Network Working Group	P. Francis	
Internet-Draft	MPI-SWS	
Intended status: Informational	X. Xu	
Expires: November 29, 2009	Huawei	
	H. Ballani	
	Cornell U.	
	D. Jen	
	UCLA	
	R. Raszuk	
	Self	
	L. Zhang	
	UCLA	
	May 28, 2009	

TOC

FIB Suppression with Virtual Aggregation
draft-ietf-grow-va-00.txt

Status of this Memo

Copyright Notice

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 deployed autonomously by an ISP (cooperation between ISPs is not required), and can co-exist with legacy routers in the ISP.

Table of Contents

---

1. Introduction

TOC

ISPs today manage constant DFRT growth in a number of ways. Most commonly, ISPs will upgrade their router hardware before DFRT growth outstrips the size of the FIB. In cases where an ISP wants to continue to use routers whose FIBs are not large enough, it may deploy them at edge locations where a full DFRT is not needed, for instance at the customer interface. Packets for which there is no route are defaulted to a "core" infrastructure that does contain the full DFRT. While this helps, it cannot be used for all edge routers, for instance those that interface with other ISPs. Alternatively, some lower-tier ISPs may simply ignore some routes, for instance /24's that fall within the aggregate of another route.

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, for instance by not committing them to the Routing Table. 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 neighbor 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] (Ballani, H., Francis, P., Cao, T., and J. Wang, “Making Routers Last Longer with ViAggre,” April 2009.). The VPs can be organized such that all routers in an ISP see FIB size decrease, or in such a way that "core" routers keep the full FIB, and "edge" routers have almost no FIB (i.e. by defining a VP of 0/0).

TOC

1.1. Scope of this Document

The scope of this document is limited to Intra-domain VA operation. In other words, the case where a single ISP autonomously operates VA internally without any coordination with neighboring ISPs.
Note that this document assumes that the VA "domain" (i.e. the unit of autonomy) is the AS (that is, different ASes run 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.
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, however, legacy routers must be able to terminate tunnels.
This document is agnostic about what type of tunnel may be used for VA, and does not specify a tunnel type per se. Rather, it refers generically to tunnels and specifies the minimum set of requirements that a given tunnel type must satisfy. Separate documents are used to specify the operation of VA for specific tunnel types.

1.2. 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.).

1.3. Terminology

TOC

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 (or, equivalently, the Routing Table). In other words, the phrase "install a route" means "install a route into the FIB", and the phrase "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. (If a Legacy router cannot terminate tunnels, then any routes that are reached via that router must be installed in all FIBs.)

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 Table: The term Routing Table is defined here the same way as in Section 3.2 of [RFC4271] (Rekhter, Y., Li, T., and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” January 2006.): "Routing information that the BGP speaker uses to forward packets (or to construct the forwarding table used for packet forwarding) is maintained in the Routing Table." As such, FIB Suppression can be achieved by not installing a route into the Routing Table

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] (Rekhter, Y., Li, T., and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” January 2006.)), 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 they hold today.

Tunnel: VA can use a variety of tunnel types: MPLS LSPs, IP-in-IP, GRE, L2TP, and so on. This document does not describe how any given tunnel information is conveyed: that is left for companion documents. This document uses the term tunnel to refer to any appropriate tunnel type.

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). A VP is not physically aggregatable, and so it is aggregated at APRs through the use of tunnels.

VP-List: A list of defines VPs. All routers must agree on the contents of this list (which is statically configured into every VA router).

1.4. Temporary Sections

TOC

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

1.4.1. Document revisions

TOC

This document was previously published as both draft-francis-idr-intra-va-01.txt and draft-francis-intra-va-01.txt.

1.4.1.1. Revisions from the 00 version (of draft-francis-intra-va-00.txt)

TOC

Added additional authors (Jen, Raszuk, Zhang), to reflect primary contributors moving forwards. In addition, a number of minor clarifications were made.

1.4.1.2. Revisions from the 01 version (of draft-francis-idr-intra-va-01.txt)

TOC

1. Changed file name from draft-francis-idr-intra-va to draft-francis-intra-va.

2. Restructured the document to make the edge suppression mode a specific sub-case of VA rather than a separate mode of operation. This includes modifying the title of the draft.

