GROW Working Group
Internet-Draft
Intended status: Informational
Expires: January 5, 2010

H. Ballani
Cornell U.
P. Francis
MPI-SWS
D. Jen
UCLA
X. Xu
Huawei
L. Zhang
UCLA

July 4, 2009

Performance of Virtual Aggregation
draft-ietf-grow-va-perf-00.txt

Status of this Memo

This Internet-Draft will expire on January 5, 2010.

Copyright Notice

Abstract

The document "FIB Suppression with Virtual Aggregation" [I-D.francis-intra-va] describes how router FIB size may be reduced. This approach entails a trade-off between path-length and load versus FIB size. It also has the potential to reduce convergence time. This document describes the results of several studies that examine these characteristics. The results of a study for a Tier-1 ISP with a relatively sophisticated deployment of VA, shows that FIB size could be reduced ten times or more with a worst-case latency penalty of 4ms and a worst-case load increase of <1.5%. Another study, examining a much simpler style of VA deployment, also for a Tier-1 ISP, shows that FIB size can be reduced by four times (in routers serving as APRs), and more than 10 times in other routers. Here, worst-case latency increase was 16 ms, though this is almost certainly an over-estimate, both because traceroute was used to make the measurement, and because popular prefixes were not considered.

Table of Contents

1. Introduction 4
  1.1. Terminology 5
  1.2. Temporary Sections 5
    1.2.1. Document revisions 5
2. The NSDI 2009 Study on FIB size versus load and stretch . . . 5
3. The IETF74 Study 8
  3.1. Evaluation Setup 8
  3.2. Evaluation Results 10
4. Convergence Time 11
  4.1. Topology Convergence Time 11
  4.2. Boot Convergence Time 13
5. Acknowledgements 14
6. References 14
  6.1. Normative References 14
  6.2. Informative References 14
Authors' Addresses 15

1. Introduction

The document "FIB Suppression with Virtual Aggregation" [I-D.francis-intra-va] describes how router FIB size may be reduced. This approach entails a trade-off between path-length and load versus FIB size. This document describes the results of two independent studies that examine these tradeoffs.

One of the studies is [nsdi09], published in NSDI 2009. In this study, the router topology and traffic matrix of a Tier-1 ISP were used to model the expected performance of relatively sophisticated VA deployments on that Tier-1 network. The primary results of this study are that FIB size could be reduced to 10% of DFZ FIB size or better, with a latency increase of no more than 4ms and a maximum load increase of <1.5%. Further, the load increase was relatively uniform across the ISP's routers. With these savings, it is estimated that the lifetime of routers with FIB space for only 1/4 million IPv4 routes could easily be extended five to ten years.

The other study [ietf74] evaluates a more straight-forward deployment style for VA. This evaluation results were presented at the 74th IETF and are summarized in this document. In a nutshell, the IETF evaluation measures the benefits that an ISP might receive under a relatively simpler deployment of VA. While the NSDI evaluation makes assumptions on packet delivery speed and topological deployments that help optimize the benefits of VA, the IETF evaluation looks at a very intuitive and easy-to-manage deployment of VA, one that may be more realistic for an ISP early on. We find that even with a very simple, intuitive, low-maintenance deployment of VA, the ISP we studied would still be able to reduce FIB sizes by 75% or more on all of its routers, at a cost of no more than 16ms additional delivery latency in the worst case. With the use of popular prefixes, this worst-case latency increase could almost certainly be reduced significantly, though this was not studied. In particular, routers in these POPs could have FIB-installed those prefixes that are reachable via the POP, thus avoiding the round-trip to another POP and back.

Both studies have their limitations. Neither study actually deployed VA per se---rather they modeled the effect that VA would have on an ISP given certain data from that ISP. The NSDI09 study estimated latency by the speed of light distance between routers, and so is almost certainly under-estimating latency increase by a small amount. The IETF74 study, on the other hand, used trace-routes to estimate latency, and so is probably over-estimating latency increase. Overall, the NSDI09 reflects the best that VA could do, but might only be achievable after considerable deployment experience. The IETF74 study, on the other hand, better reflects what an ISP might see in its initial, simple deployment.

