클라우드 환경에서 Seamless한 서비스 제공을 위한 빠른 확장성과 성능을 제공하는 적응형 자원 프로비저닝 기법

유승환*, 홍요훈**, 김성천*
*서강대학교 컴퓨터공학과
**(주) 세창인스트루먼트
e-mail : ossforever@sogang.ac.kr

Adaptive Resource Provisioning Method that provide Rapid Scalability and High Performance for Seamless I/O Intensive Application Service in Cloud Environment

Seung Hwan Yoo*, Hong Yo Hoon**, Sung Chun Kim*
*Dept of Computer Science & Engineering, Sogang University
**Dept of IT research institute , Sechang Instruments

요 약

최근 클라우드 컴퓨팅에 관련된 기술이 각광을 받으면서, 기존에 인터넷을 통해 제공되던 다양한 서비스들이 클라우드 컴퓨팅 플랫폼 환경으로 이동하고 있다. 이를 통해 사용자들에게는 좀 더 편리하고 유연한 서비스를 제공하고 서비스 제공자들에게는 기존 관리 비용의 절감을 할 수 있게 되었다. 하지만 현재 폭발적인 수요의 증가로 인해, 기존의 자원 활용도의 극대화를 목적으로 고안된 자원 분배 기법들에 대한 여러 한계점이 나타나게 되었다. 본 논문에서는 이러한 문제점들을 해결하고자 클라우드 컴퓨팅 환경에서 Qos를 기반으로 사용자나 서비스 제공자의 비용적인 측면을 고려한 자원분배를 통해 서비스를 제공시 사용자 요구를 만족시키고 동시에 서비스 공급자에게는 비용 효율적인 프로비저닝(Provisioning)기법을 제안하고자 한다. 실험 결과 기존의 자원 활용도에 중점을 둔 기법보다 사용자 요청에 대한 응답 속도가 8.35% 향상되었으며, 컴퓨터 자원 유지 관련 비용면에서도 기존 대비 11.31% 절감 효과를 가져오는 것을 확인 할 수 있었다.

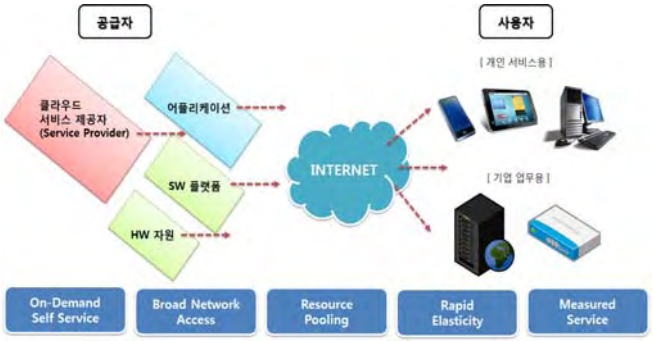
1. 서 론

클라우드 컴퓨팅(Cloud Computing)은 인터넷 기술을 활용하여 IT 자원을 서비스로 제공하는 컴퓨팅 패러다임으로 IT 자원을 필요한 만큼 빌려서 사용하고 서비스 사용 부하(workload)에 따라 실시간으로 확장성을 지원받고 그에 따른 비용을 지불하는 특성을 가진다. 기술적으로는 가상화(Virtualization), 분산 컴퓨팅(Distributed computing) 그리고 자동화(Automation) 기술을 적용하여 보다 완전한 서비스형 IT 제공 모델(Delivery Model)로서 각광받고 있다.

클라우드 컴퓨팅은 앞서 언급한 대로 컴퓨팅 패러다임의 변화를 이끌어내는 미래 新기술이다. 이를 통해 기존의 컴퓨팅 환경보다 비용(Cost)를 절감할 수 있으며, 사용자의 편익을 증진시킬 수 있다. 때문에 이와 관련된 연구가 다양하게 진행되어 왔다. 특히, 클라우드 컴퓨팅 플랫폼에서 어플리케이션(Application) 기반에서 자원 제공자(Service Provider)가 최소한의 비용(Cost)으로 사용자의 요구사항(Requirement)을 만족시키기 위한 고성능(High-Performance)과 빠른 확장성(Rapid-Scalability)을 가진 최적화된 자원 할당을 하는 적응형 프로비저닝 기법에 대한 연구를 제안하고자 한다.[1]

즉, 사용자에게 제공되는 '서비스의 수준(Quality of Service)을 보장'하면서, 전체 클라우드 시스템의 '운영 비용(Cost)을 절감'하는 것을 핵심 목적으로 서비스를 제공하기 위해 효율적인 자원분배 기법에 대해 언급하고자 한다.

