Network Working Group
Internet-Draft
Intended status: Informational
Expires: July 2, 2012

P. Francis
MPI-SWS
X. Xu
Huawei
H. Ballani
Cornell U.
D. Jen
UCLA
R. Raszuk, Ed.
NTT MCL Inc.
L. Zhang
UCLA

December 30, 2011

FIB Suppression with Virtual Aggregation
draft-ietf-grow-va-06.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 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. Virtual Aggregation (VA) allows ISPs to shrink the FIBs of any and all routers, easily by an order of magnitude with negligible increase in path length and load. FIB suppression can be deployed autonomously by an ISP without requiring cooperation between adjacent ISPs, and can co-exist with legacy routers in the ISP.

This Internet-Draft will expire on July 2, 2012.

Copyright Notice

Table of Contents

1. Introduction

ISPs today manage constant 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 may be 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 transit 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 Autonomous System (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 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 of course also does not help in cases where core routers themselves have inadequate FIB capacity.

FIB Suppression is an approach to shrinking FIB size that requires no changes to BGP, no changes to packet forwarding mechanisms in routers, and relatively minor changes to control mechanisms in routers and configuration of those mechanisms. The core idea behind FIB suppression is to run BGP as normal, and in particular to not shrink the RIB, but rather to not load certain RIB entries into the FIB. This approach minimizes changes to routers, and in particular is simpler than more general routing architectures that try to shrink both RIB and FIB. With FIB suppression, there are no changes to BGP per se. The BGP decision process does not change, the selected AS-PATH does not change, and except on rare occasion the exit router does not change. ISPs can deploy FIB suppression autonomously and with no coordination with neighboring ASes.

This document describes an approach to FIB suppression called "Virtual Aggregation" (VA). VA operates by organizing the IP (v4 or v6) address space into Virtual Prefixes (VP), and using tunnels to aggregate the (regular) sub-prefixes within each VP. The decrease in

FIB size can be dramatic, easily 5x or 10x with only a slight path length and router load increase [nsdi09].

1.1. Scope of this Document

The scope of this document is limited to intra-domain VA operation. Individual ASs autonomously operate VA internally without any coordination with neighboring ASs. For the remainder of this document, the terms ISP, AS, and domain are used interchangeably.

This document applies equally to IPv4 and IPv6.

This document is limited to the following tunnel types: MPLS Label Switched Paths (LSP), and of MPLS inner labels tunneled over either LSPs or IP headers.

VA may operate with a mix of upgraded routers and legacy routers. There are no topological restrictions placed on the mix of routers. In order to avoid loops between upgraded and legacy routers, packets are always tunneled by the VA routers to the BGP NEXT_HOPs of the matched BGP routes. If a given local ASBR (Autonomous System Border Router) is a legacy router, it must be able to terminate tunnels.

1.2. Requirements notation

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" when capitalized in this document are to be interpreted as described in [RFC2119].

1.3. Terminology

Aggregation Point Router (APR): An Aggregation Point Router (APR) is a router that aggregates a Virtual Prefix (VP) by installing routes (into the FIB) for all of the sub-prefixes within the VP. APRs advertise the VP to other routers with BGP. For each sub-prefix within the VP, APRs have a tunnel from themselves to the remote ASBR (Autonomous System Border Router) where packets for that prefix should be delivered.

Install and Suppress: The terms "install" and "suppress" are used to describe whether a RIB entry has been loaded or not loaded into the FIB. In particular, "install a route" means "install a route into the FIB", and "suppress a route" means "do not install a route into the FIB".

Legacy Router: A router that does not run VA, and has no knowledge of VA. Legacy routers, however, must be able to terminate tunnels when they are local ASBRs.

Non-APR Router: In discussing VPs, it is often necessary to distinguish between routers that are APRs for that VP, and routers that are not APRs for that VP (but of course may be APRs for other VPs not under discussion). In these cases, the term "APR" is taken to mean "a VA router that is an APR for the given VP", and the term "non-APR" is taken to mean "a VA router that is not an APR for the given VP". The term non-APR router is not used to refer to legacy routers.

Popular Prefix: A Popular Prefix is a sub-prefix that is installed in a router in addition to the sub-prefixes it holds by virtue of being a Aggregation Point Router. The Popular Prefix allows packets to follow the shortest path. Note that different routers do not need to have the same set of Popular Prefixes.

Routing Information Base (RIB): The term RIB is used rather sloppily in this document to refer either to the loc-RIB (as used in [RFC4271]), or to the combined Adj-RIBs-In, the Loc-RIB, and the Adj-RIBs-Out.

Sub-Prefix: A regular (physically aggregatable) prefix. These are equivalent to the prefixes that would normally comprise the DFRT in the absence of VA. A VA router will contain a sub-prefix entry either because the sub-prefix falls within a Virtual Prefix for which the router is an APR, or because the sub-prefix is installed as a Popular Prefix. Legacy routers hold the same sub-prefixes that they hold today.

