MODEL PARALLELISM: BUILDING AND DEPLOYING LARGE NEURAL NETWORKS (NV-MP-DEPLOY-NETW)

Course Code: 847007

Very large deep neural networks (DNNs), whether applied to natural language processing (e.g., GPT-3), computer vision (e.g., huge Vision Transformers), or speech AI (e.g., Wave2Vec 2) have certain properties that set them apart from their smaller counterparts. As DNNs become larger and are trained on progressively larger datasets, they can adapt to new tasks with just a handful of training examples, accelerating the route toward general artificial intelligence. Training models that contain tens to hundreds of billions of parameters on vast datasets isn't trivial and requires a unique combination of AI, high-performance computing (HPC), and systems knowledge.

What You'll Learn

- Train neural networks across multiple servers
- Use techniques such as activation checkpointing, gradient accumulation, and various forms of model parallelism to overcome the challenges associated with large-model memory footprint
- Capture and understand training performance characteristics to optimize model architecture
- Deploy very large multi-GPU models to production using NVIDIA Triton™ Inference Server

Visit us at www.globalknowledge.com or call us at 1-866-716-6688.

Date created: 4/8/2026 3:14:31 AM