Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 30 October 2025

L. Dunbar
Futurewei
K. Majumdar
Oracle
C. Li
Huawei Technologies
G. Mishra
Verizon
Z. Du
China Mobile

28 April 2025

BGP Extension for 5G Edge Service Metadata
draft-ietf-idr-5g-edge-service-metadata-29

Abstract

This draft describes a new Edge Metadata Path Attribute and some Sub-TLVs for egress routers to advertise the Edge Metadata about the 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 a specific User Equipment (UE).

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

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/.

Copyright Notice

Table of Contents

1. Introduction

This document describes a new Edge Metadata Path Attribute added to a BGP UPDATE message [RFC4271] for egress routers to advertise the Metadata about 5G low latency edge services directly attached to the egress routers. 5G [TS.23.501-3GPP]is characterized by having edge services closer to the Cell Towers reachable by Local Data Networks (LDN). From an IP network perspective, the 5G LDN is a limited domain [RFC8799] with edge services a few hops away from the ingress nodes. Only selective UE services are considered as 5G low latency edge services.

Note: The proposed edge service Metadata Path Attribute are not intended for the best-effort services reachable via the public Internet. The information carried by the Edge Metadata Path Attribute can be used by the ingress routers to make path selections for selective low latency services based on not only the network distance but also the running environment of the edge cloud sites. The goal is to improve latency and performance for 5G ultra-low latency services.

This extension is targeted for a single domain with a BGP Route Reflector (RR) [RFC4456] controlling the propagation of the BGP UPDATEs. The edge service Metadata Path Attribute is only attached to the low latency services (routes) hosted in the 5G edge cloud sites. These routes are only a small subset of services initiated from UEs, not for UEs accessing many internet sites.

While the proposed Edge Metadata Path Attribute is particularly beneficial for low latency services, the Edge Metadata Path Attributes can be expanded to propagate information about GPU availability, power, or other resources necessary for compute-intensive services such as AI and machine learning. This flexibility makes it a valuable tool for a wide range of applications beyond just low latency services when used within a limited domain network.

2. Conventions used in this document

The following conventions are used in this document.

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 edge services.

gNB: next generation Node B [TS.23.501-3GPP]

RTT: Round-trip Time

PSA: PDU Session Anchor (UPF) [TS.23.501-3GPP]

UE: User Equipment

UPF: User Plane Function [TS.23.501-3GPP]

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 [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Edge Metadata Influenced Ingress Node Behavior

The goal of this edge service Metadata Path Attribute is for egress routers to propagate the metrics about the running environment for a subset of edge services to ingress routers so that the ingress routers can make path selections based on not only the routing cost but also the running environment for those edge services. The BGP speakers that do not support the Edge Metadata Path Attribute can ignore the Edge Metadata Path Attribute in a BGP UPDATE Message. All

intermediate nodes can forward the entire BGP UPDATE as it is. Multiple metrics can be attached to one Metadata Path Attribute. One Metadata Path Attribute can contain computing service capability information, computing service states, computing resource states of the corresponding edge site, or more. Computing service capability information can be used to record information of the computing power node or initialization deployment information for computing service initialization. Computing service states can include one of the service connection numbers, service duration, and so on. Computing resource states can be detailed information on computing resources such as CPU/GPU. They can also be an abstract metric from these detailed parameters to indicate the resource status of the edge site. There could be more metrics about the running environment being attached to the Metadata Path Attribute; e.g., some of the metrics being discussed by the IETF CATS Working Group. This document illustrates a few examples of Sub-TLVs of the metrics under the edge service Metadata Path Attribute:

- the site physical availability index,

- the site preference index,

- the service delay predication index x, and

- the raw load measurement.

This section specifies how those Metadata impact the ingress node’s path selections.

3.1. Edge Metadata Influenced BGP Path Selection

When an ingress router receives BGP UPDATEs for the same IP prefix from multiple egress routers, all these egress routers’ loopback addresses are considered as the next hops for the IP prefix. For the selected low latency edge services, the ingress router BGP engine would call an edge service Management function that can select paths based on the edge service Metadata received. Section 5.1 has an exemplary algorithm to compute the weighted path cost based on the edge service Metadata carried by the Sub-TLV(s) specified in this document.

Section 5 has the detailed description of the edge service Metadata influenced optimal path selection.

3.2. Ingress Router Forwarding Behavior

When the ingress router receives a packet and does a lookup on the route in the FIB, it determines the destination prefix’s entire path including the optimal egress node. The ingress router encapsulates the packet destined towards the optimal egress router. For routes that carry the Metadata Path Attribute but lack the Tunnel Encapsulation Path Attribute [RFC9012], it is recommended that the ingress router encapsulate the original packet using an IP-in-IP header. This encapsulation ensures that intermediate nodes not supporting the Metadata Path Attribute do not forward the packet to unintended destinations. The outer header should set the destination address to the optimal egress router and the source address to the ingress router.

For routes without the Metadata Path Attribute, no changes are required. Packets are forwarded according to existing behavior: encapsulation is applied when Tunnel Attributes are present, and parkets are forwarded without encapsulation when they are not.

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. Forwarding packets for a particular flow to the same egress router, also known as Flow Affinity, is supported by many commercial routers. Most registered EC services have relatively short-lived flows.

How Flow Affinity is implemented is out of the scope for this document.

3.3. Forwarding Behavior when UEs Move

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 BGP Next Hop for this route. The detailed algorithm is out of the scope of this document.

4. Edge Service Metadata Encoding

4.1. Edge Metadata Path Attribute

The Edge Metadata Path Attribute is an optional non-transitive BGP Path attribute that carries metrics and Metadata about the edge services attached to the egress router. The Edge Metadata Path Attribute (TBD1) consists of a set of Sub-TLVs, and each Sub-TLV contains information for specific metrics of the edge services.

BGP Peers that intend to exchange the Edge Metadata Path Attribute should indicate this by signaling the Edge Metadata Capability (TBD2) in the Open Capabilities field with the format described in Section 4.1.5. The web of BGP peers that exchange the Edge Metadata Path Attributes forms a limited domain, either within a single AS or within a group of ASes under a single Administrative Authority.

The fields within the Edge Metadata Path Attribute and its Sub-TLVs MUST use network byte order (big-endian), where the most significant byte is transmitted first.

4.1.1. Edge Metadata Path Attribute Characteristics

Only a small subset of BGP UPDATE messages include the Edge Metadata Path Attribute. The choice of which prefix to carry the Edge Metadata Path Attribute is determined by local policies. The Edge Metadata Path Attribute can be included in a BGP UPDATE message [RFC4271] together with other BGP Path Attributes [IANA-BGP-PARAMS], such as Communities [RFC4360], NEXT_HOP, Tunnel Encapsulation Path Attribute [RFC9012], and other BGP attributes.

The Edge Metadata Path Attribute has the following characteristics:

- Non-transitive

- Boundary node filtering SHOULD be deployed to remove the BGP Edge Metadata Path Attribute at the administrative boundary to prevent the distribution of the BGP Edge Metadata Path Attribute beyond its intended scope of applicability.

- Can be packed in an UPDATE with both IPv4 and IPv6 NLRI corresponding to SAFI values 1 (Unicast)[RFC4760], 2 (Multicast) [RFC4760], 4 (MPLS Labels)[RFC8277], 65 (VPN) [RFC4364], 128 (MPLS-labeled VPN) [RFC4364] [RFC8277], 129 (Multicast VPN) [RFC6513], 133 (MPLS-based VPLS) [RFC4761], 134 (EVPN) [RFC7432], and IPv6 Anycast [RFC4786].