Tunnel: This document specifies the use of MPLS Label Switched Paths (LSP), and of MPLS inner labels tunneled over either LSPs or IP headers. While in principle other types of tunnels may be used, they are not specified here. This document uses the term tunnel to refer to the above MPLS encapsulations.

VA router: A router that operates Virtual Aggregation according to this document.

Virtual Prefix (VP): A Virtual Prefix (VP) is a prefix used to aggregate its contained regular prefixes (sub-prefixes). The set of sub-prefixes in a VP are not physically aggregatable, and so they are aggregated at APRs through the use of tunnels.

VP-List: A list of defined VPs. All routers must agree on the contents of this list.

2. Overview of Virtual Aggregation (VA)

For descriptive simplicity, this section starts by describing VA assuming that there are no legacy routers in the domain. Section 2.1 overviews the additional functions required by VA routers to accommodate legacy routers.

A key concept behind VA is to operate BGP as normal, and in particular to populate the RIB with the full DFRT, but to suppress many or most prefixes from being loaded into the FIB. By populating the RIB as normal, we avoid any changes to BGP, and changes to router operation are relatively minor. The basic idea behind VA is as follows: The address space is partitioned into large prefixes --- larger than any aggregatable prefix in use today. These prefixes are called Virtual Prefixes (VP). Different VPs do not need to be the same size. They may be a mix of /6, /7, /8 (for IPv4), and so on. Indeed, an ISP can define a single /0 VP, and use it for a core/edge type of configuration. That is, the core routers would maintain full FIBs, and edge routers could maintain default routes to the core routers, and suppress as much of the FIB as they wish. Each ISP can independently select the size of its VPs.

VPs are not themselves topologically aggregatable. VA makes the VPs aggregatable through the use of tunnels, as follows. Associated with each VP are one or more "Aggregation Point Routers" (APR). An APR (for a given VP) is a router that installs routes for all sub-prefixes (i.e. real physically aggregatable prefixes) within the VP. By "install routes" here, we mean:

1. The route for each of the sub-prefixes is loaded into the FIB, and
2. there is a tunnel from the APR to the BGP NEXT_HOP for the route.

The APR originates a BGP route to the VP. This route is distributed within the domain, but not outside the domain. With this structure in place, a packet transiting the ISP goes from the ingress router to the APR (usually via a tunnel), and then from the APR to the BGP NEXT_HOP router via a tunnel. VA can operate with MPLS LSPs, or with MPLS inner labels over LSPs or IP headers. Section 4 specifies the usage of MPLS tunnels. Other tunnel types (i.e., GRE) may be used, but are not specified in this document.

The BGP NEXT_HOP can be either the local ASBR or the remote ASBR. In the former case, an inner label is used to tunnel packets (Section 4.1). In either case, all tunnel headers are stripped by the local ASBR before the packet is delivered to the remote ASBR. In other words, the remote ASBR sees a normal IP packet, and is completely unaware of the existence of VA in the neighboring ISP.

Note that the AS-PATH is not effected at all by VA. This means among other things that AS-level policies are not effected by VA. The packet may not, however, follow the shortest path within the ISP (where shortest path is defined here as the path that would have been taken if VA were not operating), because the APR may not be on the shortest path between the ingress and egress routers. When this

happens, the packet experiences additional latency and creates extra load (by virtue of taking more hops than it otherwise would have). Note also that, with VA, a packet may occasionally take a different exit point than it otherwise would have. This can occur for instance when the exit point nearest to the selected APR is different than the exit point nearest to the router initiating the tunnel to the APR.

VA can avoid traversing the APR for selected routes by installing these routes in non-APR routers. In other words, even if an ingress router is not an APR for a given sub-prefix, it MAY install that sub-prefix into its FIB. Packets in this case are tunneled directly from the ingress to the BGP NEXT_HOP. These extra routes are called "Popular Prefixes", and are typically installed for policy reasons (e.g. customer routes are always installed), or for sub-prefixes that carry a high volume of traffic (Section 3.5.1). Different routers may have different Popular Prefixes. As such, an ISP may assign Popular Prefixes per router, per POP, or uniformly across the ISP. A given router may have zero Popular Prefixes, or the majority of its FIB may consist of Popular Prefixes. The effectiveness of Popular Prefixes to reduce traffic load relies on the fact that traffic volumes follow something like a power-law distribution: i.e. that 90% of traffic is destined to 10% of the destinations. Internet traffic measurement studies over the years have consistently shown that traffic patterns follow this distribution [nsdi09], though there is no guarantee that they always will.

