CS554 Project Ideas

MATRIX:Bench – Benchmarking the state-of-the-art Task Execution Frameworks of Many-Task Computing

Overview

Many-Task Computing (MTC) is a distributed paradigm of data-flow driven programming models, which bridges the gap between High Performance Computing (HPC) and High Throughput Computing (HTC). MTC applications are structured as graphs of loosely-coupled short tasks, with explicit input and output dependencies forming the graph edges. Tasks may be uniprocessor or multiprocessor, compute-intensive or data-intensive. The number of tasks, quantity of computing, and volumes of data may be extremely large. MTC needs scalable task execution framework to handle billions of jobs/tasks. The representative runtime systems for MTC are Charm++, Legion, Swift, Hadoop YARN, and Spark. Each one uses a task execution framework (**TEF**) similar to MATRIX, summarized in the following table. This project aims to benchmark and compare all systems with detailed profiling about their performance under the same workloads. The results will help us extend MATRIX.

Systems	Specifications	
MATRIX	Org	IIT
	TEF	MATRIX task scheduler
	Website	https://github.com/kwangiit/matrix_v2
	Papers	http://datasys.cs.iit.edu/publications/2014_BigData14_data-aware-scheduling.pdf
Charm++	Org	UIUC
	TEF	Converse
	Website	http://charm.cs.illinois.edu/
	Papers	1. http://charm.cs.illinois.edu/newPapers/14-24/sc14numa.pdf 2. http://charm.cs.illinois.edu/newPapers/13-26/paper.pdf 3. http://charm.cs.illinois.edu/newPapers/12-11/main.pdf
Legion	Org	Stanford University
	TEF	Legion task scheduler
	Website	http://legion.stanford.edu/
	Papers	1. http://legion.stanford.edu/pdfs/sc2012.pdf 2. http://legion.stanford.edu/pdfs/legion-fields.pdf
Swift	Org	UChicago & ANL
	TEF	Turbine
	Website	http://swift-lang.org/main/
	Papers	1. http://swift-lang.org/papers/pdfs/Turbine_2013.pdf 2. http://swift-lang.org/papers/pdfs/Swift_2013.pdf
YARN	Org	Hortonworks Inc.
	TEF	Hadoop scheduler for Hadoop applications
	Website	http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html
	Papers	https://canvas.instructure.com/courses/845370/files/27447474?module_item_id=5620238
Spark	Org	UC Berkeley
	TEF	Sparrow
	Website	https://github.com/radlab/sparrow
	Papers	http://dl.acm.org/citation.cfm?doid=2517349.2522716

Preferred/Required Skills

- Required: Linux, scripting language
- Preferred: C/C++, Java

Parameters

- Different workloads of fine granularity and intensity; Different scales up to 128 VMs on Amazon EC2

Metrics

- Throughput, latency, efficiency

Project Mentor

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