GPU accelerated computing for Finite Element Method
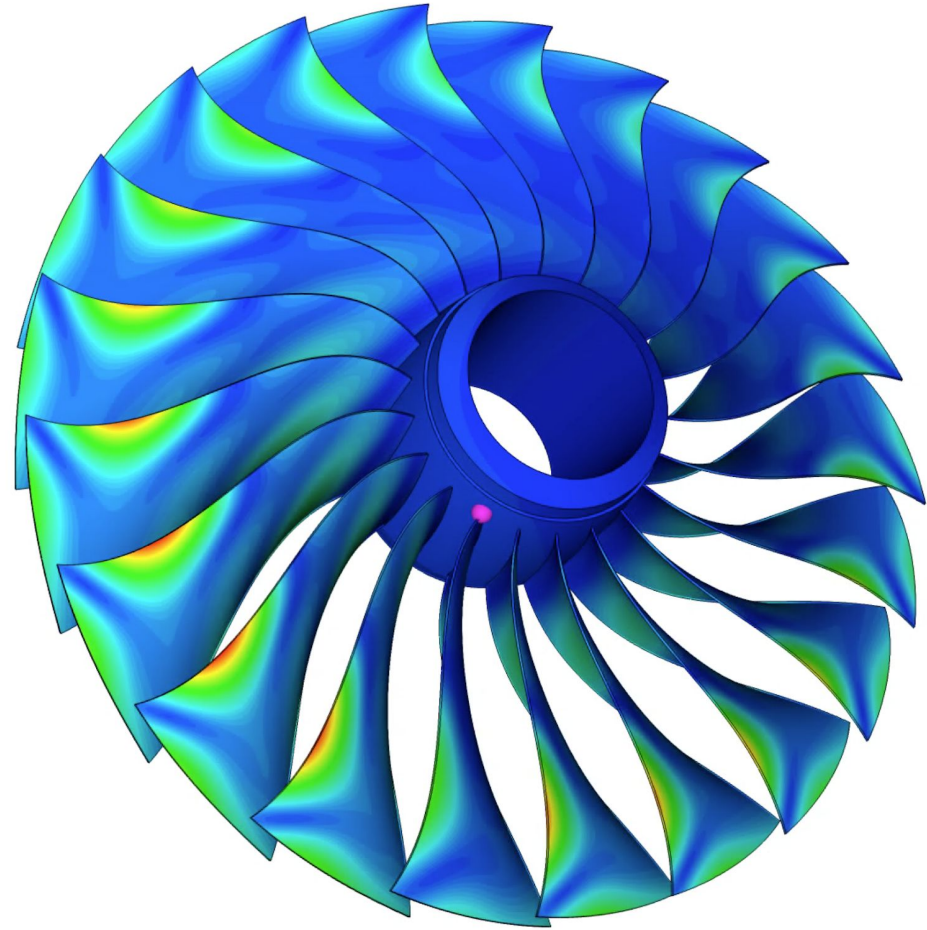
