Name: Wakar Ahmad

Supervisor: Prof. (Dr.) Bashir Alam

Department: Computer Engineering

Title of Thesis: Scheduling in Cloud for Big Data

Abstract

Big Data workflow application describe a series of computations that enable the analysis of data in a structured and distributed manner. Their importance is exacerbated in todays big data era as they become a compelling mean to process and extract knowledge from the ever-growing data produced by increasingly powerful tools such as telescopes, particle accelerators, and gravitational wave detectors. Due to their large-scale nature, scheduling algorithms are key to efficiently automate their execution in distributed environments, and as a result, to facilitate and accelerate the pace of scientific progress.

The emergence of the latest distributed system paradigm, Cloud computing, brings with it tremendous opportunities to run workflows at low costs without the need of owning any infrastructure. In particular, Infrastructure as a Service (IaaS) clouds, offer an easily accessible, flexible, and scalable infrastructure for the deployment of these scientific applications by providing access to a virtually infinite pool of resources that can be acquired, configured, and used as needed and are charged on a pay-per-use basis.

This thesis investigates novel scheduling approaches for Big Data workflow applications in IaaS clouds. They address fundamental challenges in order to fulfil a set of quality of service requirements expressed in terms of execution time, cost and energy. It advances the field by making the following key contributions:

1. A list-based scheduling algorithm with task duplication technique that enhances the system performance and minimize the makespan (running time) of the Big Data workflow applications (Montage, LIGO, and CyberShake).
2. A dynamic VM provisioning and de-provisioning based cost-efficient deadline-aware (DCEDA) based scheduling heuristic that schedule Big Data workflow applications in cost-efficient manner under user-specified deadline constraints.
3. An energy-efficient workflow scheduling algorithm named reducing energy consumption using fair pre-assignment of available budget (RECFPAB) that reduces energy consumption under client specified budget constraints.
4. A list-based scheduling algorithm that schedules Big Data Epigenomics workflow applications for minimizing their running time in heterogeneous cloud environments. The Epigenomics application is used to diagnosis of the abnormalities of the human body such as cancer.
5. An efficient scheduling heuristic is proposed named efficient IoT Big Data scheduling algorithm (E-IoT-BSA) that optimizes task scheduling problems of IoT-based Big Data application in the Cloud-Fog environment and its operating cost.