GROW Working Group R. Raszuk
Internet-Draft NTT MCL
Intended status: Informational J. Heitz
Expires: December 7, 2012 Ericsson
A. Lo
Arista
L. Zhang
UCLA
X. Xu
Huawei
June 5, 2012

Simple Virtual Aggregation (S-VA)
draft-ietf-grow-simple-va-09.txt

Abstract

The continued growth in the Default Free Routing Table (DFRT) stresses the global routing system in a number of ways. One of the most costly stresses is Forwarding Information Base (FIB) size: ISPs often must upgrade router hardware simply because the FIB has run out of space, and router vendors must design routers that have adequate FIB.

FIB suppression is an approach to relieving stress on the FIB by NOT loading selected RIB entries into the FIB. Simple Virtual Aggregation (S-VA) is a simple form of Virtual Aggregation (VA) that allows any and all edge routers to shrink their RIB and FIB requirements substantially and therefore increase their useful lifetime.

S-VA does not increase FIB requirements for core routers. S-VA is extremely easy to configure considerably more so than the various tricks done today to extend the life of edge routers. S-VA can be deployed autonomously by an ISP (cooperation between ISPs is not required), and can co-exist with legacy routers in the ISP.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on December 7, 2012.

Table of Contents

1. Introduction

ISPs today manage constant Default Free Routing Table (DFRT) growth in a number of ways. One way, of course, is for ISPs to upgrade their router hardware before DFRT growth outstrips the size of the FIB. This is too expensive for many ISPs. They would prefer to extend the lifetime of routers whose FIBs can no longer hold the full DFRT.

A common approach taken by lower-tier ISPs is to default route to their providers. Routes to customers and peer ISPs are maintained, but everything else defaults to the provider. This approach has several disadvantages. First, packets to Internet destinations may take longer-than-necessary AS paths.

This problem can be mitigated through careful configuration of partial defaults, but this can require substantial configuration overhead. A second problem with defaulting to providers is that the ISP is no longer able to provide the full DFRT to its customers. Finally, provider defaults prevents the ISP from being able to detect martian packets. As a result, the ISP transmits packets that could otherwise have been dropped over its expensive provider links.

An alternative is for the ISP to maintain full routes in its core routers, but to filter routes from edge routers that do not require a full DFRT. These edge routers can then default route to the core or exit routers. This is often possible with edge routers that interface to customer networks. The problem with this approach is that it cannot be used for all edge routers. For instance, it cannot be used for routers that connect to transits. It should also not be used for routers that connect to customers which wish to receive the full DFRT.

This draft describes a very simple technique, called Simple Virtual Aggregation (S-VA), that allows any and all edge routers to have substantially reduced FIB requirements even while still advertising and receiving the full DFRT over BGP. The basic idea is as follows. Core routers in the ISP maintain the full DFRT in the FIB and RIB. Edge routers maintain the full DFRT in the BGP protocol RIB, but suppress certain routes from being installed in RIB and FIB tables. Edge routers may install a default route to core routers, to ABRs which are installed on the Point of Presence (POP) to core boundary or to the ASBR routers.

S-VA requires no changes to BGP and no changes to any choice of forwarding mechanisms in routers. Configuration is extremely simple: S-VA must be enabled on the edge router which needs to save its RIB and FIB space. In the same time operator must inject into his intra-

domain routing a new prefix further called virtual aggregate (VA-prefix) which will be used as the aggregate forwarding reference by the edge routers performing S-VA. Everything else is automatic. ISPs can deploy FIB suppression autonomously and with no coordination with neighbor ASes.

In configurations where BGP routes are used to resolve other routes or where BGP routes are redistributed to other protocols which both happen via RIB simple-va can rather then suppressing routes before they are installed in global RIB flag them as "suppress eligible". That will allow for seamless route resolution or redistribution while in the same time FIB size will continue to be limited as previously flagged routes will not be send from RIB to FIB.

1.1. Scope of this Document

The scope of this document is limited to Intra-domain S-VA operation. In other words, the case where a single ISP autonomously operates S-VA internally without any coordination with neighboring ISPs.

Note that this document assumes that the S-VA "domain" (i.e. the unit of autonomy) is the AS (that is, different ASes run S-VA independently and without coordination). For the remainder of this document, the terms ISP, AS, and domain are used interchangeably.

This document applies equally to IPv4 and IPv6 both unicast and multicast address families.

