OPSAWG
Internet Draft
Intended status: Informational
Expires: October 22, 2014

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

April 22, 2014

Mechanisms for Optimizing LAG/ECMP Component Link Utilization in Networks

draft-ietf-opsawg-large-flow-load-balancing-11.txt

Status of this Memo

Copyright Notice

Abstract

Demands on networking infrastructure are growing exponentially due to bandwidth hungry applications such as rich media applications and inter-data center communications. In this context, it is important to optimally use the bandwidth in wired networks that extensively use link aggregation groups and equal cost multi-paths as techniques for bandwidth scaling. This draft explores some of the mechanisms useful for achieving this.

Table of Contents

1. Introduction

Networks extensively use link aggregation groups (LAG) [802.1AX] and equal cost multi-paths (ECMP) [RFC 2991] as techniques for capacity scaling. For the problems addressed by this document, network traffic can be predominantly categorized into two traffic types: long-lived large flows and other flows. These other flows, which include long-lived small flows, short-lived small flows, and short-lived large flows, are referred to as "small flows" in this document. Long-lived large flows are simply referred to as "large flows."

Stateless hash-based techniques [ITCOM, RFC 2991, RFC 2992, RFC 6790] are often used to distribute both large flows and small 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 mechanisms for optimizing LAG/ECMP component link utilization while using hash-based techniques. The mechanisms comprise the following steps -- recognizing large flows in a router; and assigning the large flows to specific LAG/ECMP component links or redistributing the small flows when a component link on the router is congested.

It is useful to keep in mind that in typical use cases for this mechanism the 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 LAG/ECMP. In other words, the number of

large flows is NOT expected to be on the order of millions of flows. Examples of such large flows would be IPsec tunnels in service provider backbone networks 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

ECMP component link: An individual nexthop within an ECMP group. An ECMP component link may itself comprise a LAG.

ECMP table: A table that is used as the nexthop of an ECMP route that comprises the set of component links and the weights associated with each of those component links. The weights are used to determine which values of the hash function map to a given component link.

LAG component link: An individual link within a LAG. A LAG component link is typically a physical link.

LAG table: A table that is used as the output port which is a LAG that comprises the set of component links and the weights associated with each of those component links. The weights are used to

determine which values of the hash function map to a given component link.

Large flow(s): Refers to long-lived large flow(s).

Small flow(s): Refers to any of, or a combination of, long-lived small flow(s), short-lived small flows, and short-lived large flow(s).

2. Flow Categorization

In general, based on the size and duration, a flow can be categorized into any one of the following four types, as shown in Figure 1:

(a) Short-lived Large Flow (SLLF),
(b) Short-lived Small Flow (SLSF),
(c) Long-lived Large Flow (LLLF), and
(d) Long-lived Small Flow (LLSF).

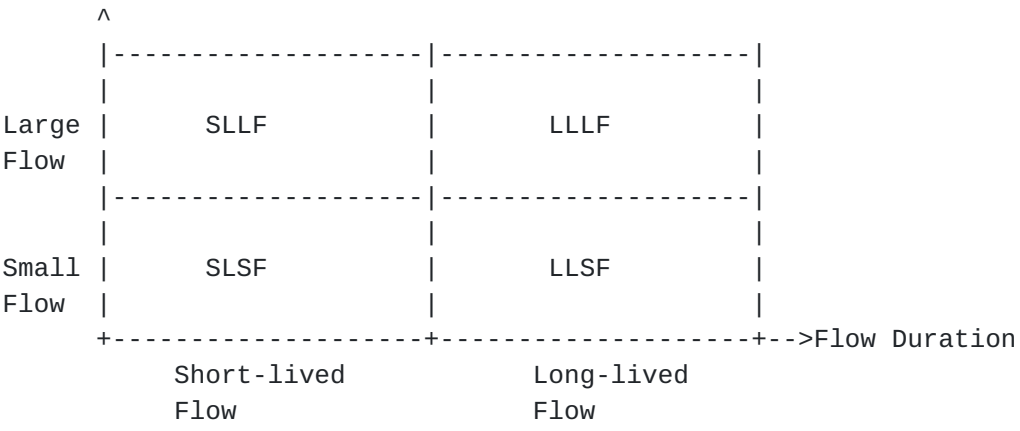


Figure 1: Flow Categorization

In this document, as mentioned earlier, we categorize long-lived large flows as "large flows", and all of the others -- long-lived small flows, short-lived small flows, and short-lived large flows as "small flows".

