LOCUST-INSPIRED META-HEURISTIC ALGORITHM FOR OPTIMISING CLOUD COMPUTING PERFORMANCE

By

MOHAMMED ALAA FADHIL

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

January 2023

FSKTM 2023 1

DEDICATIONS

This thesis is dedicated to my father, Dr. Alaa, whose unwavering love and support have been a constant source of inspiration throughout my life. His encouragement and guidance were instrumental in my decision to pursue higher education and his belief in me never wavered. His wisdom and experience have been invaluable in the completion of this work.

I would also like to dedicate this thesis to my family, whose love and support has been my foundation throughout my life. To my mother, whose love and prayers are the reason for my success. To my wife, whose dedication, love, and persistent confidence in me have taken the load off my shoulders. To my sisters, for their unflagging love and support throughout my life. I have no suitable words that can fully describe my everlasting love to them except, I love you all.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

LOCUST-INSPIRED META-HEURISTIC ALGORITHM FOR OPTIMISING CLOUD COMPUTING PERFORMANCE

By

MOHAMMED ALAA FADHIL

January 2023

Chair : Mohamed Othman, PhD
Faculty : Computer Science and Information Technology

Cloud computing offers high computational resources at a reasonable price-level. This has led to a great migration of users to cloud computing from other modes of computing. Cloud computing resources are offered on a pay-as-you-use basis, allowing users to be free from maintenance costs. The cloud paradigm has arisen due to a rapid growth in applications and data sizes. Even though cloud computing servers and resources may seem unlimited, this is not true, as increased server usage leads to increased energy consumption and carbon emissions. Therefore, minimising the number of active servers in a cloud-computing set-up can significantly improve energy consumption. Additionally, reducing the number of virtual machine migrations can improve the hardware reliability of the overall cloud computing system. Another aspect that can increase user satisfaction is the scheduling of users' tasks, as many agencies, organisations, and departments are responsible for time-critical tasks that need to be completed as soon as possible at reasonable cost.

This thesis presents three significant contributions to the field of knowledge. The first contribution entails a study on server consolidation, which employs the Locust Scheduling Meta-Heuristic Algorithm (LACE). This contribution is composed of three distinct parts. The first part involves a review of prior locust-inspired algorithms, while the second part concerns the adaptation of the algorithm to the cloud computing paradigm. The third part addresses the limitations of LACE algorithms, leading to the proposition of a novel meta-heuristic algorithm called the Locust-Inspired Algorithm (LIA) that can effectively map virtual machines (VMs) for efficient server consolidation. This algorithm can also be used for task scheduling. The proposed algorithm efficiently maps and achieves the objective function for server consolidation, optimising energy consumption, VM migrations, and server utilisation. To validate the effectiveness of the pro-

posed algorithm, it was tested via simulation using real datasets. Furthermore, a mathematical model was developed, which models the cloud computing infrastructure, capable of allocating VMs to a minimum number of servers, increasing server utilisation, and triggering necessary migrations to reduce underutilised servers. The simulation results demonstrate that the proposed algorithm outperforms existing heuristic and meta-heuristic algorithms, including the benchmarking algorithm (LACE). The proposed algorithm demonstrated a 61.8% and 81.03% reduction in energy consumption and VM migrations, respectively, compared to LACE. Additionally, the proposed algorithm exhibited superior performance compared to other state-of-the-art algorithms.

The second contribution of the thesis concerns the scheduling of independent tasks, called cloudlets. In this contribution, a novel analogy of the locust-inspired algorithm is presented in the field of cloudlet scheduling. The proposed algorithm has the ability to improve cloudlet allocation to meet the objective function. The problem is modelled as a set of events that locates an appropriate VM on which to allocate the cloudlet. The proposed algorithm's efficiency is evaluated using the CloudSim toolkit and a synthetic dataset. Results reveal that it outperforms other state-of-the-art nature-inspired algorithms such as TOPSIS-PSO, FUGE, ACO, and MACO, with average improvements of 55.6%, 66.9%, and 31.6% in makespan, waiting time, and resource utilisation, respectively.

The third contribution arises from investigating the scheduling of dependent tasks, where most of the tasks have parents and children, and the batch of tasks is called a job. These tasks are connected together based on the model structure. The scientific workflow has an immense computational requirement, which is considered data-intensive. The LIA is considered a novel algorithm that adapts the study of locust movement behaviour from biology to job scheduling in the cloud computing environment. The proposed algorithm is used with four different workflow structures (Montage, Cybershake, Inspiral, and SHIPT) and their datasets within a range of 50, 100, and 1000 tasks. The proposed algorithm is evaluated using the WorkflowSim simulation with a real dataset. From the results, the LIA improves job allocation by reducing job makespan and cutting the cost of using resources. The job scheduling of the scientific workflow can efficiently outperform state-of-the-art competitor algorithms.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

ALGORITMA META-HEURISTIK TERINSPIRASI BELALANG UNTUK MENGOPTIMUMKAN PRESTASI KOMPUTERAN AWAN

Oleh

MOHAMMED ALAA FADHIL

Januari 2023

Pengerusi : Mohamed Othman, PhD
Fakulti : Sains Komputer dan Teknolologi Maklumat

Pengkomputeran awan menawarkan perkhidmatan yang pengkomputeran tinggi pada tahap harga yang berpatutan. Ini telah membawa kepada penghijrahan besar pengguna kepada pengkomputeran awan daripada mod pengkomputeran lain. Sumber pengkomputeran awan ynag ditawarkan adalah menggunakan pendekatan bayar mengikut penggunaan (pay-as-you-use) yang membolehkan pengguna bebas daripada kos penyelenggaraan. Paradigma awan telah pesat membangun disebabkan oleh keperluan aplikasi dan saiz data yang semakin meningkat. Walaupun pelayan dan sumber pengkomputeran awan menawarkan perkhidmatan yang tidak terhad, namun implikasi sebaliknya adalah peningkatan penggunaan tenaga dan pelepasan karbon yang tidak terkawal. Oleh itu, meminimumkan bilangan pelayan yang aktif dalam persediaan pengkomputeran awan boleh mengurangkan penggunaan tenaga dengan ketara. Selain itu, dengan meminumkan bilangan pemindahan mesin maya boleh menambahbaik kebolehpercayaan perkakasan pada keseluruhan sistem pengkomputeran awan. Di samping itu, bagi meningkatkan kepuasan pengguna, kaedah penjadualan tugas pengguna boleh dilaksanakan, kerana banyak agensi, organisasi dan jabatan bertanggungjawab untuk menyelesaikan tugasan kritikal dalam jangkamasa yang pendek serta pada kos yang berpatutan.

Tesis ini membentangkan tiga sumbangan penting kepada bidang penyelidikan. Sumbangan pertama adalah kajian tentang penyatuan pelayan, yang menggunakan Algoritma Meta-Heuristik Penjadualan Locust (LACE). Sumbangan ini terdiri daripada tiga bahagian yang berbeza. Bahagian pertama melibatkan kajian semula algoritma yang diilhamkan oleh belalang sebelum ini, manakala bahagian kedua melibatkan penyesuaian algoritma kepada paradigma pengkomputeran awan. Bahagian ketiga menangani batasan algoritma LACE, yang membawa kepada cadangan algoritma meta-heuristik baru yang dipanggil Algo-

ritma Locust-Inspired (LIA). LIA berkeupayaan memetakan mesin maya (VM) dengan berkesan untuk penyatuan pelayan yang lebih cekap. Algoritma ini juga boleh digunakan untuk penjadualan tugas. Algoritma yang dicadangkan adalah cekap memetakan dan mencapai fungsi objektif untuk penyatuan pelayan, mengoptimumkan penggunaan tenaga, migrasi VM dan penggunaan pelayan. Untuk membuktikan keberkesanan algoritma yang dicadangkan, pengujian menggunakan simulasi dengan set data sebenar telah dilaksanakan. Tambahan pula, model matematik telah dibangunkan, yang memodelkan infrastruktur pengkomputeran awan, serta mampu memperuntukkan VM dengan bilangan minimum pelayan bagi meningkatkan penggunaan pelayan, dan mencetuskan migrasi yang diperlukan untuk meminumkan pelayan yang kurang digunakan. Keputusan simulasi membuktikan bahawa algoritma yang dicadangkan mengatasi algoritma heuristik dan meta-heuristik sedia ada, termasuk algoritma penanda aras (LACE). Algoritma yang dicadangkan menunjukkan pengurangan 61.8% dan 81.03% dalam penggunaan tenaga dan migrasi VM, masing-masing, berbanding LACE. Selain itu, algoritma yang dicadangkan mempamerkan prestasi unggul berbanding dengan algoritma yang lain.

Sumbangan kedua tesis ini adalah penjadualan tugas bebas yang dipanggil *cloudlets*. Dalam sumbangan ini, analogi baru algoritma yang diilhamkan oleh belalang dibentangkan dalam bidang penjadualan *cloudlet*. Algoritma yang dicadangkan mempunyai keupayaan untuk menambah baik peruntukan *cloudlet* untuk memenuhi fungsi objektif. Algoritma ini dimodelkan sebagai satu set acara bagi mencari VM yang sesuai untuk memperuntukkan *cloudlet*. Eksperimen menyeluruh dijalankan menggunakan aplikasi simulasi "CloudSim" dengan set data sintetik. Keputusan mendedahkan bahawa algoritma baru yang dicadangkan mengatasi prestasi algoritma tercanggih yang diilhamkan oleh alam semula jadi lain seperti TOPSIS-PSO, FUGE, ACO dan MACO, dengan purata peningkatan sebanyak 55.6%, 66.9% dan 31.6% dalam makespan, masa menunggu dan penggunaan sumber, masing-masing.

Sumbangan ketiga pula adalah tertumpu bagi mengkaji semula kebergantungan penjadualan tugas, di mana kebanyakan tugas mempunyai *parent* dan *child*, kumpulan tugas yang dikenali sebagai kerja (*job*). Tugas-tugas ini disambungkan bersama berdasarkan struktur model. Aliran kerja saintifik mempunyai keperluan pengiraan yang besar, yang dianggap intensif data. LIA dianggap sebagai algoritma baru yang menyesuaikan kajian tingkah laku pergerakan belalang daripada biologi kepada penjadualan kerja dalam persekitaran pengkomputeran awan. Algoritma yang dicadangkan digunakan dengan empat struktur aliran kerja yang berbeza (Montaj, Cybershake, Inspiral dan SHIPT) dan set datanya dalam julat 50, 100 dan 1000 tugasan. Algoritma yang dicadangkan diuji menggunakan simulasi 'WorkflowSim' dengan set data sebenar. Daripada keputusan itu, LIA mampu menambah baik peruntukan pekerjaan dengan mengurangkan kerja dan mengurangkan kos penggunaan sumber. Penjadualan kerja aliran kerja saintifik boleh mengatasi prestasi algoritma pesaing terkini dengan lebih efisen.

ACKNOWLEDGEMENTS

I am deeply grateful to my Lord, Allah, for granting me the strength, wisdom, and perseverance to complete this thesis. My great appreciation goes to my supervisor, Professor Dr. Mohamed Othman, for his unwavering support, guidance, and encouragement throughout the entire research and writing process. His patience, motivation, and expertise have been invaluable, and I am forever grateful for his mentorship.

I would also like to extend my thanks to the members of my supervisory committee, Dr. Zurina Mohd Hanapi and Dr. Mohamed A. Alrshah, for their valuable insights and feedback.

I am indebted to the Faculty of Computer Science and Information Technology, the staff of the Postgraduate Office, the School of Graduate Studies, the Library, and Universiti Putra Malaysia for providing me with a conducive research environment. I am also grateful to all those who have supported me throughout my pursuit of a Ph.D.

