Server Virtualization using user Behavior Model
Focus on Provisioning Concept

D. Prangchumpol

Abstract—Server provisioning is one of the most attractive topics in virtualization systems. Virtualization is a method of running multiple independent virtual operating systems on a single physical computer. It is a way of maximizing physical resources to maximize the investment in hardware. Additionally, it can help to consolidate servers, improve hardware utilization and reduce the consumption of power and physical space in the data center. However, management of heterogeneous workloads, especially for resource utilization of the server, or so called provisioning becomes a challenge. In this paper, a new concept for managing workloads based on user behavior is presented. The experimental results show that user behaviors are different in each type of service workload and time. Understanding user behaviors may improve the efficiency of management in provisioning concept. This preliminary study may be an approach to improve management of data centers running heterogeneous workloads for provisioning in virtualization system.

Keywords—association rule, provisioning, server virtualization.

I. INTRODUCTION

SERVER provisioning is the way of selecting a server from a pool of available servers; loading the appropriate software (operating system, device driver, middleware, and application); appropriately customizing and configuring the system, software to create or change a boot image for this server, and changing parameters to find associated network and storage resources [9].

Nevertheless, server provisioning is a time consuming process. If it takes too long to complete provisioning before the node can work normally, resource allocation actions won’t be able to timely catch the rapid changes in workloads, which lead to low efficiency. Recently, virtualization techniques have been proposed as a solution for maintaining reliability in data centers [10]. Provisioning often appears in the context of virtualization and cloud computing.

Virtualized systems are the masking of server resources, including the number and identity of individual physical servers, processors, and operating systems, from server users. The server administrator uses a software application to divide one physical server into multiple isolated virtual environments. The virtual environments are sometimes called virtual private servers, but they are also known as partitions, guests, instances, containers or emulations [11]. Server virtualization can be viewed as part of an overall virtualization trend in enterprise IT that includes storage virtualization, network virtualization, and workload management.

This trend is one component in the development of autonomic computing, in which the server environment will be able to manage itself based on perceived activity. Server virtualization can be used to eliminate server sprawl, to make more efficient use of server resources, to improve server availability, to assist in disaster recovery, testing and development, and to centralize server administration [12].

Utilizing resources in virtualization can reduce the number of server machines. When the number of servers is reduced, it reduces energy costs. The personnel cost will also be decreased. As a result, it is an easy-care system. Server provisioning is a key technique to improve resource utilization by changing the configurations of a shared server on demand. The server autonomic systems employ server provisioning to control the allocation of resources to maximize resource utilization and business revenue [8].

However, the problem of tuning dynamic resource allocation is a novelty. In this paper, propose a new concept for utilizing resources in virtualization system based on user behavioral models. The association rule; a technique in data mining, is employed to predict user behavior for each type of workload services and times. Understanding user behaviors may improve the efficiency of workload management and utilization resources in server virtualization.

The paper is structured as follows: related works are summarized in Section II. User behavior on heterogeneous workload is presented in Section III. Prediction user behavior is explained in Section IV. Correlation between user access and data size requirements is explained in Section V. Finally, Section VI describes conclusions and future work.

II. RELATED WORK

Now briefly review prior research about provisioning. An appliance-based autonomic provisioning framework for virtualized outsourcing data center is presented by Xiao Ying et al [10]. They introduced virtual servers into the autonomic data center and also discussed the problem of dynamic resource allocation using a queuing model based on performance estimations. Machida et al. [8] focused on a technique to shorten the provisioning processing time after the occurrence of the provisioning requested by speculative provisioning execution on virtual machine as standby.

There are numerous researches in virtualization techniques. For workload management: Steinder et al. [5] explored the usage of server virtualization technology in the autonomic management of data centers running a heterogeneous mix of workloads.

Myliski et al. [3] analyzed the influence of virtualization mechanisms of pSeries servers on dynamic resources and partition load manager utilities. Park et al. [4] identified some design considerations for constructing and managing clusters and proposed architectures to support clustering.

Examples of research related to energy cost reductions are: Tick et al. [6] emphasized the cost reducing effect of the ITS application on server virtualization through two case studies. Khanna et al. [2] showed monitoring of key performance metrics and used that data to trigger migration of Virtual Machines within physical servers, while using algorithms that attempt to minimize the cost of migration and maintain acceptable application performance levels.

D. Prangchumpol is with the Faculty of science and technology, Suan Sunandha Rajabhat University, Dusit, Bangkok 10300 Thailand (phone: +6602-160-1111; e-mail: dulywit@gmail.com).
Research related to the field of improved virtualization with varied I/O workloads includes: Kang et al. [1] used the virtual machine-aware proportional share queuing scheduler, VM-PSQ, in server virtualization environments with different I/O requirements and priorities.

III. USER BEHAVIOR IN HETEROGENEOUS WORKLOADS

The objective of this section is to demonstrate that user behaviors are different in each type of workload service and times. The studies of user behavior are investigated on three types of servers: proxy server, web server and database server. The number of user accesses and compute data size per day and hours in all servers are captured.

Fig. 1 Step to analyze user behaviors

The study consists of 3 main steps as illustrated in Fig.1. The first step is data storage from each server. The second step is the cleaning process, prepare the data for analyze. Finally, the last step is analyzing the results, capturing the number of user access and computing data size per day and hours. In Fig. 2 and Fig. 3, the average number of user accesses in the proxy server and web server for each period times are used to plot the graphs [7]. Fig. 4 shows the average number of user accesses in the database server.

Fig. 2 User accesses in proxy server (per hour)

Fig. 3 User accesses in web server (per hour)

Fig. 4 User accesses in database server (per hour)

On the other hand, Fig. 3 shows user access behavior for the web server. It can be seen that during 08.00 to 16.00, user accesses are less frequent than other times. However, according to the graph in Fig. 3, the access is peak at 19.00.

