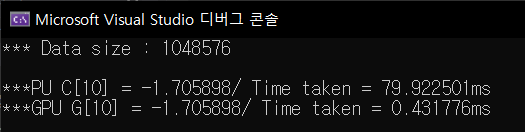
**고급소프트웨어실습 11주차 과제**

**20182204 한찬희**

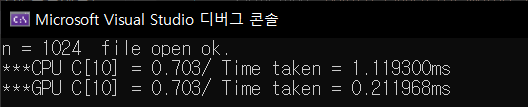
**실습 문제 1, 2, 3**

메인 메모리에서 두 배열 데이터를 읽어들여 y = Mxi 에 해당하는 백터 y를 구하는 프로그램이다.

* Init\_data() 함수와 init\_MatVec() 함수를 이용하여 입력 데이터를 설정한 후 Mat\_Vec\_Multiply()를 통해 CPU 계산을, Mat\_Vec\_Multiply\_GPU() 함수를 통해 GPU 계산을 처리한다. GPU 계산에서는 그리드를 1차원으로 설정하였고, CUDA API 함수를 이용해 device 메모리를 할당하였다. 특히 GPU 계산에 쓰이는 함수 Mat\_Vec\_Multiply\_Kernel() 함수는 병렬 계산을 위해 해당 thread의 아이디를 id 변수에 넣고, (blockDim.x \* blockIdx.x + threadIdx.x) 해당 thread의 결과 벡터를 구해 d\_VecY 에 넣는다. 결과를 구한 후 다시 Host 메모리로 복사하여 결과 배열에 저장한다.
* 실행 결과
  + Proc1



* + Proc2

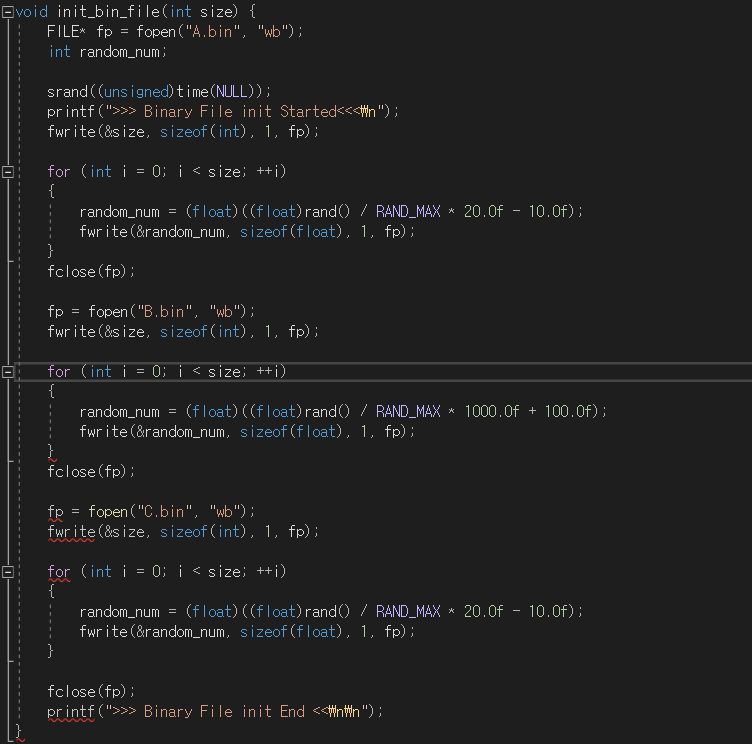


* 블록의 크기만큼 작업을 분할해 병렬 계산을 처리하는 GPU 계산의 속도가 빠른 것을 확인할 수 있다.

