클러스터 시스템상에서의 CPU 전원 관리

오수철, 김성운
한국전자통신연구원
e-mail : {ponylife, ksw}@etri.re.kr

CPU Power Management on Cluster Systems

Soo-Cheol Oh, Seong-Woon Kim
Electronics and Telecommunications Research Institute

요 약

클러스터 시스템은 가격대 성능비의 효율성 때문에 다양한 분야에서 활용되고 있으며, 구축 규모도 급속히 증가하고 있다. 특히, 인터넷을 통한 정보 검색 및 공유가 활발하게 이루어지면서, 정보를 수집, 가공 및 제공하는 대형 포털들의 규모가 급속히 증가하고 있다. 포털들은 대량의 정보를 서비스하기 위해서 대규모의 클러스터 시스템을 운영하고 있으며, 이러한 시스템을 유지 관리하는 것은 커다란 문제점중의 하나이다. 대규모 클러스터 시스템의 운영 비용중에서 전력비용이 상당히 큰 부분을 차지하고 있으며, 이를 감소시키려는 다양한 시도가 진행되고 있다. 본 논문에서는 클러스터 시스템의 전력사용량을 감소시키기 위해서 CPU 의 전력을 효율적으로 관리하는 있는 관리 메커니즘을 제안한다.

1. 서론

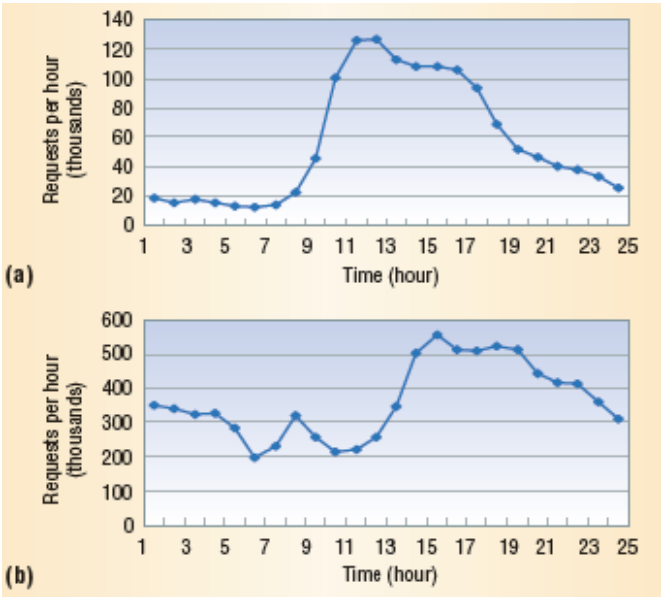
인터넷이 등장한 이후로 수많은 정보가 인터넷을 통해서 공유가 되고 있으며, 현재는 세상에서 생산되는 대부분의 정보들이 인터넷상에서 활용되고 있다. 또한, 인터넷을 활용한 정보의 생산 및 공유는 UCC 의 등장과 더불어 앞으로도 급속한 성장을 보일 것으로 예측된다. 인터넷을 통한 정보 공유의 핵심에는 정보를 수집, 저장, 가공 및 서비스하는 대형 포털들이 존재하며, 이들의 규모는 나날이 커지고 있다. 대형 포털들은 대량의 정보 처리를 수용할 수 있는 컴퓨팅 인프라를 필요로 하며, 그 규모는 앞으로도 지속적으로 증가할 것이다.

세계에서 가장 큰 규모의 포털은 구굴은 현재 약 45 만대의 서버를 운영중인 것으로 추정되며[1], 국내의 대형 포털들도 약 1 만대 수준의 서버를 운영중인 것으로 추정된다.

이러한 대규모 클러스터 시스템이 소모하는 전력량은 전체 시스템의 운영비용에 있어서 상당한 부분을 차지하고 있다.

(그림 1)은 웹 사이트를 기준으로 했을 때 하루 동안의 부하(사용자 요청)을 그래프로 나타낸 것이다[2]. (그림 1-a)는 금융 웹사이트를 나타내며 (그림 1-b)는 98 년 올림픽 웹사이트의 기록이다. 그림을 보면 매 시각마다 웹 사이트에 부과되는 부하가 다름을 알 수 있다. (그림 1-a)를 보면 11 시 – 12 시 사이에 최대 부하를 보이며 대부분의 시간에서는 최대 부하에 훨씬 못 미치는 것을 볼 수 있다. 이러한 경향은 (그림 1-b)도 동일한다.

포털 사이트들의 경우, 최대 부하 시간에 맞게 시스템을 설치하기 때문에, 최대 부하가 걸리지 않는 대부분의 시간에는 시스템 자원이 낭비되는 현상이 발생한다. 즉 설치된 시스템의 많은 부분이 작업을 수행하지 않는 상황이 발생하며, 이와 더불어 전력도 낭비되는 현상이 발생한다. 따라서, 최대 부하 시간을 제외한 시간에 낭비되는 전력을 감소시킬 필요가 있다.



(그림 1) 웹 사이트의 부하 분포

본 논문에서는 클러스터 시스템을 구성하는 각 노드에서 가장 많은 전력을 소모하는 CPU 를 대상으로 하여 클러스터 시스템 자체에서 전력을 관리하는 메커니즘을 제안한다.

2. 배경 연구

하나의 컴퓨터 시스템은 CPU, 메모리 및 하드 디스크등의 여러 가지 부품으로 구성되며, 각자 소모되는 전력이 틀리다. 이중에서 CPU 는 각 컴퓨터 시스템 전력의 30~40%를 소모하는 아주 중요한 요소이다. 따라서 CPU 의 전력을 관리함으로써 컴퓨터 시스템의 전력 소모를 효율적으로 제어할 수 있다.

DVS (Dynamic Voltage Scaling) 은 CPU 의 동작속도 및 전압을 조정함으로써 CPU 에서 소모되는 전력을 감소시키는 기술이다. 클럭에 동기화되어 동작하는 CPU 의 특성상 동작속도를 낮추게 되며 CPU 의 전력 소모가 감소한다. 또한 전압의 경우 아래의 식을 고려할 때 전압의 제곱에 비례하여 전력이 감소되는 것을 알 수 있다.

$$P = V * A = P = V * (V/R) = V^2 * R$$

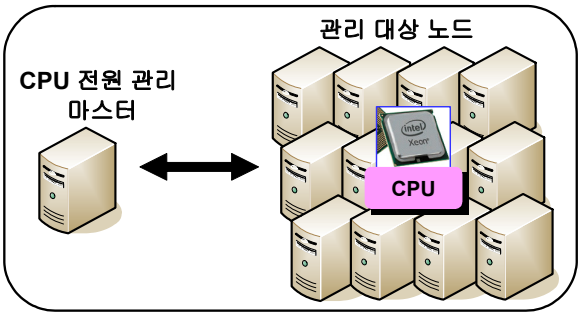
(P:전력, V:전압, A:전류, R:저항)

이러한 DVS 기술은 Intel 및 AMD 의 x86 기반 CPU 에 구현되어 있으며, Intel 은 SpeedStep 과 DBS (Demand-Based Switching) [3]이라 부르며, AMD 는 PowerNOW [4] 라고 칭하고 있다.

리눅스의 경우, 현재 커널버젼 2.6 이상에서 Intel 및 AMD 의 DVS 기술을 지원하고 있다. 이러한 관리 기술은 개별 운영체제에서 단일 노드를 대상으로 CPU 의 전원을 관리하는 것으로, 전체 클러스터 시스템 관점에서 이를 관리하는 기술은 지원하지 않는다. 따라서 본 논문에서는 리눅스의 DVS 관리 기술을 활용하여 전체 클러스터 시스템차원에서 이를 관리하는 메커니즘을 제안한다.

