Reverse Auction-based Resource Allocation Policy for Service Broker in Hybrid Cloud Environment

Sunghwan Moon, Jaekwon Kim,
Taeyoung Kim, Jongsik Lee
Department of Computer and Information Engineering, Inha University
Incheon, South Korea
e-mail: shmoon@inhaian.net, jaekwonkorea@naver.com,
silverwild@gmail.com, jslee@inha.ac.kr

***Abstract*— Hybrid cloud service utilizes public cloud and private cloud to provide its service. Furthermore, the hybrid cloud requires the resource allocation model to guarantee the Service Level Agreement (SLA), and minimize the cost. In this paper, we propose the Reverse Auction-based Resource Allocation Policy for Service Broker (RARAP). RARAP defines and utilizes the internal property of nodes on hybrid cloud environment. We simulate and evaluate the performance with the deadline compliance rate and the service usage cost. The simulation result proves the efficiency of our proposed model.**

Keywords*— *Resource Allocation Policy, Reverse Auction, Hybrid Cloud, RARAP.

I. INTRODUCTION

Globally, big data processing has become a major issue in various fields. Hence, Internet-based service is showing a tendency to rise. Therefore, a demand and importance of high-performance computing are also increasing continuously. Cloud computing utilizes the virtualization technique to construct the computing environment. It allows the cloud environment to provide high performance with distributed resources. In recent years, cloud computing has become an important part of business and industry [1].

Cloud computing is classified into three types of services. Software as a service (SaaS) aimed at providing the contents service for the user. Platform as a Service (PaaS) is concerned with processing for service requests. Infrastructure as a Service (IaaS) is interested in resource virtualization for job processing with physical resources. In addition, many services are under development for cloud computing [2].

Hybrid cloud provides data processing service using the public and private cloud. The public cloud is the paid service from external providers. On the other hand, the private cloud is the internal system with free service [3]. The collaboration with the public and private cloud may not only reduce the cost, but also increase the utilization. The service provider may also construct the resource depending on the cost. For this purpose, the system includes the service broker. The service broker automatically manages the cost to create an added value for both cloud service providers and users. This allows the hybrid cloud to minimize the cost, to utilize the various services, to manage the resource performance, and to provide the service [4]. However, hybrid cloud is vulnerable to an increasing number of service requests and complexity. Because of this, the hybrid cloud hardly provides the Service Level Agreement (SLA) for service providers and users [1]. Hence, hybrid cloud requires a new SLA-guaranteed method that minimizes cost.

In this paper, we propose the reverse auction-based resource allocation policy for service broker (RARAP) on hybrid cloud environment. RARAP defines a cost and an internal property of resources for processing a job by a deadline. RARAP utilizes the reverse auction to estimate the processing cost and allocation priority [5]. In other words, the reverse auction is to approximate the service usage cost for a resource on the hybrid cloud. Then, RARAP assigns the job to the most suitable resource using the reverse auction. RARAP ensures the SLA at a low cost in a hybrid cloud. The proposed method may be utilized for defense modeling and simulation. Battlefield data requires a large amount of computation resource to get meaningful results. Thereby, the cloud-based approach is the best choice for battlefield data analysis. And the reverse auction ensures a high efficiency for resource allocation with a reasonable cost.

The rest of this paper is structured as follows: In section 2, we briefly review the related works. Section 3 describes our key idea for cloud resource allocation policy. Section 4 explains the simulation design and results. Finally, we conclude in Section 5.

II. RELATED WORKS

A. Business Model for Resource Management on Cloud Computing

Up to now, much study has been done in the business model and job scheduling technique for cloud computing environment.

A commodity market model [6] has been proposed to connect between service providers and service users. The service provider fixes the resource fee and the parameter, which is based on the service users' usage. The commodity

 ISBN: 978-1-61208-388-9

market model has the fixed price policy for resource providing. The user does not participate in the price fixing.

An auction model [7] is most generally used in the parallel and distributed computing environment. Both service providers and service users tender the service condition. The auction model selects the service provider who suggests the most suitable condition for users' demand. Hence, the auction model shows the asymmetric feature for price fixing.

B. Cloud Resource Management and Scheduling

Resource management and scheduling is also one of the most studied topics on the cloud and distributed computing.

The cloud service often receives a complex application request from the user. The hybrid cloud utilizes the public cloud to comply with the service deadline. Because of this, Van den Bossche et al. [3] proposed a scheduling method with the cost minimizing technique. However, the cost minimizing method only considers the service cost for scheduling. This feature assigns more jobs to the free cloud service. As a result, the cost minimizing method causes a bottleneck problem on the private cloud. Hence, the variable deadline may affect the failure rate.

The ontology-based management is based on the semantic and prediction approach. The ontology-based system constructs the resource candidates with the user's requirement. The system selects the most suitable method from all the candidates to comply with the SLA [8].

