Use BGP Route Reflection and Multiple Cluster IDs

Contents

Introduction

Prerequisites

- Requirements
- Components Used

Description of BGP Route Reflection

Route Reflection Configuration Examples

- Single Cluster with Default Settings
- Single Cluster with Client-to-client Reflection Disabled
- Two Clusters, Intra-site and Inter-site Route Reflection
- Two Clusters, no Client-to-client Reflection

Cluster List and Loop Prevention

- Reflection Between Client and Non-client
- Intra-cluster Reflection
- Inter-cluster Reflection
- MCIDs and Loop Prevention

Related Information

Introduction

This document describes different scenarios of Border Gateway Protocol (BGP) route reflection and usage of multiple cluster IDs.

Prerequisites

Requirements

Prior knowledge of BGP concepts, especially clusters and route reflection, is assumed.

Components Used

This document is not restricted to specific software and hardware versions.

The information in this document was created from the devices in a specific lab environment. All of the devices used in this document started with a cleared (default) configuration. If your network is live, ensure that you understand the potential impact of any command.

Description of BGP Route Reflection

A BGP speaker is a BGP enabled router. By default, BGP speakers does not advertise iBGP-learned prefixes

to iBGP peers - this is done to maintain loop prevention. RFC4456 introduces the route reflection feature which removes the need of full mesh between iBGP speakers. When route reflector reflects a prefix, it creates/modifies an optional non-transitive attribute called CLUSTER_LIST by adding its own cluster ID to it. This attribute is used for loop prevention: when router receives update which CLUSTER_LIST contains routers own cluster ID, this update is discarded.

By default, cluster ID is set to the BGP router id value, but can be set to an arbitrary 32-bit value. Multiple cluster IDs (MCID) feature allows you to assign per-neighbor cluster IDs. Thus, there are 3 types of route reflection scenarios.

1. Between client and non-client
2. Between clients in the same cluster (intra-cluster)
3. Between clients in different clusters (inter-cluster)

Route Reflection Configuration Examples

In this section there are some router reflection scenarios and respective configuration examples.

