CS610 – Assignment 4 Report

Name: Nishant
Roll No: 251110053

November 16, 2025

1 Problem 1 – 7-Point 3D Stencil

Implementation Versions

1. Naive CUDA kernel
2. Shared-memory tiled kernel with TILE {1,2,4,8}
3. Loop transformations over tiled version
4. Pinned memory version

Profiler Results

```
nishantk25@gpu0:~/Downloads/cuda$ nvprof ./problem1a
Running CPU stencil and naive CUDA stencil kernel...

CPU time = 2.44403 ms
==2504647== NVPROF is profiling process 2504647, command: ./problem1a
Kernel only time = 1.47136 ms
Total GPU time (incl copy) = 3.19328 ms
No differences found between base and test versions
==2504647== Profiling application: ./problem1a
==2504647== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   37.53%  318.50us         1  318.50us  318.50us  318.50us  [CUDA memcpy DtoH]
                   37.52%  318.43us         1  318.43us  318.43us  318.43us  [CUDA memcpy HtoD]
                   24.95%  211.75us         1  211.75us  211.75us  211.75us  gpuStencilKernel(double const *, double*)
      API calls:   96.98%  150.69ms         4  37.673ms     523ns  150.69ms  cudaEventCreate
                    1.10%  1.7075ms         2  853.76us  491.20us  1.2163ms  cudaMemcpy
                    0.91%  1.4173ms         1  1.4173ms  1.4173ms  1.4173ms  cudaLaunchKernel
                    0.50%  778.36us       404  1.9260us     142ns  93.944us  cuDeviceGetAttribute
                    0.21%  319.83us         2  159.91us  120.46us  199.37us  cudaFree
                    0.14%  216.16us         1  216.16us  216.16us  216.16us  cudaDeviceSynchronize
                    0.12%  187.13us         2  93.566us  80.401us  106.73us  cudaMalloc
                    0.01%  22.802us         4  5.7000us  3.7820us  10.431us  cuDeviceGetName
                    0.01%  17.311us         4  4.3270us  2.3360us  8.3730us  cudaEventRecord
                    0.01%  13.293us         4  3.3230us     958ns  9.3740us  cuDeviceGetPCIBusId
                    0.00%  6.1050us         1  6.1050us  6.1050us  6.1050us  cudaEventSynchronize
                    0.00%  3.6510us         2  1.8250us  1.0440us  2.6070us  cudaEventElapsedTime
                    0.00%  1.8710us         8     233ns     151ns     595ns  cuDeviceGet
                    0.00%  1.3280us         4     332ns     237ns     518ns  cuDeviceTotalMem
                    0.00%  1.0050us         4     251ns     187ns     384ns  cuDeviceGetUuid
                    0.00%     980ns         3     326ns     169ns     580ns  cuDeviceGetCount
                    0.00%     406ns         1     406ns     406ns     406ns  cuModuleGetLoadingMode
```

