EECS 122, Lecture 16

Kevin Fall

kfall@cs.berkeley.edu

Link Costs and Metrics

- Routing protocols compute shortest/cheapest paths using some optimization criteria
- Choice of criteria has strong effect on path selection results
- Metrics are either static or dynamic

Static Cost Metrics

- *hop-count* (easy to compute)
  - reasonable for homogeneous links
  - treats DS3 (45Mb/s) & dialup (56Kb/s) equal
- manually-assigned scalars: administrator can "tweak" the metrics, but still doesn't adapt to congestion, and doesn't scale well [difficult to manage]

Traffic-Sensitive Metrics

- Original ARPAnet scheme:
  - cost proportional to queue on outgoing link
  - problems: at higher loads, fine grain measuring of queue lengths during traffic spikes could trigger frequent re-routing
  - high-cost links never used
  - high cost used to predict future high cost (but not true once traffic re-routed!)
  - no damping on changes in cost -> oscillation

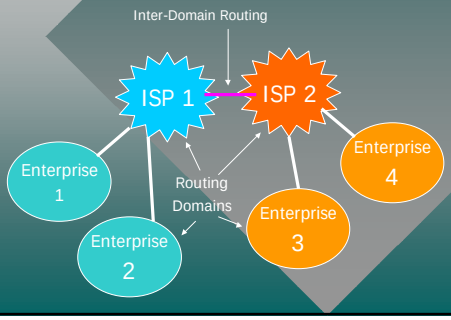
Traffic-Sensitive Metrics

- Modified ARPAnet scheme:
  - metrics f(link capacity, output queue length)
  - capacity dominates at low load
  - link costs normalized to hop-count units
  - dynamic range of link costs limited (from 127:1 to 3:1)
  - only delta of 1/2-count on each change
  - nearly eliminated oscillations, even at load

Routing Hierarchy

- Routing overhead scales with the number of nodes
  - for N nodes
    - shortest path is about O(N log N) algorithm
    - routing table size is O(N)
- So to scale, routing is built into a hierarchy [general principal]

Hierarchical Internet



Hierarchical Internet Routing

- Intra-Domain Routing Protocols
  - IGPs (Interior Gateway Protocols)
  - examples: OSPF, RIP, IGRP, EIGRP, IS-IS
- Inter-Domain Routing Protocols
  - EGPs (Exterior Gateway Protocols)
  - examples: BGP, EGP (old)

Intra- and Inter- Domain Routing Protocols

- Notion of routing "domain" or "area"; also called *Autonomous System* (AS)
- Routing computation takes place within AS, and is typically summarized at edges
- IGP selection up to administrators of each enterprise; EGP usually standardized

Routing Information Protocol (RIP1 and RIP2)

- distance-vector protocol using hop-count as metric
- infinity value is 16 hops
- announcers broadcast (or multicast) DVs every 30 seconds; time out in 180 sec
- split horizon with poisoned reverse
- RFC 1058 (RIPv1); RFC 1388 (RIPv2)

RIP History

- DV schemes were used early on in the ARPAnet (1969) and Cyclades (70's)
- RIP used within Xerox on PUP/XNS
- Internet RIP based on XNS-RIP, made available in Berkeley UNIX as "routed"

RIP Version 1

- Routing destinations are 32-bit hosts, networks, or subnets
- Routers first look for classes A, B, or C
- If subnet+host part is NULL, represents a network route, otherwise a subnet or host route
- Uses a (static) subnet mask applied to all entries

RIP Version 1 Support

- supports point-to-point links and multi-access LANs (e.g. Ethernet)
- RIP packets encapsulated in single UDP packets (not reliable, up to 512 bytes)
- DV tables sent using broadcast every 30s (or more often for triggered updates)
- updates time-out after 180s if not refreshed

RIP Routing Table Structure

- Included in RIP-maintained routing table:
  - address of (net/subnet/host) destination
  - metric associated with destination
  - address of next hop router
  - recently-updated flag
  - several timers

RIP, Version 2

- compatible upgrade to RIP v1 including subnet routing, authentication, CIDR aggregation, route tags and multicast transmission
- RFC 2453 includes background and protocol definition [std rfc]

Subnet Support

- RIP v1 supports subnet routes only within the subnetted network (using single subnet mask)
- by including subnet mask with routes, allows for subnet knowledge outside subnet
- more convenient partitioning using variable-length subnets

Authentication

- RIPv1 is completely not secure; anyone can act as a router just by sending RIP1 messages (if cost zero, everyone uses!)
- RIPv2 supports generic notion of authentication, but only "password" is defined so far (not very secure)
- At least prevents accidents reasonably well

