BGP Convergence in Virtual Private Networks

Dan Pei
AT&T Labs – Research
180 Park Ave
Florham Park, NJ
peidan@research.att.com

Jacobus Van der Merwe
AT&T Labs – Research
180 Park Ave
Florham Park, NJ
kobus@research.att.com

ABSTRACT

Multi-protocol label switching (MPLS) virtual private networks (VPNs) have had significant and growing commercial deployments. In this paper we present the first systematic study of BGP convergence in MPLS VPNs using data collected from a large tier-1 ISP. We combine several data sources to produce a methodology to accurately estimate routing convergence delays. We discovered an iBGP version of BGP path exploration, and show that the route invisibility problem occurs frequently and is one of the most significant contributing factors to BGP convergence delay in the VPNs we studied. We therefore propose and evaluate several configuration changes that can be employed to greatly improve the routing convergence time and minimize the connectivity disruption in the face of network changes.

Categories and Subject Descriptors

C.2 [**Computer Communication Networks**]: Network protocols, Network operations

General Terms

Measurement, Performance

Keywords

BGP, MPLS VPN, Routing Convergence

1. INTRODUCTION

Multi-protocol label switching (MPLS) virtual private networks (VPNs) [13] have had significant and growing commercial deployments. VPNs often carry business applications, such as VoIP, data replication, and financial transactions that do not react well to even the smallest disruptions in connectivity. Therefore, the timely convergence of routing protocols in the face of network events is critically important to the continued operation of these applications. Despite the importance of MPLS VPN networks, their routing behavior has not been studied by the research community, with the notable exception of [3] which formally analyzed the configuration conditions to ensure the correct VPN operation.

As is the case in the public Internet, Border Gateway Protocol (BGP) [11] plays a key role in MPLS VPNs. Different sites of the same VPN are connected via the provider network, for example as shown in Figure 1. Prefixes reachable in a particular VPN site are advertised to the provider network via external BGP (eBGP) sessions between routers in the VPN sites and the provider network. Multi-protocol extension to BGP [2] allows these VPN routes to be distributed via internal BGP (iBGP) throughout the provider network, and then via eBGP to other VPN sites that are part of the same VPN. The AS-level topology of VPN networks is therefore basically a hub-and-spoke topology (with some exceptions), in which the provider AS is the hub and the customer ASes are the spokes. This is in contrast to the public Internet BGP topology which have thousands of ASes in several tiers.

In the public Internet it is well-known that BGP suffers from slow convergence due to the exploration of invalid paths through different parts of the extensive AS topology. Because of the shallow AS topologies found in MPLS VPNs, much fewer AS paths are available to explore and iBGP convergence (in the provider network) therefore plays a relatively more important role in overall BGP convergence in MPLS VPNs. To the best of our knowledge, there has been no study on iBGP convergence in general, and iBGP convergence in MPLS VPNs in particular. In this paper we present an analysis of BGP convergence in MPLS VPNs using data obtained from a large Tier-1 ISP.

In our analysis we discovered an iBGP version of the *path exploration* phenomena. Because a router can receive multiple internal paths (from different route reflectors) which all go through the same egress link to reach a specific prefix, it can mistakenly choose a failed path during the convergence. Further, there is a *iBGP route invisibility* problem in which certain network configurations can also cause alternative (backup) paths to remain "hidden" from the network, until after the route for the primary path has been completely withdrawn, thus causing periods of disconnect during convergence.

Our study made use of several data sources (router configurations, forwarding table dumps, syslog messages and BGP updates) from the provider network. Combining these data sources allowed us to develop a new methodology to accurately determine the BGP convergence times caused by network events. Our measurement results show that most convergence delays we observed were relatively short (less

IMC'06, October 25–27, 2006, Rio de Janeiro, Brazil.

