Energy Consumption in Cloud Computing Environments

Kenga Mosoti Derdus[1], Dr. Vincent Omwenga O.[2], Prof. Patrick Ogao J.[3]
Email: derduskenga@gmail.com[1], vomwenga@strathmore.edu[2], ogaopj@gmail.com[3]
Faculty of Information Technology, Strathmore University, Kenya[1 & 2].
School of Computing and Information Technology, Technical University of Kenya, Kenya[3].

***Abstract* –** Datacentres are becoming indispensable infrastructure for supporting the services offered by cloud computing. Unfortunately, they consume a great deal of energy accounting for 3% of global electrical energy consumption. The effect of this is that, cloud providers experience high operating costs, which leading to increased Total Cost of Ownership (TCO) of datacentre infrastructure. Moreover, there is increased carbon dioxide emissions that affects the universe. This paper presents a survey on the various ways in which energy is consumed in datacentre infrastructure. The factors that influence energy consumption within a datacentre is presented as well.

Keywords: *Datacentre energy consumption, datacentre energy, Cloud computing*

I. INTRODUCTION

The excessive energy consumption datacentres has become a major concern to cloud computing practitioners. This is because they consume a great deal of energy accounting for 3% of global electrical energy consumption [1]. The effect of this is that, cloud providers experience high operating costs [2], which leading to increased Total Cost of Ownership (TCO) of datacentre infrastructure. The effect of high TCO is low Return on Investment (ROI). Moreover, there is increased carbon dioxide emissions that affects the universe. The reason for increased installation of datacentres is to enable cloud users to benefit from the many advantages of cloud computing such as cost-effectiveness, ease of management and on-demand scalability, as well as ensuring Quality of Service (QoS) and Service Level Agreement (SLA) [3]. According to [4], an average datacentre consumes as much energy as 25, 000 households.

Apart from low ROI, excessive energy consumpption has a negative impact on the environment, which is carbon dioxide (CO_2) emmision. According to [5], the ICT industry is estimated to contribute about 2% of global CO_2 emission, which contributes greatly to greenhouse effect – this emission is equivalent to the aviation industry. Worldwide datacenter energy consumption rose steadily steadily from year 2000 to 2010. In 2010, data center accounted for about 1.5% of total energy consumed worldwide [6]. As shown in **Figure 1**, datacenter energy consumption will continue to rise.

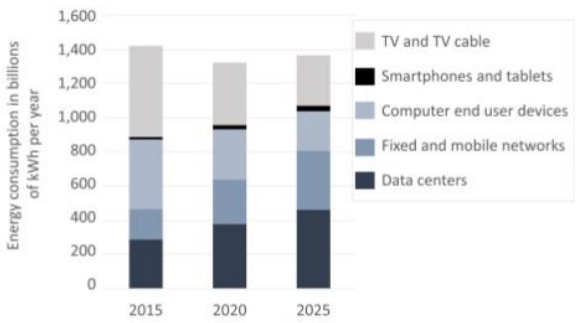

Figure 1: Forecast: Energy consumption of global cloud computing [7]

The high energy usage in the cloud is attributed to energy wastage and innefficincies related to the way electrical energy is delivered to the computing resources and the server at large and largely in the way these resources are used by applications workloads [3]. For example low server utilization and idle power westage are a major source of energy westage in a cloud computing environment.

A. What is cloud Computing?

Cloud computing is a model that provides computing resources on demand or on rental basis and so users can pay only for resources they use [3]. Therefore, customers can purchase a specific set of resources when they need it instead of renting a fixed amount of physical server. [7] defines cloud computing as "*... 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 ". By shared pool, resources are collected together and then dynamically allocated regardless of their physical location. On the other hand, network access allows the collected resources to be accessed via a network. In addition, rapid provisioning capability allows the service offering to scale so that the changing demands by cloud users are met. Cloud computing allows applications to be accessed via the internet using a browser, as well as hardware systems and systems software in the datacentres that manage user applications.

B. Virtualization in Cloud Computing

Virtualization is the main technology backing up cloud computing and it is based on physical resources abstraction in a way that several virtual resources are multiplexed on a physical one [8]. Virtualization provides high resource utilization as compared to traditional computing, flexibly, elasticity. This makes it possible to run multiple services or applications in the same PM including operating systems. A server is divided into number small servers known as Virtual Machines (VMs), which can run different applications independently and a VM can be moved from one PM to another (**Figure 2**) [8].

The hypervisor or Virtual Machine Manager (VMM) is software layer, which induces the partitioning capability and may run directly on the hardware or on a host operating system [8]. The VMM is responsible for managing physical resources. A host machine is the PM in which a VMM runs. Examples of VMMs are Xen, VMWare and KVM [8]. A VM is a representation of a real machine using a software, which provides virtual operating environment in which an operating system runs. A VM is referred to as a guest machine and it runs a guest operating system.

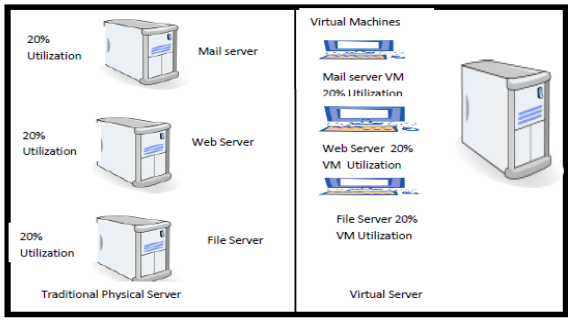

Figure 2: Traditional physical server versus virtual server [8]

As illustrated in **Figure 2**, virtualization, unlike traditional computing, can be used to run different applications hence solving the problem computing resource underutilization.

C. Cloud Computing Actors

There are four main actors in a cloud environment [9].

Cloud provider: This is the owner of the cloud service. A cloud provider has a role of managing and controlling the cloud service. The role may differ depending on the service model – IaaS, PaaS and SaaS.

***Cloud user*:** Also known as, cloud consumer, this actor uses the services offered by a cloud provider.

***Cloud broker*:** The cloud broker sits in the middle between the consumer and the provider. Their role is to help the consumer to overcome the complexity of choosing a cloud service provider. This actor may assist the consumer to combine the features of multiple cloud providers.

Cloud carrier: This actor ferries services of the cloud provider to cloud user.

D. Cloud Computing Service Models

The services provided by cloud computing can be categorized into three main layers - Software as a service (SaaS), Platform as a service (PaaS) and Infrastructure as a service (IaaS). IaaS is the lowest layer [10] and is by far the most promising model in providing cloud computing services [11]. In IaaS cloud, users provision VMs and independently run applications with mixed workloads without any control from the cloud provider. SaaS normally delivers online software services, IaaS delivers computing resources such as processor, memory, network and storage whereas PaaS delivers platform as a service where users can deploy custom software (). Each layer consumes service provided by a lower layer.

