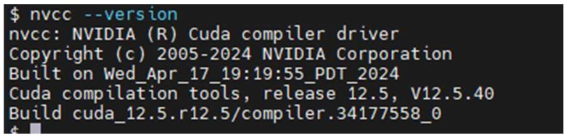
HW1 - הארכה בשל מילואים בצו 8 של שי

Shai Tayar 315034074

Gal Granot 315681593

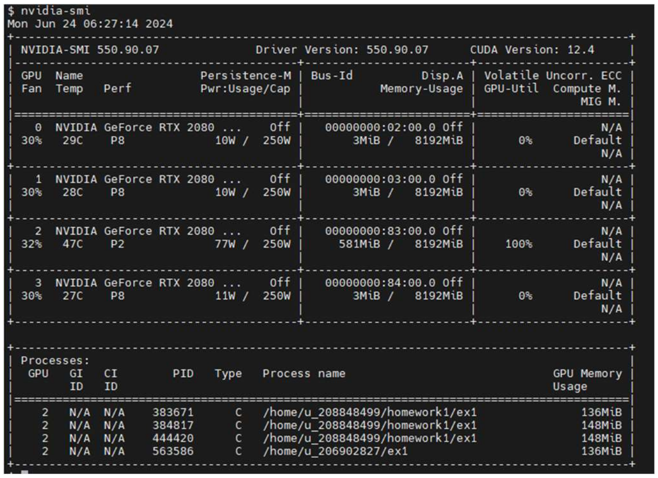
Question 1: knowing the system

a) Cuda version: 12.5.40



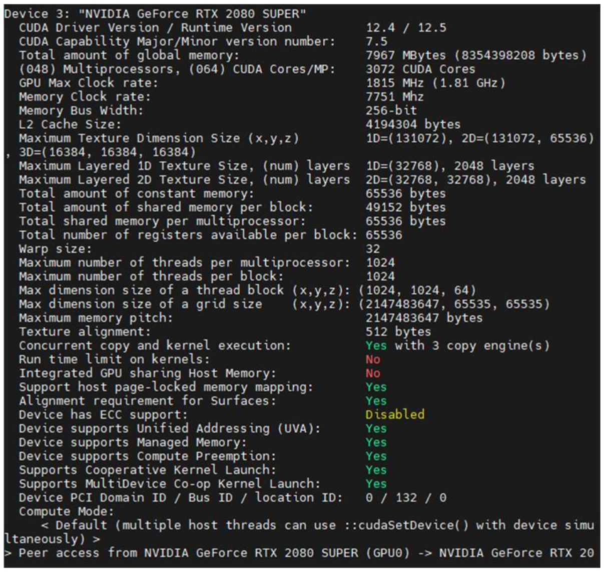
```
$ nvcc --version
nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2024 NVIDIA Corporation
Built on Wed_Apr_17_19:19:55_PDT_2024
Cuda compilation tools, release 12.5, V12.5.40
Build cuda_12.5.r12.5/compiler.34177558_0
$
```

b) GPU name: NVIDIA GeForce RTX 2080



```
$ nvidia-smi
Mon Jun 24 06:27:14 2024
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 550.90.07              Driver Version: 550.90.07      CUDA Version: 12.4     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA GeForce RTX 2080 ...    Off |   00000000:02:00.0 Off |                  N/A |
| 30%   29C    P8             10W /  250W |       3MiB /   8192MiB |      0%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+
|   1  NVIDIA GeForce RTX 2080 ...    Off |   00000000:03:00.0 Off |                  N/A |
| 30%   28C    P8             10W /  250W |       3MiB /   8192MiB |      0%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+
|   2  NVIDIA GeForce RTX 2080 ...    Off |   00000000:83:00.0 Off |                  N/A |
| 32%   47C    P2             77W /  250W |     581MiB /   8192MiB |    100%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+
|   3  NVIDIA GeForce RTX 2080 ...    Off |   00000000:84:00.0 Off |                  N/A |
| 30%   27C    P8             11W /  250W |       3MiB /   8192MiB |      0%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI        PID   Type   Process name                              GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|    2   N/A  N/A    383671      C   /home/u_208848499/homework1/ex1                136MiB |
|    2   N/A  N/A    384817      C   /home/u_208848499/homework1/ex1                148MiB |
|    2   N/A  N/A    444420      C   /home/u_208848499/homework1/ex1                148MiB |
|    2   N/A  N/A    563586      C   /home/u_206902827/ex1                          136MiB |
+-----------------------------------------------------------------------------------------+
```