- MUST contain at least one Edge Metadata Sub-TLV. Multiple Edge

Metadata Sub-TLVs can be included in a Edge Metadata Path Attribute in one BGP UPDATE message. The choice of the Sub-TLVs present in the BGP Edge Metadata Path Attribute is determined by the local policies. Multiple Sub-TLVs may be carried by a single BGP Edge Metadata Path Attribute.

4.1.2. Edge Metadata Path Attribute Processing

A BGP speaker that advertises a BGP UPDATE message received from one of its neighbors SHOULD advertise the BGP Edge Metadata Path Attribute received with the UPDATE message without modification only when forwarding to peers within the same domain. Otherwise, the Edge Metadata Path Attribute should be removed. If the UPDATE message did not come with a BGP Edge Metadata Path Attribute, the speaker MAY attach a BGP Edge Metadata Path Attribute to the UPDATE message, if configured to do so, provided that the modification adheres to the domain’s policies and security guidelines.

A BGP Peer receiving a BGP Edge Metadata Path Attribute should ignore Sub-TLVs with unknown types and process the recognized Sub-TLVs. BGP Peers should not delete any Sub-TLV from the BGP Edge Metadata Path Attribute.

To prevent forwarding loops and ensure consistent routing decisions, it is essential that all BGP peers within an Autonomous System (AS) adopt a unified approach to handling BGP Edge Metadata Path Attributes. Specifically, BGP peers should consistently ignore Sub-TLVs with unknown types while processing the recognized Sub-TLVs. Additionally, BGP peers should refrain from deleting any Sub-TLV from the BGP Edge Metadata Path attribute. This ensures that all peers have a common understanding of the routing information and reduces the risk of routing inconsistencies that could lead to forwarding loops.

4.1.3. Sub-TLVs Data Processing

By default, a BGP speaker does not report any unrecognized Sub-TLVs within a Edge Metadata Path Attribute unless configured to send a notification to its management system. The ingress node should be configured with an algorithm to combine the recognized metrics carried by the Sub-TLVs within a Edge Metadata Path Attribute of the received BGP UPDATE message.

To ensure consistent route selection, a deployment specific algorithm should be configured across all ingress nodes to factor in the Edge Metadata's contribution alongside existing policies. This will help the ingress node make informed decisions about the optimal path to the next-hop, considering both traditional routing factors and the additional insights provided by the Edge Metadata.

4.1.4. Edge Metadata Path Attribute Handling Procedure

The Edge Metadata Path Attribute MUST contain at least one Edge Metadata Sub-TLV. Multiple Edge Metadata Sub-TLVs can be included in a Edge Metadata Path Attribute in one BGP UPDATE message. The content of the Sub-TLVs present in the BGP Edge Metadata Path Attribute is determined by configuration. The domain ingress nodes should process the recognized Sub-TLVs carried by the Edge Metadata Path Attribute and ignore the unrecognized Sub-TLVs. By default, a BGP speaker does not report any unrecognized Sub-TLVs within a Edge Metadata Path Attribute unless configured to send a notification to its management system. The ingress router should be configured with an algorithm to consider the recognized metrics carried by the Sub-TLVs within a Edge Metadata Path Attribute of the received BGP UPDATE message.

4.1.5. Edge Metadata Processing Capability in BGP OPEN Message

The "Capabilities Optional Parameter" [RFC5492] allows a BGP speaker to indicate its capabilities during the BGP OPEN message exchange. The Capabilities Optional Parameter is a triple that includes a one-octet Capability Code, a one-octet Capability length, and a variable-length Capability Value.

To enable support for the Edge Metadata Path Attribute, a new Edge Metadata Processing Capability code (TBD2) is defined. This capability allows a BGP speaker to communicate its ability to process the Edge Metadata Path Attribute for specified AFI and SAFI pairs.

The Value Field of the Edge Metadata Processing Capability:

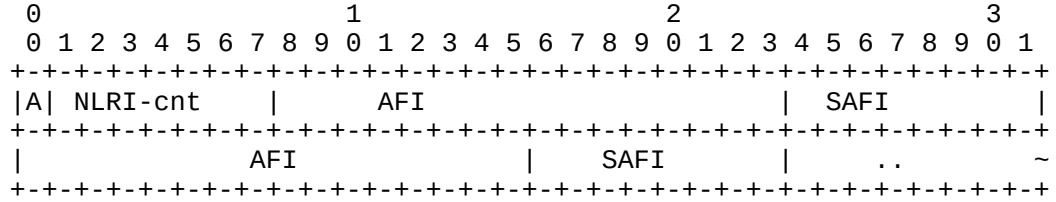

Figure 1: Edge Metadata Capability Value Field

Where:

- A Flag(1 bit): Set to 1 indicates that the Edge Metadata attribute can be attached to any AFI/SAFI. Set to 0 indicates that the Edge Metadata attribute is restricted to specific AFI/SAFI pairs listed in the remainder of the Open Capabilities.

- NLRI-CNT (7 bits): Indicates the number of AFI/SAFI pairs specified in the OPEN Capability.

- AFI (16 bits): Address Family Identifier.

- SAFI (8 bits):Sub-address Family identifier.

If a BGP speaker does not include the Edge Metadata Processing Capability in its BGP OPEN message for a specific BGP session, or if it does not receive the Edge Metadata Processing Capability from its peer on that session, it MUST NOT send any BGP UPDATE message on that session that bind the Edge Metadata Path Attribute to any prefix.

4.2. The Site Preference Index Sub-TLV

Different services might have different preference index values configured for the same site. For example, Service-A requires high computing power, Service-B requires high bandwidth among its microservices, and Service-C requires high volume storage capacity. For a DC with relatively low storage capacity but high bisectional bandwidth, its preference index value for Service-B is higher and lower for Service-C. Site Preference Index can also be used to achieve stickiness for some services.

It is out of the scope of this document how the preference index is determined or configured.

The Site Preference Index Sub-TLV has the following format:

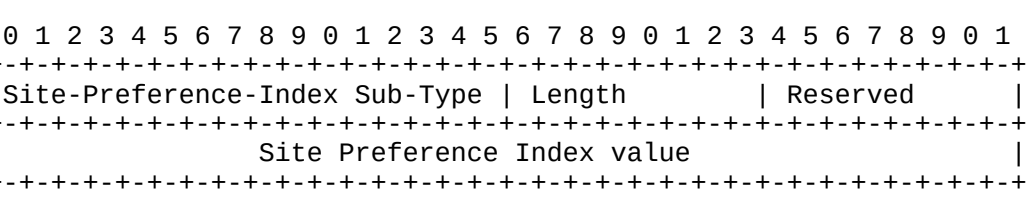

Figure 2: Site Preference Index Sub-TLV

- Site-Preference-Index Sub-Type (16 bits): 1 (specified in this document).

- Length (8 bits): Specifies the total length in octets of the value field (not including the Type and Length fields). For the Site-Preference-Index Sub-Type, the length should be set to 5.

- Reserved: Reserved for future use. In this version of the document, the Reserved field MUST be set to zero and MUST be ignored upon receipt. Received values MUST be propagated without change.

- Site Preference Index value: 1 .. (2^32-1); the higher the value, the more preference for the site. Site Preference Index value == 0 is reserved, and the Site-Preference-Index Sub-TLV should be ignored when 0 is received..