Single Cluster with Default Settings

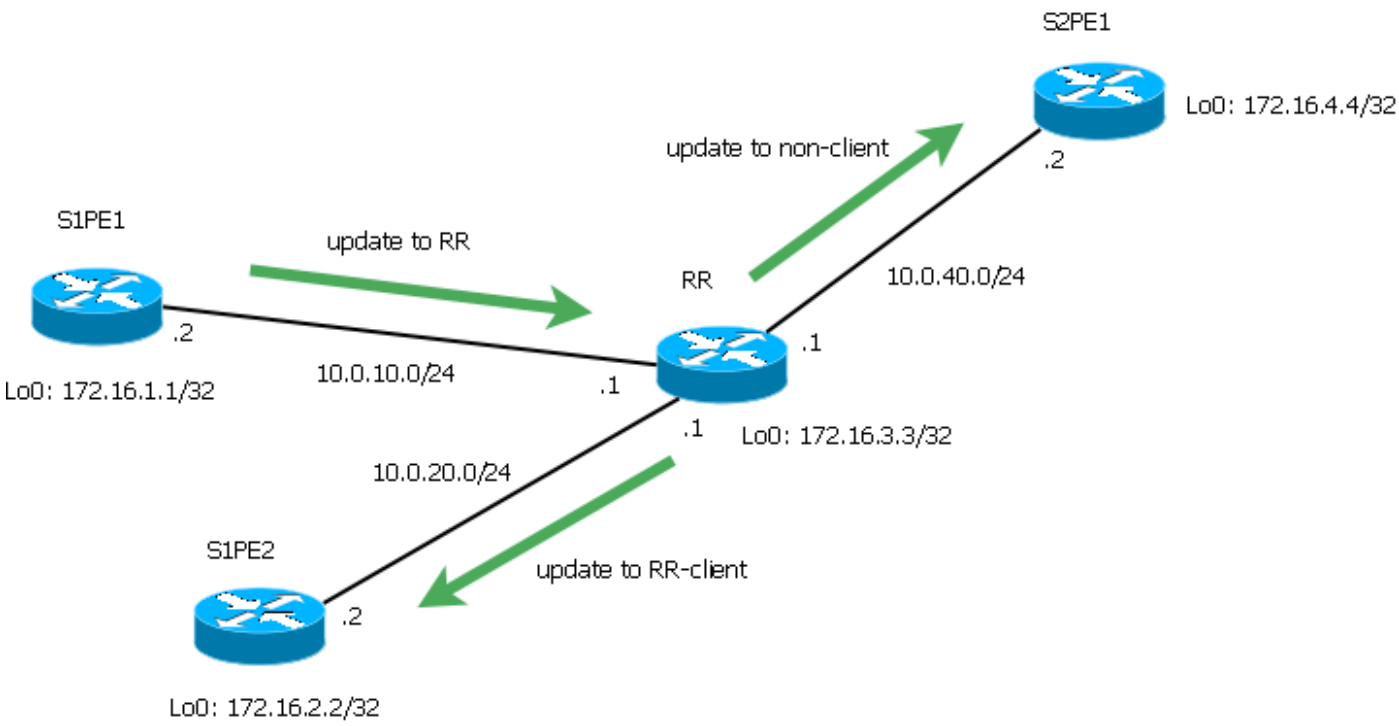


Image 1

The next configuration has been done on router RR acting as route reflector.

```
<#root>

RR#

show run | sec bgp

router bgp 1
 bgp log-neighbor-changes
 neighbor 10.0.10.2 remote-as 1
```

```
 neighbor 10.0.10.2 route-reflector-client
 neighbor 10.0.20.2 remote-as 1
 neighbor 10.0.20.2 route-reflector-client
 neighbor 10.0.40.2 remote-as 1
```

In this case, S1PE1 and S1PE2 are clients of RR while S2PE1 is non-client. In conventional designs, non-client router serves as the route reflector for routers in the next hierarchy level, but in this example, just another PE is used for simplicity.

<#root>

RR#

show ip bgp cluster-ids

Global cluster-id: 172.16.3.3 (configured: 0.0.0.0)
BGP client-to-client reflection: Configured Used
all (inter-cluster and intra-cluster):

ENABLED

intra-cluster: ENABLED

ENABLED

List of cluster-ids:
Cluster-id #-neighbors C2C-rfl-CFG C2C-rfl-USE

RR#

show ip bgp 172.16.1.1

BGP routing table entry for 172.16.1.1/32, version 2
Paths: (1 available, best #1, table default)
Advertised to update-groups:
1 2
Refresh Epoch 2
Local, (Received from a RR-client)
10.0.10.2 from 10.0.10.2 (172.16.1.1)
Origin IGP, metric 0, localpref 100, valid, internal, best
rx pathid: 0, tx pathid: 0x0

RR#

show ip bgp update-group 1

BGP version 4 update-group 1, internal, Address Family: IPv4 Unicast
BGP Update version : 4/0, messages 0
Topology: global, highest version: 4, tail marker: 4
Format state: Current working (OK, last not in list)
Refresh blocked (not in list, last not in list)
Update messages formatted 2, replicated 2, current 0, refresh 0, limit 1000
Number of NLRIs in the update sent: max 1, min 0
Minimum time between advertisement runs is 0 seconds
Has 1 member:
10.0.40.2

RR#

```
show ip bgp update-group 2

BGP version 4 update-group 2, internal, Address Family: IPv4 Unicast
  BGP Update version : 4/0, messages 0
  Route-Reflector Client
  Topology: global, highest version: 4, tail marker: 4
  Format state: Current working (OK, last not in list)
                Refresh blocked (not in list, last not in list)
  Update messages formatted 3, replicated 6, current 0, refresh 0, limit 1000
  Number of NLRIs in the update sent: max 1, min 0
  Minimum time between advertisement runs is 0 seconds
  Has 2 members:
   10.0.10.2        10.0.20.2
```

These outputs show that RR receives the 172.16.1.1/32 prefix from S1PE1 and reflects it to the client S1PE2 and non-client S2PE1. In this particular case, update is also sent back to S1PE1, but it happens because S1PE1 and S1PE2 have the same routing policy and, therefore, form the same update group.

