

VISVESVARAYA TECHNOLOGICAL UNIVERSITY
JNANA SANGAMA, BELAGAVI - 590018



A Project Report on
Development of a Predictive VM Provisioning in a Cloud Environment

Submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Engineering
in
Electronics and Communication Engineering
for the Academic Year: 2023-24

Submitted by

Sai Likith P	(1NT20EC126)
Santhosh P	(1NT20EC133)
Sudeeksha K	(1NT20EC151)

Under the Guidance of

Dr. Karunakara Rai B
Professor
Dept. of Electronics and Communication Engineering

Dr. Rajani N
Assistant Professor
Dept. of Electronics and Communication Engineering



DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

YELAHANKA, BENGALURU- 560064

VISVESVARAYA TECHNOLOGICAL UNIVERSITY
JNANA SANGAMA, BELAGAVI - 590018



A Project Report on
Development of a Predictive VM Provisioning in a Cloud Environment

Submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Engineering
in
Electronics and Communication Engineering
for the Academic Year: 2023-24

Submitted by

Sai Likith P	(1NT20EC126)
Santhosh P	(1NT20EC133)
Sudeeksha K	(1NT20EC151)

Under the Guidance of

Dr. Karunakara Rai B
Professor
Dept. of Electronics and Communication Engineering

Dr. Rajani N
Assistant Professor
Dept. of Electronics and Communication Engineering



An Autonomous Institution Affiliated to VTU

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

YELAHANKA, BENGALURU- 560064



NITTE
EDUCATION TRUST

**NITTE MEENAKSHI
INSTITUTE OF TECHNOLOGY**

An Autonomous Institution Affiliated to VTU

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
BENGALURU- 560 064**

Certificate

Certified that the project work titled “**Development of a Predictive VM Provisioning in a Cloud Environment**” is carried out by **Sai Likith P (1NT20EC126)**, **Santhosh P (1NT20EC133)** and **Sudeeksha K (1NT20EC151)**, bonafide students of Nitte Meenakshi Institute of Technology in partial fulfilment for the award of Bachelor of Engineering in Electronics and Communication Engineering of Visvesvaraya Technological University, Belagavi during the academic year 2023-2024. The project report has been approved as it satisfies the academic requirement in respect of the project work prescribed as per the autonomous scheme of Nitte Meenakshi Institute of Technology for the said degree.

Signature of the Guide

Signature of the HoD

Signature of the Principal

Dr. Karunakara Rai B
Professor
Dept of ECE, NMIT

Dr. Parameshachari B D
HoD
Dept of ECE, NMIT

Dr. H C Nagaraj
Principal
NMIT

External Viva-Voce

Name of Examiners

Signature with Date

1.

2.

Acknowledgement

The successful execution of our project gives us an opportunity to convey our gratitude to each one who have been instrumental in paving path to our continuation of this project. Whatever we have done is due to such guidance and help and we would not forget to thank them all.

We would like to thank and seek the blessings from **Dr. N R. Shetty**, Advisor, **Nitte Meenakshi Institute of Technology**, for his thrust on project-based learning and constructivist principles in our institution.

We would like to express our gratitude to the **Nitte Meenakshi Institute of Technology** and our beloved Principal **Dr. H C. Nagaraj** for providing us the support, facilities and motivation to carry out our project.

We express our deep sense of gratitude to **Dr. Parameshachari B D**, HoD, Department of Electronics and Communication Engineering, for his kind co-operation, valuable guidance and creating best learning environment for us.

We whole heartedly thank our guide **Dr. Karunakara Rai B**, Professor, Department of Electronics and Communication Engineering, for his support and guidance anytime we required.

We also thank and share this moment of happiness with our parents who rendered us enormous support during the whole tenure of our studies at Nitte Meenakshi Institute of Technology, Bengaluru.

Finally, we would like to thank all other unnamed who helped us in various ways to gain knowledge and have a good training.

Sai Likith P (1NT20EC126)

Santhosh P (1NT20EC133)

Sudeeksha K (1NT20EC151)

Place: Bengaluru

Date:

Abstract

In contemporary virtualized data centre environments, the efficient allocation of resources is a critical concern to prevent server downtimes and maintain optimal performance levels. A recently published paper introduces an innovative method aimed at mitigating resource exhaustion on ESXi servers using advanced predictive analytics techniques. This approach stands out by leveraging historical workload data to forecast potential resource saturation events, analysing past usage patterns to predict future demands accurately.

The core of this method lies in its proactive provisioning mechanism, which strategically deploys virtual machines (VMs) onto underutilized ESXi servers during anticipated high-load intervals. By anticipating resource needs based on historical trends, this pre-emptive strategy optimizes workload distribution, preventing resource bottlenecks and ensuring that servers operate within their optimal capacity thresholds.

The experimental validation of this approach demonstrates its effectiveness in enhancing the dependability and security of ESXi servers within virtualized environments. The method surpasses existing state-of-the-art techniques, showcasing its superiority in managing resource allocation challenges in dynamic computing infrastructures.

This proactive strategy significantly contributes to advancing resource management practices in virtualized infrastructures by mitigating risks associated with resource overuse. By addressing potential issues before they escalate into critical problems, this research plays a crucial role in maintaining uninterrupted operations and improving the overall reliability of computing environments, particularly in critical applications where downtime can have severe consequences. The proactive deployment of resources not only optimizes server performance but also enhances system resilience and security, setting a new benchmark for resource management in virtualized data centre environments.

Contents

Acknowledgement	i
Abstract	ii
List of Figures	iv
List of Tables	v
Chapter 1 Introduction	1
1.1 <i>Virtual Machines (VMs)</i>	2
1.2 <i>Cloud Computing</i>	3
1.3 <i>VM Provisioning</i>	4
1.4 <i>Workload Pattern</i>	5
1.5 <i>Host Resource Usage Prediction</i>	6
1.6 <i>VM Allocation</i>	6
1.7 <i>Motivation</i>	7
1.8 <i>Organization of the Report</i>	8
Chapter 2 Literature Survey	9
2.1 <i>Background Work</i>	9
2.2 <i>Open Issues and Challenges</i>	9
2.3 <i>Problem Definition</i>	10
2.4 <i>Objectives</i>	18
2.5 <i>Scope of the Work</i>	18
Chapter 3 Design Approach and Methodology	19
3.1 <i>Design Approach</i>	19
3.2 <i>Methodology</i>	20
Chapter 4 Implementation Details	22
Chapter 5 Results and Analysis	26
Chapter 6 Conclusion and Future Scope	30
Bibliography	31
Appendix-A	34
Appendix-B	35
Publication Details	36

List of Figures

Figure 1 H2H VM Provisioning.....	19
Figure 2 Proposed Solution for VM Provisioning	21
Figure 3 CPU Usage Actual vs Predicted plot.....	24
Figure 4 Sum Of Resource Usage.....	25
Figure 5 Host 1 consumption before VM Migration.....	26
Figure 6 CPU Average of Host 1 after moving VM1.....	27
Figure 7 CPU Average of Host 2 after moving VM1.....	27

List of Tables

Table 1 Host Information.....21

Table 2 VM Information.....23

Table 3 Accuracy of Models.....28

Chapter 1

Introduction

In recent years, the evolution of cloud computing has revolutionized how businesses manage their computing resources by providing on-demand access to scalable infrastructure. However, with the increasing adoption of virtual machines (VMs) in cloud environments, effectively managing and allocating resources has become a complex challenge. To address these issues, innovative technologies such as virtualization have emerged, enabling organizations to optimize resource usage and enhance overall performance. Datacentre virtualization, a fundamental component of modern cloud infrastructure, involves the creation, deployment, and management of virtualized data centers. This approach virtualizes physical servers and utilizes advanced computing technologies to efficiently manage storage, networking, and other critical infrastructure components. However, co-hosting multiple types of VMs on a limited number of servers often leads to resource conflicts between co-hosted applications, resulting in server overutilization and degraded application performance.

Moreover, many cloud services, particularly interactive applications, exhibit regularly shifting workload requirements, leading to dynamic resource demands. Dynamic server consolidation strategies, while aiming to optimize resource utilization, may inadvertently cause Service Level Agreement (SLA) violations and performance degradation. To address these challenges, predictive analytics leverages historical data to forecast future resource requirements and allocate resources preemptively. By analyzing past consumption trends, predictive VM provisioning ensures that sufficient resources are available to meet anticipated demand, thereby optimizing resource utilization and application performance. This proactive approach is particularly crucial in sectors like banking, where server availability during peak periods—such as high-volume transactions or system updates—is paramount to maintaining operations.

In response to sudden resource spikes or server failures, the hypervisor identifies suitable VMs or sets of VMs and relocates them from heavily utilized servers to underutilized ones, aiming to enhance overall system performance and reliability. Predictive VM provisioning anticipates these demand fluctuations and ensures that adequate resources are allocated in advance, preventing service disruptions and maintaining uninterrupted operations. Effective VM allocation methods rely on statistical data derived from observed job patterns and system metrics to anticipate future resource requirements and allocate VMs accordingly. This proactive approach minimizes the risk of

SLA violations and optimizes resource utilization efficiency. Additionally, strategic VM selection techniques play a vital role in resource management within cloud environments by rebalancing workloads across active nodes without causing downtime or service interruptions.

A survey of predictive VM provisioning strategies, as outlined in recent research, encompasses various prediction algorithms aimed at selecting and supplying VMs to optimize server performance and resource allocation. This survey contributes to a comprehensive assessment of cutting-edge predictive virtual machine provisioning strategies, highlighting their advantages, disadvantages, and addressing important research gaps.

The primary advancements highlighted in this paper include:

1. Thorough analysis of literature on cutting-edge predictive VM provisioning strategies, examining their strengths, weaknesses, and identifying key research gaps.
2. Discussions on security risks, server failures, and current mitigation techniques employed in cloud environments.
3. Identification of specific research challenges and gaps to reduce server downtime and improve overall cloud infrastructure reliability.

These advancements collectively contribute to advancing resource management practices in cloud computing environments, ensuring efficient resource allocation, improved performance, and enhanced reliability in critical computing applications.

1.1 Virtual Machines (VMs)

Virtual machines (VMs) are a cornerstone of modern computing infrastructure, offering a versatile and scalable approach to managing computing resources. A VM is a software-based emulation of a physical computer system, complete with its own virtualized hardware components, including CPU, memory, storage, and network interfaces. VMs enable organizations to run multiple operating systems and applications on a single physical server, maximizing resource utilization and flexibility. One of the key benefits of VMs is their ability to abstract hardware resources from underlying physical infrastructure, allowing for greater flexibility and agility in resource allocation. This abstraction enables IT administrators to provision, configure, and manage VMs independently of the underlying hardware, making it easier to scale resources up or down based on changing demands.

VMs also facilitate efficient resource utilization by enabling workload consolidation and isolation. By running multiple VMs on a single physical server,

organizations can achieve higher levels of resource utilization and reduce hardware sprawl. Additionally, VMs provide isolation between workloads, minimizing the impact of failures or security breaches on other VMs running on the same server.

Furthermore, VMs offer a cost-effective solution for IT infrastructure management. By consolidating workloads onto fewer physical servers, organizations can reduce hardware and operational costs associated with maintaining multiple physical machines. VMs also support rapid provisioning and deployment, enabling organizations to quickly spin up new VM instances to meet changing business requirements. In addition to their role in traditional data centre environments, VMs play a crucial role in cloud computing platforms. Cloud providers leverage VM technology to offer scalable and on-demand computing resources to customers, enabling organizations to quickly deploy and scale applications without the need for upfront hardware investment.

Overall, VMs are a powerful and versatile tool for managing computing resources, offering benefits such as flexibility, efficiency, and cost-effectiveness. As organizations continue to embrace digital transformation and cloud adoption, VMs will remain a fundamental building block of modern IT infrastructure.

1.2 Cloud Computing

Cloud computing is a transformative paradigm in the field of information technology, revolutionizing how businesses and individuals' access, manage, and utilize computing resources. At its core, cloud computing involves the delivery of computing services over the internet, providing on-demand access to a wide range of resources, including servers, storage, databases, networking, software, and more. One of the key advantages of cloud computing is its scalability and flexibility. Cloud providers offer a vast array of resources that can be scaled up or down dynamically to meet changing demands, allowing organizations to pay only for the resources they use. This flexibility enables businesses to quickly adapt to fluctuations in workload and respond to evolving business needs without the need for upfront capital investment in hardware or infrastructure.

Cloud computing also offers significant cost savings compared to traditional on-premises infrastructure. By leveraging the economies of scale and sharing resources across multiple users, cloud providers can offer computing services at a lower cost than organizations could achieve on their own. Additionally, cloud computing eliminates the need for organizations to invest in hardware, maintenance, and ongoing infrastructure management, further reducing operational expenses.

Another key benefit of cloud computing is its accessibility and ubiquity. Cloud services can be accessed from anywhere with an internet connection, enabling remote work, collaboration, and mobility. This accessibility empowers organizations to scale globally, reach new markets, and serve customers more effectively, regardless of geographical location. Furthermore, cloud computing offers enhanced reliability, security, and disaster recovery capabilities. Cloud providers invest heavily in robust infrastructure, redundancy, and security measures to ensure high availability and protect against data loss or breaches. This enables organizations to achieve higher levels of reliability and security than they could achieve with on-premises solutions. Cloud computing has also democratized access to advanced technologies such as artificial intelligence, machine learning, big data analytics, and the Internet of Things (IoT). Cloud providers offer a wide range of managed services and tools that enable organizations to harness these technologies without the need for specialized expertise or infrastructure.

In summary, cloud computing is a transformative force that has fundamentally changed the way organizations approach IT. By providing scalable, flexible, and cost-effective computing services, cloud computing empowers organizations to innovate, compete, and thrive in the digital economy. As technology continues to evolve, cloud computing will remain a cornerstone of modern IT infrastructure, driving innovation and enabling organizations to achieve their goals more efficiently and effectively.

1.3 VM Provisioning

Virtual machine (VM) provisioning is a pivotal aspect of modern IT infrastructure management, revolutionizing how organizations deploy and manage computing resources. At its essence, VM provisioning entails the creation, configuration, and deployment of virtual machines within a computing environment. This process allows organizations to efficiently allocate resources such as CPU, memory, storage, and networking to support diverse workloads and applications. One of the key advantages of VM provisioning is its agility and scalability. Virtual machines can be provisioned rapidly and dynamically, enabling organizations to respond quickly to changing business requirements and fluctuations in workload demand. This flexibility empowers organizations to scale resources up or down as needed, optimizing resource utilization and minimizing operational costs.

Moreover, VM provisioning facilitates workload isolation and consolidation, enabling organizations to run multiple operating systems and applications on a single physical server. By consolidating workloads onto fewer physical machines, organizations

can maximize resource utilization, reduce hardware sprawl, and achieve significant cost savings. Additionally, VM provisioning supports automation and orchestration, streamlining the process of deploying and managing virtual machines at scale. Automation tools and platforms enable organizations to define standardized templates and configurations for VM deployment, reducing manual intervention and human error. This automation enhances operational efficiency, accelerates time-to-market, and enables organizations to focus on strategic initiatives rather than routine administrative tasks. Furthermore, VM provisioning plays a crucial role in disaster recovery and business continuity planning. By maintaining virtual machine images and configurations, organizations can quickly restore systems and applications in the event of hardware failure, data loss, or other disruptive events. This enables organizations to minimize downtime, mitigate risk, and ensure uninterrupted service delivery to customers and stakeholders.

In addition to its operational benefits, VM provisioning enables organizations to embrace emerging technologies such as cloud computing, containerization, and serverless computing. Virtualization platforms provide a foundation for building and deploying cloud-native applications, enabling organizations to leverage the scalability, flexibility, and cost-effectiveness of cloud computing services.

In summary, VM provisioning is a fundamental component of modern IT infrastructure management, empowering organizations to deploy, manage, and scale computing resources efficiently and effectively. By embracing VM provisioning, organizations can enhance agility, optimize resource utilization, and drive innovation in the digital age.

1.4 Workload Pattern

Workload pattern analysis is a critical aspect of managing computing environments, providing insights into the behaviour and characteristics of tasks or activities over time. Workload patterns encompass a wide range of factors, including the frequency, intensity, duration, and variability of workloads, as well as any discernible trends or anomalies. By analysing workload patterns, organizations can gain valuable insights into the utilization of computing resources, identify peak usage periods and anticipate future demand. This analysis often involves collecting and analysing data from various sources, such as system logs, performance metrics, user interactions, and application behaviour. Advanced analytical techniques, such as time series analysis, machine learning, and statistical modelling, can be employed to uncover hidden patterns and relationships within the data. Additionally, workload pattern analysis enables organizations to identify opportunities for

workload consolidation, optimization, and efficiency improvements. By understanding the underlying patterns of workload behaviour, organizations can make informed decisions about resource provisioning, capacity planning, and infrastructure optimization, ultimately enhancing the performance, reliability, and cost-effectiveness of their computing environments.

