Network Working Group
Internet Draft
Intended status: Standard
Expires: April 31, 2021
L. Dunbar
Futurewei
K. Majumdar
CommScope

October 31, 2020

BGP NLRI App Meta Data for 5G Edge Computing Service
draft-dunbar-idr-5g-edge-compute-app-meta-data-00

Abstract

This draft describes a new BGP Network Layer Reachability Information (BGP NLRI) Path Attribute, AppMetaData, that can distribute the 5G Edge Computing App running status and environment, so that other routers in the 5G Local Data Network can make intelligent decision on optimized forwarding of flows from UEs. The goal is to improve latency and performance for 5G Edge Computing services.

The extension enables a feature, called soft anchoring, which makes one Edge Computing Server at one specific location to be more preferred than others for the same application to receive packets from a specific source (UE).

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

This document describes a new BGP Network Layer Reachability Information (BGP NLRI) Path Attribute, AppMetaData, that can distribute the 5G Edge Computing App running status and environment, so that other routers in the 5G Local Data Network can make intelligent decision on optimized forwarding of flows from UEs. The goal is to improve latency and performance for 5G Edge Computing services.

1.1. 5G Edge Computing Background

As described in [5G-EC-Metrics], one Application can have multiple Application Servers hosted in different Edge Computing data centers that are close in proximity. Those Edge Computing (mini) data centers are usually very close to, or co-located with, 5G base stations, with the goal to minimize latency and optimize the user experience.

When a UE (User Equipment) initiates application packets using the destination address from a DNS reply or from its own cache, the packets from the UE are carried in a PDU session through 5G Core [5GC] to the 5G UPF-PSA (User Plan Function - PDU Session Anchor). The UPF-PSA decapsulate the 5G GTP outer header and forwards the packets from the UEs to the Ingress router of the Edge Computing (EC) Local Data Network (LDN). The LDN for 5G EC, which is the IP Networks from 5GC perspective, is responsible for forwarding the packets to the intended destinations.

When the UE moves out of coverage of its current gNB (next generation Node B) (gNB1), handover procedures are initiated and the 5G SMF (Session Management Function) also selects a new UPF-PSA. The standard handover procedures described in 3GPP TS 23.501 and TS 23.502 are followed. When the handover process is complete, the UE has a new IP address and the IP point of attachment is to the new UPF-PSA. 5GC may maintain a path from the old UPF to new the UPF for a short period of time for SSC [Session and Service Continuity] mode 3 to make the handover process more seamless.

```
   +--+
   |UE|---\+---------+                 +------------------+
   +--+    |  5G     |    +----------+ |   S1: aa08::4450 |
   +--+    | Site +--++---+         +-----+               |
   |UE|----|  A   |PSA| Ra|         | R1 | S2: aa08::4460 |
   +--+    |      +---+---+         +-----+               |
  +---+    |         |  |           |  |   S3: aa08::4470 |
  |UE1|---/+---------+  |           |  +------------------+
  +---+                 |IP Network |       L-DN1
                        |(3GPP N6)  |
     |                  |           |  +------------------+
     | UE1              |           |  |  S1: aa08::4450  |
     | moves to         |          +-----+                |
     | Site B           |          | R3 | S2: aa08::4460  |
     v                  |          +-----+                |
                        |           |  |  S3: aa08::4470  |
                        |           |  +------------------+
                        |           |      L-DN3
   +--+                 |           |
   |UE|---\+---------+  |           |  +------------------+
   +--+    |  5G     |  |           |  |  S1: aa08::4450  |
   +--+    | Site +--++-+--+        +-----+               |
   |UE|----|  B   |PSA| Rb |        | R2 | S2: aa08::4460 |
   +--+    |      +--++----+        +-----+               |
   +--+    |         |  +-----------+  |  S3: aa08::4470  |
   |UE|---/+---------+                 +------------------+
   +--+                                     L-DN2
```

Figure 1: App Servers in different edge DCs

1.2. Problem#1: ANYCAST in 5G EC Environment

Increasingly, Anycast is used extensively by various application providers and CDNs because ANYCAST makes it possible to dynamically load balance across server locations based on network conditions.

