

Energy-Efficiency in Cloud Computing Environments: Towards Energy Savings without Performance Degradation

Ismael Solis Moreno, Jie Xu
University of Leeds, UK

ABSTRACT

Due to all the pollutants generated by it and the steady increases in its rates, energy consumption is causing serious environmental and economic problems. In this context, the growing use and adoption of ICTs is being highlighted not only as one of the principal problem sources but also as one of the principal areas that could help in the problem's reduction. Cloud computing is an emerging model for distributed utility computing and is being considered as an attractive opportunity for saving energy through central management of computational resources. For this to be successful, the design of energy-efficient mechanisms must start to play a major role. This paper argues about the importance of energy-efficient mechanisms within cloud data centers and remarks on the significance of the "energy-performance" relationship in boosting the adoption of these mechanisms in real scenarios. Principally, it provides an analysis of the current approaches and the outline of key opportunities that need to be addressed to improve the "energy-performance" relationship in this promising model.

Keywords: energy-efficiency; green computing; energy-aware; cloud computing; cloud computing challenges; energy-performance

INTRODUCTION

Nowadays, many people are devoted to a widespread adoption of Information and Communications Technologies (ICTs). However, due to the priorities of both providers and consumers, this has been focused principally on aspects such as processing speed, bandwidth, transfer rate, storage and memory capacity just to mention only a few, the environmental impact of their use has been relegated until recent years, when changing climate patterns and pollution problems have become high priority in the world's nations' agendas.

The increasing accumulation of greenhouse gases is changing the world's climate, creating serious problems such as droughts, floods and higher temperatures. In order to stop the accumulation of these gases in the atmosphere, it is necessary to stop the global growth of emissions, in which the generation of electricity plays a major role not only because of the carbon dioxide which results from the coal and oil used in this process, but also because it releases sulphurs and other pollutants into the atmosphere.

Additionally to the ICTs environmental repercussions, the worldwide economy is also being affected by the steady increases in electricity rates. The number of "smart" devices, peripherals, computers, data centers and the amount of communications are rapidly growing along with the electricity cost required to feed them. This problem is more perceptible within the industries and enterprises which have to support large

amounts of computing infrastructure normally represented by enormous data centers provided with powerful cooling systems that also require great amounts of energy to work.

In this context, cloud computing an emerging model for distributed utility computing, is becoming commercially attractive and its use is growing since it promises reducing the maintenance and management costs in comparison with traditional data centers. Clouds are normally composed by large and power-consuming data centers designed to support the elasticity and scalability required by its customers.

However, and despite that one of cloud computing commercial credentials is the reduction of energy consumption for customers, it still represents a serious problem for providers who have to deal with increasing demand and performance expectations. This creates the need for mechanisms to improve the energy-efficiency of cloud computing data centers while preserving desired levels of operation.

Green IT emerges as a new perspective for designing, developing and managing computing infrastructure aiming for more efficient processes and mechanisms to avoid waste of resources and considering the environmental implications of its use and disposal. Regarding with energy efficiency, a branch of Green IT named Energy-aware computing which is normally applied in embedded systems where strong energy constraints exist, has come forward to change the high-level computing systems point of view from “*performance-mainly*” to “*performance-energy*” balanced systems reducing the cost by an improved use of resources and the impact to the environment by diminishing the energy consumed while QoS is maintained.

Currently, some approaches have arisen to contribute to energy-efficiency improvement for data centers. Specifically, cloud computing approaches are exploiting the advantages of virtualization technology to maximize the use of underlying physical resources, dynamically resizing computing power in proportion to the customers’ requirements. However, these approaches are more focused in the dynamic aspects of the VM management life-cycle. They neglect fine-grained characteristics that in real cloud computing scenarios could lead to QoS and energy savings degradations. These characteristics represent challenges that should be addressed to boost the adoption of these mechanisms in real scenarios where customer satisfaction has priority.

This paper argues about the importance of energy-efficient mechanisms within cloud data centers and remarks the “energy-performance” relationship significance. First, it describes how ICTs are negatively impacting the environment. Then green and cloud computing are introduced. Finally, the importance of energy-efficient mechanisms in cloud computing, the analysis of current approaches and the identified opportunities in this area are presented.

ENVIRONMENTAL IMPACT OF ICTs

It is probably not a perceptible problem for most users, but ICTs affect the environment in different ways. According to (Murugesan, 2008) each of the stages of a computer’s life, from its production, use and disposal produces environmental problems. Among these problems, the excessive electrical power consumption by hardware such as servers, networks, monitors and cooling systems appears to be the most critical since it results in increased greenhouse gas emissions. However, the pollution produced during the manufacturing of computing equipment and all the e-waste generated during its disposal should be taken in consideration in order to mitigate where possible the environmental impact of the ICTs helping to construct a more sustainable environment.

According to the results of the Smart 2020 report mentioned in (Smarr, 2010), it was estimated that the ICT Industry contributed about 2 percent of the total global greenhouse gas emissions generated in 2007 and also that these will grow at a rate of approximately 6 percent per year, even assuming successful efforts to lower the industry's carbon intensity over the next decade. This means that the total emissions will roughly triple between 2002 and 2020.

In (Hickey, 2008) Gartner Research Vice President Simon Mingay, mentions in accordance with the 2020 report that the global amount of carbon dioxide emissions needs to be reduced from 60 to 80 percent by 2050. But more immediately, a 25 percent reduction is necessary by 2020 in order to diminish the environmental effects. Reducing the footprint generated by the ICTs will play a major role in the achievements of these goals. As mentioned by the Climate Group in (Webb, 2008), ICTs could save 7.8GtCO₂e or 15 percent of the global emissions in 2020.

Use of hazardous materials and e-waste generation

Electronic waste is becoming a serious fast-growing worldwide problem. Most unwanted computers and electronic equipment end up in landfills. In (Murugesan, 2008) it is mentioned that analysts predict that two-thirds of the estimated 870 million PCs manufactured in the world in the next five years will end up in landfills. The United Nations Environment Program has also estimated that 20 to 50 million tons of e-waste are generated worldwide each year and this number is increasing. Studies mentioned in (Schneider, 2010) and (Chickowski, 2007) show that this situation is more evident in some industrialized countries such as U.S. and some European nations where it has been demonstrated that the e-waste is rapidly growing in comparison with other kinds of municipal trash.

