Routing Area
Internet-Draft
Intended status: Standards Track
Expires: 4 September 2025

C. Filsfils
P. Camarillo, Ed.
Cisco Systems
D. Bernier
Bell Canada

3 March 2025

Lightweight Host Routing using LLDP
draft-filsfils-rtgwg-lightweight-host-routing-00

Abstract

Link Layer Discovery Protocol (LLDP) is widely deployed today for discovery of information across network elements like routers, switches, and hosts. This document extends LLDP to allow hosts to advertise their IP prefixes to their attached routers which can then propagate the reachability of these host prefixes into routing protocols for enabling network-wide connectivity.

Table of Contents

1. Introduction

Link Layer Discovery Protocol (LLDP) [LLDP] is widely deployed today for discovery of information across network elements like routers, switches, and hosts. LLDP is supported by most switches and routers as well as open source implementations available for hosts. The protocol is often used to discover connections between networking elements, build topology information as well as for monitoring and troubleshooting.

In a typical layer-3 data center (DC), servers (i.e., hosts) are connected to the leaf routers as show in Figure 1 below. These Layer-3 DCs, which typically run BGP routing protocol, are described in [RFC7938]. These servers run applications either natively or within containers or virtual machines (VMs). These applications are allocated IP addresses from the IP prefixes assigned to the servers. Furthermore, these server IP prefixes need to be advertised into the DC network and beyond via routing protocols to provide reachability for the application. The server IP prefixes used by applications need not be in the same subnet as the layer-3 link that connects the server hosts to leaf routers. This requires a mechanism for the leaf routers to discover the server IP prefixes connected to it.

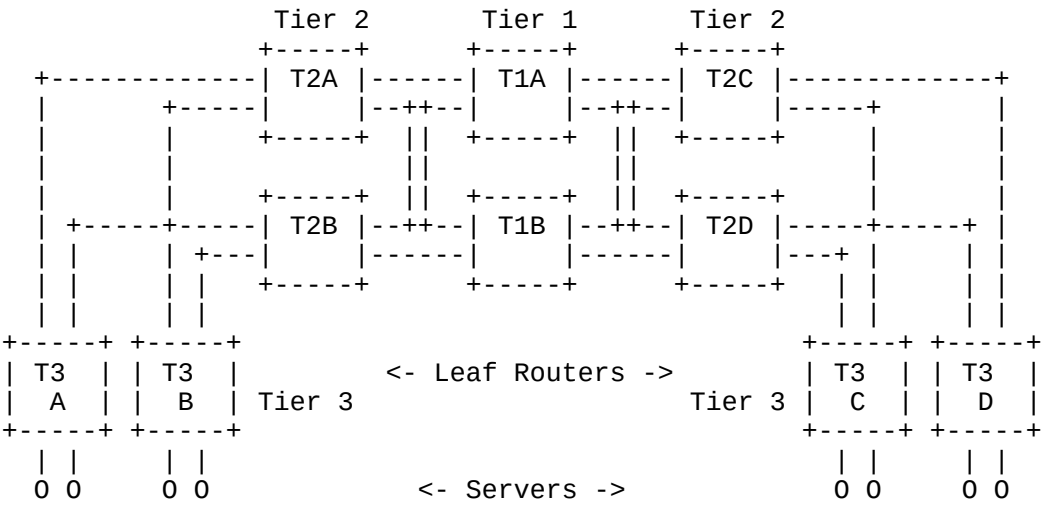

Figure 1: 5-stage Clos Layer-3 DC Topology

Typically, the advertisement of server IP prefixes is done by running a BGP stack on the server and establishing BGP sessions to BGP running on the leaf routers. This requires the provisioning of a BGP stack on the host, the configuration a BGP session between the host and the leaf router. There is also the requirement and expectation that the host is able to announce and withdraw the IP prefixes used by applications running on it in a dynamic manner. The leaf routers themselves also run DC routing protocols (BGP being a popular choice) for the further advertisement of the IP prefixes within the DC network and beyond.

