

Source Packet Routing in Networking  
Internet-Draft  
Intended status: Standards Track  
Expires: 1 January 2026

A. Stone  
Nokia  
V. P. Beeram  
Juniper Networks  
30 June 2025

Multipath Traffic Engineering for Segment Routing  
draft-stone-spring-mp-te-sr-00

## Abstract

This document describes a mechanism to achieve Multipath Traffic Engineering for Segment Routing based networks.

## Discussion Venues

This note is to be removed before publishing as an RFC.

Discussion of this document takes place on the Source Packet Routing in Networking Working Group mailing list ([spring@ietf.org](mailto:spring@ietf.org)), which is archived at <https://mailarchive.ietf.org/arch/browse/spring/>.

Source for this draft and an issue tracker can be found at <https://github.com/astone282/draft-stone-spring-mp-te-sr>.

## 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 1 January 2026.

## Copyright Notice

Copyright (c) 2025 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 to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

## Table of Contents

1. Introduction . . . . .	2
1.1. Requirements Language . . . . .	3
2. Terminology . . . . .	3
3. MP-TE vs Multiple SID lists . . . . .	3
4. MP-TE concepts with Segment Routing . . . . .	4
4.1. MPTED . . . . .	4
4.2. Junction Segment . . . . .	4
4.3. MPTE SR Policy - Tunnel with multiple ingress/egress . .	5
5. Operation . . . . .	6
5.1. Example . . . . .	7
6. Load-balancing . . . . .	10
7. Constraints . . . . .	10
8. Protection . . . . .	10
9. Other considerations . . . . .	10
9.1. Hierarchy . . . . .	11
9.2. Directly connected Junction nodes . . . . .	12
9.3. Broadcast links . . . . .	12
9.4. Local optimization . . . . .	12
9.5. Global optimization . . . . .	12
10. Security Considerations . . . . .	13
11. Manageability Considerations . . . . .	13
12. IANA Considerations . . . . .	13
13. References . . . . .	13
13.1. Normative References . . . . .	13
13.2. Informative References . . . . .	15
Acknowledgments . . . . .	15
Authors' Addresses . . . . .	15

## 1. Introduction

The document [I-D.draft-kompella-teas-mp-te] introduces a multipath traffic engineering concept that combines the benefits of both Equal-Cost Multipath (ECMP) forwarding and traffic-engineered paths. This approach uses a Directed Acyclic Graph (DAG) based forwarding mechanism, with the DAG signaled to participating network nodes. The concept is to move beyond simple ECMP paths by incorporating both ECMP and non-ECMP paths while still adhering to traffic engineering

constraints, to provide added resiliency while also permitting better usage of link bandwidth.

[I-D.draft-kompella-teas-mppte] outlines the architecture design which can be applied to both distributed and centralized signaling for various tunnel types, including MPLS, IP, and others while leaving the specific details of each out of scope.

This document proposes and discusses a centralized computation and signaling mechanism for SR-based networks, primarily utilizing existing constructs and capabilities. As MPTE evolves, new extensions to SR-based documents may be needed, both in terms of architecture and protocol-specific semantics.

The document assumes the reader is familiar with [RFC8402], [RFC9256], and [I-D.draft-kompella-teas-mppte].

### 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. Terminology

- \* For MPTE terminology such as MPTEd, DAG, MC, MID, Junction node and others see [I-D.draft-kompella-teas-mppte].
- \* For SR terminology see [RFC8402] and [RFC9256].

## 3. MP-TE vs Multiple SID lists

It's important to recognize the SR Policy information model supports multiple SID lists, effectively encoding multiple unique paths on a tunnel at the ingress. A Directed Acyclic Graph (DAG) can be represented as a collection of individual paths, each of which can be programmed as a separate SID list within an SR Policy Candidate Path. However, depending on the graph topology, the number of unique paths to encode can grow significantly. Additionally, in traditional SID list approaches, hashing is performed only at the ingress, rather than at each downstream node. In contrast, a DAG-based mechanism may allow better traffic distribution or localized tuning based on localized weight changes. Finally, the maximum segment depth (MSD) may need to be considered for long paths that deviate significantly from the shortest path.