Beyond the amount of e-waste generated, the real problem is induced because some of the computer components contain toxic materials such as lead, chromium, cadmium and mercury as described in (Murugesan, 2008). If all the ICT residues in landfills are put into the ground, these toxic materials can filter dangerous chemicals into the waterways and the environment. Furthermore, if the material in the landfills is burned, toxic gases are released into the atmosphere polluting the air and contributing to the changes in climate patterns and global warming.

In (Chickowski, 2007), it is mentioned that the e-waste management practice of many developed countries is to send it to developing countries where very low-cost labor is used to split it apart and recover components, generally with low safety conditions for the workers and absolutely no regard for the environment in the local area. This irresponsible handling has generated many serious environmental problems for these countries and gradually for the entire world.

Energy consumption by computing and cooling systems

From the environmental perspective, the growing energy consumption becomes a serious problem not only because of the carbon dioxide that results from the coal and oil used in this process but also because it releases sulphurs and other pollutants into the atmosphere. Along with the International Energy Agency (IEA) cited in (World Coal Institute, 2010), coal is the first source of energy worldwide being used to generate about 41 percent of global electricity. It is set to continue, with coal feeding 44 percent of global electricity in 2030.

The energy consumption for computing could be divided according its use in two edges, the first regarding to the energy consumed by the clients conformed by PCs,

peripherals and all types of mobile devices and the second refers to the energy consumed by servers, networks and cooling systems in data centers.

Regarding the “*client edge*”, energy consumption represents a serious problem because of the rising adoption of computers and mobile devices worldwide. In (Smarr, 2010), according to the Smart 2020 report results, it is mentioned that 57 percent of the total CO₂ emissions relating to ICTs will be produced by this sector. This is because in contrast with data centers, the “*client edge*” normally presents a lack of policies and rules to enforce the management and reduction of energy consumption relying only on end-user’s responsiveness to activate and consciously use the energy-saving mechanisms installed in their computing infrastructure. The fact is that the most of the energy consumption in this edge occurs when PCs are idle, wasting on average 85 watts/hour even with the monitor off. This occurs mostly because of the need to be continuously available on the network or simply because the users do not take care about the importance of saving energy; they prefer computers to idle instead turning them completely off in order to avoid the time used for restoring their work environments.

On the other hand, due to the need to maintain the quality of service that customers expect and the continuous expansion of the industry, energy consumption in the “*data center edge*” is increasing along with their performance increase and the rising number of them in the world as can be seen in the information presented by Sun Microsystems in (See, 2008). This increment will represent 18 percent of the total amount of CO₂ emissions generated by the ICTs for the next years. However, this number could be bigger if we consider the issue that one of the most important directions for reducing energy consumption in the “*client edge*” is moving the workloads to controlled environments such as data centers for reducing the consumption in the latter but increasing it in the former one. Additionally, processing within this type of environments always comes along with other, but not less important factor represented by cooling systems. According to (Patel, Bash, Belady, Stahl, & Sullivan, 2001), these are necessary to maintain good levels of performance due to the large amounts of waste heat generated by the massive allocation of computing infrastructure. Moreover in (Wang, 2007), it is mentioned that depending on the size, design and management of a data center the addition of cooling systems could double or triple the energy consumed, thus the environmental impact.

ECONOMICAL IMPACT OF ENERGY CONSUMPTION

Beyond the environmental issues, the growing energy consumption starts representing an economical concern in both client and data center edges because of the steady increases in electricity rates. In (Nordman & Christensen, 2010), it is mentioned that in developed nations such as U.S. the use of PCs is generating an electricity bill about \$7 billion per year plus several billion dollars more for displays. Additionally, (See, 2008) presents an electricity cost forecast which takes as its baseline the \$18,5 billions spent for supply data centers during 2005 and considers three different trends: the first is the servers growing rate at 14% a year along with U.S Energy Information Administration; the second is the increase per server consumption at 16% a year in accordance with Sun primary research; and the third is the increase in electricity cost at 12% per year provided by the U.S and Worldwide Server Installed Base 2006-2009 forecast, giving as a result that the energy costs for data centers could grow to \$250 billion worldwide for 2012.

This rise on the energy prices combined with dynamic markets and high customer demand has led energy costs to be almost 30 percent of the total operation budget for some data centers. In accordance with (Freeman, 2009), this might result in the cost to power IT exceeding its acquisition cost in a matter of years limiting business' capacity to grow and change to support customer demands.

GREEN COMPUTING

With the aim of minimizing the negative environmental impact of ICTs, emerges a different perspective to perform and use computing infrastructure named "*Green Computing*" or "*Green IT*". Its principal objective is to find a balance between good QoS levels and a diminished impact to nature resulting in a sustainable eco-friendly computing environment.

In (Naditz, 2008), Simon Mingay defines Green IT as "*The optimal use of information and communication technology for managing the environmental sustainability of enterprise operations and the supply chain, as well as that of its products, services, and resources throughout their life cycles*". However, green IT can be explained since many different perspectives depending on the job position and the interests of who defines it. Regarding this, (Molla, 2008) presents a table which contains the definitions of green computing given by experts in different sectors of the ICTs. The authors classify these definitions in four different but interconnected perspectives which include "*sourcing perspective*" related to the environmentally preferable IT purchasing; "*operation perspective*" which includes improving energy-efficiency for computing and cooling systems; "*service perspective*", that refers to the role of IT in supporting a business' overall sustainability initiatives; and "*end of IT life management perspective*", related to conscious e-waste disposal.

Beyond environmental benefits, the adoption of green computing practices and technologies result in economic and other rewards for individuals and enterprises. In (Murugesan, 2008) it is mentioned that some of these benefits could include savings in energy costs, improved systems performance, space savings in data centers, and the improvement of public image.

Energy-aware computing

Regarding energy-efficiency, a branch of Green IT named energy-aware computing has came forward to change the high-level computing systems point of view from "*performance-only*" to "*performance-energy*" balanced systems, reducing the cost through an improved use of resources and impact to the environment by diminishing the energy consumed while QoS is maintained. Energy-aware computing is a paradigm that intends to fill the gap between performance and energy waste by providing more management levels and application driven adaptability. As mentioned in (Couch & Kumar, 2008), "*The goal of energy-aware computing is not just to make algorithms run as fast as possible, but also to minimize energy requirements for computation, by treating it as a constrained resource like memory or disk*".

