CUDA Streams and Events

Module 6

10/04/2025

Annabel Lin

For this assignment, I decided to have my kernels do image operations, specifically RGB→grayscale conversion and grayscale edge detection.

- RGB→grayscale kernel:
  - Launches first
  - Uses totalThreads and blockSize to calculate numBlocks
- Grayscale edge detection kernel:
  - Launches, dependent on RGB→grayscale stream
  - Uses blockSize and image size (not totalThreads) to calculate block and grid dims

Timing:

totalThreads varying vs grayscale kernel time (only grayscale kernel uses totalThreads)

totalThreads (numBlocks=32)	Grayscale kernel time
512	2.653 ms
1024	1.410 ms
2048	0.791 ms
4096	0.906 ms
10240	0.469 ms
16384	0.366 ms
32768	0.939 ms
65536	1.244 ms

There seemed to be a sweet spot around 16384 totalThreads for the grayscale kernel. This could be due to the grid and stride calculations being optimized for this image size and num threads.

blockSize varying vs edge kernel time (edge kernel only uses totalThreads)

blockSize	Edge detection kernel time
1	1.162 ms
16	0.952 ms
32	0.938 ms
64	0.072 ms
128	0.070 ms

256	0.063 ms
512	0.066 ms
1024	0.072 ms

The improvement of using a larger blockSize seemed to plateau out as the blockSize increased (image size stays the same). This could be because the grid/blocks became pretty optimal after a certain point.

I also noticed that the greyscale kernel time + edge kernel time < total pipeline time. I think that this is because currently, the edge kernel is waiting for the greyscale kernel to finish, since it needs the greyscale as input.

To improve the use of the stream, I'd need to organize the kernels a different way, or split the input into sections to operate on so that the streams can work concurrently.

Final Project:

For my final project, I am thinking of applying image processing to a video. For example given a set of frames, process the frames with a series of operations. This can be gaussian blurring, sharpening, edge detection, or skeletonization.

The stream/event workflow in this assignment and the image processing chain can be utilized in the final project. I mentioned that these streams did not end up running concurrently because of data dependency – but since the final project will process multiple images, I can organize the code to process the images concurrently.

The code may not be 1-to-1 if I decide to expand the project with more advanced processing features and try to optimize the streams, but this framework for grid/block calculation can be reused. The event dependencies are also helpful – I think that the final project will involve this.

Resources:

Weighted grayscale: https://www.dynamsoft.com/blog/insights/image-processing/image-processing-101-color-space-conversion/#:~:text=Eyes%20are%20most%20sensitive%20to,HSV%20to%20RGB%20conversion

Edge detection: https://www.geeksforgeeks.org/computer-vision/comprehensive-guide-to-edge-detection-algorithms/#1-sobel-operator

