Assignment I： GPU Architecture

Group 12 - Chao Xiong, Chang Fu

EX1

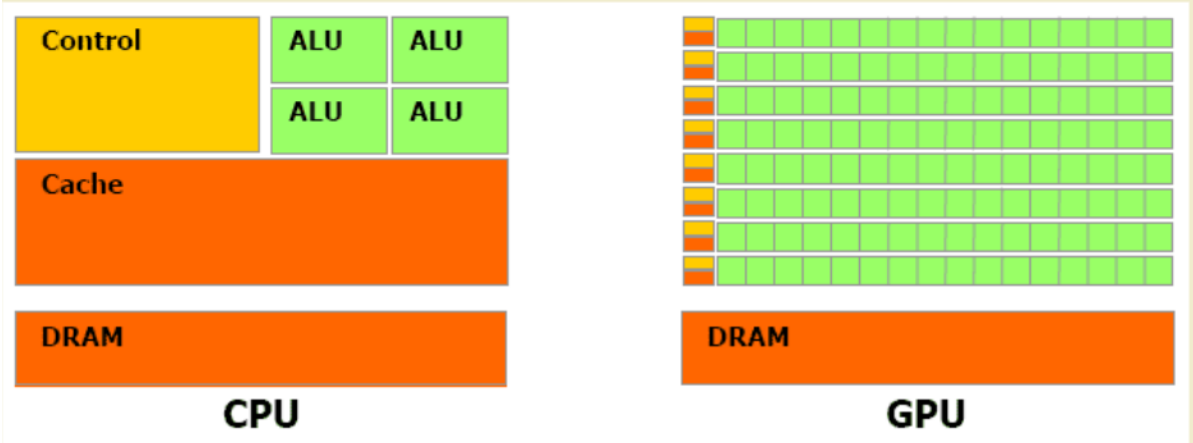
1. Why GPUs emerged as suitable hardware for computer graphics (e.g. games)?

GPUs enable high parallelism ability and calculations in CG such as lighting evaluation can all be computed independently, which makes GPU more powerful and suitable for rendering tasks.

2. Why do we talk about throughput-oriented architecture when we talk about GPUs?

Different from latency-oriented architecture like CPUs, GPUs are used for parallel computations. The number of tasks per unit time can be a good metric to measure the capability of such computations which parallelism is abundant.

3. List the main differences between GPUs and CPUs in terms of architecture.



CPU	GPU
Low compute density	High Compute density
Complex control logic	Simple control logic
Larger caches	Small caches
Shallow pipelines	Deep pipeline
Low latency tolerance	High latency tolerance

GPU Architecture. https://sites.google.com/site/daveshshingari/explorations/computer-architecture/gpu-architecture

4. **Use the Internet to find out and list the number of SMs, the number of cores per SM, the clock frequency, and the size of the memory of the NVIDIA GPU that you plan to use during the course. It might be the GPU of your laptop/workstation, or the GPUs on Tegner (NVIDIA Quadro K420 or NVIDIA Tesla K80). Please, make sure that you mention the specific model as well.**

   Telsa K80:

   - 13 SMs per processor(It has two Gk210s) and 192 cores per SM
   - 562MHz and 875MHz
   - 24G memory.

5. **The new Nvidia GPUs (Volta, Turing, and Ampere) introduce two new kinds of cores, namely the *Tensor* and *Ray Tracing* cores.**
   1. **Which operation does the Tensor core support in one clock cycle?**

      It supports matrix-multiply-and-accumulate operation on a 4×4 matrix in one GPU clock cycle.

   2. **What is the precision supported by Tensor cores?**

      Half precision for the input and matrix multiplication but single precision for accumulation,which is called mixed precision.

   3. **Why do you think that Tensor Cores were introduced in the new GPUs? For graphics or other applications?**

      It is introduced mainly for the sake of deep learning area where matrix computation is very common.

   4. **Check the Google TPU (also an accelerator) characteristic. What are the differences between GPUs with Tensor cores and TPUs?**

   - They are architecturally very different. A GPU is a processor in its own right, while a TPU is a coprocessor, which can't execute code in its own right.

- A GPU with Tensor cores is optimized for vectorized code, which features mixed-precision computing, while a TPU is designed specifically for matrix operation.
- GPUs are designed for graphical computation, while TPUs are designed for machine learning tasks.

5. **Which operation does the Nvidia Ray Tracing core support in one clock cycle?**

   The cores are designed specifically to accelerate Bounding Volume Hierarchy (BVH) traversal and ray/triangle intersection testing functions.