3. Removed MPLS tunneling details so that specific tunneling approaches can be described in separate documents.

1.4.1.3. Revisions from 00 version

TOC

*Changed intended document type from STD to BCP, as per advice from Dublin IDR meeting.

*Cleaned up the MPLS language, and specified that the full-address routes to remote ASBRs must be imported into OSPF (Section 3.2.3 (Border VA Routers)). As per Daniel Ginsburg's email http://www.ietf.org/mail-archive/web/idr/current/msg02933.html.

*Clarified that legacy routers must run MPLS. As per Daniel Ginsburg's email http://www.ietf.org/mail-archive/web/idr/current/msg02935.html.

*Fixed LOCAL_PREF bug. As per Daniel Ginsburg's email http://www.ietf.org/mail-archive/web/idr/current/msg02940.html.

*Removed the need for the extended communities attribute on VP routes, and added the requirement that all VA routers be statically configured with the complete list of VPs. As per Daniel Ginsburg's emails http://www.ietf.org/mail-archive/web/idr/current/msg02940.html and http://www.ietf.org/mail-archive/web/idr/current/msg02958.html. In addition, the procedure for adding, deleting, splitting, and merging VPs was added. As part of this, the possibility of having overlapping VPs was added.

*Added the special case of a core-edge topology with default routes to the edge as suggested by Robert Raszuk in email http://www.ietf.org/mail-archive/web/idr/current/msg02948.html. Note that this altered the structure and even title of the document.

*Clarified that FIB suppression can be achieved by not loading entries into the Routing Table, as suggested by Rajiv Asati in email http://www.ietf.org/mail-archive/web/idr/current/msg03019.html.

---

2. Overview of Virtual Aggregation (VA)

TOC

For descriptive simplicity, this section starts by describing VA assuming that there are no legacy routers in the domain. Section 2.1 (Mix of legacy and VA routers) 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 quite simple. 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 (commonly seen today). 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 (possibly via a tunnel), and then from the APR to the BGP NEXT_HOP router via a tunnel.

Normally the BGP NEXT_HOP is the remote ASBR. In this case, even though the remote ASBR is the tunnel endpoint, the tunnel header is 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. The exception to this is legacy local ASBR routers. In this case, the legacy router is the BGP NEXT_HOP, and packets are tunneled to the legacy router, which then uses a FIB lookup to deliver the packet to the appropriate remote ASBR. This applies only to legacy routers that can convey tunnel parameters and detunnel packets.

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.

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 (i.e. customer routes are always installed), or for sub-prefixes that carry a high volume of traffic (Section 3.2.5.1 (Selecting Popular Prefixes)). 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, 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 (which of course is the situation in the absence of VA). 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.

TOC

2.1. Mix of legacy and VA routers

It is important that an ISP be able to operate with a mix of "VA routers" (routers upgraded to operate VA as described in the document) 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 (and VA routers for that matter) are able to forward already-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. (This is listed as Requirement R5 in the companion tunneling documents.) Depending on the tunnel type, legacy routers may also be able to generate tunneled packets, though this is an optional requirement. (This is listed as Requirement R4 in the companion tunneling documents.) Legacy routers must use their own address as the BGP NEXT_HOP, and must FIB-install routes for which they are the BGP NEXT_HOP.

