MTL/ARL/LNL Platform NPU Benchmarking

Customer Communication

WW28, July 2024

Document Number: 782543

Legal Disclaimer

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis. You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

All product plans and roadmaps are subject to change without notice.

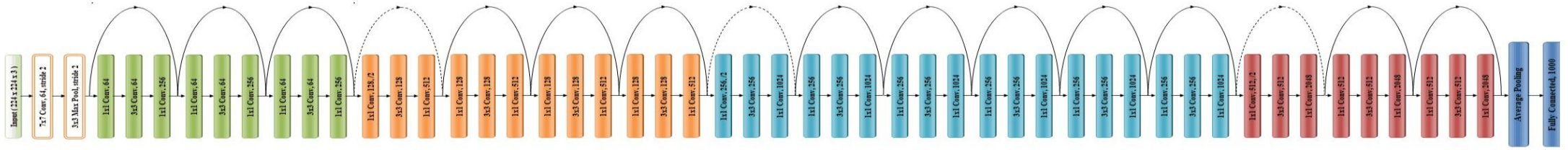
The products described may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

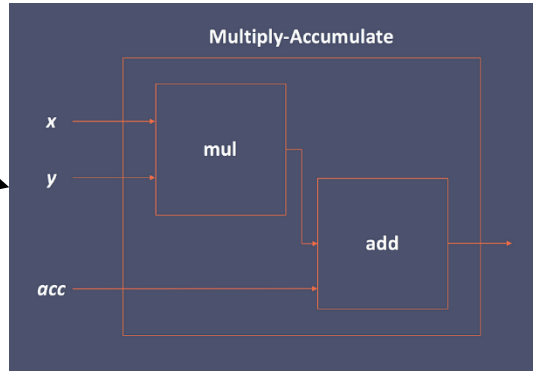
*Other names and brands may be claimed as the property of others.

Resnet50 (Meteor Lake)



Input 224x224

Workload
(MACs operations on NPU)



One multiply-accumulate is two operations

Output (frame)

- Operations per frame: constant value per network, for Resnet50 it's 8.216 GOPs

- Peak TOPS (pTOPS) = peak theoretical max performance

  pTOPS = max frequency * (MAC / clock) * 2

  NPU is 11 pTOPS

- Effective TOPS (eTOPS) = real performance on a given AI workload (the efficiency of pTOPS)

  eTOPS = (fps * each frame GOPs)/1000

  *NPU 8.2 eTOPS = (1000 * 8.216) / 1000*

- We use ResNet50: a common network + a good mix of a memory & compute bound network. Is it Perfect? -> No, but it's better than pTOPS as eTOPS shows real workload measured across many HW configs

Run Resnet50 on MTL/ARL NPU

1. Download NPU driver from RDC and install the driver
2. Install one of Python 3.8 - 3.11, open cmd terminal and follow below steps to create virtual environment, ov_24.2 under C:\Users\Public\

```
cd C:\Users\Public\
python -m venv ov_24.2
ov_24.2\Scripts\activate
python -m pip install --upgrade pip
pip install openvino==2024.2.0
```

3. Download Resnet50 model from RDC Kit#823845 (only for test purpose), unzip to C:\Users\Public
   C:\Users\Public\resnet-50-v1_5-sparse50.xml
   C:\Users\Public\resnet-50-v1_5-sparse50.bin
4. Create npu_config.json and add below line to it, and then save it at C:\Users\Public\

```
{"NPU": {"NPU_COMPILER_TYPE": "DRIVER", "NPU_COMPILATION_MODE_PARAMS": "enable-activation-sparsity=true" }}
```

5. Pre-check:

Required Files	File Path
ResNet50 int8 with 50% sparsity (IR model)	C:\Users\Public\resnet-50-v1_5-sparse50.xml C:\Users\Public\resnet-50-v1_5-sparse50.bin
NPU configuration	C:\Users\Public\npu_config.json

Run Resnet50 on MTL/ARL NPU (Cont.)

