Advanced Networking Concepts
Group Communication / Multicast

Contents - Networking - Group Communication/Multic.

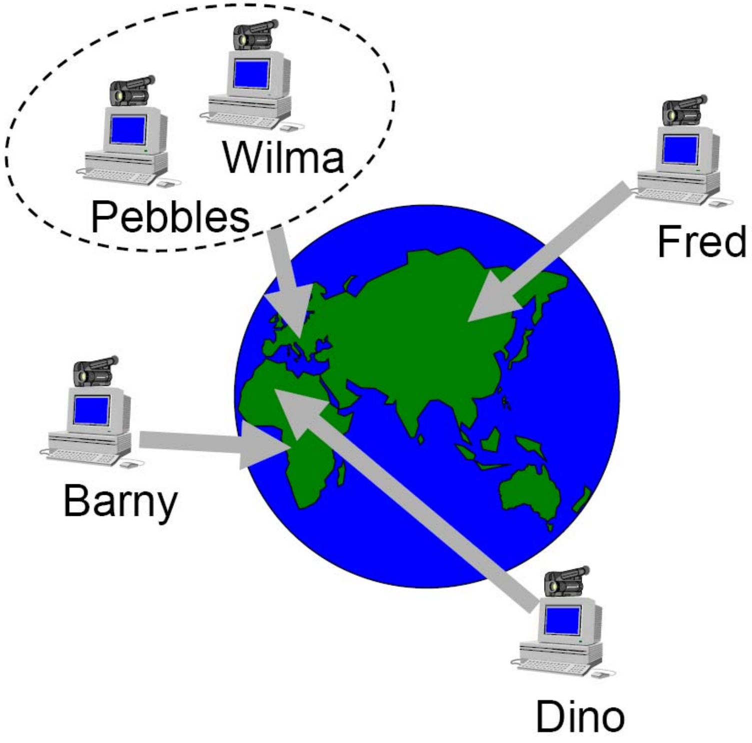
- Introduction
- Multicast Realization on different OSI Layers
- IP Multicast Details

Introduction

Definition of Group Communication

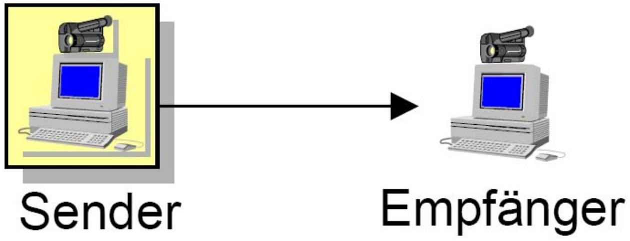
- Group Communication
  - communication in which multiple communication partners are participating in; the communication between two partners is a special case of group communication
  - participants may take different roles: sender, receiver, ...

- Example applications:
  - (Video) conferencing
  - Computer Supported Cooperative Work (CSCW)
  - software distribution services
  - information distribution services
  - parallel computing
  - distributed games

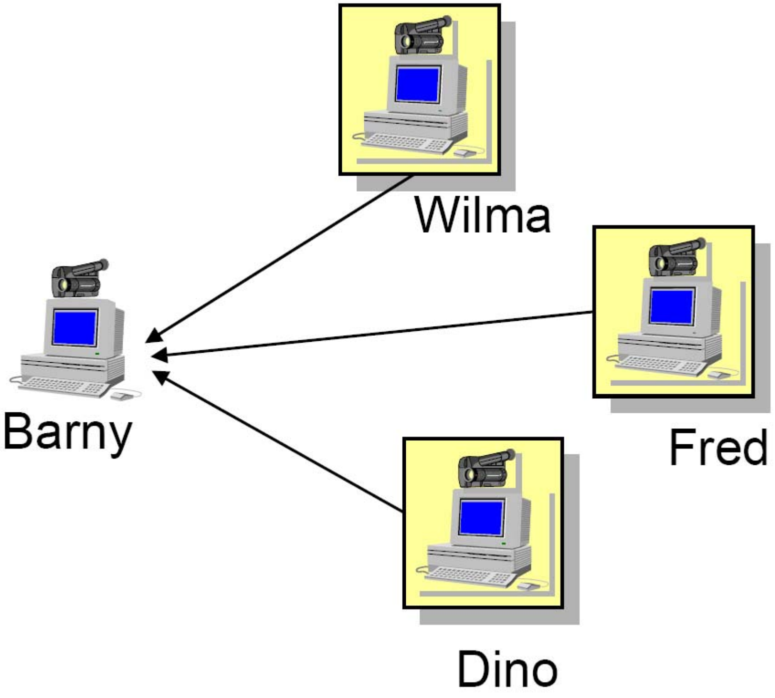


Communication Types (1)

- **Unicast (1:1 communication)**
  - 1 sender and 1 receiver
  - unidirectional user data flow
  - transmission of control data in inverse direction possible

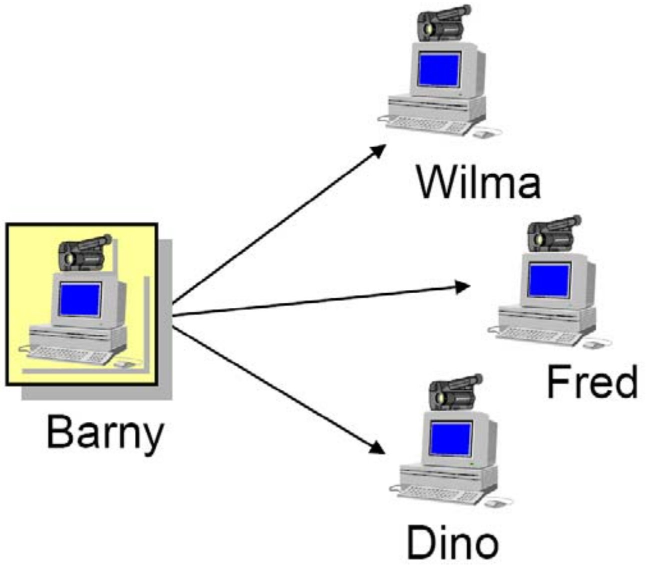


- **Concast (m:1 communication)**
  - m sender and 1 receiver
  - unidirectional user data flow
  - transmission of control data in inverse direction possible

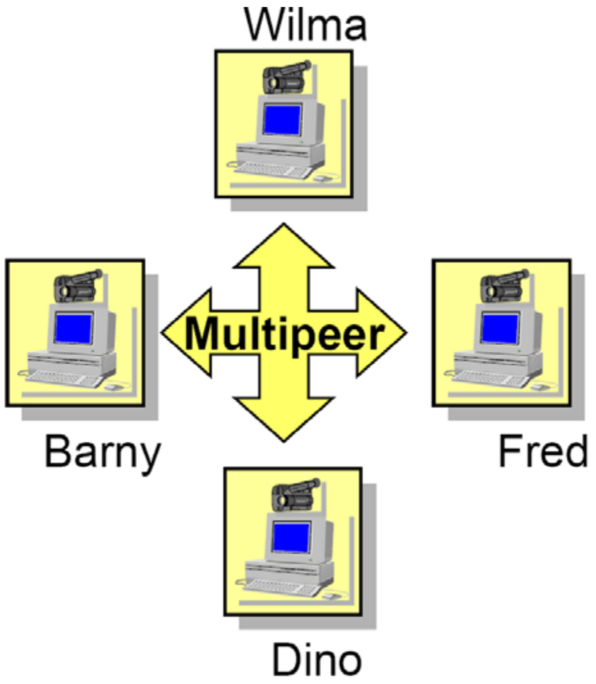


Communication Types (2)

- **Multicast (1:n communication)**
  - 1 sender and n receiver
  - unidirectional user data flow
  - transmission of control data in inverse direction possible

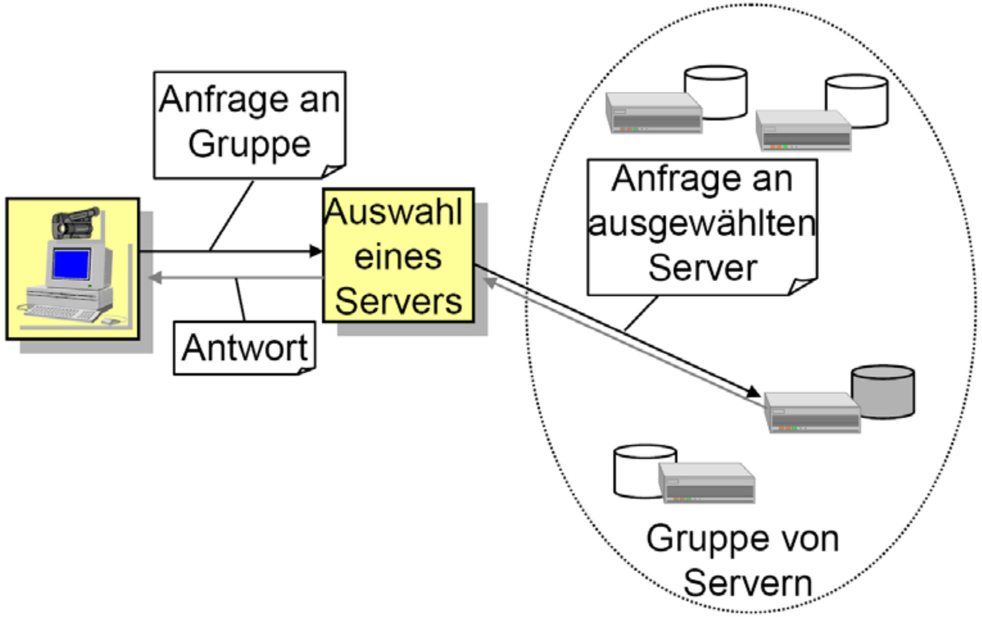


- **Multipeer (m:n communication)**
  - m sender and n receiver
  - often called multipoint communication
  - most variable form of group communication
  - often emulated via multicast



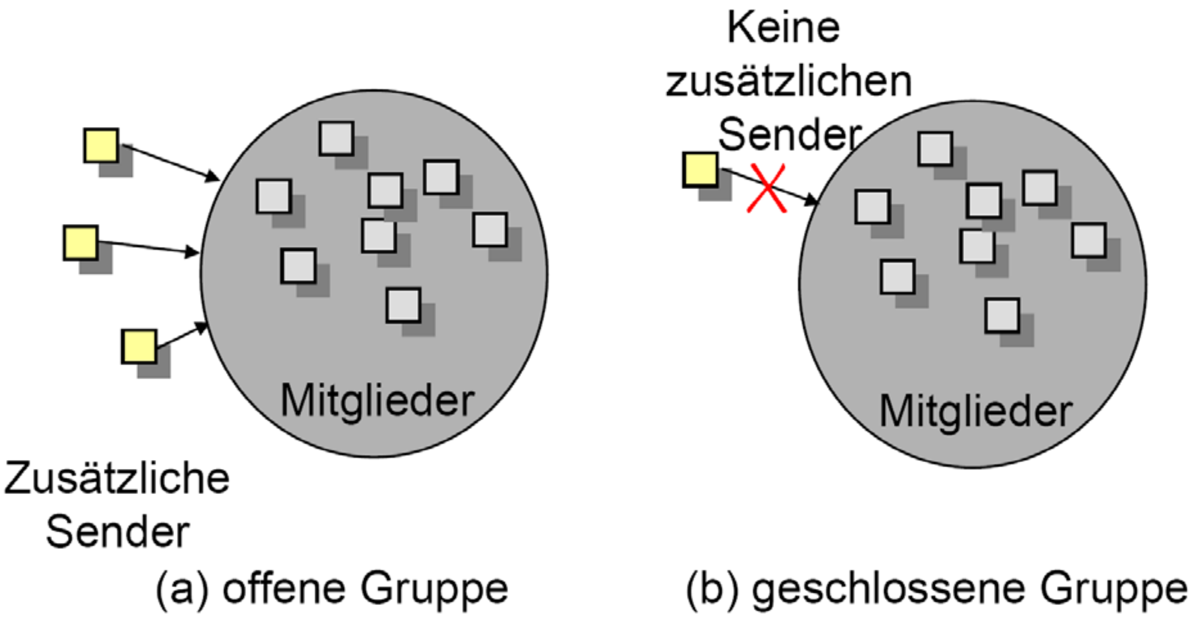
Communication Types (3)

- **Broadcast**
  - in contrast to multicast the group of receivers is not restricted in size; thus, the realization is simpler, as no groups have to be formed, addressed and maintained
  - application example: radio broadcast
- **Anycast**
  - here no group data exchange is performed, but one receiver is selected from a group of receivers; the receiver has no influence on the selection
  - application example: localization of a service in a distributed system



Properties of Groups (1)

- Openness
  - **open groups** may forward data from any sender; in **closed groups** the sender have to be members of the group



- Dynamics
  - in **static groups** the composition of the group is predefined; in **dynamic groups** the group composition may change during the communication

Properties of Groups (2)

- Life time
  - **permanent groups** exist independently of a currently active communication and thus are independent of currently active members
  - **transient groups** only exist as long as they have active members
- Security
  - security may change dynamically; different data streams may use different security concepts
- Awareness
  - in **anonymous groups** the identity of the individual members is not always known; in **known groups** the identity has to be known at all times
- Heterogeneity
  - in **heterogeneous groups** the individual members may differ in their properties; in **homogeneous groups** all members have the same resources and capabilities (e.g. data rate, image resolution)

Aspects of Group Communication (1)

- Scalability
  - the scalability of group communication of large groups is a critical problem for the technical implementation
  - aspects concerning scalability:
    - **group size**
      - large groups may contain several hundreds or thousands of participants
      - high dynamics poses a challenge to the group management
    - **robustness**
      - an exchange of control information is always necessary; this may create a high overhead for large groups and the sender may easily become the performance bottleneck
    - **awareness within the group**
      - if all members should be known, the group management has to be very powerful in case of high dynamics of the group (frequent joins and leaves)
    - **group topology**
      - geographical distribution of the group
      - heterogeneity with respect to the resources and capabilities of individual group members

Aspects of Group Communication (2)

- Due to the fact that more than two participants communicate, some services and functions have to be extended or added as new
- Examples and related issues:
  - **addressing**
    - how can the entire group be addressed effectively?
  - **routing**
    - how is the data forwarded to the group?
  - **management**
    - who is a member of a group?
    - how to efficiently manage large groups with high dynamics?
  - **reliability**
    - how to guarantee the right order of data delivery in case of more than one sender?
    - how to cope with acknowledgements from 1000 or more receivers?
  - **security**
- Consequence: group communication has to be supported in multiple layers of a communication system

Multicast Realization on different OSI Layers

Multicast on L2 - Example Ethernet Multicast (1)

- Ethernet LANs (IEEE802.3) support broadcast and multicast
- Multicast MAC addresses:

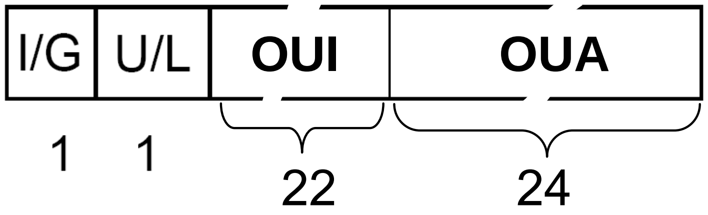
Multicast Address	Explanantion
01-00-5e-00-00-00 to 01-00-5e-7f-ff-ff	Internet (IP) multicast
01-80-C2-00-00-00	Spanning tree protocol
03-00-00-00-00-01	NetBEUI

for BPDUs

- Operation of Ethernet multicast:
  - multicast MAC addresses are used to identify stations that are associated to the multicast group on the LAN
  - Ethernet frames with multicast MAC addresses **are broadcasted within the entire LAN**
  - the station interfaces check whether the respective multicast MAC frame is destined for this station

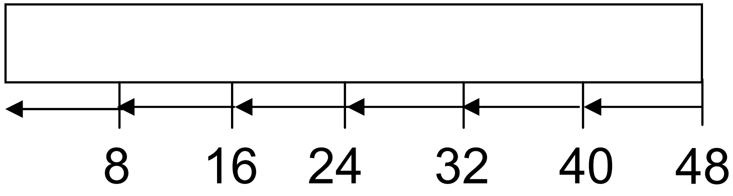
Multicast on L2 - Example Ethernet Multicast (2)

- Remarks:
  - multicast MAC addresses co-exist with the normal "hardware" MAC address of an interface
  - MAC address format:



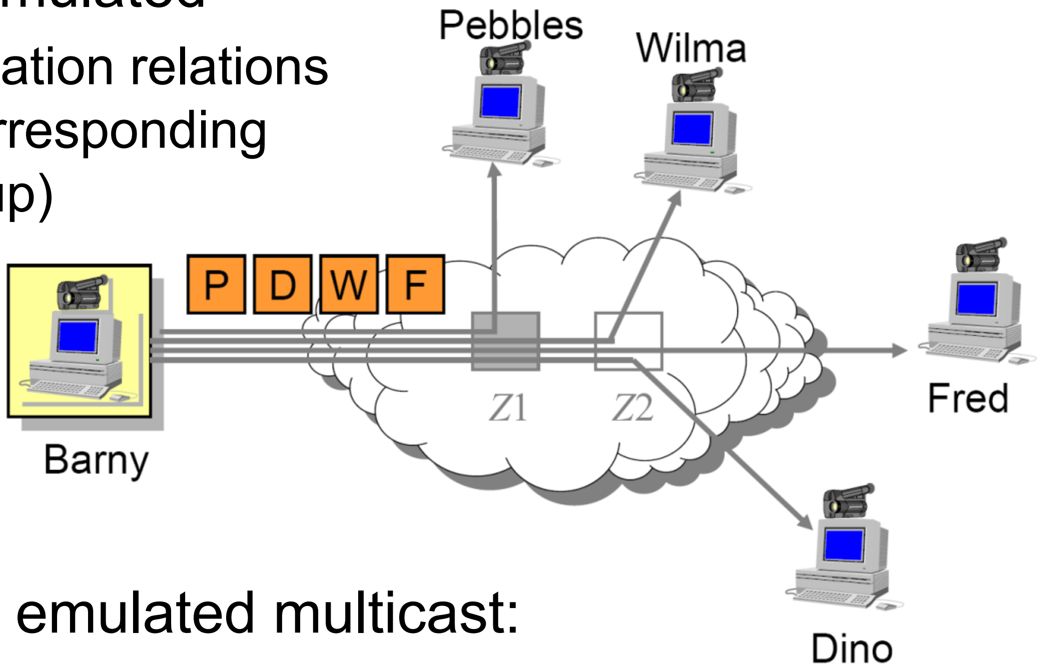
I/G: Individual (0)/Group(1)
U/L: Universal (0)/Local (1)

  - for multicast frames the I/G bit is set to 1
  - transmission order of Ethernet MAC frames on the physical medium: each octet is transmitted with the least significant bit first



Multicast on L2 - non-broadcast L2 Networks (NBMA)

- In NBMA networks multicast is emulated
  - realized by n unicast communication relations between the sender and the corresponding receivers (belonging to the group)



- Disadvantages / problems of the emulated multicast:
  - many individual communications relations
    - multiple transmission and storage of data
    - delays in sending (due to sequential sending)
    - high resource consumption
- Advantage of the emulated multicast:
  - intermediate systems require no special multicast capabilities and do not have to copy data frames

Multicast on L3 - Challenges

- Multicast addressing
  - multicast addressing (in addition to unicast addressing) is required for identifying the multicast group membership
- Management of multicast groups:
  - handling of joining and leaving group members
- Multicast routing:
  - resource consumption should be minimized; thus, the multicast packets should be routed on joint links as long as possible
  - routes vary depending on the current group members - dynamic joining and leaving of group members
  - the individual group members may have different requirements on the quality of service (QoS)
- Example: IP Multicast → see next section

Multicast on L4 - Challenges for Transport Protocols

- Multicast-capable L4 connection management
  - connection setup to multiple receivers
  - support of group addresses or name lists
  - conflict resolution during the QoS negotiation
- Extension of the L4 service interface
  - additional service elements required to add and remove participants of a multicast communication
- Multicast capable error handling
  - extension of the conventional ARQ concept required
  - adaptation of the acknowledgment mechanism to a group of receivers
  - efficient procedures for error correction
- Multicast-capable flow and congestion control

Multicast in the Application Layer

- Problem
  - multicast in the network layer (IP Multicast) is not yet applied today
- New approach
  - realization of multicast in the application layer
    - no support the legacy network infrastructure required
    - no multicast addresses and multicast address allocation necessary
    - no new mechanisms for flow and congestion control required
    - quick and easy introduction possible
  - example: Peer-to-Peer overlay networks with multicast capability
- Open questions:
  - scalability
  - QoS, in particular end-to-end delay

Group Communication / Multicast Concepts

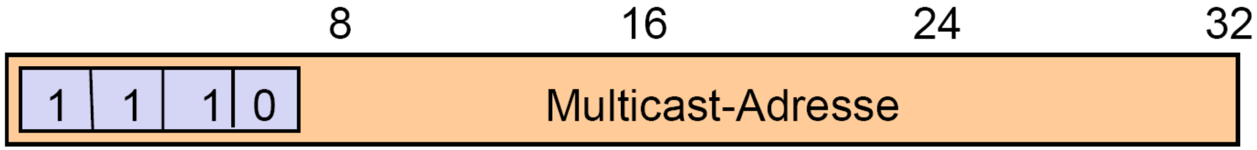
IP Multicast

Issues for IP Multicast

- IP multicast addressing
- Management of multicast groups
- Multicast routing:
  - within a (sub)net: mapping of multicast IP addresses to multicast MAC addresses
  - between (sub)nets in an AS: multicast intra-domain routing
  - between AS: multicast inter-domain routing

Multicast IP Addresses

- Range of multicast IP addresses: 224.0.0.0 - 239.255.255.255 (Class D)



- Multicast (MC) IP addresses have no further internal structure
- Several multicast IP addresses may be assigned to an interface in addition to its normal IP address; a multicast IP address represents the membership to the respective multicast group
- Multicast IP address blocks:
  - **local multicasting:** 224.0.0.0 - 224.0.0.255
    these MC addresses are not routed, i.e. they are limited to one IP subnet
  - **source-specific multicast (SSM):** 232.0.0.0 - 232.255.255.255
    using these addresses a MC source (sender) can send different data to multiple MC groups (with the same SSM MC address)
  - **administratively scoped MC (RFC 2365):** 239.0.0.0 - 239.255.255.255

Multicast IP Addresses - Dedicated MC IP Addresses

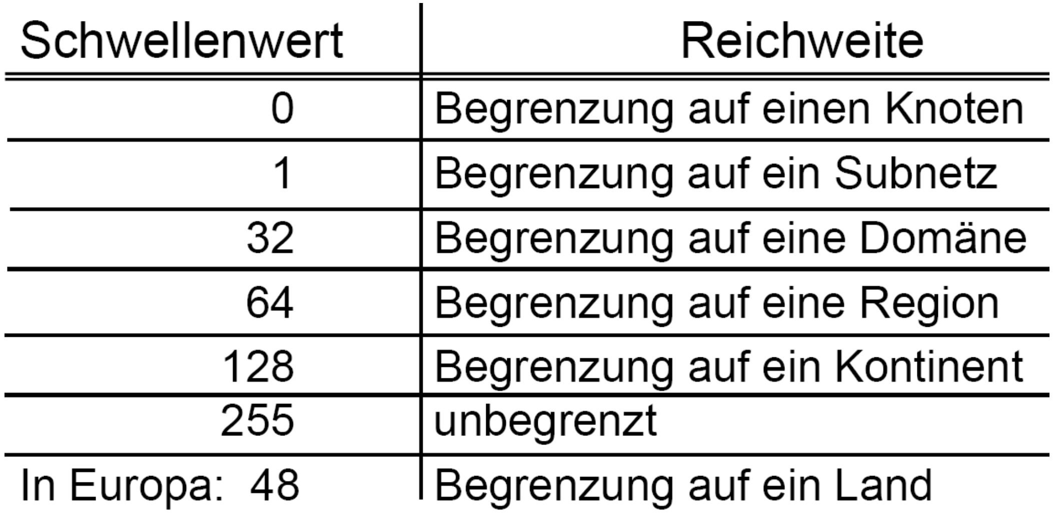
- Dedicated MC IP addresses for specific (permanent) groups:
  - all systems group (all systems on a subnet): 224.0.0.1
  - all routers group (all routers on a subnet): 224.0.0.2
  - all protocol-specific routers on a subnet (selection):
    - all DVMRP routers: 224.0.0.4
    - all OSPF routers: 224.0.0.5
    - all OSPF Designated Routers: 224.0.0.6
    - all RIPv2 routers: 224.0.0.9
    - all PIM routers: 224.0.0.13
  - application-specific MC IP addresses (selection):
    - Network Time Protocol (NTP): 224.0.1.1
    - Rhwo Daemon (RhwoD): 224.0.1.3
    - Multicast Transport Protocol (MTP): 224.0.1.9
  - for a full list see http://www.iana.org/assignments/multicast-addresses
- All other MC IP addresses are used temporarily

Multicast IP Addresses - Scope of MC IP Addresses

- Scope
  - defines the region in which the multicast data unit is forwarded
- Advantages of a limited scope
  - limitation of flooded network regions (example: multicast routing protocol DVMRP)
  - multiple use of multicast addresses in different regions of the network possible
  - increased security
- Methods for limiting the scope:
  - TTL scoping
  - administrative scoping

Multicast IP Addresses - Scope of MC IP Addresses

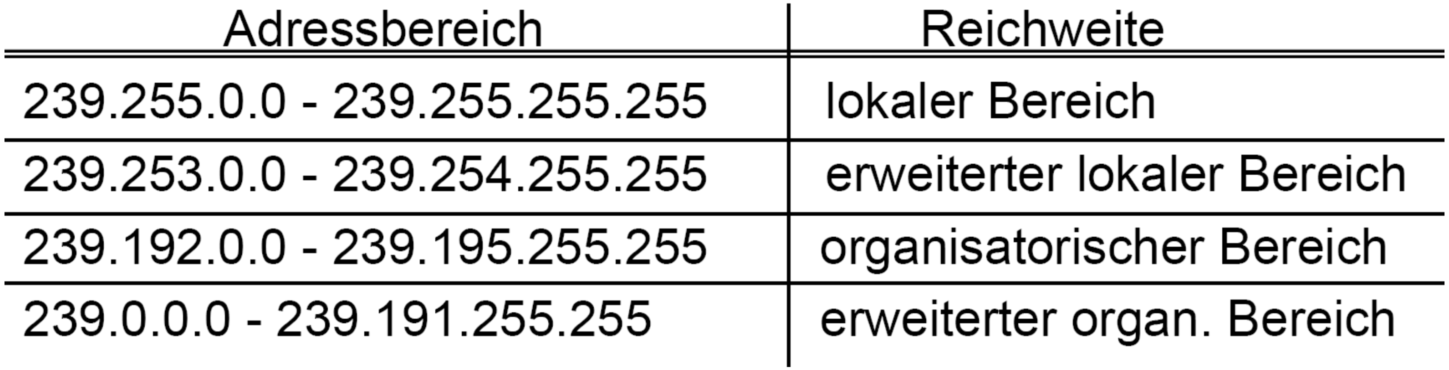
- Option 1: TTL scoping (scope limitation by TTL thresholds)
  - TTL thresholds are used to limit the scope; if the TTL value is less than the threshold, the data unit is discarded



Schwellenwert	Reichweite
0	Begrenzung auf einen Knoten
1	Begrenzung auf ein Subnetz
32	Begrenzung auf eine Domäne
64	Begrenzung auf eine Region
128	Begrenzung auf ein Kontinent
255	unbegrenzt
In Europa: 48	Begrenzung auf ein Land

Multicast IP Addresses - Scope of MC IP Addresses

- Option 2: administrative scoping (scope limitation by defining specific address blocks that are related to administrative regions, see RFC 2365)
  – multicast address blocks are related to a particular administrative region:



Adressbereich	Reichweite
239.255.0.0 - 239.255.255.255	lokaler Bereich
239.253.0.0 - 239.254.255.255	erweiterter lokaler Bereich
239.192.0.0 - 239.195.255.255	organisatorischer Bereich
239.0.0.0 - 239.191.255.255	erweiterter organ. Bereich

  – local scope: e.g. within the network of a company
  – organizational scope: specific MC addresses are assigned to an organization (example: german broadband research network D-WIN of DFN)
  – advantage: fine adjustment of the scope
  – remark: administrative regions have to be known and must not overlap

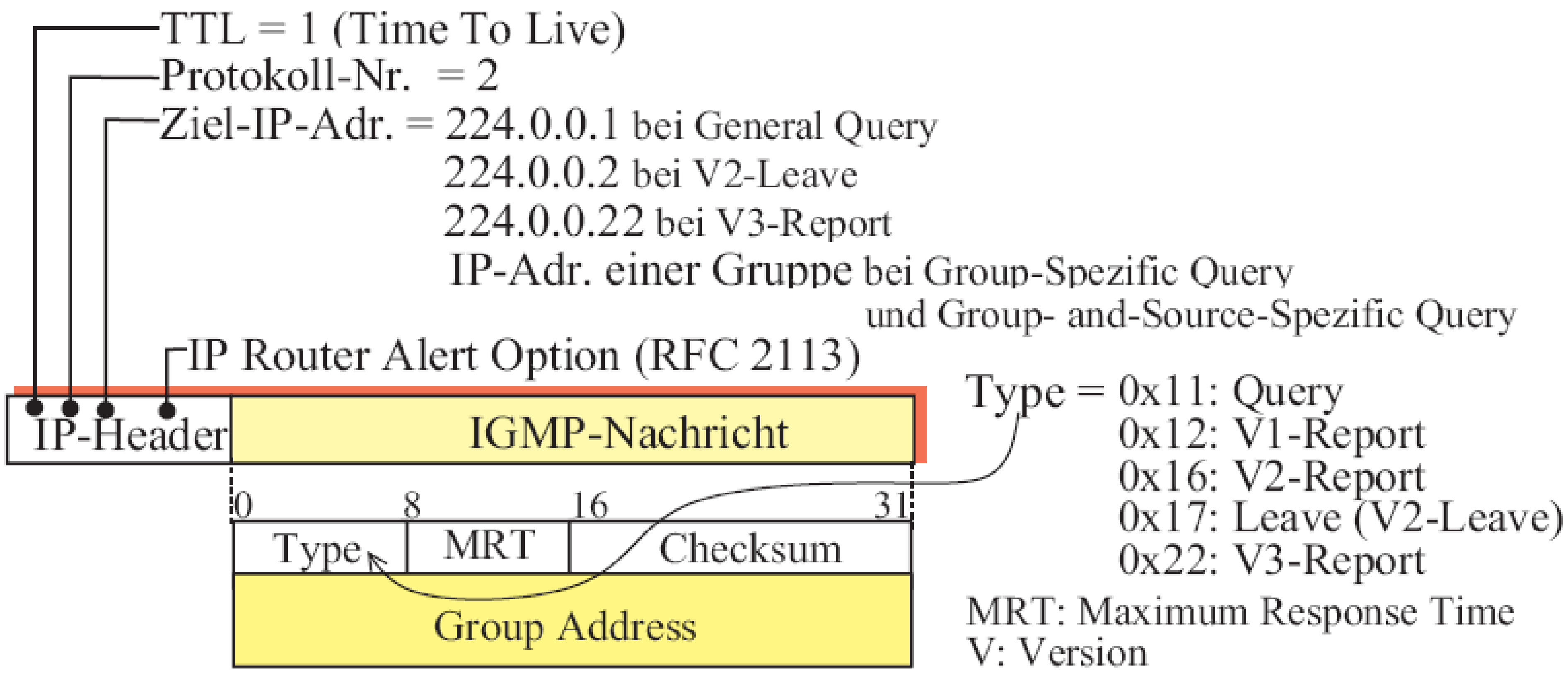
Multicast IP Addresses - MC IP Address Allocation

- Yet, no allocation procedure for MC addresses exists in the Internet
  - for source-specific multicast: not required
  - for traditional any-source multicast: address conflicts possible - the probability of a address conflict increases with increasing use
- Proposal for the allocation of MC addresses (IETF MALLOC WG) Multicast Address Allocation Architecture
  - based on administrative regions
  - objectives:
    - low probability of address conflicts
    - aggregation of addresses
  - reservation types:
    - static: fixed allocation (e.g. via Session Announcement Protocol, SAP)
    - region-related: for protocols that require an address in all administrative regions
    - dynamic: on demand allocation; limited scope; the same address can be used in different administrative regions