1.5 Host Resource Usage Prediction

Host resource prediction is a multifaceted and critical process within the realm of modern IT infrastructure management, constituting the foresight and anticipation of future resource requirements for computing hosts embedded within a networked environment. This intricate procedure hinges upon the meticulous analysis of historical data, utilization patterns, and performance metrics to extrapolate and envisage potential fluctuations in demand across key computational resources, including CPU, memory, storage, and network bandwidth. By delving into past utilization trends and discerning recurrent patterns, predictive algorithms are adept at identifying pivotal junctures characterized by heightened resource consumption, thereby facilitating the forecasting of forthcoming resource needs with a notable degree of precision. This proactive approach equips organizations with the ability to strategically allocate resources, proactively provisioning additional capacity during anticipated periods of heightened demand to circumvent potential performance bottlenecks or resource scarcities. Host resource prediction serves as the cornerstone of effective capacity planning, affording organizations the opportunity to align their infrastructure scalability with projected demand, thereby ensuring the sustenance of optimal performance and reliability. Furthermore, the employment of advanced predictive analytics techniques, such as machine learning and statistical modelling, enables organizations to refine and enhance the accuracy of resource predictions over time, thereby enabling the optimization of resource utilization, cost minimization, and augmentation of overall operational efficiency. In essence, host resource usage prediction stands as a linchpin of proficient IT infrastructure management, empowering organizations to navigate and respond adeptly to the dynamic landscape of changing workload demands, optimize resource usage allocation, and uphold the seamless functionality of their computing environments.

1.6 VM Allocation

Virtual machine (VM) allocation is a critical aspect of managing computing resources within a virtualized environment. It involves the strategic assignment of virtual machines

to physical servers or hosts based on various factors such as workload distribution, resource availability, and performance requirements. VM allocation aims to optimize resource utilization, improve performance, and ensure the efficient use of infrastructure resources. This process begins with analysing workload patterns and resource utilization data to identify the optimal placement of VMs across the available servers. By considering factors such as CPU usage, memory requirements, storage capacity, and network bandwidth, administrators can determine the most suitable host for each VM.

Furthermore, VM allocation techniques may involve dynamic resource allocation, where VMs are provisioned or migrated between hosts in real-time to adapt to changing workload demands. This dynamic allocation ensures that VMs are efficiently distributed across the infrastructure to avoid resource bottlenecks and maximize system performance. Additionally, VM allocation strategies may take into account factors such as fault tolerance, high availability, and workload balancing to ensure system reliability and resilience. Advanced VM allocation algorithms and policies may also be employed to automate the allocation process and optimize resource utilization dynamically. These algorithms may use machine learning, predictive analytics, or heuristic approaches to predict future resource demands and make proactive allocation decisions. By leveraging these techniques, organizations can achieve better resource utilization, improved performance, and increased scalability in their virtualized environments. Moreover, VM allocation plays a crucial role in cloud computing environments, where resources are shared among multiple users and organizations. Cloud service providers must efficiently allocate resources to meet the diverse needs of their customers while ensuring fair resource distribution and optimal performance. Through effective VM allocation strategies, cloud providers can enhance customer satisfaction, minimize costs, and maximize the overall efficiency of their cloud infrastructure.

In summary, VM allocation is a complex and multifaceted process that requires careful consideration of various factors to optimize resource utilization, performance, and reliability within virtualized environments. By employing advanced allocation techniques and algorithms, organizations can achieve better resource management, scalability, and cost-effectiveness in their IT infrastructure.

1.7 Motivation

The motivation behind the project described in the provided text is rooted in addressing the complex challenges of resource management within modern cloud computing environments. With the widespread adoption of virtual machines (VMs), managing

resources effectively has become increasingly difficult, leading to issues such as server overutilization and resource conflicts between co-hosted applications. The project seeks to optimize performance by pre-emptively allocating resources based on predictive analytics, aiming to meet service level agreements (SLAs) and ensure consistent application performance. Dynamic workload requirements further complicate resource allocation, making it crucial to forecast future demands accurately using historical data. In sectors like banking, maintaining reliable cloud infrastructure during peak periods is essential for uninterrupted operations, underscoring the importance of proactive resource management. The project also aims to address research gaps in predictive VM provisioning strategies, resilience, and overall reliability within cloud environments. By advancing resource management practices and mitigating critical issues, the project aims to contribute to the evolution and improvement of cloud computing infrastructure.

1.8 Organization of the Report

The project report is structured as follows:

Chapter 1 serves as an introduction, outlining the project's objectives and significance. Chapter 2 conducts a thorough literature survey, reviewing existing research and technologies related to predictive VM provisioning in cloud computing. In Chapter 3, the design and methodology of the project are detailed, explaining the approach taken to achieve the research goals. Chapter 4 focuses on implementation, providing technical insights into how the proposed solution was developed and executed. Chapter 5 presents the project's results, including data analysis and interpretations. Finally, Chapter 6 concludes the report by summarizing findings, discussing implications, and suggesting future research directions.

Chapter 2 Literature Survey

2.1 Background Work

The adoption of cloud computing has transformed resource management for businesses, offering scalable infrastructure on demand. However, as virtual machine (VM) usage grows, effective resource allocation becomes challenging. Virtualization technologies, like datacentre virtualization, help optimize resource usage by creating virtual data centres. Predictive VM provisioning leverages historical data to anticipate future resource needs, ensuring optimal performance. In critical sectors like banking, predictive VM provisioning is crucial to prevent service disruptions during peak demand. Effective VM allocation methods and VM selection techniques further enhance resource management, minimizing downtime and ensuring uninterrupted service delivery.

2.2 Open Issues and Challenges

- **Scalability:** As cloud environments continue to grow in complexity and scale, ensuring effective resource management becomes increasingly challenging. Managing a large number of VMs and diverse workloads while maintaining optimal performance and resource utilization remains a significant challenge.
- **Dynamic Workloads:** Cloud environments often experience fluctuations in workload demand, making it difficult to predict resource requirements accurately. Adapting to dynamic workloads and ensuring timely provisioning of resources to meet fluctuating demand poses a challenge for resource management systems.
- **Resource Contention:** In multi-tenant cloud environments, resource contention can occur when multiple VMs compete for the same pool of resources. Managing resource contention and ensuring fair allocation of resources among competing VMs while maintaining performance SLAs presents a complex challenge.
- **Cost Optimization:** While cloud computing offers scalability and flexibility, it also introduces cost implications. Optimizing resource allocation to minimize costs while meeting performance requirements is a continuous challenge for organizations operating in cloud environments.
- **Security and Compliance:** Ensuring the security and compliance of VMs and data in cloud environments is critical. Addressing security concerns such as data breaches, unauthorized access, and compliance requirements adds complexity to resource management efforts.

- **Technological Evolution:** The rapid pace of technological advancement in cloud computing introduces new tools, techniques, and paradigms that organizations must adapt to. Keeping abreast of emerging technologies and evolving best practices in resource management is an ongoing challenge.

2.3 Problem Definition

[1] Maximizing the Energy Efficiency of Virtualized C-RAN via Optimizing the Number of Virtual Machines

R. S. Alhumaima, R. K. Ahmed and H. S. Al-Raweshidy

In their seminal work, the authors of paper [1] addresses the challenge of optimizing the allocation of virtual machines (VMs) in cloud servers to enhance energy efficiency in cloud radio access networks. Employing Monte Carlo-based evolutionary algorithms such as particle swarm optimization, and genetic algorithms, the study aims to find the optimal number of VMs for energy-efficient operation. The model shows the impact of adding a greater number of VMs on server performance, particularly in processing resource blocks (RBs) per VM. Advantages of the model include its support for energy efficiency and improvement in quality-of-service metrics such as minimum user equipment data rate, allocated Resource blocks, and latency due to virtualization. However, shortcomings are noted in VM placement and power allocations. Additionally, the paper proposes a power model to assess power usage in virtualized server components, aiming to simplify evaluation processes.

[2] Energy-Aware VM Consolidation in Cloud Data Centers Using Utilization Prediction Model

F. Farahnakian, T. Pahikkala, P. Liljeberg, J. Plosila, N. T. Hieu and H. Tenhunen

As elucidated by the authors in paper [2] they explored VM consolidation as a means to enhance resource utilization, utilizing a regression model to forecast future CPU and memory usage. Through analysis of real workload traces from Google cluster and Planet Lab, we implement the UP-VMC approach, aiming to minimize unnecessary VM migrations. The results demonstrate that VM consolidation can significantly reduce energy consumption, up to 71.6% compared to alternative methods, by consolidating VMs onto the most heavily loaded physical machines. However, challenges remain in terms of scalability and optimizing VM placement, particularly regarding network resource utilization and traffic management. Additionally, the authors introduce the UP-VMC strategy to optimize

the energy efficiency of Cloud Radio Access Networks (C-RAN) by determining the optimal number of virtual machines.

[3] Dynamic Resource Prediction and Allocation for Cloud Data Center Using the Multiobjective Genetic Algorithm

F. -H. Tseng, X. Wang, L. -D. Chou, H. -C. Chao and V. C. M. Leung

In paper [3], a Multi objective genetic algorithm (GA) is employed to predictively estimate how resources are used and energy is consumed in cloud data centers in real time, considering CPU and memory utilization of both virtual machines (VMs) and physical machines (PMs). The study results in the development of an algorithm for placing virtual machines designed to enhance overall resource utilization and decrease energy consumption within the data center by leveraging prediction outcomes from genetic algorithms. The approach enhances PM utilization while decreasing energy consumption by minimizing the number of active PMs. However, limitations exist in the accuracy of the approach in realistic cloud environments, as it has not been tested with real-world data center traces such as those from Google. Additionally, the authors propose a novel GA-based prediction strategy to enhance forecast precision in cloud data centers and suggest a VM placement strategy based on GA predictions to optimize resource utilization and energy consumption.

[4] Predicting Multi-Attribute Host Resource Utilization Using Support Vector Regression Technique

L. Abdullah, H. Li, S. Al-Jamali, A. Al-Badwi and C. Ruan

Expounding on the findings presented in paper [4] addresses the challenge of forecasting future cloud resource utilization by employing the Support Vector Regression Technique (SVRT) with a Radial Basis Function (RBF) kernel function. This method aims to accurately predict multi-attribute host resource utilization, particularly suited for nonlinear workload patterns. The Sequential Minimal Optimization Algorithm (SMOA) is applied for training and regression estimation to enhance prediction accuracy. By utilizing SVRT with the RBF kernel function, the research suggests a method less susceptible to fluctuations in resource utilization compared to other forecasting techniques. Through experimentation with eight datasets, the research investigates variations in workload demand and resource utilization within a cloud setting, emphasizing the opportunity to minimize resource waste and enhance resource efficiency in cloud data centers.

[5] Predicted Affinity Based Virtual Machine Placement in Cloud Computing Environments

X. Fu and C. Zhou

Embarking on their analysis in paper [5] the author develops a NICBLE-based system prototype for Xen virtualization to assess the impact of hypothetical changes in VM setups on central processor quantities. NICBLE predicts application execution based on simulated calculations. The Priority-Aware VM Allocation (PAVA) technique is proposed for VM allocation, leveraging network topology information to assign critical applications to closely connected hosts. The model further explores relationships between VMs based on ARIMA-predicted resource requirements, introducing an affinity model to evaluate resource utilization volatility when VMs are placed on the same host. The resultant algorithm for virtual machine placement, rooted in predicting affinity, consolidates VMs with strong connections onto single physical machines, maximizing resource usage while staying within the limits of physical machine capacities. Extensive simulation experiments validate the algorithm's effectiveness in reducing energy consumption, VM migrations, and SLA violations based on Planet lab and Google workload traces. However, further adjustments are needed to improve the accuracy of the prediction model.

[6] An Effective Classification-Based Framework for Predicting Cloud Capacity Demand in Cloud Services

B. Xia, T. Li, Q. Zhou, Q. Li and H. Zhang

The authors in [6] have released the ground-breaking Forecast Empathy Based Virtual Machine Positioning algorithm. By using this technique, a single physical machine is created from the high-affinity virtual machines. Estimating a request within a range is pointless because the projection will continuously stray from the actual requests and serve as a guide.

[7] Application Execution Time Prediction for Effective CPU Provisioning in Virtualization Environment

H. -W. Li, Y. -S. Wu, Y. -Y. Chen, C. -M. Wang and Y. -N. Huang

Within the context of paper [7], the researchers utilize the Support Vector Relapse Strategy (SVRT), a managed factual learning technique, to foresee how the multi-property have asset will be utilized from here on out. Utilizing cloud resources to manage a non-linear workload is a perfect application for this strategy. We pick Spiral Premise Capability as the bit capability of SVRT and utilize Consecutive Negligible Streamlining Calculation (SMOA)

for the preparation and relapse assessment of the expectation strategy to build the expectation precision of SVRT.

[8] Markov Prediction Model for Host Load Detection and VM Placement in Live Migration

S. B. Melhem, A. Agarwal, N. Goel and M. Zaman

Addressing the subject matter in paper [8] they proposed the use of a Markov prediction model to forecast the future load state of hosts in cloud environments. We introduce a host load detection algorithm to identify over or under-utilized hosts, aiming to prevent immediate VM migration. A VM algorithm for placement is then employed to select candidate hosts for migrated VMs, evaluated using cloud Sim simulation. Our approach relies on dynamic utilization thresholds to enhance the VM placement process and predict results to avoid host overload shortly after migration.

[9] SLA-Aware and Energy-Efficient VM Consolidation in Cloud Data Centers Using Robust Linear Regression Prediction Model

L. Li, J. Dong, D. Zuo and J. Wu

The Median Absolute Deviation Markov Chain Hot Detection technique (MadMCHD) to improve host detection performance during live migration by determining future overutilized and underutilized states is used. By incorporating the RobustSLR prediction algorithm into the PABFD algorithm, authors altered both the existing VM placement method and the new VM algorithm for placement [9].

[10] Dynamic VM Scaling: Provisioning and Pricing through an Online Auction

X. Zhang, Z. Huang, C. Wu, Z. Li and F. C. M. Lau

In the situation of a 140% system load, the suggested method can handle up to 12% more user requests and generate up to 8% more system rewards [10].

[11] Efficient resource provisioning for elastic Cloud services based on machine learning techniques

Moreno-Vozmediano, R., Montero, R.S., Huedo, E. et al

A comprehensive analysis by [11] introduces a predictive auto-scaling mechanism for elastic cloud services, integrating time series forecasting using machine learning methods with queuing theory to enhance service response times and reduce unnecessary resource allocation. This approach utilizes SVM regression to anticipate the workload of web servers

by analyzing past data. Results indicate that SVM-based forecasting models perform better than basic models in terms of reducing over-provisioned resources. However, some SVM-based models exhibit poorer performance regarding SLA violations and unserved requests. Advantages include better prediction accuracy compared to simpler methods, as demonstrated by MAE and RMSE error measurements. Nonetheless, challenges remain in adapting predicting and performance models to the diverse components and functionalities of other cluster architectures such as MapReduce, HDFS, and Spark components.

[12] Time series-based workload prediction using the statistical hybrid model for the cloud environment

Devi, K.L., Valli, S

According to the research presented in [12] the authors employed a hybrid ARIMA–ANN model to predict future CPU and memory utilization in cloud environments, leveraging both linear and nonlinear components in the data. Analysis and experiments are conducted using workload traces from Google trace and Bit Brain compute clusters, focusing on CPU and memory usage. While the ARIMA model detects linear patterns, the ANN is effective in predicting nonlinear patterns by leveraging residuals from the ARIMA model. Advantages include improved prediction accuracy for nonlinear patterns with ANN, although reliance on training data may affect performance. To address this, the proposed hybrid model combines the strengths of both ARIMA and ANN. However, challenges remain in dynamically adjusting the sliding window size and assigning weights to recent historical data points to further enhance prediction accuracy.

[13] Approximation Algorithms to Distributed Server Allocation With Preventive Start-Time Optimization Against Server Failure

Souhei Yanase, Graduate Student Member, Fujun He and Eiji Oki

The endeavor to optimize server allocation in the face of potential failures has prompted the development of advanced algorithms aimed at minimizing delays and maximizing system efficiency. In this context of [13], recent research has proposed polynomial-time approximation algorithms leveraging Particle Swarm Optimization (PSO) to tackle the distributed server assignment problem in the event of single server failures. The NP-completeness of the problem has been established, underscoring the complexity of the task at hand. The proposed algorithms demonstrate promising approximation performances, offering significant speed enhancements compared to traditional Integer Linear Programming (ILP) approaches. Numerical analyses indicate substantial efficiency gains,

with the proposed algorithms outperforming ILP by factors ranging from 3.0×10^3 to 1.9×10^7 while slightly increasing the largest maximum delay by a factor of 1.033 on average. These findings shed light on novel strategies for efficiently addressing server allocation challenges in distributed systems, paving the way for enhanced reliability and performance in cloud computing environments.