Single Cluster with Client-to-client Reflection Disabled

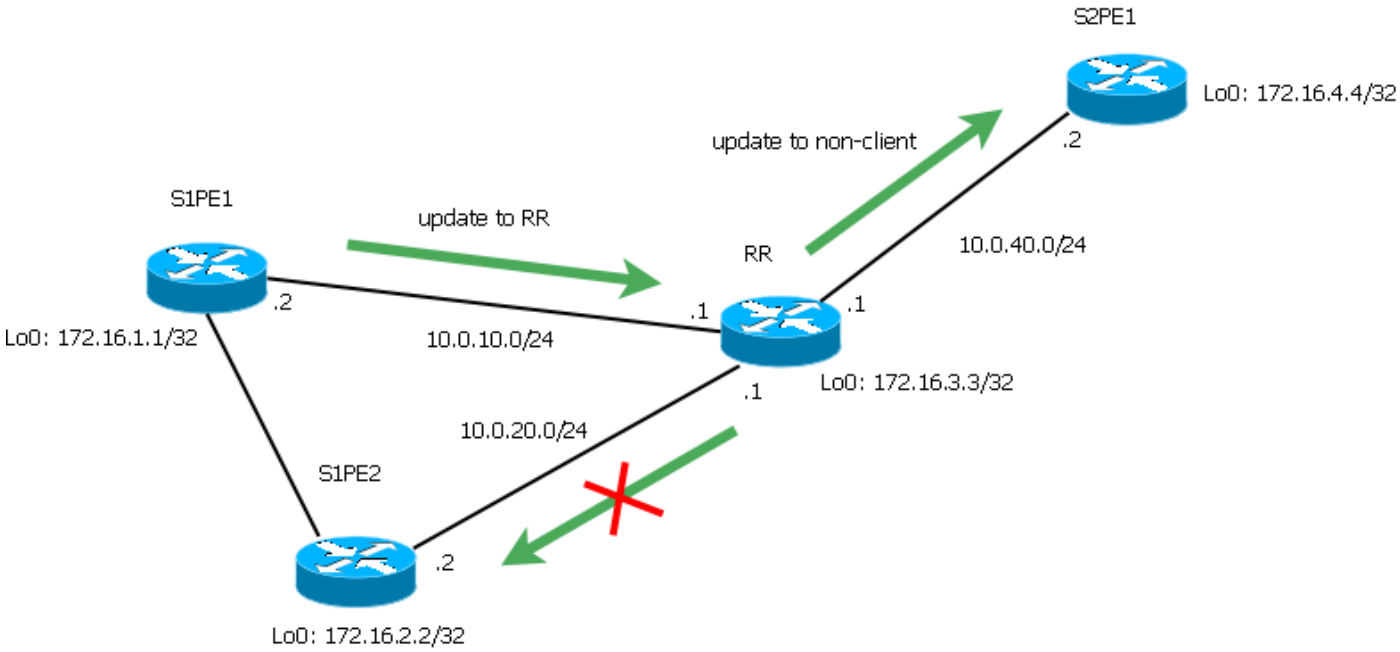


Image 2

The next configuration has been done on router RR acting as route reflector.

```
<#root>

RR#

show run | sec bgp

router bgp 1


no bgp client-to-client reflection

 bgp log-neighbor-changes
 neighbor 10.0.10.2 remote-as 1
 neighbor 10.0.10.2 route-reflector-client
```

```
 neighbor 10.0.20.2 remote-as 1
 neighbor 10.0.20.2 route-reflector-client
 neighbor 10.0.40.2 remote-as 1
```

Let us assume AS1 is partially meshed: S1PE1 and S1PE2 form iBGP neighborship (for instance, they are located on the same site and you want to optimize the way network processes updates). In this case, RR has client-to-client reflection disabled and it reflects 172.16.1.1/32 coming from S1PE1 only to the non-client S2PE1.

```
<#root>

RR#

show ip bgp cluster-ids

Global cluster-id: 172.16.3.3 (configured: 0.0.0.0)
BGP client-to-client reflection:         Configured    Used
  all (inter-cluster and intra-cluster):

DISABLED

  intra-cluster:                         ENABLED

DISABLED


List of cluster-ids:
Cluster-id     #-neighbors C2C-rfl-CFG C2C-rfl-USE


RR#

show ip bgp 172.16.1.1

BGP routing table entry for 172.16.1.1/32, version 5
Paths: (1 available, best #1, table default, RIB-failure(17))
  Advertised to update-groups:
     1
  Refresh Epoch 2
  Local, (Received from a RR-client)
    10.0.10.2 from 10.0.10.2 (172.16.1.1)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      rx pathid: 0, tx pathid: 0x0


RR#

show ip bgp update-group 1

BGP version 4 update-group 1, internal, Address Family: IPv4 Unicast
  BGP Update version : 7/0, messages 0
  Topology: global, highest version: 7, tail marker: 7
  Format state: Current working (OK, last not in list)
                Refresh blocked (not in list, last not in list)
  Update messages formatted 4, replicated 4, current 0, refresh 0, limit 1000
  Number of NLRIs in the update sent: max 1, min 0
  Minimum time between advertisement runs is 0 seconds
  Has 1 member:
   10.0.40.2
```