6. **Check the Top500 list that is reporting the 500 most powerful supercomputers in the world. How many of the first 10 most powerful supercomputers use GPUs? Report the name of the supercomputers and the GPU vendor (Nvidia, AMD, ...) and model.**

1. Marconi-100 - IBM Power System AC922, IBM POWER9 16C 3GHz, **Nvidia Volta V100**, Dual-rail Mellanox EDR Infiniband,
2. Selene - DGX A100 SuperPOD, AMD EPYC 7742 64C 2.25GHz, **NVIDIA A100**, Mellanox HDR Infiniband, Nvidia
3. HPC5 - PowerEdge C4140, Xeon Gold 6252 24C 2.1GHz, **NVIDIA Tesla V100**, Mellanox HDR Infiniband, Dell EMC
4. Sierra - IBM Power System AC922, IBM POWER9 22C 3.1GHz**, NVIDIA Volta GV100**, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox
5. Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, **NVIDIA Volta GV100**, Dual-rail Mellanox EDR Infiniband, IBM

7. **Use Google Scholar to find a scientific paper reporting about a work using GPUs in your main domain area (HPC, image processing, machine learning, ...). Report the title, authors, conference name/journal, the GPU type that has been used, and which programming approach has been employed.**

S. McNally, J. Roche and S. Caton, "*Predicting the Price of Bitcoin Using Machine Learning*," 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), Cambridge, 2018, pp. 339-343, doi: 10.1109/PDP2018.2018.00060.

The GPU they used was an NVIDIA GeForce 940M 2GB, which was 67.7% faster than the CPU. The model was built using Keras with TensorFlow backend (which involves cuDNN).

EX2

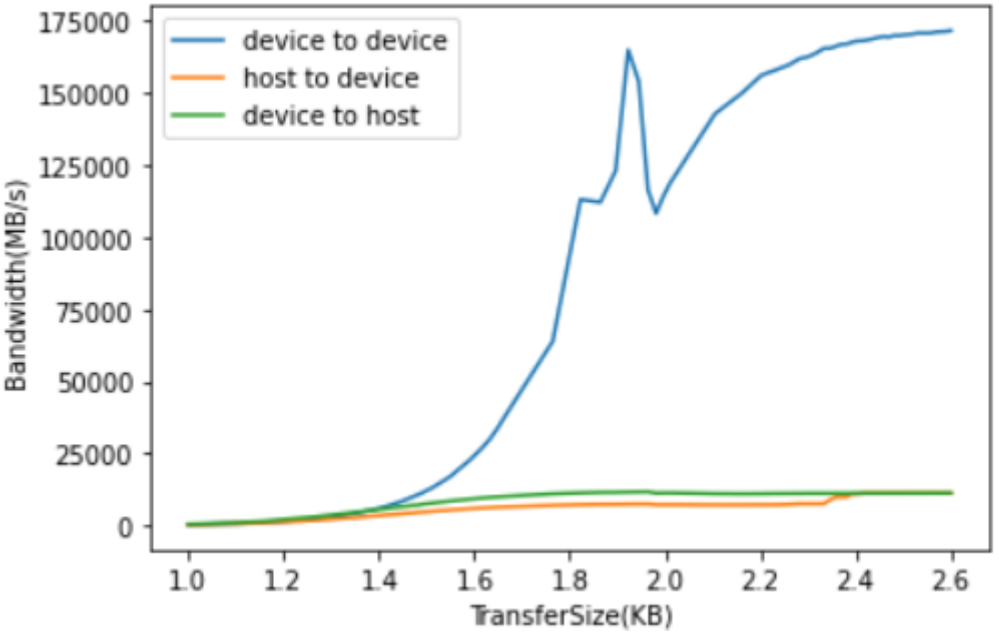
Bandwidth Test

```
Device 0: Tesla K80
Quick Mode

Host to Device Bandwidth, 1 Device(s)
PINNED Memory Transfers
  Transfer Size (Bytes)        Bandwidth(MB/s)
  33554432                     7795.5

Device to Host Bandwidth, 1 Device(s)
PINNED Memory Transfers
  Transfer Size (Bytes)        Bandwidth(MB/s)
  33554432                     9568.8

Device to Device Bandwidth, 1 Device(s)
PINNED Memory Transfers
  Transfer Size (Bytes)        Bandwidth(MB/s)
  33554432                     169779.4
```



From the results above, we can observe that there is a huge gap between the device-to-device bandwidths and others (1) for larger transfer sizes. And the bandwidths between CPU and GPU remain almost constant as the transfer size increases (2). And there is a fluctuation when transferring around 1.9K size files(3).

For (1) and (2), because the transfers within the GDRAM are super-efficient (compared to CPU-related transfers) and as the transfer size increases, GPU will create more threads to conduct the transfer in parallel, which raises the bandwidth.

And the connection between CPU and GPU is limited by the PCIe bus or NVlink (much slower compared to GPU's inner connections) and CPU can create so many threads(usually fixed) as GPU does, so it will only cause bus traffic with increasing transfer size for its serial nature.

For (3), we are not so sure what's happening here. But since this is a GPU-inner issue and we notice that "`L1 Cache is configurable from 16 KB up to 48 KB per SMX`" from the official whitepaper of GK210. Maybe there are lots of cache line switching here.