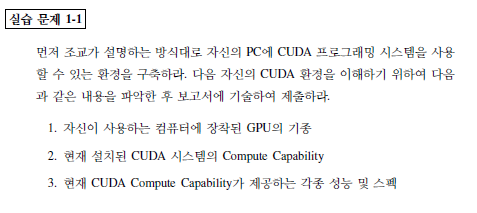
『CUDA 프로그래밍의 기초 1』

GPU 기반의 Manycore Processing의 기초

고급소프트웨어실습 3분반

20151591 이지현

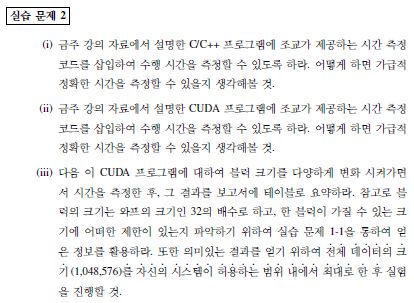


1. NVIDIA GeForce GTX 635
2. 3.5

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | **Attribute** | **GeForce GT 635 Computer Capability: 3.5 Driver Model: WDDM** |  |
|  | **Name** | **GeForce GT 635** |  |
|  | Driver | WDDM |  |
|  | DeviceIndex | 0 |  |
|  | GPU Family | GK208 |  |
|  | RmGpuId | 256 |  |
|  | Compute Major | 3 |  |
|  | Compute Minor | 5 |  |
|  | MAX\_THREADS\_PER\_BLOCK | 1024 |  |
|  | MAX\_BLOCK\_DIM\_X | 1024 |  |
|  | MAX\_BLOCK\_DIM\_Y | 1024 |  |
|  | MAX\_BLOCK\_DIM\_Z | 64 |  |
|  | MAX\_GRID\_DIM\_X | 2147483647 |  |
|  | MAX\_GRID\_DIM\_Y | 65535 |  |
|  | MAX\_GRID\_DIM\_Z | 65535 |  |
|  | MAX\_SHARED\_MEMORY\_PER\_BLOCK | 49152 |  |
|  | TOTAL\_CONSTANT\_MEMORY | 65536 |  |
|  | WARP\_SIZE | 32 |  |
|  | MAX\_PITCH | 2147483647 |  |
|  | MAX\_REGISTERS\_PER\_BLOCK | 65536 |  |
|  | CLOCK\_RATE | 967000 |  |
|  | TEXTURE\_ALIGNMENT | 512 |  |
|  | GPU\_OVERLAP | 1 |  |
|  | MULTIPROCESSOR\_COUNT | 2 |  |
|  | KERNEL\_EXEC\_TIMEOUT | 1 |  |
|  | INTEGRATED | 0 |  |
|  | CAN\_MAP\_HOST\_MEMORY | 1 |  |
|  | COMPUTE\_MODE | 0 |  |
|  | MAXIMUM\_TEXTURE1D\_WIDTH | 65536 |  |
|  | MAXIMUM\_TEXTURE2D\_WIDTH | 65536 |  |
|  | MAXIMUM\_TEXTURE2D\_HEIGHT | 65536 |  |
|  | MAXIMUM\_TEXTURE3D\_WIDTH | 4096 |  |
|  | MAXIMUM\_TEXTURE3D\_HEIGHT | 4096 |  |
|  | MAXIMUM\_TEXTURE3D\_DEPTH | 4096 |  |
|  | MAXIMUM\_TEXTURE2D\_LAYERED\_WIDTH | 16384 |  |
|  | MAXIMUM\_TEXTURE2D\_LAYERED\_HEIGHT | 16384 |  |
|  | MAXIMUM\_TEXTURE2D\_LAYERED\_LAYERS | 2048 |  |
|  | SURFACE\_ALIGNMENT | 512 |  |
|  | CONCURRENT\_KERNELS | 1 |  |
|  | ECC\_ENABLED | 0 |  |
|  | PCI\_BUS\_ID | 1 |  |
|  | PCI\_DEVICE\_ID | 0 |  |
|  | TCC\_DRIVER | 0 |  |
|  | MEMORY\_CLOCK\_RATE | 900000 |  |
|  | GLOBAL\_MEMORY\_BUS\_WIDTH | 64 |  |
|  | L2\_CACHE\_SIZE | 524288 |  |
|  | MAX\_THREADS\_PER\_MULTIPROCESSOR | 2048 |  |
|  | ASYNC\_ENGINE\_COUNT | 1 |  |
|  | UNIFIED\_ADDRESSING | 1 |  |
|  | MAXIMUM\_TEXTURE1D\_LAYERED\_WIDTH | 16384 |  |
|  | MAXIMUM\_TEXTURE1D\_LAYERED\_LAYERS | 2048 |  |
|  | CAN\_TEX2D\_GATHER | 1 |  |
|  | MAXIMUM\_TEXTURE2D\_GATHER\_WIDTH | 16384 |  |
|  | MAXIMUM\_TEXTURE2D\_GATHER\_HEIGHT | 16384 |  |
|  | MAXIMUM\_TEXTURE3D\_WIDTH\_ALTERNATE | 2048 |  |
|  | MAXIMUM\_TEXTURE3D\_HEIGHT\_ALTERNATE | 2048 |  |
