# 클라우드 컴퓨팅을 위한 적응적 가상 자원 인스턴스 할당 기법

강동기*, 김성환*, 허재원*, 윤찬현*
*한국과학기술원 전기 및 전자 공학과
e-mail:dkkang@kaist.ac.kr, s.h_kim@kaist.ac.kr, jay1@kaist.ac.kr,
chyoun@kaist.ac.kr

# Adaptive Virtualized Resource Instance Allocation for Cloud Computing

Dong-Ki Kang*, Seong-Hwan Kim*, Jae-Won Heo*, Chan-Hyun Youn*
*Dept of Electrical Engineering, KAIST

요 약

구글, 아마존 및 GoGrid 와 같은 클라우드 서비스 제공자(Cloud Service Providers)들은 서비스 사용자의 자원 사용 특성을 고려하여, 다양한 클라우드 서비스 가격 정책을 제공한다. 서비스 가격 정책은 할당되는 가상자원을 크게 온디맨드(On-demand), 예약형(Reserved) 및 스팟(Spot) 인스턴스로 구분하여 가격을 결정한다. 즉 클라우드 서비스 사용자는 자신의 응용을 고려하여 할당 받고자 하는 자원의 예상 사용 시간 및 허용 예산을 기반으로 최적화된 자원 할당을 요청해야 한다. 본 논문에서는 최적화 계산 시간 최소화 및 자원 할당 비용의 절감을 달성하면서도 사용자의 SLA를 보장할 수 있는 적응적 가상 자원 인스턴스 할당 요청 기법을 소개하고자 한다. 본 기법은 서비스 디맨드에 효율적으로 대응하면서도 응용에 따른 적절한 자원 할당을 수행할 수 있다.

## 1. 서론

클라우드 컴퓨팅 서비스를 제공하는 아마존 EC2 및 GoGrid 와 같은 대형 벤더(Vendor) 들은 일반적으로 온디맨드, 예약형 및 스팟 인스턴스의 3 가지 형태로 가격 정책을 달리하여 사용자에게 제공한다[1,2].

첫 번째로 온 디맨드 가상 자원 인스턴스(OVM : Ondemand Virtual Machine instance) 는 처리하고자 하는 응용이 상대적으로 수행 시간이 짧고 일시적으로 발생할 경우에 할당될 수 있다. 최소 단위 시간은 1 시간으로 제공되며 가격이 비싸다.

두 번째로 예약형 가상 자원 인스턴스(RVM : Reserved Virtual Machine instance) 는 처리하고자 하는 응용이 상대적으로 수행 시간이 길고 지속적으로 수행되는 경우에 할당 될 수 있다. 최소 단위 시간은 1 달에서 1 년이상으로 제공되며 단위 시간당 가격이 온 디맨드 가상 자원 인스턴스에 비하여 훨 씬 저렴하다.

세 번째로 스팟 인스턴스(Spot Instance)는 가상 자원 할당 주기에 따라서 가격이 고정되지 않고 동적으로 바뀌게 되며 입찰 가격(bid price)이 사용자와 협의되지 않는다면 할당된 자원이 자동적으로 해제된다. 가격은 앞의 OVM 이나 RVM 에 비하여 더 저렴하지만 신뢰성이 상대적으로 낮다.

기존의 가상 자원 할당에 관한 연구들은 잘 알려진 다양한 최적화 기법을 바탕으로 각 가상 자원 인스턴스의 개수를 결정한다. 그러나 클라우드 서비스 사용자의 서비스 요청 패턴은 고정되지 않고 동적으로 변동하며 그 변동폭의 고저차가 크므로 기존의 예측기반의 최적화 기법이 제대로 적용될 수 없다. 또한 클라우드 서비스 사용자는 다중으로 접속되므로 최적화를 위한 프로세싱 오버헤드가 크게 증가하게 된다. 본 논문에서는 클라우드 서비스 사용자의 서비스 수준 협약(SLA : Service Level Agreement)은 보장하면서도 자원 할당 비용을 효과적으로 줄일 수 있는 적응적 자원 할당 기법(Adaptive Resource Allocation)을 제안하고자 한다. 제안하는 기법을 통하여 다중 클라우드 서비스 사용자 환경을 반영하면서도 큰 프로세싱 오버헤드를 별도로 요구하지 않는다.