Although it is currently a trendy term, energy-aware computing is not new and it has been widely used in hardware design contexts. It emerged at circuit-level where the advent of portable and small-sized computer systems have created enormous energy constraints to increment the battery life duration. However, because growing adoption and use of ICTs have been indicated as one of the principal contributors in energy consumption (Webb, 2008), the term "*energy-aware computing*" is no longer exclusive to the circuit level and has risen to include computer and data center components. The

designers left to be only concerned about the energy consumed by circuit blocks to extend batteries' life duration and started thinking about improving energy-demanding computer components such as CPUs, monitors, memory cards and networking equipment where the amount of energy consumed represents environmental and economical problems greater than just "*drained energy sources*".

As can be observed, since its beginning energy-aware computing has been strongly related to hardware improvements. Nevertheless, these are just tools that if not well utilized or managed could result in weak or zero energy efficiency enhancements increasing the total cost of ownership (TCO). This is because as mentioned in (Carter & Rajamani, 2010), generally speaking idle or underutilized resources consume considerable energy, often almost as much as they consume when they are active, resulting in a non-negligible amount of waste.

In order to take advantage of all these hardware improvements, the development of policies and software to administrate and maximize the use of those resources has begun to play a major role. Currently, some manual strategies such turning on equipment when not in use, adjusting sleep mode and power settings in client computers, using energy monitoring software and others described in (Murugesan, 2008), (Schneider, 2010) and (Naditz, 2008) are been applied by managers in order to reduce their ICT infrastructure energy consumption. Additionally, some other specialized software-related approaches concentrated in three main categories: software development optimization, energy efficient network protocols, and virtualized data centers are emerging to enhance resources use while maintaining expected QoS levels.

Among all these current strategies and approaches, the implementation of virtualization in data centers is becoming one of the most important technologies for reducing the infrastructure cost including energy. In accordance with (Evoy & Schulze, 2008), this importance relies on the virtualization software's capability for handling and abstracting the details of sharing hardware resources with other instances maximizing their utilization, thus improving their efficiency. Additionally In (IBM, 2009), it is mentioned that there are substantial benefits for those companies or entities which implement virtualization, such as the reduction of the hardware cost and its operation; improvements in the use of resources; flexibility; responsiveness; security; and increase in the availability for disaster recovery. Because of this, the use of virtualized environments such as cloud computing data centers and virtual desktop platforms could be seen as good approaches to impulse mechanisms such as "*thin computing*" which aims to reduce the energy consumption in the "*Client Edge*" moving the workloads from an uncontrolled environment to a controlled one as data centers where rules and energy-aware mechanisms can be better applied.

As observed, energy-aware computing is not only related with the development of high-technology electronic devices, but is also an issue related to development and design of efficient software, networking protocols, and mechanisms to improve high-performance data center resource utilization.

CLOUD COMPUTING

Defined by The National Institute of Standards and Technology (NIST) in (Amrhein, Anderson, & de Andrade, 2010), "*cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction*". Additionally in (Vaquero, Rodero-Merino, Caceres, & Linder, 2009),

cloud computing is defined as “*a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers*”. However, beyond these technical definitions cloud computing is associated with a new business paradigm for distributed utility computing that change the infrastructure’s location to the network reducing the hardware and software management expenses for customers applying a “*pay-per-use*” economic model.

On need for cloud computing energy-aware mechanisms

In its basic sense, cloud computing is a client-server architecture composed by large and power-consuming data centers designed to support the elasticity and scalability required by consumers. It is becoming attractive and its use is growing since it promises cost reductions for customers in comparison with permanent investments for traditional data centers. In (Gain, 2010) referencing Gartner Research Firm, it is mentioned that “*cloud computing services revenue should total \$56.3 billion for 2009, representing a 21.3% increase compared to 2008. The market is expected to explode to \$150.1 billion in 2013*”. Additionally in accordance with (Maggiani, 2009), its use to support the increasing market of social networking and personal data storage is boosting its adoption among the general public. Both consumption scenarios are creating a huge infrastructure and energy demand that represents a challenging problem for coming years because of the environmental and economical implications discussed earlier.

Cloud computing is also considered a good platform to boost “*thin computing*” (Velte, Velte, & Elsenpeter, 2010) to reduce the energy consumption at “*client edge*” moving the loads from an uncontrolled environment to a controlled one embodied by data centers, where strong policies and energy-aware mechanisms can be applied. This could represent a non-negligible energy consumption increment for this edge. Because the data center is one of the key areas for reducing the CO₂ footprint related to ICTs identified in the smart report 2020 along with the client edge and communications, the development of energy-aware mechanisms for cloud computing is closely related to the achievement of the 2020 target emissions reduction described in the same report. Moreover, energy savings in data centers is one of the most important concerns in the industry and enterprise sector becoming in sometimes a strong constraint for business expansion and economic growth.

Although one of the cloud computing commercial credentials is the reduction of energy consumption for clients as is mentioned in (Velte, et al., 2010), it still represents a serious problem for providers who have to deal with increasing demand and performance expectations. However, due to the business context that enclose cloud computing where customer satisfaction has priority, the design and development of energy-aware mechanisms becomes a non-trivial task. Saving energy without considering QoS implications could lead to a weak adoption or failure of this type of distributed service model. This is because in accordance with (Weiss, 2007) and (Erdogmus, 2009), the success of cloud computing demands a high degree of trust; terms such as privacy, ownership, availability and of course performance become very important to boost the user adoption and growing demand. All this creates the need for mechanisms to improve energy-efficiency at cloud computing data centers preserving at the same time the desired levels of operation.

CURRENT APPROACHES FOR ENERGY EFFICIENCY IN CLOUD DATA CENTERS

The current energy-efficiency techniques for cloud computing data centers are closely linked with distribution and scheduling algorithms and can be divided into two main categories: power and workload distribution.

Power distribution approaches look for an efficient fixed power allocation to maximize the data center performance ensuring the agreed response times. Basically, this kind of approach varies the power delivered to the servers for incrementing or decrementing their number, maintaining the same power peak. This resource pool resizing could be dynamically led by load behavioral patterns or specific pre-established schedules in order to accomplish customer expected performance levels. However, and according to (Gandhi, Harchol-Balter, Das, & Lefurgy, 2009), due to uncontrolled variables encountered in power distribution such as the outside arrival rate, weather, power to frequency relationship inherent in the technology, and the minimum power consumption of a server together with the advantages provided by virtualization for workload management, workload distribution approaches have led the trends in energy-efficiency for cloud computing data centers.

In essence, workload distribution approaches try to improve energy-efficiency by resizing the resource pool according to client requirements using workload consolidation and live migration. Based on the mechanism used to induce the energy waste reduction, these could be classified as “*dynamic processor scaling*” (DPS) and “*dynamic server’s pool resizing*” (DSPR).

