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: (P364, 4.16) [10] <4.4> Assume a hypothetical GPU with the following characteristics:

- Clock rate 1.5 GHz
- Contains 16 SIMD processors, each containing 16 single-precision floating- point units
- Has 100 GB/s off-chip memory bandwidth

Without considering memory bandwidth, what is the peak single-precision floating-point throughput for this GPU in GLFOP/s, assuming that all memory latencies can be hidden? Is this throughput sustainable given the memory bandwidth limitation?

Q2: Assume a GPU architecture that contains 4 SMs (or CUs), each having 2048 registers and 16KB shared memory (SMEM). You are to launch onto the GPU one kernel, which declares an array of 4KB SMEM usage, and is compiled to use 16 registers per thread. Format of kernel launch: kernel_name<<<nblocks, blocksize>>>(args).

a. From the register perspective, how many threads can concurrently run at maximum?
b. How many blocks can be executed on each SM at maximum?
c. Suppose each thread of the kernel necessities 32 registers, and what will happen if you keep the number of threads as what you got in (a)?
d. Suppose launching parameters are <<<32, 16>>>, please calculate the GPU occupancy.
e. How to set the parameters (i.e., <<<nblocks, blocksize>>>) to maximize GPU utilization?

Q3: Please read the documents below, and list five or more architectural changes on H100 vs. A100.

- NVIDIA H100 Tensor Core GPU Architecture, https://resources.nvidia.com/en-us-tensor-core
- NVIDIA A100 Tensor Core GPU Architecture, https://resources.nvidia.com/c/ampere-architecture-white-paper?x=sfvhf4&xs=169656

Q4: Please review the paper below following the required format.
Paper:
Oreste Villa, Daniel Lustig and Zi Yan *et al.*, Need for Speed: Experiences Building a Trustworthy System-Level GPU Simulator, IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2021.
链接：https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9407154

Format:

A. Paper summary

Please provide a short summary of the paper that captures the key contributions.

==

B. Key strengths and weaknesses

Please provide up to three strengths and three weaknesses, in the form of short (+) and (-) bullets, respectively.

==

C. Comments to authors

Please provide detailed comments that support your above judgement, as well as constructive feedback to make the paper stronger. This should constitute the meat of your review.

Review 撰写参考：

[1]. O. Mutlu, Guidelines on Paper Reviews, https://course.ece.cmu.edu/~ece740/f13/lib/exe/fetch.php?media=onur-740-fall13-lecture0-3-how-to-do-the-paper-reviews.pdf

[2]. S. Krishnamurthi, How to Write Technical Paper Reviews, https://cs.brown.edu/~sk/Memos/Paper-Reviews/