In comparison, encoding a DAG's forwarding instructions across the participating Junction nodes reduces the number of individual SID lists at the ingress, but at the cost of increasing state in the network. While source-based routing aims to reduce network state, there is a trade-off between the volume and length of SID lists versus distributing that state throughout the network to achieve multipath traffic engineering use cases.

The choice between using multiple ingress segment lists or the MPTE DAG-based distribution mechanism depends on the traffic engineering requirements, overall network design, link/path metrics, and the DAG's structure.

#### 4. MP-TE concepts with Segment Routing

This document proposes the below concepts for applying MPTE in an SR environment.

##### 4.1. MPTED

The MPTED is managed by a centralized controller, such as a PCE acting as the MC. Topology discovery is performed using BGP-LS [RFC7752], while transport control plane signaling is achieved through controller-oriented protocols such as PCEP [RFC5440], [RFC8231] and BGP/BGP-LS [I-D.draft-ietf-idr-segment-routing-te-policy], [I-D.draft-ietf-idr-bgp-ls-sr-policy]. The MC computes, manages, and distributes all forwarding information to the nodes participating in the MPTE DAG, which form the MPTED.

[I-D.draft-kompella-teas-mppte] specifies that a node in the MPTED is identified by its IPv6 loopback address. However, this document allows the use of a 32-bit dotted quad router ID as an alternative. This value represents the headend address of a node participating in the DAG.

As per [I-D.draft-kompella-teas-mppte], the controller acting as the MC is responsible for assigning the MPID and incrementing the MPTED unique ID version.

##### 4.2. Junction Segment

The concept of a Junction Segment is introduced to describe the signaling and forwarding behavior of a Junction node in an SR network.

It's worth noting that the architectural use of a Junction Segment is analogous to a Replication Segment [RFC9524], but it performs forwarding based on load balancing rather than replication.

A Junction Segment is installed on nodes identified as Junction Nodes, as defined in [I-D.draft-kompella-teas-mppte].

This document version proposes that a Junction Segment is realized using the existing SR Policy construct with a single Candidate Path. Future document versions may evolve the Junction segment into its own explicit segment type as opposed to leveraging the existing SR Policy constructs. A Binding Segment is attached to an SR Policy Candidate Path with one or more SID Lists. OIF instruction signaling is achieved via segment lists, where the top SID identifies the outgoing interface(s).

Since a Junction Segment may egress to multiple downstream nodes, the endpoint of the corresponding SR Policy MUST be set to the null value (0.0.0.0). Therefore, a Junction segment is identified by its <headend, color> attribute. This document currently assumes that the color value be encoded with both the MID and version. However, further updates may include additional encoding of these independent values in the SR Policy model.

Since SR-based networks support specifying multiple egress interfaces using adjacency-SID sets and Node SIDs, the Junction Segment MAY include a SID list entry that identifies multiple outgoing interfaces. In addition to egress interfaces, [I-D.draft-kompella-teas-mppte] describes signaling ingress interfaces. The use of a Junction Segment omits the need for per-interface ingress signaling as a single Binding Segment attached to an SR Policy is used. All upstream originated traffic sent to a downstream Junction Node uses the same, single Junction Segment value which is a Binding Segment.

#### 4.3. MPTE SR Policy - Tunnel with multiple ingress/egress

[I-D.draft-kompella-teas-mppte] specifies that an MPTE Tunnel could have multiple ingress and/or multiple egress nodes. Currently, the SR Policy architecture defines an SR Policy using a {Headend, Endpoint, Color} tuple, where the Endpoint may be set to the null value (0.0.0.0), indicating multiple destinations.

For controller-initiated tunnels, the intended ingress and egress node(s) can be provided to the controller based on implementation-specific methods. These may be signaled to the network as multiple tunnels to support multi-ingress scenarios. Each tunnel MAY use a null Endpoint value to support multi-egress.

However, MPTE SR Policies that are originated or defined by network devices are typically limited to a single ingress and a single egress endpoint unless protocols such as PCEP or NETCONF are extended to encode additional intended destination node(s) for controller-based path computation. Extensions to the SR Policy architecture may be needed to support signaling of multiple intended destinations for path computation.

