CS 856
Internet Transport Performance

Network Architecture:
Mobility, Multicast and Overlays

Martin Karsten

School of Computer Science, University of Waterloo

mkarsten@uwaterloo.ca

Contents

Mobile Communication

Current Predominant Application: Voice Communication

Current System Architectures

- multiple layers of packet switching and virtual circuits
  - extension of TDMA wireless access channels

Future?

- increasing amount of data communication
- changing access technologies, e.g. WLAN
- switch architectures to all packet switching

Main Issues

- handover: latency, packet loss, overhead
- paging
  - relates to naming and addressing
- hardware restrictions: processing & power
- efficient support of intra-domain traffic

IP Mobility

Addressing

- **IP address $\rightarrow$ identification AND location**
- **separation needed for mobile communication**

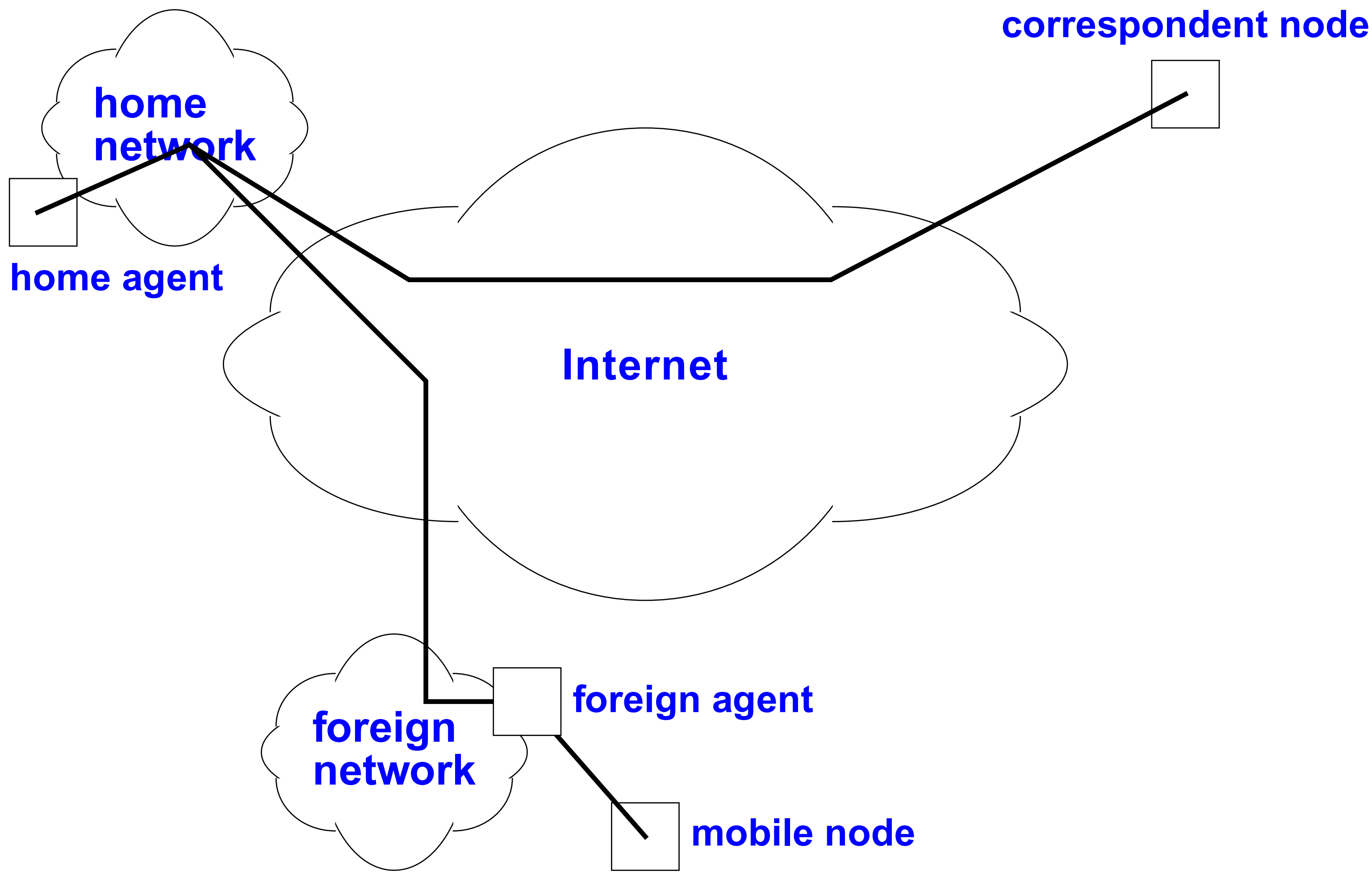
Mobile IP - Goals

- **transparency for correspondent node**
- **seamless integration into IP architecture**
- **not specifically targeted to voice communication**
  - e.g. scenario: mobile laptop connects to Internet for data communication
  - slower mobility timescale than e.g. cell phones

