i.

Threads = 4

|  |  |  |  |
| --- | --- | --- | --- |
| Matrix size/Performance | 4096x4096 | 2048x2048 | 1024x1024 |
| Run #1 | 2368.5ms | 287.732ms | 33.2671ms |
| Run #2 | 2394.46ms | 287.887ms | 32.8973ms |
| Run #3 | 2363.08ms | 287.917ms | 34.5488ms |
| Run #4 | 2380.15ms | 287.897ms | 32.7471ms |
| Run #5 | 2385.02ms | 288.016ms | 32.6755ms |
| Run #6 | 2337.72ms | 287.903ms | 32.6261ms |
| Run #7 | 2337.78ms | 287.94ms | 32.6892ms |
| Run #8 | 2372.35ms | 294.154ms | 32.6209ms |
| Run #9 | 2368.93ms | 287.943ms | 32.611ms |
| Run #10 | 2363.63ms | 287.878ms | 32.6098ms |
| Run #11 | 2370.62ms | 287.871ms | 32.5978ms |
| Run #12 | 2378.94ms | 291.052ms | 32.5901ms |
| Run #13 | 2375.54ms | 287.932ms | 32.5438ms |
| Run #14 | 2336.98ms | 287.903ms | 32.5779ms |
| Run #15 | 2375.44ms | 288.321ms | 32.5633ms |
| Run #16 | 2374.74ms | 296.406ms | 32.6123ms |
| Run #17 | 2358.36ms | 290.036ms | 32.5795ms |
| Run #18 | 2391.72ms | 289.976ms | 32.658ms |
| Run #19 | 2374.72ms | 292.478ms | 32.5529ms |
| Run #20 | 2396.86ms | 291.018ms | 32.7018ms |

ii.

Matrix size is 2048x2048

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Block Size/Performance | 16x16 | 32x32 | 64x64 | 128x128 | 256x256 |
| Run #1 | 411.139ms | 287.732ms | 337.997ms | 347.549ms | 373.445ms |
| Run #2 | 436.973ms | 287.887ms | 329.495ms | 347.62ms | 372.153ms |
| Run #3 | 418.501ms | 287.917ms | 328.294ms | 347.717ms | 372.159ms |
| Run #4 | 399.976ms | 287.897ms | 329.522ms | 347.466ms | 371.975ms |
| Run #5 | 418.545ms | 288.016ms | 329.714ms | 347.568ms | 372.086ms |
| Run #6 | 424.681ms | 287.903ms | 335.395ms | 347.638ms | 372.421ms |
| Run #7 | 394.351ms | 287.94ms | 330.131ms | 347.653ms | 371.796ms |
| Run #8 | 409.836ms | 294.154ms | 329.951ms | 347.685ms | 371.516ms |
| Run #9 | 435.195ms | 287.943ms | 331.98ms | 347.551ms | 372.193ms |
| Run #10 | 413.818ms | 287.878ms | 329.183ms | 347.562ms | 371.408ms |
| Run #11 | 419.693ms | 287.871ms | 329.139ms | 347.771ms | 371.307ms |
| Run #12 | 401.42ms | 291.052ms | 332.392ms | 347.53ms | 373.42ms |
| Run #13 | 419.655ms | 287.932ms | 330.328ms | 347.529ms | 372.711ms |
| Run #14 | 424.378ms | 287.903ms | 333.109ms | 347.684ms | 371.911ms |
| Run #15 | 417.108ms | 288.321ms | 333.723ms | 347.518ms | 371.569ms |
| Run #16 | 400.059ms | 296.406ms | 330.098ms | 347.617ms | 370.976ms |
| Run #17 | 394.389ms | 290.036ms | 330.736ms | 347.567ms | 371.708ms |
| Run #18 | 426.401ms | 289.976ms | 330.015ms | 347.547ms | 374.888ms |
| Run #19 | 424.271ms | 292.478ms | 330.33ms | 347.62ms | 371.917ms |
| Run #20 | 428.898ms | 291.018ms | 333.999ms | 347.592ms | 375.205ms |

