Chapter 4
Network Layer

A note on the use of these ppt slides:
We’re making these slides freely available to all (faculty, students, readers). They’re in PowerPoint form so you can add, modify, and delete slides (including this one) and slide content to suit your needs. They obviously represent a lot of work on our part. In return for use, we only ask the following:

- If you use these slides (e.g., in a class) in substantially unaltered form, that you mention their source (after all, we’d like people to use our book!)
- If you post any slides in substantially unaltered form on a www site, that you note that they are adapted from (or perhaps identical to) our slides, and note our copyright of this material.

Thanks and enjoy! JFK/KWR

Jim Kurose, Keith Ross
Addison-Wesley, April 2009.
Chapter 4: Network Layer

4.1 Introduction
4.2 Virtual circuit and datagram networks
4.3 What’s inside a router
4.4 IP: Internet Protocol
   - Datagram format
   - IPv4 addressing
   - ICMP
   - IPv6
4.5 Routing algorithms
   - Link state
   - Distance Vector
   - Hierarchical routing
4.6 Routing in the Internet
   - RIP
   - OSPF
   - BGP
4.7 Broadcast and multicast routing
Hierarchical Routing

Our routing study thus far - idealization
- all routers identical
- network “flat”
... not true in practice

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
Hierarchical Routing

- aggregate routers into regions, “autonomous systems” (AS)
- routers in same AS run same routing protocol
  - “intra-AS” routing protocol
  - routers in different AS can run different intra-AS routing protocol
  
  **gateway router**
  - at “edge” of its own AS
  - has link to router in another AS
Interconnected ASes

- forwarding table 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 destined outside of AS1:
  - Router should forward packet to gateway router, but which one?

AS1 must:
1. Learn which dests are reachable through AS2, which through AS3
2. Propagate this reachability info to all routers in AS1

Job of inter-AS routing!
Example: Setting forwarding table in router 1d

- Suppose AS1 learns (via inter-AS protocol) that subnet $x$ 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.
  - Installs forwarding table entry $(x,I)$
Example: Choosing among multiple ASes

- now suppose AS1 learns from inter-AS protocol that subnet $x$ is reachable from AS3 and from AS2.
- to configure forwarding table, router 1d must determine which gateway it should forward packets towards for dest $x$
  - this is also job of inter-AS routing protocol!
**Example: Choosing among multiple ASes**

- now suppose AS1 learns from 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 job of inter-AS routing protocol!
- **hot potato routing:** send packet towards closest of two routers.

---

Learn from inter-AS protocol that subnet \(x\) is reachable via multiple gateways → Use routing info from intra-AS protocol to determine costs of least-cost paths to each of the gateways → Hot potato routing: Choose the gateway that has the smallest least cost → Determine from forwarding table the interface I that leads to least-cost gateway. Enter \((x,I)\) in forwarding table.
Chapter 4: Network Layer

4.1 Introduction

4.2 Virtual circuit and datagram networks

4.3 What’s inside a router

4.4 IP: Internet Protocol
  - Datagram format
  - IPv4 addressing
  - ICMP
  - IPv6

4.5 Routing algorithms
  - Link state
  - Distance Vector
  - Hierarchical routing

4.6 Routing in the Internet
  - RIP
  - OSPF
  - BGP

4.7 Broadcast and multicast routing
Recap of inter-domain routing

- First, reachability information between ASs
- Second, choose paths and advertise internally
  - AS1 learns from inter-AS protocol that subnet $x$ is reachable from AS3 and from AS2.
  - To configure forwarding table, router 1d must determine which gateway it should forward packets towards for dest $x$.
  - This is also the job of inter-AS routing protocol!
Internet inter-AS routing: BGP

- **BGP (Border Gateway Protocol):** *the de facto inter-domain routing protocol*
  - “glue that holds the Internet together”
- **BGP provides each AS a means to:**
  - **eBGP:** obtain subnet reachability information from neighboring ASs.
  - **iBGP:** propagate reachability information to all AS-internal routers.
  - determine “good” routes to other networks based on reachability information and policy.
- **allows subnet to advertise its existence to rest of Internet:** “*I am here*”
What you need to grasp BGP?

1. To know what type of relationships exist between Ases
   Commercial relationship!
2. To understand a path vector protocol
3. To use elaborate maths to understand its correctness / convergence

Here, we will look at points 1 and 2, and briefly overview point 3
Many thanks to Renata Teixeira for her help!
What type of ASes exist?

