한국시뮬레이션학회 논문지
Vol. 23, No. 4, pp. 143-150 (2014. 12)
http://dx.doi.org/10.9709/JKSS.2014.23.4.143
ISSN 1225-5904

A Mathematical Programming Approach for Cloud Service Brokerage

Byeong-Yun Chang · Yabibal Afework Abate · Seung Hyun Yoon · Dong-Won Seo*

클라우드 서비스 중개를 위한 수리과학 모형연구

장병윤 · Yabibal A. Abate · 윤승현 · 서동원*

ABSTRACT

Cloud computing is fast becoming the wave of the future for both home and business computing. Because of this growing acceptance, we can expect an explosion of diverse cloud service providers in the coming years. However, the cloud is not a single entity, rather it is a set of many disconnected islands of application (SaaS), infrastructure (IaaS), and different platform (PaaS) services. Cloud brokering mechanisms are essential to transform the heterogeneous cloud market into a commodity-like service. Cloud service brokers (CSBs) are the new form of business entities to help to aggregate the scattered set of cloud services and make conveniently available to their diverse users. CSBs can reserve a certain percentage of their clients' (users') demand and satisfy the remaining portion in an on-demand basis. In doing so, they need to minimize cost of both reserved and on-demand instances as well as the distance of a link between the cloud service provider (CSP) and the user. The study proposes a reservation approach with a mixed integer model to optimizes the cloud service cost and quality.

Key words : Cloud computing, CSB, Reservation approach, Mixed integer programming

요 약

클라우드 컴퓨팅은 가정과 비즈니스 컴퓨팅에 있어 빠르게 확산되고 있다. 이러한 변화에 따라 수년 내에 다양한 클라우드 서비스를 제공하는 클라우드 서비스 공급자의 수가 폭발적으로 증가할 것으로 예측된다. 하지만 클라우드는 하나의 독립체가 아니고 서로 단절되어 있는 많은 응용 프로그램(SaaS)과 기반시설(IaaS), 상이한 플렛폼 서비스(PaaS)들의 집합으로 볼 수 있다. 클라우드 중개방법(brokering mechanism)은 산재되어 있는 서비스들의 집합을 통합하고 편리하게 유용한 서비스를 다양한 사용자에게 제공할 수 있게 하기 때문에 서로 상이한 클라우드 시장을 하나의 통상적 서비스로 변환시키는데 있어 필수적이다. 따라서 CSBs(클라우드 서비스 중개자들)는 사용자와 CSP(클라우드 서비스 제공자)사이의 거리를 최소화하고 예약과 on-demand 서비스의 비용을 최소화하는 운영전략을 찾을 필요가 있다. 이를 위하여 본 연구에서는 클라우드 서비스의 비용과 품질을 최적화하기 위해 예약 접근법을 고려한 수리적 모형을 제안한다.

주요어 : 클라우드 컴퓨팅, 클리우드 서비스 중개, 예약 접근법, 혼합정수계획법

1. Introduction

The use of cloud based services in order to provide online services to customers brings a new era in the ICT infrastructure and delivery. The number of providers in the cloud computing market is increasing at rapid pace. At the same time, we are observing a fragmentation of the market in terms of pricing schemes. In a simplistic

*본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [14-000-05-001, 스마트 네트워킹 핵심 기술 개발]

Received: 25 October 2014, **Revised:** 24 November 2014, **Accepted:** 2 December 2014

***Corresponding Author:** Dong-Won Seo
E-mail: dwseo@khu.ac.kr
School of Management, Kyung Hee University

scenario, a service provider (SP) decides to host the service it wants to provide to an end customer on the infrastructure provided by an infrastructure provider (IP). This is the current state of art in the use of the cloud services and is considered as a simple scenario because of the limited flexibility available to a SP to split its services into components and deploy them into infrastructures provided by multiple IPs.

However it is expected that the growth and maturity of the cloud offerings would necessitate the building of delivery models that will allow the use of multiple IPs to host multiple components in a single service[10]. At the same time, with many different clouds serving various needs, implementing a cloud services delivery infrastructure around a few offerings today raises the risk of managing services stovepipe tomorrow. And given not all services they may want to sell can reside in their own data centers, service providers must also contend with how they can deliver externally provided services, along with their own, in a manner of providing a seamless user and life cycle management experience for their customers.

