GPU for FEM

Sivasankar Arul, IT4Innovations

June/2016

Univerza *v Ljubljani*

TECHNISCHE UNIVERSITÄT WIEN

CINECA
consorzio interuniversitario

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

Objectives

Objectives

- Understand the difference between CPU and GPU

- The advantages of utilizing GPUs

Peak Performance
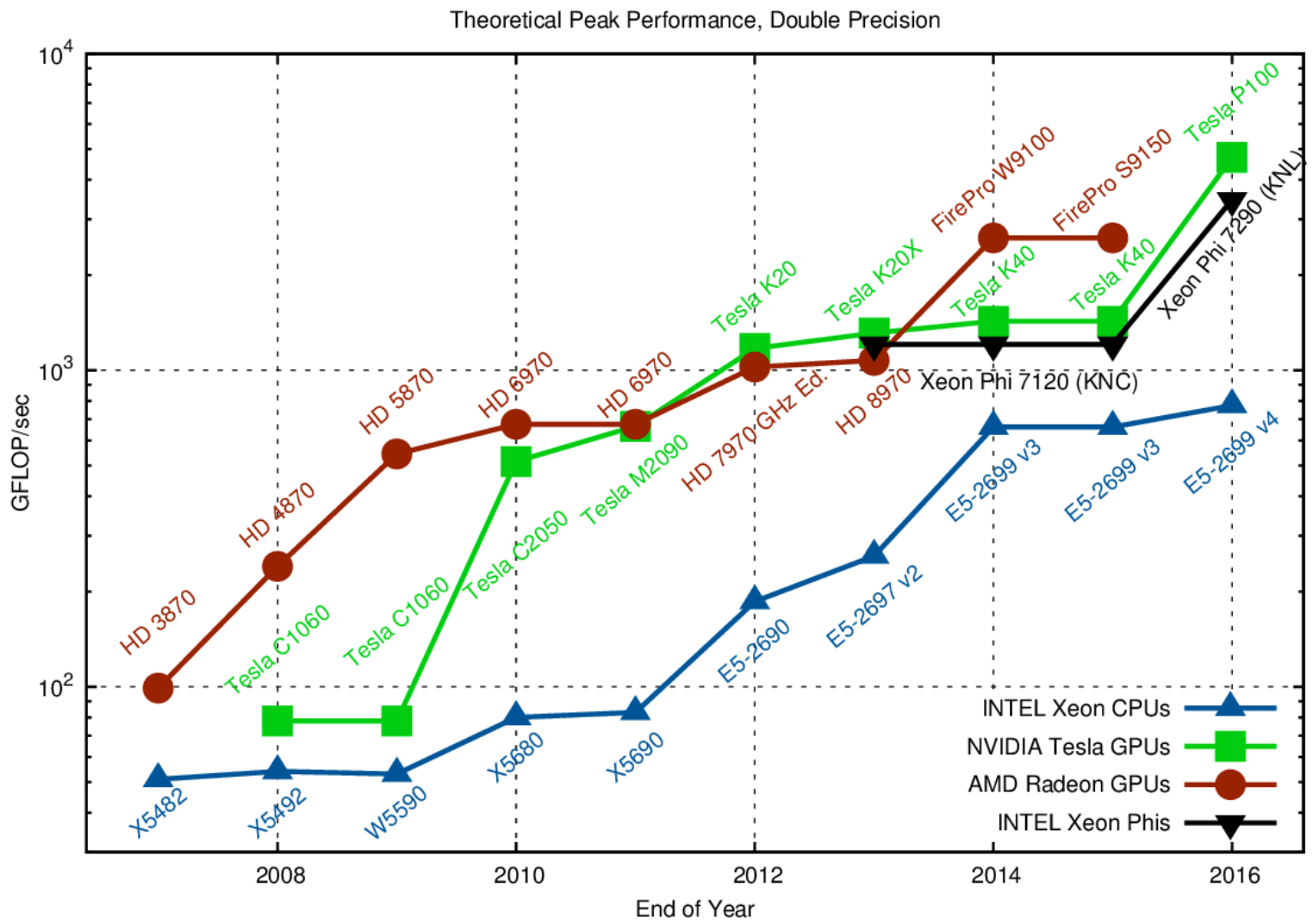
Theoretical Peak Performance, Double Precision

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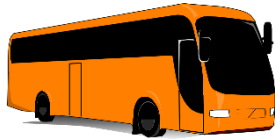
Memory Bandwidth