3. Hash-based Load Distribution in LAG/ECMP

Hash-based techniques are often used for traffic load balancing to select among multiple available paths within a LAG/ECMP group. The advantages of hash-based techniques for 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 computed using these fields is used to generate a unique number that identifies a link/path in a LAG/ECMP group. The result of the hashing procedure is a many-to-one mapping of flows to component links.

If the traffic mix 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, hash-based techniques produce 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 imbalance in the loads on individual component links.

One example is illustrated in Figure 2. In Figure 2, 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 applying the hash-based technique 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.

    o 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.

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

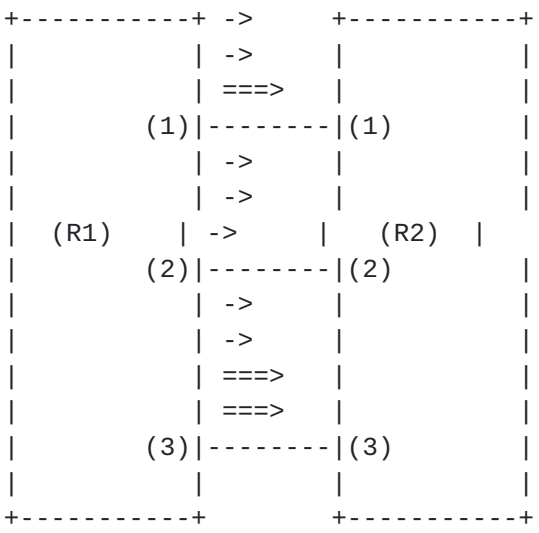


Figure 2: Unevenly Utilized Component Links

This document presents mechanisms for addressing the imbalance in load distribution resulting from commonly used hash-based techniques for LAG/ECMP that were shown in the above example. The mechanisms use large flow awareness to compensate for the imbalance in load distribution.

4. Mechanisms for Optimizing LAG/ECMP Component Link Utilization

The suggested mechanisms 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 global optimization of the 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.

4.1. Differences in LAG vs ECMP

While the mechanisms explained herein are applicable to both LAGs and ECMP groups, it is useful to note that there are some key differences between the two that may impact how effective the mechanism is. This relates, in part, to the localized information with which the scheme is intended to operate.

A LAG is usually established across links that are between 2 adjacent routers. As a result, the scope of problem of optimizing the bandwidth utilization on the component links is fairly narrow. It simply involves re-balancing the load across the component links between these two routers, and there is no impact whatsoever to other parts of the network. The scheme works equally well for unicast and multicast flows.

On the other hand, with ECMP, redistributing the load across component links that are part of the ECMP group may impact traffic patterns at all of the nodes that are downstream of the given router between itself and the destination. The local optimization may result in congestion at a downstream node. (In its simplest form, an ECMP group may be used to distribute traffic on component links that are between two adjacent routers, and in that case, the ECMP group is no different than a LAG for the purpose of this discussion. It should be noted that an ECMP component link may itself comprise a LAG, in which case the scheme may be further applied to the component links within the LAG.)

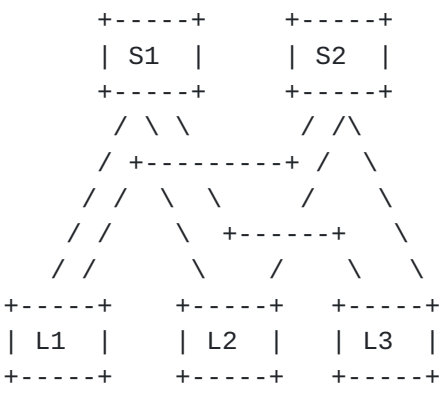


Figure 3: Two-level Fat Tree

To demonstrate the limitations of local optimization, consider a two-level fat-tree topology with three leaf nodes (L1, L2, L3) and two spine nodes (S1, S2) and assume all of the links are 10 Gbps.

