Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 3 September 2026

Yixin Lin, Ed.
Chenchen Yang, Ed.
Kuan Zhang, Ed.
Xueli An,, Ed.
Shaoyun Wu, Ed.
Huawei Technologies Co., Ltd.
Muhammad Awais Jadoon, Ed.
Sebastian Robitzsch, Ed.
InterDigital Europe Ltd.
2 March 2026

AI Agent Protocols for Multi-modality
draft-hw-protocol-agent-00

Abstract

With the advancement of AI technologies, AI Agent traffic will account for the majority of network traffic, driving an increasing demand for higher quality of multi-modal data transmissions. Current networks lack awareness of the diverse transmission quality requirements for multi-modal data within a single traffic, leading to degraded service quality and inefficient utilization of network resources. This document proposes methods to enable networks (e.g., mobile network, transport network) to recognize the characteristics and the transmission requirement of AI Agent multi-modal data and outlines necessary capabilities and features of the AI Agent Protocols.

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 3 September 2026.

Table of Contents

1. Introduction

The Agent protocols have garnered significant attention across diverse technical domains. Agent protocols are being introduced by open-source and standardization organizations, e.g., A2A protocol by Google [1], Agent protocol proposals in IETF [2] [3] and ETSI [4]. For mobile telecommunications, 3GPP has also expressed substantial interest in incorporating Agent capabilities within the future 6G network. Numerous use cases for Agent-centric scenarios have been documented within the 3GPP SA1 working group [5]. The 3GPP SA2

working group has reached consensus that the 6G network should support Agentic network and Agent communication [6]. Adaptations to End-to-End Agent protocols are needed to better support AI Agent traffic transmission and ensure the quality of such transmission, e.g., based on the characteristics of AI Agent traffic. Additionally, the 3GPP CT working groups are organizing studies on AI Agent protocols for the 6G network. In this context, IETF AI Agent protocols are highly recommended as a reference [7].

1.1. Requirements Language

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.

2. Agent Protocol Gap Analysis

2.1. Multi-modality of AI Agent Traffic

An AI Agent traffic may consist of multiple “Part” with multi-modal data (e.g., text, file, sensor data, structured data, video, audio and image) [3]. Different “Part” of an AI Agent traffic may be delivered with different transmission modes (e.g., Streaming, Push Notification, Burst).

Taking Google A2A protocol as an example [1], Message constitutes a communication unit between client and server, and Artifacts represent concrete outputs of task executions. Both Messages and Artifacts transmit content composed of one or more Parts:

* A Part serves as a container for discrete communication content segments, which MUST contain exactly one content type:text (purely textual), raw (file-based:images, videos, etc.), url (resource locator), or data (structured blobs e.g., JSON).

2.2. Indistinguishable Multi-modal AI Agent Traffic

There are two distinct methods by which an AI Agent client or server, acting as the source, can encapsulate multi-modal data.

a) The multi-modal data of an AI Agent traffic can be encapsulated into multiple lower layers streams (e.g., HTTP streams, MoQ stream, QUIC Stream):

* the underlying network lacks awareness that these streams belong to the same AI Agent traffic. Consequently, while network nodes (e.g., core network user plane function, radio access network base stations) can guarantee the QoS for individual stream during data forwarding, the overall performance (e.g., latency) for the traffic may increase.

* Moreover, in scenarios like multi-Agent systems, it may require the multi-modal data to arrive at different receiving Agent endpoints as simultaneously as possible. This characteristic renders the existing techniques susceptible to introducing significant overall performance (e.g., latency) decrease if the underlying network lacks awareness.

b) The multi-modal data of an AI Agent traffic can be encapsulated within a single lower layer stream (e.g., HTTP stream, MoQ stream, QUIC Stream), this stream exhibits inherent multi-modal characteristics:

* This manifests as significant variance in Part sizes (e.g., text snippets vs. video files) and divergent quality of service (QoS) criticality across Parts, e.g., regarding transmission urgency and latency tolerance thresholds.

