Optimal Processing Distribution for Big Data Analytics in a Cloud Environment Integrated with High-End Computing

I. Abstract

In massive computing applications that work on huge amount of data, I/O operations consume a lot of the overall execution time. Such applications belong to the class of Big Data Analytics (BDA) demanding large amounts of processing on large amounts of data. High-end computational capabilities and active storage solutions focus on pushing as much of the computation as possible closer to where the data resides. This decreases the time taken to fetch the data from the storage to the computing nodes, thus reducing the overall time taken by I/O operations. However, other components of the computations, such as data aggregation and subsequent processing of aggregated data, are frequently serial or low in amount of parallelism, thus requiring these components to execute on high-performance general-purpose processing nodes. Although the resulting distributed computing systems required for executing BDA applications feature massive parallel processing of huge amount of data on specialized computing nodes, they also feature significant complexity in scheduling multiple application components with various computational requirements for optimal execution on heterogeneous computing nodes.

This proposal considers cloud environments integrated with high-end active storage systems, and aims to develop tools and methods for optimal *distribution* of computational tasks of BDA applications in these heterogeneous computing systems. We will evaluate methods for mapping big data applications into Directed Acyclic Graphs (DAG), whose tasks can be optimally distributed across heterogeneous computational nodes to minimize overall execution time and data communication delay. We will evaluate different applications with the aim of determining the characteristics of applications that are best suited for such heterogeneous cloud environments. Specifically, we will use in our evaluation seismic applications as case studies.