

Unified Optical Networks and AI Computing Orchestration (UONACO)

CCAMP WG, IETF125

draft-tan-ccamp-uonaco-problem-statement-02

draft-hu-ccamp-uonaco-control-framework-01

Author:

Qiaojun Hu (Beijing University of Posts and Telecommunications)

Zheng Han (Beijing University of Posts and Telecommunications)

Yanxia Tan (ChinaUnicom)

Yanlei Zheng (ChinaUnicom)

Wei Wang (Beijing University of Posts and Telecommunications)

Yongli Zhao (Beijing University of Posts and Telecommunications)

Jie Zhang (Beijing University of Posts and Telecommunications)

Motivation: Why Optical for AI?

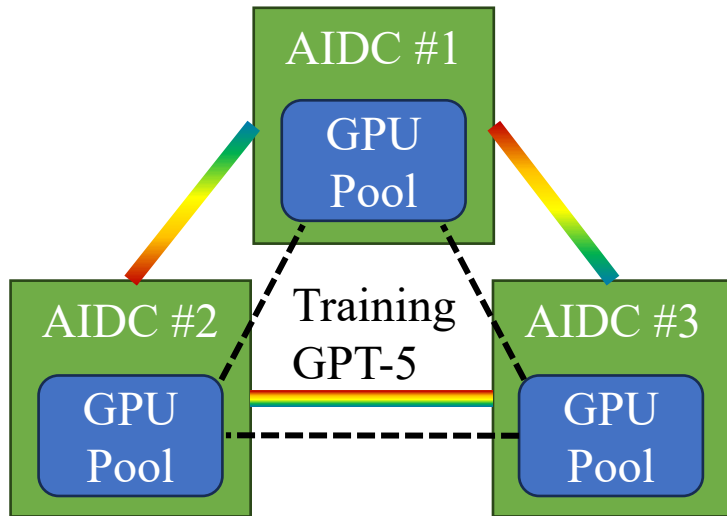
- Coordination between networks and computational resources is a hotpot within the IETF (e.g., CATS and SRv6).
- Key drivers for optical network and AI computing collaboration
 - Latency reduction & Bandwidth enhancement
 - Hard-Isolation characteristics (Crucial for AI)
- Why specific to Optical instead of general IP? (Addressing previous feedback):
 - Distributed AI training (e.g., GPU All-Reduce) demands zero-jitter and deterministic latency. Standard IP statistical multiplexing cannot guarantee this.
 - CCAMP WG possesses a solid foundation in optical network control and management, facilitating subsequent work progression.

Concept of UONACO

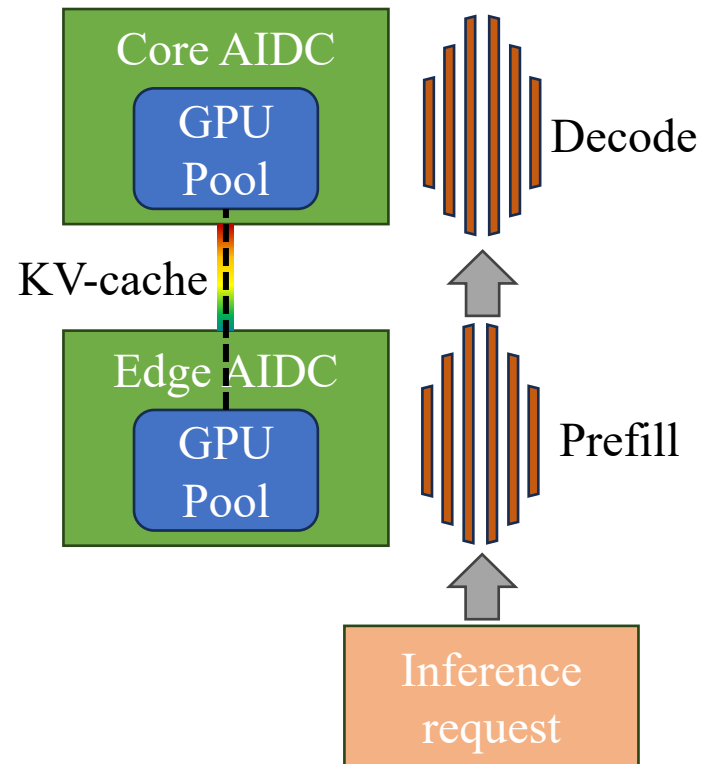
- Unified Optical Networks and AI Computing Orchestration (UONACO) is a framework that enables **bidirectional awareness**, **resource abstraction**, and **joint orchestration** across the compute-optical boundary, resolving the critical isolation between optical transport networks and AI computing infrastructure to optimize compute efficiency and network resource utilization.