Note that for routing to work properly, every packet must sooner or later reach a router that has installed a sub-prefix route that matches the packet. This would obviously be the case for a given sub-prefix if every router has installed a route for that sub-prefix. If this is not the case, then there MUST be at least one Aggregation Point Router (APR) for the sub-prefix's Virtual Prefix (VP). Ideally, every POP contains at least two APRs for every Virtual Prefix. By having APRs in every POP, the latency imposed by routing to the APR is minimal (the extra hop is within the POP). By having more than one APR, there is a redundant APR should one fail. In practice it is often not possible to have an APR for every VP in every POP. This is because some POPs may have only one or a few routers, and therefore there may not have enough cumulative FIB space in the POP to hold every sub-prefix. Note that any router ("edge", "core", etc.) MAY be an APR.

It is important that both the contents of BGP RIBs, as well as the contents of the Routing Table (as defined in Section 3.2 of [RFC4271]) not be modified by VA (other than the introduction of routes to VPs). This is because PIM-SM [RFC4601] relies on the contents of the Routing Table to build its own trees and forwarding table. Therefore, FIB suppression MUST take place between the

Routing Table and the actual FIB(s).

2.1. Mix of Legacy and VA Routers

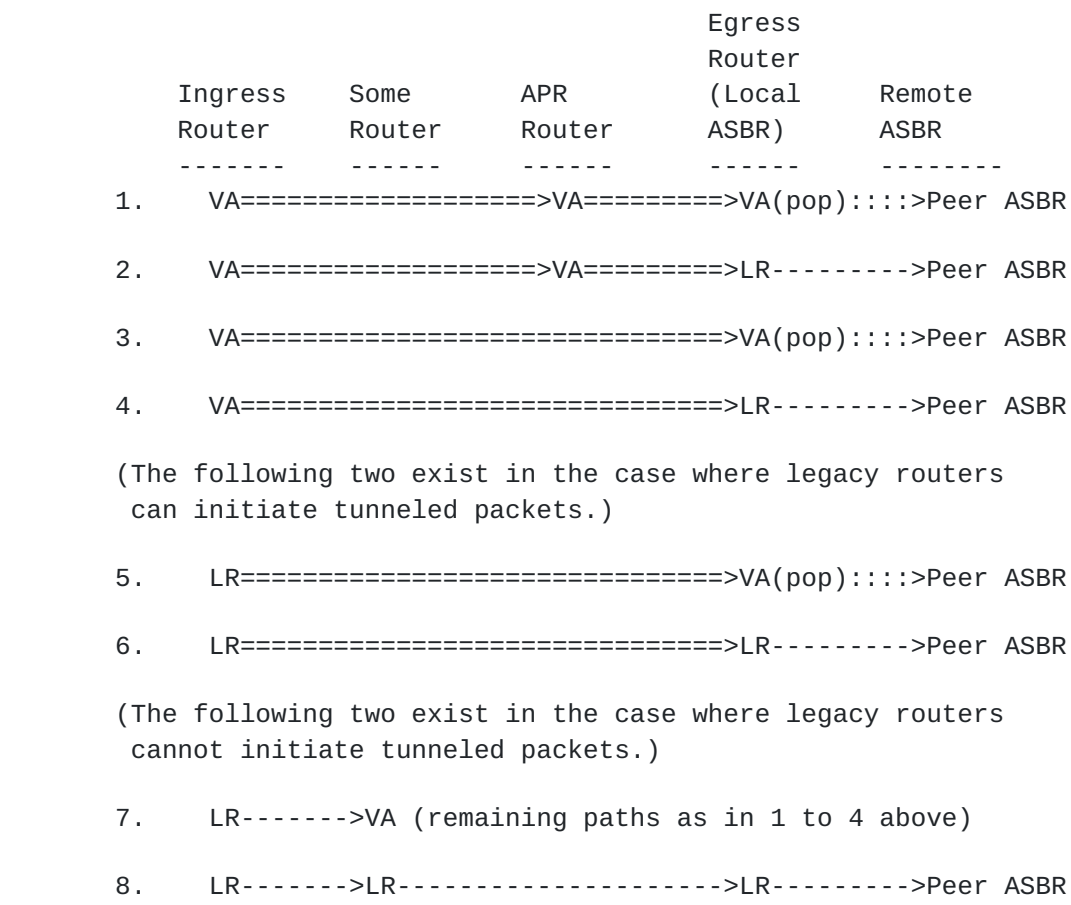
It is important that an ISP be able to operate with a mix of "VA routers" and "legacy routers". This allows ISPs to deploy VA in an incremental fashion and to continue to use routers that for whatever reason cannot be upgraded. This document allows such a mix, and indeed places no topological restrictions on that mix. It does, however, require that legacy routers are able to forward tunneled packets, are able to serve as tunnel endpoints, and are able to participate in distribution of tunnel information required to establish themselves as tunnel endpoints. Depending on the tunnel type, legacy routers MAY also be able to initiate tunneled packets, though this is an OPTIONAL requirement. Legacy routers MUST use their own address as the BGP NEXT_HOP.

2.2. Summary of Tunnels and Paths

To summarize, the following tunnels are created:

1. From all VA routers to all BGP NEXT_HOP addresses (where the BGP NEXT_HOP address is either an APR, a local ASBR, or the remote ASBR neighbor of a VA router).
2. Optionally, from all legacy routers to all BGP NEXT_HOP addresses.

