6.896 Lecture 9: Landmark and Multicast Routing (Part 1)

Hari Balakrishnan

October 7, 1998

1 Outline

- **Readings: Tsu88, DC90, BFC93**
- Landmark hierarchy
- IP multicast: overview, routing alternatives, membership (IGMP)
- CBT and PIM (next class)

2 Interdomain unicast routing

- Key problem: *scalability.* All current solutions use hierarchies in some form.
- EGP (last class) and BGP.
- Area hierarchy.
- Landmark hierarchy (new routing abstraction).

2.1 BGP

- BGP is the current exterior gateway protocol that overcomes many of EGP's drawbacks.
- Key concept: *path vectors*. Vector of ASs sent via routing message. Instead of just metric, actually send a complete vector of traversed ASs. Q: what's the advantage? Disadvantage?
- Q: What do routers have to store now? Well, have to store all reachable networks (N), one entry for each. Total space needed is number of paths + number of networks. Why? On the order of several hundred KBytes (and increasing).

- *Path attributes* modified as routing message traverses ASs. Transitive attributes (e.g., bandwidth of slowest hop) and non-transitive attributes (e.g., local preference with only internal meaning).
- Need internal as well as external peers. Tightly couple all exterior gateways within an AS to provide the right peering information.
- Uses TCP for reliable dissemination. Has pros and cons. Ask questions about pros (simplifies BGP error management) and cons (lose partial information, lose responsiveness during congestion). Latter con can be fixed with packet priorities. (Research opportunity here: try and use some form of reliable multicast rather than lots of TCPs to use bandwidth better and more importantly, to fix first con).
- BGP allows gateways to make routing decisions on the basis of policies that can be checked locally. Detecting conflicts and contradictions is hard. (Research opportunities here; some work by Sharma and Keshav at Cornell).
- A final note: BGP allowed us to move from a single tree-like backbone to a general graph as the backbone. CIDR and BGP mods in version 4 (will) allow us to temporarily tide over address exhaustion and exploding routing table sizes, until IPv6 gets here.

2.2 Area hierarchy

- Generalized hierarchical routing based on topological hierarchy.
- Key idea: Once a message for Level $k-1$ enters destination's Level k domain, it will not leave that domain.

2.3 Landmark hierarchy

- A hierarchical unicast routing protocol for dynamic topologies.
- Key idea: send packets towards a landmark in next hop domain.
- Visibility of routers diminishes as distance to them increases.
- Main controbution: A new way of thinking, a new abstraction.
- Landmark: a router, all of whose neighbors within r hops know where the router is and can get to it. (Reverse not necessarily true—different from area hierarchy).
- Define $LM_i[id]$ as landmark at level i with ID id of radius $r_i[id]$, and $0 \leq i \leq H$.
- In routing table, keep route to every landmark below your own radius.

- Addressing by concatenating landmarks. $LM_n[id_n].LM_{n-1}[id_{n-1}] \ldots LM_0[id_0]$. Note that a router can have many such addresses.
- Issues
  - Assignments of IDs and landmark election.
  - Distance-vector routing. Q: why won't link-state work?
  - How to bind permanent names to dynamic landmark IDs?

3 IP multicast

3.1 Overview

- Goal: efficient one-to-many and many-to many communication. There are three types of addresses in IP: unicast, broadcast and multicast.
- Why is it useful?
  - Reducing network and server load. E.g., stock quote dissemination, "push" applications on the Web, etc.
  - Real-time multi-party conferencing, distance learning.
  - Internet-based "cable television".
  - Resource discovery.
  - Distributed interactive simulations.
  - And many others.
- IP multicast service model
  - Best-effort (unreliable) delivery of datagram to one or more members in multicast group. No ordering or atomicity guarantees.
  - Abstraction uses a level of indirection provided by a multicast group. In IP, address range 224.0.0.0 to 239.255.255.255 (Class D).
  - Open group membership; sender unaware of receiver identities; receiver-based subscription model.
  - Two types of protocols needed: *group membership* (IGMP) and *routing* (DVMRP, MOSPF, CBT, PIM, ...).
  - Scope control achieved using hop counts (ttl).

