Route Dampening and ECMP Stability Mechanisms

This chapter explains mechanisms that enhance routing stability, including route dampening, ECMP stability techniques, BGP next-hop trigger delays, and delayed BGP route advertisements.

Route dampening

Route dampening is a BGP feature that

- reduces the propagation of unstable, flapping routes across internetworks
- assigns penalties to routes when instability is detected to temporarily suppress advertisements, and
- decays penalties over time so stable routes are reintroduced based on reuse limits.

How route dampening works

Summary

Key components involved in the process are:

- Autonomous systems: AS-1, AS-2, and AS-3.
- eBGP neighbor relationships: Between AS-1 and AS-2, and between AS-2 and AS-3.
- Route to network A: The prefix that flaps and triggers dampening.
- Dampening parameters: Initial penalty (for example, 1,000), suppression limit, reuse limit, half-life, and the history state.

Route dampening limits excessive BGP message propagation caused by route flapping. In a network with three autonomous systems, AS-1, AS-2, and AS-3, dampening penalizes unstable routes and suppresses announcements until stability is restored.

Workflow

These stages describe how route dampening works.

1. Route flap and message flow: The route to network A in AS-1 becomes unavailable. The eBGP neighbor in AS-2 sends a withdraw message, which is propagated to AS-3. When the route reappears, advertisement messages are sent again. Repeated unavailability followed by availability generates many withdraw and advertisement messages.
2. Penalty assignment: When route dampening is enabled, the router assigns an initial penalty to the flapping route (for example, 1,000) and places the route in a history state.
3. Penalty accumulation: Penalties are cumulative. If the cumulative penalty exceeds the suppression limit, the router stops advertising the route to prevent excessive churn.
4. Penalty decay: The penalty value decays with a half-life. When the penalty falls to the reuse limit, the route is re-advertised.
5. State cleanup: When the penalty decays to half of the reuse limit, the router clears the dampening information for that route.

Result

Dampening suppresses repeated announcements and withdrawals for unstable routes, reducing unnecessary BGP message propagation until the route stabilizes.

Note No penalty is applied to a BGP peer reset when route dampening is enabled, even though the reset withdraws the route.

Configure BGP route dampening

Enable and tune route dampening to minimize the impact of flapping routes.

Use address-family configuration to activate dampening with half-life, reuse, suppress, and maximum suppress time values, or attach a route policy.

Before you begin

- Identify the autonomous system number.
- Decide whether to use numeric parameters or a route policy.

Procedure

Step 1 Enter BGP configuration mode and specify the autonomous system number.

Example:

```
Router# configure
Router(config)# router bgp 120
```

Step 2 Configure the address family, IPv4 or IPv6, in unicast mode.

Example:

```
Router(config-bgp)# address-family ipv4 unicast
```

Step 3 Configure dampening parameters or attach a route policy using the **bgp dampening** [*half-life*] [*reuse suppress max-suppress-time*] | **route-policy** *route-policy-name* command.

Example:

```
Router(config-bgp-af)# bgp dampening 30 1500 10000 120
Router(config-bgp-af)# commit
```

The router suppresses advertisements for unstable routes and reuses them after penalties decay to the reuse threshold.

BGP next hop trigger delay

BGP next hop trigger delay is a BGP mechanism that

- batches next-hop change notifications to reduce CPU load and avoid unnecessary next-hop walks
- uses the Routing Information Base (RIB) classification of critical and noncritical events, and
- applies a configurable minimum batching interval per address family to control next-hop walk frequency.

How BGP next hop trigger delay works

Summary

The key components involved in the process are:

- Routing Information Base (RIB): Classifies change notifications as critical or noncritical.
- Batching interval: A configured minimum delay that governs next-hop walk scheduling.
- Address families: IPv4 and IPv6 unicast contexts where batching is applied.
- Next-hop walk scheduler: Executes deferred, batched walks based on classification and interval.

BGP next hop trigger delay improves stability by batching next-hop change notifications, reducing CPU load, and controlling how often next-hop walks run per address family.

Workflow

These stages describe how BGP next hop trigger delay works.

1. Classification: The RIB labels each next-hop change notification as critical or noncritical.
2. Interval application: The router applies the configured batching interval to defer next-hop walks for each address family.
3. Batch formation: Deferred notifications accumulate into batches for efficient processing.

4. Interleaved execution: Batched walks are interleaved across address families to prioritize work and avoid contention.
5. Stabilization: Controlled, batched processing reduces churn, improves stability, and supports faster convergence.

Result

Next-hop change notifications are batched and interleaved across address families, lowering CPU utilization and enhancing routing stability and convergence.

Guidelines for BGP next hop trigger delay

Recommendation: Use a nonzero critical next hop trigger delay

- Avoid a critical delay set to 0 in scaled environments or where next-hop changes are frequent.
- A zero delay causes high CPU utilization due to repeated next-hop walks, prevents batching, and increases wait times for address families with nonzero delays, risking traffic blackholing.
- In IPv4, a zero critical delay can slow VPNv4 convergence because IPv4 next-hop updates take precedence.

