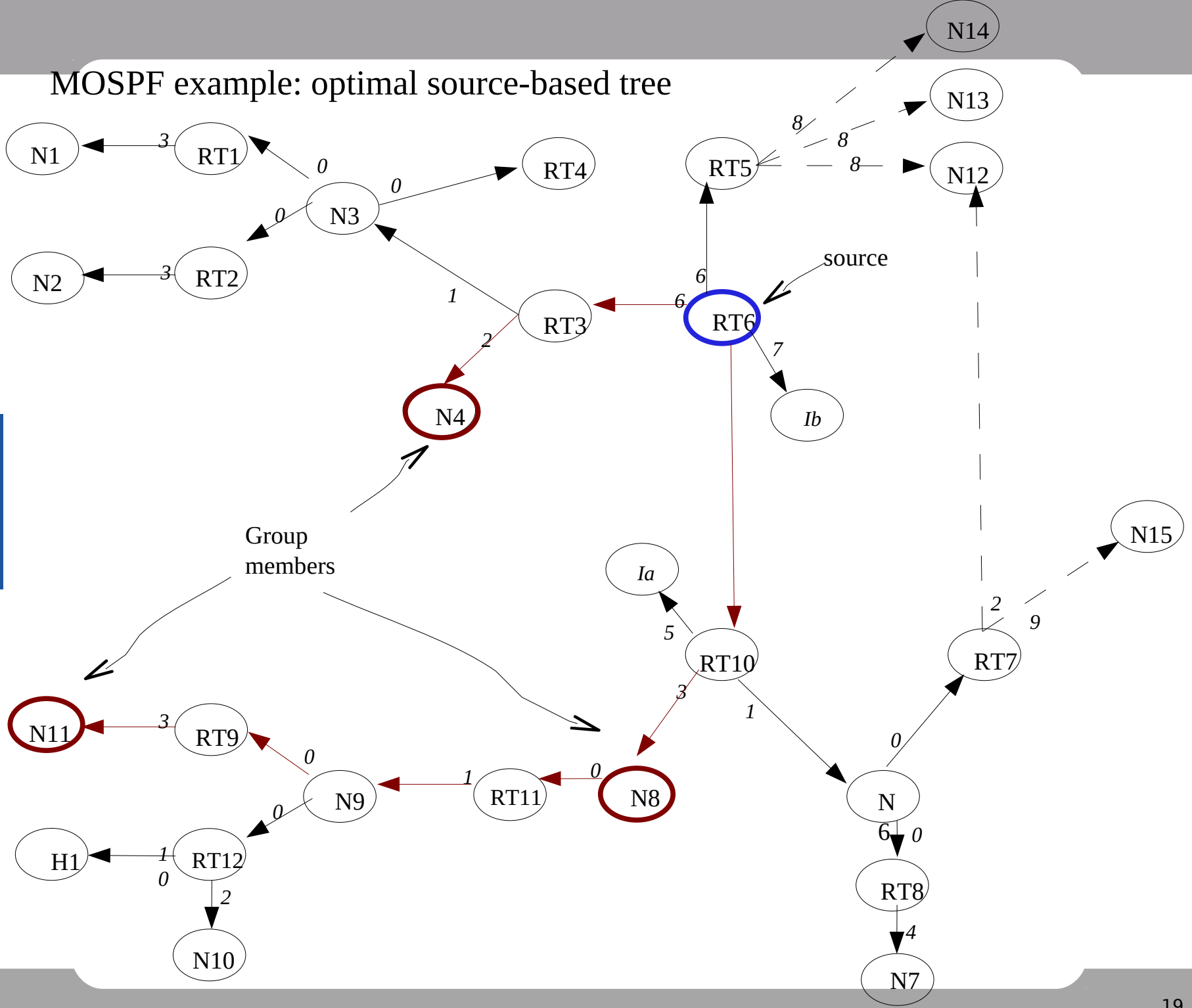
DVMRP

- Distance-Vector Multicast Routing Protocol
  - –Based on unicast distance vector (eg RIP)
- DVMRP was used in the first multicast Internet
  - –The MBone.
  - –Unicast tunnels connected multicast islands.
  - –Mainly academic use
- DVMRP is data-driven/push and uses source-based trees
- DVMRP builds Truncated Broadcast Trees
  - –Reverse Path Broadcasting
  - –Builds routing tables with source networks
  - –Uses poison reverse to detect down-link routers
- DVMRP floods trees with data
  - –Then prunes and grafts tree depending on receiver dynamics.

Link-State Multicast: MOSPF

- Add multicast to a given link-state routing protocol
- Extend LSAs with *group-membership LSA (*type 6)
  - –Only containing members of a group
- Uses the link-state database in OSPF to build delivery trees
  - –Every router knows the topology of the complete network
  - –Least-cost source-based trees using metrics
  - –One tree for all (S,G) pairs with S as source
- The delivery trees are optimal, since we use Dijkstra
- Expensive to keep all this information
  - –Cache active (S,G) pairs
  - –MOSPF is Data-driven/push: computes Dijkstra when datagram arrives
- If OSPF is used for unicast routing, it is easy(in principle) to extend it for multicast routing as well

MOSPF example: optimal source-based tree



Protocol Independent Multicasting - PIM

- PIM relies on any unicast routing tables for its RPF check
  - –Does not build its own tables (as DVMRP)
- Split into two protocols for different uses
  - –PIM-DM – PIM Dense mode
  - –PIM-SM – PIM Sparse mode