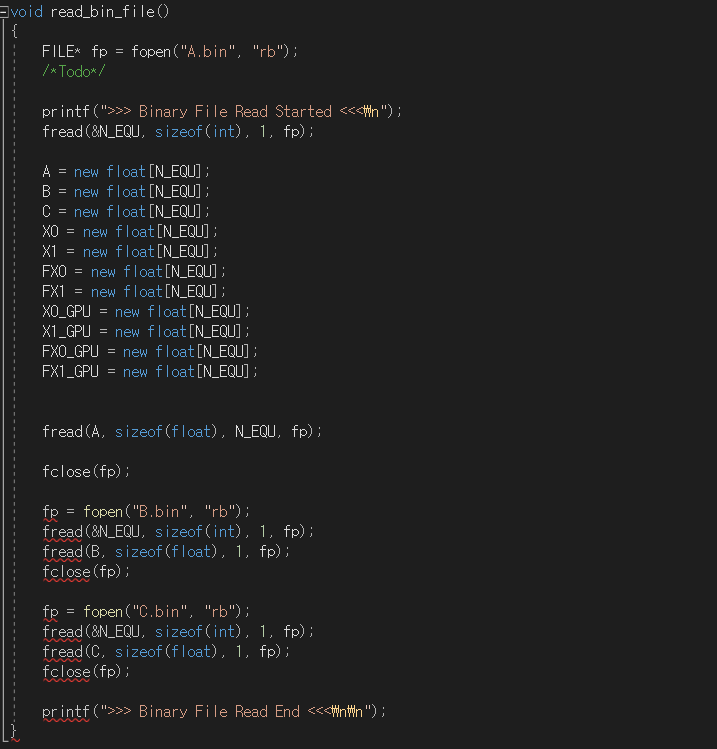
**숙제 1**

2차 방정식의 근을 구해 주는 프로그램이다.

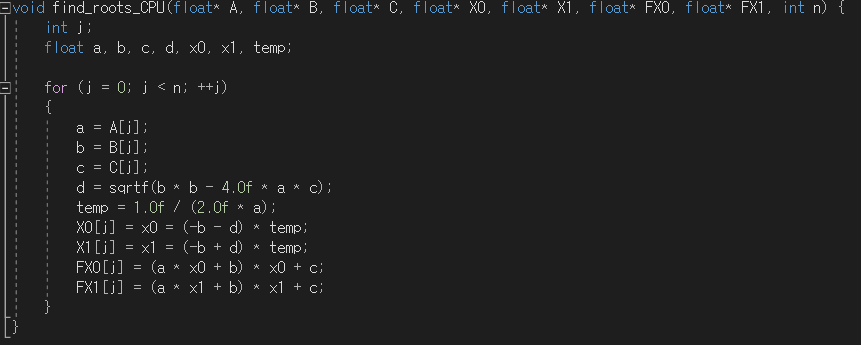
* **Init\_bin\_file**
* A.bin, B.bin, C.bin에 N\_EQUATIONS와 i번째 2차 방정식의 계수를 랜덤으로 생성하여 저장하는 함수이다.



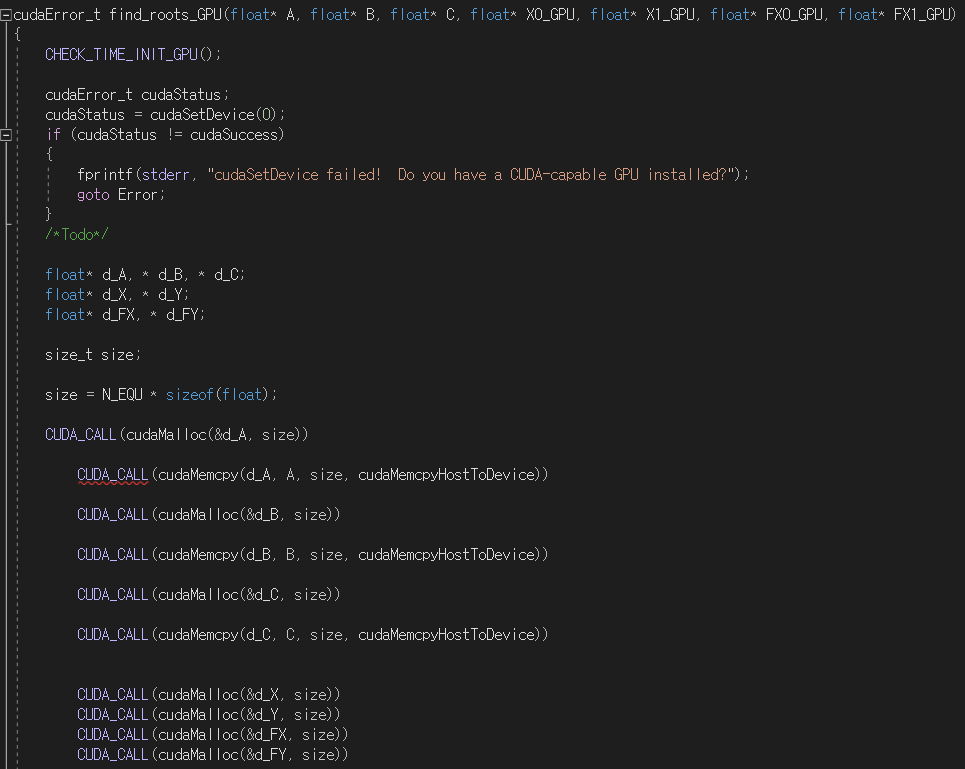
* **Read\_bin\_file()**
* A.bin, B.bin, C.bin을 읽어 들여 배열에 저장하는 함수이다. 각각 A, B, C에 저장하였다.

