**Coding Project #4: GPUs, CUDA, Vector Addition Performance Measurement**

1. What is the MFLOP/s performance gain going from the CPU-only code to the final version of your CUDA code (the one with the cudaMemPrefetchAsync() call)? Show your work on how you compute this result.
2. What is the memory bandwidth performance gain (or loss) going from the CPU-only code to the final version of your CUDA code (the one with the cudaMemPrefetchAsync() call)? Show your work on how you compute this result.
3. For the final version of your CUDA code (the one with the cudaMemPrefetchAsync() call), what is the total number of concurrent threads being run? Show your work on how you arrive at this result.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Elapsed vector addition kernel execution time in milliseconds** | **MFLOP/s** | **memory bandwidth utilized in GB/s** |
| **vecadd\_cpu** |  |  |  |
| **vecadd\_gpu\_1t.cu** |  |  |  |
| **vecadd\_gpu\_256t.cu** |  |  |  |
| **vecadd\_gpu\_256t\_manyblocks.cu** |  |  |  |
| **vecadd\_gpu\_256t\_manyblocks\_prefetch.cu** |  |  |  |

A screenshot of a computer program

Description automatically generated with low confidence

A screenshot of a computer screen

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated