At the core of Machine Learning (ML) analytics is often an expert-suggested model, whose parameters are refined by iteratively processing a training dataset until convergence. The completion time (i.e. convergence time) and quality of the learned model not only depends on the rate at which the refinements are generated but also the quality of each refinement. While data-parallel ML applications often employ a loose consistency model when updating shared model parameters to maximize throughput, the accumulated error may seriously impact the quality of refinements and thus delay completion time, a problem that usually gets worse with scale. Although more immediate propagation of updates reduces the accumulated error, this strategy is limited by physical network bandwidth.

In this talk, I will present Bosen, a system that maximizes the network communication efficiency to minimize such inconsistency error by fully utilizing the inter-machine network bandwidth under a given budget and prioritizing updates that are most significant to convergence. Experiments on various ML applications showed 2-3X improvements in convergence time compared to the previous state-of-the-art synchronization mechanism.

Presented in Partial Fulfillment of the CSD Speaking Skills Requirement.