Image writing: https://github.com/nothings/stb/blob/master/stb_image_write.h

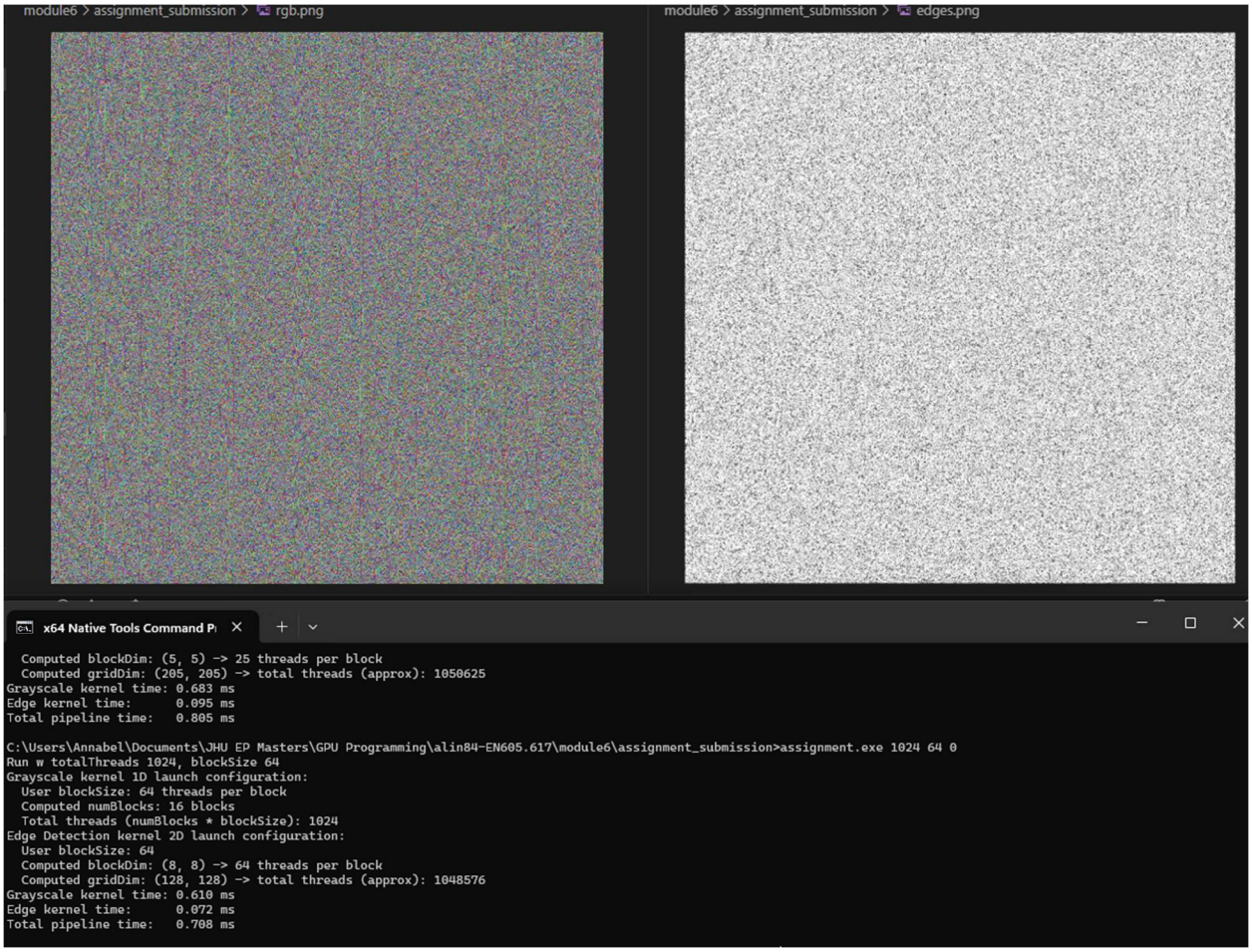
Run artifacts:

```
C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 512 32
Run w totalThreads 512, blockSize 32
Grayscale kernel 1D launch configuration:
  User blockSize: 32 threads per block
  Computed numBlocks: 16 blocks
  Total threads (numBlocks * blockSize): 512
Edge Detection kernel 2D launch configuration:
  User blockSize: 32
  Computed blockDim: (5, 5) -> 25 threads per block
  Computed gridDim: (205, 205) -> total threads (approx): 1050625
Grayscale kernel time: 2.653 ms
Edge kernel time:      0.091 ms
Total pipeline time:   2.768 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 32
Run w totalThreads 1024, blockSize 32
Grayscale kernel 1D launch configuration:
  User blockSize: 32 threads per block
  Computed numBlocks: 32 blocks
  Total threads (numBlocks * blockSize): 1024
Edge Detection kernel 2D launch configuration:
  User blockSize: 32
  Computed blockDim: (5, 5) -> 25 threads per block
  Computed gridDim: (205, 205) -> total threads (approx): 1050625
Grayscale kernel time: 1.410 ms
Edge kernel time:      0.282 ms
Total pipeline time:   2.481 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 2048 32
Run w totalThreads 2048, blockSize 32
Grayscale kernel 1D launch configuration:
  User blockSize: 32 threads per block
  Computed numBlocks: 64 blocks
  Total threads (numBlocks * blockSize): 2048
Edge Detection kernel 2D launch configuration:
  User blockSize: 32
  Computed blockDim: (5, 5) -> 25 threads per block
  Computed gridDim: (205, 205) -> total threads (approx): 1050625
Grayscale kernel time: 0.791 ms
Edge kernel time:      0.994 ms
Total pipeline time:   2.576 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 4096 32
Run w totalThreads 4096, blockSize 32
Grayscale kernel 1D launch configuration:
  User blockSize: 32 threads per block
  Computed numBlocks: 128 blocks
  Total threads (numBlocks * blockSize): 4096
Edge Detection kernel 2D launch configuration:
  User blockSize: 32
  Computed blockDim: (5, 5) -> 25 threads per block
  Computed gridDim: (205, 205) -> total threads (approx): 1050625
Grayscale kernel time: 0.906 ms
Edge kernel time:      0.092 ms
Total pipeline time:   1.024 ms
```