Support for network-originated or device-defined MPTE SR Policies with multiple ingress nodes is currently out of scope.

## 5. Operation

A path computation request or tunnel delegation notification is sent to the controller, specifying one or more ingress and egress nodes, along with constraints. This request may originate from an ingress router in the network or be provisioned directly via an API to the controller. This tunnel computation request or delegation pertains to an instance of an MPTED, which is assigned an identifier and a version.

The controller computes the DAG for ingress and all egress endpoints to determine all Junction nodes in the DAG to be used for the tunnel.

The controller signals the Junction Segment to all downstream nodes, starting with the penultimate egress hop node(s) and working upwards toward the ingress nodes.

Junction Segment deployments are in the form of a unicast SR Policy with a single Candidate Path using protocols such as PCEP, BGP, or NETCONF. The optional use of an MPTED Reflector is protocol specific. For example, PCEP sessions terminate on each and every Junction node in the topology and BGP may do the same or make use of a BGP Route Reflector. A BSID MUST be explicitly requested or signaled to the Junction node for assignment. If the controller opts for local node assigned value, it MUST wait to signal upstream Junction nodes about their Junction segments in their outgoing SID lists.

Each SR Policy contains one or more SID lists. These SID lists must include at least two segment identifiers: one SID for forwarding to the downstream Junction node and one for the Junction Segment value (BSID) of the downstream node. For directly connected neighbors, this may be the adjacency or node SID for the neighbor. For Junction nodes that are not directly connected, additional SIDs MAY be used to steer the packet along an ECMP or non-ECMP path to the downstream Junction node. It is worth noting that if the SID list comprises of only an Adjacency-SID and the Junction SID of the neighbor node, then the dataplane packet contains only one SID on egress which is the Junction SID of the neighboring node.

### 5.1. Example

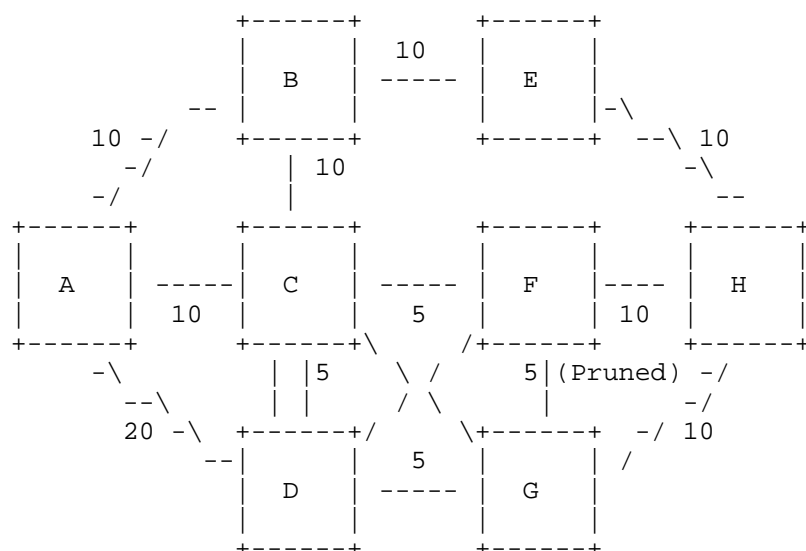


Figure 1 : Example Topology to apply DAG

Figure 1 presents a sample topology for one ingress and one egress MPTE Tunnel established as an SR Policy. The MPTE tunnel is from A to H. The figure presents the bi-directional link weights for an arbitrary metric (IGP, TE, Delay etc).

Note there is a mesh between C, D, F and G of weight 5 per link. Pruned represents an excluded link due to TE constraints. In the below, the terminology {u,v} represents a unidirectional link between U and V. The terminology Adj-SID-XY represents an adjacency SID from X to Y. Node-SID-X represents the node SID path to X.

A MPTE DAG is computed to contain nodes B,C,D,E,F,G with the links {G, F} / {F, G} pruned.