Mobile IP - Addressing

- **home address $\rightarrow$ identification**
- **care-of-address $\rightarrow$ location**
  - unique address
  - address of foreign agent (if mobile node can be reached via layer 2)

Mobile IP



Dirty Details

- mainly at: mobile node ↔ foreign agent

Mobile IP - Evaluation

Transparency ⇒ Triangle Routing

- "route optimization" (IPv4) / "binding update" (IPv6)

No Interaction with Radio Layer

- no notification mechanisms specified, no bicasting
- no fundamental obstacles either

End-to-End Operation

- triangle routing → home agent
- binding update → correspondent node
- high delay and potentially packet loss during handover
- active connectivity needed for paging

Overhead in Network

- only mobile agents are involved in connectivity
- handover → at routers: normal IP packets
- regular operation → at routers: normal IP packets

Hierarchical Mobility Solutions

Break End-to-End Association
- similar to hierarchical IP routing
- macro-mobility ~ inter-domain routing
- micro-mobility ~ intra-domain routing

Reduced Scope of Handover Updates
- improved handover latency
- reduced packet loss

Intermediate Chain Forwarding or Bicasting
- further reduce packet loss
- requires overlapping radio connectivity with multiple base stations
  - network design
  - wireless access technology

Mechanisms - Connection State
- IP in IP tunneling, e.g. Hierarchical Mobile IP
- separate routing, e.g. Cellular IP

Other Features

Passive Connectivity for Paging
- **requires support for paging area**
- **paging agent broadcasts or multicasts paging requests**
- **MN must detect paging area boundaries...**

Intra-Domain Traffic
- **often not considered in specification**
- **important in reality, e.g. "no airtime charge for calls from same network"**

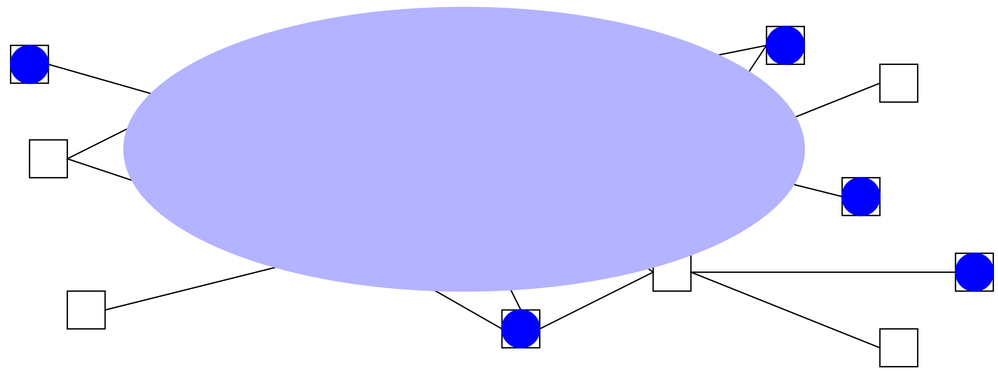
Interaction with Radio Layer
- **strong handoff radio trigger (SHRT)**

Fundamental Relationship to Mobility Architecture
- **omission of feature vs. infeasibility of feature**
  - e.g. paging in Mobile IP $\rightarrow$ always involves HA
  - e.g. SHRT in Mobile IP $\rightarrow$ would be possible, BUT: bicasting or chain?
    - chain forwarding would require changes to FA and MN
    - Hierarchical Mobile IP, Cellular IP: intermediate FA can bicast