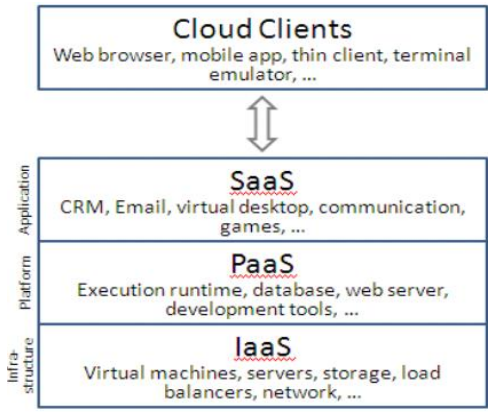

Figure 3: Cloud computing service models [12]

E. Cloud Computing Deployment Models

Cloud deployment models are private, public, community and hybrid (Sareh, 2016). In a private cloud, the compute resources are owned by one entity, normally the client. If many businesses share a business model, they may set up a cloud, which is called community cloud. When cloud infrastructure is offered to a large number of users who may have differing needs, it is called public cloud. Hybrid cloud consists of two or more cloud deployment models (**Figure 4**)**Error! Reference source not found.**.

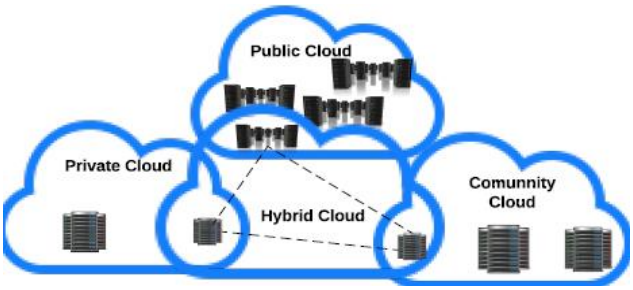

Figure 4: The four cloud deployment models: private, public, community and hybrid [3]

II. SOURCES OF ENERGY CONSUMPTION IN SERVERS AND DATACENTRES

The CPU, disk storage, memory and network are the main consumers of energy in a server [6]. The CPU consumes the largest portion of energy supplied to a server in a datacentre followed by the memory (**Figure 5**).

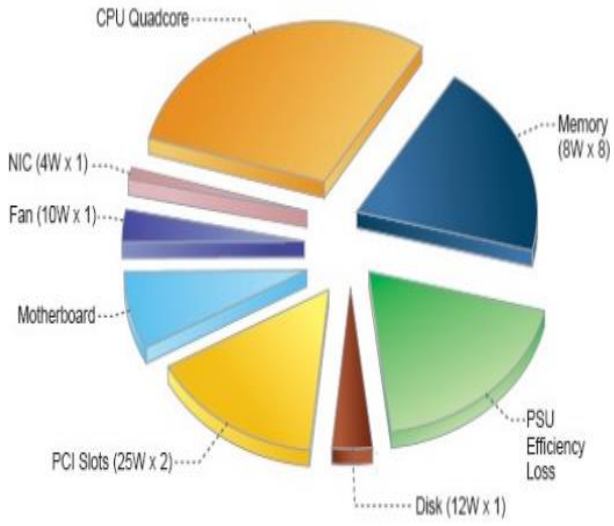

Figure 5: Server Power consumption by server component [13]

However, due to improvements in the CPU efficiency, it no longer dominates energy consumption [5]. On the other hand, energy consumed by processor greatly depends on processor types. For example, new Intel processor have power saving mechanisms [13]. Energy consumed by a datacentre can be saved up to 50% by efficiently performing VM consolidation [5]. For example, efficient VM consolidation can ensure VMs are packaged in the least number of servers so that other servers are shut down thus saving more energy. This is because an idle server consume 70% of the power when it is fully utilized [5].

Apart from IT load (CPU, disk storage, memory and network), electrical energy is also consumed by cooling and during distribution. As the datacentre servers are used, they emit heat, which need to be eliminated to avoid additional energy wastage and hardware failure [14].

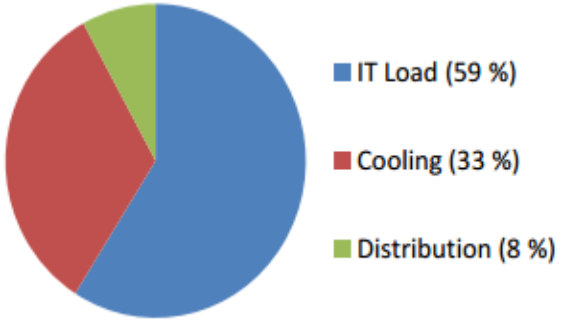

Figure 6: Energy consumption by datacentre components [15]

As shown in **Figure 6**, 33% of datacentre energy goes to cooling, which is more than 60% of that used for IT load. The amount of heat generated is a function of three factors; - frequency and voltage of the integrated circuit, technology used in manufacturing the components, efficiency of component design and most importantly, the amount of work done [13]. Removing the heat generated allows component to operate on their safe operating temperature failure of which may lead to service degradation or complete damage of the component.

III. FACTORS INFLUENCING ENERGY CONSUMPTION IN VIRTUALIZED ENVIRONMENTS

As shown previously, datacentre servers consume the most energy in a cloud computing environment. Further, it has been shown that excessive energy consumption raises environmental, system performance and monetary concerns. Therefore, it is imperative to find out the factors, which determine the amount of energy consumed by a datacentre and hence the causes of energy wastage in cloud datacentres.

A. *Level of Server Utilization*

Server utilization is the percentage of time during which a server is busy processing workload tasks and it depends on how workload patterns vary from time to time [5]. Low server utilization is a major cause of energy wastage and is causes by inefficient utilization of computing resources [9]. At high server utilization, computing resources are efficiently used and as a result, less physical server are used hence saving energy that would have been used by powering more servers. Generally, the level of server utilization determines how well energy is utilized in a server [5]. [16] reports that average server utilization for small-to-medium datacentres, with market segmentation by electricity consumption of 49%, is 10%, and 50% for High Performance Computing (HPC) datacentres, whose market segmentation by electricity consumption is 1%, and that the physical machines drew up to 90% of their peak power. Clearly, this is resource over-provisioning, which leads to increased energy consumption because many servers have to be used.

A six-month data analysed from about 5000 servers revealed that, although servers are generally not idle, their utilization never reaches 100% [5]. According to an analysis conducted by [17] on Google cluster's resource usage, 65 % of CPU and 45 % of memory goes to waste. This shows that application workloads utilizes less resources than what is provisioned- low server utilization. With high resource utilization, the number of physical servers required will be greatly reduced thus reducing the amount of energy used in datacentres.

Moreover, slim dynamic power ranges cause low server utilization because even an idle server consumed up to 70% of its peak power [3]. In this regard, it makes sense to operate at high server utilization levels. However, according to [18], there are three main challenges towards ensuring that servers are fully utilized at 100% all the time. These challenges are; diurnal patterns experienced on server workloads and load spikes, which calls for resource over-provisioning leaving servers underutilized, servers are heterogeneous and have changing configuration, thus matching diverse workloads to the servers is not trivial and at high server utilization, there is interference due to resource contention leading to performance loss. Particularly, interference has diverse effects on QoS especially on latency-critical workloads.

