OPSAWG
Internet Draft
Intended status: Informational
Expires: October 2013
April 24, 2013

R. Krishnan
S. Khanna
Brocade Communications
L. Yong
Huawei USA
A. Ghanwani
Dell
Ning So
Tata Communications
B. Khasnabish
ZTE Corporation

Mechanisms for Optimal LAG/ECMP Component Link Utilization in Networks

draft-krishnan-opsawg-large-flow-load-balancing-08.txt

Status of this Memo

Copyright Notice

Abstract

Demands on 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 wired networks that extensively use LAG/ECMP techniques for bandwidth scaling. This draft explores some of the mechanisms useful for achieving this.

Table of Contents

1. Introduction

Networks extensively use LAG/ECMP techniques for capacity 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). Stateless hash-based techniques [ITCOM, RFC 2991, RFC 2992, RFC 6790] are often used to distribute both long-lived large flows and other flows over the component links in a LAG/ECMP. However the traffic may not be evenly distributed over the component links due to the traffic pattern.

This draft describes best practices for optimal LAG/ECMP component link utilization while using hash-based techniques. These best practices comprise the following steps -- recognizing long-lived large flows in a router; and assigning the long-lived large flows to specific LAG/ECMP component links or redistributing other flows when a component link on the router is congested.

It is useful to keep in mind that the typical use case is where the long-lived large flows are those that consume a significant amount of bandwidth on a link, e.g. greater than 5% of link bandwidth. The number of such flows would necessarily be fairly small, e.g. on the order of 10's or 100's per link. In other words, the number of long-lived large flows is NOT expected to be on the order of millions of flows. Examples of such long-lived large flows would be IPSec tunnels in service provider backbones or storage backup traffic in data center networks.

1.1. Acronyms

COTS: Commercial Off-the-shelf

DOS: Denial of Service

ECMP: Equal Cost Multi-path

GRE: Generic Routing Encapsulation

LAG: Link Aggregation Group

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

1.2. 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. Hash-based Load Distribution in LAG/ECMP

Hashing techniques are often used for traffic load balancing to select among multiple available paths with LAG/ECMP. The advantages of hash-based load distribution are the preservation of the packet sequence in a flow and the real-time distribution without maintaining per-flow state in the router. Hash-based techniques use a combination of fields in the packet's headers to identify a flow, and the hash function on these fields is used to generate a unique number that identifies a link/path in a LAG/ECMP. The result of the hashing procedure is a many-to-one mapping of flows to component links.

If the traffic load constitutes flows such that the result of the hash function across these flows is fairly uniform so that a similar number of flows is mapped to each component link, if, the individual flow rates are much smaller as compared to the link capacity, and if the rate differences are not dramatic, the hash-based algorithm produces good results with respect to utilization of the individual component links. However, if one or more of these conditions are not met, hash-based techniques may result in unbalanced loads on individual component links.

One example is illustrated in Figure 1. In the figure, there are two routers, R1 and R2, and there is a LAG between them which has 3 component links (1), (2), (3). There are a total of 10 flows that

need to be distributed across the links in this LAG. The result of hashing is as follows:

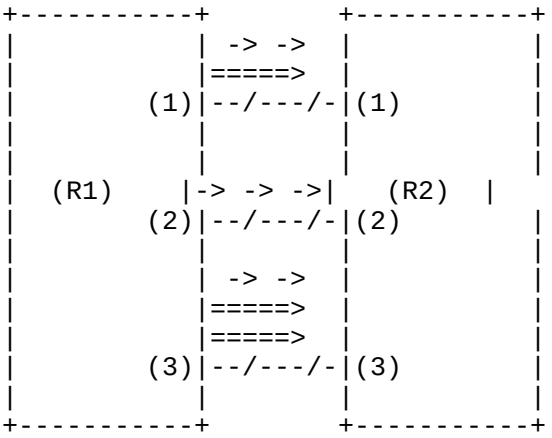
- Component link (1) has 3 flows -- 2 small flows and 1 large flow -- and the link utilization is normal.

- Component link (2) has 3 flows -- 3 small flows and no large flow -- and the link utilization is light.

  - The absence of any large flow causes the component link under-utilized.

