International Journal of Control and Automation
Vol.7, No.10 (2014), pp.405-410
http://dx.doi.org/10.14257/ijca.2014.7.10.37

A Study on the Performance Simulation Model for Estimating Hardware Scale

Kook-Hyun Choi[1], Gei-Young Kim[2], Yong-Tae Shin[3] and Jong-Bae Kim[4*]

[1,2] *Department of IT Policy and Management, Graduate School of Soongsil University, Seoul 156-743, Korea*
[3]*Department of Computer Science, Soongsil University, Seoul 156-743, Korea*
[4*]*Graduate School of Software, Soongsil University, Seoul 156-743, Korea*

[1]*khchoi@tsline.co.kr*,[2]*geiyoungkim@gmail.com*,[3]*shin@ssu.ac.kr*,[4*]*kjb123@ssu.ac.kr*

Abstract

The level of services in an information system is determined by the level of performance provided by the hardware of the system, such as the servers and the network, and the software, such as the application. Of these, the performance of the hardware is significantly affected by the specifications determined at the time of the initial implementation of the system. Therefore, when one implements a system, it should be ensured that hardware models of sufficient performance specification are selected in order to reach the targeted performance level. However, it is also apparent that high-performance hardware requires a higher investment. Therefore, from the perspective of the IT systems manager of an organization, it is important when the system is implemented to choose the hardware models that can come as close to the targeted performance level as possible within the budgetary limit. To solve this problem, organizations present instructions to calculate the appropriate scales of information system hardware. However, as these instructions are based on the capacity expected during the initial system implementation planning in the earlier stage, there are limitations in using these to establish operation plans, as deterioration in performance as the system is being operated should be anticipated. It is for this reason, in this study, that we developed an algorithm for the simulation model that can estimate the performance of the hardware not only for the initial implementation of the system but also during its operation phase.

Keywords: *Performance of the hardware, Scales, Simulation model, ITIL, Estimate*

1. Introduction

In designing and implementing an information system, the hardware represents the lion's share of the budget for the project. In order to realize a proper level of performance of the system service within the limit of the available budget, it is important to calculate the capacity of hardware with proper performance based on the requirements of the system. However, in most cases, the capacity calculation of the hardware depends on the propositions of the system suppliers or SI contractors. They, too, are in a difficult position to ensure the accuracy of their proposed data, as they had to rely on the number of expected users, the system structure based on the nature of the jobs to be performed, and their own experience with operation when they calculated the capacity [1].

* Corresponding author. Tel. : +82-10-9027-3148.
Email address: kjb123@ssu.ac.kr(Jong-Bae Kim).

ISSN: 2005-4297 IJCA

To solve this problem, TTA(http://www.tta.or.kr) and NIA(http://www.nia.or.kr), as well as other organizations, suggested the hardware scale calculation institution for information systems [2-3]. However, these instructions, too, have been based on the anticipated capacities for system implementation planning in the initial phases, resulting in limited utility in anticipating degradation of the performance during the operation of the system and the establishment of a proper operation plan.

For this reason, in this study, we intend to develop an algorithm for the simulation model that can anticipate the performance of the system not only in the initial phase of implementation but also in the operation phase of the system.

2. Related Studies

ITIL(Information Technology Infrastructure Library) defines capacity planning as the planning activities to ensure that the business requirements can be accommodated in a cost-effective manner. That is, capacity planning is required to balance the investment budget with the business requirements. In general, the capacity of the information system increases as the investment amount increases; that is, when either the number of systems is increased or the size of resources such as the CPU or the memory increases. However, the system investment budget available for the organization is inevitably limited. ITIL defines capacity management as part of the service delivery domain and presents a standardization process. Here, capacity planning can safely be regarded as a part of the entire capacity management process [4].

On the other hand, the capacity planning for an information system can encompass three different types of capacity plans based on the target criteria: the business capacity plan, the service capacity plan, and the resource capacity plan. First, business capacity planning involves the capacity planning activities to manage the services and IT infra in order to maintain the service level agreement with the customers regarding the new services to accommodate new business requirement. Secondly, service capacity planning is a series of anticipatory activities in order to establish future investment plans in terms of the operation of various services that handle the business-related tasks of an organization. Here, the estimation of the data can be put to a correlation analysis with the critical performance level that was calculated through the performance test of the target service, which finally provides valuable information for establishing the capacity plan for the service. Third, resource capacity planning is an activity to estimate the future investment plans in terms of the utilization rate of each component of the information service, such as the CPU or the memory. The utilization rate data for each component is collected over an extended period, and the trend of increased resource use over time is analyzed by calculating the workload to provide the capacity planning information on the relevant component [3].