Let L1 have two flows of 4 Gbps each towards L3, and let L2 have one flow of 7 Gbps also towards L3. If L1 balances the load optimally between S1 and S2, and L2 sends the flow via S1, then the downlink from S1 to L3 would get congested resulting in packet discards. On the other hand, if L1 had sent both its flows towards S1 and L2 had sent its flow towards S2, there would have been no congestion at either S1 or S2.

The other issue with applying this scheme to ECMP groups is that it may not apply equally to unicast and multicast traffic because of the way multicast trees are constructed.

Finally, it is possible for a single physical link to participate as a component link in multiple ECMP groups, whereas with LAGs, a link can participate as a component link of only one LAG.

4.2. Operational Overview

The various steps in optimizing 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 4.3.

Step 2) The egress component links are periodically scanned for link utilization and the imbalance for the LAG/ECMP group is monitored. If the imbalance exceeds a certain imbalance threshold, then re-balancing is triggered. Measurement of the imbalance is discussed further in 5.1. Additional criteria may also be used to determine whether or not to trigger rebalancing, such as the maximum utilization of any of the component links, in addition to the imbalance.

Step 3) As a part of rebalancing, the operator can choose to rebalance the large flows on to lightly loaded component links of the LAG/ECMP group, redistribute the small flows on the congested link to other component links of the group, or a combination of both.

All of the steps identified above can be done locally within the router itself or could involve the use of a central management entity.

Providing large flow information to a central management entity provides the capability to globally optimize flow distribution as described in Section 4.1. 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 path P1 there may be a congested link, while paths P2 and P3 may be under-utilized. This is something that 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/or P3 resulting in a more optimized flow of traffic.

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

4.3. Large Flow Recognition

4.3.1. 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, for example:

- 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 Generic Routing Encapsulation (GRE) [RFC 2784], Virtual eXtensible Local Area Network (VXLAN) [VXLAN], Network Virtualization using Generic Routing Encapsulation (NVGRE) [NVGRE], Stateless Transport Tunneling (STT) [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.

This method of flow identification is consistent with that of IPFIX [RFC 7011].

4.3.2. Criteria and Techniques for Large Flow Recognition

From a bandwidth and time duration perspective, in order to recognize 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 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 for 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 recognize a large flow are described below.

4.3.3. 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 recognition, sampling needs to be enabled on all of the egress ports in the router where such measurements are desired.

Using sFlow as an example, processing in a 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.

The option of using ingress sampling for this purpose 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 be 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. There is one situation in which this approach may not work. If there are tunnels that originate from the given router, and if the resulting tunnel comprises the large flow, then this cannot be deduced from ingress sampling at the given router. Instead, if egress sampling is unavailable, then ingress sampling from the downstream router must be used.

To illustrate the use of ingress versus egress sampling, we refer to Figure 2. Since we are looking at rebalancing flows at R1, we would need to enable egress sampling on ports (1), (2), and (3) on R1. If egress sampling is not available, and if R2 is also under the control of the same administrator, enabling ingress sampling on R2's ports (1), (2), and (3) would also work, but it would necessitate the involvement of a central management entity in order for R1 to obtain large flow information for each of its links. Finally, R1 can enable ingress sampling only on all of its ports (not just the ports that are part of the LAG/ECMP group being monitored) and that would suffice if the sampling technique annotates the samples with the egress port information.

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]. A discussion on determining the appropriate sampling frequency is available in the following reference [SAMP-BASIC].

4.3.4. Inline Data Path Measurement

Implementations may perform recognition of large flows by performing measurements on traffic in the data path of a router. Such an approach would be expected to operate at the interface speed on every interface, accounting for all packets processed by the data path of the router. An example of such an approach is described in IPFIX [RFC 5470].

Using inline data path measurement, a faster and more accurate indication of large flows mapped to each of the component links in a LAG/ECMP group may be possible (as compared to the sampling-based approach).

The advantages and disadvantages of inline data path measurement are:

Advantages:

- 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 inline data path measurement, large flows can be recognized in shorter windows on higher link speeds since every packet is accounted for [NDTM].

- Eliminates the potential dependence on an external management station for large flow recognition.

Disadvantages:

- It is more resource intensive in terms of the tables sizes required for monitoring all flows in order to perform the measurement.

