Distributed deep learning (DL) can speed up model training from days to minutes, drastically accelerating model development and deployment cycles. However, cost-conscious organizations are faced with the challenge of efficiently utilizing a shared cluster of GPUs across many models being trained simultaneously. Due to the rigidity and opaque scalability of distributed training jobs, many cluster resource schedulers used in practice fail to achieve efficient GPU utilization.

In this talk, we will introduce our ongoing work on Esper, a DL-aware resource scheduler which aims to address the shortcomings of the alternative systems used in practice. Esper predicts the scalability of each DL training job and continually re-allocates GPUs, avoiding under-utilization and ensuring resource efficiency across all jobs in the cluster. Esper also considers the statistical aspects of model convergence by adapting the training parameters of each job to its current resource allocation, which maximizes the statistical efficiency of convergence. Lastly, we present an evaluation plan for Esper on production-like workloads.