BGP convergence

marcelo bagnulo

Causes for BGP changes

- Topology changes
  - Devices going up or down
  - New routers or sessions
- BGP session failures
  - Due to equipment failures, maintenance, etc.
  - Or, due to congestion on the physical path
- Changes in routing policy
  - Change in Local Pref
  - Reconfiguration of route filters

BGP Session operation

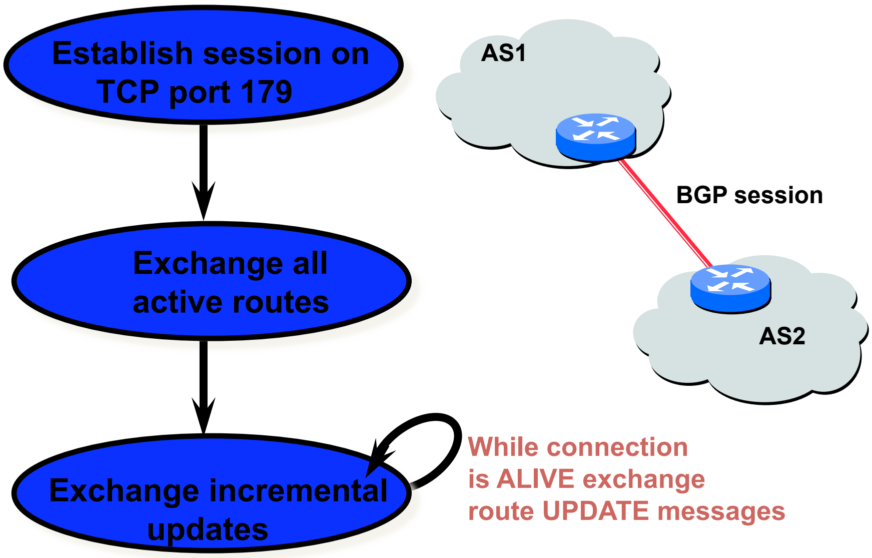
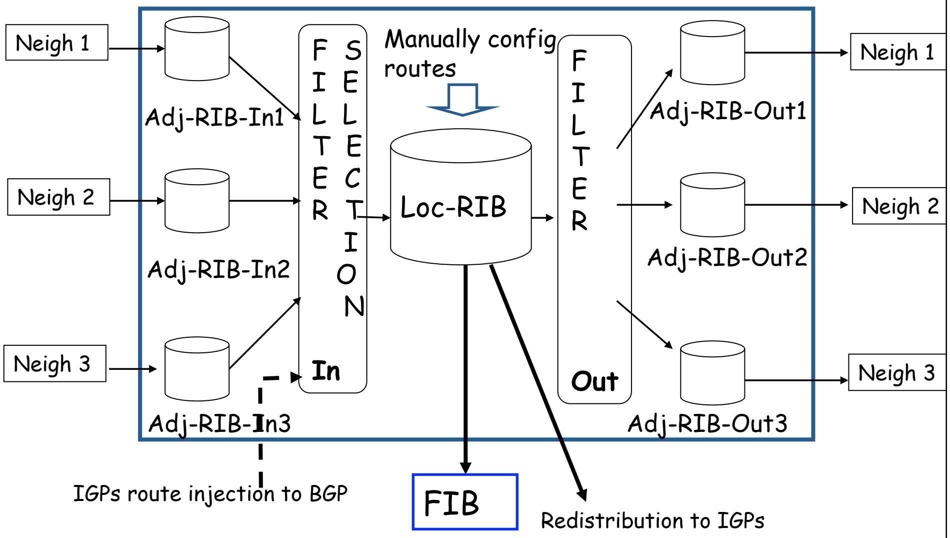


Figure from J. rexford

UPDATE messages

- An Update message can be
  - Announcement
    - Either a new prefix is announced
    - An exsiting prefix with a new attribute
      - Implicit withdraw: Exsiting route is replaced by another route
  - Withdraw
- Minimum route adverstisement interval timer (MRAI timer)
  - Minimum amount of time between two route adverstisement for the same prefix to a peer
  - Rate limts the UPDATES messages

BGP peer operation



BGP session failure

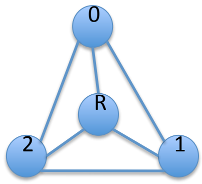
- BGP runs over TCP
  - BGP only sends updates when changes occur
  - TCP doesn't detect lost connectivity on its own
- Detecting a failure
  - Keep-alive: 60 seconds
  - Hold timer: 180 seconds
- Reacting to a failure
  - Discard all routes learned from the neighbor
  - Send new updates for any routes that change

BGP convergence

- DV protocols have delayed convergence
  - (In some cases) see count to infinity
- Path vector approach was supposed to solve the problem
  - Kind of an explicit spit horizon, since a router never use a route that contains its own AS on it
- However, measurements show that BGP has dleayed convergence as well… so?

