CS644
Advanced Networks

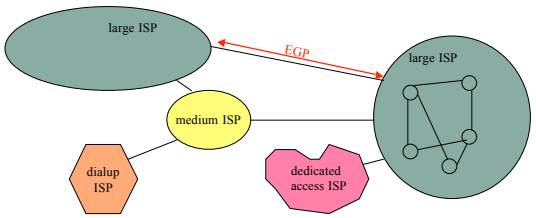
Lecture 12

BGP

Andreas Terzis

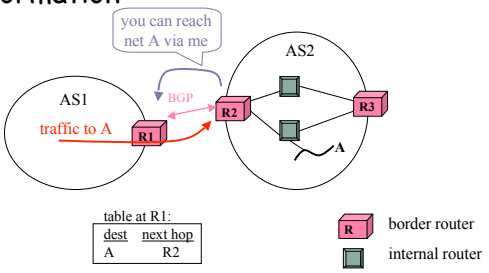
(intro slides by Nina Taft)

Internet Topology



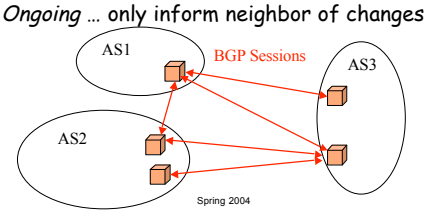
- AS (Autonomous System) - a collection of routers under the same technical and administrative domain.
- EGP (External Gateway Protocol) - used between two AS's to allow them to exchange routing information so that traffic can be forwarded across AS borders. Example: BGP

Purpose: to share connectivity information

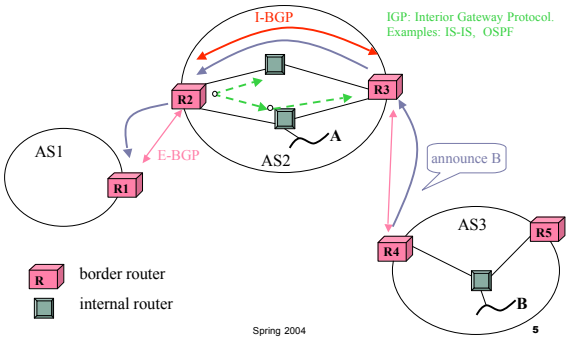


BGP Sessions

- One router can participate in many BGP sessions.
- *Initially* … node advertises ALL routes it wants neighbor to know (could be >50K routes)
- *Ongoing* … only inform neighbor of changes



Routing Protocols



Four Basic Messages

- *Open*:
  Establishes BGP session (uses TCP port #179)
- *Notification*:
  Report unusual conditions
- *Update*:
  Inform neighbor of new routes that become active
  Inform neighbor of old routes that become inactive
- *Keepalive*:
  Inform neighbor that connection is still viable

UPDATE Message

- Used to either advertise and/or withdraw prefixes
- Path attributes: list of attributes that pertain to ALL the prefixes in the Reachability Info field

FORMAT:

Withdrawn routes length (2 octets)
Withdrawn routes (variable length)
Total path attributes length (2 octets)
Path Attributes (variable length)
Reachability Information (variable length)

Advertising a prefix

- When a router advertises a prefix to one of its BGP neighbors:
  - information is valid until first router explicitly advertises that the information is no longer valid
  - BGP does not require routing information to be refreshed
  - if node A advertises a path for a prefix to node B, then node B can be sure node A is using that path itself to reach the destination.

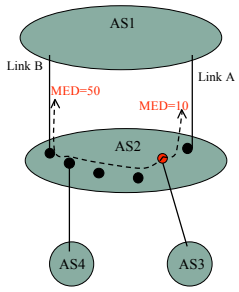
ATTRIBUTES

- ORIGIN:
  - Who originated the announcement? Where was a prefix *injected* into BGP?
  - IGP, EGP or Incomplete (often used for static routes)
- AS-PATH:
  - a list of AS's through which the announcement for a prefix has passed
  - each AS prepends its AS # to the AS-PATH attribute when forwarding an announcement
  - useful to detect and prevent loops