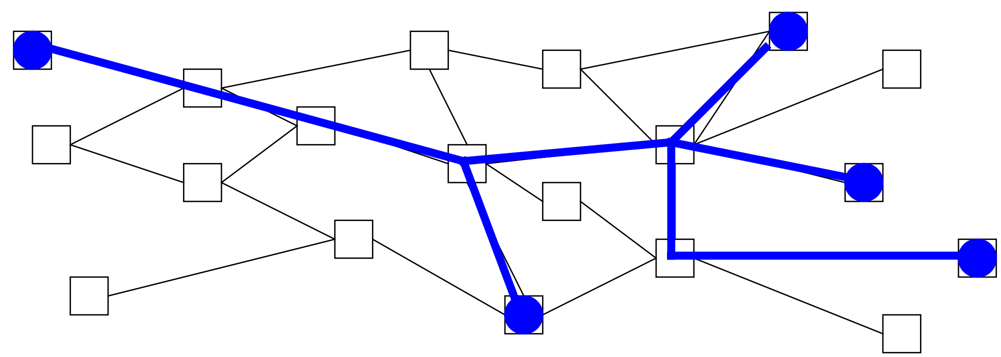
IP Multicast

Terminology

- **MULTIPOINT/GROUP COMMUNICATION (end system's viewpoint)**
  - multiple senders and/or multiple receivers
  - agnostic of actual transmission



- **MULTICAST TRANSMISSION (network's viewpoint)**
  - transmission along tree structure
  - replication of packets at branch nodes



IP Multicast

Traditional Multipoint Communication
- sender-initiated (or centrally organized) participation
- well-known participants
- static group membership
- bidirectional core-routed transmission
- individual addressing

Multicast Goals
- efficient resource utilization
- avoid traffic duplication

IP Multicast Model
- receiver-initiated join
- anonymous receivers
- dynamic membership
- independent, unidirectional transmission tree(s)
- group addressing

IP Multicast Addressing

Class D Network Addresses

1	1	1	0	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Address Range: 224.0.0.0 - 239.255.255.255

Further Partitioning

- see http://www.iana.org/assignments/multicast-addresses

Well-known Addresses

- routing protocols 224.0.0.0 - 224.0.0.255 (no data forwarding)
- all systems on subnet 224.0.0.1
- all routers on subnet 224.0.0.2
- DVMRP routers 224.0.0.4
- etc.

Multicast Routing

Multicast Tree Computation - Packet Distribution

- Flooding and Variants → Spanning Tree (per source)
- Link State → Spanning Tree (per source)
- Shared Trees
- assume (mostly) hierarchical network structure

Evaluation Criteria

- amount of generated traffic
- average path length
- computation complexity
- state complexity
- system convergence

General Trade-Off

- (+) transmission cost savings
- (-) increased system complexity
- (-) transmission cost overheads

Flooding

Characteristics/Variants

- **data-driven**
- **pure flooding**
- **(truncated) reverse path broadcasting → arrival via shortest path?**
  - truncate: pruning only at leafs
- **reverse path multicasting → recursive pruning**
  - initial join: periodic flood
  - join after prune: graft (multicast tree already exists)
- **using unicast routing information**

Evaluation

- **high transmission overhead (flooding part, amount depends on variant)**
- **sub-optimal path lengths (computation based on local state)**
- **low computation complexity**
- **state complexity (inverse to transmission overhead)**
  - low/constant: pure flooding, reverse path broadcasting
  - medium: truncated reverse path broadcasting (per group)
  - high: reverse path multicasting (per group, per sender)

⇒ **Suitable for densely populated multicast groups**

Link State

Characteristics/Variants

- **control-driven (join/leave)**
- **periodic flood of link-states**
- **with/without flood and prune**

Evaluation

- **some transmission overhead (if flood & prune is used)**
  - additionally: link-state routing overhead
- **optimal path lengths (computation based on global state)**
- **high computation complexity**
- **high state complexity (per group, per sender)**

⇒ **Suitable for intra-domain multicast routing**

Example: Multicast OSPF (RFC 1584)

- **extensions to OSPF**
- **every node calculates same optimal multicast tree (per source, per group)**
  - calculation triggered by first arriving data packet

Shared Trees

Characteristics/Variants

- **control-driven (join/leave)**
- **core-based tree**
- **Steiner tree (optimal core-based tree)**
  - NP-complete, unstable ⇒ not implemented

Evaluation

- **no transmission overhead**
- **no optimal path lengths (central point, independent of source)**
- **high computation complexity (computation based on global state)**
  - but only at central point
- **medium state complexity (mostly per group, not per source)**
  - except central point
- **central node with high load**
  - load balancing through nomination of multiple central points
- **central point of failure!**
  - often handled through backup nodes
  - but not inherently robust!

