Considerations of deploying AI services in a distributed approach

draft-hong-nmrg-ai-deploy-02

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)
S-W. Hong (ETRI), H-S. Yoon (ETRI)

nmrg Meeting@IETF 115 – London

November 7. 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
- **3rd revision : draft-hong-nmrg-ai-deploy-02 (Oct. 2022)**
  - **2nd presentation**

Updates after last meeting

- Reconfigure section 4. “Considerations for configuring a system to provide AI services”
  - 4.1. Considerations according to the functional characteristics of the Hardware
  - 4.2. Considerations according to the characteristics of the AI model
  - 4.3. Considerations according to the characteristics of the communication method

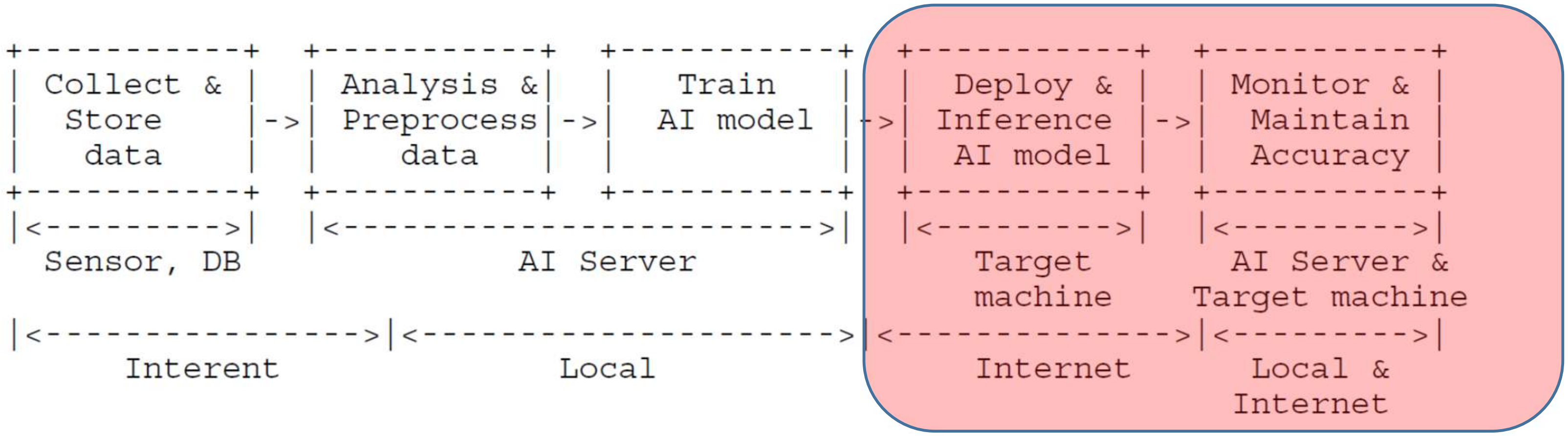
- Add a reference
  - ETSI "Mobile Edge Computing; Market Acceleration; MEC Metrics Best Practice and Guidelines" Group Specification ETSI GS MEC-IEG 006 V1.1.1 (2017-01)

- Add two authors
  - S-W. Hong (ETRI)
  - H-S. Yoon (ETRI)

Motivations

- 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 and find some possible standardization items in the nmrg

Generic procedure of AI service



```
+-----------+  +-----------+  +-----------+  +-----------+  +-----------+
| Collect & |  | Analysis &|  |   Train   |  |  Deploy & |  | Monitor & |
|   Store   |->| Preprocess|->|  AI model |->| Inference |->|  Maintain |
|    data   |  |    data   |  |           |  |  AI model |  |  Accuracy |
+-----------+  +-----------+  +-----------+  +-----------+  +-----------+
|<--------->|  |<------------------------->|  |<--------->|  |<--------->|
  Sensor, DB             AI Server              Target       AI Server &
                                                machine     Target machine

|<--------------->|<--------------------->|<-------------->|<--------->|
      Interent               Local              Internet       Local &
                                                              Internet
```

