EECE 5640: High Performance Computing

Assignment 5

Jiayun Xin

NUID: 001563582

College of Engineering

Northeastern University Boston, Massachusetts

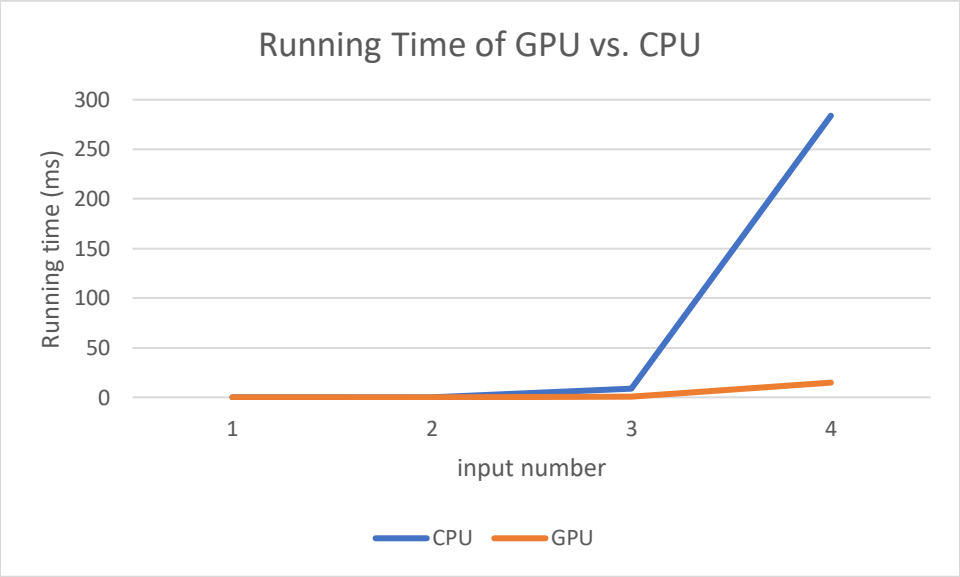
Fall, 2021

1.

a.

```
[xin.ji@c2176 a5]$ ./a.out
45 72 68 61 68 57 71 62 61 58 64 79 65 63 70 60
N = 1024
Check passed? true
GPU Time [ms]: 0.336256 0.021888 0.03824 0.026048
CPU Time [ms]: 0.105088
Speed up [xN]: 0.248769 2.74812
**********
2090 1994 2004 2074 2068 2093 2130 2074 2053 2010 1990 2102 1999 2055 2041 1991
N = 32768
Check passed? true
GPU Time [ms]: 0.295424 0.041088 0.036992 0.017728
CPU Time [ms]: 0.369824
Speed up [xN]: 0.94528 9.99741
**********
65570 65205 65542 65400 65587 65451 65654 65631 65680 65614 65775 65495 65738 65685 65320 65228
N = 1048576
Check passed? true
GPU Time [ms]: 0.53296 0.74832 0.485088 0.02944
CPU Time [ms]: 9.05363
Speed up [xN]: 5.04154 18.6639
**********
2098440 2096786 2100782 2096853 2096277 2096464 2095853 2097162 2097165 2097408 2097150 2096008 2098661 2099920 20
96379 2093091
N = 33554432
Check passed? true
GPU Time [ms]: 0.859648 16.736 14.8407 0.025152
CPU Time [ms]: 283.912
Speed up [xN]: 8.74613 19.1307
**********
```

The screenshot displays the print results of each class and CPU & GPU running time with different input numbers. The four outputs of GPU represent the running time of GPU allocate memory, copy data to device, GPU calculating time and copy data back to host. As we can see, GPU has higher performance with N number increases.