To retain existing customers, grow market share and maintain existing profit margins on core services, CSPs need an edge on their competition and to increase agility in addressing emerging market opportunities. In what is often considered their “core” business. CSPs are facing new competition from companies that are offering a range of value-added services to small, medium and large enterprises. For most providers, it's no longer a question of rather how they make that transition from a traditional telecommunication provider to a ‘technology services’ company.

In the early phase of cloud adoption, the price model was dominated by fixed prices[11]. However, cloud market trend shows that dynamic pricing schemes. In this plan, prices change according to demand in each cloud provider. In general, it is difficult for users to search cloud prices and decide where to put their resources. As the cloud computing market grows, the market complexity is also increasing as users have to deal with many different Virtual Machine (VM) types, pricing schemes and cloud interfaces. This trend has brought about the use of federated clouds[5, 8, 11] and multi-cloud deployments.

In this context, cloud brokering mechanisms are essential to transform the heterogeneous cloud market into a commodity-like service. Moreover, a cloud broker could be used to provide a uniform interface with VM’s management. Also, it could offer mechanisms required to schedule VMs placement among multiple clouds, algorithms to optimize deployments, or automatic data collection. As cloud services multiply and expand faster than the ability of cloud consumers to manage or govern, cloud services brokerages will be required to enable the trustworthy usage of services by cloud consumers. The CSB will help organizations select, manage and coordinate the various services they require[10]. The cloud service buyer is becoming an IT broker of physical/virtual applications and services. Service providers can gain by giving IT buyer’s new choices for integrating and managing across public, private, hybrid and legacy environments. CSBs will adopt a supply-chain model to track and trace multi-vendor service provision. At the same time, they will be responsible for complex product and service integrations to meet a customer’s overall service requirements.

2. Review of the concept and function of CSBs

The national institute of standards and technology (NIST) of the United States of America has a working definition of cloud service brokerage. According to NIST[8], a cloud service brokerage (CSB) is a business model in which a company or other entity adds value to one or more (generally public, hybrid, and possibly private) cloud services on behalf of one or more consumers of those services[9].

Even though many agree on the important role of CSBs in the cloud industry, how to fit their role with the existing players is very arguable for many researchers. Yet most of the researchers agree on that CSBs facilitates the dynamic sourcing of services and capacity; acts as the integrator of services to a customer; and provides workload management, including the traditional

Table 1. CSB categories based on service type

Category (Gartner Term)	Description
Customer-centric (Cloud Service Intermediation)	Responsible for fulfilling a customer's business service requirements. CSBs are also responsible for the governance and quality of service.
Solution-centric (Aggregation)	Responsible for assembling services offerings from a range of services available in the cloud. Ensures the security and integrity.
Asset centric (Arbitrage)	Provides flexibility and opportunistic choices for the service aggregator.

ITIL (Information Technology Infrastructure Library related services), compliance, and governance functions, regardless of whether they are sourced from the cloud or via traditional IT service models. In the cloud industry, brokers are responsible for and must manage the risk of failure, low service quality, inadequate security assurance, and liability between providers and consumers - all through a relationship in which the broker is the customer's single contact point for multiple cloud services.

There are three primary categories of CSB provider based on the type of service they offer, customer-centric, solution-centric and resource/asset-centric as each are described in Table 1.

The CSB business may go far beyond current engagement models and service propositions on the market today. It can offer a unique opportunity to facilitate the rapid commoditization of the existing service business and build a true service aggregation platform that can leverage a wide range of service portfolio elements[9]. In short, CSB will become a premium service offering for sophisticated service providers[4, 15].

