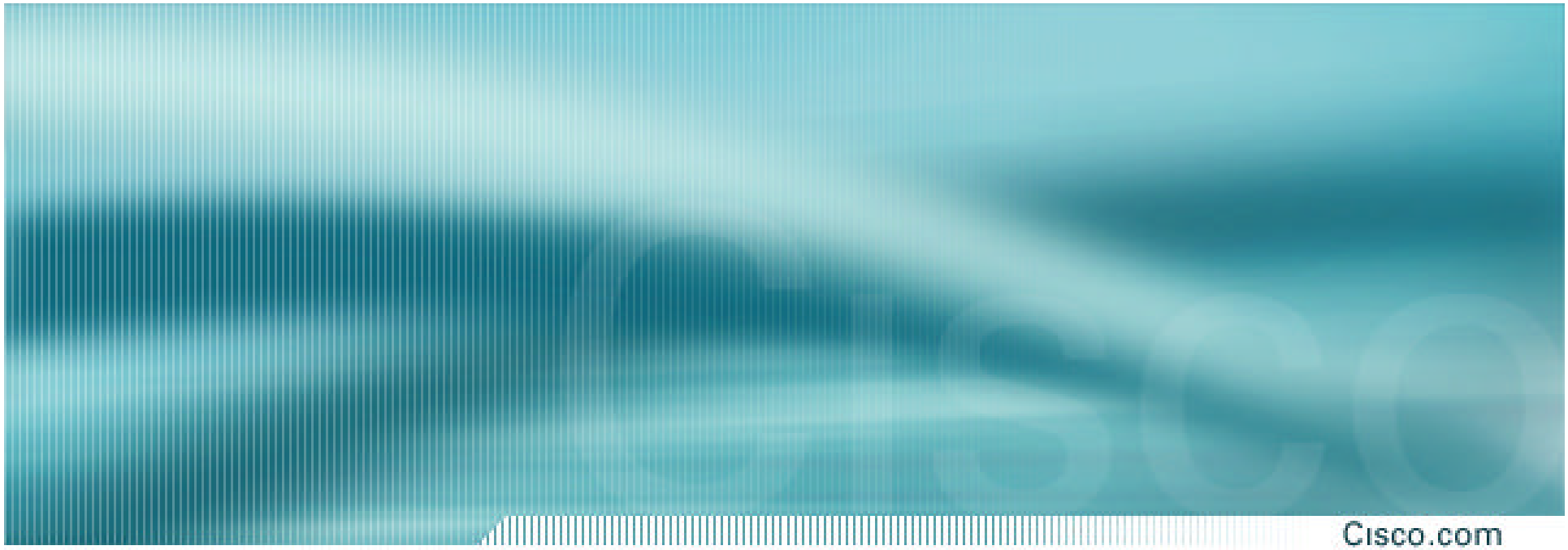


Day in the Life of a BGP Update in Cisco IOS

Philip Smith

Routing WG, RIPE 45, Barcelona

May 2003

Agenda

Cisco.com

- **Presentation tracks a BGP update through a router running Cisco IOS**

Basic BGP Processing Queues…

- **UPDATE Arrives from the peer**
  - **Goes to input hold queue**
  - **Then to tcp queue for right socket on the Route Processor (RP)**
- **BGP I/O process on RP reads from TCP sockets and puts messages on BGP InQ**
- **The BGP InQ adds the packet to the input queue for the appropriate peer**

Basic BGP Processing
BGP Process…

- **Unpacks the prefixes from the UPDATEs**
- **Begins best path computation:**

  Path is run through configured incoming filters:

  1. **miscellaneous checks (AS loop etc)**
  2. **filter-list**
  3. **route-map**
  4. **prefix-list OR distribute-list**

  Mutually exclusive

  Path is either discarded, or has attributes modified, or remains unchanged

Basic BGP Processing
Path Not Discarded (1)…

- Compared with other paths for the same prefix
- Best path chosen according to the IOS best path decision algorithm at

  www.cisco.com/warp/public/459/25.shtml

- Bestpath for a prefix marked as such in BGP database and sent to RIB for insertion

  NB: Existing path in RIB with lower admin distance means that best path will NOT be inserted

Basic BGP Processing
Path Not Discarded (2)…

- **Insertion into RIB triggers insertion into FIB**

  New entry in FIB is propagated to the LineCards FIB etc

- **Once the path is sent to the RIB, it is flagged that it needs to be sent to peers/peer-group in a BGP update**