c) After compiling and running deviceQuery:



```
Device 3: "NVIDIA GeForce RTX 2080 SUPER"
  CUDA Driver Version / Runtime Version          12.4 / 12.5
  CUDA Capability Major/Minor version number:    7.5
  Total amount of global memory:                 7967 MBytes (8354398208 bytes)
  (048) Multiprocessors, (064) CUDA Cores/MP:    3072 CUDA Cores
  GPU Max Clock rate:                            1815 MHz (1.81 GHz)
  Memory Clock rate:                             7751 Mhz
  Memory Bus Width:                              256-bit
  L2 Cache Size:                                 4194304 bytes
  Maximum Texture Dimension Size (x,y,z)         1D=(131072), 2D=(131072, 65536)
, 3D=(16384, 16384, 16384)
  Maximum Layered 1D Texture Size, (num) layers  1D=(32768), 2048 layers
  Maximum Layered 2D Texture Size, (num) layers  2D=(32768, 32768), 2048 layers
  Total amount of constant memory:               65536 bytes
  Total amount of shared memory per block:       49152 bytes
  Total shared memory per multiprocessor:        65536 bytes
  Total number of registers available per block: 65536
  Warp size:                                     32
  Maximum number of threads per multiprocessor:  1024
  Maximum number of threads per block:           1024
  Max dimension size of a thread block (x,y,z): (1024, 1024, 64)
  Max dimension size of a grid size    (x,y,z): (2147483647, 65535, 65535)
  Maximum memory pitch:                          2147483647 bytes
  Texture alignment:                             512 bytes
  Concurrent copy and kernel execution:          Yes with 3 copy engine(s)
  Run time limit on kernels:                     No
  Integrated GPU sharing Host Memory:            No
  Support host page-locked memory mapping:       Yes
  Alignment requirement for Surfaces:            Yes
  Device has ECC support:                        Disabled
  Device supports Unified Addressing (UVA):      Yes
  Device supports Managed Memory:                Yes
  Device supports Compute Preemption:            Yes
  Supports Cooperative Kernel Launch:            Yes
  Supports MultiDevice Co-op Kernel Launch:      Yes
  Device PCI Domain ID / Bus ID / location ID:   0 / 132 / 0
  Compute Mode:
     < Default (multiple host threads can use ::cudaSetDevice() with device simu
ltaneously) >
> Peer access from NVIDIA GeForce RTX 2080 SUPER (GPU0) -> NVIDIA GeForce RTX 20
```

d) The number of multiprocessors is 48.

Question 2: implement device functions

a. We implemented the prefix_sum device function, as we learned in class.

Question 3: Implement a task serial version

a. We implemented the process_image_kernel function in our code.

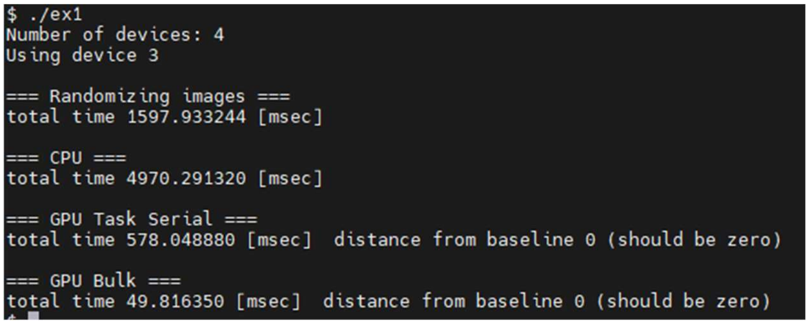
b. We used AtomicAdd in the function that computes the histogram.
In this function, all the threads in the thread block access the same array in shared memory and update it. In order to avoid race condition, in which two thread updates the same cell in the histogram simultaniously but the actual value increases in 1 (instead of 2), we use AtomicAdd instead of histogram[index]++.
AtomicAdd operation ensures that the increment operation is performed atomically, which means only one thread at a time can access the location and perform the addition.

