IGP Extension for 5G Edge Computing Service
draft-dunbar-lsr-5g-edge-compute-05

Abstract

   This draft describes using additional site capacity and
   preference related metrics to influence the SPF and using
   Flexible Algorithms to indicate the topologies those
   metrics are applied. The purpose is to differentiate
   multiple paths with similar routing distance to one
   destination in 5G Local Data Network (LDN)to achieve
   optimal performance.

   The list of Internet-Draft Shadow Directories can be
   accessed at http://www.ietf.org/shadow.html

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

Copyright Notice

Table of Contents

1. Introduction

In 5G Edge Computing (EC) environment, it is common for an
application that needs low latency to be instantiated on
multiple servers close in proximity to UEs (User
Equipment). Those applications instances can be behind one
or multiple application-layer load balancers. When they
have relatively short flows that can go to any instance,
having the cluster of them at different locations share the
same IP address can minimize the impact to DNS and achieve
optimal forwarding that leverages network conditions. E.g.,
Kubernetes for data center networking uses one single
Virtual IP address for a cluster of instances of
microservices so that the network can forward via multiple
paths towards one single destination.

This draft describes using additional site costs to
influence the shortest path computation for a specific set
of prefixes. The site costs can be a group of metrics, or
one aggregated cost computed based on a configured
algorithm. As there are a small number of egress routers
having those prefixes (or destinations) that need to
incorporate site costs in SPF computation, Flexible

Algorithms [LSR-FlexAlgo] is used to indicate the need for
the site costs to be considered for the specific
topologies. Flexible algorithms provide mechanisms for
topologies to use different IGP path algorithms.

1.1. Unbalanced Distribution due to UE Mobility

UEs' frequent moving from one 5G site to another can make
it difficult to plan where the App Servers should be
hosted. When a group of App servers at one location, which
can be behind an application-layer load balancer, are
heavily utilized, the instances for the same application at
another location can be under-utilized. The difference in
the routing distance to reach multiple sites where the
application instances are instantiated might be relatively
small in 5G LDN environment. The site capacity and
preferences can be more significant than the routing
distance from the application's latency and performance
perspective.

Since the condition can change in days or weeks, it is
difficult for the application controller to anticipate the
moving and adjusting relocation of application instances.

1.2. ANYCAST in 5G EC Environment

ANYCAST is assigning the same IP address for multiple
instances at different locations. Using ANYCAST can
eliminate the single point of failure and bottleneck at
load balancers or DNS. Another benefit is removing the
dependency on how UEs resolve IP addresses for their
applications. Some UEs (or clients) might use stale cached
IP addresses for an extended period.

But having the same IP address at multiple locations of the
5G Edge Computing environment can be problematic because
all those locations can be close in proximity. There might
be a tiny difference in the routing distance to reach an
application instance attached to a different edge router.

1.3. Scope of the Document

The draft is for scenarios where applications or micro
services are instantiated at multiple locations behind one
or multiple application layer load balancers. They have
relative short flows that can go to any instances.

Under this scenario, multiple instances for the same type
of services can be assigned with the same IP address, so
that network condition can be utilized to achieve optimal
forwarding.

From IP network perspective, application layer load
balancers and app servers all appear as IP addresses.
Throughout this document, the term "app server" can
represent the load balancer in front of a cluster of app
server instances, app server instances, or app server.

Note: for the ease of description, the EC (Edge Computing)
server, Application server, App server are used
interchangeably throughout this document.

2. Conventions used in this document

A-ER: Egress Edge 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 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 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. 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

   LDN: 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. Solution Overview

   The proposed solution is for the egress edge router (A-ER)
   with the application instances directly attached to

     - advertise the aggregated site cost via IP prefix
       reachability TLV associated with the (anycast) prefix.
       [note: the aggregated cost in this version of the
       draft is one value. Some deployment scenarios could
       have a set of values as the site cost.]

- use a Flag in the Flexible Algorithm TLV to indicate
  that the aggregated site cost needs to influence the
  SPF to reach the Prefix.

The aggregated site cost associated with a prefix (i.e.,
ANYCAST prefix) is computed based on the Capacity Index,
the Preference Index, and other constraints by a consistent
algorithm across all A-ERs. The capacity and preference
indexes are configured to the egress routers to which the
prefix is attached.

The solution assumes that the 5G EC controller or
management system is aware of the ANYCAST addresses that
need optimized forwarding. Only the addresses that match
with the ACLs configured by the 5G EC controller will have
their aggregated site cost advertised.

4. New Flags added to FAD Flags Sub-TLV

A New flag is added to indicate a constrained SPF compute
method is needed for the prefix.