In “*dynamic processor scaling*”, energy savings are gained by adjusting the operating clock to scale down the supply voltages for the circuits. This is primarily achieved using slack reclamation with the support of dynamic voltage/frequency scaling incorporated into many recent commodity processors. However, this clearly depends on the hardware component settings, which are not available in all architectures and only represents a voltage reduction which means there still exists an energy drain. On the other hand, “*dynamic server’s pool resizing*” promises the most power savings, as they ensure near-zero electricity consumption by being turned-off idle or low-utilized servers using technologies such as Wake-On-LAN (WOL). However, (Duy, Sato, & Inoguchi, 2010) mentions that these approaches have had difficulties in assuring service-level agreements due to the lack of a reliable tool for predicting future demand and weak distribution policies to assist the turning off/on decision-making process.

Although there exist some other approaches related with the reduction of energy consumption for cloud computing such as those presented in (VMware, 2009), (Ley, Bianchini, Martonosi, & Nguyen, 2009), (Nathuji, Isci, & Gorbato, 2007) and (Li et al., 2009), this paper only introduces those that include in their methodology not only energy savings but also performance preserving mechanisms and evaluations (see Table 1). While these approaches represent progress in achieving a balance between energy and performance, they still make assumptions that in real cloud scenarios could result in drawbacks. These assumptions include among others: virtual machine communication, hardware and workload heterogeneity, and the implications of aggressive workload consolidation. These should be addressed to achieve the adoption of energy-aware mechanisms in cloud environments where energy savings are required while maximizing the performance for the customers’ satisfaction.

This represents a big challenge since the complexity of cloud computing as a single entity where many variables are involved (technological and commercial) and sometimes the adjustment or improvement of determined parameters could result in the

disarrangement of others in addition with its dynamic evolution based on market demands. Dynamicity and flexibility seem to be the key, especially for energy-efficiency in cloud environments. The first step is achieved by improving the use of resources through VM consolidation and live migration, minimizing the energy waste generated by idle servers. However in accordance with (Cameron, 2010), where it is said that “*The first generations of power-management hardware and software have saved energy often at the cost of performance...*”, for cloud computing it is necessary to start exploiting its fine-grained characteristics that in addition to current resource improvement mechanisms could achieve the balance and flexibility required for the success of energy-aware mechanisms characterized by environmental and economic benefits.

Power-aware scheduling of virtual machines in DVFS-enabled clusters

In (Von Laszewski, Wang, Younge, & He, 2009) it is described a scheduling mechanism which aims to reduce the power consumption in virtualized clustered environments by dynamically reducing processor speeds. The mechanism presented is composed of three algorithms that work together in order to allocate workloads in a virtualized cluster based on the required and available processor speed in the underlying physical nodes. The algorithms continuously monitor the VMs’ status to adjust the processor speed on each node, reducing the power consumption. To achieve that, this approach uses profiles describing the available and maximum processor speed for each server in the cluster.

In aiming to maintain the performance levels, the Xen hypervisor performance governor is set to user space in order to enable the manual control of the frequencies according to the workload requirements. Additionally, the performance evaluation of varying the number of VMs and operating frequencies is presented. Here, nBench -a Linux CPU benchmark- is used to simulate intensive computing jobs and measure the CPU performance at the same time.

However, the performance results in this approach are never correlated with the energy reduction obtained. This makes it difficult to find out the level of the energy-performance balance achieved. Additionally, even though the authors mention heterogeneous clusters, the explanation of how this characteristic affects the energy-performance improvements is weak. Moreover, they assume only one type of workload with fixed behavior. This is not necessarily true in a real cloud scenario where different behavioral pattern applications can live together according to (Abdelsalam, Maly, Mukkamala, Zubair, & Kaminsky, 2009). Finally, no virtual machine live migration is used to reduce the number of servers, thus the energy savings rely only on the DVFS technology which according to (Duy, et al., 2010) is highly depending and varies on different hardware architectures.

Energy-efficient management of data centre resources for cloud computing: a vision, architectural elements, and open challenges

(Buyya, Beloglazov, & Abawajy, 2010) presents a conceptualization for the middleware layer between users and resources in a cloud environment with the aim of achieving energy savings while minimizing QoS degradation. To this end, the authors propose an architecture which is composed of 4 main components that include: the clients; a middleware layer named Green Service Allocator; the VMs; and the physical layer. They describe the Green Service Allocator as composed of a set of algorithms and monitors that work together to improve the workloads’ distribution within cloud data centers for reducing power consumption while maintaining SLAs.

They propose the use of a bin packing algorithm named Best Fit Decreasing (BFD) with some modification allowing them to sort all VMs in decreasing order of current utilization and allocate each VM to a host based on a policy of least increase of power (LEPC) consumption. Finally, they handle the optimization of current allocation in two steps: first, VMs that need to be migrated are selected and second, the chosen VMs are placed on available hosts using the Modified BFD algorithm. All this process is supported by an energy monitor which turns on/off servers in proportion to the workload demands.

With the aim of reducing the performance degradation, a threshold value is assigned to each server and is monitored triggering workload migrations based on three policies: Minimization of Migrations (MM), Highest Potential Growth (HPG), and Random Choice. All these policies try to reduce the number of SLA violations, reducing the overhead caused by live migration and potential increase of utilization.

Although the authors mention the opportunity for saving energy in heterogeneous server environments and configure their experiment with different processor capacities, they do not explain how this heterogeneity is exploited by the proposed mechanism and how it impacts the final results. Furthermore workloads with different processor requirements are included in the experiment. However they cannot be considered fully representative of workload heterogeneity since aspects such as execution time, application architecture and networking along with their energy-performance implications are not considered.

Towards energy-aware scheduling in data centers using machine learning

In (Berral et al., 2010) it is presented a theoretical approach for handling energy-aware scheduling in data centers. Here, the authors propose a framework which provides an allocation methodology using techniques that include turning on/off machines, power-aware allocation algorithms and machine learning to deal with uncertain information while the expected QoS is maintained through the avoidance of SLA violations.

In order to save energy, the strategy proposed in this paper is simple; reduce the number of active nodes by turning off those that remain inactive using workload consolidation. To achieve this, they propose a scheduling algorithm named “*dynamic backfilling*” which allows the migration of workloads among servers in order to provide a greater consolidation and thus the reduction of active nodes. These workload movements are performed with regard to certain policies that include System Occupation (SO), Current Job Performance (CJP) and Expected SLA Satisfaction (ESS) with the aim of improving the migration process and reducing SLA violations.

