THE UNIVERSITY OF WARWICK

High Performance Computing

Course Notes

GPU and CUDA - I

Dr Ligang He

GPU

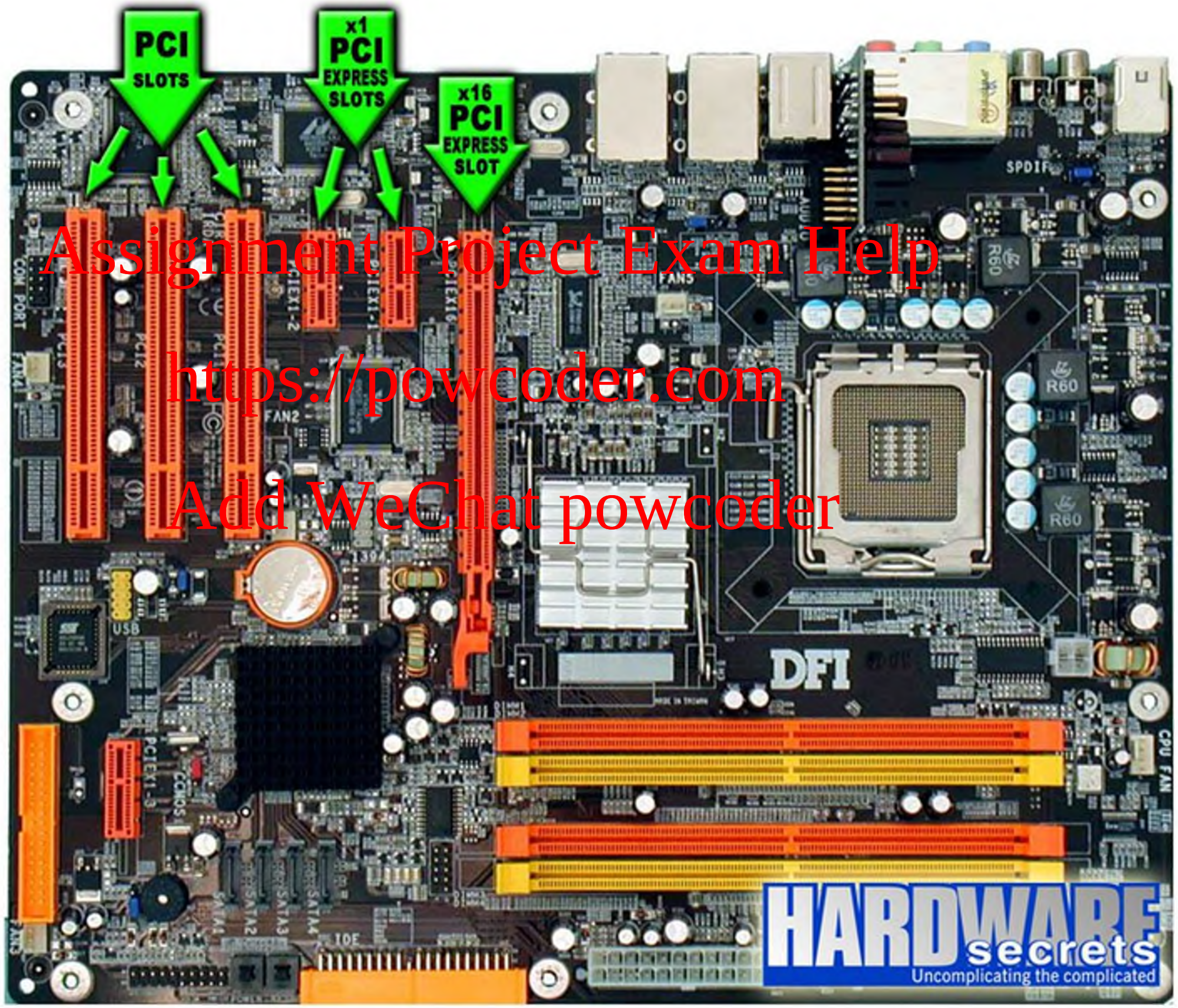
- Graphics processing unit

- Contains a large number of ALUs

  - ❑ 2560 ALUs (stream processors) in Nvidia GeForce GTX 1080

- Is a PCI-e peripheral device

PCI-e slot


PCI SLOTS
x1 PCI EXPRESS SLOTS
x16 PCI EXPRESS SLOT
DFI
HARDWARE secrets
Uncomplicating the complicated

Performance Trend

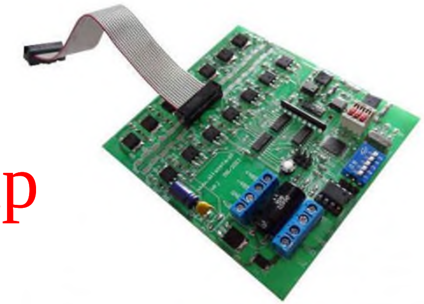
- Many-core GPU is 100x more powerful than multicore CPU

- Why is there such performance gap?