Fig. 5 shows user access behavior for the database server. Behavior in database server seems like sine wave. During 06.00 to 08.00 and 16.00 to 17.00, has an increased and 17.00 has the highest levels of user access in database server.

In Fig. 5, Fig.6 and Fig.7, days of the week and times are plotted against the data sizes requested by users to the proxy, web and database servers, respectively.

Fig. 5 Shows average data size in proxy server over a 24 hour period

Fig. 6 Shows average data size in web server over a 24 hour period
The result from Fig. 5, Fig.6 and Fig. 7 shows that user behaviors are different in type of services workload and also different in each time.

From Fig. 5, it appears that workload in the proxy server are high during the midday. Fig. 6, on the contrary, workload in web server during the mid day is less than other times [7]. Fig. 7 workload in database server seems like sine wave follow access behaviors.

IV. PREDICTION OF USER BEHAVIOR

An association rule, a data mining technique is selected to predict user accesses for all servers.

User accesses are categorized into 5 levels: ‘1’ for low level; ‘2’ for medium low level; ‘3’ for medium high level; and ‘5’ for high level. Here, the levels of access are assumed to be uniformly distributed.

The relationship in the form of LHS $\rightarrow$ RHS is applied for extracting rules. The extracted rules for LHS are based on days of week and 1-hour periods of time.

Let $D_1, D_2, ..., D_7$ be days and $T_1, T_2, ..., T_{24}$ be time. However, this research restricts the RHS as follows. Let $L_1, L_2, L_3, L_4, L_5$ be the levels of user access for the RHS that can be predicted based on the term on the LHS. Therefore, a rule $(D_i, T_j) \rightarrow L_k$ is created. Where $L_k$ occurs most frequently in the rows.

For each rule of the form LHS $\rightarrow$ RHS, define the supp and conf as the support and confidence as follows:

$$conf(LHS, RHS) = \frac{count(LHS, RHS)}{count(LHS)} \quad (1)$$

such as $conf(day, time \rightarrow level) = \frac{count(day, time and level)}{count(day, time)} \quad (2)$

$$sup(LHS, RHS) = \frac{count(LHS, RHS)}{count(All)} \quad (3)$$

such as $sup(day, time \rightarrow level) = \frac{count(day, time and level)}{count(All)} \quad (4)$

Table I shows the total association prediction model for the database server with confidence and support values.

The performance of the model was tested. In general, the data is divided into a training data set and a test data set.

Data obtained in November for 30 days are used to train the model while data acquired for 13 days in December are used to test the performance of the model. Note that the ratio of the training set and testing set is 70:30.

The performance of the predictive model for the proxy server is 86.86%. Similarly, the performance of the predictive model for the web server and the database server are measured the same way as the proxy server. The results demonstrate that the accuracy prediction of the level of user access for the web server is 87.18% and the database server is 85.26%.

V. CORRELATION

It is interesting to know how much the two variables, user access and their data size requirements in the server are correlated. A simple linear correlation is employed for the explorations.

The “$x$” is defined as the number of user accesses (independent variable) and “$y$” as data size requirements (dependent variable). The simple linear correlation equation is in the form $y = a + bx$. Where $a, b$ are calculated from the following equations:

$$b = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad (5)$$

$$a = \bar{y} - b\bar{x} \quad (6)$$

<table>
<thead>
<tr>
<th>No.</th>
<th>Rule</th>
<th>Conf (%)</th>
<th>Sup (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Monday ,12:00 AM =&gt; Low</td>
<td>100</td>
<td>0.56</td>
</tr>
<tr>
<td>2</td>
<td>Monday ,01:00 AM =&gt; Low</td>
<td>100</td>
<td>0.56</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>Monday ,09:00 AM =&gt; Low</td>
<td>75</td>
<td>0.42</td>
</tr>
<tr>
<td>11</td>
<td>Monday ,10:00 AM =&gt; Medium Low</td>
<td>50</td>
<td>0.28</td>
</tr>
<tr>
<td>12</td>
<td>Monday ,11:00 AM =&gt; Medium Low</td>
<td>50</td>
<td>0.28</td>
</tr>
<tr>
<td>13</td>
<td>Monday ,12:00 PM =&gt; Low</td>
<td>100</td>
<td>0.56</td>
</tr>
<tr>
<td>14</td>
<td>Monday ,13:00 PM =&gt; Low</td>
<td>100</td>
<td>0.56</td>
</tr>
<tr>
<td>9</td>
<td>Monday ,14:00 PM =&gt; Low</td>
<td>100</td>
<td>0.56</td>
</tr>
<tr>
<td>10</td>
<td>Monday ,15:00 PM =&gt; Medium</td>
<td>50</td>
<td>0.28</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>168</td>
<td>Sunday ,23:00 PM =&gt; Low</td>
<td>100</td>
<td>0.69</td>
</tr>
</tbody>
</table>

When $\bar{x}$ and $\bar{y}$ is the average value of $x$ and $y$, respectively.

The correlation coefficient measures the strength and direction of the linear relation between two variables. The correlation coefficient can be computed by the following formula:

$$r = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n \sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \sqrt{n \sum_{i=1}^{n} (y_i)^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

The correlation coefficient $r$ is the same as $b$ calculated from the simple linear correlation equation.

The correlation coefficient measures the strength and direction of the linear relation between two variables. The correlation coefficient can be computed by the following formula:

$$r = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n \sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \sqrt{n \sum_{i=1}^{n} (y_i)^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

The correlation coefficient $r$ is the same as $b$ calculated from the simple linear correlation equation.
Finally, the factors such as CPU usage and memory loads should take into considerations.

REFERENCES


