Green Cloud computing and Environmental Sustainability

Saurabh Kumar Garg and Rajkumar Buyya

Cloud computing and Distributed Systems (CLOUDS) Laboratory
Dept. of Computer Science and Software Engineering
The University of Melbourne, Australia
Email: {saurabhg, rbuyya}@unimelb.edu.au

Abstract: Cloud computing is a highly scalable and cost-effective infrastructure for running HPC, enterprise and Web applications. However, the growing demand of Cloud infrastructure has drastically increased the energy consumption of data centers, which has become a critical issue. High energy consumption not only translates to high operational cost, which reduces the profit margin of Cloud providers, but also leads to high carbon emissions which is not environmentally friendly. Hence, energy-efficient solutions are required to minimize the impact of Cloud computing on the environment. In order to design such solutions, deep analysis of Cloud is required with respect to their power efficiency. Thus, in this chapter, we discuss various elements of Clouds which contribute to the total energy consumption and how it is addressed in the literature. We also discuss the implication of these solutions for future research directions to enable green Cloud computing. The chapter also explains the role of Cloud users in achieving this goal.

1. Introduction

With the growth of high speed networks over the last decades, there is an alarming rise in its usage comprised of thousands of concurrent e-commerce transactions and millions of Web queries a day. This ever-increasing demand is handled through large-scale datacenters, which consolidate hundreds and thousands of servers with other infrastructure such as cooling, storage and network systems. Many internet companies such as Google, Amazon, eBay, and Yahoo are operating such huge datacenters around the world.

The commercialization of these developments is defined currently as Cloud computing [2], where computing is delivered as utility on a pay-as-you-go basis. Traditionally, business organizations used to invest huge amount of capital and time in acquisition and maintenance of computational resources. The emergence of Cloud computing is rapidly changing this *ownership-based* approach to *subscription-oriented* approach by providing access to scalable infrastructure and services on-demand. Users can store, access, and share any amount of information in Cloud. That is, small or medium enterprises/organizations do not have to worry about purchasing, configuring, administering, and maintaining their own computing infrastructure. They can focus on sharpening their core competencies by exploiting a number of Cloud computing benefits such as on-demand computing resources, faster and cheaper software development capabilities at low cost. Moreover, Cloud computing also offers enormous amount of compute power to organizations which require processing of tremendous amount of data generated almost every day. For instance, financial companies have to maintain every day the

dynamic information about their hundreds of clients, and genomics research has to manage huge volumes of gene sequencing data.

Therefore, many companies not only view Clouds as a useful on-demand service, but also a potential market opportunity. According to IDC (International Data Corporation) report [1], the global IT Cloud services spending is estimated to increase from $16 billion in 2008 to $42 billion in 2012, representing a compound annual growth rate (CAGR) of 27%. Attracted by this growth prospects, Web-based companies (Amazon, eBay, Salesforce.com), hardware vendors (HP, IBM, Cisco), telecom providers (AT&T, Verizon), software firms (EMC/VMware, Oracle/Sun, Microsoft) and others are all investing huge amount of capital in establishing Cloud datacenters. According to Google's earnings reports, the company has spent $US1.9 billion on datacenters in 2006, and $US2.4 billion in 2007[3].

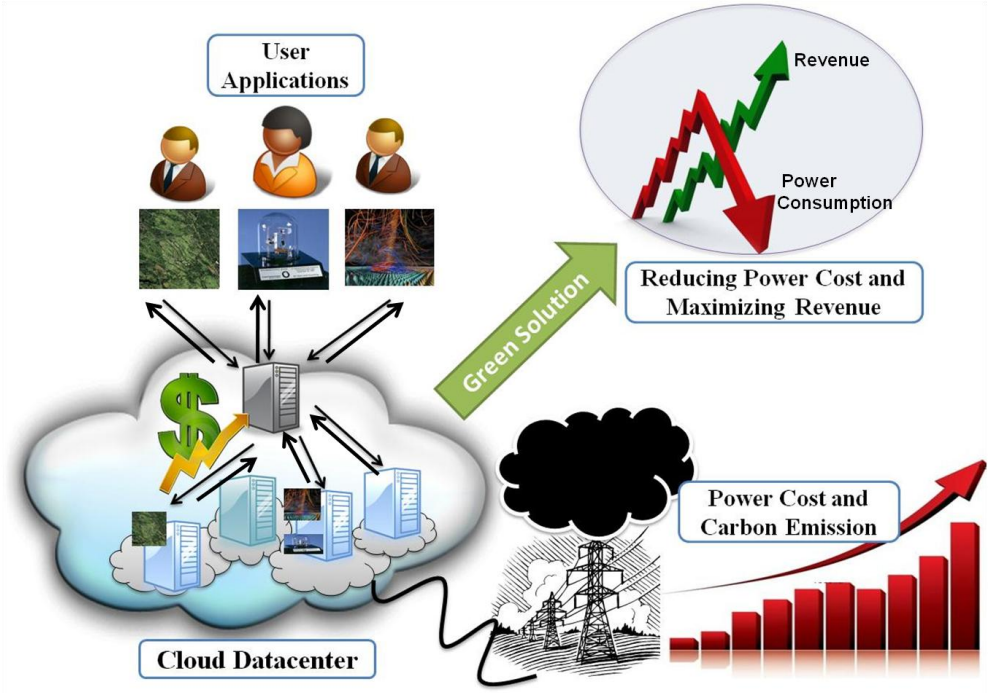


Figure 1. Cloud and Environmental Sustainability

Clouds are essentially virtualized datacenters and applications offered as services on a subscription basis as shown in Figure 1. They require high energy usage for its operation [4]. Today, a typical datacenter with 1000 racks need 10 Megawatt of power to operate [5], which results in higher operational cost. Thus, for a datacenter, the energy cost is a significant component of its operating and up-front costs. In addition, in April 2007, Gartner estimated that the Information and Communication Technologies (ICT) industry generates about 2% of the total global CO_2 emissions, which is equal to the aviation industry [5]. According to a report published by the European Union, a decrease in emission volume of 15%–30% is required before year 2020 to keep the global temperature increase below 2 ^{0}C. Thus, energy consumption and carbon emission by Cloud infrastructures has become a key environmental concern.

Some studies show that Cloud computing can actually make traditional datacenters more energy efficient by using technologies such as resource virtualization and workload consolidation. The traditional data centres running Web applications are often provisioned to handle sporadic peak loads, which can result in low resource utilization and wastage of energy. Cloud datacenter, on the other hand, can reduce the energy consumed through server consolidation, whereby different workloads can share the same physical host using virtualization and unused servers can be switched off. A recent research by Accenture [7] shows that moving business applications to Cloud can reduce carbon footprint of organizations. According to the report, small businesses saw the most dramatic reduction in emissions – up to 90 percent while using Cloud resources. Large corporations can save at least 30-60 percent in carbon emissions using Cloud applications, and mid-size businesses can save 60-90 percent.

Contrary to the above opinion, some studies, for example Greenpeace [6], observe that the Cloud phenomenon may aggravate the problem of carbon emissions and global warming. The reason given is that the collective demand for computing resources is expected to further increase dramatically in the next few years. Even the most efficiently built datacenter with the highest utilization rates will only mitigate, rather than eliminate, harmful CO_2 emissions. The reason given is that Cloud providers are more interested in electricity cost reduction rather than carbon emission. The data collected by the study is presented in Table 1 below. Clearly, none of the cloud datacenter in the table can be called as green.

Table 1. Comparison of Significant Cloud Datacenters [6]

Cloud datacenters	Location	Estimated power usage Effectiveness	% of Dirty Energy Generation	% of Renewable Electricity
Google	Lenoir	1.21	50.5% Coal, 38.7% Nuclear	3.8%
Apple	Apple, NC		50.5% Coal, 38.7% Nuclear	3.8%
Microsoft	Chicago, IL	1.22	72.8% Coal, 22.3% Nuclear	1.1%
Yahoo	La Vista, NE	1.16	73.1% Coal, 14.6% Nuclear	7%

In summary, Cloud computing, being an emerging technology also raises significant questions about its environmental sustainability. While financial benefits of Cloud computing have been analyzed widely in the literature, the energy efficiency of Cloud computing as a whole has not been analyzed. Through the use of large shared virtualized datacenters Cloud computing can offer large energy savings. However, Cloud services can also further increase the internet traffic and its growing information database which could decrease such energy savings. Thus, this chapter explores the environmental sustainability of Cloud computing by analyzing various technologies and mechanism that support this goal. Our analysis is important for users and organization that are looking at Cloud computing as a solution for their administrative, infrastructural and management problems.

Finally, we also propose and recommend a Green Cloud framework for reducing its carbon footprint in wholesome manner without sacrificing the quality of service (performance, responsiveness and availability) offered by the multiple Cloud providers.

2. What is Cloud computing?

Cloud computing is an evolving paradigm which is enabling outsourcing of all IT needs such as storage, computation and software such as office and ERP, through large Internet. The shift toward such service-oriented computing is driven primarily by ease of management and administration process involving software upgrades and bug fixes. It also allows fast application development and testing for small IT companies that cannot afford large investments on infrastructure. Most important advantage offered by Clouds is in terms of economics of scale; that is, when thousands of users share same facility, cost per user and the server utilization. To enable such facilities, Cloud computing encompasses many technologies and concepts such as virtualization, utility computing, pay as you go, no capital investment, elasticity, scalability, provisioning on demand, and IT outsourcing.

Due to such varying properties of Cloud computing, there are many informal definitions, none of which fully describes it. The literary meaning of "Cloud computing" can be "computing achieved using collection of networked resources, which are offered on subscription". In terms of qualities of real "Clouds," which do not have any definite shape or position, Cloud computing is also called "Cloud" since a Cloud server can have any configuration and can be located anywhere in the world. Internet is a fundamental medium through which these Cloud services are made accessible and delivered to end user.

The growing popularity of Cloud computing has led to several proposal defining its characteristics. Some of the definitions given by many well known scientists and organizations include:

A) Buyya et al.[2] define the Cloud computing in terms of its utility to end user: *"A Cloud is a market-oriented distributed computing system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resource(s) based on service-level agreements established through negotiation between the service provider and consumers."*

B) National Institute of Standards and Technology (NIST) [8] defines Cloud computing as follows*:"Cloud computing is a model for enabling 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. This Cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models."*

The characteristics of Clouds include on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. The available service models are classified as SaaS (Software-as-a-Service), PaaS (Platform-as-a-Service), and IaaS (Infrastructure-as-a-Service). The deployment models is categorised into public, private, community, and hybrid Clouds.

