International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 18 (2015) pp 39302-39306

A Study on Method of Calculating Cloud Computing Server Capacity for SaaS

Kook-Hyun Choi[1], Yang-Ha Chun[2], Se-Jeong Park[3], Yongtae Shin[4] and Jong-Bae Kim[5*]

[1, 2]Dept. of IT Policy and Mgmt., Graduate School of Soongsil Univ., Seoul 156-743, Korea
[3, 5]Graduate School of Software, Soongsil University, Seoul 156-743, Korea*
[4]Department of Computer Science, Soongsil University, Seoul 156-743, Korea
E-mail: [1]khchoi@tsline.co.kr, [2]yangha00@yongin.ac.kr, [3]sejung90@naver.com, [4]shin@ssu.ac.kr, [5]kjb123@ssu.ac.kr*
[5] Jong-Bae Kim (kjb123@ssu.ac.kr) is the corresponding author of this paper.*

Abstract
For cloud computing services, which require huge server resources, it is of key importance to calculate the HW capacity of service systems more systematically and accurately. Many IT service enterprises have provided SaaS by signing MOUs with enterprises equipped with a cloud infrastructure. Thus, the research on the hardware capacity calculation method for SaaS, the base of cloud services, is the foundational research that is needed to expand cloud computing into various businesses within the cloud in the future. As such, in this research, a method and criteria are proposed for calculating the capacity of hardware, out of the computing infrastructures involved in SaaS. The results of this research may be utilized as a guideline for HW capacity calculation during the establishment of a cloud computing environment.

Keywords: Cloud Computing, SaaS, Capacity, Calculation, Criteria

1 Introduction

To calculate the hardware capacity of cloud computing, it is first of all, it is necessary to present 3 service models, or the capacity calculation models suitable to for SaaS (Software as a Service), PasS (Platform as a Service), and IaaS (Infra as a Service) depending on the nature of the IT resources. However, since IaaS provides users with infra-based services and PaaS aims to provide such development environments as a platform and OS, various aspects of approach are required to calculate HW capacity. Meanwhile, since SaaS has the same process structure as the service of WEB/WAS, it may be suitable for deducing existing service characteristics and cloud characteristics. Thus, in this research, a method and criteria for calculating SaaS-based hardware capacity are presented.

2 Related Works

Several existing studies on the calculation of H/W capacity [3][4] presented the general considerations and procedures for the performance criteria and scale calculation of each server (OLTP server, WEB/WAS server) for H/W scale calculation. These studies presented the performance calculation equations and criteria values for each of elements such as computing power, memory, and storage devices, and also presented performance calculation cases on the basis of the performance calculation criteria. However, since the methods presented in the studies are based on such existing computing environments as client/server and internet/intranet systems, it is difficult to apply them to calculating the capacity of cloud computing service. SaaS is a service model in which cloud computing service providers provide software and services through a network, and service users receive the software service after access to and certification of the network [1][7]. In other words, it uses on demand services or software so that multiple users may share and use the same software service [9][10]. The characteristics of SaaS [2][6][8] are service continuity, reliability, resource pooling, internet-based service, and virtualization.

3 Hardware Capacity Calculation for Cloud Computing

In this research, the range for calculation of SaaS-based hardware capacity is limited to CPU and memory. This is because attributes such as the size of SaaS users and the kind of available applications require hardware resources be deployed more flexibly than the existing web service. For calculation of HW capacity, first of all, the process is carried out for the direction of system establishment and the investigation of basic data. Tables 1 and 2 show a case for basic-data investigation items and investigation results.