Use Cases

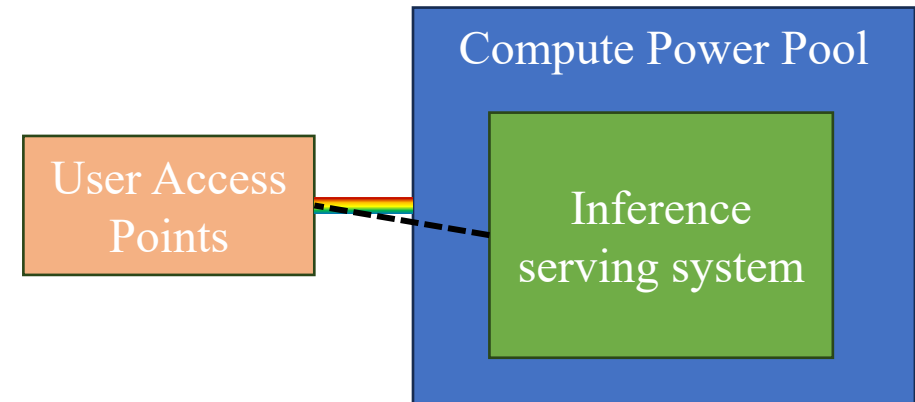
➤ Distributed AI Training



➤ Distributed AI Inference



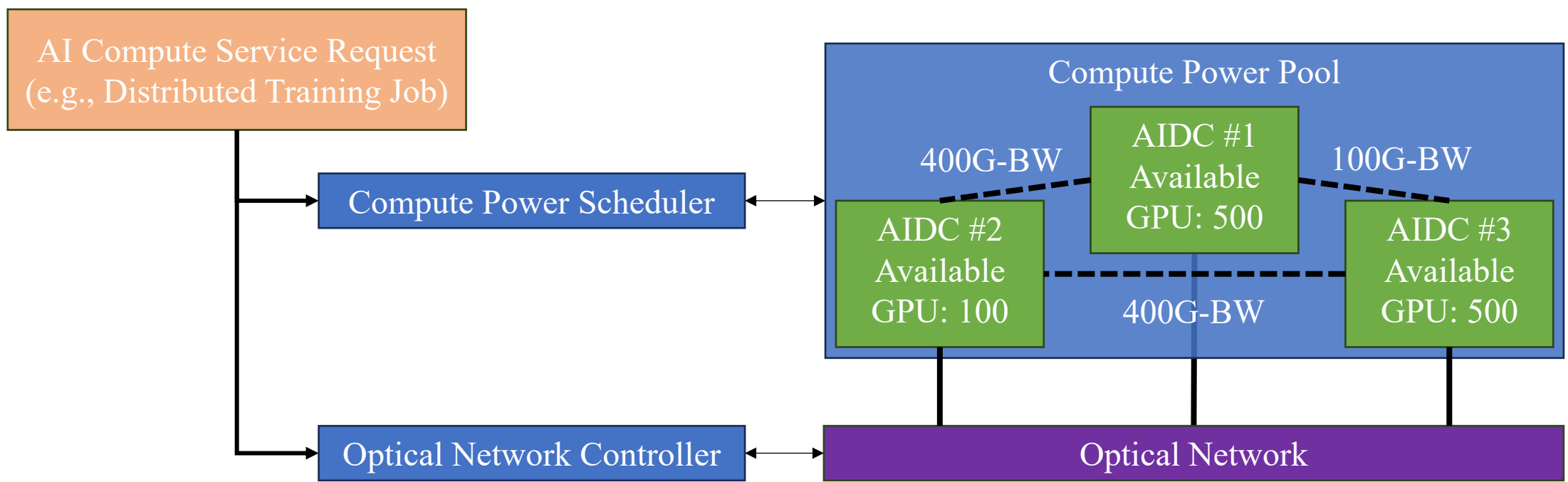
➤ Accessing Remote AI Service



Problem Statement

- Isolated Control and Management.
- Independent Resource Efficiency Evaluation.

