DCS5637 高级计算机体系结构 （Advanced Computer Architecture）, Sun Yat-sen University, Fall 2022

作业：DLP & GPU

截止时间：12 月 14 日， 23:59

提交方式：email (guoych37@mail2.sysu.edu.cn)

Q1: 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?

Q2: Please read the documents below:

- 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

a. List five or more architectural changes on H100 vs. A100
b. What is asynchronous synchronization?
c. Why NVLink is favored over PCIe? What are the NVLink/PCIe bandwidth numbers in H100 and A100?

Q3: Please review the paper below following the required format.
Paper:
Mingcong Han, Hanze Zhang and Rong Chen *et al.*, Microsecond-scale Preemption for Concurrent GPU-accelerated DNN Inferences, USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2022.
链接：https://www.usenix.org/system/files/osdi22-han.pdf

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/