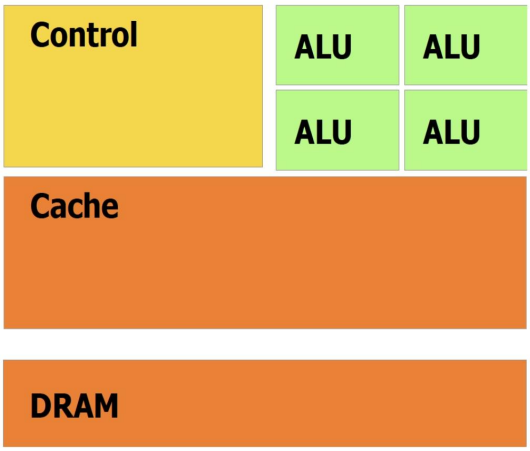
Exercise 1 - Reflection on GPU-accelerated Computing

1. Difference between CPU and GPU Architecture:

CPU:

- Large caches to move the data quickly
- Complicated control units for out-of-order execution.
- Goal is to complete serial tasks as fast as possible.
- Lots of cache
- Not that many ALUs
- Few cores



GPU:

- Designed for simple computing units
- Support high parallel workloads, to compute as much as possible in a specific period of time.
- Much more ALUs for computation
- Not as much cache
- Many cores
- Many simple processing units
- Hardware threads
- SIMD execution

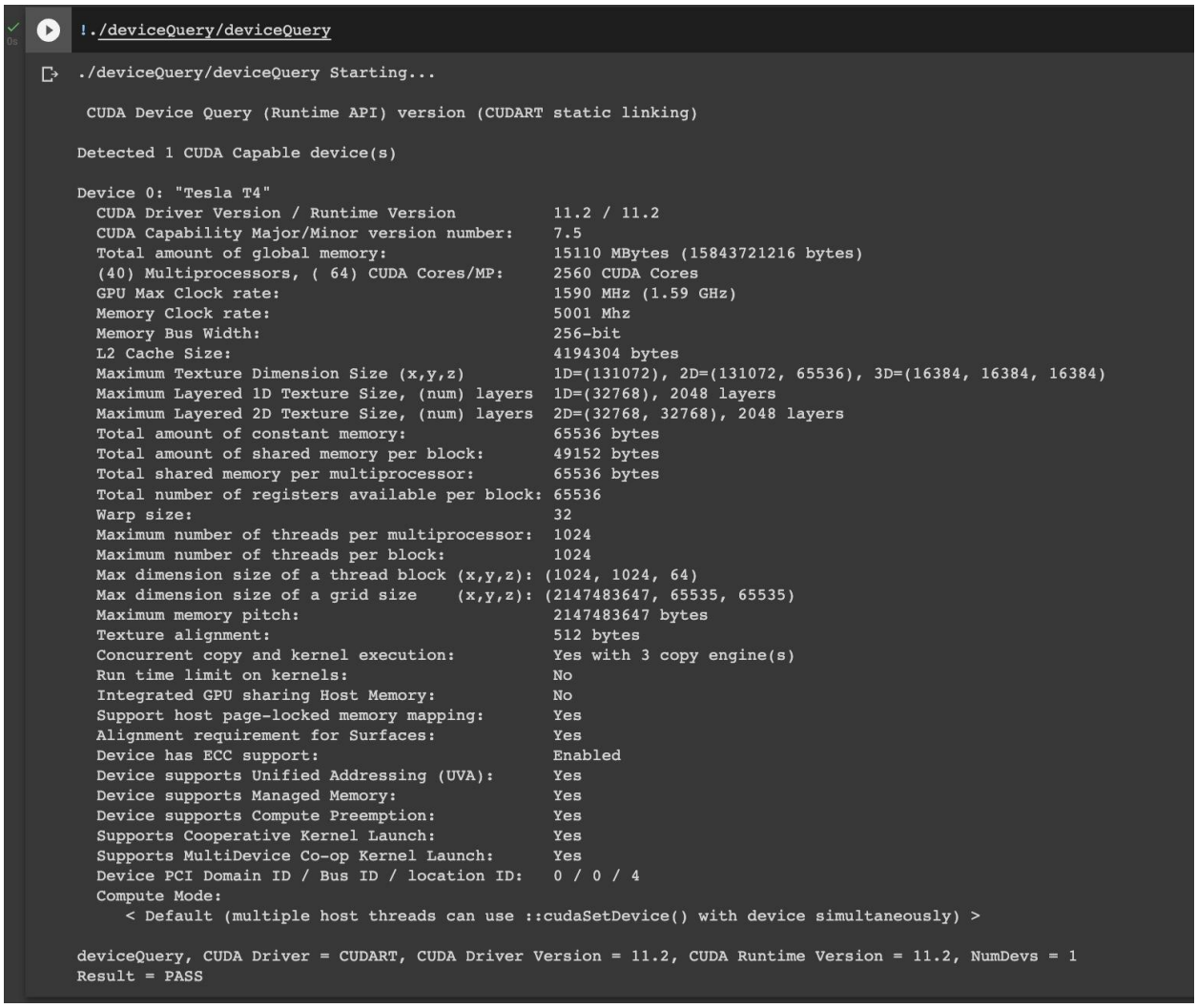
2. After careful look at the top 10 supercomputers I was able to see that 8 out 10 computers use GPU.