- Component link (3) has 4 flows -- 2 small flows and 2 large flows -- and the link capacity is exceeded resulting in congestion.

  - The presence of 2 large flows causes congestion on this component link.



Where: ->-> small flows
===> large flow

Figure 1: Unevenly Utilized Component Links

This document presents improved load distribution techniques based on the large flow awareness. The techniques compensate for unbalanced load distribution resulting from hashing as demonstrated in the above example.

3. Mechanisms for Optimal LAG/ECMP Component Link Utilization

The suggested techniques in this draft are about a local optimization solution; they are local in the sense that both the identification of large flows and re-balancing of the load can be accomplished completely within individual nodes in the network without the need for interaction with other nodes.

This approach may not yield a globally optimal placement of large flows across multiple nodes in a network, which may be desirable in some networks. On the other hand, a local approach may be adequate for some environments for the following reasons:

1) Different links within a network experience different levels of utilization and, thus, a "targeted" solution is needed for those hot-spots in the network. An example is the utilization of a LAG between two routers that needs to be optimized.

2) Some networks may lack end-to-end visibility, e.g. when a certain network, under the control of a given operator, is a transit network for traffic from other networks that are not under the control of the same operator.

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

Step 1) This involves large flow recognition in routers and maintaining the mapping of the large flow to the component link that it uses. The recognition of large flows is explained in Section 3.1.

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 large flows mapped to the congested egress component link are exported to a central management entity.

Step 3) On receiving the alert about the congested component link, the operator, through a central management entity, finds the large flows mapped to that component link and the LAG/ECMP group to which the component link belongs.

Step 4) The operator can choose to rebalance the large flows on lightly loaded component links of the LAG/ECMP group or redistribute the small flows on the congested link to other component links of the group. The operator, through a central management entity, can choose one of the following actions:

1) Indicate specific large flows to rebalance;

2) Have the router decide the best large flows to rebalance;

3) Have the router redistribute all the small flows on the congested link to other component links in the group.

The central management entity conveys the above information to the router. The load re-balancing options are explained in Section 3.2.

Steps 2) to 4) could be automated if desired.

Providing large flow information to a central management entity provides the capability to further optimize flow distribution at with multi-node visibility. Consider the following example. A router may have 3 ECMP nexthops that lead down paths P1, P2, and P3. A couple of hops downstream on P1 may be congested, while P2 and P3 may be under-utilized, which the local router does not have visibility into. With the help of a central management entity, the operator could redistribute some of the flows from P1 to P2 and P3 resulting in a more optimized flow of traffic.

The techniques described above are especially useful when bundling links of different bandwidths for e.g. 10Gbps and 100Gbps as described in [I-D.ietf-rtgwg-cl-requirement].

3.1. Large Flow Recognition

3.1.1. Flow Identification

A flow (large flow or small flow) can be defined as a sequence of packets for which ordered delivery should be maintained. 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 header: IP Protocol, IP source address, IP destination address, flow label (IPv6 only), TCP/UDP source port, TCP/UDP destination port.

. 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. The mechanisms described in this document are agnostic to the fields that are used for flow identification.

3.1.2. Criteria for Identifying a Large Flow

From a bandwidth and time duration perspective, in order to identify large flows we define an observation interval and observe the bandwidth of the flow over that interval. A flow that exceeds a certain minimum bandwidth threshold over that observation interval would be considered a large flow.

The two parameters -- the observation interval, and the minimum bandwidth threshold over that observation interval -- should be programmable in a router to facilitate handling of different use cases and traffic characteristics. For example, a flow which is at or above 10% of link bandwidth for a time period of at least 1 second could be declared a large flow [DevoFlow].

In order to avoid excessive churn in the rebalancing, once a flow has been recognized as a large flow, it should continue to be recognized as a large flow as long as the traffic received during an observation interval exceeds some fraction of the bandwidth threshold, for example 80% of the bandwidth threshold.

Various techniques to identify a large flow are described below.

3.1.3. Sampling Techniques

A number of routers support sampling techniques such as sFlow [sFlow-v5, sFlow-LAG], PSAMP [RFC 5475] and Netflow Sampling [RFC 3954]. For the purpose of large flow identification, sampling must be enabled on all of the egress ports in the router where such measurements are desired.

