CS 4/55231
Internet Engineering

Kent State University
Dept. of Computer Science

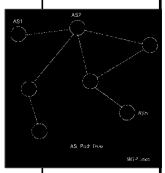
LECT-7B

Routing Between Autonomous Systems (Example: BGP4) RFC 1771

BGP4 Overview

INTERNET ENGINEERING

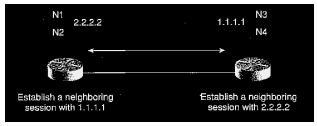
- BGP4 is a path vector protocol which carry routing information between Autonomous Systems. Path vector comes from the fact that BGP routing information carries a sequence of AS numbers, indicating the path of a route.
- BGP uses TCP as its transport (port 179). Connected Routers are called peers.
- At the beginning all routing information is sent to peer, but after that it only advertises changes. In contrast to Link State or Vector distance protocols, incremental update tremendously saves Internet Traffic.



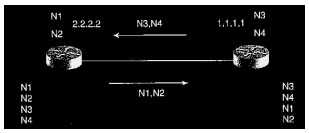
Example of Operations

INTERNET ENGINEERING

- Establish a Connection



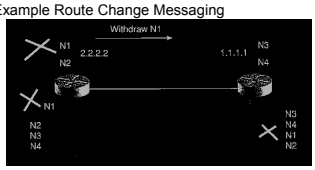
- Exchange all Routing Updates



Example of Operations (contd..)

INTERNET ENGINEERING

- Example Route Change Messaging



- Steady State After the Route Change

BGP4 Overview (2)

INTERNET ENGINEERING

- Route:
  - a unit of information that pairs a **destination** with the **attributes of a path** to that destination:
- Routes are advertised between a pair of BGP speakers in UPDATE messages:
- Routes are stored in the **Routing Information Bases** (RIBs).

Routing Information Base

- The Routing Information Base (RIB) within a BGP speaker consists of three distinct parts:
-
- Adj-RIBs-In:
  - The Adj-RIBs-In store routing information that has been learned from inbound UPDATE messages.
- Loc-RIB:
  - The Loc-RIB contains the local routing information that the BGP speaker has selected by applying its local policies to the routing information contained in its Adj-RIBs-In.
- Adj-RIBs-Out:
  - The Adj-RIBs-Out store the information that the local BGP speaker has selected for advertisement to its peers. It is carried in the local BGP speaker's UPDATE messages and advertised to its peers.

INTERNET ENGINEERING

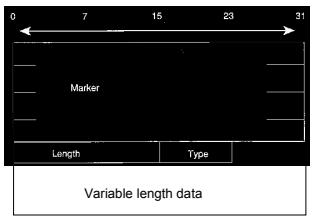
BGP Message Types

- OPEN
  - A request to open a peer connection, when a router starts.
- UPDATE
  - Routing update includes all the necessary information that BGP use to build a loop free picture of the Internet.
- NOTIFICATION
  - Notification messages are exchanged to report error
- KEEPALIVE
  - These are short messages exchanged between peers to indicate that peers are reachable.

INTERNET ENGINEERING

BGP Message Format

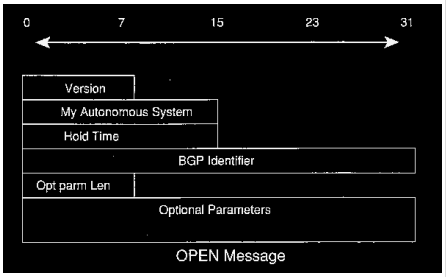
- Marker's are all 1s, unless authentication in place.



- Length ranges from 19-4096 bytes.

INTERNET ENGINEERING

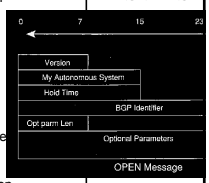
Open Message Format



INTERNET ENGINEERING

Open Message Fields (1)

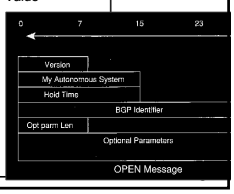
- Version:
  - indicate the version number of BGP protocol, such as BGP1, 2 or 3. BGP peers agree on a version number which is highest common among them.
- My Autonomous System:
  - a 2-byte AS ID of the router.
- Hold Time:
  - maximum number of seconds that may elapse between two KEEPALIVE or UPDATE messages. If no such message arrives past the hold time the neighbor is considered dead. Otherwise the timer is restarted upon each receipt of KEEPALIVE or UPDATE.
  - The minimum of the hold time is effective when it differs between peers.
  - 0 hold time mean, never expire timer.



