Routing on the Internet

In the beginning there was the ARPANET:
- route using GGP (Gateway-to-Gateway Protocol), a distance vector routing protocol

Problems:
- needed “flag-hour” to update routing protocol
- incompatibility across vendors

Solution: hierarchical routing
- administrative autonomy:
  - each network admin can control routing within its own network
  - internet: network of networks
- allows the Internet to scale:
  - with 200 million hosts, each router can’t store all destinations in its routing table
  - route updates alone will swamp the links

Hierarchical Routing

Aggregate routers into regions of “autonomous systems” (AS)

Routers in same AS run same routing protocol
- “intra-AS” routing protocol
- each AS uses its own link metric
- routers in different ASs can run different intra-AS routing protocol

Gateway/border router
- direct link to router in other AS(s)
- keeps in its routing table:
  - next hop to other ASs
  - all hosts within its AS
  - hosts within an AS only keep a default route to the border router

The NSFNet 1989

Area hierarchy:
- backbone/core: NSFNet
- regional networks: MichNet, BARRNET, Los Nettos, Cerfnet, JVCNet, NEARNet, etc.
- campus networks

Sketch of the Internet around 1990: point of presence (pop) NSFNet backbone Regional networks Customer networks Users Walrand NSFNet backbone

AS1 AS2 AS3 AS4 border routers 1.1 2.1 3.1 4.1

<table>
<thead>
<tr>
<th>dest</th>
<th>next</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>2.1</td>
<td>2.1</td>
</tr>
<tr>
<td>3.1</td>
<td>3.1</td>
</tr>
<tr>
<td>3.4</td>
<td>3.4</td>
</tr>
</tbody>
</table>
Commercialization (1994)
Roughly hierarchical
At center: “Tier-1” ISPs
- Tier-1 ASs are those who have peering relationship with each of the other Tier-1 ASs, e.g., UUNet, BBN/Genuity, Sprint, AT&T
- national/international coverage
- peers exchange traffic for free
- customers must pay to have their traffic carried

Tier-1 providers interconnect (peer) privately
Tier-1 providers also interconnect at public network access points (NAPs)

Tier-1 ISP: e.g., Sprint
Sprint US backbone network

Qwest IP Backbone (Late 1999)
“Tier-2” ISPs: Smaller (Often Regional) ISPs

Connect to one or more tier-1 ISPs, possibly other tier-2 ISPs

Tier-2 ISP pays tier-1 ISP for connectivity to rest of Internet
- tier-2 ISP is customer of Tier-1 provider

Tier-2 ISPs also peer privately with each other, and interconnect at NAPs

“Tier-3” ISPs and local ISPs

Last hop (“access”) network (closest to end systems)

Local and tier-3 ISPs are customers of higher tier ISPs connecting them to rest of Internet
A Packet Passes Through Many Networks

Interconnected ASes

Routing is not Symmetric

Network Access Points

To learn more about Internet AS state see:

- Geoff Huston's CIDR Report: http://www.cidr-report.org/as2.0/
- Russ Haynal's ISP page: http://navigators.com/isp.html
- CAIDA skitter maps: http://www.caida.org/research/topology/as_core_network/AS_Network.xml
ISP Backbone Links

Categories of links, by inside/outside AS:
• backbone links: connect routers inside the backbone
• edge links: connect an AS with other ASs

Categorization by commercial relationship:
• access links: customers to ISP (can be multi-homed)
• peering links: inter-AS connections
  • either direct peering or through NAPS
  • may have multiple peering relationships at different geographic locations

Categorization by traffic flow:
• ingress link
• egress link

ISP Backbone Traffic

Types of traffic:
• internal traffic: between 2 access links
• transit traffic: between 2 peering links
• inbound traffic: ingress at peering link, egress at access link
• outbound traffic: vice versa

An ISP may choose to carry transit traffic
Most traffic on large ISPs (Tier-1 ASs) is transit traffic

Internet inter-AS Routing: BGP

BGP (Border Gateway Protocol) is the de facto standard for inter-AS routing

BGP provides each AS a means to:
• obtain subnet (prefix) reachability information from neighboring ASs
• propagate the reachability information to all routers internal to the AS
• determine “good” routes to subnets based on reachability information and policy
• Inter-AS routing is policy driven, not load-sensitive, generally not QoS-based

When AS2 advertises a prefix to AS1, AS2 is promising AS1 to forward any datagrams destined to that prefix coming from AS1
• AS2 can aggregate prefixes in its advertisement

Path Attributes & BGP Routes

When advertising a prefix, advertisement includes BGP attributes

Two important attributes:
• AS-PATH: the path vector of ASs through which the advertisement for a prefix passed through
• NEXT-HOP: the specific internal-AS router to next-hop AS
  (there may be multiple exits from current AS to next-hop-AS)

Sample BGP entry:

<table>
<thead>
<tr>
<th>destination</th>
<th>NEXT-HOP</th>
<th>AS-PATH</th>
</tr>
</thead>
<tbody>
<tr>
<td>198.32.163.0/24</td>
<td>202.232.1.8</td>
<td>2497 2914 3582 4600</td>
</tr>
</tbody>
</table>

• address range 198.32.163.0/24 is in AS 4600
• to get there, send to next hop router, which has address 202.232.1.8
• the path there goes through ASs 2497, 2914, 3582, in order
### BGP Messages

Pairs of routers (BGP peers) exchange reachability information over semi-permanent TCP connections: BGP sessions

- advantage of using TCP: reliable transmission allows for incremental updates
- disadvantage: TCP congestion control mechanism slows down route updates that could decongest link!
- BGP sessions do not correspond to physical links

BGP messages:

- **OPEN**: opens TCP connection to peer and authenticates sender
- **UPDATE**: advertises a new path (or withdraws an old one)
- **KEEPALIVE**: keeps connection alive in the absence of UPDATEs; also acknowledges OPEN request
- **NOTIFICATION**: reports errors in previous message; also used to close connection

### BGP Policy Routing

Some ASs are more equal than others
Commercial relationship between ASs:

- peering: peers agree to exchange traffic for free
- customer-provider: customer pays provider for access
- backup

An AS’s export policy (which routes it will advertise):

- to a customer: all routes
- to a peer or service provider: routes to all its own prefix address ranges and to its customers’ prefixes, **not** to prefixes learned from other providers or peers
- internal routing of an AS is effected by its neighbors’ route export policy

### BGP Policy Tools

An AS may learn about more than one route to some prefix
Each AS applies its own local policies to select route

Import policies: which of the advertised routes I want to use

- always check AS-PATH against routing loop
- an AS can specify its preferred **egress** point to reach a specific other AS

Export policies: how to set attributes of routes I advertise

- always prepend itself to the AS-PATH
- multiple-exit discriminator (MED):
  - an AS can tell a neighbor its preferred **ingress** point
  - cold-potato routing: ingress closest to destination prefix
  - hot-potato routing: egress (NEXT-HOP) closest to traffic source (ignore the other guy’s MED)

### BGP Routing Policy Example

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 Example

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!

Why Separate Intra- and Inter-AS Routing?

Policy:
Inter-AS: admin wants control over how its traffic is routed, who routes through its network. Inter-AS routing is policy driven, not load-sensitive, generally not QoS-based
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

The Phone Network (for Contrast)

Phone routing:
• within central office, connect directly (local call)
• between central offices (zones), use 1 hop path
• else, send to core (long distance)
• core: use 1-hop path first, else try alternate 2-hop paths

The Internet: frequent outages, each lasting from a few minutes to a few days

Communication Networks Taxonomy

POTS Network:
• parses number dialed
• sets up a circuit between caller and callee
• sends signal to ring callee’s phone
• a circuit is set up between the two ends

Strengths:
• no end-point intelligence
• excellent voice performance

Weaknesses:
• difficult to add new services
• achieves performance and reliability by over-allocating resources (expensive)

The Internet:
• data parceled into packets
• each packet carries a destination address
• each packet is routed independently
• packets can arrive out of order
• packets may not arrive at all

Strengths:
• intelligence at end points
• decentralized control
• operates over heterogeneous access techs

Weaknesses:
• variable performance, no quality of service
• no trusted infrastructure
Packet-switched Networks: Forwarding

Datagram network:
- *destination address* in packet determines next hop
- routes may change during session
- analogy: driving, asking directions

Virtual circuit network:
- each packet carries tag (virtual circuit ID), tag determines next hop
- fixed path determined at *call setup time*, remains fixed through call
- routers maintain per-call state

MultiProtocol Label Switching (MPLS)

Initial goal: speed up IP forwarding by using fixed length label (instead of IP address) to do forwarding

- borrowing ideas from Virtual Circuit (VC) approach
- but IP datagram still keeps IP address!

Label Switching

- encapsulate a data packet
  - put an MPLS header in front of the packet
  - MPLS header includes a label
  - label switching between MPLS-capable routers

MPLS Capable Routers

A.k.a. label-switched router

Forwards packets to outgoing interface based only on label value (don’t inspect IP address)

- MPLS forwarding table distinct from IP forwarding tables

Signaling protocol needed to set up forwarding

- RSVP-TE
  - forwarding possible along paths that IP alone would not allow (e.g., source-specific routing)
  - use MPLS for traffic engineering
  - must co-exist with IP-only routers

MPLS Forwarding Tables
Status of MPLS

- Deployed in practice
  - BGP-free backbone/core
  - Virtual Private Networks
  - Traffic engineering

Challenges
- Protocol complexity
- Configuration complexity
- Difficulty of collecting measurement data

Continuing evolution
- Standards
- Operational practices and tools

VPNs With Private Addresses

MPLS tags can differentiate pink VPN from orange VPN

BGP-Free Backbone Core

Routers R2 and R3 don’t need to speak BGP