Introduction to GPU
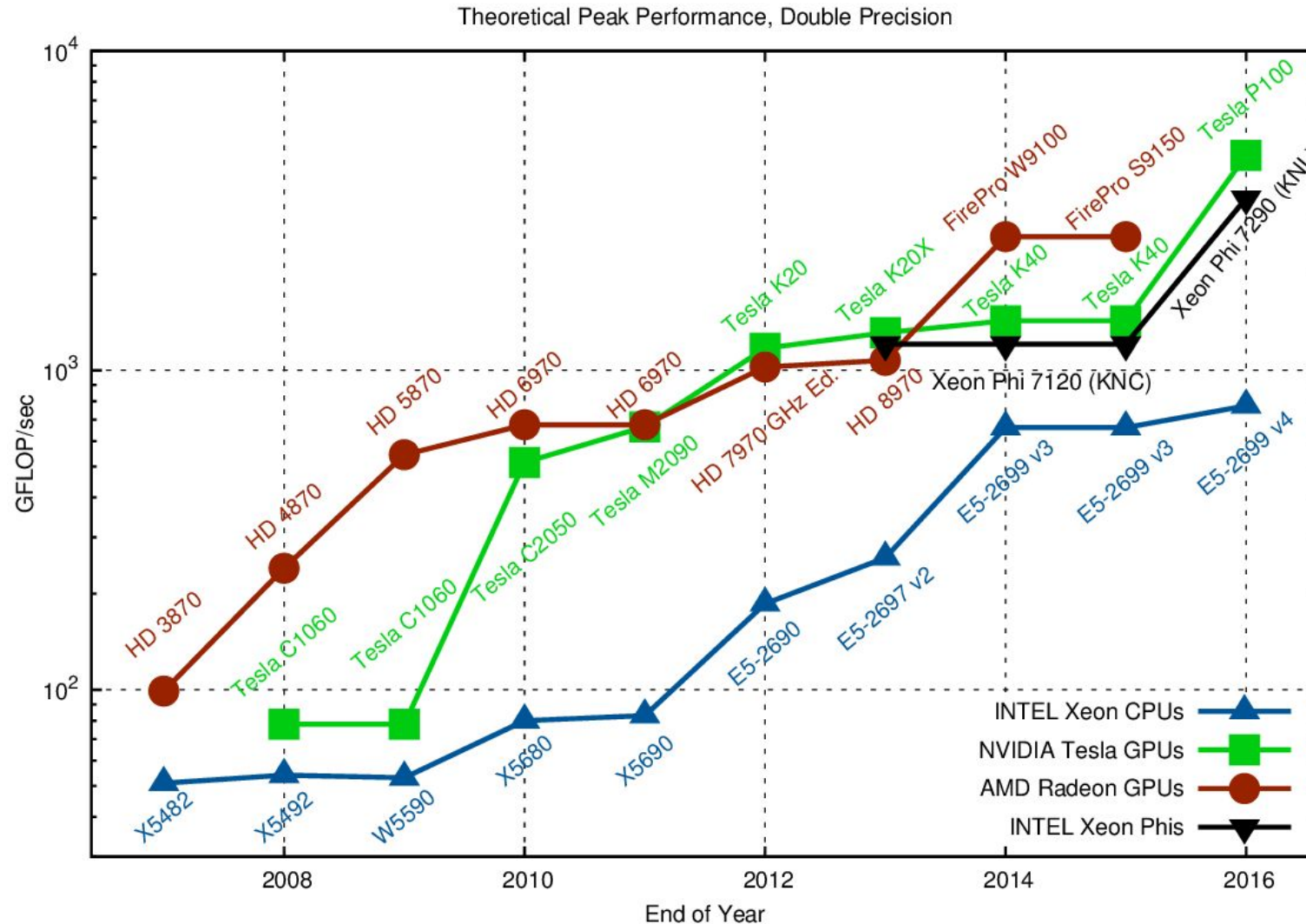
Objectives

