White Paper

Border Gateway Protocol Best Practices

By Clifton Funakura

The Internet has grown into a worldwide network supporting a wide range of business applications. Many companies depend on the Internet for day-to-day activities, whether or not e-commerce is their core competency. People are using the Internet for news, shopping, product research, customer service, and entertainment. Customer behavior is changing and a growing number of people are looking online to find products and services.

As more services converge to transport over Internet Protocol (IP), Internet reliability and performance become major concerns. While IP enables communication over the Internet, the protocol that facilitates routing on the Internet is Border Gateway Protocol (BGP) Version 4. Discussions centering on BGP often arise when Verizon® customers want to add Internet circuits for additional bandwidth or redundancy.

BGP Revealed

BGP basically determines how an Autonomous System (AS), or independent network, passes packets of data to and from another AS. Rather than depend on a calculated metric to determine the best path, BGP uses attribute information that is included in route advertisements to determine the chosen path.

BGP sessions are established on a neighbor-to-neighbor basis allowing for per-neighbor route filtering or modification of BGP attributes. The granular control over BGP attributes on a per-neighbor basis allows an AS to engineer traffic exchange points and influence inbound and outbound traffic patterns to and from a specific AS. BGP speakers (routers) exchange IP prefix (or route) advertisements and their associated attributes with neighbor BGP routers. If multiple advertisements are received for the same IP Prefix, the BGP process will inspect the attributes in a series of steps to determine the path to be used.

BGP attributes can be well known—recognized by compliant BGP implementations—or optional—recognized by some implementations. They also can be mandatory or discretionary. Although vendors may add proprietary parameters that may be used for path selection, the attributes used for path selection are based on Requests for Comment (RFC) and are typically recognized by most BGP implementations. The following table reviews the path selection algorithms that are implemented by Cisco and Juniper. If a choice must be made between multiple paths for a particular route, the router steps through the list from the top down, moving to the next step only if there is a tie.

Cisco	Juniper Networks
Path with the highest weight	
Path with the highest local preference	Path with the highest local preference
Path locally originated	
Path with the shortest AS path	Path with the shortest AS path
Path with the lowest origin code	Path with the lowest origin code
Path with the lowest MED	Path with the lowest MED
	Prefer strictly internal paths (IGP or locally generated routes)
Prefer EGBP over IBGP path	Prefer EGBP over IBGP path
Path with the lowest IGP metric to the BGP next hop	Path with the lowest IGP metric to the BGP next hop
	Path with the largest number of next hops
Prefer path that is oldest	
Path from BGP router with lowest router ID	Path from BGP router with lowest router ID
Path from lowest neighbor address	Path from lowest neighbor address

Table 1: BGP Path Selection Algorithms

This Information is reprinted with permission from Cisco and Juniper Networks. Please see Cisco and Juniper Networks technical documentation for additional detail. Additional selection rules may apply depending on router configuration options or the use of BGP confederations, route reflectors, and multiple paths.

Multiple Circuits

When adding other Internet circuits, customers often ask if BGP is required. If additional bandwidth is the primary consideration, BGP is not required. If diversity and redundancy are the primary considerations, BGP is configured to provide automatic fail-over. Four common methods to provision multiple circuits are double, bonding, shadow, and diverse. In the following examples, primary and additional circuits are provisioned to Verizon as the Internet Service Provider (ISP). Router configuration examples use the Cisco IOS Command Line Interface (CLI).

Double

If additional bandwidth is desired, a second identical circuit can be ordered for a double configuration. Verizon supports double configurations (such as double T1) for T1, T3, and OC-3 through OC-12 circuits.

