מאיצים חישוביים תרגיל 1 - Report

ברק חפר גחטן 203958442

נדב גלפר 304849821

1.1 CUDA version: release 10.2, V10.2.89

```
u_203958442@gpu-08:~/hw1$ nvcc --version
nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2019 NVIDIA Corporation
Built on Wed_Oct_23_19:24:38_PDT_2019
Cuda compilation tools, release 10.2, V10.2.89
```

1.2 GPU name: GeForce RTX 208

```
u_203958442@gpu-08:~/hw1$ nvidia-smi
Fri Apr 17 14:49:20 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 440.64       Driver Version: 440.64       CUDA Version: 10.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  GeForce RTX 208...  Off  | 00000000:65:00.0 Off |                  N/A |
| 14%   46C    P0    23W / 250W |      0MiB /  7979MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

1.3 לפי הפירוט שקיבלנו מהרצת הפקודה device query:

```
u_203958442@gpu-08:~/hw1/samples/1_Utilities/deviceQuery$ ./deviceQuery
./deviceQuery Starting...

 CUDA Device Query (Runtime API) version (CUDART static linking)

Detected 1 CUDA Capable device(s)

Device 0: "GeForce RTX 2080 SUPER"
  CUDA Driver Version / Runtime Version          10.2 / 10.2
  CUDA Capability Major/Minor version number:    7.5
  Total amount of global memory:                 7980 MBytes (8367439872 bytes)
  (48) Multiprocessors, ( 64) 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 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 Compute Preemption:            Yes
  Supports Cooperative Kernel Launch:            Yes
  Supports MultiDevice Co-op Kernel Launch:      Yes
  Device PCI Domain ID / Bus ID / location ID:   0 / 101 / 0
  Compute Mode:
     < Default (multiple host threads can use ::cudaSetDevice() with device simultaneously) >

deviceQuery, CUDA Driver = CUDART, CUDA Driver Version = 10.2, CUDA Runtime Version = 10.2, NumDevs = 1
Result = PASS
```

1.4 קיימים בכל ליבה 48 multiprocessors, ובסה"כ יש 64 ליבות, כלומר נקבל בסה"כ 3072 multiprocessors. נתונים נוספים שניתן ללמוד- גודל WARP הוא 32 חוטים, גודל threadblock מקסימלי הוא 1024 חוטים, ועל כל multiprocessor ניתן להריץ לכל היותר 1024 חוטים.

3.3 יש 256 תאים בהיסטוגרמה ו- 1024 חוטים שניגשים להיסטוגרמה. לכן נקבל שלפחות לתא אחד יגשו 4 חוטים בו זמנית. כדי להבטיח שכל קידום יתבצע כמצופה, נשתמש בפעולה אטומית. (נשים לב בפעולת ++ רגילה אינה אטומית מפני שהיא דורשת קריאה, עדכון וכתיבה)

3.4 בקוד.

3.5 בקוד.

3.6 בקוד.

3.7 השתמשנו ב- 1024 חוטים בהפעלת הקרנל. בחרנו מספר זה מתוך מטרה למקבל כמה שאפשר פעולות שצריכות להתבצע על כל התמונה. לכן בחרנו במקסימום החוטים שיכולים להיות ב- threadblock.

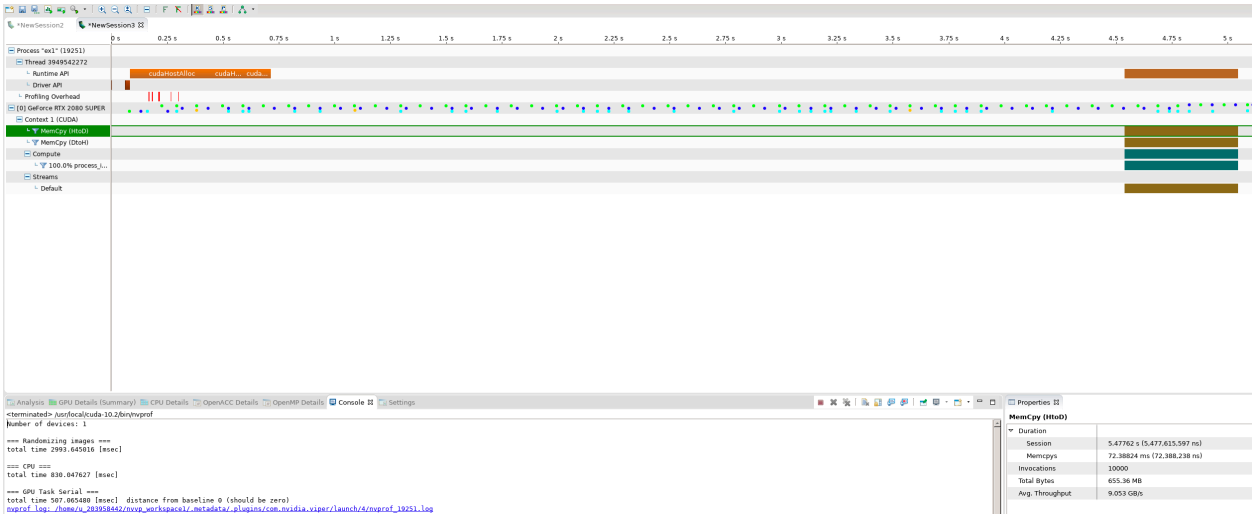
3.8 הזמנים שקיבלנו הם עבור ההרצה הנ"ל

