워크로드 섀이핑을 통한 클라우드 환경에서의 전력당 성능비 최적화 모델

김웅섭
동국대학교 컴퓨터 정보통신공학부
e-mail : woongsup@dongguk.edu

A Power-Performance Optimization Model on Cloud Environment Through Workload Shaping

Woongsup Kim
Department of Computer and Information Communication Engineering
Dongguk University, Seoul

요 약

클라우드 컴퓨팅에서는 사용량 당 과금 정책을 통해 서비스를 제공하여 사용자에게 높은 수준의 QoS 를 제공함과 동시에 비용절감의 효과를 가지고 있다. 하지만 클라우스 서비스 제공 업체에서는 최대 서비스 요구량을 만족시킬 수 있도록 시스템을 구성해야 할 필요가 있으며, 이에 맞추어 상당한 시간동안 다수의 자원을 유휴상태로 운영하여야 한다. 데이터 센터를 유휴상태로 운영될 경우 즉시 서비스 제공이 가능하다는 장점이 있으나 반대로 전력을 낭비한다는 단점을 가진다. 본 연구는 최소한의 전력소모를 하면서 QoS 를 보장할 수 있도록 하는 시스템 구축 모델을 제시하는 데 목적이 있으며 시뮬레이션 결과를 통하여 우리가 제시한 모델의 적절성을 보이려고 한다. 우리의 모델은 요청 작업 타입에 따른 traffic shaping 기법을 도입하여 작업을 저전력 컴퓨터와 고성능 컴퓨터에 분산배치하도록 하는데 목적이 있으며 가상화 기법을 통해 작업의 신속한 분산작업을 수행하는 방법을 사용한다.

1. 서론

클라우드 컴퓨팅이란 인터넷을 활용하여 IT 자원을 필요한 만큼 빌려서 사용하고 실시간 확장성을 지원받으며 사용한 만큼의 비용을 지불하는 컴퓨팅 서비스를 말한다. [1] 클라우드 컴퓨팅에서는 각 PC 단말에서 개별적으로 프로그램을 저장하던 방식에서 벗어나, 인터넷 네트워크 상에 모든 컴퓨팅 자원을 저장하여 개별 컴퓨터의 요청에 자원을 할당, 응답하는 방식을 사용한다. 클라우드 컴퓨팅에서는 서로 다른 물리적 위치에 존재하는 컴퓨터의 리소스를 가상화 기술로 통합 제공하는 것을 기본 원리로 하며, 리소스를 필요할 때 빌려쓰고 이에 대한 비용을 지급하는 방식의 서비스를 구현한다. 따라서 클라우드 컴퓨팅 기반 서비스를 제공하기 위해서는 하드웨어 인프라가 갖춰져 있는 데이터 센터 구축이 선행되어야 하며, 주문형 서비스, 동적 자원할당, 데이터 동기화, 서비스 과금 체계등의 기술이 요구된다.

그린 컴퓨팅이란 [2]에 의하면 컴퓨팅 자원 및 컴퓨터 시스템을 사용, 관리, 처리하는 데 있어서 환경에 최소한의 영향을 주거나 또는 영향을 주지 않도록 하는 컴퓨팅 시스템을 연구하거나 운영하는 것이라고 정의하고 있고, 그린 컴퓨팅을 실현하는데 가장 현실적인 방법으로 꼽히는 것은 Sleep mode 와 같은 저전력 기능을 이용하여 전력 소비를 줄이는 것이다 [3]. 예를 들면 컴퓨터 시스템을 유휴상태로 유지하는 것(일반적으로 100W 정도의 전력소비를 요구)에 비해 stand by 모드를 유지하게 될 경우 (5W 의 소비전력 소비 요구) 물리 서버 대당 95W 이상의 전력이 절약되는 것으로 알려져 있다. 더우기 물리 서버를 stand by 상태로 있는 것보다 power off 를 해주는 것이 더 많은 전력을 절약할 수 있는다. 하지만 서버를 power off 시키는 것은 해당 서버의 전력 사용량을 0 으로 만들어 전력 효율을 높일수 있으나, power 를 off 로 만들 경우 나중에 해당 서버를 power 를 on 으로 할 경우 startup 전력이 소모되고 이러한 startup 소비전력은 해당시간동안 서버를 on 으로 상태로 유지하는 것보다 더 많은 전력을 소비할 수 있다. 따라서 서비스 요구량을 예측, 전력이 확실히 절약이 되는 경우에만 off 를 하는 것이 필요하다. 또한 사용량이 예상밖으로 폭증하게 되는 경우, 즉각 필요한 자원을 확보해야 하는데 시스템 부팅시간 때문에 power 를 off 시키는 경우 즉시 자원의 확보가 안된다는 단점이 있다.