Suboptimal Resource Allocation



Requirements for Joint Orchestration

- Integrated Control and Management Architecture
 - Requires bidirectional fusion between compute and network control planes.
- Unified Abstraction and Evaluation Framework
 - A common language for Compute, Storage, and Optical states.
 - Feedback from last meeting: "There are three ways to manage a bit: compute, store, or transfer." (Stephen)
- Scheduling Algorithms for Joint Orchestration
 - Requires synchronized, domain-specific end-to-end resource allocation.

Relationship with CATS: Complementary

Dimension	CATS	UONACO
Core Objective	Steering (routing traffic on existing paths)	Orchestration(managing resource lifecycles)
Compute Action	Awareness (monitoring status)	Reservation(locking resources)
Network Action	Selection(choosing among existing pipes)	Provisioning(establishing new optical circuits)
Trigger Mechanism	Traffic/Packets arriving at Ingress	Service Intent/API request prior to traffic
Typical Use Case	User-to-Compute (e.g., AR/VR, Inference Access)	Compute-to-Compute (e.g., Distributed AI Training)

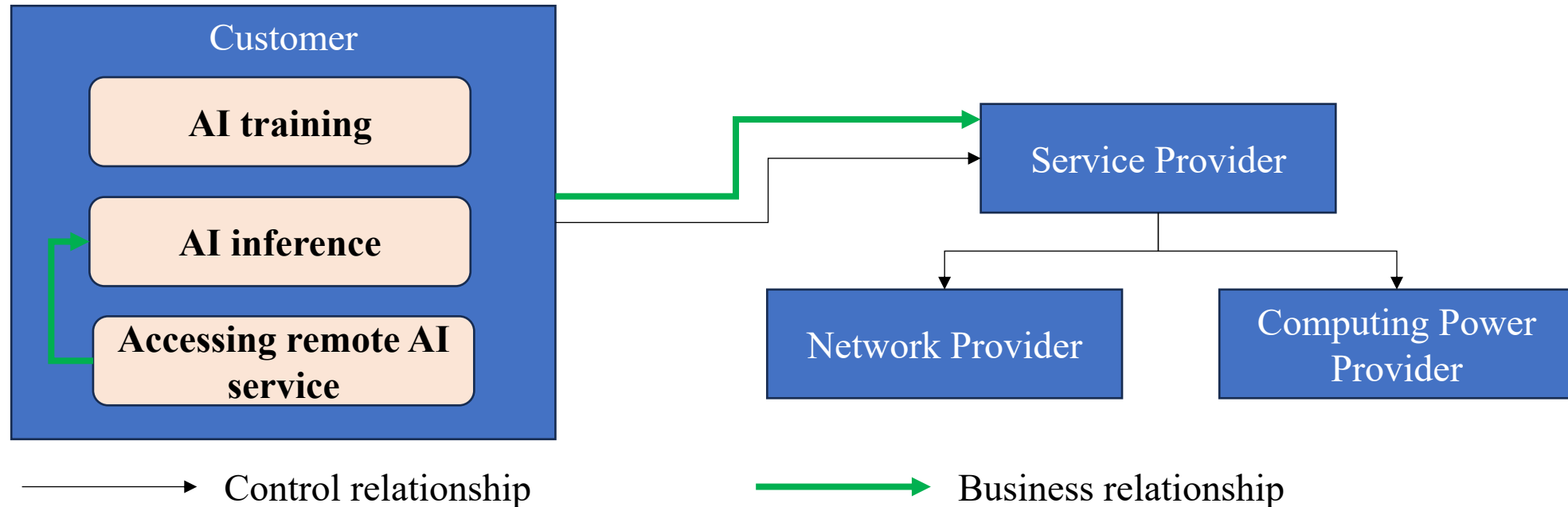
Relationship with CATS for OTN

- In CATS for OTN (draft-zhao-cats-otn-applicability-00, Sec 5.3):
 - “...paths and optical channels (e.g., ODUk/fgOTN) may be pre-provisioned... The C-PS then identifies the most suitable path...” → (Traffic Steering)
- In UONACO (draft-tan-ccamp-uonaco-problem-statement-02, Sec 2.3):
 - "For AI workloads requiring massive, dedicated bandwidth, simply steering traffic over existing shared pipes is often insufficient... We need the capability to provision new physical or logical links on demand." → (Joint Orchestration)

Service Model

➤ Customer

- AI training: specify model, deadline, performance, scale, and data privacy requirements.
- AI inference: lease computing resources to deploy and operate inference models.
- Accessing remote AI service: invoke pre-deployed inference APIs offered by third parties.