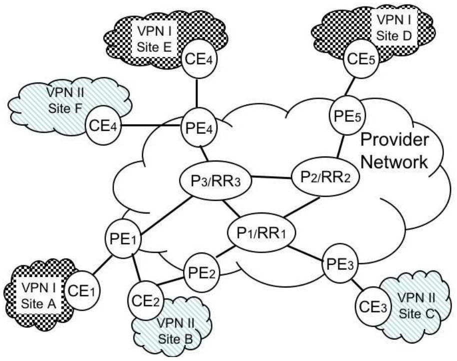

Figure 1: Components of an MPLS VPN

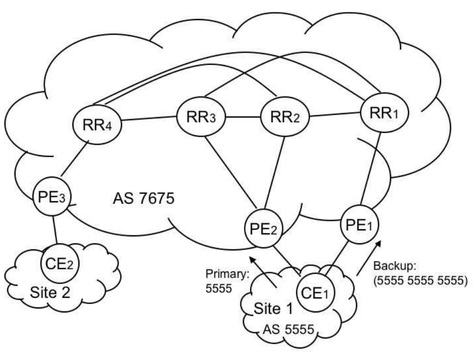

Figure 2: BGP Convergence an MPLS VPN

than 20 seconds). We also found that route invisibility occurs frequently and significantly contributes to the total convergence delay, while path exploration does not. This is different from the results in the public Internet, where the path exploration is the dominant factor [5, 9] even though invisibility of routes might occur.

Fortunately, most of the factors contributing to convergence delay can be either eliminated completely or significantly mitigated through a series of router configuration changes. Through measurement-based estimation, we show that our proposed changes can greatly reduce the convergence delay in MPLS VPN networks.

The rest of the paper is organized as follows. Section 2 provides some MPLS VPN background. Section 3 discusses the path exploration and route invisibility in MPLS VPNs using a representative example from our testbed experiment. Section 4 and Section 5 present our measurement methodology and results. Section 6 evaluates our proposed solutions, and Section 7 concludes the paper.

2. MPLS VPN BACKGROUND

We now present a brief overview of MPLS VPNs and specifically the role played by BGP. In Figure 1 we show the essential components that constitute an MPLS VPN. The figure shows two VPNs: sites A, D and E are part of VPN I and sites B, C and F are part of VPN II. A *customer-edge* (CE) router at each site connects to a *provider-edge* (PE) router in the provider network. The PE routers in turn connect to a set of *provider* (P) routers in the core of the provider network. The figure also shows a set of *route-reflectors* (RR) that distributes routes within the provider network. In order to simply the diagram we show route-reflectors to be co-resident with provider routers.

The provider network provides connectivity between the different VPN sites in such a way that separation of traffic between VPNs is maintained. A key construct to achieve this separation is the *virtual routing and forwarding* (VRF) table that is associated with each VPN on each PE. Each VRF table contains only routing information related to the VPN it is associated with. First, the VRF contains routes associated with the directly connected CE(s) which it typically obtains via an eBGP session between the PE and the CE. In much the same way as "regular" IPv4 BGP, these eBGP learned routes need to be distributed to other PEs via iBGP to allow remote PEs to reach the prefixes in question. This is achieved through the use of a VPN specific address family (VPNv4) in multiprotocol-BGP (MP-BGP) iBGP sessions within the provider network.

A salient feature of the VPNv4 address family is the eight byte *route-distinguisher* (RD) with which each VRF is configured. When routes associated with a VRF gets distributed via the MP-BGP session, this RD gets added to each IPv4 prefix to form a VPNv4 prefix (i.e., (RD:IPv4-Prefix)). Inside the provider network, the complete VPNv4 prefix is used for route comparison as part of the route selection process. This allows different VPNs to use the same address space (e.g., 10./8), without any conflict (assuming of course that they are configured with different RD values).

3. BGP CONVERGENCE IN MPLS VPNS

The key metric for BGP convergence is the time it takes for the network to converge on a stable set of routing tables after a routing change that was triggered by some network event. With reference to Figure 2, we focus on the case where the link between PE_2 and CE_1 fails, and we consider the BGP convergence from the point of view of PE_3 and the BGP messages it receives from RR_4. There are three factors that contribute to the BGP convergence time. First, it takes time for PE_2 to *detect* the failure, which can be 0 to 180 seconds (the BGP hold-timer default value). Second, BGP messages need to be distributed through the *iBGP* topology, which in practice normally consist of a route-reflector hierarchy. Third, there is an *eBGP* component between PEs and CEs. We focus on the first two factors in this paper, given the lack of eBGP update data between PEs and CEs.

