ETRI
전자통신동향분석 제24권 제4호 2009년 8월

그린 데이터센터를 위한 전력 관리 기술

The Power Management Technology for Green Datacenter

클라우드 컴퓨팅 특집

김대원 (D.W. Kim) 서버플랫폼연구팀 선임연구원
김선욱 (S.W. Kim) 서버플랫폼연구팀 선임연구원
김성운 (S.W. Kim) 서버플랫폼연구팀 팀장

목 차

언제부터인가 주위는 모두 “녹색” 물결로 뒤덮혀 버렸다. 그 중에서도 IT의 존재는 녹색 물결에 휩싸여 새로운 판로를 모색하고 있다. 데이터센터 역시 그 문제점은 수 년 전부터 대두되어 왔고, 정책적으로 혹은 업계 자체적으로 데이터센터의 문제점을 해결하고자 많은 솔루션을 내고 있는 실정이다. 그린 데이터센터를 위한 견해는 먼저 비 IT 영역으로 공조, 항온/항습 시설, 전력 배전 시설 등의 관점에서 접근하는 견해가 있고 IT 영역으로 서버 자체에 대한 관점에서 접근하는 견해가 있다. 본 고에서는 데이터센터 내에서 IT 영역에 대한 전력 관리를 위하여 해야 될 역할에 대하여 언급하고자 한다. 특히 데이터센터 내 서버 자체의 전력 절감을 위한 방법에 대하여 논하고, 앞선 업체들 특히 IBM, Google, SUN 등의 동향으로부터 그린 데이터센터를 위한 방향을 모색해 보기로 한다.

I. 데이터센터를 위한 전력 관리

일반적으로 데이터센터의 모든 시설을 설비(facility)라 하고, 이를 관리하는 것을 설비 관리(facility management)라 한다. 설비 관리에 있어서 IT 관리는 직접 IT와 관련된 인프라와 서버, 컴퓨터 등에 대한 관리를 말하고, Non-IT 관리는 데이터센터 내 필요한 시설, 즉 조명, 기계적인 시스템, 전기 발전 시설 및 발전기 등에 대한 관리를 말한다. 일반적으로 그린 데이터센터라 하면 포괄적으로 모든 분야에서 효율적인 설비 관리를 말할 수 있으나, 기본적으로 데이터센터의 에너지 소비 패턴이 IT에 집중되어 있으므로 이를 위한 전력 관리가 그린 데이터센터를 위한 가장 시급히 해결되어야 할 문제일 것이다. 본 장에서는 큰 개념의 IT와 관련된 전력 관리(PM)의 기능, 대상 및 방법에 대하여 정의하고, 이를 구현하기 위한 하드웨어적인 방법 및 소프트웨어적인 방법에 대하여 알아보도록 한다.

1. 전력 관리의 기능

전력 관리의 기능이란 전력 관리를 위하여 필요한 기능을 말한다. 전력 관리를 위해서는 다음과 같은 기능들이 요구된다.

① 전력 모니터링 기능: 관리자 혹은 OS는 시스템에 대한 전력 정보를 알 수 있어야 한다.
② 전력 예측 기능: 관리자 혹은 OS는 시스템에 대한 추정 가능한 모델을 통하여 예측 가능하여야 한다.
③ 전력 제어 기능: 관리자 혹은 OS는 시스템에 대한 예측 혹은 모니터링을 통하여 제어를 할 수 있어야 한다.

전력 관리 기능에 있어서 모니터링 기능은 현재 많은 하드웨어 센서를 이용한 표준 방법(IPMI, PSMI[1], ACPI) 등을 통하여 제공되고 있고, 제어 기능 역시 각 디바이스별로 제공되는 기능을 통한 표준 인터페이스로서 충분히 역할이 수행될 수 있다. 예측 기능은 모델을 통한 다양한 방법 등이 연구되고 있는데 예측 기능의 모델은 실제 시스템과의 유사도에 따라 어느 정도 적응 기간이 필요하게 되고 과거의 데이터나 분석 등을 통하여 실제 시스템을 예측할 수 있도록 하는 것이 일반적인 구현 사례이다.

● 용 어 해 설 ●

ACPI: 고급 구성 및 전원 인터페이스(ACPI) 규격은 HP, 인텔, 마이크로소프트, 피닉스, 그리고 도시바가 개발하고, 1996년 12월에 처음 공개된 표준 규격으로서 하드웨어 감지, 메인보드 및 장치 구성, 전원 관리를 담당하는 일반적인 인터페이스를 정의한다.

IPMI: 1998년 인텔, 델, HP, NEC 등이 시스템의 Health 정보나 관리를 위하여 컴퓨터 시스템에 정의한 표준 인터페이스이다.

DVFS: CPU의 전압 및 주파수를 조절할 수 있는 기능으로서 인텔의 Speedstep, AMD의 Powernow 기능을 말한다.

2. 전력 관리의 범위

전력 관리를 위한 기능들은 다음과 같이 데이터센터의 IT와 관련된 대상에 따라 그 범위를 구분할 수 있다.

① Single System Level: 하나의 시스템, 즉 서버나 컴퓨터를 관리 대상으로 하는 전력 관리를 뜻한다. 이는 하나의 서버에 부착되어 있는 많은 하드웨어를 저전력 기법을 사용하여 전력 관리를 하는 것을 그 대상으로 한다.
② Multi System Level: 사용자의 요구에 따라 특수 목적으로 그룹핑된 여러 대의 시스템의 전력 관리를 뜻한다. 이는 다음에 사용되는 rack level 보다 더욱 큰 수준이 될 수도 있고 혹은 더 작은 수준이 될 수도 있다.
③ Rack Level: Rack level의 데이터 관리로서 rack과 함께 전력 관리를 수행하게 되며 multi system level과는 하드웨어적인 관점에서 전력 관리에 필요한 요소들을 수반하게 되는 경우가 많다.
④ Data Center Level: 데이터센터 레벨 전체에 관한 전력 관리를 말하는 것으로 보다 대규모

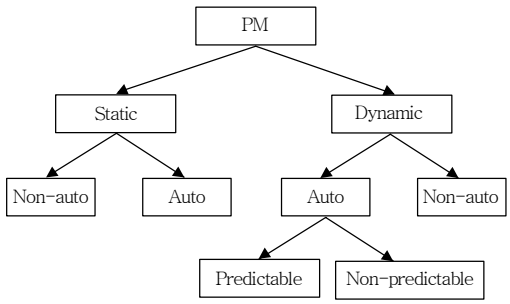
의 전력 관리 기법이 적용되며 서버나 컴퓨터에 국한되는 것이 아니라 IT와 관련된 모든 전력 관리 대상을 포함하는 것이다.

이 대상들은 데이터센터 내에서 전력 관리를 어느 범위까지 할 수 있느냐가 결정되는 요소이다.