The below Junction Segments are deployed to the network realized with an SR Policy with a single Candidate Path containing multiple segment lists. Note the following:

- \* When the DAG is computed, loops cannot exist. Therefore, in the above topology the links in direction {C, B} and {C,D} and {C,F} and {C,G} are chosen in the DAG.
- \* The path along {B,E,H} can be represented by a single Node SID. A Junction on node E is not required.
- \* Junction F is described below, but an optimization could be performed to exclude Junction F as it only has one egress link. A generalized mechanism to optimize the DAG distribution could omit Junction segments if the Junction segment contains only one egress SID list.
- \* In practice the Binding Segments MAY be all the same value. This example describes different BSID values for readability.
- \* Weights of each egress SID list is also currently omitted.



## Junction Segment B:

BSID: BSID-B

SID List 1: [Node-SID-H]

## Junction Segment F:

BSID: BSID-F

SID List 1: [Adj-SID-FH]

## Junction Segment G:

BSID: BSID-G

SID List 1: [Adj-SID-GH]

## Junction Segment C:

BSID: BSID-C

SID List 1: [Adj-SID-CB, BSID-B]

SID List 2: [Adj-SID-CF, BSID-F]

SID List 3: [Adj-SID-CG, BSID-G]

SID List 4: [Node-SID-D, BSID-D]

## Junction Segment D:

BSID: BSID-D

SID List 2: [Adj-SID-DF, BSID-F]

SID List 3: [Adj-SID-DG, BSID-G]

Then lastly, at ingress the SR Policy transport tunnel is configured with the following:

## Ingress SR Policy (Using Junction Segments):

## Candidate Path 1:

SID List 1: [Adj-SID-AB, BSID-B]

SID List 2: [Adj-SID-AC, BSID-C]

SID List 3: [Adj-SID-AD, BSID-D]

In comparison, if the above DAG was encoded at ingress then the following individual segment lists could be used to represent the above DAG. Note, some of the below could be compressed with a Node SID(s) but listed with adjacency for explicit example.

## Ingress SR Policy (Ingress only):

## Candidate Path 1:

SID List 1: [Adj-SID-AC, Adj-SID-CF, Adj-SID-FH]

SID List 2: [Adj-SID-AC, Node-SID-D, Adj-SID-DG, Adj-SID-GH]

SID List 3: [Adj-SID-AC, Node-SID-D, Adj-SID-DF, Adj-SID-FH]

SID List 4: [Adj-SID-AC, Adj-SID-CF, Adj-SID-CG, Adj-SID-GH]

SID List 5: [Adj-SID-AC, Adj-SID-CF, Adj-SID-BE, Adj-SID-EH]

SID List 6: [Adj-SID-AB, Node-SID-H]

SID List 7: [Adj-SID-AD, Adj-SID-DG, Adj-SID-GH]

SID List 8: [Adj-SID-AD, Adj-SID-DF, Adj-SID-FH]

## 6. Load-balancing

When a packet with the BSID assigned to the Junction Segment is received at its ingress, the node performs weighted ECMP forwarding among all egress SID lists associated with the SR Policy.

It is worth noting that [I-D.draft-kompella-teas-mppte] introduces the concepts of both unequal weight balancing and 0 weight to omit forwarding out of an egress interface while maintaining the instruction signaling. The 0 weight capability is not supported in the current SR Policy model and may be considered in future updates to this document.

## 7. Constraints

When the controller computes the DAG, traffic engineering constraints MUST be considered. Links which violate the constraints are pruned from the DAG. Nodes which do not form the DAG are not notified with any Junction segments.

## 8. Protection

As described in [I-D.draft-kompella-teas-mppte], as there are multiple egress interfaces (SID Lists), the loss of an interface link does not result in traffic drops as long as one egress interface (SID List) remains, although congestion may occur.

Link protection from an upstream Junction node to its downstream Junction nodes can be achieved using existing TI-LFA [I-D.draft-ietf-rtgwg-segment-routing-ti-lfa] mechanisms, applied per egress SID List. Since the top SID(s) in each SID List identify the path to the next downstream Junction node, TI-LFA is applicable.

Local computation for node protection on an upstream Junction node is not feasible, because it lacks visibility into the DAG beyond the immediate downstream Junction node as it only knows the next Junction Segment. A controller MAY be used to precompute backup SR Paths and signal these backup SID Lists to the upstream Junction segments.

