Hardware leverages for energy reduction in large scale distributed systems, 2nd Year

Editors: *Georges Da Costa and Davide Careglio*
Contributors of 2nd Year: *Davide Careglio, Georges Da Costa, Ronen I. Kat, Jean-Marc Pierson*

Technical Report : IRIT/RT–2011-1–FR

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 Hanover, 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 centers. 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 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 implemented)

- C3 (Sleep) offers improved power savings over the C1 and C2 states. The worst-case wake-up 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 by HW only.

2.2.3 P-States

P-States indicates 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 GPU

While traditional data centers are not using GPU or Cells, a current trend for the most powerful computers is to use such alternative hybrid architectures (combining CPU with Cells/GPU) to deliver ever more processing power. In the Top500 list, the Chinese Nebulae ranks second (June 2010): It is composed of Intel CPU and Nvidia GPU. In the Green500 list, we can find such data centers in the first eight positions. Indeed, from an energy point of view, it can be competitive, since the scheduled jobs finish earlier, energy (which is power x time) is spent for a shorter time and the Flops/Watts metric reaches 773 MFlops/Watts. Nvidia ships the Tesla GPU Computing Systems, consisting of 1U servers embedding 4 GPU (for instance the S2050 is delivering 2 TFlops in double precision at the cost of 900 Watts). Each GPU individually can consume as much as 250 W, for instance the Tesla C2050. The main problem with such infrastructure is when it is idle, since it is not possible to deactivate a GPU card:

When installed, it will anyway consume an important minimal amount of power (not less than 50-60 watts), and there is no such mechanisms to completely switch off GPU elements.

2.4 Vendors

2.4.1 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.

2.4.2 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 CoreDuo-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 P_n 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[Alon2004].

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.

2.4.3 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 pre-charge (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 on-die 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.

Common DDR3 runs at 1.5 Volts while Kingston manufacturer have a DDR3 that operates at 1.25 Volts for 1600 MHz, until DDR4 is actually produced in mass, requiring less energy (1 Volt) and up to 3200 MHz. Recently, a joint initiative in the European funded project EuroCloud pushes ARM Cortex A9 processors, linked with 3D DRAM to create 3D server on chip, serving as a basis for energy efficient data centers (`EuroCloud:http://www.eurocloudserver.com/`).

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% outright.

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.

Samsung Samsung aims at optimizing the base consumption of its 'Green memory' product line. Like other vendors, they do not provide dynamic way to reduce energy. Samsung uses 40nm and 30nm technology and reduced voltage to reduce this base consumption. The following data can be found in [Sam10]:

Size (Gb)	process technology (nm)	Tension (V)	Consumption (W) (for a 48Gb server active for 8 hours a day)
1	60	1.8	102
1	60	1.5	66
1	50	1.5	50
1	40	1.5	41
2	40	1.5	34
2	40	1.35	28
2	30	1.35	24
4	30	1.35	14

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].

4.4 Power consumption

Typical HDDs consume (numbers for a 2 TB Seagate Constellation ES at 7200 RPM, SATA, 140 MB/s transfer rate, 3”5) about 7 Watts when idle, and 10-11 watts when busy (read operations being more power consuming). Lower disk capacities consume less power, down to 4.6 / 9.4 / 8.2 watts (idle / read / write) for a 500 GB HDD. On these disks for instance, a PowerChoice mode (a proprietary implementation of T10 and T13 Standards [SCSI-T10, SATA-T13]) makes the disk power consumption drop down to 0.53 Watts. Smaller disks (2”5) formally only in laptops, are now getting much interest from Data Centers despite their more limited capacities at comparable performances: They run at about half the power of 3”5 disks and takes less space in the racks. For instance, the Savvio (15K.2, 15000 RPM, SAS) offers 146 GB only but consumes 4.1 watts when idle.

4.5 SDD

SDD are garnering much interests in the last years. Their most important feature is the improved access time: A multilevel cell (MLC) SSD has an access time of 0.5ms compared with an access time of 15.7 ms for a 7,200 RPM drive. Please note that the highest performances coming for SDD access rate can be limited from an application point of view in some cases (see study on the comparison metrics [IDCSurvey]). As no mechanics exist in a SSD drive, the power consumption is only a fraction of the one of a HDD. A typical Seagate SDD drive (Pulsar, 200 GB, SATA, 300 MB/s) consumes only 0.75 watt when idle and 1.3 watt in operation. This improved energy performance comes with a higher price and limited capacities, making them not really sustainable in big data centers that hosts Tera or Peta Bytes of data.