We setup a testbed corresponding to Figure 2 in which RR_1 through RR_4 and PE_1 and PE_2 were Cisco routers, while CE_1 and PE_3 were software routers running the Quagga open source routing software. Table 1 shows the representative event times in the testbed experiment. It shows the update messages sent from RR_4 to PE_3 as well as the state of RR_4's routing table in the form of *iBGP signaling path*, constructed based on two route attributes (originator and the cluster-list). The originator attribute indicates the PE who originally injects the route into the AS. The cluster-list indicates the route reflectors traversed by the updates. For example, the route learned by PE_3 and RR_4 before the link failure has the iBGP signaling path of (RR_4,RR_3,PE_2), meaning that the path was first injected by PE_2 into the AS, then propagated to RR_3's cluster, and then propagated to RR_4's cluster. We now use Table 1 to illustrate the *path exploration* and *route invisibility* problem.

time	RR_4's updates	RR_4's table
before failure	(RR_4,RR_3,PE_2)	(RR_3, PE_2), (RR_2, PE_2)
T=0	failure happens	(RR_3, PE_2), (RR_2, PE_2)
T=0.7s	(RR_4,RR_2,PE_2)	(RR_2, PE_2)
T=4.7s	withdrawal	none
T=9.7s	(RR_4,RR_1,PE_1)	(RR_1, PE_1)

Table 1: iBGP signaling path of RR_4's routes.

3.1 Path Exploration in iBGP

At $T = 0$ second, we tear down the eBGP session between CE_1 and PE_2, and PE_2 loses the primary path. Because PE_2 does not have any other path in its routing table, it sends a withdrawal message to RR_2 and RR_3 to withdraw the primary path. One might expect that RR_4 will quickly send a withdrawal to PE_3. But at $T = 0.7$ second, RR_4 announces a path (RR_4,RR_2,PE_2) to PE_3. Note that at this point in time this path is in fact invalid since PE_2 does not have a route to the destination at $T = 0.7$ second.

What happens is the following. Due to background BGP load at RR_2, it processes the withdrawal from PE_2 much slower than RR_3 does. (In the testbed background BGP traffic was provided by an Agilent router emulator connected to RR_2.) Eventually RR_2 will process the withdrawal from PE_2 and send a withdrawal to the rest of the reflectors. However, its withdrawal arrives at RR_4 somewhat later than RR_3's. As a result, RR_4 receives and processes the withdrawal from RR_3 first and computes the new (invalid) best path, (RR_2, PE_2), which is not yet withdrawn by RR_2. This "new route" is thus sent to PE_3 at $T = 0.7$ second.

Note that in the testbed environment, the above update propagation ($PE_2 \rightarrow RR_3 \rightarrow RR_4 \rightarrow PE_3$) does not involve any MRAI delay because the updates are the first to be exchanged on the sessions involved. However, after that the MRAI timers are turned on for these sessions, including the one from RR_4 to PE_3. This means that no other update can be sent over this session until M seconds later (M is by default set to 4 to 5 seconds). Therefore, when RR_2 finishes processing the withdrawal from PE_2 and sends its own withdrawal to RR_4, RR_4 realizes that there is no path in its table, and it sends a withdrawal to PE_3 at $T = 4.7$ second. In this process, RR_4 in effect "explores" an internal path (RR_2,PE_2) that is already invalidated by the failure (CE_1/PE_2 session failure) that has triggered the convergence, even though the AS path remains the same (7675 5555). This process of exploring internal paths in iBGP is similar to the eBGP AS path exploration problem [5], but, to the best of our knowledge, path exploration has never been reported in the context of iBGP or MPLS VPN.