3. 전력 관리의 방법

전력 관리 기능과 전력 관리 범위가 주어지면 이제 어떻게 그 기능을 설계할 것인지가 관건이다. 실제 전력 관리 방법은 크게 (그림 1)과 같이 분류할 수 있다. (그림 1)에서 전력 관리(PM) 방법은 크게 정적(static)과 동적(dynamic)으로 나뉜다. 정적 방법은 실제로 전력관리에 있어서 수동적으로 사용자의 요구사항을 받아 저전력 관리에 반영하는 기법을 말하고 이는 수동적인 사용자 요구사항을 적절한 시기와 장치에 반영하는 방법이 더 이상 사용자의 작업이 필요하지 않고 자동적으로 수행되느냐 그렇지 않느냐에 따라 non-auto 및 auto로 나뉠 수 있다.

그리고 동적 방법은 실시간 정보의 반영 여부에 따라 그때그때 적용이 바뀌어질 수 있도록 하는 전력 관리 방법을 말하는 것으로 여기서도 실시간 반영을 사용자의 작업 여부에 따라서 auto 및 non-auto로 나뉜다. 여기서 auto의 방법에는 다시 미래의 일을 예측하여 적용할 수 있느냐 없느냐에 따라서 predictable 혹은 non-predictable로 나눌 수 있다. (그림 1)에서 보는 것처럼 전력 관리 방법을 어떤 대상에 적용하느냐에 따라 다양한 전력 관리 시스템이 나올 수 있다.



(그림 1) 전력 관리 방법

4. 전력 관리의 구현

지금까지 전력 관리 기능, 범위, 방법에 대하여 알아 보았다. 이런 기능들이 구현되기 위하여 IT 관리는 하드웨어와 소프트웨어를 이용하여 구현된다. 일반적으로 전력 관리에 있어서 하드웨어의 역할은 중요하다. 현재 소프트웨어로 구현을 위해서는 하드웨어의 도움이 거의 필수적이다. 기본적으로 하드웨어의 도움이 없이는 전력을 줄이는 일이 쉽지는 않다. 현재 많은 장치(CPU, HDD, memory, power supply etc)에서 저전력 기법을 이용하여 고안된 제품들이 많이 나오고 있다. 전력 관리는 이를 바탕으로 설계된다. 하드웨어에서 전력 관리를 위하여 제공되어야 하는 것은 모니터링 기능과 제어 기능이다. 모니터링 기능은 BMC를 이용한 다양한 센서를 이용하여 제공된다. 그리고, CPU의 경우에는 PMU을 이용하여 제공된다. 제어 기능의 경우는 <표 1>에 나타내고 있다. 하드웨어 입장에서 보면 자체적으로 전력을 작게 쓰는 기능 방법들이 하드웨어 구현 차원에서 많이 연구되고 있고 이를 통하여 저전력 구현이 자체적으로 많이 진행되고 있다. 그러나, 시스템 차원에서 보면 이를 제어하는 방법에 대한 문제로 귀착된다. 자체 저전력 기능도 중요하지만 시스템에서 이들을 제어하는 <표 1>에서 보듯이 컴퓨터에 있어 기본적으로 사용되는 장치들이며 이들 장치를 전력 관리 차원에서 제어 가능한 기능을 살펴보면 그리 다양한 편은 아니다.

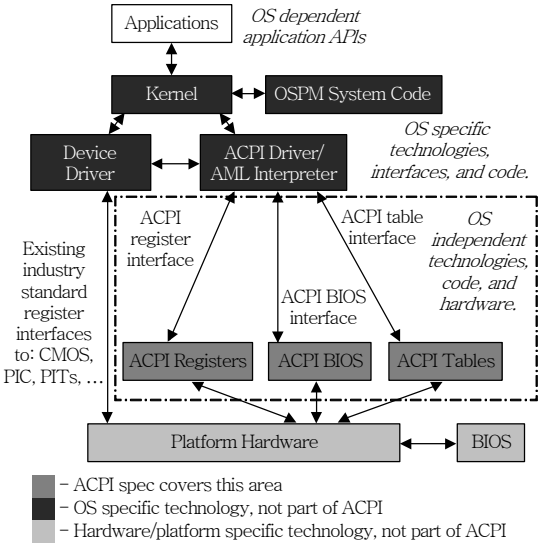
방법에 따라 전력 소모량은 유동적이게 된다. CPU의 경우 DVS 혹은 DVFS라 불리는 동적 전압 관리 혹은 동적 전압/주파수 관리가 거의 모든 CPU 상(Intel, AMD, VIA, ARM etc)에서 구현되어 있으나, 메모리의 경우는 이렇다 할 제어 방법은 없다.

〈표 1〉 장치별 제어 기능

Processor	- DVFS - Power Gating[2]
RAM	- ZettaRAM[3], Multi-Voltage 기능[4]
HDD	- Idle State 기능 - Multi Spin-Speed 기능 - Spin Down 기능
Power Supply	- 원격 On-Off 기능

삼성에서 개발한 다중 전압(1.2~1.5 V) 지원 방식이 있지만 대중화되어 있는 기능은 아니다. 그리고, 하드 디스크의 경우 Seagate, Hitachi 등이 디스크의 회전 속도를 조절하는 multi-spin speed 기능을 제공하고 idle state시 전력 소모를 줄이는 방법 등을 고안하고 있다. 그리고 위에서 열거하지 않은 많은 디바이스의 경우 자체 전력 관리 기능이 제공되는 것들이 있다. 이는 ACPI 표준 state를 지원하기 위하여 PCI express의 전력 관리 기능과 연계된 기능들이 대부분이다. 소프트웨어적인 측면에서 관리 시스템을 바라보는 것은 관리 대상 차원에서 어디에 어떤 방식으로 전력을 줄일 수 있는지에 대하여 공통된 방법들이 요구된다. 전력 관리를 위한 소프트웨어는 일단 하드웨어의 기능을 한층 견고하게 해줄 수 있다. 전력 관리를 위한 소프트웨어란 하드웨어를 제어하는 디바이스 드라이버 및 ACPI, PSMI 같은 전력 관련 인터페이스, 라이브러리를 포함하고 이를 바탕으로 구현된 시스템 차원의 관리 소프트웨어도 포함된다. 현재 각 장치별 제조사에서 제공하는 저전력 기능은 따로 장치 드라이버들이 존재한다. 그리고, OS의 전원관리 기능과 함께 단일 시스템 내에서는 많은 관리 기능이 구현되어 있는 실정이다. 그러나 이와 같은 방법들은 현재 대중적인 OS(Linux, Window)에 있어서 모든 장치들을 커버하기란 쉽지 않다. 그러므로 이를 위한 표준 인터페이스들이 존재하게 된다. ACPI는 인텔, 마이크로소프트, 도시바 등의 업체들이 표준 인터페이스를 만들어 시스템 차원에서 전력 관련 인터페이스를 제공하는 것으로 잘 알려져 있다. 이 ACPI의 경우 OS에 무관하게 인터페이스를 제공하고 각각의 장치들을 OS가 제어할 수 있도록 하고 이를 통한 인터페이스를 통하여 사용자 애플리케이션이 제공되도록 한다.

(그림 2)는 ACPI 인터페이스의 구조를 보여주고 있다. ACPI는 각 디바이스별 및 전체 시스템별 상태를 정의한다. 이들 상태들은 각각 장치의 지원 여부에 따라 다르겠지만 현재 제공되는 윈도나 리눅스 상에서 하드웨어 기능이 제공되는 한 거의 모든 기능은 제공된다.



<자료>: www.acpi.info

(그림 2) ACPI 인터페이스

Ⅱ. 정책 중심의 전력 관리

1. 정책 중심의 전력 관리의 의미

전력 관리에 있어서 시스템의 요구사항에 적합하게 전력 관리 기능을 구현하는 것은 이기종의 전력 관리 시스템에서 필수적인 요구사항이다. 각 서버, 데이터센터 마다 시설규모, 서비스 종류, 사용자 패턴 등에 따라 시스템은 서로 다를 수 있다. 이에 맞게 전력 관리 시스템을 구현하는 것이 바로 정책 중심의 전력 관리(policy-based power management)이다. 이는 단일 시스템의 적용이라기 보다 다중 시스템, 즉 클러스터 레벨에서의 전력 관리 시스템에 적용되어야 할 방법이다. 즉, 정책 중심의 전력 관리라 함은 현재 구현되어 있는 하드웨어 혹은 추가적으로 제작 가능한 저전력 하드웨어를 기반으로 클러스터 운영체제 상에서 운영자의 의도, 시스템 상태 및 서비스 상황, 사용자 패턴 등을 모니터링한 결과 혹은 모델을 통한 예측 가능한 결과 등을 바탕으로 전력 절감을 하기 위한 소프트웨어적인 기법을 말한다. 앞서 이야기 하였듯이 운영자의 의도, 시스템 모니터링 결과 혹은 모델을 통한 예측 가능한 결

과 등은 전력 관리 정책의 기반이 되고 이를 바탕으로 운영자 혹은 소프트웨어가 관리하는 방법 그 자체가 정책이라 말할 수 있고, 개발자는 이를 바탕으로 시스템에서 적용할 수 있도록 동적으로 설계하여야 한다. 그러므로 정책 중심의 전력 관리의 방법은 DPM을 이용하여 설계되는 것이 바람직하고 전력 관리 대상은 일반적인 전력 관리 대상(단일, 다중, 랙, 데이터센터 레벨) 모두에 적용 가능하다.

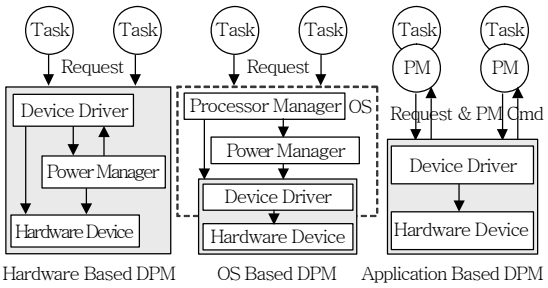
2. 정책 중심의 전력 관리의 분류

정책 중심의 전력 관리는 구현 도구에 따라 다음과 같이 분류된다.

① Hardware Based DPM Policy – 하드웨어의 activity와 workload를 관찰하여 power state를 조절

② Software Based DPM Policy
- ACPI Based DPM Policy: (그림 2)에 설명된 구조를 이용하여 정책을 설정
- OS Based DPM Policy: OS가 현재 동작중인 프로세스와 하드웨어와의 상호 작용을 알고 있다는 점을 이용(task based PM, power-oriented process scheduling)하여 정책을 결정
- Application Based DPM Policy: 각 상태에 따른 전력, transition energy, delay, future workload와 같은 파라미터로 power state를 결정

(그림 3)은 위의 3가지 정책 중심 DPM의 분류에 대한 기본적인 예를 나타내고 있다. 이 그림은 단일 시스템에 대한 구조이며 이를 확장 적용하면 다중 시스템에 대한 구조를 적용할 수 있다. 정책 중심의 전력 관리는 구현 도구에 따라 분류될 수 있지만 그 정책의 내용에 따라 다음과 같이 분류할 수 있다.



(그림 3) 정책 중심 DPM의 예[5]

① Level Based Power Policy – 전력 관리 대상에 해당되는 정책으로서 단일 노드, 그룹별 노드에 적용하는 정책으로 일반적인 전력 관리를 말한다.

② Application Based Power Policy – 현재 수행중인 애플리케이션에 따른 정책으로 현재 수행중인 프로그램 단위에 따른 정책을 결정한다.

③ Service Based Power Policy – SOA architecture 혹은 웹 서비스 등 일련의 서비스를 위한 정책으로 네트워크 등의 QoS에 따른 설정이 해당된다. Application Based Power Policy의 확장 형태이다.

④ Time Based Power Policy – day, month, year 등 시간에 따른 전력 관리 정책을 말한다.

Ⅲ. 업계의 전력 관리

현재 활발이 진행되고 있는 3개의 업체 Sun, Google, IBM의 클러스터 레벨의 전력 관리 기술과 DPM의 기술에 대하여 알아보도록 한다.

1. SUN의 Tesla Project

Sun은 Tesla Project[6]라 하여 OpenSolaris 상에서 PM을 구현하기 위한 공개 프로젝트를 2009년 2월에 시작하여 진행중에 있다. 이 프로젝트는 현재 노트북이나 랩톱 컴퓨터에 국한되어 있는 저전력 이슈를 서버에 도입시키기 위하여 진행되었고 그 목표는 최고의 성능, 최소의 전력으로 요구되는 서비스 latency를 만족시키며 최대의 시스템 신뢰성을 구현하자는 취지이다. 그리고 이를 확장하여 데이터센터 내에서 자원을 관리하는 방법을 연구하자는 것이다.

Sun의 연구 취지는 다음과 같은 근거에서 출발하였다.

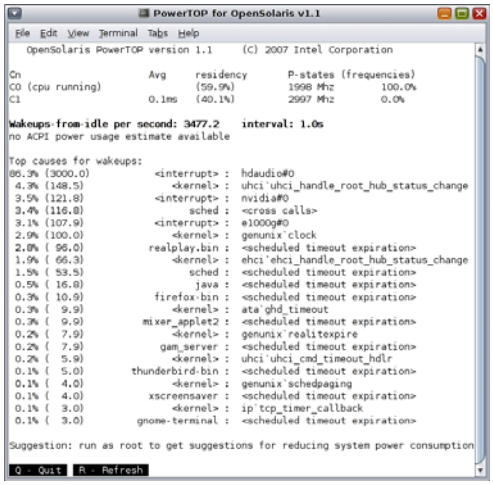
① 소유 총비용(TCO) 중 전력/냉각 비용의 증가를 무시하지 못함
② 서버 및 데이터센터의 유용성이 매우 낮음(평균 20%)
③ Idle 상태의 전력 소비가 심함(working server의 70~90% 정도 소모 시스템에 따라 차이는 있음)

위와 같은 비효율적인 문제를 개선하기 위하여 데이터센터 레벨의 프로젝트를 수행하게 되었고 플랫폼 독립적인 정책 기반의 전력 관리를 위하여 현재 진행하고 있다. 현재 2월부터 6개월 시작된 프로젝트 내용을 보면 다음과 같다.

① PowerTOP for OpenSolaris: PowerTop은 시스템의 전력 관찰을 위하여 소개된 프로그램으로 2008년 9월에 등장하여 11월에 OpenSolaris에 포팅된 툴로서 v.1.1까지 소개되었으며 이 프로젝트에서는 CPU의 power management 역할로 사용된다. 이 툴을 이용하여 얼마나 많은 시간을 idle 상태에서 있는지에 대한 체크를 할 수 있다. 그리고 어떤 소프트웨어가 시스템을 idle 상태에서 깨우는지도 체크가 가능하다. (그림 4)는 PowerTop의 OpenSolaris 실행 모습이다.

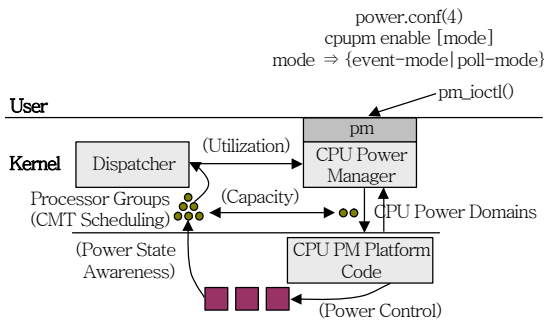
② CPU Power Management: ACPI Performance States(P-States)를 OpenSorais 커널을 통하여 DVFS를 적용하는 것으로 P-state는 추상적인 개념의 power/performance 상태를 의미한다.
- Power Aware Dispatcher: (그림 5)에서처럼 thread dispatcher는 커널에 존재하여 어떤 CPU 상에서 수행되는지 스케줄하는 기능을 제공한다. 여기에 power performance 상태를 인지하도록 한다. 동시에 CPU power management는 어떤 CPU가 idle 상태에 있는지 찾고 이를 할당하도록 한다. 이에 대한 그림은 (그림 5)와 같다.
- Deep C-State Support: x86 프로세서는 전력 소비를 줄이기 위하여 서로 다른 여러 단계의 idle 상태를 제공한다. ACPI는 이를 C-state라 정의하고 C0, C1~Cn 상태를 제공하며 이를 OS에서 제공하도록 하는 것이다.

③ Memory Power Management
- CPU Idle Notification: 위에서 말한 CPU idle 상태를 제공하기 위하여 CPU idle 상태를 메모리에게 알리기 위한 이벤트 신호를 생성한다. 이는 메모리의 내용을 유지해야 하는지 아니면 swap 해야 하는지 혹은 지워도 되는지에 대하여 각 CPU 상태와 관련이 있기 때문이다.



<자료>: http://opensolaris.org

(그림 4) PowerTop 실행모습



<자료>: http://opensolaris.org

(그림 5) Power Aware Dispatcher

④ Power Observability: 앞서 말한 PowerTop 이외에 여러 방법에 대하여 연구하고 provisioning 방법에 대하여 연구한다.

위의 단계는 현재 진행되고 있는 상황으로서 앞으로 더욱 진행될 프로젝트의 내용은 다음과 같다.

- System wide PM policy engine: 여러 플랫폼과 호환되는 policy engine을 제작하여 커널 상에 구현하도록 한다.
- Administrative facility: 관리자 중심의 정책을 구현하고 디자인 할 수 있는 툴을 제작한다.

2. Intel DCM

인텔의 경우 2008년 2월에 "Datacenter Power Management: Power Consumption Trend"[7]란 글에서 미래 데이터센터의 경향에 대하여 언급하였고, 이는 앞서 이야기한 전력관리 대상별로 다음과 같이 분석하였다. 앞서 이야기한 차이점은 데이터센터에 대한 환경적인 문제를 추가한 것이 특징이다.

① Environment Level: 에너지 전력에 관한 정부 규제 및 green datacenter 규제에 관한 획기적인 방법이 필요하다.

② Datacenter Level: 거대한 데이터센터 및 많은 modular 데이터센터의 출현은 더욱 많은 전력 공급이 필요하게 되고 이는 더욱 많은 냉각 시스템으로 새로운 데이터센터 레벨의 전력 및 냉각 시스템이 필요하다.

③ Rack Level: 랙 당 서버 밀도가 높아짐에 따라 전원의 밀집도가 커지게 되었고 주어진 공간에서 냉각 및 전원의 새로운 요구 사항이 생기기 시작하고 이는 랙 단위의 전력 및 냉각 모니터링이 필요하게 되었으며 이를 동적으로 관리하는 시스템도 요구된다.

● 용 어 해 설 ●

Power Capping: 서버, 그룹 및 데이터센터의 최고 임계치 전력을 설정하여 임계치 전력 밑으로 전력을 제한하는 방법이다.

④ Server Level: 플랫폼의 전력 소비는 플랫폼의 성능과 선형 관계에 있으므로 정책이나 부하에 따른 전력 소비를 유동적으로 제어하는 방법이 필요하다. 이에 따라 전력 및 냉각 시스템에 관한 제어 기술도 필요하게 된다.

이와 같은 개념을 바탕으로 Intel® Intelligent Power Node Manager와 Intel® Data Center Manager[7]를 만들어 중국의 China Telecom의 포털인 Baidu에서 테스트 과정을 거치고, 지난 4월 30일 공식 발표를 하였다. 위에서 이야기한 기본 요구사항을 바탕으로 랙 및 그룹 레벨의 전력 관리에 기본을 두고 설계하였다. 시스템 요구사항과 그 특징은 <표 2>와 같다.

- 다양한 종류의 데이터센터 지원
- 실시간 노드의 전원 정보, 온도 정보 모니터링 및 수집(1년까지 저장)
- 사용자가 설정한 온도, thermal 이벤트에 관한 경보 수신 및 정책에 의해 정의된 action 수행
- 사용자의 전력 임계치 설정 혹은 최소 파워 동작 기능에 따라 다중 정책 기능 제공
- 서버의 로드에 따라 유동적으로 적용되는 power capping 기능

〈표 2〉 Intel DCM의 특징

구성	요구조건
관리 대상 노드	OEM servers(Intel® Intelligent Power Node Manager 1.5, with IPMI interfaces implemented and exposed over LAN) BMC user(Intel® Data Center Manager ADMIN privilege level) Intel temperature sensors(Intel® Intelligent Power Node Manager 1.5 or IPMI 2.0 over LAN을 이용)
관리서버의 OS	Microsoft* Windows* Server 2003 32 bit and 64 bit or Red Hat* Linux EL 5.0 32 bit
관리서버의 하드웨어 사양	Install the Intel® DCM server on a system with at least: A dual-core processor of 2.6 GHz or higher 4 GB RAM , 60 GB of hard drive space Automatically installed by Intel® Data Center Manager: Sun Microsystems* Java Runtime Environment* 6 Apache* Tomcat* application server Apache Axis2* web service engine PostgreSQL 8.3 Database

<자료>: www.intel.com

- 정책으로 SLA 우선순위 설정가능
- Schedule power capping 정책들을 time of day로 스케줄링이 가능하고 데이터센터의 환경에 맞추어 관리가능(동시에 다중 정책 수행 가능)
- 관리 노드에 agent가 없음
- WSDL APIs
- 쉽게 다른 툴에 integration 가능하고 다른 서버 제품의 관리 툴 등과 함께 사용가능
- 1000개 노드까지 지원
- IPMI 2.0 BMC authentication, integrity and confidentiality code
- Power spike에 의한 자동 safeguard 기능 지원
- Schedule power capping 정책들을 time of day로 스케줄링이 가능하고 데이터센터의 환경에 맞추어 관리가능(동시에 다중 정책 수행 가능)

Baidu의 테스트는 다음과 같은 전력 관리 이슈에서 출발하였다.

① Over-allocation of Power: 실제 시스템의 전력 할당량은 실제 소모량과 큰 차이를 보인다. 이는 최악의 상황에 맞추어 항상 설정되어 있기 때문이다.

② Under-Population of rack space: 랙 공간에는 실제로 많은 여유 공간이 존재하나 서비스의 증가에 따라 요구되는 컴퓨팅 능력에 따른 시설 증가는 불가피한 실정이다.

③ Capacity Planning: 랙 레벨의 성능에 대하여 실제 부하와 전력에 관하여 측정하고 분석하기 위하여 위와 같은 테스트를 진행하였고, 성능에 대한 동적 제어나 예측과 같은 취지는 아니다.

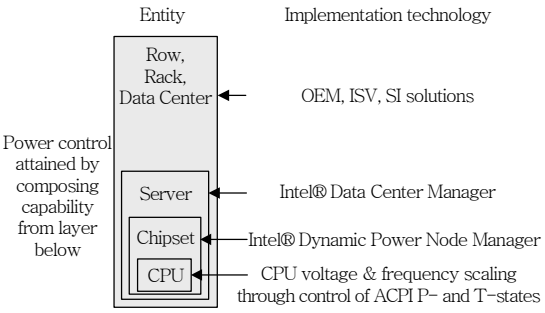
위에서 설명한 것처럼 Intel® Intelligent Power Node Manager(Node Manager)의 경우는 인텔 서버 칩셋을 이용한 OOB power management policy engine으로 프로세서의 p-state, t-state 등을 BIOS 및 OSPM으로 제어하는 구조이다. 이는 DPM을 사용하여 최고 성능 및 전력에 관한 동적 제어를 담당하는데 자세한 역할은 다음과 같다.

① 동적 전력 모니터링: +/- 10% 이내의 실제 전력소모를 측정하여 이 정보를 PSMI를 이용하여 수집하고 실시간으로 이 값을 제공한다. 이때 제공되는 방법은 IPMI를 이용한다.

② 플랫폼 Power Capping: (그림 6)과 같이 전력 대비 최대의 성능을 내는 최대 전력 값을 설정한다. 이를 설정하여 동적으로 CPU p-state를 조절한다.

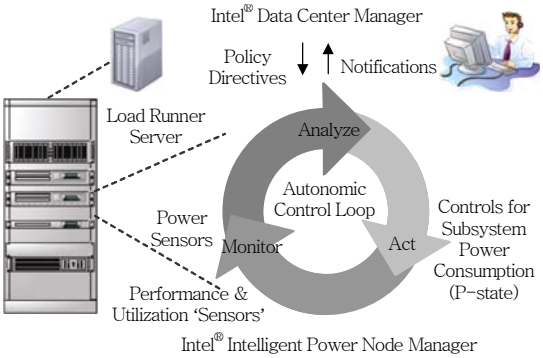
③ Power Threshold Alerting: 사용자 의도에 의한 전력 임계치를 설정하여 이를 통지하는 기능을 가진다.

(그림 6)은 인텔 DCM의 기본 구조를 나타내고, (그림 7)은 Baidu의 테스트 시스템을 나타낸다. 그림의 load runner 서버는 테스트를 위하여 랙 단위의 부하를 만들기 위하여 사용하였다.



<자료>: http://communities.intel.com

(그림 6) Intel DCM의 기본 구조



<자료>: Intelligent Power Optimization for High Server Density Racks: Intel White Paper)

(그림 7) Baidu의 테스트 구조

3. IBM Tivoli Energy Management

IBM은 2008년 5월에 전력 관리 솔루션을 기존의 Tivoli Solution에 덧붙여 출시하였다[8]. 이는 Repton이라는 회사의 third party 툴로서 제품 생산 이전에 데이터센터의 최적화를 위한 다음과 같은 이슈를 제시하였다[9].

① 각 비즈니스 서비스마다 얼마나 많은 전력이 소비되고 있고 그 리소스 타입에 대하여 분석이 필요하다.
② 리소스가 전력 사용에 있어서 주는 영향에 대하여 분석이 필요하다.
③ 전력 대 최대 성능에 대한 분석이 필요하다.
④ 데이터센터 레벨에서 전력 비용에 관한 분석이 필요하다.
⑤ 부하 스케줄링에 의한 전력 소모와 기하학적인 전력 할당에 대한 효율성에 대하여 검토하여야 한다.
⑥ 전력 절감을 위한 하드웨어 업그레이드 요소에 대하여 분석한다.
⑦ SLO, SLA에 입각한 전력 정책을 수립할 수 있어야 한다.
⑧ 전력 절감에 대한 사용자의 정책 수립에 분석 및 적용이 필요하다.

〈표 3〉 IBM Tivoli Monitoring Power Management의 사양

사양	장점
CPU 가용성 및 서비스 응답 시간	CPU Utilization을 조정하여 데이터센터에 필요한 요구량만 공급
부분별 및 전체 소요되는 전력 소모 분석 및 과금 기능	부분별 및 애플리케이션별 소비되는 전력에 관한 차등 부과 기능
다른 관리 시스템과의 인터페이스	단일 툴 및 단일 관리자를 위한 관리 기능
관리 리소스들의 단일 사용자 인터페이스	관리자에 대한 시간 및 투자 비용 단축
다른 데이터센터 및 Facility에 대한 공통데이터 인터페이스 제공	에너지 감소를 위한 최소 비용으로 최적화된 솔루션 제공
다양한 리포터 제공기능	잠정적인 전력 및 비용 절감에 대한 계획, 우선순위 및 판단기능을 제공

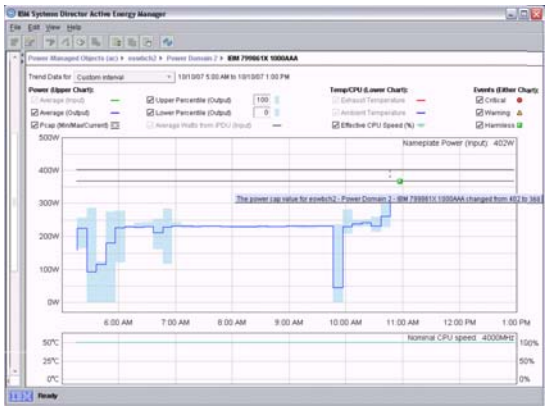
IBM의 경우 다른 업체의 관점 보다는 약간 경제적인 측면에 대하여 더욱 무게 중심을 두었다. 제품의 특징을 보면 <표 3>과 같다. IBM의 전력 관련 제품은 Repton이라는 회사의 비즈니스 파트너 형태의 솔루션으로 Tivoli와 다른 IBM 툴들과 함께 구현된 제품이다. 이는 다음과 같은 제품들이 존재한다.

