Network Working Group
Internet Draft
Intended status: Standards Track
Expires: Sept. 2010

L. Yong Ed.
P. L. Yang
Huawei
March 5, 2010

Enhanced ECMP and Large Flow Aware Transport
draft-yong-pwe3-enhance-ecmp-lfat-01.txt

Status of this Memo

Copyright Notice

Abstract

Internet Traffic has constantly shown the pattern that a very small amount of the traffic flows generate a high traffic volume while a significant amount of small flows contribute a small amount of traffic volume. Differentiating such large flow and small flow in the packet switched network enables an enhanced transport method over Equal Cost Multi Paths (ECMP). This draft describes the enhanced ECMP transport with the large flow awareness.

Table of Contents

1. Introduction

[FAT-PW] introduces the flow label on the label stack for some 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. 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 each individual transported IP flow.

However, the traffic over Internet today includes Web browsing data and audio as well as video/downloading and streaming. Video/downloading and streaming generates the very high rate flows compared to Web browsing data/audio. This causes Internet traffic clearly mixed with huge amount of small flows and small amount of very high rate flows. Internet traffic analysis [CAIDA] indicates that, today, ~2% of the top rate ranked flows takes about 30% of traffic volume while the rest of 98% flows contribute 70% of traffic volume. As Web HDTV and 3D TV will be on the Internet, the traffic volume ratio between large and small flows may be further higher. Although the flow label can improve the load balancing per the flow basis within a pseudowire, under such traffic pattern, hash based distribution is inadequate for satisfactory load balancing.

Hash based distribution ensures any flow to be mapped into only one of ECMP paths (fixed one) so the flow ordering is preserved in the transport. However, hash based distribution disperses all the possible flow identifiers over ECMP paths no matter a flow exists or not at the time and does not consider individual flow rate, i.e. it has the nature of stateless distribution. Such distribution method generates adequate load balancing if the traffic contains huge amount of flows that have similar flow rates. The simulation has shown that given Internet traffic pattern, the hash method does not evenly distribute the flows over ECMP paths. The load difference between two of ECMP paths can be significant large; the more ECMP paths exist, the more severe the un-balancing syndrome presents. This implies that hash based distribution can cause some path congested and some being partial filled only. This results that the packets are dropped at congestion point and the network reroute impacted traffic. In other words, this syndrome lowers the network performance and brings operator desires to improve load balancing over ECMP. One option to prevent such syndrome is to add more transport resource in the network. But this will lower the network utilization and increase the service cost.

This draft describes an enhanced ECMP method for such traffic pattern and also introduces the large/small flow indication on the flow label to facilitate enhanced ECMP transport in PSN. The enhanced method uses a table for a small amount of large flow distribution and hashing on all other flows. The method gets evenly load balance by maintaining a small set of large flow

states. The draft states the process procedures on PE and P routers.

The simulation result has shown that the enhanced ECMP gets much better improvement on load-balancing compared to hash based ECMP under Internet traffic pattern. The load difference among paths is less than 1%.

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

Large Flow: a group of packets that contain the same "identity" in their header and come to the network at a high rate, i.e. the packet volume per time is high.

Small Flow: a group of packets that contain the same "identity" in their header and come to the network at a low rate, i.e. the packet volume per time is very low. Single packet can be considered as a special small flow in the context of this document.

3. Large Flow Recognition

The high technology now enables router devices to inspect the received the packets and identifies the large flows from huge amount of packets that belong to many flows (both large and small). Large flow recognition process may use protocol inspection, flow volume measurement, or other methods to detect the large flows. If a router can differentiate packets that belong to the high rate flows from all the received packets, it can perform differentiated transports for the large flows and small flows in PSN as described in section 4.

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.