Basic BGP Processing
Update Our Peers…

- **When ready to update our peers with our routes:**
  - **Walk BGP table once per peer (or peer-group) looking for bestpaths**
  - **Bestpaths sent through outbound filters:**
    1. **prefix-list OR distribute-list**
    2. **ORF**
    3. **Miscellaneous (AS loop, no-export, etc)**
    4. **filter-list**
    5. **route-map OR unsuppress-map**
    6. **advertise-map (conditional advertisement)**

    Mutually exclusive
  - **Path is either discarded, or has attributes modified, or remains unchanged**

Basic BGP Processing
Forming & Sending Updates…

- Create a new update for each bestpath with a unique attribute combination
- Once reached the end of the walk (or limited by available memory), the updates are enqueued on the peer's output queue
- For peer-groups, the update is reused for each member of the peer-group
- Sending Updates
  - Updates are packed into TCP messages
  - Sent to appropriate BGP peers
    - BGP packets have IP precedence 6

Recent Changes

- **Up to 12.0(18)S:**

  Update algorithm (Normal mode) queued updates every 500 updates computed – could lead to slow convergence

- **12.0(18)S1 onwards:**

  Added new algorithm to aid faster convergence (Init mode – ignores 500 update limit)

  Aggressive on memory – selected depending on available router memory

- **12.0(21)S1 onwards:**

  Limits added to BGP's memory usage to ensure that CEF and BGP live together more peacefully

- **12.0(22)S onwards:**

  Init mode and normal mode replaced with a new algorithm which offers fast convergence but better safeguards against excessive memory usage

Details
BGP Scanner

- **BGP Scanner**
  - **This runs every 60 seconds**
  - **Timer can be changed if desired:**
    - **bgp scan-time <5-60> – units in seconds**
- **Multiprotocol BGP**
  - **Adds Import Scanner – runs every 15 seconds**
  - **Only applies to VPNv4 AF**
  - **Timer can be changed if desired:**
    - **bgp scan-time import <5-60> – units in seconds**

Details
BGP Scanner – What does it do??

- **Housekeeping!**

  Evaluates redistribution and network statements

  Conditional advertising

  Route flap damping clean up (delete old & history entries)

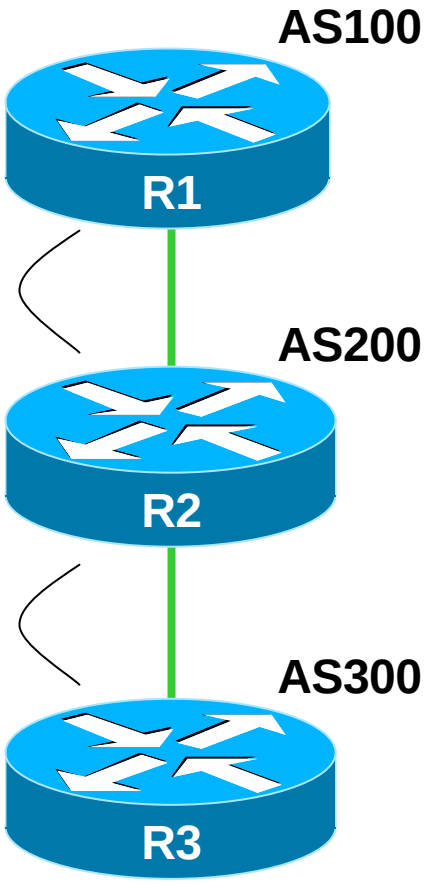
  And most importantly of all:

- **Checks validity of all entries in BGP table**

  Invalid next hop results in attempt to recompute alternative path, update to peers, etc

Very Simple Walk Through…

- R1 originates 2.0.0.0/8 into BGP
- Track the propagation of the prefix from R1 through R2 to R3