There are a number of possible paths that packets may take through an ISP, summarized in the following diagram. Here, "VA" is a VA router, "LR" is a legacy router, the symbol "==>" represents a tunneled packet (through zero or more routers), "-->" represents an untunneled packet, and "(pop)" represents stripping the tunnel header. The symbol "::>" represents the portion of the path where although the tunnel is targeted to the receiving node, the outer header has been stripped.



The first and second paths represent the case where the ingress router does not have a Popular Prefix for the destination, and MUST tunnel the packet to an APR. The third and fourth paths represent the case where the ingress router does have a Popular Prefix for the destination, and so tunnels the packet directly to the egress. The fifth and sixth paths are similar to the third and fourth paths respectively, but where the ingress is a legacy router that can initiate tunneled packets, and effectively has the Popular Prefix by virtue of holding the entire DFRT. (Note that some ISPs have only partial RIBs in their customer-facing edge routers, and default route to a router that holds the full DFRT. This case is not shown here, but works perfectly well.) Finally, paths 7 and 8 represent the case where legacy routers cannot initiate a tunneled packet.

VA prevents the routing loops that might otherwise occur when VA routers and legacy routers are mixed. In particular, VA avoids the case where a legacy router is forwarding packets towards the BGP NEXT_HOP, while a VA router is forwarding packets towards the APR, with each router thinking that the other is on the shortest path to their respective targets.

In the first four types of path, the loop is avoided because tunnels are used all the way to the egress. As a result, there is never an opportunity for a legacy router to try to route based on the destination address unless the legacy router is the egress, in which case it forwards the packet to the remote ASBR.

In the 5th and 6th cases, the ingress is a legacy router, but this router can initiate tunnels and has the full FIB, and so simply tunnels the packet to the egress router.

In the 7th and 8th cases, the legacy ingress cannot initiate tunnels, and so forwards the packet hop-by-hop towards the BGP NEXT_HOP. The packet will work its way towards the egress router, and will either progress through a series of legacy routers (in which case the IGP prevents loops), or it will eventually reach a VA router, after which it will take tunnels as in the 1st and 2nd cases.

3. Specification of VA

This section describes in detail how to operate VA.

3.1. Legacy Routers

VA can operate with a mix of VA and legacy routers. To prevent the types of loops described in Section 2.2, however, legacy routers MUST satisfy the following requirements:

1. When forwarding externally-received routes over iBGP, the BGP NEXT_HOP attribute MUST be set to the legacy router itself.
2. Legacy routers MUST be able to detunnel packets addressed to themselves at the BGP NEXT_HOP address. They MUST also be able to convey the tunnel information needed by other routers to initiate tunneled packets to them. If a legacy router cannot detunnel and convey tunnel parameters, then the AS cannot use VA.
3. Legacy routers MUST be able to forward all tunneled packets.
4. Every legacy router MUST hold its complete FIB. Note, however, that this FIB does not necessarily need to contain the full DFRT. This might be the case, for instance, if the router is an edge router that defaults to a core router.

As long as legacy routers participating in tunneling as described above there are no topological restrictions on the legacy routers. They may be freely mixed with VA routers without the possibility of forming sustained loops (Section 2.2).

3.2. Advertising and Handling Virtual Prefixes (VP)

3.2.1. Distinguishing VPs from Sub-prefixes

VA routers MUST be able to distinguish VPs from sub-prefixes. This is primarily in order to know which routes to install. In particular, non-APR routers SHOULD know which prefixes are VPs before they receive routes for those VPs, for instance when they first boot up. This is in order to avoid the situation where they unnecessarily start filling their FIBs with routes that they ultimately don't need to install (Section 3.5). This leads to the following requirement:

It MUST be possible to configure the complete list of VPs into all VA routers. This list is known as the VP-List.

3.2.2. Limitations on Virtual Prefixes

From the point of view of best-match routing semantics, VPs are treated identically to any other prefix. In other words, if the longest matching prefix is a VP, then the packet is routed towards the VP. If a packet matching a VP reaches an APR for that VP, and the APR does not have a better matching route, then the packet is discarded by the APR (just as a router that originates any prefix will discard a packet that does not have a better match).

The overall semantics of VPs, however, are slightly different from those of real prefixes. Without VA, when a router originates a route for a (real) prefix, the expectation is that the addresses within the prefix are within the originating AS (or a customer of the AS). For VPs, this is not the case. APRs originate VPs whose sub-prefixes exist in different ASes. Because of this, VPs MUST not be advertised across AS boundaries. This is done with NO_EXPORT Communities Attribute (Section 3.2.3).

It is up to individual domains to define their own VPs. VPs MUST be "larger" (span a larger address space) than any real sub-prefix. If a VP is smaller than a real prefix, then packets that match the real prefix will nevertheless be routed to an APR owning the VP, at which point the packet will be dropped if it does not match a sub-prefix within the VP (Section 6).