Flags:

```
  0 1 2 3 4 5 6 7...
  +-+-+-+-+-+-+-+-+...
  |M|P| | ...
  +-+-+-+-+-+-+-+-+...
```

P-flag: Site-Cost Metrics is included in deriving
Constrained IGP path to the prefix.

5. Minimum Interval for Aggregated Site Cost Advertisement

The aggregated site cost associated with a prefix (i.e.,
ANYCAST prefix) can be a value configured on the router to
which the prefix is attached. The aggregated site cost can
be computed based on an algorithm configured on router for
specific prefixes. The detailed algorithm of computing the
aggregated site cost is out of the scope of document.

As the cost change can impact the path computation, there
is a Minimum Interval for Metrics Change Advertisement

which is configured on the routers to avoid route
oscillations. Default is 30s.

The aggregated site cost change rate is comparable with the
rate of adding or removing application instances at
locations to adapt to the workload distribution changes.
The rate of change could be in days or weeks. On rare
occasions, there might need rate changes in hours.

6. Aggregate Site Cost Advertisement in OSPF

- IPv4: OSPFv2
  A new Aggregated Cost Sub-TLV needs to be added to
  OSPFv2 Extended Prefix TLV [RFC7684]

- IPv6: OSPFv3
  A new sub-TLV can be appended to the E-Intra-Area-
  Prefix-LSA, E-Inter-Area-Prefix-LSA, E-AS-External-
  LSA, and E-Type-7-LSA [RFC8362] to carry the
  Aggregated Cost.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|AggCostSubTLV                  | Length                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        AggCost Attribute-1 to the Prefix                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                                                              ~
|        AggCost Attribute-i to the Prefix                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
            Figure 1: Aggregated cost Advertisement in IS-IS
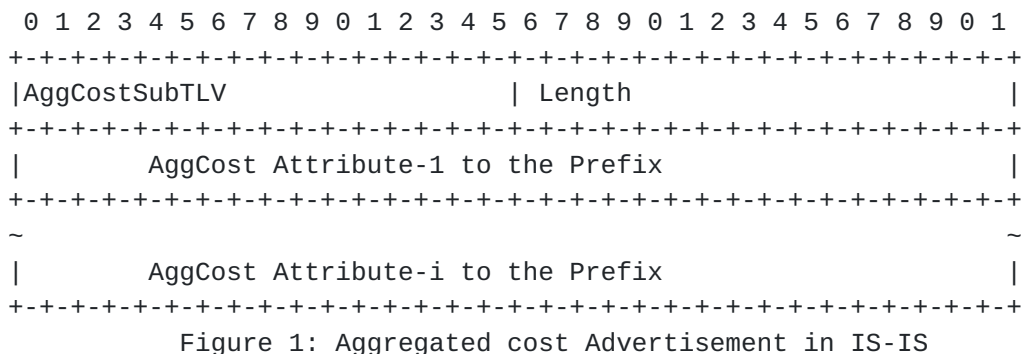
7. "Site-Cost" Advertisement in IS-IS

Aggregated Cost can be appended as subTLV to the Extended
IP Reachability TLV 135 for IPv4 [RFC5305] and 236 for IPv6
[RFC5308].

For Multi-Topology with non-zero IDs, the Aggregated Cost
SubTLV can be carried by Multi-topology TLV 235 for IPv4
and 237 for IPv6 [RFC5120].

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|AggCostSubTLV  | Length        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          AggCost Attribute-1 to the App Server               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                                                               ~
|          AggCost Attribute-i to the Prefix                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Aggregated cost Advertisement in IS-IS

8. Alternative method for Distributing Aggregated Cost

Section 6 and Section 7 demonstrate different ways for
OSPFv2, OSPFv3, and ISIS to propagate the aggregated cost.
It would be better if the aggregated cost could be
advertised the same way, regardless of OSPFv2, OSPFv3, or
ISIS.

Draft [draft-wang-lsr-stub-link-attributes] introduces the
Stub-Link TLV for OSPFv2/v3 and ISIS protocol respectively.
Considering the interfaces on an edge router that connects
to the EC servers are normally configured as passive
interfaces, these IP-layer App-metrics can also be
advertised as the attributes of the passive/stub link. The
associated prefixes can then be advertised in the "Stub-
Link TLV" that is defined in [draft-wang-lsr-stub-link-
attributes]. All the associated prefixes share the same
characteristic of the link. Other link related sub-TLVs
defined in [RFC8920] can also be attached and applied to
the calculation of path to the associated prefixes."

