

# Why We Need a New Draft



**-Current network mechanisms are not enough for AI services, especially in cross-domain environments.**

- **Current mechanisms are too coarse-grained.** QoS and DiffServ can separate traffic classes, but they cannot describe fine-grained differences inside AI traffic.
- **They do not understand traffic meaning.** The network cannot tell whether a flow carries activations, gradients, KV cache, parameters, or other AI data with different importance and timing needs.
- **ECN is mainly a congestion signal, not a scheduling method.** It can indicate congestion, but it cannot tell which traffic should go first, which traffic can wait, or which traffic can be handled in another way.
- **The problem becomes more serious across domains.** Cross-domain RTT is longer, bandwidth is more limited, and computing resources are more heterogeneous, so wrong traffic handling has a much higher cost.

# How IntelliNode Uses the Semantics



-IntelliNode turns semantic information and network state into real-time traffic handling decisions.

## Semantic Information → Network Policy

- **The application provides a small set of useful semantics.** This may include traffic class, urgency, dependency hints, compute affinity, and whether the traffic can tolerate delay, buffering, or reduced fidelity.
- **The network uses these semantics as input for policy decisions.** Instead of treating all packets in the same way, the network can understand basic differences between flows and apply more suitable handling.
- **The contract connects traffic meaning to network actions.** It builds a clear mapping from semantic information to network treatment.

Examples of semantic information:

**traffic class / urgency / dependency hint / compute affinity**

Examples of network actions:

**priority / queueing / shaping / buffering / steering**

# What IntelliNode Is and How It Uses the Semantics



**-IntelliNode is an in-network control and scheduling framework that uses semantic information and real-time network state to support better traffic handling.**

- **IntelliNode is a network-side decision point.** It is designed to observe traffic and resource conditions inside the network and support more informed traffic handling for AI services.
- **It combines semantic information with network state.** IntelliNode uses traffic meaning together with queue status, link usage, and resource conditions, instead of relying only on basic packet forwarding or congestion feedback.
- **It supports real-time in-network actions.** Based on the current situation, IntelliNode can help with priority control, scheduling, shaping, buffering, and traffic steering.
- **This makes the network more active and more service-aware.** Rather than acting only as a forwarding plane, the network can better support mixed AI traffic, cross-domain delivery, and heterogeneous computing environments.

**The semantic contract tells the network what the traffic means, and IntelliNode helps the network decide what to do.**



# What We Plan to Do Next

**-The next step is to refine the semantic model, define the protocol details, and validate the benefit in real systems.**

---

- **Refine the semantic model.** Further identify the minimum set of information that should be shared between applications and the network.
- **Define how the semantics are carried.** Study practical ways to encode and transport this information, such as metadata, extension fields, or semantic headers.
- **Clarify the boundary between endpoint and network decisions.** Decide which actions should remain at the endpoint and which actions are better handled inside the network.
- **Build prototypes and evaluate the gain.** Verify whether the approach can improve traffic handling, resource coordination, and service efficiency in AI-oriented network scenarios.

**Our goal is to move from semantic-aware traffic description to practical in-network handling for AI services.**