Internet Engineering Task Force (IETF)
Request for Comments: 7424
Category: Informational
ISSN: 2070-1721

R. Krishnan
Brocade Communications
L. Yong
Huawei USA
A. Ghanwani
Dell
N. So
Vinci Systems
B. Khasnabish
ZTE Corporation

January 2015

Mechanisms for Optimizing Link Aggregation Group (LAG) and Equal-Cost Multipath (ECMP) Component Link Utilization in Networks

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 multipaths as techniques for bandwidth scaling. This document explores some of the mechanisms useful for achieving this.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7424.

Table of Contents

1. Introduction

Networks extensively use link aggregation groups (LAGs) [802.1AX] and equal-cost multipaths (ECMPs) [RFC2991] 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] [RFC2991] [RFC2992] [RFC6790] 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 document describes mechanisms for optimizing LAG/ECMP component link utilization when using hash-based techniques. The mechanisms comprise the following steps: 1) recognizing large flows in a router, and 2) 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 these mechanisms, the large flows 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 10s or 100s 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

DoS: Denial of Service

ECMP: Equal-Cost Multipath

GRE: Generic Routing Encapsulation

IPFIX: IP Flow Information Export

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

VXLAN: Virtual eXtensible LAN

1.2. Terminology

Central management entity:
An entity that is capable of monitoring information about link utilization and flows in routers across the network and may be capable of making traffic-engineering decisions for placement of large flows. It may include the functions of a collector [RFC7011].

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

ECMP table:
A table that is used as the next hop of an ECMP route that comprises the set of ECMP component links and the weights associated with each of those ECMP component links. The input for looking up the table is the hash value for the packet, and the weights are used to determine which values of the hash function map to a given ECMP component link.

Flow (large or small):
A sequence of packets for which ordered delivery should be maintained, e.g., packets belonging to the same TCP connection.

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 LAG component links and the weights associated with each of those component links. The input for looking up the table is the hash value for the packet, and the weights are used to determine which values of the hash function map to a given LAG 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:

- short-lived large flow (SLLF),
- short-lived small flow (SLSF),
- long-lived large flow (LLLF), and
- long-lived small flow (LLSF).

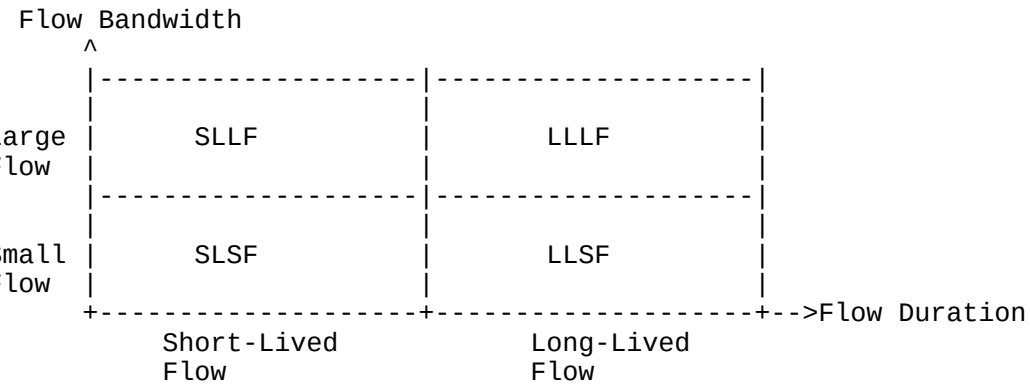

Figure 1: Flow Categorization

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 load balancing of traffic 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.

Hash-based techniques produce good results with respect to utilization of the individual 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,
- the individual flow rates are much smaller as compared to the link capacity, and
- the differences in flow rates are not dramatic.

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.

An example is illustrated in Figure 2. As shown, there are two routers, R1 and R2, and there is a LAG between them that has three component links (1), (2), and (3). A total of ten flows 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 three flows (two small flows and one large flow), and the link utilization is normal.
- Component link (2) has three flows (three small flows and no large flows), and the link utilization is light.
  - The absence of any large flow causes the component link to be underutilized.
- Component link (3) has four flows (two small flows and two large flows), and the link capacity is exceeded resulting in congestion.
  - The presence of two large flows causes congestion on this component link.

