A Guide to the A.I. Cards and A.I. Processors

Martin Rupp
SCIENTIFIC AND COMPUTER DEVELOPMENT SCD LTD

As AI is invading more and more of our daily lives, a new type of hardware is making its way: the AI-powered processors, that is to say, the processors specially designed to run Artificial Intelligence algorithms.

Some of the readers may recall the U.S. movie named 'Small Soldiers' (1998).
In the movie, a factory put by mistake new revolutionary military AI-powered processors into a group of several toys, aimed to be sold to kids. The result is that the toys are behaving fully autonomously and create - of course - panic and chaos everywhere.

While in the context of 1988, such a scenario was an anticipation, in 2020 (so 30 years later…) this is not at all unrealistic because we are seeing - among other things - the rise of A.I.-powered processors everywhere. This is just the beginning but as always with disruptive technologies, it will just grow exponentially year after year or even months after months.

Here we wish to present to the reader a quick guide of the latest AI processors available for sale as of October 2020.

GPU cards

GPU or Graphical Power Unit, while originally targeting video applications, has become a de facto standard in AI hardware, and companies traditionally providing cards for the videogame market such as Nvidia have now turned to AI hardware development.

NVIDIA Titan RTX

This is a high-end card that targets researchers in AI as well as designers of AI systems.
It can be scaled and used for a lot of applications such as market pricing prediction or data analysis in general.

The NVIDIA Titan RTX is powered by a specific architecture, the Turing architecture.
The Turing microarchitecture features among others large matrix operations for AI as well as **Deep Learning Super Sampling** (DLSS). Other similar microarchitectures include the **Volta** and **Ampere** Architectures.

The specifications of the card are truly impressive:

- 130 tensor teraflops;
- 576 Turing mixed-precision tensor cores;
- 24 GB of GDDR6 ultra-fast RAM.

The card allows the processing of huge datasets such as the ones typically met in science applications. It is fully integrated with a set of data science libraries named the RAPIDS suite and uses CUDA-X AI SDK.

Illustration: A Titan RTX card

The typical retail price of a Titan RTX card is **USD 2,500- USD 3,000.**

There are more varieties of AI-powered GPU cards provided by NVIDIA. Other AI products are also available like the NVIDIA QUADRO GV10

Radeon Instinct

Radeon Instinct is the GPU card for Deep learning manufactured by AMD.
Note that AMD also produces AI-powered microprocessors for PCs.

The brand is divided into several products: MI-6, MI-8, MI-25 and MI-50.

Here are the specifications of the MI-25 card:

- 4096 Stream Processors.
- Architecture based on Vega 10.
- 24.6 TFLOPS Half Precision (FP16)
- 12.3 TFLOPS Single Precision (FP32)
- 768 GFLOPS Double Precision (FP64)
- 16GB HBM2 RAM.

AMD provides software such as MIOpen which offers support for the biggest AI frameworks (Tensorflow, caffe, theano, etc …)

Illustration: Radeon instinct card

The typical retail price of an MI25 Radeon card is around USD18,000

IoT A.I. processors

An important application of A.I. processors is for the Internet Of Things (IoT) devices. Indeed these devices often need image recognition, and classification decisions for example and they deeply depend on deep learning methods, especially Convolutional Neural Networks (CNNs).

GAP8

GAP8 is a French revolutionary processor provided with 8 cores.
GAP8 is an ultra-low-power processor equipped with a built-in hardware Convolutional Neural Network engine. It consumes only milliwatts of power for running typical neural networks.

GapUino GAP8 development board

The typical retail price of a Gapuino GAP8 development board is around USD300

Myriad

Myriad X, developed by the company Movidius, is a vision-processing unit with a built-in neural network engine able to deliver up to 1 TOPS (Teraflops operations per second)
Myriad X also offers a specialized Deep neural Network engine. The processor can deal easily with memory bottleneck problems

Myriad X is built on 16 SHAVE cores. SHAVE cores, originally built for game physics engines, have been upgraded as vision accelerators.

One of the interests of the processor is when used in the Neural Compute Stick2 from Intel

That stick is usually used for rapid prototyping of edge AI applications. It is plugged into a standard workstation via USB.

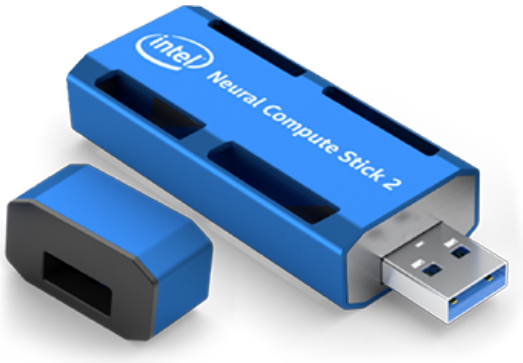

Illustration: Neural Compute Stick2

The typical retail price of a Neural Compute Stick2 is around USD99

TDA4VM

This AI processor from Texas Instrument is part of the Jacinto 7 series which is designed for the automotive industry.

