

데스크탑 클라우드 컴퓨팅 환경에서의 로드 밸런싱 기법에 관한 연구

신은환, 김지홍, 엄영익
성균관대학교 정보통신공학부
{comsky, jjilong, yieom}@ece.skku.ac.kr

A Study on Method for Load Balancing in Desktop Cloud Computing Environment

Eun Hwan Shin, Jee Hong Kim, Young Ik Eom
School of Information and Communication Engineering
Sungkyunkwan University

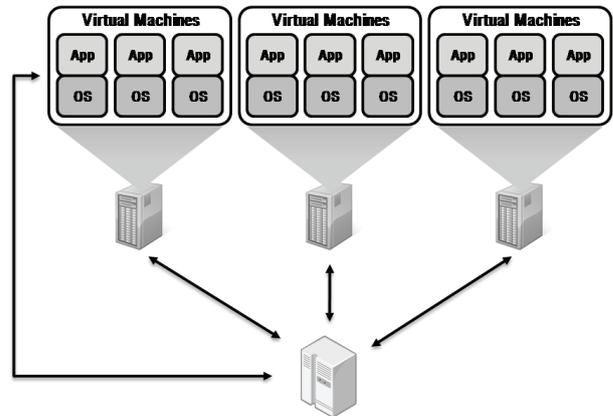
요 약

클라우드 컴퓨팅 기술의 발전으로 서버 환경에 한하여 적용되었던 가상화 기술이 데스크탑 환경으로 빠르게 확산되어가고 있다. 클라우드 컴퓨팅 환경은 가상화 기술을 이용하여 하나의 물리적 자원위에서 여러 개의 가상 머신을 운용하기 때문에 효과적인 로드 밸런싱(load balancing) 기법을 필요로 한다. 하지만 기존 서버 가상화 환경에서의 로드 밸런싱 정책은 프로세서와 네트워크 사용률을 주된 로드 측정의 기준으로 삼았다. 그에 반해 데스크탑 클라우드 컴퓨팅 환경 하에서의 데스크탑 가상 머신은 사용자 유형에 따라 다양한 종류의 작업 요청을 처리해야 하기 때문에 처리 중인 작업의 특성을 고려한 보다 유연한 로드 밸런싱 정책을 필요로 한다. 본 연구에서는 데스크탑 클라우드 컴퓨팅 환경에 적합한 로드 밸런싱 정책을 제시하고, 이를 바탕으로 제안하는 기법이 가지는 효율성을 기존 방식과 비교 및 분석 하였다.

1. 서 론

클라우드 컴퓨팅(cloud computing) 기술의 발전으로 서버 환경에 한정되어 적용되었던 가상화 기술이 데스크탑 환경으로 빠르게 확산되어가고 있다. 가상화 기술은 그림 1과 같이 하나의 물리적인 장치를 복수개의 가상 머신으로 나누어 자원을 공유함으로써 자원 활용에 대한 효율성을 높여준다. 또한 자원 공유에 따른 효과로 물리적 장치의 수가 줄어들어 시스템 구입과 유지보수에 드는 비용이 절감된다는 장점을 지니고 있다. 하지만 가상화 기술은 자원 관리의 복잡도를 증가시켜 물리적 장치의 로드 발생을 야기한다는 문제점이 있다. 이를 해결하기 위해 가상화 환경에서 발생하는 로드의 특성을 분석하기 위한 연구가 활발히 이루어지고 있다. 이러한 가상화 기술을 웹 서버, 데이터베이스 서버, 파일 서버 등과 같이 네트워크 서비스를 제공하는 환경에 적용한 것이 서버 가상화이다. 서버에 적용된 가상화 기술은 하나의 물리적 서버를 복수개의 가상 서버로 나누어 과도하게 집중된 서비스를 분산시키고 서버의 자원 활용률을 높여주는 것을 그 목적으로 한다.

이에 반해 데스크탑 클라우드 컴퓨팅 환경에서 사용자는 개별적으로 제공되는 가상 머신을 통해 개인용 컴퓨터와 동일한 환경을 온라인상으로 제공받는다. 이와 같은 데스크탑 클라우드 컴퓨팅 환경에서는 개인 사용자를 위한 애플리케이션 서비스가 중심이 되며, 요구되는 작업의 특성에 맞추어 시스템상의 자원을 동적으로 재배치하여 서비스하는 것이 가능하다.



