Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: February 20, 2016

S. Pallagatti, Ed.
P. Sarkar, Ed.
H. Gredler
Juniper Networks
S. Litkowski
Orange Business Service

August 19, 2015

IGP bandwidth based metric.
draft-spallagatti-rtgwg-bandwidth-based-metric-01

Abstract

This document describes a method to group multiple interfaces and assign metric to that group based on the cumulative bandwidth of all the interfaces in that group. Each link in a group takes same group metric irrespective of its own bandwidth.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC2119 [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on February 20, 2016.

Copyright Notice

Table of Contents

1. Introduction

A low cost path is always preferred to carry traffic from source to destination. If a application is more interested in bandwidth than the cost itself and most preferred path does not satisfy bandwidth then this could potentially lead to congestion and packet loss for an application. Bandwidth critical applications needs minimum bandwidth to be satisfied even if traffic is carried over multiple alternative paths to reach a destination.

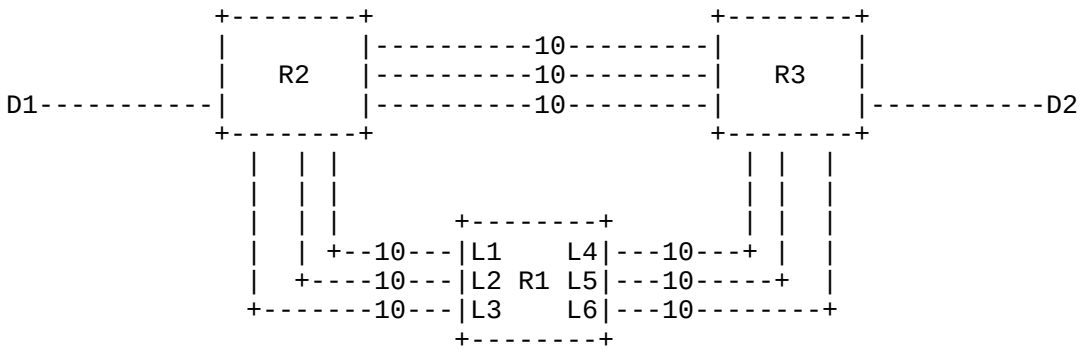

Figure 1: Example Topology

Consider the topology as show in Figure 1. The device R1 uses a links L1, L2 and L3 to carry traffic to destination D1. Similarly it uses links L4, L5 and L6 for destination D2. However in the event of links L1 and/or L2 fails traffic is still forwarded on link L3 causing traffic congestion.

In such situations operators will prefer the traffic for destination D1 is forwarded on L4, L5 and L6, as there is lesser chance of congestion. Similarly when links L4 and L5 also fails, the operator will prefer the traffic for D1 is forwarded is switched back on link L3 again. This document proposes a method called Bandwidth Based Metrics (hereafter referred as BBM), which helps achieving this desired behaviour.

BBM, on detecting a local link event, attempts to re-route traffic, based on remaining bandwidth across the links on the primary and alternate paths. When the remaining available bandwidth on the primary link(s) goes below a permissible limit (to be specified by the operator), traffic should be re-routed to one or more groups of alternative paths, and re-distributed onto multiple alternate paths with lesser likeliness of congesting them.

This document also specifies how to extend Fast Re-Route (FRR) for BBM to meet stringent re-convergence time constraints, and minimize traffic loss due to network congestion caused by standard FRR mechanisms.

2. BBM Concepts

2.1. Interface-Group

BBM method proposed in this document requires grouping of all the local links (or interfaces) attached to a node into one or more logical bundles. Such a logical grouping of multiple local interfaces is called an interface-group, and needs to be provisioned manually by the operator on each node. While assigning the local interfaces to a interface-group, all links connecting the local node to the same one-hop neighbor, SHOULD be assigned to a single interface-group. In other words the number of interface-group to be created on a node SHOULD be at the least, the number of one-hop neighbor nodes the particular node is connected to.

In Figure 1 links L1, L2, and L3 connecting R1 and R2 can be grouped into a single interface group (say IG1) on both R1 and R2. Similarly links L4, L5 and L6 connecting R1 and R3 can be grouped into another single interface-group (say IG2) on both R1 and R3.

