Network Working Group   L. Dunbar
Internet Draft   Futurewei
Intended status: Standard   K. Majumdar
Expires: June 1, 2023   Microsoft
                                                     H. Wang
                                                      Huawei
                                                   G. Mishra
                                                     Verizon
                                           December 1, 2022

BGP Extension for 5G Edge Service Metadata
draft-ietf-idr-5g-edge-service-metadata-00

Abstract

   This draft describes three new sub-TLVs for egress routers to
   advertise the Edge Service Metadata of the directly attached
   edge services (ES). The Edge Service Metadata can be used by
   the ingress routers in the 5G Local Data Network to make path
   selections not only based on the routing cost but also the
   running environment of the edge services. The goal is to
   improve latency and performance for 5G edge services.

   The extension enables an edge service at one specific location
   to be more preferred than the others with the same IP address
   (ANYCAST) to receive data flow from a specific source, like
   specific User Equipment (UE).

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

[5g-edge-Compute] describes the 5G Edge Computing background
and how BGP can be used to advertise the running status and
environment of the directly attached 5G edge services. Besides
the Radio Access, 5G is characterized by having edge services
closer to the Cell Towers reachable by Local Data Networks
(LDN) [3GPP TS 23.501]. From IP network perspective, the 5G
LDN is a limited domain with edge services a few hops away
from the ingress nodes. Only selective services by UEs are
considered as 5G Edge Services.

This document describes a new Metadata Path Attribute and
three new sub-TLVs for egress routers to advertise the Edge
Service Metadata of the directly attached edge services. The
Edge Service Metadata in this document refers to edge
services' site capacity, the site preference, and the load
index, which are further explained in Section 3. Note: the
proposed Edge Service Metadata are not intended for the
services reachable via the networks outside the 5G LDN. The
Edge Service Metadata can be used by the ingress routers in
the 5G Local Data Network to make path selection not only
based on the routing distance but also the running environment
of the edge cloud sites. The goal is to improve latency and
performance for 5G edge services.

The extension is targeted for a single domain with RR
controlling the propagation of the BGP UPDATE. The Edge
Service Metadata is only attached to the services (routes)
hosted in the 5G edge cloud sites, which are only a small

   subset of services initiated from UEs. E.g., not for UEs
   accessing many internet sites.

[2](). **Conventions used in this document**

   Application Server: An application server is a physical or
                virtual server that hosts the software system for
                the application.

   Application Server Location: Represent a cluster of servers at
                one location serving the same Application. One
                application may have a Layer 7 Load balancer,
                whose address(es) are reachable from an external
                IP network, in front of a set of application
                servers. From an IP network perspective, this
                whole group of servers is considered as the
                Application server at the location.

   Edge Application Server: used interchangeably with Application
                Server throughout this document.

   Edge Hosting Environment: An environment providing the support
                required for Edge Application Server's execution.

                NOTE: The above terminologies are the same as
                those used in 3GPP TR 23.758

   Edge DC: Edge Data Center, which provides the Hosting
                Environment for the edge services. An Edge DC
                might host 5G core functions in addition to the
                frequently used application servers.

   gNB   next generation Node B

   PSA: PDU Session Anchor (UPF)

   SSC: Session and Service Continuity

   UE: User Equipment

   UPF: User Plane Function

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.

3. BGP Protocol Extension for Edge Service Metadata

The goal of the BGP extension is for egress routers to
propagate the metrics about their running environment to
ingress routers, which are call the Edge Service Metadata
throughout the document. Here are some examples of the
metrics propagated by the egress routers:
- The site Capacity Index,
- The Site Preference Index,
- The Load Measurement Index for the attached edge services.

This section specifies how those Metadata impact the ingress
nodes' path selections.