Management of MC Groups

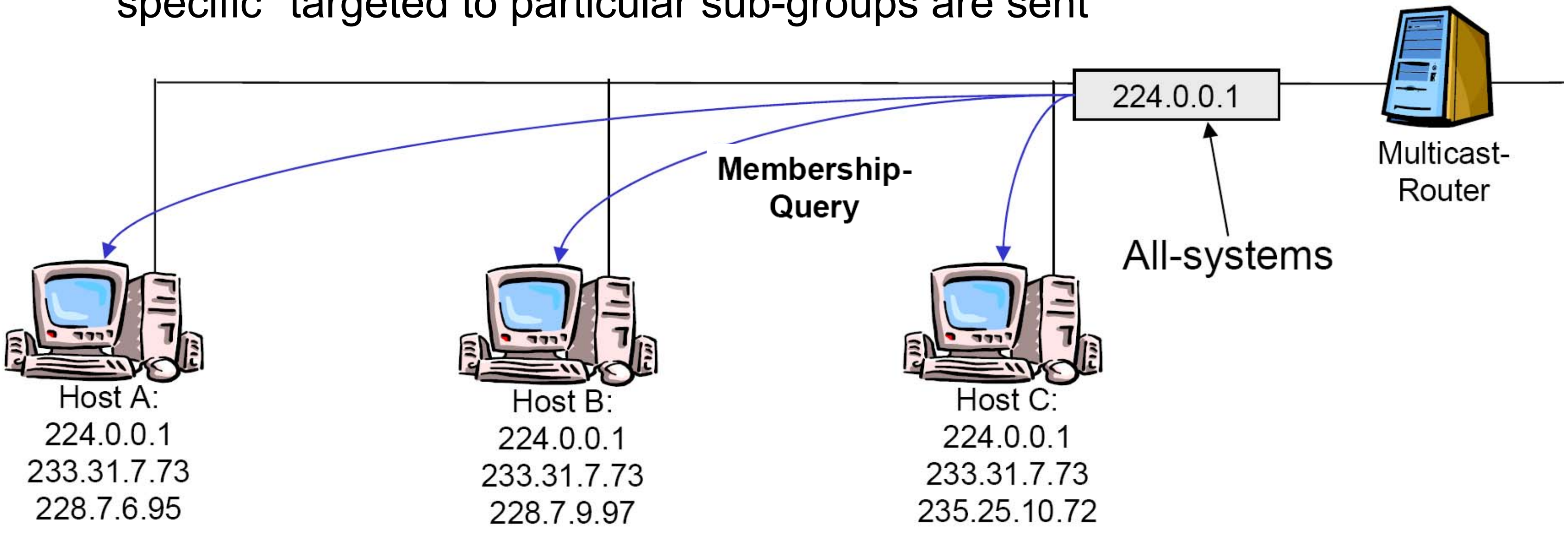
- Problem:
  - how does a router know, that it has to forward multicast data units to the connected subnets or to the end systems in the subnets?
- Solution:
  - multicast receivers inform "their" multicast routers about their group membership(s)
  - for this the following protocols can be used:
    - IPv4: IGMP (Internet Group Management Protocol, RFC 1112, 2236, 3376, 4604)
    - IPv6: Multicast Listener Discovery (MLD) for IPv6 (RFC 3810)
- General Procedure:
  - when the group membership of a end system changes, a **membership report** message with the status change is sent to the MC router
  - multicast routers also send periodic **membership query** data units to the multicast address "all-systems"; each multicast receiver in the subnet responds (after a random waiting time) one or more membership report data units which contain the addresses of the multicast groups (in which the receiver wants to participate)

Management of MC Groups - IGMP Message Format



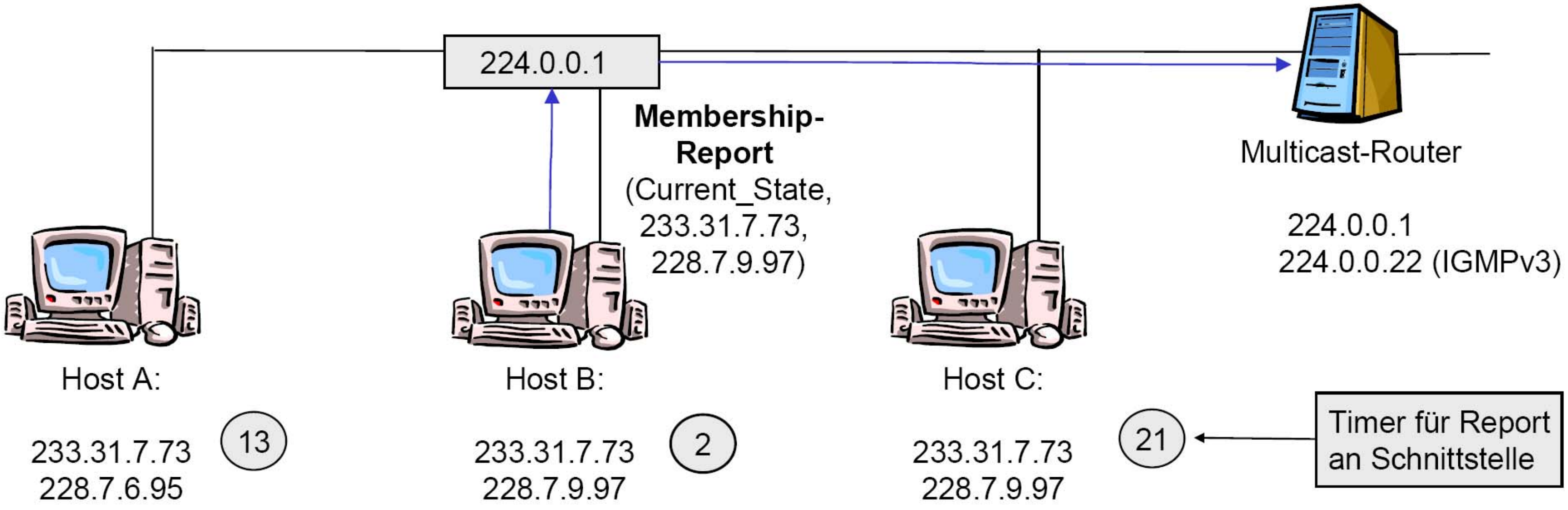
Management of MC Groups - IGMP Operation (1)

- Actions taken in the multicast router:
  - received membership reports indicating state changes are immediately processed
  - periodic sending (usually every 125s) of membership queries (of type "general query") to the "all-systems" group (224.0.0.1) in order to keep the (group membership) state information up to date
    - each multicast-capable system (host/router) is a member of "all systems" group
    - the request is sent with a TTL of 1 (scope: subnet)
    - optionally additional queries of type "group-specific" and “group-and-source-specific" targeted to particular sub-groups are sent



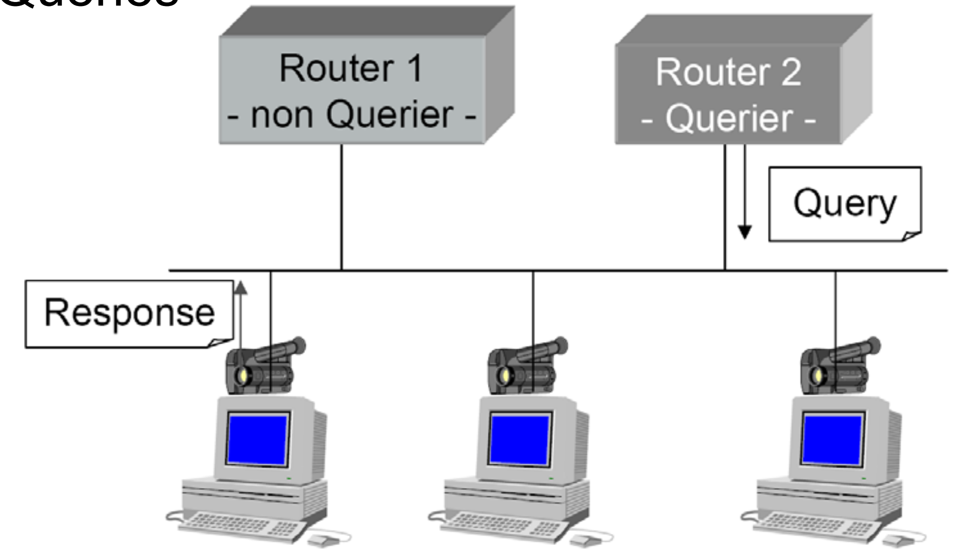
Management of MC Groups - IGMP Operation (2)

- Actions taken in the multicast host:
  - joining or leaving a multicast group leads to an immediate sending of a Membership Report (usually repeated once)
  - at receiving a membership request from the MC router a random timer is started; when the timer expires, the host responds with a Membership Report (with TTL = 1)



Management of MC Groups - IGMP Router Selection

- Problem:
  - a subnet may contain multiple multicast routers
- Solution:
  - in IGMP the multicast router can take two different roles:
    - **Querier:** in each subnet exactly one Querier (lowest IP address) exists; it is responsible for performing periodic queries within the subnet
    - **Non-Querier:** does not send periodic queries but sets the Other-Querier-Present timer; if this timer expires, the Non-Querier sends General Queries

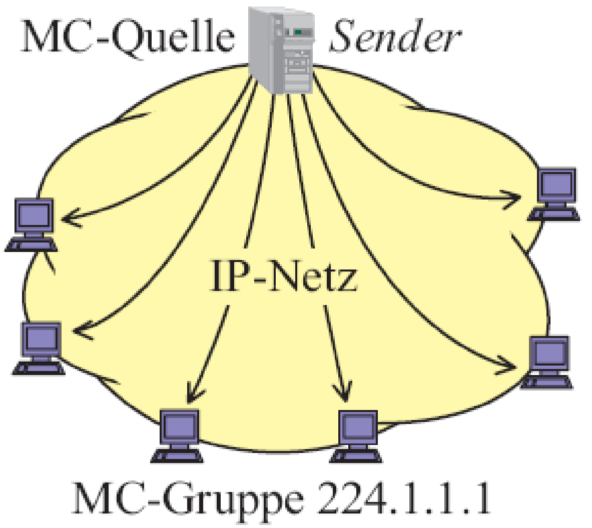


Management of MC Groups - IGMPv3 (RFC 3376,4604)

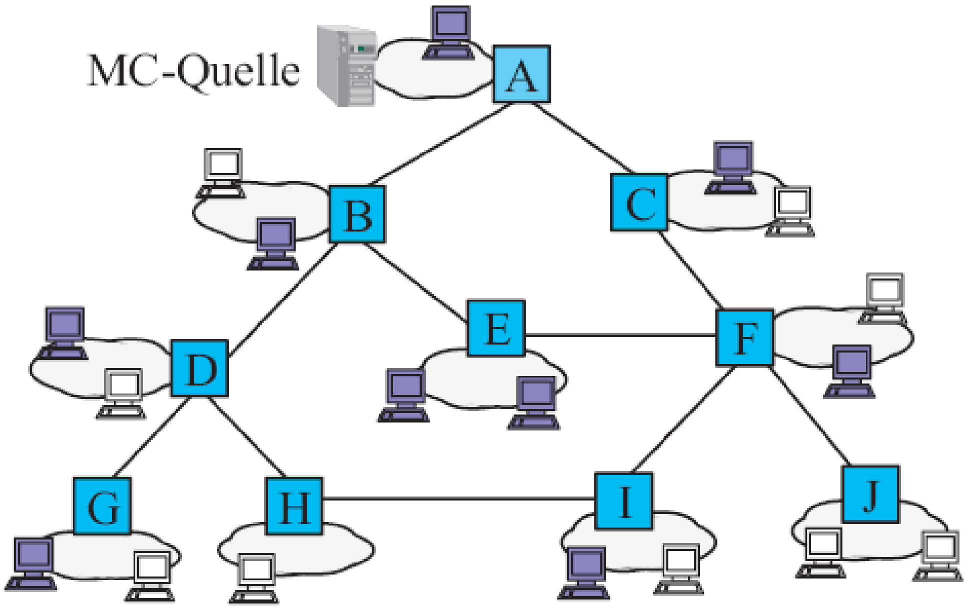
- Features of IGMPv3:
  - backward compatibility to IGMP versions 1 and 2
  - support of Source Specific Multicast (SSM): extended with "source filtering“, i.e. the ability to announce interest only for multicast packets with a specific source address (i.e. a specific station)
- Changes compared to IGMPv2:
  - states are kept for MC groups and also for source lists (SSM)
  - a programming interface now allows the specification of source lists
  - hosts repeat status change messages
  - hosts do not suppress membership reports any more (this allows for simplified implementation and explicit tracking of group membership)
  - reports can contain multiple group entries
  - reports are sent to 224.0.0.22 (this allows for simpler "snooping" by Layer 2 switches)
  - the Querier adds a robustness variable and a query interval to the queries (this allows the synchronization of Non-Queriers)

Multicast Routing - Problem Statement

Logical View

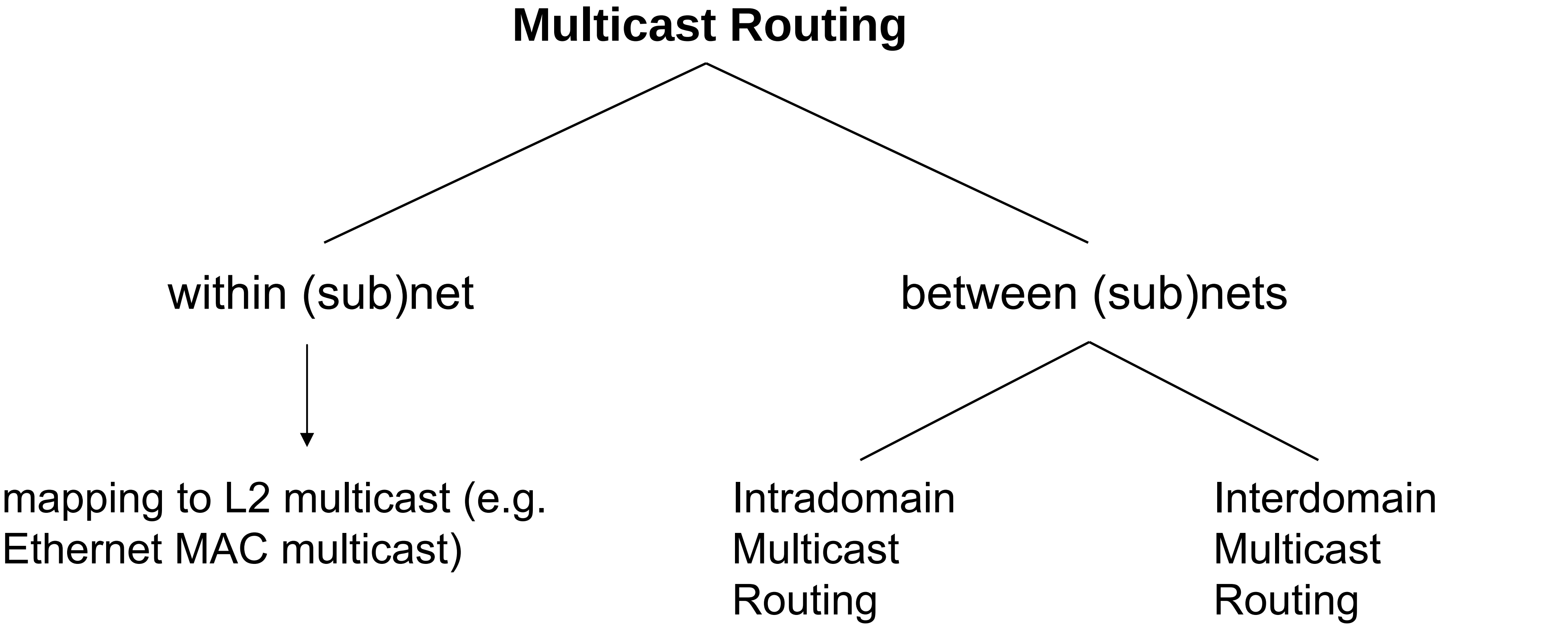


Physical View



🟦 MC-Router 🖥 Rechner der MC-Gruppe

Multicast Routing - Overview



Multicast Routing

within (sub)net

between (sub)nets

mapping to L2 multicast (e.g. Ethernet MAC multicast)