Walk Through
R1 adds configuration for 2.0.0.0/8

Operator configures prefix

```
R1(config-router)#net 2.0.0.0
10:34:48.350 BGP: Import timer expired. Walking from 1 to 1
10:35:03.350 BGP: Import timer expired. Walking from 1 to 1
10:35:18.350 BGP: Import timer expired. Walking from 1 to 1
10:35:33.350 BGP: Performing BGP general scanning
10:35:33.350 BGP(0): scanning IPv4 Unicast routing tables
10:35:33.358 BGP(IPv4 Unicast): Performing BGP Nexthop scanning for general scan
10:35:33.358 BGP(1): scanning IPv6 Unicast routing tables
10:35:33.358 BGP(IPv6 Unicast): Performing BGP Nexthop scanning for general scan
10:35:33.362 BGP(2): scanning VPNv4 Unicast routing tables
10:35:33.362 BGP(VPNv4 Unicast): Performing BGP Nexthop scanning for general scan
10:35:33.362 BGP(3): scanning IPv4 Multicast routing tables
10:35:33.370 BGP(IPv4 Multicast): Performing BGP Nexthop scanning for general scan
```

BGP scanner runs

Walk Through
R1 Scanner does its housekeeping

New net entry spotted

10:35:33.378 BGP(0): route 2.0.0.0/8 up

10:35:33.378 BGP(0): nettable_walker 2.0.0.0/8 route sourced locally

10:35:33.378 BGP(0): 192.168.4.129 computing updates, afi 0, neighbor version 64, table version 65, starting at 0.0.0.0

10:35:33.378 BGP(0): 192.168.4.129 send UPDATE (format) 12.0.0.0/8, next 192.168.4.133, metric 0, path

10:35:33.378 BGP(0): 192.168.4.129 1 updates enqueued (average=60, maximum=60)

note time!

10:35:33.378 BGP(0): 192.168.4.129 update run completed, afi 0, ran for 0ms, neighbor version 64, start version 65, throttled to 65

Update created and sent

10:35:48.378 BGP: Import timer expired. Walking from 1 to 1

10:36:03.379 BGP: Import timer expired. Walking from 1 to 1

10:36:18.379 BGP: Import timer expired. Walking from 1 to 1

10:36:33.379 BGP: Performing BGP general scanning

10:36:33.379 BGP(0): scanning IPv4 Unicast routing tables

10:36:33.387 BGP(IPv4 Unicast): Performing BGP Nexthop scanning for general scan

Walk Through
What happens on R2?

10:35:26.327 BGP: Import timer expired. Walking from 1 to 1

10:35:33.391 BGP(0): 192.168.4.133 rcvd UPDATE w/ attr: nexthop 192.168.4.133, origin i, metric 0, path 100

Compare with 10:35:33.378

10:35:33.391 BGP(0): 192.168.4.133 rcvd 2.0.0.0/8

10:35:33.391 BGP(0): Revise route installing 1 of 1 route for 2.0.0.0/8 -> 192.168.4.133(main) to main IP table

10:35:33.391 RT: add 2.0.0.0/8 via 192.168.4.133, bgp metric [20/0]

10:35:34.491 BGP(0): 192.168.9.13 computing updates, afi 0, neighbor version 33, table version 34, starting at 0.0.0.0

10:35:34.491 BGP(0): 192.168.9.13 starting fmt at 33/0 0.0.0.0, cache 0 0, abort = 1

10:35:34.491 BGP(0): 192.168.9.13 send UPDATE (format) 2.0.0.0/8, next 192.168.9.14, metric 0, path 100

10:35:34.491 BGP(0): 1 updates (average = 55, maximum = 55)

10:35:34.491 BGP(0): Set update-group 1 enq_version to 34 192.168.9.13

10:35:34.491 BGP(0): 192.168.9.13 fmt 1, enq Done vers set to 34/0 34/0 0.0.0.0

note time!

10:35:34.491 BGP(0): 192.168.9.13 update run completed, afi 0, ran for 0ms, neighbor version 34, start version 34, throttled to 34

10:35:34.491 BGP(0) : Initial update finished for 192.168.9.13

10:35:34.491 BGP(0): 1 done. vers 34 34, tbl 34 init 34

Walk Through
More from R2

10:35:34.491 BGP(0): 192.168.4.133 computing updates, afi 0, neighbor version 33, table version 34, starting at 0.0.0.0

10:35:34.491 BGP(0): 192.168.4.133 starting fmt at 33/0 0.0.0.0, cache 0 0, abort = 1

10:35:34.491 BGP(0): Set update-group 2 enq_version to 34

10:35:34.491 BGP(0): 192.168.4.133 fmt 0, enq Done vers set to 34/0 34/0 0.0.0.0

10:35:34.491 BGP(0): 192.168.4.133 update run completed, afi 0, ran for 0ms, neighbor version 34, start version 34, throttled to 34

10:35:34.491 BGP(0) : Initial update finished for 192.168.4.133

10:35:34.491 BGP(0): 2 done. vers 34 34, tbl 34 init 34

10:35:34.491 BGP(0): Reset update-group 1 versions from 34 34 to 0 0

10:35:34.491 BGP(0): Reset update-group 2 versions from 34 34 to 0 0

We get update from R3 too – 12ms to process!