* Limitation: Content of different Parts are encapsulated within a single lower layer stream (e.g., as payload of HTTP protocol, MoQ protocol, QUIC protocol), the underlying network (e.g., mobile network, transport network) remains unaware of the distinct Parts and their individual requirements, such as specific QoS needs.

The Following figure illustrates the interaction between Agent Client and Agent Server via underlying Network:

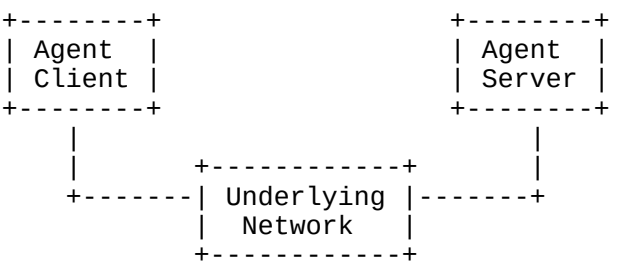

Figure 1. Interaction between Agent Client and Agent Server via underlying Network.

3. AI Agent Traffic Transmission Requirement on Underlying Networks

3.1. Quality of Service (QoS)

Multi-modal AI Agent traffic should be transmitted appropriately by underlying network. Taking 3GPP mobile network as an example of underlying network, QoS framework controls and manages transmission of traffic to ensure performance. QoS framework governs how diverse traffics are treated to meet specific performance requirements (e.g., latency, reliability, data rate, packet loss) [5], [6], [8]. It operates primarily through QoS Flows, which are granular data streams bound with specific QoS parameters and characteristics. These parameters and characteristics indicate the packet forwarding treatment (e.g., scheduling weights, admission thresholds, queue management) applied throughout the network, from the User Equipment (UE) to the mobile network (e.g., core network, radio access network). Multi-modal data of an AI Agent traffic should be transmitted with differentiated QoS to increase the overall performance and resource utilization.

3.2. Awareness of AI Agent Traffic Multi-modality

The underlying network should be aware of the AI Agent traffic characteristics. Different transmission treatment policies may be configured and bounded to the different multi-modal data of an AI Agent traffic based on their different traffic characteristics (e.g., delay budget, priority, error rate). Taking mobile network as an example, the Application Server may influence QoS treatment by notifying service requirements to the mobile network functions. Based on this, the mobile network provisions and configures the corresponding QoS enforcement policies. Subsequently, the mobile network maps incoming packets of application traffic to the appropriate QoS Flows by matching against the enforcement policies and the traffic information (e.g., typically IP 5-tuple). Similarly, the AI Agent client and server need to notify necessary information on its AI Agent traffic to the underlying network for better performance.

4. Agent Protocols for Multi-modal Traffic

4.1. Multiple Streams for Multi-modal Agent Traffic

4.1.1. Mapping Information Notification to Underlying Networks

Using method a) mentioned in 3.2, a source Agent (client or server) transmits multi-modal data of an AI Agent traffic into distinct lower layers streams based on modality and QoS requirements.

* For example, a source Agent transmits multi-modal data of an AI Agent traffic into distinct HTTP streams. Each HTTP stream is characterized by a unique IP 5-tuple and carries data of only one modality.

* As another example, multi-modal data of an AI Agent traffic is transmitted via distinct transport layer streams (e.g., MoQ streams, QUIC streams) or connections (e.g., TCP connections).

The source Agent needs to notify the mapping information to the underlying network to make alignment. Based on the mapping information, the underlying network can identify the correlation of the different lower layer streams, and execute dedicated QoS guarantee accordingly.

The AI Agent protocols in the Agent client and server should have the capabilities and features to perform the actions above, e.g., mapping multi-modal data of Agent traffic to different lower layer streams, notifying the mapping relationship to the underlying network (e.g., via control plane of mobile network), and notifying necessary information to lower layers to enable the latter to encapsulate Agent traffic assistance information. The capabilities and features of the AI Agent protocol should be highlighted and standardized, e.g., in IETF working group(s).