2.2. BBM Metric Configurations

All the interfaces under a given interface-group shall share a metric that is proportionate to the cumulative bandwidth available using the individual links under the interface-group. if a link is associated with interface-group then interface-group metric MUST override individual link metric configuration. Implementations SHOULD allow operator to specify what metric should be associated for a given total remaining available bandwidth for each interface group. Implementations SHOULD also allow operator specify the default metric to be used for each interface-group.

In Figure 1, considering all the links L1 to L6 having bandwidth capacity of 100G each, and assigned into two interface-groups IG1 and IG2 (as shown in Section 2.1), following is an example of simple BBM config for each of these interface-group.

```
IG1:
   Member-Links: L1,L2,L3
   Total-Available-BW: 200G,  Metric: 10
   Total-Available-BW: 100G,  Metric: 50
   Default-Metric: 1000

IG2:
   Member-Links: L4,L5,L6
   Total-Available-BW: 200G,  Metric: 10
   Total-Available-BW: 100G,  Metric: 50
   Default-Metric: 1000
```

Figure 2: Example BBM Configuration

2.3. BBM Terminologies

This document also defines the following attributes to be associated with each interface-group.

Attribute	Value and Significance
Intf_List	The list of interfaces assigned to this group as per configuration.
BW_Curr	Total available bandwidth across all active member interfaces of this group.
BBM_Metric_Cfg_List	This is an array of "BBM_Metric_Cfg_Entry" (defined below). The key to the list is "bandwidth" and is always sorted in descending order (i.e. entries with higher "bandwidth" appears before entries with lower "bandwidth".
BBM_Metric_Cfg_Entry	This defines a single entry in "BBM_Metric_Cfg_List" array (defined above). It is a tuple ["Bandwidth", "Metric"], and defines the metric that should be associated with the individual interfaces of this group, when the total available bandwidth for the group matches "bandwidth" range specified in this entry. Refer to Table 2 for more details.
Default_Metric	The default metric as per configuration. Default metric will be assigned to all interfaces under this group if total available bandwidth for the group Does not match the "Bandwidth" range specified in any "BBM_Metric_Cfg_Entry" for this group. Refer to Table 2 for more details.

Table 1: Interface Group Attributes

2.4. Metric Derivation

Once a interface has been assigned to a interface-group, and the corresponding BBM metric configurations has been provisioned, metric to be associated with the member interfaces can be derived as follows:

```
Sort igp.BBM_Metric_Cfg_List in descending order based on BBM_Metric_Cfg_Entry.B
andwidth
Set Intf.Metric = 0
For (all BBM_Metric_Cfg_Entry in igp.BBM_Metric_Cfg_List
     in descending order)
    - If (igp.BW_Curr >=
          igp.BBM_Metric_Cfg_List.BBM_Metric_Cfg_Entry.Bandwidth)
          - Set Intf.Metric =
                igp.BBM_Metric_Cfg_List.BBM_Metric_Cfg_Entry.Metric.
                           end the loop.
If (Intf.Metric == 0)
    - Set Intf.Metric = igp.Default_Metric.
```

Considering the BBM metric configurations for interface-group IG1 in Figure 2, Table 2 below shows how metric for individual interfaces of IG1 SHALL be computed at any point of time.

Active-Links	Total-Available-BW	BBM-Metric	Remarks
L1, L2, L3	300G	10	Total-Avaiable-BW >= 200
L1, L2, L3(down)	200G	10	Total-Avaiable-BW >= 200
L1, L2(down), L3(down)	100G	50	200 > Total-Avaiable-BW >= 100

Table 2: BBM Metric Calculation

3. Bandwidth Based Routing

Once the metric of individual interfaces are derived from the corresponding interface-group BBM configuration, the same are used in the local IGP SPF computations. In addition to using the metrics in SPF computations, the same are also advertised as the corresponding link cost (instead of the original cost associated with the individual links) in the locally-generated IGP link-state advertisements. This is done to eliminate any looping possible otherwise.

Considering the topology in Figure 1, Table 3 below shows how traffic for destination D1 shall be re-routed based on a series of events and BBM metric configurations as shown in Figure 2.

