UNIVERSITI PUTRA MALAYSIA

BENCHMARKING FRAMEWORK FOR PERFORMANCE IN LOAD BALANCING SINGLE SYSTEM IMAGE

BESTOUN S. AHMED

BENCHMARKING FRAMEWORK FOR PERFORMANCE IN LOAD BALANCING SINGLE SYSTEM IMAGE

By

BESTOUN S. AHMED

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia in Fulfilment of the Requirements for the Degree of Master of Science

February 2009

Dedication

TO MY FAMILY, MY MOTHER SPECIALLY

Abstract of the thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the degree of Master of Science

BENCHMARKING FRAMEWORK FOR PERFORMANCE IN LOAD BALANCING SINGLE SYSTEM IMAGE

By

BESTOUN S. AHMED

February 2009

Chairman : Khairulmizam Samsudin, PhD

Faculty : Engineering

Single System Image, as a distributed operating system for nodes in computer clusters has become a widely adopted clustering solution due to its complete transparency of the resource management and ease of use. An important design consideration for this environment is the load allocation and balancing which is usually handled by an automatic process migration daemon. Thus, the implementation of such mechanism becomes an important design consideration in the distributed operating system.

There is an essential need for a benchmark framework for the Single System Image clusters due to the wide range of implementation and the need for identifying the performance and behaviour of the system. The benchmark framework will enable the researchers to investigate both relative behaviour and performance of the Single System Image clusters, as well as provides the ability to compare such systems.

In this work, a carefully designed benchmark framework had been proposed to study and evaluate the performance of the load balancing single system image. The performance metrics, which takes into account the speed, nodes, network, and behaviour of the system, were formulated. The benchmark framework allows the determination of the performance degradation factors associated with system implementation and configuration. This framework has been utilized to assess the performance characteristics of an existing and successful Open Source load balancing SSI system, OpenMosix. The benchmark framework provides an understanding of how the SSI system responds under varying conditions and manages to characterize the limitation of the information dissemination algorithm of OpenMosix. The information dissemination daemon had also been improved. The performance of the improved strategy had been validated by comparing it with the original system. Finally, the results from the tests were combined into a single figure of the performance behaviour.

The experimental results obtained from the benchmark framework showed that the numbers of nodes affect the performance of the SSI cluster; this could be regarded as an important factor of performance decaying. The number of nodes can affect the performance by adding extra costs including, but not limited to, network traffic, load balancing time, and overhead. The performance of any SSI cluster can be enhanced by improving any or all of the above factors. The improved load balancing strategy shows a visible performance gain with more than five nodes. At eight nodes, a gain of nearly 50 seconds runtime, 16.13 speedup, and 12.27 % efficiency have been successfully achieved.

Abstrak thesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Master Sains

RANGKA PENANDA ARASAN BAGI PRESTASI DALAM BEBAN IMBANGAN SISTEM IMEJ TUNGGAL

Oleh

BESTOUN S. AHMED

Februari 2009

Pengerusi : Khairulmizam Samsudin, PhD

Fakulti : Kejuruteraan

Sistem Imej Tunggal sebagai sistem operasi agihan kepada nodus dalam gugusan komputer telah banyak dijadikan penyelesaian gugusan kerana ia menunjukkan ketelusan yang total terhadap pengurusan sumber selain mudah digunakan. Satu rekabentuk perkiraan yang penting bagi persekitaran ini adalah peruntukan beban dan imbangan yang lazimnya dikawal oleh proses *'migration daemon'* automatik. Oleh itu pelaksanaan mekanisma tersebut perlu dipertimbangkan dalam merekabentuk sistem operasi agihan.

Satu rangka penanda aras bagi gugusan Sistem Imej Tunggal diperlukan disebabkan oleh medan pelaksanaan yang luas dan juga terdapatnya keperluan bagi mengenalpasti pencapaian dan sifat sistem tersebut. Rangka penanda arasan ini membolehkan pengkaji menganalisa sifat dan pencapaian gugusan Sistem Imej

Tunggal ini selain membolehkan perbandingan dilakukan ke atas sistem-sistem tersebut.