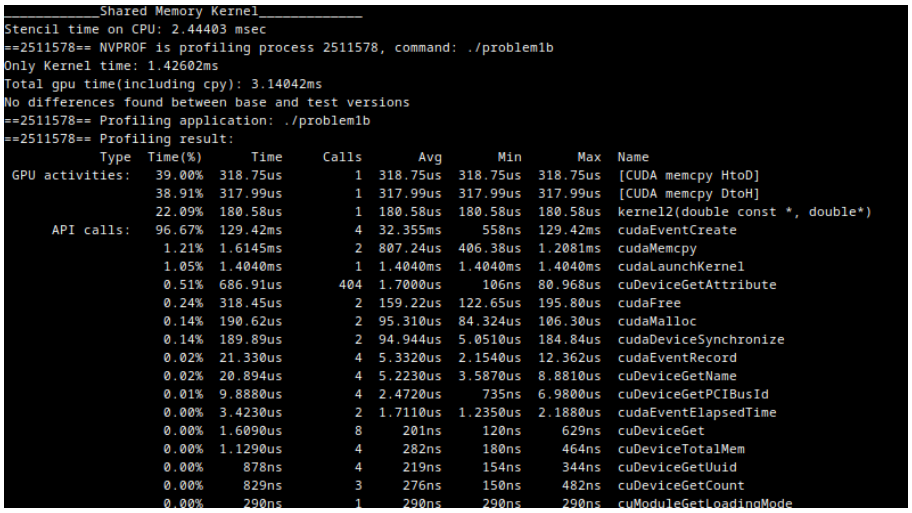
Figure 1: Output of problem1a

```
nishantk25@gpu0:~/Downloads/cuda$ nvprof ./problem1b
___________Shared Memory Kernel____________
Stencil time on CPU: 2.49004 msec
==2505558== NVPROF is profiling process 2505558, command: ./problem1b
Only Kernel time: 2.4625ms
Total gpu time(including cpy): 4.19408ms
No differences found between base and test versions
==2505558== Profiling application: ./problem1b
==2505558== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   71.87%  1.6272ms         1  1.6272ms  1.6272ms  1.6272ms  kernel2(double const *, double*)
                   14.08%  318.75us         1  318.75us  318.75us  318.75us  [CUDA memcpy HtoD]
                   14.05%  317.99us         1  317.99us  317.99us  317.99us  [CUDA memcpy DtoH]
      API calls:   96.25%  142.42ms         4  35.606ms     498ns  142.42ms  cudaEventCreate
                    1.11%  1.6375ms         2  818.73us  5.4020us  1.6321ms  cudaDeviceSynchronize
                    1.10%  1.6325ms         2  816.23us  410.02us  1.2224ms  cudaMemcpy
                    0.67%  989.44us         1  989.44us  989.44us  989.44us  cudaLaunchKernel
                    0.47%  692.58us       404  1.7140us     116ns  85.502us  cuDeviceGetAttribute
                    0.24%  350.95us         2  175.48us  150.22us  200.74us  cudaFree
                    0.13%  185.64us         2  92.820us  79.104us  106.54us  cudaMalloc
                    0.02%  24.660us         4  6.1650us  3.2270us  13.325us  cuDeviceGetName
                    0.01%  21.283us         4  5.3200us  2.2370us  12.448us  cudaEventRecord
                    0.01%  12.471us         4  3.1170us  1.1990us  6.7570us  cuDeviceGetPCIBusId
                    0.00%  3.8200us         2  1.9100us  1.4170us  2.4030us  cudaEventElapsedTime
                    0.00%  2.3580us         8     294ns     126ns     972ns  cuDeviceGet
                    0.00%  1.2900us         3     430ns     170ns     903ns  cuDeviceGetCount
                    0.00%  1.0510us         4     262ns     208ns     401ns  cuDeviceTotalMem
                    0.00%     877ns         4     219ns     173ns     324ns  cuDeviceGetUuid
                    0.00%     294ns         1     294ns     294ns     294ns  cuModuleGetLoadingMode
```

Figure 2: Output of problem1b(tile=1)