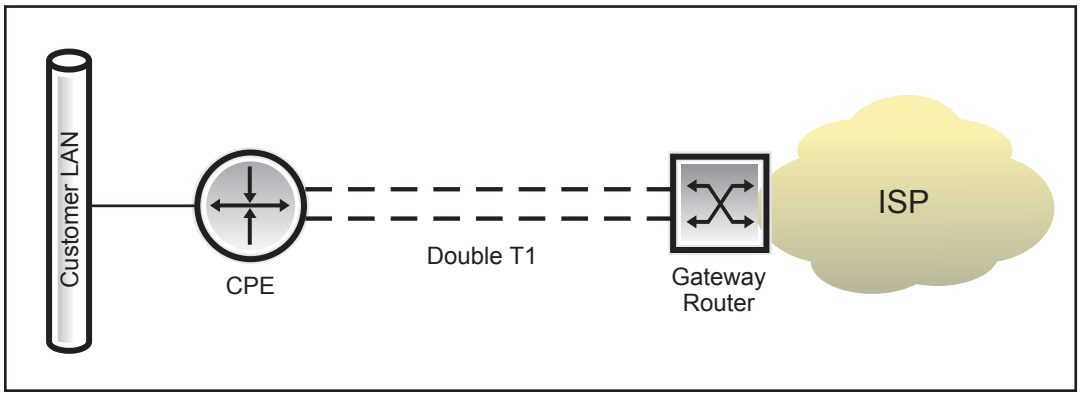

Figure 1: Double T1

The customer router, or customer premises equipment (CPE), and the Verizon gateway routers are configured with equal-cost, static-default routes. This provides a level of load-sharing, because the routers use both links to reach a destination. Link selection is typically based on destination, and a larger number of destination hosts will result in better load-sharing.

CPE configuration example (only relevant commands are shown):

IP route 0.0.0.0 0.0.0.0 s0
IP route 0.0.0.0 0.0.0.0 s1

BGP is supported over double-circuit access. The BGP session should be established between router loop-back addresses, allowing some resiliency should one of the circuits fail. Since External BGP (EBGP) sessions require directly attached neighbors, and loop-back addresses are one logical hop away, BGP multihop must be configured to establish the BGP session.

Bonding

Another option that allows for additional bandwidth is the use of a multilink protocol to logically bind multiple circuits. Use of a multilink protocol like Multilink PPP (MLPPP) or Multilink Frame Relay (MFR) is far more efficient at load-balancing than equal-cost static routing, but it requires multilink protocol support on both ends of the circuit and may require special hardware.

BGP can be configured over NxT1 circuits bonded with a multilink protocol. Verizon supports BGP over NxT1 access.

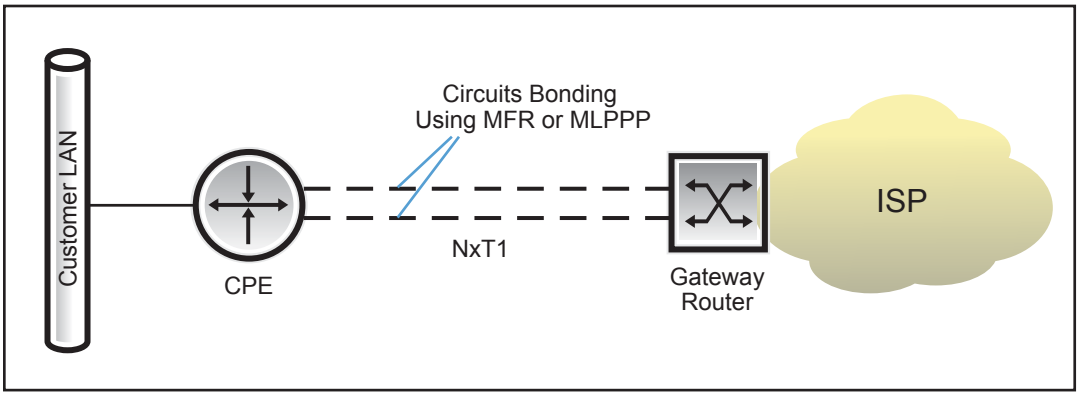

Figure 2: Logically Bonded Circuits

Verizon offers NxT1 access options up to 8xT1 (12 Mbps) for Internet Dedicated Access, or access to the Verizon Private IP Network (MPLS-based VPN service). NxT1 Internet Dedicated Access is certified using MFR. NxT1 access using MLPPP is permitted to access the Verizon Private IP network.

Shadow

If redundancy is the primary goal, a second identical circuit can be ordered for shadow service. Shadow service uses BGP to logically create a primary and back-up circuit. The back-up circuit is provisioned to an alternate gateway router at a diverse Verizon point of presence (PoP).

