加速计算进阶 —— 用多 GPU 加速 CUDA C++ 应用

充分发挥多 GPU 超强算力，大幅提升 HPC 应用性能

在高性能计算、数据科学、生物信息学和深度学习方面的计算密集型 CUDA C++ 应用，可以通过使用多个 GPU 来加速，这可以增加吞吐量和（或）减少总体运行时间。当计算和内存传输并发重叠时，计算可以扩展至多个 GPU 而不增加内存传输的开销。对于拥有多 GPU 服务器的组织，无论是在云上还是在 NVIDIA DGX 系统上，这些技术使您能够通过 GPU 加速应用程序，以实现最高性能。本课程所讲授的单节点多 GPU 技术，对于未来将应用扩展至多个节点至关重要。

本课程包含如何编写 CUDA C++ 应用程序，正确且有效地使用单一节点中的所有 GPU ，实现大幅提升应用程序性能，发挥多 GPU 系统的最佳经济效益。

课程时长	8 小时（课后可以继续访问和使用课件和实验资源）。
课程模式	讲师授课，及每位学员使用云端完全配置的 GPU 加速工作站实验练习。
课程价格	微信添加 DLI 小助手（微信号 DLIChina），沟通培训需求。
学员评测方式	基于代码技能评测，评估在单节点上正确利用多 GPU 的能力，包括如何使用复制/计算重叠执行。
培训证书	成功完成本课程和测试后，将获得 NVIDIA DLI 培训证书，证明在相关领域的能力，为职业发展提供证明。
预备知识	• 具有专业 CUDA C/ C++ 编程经验，包括使用 NVCC 编译器、内核启动、网格跨步循环、主机到设备及设备到主机的内存传输，和 CUDA 错误处理。 • 熟悉 Linux 命令行。 • 具有用 Makefiles 编译 C/ C++ 代码的经验。 建议学习如下课程，以具备本课程所需预备知识： • DLI 课程《加速计算基础 —— CUDA C/C++》 • Ubuntu Command Line for Beginners (1~5 章节) • Makefile Tutorial (至“*Simple Examples*” 章节)
课程语言	中文
工具、库和框架	CUDA C++, NVCC, Nsight Systems

学习目标

完成本课程后，您将能够了解：

- 使用并发 CUDA 流来使内存传输与GPU 计算重叠执行。
- 将工作负载扩展至多 GPU，充分利用单节点上所有可用 GPU。
- 在多 GPU 上使用数据拷贝与计算的重叠执行。
- 使用 NVIDIA Nsight Systems Visual Profiler 观察所学技术的改进效果和影响。

为何选择 NVIDIA 深度学习学院（DLI）的实战培训

- 随时随地访问云端完全配置的 GPU 加速工作站来动手实践。
- 获得实战经验指导，使用通用、行业标准的软件、工具和框架。
- 学习如何在广泛的行业中构建深度学习和加速计算应用程序，如自动驾驶汽车、数字内容创作、游戏开发、医疗医学及金融。
- 学习与行业领导者（例如洛杉矶儿童医院、梅奥医院和普华永道）合作设计的课程，获取现实应用的专业知识。
- 获得 NVIDIA 官方全球开发者培训证书，证明在相关领域的能力，助力职业发展。

课程大纲

议题	说明
介绍 (15 分钟)	> 讲师介绍 > 登录课程
使用 JupyterLab (15 分钟)	> 熟悉 GPU 加速的交互式 JupyterLab 环境
应用概览 (15 分钟)	> 从一个单 GPU 的 CUDA C++ 应用程序开始 > 使用 Nsight System 观察单 GPU 的 CUDA C++ 应用性能

学习 CUDA 流 (90 分钟)	> 学习管理并发 CUDA 流行为的规则 > 使用多个 CUDA 流执行并发的“主机到设备”和“设备到主机”内存传输 > 利用多个 CUDA 流启动 GPU 内核 > 使用 Nsight System Visual Profiler 观察多个流
休息 (60 分钟)	
使用 CUDA 流进行复制/计算的重叠执行 (90 分钟)	> 学习有效执行复制/计算的重叠执行的核心概念 > 探索在应用程序中灵活使用复制/计算重叠执行的可靠的索引策略 > 重构单 GPU CUDA C++ 应用程序以实现复制/计算的重叠执行 > 在 Nsight Systems Visual Profiler 中查看复制/计算的重叠执行
在多 GPU 上使用 CUDA C++ (60 分钟)	> 学习用 CUDA C++ 在单节点上有效使用多 GPU 的核心概念 > 探索为在应用程序中灵活使用多个 GPU 的可靠的索引策略 > 重构单 GPU 上的 CUDA C++ 应用程序以利用多个 GPU > 在 Nsight Systems Visual Profiler 中查看多 GPU 的使用情况
休息 (15 分钟)	
在多 GPU 上进行复制/计算的重叠执行 (60 分钟)	> 学习在多 GPU 上有效执行复制/计算的重叠执行的核心概念 > 探索为在多 GPU 上灵活使用复制/计算的重叠执行的可靠的索引策略 > 重构单 GPU 上的 CUDA C++ 应用程序，以在多 GPU 上执行复制/计算的重叠执行 > 观察在多 GPU 上进行复制/计算的重叠执行的性能优势 > 在 Nsight Systems Visual Profiler 中查看多 GPU 上的复制/计算的重叠执行
学习评估	> 完成测试并获取证书

(30 分钟)	
总结 (30 分钟)	> 回顾所学的关键内容 > 了解如何从 DLI 基础环境容器构建自有的训练环境 > 填写调查表

相关课程

- 《加速计算基础 —— CUDA C/C++ 》
- 《加速计算基础 —— CUDA Python》

购买培训和咨询

- 在 DLI 官网 www.nvidia.cn/dli ，页面上方导航栏处填写“联系我们”。
- 或，扫码添加 DLI 小助手，微信号 DLIChina 。