Interdomain Routing
EECS 489 Computer Networks
http://www.eecs.umich.edu/~zmao/eecs489
Z. Morley Mao
Tuesday Sept 21, 2004

Acknowledgement: Some slides adapted from Kurose&Ross and Katz&Surok

Distance Vector: Link Cost Changes

7 loop:
8 wait (until V sees a link cost change to neighbor V)
9 or until V receives update from neighbor V
10 if D(A, V) changes by d
11 for all destinations Y through V do
12 D(A, Y) = D(A, Y) + d
13 else if update D(V, Y) received from V
15 if there is a new minimum for destination Y
16 send D(A, Y) to all neighbors
17 forever

Algorithm terminates

Distance Vector: Count to Infinity

7 loop:
8 wait (until A sees a link cost change to neighbor V)
9 or until A receives update from neighbor V
10 if D(A, V) changes by d
11 for all destinations Y through V do
12 D(A, Y) = D(A, Y) + d
13 else if update D(V, Y) received from V
15 if there is a new minimum for destination Y
16 send D(A, Y) to all neighbors
17 forever

…

Distance Vector: Poisoned Reverse

• If C routes through B to get to A:
  - C tells B its (C’s) distance to A is infinite
  - Will this completely solve count to infinity problem?

Link State vs. Distance Vector

Per-node message complexity
- LS: O(e) messages
- D: number of edges
- DV: O(d) messages, many times
- d is node’s degree

Complexity/Convergence
- LS: O(n^2) computation
- DV: convergence time varies
- may be routing loops
- count-to-infinity problem

Robustness: what happens if router malfunctions?
- LS:
  - node can advertise incorrect link cost
  - each node computes only its own table
- DV:
  - node can advertise incorrect path cost
  - each node’s table used by others; error propagate through network

Are We Done?

• We now know how to route scalably
• What more is there to do?
Issues We Haven’t Addressed

- Scaling
  - Addressing
  - Router table size

- Structure
  - Autonomy
  - Policy

Scaling

- Every router must be able to forward based on “any” destination IP address
  - Given address, it needs to know “next hop” (table)
- Naive: Have an entry for each address
  - There would be 10^8 entries!
- Better: Have an entry for a range of addresses
  - But can’t do this if addresses are assigned randomly!
- Addresses allocation is a big deal

Network Structure

The Internet contains a large number of diverse networks

Autonomous Systems (AS)

- Internet is not a single network!
- The Internet is a collection of networks, each controlled by different administrations
- An autonomous system (AS) is a network under a single administrative control

Implications

- ASs want to choose own local routing algorithm
  - AS takes care of getting packets to/from their own hosts
  - Interdomain routing and Intradomain routing
- ASs want to choose own nonlocal routing policy
  - Interdomain routing must accommodate this
  - BGP is the current interdomain routing protocol

Intradomain And Interdomain
Interconnection

- IP unifies network technologies
  - allows any network to communicate with another
- BGP unifies network organizations
  - sees them into a global Internet

OSPF (Open Shortest Path First)

- "open": publicly available
- Uses Link State algorithm
  - LS packet dissemination
  - Topology map at each node
  - Route computation using Dijkstra’s algorithm
- OSPF advertisement carries one entry per neighbor router
- Advertisements disseminated to entire AS (via flooding)
  - Carried in OSPF messages directly over IP (rather than TCP or UDP)

OSPF “advanced” features (not in RIP)

- Security: all OSPF messages authenticated (to prevent malicious intrusion)
- Multiple same-cost paths allowed (only one path in RIP)
- For each link, multiple cost metrics for different TOS (e.g., satellite link cost set “low” for best effort; high for real time)
- Integrated unicast- and multicast support:
  - Multicast OSPF (MOSPF) uses same topology data base as OSPF
  - Hierarchical OSPF in large domains.

Hierarchical OSPF

- Two-level hierarchy: local area, backbone.
  - Link-state advertisements only in area
  - each node has detailed area topology; only know direction (shortest path) to nets in other areas.
- Area border routers: “summarize” distances to nets in own area, advertise to other Area Border routers.
- Backbone routers: run OSPF routing limited to backbone.
- Boundary routers: connect to other AS’s.