|  | MAXIMUM\_TEXTURE3D\_DEPTH\_ALTERNATE | 16384 |  |
|  | PCI\_DOMAIN\_ID | 0 |  |
|  | TEXTURE\_PITCH\_ALIGNMENT | 32 |  |
|  | MAXIMUM\_TEXTURECUBEMAP\_WIDTH | 16384 |  |
|  | MAXIMUM\_TEXTURECUBEMAP\_LAYERED\_WIDTH | 16384 |  |
|  | MAXIMUM\_TEXTURECUBEMAP\_LAYERED\_LAYERS | 2046 |  |
|  | MAXIMUM\_SURFACE1D\_WIDTH | 65536 |  |
|  | MAXIMUM\_SURFACE2D\_WIDTH | 65536 |  |
|  | MAXIMUM\_SURFACE2D\_HEIGHT | 32768 |  |
|  | MAXIMUM\_SURFACE3D\_WIDTH | 65536 |  |
|  | MAXIMUM\_SURFACE3D\_HEIGHT | 32768 |  |
|  | MAXIMUM\_SURFACE3D\_DEPTH | 2048 |  |
|  | MAXIMUM\_SURFACE1D\_LAYERED\_WIDTH | 65536 |  |
|  | MAXIMUM\_SURFACE1D\_LAYERED\_LAYERS | 2048 |  |
|  | MAXIMUM\_SURFACE2D\_LAYERED\_WIDTH | 65536 |  |
|  | MAXIMUM\_SURFACE2D\_LAYERED\_HEIGHT | 32768 |  |
|  | MAXIMUM\_SURFACE2D\_LAYERED\_LAYERS | 2048 |  |
|  | MAXIMUM\_SURFACECUBEMAP\_WIDTH | 32768 |  |
|  | MAXIMUM\_SURFACECUBEMAP\_LAYERED\_WIDTH | 32768 |  |
|  | MAXIMUM\_SURFACECUBEMAP\_LAYERED\_LAYERS | 2046 |  |
|  | MAXIMUM\_TEXTURE1D\_LINEAR\_WIDTH | 134217728 |  |
|  | MAXIMUM\_TEXTURE2D\_LINEAR\_WIDTH | 65000 |  |
|  | MAXIMUM\_TEXTURE2D\_LINEAR\_HEIGHT | 65000 |  |
|  | MAXIMUM\_TEXTURE2D\_LINEAR\_PITCH | 1048544 |  |
|  | MAXIMUM\_TEXTURE2D\_MIPMAPPED\_WIDTH | 16384 |  |
|  | MAXIMUM\_TEXTURE2D\_MIPMAPPED\_HEIGHT | 16384 |  |
|  | MAXIMUM\_TEXTURE1D\_MIPMAPPED\_WIDTH | 16384 |  |
|  | STREAM\_PRIORITIES\_SUPPORTED | 1 |  |
|  | GLOBAL\_L1\_CACHE\_SUPPORTED | 0 |  |
|  | LOCAL\_L1\_CACHE\_SUPPORTED | 1 |  |
|  | MAX\_SHARED\_MEMORY\_PER\_MULTIPROCESSOR | 49152 |  |
|  | MAX\_REGISTERS\_PER\_MULTIPROCESSOR | 65536 |  |
|  | MANAGED\_MEMORY | 1 |  |
|  | MULTI\_GPU\_BOARD | 0 |  |
|  | MULTI\_GPU\_BOARD\_GROUP\_ID | 0 |  |
|  | DISPLAY\_NAME | GeForce GT 635 |  |
|  | COMPUTE\_CAPABILITY\_MAJOR | 3 |  |
|  | COMPUTE\_CAPABILITY\_MINOR | 5 |  |
|  | **TOTAL\_MEMORY** | **1073741824** |  |
|  | RAM\_TYPE | 7 |  |
|  | RAM\_LOCATION | 1 |  |
|  | GPU\_PCI\_DEVICE\_ID | 310382814 |  |
|  | GPU\_PCI\_SUB\_SYSTEM\_ID | 2158236749 |  |
|  | GPU\_PCI\_REVISION\_ID | 161 |  |
|  | GPU\_PCI\_EXT\_DEVICE\_ID | 4736 |  |
|  | GPU\_PCI\_EXT\_GEN | 1 |  |
|  | GPU\_PCI\_EXT\_GPU\_GEN | 2 |  |
|  | GPU\_PCI\_EXT\_GPU\_LINK\_RATE | 5000 |  |
|  | GPU\_PCI\_EXT\_GPU\_LINK\_WIDTH | 8 |  |
|  | GPU\_PCI\_EXT\_DOWNSTREAM\_LINK\_RATE | 5000 |  |
|  | GPU\_PCI\_EXT\_DOWNSTREAM\_LINK\_WIDTH | 16 |  |
|  |  |  |  |

