I2RS Working Group
Internet Draft
Category: Informational
Expires: April 2014

R. Krishnan
Brocade Communications
A. Ghanwani
Dell
S. Kini
Ericsson
D. Mcdysan
Verizon
Diego Lopez
Telefonica

February 13, 2014

Large Flow Use Cases for I2RS PBR and QoS

draft-krishnan-i2rs-large-flow-use-case-03

Abstract

This draft discusses two use cases to help identify the requirements for policy-based routing in I2RS. Both of the use cases involve identification of certain flows and then using I2RS to program special handling for those flows.

The first use case deals with improving bandwidth efficiency. Demands on networking bandwidth are growing exponentially due to applications such as large file transfers and those with rich media. Link Aggregation Group (LAG) and Equal Cost Multipath (ECMP) are extensively deployed in networks to scale the bandwidth. However, the flow-based load balancing techniques used today make inefficient use of the bandwidth in the presence of long-lived large flows. We discuss how I2RS can be used for achieving better load balancing.

The second use case is for recognizing and mitigating Layer 3-4 based DDoS attacks. Behavioral security threats such as Distributed Denial of Service (DDoS) attacks are an ongoing problem in today's networks. DDoS attacks can be Layer 3-4 based or Layer 7 based. We discuss how such attacks can be recognized and how I2RS can be used for mitigating their effects.

Status of this Memo

Copyright Notice

Conventions used in this document

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 RFC-2119 [RFC 2119].

Table of Contents

1. Introduction

This draft describes use cases that address two problems caused by large flows. The use case consists of mitigating the problem by applying policy-based routing (PBR) on the routing elements using its I2RS. The first large flow problem is that of inefficient bandwidth usage due to hash-based load balancing in networks and the second is that of DDoS attacks.

1.1. Large Flow Load Balancing

Networks extensively deploy LAG and ECMP for bandwidth scaling. Network traffic can be predominantly categorized into two traffic types: long-lived large flows and other flows (which include long-lived small flows, short-lived small/large flows) [OPSAWG-large-flow]. Stateless hash-based techniques [ITCOM, RFC 2991, RFC 2992, and RFC 6790] are often used to distribute flows over the components in a LAG/ECMP irrespective of whether the flows are long-lived large flows or other types. In a traffic distribution consisting of long-lived large flows, the traffic load may not be evenly distributed over the components of the LAG or ECMP.

This draft describes long-lived large flow load balancing techniques for achieving the best network bandwidth utilization with LAG/ECMP and the corresponding I2RS requirements. Some of these techniques have been described in detail in [OPSAWG-large-flow]. We describe methods that can be used locally within a single router, as well as methods that can be applied across multiple network elements, where the network is under the control of single administrative entity. We refer to the former as local load balancing and the latter as global load balancing. A combination of local/global load balancing helps in achieving the best network bandwidth utilization and latency for a given network topology.

At a high-level, the technique involves recognizing large flows and rebalancing them to achieve optimal load balancing. Large flows may be recognized within a router, or using the aid of an external entity such as an IPFIX [RFC 7011] collector or a sFlow [sFlow-v5] collector. Once a large flow has been recognized, it must be signaled to an application that makes the rebalancing decision. Finally, the rebalancing decision is communicated to the routers to program the forwarding plane. In subsequent sections, we describe the requirements with recognition and rebalancing as they pertain to I2RS.

1.2. DDoS attack mitigation

Layer 3-4 based DDoS attacks are an ongoing problem in today's networks. Example of Layer 3-4 based DDoS attacks are [FDDOS]:

- SYN Flood Attack: Fake TCP connections are setup which result in table overflows in stateful devices.

- UDP Flood Attack: Servers are flooded with UDP packets that result in consumption of bandwidth and CPU. These can be used to target specific services by attacking, e.g., DNS servers and VOIP servers.