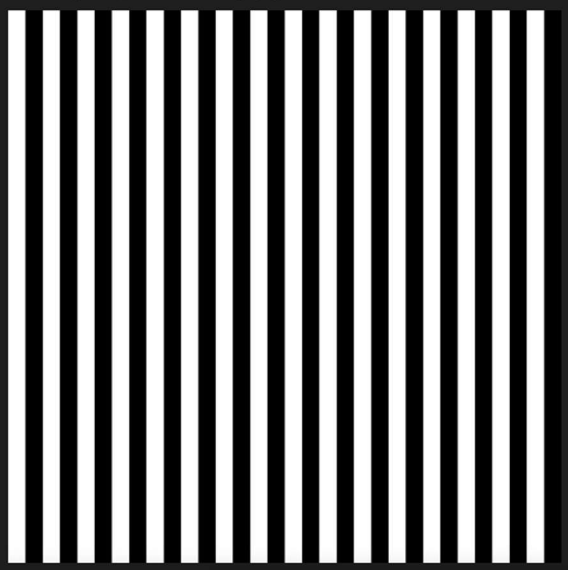
```
C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 10240 32
Run w totalThreads 10240, blockSize 32
Grayscale kernel 1D launch configuration:
  User blockSize: 32 threads per block
  Computed numBlocks: 320 blocks
  Total threads (numBlocks * blockSize): 10240
Edge Detection kernel 2D launch configuration:
  User blockSize: 32
  Computed blockDim: (5, 5) -> 25 threads per block
  Computed gridDim: (205, 205) -> total threads (approx): 1050625
Grayscale kernel time: 0.469 ms
Edge kernel time:      0.078 ms
Total pipeline time:   0.646 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 16384 32
Run w totalThreads 16384, blockSize 32
Grayscale kernel 1D launch configuration:
  User blockSize: 32 threads per block
  Computed numBlocks: 512 blocks
  Total threads (numBlocks * blockSize): 16384
Edge Detection kernel 2D launch configuration:
  User blockSize: 32
  Computed blockDim: (5, 5) -> 25 threads per block
  Computed gridDim: (205, 205) -> total threads (approx): 1050625
Grayscale kernel time: 0.366 ms
Edge kernel time:      0.096 ms
Total pipeline time:   0.524 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 32768 32
Run w totalThreads 32768, blockSize 32
Grayscale kernel 1D launch configuration:
  User blockSize: 32 threads per block
  Computed numBlocks: 1024 blocks
  Total threads (numBlocks * blockSize): 32768
Edge Detection kernel 2D launch configuration:
  User blockSize: 32
  Computed blockDim: (5, 5) -> 25 threads per block
  Computed gridDim: (205, 205) -> total threads (approx): 1050625
Grayscale kernel time: 0.939 ms
Edge kernel time:      0.079 ms
Total pipeline time:   1.091 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 65536 32
Run w totalThreads 65536, blockSize 32
Grayscale kernel 1D launch configuration:
  User blockSize: 32 threads per block
  Computed numBlocks: 2048 blocks
  Total threads (numBlocks * blockSize): 65536
Edge Detection kernel 2D launch configuration:
  User blockSize: 32
  Computed blockDim: (5, 5) -> 25 threads per block
  Computed gridDim: (205, 205) -> total threads (approx): 1050625
Grayscale kernel time: 1.244 ms
Edge kernel time:      0.090 ms
Total pipeline time:   1.359 ms
```

```
C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 1
Run w totalThreads 1024, blockSize 1
Grayscale kernel 1D launch configuration:
  User blockSize: 1 threads per block
  Computed numBlocks: 1024 blocks
  Total threads (numBlocks * blockSize): 1024
Edge Detection kernel 2D launch configuration:
  User blockSize: 1
  Computed blockDim: (1, 1) -> 1 threads per block
  Computed gridDim: (1024, 1024) -> total threads (approx): 1048576
Grayscale kernel time: 1.409 ms
Edge kernel time:      1.162 ms
Total pipeline time:   2.596 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 16
Run w totalThreads 1024, blockSize 16
Grayscale kernel 1D launch configuration:
  User blockSize: 16 threads per block
  Computed numBlocks: 64 blocks
  Total threads (numBlocks * blockSize): 1024
Edge Detection kernel 2D launch configuration:
  User blockSize: 16
  Computed blockDim: (4, 4) -> 16 threads per block
  Computed gridDim: (256, 256) -> total threads (approx): 1048576
Grayscale kernel time: 0.680 ms
Edge kernel time:      0.952 ms
Total pipeline time:   2.442 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 32
Run w totalThreads 1024, blockSize 32
Grayscale kernel 1D launch configuration:
  User blockSize: 32 threads per block
  Computed numBlocks: 32 blocks
  Total threads (numBlocks * blockSize): 1024
Edge Detection kernel 2D launch configuration:
  User blockSize: 32
  Computed blockDim: (5, 5) -> 25 threads per block
  Computed gridDim: (205, 205) -> total threads (approx): 1050625
Grayscale kernel time: 0.691 ms
Edge kernel time:      0.938 ms
Total pipeline time:   2.268 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 64
Run w totalThreads 1024, blockSize 64
Grayscale kernel 1D launch configuration:
  User blockSize: 64 threads per block
  Computed numBlocks: 16 blocks
  Total threads (numBlocks * blockSize): 1024
Edge Detection kernel 2D launch configuration:
  User blockSize: 64
  Computed blockDim: (8, 8) -> 64 threads per block
  Computed gridDim: (128, 128) -> total threads (approx): 1048576
Grayscale kernel time: 1.355 ms
Edge kernel time:      0.072 ms
Total pipeline time:   1.452 ms
```

