E6998-02: Internet Routing

Lecture 12
Exterior Gateway Protocols

John Ioannidis

AT&T Labs – Research

ji+ir@cs.columbia.edu

Announcements

Lectures 1-12 are available.

Have you been working on your project proposal?

Still looking for a TA.

The old days

- Original Arpanet.
  - Single routing domain (GGP, then SPF).
  - Every gateway (router) knew all destinations.
  - Not all that many destinations back then!
- RFC827:
  - Scaling issues identified.
    - High algorithm overhead (given the hardware).
    - Stability.
  - Software engineering issues identified.
    - Different implementations.
    - Different default parameters.
  - Administrative issues.
    - Multiple network administrators.

RFC827: EGP

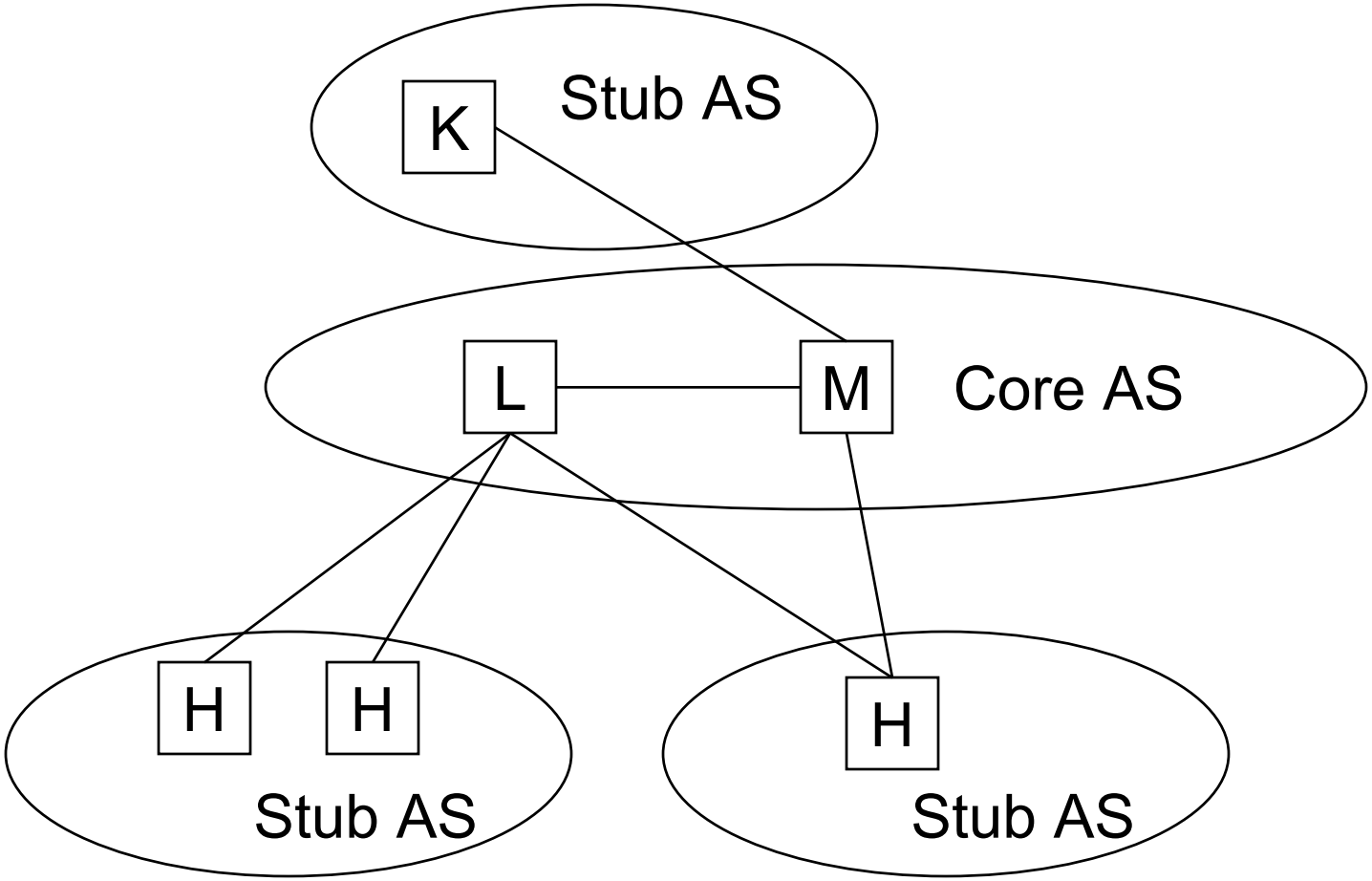
- Replace single routing domain with…
- Multiple interconnected autonomous routing domains.
  - Called “Autonomous Systems” (AS).
- Each AS managed independently.
- Identified by a 16 bit number (ASN).
  - ASN1: BBN, ASN14: Columbia, …
  - 64512 – 65535 (FC00-FFFF) are private.
- ASes run IGPs for their internal routing.
- ASes communicate using an EGP (of which “EGP” is the first one).
- IGPs are concerned with optimizing paths.
- EGPs are concerned with adhering to policy.
  - Different metrics make optimization an ill-defined problem.

