논문 15-40-09-20 The Journal of Korean Institute of Communications and Information Sciences '15-09 Vol.40 No.09
http://dx.doi.org/10.7840/kics.2015.40.9.1802

클라우드 컴퓨팅을 위한 VM 스팟 인스턴스 입찰 최적화 전략

최 영 호◆, 임 유 진°, 박 재 성*

Optimal Bidding Strategy for VM Spot Instances for Cloud Computing

Yeongho Choi◆, Yujin Lim°, Jaesung Park*

요 약

클라우드 컴퓨팅 서비스는 가상화 기술을 이용하여 물리적인 IT 자원을 VM 단위로 사용자들에게 비용을 받고 제공하는 서비스이다. 그 중 클라우드 컴퓨팅 기반 가용 자원 경매 모델은 서비스 제공자의 가용 자원을 경매를 통해서 사용자들에게 제공 하는 서비스이다. 서비스 이용자들은 제한 시간 안에 그들의 작업을 처리하기 위해 서비스 제공자에게 자원 이용에 대한 입찰가격을 제시하고 낙찰 가격 보다 높은 경우 자원을 제공 받는다. 본 논문에는 Amazone EC2에서 서비스 이용자의 작업을 완료하는데 요구되는 스팟 인스턴스에 대한 총 경비를 최소화하는 입찰 기법을 제안한다. 일반적으로, 서비스 이용자는 자원을 할당 받기 위해 높은 입찰 가격을 제시할 것이고, 그에 따라 낙찰 가격이 높아짐으로써 서비스 이용자의 실제 작업 비용은 높아지게 된다. 따라서 제안 입찰 기법을 이용하여 낙찰 가격을 낮춤으로써 서비스 이용자의 총 경비를 최소화 할 수 있다. 제안 기법의 성능 분석을 위해 실제 데이터를 이용하여 낙찰 가격과 실제 총 경비를 계산하고, 실제 낙찰 가격 기반의 입찰 기법과 비교함으로써 제안 기법의 성능을 입증하였다.

Key Words : cloud computing, auction, QoS, spot instance, bidding strategy

ABSTRACT

The cloud computing service provides physical IT resources to VM instances to users using virtual technique and the users pay cost of VM instances to service provider. The auction model based on cloud computing provides available resources of service provider to users through auction mechanism. The users bid spot instances to process their a job until its deadline time. If the bidding price of users is higher than the spot price, the user will be provided the spot instances by service provider. In this paper, we propose a new bidding strategy to minimize the total cost for job completion. Typically, the users propose bidding price as high as possible to get the spot instances and the spot price get high. we lower the spot price using proposed strategy and minimize the total cost for job completion. To evaluate the performance of our strategy, we compare the spot price and the total cost for job completion with real workload data.

※ 본 연구는 경기도의 경기도지역협력연구센터(GRRC) 사업의 일환으로 수행하였음 [(GRRC수원2015-B3), 클라우드 기반 지능형 영상 보안 감시 시스템 개발].
※ 본 연구는 2015년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업(NRF-2011-0007076)의 지원으로 수행되었습니다.
◆ First Author : University of Suwon, Department of Computer Science, ceewoo@suwon.ac.kr, 학생회원
° Corresponding Author : University of Suwon, Department of Information Media, yujin@suwon.ac.kr, 종신회원
* University of Suwon, Department of Information Security, jaesungpark@suwon.ac.kr, 종신회원
논문번호 : KICS2015-09-279, Received September 1, 2015; Revised September 18, 2015; Accepted September 18, 2015

