Routing Area Working Group
Internet-Draft
Intended status: Informational
Expires: March 23, 2013

R. White
Verisign
S. Hares
Huawei Technologies (USA)
R. Fernando
Cisco Systems

September 19, 2012

Use Cases for an Interface to the Routing System
draft-white-irs-use-case-00

Abstract

Programmatic interfaces to provide control over individual forwarding devices in a network promise to reduce operational costs while improving scaling, control, and visibility into the operation of large scale networks. To this end, several programmatic interfaces have been proposed. OpenFlow, for instance, provides a mechanism to replace the dynamic control plane processes on individual forwarding devices throughout a network with off box processes that interact with the forwarding tables on each device. Another example is NETCONF, which provides a fast and flexible mechanism to interact with device configuration and policy.

There is, however, no proposal which provides an interface to all aspects of the routing systemas a system. Such a system would not interact with the forwarding system on individual devices, but rather with the control plane processes already used to discover the best path to any given destination through the network, as well as interact with the routing information base (RIB), which feeds the forwarding table the information needed to actually switch traffic at a local level.

This document describes a set of use cases such a system could fulfill. It is designed to provide underlying support for the framework, policy, and other drafts describing the Interface to the Routing System (IRS).

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

This Internet-Draft will expire on March 23, 2013.

Copyright Notice

Table of Contents

1. Introduction

The Interface to the Routing System Framework [IRS] desribes a mechanism where the distributed control plane can be augmented by an outside control plane through an open, accessible interface, including the Routing Information Base (RIB), in individual devices. This represents a "halfway point" beteween completely replacing the traditional distributed control plane and directly configuring devices to distribute policy or modifications to routing through off-board processes. This draft proposes a set of use cases that explain where the work described in [IRS] will be useful. The goal is to inform not only the community's understanding of where IRS fits in the larger scheme of SDN proposals, but also to inform the requirements, framework, and specification of IRS to provide the best fit for the purposes which make the most sense for this type of programmatic interface.

Towards this end the authors have searched for a number of different use cases representing not only complex modifications of the control plane, including interaction with applications and network conditions, but also simpler use cases. The array of use cases presented here should provide the reader with a solid understanding of the power of an SDN solution that will augment, rather than replace, traditional distributed control planes.

Each use case is presented in its own section.

2. Optimized Exit Control

At edges where traffic exits along two or more possible paths, it is often desirable to choose a path based on more information the dynamic control plane provides. For instance, a network operator may want to take into account factos such as:

- Cost per unit of data sent, indluding time of day variations, surcharges over a specific amount of data transmitted, and surcharges for transmitting data to specific types of destinations.

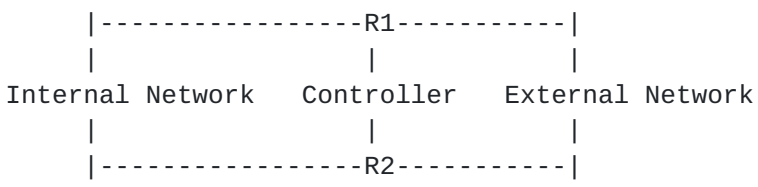
- Urgency of data traffic or flow.

- Exit point performance, including historical jitter, delay, and available bandwidth, possibly on a per destination basis.

- Availability of a specific destination through a given link at the per destination basis (more specific than the routing protocol provides).

A number of possible solutions have been proposed or deployed in the past. For instance, the necessary metrics could be added to [BGP], or any other routing protocol, to provide the necessary information, and fine-tuned algorithms could be developed and deployed. Massive changes to well known and understood distributed control plane protocols to resolve a single use case, however, are not likely to be productive for the community as a whole. It's often difficult to justify the added complexity in the database and algorithms of routing protools to solve what is considered a point case.

Another alternative has been the development of specific appliances designed to monitor the information necessary to provide an optimal edge decision, and then to use some automated configuration mechanism to transmit the decision to the edge routers. An example is illustrated in the figure below.



The controller in this network must:

- Discover the topology of the network from R1 and R2.
- Compare the current traffic flow information to policies set administratively by the network operator.
- Monitor the flow of traffic from the perspective of R1 and R2.
- Inject forwarding information to directly impact the traffic flow at the edge devices, or modify the policy of the existing distributed (dynamic) control plane already running in the network.

Many of these steps is challenging for currently available solutions.

To discover the topology at the edge rotuers, the controllers can either participate in the control plane, or walk the local routing table using a network management protocol. Neither of these options are optimal in this case because the controlling process cannot interact dynamically with the local topology information in near real time through such mechanisms.

Injecting forwarding information directly into the RIB on the individual devices in this network is possible today through the configuration of static routes through some external mechanism, such

as SNMP, NETCONF, or by direct external interaction with the devices' CLI. None of these options are attractive because:

- They modify the actual configuration of the device (unlike a dynamic routing process).