This paper aims to reduce the cost and satisfy the SLA. For this, we consider the cost, performance, job size, and deadline with the model based reverse auction.

III. REVERSE AUCTION-BASED RESOURCE ALLOCTAION POLICY

We propose RARAP to minimize the cost and ensure SLA compliance. RARAP uses the reverse auction method based on the internal property of the resource. RARAP also performs the re-scheduling technique to improve the throughput. Through this, RARAP minimizes the cost for hybrid cloud services. Figure 1 shows the architecture of RARAP.

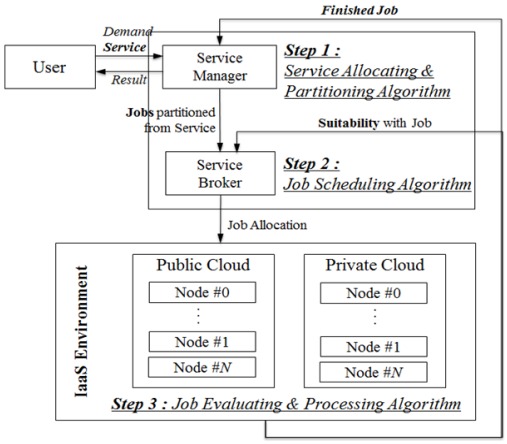

Figure 1. Architecture of RARAP

RARAP performs the procedure in a five step for resource allocation. These phases perform as follows:

A. Service request and divided into job

User is sequentially requesting services. Service manager divides the received service request in a number of jobs. For example, we assume that user requests a service provided by the save and preview the image file. Service manager divides the service into works of uploading an image file to the server, securing a storage space for image files, and creating a thumbnail for the preview. Service has the size and deadline as internal properties. Internal properties of the job follow those of the service. However, the job size is divided by the size of the service in terms of a number of jobs.

B. Delivery and classification of job

Service manager sequentially sends the job to Service broker. Service broker stores the incoming job from Service manager in the queue with the consideration of the size and the deadline. Service broker has a circular queue, and linear queues as many as the number of nodes. The job transmitted from Service broker is stored in the circular queue and waits for calculating the job suitability.

C. Job evaluation anad suitability calculation

If Service broker posts the job to be processed, all the nodes in the cloud environment return the job suitability. The job suitability is a score indicating the node efficiency of the job processing. That is, all the nodes follow the reverse auction method of competition through their performance. The score is sent to Service broker again to determine the node for job assigning. The job assigned to the node is stored in a linear queue.

D. Job processing and updating of the nodes

All nodes in the hybrid cloud are waiting to receive the job from Service broker. The linear queue of Service broker is a job queue for each node. The broker delivers the waiting job on the queue to the node, when the node is empty. The node processes the received job, and then updates to ready state in order to process the next job. The node also sends the finished job to Service manager.

E. Job merging and returning serivce

The finished job waits on Service manager for merging. When all jobs belonging to the same service have arrived, the jobs are merged into the service. The merged service is presented to the requesting user.

IV. SIMULATION DESIGN AND RESULTS

We design the hybrid cloud environments to test the effectiveness of our proposed RARAP. The environment is based on the Discrete Event System Specification (DEVS)

 ISBN: 978-1-61208-388-9

formalism [9], and measures the usage cost and the deadline compliance rate.

A. Simulation Design

We design the simulation model based on Figure 1. This simulation is designed to demonstrate the following effects of the RARAP in small hybrid cloud environment. The first is to ensure the SLA with the compliance of the job deadline. The last is the cost reduction for the same service requirements.

User sends the service request to Service Manager. Service manager divides the service into jobs, and distributes the divided jobs to the Service broker. Every node calculates the job suitability score, and returns the result to the Service Broker. Then, the Service broker finally assigns the job to the specific node. The node processes the assigned job from the Service broker. The solved job is transmitted from the node to the Service manager. Then, the Service manager merges the jobs into the service, and returns it to the user.

The hybrid cloud utilizes both the private and public nodes for service. We define the performance of both nodes for simulation as shown in Table I. The processing speed in Table I is in exact proportion to the processing time. The lower value for processing speed indicates the less time for job processing.

TABLE I. SIMULATION CONFIGURATION – NODE PERFORMANCE

Node Number	Node Type	Processing Speed	Usage Cost
0	Private	5.5	0
1	Private	10.5	0
2	Private	2.6	0
3	Public	1.1	7
4	Public	1.7	5
5	Public	2.7	2

The public node takes a service usage cost to provide the public cloud service such as Amazon Web Service [10] or Window Azure [11]. On the other hand, the private node refers to the SOHO server and network attached storage that may be held by individuals or small companies. Table II shows the usage cost policy for public node on our simulation. This pricing policy is defined on a scale from 0 to 10 according to the CPU performance, which is offered by the public cloud service. The price is based on the size and the deadline used in the service parameter for experimental environment.