Ⅰ. 서 론

클라우드 컴퓨팅 서비스는 통신 기술을 매체로 물리적인 IT 자원을 가상화하여 사용자들에게 비용을 받고 제공하는 서비스이다. 서비스 이용자들은 인터넷이 가능한 단말기만 있으면 언제, 어디서나 원하는 하드웨어를 사용할 수 있다. 또한 서비스 이용자들의 요구에 따라 IaaS(Infrastructure as a Service), PaaS(Platform as a Service), SaaS(Software as a Service) 같이 다양한 유형의 서비스를 제공한다. 특히, IaaS는 많은 기업들이 이용하고 있는 서비스로 네트워크 기능, 컴퓨터, 데이터 스토리지 등을 VM 인스턴스(Virtual Machine instance) 단위로 할당하여 관리 제어가 용이하고 확장성이 뛰어나다. 또한 서비스 이용자들은 서비스를 이용한 시간만큼 요금을 지불하기 때문에 비용이 저렴하다[1,2]. 대표적인 IaaS 서비스인 Amazon EC2(Amazon Elastic Compute Cloud)는 서비스 요금 방식에 따라 온 디맨드 인스턴스(on-demand instance), 예약 인스턴스(reserved instance), 스팟 인스턴스(spot instance)와 같은 VM 인스턴스를 제공하고 있다. 그 중 스팟 인스턴스는 온 디맨드 인스턴스와, 예약 인스턴스를 제공하고 남은 가용 자원을 경매 방식으로 서비스 이용자들에게 제공한다. 서비스 이용자들은 온 디맨드나 예약 인스턴스 기본 용량을 유지한 다음, 낙찰 가격(spot price)이 낮을 때 추가 스팟 인스턴스를 이용하여 컴퓨팅 성능을 가속화 한다.

서비스 이용자들은 특정 작업(job)을 처리하기 위하여 클라우드 컴퓨팅 서비스를 이용한다. 서비스 이용자들의 작업은 작업 특성에 따라 작업 완료 시간(job completion time)에 대한 제한 시간(job deadline time)이 존재한다. 특히, 웹 서비스, 비디오 스트리밍 서비스, 모니터링 서비스 등의 실시간 처리를 요구하는 서비스는 작업 처리 시간이 제한 시간을 초과할 경우 지연 시간에 따른 사용자의 QoS(Quality of Service)를 떨어뜨린다[3]. 이러한 서비스 이용자가 스팟 인스턴스를 이용한다면 경매 모델의 특성상 자원의 사용을 보장 받을 수 없다. 그렇기 때문에, 서비스 이용자는 제한 시간 동안 작업을 완료 할 수 있는 자원을 보장 받기 위해 높은 입찰액을 제시할 것이다. 입찰 가격이 높아짐에 따라 낙찰 가격이 상승하고 서비스 이용자의 작업 완료에 대한 총 비용은 증가할 것이다. 따라서 서비스 이용자의 QoS를 보장하고 작업 완료에 대한 총 비용을 최소화하는 입찰 기법이 요구된다.

본 논문에서는 Amazon EC2의 스팟 인스턴스 서비스 기반 가용 자원 경매 모델에서 서비스 이용자의 작업 처리를 위한 총 비용에 대한 기댓값을 최소화하는 입찰 기법을 제안하여 작업 처리에 요구되는 비용을 최소화하고 서비스 이용자의 QoS를 보장한다.

Ⅱ. 관련 환경

2.1 Amazon Web Service

Amazon의 클라우드 컴퓨팅 서비스는 분산 컴퓨팅 시스템에서 발달했다. 이것은 AWS(Amazon Web Service)로 칭해졌으며, 현 아마존 클라우드 컴퓨팅 시스템의 플랫폼을 구성하는 기반이 되었다. 아마존의 클라우드 컴퓨팅 시스템은 대표적으로 가상화된 컴퓨팅 자원을 서비스 하는 Amazon EC2와 대용량 스토리지를 서비스하는 Amazon S3로 나눠지며, 본 논문에서는 서비스 이용자에게 가상 자원을 할당하고 서비스 이용자가 비용을 지불하여 사용하는 거래 모델인 Amazon EC2에 대해서만 다루도록 한다[4].