2.1 Cloud Computing Characteristics

The key characteristics exhibited by Clouds are shown in Figure 2 and they are discussed below.

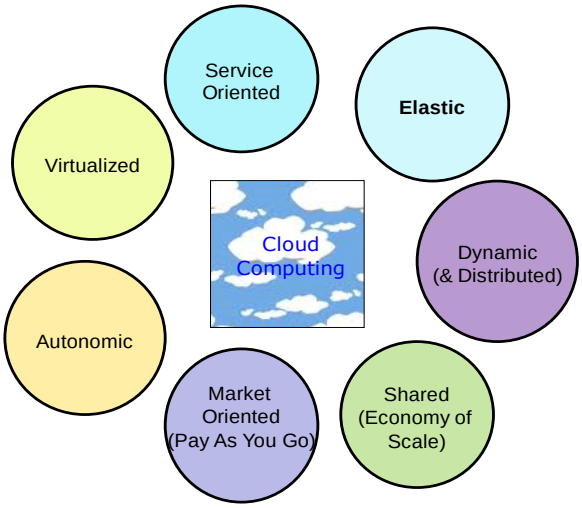


Figure 2. Characteristics of Cloud computing

- ***Virtualized:*** Resources (i.e. compute, storage, and network capacity) in Clouds are virtualized and it is achieved at various levels including VM (Virtual Machine) and Platform levels. The most basic one is at Virtual Machine (VM) level where different applications can be executed within their containers or operating systems running on the same physical machine. Platform level enables seamless mapping of applications to one or more resources offered by different Cloud infrastructure providers.
- ***Service-Oriented***: Cloud is implemented using Service-Oriented Architecture model where all the capabilities/components are available over the network as a service. Whether it is software, platform or infrastructure everything is offered as a service.
- ***Elastic:*** Resources (i.e. compute, storage, and network capacity) required for Cloud applications can be dynamically provisioned and varied i.e., increase or decrease at runtime depending on user QoS requirements. Major Cloud providers such as Amazon even provide services for automatic scale-out and scale-in based on hosted application requirements.
- **Dynamic and Distributed:** Although Cloud resources are virtualised, they are often distributed to enable the delivery of high-performance and/or reliable Cloud services. These resources are flexible and can be adapted according to customer's requirements such as software, network configuration, etc.
- ***Shared (Economy of Scale)***: Clouds are shared infrastructure where resources serve multiple customers with dynamic allocation according to their application's demand. This sharing model is also termed as "multi-tenant" model. In general, the customers neither have any direct control over physical resources nor they are aware of the resource location and with whom they are being shared.
- **Market-Oriented (Pay as you go):** In Cloud computing, customers pay for services on a pay-per-use (or pay-as-you-go) basis. The pricing model can vary depending on the QoS expectation of application. Cloud IaaS providers such as Amazon price resources using market models such as commodity or on-spot pricing models. A pricing model proposed by Allenofor & Thulasiram [49] for grid resources could be used as a base for cloud

resources. This characteristic addresses the utility dimension of cloud computing. That means, Cloud services are offered as "metered" services where providers have an accounting model for measuring the use of the services, which helps in development of different pricing plans and models. The accounting model helps in the control and optimization of resource usage.

- **Autonomic:** To provide highly reliable services, Clouds exhibit autonomic behaviour by managing themselves in case of failures or the performance degradation.

2.2 Components of Cloud Computing

Cloud computing is mainly composed of three layers which cover all the computing stack of a system. Each of these layers offers different set of services to end users as described in Figure 3. At the lowest layer, Cloud offerings are named as ***Infrastructure-as-a-Service (IaaS)*** which consists of virtual machines or physical machines, storage, and clusters. Cloud infrastructures can also be heterogeneous, integrating clusters, PCs and workstations. Moreover, the system infrastructure can also include database management systems and other storage services. The infrastructure in general is managed by an upper management layer that guarantees runtime environment customization, application isolation, accounting and quality of service. The virtualization tools, such as hypervisors, also sit in this layer to manage the resource pool and to partition physical infrastructure in the form of customized virtual machines. Depending on the end user needs, the virtualized infrastructure is pre-configured with storage and programming environment, what saves time for users who do not need to build their system from scratch.

Even though IaaS gives access to physical resources with some software configuration, for designing new applications user requires advanced tools such as Map Reduce etc. These services constitute another layer called ***Platform as a Service (PaaS)***, offering Cloud users a development platform to build their applications. Google AppEngine [9], Aneka [10], and Microsoft Azure [11] are some of the most prominent example of PaaS Clouds. In general, PaaS includes the lower layer (IaaS) as well that is bundled with the offered service. In general, pure PaaS offers only the user level middleware, which allows development and deployment of applications on any Cloud infrastructure. As noted by Appistry.com [12], the essential characteristics that identify a Platform-as-a-Service solution include:

- ***Runtime framework:*** It represents the "software stack" of the PaaS model and the most intuitive aspect that comes to the mind of people when referring to Platform-as-a- Service solutions. The runtime framework executes end-user code according to the service level policies set by the user and the provider.
- ***Abstraction*****:** PaaS solutions are distinguished by the higher level abstraction that they provide. Unlike IaaS solutions the focus is on delivering "raw" access to virtual or physical infrastructure. In the case of PaaS the focus is on the applications the Cloud must support. This means that PaaS solutions offer a way to deploy and manage applications on the Cloud rather than a bunch of virtual machines on top of which the IT infrastructure is built and configured.
- ***Cloud services***: PaaS offerings provide developers and architects with services and APIs helping them to simplify delivering of elastically scalable and highly available Cloud applications. These services are the key differentiators among competing PaaS solutions

and generally include specific component for developing applications, advanced services for application monitoring, management, and reporting.

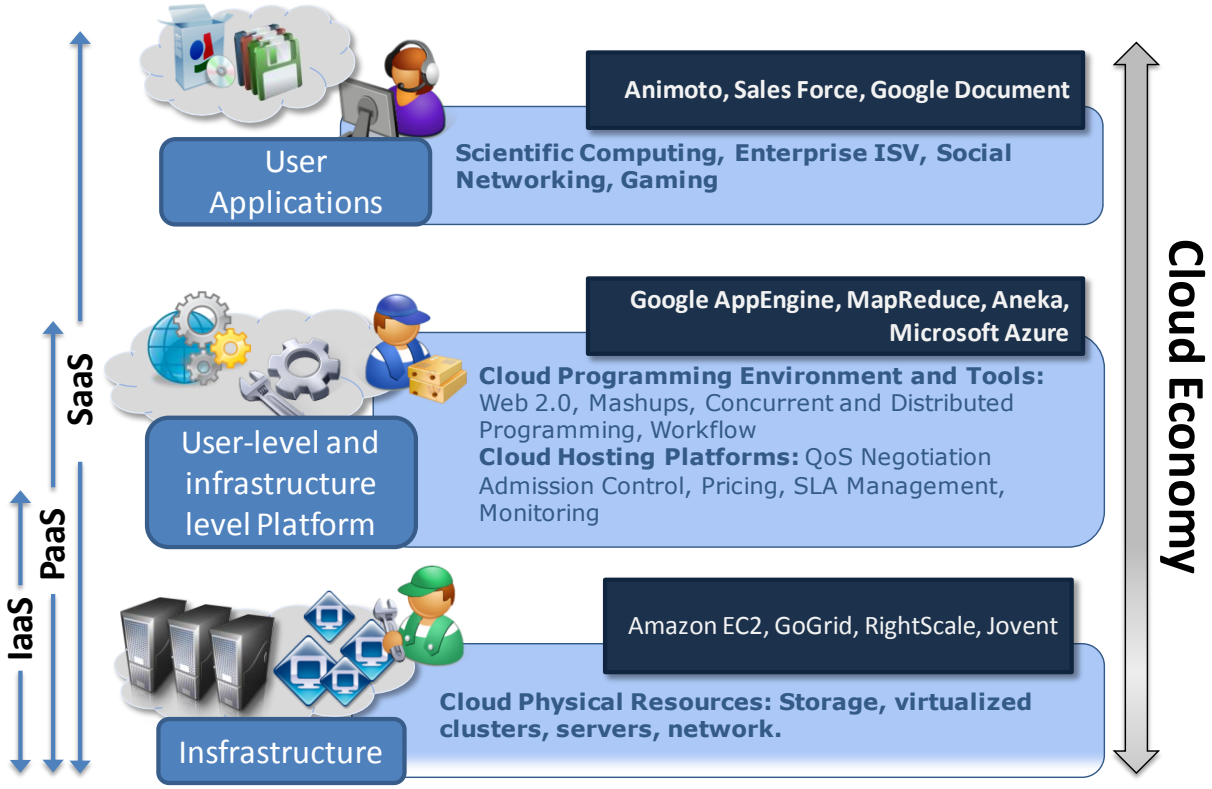


Figure 3. Cloud computing Architecture

The major advantage of PaaS is the cost saving in development, deployment, and management cycle of new applications. The PaaS providers reduces risk in terms of upgrade cost of underlying platforms and allow Cloud users to concentrate on the application development.

On topmost layer of Cloud computing Architecture, the Cloud services (Figure 3) are referred as ***Software as a Service (SaaS)*** which is a software delivery model providing on-demand access to applications. The most common examples of such service are CRM and ERPF applications that are commonly used in almost all the enterprises from small, to large business. In general, SaaS providers also constitute other layers of Cloud computing and thus, maintain the customer data and configure the applications according to customer need. This scenario results in considerable reduction in upfront cost of purchasing new software and infrastructure. The customers do not have to maintain any infrastructure or install anything within their premises. They just require high speed network to get instant access to their applications.

Multi-tenancy is another core feature of SaaS compared to traditional packaged software, allowing providers to outsource the effort of managing large hardware infrastructure, maintaining and upgrading applications, and optimizing resources by sharing the costs among the large user base. Therefore, SaaS model is particularly appealing for companies who get access to softwares configured according to their specific needs and shared between multiple users. On the customer side, only costs that will incur are the monthly software usage fee.