In a situation where a downstream Junction node experiences system or connectivity failure, the upstream Junction node will begin to blackhole the traffic until a notification can be sent to upstream nodes to remove the Junction node from the DAG.

## 9. Other considerations

### 9.1. Hierarchy

The use of Junction Segments to achieve a DAG can be used in hierarchical organization and sharing of DAGs between end-to-end tunnels.

For instance, in a multi-area or multi-instance topology, one or more shared DAGs may be created per area, connecting the border ingress node(s) and egress node(s). These DAGs can then be stitched together for use in an end-to-end SR tunnel.

Figure 2 is an example of two independent SR Policy Tunnels from Headend A and Headend B terminating on Egress X. An instance of a DAG with MID 100 can be configured between ABR-1 and the Egress X node. ABR-1 Junction Segment which ingresses the DAG has a Binding Segment attached, for example BSID-100. Therefore, the SID list on Headend A and Headend B SID lists would contain the SR Path (ex: Node-SID-ABR-1) to reach ABR 1 followed by BSID-100. The instruction set from each Headend to ABR 1 could also be another instance of a DAG, either for independent use or shared.

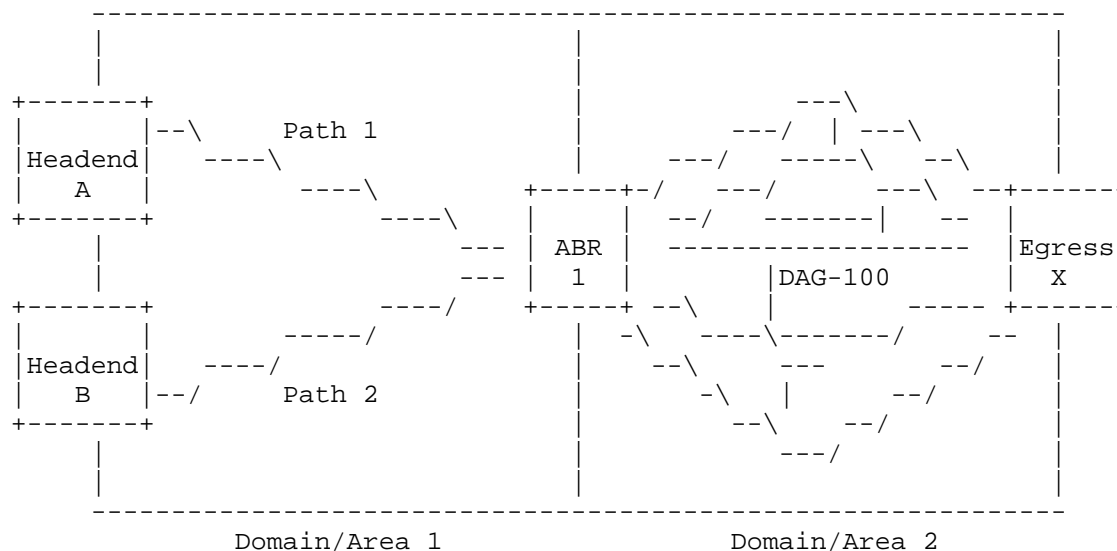


Figure 2 : Reusable DAG 100

## 9.2. Directly connected Junction nodes

As described in the Operational section, all transit nodes in a DAG MAY be signaled with a Junction Segment. Alternatively depending on the topological graph and TE requirements, two Junction segments may be interconnected via an SR Path, with a SID list, where the SR Path itself may correspond to an ECMP or TE Path.

## 9.3. Broadcast links

SR networks abstract broadcast links with the use point to point adjacency segments identifying each neighbor. Specifying a DAG which contains a broadcast link is feasible as an adjacency segment can be used to identify the neighboring Junction node on the broadcast link. The outgoing SID List of the Junction Segment simply contains the adjacency SID of the next-hop neighbor on the broadcast link.

## 9.4. Local optimization

An individual Junction Segment can be optimized by adding new SID lists for downstream neighbors in the DAG or removing downstream nodes. Additionally, the SR path between two Junction nodes whether directly or indirectly connected may also be updated. The controller utilizes existing protocol mechanisms to update the forwarding instructions of the Junction Segment.