I am eternally grateful to my parents, Dr. Alaa and my mother, for their love and support. My wife, for her unwavering dedication, love, and support, has been a constant source of inspiration and motivation. I would also like to express my love and appreciation to my daughter Aisha, for her constant love and support throughout my life.

I am also thankful to my friends, colleagues, and lab mates for their support and encouragement. I would also like to apologise to anyone who I may have inadvertently omitted in this acknowledgement.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy.

The members of the Supervisory Committee were as follows:

Mohamed bin Othman, PhD
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Zurina binti Mohd Hanapi, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Mohamed Alhadi Mahmoud Alrshah, PhD
Assistant Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

ZALILAH MOHD SHARIFF, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 8 June 2023

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

LIA	Locust-inspired algorithm
VM	Virtual machine
PM	Physical machine (server)
SLA	Service level agreement
QoS	Quality of service
PaaS	Platform as a service
IaaS	Infrastructure as a service
SaaS	Software as a service
XaaS	Anything as a service
IoT	Internet of Things
WBAN	Wireless Body Area Networks
VANET	Vehicular Ad-hoc Networks
PC	Processing capacity
VP	VM processor
PS	Powerful servers
SP	Server processor
WS	Weak servers
S	Set of servers
SST	Server static threshold
e	Set of VMs without corresponding attributes
VM_r	VM resources
S^h	Heavy loaded server
S_r	Resources for a specific server

S^l	Lightly loaded server
$Core_r$	Resources' core (CPU core)
A	Processor utilisation
BW/bw	Bandwidth
LLS	Least loaded server
PE/pe	Processor element
GMR	Global migration rule
LMR	Local migration rule
MIPS	Million instructions per second unit
f	Objective function
B	Set of processor utilisation of VM
Mig	The number of migration for migration plan
TFR	Total free resources
α	Server utilisation
CAR	Current available resources
S_s	The set of the servers that will be turned-off
U_{LLS}	The resources used in the LLS
c	CPU
$\mathbb{Z}^+$	Positive integer number
r	RAM
DAG	Directed acyclic graph
T	Set of tasks
T'	Set of tasks in non-increasing order from T
$\delta(st_j)$	The startup time of vm_j

$\delta(ru_{i,j})$	The running time of t_i on vm_j
$\delta(Thr_i)$	The throughput time of vm_j that is used by t_i
$\delta(ft_i)$	The completion time of vm_j.
Γ	The cost
$\gamma(ru_{i,j})$	The cost of executing t_i on vm_j
$\gamma(Thr_i)$	The cost of transmitting the data generated by t_i
$\gamma(ds_i)$	The cost of data storage of t_i
γ_{vt}	VM type cost

CHAPTER 1

INTRODUCTION

1.1 Background

Cloud computing is the current computer technology for delivering services to customers on demand. This technology eases access to information through various devices, for instance smartphones, laptops, and tablets. Nowadays, cloud computing is considered a worldwide trend, with many advantages. There are three models of cloud service, namely Infrastructure as a Service (IaaS), Software as a Service (SaaS), and Platform as a Service (PaaS). Many clients, industries, and so forth are migrating their data, data processing, information, etc. onto cloud computing platforms. These resources are spread all around the world for the rapid delivery of services to users (Dasgupta et al., 2013; Apostu et al., 2013). Resources are collections of physical or virtual components of bounded availability within a computer structure. Any connected device is considered as a resource as is any internal component of the system, as listed in Table 1.1 (Manvi and Shyam, 2014).

Table 1.1: Physical and logical resources

Physical resources	Logical resources
Storage	Bandwidth (BW)
Memory	Energy
CPU	Operating system
Workstations	Information security, protocols
Network elements	APIs
Sensors/actuators	Network loads, delays

(Manvi and Shyam, 2014)

According to (Kaur and Luthra, 2012; Malladi, 2015), once cloud computing emerged, many challenges were encountered such as scaling, security, Quality of Service (QoS) management, resource scheduling, data centre energy consumption, service availability, data lock-in, and competent load balancing. Nevertheless, a cloud computing platform has three logical components, namely the service providers, the service users, and the intermediary between them that may be an application, network, bandwidth, etc. The intermediary is the responsibility of software developers, while service providers and end-users are the responsibility of cloud computing researchers, as researchers must improve the ways of using services and managing them. For instance, an efficient server consolidation algorithm can reduce energy consumption in cloud computing data centres. Meanwhile, an efficient task scheduling algorithm can improve the execution time of the users' tasks and reduce the cost of using the resources and many more benefits that can be obtained. Therefore, server consolidation and task scheduling (for independent and dependent tasks) are the main concerns in cloud computing (Jadeja and Modi, 2012; Preist and Shabajee, 2010). Task scheduling is the process of assigning the load among available resources in order to improve resource utilisation and the execution time to manage tasks and reduce cost and waiting time where it can achieve high user satisfaction (Singh et al., 2016; Goyal and Verma, 2016).

On the other hand, server consolidation can play a vital role in enhancing energy consumption, while preserving the Service Level Agreement (SLA), and at the same time an efficient server consolidation algorithm can reduce the VMs that need to be migrated. Therefore, effective server consolidation and efficient task scheduling algorithms can boost the success of cloud computing environments with high QoS. A lot of research has been conducted on server consolidation and task scheduling of cloud computing; however, even though cloud computing still faces many problems. Cloud computing should have two goals: server consolidation and task scheduling, hence according to (Ivanisenko and Radivilova, 2015), the results of these goals are:

1. High resource availability
2. Increasing resource utilisation
3. Reduction in resource cost
4. Preserving the elasticity of cloud computing
5. Reduction of carbon emissions
6. Energy savings.

Server consolidation could attain the green cloud computing status, in terms of energy and carbon emissions. The literature reveals that the world's data centres consumed twice as much electricity in 2005 compared to 2000. Nevertheless, the upward trend in energy consumption slowed remarkably from 2005-2010, which was due to the economic crisis. Since 2005, the industry has made more effort to improve the efficiency of data centres and at the same time to spread virtualisation technology that improves data centre exploitation (Koomey et al., 2011; Shehabi et al., 2018).

For instance, United States data centres reported growth, where 6000 data centres consumed 61×10^9 kWh of energy in 2006, which represents 1.5% of all U.S. electricity consumption, costing $4.5 billion (Vrbsky et al., 2010). More recently, U.S. data centres consumed around 70×10^9 kWh, and the energy consumption was about 2% of the total electricity consumption of the country. At the same time, the data centres' workload exponentially increased (Shehabi et al., 2018). Data centres in the United States consume a significant amount of energy. According to the U.S. Energy Information Administration (EIA) (EIA, 2020), data centres in the United States consumed an estimated 91 billion kilowatt-hours (kWh) of electricity in 2020, which is about 1.8% of total electricity consumption in the country. This represents a significant increase from the past decade, due in part to the growing amount of data being stored and processed by data centres.

Additionally, data centres are becoming more energy-efficient with the help of new technologies and energy saving practices, but energy consumption is still a concern as the amount of data being processed increases. The growth of data centre electricity in 2022 and beyond is uncertain, but based on Moore's law, the data centre energy consumption is moving upwards from 200 TWh to 800 TWh in the upcoming 10 years as illustrated in Figure 1.1 (Koot and Wijnhoven, 2021).

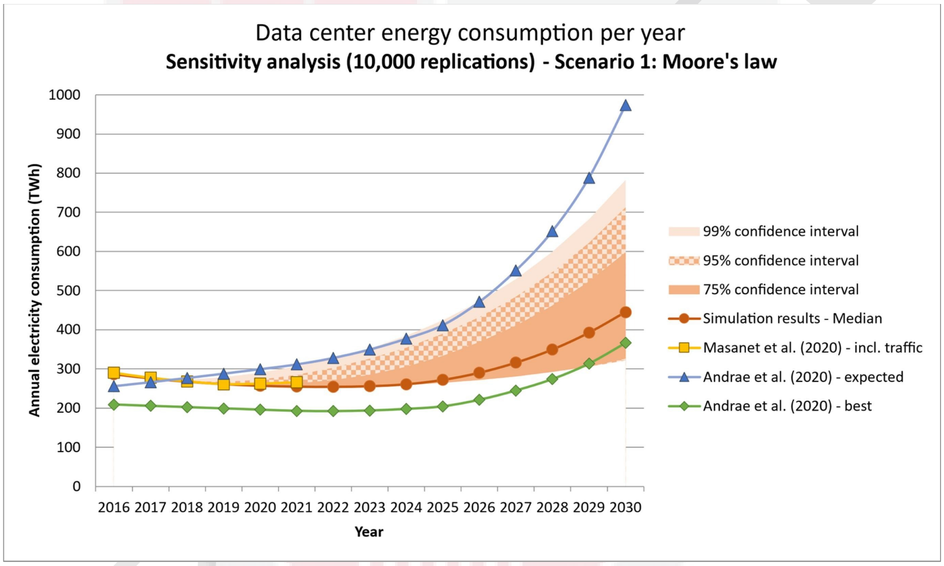

Figure 1.1: Forecasting data centre energy with Moore's law

The modelled trends chart indicates the past and the projected growth average of the electricity consumed for the years between 2016-2030. While the previous measures may not be enough for the data centres in the future, if the industry does not address this issue by using an efficient optimisation method, such as the successful stabilisation of data centre energy consumption, there will need to be innovations in the efficiency of the data centres. Researchers such as Preist and Shabajee (2010) predict the energy consumption may reach 10300 TWh per year in 2030, based on 2010 efficiency levels. Also, Koot and Wijnhoven (2021) in Figure 1.2 has presented a forecast of energy consumption of data centres that consume a huge amount of electricity that needs serious attention by researchers.

All of these enlargements in energy consumption are projected. The standby (idle) and underutilised servers could also be contributing significantly to energy wastage and carbon emissions. In (Blackburn and Hawkins, 2013) it is reported that standby servers emit 11 million tons per year of CO_2 and the total cost for standby servers is about $19 billion. Gartner research (Snyder, 2010) reported the ratio of the unutilised servers as 18% in the huge data centres, while the utilisation of the x86 servers is even lower at 12%. These results confirmed that server utilisation falls in the range of between 10-30% (Barroso and Hölzle, 2007).

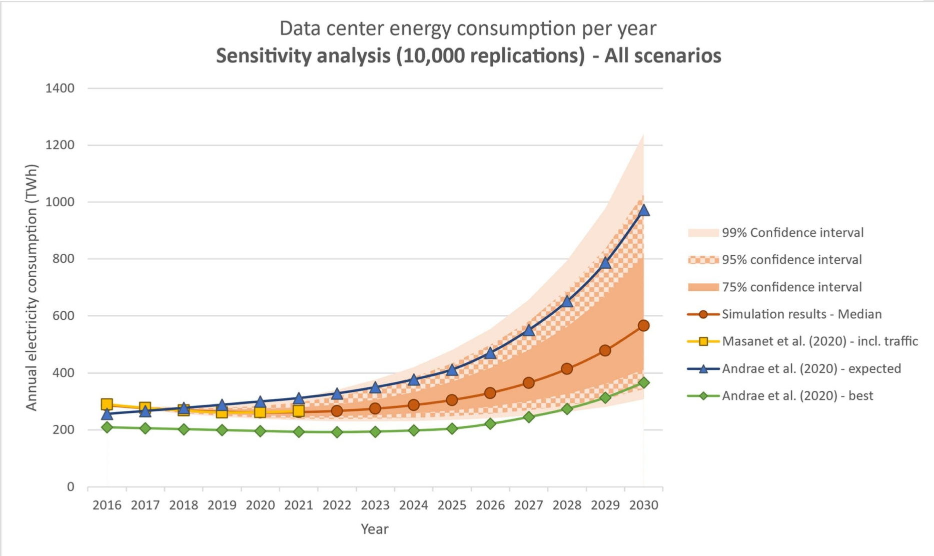

Figure 1.2: Energy Consumption per year (Koot and Wijnhoven, 2021)

As a result, efficient resource management can be utilised to reduce both operational costs and environmental effects (such as carbon emissions) while achieving system stability. On the other hand, user satisfaction can be achieved too when an efficient utilisation of the resources could be achieved that can impact the user budget and task execution. The users' tasks need efficient scheduling to achieve the efficient utilisation of cloud computing resources.

