



Draft 1: IDN (draft-li-cats-idn-00)

Draft 2: ODSI (draft-wang-cats-odsi-00)

Presenter: Hanling Wang (Pengcheng Laboratory)

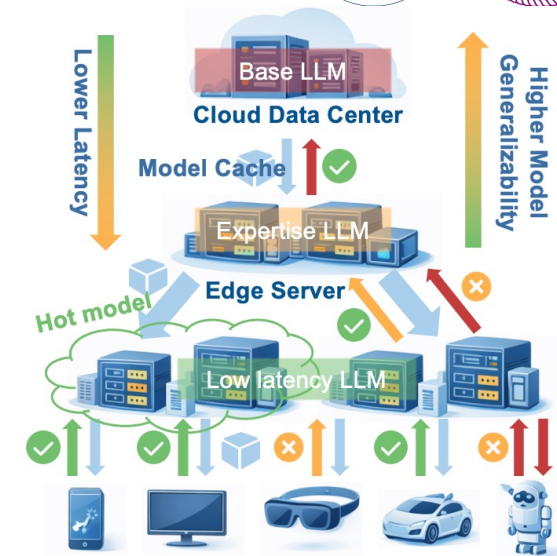
IETF 125 @ Shenzhen, March 2026

■ ■ ■ IDN (draft-li-cats-idn-00)



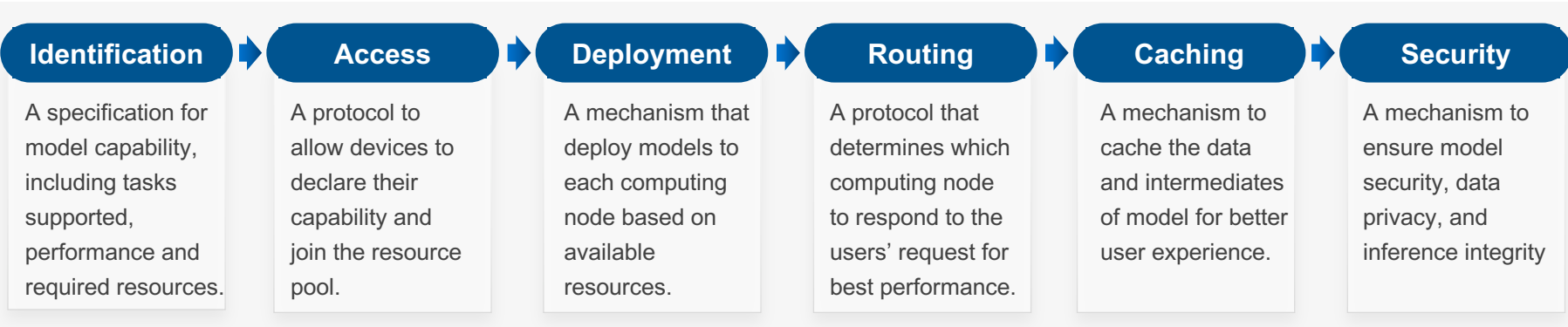
Submitted IETF draft (ongoing):

- **Title:** A Framework of Intelligence Delivery Network (IDN) for Deep Learning Inference
- **Link:** <https://datatracker.ietf.org/doc/draft-li-cats-idn/>
- **Author:** Qing Li, Hanling Wang, Yong Jiang, Mingwei Xu
- **Introduction:** Intelligence Delivery Network (IDN) is an architectural framework for deploying deep learning models across distributed cloud and edge nodes. Inspired by Content Delivery Network (CDN), IDN places model instances closer to users and dynamically select among heterogeneous models based on task requirements, latency, and system conditions.



Intelligence Delivery Network (IDN)

Components of the framework:



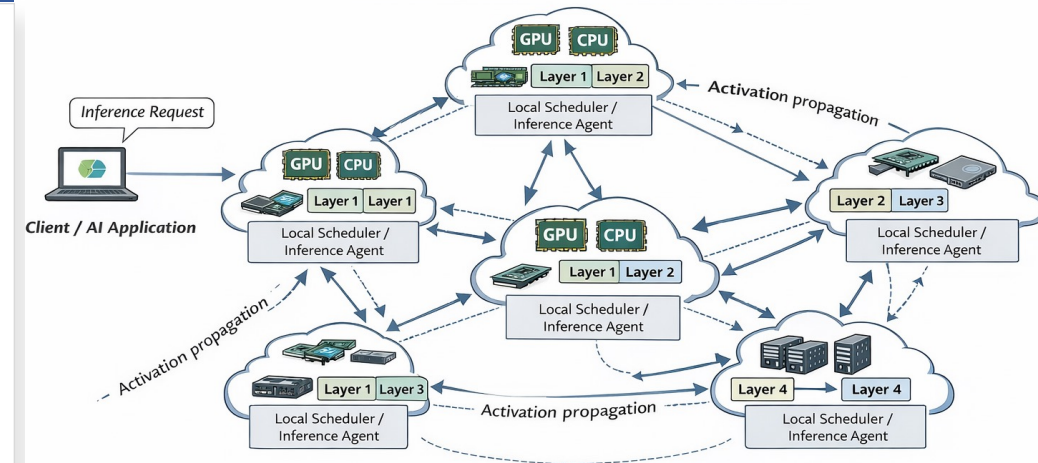
Any suggestions on

- Framework design?
- Protocol details?
- Practical application?
- Implementation?
-

ODSI (draft-wang-cats-odsi-00)

Submitted IETF draft (ongoing):

- **Title:** An Open, Decentralized, and Scalable Framework for Large Language Model Inference
- **Link:** <https://datatracker.ietf.org/doc/draft-wang-cats-odsi/>
- **Author:** Hanling Wang, Qing Li, Yong Jiang, Mingwei Xu
- **Introduction:** ODSI is a framework for executing LLM inference across independently operated and heterogeneous compute resources over the public Internet. It addresses challenges related to decentralization, coordination, scalability, and stateful execution under strict latency constraints, and discusses architectural principles and design considerations for distributed LLM inference.



Framework of ODSI



Layer-Aware Transport protocol

LLM inference is sequential, latency-sensitive, and deadline-driven. The transport protocol understands model layer boundaries, execution order, and per-layer deadlines.



Coordination Protocol

Peers differ in compute speed, memory capacity, network latency, and availability. Coordination must be decentralized, adaptive, and predictive.



Economic Protocol

Nodes are not altruistic. Compute and bandwidth have real cost. The economic protocol defines who is allowed to execute inference and how rewards are distributed.

Any suggestions on

- Framework design?
- Protocol details?
- Practical application?
- Implementation?
-



Thank you!