VRRP Failover with BGP on VPN Concentrators

Overview

This document encompasses a detailed overview of failover behavior between high availability VPN concentrators operating BGP. A reference architecture will be provided along with details on the protocols involved (TCP, BGP, VRRP). Additionally, expected behavior will be discussed and what this means for the headend WAN architecture.

Reference Architecture

In a Normal state, we have two VPN concentrators participating in the Virtual Router Redundancy Protocol (VRRP). The primary (active) being on the right with the spare (passive) on the left. EBGP sessions are bound to the shared virtual IP. (In this scenario 192.168.1.5.)

Topology

In the above architecture, the BGP Hold Timer between 192.168.1.5 and the upstream EBGP peer is 240 sec (It can be adjusted on the Meraki platform). Additionally, VRRP Heartbeats are sent every second from the Primary MX uplink IP (192.16.1.3) and the Spare MX uplink IP (192.168.1.4).
For more in-depth information regarding the VRRP mechanism on the MX, please see the HA Failover Behavior documentation.

**Failover Scenario: Primary Uplink Fails**

During failover, VRRP enables the spare VPN concentrator to assume the role of the active and establishes EBGP sessions bound to the virtual uplink IP, further sending and receiving all BGP messages (Updates, Notifications, etc.)

If the failover was caused by the primary VPN concentrator losing connectivity with the rest of the network, the TCP connection supporting the EBGP connection may appear alive at the old primary in spite of the EBGP peer having torn it down and now being connected to the spare VPN concentrator that took over the role of active. The old primary will not detect that the BGP session has failed until the underlying TCP connection is torn down.

Below is a diagram illustrating BGP convergence occurring during the failover scenario:
In the above diagram, a branch site removes a local route from being advertised in BGP; generating a BGP Update message withdrawing the 192.168.100.0/24 prefix from the route table for the spare (which is the current active). However, since the current active has its primary uplink down, no IBGP messages are received on the device's control plane. Thereby, resulting in the stale route still being present in the route table during failover.

In the above diagram, the datacenter edge router removes an IBGP learned route from its RIB. The EBGP peer generates a BGP Update message withdrawing the 10.10.10.0/24 prefix. The route table for the spare (which is the current active) is updated. However, since the current active has its primary uplink down, no EBGP messages are received on the device's control plane. Thereby, resulting in the stale route still being present in the route table during failover.
Primary Uplink Fallback Scenario

During failback, the old primary may think the EBGP connection is still established if it was not torn down during failover. The old primary will not detect that the BGP session has failed until the underlying TCP connection is torn down.

Since BGP keepalives will be transmitted every ⅔ of the Hold Timer interval (60 sec in this scenario), in the worst case, the EBGP peer will respond to such a BGP Keepalive with a TCP RST that will cause the connection to be torn down, triggering re-establishment of an EBGP session over a fresh TCP connection between the new active and the EBGP peer. As a result, any changes in route state that occurred in between the initial failover and the EBGP session being reestablished, will not be available at the new active in between the beginning of the fallback and the EBGP session being reestablished.

The below diagram illustrates the message exchange between the Active and upstream EBGP peer:

1) Primary sends Keep Alive at 60 sec (worst case scenario)
2) TCP RST is received on master
3) Spare is now receiving all BGP Control messages (Updates, Notifications etc) as described in above diagrams
4) Primary flushes all routes in RIB and BGP state is now Idle. TCP connection torn down
The below diagram illustrates the message exchange when the primary uplink recovers:

The re-establishment of the BGP session causes the primary MX at the headend to purge all BGP connections for which it received a TCP RST on. (This includes IBGP and EBGP connections.) During this time of reconvergence, branch traffic will redirect flows to its backup DC.

**Summary**

The Cisco Meraki high availability solution described in this document provides secure, reliable and scalable datacenter redundancy. This allows for the most efficient use of resources between datacenters and respective branch offices. Additionally, this document encompassed a detailed step-by-step overview when incorporating Auto VPN with BGP in your WAN architecture.