```
Number of devices: 1

=== Randomizing images ===
total time 2993.645016 [msec]

=== CPU ===
total time 830.047627 [msec]

=== GPU Task Serial ===
total time 507.065480 [msec]  distance from baseline 0 (should be zero)
```

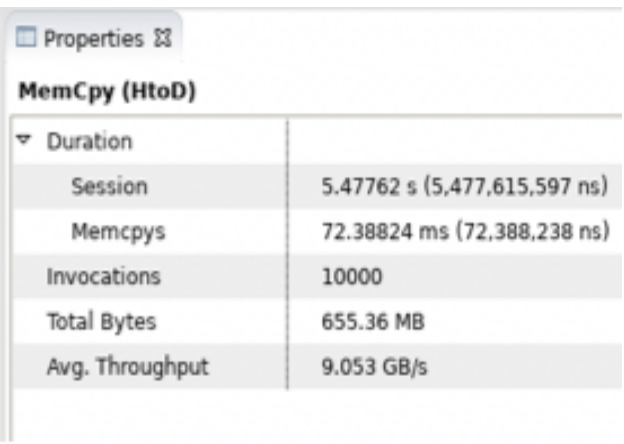
כפי שניתן לראות הזמן הכולל הינו $435.771244\ [msec]$ עבור 10,000 תמונות ולכן

$$\text{ש}throughput = \frac{10000}{0.507065} = 19721\ [\frac{image}{sec}]$$

3.10 + 3.9



אורך פעולת ההעתקה שבחרנו להציג סה"כ $7.712\ \mu s$.



Name	Start Time	Duration	Size	Throughput
Memcpy HtoD [sync]	4.48738 s	7.712 µs	6 kB	8.498 GB/s
Memcpy HtoD [sync]	4.48747 s	7.296 µs	6 kB	8.982 GB/s
Memcpy HtoD [sync]	4.48752 s	7.296 µs	6 kB	8.982 GB/s
Memcpy HtoD [sync]	4.48757 s	7.296 µs	6 kB	8.982 GB/s
Memcpy HtoD [sync]	4.48762 s	7.072 µs	6 kB	9.267 GB/s
Memcpy HtoD [sync]	4.48767 s	7.072 µs	6 kB	9.267 GB/s
Memcpy HtoD [sync]	4.48773 s	7.104 µs	6 kB	9.225 GB/s
Memcpy HtoD [sync]	4.48778 s	7.104 µs	6 kB	9.225 GB/s
Memcpy HtoD [sync]	4.48784 s	7.104 µs	6 kB	9.225 GB/s
Memcpy HtoD [sync]	4.48789 s	7.136 µs	6 kB	9.184 GB/s
Memcpy HtoD [sync]	4.48794 s	7.232 µs	6 kB	9.062 GB/s
Memcpy HtoD [sync]	4.48799 s	7.296 µs	6 kB	8.982 GB/s
Memcpy HtoD [sync]	4.48805 s	7.2 µs	6 kB	9.102 GB/s
Memcpy HtoD [sync]	4.4881 s	7.2 µs	6 kB	9.102 GB/s

4.1 בקוד.

4.2 בקוד.

4.3 בקוד.

4.4 בקוד.

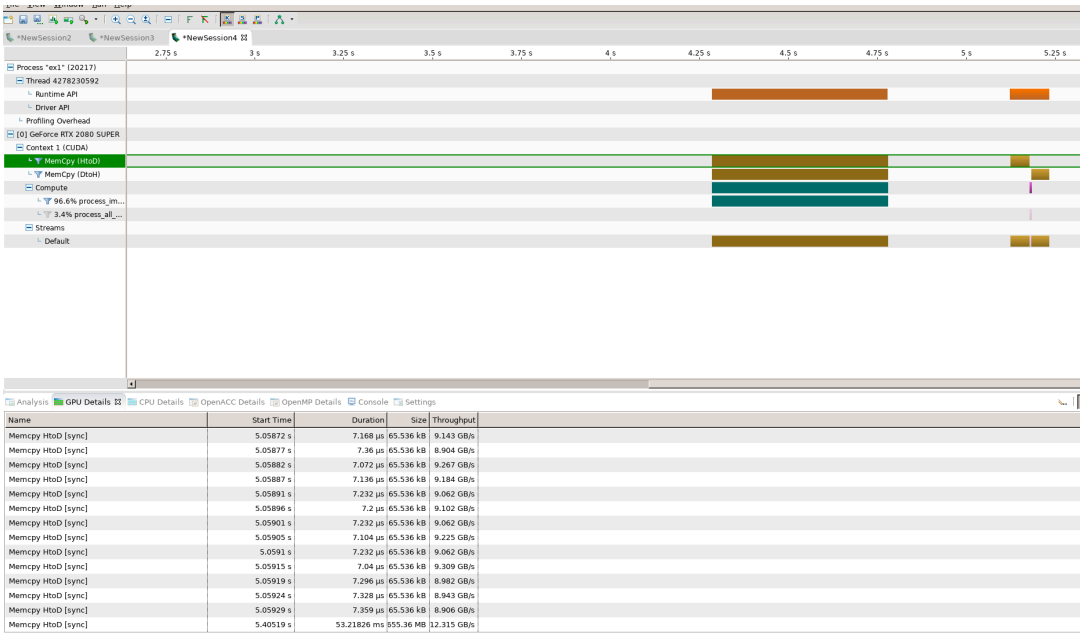
4.5

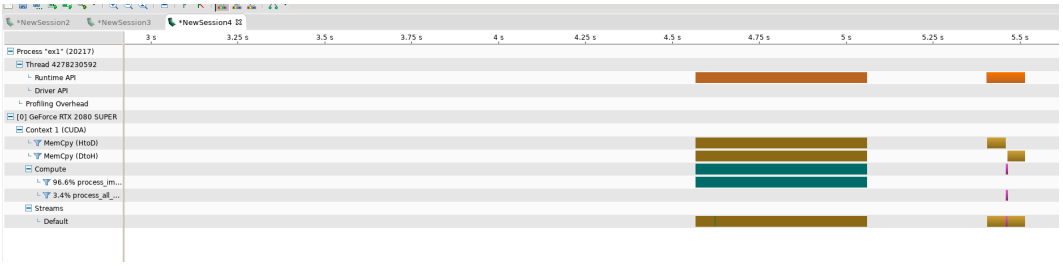
```
Number of devices: 1