Menerusi kajian ini, satu rangka penanda aras yang dibentuk dengan teliti telah dicadangkan untuk mengkaji dan menilai pencapaian beban imbangan sistem imej tunggal. Pencapaian metrik yang mengambil kira kelajuan, nodus, rangkaian dan sifat sistem tersebut telah dirumus. Rangka penanda arasan tersebut juga berupaya menentukan faktor pengurangan pencapaian yang berhubung kait dengan sistem pelaksanaan dan konfigurasi. Rangka tersebut telah digunakan dalam menilai corak pencapaian sumber terbuka sistem beban imbangan SSI (*Single System Image*) sedia ada iaitu *'OpenMosix'*. Rangka penanda aras tersebut menerangkan respon sistem SSI ini di bawah situasi yang berbeza dan ia juga berupaya menggambarkan batasan dalam algoritma penyebaran maklumat *'OpenMosix'*. Prestasi *'daemon'* penyebaran maklumat juga dapat ditingkatkan. Prestasi algoritma yang telah diperbaharui telah disahkan menerusi satu perbandingan dengan sistem asal. Akhirnya, hasil ujian telah digabungkan menjadi bentuk tunggal yang menyifatkan pencapaian.

Hasil kajian yang diperolehi dari rangka penanda arasan menunjukkan jumlah nodus mempengaruhi pencapaian gugusan SSI; ia boleh disifatkan sebagai faktor penting dalam penurunan pencapaian. Jumlah nodus tersebut berupaya mempengaruhi pencapaian dengan penambahan kos sampingan, tapi tanpa had bagi laluan rangkaian, masa beban imbangan dan perbelanjaan. Prestasi sebarang gugusan SSI boleh dipertingkatkan dengan memperbaiki sebarang atau semua faktor tersebut. Rancangan beban imbangan yang telah diperbaiki itu menunjukkan satu pencapaian

untung tampak dengan lebih daripada lima nodus. Bagi lapan nodus, keuntungan hampir 50 saat masa jalan, 16.13 kelajuan dan 12.27% kecekapan telah dicapai.

ACKNOWLEDGEMENTS

I sincerely thank Dr. Khairulmizam Samsudin, my supervisor, and Dr. Abdul Rahman Ramli, my co-supervisor, for giving me encouragement, timely advice, and guidance to complete this project. I also thank them for being flexible, adaptable, and patient during this research.

I am not less grateful to the Department of Aerospace, especially Prof. Shahnor Basri for providing facilities and equipment for building the cluster to conduct the experiments. Also, my thanks to the lab staff for their kindness and help during my research time.

I would like to thank Dr. Christen Morin from Institute of INIRIA, France and Dr. A. J. Travis from ROWETT Research Institute, Scotland for their useful discussions.

Also thanks to Mr. Florian Dilzy, OpenMosix project member and the maintainer of 2.6 kernel for his valuable advice and Prof. Mohammed Othman for initial discussion at the beginning of this research.

This thesis submitted to the Senate of University Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Science. The members of the Supervisory Committee are as follows:

Khairulmizam Samsudin, Phd
Lecturer
Faculty of Engineering
Universiti Putra Malaysia
(Chairman)

Abdul Rahman b. Ramli, Phd
Associate professor
Faculty of Engineering
Universiti Putra Malaysia
(Member)

_____
HASANAH MOHD. GHAZALI, PhD
Professor/ Dean
School of Graduate Studies
Universiti Putra Malaysia
Date: 9 April 2009

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.

_____

BESTOUN S. AHMED

Date:

TABLE OF CONTENTS

LEST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CIFS	Common Internet File System
COTS	Commodity Of The Shelf
DEC	Digital Equipment Corporation
DKG	Distributed Key Generator
DSM	Distributed Shared Memory
GFS	Global File System
HA	High Availability
HAC	High Availability Cluster
HP	High Performance
HPC	High Performance Cluster
LVM	Load Vector Management
MOSIX	Multi Operating System for unIX
MPI	Message Passing Interface
NFS	Network File System
NIC	Network Interface card
OCFS	Oracle Cluster File System
OS	Operating System
PPM	Pre-emptive Process Migration
PVM	Parallel Virtual Machine
SAN	Storage Area Network
SMP	Symmetric Multi Processors
SNMP	Simple Network Management Protocol
SSI	Single System Image

CHAPTER 1

INTRODUCTION

1.1. Overview

The computer system has undergone very little change since 1940's when von Neumann designed the modern computer. From that moment, computer systems started to develop in an accelerated mode. Because of that development, many areas started to develop applications related to and accomplished by computer. With this development, it has been found that a single computer does not have the ability to solve increasingly large and complex scientific and engineering problems because of the limitation of possible physical performance. As a result, the demand for powerful computers appeared as an urgent demand.