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 details of inline data path measurement of large flows is vendor dependent and beyond the scope of this document.

4.3.5. Use of More Than One Method for Large Flow Recognition

It is possible that a router may have line cards that support a sampling technique while other line cards support inline data path measurement of large flows. As long as there is a way for the router to reliably determine the mapping of large flows to component links

of a LAG/ECMP group, it is acceptable for the router to use more than one method for large flow recognition.

If both methods are supported, inline data path measurement may be preferable because of its speed of detection [FLOW-ACC].

4.4. Load Rebalancing Options

Below are suggested techniques for load rebalancing. 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 rebalancing 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.

4.4.1. 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 policy-based routing (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 imbalance in the utilization once again across the component links. 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 four 10 Gbps component links and there are four large flows, each of 1 Gbps. These flows are each placed on one of the component links. Subsequent, a fifth 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.

4.4.2. 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 4.4.1) 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.

4.4.3. 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 table replaces the failed component link with the reserved link.

4.4.4. Load Rebalancing Algorithms

Specific algorithms for placement of large flows are out of scope of this document. One possibility is to formulate the problem for large flow placement as the well-known bin-packing problem and make use of the various heuristics that are available for that problem [bin-pack].

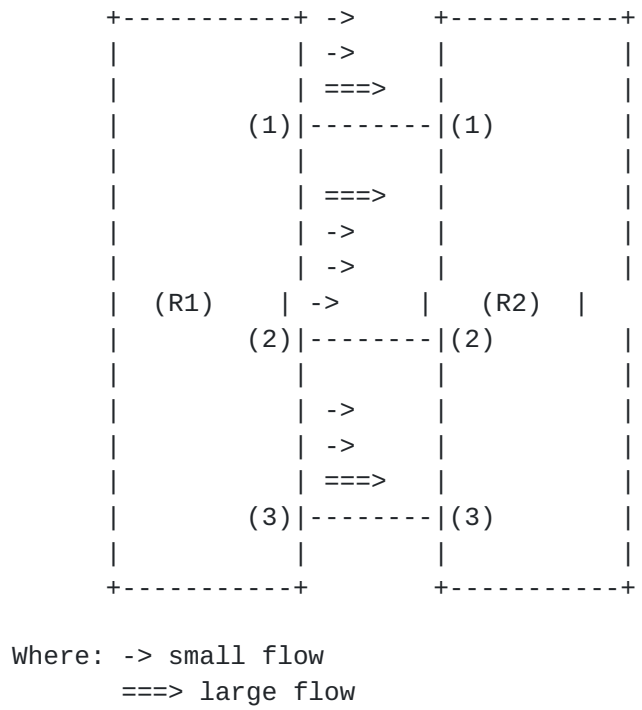
4.4.5. Load Rebalancing Example

Optimizing LAG/ECMP component utilization for the use case in Figure 2 is depicted below in Figure 4. The large flow rebalancing explained in Section 4.4 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.



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

Figure 4: Evenly Utilized Composite Links

Basically, the use of the mechanisms described in Section 4.4.1 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.

5. Information Model for Flow Rebalancing

In order to support flow rebalancing in a router from an external system, the exchange of some information is necessary between the router and the external system. This section provides an exemplary information model covering the various components needed for the purpose. The model is intended to be informational and may be used as input for development of a data model.

5.1. Configuration Parameters for Flow Rebalancing

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: A measure of the deviation of the component link utilizations from the utilization of the overall LAG/ECMP group. Since component links can be of a different speed, the imbalance can be computed as follows. Let the utilization of each component link in a LAG/ECMP group with n links of speed b_1, b_2, .., b_n, be u_1, u_2, .., u_n. The mean utilization is computed is u_ave = [(u_1 x b_1) + (u_2 x b_2) + .. + (u_n x b_n)] / [b_1 + b_2 + b_n]. The imbalance is then computed as max_{i=1..n} | u_i - u_ave | / u_ave.

- Rebalancing interval: The minimum amount of time between rebalancing events. This parameter ensures that rebalancing is not invoked too frequently as it impacts packet ordering.

These parameters may be configured on a system-wide basis or it may apply to an individual LAG. It may be applied to an ECMP group provided the component links are not shared with any other ECMP group.