Using sflow as an example, processing in an sFlow collector will provide an approximate indication of the large flows mapping to each of the component links in each LAG/ECMP group. It is possible to implement this part of the collector function in the control plane of the router reducing dependence on an external management station, assuming sufficient control plane resources are available.

If egress sampling is not available, ingress sampling can suffice since the central management entity used by the sampling technique typically has multi-node visibility and can use the samples from an immediately downstream node to make measurements for egress traffic at the local node. This may not be available if the downstream device is under the control of a different operator, or if the downstream device does not support sampling. Alternatively, since sampling techniques require that the sample annotated with the packet's egress port information, ingress sampling may suffice. However, this means that sampling would have to be enabled on all ports, rather than only on those ports where such monitoring is desired.

The advantages and disadvantages of sampling techniques are as follows.

Advantages:

. Supported in most existing routers.

. Requires minimal router resources.

Disadvantages:

. In order to minimize the error inherent in sampling, there is a minimum delay for the recognition time of large flows, and in the time that it takes to react to this information.

With sampling, the detection of large flows can be done on the order of one second [DevoFlow].

3.1.4. Automatic Hardware Recognition

Implementations may perform automatic recognition of large flows in hardware on a router. Since this is done in hardware, it is an inline solution and would be expected to operate at line rate.

Using automatic hardware recognition of large flows, a faster indication of large flows mapped to each of the component links in a LAG/ECMP group is available (as compared to the sampling approach described above).

The advantages and disadvantages of automatic hardware recognition are:

Advantages:

- Large flow detection is offloaded to hardware freeing up software resources and possible dependence on an external management station.

- As link speeds get higher, sampling rates are typically reduced to keep the number of samples manageable which places a lower bound on the detection time. With automatic hardware recognition, large flows can be detected in shorter windows on higher link speeds since every packet is accounted for in hardware [NDTM]

Disadvantages:

- Not supported in many routers.

As mentioned earlier, the observation interval for determining a large flow and the bandwidth threshold for classifying a flow as a large flow should be programmable parameters in a router.

The implementation of automatic hardware recognition of large flows is vendor dependent and beyond the scope of this document.

3.2. Load Re-balancing Options

Below are suggested techniques for load re-balancing. Equipment vendors should implement all of these techniques and allow the operator to choose one or more techniques based on their applications.

Note that regardless of the method used, perfect re-balancing of large flows may not be possible since flows arrive and depart at different times. Also, any flows that are moved from one component link to another may experience momentary packet reordering.

3.2.1. Alternative Placement of Large Flows

Within a LAG/ECMP group, the member component links with least average port utilization are identified. Some large flow(s) from the heavily loaded component links are then moved to those lightly-loaded member component links using a PBR rule in the ingress processing element(s) in the routers.

With this approach, only certain large flows are subjected to momentary flow re-ordering.

When a large flow is moved, this will increase the utilization of the link that it moved to potentially creating unbalanced utilization

once again across the link components. Therefore, when moving large flows, care must be taken to account for the existing load, and what the future load will be after large flow has been moved. Further, the appearance of new large flows may require a rearrangement of the placement of existing flows.

Consider a case where there is a LAG compromising 4 10 Gbps component links and there are 4 large flows each of 1 Gbps. These flows are each placed on one of the component links. Subsequent, a 5-th large flow of 2 Gbps is recognized and to maintain equitable load distribution, it may require placement of one of the existing 1 Gbps flow to a different component link. And this would still result in some imbalance in the utilization across the component links.

3.2.2. Redistributing Small Flows

Some large flows may consume the entire bandwidth of the component link(s). In this case, it would be desirable for the small flows to not use the congested component link(s). This can be accomplished in one of the following ways.

This method works on some existing router hardware. The idea is to prevent, or reduce the probability, that the small flow hashes into the congested component link(s).

- The LAG/ECMP table is modified to include only non-congested component link(s). Small flows hash into this table to be mapped to a destination component link. Alternatively, if certain component links are heavily loaded, but not congested, the output of the hash function can be adjusted to account for large flow loading on each of the component links.

- The PBR rules for large flows (refer to Section 3.2.1) must have strict precedence over the LAG/ECMP table lookup result.