Figure 1: AI service workflow

```
o  Data collection & Store

o  Data Analysis & Preprocess

o  AI Model Training

o  AI Model Deploy & Inference

o  Monitor & Maintain Accuracy
```

Network configuration structure to provide AI services

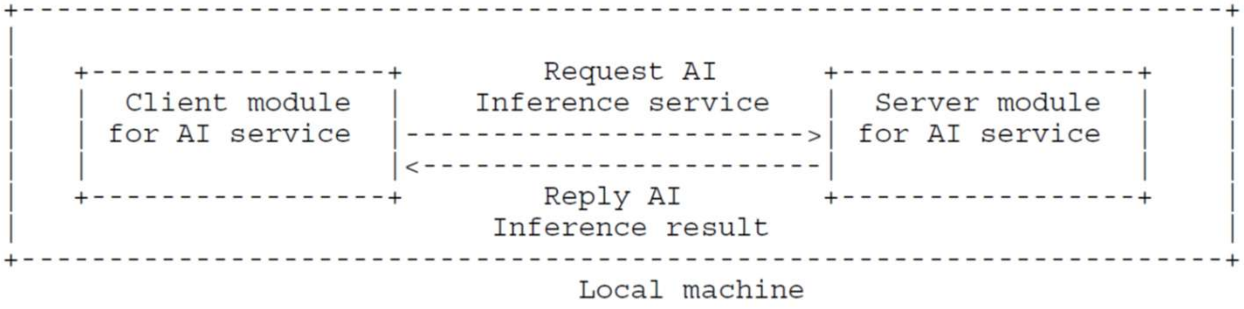

Client module for AI service
Request AI Inference service
Server module for AI service
Reply AI Inference result
Local machine