- Understand the difference between CPU and GPU

- The advantages of utilizing GPUs

Introduction to GPU
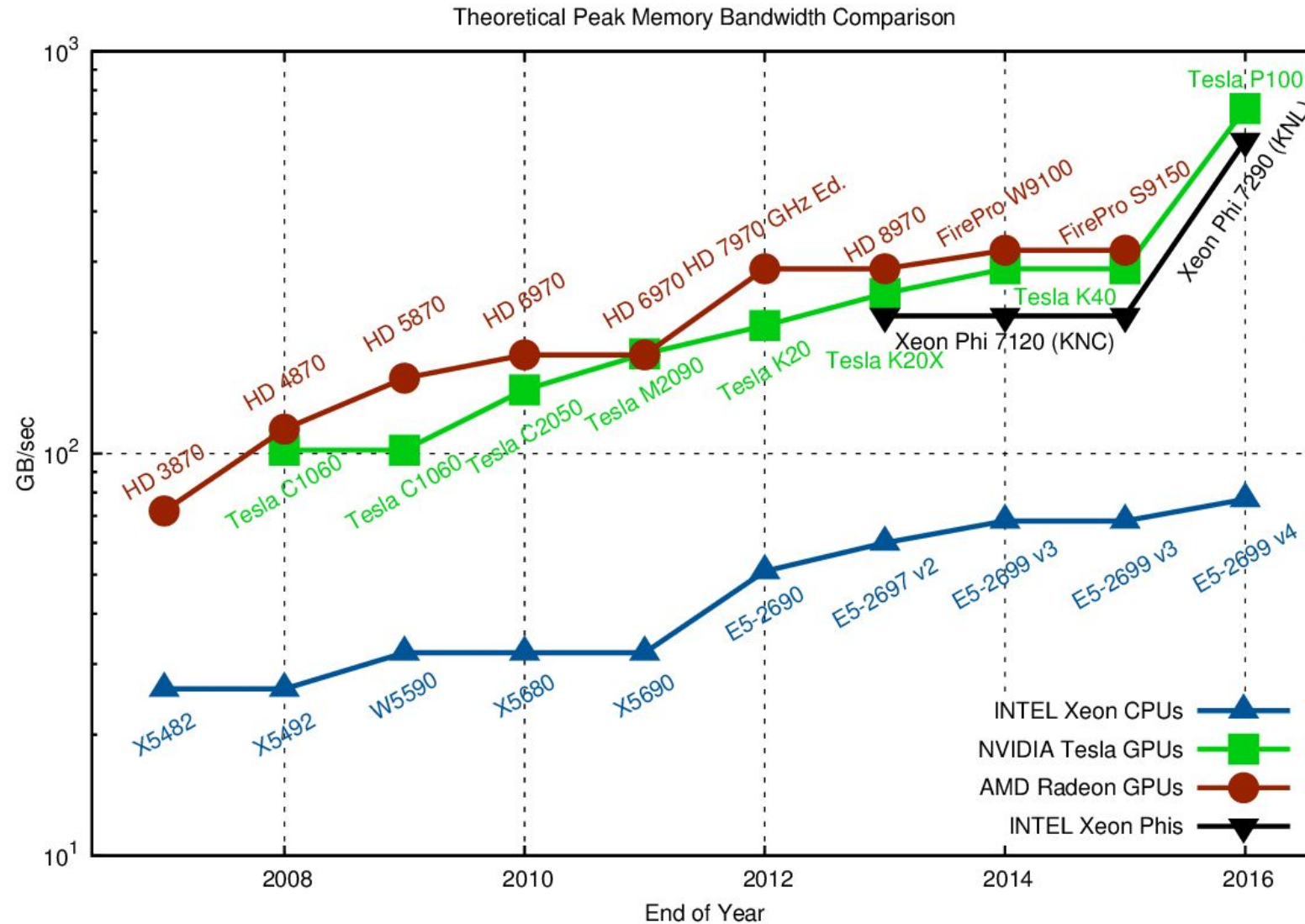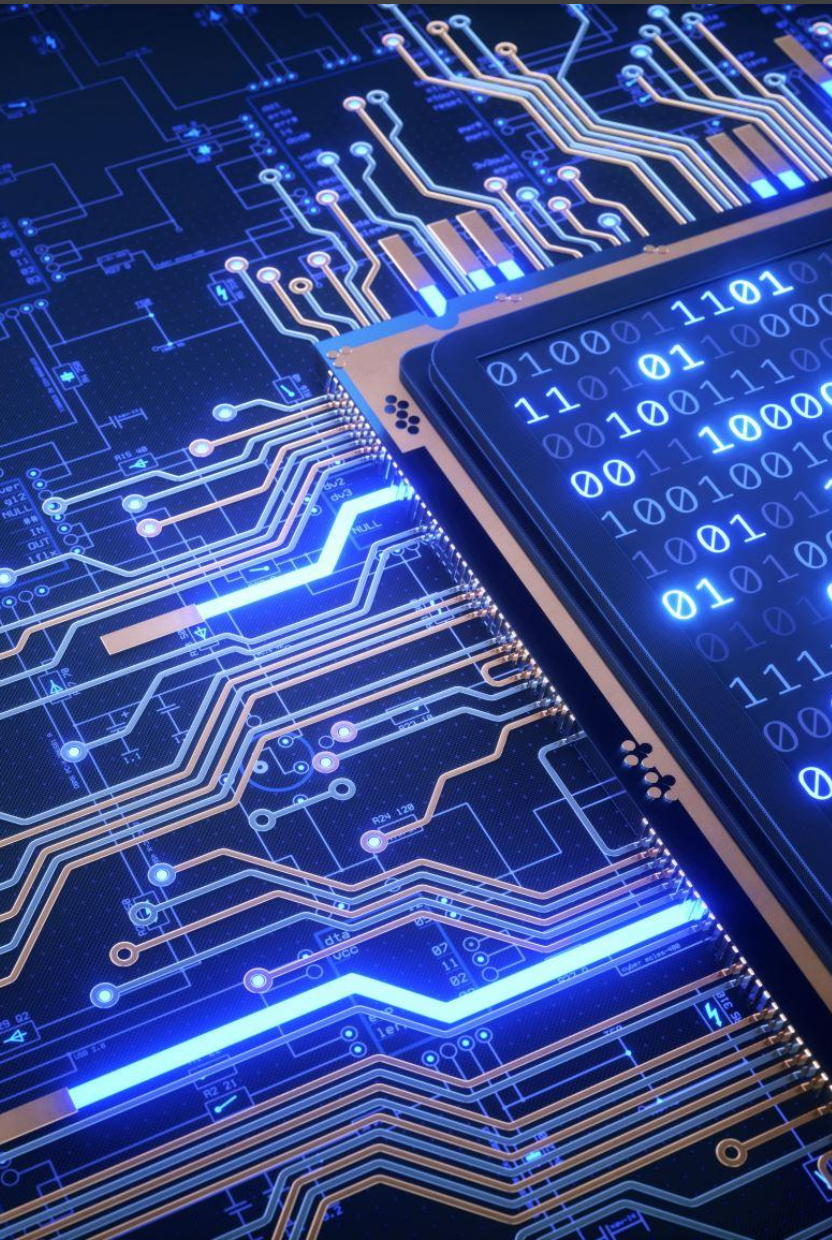
Theoretical Peak Performance, Double Precision

