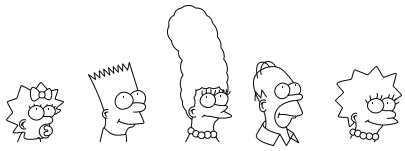
046278
Accelerators and Accelerated Systems
Assigment 2

Avraham Ayaso - 305036352, Yonatan Nakonechny - 200752061

May 25, 2020



1 CUDA Streams

1.2 Run the program in streams mode with load=0 and report the throughput in the report. We'll refer to the throughput you get here as maxLoad.

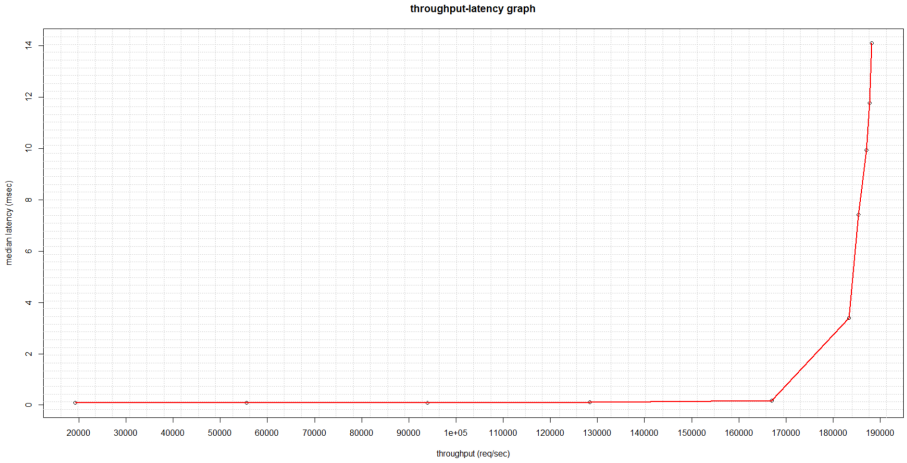
```
u_305036352@gpu-03:~/hw/hw2$ ./ex2 streams 0
Number of devices: 1

=== Randomizing images ===
total time 65.869049 [msec]

=== CPU ===
total time 47.394265 [msec]

=== Client-Server ===
mode = streams
load = 0.0 (req/sec)
distance from baseline 0 (should be zero)
throughput = 186490.5 (req/sec)
latency [msec]:
        avg         min      median  99th perc.         max
     0.0160      0.0121      0.0159      0.0179      0.0535
```

maxLoad=186490.5 (req/sec)

1.3 Vary the load from load=maxLoad/10 to load=maxLoad*2, in 10 equal steps. In each run write down the load, latency and throughput in a table in the report.

load (req/sec)	median latency (msec)	throughput (req/sec)
18649.05	0.0819	19272.8
55947.15	0.0820	55580.7
93245.25	0.0865	93921.6
130543.35	0.1031	128357.0
167841.45	0.1794	166959.2
205139.55	3.3862	183348.8
242437.65	7.4288	185343.4
279735.75	9.9496	187086.8
317033.85	11.7804	187728.8
372981	14.1029	188194.3

1.4 From the samples you collected, draw a throughput-latency graph: X-axis is the throughput, and Y-axis is the median latency. Make sure to annotate the axes with clear names, units and values. Use linear scale for the X-axis. Make sure that the sample points are marked in the graph. Add the graph to the report and explain it (what can we learn from it?)



Conclusions from the graph:

- The throughput increases as we increase the load until we reach a peak around maxLoad. Any increase in the load beyond maxLoad won't affect the throughput.
  - As we found out when we ran the program without limits, maxLoad is the maximum throughput possible, so it makes sense.
- The latency increase rate **increases** as we increase the load.
  - This also makes sense because the increase in load causes jobs to wait longer for an available stream.