본 연구는 최소한의 전력소모를 하면서 QoS 를 보장할 수 있도록 하는 시스템 구축 모델을 제시하는 데 목적이 있다. 우리가 제시하는 모델은 온라인(online)예측 방법을 사용 Job 의 특성을 파악하고 workload traffic shaping 기법을 도입하여 작업을 저전력 컴퓨터와 고성능 컴퓨터에 분산배치하도록 하고

가상화 기법을 통해 작업의 신속한 분산작업을 수행하는 방법을 사용한다.

본 연구에서 제시하는 방법은 현재 workload shaping 모델을 설정한 후 이를 가상의 traffic pattern 에 대하여 적용하며 시뮬레이터를 통하여 그 효율성을 분석하는 단계에 있으며 본 논문에서는 현재까지의 연구 결과를 제시하고 있다.

2. 클라우드 환경에서의 전력 측정 모델

일반적인 클라우드 환경에서는 전력 효율성을 높이면서 QoS 와 고수준의 서비스 가용율을 보장하기 위하여 가상화 기법을 사용한다. 가상화 기법은 다수의 물리서버를 관리할 경우 자원이용율이 낮은 경우 물리서버 일부의 전원을 끄고 소수의 물리서버를 구동시키면서 다수의 물리서버가 돌아가는 것처럼 시스템을 구동하는 것으로 본 연구에서는 가상화를 사용 시스템을 구동하는 것을 가정하여 전력 성능비 측정모델을 구성하였다. 그 전력 측정모델은 아래와 같다.

$$P_{totla} = \int_0^t \sum_{i=1}^{N} P_i^s$$

여기서 P_i^S 는 N 개의 클라우드 노드 클러스터중에 i 번째 ($0 \le i < N$) 컴퓨팅 노드가 상태 S 에 있을 때의 전력 소모량을 가리키는 것으로 본 실험에서는 컴퓨팅 노드가 {Active, Standby, Off}의 4 가지 단계를 가지고 있는 것으로 가정하며 따라서 S = {Active, Standby, Off}가 된다. 또한 일반적으로 노드의 전력 소모량은 다음과 같이 정의가 되는데 [4],

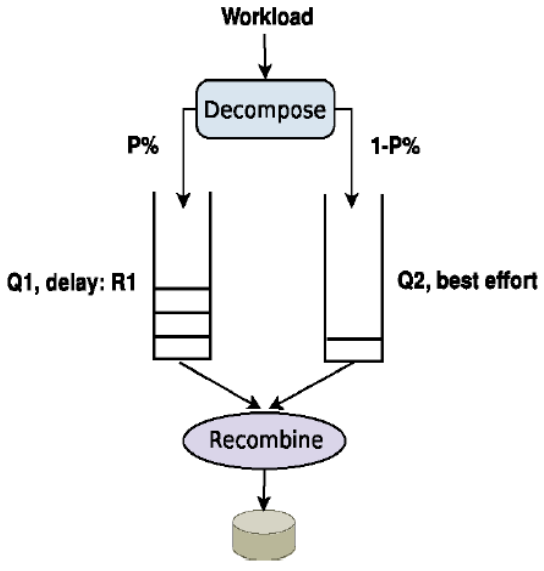
$$P_i^S = \alpha C_L V^2 f_S$$

C_L 은 CPU 의 capacity load, V 는 CPU 입력 전압, f_S 는 각 노드의 상태별 CPU 동작 클럭수를 의미한다. 본 연구에서는 편의를 위하여 클라우드 클러스터의 각 노드는 동일한 사양의 컴퓨터를 사용하는 것으로 가정하며, 따라서 V 의 값과 C_L 의 값은 컴퓨팅 노드에 상관없이 동일한 것으로 본다. 또한 컴퓨팅 노드가 Off 상태에서 Active 또는 Standby 상태로 전환할 경우 전환중에 전력 소모가 있는데 이경우 Active 상태와 Standby 상태에서의 전환은 즉시 시간소모 없이 즉시 전환하는 것으로 간주하며, Off 에서 Active 상태 또는 Standby 상태로의 전환시 Active 상태에서와 같은 전력소모를 가지는 것으로 간주한다. 본 실험에서는 Active 상태에서의 전력소모량을 95W, Standby 상태에서의 전력소모량을 5W, off 상태에서의 전력소모량을 0W 로 간주 계산하였다.