Matrix size 1024x1024

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Block Size/Performance | 16x16 | 32x32 | 64x64 | 128x128 | 256x256 |
| Run #1 | 33.8518ms | 33.2671ms | 40.896ms | 43.3774ms | 47.2969ms |
| Run #2 | 33.4527ms | 32.8973ms | 40.7637ms | 43.2724ms | 47.0633ms |
| Run #3 | 33.5109ms | 34.5488ms | 40.4743ms | 43.2511ms | 46.8366ms |
| Run #4 | 33.5751ms | 32.7471ms | 40.4076ms | 43.2625ms | 46.7481ms |
| Run #5 | 33.4667ms | 32.6755ms | 40.5156ms | 43.2826ms | 46.7237ms |
| Run #6 | 33.5762ms | 32.6261ms | 40.2509ms | 43.2646ms | 46.9262ms |
| Run #7 | 33.5452ms | 32.6892ms | 44.3232ms | 43.2993ms | 46.9398ms |
| Run #8 | 33.4682ms | 32.6209ms | 40.5112ms | 43.2796ms | 46.7344ms |
| Run #9 | 33.5412ms | 32.611ms | 40.5792ms | 43.2702ms | 47.0972ms |
| Run #10 | 33.4185ms | 32.6098ms | 40.5428ms | 43.2662ms | 47.0428ms |
| Run #11 | 33.4646ms | 32.5978ms | 40.5443ms | 43.2606ms | 47.1428ms |
| Run #12 | 33.5389ms | 32.5901ms | 40.6641ms | 43.258ms | 47.1091ms |
| Run #13 | 33.5725ms | 32.5438ms | 40.3985ms | 43.2704ms | 47.167ms |
| Run #14 | 33.5592ms | 32.5779ms | 40.5743ms | 43.272ms | 46.8824ms |
| Run #15 | 33.544ms | 32.5633ms | 40.3604ms | 43.297ms | 47.1355ms |
| Run #16 | 33.6165ms | 32.6123ms | 40.4975ms | 43.2616ms | 47.1436ms |
| Run #17 | 33.4571ms | 32.5795ms | 40.4196ms | 43.2575ms | 46.8593ms |
| Run #18 | 33.5451ms | 32.658ms | 43.7887ms | 43.2799ms | 46.8828ms |
| Run #19 | 33.5742ms | 32.5529ms | 40.3964ms | 43.2523ms | 46.8274ms |
| Run #20 | 33.5442ms | 32.7018ms | 40.5118ms | 43.2622ms | 46.9984ms |

Matrix size 4096x4096

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Block Size/Performance | 16x16 | 32x32 | 64x64 | 128x128 | 256x256 |
| Run #1 | 3790.98ms | 2368.5ms | 2742.76ms | 2790.34ms | 2974.79ms |
| Run #2 | 3756.72ms | 2394.46ms | 2749.73ms | 2787.09ms | 2973.11ms |
| Run #3 | 3708.52ms | 2363.08ms | 2751.59ms | 2796.01ms | 2977.53ms |
| Run #4 | 3650.65ms | 2380.15ms | 2749.55ms | 2795.19ms | 2973.43ms |
| Run #5 | 3811.48ms | 2385.02ms | 2750.82ms | 2782.87ms | 2974.67ms |
| Run #6 | 3742.02ms | 2337.72ms | 2743.12ms | 2789.85ms | 2975.08ms |
| Run #7 | 3857.71ms | 2337.78ms | 2747.42ms | 2788.11ms | 2977.39ms |
| Run #8 | 3690.24ms | 2372.35ms | 2746.47ms | 2791.8ms | 2976.33ms |
| Run #9 | 3729.14ms | 2368.93ms | 2746.53ms | 2790.61ms | 2990.95ms |
| Run #10 | 3673.19ms | 2363.63ms | 2750.13ms | 2787.9ms | 2977.78ms |
| Run #11 | 3699.83ms | 2370.62ms | 2739.32ms | 2788.15ms | 2996.39ms |
| Run #12 | 3753.78ms | 2378.94ms | 2741.62ms | 2790.34ms | 2980.47ms |
| Run #13 | 3851.43m | 2375.54ms | 2749.01ms | 2790.34ms | 2979.86ms |
| Run #14 | 3753.27ms | 2336.98ms | 2751.15ms | 2788.46ms | 2988.19ms |
| Run #15 | 3801.38ms | 2375.44ms | 2769.99ms | 2789.96ms | 2993.23ms |
| Run #16 | 3774.51ms | 2374.74ms | 2747.68ms | 2785.19ms | 2979.69ms |
| Run #17 | 3692.84ms | 2358.36ms | 2746.77ms | 2791.41ms | 2984.54ms |
| Run #18 | 3745.68ms | 2391.72ms | 2748.9ms | 2788.87ms | 2983.82ms |
| Run #19 | 3612.03ms | 2374.72ms | 2782.57ms | 2787.94ms | 2983.38ms |
| Run #20 | 3709.51ms | 2396.86ms | 2745.6ms | 2789.46ms | 2975.74ms |

iii.

