Network Working Group
Internet Draft
Intended status: Standard
Expires: June 16, 2022

L. Dunbar
H. Chen
Futurewei
Aijun Wang
China Telecom

January 7, 2022

IGP Extension for 5G Edge Computing Service
draft-dunbar-lsr-5g-edge-compute-03

Abstract

Routers in 5G Local Data Network (LDN) can use additional site-costs, preference, and other application related metrics in addition to the network routing distance to compute constraint-based SPF within the 5G LDN to enhance performance for selected services. This draft describes using application server related metrics to influence the SPF and Flexible Algorithms to indicate the constraints.

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.

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). When those multiple server instances share one IP address (ANYCAST), the transient network and load conditions can be incorporated in selecting an optimal path among server instances for UEs.

Flexible algorithms provide mechanisms for topologies to use different IGP path algorithms. This draft describes using Flexible Algorithms [LSR-FlexAlgo] to indicate the desired constrained SPF behavior for a subset of prefixes, in addition to the encodings for advertising the IP Layer App related metrics that can impact application servers' performance.

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 one App server is heavily utilized, other App servers of the same address close by can be under-utilized.

The difference in the routing distance to reach multiple Application Servers might be relatively small. The traffic load at the router where the App Server is attached and the site capacity, when combined, can be more significant than the routing distance from the latency and performance perspective.

Since the condition can be short-lived, it is difficult for the application controller to anticipate the moving and adjusting.

1.2. ANYCAST in 5G EC Environment

ANYCAST is assigning the same IP address for multiple servers in 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 in multiple locations of the 5G Edge Computing environment can be problematic because all those locations can be close in proximity. There might be a very small difference in the routing distance to reach an Application Server attached to a different edge router.

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 EC Servers directly attached to

- advertise the "Site-Cost" via IP prefix reachability TLV associated with the (anycast) prefix.

- use a Flag in the Flexible Algorithm TLV to indicate that "site-cost" needs to be included for the constrained SPF to reach the Prefix

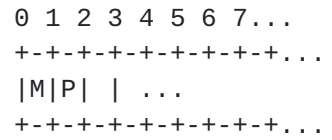
The "Site-Cost" associated with an EC server (i.e., ANYCAST prefix) is computed based on the IP layer App-related metrics [Section 12.1], such as Load Measurement, the Capacity Index, the Preference Index, and other constraints by a consistent algorithm across all A-ERs.

The solution assumes that the 5G EC controller or management system is aware of the EC ANYCAST addresses that need optimized forwarding. To minimize the processing, only the addresses that match with the ACLs configured by the 5G EC controller will have their Site-Cost collected and 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:



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

5. "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.

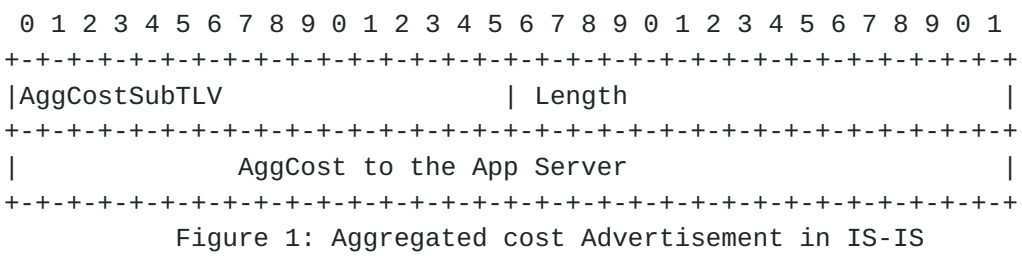


Figure 1: Aggregated cost Advertisement in IS-IS

6. "Site-Cost" Advertisement in IS-IS

Aggregated Cost appended to the IP Reachability TLV: 128, 130, or 135.

Figure 2: Aggregated cost Advertisement in IS-IS

7. 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]

8. Manageability Considerations

To be added.

9. Security Considerations

To be added.

10. 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.

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.

[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.

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-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.

