UNDERSTANDING CONVERGENCE IN MPLS VPN NETWORKS

Mukhtiar A. Shaikh (mshaikh@cisco.com)
Moiz Moizuddin (mmoizudd@cisco.com)
Agenda

• Introduction
• Convergence Definition
• Expected (Theoretical) Convergence
• Test Methodology
• Day in the Life of VPN Routing Update
• Observed Up Convergence
• Observed Down Convergence
• Design Considerations
• Summary
Importance of Convergence in L3VPN-Based Networks

- Convergence in the traditional overlay Layer 2 VPNs is pretty fast
- In the traditional Layer 2 VPN Frame or ATM-based networks, Service provider network is not a factor for Layer 3 convergence
- Customers are now moving to VPN services based on Layer 3 infrastructure (aka RFC 2547 based VPNs)
- It is necessary to understand the factors which impacts the L3VPN convergence and how it can be improved
- Convergence varies depending on the network size, PE-CE protocol, redundancy options, etc.
- Default convergence in the MPLS VPN networks could be very high in the order of 60+ secs…but not always 😊
Agenda

• Introduction
• **Convergence Definition**
• Expected (Theoretical) Convergence
• Test Methodology
• Day in the Life of VPN Routing Update
• Observed Up Convergence
• Observed Down Convergence
• Design Considerations
• Summary
Convergence Definition
What Is Convergence in MPLS VPN Networks?

• Convergence is the time it takes for the data traffic from the remote CE to reach the local CE after a topology change has occurred in the network
What Is Up Convergence??

- **Up convergence** in L3VPN environment can be defined as the time it takes for traffic to be restored between VPN sites when:
  - A new prefix is advertised and propagated from a local CE to the remote CE, or
  - A new site comes up
- Up Convergence is applicable in cases where there is a backup link which comes up only after the primary goes down
- Or If we are using some sort of conditional advertisement
- Up convergence can be loosely defined as route advertisement from CE to CE
What Is Down Convergence??

- **Down convergence** can be defined as how fast the traffic is rerouted on an alternate path due to failure either in the
  - SP network
  - Customer network
  - (Primary) PE-CE link

- Down convergence can be loosely defined as withdrawal of best path
What Are Convergence Points in a MPLS VPN Network??

- Overall VPN convergence is the sum of individual convergence points
Summary (Theoretical Convergence)

- Two sets of timers; first set consists of T1, T4, T6 and T7; second set comprises of T2, T3, T5 and T8
- First set mainly responsible for the slower convergence unless aggressively tweaked down
- Theoretically sums up to ~ 85 seconds [30 (T1)+5*2 (T4)+15(T6)+30 (T7)]
- Once different timers are tuned, convergence mainly depends on T6; min T6=5 secs
- Assuming ~“x” secs for T2, T3, T5 and T8 collectively

<table>
<thead>
<tr>
<th>PE-CE Protocol</th>
<th>Max Conv. Time (Default Settings)</th>
<th>Max Conv. Time (Timers Tweaked Scan=5, Adv=0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGP</td>
<td>~85+x Seconds</td>
<td>~5+x Seconds</td>
</tr>
<tr>
<td>OSPF</td>
<td>~25+x Seconds</td>
<td>~5+x Seconds</td>
</tr>
<tr>
<td>EIGRP</td>
<td>~25+x Seconds</td>
<td>~5+x Seconds</td>
</tr>
<tr>
<td>RIP</td>
<td>~85+x Seconds</td>
<td>~5+x Seconds</td>
</tr>
</tbody>
</table>

Remote_PE

Remote_CE

Local_CE

Local_PE
Agenda

• Introduction
• Convergence Definition
• Expected (Theoretical) Convergence
• **Test Methodology**
• Day in the Life of VPN Routing Update
• Observed Up Convergence
• Observed Down Convergence
• Design Considerations
• Summary
Test Methodology

- Testing done with reasonably large MPLS VPN network
- Test tool was used for simulating the VPN sites, generating the VPN routing information and sending traffic to the VPN prefixes

- Total number of PEs used = 100
  - Total number of vrfs created = 1000 vrfs
    - 250 BGP sessions
    - 250 RIP instances
    - 20 OSPF sessions
    - 250 EIGRP
    - Remaining sessions (~230) configured with static routing

- Same RD was used for each VPN
- 100 routes per vrf in steady state
- One additional test vrf with 1000 prefixes
- Total number of VPN routes = 1000*100 = 100k*2
- 2-RR
  - Convergence measured for the test vrf
Test Cases Carried Out…

