Hierarchical Routing

- Is a natural way for routing to scale
  - Size
  - Network Administration
  - Governance
- Exploits address aggregation and allocation
- Allows multiple metrics at different levels of the hierarchy
Two ways to interconnect IP Networks…

- **Peering**
  - The business relationship whereby ISPs reciprocally provide to each other connectivity to each others’ transit customers

- **Transit**
  - The business relationship whereby one ISP provides (usually sells) access to all destinations in it’s routing table

William B. Norton, “Internet Service Providers and Peering”
Benefits of Transit v/s Peering

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

Moving from Transit to Peering

William B. Norton, “Internet Service Providers and Peering”
Interconnected ASes

- Forwarding table is configured by both intra- and inter-AS routing algorithm
  - Intra-AS sets entries for internal dests
  - Inter-AS & Intra-As sets entries for external dests

Inter-AS tasks

- Suppose router in AS1 receives datagram for which dest is outside AS1:
  - Router should forward packet toward one of the gateway routers, but which one?
- AS1 needs:
  1. to learn which dests are reachable through AS2 and through AS3
  2. to propagate this reachability info to all routers in AS1
- Q: With 200M hosts how is each host going to know which AS a host address belongs to?
- Job of inter-AS routing!
Addressing

- Every router must be able to forward based on *any* destination IP address
  - One strategy: Have a row for each address
    - There would be $10^8$ rows!
  - Better strategy: Have a row for a range of addresses
    - If addresses are assigned at random that wouldn’t work too well
      - MAC addresses
    - Addresses allocation is a big deal.

Class-base Addressing

- Addressing reflects internet hierarchy
  - 32 bits divided into 2 parts:
    - **Class A**
      - 0/8 network host
      - ~256 networks
      - 256 hosts
    - **Class B**
      - 0/16 network host
      - ~16K networks
      - 65,536 hosts
    - **Class C**
      - 0/24 network host
      - ~2 million networks
      - 256 hosts

~2 million nets
256 hosts
Class-based addresses did not scale well

Example: an organization initially needs 100 addresses
- Allocate it a class C address
- Organization grows to need 300 addresses
- Class B address is allocated. (~64K hosts)
- That's overkill -a huge waste
- Only about 8200 class B addresses!
- Artificial Address crises

---

IP addressing: CIDR

**CIDR: Classless InterDomain Routing**
- net portion of address of arbitrary length: subnet
- address format: a.b.c.d/x, where x is # bits in subnet portion of address

```
11001000  00010111  00010000  00000000
  subnet part    host part
```

```
11001000  00010111  00010000  00000000
  subnet part    host part
```

200.23.16.0/23
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
- Is this the **longest** prefix?
  - Range is 01111111 00001111 00001001 00000000 to 01111111 00001111 00001001 00110001
  - How to write 01111111 00001111 00001001 00X?
- **Convention:** 128.23.9/26
  - /26 is called the subnet mask
- There are 32-26=6 bits for the 50 computers
  - 2^6 = 64 addresses

Example 128.5.10/23
- **Common prefix is** 23 bits: 01000000 0000101 0000101
- **Number of addresses:** 2^9 = 512
- **Prefix aggregation**
  - Combine two address ranges
  - 128.5.10/24 and 128.5.11/24:
    - 01000000 00000101 00001010 01000000 00000101 00001010 01000000 00000101 00001011
  - gives 128.5.10/23
- Routers match to longest prefix
Assigning IP address (Ideally)

- A host gets its IP address from the IP address block of its organization
- An organization gets an IP address block from its ISP’s address block
- An ISP gets its address block from its own provider OR from one of the 3 routing registries:
  - ARIN: American Registry for Internet Numbers
  - RIPE: Reseaux IP Europeens
  - APNIC: Asia Pacific Network Information Center
- Each Autonomous System (AS) is assigned a 16-bit number (65536 total)
  - Currently 10,000 AS’s in use

Address Assignment: Example

**Q:** How does network get subnet part of IP addr?

**A:** gets allocated portion of its provider ISP’s address space

<table>
<thead>
<tr>
<th>ISP’s block</th>
<th>11001000 00010111 00010000 00000000</th>
<th>200.23.16.0/20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization 0</td>
<td>11001000 00010111 00010000 00000000</td>
<td>200.23.16.0/23</td>
</tr>
<tr>
<td>Organization 1</td>
<td>11001000 00010111 00010010 00000000</td>
<td>200.23.18.0/23</td>
</tr>
<tr>
<td>Organization 2</td>
<td>11001000 00010111 00010100 00000000</td>
<td>200.23.20.0/23</td>
</tr>
<tr>
<td>...</td>
<td>.....</td>
<td>.....</td>
</tr>
<tr>
<td>Organization 7</td>
<td>11001000 00010111 00011110 00000000</td>
<td>200.23.30.0/23</td>
</tr>
</tbody>
</table>
Hierarchical addressing: route aggregation

