Large Flow Classification in Flow Aware Transport over PSN

draft-yong-pwe3-lfc-fat-pw-01.txt

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 January 11, 2010.

Copyright Notice

Copyright (c) 2010 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. Code Components extracted from this
document must include Simplified BSD License text as described in
Section 4.e of the Trust Legal Provisions and are provided without
warranty as described in the Simplified BSD License.
Abstract

Network traffic has shown the combination of few very high bit rate flows (large flow) and a huge amount of very low bit rate flows (small flow), which causes uneven load balance over ECMP or LAG. Differentiating large flow and small flow packets in IP/MPLS networks enables an enhanced ECMP transport. Enhanced ECMP applies different distribution methods to large flow packets and small flow packets to improve the load balance and congestion control. This draft proposes large flow classification scheme for flow awareness transport in PSN.

Table of Contents

1. Introduction ................................................... 2
2. Conventions used in this document .............................. 4
   2.1. Terminology ............................................. 4
3. Large Flow Recognition ........................................ 4
4. Large Flow Classification ...................................... 5
5. Backward Compatibility ....................................... 7
6. Applicability .................................................. 7
   6.1. Link Aggregation Groups .................................. 8
   6.2. The Single Large Flow Case .............................. 8
   6.3. Multi-Segment Pseudowires ................................ 8
   6.4. IP Flows ................................................ 8
   6.5. LSP with Entropy Label ................................... 9
7. Security Considerations ........................................ 9
8. IANA Considerations .......................................... 9
9. References .................................................... 9
   9.1. Normative References .................................... 9
   9.2. Informative References ................................ 10
10. Acknowledgments .............................................. 11
Appendix A. Enhanced ECMP Example ................................ 12
   10.1. Congestion Control .................................... 14
   10.2. Flow Rate Difference ................................... 14
   10.3. Backward Compatibility ................................ 15
Appendix B. Simulation Analysis .................................. 16

1. Introduction

[FAT-PW] introduces a flow label on the label stack for a pseudowires (PW) to take the advantage of ECMP transport. The method inserts a flow label on each packet at ingress PE. The ECMP process in the packet switched network (PSN) hashes the label stack that contains the flow label at the bottom. As a result, individual flows
in a PW can be transported over different ECMP paths. Since the packets that belong to the same flow have the same label value, the method gets ECMP transport benefit as well as preserves the ordering of individual transported IP flows. The method is referred to as Flow Aware Transport in PSN.

However, the aggregated traffic flows today includes Web browsing data and audio as well as video/downloading and streaming. High-quality-video streaming or file download generates the very high bit rate flows compared to Web browsing data/audio. This results in that the network traffic is clearly mixed with huge amount of low bit rate flows and few very high bit rate flows. Internet traffic analysis [CAIDA] indicates that, ~2% of the top rate ranked flows takes about 30% of traffic volume while the rest of 98% flows contribute 70% of traffic volume. Carrier network even shows fewer percentiles of large flows that contribute more to total traffic volume. As Web HDTV and 3D TV will be on the Internet, the bit rate gap between large flows and small flows will get further higher. In addition, carrier IP/MPLS network carries many L3VPN and L2VPN services. Among these services, some traffic may be from data center, which may act as a single flow that can be much higher bit rate compared to the flow generated by any application. Although the flow label in PW can provide better granularity than a PW, under such traffic pattern, hash based flow routing can not perform even load balance because the bandwidth per flow is significant unbalanced.

Experiments have shown that applying the same treatment for high bit rate flows and low bit rate flows conducts a poor performance or utilization over multiple equal cost multi paths (ECMP). If the network uses a stateful method for flow placement over the paths, a huge amount of the flows add a big burden for device to handle. If the network uses stateless method (hashing, [RFC2991] and [RFC2992]), significant rate differences make uneven load balance on ECMP paths. This results a desire to improve load balance over ECMP paths under such traffic pattern. This draft suggests applying different treatments on very high bit rate flows and low bit rate flows in flow aware transport. It drives the need for Large Flow Classification. This draft specifies Large Flow Classification scheme, which is used for ECMP load balance purpose. Other types of flow classification are for further study.

Appendix A describes one example that using large flow classification to improves load balance over ECMP. Appendix B states some simulation results. Other enhanced ECMP implementations may be implemented by using Large Flow Classification as well.
2. 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 [RFC2119].