가. Active Energy Manager

이 제품은 IBM® Systems Director Active Energy Manager 형태로 System Director와 함께 사용되고 다음과 같은 특징을 가진다.

① IT 환경의 전력 사용 및 thermal 분포에 대한 모니터링 및 관리 기능
② 데이터센터, 랙 및 섀시 레벨의 power capping 기능
③ 현 존재하는 데이터센터의 전력에 관한 효율적인 계획 기능
④ 실제 시스템에 맞는 전력 입력을 조절

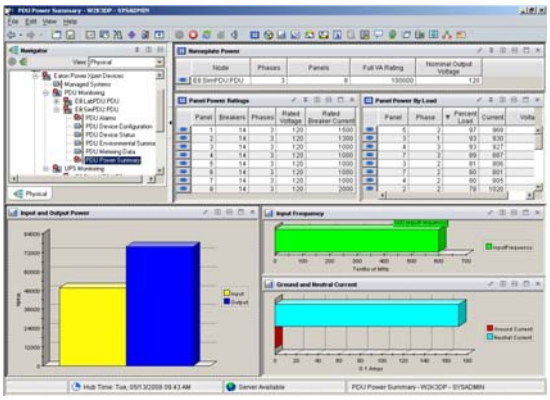
이 시스템은 PDU를 이용한 전력 모니터링 기능이 포함되어 있고 자체 IBM 시스템에 제공되고, non-IBM 시스템에도 적용이 가능하다. (그림 8)은 IBM System Direct에 포함된 Active Power Manager의 모습이다.



(그림 8) Active Energy Manager[9]

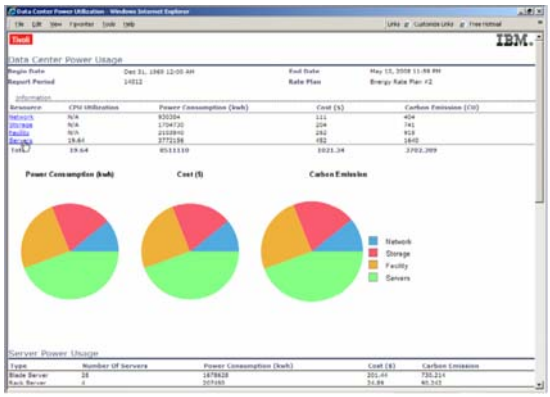
나. ITM for Energy Management

Tivoli의 제품 군에도 역시 Energy Management 툴로서 Active Energy Manger를 사용하고 ITM의 agent 형식으로 구현된다. 시스템의 전력 사용량을 모니터링 하고 개별 혹은 그룹별 시스템 온도를 모니터링 할 수 있다. 그리고 이 값들은 Tivoli Data Warehouse에 저장되고 다른 Tivoli 툴에 제공될 수 있다. 이 값들은 iPDU를 통하여 제공되는데 iPDU는 랙 단위의 전원 공급장치이며 모니터링 기능도 포함된다. (그림 9)는 랙 혹은 그룹별 전력 사용과 iPDU 모니터링 결과이다.



(그림 9) IBM PDU Monitoring[9]

다. Tivoli Data Center Optimization for Energy Management



(그림 10) Tivoli Data Center Optimization for Energy Management의 리포팅 기능[9]

Tivoli Data Center Optimization for Energy Management는 위의 기능에 리포팅 기능과 분석 기능을 포함하여 데이터센터 최적화에 사용되는 툴로서 데이터센터 전체 레벨의 전력 관리 툴이다. 이 툴은 서버뿐만 아니라 스토리지, 네트워크 및 다른 facility들에 대한 전력 사용량을 보여주고 그 리포터는 (그림 10)과 같이 나타난다.

4. Google

구글의 경우는 실제 판매되는 제품은 없으며 자신의 저전력 데이터센터 적용을 위한 분석 자료를 2007년 6월에 "Power Provisioning for a Warehouse-sized Computer"라는 제목으로 발표하였다[10]. 이 논문은 실제 구글이 제공하는 서비스의 부하를 이용하여 데이터센터 전체의 전력 사용분포를 6개월 동안 분석하여 얻은 자료이다. 물론 구글의 경우도 이런 분석이 시작된 시점은 자신들이 소유하고 있는 대규모의 데이터센터 레벨에서 전력 관리를 위한 방법을 분석하기 위하여 시작한 것이고 전력 사용이 비효율적인 것을 감안하여 데이터센터의 다음과 같은 이슈를 제기하였다.

① Staged Deployment: 데이터센터 내의 실제 facility의 설치와 실제 사용과는 차이가 있고 이는 단계적인 방법으로 새로운 facility를 사용하게 된다.

② Fragmentation: 실제 데이터센터는 매우 극대화된 전력 설비를 따른다. 즉 예로써 2.5 kW의 전력으로 실제 520 W 서버는 4대 정도 사용할 수 있는 전력이지만 데이터센터 내의 서버의 가용성은 17% 정도 밖에 되지 않는 점을 가만할 때 매우 낭비되는 전력이다.

③ Conservative Equipment Rating: 일반적으로 "Nameplate Value"라 이야기하는 표시된 전력과 실제 사용 전력에는 상당한 차이가 있다.

④ Variable Load: 실제 서버는 그 사용에 따라 다양한 전력 소모가 일어난다. 그러므로 실제 전력 provisioning이 수행되어야 하고 이에 따

른 prediction 기능도 수행되어야 한다.

⑤ Statistical Effect: 서버의 그룹 사이즈가 증가하면 실제 peak activity power level도 다르게 나타난다.

부하의 종류는 구글에서 서비스하는 WebSearch, WebMail, MapReduce이다. 이 부하를 이용하였을 경우 전력 특성을 분석하였는데 그 결과는 다음과 같다.