Typically, a large flow has the context for an entire packet switched network. It has obvious benefit that if ingress PE performs the large flow recognition and inserts a large flow indication on the packets, then all the P nodes within PSN can

distinguish the large flow packets by checking this indication. This can largely 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 the service, and insert the flow (group) identity of each packet before it is passed to the pseudowire forwarder. When ingress PE performs large flow recognition, the pseudowire forwarder [RFC3985] can perform packet inspection and detect the large flow packets. The design method for the large flow recognition is outside the scope of this document. The pseudowire forwarder can insert a large flow indication on all the packets that belong to the large flow once it is recognized as a large flow. The large flow indication encoding schema is described in section 5. Since a large flow comes and disappears when it is transported completely, the list of large flows could dynamically change.

Large Flow Recognition has the assumption that a large flow sustains for certain time on the network. This assumption applies video, streaming, and file download applications. Although application rate may vary over the time, its lowest rate value is still much high compared to the small flows. Operator can set the large flow criteria.

It is worth to mention that a large flow recognition process may or may not need a time to recognize a large flow. If it needs and even the time is very short, during this period, some packets belonging to a large flow may be treated as small flow packets, which may cause the packets for a large flow traversing different paths during the transition. Thus this may cause a bit packet disordering at a destination. To prevent the packet disordering, Large Flow Recognition Process can use some temporary caching technology to hold the large flow packets for short time at the time the flow is recognized as a new large flow. Another factor to consider is that today packet based applications at the end points normally have a buffer to deal with packet delay variance and loss/disordering, thus the seldom disordering during transport is no longer a BIG issue for Internet traffic. Some large flow recognition may not need time to detect the large flow; it does not generate the ordering issue.

4. Enhanced ECMP Process

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 the path from equal cost multi paths for all the small flow packets and uses a large flow table to select the path for all the large flow packets. Figure 1 illustrates the enhanced ECMP processing diagram.

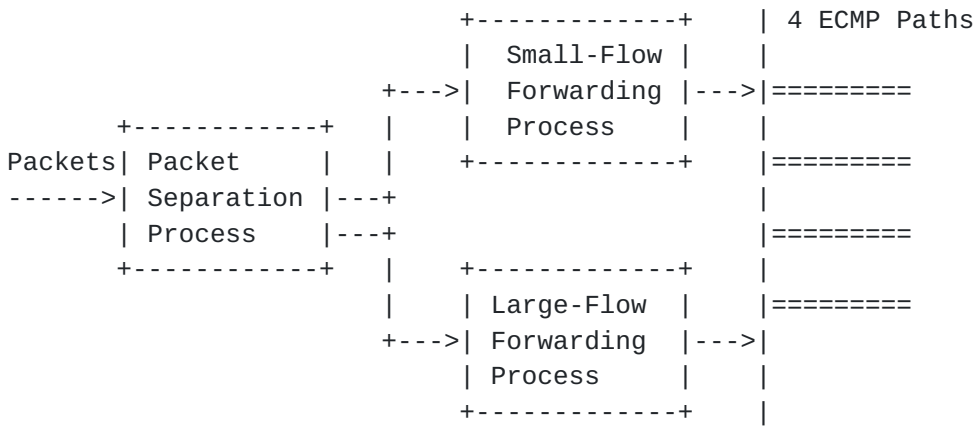


Figure 1 Enhanced ECMP Process Diagram

Figure 1 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. If a labeled packet is marked as a large flow, it will be sent to Large-Flow Forwarding Process; if not, it will be sent to Small-Flow Forwarding Process. As a result, the small flow transport path will be determined by hashing method; the large flow transport path will be determined by Large-Flow Forwarding Process. Since this draft focuses on the labeled packets, IP packet process is described in section 7.5.

Since the bottom label at the label stack (S bit = 1) can be PW label, LSP label, or Flow Label, it is necessary for Packet Separation Process to differ Flow Label from PW Label and LSP Label. Since Flow Label is never on the top of the label stack and is not processed by the forwarder, it has its unique position. The draft suggests setting FL TTL to 0. Thus, when S

bit =1 and TTL = 0, the label is Flow Label. Otherwise, it is either PW Label or LSP label.

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.