3.2 Route Invisibility Problem

At $T = 4.7$ second, RR_4 withdraws the primary path, but the backup path is not sent until $T = 9.7$ second. This additional delay is caused by the *route invisibility problem*: the backup path is "invisible" until the primary path is withdrawn. Before the failure, RR_1 prefers the primary path over the backup path, thus selects the primary path and sends it to PE_1. PE_1 compares the primary path and its own path, and selects the primary path as its best path because it is shorter than PE_1's own path (the backup path). Because a router can only announce its best path, PE_1 has to send a withdrawal to RR_1 to withdraw the backup path. That is, the backup path is "visible" only to PE_1, and is "invisible" to the rest of AS 7675. Note that the invisibility of the backup path could be by design (e.g., the VPN customer wants *all* the traffic to go through the primary path when the primary path is available), or it can be unintentional.

In order for PE_1 to announce the backup path to RR_1 and then to the rest of the network, PE_1 needs to first learn from RR_1 that the primary path is no longer valid. But, similar to RR_4, PE_1 experiences the path exploration process. RR_1 sends (RR_2,PE_2) to PE_1 at around $T = 1$ second, and then a withdrawal at around $T = 5$ second. PE_1 then selects the new best path, i.e., the backup path (5555 5555 5555). The route is propagated to RR_1, then RR_4 and PE_3. Because there has been no update on session $PE_1 \rightarrow RR_1$ and session $RR_1 \rightarrow RR_4$, the backup route is propagated over these two sessions with little delay. However, the MRAI timer on session $RR_4 \rightarrow PE_3$ is on due to the update at $T = 4.7$ second in Table 1. Therefore, the backup path cannot be propagated to PE_3 until $T = 9.7$ second.

The route invisibility problem impacts the convergence time in that routing updates needs to go through several iBGP hops to make the backup path available to the network. First, the withdrawal of the primary path needs to propagate through the reflector hierarchy to reach the PE which has the backup path. Second, the backup path is propagated through the reflector hierarchy to reach the rest of the PEs in the network.

Note that in the testbed scenario described here, there is no background BGP updates except on the session between the router emulator and RR_2. When there are background BGP updates on each session, as will be the case in any operational network, the per-neighbor MRAI timer is constantly "on" when a new update arrives at a router. Then each iBGP hop can cause a delay up to $M = 5$ seconds. In [10], we show that the worst case T_{long} and T_{down} convergence delay (failure detection and iBGP route propagation) are $(205 + n * 5)$ seconds and $(190 + n * 5)$ seconds, respectively, in the studied network.

4. MEASUREMENT METHODOLOGY

In this section, we present a methodology to accurately measure the BGP convergence time in a VPN network through correlating data from several sources (BGP, syslog, config, and PE forwarding table) that are readily available from operational networks. We will describe our data collection, event clustering, and event classification algorithms.

4.1 Correlating Data Sources

The provider network we studied collects both VPN BGP updates and syslog messages and has one level of route reflectors that form a full-mesh. One BGP collector is set up as the client of two route reflectors to collect the BGP updates from them. Among the many types of syslog messages, the layer 1, layer 2, and layer 3 messages are relevant to our study. The available information in these syslog messages, as well as that in BGP updates, are shown in Table 2. The BGP updates have prefix and RD information, but the syslog messages have only the interface/session information. Therefore, to correlate the syslog messages with the BGP messages, we use the route configurations and PE forwarding table (also shown in Table 2) to build the following two mappings: $(router, interface) \rightarrow$ *RD:prefixes* and $(router, neighbor\ ip, vrf) \rightarrow$ *RD:prefixes*.

Data Sources	types	available information
1. BGP updates	announcement	timestamp, prefix, rd, aspath, cluster-list, originator...
2. BGP updates	withdrawal	timestamp, prefix, rd
3. syslog messages	layer-1: LINK-UPDOWN	timestamp, router, per-router seqnum, interface, status
4. syslog messages	layer-2: LINEPROTOL-UPDOWN	timestamp, router, per-router seqnum, interface, status
5. syslog messages	layer-3: BGP-ADJCHANGE	timestamp, router, per-router seqnum, neighbor ip, vrf, status
6. router configurations	vrf configurations	router, vrf, rd
7. router configurations	eBGP session/interface configurations	router, interface, neighbor ip
8. PE Forwarding table	daily dump	router, prefix, vrf, nexthop ip, nexthop interface