Karl Rupp. Pictures: CPU/GPU Performance Comparison. URL: https://www.karlrupp.net/2013/06/cpu-gpu-and-mic-hardwarecharacteristics-over-time/

Introduction to GPUs – Memory Bandwidth

Theoretical Peak Memory Bandwidth Comparison

Karl Rupp. Pictures: CPU/GPU Performance Comparison. URL: https://www.karlrupp.net/2013/06/cpu-gpu-and-mic-hardwarecharacteristics-over-time/
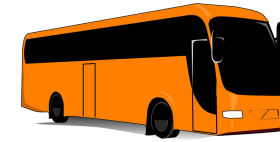
Introduction to GPU

- Latency   -   time to finish a given task
- Throughput - number of tasks in a given time

Example: transport people 25 kms and back

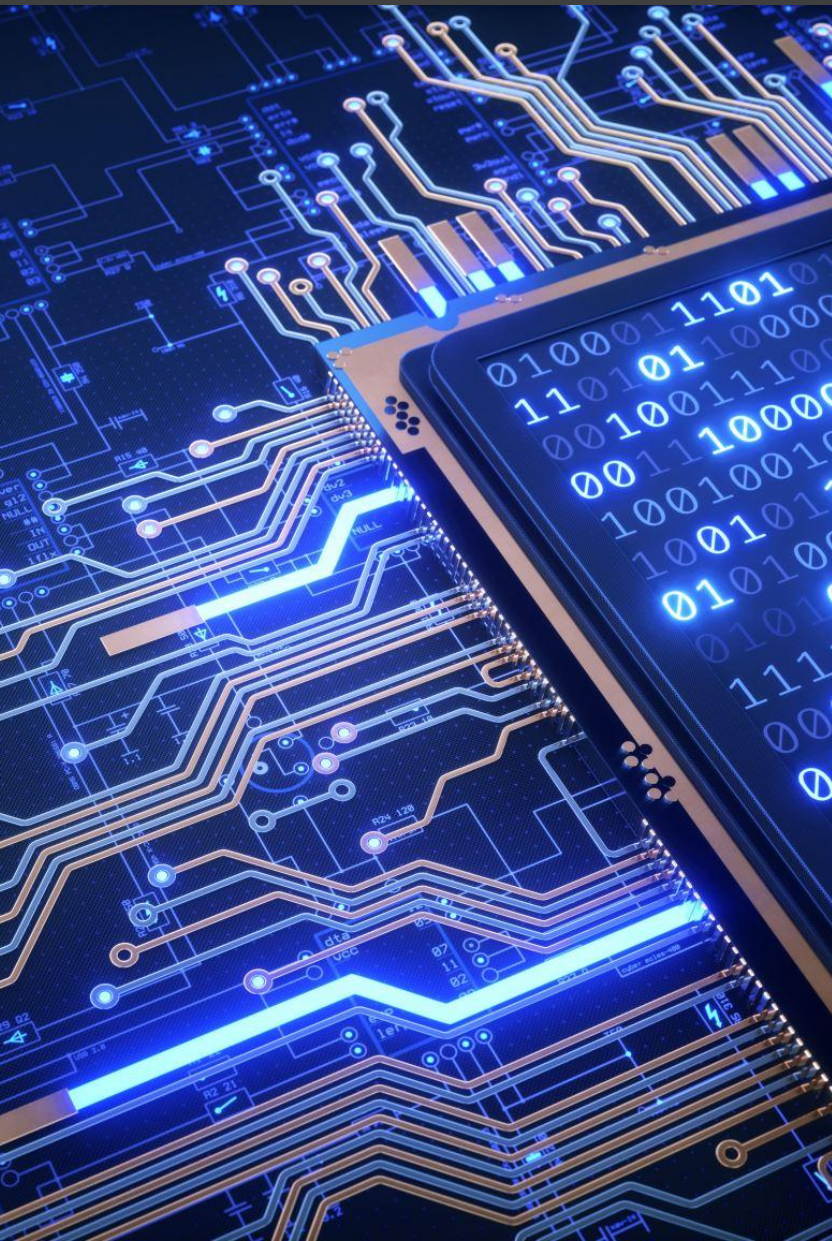| Car		Bus	
Capacity = 5		Capacity = 100	
Speed = 100 km/hr		Speed = 50 km/hr	
Latency	15 min	Latency	30 min
Throughput*	10	Throughput*	100

* For this example, throughput is measured as number of people transported in an hour