1.1. Terminology

This document adopts the terminology from [I-D.francis-intra-va]. The following terminology is additionally defined:

Latency Increase: In VA, routers don't maintain the entire DFZ FIB and hence, traffic must be routed through APRs. This, in turn, increases the length of the path traversed by packets. "Latency increase" is defined as the increase in latency for packets traversing an ISP due to the adoption of VA as compared to the status quo. Note that traversal latency can have different meanings: the NSDI09 study uses propagation latency while the IETF74 uses delay measured by traceroute.

Stretch: Same as latency increase.

Worst-case Latency Increase: The maximum increase in latency for packets destined to any routable prefix and ingressing at any ISP router.

Load Increase: VA requires APRs to forward traffic which otherwise need not be routed through them. The increase in the amount of traffic forwarded as a fraction of the original traffic forwarded by a router is termed as load increase.

Worst-case Load Increase: The maximum increase in load across all the ISP's routers.

1.2. Temporary Sections

This section contains temporary information, and will be removed in the final version.

1.2.1. Document revisions

This is the first document revision.

2. The NSDI 2009 Study on FIB size versus load and stretch

The NSDI09 paper describes and analyzes a "config-only" approach to deploying VA. Specifically, it shows how VA can be deployed with today's legacy routers without software or protocol changes. It has two types of analysis. The first studies the trade-off between FIB size and stretch and load. That analysis is presented in this section. The second looks, somewhat superficially, at router convergence time (specifically, the time it takes for a booting router to fully initialize BGP). That analysis is presented in Section 4.

The NSDI09 paper describes two approaches to VA, one where control-plane Route Reflectors (RR) are used to filter out prefixes that need

to be suppressed, and another where each data-plane router does the filtering at the boundary between the RIB and the routing table. The latter deployment is closer to the intent of the VA internet drafts, but for the purpose of FIB size/load-latency performance evaluation, either approach applies equally.

The core performance issue is to understand the trade-off between FIB size on one hand, and latency and load penalty on the other. To understand this trade-off, the NSDI09 study closely modeled how VA would perform on a Tier-1 ISP. The model was based on knowledge of the Tier-1 ISP's router-level topology including router geographic location, routing tables, and traffic matrix. This information was obtained directly from the Tier-1 ISP. Some additional information required for the study that was not directly available was inferred. This includes the latency between routers, which is approximated as the speed-of-light time for geographic distances between routers. It also includes the IGP link-weights, which were also approximated using (the inverse of) the router distance. Finally, queuing delay was not considered, though given that load increase is quite low, the queuing delay should not increase significantly. Switching time at routers was also not considered, though again this is a relatively minor component of delay. Overall, however, this study slightly under-estimates latency increase.

In the study, VA was deployed in such a way that all routers can be APRs. In other words, the different router roles (edge, core, aggregation between edge and core) were not differentiated. As a result, ALL routers in the ISP see significant FIB savings.

When deploying VA, two important questions to answer are, how are VPs structured, and how are APRs selected (i.e., what is the assignment of VP to APR)? With regards to the first question, the distribution of prefixes across the VPs directly impacts the FIB size reduction that can be achieved. If the VPs are structured such that they all have similar number of prefixes, the ISP can ensure that all its routers see substantial savings. As a contrast, if some VPs have a lot of prefixes, one (or, a few) of the ISP's routers would need to maintain them which, in turn, limits the reduction in worst-case FIB size.

The NSDI09 study considered two approaches. In one, every VP is a \7. This leads to 128 VPs: 0/7, 2/7, ..., 254/7, with the largest VP containing 22,722 prefixes or 8.9% of today's routing table. In the second, the set of VPs was selected such that the number of prefixes in each VP is relatively uniform and each VP is larger than any real prefix. This latter approach yielded 1024 VPs with the largest containing 4,551 prefixes or 1.78% of the BGP routing table. In the rest of this section, we summarize results using the latter approach.

Results using the /7 VPs can be found in the NSDI paper. The next section also present results assuming /8 VPs and other simplifying assumptions.

