Cluster Computing (2023) 26:1845–1875
https://doi.org/10.1007/s10586-022-03713-0

Energy efficiency in cloud computing data centers: a survey on software technologies

Avita Katal[1,2] · Susheela Dahiya[2] · Tanupriya Choudhury[2]

Received: 8 February 2022 / Revised: 21 June 2022 / Accepted: 7 August 2022 / Published online: 30 August 2022

Abstract
Cloud computing is a commercial and economic paradigm that has gained traction since 2006 and is presently the most significant technology in IT sector. From the notion of cloud computing to its energy efficiency, cloud has been the subject of much discussion. The energy consumption of data centres alone will rise from 200 TWh in 2016 to 2967 TWh in 2030. The data centres require a lot of power to provide services, which increases CO2 emissions. In this survey paper, software-based technologies that can be used for building green data centers and include power management at individual software level has been discussed. The paper discusses the energy efficiency in containers and problem-solving approaches used for reducing power consumption in data centers. Further, the paper also gives details about the impact of data centers on environment that includes the e-waste and the various standards opted by different countries for giving rating to the data centers. This article goes beyond just demonstrating new green cloud computing possibilities. Instead, it focuses the attention and resources of academia and society on a critical issue: long-term technological advancement. The article covers the new technologies that can be applied at the individual software level that includes techniques applied at virtualization level, operating system level and application level. It clearly defines different measures at each level to reduce the energy consumption that clearly adds value to the current environmental problem of pollution reduction. This article also addresses the difficulties, concerns, and needs that cloud data centres and cloud organisations must grasp, as well as some of the factors and case studies that influence green cloud usage.

Keywords Cloud Computing · Containerization · Data center · Load balancing · Workload categorization

1 Introduction

The last decade internet services like cloud computing and web 2.0 have changed the entire architecture of the internet ecosystem. The web, which began as a worldwide hypertext system, has developed into a distributed application platform with distinct entities for application logic and user interface. The web is the principal interface (medium) via which cloud computing distributes or makes its services available to everyone. Since time immemorial, the definition of the term web has evolved. Now web encompasses a slew of technologies and services that enable interactive sharing, collaboration, user-centred design, and application development. As a result, web 2.0 refers to the current state of internet technology in relation to the early days of the web, and it includes increased user involvement and cooperation, as well as improved communication channels. In recent years, a new computer paradigm known as cloud computing has begun to emerge. As we transitioned from web content to web of apps on the next generation web platform, web 2.0, the network cloud housed the vast bulk of user data and applications. Cloud computing is also gaining prominence as a low-cost way of software storage and distribution. Data, software and applications no longer exist on the client side in this environment; instead, they

✉ Avita Katal
avita207@gmail.com

✉ Tanupriya Choudhury
tanupriya1986@gmail.com

Susheela Dahiya
susheela.iitr@gmail.com

[1] Research Scholar, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India

[2] School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India

are viewed as abstract services and live in cloud. A cloud may be defined as a network of server-side nodes. Many underlying technologies like utility computing, Service-oriented architecture (SOA) had a direct impact on cloud computing. SOA made it possible for several services to connect with one other via a loose coupling technique in order to exchange data or organize an activity. In contrast, utility computing is a service delivery model in which a service provider makes computational resources, network administration accessible to clients on as need basis, and pays them for the usage patterns rather than a set fee. The utility model, like other types of on demand computing (such as grid computing), seeks to optimize operational efficiencies while minimising associated costs.

The word "data centre" is, by definition, an assumption. It harkens back to a time when a company's back-office computer systems were mainly devoted to data storage and were cobbled together in a basement or closet. Nobody was meant to see or notice "infrastructure," such as a sewage system or the foundation of a roadway beneath the potholes. All of these assumptions have now been debunked. The IT infrastructure of a business includes computer power, networking, and data storage. In addition, like the enterprise, it has a natural propensity to become dispersed. As with many novel notions in the IT industry, there is no universally accepted definition of what constitutes a hyperscale data centre. Hyperscale data centres are much larger than corporate data centres, and they outperform them greatly as well, owing to the benefits of economies of scale and custom engineering. The International Data Corporation (IDC), that offers new tech sector design and consultancy assistance, defines hyperscale as any data-centre with at least 5000 servers and 10,000 square feet of open field. Nevertheless, Synergy Research Group focuses on "scale-of-business criteria," that also assess a company's cloud, e-commerce, and social networking procedures instead of physical features [1]. Traditionally, data centres have used one of two techniques to provide this additional computing capacity. Horizontal scaling is inefficient in terms of energy consumption, particularly for complicated workloads. It also introduces a new issue in that each storage unit added necessitates the inclusion of the appropriate compute and network resources required to use them. Data centers require proper cooling systems to work properly which leads to increase in expenses. Hyperscale computing aids in lowering the cost of data disruptions. Systems that fail due to a lack of hyperscale computing lose money, goodwill, and the services of their IT employees who must find out why the services failed among other business operations losses. Before the system can be used again, they may need to fix compliance concerns and alert consumers. Companies might lose data for hundreds of thousands or millions of dollars. Hype scaling enables businesses to reduce downtime caused by high demand or other problems. Hyper scaling also allows IT systems to be restored considerably more quickly. The cost of cooling and maintaining the facility's temperature is one of the most significant operating expenditures a data centre confronts. A hyperscale data center optimizes airflow throughout the structure. The combination of vertical and horizontal scaling increases the utilization of energy by hyperscale data centres dramatically. Although these facilities are generally highly energy efficient, their sheer scale exerts huge power demands on the world's energy supplies. From 2015 to 2023, hyperscale data center's energy usage is anticipated to nearly quadruple, making it the world's highest proportion of data centre energy consumption [1].

The architecture of the current services of cloud is highly centralized meaning that the different types of services can be run through a single site called data centers. The data centers are increasing rapidly due to rapid advancement in cloud computing. The energy consumption of data centres alone will rise from 200 TWh in 2016 to 2967 TWh in 2030 [2]. Despite the COVID-19 problem, the worldwide market for Internet Data Centers, which was predicted at US$59.3 billion in 2020, is expected to reach US$143.4 billion by 2027, increasing at a CAGR of 13.4% between 2020 and 2027 [3]. In the United States, the Internet Data Centers business is estimated to be valued US$16 billion by 2020. China, the world 's second biggest industry, is expected to have a data centre industry of US$32 billion by 2027, with a 17.5% CAGR between 2020 and 2027 [3].

Power Usage Effectiveness (PUE), a unit of analysis for data centre power consumption efficiency, is a real-time and annual study of total facility power split by IT hardware power, or a measurement of power 'loss' flowing to non-IT devices. The ideal PUE is 1.0, which corresponds to 100% effectiveness [4]. However, it is nearly difficult to achieve. The average yearly data centre PUE in 2021 was 1.57, a small increase with the average of 1.59 in 2020, and keeping with the overall trend of PUE stagnation over the previous five years. The bulk of the energy consumed is used to power the servers; however, they generate heat and must be cooled [5]. The need for data centres is increasing due to the exponential rise in data gathering and consumption. In general, cloud computing uses a large number of data centers and servers to service a big number of clients using a pay-per-use model. Such resources cover a big area and demand a considerable amount of electricity for networking devices, cooling technologies, displays, and server farms, among other things. Making the resources green using green technology has thus become a main goal of several government and industry organizations. Green IT, from an environmental standpoint, and to deal with IT-

related environmental challenges, offers a broad number of approaches and practices through several green initiatives.Using energy more effectively is one of the most straightforward and cost-effective method to save money, decrease greenhouse gas pollutants, generate employment, and satisfy rising power demands. Improved efficiency has the potential to reduce greenhouse gas (GHG) emissions, other contaminants, and water usage. Energy efficiency can bring long-term advantages by lowering total power consumption, minimising the need for new energy generation and distribution infrastructure investment.

Cloud providers that host a range of applications must follow service-level agreements (SLAs), achieve low access latencies, meet task deadlines, and provide secure, dependable, and effective data management. Low-cost hardware designs and capacity planning tactics employed in back-end data centres for energy savings might often conflict with the commercial objectives of cloud providers. Due to the online assessment of dynamic elements such as workload allocation, resource allocation, cooling plan, inter process communication, and traffic conditions, data centre energy management is a challenging operation. Moreover, as energy prices rise and the cloud service pricing market becomes more competitive, cloud providers are being obliged to investigate energy-saving alternatives for back-end data centres [6]. Typical workload on the data center is usually about 30% and does not necessitate the use of all computer resources [7]. As a result, certain unused equipment can be turned off to achieve energy savings while meeting data center workload expectations. However, scheduling data centre resources necessitates careful consideration of data center traffic patterns [8], client SLAs [9], latency and performance concerns [7], and data replication [10].

Software level modelling is very important for energy efficiency in data centers because the different software components require power to process the tasks. The software developed should be able to benefit from advancements achieved at the hardware component level. If the software generated is not as efficient as hardware technical advances and consumes a large number of resources, overall energy consumption will remain high, negating the entire goal of building green data centres. The survey papers published until date do not include the complete details of energy efficiency techniques employed at each software layer in the data center. They do not include different mechanisms employed for modelling the containers energy consumption that is one of the emerging areas in cloud computing domain. This paper explains the various techniques employed for energy efficiency in container technology that is first a kind of effort in this direction as per author's knowledge. This paper also provides information about the environmental effect, as well as the policies/standards available for assessing energy efficiency in data centers. This survey report serves as a foundation for academics working in the field of green computing, as it covers layer-by-layer software modelling of data centres, as well as an emphasis on the many research issues to which researchers should target their efforts.

Thus, four research queries have been answered:

RQ1 What are the numerous options used at the software level like operating system, virtualization and application to reduce the usage of power by data centers?

RQ2 What are the various strategies utilised in data centres, both virtualized and non-virtualized systems, to minimise power usage?

RQ3 What are the major impacts of a data center on the environment?

RQ4 What are the major software academic difficulties for developing green data centres?

2 Related work

Despite the fact that there has been a substantial quantity of research on data centre energy usage estimation and forecasting, there have been comparatively few studies in this sector. The following papers describe software-based technologies for developing energy-efficient green data centres.

The authors of [11] presented an analysis on cloud computing energy usage. The research considered both public and private clouds, as well as the energy consumed in switching and communication, information computation, and storage. They demonstrated that power usage in transit and switching may account for a sizable portion of total energy demand in cloud computing. Their proposed method regards Cloud Computing (CC) as an equivalent of a classic logistics and supply chain issue that takes into account the power usage or expense of computing, keeping, and transporting physical goods. The authors in [12] highlighted the reasons and difficulties associated with excessive power / energy usage, as well as presented a taxonomy of energy-efficient computing system architecture at the OS, hardware, virtualization, and data centre levels. They evaluated important contributions in the area and linked them to their classification to guide future development and research initiatives. They investigated and categorised numerous ways to controlling a system's power usage from the OS level using DVFS and other power-saving strategies and algorithms. Many research efforts targeted at developing efficient algorithms for regulating CPU power usage have culminated in the

widespread acceptance of DVFS in the form of an implementation in a kernel module of the Linux operating system. In addition, the authors in [13] highlighted research difficulties connected to the competing needs of enhancing the quality of services (QoSs) supplied by cloud services while lowering energy consumption of data centre resources. They addressed the idea of creating an energy-efficient data centre controller suitable of combining data centre capabilities while reducing the effect on QoS objectives. They investigated strategies for controlling and coordinating data centre resources in order to achieve energy-efficient operations. They also offered a central controller concept and proposed resource controller cooperation. Energy-saving hardware ideas for data centre resources were also thoroughly examined. The authors in [14] discussed the different mechanism and architectures for the design of energy efficient data centers. They investigated the different power models for virtual machines, operating systems and software applications. Their systematic technique enables them to investigate a variety of challenges typical in power simulation at different stages of data centre systems, such as: (i) few modelling efforts devoted at overall data centre power consumption (ii) many cutting-edge power models rely on a few CPU or server specs; (iii) the efficacy and accuracy of these power models is still unknown. They completed the study by identifying important obstacles for future studies on building efficient and optimum data centre power models based on their findings. The authors in [15] conducted research and created a taxonomy based on pre-existing energy efficiency related surveys, i.e., research on energy saving surveys. Existing surveys were classified into five categories: those on the power consumption of all cloud-related processes, that on a particular level or component of the cloud, those on all energy-efficient methodologies, that on a specific energy-efficiency technique, and those on other energy-efficiency-related studies. A taxonomy and survey on surveys are conducted from the viewpoints of foci, views, target system, and years. The survey findings on energy consumption savings measures are then examined, laying the groundwork for their future work in the subject of energy consumption.

The survey articles described above are either incomplete or having limitations. They have not gone into length on the issues of power usage at the application, virtualization, and operating system layers of software. Furthermore, these survey studies did not give comprehensive information on the solutions that may be deployed at the data centre level and containers (operating system virtualization). These survey reports also did not get into specifics concerning environmental variables or case studies. This article is an extension of the authors' earlier work, which provides a study of hardware solutions for establishing green data centres [16]. This paper provides the detailed information about the different techniques that can be applied at the individual software levels and in-depth information about the power modelling at operating system virtualization and data center level along with the work done in different problem-solving approaches like VM migration, workload categorization, load balancing and VM placement. The article also addresses the environmental impact of data centers and ends with a discussion of the recent research challenges in the construction of green data centres.

The articles in this study were obtained from several sources, including IEEE, Springer, and Elsevier. Web of Science and Scopus are the databases used to collect publications. All of the publications included have been peer reviewed, and the bulk of them were published and 2015 and 2020. This research includes publications that focus on software-based methods for energy efficiency in data centres. This analysis excludes publications that were not peer reviewed and were published before to 2015. Studies that are not published in English and do not provide details about software innovations for energy savings in data centres are not evaluated for inclusion.

3 Motivation

Data centres are critical, energy-intensive infrastructure that provide large-scale Internet-based services. Power utilization models are essential for creating and enhancing energy-efficient processes in data centres in order to decrease excessive energy use. In recent years, the necessity of energy efficiency in data centres has increased substantially and has become more complicated. To guarantee high availability of data, all elements of the data centre design must perform their given tasks to minimise data centre downtime that requires appropriate energy support. Power supply, technical cooling, and technical security are all part of the technical infrastructure, which is the foundation of all information technology (IT) infrastructures. Any physical infrastructure outage, no matter how slight, has a major impact on the functioning of IT services. The essential qualities of a green data centre are energy efficiency and low global impact. A green or sustainable data centre is a data storage, management, and dissemination facility in which all systems, especially mechanical and electrical frameworks, improve energy efficiency. It produces less carbon footprints, saves money and increases efficiency. These eco-friendly data centres help modern enterprises save power and reduce carbon emissions. Globally, their use is rising among both major organisations and small and medium-sized businesses (SMBs). From data collection through processing,

assessment, and distribution, such data centres can efficiently fulfil the objectives for a plethora of corporate data. The objective of this manuscript is to look at the current research on green cloud computing and outline the major concerns that have been raised which consumes more power in data centers.

The software components require a large amount of power to perform their operations in the data centers. To reduce the energy usage of data center in respect to the software layer, different techniques can be applied at individual software level. The CPU core is the primary resource consumer in computation-intensive tasks (s) and cloud system's storage resources in data-intensive operations. Connected devices such as network cards, routers, switches, and others require a substantial amount of energy when performing communication-intensive operations. The operating system (OS) resides between the physical hardware and the applications layers of the data centre architecture. Most of the research is done on the hardware level for power consumption in data centers but the software level is equally important in order to reduce the power usage. Software developed should be able to exploit the advancements done at the hardware components level. If software developed is not as efficient as the hardware technology advancements and consumes a large number of resources then overall energy consumption still remains high defying the whole purpose of developing green data centers. Physical hardware is the component that consumes the IT power, while applications produce the demand for resources. Hence, looking into the details of power modelling /energy consumption at the software layer becomes equally important. Apart from the various techniques that are applied at the OS, virtualization and the application level, problem solving approaches like load balancing, workload categorization, VM placement and VM migration helps in minimizing the energy usage by consolidating the physical servers and dynamically modifying operations. These approaches prove to be effective in lowering energy usage in high performance cloud data centers.

In this article, the analysis is performed at many levels such as OS, virtualization, application and data centre to determine the energy usage by different software layers in data centers. The case studies are also included for better understanding the importance of green cloud data centers. The research challenges are discussed along with their solutions for reducing energy consumption in data centres.

4 Energy usage in data centres: a system's perspective

This segment analyses the whole data centre to the required levels based on electricity use. The data centre model utilised in this study is depicted in Fig. 1 below. Every computer system is made up of two components: hardware and software. A data centre also has two primary components: software and hardware.

These layers can be enhanced or optimized so that the power usage by data centers can be minimized. The software layer is categorized into three sublevels: OS layer, virtualization layer, application layer. For establishing the green data centre, a taxonomical method for software approaches is offered, as illustrated in Fig. 2. To fulfil the aims of green cloud computing, many strategies at the individual programme level might be used. Aside from software approaches, external factors to the data centre such as government-imposed laws and policies, organisations, and renewable energy are also considered to fulfil the goals of green cloud computing.

5 Data center power modelling at individual software level

RQ1 What are the various approaches used at the software level like operating system, virtualization and application to reduce the usage of power by data centers?

