1. 假定机器M的时钟频率为1.2GHz，某程序P在机器M上的执行时间为12秒钟。对P优化时，将其所有的乘4指令都换成了一条左移2位的指令，得到优化后的程序P’。已知在M上乘法指令的CPI为5，左移指令的CPI为2，P的执行时间是P’执行时间的1.2倍，则P中有多少条乘法指令被替换成了左移指令被执行？（20分）

显然，P’的执行时间为10秒，因此，P比P’多花了2秒钟，因此，执行时被换成左移指令的乘法指令的条数为1.2G×2/(5–2) = 800M

2、若有两个基准测试程序P1和P2在机器M1和M2上运行，假定M1和M2的价格分别是5000元和8000元，下表给出了P1和P2在M1和M2上所花的时间和指令条数。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 程序 | M1 | | M2 | |
| 指令条数 | 执行时间(ms) | 指令条数 | 执行时间(ms) |
| P1 | 200×106 | 10000 | 150×106 | 5000 |
| P2 | 300×103 | 3 | 420×103 | 6 |

请回答下列问题：（30分）

1. 对于P1，哪台机器的速度快？快多少？对于P2呢？（8分）
2. 在M1上执行P1和P2的速度分别是多少MIPS？在M2上的执行速度又各是多少？从执行速度来看，对于P2，哪台机器的速度快？快多少？（8分）
3. 假定M1和M2的时钟频率各是800MHz和1.2GHz，则在M1和M2上执行P1时的平均时钟周期数CPI各是多少？ （8分）
4. 如果某个用户需要大量使用程序P1，并且该用户主要关心系统的响应时间而不是吞吐率，那么，该用户需要大批购进机器时，应该选择M1还是M2？为什么？（8分）（提示：从性价比上考虑）
5. 如果另一个用户也需要购进大批机器，但该用户使用P1和P2一样多，主要关心的也是响应时间，那么，应该选择M1还是M2？为什么？（8分）
6. 对于P1，M2比M1快一倍；对于P2，M1比M2快一倍。
7. 对于M1，P1的速度为：；

P2的速度为：。

对于M2，P1的速度为：

P2的速度为：。

从执行速度来看，对于P2，因为100/70=1.43倍，所以M1比M2快0.43倍。

1. 在M1上执行P1时的平均时钟周期数CPI为：10×800M/(200×106)=40。

在M2上执行P1时的平均时钟周期数CPI为：5×1.2G/(150×106)=40。

1. 考虑运行P1时M1和M2的性价比，因为该用户主要关心系统的响应时间，所以性价比中的性能应考虑执行时间，其性能为执行时间的倒数。故性价比R为：

R=1/(执行时间×价格)

R越大说明性价比越高，也即，“执行时间×价格”的值越小，则性价比越高。

因为10×5000 > 5×8000，所以，M2的性价比高。应选择M2。

1. P1和P2需要同等考虑，性能有多种方式：执行时间总和、算术平均、几何平均。

若用算术平均方式，则：因为 (10+0.003)/2×5000 > (5+0.006)/2×8000，所以M2的性价比高，应选择M2。

若用几何平均方式，则：因为sqrt(10×0.003) ×5000 < sqrt(5×0.006) ×8000，所以M1的性价比高，应选择M1

1. 给定浮点数对称矩阵，
2. 试计算的浮点数操作量。（15分）（提示：考虑C的对称性）
3. 设,矩阵乘法程序在H100 GPU上计算双精度矩阵乘法花费的时间为0.1秒，考虑H100的双精度峰值性能为67 TFLOPs（ FLOPs），该程序的性能达到了硬件峰值性能的百分之多少？（15分）
4. 考虑普通矩阵乘法的FLOPS为，由于C为对称矩阵，因此只需计算矩阵的上三角或下三角部分，因此FLOPS为（可忽略对角线元素等级的计算量）。
5. FLOPS为，得到TFLOPs为,因此占14.9%的峰值性能。

4、（1）简述CPU与GPU架构的区别。（5分）

略

（2）简述GPU上SIMT与Flynn’s分类中SIMD的区别。（5分）

略

5、Roofline模型是分析程序性能一个重要工具，其主要的衡量指标为算术强度（arithmetic intensity），定义为，其中为浮点数计算量FLOPS，为需要搬运的数据量即字节数Bytes。

（1）A100 SXM4的内存带宽为1.56 TB/s，峰值双精度性能为19.5 TFLOPs；H100 SMX5的内存带宽为3.36 TB/s，峰值双精度性能为67 TFLOPs；RTX 4090的内存带宽为1.01TFLOPs，峰值双精度性能为1.29 TFLOPs。试画出A100，H100和RTX 4090的roofline模型。（9分）

类似于上课画的图，拐点的横坐标为峰值性能/内存带宽。

（2）上述哪种GPU更容易在计算任务中达到双精度峰值性能？原因是什么？（6分）

RTX 4090，根据（1）中的图形，4090需要更小的算数强度就可以达到峰值性能。

（3）考虑RTX 4090的单精度峰值性能为82.5 TFLOPs。在计算第三题中的矩阵乘法时，讨论与时，单精度与双精度矩阵乘法的性能差距，并分析原因。（10分）（提示：定量分析，但无需精确计算）

当n=10时，无论是单精度还是双精度的算术强度在roofline model中都没有达到拐点，因此性能上限在内存带宽。但是由于单精度字节数是双精度的1/2，因此总体性能单精度约为双精度的2倍。当n=10000时，单精度和双精度都已经达到拐点，因此性能上限在于峰值性能，所以单精度的性能将远超双精度的性能。