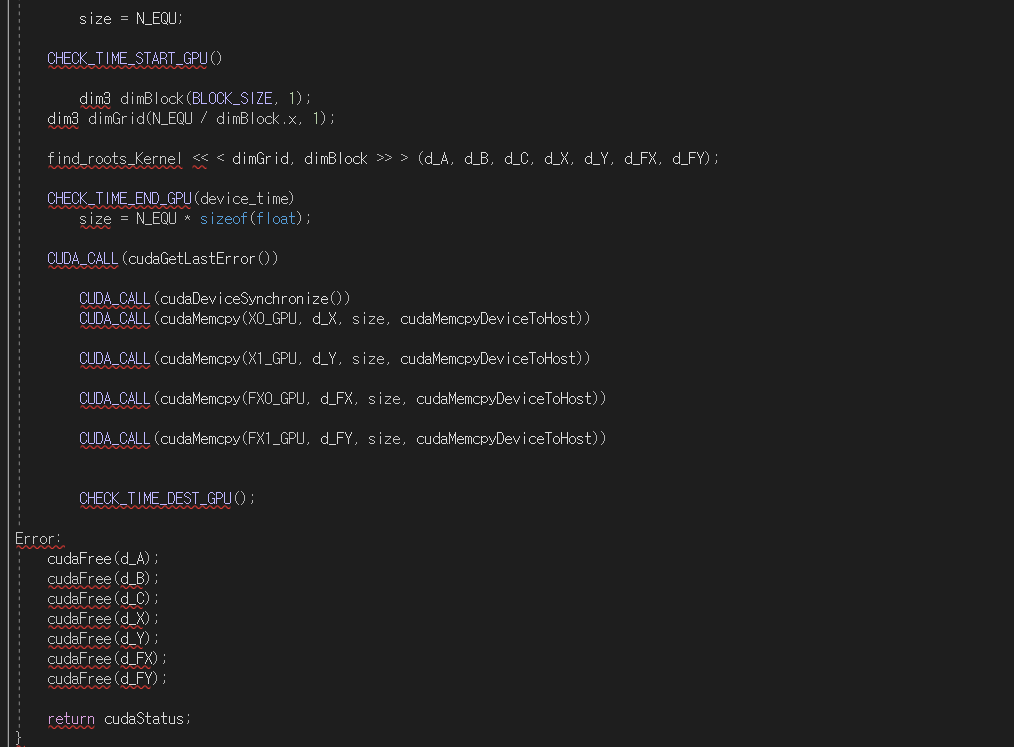


* **Find\_roots\_CPU**
* 2차 방정식의 근을 찾아 X0, X1에 저장하는 CPU 코드이다. 함수의 결과는 FX0, FX1에 저장한다.

