# CS554 Project Ideas

## CloudKon:CKMR – Accelerating MapReduce with CloudKon

### Overview

Predictions are that by the end of this decade, we will have exascale system with millions of nodes and billions of threads of execution. Unfortunately, today's job schedulers have centralized Master/Slaves architecture (e.g. Slurm, Condor, PBS, SGE), where a centralized server is in charge of the resource provisioning and job execution. This architecture has worked well in modest scales and coarse granular workloads, but it has poor scalability at the extreme scales of petascale systems with fine-granular workloads. The goal of this project is to leverage Amazon Simple Queuing Service (SQS) as a public cloud service to provide a scalable task scheduling system that supports Many Task Computing (MTC) workloads. SQS is a distributed queue service with the purpose of providing content delivery on extreme scales. CloudKon is a Distributed Task Execution Framework that runs on Amazon AWS Cloud. In this project you will be extending CloudKon to support MapReduce workload execution. Right now CloudKon only supports bag-of-task workloads invoking a variety of sleep tasks. Running Mapreduce tasks requires many features added to the system. For example the new system has to be able to support data communication through shared file systems on the framework enabling maps and reduce tasks to run on distributed data (S3 could be used as a scalable distributed storage system). There will be task dependencies and data dependencies between the tasks that have to be handled. Finally you have to compare the performance of the system with a vanilla version of MapReduce such as the Hadoop framework.

### Relevant Systems and Reading Material

Amazon SQS:

- http://aws.amazon.com/sqs/
- http://sqs-public-images.s3.amazonaws.com/Building_Scalabale_EC2_applications_with_SQS2.pdf
- http://awsdocs.s3.amazonaws.com/SQS/latest/sqs-gsg.pdf

Many Task Computing paper: I. Raicu, Y. Zhao, I. Foster, "Many-Task Computing for Grids and Supercomputers," 1st IEEE Workshop on Many-Task Computing on Grids and Supercomputers (MTAGS) 2008. http://datasys.cs.iit.edu/events/MTAGS08/MTAGS08_p25.pdf

MapReduce: **MapReduce**: simplified data processing on large clusters
https://www.usenix.org/legacy/publications/library/proceedings/osdi04/tech/full_papers/dean/dean_html/

Hadoop: http://hadoop.apache.org/

### Preferred/Required Skills

Programming language choice: Java, C/C++, Python

Skills/knowledge: You need to be familiar with MapReduce. You need to know Hadoop framework.

Other skills: REST API, Amazon EC2, Amazon SQS, Amazon DynamoDB, Linux Bash Scripting, distributed queues.

### Performance Metrics

Throughput, Latency, Efficiency, Utilization

### Project Mentor

Iman Sadooghi.

- isadoogh@iit.edu
- http://datasys.cs.iit.edu/~isadooghi/