Hardware leverages for energy reduction in large scale distributed systems

Editors: *Georges Da Costa and Davide Careglio*
Authors: *Davide Careglio, Georges Da Costa, Ronen I. Kat Avi Mendelson, Jean-Marc Pierson, Yiannakis Sazeides*

Technical Report : IRIT/RT–2010-2–FR
Cost Action IC0804

Chapter 1

Introduction

1.1 Motivation for energy aware distributed computing

It is analysed that the 1.5 billion computers in the world use about 90000 MW of electric power, which is about 10% of the global consumption. The latest in-depth survey ("Electricity Consumption and Efficiency Trends in the Enlarged European Union", Institute for Environment and Sustainability, 2007) commissioned by the European Union in 2006 about energy consumption and efficiency of equipment in buildings shows a continuous growth of energy consumption of computer end-use equipments (amongst others) over the last years. Additionally, the world energy consumption for servers has doubled over the period from 2000 to 2005. At the CeBit forum 2008 in Hannover, the worlds largest technology fair, recent shocking estimates proclaim that worldwide Internet usage needs the equivalent of 14 power stations to power the required computers and servers, producing the same amount of carbon emissions as the entire airline industry. As for an example, an operational Grid like EGEE (Enabling Grid for E-sciEnce) is constituted from more than 41,000 computer nodes, distributed in 45 countries and 240 sites. The world top 500 most powerful machines have more than 1.2 millions processors. To the raw electric consumption, one can add the energy costs in terms of air conditioning and cooling infrastructures.

The large-scale distributed systems (clusters, grids, clouds, P2P systems) are nowadays gathering transparently more and more resources to store, to compute and to communicate data and services around the world for the common benefit of many users. Cost-effective solutions are defined in terms of euros per solution. Distributed Computing, Grid Computing and most recently Cloud Computing attempt to ameliorate the personal cost of ownership by straddling ownership boundaries and taking advantage of economies of scale.

Traditionally, there has been a dearth of eco-awareness in the computing industry. Moores Law has not led to overall power savings as miniaturization would allow. Instead, greater capacity and capability have invariably taken

precedence over eco-concerns. Despite the fact that the energy dimension was taken into account since several years in mobile and embedded systems, the total collective costs of large-scale distributed technologies have not traditionally prioritized ecological concerns. Ecological impacts constitute a silent cost which until recently has largely been ignored. But recent studies charged from government institution are going to consider energy efficiency procedures mainly for server and data centres. An example is the report of the US Environment Protection Agency (EPA) delivered on August 2007 to the congress.

This Action aims at giving voice to this cost. Central to this approach is the recognition that environmental resources need to be effectively managed as an integral part of every computation - just as processing cycles, storage and bandwidth are currently routinely managed in every computation. The research topic of the Action is the investigation of realistic energy-efficient alternate solutions (software, middleware, networking) to share distributed resources.

1.2 This report

This reports provides summary and references on existing and future leverages to adapt the underlying hardware infrastructures of large scale distributed computing systems in order to decrease their energy consumption.

This report is intended to be used by other Working Groups of this action to drive their research fostering activities towards energy aware middleware for large scale distributed computing systems. This report can also be useful for other audiences: ex. academics, researchers, companies, general public, datacenter administrators, politicians,...

The report consists of several chapters each addressing one hardware resource, such as the processor, main memory, storage (disk and flash), motherboard, fan, network interface etc. The report, also includes a chapter discussing/presenting existing energy aware practices in large scale systems.

The above breakdown of resources is based on recent [Fan2007] studies by Google that show the following breakdown of power in a server with a local disk:

CPU	37%
Memory	17%
PCI slots	23%
Motherboard	12%
Disk	6%
Fan	5%

Another study performed by Lim et al. [Lim2008] obtains results similar to the above.