Intradomain Multicast Routing

Interdomain Multicast Routing

Examples:

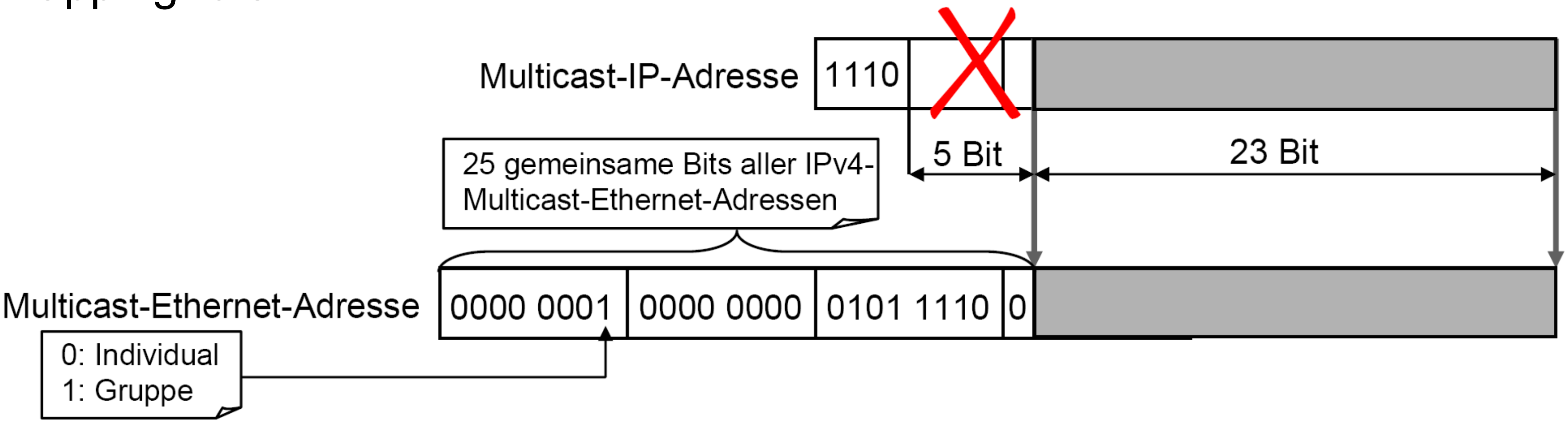
- DVMRP (RFC 1075)
- MOSPF (RFC 1584)
- PIM-SM (RFC 4601)
- PIM-DM (RFC 3973)
- CBT (RFC 2201, 2189)

Examples:

- BGMP (RFC 3913)
- MSDP (RFC 3618)

MC Routing within a (Sub)net

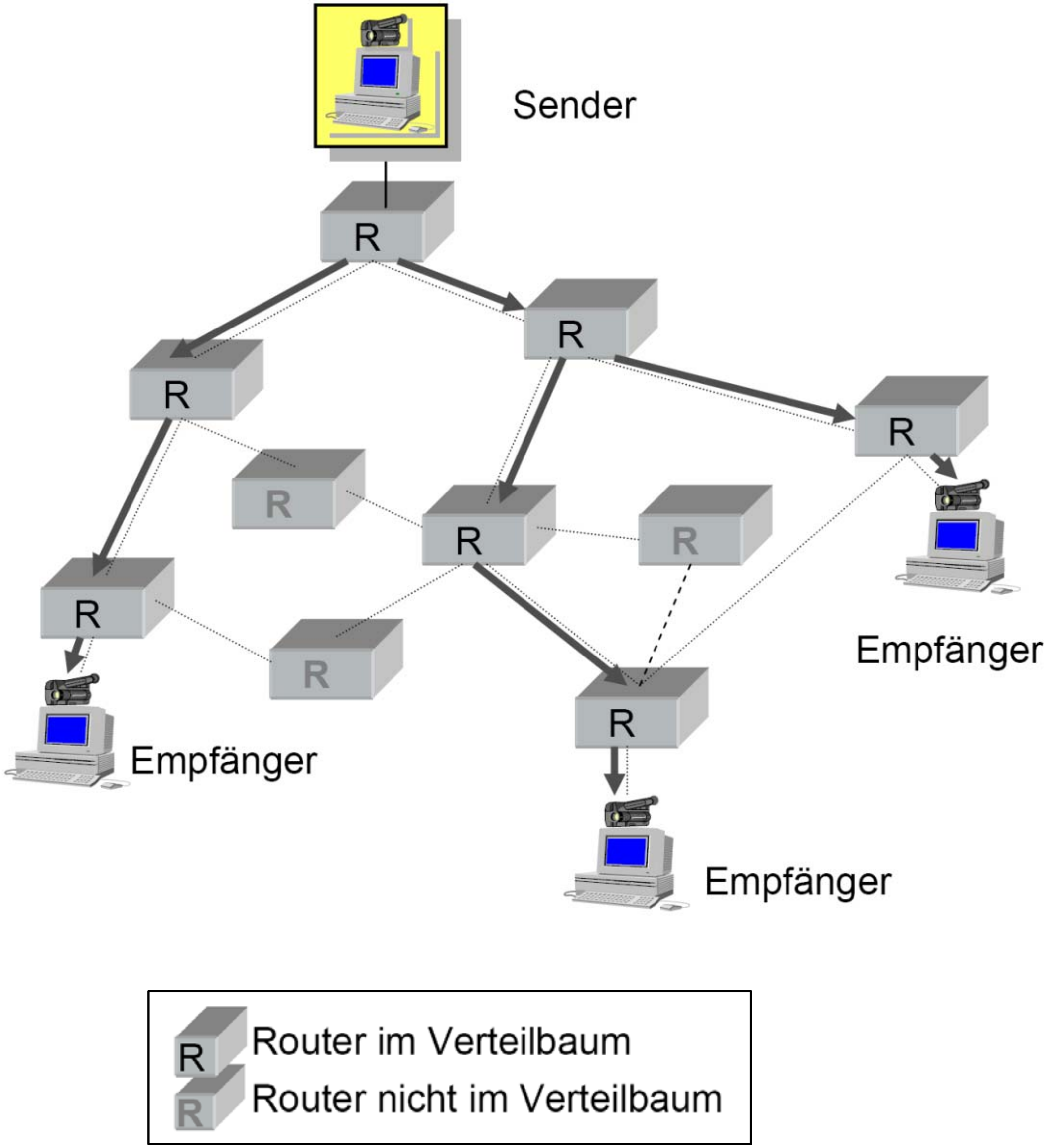
- IP multicast within a (sub)net → mapping to L2 multicast
- Example: mapping of MC IP addresses to MC MAC adresses
  - problem: the address space of multicast Ethernet is smaller (23 bits) than that of multicast IP (28 bit)
  - solution: use only 23 bits of the multicast IP address for the mapping
  - mapping rule:



Remark: therefore the first 25 bits in all Ethernet 802.3 IPv4 multicast addresses are equal: 0000 0001 0000 0000 0101 1110 0 (in hex form: 01-00-5e-x, where x is between 0 and 7)

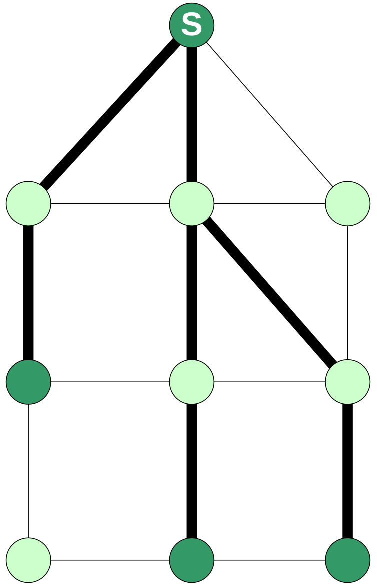
MC Routing between (Sub)nets - Basic Tasks

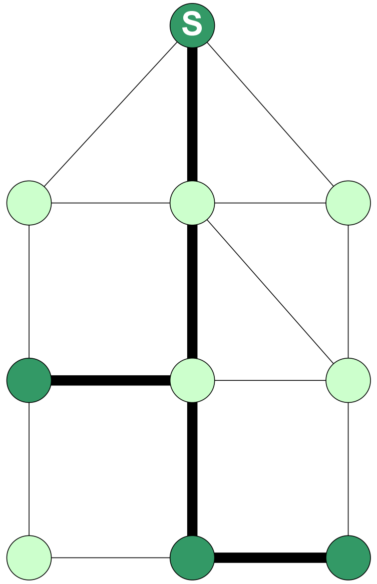
- Creation and management of **multicast trees**
  - multicast tree = tree that contains all the routers in whose (sub)nets hosts of the MC group are located
- Distribution of multicast IP packets (**MC forwarding**)
  - via multicast trees (better than simple flooding/ broadcasting)



MC Routing between (Sub)nets - Multicast Tree Variants

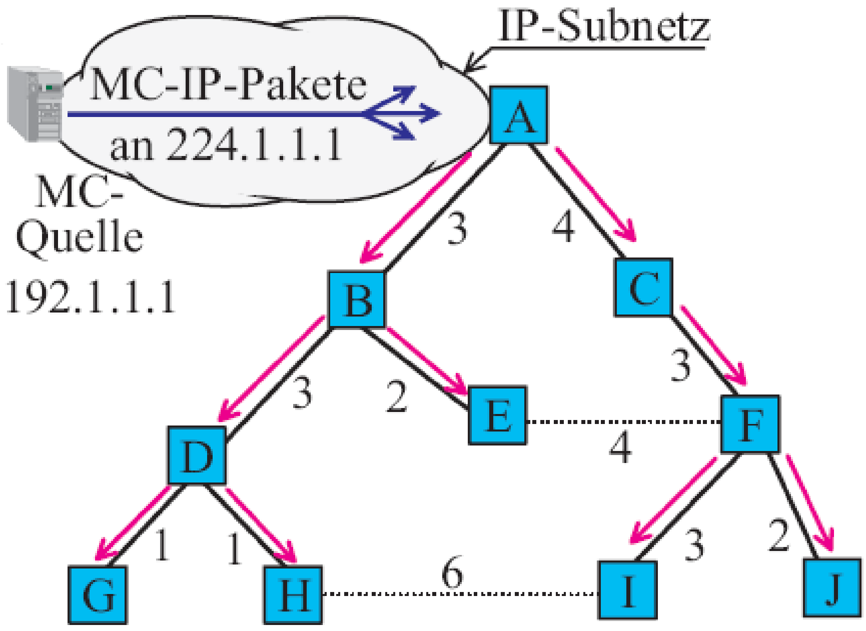
- **Option 1: source-based tree**
  - if the MC group contains more than one sender: one source-based tree per sender
  - ideal source-based tree: Shortest Path Tree = tree of shortest paths from the sender to the receivers; easy calculation via Dijkstra's algorithm
- **Option 2: shared core-based tree**
  - only one tree, shared by all senders/receivers of the MC group
  - ideal shared core-based tree: Minimum Cost Tree = minimum cost tree, which connects all senders/receivers of the MC group (Steiner tree); very complex calculation (NP hard!), therefore, often a simple heuristic solution via shortest paths to a predetermined core or center node (**Rendezvous Point**) is applied



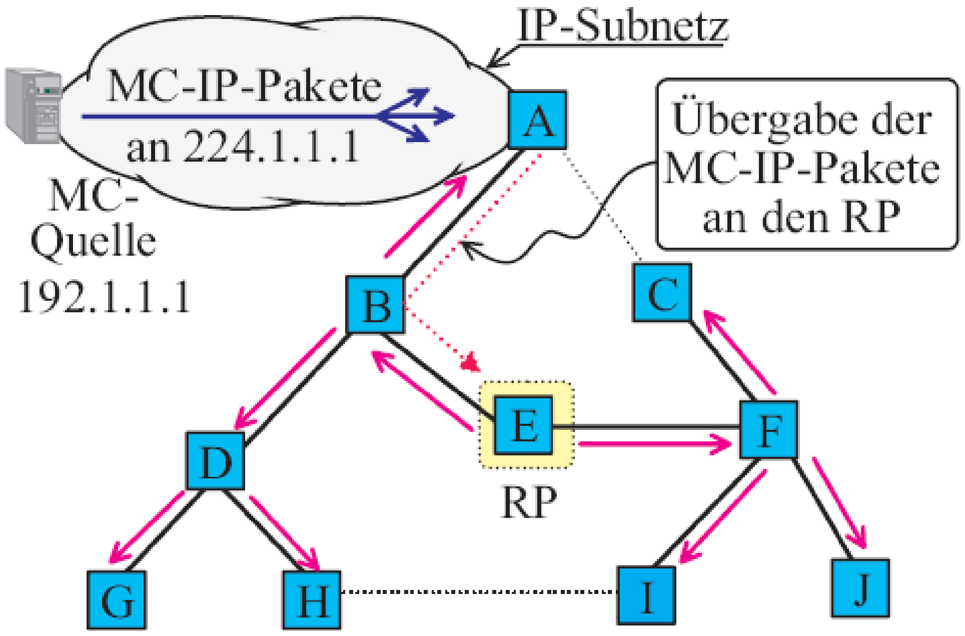


MC Routing between (Sub)nets - Multicast Tree Variants

- **Examples of multicast trees for option 1 and 2**



source-based tree



shared core-based tree
(shared multicast tree with core node as root)

MC Routing between (Sub)nets - Practical Application

- **Application of option 1 (source-based tree)**
  - determination of the tree of shortest paths from a sender to the receivers of the MC group **for each sender**
  - the tree is either constructed either implicitly by flooding starting from the source and Reverse Path Forwarding (RPF) or explicitly by registering the receivers at the source (explicit join)
  - the MC routing tables contain entries of the form (source **S**, group **G**)
- **Application of option 2 (shared core-based tree)**
  - determination of a single tree that connects all senders/receivers of the MC group; due to complexity reasons often only the shortest paths to a predetermined core or center node (**Rendezvous Point**) are determined (and not the cost-optimal Steiner tree)
  - the tree is either constructed implicitly by flooding starting from the Rendezvous Point and Reverse Path Forwarding (RPF) or explicitly by registering the receivers at the Rendezvous Point
  - the MC routing tables contain entries of the form (*, group **G**), since the sender (source) plays no role

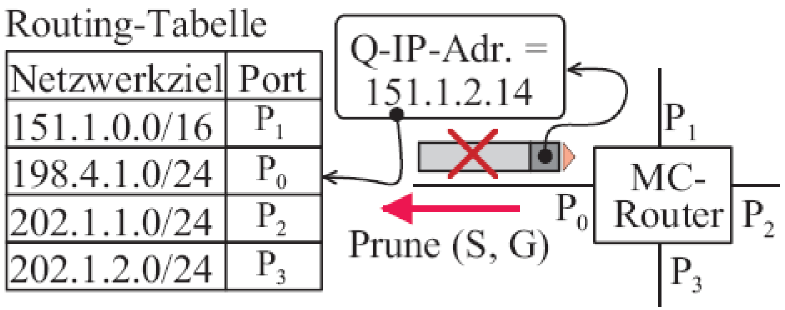
MC Routing between (Sub)nets - Source-based Trees

- Application of option 1 (source-based tree) - details:
  - a distribution tree to all receivers (= active group members except the source) is generated from each source of the MC Group
  - possible methods for constructing the multicast tree:
    - flooding of MC IP packets starting from the source and application of the **reverse path forwarding (RPF) method** (+ some improvements) → step by step construction of the multicast tree
    - explicit registration of the receivers at the source (**explicit join procedure**) → concurrent construction of the multicast tree
  - after construction of a source-specific multicast tree the routing tables in the MC routers contain entries (S, G) and the multicast traffic from the source to the receivers is routed along the multicast tree (flooding is no longer required)

MC Routing between (Sub)nets - Reverse Path Forwarding

- Reverse Path Forwarding (RPF) basic principle:
  - a MC Router remembers the source address (= address of the MC sender) of incoming MC IP packets and the interface on which the packets arrive
  - if this interface lies on the shortest route in reverse direction to the source (which is known from "normal" intra-domain routing), then MC IP packets from this source are forwarded via all other interfaces - otherwise no forwarding is performed; remark: the interface that lies on the shortest route in reverse direction is called **RPF interface**