10:35:34.503 BGP(0): 192.168.9.13 rcv UPDATE w/ attr: nexthop 192.168.9.13, origin i, originator 0.0.0.0, path 300 200 100, community , extended community

10:35:34.503 BGP(0): 192.168.9.13 rcv UPDATE about 2.0.0.0/8 -- DENIED due to: AS-PATH contains our own AS;

10:35:41.327 BGP: Import timer expired. Walking from 1 to 1

Walk Through
What happens on R3?

Compare with 10:35:33.378 from R1 and 10:35:34.491 from R2

10:35:27.910 BGP: Import timer expired. Walking from 1 to 1

10:35:34.494 BGP(0): 192.168.9.14 rcvd UPDATE w/ attr: nexthop 192.168.9.14, origin i, path 200 100

10:35:34.498 BGP(0): 192.168.9.14 rcvd 2.0.0.0/8

10:35:34.498 BGP(0): Revise route installing 1 of 1 route for 2.0.0.0/8 -> 192.168.9.14 to main IP table

10:35:34.498 RT: add 2.0.0.0/8 via 192.168.9.14, bgp metric [20/0]

10:35:34.498 BGP(0): 192.168.9.14 computing updates, afi 0, neighbor version 132, table version 134, starting at 0.0.0.0

10:35:34.498 BGP(0): 192.168.9.14 send UPDATE (format) 2.0.0.0/8, next 192.168.9.13, metric 0, path 200 100

10:35:34.502 BGP(0): 192.168.9.14 1 updates enqueued (average=57, maximum=57)

10:35:34.502 BGP(0): 192.168.9.14 update run completed, afi 0, ran for 4ms, neighbor version 132, start version 134, throttled to 134

10:35:42.910 BGP: Import timer expired. Walking from 1 to 1

10:35:57.910 BGP: Import timer expired. Walking from 1 to 1

Walk Through
Operator removes 2.0.0.0/8 from R1

Operator removes prefix

R1(config-router)#no net 2.0.0.0

10:38:18.435 BGP: Import timer expired. Walking from 1 to 1

BGP scanner runs

10:38:33.435 BGP: Performing BGP general scanning

10:38:33.435 BGP(0): scanning IPv4 Unicast routing tables

10:38:33.443 BGP(IPv4 Unicast): Performing BGP Nexthop scanning for general scan

2/8 gone!

10:38:33.443 BGP(0): nettable_scan: invalidate local path for 2.0.0.0/8

10:38:33.443 BGP(0): nettable_scan: invalidate sourced path for 2.0.0.0/8

10:38:33.443 BGP(0): no valid path for 2.0.0.0/8

10:38:33.447 BGP(1): scanning IPv6 Unicast routing tables

10:38:33.447 BGP(IPv6 Unicast): Performing BGP Nexthop scanning for general scan

10:38:33.447 BGP(2): scanning VPNv4 Unicast routing tables

10:38:33.447 BGP(VPNv4 Unicast): Performing BGP Nexthop scanning for general scan

10:38:33.447 BGP(3): scanning IPv4 Multicast routing tables

10:38:33.455 BGP(IPv4 Multicast): Performing BGP Nexthop scanning for general scan

Walk Through
R1 Scanner does its housekeeping

Cisco.com

10:38:33.463 BGP(0): nettable_walker 2.0.0.0/8 no best path

10:38:33.463 BGP(0): 192.168.4.129 computing updates, afi 0, neighbor version 65, table version 66, starting at 0.0.0.0

10:38:33.463 BGP(0): 192.168.4.129 send unreachable 2.0.0.0/8

10:38:33.463 BGP(0): 192.168.4.129 send UPDATE 2.0.0.0/8 – unreachable

10:38:33.463 BGP(0): 192.168.4.129 1 updates enqueued (average=25, maximum=25)

10:38:33.463 BGP(0): 192.168.4.129 update run completed, afi 0, ran for 0ms, neighbor version 65, start version 66, throttled to 66

note time!