6. Open cmd terminal and then run below commands to setup environment

```
C:\Users\Public\ov_24.2\Scripts\activate
cd C:\Users\Public\
```

7. Run NPU in Throughput mode

```
benchmark_app -d NPU -m resnet-50-v1_5-sparse50.xml -t 180 -layout [NCHW] -ip f16 -op f16 -hint throughput -load_config npu_config.json
```

8. Run NPU in Latency mode

```
benchmark_app -d NPU -m resnet-50-v1_5-sparse50.xml -t 180 -layout [NCHW] -ip f16 -op f16 -api sync -load_config npu_config.json
```

Result for reference:

- Configuration:
  - MTL-H 28W/ARL-S A2 RVP
  - NPU driver: 32.0.100.2540
  - OpenVINO 2024.2
  - MEMORY size: 16GB
  - Power mode: AC Best Performance
- Performance on NPU:

Execution Mode	MTL-H28	ARL-S
Throughput	8.47	9.17
Latency	6.79	7.46

Unit: eTOPs

One-click Script for Test (MTL/ARL)

- Pre-check:

Required Files	File Path
Python environment	C:\Users\Public\ov_24.2\
ResNet50 int8 with 50% sparsity (IR model)	• C:\Users\Public\resnet-50-v1_5-sparse50.xml • C:\Users\Public\resnet-50-v1_5-sparse50.bin
NPU configuration	C:\Users\Public\npu_config.json

- Create resnet50.bat and add below lines into it, and then save it at C:\Users\Public\

```
call C:\Users\Public\ov_24.2\Scripts\activate
cd C:\Users\Public\
benchmark_app -d NPU -m resnet-50-v1_5-sparse50.xml -t 180 -layout [NCHW] -ip f16 -op f16 -hint throughput -load_config npu_config.json
benchmark_app -d NPU -m resnet-50-v1_5-sparse50.xml -t 180 -layout [NCHW] -ip f16 -op f16 -api sync -load_config npu_config.json

pause
```

Run Resnet50 on LNL NPU

1. Download NPU driver from RDC and install the driver
2. Install one of Python 3.8 - 3.11, open cmd terminal and follow below steps to create virtual environment, ov_24.2 under C:\Users\Public\

```
cd C:\Users\Public\
python -m venv ov_24.2
ov_24.2\Scripts\activate
python -m pip install --upgrade pip
pip install openvino==2024.2.0
```

3. Download Resnet50 model from RDC Kit#823845 (only for test purpose), unzip to C:\Users\Public
   C:\Users\Public\resnet-50-v1_5-sparse50.xml
   C:\Users\Public\resnet-50-v1_5-sparse50.bin
4. Pre-check:

Required Files	File Path
ResNet50 int8 with 50% sparsity (IR model)	C:\Users\Public\resnet-50-v1_5-sparse50.xml C:\Users\Public\resnet-50-v1_5-sparse50.bin

Run Resnet50 on LNL NPU (Cont.)

5. Open cmd terminal and then run below commands to setup environment

```
C:\Users\Public\ov_24.2\Scripts\activate
cd C:\Users\Public\
```

6. Run NPU in Throughput mode

```
benchmark_app -d NPU -m resnet-50-v1_5-sparse50.xml -t 180 -ip f16 -op f16 -hint throughput
```

7. Run NPU in Latency mode

```
benchmark_app -d NPU -m resnet-50-v1_5-sparse50.xml -t 180 -layout [NCHW] -ip f16 -op f16 -api sync
```

Result for reference:

- Configuration:
  - LNL B0 L0MO Ultra 5
  - NPU driver: 32.0.100.2565
  - OpenVINO 2024.2
  - MEMORY size: 16GB
  - Power mode: AC Best Performance
- Performance on LNL NPU:

Execution Mode	NPU
Throughput	19.34
Latency	13.30

Unit: eTOPs

One-click Script for Test (LNL)

- Pre-check:

Required Files	File Path
Python environment	C:\Users\Public\ov_24.2\
ResNet50 int8 with 50% sparsity (IR model)	• C:\Users\Public\resnet-50-v1_5-sparse50.xml • C:\Users\Public\resnet-50-v1_5-sparse50.bin

- Create resnet50.bat and add below lines into it, and then save it at C:\Users\Public\

```
call C:\Users\Public\ov_24.1\Scripts\activate
cd C:\Users\Public\
benchmark_app -d NPU -m resnet-50-v1_5-sparse50.xml -t 180 -ip f16 -op f16 -hint throughput
benchmark_app -d NPU -m resnet-50-v1_5-sparse50.xml -t 180 -layout [NCHW] -ip f16 -op f16 -api sync

pause
```


intel