Considerations of network/ system for AI services

draft-hong-nmrg-ai-deploy-07

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

Joint meeting NMRG-ETSI ZSM@IETF 121 – Dublin

November 9. 2024

History and status

- 00 : draft-hong-nmrg-ai-deploy-00 (Mar. 2022)
- 1st revision : draft-hong-nmrg-ai-deploy-01 (Jul. 2022)
  - 1st presentation
- 2nd revision : draft-hong-nmrg-ai-deploy-02 (Oct. 2022)
- 3rd revision : draft-hong-nmrg-ai-deploy-03 (Mar. 2023)
  - Updated by comments by Alexander Clemm, Jeff Tantsura, Jeferson Campos Nobre
- 4th revision : draft-hong-nmrg-ai-deploy-04 (Jul. 2023)
  - Updated to reflect the use case of digital twin networks
- 5th revision : draft-hong-nmrg-ai-deploy-05 (Oct. 2023)
  - Updated to reflect the use case of digital twin networks and self-driving car
- 6th revision : draft-hong-nmrg-ai-deploy-06 (Jul. 2024)
  - Asking for RG adoption
- **RG adoption call (7/31 ~ 8/21)**
  - **Feedback from 8 persons in mailing list**
- **7th revision : draft-hong-nmrg-ai-deploy-07 (Oct. 2024)**
  - **Updated for addressing challenges for coupling AI and NM**

RG adoption call

- Time : July 31 ~ August 21

- Response from 8 persons
  - Jaehoon (Paul) Jeong
  - Pedro Martinez-Julia
  - PS Kim
  - Younghwan Choi
  - Minsuk Kim
  - Hyunjeong Lee
  - Bien Aime
  - Thi Nguyen

Updates since last meeting

- Add a new section to describe the addressing challenges
  - 5. Addressing challenges for coupling AI and NM
    - 5.1. Low-level challenges
    - 5.2. High-level challenges

- Add Pedro Martinez-Julia as a authors

- Fix typos

Motivations

– In the charter of NMRG

For Artificial Intelligence in Network Management (AI-NM):

1. Investigate, organize and document the major research challenges in AI for Network Management.
   Goal: provide a reference document which defines the different forms and usages of AI in network management and articulates the different goals, challenges, requirements and research directions.

Research Challenges in Coupling Artificial Intelligence and Network Management (draft-irtf-nmrg-ai-challenges)

2. Organize and animate a series of practical Network Management AI challenges/competitions.
   Goal: promote experimental research, practical knowledge and validation of AI techniques to solve network management problems and foster exchanges and cross-participation of both AI and Network Management specialists.

3. Support discussion and collaboration on techniques, (meta-)data, experimentations and best practises for the use and integration of AI with networking management approaches.
   Goal: offer a forum for the Network Management AI community to report on advances, developments and key results and introduce its efforts to the IETF. Note: Applicability of AI techniques for IBN functionalities and mechanisms is an example of potential joint activity between the Network Management AI and IBN realms.

[Source: NMRG charter]

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

Table of Contents

Table of Contents

Network configuration structure to provide AI services

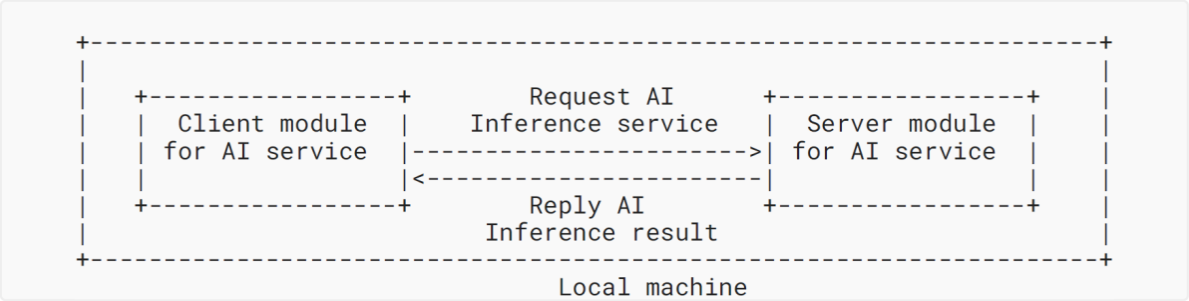

Figure 2: AI inference service on Local machine

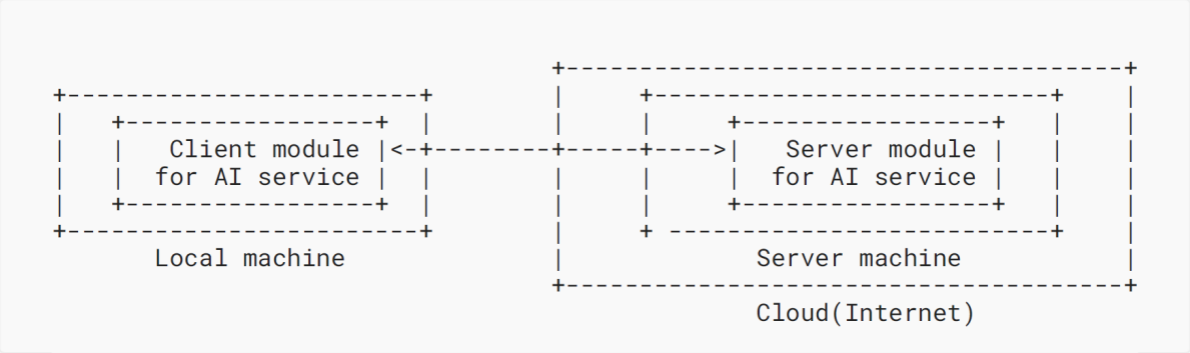

Figure 3: AI inference service on Cloud server

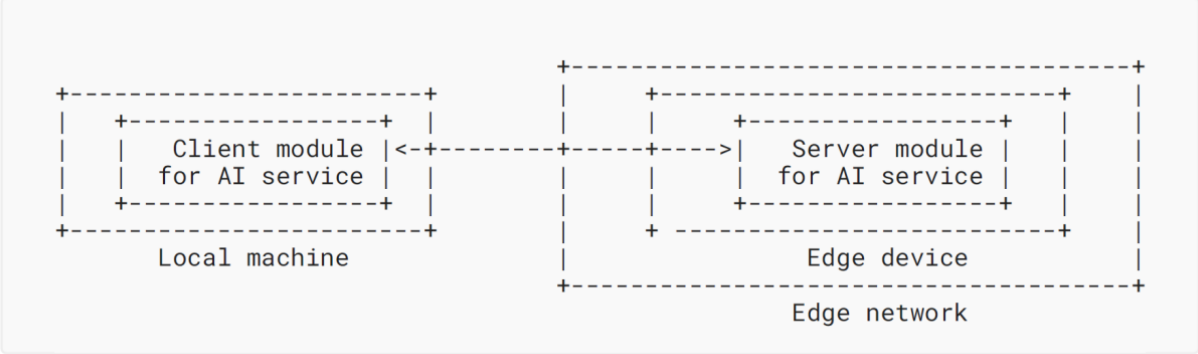

Figure 4: AI inference service on Edge device

AI inference service on vertical/horizontal servers

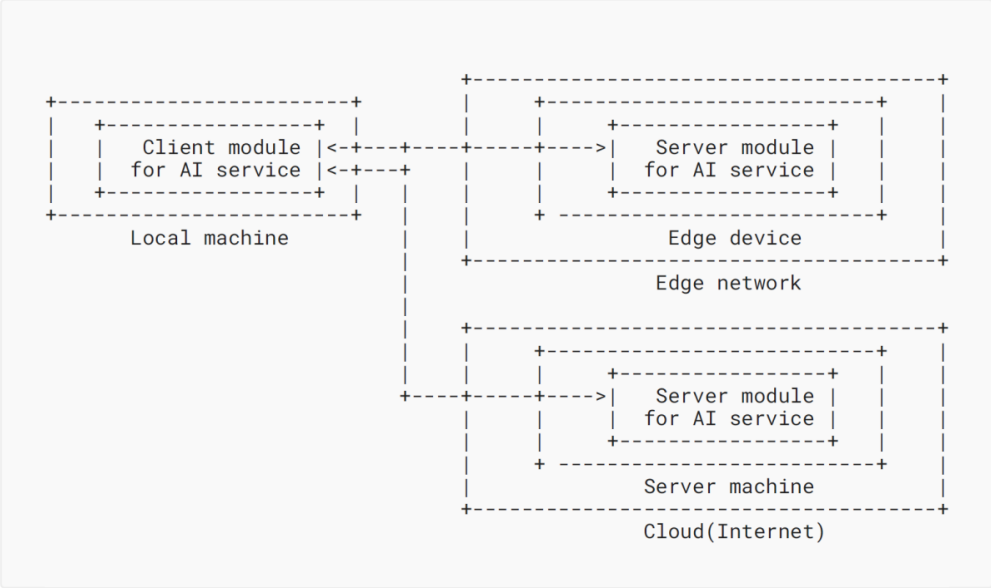

Figure 5: AI inference service on Cloud sever and Edge device

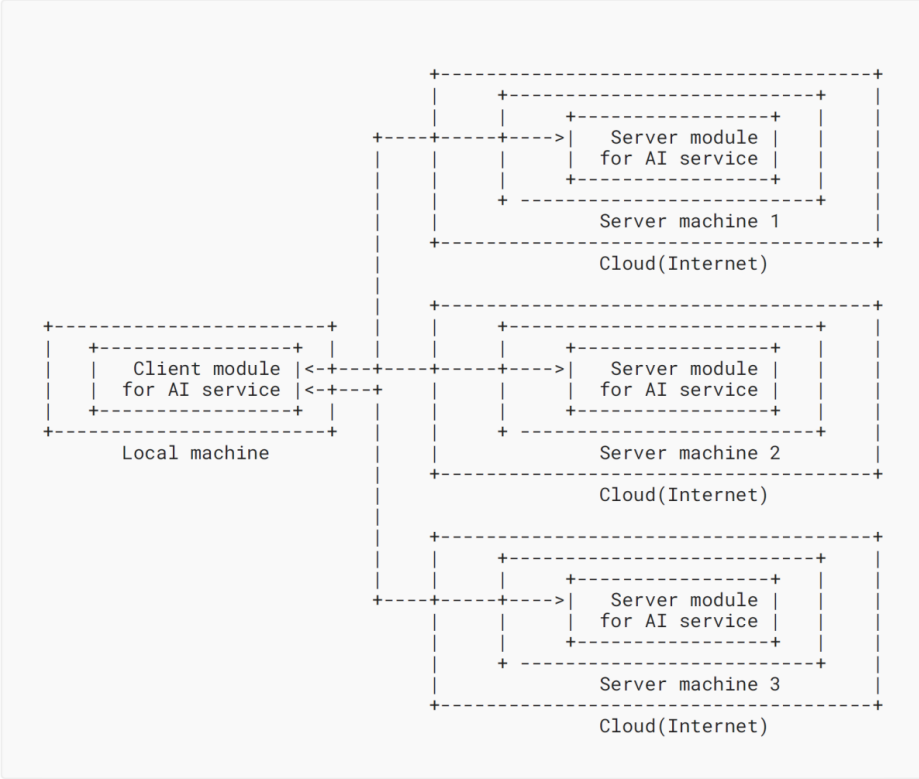

Figure 6: AI inference service on horizontal multiple servers

Network-side utilization for AI service

- Collecting and preprocessing of data and training an AI model
  - Federating learning : a machine learning technique that trains an AI model across multiple decentralized servers
    - It enables multiple network nodes to build a common machine learning model.
  - Transfer learning : a machine learning technique that focuses on storing information gained while solving one problem and applying it to a different but related problem.
    - We can utilize a network configuration to transfer common information and knowledge between different network nodes.

Considerations of the functional characteristics of the hardware

– The performance of AI inference service varies depending on how the hardware such as CPU, RAM, GPU, and network interface is configured for each cloud server and edge device.