⇒ **Suitable for sparsely populated multicast groups**

Multicast Routing Protocols

IETF Multicast Protocols

- **IGMP, RFC 1112, 2236**
  - last-hop (broadcast network) group membership
  - communication (broadcast) between hops in distribution network
- **DVMRP, RFC 1075**
- **PIM (SM & DM), RFC 2362**
- **Core Based Tree (similar to PIM-SM), RFCs 2189 & 2201**
- **M-OSPF, RFC 1584**

Common Characteristics

- **soft state: state information times out if not refreshed**
- **not necessarily striving for optimal tree**

Key Distinction

- **source/group individual tree**
  - link state vs. flood & prune
- **group shared tree**

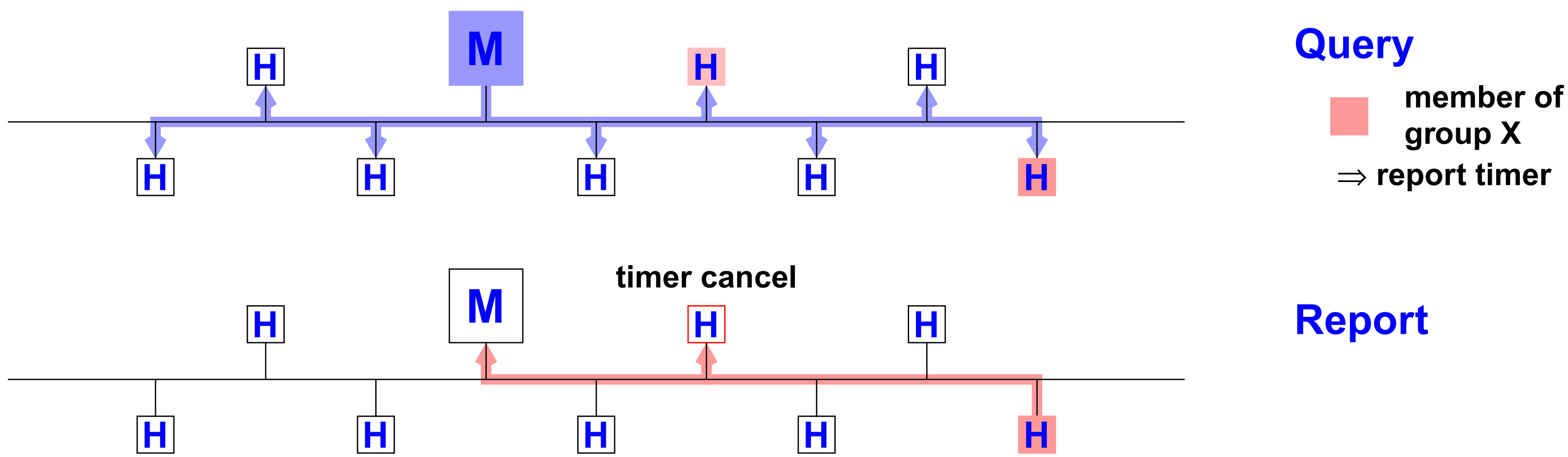
Internet Group Management Protocol

Query/Report about Group Memberships

- **primarily between router and hosts in LAN environment**
- **but also used as carrier for DVMRP**

Message Types

- **Query - sent periodically to 224.0.0.1**
- **Report - sent to respective multicast group (delayed)**
  - on query received or join



Random Timers → Reduction of Message Load

Internet Group Management Protocol

Version 2, Version 3 in progress
- **see RFC 2236, RFC 1112**
- **see http://www.ietf.org/html.charters/idmr-charter.html**

Change to v1: Generalized Terminology
- **roles in v1: "router" vs. "host"**
  - implicitly assumes 1 router and n hosts
- **roles in v2: "querier" vs. "non-querier"**
  - multiple multicast routers may exist on subnet

Extensions to v1
- **message type: group leave → better leave latency**
- **message type: group specific query**
- **flexible maximum response time setting in query**
  - set by local host configuration in v1
  - set dynamically by querier in v2
  - allows tuning state update latency & message load

⇒ **Increased Precision, Timely State Updates and Additional Tuning**

Distance-Vector Multicast Routing Protocol

Flood and Prune

- **Reverse Path Broadcasting (initially)**
  - discard packets, if not arrived along shortest path
- **Reverse Path Multicasting (`mrouted`)**
  - per-source routing, recursive pruning

Link Configuration

- **TTL threshold (decision about packet forwarding)**
- **TTL metric (governs TTL decrement)**

RIP-like Unicast Routing ("reverse")

- **routers keep state (distance) per source per previous router**
  - routing information is periodically exchanged with neighbours
- **R1 keeps state per source whether being on the shortest path to R2**
  - if not → don’t forward packets to R2 (selective forwarding → less flooding)
- **multiple routers on LAN → shortest path to source is DOMINANT**
  - others are **SUBORDINATE**
- **equal distance → lowest IP address becomes dominant**

