

ISSN 2047-3338

Efficiency in Cloud Data Centers Via a Hybrid Methodology of VM Consolidation, Clustering and Power-State Switching

Anam Ghaffar¹, Sana Yasin², Zunaira Altaf³, Faiza Shafaqat⁴, Fuhmida Suduf⁵

^{1,2}Riphah International University, Faisalabad, Pakistan

³Government College University, Faisalabad, Pakistan

^{5,6}University of Agriculture, Faisalabad, Pakistan

¹anamghaffar7@gmail.com, ²sana.yasin02@gmail.com, ³zainezunaira@hotmail.com, ⁴faiza.shafaqat@riphahfsd.edu.pk,

⁵fehmda.sadaf@riphahfsd.edu.pk

Abstract—Power consumption in data centers is a significant concern due to the ever-increasing demand for computing resources and the impact of high energy consumption. For improvement and betterment in datacenters many algorithms are introduced. Researchers work for better and efficient data centers, but there is still space for the improvement of power consumption. Due to handling large amounts of data, data centers shift on virtual machines and big data concepts. This research focuses on power-saving techniques in data centers, specifically exploring the benefits of virtualization, sleep mode, and server clustering without compromising on its performance and quality of service (QoS). This is a hybrid technique in which virtual machines can share the load of physical servers due to which costs of cooling system for physical servers and their power become less than Server clustering that involves grouping multiple servers together to work collaboratively, allowing for workload distribution and improved resource utilization and then we apply sleep mode on clustered servers that shifts the load of a server to its neighbor's server. The results demonstrate that the combination of virtualization, sleep mode, and server clustering can lead to significant power savings in data centers. We observe that the proposed technique shows significant advantages over the existing techniques so far.

Index Terms—Virtual Machines (VMs), Cloud Datacenters, Quality of Service (QoS), Clustering and Sleep Mode

I. INTRODUCTION

POWER saving in datacenters is concerned with the reduction of energy usage and increasing the efficiency of the datacenter by utilizing methods like server virtualization, consolidation, power management, and clustering. These methods contribute to lower running costs, lower environmental impact and improved overall energy efficiency [1]. Datacenters are centralized facilities equipped with servers, storage devices, networking equipment and power/cooling systems to process, store and deliver applications [2]. Cloud datacenters are more power efficient through virtualization, dynamic resource allocation and

energy-efficient infrastructure. These technologies utilize resources optimally and adapt resources to the demand and lower power usage and operating costs [5].

Datacenter electricity consumption is the amount of electricity consumed by the servers, networking equipment, cooling, and other supporting equipment. Energy losses in systems make the total power consumption larger than the power consumption by the components and efficient energy management is necessary [4].

The development of digital technologies has boosted energy requirements for various industries, including communications, education, health, and business. Both the growth of the Information and Communications Technology (ICT) infrastructure and the growing volume of data generated have played a significant role in electricity consumption. Datacenters use a significant amount of energy to meet the growing demands of ICT [4]. As energy demand rises in Pakistan, the need to optimize the efficiency of data centers becomes more critical, as doing so could lead to cost savings and greater energy sustainability. As shown in Fig. 1, the electricity consumption trend in Pakistanian datacenters has been observed from 2011 to 2022.

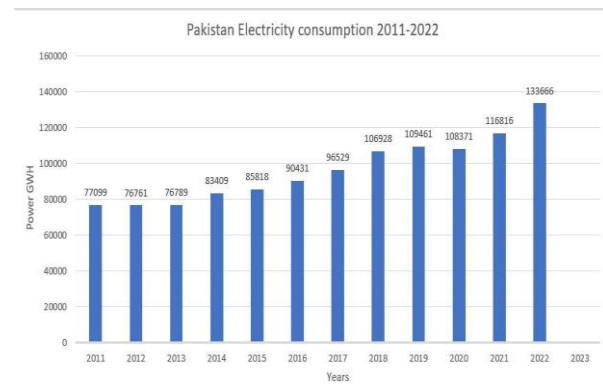


Fig. 1. Pakistan's electricity consumption of datacenters from 2011-2022

A total of 1,303 GWh of energy was used in 2020, marking a steady year-on-year increase from 1,255 GWh in 2019, as a result of the rapid expansion of data volumes and the demand for datacenter infrastructure. A datacenter is a collection of interconnected computers, servers, backups, and storage devices that are used to manage large amounts of data for government and private entities [4].

Many energy-saving methods have been proposed such as virtualization and energy-efficient cooling systems, but there are still many ways to improve energy efficiency without compromising the Quality of Service (QoS). Cloud Datacenter optimization of power consumption requires advanced algorithms and strategies [6]. In this research, we propose hybrid approach to overcome this issue which is a combination of physical server virtualization, clustering of virtualized servers and sleep mode mechanisms to achieve power efficiency.

Server virtualization is used to run multiple virtual machines on a single physical server, leading to better resource utilization, lower hardware needs, and greater scalability and resource management in datacenters [5]. Virtualization technologies like VMware, and hypervisors enable the efficient use and control of computing resources in the form of a virtual environment. Fig. 2 shows the virtualization concept, which involves the use of virtualization software like VMware and hypervisors to create and manage virtual machine environments.