INTERNET ENGINEERING

Open Message Fields (2)

- BGP Identifier:
  - An identifier indicating the sender's ID. Typically, the highest IP address of a router.
- Optional Parameters
  - A set of optional triplets in the format
  - <Parameter Type, Parameter Length, Parameter Value>
  - Field lengths: 1 byte + 1 byte + variable.
  - Example: Authentication parameters.
- Optional Parameters length
  - 1-byte unsigned integer in bytes.
  - O indicates no parameter present.

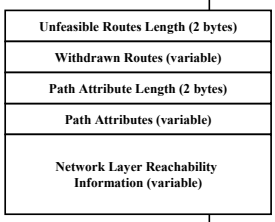


INTERNET ENGINEERING

Update Message Format

INTERNET ENGINEERING

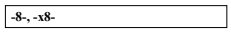
- One UPDATE message advertises one route. It is described by several path attributes.
- NLRI field contains a set of destinations about which BGP is trying to inform its other BGP neighbors.
- One UPDATE message can contain multiple routes to be withdrawn.



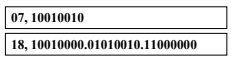
NLRI format

INTERNET ENGINEERING

- Each Address is <1 byte length, prefix (variable bytes)>



- Each Prefix is padded to byte boundary
- Below are some examples:

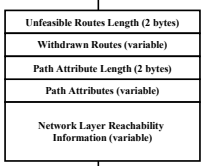


QUIZ: If MASK is 100 bits. Then How many bytes will be prefix?

Update Message Fields (1)

INTERNET ENGINEERING

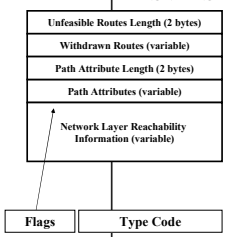
- Unfeasible Routes Length
  - This 2-octets unsigned integer indicates the total length of the Withdrawn Routes field in octets. A value of 0 indicates that no routes are being withdrawn.
- Unfeasible Routes
  - List of routes which are no longer valid.
  - The format of the routes corresponds to CIDR format
  - Format: <2 byte prefix+ IP address (variable)>.
  - The Prefix field contains IP address prefixes followed by enough trailing bits to make the end of the field fall on an octet boundary. Note that the value of trailing bits is irrelevant.
  - <18, 192.213.134.0>



Update Message Fields (2)

INTERNET ENGINEERING

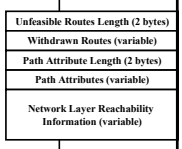
- Path Attribute Length
  - This 2-octets unsigned integer indicates the total length of the Path Attribute fields in octets. A value of 0 indicates that no reachability information is included.
- Path Attributes
  - variable length sequence of path attributes is present in every UPDATE.
  - Each path attribute is a triplet <attribute type, attribute length, attribute value> of variable length.
  - Attribute Type is a two-octet field that consists of the **Attribute Flags** octet followed by the **Attribute Type Code** octet.



Path Attribute

INTERNET ENGINEERING

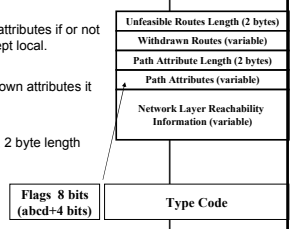
- These actually contains detail information about a path such as:
  - What is the next hop.
  - What is the cost?
  - Is it aggregated?
- These information is used in path selection process.
- There are four types:
  - Well-known mandatory.
  - Well-known discretionary.
  - Optional transitive.
  - Optional non transitive.



