יונתן בתן 302279138

אליה בן אברהם 305104580

.1

1.1. להלן הגרסה של cuda עליה אנחנו עובדים

```
u_302279138 hw1 (devel) $ 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
```

1.2. להלן פרטי ה- GPU על המכונה שהוקצתה לנו

```
u_302279138 hw1 (devel) $ nvidia-smi
Fri May  3 17:19:26 2019
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 367.57                 Driver Version: 367.57                    |
|-------------------------------+----------------------+----------------------+
| 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 GTX 780     Off  | 0000:04:00.0     N/A |                  N/A |
| 26%   34C    P8    N/A /  N/A |      2MiB /  3019MiB |     N/A      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID  Type  Process name                               Usage      |
|=============================================================================|
|    0                  Not Supported                                         |
+-----------------------------------------------------------------------------+
```

1.3. בתמונה למטה ניתן לראות פרטים חשובים אודות ה- GPU עליו אנחנו עובדים

- מדובר ב- GeForce GTX 789 (לא פסגת הטכנולוגיה 😉)
- מספר ה- SMs הוא 12

ועוד שלל פרטים חשובים כגון:

- גודל הזיכרון הגלובלי
- מספר החוטים ב- warp וכ"ו

```
u_302279138 deviceQuery (devel) $ ./deviceQuery
./deviceQuery Starting...

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

Detected 1 CUDA Capable device(s)

Device 0: "GeForce GTX 780"
  CUDA Driver Version / Runtime Version          8.0 / 7.0
  CUDA Capability Major/Minor version number:    3.5
  Total amount of global memory:                 3019 MBytes (3165782016 bytes)
  (12) Multiprocessors, (192) CUDA Cores/MP:     2304 CUDA Cores
  GPU Max Clock rate:                            902 MHz (0.90 GHz)
  Memory Clock rate:                             3004 Mhz
  Memory Bus Width:                              384-bit
  L2 Cache Size:                                 1572864 bytes
  Maximum Texture Dimension Size (x,y,z)         1D=(65536), 2D=(65536, 65536), 3D=(4096, 4096, 4096)
  Maximum Layered 1D Texture Size, (num) layers  1D=(16384), 2048 layers
  Maximum Layered 2D Texture Size, (num) layers  2D=(16384, 16384), 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:  2048
  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 1 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 PCI Domain ID / Bus ID / location ID:   0 / 4 / 0
  Compute Mode:
     < Default (multiple host threads can use ::cudaSetDevice() with device simultaneously) >

deviceQuery, CUDA Driver = CUDART, CUDA Driver Version = 8.0, CUDA Runtime Version = 7.0, NumDevs = 1, Device0 = GeForce GTX 780
Result = PASS
```

2. בקוד

.3

3.1. בקוד

3.2. בקוד

3.3. כאשר חוט מסוים רוצה לעדכן את ההיסטוגרמה בהתאם ל-byte שבאחריותו בתמונה הוא צריך לקרוא את הערך הנוכחי לתוך רגיסטר, להוסיף לערך זה 1 ולכתוב חזרה את התוצאה בהיסטוגרמה.

במידה ופעולת ה-increment לא תהיה אטומית סביר שיותר מחוט אחד, אשר רצה הגורל וערך ה-byte שלהם זהה, יקראו את אותו הערך הנוכחי בהיסטוגרמה בו"ז וכאשר יכתבו את התוצאה חזרה הם ידרסו את התוצאה אחד של השני ונקבל תוצאה שגויה.

3.4. בקוד

3.5. בחרנו להשתמש במספר החוטים המירבי של threadBlock, כלומר 1024 חוטים.

עבור החלוקה שעשינו כל חוט אחראי על $\frac{256^2}{1024} = 64$ בתים בתמונה, באופן עקרוני היינו רוצים יותר חוטים על מנת לייעל את החישוב, במצב זה נצטרך לחלק את העבודה בין threadBlocks שונים כך שכל חוט יטפל באיבר יחיד בתמונה.

כיוון שיש 10,000 תמונות אזי בשלב המקבילי (שלב 4 בתרגיל) גם ככה כל הליבות יהיו עסוקות בחישובים ולכן נקבל ביצועים דומים גם ללא פריסה של תמונה יחידה על יותר מ-threadBlock יחיד ולכן בחרנו באופציה זו כיוון שכך המימוש פשוט יותר.