3. Research Methodology

Thanks to the economies of scale, the cloud service is capable of offering such on-demand computational services at a low cost[1, 3, 13]. Cloud users usually pay for the usage (counted by the number of instance-hours incurred) in a pay-as-you-go model, and are therefore freed from the prohibitive upfront investment on infrastructure, which is usually over-provisioned to accommodate peak demands. A cloud provider prefers users with predictable and steady demands, which are friendlier to capacity planning. In fact, most cloud providers offer an additional pricing option, referred to as the reservation option, to harvest long-term risk free income. Specifically, this option allows the user to prepay a one-time reservation fee and then to reserve a computing instance for a long period (usually in the order of weeks, months or years) during which the usage is either free or charged under a significant discount[1, 3, 13].

If fully utilized, such a reserved instance can easily save its user more than 50% of the expense. However, whether and how much a user can benefit from the reservation option critically depends on its demand pattern. Due to the prepayment of reservation fees, the cost saving of a reserved instance is realized only when the accumulated instance usage during the reservation period exceeds a certain threshold (varied from 30% to 50% of the reservation period)[1, 13]. Unless heavily utilized, the achieved saving is not significant. For this reason, users with sporadic and bursty demands only launch instances on demand.

Unfortunately, on-demand instances are economically inefficient to users, not only because of the higher rates that providers charge, but also because there is a fundamental limit on how small the billing cycle can be made[4, 15]. For example, Amazon Elastic Compute Cloud (EC2) charges on-demand instances based on running hours. In this case, an instance running for only 10 minutes is billed as if it were running for a full hour[1]. Similar limitations are also imposed by other cloud providers, e.g.,[6, 7]. Such billing inefficiency becomes more salient when longer billing cycles are employed[14].

Intuitively, the cloud broker leverages the "wholesale" model and the pricing gap between reserved and on-demand instances to reduce the expenses for all the users. More importantly, the broker can optimally coordinate different users to achieve additional cost savings, along with other benefits.

The broker asks cloud users to submit their demand estimates over a certain horizon, based on which dynamic reservation decisions are made. Note that even if a user trades directly with cloud providers, it needs to estimate its future demand to decide the amount of instances to reserve for a particular time. In the case where users are unable to estimate demand at all, we propose an online reservation strategy to make decisions based on history.

Along with minimization of costs of both reserved and on-demand instances, a CSB should give a great deal of emphasis on the quality of service it can provide to its users and often, quality is a function of distance[2, 12]. Therefore, total distance of the link between providers and users has to be minimized so as to improve quality of service.

In the next section, we have discussed the Mixed Integer Programming model that minimize the total cost of the cloud service (both reserved and on-demand instances) and also the distance between the specific users and the CSB while meeting all the constraints related with capacity restrictions, bandwidth limitations and other constraints.

4. Mathematical Model

Suppose cloud users submit to the broker their demand estimates up to time t. The broker aggregates all the demands. Suppose it requires d_j instances in total to accommodate all the requests for user j at time t. The broker makes a decision to reserve r_{ij} instances at time t, with $r_{ij}(\geq 0)$ from provider $i(i \in P)$ to user $j(j \in U)$ where P and U are the sets of cloud service providers and users respectively. Note that at time t, these reserved instances may not be sufficient to accommodate the aggregate demand, DD_j of user j. The broker thus needs to launch additional on-demand instances at $o_{ij}(\geq 0)$ from provider i to user j. The CSB should also give a great deal of emphasis on the quality of service it provides. Here, we assume that their service quality depends only on the total distance. To this end, to improve the quality of service it suffices to reduce d_{ij}, the total distance between provider i and user j.

· **Variables and description:**

$$x_{ij} = \begin{cases} 1\ if\ user\ j\ is\ assigned\ to\ provider\ i \\ 0\ otherwise \end{cases}$$

c_{ij}: cost of reservation from provider i to user j

p_{ij}: cost of on-demand instances from provider i to user j

d_{ij}: distance between provider i and user j.

