

Dynamic Power Management for Cluster System¹

Dae Won Kim, Sun Wook Kim and Seong-Woon Kim
Server Platform Research Team, Electronics and Telecommunications Research Institute
Daejeon, 305-700, Korea
{won22, swkim99 and ksw}@etri.re.kr

Abstract — Many power management solutions are adapted for Cluster system. But the solutions are quite complicated and hard to adapt it for the system. In this paper, the simple algorithm and policy for the user in cluster system are introduced. This policy is based on system load and time and operates dynamically according to the policy. The controllable items are CPU frequency and system state from ACPI because the common servers and computers have the function of DVFS and ACPI. The solution consists of three tier system level: the managed node, proxy master and master node. This system has 10% to 50% power saving results depending on system status.

Keywords — Power management, Power Saving, DVFS, ACPI, cluster system.

1. Introduction

In power management system of cluster system, techniques for low power are developed for decades. Furthermore, Green IT becomes new interest in the world and government, academy and enterprise also have concerns for new industry.

From now on, power management issues are focused on the hardware issues. From chip manufacture (ASIC issues) to system design, many applications are studied and implemented. Gartner analysis showed the important of the power management in Gartner's cool vender 2008, power management issues is being magnified as a new controversy in energy saving of cluster system.[1] Although these vendors currently have no direct competition for this product, it will face competition from vendors of Software for job scheduling and virtualization, Services for facilities management and IT operations management, Server hardware, Power distribution units for data centers, Specialist niche products.

The portal sites have the large scale cluster system that collects the large information and provide the information to users through internet, and the size of the cluster system is increasing with the increase of the internet information. It is guessed that Google, one of the largest portal sites, has the cluster system with about 450K servers [2]. Google is currently developing two data centers located in the town of Dallas and Oregon. Each Korean portal site has about 10K servers.

CPU is the most power consuming component in the computer system. Thus, to control CPU power consumption is one of the most efficient power-saving mechanisms. Also, the ACPI support all servers and computers in OS easily. CPU

power management and ACPI status change mechanisms are already implemented in OS and some software are existed. But these applications are not user friendly and not applicable in active system environment. This paper presents the CPU power management and ACPI status change mechanism for new policy controllable point. Chapter 2 shows the definition of the dynamic power management and basic theory. Also In chapter 3 shows the our implementation issues and chapter 4 represent the following results

2. PM (Power Management) Issues

To apply the power management in cluster system, the power management should be necessary with these items generally.

1. Monitor the power consumption
2. Estimate the power consumption
3. Control the power consumption

Monitoring issue is very important in the power management and very hard to measure without power meter or other hardware sensors. It's quite big cost challenge for server power monitoring. Estimating power is also difficult to apply to the different types of servers. Some solution has the model for the server and computer component. But it's not easy to be adapted. And controlling server is problem on active machine because the cluster systems are the round-the-clock systems. After the needs of power management, the object should be selected. The object means the managed object.

1. Single System level
2. Multi System level
3. Rack level
4. Data Center level

Above objects are categorized by the number of the managed node. The system for power management should know how many nodes are managed.

And last, we select the way to manage the servers and computers. Figure 1 shows the categorized way of power management. Static and Dynamic power management is categorized extremely. Static power management is user

¹ This work was supported by the IT R&D program of MKE/KEIT. [10040696, A Development of Energy Saving Green Server under 1mW/MIPS for cloud service]

setting and controlling method and dynamic power management is machine setting method.

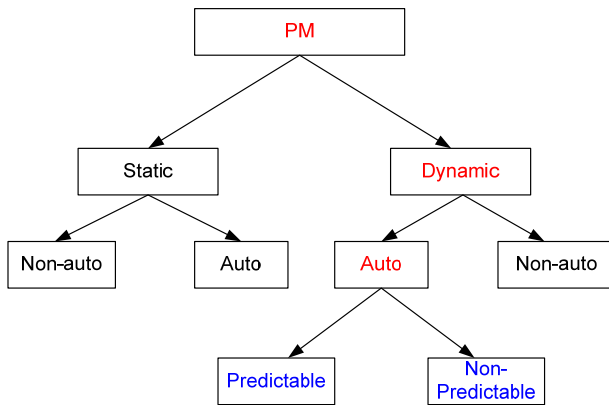


Figure 1. The method of Power management

All PM issues belongs to the three categorized issues and our implemented system reflected above three issues

3. PM Implementation

① Monitoring, Estimating and Controlling power

One computer system consists of CPU, memory, hard disk and etc. and power consumption of each component is different. In all computer components, CPU that consumes 30~40% power of one computer system is the most power consuming component. Thus, the power consumption of the computer system can be controlled effectively by managing the CPU power. The DVS has been implemented on x86-based CPU of Intel and AMD, and Intel and AMD calls this technology as SpeedStep/Demand-Based Switching (DBS) [3] [5] and PowerNow [4] [6] respectively. In DVS, frequency and voltage of each CPU can be controlled dynamically without rebooting of computer system. The CPU specific drivers are device drivers for various CPUs supporting DBS and PowerNow. The in-kernel governors are built-in kernel governors controlling CPU frequency. These governors' basic algorithm was designed by kernel developers and it cannot be controlled by users. Users that want to use its own control algorithm can adjust the CPU frequency using user-level governors.

The support of the Linux for DVS is performed in single node scope and the Linux doesn't support cluster-wide scope. Thus, this paper presents the CPU power management mechanism for the cluster systems by utilizing the DVS.

ACPI (Advanced Configuration and Power Interface) is known as the standard interface with system level. These interfaces are represented in Linux and window system.

Also monitoring information are related with the controllable point In these implementation, we select the cpu, memory and hard disk load information by below equation (1).

$$P_{server}(U_{comp}, t) = P_{start}, t + \alpha_{cpu} \cdot U_{cpu}, t + \alpha_{mem} \cdot U_{mem}, t + \alpha_{hdd} \cdot U_{hdd}, t$$

(1)#

In equation (1), Total power can be assumed by cpu, memory and hard disk usage.

② The number of Managed Node

The number of the managed node is variable. Implemented power management system can be selects from single system to data center level for user convenience.

③ The Method of power management

Implemented system has the static and dynamic power management. Static power management contains the DVFS and ACPI control simply when user needed. Also dynamic power management is load-based DPM (Dynamic power management) and Time-based DPM.

④ System Architecture

Figure 2 shows our implemented system architecture. Structure of the PMM is shows in figure 2 and it consists of power manager, operating system and hardware.

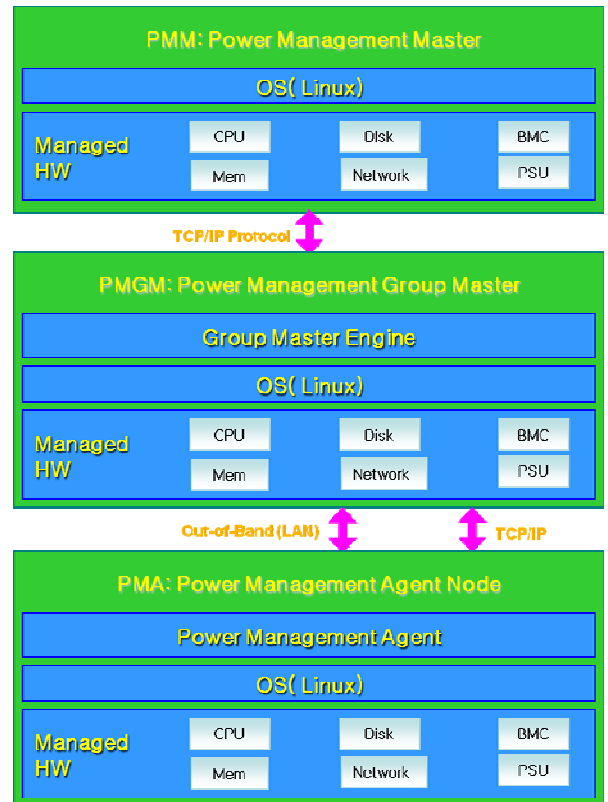


Figure 2. System Architecture

The power management architecture presented in this paper consists of one PMM (Power Management Master), multiple PMGM (Power Management Group Master) and multiple PMA (Power Management Agent) which is managed nodes. The PMM that manages the power of the cluster system sends commands related to power management to the managed nodes through the PMGM. The power management commands include power mode changes of the managed nodes, various parameters necessary to changes of the CPU frequency and ACPI state. The managed node changes its

