Network Layer: Control Plane Part II

- Routing in the Internet: Intra vs. Inter-AS Routing
  - Intra-AS: RIP and OSPF (quick recap)
- Inter-AS: BGP and Policy Routing
- Internet Control Message Protocol: ICMP
- network management and SNMP

Readings: Textbook: Chapter 5, Sections 5.4 & 5.6-5.8
Routing in the Real World

Our routing study thus far - idealization
• all routers identical
• network “flat”

How to do routing in the Internet
• scalability and policy issues

scale: with 200 million destinations:
• can’t store all dest’s in routing tables!
• routing table exchange would swamp links!

administrative autonomy
• internet = network of networks
• each network admin may want to control routing in its own network
Internet Structure

Internet: “networks of networks”!

- IXPs or private peering
- National or tier-1 ISP
  - Regional or local ISP
  - company
  - university
- National or tier-1 ISP
  - Internet eXchange Points
  - Regional ISPs
- local ISPs
- company
- Home users
- LANs
- access via WiFi hotspots

Home users
Routing in the Internet

• The Global Internet consists of Autonomous Systems (AS) interconnected with each other:
  - **Stub AS**: small corporation: one connection to other AS’s
  - **Multi-homed AS**: large corporation (no transit): multiple connections to other ASes
  - **Transit AS**: provider, hooking many ASes together

• Each AS is assigned an **AS number (ASN)**
  - Originally 16 bits, as of Dec 1, 2006: 32 bits

• Two-level routing:
  - **Intra-AS**: administrator responsible for choice of routing algorithm within network
  - **Inter-AS**: unique standard for inter-AS routing: BGP
Number of Used ASNs

Source: Geoff Huston, http://bgp.potaroo.net

32-bit ASN up to present

Time Series of RIR AS Allocations

Date

RIR Total  ARIN  RENCC  APNIC  LACNIC  AFRINIC
Number of Allocated ASNs

Source: Geoff Huston, http://bgp.potaroo.net

16-bit ASN up to present

CSci4211: Network Layer: Control Plane Part II
Growth of Destination Net Prefixes
(measured by # of BGP routes or FIB)

Source: Geoff Huston, http://bgp.potaroo.net,
Internet AS Hierarchy

Inter-AS border (exterior gateway) routers

Intra-AS interior (gateway) routers
Intra-AS vs. Inter-AS Routing

Intra-AS routing within AS A

Inter-AS routing between A and B

Intra-AS routing within AS B

Host h1

Host h2
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, update traffic

Performance:
- Intra-AS: can focus on performance
- Inter-AS: policy may dominate over performance
Intra-AS and Inter-AS Routing

“Gateways”:
• perform inter-AS routing amongst themselves
• perform intra-AS routers with other routers in their AS

inter-AS, intra-AS routing in gateway A.c

network layer
link layer
physical layer
Intra-AS Routing

- Also known as **Interior Gateway Protocols (IGP)**
- Most common Intra-AS routing protocols:
  - **RIP**: Routing Information Protocol
  - **OSPF**: Open Shortest Path First
  - **IS-IS**: Intermediate System to Intermediate System (OSI Standard)
  - **EIGRP**: Extended Interior Gateway Routing Protocol (Cisco proprietary)
RIP (Routing Information Protocol)

- Distance vector algorithm
- Included in BSD-UNIX Distribution in 1982
- Distance metric: # of hops (max = 15 hops)

- Distance vectors: exchanged among neighbors every 30 sec via Response Message (also called advertisement)
- Each advertisement: list of up to 25 destination nets within AS
RIP: Link Failure and Recovery

If no advertisement heard after 180 sec -->
neighbor/link declared dead
- routes via neighbor invalidated
- new advertisements sent to neighbors
- neighbors in turn send out new advertisements (if tables changed)
- link failure info quickly propagates to entire net
- poison reverse used to prevent ping-pong loops (infinite distance = 16 hops)
RIP Table Processing