TABLE II. SIMULATION CONFIGURATION – PUBLIC CLOUD USAGE COST

Usage Cost	1	2	3	4	5	6	7	8	9	10
Processing Speed	0.2	0.5	0.8	1.1	1.4	1.7	2.0	2.3	2.6	2.9

As mentioned above, each node calculates the job suitability score. The estimation result is based on the processing speed and cost. Table III shows score tables for each factor.

TABLE III. SCORE FOR CALCULATING SUITABILITY

Speed Score	1		2		3		4		5	
Processing Speed	22.8 ~ 25.0		20.6 ~ 22.8		18.4 ~ 20.6		16.2 ~ 18.4		14.0 ~ 16.2	
Speed Score	6		7		8		9		10	
Processing Speed	11.8 ~ 14.0		9.6 ~ 11.8		7.4 ~ 9.6		5.2 ~ 7.4		3.0 ~ 5.2	
Speed Score	11		12		13		14		15	
Processing Speed	2.6 ~ 2.9		2.3 ~ 2.6		2.0 ~ 2.3		1.7 ~ 2.0		1.4 ~ 1.7	
Speed Score	16		17		18		19		20	
Processing Speed	1.1 ~ 1.4		0.8 ~ 1.1		0.5 ~ 0.8		0.2 ~ 0.5		0.0 ~ 0.2	
CostScore	1	2	3	4	5	6	7	8	9	10
Cost	10	9	8	7	6	5	4	3	2	1
Processing Speed	0.2	0.5	0.8	1.1	1.4	1.7	2.0	2.3	2.6	2.9

SpeedScore in Table III is defined on a scale from 0 to 20 according to the performance of all the developed CPU for a personal computer [12]. We assume that the performance of CPU, which is held by nodes in the cloud, is proportional to the processing speed. However, since the public cloud is a paid service, the suitability calculation considers this feature for grading the nodes. Each node converts its own performance information to SpeedScore and CostScore by using Table III. The node calculates the job suitability using (1).

```
If(Node of Private Cloud)
Job Suitability = SpeedScore

If(Node of Public Cloud)
Job Suitability = SpeedScore - CostScore
```

(1)

In our simulation, the user requests the services from a minimum of 50 up to 500. Both size and deadline of service are based on the Wikipedia Page Traffic V3 Statistic [13], which is opened through the public data sets of Amazon Web Service. We use this public data to the processing to meet the needs of our environment.

We measure the service performance with three different models.

- First, we use the sequentially assigned model for cloud service, "round-robin". The round-robin model sequentially assigns jobs to all nodes. In other words, the job is assigned in the order of nodes, regardless

of the performance indices. Therefore, the service is returned in the order requested from the user.

- Second, we use the randomly assigned model with a table of random numbers. The random model distributes the job on the basis of the calculated suitability. However, this model will randomly select a node from the candidate group.
- Last is our proposed method RARAP. RARAP allocates the job to the node having the highest goodness of suitability as mentioned above.

B. *Simulation Results*

We measure the deadline compliance rate and the total usage cost for each comparison model. The purpose of this experiment is to verify the proposed RARAP can guarantee the SLA at an affordable cost.

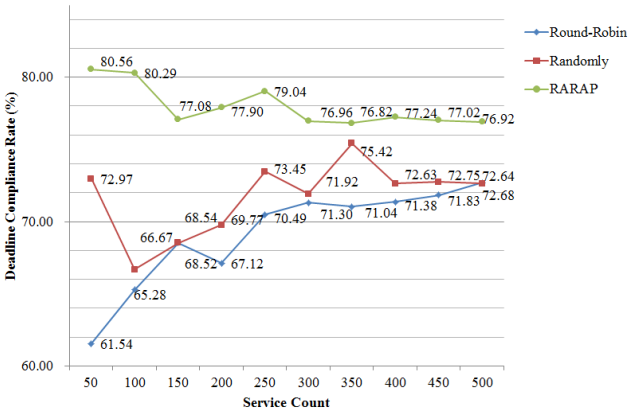

Figure 2. Result of Graph for Deadline Compliance Rate

Figure 2 shows the measured result for deadline compliance rate. As presented in (2), the deadline compliance rate is the percentage of solved jobs before the deadline. It may show the processing efficiency of each model.

$$\text{DeadlineComplianceRate}(\%) = \frac{\text{DeadlineComplianceJob}}{\text{TotalCompletedJob}} \quad (2)$$

As shown in Figure 2, the round-robin model records a deadline compliance rate of 69.118%, the random model records 71.676%, and our proposed RARAP records 77.983%. This value is an average percentage of the deadline compliance result. Our proposed RARAP considers the deadline to assign the job. As a result, our model shows superior compliance rate and less variation than other models.