Two Clusters, Intra-site and Inter-site Route Reflection

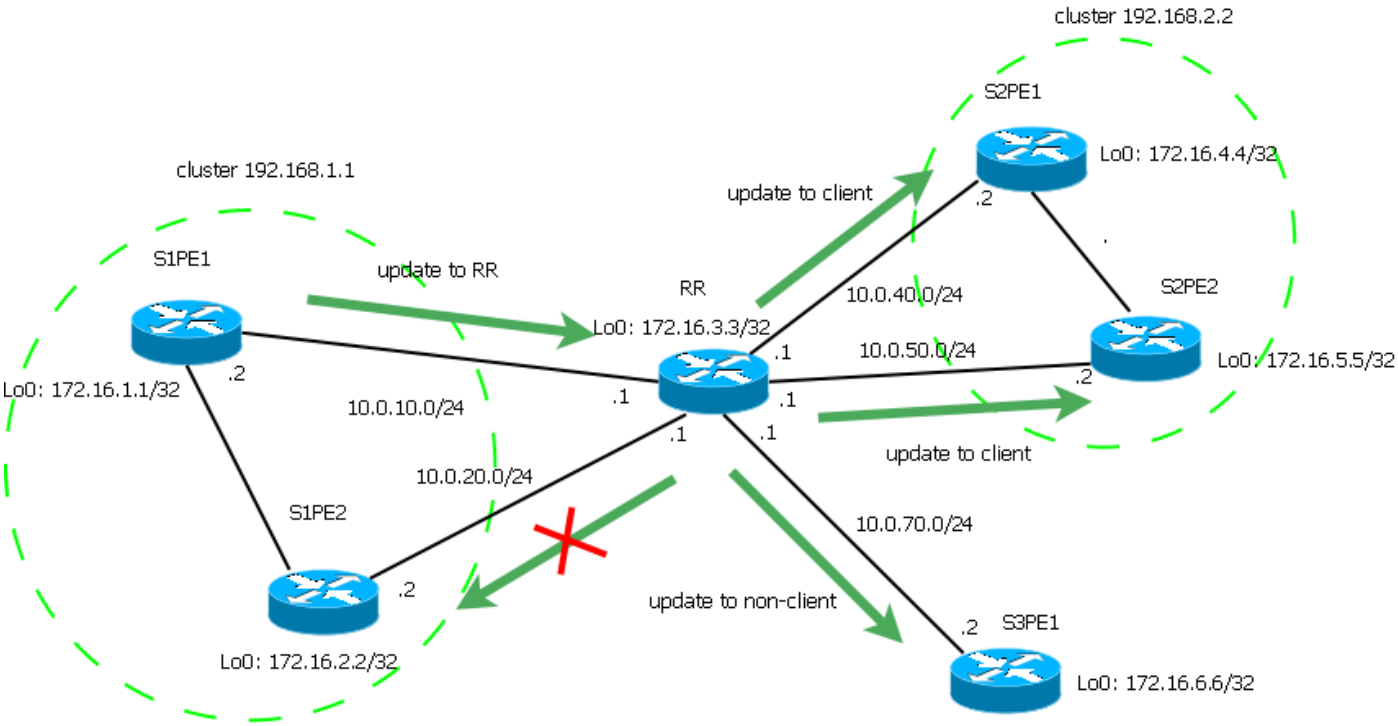


Image 3

The next configuration has been done on router RR acting as route reflector.

<#root>

RR#

show run | sec bgp

router bgp 1

no bgp client-to-client reflection intra-cluster cluster-id 192.168.1.1

```
 bgp log-neighbor-changes
 neighbor 10.0.10.2 remote-as 1
 neighbor 10.0.10.2 cluster-id 192.168.1.1
 neighbor 10.0.10.2 route-reflector-client
 neighbor 10.0.20.2 remote-as 1
 neighbor 10.0.20.2 cluster-id 192.168.1.1
 neighbor 10.0.20.2 route-reflector-client
 neighbor 10.0.40.2 remote-as 1
 neighbor 10.0.40.2 cluster-id 192.168.2.2
 neighbor 10.0.40.2 route-reflector-client
 neighbor 10.0.50.2 remote-as 1
 neighbor 10.0.50.2 cluster-id 192.168.2.2
 neighbor 10.0.50.2 route-reflector-client
 neighbor 10.0.70.2 remote-as 1
```

In this case both PEs on Site 1 form the cluster 192.168.1.1 while both PEs on Site 2 form the cluster 192.168.2.2. S3PE1 is a non-client. PEs on Site 1 have direct iBGP session, intra-cluster reflection is disabled for the cluster 192.168.1.1, but still enabled for the cluster 192.168.2.2. Inter-cluster reflection is

enabled.

