Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: 17 October 2025

H. Zhu
Huazhong University of Science and Technology

15 April 2025

Architecture for Service Flow Characteristics and Modal Mapping Based on SDN and ALTO Protocol

draft-xsaopig-nmop-service-flow-modal-mapping-03

Abstract

This Internet-Draft specifies a comprehensive framework for mapping service flow characteristics to network modal resources in multi-modal intelligent computing networks. It introduces the use of the ALTO protocol for collecting service flow data and leverages an SDN architecture to separate control and data planes. The ALTO protocol facilitates the acquisition of diverse network state information, including data from several SDN domains and dynamic network environments, directly from controllers while keeping the provider’s internal details confidential. It then transmits the controller’s decisions using a proven method. The document details methods for characteristic identification, intelligent mapping, and continuous optimization, enabling dynamic resource allocation and improved network performance. The framework is designed to support scalable, efficient, and secure operations in environments with complex network loads and diverse service requirements.

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 17 October 2025.

Table of Contents

1. 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 available 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 17 August 2025.

2. Copyright Notice

3. Introduction

This standard aims to provide a comprehensive and systematic specification for mapping service flow characteristics to network modal resources in multi-modal intelligent computing networks. By introducing the ALTO protocol to collect service flow characteristic data and adopting an SDN architecture that separates the control plane from the data plane, this standard supports the creation of stable and efficient mapping templates between application service flows and modal resources.

4. Scope

This standard applies to designers, developers, and operators of multi-modal intelligent computing networks, particularly those requiring handling of complex network loads, computing resource demands, and data transmission efficiency in vertical industries. It defines methods for extracting critical service flow characteristics from applications and achieving effective mapping to network modal resources based on these characteristics.

5. Terms and Definitions

Intelligent Computing: AI-oriented computing capabilities for training and executing AI models.

Multi-Modal Intelligent Computing Network: A network integrating multiple modalities to serve diverse application scenarios, with computing as the core and network as the foundation.

Service Flow: A continuous data transmission process generated by an application or service, including unidirectional (e.g., client-server requests) or multi-directional interactions (e.g., video conferencing).

Service Flow Characteristics: Metrics describing application behavior, including throughput, latency, packet loss rate, CPU/GPU utilization, and storage capacity usage across three dimensions: storage, network forwarding, and computing.

Elastic Perception Feature Vector: A scalable vector representation dynamically adjusting granularity to characterize multi-dimensional service flow characteristics for flexible resource allocation.

Network Modality: A specific network type or configuration optimized for functions such as high bandwidth, low latency, or concurrency.

Modal Resource: Basic units constituting multi-modal networks (e.g., computing nodes, switching devices).

Feature-Modal Mapping Mechanism: A technical framework for matching service flow characteristics with optimal modal resource combinations.

6. Abbreviations

SDN: Software-Defined Networking

ALTO: Application-Layer Traffic Optimization Protocol

7. Overview

This standard addresses challenges in multi-modal intelligent computing networks, including dynamic workloads, heterogeneous service requirements, and frequent resource state changes. Key objectives:

1. Characteristic Identification: Use AI algorithms (e.g., graph matching, reinforcement learning) to analyze service flow characteristics.

2. Intelligent Mapping: Build an SDN/ALTO-based framework for dynamic resource allocation.

3. Continuous Optimization: Implement feedback loops to refine configurations based on real-time monitoring.

8. Architecture

The architecture design refers to[SDN_ALTO_MPTCP]. The architecture comprises four layers:

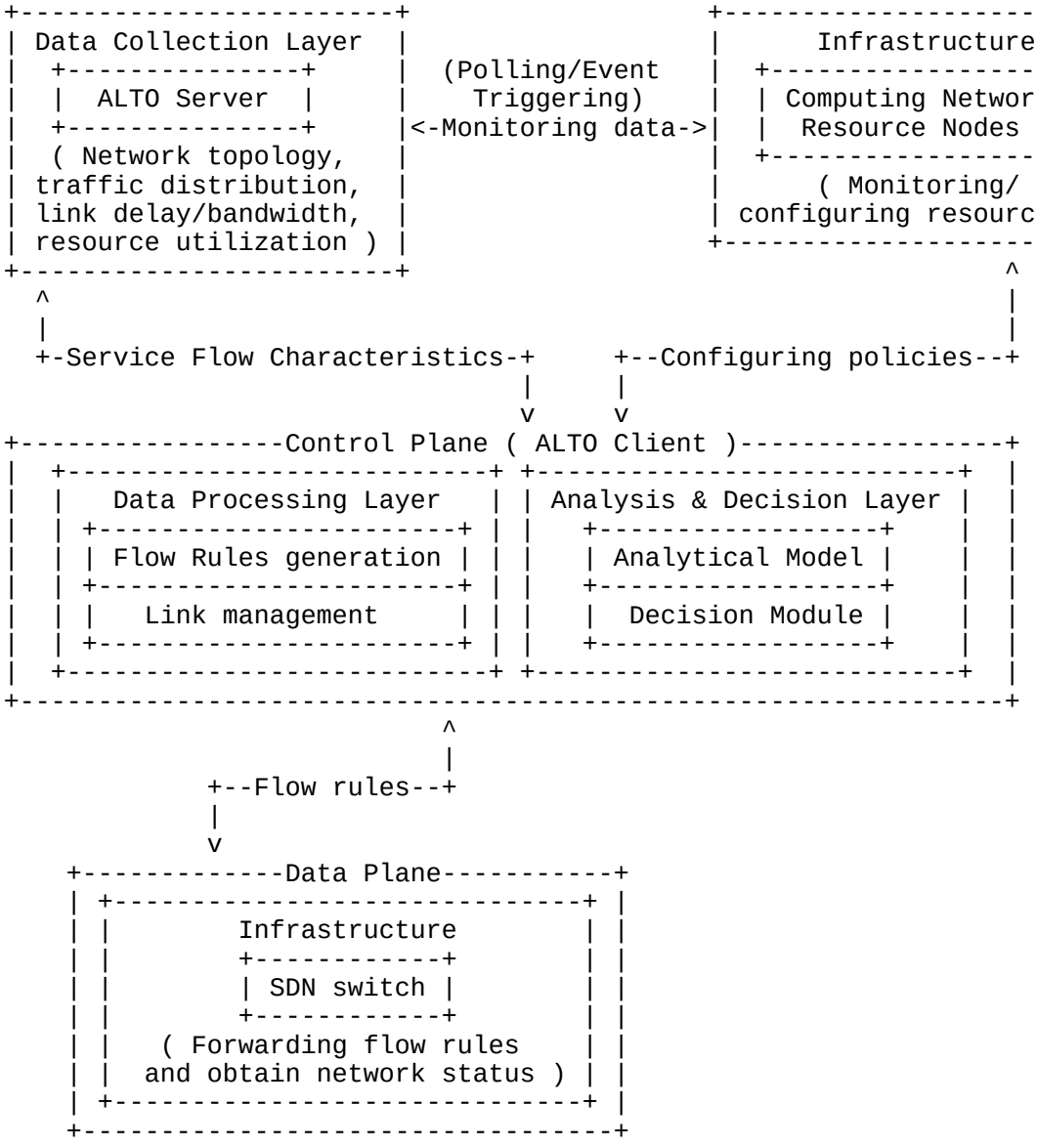

Figure 1

8.1. Infrastructure Layer

Provides hardware resources (computing nodes, switches) to support feature extraction and configuration.

8.2. Data Collection Layer

Collects real-time metrics: network topology, traffic distribution, link latency, CPU/GPU utilization, and storage I/O. Supports polling/event-driven mechanisms via SNMP, NetFlow, etc.