2.2. Summary of Tunnels and Paths

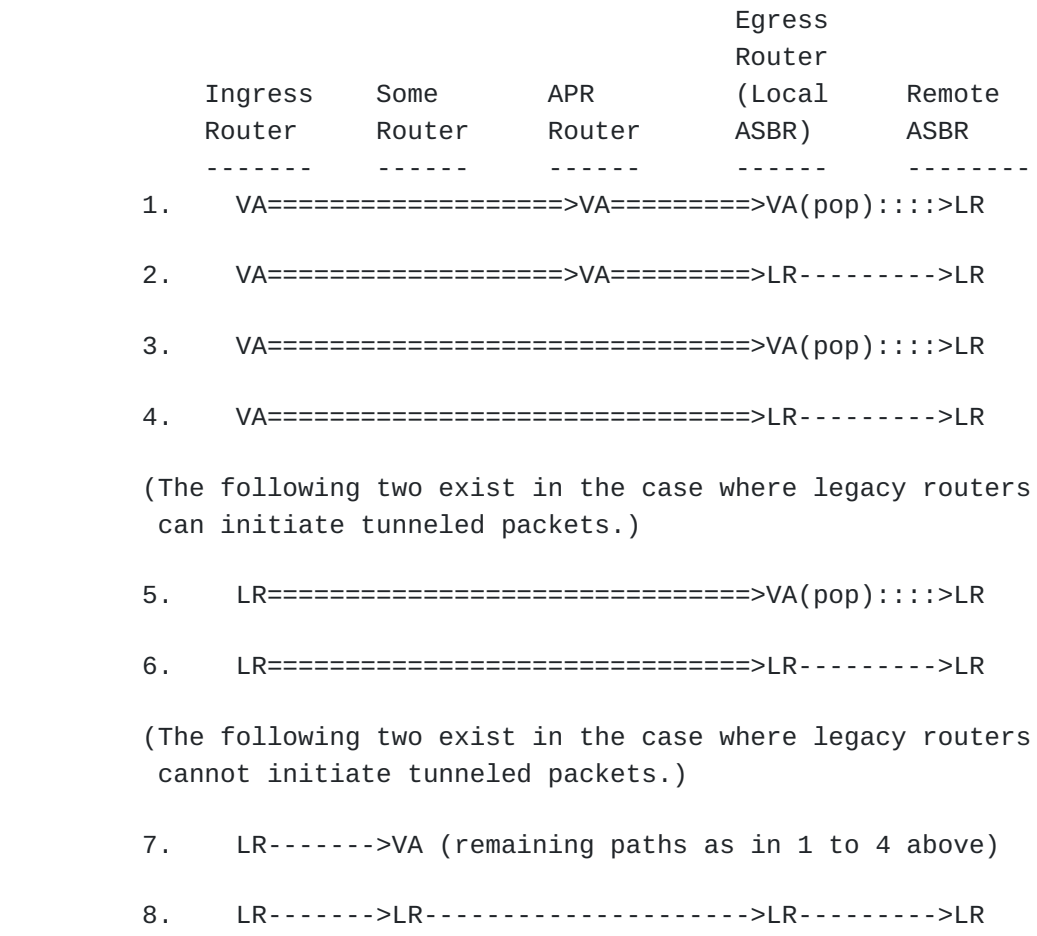
TOC

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 legacy router, or the remote ASBR neighbor of a VA router). Note that this is listed as Requirement R3 in the companion tunneling documents.

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. (Note that the remote ASBR may actually be a legacy router or a VA router---it doesn't matter (and isn't known) to the ISP.)



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, 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.) 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. The trick is avoiding 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

TOC

This section describes in detail how to operate VA. It starts with a brief discussion of requirements, followed by a specification of router support for VA.

3.1. Requirements for VA

TOC

While the core requirement is of course to be able to manage FIB size, this must be done in a way that:

- *is robust to router failure,
- *allows for traffic engineering,
- *allows for existing inter-domain routing policies,
- *operates in a predictable manner and is therefore possible to test, debug, and reason about performance (i.e. establish SLAs),
- *can be safely installed, tested, and started up,
- *Can be configured and reconfigured without service interruption,
- *can be incrementally deployed, and in particular can be operated in an AS with a mix of VA-capable and legacy routers,
- *accommodates existing security mechanisms such as ingress filtering and DoS defense,
- *does not introduce significant new security vulnerabilities.

In short, operation of VA must not significantly affect the way ISPs operate their networks today. Section 3.3 (Requirements Discussion) discusses the extent to which these requirements are met by the design presented in Section 3.2 (VA Operation).

3.2. VA Operation

TOC

In this section, the detailed operation of VA is specified.

3.2.1. Legacy Routers

TOC

VA can operate with a mix of VA and legacy routers. To avoid the types of loops described in Section 2.2 (Summary of Tunnels and Paths), 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. This is listed as "Requirement R1" in the companion tunneling documents. 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, of course, 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 (Summary of Tunnels and Paths)).