Figure 2: AI inference service on Local machine

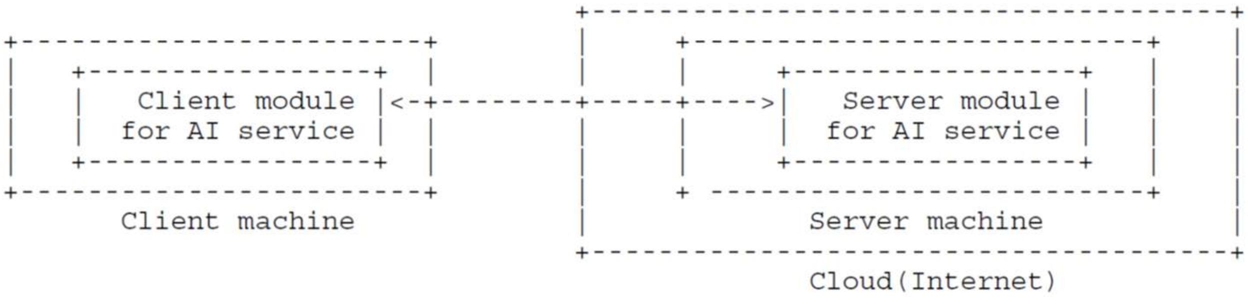


Figure 3: AI inference service on Cloud server

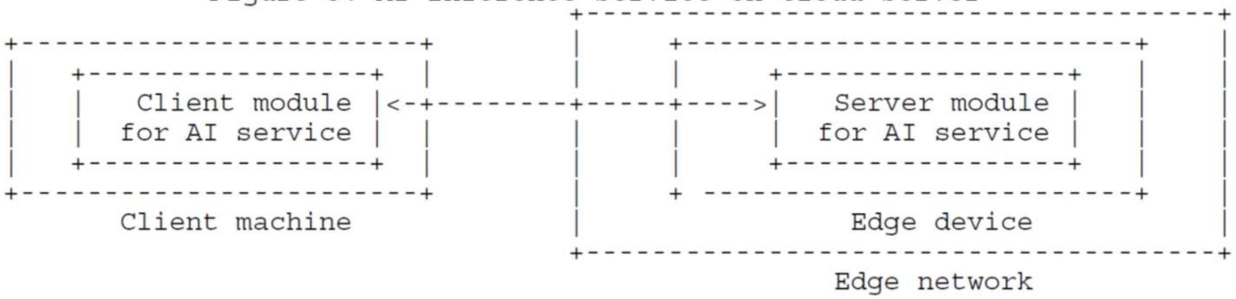
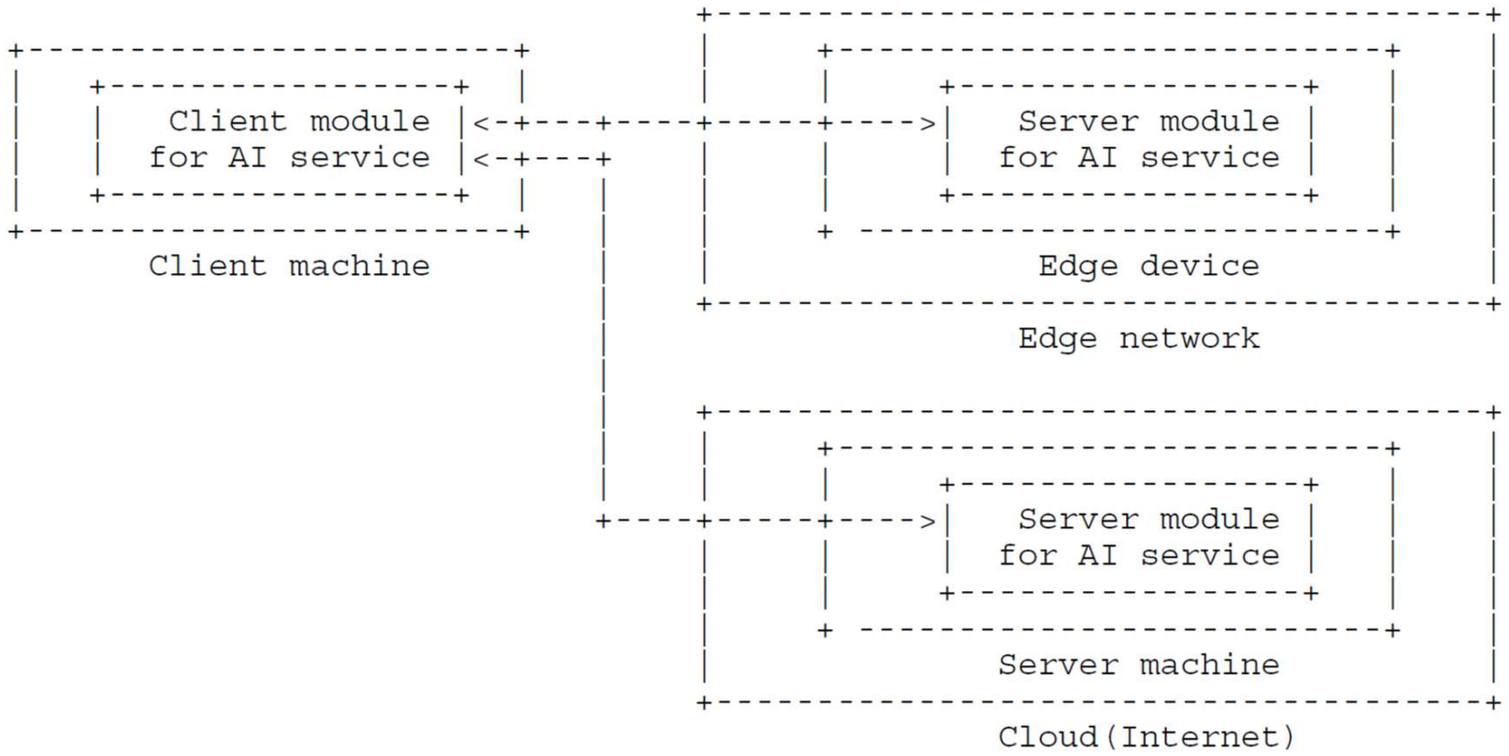