Chapter 5

Fan

As a part of the cooling infrastructure, fan can represent a really important part of energy consumption of computing elements. First it can impact during boot as usually fans start at full speed during boot sequence and then reduce their speed in case the part they cool is not overheated[SaveWatts]. This has a large impact of several energy reduction technique that switch down computers when they are not needed because those techniques then need to switch on a potentially large number of nodes together leading to a high power consumption peak.

But even during classical runs, impact can be high. On [Satoshi10], authors shows that for blades (c7000) fans consumption can go up to 16-20% of the global consumption. They derive that the low to link power consumed and fan speed (in RPM) is of the type $Power = Cte * Speed^3 + Cte$ (in their case: $power = 2.33 * 10^{-11} rpm^3 + 7.7$)

Fan can be controlled using ACPI commands. Some information can be obtained such as the rotation speed.

Chapter 6

Network Interface

6.1 IEEE P802.3az

Through the IEEE P802.3az Energy Efficient Ethernet Task Force [1], a consortium mixing academic and industries is proposing new solutions for obtaining Energy Efficient Ethernet solutions.

Today, energy consumption of Ethernet networks is not greatly linked with bandwidth utilization. So even in low or no usage context, networks equipments consume energy at high level. As a first approach, by proposing Adaptive Link Rate solutions, energy savings can be obtained by quickly changing the speed of network links in response to the amount of data that is being transmitted.

Now, for high speed Ethernet networks (1 and 10 Gbits) used in data centers, the Energy Efficient Ethernet Task Force is proposing Low power Idle modes which should allow to power down and quickly wake up specific components of Ethernet products.

[1] `http://www.ieee802.org/3/az/`

Chapter 7

Advances in network infrastructures

7.1 Current scenario

It is now held as a scientific fact that humans contribute to the global warming of planet Earth through the release of carbon dioxide (CO2), a Green House Gas (GHG), in the atmosphere. Recently, the carbon footprint of ICT was found to be comparable to that of aviation [Gartner2007]. It is estimated that 2-3% of the CO2 produced by human activity comes from ICT [Global2007][Smart2020] and a number of studies estimate an energy consumption related to ICT varying from 2% to 10% of the worldwide power consumption [Global2007]. It is worth to mention for example that Telecom Italia and France Telecom are now the second largest consumer of electricity in their country [Pileri2007][Souchon2009]. Recent initiatives gathering major IT companies started to explore the energy savings and green energy use in network infrastructure. For example, Telefonica commits to reducing 30% its energy consumption in network by 2015 [Lange2009].

But the Internet traffic is growing constantly over the years (independent of the momentary economic conditions) impacting directly to the network energy consumption which is in fact growing with the network throughput according to a power of two thirds [Tucker2009]. It is clear that increased energy consumption of the Internet will increase operational costs in the network and will exacerbate the thermal issues associated with large data centres and switching nodes.

In the current telecommunications networks, the vast majority of the energy consumption can be attributed to fixed line access networks. Today, access networks are mainly implemented with copper based technologies such as ADSL and VDSL whose energy consumption is very sensible to increased bit-rates. The trend is to replace such technologies with mobile and fiber infrastructure which is expected to increase considerably the energy efficiency in access networks. Such ongoing replacement is moving the problem to the backbone net-

works where the energy consumption for IP routers is becoming a bottleneck [LangeK2009][Tucker2009]. In Japan it is expected that by 2015, IP routers will consume 9% of the nations electricity [Nature2007].

7.2 New energy-oriented model

Such case shows that increasing the energy efficiency of the different equipments, operations or processes constituting a network infrastructure is not the solution as argued in the Khazzoom-Brookes postulate [Saunders1992]: increased energy efficiency paradoxically tends to lead to increased energy consumption (a phenomenon known as the Jevons Paradox). In fact, an improvement of the energy efficiency leads to a reduction of the overall costs, which causes an increase of the demand and consequently of the energy consumption overtaking hence the gained offset.

It is safe to say that a paradigm shift is required in the network in order to sustain the growing traffic rates while limiting and even decreasing the power consumption. Two terms must be defined at this point.