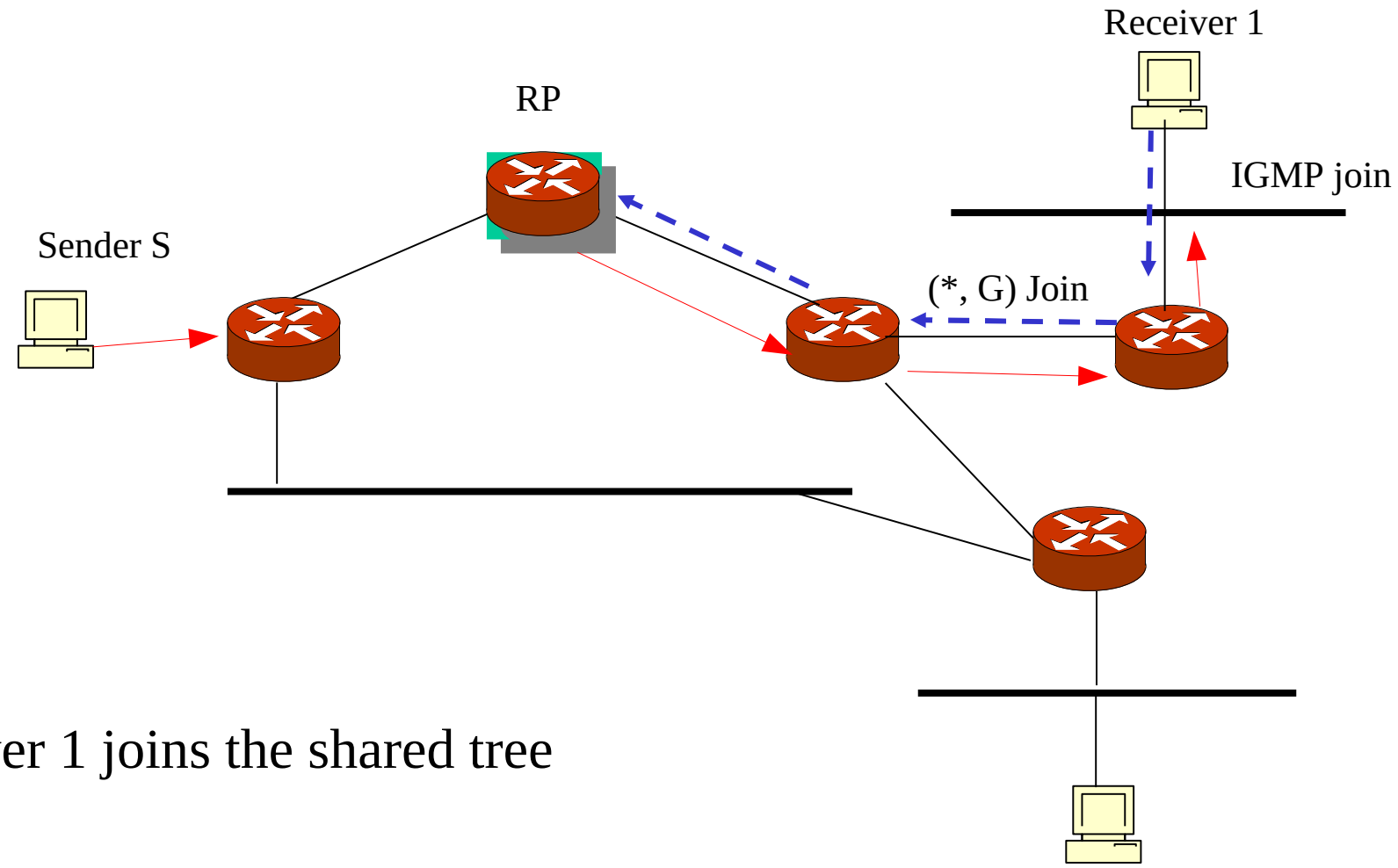
PIM – Dense Mode

- PIM - Dense Mode
  - –Flooding-and-prune strategy – similar to DVMRP
  - –Creates (S,G) in every router (even if no receivers)
  - –Ca 3-minute cycle
- Uses unicast routing tables for RPF
  - –No separate multicast routing protocol
- Builds only source-based trees
- PIM Hello messages – to detect neighbor PIM routers
- PIM Prune messages
  - –Like DVMRP, but prunes also used instead of poison reverse in DVMRP
- PIM Assert messages
  - –To elect one single forwarder on a shared link
- PIM Graft – just like DVMRP

PIM-SM

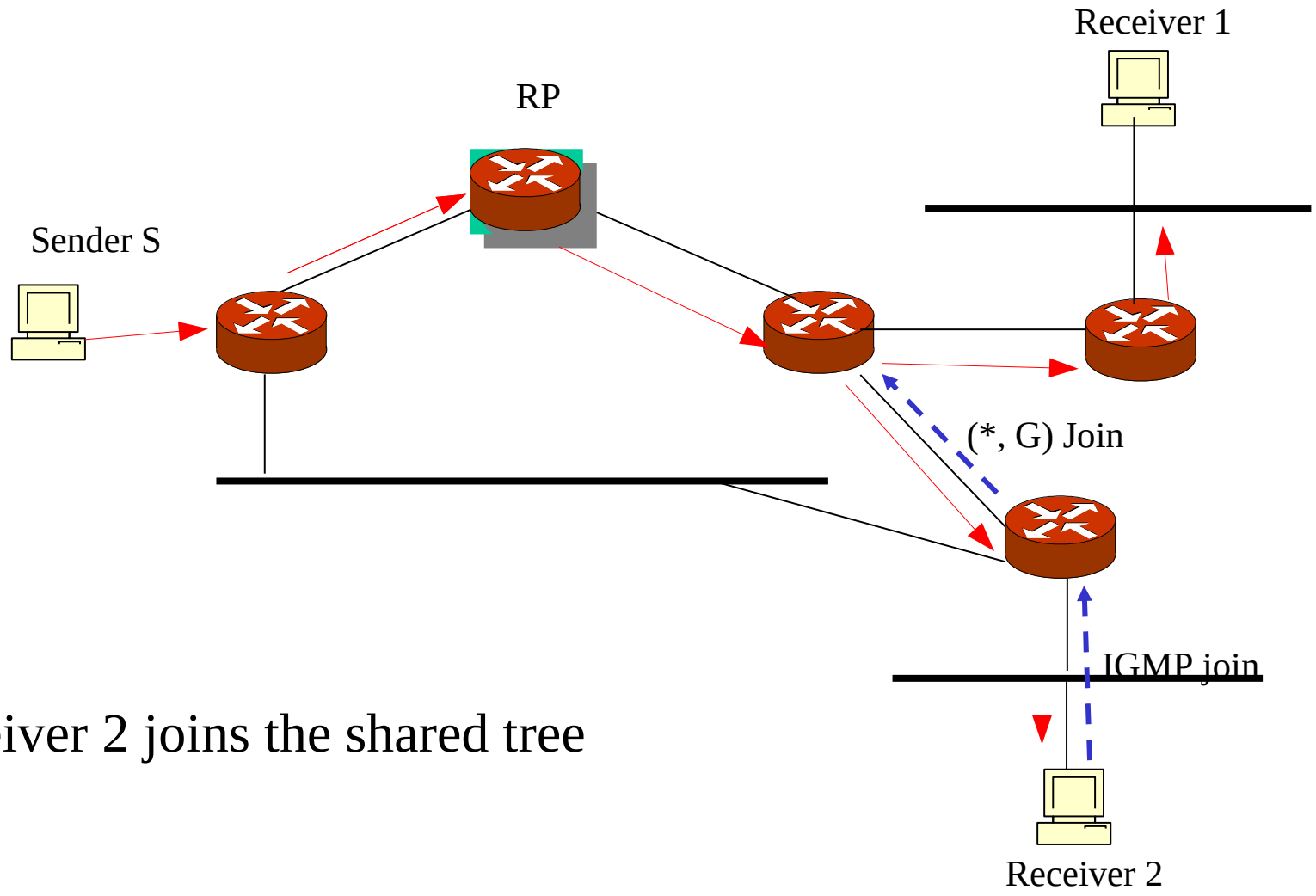
- Protocol-Independent Multicast - Sparse Mode
- Start with a shared tree
- When receivers join (via IGMP) PIM-Joins travel up to rendez-vous point to join a shared tree
- When sources start sending traffic, the sources are registered with the RP
- When new source traffic reaches receiver, a source based-tree is built
  - –Source-base tree switchover
- How do routers find the Rendez-vous point (RP)?
  - –RP discovery
- There is also a bidirectional variant of PIM-SM