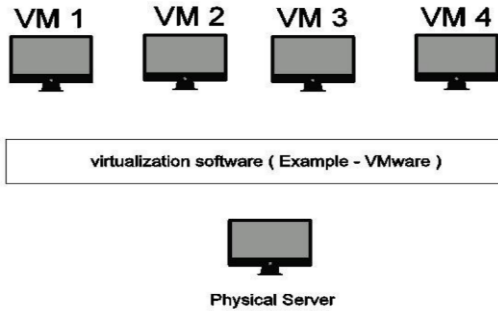


Fig. 2. Virtualization of Servers

Through the efficient distribution of workloads across multiple servers, server clustering enhances resource utilization, scalability, fault tolerance, and load balancing [10]. Additionally, sleep mode can be used to reduce energy consumption in instances when one or more servers are idle or underutilized, without impacting performance or service availability [8].

Although using power-saving approaches like virtualization, server clustering, and sleep mode mechanisms are available, keeping power down without sacrificing energy efficiency and service quality is still a major obstacle in the datacenter. Thus, a hybrid model that combines these techniques is suggested in the present work to further reduce energy consumption.

This study focuses on the following main research question: “How can power consumption in datacenters be reduced without loss of quality of services?”

II. LITERATURE REVIEW

A. Background

Energy efficiency has become a serious challenge in large datacenter because of rapid growth of big data and huge number of connected wireless devices [1]. The research on power-saving methods used in personal computing has been greatly extended and widely applied in large datacenter environments, such as sleep mode, virtualization, and renewable energy, to reduce power usage while preserving efficient system performance [2].

A datacenter is a single unit or facility with a large number of computing servers, routers, switches, etc. used to process and store significant amounts of data. These infrastructure elements consume a large amount of energy in data-centers. The power usage of the different elements within a data center is depicted in Fig. 3.

The power consumption of a datacenter is clearly divided between servers (around 40%), cooling systems, storage and communication devices. As the number of connected smart devices continues to grow, it is important to implement scalable power-saving strategies in datacenters [11].

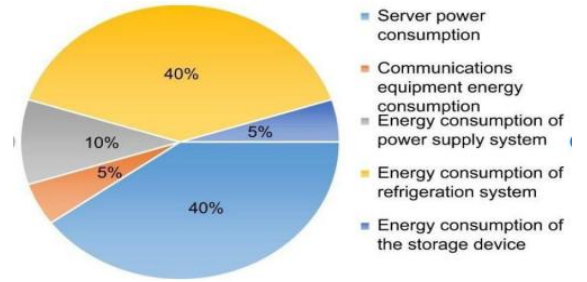


Fig. 3. Energy Efficiency of Components in a Datacenter

Internet traffic has risen by over 40% in 2020, social networking, connected devices and big data are creating a high demand for datacenter services. Although researchers have proposed solutions such as renewable energy integration and other energy-efficient techniques, the growing volume of datacenter traffic continues to drive higher energy consumption. Fig. 4 presents the energy demand of different datacenter types, including traditional, cloud, and hyperscale datacenters.

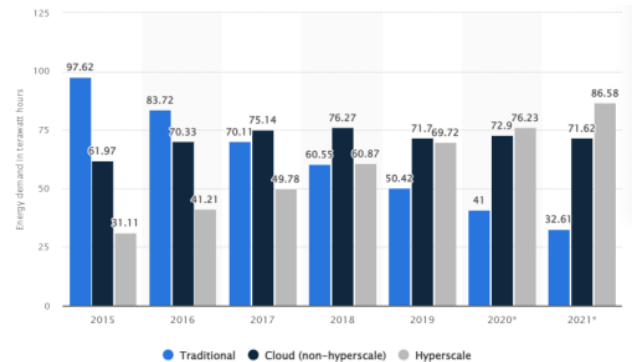


Fig. 4. Energy demand in traditional, cloud and hyper scale datacenters

Traditional datacenters are managed on-premises by an organization's internal IT team [17], whereas cloud datacenters provide remote access to computing resources over the internet. Hyperscale datacenters are large-scale facilities designed to support highly scalable applications and massive data processing. From 2015 to 2021, the use of traditional datacenters declined while hyperscale datacenters became increasingly dominant.

The main objective of this research is to decrease power usage in datacenter while maintaining the QoS. QoS provides efficient and reliable application performance by using effective resource allocation and traffic management to maintain metrics like response time, throughput, availability, reliability and security [7]. For this purpose it is proposed to use virtualization, server clustering and sleep mode methods to enhance energy efficiency.

Virtualization is the technology that allows several virtual machines to run on a single physical server, which helps to utilize resources better, decrease hardware expenses, and ensure scalability. Virtualization was originally developed for mainframe systems in the 1960s, started to be widely used with x86 virtualization in the early 2000s, and is now a fundamental component of cloud computing for the efficiency in VM deployment and management [14], [15].

Server clustering optimizes the use of resources, increases scalability, provides fault tolerance, distributes the workload across multiple servers, and simplifies management in the datacenter. Sleep mode is complementary in this regard: When servers are idle or underutilized, they enter into a low-power state to conserve energy without sacrificing system performance or the availability of services [8].

B. Literature Review

The power consumption issue of the datacenter has been widely discussed in this section for various methods of energy efficiency, quality of service and resource utilization, such as virtualization, server consolidation, clustering, sleep mode and power management. Previous research has highlighted present issues, gaps, and trends, and offered a basis for creating more effective power-saving methods for sustainable datacenter operations. While a number of works have investigated the use of 4G and 5G systems, techniques like smart antennas, scheduling algorithms, and renewable energy integration have been shown to lower energy consumption in mobile networks [23]–[26], this study concentrates on power reduction in datacenters.

