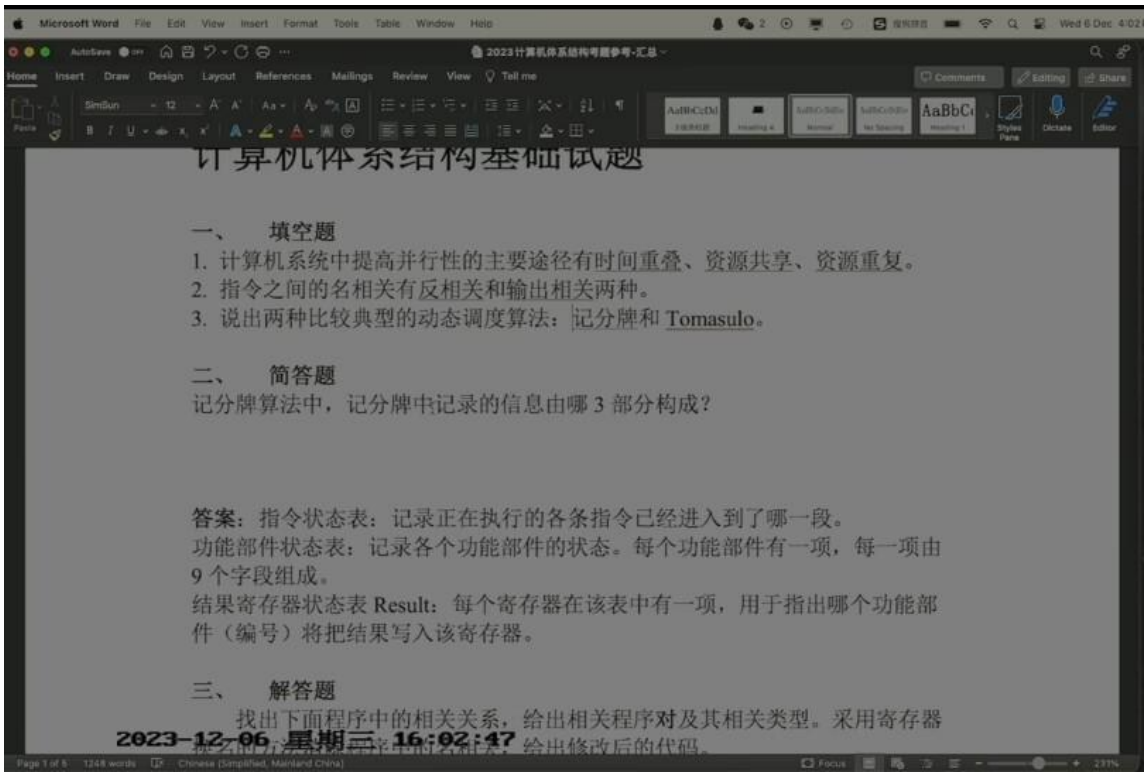
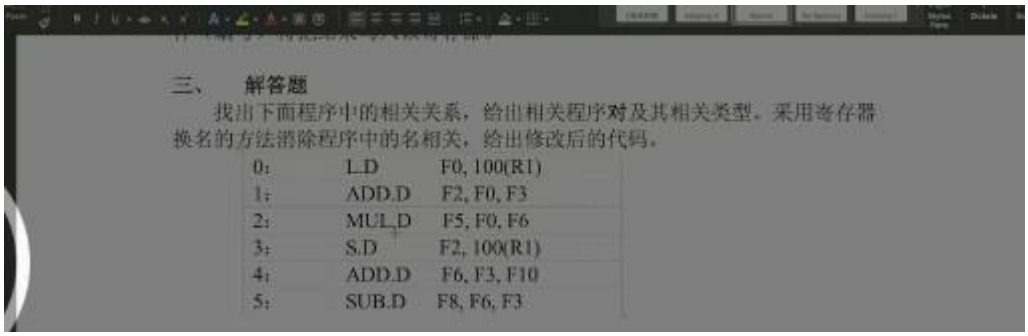
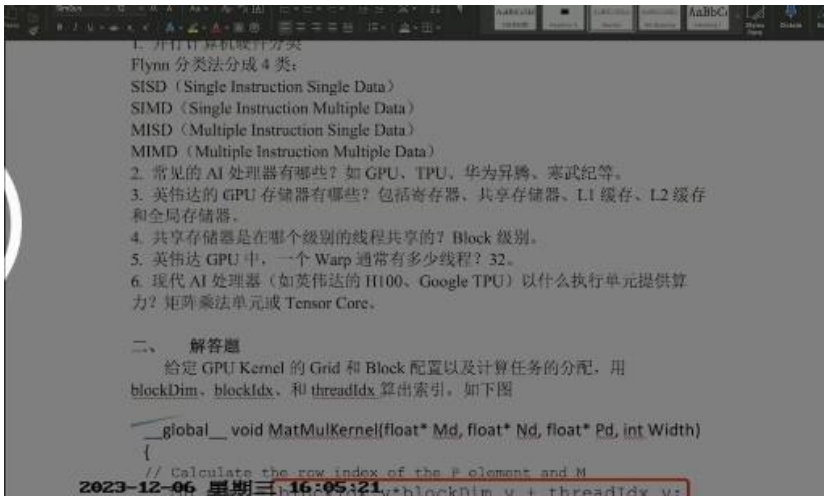
期末考试最后一节课资料