Section 6 for the advertisement of AppMetaData Metric can
also utilize the Stub-Link TLV that defined in [draft-wang-
lsr-stub-link-attributes]

9. Manageability Considerations

To be added.

10. Security Considerations

To be added.

11. IANA Considerations

The following Sub-TLV types need to be added by IANA to
FlexAlgo.

- AggCostSubTLV Type for ISIS, OSPF (TBD1): IPv4 or IPv6

P-flag added to FAD Flags Sub-TLV to indicate that the
Site-Cost Metrics is included in deriving Constrained IGP
path to the prefix.

12. References

12.1. Normative References

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

[RFC2328] J. Moy, "OSPF Version 2", RFC 2328, April 1998.

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

[RFC7684] P. Psenak, et al, "OSPFv2 Prefix/Link Attribute
          Advertisement", RFC 7684, Nov. 2015.

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

[RFC8326] A. Lindem, et al, "OSPFv3 Link State
          advertisement (LSA0 Extensibility", RFC 8362,
          April 2018.

[RFC9012] E. Rosen, et al "The BGP Tunnel Encapsulation
          Attribute", April 2021.

12.2. 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-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.

   [BGP-5G-AppMetaData] L. Dunbar, K. Majumdar, H. Wang, "BGP
             App Metadata for 5G Edge Computing Service",
             draft-dunbar-idr-5g-edge-compute-app-meta-data-
             03, work-in-progress, Sept 2020.

   [LSR-Flex-Algo] P. Psenak, et al, "IGP Flexible Algorithm",
             draft-ietf-lsr-flex-algo-17, July 2021.

   [LSR-Flex-Algo-BW] S. Hegde, et al, "Flexible Algorithms:
             Bandwidth, Delay, Metrics and Constraints",
             draft-ietf-lsr-flex-algo-bw-con-01, July 2021.

   [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.

13. Appendix A: 5G Edge Computing Background

The network connecting the 5G EC servers with the 5G Base stations consists of a small number of dedicated routers that form the 5G Local Data Network (LDN) to enhance the performance of the EC services.

When a User Equipment (UE) initiates application packets using the destination address from a DNS reply or its 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 decapsulates the 5G GTP outer header, performs NAT sometimes, before handing the packets from the UEs to the adjacent router, also known as the ingress router to the EC LDN, which 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), the handover procedure is initiated, which includes the 5G SMF (Session Management Function) selecting a new UPF-PSA [3GPP TS 23.501 and TS 23.502]. When the handover process is complete, the IP point of attachment is to the new UPF-PSA. The UE's IP address stays the same unless moving to different operator domain. 5GC may maintain a path from the old UPF to the new UPF for a short time for SSC [Session and Service Continuity] mode 3 to make the handover process more seamless.
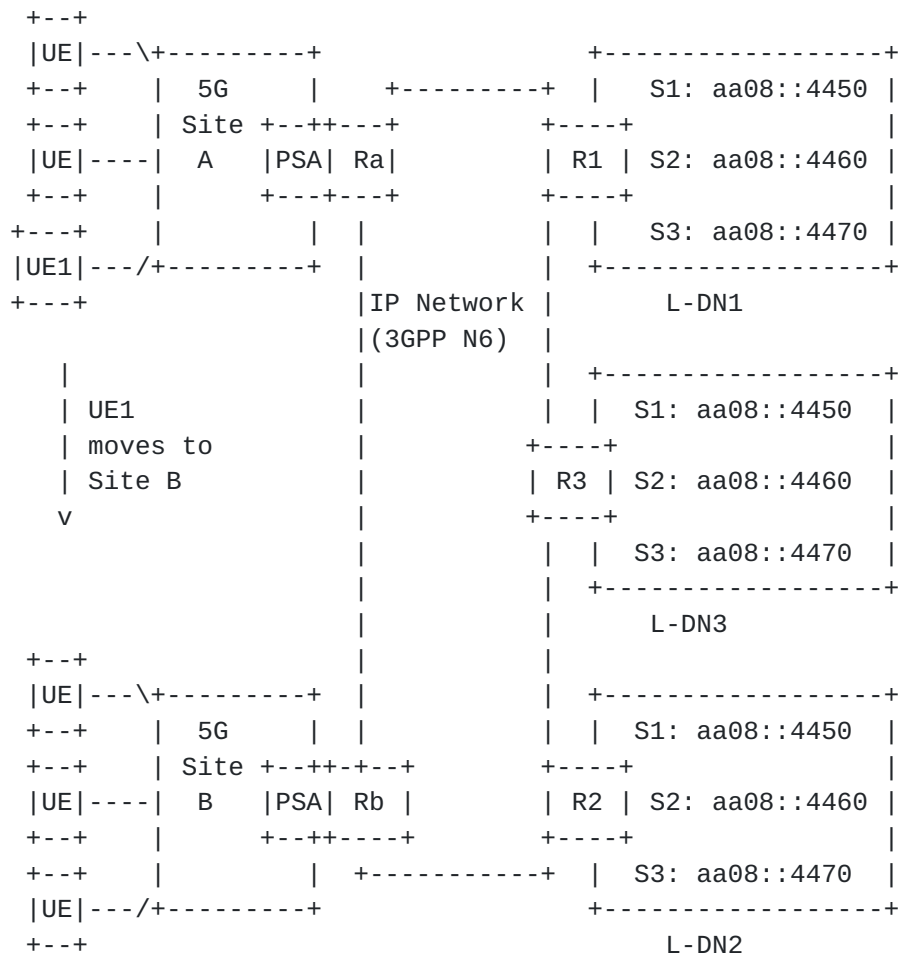
```
   +--+
  |UE|---\+---------+                    +-----------------+
  +--+    |  5G     |    +---------+  |   S1: aa08::4450 |
  +--+    | Site +--++---+        +----+                  |
  |UE|----|  A   |PSA| Ra|        | R1 | S2: aa08::4460 |
  +--+    |      +---+---+        +----+                  |
 +---+    |         | |           |  |   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 10: App Servers in different edge DCs
```