$r_{ij} \in R_j$: the amount of reserved instances from provider i to user j where R_j is the total amount of reserved instances that can be made for user j, i.e. $R_j = \sum_{i \in P} r_{ij}$

$o_{ij} \in O_j$: the amount of on-demand instances from provider i to user j where O_j is the total on-demand instances for user j which is the difference between total requirement and percentage of reservation, i.e. $O_j = DD_j - R_j$

DD_j: total demand (requirement) of user j

Ca_i: total capacity of provider i

bw_{ij}: available bandwidth capacity between provider i and user j

α: the maximum link capacity utilization rate α (say, 95%)

· **Objective:**

$$Minimize \sum_{i \in P} \sum_{j \in U} (c_{ij} r_{ij} + p_{ij} o_{ij} + d_{ij} x_{ij}) \tag{1}$$

· **Constraints:**

$$\sum_{i \in P} (r_{ij} + o_{ij}) \geq DD_j \text{ for } j \in U \tag{2}$$

$$\sum_{j \in U} (r_{ij} + o_{ij}) \leq Ca_i \text{ for } i \in P \tag{3}$$

$$\sum_{i \in P} r_{ij} \leq R_j \text{ for } j \in U \tag{4}$$

$$\sum_{i \in P} o_{ij} \leq O_j \text{ for } j \in U \tag{5}$$

$$DD_j \times x_{ij} \leq \alpha \times bw_{ij} \text{ for } i \in P, j \in U \quad (6)$$

· **Domain constraints:**

$$r_{ij} \geq 0 \text{ and } o_{ij} \geq 0 \text{ for } i \in P, j \in U \quad (7)$$

$$x_{ij} \in \{0,1\} \text{ for } i \in P, j \in U \quad (8)$$

Remarks: The objective function (1) consists three basic terms. The first two terms are to minimize the cost of both reserved and on-demand instances. The third term is to minimize the cost related to the distance between provider i to user j. To simplify we assume that this cost is incurred linearly by distance in this model.

Constraint set (2) sets the sum of reserved and on-demand instances to meet the total demand (requirement) of user j. Constraint set (3) guarantees that both reserved and on-demand instances do not exceed its maximum capacity. Constraint sets (4) and (5) insure that the sum of all reserved and on-demand instances provided from provider i to user j, respectively, do not exceed the maximum amount of reservation and on-demand instance requested from j.

Constraint set (6) is adopted to insure that if user j's requirement is assigned to provider i, the total requirement of the user should not exceed the bandwidth capacity of the link between them adjusted to the maximum link utilization rate (95% in most cases). Constraint set (7) assures that there cannot be a negative reserved and on-demand instance for each user. Finally, constraint set (8) insures that the binary variable x_{ij} is set whenever there is a traffic over the $link(i,j)$.

5. Mixed Integer Programing Model for a Hypothesized Example

Suppose there are three cloud service providers and two consumers (users) and they have one common CSB agent as illustrated in Fig. 1.

Users requirement (demand) information for the two users, capacity information for the three providers and percentage of reservation information is given in Table 2. For instance, total requirements of the first user , U_1, is 60 instances of which 90% can be reserved. The first CSP, P_1 has a total capacity of 140 instances. The unit of measurement for capacity, cost, and distance are not specified. A different unit of measure can be used for a different application of the model.

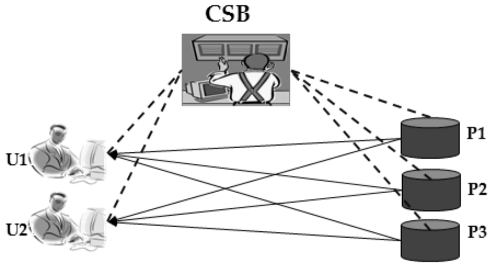

Fig. 1. Simple cloud brokerage business model

Table 2. Demand and capacity information

User	DD_j	R_j	CSP	Ca_i
U_1	60	90	P_1	140
U_2	75	88	P_2	60
			P_3	75

Table 3. Cost, link capacity and distance information

CSP	Users					
	U_1			U_2		
	c_{ij}/p_{ij}	bw_{ij}	d_{ij}	Cost (r/o)	bw_{ij}	d_{ij}
P_1	12/10	80	100	15/16	65	500
P_2	12/11	75	500	12/11	70	700
P_3	13/10	70	250	11/10	80	200

Cost information for both reserved, and on-demand instances, link capacity as well as distance information for the 2 users and 3 providers is given in Table 3. For example, it costs 12 and 10 for a reserved and on-demand instance from P_1 respectively. Bandwidth capacity of the link between P_1 and U_1 is 80 per unit of time while the distance is 100.