The literature surveyed is grouped into four broad areas: (i) virtualization techniques, (ii) server clustering, (iii) performance after virtualization, and (iv) sleep modes. Energy-efficient datacenters wireless networking has been explored in order to enhance communication protocols, network architectures, and resource allocation approaches to lower power usage [1]. A virtual machine (VM) scheduling method aimed at energy saving has been introduced to achieve optimal energy efficiency and Quality of Service (QoS) [2]. Several methods for energy-efficient resource allocation and VM migration have been proposed to improve resource

utilization and minimize power consumption in the cloud environment have also been investigated [3].

One of the major challenges in cloud datacenter is dynamically allocating workloads onto physical servers so as to achieve optimal usage of resources and minimize energy consumption in cloud datacenter [5] suggested some virtual machine consolidation techniques. There have also been processor power models analyzed to explore the relationship between the performance of the processor and its power consumption, which has proved of value to the design of energy-efficient computing systems [6]. The energy-efficient virtual machine migration strategies have been tested with the GreenCloud and Energy Checker simulators, which were used to analyze workload distribution and power consumption in a virtualized datacenters and to show that round-robin scheduling was more energy-efficient than green algorithms, and that more efficient power-aware loadbalancing techniques are needed [7]. The effectiveness of sleep mode in virtualized cloud datacenters has been studied; intelligent management of server sleep states can lead to substantial energy savings during periods of low server utilization, without sacrificing system performance [8]. In addition, an energy efficient task scheduling algorithm is proposed to dynamically schedule tasks among the active servers, and to sleep idle servers, to achieve better energy efficiency and performance than current scheduling methods [9].

Also, an energy efficient cloud datacenter architecture with the emphasis on workload consolidation, dynamic resource allocation, efficient cooling and integration of renewable energy sources has been introduced, which significantly lowers energy consumption without compromising system reliability and performance [10]. In the study discussed in [11], energy consumption reduction in cloud datacenters is the main concern, and dynamic virtual machine consolidation and migration control techniques are used. Experimental results show that the proposed intelligent VM migration can enhance resource utilization, minimize network traffic and allow for a further energy saving by powering down idle host servers.

For compute-intensive server farms, distributed auto-scaling policies have been proposed to switch between the number of active servers in response to workload demands, so as to save power significantly with acceptable system stability and response time [12].

A feedforward neural network was also used to predict the Datacenter power consumption based on cooling system, servers, and network equipment, to increase power estimation accuracy and for energy optimization [13]. While research is conducted on mobile network base stations, not datacenters, it shows that implementing the sleep mode can save about 18.7% in the daily energy usage, proving how effective the energy-saving mechanism is based on sleep.

There have been many reviews of server power consumption models for energy and thermal management in datacenters, which include additive, polynomial, non-linear, linear regression and power function models. These models include linear regression and polynomial models which have the highest accuracy of prediction [15]. Several mathematical

power consumption models for storage servers have been created based on power consumption of various hardware

components like CPU, memory, disk drives, etc., which can be used to manage power consumption in datacenters more effectively [16]. Advances in the field of power efficient datacenter technologies in the big data era have been explored, and wireless networking solutions, big data strategies, and improving efficiency without sacrificing operational costs and service quality have been discussed [17].

In addition, an energy prediction model based on neural network was proposed and optimized by an artificial bee colony algorithm to accurately predict the energy consumption of the datacenter with the parameters including server workload, cooling efficiency, and temperature to achieve proactive energy management and optimization [19].

An energy-efficient virtual machine allocation technique using an interior search algorithm has also been presented to dynamically allocate virtual machines to physical machines for energy saving in cloud datacenters with the objective of better utilization of resources [20].

At the same time, a number of power budgeting methods and their advantages and disadvantages in the context of cost effect, reliable, and energy efficient datacenter operations were discussed, such as power capping, dynamic voltage and frequency scaling (DVFS), workload consolidation, and virtual machine migration (VM migration) [21].

III. METHODOLOGY

This research explores the various energy efficient methods used in cloud datacenters and suggests a more efficient way to save power and still provide a high Quality of Service (QoS). This introduces the research methodology and proposed algorithms to solve the research problems.

The proposed methodology starts with an extensive literature review to identify current energy efficient techniques used in datacenters, identify the gaps in research, and develop the research questions.

Using these results, a datacenter system model and problem formulation are written, first by examining the power consumption of each component in the datacenter and second by creating a mathematical model to describe energy consumption. Finally, a power-efficient algorithm is proposed to minimize runtime power consumption while maintaining a high QoS. The power consumption model is analyzed from both hardware and software perspectives. Hardware modeling considers datacenter infrastructure, including cooling systems and IT components such as processors, memory, hard disks, fans, and network interfaces. Software modeling focuses on virtualization, operating systems, communication-intensive applications, and artificial intelligence techniques. Server power models are used to estimate energy consumption in datacenters [27].

Table I presents the power consumption of different server components [26].

TABLE I
POWER CONSUMPTION OF SERVER COMPONENTS

Component	Power (W)
CPU	40
Memory	09
Disk	12
Motherboard	25
Fan	10

The CPU is the primary power-consuming component in a server, accounting for approximately 40% of the total energy consumption under full load. However, since CPUs rarely operate at 100% utilization, a significant amount of energy is wasted, as CPU power consumption is not directly proportional to its utilization level.