3.1. Ingress Node BGP Path Selection Behavior

3.1.1. Edge Service Metadata Influenced BGP Path Selection

When an ingress router receives BGP updates for the same IP
address from multiple egress routers, all those egress routers
are considered as the next hops for the IP address. For the
selected edge services, the ingress router's BGP engine would
call an Edge Service Management function that can select paths
based on the Edge Service Metadata received. [5G-EC-Metrics]
has an example algorithm to compute the weighted path cost
based on the Edge Service Metadata carried by the sub-TLVs
specified in this document.

Section 5 has the detailed description of the Edge Service
Metadata influenced optimal path selection.

3.1.2. Ingress Router Forwarding Behavior

When the ingress router receives a packet and lookup the route
in the FIB, it gets the destination prefix's whole path. It
encapsulates the packet destined towards the optimal egress
node.

For subsequent packets belonging to the same flow, the ingress
router needs to forward them to the same egress router unless
the selected egress router is no longer reachable. Keeping
packets from one flow to the same egress router, a.k.a. Flow
Affinity, is supported by many commercial routers. Most
registered EC services have relatively short flows.

How Flow Affinity is implemented is out of the scope for this
document. Appendix A has one example illustrating achieving
flow affinity.

3.1.3. Forwarding Behavior when UEs moving to new 5G Sites

When a UE moves to a new 5G gNB which is anchored to the same
UPF, the packets from the UE traverse to the same ingress
router. Path selection and forwarding behavior are same as
before.

If the UE maintains the same IP address when anchored to a new
UPF, the directly connected ingress router might use the
information passed from a neighboring router to derive the
optimal Next Hop for this route. [5G-Edge-Sticky] describes
some methods for the ingress router connected to the UPF in
the new site to consider the information passed from other
ingress routers in selecting the optimal paths. The detailed
algorithm is out of the scope of this document.