5.1 Operating system level

The operating system is placed between the two layers: the application and the hardware. The main role of applications is to create the resource demand and the OS job is to manage the resources for all these applications. The main component that consumes power is the physical hardware but it is very essential to keep a check on the events that consume power at the operating system level if energy usage optimization at data centre is to be done at the software levels too. The power usage breakdown of the operating system functions is shown in Fig. 3. Data-path and pipeline topologies that allow for numerous problems and out-of-order execution were found to squander 50% of the total power of the OS processes investigated. Furthermore, the clock consumes 34% power and different levels of cache consumes the remaining power.

Operating System Power Management (OSPM) is a mechanism utilized by OS to manage the power of the underlying platform and transition of it between different power modes. OSPM allows a platform or system to adopt the most efficient power mode and is applicable to all

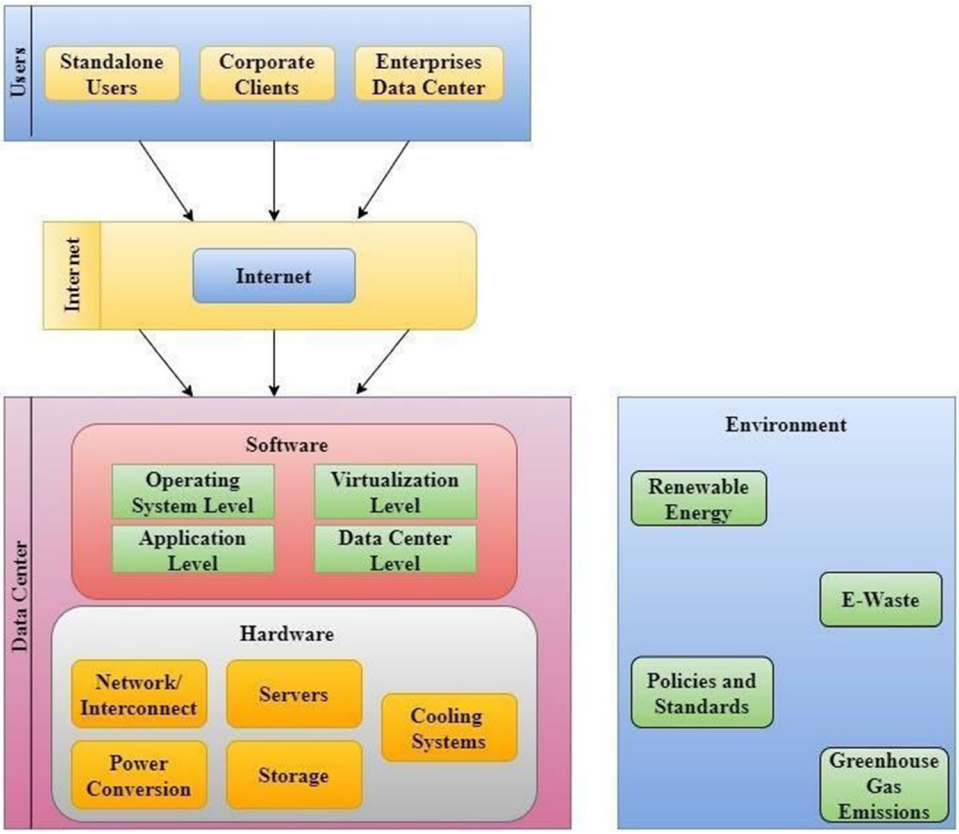


Fig. 1 A comprehensive picture of data centre energy usage modelling

devices and components inside the platform/system. OSPM is also known as OS-directed configuration and Power Management.

The trade-off between quality and power efficiency has been intensively examined and analysed, since control over running voltage and energy management has been largely shifted from the hardware and firmware level to the operating system. Herzog et al. [18] offers PolaR, a method for automatically determining energy-efficient setups, as well as a Linux implementation. PolaR proactively chooses optimal settings by integrating application profiles and system level data, and no application adjustments are required. They take into account bank shots (configuration settings unrelated to power management) in addition to controlling the system in the proper manner. OS development teams recognised the value of energy as a resource on par with time. With energy seen as just another resource available to the operating system, operating system internals (such as locking mechanisms) were changed to accommodate for this new perspective in order to produce energy-aware operating systems. Scordino et al. [19] illustrates how the deadline scheduler and the cpufreq subsystem may be changed to relax the restriction that the frequency scaling technique used only when no real-time processes are running and to create an energy-aware real-time scheduling approach. They described the architectural issues they encountered when trying to deploy the GRUB-PA algorithm on a real OS like Linux. Experiment findings on a multi-core ARM architecture demonstrated the efficacy of their suggested solution.

With the advancement of semiconductor and software technologies, the capabilities of an embedded system have grown by incorporating new features and performance. In recent years, the network has also advanced as communication infrastructure and contact with server systems has become essential. So far, TCP / IP connections between servers and embedded devices have been established by two methods. The first is a technique that includes a TCP/IP stack in embedded devices. The second is a technique of communicating via a "gateway" (to translate end-device communications). There are several server system composition options, such as putting a server in-house, establishing a server at a data centre outside of town, and utilising cloud computing. Smaller, more widespread and less well known "embedded data centres" consume half of all data centre energy, or about 1% of all energy generated in the United States. In general, embedded data centres are data centre facilities that have less than 50 kW of IT demand [20]. Server rooms, server closets, localised data centres, and several mid-tier data centres are among them.

Energy harvesting technologies based on rechargeable batteries are a popular option for addressing the issue of

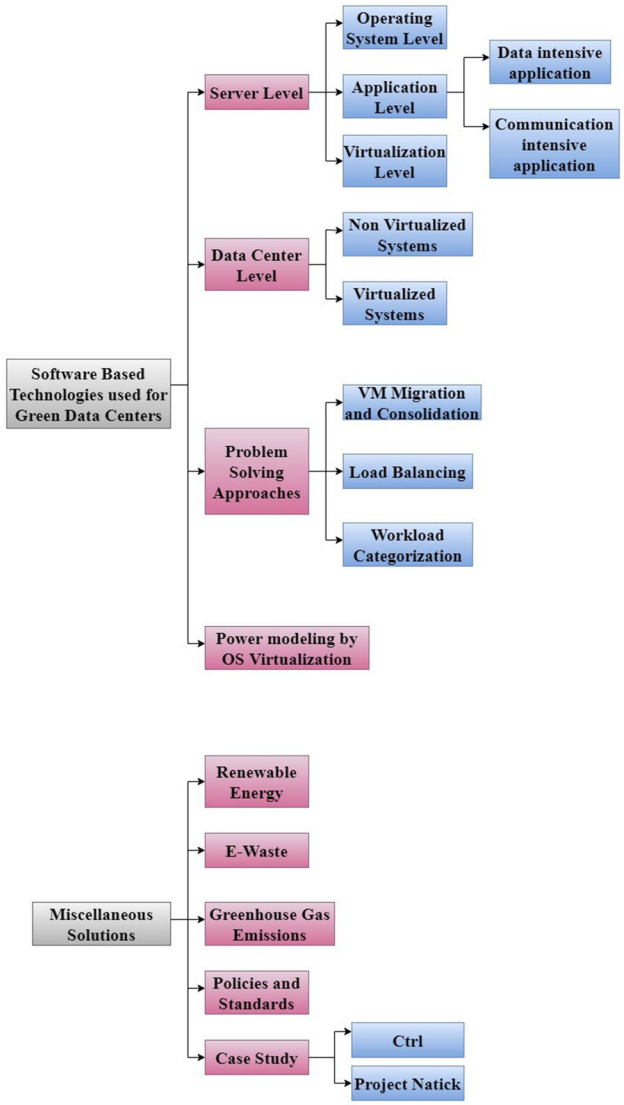


Fig. 2 A systematic summary of data centre power demand prediction at the software level

delivering continuous power to deeply implanted devices such as wireless sensor nodes. However, if the use of a node is not carefully planned, the battery may be depleted too quickly, making continuous operation of such a device unfeasible. To regulate the flow of energy, an energy-management solution is necessary. Buschhof et al. [21] presented an idea that enables the modelling of hardware energy usage and the creation of energy-aware device drivers for the embedded OS. Their drivers can account for the energy usage of each driver function call with greater than 90% accuracy. Similarly Levy et al. [22] presented Tock, a unique embedded OS for low-power systems that utilises the limited hardware-protection processes accessible on latest microcontrollers and type-safety functionalities of the Rust programming language to offer a multiprogramming ecosystem that provides software fault separation, memory protection, and efficient memory governance for dynamic applications and services written in Rust. Low-power embedded operating systems frequently use the same memory areas for both applications and the operating system. Merging applications and the kernel allows them to easily exchange references and gives efficient procedure call access to low-level functionality. This monolithic method often necessitates building and installing or upgrading a device's apps and operating system as a single unit.

5.2 Virtualization level

Virtualization uses software to construct a layer of abstraction above computer equipment, enabling the actual features of a single computer, storage, disk, and so on—to be separated into numerous virtual computers, also called as virtual machines (VMs). Each virtual machine created for a user can be allocated an individual operating system on a single physical machine that makes sure of the performance of the virtual machines and failure isolation among them. Hence, a Virtual Machine Monitor (VMM) / Hypervisor is responsible for multiplexing of resources to the virtual machine and helps in the management of the power to perform efficient operations. The two ways in which a virtual machine monitor can take part in the management of power:

- A VMM acts as a power-aware operating system. It verifies the entire performance of the system and applies the DVFS (Dynamic Voltage and Frequency Scaling) or any DCD (Dynamic Component Deactivation) techniques to the components of the system.
- The other way is to leverage the policies for the management of power and knowledge of applications at OS level. Power management calls can be mapped from different virtual machines. In addition, a coordinated system wide limits on the power can be enforced.

Virtualization technology has regained prominence in computer system architecture during the last few years. Virtual machines (VMs) provide a development route for adding new capabilities—for example, server consolidation, transparent migration, and secure computing—into a system while maintaining compatibility with existing operating systems (OSs) and applications. Multiple VMs executing on the same core in contemporary virtualized settings must adhere to a single management of power controlled by the hypervisor. These settings have different limitations. It does not enable users to specify a desired power control scheme for each virtual machine (or client). Second, it frequently affects the energy efficacy of some or all VMs, particularly when the VMs need competing energy management strategies. For mitigating above

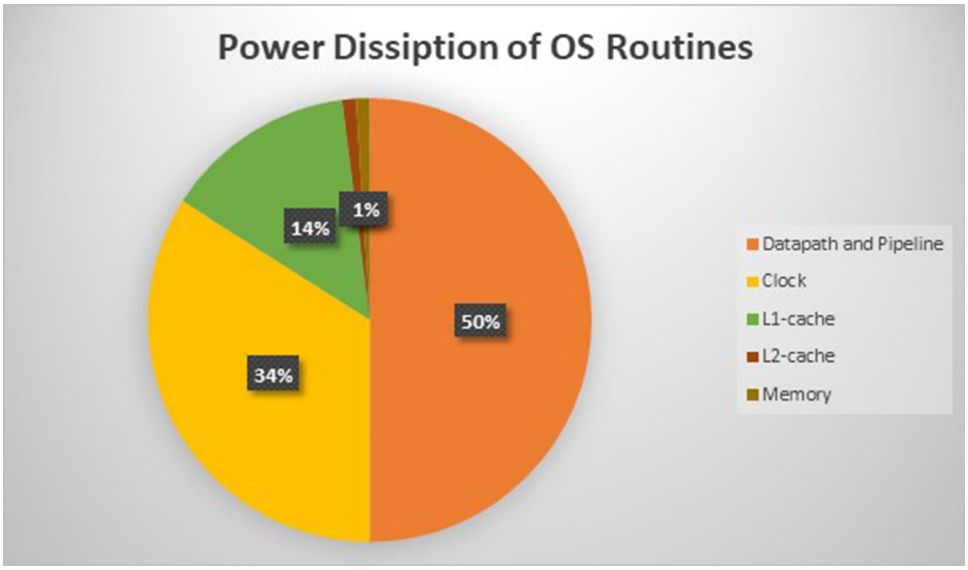


Fig. 3 Power dissipation of OS routines [17]

problems, Kang et al. [23] suggested a per-VM power control method that enables each VM's guest OS to utilise its chosen energy administration strategy and avoiding similar VMs from competing with each other's energy control strategy. When compared to the Xen hypervisor's default on demand governor, Virtual performance (VIP) minimises power usage and enhances the completion time of CPU-intensive applications by up to 27% and 32%, respectively, without breaching the SLA of latency-sensitive implementations. Furthermore, Xiao et al. [24] examined the VM scheduling model and the I/O virtualization paradigm in terms of energy-efficiency optimization. They provided a power-fairness credit sequencing approach with a novel I/O offset method to achieve speedy I/O performance while simultaneously raising energy conservation. Apart from this, Prabhakaran et al. [25] introduced VM resource calibration. They created a system to reduce the energy usage of virtual servers by utilising controlled feedback architecture as well as power monitoring services.

5.3 Application level

Energy efficiency is always a major concern in cloud computing and when it comes to the application level many recommendations have been made to optimise energy usage at the system level. However, the rising variety of contemporary workloads necessitates a better analysis at the application level to allow adaptive behaviours and to minimise global energy consumption. For achieving energy efficiency especially at the application level, Ho et al. [26] concentrated on batch applications executing on VMs in data centres. They investigate the application's characteristics, computed the energy spent on each job, and estimated the application's energy usage. The evaluation focuses on assessing software efficiency in aspects of performance and power consumption per job especially when there exists shared resources and heterogeneous environments based on profiles of energy, with the objective of determining the best resource configurations. The applications were divided into two categories: data intensive application and communication intensive application. Data-intensive applications that generate, analyse, and transmit enormous volumes of data had been executed with minimal regard for energy efficiency. Large amounts of energy may be consumed because of issues such as data management, migration, and storage. A communication-intensive application is made up of one or more interdependent services, and the communication traffic between them is typically distinct. The communication traffic requires a large amount of power and various techniques for dynamic power management that can be applied for energy efficiency.

Cloud services are referred to as Software as a Service (SaaS) on the uppermost layer of cloud computing architecture, which is a software delivery technique that offers on-demand permissions. SaaS providers, in general, provide extra layers of cloud computing, and hence keep client data and tailor apps to match customer demands. This situation reduces the initial cost of obtaining new software and infrastructure significantly. Customers are under no obligation to maintain or build infrastructures on their sites. They only need a fast network to access their apps rapidly. SaaS providers service a variety of businesses by utilising the same infrastructure and software [27]. This method is clearly more power saving than installing several copies of software on various infrastructure, which can reduce the requirement for new equipment. The lower the volatility in

demand, the better the forecast and the bigger the energy savings. SaaS companies must model and monitor the energy efficiency of their software design, execution, and deployment because they primarily sell software hosted on their own data centers or resources from IaaS providers. The SaaS provider selects data centres that are not only power saving but also close to consumers. This is particularly crucial for social networking and gaming applications, because users are often ignorant of their impact on environmental sustainability. SaaS companies can also provide Green Software Services hosted in carbon-efficient data centres with less replications.

The authors in [28] introduced a solution for dynamic software consolidation in order to decrease the number of VMs utilized. Software consolidation allows dynamically collocating different software applications on the same VM. The proposed method may be used with VM consolidation, which places several VMs on fewer actual machines. The authors of [29] proposed an energy-aware application element migration technique that calculates the load of data centre servers by taking the number of components connected to the servers, the number of rental people attempting to access the software applications, the component strike rate, and various other important factors into account when trying to decide which elements to migrate. To save energy, the server is turned off once all components of the underused servers have been moved. They used discrete event simulation to test their suggested approach.

5.3.1 Data intensive applications

Energy and power consumption are becoming increasingly significant in today's high-performance computing (HPC) systems. New cluster systems are planned to be no more than 20 MW in power [30], with the goal of attaining exascale performance as quickly as possible. The rise of big data and cloud computing has given the globe with huge opportunities as well as enormous challenges. However, the growing trend in cloud energy demand as a result of the fast-expanding volume of data to be delivered and analyzed has propelled cloud computing, along with the big data phenomenon, to become the primary source of energy consumptions and, hence, CO2 emissions. To decrease the power usage of data intensive applications in cloud data centers, the authors in [31] have presented an adoption framework for the data intensive applications whose primary goal is to minimize energy usage. The proposed framework is driven by the values of data gathered from the data streams or data sets of the applications. The authors looked at the data from different facets, from its general to its domain-specific features, and then combined them to provide a number indicating the data's importance. Furthermore, Malik et al. [32] have developed ECoST, a method for optimising energy efficiency and self-tuning for data-intensive workloads. They proved that fine-tuning settings at the application, microarchitecture, and system levels simultaneously opens up the possibility of co-locating applications at the node level and improving server energy efficiency without compromising functionality.

Energy efficiency is a critical component in the development of big supercomputers and low-cost data centers. However, adjusting a system for energy efficiency is challenging due to the competing needs of power and performance. The authors in [33] utilized Bayesian optimization (BO) to optimise a graphics processing unit (GPU) cluster system for the Green500 list, a prominent energy-efficiency rating of supercomputers. BO might obtain an excellent configuration by defining the search space beforehand with minimum information and prior experiments. As a result, BO could remove time-consuming manual tweaking and shorten the system's occupancy time for benchmarking. Furthermore, because of its influence on operating costs and processing system rate of failure, energy efficiency became a crucial component of high-performance computing. Processors are outfitted with low-power methods such as DVFS and power capping to increase the power effectiveness of such devices. These approaches must be tightly managed in relation to the load; otherwise, considerable productivity loss and/or energy usage may occur because of system overhead expenditures. The authors in [34] proposed a workload-aware runtime power-control strategy for effective V-f control. The proposed technique incorporates thread synchronisation conflict and delay due to Non-Uniform Memory Accesses to find an acceptable V-f value (NUMAs).