Path Attribute: Flags

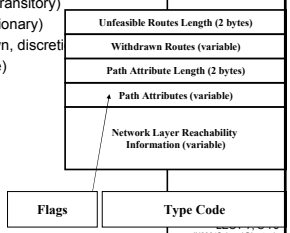
INTERNET ENGINEERING

- Optional Flag – a
  - If or not the attribute is optional (1) or well-known i.e. must (0).
- Transitive Flag - b
  - Always 1 for well-known. For optional attributes if or not the attribute should be forwarded or kept local.
- Partial Flag - c
  - 1 the information is partial. For well-known attributes it must be 0.
- Extended Length Flag – d
  - If the attribute is larger than 255 bytes, 2 byte length field, instead of 1.
- Rest 4 bits are unused.



Path Attribute: Type Codes

- 1-ORIGIN (well-known, mandatory)
- 2-AS_path (well-known, mandatory)
- 3-NEXT_HOP (well-known, mandatory)
- 4-MULTI_EXIT_DISC (optional, non-transitory)
- 5-LOCAL_PREF (well-known, discretionary)
- 6-ATOMIC_AGGREGATE (well-known, discreti
- 7-AGGREGATOR (optional, transitive)
- 8-COMMUNITY (optional, transitive)
- 9-ORIGINATOR_ID
- 10-Cluster List
- 11-Destination Preference
- 12-Advertiser
- 13-rcid_path



INTERNET ENGINEERING

Path Attribute : ORIGIN

- ORIGIN is a well-known mandatory attribute that defines the origin of the path information.
- The data octet can assume the following values:
  - 0 IGP - NLRI is interior to the originating AS
  - 1 EGP - NLRI learned via EGP
  - 2 INCOMPLETE - NLRI learned by some other means
- The ORIGIN attribute shall be generated by the autonomous system that originates the associated routing information.
- It shall be included in the UPDATE messages of all BGP speakers that choose to propagate this information to other BGP speakers.

INTERNET ENGINEERING

Path Attribute : AS_PATH

- AS_PATH is a well-known mandatory attribute that is composed of a sequence of AS path segments.
- Each AS path segment is a triplet:
  - <path segment type, path segment length, path segment value>.
- The path segment type is a 1-octet long field with values:
  - 1 AS_SET: unordered set of ASs or
  - 2 AS_SEQUENCE: ordered set of ASs a route in the UPDATE message has traversed
- The path segment length is a 1-octet long field containing the number of ASs in the path segment value field.
- The path segment value field contains one or more AS numbers, each encoded as a 2-octets long field.

INTERNET ENGINEERING

Path Attribute : AS_PATH (cont..)

- When a BGP speaker propagates a route which it has learned from another BGP speaker's UPDATE message, it shall modify the route's AS_PATH attribute based on the location of the BGP speaker to which the route will be sent:
  - When a given BGP speaker advertises the route to another BGP speaker located in its own autonomous system, the advertising speaker shall not modify the AS_PATH attribute associated with the route.
  - When a given BGP speaker advertises the route to a BGP speaker located in a neighboring autonomous system, then the advertising speaker shall update the AS_PATH attribute as follows:

INTERNET ENGINEERING

Path Attribute : AS_PATH (cont..)

- If the first path segment of the AS_PATH is :
  - Type AS_SEQUENCE, the local system shall prepend its own AS number as the last element of the sequence (put it in the leftmost position).
  - Type AS_SET, the local system shall prepend a new path segment of type AS_SEQUENCE to the AS_PATH, including its own AS number in that segment.

INTERNET ENGINEERING

Path Attribute : AS_PATH (cont..)

- When a BGP speaker originates a route then:
  - a) the originating speaker shall include its own AS number in the AS_PATH attribute of all UPDATE messages sent to BGP speakers located in neighboring autonomous systems. (In this case, the AS number of the originating speaker's autonomous system will be the only entry in the AS_PATH attribute).
  - b) the originating speaker shall include an empty AS_PATH attribute in all UPDATE messages sent to BGP speakers located in its own autonomous system. (An empty AS_PATH attribute is one whose length field contains the value zero).

INTERNET ENGINEERING

Path Attribute : NEXT_HOP

- This is a well-known mandatory attribute that defines the IP address of the border router that should be used as the next hop to the destinations listed in the NLRI field of the UPDATE message.
- If a border router belongs to the same AS as its peer, then the peer is an internal border router. Otherwise, it is an external border router.
- A BGP speaker can advertise any internal border router as the next hop provided that the interface associated with the IP address of this border router (as specified in the NEXT_HOP path attribute) shares a common subnet with both the local and remote BGP speakers.

