| Change Request Form | |
| --- | --- |
| Logo, company name  Description automatically generated | CR Hardware Performance  – Proposed Changes |
| Document Summary | |
| Official Document Number, Document Title and Version Number | TS.53 Draft Compliance Test Specification for AI Mobile Device (clean version) |
| Official Document Type | Non-binding Permanent Reference Document |
| Change Request Security Classification | Non-confidential |
| Is this a new document or a Major or Minor Change? | New Document |
| Will this Change Request result in a Major or Minor version update? | Major Version |
| This document is for | Discussion |
| Input Editor and Organisation | Reza Barazideh, James Goel, Mungal Dhanda, Qualcomm Incorporated |
| Additional Contributors |  |
| Issuing Group/Project | TSGAI |
| Approving Group/Project | TSG |
| Change Request Creation Date | 11/12/2021 |
| What are the reasons for and benefits of creating this new document or Change Request? | Propose changes for Hardware Performance Preconditions. |

© GSMA 2021. The GSM Association (“Association”) makes no representation, warranty or undertaking (express or implied) with respect to and does not accept any responsibility for, and disclaims liability for the accuracy or completeness or timeliness of the information contained in this document. The information contained in this document may be subject to change without prior notice. This document has been classified according to the GSMA [Document Confidentiality Policy](https://infocentre.gsm.org/cgi-bin/prddets.cgi?274175). GSMA meetings are conducted in full compliance with the GSMA [Antitrust Policy](https://infocentre.gsm.org/cgi-bin/docdisp.cgi?275305).

# Hardware Performance

This section defines test procedures to measure hardware performance.

If A DUT is tested based on TS.47 section 3.1 requirements then test 3.1 shall be performed, otherwise test 3.2 shall be performed.

## Hardware performance testing based on TS.47 requirements

### Test purpose

To verify that the DUT can meet the minimum requirements of TOPS and TOPS/w.

### Referenced requirements

|  |  |
| --- | --- |
| Requirement for the modified VGG 16 network | |
| TS47\_3.1\_REQ\_001 | An AI Mobile Device SHOULD have a minimum of (1) int8 TOPS. |
| TS47\_3.1\_REQ\_002 | An AI Mobile Device SHOULD have a minimum of (0.5) float16 TOPS. |
| TS47\_3.1\_REQ\_003 | An AI Mobile Device SHOULD have a minimum of (0.5) int8 TOPS/Watt. |
| TS47\_3.1\_REQ\_004 | An AI Mobile Device SHOULD have a minimum of (0.3) float16 TOPS/Watt. |

### Preconditions

* **Test Model preparation**

1. Take VGG16\_notop as the Reference Model.

2. Use the Model Conversion tool provided by the chipset vendor to convert the Reference Model to an int8 or/and a float16 model that can be run on the DUT, take this converted model as Model\_t.

3. Validate Model\_t can be used as the Test Model.(TBD)

* **Test Scripts preparation**

Scripts to pre-process the test dataset, run the test model and measure TOPS.

* **Test Dataset**

Use current most common dataset for each specific task [7-8].

|  |  |  |
| --- | --- | --- |
| **Area** | **Task** | **Dataset** |
| Vision | Image classification | ImageNet |
| Vision | Object detection | MS-COCO 2017 |
| Vision | Segmentation | ADE20K (32 classes, 512x512) |
| Language | Language processing | SQUAD 1.1 |

### Initial configuration

DUT is loaded with test scripts and test model for int8 or/and float16 TOPS, TOPS/watt measurement.

DUT is Switched OFF.

Power meter is Switched OFF.