BGP convergence model

- Assumptions
  - Each AS is a single node
  - Full mesh topology
    - Worst case
  - No filtering of routes
    - Worst case
  - No MRAI
  - FIFO ordering of messages
  - BGP processing as a single linear global queue



	Routing Tables	Msg Processing	Msgs Queued in system
0	Steady state 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 10R 2->0 20R 0->2 01R 1->2 10R 2->1 20R
2	1 and 2 receive update from 0 0(-,*1R,2R) 1(-,-,*2R) 2(01R,*1R,-)	0->1 01R 0->2 01R	1->0 10R 2->0 20R 1->0 12R 2->0 21R 1->2 10R 2->1 20R 1->2 12R 2->1 21R
3	0 and 2 receive update from 1 0(-,-,*2R) 1(-,-,*2R) 2(*01R,10R,-)	1->0 10R 1->2 10R	2->0 20R 1->0 12R 2->0 21R 0->1 02R 2->0 201R 2->1 20R 1->2 12R 2->1 21R 0->2 02R 2->1 201R
4	0 and 1 receive update from 2 0(-,-,-) 1(-,-,*20R) 2(*01R,10R,-)	2->0 20R 2->1 20R	1->0 12R 2->0 21R 0->1 02R 2->0 201R 0->1 W 1->0 120R 1->2 12R 2->1 21R 0->2 02R 2->1 201R 0->2 W 1->2 120R
5	0 and 2 receive update from 1 0(-,*12R,-) 1(-,-,*20R) 2(*01R,-,-)	1->0 12R 1->2 12R	2->0 21R 0->1 02R 2->0 201R 0->1 W 1->0 120R 0->1 012R 2->1 21R 0->2 02R 2->1 201R 0->2 W 1->2 120R 0->2 012R
..			
4	Steady state 0(-,-,-) 1(-,-,-) 2(-,-,-)		

Intuitive understanding

- Path vector approach prevents a node from reusing a route that contains its own AS in the path
  - Solves the count to infinity problem in RIP
- Does not prevent from learning a new invalid route from a heighbor
- Worst case: all will try all differnet AS paths
  - Different lenghts and different ASes in the path
- Exacerbates the counting problem
  - DV are striclty increasing
    - Only one paht is explored per path length
  - BGP is monotonically increasing
    - Multipla paths with the same path length are explored

Upper bound on convergence

- Observation 1: For a graph with n nodes, there are O((n-1)!) distinct path to reach a dst.
  - n-1 paths of length 1 to reach a dst (full mesh)
  - (n-2)(n-1) paths of length 2 to reach a dst
  - Total paths = (n-1) + (n-1)(n-2) + … + (n-1)! = O((n-1)!)
- Observation 2: When a route is withdraw, the path vector algorithm will try available path of equal or increasing path length (k-th iteration includes k edges)

Upper bound in convergence

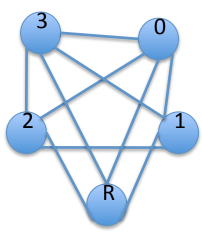
- Conditions for worst case convergence:
  - Full mesh
  - Messages are processed in sequence
  - Mesages are ordered so that the msg that invalidates current entry is processed first
    - More mesages result in updates being propagated.
    - Example node i (*013,103,-,-) and receives 1->i (1i3). The result is (*013,-,--) and no msg is propgated.
    - Example node i (*013,103,-,-) and receives 0->i (0i3). The result is (-,*103,--) and an update is progated
    - Conclusion: number of updates depends on the order (jn absence of MRAI)

Upper bound convergence

- Observation 4: If previous conditions apply, all possible paths will be explored. Once exhausted, the final withdraw will remove the route. Basis for conjecture the complexity is 0((n-1)!)
- Observation 5: number of messages is the number of states O((n-1)!) times the number of peers the update is announced i.e. (n-1)
  - Number of msg: (n-1)O((n-1)!)