- They are too persistent (routes installed through static configuration persist across device reboots).

- The controller cannot interact with the routing table in parallel with other routing processes. For instance, when a routing process attempts to install a new route in the routing table, there is often a callback or other notification to the other routing processes running on the same device; this notification provides important information the controller can take into account in its view of the current state of the routing table, and the state of the device's routing table. Interface level events also often trigger notifications from the RIB to local routing processes; these notifications would be invaluble for the controller to modify injected routing state in reaction to network topology events.

- Routes installed through the an off box controller through the CLI or XML interface are difficult to redistribute into other protocols to draw traffic to a specific exit point, and it can be difficult to fine tune how these injected routes interact with routes learned through other routing processes.

IRS can resolve these issues by providing an open interface to the local RIB on each device, allowing the controller to interact with the RIB just as a local routing process would. This would allow the controlling process to see the topology information in the RIB dynamically, receiving near real time updates for route removals, installs, and other events, and without relying on static configuration to inject forwarding information each device can use.

Summary of IRS Capabilities and Interactions:

- IRS should provide the ability to read the local RIB of each forwarding device, including the destination prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), the metric of each installed route, a route preference, and an identifier indicating the installing process.

- The ability to monitor the available routes installed in the RIB of each forwarding device, including near real time notification of route installation and removal. This information must include the destination prefix (NLRI), a table identifier (if the

forwarding device has multiple forwarding instances), the metric of the installed route, and an identifier indicating the installing process.

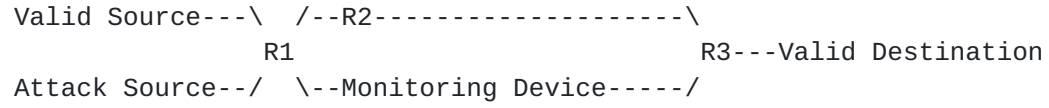
- The ability to install destination based routes in the local RIB of each forwarding device. This must include the ability to supply the destination prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), a route preference, a route metric, a next hop, an outbound interface, and a route process identifier.

- The ability to interact with various policies configured on the forwarding devices, in order to inform the policies implemented by the dynamic routing processes. This interaction SHOULD be through existing configuration mechanisms, such as NETCONF, and SHOULD be recorded in the configuration of the local device so operators are aware of the full policy implemented in the network from the running configuration.

- The ability to interact with traffic flow and other network traffic level measurement protocols and systems, in order to determine path performance, top talkers, and other information required to make an informed path decision based on locally configured policy.

3. Distributed Reaction to Network Based Attacks

Quickly modifying the control plane to reroute traffic for one destination while leaving a standard configuration in place (filters, metrics, and other policy mechanisms) is a challenge --but this is precisely the challenge of a network engineer attempting to deal with a network incursion. The ability to redirect specific flows of information or specific classes of traffic into, through, and back out of traffic analyzers on the fly is crucial in these situations. The following network diagram provides an illustration of the problem.



Modifying the cost of the link between R1 and R2 to draw the attack traffic through the monitoring device in the distributed control plane will, of necessity, also draw the valid traffic through the monitoring device. Drawing valid traffic through a monitoring device introduces delay, jitter, and other quality of service issues, as well as posing a problem for the monitoring device itself in terms of

traffic load and management.

An IRS controller could stand between the detection of the attack and the control plane to facilitate the rapid modification of control and forwarding planes to either block the traffic or redirect it to analysis devices connected to the network.

Summary of IRS Capabilities and Interactions:

- The ability to monitor the available routes installed in the RIB of each forwarding device, including near real time notification of route installation and removal. This information must include the destination prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), the metric of the installed route, and an identifier indicating the installing process.

- The ability to install source and destination based routes in the local RIB of each forwarding device. This must include the ability to supply the destination prefix (NLRI), the source prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), a route preference, a route metric, a next hop, an outbound interface, and a route process identifier.

- The ability to install a route to a null destination, effectively filtering traffic to this destination.

- The ability to interact with various policies configured on the forwarding devices, in order to inform the policies implemented by the dynamic routing processes. This interaction SHOULD be through existing configuration mechanisms, such as NETCONF, and SHOULD be recorded in the configuration of the local device so operators are aware of the full policy implemented in the network from the running configuration.

- The ability to interact with traffic flow and other network traffic level measurement protocols and systems, in order to determine path performance, top talkers, and other information required to make an informed path decision based on locally configured policy.

4. Remote Service Routing

In hub and spoke overlay networks, there is always an issue with balancing between the information held in the spoke routing table, optimal routing through the network underlying the overlay, and mobility. Most solutions in this space use some form of centralized

route server that acts as a directory of all reachable destinations and next hops, a protocol by which spoke devices and this route server communicate, and caches at the remote sites.

