Computing-Aware Traffic Steering
Internet-Draft
Intended status: Informational
Expires: 19 March 2026

H. Wang
Q. Li
Pengcheng Laboratory
Y. Jiang
Tsinghua Shenzhen International Graduate School, Pengcheng Laboratory

15 September 2025

In-Network Intelligence for Distributed Collaborative Inference Acceleration
draft-wang-cats-innetwork-infer-00

Abstract

The rapid proliferation of deep learning models has led to growing demands for low-latency and high-throughput inference across heterogeneous environments. While edge devices often host data sources, their limited compute and network resources restrict efficient model inference. Cloud servers provide abundant capacity but suffer from transmission delays and bottlenecks. Emerging programmable in-network devices (e.g., switches, FPGAs, SmartNICs) offer a unique opportunity to accelerate inference by processing tasks directly along data paths.

This document introduces an architecture for _Distributed Collaborative Inference Acceleration_. It proposes mechanisms to split, offload, and coordinate inference workloads across edge devices, in-network resources, and cloud servers, enabling reduced response time and improved utilization.

About This Document

This note is to be removed before publishing as an RFC.

The latest revision of this draft can be found at https://kongyanye.github.io/draft-wang-cats-innetwork-infer/draft-wang-cats-innetwork-infer.html. Status information for this document may be found at https://datatracker.ietf.org/doc/draft-wang-cats-innetwork-infer/.

Discussion of this document takes place on the Computing-Aware Traffic Steering Working Group mailing list (mailto:cats@ietf.org), which is archived at https://mailarchive.ietf.org/arch/browse/cats/. Subscribe at https://www.ietf.org/mailman/listinfo/cats/.

Source for this draft and an issue tracker can be found at https://github.com/kongyanye/draft-wang-cats-innetwork-infer.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Large foundation models and domain-specific deep neural networks are increasingly deployed in real-time services such as surveillance video analysis, autonomous driving, industrial inspection, and natural language interfaces. Inference for such models requires both *low latency* and *scalable throughput*.

Current deployments typically follow two paradigms:

* *Edge-only inference*, which minimizes data transmission but is constrained by limited device resources.

* *Cloud-centric inference*, which exploits large compute capacity but introduces network delays.

However, neither paradigm fully exploits the potential of programmable *in-network intelligence*, where intermediate devices along the data path can actively participate in computation. By integrating such devices into distributed collaborative inference, networks can enable *end-to-end acceleration of large-scale deep learning model inference*.

This document outlines the motivation, problem statement, and architectural considerations for _Distributed Collaborative Inference Acceleration (DCIA)_. The goal is to establish a framework where deep learning inference tasks are intelligently partitioned, scheduled, and executed across heterogeneous resources, including edge devices, in-network resources, and cloud servers.

2. Problem Statement

* *Latency bottlenecks:* Large model inference may exceed the latency tolerance of interactive applications if computed only at edge or cloud.

* *Resource fragmentation:* Heterogeneous resources (edge GPUs, in-network accelerators, cloud clusters) are not effectively coordinated.

* *Lack of steering semantics:* Existing approaches to service steering are not optimized for inference workload partitioning and scheduling.

3. Proposed Approach

The framework for DCIA includes the following:

1. *Model Partitioning and Mapping* Split large models into sub-tasks (e.g., early layers at edge, mid layers in-network, final layers in cloud) and map them based on node capabilities, load, and network conditions.

2. *In-Network Execution* Enable inference acceleration in programmable switches, FPGAs, or SmartNICs, utilizing data-plane programmability to process features in transit (e.g., feature extraction, embedding computation).

3. *Task Scheduling and Steering* Extend service capability advertisements with inference-oriented metrics (e.g., GPU/FPGA availability, model version, layer compatibility), and dynamically balance inference tasks across heterogeneous resources.

4. *Load Balancing Protocols* Support task redirection and failover when a device becomes overloaded, and explore transport-level extensions to allow adaptive task splitting along paths.

4. Use Cases

* *Video Analytics:* Smart cameras extract features locally, switches perform intermediate tensor transformations, and cloud servers handle complex classification.

* *Autonomous Vehicles:* Onboard processors execute lightweight inference, roadside units conduct mid-layer fusion, and cloud clusters finalize planning decisions.

* *Interactive AI Services:* Edge devices handle pre-processing, in-network resources accelerate embeddings, and cloud models provide final responses.

5. Conventions and Definitions

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.

6. Security Considerations

Inference partitioning must consider:

* *Data confidentiality*, ensuring sensitive inputs are not exposed in untrusted network elements.

* *Model integrity*, preventing tampering or unauthorized reuse of model partitions.

* *Policy enforcement*, allowing operators to specify where inference may or may not occur.

7. IANA Considerations

This document has no IANA actions.

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/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>.

Acknowledgments

The authors would like to thank colleagues and reviewers in the community who provided feedback on the early version of this draft.

Authors' Addresses

Hanling Wang
Pengcheng Laboratory
Email: wanghl03@pcl.ac.cn

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

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