

	Runtime (ms)	MFLOP/s	Bandwidth GB/s		
CPU Only	849.123	1.658443359	7587.182239	Problem Size (N)	
GPU: 1t, 1b	50400	98.4375	127.8264076	536870912	
GPU: 256t, 1b	2036	3.9765625	3164.268637		
GPU: 256t, Nb	1372	2.6796875	4695.663953		
GPU: 256t, Nb, prefetch	4.77	0.00931640625	1350618.647		

Number of Arithmetic Operations = N

Number of Memory Operations = 3N

Analysis Questions:

- MFLOP/s gain going from serial CPU to many-threaded GPU code?
  - Want % gain, not absolute MFLOP/s

$$(3.9765625 - 1.658443359 / 1.658443359) * 100\% = 297.65625\%$$

- Memory bandwidth utilization gain going from serial CPU to many-threaded GPU code?
  - Want % gain, not absolute GB/s

$$(3164.268637 - 7587.182239) / 7587.182239 = -58.29454813\%$$

- For your many-threaded GPU code with memory prefetch, how many concurrent threads are there in your program?

Using the formula:

$$\text{numBlocks} = (N + \text{THREADSPerBLOCK} - 1) / \text{THREADSPerBLOCK};$$

$$\text{We can find the number of blocks being} = (536870912 + 256 - 1) / 256 = 2097153 \text{ Blocks}$$

$$\text{Each block can run up to 256 threads} = 2097153 * 256 = 536871168 \text{ Threads}$$

Can be run concurrently