[14] Evaluate Solutions for Achieving High Availability or Near Zero Downtime for Cloud Native Enterprise Applications

Antra Malhotra, AMR Elsayed, Randolph Torres and Srinivas Venkatraman

The scholarly contribution of [14] the authors delve into the intricacies of ensuring high redundancy and almost zero downtime for enterprise applications within cloud environments. They highlight the significant challenge of establishing surplus and robust architectures, both at the application and database layers. Emphasizing the importance of automated failover mechanisms, particularly for applications expected to operate around the clock, the authors propose a cloud-native template architecture for enhancing availability. Through their evaluation of automatic database failover techniques, they aim to minimize disruptions during planned maintenance activities and outages. While acknowledging the variability in achieving uninterrupted service based on factors like application size and data volume, the authors provide a foundational framework for building resilient applications. They stress the need for customization of the recommended architecture and failover strategies to align with each application's unique requirements, considering factors such as cost and specific business objectives.

[15] Fault-Tolerance in the Scope of Cloud Computing

A. U. Rehman Rui L. Aguiar, and JOÃO Paulo Barracca

An in-depth exploration of cloud computing, beginning with a comprehensive background to setup a foundational grasp of the subject. It delves into fault-tolerance components and system-level metrics, emphasizing their significance and applications within cloud computing environments. The authors meticulously examine both proactive and reactive approaches to fault-tolerance in cloud computing, showcasing the state-of-the-art techniques and frameworks. By organizing and discussing current research endeavors in fault-tolerance architectures, the paper offers valuable insights into the evolving landscape of cloud computing resilience. It concludes by outlining key future research directions, underscoring the importance of continued development in fault-tolerance strategies specific to cloud computing [15].

[16] An In-Depth Correlative Study Between DRAM Errors and Server Failures in Production Data Centers

Zhinan Cheng, Shujie Han, Patrick P. C. Lee, Xin Li, Jiongzhou Liu and Zhan L

The authors in [16], presents a comprehensive data-centric analysis correlating DRAM errors with server outages, aiming to predict server outages based on DRAM error patterns. Leveraging an extensive eight-month dataset from Alibaba's production data centers, comprising over three million memory modules, we identify that correctable DRAM errors often precede server failures, highlighting the importance of regular and frequent server failure prediction intervals for accurate forecasting. Our investigation also explores various factors influencing instances of server malfunctions, encompassing component breakdowns within the memory subsystem, variations in DRAM setups, and categories of fixable DRAM inconsistencies. Notably, we extend prior work by considering multiple types of server malfunctions in the prediction of server failures, achieving significant reductions in server downtime. Our findings reveal that UE-driven failures pose challenges in prediction due to the limited occurrence of correctable errors before these failures, whereas CE-driven and miscellaneous failures exhibit higher predictability. By employing all feature groups, we enhance prediction accuracy, with tree-based prediction models demonstrating superior performance, underscoring the importance of short prediction intervals for timely failure prediction.

[17] A Large-Scale Study of I/O Workload's Impact on Disk Failure

Song Wu, Yusheng Yi, Jiang Xiao, Hai Jin, and Mao Ye

Delving into the intricate the correlation between input/output (I/O) workload characteristics and disk reliability, aiming to uncover factors influencing disk lifespan and identify detrimental I/O workloads. By proposing an innovative measure, AISR (Average I/O Service Rate), we shed light on "dangerous" I/O workloads posing significant risks to disk health. Our research represents a pioneering effort in comprehensively analyzing how input/output (I/O) workload influences disk dependability, providing valuable perspectives to improve I/O scheduling strategies within data centers. While our work marks an initial exploration in this domain, we anticipate that our findings will stimulate further research in the community, prompting a reevaluation of disk I/O workload assignments. Future endeavors will focus on incorporating additional workload metrics to extract actionable insights and implement them effectively in data center environments [17].

[18] Service-Aware Cloud-to-Cloud Migration of Multiple Virtual Machines

Jargalsaikhan Narantuya, Hannie Zang, and Hyuk Lim

As documented in [18] the research introduces a strategy for determining the sequence of migrations, utilizing network traffic data between virtual machines to reduce downtime by 50% during cloud-to-cloud migration. Employing Prim's algorithm, they identify the Minimum Spanning Tree (MST) for a weighted undirected graph, facilitating efficient migration. The controller nodes, equipped with Intel Xeon W35653.25GHz CPU, 48GB RAM, and 1TB SSD, contribute to achieving shorter service downtime during migration. The strategy accounts for migrating multiple VMs, addressing communication dependencies among them. While the approach yields shorter service downtime, it may sometimes necessitate higher security measures, potentially leading to increased service downtime.

[19] Characteristics of Co-Allocated Online Services and Batch Jobs in Internet Data Centers: A Case Study From Alibaba Cloud

Congfeng Jiang, Guangjie Han, Jiangbin Lin, Gangyong Jia, Weisong Shi, and Jian Wani

Through their meticulous study in [19], unveils an analysis among the services that are online and batch jobs, which are collocated within a production cluster in Alibaba Cloud. By clustering servers based on CPU and memory utilization correlations, it identifies opportunities for job co-allocation and resource estimation. Through examination of mean time between failures (MTBF) and completion times, it sheds light on failure distribution and workload assignment disparities. The insights gleaned from this analysis can empower data center operators to optimize resource utilization and enhance failure recovery mechanisms, ultimately fostering a deeper understanding of workload characteristics and operational efficiencies within cloud environments.

[20] New Metrics for Disk Failure Prediction That Go Beyond Prediction Accuracy

Jing Li, Rebecca J. Stones, Gang Wang, Zhongwei Li, Xiaoguang Liu, and Jianli Ding

In accordance with [20] the author evaluates the effectiveness of various disk-failure prediction methods using metrics such as timeliness and convergence. By comparing classification tree (CT), a neural network with recurrent connections and a regression tree model boosted using gradient descent models, the research highlights the nuanced performance differences across prediction experiments. While RNN exhibits superior accuracy, CT and GBRT models demonstrate advantages in resource-dependent migration rates. The findings underscore the importance of considering prediction accuracy in conjunction with practical outcomes, prompting the introduction of an enhanced GBRT

model (GBRT+). Moving forward, the exploration of urgency-weighted evaluations and adjustments to machine learning processes offer promising avenues for refining disk-failure prediction models.

2.4 Objectives

- Create VMs under specific hosts and collect workload traces.
- Identify future dates of overutilization of resources by VMs.
- Allocate VMs to under-provisioned hosts to optimize resource utilization.

2.5 Scope of the Work

The scope of the project encompasses the development and implementation of predictive VM provisioning strategies within cloud computing environments. Specifically, the project aims to enhance application performance, and ensure service availability by leveraging historical data and predictive analytics. The scope also includes conducting a comprehensive literature review, designing and implementing the proposed solution, and evaluating its effectiveness through experimentation and analysis. Additionally, the project aims to contribute to advancing research in cloud resource management and addressing critical challenges faced by organizations relying on cloud infrastructure.

Chapter 3 Design Approach and Methodology

3.1 Design Approach

The host operating system (OS) as shown in Figure 1 is the foundation of the H2H VM provisioning. This is the main operating system that is directly installed on the computer's Physical hardware. It might be Linux, macOS, or Windows.

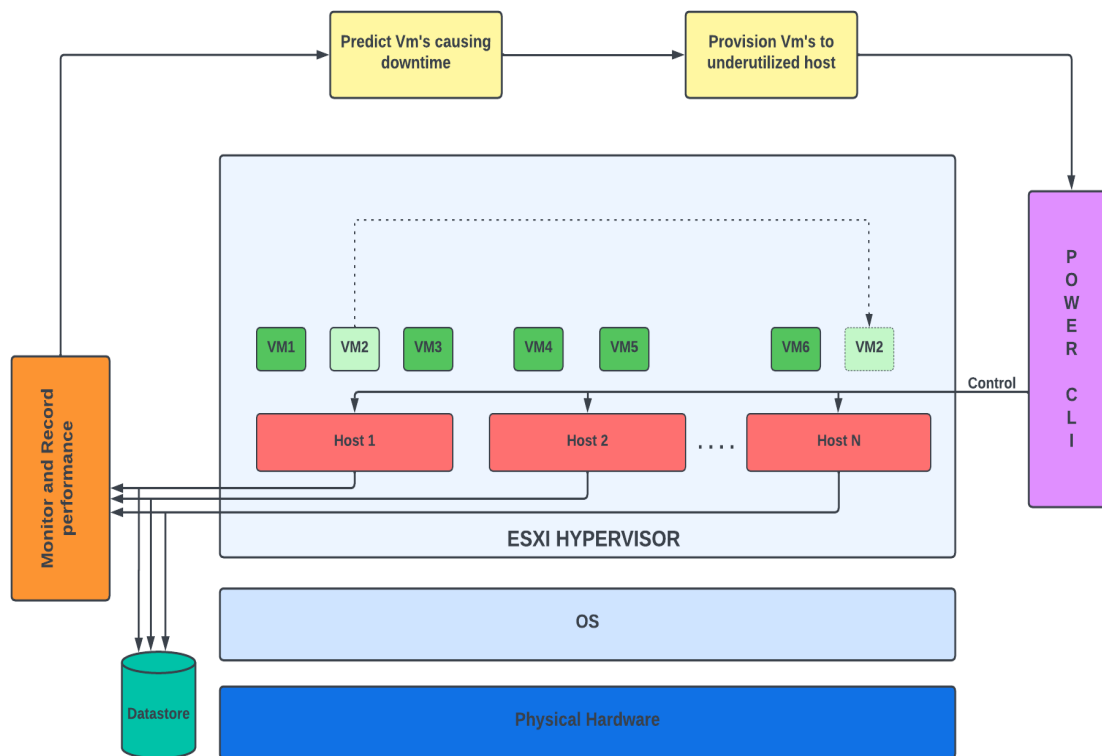


Figure 1: H2H VM Provisioning

The H2H VM provisioning serves as a versatile platform for virtualization, enabling the simultaneous execution of multiple operating systems on a single physical machine. At the core of this architecture lies the host operating system (OS), which acts as the foundational layer directly installed on the hardware. This host OS, whether it be Linux, macOS, or windows, provides the essential infrastructure necessary for virtualization to occur. Above the host OS sits the Type 2 hypervisor, functioning as an application layer. This hypervisor software facilitates the creation and management of virtual machines (VMs) within the host environment. With the Type 2 hypervisor in place, users can install and run various guest operating systems independently of each other and the host OS. This versatility allows for a diverse range of operating systems, including different versions of Windows, Linux distributions, and occasionally macOS, to be utilized within the virtualized environment. It oversees the distribution of hardware resources such as CPU cycles, memory, storage, and

network interfaces among the virtual machines as needed. Through efficient resource management, the hypervisor ensures optimal utilization of the underlying physical hardware, allowing multiple VMs to share resources without compromising performance or stability.

Central to the operation of virtual machines are datastores, which serve as repositories for virtual hard drives (VHDs) or disk image files. These datastores can be located on local storage devices or networked storage accessible by the host OS. By storing VM data and configurations in datastores, users can easily manage and deploy virtual machines across the virtualized environment. Users interact with the Type 2 hypervisor through a dedicated management interface provided by the hypervisor software. Through the management interface, users have full control over the virtualized environment, enabling them to customize and optimize their computing resources according to their specific requirements.

In summary, the Type 2 hypervisor architecture provides a robust foundation for virtualization, offering flexibility, scalability, and efficient resource utilization in diverse computing environments. By abstracting hardware resources and providing a management interface, it empowers users to create and manage virtual machines with ease, driving innovation and efficiency in modern computing systems.

3.2 Methodology

The proposed system is based on a real-time feed as inputs and follows the operational process shown in Figure 2.

- **Esxi Hypervisor Setup:** Esxi hypervisors are installed on physical hardware. Each hypervisor is configured with IP addresses (IPv4), DNS server addresses, and default gateways.
- **Datastore Creation:** Datastores are established to allocate memory to virtual machines (VMs) and houses the disk files of VMs. This allows for efficient management of resources.
- **VM Provisioning:** When a request for a new VM arises, resources are allocated based on customer demands. VMs are created and configured with unique IP addresses. ISO image files are used to load guest operating systems onto these VMs.

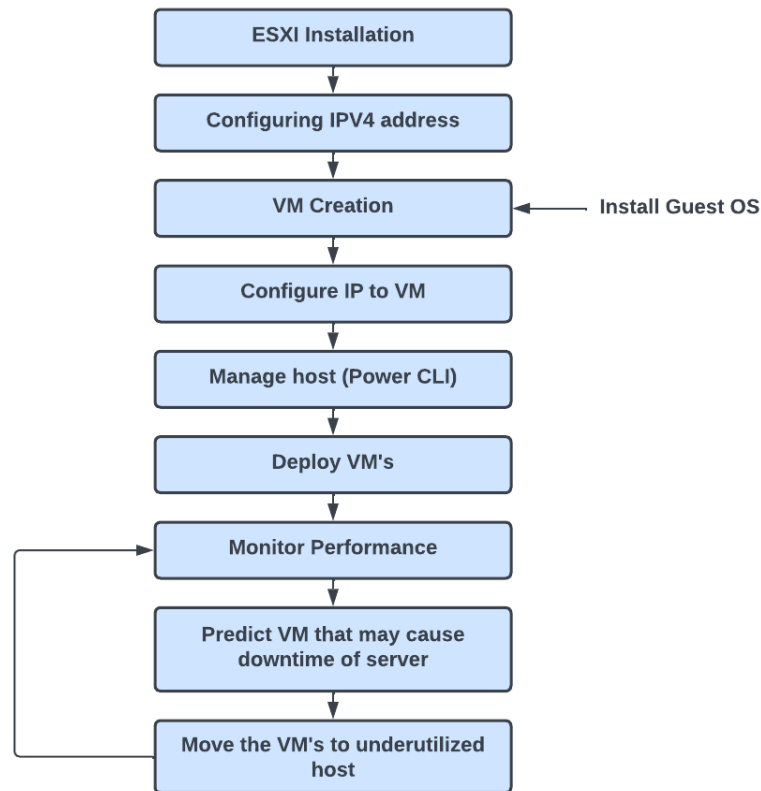


Figure 2: Proposed Solution for VM Provisioning

- **Power CLI:** This tool is used to automate and manage the Vsphere, Vcenter and VMware host client using power shell by logging in with credentials of the server.
- **Data Collection:** The usage the of host and VMs are recorded with the use of CLI script for a specific time interval for analyzing high resource usage (Memory, CPU, network, disk Usage) of the VMs in Over-Provisioned Host.
- **Template Creation:** Templates are created to expedite the provisioning process. These templates contain pre-configured settings and configurations, enabling quick deployment of new VMs within minutes.
- **VM Migration:** Identified VMs that are predicted to cause downtime are removed from host and provisioned to underutilized Esxi host. This proactive approach helps optimize resource utilization and mitigate potential disruptions to services.

Chapter 4 Implementation Details

Within the ESXi server version 8.0.2 environment, two distinct hosts are under consideration: remoteServer.localdomain and host2.localdomain. The former is designated with the static IPv4 address "192.168.204.2," while the latter is assigned the static address "192.168.204.8." Illustrated in Table 1 Host information, both hosts are endowed with 4 gigabytes of RAM and are equipped with 4 CPU cores. These resources constitute the foundational infrastructure of the virtualized environment, pivotal for the operation of virtual machines (VMs) and the management of computational resources. With their specified RAM and CPU allocations, these hosts are primed to concurrently support multiple VMs, facilitating the optimal utilization of hardware resources. Moreover, the utilization of static IPv4 addresses bolsters network stability and consistency, fostering seamless communication and connectivity throughout the ESXi server environment. This configuration lays a robust groundwork for hosting a diverse array of virtualized applications, services and management across the network.

Table 1: Host Information

Host Name	Num CPU	CPU Usage (MHz)	CPU Total (MHz)	Memory Usage (GB)	Memory Total (GB)
192.168.204.2	4	462	9980	1.59	3.99
192.168.204.8	4	156	9980	1.578	3.999

In the Esxi server environment, two virtual machines (VMs) named vm1 and WIN-2019-ser-2 are provisioned under host1, while a VM named h2_WIN-2019-ser-2 is singularly created on host2. As detailed in table 2 VM information, each VM is allocated a specific number of CPU cores, RAM, and resides within designated datastores. This strategic allocation ensures that each VM is equipped with the necessary computational resources to operate effectively within its respective host environment. For instance, vm1 and WIN-2019-ser-2 on host1 are configured with an optimal combination of CPU cores and RAM to support their individual workloads, while h2_WIN-2019-ser-2 on host2 is provisioned with resources tailored to its specific requirements. By distributing VMs across multiple hosts and appropriately allocating resources, the virtualized environment achieves enhanced performance and resource utilization. Additionally, the utilization of distinct

datastores for each VM facilitates efficient storage management, ensuring that VMs have access to the necessary storage capacity while maintaining data integrity and accessibility.

Table 2: VM Information

VM Name	VM id	ESXi Host	Data Store	Num CPU	Memory (GB)
Vm 1	10	192.168.204.2	Storage_datastore_1	2	1
WIN-2019-ser-vm-2	9	192.168.204.2	Storage_datastore_1	2	4
H2_WIN-2019-ser-vm-2	9	192.168.204.8	Storage_datastore_2	2	4

Power CLI is used to automate and manage the Vsphere, Vcenter and VMware host client using power shell by logging in with credentials of the server. The usage the of host and VMs are recorded with the use of CLI script for a specific time interval for analyzing high resource usage (Memory, CPU, network, disk Usage) of the VMs in Over-Provisioned Host. Collecting and preprocessing data from diverse sources like CPU usage, memory usage, and disk usage enables the prediction of future resource usage for VMs and the detection of those exceeding 80% resource consumption.

$$obj(y, \hat{y}) = \frac{1}{2} \sum (y_i - \hat{y}_i)^2 + \lambda \sum |w_j|^\gamma \quad (1)$$

where:

obj objective function in SG model minimizes both the training error and a regularization term to prevent overfitting.

y_i is the actual value of the i -th data point.

