Applied GPU Programming

Assignment 1: GPU Programming Environment

November 4th, 2022

Jeremy Williams

Bandwidth Test GPU-CPU on KTH computers

First of all, the following command was executed in the terminal after installing and compiling the CUDA SDK examples shared, as required.

Show info about available GPUs:

$ ***nvidia-smi***

```
Mon Oct 31 22:46:47 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 515.65.01    Driver Version: 515.65.01    CUDA Version: 11.7     |
|-------------------------------+----------------------+----------------------+
| 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 A100-PCI...  Off  | 00000000:43:00.0 Off |                    0 |
| N/A   28C    P0    35W / 250W |      0MiB / 40960MiB |      0%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   1  NVIDIA A100-PCI...  Off  | 00000000:C1:00.0 Off |                    0 |
| N/A   36C    P0    36W / 250W |      0MiB / 40960MiB |      0%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+

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

The **bandwithTest tool** was executed directly via below:

$ ./bandwidthTest

```
[CUDA Bandwidth Test] - Starting...
Running on...

 Device 0: NVIDIA A100-PCIE-40GB
 Quick Mode

 Host to Device Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)    Bandwidth(GB/s)
32000000            21.3

 Device to Host Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)    Bandwidth(GB/s)
32000000            21.3

 Device to Device Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)    Bandwidth(GB/s)
32000000            1160.0

Result = PASS

NOTE: The CUDA Samples are not meant for performance measurements. Results may vary when GPU Boost is enabled.
```

The **bandwithTest** tool was used to measure the memory bandwidth between the CPU (host) and GPU (device) and between GPUs, as well.

This execution was done on an NVIDIA A100-PCIE-40GB graphic card in PDC and KTH Computer Science Department.

Looking our CUDA bandwidth test, we can see that the both "device to host" and "host to device" memory transfer was ***21.3 GB/s*** *each with similar behavior*, while transfers from "device to device" was ***1,160 GB/s.***

What does this mean and why is this happening?

GPUs are classified as specialized electronic circuit created to quickly control and change memory usage, with a computational application, to accelerate the creation of images for a display device. They are used in embedded systems, mobile phones, personal computers, workstations, game consoles and heterogeneous system architectures **[3].**

However, GPUs have an enormous fault of getting data from/to GPU, which means this specialized device still require CPU processing capability, OS, I/O etc.

Data transfer from "device to device" are very stable and fast, because of parallel workloads attempting to maximize total throughput. Additionally, GPUs have a much, much wider buses and higher memory clock rates than any CPU.

Looking at other transfer sizes, we can run the tool in "shmoo" mode:

$ ./bandwidthTest --mode=shmoo

Host to Device:

```
[CUDA Bandwidth Test] - Starting...
Running on...