- Large, tier-1 provider with a nationwide backbone
  - At the “core” of the Internet, don’t have providers
- Medium-sized regional provider with smaller backbone
- Small network run by a single company or university
Connections Between Networks
Single-Homed Customers

- UPMC has only one connection to the Internet
Multi-Homed Customers

- Same provider: e.g., Wanadoo to FT
- Different providers: e.g., Wanadoo to FT and BT
Customer-provider relationship

- Customer needs to be reachable from everyone
- Customer does not want to provide transit service

UPMC is customer of DT
Wanadoo is a customer of FT and BT

traffic to/from UPMC
transit traffic is not allowed
Peer-peer relationship

- Peers exchange traffic between customers

FT and BT are peers
FT and DT are peers
FT doesn’t provide transit for its peers

customers exchange traffic

FT
BT
Wanadoo

UPMC

DT
Peering also allows connectivity between the customers of “Tier 1” providers.
How peering decisions are made?

- Peer
  - Reduces upstream transit costs
  - Can increase end-to-end performance
  - May be the only way to connect your customers to some part of the Internet (“Tier 1”)

- Don’t Peer
  - You would rather have customers
  - Peers are usually your competition
  - Peering relationships may require periodic renegotiation
What you need to grasp BGP?

1. To know what type of relationships exist between Ases
   - Customer Provider or Peering!
2. To understand a path vector protocol
3. To use elaborate maths to understand its correctness / convergence

Here, we will look at points 1 and 2, and briefly overview point 3
Many thanks to Renata Teixeira for her help!
Which inter-routing protocol?

- Scale, Privacy
  - Link state problematic: flood information, costly
  - Distance vector: ok

- Policy:
  - Distance vector insufficient: single cost
Use of shortest-path is restrictive

- All traffic must travel on shortest paths
- All nodes need common notion of link costs
- Incompatible with commercial relationships
  - Cust of 1 and 3 should not communicate on two peering links 1-2 and 2-3!
Which inter-routing protocol?

- Scale, Privacy
  - **Link state** problematic: flood information, costly
  - **Distance vector**: ok

- **Policy**:
  - **Distance vector** insufficient: single cost

- **Path vector**
  
  Extends distance vector by advertising not the cost but the **entire path** used
  - Avoids loop and count to infinity problem
  - Allows to have flexible import/export policy
BGP route

- Destination prefix (e.g., 128.112.0.0/16)
- Route attributes, including
  - AS path (e.g., “2 1”)
  - Next-hop IP address (e.g., 12.127.0.121)
BGP path selection

- **Simplest case**
  - Shortest AS path
  - Arbitrary tie break

- **Example**
  - Three-hop AS path preferred over a four-hop AS path
  - AS 7 prefers path through AS 6

- **But, BGP not only shortest-path routing**
  - Policy-based routing
BGP Session Operation

Establish session on TCP port 179

Exchange all active routes

Exchange incremental updates

While connection is ALIVE exchange route UPDATE messages
2 types of BGP connections

- **External BGP (eBGP)**
  - Session between routers in different ASes

- **Internal BGP (iBGP)**
  - Need to distribute BGP information within the AS
  - iBGP sessions are routed using IGP
iBGP mesh doesn’t scale

- Configuration overhead
  - N border routers means \( N(N-1)/2 \) sessions
  - One new router requires configuring all the others

- Routing overhead
  - Each router has to listen to updates from all neighbors
  - Larger routing tables, because of alternate routes
Route reflectors

- Acts like a route server
  - Routes from clients, distribute to other RRs
  - Routes from other RRs, distribute to clients
- Only sends best route
Incremental protocol

- A node learns multiple paths to destination
  - Stores all of the routes in a routing table
  - Applies policy to select a single active route
  - May advertise the route to its neighbors

- Incremental updates
  - Announcement:
    - Upon selecting a new active route, add node id to path
  - Withdrawal
    - If the active route is no longer available
BGP route processing

Receive BGP Updates → Apply Import Policies → Apply Policy = filter routes & tweak attributes → Based on Attribute Values → Best Route Selection → Best Route Table → Apply Policy = filter routes & tweak attributes → Best Routes → Apply Export Policies → Transmit BGP Updates

Install forwarding Entries for best Routes.

IP Forwarding Table
Import policy: Filtering

- **Discard some route announcements**
  - Detect configuration mistakes and attacks
- **Examples on session to a customer**
  - Discard route if prefix not owned by the customer
  - Discard route that contains other large ISP in AS path

128.112.0.0/16
Export policy: Filtering

- Discard some route announcements
  - Limit propagation of routing information
- Examples
  - Don’t announce routes from one peer to another
  - Don’t announce routes for network-management hosts
BGP Policy Configuration

- Routing policy languages are vendor-specific
  - Not part of the BGP protocol specification
  - Different languages for Cisco, Juniper, etc.
- Still, all languages have some key features
  - Policy as a list of clauses
  - Each clause matches on route attributes
  - ... and either discards or modifies the matching routes
- Configuration done by human operators
  - Implementing the policies of their AS
  - Business relationships, traffic engineering, security, ...
Best route selection:
Simplified BGP decision process

- Ignore if next hop unreachable
- Highest local preference
- Lowest AS path length
- Lowest MED (with same next hop AS)
- Prefer eBGP over iBGP
- Lowest IGP cost to egress router
- Lowest router ID of egress router
What you need to grasp BGP?