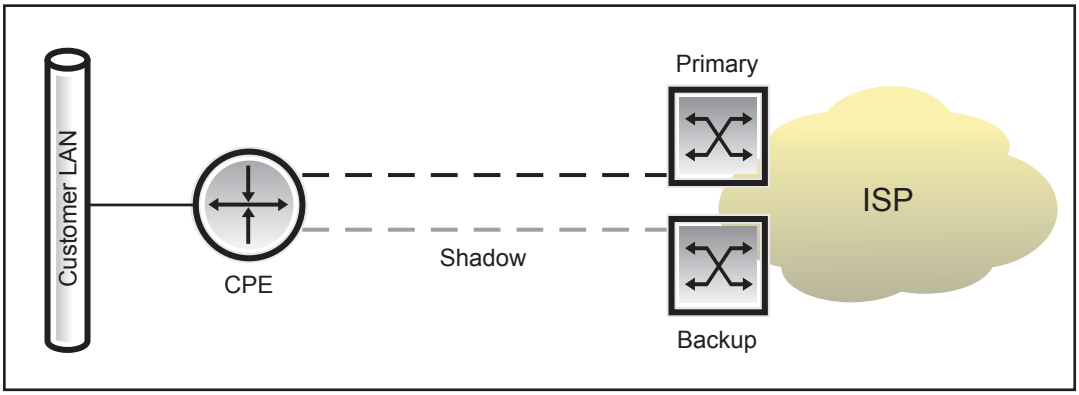

Figure 3: Shadow Circuit

To create a primary and back-up circuit, BGP sessions are established between the customer router and the two Verizon gateway routers. The customer IP prefix is advertised on both circuits, but the back-up circuit is configured with a higher Multi-Exit Discriminator (MED) attribute to make the path less-preferred.

CPE configuration example:

```
router bgp 7046
  network a.b.c.0 mask 255.255.255.0
  neighbor 1.1.1.1 remote-as 701
  neighbor 2.2.2.2 remote-as 701
  neighbor 2.2.2.2 route-map shadow in
  neighbor 2.2.2.2 route-map shadow out

route-map shadow permit 10
  set metric 10
```

AS 7046 is a generic reusable AS number for Verizon-only customers. The network statement advertises the customer IP prefix: a.b.c. 0/24. The primary BGP neighbor IP is 1.1.1.1; the back-up BGP neighbor is 2.2.2.2. Route maps are used to add a higher MED (metric) attribute to routes sent and received from the back-up neighbor (default metric=0). In the event that the primary circuit fails, the primary BGP session is removed and traffic is routed to the back-up path.

Diverse

Verizon offers a diverse service that builds on the redundancy offered through the shadow service. When diverse service is provisioned, both circuits are active and back each other up. To achieve some level of load-sharing, the customer IP prefix can be split in half so that one half of the IP block is advertised on the first circuit and the other half is advertised on the second circuit. The complementary half of the IP block is also advertised, but it is advertised with a higher MED to provide a back-up path. The higher metric causes the path to be less-preferred for inbound traffic. Inbound traffic uses one of the two circuits, depending on which half of the address block the source host belongs. In the event of a circuit, gateway, or PoP failure, all inbound traffic will use the surviving circuit.

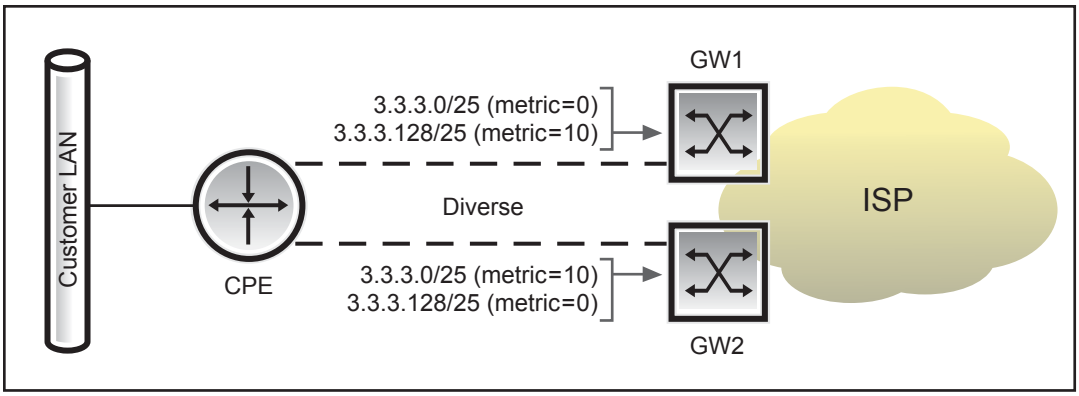

Figure 4: Diverse Service

Outbound traffic is less predictable than inbound traffic since the destination IP address of an Internet destination varies. Equal-cost static routes are configured to load share outbound traffic.

