DMM Working Group
Internet-Draft
Intended status: Informational
Expires: 30 August 2026

T. Kamata
J. Horn
Cisco Systems, Inc.
L. Jalil
Verizon
W. Cheng
China Mobile
M. Kohno
Keio University

26 February 2026

Architecture Discussion on SRv6 Mobile User plane
draft-ietf-dmm-srv6mob-arch-03

Abstract

This document describes the solution approach and its architectural benefits of transforming mobile session information into routing information, leveraging segment routing capabilities, and operating within the IP routing paradigm.

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/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 30 August 2026.

Copyright Notice

Copyright (c) 2026 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction

The existing mobile user plane is currently defined as an overlay tunnel session to a mobile anchor point (UPF: User Plane Function in 5G context).

While this approach may be suited for use cases requiring frequent mobile handover and functionality tied to session initiation/termination, it proves challenging to cost-effectively and scalably address the high traffic volumes of the 5G/Beyond 5G era and the increasingly distributed data and computing demands in the future.

The requirements for wireless systems are becoming more diverse, and there are cases , such as some IoT and FWA (Fixed Wireless Access) systems, where the frequent mobile handover is not necessarily mandatory.

This document describes the solution approach and its architectural benefits of transforming mobile session information into routing information, leveraging segment routing capabilities, and operating within the IP routing paradigm.

And this document clarifies the motivation for the MUP initiatives : [RFC9433][I-D.mhkk-dmm-mup-architecture]

2. Problem Definition

The current tunnel session based mobile user plane has the following limitations and is getting hard to support new application requirements.

* Less suited for any-to-any communication

* Less suited for edge/distributed computing

* Less suited for fixed and mobile convergence (FMC) / wireless-wireline convergence (WWC)

* Limited control of the underlay path

Mobile session information is a function of M,N (GTP-U session start point and end point), whereas routing information is a function of N (destination). Therefore, for any-to-any communications, session based paradigm yields O(N^2), whereas IP routing paradigm yields O(N).

Edge/distributed computing can be seen as a subset of any-to-any communication. IP Routing paradigm naturally supports ubiquitous computing.

As for FMC/WWC, there is currently a coordinated standardization effort between 3GPP WWC [TS.23316] and BBF [BBF407]. However, the idea is to anchor even wireline traffic in the mobile packet core, which compromises simplicity and scalability.

In addition, the anchor point that terminates tunnel sessions becomes a scaling bottleneck.

The IP routing paradigm naturally removes these tunnel session based restrictions. Segment Routing enables fast protection, policy, multi-tenancy, and provide reliability and SLA differentiation.

3. SRv6 MUP and the 5G/Beyond 5G use cases

This section describes the advantages of applying the SRv6 Mobile User Plane approach for 5G/Beyond 5G use cases. These advantage comes from the fact that it transforms mobile session information into routing information, leverages Segment Routing, and operates within the IP Routing Paradigm. Another advantage, not mentioned here, is the ability to minimize overhead through SRv6 SID Compression.

In particular, the adoption of the SRv6 uSID (NEXT-C-SID) flavor [RFC9800] significantly reduces encapsulation overhead and enhances packet efficiency. Since uSID allows multiple segments to be compressed into a single IPv6 Destination Address (DA), it optimizes forwarding performance and minimizes the MTU impact even when long Segment Lists are required. This efficiency is a critical enabler for the use cases described below.

3.1. Network Slicing

Network slicing enables network segmentation, isolation, and SLA differentiation such as latency and availability. End-to-end slicing will be achieved by mapping and coordinating IP network slicing, RAN and mobile packet core slicing.

But existing mobile user plane which is overlay tunnel does not have underlying IP network awareness, which could lead to the inability in meeting SLAs. Removing the tunnel and treating it with a IP routing paradigm simplifies the problem.

Segment Routing has a comprehensive set of slice engineering technologies. How to build network slicing using the Segment Routing technology is described in [I-D.ali-teas-spring-ns-building-blocks].

Moreover, the stateless slice identifier encoding [I-D.filsfils-spring-srv6-stateless-slice-id] can be applicable to enable per-slice forwarding policy using the IPv6 header.