- Energy efficiency refers to a technique or equipment designed or developed to reduce the ICT energy consumption without affecting the performance and taking into account the environmental impact of the used resources. Such solutions are also usually referred as eco-friendly solutions.
- Energy awareness refers to a technology or technique that adapts its behavior or performance based on the source of the energy that supplies the network. It implies direct knowledge of the quantity and quality of the renewable energy a network or equipment is expending. A direct benefit of energy aware techniques is the removal of the Khazzoom-Brookes postulate.

To become a reality, green Internet must rely on both concepts and a new energy-oriented network architecture is required, i.e. a comprehensive solution encompassing both energy-efficient devices and energy-aware paradigms acting in a systemic approach. For example, rather than bringing the electrical power to data centers (with relative power losses), it seems more appropriate to move the data centres to the source of renewable power and connect them to Internet with long reach fibre optic cable. Another alternative is to increase the use of thin clients instead of desktop PCs [Pickavet2007]; a thin client consumes less energy but requires a distributed system and a network able to support it.

7.3 Current advances in networking

From a network point of view, a way to increase the energy issue is maximizing the integration between IP/packet layer, optical transport layer and control layer in so called multi-layer approach. In the past, these layers evolved with

different constraints and without an overall optimization, which has led to the issues we are facing today such as inefficient energy strategy, limited network scalability and flexibility, reduced network manageability and increased overall network and costumer services costs. It is clear that a technique able to provide an optimal solution in several layers can cope better with the variety of possible phenomena in an overall efficient way and can benefit from the advantages of the solution in each layer.

For what regards the energy efficiency, the pioneering work in [Gupta2003] suggested the introduction of sleep mode in networks. When elements are idle their power consumption is obviously wasteful; turning these elements off reduces the power consumption. Although such functionality may not bring any advantageous when used alone, recent results (see e.g. [Pickavet2007][Tucker2008]) showed that if applied in multi-layer scenario, sleep mode can substantially reduce the energy consumption without a corresponding decrease of the network performance. For example, combining traffic grooming (at IP layer) to maximize the utilization of optical transmission links and transparent optical path (at the optical layer) to bypass routers wherever possible (so to reduce the hop count) is an optimal trade-off solution to increase energy efficiency. In such a situation in fact, the utilization of active interfaces and links are maximized while unused interfaces or even entire router and switch components can be put in sleep mode.

Nonetheless, current router architectures are not energy-aware, in the sense that their energy consumption does not scale sensibly with the traffic load. In [Chabarek2008] several router architectures have been analyzed and their energy consumptions under different traffic loads have been evaluated. Results show that the energy consumption between an idle and a heavily loaded router (with 75% of offered traffic load) vary only of 3% (about 25 W on 750 W). This happens because the router line cards, which are the most power consuming elements in a router, are always powered on even if they are totally idle. On the contrary, the energy consumption decreases to just 50% if the idle line cards are physically disconnected. Such a scenario suggests that future router architectures will be energy-aware, in the sense that they will be able to automatically switch off or dynamically down-clock independent subsystems (e.g. line cards, input/output ports, switching fabrics, buffers, etc.) according to the traffic loads in order to save energy whenever possible. Such energy-aware architectures are advocated both by standardization bodies and governmental programs [Star2010] and have been assumed by various literature sources [Gupta2003][Muhammad2010][Chabarek2008].

For what regards the energy awareness, anything less than a target of zero carbon emissions throughout the entire ICT system will be pointless due to the Khazzoom-Brookes postulate; with a zero carbon footprint any increase in consumption will still result in a total cumulative zero carbon footprint. In such direction, recent studies propose to include the information on the energy sources in the network operations. At the same time, the market is now offering renewable power supplies (mainly solar panels) for network sites (both nodes and amplifier sites) in such a way that legacy (dirty) energy sources are used

only to guarantee power supply without any interruption [Concentralia].

Chapter 8

Power supply unit

Most current computers still use an independent PSU. Due to current technological limitation, their yield are still usually limited at around 80%. In this case, 20% of energy consumed by a computer only heat the computer PSU. Moreover depending on the load on the PSU, this yield can change.