  ❑ Because of the differences in the design between GPU and CPU

Design of CPU

- The design objective of CPU is to optimize the performance of a sequential code

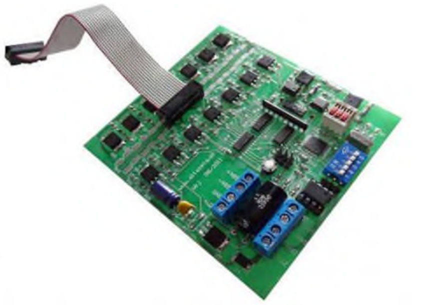


  - Has complicated control unit

    - Obtains instructions from memory
    - Interprets the instructions
    - Figure out what data are needed by instructions and where it is stored
    - Issues signals to ask other functional units (ALUs) to run the instructions

Design of CPU

- The design objective of CPU is to optimize the performance of a sequential code



  - Has complicated control unit

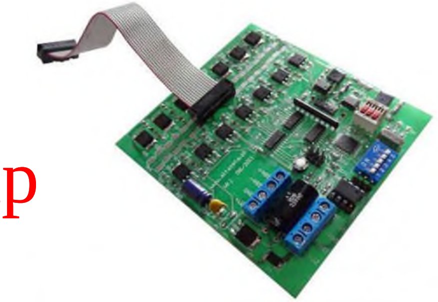
  - Complicated control unit enables

    - instructions from a single thread to execute out of their sequential order (single core) or in parallel (multicore)

    - branch prediction
    - data forwarding

Design of CPU

- The design objective of CPU is to optimize the performance of a sequential code

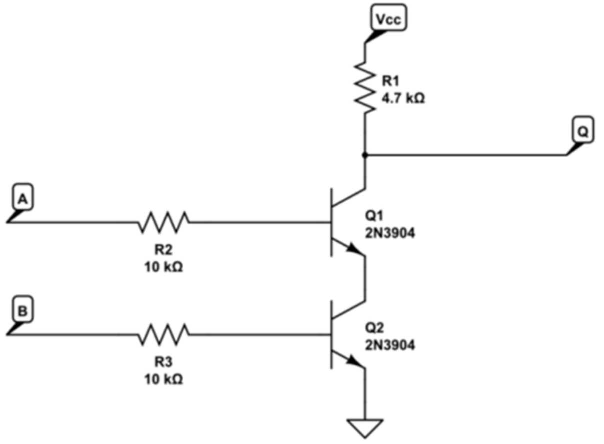


  - Has complicated control unit
  - Complicated control unit enables
  - Has large cache to reduce the instruction and data access latencies
  - Powerful ALU

Design Objective of CPU

- **Latency-oriented design**
  - □ Large on-chip caches
  - □ Complicated control unit
  - □ Complicated arithmetic logic unit
  - □ They are at the cost of increased use of chip area and power

- **Applications with one or very few threads achieve higher performance in CPU**



NAND gate with transistors

Motivation of GPU Design

- Video game industry: need to perform a massive number of floating-point calculations per video frame

- Motivate GPU vendors to maximize the chip area and power dedicated to floating point calculations

❑ Each calculation is simple: therefore simple control logic and simple ALUs

❑ Calculation is more important than cache, therefore small cache, allowing memory access to have long latency

GPU Design

- GPU has a large number of ALUs on a chip to increase the total throughput

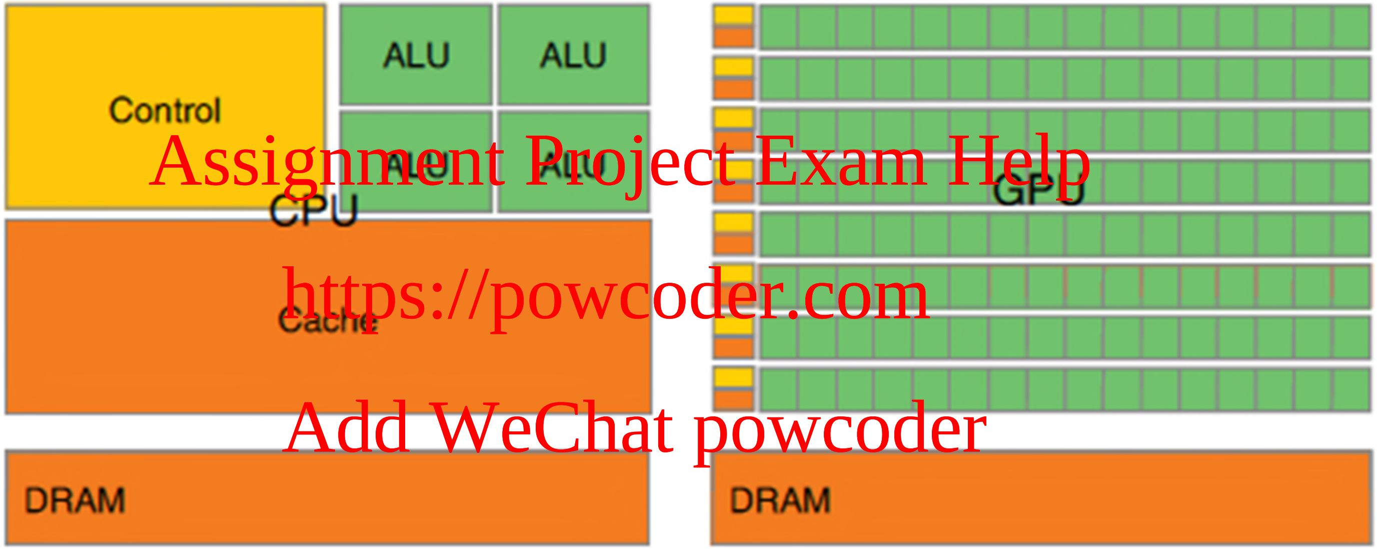
❑ The application is run with a large number of parallel threads

❑ While some threads are waiting for long-latency operations (e.g., memory access), the GPU can always find other threads to run due to the large number of threads

❑ Throughput-oriented design: maximize the total throughput of a large number of threads, allowing individual threads to take a longer time

- GPU adopts the throughput-oriented design

GPU vs. CPU in Architecture


Control
ALU
ALU
ALU
ALU
CPU
Cache
DRAM
GPU
DRAM