본 논문의 구성은 다음과 같다. 1장에 클라우드 컴퓨팅에 대한 서론에 이어 2장에서 I/O Intensive한 서비스 제공시 사용자와 공급자의 비용을 고려한(Cost-Efficient) 시스템 모델링에 기반한 자원 분배를 수행하는 적응형 프로비져닝 기법에 대해 언급하고 3장에서는 제안된 기법을 실제 클라우드 컴퓨팅 플램폼에 적용한 실험 결과를 설명할 것이다. 마지막 4장에서는 이를 통한 결론과 향후 연구 방향에 대해서 간략하게 정리할 것이다.

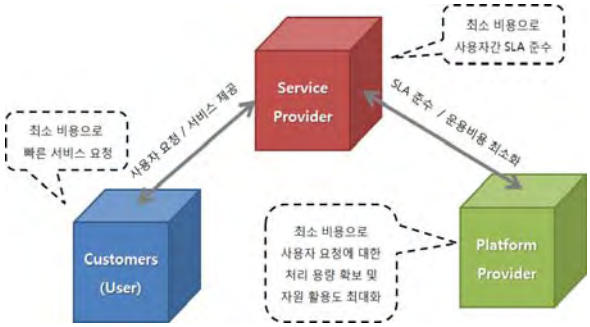


[그림 1] 클라우드 컴퓨팅 플랫폼 개념도

2. 제안 기법

제안하는 기법의 최종 목표를 이루기 위한 가장 이슈가 될 것으로는 사용자들이 요청하는 작업을 수월하게 처리할 수 있는 성능에 대한 보장과 서비스 제공자(Service Provider:SP)나 플랫폼 제공자(Platform Provider:PP)가 최소한의 비용(Cost)를 들여 작업의 요청의 크기의 변동성에 대한 적절한 대처를 위한 확장성을 가지는 자원분배 기법 연구이다. 이를 위해 본 논문에서는 수학적 방법론(Methodology)을 도입하여 클라우드 시스템 성능 모델링(Modeling)을 구성한 후 사용자들의 요구사항과 서비스 제공자들의 요구 조건간의 상관관계를 파악하여 개선된 프로비저닝 기법을 제안하고자 한다.

[그림 2]는 서비스 제공자 및 플랫폼 제공자, 사용자 간의 각각 클라우드 컴퓨팅 서비스를 위해 필요한 요구사항에 대한 설명이다.



[그림 2] SP, PP, User 간 상관관계도

이를 위하여 우선 사용자들의 요청을 처리하면서 발생되는 비선형적 특성을 가진 클라우드 컴퓨팅 작업 부하의 유형을 온 디맨드(On Demand) 호스팅 모델에 기반하여 <표 1>에서와 같이 시스템 변수(Parameter)로 정의하여 Seamless한 작업이 중요한 어플리케이션을 지원하는 가상화 자원 동적 할당 최적화를 위해 우선적으로 Memmory Streaming 기법을 적용한 시스템 모델링을 구성하고자 한다. 또한 확장성과 성능을 보장하기 위해서 Scale관련 변수(λ)를 정의하였으며, 기반이 되는 Physical Server Pool은 다양한(Hybrid) 성능의 Server로 구성되어있다고 가정한다.

