HW1 - Accelerators and Accelerated Systems

Shay Agroskin
Tomer Abrahahm

May 18, 2019

Cuda environment overview

- Cuda version

```
# nvcc --version
nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2015 NVIDIA Corporation
Built on Mon_Feb_16_22:59:02_CST_2015
Cuda compilation tools, release 7.0, V7.0.27
```

- GPU name: GeForce GTX 780
- ADD #SM here

GPU serial version

a. Why is atmicAdd is required?
A: Multiple thread can access the same cell in the histogram: if two cells have a value of 200, than both of them will access the 200th cell in the histogram array. To synchronize between them we use atomicAdd

b. How many thread did you use and why?
A: We used 256 threads so that every thread will work on a single cell in the histogram. If we'd use more, we'd have to constrain the number of threads with 'if' condition and thus creating unnecessary divergence.

c. What is the total time run time and the thoughput $\left(\frac{images}{sec}\right)$?
A: We ran the tests on 500 images.
Total run time: $193.53\,msec$
Throughput: $\frac{500}{193.53} = 2.5835\,\frac{images}{sec}$

d. Show a clear screenshot showing the execution of at least two kernel function execution and their respective memory movement.

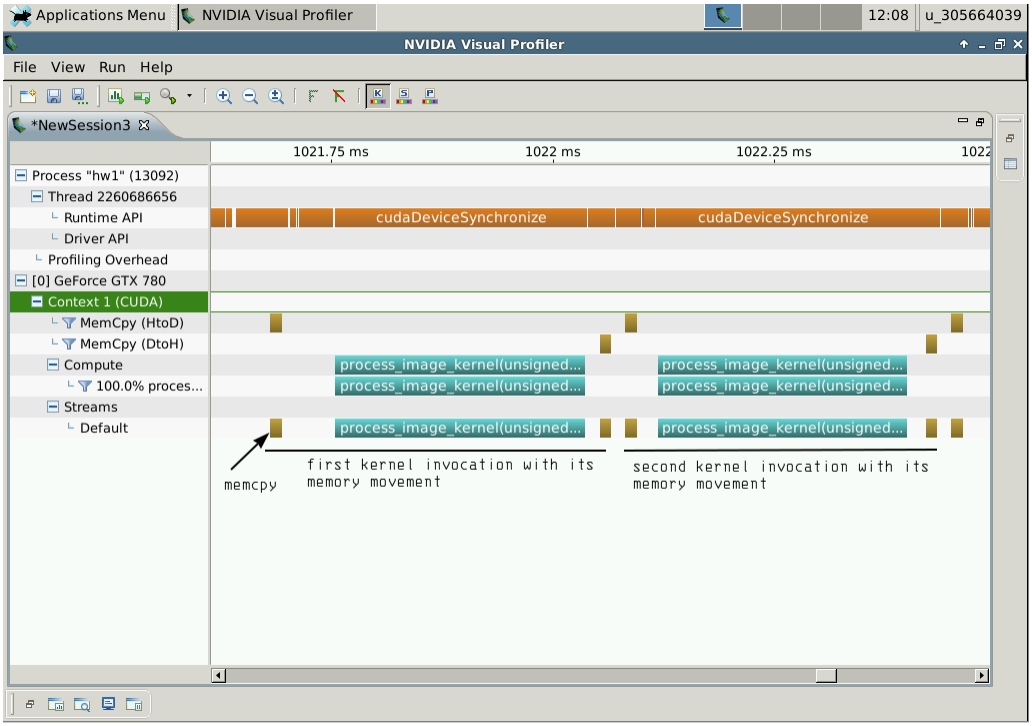

Figure 1: kernel invocation with memory movement, serial execution

e. Choose one 'memcpy' from CPU to GPU, and report its time

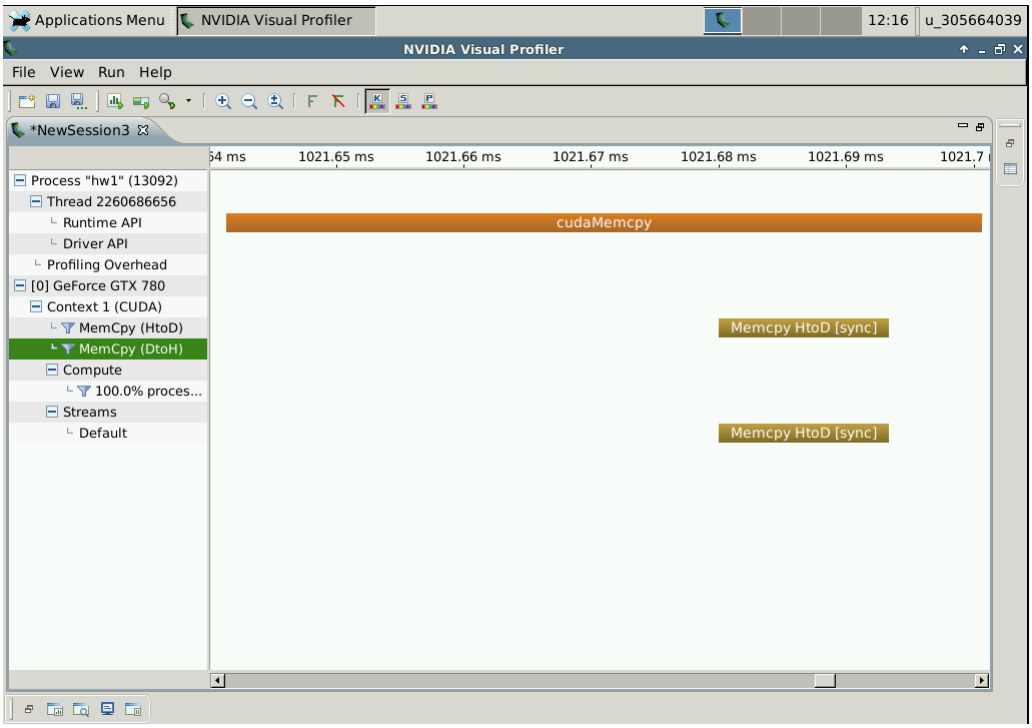

Figure 2: memory copy from CPU to GPU, serial execution

As we can see, the execution time of memory movement from CPU to GPU takes $59.265nsec$ on the CPU side and $13.345nsec$ on the account of GPU (we can deduct that the CPU is the one that makes the memory transfer, and the GPU only handles the synching).