3. Worklaod Shaping 기법

workload shaping 기술은 사용자 요구내용의 특성을 파악하여 QoS 요구단계가 높은 작업과 그렇지 않은 작업을 구분하여 QoS 요구단계가 높은 작업은 전력을 많이 먹지만 고성능의 시스템에서 작업을 수행하도록 하고, QoS 단계가 낮은 작업은 느리지만 저전력 시스템에서 작업을 진행시키며 QoS 요구단계가 낮은 작업의 경우 그 실행을 강제 연기시켜 전체 작업 요청량의 burstness 를 줄이도록 하는 방법으로 workload 의 burstness 를 줄이게 되면 예측의 오류 발생 시 생기는 과 전력 소모를 줄일 수 있으며 또한 burstness 의 정도가 줄어들 경우 service availability 정도를 유지하는데 필요한 자원의 over-provisioning 마진을 줄일 수 있는 효과도 가진다.

(그림 1)은 본 연구에서 적용한 workload shaping 기법을 도시한 것이다. 클라우드 환경에서 본 연구에서 제안하는 workload shaping 기법을 적용하기 위해서 우리는 2 종류의 Queue 를 준비하였으며 클라이언트에게 주어진 서비스 요청을 두 종류의 큐에 분산 배치하여 급히 실행해야 하는 작업들의 경우 Q2 에 약간의 delay 가 용납되는 작업들의 경우 Q1 에 배치하고 (그림 1 의 Decompose) Q1 과 Q2 에 대기중인 작업의 수를 계산하여 Q2 의 작업에 작업 우선권을 주어 먼저 실행하도록 한다. 만약 클라우드 클러스터가 단일 노드가 아니라 다중 노드로 구성되어 있다면 Q2 의 작업들은 전력소모가 많지만 높은 성능의 노드에 Q1 의 작업들은 전력소모가 적지만 낮은 성능의 노드에 작업을 할당하는 식으로 작업을 배치하는 식으로 전력



소모를 줄인다.

(그림 1) workload shaping 기법

하지만 일반적인 Cloud 환경에서 작업들을 급히 실행하여야 하는 작업과 delay 가 용납되는 작업으로 분류하기에는 어려움이 많다. 작업우선권을 위한 큐 배치를 위하여 Cloud 클라이언트의 요청에 작업 우선권을 두도록 하여 작업을 분류하는 방식을 사용할 수도 있지만 본 연구에서는 그러한 방식이 클라우드 환경

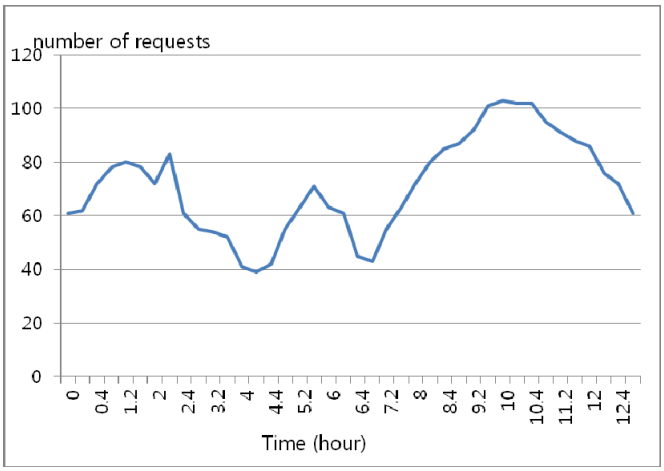
에서 일반적이지 않다고 보고 우리는 대신 작업 history 를 사용한 작업 우선권 설정을 하였다. Decompose 모듈에 요청 작업당 평균 실행시간을 관리하는 기능을 추가로 구현하여 평균 실행시간에 따라 작업 요청 시간 + 평균 작업 실행시간 * α의 시간을 요청 작업의 deadline 으로 설정했으며 dealine – 현재시간이 평균 작업 실행시간 * ρ이 되는 동안 실행이 안되고 있는 경우 Q1 의 작업을 Q2 로 우선 배치하여 작업에 우선권을 높이도록 하였다.