5.2. System Configuration and Identification Parameters

The following parameters are useful for router configuration and operation when using the mechanisms in this document.

- IP address: The IP address of a specific router that the feature is being configured on, or that the large flow placement is being applied to.

- LAG ID: Identifies the LAG on a given router. The LAG ID may be required when configuring this feature (to apply a specific set of large flow identification parameters to the LAG) and will be required when specifying flow placement to achieve the desired rebalancing.

- Component Link ID: Identifies the component link within a LAG or ECMP group. This is required when specifying flow placement to achieve the desired rebalancing.

- Component Link Weight: The relative weight to be applied to traffic for a given component link when using hash-based techniques for load distribution.

- ECMP group: Identifies a particular ECMP group. The ECMP group may be required when configuring this feature (to apply a specific set of large flow identification parameters to the ECMP group) and will be required when specifying flow placement to achieve the desired rebalancing. We note that multiple ECMP groups can share an overlapping set (or non-overlapping subset) of component links. This document does not deal with the complexity of addressing such configurations.

The feature may be configured globally for all LAGs and/or for all ECMP groups, or it may be configured specifically for a given LAG or ECMP group.

5.3. Information for Alternative Placement of Large Flows

In cases where large flow recognition is handled by an external management station (see Section 4.3.3), 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.

The IPFIX information model [RFC 7012] can be leveraged for large flow identification.

Large Flow placement is achieved by specifying the relevant flow information along with the following:

- For LAG: Router's IP address, LAG ID, LAG component link ID.
- For ECMP: Router's IP address, ECMP group, ECMP component link ID.

In the case where the ECMP component link itself comprises a LAG, we would have to specify the parameters for both the ECMP group as well as the LAG to which the large flow is being directed.

5.4. Information for Redistribution of Small Flows

Redistribution of small flows is done using the following:

- For LAG: The LAG ID and the component link IDs along with the relative weight of traffic to be assigned to each component link ID are required.

- For ECMP: The ECMP group and the ECMP Nexthop along with the relative weight of traffic to be assigned to each ECMP Nexthop are required.

It is possible to have an ECMP nexthop that itself comprises a LAG. In that case, we would have to specify the new weights for both the ECMP nexthops within the ECMP group as well as the component links within the LAG.

In the case where an ECMP component link itself comprises a LAG, we would have to specify new weights for both the component links within the ECMP group as well as the component links within the LAG.

5.5. Export of Flow Information

Exporting large flow information is required when large flow recognition is being done on a router, but the decision to rebalance is being made in an external management station. Large flow information includes flow identification and the component link ID that the flow currently is assigned to. Other information such as flow QoS and bandwidth may be exported too.

The IPFIX information model [RFC 7012] can be leveraged for large flow identification.

5.6. Monitoring information

5.6.1. Interface (link) utilization

The incoming bytes (ifInOctets), outgoing bytes (ifOutOctets) and interface speed (ifSpeed) can be measured from the Interface table (iftable) MIB [RFC 1213].

The link utilization can then be computed as follows:

Incoming link utilization = (ifInOctets 8 / ifSpeed)

Outgoing link utilization = (ifOutOctets 8 / ifSpeed)

For high speed Ethernet links, the etherStatsHighCapacityTable MIB [RFC 3273] can be used.

For scalability, it is recommended to use the counter push mechanism in [sflow-v5] for the interface counters. Doing so would help avoid counter polling through the MIB interface.

The outgoing link utilization of the component links within a LAG/ECMP group can be used to compute the imbalance (See Section 5.1) for the LAG/ECMP group.

5.6.2. Other monitoring information

Additional monitoring information that is useful includes:

- Number of times rebalancing was done.
- Time since the last rebalancing event.
- The number of large flows currently rebalanced by the scheme.
- A list of the large flows that have been rebalanced including
  - the rate of each large flow at the time of the last rebalancing for that flow,
  - the time that rebalancing was last performed for the given large flow, and
  - the interfaces that the large flows was (re)directed to.
- The settings for the weights of the interfaces within a LAG/ECMP used by the small flows which depend on hashing.

