BGP ADVANCED FEATURES AND EXTENSIONS

Tawfiq Khan
TCOM 610
George Mason University

BGP Policy Update

- BGP is stateful and built on top of TCP
- BGP only exchanges full routing table at the beginning and incremental updates afterwards

- What if you or your BGP neighbor change routing policy over an established sessions?
  - Deny some of the existing routes
  - Raise local-pref for some routes
  - Attach new community for some routes
- Old Approach: you have to clear the BGP sessions and re-establish session from the beginning again for the new policy to take effect
  - Cause network BGP instability
  - Increase router load
  - When is best time to clear the session? – Time-zone difference for global corporations

BGP soft reset

- Hard reset invalidates the cache and results in a negative impact on the network performance when the information in the cache becomes unavailable. A hard reset is also disruptive because active BGP sessions are torn down.
- A soft reset, which is performed on a per-neighbor basis, does not clear the BGP session and facilitates the application of new policies. There are two methods of performing a soft reset:
  - A dynamic inbound soft reset is used to generate inbound updates from a neighbor.
  - An outbound soft reset is used to send a new set of updates to a neighbor.
  - both BGP peers must support the soft route refresh capability

Soft reset commands

clear ip bgp * | *ip-address* | *peer-group-name* | *peer-ASN* **soft in**

clear ip bgp * | *address* | *peer-group-name* | *peer-ASN* **soft out**

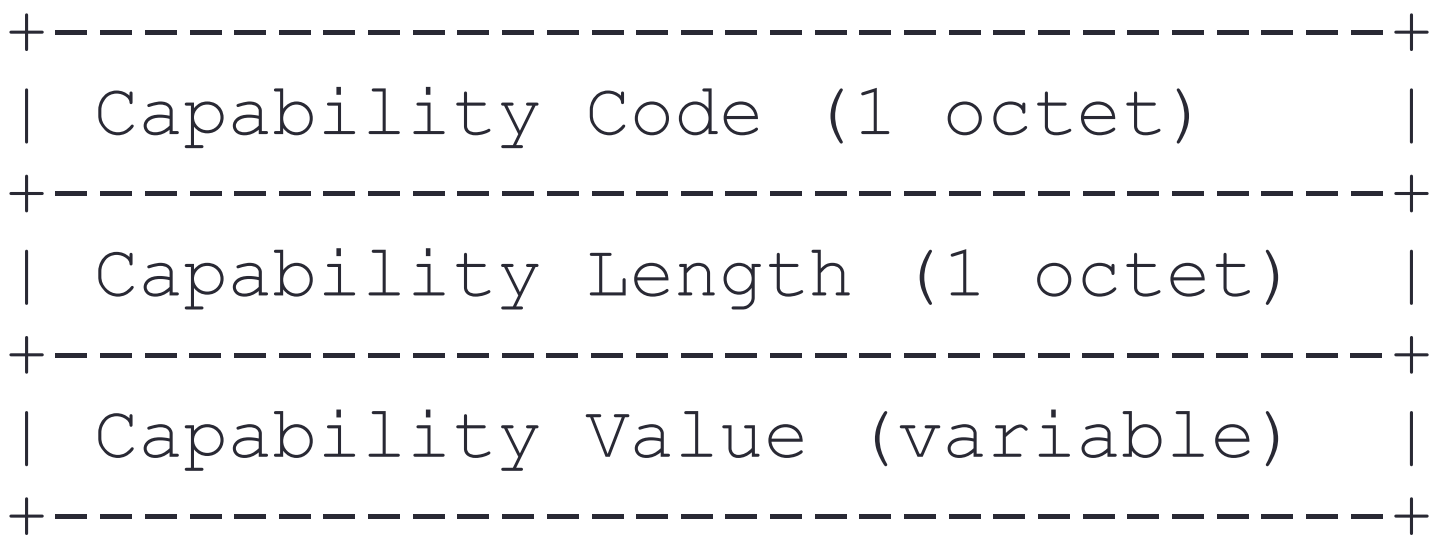
neighbor *ip-address* | *peer-group-name* soft-reconfiguration inbound

```
router bgp 100
 neighbor 131.108.1.1 remote-as 200
 neighbor 131.108.1.1 soft-reconfiguration inbound
```