UONACO Control and Management Architecture

➤ Design Goals:

➤ Converged Architecture:

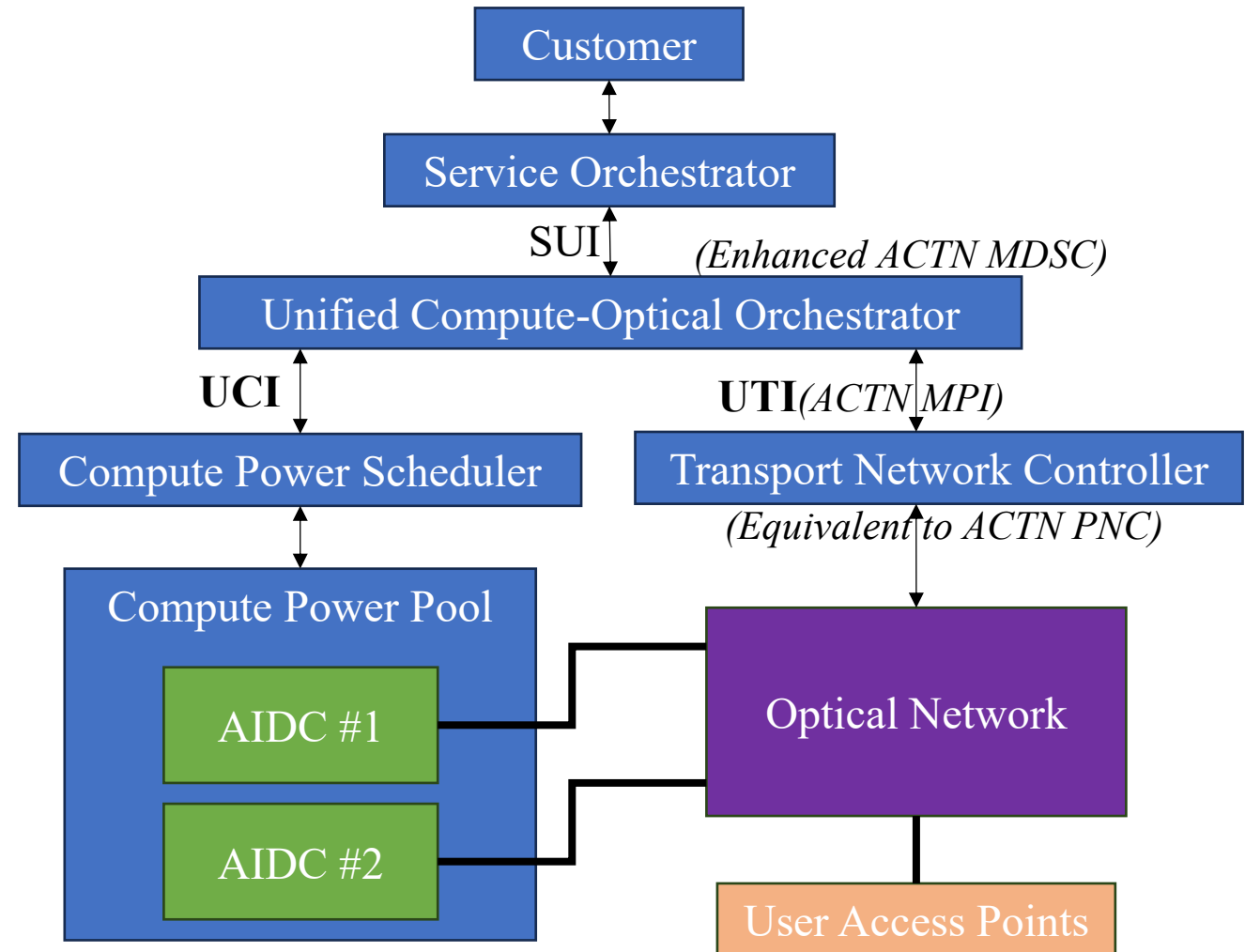
Horizontally integrate Compute and Network to break vertical silos.

➤ Unified Abstraction:

Model TE network + Compute/Storage resources together.

➤ Joint Orchestration:

Synchronize AI resource locking and optical path provisioning.



Next Step

- Next steps:
 - Improve the drafts based on WG comments and suggestions.
 - Refine the architecture, clarify the relationship with ACTN, and detail the use cases.
 - Update the draft names.
 - UONACO Interfaces and Yang Data Model Design
- Welcome more reviews, comments, and seek WG adoption.

Thank You