**What is a microprocessor(CPU)?**

A CPU at its most basic level is a device with a number of named memory cells called registers and a number of computational units called arithmetic logic units (ALU). The ALUs perform things like addition, subtraction, and other basic math operations. However, these are only connected to the CPU registers. If you want to add up two numbers, you have to get those two numbers from memory and into two registers in the CPU.

Diagram

Description automatically generated

For a long time, cheap computer systems have had the CPU and GPU integrated into the same chip (same silicon die). These have been famously slow. In the past saying, “integrated graphics” was essentially the same as saying “slow graphics.”

These where slow for severals reasons:

Separate areas of this memory got reserved for the CPU and GPU. If the CPU had a chunk of data it wanted the GPU to use, it couldn’t say “here have some of my memory.” No, the CPU had to explicitly copy the whole chunk of data over the memory area controlled by the GPU.

The second problem was that large GPUs produce a lot of heat and thus you cannot integrate them with the CPU without getting problems ridding yourself of the heat produced. Thus discrete graphics cards tend to look like the one below:

A picture containing text, fan, device, appliance

Description automatically generated

Large beasts with massive cooling fans. They have special dedicated memory designed to serve the greedy cards massive amounts of data. That is why these cards have high performance. But they have an achilles heel: Whenever they have to get data from the memory used by the CPU, this happens over a set of copper traces on the computer motherboard called a PCIe bus. Try chugging water through a super thin straw. It may get to your mouth fast, but the throughput is totally inadequate.

Having a *Unified Memory Architecture* tries to solve all these problems without having the disadvantages of old school shared memory. It can be achieved through the following ways:

1. There is no special area reserved just for the CPU or just the GPU. Memory is allocated to both processors. They can both use the same memory. No copying is needed.
2. Using memory which serves both large chunks of data and serves it fast. In computer speak that is called low latency and high throughput. Thus the need to be connected to separate types of memory is removed.
3. Through technological advancement the watt usage of the GPU has significantly reduced, so that a relatively powerful GPU can be integrated without overheating the SoC. And ARM chips produce less heat, allowing the GPU to have a higher heat budget than a GPU on the same silicon die.

There is of course a tradeoff in this strategy. Getting this high bandwidth memory (big servings) require full integration which means you take away the opportunity from customers to upgrade their memory.

Another way of increasing performance in a CPU is by adding more specialized chips doing a few specialized tasks instead of adding ever more general-purpose CPU cores. The benefit of this is that specialized chips tend to be able to perform their tasks significantly faster using much less electric current than a general-purpose CPU core.

This is not entirely new knowledge. For many years already specialized chips such as the graphical processing units (GPUs) have been sitting in Nvidia and AMD graphics cards performing operations related to graphics much faster than general-purpose CPUs. What we can do is take a more radical shift towards this direction. Rather than just having general-purpose cores and memory, the chip would contain a wide variety of specialized chips:

* Central processing unit (CPU) — the “brains” of the SoC. Runs most of the code of the operating system and your apps.
* Graphics processing unit (GPU) — handles graphics-related tasks, such as visualizing an app’s user interface and 2D/3D gaming.
* Image processing unit (ISP) — can be used to speed up common tasks done by image processing applications.
* Digital signal processor (DSP) — handles more mathematically intensive functions than a CPU. Includes decompressing music files.
* Neural processing unit (NPU) — used in high-end smartphones to accelerate machine learning (A.I.) tasks. These include voice recognition and camera processing.
* Video encoder/decoder — handles the power-efficient conversion of video files and formats.
* Secure Enclave — encryption, authentication, and security.
* Unified memory — allows the CPU, GPU, and other cores to quickly exchange information.

A user working on images and video editing using this chip’s design see speed improvements. A lot of the user’s tasks can be run directly on specialized hardware. A chip with this design can encode a large video file without breaking a sweat while an expensive PC has all its fans going full blast and still cannot keep up.