Table 1. Basic-Data Investigation Items for DB Server Establishment

Item	Description	Investigation Results (Example)
Type of System Establishment	- Single System - High-Availability System (HA System) - Parallel Structure	- Single System
Number of Users	- Total Number of Users - Rate of Simultaneous Users - Average Number of Questions per Simultaneous User (1 Day) - Duration of Peak Time out of Operating Time - Annual Rate of User Increase	- 1, 000 persons - 80% - 3 (per minute) - 4 hours - 10%
Number of Transactions	- Annual Number of Transactions - Expected Annual Rate of Transaction Increase	- Number of Transactions per User: 4/(minute) - 20%
Amount of Assigned Work	- Ratio of Assigned Work to Online Work (Amount of Data and Data Length on the basis of Large Amount of Assignments)	- 9 : 1
Data Base	- Size of Data (Rate of Data Increase of the beginning, the 1st year, the 2nd year, the 3rd year, and after the 3rd year) - Ratio of image, sound, and text files out of data - Initial Size and Within-3-Year Size of Index Table - Number of Records of the Largest Table - Size of Data Base	- Annual Rate of Increase: 20% - 1 : 1 : 8 - 60% - 1.28 million - 50TB
Data Backup	- Are data backup servers operated? - Access Patterns of Backup Equipment - Amount of Backup Data	- No - RAID-5 - 400GB
Operating Time	- Operating Time (7x24)	- 7x24

Table 2. Basic-Data Investigation Items for WEB/WAS Server Establishment

Item	Description	Investigation Results (Example)
System Use and Service Type	- Provides only web pages - Web service without frequent transactions (DB-linked) - Web service with frequent transactions (DB-linked)	- Web service with frequent transactions (DB-linked)
System Structure Type	- Single tier - 2-tier - 3-tier	- 2-tier
Number of Connections	- Average number of Connections (on the basis of 24 hours) - Maximum number of Connections (1 hour) - Annual Rate of Connection Increase	- 1, 000 persons - 700 persons - 10% (3 years are needed in consideration of the rate of increase.)
Service Factor	- Number of Simultaneous Users - Number of Operations per User - Size of Web Page - Allowed Response Time	- 800 persons - 6/second - 5K - 3 seconds ~ 5 seconds
Importance and Urgency of Work	- Importance (High, Middle, Low) - Urgency (High, Middle, Low)	- Importance: High - Urgency: High
Type of Back-End Mutual Interaction	- Read only - Update - OLTP	- OLTP

To calculate the CPU Size of the DB server, the factors considered include the number of transactions per cloud minute, the basic tpmC correction, the correction of cloud peak-time load, the correction of cloud database size, the correction of cloud application load, the correction of cloud cluster, and the standby rate of cloud system, and the detailed calculation bases and contents are given in Table 3.

Table 3. Results for Calculation of CPU for DB Server

Item	Calculation Bases	Calculation Criteria
Number of Transactions per Cloud Minute	The number of simultaneous users is calculated in consideration of the number of current users and the annual rate of increase (10%) for the next 3 years, or 800 * 1. 1 * 1. 1 * 1. 1 = 1065 persons. The number of machines allocated when used by users is calculated on the assumption that the machines are allocated to 90% of simultaneous users or 1065 * 0.9 = 959 machines. In addition, the number of transactions generated by a user per minute ranges from 3 to 5. Based on the results of the investigation, the average number of questions is 4, and so 4 is applied.	Number of Allocated Virtual Machines * Number of Transactions per User = 959 * 4 = 3,836
Basic tpmC Correction	Since there are more than 300 simultaneous users, a corrective value of 30 % is used to apply the tpmC value measured in an optimum environment to a real environment.	1. 3
Correction of Cloud Peak-Time Load	In order to smoothly operate the system during periods in which there is a high workload, a corrective value of 30% is applied when a peak-time period of 4 hours out of the operating time occurs during specific hours every day or every week.	1. 3
Correction of Cloud Database Size	Since the number of records in the largest table is 128 million and the database is larger than 500Gbyte, a corrective value of 40% is applied to correct the number of records for database tables and the total database volume.	1. 4
Correction of Cloud Application Load	A medium corrective value of 70% is applied, considering that the online work and the assigned work are achieved almost similarly, to correct the case in which the assigned work is conducted at the same time as the online work during the peak time.	1.7
Correction of Cloud Cluster	No corrective value is provided for correction when obstacles are generated in the cluster environment, as neither backup service nor clustering is conducted.	1
Standby Rate of Cloud System	A standby rate of 30% is applied to ensure stable operation of the system in the event of an unexpected work increase.	1.3
Calculation Results	CPU(tpmC Unit) = Number of Transactions per Minute * Basic tpmC Correction * Correction of Peak Time Load * Correction of Database Size * Correction of Application Structure * Correction of Application Load * Cluster Correction * System Standby Rate = 3,836 * 1. 3 * 1. 3 * 1. 4 * 1. 8 * 1. 7 * 1 * 1. 3 = 36,104 tpmC	