GPU bulk section

a. What is the total execution time and the speedup compared to the serial version?
A: Total execution time is $13.33ms$ which is speedup of $\frac{190.99ms}{13.33ms} = 14.327$ compared to the serial version

b. Attach a clear screenshot of the execution of the bulk section fron the NVIDIA profiler

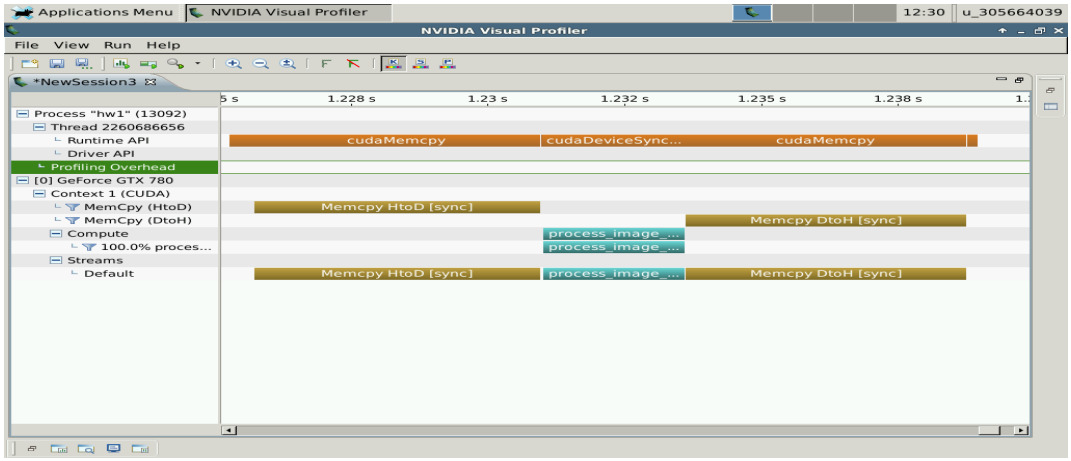

Figure 3: Profiling of the bulk execution

c. Report the time a CPU to GPU 'memcpy' takes. Compare it to the time measured in the serial implementation. Does the time grows linearly with the size of the data being copied ?

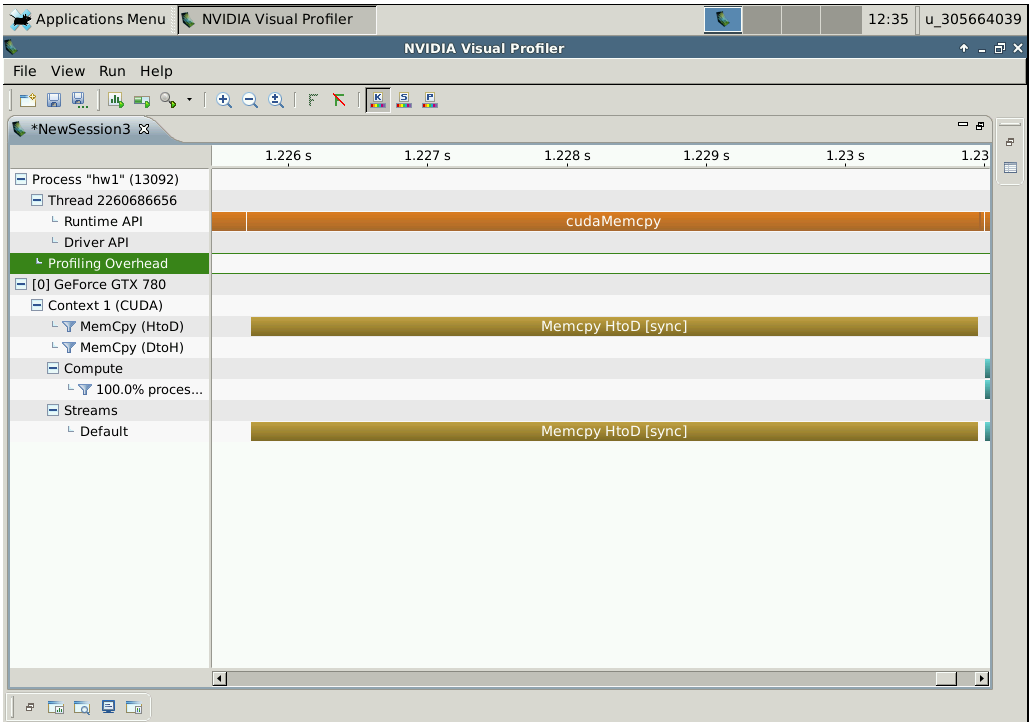

Figure 4: memory copy from CPU to GPU, bulk excution

It takes $5.265msec$ and $5.227msec$ on the CPU and GPU side respectively which is $\frac{5.265msec}{59.265nsec} = 88.84$ and $\frac{5.227msec}{13.345nsec} = 391.68$ times the time it took for a single memory copy in the serial implementation. Since we copy 500 times more pictures, it's clear that the time doesn't grow linearly.