With regards to the second question, APRs were assigned using an automated greedy algorithm that ensures that the maximum latency increase for traffic to any prefix from any ISP router is below a specified constraint while trying to minimize the FIB size across the

ISP's routers.

The analysis was split into two parts, one that looks only at FIB size versus worst case latency increase, and another that adds analysis of load. This is appropriate, because load is decreased simply by adding popular prefixes, which increases FIB size without changing worst case latency. Note that the use of popular prefixes is necessary to achieve acceptable performance in this study, so the analysis that ignores load should not be interpreted as the numbers one would see in practice.

A VA deployment such that there is an APR for every VP in each of the ISP's POPs is able to reduce the worst-case FIB size to 25% of the DFZ FIB size. In this case, all traffic redirection is within a POP and hence, the stretch imposed on traffic is virtually 0. The ISP can also choose to incur traffic stretch to further reduce router FIB requirements. The resulting FIB versus latency analysis produced a range of results but there is a performance sweet-spot at the point where worst-case stretch is capped at 4ms. In this case, the worst-case FIB size is 5% of the DMZ FIB size (in other words, 20 times reduction). Capping the latency at higher values did not significantly shrink FIB size.

Besides looking at percentage of FIB reduction, the paper analyzed how long the lifetime of current routers could extend into the future before running out of FIB space. Using DMZ growth predictions based on growth history[atnac06], Cat6500 routers that can hold 249K IPv4 FIB entries (and therefore could not hold full DFZ as of 2007), would last a decade with virtually no increased latency, and over two decades at 4ms increased latency.

The above analysis, however, ignores load. In VA, load is reduced by installing so-called "popular prefixes" into the FIB. These are the prefixes that have the heaviest traffic volumes. Without installing popular prefixes, load increases by about 40%, which is clearly unacceptable.

In the Tier-1 ISP studied, 1.5% of most popular prefixes carry 75.5% of the traffic, while 5% of the prefixes carry 90.2% of the traffic (as of late-2007). This "power-law" distribution has been the norm over many studies spanning many years. Because of this distribution, installing a relatively small number of popular prefixes improves load tremendously. Note too that prefixes that are popular at one time tend to stay popular over time. The 5% of popular prefixes that produces 90% of the traffic on one day will still produce roughly 85% of the traffic one month later. This means that an ISP could measure its popular prefixes and reconfigure its routers relatively infrequently---once per week or even less frequently.

In the analysis, when 5% of most popular prefixes are installed, the worst-case load reduces to 1.38%. These 5% popular prefixes add to the FIB size and so with a 4ms cap on the worst case latency increase, the actual FIB size is 10% of the total. In other words, overall this Tier-1 ISP could reduce FIB size by ten times while increasing worst case latency by <4ms, and worst case load by <1.5%.

Besides studying the Tier-1 ISP in detail, the NSDI09 paper also used Rocketfuel to make rough estimates of FIB size and latency for 9 other ISPs (load could not be estimated because Rocketfuel does not have the traffic matrix of these ISPs). Because Rocketfuel tends to underestimate the number of routers in an ISP, the analysis is conservative (more routers means more aggregate FIB over which to spread the routing table). In this analysis, assuming worst case latency increase of 5ms, the FIB could be reduced to 5-15% of DFZ.

3. The IETF74 Study

While the NSDI09 results might suggest that Virtual Aggregation offers a great tradeoff for ISPs, the relatively management complexity that would result from the style of deployment used in that study suggests that ISPs would not see that level of performance in initial deployments. The IETF74 study described here uses a much simpler style of VA deployment, something that better reflects what an ISP would use early on. In particular, the IETF74 study considers ease of management when assigning APRs and virtual prefixes to use. As such, it better reflects the stretch/savings tradeoff they would actually experience if they deployed VA in their networks today.

3.1. Evaluation Setup

The results of a VA evaluation will depend heavily on which VA deployment is evaluated. VA deployments can differ in the amount of VPs used, which VPs are used, how often they change, the amount of APRs used, the VP-APR mappings, and the placement of APRs. These variables must be given values to define a particular flavor of VA that's being evaluated. The NSDI study selected these variable values in order to maximize FIB savings and minimize additional packet latency. This allows us to evaluate just how good the VA stretch/savings tradeoff could potentially be, but the variable values probably described a VA deployment that would not realistically occur. For this study, selections of these variable values will be based on what we consider most simple and intuitive. In this subsection, we present the values we selected for these variables, as well as justifications for our selections.