CPE configuration example:

```
router bgp 7046
  network 3.3.3.0 mask 255.255.255.128
  network 3.3.3.128 mask 255.255.255.128
  neighbor 1.1.1.1 remote-as 701
  neighbor 1.1.1.1 route-map GW1 out
  neighbor 2.2.2.2 remote-as 701
  neighbor 2.2.2.2 route-map GW2 out
```

```
ip prefix-list 1 seq 5 permit 3.3.3.0/25
ip prefix-list 1 seq 10 deny 0.0.0.0/0 le 32

ip prefix-list 2 seq 5 permit 3.3.3.128/25
ip prefix-list 2 seq 10 deny 0.0.0.0/0 le 32

route-map GW1 permit 10
match ip address prefix-list 1
!
route-map GW1 permit 20
match ip address prefix-list 2
set metric 10
!
route-map GW1 deny 30

route-map GW2 permit 10
match ip address prefix-list 1
set metric 10
!
route-map GW2 permit 20
match ip address prefix-list 2
!
route-map GW2 deny 30

ip route 3.3.3.0 255.255.255.128 Ethernet 0
ip route 3.3.3.128 255.255.255.128 Ethernet 0

ip route 0.0.0.0 0.0.0.0 Serial 0
ip route 0.0.0.0 0.0.0.0 Serial 1
```

In this example the site IP prefix is 3.3.3.0/24, split into two /25 IP prefixes.

The Gateway 1 (GW1) IP address is 1.1.1.1. The Gateway 2 (GW2) IP address is 2.2.2.2. Route maps are used to generate specific route advertisements and add a high MED attribute to the complementary half of the IP block for back-up. Static routes to the local interface add the /25 subnets to the local routing table so they may be advertised by BGP. Static default routes to Serial 0 and Serial 1 will share outbound traffic.

In practice, it is common to split the advertisements by announcing the entire IP block with a smaller subnet (e.g., 3.3.3.0/24 and 3.3.3.128/25) rather than two smaller subnets. This preserves the larger IP block advertisement while maintaining a degree of inbound load-sharing.

Multihoming

To provide an additional level of redundancy, customers may choose to have one circuit provisioned to one ISP and another circuit provisioned to an alternate ISP. This is commonly referred to as “multihoming,” and provides circuit redundancy and some protection against an ISP network failure. BGP configuration can become complex when designing a topology that is fault-tolerant, especially when trying to fine tune traffic behavior across multiple ISP circuits.

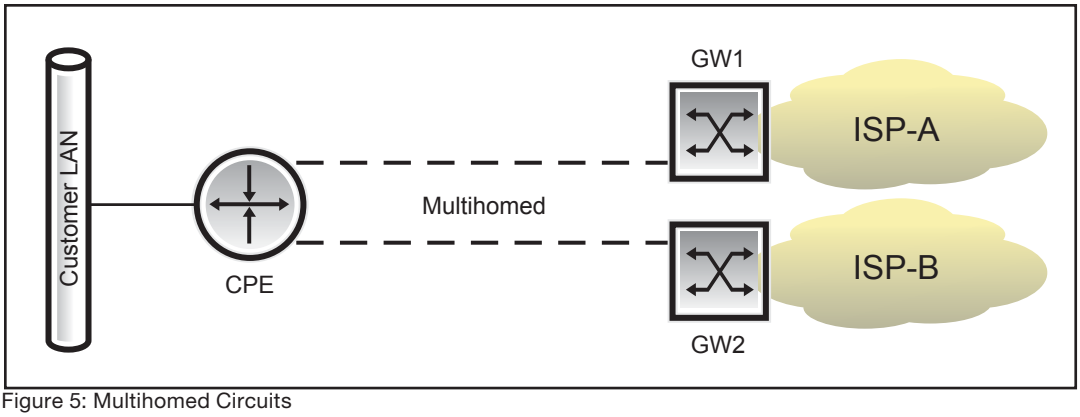

Figure 5: Multihomed Circuits

One topic that arises during a BGP discussion is the transmission of full or partial route tables from the ISP. Verizon can provide full routes, a default route, or various levels of partial route table information. If a customer is single-homed to an ISP, they do not usually require Internet route table information and can simply default route to the ISP. However, multihomed customers may benefit by receiving route tables from each provider so their BGP routers will have more information from which path selection can be made. For example, if a third party connects to the Internet through ISP-A, and a customer has BGP tables from both ISP-A and ISP-B, BGP will choose the direct path to the third party through ISP-A since the AS-path through ISP-B is longer and less-preferred.