⇒ **Multicast routing can be decoupled from unicast routing**

Distance-Vector Multicast Routing Protocol

Routing State

Source Prefix	Subnet Mask	From Gateway	Metric	Status	TTL
128.1.0.0	255.255.0.0	128.7.5.2	3	up	200
128.2.0.0	255.255.0.0	128.7.5.2	5	up	150
128.3.0.0	255.255.0.0	128.6.3.1	2	up	150
128.3.0.0	255.255.0.0	128.6.3.2	4	up	200

- **TTL: Validity Time of Routing Entry (NOT Packet TTL)**
- **Metric: Unicast distance (in hops)**

⇒ **Complexity: linear in number of sources**

Distance-Vector Multicast Routing Protocol

Forwarding State

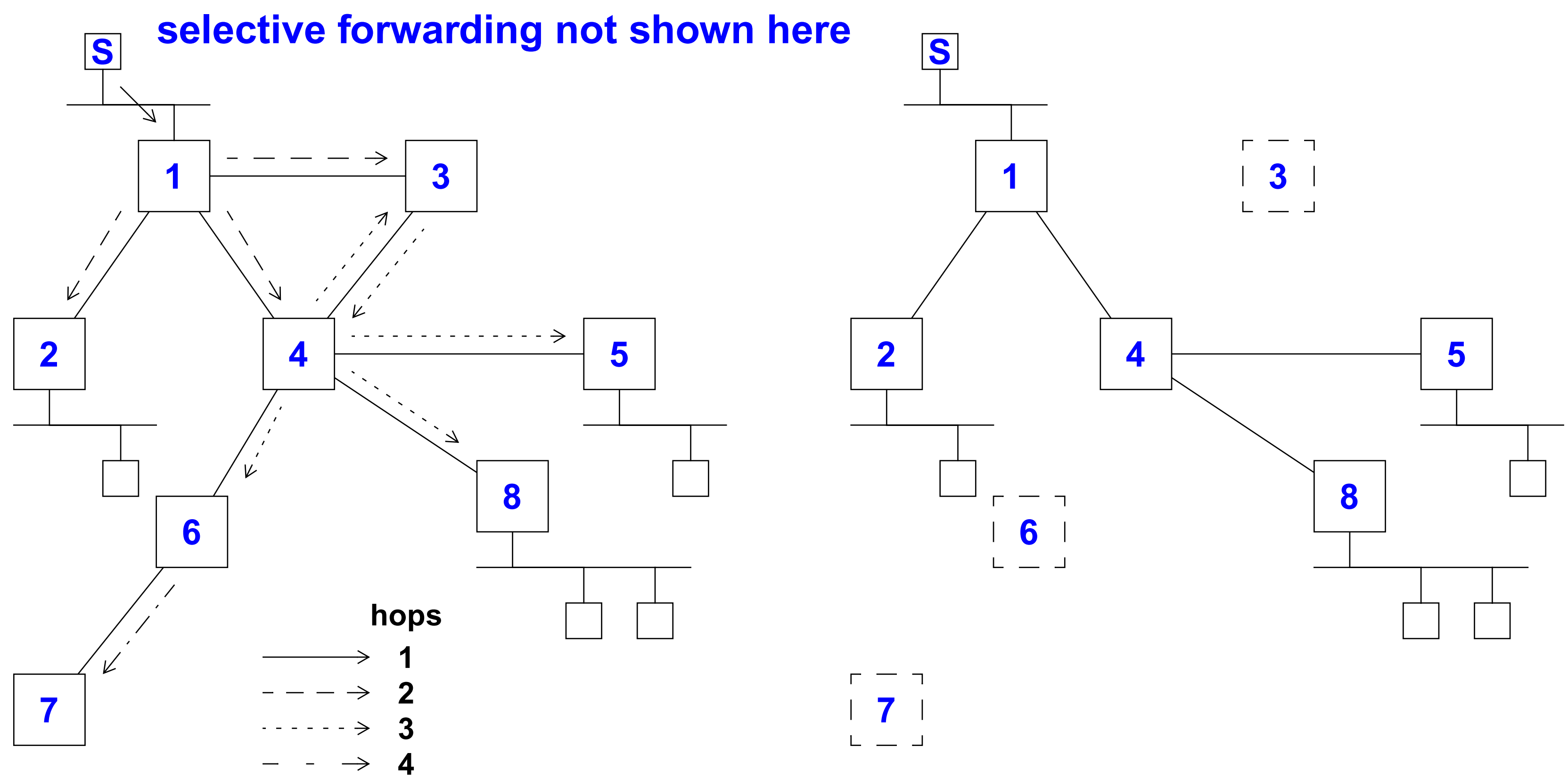
Source Prefix	Group	TTL	Inc. Interface	Out Interface
128.1.0.0	224.1.1.1	200	1 Pr	2p,3p
	224.2.2.2	100	1	2p,3
	224.3.3.3	250	1	2
128.2.0.0	224.1.1.1	150	2	2p,3

