	Original	SM	K2	I2	J2	ALL2	SM-All2
8X8	80	85	80	128.4	85	131.15	102.5
16X16	51.8	89	51.5	90.3	55	103	105.3
32X32	27.75	90	27.8	53.25	28.5	55.7	103.25

For this assignment I used an Nvidia GTX 970 in the cade lab, the original version was my ji version from the previous assignment. All numbers are measured in GFLOPS. One interesting trend I noticed was that the smaller block sizes worked better then the larger one. I found this to be odd given that my test with the pervious version was the opposite. I believe that this is because I was running on a different GPU architecture that favors the smaller block size. This might be because there have been improvements to the packing of smaller blocks into a single wrap. Another trend I noticed was that my shared memory all 2-version performance stayed relatively consistent across the block sizes. I believe that this is because the shared memory reduced the cost associated with the different block sizes. This is supported by the simple sharded memory version also having relatively the same performance. On final trend I noticed was that my All2 performed better than the SM-All2 on small block sizes but worst on larger block sizes. I believed that this is because my All2 version has coalesced memory access and while my SM-All2 has less so because it is trying to fill four distinct buffers. I also believe that given the consistency in the performance of my shared memory version that I am hitting a bottle neck in terms of my software design so that I am not taking full advantage of the hardware. One last interesting trend is the difference between i2 and the k2/j2. It is clear to me that unrolling the I loop provides a better memory access pattern and a major boost in performance. A final thing to note is that I looked up the GPU online and according to this website I am approaching the max theoretical performance. (Source https://www.techpowerup.com/gpu-specs/geforce-gtx-970.c2620)