8.3. Data Processing Layer

Constructs service feature topology graphs using adjacency matrices. Nodes represent computing/storage metrics; edges represent network forwarding metrics. Employs distributed stream processing and graph databases.

8.4. Analysis & Optimization Layer

Performs deep analysis using graph neural networks and reinforcement learning to identify optimization strategies (e.g., topology adjustments, load balancing).

9. Service Feature-Network Modal Mapping

9.1. Service Feature Definition

Defined as three vectors:

* Storage: Node-level metrics (e.g., disk I/O).

* Network Forwarding: Link-level metrics (e.g., latency).

* Compute: Node-level metrics (e.g., CPU utilization).

9.2. Modal Definition

A three-dimensional tensor: {Service Capability, Controllable Resources, Operational Logic}.

9.3. Mapping Workflow

The mapping workflow consists of:

1. Feature Extraction: Collect real-time metrics across storage, compute, and network dimensions.

2. Topology Construction: Generate feature graphs with node/edge attributes.

3. Modal Matching: Align service features with modal resources using graph-matching algorithms.

4. Optimization: Adjust configurations (e.g., path rerouting, load migration).

5. Feedback: Continuously monitor performance and update mapping rules.

Technical Requirements:

* Use distributed optimization frameworks for real-time coordination.

* Apply reinforcement learning for adaptive decision-making.

10. Security Considerations

The transmission control model employed in this document relies on the default security mechanisms provided by SDN and ALTO protocols. This draft does not alter the default encryption and authentication models as specified in [RFC7149], [RFC7285], [RFC7286] and [RFC7971]. Therefore, the overall security of the service flow mapping system depends on the secure configuration and proper deployment of these underlying protocols.

11. IANA Considerations

This memo includes no request to IANA.

12. References

12.1. Normative References

[RFC7149] Boucadair, M. and C. Jacquenet, "Software-Defined Networking: A Perspective from within a Service Provider Environment", RFC 7149, March 2014, <https://www.rfc-editor.org/info/rfc7149>.

[RFC7285] Alimi, R., Penno, R., Yang, Y., Kiesel, S., Previdi, S., Roome, W., Shalunov, S., and R. Woundy, "Application-Layer Traffic Optimization (ALTO) Protocol", RFC 7285, September 2014, <http://www.rfc-editor.org/info/rfc7285>.

[RFC7286] Kiesel, S., Stiemerling, M., Schwan, N., Scharf, M., and H. Song, "Application-Layer Traffic Optimization (ALTO) Server Discovery", RFC 7286, November 2014, <http://www.rfc-editor.org/info/rfc7286>.

[RFC7971] Stiemerling, M., Kiesel, S., Scharf, M., Seidel, H., and S. Previdi, "Application-Layer Traffic Optimization (ALTO) Deployment Considerations", RFC 7971, October 2016, <https://www.rfc-editor.org/info/rfc7971>.

12.2. Informative References

[SDN_ALTO_MPTCP] Xing, Z., Di, X., and H. Qi, "The SDN-based MPTCP-aware and MPQUIC-aware Transmission Control Model using ALTO", IEEE ICC 2012, 2024, <https://datatracker.ietf.org/doc/draft-xing-nmop-sdn-controller-aware-mptcp-mpquic/>.

[SDN_ALTO] Gurbani, V. K., Scharf, M., Lakshman, T. V., Hilt, V., and E. Marocco, "Abstracting network state in Software Defined Networks (SDN) for rendezvous services", IEEE ICC 2012, 2012, <https://doi.org/10.1109/ICC.2012.6364858>.

Authors' Addresses

Huanxing Zhu, Huazhong University of Science and Technology, Wuhan, China, Email: huanxingzhu@hust.edu.cn

Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant 2023YFB2904100.

Author's Address

Huanxing Zhu
Huazhong University of Science and Technology
Wuhan
China
Email: huanxingzhu@hust.edu.cn