Coursework 1 XU,Chengzhen

1. FLOPS: X*10*8*2*4
   number of core: 10
   I fail to find the frequency of CPU in my macbook, so here I replace it with X

```
(base) mac@xcz ~ % sysctl -a | grep machdep.cpu
machdep.cpu.cores_per_package: 10
machdep.cpu.core_count: 10
machdep.cpu.logical_per_package: 10
machdep.cpu.thread_count: 10
machdep.cpu.brand_string: Apple M1 Pro
```

There is no lscpu in mac. Here I use sysctl hw.

```
(base) mac@xcz ~ %  sysctl hw
hw.ncpu: 10
hw.byteorder: 1234
hw.memsize: 17179869184
hw.activecpu: 10
hw.perflevel0.physicalcpu: 8
hw.perflevel0.physicalcpu_max: 8
hw.perflevel0.logicalcpu: 8
hw.perflevel0.logicalcpu_max: 8
hw.perflevel0.l1icachesize: 196608
hw.perflevel0.l1dcachesize: 131072
hw.perflevel0.l2cachesize: 12582912
hw.perflevel0.cpusperl2: 4
hw.perflevel1.physicalcpu: 2
hw.perflevel1.physicalcpu_max: 2
hw.perflevel1.logicalcpu: 2
hw.perflevel1.logicalcpu_max: 2
hw.perflevel1.l1icachesize: 131072
hw.perflevel1.l1dcachesize: 65536
hw.perflevel1.l2cachesize: 4194304
hw.perflevel1.cpusperl2: 2
hw.optional.arm.FEAT_FlagM: 1
hw.optional.arm.FEAT_FlagM2: 1
hw.optional.arm.FEAT_FHM: 1
hw.optional.arm.FEAT_DotProd: 1
hw.optional.arm.FEAT_SHA3: 1
hw.optional.arm.FEAT_RDM: 1
hw.optional.arm.FEAT_LSE: 1
```

```
hw.optional.arm.FEAT_LSE: 1
hw.optional.arm.FEAT_SHA256: 1
hw.optional.arm.FEAT_SHA512: 1
hw.optional.arm.FEAT_SHA1: 1
hw.optional.arm.FEAT_AES: 1
hw.optional.arm.FEAT_PMULL: 1
hw.optional.arm.FEAT_SPECRES: 0
hw.optional.arm.FEAT_SB: 1
hw.optional.arm.FEAT_FRINTTS: 1
hw.optional.arm.FEAT_LRCPC: 1
hw.optional.arm.FEAT_LRCPC2: 1
hw.optional.arm.FEAT_FCMA: 1
hw.optional.arm.FEAT_JSCVT: 1
hw.optional.arm.FEAT_PAuth: 1
hw.optional.arm.FEAT_PAuth2: 0
hw.optional.arm.FEAT_FPAC: 0
hw.optional.arm.FEAT_DPB: 1
hw.optional.arm.FEAT_DPB2: 1
hw.optional.arm.FEAT_BF16: 0
hw.optional.arm.FEAT_I8MM: 0
hw.optional.arm.FEAT_ECV: 1
hw.optional.arm.FEAT_LSE2: 1
hw.optional.arm.FEAT_CSV2: 1
hw.optional.arm.FEAT_CSV3: 1
hw.optional.arm.FEAT_FP16: 1
hw.optional.arm.FEAT_SSBS: 1
hw.optional.arm.FEAT_BTI: 0
hw.optional.floatingpoint: 1
hw.optional.neon: 1
hw.optional.neon_hpfp: 1
hw.optional.neon_fp16: 1
hw.optional.armv8_1_atomics: 1
hw.optional.armv8_2_fhm: 1
```

Also, I failed to find the way to know whether my macbook support fma or AVX2, so I assume it supports them.

2. Here I assume the frequency is $3GHz = 3 \times 10^9/s$

minimum latency time: $2 \times \frac{0.1m}{3 \times 10^8 m/s} = 6.6 \times 10^{-10} s$
CPU clock time: $\frac{1}{3 \times 10^9} = 3.33 \times 10^{-10} s$
They are the same order.

3.

1).*Neural Network Processing Unit (NPU)*

A neural processing unit (NPU) is a specialized circuit that implements all the necessary control and arithmetic logic necessary to execute machine learning algorithms. NPUs are designed to accelerate the performance of common machine learning tasks such as image classification, machine translation, object detection, and various other predictive models. They may be part of a large SoC, a plurality of NPUs may be instantiated on a single chip, or they may be part of a dedicated neural-network accelerator. HUAWEI's flagship Kirin 970 is HUAWEI's first mobile AI computing platform featuring a dedicated Neural Processing Unit (NPU).
.

> https://en.wikichip.org/wiki/neural_processor
> https://consumer.huawei.com/en/press/news/2017/ifa2017-kirin970/

2)*Field-Programmable Gate Array (FPGA)*

A field-programmable gate array (FPGA) is an integrated circuit designed to be configured by a customer or a designer after manufacturing. FPGA designs employ very fast I/O rates and bidirectional data buses. and
the configuration is generally specified using a hardware description language (HDL) which contain a hierarchy of reconfigurable interconnects allowing blocks to be wired together and an array of programmable logic blocks which can be configured to perform complex combinational functions, or act as simple logic gates. They have ample logic gates and RAM blocks to implement complex digital computations. In most FPGAs, logic blocks also include memory elements, which may be simple flip-flops or more complete blocks of memory. Many FPGAs can be reprogrammed to implement different logic functions, allowing flexible reconfigurable computing as performed in computer software. Moreover, FPGAs are also being used as accelerators to speed up the execution of deep learning algorithms. They are also featured by the ability to update the functionality after shipping, partial reconfiguration of a portion of the design and the low non-recurring engineering costs relative to an ASIC design and it makes them more competitive to other integrated circuits. What's more, some FPGAs have analog features in addition to digital functions, including a programmable slew rate on each output pin, allowing the engineer to set low rates on lightly loaded pins that would otherwise ring or couple unacceptably, and to set higher rates on heavily loaded high-speed channels that would otherwise run too slowly. Also common are quartz-crystal oscillator driver circuitry, on-chip resistance-capacitance oscillators, and phase-locked loops with embedded voltage-controlled oscillators used for clock generation and management as well as for high-speed serializer-deserializer (SERDES) transmit clocks and receiver clock recovery.

> https://en.wikipedia.org/wiki/Field-programmable_gate_array

3)*Tensor Processing Unit (TPU)*
Tensor Processing Unit (TPU) is an AI accelerator application-specific integrated circuit (ASIC) developed by Google for neural network machine learning, using Google's own TensorFlow software.

Google began using TPUs internally in 2015, and in 2018 made them available for third party use, both as part of its cloud infrastructure and by offering a smaller version of the chip for sale.

To make comparison to GPU, TPUs are designed for a high volume of low precision computation (e.g. as little as 8-bit precision) with more input/output operations per joule, without hardware for rasterisation/texture mapping. The TPU ASICs are mounted in a heatsink assembly, which can fit in a hard drive slot within a data center rack. As we all know, different types of processors are suited for different types of machine learning models. TPUs are well suited for CNNs, while GPUs have benefits for some fully-connected neural networks, and CPUs can have advantages for RNNs.

> https://en.wikipedia.org/wiki/Tensor_Processing_Unit

4.

4: 1,5,7,8 are violated by floating point numbers.