Event	Interface-Group	Active-Links/Total-Available-BW	Total-Metric	Shortest Path
Initially	IG1	{L1, L2, L3} / 300G	10 + Dopt (R2,D1)	YES
	IG2	{L4, L5, L6} / 300G	20 + Dopt (R2,D1)	NO
L1 goes down	IG1	{L2, L3} / 200G	10 + Dopt (R2,D1)	YES
	IG2	{L4, L5, L6} / 300G	20 + Dopt (R2,D1)	NO
L2 goes down	IG1	{L2, L3} / 200G	50 + Dopt (R2,D1)	NO
	IG2	{L5, L6} / 200G	20 + Dopt (R2,D1)	YES
L4 goes down	IG1	{L3} / 100G	50 + Dopt (R2,D1)	NO
	IG2	{L5, L6} / 200G	20 + Dopt (R2,D1)	YES
L5 goes down	IG1	{L3} / 100G	50 + Dopt (R2,D1)	YES
	IG2	{L6} / 100G	60 + Dopt (R2,D1)	NO

Table 3: BBM based Routing

4. Bandwidth-based Fast Reroute

4.1. Overview

The BBM solution described in Section 2 requires IGPs running on the control plane of the network device, to detect the link failures, determine remaining available bandwidth, re-compute new optimum paths, and finally install the new best paths to the forwarding plane. This may take some time (in the order of 500 ms) for the traffic to switch to a better path.

Also, even if regular FRR mechanism using LFA [RFC5286] and Remote-LFA [I-D.ietf-rtgwg-remote-lfa] has been deployed, the alternate paths chosen is not guaranteed to meet bandwidth constraints. Also, though, [RFC5286] does not specify anything, most LFA implementations in link-state protocols running on the network devices around the world, employs use of a single backup link. Also if there are multiple primary interfaces for a specific destinations, most implementations do not install a alternate path in the forwarding plane. So in the event of the primary link (or one of the multiple primary links) going down, traffic is either switched to a single interface, or not switched to any other link at all. In the first case, there is more likeliness of the single alternate path getting congested (as it might be already carrying some primary traffic for other destinations already). In the latter case, there is more likeliness of causing a congestion on the remaining primary links (e.g. for destination D1, if both L1 and L2 goes down R1 still keeps the traffic on L3 during local repair, trying to push 300G traffic on a single 100G link L5).

Service providers who have stringent bandwidth requirements would need the device to switch the traffic during local repair to multiple alternate paths that have bandwidth constraints satisfied. When the remaining primary OR alternate paths alone cannot satisfy bandwidth requirements, it will also be desirable, to redistribute the traffic over a combination of primary AND alternate paths, during local repair as well as next SPF computations in IGP.

This document proposes a solution the above problem, based on combination of BBM logic (referred to in Section 2) and protection using LFA [RFC5286] and Remote-LFA [I-D.ietf-rtgwg-remote-lfa]. It requires a group of primary links to be protected using multiple non-best feasible alternate paths. The same group of alternate links shall also be pre-installed in forwarding table to facilitate fast re-route (FRR). The details of the solution is specified in the following sub-sections.

4.2. Assumptions and Pre-requisites

Following are some of the assumptions that the solution proposed in this document is based on.

The forwarding plane SHOULD be able handle multiple paths per route and let control plane set the preference for each path over the others. The forwarding machinery shall utilize this, to select a subset of preferred paths, and use them to forward actual traffic at any given point in time. Forwarding machinery SHOULD also load balance traffic with next-hops having same preference

All the links attached to the network device are bundled to create one or more interface-group(s). Also a link MUST belong to one and only one interface-group.

Loops are possible if protection is enabled on all three routes R1, R2 and R3 as shown in Figure 1. To avoid loops implementation MUST have downstream Path Criterion as explained in LFA [RFC5286]

For each interface-group, operator MAY enable protection by configuring the following two parameters.

Minimum-bandwidth: When the remaining bandwidth goes below this the outgoing traffic can no more be carried entirely on this bundle. Some of it shall be distributed across links of other best/non-best interface-groups.