Multiple candidate paths can be used to facilitate local optimization. However, care must be taken regarding how the MPTED version is encoded and how version increments are signaled. Specifically, if the MPTED version is encoded in the color attribute or another attribute of the SR Policy, rather than within the Candidate Path itself, multiple Candidate Paths SHOULD NOT be used when an operation requires incrementing the MPTED version to maintain consistency.

The implications of MPTED version encoding and management will be further addressed in future versions of this document.

## 9.5. Global optimization

Reoptimizing the DAG requires potentially redeploying all new Junction Segments network wide in a coordinated, make before break manner.

Since the DAG MID and version are encoded in the color field of an SR Policy, globally optimizing a DAG with make-before-break considerations requires deploying new Junction segments as SR Policies with unique color values to all Junction nodes of the new

DAG. After all of the new Junction segments are deployed, the ingress node can be updated with new SID lists which utilize the new DAG Junction segments of its's neighbors.

TODO This section needs further discussion.

TODO

## 10. Security Considerations

TODO

## 11. Manageability Considerations

This document currently proposes using the existing SR Policy construct with a color value representing the DAG MID and version. Since SR Policy color value is originally intended for ingress traffic steering on matched routers, a deployment MUST allocate a color range which will be used for MPTE and MUST NOT be assigned to any advertised routes in the network.

TODO

## 12. IANA Considerations

None at this time

## 13. References

### 13.1. Normative References

[I-D.draft-ietf-idr-bgp-ls-sr-policy]

Previdi, S., Talaulikar, K., Dong, J., Gredler, H., and J. Tantsura, "Advertisement of Segment Routing Policies using BGP Link-State", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-ls-sr-policy-17, 6 March 2025, <<https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-ls-sr-policy-17>>.

[I-D.draft-ietf-idr-segment-routing-te-policy]

Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., and D. Jain, "Advertising Segment Routing Policies in BGP", Work in Progress, Internet-Draft, draft-ietf-idr-segment-routing-te-policy-26, 23 October 2023, <<https://datatracker.ietf.org/doc/html/draft-ietf-idr-segment-routing-te-policy-26>>.

- [I-D.draft-ietf-rtgwg-segment-routing-ti-lfa]  
Bashandy, A., Litkowski, S., Filsfils, C., Francois, P., Decraene, B., and D. Voyer, "Topology Independent Fast Reroute using Segment Routing", Work in Progress, Internet-Draft, draft-ietf-rtgwg-segment-routing-ti-lfa-21, 12 February 2025, <<https://datatracker.ietf.org/doc/html/draft-ietf-rtgwg-segment-routing-ti-lfa-21>>.
- [I-D.draft-kompella-teas-mppte]  
Kompella, K., Jalil, L., Khaddam, M., and A. Smith, "Multipath Traffic Engineering", Work in Progress, Internet-Draft, draft-kompella-teas-mppte-00, 3 March 2025, <<https://datatracker.ietf.org/doc/html/draft-kompella-teas-mppte-00>>.
- [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/rfc/rfc2119>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/rfc/rfc5440>>.
- [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, <<https://www.rfc-editor.org/rfc/rfc7752>>.
- [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/rfc/rfc8174>>.
- [RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <<https://www.rfc-editor.org/rfc/rfc8231>>.
- [RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <<https://www.rfc-editor.org/rfc/rfc8402>>.

[RFC9256] Filsfils, C., Talaulikar, K., Ed., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", RFC 9256, DOI 10.17487/RFC9256, July 2022, <<https://www.rfc-editor.org/rfc/rfc9256>>.

### 13.2. Informative References

[RFC9524] Voyer, D., Ed., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "Segment Routing Replication for Multipoint Service Delivery", RFC 9524, DOI 10.17487/RFC9524, February 2024, <<https://www.rfc-editor.org/rfc/rfc9524>>.

### Acknowledgments

None at this time

### Authors' Addresses

Andrew Stone  
Nokia  
Email: [andrew.stone@nokia.com](mailto:andrew.stone@nokia.com)

Vishnu Pavan Beeram  
Juniper Networks  
Email: [vbeeram@juniper.net](mailto:vbeeram@juniper.net)