Figure 4: AI inference service on Edge device

AI inference service on Cloud server and Edge device



```
                                  +--------------------------------------+
+------------------------+        |     +---------------------------+    |
|   +-----------------+  |        |     |     +-----------------+   |    |
|   |   Client module |<-+---+----+-----+---->|   Server module |   |    |
|   |  for AI service |<-+---+    |     |     |  for AI service |   |    |
|   +-----------------+  |   |    |     |     +-----------------+   |    |
+------------------------+   |    |     + --------------------------+    |
       Client machine        |    |                Edge device           |
                             |    +--------------------------------------+
                             |                    Edge network
                             |
                             |    +--------------------------------------+
                             |    |     +---------------------------+    |
                             |    |     |     +-----------------+   |    |
                             +----+-----+---->|   Server module |   |    |
                                  |     |     |  for AI service |   |    |
                                  |     |     +-----------------+   |    |
                                  |     + --------------------------+    |
                                  |              Server machine          |
                                  +--------------------------------------+
                                                 Cloud(Internet)
```

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

Considerations according to the functional characteristics of the hardware (1/2)

- (Reference) ETSI Group Specification MEC-IEG 006 V1.1.1 (2017-01) "Mobile Edge Computing; Market Acceleration; MEC Metrics Best Practice and Guidelines“
  - It describes various metrics which can potentially be improved through deploying a service on a MEC platform
  - It can be identified in order to highlight the benefits of deploying MEC for various services and applications
  - Functional metrics
    - latency (both end-to-end, and one-way), energy efficiency, throughput, goodput, loss rate (number of dropped packets), jitter, number of out-of-order delivery packets, QoS, and MOS
  - Non-functional metrics
    - service lifecycle (instantiation, service deployment, service provisioning, service update (e.g. service scalability and elasticity), service disposal), service availability and fault tolerance (aka reliability), service processing/ computational load, global ME host load, number of API request (more generally number of events) processed/second on ME host, delay to process API request (north and south), number of failed API request

Considerations according to the functional characteristics of the hardware (2/2)

– 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

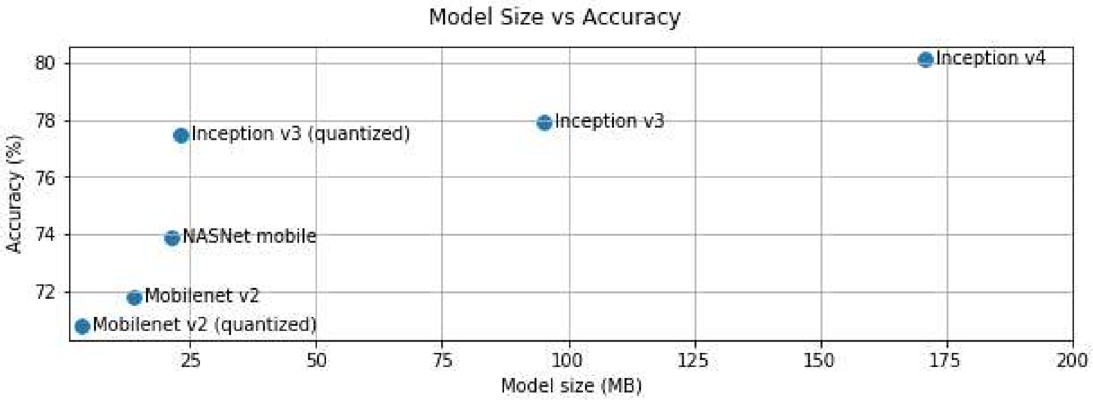
- Distant cloud server : High performance and high cost
- Near edge device : Medium performance and medium cost
- Local machine : Low performance and low cost