In order to reduce the performance degradation, machine learning techniques are introduced to predict the customer satisfaction level of each job before placing or moving them across the servers in the data center. Additionally working nodes thresholds are utilized to assist the turning on/off server frequency and adjust the overhead caused by these operations.

While in this approach the authors mention the inclusion of different workload types (grid and service) for the experiment, they do not describe how these different types are handled by the proposed mechanism. Apparently, they are treated as the same. However their results confirm that there exist significant differences in energy-performance among distinct types of workloads. Furthermore, this approach assumes homogeneous data centers. Based on the idea presented in (Nathuji, et al., 2007), cloud environments could be composed by different server architectures with different performance and energy capabilities. Considering this could lead to improvements in benefit of the energy and performance balance.

Performance evaluation of a green scheduling algorithm for energy savings in Cloud computing

(Duy, et al., 2010) presents an approach which aims to contribute in the saving energy problem by allocating VMs to the least number of turned on servers. The difference remarked in this work with respect to others that also try to reduce the energy consumption using “*dynamic servers’ pool resizing*” is the introduction of an algorithm integrating a neural network predictor for optimizing server power consumption and reducing the performance impact in cloud computing environments. This neural network is used to anticipate the future load demand on servers by considering the historical demand with the aim of reducing the turning on/off frequency, and the resulting overhead which could lead to serious performance degradation.

In this paper the authors describe the neural network, how it is composed, the training process, and how it works along with the “green” algorithm aiming to reduce the performance degradation. The simulations developed in which they used http workloads contained in the ClarkNet and NASA server logs are also presented. During these, performance was measured using the rate of drops, considering drop rate as the number of requests that exceed the capacity of each node to serve 1000 request/second for one single core.

In this approach, each time a user’s request is submitted, a process that includes VM creation and scheduling is executed; this process is called “*step*”. Here, the authors suppose that customer requests are processed within the same step being completed relatively quickly after their submission. Perhaps, this is because the type of workload utilized is represented by two types of HTTP requests (ClarkNet and NASA). In a real cloud scenario, different workload types could co-exist including those that persist for considerable periods of time (scientific, multi-tiered and multi-users applications). Considering this, in order to anticipate the number of turning on/off servers, not only predictions on the workloads’ arrival rate are needed but also on the workloads’ resource requirements and execution time. Additionally, a heterogeneous set of three different processors’ capabilities is considered. However, it is not clearly explained how this heterogeneity is exploited by the proposed approach apart from the palpable performance implications due to the restriction of allocating only one VM per core to achieve SLAs accomplishment.

Optimal power management for server farm to support green computing

In (Niyato, Chaisiri, & Sung, 2009) it is presented an approach which aims to contribute to the energy saving problem for data centers. Here the authors argue that some efforts have been made as contributions to this, but all they are based on heuristic methods in which the optimal performance and power consumption cannot be guaranteed. Because of that, they introduce a mechanism which works in two different sections of distributed data centers. First, each data center works along with an optimal power management module to make decisions about server mode switching to minimize the power consumption (turning on/off servers). Additionally, a module named job broker makes decisions on user’s assignment to a specific data center with the aim of minimizing the total cost, which is composed of network and power consumption cost.

Optimal performance levels are pursued through turning on servers in advance to reduce the workloads’ waiting time. The decision on how many servers should be reactivated is obtained by formulating and solving the constrained Markov decision process (CMDP). Additionally, optimizations at the job broker look for the best workload placement within the different data centers trying to avoid job migrations

among them that -as is mentioned by the authors- could lead to non-negligible system performance degradation.

This approach considers the allocation of only one job per server. When the job is finished, the server sends a message to the scheduler indicating its status. Then the scheduler can assign a new job or deactivate it. This, in addition to the characteristic of awaking servers in advance, could be very beneficial to performance in scenarios where all the workloads had high computing demands. However, the energy savings in real cloud scenarios could be seriously affected because of the heterogeneity of workloads and the lack of mechanisms for handling heterogeneous hardware infrastructure. The allocation of jobs with low resource demands in complete servers could represent a serious resource waste problem.

GreenCloud: a new architecture for green data center

In (Liu et al., 2009) “*GreenCloud*” is described as an approach which aims reduce the power consumption in data centers by reducing the number of turned on servers. In order to achieve that, the authors present an architecture composed of some components such as monitoring services, a migration manager, the managed environment, and the front end that provides information to users. Although the authors describe this architecture as their final proposal, in this paper they are mainly focused in describing the live migration algorithm which search optimal placement of virtual machines, minimizing the total cost; being the cost in this paper calculated considering physical machine cost, the virtual machine status and the virtual machine migration cost.

Maintenance of performance is pursued by a workload simulator which takes the resource requirements and collects real-time measurements from the data center in order to demonstrate the system performance to users, giving them the opportunity to adjust the parameters to obtain the desired performance. Because the test scenario is based on gaming applications, the performance measurements are presented in terms of Round Trip Time (RTT) which according to the authors is an essential concern in that specific type of applications.

This approach uses “*dynamic server’s pool resizing*”. However, it is not explained how the overhead caused by the turning on/off process is handled. Moreover, in this paper the authors center their focus in only one type of workload -represented by gaming applications- where RTT is a high performance constraint. It could be interesting to analyze GreenCloud’s behavior using different workload types. However, no mechanism is presented to handle workload heterogeneity. Finally, hardware heterogeneity is introduced during the calculation of migration where energy consumed by a physical machine is provided. Nevertheless, this is a static value that could be accompanied by inline monitoring to reveal the real server status in a specific time with a specific load.

Performance and energy-aware cluster level scheduling of compute-intensive jobs with unknown service times

In (Zikos & Karatza, 2010) the authors evaluate three different job allocation policies which are based on shortest queue scheduling algorithm in a scenario represented by heterogeneous servers. The main idea is to analyze the energy and performance implications for scheduling intensive-computing jobs considering different processor profiles; some of them orientated to high performance computing and others to save energy mixed in the same cluster. The job allocation in this physical resource environment is evaluated considering first an energy-efficient approach represented by the policy named Shortest Queue with Energy-Efficient Priority (SQEE) in which

processor with energy saving profiles are preferred and selected. Second, the allocation is evaluated considering a performance oriented policy named Shortest Queue with High Performance Priority (SQHP) in which processor with high computing oriented profiles are preferred and selected. And finally, another performance approach named Performance-Based Probabilistic–Shortest Queue (PBP–SQ) in which the selection probabilities are based on the computational capacity is evaluated.

