ABSTRACT

Jajoo, Akshay Ph.D., Purdue University, December 2020. Exploiting the Spatial Dimension of Big Data Jobs for Efficient Cluster Job Scheduling. Major Professor: Y. Charlie Hu.

With the growing business impact of distributed big data analytics jobs, it has become crucial to optimize their execution and resource consumption. In most cases, such jobs consist of multiple sub-entities called tasks and are executed online in a large shared distributed computing system. The ability to accurately estimate runtime properties and coordinate execution of sub-entities of a job allows a scheduler to efficiently schedule jobs for optimal scheduling.

This thesis presents the first study that highlights *spatial dimension*, an inherent property of distributed jobs, and underscores its importance in efficient cluster job scheduling. We develop two new classes of *spatial dimension* based algorithms to address the two primary challenges of cluster scheduling.

First, we propose, validate, and design two complete systems that employ learning algorithms exploiting *spatial dimension*. We demonstrate high similarity in runtime properties between sub-entities of the same job by detailed trace analysis on four different industrial cluster traces. We identify design challenges and propose principles for a sampling based learning system for two examples, first for a coflow scheduler, and second for a cluster job scheduler.

We also propose, design, and demonstrate the effectiveness of new multi-task scheduling algorithms based on effective synchronization across the spatial dimension. We underline and validate by experimental analysis the importance of synchronization between sub-entities (flows, tasks) of a distributed entity (coflow, data analytics jobs) for its efficient execution. We also highlight that by not considering sibling sub-

entities when scheduling something it may also lead to sub-optimal overall cluster performance. We propose, design, and implement a full coflow scheduler based on these assertions.