A. Power Model of Datacenter

To calculate the total power consumed by a data center (DC), we first considered the formulation. However, formulation based on just the server output and then classified as the energy consumption of the data center, which based on DC components. The formula is as follows.

$$\begin{aligned}
 i &= 0 \quad \text{if server } j \text{ is enable, } w_i = 0 \\
 p^{server} &= \{ \\
 i &= \text{idle } w_i \text{ VM dynamic} \\
 p_i + \sum_{j=1} \mu^{ij} * p_i, w_i > 0
 \end{aligned} \tag{1}$$

The server power P^{server} consists of idle power and active power. Idle power remains constant, while active power varies linearly with the total CPU use of every virtual machine (VM) that the server is hosting. To save energy, the server is turned down if no virtual machines are allocated. Here, VM_i, j represents the j th virtual machine on the i th server, w_i denotes the total number of VMs assigned to server i , and $i, \mu (VM_i, j)$ represents the utilization of the j th VM based on the combined workload of its applications. In addition, c denotes the component ID, while A_j represents the number of application components.

B. Load on Physical Servers and Virtualization

The proposed approach minimizes datacenter power consumption through three main steps: (i) selecting overloaded physical hosts, (ii) migrating virtual machines (VMs), and (iii) placing VMs on appropriate physical hosts to optimize resource utilization.

For this, a Load-Aware VM Placement architecture is proposed. The VM Manager makes efficient placement decisions by monitoring status and resource utilization of available hosts continuously. This not only helps with better resource utilization but also boosts system performance and lowers energy usage. Fig. 5 shows the proposed Load-Aware VM placement architecture.

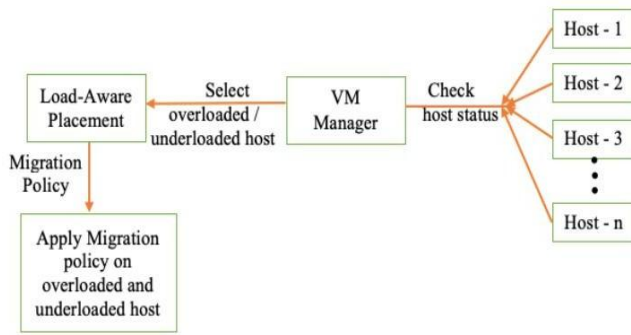


Fig. 5. System architecture proposed with load-aware VM placement system

This approach defines the hosts as overloaded or underloaded according to pre-determined threshold values. The Load-Aware VM Placement algorithm is used to deploy the VMs and to migrate them when there is an overload or an underload situation. Three primary procedures that make up the technique are task submission and virtual machine placement, load-aware virtual machine placement, and overloaded/underloaded host selection. The algorithm takes into account the usage of the CPU, memory and bandwidth to optimize system performance and minimize energy usage. These parameters are used in combination to calculate the Total Utilization Ratio (TUR) and then the total power consumption is estimated from this value.

Total utilization can be determined by the following equation:

$$U_{Total} = \frac{U_{RAM} + U_{CPU} + U_{BW}}{3} \quad (2)$$

Where,

U_{RAM} = Utilization RAM

U_{CPU} = Utilization CPU

U_{BW} = Utilization Band width

The power state dominates the power consumption of the CPU, which is the main consumer of dynamic power. Therefore, CPU utilization is used to predict the total host power consumption, whether the host is in an active or sleeping state. The power model is one that depends on the use of power and is defined as follows:

$$P_c = P_{min} + (P_{max} - P_{min}) * Total Resource Utilization$$

The power consumption model takes into account the use of RAM, CPU and bandwidth to estimate overall power consumption. It also includes the maximum power that is used when it is at its peak power P_{max} and the minimum power that is used when it is at its idle power P_{min} , to accurately calculate overall power usage.

C. Clustering & Sleep Mode Technique

Clustering is the process of connecting together multiple servers to form a single cluster in a datacenter to enhance performance, scalability and availability. It improves energy efficiency by reducing the number of servers needed to run workloads, utilising servers more, balancing workloads across

the cluster and allowing idle servers to go into sleep mode to conserve power consumption. In the proposed approach, a cluster consisting of 10 servers running on the same operating system is established and resources are grouped together into CPU pools and storage pools. CPUs are allocated in pools according to workloads and storage is allocated in pools for efficient resource management.

D. Algorithm of Sleep Mode in Datacenter

The cost of energy is still on the rise in datacenters, and many servers still use a lot of energy even when they are performing relatively little work [2]. The proposed approach is to introduce a sleep mode mechanism for idle servers, which are put into a low-power state when they are idle, and are woken up when new requests are received. The ON and SLEEP states are alternated between the servers to minimize the use of energy for unnecessary operations without compromising availability. The Cloud Management System (CMS) manages this process by scheduling tasks on running virtual machines (VMs), monitoring idle VMs, migrating tasks from selected servers to running VMs, and putting servers without any running VMs to sleep mode [30]. The sleep timer is used to set sleep time, and servers will be in sleep mode until new tasks are available to process. This improves energy efficiency without impacting performance in the datacenter.

E. Clustering Similar Requests

One of the effective energy saving techniques that are implemented in cloud datacenters is the clustering of similar user requests that eliminates redundant processing and maximizes utilization of resources [10]. Before processing the user requests, they are analyzed and clustered according to similarity. A query that is identical to the present query is found, the result stored on the query results server is retrieved rather than processing the query again. If not, the query is passed on to be processed, and the query and its answer are cached for later processing. It reduces the number of unnecessary computations, increases response efficiency and reduces energy consumption in cloud data centers.

IV. CASE STUDY OF POWER SAVING IN DATACENTERS

A. Datacenter Case Study

This case study examines the power-saving techniques implemented in the datacenter of an Internet company in Pakistan. The datacenter includes power and HVAC systems, networking equipment, and various servers, such as web, mail, database, and backup servers, which support web services, email communication, data management, billing operations, and data backup. List of servers in company datacenter shown in Table II.