3.2.2. Advertising and Handling Virtual Prefixes (VP)

TOC

3.2.2.1. Distinguishing VP's from Sub-prefixes

TOC

VA routers must be able to distinguish VP's from sub-prefixes. This is primarily in order to know which routes to install. In particular, non-APR routers must 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 FIB with routes that they ultimately don't need to install (Section 3.2.5 (Suppressing FIB Sub-prefix Routes)). This leads to the following requirement:

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

3.2.2.2. Limitations on Virtual Prefixes

TOC

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 Aggregation Point Router (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 subtly different from those of real prefixes (well, maybe not so subtly). 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, it is important that VPs not be advertised across AS boundaries.

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 5 (Security Considerations)).

(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 APR's 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 the FIB size in any router. (Section 3.2.2.5 (Adding and deleting VP's)).

3.2.2.3. Aggregation Point Routers (APR)

TOC

Any router may be configured as an Aggregation Point Router (APR) for one or more Virtual Prefixes (VP). 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 [RFC4271] (Rekhter, Y., Li, T., and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” January 2006.). 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] (Chandrasekeran, R., Traina, P., and T. Li, “BGP Communities Attribute,” August 1996.) 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 (Requirement R1).

4. If a packet is received at the APR whose best match 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 3.2.2.2 (Limitations on Virtual Prefixes)). 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.

3.2.2.3.1. Selecting APRs

TOC

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.
APRs may be (and probably should be) statically assigned. They may also, however, be dynamically assigned, for instance in response to APR failure. For instance, each router may be assigned as a backup APR for some other APR. If the other APR crashes (as indicated by the withdrawal of its routes to its VPs), the backup APR can install the appropriate sub-prefixes and advertise the VP as specified above. Note that doing so may require it to first remove some popular prefixes from its FIB to make room.
Note that, 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 static configuration only.

3.2.2.4. Non-APR Routers

TOC

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

1. Routes to VPs (identifiable using the VP-List).

2. Routes to the largest of any prefixes that contain a given VP. (Note that although this is not supposed to happen, if it does the non-APR should install it, with the effect that any addresses in the prefix not covered by VPs will be routed outside the domain.)

3. Routes to all prefixes that contain an address that is in part of the address space for which no VP is defined (i.e. as is done today without VA).

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. Note that it is strongly recommended that routers keep at least two VP routes in their RIB at all times. The main reason is that if the currently used VP route is withdrawn, the second VP route can be immediately installed, and the issue of whether to temporarily install sub-prefixes in the FIB is avoided (Section 3.2.5 (Suppressing FIB Sub-prefix Routes)). Another reason is that the IGP

can be used to even more quickly detect that the APR has crashed, again allowing the second VP route to be immediately installed.

3.2.2.5. Adding and deleting VP's

TOC

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 size either before or after the change).
Adding a VP is straightforward. The first step 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 3.2.2.4 (Non-APR Routers) 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.

3.2.3. Border VA Routers

TOC

VA routers that are border routers MUST do the following: When forwarding externally-received routes over iBGP, the BGP NEXT_HOP attribute MUST be set to the remote ASBR. They MUST establish tunnels that have the following properties (Requirement R2 in companion documents):

1. The tunnel target must be the remote ASBR BGP NEXT_HOP address. In other words, the target address used by other routers in the domain for tunneling packets is the remote ASBR address.

2. The border router must detunnel the packet before forwarding the packet to the remote ASBR. In other words, the remote ASBR receives a normal untunneled packet identical to the packet it would receive without VA.

3. The border router 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.

Note that there are a number of ways the above tunnel can be created, as documented separately. For instance, the tag on an MPLS LSP could identify the remote ASBR, and the border router could use what is effectively penultimate hop popping to deliver the packet. Or, GRE could be used whereby the outer IP header addresses the border router, and the GRE key value identifies the remote ASBR.

3.2.4. Advertising and Handling Sub-Prefixes

TOC

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 3.2.5 (Suppressing FIB Sub-prefix Routes).

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 3.2.3 (Border VA Routers).

3.2.5. Suppressing FIB Sub-prefix Routes

TOC

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 never 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. If there is an alternate route to the sub-prefix for which there is a tunnel, then that route should be selected, even if it is less attractive according to the normal BGP best path selection algorithm.

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 buildup.

4. If a non-APR router does not have a route for a known VP, then it MAY or MAY NOT install sub-prefixes within that VP. Whether or not it does is up to the vendor and the network operator. One approach is to never install such sub-prefixes, on the assumption that the network operator will engineer his network so that this rarely if ever happens.

