Considerations of deploying AI services in a distributed approach

draft-hong-nmrg-ai-deploy-01

Y-G. Hong (Daejeon Univ.), S-B. Oh (KSA), J-S. Youn (DONG-EUI Univ),
S-J. Lee (Korea University/KT), H-K. Kahng (Korea University)

nmrg Meeting@IETF 114 – Philadelphia

July 27. 2022

History and status

- 1st revision : draft-hong-nmrg-ai-deploy-00 (Mar. 2022)
- **2nd revision : draft-hong-nmrg-ai-deploy-01 (Jul. 2022)**
  - 1st presentation

Motivations

- Change of the deployment of AI services
  - Focus : training (learning) -> inference (prediction)
  - For inference, not only high-performance servers, but also small hardware, microcontroller, low-performance CPUs, and AI chipsets are optimal target device (due to cost)
- Configuration of the system in terms of AI inference service
  - For training : accuracy of the model
  - For inference :
    - Target device : Local, edge, cloud
    - Objectives : Accuracy, Latency, Network traffic, Resource utilization, etc.
    - Considerations : AI model, Serving framework, Communication method, device capacity, inference data, etc.
- Accelerate the study AI issues in the nmrg

Generic procedure of AI service

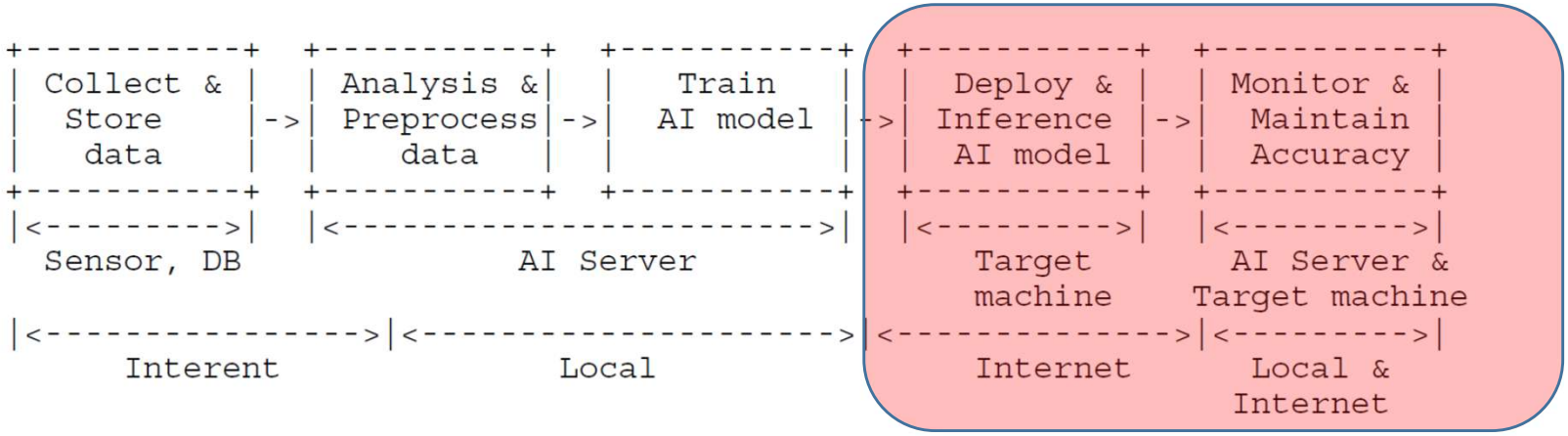


Figure 1: AI service workflow

- Data collection & Store
- Data Analysis & Preprocess
- AI Model Training
- AI Model Deploy & Inference
- Monitor & Maintain Accuracy

