OPSAWG R. Krishnan
Internet Draft S. Khanna
Intended status: Informational Brocade Communications
Expires: March 2013 September 23, 2012

Best Practices for Optimal LAG/ECMP Component Link Utilization in Provider Backbone networks

draft-krishnan-opsawg-large-flow-load-balancing-00.txt

Status of this Memo

Copyright Notice

Abstract

The demands on the networking infrastructure are growing exponentially; the drivers are bandwidth hungry rich media applications, inter data center communications etc. In this context, it is important to optimally use the bandwidth in the service provider backbone networks which extensively use LAG/ECMP techniques for bandwidth scaling. This internet draft describes the issues faced in the service provider backbone in the context of LAG/ECMP and formulates best practice recommendations for managing the bandwidth efficiently in the service provider backbone.

Table of Contents

1. Introduction.....3
2. Conventions used in this document.....3
3. Sub-optimal LAG/ECMP Component Link Utilization in the current framework.....4
4. Best practices for optimal LAG/ECMP Component Link Utilization.5
4.1. Long-lived Large Flow Identification.....7
4.1.1. Sflow/Netflow.....7
4.1.2. Automatic hardware identification.....8
4.1.2.1. Suggested Technique for Automatic Hardware Identification.....8
4.2. Long-lived Large Flow Re-balancing.....9
4.2.1. No re-balancing of short-lived small flows.....9
4.2.2. Other Techniques.....9
4.2.3. Re-balancing of long-lived large flows and short-lived small flows - an example.....9
5. Acknowledgements.....11
6. References.....12
6.1. Normative References.....12
6.2. Informative References.....12

1. Introduction

Service provider backbone networks extensively use LAG/ECMP techniques for bandwidth scaling. Network traffic can be predominantly categorized into two traffic types, long-lived large flows and short-lived small flows. Hashing techniques, which perform an approximate distribution of these flows across the LAG/ECMP component links, typically result in a sub-optimal utilization of LAG/ECMP component links. Round Robin load-balancing techniques address this problem but have the side effect of causing packet re-ordering. This internet draft recommends best practices for optimal LAG/ECMP component link utilization while using hashing techniques. These best practices comprise of the following; first is identification of long-lived large flows in routers and next is assigning the long-lived large flows to specific LAG/ECMP component links.

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 RFC 2119 [RFC2119].

3. Sub-optimal LAG/ECMP Component Link Utilization in the current framework

Hashing techniques, which perform an approximate distribution of long-lived large flows and short-lived small flows across the LAG/ECMP component links, typically results in a sub-optimal utilization of LAG/ECMP component links. This is depicted in Figure 1 with a detailed description below.

- There is a LAG between 2 routers R1 and R2. This LAG has 3 component links (1), (2), (3)
- Component link (1) has 2 short-lived small flows and 1 long-lived large flow and the link capacity is optimally utilized
- Component link (2) has 3 short-lived small flows and no long-lived large flow and the link capacity is sub-optimally utilized
  - The absence of any long-lived large flow causes the component link under-utilization
- Component link (3) has 2 short-lived small flows and 2 long-lived large flows and the link capacity is over-utilized.
  - The presence of 2 long-lived large flows causes the component link over-utilization

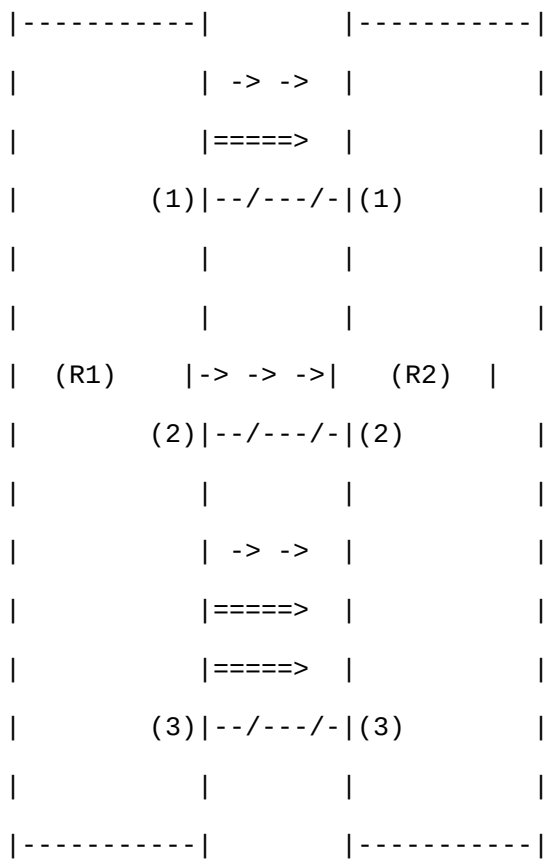

Figure 1: Long-lived Large Flows - uneven distribution across

LAG/ECMP component links

4. Best practices for optimal LAG/ECMP Component Link Utilization

The suggested techniques in this draft for optimal LAG/ECMP component link utilization are meant to put forth a locally_ optimized solution, i.e. local in the sense of both measuring and optimizing

for long-lived large flows at individual nodes in the network. This approach would not yield a globally optimal placement of a large, long-lived flow across several nodes in the network which some networks may desire/require. On the other hand, this may be adequate for some operators for the following reasons 1) Different links in the network experience different levels of utilization and, thus, a more "targeted" solution is needed for those few hot-spots in the network 2) Some networks may lack end-to-end visibility.