Effects of zero critical delay

Provide a concise, actionable summary of the operational impacts of setting the BGP next hop trigger delay critical value to 0.

- In scaled deployments or where next-hop changes are frequent, a zero critical delay causes high CPU utilization because each change notification triggers a next-hop walk for address families configured with the **nexthop trigger-delay** *critical 0* command.
- Next-hop change notifications are not batched, which prevents interleaving of next-hop walks in address families with a nonzero delay because those families wait until the zero-delay walks complete.
- Address families with nonzero critical delay values may experience extended wait times before the next-hop walk starts, which can lead to potential traffic blackholing.
- In IPv4, setting the critical delay to 0 can slow VPNv4 convergence because:
    - IPv4 address families are walked as many times as the number of critical alerts raised to BGP.
    - IPv4 next-hop updates for IPv4 prefixes take precedence over VPNv4 prefixes.

Default next hop trigger delay values

Table 1: Default values for next hop trigger delay

	Address families	**Value**	**Notes**
Default critical delay	Standard address families	3,000 ms	

	Address families	Value	Notes
Default critical delay	VPNv4, VPNv6	50 ms	Starting in Cisco IOS XR Release 7.10.1, the default critical delay in VPNv4 changed from 0 ms to 50 ms. With this change, all address families have a default nonzero critical delay value.
Default noncritical delay	All address families	10,000 ms	

Use the **show bgp all all nexthops** command to view the critical delay values per address family.

Configure BGP next hop trigger delay

Batch next-hop change notifications to reduce CPU load and avoid unnecessary next-hop walks.

The RIB classifies change notifications as critical and noncritical. A minimum batching interval controls how often next-hop walks run per address family.

Before you begin

Verify the desired delay values for critical and noncritical events per address family.

Procedure

Step 1 Enter the BGP configuration mode and specify the autonomous system number.

Example:

```
Router# configure
Router(config)# router bgp 120
```

Step 2 Specify the address family, IPv4 or IPv6, in unicast mode.

Example:

```
Router(config-bgp)# address-family ipv4 unicast
```

Step 3 Configure **nexthop trigger-delay critical delay** or **nexthop trigger-delay noncritical delay** for batching intervals.

Example:

```
Router(config-bgp-af)# nexthop trigger-delay critical 15000
Router(config-bgp-af)# commit
```

Delay BGP route advertisements

Delay BGP route advertisements is a BGP feature that

- prevents traffic loss by delaying BGP update generation until the Routing Information Base (RIB) is synchronized with the Forwarding Information Base (FIB)
- defers route advertisements during BGP start-up to avoid premature propagation, and
- allows a configurable delay from 1 to 600 seconds.

Table 2: Feature History Table

Feature Name	Release Information	Feature Description
Delay BGP Route Advertisements	Release 25.1.1	Introduced in this release on: Fixed Systems (8700 [ASIC: K100])(select variants only*) *This feature is supported on the Cisco 8712-MOD-M routers.
Delay BGP Route Advertisements	Release 24.4.1	Introduced in this release on: Fixed Systems (8200 [ASIC: P100], 8700 [ASIC: P100])(select variants only*); Modular Systems (8800 [LC ASIC: P100])(select variants only*) *This feature is supported on: • 88-LC1-36EH • 88-LC1-12TH24FH-E • 88-LC1-52Y8H-EM • 8212-48FH-M • 8711-32FH-M

Feature Name	Release Information	Feature Description
Delay BGP Route Advertisements	Release 7.5.3	You can now prevent traffic loss due to premature advertising of BGP routes and subsequent packet loss in a network. You can achieve this by setting the delay time of the BGP start-up in the router until the Routing Information Base (RIB) is synchronized with the Forward Information Base (FIB) in the routing table. This delays the BGP update generation and prevents traffic loss in a network. You can configure a minimum delay of 1 second and a maximum delay of 600 seconds. This feature introduces the **update wait-install delay startup** command.

When BGP forwards traffic, it waits for feedback from the Routing Information Base (RIB) until the RIB is ready to forward traffic. After the RIB is ready, BGP sends the route updates to the BGP neighbors and peer-groups. Advertising routes before the RIB is synchronized with the Forwarding Information Base (FIB) can cause traffic loss. To avoid this problem, the router must delay the BGP start-up process to delay the BGP route update generation until the RIB and FIB are synchronized.

To accomplish this, you can configure the **update wait-install delay startup** command to delay BGP update generation. This feature allows you to configure the minimum and maximum delay periods. Use the **show bgp process** command to view the BGP process delay since the last router reload. Set the delay to 1 to 600 seconds to prevent traffic loss.

Restrictions of delay BGP route advertisements

This feature is applicable only for the following Address Family Indicators (AFIs):

- IPv4 unicast
- IPv6 unicast
- VPNv4 unicast
- VPNv6 unicast

Configure BGP route advertisement delay

Delay BGP update generation until the Routing Information Base (RIB) is synchronized, preventing premature route advertisements and traffic loss.

Configure a start-up delay for the desired address family, IPv4, IPv6, VPNv4, or VPNv6, using the BGP address-family submode.