- RIP routing tables managed by application-level process called route-d (daemon)
- Advertisements sent in UDP packets, periodically repeated
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 (“Type-of-Services”)
  - e.g., satellite link cost set “low” for best effort; high for real time
- **Hierarchical OSPF** in large domains.
Hierarchical OSPF
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.
Inter-AS Routing in the Internet: BGP

Figure 4.5.2-new2: BGP use for inter-domain routing
BGP (Border Gateway Protocol)

- The de facto standard (BGP-4)
- **Path Vector** protocol:
  - similar to Distance Vector protocol
  - each Border Gateway broadcast to neighbors (peers) entire *path* (i.e., sequence of ASes) to destination
  - BGP routes to networks (ASes), not individual hosts
- E.g., Gateway X may announce to its neighbors it “knows” a *(AS)* path to a *destination network*, Z, via a *series of ASes*:
  
  \[
  \text{Path } (X,Z) = X,Y_1,Y_2,Y_3,...,Z
  \]
- BGP border gateways referred to as BGP *speakers*
BGP Operations: Policy Routing

Q: What does a BGP border gateway do?

• Receiving and *filtering* route advertisements from directly attached neighbor(s)
  - To accept or not accept route advertisements depends on policies (e.g., whether you “trust” your neighbors)

• Route selection (rank diff. routes to same dest. network).
  - to route to destination X, which path (of several advertised) will be taken?
  - route selection based on policies (e.g., always prefer route advertisement from “good old” neighbor Y)

• *Filtering* and sending (certain) route advertisements to neighbors
  - what/whether to advertise to your neighbors also depends on policies (e.g., don’t tell your neighbor Z that you know a route to destination X)
Customers and Providers

Customer pays provider for access to the Internet

Customer pays provider for access to the Internet
The Peering Relationship

Peers provide transit between their respective customers

Peers do not provide transit between peers

Peers (often) do not exchange $$$

---

CSci4211: Network Layer: Control Plane Part II
Peering also allows connectivity between the customers of “Tier 1” providers.
U of Minnesota (Old AS) Neighborhood

AS 1
Genuity
(was part of Level3, not part of CenturyLink)

AS 57
UMN
GigaPoP

AS 3908
SuperNet (CenturyLink)

AS 7018
AT&T

AS 1998
State of Minnesota

128.101.0.0/16

CSci4211: Network Layer: Control Plane Part II
Internet Inter-AS Routing: BGP

• **BGP (Border Gateway Protocol):** *the de facto inter-domain routing protocol*
  – “glue that holds the Internet together”

• allows subnet to advertise its existence to rest of Internet: *“I am here”* *(network reachability)*

• BGP provides each AS a means to select a route:
  – **eBGP:** obtain subnet reachability information and available routes from neighboring ASes
  – **iBGP:** propagate reachability information and available routes to all AS-internal routers.
  – determine “good” routes to other networks based on reachability information, available routes and *policy*
eBGP, iBGP Connections

gateway routers run both eBGP and iBGP protocols
BGP Messages

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

- **BGP session**: two BGP routers ("peers") exchange BGP messages over semi-permanent TCP connection:
  - advertising *paths* to different destination network prefixes (BGP is a "path vector" protocol)
  - when AS3 gateway router 3a advertises path **AS3, X** to AS2 gateway router 2c:
    - AS3 *promises* to AS2 it will forward datagrams towards X

---

**CSci4211**: Network Layer: Control Plane Part II
BGP Example

- Speaker for AS2 advertises reachability to P and Q
  - network 128.96/16, 192.4.153/24, 192.4.32/24, and 192.4.3/24, can be reached directly from AS2

  - Speaker for backbone advertises networks 128.96/16, 192.4.153/24, 192.4.32/24, and 192.4.3/24 can be reached along the path (AS1, AS2).