1.2 Motivation

The rapid development of information technology and its variety of uses has led to the emergence of cloud computing after decades of evolution of computing facilities. Previous computing technology has had many challenges and drawbacks. Therefore, the coming technology seeks to overcome or avoid those drawbacks by being more extendable, advanced, and compatible with other technologies.

Cloud computing is tied to many technologies. Examples are the Internet of

Things (IoT) (Distefano et al., 2015; Botta et al., 2016), e-Health applications with cooperating Wireless Body Area Networks (WBAN) (Diallo et al., 2014), big data management and Vehicular Ad-hoc Networks (VANET) (Botta et al., 2016). The complex diversity of approaches to cloud computing and the burden of its energy needs makes it challenging to narrow the whole field down to one optimisation algorithm. Considering energy efficiency which can be gained from server consolidation is not enough for a real application, hence this will lead to problems such as unbalanced loads for each server (Tian et al., 2018)), while optimising task allocation is another aspect that needs to be improved in order to increase user satisfaction. The improvement to task scheduling can improve execution time while preserving the hardware reliability of servers that may reduce high migration rate (i.e., migration from VM to VM and from server to server). Also, the cost required to use the resources may reduce accordingly when the scheduling is improved. Therefore, improving server consolidation and task scheduling (for independent and dependent tasks) could obtain integrated solutions based on a meta-heuristic algorithm.

1.3 Server Consolidation

Data centre services are exponentially propagated. Cloud providers present their services by virtualized Physical Machines (PMs) in an active virtual machine. This needs to be sold to clients by offering high performance and high data repository volume (Beloglazov et al., 2012). Meanwhile, the virtualisation technology of data centres is broadly employed to ease the management of PMs or "servers". However, this employment of PMs to VMs might affect the performance of the data centres if carried out incorrectly. This leads to avoiding the data centre's sprawl, energy consumption, and a large carbon footprint (Blackburn and Hawkins, 2013). On the other hand, this technology brings many benefits such as resource allocation, VM resizing, live migration, and server consolidation (Wang et al., 2012). Server consolidation is widely employed to decrease the total energy consumption in data centres as well as carbon emissions (Vogels, 2008; Ferreto et al., 2011; Wood et al., 2009).

Furthermore, the wastage of resources is at the heart of the spread of cloud computing (Pop et al., 2012). This leads to more energy wastage. Barroso and Hölzle (2007) revealed that the average server utilisation levels are between 10% and 50%. The main reason why server consolidation is a prominent topic for researchers is the issue of virtual machine live migration. Live migration is considered the best way to reduce energy consumption by reducing the number of active servers in the data centre. Figure 1.3 demonstrates a server consolidation overview by taking four servers as an example to implement VMs migration and turning off the unused servers.

Virtual machine live migration can employ VMs to move among the servers with much better system downtime to avoid SLA violations while preserving the QoS. In other words, server consolidation (Varasteh and Goudarzi, 2015) involves placing several VMs on a smaller number of PMs for enhancing resource utilisation and reducing energy consumption while using a more attractive feature for the server consolidation technique (i.e., VM live migration). Hence, this feature allows a processing VM to be relocated from one PM to another without interrupting the service. The VM migration methods might differ based on parameter variations.

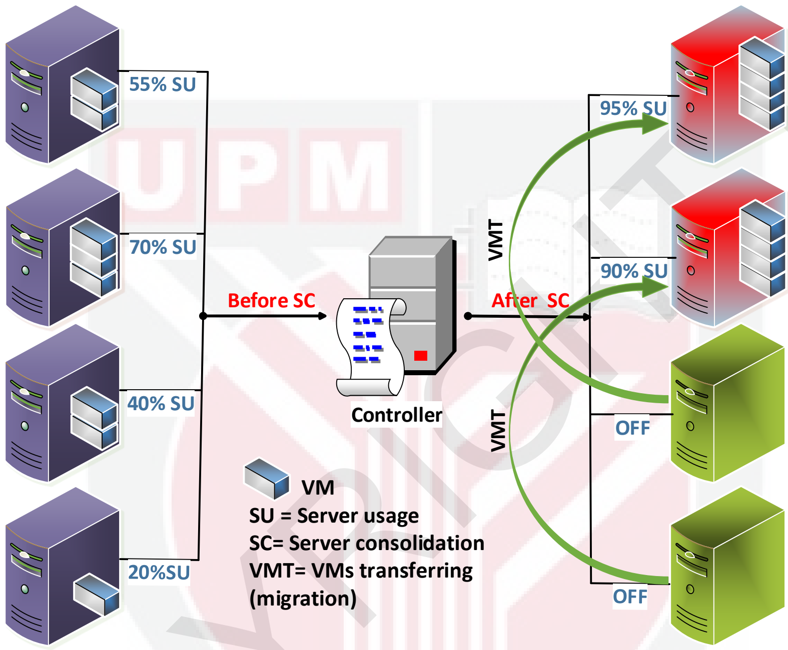

Figure 1.3: Server consolidation overview.

1.4 Cloudlet Scheduling

Cloud computing loads are unsteady, based on user requirements and the needs of resources. Load balancing is one of the main challenges in this area that cannot be neglected (Jadeja and Modi, 2012). It is the process of assigning and reassigning the load among available resources in order to achieve better utilisation to minimise the cost, energy consumption, and response time (Singh et al., 2016; Goyal and Verma, 2016). Load balancing organises the workload in a perfect manner across all the resources to achieve competent resource utilisation, user satisfaction, fair allocation of resources, as well as expanding scalability and preventing over-provisioning and bottlenecks, etc. (Milani and Navimipour, 2016).

An overview of the task scheduling model is demonstrated in Figure 1.4. The model introduces some components of the data centre such as physical components (servers) and virtualized components (i.e., VMs). The task load balancer receives clients' demands and implements a load balancing algorithm for the tasks to allocate the demands among the VMs. The load balancer selects the suitable VM that should be allocated to the upcoming demand. The data centre controller is responsible for task management. Hence, tasks are submitted to the load balancer, which implements the load balancing algorithm to select the suitable VM to manage that task or set of tasks and then the balancer will preserve the PMs' balance all the time.

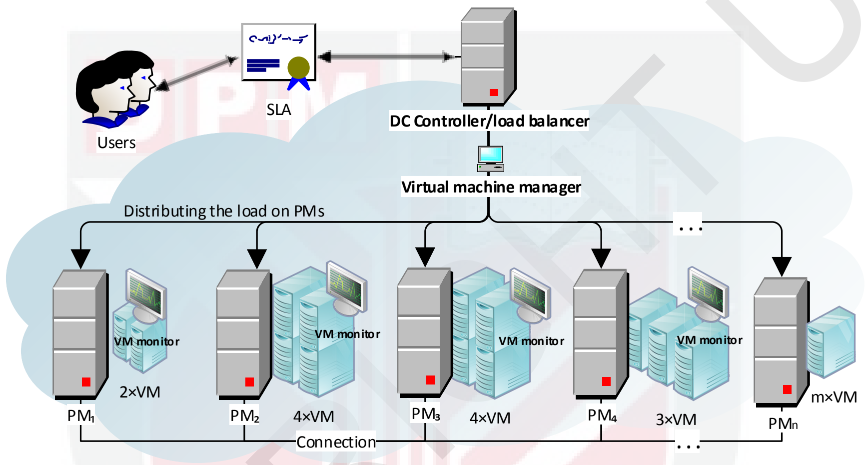

Figure 1.4: Dependent and independent task scheduling overview.

The VM manager is responsible for VMs. The dominant technology in the cloud computing environment is virtualisation, which aims to distribute expensive hardware among VMs. A virtual machine is a software application that manages systems that allow applications to run. Cloud computing users are placed all around the world and randomly submit their demands to the VMs for processing. Thus, the assignment of tasks is one of the most important concerns in cloud computing and should be considered to preserve the QoS. When some VMs are idle, overloaded, or have few tasks to manage, then the QoS will be decreased, leading to user dissatisfaction, and the user will try to migrate their work to another service provider. The Virtual Machine Monitor (VMM) or "hypervisor" is used to manage and create the VMs (LD and Krishna, 2013). VVMM presents four procedures: provision (resume), suspension (storage), multiplexing, and live migration (Hwang et al., 2013). These procedures are essential for load balancing.

1.5 Job Scheduling of Scientific Workflow

The scientific workflow is a process for accomplishing a scientific objective, usually expressed in terms of tasks and their dependencies. These tasks need an efficient platform to manage them, especially when the number of tasks is huge (e.g., hundreds or thousands of tasks) (Li et al., 2015). In scientific computing applications such as bio-informatics, physics, astronomy, etc., a workflow is the most widely used model for representing scheduled tasks (Song et al., 2017). For instance, creating a wide picture of the galaxy requires a mosaic image that contains hundreds of tasks to generate it. Also, the anatomy of an earthquake is data-intensive where the epicentre needs to analyse and evaluate a massive amount of workflow data (Callaghan et al., 2011). This data-intensive process needs high computational resources with the lowest cost such as CPUs, memory, storage units, etc.

Cloud computing supports data-intensive computation by providing resources in a cost-effective and scalable manner since the users of cloud computing can pay based on their use (Aslam et al., 2017; Ferdaus et al., 2017; Ghazouani and Slimani, 2017). Cloud computing provides the virtualisation technology which is considered the most promising solution for sharing resources among the users by shrinking resources based on the cloud user requests and this can be said to be multi-cloud computing. This can be provided by various cloud-based IaaS providers such as Google Compute Engine, Amazon EC2, Microsoft Azure, etc. (Sooezi et al., 2015). Scheduling the workflow in cloud computing is still a challenging issue to meet the QoS such as the execution time and cost.

1.6 Problem Statement

Cloud computing is a paradigm that delivers computing services to users on a pay-as-you-use basis. Due to its economical basis and other benefits, the number of cloud computing users has risen significantly, resulting in millions of instructions and data that need to be retrieved or stored in cloud storage per time unit, represented as tasks. These tasks must be scheduled to a cloud provider that has enough resources to meet the users' requirements, as regulated by agreements between cloud providers and users. Therefore, both tasks and servers must be taken into consideration. Despite fluctuations in user requirements for the number of resources needed and the number of tasks managed by VMs, an efficient algorithm is required to handle these fluctuations. Efficient server utilisation and task scheduling algorithms remain challenging issues in the cloud computing environment.

Despite various studies by (Kurdi et al., 2018; Panwar et al., 2019; Shojafar et al., 2015; Tawfeek et al., 2013; Yonggui and Ruilian, 2011; Konjaang and Xu, 2021), many challenges remain unresolved in the research domain of server consolidation and task scheduling (independent tasks and dependent tasks) in cloud computing environments. The three main problems that motivate this study are:

1. The current meta-heuristics used for server consolidation in cloud computing (i.e., LACE) have been found to suffer from local optima entrapment, resulting in a slow convergence rate. This leads to increased VM migrations, energy consumption, and high resource utilisation. Additionally, cloud server utilisation is plagued by high wastage of resources and excessive energy consumption, particularly in the case of idle servers which consume around 70% of the resources of a fully utilised server. This highlights the need for an efficient algorithm for mapping servers in cloud computing environments, particularly for idle servers.
2. The performance of cloud computing systems can be impacted by long processing times during the scheduling of cloudlets, leading to prolonged completion times. Effective scheduling of cloudlets that considers customers' QoS expectations for makespan, waiting time, and resource utilisation is critical. This underscores the importance of developing an efficient task scheduling algorithm to address these challenges in the cloud computing environment.
3. The task scheduling in scientific workflows is challenging, particularly with the rising costs and time limitations in cloud computing and the complex data processing requirements of these workflows. This leads to the need for a novel, effective algorithm that can optimise the scheduling of tasks while minimising overall costs and maximising makespan, considering heterogeneous systems, the priority and balance of tasks, child-parent tasks, and link latency.

1.7 Research aims and Objectives

The aim of this research is to enhance cloud computing performance by designing and implementing meta-heuristic algorithms that address different aspects of cloud computing, such as server consolidation, independent task scheduling, and dependent task scheduling. To achieve this main aim, the following specific objectives have been outlined:

1. Propose a mapping and consolidation algorithm for VMs that improves energy consumption in data centres, increases resource utilisation, and reduces the number of VM migrations.
2. Develop an independent task scheduling algorithm that enhances the makespan, improves resource utilisation, and reduces waiting times.
3. Create a dependent task scheduling algorithm for scientific workflow applications that improves the makespan and reduces the cost of using cloud resources.

1.8 Research Scope

The research aims to develop and evaluate algorithms for efficient resource utilisation in cloud computing environments. The scope of this research includes mapping and consolidation algorithms, cloudlet scheduling, and scientific workflow job scheduling, as follows.

1.8.1 Mapping and Consolidation Algorithms

The research aims to develop and evaluate mapping and consolidation algorithms that utilise resources in cloud computing environments effectively. The datasets used for this research are the data provided as a part of the CoMon project and benchmarking papers (Li et al., 2019; Liu et al., 2019; Li, 2019; Beloglazov et al., 2012; Kurdi et al., 2018). The data for March 22nd, 2011 are used as the benchmark. The proposed algorithms will be evaluated using 1516 VMs and 800 servers. The sub-objectives specifically for this section are:

- To minimise the number of active servers needed to run a set of VMs
- To minimise the energy consumption of the data centre
- To minimise the migration overhead between servers

1.8.2 Cloudlet Scheduling (Independent Tasks)

The research focuses on a cloudlet scheduling algorithm, which aims to minimise makespan, task waiting time, and resource utilisation. The algorithm considers tasks with a length of up to 20,000 MI and has limitations as shown in Table 5.3 and Table 5.4. This research considers only the scheduling of independent tasks using a non-preemptive scheduling allocation policy.

1.8.3 Scientific Workflow Job Scheduling (Dependent Tasks)

The research focuses on scientific workflow job scheduling, which utilises dependent tasks (i.e., tasks with relations among them, with parent and child tasks). This research aims to improve the scheduling and execution of dependent tasks in scientific workflows.

The proposed schemes rely on the open-source simulation toolkit of cloud computing called CloudSim, which is widely used in the area of mapping, consolidation, scheduling, etc. in cloud computing by several researchers (Kurdi et al., 2018; Li et al., 2019; Liu et al., 2019; Panwar et al., 2019). The research uses a CloudSim extension called WorkflowSim.

1.9 Thesis Organisation

The rest of this thesis is organised as follows: Chapter 2 presents the literature review and discusses the current state-of-the-art in cloud computing. Chapter 3 describes the research methodology used in this thesis including the research framework, experimental setup, proposed methods, performance metrics and the evaluation method. Chapter 4 explains the proposed mapping and consolidation of VMs using locust-inspired algorithms, which are designed to achieve green cloud computing. Chapter 5 applies the locust algorithm to the area of cloudlet scheduling. Chapter 6 presents the optimisation of scientific workflow job scheduling when applying the locust-inspired algorithm. Finally, Chapter 7 concludes the thesis, presents the thesis contributions, and discusses future research and open issues for researchers in this area.

REFERENCES

Abbott, B. P., Abbott, R., Adhikari, R., Ajith, P., Allen, B., Allen, G., Amin, R. S., Anderson, S. B., Anderson, W. G., Arain, M. A., and et al. (2009). LIGO: the laser interferometer gravitational-wave observatory. *Reports on Progress in Physics*, 72(7):076901.

Abdullahi, M., Ngadi, M. A., et al. (2016). Symbiotic organism search optimization based task scheduling in cloud computing environment. *Future Generation Computer Systems*, 56:640–650.

Abrishami, S., Naghibzadeh, M., and Epema, D. H. (2013). Deadline-constrained workflow scheduling algorithms for infrastructure as a service clouds. *Future generation computer systems*, 29(1):158–169.

Abrol, P., Gupta, S., and Singh, S. (2020). A qos aware resource placement approach inspired on the behavior of the social spider mating strategy in the cloud environment. *Wireless Personal Communications*, 113(4):2027–2065.

Agarwal, M. and Srivastava, G. M. S. (2021). Opposition-based learning inspired particle swarm optimization (opso) scheme for task scheduling problem in cloud computing. *Journal of Ambient Intelligence and Humanized Computing*, 12(10):9855–9875.

Ahmad, Z., Jehangiri, A. I., Ala'anzy, M. A., Othman, M., and Umar, A. I. (2021a). Fault-tolerant and data-intensive resource scheduling and management for scientific applications in cloud computing. *Sensors*, 21(21):7238.

Ahmad, Z., Nazir, B., and Umer, A. (2021b). A fault-tolerant workflow management system with quality-of-service-aware scheduling for scientific workflows in cloud computing. *International Journal of Communication Systems*, 34(1):e4649.

Akoush, S., Sohan, R., Rice, A., Moore, A. W., and Hopper, A. (2010). Predicting the performance of virtual machine migration. In *2010 IEEE international symposium on modeling, analysis and simulation of computer and telecommunication systems*, pp. 37–46. IEEE.

Al-Maamari, A. and Omara, F. A. (2015). Task scheduling using pso algorithm in cloud computing environments. *International Journal of Grid and Distributed Computing*, 8(5):245–256.

Al Nuaimi, K., Mohamed, N., Al Nuaimi, M., and Al-Jaroodi, J. (2012). A survey of load balancing in cloud computing: Challenges and algorithms. In *2012 second symposium on network cloud computing and applications*, pp. 137–142. IEEE.

Albagli-Kim, S., Schieber, B., Shachnai, H., and Tamir, T. (2016). Real-time k-bounded preemptive scheduling. In *2016 Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pp. 127–137. SIAM.

Albagli-Kim, S., Shachnai, H., and Tamir, T. (2014). Scheduling jobs with dwindling resource requirements in clouds. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pp. 601–609. IEEE.

Alboaneen, D., Tianfield, H., Zhang, Y., and Pranggono, B. (2021). A metaheuristic method for joint task scheduling and virtual machine placement in cloud data centers. *Future Generation Computer Systems*, 115:201–212.

Alghamdi, M. I. (2022). Optimization of load balancing and task scheduling in cloud computing environments using artificial neural networks-based binary particle swarm optimization (bpso). *Sustainability*, 14(19):11982.

Alhassan, S. and Abdulghani, M. (2019). A bio-inspired algorithm for virtual machines allocation in public clouds. *Procedia Computer Science*, 151:1072–1077.

Aljumah, A. and Ahanger, T. A. (2020). Cyber security threats, challenges and defence mechanisms in cloud computing. *IET Communications*, 14(7):1185–1191.

Alkhanak, E. N., Lee, S. P., Rezaei, R., and Parizi, R. M. (2016). Cost optimization approaches for scientific workflow scheduling in cloud and grid computing: A review, classifications, and open issues. *Journal of Systems and Software*, 113:1–26.

Amalarethinam, D. G. and Beena, T. L. A. (2015). Customer facilitated cost-based scheduling (cfcsc) in cloud. *Procedia Computer Science*, 46:660–667.

Apostu, A., Puican, F., Ularu, G., Suciu, G., Todoran, G., et al. (2013). Study on advantages and disadvantages of cloud computing–the advantages of telemetry applications in the cloud. *Recent advances in applied computer science and digital services*, 2103.

Ariel, G. and Ayali, A. (2015). Locust collective motion and its modeling. *PLoS Computational Biology*, 11(12):e1004522.

Aslam, S., ul Islam, S., Khan, A., Ahmed, M., Akhundzada, A., and Khan, M. K. (2017). Information collection centric techniques for cloud resource management: Taxonomy, analysis and challenges. *Journal of Network and Computer Applications*, 100:80–94.

Awad, A., El-Hefnawy, N., and Abdel_kader, H. (2015). Enhanced particle swarm optimization for task scheduling in cloud computing environments. *Procedia Computer Science*, 65:920–929.

Baccarelli, E., Scarpiniti, M., and Momenzadeh, A. (2018). Fog-supported delay-constrained energy-saving live migration of vms over multipath tcp/ip 5g connections. *IEEE Access*, 6:42327–42354.

Bagrodia, R., Meyer, R., Takai, M., Chen, Y.-a., Zeng, X., Martin, J., and Song, H. Y. (1998). Parsec: A parallel simulation environment for complex systems. *Computer*, 31(10):77–85.

Barhate, S. and Dhore, M. (2023). Evaluating data migrations with respect to interoperability in hybrid cloud. In *Machine Learning, Image Processing, Network Security and Data Sciences: Select Proceedings of 3rd International Conference on MIND 2021*, pp. 795–804. Springer.

Barroso, L. A. and Hölzle, U. (2007). The case for energy-proportional computing. *Computer*, 40(12):33–37.

Bashir, S., Mustafa, S., Ahmad, R. W., Shuja, J., Maqsood, T., and Alourani, A. (2022). Multi-factor nature inspired sla-aware energy efficient resource management for cloud environments. *Cluster Computing*, pp. 1–16.

Baxodirjonovich, K. N. and Choe, T.-Y. (2015). Dynamic task scheduling algorithm based on ant colony scheme. *International Journal of Engineering and Technology (IJET)*, 7(4):1163–1172.

Beloglazov, A., Abawajy, J., and Buyya, R. (2012). Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future generation computer systems*, 28(5):755–768.

Beloglazov, A. and Buyya, R. (2010a). Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers. *MGC@ Middleware*, 4(10.1145):1890799–803.

Beloglazov, A. and Buyya, R. (2010b). Energy efficient allocation of virtual machines in cloud data centers. In *2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, pp. 577–578. IEEE.

Beloglazov, A. and Buyya, R. (2010c). Energy efficient resource management in virtualized cloud data centers. In *2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, pp. 826–831. IEEE.

Beloglazov, A. and Buyya, R. (2012a). Managing overloaded hosts for dynamic consolidation of virtual machines in cloud data centers under quality of service constraints. *IEEE transactions on parallel and distributed systems*, 24(7):1366–1379.

Beloglazov, A. and Buyya, R. (2012b). Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency and Computation: Practice and Experience*, 24(13):1397–1420.

Bharathi, S., Chervenak, A., Deelman, E., Mehta, G., Su, M.-H., and Vahi, K. (2008). Characterization of scientific workflows. In *2008 third workshop on workflows in support of large-scale science*, pp. 1–10. IEEE.

Bhatt, A., Dimri, P., and Aggarwal, A. (2020). Self-adaptive brainstorming for jobshop scheduling in multicloud environment. *Software: Practice and Experience*, 50(8):1381–1398.