PIM-SM join example



Receiver 1 joins the shared tree

PIM-SM join example



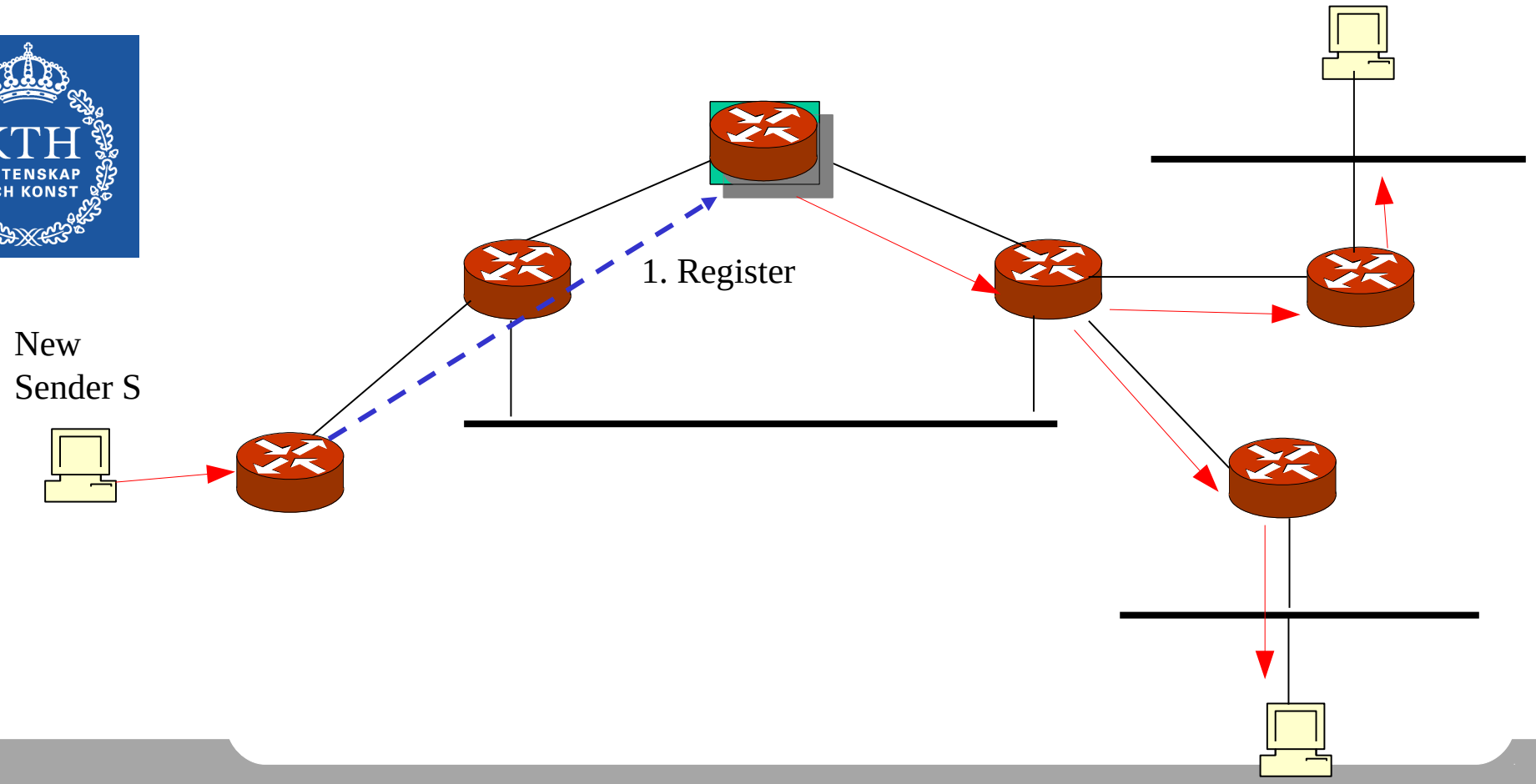
Receiver 2 joins the shared tree

Registering sources – PIM Register

- PIM-SM trees are unidirectional, sender's DR must send to RP to distribute traffic.
  - –All routers must know the Group-to-RP mapping
- In bidirectional PIM, a sender can just reach any router in the tree
- Sender's DR sends PIM Register messages towards the RP
  - –Multicast data is encapsulated within the unicast Register message
- The RP then joins the source tree!
  - –Sends a (S,G) Join towards the source
- When the native multicast packets arrive at RP:
  - –Encapsulated data is discarded
- PIM register-stop is sent to sender DR

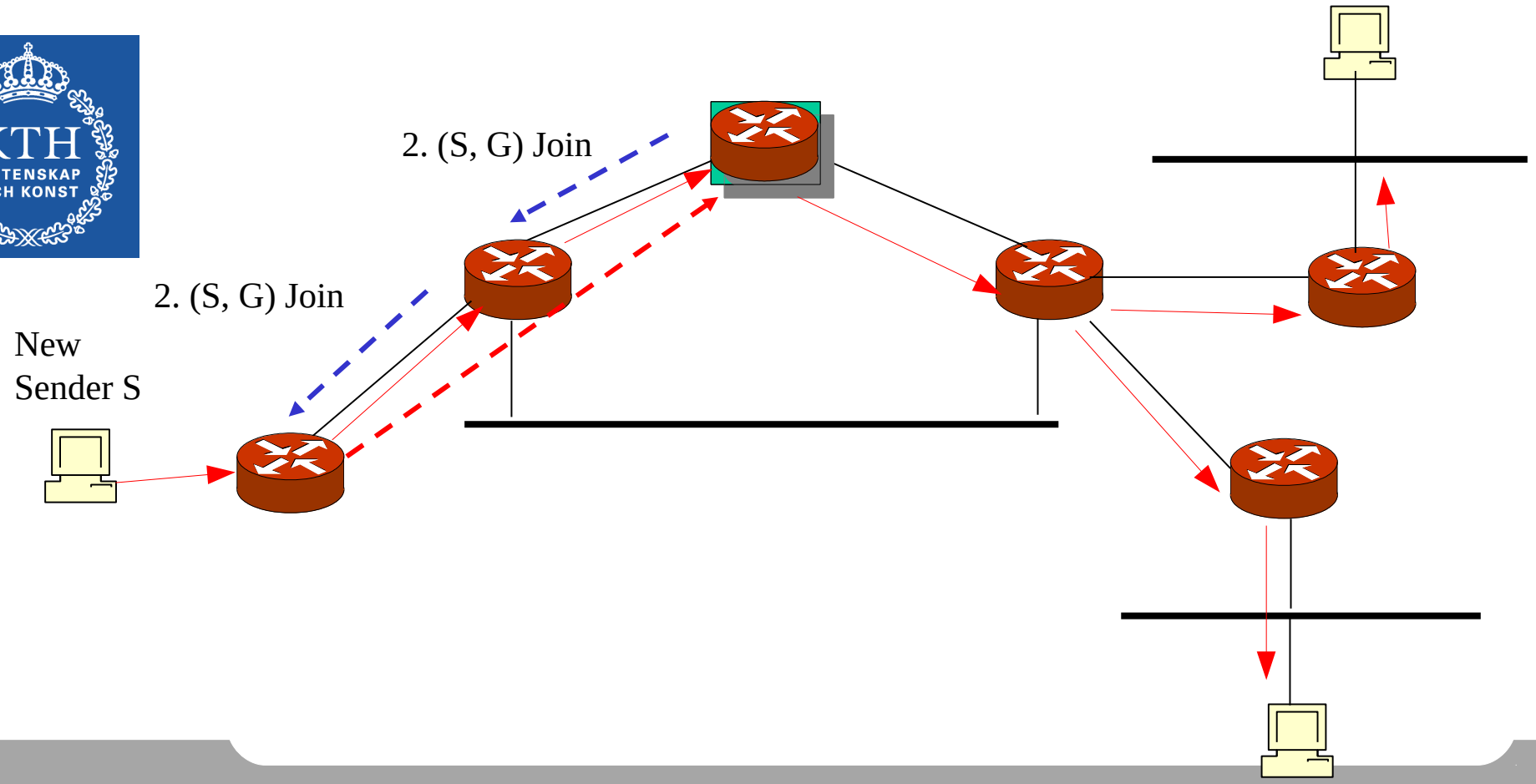
PIM-SM source registring (1)