5. Another approach is to have routers install such sub-prefixes, but taking care not to do so if the missing VP route is a transient condition. For instance, if the router is booting up, and simply has not yet received all of its routes, then it can reasonably expect to receive a VP route soon and so SHOULD NOT install the sub-prefixes. On the other hand, if a continuously operating router had only a single remaining route for the VP, and that route is withdrawn, then the router might not expect to receive a replacement VP route soon and so SHOULD install the sub-prefixes. Obviously a router can't predict the future

with certainty, so the following algorithm might be a useful way to manage whether or not to install sub-prefixes for a non-existing VP route:

*Define a timer MISSING_VP_TIMER, set for a relatively short time (say 10 seconds or so).

*Start the timer when either: 1) the last VP route is withdrawn, or 2) there are initially neither VP routes nor sub-prefix routes, and the first sub-prefix route is received.

*When the timer expires, install sub-prefix routes. Note, however, that optional routes may first need to be removed from the FIB to make room for the new sub-prefix routes. If even after removing optional routes there is no room in the FIB for sub-prefix routes, then they should remain suppressed. In other words, sub-prefix entries required by virtue of being an APR take priority over sub-prefix entries required by virtue of not having a VP route.

6. All other sub-prefix routes MAY be suppressed. Such "optional" sub-prefixes that are nevertheless installed are referred to as popular prefixes.

3.2.5.1. Selecting Popular Prefixes

TOC

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 APR's for as many VP's as possible, or by assigning relatively few APR's 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 dynamically. In other words, a router measures its own traffic volumes, and installs and removes popular prefixes in response to short term 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.

3.2.6. Core-Edge Operation

TOC

A common style of router deployment in ISPs is the "core-edge" deployment, whereby there is a core of high-capacity routers surrounded by potentially lower-capacity "edge" routers that may not carry the whole DFRT, and which default route to a core router. VA can support this style of configuration be effectively defining a single VP as 0/0, and by defining core routers to be APRs for 0/0. This results in core routers maintaining full FIBs, and edge routers having potentially extremely small FIBs. The advantage of using VA to support core-edge topologies is that, with VA, any edge router, including those peering with other ISPs, can have a small FIB. Today such routers must maintain the full DFRT in order to peer.
Vendors may wish to facilitate configuration of a core-edge style of VA for its customers that already use a core-edge topology. In other words, a vendor may wish to simplify the VA configuration task so that a customer merely needs to configure which of its routers are core and which are edge, and the appropriate VA configuration, i.e. the VP-List, tunnels, and popular prefixes, is automatically done "under the hood" so to speak. Note that, under a core-edge configuration, it isn't

strictly speaking necessary for core routers to advertise the 0/0 VP within BGP. Rather, edge routers could rely on their default route to a core router.

3.3. Requirements Discussion

TOC

This section describes the extent to which VA satisfies the list of requirements given in Section 3.1 (Requirements for VA).

3.3.1. Response to router failure

TOC

VA introduces a new failure mode in the form of Aggregation Point Router (APR) failure. There are two basic approaches to protecting against APR failure, static APR redundancy, and dynamic APR assignment (see Section 3.2.2.3.1 (Selecting APRs)). In static APR redundancy, enough APRs are assigned for each Virtual Prefix (VP) so that if one goes down, there are others to absorb its load. Failover to a static redundant APR is automatic with existing BGP mechanisms. If an APR crashes, BGP will cause packets to be routed to the next nearest APR. Nevertheless, there are three concerns here: convergence time, load increase at the redundant APR, and latency increase for diverted flows. Regarding convergence time, note that, while fast-reroute mechanisms apply to the rerouting of packets to a given APR or egress router, they don't apply to APR failure. Convergence time was discussed in Section 3.2.2.4 (Non-APR Routers), which suggested that it is likely that BGP convergence times will be adequate, and if not the IGP mechanisms may be used.

Regarding load increase, in general this is relatively small. This is because substantial reductions in FIB size can be achieved with almost negligible increase in load. For instance, [nsdi09] (Ballani, H., Francis, P., Cao, T., and J. Wang, “Making Routers Last Longer with ViAggre,” April 2009.) shows that a 5x reduction in FIB size yields a less than one percent increase in load overall. Given this, depending on the configuration of redundant APRs, failure of one APR increases the load of its backups by only a few percent. This is well within the variation seen in normal traffic loads.