Prefix	***Next hop***	***AS Path***
128.73.4.21/21	232.14.63.4	1239 701 3985 631

Attribute: Multi-Exit Discriminator (MED)

- when AS's interconnected via 2 or more links
- AS announcing prefix sets MED
- enables AS2 to indicate its preference
- AS receiving prefix uses MED to select link
- a way to specify how close a prefix is to the link it is announced on



Route Stability

- Routing instability: rapid fluctuation of network reachability information
- Route flapping: when a route is withdrawn and re-announced repeatedly in a short period of time
  - happens via UPDATE messages
- Because messages propagate to global Internet, route flapping behavior can cascade and deteriorate routing performance in many places
- Effects: increased packet loss, increased network latency, CPU overhead, loss of connectivity

Types of Routing Updates

- Forwarding instability
  - reflects legitimate topology changes
  - e.g., changes in Prefix, NEXT_HOP and/or ASPATH
  - affects forwarding paths used
- Policy fluctuation
  - reflects changes in policy
  - e.g., changes in MED, LOCAL_PREF, etc.
  - may not necessarily affect forwarding paths used
- Pathological
  - redundant messages
  - reflect neither topology nor policy changes

Anecdotes of Route Flap Storms

- April 25, 1997 - small Virginia ISP injected incorrect map into global Internet. Map said Virginia ISP had optimal connectivity to all destinations. Everyone sent their traffic to this ISP. Result: shutdown of Tier-1 ISPs for 2 hours.
- August 14, 1998 - misconfigured database server forwarded all queries to ".net" to wrong server. Result: loss of connectivity to all .net servers for few hours.
- Nov. 8, 1998 - router software bug led to malformed routing control message. Caused interoperability problem between Tier-1 ISPs. Result: persistent pathological oscillations and connectivity loss for several hours.

General Statistics

- 1996: 45K prefixes
- 1996: 3-5 million updates per day in Internet core
  - 125 updates per prefix per day
- Correlation of instability and usage
  - instability highest during business hours
  - instability lowest during nights, on weekends and in summer

Taxonomy (as per Labovitz et. al.)

Name	Type	Character
WADiff	Explicit withdrawal followed by announcement. Replace route with different path	Legitimate
AADiff	Announced twice (implicit withdrawal). Replace route with different path.	Legitimate
WADup	Explicit withdrawal followed by announcement. Replace route with same path.	Legitimate or pathological
AADup	Announced twice (implicit withdrawal). Replace route with same path.	Policy change or pathological
WWDup	Repeated duplicate withdrawals	Pathological

Per Event Type Statistics

- 1996 relative impact (approximately):
  WWDup (96%),
  AADup (2%),
  WADup (1%), AADiff(1/2%),
  WADiff (1/2%)

Who's Responsible?

- AS's
  - No single AS dominates instability statistics
  - No correlation between the size of an AS and its share of updates generated.
- Prefixes
  - Instability is evenly distributed across routes.
  - Example of measurements:
    - 75% of AADiff events come from prefixes change less than 10 times a day.
    - 80-90% of instability comes from prefixes that are announced less than 50 times/day.

Sources of Instabilities in General

- router configuration errors
- transient physical and data link problems
- software bugs
- problems with leased lines (electrical timing issues that cause false alarms of disconnect)
- router failures
- network upgrades and maintenance

Instability Problem and Cause. Example 1.

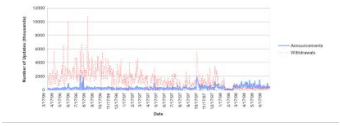
- *Problem*: 3-5 million duplicate withdrawals
- *Cause*: stateless BGP implementation
  - time-space tradeoff: no state maintained on what advertised to peers
  - when receive any change, transmit withdrawal to all peers regardless of whether previously notified or not
  - sent updates for both explicit and implicit withdrawals
- By 1998, most vendors had BGP implementations with partial state.
- Result: number of WWDups reduced by an order of magnitude

Instability Problem and Cause. Example 2

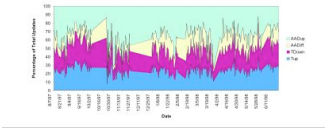
- Problem: duplicate announcements
- Cause: min-advertisement timer & stateless BGP
  - min-adv timer: wait 30 seconds. Combine all received updates in last 30 seconds into single outbound update message (if possible).
  - within 30 seconds route can be withdrawn and re-announced so that there is no net change to original announcement
- Solution: Have BGP keep some state about recently sent messages to peers. Avoid sending duplicate messages