Restore-bandwidth: When the remaining bandwidth exceeds this, the outgoing traffic can entirely be back over the members of this bundle and there is no need to use any other backup for all destinations reachable over the links of the bundle.

4.3. Additional Configuration and Attributes

This document defines the following configuration parameters to be associated with each interface-group for facilitating Bandwidth-based Fast Re-Route. Implementations MUST allow operators to configure these parameters for each interface-group on a network device that implements this solution.

Attribute	Value and Significance
Min_BW	This is the minimum bandwidth below which outgoing traffic MUST not be carried on this interface-group. It needs to load-balance across links of best/non-best interface-groups as well.
Restore_BW	This is the bandwidth above which the outgoing traffic MUST entirely be carried over the members of this interface-group not needing to load-balance across member links of other non-best interface-groups, provided it provides a path with shortest metric.

Table 4: BBM FRR Configurations

In Figure 1, considering all the links L1 to L6 having bandwidth capacity of 100G each, and assigned into two interface-groups IG1 and IG2 (as shown in Section 2.1), following is an example of simple BBM FRR config for each of these interface-group.

```
IG1:
   Member-Links: L1,L2,L3
   Total-Available-BW: 200G,  Metric: 10
   Total-Available-BW: 100G,  Metric: 50
   Default-Metric: 1000
   Protection: Enbaled
     Restore-Bandwidth:  200G
     Min-Bandwidth:      100G

IG2:
   Member-Links: L4,L5,L6
   Total-Available-BW: 200G,  Metric: 10
   Total-Available-BW: 100G,  Metric: 50
   Default-Metric: 1000
   Protection: Enbaled
     Restore-Bandwidth:  200G
     Min-Bandwidth:      100G
```

Figure 3: Example BBM FRR Configuration

This document defines the following attributes to be associated with each interface-group for facilitating Bandwidth-based Fast Re-Route.

Attribute	Value and Significance
BW_PostFail	Cumulative bandwidth through all the remaining primary next-hops considering the primary next-hop with highest bandwidth goes down.
Metric_PostFail	Bandwidth based metric after a link goes down.

Table 5: Additional Interface-Group Attributes

This solution proposed in this document also requires IGPs to define and associated the following attributes for each destination node in the IGP link-state database.

Attribute	Value and Significance
Pri_Nh_Count	Number of primary next-hops found for the destination.
Pri_BW_Curr	Cumulative bandwidth across all the remaining primary next-hops.
Pri_BW_PostFail	Cumulative bandwidth through all the remaining primary nexthops considering the primary nexthop with highest bandwidth goes down.
Pri_BW_Restore	Cumulative restore-bandwidth for all the interface-groups considered for primary nexthops.
Pri_BW_Min	Cumulative minimum-bandwidth for all the interface-groups considered for primary nexthops.

Table 6: Per-node Attributes

4.4. Enhancements to Local Repair in Forwarding Plane

Additionally, the solution proposed in this document also mandates, that the forwarding plane SHOULD implement the following enhanced local-repair logic, to facilitate BBM based fast-re-route, on detecting a link-down event.

For each affected prefix (a prefix is affected if the fated link was one of the preferred active paths used for forwarding).
- Find the actual affected path, and mark it unusable.
- For all other paths downloaded from control-plane,
  - If the preference is same as that of the affected path,
    - Modify its preference to value even lower than normal backup paths.

Finally, go through all remaining active paths
- Select a subset of paths (that share the same highest preference among all),
- Use the selected subset of paths to actually forward traffic.

Figure 4: Enhanced Local Repair in Forwarding Plane

4.5. Influencing Path Preferences

Like mentioned in Section 4.2 the solution proposed in this document relies on the preference-based local-repair logic implemented in forwarding-plane to facilitate fast re-route. This solution requires IGPs to indirectly influence the local-repair action taken by the forwarding-plane by choosing an suitable alternate path with an appropriate preference-value pre-computed and installed in the forwarding-plane, well ahead of the actual link failure event.

Table 7 below, specifies a set path-preference types that this document proposes IGP to define and use while downloading any path for a given destination in the forwarding table.

