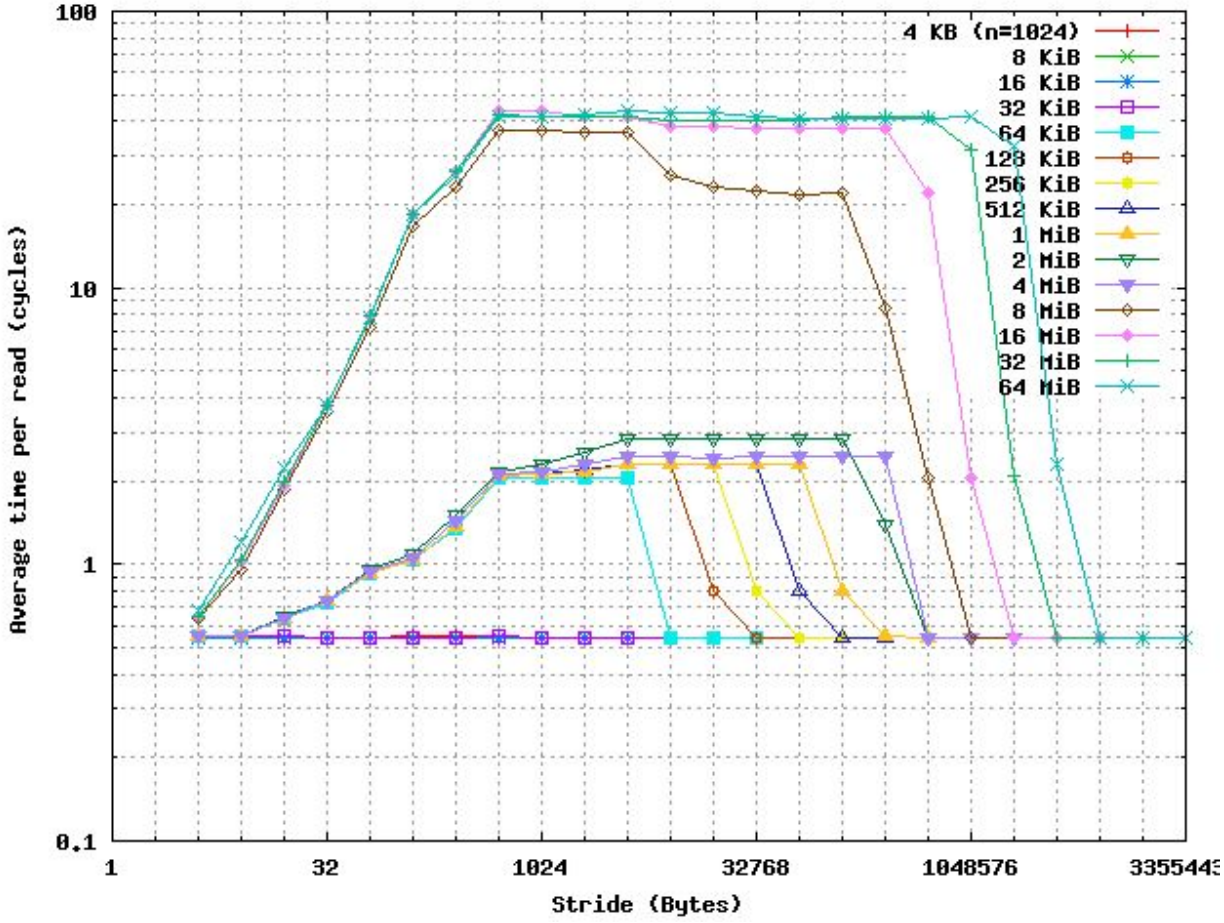
Homework 4
EECS 117

Stefan Cao (#79267250)
Andrew Yu (#23041544)

Part 1: Saavedra-Barrera benchmark

Q1.

The raw data can be found in *run.txt.*
Here is the plot of the data:



Q2.

1. There are 2 levels of caches, which is L1 and L2 caches.

2.

Level of Cache	Capacity Size	Line Size
L1	32 KiB	512 Bytes
L2	4 MiB	512 Bytes

Part 2: Cache blocked vectorized matrix multiply

Q3.

Naive Kernel Execution (using N:1024; K:1024; M:1024)**:**

Trial #	Execution time (s)
1	16.8996
2	17.4378
3	19.6573
4	27.4955
5	56.7398

Smallest execution time for naive kernel is 16.8996 seconds. Therefore the performance is $\frac{2n^3}{execution\ time * 10^9} = \frac{2*1024^3}{16.8996*10^9} = 0.1271\ GFLOPS/s$

Q4. *Please see code*

Q5.

The performance can be calculated by $\frac{2n^3}{execution\ time * 10^9}$

Cache-blocked Execution (using N:1024; K:1024; M:1024)**:**

Block Size	Execution time (s)	Performance (GFLOP/s)
1	16.0375	0.1339
2	31.4823	0.0682
4	23.6802	0.0906

8	11.2608	0.1907
16	19.4812	0.1102
32	19.6193	0.1094
64	16.2401	0.1322
128	16.7396	0.1282
256	18.0982	0.1186
512	16.1207	0.1332
1024	61.8967	0.0346

From the chart, we can see that when the block size is decreased from 1024 to 512, the execution time has a significant gap. We assume that the size of the cache will be 512.

Q6.

The performance can be calculated by $\frac{2n^3}{execution\ time * 10^9}$

SIMD vectorized Execution (using N:1024; K:1024; M:1024):

Block Size	**Execution time (s)**	**Performance (GFLOP/s)**
1	-	-
2	3.15226	0.6812
4	3.16482	0.6785
8	2.81701	0.7623
16	2.65266	0.8095
32	3.19301	0.6725
64	2.35042	0.9136
128	2.87075	0.7480
256	2.96737	0.7236
512	2.50867	0.8560
1024	2.77146	0.7748