Table 2: Available data from the provider network

4.2 Event Clustering

Because BGP path computation for different *RD:prefixes* are done separately, we conduct the measurement for different *RD:prefix* separately. Thus we convert each syslog message into m messages, one for each of the m affected prefixes. We then merge the converted syslog messages and the BGP update stream, and sort the combined stream based on timestamp. The clocks at the PEs and the BGP collector are NTP-synchronized, thus the timestamp indicates the relative timing of each message accurately enough for our purposes. We then cluster messages into events.

4.2.1 Existing BGP event clustering work

Earlier measurement work on the BGP convergence delay have focused on IPv4 BGP. These works can be classified into two types: beacon-based active measurements [5, 6, 8], in which controlled events were injected into the Internet from a small number of beacon sites (thus the event starting time is known), and time window-based passive measurement (in which the time-window value is derived based on the observed update inter-arrival time) [12, 7, 14, 4, 9].

4.2.2 Our clustering algorithm

Our clustering algorithm has two major differences from existing measurement work. First, we determine the event beginning time using syslog as the beacon because the syslog messages indicate the timing of the "root cause" of an event. Therefore, its accuracy is very close to that of the scheduled event beginning time in the active measurement. On the other hand, similar to the traditional passive measurement, it is more representative than the active measurement because it can be used to measure all the actual events triggered by the PE-CE link/session changes.

The second difference relates to how to decide the time window for determining the end of an event. Existing measurement work typically derives the time window based on the distribution of the measured update inter-arrival time. In this work, we calculate the value of the time window based on the various timer settings and the iBGP topology in the studied network. The analytical results in [10] shows that the iBGP route exchange can take at most $(5 + n) * 5 = 35$ seconds delay (where $n = 2$ is the number of reflectors that the primary PE is connected to) in the studied network. We thus use a time window of 35 seconds. Note that the time-window can be different with different router vendors.

Our algorithm classifies the events into three types [10]: *convergence, syslog-only, and update-only.* We focus on the *convergence* events in the rest of the paper. A convergence event begins with a syslog message and ends with BGP update message, and it finishes when the next message is: (i) a syslog message with different (router,interface), (ii) a syslog message with the same (router,interface) but with a direction different from the one in the starting syslog message of the event, (iii) an update message that arrives more than 35 seconds later than the last update message of the event.

4.3 BGP Convergence Event Classification

We borrow some terminology from earlier work on Internet BGP convergence [5], namely T_{down}, T_{up}, T_{long}, and T_{short}. Table 3 summarizes their definition and how we classify them in our measurement. For example, T_{long} is the event where the primary path fails, but the destination is still reachable via a less preferred path. We determine an event is T_{long} the syslog messages of the event indicates the link/session goes *down*, and the last message of the event is an *announcement*. The convergence delay for T_{long} is measured as the time it takes from the start of a syslog message to the last announcement in the event.

5. MEASUREMENT RESULTS

In this section we present the measurement results based on the data obtained from the provider network over a three month period in 2005 and answer the following questions about MPLS VPN convergence. How long are the delays of T_{long}, T_{down}, T_{up}, and T_{short} in MPLS VPN? Are they different from the eBGP convergence delays? What are different factors' contributions to the convergence delay? Is the reachability lost during T_{long} convergence, and for how long? Finally, it has been observed in IPv4 BGP that a small number of prefixes contribute most of the BGP events. Is this true in MPLS VPNs? More results are available in [10].

5.1 Results for All Types of Events

We observed around 300 thousands events during our study period, and 20% are *convergence* events. Among the convergence events, 7.3% are T_{long}, 7.4% are T_{short}, 43.4% are T_{down} and 39.9% are T_{up}, and 2% are *unclassified* (in which two "real" events overlaps). The fact that there are more T_{down} and T_{up} events than T_{long} and T_{short} are because only multi-homed prefixes can possibly have T_{long} and T_{short} events, and typically, there are much fewer multi-homed prefixes than single-homed prefixes. The fact that there are more T_{down} events than T_{up} events can be due to the fact that there are more "real" T_{up} events that are classified as syslog-only, update-only, or unclassified convergence events by our methodology as the result of event overlapping or syslog message loss (less than 1% in our study, detected by checking the per-router syslog sequence numbers).