The use of uSID facilitates efficient path steering for slice-specific policies. By compressing multiple service and topological SIDs into a compact header, it reduces the risk of fragmentation and MTU-related issues associated with deep SID stacks. This supports scalable deployment of strict SLAs—such as guaranteed latency and path isolation—across the transport network without compromising packet throughput.

3.2. Edge Computing

Edge computing, where the computing workloads and datastores are placed closer to users, is recognized as one of the key pillars to meet demanding requirements of 5G/Beyond 5G era, with regard to low latency, bandwidth efficiency, data locality and privacy.

Edge computing is more important than ever. This is because no matter how much New Radio improves access speeds, it won’t improve throughput, latency and user experiences because they are largely bound to end-to-end round trip delay.

Even with existing mobile architectures, it is possible to place UPFs in a multi-tier, or to distribute UPFs, to achieve Edge Computing. [TS.23548] and [ETSI-MEC] describes how to properly select the UPF of adequate proximity. However, complicated and signaling-heavy mechanisms are required to branch traffic or properly use different UPFs. Also, if the UPF is distributed, seamless handover has to be compromised.

When it comes to IP routing paradigum, ubiquitous computing is innately supported. uSID is particularly beneficial here, as edge computing frequently involves a high volume of small packets, such as those generated by IoT devices. Traditional SRv6 headers with long Segment Lists can impose a disproportionately high overhead on these small payloads. uSID minimizes this overhead, thereby maximizing effective throughput and ensuring efficient utilization of bandwidth-constrained links in distributed edge environments.

3.3. URLLC (Ultra-Reliable Low-Latency Communication) support

3GPP [TR.23725] investigates the key issues for meeting the URLLC requirements on latency, jitter and reliability in the 5G System. The solutions provided in the document are focused at improving the overlay protocol (GTP-U) and limits to provide a few hints into how to map such tight-SLA into the transport network. These hints are based on static configuration or static mapping for steering the overlay packet into the right transport SLA. Such solutions do not scale and hinder network economics.

Another issue that deserves special mention is the ultra-reliability issue. In order to support ultra-reliability with the tunnel session paradigm, redundant user planes paths based on dual connectivity has been proposed. The proposal has two main options.

* Dual Connectivity based end-to-end Redundant User Plane Path

* Support of redundant transmission on N3/N9 interfaces

In the case of the former, UE and hosts have RHF(Redundancy Handling Function). In sending, RFH is to replicate the traffic onto two GTP-U tunnels, and in receiving, RHF is to merge the traffic.

In the case of the latter, traffic are to be replicated and merged with the same sequence for specific QoS flow, which requires further enhancements.

And in either cases, the bigger problem is the lack of a reliable way for the redundant sessions to get through the disjoint path: even with the redundant sessions, if it ends up using the same infrastructure at some points, the redundancy is meaningless.

These issues can be solved more simply without GTP-U tunnel.

Segment routing has some advantages for URLLC traffic. First, traffic can be mapped to a disjoint path or low latency path as needed. Second, Segment routing provides an automated reliability protection mechanism known as TI-LFA, which is a sub-50ms FRR mechanism that provides protection regardless of the topology through the optimal backup path. It can be provisioned slice-aware.

uSID further strengthens URLLC support by minimizing serialization delays and optimizing hardware processing at intermediate nodes. By shortening the overall packet length and reducing the number of required hardware lookups, uSID helps minimize jitter and provides more predictable end-to-end latency.

4. Co-existence and Incremental Deployability

Mobile networks are composed of radio, mobile packet core, and IP networks (access and backbone), with separate standard organizations and communities. Therefore, in the steady state, it is difficult to innovate to a new architecture and requires coexistence and incremental deployment.

[RFC9433] defines the user plane convergence between GTP-U and SRv6, so that it can co-exist with the existing user plane as needed.

[I-D.mhkk-dmm-mup-architecture] defines the MUP architecture for Distributed Mobility Management, which can be plugged into the existing mobile service architecture. In the architecture, mobile session information is transformed to routing information, and operated in L3VPN scheme.