The above screenshot shows comparison between CPU and GPU running time. The horizon number represents N equals 1024, 32768, 1048576 and 33554432 respectively.

b.

```
[xin.ji@c2182 a5]$ ./a.out
61 77 66 61 62 63 60 62 53 72 61 74 67 60 66 59
N = 1024
Check passed? true
GPU Time [ms]: 0.344672 0.022688 0.048544 0.024512
CPU Time [ms]: 0.076608
Speed up [xN]: 0.173945 1.57811
**********************************
1992 2134 2083 2056 2092 2065 2070 2071 2012 2118 2036 1928 2051 1992 1982 2086
N = 32768
Check passed? true
GPU Time [ms]: 0.314272 0.041184 0.036768 0.017248
CPU Time [ms]: 0.06976
Speed up [xN]: 0.170366 1.8973
**********************************
65592 65152 65203 65848 65413 65090 65661 65864 65826 65517 65362 65798 65661 65324 65874 65390
N = 1048576
Check passed? true
GPU Time [ms]: 0.548064 0.671808 0.485344 0.016768
CPU Time [ms]: 1.21696
Speed up [xN]: 0.70672 2.50742
**********************************
2096152 2094090 2098363 2095619 2098706 2098958 2096925 2098839 2098580 2098172 2098836 2096586 2097700 2096817 2095704 2094340
N = 33554432
Check passed? true
GPU Time [ms]: 0.864608 16.3314 14.8195 0.020224
CPU Time [ms]: 39.8572
Speed up [xN]: 1.24414 2.6895
**********************************
```

By using OpenMP, CPU running time decreases. However, GPU still has a better performance than CPU when input number is large enough.

2.

a) Tiled vs untiled

```
GPU Time [ms]: 1.17302 9.62954 0.007616 24.6552
CPU Time [ms]: 179.364
Speed up [xN]: 5.05743 23550.9
GPU Time [ms]: 1.25504 9.42384 0.013696 24.7337
CPU Time [ms]: 179.478
Speed up [xN]: 5.06623 13104.4
```

The four outputs of GPU represent the running time of GPU allocate memory, copy data to device, GPU calculating time and copy data back to host. The first screenshot relevant to tiled implementation has a better GPU performance than the second which is non-tiled implementation.

b)

Guangli Li et al., published an article in 2019. They used a dynamic binary optimization framework to accelerate GPU computing. It is a kind of method to optimize kernels and avoid the high cost of kernel compilation [1]. Their experiment result shows the binary optimization method can accelerate GPU computing and average running time improvement is about 20%.[1]

3.

	Tesla V100	*Tesla P100*
Architecture	Volta	Pascal

Code name	GV100	GP100
Release Year	2017	2016
Cores / GPU	5120	3584
GPU Boost Clock	1530 MHz	1480 MHz
Tensor Cores / GPU	640	NA
Memory type	HBM2	HBM2
Maximum RAM amount	32 GB	16 GB
Memory clock speed	1758 MHz	1430 MHz
Memory bandwidth	900.1 GB / s	720.9 GB / s
CUDA Support	From 7.0 Version	From 6.0 Version
Floating-point performance	14,029 gflops	10,609 gflops

(source: https://www.e2enetworks.com/tesla-v100-vs-tesla-p100-key-differences/)

Comparing Tesla V100 and P100, they have many different key features including cores per GPU, maximum RAM amount, memory clock speed, memory bandwidth, etc but same memory type, HBM2. It is a high capacity and efficient CoWoS stacked memory architecture. V100 is released one year later than P100 with more cores/GPU and higher GPU boost clock, maximum RAM amount, memory clock speed, memory bandwidth and floating-point performance. V100 has better performance than P100 and is more adaptable to execute high-dense calculation.

4.
a.
number of darts: 1000000

b.

```
❯ ./double
3.14168482031249984842702360765 54
❯ ./float
3.14168477058410644531 25
```

References

[1] G. Li, L. Liu and X. Feng, "Accelerating GPU Computing at Runtime with Binary Optimization," 2019 IEEE/ACM International Symposium on Code Generation and Optimization (CGO), 2019, pp. 276-277, doi: 10.1109/CGO.2019.8661168.