作业11 GPU

GPU的利用率通常定义为处于busy状态的GPU 核心(PE，单个计算单元)占所有GPU 核心的比例。考虑下面的代码片段。每个thread执行循环中的1次迭代（包含6条指令）。假设数组A、B、C已经在寄存器中（不需要从内存读入）。该GPU的一个warp包含64个threads，该GPU包含64个核心。假设数组B的每个元素的绝对值都小于10。

```
1 for (i = 0; i < 1024; i++) {
2     A[i] = B[i] * B[i];
      if (A[i] > 0) {
3         C[i] = A[i] * B[i];
4         if (C[i] < 0) {
5             A[i] = A[i] + 1;
          }
6         A[i] = A[i] - 2;
      }
  }
```

1. 执行该代码段需要多少个warps?

```
1024/64 = 16
```

2. 执行整个代码段，最大的GPU利用率可能是多少？

100%

3. 获得最大的GPU利用率时，数组B的值有何特征？

对于每64个连续值，或者都是0，或者都是正数，或者都是负数。

4. 执行整个代码段，最小的GPU利用率可能是多少？

指令3.4.5.6（只有一个线程执行）

(64 × 2 + 1 × 4)/(64 × 6) = 132/384

指令1，2（因为这是所有线程都会执行的）

5. 获得最小的GPU利用率时，数组B的值有何特征？

仅有1个thread通过了第1个分支语句，其他threads都没通过，因此，只有一个PE busy，其余都是空闲的。也就是，每连续64个值，有1个是负数，其余都是0。