SPRING Working Group   J. Dong
Internet-Draft   Huawei Technologies
Intended status: Standards Track   S. Bryant
Expires: January 13, 2022   Futurewei Technologies
                                                       T. Miyasaka
                                                    KDDI Corporation
                                                             Y. Zhu
                                                      China Telecom
                                                             F. Qin
                                                              Z. Li
                                                       China Mobile
                                                            F. Clad
                                                      Cisco Systems
                                                      July 12, 2021

                    Introducing Resource Awareness to SR Segments
                     draft-ietf-spring-resource-aware-segments-03

Abstract

   This document describes the mechanism to associate network resource
   attributes to Segment Routing Identifiers (SIDs). Such SIDs are
   referred to as resource-aware SIDs in this document. The resource-
   aware SIDs retain their original forwarding semantics, but with the
   additional semantics to identify the set of network resources
   available for the packet processing action. The resource-aware SIDs
   can therefore be used to build SR paths or virtual networks with a
   set of reserved network resources. The proposed mechanism is
   applicable to both segment routing with MPLS data plane (SR-MPLS) and
   segment routing with IPv6 data plane (SRv6).

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 January 13, 2022.

Table of Contents

1. Introduction

   Segment Routing (SR) [RFC8402] specifies a mechanism to steer packets
   through an ordered list of segments. A segment is referred to by its
   Segment Identifier (SID). With SR, explicit source routing can be
   achieved without introducing per-path state into the network.
   Compared with RSVP-TE [RFC3209], currently SR does not have the
   capability of reserving network resources or identifying a set of
   network resources reserved for an individual or a group of services
   or customers. Although a centralized controller can have a global
   view of network state and can provision different services using
   different SR paths, in data packet forwarding it still relies on
   traditional DiffServ QoS mechanism [RFC2474] [RFC2475] to provide
   coarse-grained traffic differentiation in the network. While such
   kind of mechanism may be sufficient for some types of services, some
   customers or services may require to have a set of dedicated network

resources allocated in the network to achieve resource isolation from
other customers/services in the same network. Also note the number
of such customers or services can be larger than the number of
traffic classes available with DiffServ QoS.

Without the need of defining new SID types, this document extends the
SR paradigm by associating SIDs with network resource attributes.
These resource-aware SIDs retain their original functionality, with
the additional semantics of identifying the set of network resources
available for the packet processing action. Typical types of the
network resources include link bandwidth, buffers, and queues that
are associated with class of service, scheduling weights or time
cycles, and it is also possible to associate SR SIDs with other types
of resources (e.g., the processing and storage resources). On a
particular network segment, multiple resource-aware SIDs can be
allocated, each of which represents a subset of network resources
allocated in the network to meet the requirement of an individual or
a group of customers or services. The allocation of network
resources on network segments can be done either via local
configuration or via a centralized controller. Other approaches are
possible such as use of a control protocol signaling, but they are
for further study and out of the scope of this document. Each set of
network resources can be associated with one or multiple resource-
aware SIDs. The resource-aware SIDs can be used to build SR paths
with a set of reserved network resources, which can be used to carry
service traffic which requires dedicated network resources along the
path. The resource-aware SIDs can also be used to build SR based
virtual networks with the required network topology and resource
attributes. The proposed mechanism is applicable to SR with both
MPLS data plane (SR-MPLS) and IPv6 data plane (SRv6).

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
BCP14 RFC 2119 [RFC2119] RFC 8174 [RFC8174] when, and only when, they
appear in all capitals, as shown here.

2. Segments with Resource Awareness

In segment routing architecture [RFC8402], several types of segments
are defined to represent either topological or service instructions.
A topological segment can be a node segment or an adjacency segment.
A service segment may be associated with specific service functions
for service chaining purpose. This document introduces additional
resource semantics to these existing types of SIDs, so that the
resource-aware SIDs can be used to identify not only the topology or

service functions, but also the set of network resources allocated on
the network segments for packet processing.

This section describes the mechanisms of using SR SIDs to identify
the additional resource information associated with the SR paths or
virtual networks based on the two SR data plane instantiations: SR-
MPLS and SRv6. The mechanisms to identify the forwarding path or
network topology with SIDs as defined in [RFC8402] can be reused, and
the control plane can be based on [RFC4915], [RFC5120] and
[I-D.ietf-lsr-flex-algo].

2.1. SR-MPLS

As specified in [RFC8402], an IGP Adjacency Segment (Adj-SID) is an
SR segment attached to a unidirectional adjacency or a set of
unidirectional adjacencies. An IGP Prefix Segment (Prefix-SID) is an
SR segment attached to an IGP prefix, which identifies an instruction
to forward the packet along the path computed using the routing
algorithm in the associated topology. An IGP node segment is an IGP-
Prefix segment that identifies a specific router (e.g., a loopback).
As described in [I-D.ietf-spring-segment-routing-central-epe] and
[I-D.ietf-idr-bgpls-segment-routing-epe], BGP PeerAdj SID is used as
an instruction to steer over a local interface towards a specific
peer node in a peering Autonomous System (AS). These types of SIDs
can be extended to represent both the topological instructions and
the set of network resources allocated for packet processing
following the instruction. The MPLS instantiation of Segment Routing
is specified in [RFC8660].

A resource-aware Adj-SID represents a subset of the resources (e.g.
bandwidth, buffer and queuing resources) of a given link, thus each
resource-aware Adj-SID is associated with its own set of TE
attributes.

For one IGP link, multiple resource-aware Adj-SIDs SHOULD be
allocated, each of which is associated with a subset of the link
resources allocated on the link. For one inter-domain link, multiple
BGP PeerAdj SIDs MAY be allocated, each of which is associated with a
subset of the link resources allocated on the inter-domain link. The
resource-aware Adj-SIDs MAY be associated with a specific network
topology and/or algorithm, so that it is used only for resource-aware
SR paths computed within the topology and/or algorithm.

Note this per-segment resource allocation complies to the SR
paradigm, which avoids introducing per-path state into the network.
Several approaches can be used to partition and reserve the link
resources, such as [FLEXE], Layer-2 logical sub-interfaces, dedicated

queues, etc. The detailed mechanism of link resource partitioning is out of scope of this document.

A resource-aware Prefix-SID is associated with a network topology and/or algorithm in which the attached node participates, and in addition, a resource-aware prefix-SID is associated with a set of network resources (e.g. bandwidth, buffer and queuing resources) allocated on each node and link participating in the same topology and/or algorithm. Such set of network resources can be used for forwarding packets with this resource-aware prefix-SID along the paths computed in the associated topology and/or algorithm.

Although it is possible that each resource-aware prefix-SID is associated with a set of dedicated resources in the network, this implies the overhead with per-prefix resource reservation in both control plane signaling and data plane states, and if network resources are allocated for one prefix on all the possible paths, it is likely some resources will be wasted. A practical approach is that a common set of network resources are allocated by each network node and link participating in a topology and/or algorithm, and are associated with a group of resource-aware prefix-SIDs of the same topology and/or algorithm. Such common set of network resources constitutes a resource group. For a given <topology, algorithm> tuple, there can be one or multiple resource groups, the resource-aware prefix-SIDs which are associated with the same <topology, algorithm> tuple shares the path computation result.

This helps to reduce the dynamics in per-prefix resource allocation and adjustment, so that the network resource can be allocated based on planning and does not have to rely on dynamic signaling. While when the set of nodes and links participate in a <topology, algorithm> tuple changes, the set of network resources allocated on specific nodes and links may need to be adjusted. This means that the resources allocated to resource-aware Adj-SIDs on those links may have to be adjusted and new TE metrics for the associated Adj-SIDs re-advertised.