4.3. Site Physical Availability Index Metadata

The Site Physical Availability Index indicates the percentage of impact on a group of routes associated with a common physical characteristic, for example, a pod, a row of server racks, a floor, or an entire DC. The purpose is to use one UPDATE message to indicate a group of routes of different NLRIs impacted by a physical event. For example, a power outage to a pod can cause the Site Physical Availability Index to be 0% for all the routes in the pod. Partial fiber cut to a row of shelves can cause the Site Physical Availability Index to be 50% for all the routes in those shelves. The value is 0-100, with 100% indicating the site is fully functional, 0% indicating the site is entirely out of service, and 50% indicating the site is 50% degraded.

It is recommended to assign each route with one Site-ID. When a route is associated with multiple Site-IDs, the latest BGP UPDATE will override any previous associations. For example, one DC can use POD number as Site-ID, another DC can use Row of Shelves as the Site-ID.

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

When a physical failure occurs at an edge site (or a pod), many instances can be affected, and the associated routes (i.e., IP addresses) may not be easily aggregated. Instead of sending numerous BGP UPDATE messages to ingress routers for each impacted instance, the egress router can send a single BGP UPDATE to indicate the site’s

physical capacity availability. Based on this update, ingress routers can decide to reroute all or some of the affected instances, depending on the extent of the site's degradation. This approach significantly improves efficiency, particularly when fault detection within an edge site relies on proprietary or deployment-specific mechanisms.

The BGP UPDATE for the individual instances (i.e., the routes) can include the Capacity Availability Index solely for ingress routers to associate the routes with the Side-ID. The actual Capacity Availability Index value, i.e., the percentage for all the routes associated with the Side-ID, is generated by the egress routers with the egress routers' loopback address as the NLRI.

The Site Physical Availability Index Sub-TLV has fixed length of 8 Octets, including the Type field. Therefore a Length field is not needed.

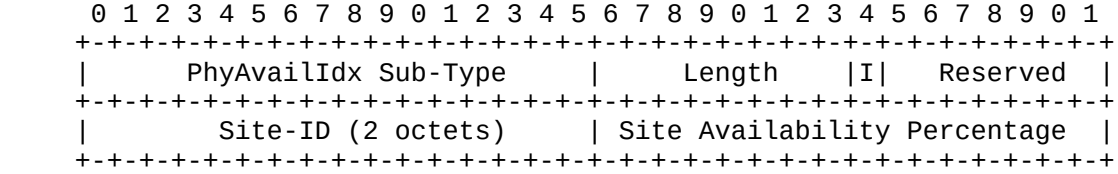

Figure 3: Site Physical Availability Index Sub-TLV

- PhyAvailIdx Sub-Type (16 bits): Indicates teh Site-Physical-Availability-Index Sub-Type=2 (Specified in this document).

- Length (8 bits): Specifies the total length in octets of the value field (not including the Type and Length fields). For the PhyAvailIdx Sub-Type, the length should be set to 5.

Route-Flag I (1 bit): is a flag bit. When set to 1, the Site Availability Index is for BGP speakers (receivers) to associate the routes with the Site-ID. The Site Availability Percentage value is ignored. When set to 0, the BGP speakers (receivers) should apply the Site Availability Index value to all the routes associated with the Site-ID.

Reserved (7 bits): Reserved for future use. The bits are set to zero upon transmission, and ignored upon reception.

- Site ID (16 bits): is an identifier for a group of routes

associated with a common physical characteristic, for example, a pod, a row of server racks, a floor, or an entire DC. The purpose is to use one UPDATE message to indicate a group of routes impacted by a physical event. Those routes might be from different address families or NLRIs. There could be multiple sites connected to one egress router (a.k.a. Edge DC GW).

- Site Availability Percentage (16 bits): When the RouteFlag-I is 1, the Site Availability Percentage is ignored by the Ingress routers. When the RouteFlag I is set to 0, the Site Availability Percentage represents the percentage of the site availability for all the routes associated with the Site-ID; e.g., 100%, 50%, or 0%. When a site goes dark, the Index is set to 0. 50 means 50% functioning. When the value is outside the 0-100% range, the value carried in this Sub-TLV is ignored.

4.3.1. Site Index Associated to Routes

An egress router sets itself as the next hop for a BGP peer before sending an UPDATE with the Edge Metadata Path Attribute that includes the Site Physical Availability Index Sub-TLV. The Site Physical Availability Index Sub-TLV (with RouteFlag-I=1) is for ingress routers to associate the Site Identifier with the prefixes.

4.3.2. BGP UPDATE with standalone Site Availability Index

A BGP UPDATE that includes the Site Availability Index Sub-TLV without specifying attached routes in the NLRI, but instead using the egress router’s loopback address in the NLRI, is referred to as a standalone Site Availability Index BGP UPDATE. When an ingress router receives such a BGP UPDATE containing the Edge Metadata Path Attribute with the standalone Site Physical Availability Index Sub-TLV from Router-X or its RR with the Originator-ID equal to Router-X, the ingress router SHOULD use the site availability index to efficiently reduce or increase the preference for all BGP routes attached to Router-X.

The BGP UPDATE with a standalone Site Availability Index is NOT intended for resolving NextHop.

4.4. Service Delay Prediction

It is desirable for an ingress router to select a site with the shortest processing time for an ultra-low latency service. However, it is not easy to predict which site has "the fastest processing time" or "the shortest processing delay" for an incoming service request because:

- The given service instance shares the same physical infrastructure with many other applications and service instances. Service requests by other applications, UEs, or applications running behavior can impact the processing time for the given service instance.

- The given service instance can be served by a cluster of servers behind a Load Balancer. To the network, the service is identified by one service ID.

- The service complexity is different. One service may call many microservices, need to access multiple backend databases, and need to go through sophisticated security scrubbing functions, etc. Another service can be processed by a few simple steps. Without the application internal logic, it is not easy to estimate the processing time for future service requests.

Even though utilization measurements, like those below, are collected by most data centers, they cannot indicate which site has the shortest processing time. A service request might be processed faster on Site-A even if Site-A is overutilized.

- Server utilization for the server where the instance is instantiated.

- The network utilization for the links to the server where the instance is instantiated.

- The number of databases that the service instance will access.

- The memory utilization of the databases.

The remaining available resource at a site is a more reasonable indication of process delay for future service requests.

- The remaining available Server resources.

- The remaining available network utilization for the links to the server where the instance is instantiated.

- The number of databases that the service instance will access.

- The remaining storage available for the databases.

The Service Delay Prediction Index is a value that predicts processing delays at the site for future service requests. The higher the value, the longer of the delay.

4.4.1. Service Delay Prediction Sub-TLV

While out of scope, we assume there is an algorithm that can derive the Service Delay Prediction Index that can be assigned to the egress router. When the Service Delay Prediction value is updated, which can be triggered by the available resources change, etc., the egress router can attach the updated Service Delay Predication value in a Sub-TLV under the Edge Metadata Path Attribute of the BGP Route UPDATE message to the ingress routers.

