ENHANCED MARINE PREDATOR ALGORITHMS FOR TASK SCHEDULING IN CLOUD DATA CENTRES

NORFAZLIN BINTI RASHID

UNIVERSITI SAINS MALAYSIA

2025

ENHANCED MARINE PREDATOR ALGORITHMS FOR TASK SCHEDULING IN CLOUD DATA CENTRES

by

NORFAZLIN BINTI RASHID

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

May 2025

ACKNOWLEDGEMENT

All praises to Allah, The Most Gracious for the grace and abundance He provides.

I would like to express gratitude to my main supervisor, Associate Professor Dr. Umi Kalsom Yusof for her unwavering support and guidance throughout this journey. Her considerate and inspiring approach has been truly invaluable, fostering an environment of growth without ever feeling overly compelled. A big thank you also to Dr Suzi Iryanti for guiding me in the final leg of submission.

To my husband, Hizam, whose support and care have been my rock, I am endlessly grateful. My children, Aiman, Nafis, Syaurah and Wasim, serve as constant sources of warmth and love, providing solace during both happy and challenging times.

To my parents, their continuous support and prayers have been a pillar of my strength. A special mention is also due to my parents-in-law who have always been there for my children in any situation.

I extend my heartfelt appreciation to Universiti Malaysia Perlis (UniMAP) for their understanding and support, providing me with the opportunity to pursue this academic endeavour.

Last but not least, to my fellow friends, colleagues and everyone in Universiti Sains Malaysia who has extended their help and kind thoughts, I thank you all and hope that our paths will continue to cross in good terms.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

M	Total Number of Task
N	Total Number of VMs
CF	Adaptive Convergence Factor
v	High Velocity Ratio
R_L	Random Levy Number
R_B	Random Brownian Number
Top_p	The Best Fitness Value
TP	Vector of Top Predator Position
R	Vector of Uniform Random Numbers
φ	Selection Threshold
t	Current Iteration
T_{max}	Maximum Iteration
θ	Maximum Simulation Try
Py_i	Current Prey i^{th}
X_{min}	Lower Boundary Value
X_{max}	Upper Boundary Value

LIST OF ABBREVIATIONS

AWS	Amazon Web Services
ACO	Ant Colony Optimization
ABC	Artificial Bee Colony
CPU	Central Processing Unit
CSP	Cloud Service Providers
CRB	Comprehensive Resource Balance
CSA	Crow Search Algorithm
FA	Firefly Algorithm
FWA	Fireworks Algorithm
FCFS	First-Come First-Served
GA	Genetic Algorithm
GSA	Golden Sine Algorithm
GPU	Graphical Processing Units
GWO	Grey Wolf Optimizer
HDFS	Hadoop Distributed File System
HDD	Hard Disk Drives
HHO	Harris Hawks Optimization
IT	Information Technology
IaaS	Infrastructure-as-a-Service
I/O	Input/Output
MPA	Marine Predator Algorithm
MI	Million Instructions
MIPS	Million Instructions Per Second

NMRA	Naked Mole-Rat Algorithm
NAS	Network-Attached Storage
NP-hard	Non-Deterministic Polynomial Time-hard
NVMe	Non-Volatile Memory Express
PSO	Particle Swarm Optimization
PaaS	Platform-as-a-Service
RAM	Random Access Memory
RB	Random Brownian number
RDA	Resource Deep Analytics
RTT	Round-Trip Time
SLA	Service Level Agreement
SJF	Shortest Job First
SaaS	Software-as-a-Service
SPEC	Standard Performance Evaluation Corporation
SAN	Storage Area Network
SOA	Sunflower Optimization Algorithm
TPU	Tensor Processing Units
TSA	Tree Seed Algorithm
VM	Virtual Machine
WOA	Whale Optimization Algorithm

ALGORITMA PEMANGSA LAUTAN TERTINGKAT UNTUK PENJADUALAN TUGAS DI PUSAT DATA AWAN

ABSTRAK

Penjadualan tugas secara berketentuan dalam pusat data awan tradisional yang homogen adalah jelas dan mudah, namun peningkatan kepelbagaian sumber telah menjadikan proses penjadualan yang cekap semakin rumit. Kaedah berketentuan menghadapi kesukaran untuk memenuhi kekangan Kualiti Perkhidmatan (Quality of Service, QoS) kerana pemetaan tugas kepada mesin maya (VM) yang tidak tepat boleh menjejaskan prestasi, yang akhirnya menyebabkan sumber VM tidak dapat digunakan secara optimum, sama ada berlebihan atau tidak mencukupi. Bagi menangani cabaran ini, para penyelidik kini beralih kepada pendekatan metaheuristik, seperti Algoritma Pemangsa Lautan (MPA). Kajian ini mencadangkan satu kaedah penjadualan tugas berasaskan MPA yang diubah suai, bagi meminimumkan masa siap tugas dan *makespan*. Tiga varian telah diperkenalkan: penjadualan tugas MPA yang diubah suai (MMPACTS), MMPACTS dengan pengecaman kapasiti VM heterogen (MMPACTS-H), dan MMPACTS dengan pengimbangan beban (MMPACTS-HLB). MMPACTS mengubah suai MPA untuk tugas diskrit dan menambah baik algoritma tersebut dengan mutasi terpilih dan mutasi Cauchy. MMPACTS-H pula menggabungkan Indeks Kapasiti VM untuk mengambil kira perbezaan keupayaan VM, manakala MMPACTS-HLB menggunakan kaedah pemilihan VM berasaskan sistem undian loteri bagi mengimbang beban kerja. Penilaian prestasi dilaksanakan menggunakan set data simulasi serta data sebenar daripada Google Cluster Trace dan GoCJ, dengan mengambil kira *makespan*, tahap penggunaan sumber, dan darjah keseimbangan beban. MMPACTS berjaya mengurangkan *makespan* sebanyak 36% berbanding