BGP Capability Negotiations

- BGP speaker receives an OPEN message with one or more unrecognized Optional Parameters, the speaker must terminate BGP peering
- MAY include an Optional Parameter, called Capabilities.
- Optional Parameter carried by the OPEN message
- If not supported, may terminate peering or establish without the optional capability
- Known Capability parameters: **MP-BGP extension (AFI/SAFI), Route Refresh, Outbound Route Filtering**

Capability Parameters



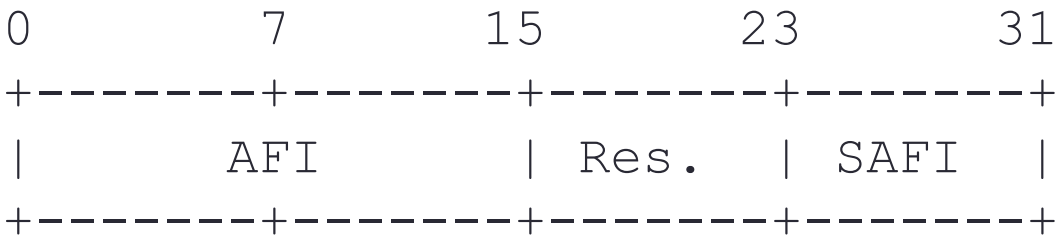
new Error Subcode – Unsupported Capability (7)

Route-refresh capability

- BGP speaker uses BGP Capabilities Advertisement [BGP-CAP]. This capability is advertised using the Capability code 2 and Capability length 0.
- BGP speaker that is capable of receiving and properly handling the ROUTE-REFRESH message should advertise the Route Refresh Capability to the peer using BGP Capabilities advertisement
- BGP speaker may send a ROUTE-REFRESH message to its peer only if it has received the Route Refresh Capability from its peer
- BGP speaker shall re-advertise to that peer the Adj-RIB-Out of the <AFI, SAFI> carried in the message, based on its outbound route filtering policy.

Route-refresh messages –RFC2918

Type: 5 - ROUTE-REFRESH
Message Format: One <AFI, SAFI> encoded as



AFI - Address Family Identifier (16 bit).
Res. - Reserved (8 bit) field. Should be set to 0 by the sender and ignored by the receiver.
SAFI - Subsequent Address Family Identifier (8 bit).

Outbound Route Filtering

- Common to use inbound policy to filter out unwanted routes
  - Route transmission takes CPU and memory
  - Internet table is growing fast
- Define a way for BGP peer to send a set of outbound filter to its neighbor
- Its neighbor to combine its local outbound filter and peer’s outbound filter to save the transmission of unwanted routes

ORF

- ORF entry is a tuple of the form <AFI/SAFI, ORF-Type, Action, Match, ORF-value>
- ORF-type: 64-Prefix, Ext-community-based?
- Action can be one of ADD, REMOVE, REMOVE-ALL
- Match" component can be one of PERMIT or DENY
- ORF entries are carried in the BGP ROUTE-REFRESH message
- One Route-REFRESH can carry multiple ORF entries in the same AFI/SAFI

| Address Family Identifier (2 octets)
| Reserved (1 octet)
| Subsequent Address Family Identifier (1 octet)
| When-to-refresh (1 octet) (IMMEDIATE|DEFER)
| ORF Type (1 octet)
| Length of ORFs (2 octets) (LEN of ORF in this type)
| First ORF entry (variable)
| N-th ORF entry (variable)
| ORF Type (1 octet)
| Length of ORFs (2 octets)
| First ORF entry (variable)
| N-th ORF entry (variable)

Route-refresh with ORF entries

| Action (2 bit)
| Match (1 bit)
| Reserved (5 bits)
| Type specific part (variable)