Lower bound on convergence

- Can we make the process to be strictly increasing in the path length?
- We include the MRAI in the problem



	Routing tables	Msg processing	Msgs queued on system
0	Steady state 0(*R,1R,2R,3R) 1(0R,*R,2R,3R) 2(0R,1R,*R,3R) 3(0R,1R,2R,*R)		
1	R withdraws route 0(-,*1R,2R,3R) 1(*0R,-,2R,3R) 2(*0R,1R,-,3R) 3(*0R,1R,2R,-)	R->0 W R->3 W R->1 W R->2 W	0->1 01R 1->0 10R 2->0 20R 3->0 30R 0->2 01R 1->2 10R 2->1 20R 3->1 30R 0->3 01R 1->3 10R 2->3 20R 3->2 30R
2	Update from 0 to 1,2,3 0(-,*1R,2R,3R) 1(-,-,*2R,3R) 2(01R,*1R,-,3R) 3(01R,*1R,2R,-)	0->1 01R 0->2 01R 0->3 01R	1->0 10R 2->0 20R 3->0 30R 1->2 10R 2->1 20R 3->1 30R 1->3 10R 2->3 20R 3->2 30R
3	Update from 1 to 0,2 and 3 0(-,-,*2R,3R) 1(-,-,*2R,3R) 2(01R,10R,-,*3R) 3(01R,10R,*2R,-)	1->0 10R 1->2 10R 1->3 10R	2->0 20R 3->0 30R 2->1 20R 3->1 30R 2->3 20R 3->2 30R
4	Update from 2 to 0,1, and 3 0(-,-,-*3R) 1(-,-,20R,*3R) 2(01R,10R,-,*3R) 3(*01R,10R,20R,-)	2->0 20R 2->1 20R 2->3 20R	3->0 30R 3->1 30R 3->2 30R
5	Update from 3 to 0,1 and 2 0(-,-,-,-) 1(-,-,*20R,30R) 2(*01R,10R,-,30R) 3(*01R,10R,20R,-)	3->0 30R 3->1 30R 3->2 30R	
	MRAI expires		0->1 W 1->0 120R 2->0 201R 3->0 301R 0->2 W 1->2 120R 2->1 201R 3->1 301R 0->3 W 1->3 120R 2->3 201R 3->2 301R
6	Withdraw from 0 0(-,-,-,-) 1(-,-,*20R) 2(-,*10R,-,30R) 3(-,*10R,20R,-)	0->1 W 0->2 W 0->3 W	1->0 120R 2->0 201R 3->0 301R 1->2 120R 2->1 201R 3->1 301R 1->3 120R 2->3 201R 3->2 301R
…			
13	Steady state 0(-,-,-,-) 1(-,-,-,-) 2(-,-,-,-) 3(-,-,-,-)		

Lower bound on convergence

- Observation 1: In the best case, at the end of a MRAI round, at most one onde will have complete withdraw
  - All nodes except 0 will choose 0R and 0 choose 1R
  - Within MRAI 0 will receive updates for all (n-2) peers, reuslting in complete withdraw

Lower bound on covergence

- Observation 2:MRAI imposes monotonically increasing metric for succesive rounds
  - At the end of an MRAI round, only higher level paths will be announced
  - Each MRAI round, all nodes process the n-1 updates from other nodes before sending the new update
- Observation 3: convergence in n-1 MRAI rounds (only applies with current assumptions, i.e. Full mesh, no filtering)

Types of updates

- Destination becomes reachable
  - Switch from no path to a new path
- Better path becomes available
  - Switch from old path to new, better path

} lower delay

- Best path becomes unavailable
  - Switch from old path to new, worse path
- Destination becomes unreachable
  - Switch from old path to no path at all

} higher delay

More questions

- What is the right MRAI timer?
- How this behaves with other assumptions
  – Not full mesh
  – Policing

References

- C. Labovitz, A. Ahuja, A. Bose, Delayed Internet Routing Convergence, SIGCOMM 2000