kaedah berketentuan seperti Tugas Terpendek Dahulu (*Shortest Jobs First, SJF)* dan Masuk Dahulu, Layan Dahulu (*First-Come First-Served, FCFS),* serta 22% lebih baik berbanding pendekatan metaheuristik lain seperti Artificial Bees Colony (ABC) dan Particle Swarm Optimization (PSO). MMPACTS-H pula menunjukkan peningkatan tambahan sebanyak 12% dalam pengurangan *makespan*. MMPACTS-HLB meningkatkan imbangan beban, mengurangkan ketidakseimbangan sebanyak 47% berbanding SJF, 51% berbanding FCFS, 28% berbanding ABC dan 34% berbanding PSO. Keputusan ini membuktikan keberkesanan pendekatan yang dicadangkan dalam persekitaran awan yang heterogen. Secara keseluruhannya, kajian ini menyumbang kepada penjadualan yang lebih cekap, pantas, dan seimbang dalam persekitaran awan heterogen, yang berpotensi meningkatkan prestasi serta menawarkan alternatif inovatif kepada penyedia perkhidmatan awan.

ENHANCED MARINE PREDATOR ALGORITHMS FOR TASK SCHEDULING IN CLOUD DATA CENTRES

ABSTRACT

Deterministic task scheduling in traditional homogeneous cloud data centres is straightforward, but increasing resource heterogeneity has made efficient scheduling more complex. Deterministic methods struggle to meet Quality of Service constraints, as improper VM-to-task mapping can degrade performance, leading to underutilization or overutilization of virtual machines (VMs). To address this challenge, researchers are turning to metaheuristic approaches, such as Marine Predator Algorithm (MPA). This study proposes a modified MPA-based task scheduling method to minimize task completion time and makespan. Three variants are introduced: modified MPA task scheduling (MMPACTS), MMPACTS with heterogeneous VM capacity identification (MMPACTS-H), and MMPACTS with load balancing (MMPACTS-HLB). MMPACTS adapts MPA for discrete task allocation and enhances it with selective mutation and also Cauchy mutation. MMPACTS-H incorporates a VM Capacity Index to account for varying VM capabilities, while MMPACTS-HLB employs a lottery-based VM selection counter to balance workloads. Performance evaluation was conducted on simulated data set and also real data based on Google Cluster Trace and GoCJ, considering makespan, resource utilization and degree of load balance. MMPACTS reduces makespan by 36% compared to deterministic methods, namely Shortest Jobs First (SJF) and First-Come First-Served (FCFS) and by 22% over other metaheuristics, namely Artificial Bees Colony (ABC) algorithm and Particle Swarm Optimization (PSO). MMPACTS-H further improves makespan by 12%. MMPACTS-HLB enhances load balancing,

reducing imbalance by 47% over SJF, 51% over FCFS, 28% over ABC and 34% over PSO. These results demonstrate the effectiveness of the proposed approach in the heterogeneous cloud environment. In summary, the research contributes to a more efficient, faster and better-balanced scheduling in heterogeneous cloud environments, potentially leading to performance improvement and an innovative alternative for cloud providers.

CHAPTER 1

INTRODUCTION

1.1 Problem Background

Cloud data centres have become essential to modern computing, with approximately 75% of organizations worldwide utilizing cloud services in some form, alongside a growing number of household consumers (Cisco, 2016, 2020). These data centres provide a wide range of on-demand information technology (IT) services with guaranteed Quality of Service (QoS), supporting various service models such as Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (Saas). By leveraging shared physical and virtual resources, cloud data centres enable applications to perform extensive computations efficiently, improving scalability and performance across diverse tasks. Additionally, cloud data centres serve as platforms for hosting user-generated content and supporting advanced data processing needs (Shafiq et al., 2021).

Cloud service requests basically consist of a set of tasks which are allocated to a server for processing. The server acts as a host, and the task processing is done by a virtual machine (VM) that resides on the host (Beloglazov & Buyya, 2013; Dalvandi et al., 2017). Processing capabilities of hosts and VMs are limited by their central processing unit (CPU) power, RAM memory allocation and also storage. Each task has a certain length that indicates how many million instructions (MI) need to be processed on a VM. The decision to map a task to a VM, and the host for that matter, depends on the task scheduling policy and algorithm set in the data centre.

Task scheduling is important as it directly impacts the performance of cloud data centres. Different objectives are achieved by employing different policies of task scheduling such as load balancing (Jafarnejad Ghomi et al., 2017; Mondal et al., 2020; Shao et al., 2018), optimized profit (Hassan et al., 2023; Heidari, 2018; Li et al., 2016), improved energy efficiency (Akhter et al., 2018; Gao, 2017; Li, 2024; Tan, 2017) and increased time-based performance (Noroozoliaee et al., 2014; Pal et al., 2021). These objectives relate to the QoS promises that a data centre gives to their clients. In the real world, most data centres apply the First-Come First-Served (FCFS) scheduling, which is a deterministic approach that is fairly straightforward and based upon multiple queues in servers (Amazon Web Services, 2024; Google Cloud, 2024; Grosof et al., 2019; Harchol-balter, 2021).