Introduction to GPU
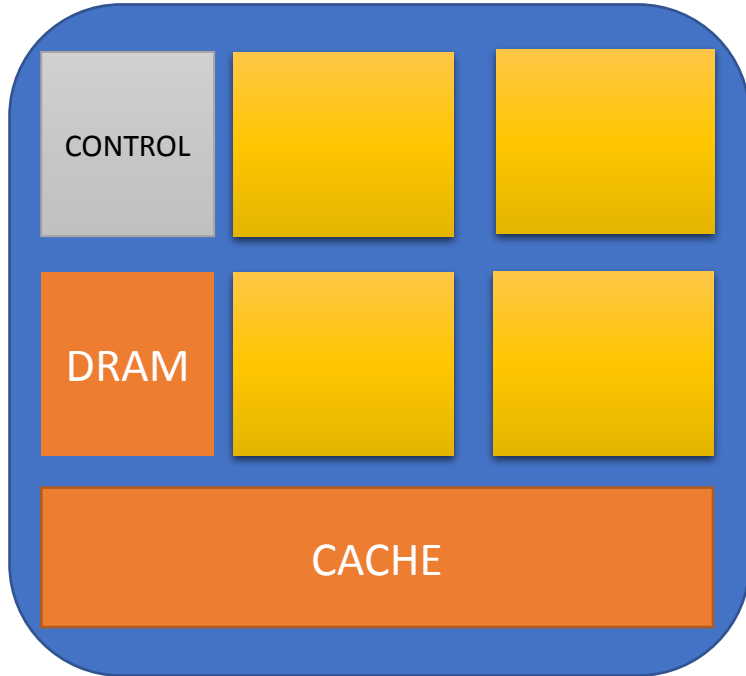
	CPU	GPU
Abbreviation	Central Processing Unit	Graphics Processing Unit
Performance	High latency - Focuses its cores on individual tasks and on finishing the tasks fast	High throughput - Delivers massive performance for tasks which can be divided and run in separate cores
Instructions	Serial instructions	Parallel instructions
Number of cores	Few powerful cores	Many smaller and weaker cores
Type of tasks	Executes wide range of commands and processes. Suitable for wide variety of tasks	More specialized cores

Introduction to GPU
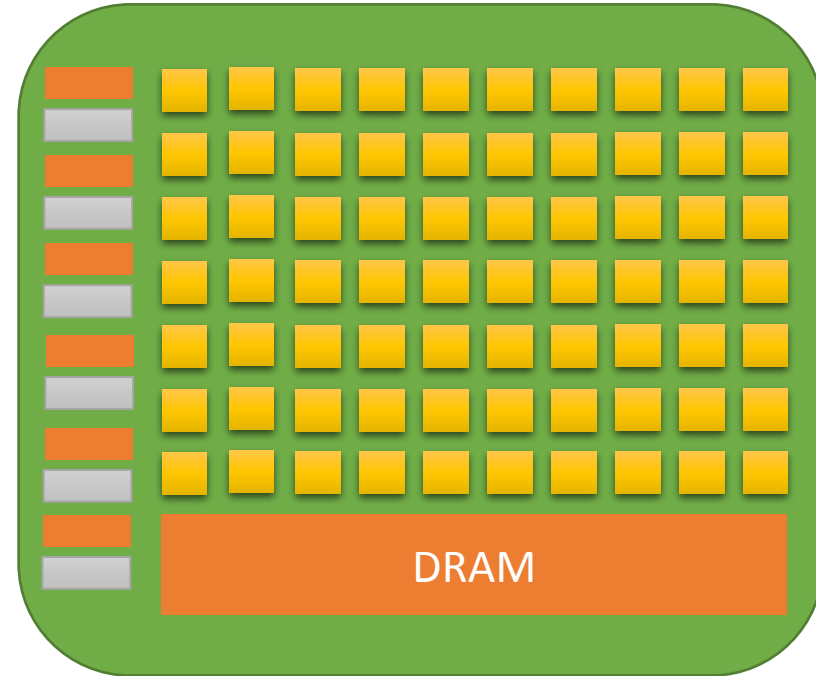
CPU

CPUs are latency oriented device

Sequential cores
CPUs can execute a sequential
code many times faster than GPUs

GPU

GPUs are throughput oriented device
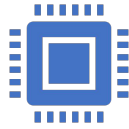

Several parallel cores
GPUs can be execute a parallel
code many times faster than
CPUs

Introduction to GPU

- GPUs:
  - ✔ Massively parallel
  - ✔ Many cores (Hundreds)
  - ✔ Many threads (Thousands)
  - ✔ Programmable

TYPES

Integrated

GPU is built into the processor.

It does not use a separate memory bank. The system memory is shared with CPU is used.

Since it is built into the processor, it uses less power and therefore creates less heat.

It is usually found in small devices like laptops, tablets, etc.

External or Discrete

It is separate from the processor.

It has it own dedicated memory that is not shared with CPUs.

It consumes more power and generates a significant amount of heat.

It provides higher performance than integrated graphics. It is most found in desktop PCs.