- Speaker can cancel previously advertised paths (by sending withdrawal messages)
Path Attributes and BGP Routes

- advertised prefix includes BGP attributes
  - prefix + attributes = “route”

- two important attributes:
  - AS-PATH: list of ASes through which prefix advertisement has passed
  - NEXT-HOP: indicates specific internal-AS router to next-hop AS

- Policy-based routing:
  - gateway receiving route advertisement uses import policy to accept/decline path (e.g., never route through AS Y).
  - AS policy also determines whether to advertise path to other other neighboring ASes
• AS2 router 2c receives path advertisement \textit{AS3,X} (via eBGP) from AS3 router 3a

• Based on AS2 policy, AS2 router 2c accepts path AS3,X, propagates (via iBGP) to all AS2 routers

• Based on AS2 policy, AS2 router 2a advertises (via eBGP) path \textit{AS2, AS3, X} to AS1 router 1c
gateway router may learn about multiple paths to destination:

- AS1 gateway router 1c learns path AS2,AS3,X from 2a
  - AS1 gateway router 1c learns path AS3,X from 3a
  - Based on policy, AS1 gateway router 1c chooses path AS3,X, and advertises path within AS1 via iBGP
How to detect loop using AS path?
BGP: AS Path Advertisement and Policy Routing

Suppose: gateway X send its path to peer gateway W

- W may or may not select path offered by X
  - cost, policy (don’t route via competitors AS), loop prevention reasons
  - Policy-based Routing Selection (using BGP attributes)

- If W selects path advertised by X, then:
  Path (W,Z) = W, Path (X,Z)

- Note: X can control incoming traffic by controlling its route advertisements to peers:
  - e.g., don’t want to route traffic to Z → don’t advertise any routes to Z
  - route filtering and export policy (by manipulating attributes)
# BGP Attributes

<table>
<thead>
<tr>
<th>Value</th>
<th>Code</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ORIGIN</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>2</td>
<td>AS_PATH</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>3</td>
<td>NEXT_HOP</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>4</td>
<td>MULTI_EXIT_DISC</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>5</td>
<td>LOCAL_PREF</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>6</td>
<td>ATOMIC_AGGREGATE</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>7</td>
<td>AGGREGATOR</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>8</td>
<td>COMMUNITY</td>
<td>[RFC1997]</td>
</tr>
<tr>
<td>9</td>
<td>ORIGINATOR_ID</td>
<td>[RFC2796]</td>
</tr>
<tr>
<td>10</td>
<td>CLUSTER_LIST</td>
<td>[RFC2796]</td>
</tr>
<tr>
<td>11</td>
<td>DPA</td>
<td>[Chen]</td>
</tr>
<tr>
<td>12</td>
<td>ADVERTISER</td>
<td>[RFC1863]</td>
</tr>
<tr>
<td>13</td>
<td>RCID_PATH / CLUSTER_ID</td>
<td>[RFC1863]</td>
</tr>
<tr>
<td>14</td>
<td>MP_REACH_NLRI</td>
<td>[RFC2283]</td>
</tr>
<tr>
<td>15</td>
<td>MP_UNREACH_NLRI</td>
<td>[RFC2283]</td>
</tr>
<tr>
<td>16</td>
<td>EXTENDED COMMUNITIES</td>
<td>[Rosen]</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>255</td>
<td>reserved for development</td>
<td></td>
</tr>
</tbody>
</table>

Most important attributes

Not all attributes need to be present in every announcement

From IANA: [http://www.iana.org/assignments/bgp-parameters](http://www.iana.org/assignments/bgp-parameters)
BGP Route Processing

Receive BGP Updates

Apply Import Policies

Apply Policy = filter routes & tweak attributes

Best Route Selection

Based on Attribute Values

Best Routes

Apply Export Policies

Install forwarding Entries for best Routes.

IP Forwarding Table

Transmit BGP Updates

Apply Policy = filter routes & tweak attributes

Open ended programming.
Constrained only by vendor configuration language

Based on Attribute Values

Best Routes

Apply Policy = filter routes & tweak attributes

CSci4211: Network Layer: Control Plane Part II
BGP Route Selection

- router may learn about more than one route to destination AS, selects route based on:
  1. local preference value attribute: policy decision
  2. shortest AS-PATH
  3. closest NEXT-HOP router: hot potato routing
  4. additional criteria ...
Tweak Tweak Tweak Tweak

- For inbound traffic
  - Filter outbound routes
  - Tweak attributes on outbound routes in the hope of influencing your neighbor’s best route selection

- For outbound traffic
  - Filter inbound routes
  - Tweak attributes on inbound routes to influence best route selection

In general, an AS has more control over outbound traffic
A simple BGP scenario

- A, B, C are provider networks
- X, W, Y are customer (of provider networks)
- X is dual-homed: attached to two networks
  - C tells X networks belonging to C, i.e., a route to them via C
  - X does not want to carry traffic from B via X to C
  - .. so X will not advertise to B any route to networks in C learned from C

Export Policy for a Provider AS?
Export Policy for a Customer AS?
BGP: Controlling Who Routes to You

A simple BGP scenario

- 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!

