Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 5 May 2026

Q. Yuan
J. Mao
B. Liu
N. Geng
X. Shang
Q. Gao
Z. Li
Huawei Technologies
1 November 2025

Use cases of the AI Network Traffic Optimization Agent
draft-yuan-rtgwg-traffic-agent-usecase-00

Abstract

This document introduces AI Network Traffic Optimization Agents as a dynamic alternative to traditional static network optimization methods. These AI entities analyze real-time network status (e.g., latency, node load) and adjust resources flexibly—deployed centrally or on devices—to enhance efficiency, ensure service quality, and cut operational costs. It defines network traffic optimization (maximizing resource use, meeting QoS) and AI agents (autonomous, learning entities that reduce manual work), then details three key application scenarios: tunnel adjustment (adaptive routing, predictive bandwidth, fault recovery), traffic steering (classification, application-aware policies, pre-emptive load balancing), and network slice adjustment (lifecycle automation, SLA compliance, slice-specific fault fixes). The document emphasizes the agents' role in enabling SLA-compliant, autonomous optimization for complex networks.

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 5 May 2026.

Table of Contents

1. Introduction

AI Network Traffic Optimization Agents are intelligent entities that analyze real-time network telemetry (e.g., bandwidth occupancy, latency, node load, packet loss) and dynamically adjust network resources on behalf of operators. Their core goals are to boost network efficiency, guarantee service quality, and lower operational costs for end-users. These agents offer flexible deployment: they can be implemented on centralized network management platforms to integrate global data for holistic optimization, or embedded in edge devices (e.g., routers, switches, IoT gateways) to respond in real time to local traffic fluctuations.

This deployment flexibility not only integrates intelligent decision-making with network infrastructure but also breaks through the rigid interaction barriers of traditional networks. Unlike conventional network devices, which rely on strict, fully standardized protocols for communication—often plagued by version incompatibility and format constraints—AI agents enable adaptive, context-aware inter-device collaboration. They natively support semi-structured data exchange and natural language interaction, allowing seamless communication across heterogeneous devices and reducing friction from fragmented protocols.

Traditional network optimization relies heavily on static configuration rules and manual adjustments, limiting it to coarse-grained issue resolution. During traffic surges (e.g., peak e-commerce sales, large-scale video conferences), this approach fails to adapt promptly, leading to increased latency or packet loss—especially detrimental to mission-critical applications requiring stable transmission. In contrast, AI Network Traffic Optimization Agents enable fine-grained, autonomous optimization: they steer traffic to underutilized paths that meet application-specific SLA requirements, or allocate exclusive resource channels for high-priority services, ensuring performance remains unaffected by non-critical traffic. Their interactive capabilities further amplify these advantages, enabling faster cross-device coordination and more agile response to dynamic network changes.

2. 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 RFC 2119[RFC2119].

3. Network Traffic Optimization and AI Agent

3.1. Network Traffic Optimization

Network traffic optimization encompasses a suite of technologies, strategies, and practices focused on monitoring, managing, and dynamically adjusting data flows across a network. Its core objectives are to maximize the efficiency of network resources (e.g., bandwidth, node capacity), mitigate issues such as congestion, latency, and packet loss, and ensure critical applications (e.g., online gaming, financial transactions, real-time video conferencing) meet their required Quality of Service (QoS) standards.

By redistributing traffic to underutilized paths, prioritizing high-priority requests, and smoothing sudden traffic surges, it transforms passive network management into proactive adjustment. This approach supports the stable operation of modern complex networks (including 5G, edge computing, and multi-vendor hybrid environments) while minimizing unnecessary operational costs—with AI-driven interaction capabilities further enhancing its adaptability to heterogeneous network ecosystems.

3.2. AI Agent

An AI Agent is an automated intelligent entity designed to act on behalf of users, systems, or organizations to achieve specific goals. Its core capabilities include:

* Perceiving contextual information (e.g., real-time network status, user behavior, environmental changes) through multi-source data collection;

* Analyzing data via advanced algorithms (e.g., machine learning, reinforcement learning) to derive actionable insights;

* Making autonomous decisions and executing tasks independently or in collaboration with other agents;

* Supporting semi-structured data exchange (e.g., schema-less telemetry metrics, partial configuration snippets) to break free from rigid format constraints;

* Enabling natural language interaction (NLI) for simplified human-device and inter-device communication.

