Network Working Group
Internet Draft
Intended status: Standard
Expires: January 25, 2023

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

July 25, 2022

BGP AppMetaData for 5G Edge Computing Service
draft-dunbar-idr-5g-edge-compute-app-meta-data-09

Abstract

This draft describes the AppMetaData encoding in the BGP Path Attribute for egress routers to advertise the running status and environment of the directly attached 5G Edge Computing (EC) instances. The AppMetaData 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 destinations. The goal is to improve latency and performance for 5G EC services.

The extension enables an EC server at one specific location to be more preferred than the others with the same IP address to receive data flows from a specific source (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 computing (EC) servers. This document describes a new subTLV, AppMetaData, for egress routers to advertise the running status and environment for the directly attached Edge Computing (EC) servers. The AppMetaData 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 destinations. The goal is to improve latency and performance for 5G Edge Computing services.

2. Conventions used in this document

A-ER: Egress Router to an Application Server, [A-ER] is used to describe the last router that the Application Server is attached. For a 5G EC environment, the A-ER can be the gateway router to a (mini) Edge Computing Data Center.

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.

EC: Edge Computing

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 Edge Computing Hosting Environment. An Edge DC might host 5G core functions in addition to the frequently used application servers.

gNB next generation Node B

L-DN: Local Data Network

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 to advertise Load & Capacity

The goal of the BGP extension is for egress routers to propagate the metrics about their running environment to ingress routers. Here are some examples of the metrics propagated by the egress routers:

- the Load Measurement Index for the attached EC Servers,

- the Capacity Index, and

- Site Preference Index.

This section specifies the Load Index Sub-TLV, Capacity Sub-TLV, and the Site Preference Sub-TLV that can be carried by the Tunnel Encap Path Attribute associated with the routes.

3.1. Ingress Node BGP Path Selection Behavior

3.1.1. AppMetaData 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 EC services, the ingress router's BGP engine would call a Plugin function that can select paths based on the AppMetaData received. The Plugin function is called Load Compute Engine throughout this document.

Suppose a destination address for 5G (S1: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 (S1:aa08::4450). The Load Compute Engine can insert a higher weight for the tunnel associated with R1 for the prefix via the tunnel.

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. Here is one example to illustrate how Flow Affinity can be achieved. This illustration is not to be standardized.

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-Sticky-Service] 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 Load Compute Engine. The forwarding plane encapsulates the packet with a tunnel 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, refer to Section 6 Soft Anchoring on how another path is selected.

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

When a UE moves to a new 5G eNB 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. Load Measurement and Site Preference AppMetaData

The Load Measurement and Site Preference AppMetaData attribute is encoded in an optional subTLV within the Tunnel Encap [RFC9012] Path Attribute.

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

4.1. 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 App Server 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Type (TBD1)           |              Length           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Measurement Period                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Aggregated Load Index to reach the App Server       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Aggregated Load Index Sub-TLV

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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Type (TBD2)         |               Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Measurement Period                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           total number of packets to the AppServer            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           total number of packets from the AppServer          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           total number of bytes to the AppServer              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           total number of bytes from the AppServer            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: Raw Load Measurement Sub-TLV

Type =TBD1: Aggregated Load Measurement Index derived from the Weighted combination of bytes/packets sent to/received from the App server:

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

Where wi is a value between 0 and 1; w1+ w2+ w3+ w4 = 1.

Type= TBD2: Raw measurements of packets/bytes to/from the App Server address.

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

4.2. The Site Preference Index sub-TLV format

The site Preference Index is used to achieve Soft Anchoring [Section 5] an application flow from a UE to a specific location when the UE moves from one 5G site to another.

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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Type (TBD4)          |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   Preference Index                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: Preference Index Sub-TLV

Note: "Site Preference Index" can be more stable for each site. If those values are configured to nodes, they might not need to be included in every BGP UPDATE.