The TDA4VM, based on the Cortex-A72 architecture, is aimed at providing driver assistance functions. For this, it is equipped with a deep-learning accelerator chip.

It can provide 8 TOPS with a dedicated Matrix Multiplication Accelerator (typically used by neural networks) which is a good performance for an IoT processor.

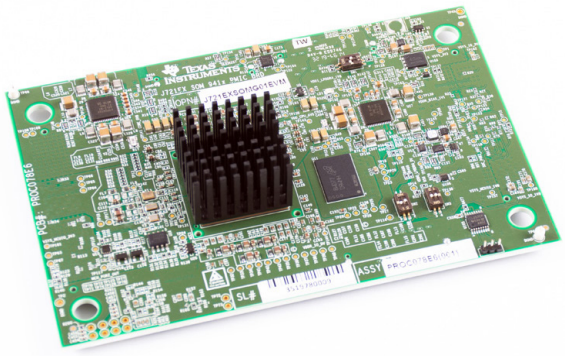

Illustration: a Jacinto7 dev board featuring the TDA4VM

KL520

The KL520 is developed by the company Kneron, based in Taiwan.
It runs natively Convolutional neural Networks and targets the image recognition market (facial recognition etc ..).

Its CNN engine can run at 0.3 TOPS which is good enough for most facial recognition applications.

Interestingly enough, the chip architecture can be modified and compression is also used for optimization of the CNN models.

Illustration: a M2AI-2280-520 card featuring the KL520

Google Tensor Processing Units

Google, which provides TensorFlow, the leading AI framework on the market, provides AI-based computing cards, the Google TPU or Tensor Processing Units. These TPUs are available for servers or edge computing. A Typical edge TPU can perform at 4 Tera-operations per second.

Tensorflow provides support for TPUs the same way it provides support for GPUs (and ‘traditional’ CPUs). It allows a neural network model to run in a TPU.

TPUs are designed to be connected to form dedicated AI computing machines.
Here is an illustration of a Cloud using TPU machines:

Illustration: TPU forming a cloud

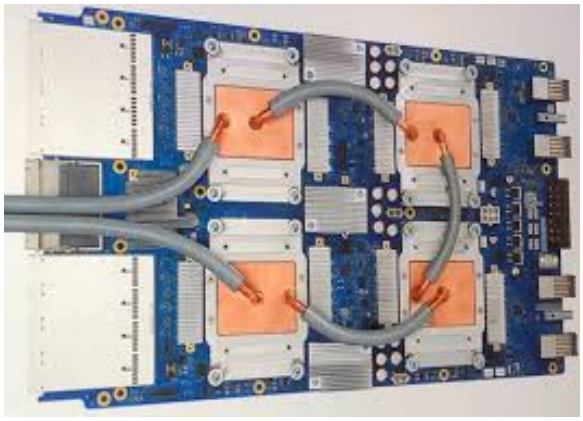

Illustration: TPUs connected

The typical retail price of a Cloud TPU v2 is around $4.95 / TPU hour

AI-powered processors for standard PCs

Intel

Intel, the famous worldwide microprocessor manufacturer, has decided, during the past recent years, to invest in AI-powered processors.

Mobileye, Agilex FPGA, or Movidius are Intel products that provide AI-accelerated processing.

Intel Deep Learning Boost

Intel Deep Learning Boost is a new processor technology that increases the speed of execution of AI deep learning for example speech recognition, facial recognition, natural language processing (NLP), etc …

Nervana neural network **processors** (NNP)

This project of Intel has been stopped.

Habana Gaudi and the Habana Goya

Intel provides highly specialized AI processors with hardware routines for training deep neural networks via the Goya and Gaudi processors.

Illustration: Habana Goya processor

2nd Gen Intel® Xeon® Scalable Processor

This new Xeon generation is equipped with Deep Learning Boost technology and can provide high performance for neural networks. For example CNNs with the standard neural networks frameworks (Keras+tensorflow / Theano/Caffe)

Intel's new AI-powered Xeon

AMD

AMD - same as Intel - has recently invested in AI processors but perhaps with slightly less strength than its rival.

AMD EPYC

EPYC is a specialized AI processor that can provide up to 5 petaflops of AI power.
It can offer 120 cores. and it is also used for [building powerful AI workstations]

The AMD ERYC

Conclusion

We provided in this article a very brief introductory guide to the recent A.I. cards and processors that can be bought on the market. Most of these A.I. processors allow modular architecture since they can be scaled in parallel.

Several A.I. workstations are built on these architectures. Their prices range from 10,000$ to 100,000$ or even more.

A.I. is the future and there will be more and more demand for specialized hardware and software that can train and run A.I. applications.

Stay tuned for more news!