\hat{y}_i is the predicted value of the i -th data point by the model.

λ is the regularization parameter controlling the strength of the penalty term.

w_j is the weight of the j -th feature in the model

γ is a hyperparameter that determines the type of regularization.

Let $T(x_i)$ represent the prediction of the i -th data point by a single tree in the ensemble.

The ensemble prediction can be formulated as:

$$\hat{y}_i = \sum f_m(x_i) \quad (2)$$

where:

m iterates over all trees in the ensemble.

$f_m(x_i)$ is the prediction of the m -th tree for the i -th data point.

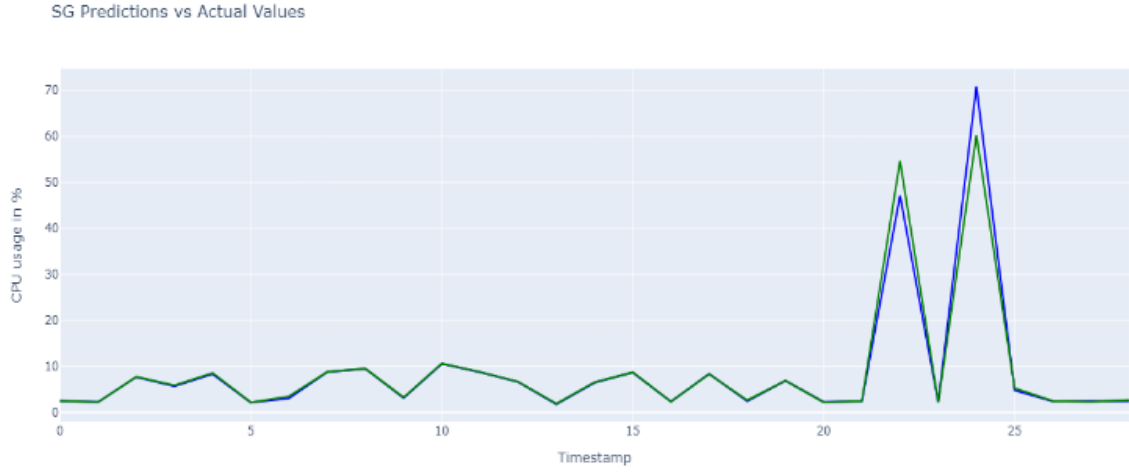


Figure 3: CPU Usage Actual vs Predicted plot

By analyzing the Figure 3, it observed the accuracy of model is better fit for the dataset. The predicted resource usage of these VM's is collected, analyzed and detected the VMs causing downtime of the Server as shown in Figure 4 sum of resource Usage of CPU, Memory and Disk Usage.

When the resource consumption of a host exceeds 80%, it evaluates the total resource usage (Rs_Usage) across all virtual machines within a specified time interval. This total usage (Rs_Usage) is calculated by summing the resource usage (Rs_Usage) of each VM at various time points (t_i) within the interval.

$$Total\ Rs_Usage(X) = \sum Rs_Usage(X, t_i) \quad (3)$$

The VM with the highest total resource usage (Rs_Usage) is identified as X_m , indicating that it's consuming the most resources.

$$Total\ Rs_Usage(X_m) >$$

$$Total\ Rs_Usage(X_1), Total\ Rs_Usage(X_2), \dots, Total\ Rs_Usage(X_n) \quad (4)$$

Where $X_1, X_2, X_3, \dots, X_n$ are n different virtual machines.

$$m \neq 1, 2, \dots, N$$

In this case, when the host's consumption is above 80%, we decide to move or provision the X_m VM to alleviate the strain on the host and maintain optimal performance.

For host ($Rs_consumption$) $\leq 80\%$

$$VM\ to\ be\ moved\ or\ provisioned = X_m \quad (5)$$

Conversely, when the host's resource consumption is below 80%, no action is taken as the resource utilization is within acceptable limits. Identified VMs that are predicted to

cause downtime are removed from host and provisioned to underutilized Esxi host. This proactive approach helps optimize resource utilization and mitigate potential disruptions to services.

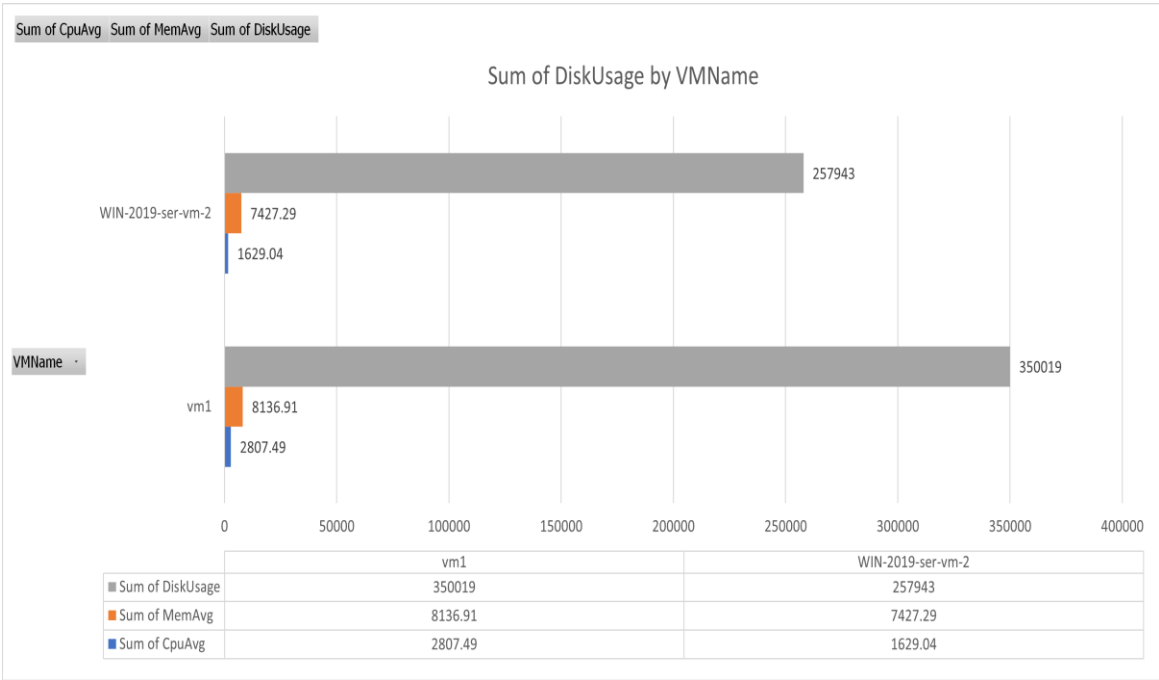


Figure 4: Sum of Resource Usage

Chapter 5 Results and Analysis

The graphical representation in Figure 5 illustrates the resource consumption pattern of host1 before the migration of virtual machines. Specifically, it highlights the CPU utilization trend, which consistently exceeds the critical threshold of 80%. This persistent elevation in CPU utilization signals a significant risk of server downtime, as it indicates that the host is operating at near maximum capacity. Such sustained high utilization levels can lead to resource contention, performance degradation, and potential system failures. Therefore, the depicted graph underscores the urgent need for proactive measures to address the overutilization of resources and mitigate the looming risk of server downtime.

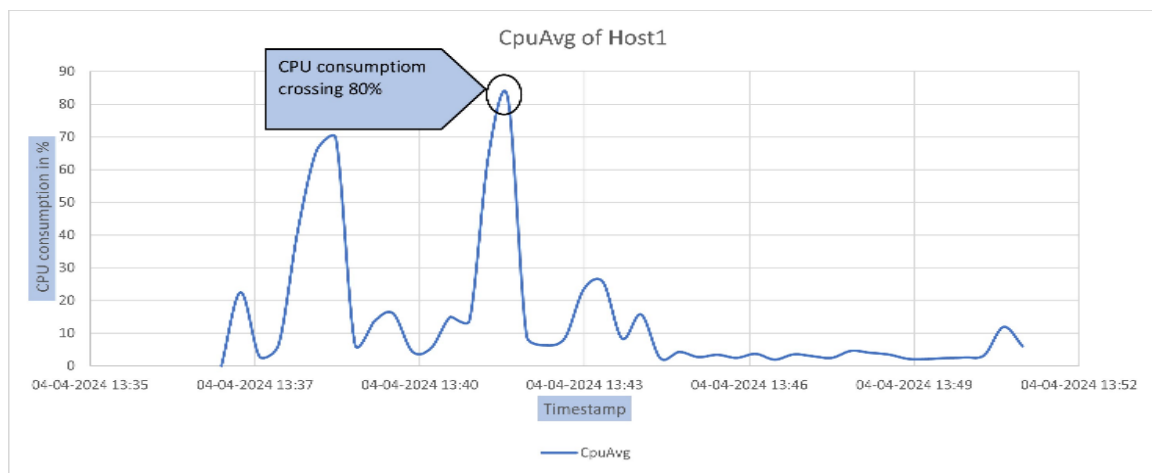


Figure 5: Host 1 consumption before VM migration

Upon studying the subsequent graph depicting the aggregated resource utilization, a clear observation emerges: the specific virtual machine (VM) responsible for surpassing the critical 80% threshold is identified as vm1. This finding underscores the urgency of addressing the overutilization of CPU resources associated with vm1 to prevent further escalation beyond the critical threshold. To mitigate the CPU usage and avert potential risks of server downtime, the imperative task at hand is to execute the migration of vm1 to host 2. By relocating vm1 to host 2, the burden on host 1's CPU resources can be alleviated, thereby restoring balance to the resource distribution and reducing the likelihood of CPU usage surpassing the critical threshold. This proactive measure aims to safeguard the stability and performance of the virtualized environment, ensuring uninterrupted operations and optimal resource utilization.

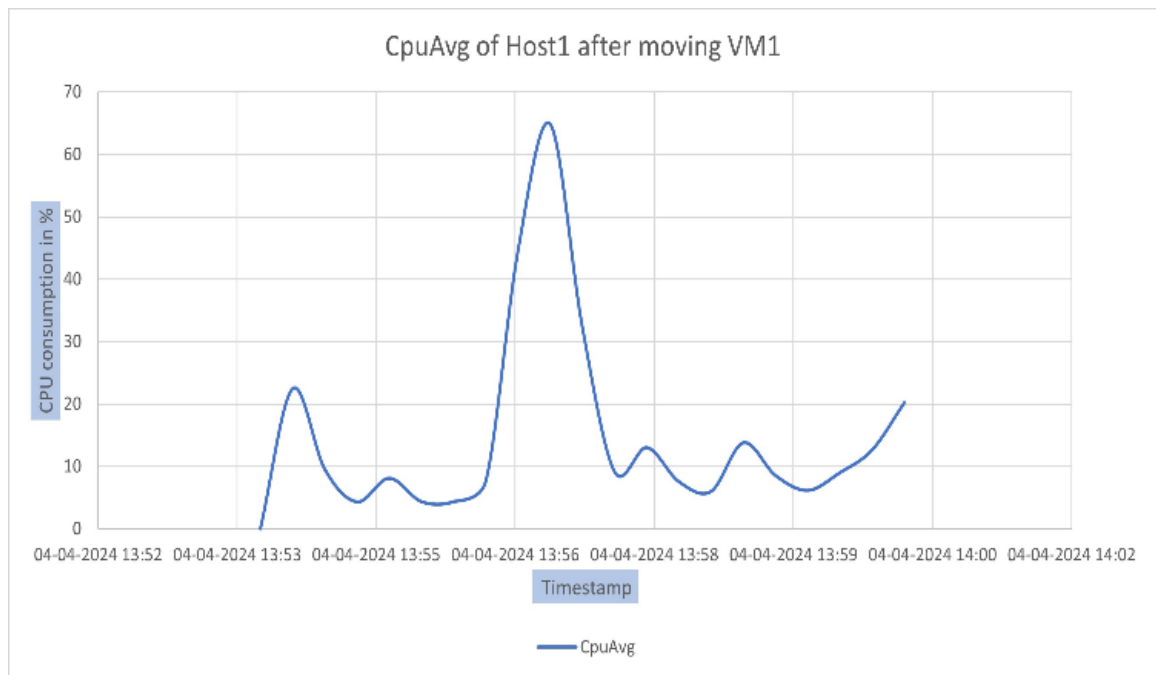


Figure 6: CPU average of Host 1 after moving VM1

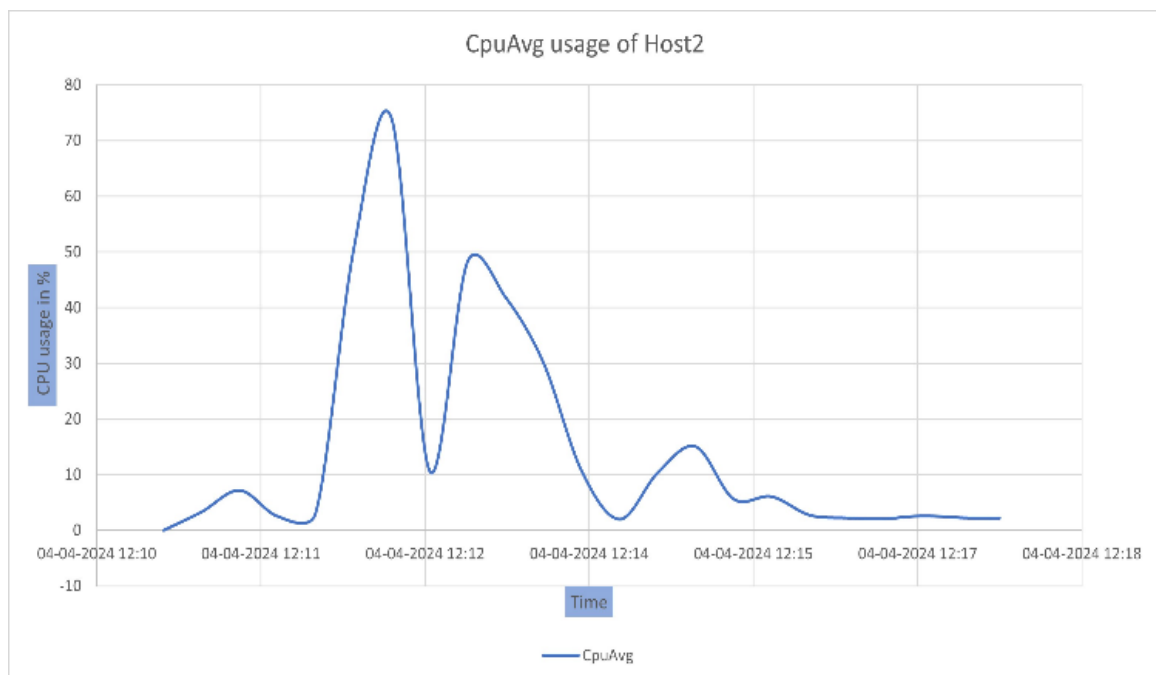


Figure 7: CPU Average of Host 2 after moving VM1

Upon studying Figure 4, it becomes apparent that vm1 is identified as the candidate for migration to host 2 to mitigate CPU usage on host 1 and ensure it remains within the acceptable threshold of 80%. Subsequent analysis, as depicted in Figure 6, reveals a notable reduction in average CPU utilization on host 1 post-migration, confirming that the migration has successfully alleviated the strain on host 1's CPU resources. Furthermore, Figure 7 provides insight into the CPU usage post-migration of vm1 to host 2,

demonstrating that despite the migration, CPU utilization remains below the critical 80% threshold. This comprehensive analysis underscores the effectiveness of migrating vm1 as the optimal solution for managing CPU usage within acceptable parameters. By proactively addressing resource contention issues through strategic VM migration, organizations can ensure the stability, performance, and reliability of their virtualized environments, thereby mitigating the risk of service disruptions and optimizing resource utilization.

Table 3: Accuracy of Models

Models	MAE	MSE	RMSE
Arima	0.78196	6.81183	2.44604
Sarima	14.2945	247.0878	15.71902
SVM	1.631639	5.892534626	2.724808
Prophet	6.177789	6.07989	7.1269
TensorFlow	8.566368	10.21722	11.7898
SG	0.7133	5.87738	2.121332

SG model, with a MAE of 0.7133, MSE of 5.87738, and RMSE of 2.424332, showed competitive performance similar to ARIMA. Despite its effectiveness, the proposed system has certain limitations. One limitation is the reliance on historical data for predictive modeling, which may not always capture sudden or unforeseen changes in workload patterns. Moreover, the accuracy of the predictions may be influenced by factors such as data quality, modeling assumptions, and the complexity of the underlying system architecture. Furthermore, the automated migration or provisioning of VMs may introduce additional overhead and complexity, requiring careful monitoring and management to ensure smooth operation.

The proposed system utilizes real-time data feeds to predict resource usage for virtual machines (VMs) hosted on Esxi hypervisors. By collecting and preprocessing data on CPU, memory, and disk usage, the system predicts future resource needs and detects VMs exceeding 80% resource consumption. The system then automatically migrates or provisions VMs to maintain optimal performance and prevent downtime. Through the analysis, it's evident that the predictive modeling approach based on historical resource usage data can effectively forecast future resource needs. By identifying VMs with high resource consumption, the system can proactively manage host resource allocation, thus ensuring efficient resource utilization and minimizing service disruptions.

The proposed system adopts a forward-thinking strategy for managing resources in virtualized environments. Through the use of predictive modeling, the system successfully tackles the complexities of resource management in Esxi hypervisor deployments, enhancing reliability, scalability, and efficiency. However, ongoing monitoring, evaluation, and refinement are essential to ensure the system's continued effectiveness and adaptability to evolving workload demands and system requirements.

Chapter 6 Conclusion and Future Scope

The experiments conducted in this study focused on evaluating the efficacy of a predictive resource provisioning approach designed to mitigate server downtimes caused by resource overutilization in ESXi servers. By leveraging historical workload data and employing a predictive analytics model to forecast critical resource demand periods, coupled with a dynamic provisioning mechanism, the study aimed to enhance the stability and resilience of virtualized environments. The results demonstrated that this proactive approach significantly reduces the likelihood of server downtimes while also improving overall system reliability and performance. Moreover, the proactive nature of the proposed approach not only prevents server downtimes but also contributes to increased overall system reliability and performance. By anticipating resource usage spikes and VM provisioning to underutilized hosts accordingly, organizations can maintain optimal system operation and minimize the risk of performance degradation or service interruptions. The study suggests that future research endeavors could focus on further refining the predictive analytics model, considering evolving workload patterns and integrating real-time adjustments into the provisioning strategy.

Overall, the objective of this approach is to enhance the stability, performance, and reliability of virtualized environments, ultimately improving the overall service delivery and user experience. Through proper implementation and ongoing monitoring, organizations can effectively manage their virtual infrastructure while meeting the demands of their customers in a dynamic and efficient manner.

Bibliography

- [1]. R. S. Alhumaima, R. K. Ahmed and H. S. Al-Raweshidy, "Maximizing the Energy Efficiency of Virtualized C-RAN via Optimizing the Number of Virtual Machines," in *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 4, pp. 992-1001, Dec. 2018, doi: 10.1109/TGCN.2018.2859407.
- [2]. F. Farahnakian, T. Pahikkala, P. Liljeberg, J. Plosila, N. T. Hieu and H. Tenhunen, "Energy-Aware VM Consolidation in Cloud Data Centers Using Utilization Prediction Model," in *IEEE Transactions on Cloud Computing*, vol. 7, no. 2, pp. 524-536, 1 April-June 2019, doi: 10.1109/TCC.2016.2617374.
- [3]. F. -H. Tseng, X. Wang, L. -D. Chou, H. -C. Chao and V. C. M. Leung, "Dynamic Resource Prediction and Allocation for Cloud Data Center Using the Multiobjective Genetic Algorithm," in *IEEE Systems Journal*, vol. 12, no. 2, pp. 1688-1699, June 2018, doi: 10.1109/JSYST.2017.2722476.
- [4]. L. Abdullah, H. Li, S. Al-Jamali, A. Al-Badwi and C. Ruan, "Predicting Multi-Attribute Host Resource Utilization Using Support Vector Regression Technique," in *IEEE Access*, vol. 8, pp. 66048-66067, 2020, doi: 10.1109/ACCESS.2020.2984056.
- [5]. X. Fu and C. Zhou, "Predicted Affinity Based Virtual Machine Placement in Cloud Computing Environments," in *IEEE Transactions on Cloud Computing*, vol. 8, no. 1, pp. 246-255, 1 Jan.-March 2020, doi: 10.1109/TCC.2017.2737624.
- [6]. B. Xia, T. Li, Q. Zhou, Q. Li and H. Zhang, "An Effective Classification- Based Framework for Predicting Cloud Capacity Demand in Cloud Services," in *IEEE Transactions on Services Computing*, vol. 14, no. 4, pp. 944-956, 1 July-Aug. 2021, doi: 10.1109/TSC.2018.2804916.
- [7]. H. -W. Li, Y. -S. Wu, Y. -Y. Chen, C. -M. Wang and Y. -N. Huang, "Application Execution Time Prediction for Effective CPU Provisioning in Virtualization Environment," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 11, pp. 3074-3088, 1 Nov. 2017, doi: 10.1109/TPDS.2017.2707543.

- [8]. S. B. Melhem, A. Agarwal, N. Goel and M. Zaman, "Markov Prediction Model for Host Load Detection and VM Placement in Live Migration," in IEEE Access, vol. 6, pp. 7190-7205, 2018, doi: 10.1109/ACCESS.2017.2785280.
- [9]. L. Li, J. Dong, D. Zuo and J. Wu, "SLA-Aware and Energy-Efficient VM Consolidation in Cloud Data Centers Using Robust Linear Regression Prediction Model," in IEEE Access, vol. 7, pp. 9490-9500, 2019, doi: 10.1109/ACCESS.2019.2891567.
- [10]. X. Zhang, Z. Huang, C. Wu, Z. Li and F. C. M. Lau, "Dynamic VM Scaling: Provisioning and Pricing through an Online Auction," in IEEE Transactions on Cloud Computing, vol. 9, no. 1, pp. 131-144, 1 Jan.-March 2021, doi: 10.1109/TCC.2018.2840999.
- [11]. Moreno-Vozmediano, R., Montero, R.S., Huedo, E. et al, "Efficient resource provisioning for elastic Cloud services based on machine learning techniques," J CloudComp 8, 5 (2019).
- [12]. Devi, K.L., Valli, S, "Time series-based workload prediction using the statistical hybrid model for the cloud environment," Computing 105, 353– 374 (2023).
- [13]. Souhei Yanase, Graduate Student Member, Fujun He and Eiji Oki, "Approximation Algorithms to Distributed Server Allocation With Preventive Start-Time Optimization Against Server Failure," in IEEE Networking Letters, Vol. 3, No. 4, December 2021.
- [14]. Antra Malhotra, AMR Elsayed, Randolph Torres and Srinivas Venkatraman, "Evaluate Solutions for Achieving High Availability or Near Zero Downtime for Cloud Native Enterprise Applications," in IEEE Access vol. 11, doi: 10.1109/ACCESS.2023.3303430.
- [15]. A. U. Rehman Rui L. Aguiar, and JOÃO Paulo Barracca, "Fault-Tolerance in the Scope of Cloud Computing," in IEEE Access vol. 10, 2022, doi: 10.1109/ACCESS.2022.3182211.
- [16]. Zhinan Cheng, Shujie Han, Patrick P. C. Lee, Xin Li, Jiongzhou Liu and Zhan Li, "An In-Depth Correlative Study Between DRAM Errors and Server Failures in Production Data Centers," in 2022 41st International Symposium on

Reliable Distributed Systems (SRDS) IEEE, DOI:
10.1109/SRDS55811.2022.00032.

- [17]. Song Wu, Yusheng Yi, Jiang Xiao, Hai Jin, and Mao Ye, “A Large-Scale Study of I/O Workload’s Impact on Disk Failure,” in IEEE Access Vol. 6, 2018, doi: 10.1109/ACCESS.2018.2866522.
- [18]. Jargalsaikhan Narantuya, Hannie Zang, and Hyuk Lim, “Service-Aware Cloud-to-Cloud Migration of Multiple Virtual Machines,” in IEEE Access Vol. 6, 2018, doi: 10.1109/ACCESS.2018.2882651.
- [19]. Congfeng Jiang, Guangjie Han, Jiangbin Lin, Gangyong Jia, Weisong Shi, and Jian Wani, “Characteristics of Co-Allocated Online Services and Batch Jobs in Internet Data Centers: A Case Study From Alibaba Cloud,” in IEEE Access Vol. 7, 2019, doi: 10.1109/ACCESS.2019.2897898.
- [20]. Jing Li, Rebecca J. Stones, Gang Wang, Zhongwei Li, Xiaoguang Liu, and Jianli Ding, “New Metrics for Disk Failure Prediction That Go Beyond Prediction Accuracy,” in IEEE Access Vol. 6, 2018, doi: 10.1109/ACCESS.2018.2884004.

Appendix-A

```
1 $esxName = '192.168.204.2'
2 $start = (Get-Date).AddDays(-7)
3 $stat = 'cpu.usage.average','mem.usage.average','disk.usage.average'
4 $esx = Get-VMHost -Name $esxName
5 $report = Get-Stat -Entity $esx -Start $start -Stat $stat -IntervalMins 120 |
    Group-Object -Property Timestamp | ForEach-Object {
        $cpuAvg = $_.Group | Where-Object { $_.MetricId -eq 'cpu.usage.average' } |
        Select-Object -ExpandProperty Value -First 1
        $memAvg = $_.Group | Where-Object { $_.MetricId -eq 'mem.usage.average' } | Select-
        Object -ExpandProperty Value -First 1
        $diskUsage = $_.Group | Where-Object { $_.MetricId -eq 'disk.usage.average'
        } | Select-Object -ExpandProperty Value -First 1

        [PSCustomObject]
        @{
            HostName =
            $esxName
            Date = $_.Name
            CpuAvg = $cpuAvg
            MemAvg = $memAvg
            DiskUsage = $diskUsage
        }
    }

20 $report | Select-Object HostName,Date,CpuAvg,MemAvg,DiskUsage |
21 Export-Csv -Path 'C:\Users\santh\OneDrive\Desktop\workload
    traces\host1\vm1AM4-04-2024.csv' -NoTypeInfoInformation -UseCulture
```

Appendix-B

```
1 # Define ESXi host IP or hostname
2 $esxName = '192.168.204.2'
3
4 # Define time range for statistics collection
5 $start = (Get-Date).AddDays(-7)
6
7 # Define performance metrics to collect
8 $stat = 'cpu.usage.average','mem.usage.average','disk.usage.average'
9
10 # Get the ESXi host
11 $esx = Get-VMHost -Name $esxName
12
13 # Initialize an empty array to store VM metrics
14 $vmMetrics = @()
15
16 # Get all VMs registered on the ESXi host
17 $vms = Get-VMHost $esx | Get-VM
18
19 # Iterate through each VM on the host
20 foreach ($vm in $vms) {
21     # Get performance statistics for the VM
22     $vmStats = Get-Stat -Entity $vm -Start $start -Stat $stat -IntervalMins 120 |
23     Group-Object -Property Timestamp | ForEach-Object {
24         $cpuAvg = $_.Group | Where-Object { $_.MetricId -eq 'cpu.usage.average' } | Select-
25         Object -ExpandProperty Value -First 1
26         $memAvg = $_.Group | Where-Object { $_.MetricId -eq 'mem.usage.average'
27         } | Select-Object -ExpandProperty Value -First 1
28         $diskUsage = $_.Group | Where-Object { $_.MetricId -eq
29         'disk.usage.average' } | Select-Object -ExpandProperty Value -First 1
30
31         [PSCustomObject]@{ VMName
32         = $vm.Name
33         Date = $_.Name
34         CpuAvg = $cpuAvg
35         MemAvg = $memAvg
36         DiskUsage = $diskUsage
37         }
38     }
39     # Add VM metrics to the array $vmMetrics
40     += $vmStats
41
42 # Export VM metrics to CSV
43 $vmMetrics | Select-Object VMName,Date,CpuAvg,MemAvg,DiskUsage |
44     Export-Csv -Path 'C:\Users\santh\OneDrive\Desktop\workload traces\host1_VMs\u21-03-
45     2024.csv' -NoTypeInfo -UseCulture 44
```

Publication Details



Sai Likith P

ICDCECE - 2024 Acceptance Notification for Paper ID 1515

ICDCECE-2024 <publicationchair@icdcece.in>

Mon, 29 Apr at 7:09 PM

To: <psailikith12@gmail.com>, <santhosh12345ind@gmail.com>, <sudeeksha24113@gmail.com>, <karunakara.rai@nmit.ac.in>, <Rajani.n@nmit.ac.in>

Cc: <drparamesh81@gmail.com>, <sunil.s@nmit.ac.in>, <thimmaraja.yg@nmit.ac.in>

Dear Author(s),

Congratulations!

We are pleased to inform you that your paper titled as "Development of a Predictive VM Provisioning in a Cloud Environment" has been Accepted for the IEEE 3rd International Conference on Distributed Computing and

Electrical Circuits And Electronics (ICDCECE-2024) in association with IEEE Bangalore Section organized by Ballari Institute of Technology and Management, Ballari, Karnataka, India. The conference proceedings will be submitted to the IEEE Xplore® digital library. kindly mail us your final camera ready paper through reply this mail with the same paper ID within 30th April 2024.

NOTE: During Camera Ready Submission, please address the reviewer comments mentioned below.

Reviewer Comments:

Format the paper strictly according to the conference template available in the conference website.

Do not modify Layout margin & size of the template.

Result finding should be included in the abstract section.

Thanks for your understanding and cooperation.

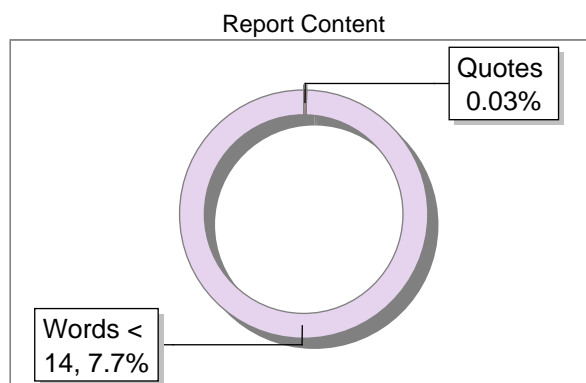
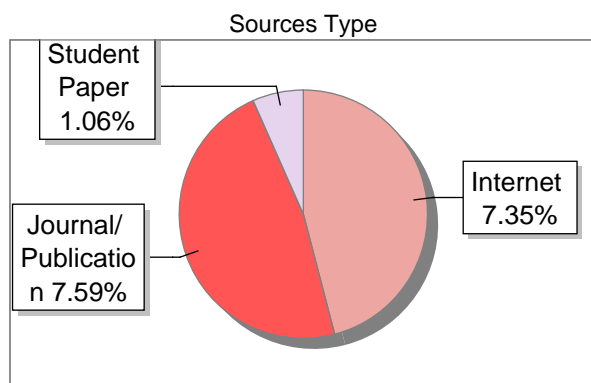
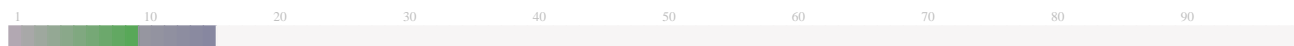
[Quoted text hidden]

Submission Information

Author Name	1nt20ec133.santhosh@nmit.ac.in
Title	Submit/Check your document for plagiarism
Paper/Submission ID	1786574
Submitted by	hod-library@nmit.ac.in
Submission Date	2024-05-11 17:53:02
Total Pages, Total Words	35, 9079
Document type	Assignment

Result Information

Similarity **16 %**



Exclude Information

Quotes	Excluded
References/Bibliography	Not Excluded
Source: Excluded < 14 Words	Not Excluded
Excluded Source	0 %
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File





DrillBit Similarity Report

16

SIMILARITY %

84

MATCHED SOURCES

B

GRADE

A-Satisfactory (0-10%)

B-Upgrade (11-40%)

C-Poor (41-60%)

D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	Markov Prediction Model for Host Load Detection and VM Placement in L, by Melhem, Suhib Bani- 2017	1	Publication
2	REPOSITORY - Submitted to VTU Examination 2 on 2024-02-15 13-16	1	Student Paper
3	shreysharma.com	1	Internet Data
4	Characteristics of Co-allocated Online Services and Batch Jobs in Int, by Jiang, Congfeng Ha- 2019	1	Publication
5	fastercapital.com	1	Internet Data
6	www.mdpi.com	<1	Publication
7	springeropen.com	<1	Internet Data
8	www.igi-global.com	<1	Internet Data
9	www.dx.doi.org	<1	Publication
10	Predicted Affinity Based Virtual Machine Placement in Clby Xiong Fu - ieeeexplore.org	<1	Publication
11	springeropen.com	<1	Publication
12	www.igi-global.com	<1	Internet Data
13	journalofcloudcomputing.springeropen.com	<1	Internet Data

14	www.dx.doi.org	<1	Publication
15	www.linkedin.com	<1	Internet Data
16	fastercapital.com	<1	Internet Data
17	fastercapital.com	<1	Internet Data
18	thesai.org	<1	Publication
19	distrowatch.com	<1	Internet Data
20	2017 Index IEEE Transactions on Parallel and Distributed Systems Vol 28 by -2018	<1	Publication
21	fastercapital.com	<1	Internet Data
22	nature.com	<1	Internet Data
23	dochero.tips	<1	Internet Data
24	eprints.whiterose.ac.uk	<1	Publication
25	fastercapital.com	<1	Internet Data
26	moam.info	<1	Internet Data
27	Thesis Submitted to Shodhganga Repository	<1	Publication
28	www.scribd.com	<1	Internet Data
29	www.simplilearn.com	<1	Internet Data
30	A dynamic VM consolidation approach based on load balancing using Pearson correl by Mapetu-2020	<1	Publication
31	e-tarjome.com	<1	Publication
32	www.dx.doi.org	<1	Publication