B. *Idle Energy Wastage*

Idle server can consume over 70% of their peak energy hence it could potentially be turned off to save energy [9]. This behaviour of severs do not represent any proportionality in increase of energy consumption with respect to system throughput. As a result, a server running at 20% can consume 80% of energy consumed by a server operating at 100% [19]. This represents a huge energy loss when servers run idle without any throughput and is usually the case for many typical servers. In this regard, one can see that this is a cause of idle energy wastage. Moreover, if an application workload does not utilize computing resources in a balanced manner, the idle components will also waste idle energy [20]. For example, if an application workload is CPU intensive, then memory idle energy goes to wastage. Therefore, it is essential that co-located VMs utilize all computing resource without leaving some being idle.

Moreover, [16] reported that as the number of servers in a datacentre continue to grow, so is the number of comatose servers. A comatose server is a server that is powered and uses electrical energy without delivering any useful service. Such servers may have been left when a certain project ended or a business process changed and since then, the servers were not removed or no one is tracking them. According to [16], an estimated 20 to 30 percent of all servers in large datacentres are idle, unused or obsolete but still consume energy. The main causes of rise of comatose servers in datacentres is lack of focus such as not budgeting time for staff to identify and remove comatose servers and aversion to risk such as IT managers fear that, by removing any previously installed servers, they may interfere with application functions that occasionally run on the servers.

C. *Adoption of Energy Efficient Solutions*

[16] reports that it is only large cloud firms that have adopted energy efficient datacentre practices. Alas, these firms account for only 5 percent of global energy consumption. The rest, 95 percent, is left to small and medium firms, which are terribly energy inefficient because of lack of adoption of energy efficient solutions and practices. Such solutions and practices include server and network consolidation, datacentre wide thermal management, purchasing and installing energy efficient hardware to replace old hardware, power planning and management (such as checking from time to time to identify and remove idle servers) and installation of energy management software [16].

Although rising energy costs is an incentive to adoption of energy efficient practices, pressure to keep up technological advancements have made many organizations to treat energy efficiency with low priority [16]. This has led to organizations not adopting even the simple and cost-effective power management software, which can monitor measure and manage both hardware level and software level energy usage. For example, energy management software offered by TSO Logic is relatively affordable and can measure datacentre power demands, active and comatose servers and energy cost, as well show how these change over time and assist in relocating application workloads and shutting down servers [21]. Nevertheless, some datacentre operators feel that by adopting automated energy usage monitoring, their employment is threatened and thus they discourage its adoption [16]. Moreover, power-saving features embedded in hardware, which can monitor hardware utilization and report to datacentre dashboards, are often disabled because of the perceived management complexity and risk associated with switching off servers. In this regard, even organization running full-scale cloud clusters do not deploy energy management solutions.

In addition, cloud providers have poor habits of procurement, which includes focusing on initial cost rather than TCO [16]. When a procurement procedure focusses only on initial purchase rather than long-term electricity costs, it may miss on energy efficient equipment in the market. For example, [22] reports that with the arrival of Intel's Sandy Bridge and Standard Performance Evaluation Corporation Power (SPECPower) benchmark, energy-proportional computing is achievable hence energy consumption by servers at idle state and low utilization can be reduced. Furthermore, [16] highlights that 80 percent of IT departments in most cloud service providers do not pay their power bills (finance department does) and so they do not see the need to make datacentre energy efficiency a priority. In addition, the IT depart do not see any incentive for implementing energy efficient practices because they are not evaluated based on the amount energy saved. In fact, IT staff have no access to power bills and most of them are more concerned with software costs. This division of accountability and split incentives are a barrier to adoption of energy efficient solutions.

D. *Server Utilization Metric*

Server utilization metric is the unit of measure of the percentage of time during which a server is busy processing workload tasks [5]. Lack of a common standardized server utilization metric has been a cause for energy wastage for many decades [16]. Increasing server utilization offers the best option for improving datacentre IT energy productivity as compared to PUE and Power Supply Efficiency (PSE) [23]. In fact, below 50 percent server utilization, a continued increase in server utilization offers the highest energy usage productivity because of the idle energy [3] [5] [18] [19].

For many years, CPU utilization has been the measure of server utilization but it is not the best since different application workloads have different CPU intensities with some of them being memory, network or I/O intensive than CPU-intensive. Besides, CPU shows the amount of work with no way of determining if that work is useful or otherwise [16]. As a result, a number of new metrics have sprung up to take care of other datacentre parameters. For example, the [24] developed a metric based on datacentre design, executing software, datacentre hardware, CPU, memory and disk as parameters. [24] also developed another metric, which attempts to measure server utilization at application level, for example tracking the number of emails sent by a server.

Other metrics include Power to Performance Effectiveness (PPE), which measures server performance per kilowatt, and SPECpower_ssj2008v1.12, which provides a means to measure power in conjunction with a performance metric. Unfortunately, there is a slow adoption of these metrics because they are complicated to implement and cannot deliver complete reports on their own without the need multiple implementations. Furthermore, different server designs have different levels of energy efficiency hence cannot work across all server designs. Therefore, average CPU utilization and average datacentre utilization (average server utilization when not in sleep mode over period), will remain in use until better metric better are developed.

E. *Energy Saving Hardware Capabilities*

A server is made up of many components such as CPU, memory, fans, power supply and disks whose power consumption efficiency manufacturers can improve by providing hardware optimization [25]. For instance, in High Performance Computing (HPC), servers frequently access storage disks thus consuming more energy. However, by spinning down the disk platters, less energy is consumed [25]. Generally, the objective is to reduce disk access

so that the disks are spun down as long as possible. In this regard, Hard Disk Drives (HDDs) are being replaced by Solid State Drives (SSDs), which becoming increasingly affordable and consume less energy. There are a number of ways in which SSDs assists in reducing power usage. SDDs utilize flash memory and do not have any moving parts, which would consume energy. On the other hand, SDDs are ideal for high-density VM environments because they provide high-speed and consistent access, which results to less time spent in storage access operations. Besides, because SSDs are ideal for high-density VM environments, it can enable a PM to hold many VMs without loss of performance [26]. Furthermore, [26] reports that SSDs require 79% less power for cooling as compared to traditional HDDs. All these benefits result to energy saving in datacentres.

In addition, manufacturers can allow individual server components to go to sleep mode independently when they are not in use [25]. For example, during computing phases, Network Interface Card (NIC) can be put to sleep state since they may not be required. Moreover, manufacturers can increase frequencies, speeds and voltages available to component making it able to adjust to current load in what is termed as proportional computing [25].

Moreover, Intel announced that the design of their new processors could be in favour of energy efficiency over speed, which technically calls to an end Moore's Law [27]. Prior to this, Intel had produced the core M series processors, which were 50 percent faster compute speed, 40 percent faster graphics performance and 20 percent longer battery life [28]. However, this family of processors have not been used in commercial servers. On the other hand, AMD developed Accelerated Processing Unit (APU), which is formally a CPU and Graphics Processing Unit (GPU) on a single chip. The aim of this design was to reduce energy consumption and it helped reduce energy consumption of between 10 and 20 percent [29].

