BGP Design and Implementation

Practical guidelines for designing and deploying a scalable BGP routing architecture

Randy Zhang, CCIE® No. 5659
Micah Bartell, CCIE No. 5069

ciscopress.com

BGP Design and Implementation

Randy Zhang, CCIE No. 5659
Micah Bartell, CCIE No. 5069

Cisco Press

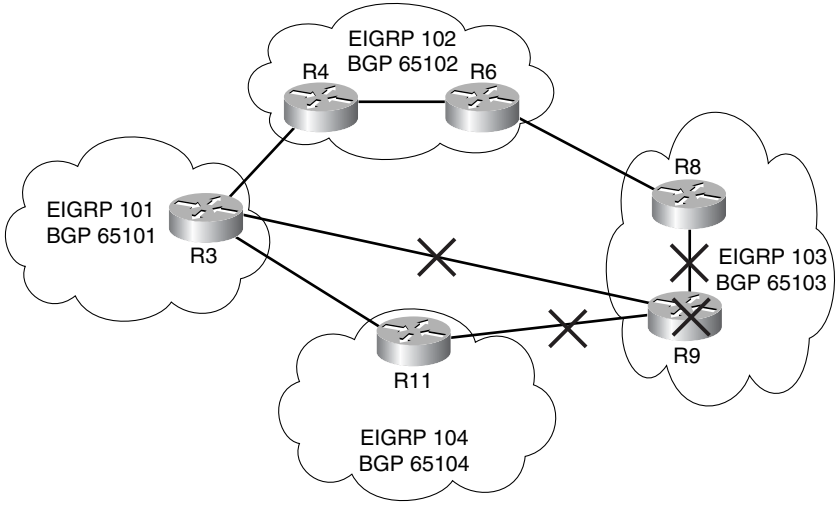
Cisco Press
800 East 96th Street, 3rd Floor
Indianapolis, IN 46240 USA

Example 5-15 *R11 Path Information for 10.2.0.0/16 (Continued)*

```
Advertised to non peer-group peers:
172.16.13.17
65103 65101 65102
  172.16.13.17 from 172.16.13.17 (172.16.9.1)
    Origin IGP, localpref 100, valid, external
65101 65102
  172.16.13.9 from 172.16.13.9 (172.16.3.1)
    Origin IGP, localpref 100, valid, external, best
```

The failure of a core router has an effect similar to a core link failure, only amplified. When the core router fails, every BGP session to that router fails. In Figure 5-9, R9 fails. The similarity of a router failure to a link failure becomes more obvious when viewed from the perspective of the other routers. R3 sees a link failure to R9, R11 sees a link failure to R9, and R8 sees a link failure to R9.

Figure 5-9 *Network Topology Core Router Failure*



For example, consider 10.2.0.0/16, which originates in AS 65102. Only R11 is actively using the path received from R9. When R11 detects the failure, the path received from R9 is removed from the BGP RIB, and the path from R3 is used.

A more interesting scenario is the reroute of prefix 10.3.0.0/16, which is originated by AS 65103. When R3 detects the failure of the BGP session to R9, it removes the prefix from its BGP RIB. Then R3 installs the path received from R4 and advertises this path to R11, which causes R11 to replace the previous path.

An important point to note from this section is that not taking advantage of an IGP for its ability to react to network changes quickly has a slight effect on the speed at which the network can reconverge. The amount of time required to reconverge can also be additive. This additive effect is a result of each BGP speaker upon receiving the new path information having to run the path-selection process and then withdraw and advertise based on the outcome.

Administrative Control

This architecture provides clear points at which administrative authority can be divided. The easiest way to divide administrative control from a routing perspective is to introduce eBGP sessions. When eBGP is used, the next hop on advertised prefixes is changed to the address of the advertising router. Only a single BGP session is required at each interconnection point. Not all the autonomous systems need an eBGP session directly between them—only those with a direct physical connection.

Routing Policy

It is sometimes desirable to prevent two regions from communicating with each other. In this design, however, every core router must have full routing information, because it might be acting as a transit router between two other regions. This disallows the use of route filtering to block connectivity between two regions. The best method of limiting connectivity is inbound packet filtering on the core router interfaces connecting with the regional network.