Amazon EC2 모델은 자신이 자원을 사용한 시간만큼의 서비스 비용을 지불한다. VM 인스턴스는 요금 지불 방식에 따라 온 디맨드 인스턴스, 예약 인스턴스, 스팟 인스턴스가 있다. 첫째로, 온 디맨드 인스턴스를 사용하면 장기 약정 없이, 소비한 컴퓨팅 파워에 대해서만 시간당 종량 과금제로 청구된다. 따라서 하드웨어를 계획, 구매, 유지 관리하는데 수반되는 비용과 복잡성이 사라지고 일반적인 비싼 고정 비용이 훨씬 저렴한 가변 비용으로 전환된다. 둘째로, 예약 인스턴스는 온 디맨드 인스턴스 요금에 할인된 가격으로 장시간 사용을 예약할 수 있다. 마지막으로, 스팟 인스턴스는 서비스 이용자들이 사용할 컴퓨팅 자원과 지불할 수 있는 최대 금액을 입찰하고 입찰 가격이 낙찰 가격을 초과하면 서비스를 제공 받는다. 본 논문에서는 스팟 인스턴스와 온 디맨드 인스턴스 방식을 고려하여 사용자의 QoS를 보장하고 작업 처리에 소요되는 비용을 최소화하는 입찰 기법을 제안한다.

2.2 낙찰 가격(spot price)

일반적으로 경매 모델의 낙찰 가격은 가용 자원과 서비스 이용자들의 입찰 가격을 고려하여 주기적으로 변경된다. 입찰 가격이 낙찰 가격 보다 큰 서비스 이용자들은 자원을 제공 받으며 입찰 가격에 상관없이 낙찰 가격을 기준으로 사용한 시간만큼의 비용을 지불한다. 또한 낙찰 받지 못한 서비스 이용자는 자원을 제공 받지 못하고 비용을 지불하지 않는다. 본 논문에

서는 서비스 제공자의 가용 자원 내에서 입찰 가격이 큰 순서대로 서비스를 제공하고, 제공 받지 못한 서비스 이용자들 중에서 가장 높은 입찰 가격을 낙찰 가격으로 결정한다[5]. 경매 모델의 특성상 높은 가격을 입찰한 서비스 이용자일수록 높은 확률로 서비스를 제공 받을 수 있다. 하지만 많은 서비스 이용자들이 서비스를 제공 받기를 원하고, 그에 따라, 서비스 이용자들이 모두 높은 가격을 입찰한다면 낙찰 가격 또한 상승한다. 본 논문에서는 낙찰 가격을 낮추는 입찰 가격을 결정하여 서비스 이용자의 작업 처리 비용을 최소화하는 기법을 제안한다.

2.3 작업 처리 시간에 따른 QoS

서비스 이용자들은 클라우드 컴퓨팅 서비스를 이용하여 다양한 작업을 처리할 수 있다. 시작과 종료 시간이 자유로운 작업의 경우 작업 완료에 대한 제한 시간이 없기 때문에 QoS에 영향을 받지 않는다. 하지만 웹 서비스, 비디오 스트리밍 서비스, 모니터링 서비스 같은 실시간 처리를 요구하는 작업은 제한 시간 안에 작업을 완료하지 못하는 경우 사용자의 QoS를 떨어뜨린다. 따라서 서비스 이용자는 작업 처리를 위한 스팟 인스턴스를 보장 못한다면 온 디맨드 이용하여 신속하게 작업을 처리하여 서비스 이용자의 QoS를 보장할 것이다. 일반적으로, 온 디맨드 인스턴스의 시간 당 가격이 스팟 인스턴스의 시간 당 가격 보다 비싸다. 온 디맨드 인스턴스만을 이용하여 작업을 처리한다면 서비스 이용자의 QoS는 보장 받을 수 있지만 작업을 처리하는데 많은 비용이 요구된다. 따라서 본 논문에서는 서비스 이용자가 먼저 스팟 인스턴스를 이용하여 보다 저렴한 가격으로 작업을 처리하고 만약 제한 시간 내에 작업을 완료 하지 못한다면 온 디맨드 인스턴스를 이용하여 신속하게 남은 작업을 완료한다[6].

