

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: April 23, 2020

R. Bush  
Arrcus & IIJ  
K. Patel  
Arrcus  
October 21, 2019

BGP Topology Discovery Requirements  
draft-ymbk-lsvr-discovery-req-03

## Abstract

For wide scale routing protocols to build their topology and reachability databases they need to discover the encapsulation data on a link, link IP layer 3 attributes, attributes for IP layer 3 and above protocols on that link, and link liveness. We refer to this as neighbor discovery. BGP-LS and its enhancements provide an API to present much of these data to BGP protocols, but do not directly collect these data. This document explores the needs and criteria for the data needed.

## 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 April 23, 2020.

## Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

Internet-Draft

BGP Topology Discovery Requirements

October 2019

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

<a href="#">1.</a>	Introduction . . . . .	<a href="#">2</a>
<a href="#">2.</a>	Architectural Considerations . . . . .	<a href="#">2</a>
<a href="#">3.</a>	Requirements . . . . .	<a href="#">4</a>
<a href="#">4.</a>	Security Considerations . . . . .	<a href="#">5</a>
<a href="#">5.</a>	IANA Considerations . . . . .	<a href="#">5</a>
<a href="#">6.</a>	Acknowledgments . . . . .	<a href="#">6</a>
<a href="#">7.</a>	References . . . . .	<a href="#">6</a>
<a href="#">7.1.</a>	Normative References . . . . .	<a href="#">6</a>
<a href="#">7.2.</a>	Informative References . . . . .	<a href="#">6</a>
	Authors' Addresses . . . . .	<a href="#">6</a>

## [1.](#) Introduction

In a massive scale datacenter or similar environment BGP([RFC4271](#)) and BGP-like protocols, e.g. BGP-SPF (see [I-D.ietf-lsvr-bgp-spf](#)), provide massive scale-out without centralization using a tried and tested scalable distributed control plane transport, offering a scalable routing solution. But BGP4, BGP-SPF, and similar protocols need layer 3 topology discovery; meaning IP encapsulations on a link, layer 3 IP addressing data on a link, attributes for IP layer 3 and above protocols on that link, and assured link liveness from the network to build and maintain the routing topology.

BGP-LS [RFC7752](#) and its extensions provide an API which BGP4 and BGP-SPF can use to get the and distribute topology data. But BGP-LS itself does not gather the data, it merely presents it. So the IP topology data must be gathered.

What topology data do BGP-like protocols actually need? What level of freshness is needed? What are the requirements for scale, extensibility, security, etc?

## [2.](#) Architectural Considerations

Massive Data Centers (MDCs) have on the order of 10,000 racks, often with two Top Of Rack (TOR) devices per rack, on the order of 40

hardware servers per rack, each with order 100 virtual services or machines, each with their own layer 3 IP address. Given service mobility, any initial IP address aggregation fragments over time. To provide this level of scaling reliably, stably, and security imposes architectural constraints on any discovery protocol.

- o Deployable - If it is not easily deployable, it is a pointless exercise. To be deployable, it must be easy to provision (e.g. zero touch provisioning, Open Config, YANG, et alia), easily reconfigured, easily measured, and easily monitored.
- o Simple - If it isn't simple, it will not scale. Simplicity requires restraint in design. 'Union Protocols' which are the sum of everyone's desires are complex disasters waiting to happen. Often they do not wait. Prefer 'Intersection Protocols' which include only those things which everyone absolutely needs.
- o Securable - Security properties should be analyzed. Again, simplicity is key; complex protocols increase in complexity over time, and security vulnerabilities increase significantly with complexity. As [\[RFC5218\]](#) 2.2.3 says "The more successful a protocol becomes, the more attractive a target it will be."
- o Extensible - As [\[RFC5218\]](#) [Section 2.2.1](#) said, successful protocols are extensible beyond the original expectation. MDC and similar needs are expanding and we are still learning about the space. Simplicity and extensibility should go a long way to adaptability; complex protocols are hard to extend, especially when they are poorly understood.
- o Implementable - It must be reasonably easy to implement and deploy. Some implications are:
  - \* Packet formats should be easy to generate and easily parsable. Type/length/Value (TLV) formats are preferred.
  - \* The protocols should be free to use and deploy; i.e. not be constrained by Intellectual Property Right (IPR) claims.
  - \* Again, simpler protocols are simpler to implement, deploy, measure, monitor, etc.

- \* Performance Problems arise if the protocol was not designed to scale.
- o Low Impact - The MDCs chose BGP for, among other reasons, it is quiet and only transmits changes, not repeated flooding of the same information. This allows great scale. The discovery protocol should be similar in this regard, not flood or chatter the same information repeatedly. It should support fast and quiet session restoration in case of link failure and restoration when there has been no actual change in end point attributes.