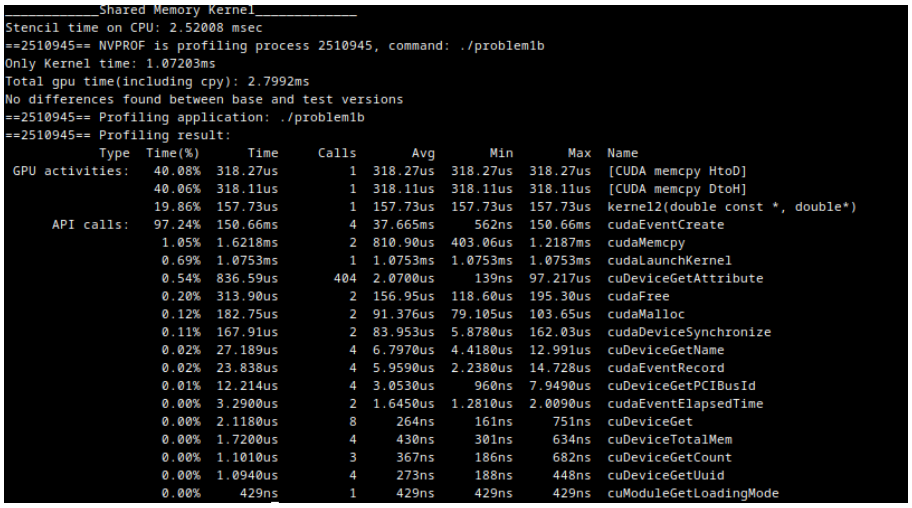
```
___________Shared Memory Kernel____________
Stencil time on CPU: 2.43497 msec
==2512107== NVPROF is profiling process 2512107, command: ./problem1b
Only Kernel time: 1.74861ms
Total gpu time(including cpy): 3.45958ms
No differences found between base and test versions
==2512107== Profiling application: ./problem1b
==2512107== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   43.06%  481.92us         1  481.92us  481.92us  481.92us  kernel2(double const *, double*)
                   28.48%  318.72us         1  318.72us  318.72us  318.72us  [CUDA memcpy DtoH]
                   28.47%  318.59us         1  318.59us  318.59us  318.59us  [CUDA memcpy HtoD]
      API calls:   96.72%  152.49ms         4  38.123ms     549ns  152.49ms  cudaEventCreate
                    1.08%  1.6983ms         2  849.17us  494.35us  1.2040ms  cudaMemcpy
                    0.90%  1.4253ms         1  1.4253ms  1.4253ms  1.4253ms  cudaLaunchKernel
                    0.62%  974.55us       404  2.4120us     195ns  114.65us  cuDeviceGetAttribute
                    0.31%  491.83us         2  245.91us  5.4740us  486.35us  cudaDeviceSynchronize
                    0.20%  315.82us         2  157.91us  119.97us  195.85us  cudaFree
                    0.12%  187.76us         2  93.882us  80.932us  106.83us  cudaMalloc
                    0.02%  31.919us         4  7.9790us  5.0720us  15.455us  cuDeviceGetName
                    0.02%  25.524us         4  6.3810us  2.5100us  15.926us  cudaEventRecord
                    0.01%  12.768us         4  3.1920us  1.3350us  7.6720us  cuDeviceGetPCIBusId
                    0.00%  3.3830us         2  1.6910us  1.0400us  2.3430us  cudaEventElapsedTime
                    0.00%  2.5190us         8     314ns     210ns     832ns  cuDeviceGet
                    0.00%  1.7140us         4     428ns     326ns     724ns  cuDeviceTotalMem
                    0.00%  1.5470us         4     386ns     255ns     772ns  cuDeviceGetUuid
                    0.00%  1.5000us         3     500ns     249ns     891ns  cuDeviceGetCount
                    0.00%  1.2390us         1  1.2390us  1.2390us  1.2390us  cuModuleGetLoadingMode
```

Figure 3: Output of problem1b(tile=2)