Before you begin

- Determine the BGP autonomous system number.
- Identify the address family to configure.
- Choose the delay value in seconds.

Procedure

Step 1 Specify the BGP autonomous system number and enter BGP configuration mode.

Example:

```
Router# configure
Router(config)# router bgp 1
```

Step 2 Specify the address-family.

Example:

```
Router(config-bgp)# address-family ipv4 unicast
```

Step 3 Schedule the delay of the BGP process to prevent routes from being advertised to peers until RIB is synchronized.

Example:

```
Router(config-bgp-af)# update wait-install delay startup 10
Router(config-bgp-af)# commit
```

Step 4 Verify the running configuration.

Example:

```
Router# show running-config router bgp 1
router bgp 1
 address-family ipv4 unicast
  update wait-install delay startup 10
!
```

Step 5 Run the **show bgp process** command to verify the delay of the BGP process update since the last router reload.

Example:

```
Router# show bgp process

BGP Process Information:
BGP is operating in STANDALONE mode
Autonomous System number format: ASPLAIN
Autonomous System: 100
Router ID: 192.168.0.2 (manually configured)
Default Cluster ID: 192.168.0.2
Active Cluster IDs:  192.168.0.2
------------------------------
------------------------------
Update wait-install enabled:
  ack request 2, ack rcvd 2, slow ack 0
  startup delay 10 secs
```

```
--More—
```

ECMP stability features

Equal-Cost Multi-Path (ECMP) stability features improve forwarding reliability during network reconfigurations, migrations, and path churn. They coordinate BGP and the FIB to avoid out-of-resource conditions, minimize packet loss, and maintain fast convergence. This section includes:

- ECMP out of resource avoidance: Delays best-path selection and hardware programming when resource thresholds are reached to prevent overload.
- ECMP ASN-based prefix download delay: Waits for all ECMP paths from a specified autonomous system (ASN) before inserting prefixes, reducing transient churn and resource spikes.

ECMP out of resource avoidance

ECMP out of resource avoidance is a network resiliency feature that

- tracks hardware resource usage inline in the FIB to provide real-time feedback
- delays BGP best-path selection, route installation into the RIB, and FIB hardware programming when utilization crosses thresholds, and
- uses dampening and Destination-based Load Balancing (DLB) mechanisms to prevent overload and minimize packet loss.

These mechanisms help optimize routing stability and hardware resource usage:

- FIB dampening: A mechanism that consolidates or caches route updates in CPU memory and delays hardware programming when resource usage reaches a configured threshold. FIB dampening is disabled by default and can be enabled through Cisco Express Forwarding (CEF) configuration.
- Dampening switchover: A mechanism that detects when route churn stabilizes and programs stable route updates into hardware. If stability is not detected within the maximum dampening duration, a forced switchover occurs.
- Destination-based Load Balancing (DLB): A protective mode that programs routes with a single forwarding path when hardware resource usage exceeds a configured threshold.

Table 3: Feature History Table

Feature Name	Release Information	Feature Description
ECMP Out of Resource Avoidance	Release 25.1.1	Introduced in this release on: Fixed Systems (8700 [ASIC: K100])(select variants only*) *This feature is supported on the Cisco 8712-MOD-M routers.

Feature Name	Release Information	Feature Description
ECMP Out of Resource Avoidance	Release 25.1.1	Introduced in this release on: Fixed Systems (8010 [ASIC: A100]) This feature is supported on: • 8011-4G24Y4H-I
ECMP Out of Resource Avoidance	Release 24.4.1	Introduced in this release on: Fixed Systems (8200 [ASIC: P100], 8700 [ASIC: P100])(select variants only*); Modular Systems (8800 [LC ASIC: P100])(select variants only*) *This feature is supported on: • 88-LC1-36EH • 88-LC1-12TH24FH-E • 88-LC1-52Y8H-EM • 8212-48FH-M • 8711-32FH-M

Feature Name	Release Information	Feature Description
ECMP Out of Resource Avoidance	Release 24.2.11	You can now ensure minimum packet loss and service disruption during network reconfigurations or migrations by preventing Equal-Cost Multi-Path (ECMP) Out of Resource (OOR) conditions. This feature allows BGP to delay route updates and FIB to delay programming the routes in hardware when resources are low, thus avoiding system overload. The feature introduces these changes: CLI: • **prefix-ecmp-delay** • **cef load-balancing recursive oor mode dampening-and-dlb** YANG Data Models: • `Cisco-IOS-XR-um-router-bgp-cfg.yang` • `Cisco-IOS-XR-ipv4-bgp-oper.yang` • `Cisco-IOS-XR-fib-common-cfg.yang` • `Cisco-IOS-XR-fib-common-oper.yang` (see GitHub, YANG Data Models Navigator)

The routers can experience transient ECMP resource shortages and traffic drops during data center migrations, maintenance events, or the introduction of new sites that temporarily increase ECMP resource usage. After the network stabilizes, the router recovers from the ECMP spike; however, traffic dropped during an out of resource (OOR) condition does not automatically recover.

