rtgwg
Internet-Draft
Intended status: Informational
Expires: 23 April 2026

S. Jian
W. Cheng
China Mobile

20 October 2025

Distributed Inference Network (DIN) Problem Statement, Use Cases, and Requirements
draft-song-rtgwg-din-usecases-requirements-00

Abstract

This document describes the problem statement, use cases, and requirements for a "Distributed Inference Network" (DIN) in the era of pervasive AI. As AI inference services become widely deployed and accessed by billions of users, applications and devices, traditional centralized cloud-based inference architectures face challenges in scalability, latency, security, and efficiency. DIN aims to address these challenges by leveraging distributed edge-cloud collaboration, intelligent scheduling, and enhanced network security to support low-latency, high-concurrency, and secure AI inference services.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 23 April 2026.

Copyright Notice

Table of Contents

1. Introduction

AI inference is rapidly evolving into a fundamental service accessed by billions of users, applications, IoT devices, and AI agents.

The rapid advancement and widespread adoption of large AI models are introducing significant changes to internet usage patterns and service requirements. These changes present new challenges that existing network need to address to effectively support the growing demands of AI inference services.

First, internet usage patterns are shifting from primarily content access to increasingly include AI model access.

Users and applications are interacting more frequently with AI models, generating distinct traffic patterns that differ from traditional web browsing or streaming. This shift requires networks to better support model inference as an important service type alongside conventional content delivery.

Second, the interaction modalities are diversifying from simple human-to-model conversations to include complex multi-modal interactions.

As AI inference costs decrease dramatically, applications, IoT devices, and autonomous systems are increasingly integrating AI capabilities through API calls and embedded model access. This expansion creates unprecedented demands for high-concurrency processing and predictable low-latency responses, as these systems often require real-time inference for critical functions including autonomous operations, industrial control, and interactive services.

Third, AI inference workloads introduce distinct traffic characteristics that impact network design.

Both north-south traffic between users and AI services, and east-west traffic among distributed AI components, are growing significantly. Moreover, the nature of AI inference communication, often organized around token generation and processing, introduces new considerations for traffic management, quality of service measurement, and resource optimization that complement traditional bit-oriented network metrics.

These developments collectively challenge current network infrastructures to adapt to the unique characteristics of AI inference workloads. Centralized approaches face limitations in supporting the distributed, latency-sensitive, and concurrent nature of modern AI services, particularly in scenarios requiring real-time performance, data privacy, and reliable service delivery.

This document outlines the problem statement, use cases, and functional requirements for a Distributed Inference Network (DIN) to enable scalable, efficient, and secure AI inference services that can address these emerging challenges.

2. Conventions and Definitions

DIN: Distributed Inference Network

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.

3. Problem Statement

The proliferation of AI inference services has exposed fundamental limitations in traditional centralized AI inference architectures.

Centralized inference deployments face severe scalability challenges when handling concurrent requests from the rapidly expanding ecosystem of users, applications, IoT devices, and AI agents. Service providers have experienced recurrent outages and performance degradation during peak loads, with concurrent inference requests projected to grow from millions to billions. The fundamental constraint of concentrating computational resources in limited geographical locations creates inherent bottlenecks that lead to service disruptions and degraded user experience under massive concurrent access.

While human-to-model conversations may tolerate moderate network latency, the emergence of diverse interaction patterns including application-to-model, device-to-model, and machine-to-model communications imposes stringent low-latency requirements that centralized architectures cannot meet.

Applications including industrial robots, autonomous systems, and real-time control platforms require low-latency responses that are fundamentally constrained by the unavoidable geographical dispersion between end devices and centralized inference facilities. This architectural limitation creates critical barriers for delay-sensitive operations across manufacturing, healthcare, transportation, and other domains where millisecond to sub-millisecond-level response times are essential.

Enterprise and industrial AI inference scenarios present unique security and compliance requirements that fundamentally conflict with centralized architectural approaches.

Sectors including finance, healthcare, and public service sectors handle sensitive data subject to strict regulatory requirements that often mandate localized processing and data sovereignty. The transmission of confidential information, model parameters, and intermediate computational data across extended network paths to centralized inference pools creates unacceptable vulnerabilities and compliance violations. These fundamental constraints render centralized inference architectures unsuitable for numerous critical applications where data sovereignty, privacy protection, and regulatory compliance represent non-negotiable requirements.

4. Use Cases

4.1. Enterprise Secure Inference Services

Enterprises in regulated sectors such as finance, healthcare, industrial and public services require strict data governance while leveraging advanced AI capabilities. In this use case, inference servers are deployed at enterprise headquarters or private cloud environments, with branch offices and field devices accessing these services through heterogeneous network paths including dedicated lines, VPNs, and public internet connections.

The scenario encompasses various enterprise applications such as AIoT equipment inspection, intelligent manufacturing, and real-time monitoring systems that demand low-latency, high-reliability, and high-security inference services. Different network paths should provide appropriate levels of cryptographic assurance and quality of service while accommodating varying bandwidth and latency characteristics across the enterprise network topology.

The primary challenge involves maintaining data sovereignty and security across diverse network access scenarios while ensuring consistent low-latency performance for delay-sensitive industrial applications.

4.2. Edge-Cloud Collaborative Model Training

Small and medium enterprises often need to dynamically procure additional AI inference capacity while facing capital constraints for full-scale inference infrastructure deployment. This use case enables flexible resource allocation where businesses maintain core computational resources on-premises while dynamically procuring additional inference capacity from AI inference providers during demand peaks.