<표 5> 시스템 변수와 사용자 요구관련 요소에 대한 변수 정의

Parameter	Description
K	Number of applications
M	Number of VMs
N	Number of physical servers
L	Number of VMs in overloaded state
DM×N	VM allocation Matrix
i	ithVM
j	jth Server
Rc	Total CPU capacity of a server
Rm	Total memory capacity of a server
RI/O	Total I/O capacity of a server
λc	Scale factor of CPU
λm	Scale factor of memory
λio	Scale factor of network I/O
Cmig	VM migration cost
memvm	VM RAM size
$\overrightarrow{RCV}$	Residual capacity vector
$\overrightarrow{TCV}$	Total capacity vector
$\overrightarrow{RRV}$	Resource requirement vector
$\overrightarrow{UCV}$	Utilized capacity vectorical servers

추가적으로, 서비스 제공자(SP)들과 플랫폼 제공자(PP)들의 클라우드 플랫폼 유지비용 절감을 위해 시스템 사용량에 대한 최적화를 고려하여야 한다. 이를 위해 식 (1), (2), (3)를 다음과 같이 정의하여 이를 만족시키는 최적의 해를 구하고자 한다. 수식 (1)은 전체적인 자원 사용의 균형을 고려하는 수식이며, (2)는 자원 활용도를 최대화를 알아보기 위해 정의된 수식이다. 마지막으로 수식 (3)은 전체적인 시스템의 재설정 비용을 최소화를 보여주는 수식이다.

$$Max \ \min_{j\in 1\cdots N} \frac{\overrightarrow{UCV_j} \bullet \overrightarrow{TCV_j}}{|\overrightarrow{UCV_j}| \bullet |\overrightarrow{TCV_j}|} \quad (1)$$

$$Max \ \frac{1}{S_r} \quad (2)$$

$$Min \sum_{i=1}^{L} C_{mig,i} \quad (3)$$

위의 시스템 성능 모델링에 대한 구성을 통해 최적의 시스템 변수 설정 값을 도출한 후, 시스템 예측 모델링을 통해 보다 정확한 시스템 설정 값을 얻고자 한다. 기본적으로 사용자 요청 작업의 변동성과 불확실성을 고려하여 본 제안 논문의 예측 모델링은 수식(4)를 통해 구성하고자 한다. E(t)는 시간 t에서 예측되는 작업량이고 O(t)는 시간 t에 측정되는 작업량이다. 또한, 가중치 α의 조정을 통해 측정되는 작업량에 따라 예측되는 작업량에 대한 적응 속도를 조절하고자 한다.

$$E(t) = \alpha \cdot E(t-1) + (1-\alpha) \cdot O(t) \quad (0 \le \alpha \le 1) \qquad (4)$$

본 논문에서 제안하는 기법은 자동 프로비저닝(Aotomous Provisioning)을 통해 필요한 순간에 동적으로 자원 할당을 수행해 서비스를 생성 및 제공하고자 한다. 이를 위해서는 실시간으로 자원의 상태를 관리(Resource Management)할 수 있는 알고리즘을 최종 목표로 한다. 추가적으로, 고려할 점으로 허용 가능한 범위의 설정은 사용자와 서비스 제공자 간에 합의한 Qos(Quality of Service)를 기준으로 정한다. 예를 들어, 활용빈도가 높은 서비스에 대한 순서를 두어 물리적으로 한정된 네트워크 자원이나 시스템 대역폭에 대한 차등을 두고자 한다.

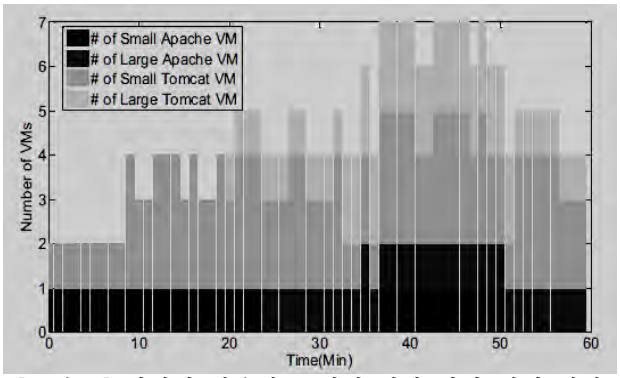
3. 실험 결과

실험은 클라우드 컴퓨팅 환경(ex> e-Bay)에서 발생하는 사용자 요청들을 발생시키는 RUBiS 벤치마크 도구를 통해서 수행하였다. 실험 가정은 웹을 통해 RTSP(프로토콜)을 활용한 미디어 컨텐츠를 사용자가 사용한다는 가정하에 수행하였다. 미디어 영상 크기는 50MB에서 2GB로 하였으며, 네트워크는 유선네트워크 환경만 고려하였다. <표 1>은 실험 환경에 대한 세부 설정에 대한 사항들을 정리한 것이다.