How ECMP OOR avoidance works

Summary

The key components involved in the process are:

- FIB inline resource tracking: Measures hardware resource consumption and reports utilization.
- BGP and RIB: Delay best-path selection and route installation when utilization crosses thresholds.
- FIB programming: Defers hardware updates to avoid overload.

- Dampening control: Consolidates updates and manages switchover timings.
- DLB mode: Provides uni-path forwarding under high resource utilization.

ECMP OOR avoidance protects forwarding capacity by monitoring resource usage and temporarily deferring route updates and hardware programming, with optional switchover to DLB when necessary.

Workflow

These stages describe how ECMP OOR avoidance works:

1. Resource monitoring: The FIB tracks hardware utilization and provides real-time feedback.
2. Threshold detection: When utilization reaches the configured threshold, BGP delays best-path selection and route installation into the RIB, and the FIB delays hardware programming.
3. Dampening activation: The FIB consolidates route updates in CPU memory and defers hardware programming to prevent overload.
4. Stability assessment: Dampening switchover checks for churn stability. If stability is detected, the FIB programs the consolidated updates. If stability is not detected within the maximum dampening duration, a forced switchover occurs.
5. DLB engagement: During a forced switchover or when utilization exceeds the DLB threshold, routes are programmed in DLB mode. New route installations may also enter DLB if resource usage is high.
6. Automatic reversion: When hardware resource usage falls below the DLB threshold, affected routes revert to ECMP forwarding.

Result

The router minimizes packet loss and service disruption by deferring route updates and hardware programming under high utilization and by switching to uni-path forwarding when required.

Conditions for DLB programming

Routes are programmed in DLB mode under these conditions:

- New route installation: If hardware resource usage exceeds the configured DLB threshold, program the route in DLB mode to avoid an OOR condition.
- Forced dampening switchover: If hardware resource usage is above the DLB threshold at the end of the maximum dampening duration, program routes in DLB mode.

Additional details:

- DLB operates in a uni-path mode, selecting a single forwarding path to protect against OOR conditions.
- The system automatically switches between DLB and ECMP based on current hardware resource utilization.

Limitations and guidelines of ECMP OOR resource accounting

ECMP OOR resource accounting limitations

- Use ECMP out-of-resource (OOR) accounting primarily in deployments without MPLS in the path. If MPLS is present and the system detects approximately 1,000 or more MPLS link-down indications

(LDIs), the platform increases the resource count to account for maximum MPLS paths only after it observes considerable usage to avoid misclassifying internal labels (for example, BFD internal label) as an MPLS deployment.

- Rely only on FIB recursive and non-recursive LDI accounting. Objects and features that reserve ECMP or members are not included, for example, Layer 2.
- Expect differences between inline FIB resource accounting and SDK resource accounting shown by the **show controller npu resource** command.
- Do not assume FIB transitions LDIs between load-balancing levels, for example, SHLDI to REC_SHLDI to PHLDI. If such a transition occurs, the system disables resource monitoring accounting and issues a warning because counters differ across levels and transitions can create inaccuracies.
- Resource accounting does not apply to management interfaces or special (drop) adjacencies.

Link utilization risks and operational guidelines

- Caution: When DLB mode is active, ECMP path spreading is not available, which can increase the risk of link over-subscription as traffic concentrates on a single path.
- Recommendation: Configure thresholds and dampening durations to balance stability with convergence. The default maximum dampening duration is 5 minutes.

Configure ECMP OOR avoidance in BGP

Configure an ECMP delay duration and a resource usage threshold to prevent out-of-resource (OOR) conditions and reduce packet loss.

The **prefix-ecmp-delay** command is supported only under global AFI/SAFI for IPv4 and IPv6. When the threshold is exceeded, programming of new routes into hardware is deferred for the configured interval.

Before you begin

- Determine the BGP autonomous system number.
- Choose the address family.
- Select the delay interval (milliseconds) and the OOR threshold (percent).

Follow these steps to configure ECMP delay duration and the resource usage threshold limit.

Procedure

Step 1 Specify the autonomous system number and enter BGP configuration mode.

Example:

```
Router# configure
Router(config)# router bgp 100
```

Step 2 Specify the address-family.

Example:

```
Router(config-bgp)# address-family ipv4 unicast
```

Step 3 Run the **prefix-ecmp-delay** *interval_value* **oor-threshold** *threshold_value* command to configure the ECMP delay duration and the OOR threshold value.

Example:

```
Router(config-bgp-af)# prefix-ecmp-delay 10000 oor-threshold 30
```

In this sample configuration, when the resource usage exceeds a threshold of 30%, programming of new routes into the hardware is delayed by 10 seconds (10000 ms).

Currently, this command is supported only in global Address Family Identifier (AFI) and Subsequent Address Family Identifiers (SAFI) for IPv4 and IPv6.

Step 4 Run the **show bgp ipv4 unicast process detail performance-statistics | b OOR** command or **show bgp ipv4 unicast process detail | b OOR** command to verify the configuration.

a) Run the **show bgp ipv4 unicast process detail performance-statistics | b OOR** command to verify the configuration.