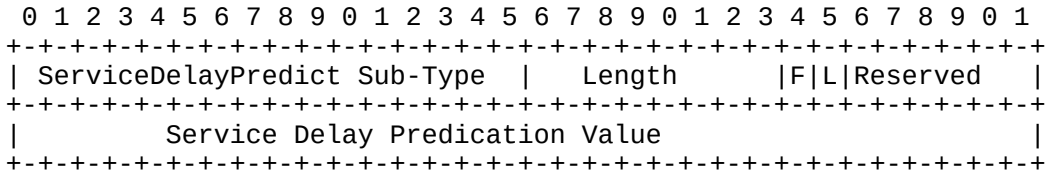

Figure 4: Service Delay Prediction Index Sub-TLV

- ServiceDelayPredict Sub-Type (16 bits): 3 (specified in this document).

- Length (8 bits): specifies the total length in octets of the value field, not including the sub-Type and Length field. The value of Length can be 5 or 9 depends on what format the Service Delay Prediction Vlaue uses.

- Flag (F) (1 bit): Indicates whether the Service Delay is a timer value (F=0) or a relative value (F=1) where a higher value represents a longer delay

- Flag (L) (1 bit): Indicates the unit of measurement for the Service Delay Prediction Value. When the F-flag is set to 0, L=0 specifies the 64-bit NTP Timestamp format, and L=1 indicates milliseconds. If the F-flag is set to 1, the L-flag value is ignored.

- Reserved (6 bits): These bits are reserved for future use and MUST be set to zero. Future documents may specify different uses for these bits.

- Service Delay Predication Value (when the Flag bit is set to 1): an integer in the range of 0-100, with 0 indicating that the service delay is negligible and 100 indicating that the site has the most significant delay compared to all other sites for the same service. When the value is outside the 0-100 range, the value carried in this Sub-TLV is ignored.

- Service Delay Predication Value (when the Flag bit is set to 0): the estimated delay time encoded in the NTP Format as defined in [RFC5905]. When the L-flag is 1, then it is a 64-bit format, otherwise it is a 32-bit short format.

4.5. Raw Measurement Sub-TLV

When ingress routers have embedded analytics tool relying on the raw measurements, it is useful for the egress router to send the raw measurement.

Raw Measurement Sub-TLV has the following format:

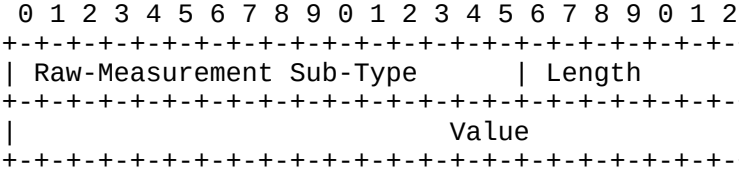

Figure 5: Service Delay Prediction Raw Measurements Sub-TLV

- Raw-Measurement Sub-Type (16 bits): 4 (specified in this document). Indicating raw measurements Metadata associated with the edge service address.

- Length (8 bits): specifies the total length, in octets, of the value field, excluding the Sub-Type and the Length fields. For the Raw-Measurement Sub-Type, the length is determined by the Value field, which carries one or more types of raw measurement.

- Reserved (8 bits): These bits are reserved for future use and MUST be set to zero. Future documents may specify different uses for these bits.

- Value: The value filed can contain multiple types of raw measurements, each represented as a Sub-Sub-TLV.

One example of a raw measurement Metadata Sub-sub-TLV is defined below to convey the total number of packets or bytes transmitted over a specified period for a particular edge service address. When a Data DC GW router cannot directly access the internal state of an edge service, the volume of incoming traffic can be a reliable indicator of its load. A sudden increase in packets or bytes can signal a surge in requests, potentially leading to performance issues or resource constraints on the service side.

To differentiate this measurement from others that may be defined in the future, this document assigns a Sub-sub-Type value of 1 to represent the total packets or bytes transmitted to an edge service address.

Future documents may define additional Sub-sub-types of raw measurement metadata. Each type of raw measurement will have a unique Sub-sub-type value assigned at the time of its specification.

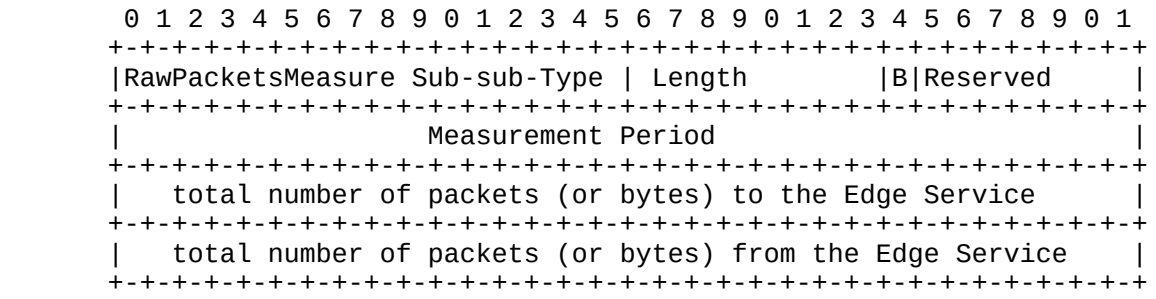

Figure 6: Packets or Bytes Measurements Sub-TLV

- RawPacketsMeasure Sub-sub-Type (8 bits): 1 (specified in this document). Indicating raw measurements of packets or bits transmitted to or from the edge service address.

- Length (8 bits): specifies the total length in octets of the value field, excluding the Sub-sub-Type and the Length fields. For the raw measurements of packets transmitted to or from the edge service address Sub-sub-Type, the length should be 22.

- B flag (1 bit): If set to 0, the raw measurement is the number of packets. If set to 1, the raw measurement is the number of bytes.

- Reserved (7 bits): These bits are reserved for future use and MUST be set to zero.

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

- Total number of packets to the Edge Service (32 bits): This field specifies the total number of packets transmitted to the edge service address over the specified measurement period.

- Total number of packets from the Edge Service (32 bits): This field specifies the total number of packets from the edge service address over the specified measurement period.

The receiver nodes can compute the needed metrics, such as the Service Delay Prediction, for the service based on the raw measurements sent from the egress router and preconfigured algorithms.

4.6. Service-Oriented Capability Sub-TLV

The service-oriented capability Sub-TLV is for distributing information regarding the capabilities of a specific service in a deployment environment. Depending on the deployment, a deployment environment can be an edge site or other types of environments. This information provides ingress routers or controllers with the available resources for the specific service in each deployment environment. It enables them to make well-informed decisions for the optimal paths to the selected deployment environment.

Currently, the Sub-TLV only has an abstract value derived from various metrics, although the specifics of this derivation are beyond the scope of this document. Importantly, this value is significant only when comparing multiple data center sites for the same service. This value is not meaningful when comparing different services, meaning the capability value relevant to Service A cannot be directly compared with that for Service B. Future enhancements may expand this sub-TLV to include more types of metrics or even raw data that represents direct metrics. This information is important in 5G network environments where efficient resource utilization is crucial for enhancing performance and service quality.

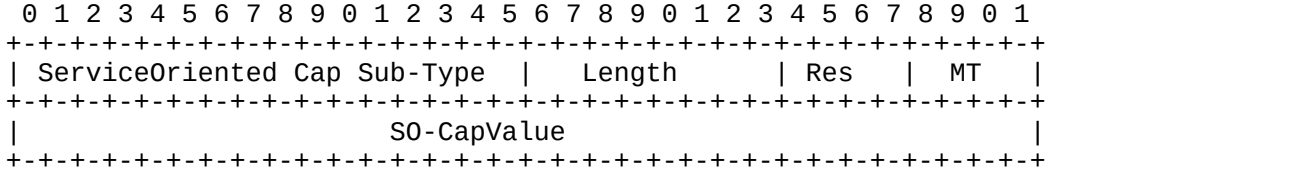

Figure 7: Service-Oriented Capability Sub-TLV

- ServiceOriented Cap Sub-Type (16 bits): 5 (specified in this document).