(그림 1) 가상화 기반 구조의 예

앞서 살펴본 바와 같이 서버 가상화 환경에서는 여러 개의 가상 서버를 이용하여 물리적 서버에 집중된 로드를 분산시킨다. 일반적으로 서버 가상화 환경에 적용되어온 로드 밸런싱 정책들은 물리적 서버상의 CPU 점유율, 메모리 사용량 및 디스크 I/O 횟수 등을 기준으로 로드를 측정한다. 하지만 일반적인 네트워크 서비스를 제공하는 서버와는 달리, 데스크탑 클라우드 컴퓨팅 환경에서는 가상화 기술을 이용하여 개별 사용자에게 하나의 완전한 개인용 데스크탑 환경을 제공한다. 따라서 데스크탑 클라우드 컴퓨팅 환경에서는 각 가상 머신 내에서 수행 중인 작업의 특성을 고려하여, 기존의 방식을 개선한 새로운 로드

밸런싱 정책이 적용될 필요가 있다[1][2].

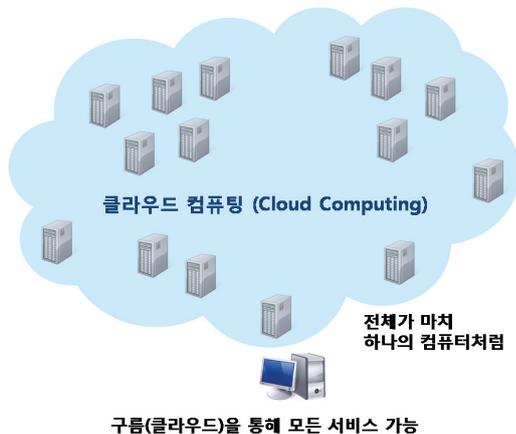
이에 본 논문에서는 데스크탑 클라우드 컴퓨팅 환경에 적합한 로드 밸런싱 기법을 제안한다. 본 논문의 2장에서는 관련 연구 및 기술 현황에 대해서 다룰 것이며, 3장에서 제안 알고리즘에 대한 소개와 더불어 기존 기법과의 비교·분석을 실시한다. 그리고 4장에서 본 논문을 마무리 한다.

2. 관련 연구 및 기술 현황

2.1 클라우드 컴퓨팅(cloud computing)

클라우드 컴퓨팅이란 인터넷 기술을 활용하여 ‘가상화된 IT 자원을 서비스’로 제공하는 컴퓨팅으로 사용자는 그림 2와 같이 클라우드 내에 존재하는 IT 자원(소프트웨어, 스토리지, 서버, 네트워크)을 필요한 만큼 빌려서 사용하고, 그에 따른 비용을 지불하는 방식의 컴퓨팅을 말한다. 클라우드 컴퓨팅 시스템 구축에 의한 기대 효과는 다음과 같다[3].

- 자원 활용률 증가 : 가상화는 물리적 자원들과 자원 풀에 대한 동적인 공유를 가능케 해줌으로써 더 높은 자원의 활용률을 얻을 수 있다.
- 관리 비용 감소 : 가상화는 관리되어야 하는 물리적 자원들을 보여줌으로써 관리 인력의 생산성을 향상시킬 수 있다. 또한 물리적 자원들의 복잡성을 숨겨 주고, 자동화·정보화·중앙화를 통해 공통된 관리 작업을 단순화시키며, 워크로드 관리의 자동화를 가능하게 해준다.
- 사용의 유연성 : 가상화는 빠르게 변화하는 비즈니스 요구 사항들을 만족시키기 위해 자원들이 동적으로 재구성되고 활용될 수 있도록 해준다.
- 보안성 향상 : 가상화는 단순한 공유 메커니즘에서는 불가능한 분리와 격리를 가능하게 해주어 데이터와 서비스에 대해서 통제되고 안전한 접근을 제공한다.
- 확장성 증가 : 리소스 분할화 및 단일화는 가상화된 자원이 개별 물리적 자원보다 더 작아지거나 더 커질 수 있도록 해 준다. 이를 통해 물리적 자원의 구성 변경이 없어도 필요한 만큼의 적절한 확장성을 얻을 수 있다.



(그림 2) 클라우드 컴퓨팅 개념도

2.2 가상화(virtualization)

가상화란 컴퓨팅 리소스의 추상화를 통해 다수의 운영체제 또는 애플리케이션 이미지가 하나의 물리적 서버를 공유하도록 함으로써 총소유비용을 절감하고 관리성을 개선할 수 있게 하는 프로세스를 말하며, 하이퍼바이저(hypervisor), 운영체제, 하드웨어의 관계에 따라 다음과 같이 분류 할 수 있다[4].