An IRS solution would use the same elements, but with a different control plane. Remote sites would register (or advertise through some standard routing protocol, such as BGP), the reachable destinations at each site, along with the address of the router (or other device) used to reach that destination. These would, as always, be stored in a route server (or several redundant route servers) at a central location.

When a remote site sends a set of packets to the central location that are eventually destined to some other remote site, the central location can forward this traffic, but at the same time simply directly insert the correct routing information into the remote site's routing table. If the location of the destination changes, the route server can directly modify the routing information at the remote site as needed.

An interesting aspect of this solution is that no new and specialized protocols are needed between the remote sites and the centralized route server(s). Normal routing protocols can be used to notify the centralized route server(s) of modifications in reachability information, and the route server(s) can respond as needed, based on local algorithms optimized for a particular application or network. For instance, short lived flows might be allowed to simply pass through the hub site with no reaction, while longer lived flows might warrant a specific route to be installed in the remote router. Algorithms can also be developed that would optimize traffic flow through the overlay, and also to remove routing entries from remote devices when they are no longer needed based on far greater intelligence than simple non-use for some period of time.

Summary of IRS Capabilities and Interactions:

- The ability to read the local RIB of each forwarding device, including the destination prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), the metric of each installed route, a route preference, and an identifier indicating the installing process.

- The ability to monitor the available routes installed in the RIB of each forwarding device, including near real time notification of route installation and removal. This information must include the destination prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), the metric of the installed route, and an identifier indicating the

installing process.

- The ability to install destination based routes in the local RIB of each forwarding device. This must include the ability to supply the destination prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), a route preference, a route metric, a next hop, an outbound interface, and a route process identifier.

5. Within Data Center Routing

Data Centers have evolved into massive topologies with thousands of server racks and millions of hosts. Data Centers use BGP with ECMP, ISIS (with multiple LAGs), or other protocols to tie the data center together. Data centers are currently designed around a three or four tier structure with: server, top-of-rack switches, aggregation switches, and router interfacing the data center to the Internet. Microsoft's usage of BGP in the data center, described in [Lapukh-BGP], examines many of these elements of data center design.

One key element of these Data Center routing infrastructures is the ability to quickly read topology information and excute configuration from a centralized location. Key to this environment is the tight feedback loop between learning about topology changes or loading changes, and instantiating new routing policy. Without IRS, may Data Centers are using extra physical topologies or logical topologies to work around the features.

For example, Microsoft's network uses BGP because the topology state could be read from BGP impementations in a consistent fashion. Microsoft might have chosen a different routing protocol (such as ISIS) if the routing protocol state had been easier to obtain. Microsoft chose BGP for the data center because routers had a good BGP interface with topology information.

An IRS solution would use the same in the elements, but with a different control plane. The IRS enable control plane could provide the Data Center 4 tier infrastructure the quick access to topology and data flow information needed for traffic flow optimization. Changes to the Data Center infrastructure done via the IRS could have a tight feedback loop.

Again, this solution would reduce the need for new and specialized protocols while giving the Data Center the control it desire. The IRS routing interface could be extended to virtual routers.

Summary of IRS Capabilities and Interactions:

- The ability to read the local RIB of each forwarding device, including the destination prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), the metric of each installed route, a route preference, and an identifier indicating the installing process.

- The ability to monitor the available routes installed in the RIB of each forwarding device, including near real time notification of route installation and removal. This information must include the destination prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), the metric of the installed route, and an identifier indicating the installing process.

- The ability to install destination based routes in the local RIB of each forwarding device. This must include the ability to supply the destination prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), a route preference, a route metric, a next hop, an outbound interface, and a route process identifier.

- The ability to read the tables of other local protocol processes running on the device. This reading action SHOULD be supported through an import/export interface which can present the information in a consistent manner across all protocol implementations, rather than using a protocol specific model for each type of available process.

- The ability to inject information directly into the local tables of other protocol processes running on the forwarding device. This injection SHOULD be supported through an import/export interface which can inject routing information in a consistent manner across all protocol implementations, rather than using a protocol specific model for each type of available process.

- The ability to interact with various policies configured on the forwarding devices, in order to inform the policies implemented by the dynamic routing processes. This interaction SHOULD be through existing configuration mechanisms, such as NETCONF, and SHOULD be recorded in the configuration of the local device so operators are aware of the full policy implemented in the network from the running configuration.

- The ability to interact with traffic flow and other network traffic level measurement protocols and systems, in order to determine path performance, top talkers, and other information required to make an informed path decision based on locally configured policy.

