Centro Euro-Mediterraneo
per i Cambiamenti Climatici

Research Papers
Issue 2011
December 2011

Scientific Computing and Operation (SCO)

Green Computing and power saving in HPC data centers

By **Osvaldo Marra**
CMCC
University of Salento, Italy
osvaldo.marra@unisalento.it

Maria Mirto
CMCC
maria.mirto@cmcc.it

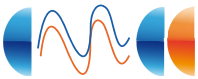
Massimo Cafaro
CMCC
University of Salento, Italy
massimo.cafaro@unisalento.it

and **Giovanni Aloisio**
CMCC
University of Salento, Italy
giovanni.aloisio@unisalento.it

SUMMARY Data centers now play an important role in modern IT infrastructures. Much research effort has been made in the field of green data center computing in order to save energy, thus making data centers environmentally friendly, for example by reducing CO_2 emission. The Scientific Computing and Operations (SCO) division of the CMCC has started a research activity exploiting proactive and reactive monitoring techniques in order to reduce the power consumption of data centers by using Green Computing technologies. The goal of this research paper is to provide an overview about the main issues and challenges to improve HPC resources management in terms of power consumption. A feasibility study on the possibility to introduce a system to measure the power consumption for the CMCC data center, and in particular for the Calypso cluster, is described.

INTRODUCTION

High performance computing (HPC) - using commodity clusters, special-purpose multiprocessors, massive data stores and high-bandwidth interconnection networks - is rapidly becoming an indispensable research tool in science and engineering. In studies on climate changes, HPC is fundamental for modeling and predicting climate changes and computation is taking an increasingly important role beside theory and experimentation as a tool for inquiry and discovery. Beyond science and engineering, large compute clusters, massive storage, and high-bandwidth connectivity have also become mission-critical infrastructure in industrial and academic settings. While the costs of computers, storage, and networking have fallen dramatically over time, the costs of the buildings, power, and cooling that support this equipment have risen dramatically. Indeed, the costs for power and infrastructure exceed the cost of the computing equipment they support. The environmental costs of information and communication technologies are also increasing; recent studies estimate that computing and communication technology sectors are responsible for 2% of global carbon emissions and these emissions are increasing at 6% annually [1]. Green computing (also known as sustainable computing) can be broadly defined as the problem of reducing the overall carbon footprint (emissions) of computing and communication infrastructure, such as data centers, by using energy-efficient design and operations. The area has garnered increasing attention in recent years, both from a technology and societal standpoint, due to the increased focus on environmental and climate change issues. Hence, there is a need to balance the dramatic growth of high-performance computing clusters and data centers in the computational sciences with green design and use so as to reduce the environmental impact. Technical issues in high-performance green computing span the spectrum from green infrastructure (energy-efficient buildings, intelligent cooling systems, green/renewable power sources) to green hardware (multi-core computing systems, energy-efficient server design, energy-efficient solid-state storage) to green software and applications (parallelizing computational science algorithms to run on modern energy-efficient multi-core clusters, intelligent load distribution and CPU switch-off). The SCO division of the CMCC has started a research activity about the use of proactive and reactive monitoring techniques in order to reduce the power consumption of data centers by using Green Computing technologies. The goal of this research paper is to provide an overview about the main issues and challenges to improve the HPC resources management in terms of power consumption. In particular, a feasibility study on the possibility to introduce a system to measure the power consumption for the CMCC data center, and in particular for the Calypso cluster is given. The research paper is articled as follows. Starting from an overview of data centers, several issues related to their management are discussed. In particular several metrics have been introduced in order to measure the power consumption (e.g., PUE). Moreover, several algorithms and techniques are described in order to balance the workload on the cluster. Hence several benchmarking tools are introduced for evaluating the system performances by measuring the power consumption when varying the workload. Finally, a feasibility study on the possibility to introduce a system to measure the power consumption for the CMCC data center, and in particular for the Calypso cluster, is provided.



TOWARDS "GREEN" DATA CENTERS

A data center hosts computational power, storage and applications required to support an enterprise business. A data center is central to modern IT infrastructure, as all enterprise content is sourced from or passes through it. Data centers can be broadly classified, on the basis of power and cooling layout, into one of the 4 tiers:

- Tier 1: Single path for power and cooling; no redundant components;
- Tier 2: Redundancy added to Tier 1, thereby improving availability;
- Tier 3: Multiple power and cooling distribution paths, of which one is active;
- Tier 4: Two active power and cooling paths, and redundant components on each path.

This classification however, is not precise and commercial data centers typically fall between Tiers 3 and 4 [21]. A higher tier implies an improvement in resource availability and reliability, but it comes at the expense of an increase in power consumption. Data centers host services that require high availability, close to 99.99%. Fault tolerance, therefore, becomes imperative. The loss of one or more components must not cause the data center to terminate its services to clients. Consequently, data centers feature hardware redundancy. Furthermore, data centers are designed for a peak load, which might be observed only occasionally, and for short bursts of time. This conservative design results in over-provisioning of hardware in a data center. All these factors combined together contribute to the high power consumption of data centers. So the electricity usage is the most expensive portion of a data center's operational costs. In fact, the U.S. Environmental Protection Agency (EPA) reported that 61 billion KWh, 1.5% of US electricity consumption, is used for data center computing [18]. Additionally, the energy consumption in data centers doubled between 2000 and 2006. It is reported that power and cooling costs are the most dominant costs in data centers [7]. Other studies ([4], [40]) have reported the global electricity costs for data centers running into billions of dollars. Thus there is a growing pressure from the general public for the IT society to offer *green* data center infrastructures and services [17]. There are two major and complementary methods [38] to build a green data center: (1) utilize green elements in the design and building process of a data center, (2) *greenify* the process of running and operating a data center in everyday usage. Various research activities have been carried out to manage data centers in a green mode (the latter method), such as reducing data center temperature [36], [42], [43], increasing server utilization [33], [35], [41], and decreasing power consumption of computing resources [22], [23], [27], [32]. A fundamental research topic for the above study is how we can define performance metrics to identify how green a data center is. Green performance metrics, for data centers, are a set of measurements that can qualitatively or quantitatively evaluate the environmental impact of a data center. Nowadays, the term *green computing* requires clear scientific definitions and it is more or less a marketing concept. Research on developing green data center metrics brings solid specifications and definitions for green data centers in the following aspects:

- identify and specify clearly how green a data center is, for example by calculating its energy efficiency or greenhouse gas emission per time unit;
- evaluate data center products and compare similar data centers;

- track green performance to increase a data center's green efficiency, and provide guidance to engineers, manufacturers, and service providers;
- research and development on future green data center technologies.