TABLE II
LIST OF SERVERS IN COMPANY'S DATACENTER

Model Name	No of Servers
HP ML30 Gen 10	5
HP Pro liant Gen 10	5

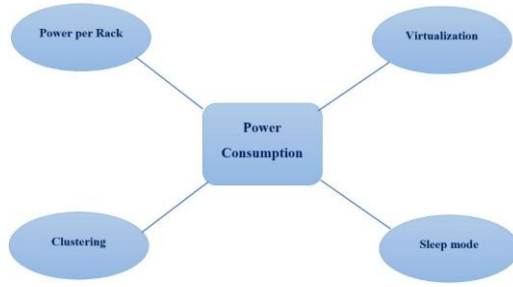


Fig. 6. Power Consumption in datacenters

The case study evaluates power-saving techniques from four perspectives: (i) power consumption of physical servers in each rack, (ii) power consumption after virtualization, (iii) power consumption after server clustering, and (iv) power consumption after placing selected servers into sleep mode. Fig. 6 shows these four evaluation stages.

B. Power Consumption of Servers Per Rack

A datacenter rack is a standardized structure that houses servers and networking equipment while providing physical support, power distribution, and cable management. Power consumption depends on the number and type of devices installed. Fig. 7 shows a rack containing 10 servers and 2 Top-of-Rack (TOR) switches, which provide network connectivity and management for the devices within the rack.

To calculate power of a rack first we need to define some variables that are

Rack = R

Server = S

Server per Rack = S/R

Power = P

Server rated power supply=Ws



Fig. 7. Power consumption per rack

Datacenter equipment is classified into primary and secondary components, and the total rack power consumption is calculated by considering the energy usage of both categories. Top-of-Rack (TOR) switches, routers, and servers are examples of primary equipment. In this study, an HP ML30 Gen10 server is assumed to consume an average of 130 W when operating at 70–80% resource utilization [27]. Table III presents the specifications and power consumption of the physical server.

TABLE III
SPECIFICATIONS AND POWER OF A PHYSICAL SERVER

Model name	HP ML30 gen 10
CPU	48
RAM	1 TB
Storage	128 GB
Cores	6 cores
Power consumption	130 W

The rack contains 10 physical servers S1, S2, S3.....S10. So, 10 servers' power will become 1300W.

The power consumption of Top-of-Rack (TOR) switches is also included in rack power calculations. TOR switches provide network connectivity for the servers within a rack. In this study, each rack contains two Juniper EX4200 TOR switches, with an average power consumption of 200 W per switch [28]. One switch is used for network management, while the other serves as a backup to ensure reliability. Fig. 8 shows two racks, R1 and R2, each equipped with 10 servers and 2 TOR switches, where backup switches provide redundancy in case of switch failure.

Secondary equipment in a datacenter includes the power system, Power Management Module (PMM), UPS, VESDA, and HVAC cooling system, which support reliable power distribution, environmental control, and fire protection. The HVAC system consumes an average of 1450 W, while the combined average power consumption of PMM, UPS, and VESDA is approximately 1050 W. The power consumption of the primary and secondary equipment is combined to determine the overall rack power consumption.

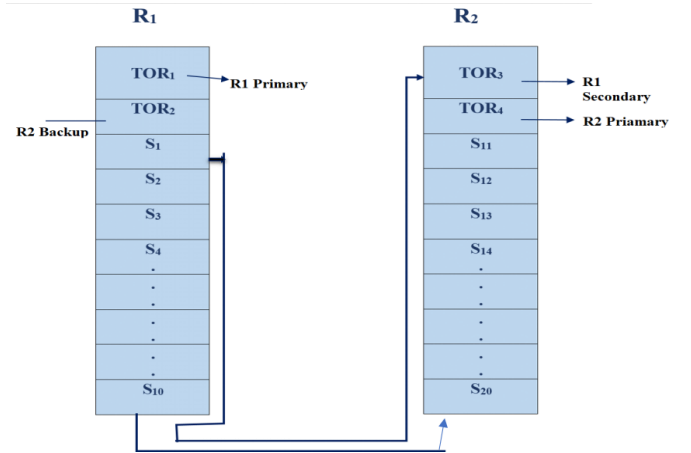


Fig. 8. Adjacent servers in racks

So, a rack overall power usage may be calculated as:

Power of a rack= P (primary equipment) +P (secondary equipment)

Power of a rack= P(ΣS) +P (ΣTOR) +P(HVAC)+P(UPS+VESDA+PMM)

Power of a rack= (130*10) +(200*2) +(1450) +(1050)

Power of a rack = 4200W.

The total power consumption of a datacenter rack is 4200 W. The annual energy cost is estimated by multiplying the rack's

power consumption by its operating time, enabling datacenter operators to evaluate yearly energy expenses. Table IV presents the annual power consumption of a rack.

TABLE IV
POWER CONSUMPTION PER WH -> PER ANNUM

Rack pdu Capacity	Rack input	Rack power in kw	Time (365 Days x 24 Hours)	Energy (kW x Hours = kWh)
32Amp single phase	230V	4.20 KW	8760hours	36792 KWH

C. Virtualization

Virtualization reduces datacenter power consumption by which allows a single physical server to execute several virtual machines (VMs). The virtual machines are managed by a hypervisor which enables applications to operate independently of the underlying hardware, improving resource utilization and reducing the number of physical servers required. Fig. 9 compares the architecture of a physical server with its virtualized environment.

Multiple VMs can run on a 48 CPU core physical server. In this case study, each server runs 10 VMs, and the rest of the CPU cores are used for background system processes for optimal performance. The server's energy after virtualization will rely on the effectiveness of the hypervisor and the amount of resources allocated to the VMs.