Although this work does not deal directly with any other mechanisms for energy savings apart from energy-efficient processor use, since the focus is centered in the policies’ evaluation, it could be complemented with both “*Dynamic Processor Scaling*” and “*Dynamic Server’s Pool Resizing*”. The performance is evaluated in terms of response time and slowdowns of a job. The relation with the specific policies and different levels of system loads are also presented.

Table 1 Energy-aware approaches with performance considerations

Approaches	Energy saving mechanism	Performance look up mechanism	Allocation -Migration policies	Energy / Performance Metrics
(Von Laszewski, et al., 2009)	Dynamic Processor Scaling	Monitoring to adjust servers processing capacity based on fixed requirements	--	Power consumed (Watts) / nBench Integer Index
(Buyya, et al., 2010)	Dynamic Server’s Pool Resizing	Threshold monitoring to prevent SLA violations	LEPC, MM, HPG and RC	Energy consumed (Kwh) / SLA violation rate (%)
(Berral, et al., 2010)	Dynamic Server’s Pool Resizing	Machine Learning to predict the resulting client satisfaction level	SO, CJP and ESS	Power consumed (Kw) / SLA accomplishment rate (%)
(Duy, et al., 2010)	Dynamic Server’s Pool Resizing	Neural network predictor to reduce the turning on/off overhead	--	Energy consumed (Kwh) / Drop rate (%)
(Niyato, et al., 2009)	Dynamic Server’s Pool Resizing	CMDP-based Algorithms to reduce the waiting time avoiding the job blocking	Network and power cost at job broker level	Power consumed (Watts) / Waiting time (seconds)
(Liu, et al., 2009)	Dynamic Server’s Pool Resizing	Workload simulator to demonstrate the system performance	Migration Cost Calculation (MCC)	Energy consumed (Kwh) / Round Trip Time (ms)
(Zikos & Karatza, 2010)	--	Shortest Queue algorithm	SQEE, SQHP and PBP-SQ	Energy consumed (Energy units) /

OPPORTUNITIES

Cloud computing is a commercial model which aims to provide computing infrastructure (software and hardware) as a service, reducing the customers' management costs. But for this to be successful, cloud providers need mechanisms not only for reducing the energy consumption to support the offered prices and demand but also for accomplishing with the required QoS to ensure customer satisfaction. Although some approaches to reduce the energy waste in cloud computing environments have been developed, these still neglect some opportunities that must be addressed to achieve a real balance between the energy consumed and the performance offered. For example:

Allocation and migration policies considering workload types and behaviors

Current approaches are focused on reducing the number of working nodes for powering down or turning off the inactive ones. In order to achieve this, virtualization is used to consolidate the greatest possible quantity of workloads in a single physical node. However, not one of these approaches considers the fact that different types of workloads (e.g. scientific, social network or enterprise applications) can be allocated into the same physical node. This unconscious aggregation can result in a negative influence among them, degrading the performance and incrementing the energy consumption at the same time. It might occur because each one of the workload types could have different resource requirements and life-time.

Furthermore, current approaches use live migration with the aim of improving workload distribution. Nevertheless, these migrations are performed considering neither the workload type nor the possible communication with others. For example, as is mentioned in (Srikantaiah, Kansal, & Zhao, 2008), migrating streaming video applications could be more expensive than migrating other types of lighter applications such as a word processor. Additionally, many of the applications hosted in cloud data centers are tiered or belong to complex workflows being in continuous communication among them. An unconscious migration and redistribution of this type of applications could result in a performance degradation caused by the increment of the network latency and a rise in the energy consumption by the use of networking equipment.

Finally, current approaches propose distribution mechanisms trying to consolidate the maximum number of workloads in a single server. This can be beneficial for reducing the number of active nodes; however, overloading servers could result in serious performance deficiencies, the increase of the waste heat and in consequence a rise in power consumption for cooling systems. As mentioned in (Liu, et al., 2009), "*... as higher utilization does equal increase power consumption and more waste heat*".

The opportunity here is the design of resource-management policies and mechanisms considering what kind of applications can be mixed in a single host as well as the communications that exist among them to perform an optimized allocation. These mechanisms should consider the optimal usage point of each server to take advantage of them achieving the desired energy – performance balance. A study for characterizing the workload types and their behaviors is also required.

Dynamic selection of physical nodes in heterogeneous environments

Current approaches consider homogeneity of physical resources giving to each node the same characteristics such as CPU speed, disk, memory, etc. Thus all the servers present the same power consumption. However as it is mentioned in (Liu, et al., 2009), (Srikantaiah, et al., 2008) and (Mello Ferreira, 2010) in a real scenario, data center heterogeneity combined with workload distribution could provide better improvements in energy savings while preserving QoS. In other words, it is important to take advantage of the different hardware characteristics for distributing workloads giving preference on those servers with better energy-performance profiles, keeping those with high energy consumption turned off waiting for peak loads or critical applications. Additionally, most of these approaches consider only the processor as the unique resource in dispute, leaving out other important components such as memory and storage which due to the massive aggregation of workloads in the same node might represent bottle-necks which reduce the performance and increase the energy consumption as mentioned in (Lee & Zomaya, 2010).

The opportunity here is to design energy-performance profiles which integrate all the necessary information to help the scheduling algorithms for distributing the workloads on those servers that in a specific moment of time provide the best tradeoff between energy consumption and performance. Additionally, mechanisms for estimating the capacity of each server and their power profiles are necessary.

Prediction mechanisms for a smart workload distribution

Some approaches, such as (Srikantaiah, et al., 2008), (Lee & Zomaya, 2010) and (Nathuji, et al., 2007) prefer not turn off/on the servers because they consider that this can cause a non-negligible overhead. Those that address turning on/off nodes have not considered the different workload types and their behavior. According to (Buyya, et al., 2010), turning resources off in a dynamic environment is risky from a QoS perspective. Due to the variability of the workload and massive aggregation, some VMs may not obtain required resources under peak load, thereby failing to meet the required QoS. Some other approaches such as (Lefèvre & Orgerie, 2010), have implemented prediction algorithms based on the average time of inter-submission time of previous jobs without considering either the resources needed by the job or its behavior relying on end user predictions for the time and resources required. This is risky since in some scenarios it can result in inaccurate predictions and finally in greater power consumption or in a performance decrease.