In the power/energy metrics Section, several performance metrics are introduced.

ENERGY USE IN A DATA CENTER

Making data centers more energy efficien a multidimensional challenge that require concerted effort to optimize power distribut cooling infrastructure, IT equipment and IT put. A data center is composed of sev components involved in the power consump (Figure 1):

- cooling equipment (e.g., computer r air-conditioner units);
- power equipment (e.g., uninterrup power supplies and power distribut units);
- IT equipment (e.g., rack optimized non-rack optimized enterprise serv blade servers, storage and networ equipment);
- miscellaneous equipment (e.g., lightii

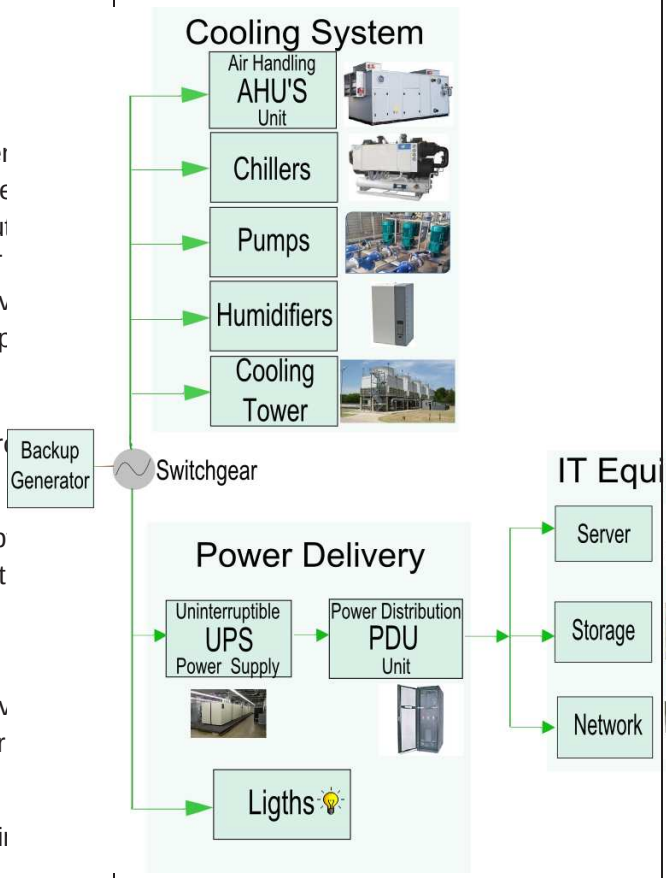


Figure 1:
Components of a Data Center

The cooling system occupies a significant part, in terms of energy consumption, in a data center. The IT components present inside a data center when running produce a high heat. To avoid overheating and consequent damage of IT components, it is necessary to cool them and keep them continuously at a constant temperature. For this purpose, a cooling system which includes chillers, pumps and fans is used. Figure 2 shows a typical cooling system, with modular conditioning units and a raised floor, which

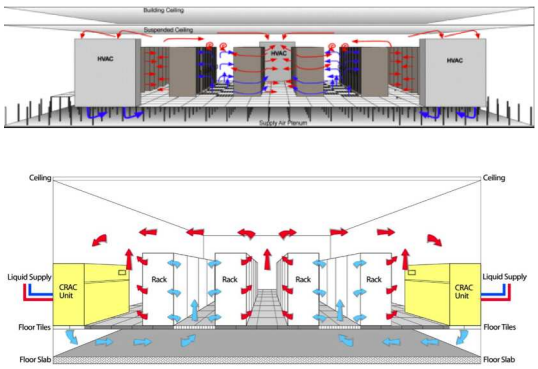


Figure 2:
A cooling system

allows the circulation of cold air. The size and power of the system is based on the expected load of the servers and is chosen during the design phase of the data center.

The air flow is evenly distributed through the CRAC (Computer Room Air Conditioning), which controls the temperature through a sensor typically set at 20°C, which measures the internal temperature. The cold air is distributed through the floor of the room with the use of fans and pumps, cooling the servers from below; the hot air, which goes up, goes back in the air conditioner to be cooled again and distributed in the data center. The combination of these components ensures optimal longevity of the equipment. At the same time, however, this ecosystem has a significant cost in the economy of the data center. The power equipment includes an energy supply system for the data center. Typically, the infrastructure includes the connection to the electricity grid, generators, backup batteries and the energy for the cooling system. The energy is recovered externally from the power supply. In the event of a power failure occurrence within the data center, a generator will supply the necessary energy to the system. Electrical device backup batteries are recharged, with the aim of maintaining a constant power even in case of interruptions both from the mains and the internal generator and to compensate for fluctuations of the network or temporary loss of power. The energy is distributed to the computing resources of the data center through the UPS (Uninterruptible Power Supply), it is properly adapted to the right voltage of the IT components, using the PDU (Power Distribution Unit). A double connection to two different PDUs may also be available, to improve the reliability of the system. The degree of redundancy in the system, therefore, is a very important consideration in the infrastructure. The data center should not run out of energy so for this there are a number of sources of supply and battery backup; the more the system is redundant, the greater the fault tolerance, but at the same time, the costs of installation and maintenance of this system will be higher. The IT equipment is the main part of a data center, which makes possible its basic operations, and contains the electronics used for computing (servers), storing data (data storage) and communicating (network). Collectively, these components process, store and transmit digital information. They require a suitable environment, where they are positioned and arranged in an optimal manner, to reduce overheating and to prevent accidents to the machines and the support staff. The servers, typically, are located in 1U or blade units, rack-mounted and interconnected via local Ethernet switches, these switches can be used in rack-level connections at 1 or 10 Gbps and have a number of active links to one or more Ethernet switch at cluster level (or data center), this second level of switching can potentially cover more than ten thousand of individual servers. The IT equipment must be decided at design time. It is very important to consider the initial power and predict the future need, to make room for new updates or inclusion of new components and at the same time, to avoid exces-