Hierarchical addressing allows efficient advertisement of routing information:

- Organization 0
  - 200.23.16.0/23
- Organization 1
  - 200.23.18.0/23
- Organization 2
  - 200.23.20.0/23
- Organization 7
  - 200.23.30.0/23

Fly-By-Night-ISP

Internet

ISPs-R-Us

“Send me anything with addresses beginning 200.23.16.0/20”

“Send me anything with addresses beginning 199.31.0.0/16”

Hierarchical addressing: more specific routes

ISPs-R-Us has a more specific route to Organization 1

Organization 0
  - 200.23.16.0/23
- Organization 2
  - 200.23.20.0/23
- Organization 7
  - 200.23.30.0/23
- Organization 1
  - 200.23.18.0/23

Fly-By-Night-ISP

Internet

ISPs-R-Us

“Send me anything with addresses beginning 200.23.16.0/20”

“Send me anything with addresses beginning 199.31.0.0/16 or 200.23.18.0/23”
Example: Setting forwarding table in router 1d

- Suppose AS1 learns (via inter-AS protocol) that subnet $x$ is reachable via AS3 (gateway 1c) but not via AS2.
- Inter-AS protocol propagates reachability info to all internal routers.
- Router 1d determines from intra-AS routing info that its interface $I$ is on the least cost path to 1c.
- Puts in forwarding table entry $(x,I)$.

Example: Choosing among multiple ASes

- Now suppose AS1 learns from the inter-AS protocol that subnet $x$ is reachable from AS3 \textit{and} from AS2.
- To configure forwarding table, router 1d must determine towards which gateway it should forward packets for dest $x$.
- This is also the job of inter-AS routing protocol!
Example: Choosing among multiple ASes

- Now suppose AS1 learns from the inter-AS protocol that subnet x is reachable from AS3 and from AS2.
- To configure forwarding table, router 1d must determine towards which gateway it should forward packets for dest x.
- This is also the job of inter-AS routing protocol!

**Hot potato routing:** send packet towards closest of two routers.

Name of the Game: Reachability

- Interdomain routing is about implementing policies of reachability
  - Routing efficiency and performance is important, but not essential
- ISPs could be competitors and do not want to share internal network statistics such as load and topology
- Use Border Gateway Protocol (BGP)
  - Border routers communicate over TCP port 179
  - A Path Vector Protocol
    - Communicate entire paths: Route Advertisements
  - A Router Can be involved multiple BGP sessions
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 ASs.
  2. Propagate reachability information to all AS-internal routers.
  3. Determine “good” routes to subnets based on reachability information and policy.
- allows subnet to advertise its existence to rest of Internet: “*I am here*”

BGP

- **Border Routers**
  - from the same AS speak IBGP
  - from different AS’s speak EBGP
- **EBGP and IBGP** are essentially the same protocol
  - IBGP can only propagate routes it has learned directly from its EBGP neighbors
  - All routers in the same AS form an IBGP mesh
  - Important to keep IBGP and EBGP in sync
I-BGP and E-BGP

BGP basics

- Pairs of routers (BGP peers) exchange routing info over semi-permanent TCP connections: BGP sessions
  - BGP sessions need not correspond to physical links.
- When AS2 advertises a prefix to AS1, AS2 is promising it will forward any datagrams destined to that prefix towards the prefix.
  - AS2 can aggregate prefixes in its advertisement
**Distributing reachability info**

- With eBGP session between 3a and 1c, AS3 sends prefix reachability info to AS1.
- 1c can then use iBGP to distribute this new prefix reach info to all routers in AS1.
- 1b can then re-advertise new reachability info to AS2 over 1b-to-2a eBGP session.
- When router learns of new prefix, creates entry for prefix in its forwarding table.

