

※ 考生請注意：本試題不可使用計算機。請於答案卷(卡)作答，於本試題紙上作答者，不予計分。

1. [20%] Please briefly show and discuss the architectures and pipelines for CPU and GPU.
2. [10%] What are (a) heterogeneous computing and (b) data parallel computing?
3. [20%] Consider two matrices A and B, where A and B are by $M \times N$ and $N \times P$, respectively. Please illustrate a C program (extended by some GPU programming framework that you are familiar with) to perform the matrix multiplication of A and B, resulting in a matrix C by $M \times P$, with CPU accelerated by GPU.
4. [20%] Synchronization is an important mechanism to control concurrent accesses to critical sections by multiple processes/threads. Please answer the following questions based on the semaphore-based software functions listed below, where the functions can be used by multiple threads to make sure that no thread could read or write the shared data while another thread is writing to it.

```
semaphore wmut=1;
semaphore rmut=1;
rcount=0;

void writer() {
    wait(wmut);
    // Write things on the shared data.
    signal(wmut);
}

void reader() {
    wait(rmut);
    rcount+=1;
    if (rcount == 1)
        wait(wmnt);
    signal(rmut);
    // Read things from the shared data.
    wait(rmut);
    rcount-=1;
    if (rcount == 0)
        signal(wmnt);
    signal(rmut);
}
```

- (1) [5%] Is the above pseudo code representing the optimal solution to the readers-writers problems? Why?
 - (2) [5%] Assume that there are three threads, a reader R1, a writer W1, and a reader R2, about to entering the critical sections as the above order. What is the actual sequence of the three threads entering the critical sections?
 - (3) [5%] What is the special hardware instruction(s) used to implement the semaphores?
 - (4) [5%] Sometimes, some embedded platforms may not have the special hardware instruction(s) support. In this case, what kind of the software implementation can be used to ensure that the lock operations, such as wait() and signal() semaphore operations, are performed atomically?
5. [30%] File systems have software caches to accelerate accesses to the frequently used data. Please answer the following questions related to file system designs.
- (1) [10%] What kinds of caches, besides processor caches, may be used while performing file data reads from the disk into a user-space application? In what situation are the file data writes not being buffered in the above software caches?
 - (2) [10%] What is the name of the virtual-memory technique used to cache both process pages and file contents? On such a system, what will happen if processes are reading large files and storing the file data into their heap memory?
 - (3) [5%] What is the situation used to describe the file data being buffered twice in different caches?
 - (4) [5%] What is the technique that is used to avoid the file data being cached twice? Please also draw the block diagram for the file I/O using the technique.