2.3 Cloud Computing Deployment Models

From above discussions, we can say that Cloud computing is a paradigm of offering on-demand services to end users. Clouds are deployed on physical infrastructure where Cloud middleware is implemented for delivering service to customers. Such an infrastructure and middleware differ in their services, administrative domain and access to users. Therefore, the Cloud deployments are classified mainly into three types: Public Cloud, Private Cloud and Hybrid Cloud (Figure 4).

2.3.1 Public Clouds

Public Cloud is the most common deployment model where services are available to anyone on Internet. To support thousand of public domain users, datacenters built by public Cloud providers are quite large comprising of thousands of servers with high speed network. Some of the famous public Clouds are Amazon Web Services (AWS), Google AppEngine, and Microsoft Azure. In this deployment, Cloud services are made available to the public in a pay-as-you-go-manner. A public Cloud can offer any of the three kinds of services: IaaS, PaaS, and SaaS. For instance, Amazon EC2 is a public Cloud providing infrastructure as a service, Google AppEngine is a public Cloud providing an application development platform as a service, and Salesforce.com is public Cloud providing software as a service. Public Cloud offers very good solutions to the customers having small enterprise or with infrequent infrastructure usage, since these Clouds provide a very good option to handle peak loads on the local infrastructure and for an effective capacity planning. The fundamental characteristic of public Clouds is its multi-tenancy, which is essentially achieved using sophisticated virtualization at various level of the software stack. Being public Clouds, Quality of Service and security are the main issues that need to be ensured in their management. Thus, a significant portion of the software infrastructure is devoted to monitor Cloud resources, to bill them according to the contract made with the user, and to keep a complete history of the Cloud usage for each customer.

2.3.2 Private Clouds

While public Clouds are quite appealing and provide a viable solution for cutting IT costs such as administration and infrastructure, there are still many scenarios where organization may want to maintain their own specialized Clouds catering to their particular needs. For instance, the health care industry maintains many confidential medical data which cannot be stored in public infrastructure. Thus, private Clouds are deployed within the premise of an organization to provide IT services to its internal users. The private Cloud services offer greater control over the infrastructure, improving security and service resilience because its access is restricted to one or few organizations. Such private deployment poses an inherent limitation to end user applications i.e. inability to scale elastically on demand as can be done using pubic Cloud services. An organization can buy more machines according to expanding needs of its users, but this cannot be done as fast and seamlessly as with public Clouds. This resulted in the emergence of hybrid deployments for Clouds where the advantages of both private and public Clouds are made available to the organization.

2.3.3 Hybrid Clouds

Hybrid Clouds is the deployment which emerged due to diffusion of both public and private Clouds' advantages. In this model, organizations outsource non-critical information and

processing to the public Cloud, while keeping critical services and data in their control. Therefore, organizations can utilize their existing IT infrastructure for maintaining sensitive information within the premises, and whenever require auto-scaling their resources using public Clouds. These resources or services are temporarily leased in peak load times and then released. The hybrid Cloud, in general, applies to services related to IT infrastructure rather than software services.

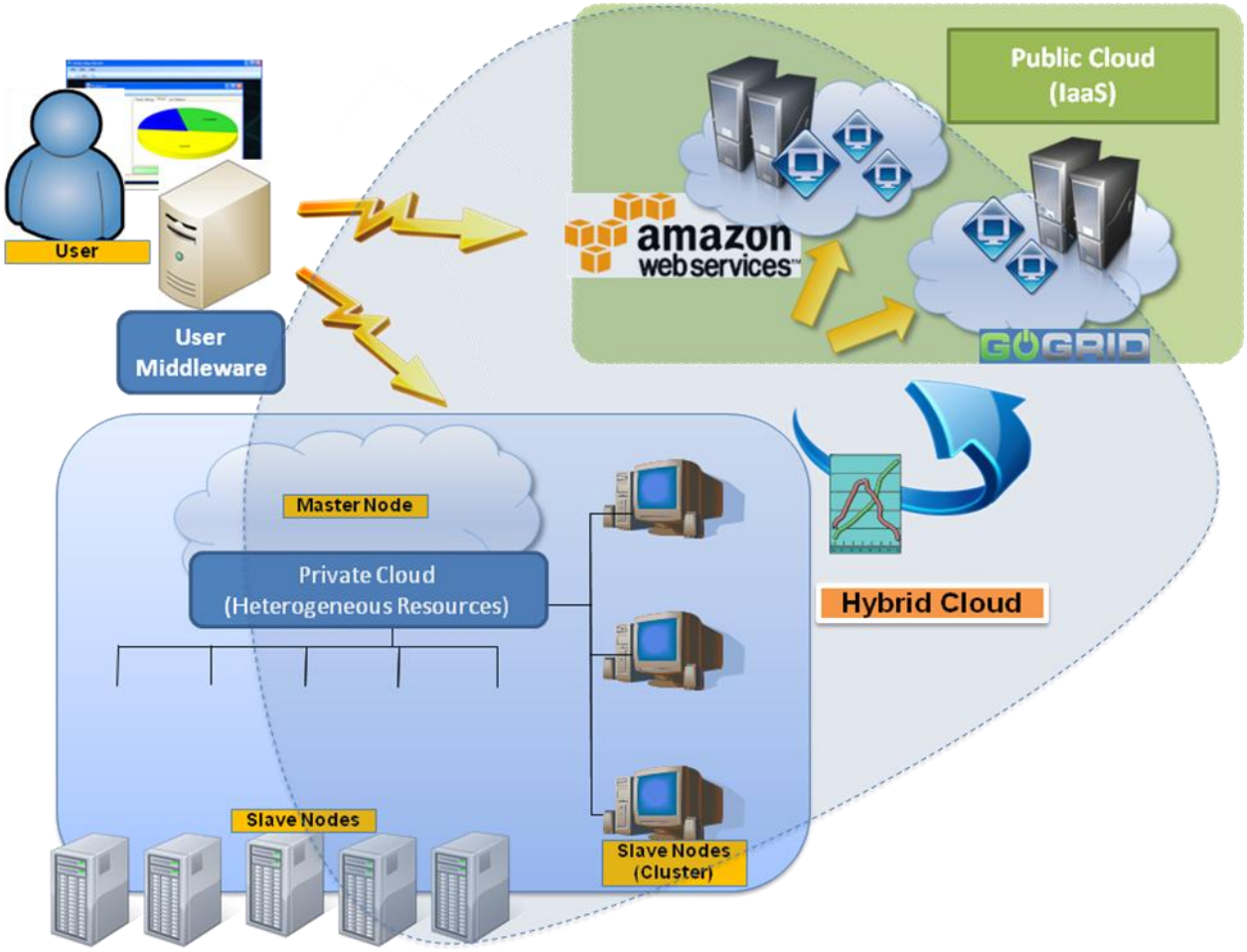


Figure 4. Deployment Models for Clouds

3. Cloud Computing and Energy Usage Model: A Typical Example

In this section, through a typical Cloud usage scenario we will analyze various elements of Clouds and their energy efficiency. Figure 5 shows an end user accessing Cloud services such as SaaS, PaaS, or IaaS over Internet. User data pass from his own device through an Internet service provider's router, which in turn connects to a Gateway router within a Cloud datacenter. Within datacenters, data goes through a local area network and are processed on virtual machines, hosting Cloud services, which may access storage servers. Each of these computing and network devices that are directly accessed to serve Cloud users contribute to energy consumption. In addition, within a Cloud datacenter, there are many other devices, such as cooling and electrical devicies, that consume power. These devices even though do not directly

help in providing Cloud service, are the major contributors to the power consumption of a Cloud datacenter. In the following section, we discuss in detail the energy consumption of these devices and applications.

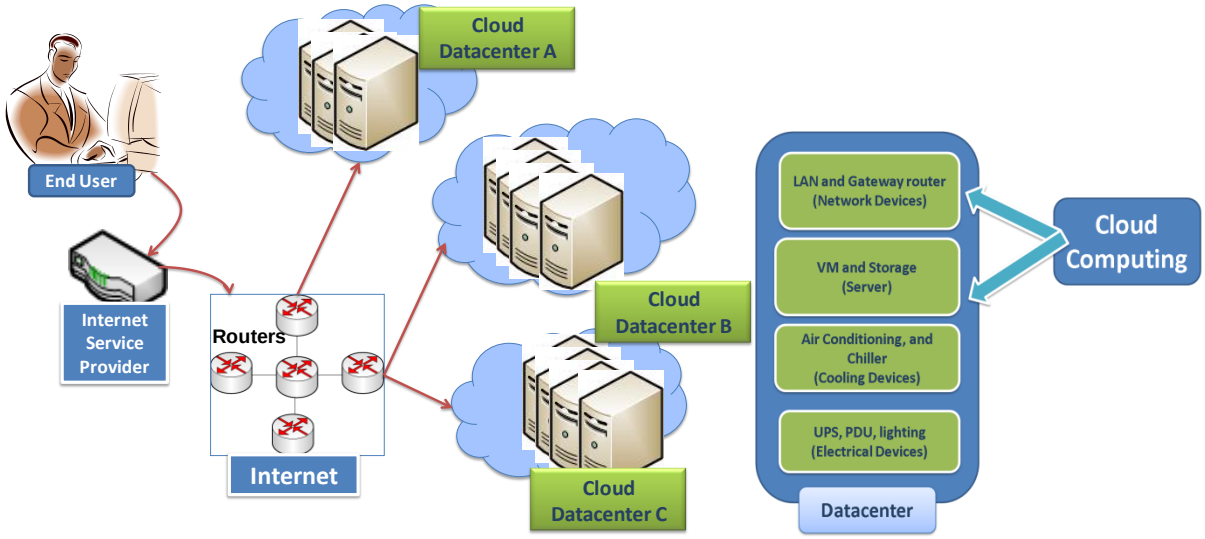


Figure 5. Cloud Usage Model.

