1. Nvidia GPU architecture
    a. GPU memory architecture: global memory, constant memory, shared memory, register file; their accessibility to kernel threads and host code.
        i. Memory coalescing when accessing global memory.
        ii. Shared memory bank conflicts
    b. Streaming multiprocessor architecture: how thread blocks are scheduled to SMs?
2. CUDA thread grid configuration and runtime behavior
    a. Thread grid configuration, thread block configuration.
    b. How the threads in a thread block communicate and synchronize?
    c. How the threads in a thread block are grouped to warps?
    d. The runtime behavior of threads in a warp.
        i. Control flow divergence.
    e. The limitations regarding the number of threads in a thread block, the number of resident wraps and the number of resident thread blocks on an SM.
    f. Understand the zero-overhead context switch on GPU; and the resource requirements of a thread and a thread block.
3. Various typical parallel applications
    a. Vector addition
    b. Matrix multiplication
    c. Convolution
    d. Pre-fix scan
    e. Histogramming
4. Floating-point representation
    a. Understand how to represent floating-point numbers following IEEE standard
    b. Convert fractions between base 10 and base 2
5. Overlapping communication and computation using streams
6. Basic concepts of OpenCL