3. CPU 전원 관리 구조

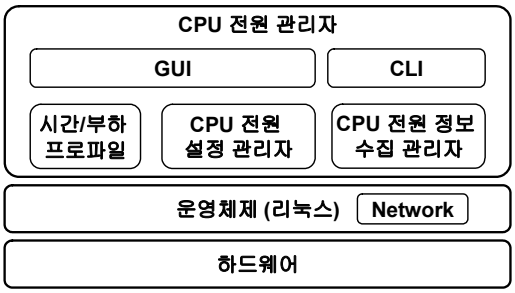
본 논문에서 제안하는 CPU 전원 관리 구조는 (그림 2)와 같이 CPU 전원 관리 마스터와 관리 대상 노드들로 구성된다. CPU 전원 관리 마스터는 전체 클러스터 시스템의 CPU 전원 관리를 담당하는 노드로, 관리 대상 노드들의 동작 모드 변경 및 CPU 동작 속도 변경에 필요한 각종 파라미터들을 설정하는 역할을 한다. 관리 대상 노드들은 클러스터 시스템에 존재하는 일반 노드들로써 CPU 전원 관리 마스터가 설정한 파라미터에 따라서 자율적으로 CPU 동작 속도를 변경하는 역할을 담당한다.



(그림 2) CPU 전원 관리 구조

4. CPU 전원 관리 마스터

CPU 전원 관리 마스터의 구조는 (그림 3)과 같으며, 하드웨어 및 운영체제상에 CPU 전원 관리자를 탑재한다. CPU 전원 관리자는 시간/부하 프로파일, CPU 전원 설정 관리자, CPU 전원 정보 수집 관리자, CLI(Command Line Interface) 및 GUI 로 구성된다. 시간/부하 프로파일은 (그림 1)과 유사하게 시간대별로 클러스터 시스템의 부하를 기록한 프로파일로 클러스터 시스템의 실제 수행 환경에서 얻어지는 내용이다. 이 프로파일은 시스템의 운영도중에 최신의 정보를 반영하기 위해서 주기적으로 갱신된다.



(그림 3) CPU 전원 관리 마스터

CPU 전원 설정 관리자는 관리 대상 노드들이 CPU 동작 속도 변경을 수행하는데 있어 필요한 각종 파라미터들을 설정하는 역할을 담당한다. 설정하는 주요 파라미터는 다음과 같다.

- CPU 전원 관리 on/off
- 최대 동작 속도 시간
- 동작 속도 변경용 부하 임계치
- 강제 동작 속도

이러한 파라미터가 관리 대상 노드에 설정되면, 관리 대상 노드는 이러한 파라미터를 기준으로 하여 자율적으로 CPU 동작 속도 변경을 수행한다.

CPU 전원 정보 수집 관리자는 주기적으로 관리 대상 노드들의 현재 동작 모드 및 CPU 동작 속도를 수집하고 이를 CPU 전원 관리 마스터의 하드디스크에 저장하는 역할을 담당한다. 이러한 수집 정보는 시스템의 동작 상황을 분석하는데 활용된다.

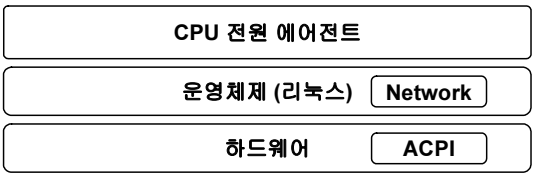
CPU 전원 관리자는 사용자 인터페이스로 커맨드라인 기반의 CLI 와 그래픽 기반의 GUI 를 지원한다.

이 인터페이스에서 지원하는 주요 기능은 다음과 같다.