3.1 User/Cloud Software Applications

The first factor that contributes to energy consumption is the way software applications are designed and implemented. The Cloud computing can be used for running applications owned by individual user or offered by the Cloud provider using SaaS. In both cases, the energy consumption depends on the application itself. If application is long running with high CPU and memory requirements then its execution will result in high energy consumption. Thus, energy consumption will be directly proportional to the application's profile. The allocation of resources based on the maximum level of CPU and memory usage will result in much higher energy consumption than actually required. The energy inefficiency in execution of an application emanates from inaccurate design and implementation. The application inefficiencies, such as suboptimal algorithms and inefficient usage of shared resources causing contention lead to higher CPU usage and, therefore, higher energy consumption. However, factors such as energy efficiency are not considered during the design of an application in most of the application domains other than for example embedded devices such as mobile phone.

3.2 Cloud Software Stack for SaaS, PaaS, IaaS Level

The Cloud software stack leads to an extra overhead in execution of end user applications. For instance, it is well known that a physical server has higher performance efficiency than a virtual machine and IaaS providers offer generally access to a virtual machine to its end users [13]. In addition, the management process in the form of accounting and monitoring requires some CPU power. Being profit oriented, service providers regularly have to adhere to Service Level Agreements (SLA) with their clients. These SLAs may take the form of time commitment for a task to be completed. Thus, Cloud provider for meeting certain level of service quality and

availability, provision extra resources than generally required. For instance, to avoid failure, fast recovery and reduction in response time, providers have to maintain several storage replicas across many datacenters. Since workflow in Web applications require several sites to give better response time to its end user, their data is replicated on many servers across the world. Therefore, it is important to explore the relationships among Cloud components and the tradeoffs between QoS and energy consumption.

3.3. Network Devices

The network system is another area of concern which consumes a non-negligible fraction of the total power consumption. The ICT energy consumption estimates [14] just for Vodafone Group radio access network was nearly 3 TWh in 2006. In Cloud computing, since resources are accessed through Internet, both applications and data are needed to be transferred to the compute node. Therefore, it requires much more data communication bandwidth between user's PC to the Cloud resources than require the application execution requirements. In some cases, if data is really large, then it may turn out to be cheaper and more carbon emission efficient to send the data by mail than to transfer through Internet.

In Cloud computing, the user data travels through many devices before it reaches a datacenter. In general, the user computer is connected to Ethernet switch of his/her ISP where traffic is aggregated. The BNG (Broadband Network Gateway) network performs traffic management and authentication functions on the packets received by Ethernet switches. These BNG routers connect to other Internet routers through provider's edge routers. The core network is further comprised of many large routers. Each of these devices consumes power according to the traffic volume. According to the study conducted by Tucker [15], public Cloud is estimated to consume about 2.7 J/b in transmission and switching in comparison to 0.46J/b for a private Cloud. They found out that power consumption in transport represents a significant proportion of the total power consumption for Cloud storage services at medium and high usage rates. Even typical network usage can result in three to four times more energy consumption in public Cloud storage than one's own storage infrastructure. Therefore, with the growth of Cloud computing usage, it is expected that energy efficiency of switches and routers will play a very significant role in what since they need to provide capacity of hundreds of terabits of bandwidth.

In the network infrastructure, the energy consumption [16] depends especially on the power efficiency and awareness of wired network, namely the network equipments or system design, topology design, and network protocol design. Most of the energy in network devices is wasted because they are designed to handle worst case scenario. Therefore, the energy consumption of these devices remains almost the same during both peak time and idle state. Many improvements are required to get high energy efficiency in these devices. For example during low utilization periods, Ethernet links can be turned off and packets can be routed around them. Further energy savings are possible at the hardware level of the routers through appropriate selection and optimization of the layout of various internal router components (i.e. buffers, links, etc.).

3.4 Datacenter

The Cloud datacenters are quite different from traditional hosting facilities. A cloud datacenter could comprise of many hundreds or thousands of networked computers with their corresponding

storage and networking subsystems, power distribution and conditioning equipment, and cooling infrastructures. Due to large number of equipments, datacenters can consume massive energy consumption and emit large amount of carbon. According to 2007 report on computing datacenters by US Environmental Protection Agency (EPA), the datacenters in US consumed about 1.5% of total energy, which costs about $4.5 billon. This high usage also translates to very high carbon emissions which was estimated to be about 80-116 Metric Megatons each year. Table 3 lists equipments typically used in datacenters with their contribution to energy consumption. It can be clearly observed that servers and storage systems are not the only infrastructure that consumes energy in the datacenter. In reality, the cooling equipments consume equivalent amount of energy as the IT systems themselves. Ranganathan [17] suggests that for every dollar spent on electricity costs in large-scale datacenters another dollar is spent on cooling.

Table 2. Percent of Power Consumption by Each Datacenter Device

Cooling device (Chiller, Computer Room Air Conditioning (CRAC))	33%+9%
IT Equipment	30%
Electrical Equipments (UPS, Power Distribution Units (PDUs), lighting)	28%

Further energy consumption occurs due to lighting, loss in the power distribution, and other electrical equipment such as UPS. In other words, the majority of power usage within a datacenter is used for other purposes than actual IT services. Thus, to achieve the maximum efficiency in power consumption and CO_2 emissions, each of these devices need to be designed and used efficiently while ensuring that their carbon footprint is reduced.

A key factor in achieving the reduction in power consumption of a datacenter is to calculate how much energy is consumed in cooling and other overheads. Standard metrics are emerging such as Power Usage Effectiveness (PUE) [19] which can be used to benchmark how much energy is being usefully deployed versus how much is spent on overhead. The PUE of a datacenter is defined as the ratio of the total power consumption of a facility (data or switching center) to the total power consumption of IT equipment (servers, storage, routers, etc.). PUE varies from datacenters depending on the place where datacenter is located and devices used in its construction. Research from the Lawrence Berkley National Labs [18] shows that 22 datacenters measured in 2008 have PUE Values in the range 1.3 to 3.0. PUE of datacenter can be useful in measuring power efficiency of datacenters and thus provide a motivation to improve its efficiency.

4. Features of Clouds enabling Green computing

Even though there is a great concern in the community that Cloud computing can result in higher energy usage by the datacenters, the Cloud computing has a green lining. There are several technologies and concepts employed by Cloud providers to achieve better utilization and efficiency than traditional computing. Therefore, comparatively lower carbon emission is expected in Cloud computing due to highly energy efficient infrastructure and reduction in the IT infrastructure itself by multi-tenancy. The key driver technology for energy efficient Clouds is "Virtualization," which allows significant improvement in energy efficiency of Cloud providers

by leveraging the economies of scale associated with large number of organizations sharing the same infrastructure. Virtualization is the process of presenting a logical grouping or subset of computing resources so that they can be accessed in ways that give benefits over the original configuration [20]. By consolidation of underutilized servers in the form of multiple virtual machines sharing same physical server at higher utilization, companies can gain high savings in the form of space, management, and energy.

According to Accenture Report [7], there are following four key factors that have enabled the Cloud computing to lower energy usage and carbon emissions from ICT. Due to these Cloud features, organizations can reduce carbon emissions by atleast 30% per user by moving their applications to the Cloud. These savings are driven by the high efficiency of large scale Cloud data centers.

1. **Dynamic Provisioning:** In traditional setting, datacenters and private infrastructure used to be maintained to fulfill worst case demand. Thus, IT companies end up deploying far more infrastructure than needed. There are various reasons for such over-provisioning: a) it is very difficult to predict the demand at a time; this is particularly true for Web applications and b) to guarantee availability of services and to maintain certain level of service quality to end users. One example of a Web service facing these problems is a Website for the Australian Open Tennis Championship [21]. The Australian Open Website each year receives a significant spike in traffic during the tournament period. The increase in traffic can amount to over 100 times its typical volume (22 million visits in a couple of weeks) [21]. To handle such peak load during short period in a year, running hundreds of server throughout the year is not really energy efficient. Thus, the infrastructure provisioned with a conservative approach results in unutilized resources. Such scenarios can be readily managed by Cloud infrastructure. The virtual machines in a Cloud infrastructure can be live migrated to another host in case user application requires more resources. Cloud providers monitor and predict the demand and thus allocate resources according to demand. Those applications that require less number of resources can be consolidated on the same server. Thus, datacenters always maintain the active servers according to current demand, which results in low energy consumption than the conservative approach of over-provisioning.

2. **Multi-tenancy:** Using multi-tenancy approach, Cloud computing infrastructure reduces overall energy usage and associated carbon emissions. The SaaS providers serve multiple companies on same infrastructure and software. This approach is obviously more energy efficient than multiple copies of software installed on different infrastructure. Furthermore, businesses have highly variable demand patterns in general, and hence multi-tenancy on the same server allows the flattening of the overall peak demand which can minimize the need for extra infrastructure. The smaller fluctuation in demand results in better prediction and results in greater energy savings.

3. **Server Utilization:** In general, on-premise infrastructure run with very low utilization, sometimes it goes down up to 5 to 10 percent of average utilization. Using virtualization technologies, multiple applications can be hosted and executed on the same server in isolation, thus lead to utilization levels up to 70%. Thus, it dramatically reduces the number

of active servers. Even though high utilization of servers results in more power consumption, server running at higher utilization can process more workload with similar power usage.

4. **Datacenter Efficiency:** As already discussed, the power efficiency of datacenters has major impact on the total energy usage of Cloud computing. By using the most energy efficient technologies, Cloud providers can significantly improve the PUE of their datacenters. Today's state-of-the-art datacenter designs for large Cloud service providers can achieve PUE levels as low as 1.1 to 1.2, which is about 40% more power efficiency than the traditional datacenters. The server design in the form of modular containers, water or air based cooling, or advanced power management through power supply optimization, are all approaches that have significantly improved PUE in datacenters. In addition, Cloud computing allows services to be moved between multiple datacenter which are running with better PUE values. This is achieved by using high speed network, virtualized services and measurement, and monitoring and accounting of datacenter.

5. Towards Energy Efficiency of Cloud computing: State-of-the-Art

5.1 Applications

SaaS model has changed the way applications and software are distributed and used. More and more companies are switching to SaaS Clouds to minimize their IT cost. Thus, it has become very important to address the energy efficiency at application level itself. However, this layer has received very little attraction since many applications are already on use and most of the new applications are mostly upgraded version of or developed using previously implemented tools. Some of the efforts in this direction are for MPI applications [22], which are designed to run directly on physical machines. Thus, their performance on virtual machine is still undefined.