Matrix size 4096x4096 Threads = 4 Matrix size 4096x4096 Threads= 1

|  |  |  |  |
| --- | --- | --- | --- |
| Block Size /Performance | 16x16 | 32x32 | 64x64 |
| Run #1 | 3790.98ms | 2368.5ms | 2742.76ms |
| Run #2 | 3756.72ms | 2394.46ms | 2749.73ms |
| Run #3 | 3708.52ms | 2363.08ms | 2751.59ms |
| Run #4 | 3650.65ms | 2380.15ms | 2749.55ms |
| Run #5 | 3811.48ms | 2385.02ms | 2750.82ms |
| Run #6 | 3742.02ms | 2337.72ms | 2743.12ms |
| Run #7 | 3857.71ms | 2337.78ms | 2747.42ms |
| Run #8 | 3690.24ms | 2372.35ms | 2746.47ms |
| Run #9 | 3729.14ms | 2368.93ms | 2746.53ms |
| Run #10 | 3673.19ms | 2363.63ms | 2750.13ms |
| Run #11 | 3699.83ms | 2370.62ms | 2739.32ms |
| Run #12 | 3753.78ms | 2378.94ms | 2741.62ms |
| Run #13 | 3851.43m | 2375.54ms | 2749.01ms |
| Run #14 | 3753.27ms | 2336.98ms | 2751.15ms |
| Run #15 | 3801.38ms | 2375.44ms | 2769.99ms |
| Run #16 | 3774.51ms | 2374.74ms | 2747.68ms |
| Run #17 | 3692.84ms | 2358.36ms | 2746.77ms |
| Run #18 | 3745.68ms | 2391.72ms | 2748.9ms |
| Run #19 | 3612.03ms | 2374.72ms | 2782.57ms |
| Run #20 | 3709.51ms | 2396.86ms | 2745.6ms |

|  |  |  |  |
| --- | --- | --- | --- |
| Block Size /Performance | 16x16 | 32x32 | 64x64 |
| Run #1 | 13786.6ms | 8849.4ms | 9617.66ms |
| Run #2 | 13769.5ms | 8856.85ms | 9622.72ms |
| Run #3 | 13776.1ms | 8928.02ms | 9616.29ms |
| Run #4 | 13789ms | 8859.87ms | 9618.53ms |
| Run #5 | 13995ms | 8844.54ms | 9623.59ms |
| Run #6 | 13911.7ms | 8850.31ms | 9613.72ms |
| Run #7 | 14258.6ms | 8851.81ms | 9613.27ms |
| Run #8 | 14212.6ms | 8846.99ms | 9617.91ms |
| Run #9 | 13795.6ms | 8848.5ms | 9620.8ms |
| Run #10 | 13785.6ms | 8843.9ms | 9630.28ms |
| Run #11 | 13922.5ms | 8846.1ms | 9639.64ms |
| Run #12 | 14031.4ms | 8847.47ms | 9623.5ms |
| Run #13 | 13808.1ms | 8851.5ms | 9625.13ms |
| Run #14 | 13785.2ms | 8849.54ms | 9622.17ms |
| Run #15 | 13774.3ms | 8849.96ms | 9615.32ms |
| Run #16 | 13779.1ms | 8851.59ms | 9614.3ms |
| Run #17 | 13794.7ms | 8844.63ms | 9618.73ms |
| Run #18 | 13805ms | 8845.98ms | 9615.72ms |
| Run #19 | 13772ms | 8847.74ms | 9618.9ms |
| Run #20 | 13784.7ms | 8851.02ms | 9611.57m |