<표 2> RUBiS Benchmark 세부 설정

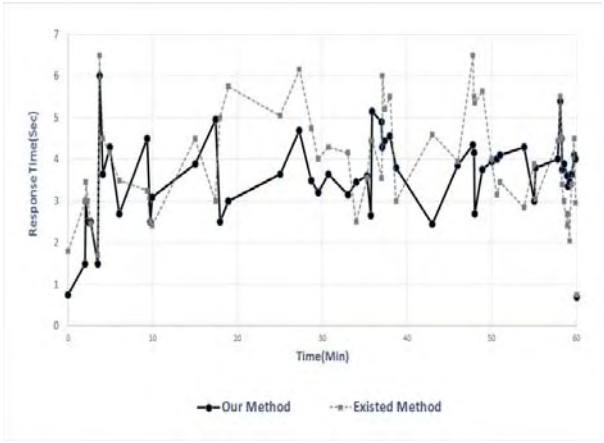
< Simulation Parameter >	
Web Server 환경 : Apache 2.0.55	
Application Server 환경 : Tomcat / Apache	
Database Server 환경 : MySQL	
Hardware 환경	4Core CPU(2.3Ghz) 128GB DDR3 RAM 4TB HDD 1Gbps LAN x2
Respond Time(Qos) : 5.0 sec	

[그림 3]는 1시간 동안 다양한 사용자 요청이 들어올 때, 그에 맞게 다양한 크기의 가상머신들이 할당되는 것을 보여주는 그래프이다. 실험결과 이를 통해 사용자 Qos를 만족시키는 비율이 91.3%에 근접하게 나타났다. 여기서 중요한 점은 Qos를 불만족시키는 자원할당의 경우이다. 이러한 문제를 극복하기 위해서 본 논문에서 제안하는 기법을 적용하여, 기존의 자원의 활용도를 극대화하고자는 기법과 실험을 통해 비교를 하였다.



[그림 3] 다양한 사용자 요청에 의한 가상 머신 할당

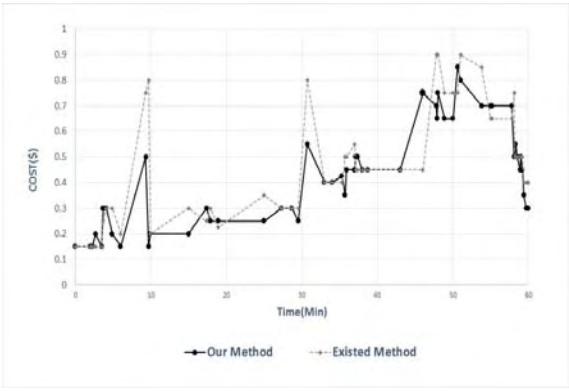
[그림 4]에서 볼 수 있듯이, 제안 기법을 통한 클라우드 컴퓨팅 자원 할당을 할 경우 보다 안정성 있는 응답시간을 얻을수 있음을 알 수 있다. 또한, 기존 자원사용량을 최대화하는 기존 기법에 비해 전체적인 평균 응답시간 역시 8.35% 빨라졌음을 알 수 있다.



[그림 4] 사용자 요청에 의한 응답시간 비교

[그림 5]는 사용자 요청에 의한 가상 머신 할당을 하기 위한 비용 발생에 대한 실험 결과이다. 그래프에서 볼 수 있듯이 전체적으로 총 발생 비용의 경우, 11.31%정도의 절감효과를 가져옴을 알 수 있었다. 다만, 새로운 요청 값이 현격하게 발생할 때 기존 기법보다 순간적으로 비용이 더 발생하는 사례가 발견되었다. 또한, 응답시간 역시 전

체적으로 안정적이지 못한 구간이 발생되는 것을 확인할 수 있었다.



