Network Working Group	P. Francis	
Internet-Draft	MPI-SWS	
Intended status: Informational	X. Xu	
Expires: April 22, 2010	Huawei	
	H. Ballani	
	Cornell U.	
	D. Jen	
	UCLA	
	R. Raszuk	
	Self	
	L. Zhang	
	UCLA	
	October 19, 2009	

TOC

Proposal for Auto-Configuration in Virtual Aggregation
draft-ietf-grow-va-auto-00.txt

Status of this Memo

Copyright Notice

Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Abstract

Virtual Aggregation as specified in [I-D.ietf-grow-va] (Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, “FIB Suppression with Virtual Aggregation,” May 2009.) requires a certain amount of configuration, namely virtual prefixes (VP), a VP list, type of tunnel, and popular prefixes. This draft proposes optional approaches to auto-configuration of popular prefixes and the VP list, and discusses the pros and cons of each. If these proposals are accepted, they will be incorporated into [I-D.ietf-grow-va] (Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, “FIB Suppression with Virtual Aggregation,” May 2009.).

Table of Contents

---

1. Introduction

TOC

Virtual Aggregation as specified in [I-D.ietf-grow-va] (Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, “FIB Suppression with Virtual Aggregation,” May 2009.) requires a certain amount of configuration, namely:

1. Each Aggregation Point Router (APR) must be configured with the VPs for which it is an APR.

2. Every router must be configured with the VP list (a list of all VPs). This allows the router to know which prefixes can and cannot be FIB-suppressed.

3. Every router should be configured with a list of prefixes that should be FIB-installed (for instance because they have large traffic volumes).

4. Every router should be configured as to the tunnel type.

Of these four items, the first and last cannot be automated. Both, however, represent a relatively small amount of configuration. The second and third are more significant, and this draft proposes mechanisms for partially or fully automating them. If any of these proposals are accepted, they will be incorporated into the main VA draft. In any event, they would be considered as optional. The manually configured VP-list would still be mandatory, though an ISP could choose not to use it if one of the options described here is available. ([I-D.ietf-grow-va] (Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, “FIB Suppression with Virtual Aggregation,” May 2009.)).
All of the approaches described in this draft involve tagging routes with a standard extended communities attribute. There are three such tags, the "should-install" tag, the "VP-route" tag, and the "can-suppress" tag. The should-install tag is for the purpose of automating the configuration of popular prefixes that are popular by virtue of having high traffic volume. The VP-route and can-suppress tags represent two alternatives for the VP-list. Note that usage of the should-install tag (popular prefixes config) is completely orthogonal with usage of either the VP-route or can-suppress tag (replacement for VP-list config).

1.1. Requirements notation

TOC

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119] (Bradner, S., “Key words for use in RFCs to Indicate Requirement Levels,” March 1997.).

TOC

2. Syntax for the tags

All three tags can be conveyed with an Extended Communities Attribute [RFC4360] (Sangli, S., Tappan, D., and Y. Rekhter, “BGP Extended Communities Attribute,” February 2006.) to be assigned by IANA. For all three tags, the Transitive Bit MUST be set to value 1 (the community is non- transitive across ASes).

---

3. Config of Popular Prefixes

TOC

Broadly speaking, a Popular Prefix is any prefix that does not have to be FIB-installed, but should never-the-less be FIB-installed. For instance:

> Prefixes for customer networks should be installed (so that traffic to customers does not incur the extra delay associated with the detour through an APR). Since customer routes are in any event tagged with a community attribute for routing policy reasons, the decision to FIB-install them is entirely local and requires no standardization.
>
> If an ASBR chooses its external peer as a next-hop for a given prefix, then it should FIB-install that prefix.

Prefixes to which there is a large volume of traffic should also be FIB-installed. This is to reduce the additional load the results from the extra hop(s) that packets must take on the APR detour. Installing these prefixes is not trivial. The volume of traffic must be measured, the high-volume prefixes identified, and routers configured to FIB-install these prefixes. Furthermore, the router where the prefix must be FIB-installed is typically different from where the high-volume is measured. Normally, the highest volume for any given prefix will be seen at the egress routers for that prefix. However, the ingress router is where FIB installation should take place.
The proposal is to identify high-volume prefixes at ASBRs and RRs (routers that forward iBGP updates), and to tag routes to these prefixes with a community attribute that effectively means "should FIB-install". How to identify high-volume prefixes is a local matter, but one way would be by examining netflow records from the router. In principle, however, a router could internally detect high-volume prefixes. Identification of high-volume prefixes need only be done for either:

1. Outgoing traffic on ASBRs peering with non-customer networks (peers or transits).

2. Route Reflectors, probably limited to traffic that is routed towards the edge.

Either way, the set of routers where this identification must take place is limited.

3.1. Operation of the should-install tag

TOC

3.1.1. Sending the should-install tag

TOC

For routers implementing this optional feature, it must be possible to configure a router to attach the community attribute (the "should-install tag") to routes for a given prefix. In practice, this may be automatically done by the system that receives and analyzes netflow records, or it may be done manually by a network administrator. Once configured as such, the router must attach the should-install tag to BGP updates containing the prefix. The update may be generated immediately after the configuration takes place, or it may be put off until the next time the update is normally transmitted.
If the configuration is removed, the router must not attach the should-install tag to subsequent updates containing the prefix. An update without the should-install tag may be generated immediately after the configuration is removed, or it may be put off until the next time the update is normally transmitted.