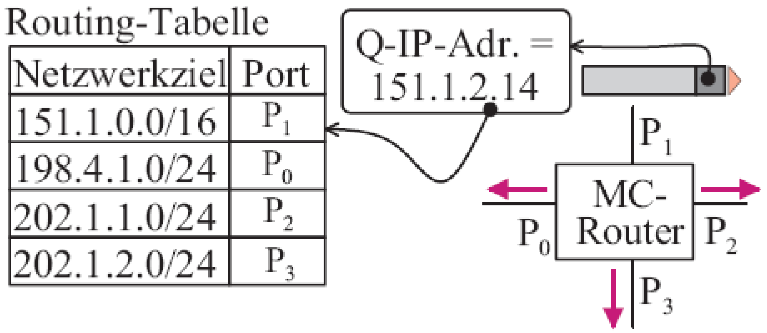
  Advantage of RPF compared to simple flooding: only MC IP packets that arrived on the shortest path from the source are forwarded



Routing-Tabelle

Netzwerkziel	Port
151.1.0.0/16	P_1
198.4.1.0/24	P_0
202.1.1.0/24	P_2
202.1.2.0/24	P_3

arrival at “wrong" interface



Routing-Tabelle

Netzwerkziel	Port
151.1.0.0/16	P_1
198.4.1.0/24	P_0
202.1.1.0/24	P_2
202.1.2.0/24	P_3

arrival at “correct" interface

MC Routing between (Sub)nets - RPF Improvements

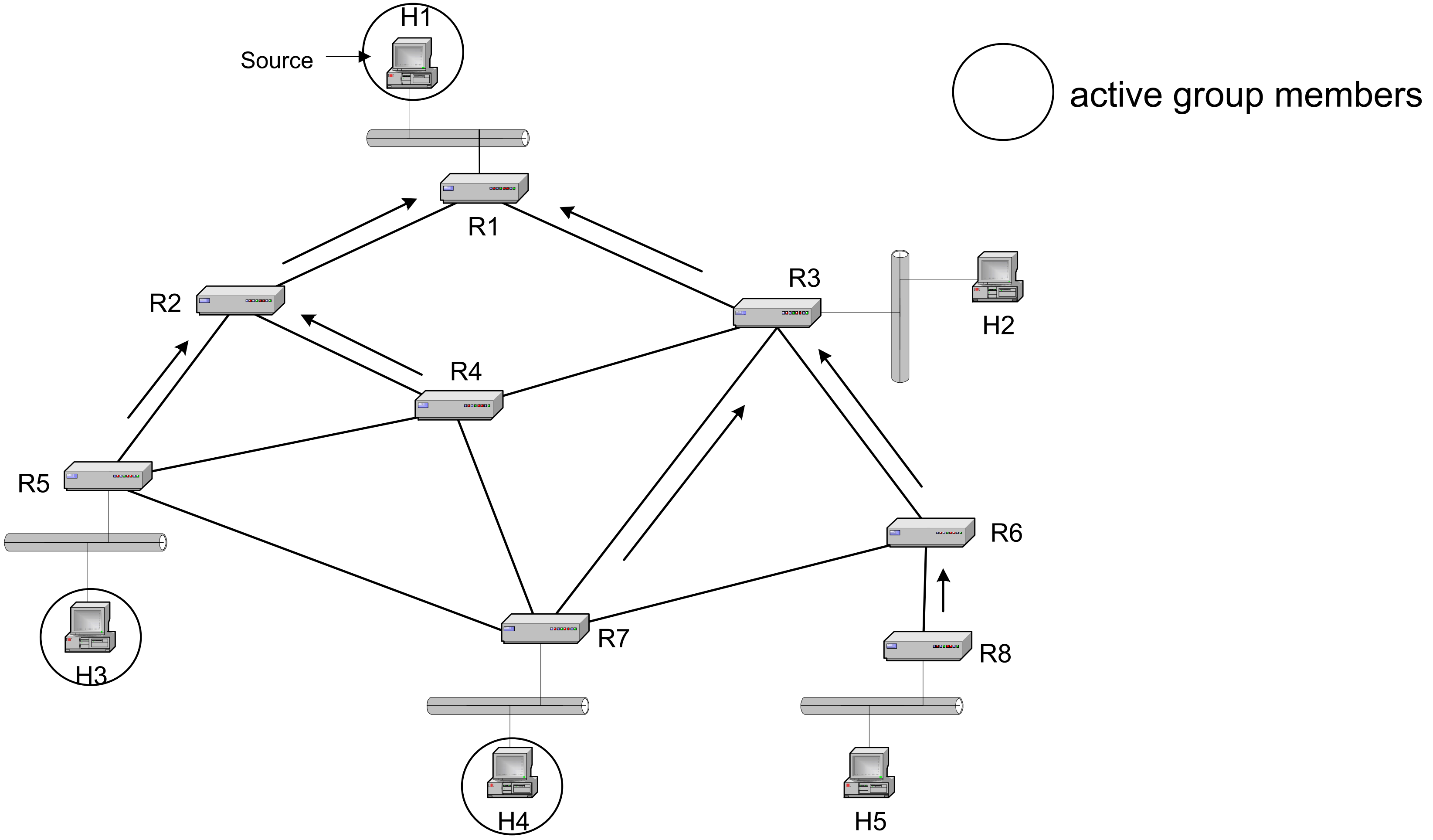
- RPF improvements:
  1) MC packets should be forwarded only via MC router interfaces which lie on the shortest path in reverse direction to the source → efficiency improvement but no consideration of the MC group structure
  2) MC packets should be forwarded only through MC router interfaces, if - additionally to 1) - there are other MC group members (of the MC group (S, G)) located in networks in upstream direction
  3) MC packets should be forwarded only through MC router interfaces, if there are other MC group members located in the respective (sub)nets
- Improvement 3) is easy to implement, since the information about MC group members in (sub)nets is derived via IGMP
- Improvements 1) and 2) are realized by introducing specific signaling messages (**Prune** or **Graft**):
  – by using the message **Prune (S, G)** a MC router tells its neighboring MC routers that they should not send him MC IP packets addressed from the source S to the group G anymore
  – the message **Graft (S, G)** has the opposite effect

MC Routing between (Sub)nets - Pruning / Grafting

- Pruning:
  - MC routers send Prune messages (Prune (S, G)) in response to received MC IP packets in the following situations:
    - if the MC packets arrived on an interface which lies on a non-shortest path in the reverse direction to the source (non-RPF interface)
    - if there are no group members in the connected (sub)nets (known through IGMP) and there are no connections to other MC routers
    - if there are no group members in the connected (sub)nets (known through IGMP) and all non-RPF interfaces to neighboring MC routers already received Prune messages
  - effect of Pruning:
    - temporary deletion of the corresponding links from the distribution tree (via deactivation of the routing table entry)
    - subsequently no MC IP packets are forwarded to these links
- Grafting:
  - a MC router sends a graft message (Graft (S, G)) via its RPF interface to cancel a previous link-deletion caused by Pruning; this is necessary if for example new members want to join the group (in networks where currently are no group members located)

MC Routing between (Sub)nets - Tree Construction (Opt. 1)

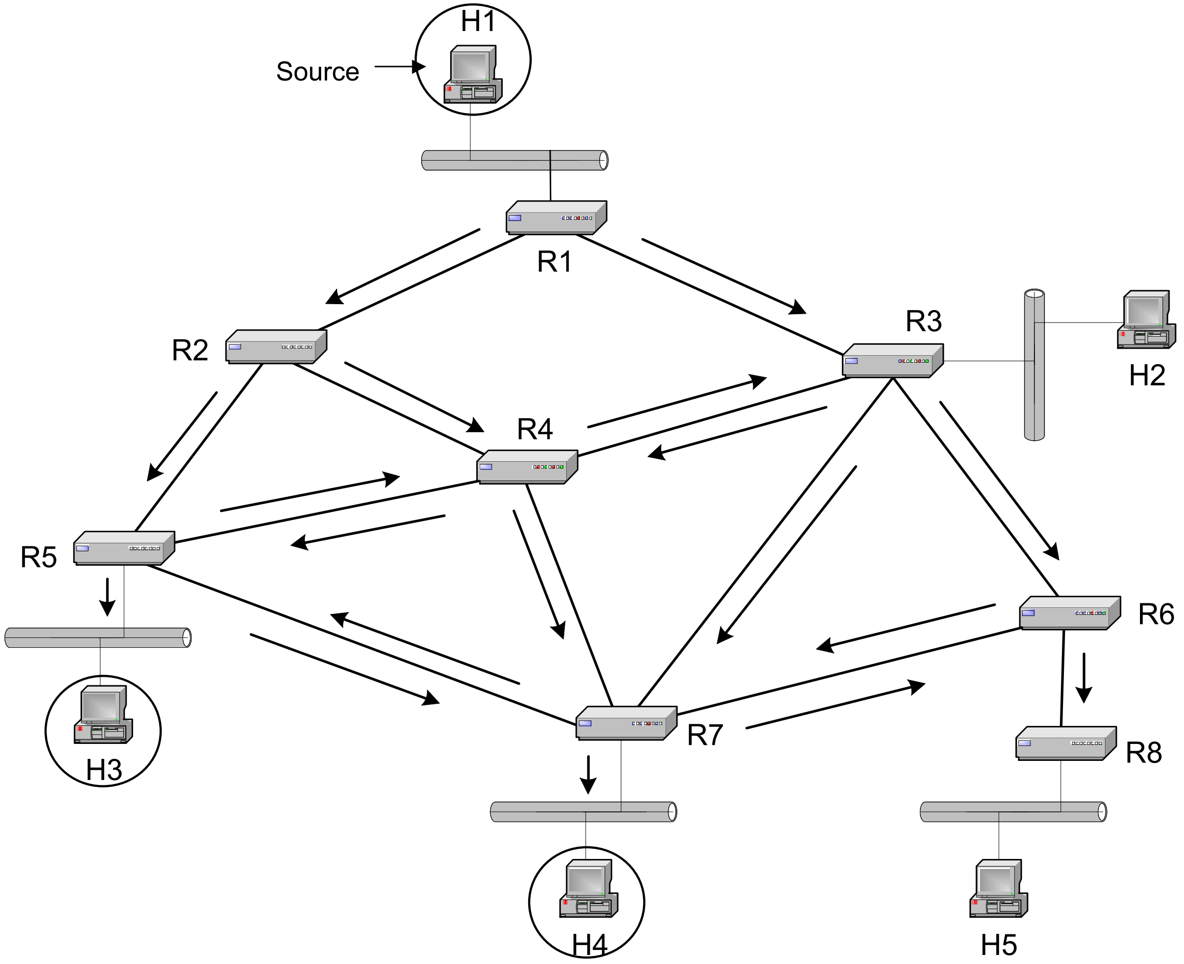
- Shortest routes to source H1:



MC Routing between (Sub)nets - Tree Construction (Opt. 1)

- RPF basic principle:

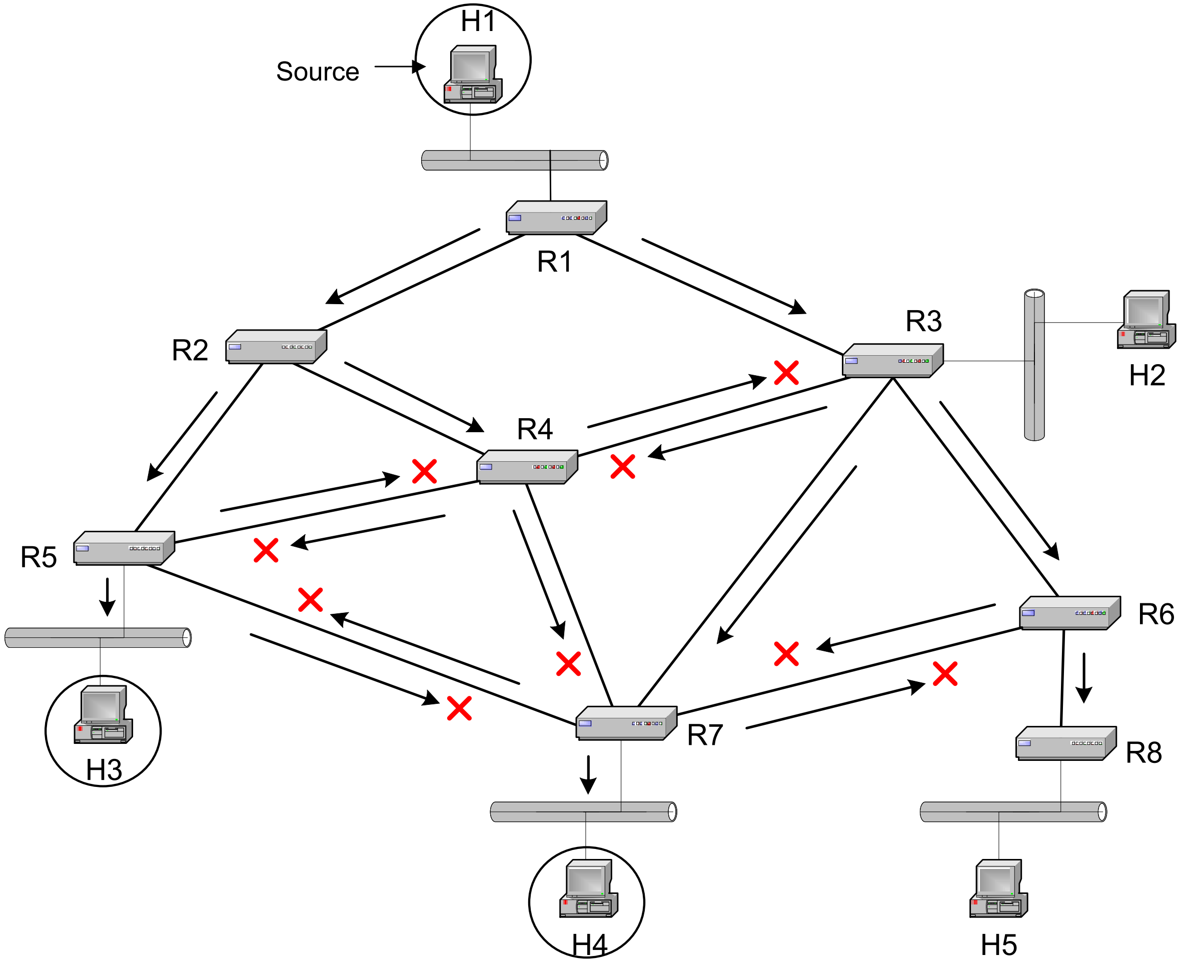
If the interface on which MC IP packets arrive, belongs to the shortest route in reverse direction to the source (RPF interface), then the MC IP packets are forwarded via all other interfaces - otherwise no forwarding happens



MC Routing between (Sub)nets - Tree Construction (Opt. 1)