Note: Large-Flow Forwarding Process can work with any hash-key generation scheme. Large-Flow distribution method using few large flows effectively compensates the uneven distribution caused by hashing and traffic pattern. It is worth to mention the differences between DiffDerv and large and small flow differentiation. Although both relate to traffic classification, they aim the different purposes. DiffDerv is for the network to perform differentiated service treatments. Large and small flow differentiation is to improve the network utilization in ECMP.

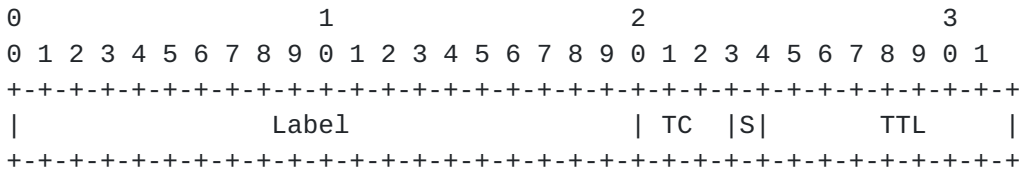
4.1. Congestion Control

The enhanced ECMP also brings an advance in congestion control. The congestion happens when the traffic volume exceeds the path capacity. Since the large flows take much more bandwidth, dropping 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 block them during the congestion. Whether it is worth to cache these blocked

flows or not is for further study and outside the scope of this document.

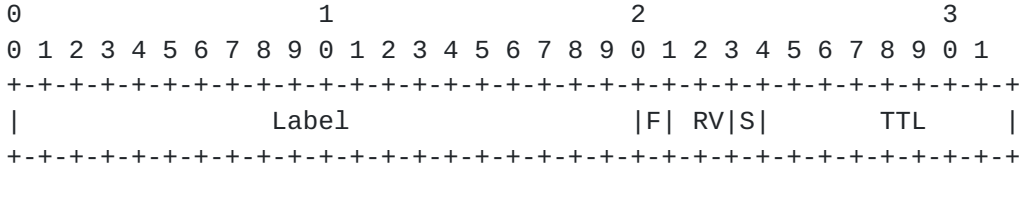
5. Large Flow Indication

This draft specifies the protocol to encode a large flow indication on the flow label specified in [FAT-PW]. Figure 2 illustrates current flow label format [RFC3031] with the amendment given in [RFC5462]. Label field is filled with the flow identity. Since the flow label is never on the top of label stack, TTL field is not used. However, to prevent any provisioning error, TTL filed is recommended to set as 1. S bit is used to indicate the bottom of stack and set to 1 for the flow label. 3 Traffic Class bits are not used in current ECMP processing now. The document suggests using the first bit in the Traffic Class bits to indicate the large flow or small flow, and suggests value 1 for the large flow and value 0 for the small flow. The two other bits reserve for the future. Figure 3 shows proposed format.



Label: Label Value, 20 bits
TC: Traffic Class, 3 bits;
S: Bottom of Stack, 1 bit
TTL: Time to Live, 8 bits

Figure 2 Current Flow Label Formant



Label: Label Value, 20 bits
F: Flow Characteristics Indication, 1 bit;

```
RV:     Reserved Bit, set to 0
S:      Bottom of Stack, 1 bit
TTL:    Time to Live, 8 bits
```

Figure 3 Flow Label Format with Large-Flow Indication

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 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 purpose. [RFC3270] [RFC5129] Since Flow Label is never used in the top of Label Stack, DiffServ function does not apply to the Flow Label. Hence the proposed Flow Label TC bit usage in this document does not impact DiffServ function.

6. Backward Compatibility

The enhanced ECMP fully support backward compatibility in PSN. If ingress PE does not support Large Flow Recognition, it SHALL set flow label F bit to 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.

7. Applicability

Carriers have desires to improve transport network capability via certain service awareness in packet transport and not be constrained in just "pipe" transport service.[FAT-PW]brings such potential by introducing the flow label in the label stack, which enables ECMP transport discriminates traffic at flow granularity. The large flow aware transport further enables ECMP transport to distinguish the large and small flows and perform different

treatments on two types of flows, which can improve the load balancing when traffic pattern contains very small percentage of large flows.