Of these, the purpose of the current system analysis is to calculate the baseline of the target system performance. By analyzing the workload and resource usage increase trend over time, the future workload and resource usage are estimated. From a statistical perspective, the longer the time span of the analysis, the more reliable the estimation data become. The most frequently used methods for modeling the workload include the simple calculation model, the linear regression model, the queuing model, and the simulation model, etc. [3].

Among these, again, the simple calculation model is the simple arithmetic calculation of the increasing trend of the usage of each component to estimate the capacity. The usage of each resource is measured over an extended period, and under the assumption that the increased amount of usage is fixed, the trend of increase in the usage rate of each resource is estimated. Based on the threshold of these resources, then, an estimation of the capacity is made. The simple calculation model does not consider the queuing that may happen to the resources, giving it a relatively lower level of reliability, which is a drawback. However, it has an advantage in that anyone can make an estimation of the future situation without difficulties.

Next, the linear regression model, which is a statistical analysis method, verifies the relationship between the independent variables and the dependent variables to determine the influence of the independent variables over the dependent variables, or to estimate the changes in the dependent variables caused by the changes in the independent variable. Based on the number of independent variables, this method is again divided into the linear regression method and the multiple regression analysis method. For capacity planning, the linear regression method, in which there is only one independent variable, is used. Here, the independent variable is defined by the time of measurement, and the dependent variable is defined by the MAX TPS during the peak times in the target system, to analyze the linear correlations between the two variables.

The estimated value from the linear regression analysis can then be put to an analysis in connection with the threshold test results that could estimate the critical conditions of the system, which will yield the estimation results for the capacity of the target system. This linear regression analysis mode can be used in connection with the results of the critical performance test results using performance test tools. In addition, since it uses an arithmetic modeling based on the workload increase trend of the target system, it may seem to be free of statistical errors. However, in the real world, the workloads are affected by a large number of independent variables, making the estimation significantly more difficult. The increase in the workload in the real world does not show a linear increase pattern due to the influences of various variables, but shows a step-wise or curved increase pattern. Due to this gap with the reality, it can also be a good idea to use a multiple regression model that considers two or more variables. However, this method takes a lot of time, and there are many limitations in terms of collecting the baseline data for each of these variables. Therefore, the linear regression model is being widely used as the capacity estimation model for information systems, for now.

For the simulation model [5-6], it is not necessary to physically construct various conditions of the information system. Instead, modeling tools could be used to run a pre-test, which will give you the results for each of the key assessment indices under the target condition. Normally, the capacity planning process using a simulation modeling is composed of 4 steps, which are modeling, predicting, changing, and analyzing. The capacity planning using simulation modeling allows you to simulate various conditions that cannot be realized in the real world through a virtual systems environment, giving you valuable estimation data regarding the changes. But on the other hand, it has its own shortcomings, in that the products used in the process are very expensive and this method cannot be applied without help from experts, preventing it from being widely used.

3. Hardware Performance Simulation Model

3.1. Designing of the hardware simulation modeling

Ahead of designing the simulation model, an examination of the transaction processing process of information systems is performed as shown in Figure 1.

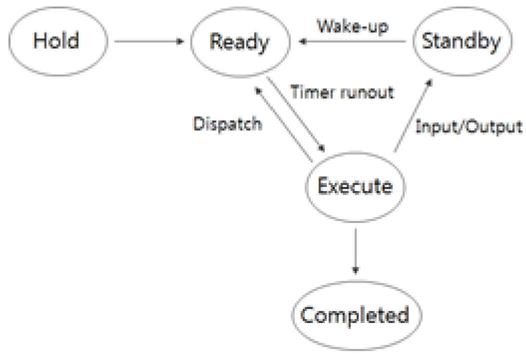

Figure 1 Transaction Processing Process

Figure 1 shows the processing procedure for when a transaction has been submitted to a server, creating a process. First, when a process is created, it is recorded in the spooling space, and then is shifted to ready status to be allocated CPU services. When CPU resources become available, the status changes to execute, with system services allocated, and information processing is performed. When events such as I/O occur, a standby state is maintained until the event ends. Once all the work that a process aims to process is completed through this series of steps, the status changes to completed, completing the transaction [7].

An examination of this series of transaction processing steps reveals that the basic mechanism for information processing operates in a similar manner as the queuing method. For this reason, to model the information processing characteristics of information systems through simulation, the queueing model was used. Queuing is a system wherein work is queued to receive services [8-9]. Accordingly, by examining the operating characteristics of information systems, a queuing-based simulation model most appropriate for information systems was defined in this study. Transaction processing by information system components, expressed in a discrete event simulation-based queuing model, is simulated using ARENA [10]. As for the types of transactions, a statistical Poisson distribution was applied to enhance the reality of the simulation and the reliability of the results. With the simulation results, a dynamic graph that changes in the process of simulation can be used to examine the trends in changes to equipment usage and the amount of transactions standing by.