Bhattacherjee, S., Das, R., Khatua, S., and Roy, S. (2020). Energy-efficient migration techniques for cloud environment: a step toward green computing. *The Journal of Supercomputing*, 76(7):5192–5220.

Blackburn, M. and Hawkins, A. (2013). Unused server survey results analysis. *The Green Grid report*. Available online: `https://silo.tips/download/unused-servers-survey-results-analysis` (accessed 16 Mar. 2022).

Blythe, J., Jain, S., Deelman, E., Gil, Y., Vahi, K., Mandal, A., and Kennedy, K. (2005). Task scheduling strategies for workflow-based applications in grids. In *CCGrid 2005. IEEE International Symposium on Cluster Computing and the Grid, 2005.*, volume 2, pp. 759–767. IEEE.

Botta, A., De Donato, W., Persico, V., and Pescapé, A. (2016). Integration of cloud computing and internet of things: a survey. *Future generation computer systems*, 56:684–700.

Braun, T. D., Siegel, H. J., Beck, N., Bölöni, L. L., Maheswaran, M., Reuther, A. I., Robertson, J. P., Theys, M. D., Yao, B., Hensgen, D., et al. (2001). A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems. *Journal of Parallel and Distributed computing*, 61(6):810–837.

Buyya, R. and Murshed, M. (2002). Gridsim: A toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing. *Concurrency and computation: practice and experience*, 14(13-15):1175–1220.

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., and Brandic, I. (2009). Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation computer systems*, 25(6):599–616.

Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A., and Buyya, R. (2011). Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and experience*, 41(1):23–50.

Callaghan, S., Deelman, E., Gunter, D., Juve, G., Maechling, P., Brooks, C., Vahi, K., Milner, K., Graves, R., Field, E., et al. (2010). Scaling up workflow-based applications. *Journal of Computer and System Sciences*, 76(6):428–446.

Callaghan, S., Maechling, P., Small, P., Milner, K., Juve, G., Jordan, T. H., Deelman, E., Mehta, G., Vahi, K., Gunter, D., et al. (2011). Metrics for heterogeneous scientific workflows: A case study of an earthquake science application. *The International Journal of High Performance Computing Applications*, 25(3):274–285.

Cao, G. (2019). Topology-aware multi-objective virtual machine dynamic consolidation for cloud datacenter. *Sustainable Computing: Informatics and Systems*, 21:179–188.

Cao, R., Tang, Z., Li, K., and Li, K. (2018). Hmgowm: A hybrid decision mechanism for automating migration of virtual machines. *IEEE Transactions on Services Computing*.

Casas, I., Taheri, J., Ranjan, R., Wang, L., and Zomaya, A. Y. (2018). Ga-eti: An enhanced genetic algorithm for the scheduling of scientific workflows in cloud environments. *Journal of computational science*, 26:318–331.

Chakraborty, D., Mankar, V. V., and Nanavati, A. A. (2007). Enabling runtime adaptation ofworkflows to external events in enterprise environments. In *IEEE International Conference on Web Services (ICWS 2007)*, pp. 1112–1119. IEEE.

Chakravarthi, K. K., Shyamala, L., and Vaidehi, V. (2020). Topsis inspired cost-efficient concurrent workflow scheduling algorithm in cloud. *Journal of King Saud University-Computer and Information Sciences*.

Changtian, Y. and Jiong, Y. (2012). Energy-aware genetic algorithms for task scheduling in cloud computing. In *2012 Seventh ChinaGrid Annual Conference*, pp. 43–48. IEEE.

Charband, Y. and Navimipour, N. J. (2016). Online knowledge sharing mechanisms: a systematic review of the state of the art literature and recommendations for future research. *Information Systems Frontiers*, 18(6):1131–1151.

Chaudhary, N. and Kalra, M. (2017). An improved harmony search algorithm with group technology model for scheduling workflows in cloud environment. In *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, pp. 73–77. IEEE.

Chen, W. and Deelman, E. (2012). Workflowsim: A toolkit for simulating scientific workflows in distributed environments. In *2012 IEEE 8th international conference on E-science*, pp. 1–8. IEEE.

Cho, K.-M., Tsai, P.-W., Tsai, C.-W., and Yang, C.-S. (2015). A hybrid metaheuristic algorithm for vm scheduling with load balancing in cloud computing. *Neural Computing and Applications*, 26(6):1297–1309.

Clark, C., Fraser, K., Hand, S., Hansen, J. G., Jul, E., Limpach, C., Pratt, I., and Warfield, A. (2005). Live migration of virtual machines. In *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation-Volume 2*, pp. 273–286.

Dai, Y., Lou, Y., and Lu, X. (2015). A task scheduling algorithm based on genetic algorithm and ant colony optimization algorithm with multi-qos constraints in cloud computing. In *2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics*, volume 2, pp. 428–431. IEEE.

Dasgupta, K., Mandal, B., Dutta, P., Mandal, J. K., and Dam, S. (2013). A genetic algorithm (ga) based load balancing strategy for cloud computing. *Procedia Technology*, 10:340–347.

Davidović, T., Šelmić, M., Teodorović, D., and Ramljak, D. (2012). Bee colony optimization for scheduling independent tasks to identical processors. *Journal of heuristics*, 18(4):549–569.

Deelman, E., Blythe, J., Gil, Y., Kesselman, C., Mehta, G., Patil, S., Su, M.-H., Vahi, K., and Livny, M. (2004). Pegasus: Mapping scientific workflows onto the grid. In *European Across Grids Conference*, pp. 11–20. Springer.

Deelman, E., Singh, G., Livny, M., Berriman, B., and Good, J. (2008). The cost of doing science on the cloud: the montage example. In *SC'08: Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, pp. 1–12. IEEE. Austin, TX, USA, 15–21 November 2008.

Deelman, E., Vahi, K., Juve, G., Rynge, M., Callaghan, S., Maechling, P. J., Mayani, R., Chen, W., Da Silva, R. F., Livny, M., et al. (2015). Pegasus, a workflow management system for science automation. *Future Generation Computer Systems*, 46:17–35.

Deng, W., Liu, F., Jin, H., Liao, X., and Liu, H. (2014). Reliability-aware server consolidation for balancing energy-lifetime tradeoff in virtualized cloud datacenters. *International Journal of Communication Systems*, 27(4):623–642.

Devi, K. L. and Valli, S. (2021). Multi-objective heuristics algorithm for dynamic resource scheduling in the cloud computing environment. *The Journal of Supercomputing*, 77:8252–8280.

Diallo, O., Rodrigues, J. J., Sene, M., and Niu, J. (2014). Real-time query processing optimization for cloud-based wireless body area networks. *Information Sciences*, 284:84–94.

Dimitri, N. (2020). Pricing cloud iaas computing services. *Journal of Cloud Computing*, 9(1):1–11.

Distefano, S., Merlino, G., and Puliafito, A. (2015). A utility paradigm for iot: The sensing cloud. *Pervasive and mobile computing*, 20:127–144.

Domanal, S. G., Guddeti, R. M. R., and Buyya, R. (2017). A hybrid bio-inspired algorithm for scheduling and resource management in cloud environment. *IEEE transactions on services computing*, 13(1):3–15.

Domanal, S. G. and Reddy, G. R. M. (2018). An efficient cost optimized scheduling for spot instances in heterogeneous cloud environment. *Future Generation Computer Systems*, 84:11–21.

Dordaie, N. and Navimipour, N. J. (2018). A hybrid particle swarm optimization and hill climbing algorithm for task scheduling in the cloud environments. *ICT Express*, 4(4):199–202.

Dorigo, M. and Di Caro, G. (1999). Ant colony optimization: a new metaheuristic. In *Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406)*, volume 2, pp. 1470–1477. IEEE.

Duffield, N. G., Goyal, P., Greenberg, A., Mishra, P., Ramakrishnan, K., and Van der Merwe, J. E. (2002). Resource management with hoses: point-to-cloud services for virtual private networks. *IEEE/ACM Transactions on Networking*, 10(5):679–692.

EIA, U. (2020). Data center energy consumption. Technical report, U.S. Energy Information Administration. Available online: `https://www.eia.gov/electricity/sales_revenue_price/data.php` (accessed 23 Jen. 2023).

El-Abd, M. (2017). Global-best brain storm optimization algorithm. *Swarm and evolutionary computation*, 37:27–44.

Elmougy, S., Sarhan, S., and Joundy, M. (2017). A novel hybrid of shortest job first and round robin with dynamic variable quantum time task scheduling technique. *Journal of Cloud computing*, 6(1):1–12.

Fard, H. M., Prodan, R., Barrionuevo, J. J. D., and Fahringer, T. (2012). A multiobjective approach for workflow scheduling in heterogeneous environments. In *2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012)*, pp. 300–309. IEEE.

Farid, M., Latip, R., Hussin, M., and Hamid, N. A. W. A. (2020). Scheduling scientific workflow using multi-objective algorithm with fuzzy resource utilization in multi-cloud environment. *IEEE Access*, 8:24309–24322.

Feller, E., Rilling, L., and Morin, C. (2011). Energy-aware ant colony based workload placement in clouds. In *Proceedings of the 2011 IEEE/ACM 12th International Conference on Grid Computing*, pp. 26–33. IEEE Computer Society.

Ferdaus, M. H., Murshed, M., Calheiros, R. N., and Buyya, R. (2017). An algorithm for network and data-aware placement of multi-tier applications in cloud data centers. *Journal of Network and Computer Applications*, 98:65–83.

Ferreto, T. C., Netto, M. A., Calheiros, R. N., and De Rose, C. A. (2011). Server consolidation with migration control for virtualized data centers. *Future Generation Computer Systems*, 27(8):1027–1034.

Filiposka, S., Mishev, A., and Gilly, K. (2019). Multidimensional hierarchical vm migration management for hpc cloud environments. *The Journal of Supercomputing*, 75(8):5324–5346.

Garg, M., Kaur, A., and Dhiman, G. (2021). A novel resource allocation and scheduling based on priority using metaheuristic for cloud computing environment. In *Impacts and Challenges of Cloud Business Intelligence*, pp. 113–134. IGI Global.

Gharehpasha, S. and Masdari, M. (2021). A discrete chaotic multi-objective scaalo optimization algorithm for an optimal virtual machine placement in cloud data center. *Journal of Ambient Intelligence and Humanized Computing*, 12:9323–9339.

Ghazouani, S. and Slimani, Y. (2017). A survey on cloud service description. *Journal of Network and Computer Applications*, 91:61–74.

Ghomi, E. J., Rahmani, A. M., and Qader, N. N. (2017). Load-balancing algorithms in cloud computing: A survey. *Journal of Network and Computer Applications*, 88:50–71.

Ghumman, N. S. and Kaur, R. (2015). Dynamic combination of improved maxmin and ant colony algorithm for load balancing in cloud system. In *2015 6th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–5. IEEE.

Goyal, S. and Verma, M. K. (2016). Load balancing techniques in cloud computing environment: a review. *International Journal of Advanced Research in Computer Science and Software Engineering*, 6(4):583–588.

Guo, L., Hu, G., Dong, Y., Luo, Y., and Zhu, Y. (2017). A game based consolidation method of virtual machines in cloud data centers with energy and load constraints. *IEEE Access*, 6:4664–4676.

Guo, W., Kuang, P., Jiang, Y., Xu, X., and Tian, W. (2019). Save: self-adaptive consolidation of virtual machines for energy efficiency of cpu-intensive applications in the cloud. *The Journal of Supercomputing*, 75(11):7076–7100.