Network configuration structure to provide AI services

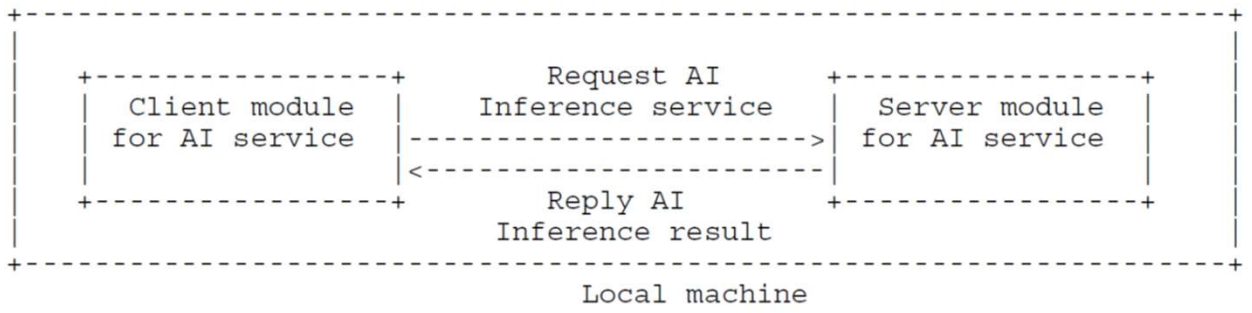


Figure 2: AI inference service on Local machine

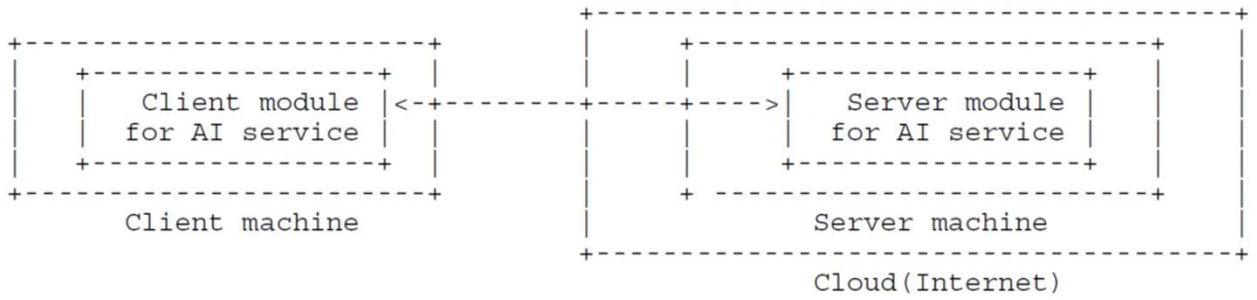


Figure 3: AI inference service on Cloud server

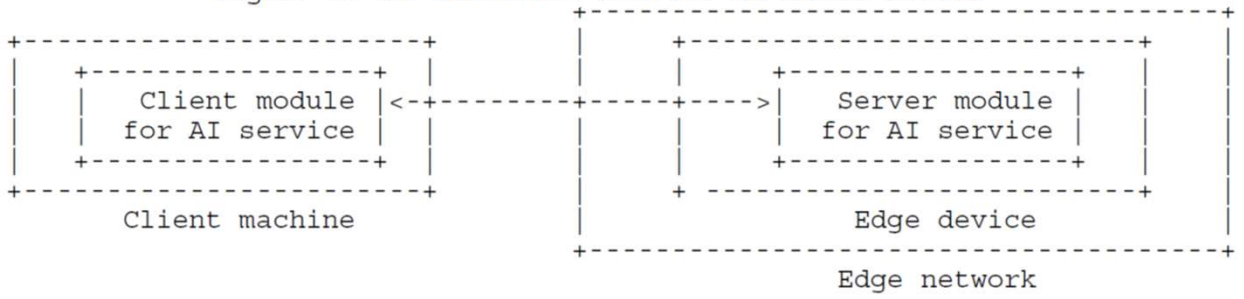


Figure 4: AI inference service on Edge device

AI inference service on Cloud server and Edge device

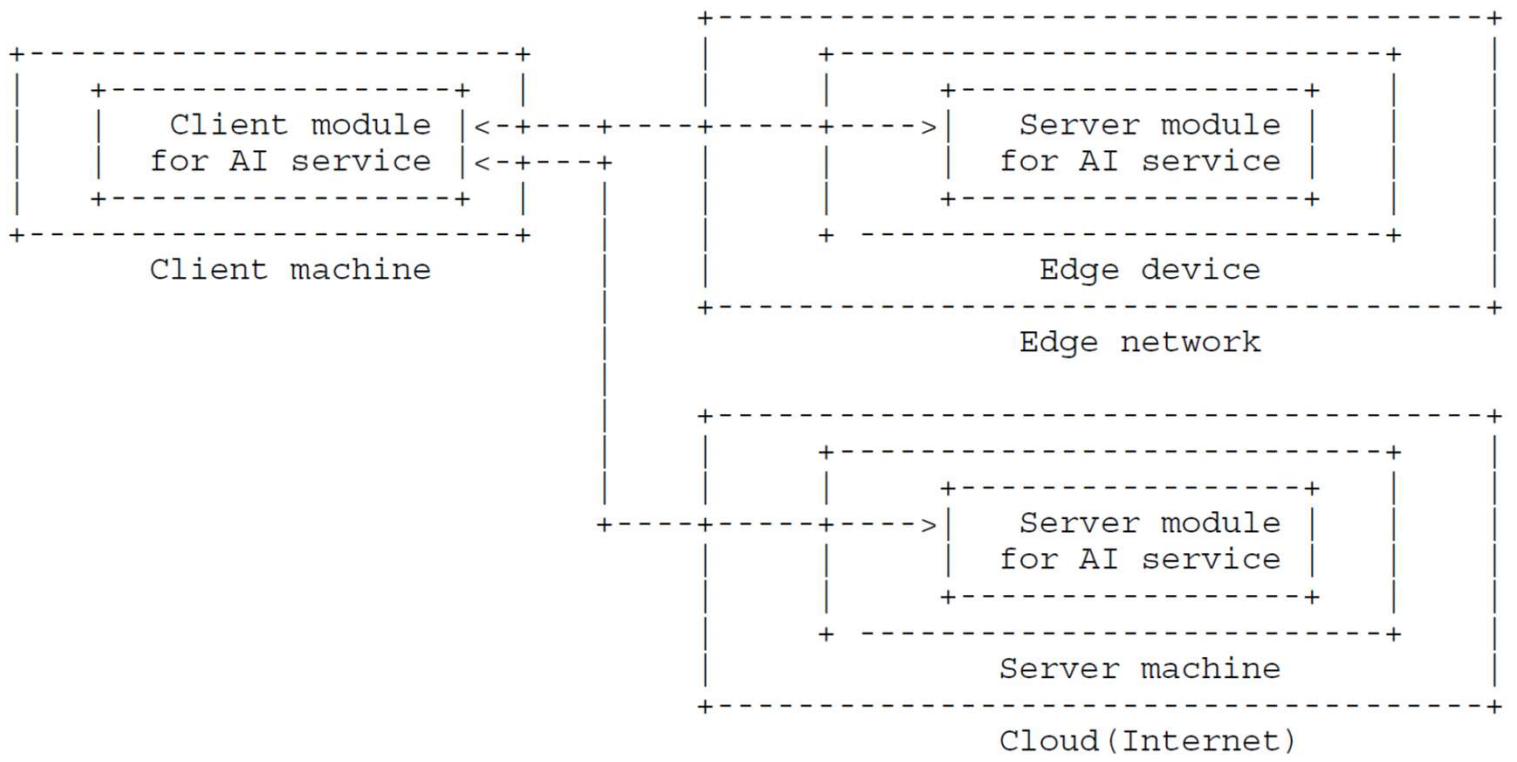


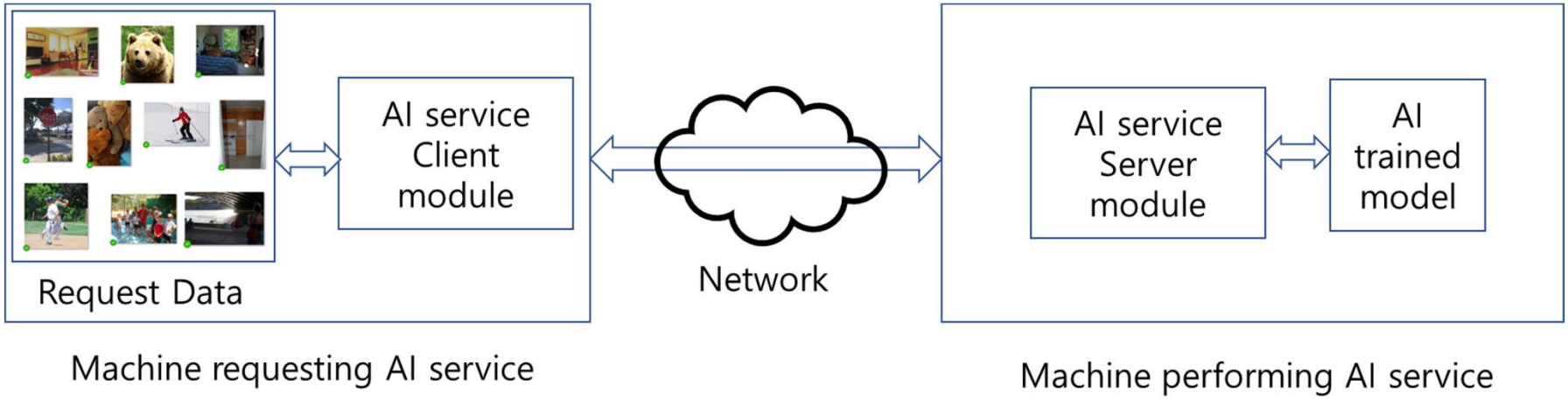
Figure 5: AI inference service on Cloud sever and Edge device

Considerations of deploying AI services in a distributed approach

- Objectives of AI services
  - Accuracy of model
  - Latency of IoT service
  - Network traffic
  - Resource utilization

- Considerations of deploying AI services
  - AI model (heavy vs. lightweight)
  - Serving framework (Web vs. Serving-targeted)
  - Communication method (REST vs. gRPC)
  - Machine capacity (CPU, RAM, etc.)
  - Inference data (realtime vs. batch, secure & non-secure, etc.)