ORF Entry Encoding

ORF operations

- BGP peer to advertise ORF capacity to its peer about its capacity for send/receive ORF type
- Upon receiving of Route-refresh message from it peer, BGP speaker combine ORF-entries and re-advertise routes from the Adj-RIB-Out for AFI/SAFI
- ORF entries of different types are the logical AND of all types in AFI/SAFI
- Lifetime of ORF is the duration of BGP session

Max-prefix

- Configure max-prefix limit from a peer to protect accidental mis-configuration or malicious attacks
- **neighbor** {*ip-address* | *peer-group-name*} {**maximum-prefix** *maximum* [*threshold*]} [**restart** *restart-interval-mins*] [**warning-only**]
- Default threshold: 75% of the max-prefix limit
- Router will try to re-establish BGP session after the configured minutes. Default: to stay down forever
- You may want to enable warning-only to notify engineers to correct the problem rather than shutting down the session

Max-prefix example

```
router bgp 101
 neighbor 192.168.6.6 maximum-prefix 500 warning-only
```

```
router bgp 101
  neighbor 192.168.6.6 maximum-prefix 2000 restart 30
```

BGP dampening (RFC2439)

- Provide a mechanism capable of reducing router processing load caused by instability
  - the BGP decision process and adding/removing forwarding entries
- To prevent sustained routing oscillations without sacrificing route convergence time for generally well behaved routes
- Two approaches: use timer adjustment or route damping

BGP Timer Approach

- Timer approach:
  - no space overhead
  - slow down every route's convergence (even good behaved routes)
  - Need very long timer (minutes or hours) to effectively reduce flapping route churn
- Timers associated with BGP route advertisement:
  - MinRouteAdvertisementInterval timer: 30 seconds
  - MinASOriginationInterval timer: 15 seconds
  - Short timers combined with route damping and BGP update packing provides good performance

Route Instability

- BGP specification originally cited fixed timers as a method to control frequent route changes and to better pack updates
  - MinRouteAdvertisementInterval - 30 seconds
  - MinASOriginationInterval - 15 seconds
- This had the bad side effect of delaying convergence, and also relied upon your peer to do the right thing
  - neighbor (*ip-address* | *peer-group-name*) advertisementinterval *seconds*
  - default is 30 seconds for EBGP, 5 seconds for IBGP