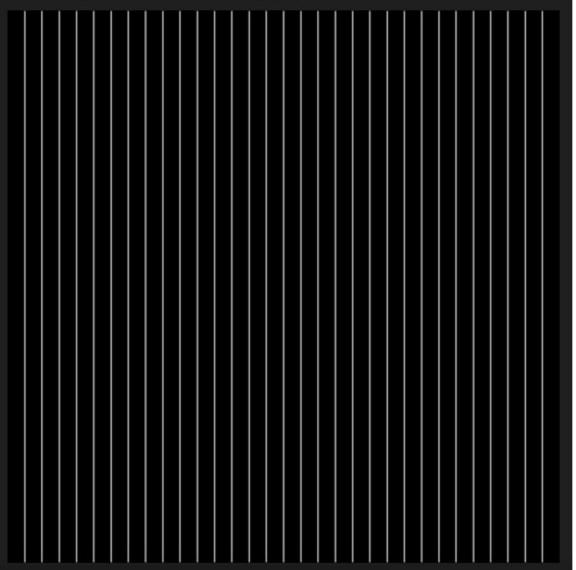
```
C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 128
Run w totalThreads 1024, blockSize 128
Grayscale kernel 1D launch configuration:
  User blockSize: 128 threads per block
  Computed numBlocks: 8 blocks
  Total threads (numBlocks * blockSize): 1024
Edge Detection kernel 2D launch configuration:
  User blockSize: 128
  Computed blockDim: (11, 11) -> 121 threads per block
  Computed gridDim: (94, 94) -> total threads (approx): 1069156
Grayscale kernel time: 1.263 ms
Edge kernel time:      0.070 ms
Total pipeline time:   1.358 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 256
Run w totalThreads 1024, blockSize 256
Grayscale kernel 1D launch configuration:
  User blockSize: 256 threads per block
  Computed numBlocks: 4 blocks
  Total threads (numBlocks * blockSize): 1024
Edge Detection kernel 2D launch configuration:
  User blockSize: 256
  Computed blockDim: (16, 16) -> 256 threads per block
  Computed gridDim: (64, 64) -> total threads (approx): 1048576
Grayscale kernel time: 0.669 ms
Edge kernel time:      0.063 ms
Total pipeline time:   0.758 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 512
Run w totalThreads 1024, blockSize 512
Grayscale kernel 1D launch configuration:
  User blockSize: 512 threads per block
  Computed numBlocks: 2 blocks
  Total threads (numBlocks * blockSize): 1024
Edge Detection kernel 2D launch configuration:
  User blockSize: 512
  Computed blockDim: (22, 22) -> 484 threads per block
  Computed gridDim: (47, 47) -> total threads (approx): 1069156
Grayscale kernel time: 0.410 ms
Edge kernel time:      0.066 ms
Total pipeline time:   0.500 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 1024
Run w totalThreads 1024, blockSize 1024
Grayscale kernel 1D launch configuration:
  User blockSize: 1024 threads per block
  Computed numBlocks: 1 blocks
  Total threads (numBlocks * blockSize): 1024
Edge Detection kernel 2D launch configuration:
  User blockSize: 1024
  Computed blockDim: (32, 32) -> 1024 threads per block
  Computed gridDim: (32, 32) -> total threads (approx): 1048576
Grayscale kernel time: 0.580 ms
Edge kernel time:      0.072 ms
Total pipeline time:   0.676 ms
```