As previously described, the BGP path selection algorithm favors ISPs with better connectivity and peering arrangements. The AS path is a well-known, mandatory BGP attribute included in BGP route advertisements. The BGP path selection process often comes down to the shortest AS path. A shorter AS path means that a packet must traverse fewer networks to reach a destination. It is reasonable to deduce that a closer destination should result in better reliability and more consistent end-to-end performance.

Redundant CPE

As customers develop a topology that is increasingly fault-tolerant, the edge router is identified as a single point of failure. Adding a second edge router can provide another layer of redundancy, but it also greatly increases complexity.

Active/Standby

The simplest scenario is an Active/Standby configuration where one router and circuit are configured as primary, and another router and circuit are configured for back-up. No route tables are received, and both routers default route to their respective ISPs.

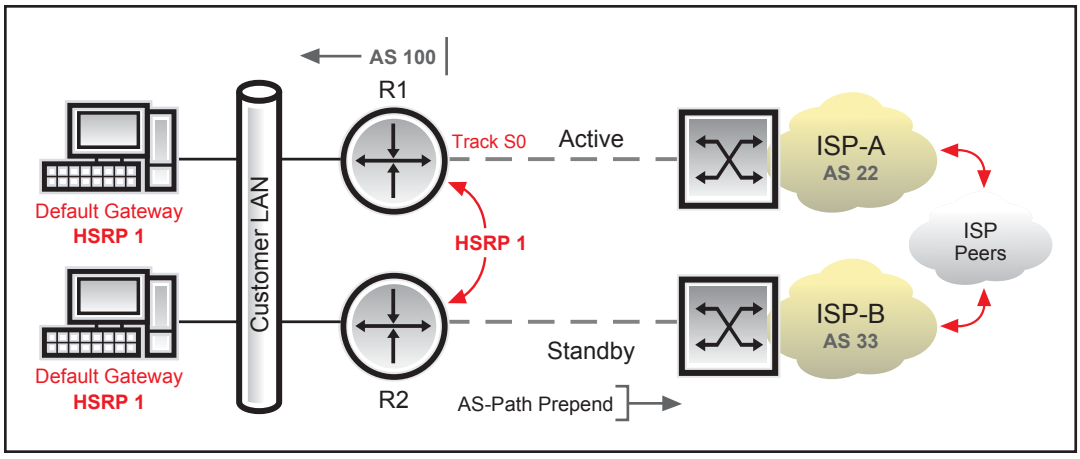

Figure 6: Dual Routers—One Active, One Standby

This scenario can make use of dynamic router fail-over protocols like Cisco's Hot Standby Routing Protocol (HSRP) or the non-proprietary Virtual Router Redundancy Protocol (VRRP). These protocols create a virtual IP address between the two routers that can be used as a single default gateway for LAN hosts. One router can be configured as active (master) and the other will be standby (back-up). If the primary router fails, the back-up router will activate. HSRP and VRRP also can be configured to track the status of the primary interface so that if it goes down, the back-up router activates.

HSRP/VRRP will send outbound traffic through the active/master router. To influence inbound traffic to use the primary circuit, the path through the back-up circuit can be made less desirable by using the prepending AS path. See the Inbound/Outbound Traffic Tuning section for more information.
CPE configuration example:

```
Router R1:
    interface serial 0
    ip address 192.168.22.1 255.255.255.0

    interface Ethernet1
      ip address 192.168.12.1 255.255.255.0
      standby 1 priority 105
      standby 1 ip 192.168.12.251
      standby 1 track Serial0

    router bgp 100
      network 192.168.12.0 mask 255.255.255.0
    neighbor 192.168.12.2 remote-as 100
    neighbor 192.168.12.2 next-hop-self
      neighbor 192.168.22.2 remote-as 22

Router R2:
    interface serial 0
    ip address 192.168.33.2 255.255.255.0

    interface Ethernet1
     ip address 192.168.12.2 255.255.255.0
     standby 1 priority 100
     standby 1 ip 192.168.12.251

    router bgp 100
     network 192.168.12.0 mask 255.255.255.0
     neighbor 192.168.12.1 remote-as 100
     neighbor 192.168.12.1 next-hop-self
     neighbor 192.168.33.3 remote-as 33
     neighbor 192.168.33.3 route-map backup out

    access-list 1 permit 192.168.12.0
    route-map backup permit 10
     match ip address 1
     set as-path prepend 100
```