3.1.2. Receiving the should-install tag

TOC

If the best-path route to a given prefix (that doesn't otherwise have to be FIB-installed), has the should-install tag, then the router locally decides whether or not to FIB-install the prefix. If there is no room in the FIB for a new prefix, the router may choose to remove an existing FIB entry (for instance, the oldest entry) to make room for the new entry.

3.2. Discussion

TOC

The time-frame over which should-install tags are attached and removed should be quite long, at least hours if not days. Evidence shows that

high-volume prefixes tend to stay high-volume on average over long periods of times (days or even weeks) [nsdi09] (Ballani, H., Francis, P., Cao, T., and J. Wang, “Making Routers Last Longer with ViAggre,” April 2009.).

There are a number of limited scenarios whereby a should-install tag is not successfully conveyed to all routers in an AS. This does not result in non-delivery of packets, only inefficiencies.

Consider the case where an AS is using Route Reflectors (RRs), and is using ASBRs to transmit should-install tags. Imagine two ASBRs, BR1 and BR2, that advertise routes to some prefix P. Further, both BR1 and BR2 are clients of the same RR. Assume that there is high-volume to prefix P at BR1 but not at BR2. As a result, BR1 attaches the should-install tag and BR2 does not. If the RR for any reason prefers the route via BR2 over BR1, then it the should-install tag will not be passed on by the RR. (Although note that a likely outcome of this is that BR2 will start to see high volumes of traffic to P, and eventually will set the should-install tag.)

Next consider the same topology as above (BR1 and BR2 both clients of the same RR), but now assume that it is the RR that is used to transmit should-install tags. Assume that the RR detects high-volume to prefix P and attaches the should-install tag for routes to P. Assume that both BR1 and BR2 choose their respective external peers as the next hop to P, and of course advertise this next hop to the RR. The RR selects and advertises a best path, say via BR1. When the RR advertises this best path to BR2, BR2 ignores it and so does not FIB-install the route. The end result here is that packets detour through an APR and then are tunneled back to the ASBR. (Though as mentioned earlier in this section, prefixes where the next hop is an external peer should be FIB-installed as a matter of local policy.)

---

4. Config of the VP list

TOC

As the current VA specification stands, routers have to know which prefixes they must FIB-install and and which they need not FIB-install. The VP-list tells them this: they must FIB-install routes to VPs, and they need not FIB-install routes to prefixes that fall within VPs for which they are not an APR. The same VP-list must be installed in every router (though it is not a problem that they differ for brief periods during modification of the VP-list). Configuration of the VP-list is not nearly as hard as configuration of popular prefixes, but it is nevert-the-less a significant task that we'd just as soon do without. There are two basic approaches to automating this configuration. One is to have APRs tag the routes to VPs that they originate, and let routers effectively reconstruct the VP-list from these tags. This approach has the advantage that no configuration what-so-ever is required to solve the problem.

The other is to have ASBRs tag the routes that need not be installed. This can be done by configuring a list of one or more "VP-ranges" in the ASBRs. This is simpler than the current configured VP-list approach in two regards. First, fewer routers need to be configured (only ASBRs interfacing with peer and provider (non-customer) networks. Second, the VP-range is simpler than the VP-list. In most cases, once an ISP is past its initial VA roll-out phase, it would consist of a single 0/0 entry.
These two approaches are discussed in the following sections.

4.1. VP-route tag

TOC

Routers that receive a route with the communities attribute indicating the VP-route tag must FIB-install the associated prefix (VP). They may FIB-suppress any sub-prefixes that fall within the VP.
Prefixes that do not fall within any known VP must be FIB-installed. During BGP initialization (i.e. before the End-of-RIB marker is received [RFC4724] (Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, “Graceful Restart Mechanism for BGP,” January 2007.)), however, the full set of VPs is not yet known. Therefore, what routers do with prefixes that do not fall within any known VP during initialization is a local matter.
There are two basic strategies, install by default and suppress by default. Each has pros and cons, though the latter is generally prefered. With install by default, some prefixes will be installed only to be removed later (when the parent VP is learned). This can actually ultimately slow down convergence, since it takes time to modify the FIB. Also, this could result in the FIB filling up with entries.
The problem with suppress by default is that entries that ultimately will be installed are not immediately installed. Instead, they are installed only after the End-of-RIB marker. This approach, however, does avoid the pitfalls of install by default, and ultimately could converge faster because FIB churn is avoided. There are also several mitigating factors that should make suppress by default work well in practice. First, if the router uses Graceful Restart [RFC4724] (Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, “Graceful Restart Mechanism for BGP,” January 2007.), then in any event forwarding can continue to take place even when the BGP session is restarted. Second, the router can have a policy whereby prefixes with a should-install tag are automatically installed. In this way, high-volume prefixes are installed and so most traffic will in fact be forwarded by the End-of-RIB. Finally, if the router has a policy that customer prefixes are always installed, then flows between customers are also correctly forwarded by the End-of-RIB.
Another issue with the VP-route tag is what to do if all APRs for a given VP stop operating (i.e. crash) and so all VP routes are

withdrawn. Strictly speaking, the router would immediately start installing the sub-prefixes within that VP. This could lead to the FIB filling up. Also, if the APR is thrashing (going up and down), then all routers in the AS could end up repeatedly adding and removing the same set of prefixes.

