ML transforms data into insights

Large amounts of data is being generated by user-software interactions, social networks, and hardware devices.

Timely insights depend on providing accurate and updated machine learning (ML) models using this data.

Large learning models, trained on large datasets often improve model accuracy [1].

Our Solution: MALT

Goal: Provide an efficient library for providing data-parallelism to existing ML applications.

MALT performs peer-to-peer machine learning. It provides abstractions for fine-grained in-memory updates using one-sided RDMA, limiting data movement costs when training models. MALT allows machine learning developers to specify the dataflow and apply communication and representation optimizations.

Network-efficient learning

In a peer-to-peer learning, instead of sending model info. to all replicas, MALT sends model updates to log(N) nodes, such that (i) the graph of all nodes is connected (ii) the model updates are disseminated uniformly across all nodes.

Traditional: all-reduce exchange of model information. As number of nodes (N) increase, the total number of updates transmitted in the network increases as O(N^2).

MALT model propagation: Each machine sends updates to log(N) nodes (to N2/2 + i and N4/4 + i) for node A. All N increases, the outbound nodes follows Halton sequence (N2, N4, N8, N16, N32,...) and the total number of updates transmitted increases as O(N log(N)).

We demonstrate that MALT outperforms single machine performance for small workloads and can efficiently train models over large datasets that span multiple machines (See our paper in EuroSys 2015 for more results).

References and Related Work