

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  
April 3, 2014

## Large Flow Use Cases for I2RS PBR and QoS

[draft-krishnan-i2rs-large-flow-use-case-04](#)

### 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 routers with 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 static hash-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

This Internet-Draft is submitted to IETF in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/1id-abstracts.txt>.

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

This Internet-Draft will expire on April, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

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.....4](#)
- [1.1. Large Flow Identification.....4](#)
- [1.2. Large Flow Load Balancing.....4](#)
- [1.3. DDoS attack mitigation.....5](#)
- [1.4. Security Appliance Bypass for Non-Malicious Large Flows...6](#)
- [1.5. Acronyms.....6](#)
- [1.6. Terminology.....7](#)
- [2. Large Flow Recognition, Signaling, and Rebalancing.....7](#)



- [2.1. Network-based Recognition of Large Flows.....](#)[7](#)
- [2.2. Off-network Notification of Large Flows.....](#)[8](#)
- [2.3. Flow Rebalancing.....](#)[8](#)
  - [2.3.1. Local Rebalancing.....](#)[8](#)
  - [2.3.2. Global Rebalancing.....](#)[9](#)
  - [2.3.3. Packet Reordering During Rebalancing.....](#)[10](#)
- [3. DDoS Detection and Mitigation.....](#)[10](#)
- [4. Summary.....](#)[11](#)
- [5. Operational Considerations.....](#)[11](#)
- [6. IANA Considerations.....](#)[11](#)
- [7. Security Considerations.....](#)[12](#)
- [8. Acknowledgements.....](#)[12](#)
- [9. References.....](#)[12](#)
  - [9.1. Normative References.....](#)[12](#)
  - [9.2. Informative References.....](#)[12](#)
- [Authors' Addresses.....](#)[13](#)

## **1. Introduction**

This draft describes use cases that address two problems caused by long-lived large flows. In these use cases, the problems in question are addressed by applying policy-based routing (PBR) rules on the routing elements using its I2RS. The first 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 Identification**

From the standpoint of a router, long-lived 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 flow," and we use the term "small flow" to refer to any of the three other types of flows (lived small flow, short-lived small/large flow).

### **1.2. Large Flow Load Balancing**

Networks extensively deploy LAG and ECMP for bandwidth scaling. 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 large flows or other types. In a traffic distribution consisting of large flows, the traffic load may not be evenly distributed over the components of the LAG or ECMP.

This draft describes 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 and global load balancing helps achieve 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.

### **[1.3. DDoS attack mitigation](#)**

Layer 3-4 based DDoS attacks are an ongoing problem in today's networks. Examples of Layer 3-4 based DDoS attacks are [[FDDOS](#)][NTP-DDoS]:

- . 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 for discarding them.
- . NTP Reflection Attack: This attack is caused by an attacker sending a specially crafted NTP query that ultimately redirects a large volume of traffic. The traffic is sent with a spoofed source address with the intention of having the NTP servers return responses to the spoofed address, which, would be the intended target.





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.

Many DDoS attacks manifest as large Layer 3-4 flows. For example, a TCP SYN Flood attack can be recognized as a large number of packets sent to the same (range of) destination IP address(es) with the SYN bit set and a relatively small number of packets with the ACK bit set.

The DDoS use case involves recognizing such 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.

#### **[1.4. Security Appliance Bypass for Non-Malicious Large Flows](#)**

In some cases, large flows are not malicious, for example the destination IP address (ranges) and port number (ranges) that correspond to a CDN server or some other trusted destination, and are thus known to not be a security threat. This information could be used to bypass a security attack detection device; for example one that is designed to collect data for Layer 7 threats such as, for example, HTTP GET Floods [[FDDOS-L7](#)] and low rate malicious port scanning. By bypassing the security appliance for such large flows, an appliance of lower capacity can be used.

Since this case involves detecting a large flow and programming a special rule to bypass the appliance (where all traffic would normally be forwarded), it presents similar requirements to those for solving the large flow load balancing problem and the DDoS attack mitigation problem discussed in [Section 1.2](#) and [Section 1.3](#) respectively. Consequently, this use case is not discussed further in this document.

#### **[1.5. 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

## **1.6. 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 [[sFlow-v5](#)] or IPFIX packet sampling [[RFC 5476](#)] 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 for the use of sFlow for detecting large flows in real-time is described in [[sflow-RT](#)].

## **[2.2. Off-network 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 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 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 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 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 sets 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 are 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**

Many DDoS attacks manifest as large Layer 3-4 flows.

For example, an NTP reflection attack [[NTP-DDoS](#)] is caused by an attacker sending a specially crafted NTP query that ultimately redirects a large volume of traffic. The traffic is sent with a spoofed source address with the intention of having the NTP servers return responses to the spoofed address, which, would be the intended target. This attack would trigger a large flow based on IP destination address, IP Protocol UDP, UDP Source Port NTP which





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.

[RFC 5476] Claise, B., "Packet Sampling (PSAMP) Protocol Specifications," March 2009.

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

[IEEE-802.1AX] IEEE Standard for Local and metropolitan area networks--Link Aggregation, November 2008.

[FDDOS-L7] Holmes, D., "Mitigating DDoS Attacks with F5 Technology", F5 White paper, 2013.

[sflow-RT] Phaal, P., "Performance optimizing hybrid OpenFlow controller," <http://blog.sflow.com/2014/03/performance-optimizing-hybrid-openflow.html>, March 2014.

[NTP-DDoS] "NTP Reflection Attacks," <https://blogs.akamai.com/2014/02/ntp-reflection-attacks.html>, February 2014

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