- PIM-register sent to RP from source
- Data is encapsulated in unicast Register messages to RP, then natively in (S,*) tree



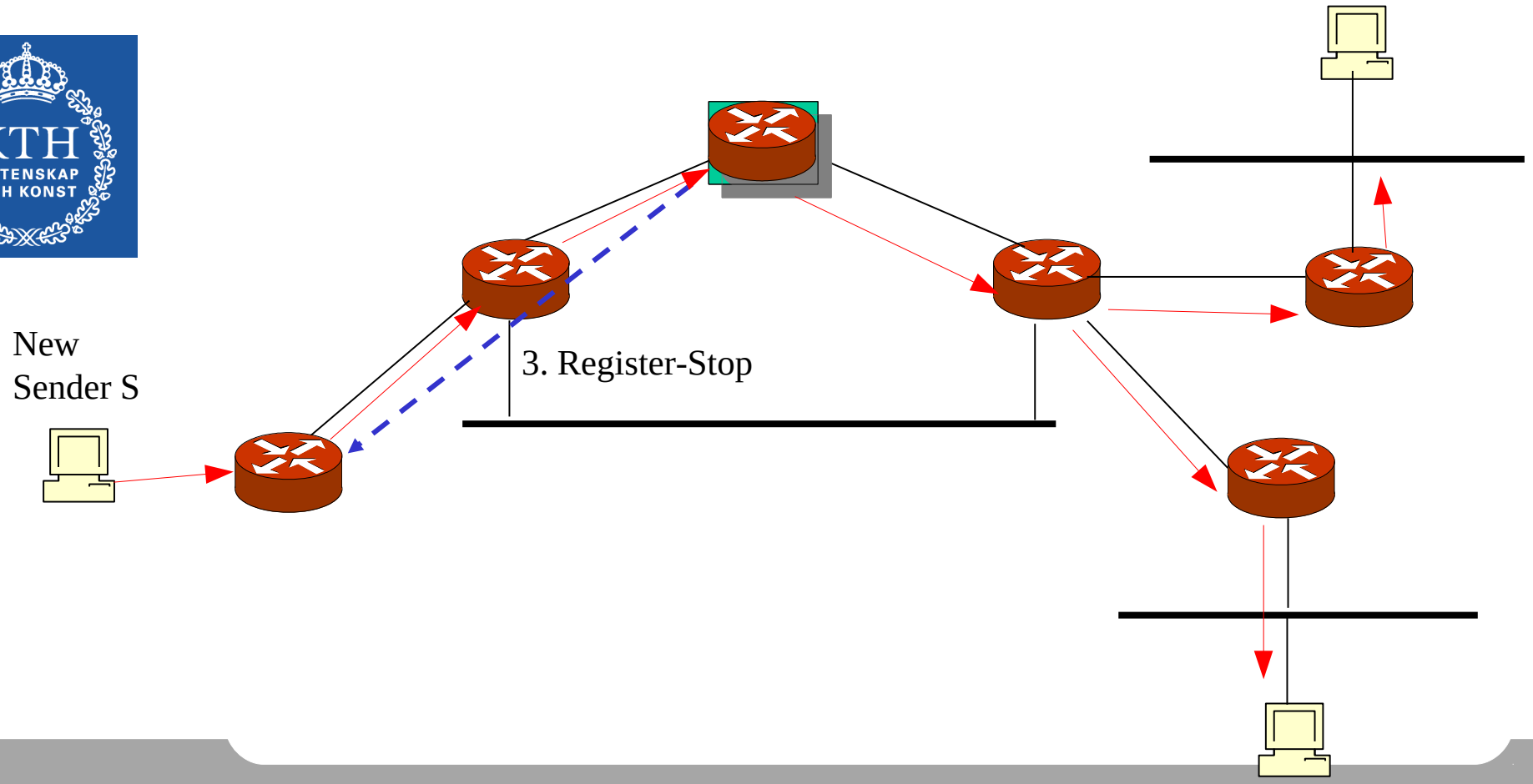
PIM-SM source registring (2)

- RP joins (S,G) tree (hop-by-hop)
- Data is sent natively from S along (S,G) tree to RP and via Register messages



PIM-SM source registring (3)

- RP sends Register-Stop to make sender stop encapsulating data.
- Data is now sent natively from S along (S,G) tree to RP.