Although the deployment of energy-efficient hardware is a crucial step, getting rid of underutilized servers is a far more effective approach, which is possible through effective consolidation [30]. Effective consolidation can be achieved through monitoring resource utilization by application workloads as well as QoS.

F. *Datacentre Thermal Management*

Thermal management is made possible by cooling by use of cooling units and fans, whose actions are controlled by the ambient temperatures in a datacentre [31]. Any increase in temperature would cause an increase in cooling energy. Therefore, the amount of heat produced by a server is an important consideration in managing energy usage in datacentres. As illustrated in **Figure 6**, 33 percent of datacentre energy is used up by the cooling unit, which is more than half of that used by IT load (59 percent).

[31] proposed a thermal-aware algorithm whose aim was to maintain uniform server tempearture by maintaining a uniform load across servers of a cluster. According to the algorithm, server tempearture should not exceed the server's treshold temperature and if it does, VM(s) is migrated to another server. In addition, the algorithm detects server underload to ensure maximum server utilization. In summary, the authors' idea is to reduce server hot spots (server with peak temperature) in a datacenter by sharing heat production, which occasionally forces the cooling units to cool the entire system, including the 'colder' servers.

In a review, [32] describes a thermal-aware resource allocation technique in which energy consumed by the cooling unit is reduced by ensuring that individual tasks are completed by their deadlines. The objective is to determine a performnce state within a server, which results to less energy consuption by the cooling while completing a workloads task by its indidual deadline. The authors reported a 9 percent reduction in power consumption on simulated experiments.

In addition, to reduce energy consumed by the cooling unit, some cloud computing providers have located clusters in cold geographical locations and udersea to benefit from free cooling such as Microsoft [33] and Google [34]. According to [33], putting a datacenter under the sea not only derives a benefit of free cooling, but also from logistics advantage because many people live close to the sea and that clean energy can be generated from sea waves to be used in the datacenter. Google has also successifully deployed datacenters close to sea such Google Hamina plant and others in Finland to benefit from cool climate [34].

G. *Computing Proportinality*

[35] cited in [3] defines proportional computing as energy efficiency technique where the energy consumption by servers is proportional to the workload. In this regard, idle servers should consume no energy. Unfortunately, energy consumption of computing units is not energy-proportional: when server load

is low, energy consumption is still high. Proportional computing is achieved by DVFS. DVFS is an energy saving technique in computer architecture that is used to save energy when server load is low [3]. In this technique, the frequency and voltage of the CPU is scaled dynamically to relate with the amount of server load. According to this approach, if the server load is at X percent of peak load, then the energy consumption should be at X percent of peak energy. Dynamic Voltage and Frequency scaling of CPU is applied for improving the energy consumption of the datacentre. The frequency of CPU is decided according to the workload by the resource controller, which is installed on each server [5].

DVFS has been used to build products available in the market such AMD Turbo Core, Intel Turbo Boost, and Intel Enhanced Speed Stepping Technology to reduce energy consumption according to workload [3]. [36] used the concept of DVFS in live VM migration. Their propossal involves monitoring CPU utilization, DVFS adjustment, and real-time migration. They report a reduction of execution time and energy consumption. Unfortunately, they note that this method has a limitation when the number of VMs in a PM approach the maximum. Moreover, DVFS is hardware-based technique and works well only on CPU bound tasks because dynamic power ranges for other components (memory, disk and network) are much narrower (< 50% for DRAM, 25% for disk drives, and 15% for network switches) [5].

[37] presents DVFS-enabled Energy-efficient Workflow Task Scheduling algorithm (DEWTS) tool, which uses DVFS and their experiments reports a 46.5% energy savings. Conversely, as DVFS is too dependent on the hardware, the resulting energy savings are low compared to other methods [9]. Although DVFS is a good solution, its savings are small because an idle server will still consume over 70% of peak energy [3].

Because of the observed failures of DVFS, powering down or switching off servers when they are not in use is a viable option and has been supported by a number of researchers [5] [9] [14].

H. *Server Power Switching*

Switching off unused or idle servers can result to significant power savings because they consume over 70% of peak energy [5] [38]. The servers would then be turned on when needed. However, deciding which server to switch off, when to do it and for how long is a complex process, which calls for careful planning. For this reason, server shutdown techniques have been developed to keep the number powered servers in line with actual workloads. Unfortunately, datacentre administrators have not embraced these techniques [16] [39]. This is because, until recently, servers were not designed to be switched off and that switching off and switching on later consumes energy and takes time [39]. Therefore, datacentre administrators have not embraced this technique for fear of interrupting with services, potential hardware failure and inability to quantify energy gains versus loss of service quality due to long booting time.

According to [39], shutdown techniques require that datacentre hardware have the ability to remotely switch off and on servers and that energy-aware algorithms should utilize this ability in a timely manner. The implementation of Advanced Configuration and Power Interface (ACPI) has five sleep states on the Linux kernel-suspend to idle, standby, suspend-to-RAM, suspend to disk and system shutdown state [39]. However, for datacentres, suspend to idle and system shutdown state are available for use. Suspend to idle allows all user spaces to be frozen and all I/O devices to be put into low energy states. System shutdown shuts down the system completely such that the system has no memory state and is not performing any task. With this option, the server consumes no energy and requires a complete boot to bring it on.

[39] describes a shutdown strategy by considering the time a server is idle and using this time to decide whether a server should be switched off. The authors defines a time T_s, which is a time threshold such that, if a server is idle for less than T_s, then it should remain idle to save energy. Additionally, T_s should be greater than the total time the server takes to switch off and on. There are two ideal shutdown policies: *knowing the future* and *aggressive shutdown* [39]. Knowing the future posits that the future dates and lengths of idle period are known for each server. On the other hand, aggressive shutdown posits that a server is shut down immediately it is idle without any prediction attempt. Although both of these policies have been found to save energy, the latter consumes more energy because idle server periods may be less than T_s.

Although shutdown techniques save energy, they cannot be used in isolation. For instance, if servers run at peak load consistently, which is rare though, energy savings from this technique will not be felt. Thus, other techniques such energy- workload consolidation need to be used.

I. *Workload Consolidation*

Workload consolidation has been studied extensively by many researchers [40] [41] [42] [43]. Dynamic consolidation reduces the number powered servers by consolidating workloads to fewer servers thus effectively improving server utilization. Additionally, when a server is idle, it is shutdown [18] [19] [3] [5]. This section describes the different approaches and techniques, which support consolidation such as VM migration and placement, VM sizing and workload characterization and how they affect energy usage in cloud computing environments.