- **TTL: Validity Time of Routing Entry (NOT Packet TTL)**
- **p: prune received**
- **Pr: prune sent**

$\Rightarrow$ **Complexity: n * m**

- **n: (average) number of sources**
- **m: number of groups**

Distance-Vector Multicast Routing Protocol



Usage of IGMP Messages for DVMRP Messages

- prune (unreliable), graft (reliable)
- routing table updates similar to RIP

Further Info

- see RFC 1075
- see draft-ietf-idmr-dvmrp-v3-11

Protocol Independent Multicast

MOSPF: Depends on OSPF

DVMRP: Dedicated Unicast Routing Protocol

Protocol Independent

- **utilize "least common denominator" of unicast routing**
- **$\rightarrow$ unicast routing table**
- **$\Rightarrow$ multicast routing must be co-located with unicast routing**
- **inhibits some optimizations**

Variants

- **Dense Mode: based on flood and prune**
- **Sparse Mode: based on shared trees**

Interoperability

- **dense mode: flood & prune $\rightarrow$ no 'join' message**
  - 'graft' only cancels earlier prune, but tree already exists
- **create 'join' at dense mode border router towards sparse-mode region**

PIM - Dense Mode

Similar to DVMRP: Flood and Prune

- **Reverse Path Multicasting**

No Separate Routing Information Exchange

- **no selective forwarding**
- **always flood packets over all interfaces (except incoming)**
    - subsequent pruning
    - more flooding than DVMRP → increased traffic load

