Assignment IV

Adam Orucu

January 7, 2024

Url to code: `https://github.com/adamorucu/gpu-programming-course`

Exercise 1

1. **Assume X=800 and Y=600. Assume that we decided to use a grid of 16X16 blocks. That is, each block is organized as a 2D 16X16 array of threads. How many warps will be generated during the execution of the kernel? How many warps will have control divergence? Please explain your answers.**

   There will be $ceil(800/16) * ceil(600/16) = 1900$ blocks. Since each block has $16 * 16 = 256$ threads there will be $256 * 1900 = 486,400$ threads in total. Given that each warp has 32 threads the number of warps is $486,400/32 = 15,200$.

   600 is not a multiple of 16, therefore both bottom edge will have divergence. $ciel(800/16) = 50$ blocks will be in divergence. This is $50 * 256/32 = 400$ warps.

2. **Now assume X=600 and Y=800 instead, how many warps will have control divergence? Please explain your answers.**

   Repeating the same calculations, 600 is not a multiple of 16, therefore both right edge will have divergence. $ciel(800/16) = 50$ blocks will be in divergence. This is $50 * 256/32 = 400$ warps.

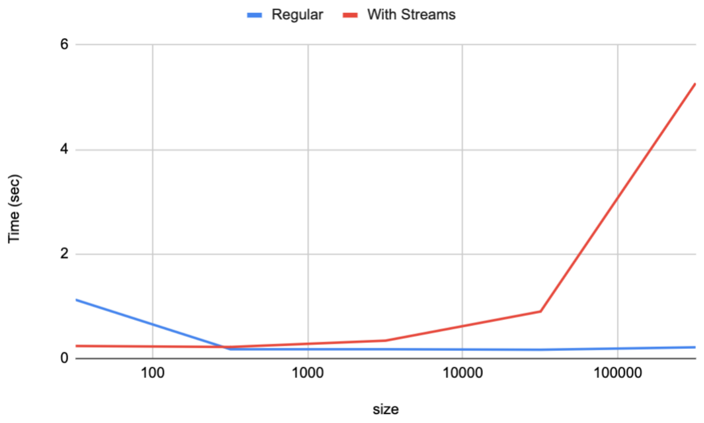
3. **Now assume X=600 and Y=799, how many warps will have control divergence? Please explain your answers.**

   Repeating the same calculations, 600 is not a multiple of 16, therefore both right edge will have divergence. $ciel(799/16) = 50$ blocks will be in divergence. This is $50 * 256/32 = 400$ warps.

Exercise 2

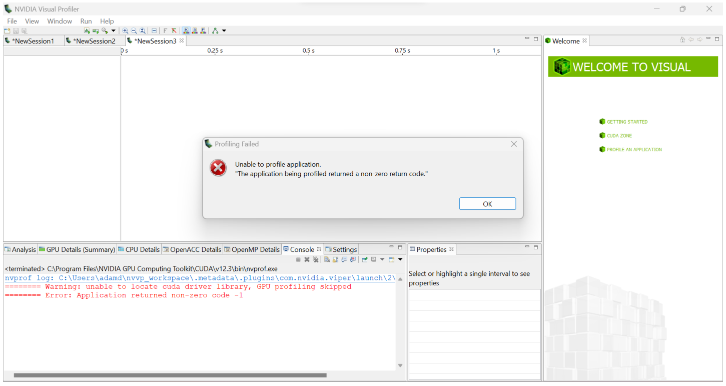
1. **Compared to the non-streamed vector addition, what performance gain do you get? Present in a plot (you may include comparison at different vector length)**

   There wasn't a performance gain in fact streamed version of the code ended up being slightly slower.



2. **Use nvprof to collect traces and the NVIDIA Visual Profiler (nvvp) to visualize the overlap of communication and computation.**

   I wasn't able to make nvvp work on two different operating systems.

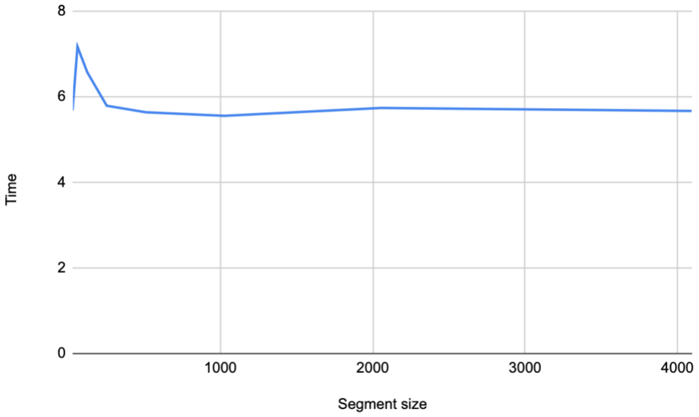


Unable to profile application.
"The application being profiled returned a non-zero return code."

```
<terminated> C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v12.3\bin\nvprof.exe
nvprof log: C:\Users\adamd\nvvp_workspace\.metadata\.plugins\com.nvidia.viper\launch\2\
======== Warning: unable to locate cuda driver library, GPU profiling skipped
======== Error: Application returned non-zero code -1
```

Select or highlight a single interval to see properties

3. **What is the impact of segment size on performance? Present in a plot (you may choose a large vector and compare 4-8 different segment sizes)**



Exercise 3

I wasn't able to generate the tests because I exceded the compute limit on Google Colab.

1. **Run the program with different dimX values. For each one, approximate the FLOPS (floating-point operation per second) achieved in computing the SMPV (sparse matrix**

multiplication). Report FLOPS at different input sizes in a FLOPS. What do you see compared to the peak throughput you report in Lab2?

2. **Run the program with dimX=128 and vary nsteps from 100 to 10000. Plot the relative error of the approximation at different nstep. What do you observe?**

3. **Compare the performance with and without the prefetching in Unified Memory. How is the performance impact?**