4. Edge Service Metadata Encoding

4.1. Metadata Path Attribute

The Metadata Path Attribute is an optional transitive BGP Path
attribute to carry the Edge Service Metadata described in this
document. Will need IANA to assign a value as the Type code
of the Path Attribute. The Metadata Path Attribute,
illustrated below, consists of a set of sub-TLVs, with each
sub-TLV containing the information corresponding to a specific
metrics of the Edge Service Metadata.
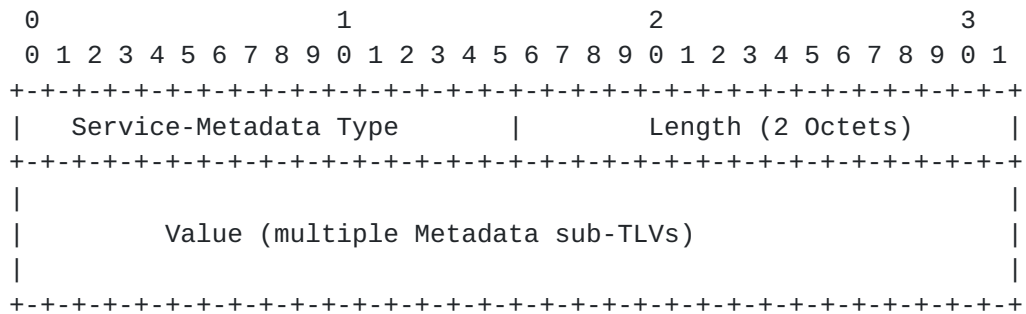
```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Service-Metadata Type      |        Length (2 Octets)      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|          Value (multiple Metadata sub-TLVs)                   |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
           Figure 1: Edge Service Metadata Path Attribute

   Service-Metadata Path Attribute Type (2 octets): identify the
   Metadata Path Attribute, to be assigned by IANA.

      o Length (2 octets): the total number of octets of the
   value field.

      o Value (variable): comprised of multiple sub-TLVs.

   There are three types of Edge Service Metadata sub-TLVs
   specified by this document for the Capacity Index Value, the
   Site Preference Index Value, and the Load Measurement.

   All values in the Sub-TLVs are unsigned 32 bits integers.

4.2. The Site Preference Index sub-TLV format

   The Site Preference Index is one of the factors integrated
   into the total cost for path selection. One Edge Cloud site
   can have fewer computing servers, less power, or lower
   internal network bandwidth than another. E.g., one micro edge
   computing center located at a remote cell site has less
   preference index value than an edge site in a metro area that
   hosts management systems, analytics functions, and security
   functions.

   The Preference Index sub-TLV has the following format:

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Site-Preference Sub-Type   |              Length           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Preference Index value                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
             Figure 2: Preference Index Sub-TLV

   Preference Index value: 1-100, with 1 being the least
   preferred, and 100 being the most preferred.

  4.3. Capacity Index Metadata

   Capacity Index indicates the capacity value for a site or a
   pod where the edge services are hosted. One Edge Site can be
   in full capacity, reduced capacity, or completely out of
   service.

   Cloud Site/Pod failures and degradation include, but not
   limited to, a site capacity degradation or entire site going
   down caused by a variety of reasons, such as fiber cut
   connecting to the site or among pods within one site, cooling
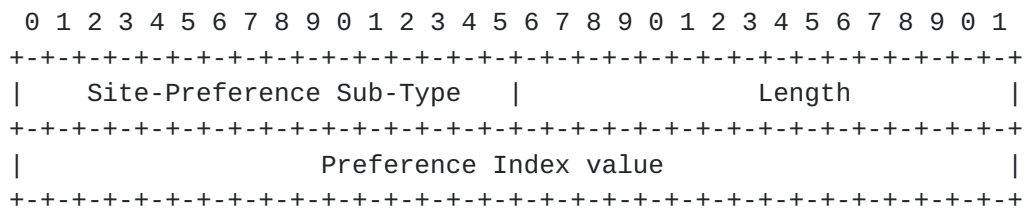   failures, insufficient backup power, cyber threats attacks,
   too many changes outside of the maintenance window, etc.
   Fiber-cut is not uncommon within a Cloud site or between
   sites.

   When those failure events happen, the Edge (egress) router
   visible to the ingress routers can be running fine. Therefore,
   the ingress routers can't use BFD to detect the failures.

   When there is a failure occurring at an edge site (or pod),
   many instances can be impacted. In addition, the routes (i.e.,
   the IP addresses) in an Edge Cloud Site might not be
   aggregated nicely. Instead of many BGP UPDATE messages for
   each instance to the impacted ingress routers, the egress
   router can send one single BGP UPDATE indicating the capacity
   of the site. The ingress routers can switch all or a portion
   of the instances that are associated with the site depending
   on how much the site is degraded.

   The Capacity Index sub-TLV:
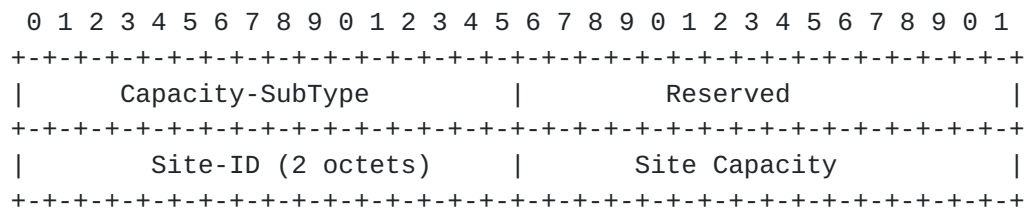
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+
   | Capacity-SubType | Reserved |
   +-+
   | Site-ID (2 octets) | Site Capacity |
   +-+
                Figure 3: Capacity Index Sub-TLV

  - Capacity subtype: (TBD by IANA)

- Site ID: identifier for a group of routes whose capacity is
  indicated by the capacity value carried in the UPDATE. There
  could be more than one sites (or Pods) connected to the egress
  router (a.k.a. Edge DC GW)

- Site Capacity: represent the percentage of the site
  availability, e.g., 100%, 50%, or 0%. When a site goes dark,
  the Index is set to 0. 50 means 50% capacity functioning.

4.3.1. Capacity Site Index attached to services

The purpose of the Capacity Site index is to advertise the
service instance's site reference identifier and the capacity
value of the site.

However, it is not necessary to include the Capacity Site
Index for every BGP Update message if there is no change to
the site-reference identifier or the Capacity value for the
service instances.

The ingress routers associate the Site reference Identifier to
the routes in the Routing table.

4.3.2. BGP UPDATE with standalone Capacity Site Index

When there are failures or degradation to a site, the
corresponding egress router can send a BGP UPDATE with the
Capacity Site Index without attaching any routes.

When an ingress router receives a BGP Update message from
Router-X with the Site-Capacity Sub-TLV without routes
attached, the new Site-Capacity value is applied to all routes
that have the Router-X as their next hops and are associated
with the Site-ID in the Sub-TLV.

4.4. Load Measurement sub-TLV format

   Two types of Load Measurement Sub-TLVs are specified. One is
   to carry the aggregated cost Index based on a weighted
   combination of the collected measurements; another one is to
   carry the raw measurements of packets/bytes to/from the Edge
   Service address. The raw measurement is useful when ingress
   routers have embedded analytics relying on the raw
   measurements.