The deployment and use of LLDP is quite common in the layer-3 DC networks as it aids in topology discovery and troubleshooting. Open source LLDP implementations are also widely deployed between the leaf routers and the hosts connected to them. This document introduces LLDP extensions to enable the hosts to advertise their prefixes that can be discovered by LLDP running on the leaf routers and used for routing of traffic to those prefixes towards the host. The routers can further advertise or withdraw these host IP prefixes discovered via LLDP into the DC routing protocols like BGP [RFC4271], IS-IS [ISO10589] or OSPF [RFC2328] [RFC5340]. This avoids the provisioning and management of BGP on the hosts towards the leaf routers, thereby simplifying operations in certain deployments.

As LLDP is not a routing protocol, the specifications in this document is applicable to layer-3 DCs where the hosts are connected via layer-3 interfaces to the leaf routers and the requirement is simply to provide server IP prefix reachability. This solution works

for both IPv4/IPv6 prefixes and enables application/container/VM orchestration mechanisms on the hosts to advertise basic routing information related to these prefixes. These orchestration mechanisms typically interact with the LLDP daemon running locally in the user space on the host to announce and withdraw prefix reachability on demand. The details of these orchestration mechanisms are outside the scope of this document.

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 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. LLDP Extensions For Lightweight Host Routing

The IPv4 and IPv6 Host Prefix TLVs are used by a host to advertise its own local IP Prefixes in the Link Layer Discovery Protocol data unit (LLDPDU) [LLDP]. The format of these TLV is as follows:

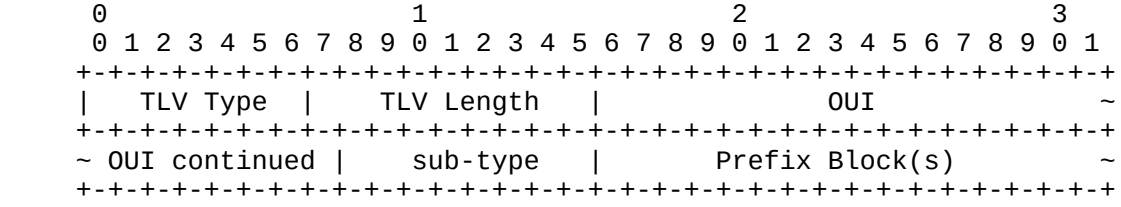

Figure 2: IPv4/IPv6 Host Prefix TLV

Where:

* TLV Type: 7 bits size carrying the value 127 that indicates vendor-specific TLV

* TLV Length: 9 bits size carrying the length of the TLV after the TLV Length field in terms of octets

* Organization Unique Identifier (OUI): 3 octet field carrying the hexadecimal value 0x00005E that indicates IANA as the organization managing the underlying allocation space

* Sub-Type: 1-octet field that carries the IANA allocated LLDP TLV sub-type TBD1 for IPv4 and TBD2 for IPv6

* Prefix Block(s): at least one or more prefix blocks as specified below

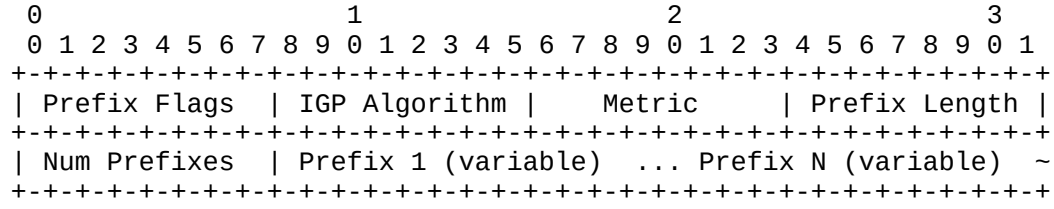

Figure 3: IPv4/IPv6 Host Prefix TLV Prefix Block

Where:

* Prefix Flags: 1-octet field carrying the IPv4 or IPv6 prefix flag as described below:

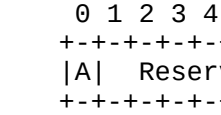

Figure 4: IPv4 Prefix Flags

- A-Flag: Anycast flag. If set, then the prefixes in the block are anycast.

- Reserved bits: Reserved for future use and MUST be zero when originated and ignored when received.

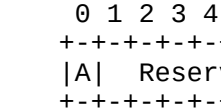

Figure 5: IPv6 Prefix Flags

- A-Flag: Anycast flag. If set, then the prefixes in the block are anycast.