Export Policy for a Peer AS?

What about route selection?
- which should you prefer? a route learned from a customer AS, a peer AS or a provider AS?

Hint: think how money flows!
Shedding Inbound Traffic with ASPATH Padding Hack

Padding will (usually) force inbound traffic from AS 1 to take primary link.

192.0.2.0/24 ASPATH = 2

Customer AS 2

Provider AS 1

Primary

Backup

192.0.2.0/24 ASPATH = 2 2 2

Padding will (usually) force inbound traffic from AS 1 to take primary link.
Padding May Not Shut Off All Traffic

AS 1
provider

AS 2
customer

AS 3
provider

192.0.2.0/24
ASPATH = 2

192.0.2.0/24
ASPATH = 2 2 2 2 2 2 2 2 2 2 2 2 2 2

primary
backup

AS 3 will send traffic on “backup” link because it prefers customer routes and local preference is considered before ASPATH length!

Padding in this way is often used as a form of load balancing
Early Exit or Hot Potato Routing: Go for the Closest Egress Point

This Router has two BGP routes to 192.44.78.0/24. Hot potato: get traffic off of your network as soon as possible. Go for egress 1!
Hot Potato Routing (from authors’ lecture notes)

• 2d learns (via iBGP) it can route to X via 2a or 2c
• *hot potato routing*: choose local gateway that has least intra-domain cost (e.g., 2d chooses 2a, even though more AS hops to X): don’t worry about inter-domain cost! (this statement is incorrect! AS-Path attribute precedes the IGP weight criteria in best route selection: hence 2d will select route the from 2c !)
• there is a way to prefer the route from 2a (using local preference attribute) - but this is no longer hot potato routing
Many customers want their provider to carry the bits!

- High bandwidth Provider backbone
- Low bandwidth customer backbone
- SFO
- NYC
- San Diego
- Heavy Content Web Farm

- 2865
- 17
- 15
- 56

- tiny http request
- huge http reply
BGP, OSPF, Forwarding Table Entries

Q: how does router set forwarding table entry to distant prefix?

- recall: 1a, 1b, 1c learn about dest X via iBGP from 1c: “path to X goes through 1c”
  - 1d: OSPF intra-domain routing: to get to 1c, forward over outgoing local interface 1

<table>
<thead>
<tr>
<th>dest</th>
<th>interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Q: how does router set forwarding table entry to distant prefix?

- recall: 1a, 1b, 1c learn about dest X via iBGP from 1c: “path to X goes through 1c”
  - 1d: OSPF intra-domain routing: to get to 1c, forward over outgoing local interface 1
  - 1a: OSPF intra-domain routing: to get to 1c, forward over outgoing local interface 2

<table>
<thead>
<tr>
<th>dest</th>
<th>interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>X</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Recap: Why Different Intra-, 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
Network Layer
Control Plane Summary

• Routing is a Key Function in the control plane
  - Basic Issues: topology, path selection, ...

• Distributed (intra-AS) routing algorithms: LS vs. DV
  - Link State (LS): How does it work?
  - Distance Vector (DV): How does it work? Issues?

• Centralized control plane and SDN controllers
  - Openflow; POX, OpenDayLight (ODL), ONOS, ..