SUPER COMPUTER	**GPU VENDOR**
Frontier	AMD
LUMI	AMD
Leonardo	Nvidia
Summit	Nvidia
Sierra	Nvidia
Perlmutter	Nvidia
Selene	Nvidia
Tianhe-2A	NUDT

3.

SuperComputer	Rmax(Pflops/s)	Power(kW)	Power Efficiency(Gflops/watts)
Frontier	1102	21100	52.22748815
Fugaku	442	29899	14.78310311
LUMI	151.9	2942	51.63154317
Summit	148.6	10096	14.71870048
Sierra	94.64	7438	12.7238505
Sunway TaihuLight	93.01	15371	6.05100514
Perlmutter	70.87	2589	27.37350328
Selena	63.46	2646	23.98337113
Tianhe-2A	61.44	18482	3.32431555
Adastra	46.1	921	50.05428882

Exercise 2 - Device Query

1. 

```
!./deviceQuery/deviceQuery

./deviceQuery/deviceQuery Starting...

 CUDA Device Query (Runtime API) version (CUDART static linking)

Detected 1 CUDA Capable device(s)

Device 0: "Tesla T4"
  CUDA Driver Version / Runtime Version          11.2 / 11.2
  CUDA Capability Major/Minor version number:    7.5
  Total amount of global memory:                 15110 MBytes (15843721216 bytes)
  (40) Multiprocessors, ( 64) CUDA Cores/MP:     2560 CUDA Cores
  GPU Max Clock rate:                            1590 MHz (1.59 GHz)
  Memory Clock rate:                             5001 Mhz
  Memory Bus Width:                              256-bit
  L2 Cache Size:                                 4194304 bytes
  Maximum Texture Dimension Size (x,y,z)         1D=(131072), 2D=(131072, 65536), 3D=(16384, 16384, 16384)
  Maximum Layered 1D Texture Size, (num) layers  1D=(32768), 2048 layers
  Maximum Layered 2D Texture Size, (num) layers  2D=(32768, 32768), 2048 layers
  Total amount of constant memory:               65536 bytes
  Total amount of shared memory per block:       49152 bytes
  Total shared memory per multiprocessor:        65536 bytes
  Total number of registers available per block: 65536
  Warp size:                                     32
  Maximum number of threads per multiprocessor:  1024
  Maximum number of threads per block:           1024
  Max dimension size of a thread block (x,y,z): (1024, 1024, 64)
  Max dimension size of a grid size    (x,y,z): (2147483647, 65535, 65535)
  Maximum memory pitch:                          2147483647 bytes
  Texture alignment:                             512 bytes
  Concurrent copy and kernel execution:          Yes with 3 copy engine(s)
  Run time limit on kernels:                     No
  Integrated GPU sharing Host Memory:            No
  Support host page-locked memory mapping:       Yes
  Alignment requirement for Surfaces:            Yes
  Device has ECC support:                        Enabled
  Device supports Unified Addressing (UVA):      Yes
  Device supports Managed Memory:                Yes
  Device supports Compute Preemption:            Yes
  Supports Cooperative Kernel Launch:            Yes
  Supports MultiDevice Co-op Kernel Launch:      Yes
  Device PCI Domain ID / Bus ID / location ID:   0 / 0 / 4
  Compute Mode:
     < Default (multiple host threads can use ::cudaSetDevice() with device simultaneously) >

deviceQuery, CUDA Driver = CUDART, CUDA Driver Version = 11.2, CUDA Runtime Version = 11.2, NumDevs = 1
Result = PASS
```

