Introduction to GPUs
GPUs

• Graphics Processing Units (GPUs) are co-processors.
  ▪ A hardware that runs programs in addition to CPU.
  ▪ Originally, those programs displayed graphics.
  ▪ GPUs run programs with a lot of computations.

• CPU uses task parallelism while GPU uses data parallelism.
GPUs versus CPUs

- Task parallelism
  - Threads perform different computations on the same or different data.
  - An array \( A \) of size \( N \) \((A[0] \ldots A[N-1])\)
  - Thread 1 computes the sum of \( A[0] \ldots A[N] \), but Thread 2 computes the average of \( A[0] \ldots A[N] \).

- Data parallelism
  - Threads perform the same computation on different parts of the same data.
  - E.g., the password cracker lab uses the data parallelism.

- Task parallelism distributes tasks. Data parallelism distributes data.
GPUs

- GPU is an array of streaming multiprocessors (SMs)
- A multithreaded program is divided into blocks of threads.
- Each block executes independent of the other blocks.
CUDA

• A programming model that uses data parallelism for speeding up program executions.
  – Used as an extension for a programming language

• Uses GPUs to execute the parallel portion of compute-intensive programs

• Originally built for graphics application, but later used for general-purpose programs
CUDA

• Programs developed by CUDA run on GPU and CPU.
  – **Host**: CPU
  – **Device**: GPU
  – CPU and GPU have their own memory spaces.

• Programs launch **kernels**
  – Kernels are functions executed on GPUs.
  – When launched, threads are generated.
  – A kernel data is divided and processed by multiple threads in parallel.
Thread Hierarchy

- Thread blocks are mapped to SMs.
- Thread blocks run in an order.
- Threads of a block run in parallel.

- The thread block is the basic unit of execution
Memory Hierarchy

Thread ←------→ Per-thread *local* memory

Thread block ←-----→ Per-block *shared* memory

Grid

Block (0, 0)  Block (1, 0)
Block (0, 1)  Block (1, 1)

Global memory
Heterogeneous System

C program

Serial code

Parallel code (kernel)

Serial code

Host

Device

Grid

Block (0, 0)  Block (1, 0)

Block (0, 1)  Block (1, 1)

Host