（个别老师课上有，请同学们期末复习时不同班别相互共享一下复习资料）







二、 解答题

给定 GPU Kernel 的 Grid 和 Block 配置以及计算任务的分配，用 blockDim、blockIdx、和 threadIdx 算出索引，如下图

```
__global__ void MatMulKernel(float* Md, float* Nd, float* Pd, int Width)
{
// Calculate the row index of the P element and M
  int Row = blockIdx.y*blockDim.y + threadIdx.y;
// Calculate the column index of the P element and N
  int Col = blockIdx.x*blockDim.x + threadIdx.x;

  if ( (Row < Width) && (Col < Width) ) {
    float Pvalue = 0.0;
// each thread computes one element of the block sub-matrix
    for (int k = 0; k < Width; ++k)
      Pvalue += Md[Row*Width+k] * Nd[k*Width+Col];
    Pd[Row*Width+Col] = Pvalue;
```

解答题

假设一辆自动驾驶汽车在高速公路上以 120km/h 的速度行驶。车辆装备有多个摄像头和雷达传感器。

数据处理和 AI 模型推理由一块性能为 2 TFLOPS（Tera Floating Point Operation Per Second）的 GPU 完成。

假设 AI 模型的计算量为 140 GFLOPs（Giga Floating Point Operations）。

感知模块除模型推理外，其余数据处理计算 0.03s。

规划模块的工作周期是 0.1s，控制模块的工作周期是 0.01s。

a) 在给定的硬件条件下，AI 模型完成一次完整数据处理的时间是多少？

b) 考虑到感知、规划、控制的整体工作流，各模块采用定时器触发工作与由上游消息实时触发工作两种模式最坏情况下的总反应时间分别是多少？

c) 基于以上计算，分别估算两种工作模式下，当 80m 外突然出现障碍物时，汽车能否成功制动。（120km/h 的制动距离为 70m）

答案

a) AI 模型数据处理时间：

AI 模型计算量：140 GFLOPs = 0.14 TFLOPs（因为 1 TFLOPs = 1000 GFLOPs）

GPU 性能：2 TFLOPS

计算时间 = AI 模型计算量 / GPU 性能

= 0.14TFLOPs / 2 TFLOPS

= 0.07s

b) 总的反应时间：

定时器触发（最坏情况下需要等待一个周期）：

AI 模型数据处理时间：0.07s

感知：0.03s

规划：10Hz

控制：100Hz

= 0.1s（传感器数据等待上一轮感知模型推理以及数据计算） + 0.07s（模型推理） + 0.03s（感知数据计算） + 0.1s（等待规划一个周期） + 0.1s（规划工作）