module6 > assignment_submission > rgb.png
module6 > assignment_submission > edges.png
x64 Native Tools Command P
Computed blockDim: (5, 5) -> 25 threads per block
Computed gridDim: (205, 205) -> total threads (approx): 1050625
Grayscale kernel time: 0.683 ms
Edge kernel time: 0.095 ms
Total pipeline time: 0.805 ms
C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 64 0
Run w totalThreads 1024, blockSize 64
Grayscale kernel 1D launch configuration:
User blockSize: 64 threads per block
Computed numBlocks: 16 blocks
Total threads (numBlocks * blockSize): 1024
Edge Detection kernel 2D launch configuration:
User blockSize: 64
Computed blockDim: (8, 8) -> 64 threads per block
Computed gridDim: (128, 128) -> total threads (approx): 1048576
Grayscale kernel time: 0.610 ms
Edge kernel time: 0.072 ms
Total pipeline time: 0.708 ms

module6 > assignment_submission > rgb.png



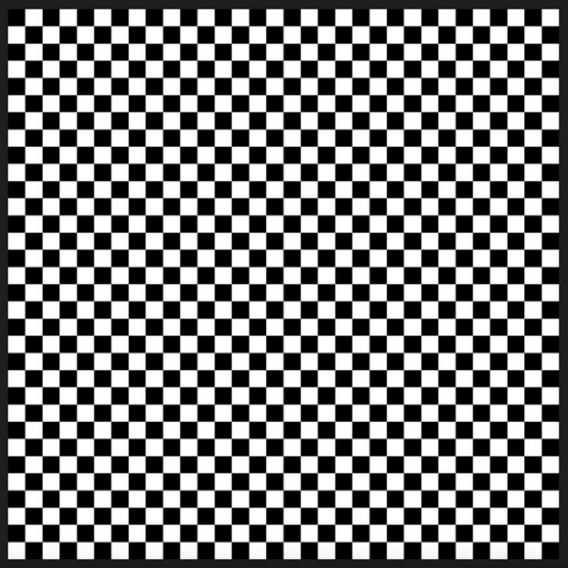
module6 > assignment_submission > edges.png



