Omar Dajani
December 7th, 2023

A. **Table**:

Elapsed Vector Addition Kernel Execution Time (in milliseconds)	MFLOP/s	Memory Banwidth Utilized in GB/s
920.028	.83747	.1304308
6,317,543,489	575.066584	.00000002
237,315,720	216.021213	.00000005
236,136,336	214.947656	.00000005
231,287,611	210.53401	.00000038

B. **Analysis Questions**:

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.