For one IGP prefix, multiple resource-aware prefix-SIDs SHOULD be allocated. Each resource-aware prefix-SID MAY be associated with a unique <topology, algorithm> tuple, in this case different <topology, algorithm> tuples can be used to distinguish the resource-aware prefix-SIDs of the same prefix. In another case, for one IGP prefix, multiple resource-aware prefix-SIDs MAY be associated with the same <topology, algorithm> tuple, then an additional distinguisher needs to be introduced to distinguish different resource-aware prefix-SIDs associated with the same <topology, algorithm> but different groups of network resources.

A group of resource-aware Adj-SID and resource-aware Prefix-SIDs can
be used to construct the SID lists, which are used to steer the
traffic to be forwarded along the explicit paths (either strict or
loose) and processed using the set of network resources identified by
the resource-aware SIDs.

In data packet forwarding, each resource-aware Adj-SID identifies
both the next-hop and the set of resources used for packet processing
on the outgoing interface. Each resource-aware Prefix-SID identifies
the path to the node which the prefix is attached to, and the common
set of network resources used for packet forwarding on network nodes
along the path. The transit nodes use the resource-aware prefix-SIDs
to determine the next-hop of the packet and the set of associated
local resources, then forward the packet to the next-hop using the
set of local resources.

When the set of network resources allocated on the egress node also
needs to be determined, It is RECOMMENDED that Penultimate Hop
Popping (PHP) [RFC3031] be disabled, or the inner service label needs
to be used to infer the set of resources to be used for packet
processing on the egress node of the SR path.

This mechanism requires to allocate additional prefix-SIDs or adj-
SIDs for network segments to identify different set of network
resources. As the number of resource groups increases, the number of
SIDs would increase accordingly, while it should be noted that there
is no per-path state introduced into the network.

2.2. SRv6

As specified in [RFC8986], an SRv6 Segment Identifier (SID) is a
128-bit value which consists of a locator (LOC) and a function
(FUNCT), optionally it may also contain additional arguments (ARG)
immediately after the FUNCT. The Locator part of the SID is routable
and leads to the node which instantiates that SID, which means the
Locator can be parsed by all nodes in the network. The FUNCT part of
the SID is an opaque identification of a local function bound to the
SID, and the ARG bits of the SID can be used to encode additional
information for the processing of the behavoir bound to the SID.
Thus the FUNCT and ARG parts can only be parsed by the node which
instantiates the SRv6 SID.

For one SRv6 node, multiple resource-aware SRv6 LOCs SHOULD be
allocated. A resource-aware LOC is associated with a network
topology and/or algorithm in which the node participates, and in
addition, a resource-aware LOC is associated with a set of local
resources (e.g. bandwidth, buffer and queueing resources) on each
node participating in the same topology and/or algorithm. Such set

of network resources are used to forward the packets with SIDs which
has the resource-aware LOC as its prefix, along the path computed
with the associated topology and/or algorithm. Similar to the
resource-aware prefix-SIDs in SR-MPLS, a practical approach is that a
common set of network resources are allocated by each network node
and link participating in a topology and/or algorithm, and are
associated with a group of resource-aware LOCs of the same topology
and/or algorithm.

For one IGP link, multiple resource-aware SRv6 End.X SIDs SHOULD be
allocated to identify different set of link resources. Each
resource-aware End.X SID SHOULD use a resource-aware LOC as its
prefix. SRv6 SIDs for other types of functions MAY also be assigned
as resource-aware SIDs, which can identify the set of network
resources allocated by the node for executing the behavoir.

A group of resource-aware SRv6 SIDs can be used to construct the SID
lists, which are used to steer the traffic to be forwarded along the
explicit paths (either strict or loose) and processed using the set
of network resources identified by the resource-aware SIDs and
Locators.

In data packet forwarding, each resource-aware End.X SID identifies
both the next-hop and the set of resources used for packet processing
on the outgoing interface. Each resource-aware Locator identifies
the path to the node which the Locator is assigned to, and the set of
network resources used for packet forwarding on network nodes along
the path. The transit nodes use the resource-aware Locators to
determine the next-hop of the packet and the set of associated local
resources, then forward the packet to the next-hop using the set of
local resources.