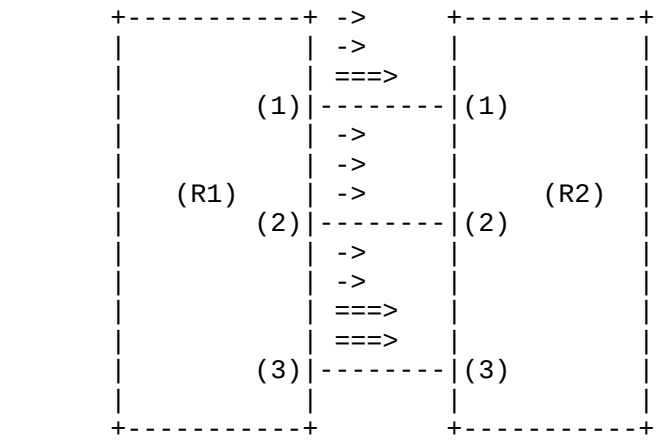

Figure 2: Unevenly Utilized Component Links

This document presents mechanisms for addressing the imbalance in load distribution resulting from commonly used hash-based techniques for LAG/ECMP that are 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 document are local optimization solutions; they are local in the sense that both the identification of large flows and rebalancing of the load can be accomplished completely within individual routers in the network without the need for interaction with other routers.

This approach may not yield a global optimization of the placement of large flows across multiple routers 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; 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 mechanisms are. This relates, in part, to the localized information with which the mechanisms are intended to operate.

A LAG is usually established across links that are between two adjacent routers. As a result, the scope of the problem of optimizing the bandwidth utilization on the component links is fairly narrow. It simply involves rebalancing 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 routers 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.)

To demonstrate the limitations of local optimization, consider a two-level Clos network topology as shown in Figure 3 with three leaf routers (L1, L2, and L3) and two spine routers (S1 and S2). 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.

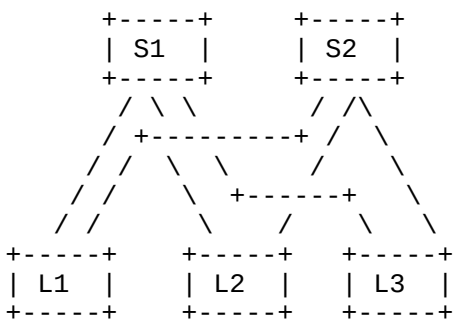

Figure 3: Two-Level Clos Network

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 step involves recognizing large flows in routers and maintaining the mapping for each large flow to the component link that it uses. 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 threshold, then rebalancing is triggered. Measurement of the imbalance is discussed further in Section 5.1. In addition to the imbalance, further criteria (such as the maximum utilization of any of the component links) may also be used to determine whether or not to trigger rebalancing. The use of sampling techniques for the measurement of egress component link utilization, including the issues of depending on ingress sampling for these measurements, are discussed in Section 4.3.3.

Step 3:
As a part of rebalancing, the operator can choose to rebalance the large flows by placing them on 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 three ECMP next hops 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 underutilized. 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 steps described above are especially useful when bundling links of different bandwidths, e.g., 10 Gbps and 100 Gbps as described in [RFC7226].

4.3. Large Flow Recognition

4.3.1. Flow Identification

Flows are typically identified using one or more fields from the packet header, for example:

- Layer 2: Source Media Access Control (MAC) address, destination MAC address, VLAN ID.
- IP header: IP protocol, IP source address, IP destination address, flow label (IPv6 only).
- Transport protocol header: Source port number, destination port number. These apply to protocols such as TCP, UDP, and the Stream Control Transmission Protocol (SCTP).
- MPLS labels.

For tunneling protocols like Generic Routing Encapsulation (GRE) [RFC2784], Virtual eXtensible LAN (VXLAN) [RFC7348], Network Virtualization using Generic Routing Encapsulation (NVGRE) [NVGRE],

Stateless Transport Tunneling (STT) [STT], Layer 2 Tunneling Protocol (L2TP) [RFC3931], etc., flow identification is possible based on inner and/or outer headers as well as fields introduced by the tunnel header, as any or all such fields may be used for load balancing decisions [RFC5640].

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 [RFC7011].

4.3.2. Criteria and Techniques for Large Flow Recognition

From the perspective of bandwidth and time duration, in order to recognize large flows, we define an observation interval and measure 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 that is at or above 10% of link bandwidth for a time period of at least one 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 in Sections 4.3.3, 4.3.4, and 4.3.5.