Figure 3 shows the distributions of the convergence delays for all types of events. Most delays (83% in T_{long} and T_{down}, 68% in T_{up} and T_{short}) are shorter than 20 seconds. There are two differences in our results compared with the

event type	description	link status	last update of the even	type of previous event
T_{down}	prefix unreachable after link failure	down	withdrawal	does not matter
T_{up}	prefix reachable again after link recovery	up	announcement	T_{down}
T_{long}	prefix reachable via backup after primary path fails	down	announcement	does not matter
T_{short}	prefix reachable via primary after primary path recovery	up	announcement	not T_{down}

Table 3: Convergence event types

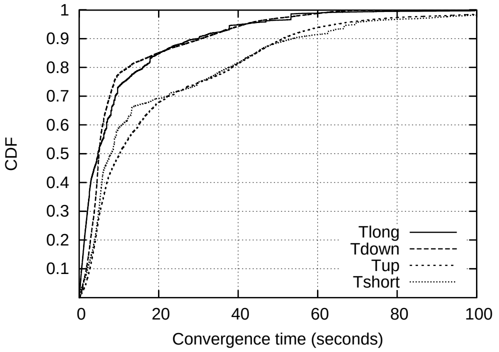

Figure 3: Cumulative distributions of convergence delays for T_{down} and T_{long} (two upper curves), T_{up} and T_{short} (two lower curves).

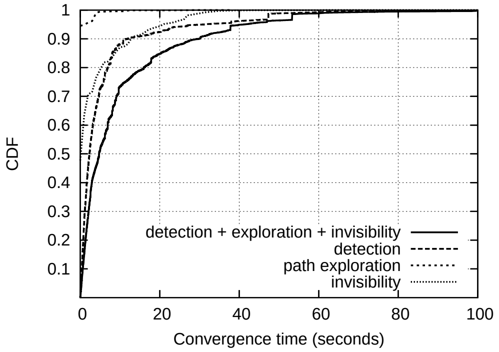

Figure 4: Cumulative distribution of delays caused by different factors in T_{long} convergence.

IPv4 BGP convergence delays [5, 8]. First, the convergence delays in MPLS VPNs are much shorter than those in IPv4 BGP, in which convergence delays on average are longer than 100 seconds for T_{down} and T_{long} are more than 30 seconds in T_{up} and T_{short} [5, 8]. This is due to several reasons. First of all, IPv4 BGP has a much bigger topology scale (thousands of ASes) than MPLS VPN networks. Thus the IPv4 BGP update propagation time along the ASes, amplified by MRAI delay, is longer than in MPLS VPN networks. Also because the simpler route reflector-based topology limits the number of alternative paths, there are much less path exploration in MPLS VPNs. Third, the default MRAI time value for iBGP is 5 seconds, while in eBGP it is 30 seconds.

Second, Figure 3 shows that the delays of T_{up} and T_{short} are longer than those of T_{down} and T_{long}. This observation is different from the eBGP convergence results in [5, 8], but they are not conflicting. In the measurements reported in [5, 8], an event starts when the initial announcement or withdrawal are advertised, but in our methodology, an event starts when a layer 1/layer2 change happens. In T_{up} and T_{short}, it takes time for the PE and CE to exchange BGP protocol messages, following the Finite State Machine specification [11], to establish a BGP session. After that, the routes are exchanged. Depending on the size and the order of the routes in the routing table, the announcement of the prefix in question can be further delayed. In practice, it is helpful to have a shorter T_{up} delay in some cases such as a simple router reboot during planned maintenance. Therefore, this suggests that further research and engineering are needed to shorten the time to establish the BGP sessions and initial table exchanges after a session is setup.