Ⅲ. 제안 모델

본 논문에서는 Amazon EC2의 스팟 인스턴스 서비스 기반 가용 자원 경매 모델에서 서비스 이용자의 작업 처리 비용에 대한 기댓값을 최소화하는 입찰 기법을 제안한다. 서비스 이용자는 스팟 인스턴스를 이용하기 위해 서비스 제공자에게 입찰한다. 서비스 제공자는 서비스 이용자들의 입찰 가격과 가용 자원을 고려하여 낙찰 가격을 결정하고 입찰 가격이 낙찰 가격 보다 큰 서비스 이용자들에게 입찰 가격이 큰 순서로 자원을 할당한다. 만약 서비스 이용자가 작업 완료에 대한 제한 시간 동안의 스팟 인스턴스를 보장 받지 못한다면 온 디맨드 인스턴스를 사용하여 작업을 완료 할 것이다. 표 1은 제안 기법에서 사용되는 파라미터를 보여준다.

서비스 이용자가 제한 시간 안에 작업을 완료하기 위해 요구되는 총 비용에 대한 기댓값 E는 다음과 같이 계산한다.

$$E = Pr_{com}\, T_{com} p_{bid} + (1 - Pr_{com})(T_{com} - T_{succ}) p_{on\_dem}, \quad (1)$$

subject to

$T_{com} > 0,\ T_{com} \leq T_{dead},$

$p_{bid}, p_{on\_dem} > 0,$

$0 \leq Pr_{com} \leq 1.$

식 (1)에서 서비스 이용자는 Pr_{com}의 확률로 작업을 완료하는 시간 T_{com} 동안 자원을 제공받는다. 서비스 이용자가 작업 완료를 위한 자원을 제공 받는다면 작업 완료 시간만큼의 시간 당 낙찰 가격 p_{spot}을 지불해야 한다. 하지만 실제 낙찰 가격은 계산 할 수 없기 때문에 시간 당 입찰 가격 p_{bid}을 지불한다고 가정한다. 반대로, 자원을 제공 받지 못하면 작업의 제한 시간이 초과되어 서비스 이용자는 온 디맨드 인스턴스를 이용하여 작업을 완료해야 한다. 따라서 남은 작업 시간만큼의 시간 당 온 디맨드 인스턴스 비용 p_{on_dem}을 지불해야 한다.

표 1. 제안 모델을 위한 파라미터 정의
Table 1. Parameters for our model.

parameter	description
E	The expected value for total cost to complete a job
p_{bid}	The bidding price per unit of time for a spot instance by user
p_{on_dem}	The price per unit of time for a on-demand instance
p_{max_spot}	The max spot price of historic spot price
p_{spot}	The spot price at current bid
T_{dead}	The deadline time for a job
T_{com}	The completion time for a job
T_{succ}	The time of successful bidding for a job
Pr_{succ}	The probability to successful bid of user
Pr_{com}	The probability to complete a job of user by deadline time

서비스 이용자의 입찰 가격이 실제 낙찰 가격 보다 크면 자원을 제공 받을 수 있다. 서비스 이용자가 자원을 제공 받을 확률은 $\Pr(|p_{bid}| \geq p_{max\_spot})$ 로 나타내고, 서비스 입찰을 위해 자신의 기존 입찰 정보를 기반으로 마르코브 부등식(Markov's Inequality)을 이용하여 간단하게 계산 할 수 있다[7].

$$Pr_{succ} = \Pr(|p_{bid}| \geq p_{spot}) \leq \frac{E[p_{bid}]}{p_{max\_spot}} \quad , \qquad (2)$$

$$\Pr(|p_{bid}| \geq p_{spot}) \fallingdotseq \min\left(\frac{E[p_{bid}]}{p_{max\_spot}}, 1\right) \quad , \qquad (3)$$

subject to

$p_{bid} \geq 0, p_{bid} \leq p_{max\_spot}$.