12. Appendix: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.

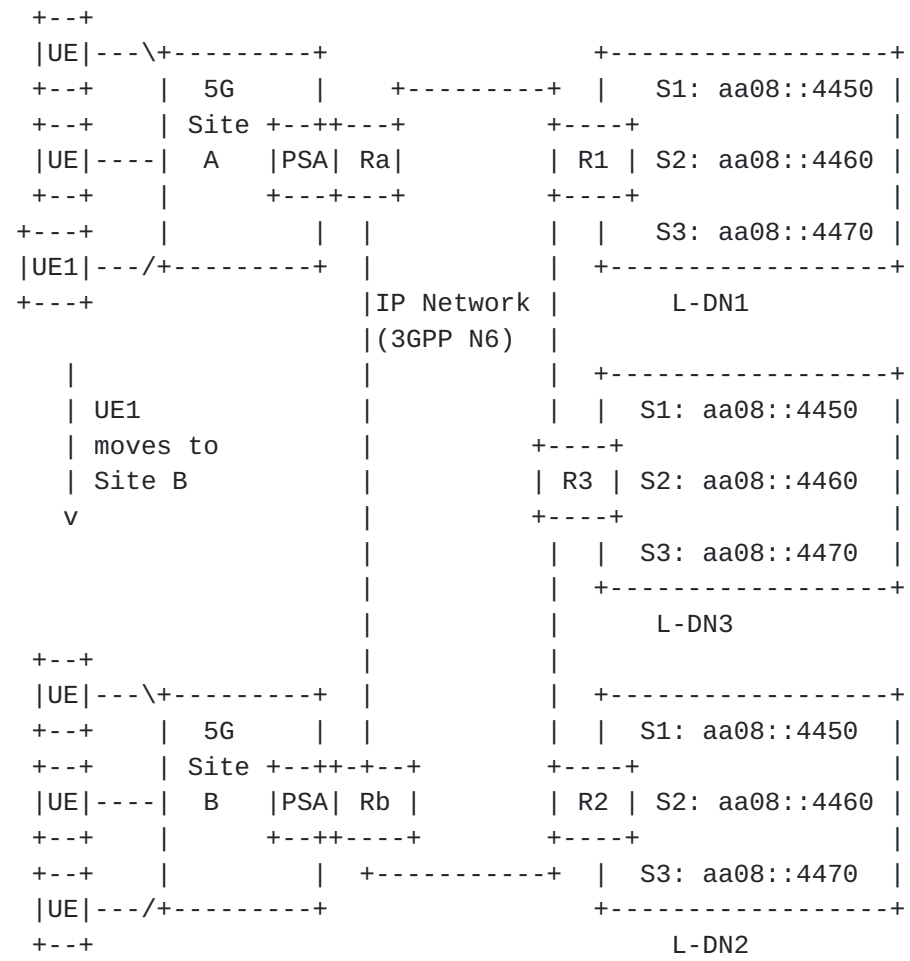


Figure 10: App Servers in different edge DCs

13. 5G EC LDN Characteristics for the Constraint SPF

13.1. IP Layer Metrics to Gauge EC Server Running Status

Most applications do not expose their internal logic to the network. Their communications are generally encrypted. Most of them do not even respond to PING or ICMP messages initiated by routers.

Here are some IP Layer App related Metrics that can gauge the servers running status and environment:

- 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.

- Load Measurement for gauging the load of the attached prefix (i.e., EC Server):
  The Load Measurement for an EC Server is a weighted combination of the number of packets/bytes to the EC server (i.e., its IP address) and the number of packets/bytes from the EC server. The Load Measurement are collected by the A-ER that has the EC Server directly attached.

  An A-ER only collects those measurement for the prefixes instructed by the Domain Controller.

For ease of description, those metrics with more to be added later are called IP Layer App Metrics (or Site-Cost) throughout the document.

13.2. App Metrics Constrained Shortest Path First

The main benefit of using ANYCAST is to leverage the network layer information to balance the traffic among multiple locations of one application server.

For the 5G EC environment, the routers in the LDN need to take consideration of various measurements of the EC servers attached to each A-ER in addition to TE metrics to compute ECMP paths to the servers.

Here is one algorithm that computes the cost to reach the App Servers attached to Site-i relative to another site, say Site-j. When the reference site, Site-j, is plugged in the formula, the cost is 1. So, if the formula returns a value less than 1, the cost to reach Site-i is less than reaching the reference site (Site-j).

$$Cost\text{-}i = \left(w * \frac{CP\text{-}j * Load\text{-}i}{CP\text{-}i * Load\text{-}j}\right) + (1-w) * \left(\frac{Pref\text{-}j * Network\text{-}Delay\text{-}i}{Pref\text{-}i * Network\text{-}Delay\text{-}j}\right)$$

Load-i: Load Index at Site-i, it is the weighted combination of the total packets or/and bytes sent to and received from the Application Server at Site-i during a fixed time period.

CP-i: capacity index at site i, a higher value means higher capacity.

Network Delay-i: Network latency measurement (RTT) to the A-ER that has the Application Server attached at the site-i.
Noted: Ingress nodes can easily measure RTT to all the egress edge nodes by existing IPPM metrics. But it is not so easy for ingress nodes to measure RTT to all the App Servers. Therefore, "Network-Delay-i", a.k.a. Network latency measurement (RTT), is between the Ingress and egress edge nodes. The cost for the egress edge nodes to reach to their attached servers is embedded in the "capacity index".

Pref-i: Preference index for site-i, a higher value means higher preference. Preference can be derived from the total path cost to reach the A-ER [RFC5305], as calculated below: 1/(total-path-cost).

w: Weight for load and site information, which is a value between 0 and 1. If smaller than 0.5, Network latency and the site Preference have more influence; otherwise, Server load and its capacity have more influence.

13.3. 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.

13.4. 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, 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