If IP address space is provided by one of the ISPs, such as ISP-A, the IP prefix provided to the customer is most likely part of a larger aggregate. Unless otherwise instructed, ISP-A probably will advertise only the larger aggregate to its peers. This creates an issue if ISP-B is advertising the more specific IP Prefix. Routers will always choose the most specific route. Traffic will always come inbound through ISP-B if ISP-A advertises the aggregate. To fix this problem, ask the ISP that provides the address space to advertise the more specific IP prefix along with the aggregate.

Active/Active

Having a second circuit sitting idle may not be cost-effective or may be difficult to justify. Customers ask for creative ways to pass traffic over both circuits while maintaining automatic fail-over. There are many variables involved when designing a fault-tolerant, load-sharing topology. The complexity may require a high level of staff

expertise. Companies have developed specialized products in response to the need for effective redundancy and load-sharing that is easy to configure and manage.

For illustration purposes, assume a fairly simple site topology that builds on the previous Active/Standby example. The site has a single IP prefix with two Internet circuits homed to different ISPs for diversity. BGP should be configured to allow traffic to pass over both Internet circuits and provide resiliency if a circuit fails.

Although there may be some creative ways to handle outbound circuit selection (e.g., policy routing, NAT, etc.), the following example builds on the Active/Standby scenario described previously and uses HSRP groups to configure active routers and provide resiliency. Two HSRP groups will be used—one for each circuit—and hosts on the LAN must use one of the HSRP IP addresses as their default gateway. Each router will be active for one group and standby for the other, complementing each other. If a circuit or router fails, the surviving router will be active for both groups.

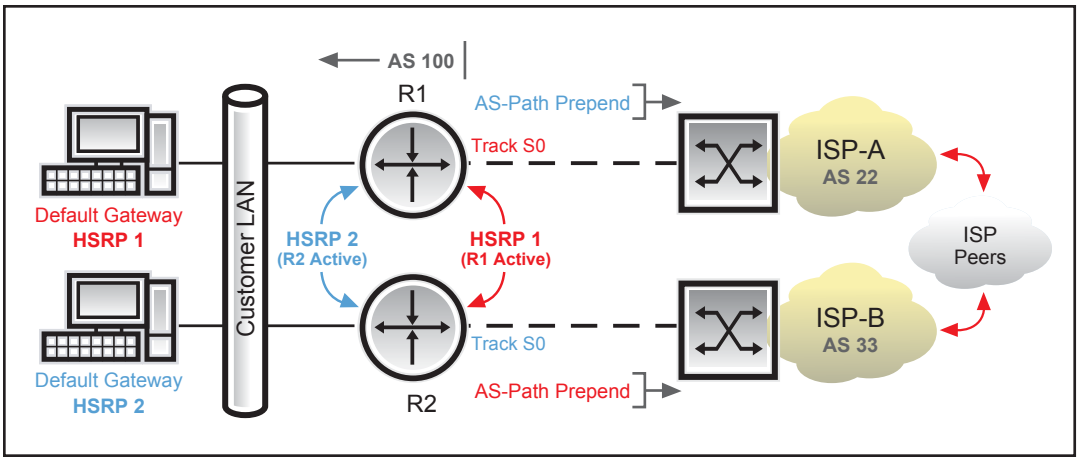

Figure 7: Dual Routers—Both Active

Ideally traffic that is sent over one Internet circuit should use the same circuit for return traffic—response packets sent from receiver back to sender. The term “asymmetric routing” describes the behavior of return traffic that comes in through a different path. This is not desired. Although the impact may be minimal if traffic simply uses two diverse gateways on a single ISP’s network, the impact on latency, jitter, and packet loss can be significant if outbound traffic traverses one ISP’s network and return traffic traverses one or more secondary ISP’s networks. In an effort to avoid asymmetric routing, AS path prepending is configured so the same circuit used for outbound traffic is preferred for inbound traffic.