INTERNET ENGINEERING

Path Attribute : NEXT_HOP (cont..)

- Do not propagate hearsay!
  - A BGP speaker can advertise any external border router as the next hop, provided that the IP address of this border router was learned from one of the BGP speaker's peers, and the interface associated with the IP address of this border router shares a common subnet with the local and remote BGP speakers.
- Do not try selling it back to the original seller!
  - A BGP speaker must never advertise an address of a peer to that peer as a NEXT_HOP, for a route that the speaker is originating.
- Pass on original story to all in your group!
  - When a BGP speaker advertises the route to a BGP speaker located in its own autonomous system, the advertising speaker shall not modify the NEXT_HOP attribute associated with the route.

INTERNET ENGINEERING

Path Attribute : NEXT_HOP (cont..)

- When a BGP speaker receives the route via an internal link, it may forward packets to the NEXT_HOP address if the address contained in the attribute is on a common subnet with the local and remote BGP speakers.
- When a BGP speaker advertises the route to a BGP speaker located in its own autonomous system, the advertising speaker shall not modify the NEXT_HOP attribute associated with the route.

INTERNET ENGINEERING

Path Attribute : MULTI_EXIT_DISC (MED)

- Used to hint another AS which of the entry ways to prefer when there are multiple router choices.
- This information is never propagated to another AS (non-transitive).

INTERNET ENGINEERING

Path Attribute : MULTI_EXIT_DISC

- This is an optional non-transitive attribute. It is a four octet unsigned number which is called a metric.
- It may be used on external (inter-AS) links to discriminate among multiple exit or entry points to the same neighboring AS.
- All other factors being equal, the exit or entry point with lower metric should be preferred.
- If received over external links, it:
  - may be propagated over internal links to other BGP speakers within the same AS.
  - but cannot be propagated to other BGP speakers in neighboring AS's.

INTERNET ENGINEERING

Path Attribute : LOCAL_PREF

- It is a well-known discretionary attribute, a four octet non-negative integer.
- Used by a BGP speaker to inform other BGP speakers in its own AS of the originating speaker's preference for an advertised route.
- A BGP speaker shall calculate the degree of preference for each external route and include the degree of preference when advertising a route to its internal peers. The higher degree of preference should be preferred.
- A BGP speaker shall not include this attribute in UPDATE messages that it sends to BGP speakers located in a neighboring AS.

INTERNET ENGINEERING

Path Attribute : ATOMIC_AGGREGATE

- ATOMIC_AGGREGATE is a well-known discretionary attribute of length 0.
- Generally used to indicate any loss of information while aggregation has been performed.
- Used by a BGP speaker to inform other BGP speakers that the local system selected a less specific route without selecting a more specific route which is included in it.
- If a BGP speaker, when presented with a set of overlapping routes from one of its peers selects the less specific route without selecting the more specific one, then the local system shall attach the ATOMIC_AGGREGATE attribute to the route when propagating it to other BGP speakers .

INTERNET ENGINEERING

Path Attribute : ATOMIC_AGGREGATE (..)

- A BGP speaker that receives a route with the ATOMIC_AGGREGATE attribute shall not remove the attribute from the route when propagating it to other speakers.
- A BGP speaker that receives a route with the ATOMIC_AGGREGATE attribute shall not make any NLRI of that route more specific when advertising this route to other BGP speakers.
- A BGP speaker that receives a route with the ATOMIC_AGGREGATE attribute needs to be cognizant of the fact that the actual path to destinations, as specified in the NLRI of the route, while having the loop-free property, may traverse ASs that are not listed in the AS_PATH attribute.

INTERNET ENGINEERING

Path Attribute : AGGREGATOR

- AGGREGATOR is an optional transitive attribute of length 6.
- The attribute contains the last AS number that formed the aggregate route (encoded as 2 octets), followed by the IP address of the BGP speaker that formed the aggregate route (encoded as 4 octets).

INTERNET ENGINEERING

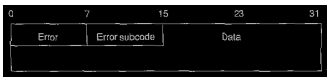
Notification Message Format

1- Message Header Error subcodes:
- 1 - Connection Not Synchronized.
- 2 - Bad Message Length.
- 3 - Bad Message Type.