<#root>

```
RR#
```

show ip bgp cluster-ids

```
Global cluster-id: 172.16.3.3 (configured: 0.0.0.0)
BGP client-to-client reflection:         Configured    Used
  all (inter-cluster and intra-cluster):
```

ENABLED

```
  intra-cluster:                         ENABLED       ENABLED

List of cluster-ids:
Cluster-id     #-neighbors C2C-rfl-CFG C2C-rfl-USE
192.168.1.1                2 DISABLED
```

DISABLED

```
192.168.2.2                2 ENABLED
```

ENABLED

```
RR#
```

show ip bgp 172.16.1.1

```
BGP routing table entry for 172.16.1.1/32, version 5
Paths: (1 available, best #1, table default, RIB-failure(17))
  Advertised to update-groups:
     3          5
  Refresh Epoch 9
  Local, (Received from a RR-client)
    10.0.10.2 from 10.0.10.2 (172.16.1.1)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      rx pathid: 0, tx pathid: 0x0


RR#
```

show ip bgp update-group 3

```
BGP version 4 update-group 3, internal, Address Family: IPv4 Unicast
  BGP Update version : 11/0, messages 0
  Topology: global, highest version: 11, tail marker: 11
  Format state: Current working (OK, last not in list)
                Refresh blocked (not in list, last not in list)
  Update messages formatted 20, replicated 20, current 0, refresh 0, limit 1000
  Number of NLRIs in the update sent: max 1, min 0
  Minimum time between advertisement runs is 0 seconds
  Has 1 member:
   10.0.70.2


RR#
```

show ip bgp update-group 5

```
BGP version 4 update-group 5, internal, Address Family: IPv4 Unicast
```

```
  BGP Update version : 11/0, messages 0
  Route-Reflector Client
  Configured with cluster-id 192.168.2.2
  Topology: global, highest version: 11, tail marker: 11
  Format state: Current working (OK, last not in list)
                Refresh blocked (not in list, last not in list)
  Update messages formatted 22, replicated 34, current 0, refresh 0, limit 1000
  Number of NLRIs in the update sent: max 3, min 0
  Minimum time between advertisement runs is 0 seconds
  Has 2 members:
   10.0.40.2        10.0.50.2
```

Prefix 172.16.1.1/32 received from S1PE1 is reflected to the clients in the cluster 192.168.2.2 and to the non-clients. At the same time, prefix 172.16.4.4/32 received from S2PE1 is reflected to all clients and non-clients.

<#root>

```
RR#
```

show ip bgp 172.16.4.4

```
BGP routing table entry for 172.16.4.4/32, version 9
Paths: (1 available, best #1, table default, RIB-failure(17))
  Advertised to update-groups:
     3          4          5
  Refresh Epoch 6
  Local, (Received from a RR-client)
    10.0.40.2 from 10.0.40.2 (172.16.4.4)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      rx pathid: 0, tx pathid: 0x0

RR#
```

show ip bgp update-group 3

```
BGP version 4 update-group 3, internal, Address Family: IPv4 Unicast
  BGP Update version : 11/0, messages 0
  Topology: global, highest version: 11, tail marker: 11
  Format state: Current working (OK, last not in list)
                Refresh blocked (not in list, last not in list)
  Update messages formatted 20, replicated 20, current 0, refresh 0, limit 1000
  Number of NLRIs in the update sent: max 1, min 0
  Minimum time between advertisement runs is 0 seconds
  Has 1 member:
   10.0.70.2

RR#
```

show ip bgp update-group 4

```
BGP version 4 update-group 4, internal, Address Family: IPv4 Unicast
  BGP Update version : 11/0, messages 0
  Route-Reflector Client
  Configured with cluster-id 192.168.1.1
  Topology: global, highest version: 11, tail marker: 11
  Format state: Current working (OK, last not in list)
                Refresh blocked (not in list, last not in list)
  Update messages formatted 26, replicated 47, current 0, refresh 0, limit 1000
  Number of NLRIs in the update sent: max 3, min 0
```

```
  Minimum time between advertisement runs is 0 seconds
  Has 2 members:
   10.0.10.2        10.0.20.2

RR#
```

show ip bgp update-group 5

```
BGP version 4 update-group 5, internal, Address Family: IPv4 Unicast
  BGP Update version : 11/0, messages 0
  Route-Reflector Client
  Configured with cluster-id 192.168.2.2
  Topology: global, highest version: 11, tail marker: 11
  Format state: Current working (OK, last not in list)
                Refresh blocked (not in list, last not in list)
  Update messages formatted 22, replicated 34, current 0, refresh 0, limit 1000
  Number of NLRIs in the update sent: max 3, min 0
  Minimum time between advertisement runs is 0 seconds
  Has 2 members:
   10.0.40.2        10.0.50.2
```