저전력 시스템 관리를 위해서는 (Q2 의 작업요청량 + Q1 의 front 작업 요청량)이 현재 가동중인 노드 capacity 합보다 높아지는 경우 standby 컴퓨팅 노드 또는 off 컴퓨팅 노드들을 활성화시켜 자원을 증가시키도록 하였으며 Q1 + Q2 의 작업량이 (현재 가동중인 노드수 -1)의 capacity 합보다 낮은 경우 자원을 회수하도록 하였다.

4. 실험

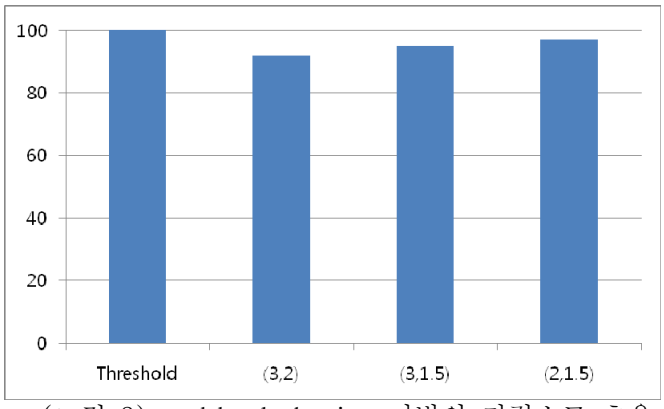
본 실험은 합성 traffic data 에 대하여 시뮬레이션 작업을 통해 전력 소모량을 계산하는 식으로 진행하였다. 합성 traffic data 의 경우 indexing 작업을 진행중에 발생하는 내부 명령어들을 기준으로 만들었으며 클라우드 시뮬레이션은 CloudSim[5]를 통하여 진행하였다. 본 실험에서 사용된 노드의 수는 총 3 개로 하였으며 active 노드의 수가 1 ~ 3 개로 동적으로 변환하도록 하였다.

(그림 2)는 본 실험에서 사용된 합성 traffic data 의 traffic 발생량이며 시간당 요청 작업의 수를 나타낸다.



(그림 2) 시간별 job requests 수

(그림 3)은 threshold 방식의 동적 노드 관리 시스템과 본 연구 방법과의 전력 절감 비교 결과는 나타낸 것으로 다양한 α, ρ 값에 따른 전력 절감 효과를 나타낸 것이다. Threshold 방식은 서버의 이용율이 사전에 지정된 threshold high 값보다 높을 경우에 active 한 컴퓨팅 노드의 수를 늘리며 threshold low 값보다 낮은 경우에 over provisioning 이라고 판단하여 컴퓨팅 노드 중 하나를 off 상태로 만드는 것을 방법을 의미한다.



(그림 3) workload shaping 기법의 전력소모 효율

(그림 3)에서 (3,2), (3,1.5), (2, 1.5)의 값들은 섹션 3 의 패러미터 α와 ρ 값을 나타낸 것이다. (그림 3)에서 보이듯이 우리의 접근 방법이 일반적인 threshold 방법에 비하여 더 효율적인 것으로 나타나며 α와 ρ 값의 변화에 따라 조금 Q2 노드에 들어가는 작업들에 변화가 있고 그에 따라 전력 효율이 변화하는 것을 알 수 있다.

5. 결론

본 연구에서 우리는 클라우드환경에서의 workload shaping 기법을 통한 전력 절감모델을 제시하였다. 시뮬레이션 결과에 따르면 우리의 방법이 전력 사용에 효율적이라고 판단된다. 우리는 실제 cloud 환경에서 workload shaping 전력 효율화 모델을 구현중에 있으며 차후에 우리 접근방법의 우수성을 보일 수 있을 것이라고 판단된다.

참고문헌

[1] 한국전자통신연구원, "모바일 클라우드 기술동향", 2010.6

[2] San Murugesan, “Harnessing Green IT: Principles and Practices,” IEEE IT Professional, January–February 2008, pp 24-33

[3] W. H. Kemp, The Renewable Energy Handbook: A Guide to Rural Energy Independence, Off-Grid and Sustainable Living, Aztext Press, 2006.

[4] R. Racu, A. Hamann, R. Ernst, B. Mochocki, X. Hu, Mothods for Power Optimization in Distributed Embedded Systems with Real-Time Requirements, CASES 06

[5] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "Cloudsim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," Software: Practice and Experience, vol. 41, pp. 25-50, 2011.