Gutierrez-Garcia, J. O. and Ramirez-Nafarrate, A. (2015). Agent-based load balancing in cloud data centers. *Cluster Computing*, 18(3):1041–1062.

Guttal, V., Romanczuk, P., Simpson, S. J., Sword, G. A., and Couzin, I. D. (2012). Cannibalism can drive the evolution of behavioural phase polyphenism in locusts. *Ecology letters*, 15(10):1158–1166.

Habib, S. M., Ries, S., and Muhlhauser, M. (2010). Cloud computing landscape and research challenges regarding trust and reputation. In *2010 7th International conference on ubiquitous intelligence & computing and 7th international conference on autonomic & trusted computing*, pp. 410–415. IEEE.

Hähnel, M., Martinovic, J., Scheithauer, G., Fischer, A., Schill, A., and Dargie, W. (2018). Extending the cutting stock problem for consolidating services with stochastic workloads. *IEEE Transactions on Parallel and Distributed Systems*, 29(11):2478–2488.

Hai, T., Zhou, J., Jawawi, D., Wang, D., Oduah, U., Biamba, C., and Jain, S. K. (2023). Task scheduling in cloud environment: optimization, security prioritization and processor selection schemes. *Journal of Cloud Computing*, 12(1):1–12.

Haidri, R. A., Katti, C. P., and Saxena, P. C. (2020). Cost effective deadline aware scheduling strategy for workflow applications on virtual machines in cloud computing. *Journal of King Saud University-Computer and Information Sciences*, 32(6):666–683.

Hall, J., Hartline, J., Karlin, A. R., Saia, J., and Wilkes, J. (2001). On algorithms for efficient data migration. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, pp. 620–629.

Houssein, E. H., Gad, A. G., Wazery, Y. M., and Suganthan, P. N. (2021). Task scheduling in cloud computing based on meta-heuristics: review, taxonomy, open challenges, and future trends. *Swarm and Evolutionary Computation*, 62:100841.

HpProLiantMl110G4Xeon3040 (Jan. 2011). Standard performance evaluation corporation (spec), hewlett-packard company proliant ml110 g3 (historical). Retrieved 16 March 2022 from `http://www.spec.org/power_ssj2008/results/res2011q1/power_ssj2008-20110127-00342.html`.

HpProLiantMl110G5Xeon3075 (Jan. 2011). The standard performance evaluation corporation (spec), hewlett-packard company proliant ml110 g5. Retrieved 16 March 2022 from `http://www.spec.org/power_ssj2008/results/res2011q1/power_ssj2008-20110124-00339.html`.

Hwang, K., Dongarra, J., and Fox, G. C. (2013). *Distributed and cloud computing: from parallel processing to the internet of things*. Morgan kaufmann.

Ivanisenko, I. N. and Radivilova, T. A. (2015). Survey of major load balancing algorithms in distributed system. In *2015 information technologies in innovation business conference (ITIB)*, pp. 89–92. IEEE.

Jadeja, Y. and Modi, K. (2012). Cloud computing-concepts, architecture and challenges. In *2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET)*, pp. 877–880. IEEE.

Jain, R. and Sharma, N. (2022). A quantum inspired hybrid ssa–gwo algorithm for sla based task scheduling to improve qos parameter in cloud computing. *Cluster Computing*, pp. 1–24.

Jena, R. (2017). Task scheduling in cloud environment: A multi-objective abc framework. *Journal of Information and Optimization Sciences*, 38(1):1–19.

Jiang, C., Wu, J., and Li, Z. (2019). Adaptive thresholds determination for saving cloud energy using three-way decisions. *Cluster Computing*, 22(4):8475–8482.

Jiang, Y., Huang, Z., and Tsang, D. H. (2016). Towards max-min fair resource allocation for stream big data analytics in shared clouds. *IEEE Transactions on Big Data*, 4(1):130–137.

Kamalinia, A. and Ghaffari, A. (2017). Hybrid task scheduling method for cloud computing by genetic and de algorithms. *Wireless personal communications*, 97(4):6301–6323.

Kanakala, V. R., Reddy, V. K., and Karthik, K. (2015). Performance analysis of load balancing techniques in cloud computing environment. In *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pp. 1–6. IEEE.

Kansal, N. J. and Chana, I. (2016). Energy-aware virtual machine migration for cloud computing-a firefly optimization approach. *Journal of Grid Computing*, 14(2):327–345.

Kaur, A., Singh, P., Singh Batth, R., and Peng Lim, C. (2022). Deep-q learning-based heterogeneous earliest finish time scheduling algorithm for scientific workflows in cloud. *Software: Practice and Experience*, 52(3):689–709.

Kaur, R. and Luthra, P. (2012). Load balancing in cloud computing. In *Proceedings of international conference on recent trends in information, telecommunication and computing, ITC*. Citeseer.

Khiyaita, A., El Bakkali, H., Zbakh, M., and El Kettani, D. (2012). Load balancing cloud computing: state of art. In *2012 National Days of Network Security and Systems*, pp. 106–109. IEEE.

Khodayarseresht, E. and Shameli-Sendi, A. (2023). A multi-objective cloud energy optimizer algorithm for federated environments. *Journal of Parallel and Distributed Computing*, 174:81–99.

Kirchoff, D. F., Xavier, M., Mastella, J., and De Rose, C. A. (2019). A preliminary study of machine learning workload prediction techniques for cloud applications. In *2019 27th Euromicro international conference on parallel, Distributed and Network-Based Processing (PDP)*, pp. 222–227. IEEE.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., and Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, 51(1):7–15.

Kommeri, J., Niemi, T., and Nurminen, J. K. (2017). Energy efficiency of dynamic management of virtual cluster with heterogeneous hardware. *The Journal of Supercomputing*, 73(5):1978–2000.

Konjaang, J. K. and Xu, L. (2021). Multi-objective workflow optimization strategy (mowos) for cloud computing. *Journal of Cloud Computing*, 10(1):1–19.

Koomey, J. et al. (2011). Growth in data center electricity use 2005 to 2010. *A report by Analytical Press, completed at the request of The New York Times*, 9(2011):161.

Koot, M. and Wijnhoven, F. (2021). Usage impact on data center electricity needs: A system dynamic forecasting model. *Applied Energy*, 291:116798. doi.org/10.1016/j.apenergy.2021.116798.

Kousalya, G., Balakrishnan, P., and Pethuru Raj, C. (2017). Workflow modeling and simulation techniques. In *Automated Workflow Scheduling in Self-Adaptive Clouds*, pp. 85–101. Springer.

Kulkarni, A. K. and Annappa, B. (2019). Context aware vm placement optimization technique for heterogeneous iaas cloud. *IEEE access*, 7:89702–89713.

Kumar, G. G. and Vivekanandan, P. (2019). Energy efficient scheduling for cloud data centers using heuristic based migration. *Cluster Computing*, 22(6):14073–14080.

Kumar, K. P. and Kousalya, K. (2020). Amelioration of task scheduling in cloud computing using crow search algorithm. *Neural Computing and Applications*, 32(10):5901–5907.

Kumar, P. M., Manogaran, G., Sundarasekar, R., Chilamkurti, N., Varatharajan, R., et al. (2018). Ant colony optimization algorithm with internet of vehicles for intelligent traffic control system. *Computer Networks*, 144:154–162.

Kumar, R., Kumar, A., and Sharma, A. (2016). A bio-inspired approach for power and performance aware resource allocation in clouds. In *MATEC Web of Conferences*, volume 57, p. 02008. EDP Sciences.

Kupiainen, E., Mäntylä, M. V., and Itkonen, J. (2015). Using metrics in agile and lean software development–a systematic literature review of industrial studies. *Information and software technology*, 62:143–163.

Kurdi, H. A., Alismail, S. M., and Hassan, M. M. (2018). Lace: a locust-inspired scheduling algorithm to reduce energy consumption in cloud datacenters. *IEEE Access*, 6:35435–35448.

Kushwah, V. S., Goyal, S. K., and Sharma, A. (2019). Meta-heuristic techniques study for fault tolerance in cloud computing environment: a survey work. In *Soft computing: theories and applications*, pp. 1–11. Springer.

Law, A. M., Kelton, W. D., and Kelton, W. D. (2007). *Simulation modeling and analysis*, volume 3. Mcgraw-hill New York.

LD, D. B. and Krishna, P. V. (2013). Honey bee behavior inspired load balancing of tasks in cloud computing environments. *Applied soft computing*, 13(5):2292–2303.

Lee, Z.-J., Su, S.-F., Chuang, C.-C., and Liu, K.-H. (2008). Genetic algorithm with ant colony optimization (ga-aco) for multiple sequence alignment. *Applied Soft Computing*, 8(1):55–78.

Li, H., Zhu, G., Cui, C., Tang, H., Dou, Y., and He, C. (2016). Energy-efficient migration and consolidation algorithm of virtual machines in data centers for cloud computing. *Computing*, 98(3):303–317.

Li, H., Zhu, G., Zhao, Y., Dai, Y., and Tian, W. (2017a). Energy-efficient and qos-aware model based resource consolidation in cloud data centers. *Cluster Computing*, 20(3):2793–2803.

Li, L., Dong, J., Zuo, D., and Wu, J. (2019). Sla-aware and energy-efficient vm consolidation in cloud data centers using robust linear regression prediction model. *IEEE Access*, 7:9490–9500.

Li, R., Zheng, Q., Li, X., and Yan, Z. (2020). Multi-objective optimization for rebalancing virtual machine placement. *Future Generation Computer Systems*, 105:824–842.

Li, X., Jiang, X., Garraghan, P., and Wu, Z. (2018a). Holistic energy and failure aware workload scheduling in cloud datacenters. *Future Generation Computer Systems*, 78:887–900.

Li, Y., Zhu, Z., and Wang, Y. (2018b). Min-max-min: A heuristic scheduling algorithm for jobs across geo-distributed datacenters. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pp. 1573–1574. IEEE.

Li, Z. (2019). An adaptive overload threshold selection process using markov decision processes of virtual machine in cloud data center. *Cluster Computing*, 22(2):3821–3833.

Li, Z., Ge, J., Hu, H., Song, W., Hu, H., and Luo, B. (2015). Cost and energy aware scheduling algorithm for scientific workflows with deadline constraint in clouds. *IEEE Transactions on Services Computing*, 11(4):713–726.

Li, Z., Yan, C., Yu, L., and Yu, X. (2018c). Energy-aware and multi-resource overload probability constraint-based virtual machine dynamic consolidation method. *Future Generation Computer Systems*, 80:139–156.

Li, Z., Yan, C., Yu, X., and Yu, N. (2017b). Bayesian network-based virtual machines consolidation method. *Future Generation Computer Systems*, 69:75–87.

Lifka, D., Foster, I., Mehringer, S., Parashar, M., Redfern, P., Stewart, C., and Tuecke, S. (2013). Xsede cloud survey report. Technical report, The Extreme Science and Engineering Discovery Environment (XSEDE). Available online: `http://www.cac.cornell.edu/technologies/xsedecloudsurveyreport.pdf` (accessed 16 Mar. 2022).

Lin, W., Liang, C., Wang, J. Z., and Buyya, R. (2014). Bandwidth-aware divisible task scheduling for cloud computing. *Software: Practice and Experience*, 44(2):163–174.

Liu, C., Krishna, C., and Gong, W.-B. (1998). Real-time system evaluation: a rational interpolation approach. In *Proceedings of the 37th IEEE Conference on Decision and Control (Cat. No. 98CH36171)*, volume 2, pp. 1680–1685. IEEE.

Liu, F., Ma, Z., Wang, B., and Lin, W. (2019). A virtual machine consolidation algorithm based on ant colony system and extreme learning machine for cloud data center. *IEEE Access*, 8:53–67.

