北京大学 - 智能硬件体系结构

2024年秋季 期末考试

1、数据依赖与寄存器重命名（25分）

```
1. R1 = R2 / R3
2. R0 = MEM[R1]
3. R1 = R2 + R3
4. R0 = R0 + R1
5. R1 = MEM[R1]
6. R1 = R0 * R1
7. R4 = R0 + R1
8. MEM[R3] = R2 // a store instruction
```

1）假设指令并行度仅受到真实依赖（true dependencies）的限制，那么这段代码中可用的指令级并行（instruction-level parallelism, ILP）是多少，（这里指令集并行度用假设我们有一个无限宽度OoO的处理器，并能够保证每个指令在单周期完成，执行这段程序得到的instr/cycle代表）？请画出依赖关系图来证明你的答案（使用指令编号表示指令）。

2）将上述汇编代码进行寄存器重命名，以消除代码段中的所有伪依赖（false dependencies），同时保持相同的功能。要求如下：第一、不能重新排序、删除或添加指令（只能更改寄存器编号）；第二、确保代码段结束时，寄存器 R0-R4 中的值与上述代码段执行后完全相同。例如，如果上述代码段执行后 R0=5，那么在你重写的代码段结束后，R0 仍然必须是 5。任何内存MEM的写操作结果也必须保持一致；第三、可使用的寄存器为 R0-R8。

2、缓存与一致性（25分）

1）在课堂上讲解的 MESI 协议中，以下哪些状态转换可能是由于另一个处理器的事务引起的?

M -> E　　S -> E　　I -> S　　M -> S　　S -> M

2) 考虑以下C代码：

```
char A[4096]; // each element is 1 byte
for(j=0;j<100000;j++){
    for(i=0;i<Y;i=i+X){
        A[i]=A[i]+1;
    }
```

}

假设只有对数组 A 的访问会进入数据缓存（其他值存储在寄存器中）。对于这段代码，如果数据缓存大小为 1 KB，缓存行（cache line）大小为 32 字节（因此存储一行是存数组A中连续的32个元素），并且是直接映射的，那么对于不同的 X 和 Y 值，预期的命中率是多少？完成以下表格并简要说明计算过程。

	X=2	X=4	X=64
Y=2048			
Y=1025			

3、乱序执行微架构（25分）

假设我们有一个乱序执行处理器，RS可容纳3条指令，ROB可容纳6条指令。由于第一个Load指令需要很长时间停顿，以下两个程序A和B都会因为等待Load完成而停滞。对于每个程序，请指出加载完成之前，最后一条将被放入 ROB 的指令，同时给出简要解释。（指令会一直停留在RS，直到它被发射到计算单元。）

Program A	Program B
R1=MEM[R2+0]	R1=MEM[R2+0]
R2=R4+4	R2=R1+4
R3=R2+R7	R3=R4+R4
R4=R1+6	R4=R2+6
R1=R2+R3	R1=R2+R3
R6=R2+R6	R6=R9+R10
R7=R6+19	R7=R3+19
R8=R3+R6	R8=R9+R3

4、AI处理器架构（25分）

1）在神经网络加速器设计中，通常会用到多级的乘加树结构来计算卷积或矩阵乘法。我们利用课堂中所学的Radix-4布斯编码乘法器，构建了一个可用于神经网络加速器的乘加单元。该乘加单元输入为4个8-bit的有符号补码数（2’s complement）X1、X2、X3、X4，输出Y = X1*X2 + X3*X4。假设X1 = -2，X2 = 8，X3 = -15，X4 = -9。参考完成第三讲第40页，完成X1*X2和X3*X4的布斯编码乘法器步骤。

2）神经网络加速器架构一般可简单分为Output Stationary、Weight Stationary、Input Stationary，请简要说明这3种架构设计方式的区别。

3）残差是当前神经网络中极为常用的机制，广泛应用于各类AI模型中。假设一个神经网络加速器采

用Weight Stationary进行设计，如何应对残差等需要跳跃多个神经网络层级的数据通信？请给出1-2个思路，并简单描述该思路的可行性和优劣（200-300字以内，配合1-2个简单图示说明即可）。