① 실제 최대 사용 전력과 이론적인 값의 차이는 40% - well-tuned large workload가 필요
② Power capping이 효율적인 방법
③ DVFS가 데이터센터 레벨에서 전력 절감 효과가 뛰어남
④ 전력 절감 효과가 뛰어난 activity range 존재
⑤ Mix load(3개의 부하가 한 서버에서 실행될 때) 일 때가 "평균 전력 ≈ Peak Power"

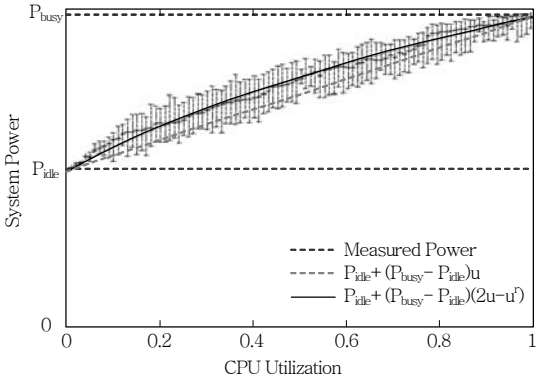
그리고 부하의존성에 관한 테스트의 경우는 <표 4>와 같다.

이는 실제 구글의 서비스 상에서 측정한 것으로서 WebSearch의 경우는 스루풋이나 큰 데이터 프로세싱의 경우 많은 전력을 사용하고 실제 시스템 상에서 주로 이슈가 되는 문제는 시간에 따른 전력 사용량의 문제(time of day issue)가 큰 비중을 차지한다. WebMail 역시 시스템 상에서는 시간에 따른 전력 사용량이 이슈이나 그 의존성은 메일 서버의 특성상 디스크 접근에 많은 의존도가 있다고 분석하였다. 그리고, 서버 전력에 관한 측정 시스템에 대하여 분석하였는데, 이는 단순한 모델을 이용하여 전력을 측정한 것이고 이 모델은 가볍고 단순한 방법이어야 하는데 그 목적을 두었다. 그래서 전력 모델은 2006년 스탠포드와 HP의 연구에서 발표된 CPU utilization 모델을 사용하여 단순화 하였고 그 결과는 (그림 11)과 같이 실제 전력과 매우 흡사하였다[11].

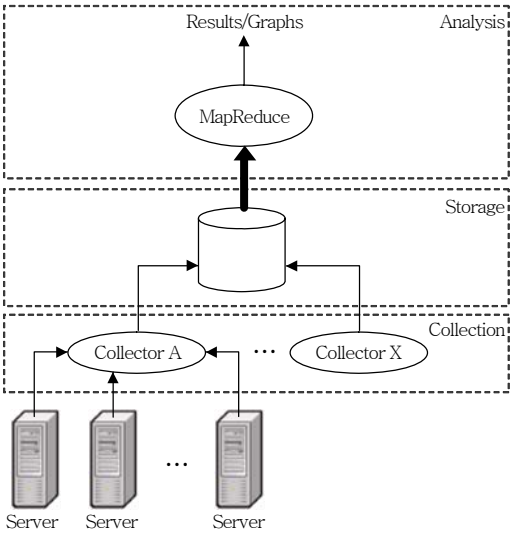
〈표 4〉 구글의 부하에 따른 의존성 분석

부하의 종류	의존성	Activity Level
WebSearch (Google)	Throughput, Large Data Processing	Time of Day Issue
WebMail (G-mail)	Disk I/O	Time of Day Issue
MapReduce	Large Offline Batchjob	User Pattern Issue

(그림 12)는 구글의 전력 관리 구조를 나타낸 것이다. 구조는 자신의 서비스 시스템과 매우 흡사하고 각 서버의 전력량과 모니터링 값을 수집하기 위한 프로그램과 이를 저장할 데이터베이스 그리고 분석에 사용되는 MapReduce 과정의 3단계 관리 구조를 지닌다. 구글에서 전력 관리를 위한 제어 방법으로 가장 효율적인 두 가지 방법을 제시하였고 이는 DVFS와 non-peak power 효율을 극대화 시키는 것이고 이것의 접근 방법으로는 idle 상태에 전력을 줄이는 방법이다. 그리고 역시 power capping 기능에 대하여서도 언급하였는데 이는 PDU 입장에



(그림 11) 구글에서 사용한 전력 모델[10]



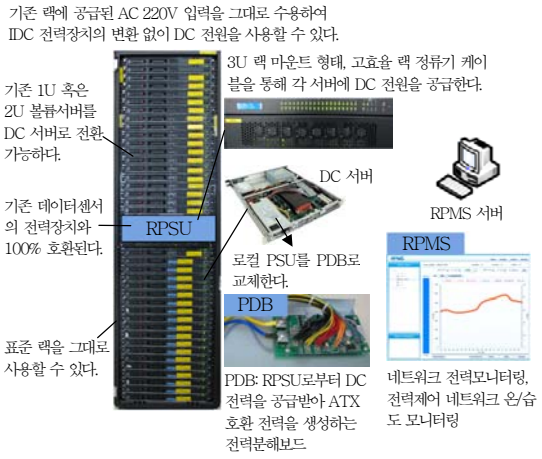
(그림 12) 구글의 전력 관리 구조[10]

서 power capping 기능을 고려한 것이다.

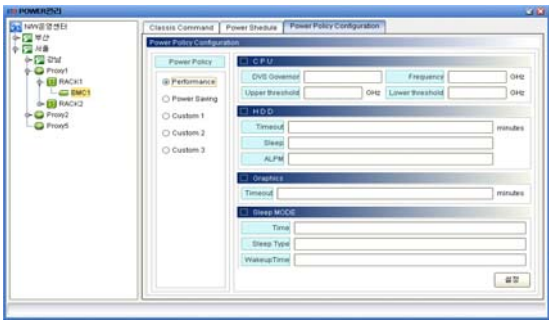
5. ETRI 전력 관리 시스템

본 연구부서에서는 GLORY 사업의 일환으로 2007년부터 저전력 플랫폼 운영에 관한 연구를 수행하고 있다. 2008년 연구 수행결과 (그림 13)과 같이 랙 단위의 DC 전류 공급 장치인 RPSU를 제작하여 전체 전력의 약 10% 정도를 절감하는 효과를 보았으며, (그림 14)와 같이 IPMI 기반의 시스템 자원 관리 시스템에 전력 절감 기능을 첨가하여 노드별 DVFS 및 ACPI S1 상태와 S4 상태를 이용하여 전력 절감 효과를 극대화 하였다.

2009년 3차년도에서는 현재 시스템을 업그레이드하여 DPM 및 power capping 기능을 추가하고 노드 단위의 DC 전원장치 개발을 진행중에 있다.



(그림 13) ETRI RPSU 및 관리 툴



(그림 14) ETRI 전력 관리 v1.0

Ⅳ. 그린 데이터센터의 방향

지금까지 데이터센터 내의 IT와 관련된 전력 절감 기술에 대하여 업계 동향에 대하여 알아보았다. 위의 업계들의 동향을 종합하여 보면 다음과 같다.