Route Tag and Next Hop ID

- Routing domain "tag" is available in each message to distinguish multiple domains running on same wire/subnet
- In addition, on multi-domain networks, the border router may specify an alternative next hop (not itself) if there is better nearby router in its domain to reach a particular destination

OSPF - Open Shortest Path

- Link-state protocol specified by IETF
- Special features supported:
  - separation of hosts and routers
  - multi-access LANs
  - non-broadcast networks
  - hierarchies ("areas")
  - multi-path (equal cost)

Broadcast Networks

- With broadcast, multi-access networks, N(N-1)/2 adjacencies would be used
- Avoid this by electing "designated router"
- Broadcast network then represented as a virtual node in routing computation

Non-Broadcast Networks

- How to perform flooding without broadcast or multicast support?
- Places burden for endpoint distribution on DR which re-sends to interested parties using unicast

Multiple Areas

- For large intranets, routing overhead can be undesirable; usually use hierarchy
- OSPF provides its own: *areas*
  - "top" area called 'backbone'
  - computation spans areas
  - area-border routers span multiple areas

External Routes

- OSPF provides the ability to import routes from other routing protocols
- Particularly useful where a router is both an IGP and EGP participant
- "Stub Areas" support the suppression of generalized external routes in favor of default (tables don't scale as size of Internet)

Protocols within OSPF

- OSPF Communication directly on IP
- All packets contain version, packet type, length, router ID, area ID, checksum and authentication data
- Really three protocols: hello, exchange, and flood

The Hello Protocol

- Periodically, routers send hello messages including their priority, the current designated router and backup designated router, neighbors they have heard from
- priority affects (B)DR election on broadcast and non-broadcast networks (BDR is used for quick failover)
- only 2-way operational links ok

DR and BDR Election

- election process runs continually with exchange of HELLO messages
- on any change, election process ensures convergence on new DR and BDR
- may have to change adjacencies to BDR (already computed) and begin computing for a new BDR

Exchange Protocol

- Used for initial synchronization of LS database entries, and after partitions heal
- Exchange "database description packets" containing ID, advertising router, sequence number, checksum, and age
- Acks for sequence number generated; simple retransmission used

Flooding Protocol

- Link state updates contain advertising router's ID, link state ID/type, and lollipop sequence space number
- ACKs to sending router, and continues flooding if sequence number is newer
- Retransmitted by sender until acknowledged

IGRP - cisco routing protocol

- response to need for routing protocol superior to RIP prior to IETF OSPF standardization
- DV scheme with special features:
  - composite metrics
  - specialized loop detection
  - multipath routing
  - handling of default routes

IGRP Composite Metrics

- (D)elay, (B)andwidth, (R)eliability, (L)oad
  - also includes (H)op-count and path MTU (these are not used in routing computation)
- Delay: path length delay to each dest
- Bandwidth: min across links to each dest
- Reliability: measured (loss prob) [1..255]
- Load: measured (loading) [1..255]

Optimization Metric

- Start with observation that time to send is:

$$T = \frac{PktSize}{Bandwidth} + Delay = \frac{P}{B} + D$$

- But avail bandwidth affected by load:

$$T = \frac{PktSize}{Bandwidth(frac)} + Delay = \frac{P}{B[(255 - L)/255]} + D$$

Optimization Metric

- But with unreliable links, may require re-sends, so multiply this by a ratio expressing reliability:

$$T = \left[\frac{PktSize}{Bandwidth(frac)} + Delay\right](frac) = \left[\frac{P}{B[(255 - L)/255]} + D\right](255/R)$$

Observations

- Measuring load should not be over very small time interval, or instabilities may arise
- In practice, several constants may be altered by the administrator to affect the relative weighting given to each component

Specialized Loop Detection

- uses split horizon and triggered updates, but not poisoned reverse
- extended with holddown (older) and route poisoning (newer)
- given triggered updates, loops normally form only due to transmission errors or slow update propagation

Holddown

- upon detection of a link failure, initiate "quarantine" period during which no updates for destination are accepted
- after at least 2 periods (180 s in IGRP), quarantine is removed and normal route selection resumes
- works, but guarantees that destination is unreachable even if other paths exist

Route Poisoning

- newer versions of IGRP replace holddown with *route poisoning*
- observe increasing cost to a destination and assume a loop has formed
- only after re-confirmation of metric is path assumed to be usable (will happen up to 1 reporting period later--90secs for IGPT, much better than 3 minutes)

EIGRP: Extended IGRP (cisco)

- none of the standard DV fixup schemes are entirely satisfactory
- use DUAL ("diffusion update algorithm") to remove transients [applies to DV & LS]
- enhance IGRP with DUAL [using incremental updates], variable length prefix masks and aggregation, and route tags

Next Time…

- finish up EIGRP with discussion of DUAL algorithm
- exterior routing protocols (BGP), including CIDR aggregation