Unlike traditional static programs, AI Agents can self-learn and iterate based on historical data, adapting to dynamic scenarios (e.g., real-time traffic path adjustment, personalized policy execution). Their key value lies in reducing manual intervention, improving task efficiency, and addressing complex problems requiring

real-time, data-driven decision-making. Critically, their flexible interaction models reduce reliance on strict standardization, minimizing version compatibility issues and enabling seamless integration of new devices— laying the foundation for rapid network iteration and scalability.

4. Usage Scenarios of the AI Network Traffic Optimization Agent

This section outlines typical application scenarios of AI Network Traffic Optimization Agents across three key network operation domains: tunnel adjustment, traffic steering into tunnels, and network slice adjustment. Leveraging AI algorithms and real-time telemetry, the agents automate optimization, enhance service reliability, and ensure SLA compliance— while their interactive capabilities (semi-structured data support, natural language interaction) amplify efficiency and scalability.

4.1. Tunnel Adjustment

AI Network Traffic Optimization Agents optimize tunnels (e.g., RSVP-TE tunnels, SRv6 tunnels) by dynamically adapting to network conditions, ensuring efficient data transmission and fault resilience. Their interactive capabilities streamline cross-device coordination, accelerating decision-making and recovery.

4.1.1. Adaptive Tunnel Routing

Agents collect real-time telemetry (e.g., link utilization, latency, packet loss) and network topology information (via protocols such as BGP-LS or IS-IS). Using machine learning-based routing algorithms (e.g., reinforcement learning for path selection), they identify optimal tunnel paths. When congestion or link degradation is detected, agents proactively recompute paths and share intent-driven instructions (via semi-structured data) with routers/switches to minimize end-to-end latency without relying on rigid protocol syntax.

4.1.2. Predictive Bandwidth Allocation

Agents analyze historical traffic patterns (e.g., diurnal peaks for enterprise services) to predict future bandwidth demands for each tunnel. Through tunnel signaling protocols (e.g., RSVP-TE), they implement dynamic adjustment: reducing allocation during off-peak periods to avoid waste, and scaling up bandwidth preemptively before traffic surges. Operators can also fine-tune prediction parameters via natural language prompts (e.g., “Increase bandwidth buffer for weekday 9 AM video conferences”), simplifying policy updates.

4.1.3. Autonomous Fault Detection and Recovery

Agents monitor real-time tunnel KPIs (e.g., availability, jitter) and use anomaly detection models (e.g., autoencoders) to identify faults (e.g., link failures). Upon detection, they automatically share fault details (via semi-structured data) with other agents and initiate recovery actions (e.g., switching traffic to pre-provisioned backups). This cross-agent collaboration reduces Mean Time to Recovery (MTTR) by eliminating manual coordination delays.

4.2. Traffic Steering into Tunnels

AI Network Traffic Optimization Agents enable fine-grained traffic steering, mapping flows to tunnels that align with their QoS requirements. Their support for multi-format data and natural language interaction simplifies policy configuration and cross-device coordination.

4.2.1. Traffic Classification and Priority Mapping

The agent can perform deep packet inspection (DPI) and flow analysis (via protocols such as NetFlow v9 or IPFIX) to classify traffic based on service type (e.g., VoIP, 4K video, bulk data), user priority (e.g., VIP users), and QoS class.Through policy-based routing (PBR) or segment routing policies, it maps classified traffic to tunnels with QoS capabilities that match the traffic’ s needs— for example, low-latency tunnels for VoIP or high-bandwidth tunnels for bulk data.

4.2.2. Application-Aware Steering Policies

Agents use application signature recognition (e.g., TLS SNI, DNS queries) to identify application-specific traffic (e.g., Zoom, AWS S3 transfers). They enforce application-specific rules: real-time applications are directed to tunnels with guaranteed latency (<50ms) and low packet loss (<0.1%), while non-critical traffic uses cost-efficient tunnels. Operators can define these rules via natural language (e.g., “Route all IoT sensor data to shared low-cost tunnels”), with agents translating prompts into executable policies— reducing configuration complexity.