The various steps in achieving optimal LAG/ECMP component link utilization in backbone networks are detailed below

Step 1) This involves identifying long-lived large flows in the egress processing elements in routers; besides the flow parameters, this also involves identifying the egress component link the flow is using. The identification of long-lived large flows is explained in detail in section 4.1.

Step 2) The egress component links are periodically scanned for link utilization. If the egress component link utilization exceeds a pre-programmed threshold, an operator alert is generated. The long-lived large flows mapping to the congested egress component link are exported to a central management entity. IETF could potentially consider a standards-based activity around, say, a data-model used to move this information from the router to the central management entity.

Step 3) On receiving the alert about the congested component link, the operator, through a central management entity finds out the long-lived large flows mapping to the component link and the LAG/ECMP group to which the component link maps to.

Step 4) The operator can choose to rebalance the long-lived large flows on lightly loaded component links of the LAG/ECMP group. The operator, through a central management entity 1) Can indicate specific long-lived large flows to rebalance 2) Let the router decide the best long-lived large flows to rebalance. The central management entity conveys the above information to the router. IETF could potentially consider a standards-based activity around, say, a data-model used to move this information from the central management entity to the router. The re-balancing of long-lived large flows is explained in detail in section 4.2.

4.1. Long-lived Large Flow Identification

A flow (long-lived large flow or short-lived small flow) can be defined using one of the following suggested formats as described below

- IP 5 tuple: IP Protocol, IP source address, IP destination address, TCP/UDP source port, TCP/UDP destination port
- IP 3 tuple: IP Protocol, IP source address, IP destination address
- MPLS Labels
- VXLAN, NVGRE
- Other formats

The best practices described in this document are agnostic to the format of the flow.

4.1.1. Sflow/Netflow

Enable Sflow/Netflow sampling on all the egress ports in the routers. Through Sflow processing in a Sflow Collector, an approximate indication of large flows mapping to each of the component links in each LAG/ECMP group is available. The advantages and disadvantages of sFlow/Netflow are detailed below.

Advantages of Sflow/Netflow

- Supported in most routers
- Minimal router resources

Disadvantages of Sflow/Netflow

- Approximate identification of long-lived large flows
- Non real-time identification of long-lived large flows based on historical analysis

The time taken to determine a candidate long-lived large flow would be dependent on the amount of sFlow samples being generated and the processing power of the external sFlow collector; this is under further study.

4.1.2. Automatic hardware identification

Implementations may choose to implement automatic identification of long-lived large flows in hardware in egress processing elements of routers. The characteristics of such an implementation would be

- Inline solution
- Minimal system resources
- Maintain line-rate performance
- Perform accounting of long-lived large flows with a high degree of accuracy

Using automatic hardware identification of long-lived large flows, an accurate indication of large flows mapping to each of the component links in a LAG/ECMP group is available. The advantages and disadvantages of automatic hardware identification are detailed below.

Advantages of Automatic Hardware Identification

- Accurate identification of long-lived large flows
- Real-time identification of long-lived large flows

Disadvantages of Automatic Hardware Identification

- Not supported in most routers

The measurement interval for determining a candidate long-lived large flow and the minimum bandwidth of the long-lived large flow would be programmable parameters in the router; this is under further study.

The implementation of automatic hardware identification of long-lived large flows is vendor dependent. Below is a suggested technique.

4.1.2.1. Suggested Technique for Automatic Hardware Identification

There are multiple hash tables, each with a different hash function. Each hash table entry has an associated counter. On packet arrival, a new flow is looked up in parallel in all the hash tables and the corresponding counter is incremented. If the counter exceeds a programmed threshold in a given time interval in all the hash table entries, a candidate long-lived-flow is learnt and programmed in a

hardware table resource like TCAM. There may be some false positives due to multiple short-lived small flows masquerading as a long-lived large flow; the amount of false positives is reduced by parallel hashing.

4.2. Long-lived Large Flow Re-balancing

Below are suggested techniques for long-lived large flow re-balancing. Our suggestion is for the router vendors to implement all these techniques and let the operator choose the right technique based on various application needs.