|  |  |
| --- | --- |
|  | * **Core의 개수: 384개** * **메모리 크기: 1073741824Byte (1GB)** |

## 실습 2



C:\Users\cspclab\Desktop\element size.JPG

모든 값은 15번의 평균 값이며, 실습의 경우 GPU가 리뉴얼 된 컴퓨터가 아니어서 block size에 제한이 있었다.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| block size=2 | CPU | GPU | block size=4 | CPU | GPU |
| avg | 23.19588 | 0.044928 | avg | 23.051 | 0.028068 |
| block size=8 | CPU | GPU | block size=16 | CPU | GPU |
| avg | 23.11482 | 0.029771 | avg | 23.18352 | 0.015785 |
| block size=32 | CPU | GPU | block size=64 | CPU | GPU |
| avg | 22.96692 | 0.013523 | avg | 23.18079 | 0.014158 |
| block size=128 | CPU | GPU | block size=256 | CPU | GPU |
| avg | 23.01989 | 0.01376 | avg | 23.1087 | 0.013583 |

블럭의 최대크기는 1024이므로 이를 넘어가면 실행이 불가능하였다.

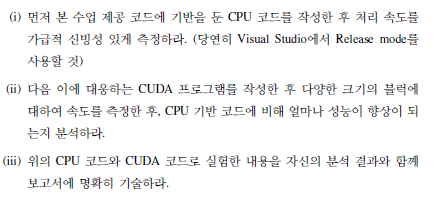
블럭의 크기를 늘리면 와프의 개수도 많아져 실행속도가 빨라질 것이라 예상했으나 블록의 크기가 커질수록 실행속도가 빨라지는 것은 아니었다.