**Sharing 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.
BGP messages

- BGP messages exchanged using TCP.
- BGP messages:
  - OPEN: opens TCP connection to peer and authenticates sender
  - UPDATE: advertises new path (or withdraws old)
  - KEEPALIVE keeps connection alive in absence of UPDATES; also ACKs OPEN request
  - NOTIFICATION: reports errors in previous msg; also used to close connection

Path attributes & BGP routes

- When advertising a prefix, advert includes BGP attributes.
  - prefix + attributes = “route”
- Two important attributes:
  - AS-PATH: contains ASs through which prefix advertisement has passed: AS 67 AS 17
  - NEXT-HOP: Indicates specific internal-AS router to next-hop AS. (There may be multiple links from current AS to next-hop-AS.)
- When gateway router receives route advertisement, uses import policy to accept/decline.
BGP: A Path-vector protocol

```sh
er-routes>show ip bgp
BGP table version is 6128791, local router ID is 4.2.34.165
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal
Origin codes: i - IGP, e - EGP, ? - incomplete
```

<table>
<thead>
<tr>
<th>Network</th>
<th>Next Hop</th>
<th>Metric</th>
<th>LocPrf</th>
<th>Weight</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>* 13.0.0.0</td>
<td>4.0.6.142</td>
<td>1000</td>
<td>50</td>
<td>0</td>
<td>701 80 i</td>
</tr>
<tr>
<td>* 14.0.0.0</td>
<td>4.24.1.35</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>i</td>
</tr>
<tr>
<td>* 112.3.21.0/23</td>
<td>192.205.32.153</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>7018 4264 6468 ?</td>
</tr>
<tr>
<td>* 128.32.0.0/16</td>
<td>192.205.32.153</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>7018 4264 6468 25 e</td>
</tr>
</tbody>
</table>

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

BGP Update Message

- Contains information about
  - New Routes
  - Withdrawn Routes: No longer valid
  - Path Attributes:
    - Path Weights
    - Multiple Exit Discriminators
    - Local Preferences
    - Etc.
- Attribute information allows policies to be implemented
**BGP route selection**

- Router may learn about more than 1 route to some prefix. Router must select route.
- **Elimination rules:**
  1. Local preference value attribute: policy decision
  2. Shortest AS-PATH
  3. Closest NEXT-HOP router: hot potato routing
  4. Additional criteria

**BGP routing policy**

- A, B, C are provider networks
- X, W, Y are customer (of provider networks)
- X is dual-homed: attached to two networks
  - X does not want to route from B via X to C
  - .. so X will not advertise to B a route to C
BGP routing policy (2)

- A advertises to B the path AW
- B advertises to X the path BAW
- Should B advertise to C the path BAW?
  - No way! B gets no “revenue” for routing CBAW since neither W nor C are B’s customers
  - B wants to force C to route to W via A
  - B wants to route only to/from its customers!

Multiexit Discriminators (MEDs)

- One AS influences the decisions of a neighboring AS
- AS_A wants to tell AS_B that network x is closer to router 2 than to router 3
  - Router 2 advertises a smaller MED value for x than Router 3
  - AS_B prefers the path to x that does not go through 6 and 3
  - AS_B does not propagate MEDS from AS_A any further
Attribute: 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

BGP at AS4:

<table>
<thead>
<tr>
<th>Destination</th>
<th>AS Path</th>
<th>Local Pref</th>
</tr>
</thead>
<tbody>
<tr>
<td>140.20.1.0/24</td>
<td>AS3 AS1</td>
<td>300</td>
</tr>
<tr>
<td>140.20.1.0/24</td>
<td>AS2 AS1</td>
<td>100</td>
</tr>
</tbody>
</table>

BGP Policies

- Multiple ways to implement a “policy”
  - Decide not propagate advertisements
    - I’m not carrying your traffic
  - Decide not to consider MEDs but use shortest hop
    - Hot potato routing
  - Prepend own AS# multiple times to fool BGP into not thinking AS further away
  - Many others…
BGP and Performance

- BGP isn’t designed for policy routing not performance
  - Hot Potato routing is most common but suboptimal
  - Performance isn’t the greatest
- 20% of internet paths inflated by at least 5 router hops
- Very susceptible to router misconfiguration
  - Blackholes: announce a route you cannot reach
    - October 1997 one router brought down the internet for 2 hours
  - Flood update messages (don’t store routes, but keep asking your neighbors to clue you in). 3-5 million useless withdrawals!
- In principle, BGP could diverge
  - Various solutions proposed to limit the set of allowable policies
  - Focuses on avoiding “policy cycles”

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
Many small networks