|  |  |  |  |
| --- | --- | --- | --- |
| Block Size/  Performance | 16x16 | 32x32 | 64x64 |
| Run #1 | 1631.1ms | 1089.24ms | 1208.84ms |
| Run #2 | 1631.26ms | 1073.35ms | 1207.65ms |
| Run #3 | 1629.73ms | 1072.64ms | 1208.2ms |
| Run #4 | 1630.49ms | 1072.11ms | 1207.98ms |
| Run #5 | 1630.29ms | 1072.17ms | 1206.93ms |
| Run #6 | 1629.98ms | 1072.15ms | 1207.2ms |
| Run #7 | 1630ms | 1072.24ms | 1207.36ms |
| Run #8 | 1634.21ms | 1072.39ms | 1205.75ms |
| Run #9 | 1632.12ms | 1071.99ms | 1205.44ms |
| Run #10 | 1633ms | 1072.41ms | 1209.45ms |
| Run #11 | 1630.34ms | 1072.11ms | 1208.12ms |
| Run #12 | 1630.01ms | 1072.91ms | 1208.1ms |
| Run #13 | 1631.22ms | 1074.13ms | 1208.34ms |
| Run #14 | 1631.02ms | 1073.54ms | 1208.35ms |
| Run #15 | 1629.61ms | 1073.26ms | 1207.54ms |
| Run #16 | 1636.59ms | 1074.99ms | 1207.81ms |
| Run #17 | 1633.02ms | 1073.63ms | 1207.95ms |
| Run #18 | 1633.49ms | 1073.36ms | 1207.84ms |
| Run #19 | 1634.73ms | 1073.08ms | 1204.56ms |
| Run #20 | 1638.31ms | 1072.75ms | 1207.59ms |

Matrix size 2048x2048 Threads = 4 Matrix size 2048x2048 Threads= 1

|  |  |  |  |
| --- | --- | --- | --- |
| Block Size/  Performance | 16x16 | 32x32 | 64x64 |
| Run #1 | 411.139ms | 287.732ms | 337.997ms |
| Run #2 | 436.973ms | 287.887ms | 329.495ms |
| Run #3 | 418.501ms | 287.917ms | 328.294ms |
| Run #4 | 399.976ms | 287.897ms | 329.522ms |
| Run #5 | 418.545ms | 288.016ms | 329.714ms |
| Run #6 | 424.681ms | 287.903ms | 335.395ms |
| Run #7 | 394.351ms | 287.94ms | 330.131ms |
| Run #8 | 409.836ms | 294.154ms | 329.951ms |
| Run #9 | 435.195ms | 287.943ms | 331.98ms |
| Run #10 | 413.818ms | 287.878ms | 329.183ms |
| Run #11 | 419.693ms | 287.871ms | 329.139ms |
| Run #12 | 401.42ms | 291.052ms | 332.392ms |
| Run #13 | 419.655ms | 287.932ms | 330.328ms |
| Run #14 | 424.378ms | 287.903ms | 333.109ms |
| Run #15 | 417.108ms | 288.321ms | 333.723ms |
| Run #16 | 400.059ms | 296.406ms | 330.098ms |
| Run #17 | 394.389ms | 290.036ms | 330.736ms |
| Run #18 | 426.401ms | 289.976ms | 330.015ms |
| Run #19 | 424.271ms | 292.478ms | 330.33ms |
| Run #20 | 428.898ms | 291.018ms | 333.999ms |

|  |  |  |  |
| --- | --- | --- | --- |
| Block Size/  Performance | 16x16 | 32x32 | 64x64 |
| Run #1 | 119.815ms | 120.655ms | 146.898ms |
| Run #2 | 118.835ms | 119.128ms | 146.246ms |
| Run #3 | 119.025ms | 117.944ms | 146.227ms |
| Run #4 | 118.78ms | 116.816ms | 146.931ms |
| Run #5 | 119.068ms | 119.299ms | 146.731ms |
| Run #6 | 118.76ms | 116.994ms | 146.121ms |
| Run #7 | 118.986ms | 118.075ms | 146.873ms |
| Run #8 | 118.953ms | 117.022ms | 146.368ms |
| Run #9 | 121.126ms | 116.575ms | 146.201ms |
| Run #10 | 120.456ms | 116.377ms | 146.568ms |
| Run #11 | 119.044ms | 120.895ms | 147.318ms |
| Run #12 | 119.207ms | 120.548ms | 146.62ms |
| Run #13 | 119.39ms | 117.614ms | 146.654ms |
| Run #14 | 119.052ms | 116.304ms | 146.254ms |
| Run #15 | 119.011ms | 116.779ms | 146.132ms |
| Run #16 | 119.146ms | 117.86ms | 146.295ms |
| Run #17 | 123.034ms | 116.847ms | 147.387ms |
| Run #18 | 122.035ms | 116.647ms | 146.806ms |
| Run #19 | 118.854ms | 120.459ms | 146.312ms |
| Run #20 | 119.197ms | 117.7ms | 146.41ms |