Regarding latency increase, some flows may see a significant increase in delay (and, specifically, an increase that puts it outside of its SLA boundary). Normally a redundant APR would be placed within the same POP, and so increased latency would be minimal (assuming that load is also quite small, and so there is no significant queuing delay). It is not always possible, however, to have an APR for every VP within every POP, much less a redundant APR within every POP, and so sometimes

failure of an APR will result in significant latency increases for a small fraction of traffic.

3.3.2. Traffic Engineering

TOC

VA complicates traffic engineering because the placement of APRs and selection of popular prefixes influences how packets flow. (Though to repeat, increased load is in any event likely to be minimal, and so the effect on traffic engineering should not be great in any event.) Since the majority of packets may be forwarded by popular prefixes (and therefore follow the shortest path), it is particularly important that popular prefixes be selected appropriately. As discussed in Section 3.2.5.1 (Selecting Popular Prefixes), there are static and dynamic approaches to this. [nsdi09] (Ballani, H., Francis, P., Cao, T., and J. Wang, "Making Routers Last Longer with ViAggre," April 2009.) shows that high-volume prefixes tend to stay high-volume for many days, and so a static strategy is probably adequate. VA can operate correctly using either RSVP-TE [RFC3209] (Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels," December 2001.) or LDP to establish tunnels.

3.3.3. Incremental and safe deploy and start-up

TOC

It must be possible to install and configure VA in a safe and incremental fashion, as well as start it up when routers reboot. This document allows for a mixture of VA and legacy routers, allows a fraction or all of the address space to fall within virtual prefixes, and allows different routers to suppress different FIB entries (including none at all). As a result, it is generally possible to deploy and test VA in an incremental fashion.

3.3.4. VA security

TOC

Regarding ingress filtering, because in VA the RIB is effectively unchanged, routers contain the same information they have today for installing ingress filters [RFC2827] (Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing," May 2000.). Presumably, installing an ingress filter in the FIB takes up some memory space. Since ingress filtering is most effective at the "edge" of the network (i.e. at the

customer interface), the number of FIB entries for ingress filtering should remain relatively small---equal to the number of prefixes owned by the customer. Whether this is true in all cases remains for further study.

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.

Regarding DoS defense system deployment, more input about specific systems is needed. It is the authors' understanding, however, that at least some of these systems use dynamically established Routing Table entries to divert victims' traffic into LSPs that carry the traffic to scrubbers. The expectation is that this mechanism simply over-rides whatever route is in place (with or without VA), and so the operation of VA should not limit the deployment of these types of DoS defense systems. Nevertheless, more study is needed here.

3.4. New Configuration

TOC

VA places new configuration requirements on ISP administrators. Namely, the administrator must:

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. We can expect network operator groups like NANOG to compile good VP-Lists that ISPs can then adopt. A good list would 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 must be enough APRs to handle reasonable APR failure scenarios. Second, APR assignment should not result in router overload. Third, particularly 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 must 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.

---

4. IANA Considerations

TOC

There are no IANA considerations.

---

5. Security Considerations

TOC

We consider the security implications of VA under two scenarios, one where VA is 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.

TOC

5.1. 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.
This document discusses intra-domain security concerns in Section 3.3.4 (VA security) which argues that any new security concerns appear to be relatively minor.
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 3.2.2.3 (Aggregation Point Routers (APR)). 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.

5.2. Mis-configured VA

TOC

VA introduces the possibility that a VP is advertised outside of an AS. This in fact should be a low probability event, 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.

---

6. Acknowledgements

TOC

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 External Communities Attribute).

7. References

TOC

7.1. Normative References

TOC

[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).
[RFC2328]	Moy, J., “OSPF Version 2,” STD 54, RFC 2328, April 1998 (TXT, XML).
[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 (TXT).
[RFC3107]	Rekhter, Y. and E. Rosen, “Carrying Label Information in BGP-4,” RFC 3107, May 2001 (TXT).
[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 (TXT).
[RFC4271]	Rekhter, Y., Li, T., and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” RFC 4271, January 2006 (TXT).

7.2. Informative References

TOC

[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