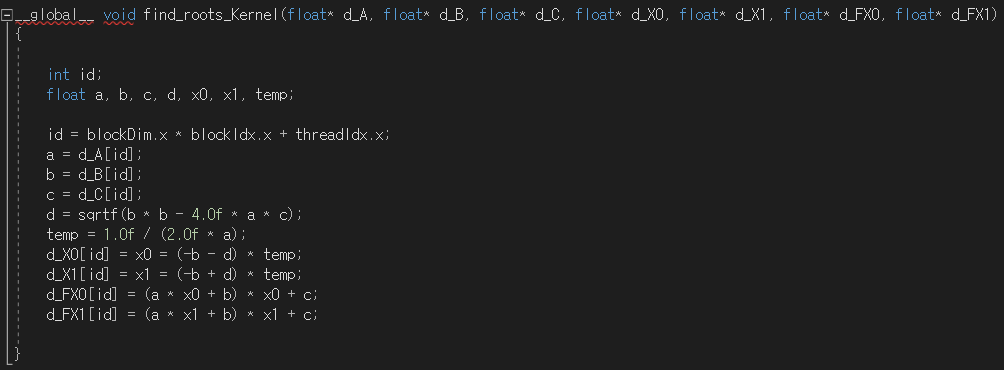


* **Find\_roots\_GPU**
* 2차 방정식의 근을 찾아 X0\_GPU, X1\_GPU에 저장하는 GPU 코드이다. 배열 복사를 위해 CUDA API 함수를 cudaMalloc, cudaMemcpy를 이용하여 device(GPU) 메모리에 메모리를 할당하고 host(CPU) 메모리에서 device 메모리로 메모리를 복사한다. 블록의 크기로 블록 안의 스레드 개수를 3차원적으로 지정해 주는 dimBlock은 (BLOCK\_SIZE, 1), 그리고 그리드의 크기로 그리드 안의 블록 개수를 3차원적으로 지정해 주는 dimGrid은 (N\_EQUATIONS / dimBlock.x, 1)로 설정해 준다. (2번째 인자 y가 1인 이유는 계산에서 쓰이는 배열이 1차원이기 때문, 3번째 인자 z는 생략됐으므로 자동으로 1로 지정이 됨)
* 이후 find\_roots\_Kernel 함수를 출력하여 커널을 수행시켜 각 thread에 대해 2차 방정식의 근과 함숫값을 계산해 준다.

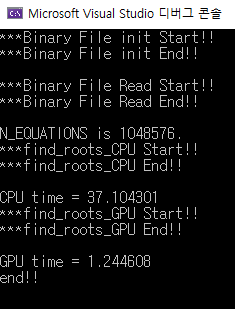




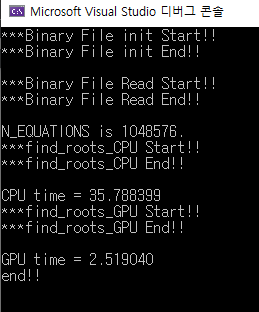
* **Find\_roots\_Kernel**
* 2차 방정식의 근을 계산해 d\_X, d\_Y에 저장하는 커널 함수이다. 이때 tid는 blockDim.x\*blockIdx.x + threadIdx.x로 지정해 주었는데, 해당 thread에 대해 병렬적으로 계산해야 하므로 해당 thread의 인덱스를 tid에 지정한 것이다.



* **실행 결과 (Debug 모드)**
* N\_EQUATIONS = 1048675 (실험하기 너무 크지도, 작지 않게 적합한 값을 임의로 설정하였다.)
* BLOCK\_SIZE = 32 / BLOCK\_SIZE = 64

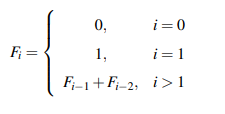
* BLOCK\_SIZE = 16 / BLOCK\_SIZE = 2

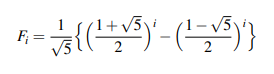
* CPU로 계산했을 때보다 GPU로 계산했을 때 모두 빠른 속도를 보였고, BLOCK\_SIZE가 작아질수록 GPU 계산 속도는 느려졌다. 블록의 크기는 작업을 분할하는 기준이므로 많은 수를 지정할수록 분할되는 작업이 많아지는 걸 의미하기 때문에 블록의 크기가 늘어날수록 실행 속도가 빨라진 것으로 보인다. 2, 16에서 32로 커졌을 때는 현저한 차이를 보였으나 32에서 64로 갔을 때는 근소하게 빨라진 것으로 보아 32, 64 등이 적절한 블록의 크기인 것으로 보인다.

**숙제 2**

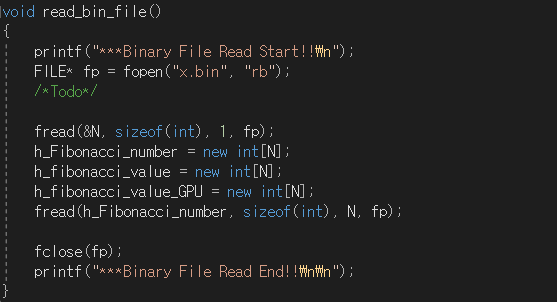
숙제 2는 입력 배열에 저장되어 있는 N개의 임의의 정수 값 각각에 대한 피보나치 수를 계산하여 출력 배열에 저장하는 프로그램이다. 피보나치 수는 다음과 같이 정의된다.



