Michał Szczygieł	Aleksander Śmierciak	06/04/2014

Zadanie 5 - Mnożenie macierzy na GPU

Programem spełniającym polecenie zadania 5. jest taki, który wykona mnożenie dwóch macierzy kwadratowych przy użyciu karty graficznej, np. za pomocą CUDA.

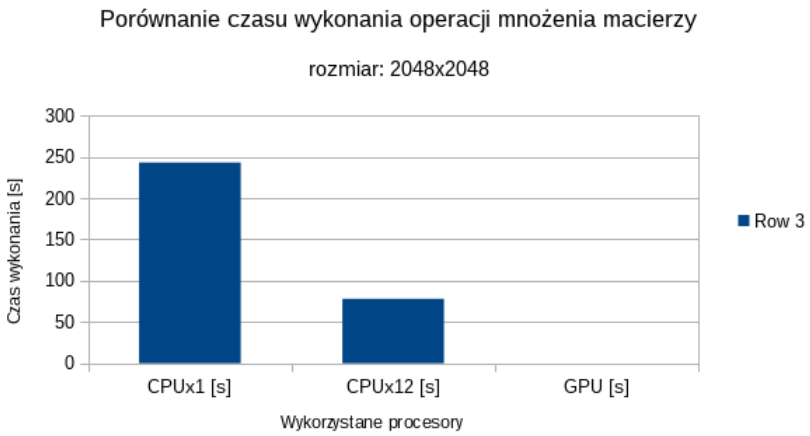
```
1   dim3 dimBlock;
2   dimBlock.x = matrixSize;
3   dimBlock.y = matrixSize;
4
5   float *hostA = initializeMatrix(matrixSize);
6   float *hostB = initializeMatrix(matrixSize);
7
8   int allocBuffer = matrixSize * matrixSize * sizeof(float);
9   float *devA = allocateDeviceMemory(allocBuffer);
10  float *devB = allocateDeviceMemory(allocBuffer);
11  float *devC = allocateDeviceMemory(allocBuffer);
12
13  copyHostMemoryToDevice(hostA, devA, allocBuffer);
14  copyHostMemoryToDevice(hostB, devB, allocBuffer);
15
16  cudaEvent_t start, stop;
17  createTimerEvents(start, stop);
18
19  startTimer(start);
20
21  matrixMultiplyKernel<<<dimBlock, threadCount>>>(devA, devB, devC,
        matrixSize);
22
23  stopTimer(stop);
24
25  cout << readExecutionTime(start, stop) << endl;
26
27  destroyTimerEvents(start, stop);
28  freeMemory(devA, hostA, devB, hostB, devC);
```

Etapy wykonania programu wykorzystującego CUDA można wydzielić następująco:

1. Ustawienie zmiennej typu struktury dim3, która przechowuje informacje o wymiarach problemu z zadania.
2. Zainicjalizowanie macierzy wejściowych, używając generatora liczb pseudo-losowych.
3. Zainicjalizowanie macierzy wejściowych przesłanych do jądra CUDA.
4. Transfer danych do macierzy jądra CUDA.
5. Wykonanie mnożenia macierzy, opatrzone zdarzeniami startu i stopu pomiaru czasu.
6. Zwolnienie pamięci.

Przebieg

Aby wykonać wykresy opisane w poleceniu należało również napisać program mnożący dwie macierze przy użyciu procesora centralnego. Taki program, korzystający z dyrektyw OpenMP, pozwolił na utworzone poniższego wykresu kolumnowego:



Przy tworzeniu takiego programu nie wykonano żadnych optymalizacji; implementacja jest „naiwnym” mnożeniem odpowiadających komórek macierzy.
Program mnożący dwie macierze przy użyciu karty graficznej pozwolił odnotować czasy wykonania rzędu 0.003 sekundy – liczba rzędów wielkości, o jakie wydajność płynąca z użycia procesora karty graficznej przewyższa wydajność płynącą z użycia procesora centralnego jest niesamowita.