4.3.3. Sampling Techniques

A number of routers support sampling techniques such as sFlow [sFlow-v5] [sFlow-LAG], Packet Sampling (PSAMP) [RFC5475], and NetFlow Sampling [RFC3954]. 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 an sFlow collector can provide an approximate indication of the mapping of large flows to each of the component links in each LAG/ECMP group. Assuming sufficient control plane resources are available, it is possible to implement this part of the collector function in the control plane of the router to reduce dependence on a central management entity.

If egress sampling is not available, ingress sampling can suffice since the central management entity used by the sampling technique typically has visibility across multiple routers in a network and can use the samples from an immediately downstream router to make measurements for egress traffic at the local router.

The option of using ingress sampling for this purpose may not be available if the downstream router 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, for this scenario, 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 only enable ingress sampling 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.

Disadvantage:

- 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 [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 [RFC5470].

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 as follows:

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].
- Inline data path measurement eliminates the potential dependence on a central management entity for large flow recognition.

Disadvantage:

- o Inline data path measurement is more resource intensive in terms of the table sizes required for monitoring all flows.

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 Multiple Methods 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. 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. Options for Load Rebalancing

The following subsections describe suggested techniques for load balancing. Equipment vendors may implement more than one technique, including those not described in this document, and allow the operator to choose between them.

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, member component links with the least average link 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 reordering.

Moving a large flow will increase the utilization of the link that it is moved to, potentially once again creating an imbalance in the utilization across the component links. Therefore, when moving a large flow, care must be taken to account for the existing load and the future load after the 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. Subsequently, 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. 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).

- 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.

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

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. When there is a failure of one or more component links, 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 links simply replaces the failed

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

4.4.4. Algorithms for Load Rebalancing

Specific algorithms for placement of large flows are out of the 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. Example of Load Rebalancing

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 4.4.1 is used. The improved link utilization is as follows:

- Component link (1) has three flows (two small flows and one large flow), and the link utilization is normal.
- Component link (2) has four flows (three small flows and one large flow), and the link utilization is normal now.
- Component link (3) has three flows (two small flows and one large flow), and the link utilization is normal now.

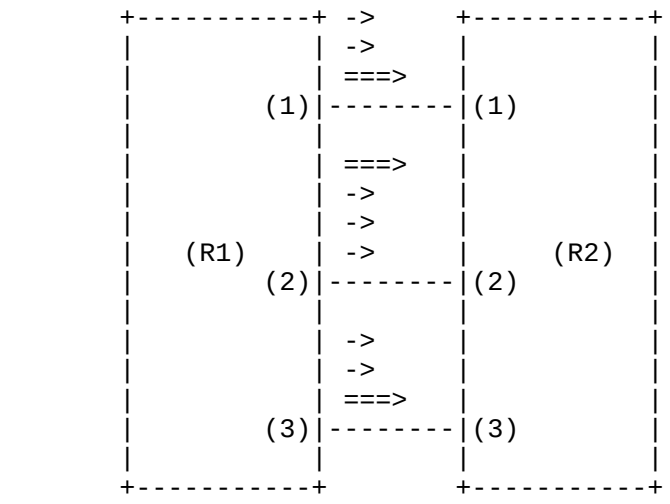

Figure 4: Evenly Utilized Composite Links

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 underutilized.

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 this purpose. The model is intended to be informational and may be used as a guide for the development of a data model.

5.1. Configuration Parameters for Flow Rebalancing

The following parameters are required for configuration of this feature:

- Large flow recognition parameters:
  - Observation interval: The observation interval is the time period in seconds over which packet arrivals are observed for the purpose of large flow recognition.

- 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 over a given observation interval exceeds this number would be recognized as a large flow.

- 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.

o 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 different speeds, 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 as

  u_ave = [(u_1 * b_1) + (u_2 * b_2) + .. + (u_n * b_n)] /
          [b_1 + b_2 + .. + b_n].

  The imbalance is then computed as

  max_{i=1..n} | u_i - u_ave |.

o 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 may apply to an individual LAG/ECMP group. They may be applied to an ECMP group, provided that 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.

o 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.

- o 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.

- o 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.

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

- o 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 a central management entity (see Section 4.3.3), an information model for flows is required to allow the import of large flow information to the router.

Typical fields used for identifying large flows were discussed in Section 4.3.1. The IPFIX information model [RFC7012] can be leveraged for large flow identification.