2.

```
GPU Max Clock rate:                            1590 MHz (1.59 GHz)
```

   - Speed of the GPU's clock. This affects how fast the gpu can solve tasks.

```
L2 Cache Size:                                 4194304 bytes
```

   - this is the shared memory that's shared by all SMs in GPU

```
Total amount of global memory:      15110 MBytes (15843721216 bytes)
```

   - This is the main memory space that's shared between the Host and the GPU. This particular GPU has around 15 GB.

3. **Memory Bandwidth** is the theoretical maximum amount of data that the bus can handle at any given time, playing a determining role in how quickly a GPU can access and utilize its framebuffer

```
Memory Bus Width:                              256-bit
Memory Clock rate:                             5001 Mhz
```

```
Effective mem clock rate = 5001*2 = 10002 bits
Memory bandwidth = Eff. mem clock * memory bus rate/8 =
10002*(256/8) = 320 Gb/s
```

4. The Memory bandwidth is consistent with what I found on the internet, it seems just a little bit higher but the calculated number supposed to be the theoretical limit.

Exercise 3 - Compare GPU Architecture

List the number of SMs, the number of cores per SM, the clock frequency and calculate their theoretical peak throughput.

Turing Architecture:

- Device: NVIDIA RTX A6000
    1. Changes: L2 Cache: 6 MB, Tensor Cores: 336, Memory: 48 GB, Memory bus width: 384 bits.
    2. Number of cores per SM: 128, Clock frequency: 1410 MHz, Performance for FP64(double) is 1,210 GFLOPS
    3. L2 caches: 6 MB, Tensor Cores: 320, Memory: 16 GB, Memory bus width: 256 bits, Number of corese per SM: 64, Clock frequency: 1.59 GHz, Performance for FP64(double): 254.4 GFLOPS

Ampere:

- Device: NVIDIA A100
    1. Changes: L2 Cache: 40 MB, Tensor Cores: 432, Memory: 80 GB, Memory bus width: 5120
    2. Number of cores per SM: 32, Clock frequence: 1410 MHz, Performance for FP64(double): 9.6 TFLOPS
    3. L2 caches: 6 MB, Tensor Cores: 320, Memory: 16 GB, Memory bus width: 256 bits, Number of corese per SM: 64, Clock frequency: 1.59 GHz, Performance for FP64(double): 254.4 GFLOPS

Hopper:

- Device: NVIDIA H100 Tensor
    1. Changes: L2 Cache: 50 MB, Tensor Cores: 456, Memory: 80 GB, Memory bus width: 5120 bits
    2. Number of cores per SM: 64, Clock frequency: 1755 MHz, Performance for FP64(double): 25.6 TFLOPS
    3. L2 caches: 6 MB, Tensor Cores: 320, Memory: 16 GB, Memory bus width: 256 bits, Number of corese per SM: 64, Clock frequency: 1.59 GHz, Performance for FP64(double): 254.4 GFLOPS

Exercise 4 - Rodinia CUDA benchmarks and Profiling

Application: k-Nearest Neighbors
Dwarf: Dense Linear Algebra

1) Compiled successfully, no changes to makefile were necessary
2)
- CUDA: time elapsed 6.432 us.
- OpenMP: time elapsed 0.027 s.

3) Yes, it makes sense that CPU is going to take slower because as previously discussed that GPU are more a parallel throughput rather then solving it step by step like CPU does.

Application: bfs
Dwarf: Graph Traversal

1) Compiled successfully, no changes to makefile were needed.
2)
- CUDA: time elapsed 0.00354073 s
- OpenMP: time elapsed 0.069879 s.

3) It still does make sense that GPU is taking a shorter time because of the parallelism of the device in comparison to CPU.

Application: PathFinder
Dwarf: Dynamic Programming

1) Compiled successfully, no changes to makefile were necessary
2)
- CUDA: time elapsed 396.73 us.
- OpenMP: time elapsed 16.1 ms.

3) Makes sense that CUDA executes way faster than openMP, gpu processes faster with the parallelism.