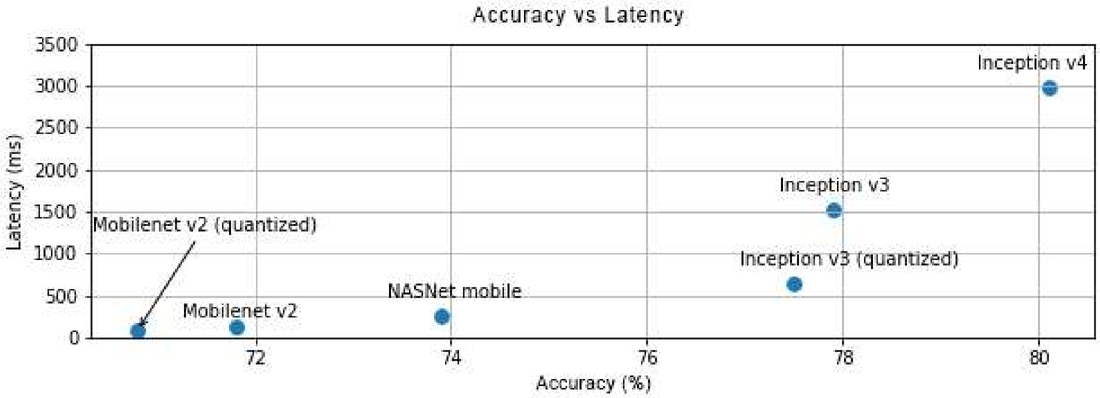
– AI inference service result in (assumption: same AI model)

- 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 according to the characteristics of the AI model (1/2)

– Model size vs. Accuracy vs. Latency





[Source : Google Tensorflow]

Considerations according to the characteristics of the AI model (2/2)

- 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

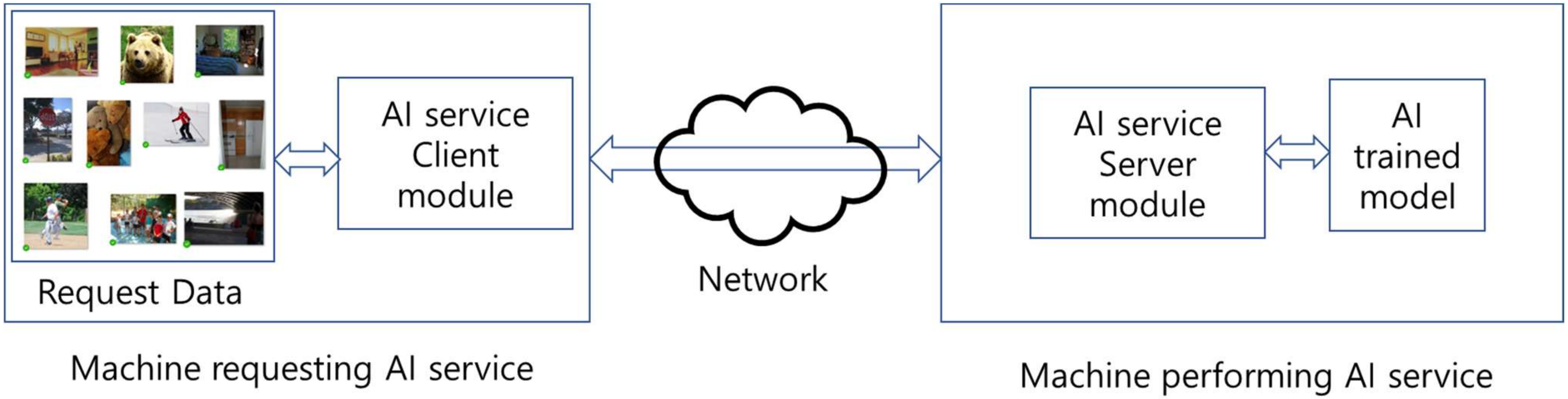
- AI inference serving framework
  - Traditional web server : ex) FastAPI, Flask, and Django
    - It can be operated on low performance machines
  - Specialized serving framework : ex) Tensorflow serving
    - It can provide high performance.