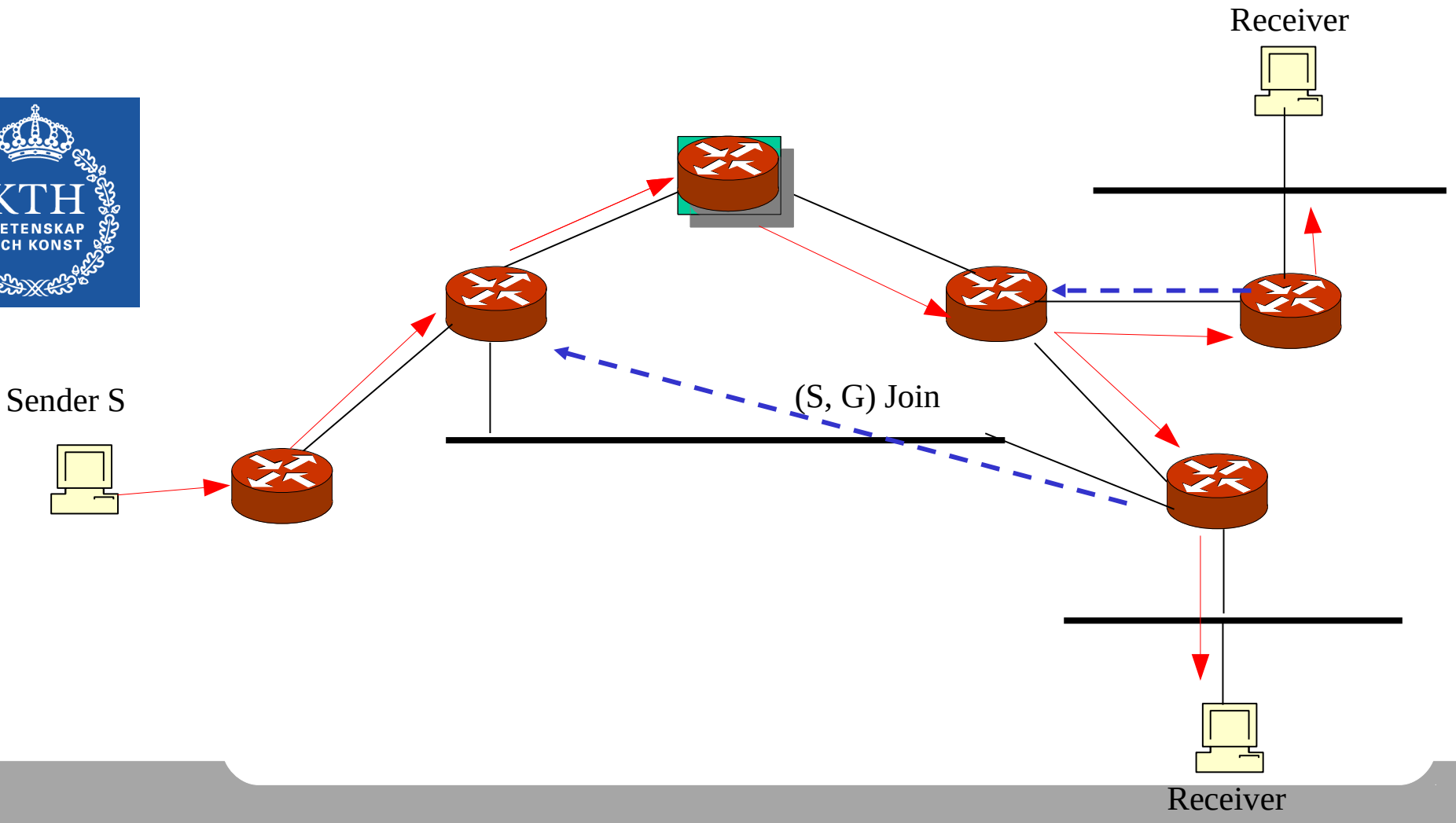


PIM-SM shortest path tree switchover

- For performance reasons, PIM-SM tries to build source-trees (SPT) after the shared tree join
  - –Most implementations: try to build source tree directly
- As soon as a receiving router (a DR) gets its first multicast data packet from a new source S, it sends an (S, G) Join upwards towards the source.
- To prune the (*,G) tree, it sends a (S,G,rpt) RP-prune at the same time (to avoid duplicates).
  - –The RP-prune is a special kind of prune
  - –The router may still be interested in other (*,G) traffic.
- Finally, the source may be pruned from the shared tree.

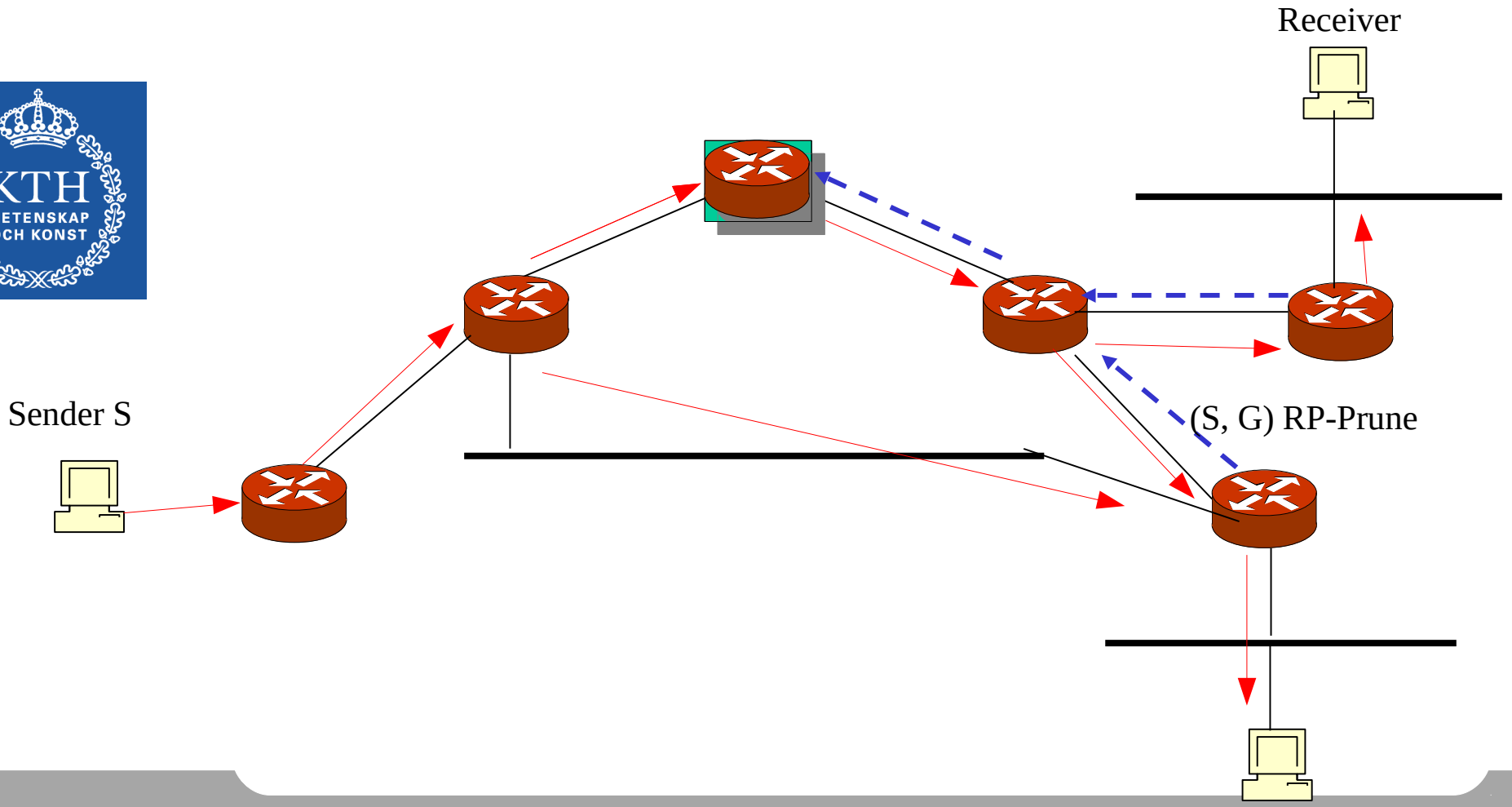
PIM-SM switchover(1)

- Receiver DRs sends (S,G) Join towards source joining the (S,G) tree.