2.1. Terminology

Flow: a group of packets that contain the same flow ‘‘identity’’ in their header. The network intends to transport the packets that belong to the same flow over the same path so flow packet sequence is preserved.

Large Flow: a flow comes to the network at the bit/packet rate that is on the top bit rate rank among all the flows

Small Flow: a flow or data that does not belong to a large flow.

3. Large Flow Recognition

The advanced technologies now enable router devices to inspect the received the packets and identifies the large flows from huge amount of received packets that belong to many and many different flows. Large flow recognition process may use protocol inspection, flow volume measurement, access list, or other methods to detect the large flows. If a router can differentiate packets that belong to the very high bit rate flows from all the received packets, it enables differentiated treatments on the large flow packets and small flow packets in the flow based routing in PSN. Both ECMP and LAG in PSN can significantly benefit from such differentiation. One example is given in Appendix A. Other methods can be implemented as well.

It is possible for hosts to insert a large flow indication on the packet header. However, there is a huge security concern for a network to perform on the customer inserted indication.

All ECMP or LAG in PSN can perform differentiated treatments on large flow packets and small flow packets. It is an obvious benefit that, if ingress PE performs the large flow recognition only and inserts a large flow indication on the packets, all the P nodes within PSN can distinguish the large flow packets by checking this indication and perform the different treatments on the packets. This can substantially reduce the implementation cost and the impact on the performance.
The native service processing function (NSP) [RFC3985] in the ingress PE can identify the flow or groups of flows in a PW service, and insert the flow label (load balance label) on each packet before it is passed to the pseudowire forwarder. Large flow recognition process described in this document can be placed at the front of the pseudowire forwarder [RFC3985]. It performs the packet inspection, detect the large flow packets, and insert a large flow indication on the load balance label in packets that belong to the large flows. The design method for the large flow recognition is outside the scope of this document.

4. Large Flow Classification

This draft specifies the protocol to encode a large flow indication on the flow label specified in [FAT-PW]. Figure 1 illustrates current flow label format specified in [RFC3032] with the amendment given in [RFC5462]. Label field is filled with the flow identification in a PW. Since the flow label never occurs on the top of label stack [FAT-PW], TTL field is not used. However, to prevent any provisioning error, [FAT-PW] recommends that TTL filed sets to 1. S bit is used to indicate the bottom of stack and set to 1 for the flow label. Three bits of Traffic Class are not used in flow aware transport. Note: Client QoS information is transformed to traffic class bits on PW label.