Using the above given parameters, we run the mathematical model assuming that there is no content constraint, i.e. provider i has all types of resources that all j users required. The summary of the results is

Table 4. Assignment result of 2 users and 3 CSPs

User	CSP	Reserved	On-demand
U_1	P_1	54	6
U_2	P_3	66	9
Total cost		1,671	

Table 5. Capacity and requirement information

CSP	Ca_i	User Requirements		
		User	DD_j	R_j
P_1	140	U_1	45	0.95
P_2	60	U_2	63	0.85
P_3	75	U_3	70	0.90
P_4	10	U_4	55	0.88
P_5	90	U_5	65	0.99
P_6	95			
P_7	110			
P_8	75			
P_9	80			
P_{10}	85			

provided in Table 4.

Requirements of user U_1 can be satisfied from provider P_1 while the requirements of user U_2 are assigned to provider P_3. The amount of reservation for U_1 and U_2 is 54 and 66 respectively where as the amount of on-demand instance is 6 and 9 for U_1 and U_2 respectively. Provider P_2 is not assigned. The total cost for the CSB is 1,671.

The above example and given information was extended to 5 users and 10 CSPs. Parameters information is given and presented as follows.

Table 5 presents information on capacities and requirements. Available capacity of each CSP, Ca_i, and the total requirement of each user, DD_j, with a reservation percentage, R_j. Fore example, provider P_1 has a total capacity of 140 instances. The total requirement of user U_1 is 45 of which 95% can be reserved with the remaining 5% to be satisfied in an on-demand basis.

Table 6. Cost of reserved and on-demand instances, link capacity, link distance, and content availability

CSP	Users	c_{ij}	p_{ij}	bw_{ij}	d_{ij}	Content
P_1	U_1	10	18	65	15	1
P_1	U_2	12	16	65	11	0
P_1	U_3	12	16	65	13	1
P_1	U_4	12	18	65	10	1
P_1	U_5	13	16	65	12	1

Table 7. Summary of the MIP results from MathProg.

User	CSP	Reserved	On-demand
U_1	P_2	42.75	2.25
U_2	P_2	53.55	9.45
U_3	P_{10}	63.00	7.00
U_4	P_6	48.40	6.60
U_5	P_{10}	58.50	6.50
Total cost		3,511.5	

Information regarding cost of reserved instances, cost of on-demand instances, link capacity, link distance, and content availability between CSP (cloud service provider) P_1 and each j user respectively is given in Table 6. For instance, to assign U_1 to P_1, it costs 10 and 18 for reserved and on-demand instances respectively; bandwidth capacity of the link between them is 65 instances at a time; distance between them is 15 and U_1 requirements are available at P_1. Content equal to 1 if the type of resource or service that a user, U_j, requirement is available at a given CSP, P_i; 0 otherwise. Therefore, the binary variable x_{ij} will be forced to 0 if content is 0.

Information regarding the rest of CSPs follows the same format. Hence, it is not provided here to avoid redundancy.

Using the GLPK MathProg programing software, the mixed integer assignment model was solved. A summary of results are presented in Table 7. Both users, U_1 and U_2 were assigned to P_2. Amount of reservation/on-demand instances to be provided to U_1 and U_2 are 42.75/2.25 and 53.55/9.45 respectively. Requirements of all users were satisfied using only P_2, P_6 and P_{10}. The minimum total cost of both reserved and on-demand instances satisfying all other constraints was found to be 3,511.5.

6. Conclusion

The future of computing lies on the hand of the cloud as many have said. However, the existing structure of the industry is that there are providers with a specific service type which could be SaaS, IaaS or PaaS. On the other spectrum of the industry, there are the service users who want to consume a combination of the above services in most cases. In such a situation, a provider can render the service in which it specializes on, and outsource other services from a third party provider. This is what the CSB framework is and how the cloud service can be simplified[3].