1) VM Migration and Placement

VM placement is defined as placing a new VM in a selected PM whereas VM migration is a dynamic consolidation technique, which involves moving a VM from one PM to another [9]. To ensure energy efficiency, the problem of VM consolidation can be divided into three sub-problems; - when to migrate, which VM to migrate and where to migrate [40] [44]. According to these sub-problems, it is important to know the right time to migrate, the right VM to migrate and the right destination PM in order to be energy efficient. One of the triggers of 'when to migrate' sub-problem is the QoS [40]. In this case, overloaded or under-loaded servers are detected for purposes of VM migration. If a migration is triggered, the next step is to determine which VM to migrate. In the case of under-loaded host (PM), all VMs are moved to another PM and such a host put to idle mode or shutdown. In the case of host overload, one or more VMs need to be migrated until the host's load balances. [40] has studid three VM selection policies namely Maximum correlation, Minimum Migration Time and Random selection. [44] has summerized some the heuristics for VM consolidation using migration as shown in **Table 1**.

Table 1: Heuristics for VM consolidation using migration [44]

Goals	Server Consolidation	Load Balancing	Hotspot Mitigation
When to Migrate?	Cold spots on PMs	Load Imbalance on PMs	Hotspots on PMs
Which VM to Migrate?	VMs from Lightly Loaded PMs	VMs from overloaded PMs	Bunch of VMs from hotspot-PM
Where to Migrate?	Higher loaded PMs	Lightly loaded PMs	PM which has enough resources to house

After VM selection, determinants for 'where to migrate' sub-problem include co-located VM interference, correlation between workloads of co-located VMs and statistical muptiplexing.

Co-located VM interfearance is as a result of resource contention where demand to a particular resource by co-located VM exceeds its supply. If VM interfearance is ignored during VM migration, performance degradation and potential energy wastage is the result [45]. Also known as joint VM provisioning, statistical muptiplexing enables a VM to borrow resources from co-located VMs while it experiences peak workloads [46]. Correlation between workloads of co-located VMs is used so as to consolidate VMs with least correlation (in terms of resource demands) such that resources underutilized by one VM can be utilized by a co-located VM at peak time [47]. According to this technique, all co-located VMs cannot peak at the same time [23].

[48] presents a power-aware algorithm (PA) for determining the most suitable PM to shut down for energy savings. They combine it with remaining utilization-aware (RUA), a VM placement algorithm. Their experiments show that there is a trade-off between energy consumption due to server utilization and SLA violations. To shut down underutilized servers, under-loaded servers are detected and then all VMs migrated to other suitable servers using live migration. The aim is to ensure that all server resources remain very high by maintaining a higher server utilization level. Unfortunately, [43] asserts that power consumption and throughput increase linearly up to a certain point of resource utilization. Aggressive utilization would cause a slight service degradation but a drop in power consumption. Additionally, the degradation caused by aggressive resource utilization causes increase in execution time, which in turn encroaches into energy saving made from reduced idle energy [20]. In this regard, the challenge is to obtain an optimal performance and energy point. Thus, utilizing resources at 100 % may not necessarily be energy efficient.

VM placement has undoubtedly been studied and considerable advances made - both energy-aware and thermal-aware [14] [40]. [49] proposes a proactive thermo-aware VM placement algorithm, which takes into current and maximum temperature (threshold temperature) before making a VM placement decision. The incoming VM is thus placed in a PM, which has the highest difference between current temperature and the threshold temperature. This is to avoid chances of hardware and software failure and most importantly to reduce energy used for cooling due to

temperature rise caused by VM activation. The same concept can be used to perform thermo-aware VM migration. However, it is hard to predict temperature rise due to VM activation as well as the temperature rise owing to the number of VM already present in the PM.

A VM placement algorithm has two main goals – ensuring QoS and energy saving [50] [51]. Poor VM placement may lead to triggering new VM migrations, energy wastage and SLA violations. VM placement is purely an *optimization problem*. To solve the optimization problem, many researchers have proposed a number of solutions which include constraint programming [38] [52], bin packing (BP) [38] [9] [17], stochastic integer programming [53], genetic algorithms [50].

Additionally, many factors such as current server load (discussed earlier), temperature (discussed earlier) and cluster location are taken into account before VM placement in decided. For instance, according to [17], cloud datacenters are made of clusters located in different locations to exploit reduced electricity costs and thus clusters at regions of lower electricity costs may be selected to host incoming VM. Once a cluster is selected, the next stept is to determine the PM in cluster that will host an incomming VM. In this case, the authors see VM consolidation problen as a BP optimization problem in which PMs are viewed as bins with different capabilities and VMs as objects with different sizes (resource demands) and the objective is to pack these objects in as few bins as possible. BP problem is known to be NP-hard [17]. Thus, heuristics such as First Fit, Next Fit and Best Fit can be used to map VM to PM. [5] proposed an algorithm to PM overload for known stationary workloads. For non-stationary workloads, the author used Markov chain model. Results showed that the method achieved energy savings and as well as QoS. [6] have examined the problem of VM selection for migration. They compared Minimization of Migrations (MM) and Highest Potential Growth (HPG) VM selection policies with the Random choice (RC). MM's objective is to minimize migration overhead (energy consumed and performance loss) whereas HPG aims at reducing CPU usage to minimize SLA violation. Their results showed that MM outperforms HP and RS.

[54] proposed a VM selection techniques based on complementary workload patterns. In this proposal, workloads that do not peak at the same time can be co-located, which resulted to reducing resource contention. This was the basis of selecting a VM for migration. Further, [54] proposed an efficient technique for VM resource provisioning based on the fact that peaks and valleys in one workload pattern do not necessarily coincide with the others. This approach exploited statistical multiplexing where unutilized resources of one VMs can be borrowed by a co-located VMs. Although this method is good, it not be suitable in multi-tenant cloud environments and further, it may not be suitable if all the VM peak simultaneously.

[55] proposed a dynamic consolidation algorithm based on constraint problem solving. The authors formulated a VM allocation problem, then applied Choco constraint solver to solve the optimization problem to satisfy constraints, which were minimizing cost of migration and minimizing the number of active nodes and available computing resources (CPU and memory). Their approach mapped tasks to nodes that are better than those found by heuristics using local optimization. Keeping the number of active nodes low saves energy and taking migration overhead into an account maintains throughput.

[56] proposed a VM resizing stategy via CPU resource. According to the authors, computing resources are added to or removed from the VM depending on the current demands. The same approach could be used to meter other VM parameters. However, it is unclear how resource contention was dealth with, incase many VMs demanded a similar resource simultaneosly.

[57] proposed a dynamic consolidation technique whose objective is to reduce idle power wastage and improve performance. The authors observed that there was an energy cost and performance overhead during server start up. In this regard, however, if a server has no task to process, it is not switched off. Instead, it is put into idle state for while (time *T*) before it swirched off during which an assessment is done to find out how long it will take for the machine to be useful again. This is to ensure that energy consumed by shutting and booting the machine does not exceed the energy consumed by idle server.

2) *VM Sizing*

VM size is the measure of computing resources –CPU, Memory and I/O – assigned to a VM [3]. For instance, IaaS cloud VM can be sized to have 1 VCPU, 1 GB memory, 2000 GB network bandwidth and 25 GB of SSD. Most IaaS cloud providers require its users to determine the resource demands of their VMs. For

unexperienced users, this is be easy. However, for unexperienced users, much resources, than required, are assigned to VMs leading to server underutilization [58], which is a major cause of energy wastage in the cloud. According to an analysis conducted by [17] on Google cluster trace on resource usage, 65 percent of CPU and 45 percent of memory is unused. Thus, new techniques need to be developed to deal with VM sizing amid unpredictable worload changes in VMs.