## 2. 가상 자원 인스턴스 할당 최적화 기법

기존의 가상 자원 인스턴스 할당 기법은 선형계획법(LP : Linear Programming) 혹은 유전자 알고리즘(GA : Genetic Algorithm) 과 같은 최적화 기법을 이용하여 수행되어졌다. 그러나 이와 같은 최적화 기법은 수행되는데 긴 프로세싱 시간과 추가 자원이 소요되므로 실제 클라우드 환경에 적용하기 적절치 않다. 본 논문에서 제안하는 가상 자원 인스턴스 할당 기법은 별도의 긴 프로세싱 시간이 소요되지 않고 동적으로 변동하는 자원 할당 요청에 대하여 적절하게 반응할 수 있다. 가상 자원 인스턴스

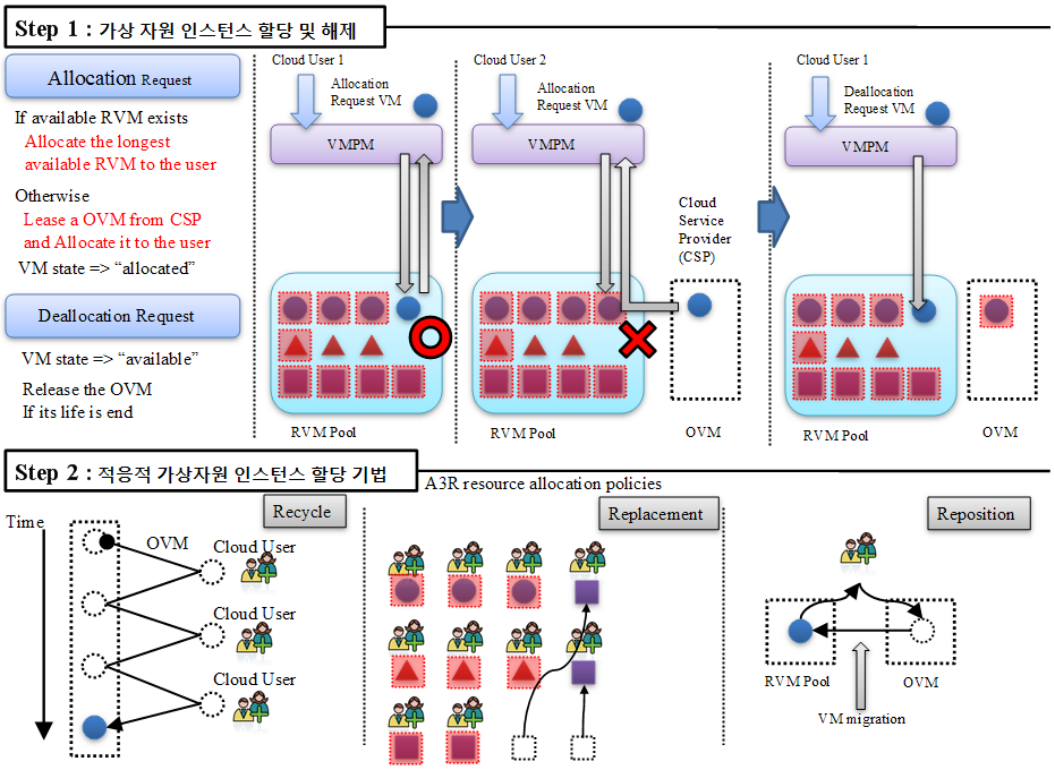


그림 1. 가상 자원 인스턴스 할당 기법

할당 기법을 3 종류로 분류하여 제안하고자 한다.

그림 1 에서는 클라우드 컴퓨팅 환경에서 OVM 및 RVM을 할당하는 과정을 보인다. Step 1 에서는 클라우드 서비스 사용자가 클라우드 브로커에게 자원 할당을 요청하는 경우, 가용한 RVM을 먼저 제공하며 가용한 RVM이 존재하지 않는 경우 OVM을 추가로 생성하여 할당하는 과정을 보이고 있으며 Step 2 에서는 생성된 OVM 및 RVM을 적응적으로 할당하는 과정을 보인다. Step 2에서 소개된 적응적 가상 자원 할당 기법은 다음과 같다.