2- OPEN Message Error subcodes:
- 1 - Unsupported Version Number.
- 2 - Bad Peer AS.
- 3 - Bad BGP Identifier. '

4 - Unsupported Optional Parameter.
- 5 - Authentication Failure.
- 6 - Unacceptable Hold Time.

3- UPDATE Message Error subcodes:
- 1 - Malformed Attribute List.
- 2 - Unrecognized Well-known Attribute.
- 3 - Missing Well-known Attribute.
- 4 - Attribute Flags Error.
- 5 - Attribute Length Error.
- 6 - Invalid ORIGIN Attribute
- 7 - AS Routing Loop.
- 8 - Invalid NEXT_HOP Attribute.
- 9 - Optional Attribute Error.
- 10 - Invalid Network Field.
- 11 - Malformed AS_PATH.



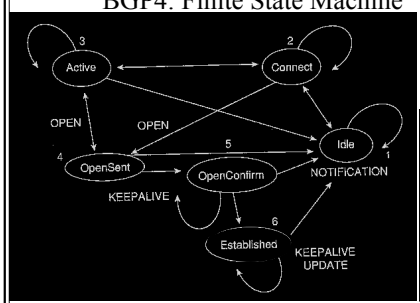
INTERNET ENGINEERING

Keep Alive Message Format

- There is no message body, just 19 bytes BGP message is sent by the BGP-speakers.
- Typically it is sent at 1/3rd of the hold-time interval to ensure proper operation.

INTERNET ENGINEERING

BGP4: Finite State Machine



EVENTS LIST:
1 - BGP Start
2 - BGP Stop
3 - BGP Transport connection open
4 - Transport connection closed
5 - Transport connection open failed
6 - BGP Transport fatal error
7 - ConnectRetry timer expired
8 - Hold Timer expired
9 - KeepAlive timer expired
10 - Receive OPEN message
11 - Receive KEEPALIVE message
12 - Receive UPDATE messages
13 - Receive NOTIFICATION message

INTERNET ENGINEERING

Overview of Operations

- Connection Opening:
  - Two systems form a TCP connection between them.
  - Exchange messages to open and confirm the connection parameters.
  - The initial data is the entire BGP routing table.
- Steady State Operation
  - BGP does not require periodic refresh . BGP speaker must retain the current version of the entire BGP routing tables of all of its peers for the duration of the connection.
  - Updates are sent when routing tables changes.
  - KeepAlive messages are sent periodically to ensure the liveness of the connection.

INTERNET ENGINEERING

Overview of Operations

- Notification messages are sent in response to errors or special conditions. If a connection encounters an error condition, a notification message is sent and the connection is closed.
- Routes are advertised between a pair of BGP speakers in UPDATE messages: the destination is the systems whose IP addresses are reported in the Network Layer Reachability Information (NLRI) field, and the the path is the information reported in the path attributes fields of the same UPDATE message.
- If a BGP speaker chooses to advertise the route, it may add to or modify the path attributes of the route before advertising it to a peer.

INTERNET ENGINEERING

Who can be BGP host?

- The hosts executing the Border Gateway Protocol need not be routers.
- A non-routing host could exchange routing information with routers via EGP or even an interior routing protocol.
- That non-routing host could then use BGP to exchange routing information with a border router in another Autonomous System.

INTERNET ENGINEERING

Building Peer Sessions

- During connection establishment by OPEN message they check AS numbers to determine who is EBGP peer and who is IGBP peer.
- Internal BGP peers have to be “logically” directly connected. i.e. they should have IP connectivity.
- External BGP peers must be physically connected. It drops any UPDATE if it is not.
  - Exception CISCO EBGP Multihop.