평균 입찰 가격 $E[p_{bid}]$ 은 서비스 이용자가 지금까지 입찰했던 가격의 평균이다. 일반적으로, 서비스 이용자는 높은 가격을 입찰하기 때문에 p_{spot}은 서비스 이용자가 낙찰 받았던 최대 낙찰 가격 p_{max_spot}으로 가정한다. 또한, $0 \leq Pr_{succ} \leq 1$ 이기 때문에 $E[p_{bid}]$ 은 p_{max_spot} 보다 작다고 가정한다.

식 (3)를 이용하여 서비스 이용자가 제한 시간 T_{dead} 내에 작업을 완료할 확률 Pr_{com}을 식 (4)와 같이 정의 하고 ${}_{T_{dead}}C_{T_{com}}$ 는 조합(combination)을 이용하여 계산한다.

$$Pr_{com} = {}_{T_{dead}}C_{T_{com}} (Pr_{succ})^{T_{com}} (1 - Pr_{succ})^{T_{dead} - T_{com}}. \qquad (4)$$

식 (3)과 식 (4)를 식 (1)에 대입하면 오목(convex)한 부분을 갖는 p_{bid}에 대한 그래프를 얻으며, 총 비용 E가 최소값을 갖는 부분의 접선의 기울기는 0이다. E의 최소값을 계산하기 위해 식 (1)을 미분하여 $dE/dp_{bid} = 0$을 만족하는 되게 하는 p_{bid}를 다음과 같이 계산한다.

$$p_{bid} = \frac{\left\{ T_{dead}p_{on\_dem} + p_{max\_spot} + T_{com}p_{max\_spot} - \sqrt{(T_{dead}p_{on\_dem} + p_{max\_spot} + T_{com}p_{max\_spot})^2 - 4T_{com}p_{on\_dem}p_{max\_spot}(1 + T_{deadl})} \right\}}{2(1 + T_{dead})} \qquad (5)$$

서비스 이용자는 작업 처리를 위한 총 비용에 대한 기댓값을 최소화하는 p_{bid}를 서비스 제공자에게 입찰하여 작업 완료에 대한 총 비용을 최소화할 수 있다.

Ⅳ. 성능 분석

본 논문에서는 Amazon EC2의 스팟 인스턴스 서비스 기반 가용 자원 경매 모델에서 서비스 이용자의 작업 처리를 위한 총 비용의 기댓값을 최소화하는 입찰 기법을 제안하였다. 제안 기법의 성능을 평가하기 위해 사용자의 요구에 대한 실측 데이터(UniLu Gaia, 2014년 5~8월까지의 데이터)를 이용하여 서비스 이용자의 작업을 완료하는데 소요되는 총 비용을 계산하였다[8]. 또한 실제 낙찰 가격 데이터(Amazon EC2, 2015년 5~8월까지의 데이터)를 기반으로 연속균등분포를 이용한 입찰 기법(random)과 성능을 비교하였다. 연속균등분포에 따른 입찰 비용은 [$0.1, $0.2] 사이의 값을 균등한 확률로 설정한다. p_{on_dem}는 Amazon EC2에서 실제 서비스 중인 us-east-m1.large 인스턴스의 시간당 온 디맨드 비용 $0.139로 설정한다[9].

