EECE 5640: High Performance Computing

Assignment 5

Jiayun Xin

NUID: 001563582

College of Engineering

Northeastern University Boston, Massachusetts

Fall, 2021

1.

a.

Graphical user interface, text, application, email

Description automatically generated

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.

Graphical user interface, text, application, email

Description automatically generated

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

Text, letter

Description automatically generated

Text

Description automatically generated

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.

|  |  |  |
| --- | --- | --- |
| esla Product | 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/)

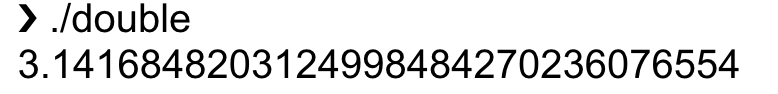
Comparing Tesla V100 and P100, they have many different key features including cores per GPU, maximum RAM amount, memory clock speed, memory bandwidth, etc but same memory type, HBM2. It is a high capacity and efficient CoWoS stacked memory architecture. V100 is released one year later than P100 with 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.



Logo

Description automatically generated with low confidence

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.