课程大作业二基础题:GPU学习智能体（agent）

- ✓选择一个赛道：摩尔、沐曦、并行；
- ✓准备数据集：
  - ✓GPU编程问答对：基于本年度GPU编程教材（Programming Massively Parallel Processors: A Hands-on Approach，2017）生成问答对；
  - ✓基础题：教材的蓝色画勾部分
  - ✓加分题：教材的红色画勾部分

课程大作业二基础题:GPU学习智能体（agent）

- ✓选择一种训练好的开源大模型：Llama、Qwen、DeepSeek、等；
- ✓选择一种大模型微调框架：
  - ✓摩尔赛道：Llama Factory、等；
  - ✓沐曦赛道：Unsloth、等；
  - ✓并行赛道：Unsloth、等；

课程大作业二基础题:GPU学习智能体（agent）

✓采用低秩适应（Low-Rank Adaptation，LoRA），通过大模型微调框架，基于你编写的GPU编程问答对，对你选的开源大模型进行微调，获得微调后的大模型文件（Hugging Face Transformers 标准格式）；

课程大作业二:GPU学习智能体（agent）

- ✓选择一种推理框架：vllm、SGLang、等，让你的GPU学习智能体工作起来哦；
  - ✓基础题：非必须用推理框架；
  - ✓加分题：采用一种推理框架。

课程大作业二加分题:GPU学习智能体（agent）

- ✓所有有利于提高考核指标成绩的方法，包括但不限于：
  - ✓采用RAG、SFT、PPO等方法，提高大模型微调的质量，提高准确率；
  - ✓采用你在本课程上学到的GPU编程的语言，包括：CUDA、PTX、Triton、TileLang、等，加速推理框架的算子，提高速度；
  - ✓可以做一个UI界面，让你的GPU学习智能体更吸引人；
  - ✓可以给你的GPU学习智能体取一个响亮的名字。

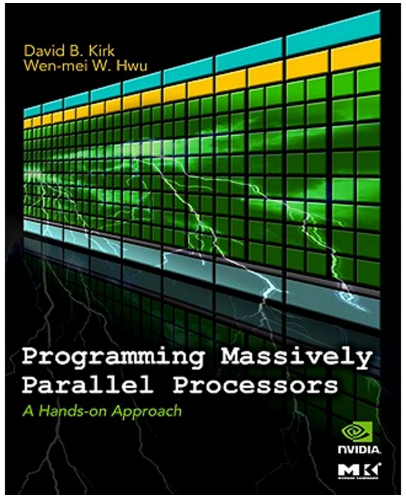
课程大作业二考核方式:GPU学习智能体（agent）

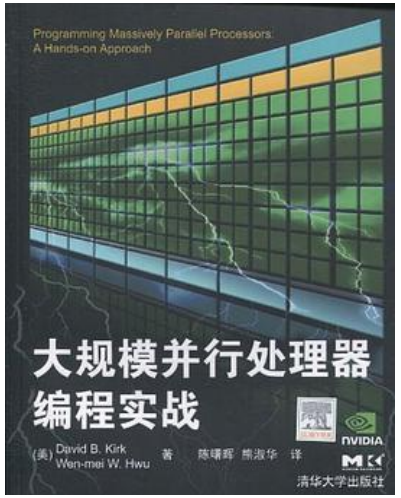
- ✓大作业二提交方式：将代码和一份描述文档，上传到本课程gitee，以便评分。

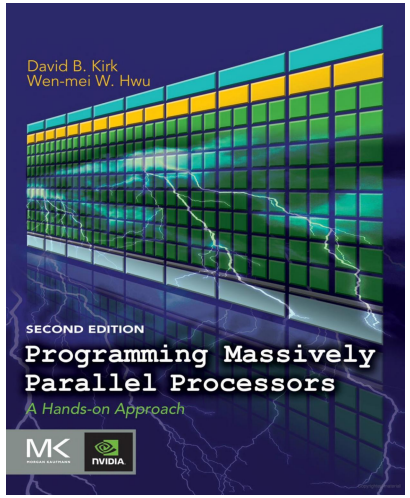
课程大作业二考核方式:GPU学习智能体（agent）

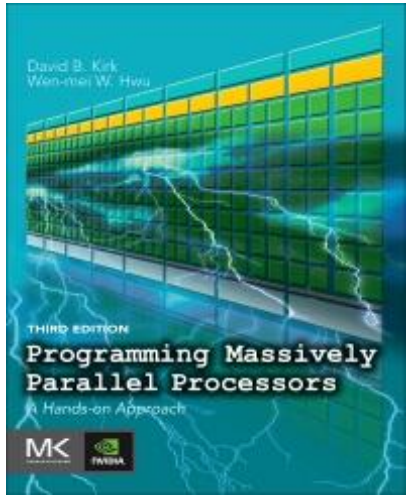
- ✓考核指标：速度（tokens/s）、准确率；
- ✓速度（tokens/s）：每个赛道的服务器测试的结果；
- ✓准确率：
  - ✓基础题：给定的测试集（仅限教材蓝色画勾的部分）；
  - ✓加分题：评委的打分。

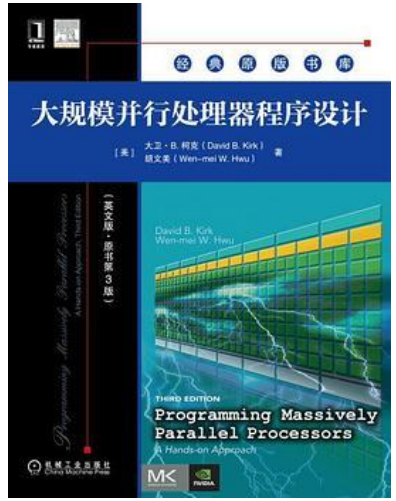
参考书：CUDA编程

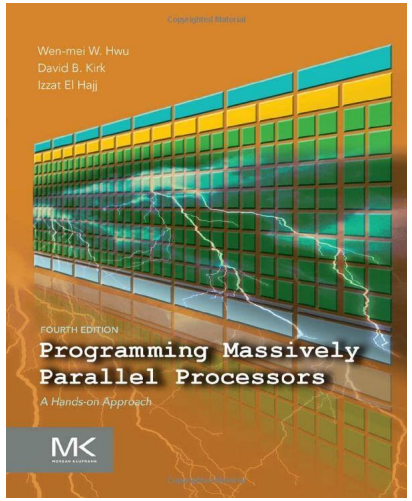












2010　　2012　　2017　　2022

✓教材：Third Edition: https://www.sciencedirect.com/book/9780128119860/programming-massively-parallel-processors

David B. Kirk, Wen-mei W. Hwu, Programming Massively Parallel Processors: A Hands-on Approach;

Contents

总结