**Lab 8 Write-Up**

**Part A – 1024 x 1024 Matrix**

**How were Matrix elements generated?**

Using the initializeArray1D() function which uses a known seed for random number generation. The same process was used for all the programs.

**Part 1 – Using Global Memory**

**GPU time (End-to-End) -** 6.966720ms

**GPU time (MMM)-** 0.195008ms

**Best CPU time -** 6.380000s

**TOL –** .00001

As it can be seen above, the execution for the MMM operation is a lot better on the GPU than on the CPU.

The following are the first 50 values of the output for the CPU & GPU execution :

CPU GPU

0: 260.16702271 260.16702271

1: 264.76211548 264.76211548

2: 256.09481812 256.09481812

3: 259.13980103 259.13980103

4: 260.20260620 260.20260620

5: 262.05215454 262.05215454

6: 263.80471802 263.80471802

7: 259.97232056 259.97232056

8: 258.40426636 258.40426636

9: 262.87518311 262.87515259

10: 268.53280640 268.53280640

11: 256.71725464 256.71725464

12: 260.61993408 260.61993408

13: 256.49404907 256.49404907

14: 263.28805542 263.28805542

15: 254.21157837 254.21157837

16: 255.62844849 255.62844849

17: 258.50457764 258.50457764

18: 254.50840759 254.50839233

19: 257.79989624 257.79989624

20: 257.38311768 257.38311768

21: 253.41467285 253.41468811

22: 260.37518311 260.37515259

23: 264.34384155 264.34384155

24: 260.64489746 260.64489746

25: 257.31823730 257.31823730

26: 266.97552490 266.97552490

27: 263.60006714 263.60009766

28: 256.80285645 256.80288696

29: 256.86828613 256.86828613

30: 261.32543945 261.32543945

31: 254.91474915 254.91474915

32: 257.95822144 257.95822144

33: 260.20166016 260.20166016

34: 251.58848572 251.58848572

35: 252.25143433 252.25141907

36: 257.06524658 257.06524658

37: 265.28146362 265.28146362

38: 259.47805786 259.47805786

39: 252.55696106 252.55696106

40: 264.55999756 264.55999756

41: 269.16696167 269.16696167

42: 266.60421753 266.60421753

43: 254.47082520 254.47082520

44: 266.10586548 266.10586548

45: 259.99789429 259.99792480

46: 258.28851318 258.28851318

47: 256.92730713 256.92727661

48: 249.55557251 249.55557251

49: 244.17973328 244.17974854

It can be seen that the values produced as output on the CPU & the GPU are roughly the same.

**Part 2 – Using Shared Memory**

**GPU time (End-to-End) -** 5.849696ms

**GPU time (MMM) -** 0.076832ms

**Best CPU time -** 6.380000s

**TOL –** .00001

As it can be seen above, using shared memory improves the execution time on the GPU.

The following are the first 50 values of the output for the CPU & GPU execution:

CPU GPU

0: 260.16702271 260.16702271

1: 264.76211548 264.76211548

2: 256.09481812 256.09481812

3: 259.13980103 259.13980103

4: 260.20260620 260.20260620

5: 262.05215454 262.05215454

6: 263.80471802 263.80471802

7: 259.97232056 259.97232056

8: 258.40426636 258.40426636

9: 262.87518311 262.87515259

10: 268.53280640 268.53280640

11: 256.71725464 256.71725464

12: 260.61993408 260.61993408

13: 256.49404907 256.49404907

14: 263.28805542 263.28805542

15: 254.21157837 254.21157837

16: 255.62844849 255.62844849

17: 258.50457764 258.50457764

18: 254.50840759 254.50839233

19: 257.79989624 257.79989624

20: 257.38311768 257.38311768

21: 253.41467285 253.41468811

22: 260.37518311 260.37515259

23: 264.34384155 264.34384155

24: 260.64489746 260.64489746

25: 257.31823730 257.31823730

26: 266.97552490 266.97552490

27: 263.60006714 263.60009766

28: 256.80285645 256.80288696

29: 256.86828613 256.86828613

30: 261.32543945 261.32543945

31: 254.91474915 254.91474915

32: 257.95822144 257.95822144

33: 260.20166016 260.20166016

34: 251.58848572 251.58848572

35: 252.25143433 252.25141907

36: 257.06524658 257.06524658

37: 265.28146362 265.28146362

38: 259.47805786 259.47805786

39: 252.55696106 252.55696106

40: 264.55999756 264.55999756

41: 269.16696167 269.16696167

42: 266.60421753 266.60421753

43: 254.47082520 254.47082520

44: 266.10586548 266.10586548

45: 259.99789429 259.99792480

46: 258.28851318 258.28851318

47: 256.92730713 256.92727661

48: 249.55557251 249.55557251

49: 244.17973328 244.17974854

Again, It can be seen that the values produced as output on the CPU & the GPU are roughly the same.