• 전가상화(full-virtualization)

전가상화는 하드웨어를 완전히 가상화하는 방식이다. 운영체제는 하드웨어에 직접 접근하는 것으로 인식 하지만 해당 하드웨어는 하이퍼바이저에 의해 에뮬레이션(emulation) 된 것이며, 운영체제 입장에서는 하드웨어를 자신이 소유한 것으로 인식한다. 하드웨어를 가상화하기 때문에 게스트 운영체제에 수정을 가할 필요가 없어, 다양한 종류의 운영 체제를 이용할 수 있다는 장점이 있다. 하지만 전가상화를 위해서는 물리적 형태의 가상화 기능이 필수적이다. Intel의 VT나 AMD의 Pacifica 같은 확장 명령어 셋이 이러한 기능을 지원한다.

• 반가상화(para-virtualization)

반가상화는 전가상화와는 달리 하드웨어를 완전히 가상화하지 않는다. 운영체제는 하드웨어에 하이퍼바이저가 제공하는 API를 통해서만 접근할 수 있어, 높은 성능의 유지가 가능하다는 장점이 있다. 하지만 운영체제 입장에서는 하드웨어를 자신이 소유하지 않은데 대한 처리를 해 줄 필요가 있기 때문에 반가상화의 실현을 위해서는 게스트 운영체제 커널의 일부분을 수정해야만하며, 이로 인해 오픈 소스 운영체제에 한해서만 적용가능하다는 단점이 있다.

2.3 로드 밸런싱(load balancing)

병렬적으로 운용되고 있는 여러 작업 사이의 로드가 최대한 균등 해지도록 처리를 분산하여 할당하는 것을 말한다. 마이크로프로세서에 균등히 처리를 부여하거나, 접속 요구를 네트워크상의 여유가 있는 서버로 회송하는 등 여러 분야에서 폭넓게 이용되고 있는 개념이다. 로드 밸런싱을 효율적으로 수행하기 위해서는 각 작업의 로드를 지속적으로 측정할 필요가 있다. 하지만 지나치게 정확한 수치를 적용시키려 할 경우, 로드 밸런싱을 위한 제어 자체가 큰 로드를 발생 시킨다. 따라서 실제로 이용되고 있는 로드 밸런싱 기법들은 정확도와 로드의 균형을 취하도록 설계되어 있다[5].

기존 서버 가상화 기술에서 로드의 특성을 분석하여 공식화한 사례 중 M. Harchol-Balter와 A. B. Downey의 연구는 서버상의 각 프로세스에 의해 요구되는 생명 주기를 고려한 로드 밸런싱 정책을 제안했다. 이 연구에서는 각 프로세스가 소모하는 CPU 시간을 정형화된 공식에 대입하여 비교적 정확한 생명 주기를 구할 수 있으며, 이를

토대로 로드 밸런싱을 하게 된다. 연구에 의하면 작업의 생명 주기 분포를 조사한 결과 대부분의 작업은 짧은 수행 시간을 가지며, 긴 수행 시간을 요구하는 작업들의 개수는 전체 작업 개수에 비해 매우 적다는 것을 알 수 있다. 따라서 장시간을 요구하는 작업을 상대적으로 로드가 적은 서버에 분배함으로써, 전체 시스템의 성능을 높일 수 있다[6].

3. 제안 알고리즘 소개

3.1 로드 밸런싱 기준 요소 비교

관련 연구에서 살펴보았듯이 이미 서버 가상화 환경에서의 로드 밸런싱을 위한 다양한 기법들이 연구 및 개발되어 있다. 이러한 일반적인 로드 밸런싱 기법들은 주로 서버 상에서 제공되는 네트워크 서비스에 대한 요청이나 해당 작업에 대한 소요 시간을 기준으로 하며, 일반적으로 고려되는 요소들은 표 1과 같다[7].

<표 1> 서버 가상화 환경에서의 로드 밸런싱 기준 요소

- CPU 점유율, 메모리 사용량
- 작업 요청에 대한 서버의 응답시간
- 서버에 요청되어진 작업의 개수

위와 같이 주로 물리적 장치가 로드 밸런싱의 기준이 되며 모든 요소들의 합이 최소인 서버에 로드를 우선적으로 분배하게 된다. 하지만 데스크탑 클라우드 컴퓨팅 환경에서는 가상 머신 내에서 처리되는 작업에 대해 보다 많은 특성들을 고려할 필요가 있으며, 제안되는 기준 요소는 표 2와 같다.

