Distributed Inference Network (DIN):

Problem Statement, Use Cases, and Requirements

draft-song-rtgwg-din-usecases-requirements-00

Jian Song
songjianyjy@chinamobile.com

Weiqiang Cheng
chengweiqiang@chinamobile.com

IETF 124 RTGWG
2025-11

The AI Inference Challenge: problem statement

- **Shift:**
  - From “content access” to “model access”
  - From “chat” to “complex multi-modal interactions”
  - From “bit traffic” to “token traffic”

- **Massive Concurrency:** Billions of users, apps, IoT devices and AI agents
- **Stringent Latency:** Industrial control, autonomous systems, healthcare inferencing require *ms to sub-ms* level response
- **Enterprise Information Security:** Data cannot leave premises in finance, healthcare, public service sectors, etc.

The Distributed Inference Network (DIN)

- **Vision:** A distributed network architecture natively supporting and optimized for AI inference services

- **Goals:**
  - For Users: Inference everywhere at edge, with optimal experience
  - For Providers: Models on-demand, efficient at scale

Use Cases

- **4.1 Enterprise Secure Inference:**
  - Branches securely access HQ-based inference via encrypted overlays
  - AIoT equipment inspection, intelligent manufacturing, and real-time monitoring systems demand low-latency, high-reliability, and high-security inference services

- **4.2 Edge-Cloud Collaborative inference:**
  - Small and medium enterprises rent cloud capacity with on-premises compute for inference and fine-tuning
  - Seamlessly integrate local and cloud-based inference resources becomes crucial for maintaining service quality

Use Cases

- **4.3 Dynamic Model Selection**:
  - Intelligently route requests to automatically select between different model sizes, specialized accelerators, and geographic locations based on real-time factors including network conditions, computational requirements, accuracy needs, and cost considerations

- **4.4 Adaptive Resource Scheduling**:
  - Large-Small Model Collaboration: Large models for reasoning, small models for fast response
  - Prefill-Decode (PD) Separation: Distribute computational stages across specialized nodes

- **4.5 Privacy-Preserving Split Inference**
  - enable sensitive computational layers to execute on-premises while utilizing cloud resources for non-sensitive operations
  - for applications processing personal identifiable information, healthcare records, financial data, or proprietary business information subject to regulatory constraints

Technical Requirements

- **Scalability & Elasticity:** Billions of concurrent sessions.
- **Performance & Determinism:** Predictable low latency, jitter, ultra-low packet loss, and performance isolation.
- **Security & Privacy:** Full-stack data protection, including physical layer.
- **Identification & Scheduling:** Fine-grained workload ID and application-aware steering.
- **Management & Observability:** End-to-end telemetry for network and inference metrics (e.g. token latency, token throughput, inference efficiency) .

中国移动
China Mobile

Thank You for Listening!

www.10086.cn