그림 1은 서비스 이용자의 입찰 가격에 따른 작업 완료를 위한 총 비용에 대한 기댓값의 변화를 보여준다. 그림 1에서 서비스 이용자의 입찰 가격이 낮아짐에 따라 스팟 인스턴스를 할당 받을 확률이 낮아지고, 그에 따른 온 디맨드 인스턴스 대한 이용률이 증가하여 기댓값이 상승한다. 또한 입찰 가격이 높아짐에 따라 낙찰 가격이 상승하여 기댓값이 상승한다. 따라서 서비스 이용자는 그림 1에서 작업 완료를 위한 총 비용의 기댓값이 최소값을 갖는 지점의 값을 입찰 가격

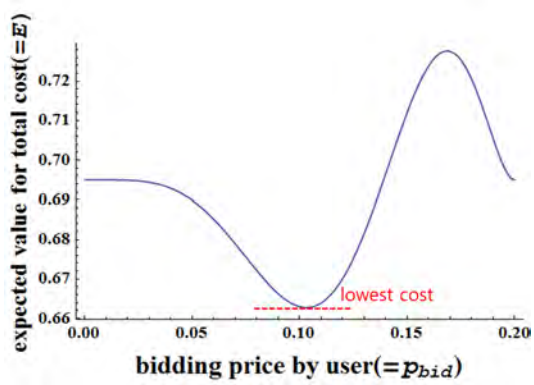

그림 1. 서비스 이용자의 입찰 가격에 따른 작업 완료를 위한 총 비용에 대한 기댓값
Fig. 1. The expected value for total cost to complete job for bidding price by user.

으로 설정한다.

그림 2는 경매 진행에 따른 낙찰 가격의 변화를 보여준다. 제안 기법을 이용하였을 때의 낙찰 가격이 연속균등분포 확률 기법으로 입찰하였을 때 보다 평균 41.6% 낮다. 낙찰 가격이 낮기 때문에 서비스 이용자들은 보다 저렴한 비용으로 자원을 이용할 수 있다.

그림 3은 서비스 이용자 별 작업 당 작업 완료를 위한 실제 비용을 보여 준다. 제한 시간 안에 낙찰 받지 못한 서비스 이용자들은 신속한 작업 처리를 위해 서비스 제공자가 제공하는 온 디맨드 인스턴스를 이용했다. 제안 기법을 이용하여 입찰했을 경우 연속균등분포 확률 기법으로 입찰하였을 때보다 평균 18.8% 낮은 비용으로 작업을 처리한다.

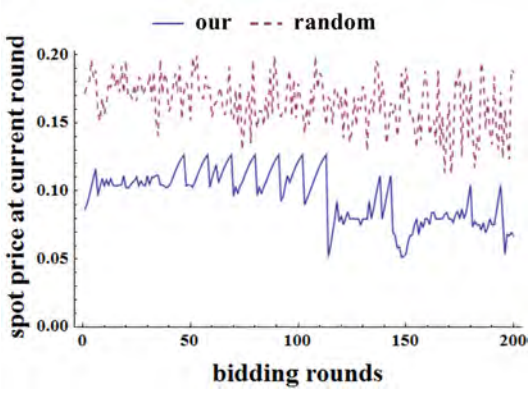

그림 2. 경매 진행에 따른 낙찰 가격
Fig. 2. The spot prices for bidding rounds.

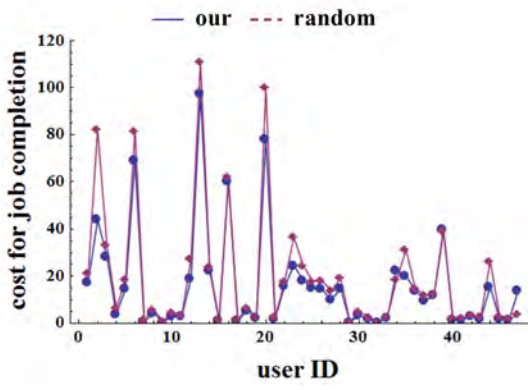

그림 3. 서비스 이용자 별 작업 당 작업 완료를 위한 비용
Fig. 3. The cost for job completion job of users.