Example:

```
Router# show bgp ipv4 unicast process detail performance-statistics | b OOR

OOR queue Info:
 Oldest Queue Num: 0
 Recent Queue Num: 0
 Prefix count HWM: 40000
 Delayed Paths count: 30680000
 Delayed Nets count: 280000
 Processed Nets count: 270000
 Last delayed Q time: May 29 22:30:23.412
 Last processed Q time: May 29 22:31:35.409
 Last OOR recovery time: ---
 Q-num  Q-size   Expiry-Time
  1     0       ---
  2     0       ---
  3     0       ---
  4     0       ---
  5     0       ---
```

b) Run the **show bgp ipv4 unicast process detail | b OOR** command to verify the configuration.

Example:

```
Router# show bgp ipv4 unicast process detail | b OOR
Fri Jun  7 17:38:18.613 UTC
 OOR Flag 0 OOR Threshold 0
 Prefix Download Delay 10000
Dampening is not enabled
```

Step 5 Run the **show bgp** *location* **detail** command to view the details of BGP prefix delays.

Example:

```
Router# show bgp 209.165.201.9/27 detail
BGP routing table entry for 209.165.201.9/27
Versions:
  Process           bRIB/RIB   SendTblVer
  Speaker           18490149     18490149
    Flags: 0x00023201+0x28010000+0x00000000 multipath;
Last Modified: Jul 30 19:17:47.643 for 18:43:25
Last Delayed at: Jul 30 19:10:32.643
Paths: (16 available, best #1)
  Advertised IPv4 Unicast paths to update-groups (with more than one peer):
    10.1 0.7 0.8
  Advertised IPv4 Unicast paths to peers (in unique update groups):
    172:23:1:79::2
```

```
  Path #1: Received by speaker 0
  Flags: 0x3000000001078001+0x00, import: 0x020
  Advertised IPv4 Unicast paths to update-groups (with more than one peer):
    10.1 0.7 0.8
  Advertised IPv4 Unicast paths to peers (in unique update groups):
    172:23:1:79::2
  9001 64313 56001 58505, (received & used)
    209.165.201.2 from 209.165.201.2 (10.1.1.1), if-handle 0x00000000
      Origin IGP, localpref 100, valid, external, best, group-best, multipath
      Received Path ID 0, Local Path ID 1, version 18490149
      Origin-AS validity: (disabled)
  Path #2: Received by speaker 0
  Flags: 0x3000000001038001+0x00, import: 0x020
  Not advertised to any peer
  9002 64313 56001 58505, (received & used)
    209.165.200.2 from 209.165.200.2 (10.1.1.2), if-handle 0x00000000
      Origin IGP, localpref 100, valid, external, group-best, multipath
      Received Path ID 0, Local Path ID 0, version 0
      Origin-AS validity: (disabled)
  Path #3: Received by speaker 0
  Flags: 0x3000000001038001+0x00, import: 0x020
  Not advertised to any peer
  9003 64313 56001 58505, (received & used)
    209.165.202.2 from 209.165.202.2 (50.1.1.3), if-handle 0x00000000
      Origin IGP, localpref 100, valid, external, group-best, multipath
      Received Path ID 0, Local Path ID 0, version 0
      Origin-AS validity: (disabled)
  Path #4: Received by speaker 0
  Flags: 0x3000000001038001+0x00, import: 0x020
  Not advertised to any peer
  9004 64313 56001 58505, (received & used)
    209.165.200.6 from 209.165.200.6 (10.1.1.4), if-handle 0x00000000
      Origin IGP, localpref 100, valid, external, group-best, multipath
      Received Path ID 0, Local Path ID 0, version 0
      Origin-AS validity: (disabled)
...
```

The sample output indicates that the BGP prefix download to the RIB has been delayed.

ECMP ASN-based prefix download delay

ASN-based prefix download delay is a BGP feature that

- delays downloading prefixes to the Routing Information Base (RIB) and Forwarding Information Base (FIB) based on autonomous system numbers (ASNs) in an Equal-Cost Multi-Path (ECMP) context
- queues new prefixes or paths until the path count per ASN matches the established neighbor count for that ASN, and
- optimizes resource utilization to reduce traffic drops and minimize network disruption during rapid route arrivals.

Table 4: Feature History Table

Feature Name	Release Information	Feature Description
ECMP out of resource avoidance using ASN-based prefix download delay	Release 25.1.1	Introduced in this release on: Fixed Systems (8200 [ASIC: Q200, P100], 8700 [ASIC: P100, K100], 8010 [ASIC: A100]); Centralized Systems (8600 [ASIC: Q200]); Modular Systems (8800 [LC ASIC: Q100, Q200, P100]) You can now ensure minimum packet loss and service disruption during network reconfigurations or migrations by preventing ECMP OOR conditions. The feature allows BGP to delay the download of BGP prefixes into the RIB and FIB until the router learns all paths from a specific ASN. This ASN-based delay dynamically optimizes resource utilization, and actively manages ECMP paths in real-time during network changes. Previously, you could apply a fixed delay to all BGP prefixes using the **prefix-ecmp-delay** command. The feature introduces these changes: CLI: • The **ecmp-delay** submode is introduced in the **address-family** command. • **show bgp as-neighbors** YANG Data Models: `Cisco-IOS-XR-um-router-bgp-cfg` (see GitHub, YANG Data Models Navigator)

When a router receives routes from multiple neighbors in the same AS, it delays RIB/FIB insertion until all paths from that AS are learned, helping prevent transient out-of-resource (OOR) conditions caused by hardware limits.