가. 적응적 재사용(Adaptive Recycle)

클라우드 브로커와 연결되어 하나의 그룹을 이루고 있는 클라우드 서비스 사용자들에게 개별적으로 OVM을 할당하는 것이 아닌 기존에 할당된 OVM을 공유하여 사용하는 기법을 제시한다. 즉 처리하고자 하는 응용의 수행시간이 오버랩되지 않고 또한 동일한 스펙의 자원을 사용하는 클라우드 서비스 사용자들은 별도의 OVM을 할당 받을 필요가 없으므로 추가적인 OVM 할당으로 인한 자원 사용 비용을 절감하면서도 기존이 할당된 가상 자원 인스턴스의 이용률을 최대화 할 수 있다.

나. 적응적 자원 대체(Adaptive Replacement)

클라우드 서비스 사용자들이 요청하는 가상 자원 인스턴스의 성능 스펙은 각자 다르지만, 일반적으로 자신이 요청한 인스턴스의 성능보다 할당된 인스턴스의 성능이 더 높을 경우 일반적으로 이는 문제가 되지 않는다. 이것을 이용하여 스펙 A를 가지는 클라우드 인스턴스가 언더 프로비저닝(Under-provisioning) 이 될 경우, 오버 프로비저닝(Over-provisioning) 되는 스펙 B 나 혹은 스펙 C 를 가지는 가상 자원 인스턴스를 클라우드 서비스 사용자에게 대신 제공할 수 있다. 해당 기법을 가용하기 위해서는 스펙 B 혹은 스펙 C 의 가상 자원 인스턴스의 성능은 스펙 A 보다 우수해야 하며 해당 스펙을 가지는 인스턴스에 대하여 예상되는 서비스 요청량을 예측할 수 있어야 한다.

다. 적응적 재위치 할당(Adaptive Reposition)

일반적으로 RVM 은 OVM 보다 가격이 저렴하므로 되도록 OVM의 개수는 최소화하고 가능한한 RVM에서 사용자의 응용을 처리하도록 하는 것이 비용을 효과적으로 절감할 수 있다. 만약 OVM에서 응용이 처리되는 도중 OVM 의 스펙과 동일한 혹은 그 이상의 스펙을 가지는 가용 RVM 이 발견되는 경우, 해당 OVM에서 처리하고 있는 응용을 RVM 에 이주(Migration) 하여 수행할 수 있다. 본 기법이 적용되기 위해서는 해당 RVM 에 대한 서비스 요청량이 정확히 예측되어져야 하며, OVM에서 RVM으로 작업을 이주하는 비용 오버헤드가 OVM 의 추가 할당 비용보다 작아야 한다.

## 3. 결 론

본 논문에서는 온디맨드 및 예약형 가상 자원 인스턴스를 기반으로 적응적 자원 할당 기법에 대한 대략적인 구조를 소개하였다. 할당시간이 상대적으로 길고 가격은 저렴한 예약형 가상 자원 인스턴스를 사용자의 응용 패턴에

따라 적절히 할당할 수 있다면 사용자의 SLA 는 보장하면서도 비용은 적절히 절감할 수 있음을 확인할 수 있다. 추후 연구에서는 본 연구에서 소개한 기법을 이론적으로 체계화하면서 이에 기반한 다양한 시뮬레이션 및 실제 시스템 실험을 통하여 성능 및 비용을 측정하고자 한다.

## Acknowledgment

이 논문은 2012 년도 정부(교육과학기술부)의 재원으로 한국연구재단-클라우드 Collaboration 기술 사업과 BK21 사업의 지원을 받아 수행된 연구임(No. 2012-0006425)

## 참고문헌

[1] Amazon EC2 (2013), http://aws.amazon.com/ec2/

[2] GoGrid (2013), http://www.gogrid.com/

[3] D. Kang, S. Kim, Y. Ren, B. Kim, W. Kim, Y. Kim, C. Youn, and C. Jeong, "Enhancing a Strategy of Virtualized Resource Assignment in Adaptive Resource Cloud Framework," Proc. ACM Int'l Conf on Ubiquitous Information Management and Communication. (ICUIMC), 2013.

[4] J. Simarro, R. Vozmediano, R. Montero, and I. Llorente, "Dynamic Placement of Virtual Machines for Cost Optimization in Multi-CLoud Environments," Proc. IEEE Int'l Conf on High Performance Computing and Simulation. (HPCS), 2011.

[5] S. Chaisiri, B. Lee, and D. Niyato, "Optimal Virtual Machine Placement across Multiple Cloud Providers," Proc. IEEE Asia-Pacific Services Computing Conf. (APSCC), 2009