10:38:48.463 BGP: Import timer expired. Walking from 1 to 1

10:39:03.463 BGP: Import timer expired. Walking from 1 to 1

10:39:18.463 BGP: Import timer expired. Walking from 1 to 1

Update created and sent

Walk Through
What happens on R2?

Compare with 10:38:33.463

10:38:26.376 BGP: Import timer expired. Walking from 1 to 1

10:38:33.476 BGP(0): 192.168.4.133 rcv UPDATE about 2.0.0.0/8 – withdrawn

10:38:33.476 BGP(0): no valid path for 2.0.0.0/8

10:38:33.476 BGP(0): nettable_walker 2.0.0.0/8 no best path

10:38:33.476 RT: del 2.0.0.0 via 192.168.4.133, bgp metric [20/0]

10:38:33.476 RT: delete network route to 2.0.0.0

10:38:34.576 BGP(0): 192.168.9.13 computing updates, afi 0, neighbor version 34, table version 35, starting at 0.0.0.0

10:38:34.576 BGP(0): 192.168.9.13 starting fmt at 34/0 0.0.0.0, cache 0 0, abort = 1

10:38:34.576 BGP(0): 192.168.9.13 send unreachable 2.0.0.0/8

10:38:34.576 BGP(0): 192.168.9.13 send UPDATE 2.0.0.0/8 – unreachable

10:38:34.576 BGP(0): 1 updates (average = 25, maximum = 25)

10:38:34.576 BGP(0): Set update-group 1 enq_version to 35 192.168.9.13

10:38:34.576 BGP(0): 192.168.9.13 fmt 1, enq Done vers set to 35/0 35/0 0.0.0.0

note time!

10:38:34.576 BGP(0): 192.168.9.13 update run completed, afi 0, ran for 0ms, neighbor version 35, start version 35, throttled to 35

10:38:34.576 BGP(0) : Initial update finished for 192.168.9.13

10:38:34.576 BGP(0): 1 done. vers 35 35, tbl 35 init 35

Walk Through
More from R2

10:38:34.576 BGP(0): 192.168.4.133 computing updates, afi 0, neighbor version 34, table version 35, starting at 0.0.0.0

10:38:34.576 BGP(0): 192.168.4.133 starting fmt at 34/0 0.0.0.0, cache 0 0, abort = 1

10:38:34.576 BGP(0): Set update-group 2 enq_version to 35

10:38:34.576 BGP(0): 192.168.4.133 fmt 0, enq Done vers set to 35/0 35/0 0.0.0.0

10:38:34.576 BGP(0): 192.168.4.133 update run completed, afi 0, ran for 0ms, neighbor version 35, start version 35, throttled to 35

10:38:34.576 BGP(0) : Initial update finished for 192.168.4.133

10:38:34.576 BGP(0): 2 done. vers 35 35, tbl 35 init 35

10:38:34.576 BGP(0): Reset update-group 1 versions from 35 35 to 0 0

10:38:34.576 BGP(0): Reset update-group 2 versions from 35 35 to 0 0

10:38:34.584 BGP(0): 192.168.9.13 rcv UPDATE about 2.0.0.0/8 – withdrawn

10:38:41.376 BGP: Import timer expired. Walking from 1 to 1

Walk Through
What happens on R3?

Compare with 10:38:33.476 from R1 and 10:38:34.576 from R2

10:38:27.959 BGP: Import timer expired. Walking from 1 to 1

10:38:34.579 BGP(0): 192.168.9.14 rcv UPDATE about 2.0.0.0/8 – withdrawn

10:38:34.579 BGP(0): no valid path for 2.0.0.0/8

10:38:34.579 BGP(0): nettable_walker 2.0.0.0/8 no best path

10:38:34.579 RT: del 2.0.0.0 via 192.168.9.14, bgp metric [20/0]

10:38:34.579 RT: delete network route to 2.0.0.0

10:38:34.583 BGP(0): 192.168.9.14 computing updates, afi 0, neighbor version 134, table version 136, starting at 0.0.0.0