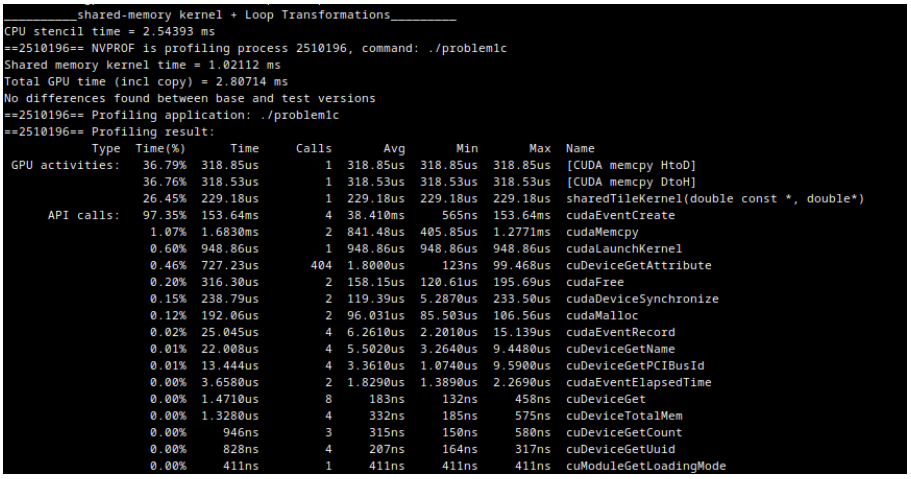
```
___________Shared Memory Kernel____________
Stencil time on CPU: 2.44403 msec
==2511578== NVPROF is profiling process 2511578, command: ./problem1b
Only Kernel time: 1.42602ms
Total gpu time(including cpy): 3.14042ms
No differences found between base and test versions
==2511578== Profiling application: ./problem1b
==2511578== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   39.00%  318.75us         1  318.75us  318.75us  318.75us  [CUDA memcpy HtoD]
                   38.91%  317.99us         1  317.99us  317.99us  317.99us  [CUDA memcpy DtoH]
                   22.09%  180.58us         1  180.58us  180.58us  180.58us  kernel2(double const *, double*)
      API calls:   96.67%  129.42ms         4  32.355ms     558ns  129.42ms  cudaEventCreate
                    1.21%  1.6145ms         2  807.24us  406.38us  1.2081ms  cudaMemcpy
                    1.05%  1.4040ms         1  1.4040ms  1.4040ms  1.4040ms  cudaLaunchKernel
                    0.51%  686.91us       404  1.7000us     106ns  80.968us  cuDeviceGetAttribute
                    0.24%  318.45us         2  159.22us  122.65us  195.80us  cudaFree
                    0.14%  190.62us         2  95.310us  84.324us  106.30us  cudaMalloc
                    0.14%  189.89us         2  94.944us  5.0510us  184.84us  cudaDeviceSynchronize
                    0.02%  21.330us         4  5.3320us  2.1540us  12.362us  cudaEventRecord
                    0.02%  20.894us         4  5.2230us  3.5870us  8.8810us  cuDeviceGetName
                    0.01%  9.8880us         4  2.4720us     735ns  6.9800us  cuDeviceGetPCIBusId
                    0.00%  3.4230us         2  1.7110us  1.2350us  2.1880us  cudaEventElapsedTime
                    0.00%  1.6090us         8     201ns     120ns     629ns  cuDeviceGet
                    0.00%  1.1290us         4     282ns     180ns     464ns  cuDeviceTotalMem
                    0.00%     878ns         4     219ns     154ns     344ns  cuDeviceGetUuid
                    0.00%     829ns         3     276ns     150ns     482ns  cuDeviceGetCount
                    0.00%     290ns         1     290ns     290ns     290ns  cuModuleGetLoadingMode
```

Figure 4: Output of problem1b(tile=4)