Unlike prefix-ecmp-delay, which applies a fixed delay to all prefixes, ecmp-delay waits for ASN-based ECMP path completion for smarter route selection and resource allocation.

This table explains the key differences between **ecmp-delay** submode and **prefix-ecmp-delay**.

Table 5: Comparison of ecmp-delay and prefix-ecmp-delay in BGP

Category	ecmp-delay	prefix-ecmp-delay
Delay mechanism	Fixed, AS-based, and platform-oor-threshold delay options.	Fixed delay for all prefixes.
Scope	Downloads paths of a prefix only after learning all the ECMP paths from a given ASN, ensuring optimal route installation.	Applies a uniform delay to all prefixes, regardless of ASN or neighbor grouping.
Configuration	Supports per-ASN filtering with AS-based delay configuration.	Applies to all prefixes globally.
Flexibility	Supports different delay types, such as ASN-based delay, and automatically adjusts delays.	Requires manual tuning of the delay interval for all prefixes.
OOR condition handling	Minimizes the probability of causing OOR issues by ensuring all ECMP paths are learnt or ready.	Might still cause OOR issues if the delay is configured incorrectly.
Impact on network convergence	Minimal impact (smart delay).	Can slow network convergence (fixed delay).

How ASN-based prefix download delay works

Summary

The key components involved in the process are:

- ASN-based grouping: Collects all ECMP paths learned from the same AS.
- Delayed queue: Holds new prefixes/paths until ASN path completion.
- RIB/FIB insertion control: Inserts prefixes after completion or after the configured delay (if applicable).

The feature groups paths by ASN and defers RIB/FIB insertion until all ECMP paths from that ASN arrive, reducing churn and transient OOR events.

Workflow

These stages describe how ASN-based prefix download delay works.

1. Wait for ECMP paths: The router waits for all ECMP paths from an ASN before installing routes.
2. Delay on incomplete sets: If the ASN-based ECMP set is incomplete, the router delays RIB/FIB installation for those prefixes to prevent premature route selection.
3. Forceful download after the configured delay: After the configured delay interval, the router forcefully downloads the prefixes, even if all ECMP paths have not arrived.

4. Completion and insertion: When all ECMP paths from the ASN are present, the router downloads the prefix set to the RIB/FIB.

Result

Batched, ASN-aware insertion reduces transient resource spikes, minimizes packet loss, and keeps ECMP routing stable and efficient.

Benefits of ASN-based prefix download delay

The key benefits of the feature are:

- Delaying RIB insertion can eliminate transient OOR conditions with FIB hardware resources.
- The delay runs automatically for BGP prefix downloads into the RIB/FIB, removing the need to tune a universal fixed delay.

Types of delay in ecmp-delay submode

Within the **ecmp-delay** submode, you can configure these delay types:

- Fixed: Delays prefixes by a set time before inserting them into the RIB or FIB.
- Platform-oor-based: Dynamically adjusts the delay based on hardware resource availability.
- AS-based: Waits for all ECMP paths from an ASN before inserting or downloading the path set or nexthop set to RIB.

Limitations and guidelines for ECMP ASN-based prefix delay

Limitations

- Configure the feature only for IPv4 and IPv6 global address families.
- Do not enable the ecmp-delay submode together with prefix-ecmp-delay in BGP.
- Apply the feature only to eBGP-learned ECMP paths; ASN-based prefix grouping is required.

Usage guidelines

- Choose the delay type carefully based on network design and traffic engineering requirements; you cannot apply multiple delay types simultaneously.
- The router downloads prefixes to the RIB after the specified delay (in milliseconds), even if the ECMP set is incomplete because not all paths from the AS have been learned.
- If you provide an AS list, the feature limits operation to the AS numbers in that list; without an AS list, the router applies the feature to all AS numbers learned on the node.

Configure ECMP ASN-based prefix download delay

Configure ASN-based delay for ECMP prefixes so the router inserts routes into the RIB/FIB only after learning all paths from a given autonomous system (ASN), reducing transient out-of-resource (OOR) events.

The ecmp-delay submode operates under global AFI/SAFI for IPv4 and IPv6. You can optionally scope the delay to a specific AS list. When configured, the router defers RIB/FIB insertion by the specified interval (in milliseconds) or until the ASN path set is complete.

Before you begin

- Identify the BGP autonomous system number.
- Choose the address family.
- Decide the delay interval (milliseconds).
- Optionally define the ASNs to include in an AS list.

Follow these steps to configure ASN-based delay:

Procedure

Step 1 In BGP configuration mode, define the address family to install multiple eBGP paths in the RIB and the forwarding table.