```
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |     subType=Aggregated-Cost    |               Length        |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                    Measurement Period                         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |         Aggregated Load Index to reach the Edge Service      |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              Figure 4: Aggregated Load Index Sub-TLV
```
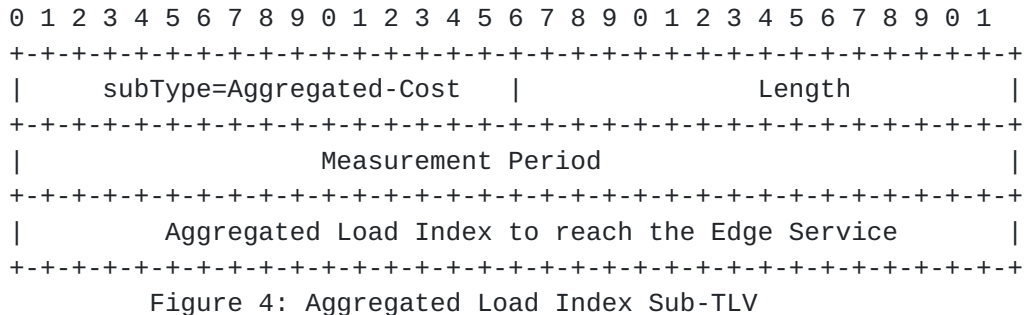
   Aggregated-Cost Sub-Type(TBD1): Aggregated Load Measurement
   Index to reach the Edge Service, which is configured or
   calculated by the egress nodes.

   Raw Load Measurement sub-TLV has the following format:

```
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |     subType= Raw-Measurements |              Length          |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                    Measurement Period                         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |          total number of packets to the Edge Service         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |          total number of packets from the Edge Service       |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |          total number of bytes to the Edge Service           |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |          total number of bytes from the Edge Service         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
              Figure 5: Raw Load Measurement Sub-TLV
```
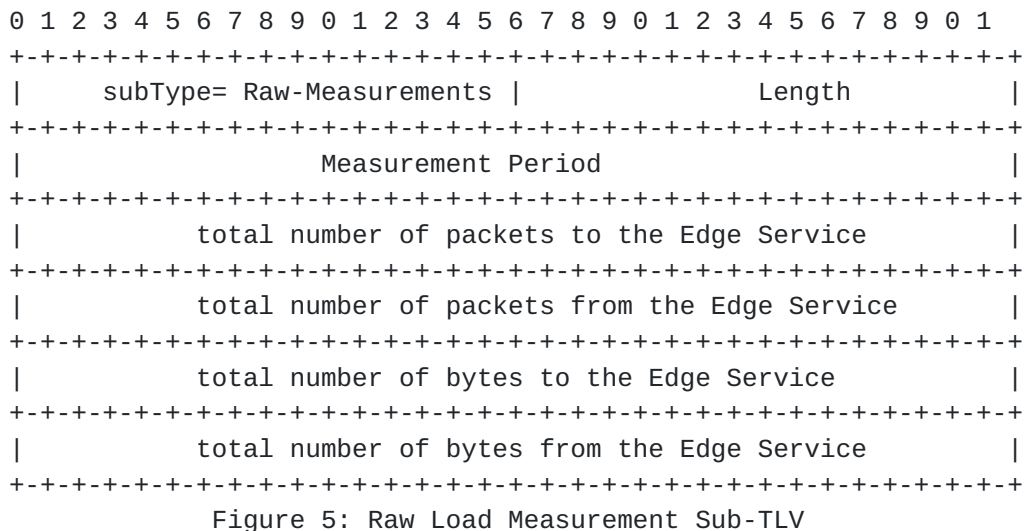
   Raw-Measurement Sub-Type (TBD2): Raw measurements of
   packets/bytes to/from the Edge Service address.

The receiver nodes can calculate the cost to reach the Edge
Service by a weighted combination of raw measurements sent
from the Edge Service, e.g.

Index=w1*ToPackets+w2*FromPackes+w3*ToBytes+w4*FromBytes

Where wi, which are configured by operators, is a value
between 0 and 1; w1+ w2+ w3+ w4 = 1.

Measure Period: BGP Update period in Seconds or user-
specified period.

5. Service Metadata Influenced Decision Process

When an ingress router receives BGP updates for the same IP
address from multiple egress routers, all those egress routers
are considered as the next hops for the IP address. For the
selected services configured to be influenced by the Edge
Service Metadata, the ingress router's BGP Decision process
would trigger the Edge Service Management function to compute
the weight to be applied to the route's next hop in the
forwarding plane. The decision process is influenced by the
Edge Service Metadata associated with the client routes, such
as Site-Capacity-Index, Site Preference, and Load Index, in
addition to the traditional BGP multipath computation
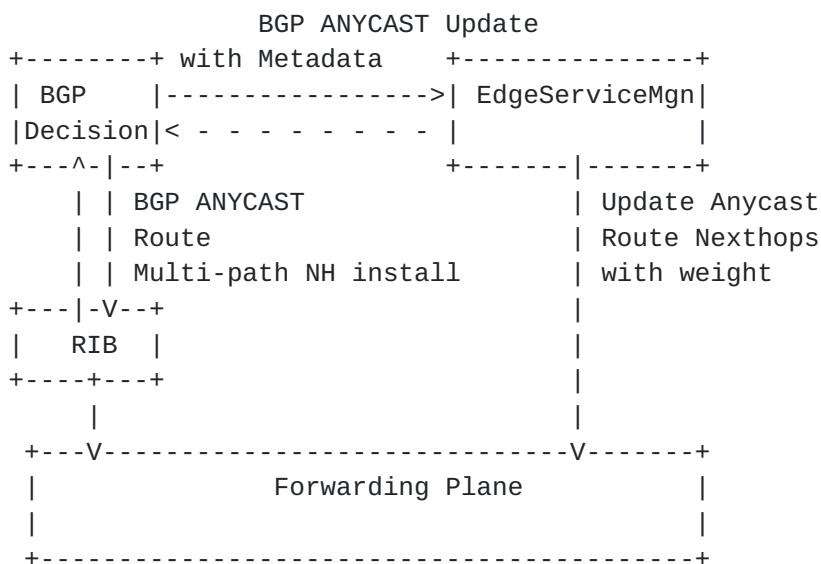algorithm, such as the Weight, Local preference, Origin, MED,
etc., shown below:

```
                      BGP ANYCAST Update
      +--------+ with Metadata     +---------------+
      | BGP    |---------------->| EdgeServiceMgn|
      |Decision|< - - - - - - - - |               |
      +---^-|--+                 +-------|-------+
          | | BGP ANYCAST                | Update Anycast
          | | Route                      | Route Nexthops
          | | Multi-path NH install      | with weight
      +---|-V--+                          |
      |   RIB  |                          |
      +----+---+                          |
           |                             |
       +---V-------------------------------V-------+
       |              Forwarding Plane             |
       |                                           |
       +-------------------------------------------+
            Figure 6: Metadata Influenced Decision
```

When any of those metadata value goes to 0, the effect is the
same as the routes becoming ineligible via the egress router
who originates the metadata UPDATE. But when any of those
metadata just degrade, there is possibility, even though
smaller, for the egress router to continue as the optimal next
hop.

Suppose a destination address for aa08::4450 can be reached by
three next hops (R1, R2, R3). Further, suppose the local BGP's
Decision Process based on the traditional network layer
policies & metrics identifies the R1 as the optimal next hop
for this destination (aa08::4450). The Edge Service Metadata
might result in R2 as the optimal next hop for the prefix and
influence the Forwarding Plane.

The Edge Service Metadata influencing next hop selection is
different from the metric (or weight) to the next hop. The
metric to a next hop can impact many (sometimes, tens of
thousands) routes that have the node as their next hop. while
as the Edge Service Metadata only impact the optimal next hop
selection for a subset of client routes that are identified as
the edge services.

When the BGP custom decision [idr-custom-decision] is used,
the Edge Service Management function would have algorithm to
combine the Edge Service Metadata attributes with the custom
decision to derive the optimal next hop for the Edge service
routes.

Note: For a BGP UPDATE message that only includes the Edge
Service Path Attribute without any NLRI, the Site Capacity
Index value is applied to all the NLRIs with the Site-ID
indicated in the Edge Service Metadata Path Attribute.

6. Edge Service Metadata Propagation Scope

Edge Service Metadata is only to be distributed to the
relevant ingress nodes of the 5G EC local data networks. Only
the ingress routers that are configured with the 5G EC
services need to receive the Edge Service Metadata for
specific Service IDs.

For each registered Edge Service, a corresponding filter group
can be formed on RR to represent the interested ingress
routers that are interested in receiving the corresponding
Edge Service Metadata information.

7. Minimum Interval for Metrics Change Advertisement

As the metrics change can impact the path selection, the Minimum Interval for Metrics Change Advertisement is configured to control the update frequency to avoid route oscillations. Default is 30s.

Significant load changes at EC data centers can be triggered by short-term gatherings of UEs, like conventions, lasting a few hours or days, which are too short to justify adjusting EC server capacities among DCs. Therefore, the load metrics change rate can be in the magnitude of hours or days.

8. Manageability Considerations

The Edge Service Metadata described in this document are only intended for propagating between Ingress and egress routers of one single BGP domain, i.e., the 5G Local Data Networks, which is a limited domain with edge services a few hops away from the ingress nodes. Only the selective services by UEs are considered as 5G Edge Services. The 5G LDN is usually managed by one operator, even though the routers can be by different vendors.

9. Security Considerations

The proposed Edge Service Metadata are advertised within the trusted domain of 5G LDN's ingress and egress routers. There are no extra security threats compared with iBGP.

10. IANA Considerations

Need IANA to assign the Metadata Path Attribute Type.

Metadata Path Attribute Type = TBD1.

Need IANA to assign three new Sub-TLV types under the Metadata Path Attribute:

Type = TBD2: Site preference value sub-TLV