With this approach the small flows that are moved would be subject to reordering.

3.2.3. Component Link Protection Considerations

If desired, certain component links may be reserved for link protection. These reserved component links are not used for any flows in the absence of any failures.. In the case when the component link(s) fail, all the flows on the failed component link(s) are moved to the reserved component link(s). The mapping table of large flows to component link simply replaces the failed component link with the

reserved link. Likewise, the LAG/ECMP hash table replaces the failed component link with the reserved link.

3.2.4. Load Re-Balancing Example

Optimal LAG/ECMP component utilization for the use case in Figure 1 is depicted below in Figure 2. The large flow rebalancing explained in Section 3.2.1 is used. The improved link utilization is as follows:

- Component link (1) has 3 flows -- 2 small flows and 1 large flow -- and the link utilization is normal.
- Component link (2) has 4 flows -- 3 small flows and 1 large flow -- and the link utilization is normal now.
- Component link (3) has 3 flows -- 2 small flows and 1 large flow -- and the link utilization is normal now.

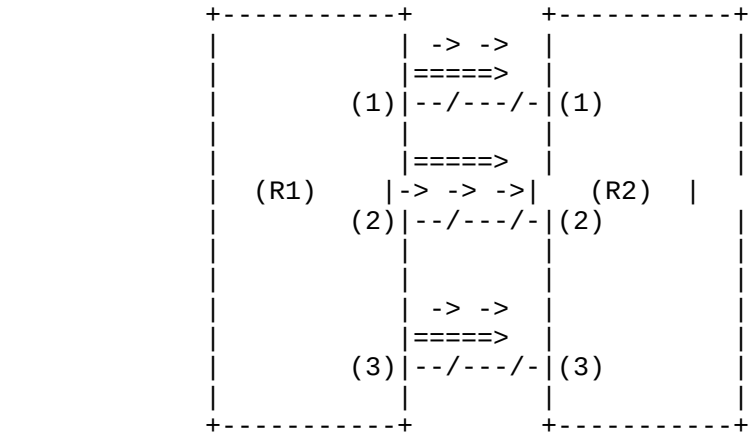


Figure 2: Evenly utilized Composite Links

Basically, the use of the mechanisms described in Section 3.2.1 resulted in a rebalancing of flows where one of the large flows on component link (3) which was previously congested was moved to component link (2) which was previously under-utilized.

4. Informat ion Model for Flow Re-balancing

4.1. Configuration Parameters

The following parameters are required the configuration of this feature:

. Large flow recognition parameters.

- o Observation interval: The observation interval is the time period in seconds over which the packet arrivals are observed for the purpose of large flow recognition.
- o Minimum bandwidth threshold: The minimum bandwidth threshold would be configured as a percentage of link speed and translated into a number of bytes over the observation interval. A flow for which the number of bytes received, for a given observation interval, exceeds this number would be recognized as a large flow.
- o Minimum bandwidth threshold for large flow maintenance: The minimum bandwidth threshold for large flow maintenance is used to provide hysteresis for large flow recognition. Once a flow is recognized as a large flow, it continues to be recognized as a large flow until it falls below this threshold. This is also configured as a percentage of link speed and is typically lower than the minimum bandwidth threshold defined above.

. Imbalance threshold: the difference between the utilization of the least utilized and most utilized component links. Expressed as a percentage of link speed.

4.2. Import of Flow Information

In cases where large flow recognition is handled by an external management station (see Section 3.1.3), an information model for flows is required to allow the import of large flow information to the router.

The following are some of the elements of information model for importing of flows:

- Layer 2: source MAC address, destination MAC address, VLAN ID.
- Layer 3 IP: IP Protocol, IP source address, IP destination address, flow label (IPv6 only), TCP/UDP source port, TCP/UDP destination port.
- MPLS Labels.

This list is not exhaustive. For example, with overlay protocols such as VXLAN and NVGRE, fields from the outer and/or inner headers may be specified. In general, all fields in the packet that can be used by forwarding decisions should be available for use when importing flow information from an external management station.

5. Operational Considerations

Flows should be re-balanced only when the imbalance in the utilization across component links exceeds a certain threshold. Frequent re-balancing to achieve precise equitable utilization across component links could be counter-productive as it may result in moving flows back and forth between the component links impacting packet ordering and system stability. This applies regardless of whether large flows or small flows are re-distributed.