(Note that, in principle there are cases where a VP could be smaller than a real prefix. This is where the egress router to the real prefix is a VA router. In this case, the APR could theoretically tunnel the packet to the appropriate remote ASBR, which would then forward the packet correctly. On the other hand, if the egress router is a legacy router, then the APR could not tunnel matching packets to the egress. This is because the egress would view the VP

as a better match, and would loop the packet back to the APR. For this reason we require that VPs be larger than any real prefixes, and that APRs never install prefixes larger than a VP in their FIBs.)

It is valid for a VP to be a subset of another VP. For example, 20/7 and 20/8 can both be VPs. In fact, this capability is necessary for "splitting" a VP without temporarily increasing the FIB size in any router. (Section 3.2.5).

3.2.3. Aggregation Point Routers (APR)

For each VP for which a router is an APR, the router does the following:

1. The APR MUST originate a BGP route to the VP. In this route, the NLRI are all of the VPs for which the router is an APR. This is true even for VPs that are a subset of another VP. The ORIGIN is set to INCOMPLETE (value 2), the AS number of the APR's AS is used in the AS_PATH, and the BGP NEXT_HOP is set to the address of the APR. The ATOMIC_AGGREGATE and AGGREGATOR attributes are not included.
2. The APR MUST attach a NO_EXPORT Communities Attribute [RFC1997] to the route.
3. The APR MUST be able to detunnel packets addressed to itself at its BGP NEXT_HOP address. It MUST also be able to convey the tunnel information needed by other routers to initiate tunneled packets to them.
4. If a packet is received at the APR whose best match route is the VP (i.e. it matches the VP but not any sub-prefixes within the VP), then the packet MUST be discarded (see Section 3.2.2). This can be accomplished by never installing a prefix larger than the VP into the FIB, or by installing the VP as a route to \dev\null.

3.2.3.1. Selecting APRs

An ISP is free to select APRs however it chooses. The details of this are outside the scope of this document. Nevertheless, a few comments are made here. In general, APRs should be selected such that the distance to the nearest APR for any VP is small---ideally within the same POP. Depending on the number of routers in a POP, and the sizes of the FIBs in the routers relative to the DFRT size, it may not be possible for all VPs to be represented in a given POP. In addition, there should be multiple APRs for each VP, again ideally in each POP, so that the failure of one does not unduly disrupt traffic.

3.2.4. Non-APR Routers

A non-APR router MUST install at least the following routes:

1. Routes to VPs (identifiable using the VP-List).
2. Routes to all sub-prefixes that are not covered by any VP in the VP-List.

If the non-APR has a tunnel to the BGP NEXT_HOP of any such route, it MUST use the tunnel to forward packets to the BGP NEXT_HOP.

When an APR fails, routers must select another APR to send packets to (if there is one). This happens, however, through normal internal BGP convergence mechanisms.

3.2.5. Adding and deleting VPs

An ISP may from time to time wish to reconfigure its VP-List. There are a number of reasons for this. For instance, early in its deployment an ISP may configure one or a small number of VPs in order to test VA. As the ISP gets more confident with VA, it may increase the number of VPs. Or, an ISP may start with a small number of large VPs (i.e. /4's or even one /0), and over time move to more smaller VPs in order to save even more FIB. In this case, the ISP will need to "split" a VP. Finally, since the address space is not uniformly populated with prefixes, the ISP may want to change the size of VPs in order to balance FIB size across routers. This can involve both splitting and merging VPs. Of course, an ISP must be able to modify its VP-List without 1) interrupting service to any destinations, or 2) temporarily increasing the size of any FIB (i.e. where the FIB size during the change is no bigger than its largest size either before or after the change).

The first step for adding a VP is to configure the APRs for the VP. This causes the APRs to originate routes for the VP. Non-APR routers will install this route according to the rules in Section 3.2.4 even though they do not yet recognize that the prefix is a VP. Subsequently the VP is added to the VP-List of non-APR routers. The Non-APR routers can then start suppressing the sub-prefixes with no loss of service.

To delete a VP, the process is reversed. First, the VP is removed from the VP-Lists of non-APRs. This causes the non-APRs to install the sub-prefixes. After all sub-prefixes have been installed, the VP may be removed from the APRs.

In many cases, it is desirable to split a VP. For instance, consider the case where two routers, Ra and Rb, are APRs for the same prefix.

It would be possible to shrink the FIB in both routers by splitting the VP into two VPs (i.e. split one /6 into two /7's), and assigning each router to one of the VPs. While this could in theory be done by first deleting the larger VP, and then adding the smaller VPs, doing so would temporarily increase the FIB size in non-APRs, which may not have adequate space for such an increase. For this reason, we allow overlapping VPs.