Type = TBD3: Site Capacity Index sub-TLV

    Type = TBD4: Aggregated Load Measurement Index derived from
    the Weighted combination of bytes/packets sent to/received
    from the Edge Service address.

    Type = TBD5: Raw measurements of packets/bytes to/from the
    Edge Service address.

[11](#). References

  11.1. Normative References

  [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
            Requirement Levels", [BCP 14](#), [RFC 2119](#), March 1997.

  [RFC4364] E. rosen, Y. Rekhter, "BGP/MPLS IP Virtual Private
            networks (VPNs)", Feb 2006.

  [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>.

  [RFC7911] D. Walton, et al, "Advertisement of Multiple Paths
            in BGP", [RFC7911](#), July 2016.

  11.2. Informative References

  [3GPP TS 23.501] 3rd Generation Partnership Project;
            Technical Specification Group Services and System
            Aspects; System architecture for the 5G System (5GS)

  [3GPP-EdgeComputing] 3GPP TR 23.748, "3rd Generation
            Partnership Project; Technical Specification Group
            Services and System Aspects; Study on enhancement of
            support for Edge Computing in 5G Core network
            (5GC)", Release 17 work in progress, Aug 2020.

   [5G-EC-Metrics] L. Dunbar, H. Song, J. Kaippallimalil, "IP
              Layer Metrics for 5G Edge Computing Service", draft-
              dunbar-ippm-5g-edge-compute-ip-layer-metrics-00,
              work-in-progress, Oct 2020.

   [5g-edge-Compute] L. Dunbar, K. Majumdar, H. Wang, and G.
              Mishra, "BGP Usage for 5G Edge Computing service
              Metadata", draft-dunbar-idr-5g-edge-compute-bgp-
              usage-00, work-in-progress, July 2022.

   [5G-Edge-Sticky] L. Dunbar, J. Kaippallimalil, "IPv6 Solution
              for 5G Edge Computing Sticky Service", draft-dunbar-
              6man-5g-ec-sticky-service-00, work-in-progress, Oct
              2020.

   [IDR-CUSTOM-DECISION] A. Retana, R. White, "BGP Custom
              Decision Process", draft-ietf-idr-custom-decision-
              08, Feb 2017.

   [SDWAN-EDGE-Discovery] L. Dunbar, S. Hares, R. Raszuk, K.
              Majumdar, "BGP UPDATE for SDWAN Edge Discovery",
              draft-ietf-idr-sdwan-edge-discovery-03, July 2022.

12. Appendix A
12.1. Example of Flow Affinity

   Here is one example to illustrate how Flow Affinity can be
   achieved. This illustration is an informational example.

   For the registered EC services, the ingress node keeps a table
   of

   - Service ID (i.e., IP address)
   - Flow-ID
   - Sticky Egress ID (egress router loopback address)
   - A timer

   The Flow-ID in this table is to identify a flow, initialized
   to NULL. How Flow-ID is constructed is out of the scope for
   this document. Here is one example of constructing the Flow-
   ID:

   - For IPv6, the Flow-ID can be the Flow-ID extracted from the
   IPv6 packet header with or without the source address.

   - For IPv4, the Flow-ID can be the combination of the Source
   Address with or without the TCP/UDP Port number.

   The Sticky Egress ID is the egress node address for the same
   flow. [5G-Edge-Sticky] describes several methods to derive the
   Sticky Egress ID.

   The Timer is always refreshed when a packet with the matching
   EC Service ID (IP address) is received by the node.

   If there is no Stick Egress ID present in the table for the EC
   Service ID, the forwarding plane can select a NextHop
   influenced by the Cost Compute Engine. The forwarding plane
   encapsulates the packet with a path to the chosen NextHop. The
   chosen NextHop and the Flow ID are recorded in the EC Service
   table entry.

   When the selected optimal NextHop (egress router) is no longer
   reachable, ingress router needs to select another path.

13. Acknowledgments

   Acknowledgements to Adrian Farrel, Alvaro Retana, Robert
   Raszuk, Sue Hares, Donald Eastlake, Dhruv Dhody, and Cheng Li
   for their review and contributions.

   This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

   Linda Dunbar
   Futurewei
   Email: ldunbar@futurewei.com

   Kausik Majumdar
   Microsoft
   Email: kmajumdar@microsoft.com

   Haibo Wang
   Huawei
   Email: rainsword.wang@huawei.com

   Gyan Mishra
   Verizon
   Email: gyan.s.mishra@verizon.com

Contributors' Addresses
   Cheng Li
   Huawei
   Email: c.l@huawei.com