- Length (8 bits): Specifies the total length in octets, excluding the sub-Type and Length fields. For the ServiceOriented Cap Sub-Type, the Length should be 5.

- Res (4 bits): These bits are reserved for future use and MUST be set to zero.

- MT (Metric Type)(4 bits): An unsigned 4 bits integer. When the MT value is set to 0, it indicates the SoCapValue field contains a normalized metric derived from multiple metric types. The rules for deriving this normalized metric are out of scope of this document and defined by per-service. Additional metric types may be defined in future documents.

- SO-CapValue (32 bits): The Service-Oriented Capability Abstract Value is an integer between 0 and 2^32-1. A larger number means higher capability, and a value of 0 indicates the site has the lowest relative capability for the service. The method used to derive this value is beyond the scope of this document.

Multiple Service-Oriented Capability Sub-TLVs with different metric types can be encoded in a Edge Metadata Path Attribute, indicating that multiple metrics are carried. However, if more than one Service-Oriented Capability Sub-TLVs with the same metric type are encoded in a Edge Metadata Path Attribute, only the first one will be processed and the others will be ignored in processing.

4.7. Service-Oriented Available Resource Sub-TLV

The "Service-Oriented Available Resource Sub-TLV" is for distributing a metric that measures the real-time avaiable resources allocated for processing specific services or applications at an edge site. This Sub-TLV complements the "Service-Oriented Capability Sub-TLV" described in Section 4.6, which addresses the static resource capability of a site for a service. While the Capability Abstract Value provides a baseline understanding of a site’s potential to handle a service, the Available Resource metric offers a dynamic perspective by quantifying how much of this capacity is currently available. This distinction is crucial for managing resource efficiency and responsiveness in network operations, ensuring that capabilities are not only available but also optimally used to meet the actual service demands.

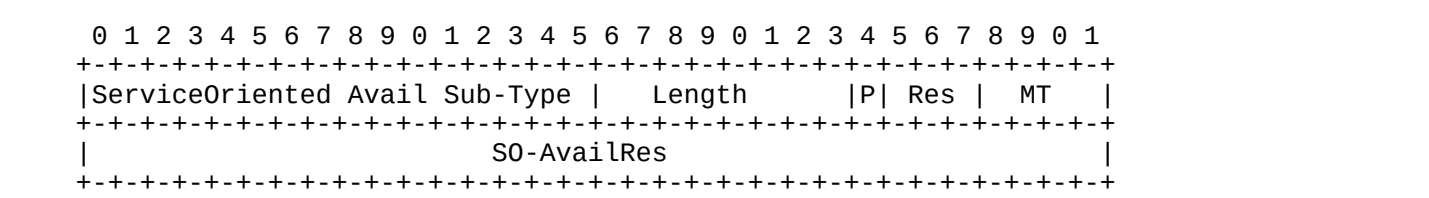

Figure 8: Service-Oriented Available Resource Sub-TLV

- ServiceOriented Avail (Service-Oriented Available Resource) Sub-Type: 6 (specified in this document).

- Length (8 bits) Specifies the total length in octets, excluding the sub-Type and the length field. For the ServiceOriented Available Resource Sub-Type, the Length should be 5.

- Flag (P): Is a single-bit Percentage flag. When it is set to 1, it indicates the value is the Service-Oriented Available Resource in percentage. When the "P" flag is set to 0, the value in this Sub-TLV is the abstract value of the available resource.

- Res (3 bits): These bits are reserved for future use and MUST be set to zero.

- MT (4 bits) Metric Type. This document defines a default metric type as value 0, indicating this is the normalized metric derived by multiple type of metrics. The rules to derive the normalized metric are out of scope of this document and defined by the service. Other Metric Types could be defined by other documents in the future.

- SO-AvailRes (32 bits): When the P-Flag bit is set to 1, Service-Oriented Available Resource Value is a percentage (0-100), with 0 indicating that 0% of the capability is available and 100 indicating that 100% of the capability is available. When the value is outside the 0-100 range, the value carried in this Sub-TLV is ignored. For example, Capacity value is 50 and the SO-AvailRes is 50 when P-flag is set, it means 50% of 50 unit of resource is available, while 25 unit of resource is available in this site for the service. When the P-flag is 0, then the value of this filed is the abstract value of the available resource. For example, When the capacity value is 50, and the SO-AvailRes is 50, it means all the resource is available.

Multiple Service-Oriented Available Resource Sub-TLVs with different metric types can be encoded in a Edge Metadata Path Attribute, indicating that multiple metrics are carried. However, if more than one Service-Oriented Available Resource Sub-TLVs with the same metric type are encoded in a Edge Metadata Path Attribute, only the first one will be processed and the others will be ignored in processing.

5. Service Metadata Propagation Scope

The propagation scope of the Edge Metadata Path Attribute needs careful consideration to ensure it does not inadvertently leak to other BGP domains. According to Section 3 of [ATTRIBUTE-ESCAPE], it is necessary for the Route Reflector (RR) to be upgraded to constrain the propagation scope when propagating the metadata path attributes. Therefore, the Edge Metadata Path Attribute originator sets the attribute as Non-transitive when sending the BGP UPDATE message to

its corresponding RR. Non-transitive attributes are only guaranteed to be dropped during BGP route propagation by implementations that do not recognize them, ensuring that the Edge Metadata path attributes do not propagate beyond the intended scope.

The RR can append the NO-ADVERTISE well-known community to the BGP UPDATE message with the Edge Metadata Path Attribute when forwarding it to the ingress routers. This signals to the ingress nodes that the associated route's Metadata Path Attribute should not be further advertised beyond their scope. This precautionary measure ensures that the receiver of the BGP UPDATE message refrains from forwarding the received update to its peers, preventing the undesired propagation of the information carried by the Metadata Path Attribute.

5.1. AS-Scope SubTLV

To address the potential issue where the NO-ADVERTISE well-known community of the BGP UPDATE message can be dropped by some routers, a new AS-Scope Sub-TLV can be included in the Metadata Path Attribute to prevent the Metadata Path Attribute from being leaked to unintended Autonomous Systems (ASes). The AS-Scope Sub-TLV will enforce stricter control over the propagation of the metadata by associating it with specific AS numbers.

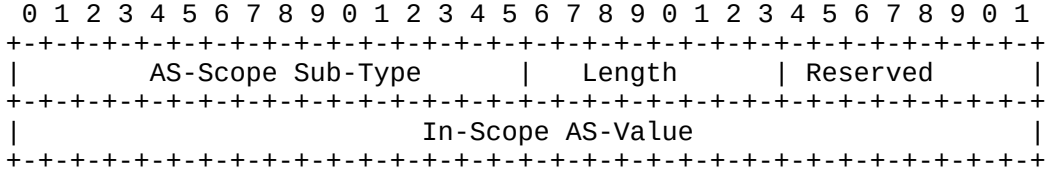

Figure 9: AS-Scope Sub-TLV

- AS-Scope Sub-Type (16 bits): 7 (specified in this document).

- Length (8 bits) Specifies the total length in octets, excluding the sub-Type and the length field. For the AS-Scope Sub-Type, the Length should be 6.

- Reserved (8 bits): These bits are reserved for future use and MUST be set to zero.