sive or unnecessary spending. This is just one the things that must be taken into account, the greatest impact on data center costs regarding maintenance is the electricity consumption, including the one due to the cooling system: the bigger the size of the IT equipment, the greater the power required to cool the components. A power and cooling analysis, also referred to as a thermal assessment, measures the relative temperatures in specific areas of a data center, as well as the ability of the data center to tolerate specific temperatures. Among other things, a power and cooling analysis can help identify hot spots, over-cooled areas that can handle greater power use density, the breakpoint of equipment loading, the effectiveness of a raised-floor strategy, and optimal equipment positioning (such as AC units) to balance temperatures across the data center. The energy consumption of IT equipment, which perform useful work in a computer center, represent on average only 50% of the total consumption, since the remainder is divided between the conditioning system (30%) and all the power supply (Figure 3). For energy is therefore critical that the share of consumption related to power and conditioning is the lowest possible, since the activity of the computer center is represented by electronic devices. It is very important to understand the values and distribution of energy flows within the computer center. The measured values of the power required for various equipments are essential for assessing the energy efficiency of data centers.

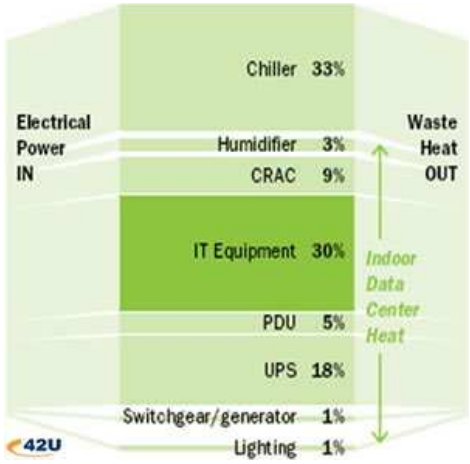


Figure 3:
Power Consumption in a data center

POWER/ENERGY METRICS

As discussed in the previous section, current data centers consume a tremendous amount of power. Power/energy metrics of various scales and various components for data center computing have been proposed and applied in data center computing practice. Figure 4 shows a summary of popular power related performance metrics. The Data Center Infrastructure Efficiency, DCiE or DCE [5], [6], is an industry accepted metric. It is defined as:

$$DCiE = \frac{IT\,Equipment\,Power}{Total\,Facility\,Power}$$

where *IT Equipment Power* includes the load associated with all of the IT equipment, i.e., computation, storage and network equipment, along with supplemental equipment such as switches, monitors, and workstations/laptops used to monitor or otherwise control the data center. *Total Facility Power* includes all IT Equipment power as described above plus everything that supports the IT equipment load such as: power delivery components i.e., UPS, switch gear, generators, PDUs, batteries and distribution losses external to the IT equipment; Cooling system components i.e., chillers, computer room air conditioning units, direct expansion air handler units, pumps, and cooling towers. Other miscellaneous component loads such as data center lighting are also included.

A DCiE value of about 0.5 is considered typical practice and 0.7 and above is better practice. Some data centers are capable of achieving 0.9

Figure 4:
Power consumption metrics taxonomy

or higher [29]. An organization for green computing provides the following standard metric for calculating the energy efficiency of the data centers, Power Usage Effectiveness (PUE) [5]:

$$PUE = \frac{1}{DCiE} = \frac{TotalFacilityPower}{ITEquipmentPower}$$

PUE shows the relation between the energy used by IT equipments and energy used by other facilities, such as cooling needed for operating the IT equipment. For example, a PUE of 2.0 indicates that for every watt of IT power, an additional watt is consumed to cool and distribute power to the IT equipment. At the present time, the PUE of a typical enterprise data center falls between the range [1 - 3].

The HVAC (Heating, Ventilation, and Air Conditioning) system of a data center typically includes the computer room air conditioning and ventilation, a large central cooling plant, and lighting and other minor loads. The HVAC System Effectiveness is the fraction of the IT equipment energy to the HVAC system energy. The HVAC system energy is the sum of the electrical energy for cooling, fan movement, and any other HVAC energy use like steam or chilled water. The HVAC System Effectiveness is calculated as follows:

$$HVAC\_Effectiveness = \frac{IT}{HVAC+(Fuel+Steam+ChilledWater)*293}$$

where IT is the annual IT Electrical Energy Use, $HVAC$ is the annual HVAC Electrical Energy Use, $Fuel$ is the annual Fuel Energy Use, $Steam$ is the annual District Steam Energy Use and $Chilled\ Water$ is the annual District Chilled Water Energy Use.

The HVAC System Effectiveness denotes the overall efficiency potential for HVAC systems. A higher value of this metric means higher potential to reduce HVAC energy use [13].

The metric of SWaP (Space, Watts and Performance) [20] characterizes a data center's energy efficiency by introducing three parameters of space, energy and performance together. The SWaP is calculated as follows:

$$SWaP = \frac{Performance}{(Space*PowerConsumption)}$$

where $Performance$ is computed using industry-standard benchmarks, $Space$ measures the height of the server in rack units (RUs) and $Power$ Determines the watts consumed by the system, using data from actual benchmark runs or vendor site planning guides.