**Part 3- Using Shared Memory & Loop Unrolling**

**GPU time(End-to-End) -** 6.803776ms

**GPU time (MMM)-** 0.077280ms

**Best CPU time -** 6.380000s

**TOL –** .00001

Surprisingly, loop unrolling makes the performance of MMM on shared memory worse, as seen with the increased execution time. However, this time is still lower than that of the version using global memory.

The following are the first 50 values of the output for the CPU & GPU execution:

CPU GPU

0: 260.16702271 260.16702271

1: 264.76211548 264.76211548

2: 256.09481812 256.09481812

3: 259.13980103 259.13980103

4: 260.20260620 260.20260620

5: 262.05215454 262.05215454

6: 263.80471802 263.80471802

7: 259.97232056 259.97232056

8: 258.40426636 258.40426636

9: 262.87518311 262.87515259

10: 268.53280640 268.53280640

11: 256.71725464 256.71725464

12: 260.61993408 260.61993408

13: 256.49404907 256.49404907

14: 263.28805542 263.28805542

15: 254.21157837 254.21157837

16: 255.62844849 255.62844849

17: 258.50457764 258.50457764

18: 254.50840759 254.50839233

19: 257.79989624 257.79989624

20: 257.38311768 257.38311768

21: 253.41467285 253.41468811

22: 260.37518311 260.37515259

23: 264.34384155 264.34384155

24: 260.64489746 260.64489746

25: 257.31823730 257.31823730

26: 266.97552490 266.97552490

27: 263.60006714 263.60009766

28: 256.80285645 256.80288696

29: 256.86828613 256.86828613

30: 261.32543945 261.32543945

31: 254.91474915 254.91474915

32: 257.95822144 257.95822144

33: 260.20166016 260.20166016

34: 251.58848572 251.58848572

35: 252.25143433 252.25141907

36: 257.06524658 257.06524658

37: 265.28146362 265.28146362

38: 259.47805786 259.47805786

39: 252.55696106 252.55696106

40: 264.55999756 264.55999756

41: 269.16696167 269.16696167

42: 266.60421753 266.60421753

43: 254.47082520 254.47082520

44: 266.10586548 266.10586548

45: 259.99789429 259.99792480

46: 258.28851318 258.28851318

47: 256.92730713 256.92727661

48: 249.55557251 249.55557251

49: 244.17973328 244.17974854

Again, the values produced as output on the CPU & the GPU are roughly the same.

**Part B – 2048 x 2048 Matrix**

**Part 1 – Using Global Memory**

**GPU time(End-to-End) -** 21.573215ms

**GPU time(MMM) -** 0.568096ms

**Best CPU time -** 49.150000s

**TOL-** 0.0001

As it can be seen above, the execution for the MMM operation is a lot better on the GPU than on the CPU, like for the 1024x1024 matrix.

The following are the first 50 values of the output for the CPU & GPU execution :

0: 511.96847534 511.96847534

1: 502.86544800 502.86541748

2: 503.09921265 503.09921265

3: 503.61126709 503.61126709

4: 504.44244385 504.44244385

5: 514.62268066 514.62268066

6: 504.18115234 504.18112183

7: 512.17102051 512.17102051

8: 512.47802734 512.47802734

9: 504.64291382 504.64291382

10: 516.60943604 516.60943604

11: 511.41964722 511.41961670

12: 496.94531250 496.94531250

13: 512.66119385 512.66119385

14: 507.61819458 507.61819458

15: 499.52340698 499.52343750

16: 504.20010376 504.20010376

17: 510.20028687 510.20025635

18: 507.44528198 507.44528198

19: 497.04547119 497.04550171

20: 506.25741577 506.25738525

21: 491.61154175 491.61160278

22: 526.00372314 526.00378418

23: 512.88958740 512.88958740

24: 511.59384155 511.59387207

25: 511.71841431 511.71841431

26: 512.91870117 512.91870117

27: 505.57281494 505.57281494

28: 500.81240845 500.81240845

29: 500.42144775 500.42144775

30: 504.47088623 504.47088623

31: 505.69696045 505.69696045

32: 495.05331421 495.05328369

33: 512.90972900 512.90966797

34: 507.62371826 507.62374878

35: 505.01126099 505.01129150

36: 505.70776367 505.70776367

37: 500.78982544 500.78982544

38: 518.94940186 518.94940186

39: 513.59509277 513.59509277

40: 504.26684570 504.26681519

41: 506.76119995 506.76116943

42: 506.62884521 506.62884521

43: 503.39712524 503.39712524

44: 507.15069580 507.15069580

45: 508.08422852 508.08425903

46: 499.66766357 499.66769409

47: 502.35678101 502.35678101

48: 494.88311768 494.88311768

49: 499.17807007 499.17807007

**Part 2 – Using Shared Memory**

**GPU time(End-to-End) -** 22.310783ms

**GPU time(MMM) -** 0.126112ms

**Best CPU time -** 49.65s

**TOL-** 0.0001

Using shared memory for the GPU in this case lends to higher end-to-end execution time than using global memory, however the MMM execution time itself is lower for shared memory execution.