Outline

- Addressing
- BGP
Assigning Addresses (Ideally)

- Host: gets IP address from its organization or ISP
- Organization: gets IP address block from ISP
- ISP: gets address block from routing registry:
  - ARIN: American Registry for Internet Numbers
  - RIPE: Réseaux IP Europeens
  - APNIC: Asia Pacific Network Information Center
  - Whois: h whois.arin.net <IP>
- Each AS is assigned a 16-bit number (65536 total)
  - Currently 10,000 AS’s in use
  - Most stub, so don’t really need own number

Original Addressing Scheme

- Class-based addressing schemes:
  - 32 bits divided into 2 parts:
    - Class A
    - Class B
    - Class C

Original Vision:
- Route on network number
- All nodes with same net # are directly connected

Classless Interdomain Routing (CIDR)

Introduced to solve two problems:
- exhaustion of IP address space
- size and growth rate of routing table

#1: Address Space Exhaustion

- Example: an organization needs 500 addresses.
  - A single class C address not enough (254 hosts).
  - Instead a class B address is allocated (~65K hosts)
  - That’s overkill, a huge waste!
- CIDR: networks assigned on arbitrary bit boundaries.
  - Requires explicit masks to be passed in routing protocols
  - Masks: identify the “network” portion of the address
- CIDR solution for example above: organization is allocated a single /23 address (equivalent of 2 class C’s).

CIDR Addressing

- Suppose fifty computers in a network are assigned IP addresses 128.23.9.0 - 128.23.9.49
  - They share the prefix 128.23.9
- Range: 01111111 00001111 00001001 00000000 to 01111111 00001111 00001010 00000000
  - How to write 01111111 00001111 00001001 00X?
- Convention: 128.23.9.0/26
  - There are 32 26-6 bits for the 50 computers
  - 2^6 = 64 addresses

More Formally

- Specify a range of addresses by a prefix: X/Y
  - The common prefix is the first Y bits of X
  - Y: 2^32 - addresses in the range
- Example 128.5.10/23
  - Common prefix is 23 bits:
    - 01000000 00000101 00000100 00000000
  - Number of addresses: 2^9 = 512
- Prefix aggregation
  - Combine two address ranges
    - 128.5.10/24 and 128.5.11/24 gives 128.5.10/23
  - Routers match to longest prefix
Problem #2: Routing Table Size

Without CIDR:

- 232.71.0.0
- 232.71.1.0
- 232.71.2.0
- 232.71.255.0

With CIDR:

- 232.71.0.0/16
- 232.71.1.0/16
- 232.71.2.0/16
- 232.71.255.0/16

Border Gateway Protocol

- ignore the details
- pay attention to the “why”

Internet inter-AS routing: BGP

- BGP (Border Gateway Protocol): the de facto standard
- BGP provides each AS a means to:
  1. Obtain subnet reachability information from neighboring ASes.
  2. Propagate the reachability information to all routers internal to the AS.
  3. Determine “good” routes to subnets based on reachability information and policy.
- Allows a subnet to advertise its existence to rest of the Internet: “I am Here”

Who speaks BGP?

- Two types of routers
  - Border router (Edge), Internal router (Core)

Purpose of BGP

- Share connectivity information across ASes

I-BGP and E-BGP
### Issues

- What basic routing algorithm should BGP use?
- How are the routes advertised?
- How are routing policies implemented?
  - policy routing: not always shortest path
- Peering, transit, and reachability

### Choice of Routing Algorithm

- Constraints:
  - scaling
  - autonomy (policy and privacy)
- Link-state?
  - requires sharing of complete network information
  - information exchanges don’t scale
  - can’t express policy
- Distance Vector?
  - scales and retains privacy
  - can’t implement policy
  - can’t avoid loops if shortest paths not taken

### Path Vector Protocol

- Distance vector algorithm with extra information
  - For each route, store the complete path (ASs)
  - No extra computation, just extra storage
- Advantages:
  - can make policy choices based on set of ASs in path
  - can easily avoid loops

### BGP Routing Table

```
BGP routing table

Network     Next Hop       Metric LocPrf Weight Path
* 10.0.0.0    4.0.6.142     1000     50      0 701 80 i
* 1.0.0.0     4.24.1.35      0  100      0 i
* 12.3.21.0/23 192.205.32.153 0     50      0 7018 4264 6468 ?
  128.32.0.0/16 192.205.32.153 0     50      0 7018 4264 6468 25 e
```

- Every route advertisement contains the entire AS path
- Can implement policies for choosing best route
- Can detect loops at an AS level

### Advertising Routes

- One router can participate in many BGP sessions.
- Initially … node advertises ALL routes it wants neighbor to know (could be > 50K routes)
- Ongoing … only inform neighbor of changes
Why different Intra- and Inter-AS routing?