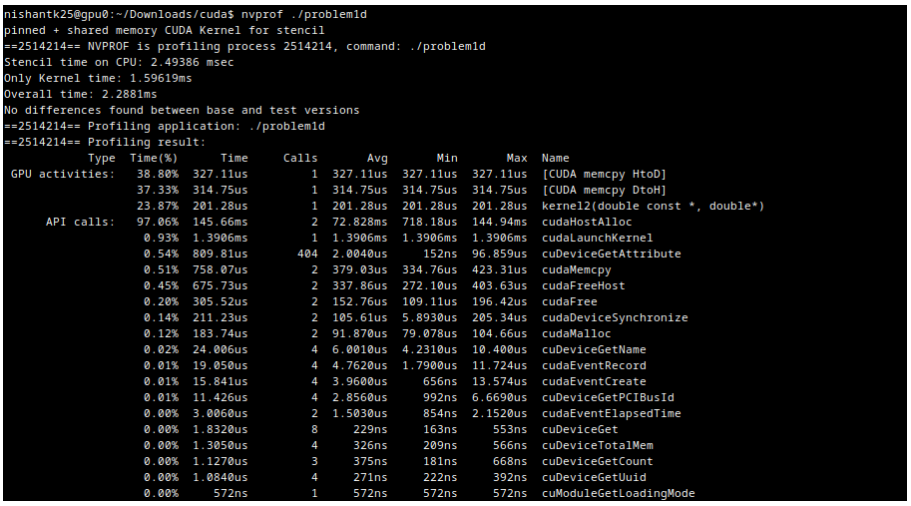
```
___________Shared Memory Kernel____________
Stencil time on CPU: 2.52008 msec
==2510945== NVPROF is profiling process 2510945, command: ./problem1b
Only Kernel time: 1.07203ms
Total gpu time(including cpy): 2.7992ms
No differences found between base and test versions
==2510945== Profiling application: ./problem1b
==2510945== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   40.08%  318.27us         1  318.27us  318.27us  318.27us  [CUDA memcpy HtoD]
                   40.06%  318.11us         1  318.11us  318.11us  318.11us  [CUDA memcpy DtoH]
                   19.86%  157.73us         1  157.73us  157.73us  157.73us  kernel2(double const *, double*)
      API calls:   97.24%  150.66ms         4  37.665ms     562ns  150.66ms  cudaEventCreate
                    1.05%  1.6218ms         2  810.90us  403.06us  1.2187ms  cudaMemcpy
                    0.69%  1.0753ms         1  1.0753ms  1.0753ms  1.0753ms  cudaLaunchKernel
                    0.54%  836.59us       404  2.0700us     139ns  97.217us  cuDeviceGetAttribute
                    0.20%  313.90us         2  156.95us  118.60us  195.30us  cudaFree
                    0.12%  182.75us         2  91.376us  79.105us  103.65us  cudaMalloc
                    0.11%  167.91us         2  83.953us  5.8780us  162.03us  cudaDeviceSynchronize
                    0.02%  27.189us         4  6.7970us  4.4180us  12.991us  cuDeviceGetName
                    0.02%  23.838us         4  5.9590us  2.2380us  14.728us  cudaEventRecord
                    0.01%  12.214us         4  3.0530us     960ns  7.9490us  cuDeviceGetPCIBusId
                    0.00%  3.2900us         2  1.6450us  1.2810us  2.0090us  cudaEventElapsedTime
                    0.00%  2.1180us         8     264ns     161ns     751ns  cuDeviceGet
                    0.00%  1.7200us         4     430ns     301ns     634ns  cuDeviceTotalMem
                    0.00%  1.1010us         3     367ns     186ns     682ns  cuDeviceGetCount
                    0.00%  1.0940us         4     273ns     188ns     448ns  cuDeviceGetUuid
                    0.00%     429ns         1     429ns     429ns     429ns  cuModuleGetLoadingMode
```

Figure 5: Output of problem1b(tile=8)



```
_________shared-memory kernel + Loop Transformations________
CPU stencil time = 2.54393 ms
==2510196== NVPROF is profiling process 2510196, command: ./problem1c
Shared memory kernel time = 1.02112 ms
Total GPU time (incl copy) = 2.80714 ms
No differences found between base and test versions
==2510196== Profiling application: ./problem1c
==2510196== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   36.79%  318.85us         1  318.85us  318.85us  318.85us  [CUDA memcpy HtoD]
                   36.76%  318.53us         1  318.53us  318.53us  318.53us  [CUDA memcpy DtoH]
                   26.45%  229.18us         1  229.18us  229.18us  229.18us  sharedTileKernel(double const *, double*)
      API calls:   97.35%  153.64ms         4  38.410ms     565ns  153.64ms  cudaEventCreate
                    1.07%  1.6830ms         2  841.48us  405.85us  1.2771ms  cudaMemcpy
                    0.60%  948.86us         1  948.86us  948.86us  948.86us  cudaLaunchKernel
                    0.46%  727.23us       404  1.8000us     123ns  99.468us  cuDeviceGetAttribute
                    0.20%  316.30us         2  158.15us  120.61us  195.69us  cudaFree
                    0.15%  238.79us         2  119.39us  5.2870us  233.50us  cudaDeviceSynchronize
                    0.12%  192.06us         2  96.031us  85.503us  106.56us  cudaMalloc
                    0.02%  25.045us         4  6.2610us  2.2010us  15.139us  cudaEventRecord
                    0.01%  22.008us         4  5.5020us  3.2640us  9.4480us  cuDeviceGetName
                    0.01%  13.444us         4  3.3610us  1.0740us  9.5900us  cuDeviceGetPCIBusId
                    0.00%  3.6580us         2  1.8290us  1.3890us  2.2690us  cudaEventElapsedTime
                    0.00%  1.4710us         8     183ns     132ns     458ns  cuDeviceGet
                    0.00%  1.3280us         4     332ns     185ns     575ns  cuDeviceTotalMem
                    0.00%     946ns         3     315ns     150ns     580ns  cuDeviceGetCount
                    0.00%     828ns         4     207ns     164ns     317ns  cuDeviceGetUuid
                    0.00%     411ns         1     411ns     411ns     411ns  cuModuleGetLoadingMode
```

