Chapter 9
Performance Evaluation of Multi-Core Multi-Cluster Architecture (MCMCA)

Norhazlina Hamid
University of Southampton, UK

Robert John Walters
University of Southampton, UK

Gary B. Wills
University of Southampton, UK

ABSTRACT

A multi-core cluster is a cluster composed of numbers of nodes where each node has a number of processors, each with more than one core within each single chip. Cluster nodes are connected via an interconnection network. Multi-cored processors are able to achieve higher performance without driving up power consumption and heat, which is the main concern in a single-core processor. A general problem in the network arises from the fact that multiple messages can be in transit at the same time on the same network links. This chapter considers the communication latencies of a multi core multi cluster architecture, investigated using simulation experiments and measurements under various working conditions.

INTRODUCTION

Chen, Wills, Gilbert, & Bacigalupo (2010) define cloud computing as an emerging business model that delivers computing services over the internet in elastic self-serviced, self-managed and cost-effective manner. Cloud computing doesn't yet have a standard definition, but a good working description of it is to say that clouds, or clusters of distributed computers, provide on-demand resources and services over a network, usually the Internet, with the scale and reliability of a data centre. Cloud computing provides a pool of computing resources which includes network, server, storage, application, service and so on that is required without huge investment on its purchase, implementation and maintenance that can be accessed through the Internet. The basic

DOI: 10.4018/978-1-4666-8210-8.ch009

principle of cloud computing is to shift the computing tasks from the local computer into the network (Sadashiv & Kumar, 2011). Resources are requested on-demand without any prior reservation and hence eliminate overprovisioning and improve resource utilization.

Cloud computing has changed the way both software and hardware are purchased and used. An increasing number of applications are becoming web-based since these are available from anywhere and from any device. Such applications are using the infrastructures of large-scale data centres and can be provisioned efficiently. Hardware, on the other side, representing basic computing resources, can also be delivered to match the specific demands without the user/consumer having to actually own them. As more organisations adopt cloud, the need for high availability platform and infrastructures, the cluster, to facilitate and distribute the load across multiple processors is evolving (Chang, Walters, & Wills, 2014). The deployment of clustered applications in Cloud infrastructures supports the capabilities of resource configuration and ensures communication between shared resources (Kosinska, Kosinski, & Zielinski, 2010).

The emergence of High Performance computing (HPC) that includes Cloud computing and Cluster computing has improved the availability of powerful computers and high speed network technologies. It can be concluded that the main target of HPC is better performance in computing. HPC aims to leverage cluster computing to solve advanced computation problems. While cluster computing has been widely used for scientific tasks, cloud computing was set out for serving business applications. Dillon et al. (2010) have pointed out that the current cloud is not geared for HPC for several reasons. Firstly, it has not yet matured enough for HPC; secondly, unlike cluster computing, cloud infrastructure only focuses on enhancing the overall system performance; and thirdly, HPC aims to enhance the performance of a specific scientific application using resources across multiple organisations. The key difference is in elasticity, where for cluster computing the capacity is often fixed, therefore running an HPC application can often require considerable human interaction, such as tuning based on a particular cluster with a fixed number of homogenous computing nodes (Schubert, Jeffery, & Neidecker-Lutz, 2010). This is contrasted with the self-service nature of cloud computing, in which it is hard to know how many physical processors are needed. In order to achieve higher availability and scalability of an application executed within cloud resources, it is important to supplement the capabilities of management services with high performance cluster computing to enable full control over communication resources.

Motivation of the Studies

In the past, there has been a trend to increase a processor's speed to get better performance. Moore's Law, which states that the number of transistors on a processor will double approximately every two years has been proven to be consistent due to the transistors getting smaller in successive processor technologies (Intel, 1997). However, reducing the transistor size and increasing clock speeds causes transistors to consume more power and generate more heat (D. Geer, 2007). These concerns limit cost-effective increases in performance which can be achieved by raising the processor speed alone. These issues gave computer engineers the idea of designing the multi-core processor, a single processor with two or more cores (Burger, 2005).

Multi-core processors are used extensively in parallel and cluster computing. As far back as 2009, more than 95% of the systems listed in the Top 500 supercomputer list used dual-core or quad-core processors (Admin, 2009). The motivation in this realm is the advances in multi-core processor technology that makes them an excellent choice to use in cluster architecture (Soryani, Analoui, & Zarrinchian, 2013). From the combination of cluster computing and multi-core processor, the