

**Model Parallelism: Building and Deploying Large Neural Networks**

**Course Number:** NVDA-106EC  
**Duration:** 1 day

**Overview**

This NVIDIA Model Parallelism training course teaches attendees how to train, optimize, and deploy large-scale models that push the boundaries of AI. Participants master cutting-edge techniques like model parallelism, inference optimization, and production deployment to tackle the real-world challenges of working with extensive deep neural networks (DNNs). By the end of this course, students confidently train large neural networks and deploy them to production.

**Prerequisites**

Attendees must have a good understanding of PyTorch and deep learning. Practice with multi-GPU training and natural language processing is useful but optional.

**Materials**

All attendees receive official courseware from NVIDIA in electronic format.

**Software Needed on Each Student PC**

The class will be conducted in a remote environment that Accelebrate will provide; students will only need a local computer with a web browser and a stable Internet connection. Any recent version of Microsoft Edge, Mozilla Firefox, or Google Chrome will be fine.

**Objectives**

* Understand the motivations and intricate nuances of training colossal neural networks
* Master fundamental techniques and frameworks for distributed training across multiple servers
* Implement advanced model parallelism strategies to overcome memory limitations and scale your models further
* Fine-tune model performance through profiling, auto-tuning, and mixture-of-experts architecture
* Implement real-world deployment tactics, including model reduction, NVIDIA libraries, and production-ready servers

**Outline**

* Introduction to Training of Large Models
  + Learn about the motivation behind and key challenges of training large models
  + Get an overview of the basic techniques and tools needed for large-scale training
  + Get an introduction to distributed training and the Slurm job scheduler
  + Train a Megatron-LM-based GPT model using data parallelism
  + Profile the training process and understand execution performance
* Model Parallelism: Advanced Topics
  + Increase the model size using a range of memory-saving techniques
  + Get an introduction to tensor and pipeline parallelism
  + Go beyond natural language processing and get an introduction to DeepSpeed
  + Auto-tune model performance
  + Learn about mixture-of-experts models
* Inference of Large Models
  + Understand the challenges of deployment associated with large models
  + Explore techniques for model reduction
  + Learn how to use NVIDIA® TensorRT™ and Faster Transformer libraries
  + Learn how to use Triton Inference Server
  + Understand the process of deploying GPT checkpoint to production
  + See an example of prompt engineering
* Conclusion