Multicore Computing

Assignment Four (Theory)

Department of Computer Engineering

Sharif University of Technology

Spring 2022

Lecturer:
Dr. Falahati

Name - Student Number:
Amirmahdi Namjoo - 97107212

1 Question One

a)

$$\text{warps} = \frac{\text{Total Thread}}{\text{Warp Size}} = \frac{1024}{32} = 32$$

We have 1024 threads because we will have one thread per outer loop iteration (based on the question description).

b) All 1024 threads execute instructions 1,2, and 4.

In the first iteration of `j`, the value of `s` is 1, so threads with `i` divisible by two will execute instruction 3. i.e., half of the threads of each warp will execute this instruction. Therefore

$$\text{SIMD Utilization} = \frac{1024 + 1024 + \frac{1024}{2} + 1024}{1024 + 1024 + 1024 + 1024} = \frac{7}{8}$$

c) All 1024 threads execute instructions 1,2 and 4.

For instruction 3, only threads with $i < 512$ will run this. The difference between this one and the previous one is that in the former, half the threads of each warp were inactive. But in this case, for half of the warps, all threads are inactive, no instructions are issued, and we have no performance loss. i.e., All threads of only half of the warps execute instruction 3.

$$\text{SIMD Utilization} = \frac{1024 + 1024 + \frac{1024}{2} + 1024}{1024 + 1024 + \frac{1024}{2} + 1024} = 1$$

d) All 1024 threads execute instructions 1,2, and 4.

For instruction 3, with $0 \leq j < 5$, all 32 warps are active, and the number of active threads per warp divides by half in each iteration. The denominator, in this case, is always 4096 with $5 \leq j < 10$, only one thread per warp is active, and some of the warps will have no active thread, and therefore scheduler will not issue any instruction for them. Therefore the denominator for these will also change.

$$\text{SIMD Utilization} = \begin{cases} \frac{3072+2^{(9-j)}}{4096}, & \text{if } 0 \leq j < 5 \\ \frac{3072+2^{(9-j)}}{3072+32\times 2^{(9-j)}}, & \text{if } 5 \leq j < 10 \end{cases}$$

e) All 1024 threads execute instructions 1,2 and 4.

For instruction 3, with $0 \leq j < 5$, all 32 threads in each warp are active, but not all of the warps are active. The number of warps active will halve each iteration. With $5 \leq j < 10$, only one warp (the one containing $i = 0$ to $i = 31$) will be active, and the number of active threads will half in each iteration. Therefore, for $0 \leq j < 5$, the value of nominator and denominator is equal, and we have 100% utilization. For $5 \leq j < 10$, the nominator changes. Therefore:

$$\text{SIMD Utilization} = \begin{cases} 1, & \text{if } 0 \leq j < 5 \\ \frac{3072+2^{(9-j)}}{3072+32}, & \text{if } 5 \leq j < 10 \end{cases}$$

f) This could not happen for $0 \leq j < 5$ because the second one has 100% utilization, while the first one has utilization of less than 1. For $5 \leq j < 10$ we can solve the equation

$$\frac{3072 + 2^{(9-j)}}{3072 + 32} = \frac{3072 + 2^{(9-j)}}{3072 + 32 \times 2^{(9-j)}}$$

$$\Rightarrow j = 9$$

Which is reasonable. In this case, only one thread of only one warp is active.

g) Code 2 will be faster; It has less intra-warp branch divergence. In other words, in most cases of Code 2, a warp is either completely active or inactive. Therefore we have high SIMD utilization, and the scheduler will not schedule completely inactive warps. In contrast, in code 1, in lots of cases, all warps are active while some threads inside warps are not active, and therefore scheduler schedules them with less than optimal utilization.