You can disable intra-site route-reflection for cluster 192.168.2.2 as well, but in this case, clients in that cluster must have full mesh of iBGP sessions:

<#root>

```
RR(config-router)#
```

no bgp client-to-client reflection intra-cluster cluster-id 192.168.2.2

```
RR#sh ip bgp cluster-ids
Global cluster-id: 172.16.3.3 (configured: 0.0.0.0)
BGP client-to-client reflection:         Configured    Used
  all (inter-cluster and intra-cluster): ENABLED
  intra-cluster:                         ENABLED       ENABLED

List of cluster-ids:
Cluster-id     #-neighbors C2C-rfl-CFG C2C-rfl-USE
192.168.1.1                2 DISABLED
```

DISABLED

```
192.168.2.2                2 DISABLED
```

DISABLED

Intra-site reflection can be also disabled for all clusters:

<#root>

```
RR(config-router)#
```

no bgp client-to-client reflection intra-cluster cluster-id any

Two Clusters, no Client-to-client Reflection

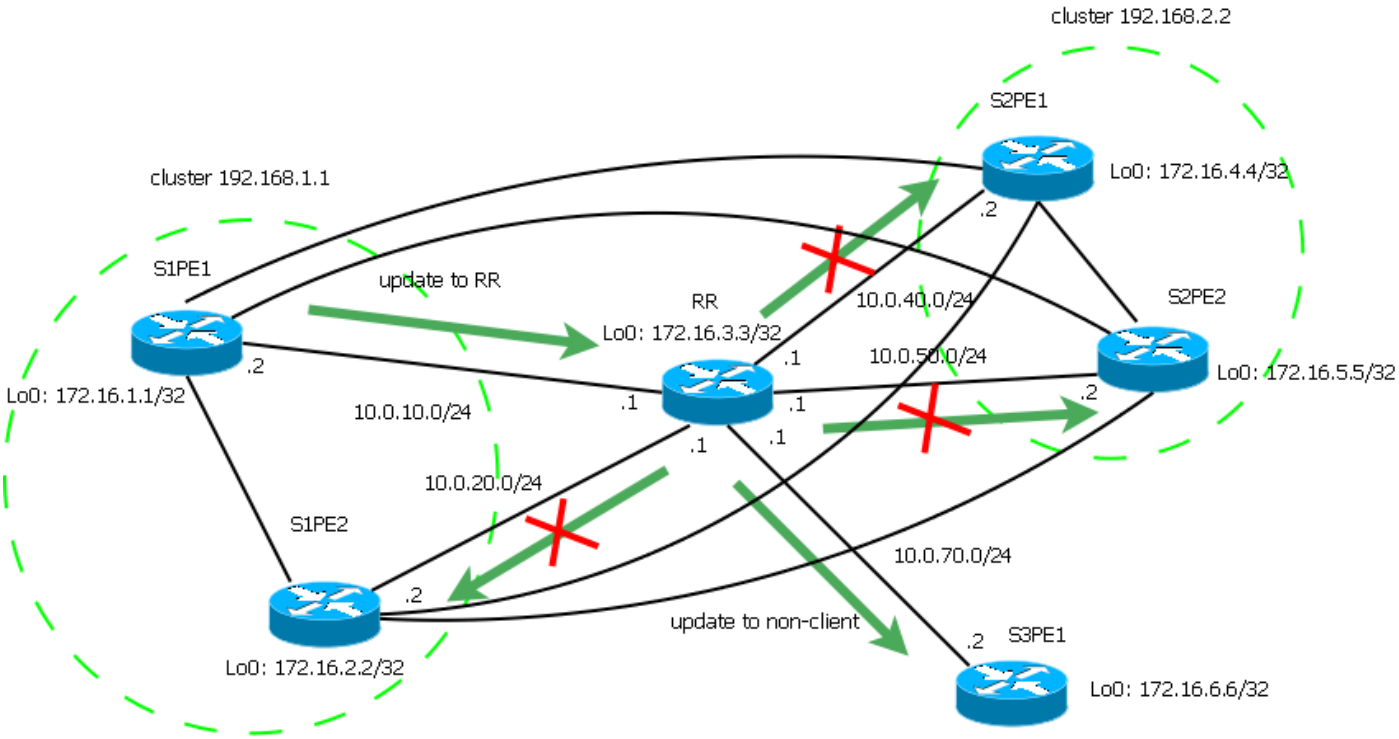


Image 4

The next configuration has been done on router RR acting as route reflector.

