**고소실 12주차 보고서**

20191571 김세영

**1. 실습**

1) (단위:ms), n: 2^26

nf=1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Thread block size | 16 | 32 | 64 | 128 | (32,8) | (32,16) |
| CPU | 180.295 | | | | | |
| NO\_SHARED | 13.042 | 7.65 | 3.949 | 3.086 | 3.092 | 3.95 |
| SHARED | 11.148 | 5.49 | 2.896 | 3.073 | 2.912 | 3.632 |
| speed up | 1.169896 | 1.393443 | 1.363605 | 1.00423 | 1.061813 | 1.087555 |

nf=4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Thread block size | 16 | 32 | 64 | 128 | (32,8) | (32,16) |
| CPU | 369.431 | | | | | |
| NO\_SHARED | 13.964 | 7.552 | 4.106 | 3.544 | 3.649 | 4.507 |
| SHARED | 9.359 | 5.657 | 3.423 | 2.726 | 2.794 | 3.805 |
| speed up | 1.49204 | 1.334983 | 1.199533 | 1.300073 | 1.306013 | 1.184494 |

nf=16

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Thread block size | 16 | 32 | 64 | 128 | (32,8) | (32,16) |
| CPU | 1497.995 | | | | | |
| NO\_SHARED | 21.615 | 11.458 | 7.133 | 6.66 | 6.987 | 7.996 |
| SHARED | 12.894 | 6.774 | 5.989 | 6.108 | 6.155 | 6.754 |
| speed up | 1.676361 | 1.691467 | 1.191017 | 1.090373 | 1.135175 | 1.183891 |

nf=64

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Thread block size | 16 | 32 | 64 | 128 | (32,8) | (32,16) |
| CPU | 4860.278 | | | | | |
| NO\_SHARED | 64.651 | 32.663 | 24.139 | 22.498 | 22.56 | 22.769 |
| SHARED | 36.458 | 18.243 | 17.826 | 17.808 | 18.467 | 19.378 |
| speed up | 1.773301 | 1.79044 | 1.354146 | 1.263365 | 1.221639 | 1.174992 |

nf=256

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Thread block size | 16 | 32 | 64 | 128 | (32,8) | (32,16) |
| CPU | 18121.266 | | | | | |
| NO\_SHARED | 181.924 | 112.904 | 84.098 | 84.104 | 81.216 | 82.007 |
| SHARED | 104.514 | 52.766 | 52.952 | 52.473 | 51.432 | 53.041 |
| speed up | 1.740666 | 2.139711 | 1.588193 | 1.602805 | 1.579095 | 1.546106 |

nf=1024

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Thread block size | 16 | 32 | 64 | 128 | (32,8) | (32,16) |
| CPU | 69762.227 | | | | | |
| NO\_SHARED | 575.381 | 338.927 | 292.189 | 308.09 | 282.548 | 301.099 |
| SHARED | 398.589 | 201.002 | 209.368 | 207.283 | 200.505 | 201.232 |
| speed up | 1.443545 | 1.686187 | 1.395576 | 1.486325 | 1.409182 | 1.496278 |

nf가 증가할수록 계산량이 많아지기 때문에 CPU와 GPU 모두에서 수행시간이 증가한다. nf가 커질수록 CPU와 GPU 모두에서 수행시간이 증가하는 비율이 늘어났는데, 이것 또한 계산량이 늘어나는 비율이 커지기 때문이다.

shared memory를 사용한 실험 결과는 모든 경우에서 shared memory를 사용하지 않는 경우보다 수행 시간이 적다. shared memory를 사용했을 때의 속도 향상은(shared memory/no shared memory, speed up) nf가 커질수록 크게 나타났다. nf가 크면 global memory access 횟수가 늘어나고, global memory access는 수백 사이클이 걸리기 때문에 대신에 shared memory를 사용하면 global memory access 시간을 줄일 수 있기 때문이다.

shared memory를 사용한 경우에도 thread block size가 warp의 배수인 경우가 16인 경우보다 빠르게 수행되었다. nf=1 ,16 일 때는 64, nf= 4, 64 일 때는 128, nf= 64, 256, 1024 일 때는 (32,8) 에서 가장 빠르게 수행되었다. shared memory를 사용하지 않더라도 대체로 shared memory와 비슷한 경향을 보인다. nf가 큰 경우(64이상)에는 2차원 thread block인 (32,8)에서 가장 빠른 속도를 보였다.

**2. 과제**

|  |  |  |
| --- | --- | --- |
| **텍스트이(가) 표시된 사진  자동 생성된 설명** | **텍스트이(가) 표시된 사진  자동 생성된 설명** | **텍스트이(가) 표시된 사진  자동 생성된 설명** |

순서대로 각각 b1(no shared), b2(shared), b3(CPU)이다. 이 세 방식의 구현 결과는 같다.

thread block size의 x, y가 이미지의 가로, 세로 길이에 나누어 떨어지는 경우만 가정한다.

|  |  |  |  |
| --- | --- | --- | --- |
| (단위: ms) | b1(no shared) | b2(shared) | b3 (CPU) |
| (16,32) | 2.19 | 3.19 | 19 |
| (16,16) | 2.13 | 2.15 |
| (8,16) | 2.12 | 2.16 |
| (8,8) | 2.17 | 2.16 |

실습에서는 shared memory를 사용하는 경우가 shared memory를 사용하지 않는 경우보다 항상 수행시간이 빨랐지만, 이 과제에서는 두 방식 사이의 차이가 거의 보이지 않았다. 또한 커널을 처음 시작한 경우 수행시간이 일정하게 측정되지 않아 측정에 어려움이 있었다. 두 방식에 뚜렷한 차이가 보이지 않는 원인 중 하나는 이 프로그램에 사용되는 이미지 파일의 크기가 충분히 크지 않다는 것이다. 이 이미지의 크기는 368px X 352px이고, 실습 때 사용한 데이터 (2^26)와 비교했을 때 매우 작은 크기이다. 원인 중 또 다른 하나는 shared memory를 사용한 코드에서 branch가 많이 사용되었다는 것이다. shared memory를 사용하는 경우는 thread가 thread block의 가장자리이거나 thread block이 전체 이미지의 가장자리인 경우 예외처리를 해야하는데, GPU는 branch에 취약하기 때문에 이 과정에서 수행시간이 오래 걸렸을 가능성이 있다.