Example:

```
Router(config)#router bgp 65536
Router(config-bgp)#address-family ipv4 unicast
Router(config-bgp-af)#maximum-paths eibgp 1024 selective route-policy mp_rpl
```

Step 2 (Optional) Run the **as-list** command in the BGP configuration mode to define a list of ASNs that must be considered for **ecmp-delay**.

Example:

```
Router(config)#router bgp 65536
Router(config-bgp)#as-list as-list1
Router(config-bgp-as-list)#100
Router(config-bgp-as-list)#300
Router(config-bgp-as-list)#500
Router(config-bgp-as-list)#600
Router(config-bgp-as-list)#commmit
```

Step 3 Run the **ecmp-delay** command to configure delay in the best path calculation for prefixes with ECMP paths based on the neighbor AS.

- Configure delay to download all BGP prefixes with ECMP paths for all ASN numbers learned on the node. In the configuration, the router delays the RIB or FIB installation for BGP prefixes by 10 milliseconds for all AS numbers that are learned on the node.

  ```
  Router(config-bgp-af)#ecmp-delay
  Router(config-bgp-af-ecmpdelay)#as-based delay 10
  ```

- Configure delay to download BGP prefixes with ECMP paths for specific ASN numbers mentioned in the as-list.

  ```
  Router(config-bgp-af)#ecmp-delay
  Router(config-bgp-af-ecmpdelay)#as-based delay 10 as-list as-list1
  ```

Step 4 Run the **show running-config** command to verify the running configuration.

Example:

```
router bgp 65536
 address-family ipv4 unicast
```

```
   maximum-paths eibgp 1024 selective route-policy mp_rpl
   ecmp-delay
    as-based delay 10
   !
  !
 !
```

Step 5 Verify the ECMP ASN-based delay for IPv4 unicast routes.

a) Run the **show bgp ipv6 unicast process** command to verify ECMP as-delay configured for IPv4 unicast routes.

Example:

```
Router#show bgp ipv4 unicast process

BGP Process Information:
BGP is operating in STANDALONE mode
Autonomous System number format: ASPLAIN
Autonomous System: 65536
Router ID: 1.1.1.1 (manually configured)
Default Cluster ID: 1.1.1.1
Active Cluster IDs:  1.1.1.1
Fast external fallover enabled
Platform Loadbalance paths max: 1024
Platform RLIMIT max: 8589934592 bytes
Maximum limit for BMP buffer size: 1638 MB
Default value for BMP buffer size: 1228 MB
Current limit for BMP buffer size: 1228 MB
Current utilization of BMP buffer limit: 0 B
Neighbor logging is enabled
Enforce first AS enabled
AS Path multipath-relax is enabled
Use SR-Policy admin/metric of color-extcomm Nexthop during path comparison: disabled
Default local preference: 100
Default keepalive: 30
Graceful restart enabled
Restart time: 1
Stale path timeout time: 0
RIB purge timeout time: 600
Non-stop routing is enabled

ExtComm Color Nexthop validation: RIB
Update delay: 1
Generic scan interval: 60
Configured Segment-routing Local Block: [0, 0]
In use Segment-routing Local Block: [15000, 15999]
Platform support mix of sr-policy and native nexthop: No
  Last insert into reset queue: Mar 17 10:03:29.542, removed at Mar 17 10:03:29.542

Address family: IPv4 Unicast
AS based ECMP Download Delay configured
OOR Flag 0 OOR Threshold 0
Prefix Download Delay 10
Selective EIBGP multipath enabled
Dampening is not enabled
Client reflection is enabled in global config
Dynamic MED is Disabled
Dynamic MED interval : 10 minutes
Dynamic MED Timer : Not Running
Dynamic MED Periodic Timer : Not Running
Scan interval: 60
Total prefixes scanned: 3811
Prefixes scanned per segment: 100000
Number of scan segments: 1
Nexthop resolution minimum prefix-length: 0 (not configured)
```

```
IPv6 Nexthop resolution minimum prefix-length: 0 (not configured)
Main Table Version: 399620
Table version synced to RIB: 399620
Table version acked by RIB: 399620
IGP notification: IGPs notified
RIB has converged: version 84
RIB table prefix-limit reached ?  [No], version 0
RPKI version 3361
RPKI soft-reconfig version 3361
Origin-AS validation is enabled for this address-family
Permanent Network Enabled
Label alloc mode: per-prefix
BGP NSR scoped sync stats:
   Scoped Sync last msg failed: 0
   Scoped Sync last msg resumed: 0
   Scoped Sync default route stopped: 0
   Scoped Sync default route resumed: 0
   Scoped Sync default route lookup failure: 0

OC-RIB Telemetry Neighbor Outbound Attributes Pool summary:
                           Alloc           Free
Pool 25:                   0               0
Pool 49:                   0               0
Pool 73:                   0               0
Pool 97:                   0               0
Pool 121:                  0               0
Pool 145:                  0               0
Pool 169:                  0               0
Pool 193:                  0               0
Pool 217:                  0               0
Pool 241:                  0               0

Number of Paths having particular number of OCRIB out attributes:
                           Paths
1 Out Attrs:               1476400096

Node                Process     Nbrs Estb Rst Upd-Rcvd Upd-Sent Nfn-Rcv Nfn-Snt
node0_RP0_CPU0      Speaker      188  148   2   177737     9562       0       0
```

The sample output is for IPv4 unicast routes configured for ECMP as-delay. In the sample ouput, **Prefix Download Delay 10** indicates that the router delays the RIB or FIB installation for BGP prefixes by 10 milliseconds for all AS numbers that are learned on the node.

b) Run the **show bgp ipv6 unicast process** command to verify ECMP as-delay configured for IPv6 unicast routes.