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/
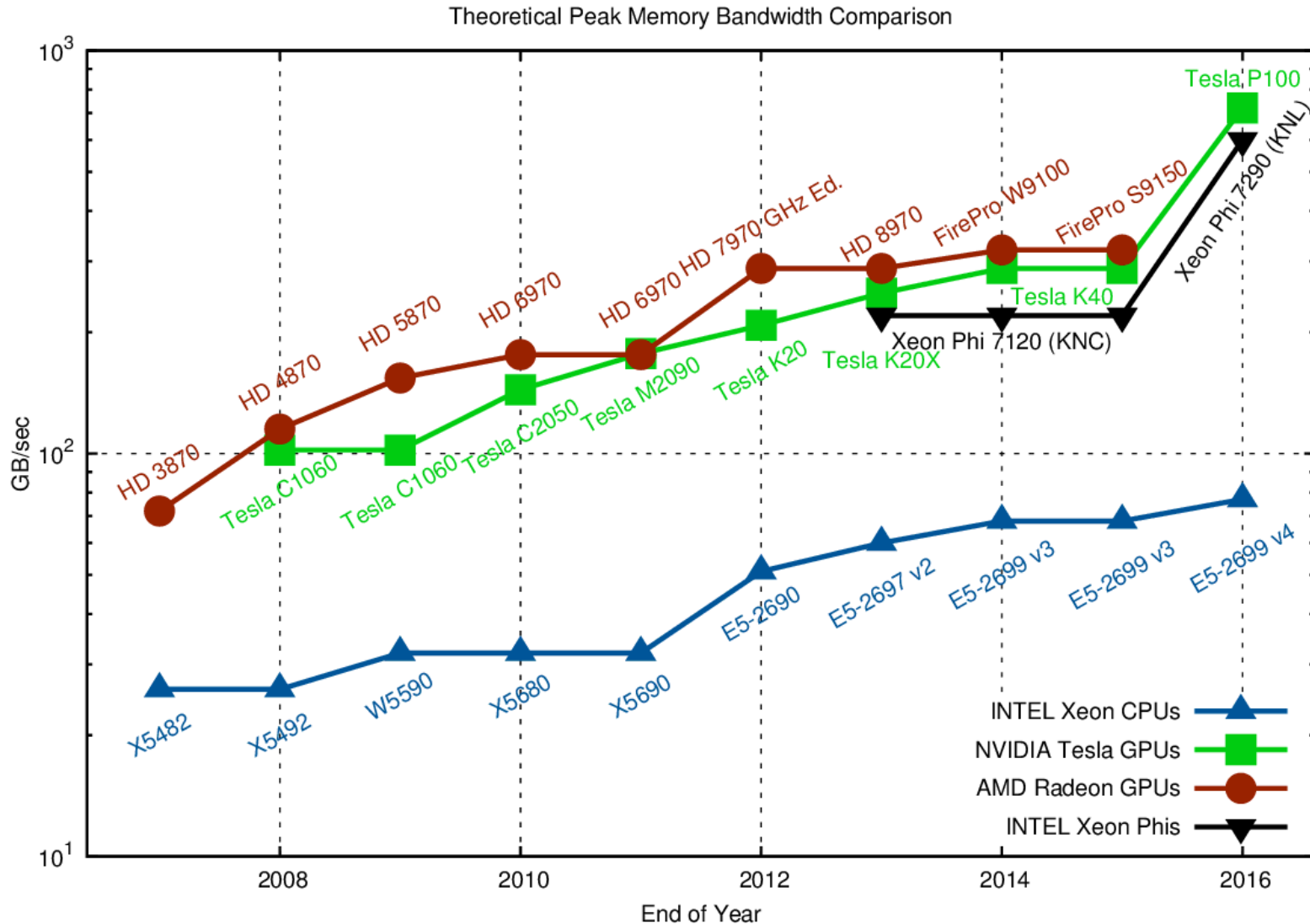
Latency and Throughput