This mechanism requires to allocate additional SRv6 Locators and SIDs
for network segments to identify different set of network resources.
As the number of resource groups increases, the number of SRv6
Locators and SIDs would increase accordingly, while it should be
noted that there is no per-path state introduced into the network.

3. Control Plane Considerations

The mechanism described in this document makes use of a centralized
controller to collect the information about the network
(configuration, state, routing databases, etc.) as well as the
service information (traffic matrix, performance statistics, etc.)
for the planning of network resources based on the service
requirement. Then the centralized controller instructs the network
nodes to allocate the network resources and associate the resources
with the resource-aware SIDs. The resource-aware SIDs can be either

explicitly provisioned by the controller, or dynamically allocated by
network nodes then reported to the controller. The controller is
also responsible for the centralized computation and optimization of
the SR paths taking the topology, algorithm and network resource
constraints into consideration. The interaction between the
controller and the network nodes can be based on Netconf/YANG
[RFC6241] [RFC7950], BGP-LS [RFC7752], BGP SR Policy
[I-D.ietf-idr-segment-routing-te-policy] or PCEP [RFC5440]. In some
scenarios, extensions to some of these protocols is needed, which are
out of the scope of this document and will be specified in separate
documents. In some cases, a centralized controller may not be used,
but this would complicate the operations and planning therefore not
suggested.

The distributed control plane is complementary to the centralized
controller. A distributed control plane can be used for the
collection and distribution of the network topology and resource
information associated with the resource-aware SIDs among network
nodes, then some of the nodes can distribute the collected
information to the centralized controller. Distributed route
computation for services with topology and/or resource constraints
may also be needed on network nodes. The distributed control plane
may be based on [RFC4915], [RFC5120], [I-D.ietf-lsr-flex-algo] or the
combination of some of them with necessary extensions.

On network nodes, the support for a resource group and the
information to associate packets with that resource group needs to be
advertised in the control plane, so that all nodes have a consistent
view of the resource group. Given that resource management is a
central function, the knowledge of the exact resources provided to a
resource group needs to be known accurately by the relevant central
control components (e.g. PCE) and the network nodes. This may be
done by configuration, alternative protocols, or by advertisements in
the IGP for collection by BGP-LS. If there are related link
advertisements, then consistency must be assured across that set of
advertisements. To advertise its support for a given resource group,
a node needs to advertise the identifier of the resource group, the
associated topology and algorithm, the resource-aware SIDs and
potentially a set of TE metrics representing the common resources
allocated to it. The details will be described in a separate
document.

4. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an
RFC.

5. Security Considerations

The security considerations of segment routing are applicable to this document.

The resource-aware SIDs may be used for provisioning of SR paths or virtual networks to carry traffic with latency as one of the SLA parameters. By disrupting the latency of such traffic an attack can be directly targeted at the customer application, or can be targeted at the network operator by causing them to violate their SLA, triggering commercial consequences. Dynamic attacks of this sort are not something that networks have traditionally guarded against, and networking techniques need to be developed to defend against this type of attack. By rigorously policing ingress traffic and carefully provisioning the resources provided to such services, this type of attack can be prevented. However care needs to be taken when providing shared resources, and when the network needs to be reconfigured as part of ongoing maintenance or in response to a failure.

The details of the underlay network MUST NOT be exposed to third parties, to prevent attacks aimed at exploiting shared network resources.

6. Contributors

Zhenbin Li
Email: lizhenbin@huawei.com

Zhibo Hu
Email: huzhibo@huawei.com

Joel Halpern
Email: jmh@joelhalpern.com

7. Acknowledgements

The authors would like to thank Mach Chen, Stefano Previdi, Charlie Perkins, Bruno Decraene, Loa Andersson, Alexander Vainshtein and John Drake for the valuable discussion and suggestions to this document.

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,
              <https://www.rfc-editor.org/info/rfc2119>.

   [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>.

   [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/info/rfc8402>.

   [RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S.,
              Decraene, B., Litkowski, S., and R. Shakir, "Segment
              Routing with the MPLS Data Plane", RFC 8660,
              DOI 10.17487/RFC8660, December 2019,
              <https://www.rfc-editor.org/info/rfc8660>.