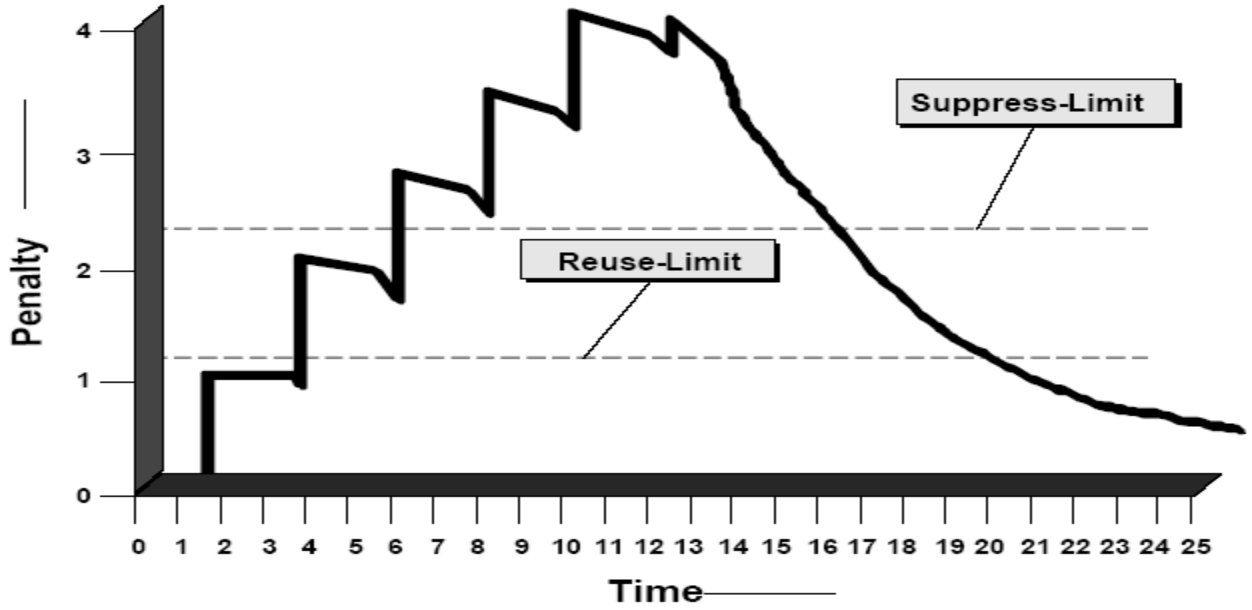
BGP Route Flap & Damping

- Keep state info for every route
- Require more space per route and more processing overhead
- How to distinguish well-behaved route and flapping routes?
  - Tolerable to wait for packing if large amount of updates received in a short time
  - Hold time for flapping route should be proportional to the expectation of the route's future stability
  - Use a route's history to predicate its future behavior
  - If a route with stable history make some quick transitions – it is okay; but if for extended time, it should be suppressed

Route damping algorithm

- Each route is assigned a figure of merit
- The merit is increased for each time that route flaps
- The routes with high value of merit over threshold are suppressed
- The merit is decreased exponentially (and by a computationally efficient algorithm)
  - Control by half-life time
- If the merit is down to certain reuse threshold, the flapped routes will be reused again

BGP Damping Parameters



Parameters

- Parameters
  - cutoff threshold
  - reuse threshold
  - maximum hold down time
  - decay half life while reachable
  - decay half life while unreachable
  - decay memory limit (used to limit internal array sizes)

Default Parameters

- Defaults
  - penalty - default 1000
  - half-life - default 15 minutes, the number of minutes it takes for the penalty to decay by 1/2
  - suppress limit - default 2000, If a route is suppressed when the penalty reaches this value
  - reuse limit - default 750, If a route is suppressed the penalty must decay to this value to be unsuppressed
  - maximum suppress time (mins) - default 4 x half-life, The maximum number of minutes a route may be suppressed
- Example
  - bgp dampening 20 1000 10000 150

Damping Parameters

- Calculated parameters:
  - max-penalty – The maximum penalty a route may have that will allow the penalty to decay to reuse-limit within max-suppress-time
  - max-penalty = reuse-limit * 2^(max-suppress-time/half-life)
  - *If half-life is 30, reuse-limit is 800, and max-suppress-time is 60 then the max-penalty would be 3200. If we allowed the penalty to reach 3201 it would be impossible for the penalty to decay to 800 within 60 minutes.*
  - *IOS will generate a warning message if the max-penalty is above 20,000 or less than the suppress-limit.*

Configuring BGP Damping

router bgp 1000

bgp dampening [*half-life reuse suppress max-suppresstime*] [route-map *map*]

for global BGP route damping parameters

set dampening (in route-map configuration, to selectively apply dampening

show ip bgp dampened-paths
show ip bgp flap-statistics
clear ip bgp dampening (in CLI)
clear ip bgp flap-statistics

BGP Damping

- A route can only be suppressed when receiving an advertisement. Not when receiving a WITHDRAW.
- Attribute changes count as a half flap
- In order for a route to be suppressed the following must be true:
  - The penalty must be greater than the suppress-limit
  - An advertisement for the route must be received while the penalty is greater than the suppress-limit
  - *A route will not automatically be suppressed if the suppresslimit is 1000 and the penalty reaches 1200. The route will only be suppressed if an advertisement is received while the penalty is decaying from 1200 down to 1000.*

Damping Example

- Small suppress window:
  - Half-life of 30 minutes, reuse-limit of 800, suppress limit of 3000, and max-suppress-time of 60 max-penalty is 3200
- Advertisement must be received while penalty is decaying from 3200 down to 3000 for the route to be suppressed
- A 3 min 45 second (rough numbers) window exist for an advertisement to be received while decaying from 3200 to 3000.

