SUMMARY

Cloud computing is described as a model for offering services to computer resources including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) by the National Institute of Standards and Technology (NIST) [1]. IaaS is a type of cloud computing services which furnishes its services through resources which are virtualized that includes virtual storage, computers, infrastructures and other hardware assets. The requirement of these resources can be smoothly increased up or down in real time as per the demand. PaaS, a cloud service provider (CSP), offers hardware and software tools such applications, operating systems, services, control structures, transactions, and physical and virtual computers. Software as a Service (SaaS), often known as on-demand software, is a service that a cloud provider offers that includes an entire operating system, including all management, applications, and user interface (i.e., Google Apps, Windows Azure Platform, SalesForce.com, Office 365). In a private cloud, an organisation deploys the infrastructure; in a community cloud, several entities share the infrastructure; in a public cloud, an organisation owns the infrastructure and provides cloud services; and a combination of public and private clouds is hybrid cloud. Cloud computing provides on-demand self-service, broadband network connectivity, and rapid flexibility, where capabilities are instantly offered and dismissed by cloud users.

Cloud computing is a utility-based computing paradigm. It provides a seamless acquisition of computing, storage and network resources to users on a payment basis, in a similar fashion like that of water, gas and electricity. Cloud computing is highly embraced by individuals and business organizations due to the advantages it offers, few of them be like-

- Zero or no opening investment for accessing computing resources like computing, memory and/or network as these are measured in metrical units and provided to cloud users on a demand basis.
- Cloud resources or services are available round the clock with negligible downtime and can be accessed from any location.

- Cloud users' privacy and security features are well-maintained. Moreover, cloud functioning is transparent to its users.
- Cloud users are charged according to their utilization of cloud resources, thereby, extending the advantage of scalability.
- Cloud services are made available on a payment basis and are categorized as software, platform or infrastructure services

The key mechanism behind cloud computing is virtualization which creates an abstraction of actual cloud resources. These abstracted resources are enveloped in VMs with same functions and interfaces as that of real/actual resources. An important aspect of virtualization is that it is transparent to the cloud user. These virtual resources are offered in different sizes, performances and costs. The entire life cycle of a virtual machine is controlled a virtual machine monitor (VMM) or a hypervisor software. A physical machine or server contains numerous VMs and is known as a host. VMs residing on a host are therefore termed as guests. Hypervisor acts as a bridge between the actual resources of a host and its multiple guests by giving a virtual operating environment to execute the guest VM tasks.

The core component of virtualization and cloud computing is RM System (RMS). The resource acquisition and release process are handled by RMS. Using RMS, both physical and virtual resources are managed. Whenever a request is raised from a user for resources, RMS examines the state of the actual resources to determine the availability of the resources sufficient enough to fulfil the users' request. The uninterrupted surveillance of the resources allotted to users while optimizing the system is handled by RMS. With the completion of task, all resources are released and added to the pool of accessible resources. This task is carried out in accordance with the SLA which has been established between the client and the CSP. SLA provides information about the service level required by a tenant. It also includes information on the payment mechanism and the fine for SLA infractions. Since there are numerous competing goals that need to be optimised, such as maintaining performance, cutting expenses, and enhancing security, it can be described as a multi-objective problem.

Because the cloud computing paradigm is market-oriented, standard system-centric RM architectures are inadequate for such systems. These types of architectures treat all

requests equally and don't offer any enticement to the CSP. So, for cloud environments that can support resource provisioning whenever in demand and sharing the resource among users, certain market-oriented methodologies have been developed. These RM strategies give customers and CSP financial incentives. On a pay-per-use basis, CSP distribute their resources among numerous renters.