<#root>

```
RR#
```

show run | sec bgp

```
router bgp 1
```

no bgp client-to-client reflection

```
 bgp log-neighbor-changes
 neighbor 10.0.10.2 remote-as 1
 neighbor 10.0.10.2 cluster-id 192.168.1.1
 neighbor 10.0.10.2 route-reflector-client
 neighbor 10.0.20.2 remote-as 1
 neighbor 10.0.20.2 cluster-id 192.168.1.1
 neighbor 10.0.20.2 route-reflector-client
 neighbor 10.0.40.2 remote-as 1
 neighbor 10.0.40.2 cluster-id 192.168.2.2
 neighbor 10.0.40.2 route-reflector-client
 neighbor 10.0.50.2 remote-as 1
 neighbor 10.0.50.2 cluster-id 192.168.2.2
 neighbor 10.0.50.2 route-reflector-client
 neighbor 10.0.70.2 remote-as 1
```

It is possible to disable both intra-cluster and inter-cluster reflection. In this case, only reflection between clients and non-clients is performed.

<#root>

RR#

show ip bgp cluster-ids

Global cluster-id: 172.16.3.3 (configured: 0.0.0.0)
BGP client-to-client reflection: Configured Used
all (inter-cluster and intra-cluster):

DISABLED

intra-cluster: ENABLED

DISABLED

List of cluster-ids:
Cluster-id #-neighbors C2C-rfl-CFG C2C-rfl-USE
192.168.1.1 2 ENABLED

DISABLED

192.168.2.2 2 ENABLED

DISABLED

RR#

show ip bgp 172.16.1.1

BGP routing table entry for 172.16.1.1/32, version 5
Paths: (1 available, best #1, table default, RIB-failure(17))
Advertised to update-groups:
3
Refresh Epoch 9
Local, (Received from a RR-client)
10.0.10.2 from 10.0.10.2 (172.16.1.1)
Origin IGP, metric 0, localpref 100, valid, internal, best
rx pathid: 0, tx pathid: 0x0

RR#

show ip bgp 172.16.4.4

BGP routing table entry for 172.16.4.4/32, version 9
Paths: (1 available, best #1, table default, RIB-failure(17))
Advertised to update-groups:
3
Refresh Epoch 6
Local, (Received from a RR-client)
10.0.40.2 from 10.0.40.2 (172.16.4.4)
Origin IGP, metric 0, localpref 100, valid, internal, best
rx pathid: 0, tx pathid: 0x0

RR#

show ip bgp update-group 3

```
BGP version 4 update-group 3, internal, Address Family: IPv4 Unicast
  BGP Update version : 11/0, messages 0
  Topology: global, highest version: 11, tail marker: 11
  Format state: Current working (OK, last not in list)
                Refresh blocked (not in list, last not in list)
  Update messages formatted 20, replicated 20, current 0, refresh 0, limit 1000
  Number of NLRIs in the update sent: max 1, min 0
  Minimum time between advertisement runs is 0 seconds
  Has 1 member:
   10.0.70.2
```

Prefixes 172.16.1.1/32 and 172.16.4.4/32 are originated by clusters 192.168.1.1 and 192.168.2.2, respectively. Both these prefixes are reflected only to the non-client S3PE1. In this case, all clients must be fully meshed. Generally, in this particular scenario, MCIDs do not really make sense (the same behavior could be achieved with single cluster), but they can still be used if you want to have different cluster lists for routes from different neighbours.



Note: It is not possible to enable intra-cluster reflection (either for specific cluster or for all clusters) while inter-cluster reflection is disabled.

Cluster List and Loop Prevention

When RR reflects a prefix, it adds cluster ID to the optional non-transitive attribute CLUSTER_LIST. Also, it sets the optional non-transitive attribute ORIGINATOR_ID to the router ID of the peer, that has advertised the prefix to the RR.

When MCIDs are used and RR reflects the prefix, it uses cluster ID configured for the peer which has advertised that prefix to the RR. If that peer does not have specific cluster ID configured, global cluster ID is used.

Let us see some examples. RR has all forms of route reflection enabled. Global cluster ID is 172.16.3.3, cluster IDs 192.168.1.1 and 192.168.2.2 are set to PEs on site 1 and site 2 respectively (refer to topology diagram above).