MapReduce is used for data processing in modern data centers. It is known as the programming model that can be used for the processing and generation of large data items. The MapReduce programming model processes huge amounts of data by executing a series of data-parallel jobs that work on distinct sections of the data set. MapReduce platforms, which are runtime environments, allow customers to scale up their programmes fast and easily. In order to optimize the energy efficiency for MapReduce, Tiwari et al. [35] have proposed a configurator based on performance and energy models to enhance MapReduce system energy efficiency. It considers the dependence of the energy consumption and performance of a cluster on MapReduce parameters. Their proposed solution improves the energy efficiency of up to 50% in two structurally distinct clusters of typical MapReduce applications.

5.3.2 Communication intensive applications

Communication intensive application programs are made up of a series of tasks that share a vast number of messages over the process of computing. These applications are designed by utilizing the Message Passing Interface (MPI). Dynamic end-to-end request needs and uneven route power effectiveness, as well as uneven and time-varying link usage, throughput and delay limits for service needs, all offer challenges to power effective connections. The authors in [36] proposed a multi-constraint optimization framework for improving energy efficiency in cloud computing technology including geographically dispersed data centres linked by cloud networks. Their technique improves energy savings in both data centres and cloud networks. An intelligent heuristic technique is provided to handle this model for dynamic request demands among data centres as well as among data centres and consumers. Furthermore, the authors in [37] established a simultaneous optimisation of server power usage, network connectivity, and migration expense with workload and host heterogeneity constrained by resource and bandwidth restrictions in VM placement. Although Integer Quadratic Program (IQP) can only be addressed for relatively small systems and but it has been decomposed into master and price sub problems that can be solved using the column generation approach for larger systems.

6 Power modelling at operating system virtualization

Virtualization is regarded as the most important technique for initiating modern clouds by sharing the physical resources among applications and the users. Virtualization allows the efficient use of resources like software, hardware, energy, etc. by consolidating many underutilized machines on to a single system. Virtualization is divided into five categories: application, server, desktop, network, storage and based on the execution environment. The detailed classification of the virtualization techniques is shown in Fig. 4.

The conventional virtualization can be further divided into two different categories: Para and Full virtualization. Full virtualization can be defined as the creation of virtual processor, storage devices, memory and I/O devices in order to run the various guest operating systems on a single machine so that the guest OS is not aware about the presence of virtualization. In case of full virtualization, the goal is to run the unmodified binaries of the operating system. The code of the operating system remains unchanged, that is why it is not aware of the fact that it does not have the required permissions to run privileged instructions [38]. This gives rise to problems in certain architectures(x86) as some privileged instructions may silently fail. Hypervisor resorts to a binary translation mechanism where validation is done on the set of instructions that may fail silently to resolve the above-mentioned problem. The other approach of conventional virtualization is Paravirtualization (PV). Paravirtualization is a kind of CPU virtualization in which instructions are handled at compile time via hyper calls. Instead of trying to imitate an entire hardware eco system, PV is a virtualization technology advancement in which a guest OS is reconfigured even before to setup within a VM to allow all guest OS inside the scheme to share resources and effectively cooperate.

The other approach to virtualization is containerization that is also known as the virtualization at OS level. Virtualization technology utilizes the hypervisor that helps in emulating the hardware resources in order to run the guest operating systems on top of it. The concept behind this was that an application running on the hardware seldom makes use of the entire resources. Virtualization creates copies of the functionality of the physical resources that includes the computational, storage, memory, networking resources that run an application. Containerization, a new concept introduced lately, is on the verge of development and growth. Containers also aid in lowering administration expenses. Since they use the same OS, just one needs to be monitored and fed for security patches, and so on.

Virtualization allows several operating systems on a single physical server's hardware, while containerization enables to install many programs that use the same OS on the same virtual machine or host. The architectural difference between the virtualization and containerization is as shown in Fig. 5.

Containerization is a lightweight virtualization solution that facilitates the distribution and operation of application services across platforms such as edge/fog, cloud, and IoT. Containerization is changing the working of industries because it is storage and resource efficient, performance efficient, cost efficient, portable, energy efficient and extremely quick during boot up. Although the traditional VMs enhance the efficiency of the physical servers, they incur a fair amount of overhead in costs and effort. A container model enables the data center's owners to simply deliver the code they need to perform the function of the application without all the extra dependencies. This leads to the efficient use of the resources within the data center. With the traditional virtual machines, the guest operating system rather than the actual mission of the application utilizes a major portion of the resources. The lighter footprint of the containers has many advantages throughout the data center. A container model needs fewer racks, less energy for cooling and power, less software licenses, less

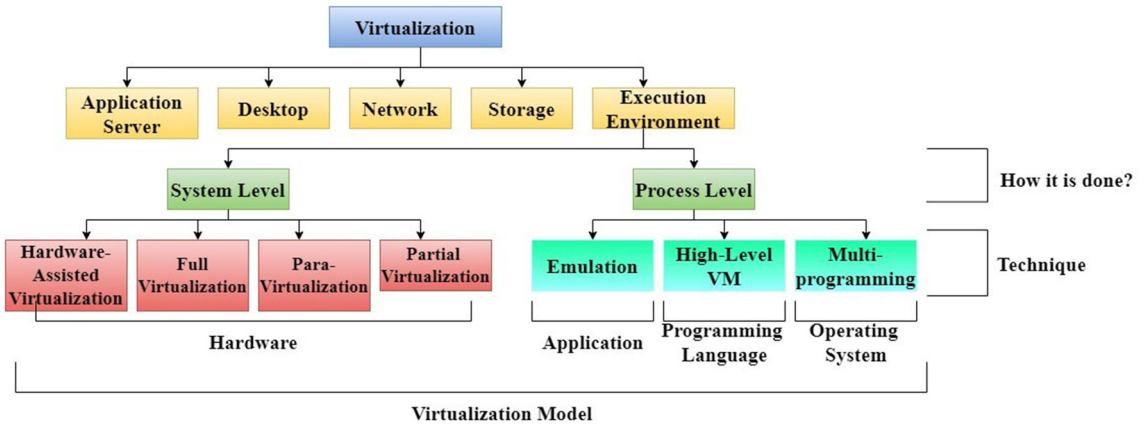


Fig. 4 Types of virtualizations

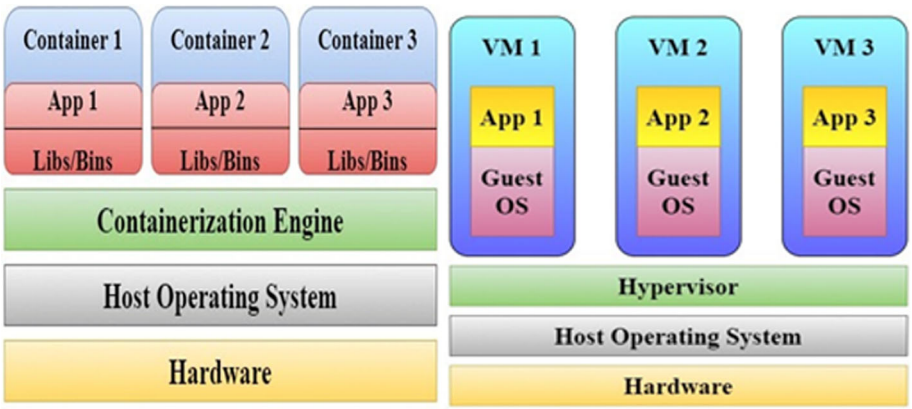


Fig. 5 Architectural difference between virtualization and containerization

maintenance. Containers offer a higher level of service quality than other virtualization technologies. Furthermore, because they require fewer resources than virtual machines, additional entries are anticipated and will be integrated on the same server, decreasing energy usage because fewer servers are planned to operate the same number of services. Docker, when configured to a maximum latency of 3000ms, can operate up to 21% more services than KVM. Docker provides this service in this setup while consuming 11.33% less energy than KVM [39].

In containerization, containers execute onto the shared operating system kernel in isolation. One of the major differences between containerization and hypervisor-based virtualization is that in containerization, the objects that are virtualized are limited to the resources of the global kernel that enables containerization to start various virtual environments onto the common host kernel. The created virtual machines are resource intensive and do not allow individual application's functionalities/components to run in isolated environments. The execution of an individual component or application in an isolated environment needs a separate virtual machine.

Migration of applications running in virtual machines to another data center or machine/server also needs the whole OS to be migrated along with it. Virtualization technology is developed to exploit the existing resources but the operations of the workload in virtualization do not consume all the resources available to them, which leads to a significant wastage of resources also. In addition, the virtual machines do not incorporate the leftover resources in capacity planning and distribution across all the virtual machines and workloads. On the other hand, containerization enables the individual functionality of an application to run independently making it possible for different workloads to run on the same physical resource. These can execute on bare metals or on top of hypervisors or cloud infrastructure too. Containers have the capability to create isolated OS environments within the same host, different functionalities of the same application run by sharing the Linux Kernel in containerization [40]. Table 1 shows the difference between Virtualization and Containerization.

The performance of containers and virtual machines is compared in experimentation conducted by various researchers. The experiments are performed on Docker and KVM. Below are the results of the comparison of the

Table 1 Difference between virtualization and containerization

Parameters	Virtual machine	Container
Guest operating system	Hypervisor allows multiple and distinct OS to run on the same host. Each VM is allocated with a specified amount of memory on which the Kernel functions.	All the guests share the same base OS and its Kernel. The image of the Kernel is loaded into the physical memory.
Security	Security of the VMs depends on how the hypervisor is implemented.	Container software like Docker have built-in security features that can be leveraged.
Performance	The VMs have a small overhead when compared to containers as the translation of machine instructions occurs from the host OS to guest OS.	There is little to no overhead in using containers as the applications are executed in the base OS itself.
Isolation	Hypervisor isolates each VM from host OS as well as from other VMs. This means files, libraries etc. cannot be shared between guests and the host.	Each container has its own set of file systems that can be shared between other applications.
Startup time	VM takes sufficient time to boot.	Containers take less time to boot as compared to VMs.
Storage	VMs take ample storage as the whole Kernel and the secondary programs associated with the OS need to be installed.	Since the base OS is shared, containers require less storage.

containers and virtual machines on the basis of different parameters [41]:

Throughput metric is used to calculate the output of a task when the CPU is exposed to a compression High Performance Computing (HPC) test. It has been seen that native and Docker compression performance is comparable, but KVM is slower. HPC performance is comparable on native and Docker but very sluggish on KVM due to abstraction, which acts as negative in this scenario. The CPU schedulers have no effect on the processor in either the native or Docker arrangement; therefore, there is no difference in performance. The parameter bandwidth is used to measure the speed of memory access operations. According to many benchmarks designed to test memory in linear and random-access approaches, the efficiency of native, Docker, and KVM systems is almost similar for a variety of workloads with very little variation. The testing was carried out on a single node using large datasets. Container-based systems returned unused memory to the host, resulting in more efficient memory utilisation. Virtualization systems suffered from double cache since the host and virtual computer used the same memory blocks. Bandwidth is used to assess network communication performance. Bulk data transfer using a single TCP connection, similar to the client-server architecture, is the communication situation. Because the TCP/IP stack has distinct regulations for sending and receiving data, the data transfer rate is measured in both directions. The NIC, which utilises CPU cycles to measure overhead, is the primary component that causes a bottleneck in performance. In terms of performance, Docker employs bridging and Network Address Translation (NAT), which lengthens the path. Dockers that do not utilise NAT operate similarly to native systems. KVM performance can be increased if the VM can interact directly with the host, bypassing the in-between layers. Latency is another network metric that may be used to assess performance. Throughput is also used to assess the efficiency of disc operations. As previously stated, Docker and KVM add relatively little overhead when compared to native; however, there is a significant performance difference in KVM's case due to a potential bottleneck in fibre channel. Docker has no cost for random read and write operations, while KVM's performance suffers considerably. The system's I/O scheduler has an impact on disc performance.

Containers are gaining popularity and will be a significant deployment strategy in cloud computing. Consolidation techniques are widely employed in the cloud environment to maximise resource usage and minimise power consumption. To minimize the power consumption through container consolidation, Piraghaj et al. [42] have presented the problem of container consolidation and have compared the different algorithms. They evaluated their performance against parameters like, SLA violations, consumption of energy, average rate of transferring containers, and the typical number of VMs generated. The consumption of power by data centers at time t can be calculated as:

$$P_{dc}(t) = \sum_{i=1}^{N_S} P_i(t) \tag{1}$$

Pdc(t) denotes the consumption of power by the data center during t time, Ns denotes Number of servers and Pi(t) denotes the consumption of power by Server i during the time t. The metric for SLA is calculated as the fraction of the difference among the allocated and the requested CPU for each VM [43].

$$SLA = \sum_{i=1}^{N_S} \sum_{j=1}^{N_{VM}} \sum_{p=1}^{N_V} \frac{CPU_r(vm_{j,i}, t_p) - CPU_a(vm_{j,i}, t_p)}{CPU_r(vm_{j,i}, t_p)} \quad (2)$$

Ns represents Number of servers, Nvm represents Number of VMs, Nv represents Number of SLA Violations, CPUr(vmji, tp) represents the amount of CPU needed by VM j on server i during tp time, CPUa(vmji, tp) represents the amount of CPU amount assigned to VM j during tp time.

For the reduction of the consumption of power by data center that consists of Nvm VMs, M containers, and Ns servers, they designed the equation as:

$$min(P_{dc}(t) = \sum_{i=1}^{N_S} P_i(t)) \quad (3)$$

In the above equation, Pdc(t) denotes the consumption of power by data center during the t time, the consumption of power by server i at time t is denoted by Pi(t), and Ns represents Number of servers.

Since container-based service aggregation is a tough process, the cloud data centre consumes a lot of power due to a lack of management over the data centre systems. Because containers need a smaller resource footprint, consolidating them in servers may result in limited resource availability. Nath et al. [44] have designed an energy efficient service based on the consolidation of the containers. The authors have formulated a service consolidation problem as an optimization problem by considering the minimizing the power usage by the data centers.

In data centers, there are n servers and m containers in the system. They represent N = {Ni: i ∈ (1. . n)}as sets of data center servers and Cn = {Cnj : j ∈ (1,. ., m)} as sets of data center containers. They have defined the total consumption of energy by a server as:

$$E_{Ni} = E_{Ni}^{rev-data} + E_{Ni}^{rcv-frm-manager} + (1 + \alpha + \beta) \times E_{Ni}^{comp} + \beta \times (E_{Ni}^{rev-frm-node} + E_{Ni}^{send-to-node}) + \alpha(E_{Ni}^{offl-node} + E_{Ni}^{rcv-frm-node}) + E_{Ni}^{send-to-client} \quad (4)$$

In the above equation α ($0 \leq \alpha < 1$) denotes the data offload percentage to other server (worker) and β ($0 \leq \beta < 1$) denotes the data received percentage from another server (worker).

Collecting and delivering bits in a host uses energy. It can be defined as:

$$E_{Ni}^{offl-node} = E_{Ni}^{send} = E_{Ni}^{rcv} = E_{bit} \times B \quad (5)$$

In the above equation, Ebit denotes the energy consumption by sending one bit of data and B denotes the bits received or sent by the user.

The central processing unit (CPU) is the main component that consumes the power in the data center. The energy consumed for computation is [45]:-

$$E_{Ni}^{comp} = \left(E_{Ni}^{max-comp} - E_{Ni}^{idle-comp}\right) \times U_{Ni}^{CPU} + E_{Ni}^{idle-comp} \quad (6)$$

The container-based cloud-computing concept has grown through time as a versatile and power efficient resource-use approach. Cloud providers strive to enhance utilisation of resources and resource use when executing container aggregation, which includes VM selection and placement. Shi et al. [46] have designed TMPSO, for energy aware consolidation of containers. The proposed algorithm integrates the heuristic and greedy optimization mechanism to get the balance between the computation and performance cost.

In a cloud, the authors have assumed a set of physical machines PM = {PM1, …, PMm, …, PMc}, a set of virtual machines VM = {VM1, VM2, …, VMi, …, VMv}. Each virtual machine VMi has CPU Ci, memory Mi and operating system Oi, i.e., VMi (Ci, Mi, Oi). They have assumed that the capacity of each virtual machine is the same so for each PMm ∈ PM, it can be demonstrated as PMm(CC, CM). A CSP first associates each of the applications to a container that satisfies their needs of the resource that includes CPU c, Memory m and operating system o. These containers are demonstrated by triple Cj (cj, mj, oj). They have assumed that each Cj is assigned to a VM. VMi ∈ VM to satisfy the needs of the resources including CPU cj, Memory mj and operating system oj. They represent allocation λj : Cj → VMi.

Finally, each VMi is allocated to a PMm, the allocation can be denoted as γi : VMi → PMm.

For the reduction in the consumption of power by data center, the authors have designed objective function for the consolidation of containers as:

$$minimizeE = \sum_{m=1}^{c} P_m.z^m \quad (7)$$

In the above equation, Pm represents the consumption of energy by physical machine PMm and zm represents the binary variable, zm ∈ {0, 1}, indicating whether PMm is active.