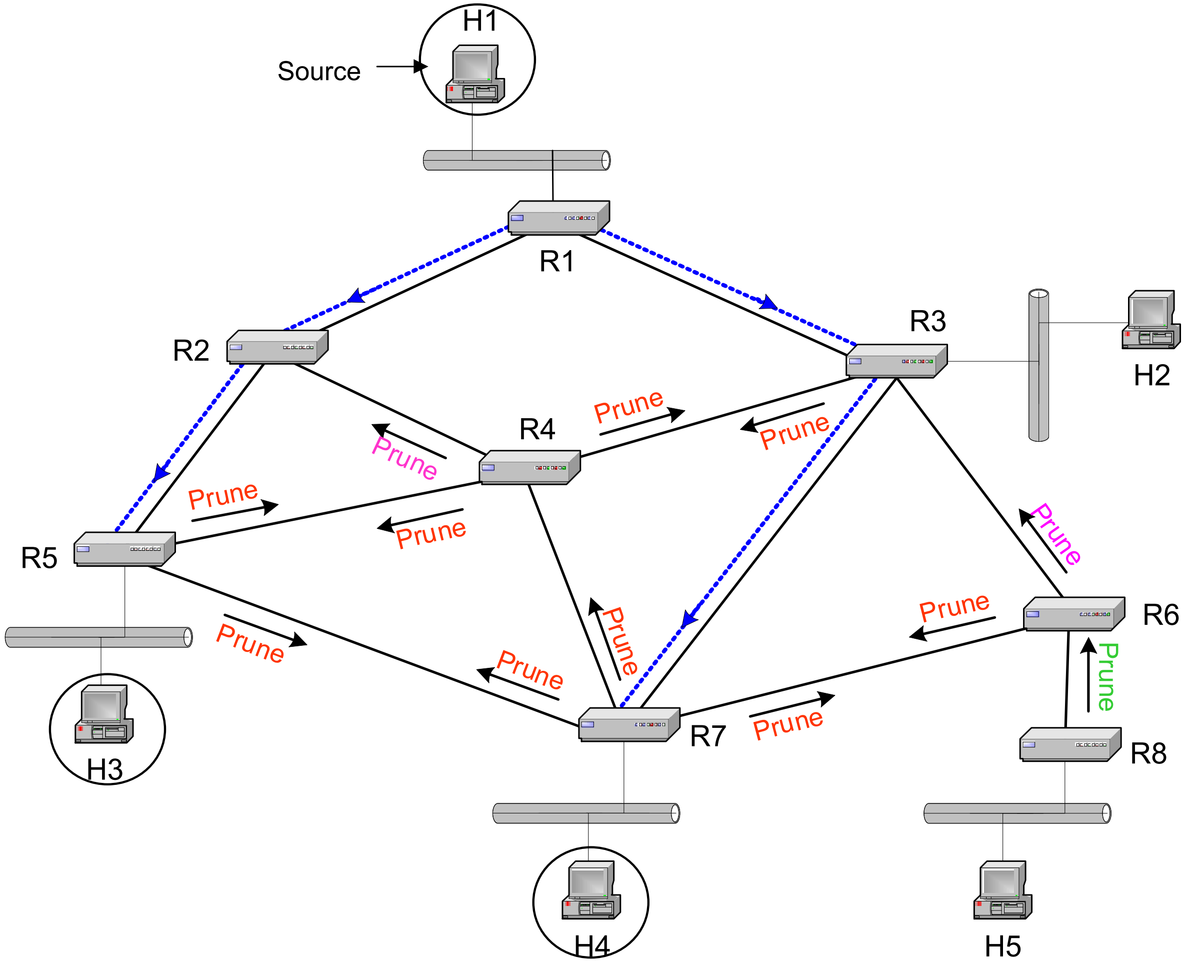
- RPF basic principle:

MC IP packets marked with ✗ are not received on RPF interfaces, and therefore are not forwarded by the MC routers



MC Routing between (Sub)nets - Tree Construction (Opt. 1)

- RPF improved with Pruning:



Prune messages are sent:

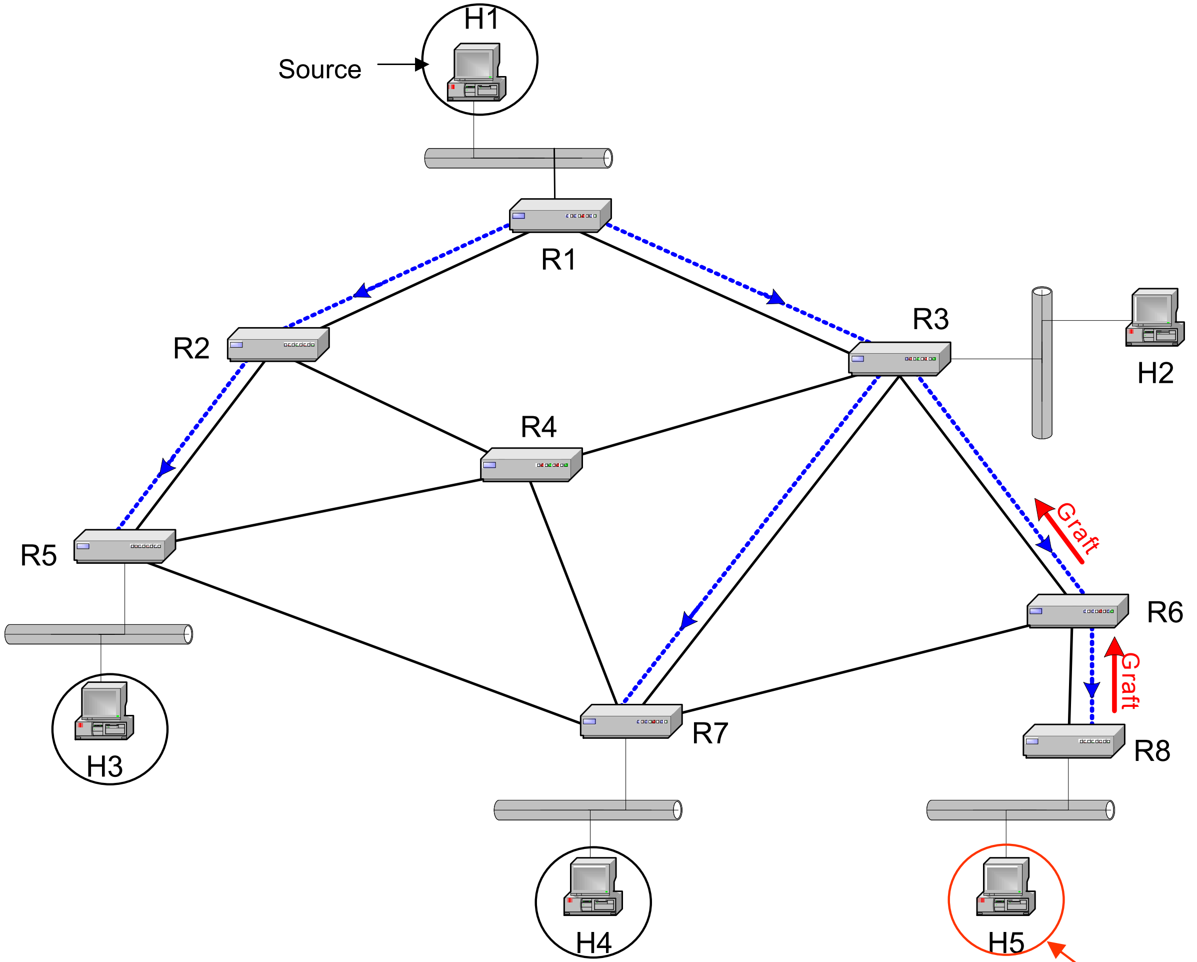
- if MC packets arrived on an interface which lies on a non-shortest path in the reverse direction to the source (non-RPF interface)
- if there are no group members in the connected (sub)nets and there are no connections to other MC routers
- if there are no group members in the connected (sub)nets and all non-RPF interfaces to neighboring MC routers already received Prune messages

— resulting multicast tree (Source-based Tree) for source H1 and active group members H3 and H4

MC Routing between (Sub)nets - Tree Construction (Opt. 1)

- Grafting:

A MC router sends a Graft message via its RPF interface to cancel a previous link-deletion caused by Pruning

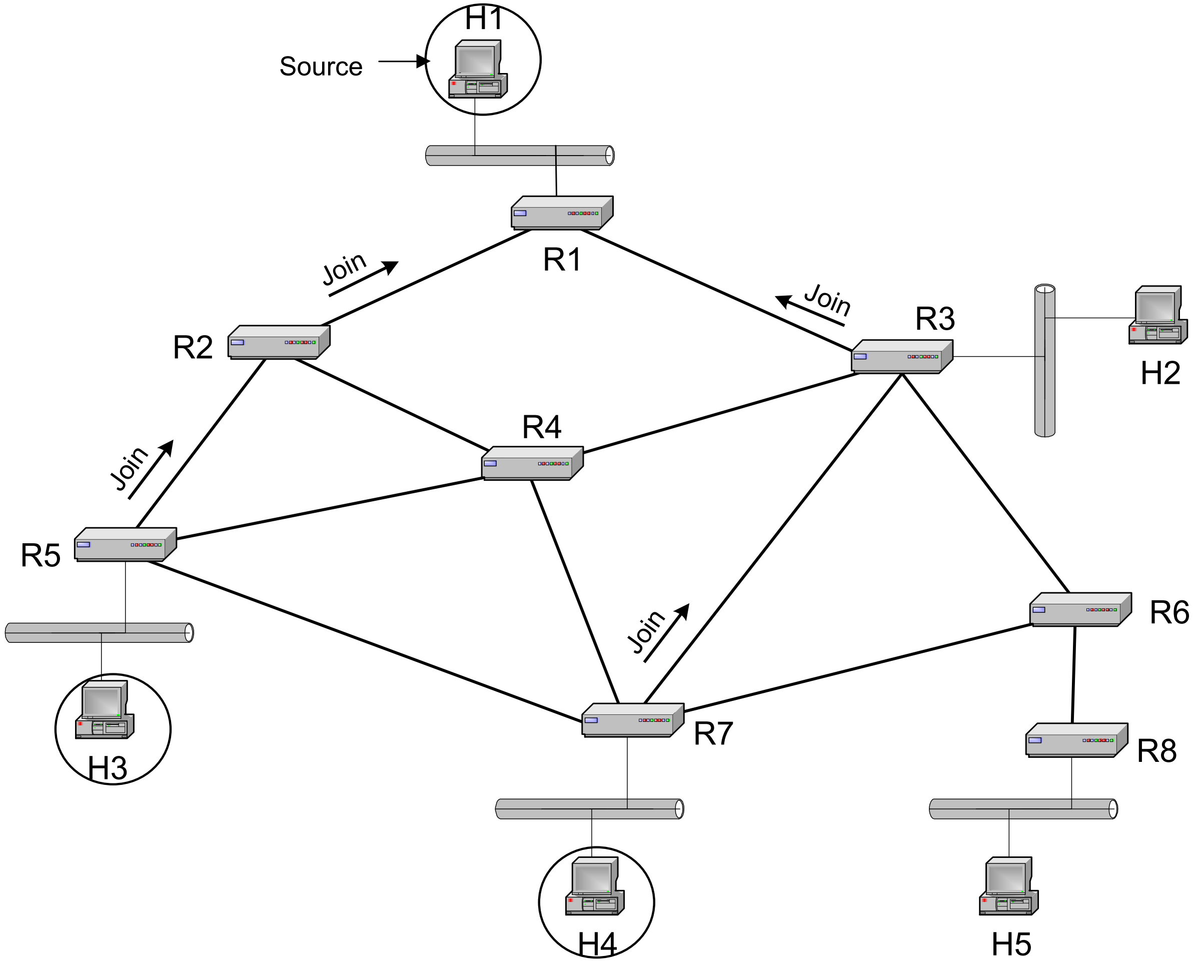


🟦 resulting multicast tree (Source-based Tree) for source H1 and active group members H3, H4 and H5

new group member

MC Routing between (Sub)nets - Tree Construction (Opt. 1)

- Alternative tree construction with "Explicit-Join" (instead of RPF):



Explicit-Join:

- MC routers whose (sub)nets contain active group members send "Explicit-Join" messages to the source
- these messages are forwarded on the shortest paths (wrt. the intra-domain routing) to the source - in the traversed routers the MC routing table entries (S, G) are generated
- **Explicit-join works only if the group members know the source!**

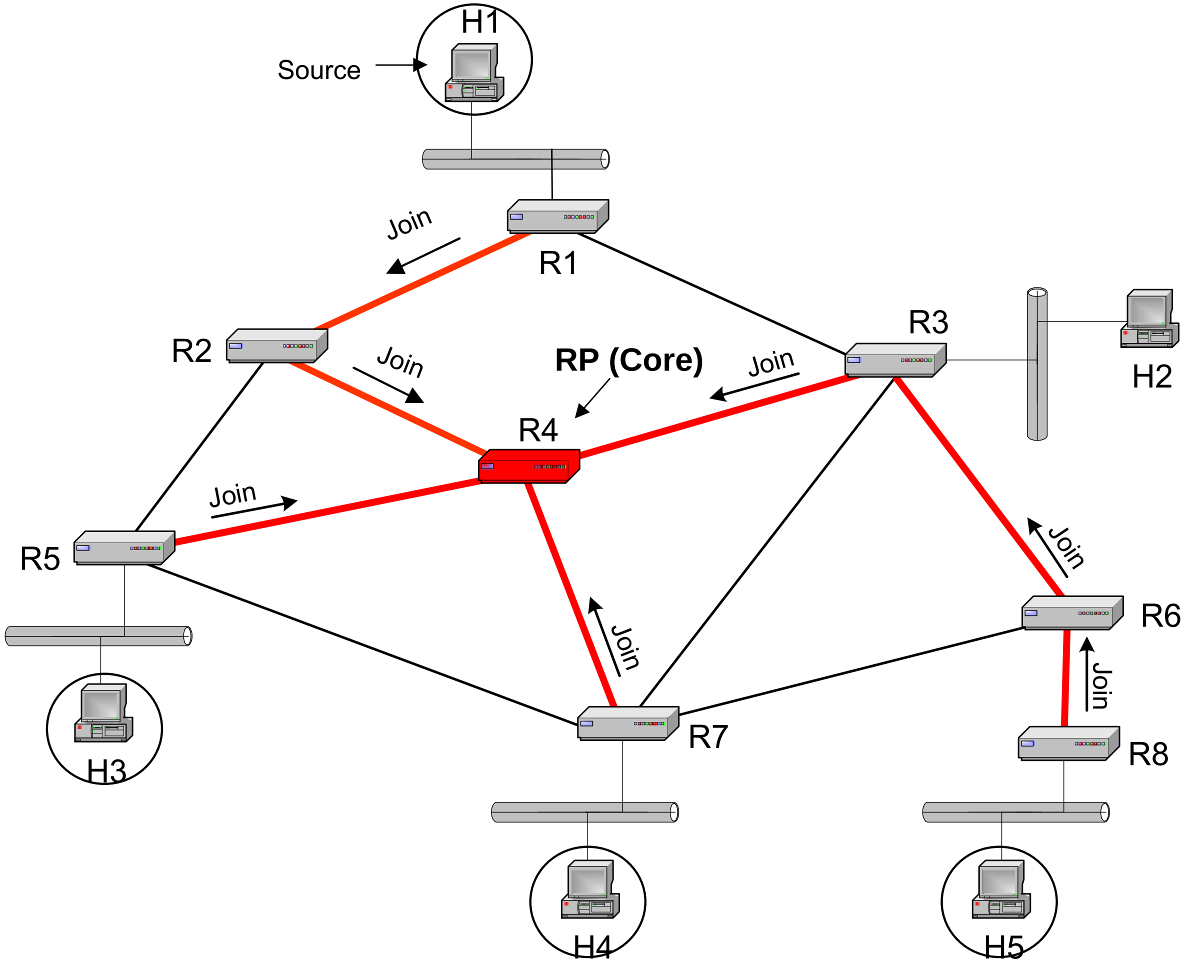
MC Routing between (Sub)nets - Shared Core-based Trees

- Application of option 2 (core-based shared tree) - details:
  - for a MC group one (or more) core/center node(s) (Rendezvous Point(s)) are selected
  - the group members register at the Rendezvous Point
  - starting from the Rendezvous Point a distribution tree to all MC group members is generated
  - possible methods for constructing the multicast tree:
    - initially MC IP packets are flooded from the Rendezvous Point and **Reverse Path Forwarding (RPF)** (+ some improvements) is applied → gradual establishment of the multicast tree
    - upon registration of the MC group members at the Rendezvous Point the distribution tree is constructed immediately (**Join procedure**); this is the preferred method
  - after construction of the core-based multicast tree the routing tables in the MC routers contain entries (*, G) and the multicast traffic from any source is forwarded to the Rendezvous Point and from there to the receivers via the distribution tree

MC Routing between (Sub)nets - Tree Construction (Opt. 2)

- Constructing the multicast tree with "Join":

Upon registration of the MC group members at the Rendezvous Point (RP) the multicast tree is construced immediately (Join procedure); the routing tables in the MC routers then contain entries (*, G)