Application Server location selection using Anycast address leverages the proximity information present in the network (routing) layer and eliminates the single point of failure and bottleneck at the DNS resolvers and application layer load balancers. Another benefit of using ANYCAST address is removing the dependency on UEs. Some UEs (or clients) might use their cached IP addresses instead of querying DNS for extended period.

But, having multiple locations of the same ANYCAST address in 5G Edge Computing environment can be problematic because all those edge computing Data Centers can be close in proximity. There might be very little difference in the routing cost to reach the Application Servers in different Edge DCs.

BGP is an integral part in the way IP Anycast usually functions. Within BGP routing there are multiple routes for the same IP address which are pointing to different locations.

This draft describes the BGP UPDATE extension to allow the App Servers Running status and environment to be included in the BGP UPDATE messages, so that other routers can select more optimal ANYCAST location based on the combination of network delay, the App Server load index, the location capacity index and the location preference.

1.3. Problem #2: Unbalanced Anycast Distribution due to UE Mobility

Another problem of using ANYCAST address for multiple Application Servers of the same application in 5G environment is that UEs' frequent moving from one 5G site to another, which can make it difficult to plan where the App Server should be hosted. When one App server is heavily utilized, other App servers of the same address close-by can be very underutilized. Since the condition can be short lived, it is difficult for the application controller to anticipate the move and adjust.

1.4. Problem 3: Application Server Relocation

When an Application Server is added to, moved, or deleted from a 5G Edge Computing Data Center, the routing protocol needs to propagate the changes to 5G PSA or the PSA adjacent routers. After the change, the cost associated with the site [5G-EC-Metrics] might change as well.

Note: for the ease of description, the Edge Application Server and Application Server are used interchangeably throughout this document.

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 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 host 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 external IP network, in front of a set of application servers. From IP network perspective, this whole group of servers are 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 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. It might be co-located with 5G
Base Station and not only host 5G core functions,
but also host frequently used Edge server instances.

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. Usage of App Meta Data for 5G Edge Computing

3.1. Overview

From IP Layer, the Application Servers are identified by their IP (ANYCAST) addresses. The 5G Edge Computing controller or management system is aware of the ANYCAST addresses of the

Applications that need optimized forwarding in 5G EC environment. The 5G Edge Computing controller or management system can configure the ACLs to filter out those applications on the routers adjacent to the 5G PSA and the routers to which the Application Servers are directly attached.

The proposed solution is for the routers, i.e. A-ER, that have direct links to the Application Servers to collect various measurements about the Servers' running status [5G-EC-Metrics] and advertise the metrics to other routers in 5G EC LDN (Local Data Network).

3.2. IP Layer Metrics to Gauge Application Behavior

[5G-EC-Metrics] describes the IP Layer Metrics that can gauge the application servers running status and environment:

- IP-Layer Metric for App Server Load Measurement:
  The Load Measurement to an App Server is a weighted combination of the number of packets/bites to the App Server and the number of packets/bytes from the App Server which are collected by the A-ER to which the App Server is directly attached.
  The A-ER is configured with an ACL that can filter out the packets for the Application Server.
- Capacity Index
  Capacity Index is used to differentiate the running environment of the application server. Some data centers can have hundreds, or thousands, of servers behind an Application Server's App Layer Load Balancer that is reachable from external world. Other data centers can have very small number of servers for the application server. "Capacity Index", which is a numeric number, is used to represent the capacity of the application server in a specific location.
- Site preference index:
  [IPv6-StickyService] describes a scenario that some sites are more preferred for handling an application server than others for flows from a specific UE.

In this document, the term "Application Server Egress Router" [A-ER] is used to describe the last router that an Application Server is attached. For 5G EC environment, the A-ER can be the gateway router to the EC DC where multiple Application servers' instance are hosted.

From IP Layer, an Application Server is identified by its IP (ANYCAST) Address. Those IP addresses are called the Application Server IDs throughout this document.