=== Randomizing images ===
total time 2984.240230 [msec]

=== CPU ===
total time 830.170821 [msec]

=== GPU Task Serial ===
total time 492.691936 [msec]  distance from baseline 0 (should be zero)

=== GPU Bulk ===
total time 109.361672 [msec]  distance from baseline 0 (should be zero)
```

עפ"י הזמן החדש שקיבלנו נחשב את ה-speedup:

$$GPU\ Bulk\ Time = 109.361\ ms$$
$$Speedup = \frac{GPU\ Serial\ Time}{GPU\ Bulk\ Time} = 4.505$$

4.7+4.6





Name	Start Time	Duration	Size	Throughput
Memcpy HtoD [sync]	5.05872 s	7.168 µs	65.536 kB	9.143 GB/s
Memcpy HtoD [sync]	5.05877 s	7.36 µs	65.536 kB	8.904 GB/s
Memcpy HtoD [sync]	5.05882 s	7.072 µs	65.536 kB	9.267 GB/s
Memcpy HtoD [sync]	5.05887 s	7.136 µs	65.536 kB	9.184 GB/s
Memcpy HtoD [sync]	5.05891 s	7.232 µs	65.536 kB	9.062 GB/s
Memcpy HtoD [sync]	5.05896 s	7.2 µs	65.536 kB	9.102 GB/s
Memcpy HtoD [sync]	5.05901 s	7.232 µs	65.536 kB	9.062 GB/s
Memcpy HtoD [sync]	5.05905 s	7.104 µs	65.536 kB	9.225 GB/s
Memcpy HtoD [sync]	5.0591 s	7.232 µs	65.536 kB	9.062 GB/s
Memcpy HtoD [sync]	5.05915 s	7.04 µs	65.536 kB	9.309 GB/s
Memcpy HtoD [sync]	5.05919 s	7.296 µs	65.536 kB	8.982 GB/s
Memcpy HtoD [sync]	5.05924 s	7.328 µs	65.536 kB	8.943 GB/s
Memcpy HtoD [sync]	5.05929 s	7.359 µs	65.536 kB	8.906 GB/s
Memcpy HtoD [sync]	5.40519 s	53.21826 ms	655.36 MB	12.315 GB/s

כפי שניתן לראות אורך פעולת העתקת הזיכרון כולה אורכת $53.21\ \mu s$ כלומר פי $\frac{53.12\ [msec]}{7.712\ [\mu sec]} = 6887$, כלומר יחסית קרוב לפי $N_{IMAGES} = 10{,}000$, לכן הזמן לא גדל לינארית בדיוק אך דיי קרוב לכך וזאת כיוון שה- $overhead$ של פעולת ה- DMA מבוצע פעם אחת במקום $10{,}000$ פעמים.