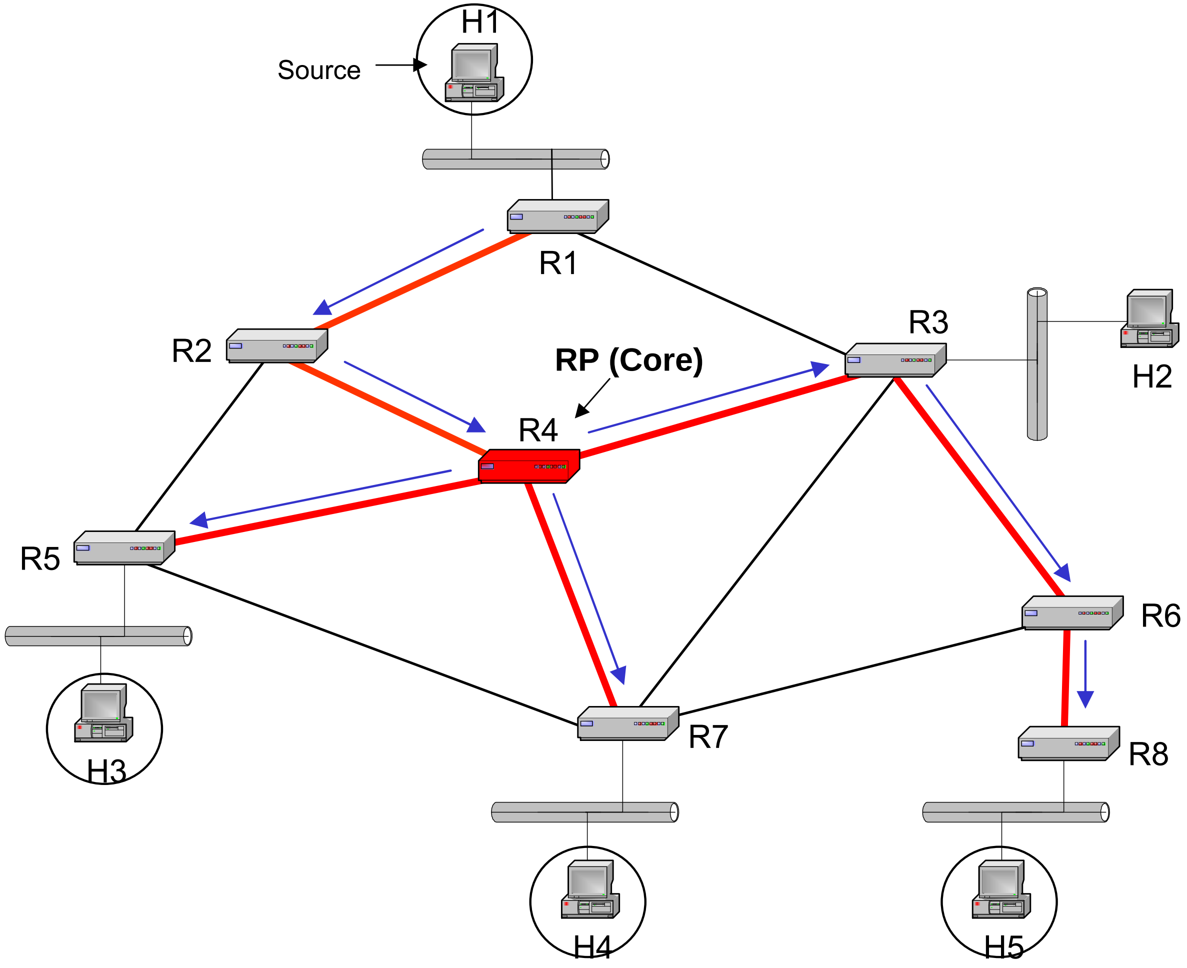


🟥 resulting multicast tree (Shared core-based tree) for active group members H1, H3, H4 and H5

MC Routing between (Sub)nets - Tree Construction (Opt. 2)

- Forwarding in Shared Core-based Trees:

The source initially sends MC IP packets to the Rendezvous Point; from there they are forwarded to the receivers via the multicast tree

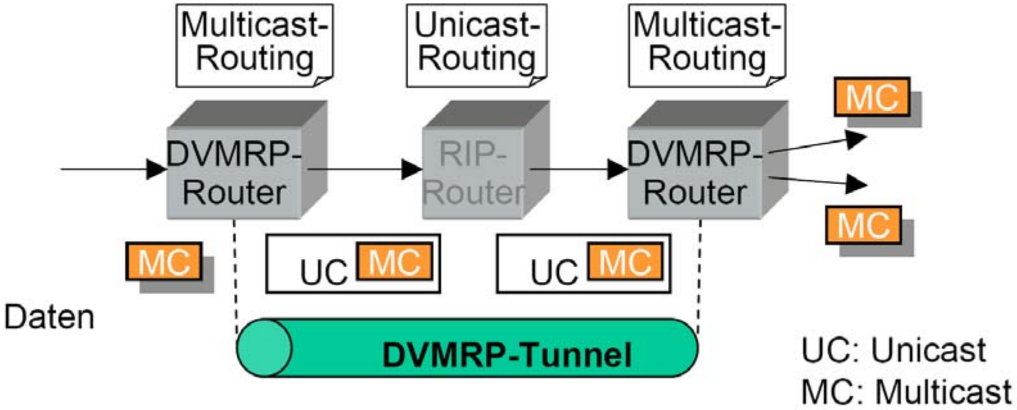


Multicast Routing Protocols

- Intra-domain MC routing protocols:
  - Distance Vector Multicast Routing Protocol (DVMRP), RFC 1075
    - MC extension of RIP based on source-based trees (tree construction by flooding/RPF and pruning) - first MC routing protocol
  - Multicast Open Shortest Path First (MOSPF), RFC 1584
    - MC extension of OSPF based on source-based trees
  - Core Based Tree (CBT), RFCs 2189, 2201
    - first MC routing protocol based on shared core-based trees
  - Protocol Independent Multicast (PIM)
    - two modes: PIM Dense Mode (PIM-DM) and PIM Sparse Mode (PIM-SM)
    - PIM-DM: based on source-based trees (tree construction by Flooding/RPF and Pruning), RFC 3973
    - PIM-SM: based on shared core-based trees and source-based trees (tree construction by "Explicit-Join"), RFC 4601
- Inter-domain MC routing protocols:
  - Border Gateway Multicast Routing Protocol (BGMP), RFC 3913
  - Multicast Source Discovery Protocol (MSDP), RFC 3618, 4611

Intradomain MC Routing - DVMRP

- DVMRP is based on a multicast extension of RIP
  - source-based method
  - RIP calculates the shortest path to source
- A DVMRP router uses two independent routing protocols
  - unicast routing protocol (RIP)
  - DVMRP for multicast routing
- For a practical application in the Internet (which works without MC support) tunneling is used
  - example: tunnel between 2 DVMRP routers



Intradomain MC Routing - MOSPF

- MOSPF is based on a multicast extension of OSPF
  - source-based method
  - MOSPF applies (like OSPF) the link-state principle
- The multicast extensions of OSPF are backward compatible
  - MOSPF routers can interwork with OSPF routers (unicast traffic) - RIP and DVMRP are not interoperable
  - in case of joint operation of OSPF and MOSPF in a network, a MOSPF router has to be elected as Designated Router, so that the multicast traffic can be forwarded to the appropriate (sub)net
- MOSPF extensions compared to OSPF:
  - the local group membership must be known by the routers
  - for each pair of sender S and group G a separate multicast tree has to be calculated

Intradomain MC Routing - PIM

- PIM uses the existing unicast routing protocol and is independent of the specific type of the unicast routing protocol
- PIM supports two different scenarios
  - spatially dispersed groups with low member density
  - spatially less dispersed groups with a high member density
- Objectives:
  - minimizing the states that have to be stored in the routers
  - minimizing the processing effort of control and user data
  - minimizing the network capacity required for control and user data
- PIM is an umbrella term for two different protocols:
  - **PIM Dense Mode (PIM-DM)**
    - for groups with a high member density
    - based on flooding and pruning
  - **PIM Sparse Mode (PIM-SM)**
    - for groups with low member density
    - based on rendezvous points

Intradomain MC Routing - PIM-SM Operation

- PIM-SM works in 3 phases:

  Phase 1: Use of a shared core-based tree

  Phase 2: Transition to source-based trees

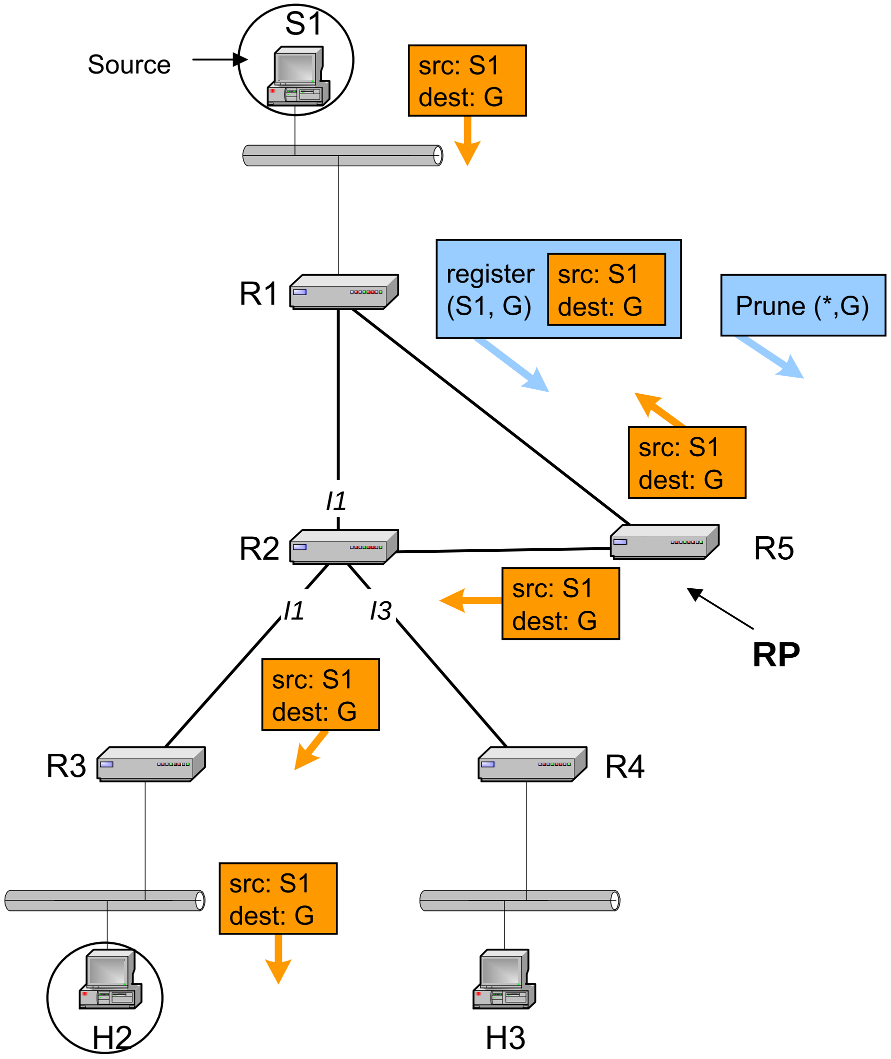
  Phase 3: Use of source-based trees

Intradomain MC Routing - PIM-SM Operation

- Phase 1: use of a shared core-based tree (register-phase)
  - for MC support in each (sub)net that contains MC group members, a Designated Router (DR) is required; DR and group members communicate via IGMP
  - the MC group members know the Rendezvous Point (RP) and register via a Join (*, G) message; at the same time the shared core-based tree (with RP as root) is set up
  - the MC source sends IP packets to the RP - from there they are distributed over the shared core-based tree; the MC IP packets are transported in so-called register messages (with which the MC source has registered at the RP); the MC IP packets are distributed from the RP via the shared core-based tree to all MC group members - also (unnecessarily) back to the MC source; the source (or the DR of the (sub)net) notes this and sends a Prune (*, G) message to the RP to avoid further reception of MC IP packets (i.e. the route between source and RP is removed from the shared Core-based Tree)

Intradomain MC Routing - PIM-SM Operation

- Phase 1: use of a shared core-based tree (register-phase)

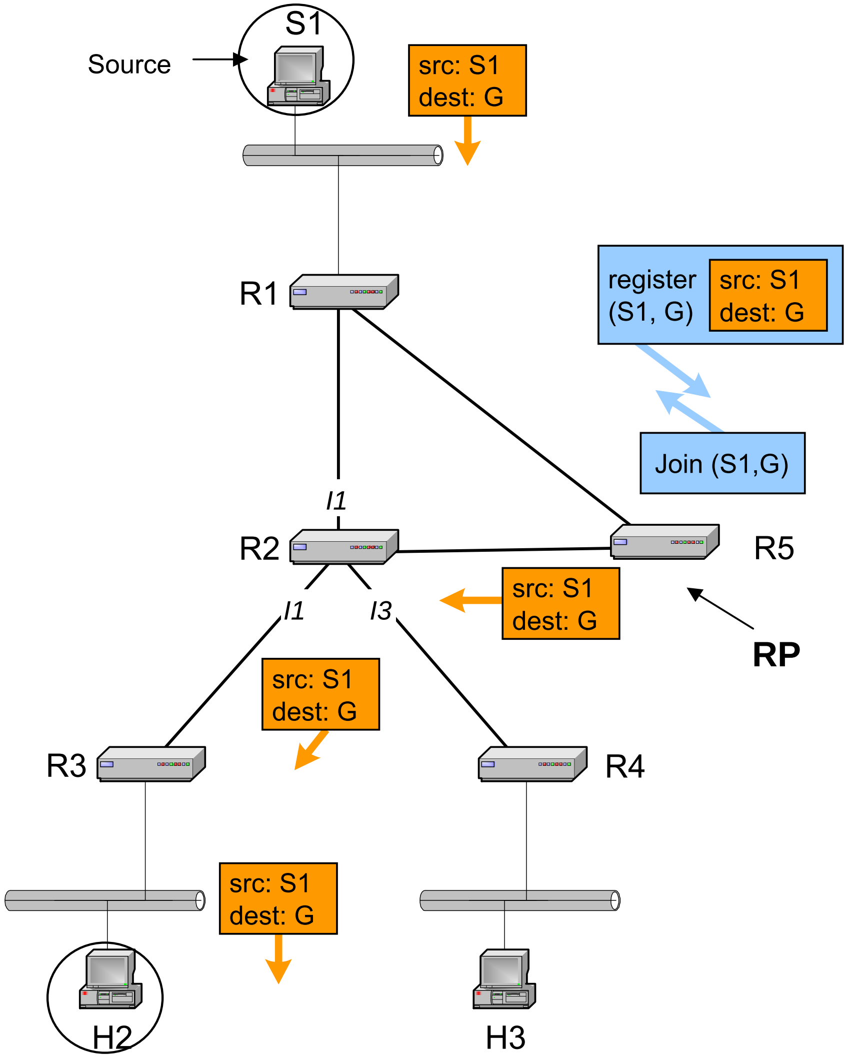


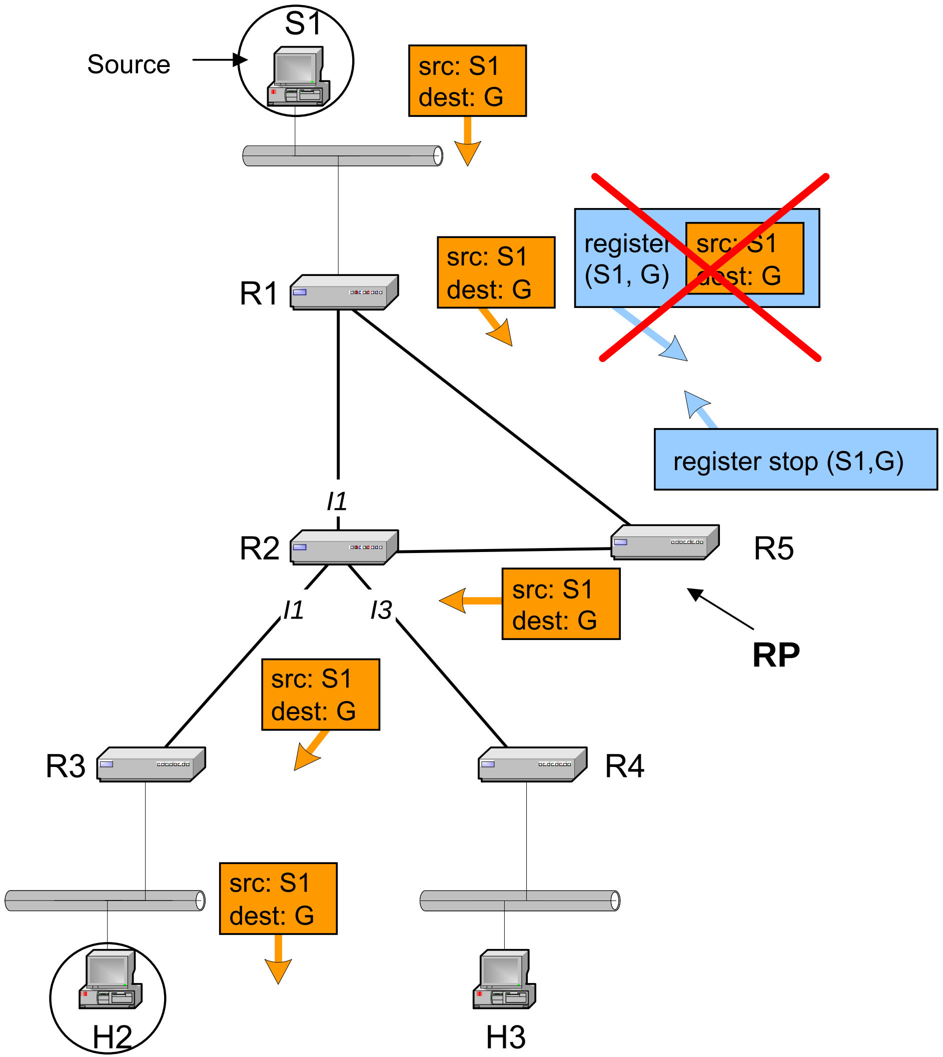
Intradomain MC Routing - PIM-SM Operation