- Christmas Tree Flood Attack: TCP packets from non-existent connections with flags other than the SYN flag sent to servers result in consumption of more CPU than normal packets because of the effort required to discard them.

Typically, the above attacks are not from a single host or source IP address; multiple hosts with different source IP addresses working in tandem cause these attacks - hence the term Distributed DoS or DDoS.

The DDoS use case involves recognizing large flows and performing various types QoS actions on the recognized flows based on configured policies. Large flows may be recognized within a router, or using the aid of an external entity such as an IPFIX [RFC 7011] collector or a sFlow [sFlow-v5] collector. In subsequent sections, we describe the requirements with respect to recognition and QoS actions as they pertain to I2RS.

1.3. Large Flow Identification

From the standpoint of a router, large flows are typically identified using one or more fields from the packet header from the following list:

- Layer 2: source MAC address, destination MAC address, VLAN ID.

- IP/TCP/UDP header: IP Protocol, IP source address, IP destination address, flow label (IPv6 only), TCP/UDP source port, TCP/UDP destination port, TCP Flags.

- MPLS Labels.

For tunneling protocols like GRE, VXLAN, NVGRE, STT, etc., flow identification is possible based on inner and/or outer headers. The

above list is not exhaustive. This definition of a flow is consistent with [RFC 7011].

In the remainder of this document, consistent with [OPSAWG-large-flow], we use the term "large flow" to refer to "long-lived large flows," and we use the term "small flow" to refer to any of the three other types of flows identified above.

1.4. Acronyms

COTS: Commercial Off-the-shelf

DoS: Denial of Service

DDoS: Distributed Denial of Service

ECMP: Equal Cost Multi-path

GRE: Generic Routing Encapsulation

LAG: Link Aggregation Group

LSR: Label Switch Router

MPLS: Multiprotocol Label Switching

NVGRE: Network Virtualization using Generic Routing Encapsulation

PBR: Policy Based Routing

QoS: Quality of Service

STT: Stateless Transport Tunneling

TCAM: Ternary Content Addressable Memory

VXLAN: Virtual Extensible LAN

1.5. Terminology

Large flow(s): long-lived large flow(s)

Small flow(s): long-lived small flow(s) and short-lived small/large flow(s)

2. Large Flow Recognition, Signaling, and Rebalancing

2.1. Network-based Recognition of Large Flows

The first step is recognizing large flows. There are two ways for recognizing large flows as described in [OPSAWG-large-flow].

The first method is automatic hardware-based recognition in which the large flows are identified in hardware. Once a large flow is recognized, it needs to be communicated to an application that is capable of making rebalancing decisions. This communication is out of scope for I2RS and can be handled using protocols such as IPFIX [RFC 7011].

The next method is where sFlow or IPFIX packet sampling [PSAMP] can be used to convey packet samples to an external entity such as sFlow or IPFIX collector. The external entity recognizes large flows and this entity signals the large flows to another application that is capable of making rebalancing decisions. Once again, this communication is out of scope of the I2RS. An example of software which can be used to recognize large flows in real-time is inMon sFlow-RT [sFlow-RT]; sFlow-RT is a component of the external sFlow collector entity.

2.2. Off-network based notification of Large Flows

Instead of having the network recognize large flows, the large flow can be notified by an application that has awareness of large flows, e.g. a backup operation, and may perhaps indicate other parameters such as the latency desired. Such flows would once again need to be notified to the application capable of routing or rebalancing decisions. This communication is also outside the scope of I2RS.

2.3. Flow Rebalancing

2.3.1. Local Rebalancing

In the case of local rebalancing, the utilization of the component links that are part of the LAG or ECMP are monitored and the flows are redistributed among the member links to ensure optimal load balancing across all of the component links. Typically, this involves redirecting large flows to specific ECMP or LAG components, and potentially adjusting the weights used to distribute small flows across these components, using mechanisms specified in [OPSAWG-large-flow].

This approach works regardless of whether the underlying network is IP or MPLS.