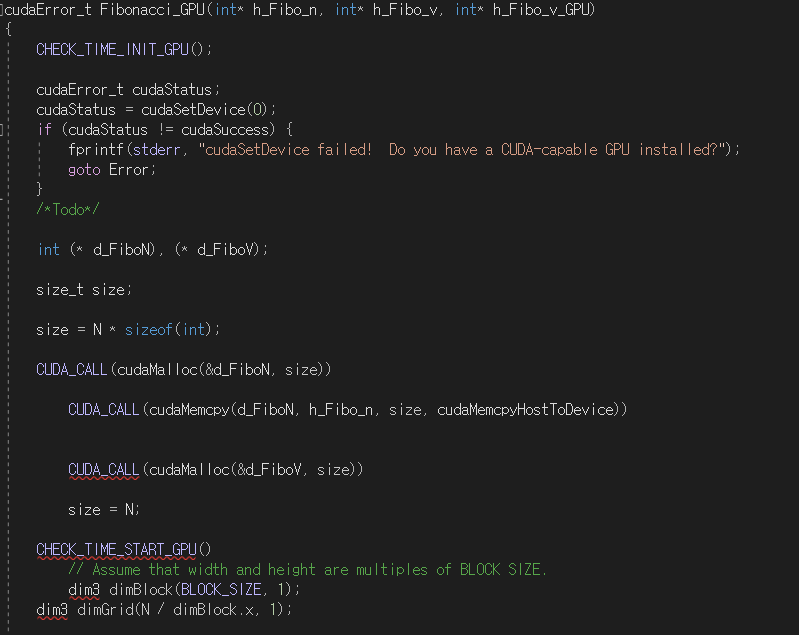
임의의 n에 대해 Fn은 다음과 같이 표현할 수 있다.

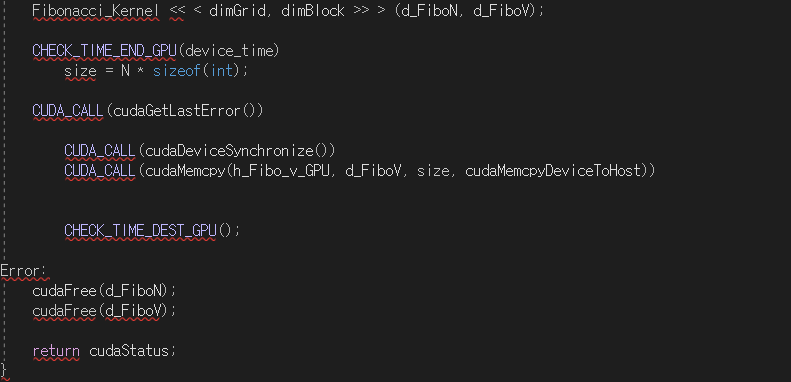


* **Init\_bin\_file(int size)**
* 정수인 난수를 생성하여 x.bin 파일에 저장해 주는 함수이다.
* **Read\_bin\_file()**

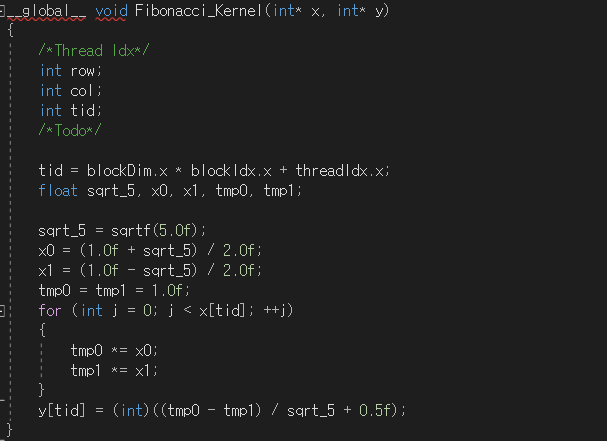


* x.bin에서 배열의 크기와 입력 배열을 읽어 들여 N, h\_Fibonacci\_number에 저장한다.
* **Fibonacci\_CPU(int \*h\_Fibo\_n, int \*h\_Fibo\_v)**
* 피보나치 수를 계산하는 CPU 코드이다. 결과는 h\_Fibo\_v에 저장한다.
* **Fibonacci\_GPU(int \*h\_Fibo\_n, int \*h\_Fibo\_v, int \*h\_Fibo\_v\_GPU)**

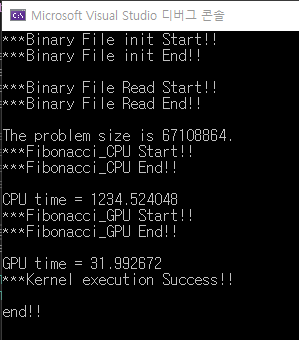
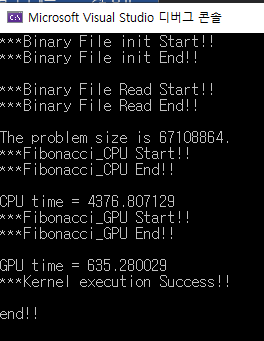