PIM - Sparse Mode

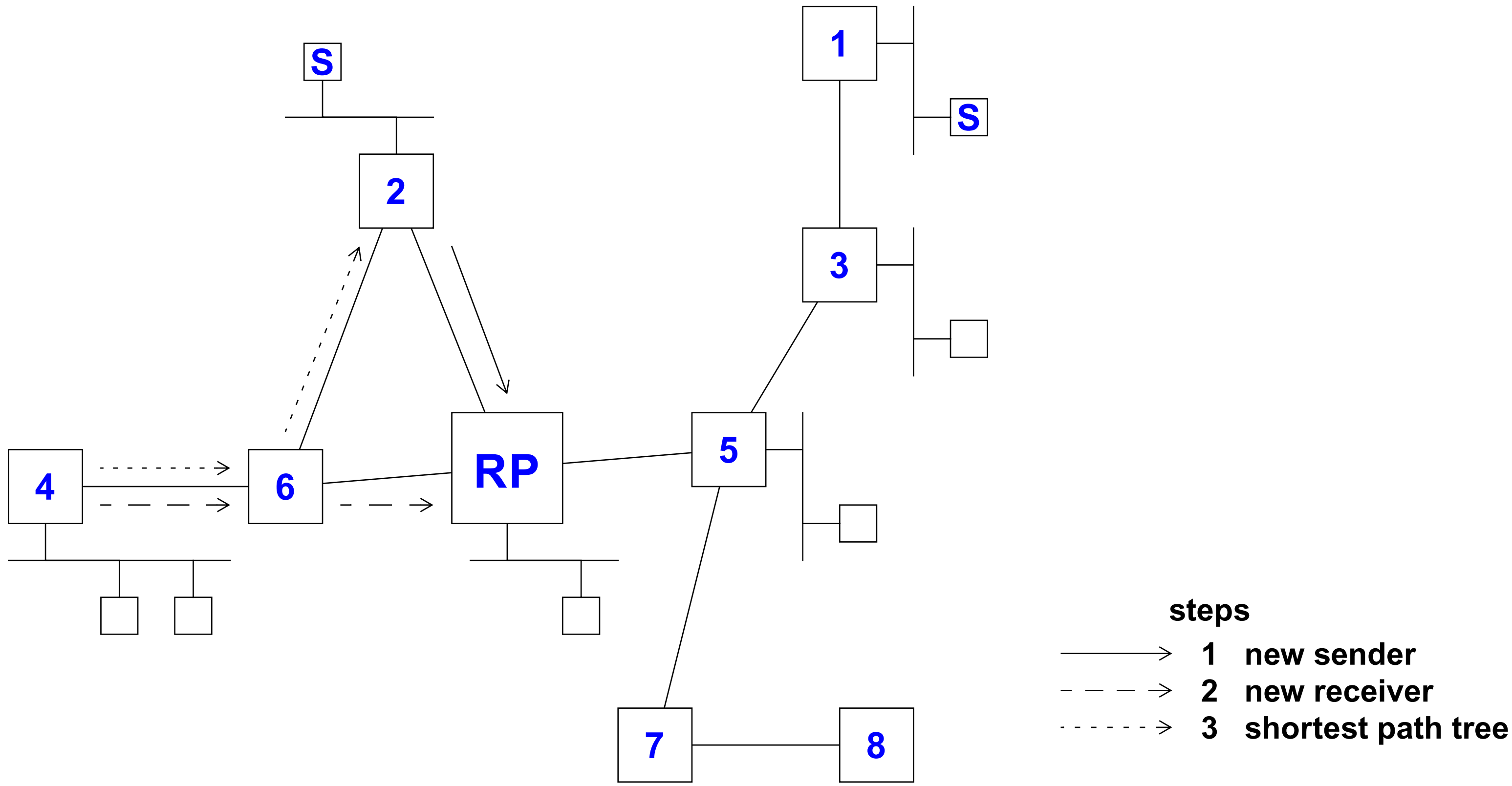
Shared Tree

- **central point: RENDEZVOUS POINT (RP)**
- **distributed construction of shared trees**
  - each router maintains list of RPs
    - hash-based mapping: group $\rightarrow$ RP
  - receiver: join request is sent to RP
  - intermediate nodes (RP $\rightarrow$ receiver) create (*,group) forwarding state
    - check join with unicast routing information
  - sender: encapsulate first data packet in control message (SM register)
  - RP responds with join to source
  - intermediate nodes (source $\rightarrow$ RP) create (source, group) forwarding state
    - check join with unicast routing information
- **not shortest path**

Source-based Shortest Path Tree

- **can be requested by receiver**
- **can be initiated by RP**
- **corresponding prunes in shared tree**

PIM - Sparse Mode



Further Info

- **see RFC 2362**

Emerging Approaches

Source Discovery (MSDP)

- **find path to source**
- **connect shared-trees across multiple domains**
- **use information to optimize multicast tree**
- **http://www.ietf.org/html.charters/OLD/msdp-charter.html**

Source-Specific Multicast

- **http://www.ietf.org/html.charters/ssm-charter.html**
- **receiver must know source address**
  - dedicated address space: 232.0.0.0/8
  - rules for allocating addresses
  - URD: URL-based rendezvous protocol for unaware receivers

Border Gateway Multicast Protocol (BGMP)

- **inter-domain multicast routing**
- **http://www.ietf.org/html.charters/bgmp-charter.html**

Multicast & Naming

Naming and Address Allocation

- **no natural hierarchy as in IP addresses**
  - flat address space with some restrictions
- **no controlled address range allocation**

Some Global Administration

- **different address ranges used for different distribution range**

Dynamic Session Directory

- **group announcements are multicast (broadcast) in special group**
- **soft-state $\rightarrow$ announcement expires if not refreshed**
- **advance announcements**
- **scope of announcement can be limited by TTL**

