Network Working Group
Internet Draft
Intended status: Standard
Expires: February 24, 2023

L. Dunbar
Futurewei
K. Majumdar
Microsoft
H. Wang
Huawei
G. Mishra
Verizon

August 24, 2022

BGP Extension for 5G Edge Service Metadata
draft-dunbar-idr-5g-edge-compute-app-meta-data-14

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 selection 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 flows from a specific source, like specific User Equipment (UE).

This Internet-Draft will expire on April 7, 2021.

Table of Contents

1. Introduction..... 3
2. Conventions used in this document..... 4
3. BGP Protocol Extension for Edge Service Metadata..... 5
   3.1. Ingress Node BGP Path Selection Behavior..... 5
      3.1.1. Edge Service Metadata Influenced BGP Path Selection..... 5
      3.1.2. Ingress Router Forwarding Behavior..... 6
      3.1.3. Forwarding Behavior when UEs moving to new 5G Sites..... 6
4. Edge Service Metadata Encoding..... 6
   4.1. Metadata Path Attribute..... 6
   4.2. The Site Preference Index sub-TLV format..... 7

4.3. Capacity Index Metadata..... 8
4.3.1. Capacity Site Index attached to services..... 9
4.3.2. BGP UPDATE with standalone Capacity Site Index.. 9
4.4. Load Measurement sub-TLV format..... 10
5. Consideration for Optimal Paths Selection..... 11
6. Edge Service Metadata Propagation Scope..... 11
7. Minimum Interval for Metrics Change Advertisement..... 12
8. Manageability Considerations..... 12
9. Security Considerations..... 12
10. IANA Considerations..... 12
11. References..... 13
11.1. Normative References..... 13
11.2. Informative References..... 13
12. Appendix A..... 14
12.1. Example of Flow Affinity..... 14
13. Acknowledgments..... 15

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 a Plugin 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. The Plugin function is called Cost Compute Engine throughout this document.

Suppose a destination address for a service (aa08::4450) can be reached by three next hops (R1, R2, R3). Further, suppose the local BGP's Compute Engine Identifies the R1 as the optimal next hop for flows to be sent to this destination (aa08::4450). The Cost Compute Engine can insert a higher weight for the path towards R1 for the prefix. Suppose BGP Add Path is supported [RFC7911], all three paths can be added to the FIB who can choose the optimal paths for the received data packets.

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.

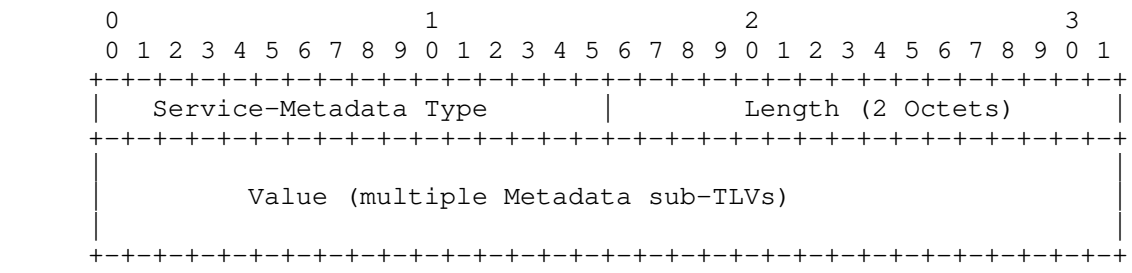


Figure 1: Edge Service Metadata Path Attribute

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

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

- 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:

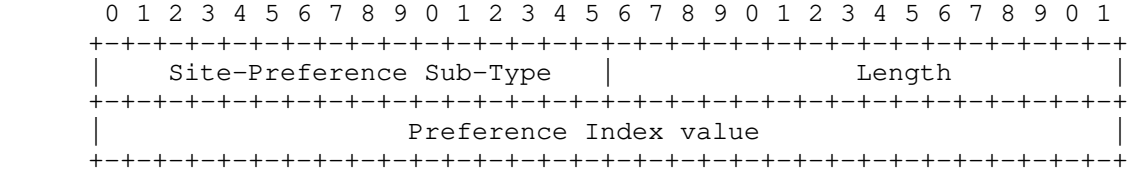


Figure 2: Preference Index Sub-TLV

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

4.3. Capacity Index Metadata

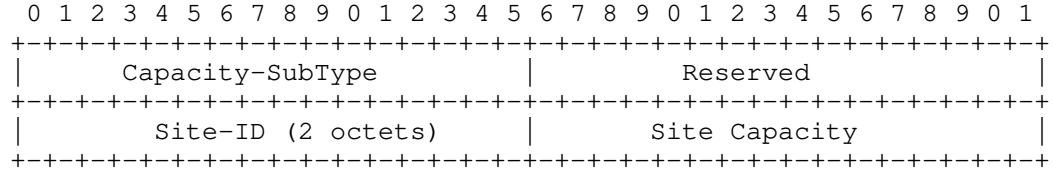
Capacity Index indicates the capacity value for a site or a pod where the edge services are hosted. One Edge Site can be in full capacity, reduced capacity, or completely out of service.

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:



- 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.

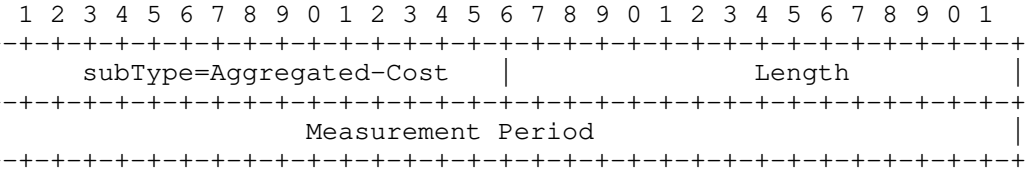


Figure 2: Aggregated Load Index Sub-TLV

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:

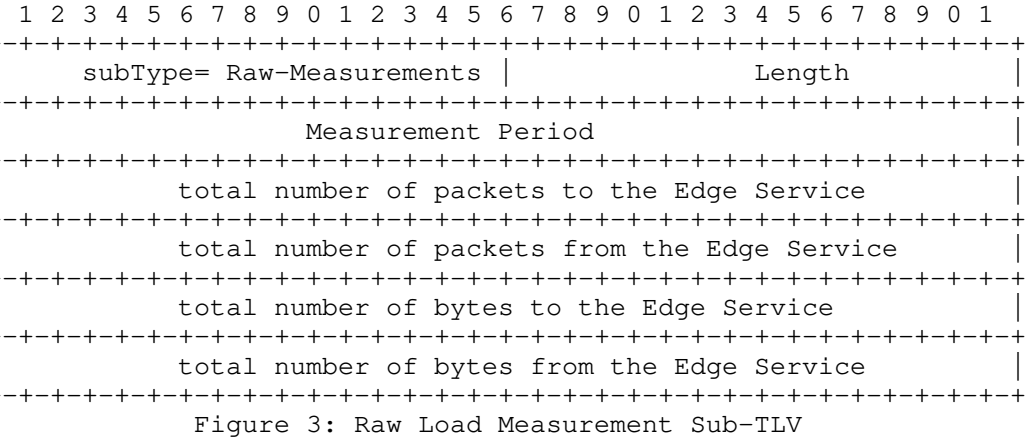


Figure 3: Raw Load Measurement Sub-TLV

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. Consideration for Optimal Paths Selection

When an ingress router receives BGP updates for the same IP address from multiple routers, all those egress routers are considered as the potential paths (or next hops) for the IP address (i.e., if the BGP Add Path is supported). For the selected services, the ingress router's BGP engine would call a Plugin function that can select paths based on the cost associated with the client route received, such as Site-Capacity-Index, Site Preference, load index, and network cost. The Plugin function is called Cost Compute Engine throughout this document. When any of those factors goes to 0, the effect is the same as the egress router not reachable, which triggers the ingress nodes to switch to another egress router. But when any of those factors just degrade, the effect could be a path to another egress router becoming more optimal.

Suppose a destination address for aa08::4450 can be reached by three next hops (R1, R2, R3). Further, suppose the local BGP's Compute Engine Identifies the R1 as the optimal next hop for flows to be sent to this destination (aa08::4450). The Cost Compute Engine can insert a higher weight for the path towards R1 for the prefix.

Note: The Edge Service Metadata Path Attribute are applicable to different NLRIs.

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.

[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, 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