For solving such problem, researchers proposed three ways [1] , [2]: firstly, is to work harder by taking large amount of time for processing. Secondly, is to work smarter, by means of using better algorithm for processing. Although the two ideas are important a third idea was proposed, that is by working as a group, using more than one computer or processor to solve a specific problem cooperatively. The latter idea led the researchers to the SMP (Symmetric Multiprocessors) systems while the development of network led to cluster systems.

Although SMP computers had powerful ability in comparison to other systems, there has been increasing trend to move away from expensive and proprietary SMP towards network of workstation. This is because SMP suffers several problems such as high cost, complicated programming procedure, and the limitation of growth because of the heat issues and bus communication [3].

Cluster theory in computers is best described as the interaction of a number of off-the-shelf commodity (COTS) computers and resources integrated through hardware, networks, and software to behave as a single computer. Whereas SMP means all processors are tied to common global memory [2].

With rapid advance in networking software and hardware technology, clusters of workstations / PCs are becoming the most important infrastructure for data transactions and engineering due to its popular platform for executing computationally intensive application and providing high availability. Many cluster systems have emerged; one of the most important architecture recently is the single system image (SSI).

The single system image architecture was developed to provide a unified system view as well as globally connected processor, file system, and network. The characteristics of SSI allow user to access system resources transparently irrespective of where they are available [4]. As it will be discussed in the following chapter, the load balancing single system image clusters dominate research work in this environment.

1.2. Problem Statement

The cost of building cluster introduces problems for those who want to get benefits of its features. The open source single system image clusters aided by commodity of the shelf (COTS) hardware, have appeared as an alternative.

Performance evaluation, benchmarking, and knowing the behavior of this kind of cluster is far more than curiosity about how fast the system runs an application. Most of the works in this field took the form of simulations, due to the complexity of evaluating and studying load balancing SSI. The limitations of simulation research urge the need for empirical study of implemented systems. Existing research on performance evaluation and benchmarking focus on the Beowulf clusters that has a different approach due to its dependencies on MPI [5], PVM [6] , [7] or any other middleware libraries [8]. These researches assume a simple static job assignment algorithm that completely ignores the current state of dynamic load balancing algorithm in SSI based system [9]. Thus, with simple time measurement and current benchmarks, the performance cannot be summarized and known in Load balancing single system image systems. Furthermore, unlike ordinary clusters like Beowulf, only a small effort has been made to define a benchmarking methodology and performance metrics in load balancing SSI.

To date, there is no explicit experimental study and real performance measurement for this kind of clusters. Literature shows that the factors that affect the performance in an SSI system are not clear. This research is carried out to fill this gap. This is very

important for understanding and studying the implemented load balancing mechanisms, and identifying the factors that affect the performance.

1.3. Aims and Objectives

The aim of the research is to obtain a benchmark framework for performance enhancement in load balancing single system image cluster. To realize this aim the following objectives are adopted:

i. To build a load balancing single system image cluster using OpenMosix.

ii. To enhance the system performance of load balancing single system image based on load information management.

iii. To evaluate and characterize the performance of the developed load balancing single system image.

1.4. Thesis Layout and Research Flow

Figure 1.1 shows the research flow that organizes the thesis. It highlights the topics included in the coming three chapters.