- Reserved bits: Reserved for future use and MUST be zero when originated and ignored when received.

- L-Flag: Locator Flag. If set, then the prefixes in the block are SRv6 Locators [RFC8986].

* IGP Algorithm: 1-octet value providing the algorithm associated with the prefixes in the block. The algorithm values are from IGP Algorithm Types registry under the IANA Interior Gateway Protocol (IGP) Parameters.

* Metric: 1-octet carrying the metric value from the range 0 to 255 associated with the prefixes in the block.

* Prefix Length: 1-octet carrying the length in bits of each of the prefixes in the block. The valid values are 1-32 for IPv4 and 1-128 for IPv6.

* List of Prefixes: One or more prefixes where each prefix is encoded up to the number of bits as indicated by the Prefix Length followed by the minimum number of trailing bits needed to make the end of each prefix field falls on an octet boundary. Any trailing bits MUST be set to 0. Thus, each prefix field contains the most significant octets of the prefix, i.e., 1 octet for prefix length 1 up to 8, 2 octets for prefix length 9 to 16, 3 octets for prefix length 17 up to 24, 4 octets for prefix length 25 up to 32, and so on.

To ensure efficient encoding of the LLDPDU, the following rules apply:

* All prefixes that have the same properties (i.e., prefix length, flags, metric, and algorithm) MUST be encoded in a single prefix block unless doing so makes the block larger than the size that can be accommodated in a single IPv4/IPv6 Host Prefix TLV.

* More than one instance of the IPv4 or IPv6 Host Prefix TLV MUST NOT be used unless the prefix blocks to be advertised do not fit into a single TLV instance.

If the same prefix is present in multiple TLV instances or prefix blocks, only the first occurrence of that prefix in the LLDPDU MUST be considered and the rest MUST be ignored.

LLDP has been extended to support multi-frame LLDPDUs [LLDP-MULTIFRAME]. The above rules also apply to multi-frame LLDPDUs.

3. Procedures

The LLDP extensions for lightweight host routing in this document enable the advertisement of the prefixes from the host towards its directly connected router. The host includes its local prefixes in the LLDPDU that it sends on its port(s) connected to the router. The router on receiving these LLDPDU discovers the host prefixes and programs them in its forwarding table with the outgoing interface pointing towards the port over which the LLDPDU was received and with the nexthop as the IP address on the host on that port. This nexthop address MAY be the one that is received via the Management Address TLV of LLDP or discovered via a protocol like ARP or ND on that specific interface. The same prefix may be learnt via LLDP from the same or different hosts over different ports; these MAY be installed as an equal cost multipath (ECMP) route by the router.

Further the router MAY advertise these host prefixes learnt via LLDP into other protocols like BGP, OSPF, or IS-IS via route redistribution. The details of route redistribution mechanism for conveying information like metric and algorithm along with the prefix are local to the router implementation and outside the scope of this draft.

LLDPDUs are sent periodically by the host and the router including the host IP prefixes that are active on the host. The host MAY also trigger an LLDPDU on demand when the set of IP host prefixes that are active change (e.g., when a prefix is removed, or a new prefix is provisioned).

The processing of the host IP prefix information on the receiving router side follows the LLDP specification [LLDP]. This includes performing a mark and sweep operation between the existing set of host IP prefixes learnt on a specific port previously against the set of IP prefixes received on the same port in a subsequent LLDPDU. Any newly learnt prefixes are installed in the forwarding and made available for advertisement into other routing protocols via redistribution. Any prefixes that are no longer being received via LLDPDU on that port are deleted from the forwarding and withdrawn from routing protocols where they might have been previously redistributed into.

The semantics of the mandatory Time To Live TLV of LLDP [LLDP] also affect the IP host prefix information learnt via LLDP. This includes removing all the learnt IP prefixes if an LLDPDU is not received within the period specified in the previous LLDPDU. Additionally, an implementation SHOULD delete the learnt host IP prefixes as soon as the port over which they are learnt goes down.

There is no change the LLDPDUs sent from the routers towards the host.

4. IANA Considerations

This document requests IANA to allocate code points from the "Link Layer Discovery Protocol (LLDP) TLV Subtypes" registry of the "IANA OUI Ethernet Numbers" registry group.