Exterior Gateway Protocol

- RFC 827, 888, 904.
- IP Protocol 8
- *Neighbors* (or *peers*): routers exchanging EGP messages.
  - *Interior neighbors*: in the same AS.
  - *Exterior neighbors*: in different ASes.
- All EGP routers accept messages about other ASes.
- *Stub gateways* send messages only about their own AS.
- *Core gateways* send messages about all ASes.

EGP topology

- One Core AS to which Stub ASes connect.
- Avoids loops.



EGP Neighbor Acquisition/Reachability

- Neighbor addresses manually configured.
- There is an *active* and a *passive* neighbor.
- *Neighbor Acquisition Request* unicast to neighbor.
  - *Hello interval* and *Poll interval* specified.
- *Neighbor Acquisition Confirm* and *Refuse*.
- *Neighbor Cease* / *Neighbor Cease Ack*.
- Relationship maintained with periodic *Hello/I-Heard-You* messages.

- Nothing surprising here!

EGP Network Reachability Protocol

- One neighbor sends a *Poll* message
  - Contains a sequence number.
- The other responds with an *Update* message.
  - Echoes the s/n.
  - Includes list of reachable networks.
- Hello/IHU messages contain the same s/n until an update is received.
  - S/N is then incremented.
- Unsolicited updates are an option.
- Notion of indirect (proxy) updates.
  - Route server.
- Details are not important.

Limitations of EGP

- Inability to detect routing loops.
  - Metrics don't really mean much.
  - Count-to-infinity too slow.
- Must be engineered loop-free.
- Policy was kludged when NSFNET dictated AUPs.
- Little interaction with IGP to pick best routes.
- Very slow to advertise topology changes.
- Classful.

- Abandoned in favor of BGP(-1, -2, -3, -4).

BGP-4 Overview

- RFC1771.
- BGP runs over TCP (port 179).
- BGP happens between exactly two nodes.
  - *BGP Session* between *BGP Peers*.
    - *BGP Speakers*.
  - A router can have multiple sessions (with multiple peers).
- Maintains the concept of Autonomous System.
- Allows arbitrary AS connectivity.
  - Transit ASes.
  - Non-transit ASes.
  - No such thing as “backbone”.
- Objective: find optimal AS paths satisfying policy constraints.

BGP-4 Overview, cont'd

- In a nutshell:
  - Establish connection with peer.
  - Exchange all routes.
  - While link stays up
    - Exchange incremental updates.
- Routes are not refreshed.
  - A route is considered valid until it is changed or withdrawn.
  - Or until the BGP session is terminated.

