Chapter 88
Performance Evaluation of Data Intensive Computing In the Cloud

Sanjay P. Ahuja
School of Computing, University of North Florida, USA

Bhagavathi Kaza
School of Computing, University of North Florida, USA

ABSTRACT

Big data is a topic of active research in the cloud community. With increasing demand for data storage in the cloud, study of data-intensive applications is becoming a primary focus. Data-intensive applications involve high CPU usage for processing large volumes of data on the scale of terabytes or petabytes. While some research exists for the performance effect of data intensive applications in the cloud, none of the research compares the Amazon Elastic Compute Cloud (Amazon EC2) and Google Compute Engine (GCE) clouds using multiple benchmarks. This study performs extensive research on the Amazon EC2 and GCE clouds using the TeraSort, MalStone and CreditStone benchmarks on Hadoop and Sector data layers. Data collected for the Amazon EC2 and GCE clouds measure performance as the number of nodes is varied. This study shows that GCE is more efficient for data-intensive applications compared to Amazon EC2.

1. INTRODUCTION

Cloud computing has become a viable solution for researchers and organizations for the on growing demanding needs. With the amount of data increasing exponentially across various fields of research like IT, social networking, Science, Engineering applications etc., dependency on the cloud is increasing. There is a need for the researchers to evaluate the performance of the cloud and study the metrics affecting the performance.

The present work evaluates the performance of two public clouds Amazon EC2 and GCE which are part of IaaS layer of the cloud. Three data-intensive benchmarks TeraSort, MalStone and CreditStone were used to benchmark the cloud. High CPU instances are chosen for the clouds as data intensive applications need more computing power than memory. Performance of the cloud is studied by varying the data sizes from 1GB, 10GB, 100GB and 1TB across the nodes 1 through 8. Response time is considered to be the primary

DOI: 10.4018/978-1-4666-6539-2.ch088

metric in evaluating the performance for big data applications.

Cloud offers the hardware and software necessary to support an application while providing storage, performance, security and maintenance. Clouds are classified into Public. Private and Hybrid clouds based on the deployment models and Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) based on the service models.

Amazon EC2 is an IaaS cloud service that provides a resizable computing capacity. EC2 supports various operating systems and instance types and Amazon EC2 defines the minimum processing unit, referred to as EC2 Compute Unit (ECU), which is the equivalent CPU capacity of a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor (AWS13, 2013).

Google Compute Engine (GCE) is an open source IaaS cloud service. GCE is a suitable alternative to the Amazon EC2 cloud service. GCE defines the minimum processing unit, referred to as Google Compute Engine Unit (GCEU), which is the equivalent CPU capacity of a 1.0-1.2 GHz 2007 Opteron processor. GCE uses 2.75 GCEU's to represent the minimum processing power of one logical core.

Big data refers to the collection of large, complex data sets, which can be structured or unstructured, and are difficult to process using traditional relational database management tools. Big data refers to large volumes of data which can be terabytes, petabytes or even xetabytes of data. Apache Hadoop and Sector are open source frameworks used to process big data to produce useful information.

Apache Hadoop is a well known open source framework used for data intensive applications. Apache Hadoop utilizes Master-slave system architecture in which the single master node is responsible for storing and managing the metadata and the multiple slave (worker) nodes process and store the data. Hadoop uses the Hadoop Distribution File System (HDFS), which is a block-based distributed file system, to distribute an application across the nodes in a cluster. Apache Hadoop ensures fault tolerance to prevent data loss in the event of a system failure by storing the same data on three unrelated nodes, by default; however, the number of nodes used for fault tolerance (referred to as the Replication Factor) is configurable.

MapReduce is a programming model used to process large data sets across a distributed collection of nodes in a cluster. Map () and Reduce () are two different functions in which Map () works on a set of inputs to generate the key-value pairs and Reduce () works on the output produced by Map () and sorts them to produce a single output.

Sector, a valid alternative for Hadoop for data intensive applications uses Sphere processing framework. Sector also uses master-slave architecture and ensures fault tolerance. Sector is widely used for WAN since it uses User Datagram Protocol (UDP) which is considered to be faster than TCP across wide area networks.

The remaining sections in the paper discuss the related works in section II, our experimentation in section III followed by results discussion in section IV and conclusions in section V.

2. MOTIVATIONS AND RELATED WORKS

2.1. TeraSort Benchmark

Gu *et al* performed a study on Apache Hadoop and Sector using the TeraSort benchmark and concluded that Sector is approximately two times faster than Apache Hadoop (Gu et al, 2009). The study, consisting of four 32-node racks located in three geographic regions of the United States (Baltimore, Chicago and San Diego) also indicated that Sector scaled better than Apache Hadoop as the number of nodes increased. Another study performed using the TeraSort benchmark run on 118 nodes located in the same region, required 1526sec with Sector and 3702 sec with Hadoop