CS554 Project Ideas

MATRIX: HadoopSim – Understanding the Scalability of the Hadoop Framework through Simulations

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 for certain workloads, we have limited knowledge regarding the scalability of the Hadoop framework for data-intensive workloads with different execution lengths (e.g. tens of seconds, seconds, sub-second). This project aims to explore the Hadoop scalability through simulations up to extreme-scales of hundreds of thousands to millions of nodes. We will simulate the whole Hadoop framework stack (Hadoop application, YARN resource manager, Hadoop scheduler, HDFS file system), on top of the discrete event simulator, PeerSim. We will first validate the simulations against running the real Hadoop system at moderate scales (up to 128 nodes), and then make an effort to scale the simulation up to extreme-scales on a shared-memory single node that has 256GB memory. The workloads that are going to be evaluated will have various data volumes and task execution lengths, ranging from sub-seconds to tens of seconds. Furthermore, we will compare the Hadoop simulation framework with the SimMatrix simulator, which a discrete event simulator of distributed many-task computing execution framework (i.e. MATRIX).

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

SimMatrix:

- Paper: http://datasys.cs.iit.edu/publications/2013_HPC13-SimMatrix.pdf
- Source code: https://github.com/kwangiit/SimMatrix

PeerSim:

- Website: http://peersim.sourceforge.net/

Methodology

Discrete event simulations on top of the PeerSim simulator that simulates the full Hadoop stack

Preferred/Required Skills

Required: Java, Linux, Scripting language

Parameters

Different workloads, different scales

Metrics

Throughput (tasks/sec), efficiency, speedup

Project Mentor

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