Chapter 77
The Impact of Virtualization on High Performance Computing Clustering in the Cloud

Ouidad Achahbar
Al Akhawayn University, Morocco

Mohamed Riduan Abid
Al Akhawayn University, Morocco

ABSTRACT

The ongoing pervasiveness of Internet access is intensively increasing Big Data production. This, in turn, increases demand on compute power to process this massive data, and thus rendering High Performance Computing (HPC) into a high solicited service. Based on the paradigm of providing computing as a utility, the Cloud is offering user-friendly infrastructures for processing Big Data, e.g., High Performance Computing as a Service (HPCaaS). Still, HPCaaS performance is tightly coupled with the underlying virtualization technique since the latter is responsible for the creation of virtual machines that carry out data processing jobs. In this paper, the authors evaluate the impact of virtualization on HPCaaS. They track HPC performance under different Cloud virtualization platforms, namely KVM and VMware-ESXi, and compare it against physical clusters. Each tested cluster provided different performance trends. Yet, the overall analysis of the findings proved that the selection of virtualization technology can lead to significant improvements when handling HPCaaS.

INTRODUCTION

Big data and Cloud computing are emerging as new promising IT fields that are substantially changing the way humans dealt with data forever. During the last decade, data generation grew exponentially. IBM estimated data generation rate to 2.5 quintillion bytes per day, and that 90% of the data in the world today has been generated during the last two years (Manish et al., 2013).

DOI: 10.4018/978-1-4666-9840-6.ch077

The latest advances in Internet access (e.g. WiFi, WiMax, Bluetooth, 3G, and 4G) have substantially contributed to the massive generation of Big Data. Besides, the quick proliferation of the WSNs (Wireless Sensors Networks) technology did further boost the *data capture* levels.

Indeed, as Big Data grows in terms of volume, velocity and value, the current technologies for storing, processing and analyzing data have become inefficient and insufficient. A Gartner survey stated that data growth is considered as the largest challenge for organizations (2013). Stating this, HPC has started to be widely integrated in processing big data related to problems that require high computation capabilities, high bandwidth, and low latency network (Chee et al., 2005). HPC, by itself, has been integrated with new and evolving technologies, including Cloud computing platforms (e.g. OpenStack (The OpenStack Cloud Software)) and distributed and parallel systems (e.g. MapReduce and Hadoop). Merging HPC with these new technologies has led to a new HPC model, named HPC as a Service (HPCaaS). The latter is considered as an emerging computing model where end users have on-demand access to pre-existing needed technologies that provide high performance and scalable HPC computing environment (Ye et al., 2010). HPCaaS provides unlimited benefits because of the better quality of services, including (1) high scalability, (2) low cost, and (3) low latency (Umakishore and Venkateswaran, 1994).

Cloud computing is promising, in this context, as it provides organizations with the ability to analyze and store data economically and efficiently. Cloud computing is defined by National Institute of Standards and Technology (NIST) (2011) as a model for providing on-demand access to shared resources using minimum management efforts. NIST (2011) set five characteristics that define Cloud computing, including: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. Furthermore, based on NIST definition, Cloud computing provides the following basic services: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS).

Virtualization is deemed as the core enabling technology behind Cloud computing. When a user requests a Cloud service (e.g., SaaS, PaaS, or IaaS), the Cloud computing platform "forks" the corresponding virtual machines. The latter are created instantly, upon service request, and are "destroyed" once the user releases the relevant services. This fact leverages the "pay-per-use" feature of the Cloud. Since, Cloud computing platforms use different virtualization techniques, varying in their architectures and design, this ought to impact the overall performance of the Cloud services.

Parallel and distributed systems have also a significant role in enhancing the performance of HPC. One of the most known and adopted parallel systems is MapReduce paradigm (Jeffrey and Sanjay, 2004) that was developed by Google to meet the growth of their web search indexing. MapReduce computations are performed with the support of data storage system known as Google File System (GFS). The success of both MapReduce and GFS inspired the development of Hadoop (Apache Hadoop). This implements both MapReduce and Hadoop Distributed File System (HDFS) to distribute Big Data across HPC clusters (Molina-Estolano et al., 2009; Cranor et al., 2012). Nowadays, Hadoop is widely adopted by big players in the market because of its scalability, reliability and low cost of implementation.

At present, the use of HPC in the Cloud is still limited. The first step towards this research was done by the Department of Energy National Laboratories (DOE), which started exploring the use of Cloud services for scientific computing (Xiaotao et al., 2010). Stating this, HPCaaS still needs more investigation to decide on appropriate environments that can better fit big data requirements.