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Label                  | TC  |S|       TTL     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Label:  Label Value, 20 bits
TC:     Traffic Class, 3 bits; no define
S:      Bottom of Stack, 1 bit; set to 1
TTL:    Time to Live, 8 bits; set to 1
```

Figure 1 Current Flow Label Semantics

The document suggests using 1 bit in Flow Label Traffic Class to differentiate the large flow from other flows in a PW (referred to as small flows), and suggests value 1 to represent the large flow and value 0 for the small flow. The two other bits are reserved for the future. Since flow label, PW label, and LSP label may all be placed at the bottom of label stack, for differentiated flow treatments in PSN, it is necessary for LSR to distinguish the flow label from both PW label and LSP label to support backward
compatibility. Given limited option on the flow label, the draft proposes setting TTL value to 0 on the flow label, which can be used to differentiate it from PW label and LSP label at the bottom stack. (However, open for other suggestion) Since the flow label and PW label are pushed into label stack at ingress PE and removed at egress PE [FAT-PW], the flow label SHALL never occur on the top label stack entry, which means that its TTL can be neither incoming TTL nor outgoing TTL [RFC3032]. If FL TTL accidentally occurred as outgoing TTL, the packets would be dropped locally based on RFC3032. Hence it can never be incoming TTL. Figure 2 shows proposed flow label semantics.

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Label                  | TC  |S|       TTL     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Label: Label Value, 20 bits
TC: Traffic Class, 3 bit;
    First bit: for flow classification,
    Set 1 for large flow, 0 for small flow
    Other 2 bit: reserve for future; set to 0
S: Bottom of Stack, 1 bit; set to 1
TTL: Time to Live, 8 bits; set to 0

Figure 2 Flow Label Semantics with Large Flow Classification

When Flow Label is used on a PW, ingress PE can insert the flow label and a large flow indication on each packet; egress PE will trim off the PW label and flow label before sending the packets to the right AC. The procedure for informing flow label presence and label insertion procedure remains the same as [FAT-PW].

TC bits in MPLS Label are typically used for DiffServ function which uses TC bits on the label that occurs on the top of label stack entry. [RFC3270] [RFC5129] Since Flow Label never occurs on the top Label Stack entry and only be used in flow based routing, DiffServ function does not apply to TC bits on the Flow Label. Hence the proposed Flow Label TC bit usage in this document does not conflict with DiffServ function.

It is worth to mention the differences between DiffServ and large flow classification. Although both relate to traffic classification, they aim on the different purposes. DiffServ is for the network to perform differentiated service treatments, i.e. different queuing
processes. Large Flow Classification apply to flow based routing in ECMP for load balance.

Although control plan/signaling can be also used to inform P node about large flow if it is carried over one PW and one LSP, large flow classification in data plane serves different purposes. Today, signaling a PW is used per a service basis such as VPWS or VPLS and it can be multiplexed with other PWs into one LSP that is transported over PSN. The control plane at P node is only aware of the LSP (top label) but is not aware of the PW and Flows in the PW at all. Only ECMP function perform flow based routing, uses of the label stack, and is aware of inner label on the label stack. A service may contain many flows. Carriers need to set up a PW or LSP per a service basis not per individual flow basis. This is the case that the flow label is used and the large flow classification is necessary. Using one LSP to carry one large flow can apply to some cases. In this case, the control plane can determine which ECMP path the LSP should take.

5. Backward Compatibility

The method fully supports backward compatibility in PSN. If a PW is not configured with Flow Label at the PE that supports large flow recognition process, the process should be disabled. If ingress PE does not support Large Flow Recognition and PW flow label is used, it SHALL set large flow classification bit as 0. If Flow label is used, its TTL value SHALL be set to 0. When intermediate router does not support the differentiated treatment, it uses hashing in ECMP or LAG process. If intermediate router supports the differentiated treatment, it performs large flow treatment if the received packets belong to the large flow.

6. Applicability

Carriers have desires to improve transport network capability via certain service awareness in packet transport and are not to be constrained in just ‘’pipe’’ transport service. [FAT-PW] brings such potential by introducing the flow label in the label stack, in which ECMP transport discriminates traffic at flow granularity. The large flow classification further enables ECMP transport to perform different treatments on large flows and small flows, which can improve the load balance when flow bit rates vary substantially.