This example splits the site’s IP block to load share inbound traffic, as described in the Diverse section; however, be sure to check with your ISP to determine the smallest IP prefix that can be advertised and passed to peers. Most ISPs will filter IP prefix advertisements to peers that are smaller than /24.

CPE configuration example:

```
Router R1:
    iinterface serial 0
    ip address 192.168.22.1 255.255.255.0

    interface Ethernet1
      ip address 192.168.12.1 255.255.255.0
      standby 1 priority 105
      standby 1 ip 192.168.12.251
      standby 1 track Serial0
      standby 2 priority 100
      standby 2 ip 192.168.12.252

    router bgp 100
      network 192.168.12.0 mask 255.255.255.128
      network 192.168.12.128 mask 255.255.255.128
      neighbor 192.168.12.2 remote-as 100
```

```
    neighbor 192.168.12.2 next-hop-self
    neighbor 192.168.22.2 remote-as 22
    neighbor 192.168.22.2 route-map PREPEND-BLUE out

  ip prefix-list 1 seq 5 permit 192.168.12.128/25
  ip prefix-list 1 seq 10 deny 0.0.0.0/0 le 32

  route-map PREPEND-BLUE permit 10
     match ip address prefix-list 1
     set as-path prepend 100 100 100

  route-map PREPEND-BLUE permit 20

  ip route 192.168.12.0 255.255.255.128 Ethernet1
  ip route 192.168.12.128 255.255.255.128 Ethernet1

Router R2:
  interface serial 0
  ip address 192.168.33.2 255.255.255.0

  interface Ethernet1
    ip address 192.168.12.2 255.255.255.0
    standby 1 priority 100
    standby 1 ip 192.168.12.251
    standby 2 priority 105
    standby 2 ip 192.168.12.252
    standby 2 track Serial0

  router bgp 100
    network 192.168.12.0 mask 255.255.255.128
    network 192.168.12.128 mask 255.255.255.128
    neighbor 192.168.12.1 remote-as 100
    neighbor 192.168.12.1 next-hop-self
    neighbor 192.168.33.3 remote-as 33
    neighbor 192.168.33.3 route-map PREPEND-RED out

  ip prefix-list 2 seq 5 permit 192.168.12.0/25
  ip prefix-list 2 seq 10 deny 0.0.0.0/0 le 32

  route-map PREPEND-RED permit 10
    match ip address prefix-list 2
    set as-path prepend 100 100 100

  route-map PREPEND-RED permit 20

  ip route 192.168.12.0 255.255.255.128 Ethernet1
  ip route 192.168.12.128 255.255.255.128 Ethernet1
```

In this example, the site (AS 100) IP Prefix is 192.168.12.0/24. This was split into two /25 IP prefixes and advertised on both circuits, with one IP prefix set with a higher MED (for back-up). Two HSRP groups were created, HSRP 1 (Red, R1 Active) and HSRP 2 (Blue, R2 Active). Hosts on the LAN are configured to use either the IP of HSRP group 1 or group 2 as their default gateway.

Outbound traffic using HSRP 1 will be processed by R1. If default routes are sent by both ISPs, outbound traffic from R1 will be sent to ISP-A since the EBGP route to ISP-A is preferred over the Internal BGP (IBGP) route learned from R2—all other BGP attributes being equal. Similarly, outbound traffic using HSRP 2 will be processed by R2 and will be sent to ISP-B.

AS path prepending is configured to influence inbound traffic to use the same circuit as outbound traffic. Advertisements for the back-up IP prefix have AS 100 prepended to make the AS path longer and the path less desirable.

If route tables are sent by both ISPs, it is possible for R1 to send packets to ISP-B through R2 if the route to ISP-B has a shorter AS-Path. Although the path from the site to the Internet destination may be shorter, the return traffic may use an alternate path (asymmetric routing) since the return traffic path is based on the source IP.