The SWaP metric gives users an effective cross-comparison and total view of a server's overall efficiency. Users are able to accurately compare the performance of different servers and determine which ones deliver the optimum performance for their needs. The SWaP can help users better plan for current and future needs and control their data center costs.

The DCeP [14], [10] is proposed to characterize the resource consumed for useful work in a data center. The DCeP is calculated as follows:

$$DCeP = \frac{Useful\_work\_produced}{Total\_energy\_consumed\_to\_produce\_that\_work}$$

Useful work is the tasks performed by the hardware within an assessment window. The calculation of total energy consumed is the kWh of the hardware times the PUE of the facility.

MEASURING POWER CONSUMPTION

There are a variety of computer power consumption benchmarking techniques in practice

today. Even though many techniques are available, there is no universal *one size fits all* technique that is appropriate for every benchmarking situation. Different methods of benchmarking must be completed for different usage patterns. Additionally, measuring computer power consumption is different than general computer benchmarking because a tool is usually required to measure how much electricity is being consumed by a running machine. General computer benchmarks such as CPU or video benchmarks do not require any special tools. Since power consumption benchmarks require electricity measuring devices, it makes it harder for people to participate in the power consumption benchmarking area. The requirement and cost of a tool adds a burden to those who wish to run their own computer power consumption benchmarks. Here two electricity consumption meters are described, the affordable KILL A WATT meter and the more expensive Watts up? meter. There are two fundamental ways of measuring power consumption: measuring power consumption at one moment in time and measuring power consumption over time. Each method has its pros and cons. Measuring power consumption over one moment in time is useful when measuring a device that is using a constant amount of power. Less time is needed to take a sample when measuring a device just in a time instant, rather than when measuring a device over some period of time. A disadvantage of measuring over one moment in time is that if the device being measured has power fluctuations, it is not possible to get an accurate average of how much power the device uses. For example a refrigerator may use a lot of power when the compressor is activated but it may use very little power when the compressor is turned off. It may use a little bit more power when the refrigerator door is opened and the light turns on. The power consumption of a refrigerator fluctuates over the day. In order to get a better idea of how much power a refrigerator consumes, it is important to measure the load over some time period that represents typical refrigerator use patterns. In the case of a refrigerator it may be necessary to measure a weeks worth of use in order to get a clear picture of how much power the refrigerator consumes on average. Similarly computers fluctuate in their power consumption as they carry out various tasks. When comparing a computer's overall power usage, it is crucial to compare power usage over time [9]. Simply measuring a computer's power consumption when it is idling will not factor in how much power may be consumed while it has a load placed on it. A good way of comparing two systems is to compare real tasks on each system identically from start until finish [9]. This reveals actual power requirements for tasks. It also exposes differences between systems. Here is an example: "Let's say you are comparing system A that uses 45 watts of power under load and system B that uses 55 watts of power under load. If system B finishes the same task as system A in half time (because it has a better performing CPU), system B is actually a more power efficient system for the benchmarked task than system A even though it uses more power under load. Since it takes a lot less time to finish, less power is spent [9]". Although Idle benchmarks can be compared between different reviewers and magazines for identical systems, the load benchmarks depend on the specific tests conducted by the reviewers. If two reviewers have identical systems but have different benchmarking and measuring schemes, the load benchmarks cannot be compared. Luckily Idle should be comparable because in order to reach an idle state, the system has to be left doing effectively nothing or very little. The systems being compared should come installed with the same hardware, operating system, service packs, patches, soft-

Figure 5:
P3 International KILL A WATT

ware, and background processes. This would ensure that comparisons could be made between benchmarks performed on two different systems.

KILL A WATT Power Consumption Meter

KILL A WATT (Figure 5) is a simple meter made by P3 International [11]. In order to use KILL A WATT, it is required to plug it into the wall and then plug a device into KILL A WATT. It can be connected directly to the wall plug or can be connected to an extension cable or power strip which is then connected to the wall plug. It is easier to work with KILL A WATT with an AC extension cable so that it could be positioned closer to the experimenter for easier viewing. KILL A WATT has several modes including Volt (voltage), Amp (amperes), Watt (wattage), VA (volt-amperes), Hz (hertz), PF (power factor), KWH (kilowatt hours), and Hour (the length of time the device has been connected to a power source). The voltage tends to approximately 120 volts for most devices and appliances. Some heavy appliances like electric stoves and electric dryers use 240 volts of alternating current (VAC). KILL A WATT is not able to be used with 240 VAC wall plugs. The Watt mode displays how many watts are being consumed through the device at the moment of time the wattage value is recorded. This value tends to fluctuate with most devices as the electrical loads vary. The wattage is what we were mostly concerned with in this work. The Hz mode measures hertz or how many times the alternating current cycles per second (usually 60 times). The PF mode displays Power Factor. The KWH mode displays the amount of kilowatt hours consumed since the KILL A WATT meter was connected to a power source. This mode is very useful in determining how much power a device or appliance uses over some time period. The last mode is the “Hour” mode, which displays the amount of time the KILL A WATT meter has been connected to a power source. Measuring power use over time is preferred because this type of measurement will capture all the fluctuations of power usage as the electrical load varies. Measuring power consumption by looking at the amounts of watts used at any given moment in time is not representative of true power consumption of devices whose loads vary, such as computers.