c. The global memory accesses are coalesced regardless of the number of the threads. The image's size is $N \times N$ and $N = k^2 (k \in \mathbb{N}), N \geq 64$. Therefore, the GPU access to the global memory is in full lines, since the threads access contiguous memory locations. Means that the memory accesses are coalesced, regardless the number of threads.

d. The state needed: uchar* all_in, uchar* all_out, uchar* map.
e. We allocated and released the necessary resources in the code.

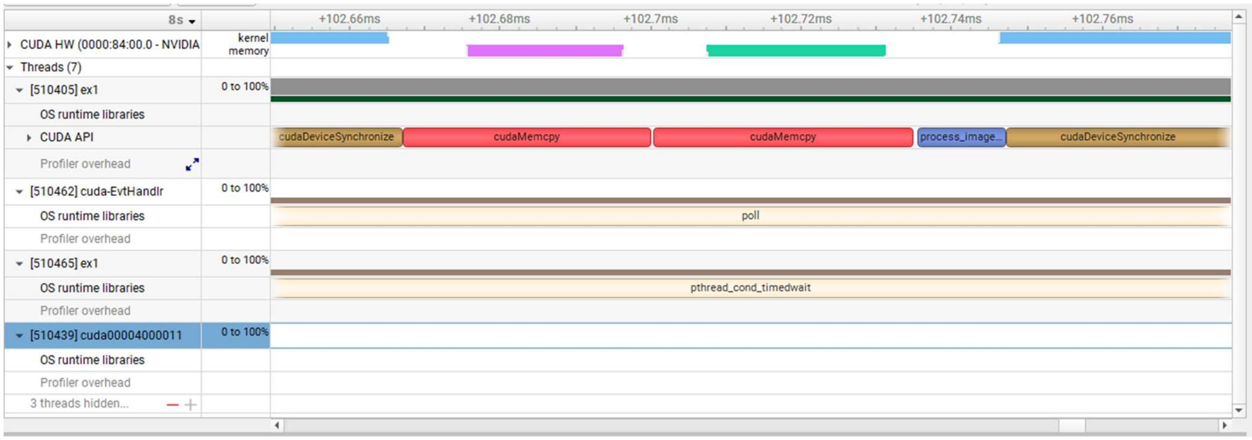
g. Each thread block can contain up to 1024 threads, as seen in question 1c. The number of threads should be a multiple of the tile width, but also not lower than COLOR_RANGE (256).
Moreover, the expression: $\frac{PIXELS\_PER\_TILE}{THREAD\_NUM}$ should be an integer.
We decided to use 256 threads per thread block.

h. The run time is $961.73\ msec$ and the throughput is $\frac{1000}{578.04} \cdot 1000 = 1729.98 \frac{images}{sec}$

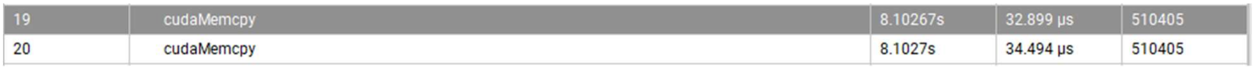


```
$ ./ex1
Number of devices: 4
Using device 3

=== Randomizing images ===
total time 1597.933244 [msec]

=== CPU ===
total time 4970.291320 [msec]

=== GPU Task Serial ===
total time 578.048880 [msec]  distance from baseline 0 (should be zero)

=== GPU Bulk ===
total time 49.816350 [msec]  distance from baseline 0 (should be zero)
$
```

i. Execution diagram:



j. Memcpy from CPU to GPU, duration: 32.899 us



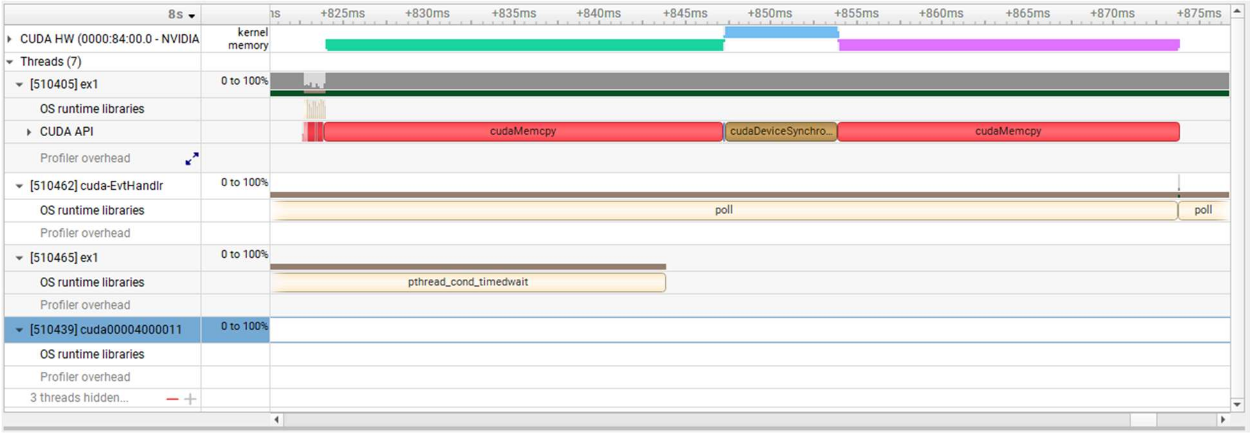
19	cudaMemcpy	8.10267s	32.899 µs	510405
20	cudaMemcpy	8.1027s	34.494 µs	510405

Question 4: Implement a bulk synchronous version

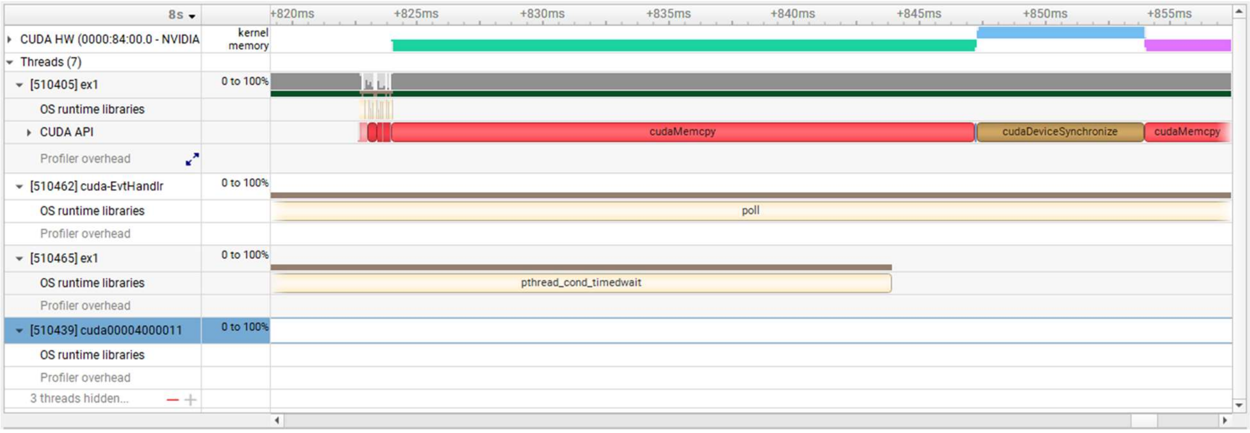
f. From image in Question 3h: Execution time: 49.816 msec.

Speedup: $\frac{578.048}{49.816} = 11.603$

g. Execution diagram:



h. CPU to GPU memcpy in the diagram:



Duration: 28.287 ms

$$\frac{Memcpy_{Bulk}}{Memcpy_{serial}} = \frac{28.287ms}{32.899\ \mu s} = 859.81$$

The bulk version copied X1000 more images, which means the time does not grow linearly with the size of data being copied.