Considerations of network/system for AI services

draft-irtf-nmrg-ai-deploy-01

Y-G. Hong (Daejeon Univ.), J-S. Youn (DONG-EUI Univ.),
S-W. Hong (ETRI), P. Martinez-Julia (NICT), Q. Wu (Huawei)

nmrg Meeting@IETF 123 – Madrid
July 25. 2025

History and status

- 00 version I-D : draft-hong-nmrg-ai-deploy-00 (Mar. 2022)
- **RG adoption call (7/31/2024 ~ 8/21/2024)**
- 7th revision I-D : draft-hong-nmrg-ai-deploy-07 (Oct. 2024)

- **00 version R-D : draft-irtf-nmrg-ai-deploy-00 (Nov. 2024)**
- 01 version R-D : draft-irtf-nmrg-ai-deploy-01 (Mar. 2025)

Object of this draft

- Share experiences and implementation results to find optimal network/system for AI services
  - To find what is **important information** to provide optimal AI services
  - To find how to **deliver these information** between related devices
  - To find how to **manage these information**
- Find common components to provide optimal AI services
  - Common information (similar to MIB)
  - Common system to provide AI services
  - Common network architecture to provide AI services
  - Common protocols to exchange information for AI services
- Find useful use cases
  - Self-driving cars
  - Network digital twin

Abstract

- As the development of AI technology has matured and AI technology has begun to be applied in various fields, AI technology is changing from running only on **very high performance servers** to **commodity servers with small affordable hardware, including microcontrollers, low performance CPUs, and AI chipsets**.
- This document considers **how to configure the network and system in terms of AI inference service to provide AI service in a distributed manner**.
- It also describes the points to consider **in the environment where a client connects to a cloud server and an edge device and requests an AI service**. Some use cases of deploying network-based AI services, such as **self-driving vehicles and network digital twins**, are described.

Table of Contents

Table of Contents

Updates since last meeting (1/5)

- Add/modify text to focus the addressing challenges
  - 3.3. AI inference service on Edge device
  - 3.5. AI inference service on horizontal multiple servers
  - 4.3. Considerations for the characteristics of the communication method
  - 6.2. Deploying AI services for network digital twins
- Add Qin Wu as a authors
- Fix typos

Updates since last meeting (2/5)

- 3.3. AI inference service on Edge device

The following figure shows the case where the client module that requests AI service and the AI server module that directly performs AI service are separated, and the AI server module is located in the edge device. As more data gets generated at the edge, it is necessary to perform AI processing at the edge for performance, privacy and cost of moving data to a central location, e.g., some enterprises who owns both local machine and Edge network won't be comfortable uploading their confidential data sets into the central cloud, instead, these enterprises retrieve AI model from the Cloud and use such AI model for AI inference at the edge device locally.

Updates since last meeting (3/5)

- 3.5. AI inference service on horizontal multiple servers

In this network configuration, AI service may have different performance according to the load level of the server, computing capability of the server machine and link-state between the local machine and the server machines of the horizontal level. Thus, to look for the server machine that can support the best AI service, it is necessary for the network element that can monitor network link-state and current state of the computing capability of the server machines and the network load-balance that can perform a scheduling policy of load balancing. Therefore, the client requesting the AI inference service, those Requests are shared between the Edge devices, Cloud servers and routed to a single Cloud Server machine in the network based on the load balancer's decision. Alternatively, Those Requests are processed by all the Cloud Server machines which contribute to the final inference result.

Updates since last meeting (4/5)

- 4.3. Considerations for the characteristics of the communication method

The communication method for transferring data to request AI inference service is also an important decision in constructing an AI system. Using the traditional REST method, it can be used for various machines and services, but its performance is inferior to gRPC. There are many advantages to using gRPC for AI inference services because gRPC uses HTTP/2 and Protocol Buffers for communication, offering low latency and efficient data serialization and enabling large-capacity data transfer and efficient data transfer compared to REST. Alternatively, QUIC protocol or other future transport protocol also can be used to request AI inference service. Which transport protocol is used is not in the scope of this document.

Updates since last meeting (5/5)

- 6.2. Deploying AI services for network digital twins

Network digital twin also need to build distributed AI services. The purpose of a network digital twin is described in [I-D.irtf-nmrg-network-digital-twin-arch]. In particular, the network digital twin provides network operators with technology that enables data driven network management and allows real time interaction between physical network and twin network. To achieve this, the network digital twin will use AI capabilities for various purposes such as scenario planning, impact analysis and change management.

Various AI functions will be applied for optimal network operation and management. However, the actual physical network consists of various different network devices and has a complex structure for various different type of data such as topology data, configuration data, state data. In addition, in a large-scale network environment, the network overhead is very large and expensive to collect and store information from many network devices in a centralized manner, and to create and change network management policies based on it.

Next steps

- Add more text from experiments/experience

- Asking NMRG to guide this draft

Thanks!!

Questions to NMRG