33	www.leewayhertz.com	<1	Internet Data
34	climateerinvest.blogspot.com	<1	Internet Data
35	Incremental prediction model of disk failures based on the density metric of edg by Gao-2019	<1	Publication
36	moam.info	<1	Internet Data
37	etasr.com	<1	Publication
38	moam.info	<1	Internet Data
39	springeropen.com	<1	Publication
40	www.freepatentsonline.com	<1	Internet Data
41	www.linkedin.com	<1	Internet Data
42	cloudbus.org	<1	Publication
43	Dynamic Control of Data Streaming and Processing in a Virtualized Envi by J-2012	<1	Publication
44	moam.info	<1	Internet Data
45	moam.info	<1	Internet Data
46	www.dx.doi.org	<1	Publication
47	IEEE 2013 IEEE Conference on Computer Communications Workshops (INFO by	<1	Publication
48	downloads.hindawi.com	<1	Publication
49	fastercapital.com	<1	Internet Data
50	www.crisis-control.com	<1	Internet Data

51	www.dx.doi.org	<1	Publication
52	BurScale Using Burstable Instances for Cost-Effective Autoscaling in the Public by Baarzi-2019	<1	Publication
53	docplayer.net	<1	Internet Data
54	docplayer.net	<1	Internet Data
55	eeslindia.org	<1	Publication
56	Energy Efficient Virtual Machine Placement in Data Center- www.ijcaonline.org	<1	Publication
57	IEEE 2019 Design, Automation Test in Europe Conference Exhibitio	<1	Publication
58	livrosdeamor.com.br	<1	Internet Data
59	online.hbs.edu	<1	Internet Data
60	utilitiesone.com	<1	Internet Data
61	utilitiesone.com	<1	Internet Data
62	www.dx.doi.org	<1	Publication
63	www.openaccessojs.com	<1	Publication
64	aprd.in	<1	Internet Data
65	austlii.edu.au	<1	Internet Data
66	fastercapital.com	<1	Internet Data
67	ijrpr.com	<1	Publication
68	Improving magnetotelluric data-processing methods by Epishkin-2016	<1	Publication
69	Intervenors with Interests and Power by Michae-1997	<1	Publication

70	jessup.edu	<1	Internet Data
71	library.sadjad.ac.ir	<1	Publication
72	moam.info	<1	Internet Data
73	Performance comparison of bubble point pressure from oil PVT data Several neuro by Ghorbani-2019	<1	Publication
74	Procurement and its Role in Corporate Strategy An Overview of the Wine and Spir by Sutton-1989	<1	Publication
75	REPOSITORY - Submitted to Bannari Amman Institute of Technology on 2023-02-10 10-43	<1	Student Paper
76	Resource Management in a Containerized Cloud Status and Challenges by Maenhaut-2019	<1	Publication
77	Semantics in Mobile Sensing by Yan-2014	<1	Publication
78	springeropen.com	<1	Publication
79	Submitted to Visvesvaraya Technological University, Belagavi	<1	Student Paper
80	Thesis Submitted to Shodhganga, shodhganga.inflibnet.ac.in	<1	Publication
81	www.blog-qhse.com	<1	Internet Data
82	www.marshall.usc.edu	<1	Internet Data
83	www.ncbi.nlm.nih.gov	<1	Internet Data
84	www.southindianbank.com	<1	Publication

Development of a Predictive VM Provisioning in a Cloud Environment

Sai Likith P

*Electronics and Communication
Engineering*

Nitte Meenakshi Institute of Technology
Bengaluru, India
psailikith12@gmail.com

Santhosh P

*Electronics and Communication
Engineering*

Nitte Meenakshi Institute of Technology
Bengaluru, India
santhosh12345ind@gmail.com

Sudeeksha K

*Electronics and Communication
Engineering*

Nitte Meenakshi Institute of Technology
Bengaluru, India
sudeeksha24113@gmail.com

Karunakara Rai B

Electronics and Communication Engineering
Nitte Meenakshi Institute of Technology

Bengaluru, India
karunakara.raai@nmit.ac.in

Rajani N

Electronics and Communication Engineering
Nitte Meenakshi Institute of Technology

Bengaluru, India
Rajani.n@nmit.ac.in

Abstract— In contemporary virtualized data center environments, efficient resource allocation in virtualized data center environments is crucial for preventing server downtimes. This paper introduces a method to mitigate resource exhaustion on ESXi servers using predictive analytics. Leveraging historical workload data, the proposed approach forecasts potential resource saturation events by analyzing past usage patterns. A proactive provisioning mechanism is outlined to deploy virtual machines (VMs) onto new ESXi servers during predicted high-load intervals. This preemptive strategy strategically distributes workloads, averting server downtime and ensuring optimal resource utilization. Experimental results validate the effectiveness of the proposed method in improving the dependability and security of ESXi servers in virtualized environments and outperforms better than other state-of-the-arts. The proposed method is implemented using ESXi host client and the result shows effective performance and improved stability of the overall system.

Keywords— VM, Cloud Computing, VM provisioning, workload pattern, host resource prediction, VM allocation.

I. INTRODUCTION

In recent years, the widespread adoption of cloud computing has transfigured how businesses manage their computing resources, offering on-demand access to scalable infrastructure. However, as the utilization of virtual machines (VMs) proliferates, effectively managing and allocating resources within a cloud environment becomes increasingly challenging. To address this, innovative technologies like virtualization have emerged, enabling businesses to optimize resource usage and enhance performance. Datacenter virtualization, a cornerstone of modern cloud infrastructure, involves the creation, deployment, and management of virtualized data centers. By virtualizing physical servers and leveraging advanced computing technologies, organizations can efficiently manage storage, networking, and other critical infrastructure components. Researchers therefore face difficulties when co-hosting many types of virtual machines (VMs) on a small number of servers due to resource conflict between co-hosted applications, which causes servers to be overutilized, which in turn degrades application performance. Furthermore, a lot of cloud services, such as interactive apps, have regularly shifting workload requirements, which lead to

dynamic resource demand. If dynamic server consolidation is employed, this might cause SLA violations and performance degradation. This approach leverages historical data to forecast future resource requirements and allocate resources pre-emptively. By analysing past consumption trends, predictive VM provisioning assures that sufficient resources are available to meet anticipated demand, optimizing resource utilization and application performance. In sectors like banking, where server availability is paramount during peak periods of activity, such as high-volume transactions or system updates, cloud environments may experience sudden spikes in resource demand and there is high chances of server-failures. To address the aforementioned concerns, the hypervisor identifies suitable virtual machines or sets of virtual machines and relocates them from heavily utilized servers to underutilized ones, aiming to enhance overall performance. Predictive VM provisioning anticipates these spikes, checks and assures that adequate resources are allocated in advance, preventing service disruptions and maintaining uninterrupted operations.

To achieve this, effective VM allocation methods are essential. By employing statistical data based on observed job patterns and system metrics, organizations can anticipate future resource requirements and allocate VMs accordingly. This proactive approach minimizes the risk of Service Level Agreement (SLA) violations and optimizes resource utilization efficiency. Additionally, VM selection techniques play a vital role in resource management within cloud environments. By strategically choosing VMs for migration based on workload distribution and resource availability, organizations can rebalance the load across the system's active nodes without causing downtime or service interruptions. In [12] presents a survey of predictive VM provisioning strategies, encompassing existing prediction algorithms that aid and support in selecting virtual machines (VMs) to be supplied to other servers. In this study, we give a thorough assessment on cutting-edge predictive virtual machine provisioning strategies and overcome the shortcomings of previous surveys.

The primary advancements outlined in this paper are as follows:

1. A thorough analysis of the literature on cutting-edge predictive virtual machine provisioning strategies, outlining their advantages, disadvantages, and important research gaps.
2. Discussions on different security risks, server failures and current mitigation techniques.
3. Identification of specific gaps and research challenges to reduce the downtime of servers.

The rest of this paper is organized as follows. Section II summarizes related works. Section III presents a proposed method, which provides a solution to prevent down time of servers. Section IV includes software requirements, host workload analysis, comparative analysis, key findings under results and discussion. We present our performance evaluation in Section IV, while the conclusion and future work are presented in Section V.

II. RELATED WORK

In their seminal work, the authors of paper [1] addresses the challenge of optimizing the allocation of virtual machines (VMs) in cloud servers to enhance energy efficiency in cloud radio access networks. Employing Monte Carlo-based evolutionary algorithms such as particle swarm optimization, and genetic algorithms, the study aims to find the optimal number of VMs for energy-efficient operation. The model shows the impact of adding a greater number of VMs on server performance, particularly in processing resource blocks (RBs) per VM. Advantages of the model include its support for energy efficiency and improvement in quality-of-service metrics such as minimum user equipment data rate, allocated Resource blocks, and latency due to virtualization. However, shortcomings are noted in VM placement and power allocations. Additionally, the paper proposes a power model to assess power usage in virtualized server components, aiming to simplify evaluation processes. As elucidated by the authors in paper [2] they explored VM consolidation as a means to enhance resource utilization, utilizing a regression model to forecast future CPU and memory usage. Through analysis of real workload traces from Google cluster and Planet Lab, we implement the UP-VMC approach, aiming to minimize unnecessary VM migrations. The results demonstrate that VM consolidation can significantly reduce energy consumption, up to 71.6% compared to alternative methods, by consolidating VMs onto the most heavily loaded physical machines. However, challenges remain in terms of scalability and optimizing VM placement, particularly regarding network resource utilization and traffic management. Additionally, the authors introduce the UP-VMC strategy to optimize the energy efficiency of Cloud Radio Access Networks (C-RAN) by determining the optimal number of virtual machines.

In paper [3], a Multi objective genetic algorithm (GA) is employed to predictively estimate how resources are used and energy is consumed in cloud data centers in real time, considering CPU and memory utilization of both virtual machines (VMs) and physical machines (PMs). The study results in the development of an algorithm for placing virtual machines designed to enhance overall resource utilization and decrease energy consumption within the data center by leveraging prediction outcomes from genetic algorithms. The approach enhances PM utilization while decreasing energy consumption by minimizing the number of active PMs. However, limitations exist in the accuracy of the approach in

realistic cloud environments, as it has not been tested with real-world data center traces such as those from Google. Additionally, the authors propose a novel GA-based prediction strategy to enhance forecast precision in cloud data centers and suggest a VM placement strategy based on GA predictions to optimize resource utilization and energy consumption. Expounding on the findings presented in paper [4] addresses the challenge of forecasting future cloud resource utilization by employing the Support Vector Regression Technique (SVRT) with a Radial Basis Function (RBF) kernel function. This method aims to accurately predict multi-attribute host resource utilization, particularly suited for nonlinear workload patterns. The Sequential Minimal Optimization Algorithm (SMOA) is applied for training and regression estimation to enhance prediction accuracy. By utilizing SVRT with the RBF kernel function, the research suggests a method less susceptible to fluctuations in resource utilization compared to other forecasting techniques. Through experimentation with eight datasets, the research investigates variations in workload demand and resource utilization within a cloud setting, emphasizing the opportunity to minimize resource waste and enhance resource efficiency in cloud data centers.

Embarking on their analysis in paper [5] the author develops a NICBLE-based system prototype for Xen virtualization to assess the impact of hypothetical changes in VM setups on central processor quantities. NICBLE predicts application execution based on simulated calculations. The Priority-Aware VM Allocation (PAVA) technique is proposed for VM allocation, leveraging network topology information to assign critical applications to closely connected hosts. The model further explores relationships between VMs based on ARIMA-predicted resource requirements, introducing an affinity model to evaluate resource utilization volatility when VMs are placed on the same host. The resultant algorithm for virtual machine placement, rooted in predicting affinity, consolidates VMs with strong connections onto single physical machines, maximizing resource usage while staying within the limits of physical machine capacities. Extensive simulation experiments validate the algorithm's effectiveness in reducing energy consumption, VM migrations, and SLA violations based on Planet lab and Google workload traces. However, further adjustments are needed to improve the accuracy of the prediction model. The authors in [6] have released the ground-breaking Forecast Empathy Based Virtual Machine Positioning algorithm. By using this technique, a single physical machine is created from the high-affinity virtual machines. Estimating a request within a range is pointless because the projection will continuously stray from the actual requests and serve as a guide.

Within the context of paper [7], the researchers utilize the Support Vector Relapse Strategy (SVRT), a managed factual learning technique, to foresee how the multi-property have asset will be utilized from here on out. Utilizing cloud resources to manage a non-linear workload is a perfect application for this strategy. We pick Spiral Premise Capability as the bit capability of SVRT and utilize Consecutive Negligible Streamlining Calculation (SMOA) for the preparation and relapse assessment of the expectation strategy to build the expectation precision of SVRT. Addressing the subject matter in paper [8] they proposed the use of a Markov prediction model to forecast the future load state of hosts in cloud environments. We introduce a host load detection algorithm to identify over or under-utilized hosts,

aiming to prevent immediate VM migration. A VM algorithm for placement is then employed to select candidate hosts for migrated VMs, evaluated using cloud Sim simulation. Our approach relies on dynamic utilization thresholds to enhance the VM placement process and predict results to avoid host overload shortly after migration. Additionally, the Median Absolute Deviation Markov Chain Hot Detection technique (MadMCHD) to improve host detection performance during live migration by determining future overutilized and underutilized states is used. By incorporating the RobustSLR prediction algorithm into the PABFD algorithm, authors altered both the existing VM placement method and the new VM algorithm for placement [9].

In the situation of a 140% system load, the suggested method can handle up to 12% more user requests and generate up to 8% more system rewards [10]. A comprehensive analysis by [11] introduces a predictive auto-scaling mechanism for elastic cloud services, integrating time series forecasting using machine learning methods with queuing theory to enhance service response times and reduce unnecessary resource allocation. This approach utilizes SVM regression to anticipate the workload of web servers by analyzing past data. Results indicate that SVM-based forecasting models perform better than basic models in terms of reducing over-provisioned resources. However, some SVM-based models exhibit poorer performance regarding SLA violations and unserved requests. Advantages include better prediction accuracy compared to simpler methods, as demonstrated by MAE and RMSE error measurements. Nonetheless, challenges remain in adapting predicting and performance models to the diverse components and functionalities of other cluster architectures such as MapReduce, HDFS, and Spark components. According to the research presented in [12] the authors employed a hybrid ARIMA-ANN model to predict future CPU and memory utilization in cloud environments, leveraging both linear and nonlinear components in the data. Analysis and experiments are conducted using workload traces from Google trace and Bit Brain compute clusters, focusing on CPU and memory usage. While the ARIMA model detects linear patterns, the ANN is effective in predicting nonlinear patterns by leveraging residuals from the ARIMA model. Advantages include improved prediction accuracy for nonlinear patterns with ANN, although reliance on training data may affect performance. To address this, the proposed hybrid model combines the strengths of both ARIMA and ANN. However, challenges remain in dynamically adjusting the sliding window size and assigning weights to recent historical data points to further enhance prediction accuracy.

The endeavor to optimize server allocation in the face of potential failures has prompted the development of advanced algorithms aimed at minimizing delays and maximizing system efficiency. In this context of [13], recent research has proposed polynomial-time approximation algorithms leveraging Particle Swarm Optimization (PSO) to tackle the distributed server assignment problem in the event of single server failures. The NP-completeness of the problem has been established, underscoring the complexity of the task at hand. The proposed algorithms demonstrate promising approximation performances, offering significant speed enhancements compared to traditional Integer Linear Programming (ILP) approaches. Numerical analyses indicate substantial efficiency gains, with the proposed algorithms outperforming ILP by factors ranging from 3.0×10^3 to 1.9×10^7 while slightly increasing the largest maximum

delay by a factor of 1.033 on average. These findings shed light on novel strategies for efficiently addressing server allocation challenges in distributed systems, paving the way for enhanced reliability and performance in cloud computing environments. The scholarly contribution of [14] the authors delve into the intricacies of ensuring high redundancy and almost zero downtime for enterprise applications within cloud environments. They highlight the significant challenge of establishing surplus and robust architectures, both at the application and database layers. Emphasizing the importance of automated failover mechanisms, particularly for applications expected to operate around the clock, the authors propose a cloud-native template architecture for enhancing availability. Through their evaluation of automatic database failover techniques, they aim to minimize disruptions during planned maintenance activities and outages. While acknowledging the variability in achieving uninterrupted service based on factors like application size and data volume, the authors provide a foundational framework for building resilient applications. They stress the need for customization of the recommended architecture and failover strategies to align with each application's unique requirements, considering factors such as cost and specific business objectives.