6. Operational Considerations

6.1. Rebalancing Frequency

Flows should be rebalanced only when the imbalance in the utilization across component links exceeds a certain threshold. Frequent rebalancing 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 redistributed. It should be noted that reordering is a concern for TCP flows with even a few packets because three out-of-order packets would trigger sufficient duplicate ACKs to the sender resulting in a retransmission [RFC 5681].

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.

6.2. Handling Route Changes

Large flow rebalancing must be aware of any changes to the FIB. In cases where the nexthop of a route no longer to points to the LAG, or to an ECMP group, any PBR entries added as described in Section 4.4.1 and 4.4.2 must be withdrawn in order to avoid the creation of forwarding loops.

7. IANA Considerations

This memo includes no request to IANA.

8. 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.

9. Contributing Authors

Sanjay Khanna
Cisco Systems
Email: sanjakha@gmail.com

10. Acknowledgements

The authors would like to thank the following individuals for their review and valuable feedback on earlier versions of this document: Shane Amante, Fred Baker, Michael Bugenhagen, Zhen Cao, Brian Carpenter, Benoit Claise, Michael Fargano, Wes George, Sriganesh Kini, Roman Krzanowski, Andrew Malis, Dave McDysan, Pete Moyer, Peter Phaal, Dan Romascanu, Curtis Villamizar, Jianrong Wong, George Yum, and Weifeng Zhang.

11. References

11.1. Normative References

11.2. Informative References

[802.1AX] IEEE Standards Association, "IEEE Std 802.1AX-2008 IEEE Standard for Local and Metropolitan Area Networks - Link Aggregation", 2008.

[bin-pack] Coffman, Jr., E., M. Garey, and D. Johnson. Approximation Algorithms for Bin-Packing -- An Updated Survey. In Algorithm Design for Computer System Design, ed. by Ausiello, Lucertini, and Serafini. Springer-Verlag, 1984.

[CAIDA] Caida Internet Traffic Analysis, http://www.caida.org/home.
[DevoFlow] Mogul, J., et al., "DevoFlow: Cost-Effective Flow Management for High Performance Enterprise Networks," Proceedings of the ACM SIGCOMM, August 2011.

[FLOW-ACC] Zseby, T., et al., "Packet sampling for flow accounting: challenges and limitations," Proceedings of the 9th international conference on Passive and active network measurement, 2008.

[ID.ietf-rtgwg-cl-requirement] Villamizar, C. et al., "Requirements for MPLS over a Composite Link," September 2013.

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

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

[NVGRE] Sridharan, M. et al., "NVGRE: Network Virtualization using Generic Routing Encapsulation," draft-sridharan-virtualization-nvgre-04, February 2014.

[RFC 2784] Farinacci, D. et al., "Generic Routing Encapsulation (GRE)," March 2000.

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

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

[RFC 1213] McCloghrie, K., "Management Information Base for Network Management of TCP/IP-based internets: MIB-II," March 1991.

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

[RFC 3273] Waldbusser, S., "Remote Network Monitoring Management Information Base for High Capacity Networks," July 2002.

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

[RFC 5470] G. Sadasivan et al., "Architecture for IP Flow Information Export," March 2009.

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

[RFC 5681] Allman, M. et al., "TCP Congestion Control," September 2009.

[RFC 7011] Claise, B. et al., "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information," September 2013.

[RFC 7012] Claise, B. and B. Trammell, "Information Model for IP Flow Information Export (IPFIX)," September 2013.

[SAMP-BASIC] Phaal, P. and S. Panchen, "Packet Sampling Basics," http://www.sflow.org/packetSamplingBasics/.

[sFlow-LAG] Phaal, P. and A. Ghanwani, "sFlow LAG counters structure," http://www.sflow.org/sflow_lag.txt, September 2012.

[sFlow-v5] Phaal, P. and M. Lavine, "sFlow version 5," http://www.sflow.org/sflow_version_5.txt, July 2004.

[STT] Davie, B. (ed) and J. Gross, "A Stateless Transport Tunneling Protocol for Network Virtualization (STT)," draft-davie-stt-06, March 2014.

[VXLAN] Mahalingam, M. et al., "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks," draft-mahalingam-dutt-dcops-vxlan-09, April 2014.

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

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: ramkri123@gmail.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: vumip1@gmail.com