データ転送量を考慮したHadoop性能改善方式の提案と評価

A Study of Data Transfer Aware Hadoop Task scheduling

学籍番号：201221609
氏名：渡邊 飛雄馬
Hyuma WATANABE

現在広く使われている大規模データ分散処理フレームワークであるHadoopはMapReduceと呼ばれる分散処理の仕組みを採用している。MapReduceのTaskは大量のデータ入出力を発生させるため、TaskTrackerのHDD I/Oやネットワーク I/Oの輻輳を引き起こし、しばしばその処理速度が低下する。ところが、既存のHadoopのTask schedulerはTaskTrackerのI/O使用率に拘わらず、空きTask slotの有無のみでTask割り当ての判断を行ってしまう。このため、クラスタ内でI/Oリソース使用率の不均衡が発生し、Hadoopの性能が低下する。

本論文では、既存のTask schedulerを用いた予備実験を行い、HadoopのJob実行において、データI/Oを原因としてJob実行速度が低下する仕組みを示す。

さらに、その結果に基づき、TaskTrackerのネットワークI/O使用率とHDD I/O使用率を考慮したTask 割り当てを行う事でJob実行時間を短縮するHadoop Task scheduling方式を複数提案する。また、研究室内に構築した小規模なHadoopクラスタとAmazon EC2上に構築した中規模なHadoopクラスタを用いて提案方式の性能を評価する。提案したTask scheduling方式を組み合わせることで、平均6%のJob実行時間短縮を実現した。

研究指導教員：川原崎 雅敏
副研究指導教員：森嶋 厚行