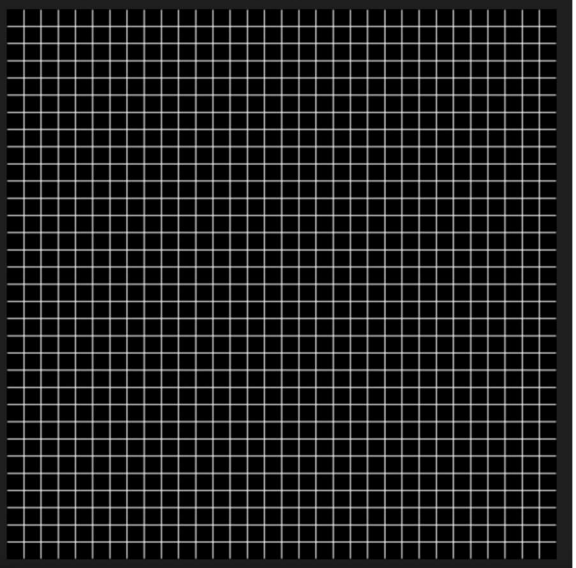
x64 Native Tools Command P

```
  Computed blockDim: (8, 8) -> 64 threads per block
  Computed gridDim: (128, 128) -> total threads (approx): 1048576
Grayscale kernel time: 0.610 ms
Edge kernel time:      0.072 ms
Total pipeline time:   0.708 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 64 1
Run w totalThreads 1024, blockSize 64
Grayscale kernel 1D launch configuration:
  User blockSize: 64 threads per block
  Computed numBlocks: 16 blocks
  Total threads (numBlocks * blockSize): 1024
Edge Detection kernel 2D launch configuration:
  User blockSize: 64
  Computed blockDim: (8, 8) -> 64 threads per block
  Computed gridDim: (128, 128) -> total threads (approx): 1048576
Grayscale kernel time: 0.670 ms
Edge kernel time:      0.071 ms
Total pipeline time:   0.765 ms
```