An example of AI system for Object detection services



Latency of object detection services in each device

System Information

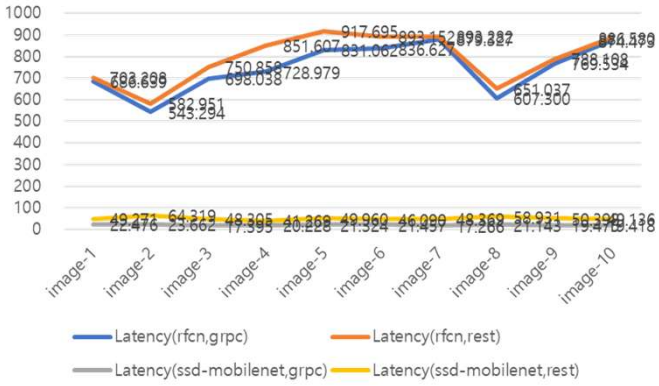
System Information	
Operating System	Ubuntu 20.04.3 LTS
Model	LG Electronics 14TD90P-GX70K
Motherboard	LG Electronics 14T90P
CPU Information	
Name	Intel Core i7-1165G7
Topology	1 Processor, 4 Cores, 8 Threads
Base Frequency	4.70 GHz
L1 Instruction Cache	32.0 KB x 4
L1 Data Cache	48.0 KB x 4
L2 Cache	1.25 MB x 4
L3 Cache	12.0 MB x 1
Memory Information	
Memory	15.44 GB

Geekbench 5 Score

1660	5617
Single-Core Score	Multi-Core Score

Geekbench 5.4.4 Tryout for Linux x86 (64-bit)

<Local device>



System Information

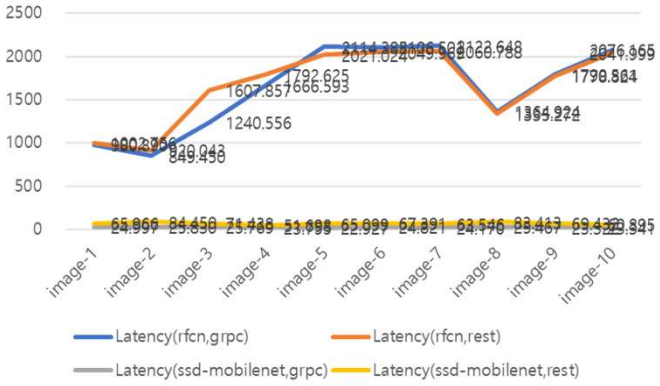
System Information	
Operating System	Ubuntu 20.04.3 LTS
Model	LENOVO 20U9S19800
Motherboard	LENOVO 20U9S19800
CPU Information	
Name	Intel Core i7-10510U
Topology	1 Processor, 4 Cores, 8 Threads
Base Frequency	4.90 GHz
L1 Instruction Cache	32.0 KB x 4
L1 Data Cache	32.0 KB x 4
L2 Cache	256 KB x 4
L3 Cache	8.00 MB x 1
Memory Information	
Memory	15.30 GB

Geekbench 5 Score

1175	3589
Single-Core Score	Multi-Core Score

Geekbench 5.4.4 Tryout for Linux x86 (64-bit)

<Edge device>



System Information

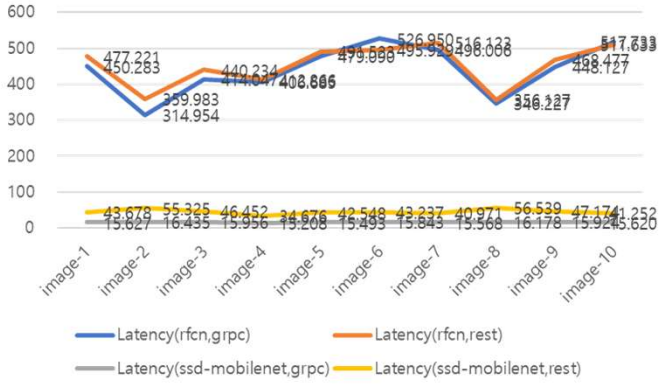
System Information	
Operating System	Ubuntu 20.04.3 LTS
Model	ASUS System Product Name
Motherboard	ASUSTeK COMPUTER INC. TUF GAMING Z490-PLUS
CPU Information	
Name	Intel Core i7-10700K
Topology	1 Processor, 8 Cores, 16 Threads
Base Frequency	5.10 GHz
L1 Instruction Cache	32.0 KB x 8
L1 Data Cache	32.0 KB x 8
L2 Cache	256 KB x 8
L3 Cache	16.0 MB x 1
Memory Information	
Memory	94.19 GB

Geekbench 5 Score

1465	8078
Single-Core Score	Multi-Core Score

Geekbench 5.4.4 Tryout for Linux x86 (64-bit)

<Cloud server>



Thanks!!

Questions & Comments