Conventionally, the configuration of servers and other resources in data centres is relatively similar. With similar processing capacity of each VM, denoted as million instructions per second (MIPS), a task can be processed in the same duration by any VM that is available and not busy. For instance, in a data centre with five VMs where all its VM processing capacity is 500 MIPS, a task of length 2000 MI can be completed within 4 seconds by any of the five VMs. This homogeneous trait permits the application of deterministic task scheduling policies to be employed with favourable outcomes within a short time (Beloglazov & Buyya, 2011; Prity et al., 2023).

However, with the upscaling capability of cloud data centres, many data centres now have heterogeneous configurations of machines (Behera & Sobhanayak, 2023; Crago & Walters, 2015; Kim, 2009). Hence, instead of processing the task of length 2000 MI on a 500 MIPS VM for 4 seconds, assigning the task to a VM with 1000 MIPS will cut the processing time in half. Heterogeneous resources in cloud data

centres pose challenges primarily due to the varying capabilities and capacities of the underlying hardware (Sohani & Jain, 2021). This diversity encompasses differences in processing power, memory capacity, network bandwidth, and specialized hardware across different VMs or physical servers within the data centre.

In addition to the diversity in resources, cloud tasks exhibit heterogeneity as well. Cloud data centres receive and process a wide range of tasks, spanning various services such as web queries, video streaming, and online word processing. Modern data centres are also supporting big data platforms such as Spark (Z. Fu & Tang, 2022) and Hadoop (Soualhia et al., 2020). Varying resource requirements are expected in applications such as machine learning, data analysis and graph analysis. Tasks can either be computationally intensive, memory-intensive, floating-point-intensive, data-intensive, and so on. For example, a machine learning task may initially require significant input/output (I/O) resources for data processing but later transitions to being memory- and compute-bound during subsequent stages. To accommodate such diverse demands, data centres deploy a multitude of heterogeneous devices, ensuring efficient resource allocation for big data analytics frameworks within the cluster.

These devices include computation devices such as hosts and VMs, storage devices and network devices. Storage devices include traditional Hard Disk Drives (HDDs), Solid-State Drives (SSDs) and emerging technologies such as Non-Volatile Memory Express (NVMe) drives. Data centres may additionally utilize Network-Attached Storage (NAS) and Storage Area Network (SAN) solutions, along with distributed file systems like Hadoop Distributed File System (HDFS). By leveraging this array of storage options, data centres can effectively balance performance,

capacity and cost considerations to support the diverse workloads of big data analytics applications.

The delivery of cloud services is regulated by a defined set of QoS criteria, ensuring reliability and performance benchmarks pledged by the cloud service provider (Kalekuri et al., 2016). These specific parameters, formally stated within a Service Level Agreement (SLA) document, encompass assurances to users regarding the completion of their service requests with optimal user experience (Laghari et al., 2018; Lang et al., 2018). The defined set of QoS criteria includes stringent guarantees on response time, uptime, data transfer rates, security measures and scalability, all aiming to ensure a consistent and satisfactory service delivery to users.

1.2 Issues

The growing popularity of cloud services means a substantial number of tasks requires scheduling to hosts and VMs in cloud data centres each day. For example, Google processes approximately 20 million tasks over 12 000 hosts in the duration of 29 days (Liu & Cho, 2012; Wilkes, 2020). One of the major issues that arises is that task scheduling has evolved into an NP-hard combinatorial optimization problem, attributable to the dynamic nature of tasks and the heterogeneity of resources (Bo et al., 2022).

Examining the compatibility of each task with individual VMs is a time-intensive process. The scheduler needs to ensure that user requests are fulfilled in a timely manner, which usually means the fastest time possible (Liu et al., 2021; Sambangi & Gondi, 2019). Tasks need to be mapped in such a way that allows them to complete processing optimally within an acceptable duration, lest to not violate the QoS requirement stated in the Service Level Agreement (SLA).

Due to the heterogeneous nature of incoming tasks, task scheduling is also subject to the risk of unbalanced resource utilization (Loncar, 2021; Newhall et al., 2016). In an attempt to maintain simplicity, current algorithms do not consider the varying resource requirements of tasks. As a result, VMs with lower resource specifications might be underutilized, operating at a small fraction of their capacity, while other 'more popular' VMs are heavily loaded. This underutilization scenario leads to inefficient resource allocation and wasted resources. On the other hand, a sudden surge in demand of resource-intensive application might keep on assigning additional tasks to VMs that are already running near maximum capacity. This overloading may cause performance degradation and a higher risk of system failures (Jena et al., 2021; Monil & Rahman, 2015). In extreme cases, a server may fail due to the excessive load, thus impacting the availability and reliability of the hosted applications.

In contemporary cloud data centres like those operated by Google (Google Cloud, 2024) and Amazon (Amazon Web Services, 2024), task scheduling often employs FCFS approach. While FCFS is straightforward and easy to implement, it may not always be the most efficient method for task allocation (Jafarnejad Ghomi et al., 2017; Kumar & Kumar, 2019). As data centres continue to grow in scale and complexity, the limitations of FCFS become more apparent, particularly in handling dynamic workloads and optimizing resource utilization.

