

计算机系统结构习题

智能 1602 201607020301 孟祥炜

练习 1.8

架构师面对的一个挑战是，今天拟定的设计方案可能需要几年的时间进行实施、验证和测试，然后才能上市。这就意味着架构师必须提前为几年之后的技术进步制定计划。有时，这是很难做到的

a.根据摩尔定律观测到的器件发展趋势，到2015年，一个芯片上的晶体管数目应当是2005年的多少倍？

b.时钟频率的增加也一度反映了这一趋势。如果时钟频率仍以20世纪90年代的相同速度攀升，2015年的时钟速率大约是多少

c.以目前的增长速率，2015年的时钟频率是多少？

d.是什么限制了时钟频率的增长速度？为了提升性能，架构师现在能用多出来的晶体管做些什么？

e.DRAM容量的增长速度也已变缓。20年来，DRAM容量每年提高60%这一速率下降到每年40%，现在的改进速率为每年25%~40%。如果这一趋势继续下去，2020年的DRAM容量增速大约是多少

答：

(a) $1.35^{10} = 20.1$ 倍

(b) $3200 * 1.4^{12} = 181420.5$ 倍

(c) $3200 * 1.01^{12} = 3605.8$ 倍

(d) 功率密度，即在越来越小的区域内消耗的功率，已经产生了太多的热量以使散热器难以完成散热。这限制了芯片上晶体管的活动。

现在的制造商不再把精力放在提高时钟频率上，而是在芯片上放置多个内核。

(e) 15%~25%

练习2.19

每当计算机空闲时，既可以将其置于待机状态（DRAM仍然处于活动状态），也可以让它休眠。为了使其进入休眠状态，假定必须仅将DRAM的内容复制到永久性介质中，比如闪存中。如果将大小为64字节的级存行读写至闪存需要2.56μJ，读写至DRAM需要0.5nJ，如果8GB DRAM的空闲功耗为1.6W，那么一个系统空闲多长时间后才能从休眠中获益？假定主存储器的容量为8GB.

答:

当 DRAM 中保存的静态能量至少等于从 DRAM 复制到闪存再返回 DRAM 所需的能量时, Hibernating 将非常有用。与 Flash 相比, DRAM 的动态读取/写入能量可以忽略不计。

$$\text{所以 } T = \frac{8 \times 10^9 \times 2 \times 2.56 \times 10^{-6}}{64 \times 1.6} = 400s$$

即空闲 400s 以上才能使得休眠的收益比待机的收益大

练习 4.13

假定有一种包含10个SIMD处理器的GPU体系结构。每条SIMD指令的宽度为32, 每个SIMD处理器包含8个车道, 用于执行单精度运算和载入/存储指令, 也就是说, 每个非分支SIMD指令每4个时钟周期可以生成32个结果。假定内核的分支将导致平均80%的线程为活动的。假定在所执行的全部SIMD指令中, 70%为单精度运算、20%为载入/存储。由于并不包含所有存储器延迟, 所以假定SIMD指令平均发射率为0.85。假定GPU的时钟速度为1.5 GHz

a. 计算这个内核在这个GPU上的吞吐量, 单位为 GFLOPs/s。

b. 假定我们有以下选项:

(1) 将单精度车道数增大至16

(2) 将SIMD处理器数增大至15 (假定这一改变不会影响所有其他性能度量, 代码会扩展到增加的处理器上)。

(3) 添加缓存可以有效地将存储器延迟缩减40%, 这样会将指令发射率增加至0.95, 对于这些改进中的每一项。

吞吐量的加速比为多少?

答:

(a) $1.5 \text{ GHz} \times 0.8 \times 0.85 \times 0.7 \times 10 \times 32 / 4 = 57.12 \text{ GFLOPs/s}$

(b) (1) $1.5 \text{ GHz} \times 0.8 \times 0.85 \times 0.7 \times 10 \times 32 / 2 = 114.24 \text{ GFLOPs/s}$

加速比 = $114.24 / 57.12 = 2$

(2) $1.5 \text{ GHz} \times 0.8 \times 0.85 \times 0.7 \times 15 \times 32 / 4 = 85.68 \text{ GFLOPs/s}$

加速比 = $85.68 / 57.12 = 1.5$

(3) $1.5 \text{ GHz} \times 0.8 \times 0.95 \times 0.7 \times 10 \times 32 / 4 = 63.84 \text{ GFLOPs/s}$

加速比 = $63.84 / 57.12 = 1.11$

注: 频率 × 80% 的线程为活动的 × 发射率 × 只有运算可以使用 GPU 计算, 70% × 核心数 × 单个 SIMD 处理器的车道数

练习 4.15

列出并介绍至少 4 种可以影响 GPU 内核性能的因素。换句话说，哪些由内核代码导致的运行时行为会降低内核执行时的资源利用率？

答：

- a.分支情况：当线程遵循不同的控制路径时，会导致 SIMD 通道被屏蔽
- b.覆盖内存延迟：足够数量的活动线程可以隐藏内存延迟并提高指令发布率
- c.合并的片外存储器引用：应在 SIMD 线程组内连续组织存储器访问
- d.使用片上存储器：具有局部性的存储器参考应该利用片上存储器，应该组织对 SIMD 线程内部的片上存储器的引用以避免存储体冲突。

练习 5.35

在增加处理器的数目时，多处理器和集群的性能通常也会提高，理想情况下应当是 n 个处理器提高 n 倍。这一有偏基准测试的目标是让程序在增加处理器时的性能恶化。例如，这就意味着当多处理器或集群仅有一个处理器时，程序的运行速度最快，有2个处理器时，速度较慢，4个处理器时要比2个处理器还慢，以此类推。在每种组织结构中，是哪些关键的性能特性导致逆线性加速比的？

答：

可能的原因有

1. 在多处理器情况下，cache 的未命中和访问之间的连续无效可能比实际执行计算的时间贡献更多，导致处理器增加时总体时间逆线性增加
2. 对于多处理器或集群系统，更新公共变量所需的成本远大于计算带来的收益，从而导致逆线性加速比的出现