Faculty of Computer Science and Information Technology

Energy Efficient High-Performance Computing Aware Proactive Dynamic Virtual Machine Consolidation Technique in Cloud Computing

Rukshanda Kamran Butt

Doctor of Philosophy
2022

Energy Efficient High-Performance Aware Proactive Dynamic Virtual Machine Consolidation Technique in Cloud Computing

Rukshanda Kamran Butt

A thesis submitted

In fulfilment of the requirements for the degree of Doctor of Philosophy

(Computer Science)

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK

2022

DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Malaysia Sarawak. Except where due acknowledgements have been made, the work is that of the author alone. The thesis has not been accepted for any degree and is not concurrently submitted in the candidature of any other degree.

Rukshanda

……………………………

Signature

Name: Rukshanda Kamran Butt

Matric No.: 14010030

Faculty of Computer Science and Information Technology

Universiti Malaysia Sarawak

Date : 17 August 2022

ACKNOWLEDGEMENT

First of All, I thank ALLAH (SWT), the God Almighty, for granting me the health, knowledge, strength, ability, and patience to accomplish this research and blessing me with sympathetic and supportive supervisors and family members.

I am glad to express tremendous gratitude to my supervisor Dr Noor Alamshah B. Bolhassan, for his sympathetic character, knowledge sharing, ideas and continuous support from the first until the last day of this study. His sincere behaviour and constructive feedback enabled me to achieve significant research milestones within the required time frame.

I would also like to thank my Co-supervisor, Dr Nadianatra Binti Musa, for enriching me with innovative ideas and skills by sharing her expertise and knowledge. Due to her unlimited support for reviewing, improving and evaluating my research, I completed my thesis sustaining the high standard of practice in UNIMAS.

At various stages during this study, I faced several undesirable challenges which overburdened me with mental and physical stress. However, this never stopped me from progressing further due to encouraging, moral, and financial support from my husband Kamran butt and my son Muhammad Shayan. I am incredibly thankful to them for their understanding, kindness, belief, and trust in me.

I also express my profound appreciation to my father, mother, brothers and sisters for their unlimited prayers regarding my success during this study. This research would have never been possible without their wishes for my success.

Finally, I would like to thank the management of the Universiti Malaysia Sarawak for making it possible for me to complete my study here in Sarawak. Thank you all.

ABSTRACT

High-Performance Computing (HPC) applications have gained extensive interest in cloud computing. Cloud computing provides HPC applications with enormous on-demand resources; therefore, many studies have tried to investigate the currently available cloud service architectures for running HPC applications in the most effective approach. There is yet a promising research direction to explore techniques to efficiently schedule a mix of HPC and non-HPC applications to increase resource utilization in the cloud data centre. However, there are notable challenges that involve cross-application interference, energy issues, and maintaining service-level agreements (SLAs) in multiple metrics, e.g. response time (web application) vs execution time (HPC application). In High-performance Computing (HPC) applications, the resource usage pattern varies over time, making it challenging to track resource utilization. Better workload predictions will lower energy costs by accurately forecasting future workload demands. For this purpose, proactive dynamic Virtual Machine (VM) consolidation is among the most effective approaches to minimize energy consumption and enhance resource usage while balancing the efficiency of HPC and non-HPC applications in a cloud data centre. Since the VM consolidation problem is strictly NP-hard, many heuristic algorithms have been proposed to tackle the problem. However, these algorithms do not consider the workloads' different nature, incredibly individual classes of HPC workload in the cloud data centre. Thus, this thesis explores the opportunities and challenges of proactive, dynamic VM consolidation for HPC and non-HPC applications on the cloud datacenter. Additionally, this thesis proposes and implements an HPC-aware and energy-efficient scheduling algorithm for proactive, dynamic VM consolidation, which achieves better resource utilization and limits cross-application interference through careful co-location of HPC and Non-HPC applications on the same pool of resources in a cloud datacenter. Experiments were conducted using CloudSim and real HPC workload traces of Metacentrum HPC Workload. The proposed approach, Energy-Aware Multi-Dimensional Online Bin Packing (EAMDOBP), was tested against Power-aware best fit decreasing algorithm (PABFD), Modified Worst Fit decreasing algorithm (MWFD) and Hybrid Local Regression Host Overload Detection (HLRHOD). Results indicate a relative improvement of 60% in Simulation time and a 73% increase in throughput. Additionally, the utilization of the CPU, RAM, and bandwidth increased, respectively, by 77%, 84%, and 70%. The results reveal that the proposed approach significantly improves all the critical metrics used over other heuristic algorithms.