INTERNET ENGINEERING

Authentication

- BGP message header allows authentication.
- It is used so that hackers can not pose as peer and inject wrong routing information.
- It used MD5. It involves a combination of password and key mechanism.
-

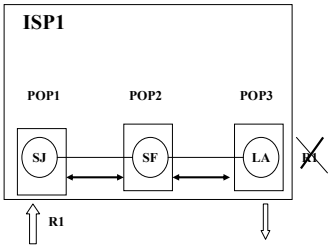
INTERNET ENGINEERING

BGP Continuity Inside an AS

- To avoid creating routing loops inside an AS, BGP does not advertise to internal BGP peers routes that are learned via other IBGP peers.
- Thus, every BGP routers within an AS should be fully connected.
-

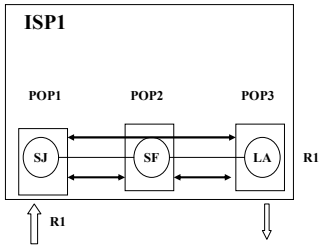
INTERNET ENGINEERING

BGP Continuity Inside an AS



INTERNET ENGINEERING

BGP Continuity Inside an AS



INTERNET ENGINEERING

Synchronization of multiple BGP Speakers in an AS

- If a particular AS has multiple BGP speakers and is providing transit service for other ASs, then care must be taken to ensure a consistent view of routing within the AS by the interior routing protocol.
  - Let the BGP speakers arrive at an agreement as to which border routers will serve as exit/entry points for particular destinations outside the AS using a common set of policies,.
  - This information is communicated to the AS's internal routers, possibly via the interior routing protocol.
  - Care must be taken to ensure that the interior routers have all been updated with transit information before the BGP speakers announce to other ASs that transit service is being provided.

INTERNET ENGINEERING

Source of Routing Update (1/3)

- Dynamic Injection of Routes
  - Purely dynamic [**redistribute**]
  - Routes that arrive in the local routing table via IGPs such as RIP, IGRP, OSPF, EIGRP, ISIS are all automatically forwarded.
  - If an AS is running multiple of the IGPs then it can use "protocol distance" to tell which route to accept when there competing route information.
  - Not a very good approach. Privacy as can be compromised.

INTERNET ENGINEERING

Source of Routing Update (2/3)

- Semi-Dynamic Injection of Routes
  - A list is maintained which indicated routes for which network can be advertised. [use **network command**]
  - If the routes that arrive in the local routing table are in this list only then these are forwarded.
  - List is typically size limited.

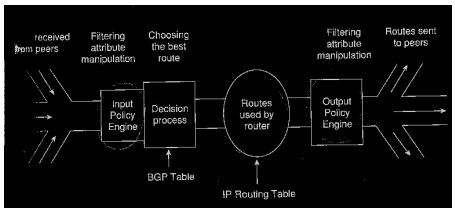
INTERNET ENGINEERING

Source of Routing Update (3/3)

- Static Injection of Routes
  - Administrator manually defines the static routes in the routing table then use either **network** or **redistribute** technique to propagate them.
  - The routes never disappear from routing table.
  - Increases routing stability.
  - But needed expensive admin time.

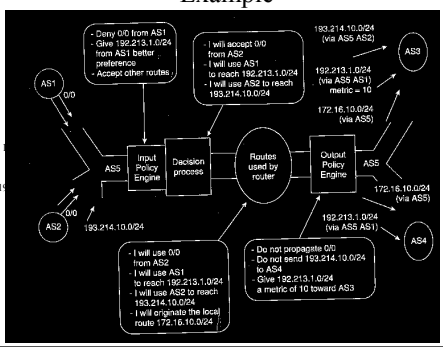
INTERNET ENGINEERING

Putting it All together



INTERNET ENGINEERING

Example



INTERNET ENGINEERING

Summary

- In this section we have seen the basic inter-domain routing architecture. However, today's internet operation is much more complex.
- Advanced Internet Systems Design (AID2003, Fall 2003) will discuss further advance techniques for synchronization, load-balancing, fault-tolerant, and high throughput technical operation of an AS or of the Internet, in general.
- In this section we have seen the basic structure of systems needed so that multiple autonomous systems can operate effectively. However, as the Internet is evolving, further structural refinements and inventions will be needed, specially with its internationalization.

INTERNET ENGINEERING

Next Topic:
ISP Traffic Management

Mid Term

- 4 Questions Total:
  - 1 True-False
  - 1 P2P/LAN/WAN/SWITCHING
  - 1 TCP/IP/Routing & Addressing
- Open Book 60 min.

INTERNET ENGINEERING