1. To know what type of relationships exist between Ases
   Customer Provider or Peering!
2. To understand a path vector protocol
   Path exchange, import/export policy
3. To use elaborate maths to understand its correctness / convergence

Many thanks to Renata Teixeira for her help!
Import policy: Local preference

- Favor one path over another
  - Override the influence of AS path length
  - Apply local policies to prefer a path
- Example: prefer customer over peer
Example: Customer to provider

<table>
<thead>
<tr>
<th>router</th>
<th>import policies</th>
<th>route selection</th>
<th>export policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>local pref = 100</td>
<td>select Univ route</td>
<td>send to other iBGP neighbors</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>select A’s route</td>
<td>send to other eBGP neighbors</td>
</tr>
</tbody>
</table>

132.239.17.0/24
Example: Peers

Suppose Medium1, Big, and Large are peers.

**route** | **import policies** | **route selection** | **export policies**
---|---|---|---
A | local pref = 90 | select M1 route | send to other iBGP routers don’t send
B | | select A’s route | send to customers
C | | select A’s route |
Example: Customers vs. peers

Suppose:
- M1 is a customer of Big and Large
- Big and Large are peers

<table>
<thead>
<tr>
<th>Router A</th>
<th>Import Policies</th>
<th>Route Selection</th>
<th>Export Policies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>local pref (M1) = 100</td>
<td>select M1 route</td>
<td>send to other iBGP and eBGP neighbors</td>
</tr>
<tr>
<td></td>
<td>local pref (L) = 80</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

132.239.0.0/16
Asymmetric routing
BGP converges slowly, if at all

- Path vector avoids count-to-infinity
  - But, ASes still must explore many alternate paths
  - ... to find the highest-ranked path that is still available

- Policies need to be validated
  - Condition to avoid loop, condition to obtain convergence
  - Based on algebraic properties of path’s attributes

- Fortunately, in practice
  - Most popular destinations have very stable BGP routes
  - And most instability lies in a few unpopular destinations

- Still, lower BGP convergence delay is a goal
  - Can be tens of seconds to tens of minutes
  - High for important interactive applications
  - ... or even conventional application, like Web browsing
Conclusions

- BGP is solving a hard problem
  - Routing protocol operating at a global scale
  - With tens of thousands of independent networks
  - That each have their own policy goals
  - And all want fast convergence

- Key features of BGP
  - Prefix-based path-vector protocol
  - Incremental updates (announcements and withdrawals)
  - Policies applied at import and export of routes
  - Internal BGP to distribute information within an AS
  - Interaction with the IGP to compute forwarding tables
Recommended readings

- Tim Griffin’s inter-domain routing page:
  - http://www.cl.cam.ac.uk/~tgg22/interdomain/
- Non-convergence of BGP: policy conflicts
- Delayed convergence
- Food for thought...
BGP basics

- **BGP session**: two BGP routers ("peers") exchange BGP messages:
  - advertising *paths* to different destination network prefixes ("path vector" protocol)
  - exchanged over semi-permanent TCP connections
- **when AS3 advertises a prefix to AS1**:
  - AS3 *promises* it will forward datagrams towards that prefix
  - AS3 can aggregate prefixes in its advertisement
BGP basics: distributing path information

- using eBGP session between 3a and 1c, AS3 sends prefix reachability info to AS1.
  - 1c can then use iBGP to distribute new prefix 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, it creates entry for prefix in its forwarding table.
Path attributes & BGP routes

- advertised prefix includes BGP attributes
  - prefix + attributes = “route”
- two important attributes:
  - **AS-PATH**: contains ASs through which prefix advertisement has passed: e.g., AS 67, AS 17
  - **NEXT-HOP**: indicates specific internal-AS router to next-hop AS. (may be multiple links from current AS to next-hop-AS)
- gateway router receiving route advertisement uses import policy to accept/decline
  - e.g., never route through AS x
  - *policy-based* routing
BGP route selection

- router may learn about more than 1 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
BGP messages

- BGP messages exchanged between peers over TCP connection

- 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
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 path AW to B
- B advertises path BAW to X
- **Should B advertise path BAW to C?**
  - 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 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