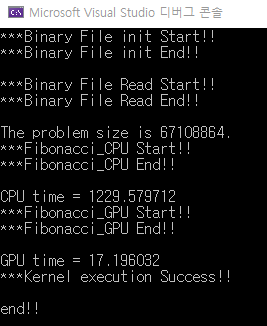
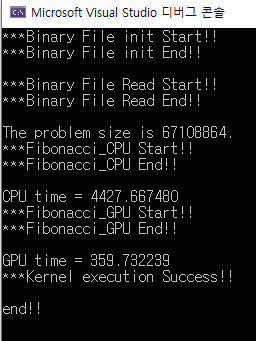
* 피보나치 수를 계산하는 GPU 코드이다. 입력 배열 복사를 위해 CUDA API 함수를 cudaMalloc, cudaMemcpy를 이용하여 device(GPU) 메모리에 메모리를 할당하고 host(CPU) 메모리에서 device 메모리로 메모리를 복사한다. (h\_Fibo\_n -> d\_FiboN) 또한 출력 배열 d\_FiboV도 cudaMalloc 함수를 이용하여 메모리 할당을 해 준다. 블록의 크기로 블록 안의 스레드 개수를 3차원적으로 지정해 주는 dimBlock은 (BLOCK\_SIZE, 1), 그리고 그리드의 크기로 그리드 안의 블록 개수를 3차원적으로 지정해 주는 dimGrid은 (N / dimBlock.x, 1)로 설정해 준다. (2번째 인자 y가 1인 이유는 계산에서 쓰이는 배열이 1차원이기 때문, 3번째 인자 z는 생략됐으므로 자동으로 1로 지정이 됨) 이후 Fibonacci\_Kernel 함수를 출력하여 커널을 수행시켜 각 thread에 대해 피보나치 수를 계산해 준다.
* 시간 측정은 커널로 계산할 때만 필요하므로 해당 함수 앞뒤에 시간 측정 함수를 배치하였다.
* **Fibonacci\_Kernel(int \*x, int \*y)**



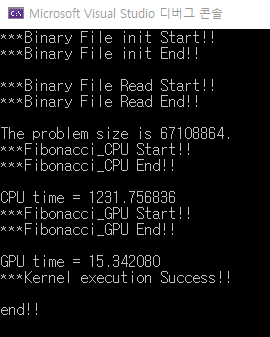
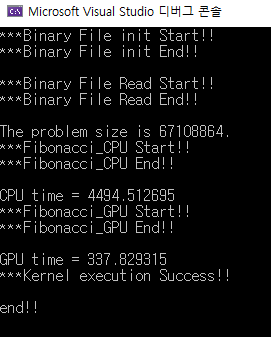
* 각 thread에 대해 피보나치 수를 계산하는 GPU 함수이다. tid를 설정한 후 해당 thread에 대한 출력 배열을 계산한다.
* **BLOCK\_SIZE에 대한 실행 결과 (왼쪽: Debug 모드, 오른쪽: Release 모드)**
* BLOCK\_SIZE 16



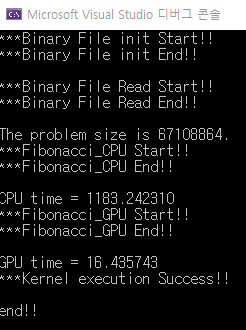
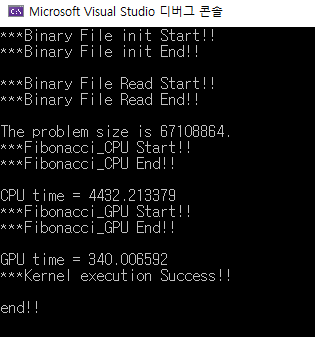
* BLOCK\_SIZE 32