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

- o For LAG: router’s IP address, LAG ID, LAG component link ID.

- o 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 next hop along with the relative weight of traffic to be assigned to each ECMP next hop are required.

It is possible to have an ECMP next hop that itself comprises a LAG. In that case, we would have to specify the new weights for both the ECMP component links and the LAG component links.

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 a central management entity. Large flow information includes flow identification and the component link ID that the flow is currently assigned to. Other information such as flow QoS and bandwidth may be exported too.

The IPFIX information model [RFC7012] 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 obtained, for example, from the Interfaces table (ifTable) in the MIB module defined in [RFC1213].

The link utilization can then be computed as follows:

Incoming link utilization = (delta_ifInOctets * 8) / (ifSpeed * T)

Outgoing link utilization = (delta_ifOutOctets * 8) / (ifSpeed * T)

Where T is the interval over which the utilization is being measured, delta_ifInOctets is the change in ifInOctets over that interval, and delta_ifOutOctets is the change in ifOutOctets over that interval.

For high-speed Ethernet links, the etherStatsHighCapacityTable in the MIB module defined in [RFC3273] can be used.

Similar results may be achieved using the corresponding objects of other interface management data models such as YANG [RFC7223] if those are used instead of MIBs.

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 group used by the small flows that 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 counterproductive 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 [RFC5681].

The operator would have to experiment with various values of the large flow recognition parameters (minimum bandwidth threshold, minimum bandwidth threshold for large flow maintenance, and 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 Forwarding Information Base (FIB). In cases where the next hop of a route no longer to points to the LAG or to an ECMP group, any PBR entries added as described in Sections 4.4.1 and 4.4.2 must be withdrawn in order to avoid the creation of forwarding loops.

6.3. Forwarding Resources

Hash-based techniques used for load balancing with LAG/ECMP are usually stateless. The mechanisms described in this document require additional resources in the forwarding plane of routers for creating PBR rules that are capable of overriding the forwarding decision from the hash-based approach. These resources may limit the number of flows that can be rebalanced and may also impact the latency experienced by packets due to the additional lookups that are required.

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 that causes a hash imbalance resulting in heavy overloading of large flows to certain LAG/ECMP component links.

An attacker with knowledge of the large flow recognition algorithm and any stateless distribution method can generate flows that are distributed in a way that overloads a specific path. This could be used to cause the creation of PBR rules that exhaust the available PBR rule capacity on routers in the network. If PBR rules are consequently discarded, this could result in congestion on the attacker-selected path. Alternatively, tracking large numbers of PBR rules could result in performance degradation.

8. References

8.1. Normative References

[802.1AX] IEEE, "IEEE Standard for Local and metropolitan area networks - Link Aggregation", IEEE Std 802.1AX-2008, 2008.

[RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, November 2000, <http://www.rfc-editor.org/info/rfc2991>.

[RFC7011] Claise, B., Ed., Trammell, B., Ed., and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information", STD 77, RFC 7011, September 2013, <http://www.rfc-editor.org/info/rfc7011>.

[RFC7012] Claise, B., Ed., and B. Trammell, Ed., "Information Model for IP Flow Information Export (IPFIX)", RFC 7012, September 2013, <http://www.rfc-editor.org/info/rfc7012>.

8.2. Informative References

[BIN-PACK] Coffman, Jr., E., Garey, M., and D. Johnson. "Approximation Algorithms for Bin-Packing -- An Updated Survey" (in "Algorithm Design for Computer System Design"), Springer, 1984.

[CAIDA] "Caida Traffic Analysis Research", <http://www.caida.org/research/traffic-analysis/>.

[DEVOFLOW] Mogul, J., Tourrilhes, J., Yalagandula, P., Sharma, P., Curtis, R., and S. Banerjee, "DevoFlow: Cost-Effective Flow Management for High Performance Enterprise Networks", Proceedings of the ACM SIGCOMM, 2010.

[FLOW-ACC] Zseby, T., Hirsch, T., and B. Claise, "Packet Sampling for Flow Accounting: Challenges and Limitations", Proceedings of the 9th international Passive and Active Measurement Conference, 2008.

[ITCOM] Jo, J., Kim, Y., Chao, H., and F. Merat, "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] Garg, P. and Y. Wang, "NVGRE: Network Virtualization using Generic Routing Encapsulation", Work in Progress, draft-sridharan-virtualization-nvgre-07, November 2014.