– AI inference service can be deployed in the following locations
  - Distant cloud server : High performance and high cost
  - Near edge device : Medium performance and medium cost
  - Local machine : Low performance and low cost

– AI inference service result in (assumption: same AI model)
  - Distant cloud server : High accuracy, short inference time, and long delay to transmit
  - Near edge device : Medium accuracy, medium inference time, and medium delay to transmit
  - Local machine : Low accuracy, long inference time, and short delay to transmit

Considerations of the characteristics of the AI model

- AI inference service can be deployed in the following locations
  - Distant cloud server : Heavy AI model, high accuracy, Big size, long inference time
  - Near edge device : Medium AI model, medium accuracy, medium size, medium inference time
  - Local machine : Light AI model, low accuracy, small size, short inference time

- AI inference serving framework
  - Traditional web server : ex) FastAPI, Flask, and Django
    - It can be operated on low performance machines
  - Specialized serving framework : ex) Tensorflow serving
    - It can provide high performance.

Considerations of the characteristics of the communication method

- AI inference service can be utilized
  - Traditional REST method
    - Common and easily deployed
  - Specified communication method (e.g., gRPC)
    - Better performance but need some works

- AI inference data can be classified
  - Real-time vs. Batch
  - Secure & non-secure

Use cases of deploying AI services in a distributed method

Deploying AI services in Self-driving car

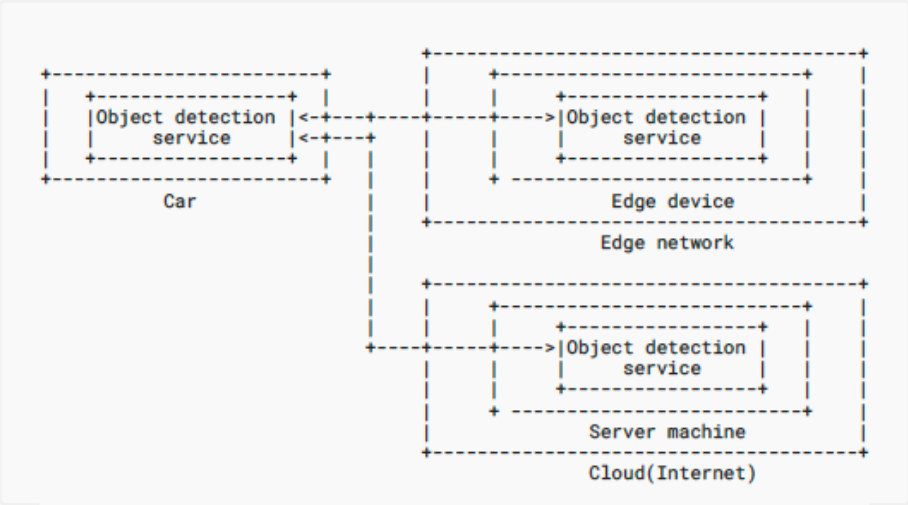

Figure 7: Distributed object detection service in self-driving car

Deploying AI services in NDT

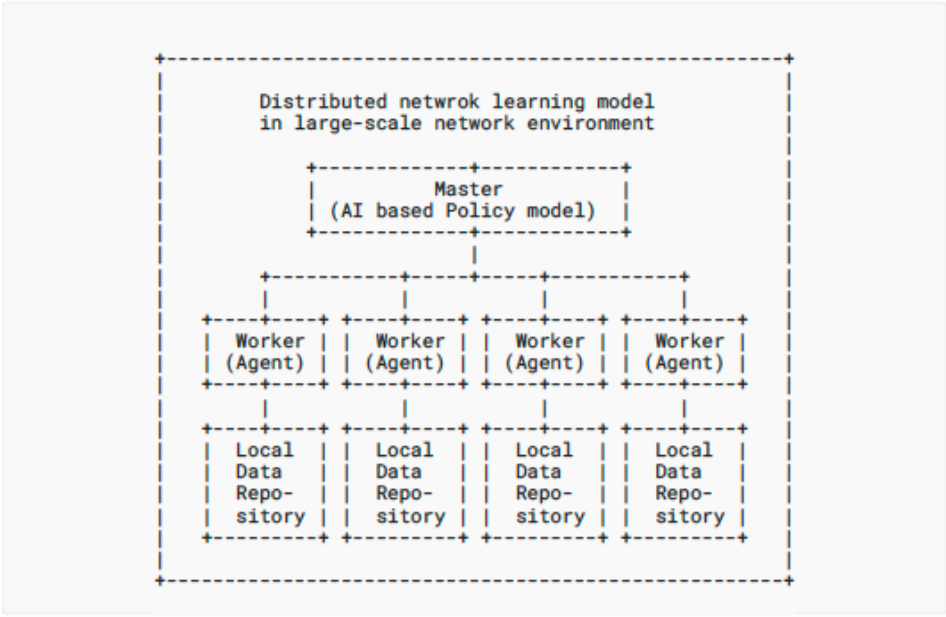

Figure 8: Distributed learning model of network learning for Digital twin network