We start by describing where we decided to place the APRs for our study. Again, our aim is for simplicity and ease of management. The topology of the tier-1 ISP we used for this evaluation revealed that the ISP consists of a many small PoPs with a few routers, and a few large PoPs consisting of many routers. While exact numbers cannot be revealed due to confidentiality agreements, the discrepancy between the number of routers in small and large PoPs is significant. The discrepancy between the number of large and small PoPs was also significant. Based on these observations, we decided that it would be intuitive to have an APR for every VP at each large PoP. Furthermore, only large PoPs should contain APRs, which makes them easy to locate. When forwarding packets, nodes in large PoPs should obviously use their local APRs, while routers in small PoPs should use the APRs from their nearest major PoP. Unlike the VA deployment used for the NSDI evaluation, APR placement does not change with the routing table. This significantly reduces troubleshooting efforts compared to the architecture proposed by the NSDI evaluation. We also had to decide how many APRs each major PoP should have. This question is independent of the number of VPs an ISP decides to use. For example, an ISP could choose to use 8 different VPs. However, the ISP could have 1 APR be responsible for all 8 VPs, or assign 2 VPs to each of 4 different APRs, or assign 5 VPs to one APR and the remaining 3 to another APR, etc. We decided that each large PoP would evenly distribute the FIB storage requirements amongst 8 different APRs. We chose 8 APRs in particular simply because each large PoP in our studied tier-1 ISP had at least 8 routers storing complete global routing tables, and these machines would have the storage capacities to be APRs.

The next decision involved which virtual prefixes to use. The more virtual prefixes you have, the finer the granularity for assigning virtual prefixes to APRs. This increases an ISP's flexibility to divide FIB storage evenly amongst its different APRs, and thus keep worst-case FIB size at a minimum. Of course, in order for VA to work properly, VPs must be longer than any real prefixes. For these reasons, the NSDI study maximized the number of VPs used by looking at the global routing table on a certain date, and carefully adding virtual prefixes that were never longer than any of the real prefixes it covered, which yielded 1024 prefixes. However, note that as the global routing table changes, the VPs used may have to change as

well. A static VP list would be much easier to manage. Therefore, we decided to use VPs of length 8, which allows us to cover the v4 space using 256 VPs. While this number is smaller than the 1024 used in the NSDI study, this allows us to maintain a static VP list where each VP is still never longer than any of the real prefixes it covers.

We have now decided on a VP list and a placement of APRs we decided would be simple and easy to manage. Our final task to complete the description of our VA deployment was to decide which VPs to assign to which APRs. In accordance to the concept of manageability, we used a simple and straightforward greedy algorithm for assigning virtual prefixes to APRs. The aim was for this algorithm to be computationally cheap and quickly run at regular intervals to still evenly distribute FIB storage responsibilities amongst the APRs. The algorithm basically assigns VPs to an APR one by one until the APR has VPs covering at least 1/8 of the DFZ. We then starts assigning the remaining VPs to another APR. The algorithm is as follows:

```
Select one of the eight APRs
For VPs 0/8 thru 255/8:

   Assign VP to selected APR
   If number of entries in APR > 1/8 of GRT: Select a previously
   unselected APR
```

We have now assigned values to all of the variables required to describe a particular deployment of VA. We can now evaluate the costs and benefits of this VA deployment for the ISP.

3.2. Evaluation Results

FIB savings for VA routers are calculated by counting the number of FIB entries in the VA router and comparing this to the size of the global routing table.

To distribute FIB storage requirements evenly amongst the 8 different APRs, we ran the algorithm described above on global routing table contents for the first days of the months between 06/08 and 02/09.

The algorithm, while simple, works quite well. In the worst case, an APR was assigned to cover 14% of the global routing table, as opposed to the ideal amount of 12.5%(1/8th).

Under VA, routers have to also store peer-to-label mapping for each external router that peers with the ISP. An operator for the ISP we studied estimated the number of external peerings at around 20,000.