module6 > assignment_submission > rgb.png

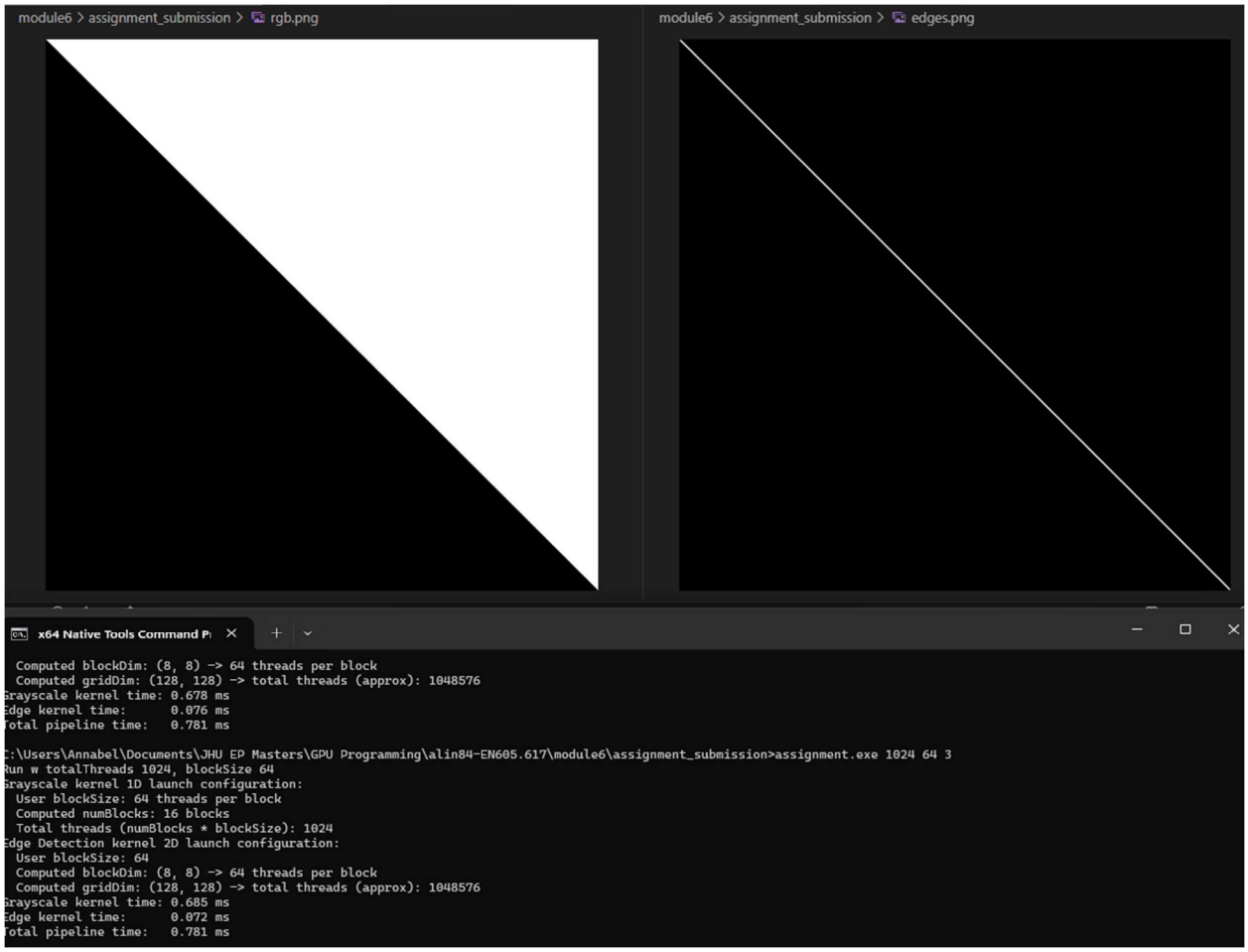


module6 > assignment_submission > edges.png



x64 Native Tools Command P

```
  Computed blockDim: (8, 8) -> 64 threads per block
  Computed gridDim: (128, 128) -> total threads (approx): 1048576
Grayscale kernel time: 0.670 ms
Edge kernel time:      0.071 ms
Total pipeline time:   0.765 ms

C:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 64 2
Run w totalThreads 1024, blockSize 64
Grayscale kernel 1D launch configuration:
  User blockSize: 64 threads per block
  Computed numBlocks: 16 blocks
  Total threads (numBlocks * blockSize): 1024
Edge Detection kernel 2D launch configuration:
  User blockSize: 64
  Computed blockDim: (8, 8) -> 64 threads per block
  Computed gridDim: (128, 128) -> total threads (approx): 1048576
Grayscale kernel time: 0.678 ms
Edge kernel time:      0.076 ms
Total pipeline time:   0.781 ms
```


module6 > assignment_submission > rgb.png
module6 > assignment_submission > edges.png
x64 Native Tools Command Pr
Computed blockDim: (8, 8) -> 64 threads per block
Computed gridDim: (128, 128) -> total threads (approx): 1048576
rayscale kernel time: 0.678 ms
dge kernel time: 0.076 ms
otal pipeline time: 0.781 ms
:\Users\Annabel\Documents\JHU EP Masters\GPU Programming\alin84-EN605.617\module6\assignment_submission>assignment.exe 1024 64 3
un w totalThreads 1024, blockSize 64
rayscale kernel 1D launch configuration:
User blockSize: 64 threads per block
Computed numBlocks: 16 blocks
Total threads (numBlocks * blockSize): 1024
dge Detection kernel 2D launch configuration:
User blockSize: 64
Computed blockDim: (8, 8) -> 64 threads per block
Computed gridDim: (128, 128) -> total threads (approx): 1048576
rayscale kernel time: 0.685 ms
dge kernel time: 0.072 ms
otal pipeline time: 0.781 ms