Workgroup: Network Working Group
Internet-Draft:
draft-decraene-lsr-lag-indication-02
Published: 31 January 2022
Intended Status: Standards Track
Expires: 4 August 2022
Authors: B. Decraene
Orange
S. Hegde
Juniper Networks Inc.
J. Halpern
Ericsson

LAG indication

Abstract

This document defines a new link flag to advertise that a layer-three link is composed of multiple layer-two sub-links, such as when this link is a Link Aggregation Group (LAG). This allows a large single flow (an elephant flow) to be aware that the link capacity will be lower than expected as this single flow is not load-balanced across the multiple layer-two sub-links. A path computation logic may use that information to route that elephant flow along a different path.

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 https://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 4 August 2022.

Copyright Notice

Table of Contents

1. Introduction

An IP link may be composed a multiple layer two sub-links not visible to the IGP routing topology. When traffic crossing that IP link is load-balanced on a per flow basis, a large elephant flow will only benefit from the capacity of a single sub-link. This is an issue for the routing logic which only see the aggregated bandwidth of the IP link, and hence may incorrectly route a large flow over a link which is incapable of transporting that flow.

This document defines a new link flag to signal that an IP link is a Link Aggregate Group composed of multiple layer two sub-links. This flag may be automatically be set by routing nodes connected to such links, without requiring manual tagging by the network operator. A path computation logic such as a PCE or a CSPF computation on the ingress, may use that information to avoid such links for elephant flows.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 RFC 2119 [RFC2119] RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Protocol extensions

2.1. IS-IS extension

To advertise that a layer-three link is composed of multiple layer-two sub-components this document defines a new bit in the IS-IS link-attribute sub-TLV RFC 5029 [RFC5029].

L2 LAG (Link Aggregation Group) TBD1. When set, this layer-three link is composed of multiple layer-two sub-components performing per flow load balancing.

2.2. OSPF extension

To advertise that a layer-three link is composed of multiple layer-two sub-components this document defines a new bit in the OSPF Link Attributes Bits TLV [I-D.ietf-lsr-dynamic-flooding].

L2 LAG (Link Aggregation Group) TBD2. When set, this layer-three link is composed of multiple layer-two sub-components performing per flow load balancing.

3. Operational considerations

A node supporting this extension SHOULD automatically advertise the L2 LAG flag for IP links composed of multiple layer-two sub-components. Configuration knob MAY be provided to override this default.

In order to handle nodes not supporting this extension, network operator may need to use an admin group (color) [RFC5305] [RFC7308] in order to flag those links on legacy nodes.

3.1. Usage

The information provided by this flag can be used in several different ways, depending upon the technology choices and needs of the operator.

If the operator's usage of LAGs is fairly consistent, one could have a variation on a bandwidth limited flex-algo that specifies minimum bandwidth and the LAG flag not being set. This could then be selected by encapsulating head ends for streams which are judged to need to avoid the LAGs. Likely this would be coupled with a configured value representing the likely limit of LAG components for selecting when to use this flex-algo instance. Note that extending flex-algo requires every node to upgrade.

Another option is if the operator is using traffic engineering (either with a PCE or the head end doing the path selection). The

path selector can select points in e.g. a segment routed path so as to avoid links marked as being LAGs for elephant flows. This can be coupled with a more flexible heuristic for limits than the above. The path selector can look at the advertised link bandwidth, and the presence of the LAG flag, and frequently reliably infer the LAG component size. Thus, it would only need to avoid LAGs where the component is expected to be too small for the large flow being placed.

[Editor's note: This does suggest a possible extension if the working group is interested. We could add a new sub-TLV indicating the lowest bandwidth of the LAG components of a given LAG. This is additional complexity and the question is whether the use cases where this would give noticeably more accurate path estimates and better elephant flow placement are likely.]

4. IANA Considerations

4.1. IS-IS

IANA is requested to allocate one bit value from the registry: link-attribute bit values for sub-TLV 19 of TLV 22 (Extended IS reachability TLV).

Value	Name
TBD1	L2 LAG (Link Aggregation Group)

Figure 1

4.2. OSPF

IANA is requested to allocate one bit number from the registry: OSPF Link Attributes Sub-TLV Bit Values.

Bit Number	Description
TBD2	L2 LAG (Link Aggregation Group)

Figure 2

5. Security Considerations

This extension advertises additional information and capabilities about a link.

An attacker having access to this information would gain knowledge that this link has sub components and that sending a large amount of traffic via a single flow (hence not a DOS) is more likely to overload that sub-component. On the other hand, this overloading

would be limited to this specific sub-component and hence not affect other sub-component.

An attacker been capable of adding this information may gain ability to change the routing of some flow crossing the links, typically large elephant flows specifically configured to avoid such link.

An attacker been capable of removing this information may gain the ability to change the routing and direct a large elephant flow on this link, which would overload a sub component of this link and likely create packet drop for this specific flow.

However, in those two cases, the attacker would equally have the capability to change other routing information such as the link metric, link usability and any link characteristics. Hence this new information does not add new security considerations. Besides, as with others TLV advertisements, the use of a cryptographic authentication as defined in [RFC5304] or [RFC5310] allows the authentication of the peer and the integrity of the message and remove the ability for an attacker to modify such information.

.

6. Acknowledgments

TBD.

7. References

7.1. Normative References

[I-D.ietf-lsr-dynamic-flooding]
Li, T., Przygienda, T., Psenak, P., Ginsberg, L., Chen, H., Cooper, D., Jalil, L., Dontula, S., and G. S. Mishra, "Dynamic Flooding on Dense Graphs", Work in Progress, Internet-Draft, draft-ietf-lsr-dynamic-flooding-10, 7 December 2021, <https://www.ietf.org/archive/id/draft-ietf-lsr-dynamic-flooding-10.txt>.

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

[RFC5029] Vasseur, JP. and S. Previdi, "Definition of an IS-IS Link Attribute Sub-TLV", RFC 5029, DOI 10.17487/RFC5029,

September 2007, <https://www.rfc-editor.org/info/rfc5029>.

[RFC5304] Li, T. and R. Atkinson, "IS-IS Cryptographic Authentication", RFC 5304, DOI 10.17487/RFC5304, October 2008, <https://www.rfc-editor.org/info/rfc5304>.

[RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, DOI 10.17487/RFC5310, February 2009, <https://www.rfc-editor.org/info/rfc5310>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

7.2. Informative References

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC7308] Osborne, E., "Extended Administrative Groups in MPLS Traffic Engineering (MPLS-TE)", RFC 7308, DOI 10.17487/RFC7308, July 2014, <https://www.rfc-editor.org/info/rfc7308>.

Appendix A. Changes / Author Notes

[RFC Editor: Please remove this section before publication]

00: Initial version.

01: Refresh.

Authors' Addresses

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Shraddha Hegde
Juniper Networks Inc.
Exora Business Park
Bangalore 560103
KA
India

Email: shraddha@juniper.net

Joel Halpern
Ericsson

Email: joel.halpern@ericsson.com