The operator would have to experiment with various values of the large flow recognition parameters (minimum bandwidth threshold, observation interval) and the imbalance threshold across component links to tune the solution for their environment.

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

This document does not directly impact the security of the Internet infrastructure or its applications. In fact, it could help if there is a DOS attack pattern which causes a hash imbalance resulting in heavy overloading of large flows to certain LAG/ECMP component links.

8. Acknowledgements

The authors would like to thank the following individuals for their review and valuable feedback on earlier versions of this document: Shane Amante, Curtis Villamizar, Fred Baker, Wes George, Brian Carpenter, George Yum, Michael Fargano, Michael Bugenhagen, Jianrong Wong, Peter Phaal, Roman Krzanowski and Weifeng Zhang.

9. References

9.1. Normative References

9.2. Informative References

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

[RFC 6790] Kompella, K. et al., "The Use of Entropy Labels in MPLS Forwarding", November 2012.

[CAIDA] Caida Internet Traffic Analysis, http://www.caida.org/home.

[YONG] Yong, L., "Enhanced ECMP and Large Flow Aware Transport", draft-yong-pwe3-enhance-ecmp-lfat-01, September 2010.

[ITCOM] Jo, J., et al., "Internet traffic load balancing using dynamic hashing with flow volume", SPIE ITCOM, 2002.

[RFC 2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast", November 2000.

[RFC 2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", November 2000.

[RFC 5475] Zseby, T., et al., "Sampling and Filtering Techniques for IP Packet Selection", March 2009.

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

[sFlow-LAG] Phaal, P. and A. Ghanwani, "sFlow LAG counters structure", September 2012.

[RFC 3954] Claise, B., "Cisco Systems NetFlow Services Export Version 9", October 2004

[DevoFlow] Mogul, J., et al., "DevoFlow: Cost-Effective Flow Management for High Performance Enterprise Networks", Proceedings of the ACM SIGCOMM, August 2011.

[NDTM] Estan, C. and G. Varghese, "New directions in traffic measurement and accounting", Proceedings of ACM SIGCOMM, August 2002.

Appendix A. Internet Traffic Analysis and Load Balancing Simulation

Internet traffic [CAIDA] has been analyzed to obtain flow statistics such as the number of packets in a flow and the flow duration. The five tuples in the packet header (IP addresses, TCP/UDP Ports, and IP protocol) are used for flow identification. The analysis indicates that < ~2% of the flows take ~30% of total traffic volume while the rest of the flows (> ~98%) contributes ~70% [YONG].

The simulation has shown that given Internet traffic pattern, the hash-based technique does not evenly distribute the flows over ECMP paths. Some paths may be > 90% loaded while others are < 40% loaded. The more ECMP paths exist, the more severe the misbalancing. This implies that hash-based distribution can cause some paths to become congested while other paths are underutilized [YONG].

The simulation also shows substantial improvement by using the large flow-aware hash-based distribution technique described in this document. In using the same simulated traffic, the improved

rebalancing can achieve < 10% load differences among the paths. It proves how large flow-aware hash-based distribution can effectively compensate the uneven load balancing caused by hashing and the traffic characteristics [YONG].

Authors' Addresses

Ram Krishnan
Brocade Communications
San Jose, 95134, USA
Phone: +1-408-406-7890
Email: ramk@brocade.com

Sanjay Khanna
Brocade Communications
San Jose, 95134, USA
Phone: +1-408-333-4850
Email: skhanna@brocade.com

Lucy Yong
Huawei USA
5340 Legacy Drive
Plano, TX 75025, USA
Phone: +1-469-277-5837
Email: lucy.yong@huawei.com

Anoop Ghanwani
Dell
San Jose, CA 95134
Phone: +1-408-571-3228
Email: anoop@alumni.duke.edu

Ning So
Tata Communications
Plano, TX 75082, USA
Phone: +1-972-955-0914
Email: ning.so@tatacommunications.com

Bhumip Khasnabish
ZTE Corporation

New Jersey, 07960, USA
Phone: +1-781-752-8003
Email: bhumip.khasnabish@zteusa.com