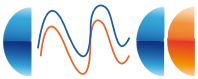
Watts Up? Power Consumption Meter The Watts up? power consumption meter (Figure 6) features the same functionality as the KILL A WATT meter and more. In addition to the KILL A WATT features, the Watts up? meter can record power measurements over time in watt hours sensitively enough to be used for computer power consumption measurements. In addition to recording the amount of power consumed over time, the meter comes with memory which saves the data that is recorded during measurements. The data can then be accessed by loading software from the watts

Figure 6:
Watts up? Power Consumption Meter

up? website [12] and connecting the meter to a Windows computer with an included USB cable. There are several Watts up? meter models available. Among them the Watts up? PRO es model, which includes extra memory for saving measurements that are taken over long periods of time. The Watts up? meter has several advantages over the KILL A WATT meter. It has more complex features than the KILL A WATT meter such as more sensitive measurements and memory for data storage. It can connect to a Windows machine with a USB cable and transfer data which can then be graphed and used for analysis. It has an extension cable that gives the user a lot of leeway to position the meter in a viewable spot. The watts up? power consumption meter's functionality made it a great tool for professional power consumption benchmarking. There are also several important disadvantages to note about the Watts up? meter. The Watts up? meter costs several times more money than the KILL A WATT meter, around 90-220 dollars as of the time of this writing. The Watts Up? meters are more complicated and might be harder for some people to learn how to use. Despite these drawbacks the watts up? meter proved to be a better tool for our research in order to carry out serious computer power consumption benchmarks. Without a watt hour measuring option on a power meter, it is not possible to accurately measure how much power a computer uses when its power fluctuates unless the benchmarks lasts for hours or days.

COMPUTER POWER CONSUMPTION BENCHMARKING

One area of the computing sciences that is becoming more and more important is computer power consumption benchmarking. Benchmarking is a general and widely known approach where a standard load, or benchmark, is used to measure some desired aspects, behavior or other characteristics of the thing being observed. In computing, benchmarks are typically computer programs that are run on a system, enabling accurate and repeatable measurement of some specific characteristic of interest. A performance metric is typically associated with a benchmark or a workload, as well as a well-defined execution environment. A benchmark is used to emulate real-world computing tasks. A data center benchmark is the act of running a set of programs, or other operations, in order to assess the relative performance of data center performance metrics. Typically data center benchmarking is carried out under a certain workload, which is artificially generated or produced in real life, according to the following workflow: real or artificial workload $\rightarrow$ benchmark $\rightarrow$ performance metrics. Although some benchmark workloads are available, for example, server-side Java under various loads for SPECpower [34], JouleSort



[39], data center benchmarking is generally taken under normal data center practice and workload. In this section, two types of benchmarks are described: server level benchmarks and data center level benchmarks. Server level benchmarks, like GCPI, Green500 and JouleSort, normally measure energy consumption in a compute server or a cluster. Data center level benchmarks, such as, HVAC Effectiveness Index and Computer Power Consumption Index from LBNL, present a measurement for the entire data center.

Computer server level benchmark This section introduces several performance metrics for computer server level performance evaluation: GCPI, SPECpower, Green500 project, and JouleSort. SiCortex [19] proposes an inclusive metric for comparing energy efficiency in the High-Productivity Computing segment. The Green Computing Performance Index (GCPI) [15] is an inclusive metric for comparing energy efficiency in High-Productivity Computing segment proposed by SiCortex [19]. The GCPI analyzes computing performance per kWatt across a spectrum of industry-standard benchmarks, providing organizations with much-needed guidance in the era of out-of-control data center energy consumption. It has been declared that the GCPI is the first index to measure, analyze and rank computers on a broad range of performance metrics relative to energy consumed. In 2006, the SPEC community started to establish SPECpower [34], an initiative to augment existing SPEC benchmarks with a power/energy measurements. SPECpower's workload is a Java application that generates and completes a mix of transactions; the reported throughput is the number of transactions completed per second over a fixed period. SPECpower reports the energy efficiency in terms of overall operations per watt. This metric represents the sum of the performance measured at each target load level divided by the sum of the average power (in watts) at each target load. The Green500 project [26] aims at increasing awareness about the environmental impact and long-term sustainability of high-end supercomputing by providing a ranking of the most energy-efficient supercomputers in the world. The Green500 list uses *performance per watt* (PPW) as its metric to rank the energy efficiency of supercomputers. The performance is defined as the achieved maximum GFLOPS (Giga FLoating-point OPerations per Second) performance, by the Linpack [16] benchmark on the entire system. The power is defined as the average system power consumption during the execution of Linpack with a defined problem size. JouleSort [39] is an I/O-centric benchmark that measures the energy efficiency of systems at peak use. It is an extension of the sort benchmark, which is used to measure the performance and cost-performance of computer systems. The JouleSort benchmark measures the cost of doing some amount of work, which reflects some measure of power use, e.g., average power, peak power, total energy, and energy-delay.

Data center level benchmarking Data center level benchmarks are used to evaluate how green a data center is and are also used to compare similar data centers. Lawrence Berkeley National Laboratory (LBNL) releases a set of efforts and practices for data center benchmarking [3]. These practices include improved air management, emphasizing control and isolation of hot and cold air streams; rightsizing central plants and ventilation systems to operate efficiently both at inception and as the data center load increases over time; optimizing central chiller plants designed and controlled to maximize overall cooling plant efficiency, central air-handling units in lieu of distributed units. Figure 7 shows the formal process for benchmarking

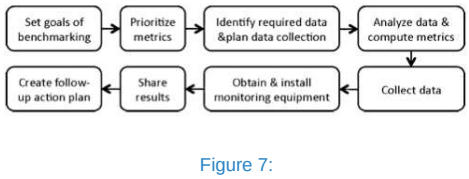


Figure 7:
Data center benchmarking process

a data center [13]. Main software packages for benchmarking and designing energy efficient buildings and data centers are: Calarch, Comis, DoE-2, EnergyPlus, Genopt, home energy saver.

CLUSTER POWER MANAGEMENT ALGORITHMS

Two basic power management mechanisms are dynamic voltage scaling and node vary-on/vary-off. In the following these mechanisms are briefly described. Also hybrid approaches that combine these techniques have been investigated.

