Intel® Edge AI for IoT Developers Nanodegree

Glossary – Course 2 – Choosing the Right Hardware

CPU

Central Processing Unit (CPU)

The subset of the device that performs the basic operations of running the operating system, reading and writing from memory, or interfacing with other devices. Often used synonymously with *microprocessor* or *core*.

Clock speed

The rate at which a processor completes a certain number of cycles, typically specified in the form of cycles per second (or Hertz).

Instruction set

The set of instructions supported by a given processor.

Multicore processor

A computer chip that has multiple CPUs (called cores).

Multiprocessing

The act of executing multiple processes simultaneously.

Multithreading

The act of executing multiple threads simultaneously.

Process

A copy of a computer program that is being actively executed by the CPU.

Thermal Design Power (TDP)

TDP refers to the maximum power required—or heat generated—by a component, such as a CPU.

Thread of execution

A thread is a stream of instructions. It can also be thought of as a subprocess, in the sense that a single process is typically made up of multiple threads.

IGPU

Configurable power consumption

On an IGPU, the clock rate for the slice and unslice can be controlled separately. This means that unused sections in a GPU can be powered down to reduce power consumption.

Execution Units (EUs)

EUs are processors optimised for multi-threading. Each EU can run up to seven threads simultaneously.

Integrated Graphics Processing Unit (IGPU)

A GPU that is located on a processor alongside the CPU cores and that shares memory with them.

Improved performance when batch processing

IGPUs generally can handle a much larger number of processes at once (as compared to a CPU). Thus, if our data is divided into batches, an IGPU can process multiple batches simultaneously, which can sometimes give a significant boost in performance.

Model precision and speed

On IGPUs, the Execution Unit instruction set and hardware are optimized for 16bit floating point data types. This improves inference speed, as we can process twice as many 16bit operands per clock cycle as we can when using 32 bit operands.

OpenCL Startup Time

When loading a model, OpenCL uses a just-in-time compiler to compile the code for whatever hardware is being used. With an IGPU, this can lead to significantly longer model load times when compared to the same OpenVINO application running on just a CPU.

Slice and unslice

A slice is a collection of 24 EUs. Slices handle computational tasks, such as running inference with OpenVINO. The unslice is the remainder of the IGPU. It's main functions are to support video playback functions.

VPU

Cost

Compared to other AI accelerators, the NCS2 is an inexpensive option, typically costing around $70 to $100.

AI Accelerator

Hardware specifically designed to handle AI requirements and speed up processes used in AI and machine learning.

FPS vs. power tradeoff

The NCS2 is meant to be a low-power device so that it can be easily deployed at the edge; however, one drawback of this is that it cannot process as many frames per second (FPS) as some other devices.

Imaging accelerators

Specialized accelerators in VPUs that have specific kernels used for image processing operations.

Interface / Form factor

Intel VPUs come in a number of different form factors. In the case of the NCS2, the device has a convenient USB3.1 plug and play interface. Note that the NCS2 can be used on systems with only a USB2 port, but the inference will run slower due to I/O throttling.

Interface unit

The part of the VPU that interacts with the host device.

Low power consumption

The processor in the NCS2 (the Myriad X) has a very low power consumption of only 1-2 watts.

Multi-Device plugin (MULTI)

A plugin for sharing load across multiple devices. Can give improved throughput and more consistent performance, assuming there is enough data to share across devices. Example syntax:
MULTI: MYRIAD, CPU, GPU

Neural compute engine

Modern Intel VPUs feature a neural compute engine, which is a dedicated hardware accelerator optimized for running deep learning neural networks at low power without any loss in accuracy.

Neural Compute Stick 2 (NCS2)

A specific implementation of a VPU (in this case the Myriad-X) with 4GB of memory and a USB3 form factor. This is the lowest cost and lowest performing type of accelerator.

On-chip CPUs

VPUs have specialized on-chip CPUs. The Myriad X VPU has two: one used to run the host interface and the other is used for on-chip coordination between the Neural Compute Engine (NCE), the vector processor, and the imaging accelerators.

Precision

The NCS2 only supports FP16 model precision.

Vector processors

Processors that work on a vector or an array of 1D data. They can be contrasted with scalar processors, which often work on single data items.

Vision Processing Units (VPUs)

Accelerators that are specialized for AI tasks related to computer vision—such as Convolutional Neural Networks (CNNs) and image processing.

FPGA

Application-Specific Integrated Circuits (ASICs)

Chips that are hardwired during manufacturing in order to be optimally efficient for a specific need.

Bitstream

Bitstreams (i.e., binary sequences) are used as input to an FPGA to do the actual configuration of the logic blocks and interconnects; in other words, bitstreams are used to program FPGAs.

Field-Programmable Gate Arrays (FPGAs)

Chips designed with maximum flexibility, so that they can be reprogrammed as needed in the field (i.e., after manufacturing and deployment).

Heterogeneous (HETERO) plugin

Plugin that allows you to specify the primary device, as well as one or more fallback devices that should be used in the event that the primary device does not support the layers in your model.

High performance, low latency

Once programmed with a suitable bitstream, FPGAs can execute neural networks with high performance and very little latency.

Large networks

One feature of FPGAs that makes them especially useful in deep learning is that they can support large networks, with a capacity to handle networks that have more than 2 million parameters.

Programmable I/O Blocks

Programmable I/O Blocks connect an FPGA tile to an external circuit for input and output. These external circuits are external to the current tile, but still internal to the overall FPGA. They can be other tiles, Digital Signal Processing blocks (DSPs), memory blocks, or even more I/O blocks.

Register Transfer Level / Register Transfer Language

The lowest level of abstraction for programming an FPGA, which—as the name suggests—involves specifying the configuration of the actual hardware registers.

Flexibility

FPGAs are flexible in that they are reprogrammable, support various precision options, and allow for the bitstreams being used to be updated without changing the hardware.

Configurable Logic Blocks (CLBs)

Configurable Logic Blocks (CLBs) form the core of the FPGA. Each block can implement its own function using look up tables. The logic blocks also contain flip flops, transistor pairs, and multiplexers.

Long lifespan

FPGAs have a long lifespan. FPGAs that use devices from Intel’s Internet of Things Group have a guaranteed availability of 10 years, from start of production.

Programmable Interconnects

Programmable Interconnects steer the input and outputs of CLBs in an FPGA. Each CLB is interconnected in four directions—and we achieve the ability to program the logic through the ability to switch these connections on and off.

Robust

FPGAs are designed to have 100% on-time performance, meaning they can be continuously running 24 hours a day, 7 days a week, 365 days a year. They are also able to function over a wide range of temperatures, from 0° C to 60° C. This means that FPGAs can be deployed in harsh environments like factory floors and still perform optimally.