<표 2> 제안 알고리즘의 로드 밸런싱 기준 요소

- CPU 점유율, 메모리 사용량
- I/O(Disk, VGA, Network) 횟수
- system call 횟수
- context switching 횟수
- 가상 머신의 대기·응답 시간
- 가상 머신의 자원 요청 횟수

표 2에서는 system call 횟수, context switching 횟수, 가상 머신의 대기·응답 시간, 가상 머신의 자원 요청 횟수가 새로운 기준 요소로 추가 되었다. 이에 더하여 각 가상 머신 내의 작업 특성을 보다 세밀하게 측정하기 위해 처리 중인 작업 종류에 따른 로드 특성을 표 3에 분류해 보았다.

<표 3> 가상 머신 내의 작업 종류에 따른 로드 특성 분류

작업 종류	주요 발생 로드
Web Browsing	Network I/O, Disk I/O
Documentation	Memory, Disk I/O
Game	CPU, Memory, VGA I/O, Network I/O
Multimedia	CPU, Memory, Network I/O

3.2 제안 알고리즘 계산 과정

제안 알고리즘에서는 추가적으로 표 2에서 제안된 각 요소가 가상 머신이 발생시키는 로드량에 영향을 미치는 비율을 고려하였으며, 이는 클라우드 컴퓨팅 환경 하에서 시스템 자원이 각 가상 머신에 동적으로 재배치 가능하다는 점을 감안한 것이다. 즉, 처리하고자 하는 작업량에 따라 할당 받는 자원의 양이 달라지는 클라우드 컴퓨팅 서비스의 특성을 반영하였다. 제안 알고리즘의 계산 과정은 총 3단계로 구성되어 있으며, 첫 번째 단계는 개별 가상 머신 로드량 계산 과정으로 아래 식 (1)과 같다.

$$R_{all}(x) = R_{cpu}(x) + R_{mem}(x) + R_{disk}(x) + R_{vga}(x) + R_{net}(x) + R_{sys}(x) + R_{con}(x) \quad (1)$$

R_x = utilization rate for factor $x \times$ affected rate of rate factor x for specific virtual machine

R_{all} = load rate for all factors, R_{cpu} = load rate for CPU usage, R_{mem} = load rate for memory usage, R_{disk} = load rate for disk I/O, R_{vga} = load rate for graphic I/O, R_{net} = load rate for network I/O, R_{sys} = load rate for system call, R_{con} = load rate for context switching

식 (1)에서는 표 2에서 제시한 각 요소의 현재 수치에 해당 요소가 가상 머신의 로드량에 영향을 미치는 비율을 곱한 값들을 모두 더함으로써 개별 가상 머신의 로드량을 계산한다. 두 번째 단계는 식 (1)에서 구한 값을 이용하여 로드 특성 분류에 따른 우선순위를 계산하는 과정이다.

$$P(x) = \frac{R_{w|d|g|m}(x)}{R_{all}(x)} \quad (2)$$

$P(x)$ = priority of virtual machine x

$R_{w|d|g|m}$ = rate of specific factors

(w: web browsing, d: documentation, g: game, m: multimedia)

식 (2)에서는 식 (1)에서 계산한 개별 가상 머신의 로드량을 표 3의 특성 분류와 같이 웹 탐색, 문서 편집, 게임, 멀티미디어 관련 작업 등 현재 처리 중인 작업에서 주로 요구하는 로드량의 합으로 나누어 해당 가상 머신의

우선순위를 계산하게 된다. 마지막 세 번째 단계에서는 식 (2)에서 구한 값을 이용하여 우선순위에 따른 로드량을 계산한다.

$$L(x) = \sum_{k=1}^n \frac{1}{P(n)} \times P(x) \quad (3)$$

$L(x)$ = load on virtual machine x
 n = number of virtual machines on the physical machine

식 (3)에서는 앞서 식 (2)에서 계산한 특정 가상 머신의 우선순위를 물리적 장치 내에 존재하는 모든 가상 머신들의 우선순위의 합으로 나누어 주며, 최종적으로 이 값이 제안 알고리즘을 적용한 로드 밸런싱 기법의 기준이 된다. 이와 같이 식 (1) ~ (3)을 이용한 제안 알고리즘의 계산 과정을 통해 해당 가상 머신이 전체 시스템 상에서 차지하는 로드량을 보다 정밀하게 측정하는 것이 가능하다.