- Phase 2: transition to source-based trees - part 1 (register-stop phase)
  - the use of a shared core-based tree may lead to a non-optimal distribution of MC IP packets from a source - hence PIM-SM allows a gradual transition to a source-based tree during the ongoing distribution
  - the transition is initiated by the RP by sending the message Join (S, G) to the source - this indicates that the RP intends to join the source-based tree (with the source S as root); the message Join (S, G) is forwarded on the shortest path (according to intra-domain routing) to the source and causes a change of the routing table entries wrt. the source-based tree in the traversed MC routers
  - the RP now receives MC IP packets from the source S in two ways: encapsulated in Register messages and via the source-based tree; to stop the reception of Register messages the RP sends the message Register Stop (S, G) to the source
  - now MC IP packets are sent from the source to the RP via the source-based tree and from there they are distributed via the shared core-based tree

Intradomain MC Routing - PIM-SM Operation

- Phase 2: transition to source-based trees - part 1 (register-stop phase)



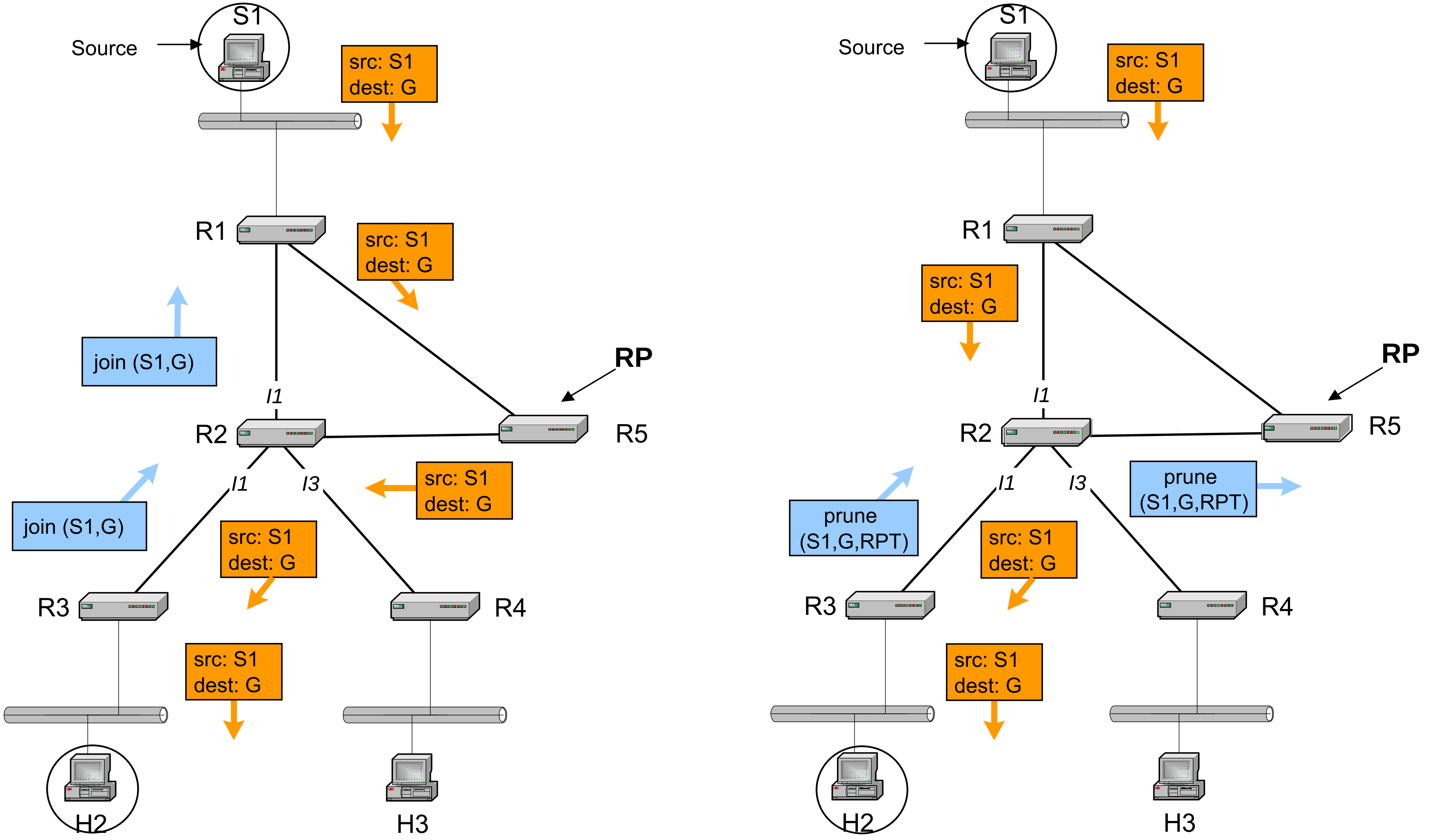


Intradomain MC Routing - PIM-SM Operation

- Phase 2: transition to source-based trees - part 2
  - the full transition to a source-based tree is triggered by MC routers in whose (sub)nets receivers (group members) are located; if the MC traffic to the receivers (which is still forwarded on the core-based shared tree) exceeds a certain threshold, the MC routers send an Explicit-Join (S, G) message to the source; this message traverses along the shortest path (according to intra-domain routing) to the source and causes a change of the routing table entries wrt. the source-based tree in the traversed MC routers
  - during the transition phase it is possible that some MC routers receive MC IP packets twice (via the shared core-based tree the and source-based tree); with the message Prune (S, G, RPT) to the RP the MC routers announce their wish not to receive further MC IP packets over the shared core-based tree; at the end only the source-based tree remains as multicast tree

Intradomain MC Routing - PIM-SM Operation

- Phase 2: transition to source-based trees - part 2



Intradomain MC Routing - PIM-DM

- Assumption
  - the group members are located in almost all subnets
- Mechanism
  - flooding (with reverse-path forwarding) and pruning (similar to DVMRP)
  - pruning is associated with (S, G)
  - remark: the same message format as for PIM-SM is used
- Neighbor detection
  - by periodic hello messages
- Differences compared to DVMRP
  - not dependent on procedures for topology discovery
  - use of the existing unicast routing protocol

Interdomain MC Routing - BGMP

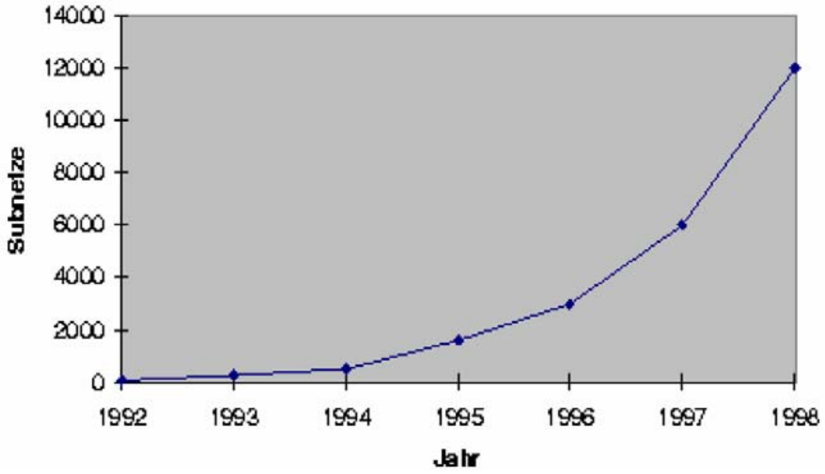
- BGMP is an inter-domain multicast protocol with some similarities to BGP
- BGMP supports the following multicast trees:
  - unidirectional source-based multicast trees
  - unidirectional shared multicast trees
  - bi-directional shared multicast trees
- MC IP addresses are distributed by the MBGP (Multiprotocol Extensions for BGP-4) protocol

Interdomain MC Routing - MSDP

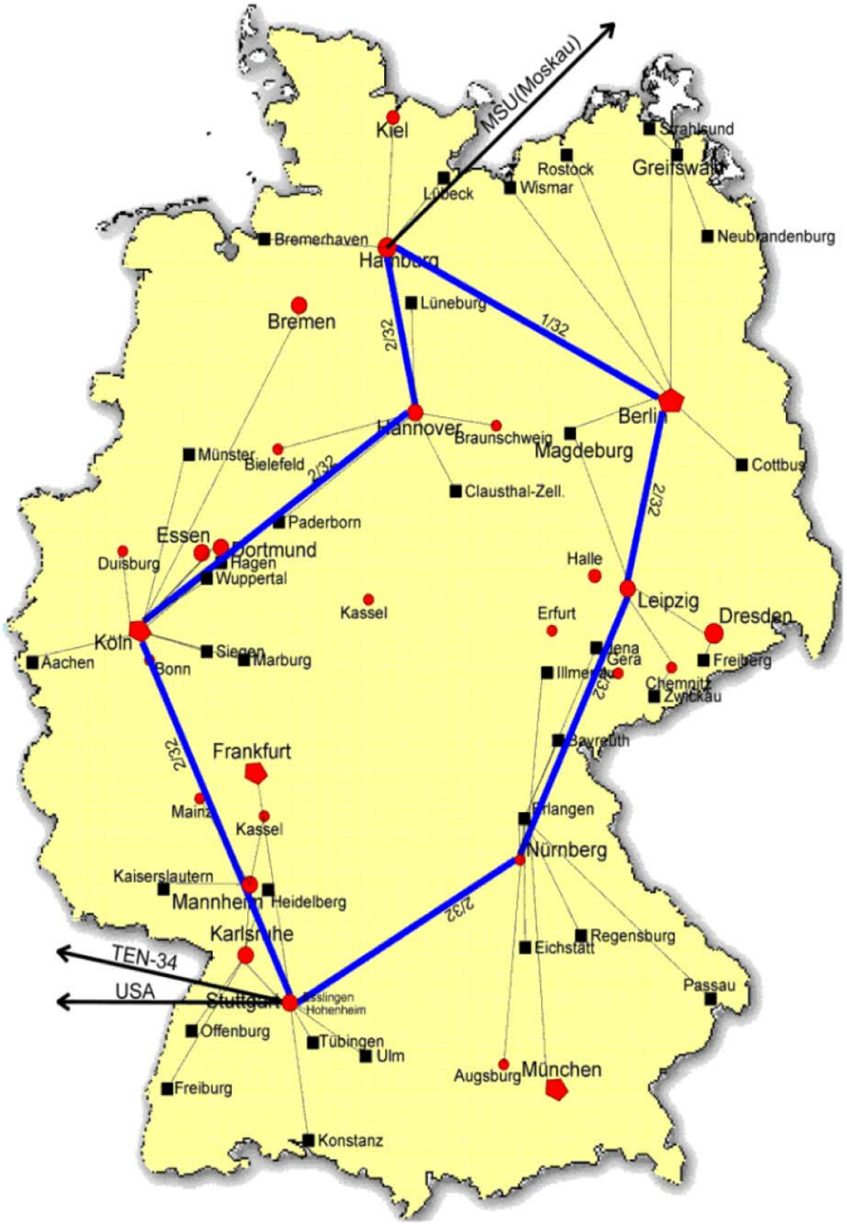
- MSDP is regarded as an intermediate strategy and not a long-term solution
- Domains (AS) are using PIM-SM and contain a full set of Rendezvous Points (i.e. for all multicast groups)
- MSDP allows a loose interconnection of Rendezvous Points between the domains
- In case a sender in one domain is becoming active, all Rendezvous Points which are connected via MSDP are notified via a Source Active message
- Neighboring Rendezvous Points send a sender-specific Join message to the sender
- Disadvantage: poor scalability, since each Rendezvous Point has to be notified about the activity of the sender

MBone (Multicast Backbone On the interNEt)

- Motivation: no support of group communication in the Internet yet
  - no multicast address management
  - no group management
- The introduction of group communication requires MC-enabled end systems (hosts) and intermediate systems (routers)
- Therefore: only gradual introduction of group communication
- Concept: MBone as overlay network
- MBone was launched in March 1992 to enable a multicast audio transmission during an IETF meeting
- Growth of the MBone:



German MBone (DFN Multicast)

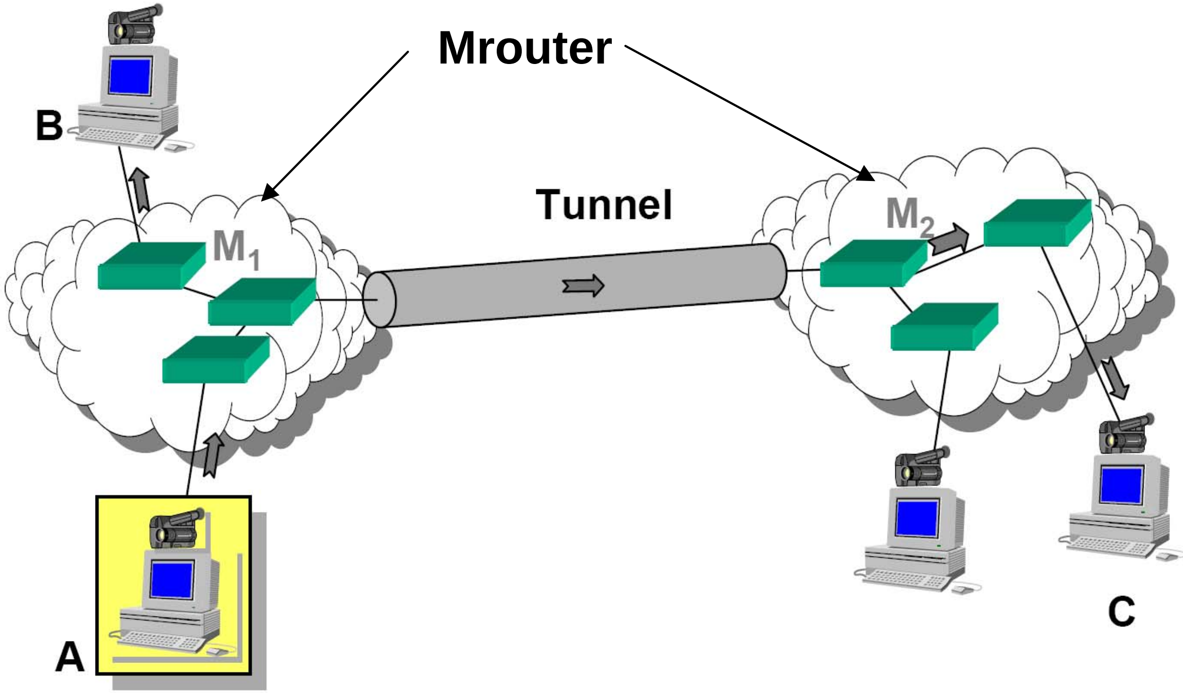


German MBone (multicast DFN)

- standard service in the German scientific network (G-WiN)
- native IP multicast in the backbone (G-WiN) up to the customer edge routers
- Protocols:
  - intra-domain
    - OSPF
    - PIM-SM, PIM-DM
  - inter-domain
    - BGMP / MBGP
    - MSDP (Multicast Source Discovery Protocol)
- approximately 76 participants

Architecture of the MBone

- Components of the MBone architecture
  - multicast domains with multicast-enabled systems
  - tunneling to interconnect multicast domains
    - virtual connections between multicast-enabled tunnel endpoints
    - the tunnel endpoints are denoted as MRouters
    - the tunnels are unidirectional
    - example:


Mrouter
B
Tunnel
M_1
M_2
A
C