V. 결 론

본 논문에서는 Amazon EC2의 스팟 인스턴스 서비스 기반 가용 자원 경매 모델에서 서비스 이용자의 QoS를 보장하고 작업 처리를 위한 총 비용을 최소화하는 입찰 기법을 제안하였다. 일반적으로, 서비스 이용자들은 서비스를 이용하기 위해 보다 높은 가격을 입찰한다. 하지만 입찰 가격이 높아짐에 따라 낙찰 가격이 높아지고 서비스 이용자들의 작업을 처리하는데 더 많은 비용이 요구된다. 따라서 본 논문에서는 서비스 이용자가 제한 시간 내에 작업을 완료할 확률을 기반으로 스팟 인스턴스와 온 디맨드 인스턴스를 사용하여 서비스 이용자의 QoS를 보장하고 작업 완료를 위한 총 비용을 최소화하는 입찰 가격을 계산했다. 또한 제안 기법의 성능 평가를 위하여 실제 낙찰 가격 기반의 입찰 기법과 비교함으로써 제안 모델의 입찰 가격에 따른 낙찰 가격과 작업 처리 비용이 낮다는 것을 증명하였다.

References

[1] H. Kim and H. Kim, "Control algorithm for virtual machine-level fairness in virtualized cloud data center," *J. KICS*, vol. 38C, no. 6, pp. 512-520, Jun. 2013.

[2] M. Kim and M. Park, "Energy-aware virtual machine deployment method for cloud computing," *J. KICS*, vol. 40, no. 1, pp. 61-69, Jan. 2015.

[3] Y. Choi, Y. Lim, and J. Park, "Reinforcement learning approach for resource allocation in cloud computing," *J. KICS*, vol. 40, no. 4, pp. 654-658, Apr. 2015.

[4] S. Lee, T. Kim, and J. Lee, "Resource availability-based multi auction model for cloud service reservation and resource brokering system," *JKSS*, vol. 23, no. 1, pp. 1-10, Mar. 2014.

[5] S. Zaman and D. Grosu, "A combinatorial auction-based mechanism for dynamic VM provisioning and allocation in clouds," *IEEE Trans. Cloud Comput.*, vol. 1, no. 2, pp. 129-141, Oct. 2013.

[6] H. Zhang, B. Li, H. Jiang, F. Liu, A. V. Vasilakos, and J. Liu, "A framework for

truthful online auctions in cloud computing with heterogeneous user demands," in *Proc. IEEE INFOCOM 2013*, pp. 14-19, Turin, Italy, April 2013.

[7] P. Athanasions and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, Prentice Hall, 2002.

[8] D. Feitelson, *Parallel workloads archive: Logs*, Retrieved Aug. 15, 2015, from http://www.cs.huji.ac.il/labs/parallel/workload.

[9] M. Abundo, V. D. Valerio, V. Cardellini, and F. L. Presti, "QoS-aware bidding strategies for VM spot instances: a reinforcement learning approach applied to periodic long running jobs," in *Proc. IFIP/IEEE Int. Symp. Integrated Netw. Management (IM),* pp. 53-61, Ottawa, Canada, May 2015.

최 영 호 (Yeongho Choi)

2013년 2월 : 수원대학교 정보미디어학과 졸업

2013년 9월~현재 : 수원대학교 컴퓨터학과 석사과정

<관심분야> Cloud Computing, Auction System, Resource Allocation

임 유 진 (Yujin Lim)

2000년 2월 : 숙명여자대학교 전산학과 박사

2013년 2월 : Tohoku University, Dept, of Information Sciences 박사

2014년~현재 : 수원대학교 정보미디어학과 교수

<관심분야> Wireless Communication, Cloud Computing

박 재 성 (Jaesung Park)

2001년 2월 : 연세대학교 전기, 전자공학과 박사

2001년~2002년 : University of Minnesota (PostDoc.)

2002년~2005년 : LG전자(선임연구원)

2014년~현재 : 수원대학교 정보보호학과 교수