Why GPUs have evolved as the general-purpose computing devices:

A **supercomputer** is a computer with **high level of performance** compared to a ***general-purpose computer***.

**JUST KNOW ABOUT THE EXISTENCE OF THE NVIDIA TESLA GPUS.**

SKIP TO MINUTE 3:46

**What is it GPU?**

GPU or graphic processing unit is a **specialized computer unit** which ***addresses the demand of rendering real-time high-resolution 3D graphics***.

A close-up of a camera

Description automatically generated with medium confidenceTo address this demand, GPUs are designed in a different way than traditional CPUs.

**Differences between CPU and GPU:**

A picture containing graphical user interface

Description automatically generated

CPUs are enhanced to process instruction quickly. It is a latency device, and it has very small number of computing cores.

The CPU shown in this diagram have only **two cores** or **two computing units.**

Now these cores include *advanced optimization hardware* like **branch predictors** and **multi-level caches**. The **MAIN** focus is to ***run the instruction faster***.

A picture containing graphical user interface

Description automatically generated

GPUs are designed ***with light thousands of cores*** to ***render thousands of pixels at once***.

These cores do not include mechanisms like **branch predictors** and have ***relatively less clock speed than CPUs***. However, GPUs have ***thousands of cores*** (have a massively parallel architecture) so that it **can execute thousands of threads in parallel**. We call this type of devices as ***throughput devices***.

Table

Description automatically generatedSince the main purpose of GPU is to work on thousands of pixels at same time the **design** *includes the hardware to perform* ***context switching in zero clock cycle***.

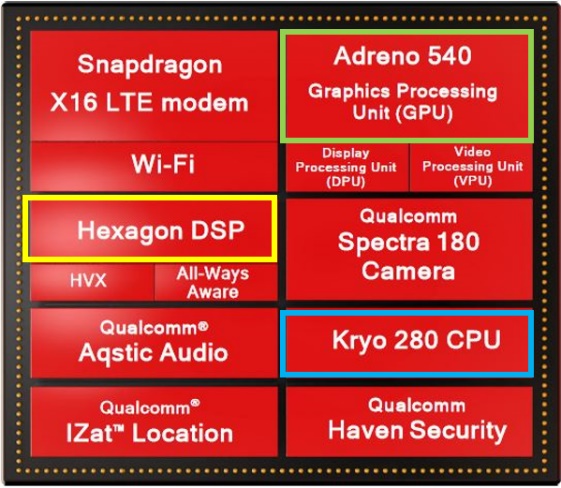
But in ***CPU context switching happen using software*** so **it will consume hundreds of clock cycles in CPU**. So, *if you have much more threads running in CPU than available cores*, ***performance will be degraded due to the context switching***.

But in GPUs, ***the presence of more threads than the available computing cores*** will ***enable much efficient execution***, as you will see in the upcoming videos.

***Heterogeneous Computing***

A ***Heterogeneous computing system*** refers to a **system** that ***uses more than one kind of processors or cores***. These systems **give us an efficient way of performing highly parallel computations**.

This diagram shows you the components of a very cheap smartphone available in current market.

There are multiple computer processors in this diagram. We have:

* DSP or digital signal processing (Hexagon DSP)
* The main processor itself (Kryo 280 CPU)
* A graphic processor (Adreno 540 GPU)

**The current trend is to use all of these available computing power efficiently in general purpose way**.

**Ex**: *we can perform some computations* which ***are intended to execute in CPU*** on ***the DSP or the GPU when they are free***. This **allows us to harness(control) the total computing power of our overall device**. This kind of **multiprocessor environment** ***enables us to perform tasks using many types of processors***.