Accelerating CUDA C++ Applications with Multiple GPUs

Computationally-intensive CUDA C++ applications in high performance computing, data science, bioinformatics, and deep learning (DL) can be accelerated by using multiple GPUs, which can increase throughput and/or decrease your total runtime. When combined with the concurrent overlap of computation and memory transfers, computation can be scaled across multiple GPUs without increasing the cost of memory transfers. For organizations with multi-GPU servers, whether in the cloud or on NVIDIA® DGX™ systems, these techniques enable you to achieve peak performance from GPU-accelerated applications. And it's important to implement these single-node multi-GPU techniques before scaling your applications across multiple nodes.

This workshop covers how to write CUDA C++ applications that efficiently and correctly utilize all available GPUs in a single node, dramatically improving the performance of your applications, and making the most cost-effective use of systems with multiple GPUs.

Learning Objectives

By participating in this workshop, you'll learn how to:

- Use concurrent CUDA Streams to overlap memory transfers with GPU computation.
- Utilize all available GPUs on a single node to scale workloads across all available GPUs.
- Combine the use of copy/compute overlap with multiple GPUs.
- Rely on the NVIDIA® Nsight™ Systems Visual Profiler timeline to observe improvement opportunities and the impact of the techniques covered in the workshop.

Workshop Information and Prerequisites:

Duration:	8 hours
Price:	**Contact us for pricing**
Prerequisites:	- Professional experience programming CUDA C/C++ applications, including the use of the nvcc compiler, kernel launches, grid-stride loops, host-to-device and device-to-host memory transfers, and CUDA error handling - Familiarity with the Linux command line - Experience using Makefiles to compile C/C++ code Suggested resources to satisfy prerequisites: **Fundamentals of Accelerated Computing with CUDA C/C++**, **Ubuntu Command Line for Beginners** (sections 1 through 5), **Makefile Tutorial** (through *Simple Examples* section)
Tools, libraries, and frameworks:	**CUDA C++**, **nvcc**, **Nsight Systems**
Assessment type:	Skills-based cumulative assessments evaluate your ability to correctly leverage multiple GPUs on a single node, including the use of copy/compute overlap.
Certificate:	Upon successful completion of the assessments, you will receive an NVIDIA Deep Learning Institute (DLI) certificate to recognize your subject matter competency and support your professional career growth.
Hardware/software requirements:	You'll need a desktop or laptop computer capable of running the latest version of Chrome or Firefox. You'll be provided with dedicated access to a fully configured, GPU-accelerated workstation in the cloud. **Please complete the hands-on learning orientation at courses.nvidia.com/join prior to joining the course.**
Language:	English

Sample Workshop Outline

Introduction (15 mins)	> Meet the instructor.
Using JupyterLab (15 mins)	> Get familiar with your GPU-accelerated interactive JupyterLab environment.
Application Overview (15 mins)	> Orient yourself with a single GPU CUDA C++ application that will be the starting point for the course > Observe the current performance of the single GPU CUDA C++ application using the Nsight Systems
Introduction to CUDA Streams (90 mins)	> Learn the rules that govern concurrent CUDA Stream behavior > Use multiple CUDA streams to perform concurrent host-to-device and device-to-host memory transfers > Utilize multiple CUDA streams for launching GPU kernels > Observe multiple streams in the Nsight Systems Visual Profiler timeline view
Break (60 mins)	
Copy/Compute Overlap with CUDA Streams (90 mins)	> Learn the key concepts for effectively performing copy/compute overlap > Explore robust indexing strategies for the flexible use of copy/compute overlap in applications > Refactor the single-GPU CUDA C++ application to perform copy/compute overlap > See copy/compute overlap in the Nsight Systems visual profiler timeline
Multiple GPUs with CUDA C++ (60 mins)	> Learn the key concepts for effectively using multiple GPUs on a single node with CUDA C++ > Explore robust indexing strategies for the flexible use of multiple GPUs in applications > Refactor the single-GPU CUDA C++ application to utilize multiple GPUs > See multiple GPU utilization in the Nsight Systems Visual Profiler timeline
Break (15 mins)	
Copy/Compute Overlap with Multiple GPUs (60 mins)	> Learn the key concepts for effectively performing copy/compute overlap on multiple GPUs > Explore robust indexing strategies for the flexible use of copy/compute overlap on multiple GPUs > Refactor the single-GPU CUDA C++ application to perform copy/compute overlap on multiple GPUs > Observe performance benefits for copy/compute overlap on multiple GPUs > See copy/compute overlap on multiple GPUs in the Nsight Systems visual profiler timeline
Course Assessment (30 mins)	> Complete the assessment and earn a certificate.
Final Review (30 mins)	> Review key learnings. > Learn to build your own training environment from the DLI base environment container. > Complete the workshop survey.

Why Choose NVIDIA Deep Learning Institute for Hands-On Training?

- Access workshops from anywhere with just your desktop/laptop and an internet connection. Each participant will have access to a fully configured, GPU-accelerated server in the cloud.
- Obtain hands-on experience with the most widely used, industry-standard software, tools, and frameworks.
- Learn to build deep learning and accelerated computing applications for industries, such as healthcare, robotics, manufacturing, accelerated computing, and more.
- Gain real-world experience through content designed in collaboration with industry leaders, such as the Children's Hospital of Los Angeles, Mayo Clinic, and PwC.
- Earn an NVIDIA DLI certificate to demonstrate your subject matter competency and support your career growth.

For the latest DLI workshops and trainings, visit **www.nvidia.com/dli**

For questions, contact us at **nvdli@nvidia.com**