5.2 Breaking down T_{long} convergence

We now break down the T_{long} convergence factors. We measure the session failure detection time as the time from the beginning of the event to the time the first BGP update is received. To measure the *path exploration* time, we first found the time (say, T_3) for the first update with an aspath equal to the primary aspath, and the originator/BGP nexthop equal to the primary PE. Then we find the time for the withdrawal message (say T_4). Then the path exploration time is defined as $T_4 - T_3$. The *route invisibility*'s contribution is $T_5 - T_4$, where T_5 is the end of the convergence.

Figure 4 shows distributions of the delay contribution of the different factors contributing to T_{long}. It shows that failure detection contributes the most, closely followed by route invisibility (both contribute around 10 seconds in 90% of events). On the other hand, path exploration contributes least in MPLS VPNs, as opposed to the Internet environment where it is the dominant factor. As we mentioned earlier, in MPLS VPNs there are much fewer paths to "explore" due to its smaller and simpler route reflector-based topology. In addition, we found that that 30% of the events have route failure (during which either there is no route or the failed primary path is used) for longer than 9 seconds.

These results show that it is very important to improve the T_{long} convergence delay to minimize significant disruptions to important applications such as VoIP, and that in order to improve T_{long} convergence, shortening failure detection time and solving route invisibility are more important than solving path exploration.

5.3 Event Contribution by Network Entity

In IPv4 BGP, it has been observed that the majority of the events are caused by a small number of very unstable prefixes [12]. In this section, we investigate whether similar observations hold for VPNv4 prefixes, as well as for other MPLS VPN specific "network entities" namely, VPNs, PEs, and PE-CE interfaces. We first ranked (from high to low) the entities based on the number of events per entity. Then we determined the cumulative ranked contribution of each of the four network entities to the total convergence event count. We found significant "popularity" in the network entities that are involved with convergence events. More specifically, only 18.6% of prefixes contributed to 90% of the

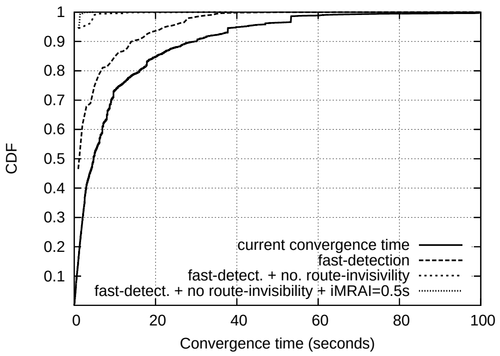

Figure 5: Estimation of T_{long} Improvement

convergence events. The number is approximately the same (17%) for RDs, while the PE-CE interface granularity has an even more skewed distribution with only 6.6% of interfaces contributing to 90% of the events. Considering the router level granularity, we see a somewhat less skewed distribution where 42.9% of PEs contribute to 90% of events.

6. SOLUTION AND EVALUATION

In this section we summarize and evaluate our proposed solutions [10] to the delayed convergence in MPLS VPN. To shorten the failure detection time, "Next-Hop Tracking (NHT)" [1] or similar features should be enabled on PE routers. The route invisibility problem can be easily eliminated by configuration changes to force the backup route to be distributed to the remote PEs. That is, we can configure PEs such that they *always* prefer their locally learned eBGP route (over iBGP routes), and configure the VRFs from different sites of a VPN with different RDs.

In Figure 5, we estimate the improvement of the our solutions based on our T_{long} measurement results. Supposing we have a fast detection mechanism (e.g., like Next-hop tracking) in place we can expect to see detection times around 1 second. We can thus estimate the potential improvement by replacing the measured failure detection time with this value. The "fast-detection" curve in Figure 5 shows significant improvements over the current convergence delay.

However, the fast-detection curve also shows that 20% of the delays are longer than 6 seconds, and 10% are longer than 15 seconds. This shows the necessity of eliminating route invisibility. The curve "fast-detection + no route-invisibility" estimates the improvement when route invisibility is eliminated in addition to the fast failure detection. It shows that the convergence time is significantly reduced (what is left is the path exploration contribution plus the assumed 1 second fast detection delay). The final curve in Figure 5 shows the improvement of using a shorter iBGP MRAI value (0.5 second) by dividing the measured path exploration time by 10(= 5/0.5). The figure shows that the additional improvement by using a smaller MRAI value is insignificant, and thus may not warrant the additional load incurred by such an approach.