An in-depth exploration of cloud computing, beginning with a comprehensive background to setup a foundational grasp of the subject. It delves into fault-tolerance components and system-level metrics, emphasizing their significance and applications within cloud computing environments. The authors meticulously examine both proactive and reactive approaches to fault-tolerance in cloud computing, showcasing the state-of-the-art techniques and frameworks. By organizing and discussing current research endeavors in fault-tolerance architectures, the paper offers valuable insights into the evolving landscape of cloud computing resilience. It concludes by outlining key future research directions, underscoring the importance of continued development in fault-tolerance strategies specific to cloud computing [15]. The authors in [16], presents a comprehensive data-centric analysis correlating DRAM errors with server outages, aiming to predict server outages based on DRAM error patterns. Leveraging an extensive eight-month dataset from Alibaba's production data centers, comprising over three million memory modules, we identify that correctable DRAM errors often precede server failures, highlighting the importance of regular and frequent server failure prediction intervals for accurate forecasting. Our investigation also explores various factors influencing instances of server malfunctions, encompassing component breakdowns within the memory subsystem, variations in DRAM setups, and categories of fixable DRAM inconsistencies. Notably, we extend prior work by considering multiple types of server malfunctions in the prediction of server failures, achieving significant reductions in server downtime. Our findings reveal that UE-driven failures pose challenges in prediction due to the limited occurrence of correctable errors before these failures, whereas CE-driven and miscellaneous failures exhibit higher predictability. By employing all feature groups, we enhance prediction accuracy, with tree-based prediction models demonstrating superior performance, underscoring the importance of short prediction intervals for timely failure prediction.

Delving into the intricate the correlation between input/output (I/O) workload characteristics and disk

reliability, aiming to uncover factors influencing disk lifespan and identify detrimental I/O workloads. By proposing an innovative measure, AISR (Average I/O Service Rate), we shed light on "dangerous" I/O workloads posing significant risks to disk health. Our research represents a pioneering effort in comprehensively analyzing how input/output (I/O) workload influences disk dependability, providing valuable perspectives to improve I/O scheduling strategies within data centers. While our work marks an initial exploration in this domain, we anticipate that our findings will stimulate further research in the community, prompting a reevaluation of disk I/O workload assignments. Future endeavors will focus on incorporating additional workload metrics to extract actionable insights and implement them effectively in data center environments [17]. As documented in [18] the research introduces a strategy for determining the sequence of migrations, utilizing network traffic data between virtual machines to reduce downtime by 50% during cloud-to-cloud migration. Employing Prim's algorithm, they identify the Minimum Spanning Tree (MST) for a weighted undirected graph, facilitating efficient migration. The controller nodes, equipped with Intel Xeon W35653.25GHz CPU, 48GB RAM, and 1TB SSD, contribute to achieving shorter service downtime during migration. The strategy accounts for migrating multiple VMs, addressing communication dependencies among them. While the approach yields shorter service downtime, it may sometimes necessitate higher security measures, potentially leading to increased service downtime.

Through their meticulous study in [19], unveils an analysis among the services that are online and batch jobs, which are collocated within a production cluster in Alibaba Cloud. By clustering servers based on CPU and memory utilization correlations, it identifies opportunities for job co-allocation and resource estimation. Through examination of mean time between failures (MTBF) and completion times, it sheds light on failure distribution and workload assignment disparities. The insights gleaned from this analysis can empower data center operators to optimize resource utilization and enhance failure recovery mechanisms, ultimately fostering a deeper understanding of workload characteristics and operational efficiencies within cloud environments. In accordance with [20] the author evaluates the effectiveness of various disk-failure prediction methods using metrics such as timeliness and convergence. By comparing classification tree (CT), a neural network with recurrent connections and a regression tree model boosted using gradient descent models, the research highlights the nuanced performance differences across prediction experiments. While RNN exhibits superior accuracy, CT and GBRT models demonstrate advantages in resource-dependent migration rates. The findings underscore the importance of considering prediction accuracy in conjunction with practical outcomes, prompting the introduction of an enhanced GBRT model (GBRT+). Moving forward, the exploration of urgency-weighted evaluations and adjustments to machine learning processes offer promising avenues for refining disk-failure prediction models.

III. PROPOSED METHOD

The host operating system (OS) as shown in Figure 1 *H2H (Host to Host) VM Provisioning* is the foundation of the Type 2 hypervisor architecture. This is the main operating system that is directly installed on the computer's Physical hardware. It might be Linux, macOS, or Windows.

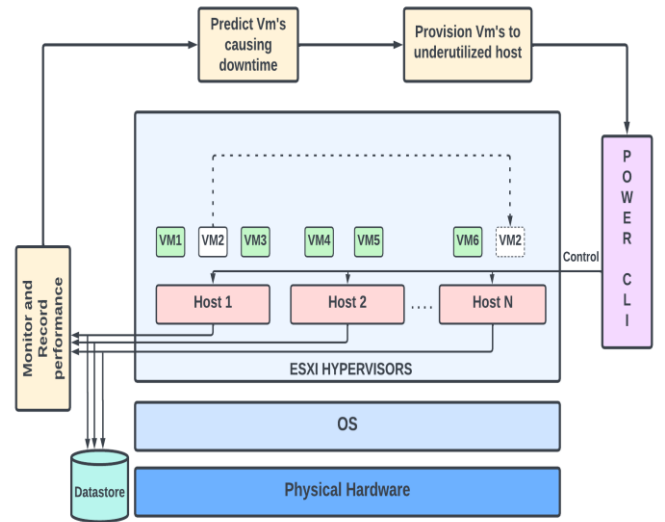


Figure 1: *H2H* VM Provisioning

The Type 2 hypervisor operates as an application on top of the host operating system. Multiple guest operating systems can operate concurrently on the same physical hardware by the virtualization layer created by hypervisor software like VMware Workstation Player. One or more guest operating systems may be installed as virtual machines inside the Type 2 hypervisor environment. Both the host OS and these guest OS instances run independently of one another. Different operating systems, including different Windows versions, Linux distributions, and (in certain situations) macOS, can be installed by users.

In order to allocate parts of the hardware resources such as CPU, memory, storage, and network interfaces to each virtual machine as required, the hypervisor manages these resources. As a result, numerous virtual machines (VMs) can share the physical resources without interfering with one another. Datastores are used by guest operating systems to store virtual hard drives (VHDs) or virtual machine disk images. They are commonly represented as disk image files. These files may be kept on local or network storage or on any other storage device that the host operating system may access. Users communicate with the Type 2 hypervisor via a management interface supplied by the hypervisor software. This interface allows users to set up, configure, start, stop, and delete virtual machines, as well as manage their settings and resources.

The proposed system is based on a real-time feed as inputs and follows the operational process shown in Figure 2.

- ESXi Hypervisor Setup
- Datastore Creation
- VM Creation and Provisioning
- Template Creation
- Installation of Power CLI
- Data Collection and Prediction
- VM Migration

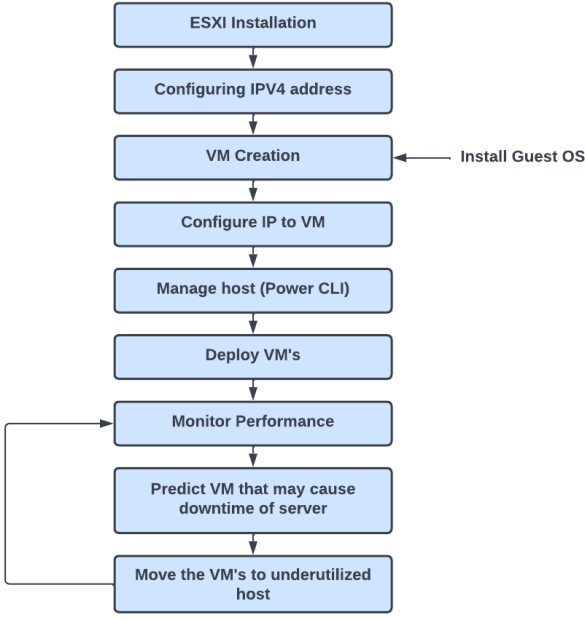


Figure 2: Proposed Solution for VM Provisioning

ESXi hypervisors are installed on physical hardware. Each hypervisor is configured with IP addresses (IPv4), DNS server addresses, and default gateways. Datastores are established to allocate memory to virtual machines (VMs) and houses the disk files of VMs. This allows for efficient management of resources. Let's consider 2 hosts (ESXi server version 8.0.2) host 1 as remoteServer.localdomain configured with static ipv4 address "192.168.204.2" and host2 as host2.localdomain with static address "192.168.204.8" each with 4GB of RAM and 4 CPU cores as shown in Table 1 host information.

Table 1: Host Information

Host Name	Num CPU	CPU Usage (Mhz)	CPU Total (Mhz)	Memory Usage (GB)	Memory Total (GB)
192.168.204.2	4	462	9980	1.59	3.999
192.168.204.8	4	156	9980	1.578	3.999

When a request for a new VM arises, resources are allocated based on customer demands. VMs are created and configured with unique IP addresses. ISO image files are used to load guest operating systems onto these VMs. Templates are created to expedite the provisioning process. These templates contain pre-configured settings and configurations, enabling quick deployment of new VMs within minutes. Two VMs vm1 and WIN-2019-ser-2 are created under host1, VM named h2_WIN-2019-ser-2 alone is created in host2. The number of CPU cores, RAM and datastore of these VMs are shown in Table 2 VM information.

Table 2: VM Information

VM Name	VM id	EXSi Host	Data Store	Num CPU	Memory (GB)
vm1	10	192.168.204.2	Storage_datastore_1	2	1
WIN-2019-ser-vm-2	9	192.168.204.2	Storage_datastore_1	2	4
h2_WIN-2019-ser-vm-2	9	192.168.204.8	Storage_datastore_2	2	4

Power CLI is used to automate and manage the Vsphere, Vcenter and VMware host client using power shell by logging in with credentials of the server. The usage of host and VMs are recorded with the use of CLI script for a specific time interval for analyzing high resource usage (Memory, CPU, network, disk Usage) of the VMs in Over-Provisioned Host. Collecting and preprocessing data from diverse sources like CPU usage, memory usage, and disk usage enables the prediction of future resource usage for VMs and the detection of those exceeding 80% resource consumption.

$$obj(y, \hat{y}) = \frac{1}{2} \sum (y_i - \hat{y}_i)^2 + \lambda \sum |w_j|^\gamma \quad (1)$$

where:

obj objective function in SG model minimizes both the training error and a regularization term to prevent overfitting. y_i is the actual value of the i-th data point.

\hat{y}_i is the predicted value of the i-th data point by the model.

λ is the regularization parameter controlling the strength of the penalty term.

w_j is the weight of the j-th feature in the model

γ is a hyperparameter that determines the type of regularization.

Let $T(x_i)$ represent the prediction of the i-th data point by a single tree in the ensemble. The ensemble prediction can be formulated as:

$$\hat{y}_i = \sum f_m(x_i) \quad (2)$$

where:

m iterates over all trees in the ensemble.

$f_m(x_i)$ is the prediction of the m-th tree for the i-th data point.

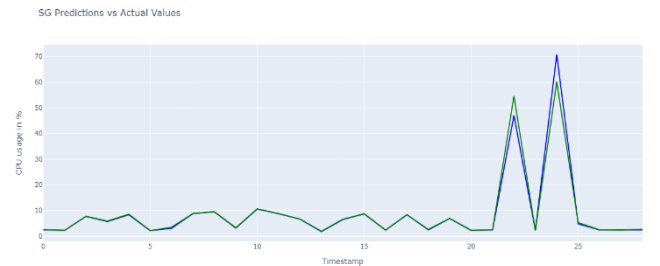


Figure 3: CPU Usage Actual vs Predicted plot

By analyzing the Figure 3, its observed the accuracy of model is better fit for the dataset. The predicted resource usage of these VM's are collected, analyzed and detected the VMs causing downtime of the Server as shown in Figure 4 sum of resource Usage of CPU, Memory and Disk Usage.

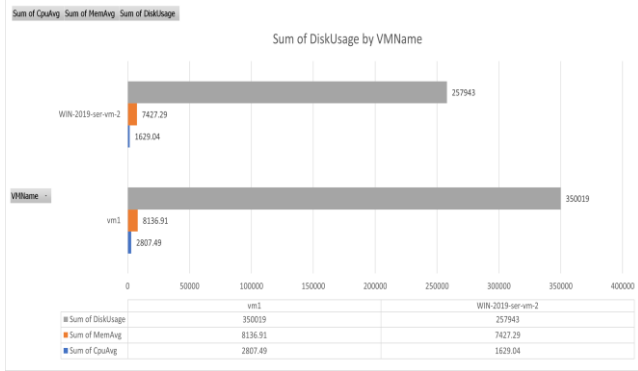


Figure 4: Sum of Resource Usage

When the resource consumption of a host exceeds 80%, it evaluates the total resource usage (R_s_Usage) across all virtual machines within a specified time interval. This total usage (R_s_Usage) is calculated by summing the resource usage (R_s_Usage) of each VM at various time points (t_i) within the interval.

$$\text{Total } R_s_Usage(X) = \sum R_s_Usage(X, t_i) \quad (3)$$

The VM with the highest total resource usage (R_s_Usage) is identified as X_m , indicating that it's consuming the most resources.

$$\text{Total } R_s_Usage(X_m) > \text{Total } R_s_Usage(X_1), \text{Total } R_s_Usage(X_2), \dots, \text{Total } R_s_Usage(X_n) \quad (4)$$

Where $X_1, X_2, X_3, \dots, X_n$ are n different virtual machines. $m \neq 1, 2, \dots, N$

In this case, when the host's consumption is above 80%, we decide to move or provision the X_m VM to alleviate the strain on the host and maintain optimal performance.

$$\text{For host } (R_s_consumption) \leq 80\% \\ \text{VM to be moved or provisioned} = X_m \quad (5)$$

Conversely, when the host's resource consumption is below 80%, no action is taken as the resource utilization is within acceptable limits. Identified VMs that are predicted to cause downtime are removed from host and provisioned to underutilized ESXi host. This proactive approach helps optimize resource utilization and mitigate potential disruptions to services.

IV. RESULTS

A. Software Requirements

The system requirements include a minimum of 12 GB RAM and 256 GB storage space. Essential software components comprise VMware Workstation Player for virtualization, a hypervisor like VMware ESXi for resource management, ISO image files for OS installation, VMware

VCenter Appliance for centralized management, and Power CLI for automation.

B. Host Workload Analysis

The graph in Figure 5 represents that the CPU utilization consistently exceeds the 80% threshold. CPU utilization refers to the percentage of the CPU's processing capacity that is being used at any given time. In this context, surpassing the 80% threshold indicates that host1's CPU is being heavily utilized, potentially reaching its maximum capacity frequently. Such high CPU utilization poses a significant risk to the server's stability and performance. When the CPU is consistently operating at or near its maximum capacity, it leaves little room for handling additional processing tasks or spikes in workload. As a result, there is an increased likelihood of server downtime or performance degradation. To mitigate this risk and ensure the continued smooth operation of the server, it's advisable to take proactive measures such as virtual machine migration. By migrating some of the virtual machines from host1 to other hosts with lower resource utilization, the burden on host1's CPU can be alleviated, reducing the risk of downtime and ensuring optimal performance for all virtual machines hosted on the server.

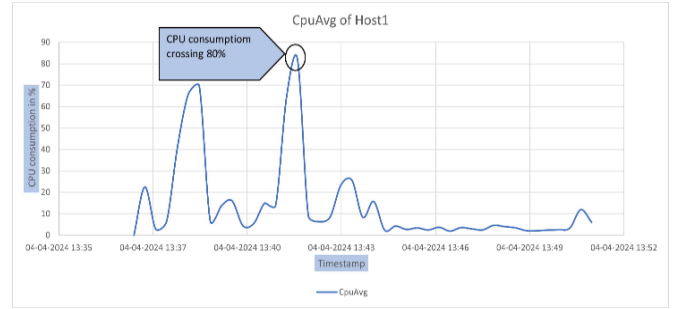


Figure 5: Host 1 consumption before VM migration

Migrating vm1 to host2 involves relocating its workload and associated resources from host1 to host2. By doing so, the CPU usage on host1 will decrease, mitigating the risk of server downtime and ensuring optimal resource allocation across the infrastructure. Executing this migration is essential for maintaining the stability and performance of the server environment. It allows for better distribution of workload and prevents any single virtual machine from monopolizing resources as shown in Figure 6, thereby promoting overall system efficiency and reliability.