Figure 6: Output of problem1c



```
nishantk25@gpu0:~/Downloads/cuda$ nvprof ./problem1d
pinned + shared memory CUDA Kernel for stencil
==2514214== NVPROF is profiling process 2514214, command: ./problem1d
Stencil time on CPU: 2.49386 msec
Only Kernel time: 1.59619ms
Overall time: 2.2881ms
No differences found between base and test versions
==2514214== Profiling application: ./problem1d
==2514214== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   38.80%  327.11us         1  327.11us  327.11us  327.11us  [CUDA memcpy HtoD]
                   37.33%  314.75us         1  314.75us  314.75us  314.75us  [CUDA memcpy DtoH]
                   23.87%  201.28us         1  201.28us  201.28us  201.28us  kernel2(double const *, double*)
      API calls:   97.06%  145.66ms         2  72.828ms  718.18us  144.94ms  cudaHostAlloc
                    0.93%  1.3906ms         1  1.3906ms  1.3906ms  1.3906ms  cudaLaunchKernel
                    0.54%  809.81us       404  2.0040us     152ns  96.859us  cuDeviceGetAttribute
                    0.51%  758.07us         2  379.03us  334.76us  423.31us  cudaMemcpy
                    0.45%  675.73us         2  337.86us  272.10us  403.63us  cudaFreeHost
                    0.20%  305.52us         2  152.76us  109.11us  196.42us  cudaFree
                    0.14%  211.23us         2  105.61us  5.8930us  205.34us  cudaDeviceSynchronize
                    0.12%  183.74us         2  91.870us  79.078us  104.66us  cudaMalloc
                    0.02%  24.006us         4  6.0010us  4.2310us  10.400us  cuDeviceGetName
                    0.01%  19.050us         4  4.7620us  1.7900us  11.724us  cudaEventRecord
                    0.01%  15.841us         4  3.9600us     656ns  13.574us  cudaEventCreate
                    0.01%  11.426us         4  2.8560us     992ns  6.6690us  cuDeviceGetPCIBusId
                    0.00%  3.0060us         2  1.5030us     854ns  2.1520us  cudaEventElapsedTime
                    0.00%  1.8320us         8     229ns     163ns     553ns  cuDeviceGet
                    0.00%  1.3050us         4     326ns     209ns     566ns  cuDeviceTotalMem
                    0.00%  1.1270us         3     375ns     181ns     668ns  cuDeviceGetCount
                    0.00%  1.0840us         4     271ns     222ns     392ns  cuDeviceGetUuid
                    0.00%     572ns         1     572ns     572ns     572ns  cuModuleGetLoadingMode
```

Figure 7: Output of problem1d

Observations

- Increasing tile size reduces kernel time until shared memory limits are reached.
- TILE=1 performs poorly due to overhead without reuse.
- Pinned memory significantly reduces memcpy overhead.