Based on these results, we conclude that the combination of fast failure detection and route invisibility elimination are necessary and sufficient configuration changes to achieve a short convergence delay in MPLS VPN networks.

7. CONCLUSION

In this paper we have presented the first systematic study of BGP convergence in MPLS VPNs. We used several data sources from a Tier-1 ISP and developed a methodology that allows accurate estimation of BGP convergence delays. We identified contributing factors such as path exploration in iBGP and the information hiding in the form of the route invisibility problem. Our analysis show that, among the contributing factors, failure detection and route invisibility have the most significant impact on convergence delays. Fortunately, suitable solutions exist to eliminate or mitigate these problems. Our measurement-based estimation shows that applying our proposed configuration-only changes to MPLS VPNs can significantly reduce the convergence delay.

Our work has also uncovered a number of results. For example, the fact that a relative small percentage of entities (prefixes, RDs, PEs, interfaces) are responsible for the bulk of events might indicate some underlying problems. Given the importance of MPLS VPNs to the commercial world, these results warrant further investigation.

8. ACKNOWLEDGMENT

We thank Chris Chase, Alex Gerber, Joel Gottlieb, Carsten Lund, Aman Shaikh, and Jen Yates for their help with the data we used in this study. We also thank Chris Chase, Kevin D'souza, Aman Shaikh, Jim Uttaro, and IMC committee for their helpful comments on earlier versions of this paper.

9. REFERENCES

[1] Bgp next-hop tracking. Technical report. http://www.cisco.com/univercd/cc/td/doc/product/software/ios124/124cg/hirp-c/ch05/h-bnht.pdf.

[2] T. Bates, Y. Rekhter, R. Chandra, and D. Katz. Multiprotocol extensions for bgp-4. ITEF RFC 2858, June 2000.

[3] R. Bush and T. Griffin. Integrity for virtual private routed networks. In *Proceedings of the IEEE INFOCOM*, April 2003.

[4] A. Feldmann, O. Maennel, Z. M. Mao, A. Berger, and B. Maggs. Locating internet routing instabilities. In *Proceedings of ACM Sigcomm*, August 2004.

[5] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. Delayed Internet Routing Convergence. In *Proceedings of ACM Sigcomm*, August 2000.

[6] C. Labovitz, R. Wattenhofer, S. Venkatachary, and A. Ahuja. The Impact of Internet Policy and Topology on Delayed Routing Convergence. In *Proceedings of the IEEE INFOCOM*, April 2001.

[7] O. Maennel and A. Feldmann. Realistic BGP traffic for test labs. In *Proc. of ACM SIGCOMM*, 2002.

[8] Z. Mao, R. Bush, T. Griffin, and M. Roughan. BGP Beacon. In *Proceedings of ACM IMC 2003*, October 2003.

[9] R. Oliveira, B. Zhang, D. Pei, R. Izhak-Ratzin, and L. Zhang. Quantifing path exploration in the internet. In *Proceedings of ACM IMC 2006*, October 2006.

[10] D. Pei and J. Van der Merwe. Bgp convergence in virtual private networks. Techincal Report TD-6QCNCP, AT&T Labs–Research, June 2006.

[11] Y. Rekhter and T. Li. Border Gateway Protocol 4. RFC 4271, SRI Network Information Center, Jan 2006.

[12] J. Rexford, J. Wang, Z. Xiao, and Y. Zhang. BGPRouting Stability of Popular Destinations. In *Proceedings of ACM IMW 2002*, October 2002.

[13] E. Rosen and Y. Rekhter. Bgp/mpls ip virtual private networks (vpns). IETF RFC 4364, February 2006.

[14] J. Wu, Z. Mao, J. Rexford, and J. Wang. Finding a needle in a haystack: Pinpointing significant BGP routing changes in an IP network. In *NSDI 2005*, May 2005.