 Device 0: NVIDIA A100-PCIE-40GB
 Shmoo Mode

.....................................................................
 Host to Device Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)    Bandwidth(GB/s)
1000            0.3
2000            0.6
3000            0.9
4000            1.2
5000            1.4
6000            1.7
7000            2.0
8000            2.2
9000            2.4
10000           2.7
11000           2.9
12000           3.1
13000           3.3
14000           3.5
15000           3.7
16000           4.0
17000           4.1
18000           4.3
19000           4.5
20000           4.7
22000           5.1
24000           5.4
26000           5.7
28000           6.0
30000           6.2
```

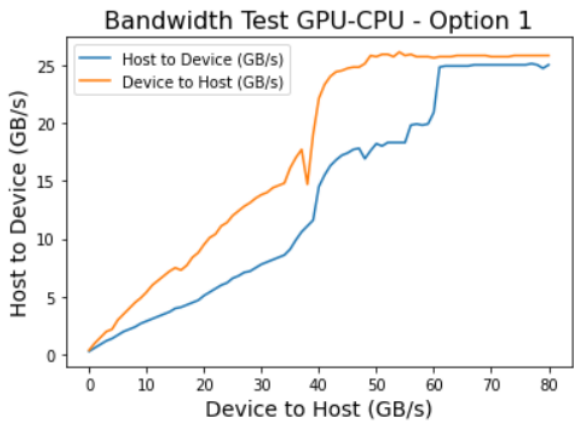
Device to Host:

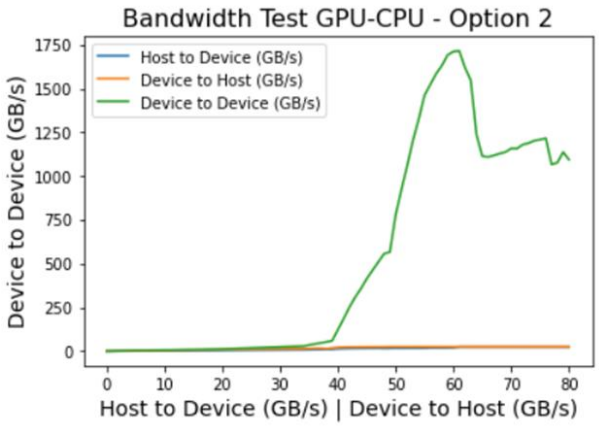
```
93  ..........................................................................
94   Device to Host Bandwidth, 1 Device(s)
95   PINNED Memory Transfers
96     Transfer Size (Bytes)    Bandwidth(GB/s)
97  1000            0.4
98  2000            1.0
99  3000            1.5
100 4000            2.0
101 5000            2.2
102 6000            3.0
103 7000            3.5
104 8000            4.0
105 9000            4.5
106 10000           4.9
107 11000           5.4
108 12000           6.0
109 13000           6.4
110 14000           6.8
111 15000           7.2
112 16000           7.5
113 17000           7.3
114 18000           7.7
115 19000           8.4
116 20000           8.8
117 22000           9.5
118 24000           10.1
119 26000           10.4
120 28000           11.1
121 30000           11.4
122 32000           12.0
123 34000           12.4
124 36000           12.8
125 38000           13.1
126 40000           13.5
```

Device to Device:

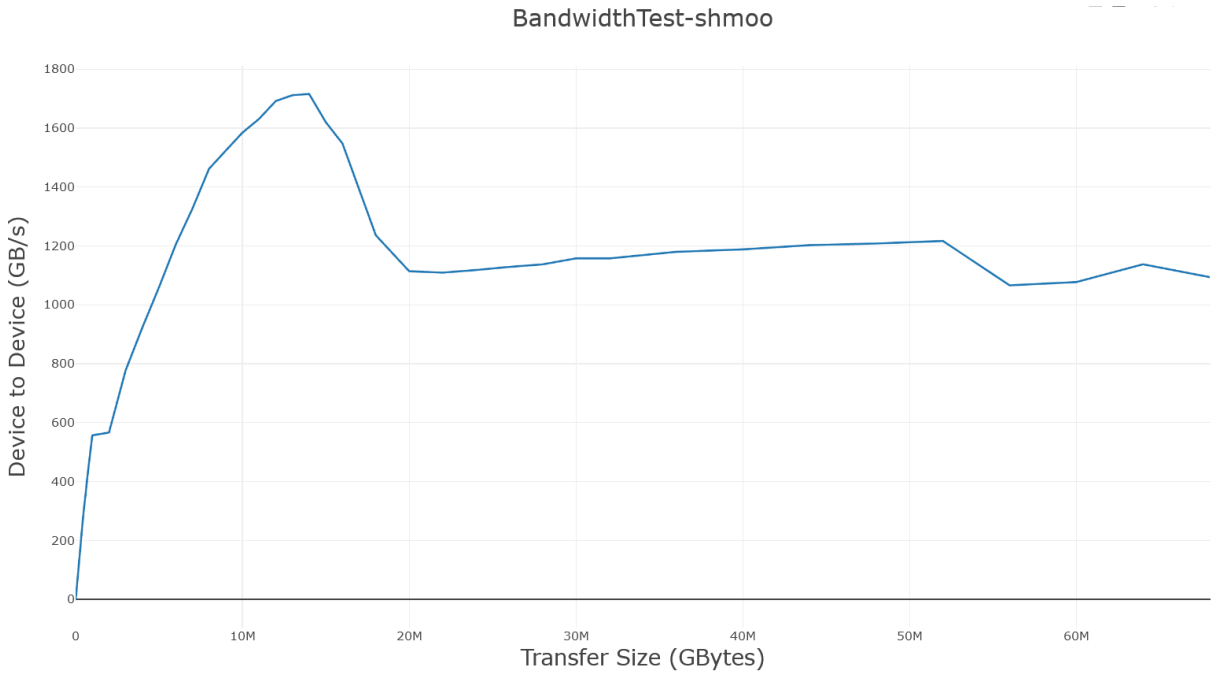
```
179 ..........................................................................
180  Device to Device Bandwidth, 1 Device(s)
181  PINNED Memory Transfers
182    Transfer Size (Bytes)    Bandwidth(GB/s)
183 1000            0.5
184 2000            1.2
185 3000            2.0
186 4000            2.6
187 5000            3.2
188 6000            4.0
189 7000            4.6
190 8000            5.2
191 9000            6.1
192 10000           6.6
193 11000           7.3
194 12000           7.1
195 13000           7.7
196 14000           8.2
197 15000           8.9
198 16000           8.8
199 17000           10.0
200 18000           10.5
201 19000           11.2
202 20000           11.6
203 22000           12.9
204 24000           14.1
205 26000           15.3
206 28000           16.7
207 30000           17.5
208 32000           18.9
209 34000           19.9
210 36000           21.2
211 38000           22.5
212 40000           23.3
```

Looking at the graphs below, we can see that GPU (device) bandwidth is doing well, where the higher the speed the faster the data is sent. Option 1 and 2 shows GPU usage and increase in bandwidth.





The **BandwidthTest-shmoo** graph below shows an increase in bandwidth, as well, however it can be seen in all graphs (Option 1, 2 and below) there seems to be a flattening of GPU capability; which could be lead to over-utilization and reduce levels of data transfer needs.



References

1. Notes on Intro to GPUs (Stefano Markidis and Sergio Rivas-Gomez)
2. Notes on GPU = a Throughput-Oriented Processor (Stefano Markidis and Sergio Rivas-Gomez)
3. **Wikipedia contributors.** "Graphics processing unit." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 4 Nov. 2022. Retrieved Fri. 4 November. 2022.