It turns out that even with this simple deployment of VA, the largest

storage requirements for an APR still reduces its FIB storage requirements by 75%. For non-APR routers, FIB storage requirements were reduced by 93%.

For each PoP, we calculated the worst-case stretch that a packet ingressing from the PoP could experience. Stretch is defined as the additional time the packet takes to exit the ISP due to suboptimal paths introduced by the VA architecture. Worst-case stretch occurs when the nearest APR for the packet is in the opposite direction of the egress router out of the ISP, and thus the packet is stretched a round-trip distance from the ingress PoP to the APR and back.

When calculating the worst-case stretch for each PoP, we did not simply assume speed of light and shortest distance between PoPs as the NSDI evaluation did. We wanted to capture the actual delay that a user would experience for the packet, which would include processing time as well as propagation time. Thus we used traceroute to capture the true stretch experienced by packets due to VA. To determine the worst case stretch for a given PoP, we tracerouted from this PoP to all of the major PoPs. This allowed us to map the PoP to its 'nearest' major PoP. We then used traceroute to determine the time it takes to go to the nearest major PoP and back, thus giving us the worst-case stretch for that PoP. We did this for every PoP, and found that all PoPs can send a packet to a large PoP in 8ms or less, and thus the worst-case stretch for any PoP in our studied ISP is 16ms. Furthermore, 70% of PoPs experience a worst-case stretch of 8ms or less, and over 30% of all PoPs experience no stretch at all. This is because they were either a major PoP, or they naturally defaulted to a major PoP for all of their traffic anyway, so VA did not change their natural delivery path whatsoever.

4. Convergence Time

Regarding convergence time, in general we expect convergence with VA to be faster than convergence without VA. The basic argument is simple: updating FIB entries takes time. If there are fewer FIB entries to update, then convergence goes faster. There are two forms of convergence discussed here: convergence time for a given router when a link or some other router goes down or comes up, and time to fully initialize BGP when a router boots up. We call the first topology-convergence, and the second boot-convergence. Each is discussed in turn.

4.1. Topology Convergence Time

A simple experiment was run to demonstrate the improved topology-convergence aspect of VA. Before discussing the experiment, it is important to note that the benefit demonstrated by this experiment could also be had using the "Prefix Independent Convergence" mechanism. In other words, VA isn't the only way, or even the simplest way, to achieve this kind of fast convergence.

The following topology was used in the experiment.

```
                    _.--------.
               ,--''           `---.             ,-------.
             ,'                     `.        ,-'   AS2   `-.
           ,' +----+         +----+   `.     /     +----+    \
          /   | R1 |.........| R3 |.....\...(......| R2 |     )
         /    +----+         +----+      \   \     +----+    /
        ;         \            /          :   `-.         ,-'
        |          \          /           |      `-------'
        :           \        /            ;
         \           +----+ /       AS1  /
          \          | R4 |             /
           `.        +----+           ,'
             `.                     ,'
               `---.           _.--'
                    `--------''
```

Router R2 advertised a number of routes (ranging from 5000 to 220K) to R3, which in turn distributed them in AS1 with full-mesh iBGP. Packets were transmitted to R1 (via a test-box not shown) for destinations within the advertised routes. At time T0, the link R1--R3 is taken down. This results in a period of time during which some fraction of transmitted packets are dropped at R1. The elapsed time until all transmitted packets are successfully received at R2 is then measured.

The following table shows the elapsed time for a varying number of advertised routes, for both with VA and without VA.

Number of Routes	With VA	Without VA
5000	2 sec	2 sec
50000	2 sec	10 sec
100000	2 sec	17 sec
150000	3 sec	24 sec
200000	3 sec	30 sec
220000	3 sec	35 sec

As can be seen from the table, convergence time for the VA scenario is pretty much independent of routing table size. This is because the only change in the FIB is that of the best igp next-hop to the APR. In the non-VA case, however, the igp next-hop needs to be changed for all prefixes before routes to all prefixes converge.