VM sizing techniques can be categorised as static or dynamic [3]. Static techniques involves fixing VM sizes and consolidating them in fewer PMs possible or characterizing workloads, then sizing VMs according to application worload. Unfortunately, static techniques are not the best because application workloads resource demands change frequently. This situation can be handled by using dynamic VM sizing. In dynamic VM sizing, VM configurations are adjusted at runtime to meet VM applications resource demands. The ultimate objective of dynamic VM sizing is to reduce resource underprovisioning and overprovisioning.

Underprovisioning, which leads to resource underutilization, can be avoided by using a techniques known as resource overcommitment [17]. This technique involves allocating resources to the VMs than a host PM can afford. For instance, allocating 4 VMs 2GB each of RAM on a PM with 6GB of RAM. Overcommitment assumes that no VM will utilize all the resources that is allocated to it, thus more VMs can be placed in one PM, hence reducing energy consumed as fewer PMs. One downside of resource overcommitment is when the total resource request by VMs exceed what the PM can provide – overload. In overload situations, VM migration is triggered to avoid service degradation thus reducing chances of SLA violation. According to [17] it is difficult to determine the level of resource overcommitment. To address this, the same author proposes an approach called prediction. In this case, one needs to predict future aggregate resource demands for the VMs, which assists in estimating a resoanable overcommitment level.

3) VM Workload Characterization and Mixing

[59] defines workload as "*a specific amount of work computed or processed within the datacentre with defined resource consumption patterns*". A typical system workload may include tasks to be performed and the users submitting requests. Understanding system workload is the basis of understanding resource demands. When applications runs in the servers, the system is able to record application resource demand in form of logs, which may be analysed for use in resource planning. Because of security concerns, real back-end application logs are not publicly available. However, in 2009 and 2012, Google released its first and second version of production back-end trace logs respectively. The second version is more detailed as it contains machine details and resource demands [60].

Workload attributes are used for modelling and are based on resource usage of jobs and tasks or their intensity [61]. When selecting a technique for workload modelling, a number of workload attributes have to be considered. Typically, a workload arriving as user request may possess the time when such request was made as well as the amount of resources computing resources (CPU, I/O and memory).

Workloads are characterized to learn their behaviour. Characterization is based on the workload attributes. Advanced numerical and statistical techniques are required to capture workload heterogeneity or homogeneity to result to realistic models. A good example is clustering and fitting. Clustering is used to identify workloads that exhibit similar behaviour. A commonly used clustering algorithm is K-means [62]. A time series approach can also be used to determine resource usage patterns. Besides, clustering technique can identify groups of VMs with workload patterns whereas Hidden Markov predicts the changes of these patterns [61] [63]. The concept of predicting the changes of these patterns is applied in dynamic VM sizing. Extensive research on workload modelling has been done on publicly available Google cluster trace logs.

Although a cloud enter is not expected to run at its maximum – peak load - , the concept of shutting down underutilized machine does not have any energy at peak load. Thus, workload profile can be used to make a decision on the technique to use in saving energy in the cloud. For example, the CPU frequency and voltage can be adjusted depending on the workload – DVFS as discussed earlier. Further, the workload profile of co-located VMs can determine energy efficiency in a PM because of workload interference due to resource contention. Workload characterization also assists in designing multiplexing, interference-aware and correlation-aware algorithms and dynamic VM sizing for saving energy in cloud environments as discussed earlier.

[64] investigated the effect of different workloads on server power consumption in a private cloud. The authors found that placing many CPU-intensive VMs workloads in the same PM will have a detrimental effect in terms of perfomance and power consumption. According to the authors, it is wise to pair VMs, which do not consume a large amount of a similar resource. For instance, pairing a CPU intensive VM with a disk intensive VM in the same PM..

[60]have characterised workloads on a Google cluster. The frequency and pattern of job and task-level workload behavior, and how the overall cluster resources are utilized is studied. The success rate of jobs and tasks are studied i.e. successiful tasks and jobs and those that eventually fail or get killed. The authors have concluded that if the resource scheduler is offered hints about the nature and periodicity of the submitted jobs, it may specialize the resource management decisions in order to save energy and reduce performance variations of important jobs.

Using Windows Live Messenger and Windows Azure workload traces, [65] described an energy-aware server provisioning strategy that predicts near future resource demands via load patterns analysis, auto-correlations and cluster utilization. Their objective was to minimize unmet resource demands while reducing energy usage and cost of hosting clusters. Their strategy was tested on a three data center workloads and results showed that energy savings are close to optimal.

[43] investigated the effect of workload profile on power consumption. The authors used TPC-W workload generator tool and varied client behavior using *browsing* and *ordering* profiles. They observed that power consumption is greater under ordering than under browsing while both have the same throughput. Thus, it is wise to determine an optimal workload mix (ordering and browsing), which delivers energy savings.

[62] have proposed a model for energy saving in IaaS cloud via migration. They use K-means to cluster workloads, which is the basis of detecting PM overload and underload that triggers VM migration. Using this chracterization, their model is able to determine resource demands in real time so that VM scheduling is done effiently hence reducing energy consuption. In their experments, they use Google cluster trace logs.

[66] describes an architecture that charaterizes dynamic energy consumption (energy consumed by Cloud tasks) of tasks (communication, storage and computation) by analysing the characteristics of the tasks and the impact of system configurations. The architecture investigates the energy consumption patterns of tasks under different systems configurations. Once an appropriate energy-saving configuration is detected, communication is sent to an appropriate cloud environment monitoring framework for comparison with other configurations. The confirations that achives least energy is activated and adopted.

[67] proposed a new architecture, which allocates groups of tasks to customized VMs based on task characteristics in container-based clouds. This mapping is based on actual task resource usage patterns obtained from an analysis of real usage trace logs instead of the resources requested by cloud users. The authors used second version of Google cluster trace logs and X-means for clustering algorithm (a variant of K-Means clustering algorithm). [58] and [68] proposed a workload mix aware resource provisioning technique, which can predict non-stationary workloads. The technique predicts future server capacity and this was important in avoiding resource over-provisioning and under-provisioning. Although the primary objective of the authors was not energy efficiency, efficient recource consolidation can achieve energy saving, for example where avoiding over-provisioning can reduce the number of active PMs.

[17] observed that one of the reasons of energy wastage in a datacenter is having idle servers, which are consuming electrical energy but not delivering any useful service. Therefore, it makes sense to switch off idle servers. However, switching off a server just because it is idle will not necessarily save energy because the amount of energy consumed by switching off and on, may exceed energy consumed by letting the server remain in active state. This is partly because the cloud manager is unware of the length of time the server will remain idle before it needed again to provide service. Therefore, it makes sense that if the server is not needed for a long time, switching it off will save energy. In this regard, the authors proposed a workload predictor to assist in estimating future workloads, which will in turn determine how long the server is expected to remain idle. However, predicting workloads is not trivial because of the frequency of client requests and different combinations of resource requests. For this reasons, the authors categorised requests based on their charactericts (frequence and resource demands) via clustering and each category has a unique predictor. The

resulting clusters are a basis for resource estimation.