Chapter 2

Processor

2.1 Context

Power related issues consider as one of the most important aspects of designing modern processors since it affects many design aspects of the entire system. When discussing power issues, we need to consider different aspects of the problem:

- Energy consumption how much energy (power over time) the system consumes when execute a piece of work (workload). This parameter mainly affects battery life of mobile systems and the cost of operation of other systems such as servers, data centers and cloud computers.
- Power consumption how much power the processor (or the system) consumes at a certain point of time. This parameter affects the power delivery subsystem and in many cases, the cooling system, Since the die must not exceed max temperature due to chemical and reliability limitation.
- Power density the distribution of power consumption to different subsystems. This parameter has a significant impact of the internal design and the max temperature of certain parts of the system. This aspect is out of the scope of this presentation.
- Dynamic power vs. leakage power; dynamic power is defined as power the system consumes while working and leakage power is defined as a power the system, or sub-system consumes while not doing any active work. Leakage power starts to be very significant in modern architectures. Many techniques proposed to reduce leakage power, all of them require significant latency when moving from sleep mode to active mode.

The processor is very sensitive to each of these aspects of power management; in many systems the processor consumes a significant part of the entire systems energy consumption, the cost of cooling can be very significant.

But on the top of it, the amount of heat the processor produced, is proportional to the power it consumes and the power consumption (P) depends on its frequency (f) and voltage (V) since $P = \alpha * f * V^2$, thus in order to prevent the system from overheating, we may need to reduce its frequency and so to lose performance.

The understanding that the power consumption of the processor impacts the direct and indirect cost of the system and its performance cause power to be first class citizen in any modern design. But controlling power true Hardware only mechanisms was found not to be optimal, so many HW/SW techniques were developed such as: AMDs PowerNow! and Intels SpeedStep, that help to control the Voltage and the Frequency of the processor as a function of the workload being executed. But the most common techniques to control the different aspects of the power related issues in processors and the entire system, was developed as a consortium between Intel, Microsoft, HP Phoenix and Toshiba and is called Advanced Configuration and Power Interface (ACPI).

2.2 ACPI

The ACPI specification [ACPI] is quite complicated and contains 700 hundreds of pages that cover many SW and HW related issues. In this report we will focus only on the main features that impact the operation modes of the CPU (processor) and include three mechanisms, termed Thermal control Zone, Power state (P-state) and CPUs state (C-state).

2.2.1 Thermal control Zone

ACPI defines a set of events to prevent the system form getting over-heated. These events include Trip_points which are dynamically defined events that indicate to the OS to change the speed of the fan or to change the speed of the CPU, and a Critical Shutdown event that that indicates that the system MUST shut down immediately to prevent damages [Alon2004].

2.2.2 C-States

They control how deep the CPU sleeps when is not active. The deeper the CPU goes, less leakage power it consumes but it also significantly increases the latency of the system to wake-up. Thus the ACPI defines 4 states (as we will see later HW companies extend it)

- C0 is the operating state, it consumes dynamic power
- C1 (Halt) is the state where the processor is not executing anything but can come back at C0 in a few cycles (but the saving is minimal)
- C2 (Stop-Clock) is a deeper sleep state that consumes less leakage power than C1 at the cost of a slower wake-up (this state is optional and usually not implemted)

- C3 (Sleep) offers improved power savings over the C1 and C2 states. The worst-case wakeup latency for this state is provided via the ACPI system firmware and the operating software can use this information to determine when the C3 can be used and when higher states must be used to guarantee critical response time

All modern processors extend the notion of C3 to further refinements. For example, I7 (Intel) implements the notion of C6 state. But from the ACPI point of view (SW/HW interfaces) only 3 of them exist and the rest are handled y HW only.