Dynamic Voltage and Frequency Scaling One technique being explored is the use of Dynamic Voltage and Frequency Scaling (DVFS) within Clusters and Supercomputers [30], [31]. By using DVFS one can lower the operating frequency and voltage, which results in decreased power consumption of a given computing resource considerably. High-end computing communities such as cluster computing and supercomputing in large data centers, have applied DVFS techniques to reduce power consumption and achieve high reliability and availability [28], [24], [8]. A power-aware cluster is defined as a compute cluster where compute nodes support multiple power/performance modes, for example, processors with frequencies that can be turned up or down. This technique was originally used in portable and laptop systems to conserve battery power, and has since migrated to the latest server chipsets. Current technologies exist within the CPU market such as Intel's SpeedStep and AMD's PowerNow! technologies. These dynamically raise and lower both frequency and CPU voltage using ACPI P-states [2]. In [25], DVFS techniques are used to scale down the frequency by 400Mhz while sustaining only a 5% performance loss, resulting in a 20% reduction in power. A power-aware cluster supports multiple power and performance modes, allowing for the creation of an efficient scheduling system that minimizes power consumption of a system while attempting to maximize performance. When looking to create a DVFS scheduling system for a data center, there are a few rules of thumb to build a scheduling algorithm which schedules virtual machines in a cluster while minimizing the power consumption:

1. Minimize the processor supply voltage by scaling down the processor frequency;
2. Schedule virtual machines to processing elements with low voltages and try not to scale Processing Elements (PEs) to high voltages.

Based on the performance model defined above, Rule 1 is obvious as the power consumption could be reduced when supplied voltages are minimized. Then Rule 2 is applied: by scheduling virtual machines to processing elements with low voltages and trying not to operate PEs with high voltages to support virtual machines. A scheduling algorithm for virtual machines in a DVFS-enabled cluster [45] is shown in Figure 8 where incoming virtual machine requests arrive at the cluster and are placed in a sorted queue. This system has been modeled, created and described in [44]; in this Section we will explain in detail that scheduling mechanism. The scheduling algorithm runs as a daemon in a cluster with a predefined schedule interval, INTERVAL. During

Figure 8:
Working scenario of a DVFS-enabled cluster scheduling

the scheduling interval, incoming virtual machines arrive at the scheduler and will be scheduled at the next schedule round. For each virtual machine (VM), the algorithm checks the PE operating point set from low voltage level to high voltage level. The PE with lowest possible voltage level is found; if this PE can fulfill the virtual machine requirement, the VM can be scheduled on this PE. If no PE can schedule the VM, a PE must be selected to operate with higher voltage. This is done using the processor with the highest potential processor speed, which is then adjusted to run at the lowest speed that fulfills the VM requirement, and the VM is then scheduled on that PE. After one schedule interval elapses, some virtual machines may have finished their execution; thus this algorithm attempts to reduce a number of PE's supply voltages if they are not fully utilized.

Vary-on/Vary-off Node vary-on/vary-off (VOVO) takes whole nodes offline when the workload can be adequately served by a subset of the nodes in the cluster. Machines that are taken off-line may be powered off or placed in a low power state. Machines that are off-lined are placed back online should the workload increase. Node varyon/vary-off is an inter-node power management mechanism. This policy, originally proposed by Pinheiro et. al. [37], turns off server nodes so that only the minimum number of servers required to support the workload are kept active. Nodes are brought online as and when required. VOVO does not use any intra-node voltage scaling, and can therefore be implemented in a cluster that uses standard high-performance processors without dynamic voltage scaling. However, some hardware support, such as a Wake-On-LAN network interface, is needed to signal a server to transition from inactive to active state. Node vary-on/vary-off can be implemented as a software service running on one of the cluster nodes (or a separate support server) to determine whether nodes must be taken offline or brought online and requires software probes running on each of the cluster nodes to provide information on the utilization of that node. The load distribution mechanism must be aware of the state of the servers in the cluster so that it does not direct requests to inactive nodes, or to nodes that have been selected for vary-off but have not yet completed all received requests.

Given the features of the Euro-Mediterranean Centre for Climate Change (CMCC) clusters, at the moment it is not possible to apply DVFS algorithms. So, several Vary-on/Vary-off mechanisms will be investigated.

A PROPOSAL FOR A METERING SYSTEM FOR THE POWER CONSUMPTION OF CALYPSO

The Supercomputing infrastructure provided by CMCC consists of two cluster systems, which together reach a theoretical peak performance of more than 30 TFlops. These systems are complemented by an high-capacity and high-performance storage infrastructure that, overall, provides a useful storage space, which amounts to more than 450 Tbytes. The two

supercomputing clusters differ in the type of architecture: the first one is a vector system consisting of NEC SX-8R series nodes and SX-9 nodes, while the second one is a scalar system consisting of 30 IBM POWER 6 nodes divided into 3 frames. Each node contains 16 Power 6 dual-core processors, operating at a frequency of 4.7 GHz. Therefore, the number of physical cores amounts to 960 cores, capable of providing a total computing power (theoretical peak performance) of about 18 TFLOPS. To ensure the continuity of the power to such systems - and the consequent preservation of data - the CMCC has a 500KVA Static UPS power system.

The power panel is already prepared for a second parallel UPS able to provide total protection for 1000kVA for future additions (Figure 9). The CMCC has a dedicated transformer substation from which three lines of service reach the CMCC. A 30 kW line is dedicated to the ordinary user, another of 350 kW is dedicated to users of air conditioning, and the third one - the most important - is the preferred line dedicated to computer systems. The presence of a single supply line and the presence of some components of redundant cooling allow classifying the CMCC Supercomputing Center as a Tier II data center. Inside the CMCC there are two main areas (Figure 10) well separated and with well defined roles.

In the first environment called *Control Room*, all the hosted electrical panels supply power to various power lines of computer systems, storage and services. The system cooling is located in the second environment (greater than 200 square meters), called *Computing Room*. Inside the Control Room there are several electrical panels whose main task is to distribute electrical power to computer systems and storage. There are three main power supply lines: the power supply line of NEC systems, the power supply line of IBM systems and, finally, a supply line dedicated to services. The main control panels of each line are equipped with meters which record the total consumption. Regarding Calypso the power of each individual frame is managed by four three-phase lines (12 in total) (Figure 10).