Code Point	Description	Reference
TBD1	IPv4 Host Prefix TLV	this document
TBD2	IPv6 Host Prefix TLV	this document

Figure 6: LLDP Extensions Code Points

This document also requests the creation of two registries called "LLDP IPv4 Host Prefix Flags" and "LLDP IPv6 Host Prefix Flags" under the "IANA OUI Ethernet Numbers" registry group. The allocation policy for these registries is "Expert Review" according to [RFC8126] with the guidance for Designated Experts being the same as for the LLDP TLV Subtypes registry in [RFC9542].

The initial allocations are as follows:

Bit	Description	Reference
0	Anycast (A-Flag)	This document
1-7	Unassigned	

Figure 7: LLDP IPv4 Host Prefix Flags

Bit	Description	Reference
0	Anycast (A-Flag)	This document
1-6	Unassigned	
7	SRv6 Locator (L-Flag)	This document

Figure 8: LLDP IPv6 Host Prefix Flags

5. Manageability Considerations

The extensions in this document MUST NOT be enabled by default. Implementations on both the host and router side MUST provide a per-port configuration option to enable this feature. The implementation on the router side SHOULD log the activity of prefix discovery for monitoring and troubleshooting purposes.

6. Security Considerations

The extensions in this document introduce additional information in LLDP. The IEEE 802.1AE [MACsec] standard can be used for encryption and/or authentication to provide privacy and integrity. MACsec utilizes the Galois/Counter Mode Advanced Encryption Standard (AES-GCM) for authenticated encryption and Galois Message Authentication Code (GMAC) if only authentication, but not encryption is required.

The MACsec Key Agreement (MKA) is included as part of the IEEE 802.1X-20200 Port-Based Network Access Control Standard [MKA]. The purpose of MKA is to provide a method for discovering MACsec peers and negotiating the security keys needed to secure the link.

A rogue host may inject arbitrary and invalid prefixes into its connected router that could result in diversion of traffic and disruption for applications and services. This feature is expected to be used in environments where the router and the hosts are secured and within a single administrative control - e.g., a DC.

7. Acknowledgements

The authors of this document would like to acknowledge the review and inputs provided by Ketan Talaulikar during the early stages of this work.

8. References

8.1. Normative References

[LLDP] IEEE, "IEEE Standard for Local and metropolitan area networks - Station and Media Access Control Connectivity Discovery", IEEE 802.1AB-2016, DOI 10.1109/IEEESTD.2016.7433915, 11 March 2016, <https://doi.org/10.1109/IEEESTD.2016.7433915>.

[LLDP-MULTIFRAME] IEEE, "IEEE Standard for Local and metropolitan area networks-- Station and Media Access Control Connectivity Discovery Amendment 2: Support for Multiframe Protocol

Data Units", IEEE 802.1ABdh-2021,
DOI 10.1109/IEEESTD.2022.9760302, 19 April 2022,
<https://doi.org/10.1109/IEEESTD.2022.9760302>.

[MACsec] IEEE, "IEEE Standard for Local and metropolitan area networks - Media Access Control (MAC) Security", IEEE Standard 802.1AE-2018, 27 September 2018.

[MKA] IEEE, "IEEE Standard for Local and metropolitan area networks - Port Based Network Access Control", IEEE Standard 802.1X-2020, 30 January 2020.

[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>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[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>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

[RFC9542] Eastlake 3rd, D., Abley, J., and Y. Li, "IANA Considerations and IETF Protocol and Documentation Usage for IEEE 802 Parameters", BCP 141, RFC 9542, DOI 10.17487/RFC9542, April 2024, <https://www.rfc-editor.org/info/rfc9542>.

8.2. Informative References

[ISO10589] International Organization for Standardization, "Intermediate System to Intermediate System intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service (ISO 8473)", ISO/IEC 10589, November 2002.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

Authors' Addresses

Clarence Filsfils
Cisco Systems
Belgium
Email: cf@cisco.com

Pablo Camarillo (editor)
Cisco Systems
Spain
Email: pcamaril@cisco.com

Daniel Bernier
Bell Canada
Canada
Email: daniel.bernier@bell.ca