The following are the first 50 values of the output for the CPU & GPU execution :

0: 511.96847534 511.96847534

1: 502.86544800 502.86541748

2: 503.09921265 503.09921265

3: 503.61126709 503.61126709

4: 504.44244385 504.44244385

5: 514.62268066 514.62268066

6: 504.18115234 504.18112183

7: 512.17102051 512.17102051

8: 512.47802734 512.47802734

9: 504.64291382 504.64291382

10: 516.60943604 516.60943604

11: 511.41964722 511.41961670

12: 496.94531250 496.94531250

13: 512.66119385 512.66119385

14: 507.61819458 507.61819458

15: 499.52340698 499.52343750

16: 504.20010376 504.20010376

17: 510.20028687 510.20025635

18: 507.44528198 507.44528198

19: 497.04547119 497.04550171

20: 506.25741577 506.25738525

21: 491.61154175 491.61160278

22: 526.00372314 526.00378418

23: 512.88958740 512.88958740

24: 511.59384155 511.59387207

25: 511.71841431 511.71841431

26: 512.91870117 512.91870117

27: 505.57281494 505.57281494

28: 500.81240845 500.81240845

29: 500.42144775 500.42144775

30: 504.47088623 504.47088623

31: 505.69696045 505.69696045

32: 495.05331421 495.05328369

33: 512.90972900 512.90966797

34: 507.62371826 507.62374878

35: 505.01126099 505.01129150

36: 505.70776367 505.70776367

37: 500.78982544 500.78982544

38: 518.94940186 518.94940186

39: 513.59509277 513.59509277

40: 504.26684570 504.26681519

41: 506.76119995 506.76116943

42: 506.62884521 506.62884521

43: 503.39712524 503.39712524

44: 507.15069580 507.15069580

45: 508.08422852 508.08425903

46: 499.66766357 499.66769409

47: 502.35678101 502.35678101

48: 494.88311768 494.88311768

49: 499.17807007 499.17807007

**Part 3- Using Shared Memory & Loop Unrolling**

**GPU time(End-to-End) -** 21.468704ms

**GPU time(MMM) -** 0.133952ms

**Best CPU time -** 48.630000s

**TOL-** 0.0001

As seen with the case of the matrix size of 1024x1024, unrolling in this case also leads to an increased execution time for the MMM from the version in Part 2. However the end-to-end GPU execution time is slightly lower in this case.

The following are the first 50 values of the output for the CPU & GPU execution :

0: 511.96847534 511.96847534

1: 502.86544800 502.86541748

2: 503.09921265 503.09921265

3: 503.61126709 503.61126709

4: 504.44244385 504.44244385

5: 514.62268066 514.62268066

6: 504.18115234 504.18112183

7: 512.17102051 512.17102051

8: 512.47802734 512.47802734

9: 504.64291382 504.64291382

10: 516.60943604 516.60943604

11: 511.41964722 511.41961670

12: 496.94531250 496.94531250

13: 512.66119385 512.66119385

14: 507.61819458 507.61819458

15: 499.52340698 499.52343750

16: 504.20010376 504.20010376

17: 510.20028687 510.20025635

18: 507.44528198 507.44528198

19: 497.04547119 497.04550171

20: 506.25741577 506.25738525

21: 491.61154175 491.61160278

22: 526.00372314 526.00378418

23: 512.88958740 512.88958740

24: 511.59384155 511.59387207

25: 511.71841431 511.71841431

26: 512.91870117 512.91870117

27: 505.57281494 505.57281494

28: 500.81240845 500.81240845

29: 500.42144775 500.42144775

30: 504.47088623 504.47088623

31: 505.69696045 505.69696045

32: 495.05331421 495.05328369

33: 512.90972900 512.90966797

34: 507.62371826 507.62374878

35: 505.01126099 505.01129150

36: 505.70776367 505.70776367

37: 500.78982544 500.78982544

38: 518.94940186 518.94940186

39: 513.59509277 513.59509277

40: 504.26684570 504.26681519

41: 506.76119995 506.76116943

42: 506.62884521 506.62884521

43: 503.39712524 503.39712524

44: 507.15069580 507.15069580

45: 508.08422852 508.08425903

46: 499.66766357 499.66769409

47: 502.35678101 502.35678101

48: 494.88311768 494.88311768

49: 499.17807007 499.17807007