Keywords: HPC (High-Performance Computing); power management; proactive virtual machine consolidation; host overload detection; multiple regressions.

Teknik Penggabungan Mesin Maya Dinamik Proaktif yang Cekap Tenaga dan Sedar Pengkomputeran Berprestasi Tinggi dalam Pengkomputeran Awan

ABSTRAK

Aplikasi Pengkomputeran Berprestasi Tinggi (HPC) telah mendapat minat yang meluas dalam pengkomputeran awan. Pengkomputeran awan menyediakan aplikasi HPC dengan sumber atas permintaan yang besar; oleh itu, banyak kajian telah cuba menyiasat seni bina perkhidmatan awan yang tersedia pada masa ini untuk menjalankan aplikasi HPC dalam pendekatan yang paling berkesan. Masih terdapat hala tuju penyelidikan yang menjanjikan untuk meneroka teknik untuk menjadualkan gabungan aplikasi HPC dan bukan HPC dengan cekap untuk meningkatkan penggunaan sumber dalam pusat data awan. Walau bagaimanapun, terdapat cabaran ketara yang melibatkan gangguan merentas aplikasi, isu tenaga dan mengekalkan perjanjian peringkat perkhidmatan (SLA) dalam berbilang metrik, mis. masa tindak balas (aplikasi web) vs masa pelaksanaan (aplikasi HPC). Dalam aplikasi Pengkomputeran Berprestasi Tinggi (HPC), corak penggunaan sumber berubah mengikut masa, menjadikannya mencabar untuk menjejak penggunaan sumber. Ramalan beban kerja yang lebih baik akan mengurangkan kos tenaga dengan meramalkan permintaan beban kerja masa hadapan dengan tepat. Untuk tujuan ini, penyatuan Mesin Maya (VM) dinamik proaktif adalah antara pendekatan paling berkesan untuk meminimumkan penggunaan tenaga dan meningkatkan penggunaan sumber sambil mengimbangi kecekapan aplikasi HPC dan bukan HPC dalam pusat data awan. Memandangkan masalah penyatuan VM adalah NP-hard, banyak algoritma heuristik telah dicadangkan untuk menangani masalah tersebut. Walau bagaimanapun, algoritma ini tidak mengambil kira sifat beban kerja yang berbeza, iaitu kelas individu beban kerja HPC yang luar biasa dalam pusat data awan. Oleh itu, tesis ini meneroka *peluang dan cabaran penyatuan VM yang proaktif dan dinamik untuk aplikasi HPC dan bukan HPC pada pusat data awan. Selain itu, tesis ini mencadangkan dan melaksanakan algoritma penjadualan yang sedar HPC dan cekap tenaga untuk penyatuan VM dinamik yang proaktif, yang mencapai penggunaan sumber yang lebih baik dan mengehadkan gangguan merentas aplikasi melalui lokasi bersama yang teliti bagi aplikasi HPC dan Bukan HPC pada yang sama. kumpulan sumber dalam pusat data awan. Eksperimen telah dijalankan menggunakan CloudSim dan surih beban kerja HPC sebenar bagi Metacentrum HPC Workload. Pendekatan yang dicadangkan, Pembungkusan Tong Sampah Dalam Talian Berbilang Dimensi Sedar Tenaga (EAMDOBP), telah diuji terhadap Algoritma Penurunan Muatan Terbaik Sedar Kuasa (PABFD), Algoritma Penurunan Sesuai Tersuai (MWFD) dan Pengesanan Lebihan Hos Regresi Tempatan Hibrid (HLRHOD). Keputusan menunjukkan peningkatan relatif sebanyak 60% dalam masa Simulasi dan peningkatan 73% dalam pemprosesan. Selain itu, penggunaan CPU, RAM, dan lebar jalur meningkat, masing-masing, sebanyak 77%, 84% dan 70%. Hasilnya mendedahkan bahawa pendekatan yang dicadangkan meningkatkan dengan ketara semua metrik kritikal yang digunakan berbanding algoritma heuristik yang lain.*

Kata kunci: *HPC (Pengkomputeran Berprestasi Tinggi); pengurusan tenaga; penyatuan mesin maya proaktif; pengesanan lebihan hos; regresi berganda.*