The method described in this document requires the new capability from 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 P or PE routers perform the enhanced ECMP function. Since 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. This allows operator to gradually upgrade the network.

7.1. Link Aggregation Groups

A Link Aggregation Group (LAG) is used to bond together several physical circuits 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 enhanced ECMP proposed in this document can assist in producing a more uniform flow distribution and controlling the congestion in LAG.

7.2. 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 for a single large flow. Then ingress PE can insert a large flow indication. P routers in PSN can treat it as a large flow.

7.3. Flow Rate Difference

The enhanced ECMP method uses the different treatments between large flows and small flows. Neither of treatments considers the flow rate in the distribution process. This is because that even load balance is achieved by hashing on the small flows and selecting the least used the path for a new "live" flow. The latter distribution using few large flows effectively compensates the uneven balance caused by the former and is unnecessary to consider individual flow rate. This is nice that enhanced ECMP keeps the nature of statistical balancing. Therefore, the enhanced ECMP method works well even flow rates are broad.

7.4. Multi-Segment Pseudowires

The flow label mechanism described in this document works on multi-segment PWs [MS-PW] without requiring 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.

7.5. IP Flows

Today's ECMP method applies to both IP flows and MPLS labeled flows 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. 3-tuple: source, destination, and flow label is used in ECMP.[RFC3697] 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) to indicate small and large flow. By default, all TC bits are set as 0 [RFC2460], which is compatible to this solution. 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 indication 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 and small IP flow distinction, the enhanced ECMP treats all IP flows as small flows.

7.6. LSP with Entropy Label

Entropy Label [Entropy] is inserted in LSP traffic at ingress LSR to gain better ECMP load balancing at transit LSRs. Entropy label is very similar as PW flow label and is used to differentiate "microflow" within a LSP so ECMP process can get better dispersion granularity. Enhanced ECMP and Large Flow Aware Transport can apply to LSP with entropy label. Traffic class field in the Entropy can use the same encoding scheme described in this document. If ingress LSR does not support large flow recognition, then it SHOULD set Large Flow indication bit to 0.

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

8. Security Considerations

Since the number of large flows is very small compared to the number of small flows; packet switched routers only need to maintain a small size of table or flow states. Notes operator can use the large flow criteria to control the large flow volume. The method won't create the scalability and performance issue.

9. IANA Considerations

IANA is for the further study.

10. References

10.1. Normative References

[RFC2460] Deering, S., Hinden, R., "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1995.

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

[RFC3031] Rosen, E., Viswanathan, A., and Callon, R., "Multiple protocol Label Switching Architecture", RFC3031, January 2001.

[RFC3270] Faucheur F. Le, etc, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC3270, May 2002

[RFC3697] Rajahalme, J., Conta, A., Carpenter, B., and S. Deering, "IPv6 Flow Label Specification", RFC 3697, March 2004.

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

[RFC4928] Swallow, G., Bryant, S., and Andersson, L., "Avoiding Equal Cost Multipath Treatment in MPLS Network", RFC4928, June 2007.

[RFC5129] Davie, B., Briscoe, B., and Jay, J., "Explicit Congestion Marking in MPLS", RFC5129, January 2008

[RFC5462] Andersson, L. etc, "Multiprotocol Label Switching (MPLS) Label Stack Entry: EXP Field Renamed to Traffic Class Field", October 2009

10.2. 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-03, (work in progress), Jan. 2010

[Entropy] Kompella K, Amante S., "The use Entropy Labels in MPLS Forwarding", draft-kompella-mpls-entropy-label-01, January 2009

[MS-PW] Bocci, M. Bryant, S., "An Architecture fro Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC5659 October 2009

[FLOW-ECMP] Carpenter B., "Using the IPv6 flow label for equal cost multipath routing in tunnels", draft-carpenter-flow-ecmp-01, February 18, 2010

[CAIDA] Caida Internet Traffic Analysis, www.caida.org/data/monitor

11. Acknowledgments

Authors like to thank Stewart Bryan, Frederic Jounay, Simon Delord, Raymond Key for the review and suggestions.

Appendix A. 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.

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