The opportunity here is to develop mechanisms to predict the resources that a specific task will consume in order to allocate it correctly while keeping awake the proper number of servers, thereby avoiding the overhead generated during the shutting up/down process and achieving the QoS. Also in (Buyya, et al., 2010), it is considered essential to carry out a study of cloud services and their workloads in order to identify common behaviors, patterns, and explore load forecasting approaches that can potentially lead to more efficient resource provisioning and consequent energy efficiency.

Improved resource monitoring considering hardware's performance

Resource monitoring is another important issue that should be addressed in order to improve energy efficiency in cloud computing environments. Some of the approaches presented here introduce resource monitoring to support an energy efficient workload

distribution. However, all these monitors are just concerned about the energy consumed by the nodes without considering the underlying hardware's performance behavior. Monitoring mechanisms should contemplate optimal usage points for each node in cloud data centers to ensure a good balance between energy savings and performance.

The opportunity here is to design and develop energy–performance mechanisms for monitoring resources in cloud computing environments to support workload scheduling and distribution algorithms considering additionally the relying infrastructure's performance and the workload heterogeneity that exists in cloud computing environments.

Live migration's overhead reduction mechanisms

There exist other problems directly related to live migration use. For example in some approaches such as (Lee & Zomaya, 2010) and (Lefèvre & Orgerie, 2010) live migration seems not to be significant. This is because the migration process consumes a large amount of energy since it requires substantial attention from the hypervisor. According to (Lefèvre & Orgerie, 2010), hypervisor should copy the memory pages and send them to the new host. Additionally, if several migrations are required at the same time on the same node they are queued and processed one by one consuming a lot of energy. Finally, most approaches assume that the data centers are confined to a particular physical location. However, they recognize that a data center can span across multiple geographic locations (in a MAN or WAN for example). In this kind of scenario, migration becomes more complex due to the IP addressing problems impacting the virtual machines' performance adding a non-negligible overhead.

The opportunity here is to design and develop mechanisms to reduce the number of live migrations. Next generation of energy-aware mechanisms for cloud computing should avoid the continuous workload movements just to liberate resources. These migrations should be performed based on smart policies considering not only energy savings but also performance issues to accomplish the expected QoS. Additionally, mechanisms to reduce the overhead at hypervisor level and across different domains are also required.

CONCLUSIONS

Cloud computing is becoming one of the most important trends in service oriented and distributed systems. One of its strongest credentials is the “green” alternative offered to the customers. However, by itself cloud computing should not be considered as a green approach since in its core concept it is just a movement of loads and infrastructure from one place to another, including the energy consumed. What cloud computing represents is an attractive commercial opportunity to reduce the energy consumption at the “client edge” increasing it at data centers, where energy efficient mechanisms could be better applied. Because of this, some approaches have arisen trying to minimize energy consumption at cloud computing data centers. However, cloud providers are in need of mechanisms not only for reducing energy consumption to support the offered prices and demand but also for accomplishing with the required QoS to ensure the customer satisfaction.

In this paper the importance of energy savings without degrading the performance in cloud computing was discussed, since more than a technological advance it represents a business model where the satisfaction of customers has high priority. The state of art in energy-aware computing for cloud environments shows that the initials efforts for saving energy have started primarily focused in the reduction of energy waste generated

by idle servers mainly supported by VM consolidation and live migration. These, in conjunction with scheduling algorithms have boosted up two main trends: “dynamic server’s pool resizing” and “dynamic processor scaling”.

Most of the approaches embraced in these trends have been designed aiming to achieve the highest possible energy savings with weak performance considerations. However, some others have been proposed with a strong concern for performance preservation introducing policies and evaluations in their methodologies. These approaches were described and analyzed in this paper.

From this analysis it is possible to conclude that there still exist some gaps that must be covered to achieve the energy-performance balance that is necessary in cloud computing environments. These represent a challenge related to the exploitation of cloud computing fine-grained characteristics that in addition with current approaches could lead to the success of energy-aware mechanisms, characterized by environmental and economical benefits. The introduction of some additional variables such as workload and hardware heterogeneity, workload networking and server’s optimal utilization point that exist in real cloud scenario as well as mechanisms to handle them, still represent opportunities that should be addressed looking for the adoption of energy-aware mechanisms in cloud environments where energy savings are required maximizing the performance for the customer’s satisfaction.

REFERENCES

- Abdelsalam, H. S., Maly, K., Mukkamala, R., Zubair, M., & Kaminsky, D. (2009). *Analysis of Energy Efficiency in Clouds*. Paper presented at the Proc. of Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns.
- Amrhein, D., Anderson, P., & de Andrade, A. (2010). *Cloud Computing Use Case* (White paper,): Cloud Computing Use Case Discussion Group.
- Berral, J. L., Goiri, Í., Nou, R., Julià, F., Guitart, J., Gavaldà, R., et al. (2010, April 13). *Towards energy-aware scheduling in data centers using machine learning*. Paper presented at the Proc. of the 1st International Conference on Energy-Efficient Computing and Networking, Passau, Germany.
- Buyya, R., Beloglazov, A., & Abawajy, J. (2010, July 12-15). *Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges*. Paper presented at the Proc. of the 2010 International Conference on Parallel and Distributed Processing Techniques and Applications, Las Vegas, NV, USA.
- Cameron, K. W. (2010, May 2010). The Challenges of Energy-Proportional Computing. *Computer Mag.*, 43, 82-83.
- Carter, J., & Rajamani, K. (2010, July 19). Designing Energy-Efficient Servers and Data Centers. *Computer Mag.*, 43, 76-78.
- Couch, A. L., & Kumar, K. (2008). *Workshop on Power Aware Computing and Systems* (Sum. Rep.). San Diego, CA, USA.
- Chickowski, E. (2007, June 22). Safely Eliminating E-waste. *Processor Mag.*, 29, 12.
- Duy, T. V. T., Sato, Y., & Inoguchi, Y. (2010). *Performance evaluation of a Green Scheduling Algorithm for energy savings in Cloud computing*. Paper presented at the IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW). Retrieved from <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5470908>
- Erdogmus, H. (2009). Cloud Computing: Does Nirvana Hide behind the Nebula? *IEEE Softw.*, 26(2), 4-6.