BGP-4 Overview, cont'd

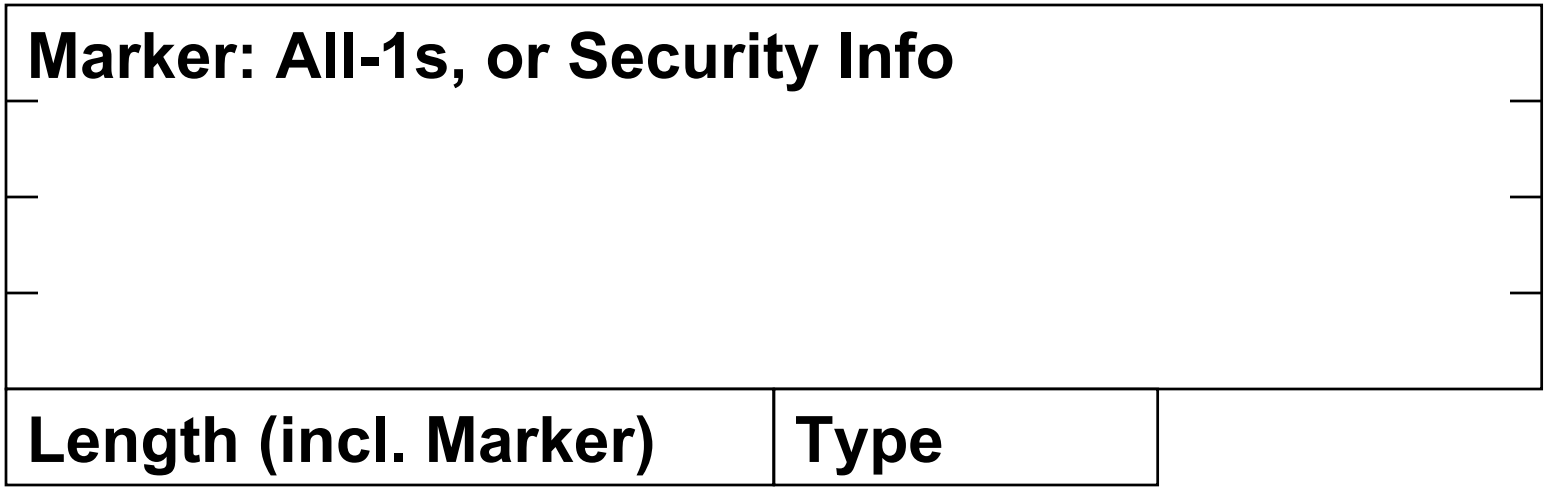
- Advertisements are about reachability.
  - A advertises to B a path for N.
  - B is assured that A uses that path to reach N.
- Path-Vector:
  - Almost like DV, except complete paths are advertised.
    - Loops are prevented this way.
- Attributes:
  - That's what makes BGP so flexible and extensible …
  - and prone to misconfigurations.
  - Next hops, various metrics, path, …
  - Lots of new attributes defined since RFC1771.

Bringing up BGP

- *BGP Peers*: endpoints of a *BGP Session*.
- BGP Peers are configured.
  - No automatic discovery.
- Start at *Idle* state.
- Attempt TCP connection: *Connect* state.
- While establishing TCP connection: *Active* state.

- Now BGP messages can be sent.
  - While TCP connection is up.