To split a VP, first the two smaller VPs are added to the VP-Lists of all non-APR routers (in addition to the larger superset VP). Next, the smaller VPs are added to the selected APRs (which may or may not be APRs for the larger VP). Because the smaller VPs are a better match than the larger VP, this will cause the non-APR routers to forward packets to the APRs for the smaller VPs. Next, the larger VP can be removed from the VP-Lists of all non-APR routers. Finally, the larger VP can be removed from its APRs.

To merge two VPs, the new larger VP is configured in all non-APRs. This has no effect on FIB size or APR selection, since the smaller VPs are better matches. Next the larger VP is configured in its selected APRs. Next the smaller VPs are deleted from all non-APRs. Finally, the smaller VPs are deleted from their corresponding APRs.

3.3. Border VA Routers

A VA router that is an ASBR MUST do the following:

1. When forwarding externally-received routes over iBGP, if a tunnel with an inner label is used, the ASBR MUST set the BGP NEXT_HOP attribute to itself. Otherwise, the BGP NEXT_HOP attribute is left unchanged.
2. They MUST establish tunnels as described in Section 4.
3. The ASBR MUST detunnel the packet before forwarding the packet to the remote ASBR.
4. The ASBR MUST be able to forward the packet without a FIB lookup. In other words, the tunnel information itself contains all the information needed by the border router to know which remote ASBR should receive the packet.

3.4. Advertising and Handling Sub-Prefixes

Sub-prefixes are advertised and handled by BGP as normal. VA does not effect this behavior. The only difference in the handling of sub-prefixes is that they might not be installed in the FIB, as described in Section 3.5.

In those cases where the route is installed, packets forwarded to prefixes external to the AS MUST be transmitted via the tunnel

established as described in Section 3.3.

3.5. Suppressing FIB Sub-prefix Routes

Any route not for a known VP (i.e. not in the VP-List) is taken to be a sub-prefix. The following rules are used to determine if a sub-prefix route can be suppressed.

1. A VA router MUST NOT FIB-install a sub-prefix route for which there is no tunnel to the BGP NEXT_HOP address. This is to prevent a loop whereby the APR forwards the packet hop-by-hop towards the next hop, but a router on the path that has FIB-suppressed the sub-prefix forwards it back to the APR.
2. If the router is an APR, a route for every sub-prefix within the VP MUST be FIB-installed (subject to the above limitation that there be a tunnel).
3. If a non-APR router has a sub-prefix route that does not fall within any VP (as determined by the VP-List), then the route MUST be installed. This may occur because the ISP hasn't defined a VP covering that prefix, for instance during an incremental deployment build-up.
4. If an ASBR is using strict uRPF to do ingress filtering, then it MUST install routes for which the remote ASBR is the BGP NEXT_HOP [RFC2827]. Note that only an APR may do loose uRPF filtering, and then only for routes to sub-prefixes within its VPs.
5. All other sub-prefix routes MAY be suppressed. Such "optional" sub-prefixes that are nevertheless installed are referred to as Popular Prefixes. Note, however, that whether or not to install a given sub-prefix SHOULD NOT be based on whether or not there is an active route to a VP in the VP-List. This avoids the situation whereby, during BGP initialization, the router receives some sub-prefix routes before receiving the corresponding VP route, with the result that it installs routes in its FIB that it will only remove a short time later, possibly even overflowing its FIB.

3.5.1. Selecting Popular Prefixes

Individual routers MAY independently choose which sub-prefixes are Popular Prefixes. There is no need for different routers to install the same sub-prefixes. There is therefore significant leeway as to how routers select Popular Prefixes. As a general rule, routers should fill the FIB as much as possible, because the cost of doing so is relatively small, and more FIB entries leads to fewer packets taking a longer path. Broadly speaking, an ISP may choose to fill the FIB by making routers APRs for as many VPs as possible, or by assigning relatively few APRs and rather filling the FIB with Popular Prefixes. Several basic approaches to selecting Popular Prefixes are

outlined here. Router vendors are free to implement whatever approaches they want.

1. Policy-based: The simplest approach for network administrators is to have broad policies that routers use to determine which sub-prefixes are designated as popular. An obvious policy would be a "customer routes" policy, whereby all customer routes are installed (as identified for instance by appropriate community attribute tags). Another policy would be for a router to install prefixes originated by specific ASes. For instance, two ISPs could mutually agree to install each other's originated prefixes. A third policy might be to install prefixes with the shortest AS-PATH.
2. Static list: Another approach would be to configure static lists of specific prefixes to install. For instance, prefixes associated with an SLA might be configured. Or, a list of prefixes for the most popular websites might be installed.
3. High-volume prefixes: By installing high-volume prefixes as Popular Prefixes, the latency and load associated with the longer path required by VA is minimized. One approach would be for an ISP to measure its traffic volume over time (days or a few weeks), and statically configure high-volume prefixes as Popular Prefixes. There is strong evidence that prefixes that are high-volume tend to remain high-volume over multi-day or multi-week timeframes (though not necessarily at short timeframes like minutes or seconds). High-volume prefixes MAY also be installed automatically. For this, a router measures its own traffic volumes, and installs and removes Popular Prefixes in response to changes in traffic load. The downside of this approach is that it complicates debugging network problems. If packets are being dropped somewhere in the network, it is more difficult to find out where if the selected path can change dynamically.