[RFC2784] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000, <http://www.rfc-editor.org/info/rfc2784>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012, <http://www.rfc-editor.org/info/rfc6790>.

[RFC1213] McCloghrie, K. and M. Rose, "Management Information Base for Network Management of TCP/IP-based internets: MIB-II", STD 17, RFC 1213, March 1991, <http://www.rfc-editor.org/info/rfc1213>.

[RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, November 2000, <http://www.rfc-editor.org/info/rfc2992>.

[RFC3273] Waldbusser, S., "Remote Network Monitoring Management Information Base for High Capacity Networks", RFC 3273, July 2002, <http://www.rfc-editor.org/info/rfc3273>.

[RFC3931] Lau, J., Ed., Townsley, M., Ed., and I. Goyret, Ed., "Layer Two Tunneling Protocol - Version 3 (L2TPv3)", RFC 3931, March 2005, <http://www.rfc-editor.org/info/rfc3931>.

[RFC3954] Claise, B., Ed., "Cisco Systems NetFlow Services Export Version 9", RFC 3954, October 2004, <http://www.rfc-editor.org/info/rfc3954>.

[RFC5470] Sadasivan, G., Brownlee, N., Claise, B., and J. Quittek, "Architecture for IP Flow Information Export", RFC 5470, March 2009, <http://www.rfc-editor.org/info/rfc5470>.

[RFC5475] Zseby, T., Molina, M., Duffield, N., Niccolini, S., and F. Raspall, "Sampling and Filtering Techniques for IP Packet Selection", RFC 5475, March 2009, <http://www.rfc-editor.org/info/rfc5475>.

[RFC5640] Filsfils, C., Mohapatra, P., and C. Pignataro, "Load-Balancing for Mesh Softwires", RFC 5640, August 2009, <http://www.rfc-editor.org/info/rfc5640>.

[RFC5681] Allman, M., Paxson, V., and E. Blanton, "TCP Congestion Control", RFC 5681, September 2009, <http://www.rfc-editor.org/info/rfc5681>.

[RFC7223] Bjorklund, M., "A YANG Data Model for Interface Management", RFC 7223, May 2014, <http://www.rfc-editor.org/info/rfc7223>.

[RFC7226] Villamizar, C., Ed., McDysan, D., Ed., Ning, S., Malis, A., and L. Yong, "Requirements for Advanced Multipath in MPLS Networks", RFC 7226, May 2014, <http://www.rfc-editor.org/info/rfc7226>.

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

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

[sFlow-LAG] Phaal, P. and A. Ghanwani, "sFlow LAG Counters Structure", September 2012, <http://www.sflow.org/sflow_lag.txt>.

[STT] Davie, B., Ed., and J. Gross, "A Stateless Transport Tunneling Protocol for Network Virtualization (STT)", Work in Progress, draft-davie-stt-06, April 2014.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, August 2014, <http://www.rfc-editor.org/info/rfc7348>.

[YONG] Yong, L. and P. Yang, "Enhanced ECMP and Large Flow Aware Transport", Work in Progress, draft-yong-pwe3-enhance-ecmp-lfat-01, March 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 5-tuple in the packet header (IP source address, IP destination address, transport protocol source port number, transport protocol destination port number, and IP protocol) is 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 patterns, the hash-based technique does not evenly distribute flows over ECMP paths. Some paths may be > 90% loaded while others are < 40% loaded. The greater the number of ECMP paths, the more severe is the imbalance in the load distribution. 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].

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. As a part of the IETF Last Call process, valuable comments were received from Martin Thomson and Carlos Pignataro.

Contributors

Sanjay Khanna
Cisco Systems
EMail: sanjakha@gmail.com

Authors' Addresses

Ram Krishnan
Brocade Communications
San Jose, CA 95134
United States
Phone: +1-408-406-7890
EMail: ramkri123@gmail.com

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

Anoop Ghanwani
Dell
5450 Great America Pkwy
Santa Clara, CA 95054
United States
Phone: +1-408-571-3228
EMail: anoop@alumni.duke.edu

Ning So
Vinci Systems
2613 Fairbourne Cir
Plano, TX 75093
United States
EMail: ningso@yahoo.com

Bhumip Khasnabish
ZTE Corporation
New Jersey 07960
United States
Phone: +1-781-752-8003
EMail: vumip1@gmail.com