Simulation and Scientific Computing
Assignment 1

Code to optimize matrix-matrix multiplication

-Aishvarya Kumar Jain

22060400

-Narayanan Achuthan

22078846

-Prachi Chirputkar

22064059

Performance Plots:

All performance characteristics are plotted with Matrix size in logscale base 2 and performance characteristics in logscale base 10 for various cases.

a) Naive matrix multiplication.

b) Code implementing padding alone.

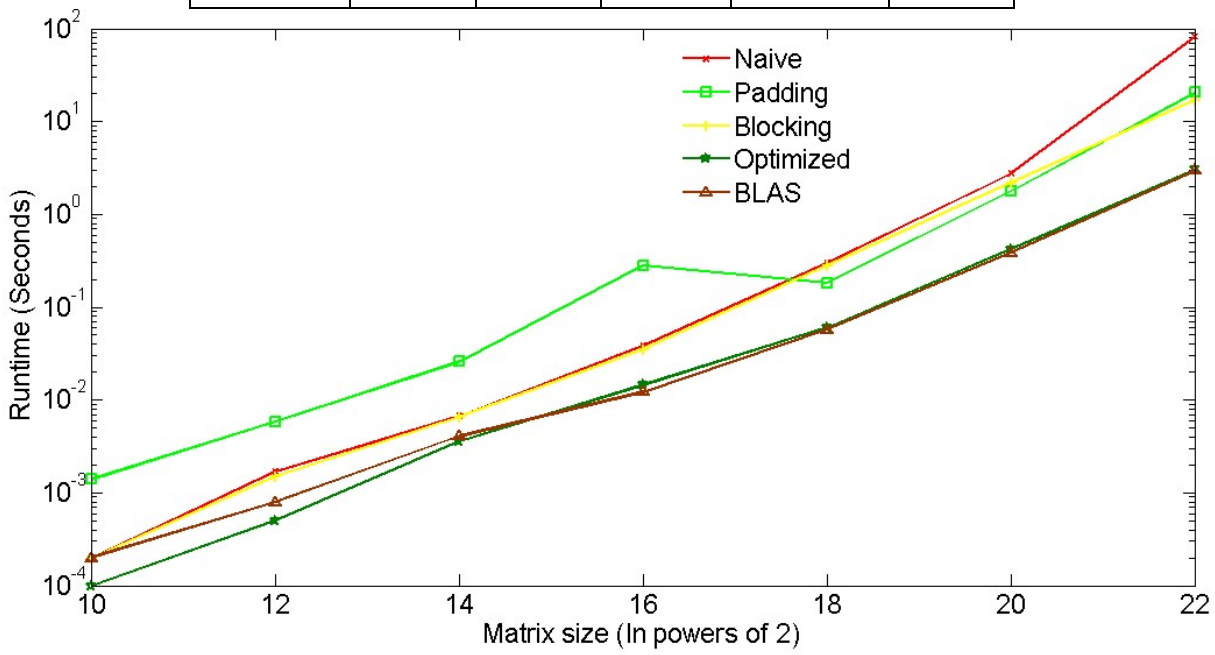
c) Code implementing blocking alone.

d) Optimized code implementing

i) Padding, ii) Blocking and Unrolling and iii) Loop interchange

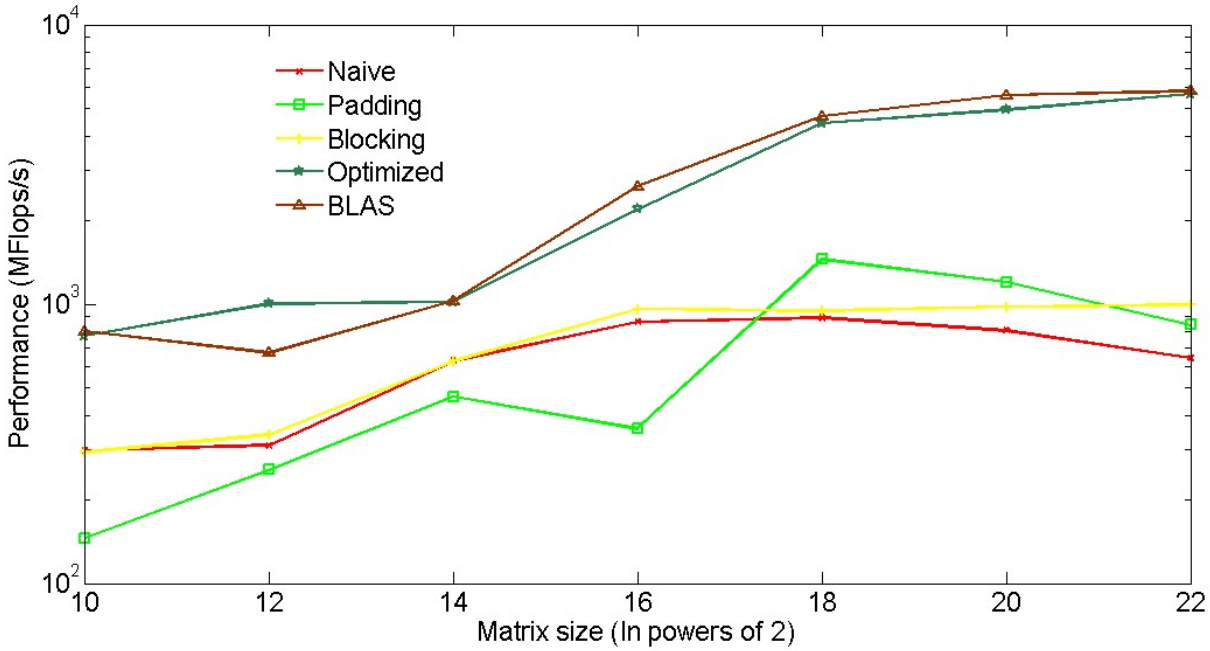
1) Runtime vs Matrix size:

Matrix Size	Runtime in Seconds				
	Naïve	Padding	Blocking	Optimized	BLAS
32x32	0.0002	0.0014	0.0002	0.0001	0.0002
64x64	0.0017	0.0058	0.0015	0.0005	0.0008
128x128	0.0067	0.026	0.0066	0.0036	0.0041
256x256	0.0384	0.2819	0.035	0.0146	0.0122
512x512	0.3004	0.1827	0.2816	0.0603	0.057
1024x1024	2.7347	1.7724	2.1892	0.4207	0.3835
2048x2048	81.7576	20.4302	17.011	3.0393	2.9653



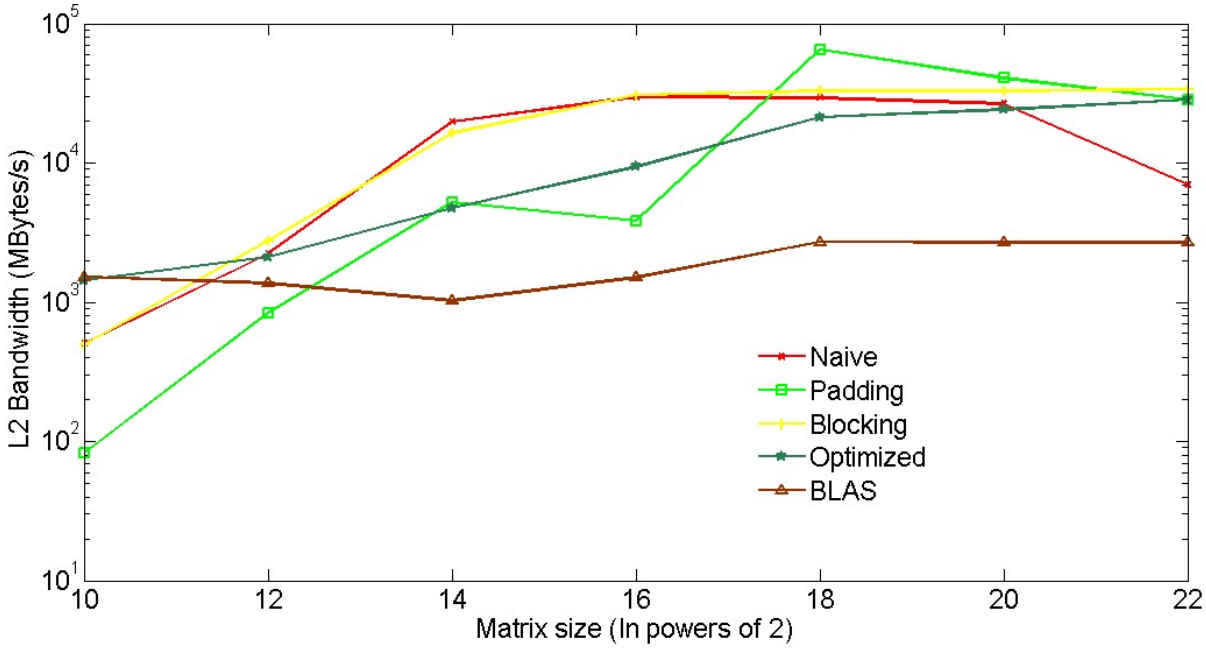
2) Performance (MFlops/s) vs Matrix Size:

Matrix Size	Performance in MFlops/s				
	Naïve	Padding	Blocking	Optimized	BLAS
32x32	299.31	144.6863	294.756	766.9995	798.2248
64x64	312.411	255.02	340.15	1004.42	669.58
128x128	626.03	465.93	624.8	1014.92	1028.24
256x256	863.6688	358.0715	957.8095	2186	2645.498
512x512	893.5177	1446.0774	947.64	4448.2181	4702.33
1024x1024	804.9273	1199.9846	978.8534	4972.9583	5597.65
2048x2048	641.3415	840.904	990.4511	5652.5337	5797.98



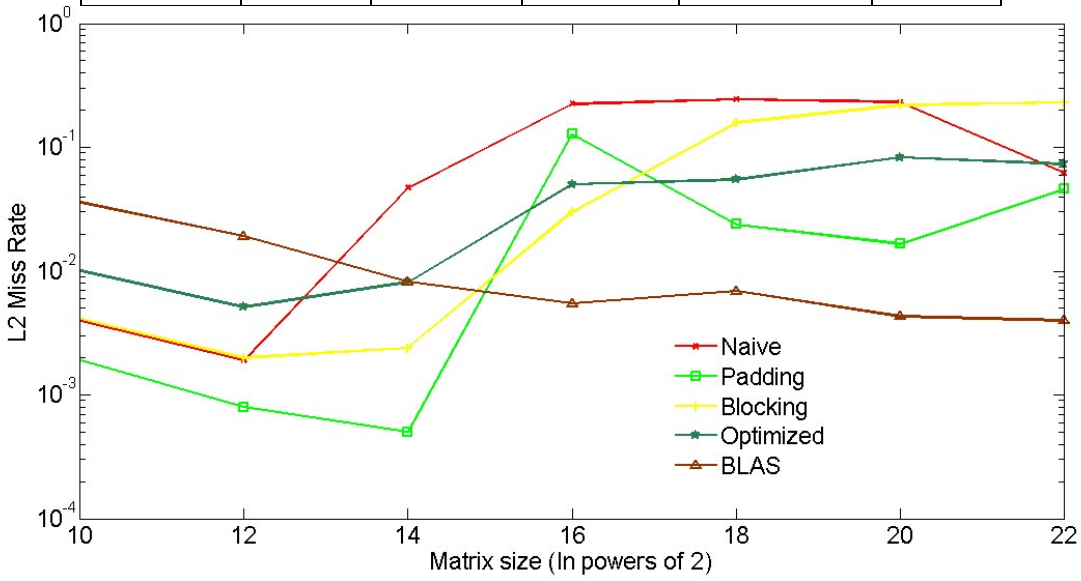
3) L2 bandwidth (MBytes/s) vs Matrix Size:

Matrix Size	L2 Bandwidth in Mbytes/s				
	Naïve	Padding	Blocking	Optimized	BLAS
32x32	507.0074	81.8856	496.6517	1439.0105	1515.0546
64x64	2224.44	832.06	2752.31	2115.38	1369.95
128x128	19703.49	5221.1586	16362.29	4729.81	1022.84
256x256	29780.7392	3836.5731	30858.0817	9373.44	1502.9707
512x512	29411.303	64972.6359	32972.2993	21358.2034	2722.65
1024x1024	26687.0831	40743.697	32779.4754	24126.689	2692.2308
2048x2048	6959.5441	28550.6819	33697.7671	28430.3685	2691.7823



4) L2 miss rate vs Matrix Size:

Matrix Size	L2 Miss Rate				
	Naïve	Padding	Blocking	Optimized	BLAS
32x32	0.004	0.0019	0.0041	0.0101	0.036
64x64	0.0019	0.0008	0.002	0.0051	0.0192
128x128	0.0471	0.0005	0.0024	0.0081	0.0082
256x256	0.2245	0.1279	0.0302	0.0501	0.0055
512x512	0.2436	0.0238	0.1583	0.0548	0.0069
1024x1024	0.2315	0.0166	0.2186	0.0828	0.0043
2048x2048	0.0619	0.0461	0.2293	0.0731	0.004



Inferences:

The performance characteristics of the optimized code is better when compared with naive implementation of matrix matrix multiplication.

The number of flops (MFlops/s) is significantly higher for the optimized code as it utilizes the temporal locality of the L2 cache. This can be inferred from the higher L2 bandwidth for the optimized case. Data transfer from the main memory is reduced significantly as blocking ensures data reuse. Loop interchange ensures lower cache misses.