IV. CONCLUSION

This paper provides an highlight on the sources of energy consumption and the factors that influence energy consumption in cloud computing environments. As future work, a conceptual model will be developed, which will aid in developing energy efficient solution for the cloud environments based on the factors identified.

REFERENCES

[1] A. Rallo, "Industry Outlook: Data Center Energy Efficiency," 2014. [Online]. Available: http://www.datacenterjournal.com/industry-outlook-data-center-energy-efficiency/. [Accessed 4 August 2015].

[2] G. Albert, H. James, A. M. David and P. Parveen, "The cost of a cloud: research problems in data center networks," *The ACM Digital Library is published by the Association for Computing Machinery*, vol. 39, no. 1, 2009.

[3] F. P. Sareh, "Energy-Efficient Management of Resources in Enterprise and Container-based Clouds," The University of Melbourne , 2016.

[4] G. K. V. Rao and K. Premchand, "Scheduling Virtual Machines across Data Centres in accordance to availability of Renewable Sources of Energy," *International Journal Of Engineering And Computer Science*, vol. 5, no. 10, 2016.

[5] B. Anton, "Energy-Efficient Management of Virtual Machines Data Centers for Cloud Computing," THE UNIVERSITY OF MELBOURNE, 2013.

[6] B. Anton and B. Rajkumar, "Energy Efficient Resource Management in Virtualized Cloud Data Centers," in *10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, 2010.

[7] NIST, "The NIST Definition of Cloud Computing," U.S. Department of Commerce, 2011.

[8] R. M. Sharma, "The Impact of Virtulization in Cloud Computing," *International Journal of Recent Development in Engineering and Technology*, vol. 3, no. 1, 2014.

[9] G. Chaima, "Energy efficient resource allocation in cloud computing Enviroment.," Institut National des T´el´ecommunications, Paris, 2014.

[10] R. Neha and J. Rishabh, "CLOUD COMPUTING: ARCHITECTURE AND CONCEPT OF VIRTUALIZATION," *International Journal of Science, Technology & Management*, vol. 4, no. 1, 2015.

[11] B. Esha, Y. J. S. and I. Biju, "Energy Efficient Virtual Machine Placement using Enhanced Firefly Algorithm," *Multiagent and Grid Systems - An International journal*, pp. 167-198, 2016.

[12] E. Gorelik, "Cloud Computing Models," MIT, 2013.

[13] Intel, "The problem of power consumption in servers.," Intel , 2009.

[14] Z. Jiaqi, M. Yousri, T. Jie, J. Foued, L. Qinghuai and S. Achim, "Using a vision cognitive algorithm to schedule virtual machines," *International Journal of Applied Mathematics and Computer Science*, vol. 24, no. 3, 2014.

[15] G. Akhil and C. Navdeep, "A Proposed Approach for Efficient Energy Utilization in Cloud Data Center," *International Journal of Computer Applications (0975 – 8887)*, vol. 115, no. 11, 2015.

[16] Natural Resources Defense Council (NRDC), "Data Center Efficiency Assessment," NRDC, 2014.

[17] M. Dabbagh, B. Hamdaoui, M. Guizani and A. Rayes, "Toward energy-efficient cloud computing: Prediction, consolidation, and overcommitment," *IEEE Network*, vol. 29, no. 2, 2015.

[18] C. Delimitrou, "Improving Resource Efficiency In Cloud Computing.," Stanford University, 2015.

[19] T. Mastelic, A. Oleksiak, H. Claussen, I. Brandic, J.-M. Pierson and A. V. Vasilakos, "Cloud computing: survey on energy efficiency," *ACM Computing Surveys*, vol. 47, no. 2, 2015.

[20] A. Mirabel and R. Siddiqui, "Energy Aware Consolidation in Cloud Computing," 2015 .

[21] TSO Logic, "TSO Logic," 2017. [Online]. Available: http://tsologic.com/. [Accessed 4 January 2017].

[22] B. Subramaniam and W.-c. Feng, "Towards Energy-Proportional Computing Using Subsystem-Level Power Management," in *Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering*, Prague, Czech Republic, 2013.

[23] V. Sanjeev, "INNOVATIONS IN TECHNOLOGY: CLOUD COMPUTING AND ENERGY EFFICIENCY," *International Journal of Engineering and Management Sciences*, vol. 6, no. 2, 2015.

[24] Green Grid, "THE GREEN GRID DATA CENTER COMPUTE EFFICIENCY METRIC: DCcE," 2010.

[25] O. Anne-C´ecile, M. D. d. Assuncao, L. Lef`evre and R. Tut´oi, "A Survey on Techniques for Improving the Energy efficiency of large scale distributed systems," ACM Computing Surveys, 2013.

[26] Intel, "Data Center Strategy Leading Intel's Business Transformation.," Intel, 2016.

[27] M. Anderson, "Intel says chips to become slower but more energy efficient," 2016. [Online]. Available: https://thestack.com/iot/2016/02/05/intel-william-holt-moores-law-slower-energy-efficient-chips/. [Accessed 26 January 2017].

[28] T. Team, "Intel Launches Its Most Energy-Efficient Processor To Spur PC Demand," 2014. [Online]. Available: http://www.forbes.com/sites/greatspeculations/2014/09/10/intel-launches-its-most-energy-efficient-processor-to-spur-pc-demand/#1c55373c7d07. [Accessed 26 January 2017].

[29] AMD, "AMD A-Series Desktop APUs," 2017. [Online]. Available: http://www.amd.com/en-us/products/processors/desktop/a-series-apu. [Accessed 26 January 2017].

[30] F. Elijorde and J. Lee, "Attaining Reliability and Energy Efficiency in Cloud Data Centers Through Workload Profiling and SLA-Aware VM Assignment," *International Journal Advance Soft Computer Applications* , vol. 7, no. 1, 2015.

[31] V. Villebonnet and D. C. Georges, "Thermal-aware cloud middleware to reduce cooling needs," in *IEEE International Conference on Collaboration Technologies and Infrastructures - WETICE*, Parma, Italy. , 2014 .

[32] M. Neeraj, S. Manpreet and K. R. Sanjeev, "RESOURCE SCHEDULING IN CLOUD ENVIRONMET: A SURVEY," *Advances in Science and Technology*, vol. 10, no. 30, 2016.

[33] Microsoft, "Project Natick," 2016. [Online]. Available: http://natick.research.microsoft.com/. [Accessed 25 October 2016].

[34] Oxford Research, "Finland's Giant Data Center Opportunity," Oxford Research, 2015.

[35] B. Luiz and H. Urs, "The Case for Energy-Proportional Computing," IEEE Computer Society , 2007.

[36] V. J. Patel and H. A. Bheda, "Reducing Energy Consumption with Dvfs for Real-Time Services in Cloud Computing," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 16, no. 3, 2014.