* BLOCK\_SIZE 64



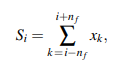
* BLOCK\_SIZE 128



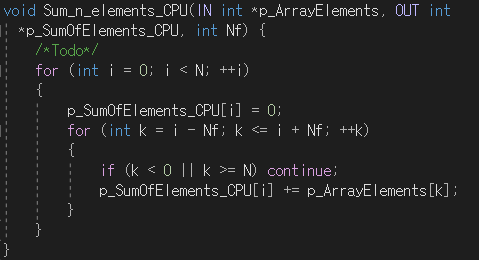
* 모든 실행 결과를 통해 GPU 계산 속도가 CPU 계산 속도보다 현저히 빠르다는 것을 확인할 수 있었다. 동일한 방식의 계산이 서로 다른 데이터에 대하여 반복이 되고 있고, 그 계산이 서로 독립적일 때 CUDA 프로그램을 이용하여 병렬 계산을 수행할 수 있는데, 이 때문에 CPU보다 확실히 빠르게 계산을 완료할 수 있는 것이다. 또한, BLOCK\_SIZE가 커질수록 GPU 계산 속도가 점점 더 빨라지는 것을 확인할 수 있었다. 블록의 수는 작업을 분할하는 기준이므로 많은 수를 지정할수록 분할되는 작업이 많아지는 걸 의미하기 때문에 블록의 수가 늘어날수록 실행 속도가 빨라진 것으로 보인다. 그러나 32 이후로는 계산 속도가 비등비등한 결과로 나왔다.

**숙제 3**

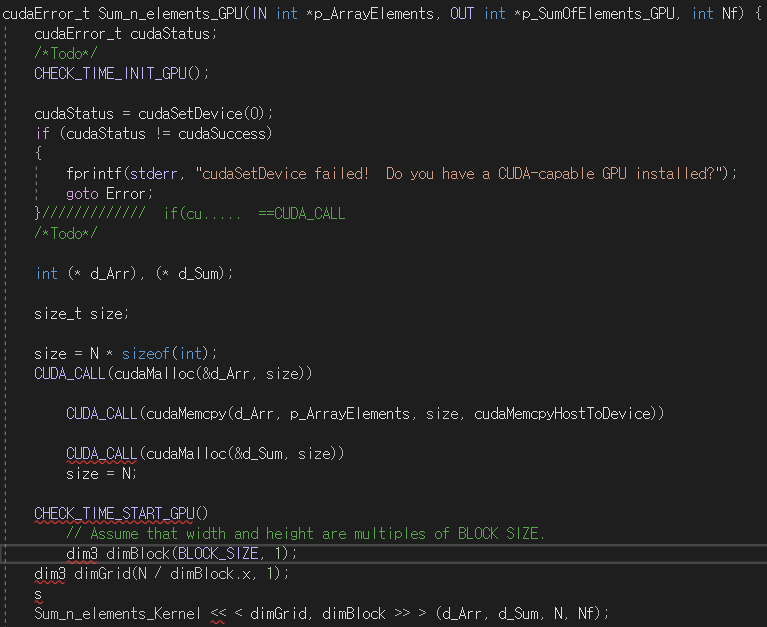
-100과 100사이의 값을 가지는 n개의 정수로 구성된 수열 X에 대하여 다음과 같은 S 수열을 구하는 프로그램이다.

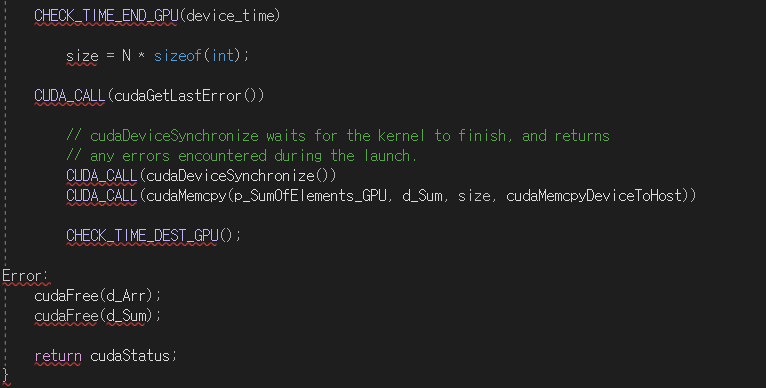


* **Read\_bin\_file(), init\_bin\_file()**
* 입력배열을 랜덤으로 생성해 Cuda\_HW3\_input.bin 파일에 저장하고 그것을 읽어들여와 h\_ArrayElements에 저장하는 함수들이다.
* **Sum\_n\_elements\_CPU**