The below equation shows the connection among the consumption of energy and CPU utilization.

$$P_m = \begin{cases} P^{idle} + \left(P^{busy} - P^{idle}\right).u_{cpu}^m, if N_{vm} > 0 \\ 0, if N_{vm} = 0 \end{cases} \quad (8)$$

where the CPU utilization is, u_{cpu}^mPidle and Pbusy are the consumption of power by PM during the utilization is 0% and 100% respectively, Nvm denotes all the VMs allocated to the PM.

The utilization of memory by PMm is calculated by Eq. 10, OMi denotes the overhead of memory of VMi.

$$u_{cpu}^{m} = \sum_{i=1}^{N_{vm}} (OC_i + C_i).y_i^m \tag{9}$$

$$u_{mem}^{m} = \sum_{i=1}^{N_{vm}} (OM_i + M_i).y_i^m \tag{10}$$

Apart from the literature that is mentioned above, there are various other works that help in the placement of containers, the placement of virtual machines so that the power consumed is less, and resource utilization is efficient in the container as a service environment. The authors in [47] have addressed the issues of container placement and VM placement in the two stages. To solve the container placement problem, they merged the two strategies. The proposed solution is divided into four decision-making processes: VM selection and creation, PM selection and creation. A hybrid technique, genetic programming-based hyper-heuristics (GPHH) and human-designed rules was used to tackle the two-level container allocation problem. To assess container allocation, they took into account accumulated power. The energy consumed P_d^t of all active PMs over the time period t1, t2 is added to calculate the accumulated energy (see Eq. 11). In other words, the authors add the energy usage of all PMs at each time interval ti. Fan's [48] energy model of a PM (Eq. 12) is a commonly used model. In their energy model, Pidle and Pmax represent the energy use while a PM is idle and fully utilized. However, cpu(d) indicates a PMd's CPU use at time t. The purpose of container assignment is to decrease net power consumption, i.e., M in Accumulated Energy (AE).

$$AE = \int_{t=t_1}^{t=t_2} \sum_{d=1}^{D} P_d^t \tag{11}$$

$$P_d^t = P_d^{idle} + \left(P_d^{max} - P_d^{idle}\right).u_{cpu}^t(d) \tag{12}$$

In the problem, they examined three sorts of restrictions. To start, the total capacity use of containers need not surpass the threshold of the target VM. The aggregated energy needs of VMs must not surpass the target PM's capacity. Second, a container may only be assigned once. Third, they use an affinity restriction to limit container deployment to OS-compatible VMs alone.

Furthermore, Chen et al. [49] presented many-to-one stable matching method and a container placement technique MLSM. They started with an early container hosting technology to shorten migration durations by the use of a trustworthy matching mechanism. This programme utilises resemblance algorithms as a finding choice strategy for containers and VMs. The resource usage rate is used to order the virtual machine preference list. According to the simulation findings, the algorithm may cut energy usage by an average of 12.8% when compared to the First Fit method.

In the technique proposed by Chen et al. [49], the energy usage is by Eq. 1:

They often assess server power utilization utilizing the CPU efficiency ratio, as the CPU is the most commonly utilised element of energy expenditures in terms of server utilisation. The CPU utilization ratio of each server is equal to the

$$\sum_{j=1}^{N_{vm}} \sum_{k=1}^{N_c} U_c(k,j,t)(t) \tag{13}$$

The server energy consumption is estimated using the Eq. 8:

They assume M containers, Nvm VMs, and Ns servers in the containerized cloud computing paradigm, and the energy consumption problem may be stated as Eq. 3.

They assume M containers, N VMs, and K servers in the containerized cloud computing paradigm, and the energy consumption problem may be stated as Eq. 1:

Apart from this, Al-Moalmi et al. [50] have addressed the issues related to the placement of container and VM in Container as a Service (CaaS) environment by considering the optimization of the power consumption and resource utilization. They proposed an algorithm based on the Whale Optimization Algorithm (WOA) to solve the two problems that are container placement and VM placement. Each VM and PM in this cloud data centre may be assigned to any single type of container and VM. To put it different way, every type of container may be hosted by a single VM, and any type of VM can be hosted by a single PM, according to the constraints stated below:

$$\forall(Cont_a, V_b) and (Cont_c, V_d) if Cont_a = Cont_c, then V_b = V_c \tag{14}$$

where $Cont_a$and $Cont_c$ are container identifiers and Va and Vd are VM identifier.

$$\forall(V_a, P_b) and (V_c, P_d) if V_a = V_c, then P_b = P_c \tag{15}$$

where $V_a$and V_c are VM identifiers and Pa and Pd are PM identifier

$$\sum_{l=1}^{L} Cont_{cpu_l} \leq V_{cpu_l}, \forall cpu_l \in C, and V_i \in V) \tag{16}$$

$$\sum_{l=1}^{L} Cont_{ram_l} \leq V_{ram_l}, \forall Cont_l \in C, and V_i \in V) \tag{17}$$

$$\sum_{i=1}^{M} V_{cpu_i} \leq P_{cpu_{capacity_j}}, \forall V_i \in V, and P_j \in P) \quad (18)$$

$$\sum_{i=1}^{M} V_{ram_i} \leq P_{ram_{capacity_j}}, \forall V_i \in V, and P_j \in P) \quad (19)$$

Equations (14) and (15) specify the prerequisites for presenting containers to VMs and VMs to PMs, respectively, while Eqs. (16)–(19) ensure that a collection of containers' total consumed capacity does not surpass the host VM's CPU and memory capacities. Similarly, the overall resources used by a group of VMs should not exceeding the host PM's Memory and CPU capabilities.

7 Dynamic power management at data center level

RQ2 *What are the various strategies utilised in data centres, both virtualized and non-virtualized systems, to minimise power usage?*

When developing higher-level power models for data centres, it is important to understand the intricacies of the lower-level elements that account for the total power use of the data centre. A technique is often focused on workload reduction among physical nodes in data centres. The goal is to assign as little physical resources as feasible to requests / virtual machines while shutting off or placing unused resources to sleep / hibernate. The allocation difficulty is twofold: first, new requests must be allotted; second, the performance of present services / VMs must be constantly checked, and if required, the allocation must be altered to give the best possible power-performance trade-off concerning stated QoS. This section delves into data centre power models.

7.1 Non-Virtualized

Power consumption is increasing in large-scale systems such as data centers and supercomputers. These systems are typically measured based on peak demand. Since the power consumption of these devices is not proportional, their energy usage stays high even when the workload is low. Shutdown processes have been established to match the number of servers that are actively engaged in the workload processing. However, because of the potential influence on performance and hardware concerns, data centre administrators are cautious to utilise such tactics. Furthermore, the energy advantage is usually overestimated. The authors in [51] have evaluated the potential benefits of shutdown procedures by accounting for shutdown and boot up costs in terms of both time and energy. They investigated the energy savings provided by suspend-to-disk and suspend-to-RAM approaches, as well as the influence of shutdown processes on the energy consumption of future models with various CPUs. Furthermore, the authors in [52] presented different shutdown models that may be utilised under current and future supercomputer limitations, considering the impact of closing down and waking up networks as well as the idle and off states seen after such procedures as they influence power usage of the resources.

7.2 Virtualized systems

Virtualization and Cloud computing are enabling innovations for the creation of resource planning algorithms that are energy-aware in virtualized data centres. Indeed, one of the primary problems for large data centres is to reduce power usage, both to save money and to reduce environmental effects. The authors in [53] developed a one-of-a-form of combined server and network reduction model that considers the power consumption of both switches capable of transmitting traffic and servers hosting virtual machines. Under QoS constraints, it shuts access points and sends information to the least energy-consuming host over the most energy-efficient route. Due to the model's complexity, a quick Simulated Annealing-based Resource Consolidation (SARC) approach is provided. Furthermore, the authors in [54] developed the energy-aware fault-tolerant dynamic scheduling system (EFDTS), a dynamic task assignment and scheduling method that uses a fault tolerant mechanism to maximise resource usage and minimize energy consumption. In the task assignment scheme, a task classification approach is designed to divide incoming tasks into separate classes and then redistribute them to the most appropriate VMs based on their classes in order to minimize mean response time while minimizing energy usage. Apart from this, the authors in [55] presented an energy-conscious management method for virtualized data centres that is based on dynamically adjusting and scaling computer capability to workload factors. They created a new ontological model for describing the energy and performance aspects of data center operations. To address the issue of energy usage, the authors of [56] proposed the Energy and Performance-Efficient Task Scheduling Algorithm (EPETTS) in a heterogeneous virtualized cloud. The proposed algorithm is divided into two stages: initial scheduling, that aims to minimize completion time and meet task deadlines while not keeping in account power consumption. The second stage is task reassignment planning, which enables for the best execution location within the timeline limit while using the least amount of energy.

8 Problem solving approaches

Virtual machine migration, load balancing and workload categorization are the various problem-solving techniques that can be used for the reduction in the power consumption. These techniques are used to migrate the virtual machine when the threshold is attained for the particular server, balance the load among the different VMs and categorize workload type before placing them onto the server. The various ML algorithms are used on top of these approaches to efficiently manage the power consumption in data centers.

8.1 Virtual machine Migration

Virtual machine migration can be defined as sending the VM from one host to another by remaining connected with the application or client. The Virtual Machine Migration (VMM) can be categorized as live migration and non-live migration of virtual machine as shown in Figs. 6 and 7.

Live VM migration refers to the moving of VMs from one server to another when the host system stays active. There are two forms of live virtual machine migration: pre-copy live VMM and post-copy live VMM. Non-live virtual machine migration is defined as migrating a VM from one server to another by turning off the virtual machine on the host server. Non-live migration stops or shuts down the VM prior to transfer, depending on whether it wants to continue running services after transfer. When a virtual machine is terminated, its operating states are wrapped and transferred to the destination location. Live migration is the process of migrating a functioning VM or application among PMs without interrupting the client or service.

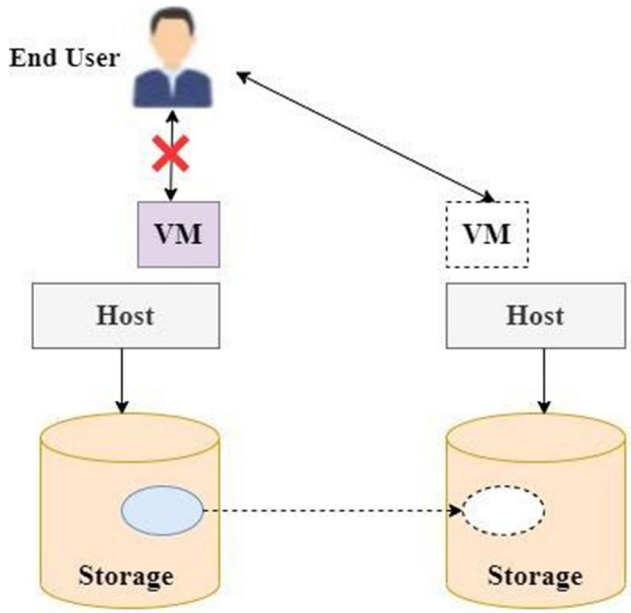


Fig. 6 Non-Live VM migration

Consolidation of virtual machines (VMs) is a typical technique for lowering energy usage based on peak, off-peak, or average CPU use of VMs in order to execute them on the least number of servers while preserving service quality (QoS). There are different techniques for the live migration like pre copy migration [57], post copy migration [58], hybrid VM migration [59], dynamic self-ballooning [60], Adaptive Worst Fit Decreasing [61], Check pointing/recovery and trace/replay technology [62], Composed Image Cloning (CIC) methodology [63], Memory management based live migration [64], Stable Matching [65], Matrix Bitmap Algorithm [66], Time Series based Pre-Copy Approach [67], Memory Ballooning [68], WSClock Replacement Algorithm [69], Live Migration using LRU and Splay Tree [70]. Apart from these, the various machine-learning approaches are also used to migrate the VM from one host to other. The techniques like autoregressive integrated moving average [71], support vector regression [72], linear regression, SVR with bootstrap aggregation [73] were also used for VM migration. These approaches are used to forecast and manage resources effectively in the data center, as well as to calculate the energy consumption. Moreover, metaheuristics are also used for the migration of virtual machines. The techniques like Firefly Optimization [74], Particle Swarm Optimization [75], Ant Colony Optimization [76], Biogeography-Based Optimisation [77], Discrete Bacterial Foraging Algorithm [78] are also used for the migration of virtual machines. These approaches optimise energy usage, QoS, resource use, or all three.

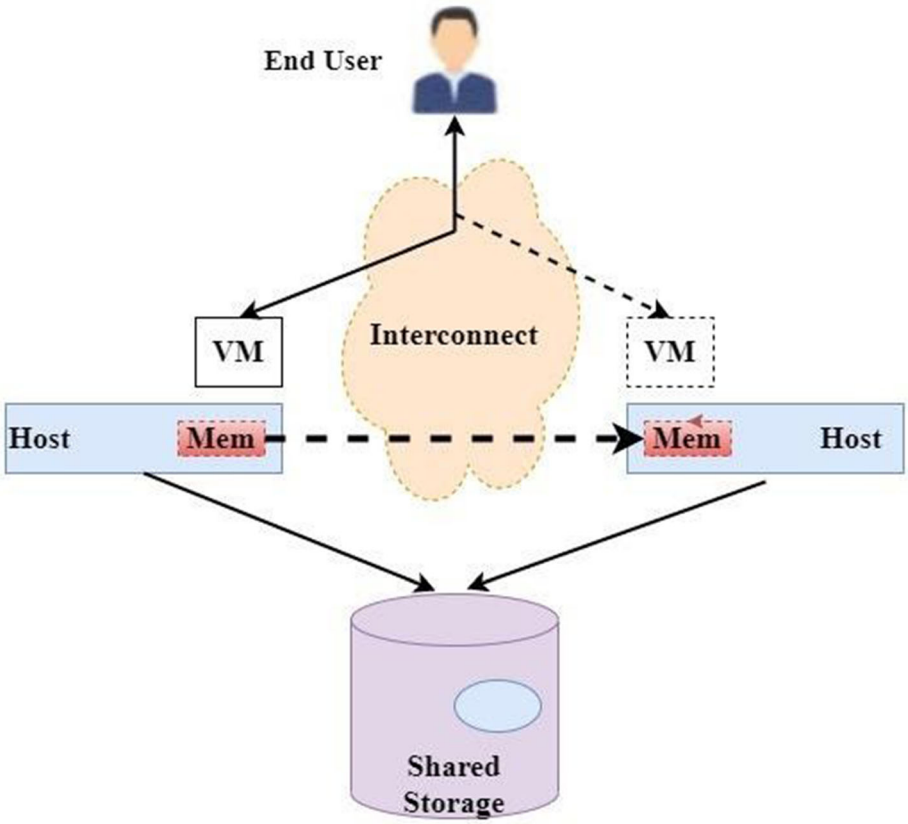
The purple box in (8) represents a VM that has been shut down or terminated on the originating host.

8.2 Load balancing

The practice of equally splitting workload in a distributed environment such that no processor is overloaded, under loaded, or idle is known as load balancing. Load balancing assists in the acceleration of various constrained parameters such as execution speed, response time, device reliability, and so on. Load balancers are highly efficient where huge workloads will quickly overload a single computer or SLAs need high levels of service efficiency and response times for certain business processes. The users deliver multiple requests and load balancer that is installed prior to the cloud server handles these requests. Load balancer distributes the incoming workload to the different cloud servers. Figure 8 shows the mechanism of load balancing.

Load balancing (LB) gives a well-organized solution to a wide range of difficulties in a cloud environment. LB plays an essential factor in the system's efficiency and

Fig. 7 Live VM migration



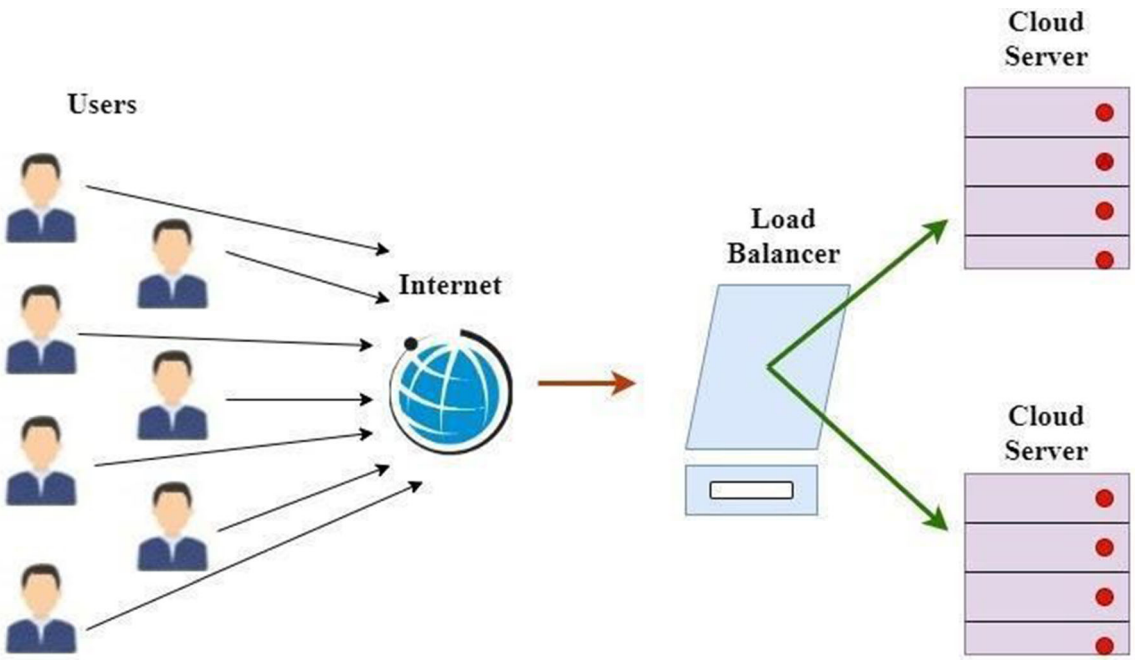


Fig. 8 Load Balancing mechanism in cloud computing environment

robustness. LB in cloud computing is one of the most difficult and valuable research topics for spreading work across virtual machines in data centres. As a result, a method for improving system efficiency by balancing workload among VMs is required. Load balancing strategies come in a variety of forms that balance the requests of the resources. These are Round Robin [79], Equally Spread Current Execution Algorithm [80], Throttled Load

Balancing Algorithm [81], Biased Random Sampling [82], Min-Min Algorithm [83], Max-Min Algorithm [84] and Token Routing [85]. The above-mentioned techniques were not able to dynamically balance the workload in a cloud-computing environment; therefore, the machine learning approaches were introduced. There are various machine learning algorithms like K-Nearest Neighbors [86], deep neural networks [87], multi-layer perceptron [88], Simulated Annealing [89] that were used for LB in cloud environments. These approaches enable accurate and practical decision making the resource allotment to inbound requests, resulting in the selection of the most relevant applications to finish. There are two types of metaheuristics algorithms: nature-inspired algorithms and evolution-based algorithms. There are various nature inspired algorithms like Swarm behaviour-based algorithm [90]: Ant colony optimization [91], Particle Swarm Optimization (PSO) [92], Artificial bee colony (ABC) optimization [93], modified Particle swarm optimization (MPSO) and improved Q-learning algorithm [94], Bat Algorithm [95], Crow inspired metaheuristic algorithm [96], Shuffled Frog Leaping Algorithm [97], Honey bee behaviour [98]. Depending on task levels, these strategies give tasks to VMs while guaranteeing equitable load sharing and generate extra number of available slots in series or parallel mode. Moreover, for complicated and huge sample space issues with a hazy sample space various evolution-based algorithms are employed. Techniques like Genetic Algorithm (GA) [99] are also used for LB.

8.3 Workload categorization and prediction

Workload is the total amount of effort done by a targeted server for a fixed period. Workload is classified before being transferred to the virtual machines that minimizes the probability of server overutilization, eliminates the need for virtual machine relocation, and thereby improves energy consumption. Web applications, web servers, distributed data storage, containerized microservices, and other workloads that require broad processing capacity are prevalent. In cloud data centers, workload classification and characterisation are used for resource planning, application performance management, capacity sizing, and projecting future resource requirements. An accurate estimate of future resource demand assists in meeting QoS requirements and guaranteeing effective resource utilisation. Workloads are categorised based on computational paradigms, technology stack, resources, and applications, as seen in Fig. 9. Based on the processing methodology, workloads are divided into two types: batch workloads and interactive workloads. Based on resource requirement, the workload is classified as Memory, CPU, IO and database. These requirements include scalability, flexibility, extensibility, and administration. Moreover, the cloud must provide capabilities that meet the best-in-class demands of the organisation, such as privacy, realistic reliability, and economy. Computer equipment are organised differently in different computing environments, and they share data between themselves to analyze and solve problems. One computing ecosystem is made up of different computational resources, software, and networks that help with computation, sharing, and problem solving. Workload can be classified into three types based on the generation: Synthetic, Real and Cloud. Furthermore, based on the application, the workload is classified into four categories: Web, Social Network, Video Service etc.

The load-balancing problem falls under the category of NP-complete. As a result, application developers frequently employ heuristic or stochastic approaches to solve it. Initially, the characterization of workload was done using statistical methods like mean and standard deviation, Auto Correlation Function, Pearson Coefficient of Correlation, Coefficient of Variation, and Peak to Mean Ratio [100]. These approaches were utilised for a thorough characterisation of both requested and actually used resources, including data relating to CPU, memory, disc, and network resources. Parameters like, CPU usage, memory consumption, network bandwidth, storage bandwidth, and job length were frequently included in data center traces. Not all qualities are equally significant in workload categorization and characterization. When all of the features are considered, the model's complexity increases. This being one of the drawbacks of statistical methods, clustering (unsupervised learning) was used to classify the workloads. For better categorization of workload, different types of clustering like Hierarchical clustering [101], Density based clustering [102] etc. were used. Apart from the unsupervised clustering techniques there are various other supervised learning techniques like Support Vector Machine (SVM) [103], Stochastic Gradient Descent (SGD) [104], Logistic Regression (LR) [105], Random Forest (RF) [106], Multi-Layer Perceptron (MLP) [107], Backpropagation neural network [108] were used. After categorising the workload, the prediction is performed in order to forecast the future workload. Forecasting of workload, management of resources dynamically and scaling proactively may all aid in the achievement of a variety of essential objectives. Accurate forecasting of near-term workload, for example, has a direct influence on response time, SLA violations over-provisioning, and under-provisioning concerns. Effective workload management improves system scalability and throughput. Furthermore, by limiting over-provisioning of virtual resources, cloud DC power consumption, cost, and the number of unsuccessful requests may be reduced, and satisfaction of customer can be enhanced.

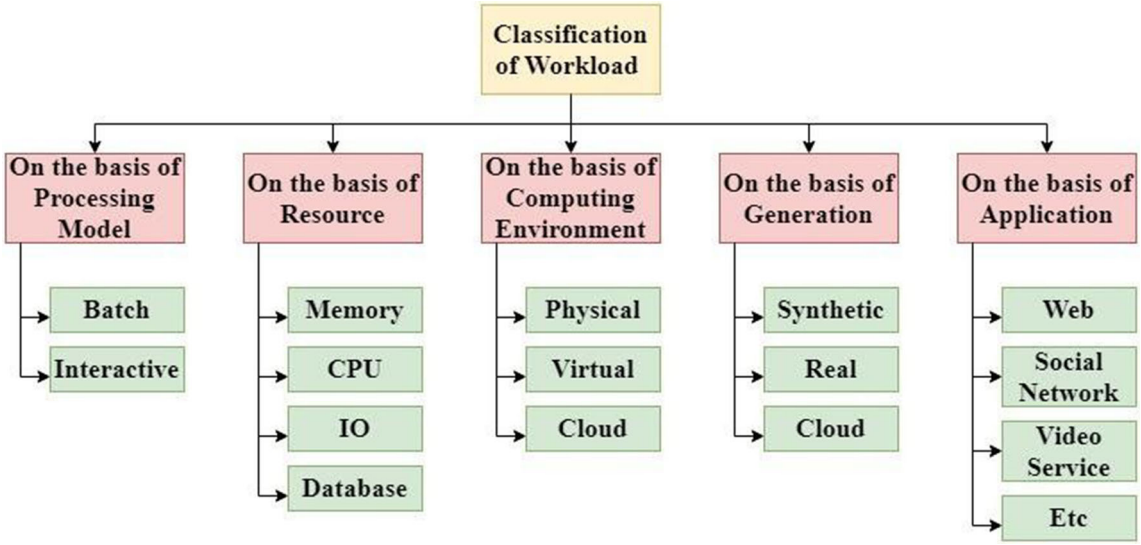


Fig. 9 Classification of workload

There are many techniques that can be utilized to predict the future workload. These include: regression-based schemes that includes ARIMA-based schemes [109], Support vector regression-based schemes [110]; Classifier-based schemes that includes SVM-based schemes [111], Random forest-based schemes [106], Artificial neural network-based schemes [112], Bayesian-based schemes [113], Deep learning-based schemes [114]; Stochastic-based workload prediction schemes that includes Hidden Markov model-based schemes [115], Queuing model-based schemes [116]. Apart from this there are various other approaches like Grey predicting-based schemes [117], Autocorrelation clustering-based schemes [118], Chaos-based schemes [119], Kalman filter model-based schemes [120], Wavelet-based schemes [121], Collaborative filtering-based schemes [122] and Ensemble-based schemes [123] that can be used to predict the workload. The combination of the above-mentioned techniques are also used to predict the future workload. These include SVR and Kalman filter [124], ARIMA and RNN [125], ARIMA and wavelet decomposition [126] (Fig. 9).

8.4 VM placement

Virtual machine placement refers to the process of determining the appropriate PM for a certain VM. As a consequence, a VM placement algorithm finds the ideal VM to PM connection, whether it is a new VM placement or a VM migration for placement re-optimization. A VM placement method may be roughly classified into two categories based on the aim of placement: Power-based [127] and QoS-based [127]. VM Placement strategies are primarily categorised as under, based on the type of principal strategy employed to achieve a suitable VM-PM mapping: Constraint Programming [128], Stochastic Integer Programming [129]. The above-mentioned approaches are not suitable for today's scenario, as these approaches cannot predict the future based on the previous history. For the suitable placement, machine-learning approaches can be used for the placement of virtual machines. Reinforcement Learning [130], Artificial Neural Network [131], and Fuzzy reinforcement learning [132] are some of them. Population based techniques begin with a collection of single solutions that grow from one generation to another. This category is centred on exploration and provides for greater variety in the search process. In order to discover the Pareto optimum solutions, population-based techniques employ the idea of dominance in their screening process. The techniques used are Genetic [133], Ant Colony Optimization (ACO) [134], Memetic [135], Firefly [136], Whale optimization [137], Sine-Cosine Algorithm and the Salp Swarm Algorithm [138]. Single solution–based algorithms begin with a single solution, which is then modified and transformed throughout the optimization process. These algorithms are exploitation-focused, which means they try to enhance the search strength in certain locations. The techniques include EAGLE algorithm [139], Imperialist competitive algorithm [140], Krill herd algorithm [141].

9 Environment

RQ3 *What are the major impacts of a data center on the environment?*

9.1 Renewable energy

In the last few years, it can be seen that there is an exponential rise in data centers developed by different companies to provide services like cloud computing. They consume very large amounts of electricity for normal functioning. Apart from the high consumption of electricity, the increase in the consumption of energy by cloud data centers results in adverse effects on the environment. In many parts of the world, the electricity is produced by burning coal that leads to negative results like increased carbon dioxide emission and increase in pollution.

The growing use of renewable energy plants, in particular, presents a tremendous potential for more effective administration of dispersed data centres. Dynamic workload allocation and migration across data centres might help to save costs by shifting workload to regions where energy is cleaner or cooling costs are lesser. After obtaining the user's query, the cloud hosting has the option of selecting the target region depending on a variety of factors. The authors in [142] have worked upon the already proposed technique called EcoMultiCloud. They looked at the example of a complex network made up of data centres (DCs) spread around the country, with renewable energy producers co-located with cloud services to reduce the amount of electricity purchased by the power grid. Since renewable energy sources are infrequent, infrastructure load control solutions must be customised to the intermittent nature of the sources. Furthermore, the authors in [143] have addressed the problem of reducing energy costs for geographically distant data centres while maintaining assured service quality (i.e., service latency) under time-varying system dynamics. They proposed a green geographical load balancing (GreenGLB) online solution for interactive and indivisible work distribution based on the greedy algorithm design approach. An indivisible job is something that cannot be divided further and must be allocated to a single data centre.

9.2 E-Waste

The rapid escalation in the use of electronic devices at the consumer level along with the growth of enterprise-class and hyperscale computing has added to the issue of e-waste. It is a combined responsibility of consumers, manufacturers, enterprises and governments to ensure that this waste is being minimized, reused and recycled properly. Waste Electrical and Electronic Component (WEEE) passes regulations at country, state or province level aiming to promote the reuse and recycling of e waste leading to reduction of consumption of such resources and the amount of e-waste going to landfill.

Most common e-waste includes LCD monitors, LCD and Plasma televisions, and computers with Cathode Ray tubes. However, this does not mean that other electrical and electronic equipment does not fall into this category. Actually, any piece of electronic equipment is e-waste. Almost all technology-based industries produce e waste. However, the footprint of the data center in this domain is relatively small. Data center consists of components like generators, Uninterrupted Power Supply (UPS) etc. These components have a long primary lifecycle of about 5–10 years and are also repurposed into non 24 $\times$ 7 roles before they can be recycled.

9.3 Carbon footprint (Greenhouse gas emissions)

Cloud data centers consist of a huge number of rows of electricity consuming servers having network, storage, power supply systems along with gigantic HVAC (heating, ventilation and air conditioning) units that avoid overheating. However, these data centers appear to be clean but they are not contributing to the green initiative.

Total Greenhouse Gas Emissions (GHG) attributed to data centres in 2018 were 3.15 $\times$ 107 tons CO2-eq, accounting for about 0.5% of total GHG emissions in the United States [144]. A little more than half (52%) of total data centre emissions is attributable to the Northeast, Southeast, and Central United States, which have a high concentration of thermoelectric power plants as well as a big number of data centres. Almost 30% of the emissions from the data centre sector occur in the Central United States, which depends significantly on coal and natural gas to satisfy its energy needs.

Looking at these estimates, it becomes quite evident that companies need to reduce GHG emissions. Many companies and countries are already trying in this direction and have introduced improvements to the existing data centers. It is estimated that most of the company's servers are having a utilization of only 10–15% and 30% of these corporate servers are zombies in the sense that they are inactive yet use electricity while doing so (data centres play a key role in reducing GHG emissions) [145].

9.4 Case studies

Zero carbon, zero emissions, and zero waste are the goals of the next generation of sustainable data centers. It all starts with sustainable energy. Solar energy has no carbon impact, generates no pollutants, and may grow fast. Solar-powered data centers are not only environmentally green, but also financially effective. It reduces the energy expense by about 70%, a benefit that is immediately passed on to customers, resulting in significant cost savings [146]. The

relative stability of solar power expenses vs the ever-increasing cost of regular energy provides organisations with significant headroom to engage in innovation. Scalability is also not a problem. The below mentioned case studies uses the renewable energy to power the data centers. The CtrilS data center make use of the solar energy to power the data center and it is the first data center in the world which is awarded with the LEED Platinum' certification from United States Green Building Council whereas the project Natick by Microsoft is the first underwater data center which is 100% powered by solar and wind. The project proves that the underwater data center is feasible as well as logistically, environmentally and economically practical.

9.4.1 CtrilS data center [146]

CtrilS data centre is one of the data centers in Mumbai that has taken initiative in the building of green data centers and has been awarded with the highest rating - platinum - under LEED V4 existing Building (O + M) category. It is the first data centre in the entire world to get this recognition from the U.S Green Building Council (USGBC).

CtrilS has identified and worked upon many key areas like operations, design and technology. Some of them are as:

- Many of the enhancements done like consumption of power by cooling system, enhancing the pumping systems efficiency, cooling towers automation and effectiveness, helped them to gain savings of 15 lakhs kWh/annum which means saving of approximately 1.5 crore rupees on the entire operations.
- Mercury free lights used resulted in saving of 0.5 lakh kWh/annum (cost saving of 5 lakhs/year).
- Proper centralized monitoring system was used to monitor the electricity utilization and report anomalies when there is excessive use.
- Water saving fixtures were used leading to the saving of 5KL of water per day that was about 24% of water reduction in the overall water requirement.
- Rainwater treatment for storing rainwater during the rainy season was also used.
- Green seal products were used for cleaning purposes in order to avoid toxic substances.
- CFC-Free Refrigerants in cooling systems were used to minimize greenhouse gas emissions.
- Strategic ventilation points were introduced in order to address the quality of indoor air and increase in productivity. These were created to get the levels above the baseline of Ashrae Standard 62.1 for the needs of the fresh air.
- Extensive Training Programs were held to aware the employees about the importance of incorporating the green compliance features.

9.4.2 Project natick by microsoft [147]

Nearly 50% of the world's people lives within a few miles of the shore. By locating data centers near coastal cities, data transit times to coastal towns would be reduced, resulting in faster and smoother online surfing, video streaming, and gaming. Furthermore, not only do the obstacles connected with creating data centres on land disappear, but the ocean also delivers a rather stable environment for these underwater pods. Finally, even at intermediate depths of 10–200 m, the water maintains a temperature range of 14–18 degrees Celsius, making it ideal for cooling data centres. As a result, cooling costs are lowered.

Microsoft project Natick is a research project whose main aim is to determine the credibility of the subsea data centers that can be powered by offshore renewable energy. On the seafloor off the coast of Scotland, Microsoft's underwater data centre project used wave energy and post-quantum encryption. The concept of an underwater data center was put forward at Microsoft in the year 2014 during the ThinkWeek that is related to the employees to discuss the extraordinary ideas. The project aims to provide the lightning quick services of the cloud to the coastal population and save energy. The cool surface of the sea enables the energy efficient design of data centers. The experimental green energy solutions under research at the European Marine Energy Centre and the wind and solar energy used to power the grid are some of the main reasons for Northern Isles installation by the Project Natick team at the 'The Orkney Islands'.

9.5 Policies and standards

Energy is one of the most essential components of a data center's running expenses, yet rising energy costs and associated operational costs pose hurdles to corporate competitiveness. As a result, it is vital to minimise energy consumption in data centres, and improved energy efficiency has been identified as an acceptable tool for this aim. There are several policies in place to address energy usage in data centres [148]. Some of these are mentioned below in Table 2 with their key features.