[그림 5] 사용자 요청에 의한 발생비용 비교

3. 결 론

기존 클라우드 컴퓨팅 플랫폼 환경에서 자원 할당하는 프로비져닝 기법들은 자원의 활용도(Utilization)을 극대화 하는데에 중점을 두고 연구되어져 왔다. 하지만 최근의 클라우드 컴퓨팅을 활용한 서비스들은 비용적인 측면을 고려하여 사용자들에게 신속한 서비스를 제공하야 할 필요성이 점차 커지고 있다. 때문에 본 논문에서는 사용자들에게 객관적인 수치화 된 신속한 서비스를 제공하기 위해 특히 미디어 컨텐츠 관련된 서비스에 대한 Qos를 고려한 적응형 프로비져닝 기법을 제안하고자 하였다. 추후 연구로는 사용자 요청이 종류가 좀 더 다양하지고 비정형적인 특징을 가지게 될 때, 발생하는 문제점들을 해결하기 위하여 비용적인 측면을 장기적인 관점과 단기적인 관점에서 Qos를 보장하면서 동시에 비용 효율적인 클라우드 자원 분배 기법에 대해서 연구를 진행하고자 한다.

ACKNOWLEDGMENTS

"이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원(2012R1A1A2009558)과 부분적으로 2013년도 정부(중소기업청)의 재원으로 혁신기술개발사업의 지원(S20557016)을 받아 수행된 연구임."

참고문헌

[1] "클라우드 컴퓨팅 이슈 및 현황", 한국산업기술평가관리원, 2011.12.

[2] V. Almeida, M. Arlitt, and J. Rolia. Analyzing, "a Web-Based System's Performance Measures at Multiple Time Scales". SIGMETRICS Performance Evaluation Review, 30(2):3-9, 2002.

[3] E. Kalyvianaki, T. Charalambous, and S. Hand, "Self-adaptive and self-configured CPU resource provisioning for virtualized servers using kalman filters", Proceedings of the 6th international conference on Autonomic computing, Barcelona, Spain, June 15-19, 2009. pp. 117-126

[4] W. E. Walsh, G. Tesauro, and J. O. Kephart, "Utility functions in autonomic systems", Proceedings of the First IEEE International Conference on Autonomic Computing, New York, NY, USA, May 17-18, 2004.

[5] B. Urgaonkar, P. Shenoy, and A. Chandra, "Agile dynamic allocation of multi-tier Internet application", ACM Trans. on Autonomous and Adaptive Systems, 2008, vol. 3, pp. 1-39.

[6] D. Ardagna, M. Trubian, and L. Zhang, "SLA based profit optimization in multi-tier systems", Proceedings of the 4th IEEE International Symposium on Network Computing and Applications, Cambridge, Massachusetts, USA, July 27-29, 2005.

[7] J. Zhang, M. Yousif, and R. Carpenter, et al, "Application resource demand phase analysis and prediction in support of dynamic resource provisioning", Proceedings of the 4th International Conference on Autonomic Computing, 2007, pp. 12-12.

[8] Herbst, N., Huber, N., Kounev, S., & Amrehn, E. (2013). "Self-adaptive workload classification and forecasting for proactive resource provisioning." ICPE 13, Proceedings of 4th ACM/SPEC International Conference on Performance Engineering, Pages 187-198, ACM, New York, USA, 2013.

[9] Huu, T. T., & Tham, C.-K. (2013). "An Auction-Based Resource Allocation Model for Green Cloud Computing." 2013 IEEE International Conference on Cloud Engineering (IC2E), 269.278.

[10] Yanggratoke, R. (2011). "Gossip-based resource allocation for green computing in large clouds." 7th International Conference on Network and Service Management (CNSM),pp1-9, 2011, Paris.

[11] Moreno, I., & Xu, J. (2011). "Customer-aware resource overallocation to improve energy efficiency in realtime Cloud Computing data centers," SOCA' 11 Proceedings of the 2011 IEEE International Conference on Service-Oriented Computing and Applications, pages 1-8, IEEE Computer Society Washington, DC, USA, 2011.