Justin Bahr
NUID: 001592707

High Performance Computing
Homework 3 - Question 4

a. Compared to my dense implementation from question 3, the OpenBLAS implementation is about 68 times faster. This is a massive speedup obtained from using level 3 BLAS. Like my dense implementation, OpenBLAS uses multithreading and cache optimizations such as loop tiling. However, the OpenBLAS library has been highly optimized to recognize and exploit the intricacies of a specific CPU architecture. This is done by tuning advantageous assembly instructions that can minimize runtime.

Dense	Runtime (ms)		**OpenBLAS**	Runtime (us)
1	56.267505		1	753.713
2	56.255503		2	728.797
3	56.241767		3	786.633
4	56.213329		4	744.22
5	56.286368		5	1105.755
Average	56.2528944		**Average**	823.8236
Fastest	56.213329		Fastest	728.797

Figure 1. Dense and OpenBLAS Runtimes

b. First I ran my program on the Explorer Cluster with a CPU node (Intel ® Xeon ® CPU E5-2680 v4 @ 2.40GHz, 14 cores per socket, 2 sockets with 527730564 KB of memory, and Linux 15.4.0). Second I ran my program on the Explorer Cluster with an AVX512 node (--constraint=cascadelake) (Intel ® Xeon ® Platinum 8276 CPU @ 2.20GHz, 28 cores per socket, 2 sockets, with 196090404 KB of memory, and Linux 5.14.0).

Xeon E5-2680	Runtime (us)		**Platinum 8276 (AVX512)**	Runtime (us)
1	753.713		1	1734.075
2	728.797		2	1987.034
3	786.633		3	1500.652
4	744.22		4	2278.455
5	1105.755		5	1355.663
Average	823.8236		**Average**	1771.1758
Fastest	728.797		Fastest	1355.663

Figure 2. Run on Explorer CPUs with and without AVX Support

The OpenBLAS program ran about twice as fast on the Xeon E5-2680 compared to the Xeon Platinum 8276. While I was hoping to see a performance increase while combining OpenBLAS and AVX512, this did not occur. I checked the assembly file for Q4_avx, and I did see vector instructions, but it is possible that there is a more fitting OpenBLAS function call to better utilize AVX512 support. I believe that the performance of the E5-2680 could be due to a combination of the faster clock and a smaller L2 data cache (faster cache latency). However, the Platinum 8276 has twice as many cores and AVX512 support, but this did not translate to a speedup.