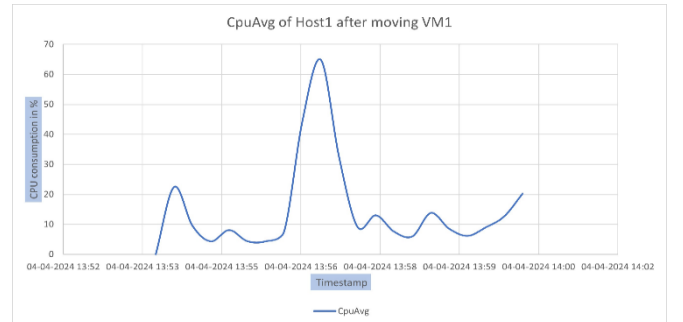


Figure 6: CPU average of Host 1 after moving vm1

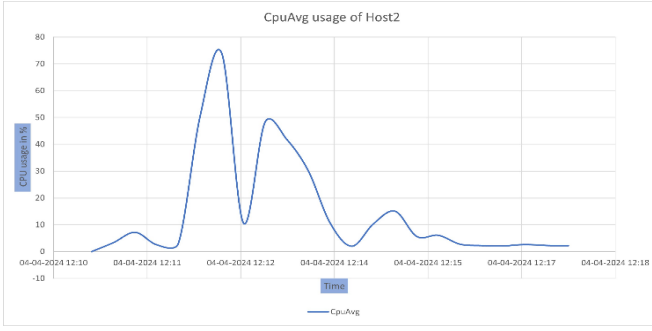


Figure 7: CPU Average of Host 2 after moving vm1

As depicted in Figure 6, the average CPU utilization on host 1 post-migration has indeed decreased to below 80%. Subsequently, Figure 7 illustrates the CPU usage post-migration of VM1 to host 2, revealing that despite the migration, CPU utilization remains below the 80% threshold. Consequently, it is evident that migrating vm1 is the optimal solution to effectively manage CPU usage within acceptable parameters.

C. Discussion

The performance of various forecasting models, including ARIMA, SARIMA, SVM, Prophet, TensorFlow, and SG, was evaluated based on metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) shown in Table 3.

Table 3: Accuracy of models

Models	MAE	MSE	RMSE
Arima	0.78196	6.81183	2.44604
Sarima	14.2945	247.0878	15.71902
SVM	1.631639	5.892534626	2.724808
Prophet	6.177789	6.07989	7.1269
TensorFlow	8.566368	10.21722	11.7898
SG	0.7133	5.87738	2.424332

SG model, with a MAE of 0.7133, MSE of 5.87738, and RMSE of 2.424332, showed competitive performance similar to ARIMA. Despite its effectiveness, the proposed system has certain limitations. One limitation is the reliance on historical data for predictive modeling, which may not always capture sudden or unforeseen changes in workload patterns. Moreover, the accuracy of the predictions may be influenced by factors such as data quality, modeling assumptions, and the complexity of the underlying system architecture. Furthermore, the automated migration or provisioning of VMs may introduce additional overhead and complexity, requiring careful monitoring and management to ensure smooth operation.

The proposed system utilizes real-time data feeds to predict resource usage for virtual machines (VMs) hosted on ESXi hypervisors. By collecting and preprocessing data on CPU, memory, and disk usage, the system predicts future resource needs and detects VMs exceeding 80% resource consumption. The system then automatically migrates or

provisions VMs to maintain optimal performance and prevent downtime. Through the analysis, it's evident that the predictive modeling approach based on historical resource usage data can effectively forecast future resource needs. By identifying VMs with high resource consumption, the system can proactively manage host resource allocation, thus ensuring efficient resource utilization and minimizing service disruptions.

The proposed system offers a proactive and automated approach to resource management in virtualized environments. By leveraging predictive modeling and automation technologies, the system effectively addresses the challenges of resource allocation and optimization, thereby improving the reliability, scalability, and efficiency of ESXi hypervisor deployments. However, ongoing monitoring, evaluation, and refinement are essential to ensure the system's continued effectiveness and adaptability to evolving workload demands and system requirements.

V. CONCLUSION

The experiments conducted in this study aimed to assess the effectiveness of a predictive resource provisioning approach in preventing server downtimes due to resource overutilization in ESXi servers. The results demonstrate that leveraging historical workload data and deploying a predictive analytics model for forecasting critical periods, coupled with a dynamic provisioning mechanism, significantly enhances the stability and resilience of virtualized environments. The proactive nature of the proposed approach not only prevents server downtimes but also contributes to increased overall system reliability and performance. Future research could explore additional refinements to the predictive analytics model, considering evolving workload patterns and incorporating real-time adjustments to the provisioning strategy. Overall, this aims to enhance the stability, performance, and reliability of virtualized environments, thereby improving the overall service delivery and user experience. With proper implementation and ongoing monitoring, this solution can help organizations effectively manage their virtual infrastructure while meeting the demands of their customers.

REFERENCES

- [1]. R. S. Alhumaima, R. K. Ahmed and H. S. Al-Raweshdy, "Maximizing the Energy Efficiency of Virtualized C-RAN via Optimizing the Number of Virtual Machines," in IEEE Transactions on Green Communications and Networking, vol. 2, no. 4, pp. 992-1001, Dec. 2018, doi: 10.1109/TGCN.2018.2859407.
- [2]. F. Farahnakian, T. Pahikkala, P. Liljeberg, J. Plosila, N. T. Hieu and H. Tenhunen, "Energy-Aware VM Consolidation in Cloud Data Centers Using Utilization Prediction Model," in IEEE Transactions on Cloud Computing, vol. 7, no. 2, pp. 524-536, 1 April-June 2019, doi: 10.1109/TCC.2016.2617374.
- [3]. F. -H. Tseng, X. Wang, L. -D. Chou, H. -C. Chao and V. C. M. Leung, "Dynamic Resource Prediction and Allocation for Cloud Data Center Using the Multiobjective Genetic Algorithm," in IEEE Systems Journal, vol. 12, no. 2, pp. 1688-1699, June 2018, doi: 10.1109/JSYST.2017.2722476.
- [4]. L. Abdullah, H. Li, S. Al-Jamali, A. Al-Badwi and C. Ruan, "Predicting Multi-Attribute Host Resource Utilization Using Support Vector Regression Technique," in IEEE Access, vol. 8, pp. 66048-66067, 2020, doi: 10.1109/ACCESS.2020.2984056.

- [5]. X. Fu and C. Zhou, "Predicted Affinity Based Virtual Machine Placement in Cloud Computing Environments," in *IEEE Transactions on Cloud Computing*, vol. 8, no. 1, pp. 246-255, 1 Jan.-March 2020, doi: 10.1109/TCC.2017.2737624.
- [6]. B. Xia, T. Li, Q. Zhou, Q. Li and H. Zhang, "An Effective Classification-Based Framework for Predicting Cloud Capacity Demand in Cloud Services," in *IEEE Transactions on Services Computing*, vol. 14, no. 4, pp. 944-956, 1 July-Aug. 2021, doi: 10.1109/TSC.2018.2804916.
- [7]. H. -W. Li, Y. -S. Wu, Y. -Y. Chen, C. -M. Wang and Y. -N. Huang, "Application Execution Time Prediction for Effective CPU Provisioning in Virtualization Environment," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 11, pp. 3074-3088, 1 Nov. 2017, doi: 10.1109/TPDS.2017.2707543
- [8]. S. B. Melhem, A. Agarwal, N. Goel and M. Zaman, "Markov Prediction Model for Host Load Detection and VM Placement in Live Migration," in *IEEE Access*, vol. 6, pp. 7190-7205, 2018, doi: 10.1109/ACCESS.2017.2785280.
- [9]. L. Li, J. Dong, D. Zuo and J. Wu, "SLA-Aware and Energy-Efficient VM Consolidation in Cloud Data Centers Using Robust Linear Regression Prediction Model," in *IEEE Access*, vol. 7, pp. 9490-9500, 2019, doi: 10.1109/ACCESS.2019.2891567.
- [10]. X. Zhang, Z. Huang, C. Wu, Z. Li and F. C. M. Lau, "Dynamic VM Scaling: Provisioning and Pricing through an Online Auction," in *IEEE Transactions on Cloud Computing*, vol. 9, no. 1, pp. 131-144, 1 Jan.-March 2021, doi: 10.1109/TCC.2018.2840999.
- [11]. Moreno-Vozmediano, R., Montero, R.S., Huedo, E. et al, "Efficient resource provisioning for elastic Cloud services based on machine learning techniques," *J CloudComp* 8, 5 (2019).
- [12]. Devi, K.L., Valli, S, "Time series-based workload prediction using the statistical hybrid model for the cloud environment," *Computing* 105, 353-374 (2023).
- [13]. Souhei Yanase, Graduate Student Member, Fujun He and Eiji Oki, "Approximation Algorithms to Distributed Server Allocation With Preventive Start-Time Optimization Against Server Failure," in *IEEE Networking Letters*, Vol. 3, No. 4, December 2021
- [14]. Antra Malhotra, AMR Elsayed, Randolph Torres and Srinivas Venkatraman, "Evaluate Solutions for Achieving High Availability or Near Zero Downtime for Cloud Native Enterprise Applications," in *IEEE Access* vol. 11, doi: 10.1109/ACCESS.2023.3303430.
- [15]. A. U. Rehman Rui L. Aguiar, and JOÃO Paulo Barracca, "Fault-Tolerance in the Scope of Cloud Computing," in *IEEE Access* vol. 10, 2022, doi: 10.1109/ACCESS.2022.3182211.
- [16]. Zhinan Cheng, Shujie Han, Patrick P. C. Lee, Xin Li, Jiongzhou Liu and Zhan Li, "An In-Depth Correlative Study Between DRAM Errors and Server Failures in Production Data Centers," in 2022 41st International Symposium on Reliable Distributed Systems (SRDS) IEEE, DOI: 10.1109/SRDS55811.2022.00032
- [17]. Song Wu, Yusheng Yi, Jiang Xiao, Hai Jin, and Mao Ye, "A Large-Scale Study of I/O Workload's Impact on Disk Failure," in *IEEE Access* Vol. 6, 2018, doi: 10.1109/ACCESS.2018.2866522.
- [18]. Jargalsaikhan Narantuya, Hannie Zang, and Hyuk Lim, "Service-Aware Cloud-to-Cloud Migration of Multiple Virtual Machines," in *IEEE Access* Vol. 6, 2018, doi: 10.1109/ACCESS.2018.2882651.
- [19]. Congfeng Jiang, Guangjie Han, Jiangbin Lin, Gangyong Jia, Weisong Shi, and Jian Wani, "Characteristics of Co-Allocated Online Services and Batch Jobs in Internet Data Centers: A Case Study from Alibaba Cloud," in *IEEE Access* Vol. 7, 2019, doi: 10.1109/ACCESS.2019.2897898.
- [20]. Jing Li, Rebecca J. Stones, Gang Wang, Zhongwei Li, Xiaoguang Liu, and Jianli Ding, "New Metrics for Disk Failure Prediction That Go Beyond Prediction Accuracy," in *IEEE Access* Vol. 6, 2018, doi: 10.1109/ACCESS.2018.2884004

Picture with Guide and Team





NITTE MEENAKSHI
INSTITUTE OF TECHNOLOGY

Friday, 05th April 2024

IEEE YESIST12 PRELIMS
"KAUSHALYA"- OPEN HOUSE PROJECT EXPO

Certificate of Appreciation

This certificate is presented to Dr. Karunkara Rai B

for his/her contribution towards the **GUIDANCE** for the project titled
Development of Predictive VM provisioning
in a cloud environment
for IEEE YESIST12 Prelims 'KAUSHALYA'-Open House Project Expo - 2024
organized by the Department of ECE and IEEE Student Branch,
Nitte Meenakshi Institute of Technology, Bengaluru, in association with
IEEE Bangalore Section on 5th April 2024.

Dr. Parameshchari B D
Prof. & Head, Dept. of ECE
NMIT, Bengaluru

Dr. H C Nagaraj
Principal
NMIT, Bengaluru



Ballari Institute of Technology & Management, Ballari

Autonomous Institute under VTU, Belagavi (A unit of T.E.H.R.D. Trust © An ISO 9001:2015 Certified Institution)

Certificate of Participation

This is to recognise the contribution of

Karunakara Rai B

Development of a Predictive VM Provisioning in a Cloud Environment
as **PARTICIPATION** for **3rd IEEE International Conference on Distributed
Computing and Electrical Circuits and Electronics (ICDCECE-2024)** organized by
IEEE SB Ballari Institute of Technology and Management, Ballari, 26-27 April, 2024

Dr. Yadavalli Basavaraj
Principal
BITM, Ballari

Mr. Y.J. Prithviraj Bhupal
Director, BITM, Ballari & Pro Chancellor,
Kishinda University, Ballari




NITTE MEENAKSHI
INSTITUTE OF TECHNOLOGY


Friday, 05th April 2024

IEEE YESIST12 PRELIMS
"KAUSHALYA" - OPEN HOUSE PROJECT EXPO

Certificate of Participation

This is to certify that SUDEEKSHA K
of NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY
has presented the project entitled Development of Predictive
VM Provisioning in a cloud environment.
in IEEE YESIST12 Prelims 'KAUSHALYA' - Open House Project Expo - 2024
organized by the Department of ECE and IEEE Student Branch,
Nitte Meenakshi Institute of Technology, in association with IEEE
Bangalore Section on 5th April 2024.


Dr. Parameshchary B D
Prof. & Head, Dept. of ECE
NMIT, Bengaluru


Dr. H C Nagaraj
Principal
NMIT, Bengaluru



Ballari Institute of Technology & Management, Ballari

Autonomous Institute under VTU, Belagavi (A unit of T.E.H.R.D. Trust © An ISO 9001:2015 Certified Institution)

Certificate of Participation

This is to recognise the contribution of

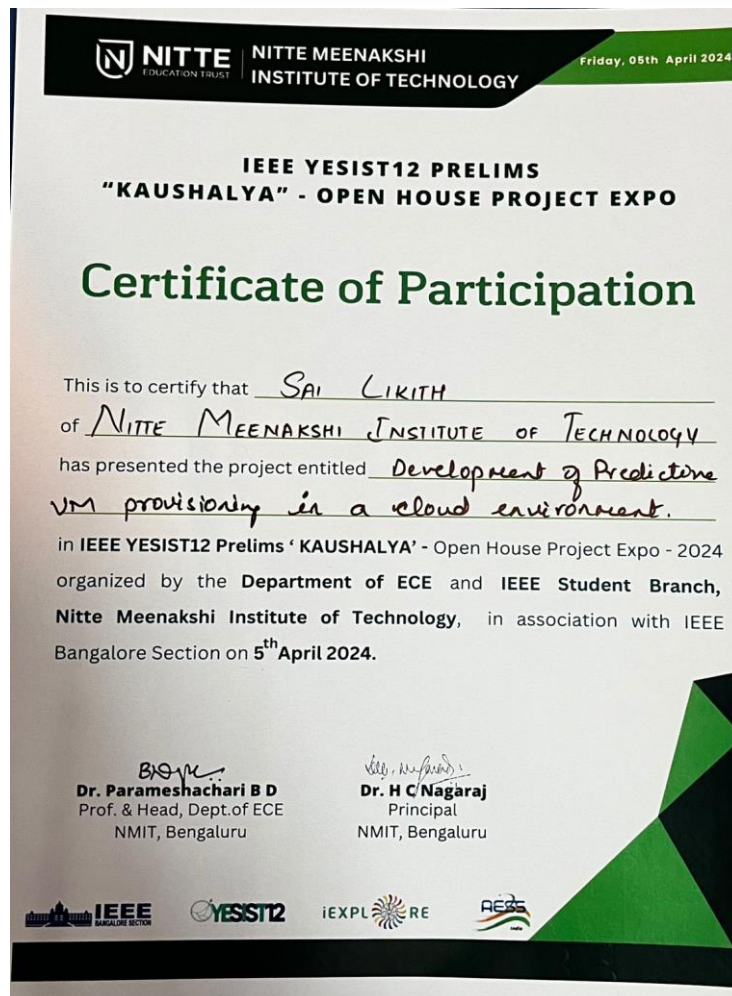
Sudeeksha K

Development of a Predictive VM Provisioning in a Cloud Environment
as **PARTICIPATION** for **3rd IEEE International Conference on Distributed
Computing and Electrical Circuits and Electronics (ICDCECE-2024)** organized by
IEEE SB Ballari Institute of Technology and Management, Ballari, 26-27 April, 2024


Dr. Yadavalli Basavaraj
Principal
BITM, Ballari


Mr. Y.J. Prithviraj Bhupal
Director, BITM, Ballari & Pro Chancellor,
Kishkinda University, Ballari







Ballari Institute of Technology & Management, Ballari

Autonomous Institute under VTU, Belagavi (A unit of T.E.H.R.D. Trust © An ISO 9001:2015 Certified Institution)

Certificate of Participation

This is to recognise the contribution of

Rajani N

Development of a Predictive VM Provisioning in a Cloud Environment

as **PARTICIPATION** for **3rd IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE-2024)** organized by IEEE SB Ballari Institute of Technology and Management, Ballari, 26-27 April, 2024

Dr. Yadavalli Basavaraj
Principal
BITM, Ballari

Mr. Y.J. Prithviraj Bhupal
Director, BITM, Ballari & Pro Chancellor,
Kishkinda University, Ballari

Contact Information

Guide: Dr. Karunakara Rai B
E-Mail: karunakara.rai@nmit.ac.in
Ph. No: 9844286965

Guide: Dr. Rajani N
E-Mail: Rajani.n@nmit.ac.in
Ph. No: 9740798868

Sai Likith P
USN: 1NT20EC126
E-Mail: psailikith12@gmail.com
Ph. No: 9113504205

Santhosh P
USN: 1NT20EC133
E-Mail: santhosh12345ind@gmail.com
Ph. No: 9880505575

Sudeeksha K
USN: 1NT20EC151
E-Mail: sudeeksha24113@gmail.com
Ph. No: 6366605221