TABLE OF CONTENTS

		Page
DECLARATION		i
ACKNOWLEDGEMENT		ii
ABSTRACT		iv
ABSTRAK		vi
TABLE OF CONTENTS		viii
LIST OF TABLES		xii
LIST OF FIGURES		xiii
LIST OF ABBREVIATIONS		xiv
CHAPTER 1	**INTRODUCTION**	1
1.1	Study Background	1
1.2	Problem Background	4
1.3	Problem Statement	7
1.4	Objectives	9
1.5	Scope	10
1.6	Significance of the Study	11
1.7	Organization of Thesis	12
CHAPTER 2	**LITERATURE REVIEW**	13
2.1	Overview	13

2.2 Definitions of HPC and Cloud 13
2.2.1 High-Performance Computing (HPC) 13
2.2.2 Cloud 14
2.3 Cloud Workload 16
2.4 Types of Cloud Services 17
2.5 Virtualization 23
2.6 Virtual Machine Allocations 24
2.6.1 Static VM Provisioning 25
2.6.2 Dynamic VM Consolidation 26
2.7 Elasticity 28
2.8 Resource Allocation 29
2.9 HPC in the Cloud 30
2.10 Viability Studies on HPC Cloud 32
2.11 Studies on Facilitating the Use of HPC Application in the Cloud 34
2.12 Optimizing the Performance of HPC Clouds 35
2.13 Challenges Facing HPC Cloud 36
2.13.1 Cost vs Performance 38
2.13.2 Service Level Agreement 39
2.13.3 Energy Consumption in HPC Cloud Data centres 40
2.14 Solutions to the Challenges Faced by HPC Cloud 45

2.14.1 HPC-Aware Schedulers 46

2.14.2 VM Consolidation 47

2.14.3 Hybrid Local Regression Host Overload Detection Method (HLRHOD) 48

2.14.4 HPC Cloud Workload Prediction 49

2.15 Precis and Analysis of Literature Review 56

2.16 Summary 64

CHAPTER 3 RESEARCH METHODOLOGY 65

3.1 Overview 65

3.2 The Literature Review 67

3.3 Analysis 68

3.4 Design of the Proposed Energy-Aware Multi-Dimensional Online Bin Packing (EAMDOBP) Technique 70

3.4.1 Energy-efficient and HPC-Aware Proactive Dynamic VM Consolidation in Cloud Computing 71

3.4.2 Energy-Aware Multi-Dimensional Online Bin Packing (EAMDOBP) Architecture 73

3.5 Implementation 77

3.6 Evaluation Methodology and Experimental Setup 79

3.6.1 Power Model 82

3.6.2 Performance Metrics 83

3.6.3 Workload 85

3.6.4 CloudSim Architecture and Modelling 85

3.7 Summary 87

CHAPTER 4 EVALUATION AND RESULTS 89

4.1 Overview 89

4.2 Simulation Results and Analysis 89

4.2.1 Number of VMs 90

4.2.2 Algorithms Comparative Analysis 90

4.3 Contribution and Significance 94

CHAPTER 5 CONCLUSION AND RECOMMENDATIONS 96

5.1 Introduction 96

5.2 Conclusion 96

5.3 Future Research Recommendations 98

5.4 Closing Remarks 98

REFERENCES 100

LIST OF TABLES

		Page
Table 2.1:	Comparisons of the Top-3 Cloud Vendors Based on Strength and Weakness	22
Table 2.2:	Summarized Findings of Some Feasibility Studies (Netto et al., 2018)	33
Table 2.3:	Summarized the Findings of Studies on the Usability of HPC Cloud (Netto et al., 2018)	35
Table 2.4:	Literature Review for Gap Analysis	57
Table 3.1:	The Original and Extended Classes of CloudSim	81
Table 3.2:	Power Consumed by the Chosen Hosts at Various Load Levels in Watts	83
Table 3.3:	PUE Efficiency Values	84