Various power efficient techniques [24][25] for software designs are proposed in the literature but these are mostly for embedded devices. In the development of commercial and enterprise applications which are designed for PC environment, generally energy efficiency is neglected. Mayo et al. [23] presented in their study that even simple tasks such as listening to music can consume significantly different amounts of energy on a variety of heterogeneous devices. As these tasks have the same purpose on each device, the results show that the implementation of the task and the system upon which it is performed can have a dramatic impact on efficiency. Therefore, to achieve energy efficiency at application level, SaaS providers should pay attention in deploying software on right kind of infrastructure which can execute the software most efficiently. This necessitates the research and analysis of trade-off between performance and energy consumption due to execution of software on multiple platforms and hardware. In addition, the energy consumption at the compiler level and code level should be considered by software developers in the design of their future application implementations using various energy-efficient techniques proposed in the literature.

5.2 Cloud Software Stack: Virtualization and Provisioning

In the Cloud stack, most works in the literature address the challenges at the IaaS provider level where research focus is on scheduling and resource management to reduce the amount of active

resources executing the workload of user applications. The consolidation of VMs, VM migration, scheduling, demand projection, heat management and temperature-aware allocation, and load balancing are used as basic techniques for minimizing power consumption. As discussed in previous section, virtualization plays an important role in these techniques due to its several features such as consolidation, live migration, and performance isolation. Consolidation helps in managing the trade-off between performance, resource utilization, and energy consumption [26]. Similarly, VM migration [48] allows flexible and dynamic resource management while facilitating fault management and lower maintenance cost. Additionally, the advancement in virtualization technology has led to significant reduction in VM overhead which improves further the energy efficiency of Cloud infrastructure.

Abdelsalam et al. [31] proposed a power efficient technique to improve the management of Cloud computing environments. They formulated the management problem in the form of an optimization model aiming at minimization of the total energy consumption of the Cloud, taking SLAs into account. The current issue of under utilization and over-provisioning of servers was highlighted by Ranganathan et al. [41]. They present a peak power budget management solution to avoid excessive over-provisioning considering DVS and memory/disk scaling. There are several other research work which focus on minimizing the over provisioning using consolidation of virtualized server [27]. Majority of these works use monitoring and estimation of resource utilization by applications based on the arrival rate of requests. However, due to multiple levels of abstractions, it is really hard to maintain deployment data of each virtual machine within a Cloud datacenter. Thus, various indirect load estimation techniques are used for consolidation of VMs.

Although above consolidation methods can reduce the overall number of resources used to serve user applications, the migration and relocation of VMs for matching application demand can impact the QoS service requirements of the user. Since Cloud providers need to satisfy a certain level of service, some work focused on minimizing the energy consumption while reducing the number of SLA violations. One of the first works that dealt with performance and energy trade-off was by Chase et al. [27] who introduced MUSE, an economy-based system of resource allocation. They proposed a bidding system to deliver the required performance level and switching off unused servers. Kephart et al. [29] addressed the coordination of multiple autonomic managers for power/performance tradeoffs using a utility function approach in a non-virtualized environment. Song et al. [30] proposed an adaptive and dynamic scheme for efficient sharing of a server by adjusting resources (specifically, CPU and memory) between virtual machines. At the operating system level, Nathuji et al. [28] proposed a power management system called VirtualPower integrating the power management and virtualization technologies. VirtualPower allows the isolated and independent operation of virtual machine to reduce the energy consumption. The soft states are intercepted by Xen hypervisor and are mapped to changes in the underlying hardware such as CPU frequency scaling according to the virtual power management rules.

In addition, there are works on improving the energy efficiency of storage systems. Kaushik et al. [32] presented an energy conserving self-adaptive Commodity Green Cloud storage called Lightning. The Lightning file system divides the Storage servers into Cold and Hot logical zones using data classification. These servers are then switched to inactive states for energy saving.

Verma et al [33] proposed an optimization for storage virtualization called Sample-Replicate-Consoidate Mapping (SRCMAP) which enables the energy proportionality for dynamic I/O workloads by consolidating the cumulative workload on a subset of physical volumes proportional to the I/O workload intensity. Gurumurthi et al. [39] proposed intra-disk parallelism on high capacity drives to improve disk bandwidth without increasing power consumption. Soror et al. [34] addressed the problem of optimizing the performance of database management systems by controlling the configurations of the virtual machines in which they run.

Since power is dissipated in Cloud datacenter due to heat generated by the servers, several work also have been proposed for dynamic scheduling of VMs and applications which take into account the thermal states or the heat dissipation in a data centre. The consideration of thermal factor in scheduling also improves the reliability of underline infrastructure. Tang et al. [36] formulated the problem using a mathematical model for maximizing the cooling efficiency of a data center. Heath et al. [40] proposed emulation tools for investigating the thermal implications of power management. Ramos et al. [37] proposed a software prediction infrastructure called C-Oracle that makes online predictions for data center thermal management based on load redistribution and DVS. Moore et al. [35] proposed a method for automatic reconfiguration of thermal load management system taking into account thermal behavior for improving cooling efficiency and power consumption. They also propose thermal management solutions focusing on scheduling workloads considering temperature-aware workload placement. Bash et al. [36] propose a workload placement policy for a datacenter that allocate resources in the areas which are easier to cool resulting in cooling power savings. Raghavendra et al. [38] propose a framework which coordinates and unifies five individual power management solutions (consisting of HW/SW mechanisms).

5.3 Datacenter level: Cooling, Hardware, Network, and Storage

The rising energy costs, cost savings and a desire to get more out of existing investments are making today's Cloud providers to adopt best practices to make datacenters operation green. To build energy efficient datacenter, several best practices has been proposed to improve efficiency of each device from electrical systems to processor level.

First level is the smart construction of the datacenter and choosing of its location. There are two major factors in that one is energy supply and other is energy efficiency of equipments. Hence, the datacenters are being constructed in such a way that electricity can be generated using renewable sources such as sun and wind. Currently the datacenter location is decided based on their geographical features; climate, fibre-optic connectivity and access to a plentiful supply of affordable energy. Since main concern of Cloud providers is business, energy source is also seen mostly in terms of cost not carbon emissions.

Another area of concern within a datacenter is its cooling system that contributes to almost 1/3 of total energy consumption. Some research studies [17][35] have shown that uneven temperature within datacenter can also lead significant decline in reliability of IT systems. In datacenter cooling, two types of approaches are used: air and water based cooling systems. In both approaches, it is necessary that they directly cool the hot equipment rather than entire room area. Thus newer energy efficient cooling systems are proposed based on liquid cooling, nano fluid-

cooling systems, and in-server, in-rack, and in-row cooling by companies such as SprayCool. Other than that, the outside temperature/climate can have direct impact on the energy requirement of cooling system. Some systems have been constructed where external cool air is used to remove heat from the datacenter [42].

Another level at which datacenter's power efficiency is addressed is on the deployment of new power efficient servers and processors. Low energy processors can reduce the power usage of IT systems in a great degree. Many new energy efficient server models are available currently in market from vendors such as AMD, Intel, and others; each of them offering good performance/watt system. These server architecture enable slowing down CPU clock speeds (clock gating), or powering off parts of the chips (power gating), if they are idle. Further enhancement in energy saving and increasing computing per watt can be achieved by using multi-core processors. For instance Sun's multicore chips, each 32-thread Niagara chip, UltraSPARC 1, consumes about 60 watts, while the two Niagara chips have 64 threads and run at about 80 watts. However, the exploitation of such power efficiency of multi-core system requires software which can run on multi-CPU environment. Here, virtualization technologies play an important role. Similarly, consolidation of storage system helps to further reduce the energy requirements of IT Systems. For example, Storage Area Networks (SAN) allow building of an efficient storage network that consolidates all storage. The use of energy efficient disks such as tiered storage (Solid-State, SATA, SAS) allows better energy efficiency.

The power supply unit is another infrastructure which needs to be designed in an energy efficient manner. Their task is to feed the server resources with power by converting the high-voltage alternating current (AC) from the power grid to a low-voltage direct current (DC) which most of the electric circuits (e.g. computers) require. These circuits inside Power Supply Unit (PSU) inevitably lose some energy in the form of heat, which is dissipated by additional fans inside PSU. The energy efficiency of a PSU mainly depends on its load, number of circuits and other conditions (e.g. temperature). Hence, a PSU which is labeled to be 80% efficient is not necessarily that efficient for all power loads. For example, low power loads tend to be the most energy inefficient ones. Thus, a PSU can be just 60% efficient at 20% of power load. Some studies have found that PSUs are one of the most inefficient components in today's data centers as many servers are still shipped with low quality 60 to 70 percent efficient power supplies. One possible solution offered is to replace all PSUs by ENERGY STAR certified ones. This certificate is given to PSUs which guarantee a minimum 80% efficiency at any power load.

5.4 Monitoring/Metering

It is said that *you cannot improve what you do not measure*. It is essential to construct power models that allow the system to know the energy consumed by a particular device, and how it can be reduced. To measure the unified efficiency of a datacenter and improve its' performance per-watt, the Green Grid has proposed two specific metrics known as the Power Usage Effectiveness (PUE) and Datacenter Infrastructure Efficiency (DciE) [19].

- **PUE** = Total Facility Power/IT Equipment Power
- **DciE** = 1/PUE = IT Equipment Power/Total Facility Power x 100%

Here, the Total Facility Power is defined as the power measured at the utility meter that is dedicated solely to the datacenter power. The IT Equipment Power is defined as the power consumed in the management, processing, and storage or routing of data within the datacenter. PUE and DCIE are most common metrics designed to compare the efficiency of datacenters. There are many systems in the marketplace for such measurements. For instance Sun[SM] Eco Services measures at a higher level rather than attempting to measure each individual device's power consumption. For measuring and modeling the power usage of storage system, Researchers from IBM [44] have proposed a scalable, enterprise storage modelling framework called STAMP. It side steps the need for detailed traces by using interval performance statistics and a power table for each disk model. STAMP takes into account controller caching and algorithms, including protection schemes, and adjusts the workload accordingly. To measure the power consumed by a server (e.g. PowerEdge R610) the Intelligent Platform Management Interface (IPMI) [43] is proposed. This framework provides a uniform way to access the power-monitoring sensors available on recent servers. This interface being independent of the operating system can be accessed despite of operating system failures and without the need of the servers to be powered on (i.e. connection to the power grid is enough). Further, intelligent power distribution units (PDUs), traditional power meters (e.g. Watts Up Pro power meter) and ACPI enabled power supplies can be used to measure the power consumption of the whole server.

