CS554 Project Ideas

MATRIX:Hadoop – Improving Hadoop Performance and Scalability through Distributed Data-Aware Scheduling

Overview

Data driven programming models like MapReduce have gained the popularity in large-scale data processing. Although great efforts through the Hadoop implementation and framework decoupling (e.g. YARN, Mesos) have allowed MapReduce to scale to tens of thousands of commodity cluster processors, the centralized designs of the resource manager, task scheduler and metadata management of HDFS file system adversely prevent Hadoop from scaling to larger scales towards tomorrow's extreme-scale systems. This work investigate the potential of integrating the MATRIX scheduling system to the Hadoop framework. MATRIX will replace the Hadoop scheduler and enable distributed data-aware scheduling of data-intensive Hadoop applications. This work will build the interfaces that allow MATRIX talks to the YARN resource manager, the Hadoop application manager, and the HDFS file system. A distributed key-value store (e.g. ZHT) will be used to keep the task metadata in a distributed and scalable way. We will start from running benchmarking workloads, such as BOT; and then try to run typical Hadoop workloads, such as WordCount, TeraSort, RandomWriter and Grep; and finally make an effort to support real applications.

Relevant Systems and Reading Material

YARN:

Website: http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html
Paper: https://canvas.instructure.com/courses/845370/files/27447474?module_item_id=5620238

MATRIX:

Website: https://github.com/kwangiit/matrix_v2
Papers: assigned on the course webpage

ZHT:

Paper: assigned in the course webpage
Source code: https://github.com/mierl/ZHT

Methodology

Implementing the interfaces between MATRIX and YARN, between MATRIX and HDFS

Preferred/Required Skills

Required: Java, C/C++, Linux, Scripting language,

Parameters

Different workloads (e.g. WordCount, TeraSort, RandomWriter and Grep); Different scales up to 128 VMs on Amazon EC2

Metrics

Throughput (tasks/sec), efficiency, speedup

Project Mentor

Ke Wang, http://datasys.cs.iit.edu/~kewang/