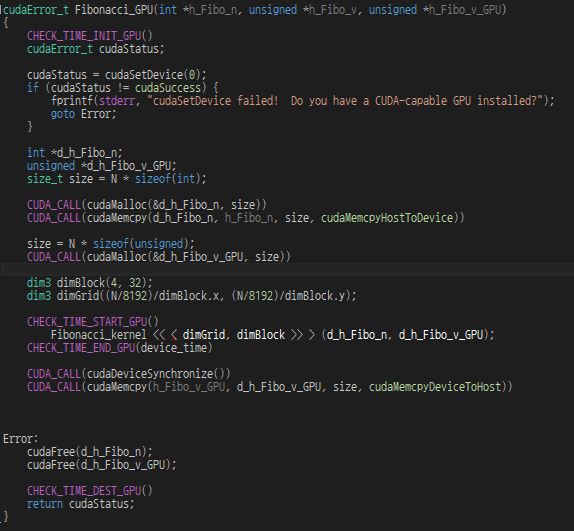
## 실습 3

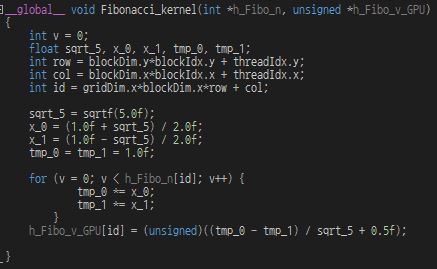
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| block size=2 | CPU | GPU | block size=4 | CPU | GPU |
| avg | 5.445120733 | 13.133024 | avg | 18.326488 | 13.609779 |
| block size=8 | CPU | GPU | block size=16 | CPU | GPU |
| avg | 68.34754893 | 49.21799027 | avg | 368.6916993 | 851.4683594 |
| block size=32 | CPU | GPU |  |  |  |
| avg | 1637.511527 | 7974.461621 |  |  |  |

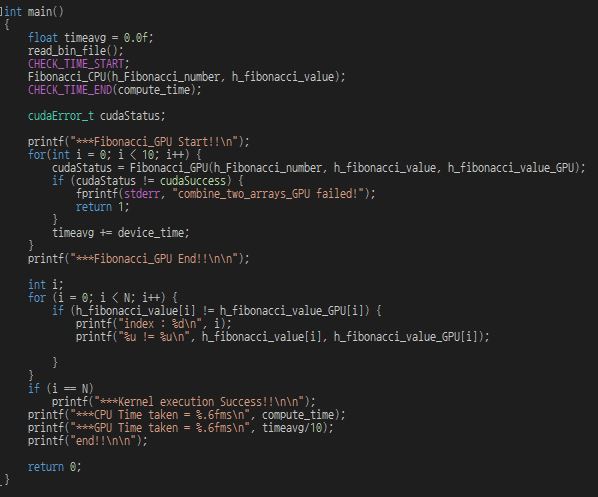
정사각형으로 나눌 경우, block size = 64 부터는 실행되지 않는다. block size = 4부터 CPU보다 좋은 성능을 갖지만, block size = 16부터는 CPU보다 성능이 좋지 않다. 따라서 가장 적절한 block size = 8이다.