Considerations according to the characteristics of the communication method

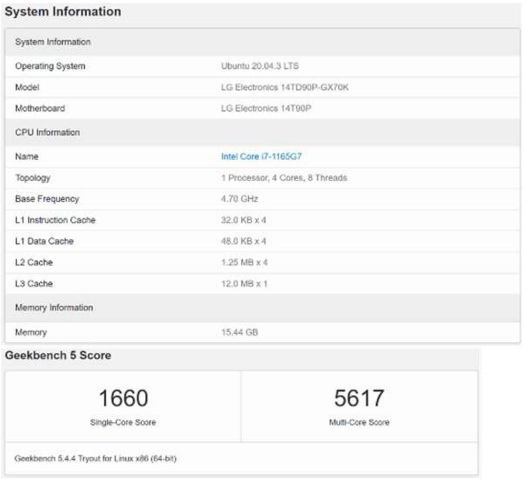
- AI inference service can be utilized
  - Traditional REST method
    - Common and easily deployed
  - Specified communication method (e.g., gRPC)
    - Better performance but need some works

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

An example of AI system for Object detection services



Latency of object detection services in each device



System Information

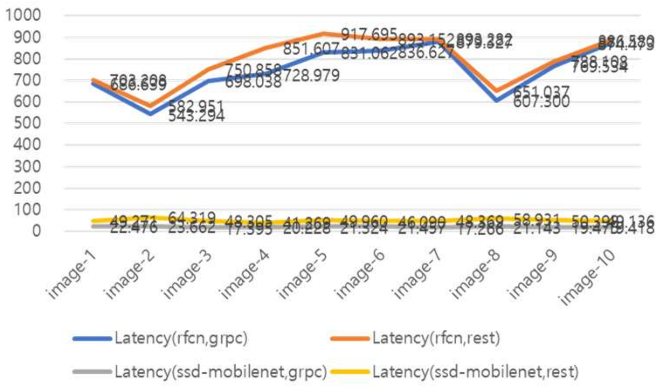
System Information	
Operating System	Ubuntu 20.04.3 LTS
Model	LG Electronics 14TD90P-GX70K
Motherboard	LG Electronics 14T90P
CPU Information	
Name	Intel Core i7-1165G7
Topology	1 Processor, 4 Cores, 8 Threads
Base Frequency	4.70 GHz
L1 Instruction Cache	32.0 KB x 4
L1 Data Cache	48.0 KB x 4
L2 Cache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

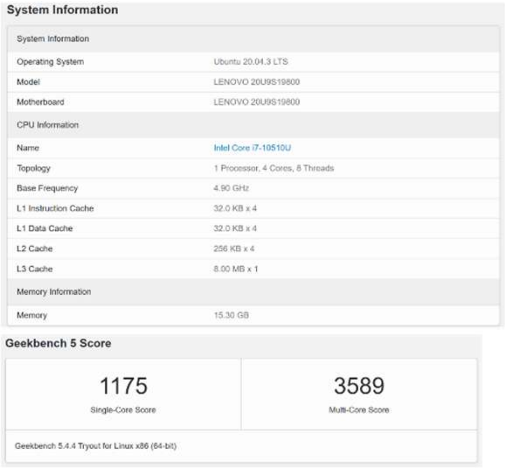
Geekbench 5 Score

1660 Single-Core Score	5617 Multi-Core Score

Geekbench 5.4.4 Tryout for Linux x86 (64-bit)