The number of simultaneous users, the system region, and the memory required per user, the buffer cache correction, cluster correction, and the system standby rate should be calculated in order to calculate the scale of memory for DB server, and the bases of calculation and the contents of calculation are given in Table 4.

Table 4. Results for Calculation of DB Server Memory

Item	Contents of Calculation	Calculation Criteria
System Region	OS, DBMS Engine, Middleware Engine , Required Space of Other utilities: Basic OS + Service (Transaction and Data Base) + Other Utilities (including RAID)	128MB + 1152MB(1024MB + 128MB) + 128MB = 1408MB
Number of Allocated Virtual Machines	Number of Simultaneous Users * Rate of Increase (10%) * 3 Years * Allocation Rate of Virtual Machines (90%)	959 machines
Memory Required per User	The memory per user required to use DBMS (2MB) is applied.	2MB
Buffer Cache Correction	The general value of 20% is applied.	1.2
System Standby Rate	A standby rate of 30% is applied to prepare for unexpected situations and expansions.	1.3
Calculation Results	Memory = (System Region + Number of Virtual Machines * Memory Required per User) * Buffer Cache Correction * System Standby Rate = (1408 + 959 * 2) * 1. 2 * 1. 3 = 5,189MB	

Next, in the case of WEB/WAS server, the scale of the calculation target server with 2 tiers is measured in consideration of 8 measurement items, as shown in Table 5.

Table 5. Results for CPU Calculation of WEB/WAS Server

Item	Bases of Calculation	Calculation Criteria
Number of Simultaneous Users	Based on the current number of 800 simultaneous users, the number of simultaneous users after 3 years is 800 * 1.1 * 1.1 * 1.1=1065 persons.	1,065 persons
Number of Allocated Virtual Machines	Considering that virtual machines are allocated to 90% of the simultaneous users, the number of allocated virtual machines is calculated as 1065 * 0.9 = 956.	959 machines
Correction of Application Load Requested to Server	The corrective value for the load, which varies depending on the kind of applications requested, cannot be specifically selected and thus a general value of 10% is applied.	1.1
Correction of Cloud Interface Load	The corrective value for the load generated at the interface when a server communicates with other servers cannot be specifically selected and so a general value of 10% is applied.	1.1
Correction of Cloud Peak-Time Load	A corrective value of 30% is applied to correct the overload generated by many sudden connections during specific hours.	1.3
Standby Rate of Cloud System	A corrective value of 30% is applied to the entire work for a stable operation of the system.	1.3
Number of Operations per Cloud User	The value of 6 is applied since the web service with frequent transactions handles mainly application logics.	6
Correction of Cloud Cluster	No corrective value is provided to correct the case of obstacles generated in the cluster environment, as neither backup service nor clustering is conducted.	1
Calculation Results	OPS = Number of Virtual Machines * Correction of Application Load Requested * Correction of Interface Load * Correction of Peak-Time Load * System Standby Rate * Number of Operations per User * Cluster Correction = 959 * 1.1 * 1.1 * 1.3 * 1.3 * 6 * 1 = 11,766	

The number of simultaneous users, the system region, the memory required per user, the buffer cache correction, and the system standby rate should be calculated in order to calculate the memory scale of WEB/WAS server on the basis of the data investigated at the stage of basic data and work analysis. The calculation bases and the calculation contents are given in Table 6.