3.6. New Configuration

VA places new configuration requirements on ISP administrators. Namely, the administrator does the following.

1. Select VPs, and configure the VP-List into all VA routers. As a general rule, having a larger number of relatively small prefixes gives administrators the most flexibility in terms of filling available FIB with sub-prefixes, and in terms of balancing load across routers. Once an administrator has selected a VP-List, it is just as easy to configure routers with a large list as a small list. A good list might be one where the number of VPs is relatively large, say 100 or so (noting again that each VP must be smaller than a real prefix), and the number of sub-prefixes within each VP is roughly the same.

2. Select and configure APRs. There are three primary considerations here. First, there needs to be enough APRs to failover to should one or more APRs crash. Second, APR assignment should not result in router overload. Third, excessively long paths should be avoided. Ideally there should be two APRs for each VP within each PoP, but this may not be possible for small PoPs. Failing this, there should be at least two APRs in each geographical region, so as to minimize path length increase. Routers should have the appropriate counters to allow administrators to know the volume of APR traffic each router is handling so as to adjust load by adding or removing APR assignments.
3. Select and configure Popular Prefixes or Popular Prefix policies. There are two general goals here. The first is to minimize load overall by minimizing the number of packets that take longer paths. The second is to insure that specific selected prefixes don't have overly long paths. These goals have to be weighed against the administrative overhead of configuring potentially thousands of Popular Prefixes. As one example a small ISP may wish to keep it simple by doing nothing more than indicating that customer routes should be installed. In this case, the administrator could otherwise assign as many APRs as possible while leaving enough FIB space for customer routes. As another example, a large ISP could build a management system that takes into consideration the traffic matrix, customer SLAs, robustness requirements, FIB sizes, topology, and router capacity, and periodically automatically computes APR and Popular Prefix assignments.

3.7. Interaction with Traffic Engineering

In VA, some traffic traverses an APR as an intermediate "hop", and some does not. For that traffic that does not, there is no difference between how that traffic is handled and how it is handled in a non-VA network with edge-to-edge tunnels. As a result, there should be no difference in how traffic engineering operates on that traffic.

For traffic that does traverse APR "hop", the following holds: Any traffic engineering decisions that affect the BGP NEXT_HOP must be made at the APR. Traffic engineering decisions that effects the router path through the AS may be handled in one of two ways. First, the path decision may simply be made twice independently, once for the ingress-to-APR tunnel, and once for the APR-to-egress tunnel. This approach requires no changes to the traffic engineering mechanisms per se, but it may not make optimal path selection decisions. Second, the traffic engineering decision may take into account both tunnels, even to the point of choosing among multiple

transit APRs. This approach may be more optimal, but is more complex and requires changes to existing mechanisms.

Overall, if the majority of traffic does not involve an APR "hop", for instance through the use of popular prefixes, then VA in any event has a minimal impact on traffic engineering, and so the impact of VA may potentially be ignored.

4. Usage of MPLS Tunnels

VA utilizes a straight-forward application of MPLS. The tunnels are MPLS Label Switched Paths (LSP), and are signaled using either the Label Distribution Protocol (LDP) [RFC5036] or RSVP-TE [RFC3209]. Both VA and legacy routers MUST participate in this signaling.

APRs and ASBRs initiate tunnels. In both cases, Downstream Unsolicited tunnels are initiated to all IGP neighbors with the full BGP NEXT_HOP address as the Forwarding Equivalence Class (FEC). In the case of APRs, the BGP NEXT_HOP is the APR's own address. In the case of legacy ASBRs, the BGP NEXT_HOP is the ASBR's own address. In the case of VA ASBRs, the BGP NEXT_HOP is that of the remote ASBR.

Existing Penultimate Hop Popping (PHP) mechanisms in the data plane can be used for forwarding packets to remote ASBRs.

4.1. Usage of Inner Label

Besides using a separate LSP to identify the remote ASBR as described above, it is also possible to use an inner label to identify the remote ASBR. Either an outer label or an IP tunnel identifies the local ASBR.

When a local ASBR advertises a route into iBGP, it sets the NEXT_HOP to itself, and assigns a label to the route. This label is used as the inner label, and identifies the remote ASBR from which the route was received [RFC3107].

The presence of the inner label in the iBGP update acts as the signal to the receiving router that an inner label MUST be used in packets tunneled to the NEXT_HOP address. If there is an LSP established targeted to the NEXT_HOP address, then it is used to tunnel the packet to the NEXT_HOP address. Otherwise, an IP header address to the NEXT_HOP address is used.

5. IANA Considerations

There are no IANA considerations.