<Local device>





System Information

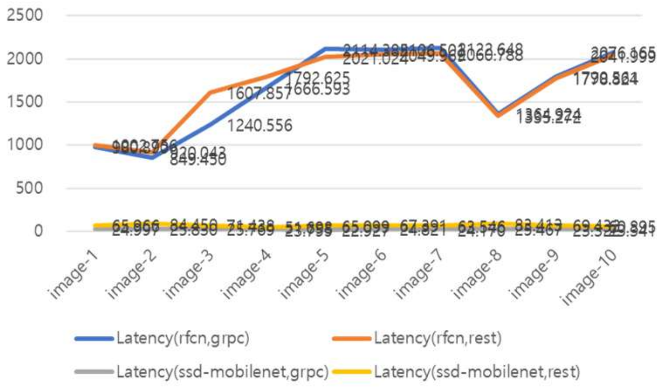
System Information	
Operating System	Ubuntu 20.04.3 LTS
Model	LENOVO 20U9S19800
Motherboard	LENOVO 20U9S19800
CPU Information	
Name	Intel Core i7-10510U
Topology	1 Processor, 4 Cores, 8 Threads
Base Frequency	4.90 GHz
L1 Instruction Cache	32.0 KB x 4
L1 Data Cache	32.0 KB x 4
L2 Cache	256 KB x 4
L3 Cache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

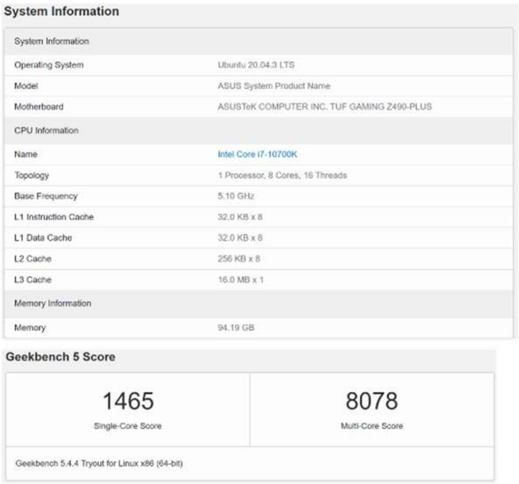
Geekbench 5 Score

1175 Single-Core Score	3589 Multi-Core Score

Geekbench 5.4.4 Tryout for Linux x86 (64-bit)

<Edge device>





System Information

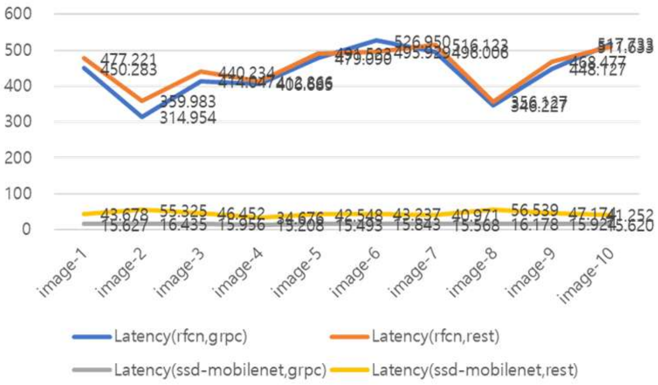
System Information	
Operating System	Ubuntu 20.04.3 LTS
Model	ASUS System Product Name
Motherboard	ASUSTeK COMPUTER INC. TUF GAMING Z490-PLUS
CPU Information	
Name	Intel Core i7-10700K
Topology	1 Processor, 8 Cores, 16 Threads
Base Frequency	5.10 GHz
L1 Instruction Cache	32.0 KB x 8
L1 Data Cache	32.0 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 Single-Core Score	8078 Multi-Core Score

Geekbench 5.4.4 Tryout for Linux x86 (64-bit)

<Cloud server>



Relationship to “Challenge document”

- In the NMRG, the “Challenge document” (draft-francois-nmrg-ai-challenges) is the main document for handling AI issues

- This draft is also related to the “Challenge document” and some texts can be added or merged
  - Distributed AI service
  - Lightweight AI service
  - Deployment of AI service

- This draft can be developed as a different document to focus on AI inference (Deployment of AI services)
  - The “Challenge document” includes many items

Thanks!!

Questions & Comments