S-VA may operate with a mix of upgraded routers and legacy routers. There are no topological restrictions placed on the mix of routers. S-VA functionality is local to the router on which it is enabled and routing correctness is guaranteed.

Note that S-VA is a greatly simplified variant of "full VA" [I-D.ietf-grow-va]. With full VA, all routers (core or otherwise) can have reduced FIBs. However, full VA requires substantial new configuration and operational complexity compared to S-VA. Full VA also requires the use of MPLS LSPs between all routers. Note that S-VA was formerly specified in [I-D.ietf-grow-va]. It has been moved to this separate draft to simplify its understanding.

1.2. Requirements notation

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].

1.3. Terminology

RIB/FIB-Installing Router (FIR): An router that does not suppress any routes, and advertises itself as a default route for 0/0. Typically a core router, POP to core boundary router or an ASBR would be configured as an FIR.

RIB/FIB-Suppressing Router (FSR): An S-VA router that installs a route to 0/0, and may suppress other routes. Typically an edge router would be configured as an FSR.

Install and Suppress: The terms "install" and "suppress" are used to describe whether a protocol local RIB entry has been loaded or not loaded into the global RIB and FIB. In other words, the phrase "install a route" means "install a route into the global RIB and FIB", and the phrase "suppress a route" means "do not install a route from BGP into the global RIB and FIB".

Legacy Router: A router that does not run S-VA, and has no knowledge of S-VA.

Global Routing Information Base (RIB): The term global RIB is used to indicate the router's main routing information base. That RIB is normally used to populate FIB tables of the router. It needs to be highlighted that unless FIB compression is used global RIB and FIB tables are in sync.

Local/Protocol Routing Information Base (loc-RIB): The term local RIB is used to indicate the protocol's table where product of SPF or BGP best path selection is kept before being installed in global RIB. For example, in some protocol implementations BGP loc-RIB can be further divided into Adj-RIBs-In, the Loc-RIB, and the Adj-RIBs-Out.

2. Operation of S-VA

There are three types of routers in S-VA, FIB-Installing routers (FIR), FIB-Suppressing routers (FSR), and optionally legacy routers. While any router can be an FIR or an FSR (there are no topology constraints), the most simple form of deployment is for AS border or POP border routers to be configured as FIRs, and for customer facing edge routers respectively in the AS or in the POP to be configured as FSRs.

There are two basic network deployment scenarios for S-VA - with and without presence of a default route. In both cases simple VA operates in an identical way however when default route is found and is eligible to become a less specific prefix by router's configuration the S-VA may use it. That should not prevent detection of any other potential prefix with different next hop as the next hop of default route.

In the event of FIRs originating a default BGP route to NLRI 0/0 [RFC4271]. The ORIGIN is set to INCOMPLETE (value 2) and the BGP NEXT_HOP is set to match the other BGP routes which are also advertised by said FIR. The ATOMIC_AGGREGATE and AGGREGATOR attributes are not included. The FIR MUST attach a NO_EXPORT Community Attribute [RFC1997] to the default route.

FIRs should not FIB-suppress any routes. They may, however, still use some form of local FIB compression algorithm if deemed necessary.

FSRs must detect the VA prefix or prefixes (including 0/0) and install it both in loc-RIB, RIB and FIB. Following that FSR may suppress any more specific routes which carry the same next hop as the VA prefix. To guarantee semantical correctness FSR by default should also be able to detect installation of not matching next hop route and reinstall all the more specifics which were previously eligible for suppression to maintain semantical forwarding correctness.

Generally, any more specific route which carries the same next hop as the VA-prefix is eligible for suppression. However, provided that there was at least one less specific prefix with different next hop between VA-prefix and the prefixes which got suppressed then all previously suppressed prefixes must be reinstalled.

Similarly when IBGP multipath is enabled and when multiple VA prefixes are detected which are multipath candidates under given network condition only those more specific prefixes are subject to suppression which have the identical set of next hops as multipath set of VA prefixes.

