作业（3）：GPU Profiling

截止时间：12 月 26 日，23:59

提交方式：超算习堂（https://easyhpc.net/course/133）

本练习目的是对 GPU 程序调优（profiling）过程进行熟悉；鉴于并非所有同学都有 GPU 或相同型号的 GPU 可以访问，本练习提供使用 NVIDIA Nsight Compute 工具对以下应用在 A100 GPU 上收集的 profiling 结果（见压缩包 data_csv.zip 中.csv 文件）。你所需要做的是对这些提供的结果文件进行分析，包括数据抽取、汇总及画图等。

涉及到的应用均来自于 rodinia benchmark[3][4]，具体为：

序号	应用名称	结果文件	应用描述
1	BFS	bfs.csv	广度优先搜索算法模拟
2	BP	bp.csv	反向传播算法
3	CFD	cfd.csv	计算流体力学模拟
4	Hotspot	hotspot.csv	二维温度传播模拟
5	K-Means	kmeans.csv	K-均值算法
6	MM	mm.csv	矩阵乘法
7	SRAD	srad.csv	图像斑点去除算法

给定上述应用程序的 profiling 结果 文件(.csv)，你需要：

a. [10 分] 分析每个应用中的核函数启动参数（gridsize，blocksize），以表格或柱状图方式展示；

b. [20 分] 对于包含多个核函数的应用，分析每个核函数的运行时间占比，画出饼图，并找出该程序的热点函数（即最耗时函数）；

c. [35 分] 对应用指标进行分析，可供参考的指标：

nvprof metric	ncu metric
ipc	sm__inst_executed.avg.per_cycle_active
dram_utilization	gpu__dram_throughput.avg.pct_of_peak_sustained_elapsed
achieved_occupancy	sm__warps_active.avg.pct_of_peak_sustained_active
sm_effiency	sm__cycles_active.avg

各个指标的含义见参考链接 [1] [2]，分析不同应用之间的指标差异，进而推测该应用的类型（计算密集型 或 访存密集型 或延迟型），有些应用可能兼具多种属性，分析的指标不局限于以上所提供，也可以根据 csv 文件中其他的指标进行分析，给出合理解释即可；

d. [35 分] 根据上述指标，选择其中的任意两个应用，分析应用的瓶颈和可能优化的方向，可以结合代码进辅助分析；

e. [选做题，10 分] 理解 occupancy 的含义 [5]，结合核函数的启动线程数量、共享内存大小、寄存器大小以及 Ampere 架构下的最大硬件支持 [6]，选择任意两个应用分析 occupancy 低下或者可以被充分利用的原因；

f. [选做题，10 分] 自己选择一个 GPU 任务（不局限于 Rodinia [4]），进行分析，包括指令执行，访存情况，性能瓶颈。

参考资料：

[1]. NVIDIA nvprof 工具性能指标及其含义参考：https://docs.nvidia.com/cuda/profiler-users-guide/index.html#metrics-reference-7x

[2]. NVIDIA 新版 Profiling 工具 Nsight Compute 指标与 nvprof 指标对应关系：https://docs.nvidia.com/nsight-compute/NsightComputeCli/index.html#nvprof-metric-comparison

[3]. S. Che, M. Boyer *et al.* Rodinia: A Benchmark Suite for Heterogeneous Computing：https://www.cs.virginia.edu/~skadron/Papers/rodinia_iiswc09.pdf

[4]. Rodinia benchmark：https://github.com/yuhc/gpu-rodinia

[5]. Achieved occupancy 解释：https://docs.nvidia.com/gameworks/content/developertools/desktop/analysis/report/cudaexperiments/kernellevel/achievedoccupancy.htm

[6]. NVIDIA Ampere architecture whitepaper：https://images.nvidia.cn/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf（重点关注 P36 Table 4 和 P43 Table 5）

★Acknowledgements ★：
Thanks Yue Weng and Xuanteng Huang for collecting the data and composing the homework.
It would not be possible to have the document as-is without their much time and effort.