   [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>.

8.2. Informative References

   [FLEXE]   "Flex Ethernet Implementation Agreement", March 2016,
              <http://www.oiforum.com/wp-content/uploads/OIF-FLEXE-
              01.0.pdf>.

   [I-D.ietf-idr-bgpls-segment-routing-epe]
              Previdi, S., Talaulikar, K., Filsfils, C., Patel, K., Ray,
              S., and J. Dong, "BGP-LS extensions for Segment Routing
              BGP Egress Peer Engineering", draft-ietf-idr-bgpls-
              segment-routing-epe-19 (work in progress), May 2019.

   [I-D.ietf-idr-segment-routing-te-policy]
              Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P.,
              Rosen, E., Jain, D., and S. Lin, "Advertising Segment
              Routing Policies in BGP", draft-ietf-idr-segment-routing-
              te-policy-11 (work in progress), November 2020.

   [I-D.ietf-lsr-flex-algo]
              Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and
              A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-
              algo-15 (work in progress), April 2021.

   [I-D.ietf-spring-segment-routing-central-epe]
             Filsfils, C., Previdi, S., Dawra, G., Aries, E., and D.
             Afanasiev, "Segment Routing Centralized BGP Egress Peer
             Engineering", draft-ietf-spring-segment-routing-central-
             epe-10 (work in progress), December 2017.

   [I-D.ietf-spring-segment-routing-policy]
             Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and
             P. Mattes, "Segment Routing Policy Architecture", draft-
             ietf-spring-segment-routing-policy-11 (work in progress),
             April 2021.

   [RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black,
             "Definition of the Differentiated Services Field (DS
             Field) in the IPv4 and IPv6 Headers", RFC 2474,
             DOI 10.17487/RFC2474, December 1998,
             <https://www.rfc-editor.org/info/rfc2474>.

   [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z.,
             and W. Weiss, "An Architecture for Differentiated
             Services", RFC 2475, DOI 10.17487/RFC2475, December 1998,
             <https://www.rfc-editor.org/info/rfc2475>.

   [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol
             Label Switching Architecture", RFC 3031,
             DOI 10.17487/RFC3031, January 2001,
             <https://www.rfc-editor.org/info/rfc3031>.

   [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V.,
             and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP
             Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001,
             <https://www.rfc-editor.org/info/rfc3209>.

   [RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P.
             Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF",
             RFC 4915, DOI 10.17487/RFC4915, June 2007,
             <https://www.rfc-editor.org/info/rfc4915>.

   [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi
             Topology (MT) Routing in Intermediate System to
             Intermediate Systems (IS-ISs)", RFC 5120,
             DOI 10.17487/RFC5120, February 2008,
             <https://www.rfc-editor.org/info/rfc5120>.

   [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/info/rfc5440>.

   [RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed.,
              and A. Bierman, Ed., "Network Configuration Protocol
              (NETCONF)", RFC 6241, DOI 10.17487/RFC6241, June 2011,
              <https://www.rfc-editor.org/info/rfc6241>.

   [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/info/rfc7752>.

   [RFC7950] Bjorklund, M., Ed., "The YANG 1.1 Data Modeling Language",
              RFC 7950, DOI 10.17487/RFC7950, August 2016,
              <https://www.rfc-editor.org/info/rfc7950>.

Authors' Addresses

   Jie Dong
   Huawei Technologies

   Email: jie.dong@huawei.com

   Stewart Bryant
   Futurewei Technologies

   Email: stewart.bryant@gmail.com

   Takuya Miyasaka
   KDDI Corporation

   Email: ta-miyasaka@kddi.com

   Yongqing Zhu
   China Telecom

   Email: zhuyq8@chinatelecom.cn

   Fengwei Qin
   China Mobile

   Email: qinfengwei@chinamobile.com

Zhenqiang Li
China Mobile

Email: li_zhenqiang@hotmail.com

Francois Clad
Cisco Systems

Email: fclad@cisco.com