The estimated hypervisor overhead is 20% [29] and the energy consumption is divided between all virtual machines for the total energy consumption. To calculate the power consumption, we'll follow these steps:

Step 1: Determine the total power use of the server and rack (when no VMs are loaded):

Server Power = 130W

Rack power= 4200W

Step 2: Adjust for the hypervisor overhead:

The total power of the Rack is calculated as the sum of the power of its Server(s) and the Hypervisor Overhead. The Rack power is calculated as:

Server Power + (Server Power * Hypervisor Overhead) =
(4200W + 0.2*4200W = 5040W).

To keep things running as they are, the overhead of background tasks is added to our server as 10% overhead.

So, if you put a hypervisor and background tasks on the racks, then they will be:

=4200W+(4200*0.2) +(4200*0.1)
= 5460 W

Step 3: Distribute the power equally between 100 VMs – We have 10 servers in a rack and 10 virtual machines per server. So the total VMs will be: 100:

Power per VM = Rack Power with Hypervisor / Number of VMs

Power per VM = 5460 / 100 = 54.6W

The total virtual machine (VM) power consumption is roughly 54.6 W for the assumed equal power distribution, but it is likely to be different depending on workload and server

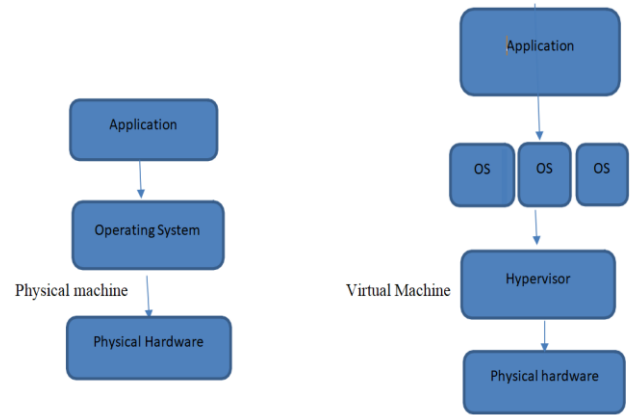


Fig. 9. Virtualization

efficiency. As compared, a physical Core i3 server consumes an average of 65 W, which equates to 100 physical i3 servers consuming around 6500 W in total. Table V shows the energy savings that are possible with virtualization as compared to running the same machines physically.

TABLE V
A COMPARISON TABLE OF ACTUAL I-3 POWER CONSUMPTION VS. VIRTUALIZED I-3 POWER CONSUMPTION

Virtual machines	Actual i-3
1 VM power= 54.6W	1 core i-3 power = 65W
100 VMs power=5460W	12 core i-3 power= 65*100= 6500W

Thus, after applying virtualization, the power of a server reduces to 16%.

D. Clustering Technique

Server clustering combines multiple servers into a single resource pool to improve performance, scalability, resource utilization, and availability. It reduces power consumption by consolidating workloads onto fewer servers, balancing workloads efficiently, maximizing server utilization, and placing idle servers into sleep mode when demand is low.

In the proposed approach, 10 servers are grouped into a single cluster by installing a hypervisor and creating shared CPU, RAM, storage, and networking resource pools. These pooled resources enable efficient allocation to virtual machines according to workload requirements, improving resource utilization and reducing overall energy consumption.

E. Sleep mode

The proposed technique aims to minimize server energy consumption by dynamically scheduling tasks while maintaining the performance requirements of cloud applications. It consists of two main components: (i) task scheduling and (ii) server ON/SLEEP mode management. The task scheduling mechanism allocates tasks based on their resource requirements, deadlines, and dependencies to achieve energy-efficient resource utilization. Fig. 10 presents the number of incoming requests or tasks during each hour, which is used to guide the scheduling process.

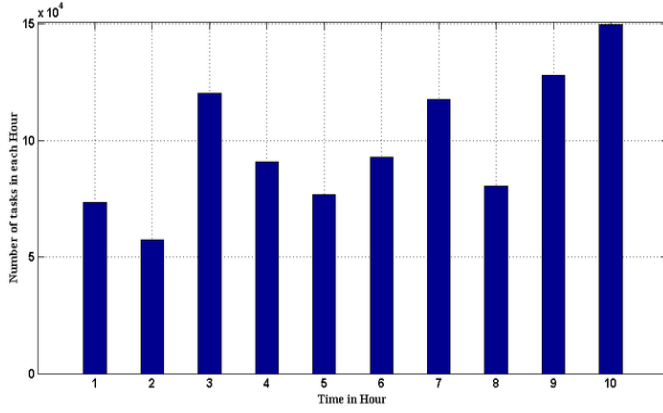


Fig. 10. Number of requests/ tasks

The proposed mechanism dynamically switches servers between ON and SLEEP modes according to workload demand. During periods of low utilization, underutilized servers are placed into sleep mode to reduce energy consumption, while increasing workloads trigger the activation of idle servers to process incoming tasks efficiently. Fig. 11 presents the server switching mechanism between the ON and SLEEP states.

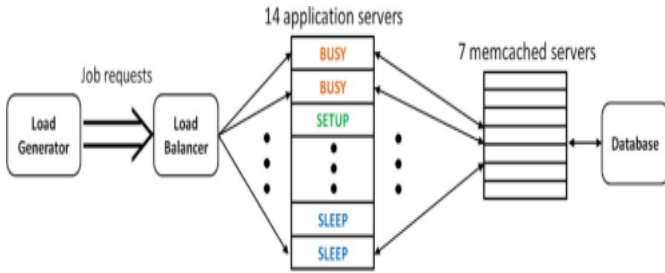


Fig. 11. Server Switching on Sleep Mode

We have 10 servers that are always in active state, but at a certain time (10pm-6am) the load on servers becomes less so that we can put some servers in sleep mode.