2.2.3 P-States

P-States indicate how fast the processor should run when in C0 state. System can define a table of frequency/voltage operational points and OS/SW can define at what operational work the system will work

- P_0 is the max frequency/voltage state
- P_1 is a state where frequency and voltage are reduced
- P_n is a state where frequency and voltage are reduced compared to P_{n-1}

The way OS handled P states is by applying dynamic learning algorithms, It sample the system every period of time (usually 100MS) and determine the utilization of the system during that period of time. If found that the system was busy most of the time, it reduces its P state (faster) and if found that the CPU was at sleep state most of the time, it increases its P-state (slower). By doing that the systems tries to optimize between the power the system consumes to the performance it can get. Please note that T states are independent of P-State and may impact the absolute frequency the system can run at P_0.

2.3 Vendors

2.3.1 AMD

Recent AMD processors use PowerNow! and Cool'N'Quiet technology. Those two technology provide means to change processor frequency and voltage. PowerNow! is aimed at laptops, and Cool'N'Quiet is aimed at desktop and servers.

AMD announced AMD Turbo Core technology [Llano, Llano2]. AMD Turbo CORE aims at optimizing use of multi-core by tuning frequency when some cores are idle. It has an aim of a certain maximum power, and when power consumption is lower than a certain threshold, it boosts the frequency of one core while reducing idle cores frequency by a few hundred MHz.

2.3.2 Intel

Recently much information was published [PmI7] regarding the power management of the new I7 processors family. This section extends the discussion on some of these features to provide a better picture on how power is managed in modern cores.

The implementation of the Throttling mechanism in I7 is not well document, so we will based the discussion on the implementation of CoreDue-2 as appeared in [Alon2006]. Here two mechanisms were discussed, the two levels mechanism and the dynamic throttling mechanism.

The two levels mechanism (implemented in P4 family or processors) defines two operations points normal and halt. While in normal mode the system works corresponding to the Pn state. When the system reaches Max-temp trip point (this is usually at 90C or 100C) it indicates the OS to change the state to halt. While in Halt state, the system waits until cool down below a specified point and change to normal again. If for any reason the system reaches the critical-shutdown point, the HW shut the system down immediately to prevent any damages.

The more sophisticated mechanism descried there is a dynamic throttling, here, at any operational point, the system defines two trip points, upper and lower. When the temperature cross the upper trip point, an HW mechanism force the system to slow down (redefine the values of the P-state table) and when the temperature cross the low-trip point, it allow the system to work faster (limited by max frequency.

I7 Core extends the implementation of the C states and defines new state termed C6. (since it is not exposed to the ACPI it is purely handled by HW.).

The states I7 Core implements are:

- C0: when the microprocessor is in the active state (some P-State)
- C1: no instructions are being executed; controller clock-gate all gates pertaining to the core pipeline. Clock gating is accomplished by logically ANDing the clock signal of a particular clock domain with a conditional control signal
- C3: the core phase-locked-loops (PLLs) are turned off, and all the core caches are flushed. A core in C3 is considered an inactive core. The time it takes to the core to wake up is significantly longer that C1 since the PLLs are linear-feedback based control systems, which need to be turned back on, time must be allocated for the PLLs to lock (stabilize) to the correct frequency return to full speed. More than that, the time it take the system to return to full utilization is even longer, due to the cold start of the cache.
- C6: the most power efficient state, the core PLLs are turned off, the core caches are flushed and the core state is saved to the Last Level Cache (LLC). The power gate transistors activated to reduce leakage power consumption to a particular core to near to zero Watts. A core in idle state

C6 is considered an inactive core. The wakeup time a core in idle state C6 is the longest since the core state must be restored from the LLC, the core PLLs must be re-locked, the power gates must be deactivated, and the caches starts from clean state.

Please note that waking up from C6 may consume significant power, so the system need to make sure that the power it saves is greater than the power it waste for entering and exiting from the state. To prevent this, i7 Core, includes an auto-demote capability that uses intelligent heuristics to optimize the use of this aggressive state.

I7 Core also presents a revolutionary approach on how to manage P states and the overall thermal budget of the core. Intel discovered [Gelsi2008] that when only a single core is active it never use the entire power and thermal envelop allowed for the 4 CPU die. Thus they introduce the notion of Intel Turbo Boost Technology that allows a core to increase its frequency above the maximum allowed frequency (the official frequency of the core) if thermal and power head rooms allows it.

2.3.3 Apple

Even if Apple is not a chip maker, it is currently the only one to provide a tool for it current OS (MacOSX) for stopping a core on a multi-core processor.

Chapter 3

Memory (DRAM)

3.1 Context

According to [FAN2007][Lim2008], the memory is one of most consuming device of a computer, counting for 17% of energy consumption in a typical server with a local disk. Moreover, CPU and memory are the main contributors to the dynamic power, while other components have very small dynamic range.

In this chapter, we only focus on DRAM memory.

3.2 Power consumption

To know the power consumption of a DRAM, it is necessary to understand the basic functionality of the device. The master operation of the DRAM is controlled by clock enable (CKE). If CKE is LOW, the input buffers are turned off. To allow the DRAM to receive commands, CKE must be HIGH, thus enabling the input buffers and propagates the command/address into the logic/decoders on the DRAM.

During normal operation, the first command sent to the DRAM is typically an active (ACT) command. This command selects a bank and row address. For every ACT command, there is a corresponding precharge (PRE) command. The ACT command opens a row, and the PRE closes the row.

In the active state, the DRAM device can perform READs and WRITEs.

Background Power During normal operation, the DRAM always consumes background power. When CKE is LOW, most inputs are disabled. This is the lowest power state in which the device can operate. When CKE goes HIGH, commands start propagating through the DRAM command decoders, and the activity increases the power consumption.

Activate Power To allow a DRAM to READ or WRITE data, a bank and row must first be selected using an ACT command. Following an ACT command, the device uses a significant amount of current to decode the

command/address and then transfer the data from the DRAM array to the sense amplifiers. When this is complete, the DRAM is maintained in an active state until a PRE command is issued.

Write and Read Power After a bank is open, data can be either read from or written to the DRAM. The two cases are similar

I/O Termination Power This is the power consumed by the output driver or ondie termination.

Refresh Power Refresh is the final power component that must be calculated for the device to retain data integrity. DDR3 memory cells store data information in small capacitors that lose their charge over time and must be recharged. The process of recharging these cells is called refresh.

3.3 Energy efficiency techniques

Current solutions are mainly based on lowering the voltage of DRAM which can surprisingly reduce the power use of the CPU-memory subsystem quite significantly.

Other solutions exploit the multiple power states such as active, standby, nap and power-down of the DRAM manufacturers. The chip must be in the active state to service a request. The remaining states are in order of decreasing power consumption but increasing time to transition back to active. Energy efficiency can be improved by placing the chips in a lower power state when not used. The challenge is to understand the characteristics of memory access patterns in a cache-based memory architecture and how those patterns affect the design of power-management controller policies to control the transition among power states.

Other research solutions aim at reducing the refreshing time to lowering the energy consumption.

3.4 Vendors

Kingston Kingston recently dropped its latest 'LoVo' (low voltage) HyperX DDR3 High-Performance memory product line that will run at anything down to 1.25V at 1,333MHz, or even 1,866MHz at 1.35V with its built-in XMP profile. The flagship product, running an ultra-low 1.25 volts at 1600MHz, is the lowest voltage to date for desktop PCs.

Micron Microns energy-efficient Aspen Memory product line features 1.5V DDR2 and 1.35V DDR3 reduced chip count (RCC) modules, specifically designed to lower data center server power consumption.

Micron claims that when the 1.5V modules, for example, are implemented into data center server systems in place of 1.8V solutions, the reduced voltage cuts power consumption by 16

The reduced chip count also factors into overall savings. RCC FBDIMMs deliver the same performance and memory capacity with half the number of components. And because there are fewer modules, less heat is generated so cooling costs are lower.

Chapter 4

Disk/Flash

Disk drives are the primary storage medium in todays storage systems. The disk mechanical design is the dominating factor in its energy consumption. In order to be able to quickly serve I/O requests, the disk platters must always be spinning. The disk controls the platters spin and maintains a communication channel with the host. These two factors are the main contributors to the constant portion of the disk energy consumption, which amounts to about 2/3 of the total energy consumption under load.

Disk drive technology allows three main control knobs:

- Spindle speed
- Seek speed
- Disk power mode

4.1 Spindle speed

The energy consumption of the spindle motor is quadratic to the platters RPM (Revolutions Per Minute). Therefore, a reduction in RPM has a dramatic effect on the energy consumption. The term DRPM (Dynamic RPM) [Carrera2003, Gurumurthi2003Sivasubramaniam, Li2004, Pinheiro2004] refers to the ability to vary the spindle RPM which allows the disk to serve I/O requests at different RPMs and data transfer rates. Unfortunately manufacturing disks that support DRPM is not easy and there are no available disk drives that support DRPM.

However, allowing the disk to reduce its RPM during idle is easy, as the disk head can be parked outside the platters during this idle time. Moving the heads outside the platter is required before slowing down the spindle RPM.

Some disk vendors allow the disk RPM to be reduced by about a quarter of its operational RPM (e.g., from 7200 to 5400 RPM when idle), thus reducing the constant energy consumption. In some cases, this can reduce the energy consumption for idle by almost half of the regular idle energy consumption.

4.2 Seek speed

Disk drives can control the disk head acceleration, deceleration and velocity by applying different currents to the voice-coil motor that moves the disk head. Vendors such as Seagate and Western Digital introduced a Just-In-Time (JIT) seek mode [Seagate2000]. With Seagate's JIT or Western Digital's IntelliSeek mode, the acceleration and speed of the disk head is adjusted so that the disk head will arrive at its destination in time for the data to be located beneath the disk head. This as opposed to normal mode, where the disk head may arrive too early, and will wait until the data is beneath the disk head. This method of slowing down the disk head leads to reduced acoustics and energy consumption without any performance degradation.

The SATA specification [ATAPI2002] includes a standard Automatic Acoustic Management (AAM) feature. This feature allows vendors to define various acoustic modes for the disk. Currently, vendors include only two modes, a normal acoustic mode and a quiet acoustic mode. In the quiet mode the disk performs seek operations at a reduced velocity (compared to normal mode); as a result the peak energy consumption of the disk drive is reduced.

4.3 Power modes

Power modes mainly have to do with the state of the disk when idle. Various vendors have increased the number of available power modes, for example, unloading and parking the heads, which reduces friction, or slowing down to a low RPM idle mode.

The SATA specification defines an Advanced Power Management (APM) feature, which supports moving a disk from one power mode to another, following a predefined settings (e.g., a given idle period), without the need to receive a specific command from the host. In addition to placing the disk at a lower power mode, recent works focus on putting the communication link between the host and the disk in a lower power mode. This may be very beneficial as maintaining the communication link consumes a considerable amount of energy this is especially noticeable when a disk enters a lower power mode, but still needs to be able to receive commands, such as a spin-up command, from the host.

A complementary approach to the above is maximizing the idle interval time between non-idle periods [Gurumurthi2003Zhang, Li2004]. This allows for longer intervals in which the disk can be placed in a lower power mode or turned off. The concept of Massive Array of Idle Disks (MAID), where most of the system disk drives are turned off, is the result of this approach [Colarelli2002, Pinheiro2004].

Chapter 5

Current Practices in Large Scale Distributed Systems

5.1 Evaluation

The increased pressure of energy consumption awareness leaded to the creation of new tools to evaluate and monitor whole data centers power consumption. The GREEN-GRID consortium established a number of useful documents[1] for designing data centers, measuring, adjusting and so on.

5.1.1 Buildings

As the environmental pressure rise, news buildings are designed with the energy management as a priority. By instance, EnergyStar helps evaluating the energy impact of a building [2]. It provides the EnergyStar label to buildings that achieve a 75 out of 100 points after evaluation. IBM provides a tool to evaluate energy efficiency of IT infrastructure [3]

- Metrics: To evaluate the quality of a data center in relation to energy several metrics exists:
  - Perf/Watt. This metric is mainly used to evaluate only the computing nodes. By instance Green500 [Green500] uses it to ranks the most powerful supercomputers (mainly clusters). It does not encompass the whole energy consumption of the room (such as AC) but only the consumption of the computing nodes themselves.
  - PUE (Power usage effectiveness). This value is complementary to the previous one. It evaluates the ratio between the total energy

[1] `http://www.thegreengrid.org/library-and-tools`

[2] `http://www.energystar.gov/index.cfm?c=evaluate_performance.bus_portfoliomanager_benchmarking`

[3] `http://ibmgreen.bathwick.com/`

consumed by the data center facility and the energy provided to the computing element [4]. In 2006, a classical PUE was about 2.0 [Malone2006], meaning that half of the energy consumed was to be used for cooling, lightning, ... but not computing.

- Real time monitoring

One of the most important improvement comes from feedback. The more data are available about power usage, the easier it is to optimize a data center consumption [5].

5.2 Context aware building

First of all, servers are not afraid of the dark: Lightling is unuseful! As self-evident this statement seems, it is common to see a full lightning in data centers. Occupacy sensors and/or economic bulbs can save a lot of energy without a extensive cost [6].

A common believed idea is that a data center in Groenland will consume less than a data center in Sahara, since the external temperature is on average lower. But it has been shown (for instance in the Energy Star study [EnergyStar2010], slide 23) that the external temperature has little impact on the overall electricity consumption of data centers. This study is not expliciting exactly the infrastructure of the builiding and the cooling of the server rooms. Indeed, if air circulation coming from outside is in the game, the difference will be significant while if traditional air conditionning is the rule then outside temperature has little influence.

More and more data centres are built so that they are using renewable energy. Solar panels (AISO [7], Phoenix [8], Intel [9], Sun [10], Google [11], ...), wind mills (Google [12], OWC [13], Green House Data [14], Baryonyx [15]) are producing part of

[4] http://www.google.com/corporate/green/datacenters/measuring.html
[5] http://technet.microsoft.com/en-us/magazine/2009.gr.datacenter.aspx
[6] http://hightech.lbl.gov/DCTraining/strategies/light.html
[7] http://www.aiso.net/technology-network-sun.html
[8] http://www.datacenterknowledge.com/archives/2009/06/16/solar-power-at-data-center-scale/
[9] http://www.datacenterknowledge.com/archives/2009/01/19/intel-testing-solar-power-for-data-centers/
[10] http://www.datacenterknowledge.com/archives/2008/05/22/the-solar-powered-blackbox/
[11] http://www.google.com/corporate/green/clean-energy.html
[12] http://www.datacenterknowledge.com/archives/2007/11/29/googles-data-center-windmill-farm/
[13] http://www.datacenterknowledge.com/archives/2009/12/21/data-center-powered-entirely-by-the-wind/
[14] http://www.datacenterknowledge.com/archives/2007/11/29/wind-powered-data-center-in-wyoming/
[15] http://www.datacenterknowledge.com/archives/2009/07/20/wind-powered-data-center-planned/

the electricity needed by the data centres (in one case, all the electricity: [16]). Most of the experiences are small size experiences, mainly due to the fact that the cost of these energy productions are still higher than normal electricity for the consumer.

Solutions are also developed to consume renewable electricity in data centers when the cost of electricity is high (typically during daytime) and use chillers during nights. Doing so, the cold that was produced and kept during night can be additionally used with the "free" electricity during day time [17]. This difference of electricity generation and usage can also reflect on the data centers usage itself, offloading the data centres whether during day time (when classicl electricity is the rule) or during nights (when solar panels are in the game).