- Test case I—Default timers
  BGP import scanner = 15
  Advertisement interval = 30 (EBGP) and 5 (IBGP)

- Test case II—Tweak BGP advertisement interval
  BGP import scanner = 15
  Advertisement interval = 0

  ```
  router bgp 65001
  Address-family vpnv4
  neighbor a.b.c.d advertisement-interval 0
  ```

- Test case III—Tweak BGP import scanner
  BGP import scanner = 5
  ```
  router bgp 65001
  Address-family vpnv4
  bgp import scan 5
  Advertisement interval = default
  ```

- Test case IV—Tweak BGP advertisement and import Scanner timers
  BGP import scanner = 5
  Advertisement Interval = 0

  ```
  router bgp 65001
  Address-family vpnv4
  bgp import scan 5
  neighbor a.b.c.d advertisement-interval 0
  ```
Agenda

- Introduction
- Convergence Definition
- Expected (Theoretical) Convergence
- Test Methodology
- Day in the Life of VPN Routing Update
- Observed Up Convergence
- Observed Down Convergence
- Design Considerations
- Summary
Day in the Life of a VPN Update

- This section shows route propagation for a network using EIGRP as the PE-CE routing protocol.
- Other routing protocols would exhibit similar behavior with few exceptions (explained later on).
- Default BGP timers are used; adv-interval = 5s, import scanner = 15s.
- Up Convergence discussed as part of this case study.
Day in the Life of a VPN Update (Cont.)

- Routes are first installed in the VRF routing table
- If PE-CE protocol is non-BGP (in this case EIGRP), additional time (can be negligible for smaller number of prefixes) is needed to redistribute these routes into MP-BGP and generating VPNV4 prefixes
- Not all the prefixes are installed in the routing table at the same time as updating RT, FIB/LFIB takes some time
- Once update is sent with some prefixes (First batch) to RR, the iBGP adv-int (5 secs) kicks in on local PE

After Receiving the First EIGRP Update, within 100 Msec (Approx) in This Example First Batch of IBGP Updates are Advertised to RR

Jan 2 10:35:01.469 PST: BGP(2): 192.168.10.12 send UPDATE (prepend, chgflags: 0x820) 100:1:10.0.0.0/24

Jan 2 10:35:01.341 PST: RT(vpn1): add 10.0.0.0/24 via 192.168.1.54, eigrp metric [170/2560002816]
Day in the Life of a VPN Update (Cont.)

- When first batch is received on RR, routes are immediately sent to the remote PE

When processing the first batch of BGP updates, the route reflector (RR) sends the routes to the remote PE immediately. The first batch is processed within 5 seconds, and then the next batch is sent after 15 seconds. The BGP updates include the prefix 100:1:10.0.0.0/24, which is sent to the remote PE.

- In the mean time, more EIGRP prefixes are received from the CE, processed and installed in the VRF table on local PE router

While the BGP updates are being processed, additional EIGRP prefixes are received from the CE. These prefixes are processed and installed in the VRF table on the local PE router. However, these prefixes are subjected to the ad-interval and have to wait for a total of 5 seconds before they are sent to RR.

- But these prefixes are subjected to the adv-interval and have to wait for a total of 5 secs before they are sent to RR

After the first batch of updates is processed, more EIGRP prefixes are received and processed by the local PE router. These prefixes are subjected to the ad-interval and have to wait for a total of 5 seconds before they are sent to the RR.
Day in the Life of a VPN Update (Cont.)

- After the import scanner timer expires, remote PE installs the routes in the VRF table; FIB/LFIB gets updated both on the RP and LCs.
- Remote PE advertises the prefixes towards CE.

**Remote PE**

Jan 2 10:35:03.522 PST: BGP: ... start import cfg version = 0
Jan 2 10:35:03.986 PST: RT(vpn1): add 10.0.0.0/24 via 192.168.1.11, bgp metric [200/2560002816]

Compare with the Timestamp 10:35:02.000 When RR Sent the Update

**Remote_CE**

Remote_PE# Show ip bgp vpnv4 all 10.0.0.0
BGP routing table entry for 100:1:10.0.0.0./24, version 22
Paths: (1 available, best #1, table vpna)
Flag: 0x820
Not advertised to any peer 65501
192.168.1.11 (metric 30) from 192.168.10.12 (192.168.10.12)
Origin incomplete, metric 0, localpref 100, valid, internal, best
Extended Community: RT:100:1

**Local_CE**

Jan 2 10:35:04.042 PST: RT: add 10.0.0.0/24 via 193.1.1.1, eigrp metric [170/2560005376]

Routes Once Received at Remote CE are Processed and Installed in the RT Immediately
Day in the Life of a VPN Update (Cont.)