Origins of Internet Routing Instability

- Follow-up study in 1999
  - Volume of inter-domain routing messages decreased by an order of magnitude
  - Mainly due to the reduction of pathological Ww_{dup}
  - Software upgrade on routing vendor's code
  - Number of announcements doubled to a total of 430K per day

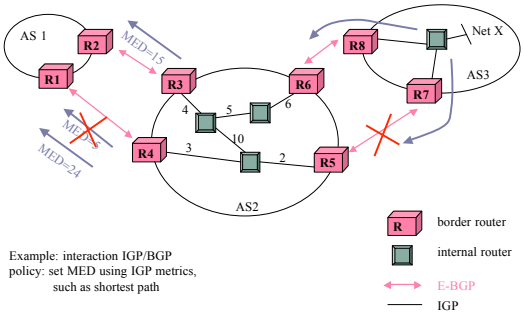


Breakdown of BGP updates



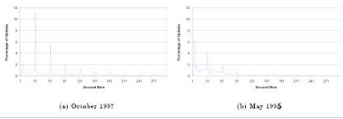
- 40% is T_{up} and T_{down}
- 15% is AA_{diff}
- The majority is AA_{dup}

Instability Problem and Cause



Example: interaction IGP/BGP policy: set MED using IGP metrics, such as shortest path

Update Frequency

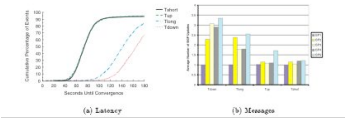


- Most updates happen with an interval of 30 or 60 sec
  - Due to *unjittered* timer in BGP implementation

Where are we

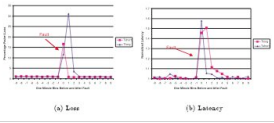
- Initially numbers of updates very high
  - WW_{dup}: FIXED
  - AA_{diff} and AA_{dup}: FIXED
  - What about T_{up} and T_{down}?
    - 40% of routing updates observed
    - Related question: how long does it take for a route to *converge*?
- Delayed Internet Routing Convergence [LABJ00]

Observed Convergence Latency



- 20% of T_{long} and 40% of T_{down} oscillate for more than 3 minutes
- T_{down} and T_{long} converge more slowly than T_{up} and T_{short}
- T_{down} and T_{long} trigger twice the number of update messages than T_{up} and T_{short}
- Convergence time is not a function of geographic or network distance

E2E Effects



- Packet loss reaches 17% to 32% during convergence
- Latency increases by 50% during convergence period

BGP Convergence Model

- BGP explores path of increasing length

Stage	Routing Tables	Mesg Processing	Mesg queued
0	0(*R,1R,2R) 1(0R,*R,2R) 2(0R,1R,*R)		
1	R withdraws route 0(-,*1R,2R) 1(*0R,-,2R) 2(*0R,1R,-)	R->0 W R->1 W R->2 W	0->1 01R 1->0 01R 2->0 20R 0->2 01R 1->2 10R 2->1 20R
2	1 and 2 receive new annc from 0 0(-,*1R,2R) 1(-,-,*2R) 2(01R,*1R,-)	0->1 01R 0->2 01R	1->0 01R 2->0 20R 1->0 12R 2->0 21R 1->2 10R 2->1 20R 2->0 12R 2->1 21R
...			
48	Steady state 0(-,-,-) 1(-,-,-) 2(-,-,-)		

Upper bound on Convergence

- For a complete graph of n nodes there exist O((n-1)!) distinct paths to reach a particular destination
- Path vector algorithm attempts to find an alternate path by iterating on the available paths of equal or increasing length

Effect or MRAI

- Minimum Routing Advertisement Interval:
  - Send one advertisement per interval to (peer,prefix) pair
- Effect is that paths are strictly increasing
- So why is T_{down}, T_{long} slower that T_{up},T_{short}?
  - BGP does *Path Exploration* in the first case while not in the second case