10:38:34.583 BGP(0): 192.168.9.14 send unreachable 2.0.0.0/8

10:38:34.583 BGP(0): 192.168.9.14 send UPDATE 2.0.0.0/8 – unreachable

10:38:34.583 BGP(0): 192.168.9.14 1 updates enqueued (average=25, maximum=25)

10:38:34.583 BGP(0): 192.168.9.14 update run completed, afi 0, ran for 0ms, neighbor version 134, start version 136, throttled to 136

10:38:42.959 BGP: Import timer expired. Walking from 1 to 1

10:38:57.959 BGP: Import timer expired. Walking from 1 to 1

10:39:12.959 BGP: Performing BGP general scanning

Walk Through Summary

- **Update coming into IOS router is processed immediately, and sent onwards if appropriate:**
  - **Time required depends on number of prefixes, speed of processor, etc**
  - **General case dependent on “advertisement interval”**
- **Locally configured routes await the BGP scanner…**
  - **…when BGP network statement exists, and route is introduced by static, connected, IGP, etc**

Walk Through
Flap Damping on R2 and R3

- After 3 withdraw and announce of 2.0.0.0/8
- R1

  16:38:42.236 BGP(0): route 2.0.0.0/8 up

  16:38:42.236 BGP(0): nettable_walker 2.0.0.0/8 route sourced locally

  16:38:43.236 BGP(0): 192.168.4.129 computing updates, afi 0, neighbor version 87, table version 88, starting at 0.0.0.0

  16:38:43.236 BGP(0): 192.168.4.129 send UPDATE (format) 2.0.0.0/8, next 192.168.4.133, metric 0, path

  16:38:43.236 BGP(0): 192.168.4.129 1 updates enqueued (average=60, maximum=60)

  16:38:43.236 BGP(0): 192.168.4.129 update run completed, afi 0, ran for 0ms, neighbor version 87, start version 88, throttled to 88

- R2

  16:38:43.243 BGP(0): 192.168.4.133 rcvd UPDATE w/ attr: nexthop 192.168.4.133, origin i, metric 0, path 100

  16:38:43.247 BGP(0): 192.168.4.133 rcvd 2.0.0.0/8

  16:38:43.247 BGP(0): no valid path for 2.0.0.0/8

Walk Through
Flap Damping on R2 and R3

- R2 sees three flaps from R1, so prefix is suppressed when the update arrives on the router

  BGP scanner only updates the flap damping counters

  BGP Process handles whether the prefix is suppressed or not as the update arrives

- R3 sees no update – R2 suppresses the announcement:

```
R2#sh ip bgp damp flap
   Network          From            Flaps Duration Reuse    Path
*d 2.0.0.0          192.168.4.133   3     00:05:29 00:04:39 100
R3#sh ip bgp damp flap
   Network          From            Flaps Duration Reuse    Path
 h 2.0.0.0          192.168.9.14    3     00:05:36          200 100
```

Flap Damping Summary

- **If prefix is flapping**

  Updates are suppressed once suppress limit reached – done by the BGP Process

- **If prefix is not flapping**

  Updates sent onwards immediately after they are heard from the neighbour and processed

  General case dependent on “advertisement interval”

Other timers: Update Delay

- **IOS has a timer to handle graceful startup of BGP peerings:**
- **Update-delay:**

  bgp update-delay <1-3600> – default is 120 seconds

  Determines how long BGP process waits before computing bestpaths, updating the routing table, and sending updates

Other timers: Advertisement Interval

- **Advertisement interval:**

  neighbor x.x.x.x min-advertise <1-3600>

  eBGP defaults to 30 seconds

  iBGP and eiBGP defaults to 5 seconds

  This value is the amount of time the router will wait before it generates more updates relating to a prefix change

Other timers: Advertisement Interval (cont)

- **Example of a BGP Update**
  - **Walk BGP table & Detect Changes**
  - **Generate Update & send the packets**
  - **Start minimum adv interval timer when finished**
  - **Walk the table again**
  - **If any changes detected, DO NOTHING!**
  - **We wait until minimum adv interval timer is finished before sending any more updates**

Summary

- **Prefix update propagation delay in IOS is purely due to the speed of the processor and the number of updates to be handled – plus:**
  - **Min-advertise interval prevents “churn”**
  - **Update-delay waits until BGP process is running before generating updates**
- **Examples for this presentation were tested on 12.2(11)T, 12.0(24)S and 12.2(14)S**