Another trend are the movable data centers. For instance, IBM with portable modular data center (PMDC) [18]. It is advertized that "PMDCs have a power usage effectiveness (PUE) of 1.3, including the IT components and physical infrastructure such as chillers, UPS and other components. That compares to a PUE of 2.3 or higher for most existing data centers, and a PUE of 1.5-1.7 for some of the newer ground-based data centers.". Interestingly, Sun proposes a portable solution powered by solar panels [19].

5.3 Cooling

An important part of the data centres energy consumption is wasted for cooling the running components. As explained above, the typical PUE of a data centre was about 2.0 in 2006, meaning that one watt for the infrastructure is wasted for each watt used to compute. AMong this waste, part of it is due to the cooling.

The first aspect on this is to determine the optimal operational temperature for a data centres. Recent studies tend to exhibit that data centres are often too cold[20] and could operate at higher temperature (with some limits): A consensus is agreed by the industry to maintain an ambient temperature range of 20 to 24C, while the limit is set to 30C. A study jointly published by Intel, IBM, HP and Lieberth [21] shows that most data centers are cooled at 20C while they could operate at 26C [ASHRAE2008].

Several techniques exist and often coexist to cool down the server rooms. Traditionally, air conditionning has been used ever and ever for cooling the infrastructure. Problems arise when the air circulation has not been optimally studied between the racks in the rooms. Some hot spots can exist, and a full investigation taking into account CFD models, cold and hot aisle locations,

[16] http://www.datacenterknowledge.com/archives/2009/06/16/solar-power-at-data-center-scale/

[17] http://www.datacenterknowledge.com/archives/2009/06/16/solar-power-at-data-center-scale/

[18] http://www.environmentalleader.com/2009/12/03/ibm-advances-data-center-efficiencies/

[19] http://www.datacenterknowledge.com/archives/2008/05/22/the-solar-powered-blackbox/

[20] http://www.greenbiz.com/blog/2009/09/01/your-data-center-much-too-cold

[21] http://download.intel.com/pressroom/archive/reference/IPACK2009.pdf

must be done. Some vendors (HP with Dynamic Smart Cooling [22], DegreeC with AdaptivCool [23]) are offering tools to monitor and adjust cooling according to heat dispersion and air circulation.

Another way witnessed is to use cold air column, where the heated air is directed from behind the racks to ease the air circulation. Such an approach can be seen at the Barcelona Marenostrum for instance.

Water cooling is being more and more used, since the efficiency of heat dispersion with water is much higher than with air. In these solutions, water circulates behind the racks and capture the heat and direct it away from the server, before being chilled again and sent back colder. For instance, the CALMIP machine in Toulouse is working with this system.

5.4 Uses cases and example of current practices

As energy awareness gains momentum, several uses cases have been fully documented:

- An US best practices repository[Greenberg2006], after an extensive benchmarking of 22 data centers: `http://hightech.lbl.gov/DCTraining/Best-Practices.html`
- In `http://www1.eere.energy.gov/industry/saveenergynow/pdfs/doe_data_centers_presentation.pdf` the US Department of Energy shows a joint study with LucasFilm and Verizon.
- In `http://www-01.ibm.com/software/success/cssdb.nsf/solutionareaL2VW?OpenView&Count=30&RestrictToCategory=corp_Energyefficiency&cty=en_us` IBM provides information about uses cases where its technology improved energy efficiency.
- In `http://www.microsoft.com/environment/news_resources/case_studies.aspx` Microsoft shows cases where its technology helped reduce carbon footprint

[22] `http://www.hp.com/hpinfo/newsroom/press/2006/061129xa.html`
[23] `http://www.adaptivcool.com/`

Chapter 6

References

[ACPI] ACPI specification - http://www.acpi.info/spec.htm

[Alon2004] Efi Rotem, Alon Naveh, Micha Moffie and Avi Mendelson, "Analysis of Thermal Monitor features of the Intel Pentium M Processor in TACS workshop, at ISCA-31, June 2004.