① 시기적으로 2~3년 전부터 본격적인 연구가 시작되었다.
② 전력 관리 대상으로는 서버에서 랙 그리고 데이터센터 레벨로 확장되고 있다.
③ 전력 관리 방법은 정적 방법에서 동적 방법으로 변하고 있다.
④ 전력 관리를 위한 도구는 하드웨어에서 소프트웨어로 변모하고 있다.
⑤ 동적 방법에 따른 전력 예측 기능을 포함하는 경우가 많다.
⑥ 전력 예측 및 분석은 주로 CPU, 메모리의 사용량 정도가 대부분이다.
⑦ 제어 기능으로서는 주로 DVFS, power capping 방법을 많이 사용한다.
⑧ 전력 제어 인터페이스로는 ACPI를 사용하거나 자체 하드웨어 제작인 경우 해당 드라이버를 사용한다.
⑨ 현재 정책으로는 time-based policy가 가장 많이 사용된다.

위의 내용을 살펴본 바와 같이 많은 업체들이 현재 그린 데이터센터를 위한 전력 절감 방법을 연구중에 있다. 그러나 이런 연구들은 현재 여러 방면으로 문제점을 안고 있다. 그 문제점을 점검하여 보면 다음과 같다.

① 데이터센터 내 이기종 서버들의 호환 문제 – 데이터센터 내의 서로 다른 서버들이 존재하고 서로 다른 기능들이 존재하여 공통의 인터페이스를 가지기가 힘들다. 그러나 현재 제공되고 있는 표준 인터페이스를 갖춘 제품들이 출시되고 있는 실정이다.
② 전력 절감 효과의 회의적인 반응 – 실제 IT 장비들과 서버의 전력 절감 효과에 대한 회의감

이 팽배해 있다. 이는 세계적으로 많은 노력을 기울이고 언론에 노출된 긍정적 데이터들로 인해 인식 전환이 이루어지고 있다.

③ 도입시 안정성 문제 – 많은 데이터센터의 고민이 도입시 현재 서비스 상태를 유지하면서 빠르게 적응하여 효과를 볼 수 있는 단순한 시스템을 원하고 있다. 많은 업체들이 자신들 솔루션에 첨가된 단순한 방법을 취하고 있으며 공개 소프트웨어 형태의 솔루션들도 속속 나오고 있는 상황이다.

④ 도입 비용 문제 – 전력 절감을 위한 저렴하고 단순한 시스템을 원하고 있다. 이 문제 역시 많은 메이저 업체들의 공개 소프트웨어 형태로 발전한다는 것이 긍정적 영향을 줄 수 있다.

⑤ 업체 독립적인 개발 – 각 툴이나 하드웨어들은 업체들마다 자체 솔루션을 가지고 있고 이를 데이터센터 내에 적용하는 것은 한계가 있다. 그러나 자체 솔루션을 지원하면서도 표준 인터페이스에도 역시 관심을 두고 있는 상황이다.

⑥ 종합적인 facility 관리 – IT 분야 이외에 non-IT 분야의 전력 절감 효과도 간과할 수 없는 부분이므로 이에 대한 종합적인 asset management가 필요한 시점이다.

⑦ New technology에 대한 대비 – 앞으로 데이터센터 내 새로운 트렌드로 부상할 다양한 솔루션(예로 가상화 솔루션) 등에도 적응 가능하여야 한다.

위의 문제는 데이터센터가 앞으로 나아가야 할 방향의 관점에서 본다면 필수적으로 해결해야 할 문제이고 앞으로 데이터센터가 나아가야 할 방향에 속한다. 현재 위의 문제를 시기적으로 이르다고 판단하는 사람들이 많다. 이것은 현재 대규모, 소형 등 산재해 있는 데이터센터를 쉽게 바꾸기는 어렵다는 견해에서 출발하는 것이다. 현재 향후 몇 년 간의 솔루션은 새롭게 지어지는 데이터센터에 초점을 두면서 현재 데이터센터에 쉽게 접근 가능한 간단한 솔루션 형태가 나타날 것으로 사료된다.

Ⅴ. 결론

지금까지 데이터센터 내 IT와 관련된 전력 관리에 대하여 그 방법과 문제점에 대하여 알아 보았다. 전력 관리를 위한 방법은 아주 다양하다. 이렇게 다양한 관리 방법들은 전력 관리에 있어서 문제점을 주기도 하지만 다양한 노력들과 더불어 그 효율성을 높여가고 있고, 시기적으로 많은 해결책을 보이고 있다. 세계적인 추세에 발맞춘 인식의 전환과 이로 인한 표준화 활동 및 솔루션 개발의 3박자가 갖추어지고 있는 실정이어서 데이터센터의 효율성은 증대될 것으로 생각된다.

약어 정리

ACPI	Advanced Configuration and Power Interface
DCM	Data Center Management
DPM	Dynamic Power Management
DVFS	Dynamic Voltage Frequency Scaling
IPMI	Intelligent Platform Management Interface
ITM	IBM Tivoli Monitoring
OOB	Out-of-Band
OS	Operating System
PM	Power Management
PMU	Performance Monitoring Unit
PSMI	Power Supply Management Interface
QoS	Quality of Service
SLA	Service Level Agreement
SOA	Service Oriented Architecture
TCO	Total Cost of Ownership
WSDL	Web Services Description Language

참 고 문 헌

[1] http://download.intel.com/technology/itj/2003/volume07issue02/art04_validation/vol7iss2_art04.pdf

[2] Low Power Methodology Manual for System-on-Chip Design, Springer US, 2007.

[3] Ravi K. Venkatesan Ahmed, S. AlZawawi, Krishnan Sivasubramanian, and Eric Rotenberg, "ZettaRAM: A Power-Scalable DRAM Alternative through Charge-Voltage Decoupling," *Computers, IEEE Transactions on.,* Vol.56, Issue 2, Feb. 2007, pp.147-160.
[4] http://www.samsung.com/global/business/semicondutor/products/dram/Products_Graphics-Memory.html
[5] Wissam Chedid, Chansu Yu, and Ben Lee, "Power Analysis and Optimization Techniques for Energy Efficient Computer Systems," *Advances in Computers 63*, 2005, pp.130-165.
[6] www.opensolaris.org/os/project/tesla/
[7] http://communities.intel.com/community/openportit/server/blog/2008/02/20/datacenter-power-management-power-consumption-trend
[8] http://www01.ibm.com/software/tivoli/products/monitor-energy-management/
[9] http://www.repton.co.uk/library/ibm_energy_monitoring_management_solution.pdf
[10] Xiaobo Fan and Wolf-Dietrich Weber, "Power Provisioning for a Warehouse-sized Computer," *In Proc. of the ACM Int'l Symp. on Computer Architecture,* San Diego, CA, June 2007.
[11] D. Economou, S. Rivoire, C. Kozyrakis, and P. Ranganathan, "Full-System Power Analysis and Modeling for Server Environments," Workshop on Modeling, Benchmarking, and Simulation(MoBS), June 18, 2006.