Moreover, it is possible that the Agent protocol layer and/or lower layers need some enhancements, e.g., to encapsulate AI Agent assistance information (e.g., traffic characteristics, multi-modal data types, correlation information).

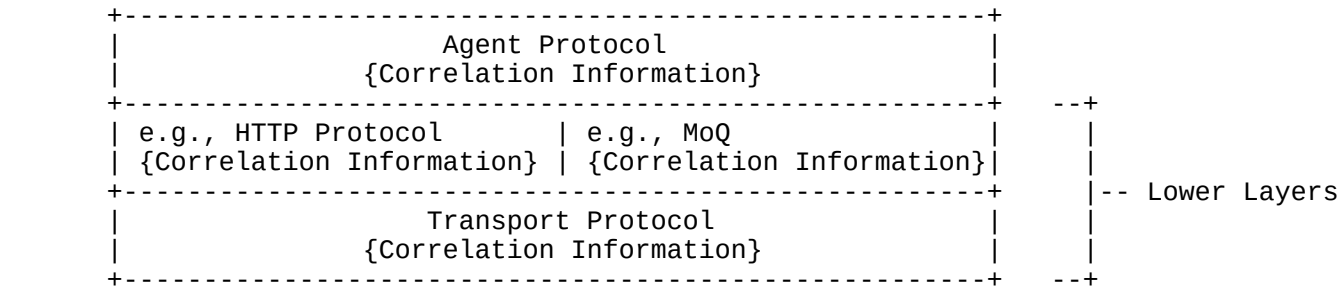

Figure 2. Enhancements in Agent protocol layer and/or lower layers.

4.1.2. Inter-stream Relationship Indicator

As outlined in 2.2, there is a potential concern associated with correlated multi-stream transmission. In this case, an inter-stream relationship indicator, as an example of the correlation information, is introduced to explicitly signal that multiple distinct streams belong to the same AI Agent traffic. This indicator could take a form of a unique ID that determined by source Agent(s), or it can be assigned by the underlying network. The sending entity encapsulates this indicator within the headers of all associated streams. Upon processing streams containing this indicator, the underlying network transmission nodes can leverage it to perform coordinated transmission operations. For example, base stations in mobile network can hold data from related streams to facilitate synchronization, and apply scheduling algorithms that consider the temporal relationship between streams. This enables the base station to orchestrate the transmission of data from these related streams to minimize arrival time variance to the receiver(s), ensuring data arrives as simultaneously as possible at intended endpoint(s).

This inter-stream relationship indicator can be encapsulated in the header of Agent Protocol Layer or lower layers, e.g., HTTP Layer, Transport Layer, etc.

Figure 3. *Correlation Information* in the HTTP header.

4.2. Single Stream with Multi-modality

Using method b) mentioned in 2.2, the source Agent can encapsulate multi-modal data of an AI Agent traffic (e.g., as separate Parts) within a single lower layer stream. One of the limitations of this method is that the underlying network remains unaware of individual Parts and their associated QoS requirements.

To address this limitation, the source Agent is required to perform Part segmentation during encapsulation: it should separate different Parts with clear boundaries, and embed supplementary metadata (e.g., traffic information, QoS requirements) corresponding to each Part within the packet header or sub-headers. This supplementary metadata

enables the underlying network to identify assistance information on specific Parts (e.g., their different traffic characteristics and respective QoS requirements).

a) In one implementation, supplementary metadata of each Part can be encapsulated in the packet header (e.g., Agent protocol header, or lower layer header) as in Figure 4. For example, the header contains fields such as:

* the total number of Parts,

* sequence number of each Part and the corresponding traffic information (e.g. size of a Part, modality of a Part, and transmission mode of a Part), QoS requirements (e.g., latency, error rate, priority, privacy), etc.

The contents of different Parts will be encapsulated in the payload in sequence with the same order of the supplementary metadata in the header.