CPU frequency according to the commands and parameters issued by the PMM. PMGM has the role of transferring the data from PMM to PMA and lessens the burden of PMM for a huge data connection and traffics.

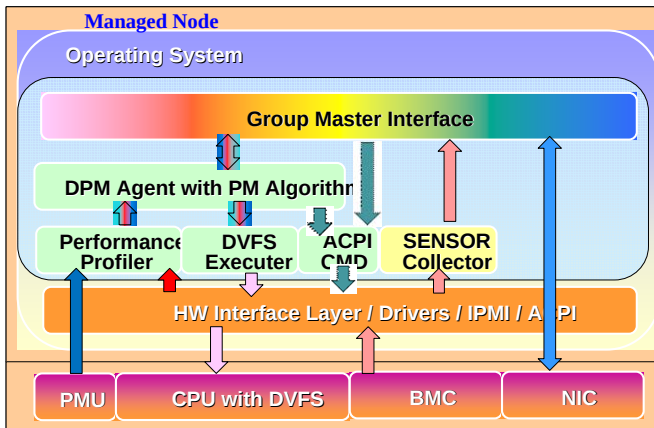


Figure 3. System Architecture

Figure 3 shows the agent architecture. It consists of performance profile, DVFS Executer, ACPI Command interface and sensor collector. Sensor collector gathers the information data from server and computer. And performance profiler calculates the server node to apply controllable point. The managed nodes have simple power agent that changes the CPU frequency and ACPI Status according to the parameters and the agent impose little load to the managed node for the power management. The power agent controls its CPU's frequency and using algorithm

- ACPI Status and CPU DVFS

The CPU of the managed nodes always works at maximum clock frequency when the power management is off. It is not controllable point. But single or datacenter level is applied in static power management condition

- Time-based DPM

When the dynamic power management is on, the power agent checks time parameters. If the current time belongs to maximum clock frequency period, the CPU frequency is set to maximum. Also, other conditions are checked and applied conditions.

- Load -Based DPM

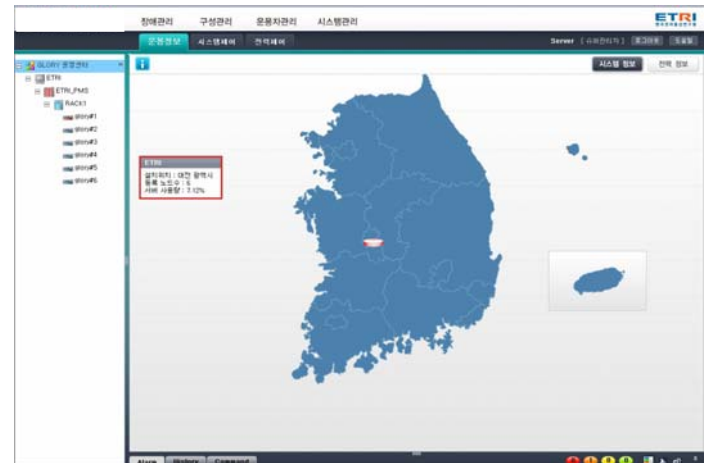
In this mode, the power agent checks the load of node periodically. If the current load is over the load upper threshold, this agent changes immediately the CPU frequency to maximum. If the current load is under the load lower threshold, the CPU frequency is changed to next lower frequency. Load value is calculated by equation (1).

Solutions are implemented in Linux server and it is applied to linux and window OS.