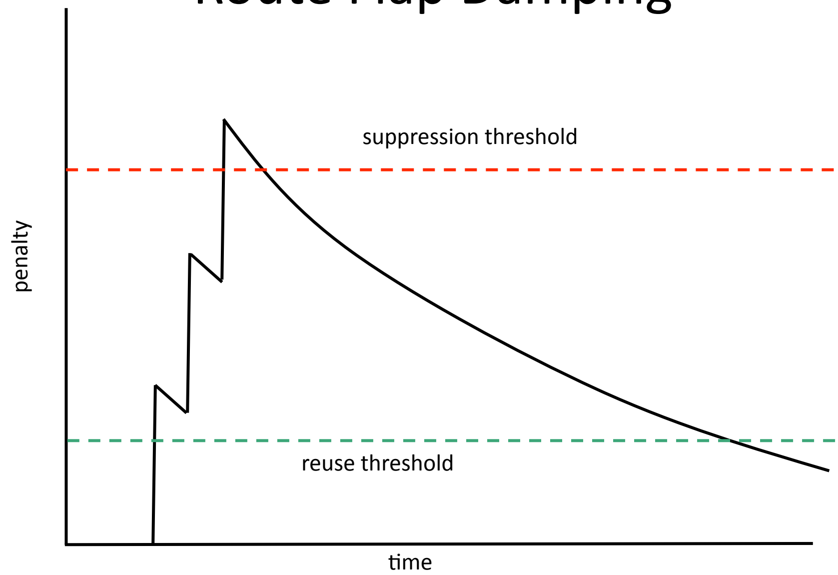
Route Flap Damping

- Mechanism to deal with route flaps
- Assumed caused of flaps
  - Router reconfiguration
  - Unstable links
- Result: additional BGP updates
  - More route computation, more work for routers
- MRAI suppress updates in short timescales
  - 30 secs
- Proposed solution: Route Flap Damping
  - Not consider routes that are flapping

Route Flap Damping

- RFD mechanism
- For each prefix and for each peer neighbor, the BGP router maintains penaly P(p,n)
  - If there is a change in the route announced by the peer, the penalty is increased (fixed)
  - P(p,n) decays exponentially P(p,n,t)=ke^(-at)
- If P(p,n) is higher than the suppression threshold, then is marked, included in Adj-RIB-In and not considered when calculating Loc-RIB
- P(p,n) continues being calculated and when its value is lower than Reuse threshold, the route is included in the Loc-RIB calculation

Route Flap Damping



Route Flap Damping

- Configurable Parameters
  - Suppression threshold
  - Reuse threshold
  - a – usually expressed as H half life i.e. The time for the penalty to decay to half of its value
- Reccomendations
  - More specific prefixes should be damped more aggresively
  - Routes should not damp until 4 flaps
  - Reccomended values such that a /20, min time of 10 min and max of 30 min

Usual values for commercial routers

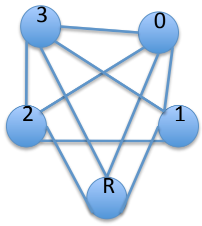
RFD parameter	Cisco	Juniper
Withdraw penalty	1000	1000
Reacvertisement penalty	0	1000
Attribute change penalty	500	500
Cutoff threshold	2000	3000
Half life (min)	15	15
Reuse threshold	750	750
Max suppress time (min)	60	60

Interaction with path hunting

- BGP model
  - Route selection based on AS path length
  - MRAI set to 30 secs
    - Not applies to withdraws
  - No sender side loop detection
- Msg propagation and delay negligible compared to MRAI

Withdrawal triggered suppression

- Consider the case a node R withdraws a route R and then announce it back.
- 5-node clicke topology
- Path hunting with MRAI will imply 4 rounds till convergence



Withdrawal triggered suppression

- 4 MRAI rounds account for 2 min
- Each round account for 500 penalty for the attribute change and 1000 for the withdraw
- Total penalty of 2500
  - Minus the decrease of the 2 min
  - In juniper, 3500 due the readvertisement
- In both cases, the value is higher than the cutoff threshold
- The route is damped for 15 min

Questions

- How to filter real flaps and allow path exploration needed for convergence?
  - Add more information in the updates
    - In path exploration, the different routes advertised are less and less preferred. We could identify path exploration through this. Note that don't need to be longer routers, due to local policy
  - Adjust the timers
  - Do other than supressing
  - Do something more clever than a timer

Reference

- Z. Morley Mao, R. Govindan, G. Varghese, R. Katz, Route Flap damping exacerbates Internet routing convergence, SIGCOMM 2002

Assignment

- The Impact of Internet Policy and Topology on Delayed Routing Convergence Craig Labovitz, Ahba Roger Wattenhofer, Srinivasan Venkatachary
- http://www.cs.ucsb.edu/~ravenben/papers/networking/labovitz01impact.pdf