- In-Scope AS-Value (32 bits): AS value that is recognized by the BGP speaker in the domain.

5.1.1. AS-Scope Value Checking Procedure

When a router receives a BGP UPDATE message containing the AS-Scope Sub-TLV, it must perform the following steps to process the AS-Scope value:

- AS Recognition: The router will check the AS value in the AS-Scope Sub-TLV.

- If the AS value matches the local AS or a recognized AS in its configuration, the router will process the update as usual. If the AS value does not match or is not recognized, the router SHOULD NOT process the Edge Metadata Path Attribute values in the BGP UDPATE and SHOULD NOT propagate the received BGP UPDATE to other nodes. I.e., treat-as-withdraw behavior will be used.

Example Usage:

Consider a scenario where a router in AS 65001 advertises a BGP UPDATE message with the AS-Scope Sub-TLV set to AS 65001. When another router in AS 65002 receives this UPDATE, it will check the AS-Scope Sub-TLV value:

Since AS 65002 does not match the AS value 65001, the router in AS 65002 will drop the UPDATE, preventing the metadata from leaking into AS 65002.

This mechanism ensures that the metadata remains confined to the intended ASes, enhancing the security and control over the propagation of BGP metadata.

6. Policy Based Metadata Integration

This section describes how the information carried in the Edge Metadata Path Attribute is integrated into the BGP route selection process. RR and Ingress nodes can incorporate metadata into their route selection, depending on the network deployment and local policy configuration. To ensure compliance with Section 9.1.1 of [RFC4271], metadata-based preferences must be applied after the LOCAL_PREF attribute is set for iBGP routes or after local policies are applied for eBGP routes.

Deployment Specific Attribute Selection:

Each deployment, by the local policy, chooses the subset of available metadata attributes to use in setting the local preference for the route. This tailors the route selection process to the specific needs and policies of the network. Both RRs and Ingress nodes can selectively integrate metadata attributes into their computations based on these policies.

To ensure consistent routing decisions when integrating Edge Metadata Path Attributes, two deployment models are described:

- Centralized RR Model:

  If the RR is acting as the deployment’s "server" for best paths, it is recommended that routers in the AS ONLY peer through the RR. This ensures that the RR serves as the single point of policy-based computation, and all ingress routers receive consistent routes that account for the Edge Metadata Path Attribute.

- Consistent Distributed Model:

  If routers in the AS are partially meshed and allowed to exchange iBGP routes directly, the RR must be treated as just another node. In this case: All nodes, including the RR, must implement the same policy for integrating the Edge Metadata Path Attribute and computing route preference. The procedure for combining metadata and traditional BGP attributes should be consistent across all nodes, ensuring that all routers converge on the same "best" path when presented with the same set of routes and metadata.

Influence on the BGP Decision Process:

- At the Route Reflector (RR):

  In deployments where RRs are responsible for pre selecting routes, the RR integrates metadata and traditional BGP attributes when determining the "best" route. The RR reflects only the selected route to its client routers (e.g., Ingress PEs), ensuring alignment with service specific requirements. To mitigate hot potato routing issues, deployments SHOULD consider Optimal Route Reflection (ORR) as specified in [RFC9107], which enables the RR to compute and advertise routes based on ingress routers’ perspectives rather than the RR’s own location.

- At the Ingress Node:

  When the RR reflects multiple routes (e.g., using Add Paths), the

Ingress node receives all candidate routes. It then integrates metadata attributes with traditional BGP attributes to compute the preference for a route. This allows the Ingress node to make service specific routing decisions based on its local policy and visibility into metadata.

Policy Driven Combined Preference Evaluation:

The preference for a route is computed based on a weighted combination of metadata attributes and traditional BGP attributes. The weights are determined by local policy:

- Metadata and traditional BGP attributes are integrated into a single preference value using a deployment specific algorithm.

- Either the RR or the Ingress node selects the route with the highest computed preference value for reflection or traffic steering.

Handling Degraded Metrics:

When critical metadata metrics, such as the Capacity Availability Index or Service Delay Prediction Index, degrade beyond a configured threshold, local BGP policy may treat the affected route as ineligible for traffic steering. This behavior is equivalent to BGP local policy declaring the route is not eligible for route selection. This ensures that traffic is not routed to service instances that not capable to process the services, preserving the quality of service for critical applications.

Example Scenarios for Policy Based Route Selection:

A BGP peer uses local policy run over the route (prefix plus attribute) to select the best route and then use tie breaking based on [RFC4271]. This section simply provides 3 examples of how local policy might weigh the Metadata metrics during that policy selection.

Scenario 1: Local Policy Prioritizes Metadata Metrics Over Traditional BGP Attributes.

The local policy assigns a higher weight to metadata metrics when computing the preference for routes. The selection process follows these steps:

- Compute a preference value based on the weighted combination of metadata attributes and traditional BGP attributes, with metadata metrics having higher weight.

- Prefer the route with the highest computed preference value.

- Resolve remaining ties using traditional BGP tie breaking criteria (e.g., eBGP over iBGP, lowest IGP metric, oldest route, lowest route ID).

Scenario 2: Local Policy Weighs Metadata Metrics and Traditional BGP Attributes Equally.

The local policy assigns equal weight to metadata and traditional BGP attributes during preference computation. The selection process is as follows:

- Compute a preference value by equally weighing metadata derived metrics and traditional BGP attributes.

- Prefer the route with the highest computed preference value.

- Resolve remaining ties using traditional BGP tie breaking criteria.

Scenario 3: Local Policy Prioritizes Traditional BGP Attributes Over Metadata Metrics

The local policy assigns a higher weight to traditional BGP attributes. The selection process follows these steps:

- Compute a preference value based on the weighted combination of metadata attributes and traditional BGP attributes, with traditional attributes having higher weight.

- Prefer the route with the highest computed preference value.

- Resolve remaining ties using traditional BGP tie breaking criteria.

Equal Cost Multi Path (ECMP) in BGP Route Selection:

When the BGP decision process identifies multiple paths with equal preference after considering both Edge Metadata Path Attributes and traditional BGP attributes, BGP can pass these paths to the forwarding engine to enable ECMP.

This Policy Based Metadata Integration approach enables network operators to incorporate Edge Metadata Path Attributes into BGP route selection based on their specific operational goals and requirements, while maintaining compatibility with traditional BGP operations.

7. Minimum Interval for Metrics Change Advertisement

Route Churn Considerations

While the mechanism detailed in this document aims to provide dynamic metrics like Capacity Availability Index, Site Delay Prediction Index, Service Delay Prediction Index, and Raw Measurement to optimize path selection, it is essential to consider the broader implications of metric-induced churn. Particularly, in the context of routes used for BGP nexthop resolution (e.g., labeled unicast), frequent changes in these metrics can lead to significant churn not only for the prefixes carrying the data but also for dependent routes.

In normal operation, the metadata associated with a prefix is propagated along with BGP UPDATE messages as per standard BGP behavior. The advertisement interval is governed by the underlying BGP mechanisms, such as the MRAI timer (typically 30 seconds for iBGP). This document does not propose a new periodic advertisement mechanism independent of routing updates. If metadata attributes (e.g., compute availability, service locality) change, a BGP UPDATE is triggered accordingly. If there is no change to the advertised metadata, no additional UPDATE is sent, in order to avoid unnecessary update churn and to comply with BGP best practices. Any active or proactive refresh mechanisms for metadata would require explicit triggers and change detection mechanisms, which are outside the scope of this document.