<#root>

RR#

show ip bgp cluster-ids

Global cluster-id:

172.16.3.3

 (configured: 0.0.0.0)
BGP client-to-client reflection: Configured   Used
  all (inter-cluster and intra-cluster): ENABLED
  intra-cluster: ENABLED   ENABLED

List of cluster-ids:
Cluster-id   #-neighbors C2C-rfl-CFG C2C-rfl-USE

192.168.1.1

               2 ENABLED   ENABLED

```
192.168.2.2

                2 ENABLED     ENABLED
```

Reflection Between Client and Non-client

```
<#root>

S2PE3#

show ip bgp 172.16.1.1

BGP routing table entry for 172.16.1.1/32, version 2
Paths: (1 available, best #1, table default, RIB-failure(17))
  Not advertised to any peer
  Refresh Epoch 1
  Local
    10.0.10.2 (metric 20) from 10.0.70.1 (172.16.3.3)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Originator: 172.16.1.1, Cluster list:

192.168.1.1

      rx pathid: 0, tx pathid: 0x0


S2PE3#

show ip bgp 172.16.4.4

BGP routing table entry for 172.16.4.4/32, version 4
Paths: (1 available, best #1, table default, RIB-failure(17))
  Not advertised to any peer
  Refresh Epoch 1
  Local
    10.0.40.2 (metric 20) from 10.0.70.1 (172.16.3.3)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Originator: 172.16.4.4, Cluster list:

192.168.2.2

      rx pathid: 0, tx pathid: 0x0
```

Non-client S2PE3 receives prefix 172.16.1.1/32 originated by the cluster 192.168.1.1 - cluster ID 192.168.1.1 is added to the cluster list. It also receives prefix 172.16.4.4/32 originated by the cluster 192.168.2.2 - cluster ID 192.168.2.2 is added to the cluster list.

```
<#root>

S1PE1#

show ip bgp 172.16.6.6

BGP routing table entry for 172.16.6.6/32, version 5
Paths: (1 available, best #1, table default, RIB-failure(17))
  Not advertised to any peer
  Refresh Epoch 1
  Local
```

```
    10.0.70.2 (metric 20) from 10.0.10.1 (172.16.3.3)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Originator: 172.16.6.6, Cluster list:
```

172.16.3.3

```
      rx pathid: 0, tx pathid: 0x0
```

Client S1PE1 receives prefix 172.16.6.6/32 originated by a non-client - the global cluster ID 172.16.3.3 is added to the cluster list.

Intra-cluster Reflection

<#root>

S1PE2#

show ip bgp 172.16.1.1/32

```
BGP routing table entry for 172.16.1.1/32, version 8
Paths: (1 available, best #1, table default, RIB-failure(17))
  Not advertised to any peer
  Refresh Epoch 1
  Local
    10.0.10.2 (metric 20) from 10.0.20.1 (172.16.3.3)
      Origin IGP, metric 0, localpref 100, valid, internal
      Originator: 172.16.1.1, Cluster list: 192.168.1.1
      rx pathid: 0, tx pathid: 0
```

S1PE2 belongs to the cluster 192.168.1.1 and receives prefix 172.16.1.1/32 originated by S1PE1 which also belongs to the cluster 192.168.1.1. Cluster ID 192.168.1.1 is added to the cluster list.

Inter-cluster Reflection

<#root>

S2PE1#

show ip bgp 172.16.1.1/32

```
BGP routing table entry for 172.16.1.1/32, version 4
Paths: (1 available, best #1, table default, RIB-failure(17))
  Not advertised to any peer
  Refresh Epoch 1
  Local
    10.0.10.2 (metric 20) from 10.0.40.1 (172.16.3.3)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Originator: 172.16.1.1, Cluster list:
```

192.168.1.1

```
      rx pathid: 0, tx pathid: 0x0
```

S1PE1#

show ip bgp 172.16.4.4/32

```
BGP routing table entry for 172.16.4.4/32, version 4
Paths: (1 available, best #1, table default, RIB-failure(17))
  Not advertised to any peer
  Refresh Epoch 1
  Local
    10.0.40.2 (metric 20) from 10.0.10.1 (172.16.3.3)
      Origin IGP, metric 0, localpref 100, valid, internal, best
      Originator: 172.16.4.4, Cluster list:
```

192.168.2.2

```
      rx pathid: 0, tx pathid: 0x0
```

S2PE1 belongs to the cluster 192.168.2.2 and receives prefix 172.16.1.1/32 originated by cluster 192.168.1.1 - cluster ID is set to 192.168.1.1.

S1PE1 belongs to the cluster 192.168.1.1 and receives prefix 172.16.4.4/32 originated by cluster 192.168.2.2 - cluster ID is set to 192.168.2.2.

MCIDs and Loop Prevention

If router receives the update for the prefix which cluster list contains routers own cluster ID, the update is discarded. If MCIDs are used, update that contains any of configured cluster IDs (either global or per-neighbor) would be discarded.

Related Information

- RFC 4456
- BGP-Multiple Cluster IDs