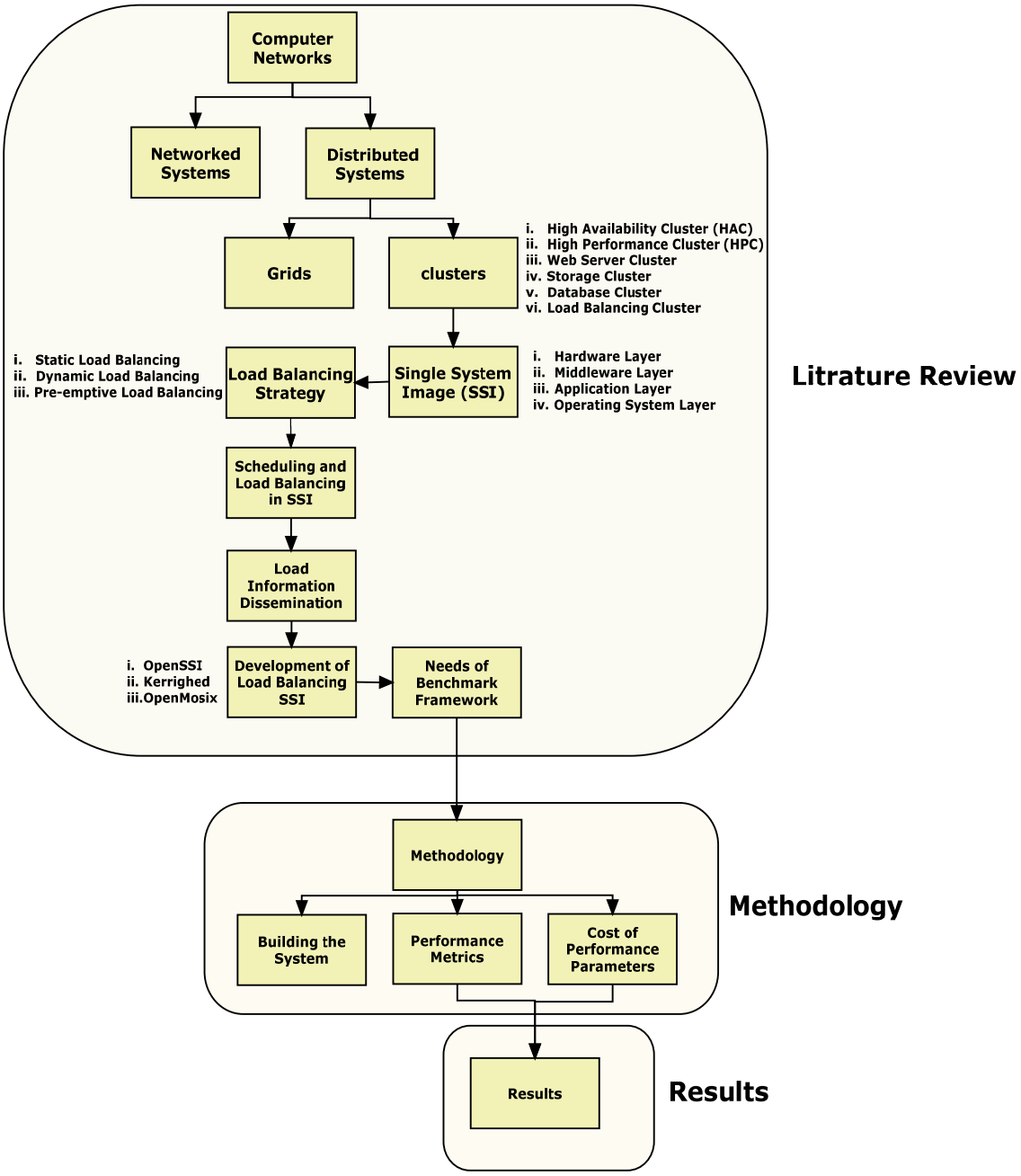

Figure 1.1 Research Flow

This thesis is organized into five chapters. In this chapter, an introduction and general overview of the research, brief background, objectives, and problem statement of the research have been given.

Chapter 2 aims to provide literature review of various related articles of the research. The chapter starts with the computer networks idea moving into introducing distributed systems, clusters, single system image (SSI), and load balancing strategy . Furthermore, the chapter discusses and examines the current literature of load balancing ideas and related researches ultimately introducing implemented systems. The chapter has concluded the needs of benchmark and performance evaluation framework. It gives a brief discussion on benchmarking, techniques and tools used in the framework as well.

In Chapter 3, a description of the empirical methodology that used to achieve the aim of this research is given. This is done by describing the test bed cluster that built for the purpose of this research. Then, going through the installation procedure, and finally describing the experimental procedure.

Chapter 4 gives the results of the experiments and testing of the system, which are all related to the framework. The chapter is given the results that were obtained due to an improvement of the load balancing strategy as well. The presentation of the results accompanies the discussion as well.

Finally, Chapter 5 concludes the thesis and its contribution. In addition it proposes future work and continuation in this research area.

CHAPTER 2

LITERATURE REVIEW

2.1. Introduction

This chapter begins with a general idea of networked and distributed systems. A single system image represents a very specific kind of clusters and distributed systems. For such reason, the chapter gives a brief classification of cluster systems. Then, a classification of SSI systems has been given.

In SSI system, like most of the distributed systems, situation arises when the node tried to get help from other idle processors within same system. Such situation represents an important part of this research. For this reason, the chapter presents different load balancing strategies in addition to the use of these strategies in SSI systems with the aid of scheduling.

Scheduling, as the main component of load balancing in SSI, needs an efficient knowledge about the system components and resources. Information dissemination mechanism plays the major part of this knowledge providing process. For this reason, information dissemination mechanism has been demonstrated in this chapter.