- Policy:
  - Inter-AS: admin wants control over how its traffic routed, who routes through its net.
  - Intra-AS: single admin, so no policy decisions needed
- Scale:
  - hierarchical routing saves table size, reduced update traffic
- Performance:
  - Intra-AS: can focus on performance
  - Inter-AS: policy may dominate over performance

Basic Messages in BGP

- Open:
  - Establishes BGP session (uses TCP port #179)
  - BGP uses TCP
- Notification:
  - Report unusual conditions
- Update:
  - Inform neighbor of new routes that become active
  - Inform neighbor of old routes that become inactive
- Keepalive:
  - Inform neighbor that connection is still viable

Routes Have Attributes

- When a route is “advertised” it is described in terms of attributes:
  - next hop, AS-path, etc.
  - We will discuss: Origin, MED, Local Preference
- Origin:
  - Who originated the announcement? Where was a prefix injected into BGP?
  - IGP, EGP or Incomplete (often used for static routes)

Multi-Exit Discriminator (MED)

- When AS’s interconnected via 2 or more links
- AS announcing prefix sets MED (AS2 in picture)
- AS receiving prefix uses MED to select link
- A way to specify how close a prefix is to the link it is announced on

Local Preference

- Used to indicate preference among multiple paths for the same prefix anywhere in the Internet.
- The higher the value the more preferred
- Exchanged between IBGP peers only. Local to the AS.
- Often used to select a specific exit point for a particular destination
Choosing Best Route

- Choose route with highest **LOCAL_PREF**
  - Preference-based routing
- Multiple choices: select route with shortest **hop-count**
- Multiple choices for same neighboring AS: choose path with min MED value
- Choose route based on lowest origin type
  - **IGP < EGP < INCOMPLETE**
- Among IGP paths, choose one with lowest cost
- Finally use router ID to break the tie.

Routing Process Overview

Import and Export Policies

- **Inbound** filtering controls **outbound** traffic
  - Filters route updates received from other peers
  - Filtering based on IP prefixes, AS_PATH, community
- **Outbound** filtering controls **inbound** traffic
  - Forwarding a route means others may choose to reach the prefix through you
  - Not forwarding a route means others must use another router to reach the prefix

Transit vs. Nontransit AS

- **Transit AS**: does carry transit traffic
  - Advertises its own routes PLUS routes learned from other AS’s
- **Nontransit AS**: does not carry transit traffic
  - Advertise own routes only
  - Do not propagate routes learned from other AS’s

AS Relationships and Export Rules

- **AS** has customers, providers and peers
- Relationships between AS pairs:
  - Customer-provider
  - Peer-to-peer
- Type of relationship influences policies

Customer-Transit Problem

- Assume that the small ISP is a customer of two large ISPs
- If customer ISP does not obey export rules
  - Forwards advertisements from one large ISP to another
- Carries huge volume of transit traffic between two large ISPs
Is Reachability Guaranteed?

- In normal routing, if graph is connected then reachability is assured
- With policy routing, not always

Peering & Transit

- Peering
  - The business relationship whereby ISPs reciprocally provide to each other connectivity to each other’s transit customers
- Transit (customer)
  - The business relationship whereby one ISP provides (usually sells) access to all destinations in its routing table

Peering

Figure from William B. Norton, “Internet Service Providers and Peering”

West and East Peer with USNet but they can’t reach each other

Transit

Figure from William B. Norton, “Internet Service Providers and Peering”

Reachability?

- Depends on series of economic relationships
- Not a technical solution, but an economic one

BGP and Performance

- BGP designed for policy not performance
  - Hot Potato routing common but suboptimal
  - 20% of internet paths inflated by at least 5 router hops
- Susceptible to router misconfiguration
  - Blackholes: announce a route you cannot reach
- Incompatible policies
  - Solutions to limit the set of allowable policies
Research Aside

- What is one of the most important Internet design principles?

- In what way is BGP robust?
  - Can it protect against node failures?
  - Can it protect against spoofing?
  - Can it protect against lying?

- Semantic failures versus syntactic failures
  - BGP can check who you are, but not what you say