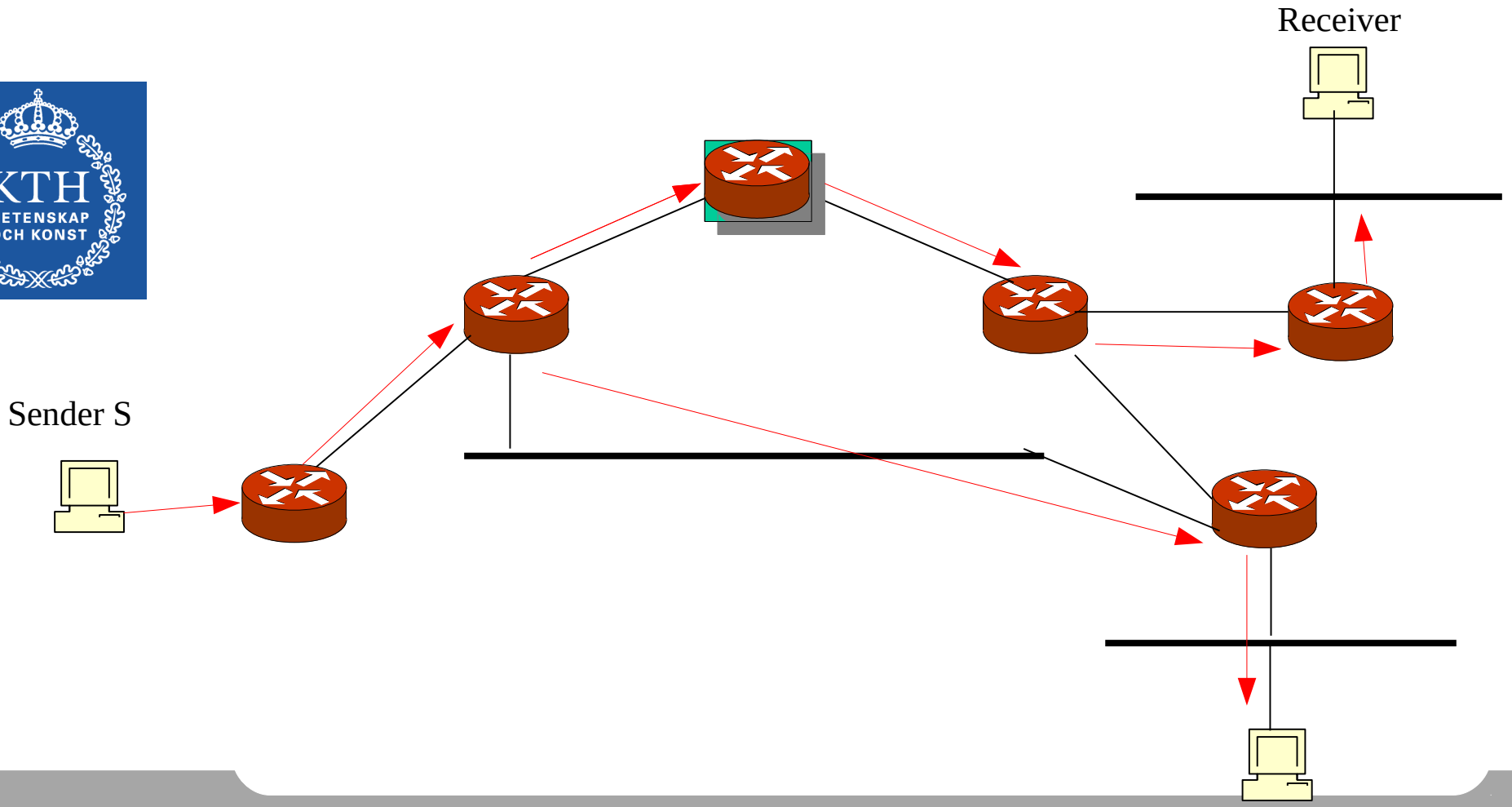
PIM-SM switchover (2)

- Traffic starts flowing down SPT also
- Receiver DR:s send (S,G,rpt) Prune towards RP to leave RPT tree and avoid duplicates



PIM-SM Switchover(3)

- Traffic flows down (S,G) tree only
- Note traffic still flows to RP (for new receivers joining RPT)



Source-specific joins

- If a router receives an IGMP v3 source-specific join
- The router can skip the RPT join and go directly to the SPT join!
- The RP and shared tree mechanism is really just a synchronization mechanism for senders of a group to make contact with receivers of that group,...

PIM Hello protocol

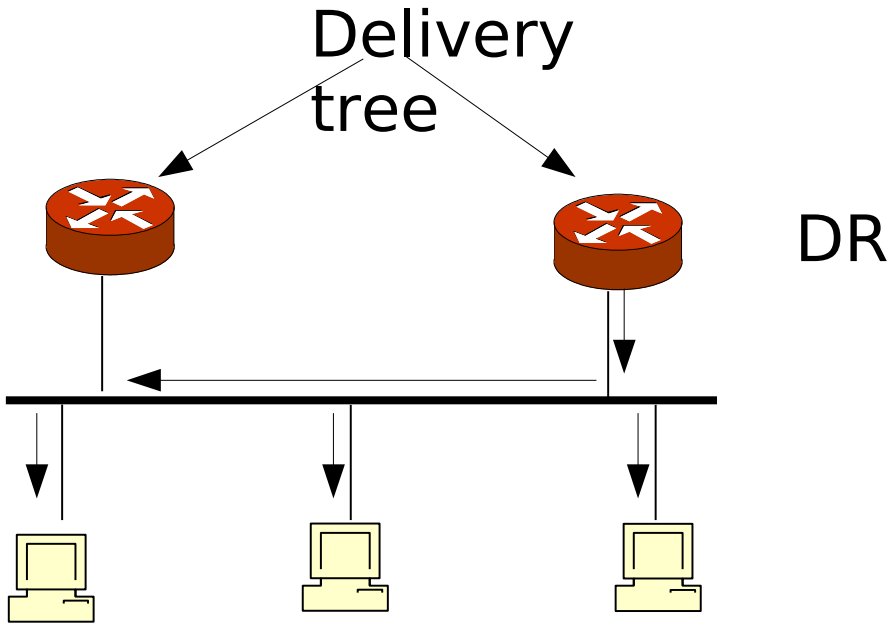
PIM has a Hello protocol

PIM Hellos are sent on 224.0.0.13 to shared media in order to

- –Form neighbor adjacencies
- –Select DR on a shared media (for handling hosts on a network)
- –Options negotiation.

