Computing-Aware Traffic Steering Q. Li
Internet-Draft T. gao
Intended status: Standards Track Pengcheng Laboratory
Expires: 2 September 2026 Y. Jiang
Tsinghua Shenzhen International Graduate School & Pengcheng Laboratory
1 March 2026

Semantic-Driven Traffic Shaping Contract for AI Networks
draft-li-cats-aisemantic-contract-00

Abstract

This document defines a "Semantic-Driven Shaping Contract". Traditional network protocols treat AI training and inference traffic as opaque byte streams, leading to highly inefficient scheduling. This contract allows applications or distributed training frameworks to explicitly pass "minimum necessary semantics" to the underlying network. In exchange, the network commits to executing fine-grained, differentiated forwarding and resource allocation actions for tensor flows with diverse semantics, based on predefined rules and global real-time states. This model significantly improves overall resource utilization and task completion times in heterogeneous computing networks, cross-domain intelligent computing centers, and integrated training-inference scenarios.

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 2 September 2026.

Copyright Notice

Table of Contents

1. Problem Statement: Limitations of Existing Network Mechanisms

In the era of large AI models, the "importance" of traffic dynamically shifts with the model's phase and exhibits a high degree of computability. Existing traffic control and Quality of Service (QoS) mechanisms suffer from fundamental flaws in this context:

* *Coarse QoS Granularity and Invalid Implicit Assumptions:* Traditional QoS assumes that traffic within the same class has negligible variance and its importance remains stable at the session level. However, in AI scenarios, QoS fails to differentiate between "early-layer activations" and "late-layer activations," nor can it distinguish between the "KV Cache of early tokens" and "tail tokens."

* *Static and Incomputable DiffServ Semantics:* Differentiated Services (DiffServ/DSCP) rely on static markings that the network blindly executes. It cannot express dynamic computing semantics, such as "this flow is quantizable during congestion," "this flow tolerates a 5ms store-and-forward delay," or "this flow requires absolute preemption."

* *Passive and Dimensionless ECN Feedback:* Existing Explicit Congestion Notification (ECN) mechanisms operate on the assumption that "end-systems know best how to respond to congestion" and that "rate reduction is the only correct response." It possesses zero understanding of computing semantics, treating activations, gradients, blocking, and non-blocking operations equally. In AI inference, the correct response to congestion is often "precision degradation (quantization)" or "prioritizing the draining of critical tensors," rather than blind rate reduction.

2. Cross-Domain Amplification of Challenges

In cross-domain intelligent computing networks characterized by multi-tasking, multi-tenancy, and integrated training and inference, the aforementioned flaws are severely amplified:

* *Time-scale Mismatch:* Cross-domain Round-Trip Times (RTT) reach the millisecond level, easily exceeding the "effective value window" of sensitive tensors like late-layer activations. The network MUST make differentiation and routing decisions instantaneously during forwarding; post-facto congestion control feedback is entirely ineffective.

* *Resource & Path Asymmetry:* Cross-domain links are scarce, high-cost resources. Delay-tolerant and compressible intermediate activations absolutely MUST NOT compete equally for cross-domain bandwidth with critical gradients that require immediate delivery.

* *Tight Compute-Network Coupling:* Traffic steering is no longer merely about "delivery to a fixed destination." It requires dynamic selection based on compute heterogeneity (e.g., local GPU vs. remote FPGA). A lack of semantic understanding leads to a severe mismatch between computing power and network resources.

3. The Semantic-Driven Mapping Loop: The Contract

The core of this draft is to establish a closed-loop mapping mechanism from "application-layer semantic input" to "network-side action commitment."

3.1. Semantic Information Model (Metadata Model)

The application layer MUST expose "exchangeable Semantic Metadata" to the network. Based on the commonalities and specifics of training and inference tasks, this is categorized as follows:

* *Traffic Class:* Explicitly identifies the data type (e.g., Activation, Gradient, KV Cache, Parameter, Collaborative State Synchronization).

* *Urgency & Dependency:* Provides coarse-grained dependency hints (e.g., Early-token vs. Late-token) and the current layer or stage of the model (Layer ID / Pipeline Stage).

* *Tolerance & Sensitivity:*

  - *Fidelity/Accuracy Sensitivity:* Indicates whether in-network low-precision quantization is permitted.

  - *Loss/Latency Tolerance:* Indicates whether the flow permits buffering (store-and-forward) or dropping.

* *Compute Affinity:* Indicates the preferred characteristics of the underlying computing power (e.g., GPU, FPGA, CPU, or specific operator acceleration hardware).

3.2. Network Policy / Action Set

Upon receiving the aforementioned semantics, network nodes with global state awareness can execute a set of policies that transcend traditional routing:

* *Queueing / Scheduling:* Identifies flow states to guarantee absolute preemption for highly time-sensitive traffic.

* *Buffering / Store-and-forward:* Utilizes the storage resources of network devices to temporarily delay flows with high latency tolerance (e.g., large-block parameter pulls); it also implements cache multiplexing for inference requests from different users, directly optimizing hardware throughput without altering the model structure.

* *Shaping & In-network Quantization:* Triggers in-network low-precision quantization and sparsity strategies during congestion, rather than relying on simple packet dropping.

* *Steering:* Intelligently guides task flows to the most appropriate heterogeneous computing nodes based on Compute Affinity.

4. Extended Use Case: Top-K Routing Semantics for MoE Architecture

For dynamic computing architectures like Mixture-of-Experts (MoE), this contract supports the definition of more complex routing metadata for intelligent scheduling in the network data plane:

* *Model Router Metadata:* Carries Token ID / Query vector summaries, Top-K candidate expert lists, weights/confidence levels, and positional markers (Token_pos).

* *System State Semantics:* Network nodes maintain real-time metrics for each expert node, including backlog queues, computing load, network latency, and bandwidth utilization.

By matching these two semantics, the network can instantaneously determine which Expert node with the lightest load should receive the Token flow at the moment of forwarding.

5. Deployment Considerations

5.1. Decision Location: Why In-Network?

Compared to edge devices (GPUs/NICs) that only possess local queuing information, in-network nodes (e.g., Core/Spine Switches) maintain a global perspective. The network can perceive concurrent multi-tenant tasks and real-time multipath congestion states. Crucially, it can make immediate decisions to buffer, slice, or reroute cross-domain traffic before it enters high-cost bottleneck links.

5.2. RDMA / RoCEv2 Integration

Intelligent computing centers rely heavily on RDMA. The Semantic Header defined in this contract will be designed as Extension Headers for RoCEv2/UDP packets, or carried using specific reserved fields. This enables supporting hardware (such as the FPGA and parsing pipelines in the IntelliNode architecture) to extract metadata and execute policies at line rate (e.g., 400Gbps).

6. Security Considerations

To ensure the integrity of the Semantic-Driven Shaping Contract, the system MUST:

* *Authentication and Anti-Spoofing:* Prevent malicious tenants from tampering with the Urgency level or forging network states to unfairly preempt high-priority queues.

7. IANA Considerations

This document requests that IANA allocate specific protocol numbers or RoCEv2 option type spaces for the AI Semantic Header to facilitate standardized deployment.

Authors' Addresses

Qing Li
Pengcheng Laboratory
Email: liq@pcl.ac.cn

Teng gao
Pengcheng Laboratory
Email: gaot@pcl.ac.cn

Yong Jiang
Tsinghua Shenzhen International Graduate School & Pengcheng Laboratory
Email: jiangy@sz.tsinghua.edu.cn