$\Rightarrow$ **Collisions possible and require manual intervention.**

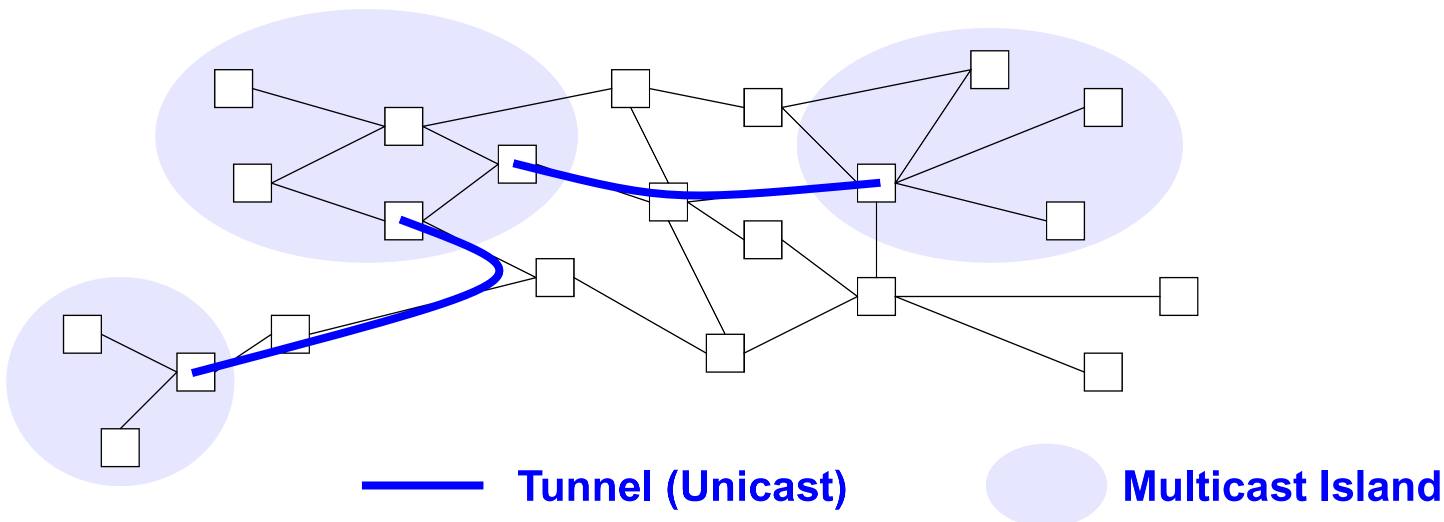
MBone

Experimental Overlay Network

- **connecting multicast-capable islands**
- **edge-to-edge tunneling**
  - routing protocol messages
  - data packets

Global Multicast Testbed

- **multicast transport protocols**
- **multicast applications**



IP Multicast - Evaluation

Or: Why IP Multicast "failed"...

Technical Problems

- integration with unicast IP → no flexible design
- IPv4 - limited address range: ~ 1 Mio group addresses
- uncontrolled address allocation
  - hacks for good utilization of address range: address designation, TTL

General Problems

- most interesting applications: games, multimedia
- no guaranteed transmission quality
- ⇒ little demand
- significant deployment cost for providers

⇒ Failure or Postponement?

- IPv6 removes address limitations
- multicast overlay networks

Multicast Overlays

Separation of Multicast from Basic IP Forwarding

- **"overlay" in terms of network structure → already in IP multicast**
  - not every node is a multicast router
  - unicast tunneling integral part of multicast approach
- **"overlay" in terms of edge vs. core**
  - main cost metric: access bandwidth
- **"overlay" in terms of protocol layering → on top of IP**
- **"overlay" in terms of implementation layering**
  - user-level process, vs.
  - low-level - kernel or hardware implementation
  - modularity
- **distinct & orthogonal design concepts, same terminology**

Trade-Offs

- **network structure → deployment vs. path selection efficiency**
- **edge vs. core → deployment with network efficiency**
- **protocol layering → implementation/deployment vs. protocol overhead**
- **implementation layering → flexibility vs. execution cost**

Routing in Edge-System Overlays

Parameters

- **degree of vertices in routing graph**
  - packet replication overhead
  - acess link bandwidth requirements
- **diameter of routing graph**
  - transmission delay

Suggested Algorithms

- **fix degree, minimize diameter**
  - max workload at node, find best worst-case delay
- **fix diameter, balance degree**
  - max worst-case delay, find best workload distribution
- **NP-hard / NP-complete problems → heuristic algorithms needed**

Comparison with IP Multicast

- **IP Multicast: find spanning tree with shortest paths to receivers**
- **IP Multicast: source-specific routing**
- **replication workload and replication efficiency → lesser concern**
  - possible with link-state protocols, but only with high computation complexity

Example: Balanced Degree Allocation

Goal: Balance Degree subject to Maximum Diameter

- $d_{max}(v)$: replication capacity (configuration parameter)
- Note: number of nodes → number of edges → fixed sum of degrees

1. Degree Allocation

- increase degree of node with most available capacity
- balance remaining capacity at nodes (residual degree)

2. Find Edges, subject to Degree Allocation

- try to satisfy diameter condition
- several algorithms possible, no perfect choice

3. Restart at 2.

- if diameter constraint is not met
- relax degree allocation

Explain Example in Paper!

Discussion

Routing in Overlay Networks

- relation to network structure and underlying routing
- distributed route computation
- large-scale groups

Mobile and Multicast Communication

- differences
- commonalities