Compilation commands

```
make problem1
./problem1a.out
```

```
./problem1b.out
./problem1c.out
./problem1d.out
```

2 Problem 2 – Prefix Sum

Versions

1. Copy-then-execute
2. UVM-based version using `cudaMallocManaged`

```
nishantk25@gpu3:~/Downloads/cuda$ ./problem2
CPU last element = 4194304
Chunks: 1
Total time (copy then execute): 35.7183 ms
No differences found between base and test versions
________UVM kernel________
Total time (uvm): 0.893856
No differences found between base and test versions
UVM prefix sum completed.
```

Figure 8: Output of program for $N = 1 \ll 22$

```
nishantk25@gpu3:~/Downloads/cuda$ ./problem2
CPU last element = 2147483648
Chunks: 6
Total time (copy then execute): 17515.5 ms
No differences found between base and test versions
________UVM kernel________
Total time (uvm): 148.341
No differences found between base and test versions
UVM prefix sum completed.
```

Figure 9: Output of program for $N = 1 \ll 31$

Compilation commands

```
make problem2
./problem2.out
```

3 Problem 3 – 10D Loop Nest

All versions produce the same correct output: **Result pnts = 11608**

Kernel Versions

- Baseline CUDA kernel (correctness focused)
- Optimized kernel (Shared Memory Tiling, Loop Unrolling)
- UVM-based version (prefetch, memory advice)
- Thrust-based version

Execution Instructions

The directory `problem3-dir` contains all source files along with the auxiliary files `disp.txt` and `grid.txt`. All executable files for this problem can be generated using `make`:

```
make problem3
```

This command compiles the CPU baseline and all four CUDA implementations, producing the following executables:

- `problem3-v0.out` – CPU baseline version
- `problem3a.out`
- `problem3b.out`
- `problem3c.out`
- `problem3d.out`

Once compiled, each version can be executed individually, for example:

```
./problem3a.out
```

Each executable produces an output text file of the form:

`results-v0.txt`, `results-va.txt`, `results-vb.txt`, `results-vc.txt`, `results-vd.txt`

A diff check is performed to ensure that all CUDA implementations match the baseline CPU output exactly.

Implementation

All versions process the large 10D iteration space using chunking to avoid excessive memory use. For each chunk, the kernel stores the linearized iteration indices of all valid points in the first `n` positions of the output buffer.

Performance

Version	Time (s)	Notes
CPU	–	baseline
Naive CUDA	125.2	
Optimized CUDA	100.3	
UVM	99.2	
Thrust	121.1	

Compilation commands

```
make problem3
./problem3a.out
./problem3b.out
./problem3c.out
./prloblem3d.out
```

4 Problem 4 – 2D and 3D Convolution

Kernel Versions

1. Basic global memory version
2. Optimized version:
    - Shared memory tiling
    - Constant memory for filter
    - Loop unrolling

```
nishantk25@gpu3:~/Downloads/cuda$ ./problem4.out
__________2D Convolution__________
CPU convolution time: 0.0650883 ms
Basic Kernel time: 2.39821ms
Basic Kernel time including memory transfers: 4.82426ms
No differences between gpu and cpu results.
Optimized Kernel time: 0.01024ms
Optimized Kernel time including memory transfers: 2.35363ms
No differences between gpu and cpu results.
__________3D Convolution__________
CPU convolution time: 10.9 ms
Basic Kernel time: 0.261664ms
Basic Kernel time including memory transfers: 1.03843ms
No differences between gpu and cpu results.
Optimized Kernel time: 1.34461ms
Optimized Kernel time including memory transfers: 2.07398ms
No differences between gpu and cpu results.
```

Figure 10: Output of program for N=64

Performance

Table 1: Performance Results for Problem 4 (N=64, FILTER-SIZE=3 , TILE=8)

Kernel	CPU Time (ms)	Kernel Time (ms)	Kernel + Transfer (ms)
2D Basic	0.27	1.44	1.525
2D Optimized	0.27	0.026	0.066
3D Basic	11.32	0.26	0.931
3D Optimized	11.32	0.21	0.965

Compilation commands

```
make problem4
./problem4.out
```