Internal/External BGP Core Architecture

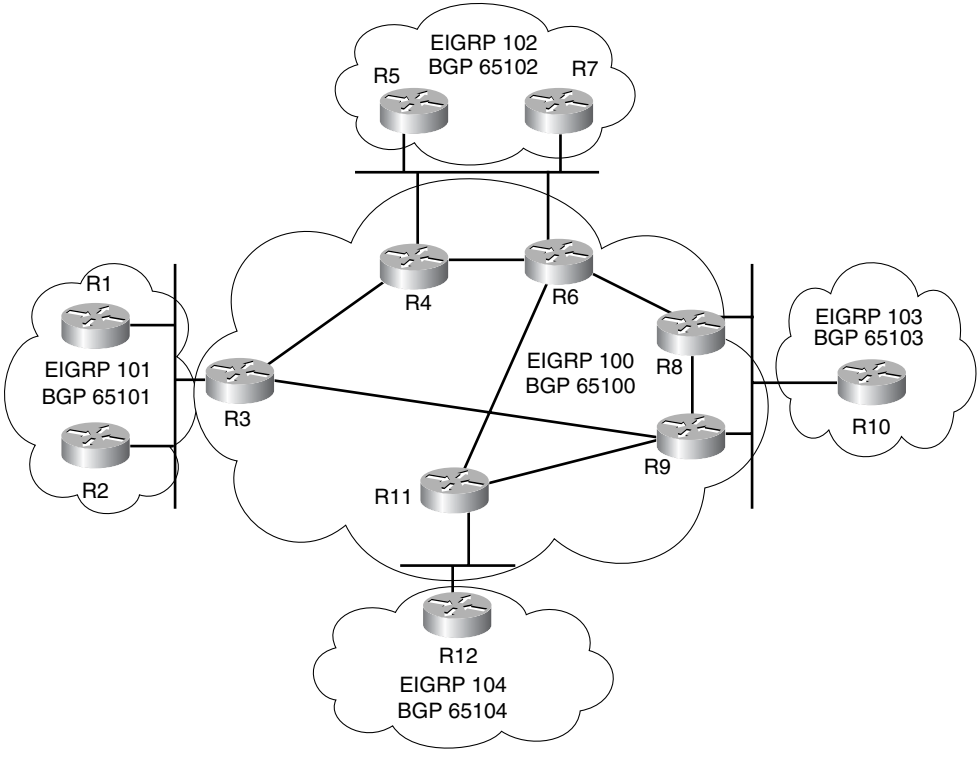
The internal/external BGP core architecture employs an iBGP core, with external BGP as the mechanism by which regions attach to the core. Figure 5-10 shows an example. This architecture provides prefix reduction in regional IGP processes, clear delineation of administrative boundaries, and flexible policy control. It also bounds the scope of regional IGP instabilities.

The internal/external BGP scenario at first appears to be the most complex scenario because of the number of components. However, the end result is a BGP architecture that is easier to work with when defining policy, troubleshooting, or expanding the network.

The regional IGP process provides reachability throughout the entire regional network. This process carries full routing and topological information for the region. The regional IGP process also provides next-hop resolution for the iBGP-learned prefixes between the regional border routers in addition to reachability for the iBGP peers, as

required in redundant regional environments. A default route is injected into the regional IGP process on each of the regional border routers.

Figure 5-10 *Internal/External BGP Architecture*



The regional border router is a new concept introduced in this architecture. The DMZ between the core and a region is the connection between the regional border router and the core router. The regional border router exists entirely in the region and connects to the network core via eBGP. This separates regional routing from core routing. In the previous architectures, the regional border router functionality was shared with the core router functions on the same device.

The core IGP is used to provide next-hop resolution and reachability for the iBGP peering sessions between core routers. The core IGP contains only the core routers, core links, and loopback interfaces on the core routers. The core IGP process does not participate in any redistribution between protocols.

There are multiple aspects to the BGP portion of this architecture. There is a full iBGP mesh between the core routers. These iBGP sessions are sourced from the loopback interfaces on the core routers and are configured with **next-hop-self**. The use of **next-hop-self** removes the need to inject the subnets on the links connecting to the regional border routers for next-hop resolution. Sourcing the iBGP sessions from the loopback interfaces allows the sessions to remain active in the case of a core link failure that can be routed around.

Each region has its own BGP AS, and the core has its own AS. Network prefixes for each region are injected into the regional BGP process using **network** statements on the regional border routers. The use of **network** statements allows for controlled injection of prefixes that can be reached in the IGP. It is possible to directly redistribute from the IGP into BGP on the regional border routers. Although this practice is discouraged, if the number of prefixes involved makes using **network** statements administratively unfeasible, redistribution becomes an option. It cannot be stressed enough that prefix filtering should be applied any time redistribution is performed between protocols. Another option is to aggregate the prefixes with a static route to Null0 and use the **network** command to inject the aggregate.