[Alon2006] Alon Naveh, Efraim Roterm, Avi Mendelson, Simcha Gochman, Rajshree Chabukswar, Karthik Krishnan, Arun Kumar , "Power and Thermal Management in the Intel Core Duo Processor Architecture" in Intel Technology Journal, Volume 10, issue 02, pp 109-122, 2006.

[ASHRAE2008] 2008 ASHRAE Environmental Guidelines for Datacom Equipment, -Expanding the Recommended Environmental Envelope-, ASHRAE TC 9.9, ASHRAE, 2008

[ATAPI 2002] INCITS 361-2002 (1410D): AT attachment - 6 with packet interface (ATA/ATAPI - 6), 2002.

[Carrera2003] E. V. Carrera, E. Pinheiro, and R. Bianchini, Conserving disk energy in network servers, in Proceedings of the 17th Annual International Conference on Supercomputing, June 2003, pp. 8697.

[Colarelli2002] D. Colarelli and D. Grunwald, Massive arrays of idle disks for storage archives, in Proceedings of the 2002 ACM/IEEE conference on High Performance Networking and Computing, November 2002, pp. 111.

[EnergyStar2010] http://www.thegreengrid.org/~/media/TechForumPresentations2010/ENERGYSTARforDataCenters.ashx?lang=en

[Fan2007] Fan, X., Weber, W., and Barroso, L. A. 2007. Power provisioning for a warehouse-sized computer. In Proceedings of the 34th Annual international Symposium on Computer Architecture (San Diego, California, USA, June 09 - 13, 2007). ISCA '07.

[Gelsi2008] Gelsinger talk at IDF-08 on Nehalem power management: http://news.zdnet.com/2422-19178_22-216954.html

[Green500] http://www.green500.org/

[Greenberg2006] Greenberg, S., Mills, E., Tschudi, B., Rumsey, P., and Myatt, B. (2006, August). Best practices for data centers: Lessons learned from

benchmarking 22 data centers. Paper presented at the ACEEE Summer Study on Energy Efficiency in Building, Asilomar, California.

[Gurumurthi2003Sivasubramaniam] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke, DRPM: Dynamic speed control for power management in server class disks, in Proceedings of the 30th Annual International Symposium on Computer Architecture, June 2003, pp. 169181.

[Gurumurthi2003Zhang] S. Gurumurthi, J. Zhang, A. Sivasubramaniam, M. Kandemir, H. Franke, N. Vijaykrishnan, and M. J. Irwin, Interplay of energy and performance for disk arrays running transaction processing workloads, in Proceedings of the International Symposium on Performance Analysis of Systems and Software, March 2003, pp. 123132.

[Li2004] X. Li, Z. Li, F. David, P. Zhou, Y. Zhou, S. Adve, and S. Kumar, Performance directed energy management for main memory and disks, in Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems. ACM Press, 2004, pp. 271283.

[Lim2008] Lim, K., Ranganathan, P., Chang, J., Patel, C., Mudge, T., and Reinhardt, S. 2008. Understanding and Designing New Server Architectures for Emerging Warehouse-Computing Environments. SIGARCH Comput. Archit. News 36, 3 (Jun. 2008)

[Llano] AMD Reveals More Llano Details at ISSCC: 32nm, Power Gating, 4-cores, Turbo?:

[Llano2] `http://www.anandtech.com/show/2933`

[Malone2006] Malone, C., Belady, C., 2006, Metrics to Characterize Data Center & IT Equipment Energy Use, Proceedings of 2006 Digital Power Forum, Richardson, TX.

[Pinheiro2004] E. Pinheiro and R. Bianchini, Energy conservation techniques for disk array-based servers, in Proceedings of the 18th Annual International Conference on Supercomputing, June 2004, pp. 6878.

[PmI7] Power management of Intel I7 cores - `http://cs466.andersonje.com/public/pm.pdf`

[Seagate2000] Seagates sound barrier technology (SBT), 2000, `http://www.seagate.com/docs/pdf/whitepaper/sound_barrier.pdf`