3.3. To Equalize among Multiple ANYCAST Locations

The main benefit of using ANYCAST is to leverage the network layer information to equalize the traffic among multiple Application Server locations of the same Application, which is identified by its ANYCAST addresses.

For 5G Edge Computing environment, the ingress routers to the LDN needs to be notified of the Load Index and Capacity Index of the App Servers at different EC data centers to make the intelligent decision on where to forward the traffic for the application from UEs.

[5G-EC-Metrics] describes the algorithms that can be used by the routers directly attached to the 5G PSA to compare the cost to reach the App Servers between the Site-i or Site-j:

```
         alpha*(LoadIndex-i*Beta-i)   (1-alpha)*(Delay-i*gamma-i)
Cost=min(--------------- ----------  + ----------------------------)
          (LoadIndex-j * Beta-j)          ( Delay-j *gamm-j)
```

LoadIndex-i: weighted combination of the total bytes (or/and packets) sent to/received from the Application Server at Site-i during a fixed time period.

Beta-i (larger value means higher capacity): capacity index at the site i.

Delay-i: Network latency measurement (RTT) to the A-ER that has the Application Server attached at the site-i.

gamma (larger value means higher preference): Network Preference index for the site-I.

alpha (a value between 0 and 1: Weight for load & site Index. If smaller than 0.5, Network latency has more influence; otherwise, Server load has more influence).

3.4. BGP Protocol Extension to advertise Load & Capacity

Goal of the protocol extension:

- Propagate the Load Measurement Index for the attached App Servers to other routers in the LDN.

- Propagate the Capacity Index &

- Propagate Site Preference Index.

The BGP extension is to add the Load Index Sub-TLV, Capacity Sub-TLV, and the Site Preference Sub-TLV in the NLRI associated with the routes.

3.5. Reason for using BGP Based Solution:

To Be Added

4. The NLRI Path Attribute for App Meta Data

The App Meta Data attribute is an optional transitive BGP Path attribute to carry application specific data, such as running status, capacity and site preference. Will need IANA to assign a value as the type code of the attribute. The attribute is composed of a set of Type-Length-Value (TLV) encodings. Each TLV contains information corresponding to metrics to a specific Application Server. An App Meta Data TLV, is structured as shown in Figure 1:

```
      0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     | AppMetaData Type (2 Octets)   |        Length (2 Octets)      |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                                                               |
     |                            Value                              |
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: App Meta Data TLV Value Field

AppMetaData Type (2 octets): identifies a type of Application related metadata. The field contains values from the IANA Registry "BGP AppMetaData Types". To be added.

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

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

Each sub-TLV consists of three fields: a 1-octet type, a 1-octet or 2-octet length field (depending on the type), and zero or more octets of value. A sub-TLV is structured as shown in Figure 2:

```
+--------------------------------+
| Sub-TLV Type (1 Octet)         |
+--------------------------------+
| Sub-TLV Length (1 or 2 Octets) |
+--------------------------------+
| Sub-TLV Value (Variable)       |
+--------------------------------+
```

Figure 3: App Metadata Sub-TLV Value Field

o Sub-TLV Type (1 octet): each sub-TLV type defines a certain property about the AppMetaData TLV that contains this sub-TLV. The field contains values from the IANA Registry "BGP AppMetaData Attribute Sub-TLVs".

o Sub-TLV Length (1 or 2 octets): the total number of octets of the sub-TLV value field. The Sub-TLV Length field contains 1 octet if the Sub-TLV Type field contains a value in the range from 0-127. The Sub-TLV Length field contains two octets if the Sub-TLV Type field contains a value in the range from 128-255.

o Sub-TLV Value (variable): encodings of the value field depend on the sub-TLV type as enumerated above. The following sub-sections define the encoding in detail.

4.1. Load Measurement sub-TLV format

Two types of Load Measurement Sub-TLVs are specified. One is to carry the measurements of packets/bytes to/from the App Server address, another one is to carry the aggregated cost Index based on weighted combination of the collected measurements.

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 4: Load Measurement Sub-TLV

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Type (TBD3)           |              Length           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   Measurement Period                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Aggregated Load Index to reach the App Server       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: Aggregated Load Index Sub-TLV

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

Type =TBD3: 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 w1+ w2+ w3+ w4 = 1 and 0< wi <1;

Measure Period: BGP Update period or user specified period

4.2. Capacity Index sub-TLV format

The Capacity 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          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   Capacity Index                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Note: "Capacity 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.

4.3. 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 (TBD5)         |               Length          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   Preference Index                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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. Soft Anchoring of an ANYCAST Flow

"Sticky Service" in the 3GPP Edge Computing specification (3GPP TR 23.748) requires a UE to a specific ANYCAST location when the UE moves from one 5G Site to another.

"Soft Anchoring" is referring to forwarding the Application flow from the UE to the a preferred location for the ANYCAST address, when the preferred location is in good condition. But if there is any failure at the preferred location, the Application flow from the UE need to be forwarded to another location that host the same application.

This section describes a solution that can softly anchor an application flow from a UE to a preferred location.

Lets' assume one application "App.net" is instantiated on four servers that are attached to four different routers R1, R2, R3, and R4 respectively. It is desired for packets to the "App.net" from UE-1 to stick with one server, say the App Server attached to R1, even when the UE moves from one 5G site to another. When there is failure at R1 or the Application Server attached to R1, the packets of the flow "App.net" from UE-1 need to be forwarded to the Application Server attached to R2, R3, or R4.

We call this kind of sticky service "Soft Anchoring", meaning that anchoring to the site of R1 is preferred, but other sites can be chosen when the preferred site encounters failure.

Here is details of this solution:

- Assign a group of ANYCAST addresses to one application. For example, "App.net" is assigned with 4 ANYCAST addresses, L1, L2, L3, and L4. L1/L2/L3/L4 represents the location preferred ANYCAST addresses.
- For the App.net Server attached to a router, the router has four Stub links to the same Server, L1, L2, L3, and L4 respectively. The cost to L1, L2, L3 and L4 is assigned differently for different routers. For example,

  o When attached to R1, the L1 has the lowest cost, say 10, when attached to R2, R3, and R4, the L1 can have higher cost, say 30.
  o ANYCAST L2 has the lowest cost when attached to R2, higher cost when attached to R1, R3, R4 respectively.
  o ANYCAST L3 has the lowest cost when attached to R3, higher cost when attached to R1, R2, R4 respectively, and
  o ANYCAST L4 has the lowest cost when attached to R4, higher cost when attached to R1, R2, R3 respectively

- When a UE queries for the "App.net" for the first time, the DNS replies the location preferred ANYCAST address, say L1, based on where the query is initiated.
- When the UE moves from one 5G site-A to Site-B, UE continues sending packets of the "App.net" to ANYCAST address L1. The routers will continue sending packets to R1 because the total cost for the App.net instance for ANYCAST L1 is lowest at R1. If any failure occurs making R1 not reachable, the packets of the "App.net" from UE-1 will be sent to R2, R3, or R4 (depending on the total cost to reach each of them).

If the Application Server supports the HTTP redirect, more optimal forwarding can be achieved.

- When a UE queries for the "App.net" for the first time, the global DNS replies the ANYCAST address G1, which has the same cost regardless where the Application Servers are attached.
- When the UE initiates the communication to G1, the packets from the UE will be sent to the Application Server that has the lowest cost, say the Server attached to R1. The Application server is instructed with HTTPs Redirect to respond back a location specific URL, say App.net-Loc1. The client on the UE will query the DNS for App.net-Loc1 and get the response of ANYCAST L1. The subsequent packets from the UE-1 for App.net are sent to L1.

6. Manageability Considerations

To be added.

7. Security Considerations

To be added.

8. IANA Considerations

To be added.

9. References

9.1. 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

9.2. 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-StickyService] 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.

10. Acknowledgments

Acknowledgements to 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
CommScope
350 W Java Drive, Sunnyvale, CA 94089
Email: kausik.majumdar@commscope.com