- Latency   - time to finish a given task

- Throughput - number of tasks in a given time
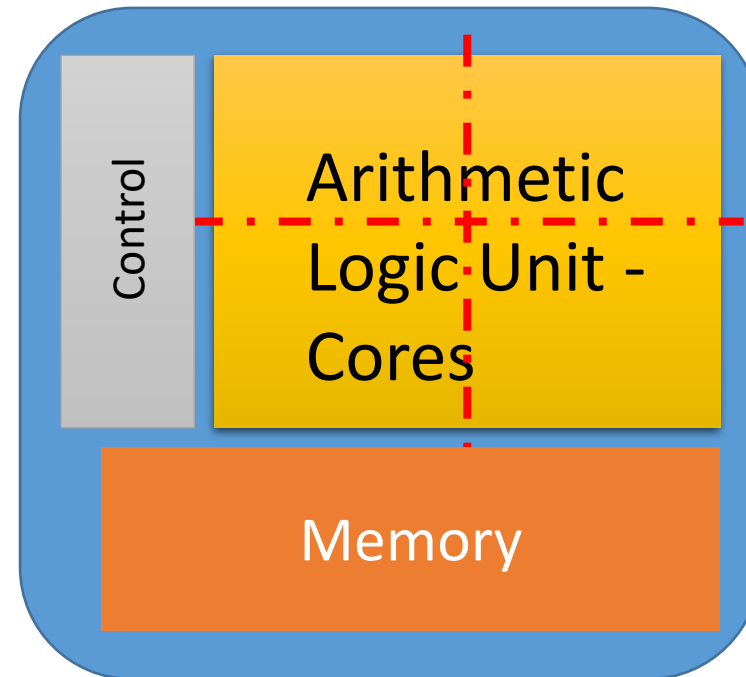
Example: transport people 25 kms and back

Car		Bus	
Capacity – people transported	5	Capacity – people transported	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
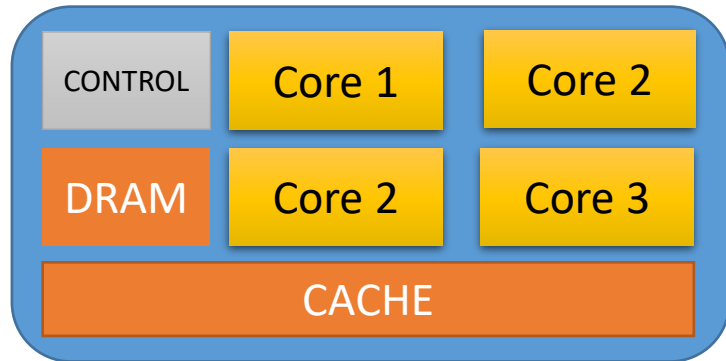
CPU vs GPU

|	CPU	GPU
Abbreviation	Central Processing Unit	Graphics Processing Unit
Performance	Low 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
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. Suitable for parallel data processing

Basic Layout of Processing Unit

CPU vs GPU

CPUs are latency oriented device

CONTROL	Core 1	Core 2
DRAM	Core 2	Core 3
CACHE		

GPUs are throughput oriented device

CONTROL

CONTROL

CONTROL

DRAM

- Few big cores

- Each core has a dedicated unit for control
- Can handle complicated programs

- Any type of workload can be run efficiently

- CPUs can execute a complex workload many times faster than GPUs due to dedicated control unit.
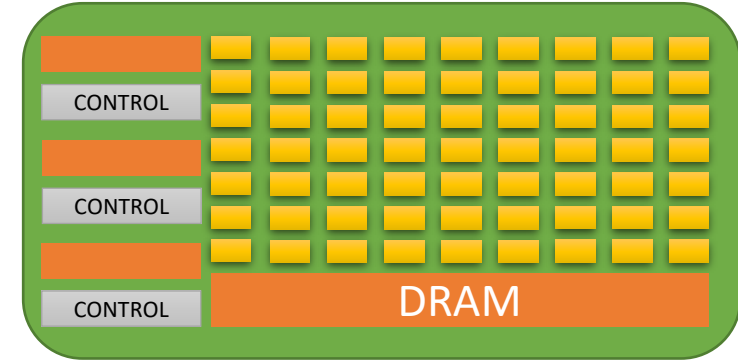
- Several small cores

- Control unit for a set of cores

- Only basic operations can be run efficiently

- GPUs can execute only simple workload many times faster than CPUs due to large number of cores and limited control unit.

GPU

https://www.anandtech.com/tag/gpus

- GPU has good computational power

- It has significantly faster memory – High Bandwidth Memory.

- Therefore, a computation that needs a lot of memory access is efficient in GPU.

Tensor Core : reduced precision useful for machine learning applications
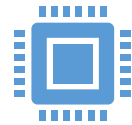FP32   : float
FP64   : double

SFU   : Special Function Units - very few

Types of GPUs

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.

Thank you for your attention!

http://sctrain.eu/