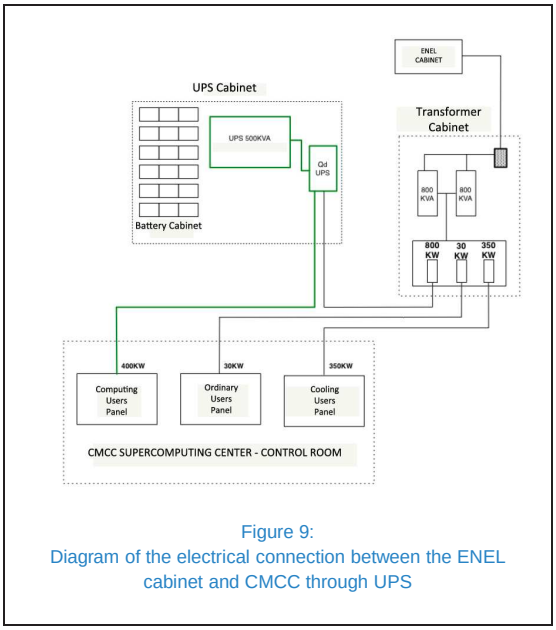


Figure 9:
Diagram of the electrical connection between the ENEL cabinet and CMCC through UPS

Since we intend to analyze the energy consumption of Calypso, a possibility is to perform the metering of the consumption of each IBM (frame) rack by measuring the three-phase supply. Every IBM frame provides 4-phase electrical connections that feed to 12 PDUs. These PDUs in turn feed the IBM P575 computing nodes in addition to two control units. Hence, it is possible to place a power panel in the Computing Room for measuring the energy consumption of each of the 4 lines per frame, thus obtaining a measurement of the total consumption of each IBM Calypso rack (Figure 11).

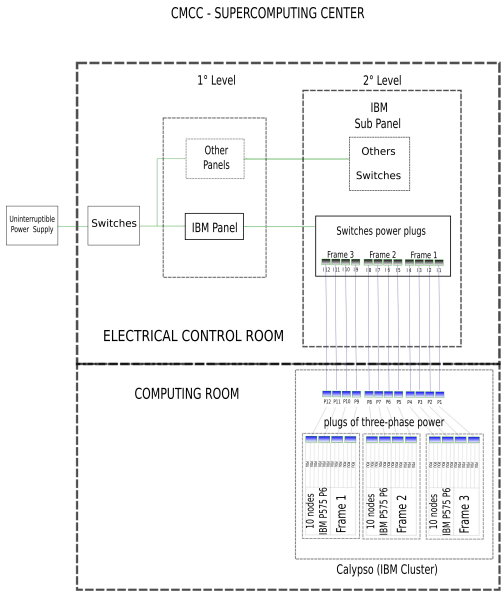


Figure 10:
Scheme of electrical connection between the control room and Calypso

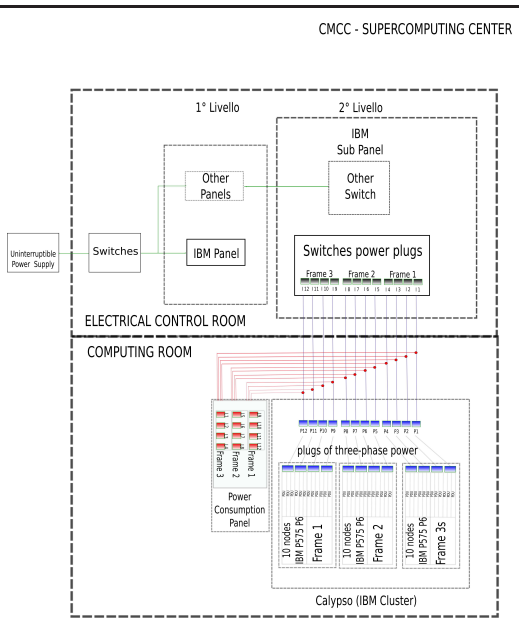


Figure 11:
Scheme of electrical connection between the control room and Calypso with insertion of a panel dedicated to measure the power consumption of each frame

Bibliography

[1] High Performance Green Computing, A Proposal for New Investments in Faculty Hiring in the Departments of Astronomy and Computer Science. http://www.umass.edu/oapa/oapa/proposals/high_performance_green_computing.pdf.

[2] Data Center Power Management and Benefits to Modular Computing, 2003. Intel Developer Forum.

[3] Best practices for data centers: lessons learned from benchmarking 22 data centers, 2006. Technical report, Lawrence Berkeley National Laboratory.

[4] Estimating total power consumption by servers in the US and the world, 2007. Tech. rep., Lawrence Berkeley National Laboratory.

[5] The green grid data center efficiency metrics: PUE and DCIE, 2007. Technical report, The Green Grid.

[6] The Green Grid metrics: data center infrastructure efficiency (DCiE) detailed analysis, 2007. Technical report, The Green Grid.

[7] The green grids opportunity: decreasing datacenter and other IT energy usage patterns, 2007. Technical report, the green grid.

[8] The Green500 List: Encouraging Sustainable Supercomputing, 2007. IEEE Computer Society.

[9] The Truth About PC Power Consumption, 2007. Tom's Hardware, Bestofmedia Network.

[10] A framework for data center energy productivity, 2008. Technical report, The GreenGrid.

[11] P3 - Kill A Watt, 2008.

[12] Watts Up? Products: Meters, Electronic Educational Devices, 2008.

[13] Self-benchmarking guide for data centers: metrics, benchmarks, actions, 2009. Technical report, Lawrence Berkeley National Laboratory, Berkeley, California.

[14] 42U DCeP: Data Center energy Productivity, 2011.