- **The MBone:** interconnected set of subnetworks and routers that support the delivery ofIP multicast traffic. Several thousand subnets today. Virtual network laid on top of the physical Internet, using *tunnels* via IP-in-IP encapsulation. Of course, there are islands of completely multicast connectivity too. DRAW PICTURE!
- Multicast routers talk a routing protocolto efficiently route packets.
- At the leaves, last-level (leaf) routers and hosts handle membership using a group membership protocol (IGMP).

3.2 Routing options

- **Reverse Path Flooding (RPF):** Simplest technique.
  1. Packet arrives at router.
  2. First time? If so, forward on all interfaces but incoming. Else, discard.
  3. Simple.
  4. Q: drawbacks? Duplicates. Router buffer space.
- **Spanning trees:** Select a subset of the Internet topology to form a spanning tree.
  1. Packet arrives at router.
  2. For each outgoing interface that's on the spanning tree, forward packet. Don't do so on other interfaces.
  3. Simple.
  4. Q: drawbacks? Traffic concentration. Sub-optimal paths to destinations.
- **Reverse path Broadcasting (RPB):** Idea—build group-specific spanning trees for each source subnetwork. Different tree for each source-group pair.
  1. Packet arrives at router.
  2. If packet arrives from router that is on shortest path to source, then forward on all other interfaces.
  3. Else, discard packet.
  4. Optimization: forward only to children downstream for whom you're on the shortest path to sender. Q: how do you know this?
  5. Show example using picture. Avoids duplicates.
  6. If two routers are equidistant to source, who "wins?"

7. Advantages: No special packet list maintanence like in flooding. Low latency (shortest path) paths.
8. Q: drawbacks? does not take group information while building (source,group) trees.

- Truncated Reverse Path Broadcasting (TRPB): improvement to RPF. At the leaf level, check if there's any interest in group membership. This is done using IGMP. If no interest, *truncate* delivery tree.

  – How does IGMP work? Periodic messages from mrouter *querying* LAN for membership in groups. Soft-state with periodic refreshes.
  – A host interested in a group replies to upstream querying router.
  – Problem: scaling due to implosion.
  – Solution: randomization + multicast damping. Powerful idea.
  – Receiver-initiated join for lower latency (don't have to wait for periodic query to arrive, be proactive while joining).
  – IGMP v1: mrouter queries 224.0.0.1 with TTL 1. Hosts reply (as above) for all groups they subscribe to. After threshold number of queries with no replies, remove group from query list.
  – IGMP v2: Handle multiple potential queriers using address-based tie-break. Adds a *group-specific* query message. Define a *leave group* mesage to lower leave latency. Such a leave forces the corresponding query. Currently deployed in most mrouters.
  – IGMP v3: Support to receive traffic only from specific sources in the multicast group. Can explicitly identify "inclusion" and "exclusion" sources. Also enhance group-leave to group-source-leave.

- **Reverse Path Multicasting (RPM):** Enhances TRPF to create a delivery tree that spans only subnetworks with group members, such that it uses only SPTs to source.

  1. Use TRPB to construct source-specific trees to *leaf routers.*
  2. From leaf routers, use Truncated RPB. If no interest, propagate *prune* message up the tree, one hop at a time.
  3. Age prunes and time them out: on timeout, drop prune and reinstantiate flow. Downstream routers reprune. Uses *soft-state*.
  4. Q: what do you end up with? A tree with only *live* branches.
  5. Q: drawbacks? Scaling. State at edge routers. (source,group) state in routers. First packet flooded everywhere. Occasional (and periodic) flooding occurs because of soft-state model.
  6. How to pick prune timers? If too long, then long join time. If too short, then frequent flooding.
  7. Solution: explicit prune-cancel messages, or *grafts.*

8. Grafts need to be reliable. Why?

- **DVMRP:** Periodic exchange of routing information with neighboring mrouters. Info includes (source-subnet from-gateway metric inport outports). Forwarding table consists of source-subnet groups inport outports and prune information.
- Hierarchical DVMRP for better scaling.
- **MOSPF:** OSPF extensions for multicast.