3.2. Composition of the Simulation Model

To perform server modeling, server runtime was simplified into service time, wait time, and queue time. The input values for modeling were also restricted to tpmC and usage. To perform simulation using the discrete event simulation technique, the service time of the server must be deduced. Here, transactions that are input values for the system architecture, being logical values, make the deducing of server service time difficult.

The server system was modeled using service time, that is, the time required by the CPU to process a transaction, standby time that occurs at the NIC(Network Interface Card), memory, and disk I/O(Input/Output), and queuing time without service allocation. Results of previous tests show that transaction intervals and server usage can be known when data is input, from which service time can be deduced based on the queuing model. This model uses the results of previous performance tests to deduce the service time for a certain transaction, them simulates the situation as-is or with a system change applied. Therefore, its input is the result from prior performance testing. Here, the occurrence interval of the specific transaction, the total runtime of the transaction, and the usage rates for each server and equipment must be provided as input. The service time for the transaction deduced in this manner is significant in that a logical transaction has been quantified, allowing for numerical simulation. Meanwhile, service time, as shown in equation (1) below, is computed using the input CPU usage and the average transaction arrival time.

$$CPU\ usage\ (\%) = \frac{Service\ Time \times 100}{Average\ Transaction\ Arrival\ Time} \quad (1)$$

Events occurring with a Poisson distribution are stored in a queue, and simulation is performed according to the queuing model. Following simulation, current server service time is deduced based on tpmC and previous usage rate. Using the CPU load according to the occurrence of transactions, the new usage rate is deduced.

4. Conclusion

In this study, we developed an algorithm for the simulation model that can estimate the performance of the hardware system and perform a visual simulation of it. It is expected that the findings of this study will allow estimation of the capacity of the hardware required in the operation phase, making it possible to establish cost plans for the initial implementation and subsequent expansion of systems, and contribute to maintaining a proper level of systems services.

References

[1] K.-H. Choi, G.-Y. Kim, Y.-T. Shin and J.-B. Kim, "A Study on the Simulation Model of Hardware Performance", 2014 International Conference on Future Information & Communication Engineering (ICFICE), **(2014)**, pp 335-338.
[2] W. S. Kim, "Information System Performance Management Guidelines," TTA, **(2007)**, December.
[3] S. J. Moon, "Guidelines for Calculation of Hardware Capacity for Information Systems", TTA, **(2008)**, December.
[4] L. Klosterboer, "ITIL Capacity Management", 1st Edition, IBM Press, **(2011)**.
[5] B.-S. Lee, K.-S. Cho, J.-K. Yang, B.-C. Ahn and Y.-W. Seo, "A Research on Implementation of Information System Simulator", 2010 Internationl Conference of Computer Measurement Group, **(2010)**, Dec.
[6] IBM High Performance On Demand Solutions team and eBay, "HVWS Simulator : eBay Case Study", IBM Corporation, **(2004)**.
[7] Silberschatz, "Operating System Concepts", 8th Edition, Wiley, **(2008)**.
[8] C. MoonGak, "Translated by Seung Ho Shin et al. Introduction to Queuing Networks", 2nd Edition, **(2000)**.
[9] H. Lee, "Queuing Systems for Engineering Students", Hongreung Scientific Publishing, **(2008)**.
[10] Translated by Gyeong Moon, Simulation Using ARENA, 4th Edition, McGraw-Hill Korea, **(2007)**.

Authors

Kook-Hyun Choi

He received his bachelor`s degree in Computer Science from In-ha University in Korea, (1995). He worked in the IT field as a System engineer over 12 years. Now He is CEO of TSLINE System Co., LTD. since 2003.

Gei-Young Kim,

He obtained LG Masters of Business Administration (2005). He received his master's degree in health management information system from Yonsei University (2011). He is a PhD student studying Information Technology Policy & Management at the Soongsil University, while working as the head of public IT service department in LIG system. His current research focuses on topics such as MIS, Information Security, and mobile internet, and Public Information Policy

Yong-Tae Shin

He is a Ph.D., professor in the School of Computer Science and Engineering, Soongsil University, Seoul, Korea. His research interests focus on Multicast, IoT, Information Security, Content Security, Mobile Internet, Next Generation Internet.

Jong-Bae Kim

He received his bachelor's degree of Business Administration in University of Seoul, Seoul(1995) and master's degree(2002), doctor’s degree of Computer Science in Soongsil University, Seoul(2006). Now he is a professor in the Graduate School of Software, Soongsil University, Seoul, Korea. His research interests focus on Software Engineering, and Open Source Software.