Preference-Type	Significance
Pri_Nh_Pref	Preference type for normal primary paths.
Bkup_Nh_Pref_High	Preference type for paths, which are preferred, more than normal backup paths but less compared to normal primary paths.
Bkup_Nh_Pref_Normal	Preference type for normal backup paths.

Table 7: Path-Preference Types

4.6. Path Selection and Preference

Based on the above assumptions, additional configuration parameters and attributes the document proposes IGPs to implement the following logic for computing primary and alternate paths for each destination, and determine their corresponding path-preference value as well

```
Step 1:
=======
- For each interface-group "igp"
  - Update "igp.BW_Curr" by adding the bandwidths
    of the individual active member interfaces.
  - Update "igp.BW_PostFail", assuming one of the active
    member interfaces with highest bandwidth goes down next.

Step 2:
=======
- For each destination node "D" in the network (Pass-1)
  - Update D.Pri_BW_Restore and D.Pri_BW_Min from the SPF results.
  - Reset D.Pri_Nh_Count to 0.
  - For each corresponding primary path N,
    - Set "igp" -> Interface-group N belongs to.
    - If igp.BW_Curr > D.Min_BW
      - Set preference of N to Pri_Nh_Pref.
      - Increment the D.Pri_Nh_Count by 1.
    - Else
      - Set preference of N to Bkup_Nh_Pref_Normal.

Step 3:
=======
- For each destination node "D" in the network (Pass-2)
  - Update D.Pri_BW_Restore and D.Pri_BW_Min.
  - For each corresponding primary path N,
    - Set "Pri_Igp" -> Interface-group N belongs to.
    - If protection configured on "Pri_Igp"
      - If igp.Pri_BW_PostFail < D.Pri_BW_Restore,
        OR igp.Pri_BW_PostFail <= D.Pri_BW_Min
        - For all backup paths M,
          - Set "Alt_Igp" -> Interface-group N belongs to.
          - Select M for installing in forwarding plane.
          - If D.Pri_Nh_Count == 0
            - If Alt_igp.BW_Curr >= D.Pri_BW_Restore,
              AND Alt_Igp.BW_Curr > D.Pri_BW_Min
              - Set preference of M to Pri_Nh_Pref.
            - Else
              - Set preference of M to Bkup_Nh_Pref_Normal.
          - Else
            - If Alt_Igp.BW_Curr >= D.Pri_BW_Restore,
              AND Alt_Igp.BW_Curr > D.Pri_BW_Min
              - Set preference of M to Bkup_Nh_Pref_High.
            - Else
              - Set preference of M to Bkup_Nh_Pref_Normal.
```

5. Limitations

The BBM method proposed in this document does NOT ensure end to end bandwidth requirement. It, only ensures that the metric is altered only on local interfaces, based on the BBM metric configurations and remaining available bandwidth.

The solution proposed in this documents attempts to provide protection for single link failures only. It always assumes that link with the highest individual bandwidth capacity shall fail next. In case if any other link with lesser individual bandwidth capacity fails instead, the local repair action taken by the forwarding plane may not be exactly as expected, even though the forwarding plane will still take care of protecting the traffic.

6. Security Consideration

Changes suggested in the draft does not raise any security concerns.

7. IANA Consideration

This draft does not have any request from IANA.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

8.2. Informative References

[I-D.ietf-rtgwg-remote-lfa] Bryant, S., Filsfils, C., Previdi, S., Shand, M., and N. So, "Remote Loop-Free Alternate (LFA) Fast Re-Route (FRR)", draft-ietf-rtgwg-remote-lfa-11 (work in progress), January 2015.

[RFC5286] Atlas, A., Ed. and A. Zinin, Ed., "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, DOI 10.17487/RFC5286, September 2008, <http://www.rfc-editor.org/info/rfc5286>.

Authors' Addresses

Santosh Pallagatti (editor)
Juniper Networks
Embassy Business Park
Bangalore, KA 560093
India

Email: santoshpk@juniper.net

Pushpasis Sarkar (editor)
Juniper Networks
Embassy Business Park
Bangalore, KA 560093
India

Email: psarkar@juniper.net

Hannes Gredler
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, California 94089-1206
USA

Email: hannes@juniper.net

Stephane Litkowski
Orange Business Service

Email: stephane.litkowski@orange.com