So, we put 3 servers in sleep mode during a specific time duration, because the load on these servers is so less, so we can put it on sleep mode. Each active server has a power rating of 54.6W. Servers in sleep mode consume negligible power (e.g., less than 1W).

Step 1: Calculate the total power consumption of the 10 active servers:

Total Power Consumption (Active) = Number of Active Servers * Power per Server
 Total Power Consumption (Active) = 10 * 54.6W = 546W

Step 2: Calculate the power consumption when 3 servers are in sleep mode:

Power Consumption = Total Power Consumption (Active) - (Power per Server * Number of Servers in Sleep Mode)

Power Consumption = 546W - (54.6 W * 3) = 546W - 163.8W = 382.2W

By putting 3 servers in sleep mode, the power consumption of the remaining active servers is reduced to 382.2W.

Thus, after applying clustering of servers and putting 3 out of 10 servers, we are able to reduce power of rack to 30%.

F. Steps Involved in our Suggested Approach

Initialize: Start the algorithm by calculating the power consumption and specifications of the physical servers.

Load Monitoring: Continuously monitor CPU usage, memory utilization, and workload demands to identify resource usage.

Virtual Machine Creation: Set up each physical server with the necessary number (N) of virtual machines.

Server Similarity Evaluation: To find appropriate servers for clustering, assess server similarity based on resource use and performance traits.

Server Grouping: Assign comparable servers to clusters while taking fault tolerance, scalability, and workload distribution into account.

Sleep Mode Identification: Use workload and CPU usage thresholds to identify idle or underused servers within each cluster.

VM Migration and Load Balancing: Before putting servers into sleep mode, move virtual machines from idle to active servers and balance the burden.

Sleep Mode Activation: Switch identified idle servers into sleep mode using power management mechanisms while maintaining data availability.

Server Re-activation: Wake sleeping servers when workload exceeds predefined thresholds to handle increasing demand.

Wake-Up Process: Ensure servers successfully transition from sleep mode to the active state and are ready to execute incoming tasks.

Termination: Evaluate the proposed approach by measuring power savings, performance, availability, and user satisfaction.

Case Study Outcome: The proposed methodology calculates the power consumption of physical servers, applies virtualization, forms VM clusters, and places idle virtual machines into sleep mode during low workload periods. This hybrid approach improves datacenter efficiency and achieves approximately 30% power savings. Table VI shows the results.

TABLE VI
RESULTS AFTER APPLYING POWER SAVING TECHNIQUES

Techniques Applied	Results
Power of rack of physical server	6500W
After applying virtualization	5460W 16% power save
After applying Clustering + Sleep mode	3822W 30% power save

V. CONCLUSION

The rapid growth of internet users has significantly increased the demand for energy-efficient datacenters. Although existing techniques have improved power management, there is still a need for more effective power-saving approaches. Thus, to lower energy usage while preserving Quality of Service (QoS), this study examined previous research and suggested a hybrid strategy that incorporates virtualization, server clustering, and sleep mode. By migrating workloads to virtual machines and placing idle servers into sleep mode, the proposed method improves resource utilization, optimizes system performance, and achieves approximately 30% power savings.