How to deal with this is a local matter. There are two questions the router must answer:

1. How should hysteresis be applied to the (implicit) VP list to avoid FIB churn?

2. How are FIB entries prioritized in the case where the FIB is full?

Regarding VP list hysteresis, perhaps the simplest thing to do is to use standard route flap damping on the VP routes [RFC2439] (Villamizar, C., Chandra, R., and R. Govindan, "BGP Route Flap Damping," November 1998.). Alternatively, the router could simply not install sub-prefixes for a recently known VP for some period of time (minutes) after which the VP route was withdrawn, or only install sub-prefixes slowly (to minimize the impact of churn).

Regarding FIB entry prioritization, routers must in any event install VP routes and sub-prefixes within the VPs for which the router is an APR. If the FIB does not have room for at least these entries, then VA has simply been configured incorrectly in the AS, and the administrator must fix this. Beyond these necessary FIB entries, prioritization is a local matter. A reasonable prioritization, however, is the following: 1) customer routes, 2) routes with should-install tag, 3) routes for sub-prefixes of recently withdrawn VPs, 4) other.

4.2. Can suppress tag

TOC

With this approach, some set of ASBRs are configured with a "VP range". This is the ranges of IP address that are covered by all VPs. In a mature deployment of VA, the range would amount to all IP addresses, in which case the VP range is simply 0/0. Early in VA deployment, when an ISP is still in the testing or roll-out phase, the VP range would consist of multiple entries. At a minimum, the set of ASBRs so configured are those with peers in peer or transit ASes. If the AS has a policy that customer routes are always FIB-installed, then it is not necessary to configure routers that connect to customer ASes.

VP-range configured ASBRs must tag any route whose prefix falls within the VP range with a "can-suppress" tag, with the following exceptions:

1. Routers must never tag a VP route with can-suppress.

2. If the ISP has a policy of FIB-installing customer routes, then routes received from customers should not be tagged with can-suppress.

A router receiving a route with a can-suppress tag first determines if it must FIB-install the prefix. It would have to do this for instance if the prefix falls within a VP for which it is an APR. If the router does not have to install the prefix, then it may suppress the prefix at its own discretion.
When the can-suppress approach is used, then routers must FIB-install any prefixes not tagged as can-suppress. The primary reason for this is so that VP routes are always installed.
Note that in the case where all VP routes for a given VP are withdrawn, routers would not be able to FIB-install the (now unreachable) sub-prefixes. This is because, with the can-suppress approach, routers do not actually know which routes are VPs.

---

5. IANA Considerations

TOC

IANA must assign type values for the Extended Communities Attributes that convey the tags.

---

6. Security Considerations

TOC

As of this writing, there are no known new security threats introduced by this draft.

---

7. References

TOC

7.1. Normative References

TOC

[I-D.ietf-grow-va]	Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, “FIB Suppression with Virtual Aggregation,” draft-ietf-grow-va-00 (work in progress), May 2009 (TXT).
[RFC1997]	Chandrasekeran, R., Traina, P., and T. Li, “BGP Communities Attribute,” RFC 1997, August 1996 (TXT).
[RFC2119]	

	Bradner, S., “Key words for use in RFCs to Indicate Requirement Levels,” BCP 14, RFC 2119, March 1997 (TXT, HTML, XML).
[RFC4360]	Sangli, S., Tappan, D., and Y. Rekhter, “BGP Extended Communities Attribute,” RFC 4360, February 2006 (TXT).
[RFC4724]	Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, “Graceful Restart Mechanism for BGP,” RFC 4724, January 2007 (TXT).

7.2. Informative References

TOC

[RFC2439]	Villamizar, C., Chandra, R., and R. Govindan, “BGP Route Flap Damping,” RFC 2439, November 1998 (TXT, HTML, XML).
[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

TOC

	Paul Francis
	Max Planck Institute for Software Systems
	Gottlieb-Daimler-Strasse
	Kaiserslautern 67633
	Germany
Phone:	+49 631 930 39600
Email:	francis@mpi-sws.org
	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
	Hitesh Ballani
	Cornell University
	4130 Upson Hall
	Ithaca, NY 14853
	US
Phone:	+1 607 279 6780
Email:	hitesh@cs.cornell.edu

	Dan Jen
	UCLA
	4805 Boelter Hall
	Los Angeles, CA 90095
	US
Phone:	
Email:	jenster@cs.ucla.edu
	Robert Raszuk
	Self
Phone:	
Email:	robert@raszuk.net
	Lixia Zhang
	UCLA
	3713 Boelter Hall
	Los Angeles, CA 90095
	US
Phone:	
Email:	lixia@cs.ucla.edu