The 80 Plus[1] initiative tries to improve awareness of the efficiency of PSU by multi-level certification:

80 PLUS Certification	115V Internal Non-Redundant			230V Internal Redundant		
% of Rated Load	20%	50%	100%	20%	50%	100%
80 PLUS	80%	80%	80%	N/A		
80 PLUS Bronze	82%	85%	82%	81%	85%	81%
80 PLUS Silver	85%	88%	85%	85%	89%	85%
80 PLUS Gold	87%	90%	87%	88%	92%	88%
80 PLUS Platinum	90%	92%	89%	90%	94%	91%

[1] `http://www.80plus.org`

Chapter 9

Current Practices in Large Scale Distributed Systems

9.1 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.

9.1.1 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 [4] 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.

[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/`

[4] `http://www.green500.org/`

– PUE (Power usage effectiveness). This value is complementary to the previous one. It evaluates the ratio between the total energy consumed by the data center facility and the energy provided to the computing element [5]. 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. The newest Yahoo center [6] constructed near the Niagara falls uses circulating exterior air to cool the servers, and is able to achieve a PUE of around 1.1.

- 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 [7].

9.2 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 [8].

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 [9], slide 23) that the external temperature has little impact on the overall electricity consumption of data centers. This study does not explicit exactly the infrastructure of the building 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 conditioning is the rule then outside temperature has little influence.

More and more data centers are built so that they are using renewable energy. Solar panels (AISO [10], Phoenix [11], Intel [12], Sun [13], Google [14], ...), wind mills

[5] http://www.google.com/corporate/green/datacenters/measuring.html

[6] http://green.yahoo.com/blog/ecogeek/1125/yahoo-data-center-will-be-powered-by-niagara-falls.html

[7] http://technet.microsoft.com/en-us/magazine/2009.gr.datacenter.aspx

[8] http://hightech.lbl.gov/DCTraining/strategies/light.html

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

[10] http://www.aiso.net/technology-network-sun.html

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

[12] http://www.datacenterknowledge.com/archives/2009/01/19/intel-testing-solar-power-for-data-centers/

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

[14] http://www.google.com/corporate/green/clean-energy.html

(Google [15], OWC [16], Green House Data [17], Baryonyx [18]) are producing part of the electricity needed by the data centers (in one case, all the electricity: [19]). 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 [20].

This difference of electricity generation and usage can also reflect on the data centers usage itself, offloading the data centers whether during day time (when classical 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) [21]. It is advertised 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 [22].

9.3 Cooling

An important part of the data centers energy consumption is wasted for cooling the running components. As explained above, the typical PUE of a data center 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 centers. Recent studies tend to exhibit that data centers are often too cold[23] 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

[15] http://www.datacenterknowledge.com/archives/2007/11/29/googles-data-center-windmill-farm/

[16] http://www.datacenterknowledge.com/archives/2009/12/21/data-center-powered-entirely-by-the-wind/

[17] http://www.datacenterknowledge.com/archives/2007/11/29/wind-powered-data-center-in-wyoming/

[18] http://www.datacenterknowledge.com/archives/2009/07/20/wind-powered-data-center-planned/

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

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

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

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

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

and Lieberth [24] 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 conditioning 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, must be done. Some vendors (HP with Dynamic Smart Cooling [25], DegreeC with AdaptivCool [26]) 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.

9.4 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: [27]
- In [28] the US Department of Energy shows a joint study with LucasFilm and Verizon.
- In [29] IBM provides information about uses cases where its technology improved energy efficiency.
- In [30] Microsoft shows cases where its technology helped reduce carbon footprint
- In [31] Accenture and major leaders are forecasting the future. (July 2008)

[24] http://download.intel.com/pressroom/archive/reference/IPACK2009.pdf
[25] http://www.hp.com/hpinfo/newsroom/press/2006/061129xa.html
[26] http://www.adaptivcool.com/
[27] http://hightech.lbl.gov/DCTraining/Best-Practices.html
[28] http://www1.eere.energy.gov/industry/saveenergynow/pdfs/doe_data_centers_presentation.pdf
[29] http://www-01.ibm.com/software/success/cssdb.nsf/solutionareaL2VW?OpenView&Count=30&RestrictToCategory=corp_Energyefficiency&cty=en_us
[30] http://www.microsoft.com/environment/news_resources/case_studies.aspx
[31] https://microsite.accenture.com/svlgreport/Documents/pdf/SVLG_Report.pdf

9.5 References

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

[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.

[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.

[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.