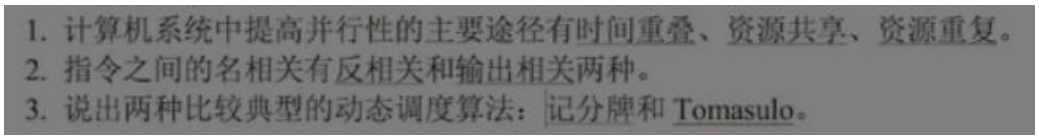
2023-12-06 星期三 16:10:30

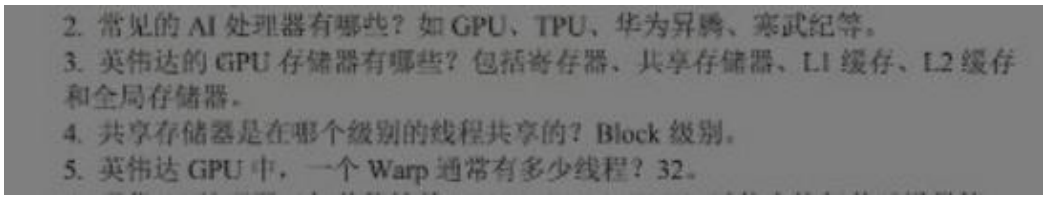
2023 级-计算机体系结构-回忆版

大致与上述内容题型相同，有改编题也有原题也有新题。

By zyj ,zjc

一：填空题(20 分）：

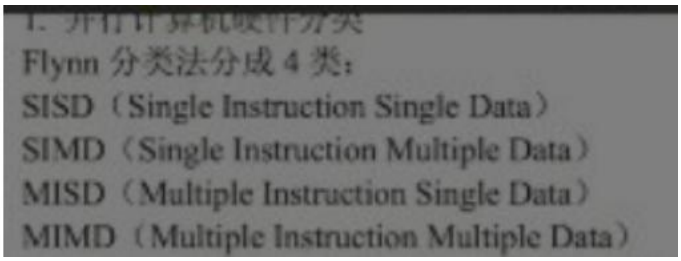

1. 计算机系统中提高并行性的主要途径有时间重叠、资源共享、资源重复。
2. 指令之间的名相关有反相关和输出相关两种。
3. 说出两种比较典型的动态调度算法：记分牌和 Tomasulo。

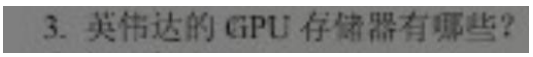

2. 常见的 AI 处理器有哪些？如 GPU、TPU、华为昇腾、寒武纪等。
3. 英伟达的 GPU 存储器有哪些？包括寄存器、共享存储器、L1 缓存、L2 缓存和全局存储器。
4. 共享存储器是在哪个级别的线程共享的？Block 级别。
5. 英伟达 GPU 中，一个 Warp 通常有多少线程？32。

最后一题：自动驾驶常用的传感器有哪些：_____, _____, _____.

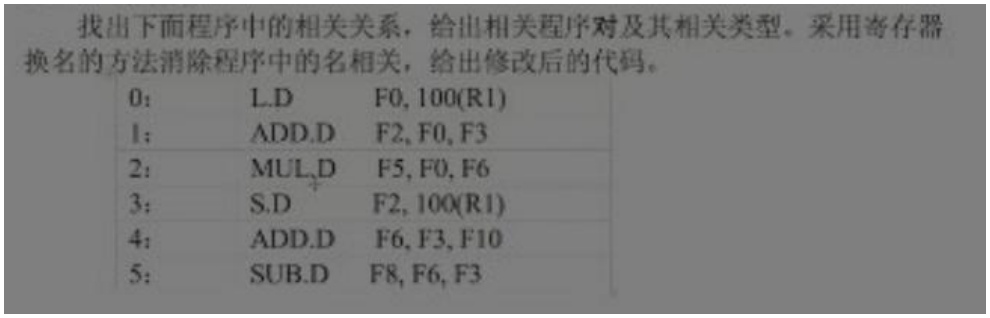
二：简答题(30 分）：

1. Flynn 并行计算机硬件分类：（5 分）


Flynn 分类法分成 4 类：
SISD（Single Instruction Single Data）
SIMD（Single Instruction Multiple Data）
MISD（Multiple Instruction Single Data）
MIMD（Multiple Instruction Multiple Data）