REFERENCES

- [1] X. Cao, L. Liu, Y. Cheng, and X. Shen, "Towards energy-efficient wireless networking in the big data era: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 303–332, First Quart. 2018, doi: 10.1109/COMST.2017.2771534.
- [2] X. Li, P. Garraghan, X. Jiang, Z. Wu, and J. Xu, "Holistic Virtual Machine Scheduling in Cloud Datacenters towards Minimizing Total Energy," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 6, pp. 1317–1331, Jun. 2018, doi: 10.1109/TPDS.2017.2688445.
- [3] A. Choudhary, M. C. Govil, G. Singh, and L. K. Awasthi, "Energy-efficient resource allocation approaches with optimum virtual machine migrations in the cloud environment," in *Proc. 4th Int. Conf. Parallel, Distributed and Grid Comput. (PDGC)*, 2016, pp. 182–187, doi: 10.1109/PDGC.2016.7913141.
- [4] Pakistan Electricity Generation and Consumption. [Online]. Available: <https://www.ceicdata.com/en/pakistan/electricity-generation-and-consumption>. [Accessed: 00-Month-Year].
- [5] Y. Chang, C. Gu, and F. Luo, "Energy-efficient virtual machine consolidation in cloud data centers," in *Proc. 4th Int. Conf. Systems and Informatics (ICSAI)*, 2017, pp. 401–406, doi: 10.1109/ICSAI.2017.8248325.
- [6] C. Z. Rădulescu and D. M. Rădulescu, "A performance and power consumption analysis based on processor power models," in *Proc. 12th Int. Conf. Electronics, Computers and Artificial Intelligence (ECAI)*, 2020, pp. 1–4, doi: 10.1109/ECAI50035.2020.9223124.
- [7] N. Kortas and H. Youssef, "Performance evaluation of virtual machines migration within a data center," in *Proc. Int. Conf. Engineering & MIS (ICE)*, 2017, pp. 1–7, doi: 10.1109/ICEMIS.2017.8272984.
- [8] A. Gandhi, M. Harchol-Balter, and M. A. Kozuch, "Are sleep states effective in data centers?" in *Proc. Int. Green Computing Conf. (IGCC)*, 2012, pp. 1–10, doi: 10.1109/IGCC.2012.6322260.
- [9] C. Yin, J. Liu, and S. Jin, "An Energy-Efficient Task Scheduling Mechanism with Switching On/Sleep Mode of Servers in Virtualized Cloud Data Centers," *Math. Problems Eng.*, vol. 2020, pp. 1–11, 2020, doi: 10.1155/2020/4176308.
- [10] S. Prathiba and S. Sankar, "Architecture to Minimize Energy Consumption in Cloud Datacenter," in *Proc. Int. Conf. Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 1044–1048, doi: 10.1109/ICCS45141.2019.9065682.
- [11] M. A. Haque Monil, R. Qasim, and R. M. Rahman, "Energy-aware VM consolidation approach using a combination of heuristics and migration control," in *Proc. 9th Int. Conf. Digital Information Management (ICDIM)*, 2014, pp. 74–79, doi: 10.1109/ICDIM.2014.6991413.
- [12] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M. A. Kozuch, "Distributed, Robust Auto-Scaling Policies for Power Management in Compute Intensive Server Farms," in *Proc. 6th Open Cirrus Summit*, Atlanta, GA, USA, 2011, pp. 1–5, doi: 10.1109/OCS.2011.6.
- [13] T. Bhattacharya and X. Qin, "Modeling Energy Efficiency of Future Green Datacenters," in *Proc. 11th Int. Green and Sustainable Computing Workshops (IGSC)*, 2020, pp. 1–3, doi: 10.1109/IGSC51522.2020.9291049.
- [14] A. Ali and S. E. Elavoubi, "Design and performance evaluation of site sleep mode in LTE mobile networks," in *Proc. 26th Int. Teletraffic Congr. (ITC)*, 2014, pp. 1–6, doi: 10.1109/ITC.2014.6932966.
- [15] C. Yin, J. Liu, and S. Jin, "A review of power consumption models of servers in data centers," 2020, doi: 10.1155/2020/4176308.
- [16] T. Inoue, A. Aikebaier, T. Enokido, and M. Takizawa, "A Power Consumption Model of a Storage Server," in *Proc. 14th Int. Conf. Network-Based Information Systems (NBIS)*, 2011, pp. 382–387, doi: 10.1109/NBIS.2011.64.
- [17] T. Bhattacharya and X. Qin, "Modeling Energy Efficiency of Future Green Data centers," in *Proc. 11th Int. Green and Sustainable Computing Workshops (IGSC)*, 2020, pp. 1–3, doi: 10.1109/IGSC51522.2020.9291049.
- [18] Google. [Online]. Available: <https://www.google.com/>. [Accessed: Dec. 30, 2021].
- [19] S. Zhang, X. Ye, and Y. Ren, "Prediction and Optimization on Energy Consumption of Data Center Based on Multi-layer Feedforward Neural Network," in *Proc. IEEE 8th Int. Conf. Advanced Power System Automation and Protection (APAP)*, 2019, pp. 1480–1483, doi: 10.1109/APAP47170.2019.9224637.
- [20] M. J. Usman, A. Samad, Ismail, H. Chizari, and A. Aliyu, "Energy-Efficient virtual machine allocation technique using interior search algorithm for cloud datacenter," in *Proc. 6th ICT Int. Student Project Conf. (ICT-ISPC)*, 2017, pp. 1–4, doi: 10.1109/ICT-ISPC.2017.8075327.
- [21] X. Zhan and S. Reda, "Power Budgeting Techniques for Data Centers," *IEEE Trans. Comput.*, vol. 64, no. 8, pp. 2267–2278, Aug. 2015, doi: 10.1109/TC.2014.2357810.
- [22] F. Al-Hazemi, J. Lorincz, A. F. Y. Mohammed, and F. Salamh, "Reducing Data Center Power Losses through UPS Serial Consolidation," in *Proc. Int. Conf. Software, Telecommunications and Computer Networks (SoftCOM)*, 2019, pp. 1–6, doi: 10.23919/SOFTCOM.2019.8903708.
- [23] A. Ali and S. E. Elavoubi, "Design and performance evaluation of site sleep mode in LTE mobile networks," in *Proc. 26th Int. Teletraffic Congr. (ITC)*, 2014, pp. 1–6, doi: 10.1109/ITC.2014.6932966.
- [24] X. He, X. Fu, Y. Yang, and J. Zhang, "Mobile Sink path selection and routing optimization algorithm for network energy balance," in *Proc. Int. Symp. Sensing and Instrumentation in IoT Era (ISSI)*, 2018, pp. 1–5, doi: 10.1109/ISSI.2018.8538112.

- [25] V. Inzillo, F. De Rango, and A. A. Quintana, "A sectorized directional MAC proposal for mitigating deafness and energy consumption in mobile ad hoc networks," in Proc. 15th Int. Conf., 2018.
- [26] N. Hariprasad and M. Biju, "Optimized rate and power control-based energy-saving transmission for LTE base-station," in Proc. Int. Conf. Circuit, Power and Computing Technologies (ICCPCT), 2016, pp. 1–6, doi: 10.1109/ICCPCT.2016.7530372.
- [27] HPE ProLiant ML30 Gen10 Review. [Online]. Available: <https://www.servethehome.com/hpe-proliant-ml30-gen10-review/4/>.
- [28] Juniper Community, "Community Discussion." [Online]. Available: <https://community.juniper.net/communities/community-home/digestviewer/viewthread?MID=70997>.
- [29] Microsoft, "Storage I/O Performance Tuning for Hyper-V," Microsoft Learn. [Online]. Available: <https://learn.microsoft.com/en-us/windows-server/administration/performance-tuning/role/hyper-v-server/storage-io-performance>.