LIST OF FIGURES

		Page
Figure 2.1:	Convergence of Various Advances Leading to the Advent of Cloud Computing (Adamuthe & Thampi, 2014)	15
Figure 2.2:	IBM Presentations on Diverse Variations of Cloud Computing Workloads (Mateo, 2011)	17
Figure 2.3:	Cloud Computing Service Models: IaaS, PaaS and SaaS (Muratore et al., 2018)	18
Figure 2.4:	A Typical Data Centres Network	24
Figure 3.1:	Research Methodology and Outcomes	66
Figure 3.2:	Empirical Research Methodology	67
Figure 3.3:	Energy Comparison of MDOBP vs Default Scheduler	69
Figure 3.4:	Inputs and Outputs of (a) Static VM Consolidation and (b) Dynamic VM Consolidation	71
Figure 3.5:	The Proposed Energy-Aware Multi-Dimensional Online Bin Packing (EAMDOBP) Architecture	73
Figure 3.6:	Energy-Aware Multi-Dimensional Online Bin Packing (EAMDOBP)	77
Figure 3.7:	CloudSim Architecture (Redkar & Guidici, 2011)	86
Figure 3.8:	Communication Flow between Various Entities in CloudSim	87
Figure 4.1:	Algorithms Comparison for Metacentrum HPC Workload, where the Simulation time and Throughput metrics are depicted in Log-Scale	91
Figure 4.2:	Algorithms Comparison for Metacentrum HPC Workload, where the CPU, RAM and Bandwidth Utilizationmetrics are depicted in Log-Scale	92
Figure 4.3:	EAMDOBP Results	94

LIST OF ABBREVIATIONS

AWS	Amazon Web Service
CPU	Central Processing Unit
DVFS	Dynamic Voltage and Frequency Scaling
HaaS	Hardware as a Service
HPC	High-Performance Computing
IVMPP	Interference-Aware Virtual Machine Placement Problem
MPI	Message Passing Interface
NaaS	Network as a Service
CRM	Customer Relation Management
PaaS	Infrastructure as a Service
PUE	Power Usage Efficiency
QoS	Quality of Service
QoE	Quality of Expectations
SaaS	Software as a Service
SLA	Service Level Agreement
STaaS	Storage as a Service
TCA	Total Cost of Acquisition
VM	Virtual Machine
VMM	Virtual Machine Monitor

CHAPTER 1

INTRODUCTION

1.1 Study Background

The popularity of cloud computing has increased because of its many advantages available to customers. Cloud computing provides on-demand computing services, including infrastructure, platforms, software, and database for paying users (Sundarakani, 2019). It is considered an evolution enabling workload management and remote resource interaction. Cloud services' increasing popularity is mainly because of their unique features, including the possibility of resource sharing among cloud providers and users in a fee-based model. This feature is helpful for the users as it enables them to have a lower cost of ownership and an increasable capacity of dedicated infrastructures when necessary (Guyon, 2018).

High-Performance Computing (HPC) refers to using supercomputers, grid environments, and clusters of computers to solve computation-intensive problems (Regassa, (2022). HPC applications are regularly used for weather forecasting, aircraft crash simulation, and computational fluid dynamics for aerodynamics studies, among other computation-intensive purposes (Marques, 2021). Although HPC systems are challenging to obtain by research communities because of their high cost, advancement in computing technologies has enabled the use of HPC applications in the cloud, a revolutionary computing paradigm (Mohammadian, 2020). Running HPC applications on the cloud is a significant advancement that offers many advantages to the users.

In this sense, the HPC cloud uses cloud resources to administer HPC applications (Mell & Grance, 2011). The main advantage of HPC cloud computing is that it provides HPC infrastructure to individuals and businesses with no physical space, financial resources or technical and operational ability to construct their HPC infrastructure. Thus, small start-ups can obtain enormous computing power (Fernandes et al., 2016). Nevertheless, combining these two domains, HPC and cloud, is not short of shortcomings. Executing HPC applications on the cloud faces many challenges, including changing workloads. HPC operates in dedicated hardware with batch processing. HPC applications usually need direct physical access to the hardware (i.e. multi-core machines, grid and supercomputers) to have the best achievable performance. Accordingly, the traditional HPS platforms are fine-tuned perfectly to achieve the best parallel processing performance. Transferring HPC applications directly to the cloud will likely cause performance penalties and thus extinguish the advantage gained from the cloud services. Also, the increasing complexity of modern data centres increases concerns over energy usage, power efficiency, thermal stresses and cooling needs about the operating costs and their effect on the environment and societies (Singh & Chana, 2016). Specifically, a large-scale HPC application would ideally need dedicated cloud resources (compute and network). A cloud provider's per-hour charge for dedicating the resources would quickly make the proposition uneconomical for customers.