At the RIB level, the nexthop information is typically resolved over an IP interface. However, the IP interface can be realized over a L2 LAG. For this use case the nexthop of a PBR route should be resolvable to the granularity of a component of a L2 LAG.

To achieve this, there are two requirements for I2RS:

- For redirecting large flows to a specific component, a PBR entry should be programmable for the flow with its nexthop that identifies the specific LAG or ECMP component.

- For adjusting the weights used to distribute traffic across components of the LAG or ECMP, a mechanism a programmable mechanism should be provided that identifies ECMP entries and is able to associate weights that can be programmed for each of the components. To do this in a scalable fashion, it would be useful to have the notion of an ECMP nexthop that is used by multiple routes.

2.3.2. Global Rebalancing

2.3.2.1. IP Networks

For IP networks, this involves programming a globally optimal path for the large flow. The globally optimal path is programmed in the IP network using hop-by-hop PBR rules.

For IP networks, this involves creating a globally optimal path [HEDERA-dynamic-flow-scheduling] using a network management entity which hosts an I2RS client. The globally optimal path is programmed in the IP network using hop-by-hop PBR rules. The weights of the ECMP table for different nexthops should be adjusted to factor the long-lived large flows - this is explained below with an example.

As an example, consider a 4 way ECMP at node n1 with IP nexthops n11, n12, n13, n14 using links l1, l2, l3, l4 each of capacity 10 Gbps. Say, a long-lived large flow of average bandwidth 2 Gbps is admitted to one of the links l3. The ECMP nexthop table needs to be adjusted to approximately account for the long-lived large flow so that the other flows do not overload link l3 which is already used by the large flow. The ECMP nexthop table will be programmed as w1*n11, w2*n12, w3*n13, w4*n14 where w1=w2=w4=1 and w3=0.8; this needs to be done for all the routes using the same set of nexthops.

Now, if there are other set of nexthops from node n1 using link l3, they should also be adjusted. Say, there is another set of IP ECMP nexthops n13, n14, n15, n16 using links l3, l4, l5, l6. The ECMP nexthop table will be programmed as w1*n13, w2*n14, w3*n15, w4*n16 where w2=w3=w4=1 and w1=0.8; this needs to be done for all the routes using the same set of nexthops. In practice, there could be multiple large flows on a single link and the ECMP nexthop table must be adjusted to factor all of these flows.

As mentioned in Section 2.3.1. , there should be a way of addressing an ECMP group, so that all routes sharing an ECMP group are addressed together.

2.3.2.2. MPLS Networks

There are several ways to address global load rebalancing in MPLS networks. For example:

- Have multiple LSPs between ingress and egress routers. In this case, having a PBR entry at the edge LSR that forwards the large flow to specific LSP known to have the necessary bandwidth is needed.

- Program a new LSP for a given large flow.

Here the requirements for I2RS should be to provide the ability to program PBR entries at the edge LSR, and to program new LSPs in the network.

2.3.3. Packet Reordering During Rebalancing

During rebalancing events, as flows are moved from one component link of a LAG to another, or from one ECMP nexthop to another, there is a possibility of packets getting reordered.

In the case of link aggregation, IEEE 802.1AX [IEEE-802.1AX] defines a Marker Protocol which can be invoked at times when rebalancing occurs before flows are moved.

Another possibility is to make the forwarding logic aware of flows whose packets are sensitive to ordering and avoid moving those flows. This can be done in the following way. Consider an ECMP group with n nexthops. We define 2 ECMP separate ECMP groups with these n nexthops. The first ECMP group (G1) would be static; i.e. its weights would not be changed. The second ECMP group (G2), which is dynamic, would have its weights adjusted in accordance with rebalancing events as described above. Now when a packet arrives,

it is classified as whether it belongs to a reordering sensitive flow or not. If it belongs to a reordering sensitive flow, then a lookup is done in a FIB which yields the static ECMP group G1. Otherwise, the lookup is done in a different FIB which would yield the dynamic ECMP group G2. This makes the assumption that the ordering sensitive flows are relatively low bandwidth and would therefore not impact the rebalancing scheme in a significant way.