Recent research trends have thus shifted towards exploring metaheuristic algorithms for task scheduling (Shafiq et al., 2021; Singh et al., 2021). Metaheuristic approaches offer the potential to address the shortcomings of traditional scheduling methods by providing adaptive and heuristic-driven solutions. This research direction

underscores the importance of developing advanced scheduling techniques to meet the evolving demands of modern cloud computing infrastructures. A variety of metaheuristic algorithms have been proposed for cloud task scheduling, including the Artificial Bee Colony (ABC) algorithm (Kruekaew & Kimpan, 2022), Ant Colony Algorithm (Ajmal et al., 2021; Sharma et al., 2022), Whale Optimization Algorithm (Gupta & Singh, 2024; Mangalampalli et al., 2023; Mohammadzadeh et al., 2024) and many others.

Another promising algorithm is the Marine Predator Algorithm (MPA) which mimics the hunting behaviour of marine animals such as sharks, whales and sea lions (Mugemanyi et al., 2023). MPA is particularly well-suited for task scheduling in cloud data centres where the predators resemble incoming tasks that explore optimal task-to-VM assignments across a large, distributed cloud infrastructure around the globe, similar to the scattered distribution of prey in the ocean. MPA demonstrates a stronger capability to escape local optima and consistently find the best optimal across diverse settings, outperforming several other metaheuristic algorithms (Aydemir, 2023; Faramarzi et al., 2020).

1.3 Motivation

The increasing reliance on cloud services has made robust and efficient task scheduling a critical component of cloud data centres. The limitations of traditional scheduling approaches, particularly FCFS and SJF, in modern cloud environments necessitate the exploration of more advanced techniques. The growing scale and complexity of cloud data centres, coupled with the dynamic nature of workloads and the heterogeneity of resources, demand sophisticated solutions to ensure optimal performance, resource utilization and reliability. This is underscored by several high-

profile incidents in public clouds that highlight the potential consequences of inadequate scheduling and resource management.

For example, AWS has experienced outages in several occasions, affecting their big names clients in consumer services such as Netflix and DoorDash (The Guardian, 2021). While often attributed to a combination of factors such as power failure and misconfigurations, these outages could have been exacerbated by inefficient task scheduling. Network congestion (Bhatele et al., 2015; Wang et al., 2017; Zhang et al., 2023), a common factor in cloud outages, can significantly be influenced by how tasks are distributed and how resources are allocated. If a surge in requests is not handled by distributing tasks efficiently, network segments can become overloaded and contribute to cascading failures.

Similarly, Azure outages, though often linked to infrastructure issues or software bugs, demonstrate the importance of dynamic rescheduling and workload rebalancing. The ability of the system to quickly adapt and redistribute tasks in response to failures is crucial for minimizing downtime. Inefficient scheduling can hinder recovery and prolong outages. Even Google Cloud has faced disruptions. While their post-mortem analysis often points to specific causes, the underlying scheduling mechanisms play a crucial role in how well the system handles unexpected events and recovers from failure (Salamone, 2022).

These incidents collectively highlight the critical role of intelligent task scheduling in building resilient cloud infrastructures. By developing advanced scheduling strategies that consider heterogeneity and dynamic workloads, cloud systems can achieve greater efficiency, reliability and scalability. This research directly supports the objectives of Sustainable Development Goal 9 (SDG 9), which

emphasizes building resilient infrastructure, promoting inclusive and sustainable industrialization, and fostering innovation. Enhancing cloud resource management not only strengthens the backbone of the digital economy but also contributes to creating more robust, innovative, and sustainable infrastructures essential for supporting future industrial and technological growth.

1.4 Problem Statement

Cloud service users expect their requests to be completed as fast as possible and any deviation from the specified time service level objectives is not well tolerated (Pal et al., 2021). To meet these expectations, cloud data centres attempt to match each incoming task to the most suitable VM with sufficient CPU capability to ensure timely task completion. Performance is usually measured by the average task completion time and the total time to complete all tasks (makespan). However, the matching feat becomes increasingly complex due to the large search space resulting from the large number of available VMs.

Traditional task scheduling methods, which often rely on deterministic methods such as FCFS or SJF are becoming inadequate for handling this complexity, leading to significant time delays in finding optimal VM-to-task matches (Tamilarasu & Singaravel, 2023). Consequently, researchers have turned towards metaheuristic approaches to tackle the VM-to-task assignment problem. These methods offer promising ways to find near-optimal solutions within reasonable timeframes.

However, despite the promising performance of metaheuristic algorithms in cloud task scheduling, several challenges remain. Many existing metaheuristic approaches, including the popular methods like ABC, ACO and WOA, often struggle with issues such as premature convergence and inconsistent performance across highly

dynamic environments (Kruekaew & Kimpan, 2022; Mangalampalli et al., 2023; Sharma et al., 2022), which requires them to be finetuned in different situations.

The MPA, while recently proposed as a powerful nature-inspired method capable of mitigating premature convergence issues (Faramarzi et al., 2020), still exhibits certain limitations when applied to cloud task scheduling. Notably, MPA was originally designed for continuous optimization problems, and requires adaptation to effectively handle discrete populations (Abdel-Basset et al., 2020). Although MPA achieves high accuracy through its meticulous update mechanisms, this also results in high time complexity, leading to slower convergence speed (Sadiq et al., 2022), which is a significant drawback in the fast-paced environment of cloud computing. Therefore, further modifications are necessary to enhance MPA for cloud task scheduling.