Table 6. Bases for Calculation of WEB/WAS memory

Item	Calculation Bases	Calculation Criteria
Number of Simultaneous Users	Number of Simultaneous Users * Rate of Increase (10%) * 3 years	800 * 1.1 * 1.1 * 1.1 = 1,065 persons
Number of Virtual Machines	Considering that virtual machines are allocated to 90% of the simultaneous users, the number of allocated virtual machines is calculated as 1065 * 0.9 = 956.	1065 * 0.9 = 959 machines
System Region	OS, DBMS Engine, Middleware Engine , Required Space of Other Utilities: Basic OS + Service (Transaction) + Other Utilities (including RAID)	128MB + 128MB +128MB = 384MB
Memory Required per User	The memory per user required to use applications, middleware, and DBMS	2MB
Buffer Cache Correction	The general value of 20% is applied.	1.2
System Standby Rate	A standby rate of 30% is applied to prepare for unexpected situations and expansions.	1.3
Calculation Results	Memory (MB) = {System Region + (Memory Required per User * Number of Virtual Machines)} * Buffer Cache Correction * System Standby Rate = {384MB + (959 * 2MB) } * 1. 2 * 1. 3 = 3,591MB	

As shown in Table 7, the results for calculation of the final capacity are obtained by multiplying the calculated value of the target server by the architecture correction value.

Table 7. Results for Calculation of Final Capacity

Division	DB Server		WEB/WAS Server	
	Calculated Value	Decided Value	Calculated Value	Decided Value
CPU	36,104 tpmC		11,766	18,826
Memory	10GB (10,240MB)		8GB (8,192MB)	

4 Conclusions

Research on the calculation of SaaS hardware capacity, the foundation for cloud services, is essential research that is

required to expand into various projects within the cloud. In addition, research will also be conducted on the performance measurement of each service to allocate the cloud resource to each customer who uses SaaS on the basis of this research. This research is thought to be similar to the performance calculation technique of large servers from the perspective of business operators, but it will serve as basic research to be used by users or subscribers who use cloud service.

References

[1] Kook-Hyun Choi, Yang-Ha Chun, Se-Jeong Park, Yongtae Shin and Jong-Bae Kim, "Method for Calculation of Cloud Computing Server Capacity", Proc. of ASTL, 87, (2015), 38-41.

[2] Ronnie D. Caytiles, Sunguk Lee and Byungjoo Park, "Cloud Computing: The Next Computing Paradigm", IJMUE, 7, 2, (2012), 297-302.

[3] Jonghei Ra, Kwangdon Choi, Haeyong Jung, "The Study on Hardware Sizing Method Based on the Calculating", Journal of Information Technology Services, 5, 1, (2006), 47-59.

[4] Jonghei Ra, Kwangdon Choi, "An Exploratory Study on Capacity Sizing Method for Information System : Focus on H/W Sizing in Public Sector", Journal of Information Technology Services, 3, 2, (2004), 9-23.

[5] Canturk Isci, Alper Buyuktosunoglu, and Margaret Martonosi, "LONG-TERM WORKLOAD PHASES : DURATION PREDICTION AND APPLICATION TO DVFS", IEEE Micro, 25, 5, (2005), 39-51.

[6] Cloud Security Alliance, "Top Threats To Cloud Computing V1.0." http://www.cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf

[7] P.Mell and T.Grance, "The NIST Definition of Cloud Computing" http://csrc.nist.gov/groups/SNS/cloud-computing/

[8] Cloud Security Alliance, "Security Guidance for Critical Areas of Focus in Cloud Computing V2.1." http://www.cloudsecurityalliance.org/csaguide.pdf

[9] S. Lee, "Security Considerations for Public Mobile Cloud Computing", IJAST, 44, 8, (2012), 81-88.

[10] Atiq ur Rehman, M.Hussain, "Efficient Cloud Data Confidentiality for DaaS", IJAST, 35, 1,(2011), 1-10.

Received: April 24th, 2015