10 Research challenges

RQ4 *What are the major software academic difficulties for developing green data centres?*

Green data centers have been an active area of research and the major challenges to achieve green data center includes the challenges in container technology, VM

Table 2 Key Features of various policies and standards for data centre energy consumption

Policy name	Country	Year	Key features
ENERGY STAR Rating by Environment Protection Agency (EPA)	USA	1992	The ENERGY STAR score for data centers is applicable to facilities that are specially constructed and equipped to fulfil the demands of high-density computer equipment, such as server racks, which are utilised for data storage and processing. The goal of the ENERGY STAR score is to offer a fair assessment of a facility's power efficiency in compared to peers, while considering climate, weather, and business operations at the location into account. It entails measuring and optimizing energy usage on a consistent basis.
American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE)	USA	1895	This standard establishes the minimal energy saving requirements for structures such as data centres and telecommunications buildings, in terms of architecture, development, service, and maintenance, as well as the use of on-site or off-site renewable energy supplies.
National Australian Built Environment Rating System (NABERS)	Australia	1998	This policy measures the performance related to the environment on a scale of 1 to 6 stars. The three categories decide the final rating.
Green Mark	Singapore	2005	This rating system is based in Singapore that gives points features like efficiency of energy and best practices.
Green Building Index (GBI)	Malaysia	2009	This is based on current ranking tools, such as the Green Mark, but it has been heavily updated to account for Malaysia's tropical weather, environmental background, and cultural and social needs.
Certified Energy Efficient data center Audit (CEEDA)	United Kingdom	NR	This rating system is based in UK that assesses the best practices that should be followed for building energy efficient data centers.
Building Research Establishment Environmental Assessment for Data Centers (BREAM)	United Kingdom	1990	This rating system is based in the UK which is also followed by Hong Kong. It is based on the data center score across the ten categories.
European Code of Conduct by the European Commission	Europe	2008	This aids in goal planning, monitoring energy use, creating attention, and promoting energy-efficient best practises to improve data center energy quality.
Blue Angel Eco-Label	Germany	1978	This norm is provided to every environmentally aware organization that is committed to implementing a long-term strategy to improve the resource and energy usage of its data centre in relation to the IT services to be distributed, as well as performing daily testing to improve data centre operations.
International Standards Organization 50001:2011	Europe	2011	This international standard outlines the Energy Management Systems (EMS) standards that data centers must meet in order to formulate and enforce an energy strategy, as well as set goals, priorities, and action plans that take legal requirements and facts into account.

migration, virtualization, load balancing and workload categorization.

Table 3 shows the research challenges in containers.

Although operating system virtualization is becoming more popular, it commonly adds layers of inference and abstraction beneath the application code to an already intensively layered software stack that contains assistance for legacy physical protocols and unimportant enhancements. If the current rate of layer inclusion persists (due to the distributed presence of logic across various software elements written in different languages), the problem will worsen in the coming years, as future software developers will have to dig through multitudes of layers to debug even the relatively simple applications. Developing a security-aware scheduler to mitigate security concerns associated with containers when distributed across cloud architectures might be an intriguing future research topic.

Table 4 shows the research challenges in the domain of VM Migration.

Another critical area of research is VM aggregation and resource redistribution via VM migrations, with an emphasis on both power monitoring and network overhead. Considerations on VM placement that are purely focused on increasing server resource usage and lowering power usage may result in data centre designs that are neither traffic-aware nor network efficient, resulting to more SLA breaches. Consequently, VM allocation techniques that take into account both VM resource demands and inter-VM traffic load may be able to make more precise and economical placement decisions [151].

Table 3 Research challenges in containers

Domain	Problems & Solutions
Containers	Containers pose a number of security threats like attack through the untrusted images. These images may contain backdoors. In addition, the images may have configuration defects that may provide them unnecessary privileges that will allow the application to have control over the container. The DoS attack on the containers can also create a serious security threat in which a vulnerable application attacks the neighbouring containers and uses the large number of resources that affects the performance of the container.
	Using the trusted images and verifying the images with the signatures is one such solution available to counter security threats. Periodic scanning of applications/ images and running applications with minimum privileges are some other measures that can be taken.
	The industry requires a comprehensive container orchestration architecture. Aside from those, there are a few other issues to consider when it comes to container deployment and orchestration: During stateful container deployment and orchestration, dynamic life-cycle management is a major concern. This is because stateful distributed services often have many dependencies and dynamic implementation stages (for example: Kubernetes deployment has 21 steps). Furthermore, orchestration operators must handle the orchestrator in real time, as well as deal with numerous updates, crashes, interface adjustments, scaling, and other issues. Because of the following basic issues, multitenancy becomes even more difficult when multiple tenants with their own conditions, isolation, and protection are included. It is critical to keep an eye on/resist any single tenant from monopolizing a cluster's services.
	Advanced automation and orchestration systems are the most significant technologies for aiding IT professionals in operating large-scale cloud data centres. Many off-the-shelf solutions are already accessible as open-source software, but it has been seen that almost every organisation is creating its own bespoke orchestration solution to meet its specific use cases. Because of this strategy, the container orchestrator market is extremely fragmented, making it difficult for new small IT businesses to implement them in production.
	The container networking poses a number of problems. In the containerized environment, there is no static IP as containers spin up and down continuously. Traditional IP networking is inefficient and difficult to automate. Static addresses need to be manually configured. Use of DHCP or anything similar must be done in order to address the issues automatically but in this case, service discovery and address assignment might take several seconds or even minutes. That is insufficient in a containerized environment that is constantly changing. Overlay networks, service discovery, IDs, and labels enable traffic to be routed across containerized networks without the need for traditional IP networking

When apps like CPU and memory intensive are co-located on the same physical server, resource competition for some capabilities may occur, while others may be underutilised. Furthermore, such resource contention will have a significant impact on application performance, resulting in SLA breaches and economic loss. As a result, it is critical to comprehend the host's behaviour and resource utilisation patterns. Overhead for VM relocation and reconfiguration might have a negative impact on scalability and data centre bandwidth consumption, as well as performance of the application. Because of this, VM strategies for placement and scheduling that are uninformed about VM migration and reconfiguration overhead may significantly clog the network, induce unnoticed SLA breaches. Incorporation considering the expected migration overhead using placement techniques and VM placement optimization and migration by balancing utilisation in terms of network resources, migration overhead, and so on, aspects of energy consumption are still to be explored.

Table 5 shows the research challenges in virtualization.The entire virtualization process increases the total complexity of the machine, making testing, compilation, and management difficult. This may be avoided by running the emulated server solution on a desktop virtual machine and utilising the host as the client. As there are many VMs in virtualization, proper isolation is critical for security. The isolation approach is commonly used in huge software stacks. Under the direct supervision of cloud providers, it may be vulnerable to VMM vulnerability exploitation and insider assaults. Utilization of verified thin virtualization instead of employing a conventional framework, a Trusted Computing Base (TCB) can solve this challenge.

Table 6 shows the research challenges in the domain of load balancing.

Due to the sheer exponential development in demand for cloud services, efficient utilisation of energy and processing resources has become a major challenge. Load balancing improves resource efficiency, quality, and energy savings by optimally distributing the load in the data center among diverse computing units. It has been observed that the methods under consideration frequently function to improve QoS, resource consumption, and energy conservation. Current LB algorithms have a number of flaws, such as energy and resource waste, inadequate frequencies management, and static impediments. As a result, there is a lot of space for growth. More efficient and adaptable LB algorithms should be developed to offer clients with excellent services at the lowest possible cost in order to

Table 4 Research challenges in the domain of VM Migration

Domain	Problems & Solutions
VM Migration	One of the important problems is to enhance the efficiency and effectiveness of the VM migration because it is widely used in the enterprise environments. Most of the work done in this domain is application-oblivious. It could be possible to design more flexible migration strategies that reduce migration costs by disclosing the application's characteristics and efficiency goals. Research should be done for the different types of resources and the different types of the workloads.
	Virtualized servers are intended to separate the underlying hardware from the applications that run on it. Workload migration is made possible by this concept. Although hardware interruptions between the source and destination servers are uncommon, they can inhibit a successful transfer.
	It is critical to begin troubleshooting by assessing the server hardware and configuration. For example, in order for a workload transfer to be successful, the source and destination servers must have the same CPU.
	One of the main issues that the virtualization architecture must overcome is security. While Xen lacks adequate protection countermeasures against tampering and exploitation threats, several techniques exist, that can be used to mitigate these problems [149]. To provide a more stable platform for supporting VM migration, more sophisticated and collaborative defences are needed.
	50% of all x86 data center workloads are in the form of virtual machine (Gartner research) [150], and this percentage will continue to rise in the coming years. This would result in a growth in the number of virtual machines (VMs) housed in data centers, which will increase operational complexity. In order to identify reasonable trade-offs between costs and advantages of VM migration, multiple competing goals and criteria related to efficiency and energy usage should be taken into account. Finding a flexible approach for tracking and determining appropriate policies to guide migration decisions is an issue that should be overcome. To resolve the issues of complexity and scalability, automated VM migration management techniques must be specified. Large-scale VM implementation necessitates the creation of sophisticated migration management techniques.
	If a task lacks the appropriate computing capacity, it cannot be migrated to a destination server. When the destination node does not have enough CPU cores, memory space, or NIC ports, or when storage is limited, and cannot reserve resources for the new demand, migration issues might arise. This is becoming a more prevalent issue as physical server numbers are decreasing and workload consolidation levels are rising.
	Workload migration between physical servers is a necessary element of a virtualized system, but the procedure is laden with potential pitfalls. Factors such as hypervisor flaws, migration settings, unanticipated hardware requirements, network connectivity difficulties and configuration errors, resource limitations, and SAN configurations can all work together to hinder effective workload transfers.

Table 5 Research challenges in virtualization

Domain	Problems & Solutions
Virtualization	Virtualization consolidates multiple clients on a single physical machine making it a Single Point of Failure (SPOF). However, along with this, the hypervisor supporting these multiple clients also becomes a SPOF. A bug in the hypervisor may affect some or all of the clients. To ensure that individual component failures do not result in service loss, highly available systems are configured without Single Points of Failure (SPOF). The basic approach to avoiding SPOFs is to offer duplicate components for each required resource, allowing service to continue even if a component fails.
	As multiple clients are on the same machine, side channel attacks are common between VMs. Attacks can be performed to extract sensitive information. Performance degradation by resource exhaustion is some of the common security attacks seen in this domain. Many solutions have been provided which can work at hardware or application level. Mitigation at the hypervisor level, on the other hand, may be more advantageous because it covers all clients and needs no interaction between them.
	The capacity to create as many virtual machines as needed may result in the development of more VMs than the business requires. VM sprawl may appear to be innocuous, but it can worsen resource distribution issues by diverting resources to VMs that aren't even being utilised, while those that are being used and required suffer from reduced functionality. Businesses may avoid VM sprawl by restricting the number of VMs that are genuinely needed and adding more as needed.
	Nested Virtualization refers to running a hypervisor inside another. It finds many applications in the domain of testing, security and fault tolerance. However, many of the hypervisors do not support this feature. The hypervisors face stability and performance issues while supporting nested virtualization.
	Congestion is a common and well-known problem with VMs. Prior to virtualization, a standalone executable on a central computer would often only consume a small portion of the computer's bandwidth usage. However, several VMs on a virtualized server, each requiring network bandwidth, gradually exceeds the virtual environment's NIC port (typically one in a server). One way to resolve the congestion issue is by adding more NIC's to the VM server. Another is by utilizing solutions that balance VMs across multiple servers. One such example being VMware DRS cluster of ESXi hosts.

optimise resource efficiency, energy conservation, and production. Adaptive LB will enable for traffic control between quick activities, efficient resource consumption, and a combination of centralised and distributed control mechanisms. Energy conservation is a key aspect in promoting economic growth in situations when greater resource demand results from decreased resource acquisition.

Table 7 shows the research challenges in workload categorization.

Workload characterisation is a trivial research issue due to the prevalence of workloads with distinctive, nontrivial, and poorly understood qualities. The workloads related to fog and vehicular clouds, big data, anonymous social networks, and sensor networks, to mention a few, are addressed in open research problems.

The practise of safeguarding workloads that migrate among cloud environments is known as cloud workload protection. For a cloud-based application to function correctly and without posing security problems, the entire workload must be operational. As a result, cloud workload security and app service workload security are essentially different from desktop application safety. Workload safety is especially challenging in hybrid data centre designs that use any from physical on-premises desktops to various public cloud infrastructure as a service (IaaS) setting to container-based application architectures. Cloud workload security is particularly difficult because, when workloads travel between providers and hosts, responsibility for securing the workload must be pooled. One of the significant future research objectives is the integration of auto-scaling techniques with IDS and IPS systems to better manage Distributed Denial-of-Service (DDoS) and Yo–Yo attacks. To deal with malicious behaviour, auto scaling systems often transform DDoS attacks to Economic Denial of Sustainability (EDoS) attacks. Recognizing DDoS workload from user workload is an unresolved topic that should be addressed in future study.

Table 6 Research challenges in the domain of load balancing

Domain	Problems & Solutions
Load balancing	One of the most difficult challenges in the load-balancing algorithm is determining which technique is utilised to assess the workload of a certain node. The total number of processes operating on computer are used as a measurement of workload in the majority of load balancing algorithms. Load depends on many parameters like: no of processes, demands of the processes running, instruction mix, architecture/speed of the processor, CPU utilization etc. Research should focus on all the parameters before proceeding in the direction of designing and developing load-balancing algorithms. Some other factors that play an important role in research in this domain include location policy for load-balancing algorithms (Threshold based, bidding based, pairing etc.), process transfer policy for load-balancing algorithms (static threshold vs. dynamic threshold) etc.
	Many load balancing algorithms targeting a single and multiple objectives have been proposed by researchers. However, many of these do not consider algorithm complexity. Similarly, the load balancing process in data centers requires VM migrations, where migration cost is involved. Migration cost has not been considered as an important metric in many proposed load-balancing algorithms. The following metrics have been considered more frequently: latency, resource efficiency, computation time, interoperability, and communication cost. The less considered metrics include: throughput, overhead, fault tolerance, degree of balance, migration time etc.

Table 7 Research challenges in workload categorization

Domain	Problems & Solutions
Workload categorization	Workload is categorised based on many factors such as submitting time, completion time, inter-session interval, and so on, however there are several types of characteristics that need to be addressed during workload classification.
	Due to the huge magnitude and complexity of the workload, in-depth statistical analysis and workload classification inside a large-scale production cloud is difficult.
	Due to the behavioural characteristics of workload in the setting of cloud computing, there is a shortage of techniques to define the workload.

11 Conclusion

Data centres' energy usage is becoming a significant problem. Advances in server equipment technologies and growing demand for processing power have resulted in higher workload and hence has resulted in higher data centre power usage. Data centers are the core of today's Internet and cloud computing networks. Data centers' growing demand for electrical energy necessitates accounting for the immense amount of energy they consume. In this context, data center energy modelling and prediction are critical. This survey article conducted a thorough examination of the present status of software solutions that aids in data centre power consumption reduction. The paper is divided into five separate sections. First, approaches for applying software virtualization are described, followed by ways for applying operating system virtualization. Furthermore, methodologies used in data centres as well as other problem-solving approaches for energy saving have been examined. Environmental factors have also been studied in order to reduce electricity use. The article also covers the significance of containerization in decreasing data centre power consumption and concludes with research challenges for sustainable data centre building. The container technology is rising and requires more research in the domain of energy efficiency. Containers, when properly built, allow a host to utilise nearly all available resources. Isolated containers can function independently of other containers, allowing a single host to perform many functions. The future work includes development of an efficient technique for placement of containers/tasks onto physical machines considering: CPU multicores, memory, storage & network together. Container migration technique will also be developed to reduce energy usage while preserving needed service quality (QoS).

Author contributions All authors contributed to the study conception and design. Avita Katal, Susheela Dahiya and Tanupriya Choudhury performed material preparation, data collection and analysis. Avita Katal wrote the first draft of the manuscript and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding No funding was received to assist with the preparation of this manuscript.