### Test procedure

|  |  |  |
| --- | --- | --- |
| Step | Test procedure | Expected result |
| 1 | Switch the power meter on and connect it to DUT for power measurement. | The power meter is on. |
| 2 | Switch DUT on and adjust the screen brightness to the lowest level, turn off the Bluetooth, mute the DUT and turn on the flight mode. | DUT is on and is in flight mode. |
| 3 | Record the current and voltage. | The current curve and the voltage are displayed. |
| 4 | Wait until the current is stable, i.e. the current curve is stable [+/-5%]. | The current is stable. |
| 5 | Record the background current and the voltage for 60 seconds, compute the average value. | The value of average background current and average voltage are obtained. |
| 6 | Run the test scripts for int8 Model\_t, record the inference time and compute the average inference current. | The inference time and the average inference current value are obtained. |
| 7 | Compute int8 TOPS and compare the result with the value specified in the requirement TS.47\_3.1\_REQ\_001. | The int8 TOPS result meets requirement TS.47\_3.1\_REQ\_001. |
| 8 | Compute int8 TOPS/Watt and compare the result with the value specified in the requirement TS.47\_3.1\_REQ\_003. | The int8 TOPS/Watt result meets requirement TS.47\_3.1\_REQ\_003. |
| 9 | Stop recording the current. | The current curve stops recording. |
| 10 | Change the test model to float16 Model\_t, repeat step 3 to 9 for float16 TOPS and float 16 TOPS/Watt measurement. | The float16 TOPS result meets requirement TS.47\_3.1\_REQ\_002.  The float16 TOPS/Watt result meets requirement TS.47\_3.1\_REQ\_004. |

## Hardware performance testing with alternative model and strategy

### Test purpose

To verify that the DUT can meet the minimum performance requirements using alternative model and metrics.

### Referenced requirements

There are not specific requirements as ML commons is used for benchmarking. We need to define requirements for ML commons proposed metrics such as Latency, Frames/Sec, Samples/Sec.

### Preconditions

* **Test Model preparation**

1. Use ML Commons approach and models for each specific task [7-8].

|  |  |  |
| --- | --- | --- |
| **Area** | **Task** | **Model** |
| Vision | Image classification | MobileNetEdgeTPU |
| Vision | Object detection | MobileDETs |
| Vision | Segmentation | DeepLabV3+ (MobileNetV2) |
| Language | Language processing | Mobile-BERT |

2. Use ML Commons approach for model conversion [7].

3. Final model validation can be done based on ML Commons approach and provided score or we need to define new requirements for proposed metrics.

* **Test Scripts preparation**

All software and requirements are available, and they are open source [7-8].

* **Test Dataset**

Use current most common dataset for each specific task.

|  |  |  |  |
| --- | --- | --- | --- |
| **Area** | **Task** | **Model** | **Dataset** |
| Vision | Image classification | MobileNetEdgeTPU | ImageNet |
| Vision | Object detection | MobileDETs | MS-COCO 2017 |
| Vision | Segmentation | DeepLabV3+ (MobileNetV2) | ADE20K (32 classes, 512x512) |
| Language | Language processing | Mobile-BERT | SQUAD 1.1 |

### Test procedure

### Performance measurement

There are predefined metrics and test procedures for different scenarios according to ML Commons approach [8].

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scenario** | **Query Generation** | **Duration** | **Samples/query** | **Performance Metric** |
| Single stream | LoadGen sends next query as soon as DUT completes the previous query | 1024 queries and 60 seconds | 1 | 90%-ile measured latency |
| Multiple stream | LoadGen sends a new query every latency constraint if the DUT has completed the prior query, otherwise the new query is dropped and is counted as one overtime query | 270,336 queries and 60 seconds | Variable, see metric | Maximum number of inferences per query supported |
| Server | LoadGen sends new queries to the DUT according to a Poisson distribution | 270,336 queries and 60 seconds | 1 | Maximum Poisson throughput parameter supported |
| Offline | LoadGen sends all queries to the DUT at start | 1 query and 60 seconds | At least 24,576 | Measured throughput |

## 1.5 References

Requirements shall be based on the exact versions as indicated below. However if the manufacturers use a later release and/or version this should be indicated. The GSMA will continually align with other SDOs as appropriate.

| Ref | Doc Number | Title |
| --- | --- | --- |
|  | GSMA PRD TS.47 | AI Mobile Device Specification, Version 1.0, September 2019 |
| [2] | ISO-IEC-19795-1 | Information technology — Biometric performance testing and reporting —Part 1: Principles and framework |
| [3] |  | FIDO Biometrics Requirements (2020) |
| [4] | GSMA PRD TS.29 | Smartphone Performance Test Case Guideline Version 6.0 or later |
| [5] | RFC 2119 | “Key words for use in RFCs to Indicate Requirement Levels”, S. Bradner, March 1997. Available at <http://www.ietf.org/rfc/rfc2119.txt> |
| [6] | RFC8174 | Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words  <https://www.rfc-editor.org/info/rfc8174> |
| [7] |  | <https://github.com/mlcommons> |
| [8] |  | MLPerf Inference Benchmark, https://arxiv.org/abs/1911.02549 |