6. Temporary Overlays between Data Centers

Data Centers within one organization may operate as one single entity even though the Data Centers are geographically distributed fashion. Applications are load balanced within Data Centers and between data centers to take advantage of cost economics in power, storage, and server availability for compute resources. Applications are also transfer to alternate data centers in case of failures within a data center. To reduce time during failure, Data Centers often replicate user storage between two or more data centers. During the tranfer of stored information prior to a Data Center to Data Center move, the Data Center controllers need to dynamically aquire a large amount of inter-data center bandwidth through an overlay network, often during off hours.

IRS could provide the connection between the overlay network configuration, local policies, and the control plane to dynamically bring a large bandwidth inter-data center overlay or channel into use, and then to remove it from use when the data transfer is completed.

Similarly, during a fail-over, a control process within data centers interacts with a group host process and the network to seamless move the processing to another data center. During the fail-over case, additional process state may need to be moved as well to restart the system. The difference between these data-to-data center moves is immediate and urgent need to move systems. If an application (such as medical or banking services) pays to have this type of fail-over, it is likely the service will pay for preemption on network bandwidth. IRS can allow the Data Center network and the Network connecting the data center to prempt other best-effort traffic to send this priority data flow. After the high priority data flow has finished, networks can return to their previous condition

Summary of IRS Capabilities and Interactions:

- The ability to read the local RIB of each forwarding device, including the destination prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), the metric of each installed route, a route preference, and an identifier indicating the installing process.

- The ability to monitor the available routes installed in the RIB of each forwarding device, including near real time notification of route installation and removal. This information must include the destination prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), the metric of the installed route, and an identifier indicating the

installing process.

- The ability to install destination based routes in the local RIB of each forwarding device. This must include the ability to supply the destination prefix (NLRI), a table identifier (if the forwarding device has multiple forwarding instances), a route preference, a route metric, a next hop, an outbound interface, and a route process identifier.

- The ability to interact with various policies configured on the forwarding devices, in order to inform the policies implemented by the dynamic routing processes. This interaction SHOULD be through existing configuration mechanisms, such as NETCONF, and SHOULD be recorded in the configuration of the local device so operators are aware of the full policy implemented in the network from the running configuration.

- The ability to interact with policies and configurations on the forwarding devices using time based processing, either through timed auto-rollback or some other mechanism. This interaction SHOULD be through existing configuration mechanisms, such as NETCONF, and SHOULD be recorded in the configuration of the local device so operators are aware of the full policy implemented in the network from the running configuration.

- The ability to interact with traffic flow and other network traffic level measurement protocols and systems, in order to determine path performance, top talkers, and other information required to make an informed path decision based on locally configured policy.

7. Central membership computation for MPLS based VPNs

MPLS based VPNs use route target extended communities to express membership information. Every PE router holds incoming BGP NLRI and processes them to determine membership and then import the NLRI into the appropriate MPLS/VPN routing tables. This consumes resources, both memory and compute on each of the PE devices.

An alternative approach is to monitor routing updates on every PE from the attached CEs and then compute membership in a central manner. Once computed the routes are pushed to the VPN RIBs of the participating PEs.

This centralization of membership control has a few advantages.

- The membership mechanism (route-targets) need not be configured in each of the PEs and can be expressed once centrally.

- No resources in the PEs need to be spent to categorize routes into the VRF tables that they belong and to filter out unwanted state.

- Doing it centrally means the availability of almost unlimited compute capacity to compute membership and hence can be done in a scaleable manner.

- More sophisticated routing policies and filters can be applied during the central import/export process than can be expressed and performed using the traditional route target mechanism.

- Routes can be selectively pushed only to the participating PE's further reducing the memory load on the individual routers in the network. This further obviates for a distributed mechanisms such as rt constraints to reduce unnecessary path state in the routers.

Note that centrally compution of membership can be applied to other scenarios as well such as VPLS, MVPNs, MAC VPNs etc. Depending on the scenario, what gets monitored from the CE might vary. Central computation will especially help VPLS where multi-homing and load balancing using distributed techniques has particularly been a challenge.

Also note that one of the biggest promises of central route computation is simplification and reduction of computation and memory load on all devices in the network. This use case is just one example that illustrates these benefits of central computation very well.

Summary of IRS Capabilities and Interactions:

- The ability to read the loc-RIB-In BGP table that gets all the routes that the CE has provided to a PE router.

- The ability to install destination based routes in the local RIB of the PE devices. This must include the ability to supply the destination prefix (NLRI), a table identifier, a route preference, a route metric, a next-hop tunnel through which traffic would be carried

8. Normative References

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

Authors' Addresses

Russ White
Verisign
12061 Bluemont Way
Reston, VA 20190
USA

Email: riwhite@verisign.com

Susan Hares
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: Susan.Hares@huawei.com

Rex E. Fernando
Cisco Systems
170 W Tasman Dr
San Jose, CA 95134
USA

Email: rex@cisco.com