[37] Z. Tang, L. Qi, Z. Cheng, K. Li, S. U. Khan and K. Li, "An Energy-Efficient Task Scheduling Algorithm in DVFS-enabled Cloud Environment," *Journal of Grid Computing* , vol. 14, no. 1, 2016.

[38] H. Zhang, L. Ye, X. Du and M. Guizani, "Protecting private cloud located within public cloud," in *Global Communications Conference (GLOBECOM)*, Atlanta, GA , 2013.

[39] I. Rais, A.-C. Orgerie and M. Quinson, "Impact of Shutdown Techniques for Energy-Efficient Cloud Data Centers," in *16th International Conference on Algorithms and* , Granada, Spain , 2017.

[40] B. Anton, A. Jemal and B. Rajkumar, "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing," *ELSEVIER*, vol. 28, 2011.

[41] K. S. Ashwin, R. Rahul, Dheepan and K. Sendhil, "An Optimal Ant Colony Algorithm for Efficient VM Placement," *Indian Journal of Science and Technology*, vol. 8, no. S2, p. 156–159, 2015.

[42] S. Mohsen, S. Hadi and N. Mahsa, "Power-efficient distributed scheduling of virtual machines using workload-aware consolidation techniques," *The Journal of Supercomputing* , 2011.

[43] C.-Z. Mar, S. Lavinia, A.-C. Orgerie and P. Guillaume, "An experiment-driven energy consumption model for virtual machine management systems," 2016.

[44] A. Choudhary, S. Rana and K. J. Matahai, "A Critical Analysis of Energy Efficient Virtual Machine Placement Techniques and its Optimization in a Cloud Computing Environment," in *1st International Conference on Information Security & Privacy* , 2016.

[45] X. Fei, L. Fangming, L. Linghui, J. Hai, B. Li and L. Baochun, "iAware: Making Live Migration of Virtual Machines Interference-Aware in the Cloud," *IEEE TRANSACTIONS ON COMPUTERS*, pp. 3012 - 3025, 2014.

[46] F. P. Sareh, "Energy-Efficient Management of Resources in Enterprise and Container-based Clouds," The University of Melbourne , 2016.

[47] K. Jungsoo, Martino, D. A. Ruggiero and L. Marcel, "Correlation-Aware Virtual Machine Allocation for Energy-Efficient Datacenters," in *13 Proceedings of the Conference on Design, Automation and Test in Europe*, 2013.

[48] G. Han, W. Que, G. Jia and L. Shu, "An Efficient Virtual Machine Consolidation Scheme for Multimedia Cloud Computing," *Sensors journal* , vol. 16, no. 2.

[49] K. Supriya, K. Rajesh and S. Anju, "Prediction Based Proactive Thermal Virtual Machine Scheduling in Green Clouds," *The Scientific World Journal*, pp. 1-13, 2014.

[50] U. Zoha and S. Shailendra, "A Survey of Virtual Machine Placement Techniques in a Cloud Data Center," *International Conference on Information Security & Privacy*, p. 491 – 498 , 2015.

[51] J. R. Chirag, "A Survey on Different Virtual Machine Placement Algorithms," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 2, no. 2, 2014.

[52] D. Jiankang, W. Hongbo and C. Shiduan, "Energy-performance tradeoffs in IaaS cloud with virtual machine scheduling," China communications , 2015.

[53] B. Speitkamp and M. Bichler, "A Mathematical Programming Approach for Server Consolidation Problems in Virtualized Data Centers," *IEEE Transactions on Services Computing* , vol. 3, no. 4, pp. 266 - 278, 2010.

[54] X. Meng, C. Isci, J. Kephart, L. Zhang, E. Bouillet and D. Pendarakis, "Efficient Resource Provisioning in Compute Clouds via VM Multiplexing," in *Proceedings of the 7th International Conference on Autonomic Computing*, VA, USA , 2010c.

[55] F. Hermenier, X. Lorca, J.-M. Menaud, G. Muller and J. L. Lawall, "Entropy: a Consolidation Manager for Clusters," in *5th International Conference on Virtual Execution Environments*, 2009.

[56] S. Saravanankumar, M. Ellappan and N. Mehanathen, "CPU Resizing Vertical Scaling on Cloud," *International Journal of Future Computer and Communication*, vol. 4, no. 1, 2015.

[57] A. Gokul and S. Priya, "Energy Optimization in Cloud Computing by EGC Algorithm," in *International Conference on Innovations in Engineering and Technology* , 2016 .

[58] P. Jemishkumar, I.-L. Y. Vasu, B. Farokh, Jindal, X. Jie and G. Peter, "Workload Estimation for Improving Resource Management Decisions in the Cloud.," in *2015 IEEE TwelfthInternational Symbosium on Autonomous Decentralized Systems* , 2015.

[59] S. M. Ismael, Y. Renyu, X. Jie and W. Tianyu, "Improved Energy-Efficiency in Cloud Datacenters with Interference-Aware Virtual Machine Placement," in *Autonomous Decentralized Systems (ISADS), 2013 IEEE Eleventh International Symposium*, 2013.

[60] L. Zitao and C. Sangyeun, "Characterizing Machines and Workloads on a Google Cluster," in *41st International Conference on Parallel Processing Workshops*, 2012.

[61] C. C. Maria, L. D. V. Marco, M. Luisa, P. Dana, I. M. Momin and D. T. Tabash, "Workloads in the Clouds," 2016.

[62] Q. Xia, Y. Lan and L. Zhao, "Energy-saving analysis of Cloud workload based on K-means clustering," in *Computing, Communications and IT Applications Conference (ComComAp)* , 2014.

[63] A. Khan, X. Yan, S. Tao and i. Anerousis, "Workload characterization and prediction in the cloud: A multiple time series approach," in *Network Operations and Management Symposium (NOMS), 2012 IEEE*, 2012.

[64] J. Smith and I. Sommerville, "Workload Classification & Software Energy Measurement for Efficient Scheduling on Private Cloud Platforms," in *Conference'10 University of St Andrews*, 2011.

[65] B. Guenter, N. Jain and C. Williams, "Managing Cost, Performance, and Reliability Tradeoffs for Energy-Aware Server Provisioning," Microsoft , 2011.

[66] A. Uchechukwu, k. Li and Y. Shen, "Improving Cloud Computing Energy Efficiency," in *Cloud Computing Congress (APCloudCC)*, 2012.

[67] F. P. Sareh, R. N. Calheiros, J. Chan, A. V. Dastjerdi and R. Buyya, "Virtual Machine Customization and Task Mapping Architecture for Efficient Allocation of Cloud Data Center Resources," *The Computer Journal*, 2015.

[68] R. Singh, U. Sharma, E. Cecchet and P. Shenoy, "Autonomic Mix-Aware Provisioning for Non-Stationary Data Center Workloads," in *Proceedings of the 7th international conference on Autonomic computing*, 2010.

[69] R. Hintemann and J. Clausen, "Green cloud? the current and future development of energy consumption by data centers, networks and end user devices," 2016.