The heterogeneous nature of cloud data centres introduces significant variability in VM performance characteristics. For instance, one VM might have higher computational power but limited memory, or two VMs possess the same computational power but different memory dan disk space. Assigning tasks to mismatched resources can lead to suboptimal performances or resource wastage. Recent literatures are increasingly acknowledging the heterogeneity of cloud data centres. However, many works still focus predominantly on computational power (CPU) as the primary attribute of VMs (Ghafari et al., 2022; Liang et al., 2020), often overlooking the equally significant aspects of memory and disk space within the VM architecture (Hai et al., 2023).

The need to address heterogeneity is more apparent with big data applications offered in cloud data centres. Service providers obtain IaaS instances from cloud providers with reserved options and use them to execute big data analysis jobs from

their users. However, big data platforms, e.g., Spark, are typically oblivious of the dynamic resource demands of applications and the underlying heterogeneity, as the unit of per-task resource allocation is a homogeneous container, i.e., the abstraction does not capture heterogeneity of resources such as CPU cores, RAM, and disk Input/Output (I/O). This fundamental mismatch between the high-level software platforms and the underlying hardware results in degraded performance, and wastage of resources due to inability to efficiently utilize different resources (Xu et al., 2018).

The majority of authors focus primarily on computing resources only, since the main activity of the task is on CPUs. However, memory resources and storage resources are also important. Sufficient RAM is necessary to handle large datasets efficiently and to support quick data access during processing. Additionally, the frequent I/O operations required for big data analytics tasks make data locality more crucial, as local I/O can minimize task execution time more effectively than remote ones (Bouhouch et al., 2023). In other words, choosing a VM that has sufficient memory and storage capacity to hold the data while being processed will speed up the execution of tasks.

Additionally, imbalanced task load distribution poses a threat to the performance of a cloud data centre which may lead to violation of QoS. Hence, maintaining a good balance between overutilization and underutilization of resources is also a challenge in cloud computing (Jain et al., 2019). Many methods of load balancing depend on the different situations and characteristics of the tasks and goals of the data centre. Overutilization may lead to performance bottlenecks, increased latency, and potential service degradation. Underutilization, on the other hand, might result in wasted resources, higher operational costs and reduced overall system

efficiency. Hence, a good task scheduling should be able to provide optimal resource utilization (Mekala & Viswanathan, 2019; Prakash et al., 2021) and lower degree of imbalance (Lou et al., 2023) in the data centre as a whole.

Thus, the main problem identified is that existing task scheduling methods inadequately address the multi-dimensional heterogeneity of VM resources (CPU and memory) and needs better way to achieve optimal load balancing across cloud data centres, leading to suboptimal task completion times, resource wastage and potential QoS violations. To this end, the main research question of this study is as follows:

"How can incoming tasks with diverse processing requirements be distributed among heterogeneous virtual machines in a cloud data centre to minimize task completion time and avoid resource imbalance?"

1.5 Research Objectives

The main goal of this study is to improve task execution performance in cloud data centres with heterogeneous capacity and requirement. To achieve this, the objectives are as follows:

1. To design a task scheduling method for cloud data centres based on modified Marine Predator Algorithm (MPA) to minimize task completion time and makespan (MMPACTS).

2. To enhance MMPACTS with a VM capacity index to consider the CPU and memory in heterogeneous cloud data centres (MMPACTS-H).

3. To enhance MMPACTS-H with a capability of balancing the load to minimize degree of imbalance and avoid underutilized or overutilized VMs (MMPACTS-HLB).

1.6 Scope and Limitations

This thesis focuses on the development and evaluation of an enhanced Marine Predator Algorithm for task scheduling in cloud data centres. The scope of this study encompasses the following key aspects:

(i) The focus is on optimizing the allocation of tasks to VMs based on their computing capacities and other relevant factors.

(ii) The study specifically targets the allocation of tasks to VMs within cloud data centre environments. Each VM is capable of processing multiple tasks, subject to its computing capacity, which includes CPU power, memory, and disk size.

(iii) The tasks considered in this study are assumed to be independent and can be executed in any sequence. The length of each task is known beforehand, facilitating the scheduling process.

(iv) The VMs in the cloud data centres are heterogeneous in nature, with varying capacities in terms of CPU power, memory, and disk size. The enhanced MPA is designed to effectively allocate tasks to VMs while considering these differences in capacity.

(v) Single Machine Assignment: Each task is assigned to a single VM, and no task splitting is involved. Once a task is assigned to a VM, it is

completed by that VM alone, without further distribution or partitioning.

1.7 Thesis Organization

This thesis is organized into seven chapters, including this first introductory chapter. Brief descriptions of the content of each chapter are given as follows:

(i) Chapter 2 outlines the review of current advances in the domain of cloud task scheduling and cloud load balancing. This chapter also provides some insights into the theoretical background of the focused domain problems, challenges, trends and directions that motivate the pursuit of this study.

(ii) Chapter 3 describes the research methodology employed in this research including the research framework, data sources, instrumentation, problem description, performance measurements, as well as experimental design and analysis conducted in this research.

(iii) Chapter 4 describes the design of modified MPA approach to accommodate the task scheduling problem in cloud data centres. The procedure of parameter tuning is also discussed. The numerical results of simulation experiments done is presented in this chapter.

(iv) Chapter 5 touches on the incorporation of VM capacity index that identifies the capacity of each VM and the suitability of the VM for each task.