## 숙제 2





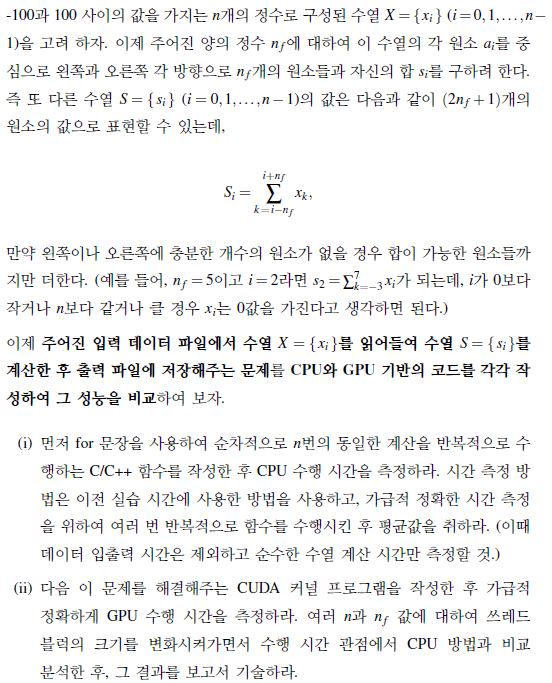


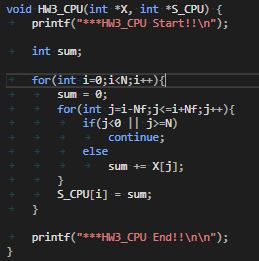


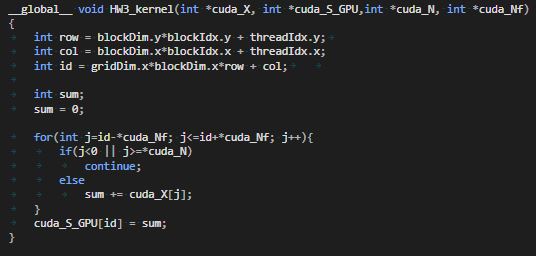
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| block size =8 | CPU | GPU | block size =16 | CPU | GPU |
| avg | 1270.844011 | 65.8428768 | avg | 1273.151815 | 65.755047 |
| block size =32 | CPU | GPU | block size = 4, 8 | CPU | GPU |
| avg | 1271.259619 | 65.7561826 | avg | 1269.494108 | 65.72501873 |
| block size = 8, 16 | CPU | GPU | block size = 16, 32 | CPU | GPU |
| avg | 1269.395524 | 65.7168 | avg | 1270.47806 | 65.71960813 |
| block size = 4, 32 | CPU | GPU |  |  |  |
| avg | 1306.475936 | 65.712619 |  |  |  |

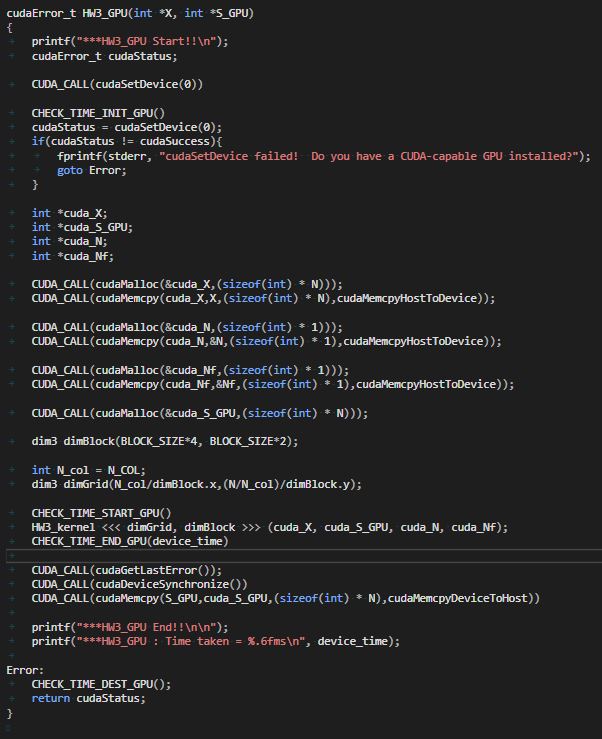
CPU보다 GPU의 평균 속도가 더 빨랐다. 정사각형뿐만 아니라, 직사각형 등 여러 block size를 입력해보았으나 대체로 CPU는 각 1270, GPU는 65ms를 기록하였다.

## 숙제 3









|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| block size=8,8 | CPU | GPU | block size=8,16 | CPU | GPU |
| avg | 236.84026 | 77.23095 | avg | 227.11459 | 53.80145 |
| block size=8,32 | CPU | GPU | block size=16,8 | CPU | GPU |
| avg | 226.23578 | 56.15797 | avg | 225.17954 | 48.015785 |
| block size=16,16 | CPU | GPU | block size=16,32 | CPU | GPU |
| avg | 227.86793 | 49.14735 | avg | 226.13529 | 49.13415 |
| block size=32,16 | CPU | GPU | block size=32,32 | CPU | GPU |
| avg | 226.01402 | 49.01328 | avg | 230.31078 | 51.017148 |

우선 최대 크기 N을 확인하였다.

#define N\_initial 1<<26

다양한 block size에서 실험을 해보고 CPU와 GPU의 성능을 binary 파일을 읽어 output하는 함수를 만들어 확인해보았다. 블록사이즈에 따라 다이나믹한 변화가 있을 것이라고 예상하였지만 아주 큰 변동은 없었다. 그러나 block size = 8,8가 77ms block size = 32,16가 49ms를 기록한 것으로 보아 적절한 block size를 사용하는 것이 앞으로의 실험에 있어서 중요함을 깨달았다.