This behavior is analogous to the impacts observed with RSVP auto-bandwidth, which can introduce considerable instability within a network. Such route churn can propagate through the network, causing a cascade of UPDATEs and potential route flaps, thereby affecting overall network stability and performance.

To mitigate these effects, network operators should carefully manage the advertisement intervals of these dynamic metrics, ensuring they are set to avoid unnecessary churn. The default minimum interval for metrics change advertisement, set at 30 seconds, is designed to balance responsiveness with stability. However, in scenarios with higher sensitivity to route stability, operators may consider increasing this interval further to reduce the frequency of UPDATEs.

Significant load changes at EC data centers can be triggered by short-term gatherings of UEs, like conventions, lasting a few hours or days. Therefore, a high metrics change rate can persist for hours or days.

8. Validation and Error Handling

The Edge Metadata Path Attribute is an optional non-transitive BGP Path attribute that carries metrics and metadata about the edge services attached to the egress router. The Edge Metadata Path Attribute, to be assigned by IANA , consists of a set of Sub-TLVs, and each Sub-TLV contains information for specific metrics of the edge services.

When more than one sub-TLV is present in a Metadata Path Attribute, they are processed independently. Suppose a Edge Metadata Path Attribute can be parsed correctly but contains a Sub-TLV whose type is not recognized by a particular BGP speaker; that BGP speaker MUST NOT consider the attribute malformed. Instead, it MUST interpret the attribute as if that Sub-TLV had not been present. Logging the error locally or to a management system is optional. If the route carrying the Edge Metadata path attribute is propagated with the attribute, the unrecognized Sub-TLV remains in the attribute.

9. Manageability Considerations

The edge service Metadata described in this document are only intended for propagating between ingress and egress routers of one single BGP Administrative Domain [RFC1136]. A single BGP Administrative Domain can consist of one AS or multiple ASes.

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.

10. Security Considerations

The proposed edge service Metadata are advertised within the trusted domain of 5G LDN’s ingress and egress routers. The ingress routers should not propagate the edge service Metadata to any nodes that are not within the trusted domain.

To prevent the BGP UPDATE receivers (a.k.a. ingress routers in this document) from leaking the Edge Metadata Path Attribute by accident to nodes outside the trusted domain [ATTRIBUTE-ESCAPE], the following practice should be enforced:

- The Edge Metadata Path Attribute is non-transitive. Per [RFC4271], non-transitive Path Attributes are dropped during BGP route propagation by implementations that do not recognize them.

- Route Reflectors can append the NO-ADVERTISE well-known community

to the BGP UPDATE message with Edge Metadata Path Attribute when forwarding to the ingress routers. By doing so, the Route Reflector signals to ingress nodes that the routes with the Edge Metadata Path Attribute should not be further advertised beyond their scope. This precautionary measure ensures that the receiver of the BGP UPDATE message refrains from forwarding the received UPDATE to its peers, preventing the undesired propagation of the information carried by the Edge Metadata Path Attribute.

BGP Route Filtering or BGP Route Policies [RFC5291] can also be used to ensure that BGP UPDATE messages with Edge Metadata Path Attribute attached do not get forwarded out of the administrative domain. BGP route filtering [RFC5291] allows network administrators to control the advertisements and acceptance of BGP routes, ensuring that specific routes do not leak outside the intended administrative domain. Here are the steps to achieve this:

- Use Route Filtering: Implement route filtering policies on the ingress routers to restrict the propagation of BGP UPDATE messages for the registered 5G edge services beyond the administrative domain. You can use access control lists (ACLs), prefix lists, or route maps to filter the BGP routes classified as the 5G edge services, which need the Edge Metadata Path Attributes to be distributed from egress routers to ingress routers.

- Filter by Prefix: Use prefix filtering to specify which IP prefixes should be advertised to peers and which should be suppressed. This step ensures that only authorized routes are sent to external peers.

- Use Route Maps: Route maps provide a flexible way to filter and manipulate BGP route advertisements. You can create route maps to match specific conditions and then apply them to the BGP configuration.

11. IANA Considerations

11.1. Edge Metadata Path Attribute

IANA has assigned value 42 to the "Edge Metadata Path Attribute" in the "BGP Path Attributes" registry. The reference for this assignment is [this document].

Value	Description	Reference
42	Edge Metadata Path Attribute	[this document]

11.2. Edge Metadata Capability Code

IANA is requested to assign a Capability Code from the "BGP Capability Codes" registry, within the range 64-238, for the Edge Metadata Capability in the BGP OPEN message, following the First Come, First Served (FCFS) policy.

Value	Description	Reference
TBD2	Edge Metadata Capability in BGP OPEN	[This document]

11.3. Edge Metadata Path Attribute Sub-Types

IANA is requested to create a new sub-registry under the Edge Metadata Path Attribute registry as follows:

Name: Sub-TLVs under the "Edge Metadata Path Attribute"

Registration Procedure: Expert Review [RFC8126].

Detailed Expert Review procedure will be added per [RFC8126].

Reference: [this document]

Sub-Type	Description	Reference
0	reserved	[this document]
1	Site Preference Index	[this document:4.3]
2	Site Physical Avail Index	[this document:4.4]
3	Service Delay Predication	[this document:4.5]
4	Raw Measurement	[this document:4.6]
5	Service-Oriented Capability	[this document:4.7]
6	Service-Oriented Available Resource	[this document:4.8]
7	AS-Scope	[this document:5.1]
8-65534	unassigned	
65535	reserved	[this document]

12. Contributors

Changwang Lin

New H3C Technologies

China

Email: linchangwang.04414@h3c.com

13. Acknowledgements

Acknowledgements to Jeff Haas, Tom Petch, Adrian Farrel, Alvaro Retana, Robert Raszuk, Sue Hares, Shunwan Zhuang, Donald Eastlake, Dhruv Dhody, Cheng Li, DongYu Yuan, and Vincent Shi for their suggestions and contributions.

14. References

14.1. 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>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <https://www.rfc-editor.org/info/rfc4761>.

[RFC4786] Abley, J. and K. Lindqvist, "Operation of Anycast Services", BCP 126, RFC 4786, DOI 10.17487/RFC4786, December 2006, <https://www.rfc-editor.org/info/rfc4786>.

[RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, DOI 10.17487/RFC5291, August 2008, <https://www.rfc-editor.org/info/rfc5291>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

[RFC5905] Mills, D., Martin, J., Ed., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, DOI 10.17487/RFC5905, June 2010, <https://www.rfc-editor.org/info/rfc5905>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/ BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <https://www.rfc-editor.org/info/rfc6513>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

14.2. Informative References

[ATTRIBUTE-ESCAPE] J. Haas, "BGP Attribute Escape", July 2023, <https://datatracker.ietf.org/doc/draft-haas-idr-bgp-attribute-escape/>.

[IANA-BGP-PARAMS] IANA, "BGP Path Attributes", BGP Path Attributes https://www.iana.org/assignments/bgp-parameters/.

[RFC1136] Hares, S. and D. Katz, "Administrative Domains and Routing Domains: A model for routing in the Internet", RFC 1136, DOI 10.17487/RFC1136, December 1989, <https://www.rfc-editor.org/info/rfc1136>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[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>.

[RFC8799] Carpenter, B. and B. Liu, "Limited Domains and Internet Protocols", RFC 8799, DOI 10.17487/RFC8799, July 2020, <https://www.rfc-editor.org/info/rfc8799>.

