CHAPTER I

INTRODUCTION

Virtualization has widely been known in the field of energy saving technology. There are many related products that use this technique, and every enterprise improves its products to compete in the market. One use of such a technique is server virtualization which simulates not only number of virtual CPUs but also the CPU organization. It imitates a virtual machine with multiple computers, each of which runs its own operating systems and thereby working as though it was a multiple platform system. For resource usage, virtualization can be used to simultaneously manage heterogeneous workloads of different applications. The resources usage in virtualization is necessary for an appropriate management because it is a system that integrates with a variety workloads. It is important to know the resource usage of each application for managing heterogeneous workload. Normally, each application use resources such as CPU memory in its own pattern. There are many researches related with virtualization techniques, one of them is to improve workload management. However, the problem of tuning dynamic resource allocation is still a novel arena.

The typical concerned resources in every application are the number of assigned CPUs and size of allotted memory units to achieve the highest resource utilization and user satisfaction. An efficient resource management plan to achieve both aspects must be based on the pattern of past resource requests and prediction of future requests. Actually, the performance of any system must be evaluated in both aspects on system and user's side. The maximum satisfaction of both sides should be as high as possible with minimum compromising factors.

1.1 Problem Identification and Motivation

Previously, there were many researches focusing on how to improve real-time monitoring workload and allocate resources to the system. None of them involved user satisfaction, which could be in terms of response time and execution cost as part of their objectives. However, the real-time monitoring approach is too costly in practice because it must continuously run. Furthermore, the system must waste some execution cycles to monitor the events prior to allocating resource. This obviously increases the undesirable workload of the system. To alleviate this situation, the amount of requested resources must be estimated and allocated in advance.

This research presents a method to estimate the required number of CPUs and amount of memory usage in advance from the past data. The estimated resources must satisfy two previously mentioned aspects as much as possible. The following constraints are imposed in this study.

1. The behavior of deploying a computing system of any user is unknown.

2. Requested resources are determined by the running processes of the user's submitted tasks. The user has no privilege to demand the requested resources to the computing system.

The above constraints imply that estimating the requested resources within a short period of time can achieve better performance of resource management than a longer period. Therefore, the proposed method confines the estimation of resource usage to one hour in advance based on past data. The estimated resources may satisfy the response time but they may reduce resource utilization. Therefore, these estimated resources must be further adjusted as a consequent process to compromise for resource utilization.

1.2 Research Objectives

The objectives are as follows:

1. To predict hardware resources consumption focusing on CPU and memory.
2. To improve workload management in server virtualization.
3. To optimize hardware resources consumption focusing on CPU and memory.

1.3 Scope of the Study

Scope of work can be described as follows:

1. Predict user access and workload for both proxy and web server.
2. Predict hardware resources consumption focusing on CPU and memory on three different servers.

3. Compare predictive model using three different prediction algorithms.
4. Test the resource utilization focusing on CPU and memory.

1.4 Problem Statements

The following problems are investigated:

Problem1: How can one predict user behaviors in each service and duration of service?

Problem2: How can one predict the consumption of hardware resources required in each server service?

Problem3: How can one improve heterogeneous workload management in server virtualization?

1.5 Research Contribution

The expected outcomes will be as follows:

1. Introduce a new concept for managing workload in server virtualization.
2. Optimize hardware resources of the virtualization system.
3. Reduce investment and operation cost expended by the data center.

1.6 Related Definitions

Server Virtualization is "the masking of server resources, including the number and identity of individual physical servers, processors, and operating systems, from server users and enables to run multiple operating systems on a single physical server" [1], [2], [3].

Virtual machine (VM) is "the separation of virtual computing environments using virtualization technology and physical resources of a computing platform in the form of separate logical resources or computing environments" [4].

Resource utilization is "the use of a resource in such a way that increases throughput. The aim is to use these assets efficiently so as to maximize customer service levels" [5].

1.7 Organization of the Dissertation

The rest of this dissertation is organized in five chapters as follows. Chapter 2 recounts some the related work and basic knowledge such as virtualization backgrounds, exponential smoothing technique, association rule, ARIMA model, and resource utilization. Chapter 3 discusses the research processes encompassing problem formulation in mathematical terms, namely, objective function and constraints. Chapter 4 summarizes the experiment and results. Chapter 5 concludes the study and suggests appropriate future work.