3.3 기존 기법과 제안 알고리즘의 비교·분석

일반적인 컴퓨터 시스템에서는 물리적인 메모리량이 증가할 경우 CPU의 로드 발생 비율이 감소하고 이에 따라 특정 프로세스에 의해 발생하는 로드량에도 영향을 미치게 된다. 기존의 서버 가상화 환경에서는 각 가상 머신에 할당되는 물리적 자원량이 해당 가상 머신이 제공하는 네트워크 서비스에 맞추어 고정적이었다. 하지만 클라우드 컴퓨팅 환경에서는 가상 머신에 할당되는 물리적 자원의 양이 사용자가 요구하는 작업의 특성에 따라 수시로 변화한다. 클라우드 컴퓨팅 서비스가 가지는 이러한 동적 자원 재배치의 특성은 표 2에서 나타난 로드 밸런싱의 기준이 되는 각 요소들 또한 동적으로 변화시킨다. 제안 알고리즘에서는 이러한 각 요소들의 변화를 식 (1)의 계산 과정에서 R_x 값에 반영하였다. 따라서 제안 알고리즘은 기존 로드 밸런싱 기법과 비교하여 더욱 정밀한 로드 측정이 가능하며, 높은 신뢰도를 지닌다. 표 4에서 기존 로드 밸런싱 기법과 제안 알고리즘을 비교하였다.

<표 4> 기존 기법과 제안 알고리즘 비교

기준 요소	기존 기법	제안 알고리즘
CPU 점유율	○	○
메모리 사용량	○	○
I/O 횟수	○	○
system call 횟수	×	○
context switching 횟수	×	○
가상 머신 대기 시간	×	○
가상 머신 응답 시간	×	○
가상 머신 자원 요청 횟수	×	○
프로세스 단위 로드 측정	×	○
동적 자원 재배치 시 로드 변화 반영	×	○

표 4에서는 10가지 로드 밸런싱 기준 요소를 대상으로 비교를 실시하였다. 기준 요소를 만족하는 개수가 많을수록 보다 정밀한 로드 제어가 가능하며, 신뢰도 또한 높다고 할 수 있다. 표 4에서도 볼 수 있듯이, 기존 서버 가상화 환경에서의 로드 밸런싱 기법은 10개의 비교 항목 중 3가지 항목만을 만족하는 반면, 제안 알고리즘을 적용한 로드 밸런싱 기법은 모든 항목을 만족함을 볼 수 있다. 특히 마지막 비교 항목인 ‘동적 자원 재배치 시 로드 변화 반영’을 만족함으로써 제안 알고리즘이 데스크탑 클라우드 컴퓨팅 환경에 적합하다는 것을 알 수 있다.

4. 결 론

본 연구에서는 데스크탑 클라우드 컴퓨팅 환경을 위한 개선된 로드 밸런싱 알고리즘을 제안하였다. 클라우드 컴퓨팅 환경은 시스템 자원의 동적 재배치가 가능하다는 장점이 있지만 이에 따른 보다 세밀하고 유연한 로드 밸런싱 정책을 필요로 한다. 이에 본 연구에서는 기존 서버 가상화 환경에서의 로드 밸런싱 기법을 개선한 알고리즘을 제시하고, 제안 알고리즘을 기존 기법과 비교·분석 하였다.

참고문헌

[1] Michael Miller, “Cloud Computing: Web-Based Applications That Change the Way You Work and Collaborate Online,” Landmark Ltd.

[2] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, M. Zaharia, “Above the Clouds: A Berkeley View of Cloud computing,” Technical Report No. UCB/EECS-2009-28, University of California at Berkley, USA.

[3] 한국 IBM 시스템 테크놀로지 그룹, 「가상화 기술의 새로운 패러다임」, 한국경제신문

[4] Oracle Virtualization Technology Center, <http://www.oracle.com/technology/global/kr/tech/virtualization/index.html>

[5] Wikipedia, “Load balancing,” [http://en.wikipedia.org/wiki/Load_balancing_\(computing\)](http://en.wikipedia.org/wiki/Load_balancing_(computing))

[6] Mor harchol-balter and Allen B. Downey, “Exploiting Process Lifetime Distributions for Dynamic Load Balancing,” ACM Transactions on Computer Systems, vol.15, No.3, page253-285.

[7] Deepak C. Shadrach, Kiran S. Balagani, Vir V. Phooha, “A Weighted Metric Based Adaptive Algorithm for Web Server Load Balancing,” Intelligent Information Technology Applications, 2007 Workshop on, pp. 449-452, 2009 Third International Symposium on Intelligent Information Technology Application.