4.2.3. Pre-emptive Traffic Load Balancing

Agents forecast traffic hotspots (e.g., regional surges from events) using time-series models (e.g., LSTM). They implement pre-emptive steering to distribute predicted heavy traffic across parallel tunnels, preventing overload. Agents share load distribution plans with edge devices via semi-structured data, ensuring uniform resource utilization across the network without strict protocol alignment.

4.3. Network Slice Adjustment

AI Network Traffic Optimization Agents support the lifecycle management and optimization of Network Slice Instances (NSIs), focusing on resource efficiency, SLA compliance, and fault resilience. Their flexible interaction models enable seamless collaboration between slice-specific agents, accelerating slice deployment and adjustment.

4.3.1. NSI Lifecycle Automation

Agents participate in NSI design by using slice requirements (e.g., bandwidth, latency, isolation) to recommend optimal resource allocation (e.g., CPU, bandwidth, tunnel assignments) and topology configurations (e.g., dedicated vs. shared tunnels). They automate instantiation and termination: during deployment, agents coordinate across devices to deploy required tunnels and steering rules (via semi-structured data exchange); upon termination, they release resources to prevent leakage. This cross-agent collaboration reduces reliance on standardized interfaces, enabling faster slice deployment.

4.3.2. Closed-Loop SLA Compliance

Agents monitor slice-level KPIs (e.g., throughput, latency) in real time and compare them against SLA thresholds (e.g., 100 Mbps minimum throughput, 100ms maximum latency). When SLA violations are predicted or detected, they trigger closed-loop adjustments (e.g., augmenting tunnel bandwidth, optimizing routing paths). Operators can also set SLA thresholds via natural language (e.g., “Ensure industrial IoT slice latency stays below 80ms”), making policy updates intuitive and agile.

4.3.3. Slice-Specific Fault Remediation

The agent can analyze multi-dimensional slice alarms (e.g., tunnel faults within the slice, resource shortages) via correlation models that integrate slice topology and historical fault data. It enables slice-aware fault recovery: it identifies the root cause of slice degradation (e.g., a failed tunnel in the slice’s path) and executes slice-specific remediation (e.g., re-provisioning a dedicated backup tunnel for the slice), thereby minimizing impact on the slice’s services.

5. Conclusion

This document systematically elaborates on AI Network Traffic Optimization Agents, covering their role in addressing traditional network limitations, core definitions of network traffic optimization and AI agents, and practical application scenarios. Beyond dynamic resource allocation and SLA-compliant optimization, these agents deliver transformative value through enhanced inter-device interaction capabilities.

By supporting semi-structured data exchange, AI agents break free from the rigid format constraints of traditional network protocols, enabling seamless communication across heterogeneous devices and vendors. Natural language interaction simplifies policy configuration and human-device collaboration, lowering operational barriers. These features reduce reliance on strict standardization and mitigate version compatibility issues, fostering network scalability and enabling rapid iteration of optimization strategies.

In complex environments such as 5G, edge computing, and multi-vendor hybrid networks, AI Network Traffic Optimization Agents serve as a cornerstone of next-generation intelligent networks. They not only automate fine-grained optimization to enhance efficiency and service quality but also through flexible interaction models, enable agile response to dynamic traffic patterns and emerging service requirements— future-proofing networks against technological evolution while minimizing operational costs. As network ecosystems grow more complex, the interactive and adaptive capabilities of these AI agents will become increasingly critical to unlocking the full potential of intelligent network management.

6. Security Considerations

TBD.

7. IANA Considerations

TBD.

8. Normative References

[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/info/rfc2119>.

Authors' Addresses

Quan Yuan
Huawei Technologies
Beijing
100095
China
Email: yuanquan25@huawei.com

Jianwei Mao
Huawei Technologies
Beijing
100095
China
Email: maojianwei@huawei.com

Bing Liu
Huawei Technologies
Beijing
100095
China
Email: leo.liubing@huawei.com

Nan Geng
Huawei Technologies
Beijing
100095
China
Email: gengnan@huawei.com

Xiaotong Shang
Huawei Technologies
Beijing
100095
China
Email: shangxiaotong@huawei.com

Qiangzhou Gao
Huawei Technologies
Beijing
100095
China
Email: gaoqiangzhou@huawei.com

Zhenbin
Huawei Technologies
Beijing
100095
China
Email: robinli314@163.com