[Chabarek2008] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, and S. Wright, Power awareness in network design and routing, in Proc. IEEE INFOCOM, 2008.

[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.

[Concentralia] Concentralia, http://www.concentralia.net/

[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.

[Gartner2007] Gartner press release, 2007
http://www.gartner.com/it/page.jsp?id=503867.

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

[Global2007] An inefficient Truth by the Global Action Plan,
http://www.globalactionplan.org.uk/

[Greenberg2006] Greenberg, S., Mills, E., Tschudi, B., Rumsey, P., & 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.

[Gupta2003] M. Gupta, S. Singh, Greening of the Internet, in Proc. ACM SIGCOMM 2003, Karlsruhe, Germany, Aug. 2003.

[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.

[IDC-Survey] John Rydning, David Reinsel, and Jeff Janukowicz. White paper: The need to standardize storage device performance metrics, September 2008.

[Lange2009] C. Lange, Energy-related Aspects in Backbone Networks, in Proc. ECOC 2009, Vienna, Austria, Sep. 2009.

[LangeK2009] C. Lange, D. Kosiankowski, C. Gerlach, F. Westphal, A. Gladisch, "Energy Consumption of Telecommunication Networks, in Proc. ECOC 2009, Vienna, Austria, Sep. 2009.

[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.

[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)

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

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

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

[Muhammad2010] A. Muhammad, Paolo Monti, Isabella Cerutti, Lena Wosinska, Piero Castoldi, Anna Tzanakaki, Energy-Efficient WDM Network Planning with Protection Resources in Sleep Mode, accepted for Globecom 2010, ONS01.

[Nature2007] Nature Photonics technology conference 2007, Tokyo, Japan, Oct. 2007.

[Pickavet2007] M. Pickavet, R. Van Caenegem, S. Demeyer, P. Audenaert, D. Colle, P. Demeester, R. Leppla, M. Jaeger, A. Gladisch, and H.-M. Foisel, Energy footprint of ICT, in Broadband Europe 2007, Dec. 2007.

[Pileri2007] S.Pileri, Energy and Communication: engine of the human progress, INTELEC 2007 keynote, Rome, Italy, Sep. 2007.

[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.

[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`

[SATA-10] SCSI Specification : INCITS Technical Committee T10 subcommittee SCSI. www.t13.org/.

[SATA-13] SATA Specification : INCITS Technical Committee T13 subcommittee ATA. www.t13.org/.

[Sam10] Green Memory Moving into the Driver's Seat, Sylvie Kadivar, at Intel Developer Forum 2009

[Satoshi10] "Measuring and Modeling of a Data Center", Satoshi Itoh, Yuetsu Kodama, Hiroshi Nakamura, Naohiko Mori, Toshiyuki Shimizu and Satoshi Sekiguchi

[Saunders1992] Harry D. Saunders, The Khazzoom-Brookes postulate and neoclassical growth, The Energy Journal, Oct. 1992.

[SaveWatts] Anne-Cecile Orgerie, Laurent Lefevre, and Jean-Patrick Gelas. "Save Watts in your Grid: Green Strategies for Energy-Aware Framework in Large Scale Distributed Systems", ICPADS 2008 : The 14th IEEE International Conference on Parallel and Distributed Systems, Melbourne, Australia, December 2008

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

[Smart2020] SMART 2020: Enabling the low carbon economy in the information age. The climate group, 2008.

[Souchon2009] L. Souchon Foll, TIC et nergtique : Techniques destimation de consommation sur la hauteur, la structure et lvolution de limpact des TIC en France, Ph.D. dissertation, Orange Labs/Institut National des Tlcommunications, 2009.

[Star2010] Energy Star, Small network equipment,
http://www.energystar.gov/index.cfm?c=new_specs.small_network_equip.

[Tucker2008] R.S. Tucker, J. Baliga, R. Ayre, K . Hinton, W.V. Sorin, Energy consumption in IP networks, in Proc. ECOC 2008, Brusseles, Belgium, Sep. 2008.

[Tucker2009] R. S. Tucker, R. Parthiban, J. Baliga, K. Hinton, R.W. A. Ayre, W.V. Sorin, Evolution of WDM Optical IP Networks: A Cost and Energy Perspective, IEEE/OSA J. Lightw. Technol., vol. 27, no. 3, Feb. 2009.