The method described in this document requires the new capability in the PSN and applies to packet switched routers. It requires ingress PE to perform the large flow recognition and inserts a large flow indication on the flow label; and ECMP process on P or PE routers
perform the different treatments on large flow and small flow packets. Each router node performs ECMP function independently, a packet switched network can work well even when some nodes support the enhanced ECMP capability and some do not. Note: P nodes without different treatment capability remain the same level load balance performance as before. This lets operator to gradually upgrade the network.

6.1. Link Aggregation Groups

A Link Aggregation Group (LAG) is used to group several physical links together between two adjacent nodes so they appear to higher-layer protocols as a single, higher bandwidth ‘‘virtual’’ pipe. These may co-exist in various parts of a given network. The large flow classification proposed in this document can facilitate flow distribution in LAG.

6.2. The Single Large Flow Case

[FAT-PW] has suggested several options for the single large flow in a PW. With the enhanced ECMP capability, it has beneficial to insert a flow label even there is one single large flow in a PW. Then ingress PE can insert a large flow indication. P routers in PSN can treat this PW as a large flow. In this case, the control plane does not need to determine which ECMP path for this PW/LSP.

6.3. Multi-Segment Pseudowires

The flow label mechanism described in this document works on multi-segment PWs [MS-PW] without requiring any modification to the Switched PEs (S-PEs). This is because the flow label is transparent to the label swap operation. There is no need to perform Large Flow Recognition at Switched PEs.

6.4. IP Flows

Today’s ECMP process applies to both IP flows and MPLS labeled traffic in PSN. Typically, Hash method uses IP source and destination address pair plus other elements to discriminate IP flows and distribute them over ECMP paths. If PE can insert a large flow indication in the packets of IP flows, the proposed method can apply to IP flows as well. IPv6 protocol [RFC2460] already has the flow label field. [RFC3697] describes flow label specification. Source address and Flow Label can be used in ECMP process. The method proposed here meets the restriction on the flow label. [FLOW-ECMP] IETF just needs to specify one bit in TC field (8 bits) for large flow classification. [LFC-IPv6] By default, all TC bits are
set as 0 [RFC2460], which is compatible to this solution. P nodes in PSN can easily differentiate IPv6 packets and check if the flow label is used or not. Although IPv4 protocol does not have such flow label, IETF can decide if it is necessary to improve IPv4 protocol to have the large flow classification or just wait the time for IPv6 to take over. The IP large flow recognition and indication is outside the scope of this document.

The Packet Separation Process in the enhanced ECMP uses the first nibble to differentiate IP flows and non IP flows before evaluating the large flow indication. When PSN does not support large IP flow classification, the enhanced ECMP treats all IP flows as small flows.

6.5. LSP with Entropy Label

Entropy Label [Entropy] is inserted in LSP traffic at ingress LSR to gain better ECMP load balancing at transit LSRs. The purpose of Entropy Label is the same as PW flow label and is used to differentiate "microflow" within a LSP for ECMP process. Enhanced ECMP and Large Flow Classification can apply to LSP with entropy label. Traffic class field in the Entropy can use the same semantics as PW flow label described in this document. If ingress LSR does not support large flow recognition, then it SHOULD set Large Flow Classification bit to 0. Entropy Label TTL value is set to 0. LSP with Entropy Label has beneficial to apply L3VPN services, in which entropy label represents one or a group recognized IP flows.

The same approach applies to Application Label [RFC4928] as well.

7. Security Considerations

For the further study.

8. IANA Considerations

IANA is for the further study.

9. References

9.1. Normative References


[RFC3985] Bryant, S., Pate P., "Multiprotocol Label Switching (MPLS) Label Stack Entry: 'EXP' Field Renamed to 'Traffic Class' Field", RFC3985, March 2005


9.2. Informative References


[FAT-PW] Bryant, S., Drafz, U Kompella, V., etc, "Flow Aware Transport of Pseudowires over an MPLS PSN", draft-ietf-pwe3-fat-pw-04, July. 2010

Acknowledgments

Authors like to thank Stewart Bryan, Frederic Jounay, Simon Delord, Raymond Key for the review and suggestions.
Appendix A. Enhanced ECMP Example

Label switched routers can implement the enhanced ECMP for distributing flows over ECMP paths. The enhanced ECMP process separates the packets that belong to a large flow from the packets that belong to a small flow and applies different treatments on these two types of packets. The process uses hashing to select ECMP paths for all the small flow packets and uses a large flow table to select the path for all the large flow packets. Figure 3 illustrates the enhanced ECMP processing diagram.