Inbound/Outbound Traffic Tuning

If a site advertises an IP prefix out two or more Internet circuits homed to different ISPs, it is highly unlikely that there will be a perfect balance of inbound traffic across all circuits. Inbound traffic may come from anywhere in the worldwide Internet. If a source host is on the same ISP network as the site's Internet circuit, the traffic will remain on that ISP's network and come in on the corresponding circuit. Sources that are more distant may use a path that is determined by the best ISP peering arrangement (shortest AS path).

To influence inbound traffic, the MED attribute can be used as previously described. Note that the MED attribute is nontransitive and will not be passed through to other Autonomous Systems. Another option is AS-path prepending, where the AS path is made artificially longer to cause a path to be less-preferred. AS path prepending adds one or more extra AS numbers that will be added to the cumulative AS-path information passed to peers.

A preferred outbound exit point for an AS can be specified using the local preference attribute. The highest local preference is chosen as the exit point. Different exit points can be assigned a higher or lower preference depending on network policy. If an exit circuit or router fails, the next exit with the next highest local preference is chosen. By setting local preference values for specific IP prefixes, outbound traffic can be engineered intelligently.

Cisco routers typically use route maps to apply attributes to IP prefixes. A route map can directly set attributes on route advertisements or can be configured as a series of statements to create a logical "if-then-else" framework (as shown in the examples).

Verizon provides a set of BGP community values that can be sent with route advertisements and will subsequently set BGP attributes, limit route advertisements, or black hole (drop) routes within the Verizon network. For example, if a route advertisement is tagged with the community 701:120, the local preference on the Verizon gateway router will be set to 120. This value is higher than the default of 100, so the circuit will be the preferred exit point from Verizon to the customer's AS.

Partial routing may also be used in an effort to send traffic over multiple circuits. If one ISP sends partial route tables and a default route is sent from the second ISP, traffic will use the first ISP for destinations in the partial route table, and the second ISP will be used to reach all other destinations. Be aware that there is a high level of unpredictability associated with this scenario. Asymmetric routing is likely.

Verizon's Offerings

Verizon provides shadow, double, and diverse services, along with Internet bandwidth options ranging from fractional T1 to OC-48. Verizon can supply full route tables, a default route, or a range of partial route table options as desired.

It is important to understand BGP and its role in the Internet to effectively design for resiliency, consistent performance, and implementation of network policy. Verizon offers a set of community values that can be used to set BGP attributes, contain route advertisements, or even reduce the impact of a Distributed Denial of Service (DDoS) attack.

The global Internet has become a key component of conducting business and provides an efficient network for transfer of application data, voice, and video. The Internet also provides a medium for businesses to reach consumers in a global marketplace. Many businesses rely on Internet connectivity, which is why, more than ever, reliability and performance are critical factors. Verizon gives its customers that reliability and performance. Let Verizon be the solution to your business requirements.

Appendix

Verizon Community Strings

Verizon Internet Dedicated customers that run BGP can send community strings with their route advertisements to modify the Verizon Local Preference, increase the AS path length, control which routes are advertised to peers, contain routes within a continental AS, or black hole (drop) traffic at the Verizon network edge, to reduce the impact of a DDoS attack.

The following table lists the community tags that would be used for AS 701, associated BGP attribute and value (if applicable), and resulting action.

Community Tag	Attribute	Value	Action
701:80	Local Pref	80	Least preferred
701:90	Local Pref	90	Less preferred
701:100	Local Pref	100	Default
701:110	Local Pref	110	More preferred
701:120	Local Pref	120	Most preferred
701:1	AS Path	701 701	Prepend the Verizon AS 701 one time
701:2	AS Path	701 701 701	Prepend the Verizon AS 701 two times
701:3	AS Path	701 701 701 701	Prepend the Verizon AS 701 three times
701:20			Keep route within Verizon Autonomous Systems (do not send to peers)
701:30			Refer to Note 1 (keep route within regional Verizon AS and do not send to peers)
701:9999			Refer to Note 2 (black hole the route at Verizon network edge devices)

Table A-1

Note 1: Verizon has several AS numbers assigned to different regions of the world:

701 North America
702 Europe/Middle East/Africa
703 Asia-Pacific
14551 Latin America/South America

The AS portion of the community tag values should be changed appropriately for use in different Verizon regions.

Note 2: For technical reasons, customers who wish to use the black hole community must enable EBGP multi-hop (although BGP is still configured between the circuit addresses).

verizonbusiness
We never stop working for you.