In most cases, servers work at 10-50% of their total capacity, causing additional costs on over-provisioning and, in this manner, additional Total Cost of Acquisition (TCA) (Egwutuoha et al., 2014). Since the end of the 2000s, cloud computing has become the subject of many studies as it drew the attention of researchers to its increasing maturity level. This thesis hypothesises that the presently available clouds are compatible with only some HPC applications, for which clouds are considered more cost-effective than the traditional dedicated HPC platforms when applying intelligent scheduling. Based on the evaluation of performance and analysis, it was found that many HPC applications are not perfectly compatible with cloud environments (Tavares, 2021). The poor interconnect and I/O performance in the cloud, network virtualization overhead, HPC-agnostic cloud schedulers (Gupta et al., 2018), and the inherent heterogeneity and multi-tenancy in the cloud are some obstacles to the effective running of HPC applications in the cloud (Zhou et al., 2021). This study attempts to bridge the gap between HPC and clouds by proposing a complementary approach of clouds HPC aware through virtual machine consolidation.

Power awareness is a problematic issue in HPC as it is a crucial factor in operational costs. For efficient management of resources, virtual machine consolidation is proposed to control the overloading and underloading of physical machines (PMs) caused by the random distribution of VMs. Mainly, the VM consolidation process handles the movement of some VMs from the overloaded physical machines (PMs) to the under-loaded PMs. Dynamic Virtual Machine consolidation is considered an efficient process to minimize energy consumption and enhance resource utilization in the data centres while maintaining the quality performance of HPC and non-HPC applications on the same platform. Because the problem of VM consolidation is mainly related to NP-hard, several heuristic algorithms have been suggested to solve the issue. However, all the proposed algorithms do not consider the variable nature of workloads in HPC and non-HPC applications in the cloud.

This introductory chapter presents the background to the study, which is focused on the High-Performance Computing (HPC) application performance in the cloud and explores the best techniques that would enable the effective utilization of a shared pool of resources for efficient execution of HPC and Non-HPC applications in the cloud. Proactive, dynamic application-aware VM consolidation strategies are examined to check how to be tailored to application characteristics. The chapter starts with an overview of the topic under study, followed by the problem background, problem statement, research objectives, research scope, significance, thesis organization and summary.

1.2 Problem Background

Running HPC applications on the cloud is accompanied by the problem of wasting energy and resources. The exponential demand for the cloud and its dynamic nature increase pressure on resource consumption. It is estimated that servers consume 0.5% of the total electricity usage worldwide. According to the International Energy Agency, data centres utilize 1% of all global energy demand and will consume 1/5 of the world's power supply by 2025 (IEA – International Energy Agency, n.d.). Data centre power utilization efficiency is evaluated in units of Power Usage Effectiveness (PUE). A PUE of 1.0 indicates complete efficiency (i.e., all consumed energy is used only on IT equipment, with no power distribution losses). However, it is nearly hard to achieve. According to the most recent annual Uptime Institute survey, the average power use effectiveness (PUE) ratio for a data centre in 2021 was 1.58, only modestly better than seven years previously (George, 2022).

The main drawback of cloud computing is its lack of a stringent energy efficiency regime (Sigh & Bal, 2017). This flaw is even aggravated when executing HPC applications on the cloud because they use more resources and computing power than conventional cloud applications (Netto et al., 2018). HPC applications need more computing power because they need a high CPU capacity, extensive memory and high network speed to execute correctly, and a particular execution mechanism. HPC applications do not run 24×7; instead, they usually are executed in batches (Netto et al., 2018). Simply put, a set of jobs are run by the users who wait until results are generated to decide the need for new jobs. Also, stacking several workloads on the same physical resources is not a proper approach for HPC applications, meaning they need a dedicated cloud that costs a lot of resources (Netto et al., 2018). Accordingly, running an HPC application on the cloud requires special resource allocation and optimizations in infrastructure.

For the HPC cloud applications to perform optimally, they require sufficient resources, which means the provision of adequate resources that are not less or more than needed. Maintaining sufficient resources is a bit difficult because the workload changes from time to time. Because of the users' usage patterns variance, a workload difference is likely to occur. The system must be adaptable to the workload changes to avoid the harm of overprovision or under-provision of resources. On the one hand, resources under-provision leads to violating the Service level agreement (SLA), thus less quality of services and customer dissatisfaction. On the other hand, the over-provision of cloud resources causes energy and cost waste. In this case, the cloud service costs a lot of energy and resources instead of being cost-effective.