Liu, G., Li, J., and Xu, J. (2013). An improved min-min algorithm in cloud computing. In *Proceedings of the 2012 International Conference of Modern Computer Science and Applications*, pp. 47–52. Springer.

Liu, J., Pacitti, E., Valduriez, P., and Mattoso, M. (2015). A survey of data-intensive scientific workflow management. *Journal of Grid Computing*, 13(4):457–493.

Liu, X., Li, X., Shi, X., Huang, K., and Liu, Y. (2012). A multi-type ant colony optimization (maco) method for optimal land use allocation in large areas. *International Journal of Geographical Information Science*, 26(7):1325–1343.

Lovász, G., Niedermeier, F., and De Meer, H. (2013). Performance tradeoffs of energy-aware virtual machine consolidation. *Cluster Computing*, 16(3):481–496.

Maggiani, R. (2009). Cloud computing is changing how we communicate. In *2009 IEEE International Professional Communication Conference*, pp. 1–4. IEEE.

Malla, P. A. and Sheikh, S. (2023). Analysis of qos aware energy-efficient resource provisioning techniques in cloud computing. *International Journal of Communication Systems*, p. e5359.

Malladi, R. R. (2015). An approach to load balancing in cloud computing. *Int. J. Innov. Res. Sci., Eng. Technol*, 4(5):3769–3777.

Mangalampalli, S., Karri, G. R., and Rajkumar, K. V. (2023). Evmpcsa: Efficient vm packing mechanism in cloud computing using chaotic social spider algorithm. *Procedia Computer Science*, 218:554–562.

Manvi, S. S. and Shyam, G. K. (2014). Resource management for infrastructure as a service (iaas) in cloud computing: A survey. *Journal of network and computer applications*, 41:424–440.

Masadeh, R., Alsharman, N., Sharieh, A., Mahafzah, B. A., and Abdulrahman, A. (2021). Task scheduling on cloud computing based on sea lion optimization algorithm. *International Journal of Web Information Systems*, 17(2):99–116.

Masdari, M. and Khezri, H. (2020). Efficient vm migrations using forecasting techniques in cloud computing: a comprehensive review. *Cluster Computing*, pp. 1–30.

Masdari, M., ValiKardan, S., Shahi, Z., and Azar, S. I. (2016). Towards workflow scheduling in cloud computing: a comprehensive analysis. *Journal of Network and Computer Applications*, 66:64–82.

Mathew, T., Sekaran, K. C., and Jose, J. (2014). Study and analysis of various task scheduling algorithms in the cloud computing environment. In *2014 International conference on advances in computing, communications and informatics (ICACCI)*, pp. 658–664. IEEE.

Mekala, M. S. and Viswanathan, P. (2019). Energy-efficient virtual machine selection based on resource ranking and utilization factor approach in cloud computing for iot. *Computers & Electrical Engineering*, 73:227–244.

Mell, P. and Grance, T. (2011). The nist definition of cloud computing. *Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD*. Available online: `https://doi.org/10.6028/NIST.SP.800-145` (accessed 16 Mar. 2022).

Meng, X., Isci, C., Kephart, J., Zhang, L., Bouillet, E., and Pendarakis, D. (2010a). Efficient resource provisioning in compute clouds via vm multiplexing. In *Proceedings of the 7th international conference on Autonomic computing*, pp. 11–20.

Meng, X., Pappas, V., and Zhang, L. (2010b). Improving the scalability of data center networks with traffic-aware virtual machine placement. In *2010 Proceedings IEEE INFOCOM*, pp. 1–9. IEEE.

Milan, S. T., Rajabion, L., Darwesh, A., Hosseinzadeh, M., and Navimipour, N. J. (2020). Priority-based task scheduling method over cloudlet using a swarm intelligence algorithm. *Cluster Computing*, 23(2):663–671.

Milani, A. S. and Navimipour, N. J. (2016). Load balancing mechanisms and techniques in the cloud environments: Systematic literature review and future trends. *Journal of Network and Computer Applications*, 71:86–98.

Mishra, K. and Majhi, S. K. (2021). A binary bird swarm optimization based load balancing algorithm for cloud computing environment. *Open Computer Science*, 11(1):146–160.

Mishra, N. K., Himthani, P., and Dubey, G. P. (2021). Priority-based shortest job first broker policy for cloud computing environments. In *Proceedings of International Conference on Communication and Computational Technologies: ICCCT 2021*, pp. 279–290. Springer.

Moghaddam, F. F., Moghaddam, R. F., and Cheriet, M. (2015). Carbon-aware distributed cloud: multi-level grouping genetic algorithm. *Cluster Computing*, 18(1):477–491.

Mork, R., Martin, P., and Zhao, Z. (2015). Contemporary challenges for data-intensive scientific workflow management systems. In *Proceedings of the 10th Workshop on Workflows in Support of Large-Scale Science*, pp. 1–11.

Ms, S., PM, J. P., and Alappatt, V. (2020). Profit maximization based task scheduling in hybrid clouds using whale optimization technique. *Information Security Journal: A Global Perspective*, 29(4):155–168.

Muthulakshmi, B. and Somasundaram, K. (2019). A hybrid abc-sa based optimized scheduling and resource allocation for cloud environment. *Cluster Computing*, 22(5):10769–10777.

Naghibzadeh, M. (2016). Modeling and scheduling hybrid workflows of tasks and task interaction graphs on the cloud. *Future Generation Computer Systems*, 65:33–45.

Nashaat, H., Ashry, N., and Rizk, R. (2019). Smart elastic scheduling algorithm for virtual machine migration in cloud computing. *The Journal of Supercomputing*, 75(7):3842–3865.

Nasr, A. A., El-Bahnasawy, N. A., Attiya, G., and El-Sayed, A. (2019). Cost-effective algorithm for workflow scheduling in cloud computing under deadline constraint. *Arabian Journal for Science and Engineering*, 44(4):3765–3780.

Navimipour, N. J. and Charband, Y. (2016). Knowledge sharing mechanisms and techniques in project teams: Literature review, classification, and current trends. *Computers in Human Behavior*, 62:730–742.

Nazir, B. and Ahmad, Z. (2022). Dynamic energy efficient load balancing strategy for computational grid. *Concurrency and Computation: Practice and Experience*, 34(1):e6484.

Nazir, B. et al. (2018). Qos-aware vm placement and migration for hybrid cloud infrastructure. *The Journal of Supercomputing*, 74(9):4623–4646.

Oppenheimer, D., Ganapathi, A., and Patterson, D. A. (2003). Why do internet services fail, and what can be done about it? In *USENIX symposium on internet technologies and systems*, volume 67. Seattle, WA.

Palta, R. and Jeet, R. (2014). Load balancing in the cloud computing using virtual machine migration: A review. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 3(5):437–441.

Panwar, N., Negi, S., Rauthan, M. M. S., and Vaisla, K. S. (2019). Topsis–pso inspired non-preemptive tasks scheduling algorithm in cloud environment. *Cluster Computing*, 22(4):1379–1396.

Park, K. and Pai, V. S. (2006). Comon: a mostly-scalable monitoring system for planetlab. *ACM SIGOPS Operating Systems Review*, 40(1):65–74.

Paya, A. and Marinescu, D. C. (2015). Energy-aware load balancing and application scaling for the cloud ecosystem. *IEEE Transactions on Cloud Computing*, 5(1):15–27.

Poola, D., Ramamohanarao, K., and Buyya, R. (2014). Fault-tolerant workflow scheduling using spot instances on clouds. *Procedia Computer Science*, 29:523–533.

Pop, C. B., Anghel, I., Cioara, T., Salomie, I., and Vartic, I. (2012). A swarm-inspired data center consolidation methodology. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, pp. 1–7.

Preist, C. and Shabajee, P. (2010). Energy use in the media cloud: Behaviour change, or technofix? In *2010 IEEE Second International Conference on Cloud Computing Technology and Science*, pp. 581–586. IEEE.

Qian, J., Li, J., Ma, R., Lin, L., and Guan, H. (2019). Lg-ram: Load-aware global resource affinity management for virtualized multicore systems. *Journal of Systems Architecture*, 98:114–125.

Rahman, A. U., Khan, F. G., and Jadoon, W. (2016). Energy efficiency techniques in cloud computing. *International Journal of Computer Science and Information Security*, 14(6):317.

Ranjbari, M. and Torkestani, J. A. (2018). A learning automata-based algorithm for energy and sla efficient consolidation of virtual machines in cloud data centers. *Journal of Parallel and Distributed Computing*, 113:55–62.

Rathore, M., Rai, S., Saluja, N., et al. (2015). Load balancing of virtual machine using honey bee galvanizing algorithm in cloud. *IJCSIT*, 6(4):4128–4132.

Ray, S. and De Sarkar, A. (2012). Execution analysis of load balancing algorithms in cloud computing environment. *International Journal on Cloud Computing: Services and Architecture (IJCCSA)*, 2(5):1–13.

Riahi, M. and Krichen, S. (2018). A multi-objective decision support framework for virtual machine placement in cloud data centers: A real case study. *The Journal of Supercomputing*, 74(7):2984–3015.

Riedel, M., Wolf, F., Kranzlmüller, D., Streit, A., and Lippert, T. (2009). Research advances by using interoperable e-science infrastructures. *Cluster computing*, 12(4):357–372.

Rukmini, S. and Shridevi, S. (2023). An optimal solution to reduce virtual machine migration sla using host power. *Measurement: Sensors*, 25:100628.

Sadeeq, M. M., Abdulkareem, N. M., Zeebaree, S. R., Ahmed, D. M., Sami, A. S., and Zebari, R. R. (2021). Iot and cloud computing issues, challenges and opportunities: A review. *Qubahan Academic Journal*, 1(2):1–7.

Saeed, W., Ahmad, Z., Jehangiri, A. I., Mohamed, N., Umar, A. I., and Ahmad, J. (2021). A fault tolerant data management scheme for healthcare internet of things in fog computing. *KSII Transactions on Internet and Information Systems (TIIS)*, 15(1):35–57.

Sanaj, M. and Prathap, P. J. (2020). Nature inspired chaotic squirrel search algorithm (cssa) for multi objective task scheduling in an iaas cloud computing atmosphere. *Engineering Science and Technology, an International Journal*, 23(4):891–902.

Santhosh, B., PK, A., and Manjaiah, D. (2014). Comparative study of workflow scheduling algorithms in cloud computing. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(5).

Sargent, R. G. (2010). Verification and validation of simulation models. In *Proceedings of the 2010 winter simulation conference*, pp. 166–183. IEEE.

Şaykol, E., Başer, A. S., and Sari, K. (2016). Analyzing the important factors for cloud service provider selection among the it firms in turkey. In *International Conference of Eurasian Economies, Kaposvár, HUNGARY*.

Seth, S. and Singh, N. (2017). Dynamic threshold-based dynamic resource allocation using multiple vm migration for cloud computing systems. In *Information, Communication and Computing Technology: Second International Conference, ICICCT 2017, New Delhi, India, May 13, 2017, Revised Selected Papers 2*, pp. 106–116. Springer.

Sharma, M. and Garg, R. (2020). Higa: Harmony-inspired genetic algorithm for rack-aware energy-efficient task scheduling in cloud data centers. *Engineering Science and Technology, an International Journal*, 23(1):211–224.

Shehabi, A., Smith, S. J., Masanet, E., and Koomey, J. (2018). Data center growth in the united states: decoupling the demand for services from electricity use. *Environmental Research Letters*, 13(12):124030.

Shi, Y. (2011). Brain storm optimization algorithm. In *International conference in swarm intelligence*, pp. 303–309. Springer.