The chapter-wise summary of the research is given below in brief:

CHAPTER 1
INTRODCUTION

This chapter provides an introduction to cloud computing technology including the discussion about its evolution and architecture. A number of widely accepted and most cited definitions of cloud computing are also given. It further described the converged technologies in cloud computing and its characteristics. It furnishes the brief description on cloud deployment models and services. Cloud computing is a special type of distributed computing in which any type of resource, physical or virtual, can be made available to the users, worldwide, by means of powerful technology called virtualization of resources. Virtualization is given due importance with discussion on its types and utilities. Cloud computing is a special type of distributed computing in which any type of resource, physical or virtual, can be made available to the users, worldwide, by means of powerful technology called virtualization of resources. Virtualization is given due importance with discussion on its types and utilities.

CHAPTER II
REVIEW OF LITERATURE

The chapter provides the details of the existing literature on cloud computing and furnishes an elaborated the review of the existing work various issues of resource management like virtual machine allocation, load balancing, scaling, provisioning and security issues. The rapid adoption of cloud computing is done by the society emphasis the need efficient and feasible resource management techniques for managing the cloud

resources. An extensive study of related research papers on resource management techniques suggests that the objectives and implementation of resource management in cloud networks are very different from classical networks. For example, in classical networks resources are only physical while in cloud, resources are physical as well as virtual. Therefore, a different approach is required for cloud computing to manage resources effectively. Several reputed journals, e-books, etc. are consulted for understanding the new research problems.

CHAPTER III
VIRTUAL MACHINE ALLOCATION USING BINARY WHALE OPTIMIZATION

An efficient VM allocation technique is one of the major concerns of the utility-based computing. The allocation of VM is significantly impacted by resource usage which includes CPU, memory and quantity of server's utilization. This highly influence the expense and energy usage in the procedure. A genetic based allocation technique is applied for the allocation of the VMs to the server. Due to its complexity another method that is binary whale optimization algorithm is applied for the allotment of VMs to server. The whale algorithm is inspired from the hunting process of humpback whale. Since the placement problem is discrete in nature, applying the whale optimization is not suggestable due to its usability for continuous problem, hence binary version of whale algorithm is applied. The technique has efficacious exploration and exploitation procedure which enhances its capability to generate efficient results. The technique not only reduces the amount of servers but also the expenditure of the CSP and users as it highly contributes in minimizing the consummation of energy and thus the expense of infrastructure is also highly reduced.

The content of this chapter is published in:

1. Ankita Srivastava, Narander Kumar, "Virtual Machine Allocation Using Genetic-Based Algorithm in Cloud Infrastructure", Second International Conference on Computational Electronics for Wireless Communications. Lecture Notes in Networks and Systems, Springer, vol 554, pp 273-282, 2023.

2. Ankita Srivastava, Narander Kumar, "Multi-Objective Binary Whale Optimization-Based Virtual Machine Allocation in Cloud Environments", International Journal of Swarm Intelligence Research (IJSIR), IGI Global, Vol. 14, No. 1, pp 1-23, 2023.

CHAPTER IV
LOAD BALANCING USING FIREFLY AND HONEYBEE MECHANISM

Load Balancing is stated as the administered dispersal of workloads among the resources. It facilitates the cloud providers to supervise the workload and application demands by rationing the resources among the servers, networks, or computers. This work has suggested a novel LB algorithm, encompassing the fusion of the Firefly algorithm and improved Bee algorithm. The methodology has adopted the Firefly algorithm with the motive of exploration of potential search space and further Bee algorithm technique is applied to eliminate the situation of over-usage and under-usage of resources and assign the workloads with appropriate resources. The BA technique works by carrying out legitimate task allocation to those VMs whose load is below its capacity and segregates the VMs into the overloaded, underloaded, and balanced categories, further migrating the tasks from overloaded and underloaded VMs for maintaining the system's balance.