- Advertisement_interval expires on the local PE router and as a result it announces the second batch of routes to RR.
- Not all the updates could be processed before we suspend the process; advertisement-interval kicks in again (5s).

RR Receives the Second Batch of Updates
*Jan 2 10:35:07.910: BGP(2): 192.168.1.11 rcvd 100:1:10.0.100.0/24

Remote PE Receives the Second Batch of Updates
Jan 2 10:35:08.238 PST: BGP(2): 192.168.10.12 rcvd UPDATE w/ attr: nexthop 192.168.1.11

Remote PE Receives the Second Batch of Updates
Jan 2 10:35:08.238 PST: BGP(2): 192.168.10.12 rcvd UPDATE w/ attr: nexthop 192.168.1.11

- But routes don’t get installed in the routing table but wait for up to 15 secs; (reason for spike in the graph discussed later)
Day in the Life of a VPN Update (Cont.)

- While remote PE is waiting for import-scan to expire, a third batch of updates is received.
- But again these updates are subjected to the scan-interval and wait for import-scanner (up to 15s) before they are installed in the VRF Routing Table on remote PE.
- Remaining prefixes get installed in routing table and are advertised to the remote CE

Local PE

BGP

VRF

Remote PE

BGP

VRF

3rd Batch

3rd Batch

Remote CE Receives All the Prefixes
Jan 2 10:35:13.380 : BGP(2): 192.168.1.11 rcvd 100:1:10.3.132.0/24

RR Receives the Third Batch of Updates
Jan 2 10:35:13.380 : BGP(2): 192.168.1.11 rcvd 100:1:10.3.132.0/24

Remote CE Receives All the Prefixes
Jan 2 10:35:20.526 PST: BGP: Import walker start version 4472514, end version 4473513

Jan 2 10:35:32.266 PST: RT: add 10.0.100.0/24 via 193.1.1.1, eigrp metric [170/2560005376]
Jan 2 10:35:22.998 PST: RT: add 10.3.132.0/24 via 193.1.1.1, eigrp metric [170/2560005376]
Agenda

- Introduction
- Convergence Definition
- Expected (Theoretical) Convergence
- Test Methodology
- Day in the Life of VPN Routing Update
- **Observed Up Convergence**
- **Observed Down Convergence**
- Design Considerations
- Summary
PE-CE=EIGRP/OSPF/BGP: Test Case I: BGP Import-Scan/Adv-Interval=Default (Cont.)

- Diagrams show the average, maximum, minimum, and median values for all prefixes measured across 100 iterations.
- Maximum convergence for each protocol is pretty close to the expected results.
- Average/median close to 30 secs for BGP and little over 20 secs for other protocols.
- Minimum convergence ranges from <1 sec for the first prefix to over 10 seconds for the last prefix.
- The difference is not linear but on the average, a jump of 10 Secs between the convergence of first and the last (1000th) prefix is seen.

<table>
<thead>
<tr>
<th>PE-CE Protocol</th>
<th>Max Conv. Time (Default Settings)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGP</td>
<td>~85+x Seconds</td>
</tr>
<tr>
<td>OSPF</td>
<td>~25+x Seconds</td>
</tr>
<tr>
<td>EIGRP</td>
<td>~25+x Seconds</td>
</tr>
</tbody>
</table>

Ref: Theoretical
What Are Those Jumps??

• Straight lines indicate that all prefixes converged almost at the same time
• Jumps indicate that some prefixes converged before others
• Jumps are because router is either waiting for advertisement interval or bgp import scanner interval or waiting for both timers to expire
PE-CE=EIGRP/OSPF/BGP:
Test Case IV: Import-Scan=5, Adv-Interval=0

- The last scenario offers the best convergence times
- The max is pretty close to 10 secs while average is ~5+ seconds