5.5. Network Infrastructure

As discussed previously, at network level, the energy efficiency is achieved either at the node level (i.e. network interface card) or at the infrastructure level (i.e. switches and routers). The energy efficiency issues in networking is usually referred to as "green networking", which relates to embedding energy-awareness in the design, in the devices and in the protocols of networks. There are four classes of solutions offered in literature, namely resource consolidation, virtualization, selective connectedness, and proportional computing. Resource consolidation helps in regrouping the under-utilized devices to reduce the global consumption. Similar to consolidation, selective connectedness of devices [44][45] consists of distributed mechanisms which allow the single pieces of equipment to go idle for some time, as transparently as possible from the rest of the networked devices. The difference between resource consolidation and selective connectedness is that the consolidation applies to resources that are shared within the network infrastructure while selective connectedness allows turning off unused resources at the edge of the network. Virtualization as discussed before allows more than one service to operate on the same piece of hardware, thus improving the hardware utilization. Proportional computing [45] can be applied to a system as a whole, to network protocols, as well as to individual devices and components. Dynamic Voltage Scaling and Adaptive Link Rate are typical examples of proportional computing. Dynamic Voltage Scaling [45] reduces the energy state of the CPU as a function of a system load, while Adaptive Link Rate applies a similar concept to network interfaces, reducing their capacity, and thus their consumption, as a function of the link load. The survey by Bianzino et al. [46] gives more details about the work in the area of Green networking.

6. Green Cloud Architecture

From the above study of current efforts in making Cloud computing energy efficient, it shows that even though researchers have made various components of Cloud efficient in terms of power

and performance, still they lack a unified picture. Most of efforts for sustainability of Cloud computing have missed the network contribution. If the file sizes are quite large, network will become a major contributor to energy consumption; thus it will be greener to run application locally than in Clouds. Furthermore, many work focused on just particular component of Cloud computing while neglecting effect of other, which may not result in overall energy efficiency. For example, VM consolidation may reduce number of active servers but it will put excessive load on few servers where heat distribution can become a major issue. Some other works just focus on redistribution of workload to support energy efficient cooling without considering the effect of virtualization. In addition, Cloud providers, being profit oriented, are looking for solutions which can reduce the power consumption and thus, carbon emission without hurting their market. Therefore, we provide a unified solution to enable Green Cloud computing. We propose a Green Cloud framework, which takes into account these goals of provider while curbing the energy consumption of Clouds. The high level view of the green Cloud architecture is given in Figure 7. The goal of this architecture is to make Cloud green from both user and provider's perspective.

In the Green Cloud architecture, users submit their Cloud service requests through a new middleware Green Broker that manages the selection of the greenest Cloud provider to serve the user's request. A user service request can be of three types i.e., software, platform or infrastructure. The Cloud providers can register their services in the form of 'green offers' to a public directory which is accessed by Green Broker. The green offers consist of green services, pricing and time when it should be accessed for least carbon emission. Green Broker gets the current status of energy parameters for using various Cloud services from Carbon Emission Directory. The Carbon Emission Directory maintains all the data related to energy efficiency of Cloud service. This data may include PUE and cooling efficiency of Cloud datacenter which is providing the service, the network cost and carbon emission rate of electricity, Green Broker calculates the carbon emission of all the Cloud providers who are offering the requested Cloud service. Then, it selects the set of services that will result in least carbon emission and buy these services on behalf users.

The Green Cloud framework is designed such that it keeps track of overall energy usage of serving a user request. It relies on two main components, Carbon Emission Directory and Green Cloud offers, which keep track of energy efficiency of each Cloud provider and also give incentive to Cloud providers to make their service "Green". From user side, the Green Broker plays a crucial role in monitoring and selecting the Cloud services based on the user QoS requirements, and ensuring minimum carbon emission for serving a user. In general, a user can use Cloud to access any of these three types of services (SaaS, PaaS, and IaaS), and therefore process of serving them should also be energy efficient. In other words, from the Cloud provider side, each Cloud layer needs to be "Green" conscious.

- **SaaS Level:** Since SaaS providers mainly offer software installed on their own datacenters or resources from IaaS providers, the SaaS providers need to model and measure energy efficiency of their software design, implementation, and deployment.
  For serving users, the SaaS provider chooses the datacenters which are not only energy efficient but also near to users. The minimum number of replicas of user's confidential data should be maintained using energy-efficient storage.

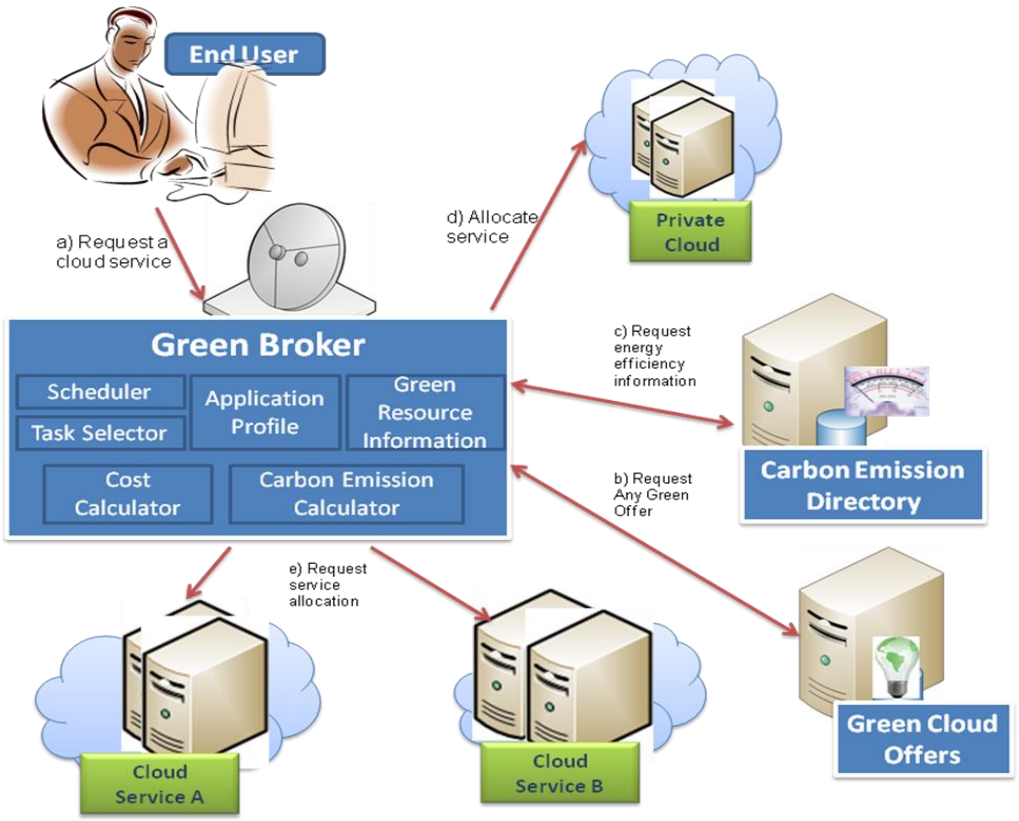


Figure 7. Green Cloud Architecture

- **PaaS level:** PaaS providers offer in general the platform services for application development. The platform facilitates the development of applications which ensures system wide energy efficiency. This can be done by inclusion of various energy profiling tools such as JouleSort [5]. It is a software energy efficiency benchmark that measures the energy required to perform an external sort. In addition, platforms itself can be designed to have various code level optimizations which can cooperate with underlying complier in energy efficient execution of applications. Other than application development, Cloud platforms also allow the deployment of user applications on Hybrid Cloud. In this case, to achieve maximum energy efficiency, the platforms profile the application and decide which portion of application or data should be processed in house and in Cloud.

- **IaaS level:** Providers in this layer plays most crucial role in the success of whole Green Architecture since IaaS level not only offer independent infrastructure services but also support other services offered by Clouds. They use latest technologies for IT and cooling systems to have most energy efficient infrastructure. By using virtualization and consolidation, the energy consumption is further reduced by switching-off unutilized server. Various energy meters and sensors are installed to calculate the current energy efficiency of each IaaS providers and their sites. This information is advertised regularly by Cloud providers in Carbon Emission Directory. Various green scheduling and resource provisioning policies will ensure minimum energy usage. In addition, the Cloud

provider designs various green offers and pricing schemes for providing incentive to users to use their services during off-peak or maximum energy-efficiency hours.

6. Case Study: IaaS Provider

In this section, we describe a case study example to illustrate the working of the proposed Green Architecture in order to highlight the importance of considering the unifying picture to reduce the energy and carbon emissions by Cloud infrastructure. The case study focuses on IaaS service providers. Our experimental platform consists of multiple Cloud providers who offer computational resources to execute user's HPC applications. A user request consists of application, its estimated length in time and number of resources required. These applications are submitted to the Green broker who acts as an interface to the Cloud infrastructure and schedules applications on behalf of users as shown in Figure 7. The Green Broker interprets and analyzes the service requirements of a submitted application and decides where to execute it. As discussed, Green Broker's main objective is to schedule applications such that the CO_2 emissions are reduced and the profit is increased, while the Quality of Service (QoS) requirements of the applications are met. As Cloud data centers are located in different geographical regions, they have different CO_2 emission rates and energy costs depending on regional constraints. Each datacenter is responsible for updating this information to Carbon Emission Directory for facilitating the energy-efficient scheduling. The list of energy related parameters is given in Table 3.

Table 3. Carbon Emission Related Parameter of a Datacenter

Parameter	Notation
Carbon emission rate (kg/kWh)	$r_i^{CO_2}$
Average COP	COP_i
Electricity price ($/kWh)	p_i^e
Data transfer price ($/GB) for up-load/download	p_i^{DT}
CPU power	$P_i = \beta_i + \alpha_i f^3$
CPU frequency range	$[f_i^{min}, f_i^{max}]$
Time slots (start time, end time, number of CPUs)	(t_s, t_e, n)

In order to validate our framework and to prove that it achieves better efficiency in terms of carbon emission, we have studied five policies (Green and profit-oriented) employed for scheduling by Green Broker.

