作业(2)： DLP & GPU

截至时间：2022.11.25/周五 23:59:59
提交方式：超算习堂（https://easyhpc.net/course/157）

Q1-Sol:

Peak throughput = 1.5 x 16 x 16 = 384 GFLOPS/s (or, 1.5 x 16 x 16 x 2 = 768 GFLOPS/s)
Assume each single-precision operation requires 2 four-byte operands and outputs 1 four-byte result, sustaining this throughput would require 12 bytes/FLOP x 384 GFLOPS/s = 4.6 TB/s of bandwidth. As such, the throughput is not sustainable on the 100 GB/s off-chip bandwidth, but can still be achieved in short bursts when using on-chip memory.

Q2-Sol:

a. 4 x (2048 / 16) = 4 x 128 = 512
b. 16KB / 4KB = 4
c. register spilling, leading to use local memory which is likely to downgrade performance.
d. 50%, only 16 blocks can be supported. 16x16 / 512 = 50%
e. <<<16, 32>>>, <<<8, 64>>>, <<<4, 128>>>

Q3-Sol:

Any differences on core/instruction/link are acceptable. Here are some examples:

a. new 8-bit FP8 floating point format on H100 tensor core
b. new DPX instructions to accelerate dynamic programming
c. distributed shared memory
d. larger L2 cache
e. new thread block cluster
f. HBM3 memory system
g. Faster NVLink
h. … …

Q4-Sol:

N.A.