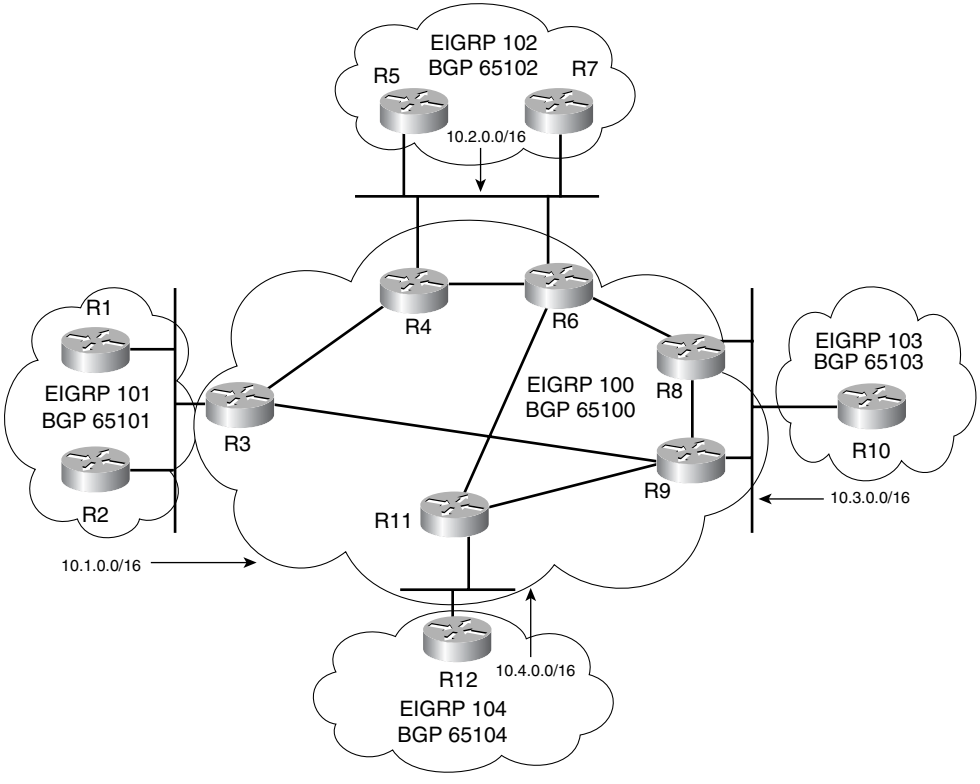
The regional BGP autonomous systems connect to the BGP core AS through the use of eBGP. Each regional border router peers via eBGP with all the core routers it is directly connected to when redundant routers are in place.

Path Selection

The path selection in the core routers is very similar to that in the iBGP-only architecture scenario. The BGP decision process typically uses the IGP metric to the next hop as the decision point.

The prefix 10.2.0.0/16 is injected by both R5 and R7 into BGP 65102. The prefix is then advertised by R5 to both R4 and R6 via eBGP. The prefix is also advertised by R7 to R4 and R6 via eBGP. The routers R4 and R6 both have to make a path selection, which results in a decision based on router ID, unless the path-selection process is manipulated through modification of one of the BGP attributes, or the BGP multipath feature is enabled. The prefix advertisements are shown in Figure 5-11. Examples 5-16 and 5-17 show the path selection. Both R4 and R6 select the path from R5.

Figure 5-11 *Prefix Advertisements*



Example 5-16 *R4 Selecting the Path from R5*

```
R4#show ip bgp 10.2.0.0
BGP routing table entry for 10.2.0.0/16, version 6
Paths: (3 available, best #3, table Default-IP-Routing-Table)
Flag: 0x200
  Advertised to peer-groups:
     internal
  Advertised to non peer-group peers:
  172.17.2.2
```

continues

Example 5-16 *R4 Selecting the Path from R5 (Continued)*

```
  65102
    172.16.6.1 (metric 409600) from 172.16.6.1 (172.16.6.1)
      Origin IGP, metric 0, localpref 100, valid, internal
  65102
    172.17.2.2 from 172.17.2.2 (172.16.7.1)
      Origin IGP, metric 0, localpref 100, valid, external
  65102
    172.17.2.1 from 172.17.2.1 (172.16.5.1)
      Origin IGP, metric 0, localpref 100, valid, external, best
```

Example 5-17 *R6 Selecting the Path from R5*

```
R6#show ip bgp 10.2.0.0
BGP routing table entry for 10.2.0.0/16, version 8
Paths: (3 available, best #3, table Default-IP-Routing-Table)
Flag: 0x210
  Advertised to peer-groups:
     internal
  Advertised to non peer-group peers:
  172.17.2.2
  65102
    172.17.2.2 from 172.17.2.2 (172.16.7.1)
      Origin IGP, metric 0, localpref 100, valid, external
  65102
    172.16.4.1 (metric 409600) from 172.16.4.1 (172.16.4.1)
      Origin IGP, metric 0, localpref 100, valid, internal
  65102
    172.17.2.1 from 172.17.2.1 (172.16.5.1)
      Origin IGP, metric 0, localpref 100, valid, external, best
```

Enabling eBGP multipath on both R4 and R6 allows traffic to be load-shared between R5 and R7 instead of only one of them being chosen to receive all traffic inbound to AS 65102. Examples 5-18 and 5-19 show the changes.

Example 5-18 *R4 Load-Balancing Between R5 and R7*

```
R4#show ip bgp 10.2.0.0
BGP routing table entry for 10.2.0.0/16, version 10
Paths: (3 available, best #2, table Default-IP-Routing-Table)
  Advertised to peer-groups:
     internal
  Advertised to non peer-group peers:
  172.17.2.1
  65102
    172.16.6.1 (metric 409600) from 172.16.6.1 (172.16.6.1)
      Origin IGP, metric 0, localpref 100, valid, internal
```