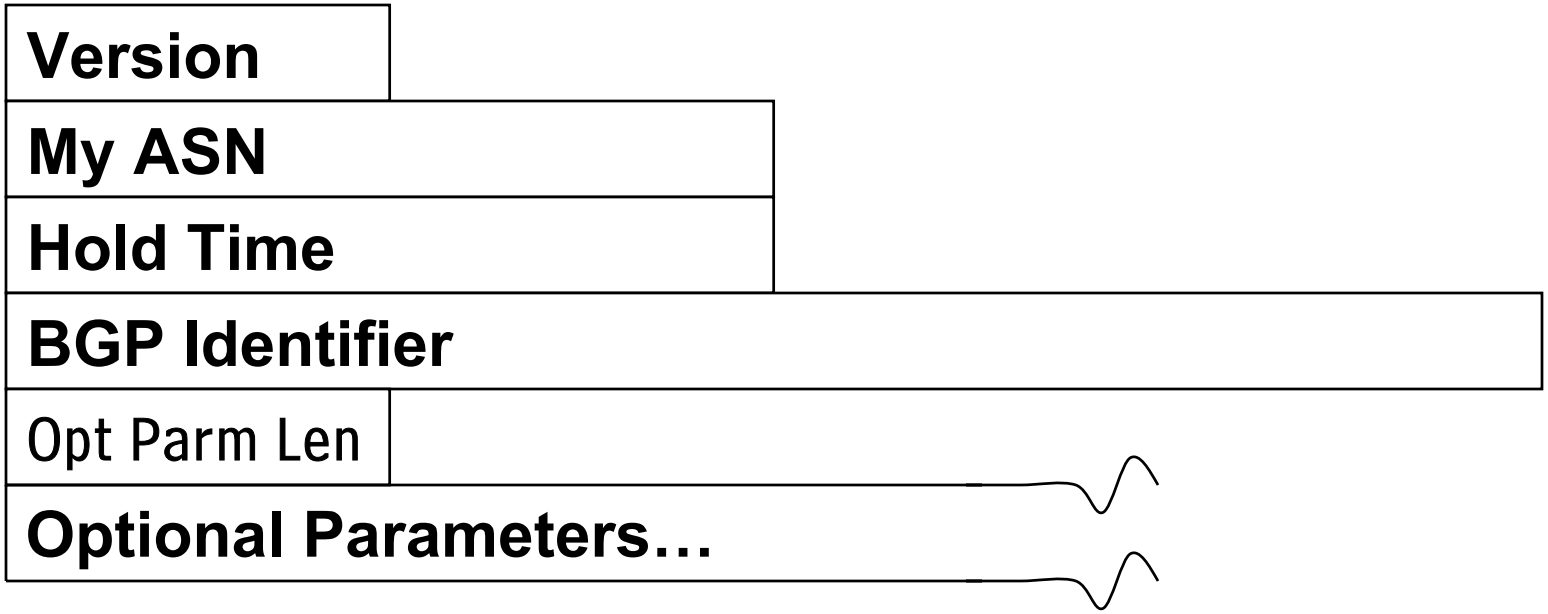
BGP Message Common Header



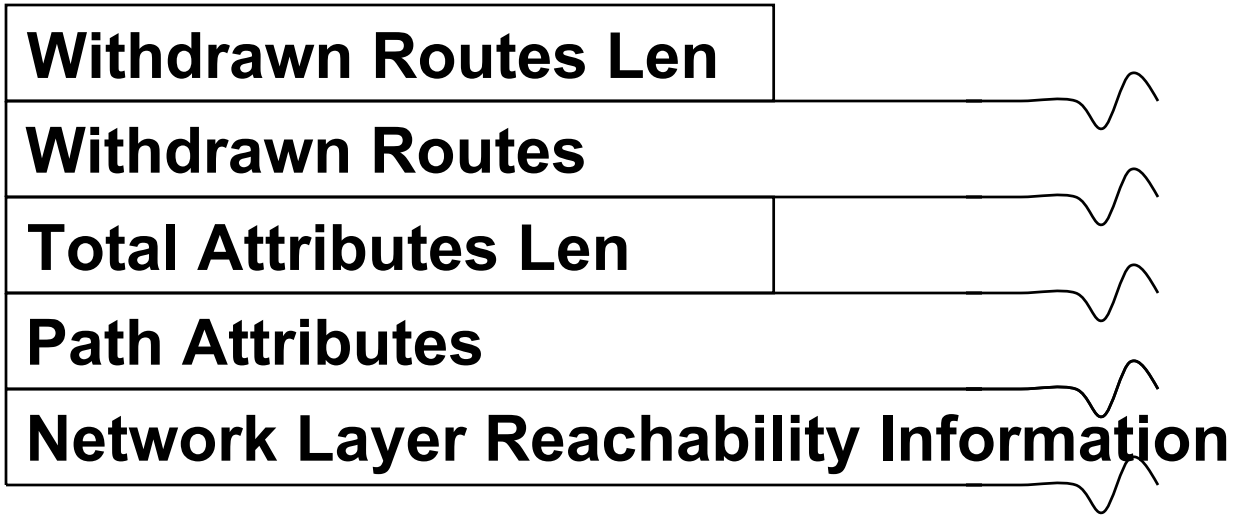
- Type is one of:
  - OPEN (1)
  - UPDATE (2)
  - NOTIFICATION (3)
  - KEEPALIVE (4)

BGP OPEN

- BGP speakers identify each other.
  - And verify that they are who they are supposed to be.
- Verify they speak the same version of BGP.
- Inform each other of their ID.
- Exchange/negotiate optional parameters.