3. DDoS Detection and Mitigation

Layer 3-4 based DDoS attacks can be mapped to large flows in the network. Consider the following example of a TCP SYN attack. A TCP SYN packet from a single source IP address can be mapped to a Layer 4 flow based on the following: IP source/destination addresses, TCP/UDP source/destination ports, IP protocol, TCP SYN Flag. For the purpose of DDoS it is not useful to observe the above Layer 4 flow in the network. Say, we observed a large flow based on IP destination addresses, TCP/UDP destination port, IP protocol, and TCP SYN flag in the network. In the case of a DDoS attack such a flow would cause a significant event in the network in terms of exceeding a pre-defined bandwidth threshold over an observation interval.

Once the large flows causing the DDoS attacks are recognized in the network, various types of Quality of Service (QoS) actions such as rate-limiting, re-marking, or discarding can be performed on the flows based on configured policies. Besides the QoS actions, we need the capability to redirect the large flow to a DDoS scrubber appliance for further examination (typically layer 7) of the traffic - this can be accomplished through nexthop redirection (the nexthop may be directly connected to the router or indirectly through a tunnel). The QoS action is independent of the nexthop redirection action. From an I2RS requirement perspective, it should be possible to program either of these actions independently of the other. This would help in preventing resource exhaustion (CPU, memory etc.) on devices such as servers and unfair access to network resources in a multitenant network.

4. Summary

We have described the problems of large flow load balancing and DDoS mitigation using I2RS. In both cases, the problem translates to that of detection large flows that meet certain criteria. The detection can be done without I2RS using tools such as IPFIX and sFlow.

Once a large flow has been detected, I2RS must be used to modify the forwarding tables in the router.

- In the case of large flow load balancing, this may involve redirecting the large flow to a particular member with the LAG or ECMP group and readjusting the weights of the other members to account for the large flow.

- In the case of DDoS mitigation, the action involves rate limiting, remarking or potentially discarding the large flow in question.

5. Operational Considerations

Operational considerations would be similar to those specified in [OPSAWG-large-flow].

6. IANA Considerations

None.

7. Security Considerations

This draft specifies a use case for I2RS and does not introduce any new security requirements beyond those already under consideration for I2RS.

8. Acknowledgements

9. References

9.1. Normative References

9.2. Informative References

[OPSAWG-large-flow] Krishnan, R. et al., "Mechanisms for Optimal LAG/ECMP Component Link Utilization in Networks," February 2014.

[HEDERA-dynamic-flow-scheduling] Al-Fares, M. et al., "Hedera: Dynamic Flow Scheduling for Data Center Networks", December 2009

[sFlow-v5] Phaal, P. and M. Lavine, "sFlow version 5," July 2004.

[RFC 7011] Claise, B., "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information,", September 2013

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

[sFlow-RT] http://www.inmon.com/products/sFlow-RT.php

[PSAMP] Claise, B., "Packet Sampling (PSAMP) Protocol Specifications", March 2009

[FDDOS] David Holmes, "The DDoS Threat Spectrum", F5 White paper 2012

[IEEE-802.1AX] IEEE Standard for Local and metropolitan area networks--Link Aggregation

Authors' Addresses

Ram Krishnan
Brocade Communications
ramk@brocade.com

Anoop Ghanwani
Dell
anoop@alumni.duke.edu

Sriganesh Kini
Ericsson
sriganesh.kini@ericsson.com

Dave Mcdysan
Verizon
dave.mcdysan@verizon.com

Diego Lopez
Telefonica I+D
Don Ramon de la Cruz, 82 Street
Madrid, 28006, Spain
+34 913 129 041
diego@tid.es