(v) Chapter 6 discusses the enhancement of the task scheduler with load balancing capability, by incorporating a number of limits for each VM to be selected, based on lottery allocation. This feature ensures that each VM is chosen accordingly and not overloaded or underloaded at any point of time.

(vi) Chapter 7 presents the conclusion, research contributions and recommendation of future work that can be undertaken from this work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Optimization of cloud data centres are vital in accommodating the large demands of modern computational tasks. This chapter reviews the key areas relevant to task scheduling and load balancing in cloud environments. Section 2.2 discusses the fundamentals of cloud data centres, while Section 2.3 focuses on cloud task scheduling. Section 2.4 examines load balancing techniques. Section 2.5 reviews related work, covering both deterministic and metaheuristic approaches. Section 2.6 identifies research gaps and discusses current challenges. Section 2.7 introduces the Marine Predators Algorithm (MPA), detailing its working principles, advantages, and limitations. Finally, Section 2.8 provides a summary of the chapter.

2.2 Cloud Data Centres

Cloud data centres are a type of distributed computing facility used to house and manage a large number of resources and related infrastructure required to support cloud computing services. Clients can opt for specific usage of resources based on their need, without worrying about capital expenditure and maintenance (Ison, 2020). Expanding or downsizing operations can be done at any point of time, and the costs involved are based on pay-per-use model (Chopra, 2018; Kim, 2009).

Traditionally, a data centre is very costly as all the resources must be procured by the organization and maintained. The physical facility consists of different equipment such as computers and servers as hosts, network devices such as routers and switches, and application delivery controllers (Barroso & Hölzle, 2009). To keep all the

hardware and software operational, other infrastructure are in place which include ventilation and air-conditions, power supplies and security. A dedicated space is necessary and most of the time takes up a lot of space. With cloud data centres, organizations and even individuals can access information services with a lower price, paying only for the services that they use according to their needs and without geographical constraints (Chopra, 2018). All services are provided seamlessly to users, even though the data centres are distributed in various locations.

The primary aim of the cloud is to utilize the distributed resources effectively in order to attain high throughput and performance (Alipour et al., 2016; Atmaca et al., 2016). It enables the cloud to resolve the problems that require high computing power. It also allows distribution of the resources all around the world to execute the tasks at different data centres to provide faster services to clients. This provides business opportunities to both cloud service providers (CSPs) to establish new data centres and cloud users to put their business on the cost-efficient cloud. Cloud computing allows users to scale in or out virtual resources dynamically without human interaction according to their computing requirements (Alam & Ahmad Khan, 2017).

Service requests on the cloud are represented by jobs and tasks; a job can have one or more tasks (Chen & Katz, 2010; Wilkes, 2020). Jobs and tasks are directed to available host and VMs according to a defined policy as visualized in Figure 2.1. Multiple VMs could be allocated in a single host using virtualization technology (Liu & Cho, 2012). These virtual machines act as individual physical machines, along with their individual operating systems and shared system resources. Multiple tasks run on these machines in parallel. When a user submits a request to the cloud, they are normally interested in the minimization of the overall execution time (Ardagna et al., 2014; Yang

et al., 2019). Cloud service providers are also aware of the proper utilization of available system resources and execution time (Naha & Othman, 2015).

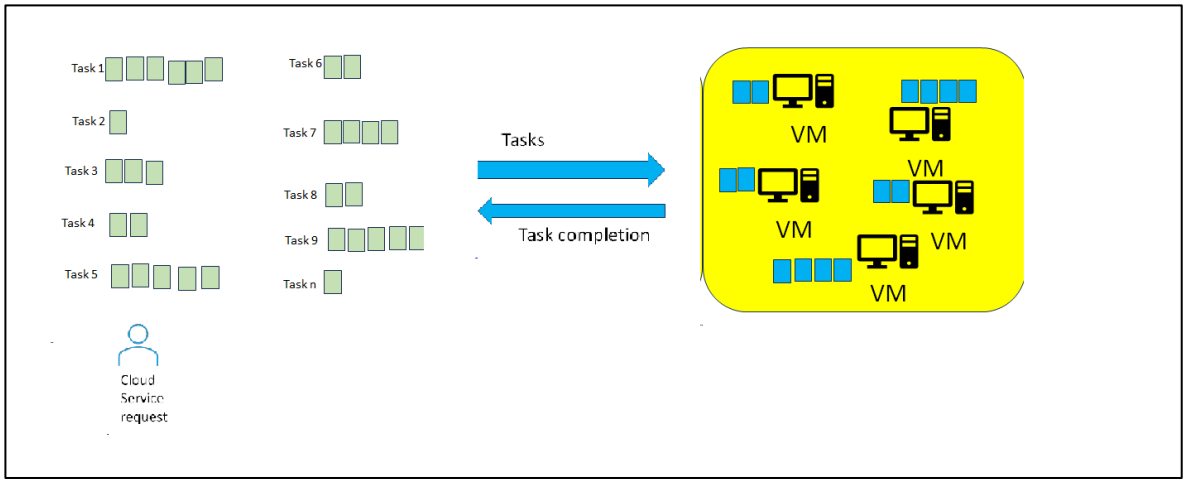

Figure 2.1 Visualization of Jobs and Tasks Mapping to Cloud Host and VMs

2.2.1 Types of Cloud Services

Cloud data centres offer a wide range of off-site services to remote clients, including Platform-as-a-Service (PaaS), Software-as-a-Service (SaaS), and Infrastructure-as-a-Service (IaaS). These services are fundamental components of the cloud computing model, which has become a critical technology in the digital age.