Addressing challenges for coupling AI and NM(1/5)

- Difficult problems in network management in the challenge doc.
  - C1: A very large solution space, combinatorially exploding with the size of the problem domain. This makes it impractical to explore and test every solution
  - C2: Uncertainty and unpredictability along multiple dimensions, including the context in which the solution is applied, behavior of users and traffic, lack of visibility into network state, and more.
  - C3: The need to provide answers (i.e. compute solutions, deliver verdicts, make decisions) in constrained or deterministic time. In many cases, context changes dynamically and decisions need to be made quickly to be of use.
  - C4: Data-dependent solutions. To solve a problem accurately, it can be necessary to rely on large volumes of data, having to deal with issues that range from data heterogeneity to incomplete data to general challenges of dealing with high data velocity.
  - C5: Need to be integrated with existing automatic and human processes.
  - C6: Solutions MUST be cost-effective as resources (bandwidth, CPU, human, etc.) can be limited, notably when part of processing is distributed at the network edge or within the network.

Addressing challenges for coupling AI and NM(2/5)

- Addressing low-level challenges
  - The first challenge (C1) concerns to the combinatorial explosion of the solution space in relation to the size of the problem.
    - This draft proposes to tackle part of such complexity by separating the AI work in multiple elements of the network. The separation is asymmetrical, so that some elements (cloud side) will contribute more computation power to the distributed service.
  - The second challenge (C2) regards the dimensional and context uncertainty and unpredictability.
    - This draft focuses on cloistering the AI models and data within pre-defined client-edge-cloud-server structures.
  - The third challenge (C3) is to ensure that AI to is able to provide prompt responses and decisions to management questions.
    - The key aspect discussed in this draft to resolve this challenge is the integration of end-node, in-network, edge, a cloud computing services.
    - This allows AI computations to be performed in the best place possible.

Addressing challenges for coupling AI and NM(3/5)

- Addressing low-level challenges
  - The fourth challenge (C4) states the difficulties related to resolving data imperfections and scalability of techniques.
    - This challenge is specific to each information domain and problem.
    - Instantiating a distributed AI system in a cloud continuum enables the AI system to have many functions to deal with data homogenization, resolving high data rates, etc.
  - The fifth challenge (C5) concerns the integration of AI services with existing automation and human processes.
    - The flexibility of the structures presented in this draft allow them to be connected to existing systems, being aligned and somewhat interconnected with AINEMA [I-D.pedro-nmrg-ai-framework].
  - The sixth challenge (C6) exposes the need for cost-effective solutions.
    - Enabling AI services to be deployed in the cloud continuum supports the maximization of effectiveness by dynamically relocating services to the cheapest provider point that is able to accomplish the computation requirements.

Addressing challenges for coupling AI and NM(4/5)

- Addressing high-level challenges
  - The first challenge (H1) relies in the observation that AI techniques were developed in a different area---imaging---, so they must be extensively tailored for network problems.
    - Although this is a quite complex challenge, this draft supports its resolution by enabling AI models and algorithms to evolve separately.
  - The second challenge (H2) conveys the mismatch that exists from the original data and internal data used in AI models.
    - Although the present document does cover this mismatch, the structures presented in this draft help alleviate the burden by relocating some AI processes as close as possible to the end-points (e.g., vehicles and NDT).

Addressing challenges for coupling AI and NM(5/5)

- Addressing high-level challenges
  - The third challenge (H3) consists on the level of acceptance that an AI system experiences from administrators and operators.
    - It is agreed that giving full control of AI operations to administrators and operators increases such level of acceptance.
    - The structures presented in this draft support the involvement of administrators and operators in AI system processes through the provision of policies and network intents.

Thanks!!

Questions to NMRG