The content of this chapter is published in:

1. Ankita Srivastava, Narander Kumar, "A Discrete Firefly-based Task Scheduling Algorithm for Cloud Infrastructure", International Conference on Data Analytics and Computing (ICDAC-2022), Springer Singapore, 2023. (In Press)
2. Ankita Srivastava, Narander Kumar "An Efficient Firefly and Honeybee based Load Balancing Mechanism in Cloud Infrastructure", Cluster Computing, 2023. **(SCI)** (Accepted)

CHAPTER V

QUEUING BASED SCALING APPROACH FOR CONTAINERIZED CLOUD

Cloud computing is a new technology that has acquired universal acceptance in the scientific community as well as huge organisations such as government and industry. Lightweight containers have grown in prominence as a result of the exceptionally complicated nature of VM virtualization, and approaches for provisioning resources to these containers have grabbed academics' attention. There are not enough models or techniques in literature that allows scalability in real time in fewer amount of resources for container clouds while sorting out the performance issue and QoS. The dynamic scalability of cloud services enables them to provide timely, on-demand, and computational resources with dynamic modification to end users. This chapter offered a technique that used queuing model for application of scaling process dynamically and aids in scaling the resources of the containers. The methodology helps in mitigating the cost and reduces the penalties of SLA. The study tries to improve virtual resource utilisation while meeting SLA requirements for drop rate, response time, system throughput, and container count.

The content of this chapter is published in:

1. Ankita Srivastava, Narander Kumar, “Queueing Model based Dynamic Scalability for Containerized Cloud”, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 14, No.1, pp 1-12, 2023.

CHAPTER VI

ENERGY EFFICIENT RESOURCE PROVISIONING THROUGH PSO-ANN BASED APPROACH

Computing has been revolutionised by high-performance cloud computing. Resource provisioning has developed into a crucial obstacle in the high-end service paradigm that facilitates computing. Due to its ability to reduce the cost of the cloud system, provisioning which is efficient in downsizing the energy usage has emerged as the

biggest success in cloud. A method has been put out to address the energy issue by provisioning the jobs to virtual machines. The method attempted to reduce the consumption of system's energy, total time, and average initiation time in order to increase the effectiveness of the scheduling technique in cloud application in real time. The Particle Swarm Optimization (PSO) method is exploited in the technique, which draws inspiration from Artificial Neural Networks (ANN). By reducing energy use, it completes the provisioning in the cloud datacentre.

The content of this chapter is published in:

1. Ankita Srivastava, Narander Kumar, "An energy efficient robust resource provisioning based on improved PSO-ANN", International Journal of Information Technology, Springer, Vol. 15, No 1, pp 107-117, 2022.

CHAPTER VII
SECURITY TECHNIQUE AGAINST CO-RESIDENCE ATTACK IN VM ALLOCATION

With the increasing number of co-residence attacks, the security of the multi-tenant public IaaS cloud environment has become a growing concern. The co-residence attacker creates a side channel to retrieve the secured data. These attacks help the adversary to leak out the sensitive information of the user with whom it is co-located. This work has also discussed a secured VM placement technique, Previous Server and Co-resident users First (PSCF), which focuses on facilitating security against the co-residence attack by minimizing the probability of co-locating the malicious user with the authentic user. Co-location resistance and core utilization metrics are utilized to evaluate the algorithm's performance. Further, simulation has been performed to analyse the performance of the algorithm.

The content of this chapter is published in:

1. Ankita Srivastava, Narander Kumar, "A Secure VM Placement Strategy to defend against Co-residence Attack in Cloud Datacenters", International

Journal of Computer Network and Information Security (IJCNIS), 2023. (In Press)

CHAPTER VIII
CONCLUSIONS AND FUTURE PERSPECTIVES

This chapter is devoted to the conclusions of the presented research work and the future perspectives in the area of cloud resource management. Cloud computing systems have a long way to go. Future computing world will strive on virtual resources promising seamless and continuous services. Hence, the issue of cloud resources management must be addressed in a timely and priority fashion. In the present research work, several issues of cloud resource management have been identified and independent solutions to each issue have been proposed. These are –

- VM Allocation
- Load Balancing
- Scaling
- Energy Efficiency
- Security

The solutions provided for each sub-problem are feasible, scalable and dynamic in nature and efficiently manage cloud resources as validated by the simulation results. As future perspectives, this work can be extended for newer DCN architectures and interoperable clouds which enable a cloud user with the flexibility of shifting his/her acquired resources between datacentres and no disruption in serviceability.