The HPC research community has been giving special attention to the cost and computational performance of different cloud services (Egwutuoha et al., 2014). The available VM scheduling mechanisms depend on user inputs and static partitioning of clusters into zones of availability for various application types (such as HPC and non-HPC).

However, this process is problematic because a large-scale HPC application would ideally require a dedicated cloud resource, dramatically increasing operating costs. To overcome this problem, we aim to identify suitable application combinations whose execution profiles complement each other well and can be consolidated on the same hardware resources without harming the overall HPC performance. This way, the solution's business value is maximized by utilizing the hardware, lowering the cost for HPC applications while maintaining performance and profitability (Parashar et al., 2013).

Dynamic consolidation of Virtual Machines (Goscinski et al., 2012) is considered one of the best ways to improve resource utilization while optimizing energy efficiency in cloud data centres. This methodology influences the dynamism of Cloud workloads helping to reduce the quantity of active physical hosts needed to manage the workload while turning the idle hosts to low-power modes to reduce the total energy consumption. The hosts are active only when resources are needed. However, while reducing energy consumption is advantageous, it is essential to satisfy Quality of Service (QoS) requests.

In addition, massive amounts of data are being transferred between the infrastructure and consumers in Cloud applications; therefore, it is essential to view the computing and network aspects of energy efficiency carefully. Migrating VMs based on CPU utilization alone is not accurate enough because they will not be accurate for the network, Input and Output (I/O) and memory-intensive applications (Fernandes et al., 2016). Therefore, developing a solution to enhance energy efficiency based on hybrid factors (CPU, Memory, Bandwidth) is essential.

Many challenges emerge when running HPC applications in the cloud, such as changeable workload, assignment of nodes, random layout, unstable network flows, the shortage of low-latency interconnects, virtualization overhead optimization and the application-agnostic cloud schedulers (such as MOAB or LoadLeveler) (Skovira, 2000). For an efficient deployment of the HPC application, sharing resources is required. Currently, user inputs and static partitioning of clusters into availability zones for HPC and non-HPC applications are the base of VM scheduling mechanisms, making scheduling the main challenge to efficiently utilising resources. Fixed scheduling modules and strategies are now embedded in the structure of cloud platforms, which sometimes lead to underutilization of cloud resources harming the HPC performance. In the cloud, Software as a service (SaaS) workloads rely on request-level parallelism, which makes them different from HPC workloads that rely on a thread or task-level parallelism. As a result, HPC workloads are considered more communication-intensive and more sensitive to communication latency (Sigh & Bal, 2017). For this reason, more advanced scheduling techniques are required to use a shared pool of resources, providing a reconfigurable system.

1.3 Problem Statement

As discussed in the problem background, modern data centres are getting more complicated, which complicates the problem of energy usage, power efficiency, maximizing the performance of the workload and efficient utilization of resources. Thereby, there is a crucial need for dynamic state-of-the-art techniques to schedule a fusion of HPC and non-HPC applications smartly to maximize resource utilization and energy efficiency. Nevertheless, existing HPC cloud providers still rely on a separate dedicated pool of resources for HPC and non-HPC to maintain the high performance for HPC applications. Furthermore, depending on the variation in the input flux of HPC workload results in either under-utilized or over-utilized provision of resources in a cloud data centre. The trade-off between improved HPC performance vs better resource usage makes scheduling for HPC and non-HPC applications challenging. Considering these issues, the problem statement of the research is mentioned as follows:

How to develop dynamic scheduling techniques to classify suitable co-allocated applications (HPC and non-HPC) that can be executed on the same pool of resources while guaranteeing optimum utilization of resources, thus avoiding cross-interference, concurrently ensuring the high performance of applications and at the same time energy efficient?

In this study, it is hypothesized that the placement of VMs to physical machines, while considering the dynamic nature of applications and energy, will efficiently utilise virtualized cloud data centres.

Accordingly, several research questions can be pulled out from the problem statements are stated as follows:

i. How the suggested proactive, dynamic VM scheduling technique aids in overcoming the trade-off between optimum utilization while boosting the performance of HPC and non-HPC applications using a single pool of cloud data centre resources?

ii. How can the suggested proactive dynamic VM scheduling technique help with energy efficiency issues so that cloud data centre resources can be used to their maximum capacity while satisfying SLA?