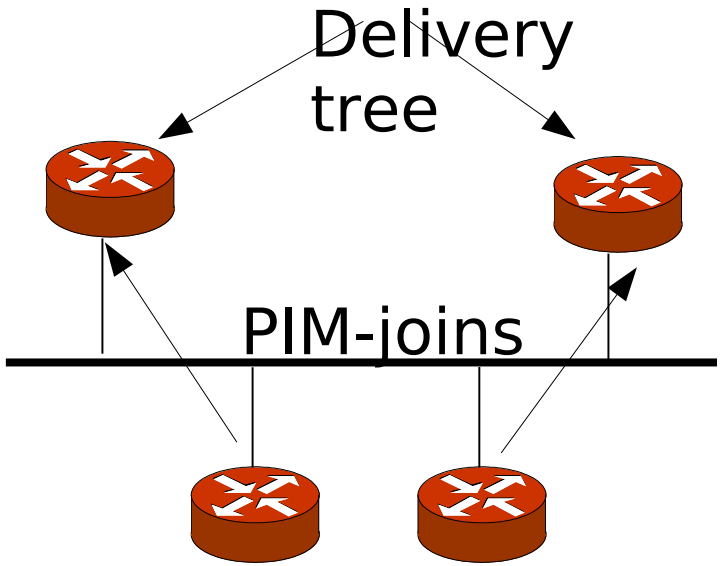
Designated Router

- Many multicast routers on one link. IGMP elects one Querier
- But multicast protocol also elects a Designated Router on each shared segment using HELLO.
- DR handles multicast routing on behalf of the link, and forwards data.
  - Typically same as IGMP Querier
- DR also detects new senders and encapsulates using register messages



PIM Assert

- On a transit link (shared media) one router must be the upstream router for a tree
- Routers may have inconsistent unicast routing tables causing different routers to believe different routers are upstream
- This would cause the tree to have several upstream routers -> duplicates.
- PIM Assert messages are sent to vote for one single forwarder per tree for the shared media.
- All routers on that shared link use the winner to send joins to.



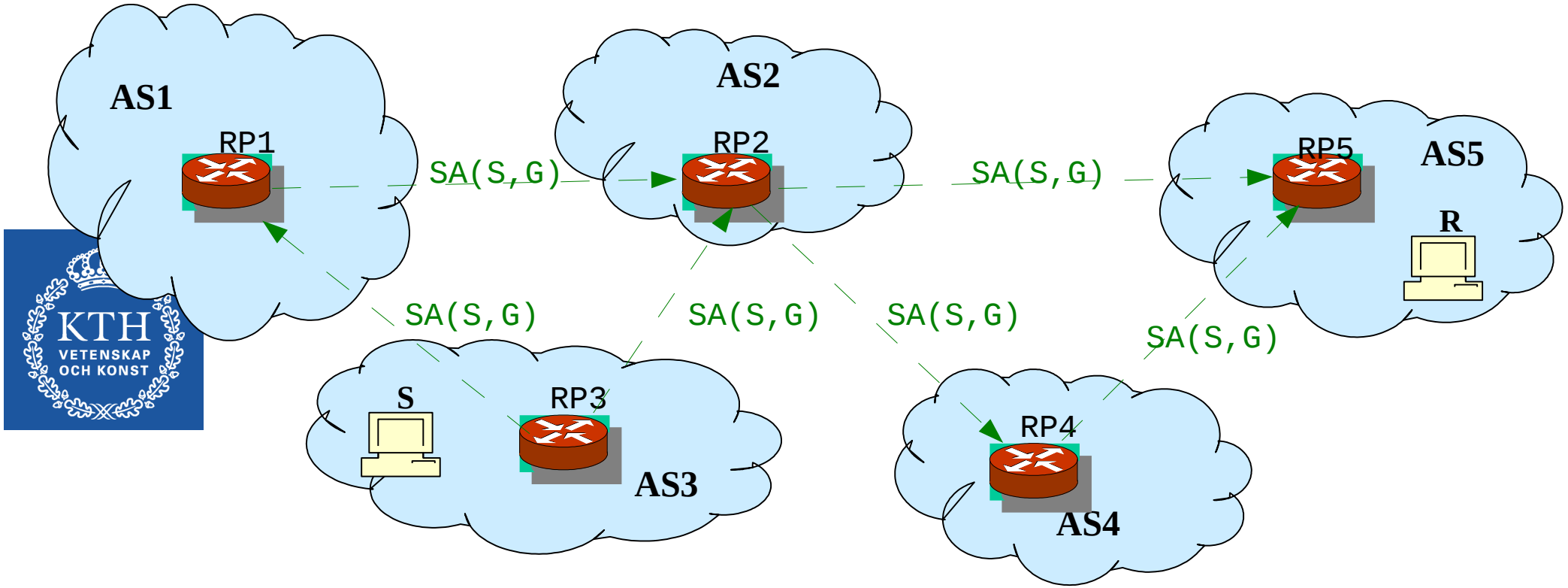
PIM Rendez-vous points

- The placement of the RP can be important, since all (at least initial) traffic need to pass the RP.
- Group-to-RP mapping
  - –Same RP for all groups
  - –Different RP's for different groups
- Manual configuration
- Anycast-RP
  - –Use anycast address to identify RP
- But then the RPs need to communicate using
  - –the Multicast Source Discovery Protocol (MSDP)