6. Security Considerations

We consider the security implications of VA under two scenarios, one where VA is assumed to be configured and operated correctly, and one where it is mis-configured. A cornerstone of VA operation is that the basic behavior of BGP doesn't change, especially inter-domain. Among other things, this makes it easier to reason about security.

6.1. Properly Configured VA

If VA is configured and operated properly, then the external behavior of an AS does not change. The same upstream ASes are selected, and the same prefixes and AS-PATHs are advertised. Therefore, a properly configured VA domain has no security impact on other domains.

If another ISP starts advertising a prefix that is larger than a given VP, this prefix will be ignored by APRs that have a VP that falls within the larger prefix (Section 3.2.3). As a result, packets that might otherwise have been routed to the new larger prefix will be dropped at the APRs. Note that the trend in the Internet is towards large prefixes being broken up into smaller ones, not the reverse. Therefore, such a larger prefix is likely to be invalid. If it is determined without a doubt that the larger prefix is valid, then the ISP will have to reconfigure its VPs.

VA does not change an ISP's ability to do ingress filtering using strict uRPF (Section 3.5).

Regarding DoS attacks, there are two issues that need to be considered. First, does VA result in new types of DoS attacks? Second, does VA make it more difficult to deploy DoS defense systems. Regarding the first issue, one possibility is that an attacker targets a given router by flooding the network with traffic to prefixes that are not popular, and for which that router is an APR. This would cause a disproportionate amount of traffic to be forwarded to the APR(s). While it is up to individual ISPs to decide if this attack is a concern, it does not strike the authors that this attack is likely to significantly worsen the DoS problem.

Many DoS defense systems use dynamically established Routing Table entries to divert victims' traffic into LSPs that carry the traffic to scrubbers. This mechanism works with VA---it simply over-rides whatever route is in place. This mechanism works equally well with

APRs and non-APRs.

6.2. Mis-configured VA

VA introduces the possibility that a VP is advertised outside of an AS. This in fact should be a low probability event since routers filter these, but it is considered here none-the-less.

If an AS leaks a large VP (i.e. larger than any real prefixes), then the impact is minimal. Smaller prefixes will be preferred because of best-match semantics, and so the only impact is that packets that otherwise have no matching routes will be sent to the misbehaving AS and dropped there. If an AS leaks a small VP (i.e. smaller than a real prefix), then packets to that AS will be hijacked by the misbehaving AS and dropped. (This can happen with or without VA, and so doesn't represent a new security problem per se.)

Although VPs MUST be larger than real prefixes, there is intentionally no mechanism designed to automatically insure that this is the case. Such a mechanisms would be dangerous. For instance, if an ISP somewhere advertised a very large prefix (a /4, say), then this would cause APRs to throw out all VPs that are smaller than this. For this reason, VPs must be set through configuration only.

7. Acknowledgements

The authors would like to acknowledge the efforts of Xinyang Zhang and Jia Wang, who worked on CRIO (Core Router Integrated Overlay), an early inter-domain variant of FIB suppression, and the efforts of Hitesh Ballani and Tuan Cao, who worked on the configuration-only variant of VA that works with legacy routers. We would also like to thank Scott Brim, Daniel Ginsburg, and Rajiv Asati for their helpful comments. In particular, Daniel's comments significantly simplified the spec (eliminating the need for a new Extended Communities Attribute). Finally, we would like to thank Wes George, Med Boucadair, and Bruno Decraene for their reviews and suggestions.

8. References

8.1. 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.

[RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

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

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

8.2. Informative References

[I-D.ietf-grow-simple-va]
Francis, P., Xu, X., Ballani, H., Raszuk, R., and L. Zhang, "Simple Virtual Aggregation (S-VA)", draft-ietf-grow-simple-va-00 (work in progress), March 2010.

[I-D.ietf-grow-va-gre]
Francis, P., Raszuk, R., and X. Xu, "GRE and IP-in-IP Tunnels for Virtual Aggregation", draft-ietf-grow-va-gre-00 (work in progress), July 2009.

[I-D.ietf-grow-va-mpls]
Francis, P. and X. Xu, "MPLS Tunnels for Virtual Aggregation", draft-ietf-grow-va-mpls-00 (work in progress), May 2009.

[I-D.ietf-grow-va-mpls-innerlabel]
Xu, X. and P. Francis, "Proposal to use an inner MPLS label to identify the remote ASBR VA", draft-ietf-grow-va-mpls-innerlabel-00 (work in progress), September 2009.

[nsdi09] Ballani, H., Francis, P., Cao, T., and J. Wang, "Making Routers Last Longer with ViAggre", ACM Usenix NSDI 2009 ht tp://www.usenix.org/events/nsdi09/tech/full_papers/

ballani/ballani.pdf, April 2009.

Authors' Addresses

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 (editor)
NTT MCL Inc.
101 S Ellsworth Avenue Suite 350
San Mateo, CA 94401
US

Email: robert@raszuk.net

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

Phone:
Email: lixia@cs.ucla.edu