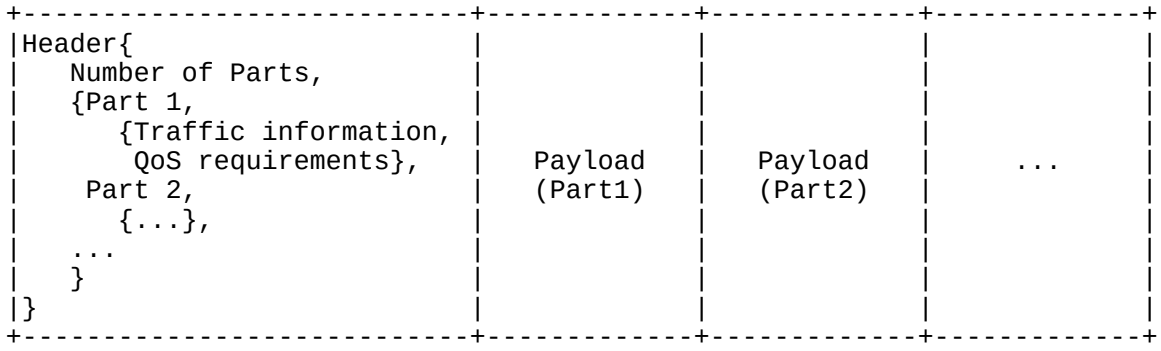

Figure 4. Supplementary metadata in the *header*.

b) In another implementation, supplementary metadata of each Part can be encapsulated in the sub-header of the sub-PDU as in Figure 5. For example, each sub-header contains fields such as:

* sequence number of each Part and the corresponding traffic information (e.g. size of a Part, modality of a Part, and transmission mode of a Part), QoS requirements (e.g., latency, error rate, priority, privacy), etc.

The contents of different Parts, on the other hand, will be encapsulated in the payload following the corresponding sub-header. In this case, a common header (e.g., in front of the subPDUs) can contain the field to indicate the total number of Parts of an AI Agent traffic.

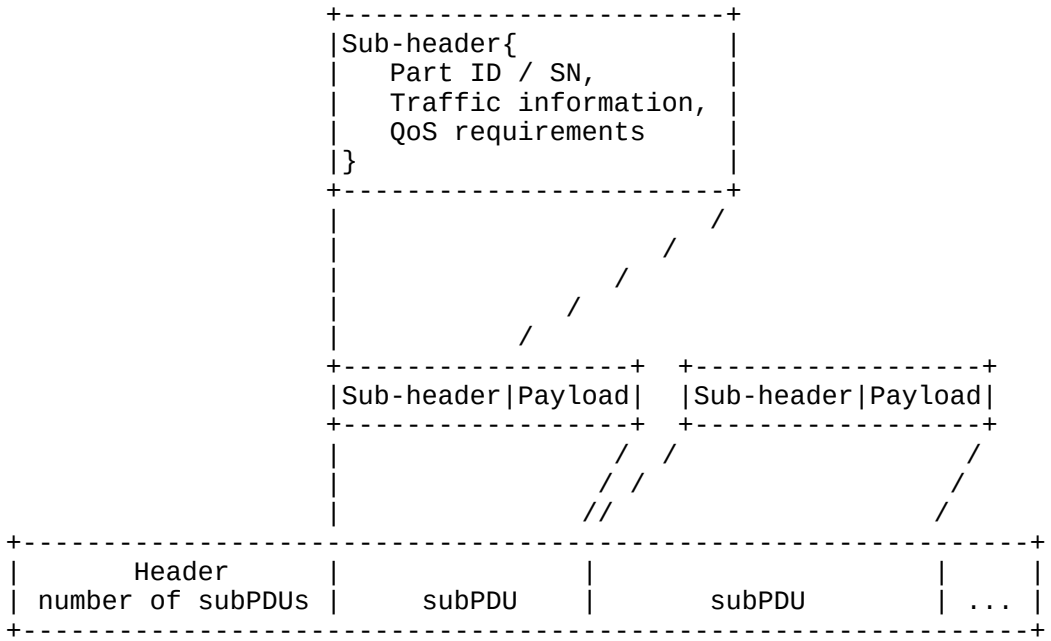

Figure 5. Supplementary metadata in the *sub-headers*.