MSDP – Multicast Source Discovery Protocol

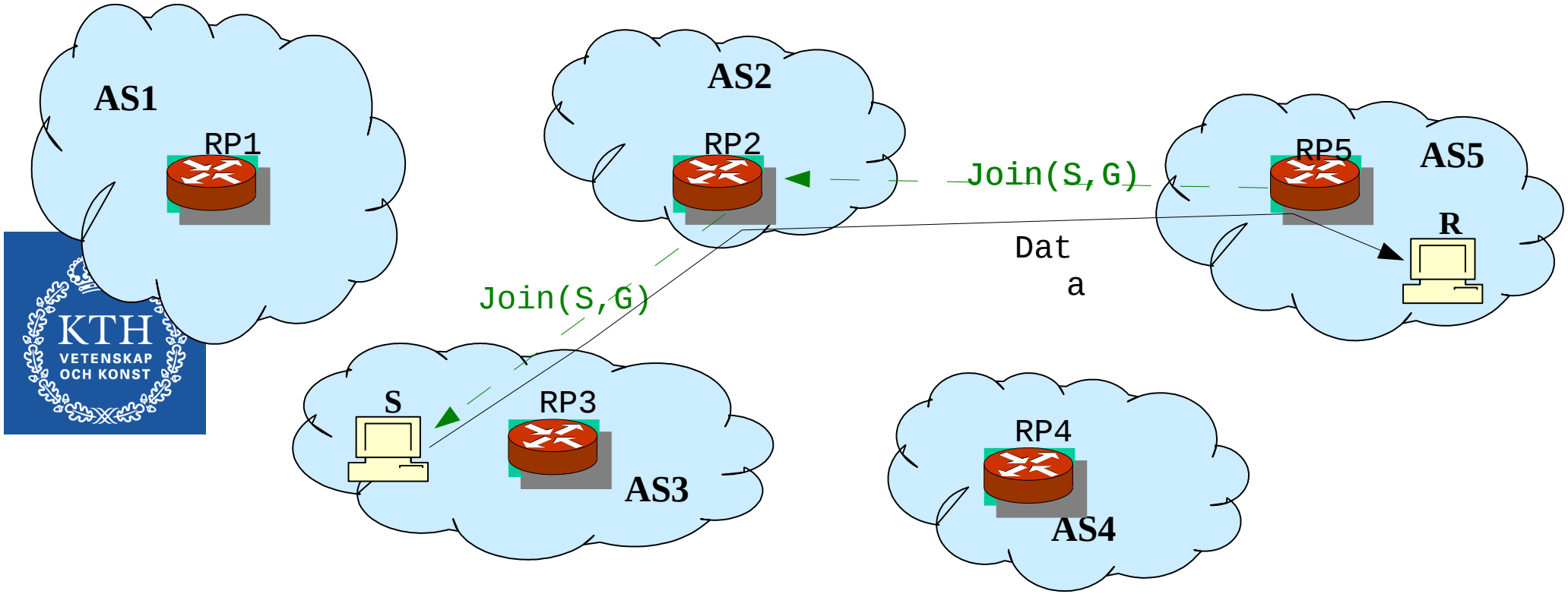
- Discovering sources is a major problem in PIM, especially in the inter-domain routing case:
  - –One AS manages an independent RP for that domain – following the autonomous systems architecture
  - –But what if there are senders in other domains?
- MSDP operation:
  - –The RPs in different domains communicate with each other on a peer-to-peer basis
  - –The RP keeps track of all senders in one domain and sends MSDP Source Active (SA) messages to the RPs in the other domains
  - –If an RP has receivers in a group G, and discovers a sender S in another domain, it performs a (S,G) Join towards S.
  - –SA messages are reliably flooded to all RP peers
    - Concern for scalability

MSDP Example(1)



- Assume a group G and a sender S in AS3 and receiver R in AS5
- S starts sending to G and registers with RP3 in AS3 and R has joined the shared tree in AS5
- RP3 floods Source Active messages for S

MSDP Example(2)



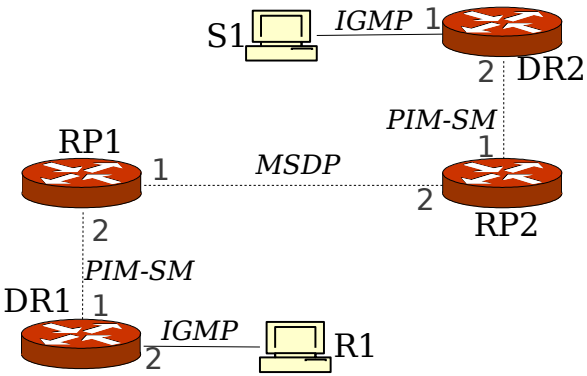
- RP5 receives the SA (S,G), and since it has receivers in its domain,
- It joins the source-specific tree (S,G)
- Data flows from S to R via RP5, and then switchover to source-specific tree

Anycast-RP and MSDP

- In anycast-RP the RP is an anycast address which is used by many routers.
  - Every RP serves a part of the network.
- But the RP:s need to synchronize their state to detect all senders/receivers.
- Anycast-RP:s synchronizes their state with MSDP
  - Also within a domain.
  - This is common practice and used in the lab

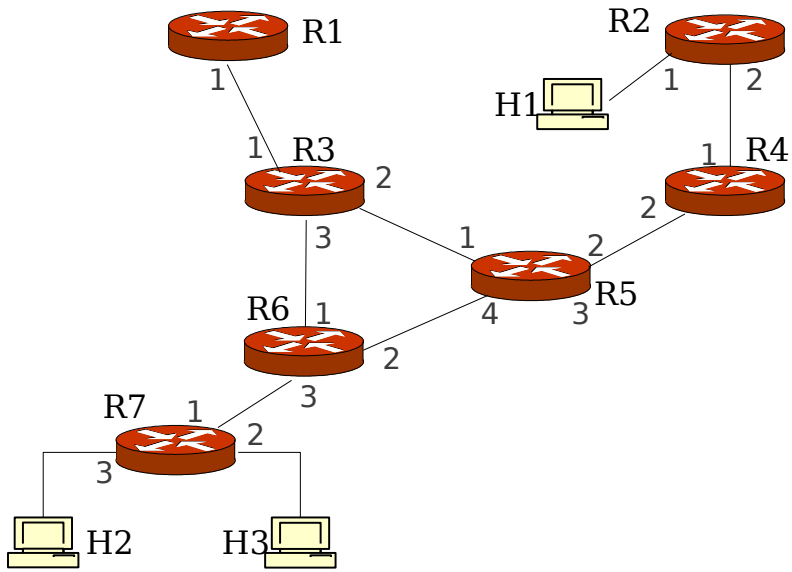
Exercise 2: Anycast-RP, PIM and MDSP

- Four routers running PIM-SM, Anycast-RP and MSDP
- RP1 and RP2 are RPs and run MSDP
- A sender S1 sends multicast traffic to group G
- A receiver R1 receives traffic.
- List the multicast forwarding entries
  - Before MSDP synchronization
  - After MSDP sync but before tree switchover
  - After tree switchover
- Solution on web after the lecture.



Router	*(Sender, Group)*	*Outgoing i/f list*

Homework 3



- List multicast forwarding entries of routers after MSDP sync but before tree switchover
- Deadline: 26/4 13:15 (Lecture 7)

Source-specific multicast

- Regular IP multicast model (Any-source multicast)
  - –A receiver R joins a group G (*, G)
  - –R receives all traffic destined to G
  - –Senders unknown a priori -> The routing protocol must detect senders
- Source-specific multicast
  - –A receiver R joins a group G for sender S only
  - –R receives traffic only from S destined to G
  - –The routing protocol need not detect sender
- SSM uses multicast addresses with prefix 232/8
- Requires changes to socket API
- IGMPv3
- But detection of senders not necessary