5. Security Considerations

The deployment of this architecture is targeted in an administrative domain, and the functionality is domain specific.

6. IANA Considerations

This memo includes no request to IANA.

7. References

7.1. Normative References

[RFC9433] Matsushima, S., Ed., Filsfils, C., Kohno, M., Camarillo, P., Ed., and D. Voyer, "Segment Routing over IPv6 for the Mobile User Plane", RFC 9433, DOI 10.17487/RFC9433, July 2023, <https://www.rfc-editor.org/rfc/rfc9433>.

[RFC9800] Cheng, W., Ed., Filsfils, C., Li, Z., Decraene, B., and F. Clad, Ed., "Compressed SRv6 Segment List Encoding", RFC 9800, DOI 10.17487/RFC9800, June 2025, <https://www.rfc-editor.org/rfc/rfc9800>.

7.2. Informative References

[BBF407] BBF, "5G Wireless Wireline Convergence Architecture", BBF TR-407 Issue:1, August 2020.

[ETSI-MEC] ETSI, "MEC in 5G Networks", ETSI White Paper No.28, June 2018.

[I-D.ali-teas-spring-ns-building-blocks] Ali, Z., Filsfils, C., Camarillo, P., Voyer, D., Matsushima, S., Rokui, R., Dhamija, and P. Maheshwari, "Building blocks for Network Slice Realization in Segment Routing Network", Work in Progress, Internet-Draft, draft-ali-teas-spring-ns-building-blocks-03, 7 September 2022, <https://datatracker.ietf.org/doc/html/draft-ali-teas-spring-ns-building-blocks-03>.

[I-D.filsfils-spring-srv6-stateless-slice-id] Filsfils, C., Clad, F., Camarillo, P., Raza, S. K., Voyer, D., and R. Rokui, "Stateless and Scalable Network Slice Identification for SRv6", Work in Progress, Internet-Draft, draft-filsfils-spring-srv6-stateless-slice-id-10, 22 July 2024, <https://datatracker.ietf.org/doc/html/draft-filsfils-spring-srv6-stateless-slice-id-10>.

[I-D.mhkk-dmm-mup-architecture] Matsushima, S., Horiba, K., Kawakami, Y., Murakami, T., Patel, K., and J. Horn, "Mobile User Plane Architecture for Distributed Mobility Management", Work in Progress, Internet-Draft, draft-mhkk-dmm-mup-architecture-02, 2 March 2025, <https://datatracker.ietf.org/doc/html/draft-mhkk-dmm-mup-architecture-02>.

[TR.23725] 3GPP, "Study on enhancement of Ultra-Reliable Low-Latency Communication (URLLC) support in the 5G Core network (5GC)", 3GPP TR 23.725 16.2.0, June 2019.

[TS.23316] 3GPP, "Wireless and wireline convergence access support for the 5G System (5GS)", 3GPP TS 23.316 16.7.0, September 2021.

[TS.23548] 3GPP, "5G system Enhacements for Edge Computing", 3GPP TS 23.548 17.0.0, September 2021.

Acknowledgements

Authors would like to thank Satoru Matsushima, Shunsuke Homma, Yuji Tochio and Jeffrey Zhang, for their insights and comments.

Contributors

Francois Clad
Cisco Systems, Inc.
France
Email: fclad.ietf@gmail.com

Pablo Camarillo Garvia
Cisco Systems, Inc.
Spain
Email: pcamaril@cisco.com

Zafar Ali
Cisco Systems, Inc.
United States of America
Email: zali@cisco.com

Authors’ Addresses

Teppei Kamata
Cisco Systems, Inc.
Japan
Email: tkamata@cisco.com

Jakub Horn
Cisco Systems, Inc.
Czech Republic
Email: jakuhorn@cisco.com

Luay Jalil
Verizon
United States of America
Email: luay.jalil@verizon.com

Weiqiang Cheng
China Mobile
China
Email: chengweiqiang@chinamobile.com

Miya Kohno
Keio University
Japan
Email: miya_kohno@keio.jp