Moreover, the source Agent may also inform the underlying network in advance about the supplementary metadata (e.g., traffic information, QoS requirements) corresponding to each distinct Part. The underlying network then provision QoS parameters for each Part. Upon receiving the stream, the underlying network can distinguish these multi-modal Parts based on their specific characteristics and associated QoS requirements.

Upon data arrival at the receiving endpoint, the entity of underlying network can perform reassembly of data packets associated with an AI Agent traffic prior to delivery to the receiving Agent client/server. Or, the entity of Agent protocol layer in the receiving Agent client/server can reassemble the data packets. Which one is better can be left for further study.

5. Summary

In this paper, we address the challenge wherein the underlying network lacks awareness of multi-modality of AI Agent traffic. We proposed two different ways for the source Agent client or server to encapsulate multi-modal data: one is to encapsulate multi-modal data into multiple streams, another one is to encapsulate multi-modal data into a single stream. In the former case, the source Agent notifies stream correlation information to the underlying network. In the latter case, the source Agent separates multi-modal data into different Parts, and embed supplementary metadata (e.g., traffic information and QoS requirements) within the header and/or sub-headers of Agent protocol layer or lower layers. For both methods, upon receiving the stream(s), underlying network can apply differential treatment to individual Parts, e.g., based on their specific QoS requirements.

6. Reference

[1] A2A Protocol. "Specification". https://a2a-protocol.org/v0.2.5/specification/, 2025.

[2] Jonathan Rosenberg , Cullen Fluffy Jennings, “Framework, Use Cases and Requirements for AI Agent Protocols” , https://datatracker.ietf.org/doc/draft-rosenberg-ai-protocols/, 2025.05

[3] Chenchen Yang , Huanhuan Huang, , Arashmid Akhavain , Faye Liu , Xueli An, , Weijun Xing, , Jinyan Li , Aijun Wang , Yang Wencong, “Requirements and Enabling Technologies of Agent Protocols for 6G Networks.”

[4] ETSI ENI ISG 055 Early Draft, “Use Cases and Requirements for AI Agents based Core Network” , 2025.06.04.

[5] 3GPP TR 22.870 (V1.1.0): “Study on 6G Use Cases and Service Requirements; Stage 1 (Release 20)” .

[6] 3GPP TR 23.801-01 (V0.3.0): “Study on Architecture for 6G System; Stage 2 (Release 20)” .

[7] C3-260051, C4-260225, C1-260122, “New SID on Study on the Protocol for AI in the 6G System” , 2026.02.

[8] 3GPP TR 23.501 (V20.0.0): “System architecture for the 5G System (5GS); Stage 2 (Release 20)” .

7. Security Considerations

This document does not introduce any new security considerations.

8. IANA Considerations

This document has no IANA actions.

9. Normative References

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/rfc/rfc8986>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/rfc/rfc2119>.

[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/rfc/rfc8174>.

Authors' Addresses

Yixin Lin (editor)
Huawei Technologies Co., Ltd.
Shanghai
China
Email: linyixin6@huawei.com

Chenchen Yang (editor)
Huawei Technologies Co., Ltd.
Shanghai
China
Email: yangchenchen7@huawei.com

Kuan Zhang (editor)
Huawei Technologies Co., Ltd.
Shanghai
China
Email: zhangkuan3@huawei.com

Xueli An, (editor)
Huawei Technologies Co., Ltd.
Munich
Germany
Email: Xueli.An@huawei.com

Shaoyun Wu (editor)
Huawei Technologies Co., Ltd.
Shanghai
China
Email: wushaoyun@huawei.com

Muhammad Awais Jadoon (editor)
InterDigital Europe Ltd.
London
United Kingdom
Email: Muhammad.AwaisJadoon@InterDigital.com

Sebastian Robitzsch (editor)
InterDigital Europe Ltd.
London
United Kingdom
Email: Sebastian.Robitzsch@InterDigital.com