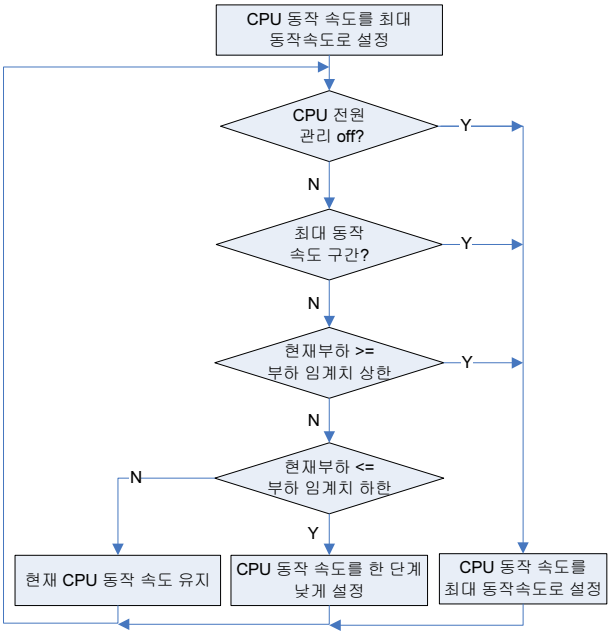
- 동작 속도 변경 파라미터 설정
- 관리 대상 노드들을 그룹 단위로 관리
- 수집된 전원 정보를 사용자에게 제공
- 시간/부하 프로파일의 생성/수정 지원

5. 관리 대상 노드

관리 대상 노드에는 (그림 4)와 같은 CPU 전원 관리 에이전트가 수행되며, CPU 전원 관리 마스터가 설정한 파라미터에 따라서 CPU 동작 속도 변경을 (그림 5)와 같이 수행한다. 이러한 작업은 정해진 주기마다 반복적으로 수행된다.



(그림 4) 관리 대상 노드



(그림 5) CPU 동작 속도 변경

- CPU 전원 관리 on/off

CPU 전원 관리가 off 되면 관리 대상 노드들의 CPU 는 항상 최대 동작속도로 동작한다. CPU 전원 관리가 on 되면 관리 대상 노드들은 시간, 부하 임계치등의 다른 파라미터에 따라서 자율적으로 CPU 동작속도를 변경한다.

- 시간

CPU 전원 관리가 on 되면 관리 대상 노드는 시간 파라미터를 검사한다. 현재 시간이 최대 동작 속도구간으로 설정되어 있으면, CPU 는 항상 최대 동작속도로 동작하도록 설정한다. 최대 동작 속도 구간이 아니면 다음에 나오는 부하 임계치에 따라서 동작한다.

- 부하 임계치

본 모드에서는 현재 노드의 부하를 주기적으로 조사한다. 시스템의 현재 부하가 부하 임계치 상한을 초과하면 즉시 최대 동작 속도로 동작하여 사용자의 요청을 처리하도록 한다. 시스템의 부하가 부하 임계치 하한보다 낮으면 CPU 동작 속도를 한 단계 낮은 동작 속도로 낮게 설정한다. 예를 들어서 부하 임계치 상한은 80, 부하 임계치 하한은 20 으로 설정할 수 있다.

(그림 1)에 표시된 시간 대비 부하를 고려하면 클러스터 시스템은 대부분의 시간에 CPU 전원 관리가 on 되고 현재부하와 부하 임계치를 비교하여 현재 CPU 의 동작 속도를 결정하는 모드로 운용될 것이다. 즉 대부분의 시간에서 CPU 는 최대 동작속도 보다 낮은 속도로 동작할 것이며, 이로 인한 전력 감소 효과를 볼 수 있을 것이다.

6. 결론

본 논문에서는 클러스터 시스템에서 소모되는 전력량을 감소시키기 위해서 각 노드들의 CPU 전원 관리를 원격으로 제어할 수 있는 프레임워크를 제안하였다. 본 연구팀은 현재 제안한 구조에 맞춰 시스템 구현을 진행중이다. 또한 구현이 완료된 후에는 실제 실험을 통하여 전력 감소 효과를 측정하며, 각종 파라미터 튜닝을 수행할 것이다.

참고문헌

[1] "Google Platform", Wikipedia, http://en.wikipedia.org/wiki/Google_platform#_note-google_arch.
[2] Charles Lefurgy, Karthick Rajamani, Freeman Rawson, Wes Felter, Michael Kistler, Tom W. Keller, "Energy Management for Commercial Servers", IEEE Computer, pp. 39-48, December, 2003.
[3] "SpeedStep", Wikipedia, http://en.wikipedia.org/wiki/SpeedStep.
[4] "AMD PowerNow Technology", AMD.