The hybrid deployment model allows sensitive data to remain within enterprise boundaries while leveraging elastic cloud resources for computationally intensive operations. As enterprise business requirements fluctuate, the ability to seamlessly integrate local and cloud-based inference resources becomes crucial for maintaining service quality while controlling operational costs.

The network should support efficient coordination between distributed computational nodes, ensuring stable performance during resource scaling operations and maintaining inference pipeline continuity despite variations in network conditions across different service providers.

4.3. Dynamic Model Selection and Coordination

The transition from content access to model inference access necessitates intelligent model selection mechanisms that dynamically route requests to optimal computational resources. This use case addresses scenarios where applications should 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.

The inference infrastructure should support real-time assessment of available resources, intelligent traffic steering based on application characteristics, and graceful degradation during resource constraints.

Key requirements include maintaining service continuity during model switching, optimizing the balance between response time and inference quality, and ensuring consistent user experience across varying operational conditions. This capability is particularly important for applications serving diverse user bases with fluctuating demand patterns and heterogeneous device capabilities.

4.4. Adaptive Inference Resource Scheduling and Coordination

The evolution from content access to model inference necessitates intelligent resource coordination across different computational paradigms. This use case addresses scenarios where inference workloads require adaptive resource allocation strategies to balance performance, cost, and efficiency across distributed environments.

Large-small model collaboration represents a key approach for balancing inference accuracy and response latency. In this pattern, large models handle complex reasoning tasks while small models provide efficient specialized processing, requiring the network to deliver low-latency connectivity and dynamic traffic steering between distributed model instances. The network should ensure efficient synchronization and coherent data exchange to maintain service quality across the collaborative ecosystem.

Prefill-decode separation architecture provides an optimized framework for streaming inference tasks. This pattern distributes computational stages across specialized nodes, with prefilling and decoding phases executing on optimized resources. The network should provide high-bandwidth connections for intermediate data transfer and reliable transport mechanisms to maintain processing pipeline continuity, enabling scalable handling of concurrent sessions while meeting real-time latency requirements.

The network infrastructure should support dynamic workload distribution, intelligent traffic steering, and efficient synchronization across distributed nodes. This comprehensive approach ensures optimal user experience while maximizing resource utilization efficiency across the inference ecosystem.

4.5. Privacy-Preserving Split Inference

For applications requiring strict data privacy compliance, model partitioning techniques enable sensitive computational layers to execute on-premises while utilizing cloud resources for non-sensitive operations. This approach is particularly relevant for applications processing personal identifiable information, healthcare records, financial data, or proprietary business information subject to regulatory constraints.

The network should support efficient transmission of intermediate computational results between edge and cloud with predictable performance characteristics to maintain inference pipeline continuity. Challenges include maintaining inference quality despite network variations, managing computational dependencies across distributed nodes, and ensuring end-to-end security while maximizing resource utilization efficiency across the partitioned model architecture.

5. Requirements

5.1. Scalability and Elasticity Requirements

Distributed Inference Network should support seamless scaling to accommodate billions of concurrent inference sessions while maintaining consistent performance levels. The network should provide mechanisms for dynamic discovery and integration of new inference nodes, with automatic load distribution across available resources. Elastic scaling should respond to diurnal patterns and sudden demand spikes without service disruption.

5.2. Performance and Determinism Requirements

AI inference workloads require consistent and predictable network performance to ensure reliable service delivery. The network should provide strict Service Level Agreement (SLA) guarantees for latency, jitter, and packet loss to support various distributed inference scenarios. Bandwidth provisioning should accommodate bursty traffic patterns characteristic of model parameter exchanges and intermediate data synchronization, with performance isolation between different inference workloads.

5.3. Security and Privacy Requirements

Comprehensive security mechanisms should protect AI models, parameters, and data throughout their transmission across network links. Cryptographic protection should extend to physical layer transmissions without introducing significant overhead or latency degradation. Privacy-preserving techniques should prevent leakage of sensitive information through intermediate representations while supporting efficient distributed inference.

5.4. Identification and Scheduling Requirements

The network should support fine-grained identification of inference workloads to enable appropriate resource allocation and path selection. Application-aware networking capabilities should allow inference requests to be steered to optimal endpoints based on current load, network conditions, and computational requirements. Both centralized and distributed scheduling approaches should be supported to accommodate different deployment scenarios and organizational preferences.

5.5. Management and Observability Requirements

The network should provide comprehensive telemetry for performance monitoring, fault detection, and capacity planning. Metrics should include inference-specific measurements such as token latency, throughput, and computational efficiency in addition to traditional network performance indicators. Management interfaces should support automated optimization and troubleshooting across the combined compute-network infrastructure.

6. Security Considerations

This document highlights security as a fundamental requirement for DIN. The distributed nature of inference workloads creates new attack vectors including model extraction, data reconstruction from intermediate outputs, and adversarial manipulation of inference results. Security mechanisms should operate at multiple layers while maintaining the performance characteristics necessary for efficient inference. Physical layer encryption techniques show promise for protecting transmissions without the overhead of traditional cryptographic approaches.

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 the contributors from China Mobile Research Institute for their valuable inputs and discussions.

Authors' Addresses

Song Jian
China Mobile
Email: songjianyjy@chinamobile.com

Weiqiang Cheng
China Mobile
Email: chengweiqiang@chinamobile.com