3.6. כפי שניתן לראות הזמן הכולל הינו $1443\ [msec]$ עבור 10,000 תמונות ולכן $throughput = 6930\ [\frac{image}{sec}]$

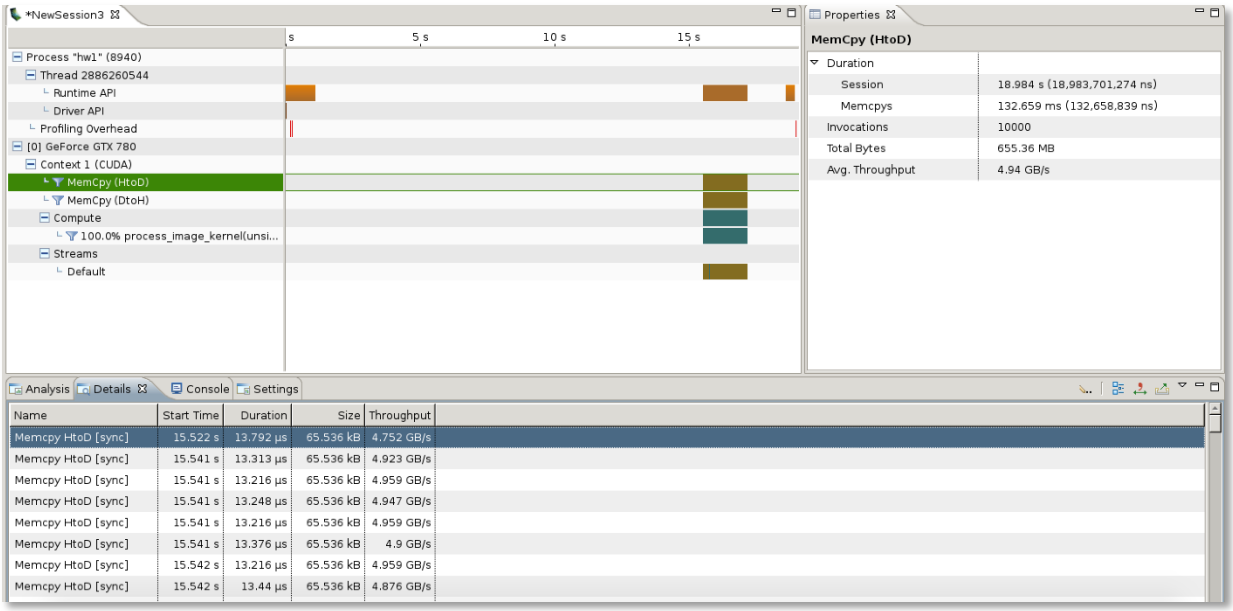
```
u_302279138 hw1 (devel) $ ./hw1

=== CPU ===
total time 1385.219971 [msec]

=== GPU Task Serial ===
total time 1443.549072 [msec]  distance from baseline 0 (should be zero)
```

3.7. התשובה ביחד עם סעיף 3.8

3.8. כפי שניתן לראות אורך פעולת העתקת הזיכרון שבחרנו להציג אורכת $13.79\ [\mu s]$



.4

4.1. בקוד

4.2. בקוד

4.3. בקוד

4.4. כפי שניתן לראות קיבלנו תוצאה של $259\ [msec]$, כלומר $speedup = \frac{1420}{259} = 5.48$ ביחס לסעיף 3.

כעת יש לנו פי 12 יותר ליבות לכן לכאורה היינו יכולים לצפות לקבל $speedup = 12$ אך זה כמובן לא נכון כפי שראינו לפי Amdahl's law שכן יש לקחת בחשבון פעולות העתקת זיכרון, ניהול ה- threadBlocks (איזה threadBlock הולך לאיזה SM) ו- overhead נוסף אשר מקשה עלינו לקבל $speedup$ לינארי.

```
u_302279138 hw1 (devel) $ ./hw1

=== CPU ===
total time 1226.776123 [msec]

=== GPU Task Serial ===
total time 1420.125000 [msec]  distance from baseline 0 (should be zero)

=== GPU Bulk ===
total time 259.284912 [msec]  distance from baseline 0 (should be zero)
```

4.5. התשובה ביחד עם סעיף 4.6

4.6. כפי שניתן לראות אורך פעולת העתקת הזיכרון כולה אורכת $104.79\ [msec]$ כלומר פי $\frac{104.79\ [msec]}{13.74\ [\mu sec]} = 7626$, כלומר יחסית קרוב לפי $N_{IMAGES} = 10{,}000$, לכן הזמן לא גדל לינארית בדיוק אך די קרוב לכך וזאת כיוון שה- $overhead$ של פעולת ה- DMA מבוצע פעם אחת במקום $10{,}000$ פעמים.