- Evoy, G. V. M., & Schulze, B. (2008). *Using clouds to address grid limitations*. Paper presented at the Proceedings of the 6th international workshop on Middleware for grid computing. Retrieved from <http://portal.acm.org/citation.cfm?id=1462715#>
- Freeman, L. (2009). *Reducing Data Center Power Consumption Through Efficient Storage* (White paper No. WP-7010-0709): NetApp, Inc.
- Gain, B. (2010, January 1). Cloud Computing & SaaS In 2010 *Processor Mag.*, 32, 12.
- Gandhi, A., Harchol-Balter, M., Das, R., & Lefurgy, C. (2009). *Optimal power allocation in server farms*. Paper presented at the Proc. of the eleventh Joint International Conference on Measurement and Modeling of Computer Systems Seattle, WA, USA.
- Hickey, A. R. (2008, 2008 June 18). Gartner: Green IT Needs To Be On Midsize CIOs' Radar Screens. Retrieved April 16, 2010, from <http://www.crn.com/hardware/208700292>
- IBM. (2009). *Power Systems: Introduction to Virtualization* (Tech. rep. No. 5733-SSI): IBM Corp.
- Lee, Y. C., & Zomaya, A. Y. (2010). Energy efficient utilization of resources in cloud computing systems *The Journal of Supercomputing*, 53, 1-13.
- Lefèvre, L., & Orgerie, A.-C. (2010). Designing and evaluating an energy efficient Cloud. *The Journal of Supercomputing*, 51(3), 352 - 373.
- Ley, K., Bianchini, R., Martonosi, M., & Nguyeny, T. D. (2009). *Cost- and Energy-Aware Load Distribution Across Data Centers*. Paper presented at the 22nd ACM Symposium on Operating Systems Principles. Retrieved from <http://www.cs.rutgers.edu/~ricardob/papers/hotpower09.pdf>
- Li, B., Li, J., Huai, J., Wo, T., Li, Q., & Zhong, L. (2009). *EnaCloud: An Energy-saving Application Live Placement Approach for Cloud Computing Environments*. Paper presented at the International Conference on Cloud Computing. Retrieved from <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5284078>
- Liu, L., Wang, H., Liu, X., Jin, X., He, W. B., Wang, Q. B., et al. (2009). *GreenCloud: a new architecture for green data center*. Paper presented at the Proc. of the sixth International Conference on Autonomic Computing Barcelona, Spain.
- Maggiari, R. (2009). *Cloud computing is changing how we communicate*. Paper presented at the Proc. of the IEEE International Professional Communication Conference.
- Mello Ferreira, A. (2010). *An energy-aware approach for service performance evaluation*. Paper presented at the International Conference on Energy-Efficient Computing and Networking. Retrieved from An energy-aware approach for service performance evaluation
- Molla, A. (2008, December 3-5). *GITAM: A Model for the Adoption of Green IT*. Paper presented at the Proc. of the ACIS Australian Conference on Information Systems, Christchurch, New Zealand.
- Murugesan, S. (2008, January 2008). Harnessing Green IT: Principles and Practices. *IT Professionals Mag.*, 10, 24-33.
- Naditz, A. (2008, October 2008). Green IT 101: Technology Helps Businesses and Colleges Become Enviro-Friendly. *Sustainability: The Journal of Record Mag.*, 1, 315-318.
- Nathuji, R., Isci, C., & Gorbatov, E. (2007). *Exploiting Platform Heterogeneity for Power Efficient Data Centers*. Paper presented at the Proc. of the IEEE International Conference on Autonomic Computing Washington, DC, USA

- Niyato, D., Chaisiri, S., & Sung, L. B. (2009). *Optimal Power Management for Server Farm to Support Green Computing*. Paper presented at the 9th IEEE/ACM International Symposium on Cluster Computing and the Grid. Retrieved from <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5071858>
- Nordman, B., & Christensen, K. (2010, January 2010). Proxying: Next Step in Reducing IT Energy Use. *Computer Mag.*, 43, 91-93.
- Patel, C. D., Bash, C. E., Belady, C., Stahl, L., & Sullivan, D. (2001, July 8-13). *Computational Fluid Dynamics Modeling of High Compute Density Data Centers to Assure System Inlet Air Specifications*. Paper presented at the Proc. of the Pacific Rim/ASME International Electronic Packaging Technical Conference and Exhibition, Kauai, Hawaii, USA.
- Schneider, E. (2010, 10/12/2009). Go Green, Save Green The Benefits of Eco-Friendly Computing. Retrieved April 13, 2010, from http://www.apcmedia.com/salestools/SLAT-7DCQ5J_R0_EN.pdf
- See, S. (2008). Is there a pathway to a Green Grid ?? Retrieved April 20, 2010, from <http://www.ibergrid.eu/2008/presentations/Dia%2013/4.pdf>
- Smarr, L. (2010, January 26). Project GreenLight: Optimizing Cyber-Infrastructure For a Carbon-Constrained World. *Computer Mag.*, 43, 22-27.
- Srikantaiah, S., Kansal, A., & Zhao, F. (2008). *Energy Aware Consolidation for Cloud Computing*. Paper presented at the Proc. of the USENIX HotPower'08: Workshop on Power Aware Computing and Systems.
- Vaquero, L. M., Rodero-Merino, L., Caceres, J., & Linder, M. (2009, January 2009). *A Break in the Clouds: Towards a Cloud Definition*. Paper presented at the Proc. of the ACM SIGCOMM Computer Communication Review.
- Velte, A. T., Velte, T. J., & Elsenpeter, R. (2010). Chapter One: Cloud Computing Basics *Cloud Computing: A Practical Approach* (pp. 3-22): McGraw-Hill.
- VMware. (2009). *VMware Distributed Power Management Concepts and Use* (Tech. Rep. No. IN-073-PRD-01-01). Palo Alto, CA, USA: VMware Inc.
- Von Laszewski, G., Wang, L., Younge, A. J., & He, X. (2009). *Power-Aware Scheduling of Virtual Machines in DVFS-enabled Clusters*. Paper presented at the Proc. of IEEE International Conference on Cluster Computing 2009, New Orleans, LA, USA.
- Wang, D. (2007). *Meeting Green Computing Challenges*. Paper presented at the High Density packaging and Microsystem Integration, 2007. HDP '07. International Symposium on.
- Webb, M. (2008). *SMART 2020: Enabling the low carbon economy in the information age* (Tech. Rep.,): The Climate Group.
- Weiss, A. (2007). Computing in the clouds. *netWorker*, 11(4), 16-25.
- World Coal Institute. (2010). Coal and Electricity. Retrieved September 16, from <http://www.worldcoal.org/coal/uses-of-coal/coal-electricity/>
- Zikos, S., & Karatza, H. D. (2010). Performance and energy aware cluster-level scheduling of compute-intensive jobs with unknown service times. *Simulation Modelling Practice and Theory, In Press, Corrected Proof*.