CUDA Project

CLE - Assignment 3
June 2023

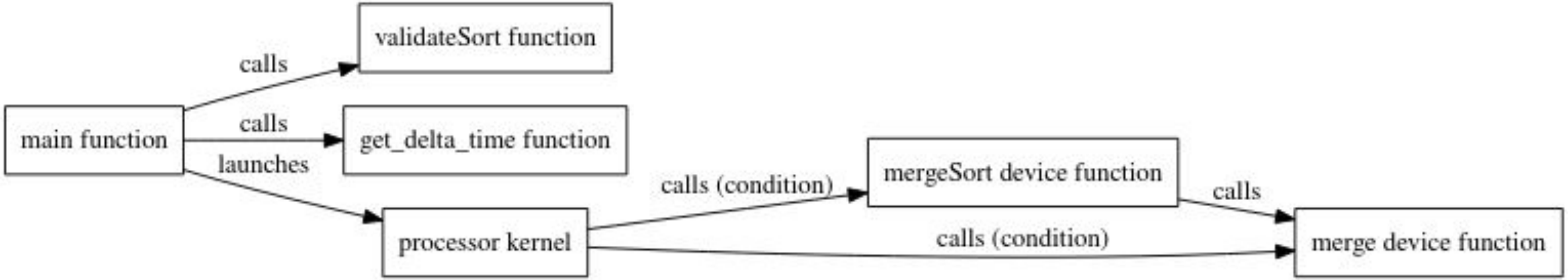
José Trigo 98597
Pedro Monteiro 97484

Implementation

Objective: sort an array of integers in ascending order using the CUDA library to take advantage of the GPU.

1. Read the input binary file containing integers
2. Set Up the Device
3. Copy the data from the host memory to the device memory
4. Define the dimensions for the kernel grid and blocks
5. Invoke CUDA kernel processor to iteratively merge the sorted subsequences
6. Measure the kernel execution time
7. Copy back the the device memory to the host memory
8. Validate the sorted array

Implementation



Row VS Column Processing

Objective: sort an array of integers in ascending order using GPU programming.

1. Row

$$
\begin{aligned}
&0 \le idx < (N >> iter) \\
&subseq \;=\; seq + N * (1 << iter) * idx \quad \Rightarrow \\
&\quad \Rightarrow \quad subseq[i] \;=\; seq[N * (1 << iter) * idx + i]\;, \\
&\qquad\qquad\qquad\qquad \text{with} \quad 0 \le i < (1 << iter) * N
\end{aligned}
$$

2. Column

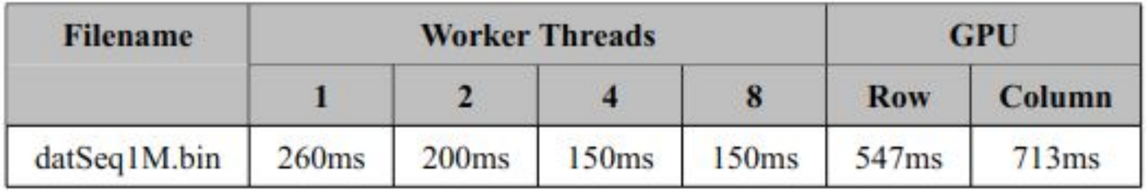
$$
\begin{aligned}
&0 \le idx < (N >> iter) \\
&subseq \;=\; seq + (1 << iter) * idx \quad \Rightarrow \\
&\quad \Rightarrow \quad subseq[i] \;=\; seq[(1 << iter) * idx + N * (i \bmod N) + (i \operatorname{div} N)]\;, \\
&\qquad\qquad\qquad\qquad \text{with} \quad 0 \le i < (1 << iter) * N
\end{aligned}
$$

Results

These are the average processing times (based on 10 runs) for executing the program on datSeq1M.bin for both the CPU and GPU approach. The CPU and GPU specifications are listed below:

- Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz
- Thread(s) per core: 2
- Core(s) per socket: 6
- CPU max MHz: 4100,0000

- NVIDIA Corporation TU116 [GeForce GTX 1660 Ti]



Filename	Worker Threads				GPU	
	1	2	4	8	Row	Column
datSeq1M.bin	260ms	200ms	150ms	150ms	547ms	713ms

Conclusions

Is it worthwhile to use the GPU to solve this kind of problem?

- GPU performance is worse than CPU performance in both row processing and column processing, making the use of GPU in this case not worthwhile.

- Even a single-threaded CPU run is faster than GPU runs because of memory accesses.