Shojafar, M., Javanmardi, S., Abolfazli, S., and Cordeschi, N. (2015). Fuge: A joint meta-heuristic approach to cloud job scheduling algorithm using fuzzy theory and a genetic method. *Cluster Computing*, 18(2):829–844.

Sidhu, A. K. and Kinger, S. (2013). Analysis of load balancing techniques in cloud computing. *International Journal of computers & technology*, 4(2):737–741.

Simão, J. and Veiga, L. (2014). Partial utility-driven scheduling for flexible sla and pricing arbitration in clouds. *IEEE Transactions on Cloud Computing*, 4(4):467–480.

Singh, H., Bhasin, A., and Kaveri, P. (2019). Secure: Efficient resource scheduling by swarm in cloud computing. *Journal of Discrete Mathematical Sciences and Cryptography*, 22(2):127–137.

Singh, L. and Singh, S. (2014). Deadline and cost based ant colony optimization algorithm for scheduling workflow applications in hybrid cloud. *Journal of Scientific & Engineering Research*, 5(10):1417–1420.

Singh, P., Baaga, P., and Gupta, S. (2016). Assorted load balancing algorithms in cloud computing: A survey. *International Journal of Computer Applications*, 143(7):34–40.

Snyder, B. (2010). Server virtualization has stalled, despite the hype. *InfoWorld*.

Soltani, Z. and Navimipour, N. J. (2016). Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research. *Computers in Human Behavior*, 61:667–688.

Song, W., Chen, F., Jacobsen, H.-A., Xia, X., Ye, C., and Ma, X. (2017). Scientific workflow mining in clouds. *IEEE Transactions on Parallel and Distributed Systems*, 28(10):2979–2992.

Song, W., Xiao, Z., Chen, Q., and Luo, H. (2013). Adaptive resource provisioning for the cloud using online bin packing. *IEEE Transactions on Computers*, 63(11):2647–2660.

Sooezi, N., Abrishami, S., and Lotfian, M. (2015). Scheduling data-driven workflows in multi-cloud environment. In *2015 IEEE 7th international conference on cloud computing technology and science (CloudCom)*, pp. 163–167. IEEE.

Stage, A. and Setzer, T. (2009). Network-aware migration control and scheduling of differentiated virtual machine workloads. In *2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing*, pp. 9–14. IEEE.

Sun, W., Ji, Z., Sun, J., Zhang, N., and Hu, Y. (2015). Saaco: a self adaptive ant colony optimization in cloud computing. In *2015 IEEE Fifth International Conference on Big Data and Cloud Computing*, pp. 148–153. IEEE.

Suresh, S. and Sakthivel, S. (2017). A novel performance constrained power management framework for cloud computing using an adaptive node scaling approach. *Computers & Electrical Engineering*, 60:30–44.

Takahashi, S., Nakada, H., Takefusa, A., Kudoh, T., Shigeno, M., and Yoshise, A. (2012). Virtual machine packing algorithms for lower power consumption. In *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*, pp. 161–168. IEEE.

Talwani, S. and Singla, J. (2021). Enhanced bee colony approach for reducing the energy consumption during vm migration in cloud computing environment. In *IOP Conference Series: Materials Science and Engineering*, volume 1022, p. 012069. IOP Publishing.

Talwani, S., Singla, J., Mathur, G., Malik, N., Jhanjhi, N., Masud, M., and Aljahdali, S. (2022). Machine-learning-based approach for virtual machine allocation and migration. *Electronics*, 11(19):3249.

Tawfeek, M. A., El-Sisi, A., Keshk, A. E., and Torkey, F. A. (2013). Cloud task scheduling based on ant colony optimization. In *2013 8th international conference on computer engineering & systems (ICCES)*, pp. 64–69. IEEE.

Tekawade, A. and Banerjee, S. (2023). Wanms: A makespan, energy, and reliability aware scheduling algorithm for workflow scheduling in multi-processor systems. In *Distributed Computing and Intelligent Technology: 19th International Conference, ICDCIT 2023, Bhubaneswar, India, January 18–22, 2023, Proceedings*, pp. 20–35. Springer.

Thakur, A. and Goraya, M. S. (2017). A taxonomic survey on load balancing in cloud. *Journal of Network and Computer Applications*, 98:43–57.

Tian, W., He, M., Guo, W., Huang, W., Shi, X., Shang, M., Toosi, A. N., and Buyya, R. (2018). On minimizing total energy consumption in the scheduling of virtual machine reservations. *Journal of Network and Computer Applications*, 113:64–74.

Tighe, M. and Bauer, M. (2017). Topology and application aware dynamic vm management in the cloud. *Journal of Grid Computing*, 15(2):273–294.

Tiwari, D. and Tiwari, D. (2014). An efficient hybrid sjf and priority based scheduling of jobs in cloud computing. *Int. J. Mod. Eng. Manag. Res*, 2(4):26.

Topaz, C. M., Bernoff, A. J., Logan, S., and Toolson, W. (2008). A model for rolling swarms of locusts. *The European Physical Journal Special Topics*, 157(1):93–109.

Tsakalozos, K., Verroios, V., Roussopoulos, M., and Delis, A. (2017). Live vm migration under time-constraints in share-nothing iaas-clouds. *IEEE Transactions on Parallel and Distributed Systems*, 28(8):2285–2298.

Ullah, A., Li, J., Shen, Y., and Hussain, A. (2018). A control theoretical view of cloud elasticity: taxonomy, survey and challenges. *Cluster Computing*, 21(4):1735–1764.

Vaquero, L. M., Rodero-Merino, L., Caceres, J., and Lindner, M. (2008). A break in the clouds: towards a cloud definition.

Varasteh, A. and Goudarzi, M. (2015). Server consolidation techniques in virtualized data centers: A survey. *IEEE Systems Journal*, 11(2):772–783.

Vasić, N., Barisits, M., Salzgeber, V., and Kostic, D. (2009). Making cluster applications energy-aware. In *Proceedings of the 1st workshop on Automated control for datacenters and clouds*, pp. 37–42.

Vemula, D. R., Morampudi, M. K., Maurya, S., Abdul, A., Hussain, M. M., and Kavati, I. (2022). Enhanced resource provisioning and migrating virtual machines in heterogeneous cloud data center. *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12.

Vijaya, C. and Srinivasan, P. (2020). A hybrid technique for server consolidation in cloud computing environment. *Cybernetics and Information Technologies*, 20(1):36–52.

Vogels, W. (2008). Beyond server consolidation: Server consolidation helps companies improve resource utilization, but virtualization can help in other ways, too. *Queue*, 6(1):20–26.

Vrbsky, S. V., Lei, M., Smith, K., and Byrd, J. (2010). Data replication and power consumption in data grids. In *2010 IEEE Second International Conference on Cloud Computing Technology and Science*, pp. 288–295. IEEE.

Wadhwa, H. and Aron, R. (2023). Optimized task scheduling and preemption for distributed resource management in fog-assisted iot environment. *The Journal of Supercomputing*, 79(2):2212–2250.

Wang, W., Chen, H., and Chen, X. (2012). An availability-aware virtual machine placement approach for dynamic scaling of cloud applications. In *2012 9th International Conference on Ubiquitous Intelligence and Computing and 9th International Conference on Autonomic and Trusted Computing*, pp. 509–516. IEEE.

Welke, S. R., Johnson, B. W., and Aylor, J. H. (1995). Reliability modeling of hardware/software systems. *IEEE Transactions on Reliability*, 44(3):413–418.

Wood, T., Tarasuk-Levin, G., Shenoy, P., Desnoyers, P., Cecchet, E., and Corner, M. D. (2009). Memory buddies: exploiting page sharing for smart colocation in virtualized data centers. *ACM SIGOPS Operating Systems Review*, 43(3):27–36.

Wu, C. Q., Lin, X., Yu, D., Xu, W., and Li, L. (2014). End-to-end delay minimization for scientific workflows in clouds under budget constraint. *IEEE Transactions on Cloud Computing*, 3(2):169–181.

Wu, H., Deng, S., Li, W., Yin, J., Yang, Q., Wu, Z., and Zomaya, A. Y. (2017). Revenue-driven service provisioning for resource sharing in mobile cloud computing. In *International Conference on Service-Oriented Computing*, pp. 625–640. Springer.

Xavier, V. A. and Annadurai, S. (2019). Chaotic social spider algorithm for load balance aware task scheduling in cloud computing. *Cluster Computing*, 22(1):287–297.

Xiao, Z., Song, W., and Chen, Q. (2012). Dynamic resource allocation using virtual machines for cloud computing environment. *IEEE transactions on parallel and distributed systems*, 24(6):1107–1117.

Yadav, S. B. S. and Kalra, M. (2020). Energy-aware vm migration in cloud computing. In *Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India*, pp. 353–364. Springer, Singapore.

Yan, H., Wang, H., Li, X., Wang, Y., Li, D., Zhang, Y., Xie, Y., Liu, Z., Cao, W., and Yu, F. (2016). Cost-efficient consolidating service for aliyun's cloud-scale computing. *IEEE Transactions on Services Computing*, 12(1):117–130.

Yang, C.-T., Chen, S.-T., Liu, J.-C., Chan, Y.-W., Chen, C.-C., and Verma, V. K. (2019). An energy-efficient cloud system with novel dynamic resource allocation methods. *The Journal of Supercomputing*, 75(8):4408–4429.

Yang, X.-S. (2014). *Nature-inspired optimization algorithms*. Elsevier.

Yonggui, W. and Ruilian, H. (2011). Study on cloud computing task schedule strategy based on maco algorithm. *Computer Measurement & Control*, 5.

Yu, J., Kirley, M., and Buyya, R. (2007). Multi-objective planning for workflow execution on grids. In *2007 8th IEEE/ACM International Conference on Grid Computing*, pp. 10–17. IEEE.

Yu, S., Wang, C., Ren, K., and Lou, W. (2010). Achieving secure, scalable, and fine-grained data access control in cloud computing. In *2010 Proceedings IEEE INFOCOM*, pp. 1–9. Ieee.

Zhou, Q., Xu, M., Gill, S. S., Gao, C., Tian, W., Xu, C., and Buyya, R. (2020). Energy efficient algorithms based on vm consolidation for cloud computing: comparisons and evaluations. In *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, pp. 489–498. IEEE.

Zhou, X., Zhang, G., Sun, J., Zhou, J., Wei, T., and Hu, S. (2019). Minimizing cost and makespan for workflow scheduling in cloud using fuzzy dominance sort based heft. *Future Generation Computer Systems*, 93:278–289.

Zhu, W., Zhuang, Y., and Zhang, L. (2017). A three-dimensional virtual resource scheduling method for energy saving in cloud computing. *Future Generation Computer Systems*, 69:66–74.

Zhu, X., Wang, J., Guo, H., Zhu, D., Yang, L. T., and Liu, L. (2016). Fault-tolerant scheduling for real-time scientific workflows with elastic resource provisioning in virtualized clouds. *IEEE Transactions on Parallel and Distributed Systems*, 27(12):3501–3517.

Zubair, A. A., Razak, S. A., Ngadi, M. A., Al-Dhaqm, A., Yafooz, W. M., Emara, A.-H. M., Saad, A., and Al-Aqrabi, H. (2022). A cloud computing-based modified symbiotic organisms search algorithm (ai) for optimal task scheduling. *Sensors*, 22(4):1674.

Zuo, L., Shu, L., Dong, S., Zhu, C., and Hara, T. (2015). A multi-objective optimization scheduling method based on the ant colony algorithm in cloud computing. *Ieee Access*, 3:2687–2699.