- o Compatible - It must be compatible with the various routing technologies used in MDCs. The new discovery protocol will discover the IP layer 3 encapsulations, learn layer 3 addressing data from the network, confirm link liveness, in order to allow upper layer routing protocol(s), e.g. BGP4, BGP-SPF, and BGP-LS, to build maintain and distribute the topology.

### 3. Requirements

The target for the discovery protocol(s) is a massive datacenter scale deployment using BGP or similar routing, e.g. BGP4 or [\[I-D.ietf-lsvr-bgp-spf\]](#); but should be generally usable by other routing protocols, e.g. EVPN, and in other similar environments.

The IETF is very good at finding corner cases which expand needs and complicate protocols. This effort should resist this tendency.

It would be easiest for the BGP-like protocols to consume the data if they are presented via the BGP-LS [\[RFC7752\]](#) API as used in [\[I-D.ietf-lsvr-bgp-spf\]](#) [Section 4](#).

BGP-like protocols will need at least the following information about the topology:

Node Identity: Each node in the topology must have an identity/identifier which must be unique in the topology.

A node must have one or more links to other nodes or it is, *ab*definito, not in the topology.

A node has IP layer 3 attributes such as encapsulations and IP addresses.

**Link Identity:** A link is between two nodes. Each end of a link is a node/device interface.

Each link in the topology must be uniquely identified and the identities of the nodes on the link must be identified. This includes LAGged links.

As MAC addresses are not unique in actual deployment, they may not be assumed to uniquely identify a link. Multiple VLANs between a port pair on two devices are a simple example of this. Link Aggregation Groups (LAGs), Multi-Chassis LAGs etc. must be accommodated.

A link might be on a tunnel interface; though the tunnel type may be restricted..

**Link Liveness:** Because adjacencies and topology changes must be quickly detected, The stability of each link should be able to be monitored and reported. As this can be noisy, it must be able to be tuned by the operator, and expensive operations should be minimized.

**Encapsulations:** The encapsulation(s) (IPv4, IPv6, ...) on each link must be known. One or more of the common AFI/SAFIs must be supported on each link, IPv4, IPv6, MPLS, etc.

It is assumed that the set of encapsulations is the same across the entire topology.

**Addresses:** The available addresses on the node interfaces for each encapsulation must be known. More than one address for an encapsulation type must be supported.

As BGP-like protocols will be peering between the nodes, there may be a preferred encapsulation and address on an link, or a loopback interface may be used.

**Upper Layer Protocol Parameters** To facilitate peering of upper layer

protocols across a link, e.g. BGP, the protocol should support signaling of the parameters for these protocols, e.g. peer AS number, peering address(es), etc.

Mobility: Fast detection of [micro-]service mobility must be supported.

EVPNs: EVPN end-point discovery must be supported.

#### [4.](#) Security Considerations

While this document has no security considerations per se, it does make a plea for securability in protocol design.

Mis-wires, malicious devices being plugged into ports, and monkey in the middle attacks should be considered.

There should at least be assurance that the end-point a device opened a session with six months ago is the same one sending PDUs today.

#### [5.](#) IANA Considerations

This document has no IANA considerations.

#### [6.](#) Acknowledgments

The authors thank Victor Kuarsingh and Gunter Van De Velde for reviews.

#### [7.](#) References

##### [7.1.](#) Normative References

[I-D.ietf-lsvr-bgp-spf]

Patel, K., Lindem, A., Zandi, S., and W. Henderickx,  
"Shortest Path Routing Extensions for BGP Protocol",  
[draft-ietf-lsvr-bgp-spf-06](#) (work in progress), September  
2019.

- [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, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", [RFC 7752](#), DOI 10.17487/RFC7752, March 2016, <<http://www.rfc-editor.org/info/rfc7752>>.

## 7.2. Informative References

- [RFC5218] Thaler, D. and B. Aboba, "What Makes for a Successful Protocol?", [RFC 5218](#), DOI 10.17487/RFC5218, July 2008, <<http://www.rfc-editor.org/info/rfc5218>>.

### Authors' Addresses

Randy Bush  
Arrcus & IIJ  
5147 Crystal Springs  
Bainbridge Island, WA 98110  
United States of America

Email: [randy@psg.com](mailto:randy@psg.com)

Keyur Patel  
Arrcus  
2077 Gateway Place, Suite 400  
San Jose, CA 95119  
United States of America

Email: [keyur@arrcus.com](mailto:keyur@arrcus.com)