Moreover, using the reservation technique, the broker can aggregate the forecasted demand of its clients and make a reservation from multiple providers. By doing so, the cost of the service can be minimized since providers can maintain a fixed capacity as far as they have sufficient information about the requirements of their users. Quality of the service is also considered in this study as it depends on distance of the link between a provider and the user and the utilization rate of this link. The mathematical model we have proposed above optimizes both cost and quality of the cloud service.

References

1. Amazon EC2 Pricing http://aws.amazon.com/ec2/pricing/.
2. Amiri, A. and Pirkul, H., “Routing and capacity assignment in backbone communication networks under time varying traffic conditions”, European Journal of Operational Research, Vol. 117, No. 1, pp. 15-29, 1999.
3. Armbrust, M., Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A.I., Stoica, A. and Zaharia, M., “A view of cloud computing”, Communications of the ACM, 2010.
4. Calcavecchia, N.M., Celesti, A. and Nitto, D.E., “Understanding decentralized and dynamic brokerage in federated cloud environment”, Milan Politech, 2012.
5. David, M.S., “Hype cycle for cloud computing”, Gartner, 2011.
6. ElasticHosts, http://www.elastichosts.com/.
7. GoGrid Cloud Hosting, http://www.gogrid.com.
8. Liu, F., Tong, J., Mao, J., Bohn, R., Messina, J., Badger, I. and Leaf, D., “NIST cloud computing architecture”, The national institute of standards and technology, USA, 2011.
9. Morison, D., “The evolution of cloud service brokerage”, 2012 experts forum, 2012.
10. Nair, S.K., Pawer, P., Sajjad, A., Kiran, M. and Jiang, M., “Requirements and architecture of a cloud broker”, Optimis 7th framework program, 2010.
11. Schubert, L., “The Future of Cloud Computing”, 2010.
12. Simarro, J.L.L., Vozmediano, R.M., Montero, R.S. and Llorente, I.M., “Scheduling strategies for optimal service deployment across multiple clouds”, Future Generation Computer Systems, Vol. 29, No. 6, pp. 1431-1441, 2013.
13. Stoica, A.I. and Zaharia, M., “A view of cloud computing”, Communications of the ACM, 2010.
14. VPS.NET, http://vps.net.
15. Wang, W., Nui, D., Li, B. and Liang, B., “Dynamic cloud resource reservation via cloud Brokerage”, University of Toronto, 2013.

장 병 윤 (bychang@ajou.ac.kr)

1995 성균관대학교 산업공학과 학사
2000 Georgia Tech. Operations Research 석사
2002 Georgia Tech. Applied Statistics 석사
2004 Georgia Tech. Industrial and Systems Engineering 박사
2004～2006 Georgia Tech. Post Doc.
2006～2009 KT 네트워크 연구소 선임 연구원
2009～현재 아주대학교 경영대학 경영학과 부교수

관심분야 : 정보통신경영, BPM, OR/OM, SCM, Simulation, Applied Statistics

Abate, Yabibal Afework (abate@ajou.ac.kr)

2010 Addis Ababa University Procurement and Supply Chain Management, 학사
2012 아주대학교 International Business MBA
2013～현재 아주대학교 Management Science and Operations Management, 박사과정

관심분야 : SCM, Operations Management, Management Science, Applied Statistics, Organizational Behavior

윤 승 현 (shpyoon@etri.re.kr)

1991 성균관대학교 산업공학과 학사
1993 성균관대학교 대학원 산업공학과 석사
1997 성균관대학교 대학원 산업공학과 박사
1997～현재 한국전자통신연구원 책임연구원

관심분야 : 통신망 설계, 망관리, 통신망 최적화, 스마트노드

서 동 원 (dwseo@khu.ac.kr)

1991 성균관대학교 산업공학과 학사
1996 성균관대학교 대학원 산업공학과 석사
2002 Georgia Institute of Technology, 산업공학 박사
2003～현재 경희대학교 경영대학 경영학부 교수

관심분야 : 확률과정론, Series Expansion, (Max,+)-algebra, 시뮬레이션