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Code Availability Not Applicable.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethical approval This paper does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Fiona, B., Ballarat, C.: International Review of Energy Efficiency in Data Centres Acknowledgements. (2021)
2. Koot, M., Wijnhoven, F.: Usage impact on data center electricity needs: A system dynamic forecasting model. Appl. Energy. **291**, 116798 (2021)
3. Analysts, G.I.: I. *Internet Data Centers - Global Market Trajectory & Analytics*. (2021)
4. Chester, S.: What Is Power Usage Effectiveness (PUE)? (2019). https://www.colocationamerica.com/blog/what-is-pue
5. The future of: data center power consumption – 5 essential facts | Danfoss. https://www.danfoss.com/en/about-danfoss/insights-for-tomorrow/integrated-energy-systems/data-center-power-consumption/
6. US20080086731A1 - Method: and system for managing resources in a data center - Google Patents. https://patents.google.com/patent/US20080086731
7. Kliazovich, D., Bouvry, P., Khan, S.U. DENS: Data center energy-efficient network-aware scheduling. *Proceedings – 2010 IEEE/ACM International Conference on Green Computing and Communications, GreenCom* 2010 IEEE/ACM International Conference on Cyber, Physical and Social Computing, CPSCom 2010 69–75 (2010) https://doi.org/10.1109/GREENCOM-CPSCOM.2010.31
8. Heller, B., et al. ElasticTree: Saving Energy in Data Center Networks. *IN NSDI* (2010)
9. Khargharia, B., et al. Autonomic power & performance management for large-scale data centers. *Proceedings – 21st International Parallel and Distributed Processing Symposium, IPDPS 2007; Abstracts and CD-ROM* (2007). https://doi.org/10.1109/IPDPS.2007.370510
10. SRCMap: Energy Proportional Storage Using Dynamic Consolidation. https://www.researchgate.net/publication/221353706_SRCMap_Energy_Proportional_Storage_Using_Dynamic_Consolidation
11. Baliga, J., Ayre, R.W.A., Hinton, K., Tucker, R.S. Green cloud computing: Balancing energy in processing, storage, and transport. *Proceedings of the IEEE* **99**, 149–167 (2011)
12. Beloglazov, A., Buyya, R., Lee, Y.C., Zomaya, A.A.: Taxonomy and survey of energy-efficient data centers and cloud computing systems. Adv. Computers **82**, 47–111 (2011)
13. Shuja, J., et al.: Survey of techniques and architectures for designing energy-efficient data centers. IEEE Syst. J. **10**, 507–519 (2016)
14. Dayarathna, M., Wen, Y., Fan, R.: Data center energy consumption modeling: A survey. IEEE Commun. Surv. Tutorials. **18**, 732–794 (2016)

15. You, X., Li, Y., Zheng, M., Zhu, C., Yu, L.: A survey and taxonomy of energy efficiency relevant surveys in cloud-related environments. IEEE Access. **5**, 14066–14078 (2017)
16. Katal, A., Dahiya, S., Choudhury, T.: Energy efficiency in cloud computing data center: a survey on hardware technologies. Cluster Comput. **2021**, 1–31 (2021). https://doi.org/10.1007/S10586-021-03431-Z
17. LiTao, Kurian, J.: Run-time modeling and estimation of operating system power consumption. ACM SIGMETRICS Performance Evaluation Review. **31**, 160–171 (2003)
18. Herzog, B., Hügel, F., Reif, S., Hönig, T., Schröder-Preikschat, W.: Automated selection of energy-efficient operating system configurations. Energy (2021). https://doi.org/10.1145/3447555.3465327
19. Scordino, C., Abeni, L., Lelli, J.: Energy-aware real-time scheduling in the linux kernel. Proc. ACM Sympos. Appl. Comput. (2018). https://doi.org/10.1145/3167132.3167198
20. Embedded Data Centers: | Products | ENERGY STAR. https://www.energystar.gov/products/office_equipment/data_center_storage/data_center_energy_efficiency/embedded_data_centers
21. BuschhoffMarkus, F.R., SpinczykOlaf: Energy-aware device drivers for embedded operating systems. ACM SIGBED Review. **16**, 8–13 (2019)
22. Levy, A., et al. Multiprogramming a 64 kB Computer Safely and Efficiently. Proceedings of the 26th Symposium on Operating Systems Principles (2017) https://doi.org/10.1145/3132747
23. Kang, D.G.I.S.T., Alian, K.-D., Kim, M., Huh, D.G.I.S.T.D. KAIST, J. & Sung Kim, N. VIP: Virtual Performance-State for Efficient Power Man-agement of Virtual Machines. *Proceedings of the ACM Symposium on Cloud Computing '18* (2021)
24. Xiao, P., Ni, Z., Liu, D., Hu, Z.: Improving the energy-efficiency of virtual machines by I/O compensation. J. Supercomputing. **77**, 11135–11159 (2021)
25. Prabhakaran, G., Selvakumar, S.: An diverse approach on virtual machines administration and power control in multi-level implicit servers. J. Ambient Intell. Humaniz. Comput. (2021). https://doi.org/10.1007/S12652-021-03013-2
26. Ho, T.T.N., Gribaudo, M., Pernici, B.: Characterizing Energy per Job in Cloud Applications. Electron. 2016. **5**, 90 (2016)
27. Kumar, S., Buyya, R.: Green cloud computing and environmental sustainability harnessing green. Principles Practices (2012). https://doi.org/10.1002/9781118305393.CH16
28. Tchana, A., et al. Software consolidation as an efficient energy and cost saving solution for a SaaS/PaaS cloud model. Lecture Notes Comput. Sci. **9233**, 305–316 (2015)
29. Samrajesh, M.D., Gopalan, N.P. Component based energy aware multi-tenant application in software as-a service. *15th International Conference on Advanced Computing Technologies, ICACT 2013* (2013). https://doi.org/10.1109/ICACT.2013.6710502
30. Czarnul, P., Proficz, J., Krzywaniak, A. Energy-Aware High-Performance Computing: Survey of State-of-the-Art Tools, Techniques, and Environments. *Scientific Programming* (2019)
31. Ho, T.T.N., Pernici, B.: A data-value-driven adaptation framework for energy efficiency for data intensive applications in clouds. IEEE Conf. Technol. Sustainabil. (2015). https://doi.org/10.1109/SUSTECH.2015.7314320
32. Malik, M., et al. ECoST: Energy-efficient co-locating and self-tuning mapreduce applications. *ACM International Conference Proceeding Series* (2019). https://doi.org/10.1145/3337821.3337834
33. Miyazaki, T. Bayesian Optimization of HPC Systems for Energy Efficiency. Lecture Notes Comput. Sci. **10876**: 44–62 (2018)
34. Reddy Basireddy, K., Wachter, E.W., Al-Hashimi, B.M., Merrett, G. Workload-Aware runtime energy management for HPC Systems. *Proceedings – 2018 International Conference on High Performance Computing and Simulation, HPCS* 292–299 (2018) https://doi.org/10.1109/HPCS.2018.00057
35. Tiwari, N., Bellur, U., Sarkar, S., Indrawan, M.: Optimizing MapReduce for energy efficiency. Software: Pract. Experience. **48**, 1660–1687 (2018)
36. Jiang, D., Wang, Y., Lv, Z., Wang, W., Wang, H.: An Energy-Efficient Networking Approach in Cloud Services for IIoT Networks. IEEE J. Sel. Areas Commun. **38**, 928–941 (2020)
37. Vakilinia, S.: Energy efficient temporal load aware resource allocation in cloud computing datacenters. J. Cloud Comput. **7**, 1–24 (2018)
38. Barrett, D., Kipper, G.: How virtualization happens. Virtualiz. Forensics. (2010). https://doi.org/10.1016/B978-1-59749-557-8.00001-1
39. Cuadrado-Cordero, I., Orgerie, A.C., Menaud, J.M. Comparative experimental analysis of the quality-of-service and energy-efficiency of VMs and containers' consolidation for cloud applications. *25th International Conference on Software, Telecommunications and Computer Networks, SoftCOM 2017* (2017) https://doi.org/10.23919/SOFTCOM.2017.8115516
40. Huang, D., Wu, H.: Virtualization. Mob. Cloud Comput. (2018). https://doi.org/10.1016/B978-0-12-809641-3.00003-X
41. Ramchandra Desai, P.A. Survey of Performance Comparison between Virtual Machines and Containers. Int. J. Comput. Sci. Eng. (2016)
42. Piraghaj, S.F., Dastjerdi, A.V., Calheiros, R.N., Buyya, R.A. Framework and Algorithm for Energy Efficient Container Consolidation in Cloud Data Centers. *Proceedings – 2015 IEEE International Conference on Data Science and Data Intensive Systems; 8th IEEE International Conference Cyber, Physical and Social Computing; 11th IEEE International Conference on Green Computing and Communications and 8th IEEE Inte* 368–375 (2015). https://doi.org/10.1109/DSDIS.2015.67
43. Beloglazov, A., Buyya, R.: Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers. Concurrency Comput. Pract. Experience. **24**, 1397–1420 (2012)
44. Nath, S.B., Addya, S.K., Chakraborty, S., Ghosh, S.K. Green Containerized Service Consolidation in Cloud. IEEE International Conference on Communications (2020)
45. Ferdaus, M.H., Murshed, M., Calheiros, R.N., Buyya, R. Virtual machine consolidation in cloud data centers using ACO metaheuristic. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **8632**, 306–317 (2014)
46. Shi, T., Ma, H., Chen, G. Energy-Aware Container Consolidation Based on PSO in Cloud Data Centers. IEEE Congress on Evolutionary Computation, CEC 2018 - Proceedings (2018). https://doi.org/10.1109/CEC.2018.8477708
47. Tan, B., Ma, H., Mei, Y.A., Hybrid Genetic Programming Hyper-Heuristic Approach for Online Two-level Resource Allocation in Container-based Clouds. 2019 IEEE Congress on Evolutionary Computation, CEC 2019 - Proceedings 2681–2688 (2019). https://doi.org/10.1109/CEC.2019.8790220
48. Fan, X., Weber, W.D., Barroso, L.A. Power provisioning for a warehouse-sized computer. Proceedings - International Symposium on Computer Architecture 13–23 (2007). https://doi.org/10.1145/1250662.1250665
49. Chen, F., Zhou, X., Shi, C. The container deployment strategy based on stable matching. IEEE 4th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA **2019** 215–221 (2019). https://doi.org/10.1109/ICCCBDA.2019.8725707

50. Al-Moalmi, A., Luo, J., Salah, A., Li, K., Yin, L.: A whale optimization system for energy-efficient container placement in data centers. Expert Syst. Appl. **164**, 113719 (2021)
51. Raïs, I., Orgerie, A.-C., Quinson, M., Lefèvre, L.: Quantifying the impact of shutdown techniques for energy-efficient data centers. Concurrency and Computation: Practice and Experience. **30**, e4471 (2018)
52. Benoit, A., Lefèvre, L., Orgerie, A.-C., Raïs, I. Reducing the energy consumption of large-scale computing systems through combined shutdown policies with multiple constraints (2017). https://doi.org/10.1177/1094342017714530
53. Marotta, A., Avallone, S., Kassler, A.A.: Joint power efficient server and network consolidation approach for virtualized data centers. Comput. Netw. **130**, 65–80 (2018)
54. Marahatta, A., et al.: Classification-based and energy-efficient dynamic task scheduling scheme for virtualized cloud data center. IEEE Trans. Cloud Comput. **1–1** (2019). https://doi.org/10.1109/TCC.2019.2918226
55. Cioara, T., Anghel, I., Salomie, I.: Methodology for energy aware adaptive management of virtualized data centers. Energ. Effi. **10**, 475–498 (2017)
56. Hussain, M., et al.: Energy and performance-efficient task scheduling in heterogeneous virtualized cloud computing. Sustainable Computing: Informatics and Systems. **30**, 100517 (2021)
57. Shukla, R., Gupta, R.K., Kashyap, R.A.: Multiphase pre-copy strategy for the virtual machine migration in cloud. Smart Innov. Syst. Technol. **104**, 437–446 (2019)
58. Jalaei, N., Safi-Esfahani, F. VCSP: virtual CPU scheduling for post-copy live migration of virtual machines. International Journal of Information Technology 2020 13:1 **13**, 239–250 (2020)
59. Kaur, R.A.: Hybrid approach for virtual machine migration in cloud computing environment. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **7**, 30 (2017)
60. Hines, M.R., Gopalan, K. Post-copy based live virtual machine migration using pre-paging and dynamic self-ballooning. Proceedings of the 2009 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments, VEE'09 51–60 (2009). https://doi.org/10.1145/1508293.1508301
61. Nashaat, H., Ashry, N., Rizk, R. Smart elastic scheduling algorithm for virtual machine migration in cloud computing. The Journal of Supercomputing **75**, 3842–3865 (2019)
62. Liu, H., Jin, H., Liao, X., Yu, C., Xu, C.Z.: Live virtual machine migration via asynchronous replication and state synchronization. IEEE Trans. Parallel Distrib. Syst. **22**, 1986–1999 (2011)
63. Celesti, A., Tusa, F., Villari, M., Puliafito, A. Improving virtual machine migration in federated cloud environments. Proceedings – 2nd International Conference on Evolving Internet, Internet 1st International Conference on Access Networks, Services and Technologies, Access 2010 61–67 (2010). https://doi.org/10.1109/INTERNET.2010.20
64. Bloch, T., Sridaran, R., Prashanth, C.: Understanding Live Migration Techniques Intended for Resource Interference Minimization in Virtualized Cloud Environment. Adv. Intell. Syst. Comput. **654**, 487–497 (2018)
65. Kella, A., Belalem, G.: A stable matching algorithm for VM migration to improve energy consumption and QOS in cloud infrastructures. Cloud Technology: Concepts, Methodologies, Tools, and Applications. **2**, 606–623 (2014)
66. Hu, B., Lei, Z., Lei, Y., Xu, D., Li, J. A time-series based precopy approach for live migration of virtual machines. Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS 947–952 (2011). https://doi.org/10.1109/ICPADS.2011.19
67. Ruchi, T. & Avita Katal. An Optimized Time Series based Two Phase Strategy Pre-Copy Algorithm for Live Virtual Machine Migration. Internat. J. Eng. Res. V6, (2017)
68. Chashoo, S.F., Malhotra, D. VM-Mig-framework: Virtual machine migration with and without ballooning. PDGC 2018–2018 5th International Conference on Parallel, Distributed and Grid Computing 368–373 (2018). https://doi.org/10.1109/PDGC.2018.8745993
69. Sagana, C., Geetha, M., Suganthe, R.C. Performance enhancement in live migration for cloud computing environments. Int. Conf. Informat. Commun. Embedded Syst, ICICES 2013 361–366 (2013). https://doi.org/10.1109/ICICES.2013.6508339
70. Rajapackiyam, E., Subramanian, A.V., Arumugam, U.: Commons Attribution (CC-BY) 3.0 license. J. Comput. Sci. (2020). https://doi.org/10.3844/jcssp.2020.543.550
71. Patel, M., Chaudhary, S., Garg, S. Machine learning based statistical prediction model for improving performance of live virtual machine migration. J. Eng. (United Kingdom) (2016)
72. Tseng, F.H., Chen, X., Chou, L., Chao, H.C., Chen, S.: Support vector machine approach for virtual machine migration in cloud data center. Multimedia Tools Appl. **74**, 3419–3440 (2015)
73. Jo, C., Cho, Y., Egger, B. A machine learning approach to live migration modeling. **14**, (2017)
74. Kansal, N.J., Chana, I.: Energy-aware virtual machine migration for cloud computing a firefly optimization approach. J. Grid Comput. **14**, 327–345 (2016)
75. Rodrigues, T.G., Suto, K., Nishiyama, H., Kato, N. A PSO model with VM migration and transmission power control for low Service Delay in the multiple cloudlets ECC scenario. *IEEE International Conference on Communications* (2017). https://doi.org/10.1109/ICC.2017.7996358
76. Hossain, M.K., Rahman, M., Hossain, A., Rahman, S.Y., Islam, M.M. Active Idle Virtual Machine Migration Algorithm-a new Ant Colony Optimization approach to consolidate Virtual Machines and ensure Green Cloud Computing. *ETCCE - International Conference on Emerging Technology in Computing, Communication and Electronics* (2020). https://doi.org/10.1109/ETCCE51779.2020.9350915
77. ZhengQinghua, et al.: Virtual machine consolidated placement based on multi-objective biogeography-based optimization. Future Generation Comput. Sys. **54**, 95–122 (2016)
78. Sha, J., et al.: A method for virtual machine migration in cloud computing using a collective behavior-based metaheuristics algorithm. Concurr. Comput. **32**, e5441 (2020)
79. Ghosh, S., Banerjee, C. Dynamic time quantum priority based round robin for load balancing in cloud environment. Proceedings – 2018 4th IEEE International Conference on Research in Computational Intelligence and Communication Networks, ICRCICN 33–37 (2018) https://doi.org/10.1109/ICRCICN.2018.8718694
80. Falisha, I.N., Purboyo, T.W., Latuconsina, R.: Experimental model for load balancing in cloud computing using equally spread current execution load algorithm. Int. J. Appl. Eng. Res. **13**, 1134–1138 (2018)
81. Patel, D., Rajawat, A. Efficient Throttled Load Balancing Algorithm in Cloud Environment.International Journal of Modern Trends in Engineering and Research (2015)
82. Manakattu, S.S., Kumar, S.D.M. An improved biased random sampling algorithm for load balancing in cloud based systems. ACM International Conference Proceeding Series 459–462 (2012). https://doi.org/10.1145/2345396.2345472
83. Chen, H., Wang, F., Helian, N., Akanmu, G. User-priority guided min-min scheduling algorithm for load balancing in cloud computing. National Conference on Parallel Computing Technologies, PARCOMPTECH (2013). https://doi.org/10.1109/PARCOMPTECH.2013.6621389