2.（7 分） 
3. 英伟达的 GPU 存储器有哪些？
（请描述 gpu 的存储层次）

3：指令相关和调度（8 分）


找出下面程序中的相关关系，给出相关程序对及其相关类型。采用寄存器换名的方法消除程序中的名相关，给出修改后的代码。

0:	L.D	F0, 100(R1)
1:	ADD.D	F2, F0, F3
2:	MUL.D	F5, F0, F6
3:	S.D	F2, 100(R1)
4:	ADD.D	F6, F3, F10
5:	SUB.D	F8, F6, F3

4：给定 GPU Kernel 的 Grid 和 Block 配置以及计算任务，请补充以下代码(注：题目只有这么多信息，与样题不同）(10 分）：

```
__global__ void MatMulKernel(float* Md, float* Nd, float* Pd, int Width)
{
    // Calculate the row index of the P element and M
    int Row = _____________________________________;
    // Calculate the column index of the P element and N
    int Col = _____________________________________;
    if ( (Row < Width) && (Col < Width) ) {
    float Pvalue = 0.0;
    // each thread computes one element of the block sub-matrix
    for (int k = 0; k < Width; ++k)
    Pvalue += Md[____________] * Nd[____________];
    Pd[____________] = Pvalue;
    }
}
```

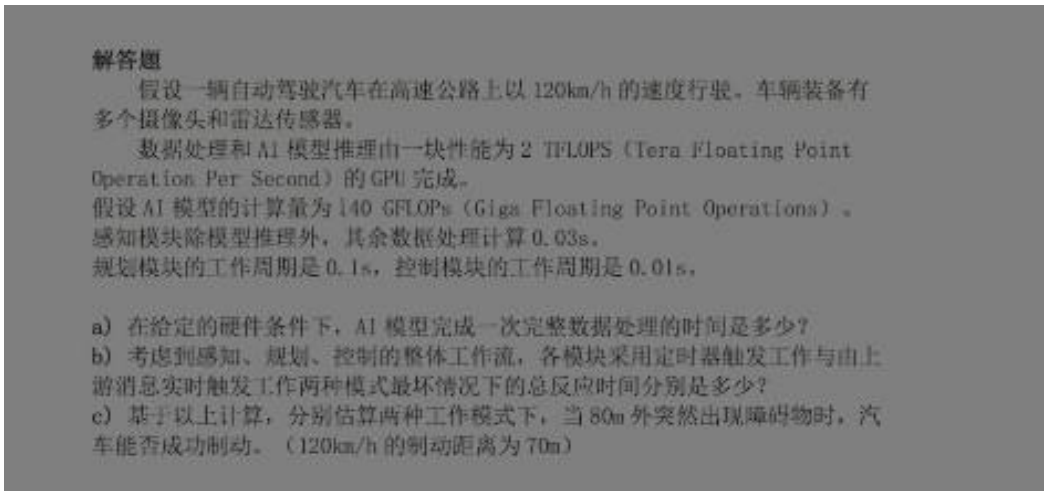
三：解答题：

1. 请描述评估分支预测有效性的指标（10 分）

2.请画图描述分支预测器在取指、译码、执行阶段做了哪些工作。(10 分）

3.自动驾驶当中用到了哪些传感器及其功能？（10 分）

4.原题（20 分）



解答题

假设一辆自动驾驶汽车在高速公路上以 120km/h 的速度行驶，车辆装备有多个摄像头和雷达传感器。

数据处理和 AI 模型推理由一块性能为 2 TFLOPS（Tera Floating Point Operation Per Second）的 GPU 完成。

假设 AI 模型的计算量为 140 GFLOPs（Giga Floating Point Operations）。

感知模块除模型推理外，其余数据处理计算 0.03s。

规划模块的工作周期是 0.1s，控制模块的工作周期是 0.01s。

a) 在给定的硬件条件下，AI 模型完成一次完整数据处理的时间是多少？

b) 考虑到感知、规划、控制的整体工作流，各模块采用定时器触发工作与由上游消息实时触发工作两种模式最坏情况下的总反应时间分别是多少？

c) 基于以上计算，分别估算两种工作模式下，当 80m 外突然出现障碍物时，汽车能否成功制动。（120km/h 的制动距离为 70m）