4.2.1. No re-balancing of short-lived small flows

In the LAG/ECMP group, choose other member component links with least average port utilization. Move the long-lived large flow(s) from the heavily loaded component link to the new member component links using a Policy based routing (PBR) rule in the ingress processing element(s) in the routers. The benefits of this algorithm are

- Short-lived small flows are not subjected to flow re-ordering
- Only certain long-lived large flows are subjected to flow re-ordering

4.2.2. Other Techniques

It is possible use other algorithms, for example, removing a member component link from the LAG/ECMP group and using it only for long-lived large flows.

4.2.3. Re-balancing of long-lived large flows and short-lived small flows - an example

Optimal LAG/ECMP component utilization for the use case in Figure 1, is depicted below in Figure 2. This is achieved as follows

Step 1) Long-lived large flows are identified in the egress processing elements of router R1 using techniques suggested in Section 4.1.

Step 2) An operator alert is generated indicating that egress component link (3) in router R1 is congested. The long-lived large flows mapping to the congested egress component link are exported from the router to a central management entity.

Step 3) On receiving the alert about the congested component link (3), the operator, through a central management entity finds out the long-lived large flows mapping to the component link and the LAG/ECMP group to which the component link maps to.

Step 4) The operator, through a central management entity, can choose to rebalance the long-lived large flows on lightly loaded component links of the LAG/ECMP group using the suggested techniques in Section 4.2. In the router, a long-lived large flow is moved from component link (3) to component link (2) by using a PBR rule in the ingress processing element(s) in the routers.

Detailed description for Figure 2 is as follows

- There is a LAG between 2 routers R1 and R2. This LAG has 3 component links (1), (2), (3)
- Component link (1) has 2 short-lived small flows and 1 long-lived large flow and the link capacity is optimally utilized
- Component link (2) has 3 short-lived small flows and 1 long-lived large flow and the link capacity is optimally utilized
- Component link (3) has 2 short-lived small flows and 1 long-lived large flow and the link capacity is optimally utilized

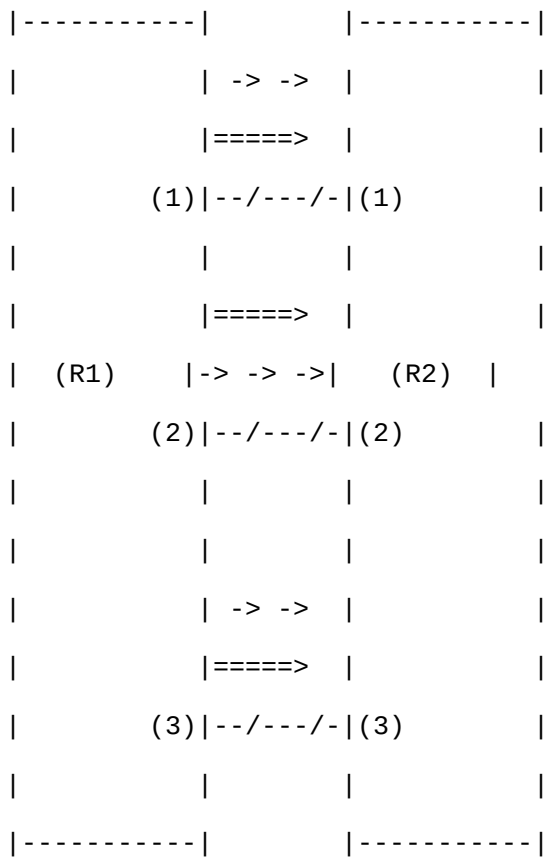

Figure 2: Long-lived Large Flows - even distribution across LAG/ECMP component links

5. Acknowledgements

The authors would like to thank Shane Amante for his input.

6. References

6.1. Normative References

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

[2] Crocker, D. and Overell, P.(Editors), "Augmented BNF for Syntax Specifications: ABNF", RFC 2234, Internet Mail Consortium and Demon Internet Ltd., November 1997.

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

[RFC2234] Crocker, D. and Overell, P.(Editors), "Augmented BNF for Syntax Specifications: ABNF", RFC 2234, Internet Mail Consortium and Demon Internet Ltd., November 1997.

6.2. Informative References

[I-D.ietf-rtgwg-cl-requirement] C. Villamizar et al., "Requirements for MPLS Over a Composite Link", June 2012

[I-D.ietf-mpls-entropy-label] K. Kompella et al., "The Use of Entropy Labels in MPLS Forwarding", July 2012

Authors' Addresses

Ram Krishnan

Brocade Communications

San Jose, 95134, USA

Phone: +001-408-406-7890

Email: ramk@brocade.com

Sanjay Khanna

Brocade Communications

San Jose, 95134, USA

Phone: +001-408-333-4850

Email: skhanna@brocade.com