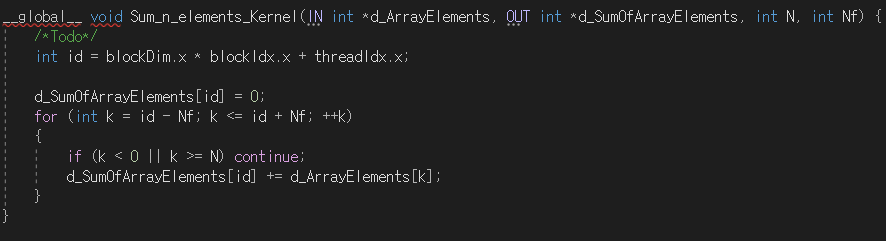


* 배열의 index-Nf부터 index+Nf 데이터까지의 합을 계산하는 CPU 프로그램이다. K가 0보다 작거나 N 이상일 경우 0으로 간주해 더하지 않고 다음으로 넘어간다. 출력 배열은 p\_SumOfElements\_CPU에 저장된다.
* **Sum\_n\_elements\_GPU**





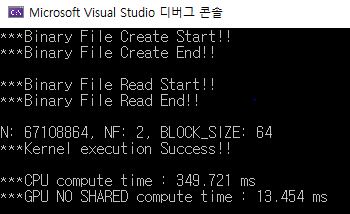
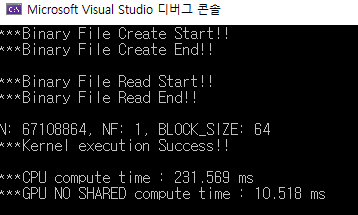
* 데이터를 계산하는 GPU 코드이다. 숙제 2와 같이 cudaMalloc, cudaMemcpy를 이용하여 d\_Arr, d\_Sum을 device(GPU) 메모리에 할당한다. 이후 Sum\_n\_elements\_kernel 함수를 호출해 d\_Sum을 계산하여 넣는다. 숙제 2와 같이 1차원 그리드로 설정하였다.
* **Sum\_n\_elements\_Kernel**



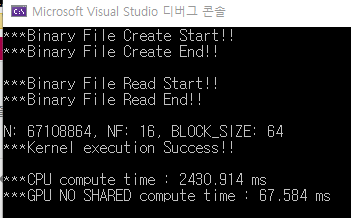
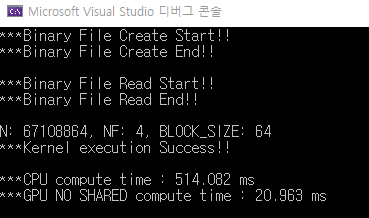
* D\_Sum을 계산하는 커널 함수이다.
* **실행 결과 (Release 모드)**

같은 BLOCK\_SIZE에 대하여 Nf가 다를 때

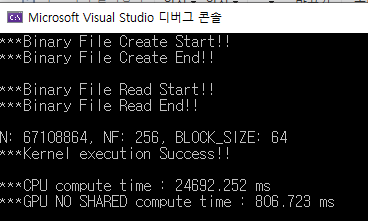
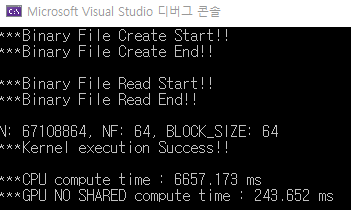
* Nf: 1, BLOCK\_SIZE: 64 / Nf: 2, BLOCK\_SIZE: 64



* Nf: 4, BLOCK\_SIZE: 64 / Nf: 16, BLOCK\_SIZE: 64



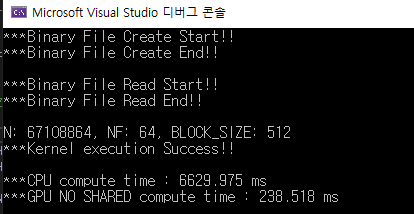
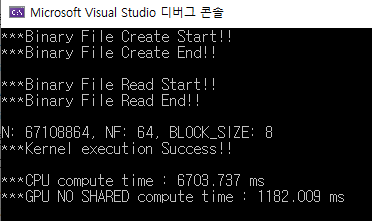
* Nf: 64, BLOCK\_SIZE: 64 / Nf: 256, BLOCK\_SIZE: 64



* Nf가 커질수록 CPU와 GPU 계산 속도는 현저히 증가하는 것을 확인할 수 있었다. Nf가 커질수록 합을 구하는 계산량도 늘어나기 때문에 처리하는 시간도 오래 걸리는 것으로 보인다.

같은 Nf에 대하여 BLOCK\_SIZE가 다를 때

* Nf: 64, BLOCK\_SIZE: 8 / Nf: 64, BLOCK\_SIZE: 512



* 숙제 2와 같이 블록의 수가 커질수록 GPU의 처리하는 시간도 줄어드는 것을 확인할 수 있었다. 블록의 수가 많아질수록 분배의 수가 많아지는 것이므로 처리 속도가 빨라진 것으로 보인다.