1

3.2M\*8(cores)\*4(SIMD)\*2(fma)\*4(avx2) = 0.8192 G FLOPS

支持avx2 支持fma

cat /proc/cpuinfo

cpu MHz : **3194.011**

cpu cores : **8**

flags : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush mmx fxsr sse sse2 ht syscall nx mmxext fxsr\_opt pdpe1gb rdtscp lm constant\_tsc rep\_good nopl tsc\_reliable nonstop

\_tsc cpuid extd\_apicid pni pclmulqdq ssse3 **fma** cx16 sse4\_1 sse4\_2 movbe popcnt aes xsave avx f16c rdrand hypervisor lahf\_lm cmp\_legacy svm cr8\_legacy abm sse4a misalignsse 3dnowprefetch osvw topoext perfctr\_core

ssbd ibrs ibpb stibp vmmcall fsgsbase bmi1 **avx2** smep bmi2 erms invpcid rdseed adx smap clflushopt clwb sha\_ni xsaveopt xsavec xgetbv1 xsaves clzero xsaveerptr arat npt nrip\_save tsc\_scale vmcb\_clean flushbyasid

decodeassists pausefilter pfthreshold v\_vmsave\_vmload umip vaes vpclmulqdq rdpid fsrm

2 cpu time 1s/3.2g = 0.3ns

latency = 10cm/(3\*10^8m /s) = 0.33ns

They are about the same level but the speed of the information cannot even reach the speed of light. So the calculation will not be very accuracy. There will be longer latency time.

1 Features of NPU

NPU is one kind of Deep Learning Processors (DLPs) used in certain mobile devices. It is a specific circuit which implements necessary control and arithmetic logic necessary to execute [machine learning](https://en.wikichip.org/w/index.php?title=machine_learning&action=edit&redlink=1) algorithms. Most DLPs employ a large number of computing components to leverage high data-level parallelism, a larger on-chip buffer or memory to leverage the data reuse patterns, and limited data-width operators for error-resilience of deep learning.

Since designed for [machine learning](https://en.wikichip.org/w/index.php?title=machine_learning&action=edit&redlink=1) algorithms that include a large amount of [multiply-accumulate operations](https://en.wikichip.org/w/index.php?title=multiply-accumulate_operations&action=edit&redlink=1), it processes instructions in a highly parallelizable way, which makes it quite different from CPU processing instructions. Moreover, unlike a GPU, NPUs can benefit from vastly simpler logic because their workloads tend to exhibit high regularity in the computational patterns of [deep neural networks](https://en.wikichip.org/w/index.php?title=deep_neural_networks&action=edit&redlink=1).

Due to the difference demands from ML algorithms, there are two general classification of NPU. One is about training and the other focuses on inference. The former focuses on accelerating the curating of new models, and the later concentrates on accelerating inference operate on complete models. Some argue that they are not rigorously the same thing, but they do have their overlaps.2 Features of FPGA

FPGA is the abbreviation for field programmable gate array. It is an integrated circuit whose most important feature is that it contains an array of programmable logic blocks, and a hierarchy of reconfigurable interconnects. So this makes it to be easily customized by the designers or customers with a hardware description language.

As for the application of FPGA, an FPGA can be used to solve any problem which is computable. that FPGAs can be used to implement a soft micro-processor. Their advantage lies in that they are significantly faster for some applications because of their doing a good work of parallelization and their optimality with respect to the number of gates used for certain processes.

Another trend in the use of FPGAs is hardware acceleration. People use FPGAs to accelerate the procedure of some algorithms and sometimes it will also share the computation with other kinds of processors.

And despite of the low burden for commercial use, advantages of FPGAs also include the ability to be re-programmed when they are already in the field to fix bugs, making the procedure more flexible.

3 Features of TPU

Tensor Processing Unit (TPU) is an [AI accelerator](https://en.wikipedia.org/wiki/AI_accelerator) [application-specific integrated circuit](https://en.wikipedia.org/wiki/Application-specific_integrated_circuit) (ASIC) developed by [Google](https://en.wikipedia.org/wiki/Google) for [neural network](https://en.wikipedia.org/wiki/Artificial_neural_network) [machine learning](https://en.wikipedia.org/wiki/Machine_learning). TPU accelerates Machine Learning algorithms with TensorFlow, and doing complicated matrix or tensor computation in an efficient way.

Compared to traditional GPU, TPU are designed for a high volume of low precision computation. From the energy-saving perspective, TPU can do the same input and output operations in less energy unit and do not need any hardware for rasterization or texture mapping. These 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. In their architectures, GPU is a processor itself which implements instructions while TPU assists computation of matrix with help from other processors.

Reference:

<https://en.wikipedia.org/wiki/Deep_learning_processor>

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

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

<https://en.wikichip.org/wiki/neural_processor>

4

There are five axioms being violated by floating point numbers.

1 x

2 x

3 o

4 o

5 x

6 o

7 x

8 x

9 o

About 1,4, 5: please refer to

<https://en.wikipedia.org/wiki/Floating-point_arithmetic#Accuracy_problems>

While floating-point addition and multiplication are both commutative (a + b = b + a and a × b = b × a), they are not necessarily associative. They are also not necessarily distributive

About 2: There are 0 and -0;

关于7：

(我测试了 1/3.33333333333333333\*3.33333333333333333 和 1/3.33333333333333333\*3.333333333333333333 后面比前面多一个3)