5. Capacity Index AppMetaData

Capacity Index indicates the capacity value for the site or pod where the EC service instances are instantiated. One Edge Cloud site (or Pod) can have larger capacity than another one. One Edge Site can be in full capacity, or reduced capacity. When there is a failure occurring at an Edge site (or pod), many instances can be impacted. 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.

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 Cloud (egress) router visible to the ingress routers can be running fine. Therefore, the ingress routers can't use BFD to detect the failures.

When a site capacity degrades or goes dark, there can be many routes impacted. In addition, the routes (i.e., the IP addresses) in an Edge Cloud Site might not be aggregated nicely, triggering very large number of BGP UPDATE messages (see RFC4271) when a failure occurs.

The Capacity Index can be carried by an opaque Extended Community with a new subtype (Site Capacity Cost), by Tunnel-Encap subTLV, or wide community:

The Opaque Extended community has (with High value = 0x03).

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| 0x03 (1 octet)| Capacity-index|   Reserved    | Usage-Index   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Side-reference (2 octets)     |   Site Capacity Index         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- Sub Type: Capacity-Index subtype (TBD by IANA)

- Usage-Index: Indicating if the Site Capacity Index is absolute value, relative to all the sites/pods attached to the BGP speaker, or percentage, etc.

- Site Reference: The Site-reference can represent a group of routes within one site/pod, or locally significant Site/Pod identifer to represent capacity for all the routes (instances) in the site/pod. There could be many sites/pods connected to the egress router (a.k.a. Edge DC GW)

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

5.1. Service Instance Attached Capacity Site Index

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 Extended Community for every BGP Update message if there is no change to the site-reference identifier or value for the service instances.

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

5.2. 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 Extended Community without attaching any routes.

When an ingress router receives a BGP Update message from Router-X with the Site-Capacity Extended Community (Received-Site-Reference=t) but without specific routes attached, the ingress router performs the following steps:

```
     For (i=0; i<RoutingTableSize; i++)
     {
        If (RoutingTable[i].NextHop == Router-X)
        {
           If (RoutingTable[i].Site-Reference == Received-Site-
           Reference-ID)
           {
              RoutingTable[i].Site-capacity = newly-received-
              site-capacity;
           }
        }

     }
```

The new Site-Capacity value is applied to all routes that are associated with the Site-Reference ID with the NextHop being the Router-X.

When a CPE 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, load index site preference, and network cost. The Plugin function is called Cost Compute Engine throughout this document.

Suppose a destination address for S1: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 (S1:aa08::4450). The Cost Compute Engine can insert a higher weight for the tunnel associated with R1 for the prefix via the tunnel.

6. AppMetaData Propagation Scope

AppMetaData 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 AppMetaData for specific Service IDs.

For each registered EC service, a corresponding filter group can be formed on RR to represent the interested ingress routers that are interested in receiving the corresponding AppMetaData 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

To be added.

9. Security Considerations

To be added.

10. IANA Considerations

Here are new Sub-TLV types requiring IANA registration:

Type = TBD1: Aggregated Load Measurement Index derived from the Weighted combination of bytes/packets sent to/received from the App server.

Type = TBD2: Raw measurements of packets/bytes to/from the App Server address.

Type = TBD3: Capacity value sub-TLV

Type = TBD4: Site preference value sub-TLV

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>.

[RFC8200] s. Deering R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", July 2017

11.2. Informative References

[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.

[RFC5521] P. Mohapatra, E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", April 2009.

[BGP-SDWAN-Port] L. Dunbar, H. Wang, W. Hao, "BGP Extension for SDWAN Overlay Networks", draft-dunbar-idr-bgp-sdwan-overlay-ext-03, work-in-progress, Nov 2018.

[SDWAN-EDGE-Discovery] L. Dunbar, S. Hares, R. Raszuk, K. Majumdar, "BGP UPDATE for SDWAN Edge Discovery", draft-dunbar-idr-sdwan-edge-discovery-00, work-in-progress, July 2020.

[Tunnel-Encap] E. Rosen, et al "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10, Aug 2018.

12. Acknowledgments

Acknowledgements to Sue Hares, Donald Eastlake 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