• Routing in the Internet:
  - Intra-AS vs. Inter-AS routing
  - Distributed intra-AS routing protocols: RIP and OSPF

• Inter-domain (inter-AS) routing: BGP and Policy Routing
Routing & Forwarding: Logical View of a Router
IP Forwarding & IP/ICMP Protocol

Routing protocols
- path selection
- RIP, OSPF, BGP

IP protocol
- addressing conventions
- packet handling conventions

ICMP protocol
- error reporting
- router “signaling”

Transport layer: TCP, UDP

Network layer

Data Link layer (Ethernet, WiFi, PPP, …)

Physical Layer (SONET, …)
## ICMP: Internet Control Message Protocol

- used by hosts & routers to communicate network-level information
  - error reporting: unreachable host, network, port, protocol
  - echo request/reply (used by ping)
- network-layer “above” IP:
  - ICMP msgs carried in IP datagrams
- ICMP message: type, code plus first 8 bytes of IP datagram causing error

<table>
<thead>
<tr>
<th>Type</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>echo reply (ping)</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>dest. network unreachable</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>dest host unreachable</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>dest protocol unreachable</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>dest port unreachable</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>dest network unknown</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>dest host unknown</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>source quench (congestion control - not used)</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>echo request (ping)</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>route advertisement</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>router discovery</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>TTL expired</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>bad IP header</td>
</tr>
</tbody>
</table>
Traceroute and ICMP

• source sends series of UDP segments to destination
  - first set has TTL =1
  - second set has TTL=2, etc.
  - unlikely port number

• when datagram in nth set arrives to nth router:
  - router discards datagram and sends source ICMP message (type 11, code 0)
  - ICMP message include name of router & IP address

• when ICMP message arrives, source records RTTs

stopping criteria:
  ▪ UDP segment eventually arrives at destination host
  ▪ destination returns ICMP “port unreachable” message (type 3, code 3)
  ▪ source stops
What is Network Management?

- autonomous systems (aka “network”): 1000s of interacting hardware/software components
- other complex systems requiring monitoring, control:
  - jet airplane
  - nuclear power plant
  - others?

"Network management includes the deployment, integration and coordination of the hardware, software, and human elements to monitor, test, poll, configure, analyze, evaluate, and control the network and element resources to meet the real-time, operational performance, and Quality of Service requirements at a reasonable cost."
Infrastructure for Network Management

definitions:

managed devices contain managed objects whose data is gathered into a Management Information Base (MIB)
SNMP Protocol

Two ways to convey MIB info, commands:

request/response mode

trap mode
## SNMP Protocol: Message Types

<table>
<thead>
<tr>
<th>Message type</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>GetRequest</td>
<td>manager-to-agent: “get me data” (data instance, next data in list, block of data)</td>
</tr>
<tr>
<td>GetNextRequest</td>
<td></td>
</tr>
<tr>
<td>GetBulkRequest</td>
<td></td>
</tr>
<tr>
<td>InformRequest</td>
<td>manager-to-manager: here’s MIB value</td>
</tr>
<tr>
<td>SetRequest</td>
<td>manager-to-agent: set MIB value</td>
</tr>
<tr>
<td>Response</td>
<td>Agent-to-manager: value, response to Request</td>
</tr>
<tr>
<td>Trap</td>
<td>Agent-to-manager: inform manager of exceptional event</td>
</tr>
</tbody>
</table>

*CSci4211: Network Layer: Control Plane Part II*
SNMP Protocol: Message Formats

More on network management: see earlier editions of textbook!
Network Layer Summary

• Network Layer Functions and Service Models
  - Addressing, Routing and Forwarding
  - Virtual Circuit vs. Datagram; Programmable Data Plane via SDN
  - Distributed vs. Centralized Control Plane
• IP Addressing Scheme: CIDR; DHCP
• IP Forwarding and IP Protocol
  - IP Datagram Forwarding Model: dest. in same net vs. diff. net
  - IPv4: Datagram Format, IP Fragmentation, ...; IPv6
• Network Layer Routing
  - Fundamental Issues
  - Two Basic Distributed Algorithms: LS and DV
  - Routing in the Internet: Intra-AS vs. Inter-AS routing
    • Intra-AS: RIP and OSPF (distributed routing protocols)
    • Inter-AS: BGP and Policy Routing
• Openflow switches, SDN Controllers & Centralized Control Plane
• ICMP & SNMP