Matrix size 1024x1024 Threads = 4 Matrix size 1024x1024Threads= 1

|  |  |  |  |
| --- | --- | --- | --- |
| Block Size/  Performance | 16x16 | 32x32 | 64x64 |
| Run #1 | 33.8518ms | 33.2671ms | 40.896ms |
| Run #2 | 33.4527ms | 32.8973ms | 40.7637ms |
| Run #3 | 33.5109ms | 34.5488ms | 40.4743ms |
| Run #4 | 33.5751ms | 32.7471ms | 40.4076ms |
| Run #5 | 33.4667ms | 32.6755ms | 40.5156ms |
| Run #6 | 33.5762ms | 32.6261ms | 40.2509ms |
| Run #7 | 33.5452ms | 32.6892ms | 44.3232ms |
| Run #8 | 33.4682ms | 32.6209ms | 40.5112ms |
| Run #9 | 33.5412ms | 32.611ms | 40.5792ms |
| Run #10 | 33.4185ms | 32.6098ms | 40.5428ms |
| Run #11 | 33.4646ms | 32.5978ms | 40.5443ms |
| Run #12 | 33.5389ms | 32.5901ms | 40.6641ms |
| Run #13 | 33.5725ms | 32.5438ms | 40.3985ms |
| Run #14 | 33.5592ms | 32.5779ms | 40.5743ms |
| Run #15 | 33.544ms | 32.5633ms | 40.3604ms |
| Run #16 | 33.6165ms | 32.6123ms | 40.4975ms |
| Run #17 | 33.4571ms | 32.5795ms | 40.4196ms |
| Run #18 | 33.5451ms | 32.658ms | 43.7887ms |
| Run #19 | 33.5742ms | 32.5529ms | 40.3964ms |
| Run #20 | 33.5442ms | 32.7018ms | 40.5118ms |

Comparison with MKL

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Matric size | 1024x1024 | 2048x2048 | 4096x4096 | 1024x1024 | 2048x2048 | 4096x4096 |
| MKL ? | NO | NO | NO | YES | YES | YES |
| Block size | 16x16 | 16x16 | 16x16 | N/A | N/A | N/A |
| Thread = 4 performance | 33.54ms | 415.96ms | 3740.24ms | 7.31ms | 53.16ms | 410.99ms |
| Thread = 1 performance | 119.58ms | 1632.02ms | 13866.86ms | 23.72ms | 185.17ms | 1544.73ms |
| Block size | 32x32 | 32x32 | 32x32 | N/A | N/A | N/A |
| Thread = 4 performance | 32.76ms | 289.41ms | 2370.27ms | N/A | N/A | N/A |
| Thread = 1 performance | 118.02ms | 1073.72ms | 8853.28ms | N/A | N/A | N/A |
| Block size | 64x64 | 64x64 | 64x64 | N/A | N/A | N/A |
| Thread = 4 performance | 40.87ms | 331.27ms | 2750.03ms | N/A | N/A | N/A |
| Thread = 1 performance | 146.56ms | 1207.55ms | 9619.98ms | N/A | N/A | N/A |

**-** **Does operating at certain (large or small) matrix sizes seem to make it less important what the size of blocks is (for performance)?**

Yes, take 1024x1024 matrix size and block size 16x16 and compare it to 32x32

Performance (Thread = 4) of 16x16 : 33.54ms

Performance (Thread = 4) of 32x32 : 32.76ms

Thus, the performance at a small resolution (1024x1024) and small block size (16x16 and 32x32) is negligible due to the performance differing by 2% only.

**- Are you seeing the same changes in parameters such as block size helping performance in some resolutions and hindering performance in others?**

Yes, the block size 32x32 yields the best performance for any resolution. I think that this is happening because of the cache size. For example, if the cache size was larger the best block size might be 64x64.

**- For some of the resolutions/block sizes in your experiment, are you witnessing speed-up that is not quite proportional with the number of cores used when using all vs. one thread?**

No, from part (iii), all the tests are scaling according to the number of threads we have.

**Additional Commentary:**

The performance increase from 1 thread to 4 threads is always a 3.5-3.7 times increase. I believe that it is not a 4 times increase because of the work it needs the to split the job into 4 threads.

**Conclusion:**

Block size helps with the performance by around 0.3-0.5 times increase.

The perfect block size is determined by the cache size (i.e. larger cache size -> perfect block size is larger)

Threads increase the performance relative to the amount if threads(i.e. 2 threads will have around 2 times better performance than 1 thread)