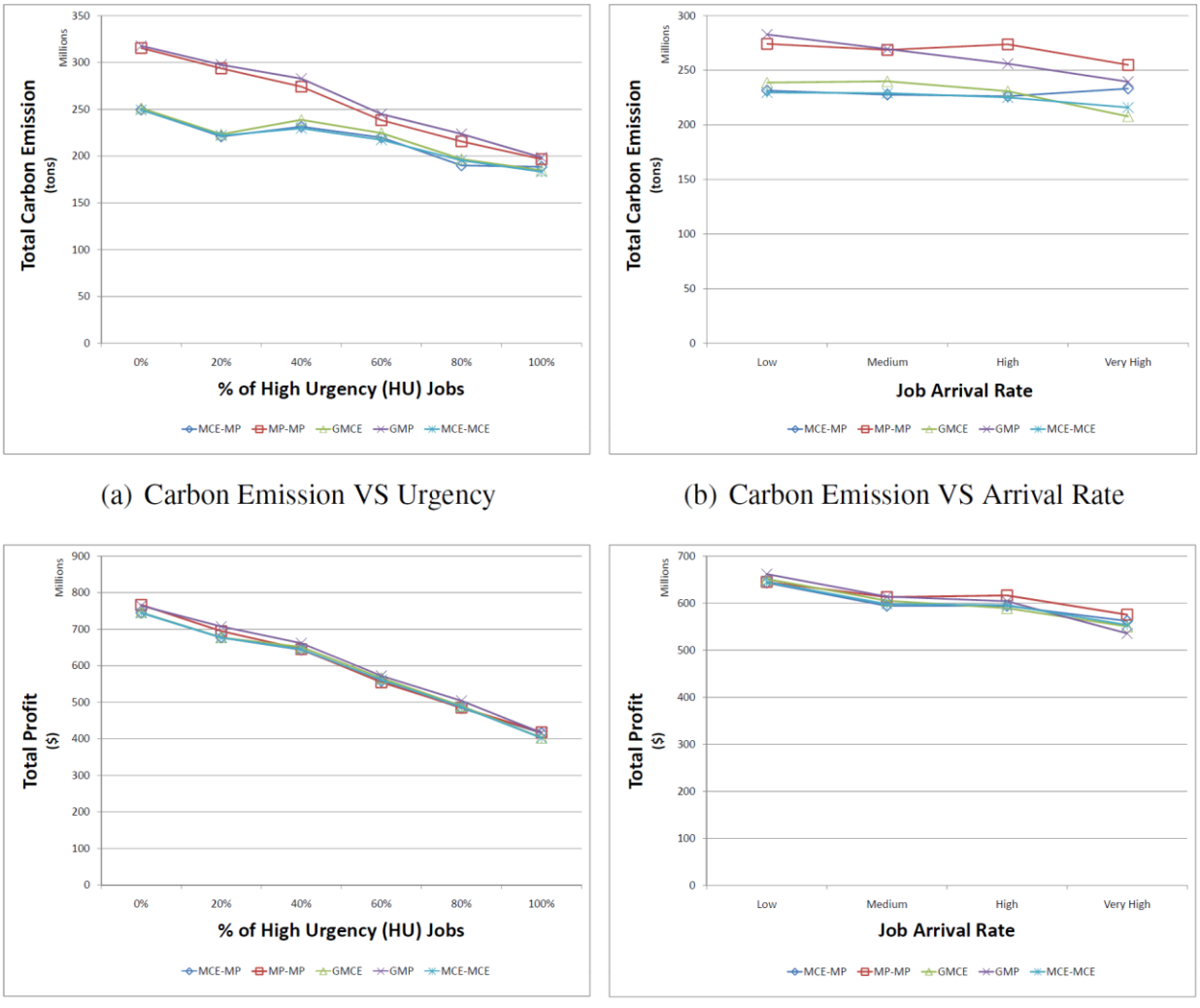
a) *Greedy Minimum Carbon Emission (GMCE):* In this policy, user applications are assigned to Cloud providers in greedy manner based on their carbon emission.

b) *Minimum Carbon Emission - Minimum Carbon Emission (MCE-MCE):* This is a double greedy policy where applications are assigned to the Cloud providers with minimum

Carbon emission due to their datacenter location and Carbon emission due to application execution.

c) *Greedy Maximum Profit (GMP):* In this policy, user applications are assigned in greedy manner to a provider who execute the application fastest and get maximum profit. .
d) *Maximum Profit - Maximum Profit (MP-MP*): This is double greedy policy considering profit made by Cloud providers and application finishes by its deadline.
e) *Minimising Carbon Emission and Maximising Profit (MCE-MP):* In this policy, the broker tries to schedule the applications to those providers which results in minimization of total carbon emission and maximization of profit.

Above GMCE, MCE-MCE and MCE-MP are "Green" policies while MP-MP and GMP are profit-oriented policies. A more extensive detail on modeling of energy efficiency of a Cloud datacenter, experimental data and results is available in previous work [47]. Here, we present some important results to illustrate the validity of our presented framework.



(a) Carbon Emission VS Urgency

(b) Carbon Emission VS Arrival Rate

(c) Profit VS Urgency

(d) Profit VS Arrival Rate

Figure 8. Carbon Emission and Profit of Provider Using Green Cloud Framework

Figure 8 shows the course of experiments conducted with varying user's urgency for executing his application and job arrival rate. The metrics of total carbon emission and total profit are used

since the resource provider needs to know the collective loss in carbon emission and gain in profit across all datacenters. From these results three main inferences can be made.

- Green policies reduce the carbon emission by almost 20% in comparison to profit based policies. This observation emphasizes the inclusion of overall carbon efficiency of all the Cloud providers in scheduling decisions.
- With the increase in user's urgency to execute the application, the gain in carbon emission reduces almost linearly. This clearly shows how important is the role of user in making Cloud computing in general "Green". If users are more patient and schedule the applications when the datacenters are running at higher energy efficiency, more energy and carbon gain can be made. Thus, in our framework, we introduce the need of Green Cloud Offers from providers.
- The green policies also have minimal effect on the provider's profit. This clearly shows that by using energy efficient solutions such as Green Cloud Framework both Cloud providers and users can benefit.

7. Conclusions and Future Directions

Cloud computing business potential and contribution to already aggravating carbon emission from ICT, has lead to a series of discussion whether Cloud computing is really green. It is forecasted that the environmental footprint from data centers will triple between 2002 and 2020, which is currently 7.8 billion tons of CO_2 per year. There are reports on Green IT analysis of Clouds and datacenters that show that Cloud computing is "Green", while others show that it will lead to alarming increase in Carbon emission. Thus, in this chapter, we first analyzed the benefits offered by Cloud computing by studying its fundamental definitions and benefits, the services it offers to end users, and its deployment model. Then, we discussed the components of Clouds that contribute to carbon emission and the features of Clouds that make it "Green". We also discussed several research efforts and technologies that increase the energy efficiency of various aspects of Clouds. For this study, we identified several unexplored areas that can help in maximizing the energy efficiency of Clouds from a holistic perspective. After analyzing the shortcoming of previous solutions, we proposed a Green Cloud Framework and presented some results for its validation. Even though our Green Cloud framework embeds various features to make Cloud computing much more Green, there are still many technological solutions are required to make it a reality:

- First efforts are required in designing software at various levels (OS, compiler, algorithm and application) that facilitates system wide energy efficiency. Although SaaS providers may still use already implemented software, they need to analyze the runtime behavior of applications. The gathered empirical data can be used in energy efficient scheduling and resource provisioning. The compiler and operating systems need to be designed in such a way that resources can be allocated to application based on the required level of performance, and thus performance versus energy consumption tradeoff can be managed.
- To enable the green Cloud datacenters, the Cloud providers need to understand and measure existing datacenter power and cooling designs, power consumptions of servers and their cooling requirements, and equipment resource utilization to achieve maximum efficiency. In addition, modeling tools are required to measure the energy usage of all

the components and services of Cloud, from user PC to datacenter where Cloud services are hosted.

- For designing the holistic solutions in the scheduling and resource provisioning of applications within the datacenter, all the factors such as cooling, network, memory, and CPU should be considered. For instance, consolidation of VMs even though effective technique to minimize overall power usage of datacenter, also raises the issue related to necessary redundancy and placement geo-diversity required to be maintained to fulfill SLAs with users. It is obvious that last thing Cloud provider will want is to lose their reputation by their bad service or violation of promised service requirements.
- Last but not the least, the responsibility also goes to both providers and customers to make sure that emerging technologies do not bring irreversible changes which can bring threat to the health of human society. The way end users interact with the application also has a very real cost and impact. For example, purging of unsolicited emails can eliminates energy wasted in storage and network. Similarly, if Cloud providers want to provide a truly green and renewable Cloud, they must deploy their datacenters near renewable energy sources and maximize the Green energy usage in their already established datacenters. Before adding new technologies such as virtualization, proper analysis of overhead should be done real benefit in terms of energy efficiency.

In conclusion, by simply improving the efficiency of equipment, Cloud computing cannot be claimed to be Green. What is important is to make its usage more carbon efficient both from user and provider's perspective. Cloud Providers need to reduce the electricity demand of Clouds and take major steps in using renewable energy sources rather than just looking for cost minimization.

Acknowledgements

We thank Professor Thulasiram Ruppa, Dr. Rodrigo Calheiros and Professor Rod Tucker, for their comments on improving this chapter.

References

[1] Gleeson, E. 2009. Computing industry set for a shocking change. Retrieved January 10, 2010 from http://www.moneyweek.com/investment-advice/computing-industry-set-for-a-shocking-change-43226.aspx

[2] Buyya, R., Yeo, C.S. and Venugopal, S. 2008. Market-oriented Cloud computing: Vision, hype, and reality for delivering it services as computing utilities. *Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications*, Los Alamitos, CA, USA.

