[CSEG483/CSE5483] 기초 GPU 프로그래밍

HW 2: Matrix Multiplication

담당교수: 서강대학교 컴퓨터공학과 임 인 성

May 13, 2025

제출 마감: 5월 31일 (토) 오후 8시 정각 이전에 조교가 사이버 캠퍼스에 공지한 방식으로 제출 (LATE 최대 24 시간)

이번 숙제에서는 두 행렬 A와 B를 파일에서 읽어 들인 후 서로 곱하여 생성한 행렬 C를 파일에 저장하는 하는 문제를 GPU를 사용하여 해결하여 본다. 이를 위하여 다음 7 가지 방법을 구현하여 실험하라.

1. [**방법 1**] Device에서 shared memory를 사용하지 않는 기본적인 함수를 가급적 효율적으로 구현하라. (MM_DEVICE_GM)
2. [**방법 2**] Device에서 shared memory를 사용하는 함수를 가급적 효율적으로 구현하라. (MM_DEVICE_SM)
3. [**방법 3**] Device에서 shared memory를 사용하고 "More-Work-per-Thread" 아이디어를 사용하여 최적화한 함수를 가급적 효율적으로 구현하라. (MM_DEVICE_SM_MWPT)
4. [**방법 4**] Device에서 Tensor core를 기반으로 shared memory를 사용하지 않는 함수를 가급적 효율적으로 구현하라. (MM_DEVICE_TC_GM)
5. [**방법 5**] Device에서 Tensor core를 기반으로 shared memory를 사용하여 최적화한 함수를 가급적 효율적으로 구현하라. (MM_DEVICE_TC_SM)
6. [**방법 6**] Device에서 cuBlas에 기반하여 CUDA Core를 활용하는 함수를 가급적 효율적으로 구현하라. (MM_DEVICE_CUBLAS_CC)
7. [**방법 7**] Device에서 cuBlas에 에 기반하여 Tensor Core를 활용하는 함수를 가급적 효율적으로 구현하라. (MM_DEVICE_CUBLAS_TC)

참고

1. 행렬 A, B, 그리고 C는 다음과 같은 형식으로 저장된다.
   - 행렬의 크기가 m행 n열이라 할 때, 파일의 첫 4 바이트에는 m, 그리고 다음 4 바이트에는 n 값이 `int` 형식으로 binary 형태로 저장된다.
   - 다음 총 mn개의 원소가 row-major 형태로 나열이 되는데, 각 원소 값은 4 바이트에 `float` 형식으로 binary 형태로 저장된다.

2. 입력 행렬의 각 원소는 0과 1 사이의 uniform random number로 생성된다.

3. (채점의 편의 상) 자신의 C++ 코드의 앞 부분에서 다음과 같이 세 파일의 이름을 매크로를 사용하여 정의하라 (파일의 이름은 임의로 정할 것).

```
#define  FILE_A  my_file_A.bin
#define  FILE_B  my_file_B.bin
#define  FILE_C_1  my_file_C_1.bin
           ⋮
#define  FILE_C_7  my_file_C_7.bin
```

4. [**방법 1**], [**방법 2**], [**방법 3**], 그리고 [**방법 6**]은 모두 FP32 정밀도로 계산을 하고, 나머지 방법은 mixed-precision을 사용하되 어떤 방식을 사용하였는지 보고서에 밝힐 것.

5. 여러분의 C++ 코드는 한 번의 수행을 통하여 7개의 파일을 생성하여야 하며, 콘솔 윈도우에 각 방법의 커널 수행시간 (ms)을 순서대로 출력해야 한다.

6. 조교는 CPU에서 double precision을 사용하여 계산한 결과를 정답으로 하여 여러분의 각 방법이 생성한 C 행렬의 정확도를 확인할 예정임.

7. 자신의 코드는 임의의 크기의 행렬을 처리하여야 하나, 편의상 채점은 A 행렬은 2,048행 4,096열, 그리고 B 행렬은 4,096행 1,024열로 할 예정임.

8. 자신의 실험 결과를 바탕으로 발견한 내용과 분석한 내용을 보고서에 분명히 기술할 것.//

제출물: 자신이 구현한 코드를 이름이 HW_2_학번인 디렉터리 아래의 Visual Studio 프로젝트를 생성한 후, zip으로 압축하여 제출할 것.

1. **자신이 작성한 코드:** Visual Studio 2022를 통하여 확인 할 수 있도록 위의 directory를 제출하되 .vs 파일 등 코드 수행에 불필요한 파일들은 반드시 제거한 후 제출할 것.

2. **프로그램 실행 결과:** 자신의 코드를 실행한 결과를 증빙할 수 있는 자료 (예를 들어, 콘솔 윈도우의 내용을 캡춰한 영상)를 보고서에 포함할 것.

3. **보고서:** 자신의 실험 결과를 바탕으로 분석한 내용을 기술할 것.