Ref: Theoretical

<table>
<thead>
<tr>
<th>PE-CE Protocol</th>
<th>Max Conv. Time (Default Settings)</th>
<th>Max Conv. Time (Timers Tweaked Scan=5, Adv=0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGP</td>
<td>~85+x Seconds</td>
<td>~5+x Seconds</td>
</tr>
<tr>
<td>OSPF</td>
<td>~25+x Seconds</td>
<td>~5+x Seconds</td>
</tr>
<tr>
<td>EIGRP</td>
<td>~25+x Seconds</td>
<td>~5+x Seconds</td>
</tr>
</tbody>
</table>
Summary Observed Up Convergence

- Most of the results are within the max theoretical limits
- Important observation is that cumulative convergence is not necessarily the simple addition of timers
- Especially there can multiple occurrences of T1, T4, T6, or T7 before all the prefixes have converged
- Tweaked timers improve convergence
Agenda

• Introduction
• Convergence Definition
• Expected (Theoretical) Convergence
• Test Methodology
• Day in the Life of VPN Routing Update
• Observed Up Convergence
• Observed Down Convergence
• Design Considerations
• Summary
## Failure Scenarios

- PE Failure
- RR Failure
- CE (Link/Node) Failure
- Failure in the MPLS Core

<table>
<thead>
<tr>
<th>Failure Scenario</th>
<th>Expected Max Convergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary PE-Failure</td>
<td>~ 65 Secs(_1)</td>
</tr>
<tr>
<td>RR Failure</td>
<td>~ 15Secs(_2)</td>
</tr>
<tr>
<td>CE (node/link) Failure</td>
<td>~ 60 Secs(_3)</td>
</tr>
<tr>
<td></td>
<td>~ 5 Secs(_4)</td>
</tr>
<tr>
<td>MPLS Core (Link/Node Failure)</td>
<td>~ 60 Secs</td>
</tr>
</tbody>
</table>

Later Slides Will Explain How We Got the Maximum Value

1. Assuming PE, RR Did Not Send a Notification to RR or to PE
2. Assuming Dual RRs and different RD case
3. Assuming CE Did Not Send a Notification to the PE Router
4. Assuming the PE-CE Link Failure Was Immediately Detected by the PE Router
PE Router Failure Scenario

- In this case we measure how long it takes Remote_PE to select Local_PE1 as the bestpath for prefix N2 when Local_PE2 goes down (provided PE2 was preferred path)
- When Local_PE2 goes down it takes BGP scan time (default 60s) for RR to detect that next-hop for N2 is gone down
- If Local_PE2 doesn’t crash but rather reloads then it may send a BGP notification to RRs to close the BGP session
- In this case RR would send an immediate withdraw to Remote_PE
Core Link Failure Scenario

- RR1 is reflecting PE2 as the bestpath and RR2 is reflecting PE1 as the best path; Remote_PE chooses the path from RR2 (i.e. PE1) as the bestpath.
- The BGP session between the RRs and PE-1 may not go down for 3 minutes (default holdtimer assumed).
- When next-hop inaccessibility is detected by the BGP scanner process (runs every 60 secs), remote PE would switch over to the alternate path.
Agenda

• Introduction
• Convergence Definition
• Expected (Theoretical) Convergence
• Test Methodology
• Day in the Life of VPN Routing Update
• Observed Up Convergence
• Observed Down Convergence
• Design Considerations
• Summary
Design Considerations

• Down convergence could be improved by using the NHT (next-hop tracking) feature.

• Risk of instabilities caused by BGP/routing churns as result of lowering to minimum values

  Careful setting of the advertisement interval both for both IBGP and EBGP sessions is needed

  Keeping the advertisement interval to 1 sec both for the IBGP and EBGP could prevent the unnecessary churn and at the same time could improve the convergence significantly
Design Considerations

- Fast IGP timers help improving the overall convergence in case of failure in the SP core
- Conditionally advertise only PE and P loopback addresses to reduce the number of prefix+label rewrites in the core failure event
- Use of default BGP behavior (bgp fast-fall-over)
- Use interface dampening and route dampening for customer links/sessions to prevent the churns
Agenda

• Introduction
• Convergence Definition
• Expected (Theoretical) Convergence
• Test Methodology
• Day in the Life of VPN Routing Update
• Observed Up Convergence
• Observed Down Convergence
• Design Considerations
• Summary
Summary

• Possible to get less than 5s convergence for small number of prefixes

• Maximum up convergence could be reduced down to ~5 secs when both the advertisement and import scanner are lowered to their min possible values

• While BGP is little slower to react, no major difference in the convergence across various PE-CE protocols once BGP timers tweaked

• For large number of prefixes convergence may happen in multiple batches
Q AND A