Figure 3 shows the other measured result for processing cost. This result is to present the price effectiveness for each model. We measure the processing cost in the same throughput for fair comparison.

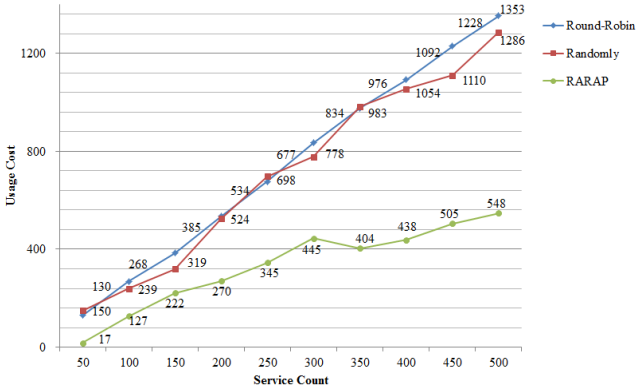

Figure 3. Result of Graph for Usage Cost

As shown in Figure 3, the round-robin model records 747.70, the random model records 714.10, and our proposed model RARAP records 332.10. In our design, only the public node charges the service usage cost with the price policy shown in Table III. Our model tries to minimize the processing cost. RARAP avoids the public node under the same conditions. The public node is inevitable choice for our model. This method induces the least processing cost for RARAP.

V. CONCLUSION AND FUTURE WORK

This paper proposes reverse auction-based resource allocation policy for service broker in hybrid cloud environment. RARAP utilizes the reverse auction method, and assigns the resource with three steps. RARAP controls the job schedule with the job suitability score, which is based on the processing speed and service usage cost. It may improve the efficiency, and decrease the cost and the number of SLA violations.

Future work will concentrate on the data partitioning. The interval-based partitioning management can increase the utilization per cost for cloud resources. Our study will be used in the analysis of battlefield data. The defense simulation has traditionally required a large amount of processing resources. The proposed model is expected to be able to meet the analysis data required for war game simulation.

ACKNOWLEDGMENT

This research was supported by Defense Acquisition Program Administration and Agency for Defense Development under the contract UD140022PD, Korea, and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2002751).

REFERENCES

[1] J. Koh, H. Kang, and Y. Kim, “Adaptive policy-based task scheduling for scientific applications in hybrid cloud,” Journal

of KIISE: Computing Practices and Letters, 19th vol.11, 2013, pp. 572-579.

[2] J. Kim and J. Lee, "Fuzzy logic-driven virtual machine resource evaluation method for cloud provisioning service," Journal of The Korea Society for Simulation, 22nd vol.1, 2013, pp. 77-86.

[3] R. Van den Bossche, K. Vanmechelen, and J. Broeckhove, "Cost-optimal scheduling in hybrid iaas clouds for deadline constrained workloads," Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference, 2010, pp. 228-235.

[4] J. Kim, D. Kang, N. Kim, J. Lee, and S. Jung, "Cloud service broker managing and integrating multiple heterogeneous clouds," Communications of KIISE, 32nd vol.2, 2014, pp. 52-58.

[5] S. D. Jap, "The impact of online reverse auction design on buyer-supplier relationships," Journal of Marketing, 71st vol.1, 2007, pp. 146-159.

[6] Y. Amir, B. Awerbuch, A. Barak, R. S. Borgstrom, and A. Keren, "An opportunity cost approach for job assignment in a scalable computing cluster," Parallel and Distributed Systems, IEEE Transactions, 11st vol.7, 2000, pp. 760-768.

[7] M. Stonebraker et al., "An economic paradigm for query processing and data migration in Mariposa," Parallel and Distributed Information Systems, 1994., Proceedings of the Third International Conference, 1994, pp. 58-67.

[8] Y. B. Ma, S. H. Jang, and J. S. Lee, "Ontology-based resource management for cloud computing," Intelligent Information and Database Systems, Springer Berlin Heidelberg, 2011, pp. 343-352.

[9] B. P. Zeigler, H. Praehofer, and T. G. Kim, "Theory of modeling and simulation: integrating discrete event and continuous complex dynamic systems," Academic Press, 2000, pp. 76-96.

[10] Amazon Web Service. [Online]. Available from: http://aws.amazon.com/ 2013.11.01

[11] Windows Azure. [Online]. Available from: http://www.windowsazure.com/ 2013.11.01

[12] CPU Benchmarks by PassMark Software. [Online]. Available from: http://www.cpubenchmark.net/cpu_list.php 2014.03.04

[13] Wikipedia Page Traffic V3 Statistic. [Online]. Available from: https://aws.amazon.com/datasets/6025882142118545 2013.10.05

 ISBN: 978-1-61208-388-9