

# Homework 6

PB17000297 罗晏宸

May 18 2020

## 1 Exercise 4.9

假定处理器运行频率为 700 MHz，最大向量长度为 64，载入/存储单元的启动开销为 15 个时钟周期，乘法单元为 8 个时钟周期，加法/减法单元为 5 个时钟周期。在该处理器上进行如下运算，将两个包含单精度复数值的向量相乘：

```
for (i = 0; i < 300; i++) {  
    c_re[i] = a_re[i] * b_re[i] - a_im[i] * b_im[i];  
    c_im[i] = a_re[i] * b_im[i] + a_im[i] * b_re[i];  
}
```

(1) 这个内核的运算密度为多少（注：运算密度指运行程序时执行的浮点运算数除以主存储器中访问的字节数）？

(2) 将此循环转换为使用条带挖掘（Strip Mining）的 VMIPS 汇编代码。

(3) 假定采用链接和单一存储器流水线，需要多少次钟鸣？每个复数结果值需要多少个时钟周期（包括启动开销在内）？

(4) 如果向量序列被链接在一起，每个复数结果值需要多少个时钟周期（包含开销）？

(5) 现在假定处理器有三条存储器流水线和链接。如果该循环的访问过程中没有组冲突，每个结果需要多少个时钟周期？

解

(1) 执行 6 次浮点运算,读 4 个浮点数,写 2 个浮点数,访问  $(4+2) \times 4 = 24$  个字节。内核运算密度为

$$\frac{6}{(4+2) \times 4} = \frac{1}{4}$$

(2)

```

li          $VL, 44          ; 前 44 步操作
li          $r1, 0           ; 初始化下标
loop:  lv    $v1, a_re + $r1 ; load a_re
        lv    $v3, b_re + $r1 ; load b_re
        mulvv.s $v5, $v1, $v3 ; a_re * b_re
        lv    $v2, a_im + $r1 ; load a_im
        lv    $v4, b_im + $r1 ; load b_im
        mulvv.s $v6, $v2, $v4 ; a_im * b_im
        subvv.s $v5, $v5, $v6 ; a_re * b_re
                                - a_im * b_im
        sv     $v5, c_re + $r1 ; store c_re
        mulvv.s $v5, $v1, $v4 ; a_re * b_im
        mulvv.s $v6, $v2, $v3 ; a_im * b_re
        addvv.s $v5, $v5, $v6 ; a_re * b_im
                                + a_im * b_re
        sv     $v5, c_im + $r1 ; store c_im
        bne    $r1, 0, else    ; 是否首次循环
        addi   $r1, $r1, #176 ; 首次循环
        j loop                                ; 跳转下次循环
else:  addi   $r1, $r1, #256 ; 非首次循环
skip:  blt    $r1, 1200, loop ; 跳转下次循环

```

(3)

```

1  mulvv.s    lv
2  lv         mulvv.s

```

```

3  subvv.s      sv
4  mulvv.s      lv      ; load 下一个向量
5  mulvv.s      lv      ; load 下一个向量
6  addvv.s      sv

```

共需要 6 次钟鸣，每个复数结果值需要的时钟周期为

$$\frac{6 \times 64 + 15 \times 6 + 8 \times 4 + 5 \times 2}{2 \times 64} = \frac{516}{128} = \frac{129}{32} = 4.03125$$

(5)

```

1  mulvv.s
2  mulvv.s
3  subvv.s      sv
4  mulvv.s
5  mulvv.s      lv
6  addvv.s      sv  lv  lv  lv ; load
                        下一个向量

```

共需要 6 次钟鸣，因此每个复数结果值需要的时钟周期为 4.03125 与上相同。

## 2 Exercise 4.16

假定一个虚设 GPU 具有以下特性：

- 时钟频率为 1.5 GHz
- 包含 16 个 SIMD 处理器，每个处理器包含 16 个单精度浮点单元
- 片外存储器带宽为 100 GB/s

(1) 不考虑存储器带宽，假定所有存储器延迟可以隐藏，则这一 GPU 的峰值单精度浮点吞吐量为多少 GFLOP/s？

(2) 在给定存储器带宽限制下，这一吞吐量是否可持续？

解

(1)

$$1.5 \text{ GHz} \times 16 \times 16 = 384 \text{ GFLOP/s}$$

(2) 每个单精度运算需要读 2 个操作数，写 1 个操作数，访问  $(2 + 1) \times 4 = 12$  个字节，需要  $12 \text{ B} \times 384 \text{ GFLOP/s} = 4608 \text{ GB/s}$ ，因此吞吐量不可持续。

### 3 Exercise 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，每个非分岔 SIMD 指令每 4 个时钟周期可以生成 32 个结果。假定内核的分岔分支将导致平均 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。

(1) 计算这个内核在这个 GPU 上的吞吐量，单位为 GFLOP/s。

(2) 对于以下改进中的每一项，吞吐量的加速比为多少？

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。

解

(1)

$$1.5 \text{ GHz} \times 80\% \times 85\% \times 70\% \times 10 \times 8 = 57.12 \text{ GFLOP/s}$$

(2)

① 加速比为

$$\begin{aligned}& \frac{1.5 \text{ GHz} \times 80\% \times 85\% \times 70\% \times 10 \times 16}{57.12 \text{ GFLOP/s}} \\&= \frac{114.24 \text{ GFLOP/s}}{57.12 \text{ GFLOP/s}} \\&= 2\end{aligned}$$

② 加速比为

$$\begin{aligned}& \frac{1.5 \text{ GHz} \times 80\% \times 85\% \times 70\% \times 15 \times 8}{57.12 \text{ GFLOP/s}} \\&= \frac{85.68 \text{ GFLOP/s}}{57.12 \text{ GFLOP/s}} \\&= 1.5\end{aligned}$$

③ 加速比为

$$\begin{aligned}& \frac{1.5 \text{ GHz} \times 80\% \times 95\% \times 70\% \times 10 \times 8}{57.12 \text{ GFLOP/s}} \\&= \frac{63.84 \text{ GFLOP/s}}{57.12 \text{ GFLOP/s}} \\&= \frac{19}{17} \\&= 1.118\end{aligned}$$