[15] Green Computing Performance Index, 2011. http://www.sicortex.com/green_index.

[16] Green Computing Performance Index, 2011. http://www.netlib.org/linpack/.

[17] IEEE TCSC technical area of green computing, 2011.

[18] Report to Congress on Server and Data Center Energy Efficiency, 2011. www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf.

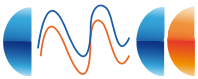
[19] SiCortex, 2011. http://www.sicortex.com/.

[20] SWaP (Space, Watts and Performance) Metric, 2011.

[21] L. A. Barroso and U. Holzle. The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines. *Synthesis Lectures on Computer Architecture*, 1(1):1–108, 2009.

[22] Y Chen, A Das, W Qin, A Sivasubramaniam, Q Wang, and N Gautam. Managing server energy and operational costs in hosting centers. In *SIGMETRICS'05: proceedings of the 2005 ACM SIGMETRICS international conference on measurement and modeling of computer systems*, pages 303–314, 2005.

[23] R Das, JO Kephart, C Lefurgy, G Tesauro, DW Levine, and H Chan. Autonomic multi-agent management of power and performance in data centers. In *AAMAS Õ08:*



proceedings of the 7th international joint conference on autonomous agents and multiagent systems, pages 107–114, 2008.

[24] W. Feng, A. Ching, C.H. Hsu, and V. Tech. Green Supercomputing in a Desktop Box. In *IIn IEEE International Parallel and Distributed Processing Symposium*, pages 1–8, 2007.

[25] W. Feng, X. Feng, and R. Ge. Green Supercomputing Comes of Age. *IT PROFESSIONA*, 10:17, 2008.

[26] WC Feng and KW Cameron. The Green500 list: encouraging sustainable supercomputing. *Computer*, 40:50–55, 2007.

[27] VW Freeh and DK Lowenthal. Using multiple energy gears in mpi programs on a power-scalable cluster. In *Pingali K, Yelick KA, Grimshaw AS (eds) PPOPP. ACM Press, New York*, pages 164–173, 2005.

[28] I. Gorton, P Greeeld, A Szalay, and R. Williams. Data-Intensive Computing in the 21st Century. *Computer*, 41:30–32, 2008.

[29] S Greenberg, A. Khanna, and W Tschudi. High performance computing with high efficiency. *AmSoc Heat Refrig Air-cond Eng (ASHRAE) Trans*, page 179, 2009.

[30] Chung hsing Hsu and Wu chun Feng. A feasibility analysis of power awareness in commodity-based high-performance clusters. In *In Proceedings of IEEE International Conference on Cluster Computing*, pages 1–10, 2005.

[31] C. Hsu and W. Feng. A power-aware runtime system for high-performance computing. In *In Proceedings of the 2005 ACM/IEEE conference on Supercomputing*, 2005.

[32] N Jiang and M Parashar. Enabling autonomic power-aware management of instrumented data centers. In *IPDPS Õ09: proceedings of the 2009 IEEE international symposium on parallel & distributed processing*, pages 1–8, 2009.

[33] G Jung, KR Joshi, MA Hiltunen, RD Schlichting, and C Pu. A cost-sensitive adaptation engine for server consolidation of multitier applications. In *ACM/IFIP/USENIX 10th international middleware conference*, pages 163–183, 2009.

[34] KD Lange. Identifying shades of green: the specpower benchmarks. *Computer*, 42:95–97, 2009.

[35] S Mehta and A Neogi. Recon: a tool to recommend dynamic server consolidation in multi-cluster data centers. In *IEEE/IFIP network operations and management symposium: pervasive management for ubioquitous networks and services*, pages 363–370, 2008.

[36] CD Patel, R Sharma, CE Bash, and A Beitelmal. Thermal considerations in cooling large scale high compute density data centers. In *Proceedins of Thermal and thermomechanical phenomena in electronic systems - ITHERM 2002*, pages 767–776, 2002.

[37] E. Pinheiro, R. Bianchini, E.V. Carrera, and T. Heath. Load Balancing and Unbalancing for Power and Performance in Cluster-Based Systems. In *In Workshop on Compilers and Operating Systems for Low Power*, pages 180–195, 2001.

[38] S. Rivoire, MA Shah, P Ranganathan, and J Meza C. Kozyrakis. Models and metrics to enable energy-efficiency optimizations. *Computer*, 40(12):39–48, 2007.

[39] S Rivoire, MA Shah, P Ranganathan, and C Kozyrakisr. Joulesort: a balanced energy-efficiency benchmark. In *SIGMOD'07: proceedings of the ACM SIGMOD international conference on management of data*, pages 365–376, 2007.

[40] A Shah and N Krishnan. Optimization of global data center thermal management workload for minimal environmental and economic burden. *IEEE Transactions on Components and Packaging Technologies*, 31(1):39–45, 2008.

[41] A Singh, MR Korupolu, and A Mohapatra. Server-storage virtualization: integration and load balancing in data centers. In *Proceedings of the ACM/IEEE conference on high performance computing*, page 53, 2008.

[42] Q Tang, SKS Gupta, and G Varsamopoulos. Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: a cyber-physical approach. *IEEE Trans Parallel Distrib Syst*, 19(11):1458–1472, 2008.

[43] DC Vanderster, A Baniasadi, and NJ Dimopoulos. Exploiting task temperature profiling in temperature-aware task scheduling for computational clusters. In *Asia-Pacific computer systems architecture conference*, pages 175–185, 2007.

[44] Gregor von Laszewski, Lizhe Wang, Andrew J. Younge, and Xi He. Power-Aware Scheduling of Virtual Machines in DVFS-enabled Clusters. In *IEEE Cluster 2009*, 2009.

[45] Andrew J Younge, Gregor Von Laszewski, Lizhe Wang, and Geoffrey C Fox. Providing a Green Framework for Cloud Based Data Centers. In *The Handbook of EnergyAware Green Computing*, 2011.