[TS.23.501-3GPP]
3rd Generation Partnership Project (3GPP), "System Architecture for 5G System; Stage 2, 3GPP TS 23.501 v2.0.1", December 2017.

Appendix A. Service Delay Prediction Based on Load Measurement

When data centers detailed running status are not exposed to the network operator, historic traffic patterns through the egress routers can be utilized to predict the load to a specific service. For example, when traffic volume to one service at one data center suddenly increases a huge percentage compared with the past 24 hours average, it is likely caused by a larger than normal demand for the service. When this happens, another data center with lower-than-average traffic volume for the same service might have a shorter processing time for the same service.

Here are some measurements that can be utilized to derive the Service Delay Predication for a service ID:

- Total number of packets to the attached service instance (ToPackets);

- Total number of packets from the attached service instance (FromPackets);

- Total number of bytes to the attached service instance (ToBytes);

- Total number of bytes from the attached service instance (FromBytes);

- The actual load measurement to the service instance attached to an egress router can be based on one of the metrics above or including all four metrics with different weights applied to each, such as:

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

  Where w1/w2/w3/w4 are between 0-1. w1+ w2+ w3+ w4 = 1;

  The weights of each metric contributing to the index of the service instance attached to an egress router can be configured or learned by self-adjusting based on user feedbacks.

The Service Delay Prediction Index can be derived from LoadIndex/24Hour-Average. A higher value means a longer delay prediction. The egress router can use the ServiceDelayPred sub-TLV to indicate to the ingress routers of the delay prediction derived from the traffic pattern.

Note: The proposed IP layer load measurement is only an estimate based on the amount of traffic through the egress router, which might not truly reflect the load of the servers attached to the egress routers. They are listed here only for some special deployments where those metrics are helpful to the ingress routers in selecting the optimal paths.

Appendix B. Service Metadata Influenced Decision Process

B.1. Egress Router Behavior

Multiple instances of the same service could be attached to one egress router. When all instances of the same service are grouped behind one application layer load balancer, they appear as one single route to the egress router, i.e., the application loader balancer’s prefix. Under this scenario, the compute metrics for all those instances behind one application layer balancer are aggregated under the application load balancer’s prefix. In this case, the compute metrics aggregated by the Load Balancer are visible to the egress router as associated with the Load Balancer’s prefix. However, how the application layer Load Balancers distribute the traffic among different instances is out of the scope of this document. When multiple instances of the same service have different paths or links reachable from the egress router, multiple groups of metrics from respective paths could be exposed to the egress router. The egress router can have preconfigured policies on aggregating various metrics from different paths and the corresponding policies in selecting a path for forwarding the packets received from ingress routers. The aggregated metrics can be carried in the BGP UPDATE messages instead of detailed measurements to reduce the entries advertised by the control plane and dampen the routes update in the forwarding plane. Upon receiving packets from ingress routers, the egress router can use its policies to choose an optimal path to one service instance. It is out of the scope of this document how the measurements are aggregated on egress routers and how ingress routers are configured with the algorithms to integrate the aggregated metrics with network layer metrics.

Many measurements could impact and correspondingly reflect service performance. In order to simplify an optimal selection process, egress routers can have preconfigured policies or algorithms to aggregate multiple metrics into one simple one to ingress routers.

Though out of the scope of this document, an egress router can also have an algorithm to convert multiple metrics to network metrics, an IGP cost for each instance, to pass to ingress nodes. This decision-making process integrates network metrics computed by traditional IGP/BGP and the service delay metrics from egress routers to achieve a well-informed and adaptive routing approach. This intelligent orchestration at the edge enhances the service's overall performance and optimizes resource utilization across the distributed infrastructure. When the egress has merged the compute metrics from the local sites behind it, it can include one or more aggregated compute metrics in the Metadata Path Attribute in the BGP UPDATE to the Ingress. Also, an identifier or flag can be carried to indicate that the metrics are merged ones. After receiving the routes for the Service ID with the identifier, the ingress would do the route selection based on pre-configured algorithms (see Section 3 of this document).

B.2. Integrating Network Delay with the Service Metrics

As the service metrics and network delays are in different units, here is an exemplary algorithm for an ingress router to compare the cost to reach the service instances at Site-i or Site-j.

$$Cost\text{-}i = \min\left(w * \frac{ServD\text{-}i * CP\text{-}j}{ServD\text{-}j * CP\text{-}i} + (1-w) * \frac{Pref\text{-}j * NetD\text{-}i}{Pref\text{-}i * NetD\text{-}j}\right)$$

CP-i: Capacity Availability Index at Site-i. A higher value means higher capacity available.

NetD-i: Network latency measurement (RTT) to the Egress Router at the site-i.

Pref-i: Preference Index for Site-i, a higher value means higher preference.

ServD-i: Service Delay Predication Index at Site-i for the service, i.e., the ANYCAST address [RFC4786] for the service.

w: Weight is a value between 0 and 1. If smaller than 0.5, Network latency and the site Preference have more influence; otherwise, Service Delay and capacity availability have more influence.

When a set of service Metadata is converted to a simple metric, a decision process is determined by the metric semantics and deployment situations. The goal is to integrate the conventional network decision process with the service Metadata into a unified decision-making process for path selection.

B.3. Integrating with BGP Route Selection

Not all metadata attributes specified in this document are intended for use in every deployment. Each deployment may choose to consider only a subset of the available metadata attributes based on its specific service requirements.

- Deployment-Specific Attribute Selection:

A deployment may prioritize only certain metadata attributes relevant to its operational needs. For example, one deployment might only use the Service Delay Prediction Index for latency-sensitive applications, while another might focus solely on the Capacity Availability Index to manage resource availability.

- Influence on BGP Decision Process:

The edge service Metadata influences next-hop selection differently from traditional BGP metrics (e.g., Local Preference, MED). Unlike a general next-hop metric that can affect many routes, edge service Metadata selectively impacts optimal next-hop selection for specific routes configured to consider these service-specific attributes. This targeted influence allows for optimized path selection without disrupting broader route decisions.

- Handling Degraded Metrics (Policy-Based):

If a service-specific metric degrades beyond a configured threshold (e.g., the Service Delay Prediction Index exceeds an acceptable delay threshold or the Capacity Availability Index drops below a required level), the ingress router will treat that route as ineligible for traffic steering. This is similar to a BGP route withdrawal, where the degraded route is deprioritized or ignored, even if traditional BGP attributes would otherwise favor it. This ensures that traffic is directed only to service instances that meet the defined performance criteria.

- Fallback to Non-Metadata Routes:

If no suitable routes with the required metadata are available, the BGP decision process defaults to traditional attribute evaluation [RFC4271], ensuring consistent routing even when metadata-specific paths are absent.

This approach provides flexibility and adaptability in routing decisions, allowing each deployment to apply relevant metadata attributes and enforce performance thresholds for improved service quality.

Authors' Addresses

Linda Dunbar
Futurewei
Dallas, TX,
United States of America
Email: ldunbar@futurewei.com

Kausik Majumdar
Oracle
California,
United States of America
Email: kausik.majumdar@oracle.com

Cheng Li
Huawei Technologies
Beijing
China
Email: c.l@huawei.com

Gyan Mishra
Verizon
United States of America
Email: gyan.s.mishra@verizon.com

Zongpeng Du
China Mobile
Beijing
China
Email: duzongpeng@chinamobile.com