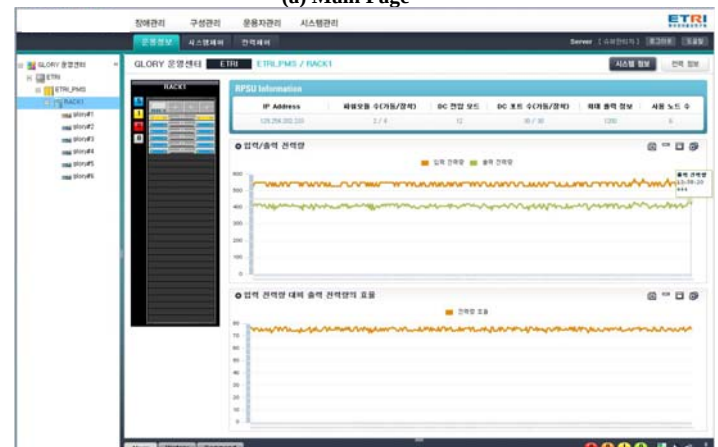
4. PM Results

Figure 4 shows GUI of our system. The main functions of GUI consists basically of

- Show the collected power information to user
- Manage the managed nodes into single node to data center
- Control the power state by setting the parameters
- Generation, update and modification of time/load profile



(a) Main Page



(b) Power Information in Group



(c) Managed Node Monitoring
Figure 4. Power Management GUI

This system is applied to our cluster system with 6 nodes. Each rack has power meter to measure the power
In our System, LDPM is applied to 6 example node. Before the application, Total power is 444W and after the application 396W with 10% power saving as shown in Figure 5



Figure 5. LDPM Application

Also, we applied to LDPM with ACPI in Figure 6. After ACPI status changing is applied, 216W power saving for 6 node group.

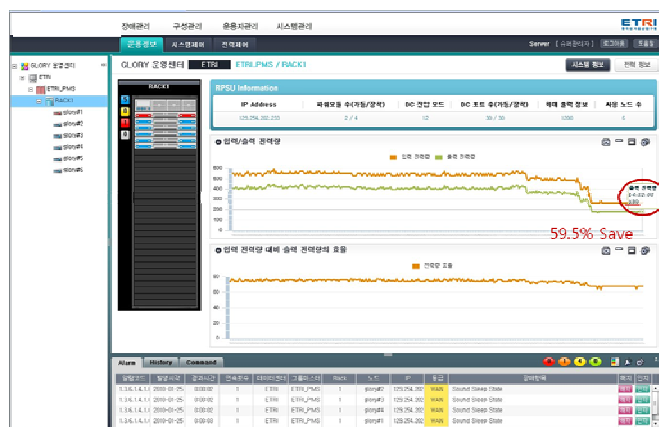


Figure 5. LDPM Application with ACPI

6. Conclusions

This paper proposed the dynamic power management mechanism that reduces the power consumption in the cluster system by controlling the CPU frequency and ACPI State remotely. The power management architecture presented in this paper consists of one PMM and multiple PMGM and multiple managed nodes. The PMM sends power management commands to the PMA and the PMA changes its CPU frequency or ACPI State according to the commands and parameters issued by the PMM. We are now implementing this proposed mechanism for the cluster system. Also, when completion of implementation, we will measure power saving effect of this mechanism and perform parameter tuning.

REFERENCES

- [1] Gartner Analysis (Gartner' s Coolvender 2008)
- [2] "Google Platform", Wikipedia, http://en.wikipedia.org/wiki/Google_platform#_note-google_arch.
- [3] "SpeedStep", Wikipedia, <http://en.wikipedia.org/wiki/SpeedStep>.
- [4] "AMD PowerNow Technology", AMD.
- [5] Enhanced Intel SpeedStep Technology and Demand-Based Switching on Linux, <http://www3.intel.com/cd/ids/developer/asmo-na/eng/195910.htm?page=1>
- [6] Howto PowerNow, http://gentoo-wiki.com/HOWTO_PowerNow!