13.1. Metrics to change traffic flow patterns

   When UEs pattern changes, the Application controller can
   instantiate more instances at certain locations to
   accommodate higher demand.

   However, network layer can offer a simpler solution. By
   adjusting the site cost for the prefix at specific egress
   routers, IGP distribution of those site cost plus the flex
   algorithm can increase (or decrease) flows for the specific
   prefixes towards the certain locations.

     - Capacity Index:
       a numeric number, configured on all A-ERs in the
       domain consistently, is used to represent the capacity
       of an EC server attached to an A-ER. The IP addresses
       exposed to the A-ER can be the App Layer Load
       balancers that have many instances attached. At other
       sites, the IP address exposed is the server itself.
     - Site preference index:
       Is used to describe some sites are more preferred than
       others. For example, a site with less leasing cost has
       a higher preference value. Note: the preference value
       is configured on all A-ERs in the domain consistently
       by the Domain Controller.

 13.2. Reason for using IGP Based Solution

    Here are some benefits of using IGP to propagate the IP
    Layer App-Metrics:
    - Intermediate routers can utilize the aggregated cost to
      reach the EC Servers attached to different egress edge
      nodes, especially:
        - The path to the optimal egress edge node can be
           more accurate or shorter.
        - Convergence is shorter when there is any failure
           along the way towards the optimal ANYCAST server.
        - When there is any failure at the intended ANYCAST
           server, all the packets in transit can be optimally
           forwarded to another App Server attached to a
           different egress edge router.
    - Doesn't need the ingress nodes to establish tunnels with
      egress edge nodes.

    There are limitations of using IGP too, such as:

    - The IGP approach might not suit well to 5G EC LDN
      operated by multiple ISPs.
      For LDN operated by multiple IPSs, BGP should be used.
      [BGP-5G-AppMetaData] describes the BGP UPDATE message to
      propagate IP Layer App-Metrics crossing multiple ISPs.

 13.3. Flow Affinity to an ANYCAST server

   When multiple servers with the same IP address (ANYCAST)
   are attached to different A-ERs, Flow Affinity means
   routers sending the packets of the same flow to the same A-
   ER even if the cost towards the A-ER is no longer optimal.

   Many commercial routers support some forms of flow affinity
   to ensure packets belonging to one flow be forwarded along
   the same path.

   Editor's note: for IPv6 traffic, Flow Affinity can be
   achieved by routers forwarding the packets with the same
   Flow Label extracted from the IPv6 Header along the same
   path.

[14]. Acknowledgments

   Acknowledgements to Peter Psenak, Les Ginsberg, Robert
   Raszuk, Acee Lindem, Shraddha Hegde, Tony Li, Gyan Mishra,
   Jeff Tantsura, and Donald Eastlake for their review and
   suggestions.

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

Authors' Addresses

   Linda Dunbar
   Futurewei
   Email: ldunbar@futurewei.com

   Huaimo Chen
   Futurewei
   Email: huaimo.chen@futurewei.com

   Aijun Wang
   China Telecom
   Email: wangaj3@chinatelecom.cn