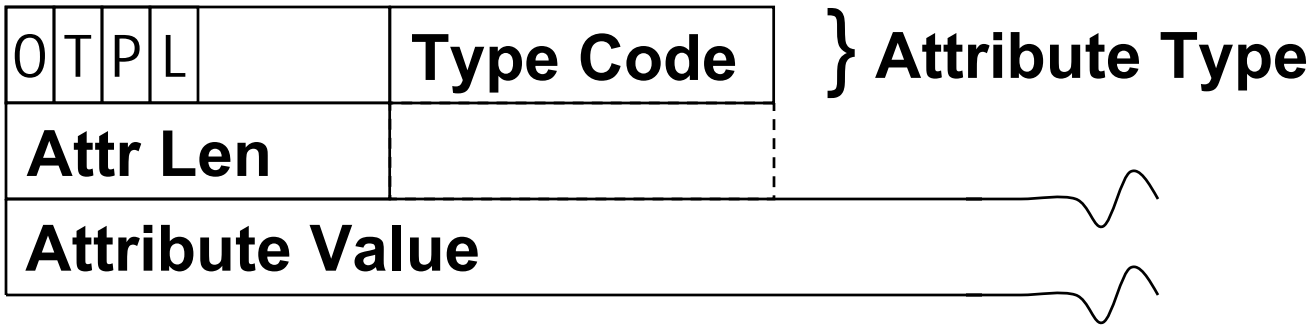
BGP UPDATE


Withdrawn Routes Len
Withdrawn Routes
Total Attributes Len
Path Attributes
Network Layer Reachability Information

Withdrawn Routes

- List of IP prefixes to withdraw.
- Length is the prefix length.
- Prefix is padded to a multiple of 8 bits.
  - Pad bits ignored.

Path Attributes



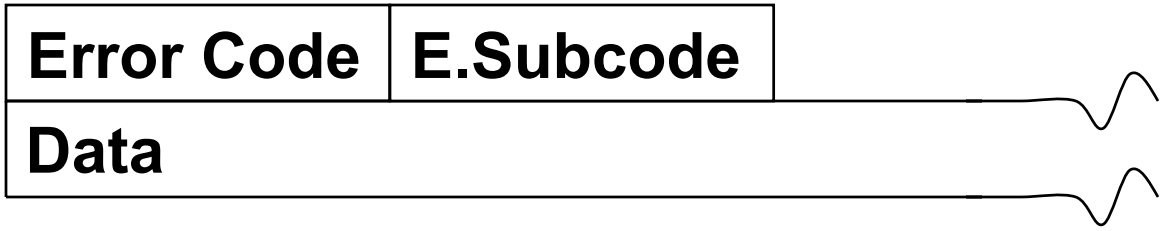
- O: Optional/Well Known
- T: Transitive/Nontransitive (passed on to peers)
- P: Partial: only some routers in the path understand an Optional and Transitive attribute.
  - If O=0 and T=0 then P must be 0.
- L: Extended Length: L=1 means length field is 2 bytes.

- Attributes apply to all advertised prefixes in the UPDATE message.

Network Layer Reachability Information

- List of advertised prefixes.
- All attributes apply to all prefixes.
- Prefixes with different attributes are advertised in separate UPDATE messages.

BGP NOTIFICATION



- Report errors about:
  - Format of received message.
  - Unexpected state.
  - Timers expiring.
- The TCP connection is closed right after the NOTIFICATION.
  - All notifications are fatal!

BGP KEEPALIVE

- Sent if there have been no updates in the last HoldTimer seconds.
- Syntactically, just a BGP header with Type=4

(About Keepalives)

- Some TCP implementations have the notion of a keepalive:
  - Packet sent periodically to probe the connection.
- What it does keep alive is the underlying link IF the underlying link depends on continuous traffic to stay up (e.g., dialup).

- TCP state is kept only at the endpoints.
  - Intermediate hops do not need to be refreshed.
- If intermediate links go away temporarily, TCP will keep retransmitting until they come back up.
- In most cases, tearing down a link when no other data traffic would have flowed anyway is wasteful.

- Hence the term “makedeads”.

Keepalive

- In BGP, we DO want a Makedead!
- A failed link indicates that routing should change.
  - Since BGP messages are exchanged over the same link that all other traffic would be routed.
  - (There is an exception to this, don't worry about it yet.)
- Detects if the link has failed, and tears the session down.
- A torn-down BGP session causes routes to be withdrawn
  - This is the desired behavior.

Conceptual Model of Operation

- BGP is about advertising prefixes.
  - Some prefixes are learned from BGP neighbors.
  - Some more prefixes are also learned from the IGP.
  - Some of these prefixes are advertised to neighbors.
- RIB: Routing Information Base.
- Each router keeps:
  - One **Adj-RIB-In** for each peer.
    - Stores prefixes learned from each peer.
  - Prefixes from all the **Adj-RIB-In**s are selected for use.
  - Stored in the **Loc-RIB**.
    - One per router.
  - One **Adj-RIB-Out** for each peer.
    - Stores prefixes to be advertised to each peer.