[3] New Datacenter Locations. 2008. http://royal.pingdom.com/2008/04/11/map-of-all-google-data-center-locations/

[4] Bianchini, R., and Rajamony, R. 2004, Power and energy management for server systems, Computer, 37 (11) 68-74.

[5] Rivoire, S., Shah, M. A., Ranganathan, P., and Kozyrakis, C. 2007. Joulesort: a balanced energy-efficiency benchmark, Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, NY, USA.

[6] Greenpeace International. 2010. Make IT Green http://www.greenpeace.org/international/en/publications/reports/make-it-green-Cloudcomputing/
[7] Accenture Microsoft Report. 2010. Cloud computing and Sustainability: The Environmental Benefits of Moving to the Cloud, http://www.wspenvironmental.com/media/docs/newsroom/Cloud_computing_and_Sustainability_-_Whitepaper_-_Nov_2010.pdf.
[8] Mell, P. and Grance, T. 2009. The NIST Definition of Cloud computing, National Institute of Standards and Technology.
[9] Google App Engine. 2010. http://code.google.com/appengine/.
[10] Vecchiola, C., Chu, X. and Buyya, R. 2009. Aneka: A Software Platform for .NET-based Cloud Computing. In *High Performance & Large Scale computing*, Advances in Parallel Computing, ed. W. Gentzsch, L. Grandinetti and G. Joubert, IOS Press.
[11] Microsoft Azure. 2011. www.microsoft.com/windowsazure/
[12] Charrington, S. 2010, Characteristics of Platform as a Service, Cloud Pulse blog, http://Cloudpulseblog.com/2010/02/the-essential-characteristics-of-paas.
[13] Cherkasova, L. and Gardner, R. 2005, Measuring CPU overhead for I/O processing in the Xen virtual machine monitor. *Proceeding of 2005 Annual Technical Conference on USENIX, Anaheim*, CA, USA.
[14] Sarokin, D. 2007. Question: Energy use of Internet, http://uclue.com/?xq=724.
[15] Baliga J., Ayre R., Hinton K., and Tucker R. S. 2010. Green Cloud computing: Balancing energy in processing, storage and transport. *Proceedings of the IEEE*, 99(1) 149-167.
[16] Chabarek, J., Sommers, J., Barford, P., Estan, C., Tsiang, D., and Wright, S. 2008. Power Awareness in Network Design and Routing. *Proceedings of 27th IEEE INFOCOM, Pheonix*, AZ, USA.
[17] Ranganathan P, 2010, Recipe for efficiency: principles of power-aware computing. *Communication. ACM*, 53(4):60–67.
[18] Greenberg, S., Mills, E., Tschudi, B., Rumsey, P., and Myatt, B., 2008, Best Practices for Data Centers: Lessons Learned from Benchmarking 22 Data Centers. ACEEE Summer Study on Energy Efficiency in Buildings. Retrieved September 4, 2008, from http://eetd.lbl.gov/emills/PUBS/PDF/ACEEE-datacenters.pdf
[19] Rawson, A., Pfleuger, J., and Cader, T., 2008. Green Grid Data Center Power Efficiency Metrics. Consortium Green Grid.
[20] Smith, J. and Nair, R. 2003. *Virtual Machines: Versatile Platforms for Systems and Processes*. Morgan Kaufmann: Los Altos, CA.
[21] IBM. 2008. ”Take the tennis to 1.9 billion viewersworldwide? Done.,” 04/05/08, 2008; http://www- 07.ibm.com/innovation/au/ausopen/pdf/CaseStudy 01.pdf
[22] Freeh, V. W., Pan, F., Kappiah, N., Lowenthal, D. K., and Springer, R. 2005. Exploring the energy-time trade-off in MPI programs on a power-scalable cluster, *Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium*, CA, USA.
[23] Mayo, R. N. and Ranganathan P., 2005. Energy consumption in mobile devices: Why future systems need requirements-aware energy scale-down. *Proceedings of 3rd International Workshop on Power-Aware Computer Systems*, San Diego, CA, USA.
[24] Saxe, E. 2008. Power Efficient Software. *Communication of the ACM*. 53(2) 44-48.
[25] Großschädl, Avanzi R.M., Savas E., and Tillich S. 2005. Energy-efficient software implementation of long integer modular arithmetic, *Proceedings of 7th Workshop on Cryptographic Hardware and Embedded Systems*, Edinburg, Scotland.

[26] Srikantaiah, S., Kansal, A., and Zhao, F. 2008. Energy aware consolidation for Cloud computing. *Proceedings of HotPower '08 Workshop on Power Aware computing and Systems*, San Diego, CA, USA.

[27] Chase, J.S., Anderson, D.C., Thakar, P.N., Vahdat, A.M., and Doyle, R.P. 2001. Managing energy and server resources in hosting centers. *Proceedings of 18th ACM Symposium on Operating Systems Principles (SOSP '01)*, Banff, Canada.

[28] Nathuji, R. and Schwan, K. 2007. VirtualPower: coordinated power management in virtualized enterprise systems. *Proceedings of 21st ACM SIGOPS Symposium on Operating Systems Principles*, Stevenson, WA, USA.

[29] Kephart, J. O., Chan, H., Das, R., Levine, D. W., Tesauro, G., Rawson, F., and Lefurgy, C. 2007. Coordinating multiple autonomic managers to achieve specified power-performance tradeoffs. *Proceedings of 4th International Conference on Autonomic Computing*, Florida, USA.

[30] Song, Y., Sun, Y., Wang, H., and Song, X. 2007. An adaptive resource flowing scheme amongst VMs in a VM-based utility computing. *Proceedings of IEEE International Conference on Computer and Information Technology*, Fukushima, Japan.

[31] Abdelsalam, H., Maly, K., Mukkamala, R., Zubair, M., and Kaminsky, D. 2009. Towards energy efficient change management in a Cloud computing environment, *Proceedings of 3rd International Conference on Autonomous Infrastructure, Management and Security*, The Netherlands.

[32] Kaushik, R. T., Cherkasova, L., Campbell, R., and Nahrstedt, K., 2010. Lightning: self-adaptive, energy-conserving, multi-zoned, commodity green Cloud storage system. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed computing* (HPDC '10). ACM, New York, NY, USA.

[33] Verma, A., Koller, R., Useche, L., and Rangaswami, R., 2010. SRCMap: energy proportional storage using dynamic consolidation. 2010, *Proceedings of the 8th USENIX conference on File and storage technologies (FAST'10)*, San Jose, California

[34] Soror, A. A., Minhas, U. F., Aboulnaga, A., Salem, K., Kokosielis, P., and Kamath, S. 2008. Automatic virtual machine configuration for database workloads. *Proceedings of ACM SIGMOD International Conference on Management of data*, Vancouver, Canada

[35] Moore, J. D., Chase, J. S., and Ranganathan, P. 2006. Weatherman: Automated, online and predictive thermal mapping and management for datacenters, *Proceedings of the 3rd* International Conference on Autonomic computing, Dublin, Ireland.

[36] Bash, C. and Forman, G. 2007. Cool job allocation: Measuring the power savings of placing jobs at cooling-efficient locations in the datacenter, *Proceeding of 2007 Annual Technical Conference on USENIX*, Santa Clara, CA, USA.

[37] Ramos, L. and Bianchini, R., 2008, C-oracle: Predictive thermal management for data centers, Proceedings of 14th International Symposium on High- Performance Computer Architecture, Salt Lake City, UT, USA.

[38] Raghavendra, R., Ranganathan, P., Talwar, V., Wang, Z., and Zhu Z. 2008. No "power" struggles: coordinated multi-level power management for the datacenter, *SIGOPS Operating System Review*, 42(2) 48–59.

[39] Gurumurthi, S., Stan, M. R., and Sankar, S., 2009, Using intra-disk parallelism to build energy-efficient storage systems, *IEEE Micro Special Issue on Top Picks from the Computer Architecture Conferences of 2008*.

[40] Heath, T., Centeno, A. P., George, P., Ramos, L., Jaluria, Y., and Bianchini, R. 2006. Mercury and Freon: temperature emulation and management for server systems. *Proceedings of Twelfth* International Conference *on* Architectural Support for Programming Languages and Operating Systems, San Jose, CA, USA.

[41] Ranganathan, P., Leech, P., Irwin, D., and Chase, J. 2006. Ensemble level power management for dense blade servers, *SIGARCH Computer Architecture News*, 34(2)66–77.

[42] Woods, A. 2010. Cooling the data center. *Communications of the ACM*, 53(4):36-42.

[43] Giri, Ravi A. 2010. Increasing Datacenter efficiency with server power measurements. *http://*download.intel.com/it/pdf/Server_Power_Measurement_final.pdf

[44] Allalouf, M., Arbitman, Y., Factor, M., Kat, R. I., Meth, K., and Naor, D. 2009. Storage modelling for power estimation. *In Proceedings of 2009 Israeli Experimental Systems Conference (SYSTOR '09)*, Isreal.

[45] Allman, M., Christensen, K., Nordman, B., and Paxson, V. 2007. Enabling an Energy-Efficient Future Internet Through Selectively Connected End Systems, *Proceedings of the Sixth ACM Workshop on Hot Topics in Networks (HotNets-VI)*, Atlanta, Georgia, USA.

[46] Bianzino, P., Chaudet, C., Rossi, D., and Rougier, J. (2011). *A Survey of Green Networking Research, IEEE Communications Surveys and Tutorials, IEEE, USA (in press.)*

[47] Garg, S. K., Yeo, C. S., Anandasivam, A., and Buyya, R. 2011. Environment-conscious scheduling of HPC applications on distributed Cloud-oriented datacenters, *Journal of Parallel and Distributed computing (JPDC)*, 71(6):732-749.

[48] Beloglazov, A, Buyya, R, Lee, YC, and Zomaya, A. 2011. A Taxonomy and Survey of Energy-Efficient Data Centers and Cloud computing Systems, *Advances in Computers*, M. Zelkowitz (ed), ISBN 13: 978-0-12-012141-0, Elsevier, Amsterdam, The Netherlands.

[49] Allenotor, D., Thulasiram, R. K. 2008. Grid resources pricing: A novel financial option based quality of service-profit quasi-static equilibrium model, *Proceedings of the 8th ACM/IEEE International Conference on Grid Computing*, Tsukuba, Japan.