Example II

- No window:
  - Half-life of 30 minutes, reuse-limit of 750, suppress limit of 3000, and max-suppress-time of 60
  - max-penalty = 750 * 2^(60/30) = 3000
  - Here the max-penalty is equal to the suppress-limit
- The penalty can only go as high as 3000.
- The decay begins immediately, so the penalty will be lower than 3000 by the time an advertisement is received.
- A route could consistently flap several times a minute and never be suppressed

Clearing Damping Stats

- Clearing the flap statistics on a route...
  - clear ip bgp flap-statistics [(regexp *regexp*) | (filter-list *list*) |)(*address mask*)]
  - clear ip bgp *address* flap-statistics
- Parameters
  - regexp is an aspath regexp
  - filter-list is an aspath list number
  - address and mask can identify an address to be cleared

Show flap stats

- Displaying the BGP flap statistics
  - show ip bgp flap-statistics [(regexp *regexp*)|(filter-list *list*)|(*address mask* [longer])]
- Parameters
  - regexp - aspath regexp
  - filter-list - as path access-list
  - address/mask
  - longer - do the more specifics, too

Example

```
router bgp 1239
 bgp dampening route-map flap-dampen
ip as-path access-list 77 deny ^$
ip as-path access-list 77 deny ^617[4-7]_
ip as-path access-list 77 permit .*
access-list 119 permit ip any 255.255.224.0 0.0.31.255
access-list 124 permit ip any 255.255.255.0 0.0.0.255
route-map flap-dampen permit 10
 match ip address 124
 match as-path 77
 set dampening 45 125 2000 255
route-map flap-dampen permit 20
 match ip address 119
 match as-path 77
 set dampening 30 750 2000 45
route-map flap-dampen permit 30
 match as-path 77
 set dampening 15 750 2000 30
```

Damping or not?

- Turn on damping might cause flapping routes not reachable for extended time
- If routers can handle the overhead of flapping routes, why do you need to turn on damping?
  - If damped, cause outages for those routes
  - Study also show that damping cause extended BGP convergence time
- Might be best to do selective damping on per-peer base
  - Don’t damp customer routes, only damp peer routes?
- Also important to aggregate routes if damping is turned on
  - So that individual router instability only damps specific routes

AS Number Extension

- ASN: global unique number for a BGP domain, 2 bytes
- ASN is about 50% allocated and we need to prepare for future network expansion
- Extend ASN from 2 bytes to 4 bytes
- BGP carries ASN in OPEN (my ASN), AS_PATH attribute, Aggregator (ASN:IPAddr), and Communities

ASN Extension

- Use BGP capacity to advertise the capacity to support 4-byte ASN
- AS_PATH is encoded as 4-byte ASN
- Define new optional transitive attribute: NEW_AS_PATH to carry 4-byte AS_PATH for current 2-byte ASN speaker
- Define new optional transitive attribute: NEW_AGGRGATOR
- From 2-byte ASN to 4-byte ASN: pad with 0 in high order bits

Compatibility Operations

- If support 4-byte ASN capacity, all ASN and AS_PATH are encoded in 4-byte ASN
- New speaker use special reserved 2 octets ASN (AS_TRANS) to establish session with Old speaker
- New speaker will send 2 octet AS_PATH as well as NEW_AS_PATH; Non-mappable 4 octets ASN is replaced by AS_TRANS in 2-octets AS_PATH
- Need to process AS_PATH as well as NEW_AS_PATH at the same time to construct correct info

Other Extensions

- BGP Graceful restart: Another BGP capacity to help your neighbor to restart BGP process
  - “I will be back, keep forwarding”
- AS_PATH_LIMIT: when leaking more specific out for traffic engineering, announce with AS_PATH_LIMIT to prevent it from polluting global BGP tables