84. Hung, T.C., Hy, P.T., Hieu, L.N., Phi, N.X. MMSIA: Improved max-min scheduling algorithm for load balancing on cloud computing. ACM International Conference Proceeding Series 60–64 (2019). https://doi.org/10.1145/3310986.3311017
85. Ananthakrishnan, B. An-Efficient-Approach-for-Load-Balancing-in-Cloud-Environment.doc. Int. J. Sci. Eng. Res. **6**, (2015)
86. Banerjee, A., Chatterjee, G., Chakraborty, D., Majumder, S.: Cluster based intelligent load balancing algorithm applied in cloud computing using KNN. SSRN Electron. J. (2019). https://doi.org/10.2139/SSRN.3503518
87. Kaur, A., Kaur, B., Singh, P., Devgan, M.S., Toor, H.K.: Load balancing optimization based on deep learning approach in cloud environment. Int. J. Inform. Technol. Comput. Sci. **12**, 8–18 (2020)
88. Chen, J. Machine learning for load balancing in the linux kernel. Proceedings of the 11th ACM SIGOPS Asia-Pacific Workshop on Systems **20**
89. Mondal, B., Choudhury, A. Simulated annealing (SA) based load balancing strategy for cloud computing. Int. J. Comput. Sci. Informat. Technologies (2015)
90. Singhal, U., Jain, S.: An analysis of swarm intelligence based load balancing algorithms in a cloud computing environment. Int. J. Hybrid Inform. Technol. **8**, 249–256 (2015)
91. Gupta, A., Garg, R. Load Balancing Based Task Scheduling with ACO in Cloud Computing. International Conference on Computer and Applications, ICCA 2017 174–179 (2017) doi: (2017). https://doi.org/10.1109/COMAPP.2017.8079781
92. Acharya, J., Mehta, M., Saini, B. Particle swarm optimization based load balancing in cloud computing. *Proceedings of the International Conference on Communication and Electronics Systems, ICCES* (2016) doi: (2016). https://doi.org/10.1109/CESYS.2016.7889943
93. Ullah, A., Nawi, N.M., Uddin, J., Baseer, S., Rashed, A.H.: Artificial bee colony algorithm used for load balancing in cloud computing: review. IAES Int. J. Artif. Intell. (IJ-AI). **8**, 156–167 (2019)
94. Jena, U.K., Das, P.K., Kabat, M.R.: Hybridization of meta-heuristic algorithm for load balancing in cloud computing environment. J. King Saud Univ. - Comput. Inform. Sci. (2020). doi:https://doi.org/10.1016/J.JKSUCI.2020.01.012
95. Sharma, S., Luhach, A., Kr, Sheik Abdhullah, S. An Optimal Load Balancing Technique for Cloud Computing Environment using Bat Algorithm.Indian Journal of Science and Technology**9**, (2016)
96. Crow Search based Scheduling Algorithm for Load Balancing in Cloud Environment: Int. J. Innovative Technol. Exploring Eng. **8**, 1058–1064 (2019)
97. Wang, Q., Liu, D. Research on Load Balancing Method in Cloud Computing. *Proceedings of IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2018* 1489–1493 (2018) doi: (2018). https://doi.org/10.1109/IAEAC.2018.8577591
98. Hashem, W., Nashaat, H., Rizk, R.: Honey bee based load balancing in cloud computing. KSII Trans. Internet Inf. Syst. **11**, 5694–5711 (2017)
99. Makasarwala, H.A., Hazari, P. Using genetic algorithm for load balancing in cloud computing. *Proceedings of the 8th International Conference on Electronics, Computers and Artificial* Intelligence, *ECAI* (2017) doi: (2016). https://doi.org/10.1109/ECAI.2016.7861166
100. Derdus, K.M., Omwenga, V., Ogao, P. Statistical Techniques for Characterizing Cloud Workloads: A Survey.International Journal of Computer and Information Technology2279–0764(2019)
101. Ismaeel, S., Al-Khazraji, A., Miri, A. An efficient workload clustering framework for large-scale data centers. *8th International Conference on Modeling Simulation and Applied* Optimization, *ICMSAO 2019* (2019) doi: (2019). https://doi.org/10.1109/ICMSAO.2019.8880305
102. Yousif, S.A., Al-Dulaimy, A. Clustering Cloud Workload Traces to Improve the Performance of Cloud Data Centers. *Proceedings of the World Congress on Engineering* (2017)
103. Zhao, X., Yin, J., Chen, Z., He, S. Workload classification model for specializing virtual machine operating system. *IEEE International Conference on Cloud Computing, CLOUD* 343–350 doi: (2013). https://doi.org/10.1109/CLOUD.2013.144
104. Li, S., Ben-Nun, T., Girolamo, S., di, Alistarh, D., Hoefler, T. Taming Unbalanced Training Workloads in Deep Learning with Partial Collective Operations. (2020)
105. Mathematics, K.K.-T.J. of C. and & undefined. Forecasting of Cloud Computing Services Workload using Machine Learning. *turcomat.org* **12**, 4841–4846 (2021). (2021)
106. Cetinski, K., Juric, M.B.: AME-WPC: Advanced model for efficient workload prediction in the cloud. J. Netw. Comput. Appl. **55**, 191–201 (2015)
107. Shekhawat, V.S., Gautam, A., Thakrar, A. Datacenter Workload Classification and Characterization: An Empirical Approach. *13th International Conference on Industrial and Information Systems, ICIIS 2018 - Proceedings* 1–7 (2018) doi: (2018). https://doi.org/10.1109/ICIINFS.2018.8721402
108. Sun, Q., Tan, Z., Zhou, X.: Workload prediction of cloud computing based on SVM and BP neural networks. J. Intell. Fuzzy Syst. **39**, 2861–2867 (2020)
109. Kumar, A.S., Mazumdar, S. Forecasting HPC workload using ARMA models and SSA. *Proceedings – 2016 15th International Conference on Information Technology, ICIT* 294–297 (2017) doi: (2016). https://doi.org/10.1109/ICIT.2016.52
110. Barati, M., Sharifian, S.: A hybrid heuristic-based tuned support vector regression model for cloud load prediction. J. Supercomputing. **71**, 4235–4259 (2015)
111. Zhong, W., Zhuang, Y., Sun, J., Gu, J. A load prediction model for cloud computing using PSO-based weighted wavelet support vector machine. *Applied Intelligence 2018 48:11* **48**, 4072–4083 (2018)
112. Yang, Q., et al. Multi-step-ahead host load prediction using autoencoder and echo state networks in cloud computing. *The Journal of Supercomputing 2015 71:8* **71**, 3037–3053 (2015)
113. Tian, C., et al.: Minimizing Content Reorganization and Tolerating Imperfect Workload Prediction for Cloud-Based Video-on-Demand Services. IEEE Trans. Serv. Comput. **9**, 926–939 (2016)
114. Zhang, Q., Yang, L.T., Yan, Z., Chen, Z., Li, P.: An Efficient Deep Learning Model to Predict Cloud Workload for Industry Informatics. IEEE Trans. Industr. Inf. **14**, 3170–3178 (2018)
115. Li, S. A workload prediction-based multi-VM provisioning mechanism in cloud computing. 1–6 (2013)
116. Jiang, J., Lu, J., Zhang, G., Long, G. Optimal cloud resource auto-scaling for web applications. *Proceedings – 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2013* 58–65 doi: (2013). https://doi.org/10.1109/CCGRID.2013.73
117. Jheng, J.J., Tseng, F.H., Chao, H.C., Chou, L. der. A novel VM workload prediction using grey forecasting model in cloud data center. *International Conference on Information Networking* 40–45 doi: (2014). https://doi.org/10.1109/ICOIN.2014.6799662
118. Kluge, F., Uhrig, S., Mische, J., Satzger, B., Ungerer, T. Dynamic workload prediction for soft real-time applications. *Proceedings – 10th IEEE International Conference on Computer and Information Technology, CIT*- 7th IEEE International Conference on Embedded Software and Systems, ICESS-2010, ScalCom-2010 1841–1848 (2010) doi: (2010). https://doi.org/10.1109/CIT.2010.317

119. Qazi, K., Li, Y., Sohn, A. PoWER - Prediction of workload for energy efficient relocation of virtual machines. *Proceedings of the 4th Annual Symposium on Cloud Computing, SoCC 2013* doi: (2013). https://doi.org/10.1145/2523616.2525938
120. Hu, Y., Deng, B., Peng, F., Wang, D. Workload prediction for cloud computing elasticity mechanism. *Proceedings of IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2016* 244–249 (2016) doi: (2016). https://doi.org/10.1109/ICCCBDA.2016.7529565
121. Lyu, H., et al. Load forecast of resource scheduler in cloud architecture. *PIC - Proceedings of the 2016 IEEE International Conference on Progress in Informatics and Computing* 508–512 (2017) doi: (2016). https://doi.org/10.1109/PIC.2016.7949553
122. Zhang, L., Zhang, Y., Jamshidi, P., Xu, L., Pahl, C. Workload patterns for quality-driven dynamic cloud service configuration and auto-scaling. *Proceedings – 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, UCC 2014* 156–165 doi: (2014). https://doi.org/10.1109/UCC.2014.24
123. Cao, J., Fu, J., Li, M., Chen, J.: CPU load prediction for cloud environment based on a dynamic ensemble model. Software: Pract. Experience. **44**, 793–804 (2014)
124. Hu, R., Jiang, J., Liu, G., Wang, L., KSwSVR: A new load forecasting method for efficient resources provisioning in cloud. in *Proceedings - IEEE 10th International Conference on Services Computing, SCC 2013* 120–127 doi: (2013). https://doi.org/10.1109/SCC.2013.67
125. Janardhanan, D., Barrett, E. CPU workload forecasting of machines in data centers using LSTM recurrent neural networks and ARIMA models. *12th International Conference for Internet Technology and Secured Transactions, ICITST 2017* 55–60 (2018) doi: (2017). https://doi.org/10.23919/ICITST.2017.8356346
126. Fard, A.K., Akbari-Zadeh, M.R.: A hybrid method based on wavelet, ANN and ARIMA model for short-term load forecasting. J. Experimental Theoretical Artif. Intell. **26**, 167–182 (2014)
127. Usmani, Z., Singh, S.A.: Survey of Virtual Machine Placement Techniques in a Cloud Data Center. Phys. Procedia. **78**, 491–498 (2016)
128. Yu, Y., Gao, Y.: Constraint Programming-Based Virtual Machines Placement Algorithm in Datacenter. IFIP Adv. Inform. Communication Technol. **385 AICT**, 295–304 (2012)
129. Lin, M.-H., Tsai, J.-F., Hu, Y.-C., Su, T.-H. Optimal Allocation of Virtual Machines in Cloud Computing. *Symmetry* Vol. 10, Page 756 **10**, 756 (2018). (2018)
130. Long, S., et al. A Reinforcement Learning-Based Virtual Machine Placement Strategy in Cloud Data Centers. *Proceedings – 2020 IEEE 22nd International Conference on High Performance Computing and Communications, IEEE 18th International Conference on Smart City and IEEE 6th International Conference on Data Science and Systems, HPCC-SmartCity-DSS 2020* 223–230 doi: (2020). https://doi.org/10.1109/HPCC-SMARTCITY-DSS50907.2020.00028
131. Shalu, Singh, D. Artificial neural network-based virtual machine allocation in cloud computing. (2021).
132. Jumnal, A., Dilip Kumar, S.M. Optimal VM placement approach using fuzzy reinforcement learning for cloud data centers. *Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021* 29–35 doi: (2021). https://doi.org/10.1109/ICICV50876.2021.9388424
133. Kaaouache, M.A., Bouamama, S.: An energy-efficient VM placement method for cloud data centers using a hybrid genetic algorithm. J. Syst. Inform. Technol. **20**, 430–445 (2018)
134. Tawfeek, M.A., El-Sisi, A.B., Keshk, A.E., Torkey, F.A.: Virtual Machine Placement Based on Ant Colony Optimization for Minimizing Resource Wastage. Commun. Comput. Inform. Sci. **488**, 153–164 (2014)
135. Pires, F.L., Barán, B. Multi-objective virtual machine placement with service level agreement: A memetic algorithm approach. *Proceedings – 2013 IEEE/ACM 6th International Conference on Utility and Cloud* Computing, *UCC 2013* 203–210 doi: (2013). https://doi.org/10.1109/UCC.2013.44
136. Li, X.K., Gu, C.H., Yang, Z.P., Chang, Y.H. Virtual machine placement strategy based on discrete firefly algorithm in cloud environments. 12th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP (2015). https://doi.org/10.1109/ICCWAMTIP.2015.7493907
137. Abdel-Basset, M., Abdle-Fatah, L., Sangaiah, A.K. An improved Lévy based whale optimization algorithm for bandwidth-efficient virtual machine placement in cloud computing environment. Cluster Computing **22**, 8319–8334 (2018)
138. Gharehpasha, S., Masdari, M., Jafarian, A. (2020) Power efficient virtual machine placement in cloud data centers with a discrete and chaotic hybrid optimization algorithm. Cluster Computing **24**, 1293–1315
139. Li, X., Qian, Z., Lu, S., Wu, J.: Energy efficient virtual machine placement algorithm with balanced and improved resource utilization in a data center. Math. Comput. Model. **58**, 1222–1235 (2013)
140. Jamali, S., Malektaji, S., Analoui, M.: An imperialist competitive algorithm for virtual machine placement in cloud computing. J. Experimental Theoretical Artif. Intell. **29**, 575–596 (2017)
141. Baalamurugan, K.M., Vijay Bhanu, S.: A multi-objective krill herd algorithm for virtual machine placement in cloud computing. J. Supercomputing. **76**, 4525–4542 (2020)
142. Laganà, D., Mastroianni, C., Meo, M., Renga, D. Reducing the operational cost of cloud data centers through renewable energy. Algorithms **11** (2018)
143. Khalil, M.I.K., Ahmad, I., Almazroi, A.A.: Energy Efficient Indivisible Workload Distribution in Geographically Distributed Data Centers. IEEE Access. **7**, 82672–82680 (2019)
144. Abu Bakar Siddik, M., Shehabi, A., Marston, L.: The environmental footprint of data centers in the United States. Environ. Res. Lett. **16**, 64017 (2021)
145. Improving Data Center Power Consumption & Energy Efficiency:. https://www.vxchnge.com/blog/growing-energy-demands-of-data-centers
146. Solar Powered Datacenters Drive Sustainable Growth - CtrlS Blog:. https://www.ctrls.in/blog/solar-powered-datacenters-drive-sustainable-growth/
147. Project Natick Phase 2:. https://natick.research.microsoft.com/
148. Data Center Energy Efficiency Standards in India: : Preliminary Findings from Global Practices | Energy Technology Area. https://eta.lbl.gov/publications/data-center-energy-efficiency
149. Mostafavi, M., Kabiri, P.: Detection of repetitive and irregular hypercall attacks from guest virtual machines to Xen hypervisor. Iran. J. Comput. Sci. 2018. **1**(2 1), 89–97 (2018)
150. Virtual machines to run 50% of workloads by 2012: :Gartner. https://www.computerweekly.com/news/1372216/Virtual-machines-to-run-50-of-workloads-by-2012-Gartner
151. Ferdaus, M.H., Murshed, M., Calheiros, R.N., Buyya, R.: Network-aware virtual machine placement and migration in cloud data centers. Emerging Res. Cloud Distrib. Comput. Syst. (2015). https://doi.org/10.4018/978-1-4666-8213-9.CH002

Avita Katal Avita Katal is working as an Assistant Professor in School of Computer Science, University of Petroleum and Energy Studies Dehradun, Uttarakhand. She has received her B. E degree from University of Jammu in Computer Science Engineering in 2010 and M. Tech degree in the year 2013. She is currently pursuing her Ph.D. in the area of Cloud Computing. Her research interest is in the area of Cloud Computing, Mobile Ad hoc Networks, IoT, and Artificial Intelligence. She has published various research papers in renowned conferences and journals and has also served as a reviewer in various conferences and journals. She is an active member of IEEE, IEEE Women in Engineering and ACM. She is actively involved in all areas of education including research, curriculum development, teacher mentoring, student career preparation and community work with a genuine interest in student's cognitive and social growth.

Susheela Dahiya Susheela Dahiya is working as Assistant Professor in School of Computer Science, University of Petroleum and Energy Studies (UPES), Dehradun. She had completed her M.Tech. and Ph.D. degree from IIT, Roorkee. She has more than 8 years of academic/research/industry experience. Her research work is focused on satellite image processing, video processing, cyber security, and cloud computing and deep learning. She has published several research papers in SCI and Scopus indexed journals and conferences.

Tanupriya Choudhury received his bachelor's degree in CSE from West Bengal University of Technology, Kolkata, India, and master's Degree in CSE from Dr. M.G.R University, Chennai, India. He has received his Ph.D. degree in the year 2016. He has Ten years' experience in teaching as well as in Research. Currently he is working as an Associate Professor in dept. of CSE at UPES Dehradun. Recently he has received Global Outreach Education Award for Excellence in bestYoung researcher Award in GOECA 2018 His areas of interests include human computing, softcomputing, Cloud computing, Data Mining etc. He has filed 14 patents till date and received 16copyrights from MHRD for his own software. He has been associated with many conferences inIndia and abroad.He has authored more than 85 research papers till date. He has delivered invitedtalk and guest lecture in Jamia Millia Islamia University, Maharaja Agersen College of DelhiUniversity, Duy Tan University Vietnam etc. He has been associated with many conferencesthroughout India as TPC member and session chair etc. He is a lifetime member of IETA, memberof IEEE, and member of IET (UK) and other renowned technical societies. He is associated withCorporate and he is Technical Adviser of DeetyaSoft Pvt. Ltd. Noida, IVRGURU, and Mydigital360etc. He is holding the post of Secretary in IETA (Indian Engineering Teacher's Association-India),He is also holding the Advisor Position in INDO-UK Confederation of Science, Technology andResearch Ltd. London, UK and International Association of Professional and Fellow Engineers-Delaware-USA.