```
+-------------+    | 4 ECMP Paths
       |    |
+-------------+    |    |
       |    | Small-Flow
       |---->| Forwarding
       |     | Process
       |     +=============

Packets| Packet | +------------+    |
-------|-------|------------|    |
        |       |   +-------------+    |
        |       |   | Small-Flow |
        |       |   | Forwarding  |
        |       |   | Process     |
        |       |   +-------------+    |

       |    | Packet Separation
       |---->|   +-------------+    |
       |     |   | Small-Flow |
       |     |   | Forwarding  |
       |     |   | Process     |
       |     |   +-------------+    |

       |    | Large-Flow
       |---->| Forwarding
       |     | Process
       +-------------+    |
```

Figure 3 Enhanced ECMP Process Diagram

Figure 3 depicts three function elements. There are four equal cost paths shown as an example. Small-Flow Forwarding Process is used for forwarding all the small flow packets, which can be the same as existing ECMP process. Packet Separation Process and Larger-Flow Forwarding Process are the new elements in the enhanced ECMP proposed in this document.

The Packet Separation Process receives all the transported packets and evaluates all the income packets; it uses the first nibble to distinguish labeled packets or IP packets. Following is the algorithm for the packet separation process.

If the first nibble on the packet header == 0 or 1 then

```
// Labeled Packet
Find the bottom label
```
If TTL of the bottom label on the packet == 0 then

// Flow Label packet

If Large Flow Classification bit == 1 then

Send packet to large-flow forwarding process

Else

Send packet to small-flow forwarding process

End If

Else

// non-flow-labeled packet

Send the packet to small-flow forwarding process

// option: operator may want to apply large-flow forwarding process to it

End If

Else If the first nibble on the packet header == 0x6

// IPv6 packet

If Large Flow Classification bit == 1 then

Send the packet to large-flow forwarding process

Else

Send the packet to small-flow forwarding process

End If

Else // Ipv4 packet

Send to small-flow forwarding process

End If
Large-Flow Forwarding Process uses a flow table for packet forwarding. The flow table has an entry for each "live" flow. When the process receives a packet, it retrieves the flow ID from the packet and performs the table lookup by using flow ID. It forwards the packets to the path indicated in the table. If the process does not find an entry that matches the flow ID on a packet, it calls the path selection algorithm. The algorithm can select a path for the flow, say A, based on current path load, i.e. select the path that has least load at the time. Then the process forwards the packet to the selected path and inserts a new entry for the flow A in the table. The following packets of flow A will be forwarded to the path indicated in the table. When a flow is transported completely, the process no longer receives the packets that belong to the flow; the age function in the process can delete the flow entry from the table, which prevents the table size from the unnecessary growth. The age process frequency is configurable based on operation needs. If one of ECMP paths is down the algorithm will map impacted large flows to other ECMP paths. If a new ECMP path is added, the new flows can be assigned to the new path; it is optional for the process to perform the "live" large flow reassignment since the "live" flows may disappear itself anyway. The design method of Large-Flow Forwarding Process is outside the scope of this document.

Besides above example, another example is to split ECMP paths into two groups, and have one group for small flows and one for large flows; use hash for small flows and round robin for the large flows. Other implementation are possible as well.

10.1. Congestion Control

The enhanced ECMP also brings an advance in congestion control. The congestion happens when the traffic volume exceeds the path load threshold configured by operators. Since the large flows contribute much more volume than small flows, changing few large flows can efficiently rescue the congestion condition and keep the rest of services running normally. As a result, the congestion control only impacts few services. Large-Flow Process can easily select the large flows and send them to other paths or block them during the congestion. Flow QoS can be considered in the selection process. Notes: flow QoS indication is encoded on the top label of the label stack. Whether it is worth to cache these blocked flows or not is for further study.

10.2. Flow Rate Difference

The enhanced ECMP method uses the different treatments between large flows and small flows. Neither of treatments considers individual
flow rates in the routing process. This is because that improved load balance is achieved by hashing on the small flows and selecting the least loaded path for a new ‘live’ flow. The latter distribution using few large flows effectively compensates the uneven balance caused by the hashing and it does not need to consider individual flow rate. This is nice for the enhanced ECMP to keep the nature of statistical balancing. Therefore, the enhanced ECMP method works well even flow rates are not balanced.

10.3. Backward Compatibility

The method fully supports backward compatibility in PSN. If ingress PE does not support Large Flow Recognition, it SHALL set large flow indication bit as 0. Then all the flows are treated as small flows in PSN. P routers with existing ECMP or enhanced ECMP capability use hashing to discriminate the flows and distribute those flows over ECMP paths. If ingress PE supports Large Flow Recognition, it will insert the indication on the flow label. The P routers with existing ECMP capability will ignore the indication and just perform hashing on all the flows. The P routers with enhanced ECMP capability will separate the large and small flows and perform different treatments as proposed in this document. Although P router with existing ECMP capability gets uneven load balancing over its ECMP paths, it maintains the same performance as today’s network. If ingress PE does not support the flow label on PW, when enhanced ECMP applies, it will treat PW packets or LSP packets as small flow packets.
Appendix B. Simulation Analysis

We create Internet Traffic Generator based on observed Internet Traffic pattern. The generator randomly generates 98% of small traffic flows and 2% of large traffic flows up to 10G traffic. The traffic volume for the large flows and small flows are 30% and 70%. Simulator uses hash based distribution to disperse the traffic over 4 paths and 10 paths, respectively; and also uses enhanced ECMP method to disperse the traffic over 4 paths and 10 paths. The results show the performance between ECMP and enhanced ECMP from 6 simulations. Enhanced ECMP gets <1% load differences among paths while ECMP have up to 15% load differences. It shows how the simple distribution on few large flows can effectively compensate the uneven load balance caused by hashing and the traffic pattern.

We also use carrier suggested traffic matrix for the simulation, the improvement even better because carrier network has some significant high bit rate flows whose rate is much higher than the rate generated by any application.
Authors’ Addresses

Lucy Yong
Huawei Technologies Co., Ltd.
1700 Alma Dr.
Plano, TX 75075
US

Phone: +14692295387
Email: lucyyong@huawei.com

Peilin Yang
Huawei Technologies Co., Ltd.
No.91, Baixia Road, Nanjing 210001
P. R. China

Phone: +86-25-84565881
EMail: yangpeilin@huawei.com