Example:

```
Router#show bgp ipv6 unicast process
Mon Mar 17 17:23:59.146 UTC

BGP Process Information:
BGP is operating in STANDALONE mode
Autonomous System number format: ASPLAIN
Autonomous System: 65536
Router ID: 1.1.1.1 (manually configured)
Default Cluster ID: 1.1.1.1
Active Cluster IDs:  1.1.1.1
Fast external fallover enabled
Platform Loadbalance paths max: 1024
Platform RLIMIT max: 8589934592 bytes
Maximum limit for BMP buffer size: 1638 MB
Default value for BMP buffer size: 1228 MB
Current limit for BMP buffer size: 1228 MB
Current utilization of BMP buffer limit: 0 B
Neighbor logging is enabled
```

```
Enforce first AS enabled
AS Path multipath-relax is enabled
Use SR-Policy admin/metric of color-extcomm Nexthop during path comparison: disabled
Default local preference: 100
Default keepalive: 30
Graceful restart enabled
Restart time: 1
Stale path timeout time: 0
RIB purge timeout time: 600
Non-stop routing is enabled

ExtComm Color Nexthop validation: RIB
Update delay: 1
Generic scan interval: 60
Configured Segment-routing Local Block: [0, 0]
In use Segment-routing Local Block: [15000, 15999]
Platform support mix of sr-policy and native nexthop: No
  Last insert into reset queue: Mar 17 10:03:29.542, removed at Mar 17 10:03:29.542

Address family: IPv6 Unicast
AS based ECMP Download Delay configured
OOR Flag 0 OOR Threshold 0
Prefix Download Delay 10
Selective EIBGP multipath enabled
Dampening is not enabled
Client reflection is enabled in global config
Dynamic MED is Disabled
Dynamic MED interval : 10 minutes
Dynamic MED Timer : Not Running
Dynamic MED Periodic Timer : Not Running
Scan interval: 60
Total prefixes scanned: 3090
Prefixes scanned per segment: 100000
Number of scan segments: 1
Nexthop resolution minimum prefix-length: 0 (not configured)
IPv6 Nexthop resolution minimum prefix-length: 0 (not configured)
Main Table Version: 407943
Table version synced to RIB: 407943
Table version acked by RIB: 407943
RIB has converged: version 43
RIB table prefix-limit reached ?  [No], version 0
RPKI version 3361
RPKI soft-reconfig version 3361
Origin-AS validation is enabled for this address-family
Permanent Network Enabled
Label alloc mode: per-prefix
BGP NSR scoped sync stats:
   Scoped Sync last msg failed: 0
   Scoped Sync last msg resumed: 0
   Scoped Sync default route stopped: 0
   Scoped Sync default route resumed: 0
   Scoped Sync default route lookup failure: 0

OC-RIB Telemetry Neighbor Outbound Attributes Pool summary:
                           Alloc           Free
Pool 25:                   0               0
Pool 49:                   0               0
Pool 73:                   0               0
Pool 97:                   0               0
Pool 121:                  0               0
Pool 145:                  0               0
Pool 169:                  0               0
Pool 193:                  0               0
Pool 217:                  0               0
```

```
Pool 241:                    0               0

Number of Paths having particular number of OCRIB out attributes:
                             Paths
1 Out Attrs:                 3371410339

Node                Process     Nbrs Estb Rst Upd-Rcvd Upd-Sent Nfn-Rcv Nfn-Snt
node0_RP0_CPU0      Speaker      188  148   2   177737     9562       0       0
```

c) Run the **show bgp as-neighbors** command to view BGP neighbor relationships grouped by AS.

Example:

```
Router#show bgp as-neighbors
Wed Nov 20 21:10:58.133 UTC
AS: 4294967291, Neighbors: 64, Established: 0
Last updated: Nov 20 00:25:35.285
Neighbor                                        State                   Last state change
31.0.2.2                                         Idle                    Nov 20 19:33:34.119
…
31.0.65.2                                       Idle                    Nov 20 19:33:34.025

AS: 4294967292, Neighbors: 64, Established: 0
Last updated: Nov 20 00:25:35.285
Neighbor                                        State                   Last state change
32.0.2.3                                        Active                  Nov 20 19:33:39.613
…
32.0.65.3                                       Active                  Nov 20 19:33:39.540
```

The router defers RIB/FIB insertion for prefixes with ECMP paths by ASN, either until all paths from the ASN are learned or until the configured delay expires, reducing transient resource spikes and improving stability.