We illustrate the expected behavior on the figure below. This figure shows an autonomous system with a FIR FIR1 and an FSR FSR1. FSR1 is an ASBR and is connected to two remote ASBRs, EP1 and EP2.

```
+-------------------------------------------+
|      Autonomous System                    |   +----+
|                                           |   |EP1 |
|                                       /---+---|    |
|   To   ----\ +----+          +----+ /     |   +----+
| Other       \|FIR1|----------|FSR1|/      |
|Routers      /|    |          |    |\      |
|        ----/ +----+          +----+ \     |   +----+
|                                       \---+---|EP2 |
|                                           |   |    |
|                                           |   +----+
+-------------------------------------------+
```


Suppose that FSR1 has been enabled to perform S-VA. Originally it receives all routes from FIR1 (doing next hop self) as well as directly connected EBGP peers EP1 and EP2. FIR1 now will advertise a VA prefix 0/0 with next hop set to himself. That will trigger detection of such prefix on FSR1 and suppression all routes which have the same next hop as VA prefix and which otherwise would be installed in RIB and FIB. However it needs to be observed that FSR1 will not suppress any EBGP routes received from his peers EP1 and EP2 due to next hop being different from the one assigned to VA-prefix.

3. Deployment considerations

The simplest deployment model of S-VA is its use within the POP. In such model the POP to core boundary routers (usually RRs in the data path) would act as FIRs and would inject VA-prefix 0/0 to all of its clients within the POP. In such model of operation an observation can be made that such ABRs do have full routing knowledge and client to ABR distance is negligible as compared with client to intra-domain exit distance.

Therefore under the above intra POP S-VA deployment model clients can be configured that even in the event of lack of ABR to ABR advertisement symmetry there is still no need to monitor if more specific unsuppressed route would cover suppressed one. Thus in this particular deployment model there is no need to detect and reinstall the previously suppressed ones.

Another deployment consideration should be given to networks which may utilize route reflection. In the event of enabling IBGP multipath a special care must be taken that both outbound prefixes as well as VA-prefixes would pass via said route reflectors to their clients.

In order to address the above aspects the following solutions could be considered:

- Use of intra-POP S-VA
- Full mesh Small or medium side networks where S-VA can be deployed are normally fully meshed and do not use route reflection. It also needs to pointed out that some large networks are also fully meshed today.
- Use of add-paths Use of add-paths new BGP encoding will allow to distribute more then one overall best path from RR to each client.

- Alternate advertisement of VA-prefix S-VA prefix does not need to be advertised in BGP. The BGP suppression will happen as long as we configure the S-VA with next hop(s) and implementation verifies that such VA-prefix is installed in the RIB and FIB.

BGP routes may be used to resolve nexthops for static routes or other BGP routes. Because the default route does not imply reachability of any destination, a router can be configured not to resolve nexthops using the default route. In this case, simple-va should not suppress a route that may be used to resolve a nexthop for another route.

Selected BGP routes in the RIB may be redistributed to other protocols. If they no longer exist in the RIB, they will not be redistributed. This is especially important when the conditional redistribution is taking place based on the length of the prefix, community value etc .. In those cases where redistribution policy is in place simple-va code should refrain from suppressing prefixes matching such policy.

A router may originate a network route or an aggregate route into BGP. Some addresses covered by such a route may not exist. If this router were to receive a packet for an unreachable address within an originated route, it must not send that packet to the default route. There are several ways to achieve this. One is to have the FIR aggregate the routes instead of the FSR. Another is to install a blackhole route for the nonexistent addresses on the originating router. This issue is not specific to simple-va, but applicable to the general use of default routes.

4. IANA Considerations

There are no IANA considerations.

5. Security Considerations

The authors are not aware of any new security considerations due to S-VA.

6. Acknowledgements

The concept for Virtual Aggregation comes from Paul Francis. In this document authors only simplified some aspects of its behavior to allow simpler adoption by some operators.

Authors would like to thank Clarence Filsfils, Nick Hilliard, S.

Moonesamy and Tom Petch for their review and valuable input.

7. References

7.1. Normative References

[RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.

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

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

7.2. Informative References

[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-06 (work in progress), December 2011.

Authors' Addresses

Robert Raszuk
NTT MCL
101 S Ellsworth Avenue Suite 350
San Mateo, CA 94401
US

Email: robert@raszuk.net

Jakob Heitz
Ericsson
300 Holger Way
San Jose, CA 95135

USA

Phone:
Email: jakob.heitz@ericsson.com

Alton Lo
Arista Networks
5470 Great America Parkway
Santa Clara, CA 95054
USA

Phone:
Email: altonlo@aristanetworks.com

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

Phone:
Email: lixia@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