It should be noted that this experiment intentionally puts VA topology-convergence in the best light. Router R1 only has a single FIB entry that it needs to update: that of the VP. In practice, any given router may have both VP sub-prefixes (those needed by virtue of being an APR) and popular prefixes in the FIB. Certainly at least the new next-hops to the VP sub-prefixes and VPs would have to be FIB-installed before convergence. Realistically, however, the popular prefixes should be installed too, since until they are load is increased. Therefore, in practice the improvement in convergence time will be less than shown here.

4.2. Boot Convergence Time

The NSDI09 paper [nsdi09] experimented with the time it takes for a router to boot and fully populate its routing tables and advertise all updates (i.e. initialize BGP). It is important to note that the NSDI09 paper did not implement the VA draft [I-D.francis-intra-va] per se. Rather it deployed VA using legacy routers configured to implement VA. In fact, the NSDI09 paper experimented with two distinct configurations of VA. In one, control-plane Route Reflectors (RR) were used to filter out the prefixes not needed by data-plane routers (i.e. prefixes that could be FIB-suppressed according to the rules of VA). In this setting, both the RIBs and FIBs of data-plane routers were shrunk.

The second configuration approach better reflects the spirit of the VA draft. Here, data-plane routers ran BGP as normal (i.e. the RIBs were populated more-or-less as they would be in the absence of VA), and did FIB-suppression by setting the administrative distance for suppressible prefixes to 255. In the Cisco routers used in the experiment, such prefixes were then FIB-suppressed, but otherwise treated normally.

In the test setup, the VA routers' FIBs held roughly 1/2 of the full DFZ. VA as deployed with control-plane RRs was able to initialize BGP about twice as fast as regular (full DFZ) routers (for example, 124 seconds versus 273 seconds). With the "admin-distance = 255" approach, it took roughly twice as long for the VA routers to initialize compared to the regular routers (for example, 487 seconds versus 273 seconds). The reason for this is that these Cisco routers were not designed to deal with large admin-distance access lists efficiently. We believe that this inefficiency can be overcome, but in fact there are at this time no hard numbers to back up this supposition.

5. Acknowledgements

Tuan Cao, a PhD student at Cornell, did much of the heavy lifting on the NSDI09 measurement study.

6. References

6.1. Normative References

[I-D.francis-intra-va]
Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, "FIB Suppression with Virtual Aggregation", draft-francis-intra-va-01 (work in progress), April 2009.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

6.2. Informative References

[atnac06] Huston, G. and G. Armitage, "Projection Future IPv4 Router Requirements from Trends in Dynamic BGP Behaviour", ATNAC 2006 caia.swin.edu.au/pubs/ATNAC06/Huston_1m.pdf, December 2006.

[ietf74] Jen, D., "Scaling FIBs with Virtual Aggregation: How Much Stretch? How Much FIB savings?", IRTF Routing Research

Group Meeting http://www.ietf.org/proceedings/09mar/slides/RRG-2.pdf, March 2009.

[nsdi09] Ballani, H., Francis, P., Cao, T., and J. Wang, "Making

Routers Last Longer with ViAggre", ACM Usenix NSDI 2009 http://www.usenix.org/events/nsdi09/tech/full_papers/ballani/ballani.pdf, April 2009.

Authors' Addresses

Hitesh Ballani
Cornell University
4130 Upson Hall
Ithaca, NY 14853
US

Phone: +1 607 279 6780
Email: hitesh@cs.cornell.edu

Paul Francis
Max Planck Institute for Software Systems
Gottlieb-Daimler-Strasse
Kaiserslautern 67633
Germany

Phone: +49 631 930 39600
Email: francis@mpi-sws.org

Dan Jen
UCLA
4805 Boelter Hall
Los Angeles, CA 90095
US

Phone:
Email: jenster@cs.ucla.edu

Xiaohu Xu
Huawei Technologies
No.3 Xinxi Rd., Shang-Di Information Industry Base, Hai-Dian District
Beijing, Beijing 100085
P.R.China

Phone: +86 10 82836073
Email: xuxh@huawei.com

Lixia Zhang
UCLA
3713 Boelter Hall
Los Angeles, CA 90095
US

Phone:
Email: lixia@cs.ucla.edu