Infrastructure as a Service (IaaS) is a fundamental cloud computing model that provides virtualized computing resources over the internet. In the context of cloud computing, IaaS offers a comprehensive and scalable infrastructure, allowing users to access and manage fundamental computing components without the need for physical hardware ownership or maintenance (Badshah et al., 2023). In an IaaS environment, users are typically provided with virtualized resources such as virtual machines, storage, and networking components. These resources are hosted on the infrastructure provider's data centres and can be dynamically scaled based on the user's requirements. The key

features of IaaS include on-demand provisioning, elasticity, and resource pooling, enabling users to allocate and de-allocate resources according to their computational needs.

One of the primary advantages of IaaS is its cost-effectiveness, as users only pay for the resources they consume (Oracle, 2023). This model eliminates the need for large upfront investments in physical hardware and allows for more efficient resource utilization. Additionally, IaaS provides flexibility and agility, enabling rapid deployment of applications and services without the constraints of physical infrastructure. IaaS is particularly relevant in the context of modern computing paradigms, supporting various applications ranging from simple web hosting to complex data analytics and machine learning tasks. Its versatility makes it a crucial component in the contemporary IT landscape, facilitating the development and deployment of scalable and resilient systems.

Platform-as-a-Service (PaaS) is a cloud service that provides a platform and environment for developers to build, deploy, and manage applications without having to worry about the underlying infrastructure (Aziza & Krichen, 2018; Oracle, 2023). PaaS offerings typically include development tools, databases, and other resources necessary for application development. On the other hand, Software-as-a-Service (SaaS) delivers software applications over the internet on a subscription basis, eliminating the need for users to install and maintain software locally (Wang et al., 2022). Popular examples of SaaS include Google Workspace (formerly G Suite), Microsoft 365, and Salesforce.

Traditionally deployed in dedicated on-premises computing cluster, big data services are now also offered as a cloud service. This platform, exemplified by systems

like Apache Spark and Apache Hadoop, plays a pivotal role in cloud data centres by providing scalable and distributed frameworks for processing vast volumes of data (Bouhouch et al., 2023; Sharma & Shamkuwar, 2019). Typically associated with IaaS and PaaS, depending on how they are deployed, these platforms are designed to handle the three key challenges associated with big data: volume, velocity, and variety. In the context of cloud data centres, Apache Spark and Hadoop offer distributed storage and processing capabilities, enabling the efficient handling of massive datasets across clusters of interconnected servers. These platforms leverage parallel processing and fault tolerance mechanisms to enhance performance and reliability, crucial aspects when dealing with large-scale data processing tasks.

Apache Hadoop, often considered the pioneer in big data processing, employs the Hadoop Distributed File System (HDFS) for distributed storage and the MapReduce programming model for distributed processing (Li & Hei, 2022). It allows users to break down complex data processing tasks into smaller, more manageable sub-tasks, which are then distributed across the cluster for parallel execution. Apache Spark, on the other hand, is known for its in-memory processing capabilities, offering faster data processing by keeping intermediate data in memory rather than writing it to disk (Fu & Tang, 2022; Mittal et al., 2019). Spark also provides a more versatile and expressive programming model compared to Hadoop's MapReduce, allowing developers to write complex data processing workflows more efficiently. In a cloud data centre environment, these big data platforms bring several advantages. They enable elastic scalability, allowing users to dynamically adjust computing resources based on the workload (Hewapathirana & Silva, 2021). Cloud-based deployment also facilitates easy access to a variety of storage options (Gupta, 2021; Pan et al., 2018; Qian et al., 2019), and users can take advantage of managed services for seamless integration with other cloud services.

2.2.2 Resources in Cloud Data Centres

As mentioned earlier, the number of resources that makes a data centre is large, and consists of several categories, namely computing resources, storage resources and networking elements.

Computing resources are composed of numerous servers, usually aggregated together and managed through virtualization technologies. This arrangement facilitates the effective distribution of processing power to cater to varying workloads and user demands (Delimitrou & Kozyrakis, 2016; Xu & Buyya, 2019). The scalability and on-demand provisioning features of these resources are vital in accommodating the dynamic computational requirements of cloud-based applications and services (Mao et al., 2023).

Storage resources within cloud data centres are a cornerstone for accommodating vast amounts of data, employing distributed file systems and storage architectures (Gupta, 2021; Klimovic et al., 2016). This ensures reliability, fault tolerance, and the ability to scale storage capacity in response to escalating data volumes (Barroso & Hölzle, 2009). Networking resources form the connective tissue of these data centres, facilitating data transmission and communication between the diverse components and across geographic regions. High-speed, redundant networking infrastructures are integral in maintaining seamless connectivity and enabling rapid data transfer, underscoring the importance of robust networking resources in cloud data centres (Kouyoumdjieva & Karlsson, 2016).

The advent of specialized resources, such as Graphical Processing Units (GPUs) and Tensor Processing Units (TPUs), has gained prominence within these centres, enabling accelerated performance for specific workloads like machine learning,

scientific simulations, and high-performance computing. The utilization of these specialized resources is instrumental in driving innovation and supporting advanced applications that demand enhanced processing capabilities.

