6.896 Lecture 9: Unicast Routing

Hari Balakrishnan

October 7, 1998

1 Outline

- **Readings: Nar89 (historical overview), Tsu88**
- Distance-vector and RIP
- Link-state and OSPF
- Administrivia
- Exterior gateway protocols: EGP and BGP
- Dynamic topologies: Landmark hierarchical routing

2 Distance-vector protocols and RIP

Q: Difference between routing and forwarding?

2.1 Theory (with some practical elements)

- Q: How does the Bellman-Ford algorithm work? Essentially does dynamic programming: $d_{ij} = \min_k(d_{ik} + d_{kj})$.
- Distributed version
  - Advertise distance vectors to all your neighbors: [destination metric link].
  - Pick best (shortest) path to a given destination and incorporate into your routing table.
  - In the absence of failures and partitions, converges and is loopless.

- But what happens if a link fails? In many cases, this is figured out OK and things heal, converging to the new routing state.
- But not always. Might end up with the *bouncing effect*. If the "other end" of the broken link doesn't get to send its routing update first, you end up with a ping-pong effect, if an alternate path has a much higher starting metric. (Draw picture of this).
- Q: How does this get resolved? TTLs kill packet eventually. And routing eventually heals itself after a while.
- And it gets worse. A network partition could cause a *counting to infinity* problem.
- How does this get resolved? Well, you need to set "infinity" to a small positive number! Like 16.
- Two solutions that alleviate these problems: *split horizons* and *triggered updates*.
- Split horizons: If you're routing to a destination on a certain link, don't advertise a route to it on that link. And you can really ensure that no two-hop loops arise using *poisonous reverse:* advertise a route on that link but set its metric to infinity.
- But three or higher-hop loops aren't killed at all. Show example.
- Triggered updates: in addition to periodic transmissions of routes, send updates immediately upon changes.

2.2 Practice: RIP

- RIP is simple: only two types of messages.
- A distance-vector IGP that uses both these schemes (UNIX *routed* program). In addition, age routes not updated in 180s (6 x 30s).
- Upon startup, download neighbors' tables using a request message.
- Works OK for small networks, RIPv2 understands subnetting among other enhancements.
- Later in the term, we will look at how synchronization occurs in this, affecting TCP. Lesson: need explicit randomization in announcement protocols.

3 Link-state and OSPF

- What's wrong with RIP?

  1. Computing new routes is slow (counting to infinity).
  2. Long-lasting "transients" exacerbate congestion.
  3. Can't handle more than small-sized networks.

- Link-state protocols try to overcome these problems.
- Observation: problems because of incomplete information and distributed route computation.
- So, why don't we disseminate entire topology map and let individual nodes compute routes?
- Two issues: how to disseminate map (hard) and how to compute routes at individual nodes (not-so-hard).
- *Dissemination:* representing database (adjacency list), flooding map information, bringing up adjacencies (i.e, resolving databases when a partition heals; need database versions to do this), protecting against corruption.
- *Route computation:* Use Dijkstra's shortest path algorithm. Q: ask class to explain this.
- So why is it called OSPF?
- It's more complicated, so why is it better?

  – Fast, loopless convergence.
  – Supports multiple metrics.
  – Supports dispersity (multiple paths).

3.1 Other IGPs

- Many other IGPs exist (or existed in the past)
- Examples include the Arpanet GGP, Hello protocol, Dual IS-IS, and Cisco's proprietary (E)IGRP.
- Also look at Garcia-Luna et al.'s work on loop-free distance-vector algorithms (some are in EIGRP).

4 Administrivia

- Project proposals due today by 5.
- Try to be as concrete and crisp, rather than identify a huge research area and say you will explore it! I am not as interested in design papers with no substantiation. Substantiation of any design (if you're doing a design paper) should be done via simulation, measurements or analytical proofs.
- Will probably end up sliding landmark to next week and catch up slack during forwarding/router design lecture.
- Useful project tools: tcpdump, traceroute, mtrace, ns. How many want a (meta-)tutorial on ns?

5 Exterior gateway protocols: EGP and BGP

5.1 What is an exterior gateway protocol?

- Problem: network grew too big in the mid-80s.
- Needed a hierarchical organization: achieved using *autonomous systems (AS)*.
- Q: What is an AS? ("A set of routers and networks under the same administration"). Each AS has a unique 16-bit AS number.
- Inside an AS, run any IGP (RIP , OSPF, ...).
- Exterior gateway protocols allow ASs to exchange routing (in particular, reachability) information.
- Yesterday, EGP; today, BGP (now in version 4).

5.2 EGP

- Draw picture of two ASs and exterior gateways with inter-AS link.
- EGP exchanges *reachability* information between ASs.
- Three procedures:
  1. Neighbor acquisition.
  2. Neighbor reachability.
  3. Network reachability.

- Neighbor acquisition: Agree to be peers and route for each other. Needs a formal agreement / contract and explicit configuration. Uses a 2-way handshake to achieve this. Neighbor cease with ACK to terminate peering.
- Neighbor reachability: more than just link connectivity. Use "hello—I-H-U" protocol with different response codes.
- Network reachability: Exchange list of interior reachable networks with peer. Poll interval decides frequency of this exchange.
- Some issues:
  - Backdoor routes between ASs. Need to be careful what you advertise to peers.
  - Distance advertisement: What metric should gateways advertise to peers? If AS uses OSPF, EGP announces external link-state records ("E" bit in OSPF records). What if AS uses RIP? Makes it nearly impossible to choose between routes; you end up with pretty much only one option.
  - False information problem: a "bad" exterior gateway could wreck system.
  - Hard to do policy routing: choosing between ISPs or ASs to route through.
  - Loops: doesn't actually avoid loops.
  - Long messages when there are a large number of networks.

5.3 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.

6 Landmark hierarchy