## 第二章习题

**一. 填空题：**

1. 中央处理器由 、 和 组成，通常称为处理器。
2. 利用超大规模集成电路制成的处理器芯片称为 。
3. 指令由 和 组成，通过哈夫曼压缩的方法，对 进行代码压缩，达到优化指令字长的目的。
4. CPU最常见的两种架构是 和 。
5. 多核CPU在计算时需要通信，内部互连方式主要有 、 、和 等架构，使得处理器性能得到充分利用。
6. 图形处理器（GPU）专门面对 的计算任务，GPU的核达到上千个，在进行浮点数计算时表现出比CPU高得多的性能。
7. CUDA是面向通用计算的 的并行计算平台和编程模型，定义了“主机-设备”的编程模式。
8. 多核CPU-GPU计算平台对计算资源的调度方式主要有 和 ，以及较难控制的 。
9. 并行计算应用可分为 和 。
10. 数据表示指能由 直接辨认的数据类型。
11. OpenMP是由OpenMP Architecture Review Board牵头提出的，用于 的多处理器程序设计的一套指导性编译处理方案。

**二. 选择题：**

1. 以下不属于处理器的任务的是（ ）。

A.生成指令 B.取出指令

C.解释指令 D.执行指令

1. 时钟频率是指同步电路中时钟的基础频率，它以（ ）来度量，单位为赫兹。

A.每秒指令周期 B.每秒时钟周期

C.每秒机器周期 D.每秒状态周期

1. CPU倾向并行运算的设计越来越多，以下不属于并行实现方式的是（ ）。

A.指令级并行 B.线程级并行

C.时钟周期并行 D.数据并行

1. 多核技术是指把多个CPU内核集成到（ ）中，使得多个不同的线程可以在不同的内核中同时执行。

A.单个集成电路芯片 B.多个集成电路芯片

C.单个非集成电路芯片 D.多个非集成电路芯片

1. OpenMP的编程模型以（ ）为基础，通过编译执导语句和相关的API函数，采用Fork-Join的形式，使程序通过多线程在多核上并行执行，提高计算速度。

A.进程 B.线程

C.程序代码 D.指令系统

1. 合理定义相应的线程参数能够加速GPU的并行，以下不属于CUDA编程时对线程的管理模式的是（ ）。

A.线程 B.线程块

C.网格 D.线程周期

1. （ ）是资源分配的最小单位，（ ）是CPU调度的最小单位。

A.进程；线程 B.线程；进程

C.进程；管程 D.管程；线程

1. OpenMP提供了一组API函数用于控制并发线程的某些行为，以下（ ）是omp\_set\_num\_threads函数的作用。

A.返回线程号 B.设置后续并行域中的线程个数

C.返回当前并行域中的线程数 D.返回系统中处理器的个数

1. （ ）是GPU执行指令的基本单位。

A.流处理器 B.指令控制单元

C.流多处理器 D.寄存器和缓存

1. GPU是通过大量（ ）并行执行达到高速计算的目的。

A.进程 B.程序

C.管程 D.线程

**三. 问答题：**

1. CPU常用的架构有哪两种？分别有何特点？
2. CPU是由哪几个部件组成？各部件有何作用？：
3. CPU的并行运算实现方式有哪几种？各有什么特点？
4. GPU的存储层次由哪几个部分组成？每个部分的特点如何？：
5. 简述CUDA程序的一般执行流程：

**四. 计算题：**

1. 有一组指令的操作码共分为8类，它们出现的概率如下表所示：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 |
| 0.40 | 0.05 | 0.15 | 0.03 | 0.05 | 0.30 | 0.01 | 0.01 |

请给出此组指令的哈夫曼编码,并求出最短编码长度。

**五. 编程题：**

1.OpenMP编程练习：尝试使用OpenMP编程，用2种方法实现两个向量点积的运算。可自定义初始化两个向量，并输出最后结果。

1. CUDA编程练习：使用CUDA编程，分别利用CPU和GPU输出“hello world”。