2.2.3 Heterogeneity in a Cloud Data Centre

A cloud data centre can be considered heterogeneous in two ways: (1) the platforms used, and (2) the infrastructure level (Varghese & Buyya, 2018). The first is in the context of multi-clouds, in which platforms that offer and manage infrastructure and services of multiple cloud providers are considered to be a heterogeneous cloud. Heterogeneity arises from using hypervisors and software suites from multiple vendors. The second is related to heterogeneity at the infrastructure level, in which different types of processors, storage, and other resources are combined to offer VMs with varied compute capability.

Heterogeneous resources in cloud data centres pose challenges primarily due to the varying capabilities and capacities of the underlying hardware (Shirvani & Talouki, 2021). This diversity encompasses differences in processing power, memory capacity, network bandwidth, and specialized hardware (such as GPUs or TPUs) across different VMs or physical servers within the data centre. Added with the large range of diversified nature of tasks specifications, task variance is also another factor to be considered to ensure efficiently to these varied resources (Hai et al., 2023). While it is possible to allocate tasks to any available VM, ensuring optimal performance becomes complex due to the following reasons:

Resource Variability: Different VMs may have varying performance characteristics (Xu & Buyya, 2019). For instance, one VM might possess higher computational power but limited memory, while another might have abundant memory

but slower processors. Assigning tasks to mismatched resources can lead to suboptimal performance or resource wastage.

Resource Constraints: Certain tasks might require specific hardware configurations or software environments, like GPUs for machine learning or specialized software stacks. Not all VMs might have these resources available, making it challenging to allocate tasks that demand such specific environments (Pirozmand et al., 2022).

Performance Heterogeneity: VMs within a cloud data centre often experience performance variations due to factors like resource contention, interference from neighbouring VMs, or differing workload patterns (Kim et al., 2011; Saxena & Singh, 2023). Assigning tasks without considering these variations can result in unpredictable performance outcomes.

Optimization Complexity: Finding an optimal allocation strategy among heterogeneous resources is computationally complex (Frangioni et al., 2016; Kashani & Jahanshahi, 2009). It involves considering multiple factors such as workload characteristics, resource availability, and real-time demand, which can be challenging to manage efficiently.

Efficient task allocation in heterogeneous cloud environments often requires sophisticated resource management techniques (Ge et al., 2018; Jennings & Stadler, 2015). These techniques involve intelligent workload placement algorithms (Molka & Casale, 2016), dynamic resource provisioning (Priya et al., 2019; S. Singh & Chana, 2016), and workload scheduling (Postoaca & Pop, 2018) strategies that consider the heterogeneity of resources to optimize performance while meeting user-defined constraints.

Research in this area explores various optimization models, such as genetic algorithms (Goudarzi et al., 2017), machine learning-based approaches (Alsubaei et al., 2024), and heuristics (Aron & Abraham, 2022) to address these challenges. As cloud data centres continue to grow in scale and complexity, addressing the issue of heterogeneous resources remains a focal point in ensuring efficient resource utilization and performance in these environments.

2.2.4 Geographical Distribution

As cloud services gain traction, there is a discernible trend of data centres being deployed across dispersed global locations (Amazon Web Services, 2023; Google, 2023b). This geo-distributed approach offers a range of advantages such as continuous and efficient delivery of services. Clients also benefit from enhanced performance and cost-effectiveness, attributed to the proximity of the data centres. The deliberate placement of data centres in different regions serves to mitigate the risk associated with natural disasters in specific locales. Besides that, data centres can exploit certain environmental traits to their advantage.

Big companies like Google (Google, 2023b), Apple (Zhang, 2022) and Twitter (Sullivan et al., 2022) have data centres located all over the world to support their global operations. The exact locations and numbers are subject to change due to various factors such as expansion strategies, regulatory requirements and technological advancements. For instance, Google's extensive network of data centres includes locations in North America, Europe, Asia and other regions, allowing it to deliver services efficiently to users worldwide. These companies prioritize the strategic placement of data centres to minimize latency, enhance data security, and comply with local regulations while catering to the diverse needs of their global user base.

2.2.5 Virtual Machines Ranking

Cloud virtual machines ranking refers to the evaluation and comparison of VMs offered by various cloud service providers. Ranking VMs enables users to make informed decisions based on their specific requirements, balancing factors such as performance, cost, scalability, reliability, security and user experience. When making comparisons, it is crucial to have a certain benchmark or score system that can be used to evaluate and compare between the machines.

Benchmark scoring is not entirely foreign in computing. CoreMark (EEMBC, 2009; Kolev & Helpa, 2023) is one of the benchmark suites that serves as a standard in assessing the performance of computer processors or microcontrollers. Introduced by the Embedded Microprocessor Benchmark Consortium (EEMBC) in 2009, the functionality of a processor core is represented by a single number score, providing a simple metric aiding developers in hardware selection. It focuses on evaluating core CPU functionality in embedded systems, including integer performance, memory operations, and control flow.

Unlike CoreMark, SPEC CPU 2017 (Standard Performance Evaluation Corporation, 2017) is a suite of benchmarks covering a wider range of computation tasks, such as integer and floating-point operations and multimedia processing, and memory access patterns. Established by the Standard Performance Evaluation Corporation (SPEC), its comprehensive evaluation of CPU performance across diverse workloads makes it suitable for a wide range of applications and industries. However, running this benchmark requires significant computational resources and time which might limit its practicality for smaller users or environments (Tousi & Lujan, 2022).