

# A Review of Enhanced Algorithm for Dynamic Load Balancing in Cloud Computing Environment

Avneet Kaur

Dept. of Computer Science, Sri Guru Harkrishan College of Management and Technology, Patiala, Punjab India

## Abstract

Load balancing has been considered as one of the most important aspect of cloud computing in recent times. An increase in the number of users around the world has resulted in a large number of requests at a rapid rate. Researchers around the world have designed many algorithms to carry out the client's request at distributed cloud servers. Based on this, the cloud computing paradigm will automate configuration of servers in order to achieve efficient load balancing. Henceforth, selection of virtual machines has to be scheduled efficiently based on the load balancing algorithm. In this paper, a load balancing algorithm is proposed based on the availability of the VM. Specifically, the Availability Index (AI) value is evaluated for every VM over a given period of time, and therefore a task is assigned to that machine based on the AI value. In order to validate the proposed model, it is compared with three famous load balancing algorithms are compared, namely Round Robin, Throttled and Active Monitoring. The performance of each algorithm was evaluated using Cloud Analyst. Simulation results show that the proposed algorithm is more efficient in load balancing over virtual machines as compared to other algorithms.

## Keywords

Cloud computing, Modified throttled, Virtual machine, Throttled algorithm, Round-robin algorithm, Active monitoring

## I. Introduction

Because of the advancement in Information and Communication Technology (ICT) over past few years, Computing has been considered as a utility like water, electricity, gas and telephony [1]. These utilities are available at any time to the consumers based on their requirement. Consumers pay service providers based on their usage.

Like all the other existing utilities, Computing utility is the basic computing service that meets the day to day needs of the general community. To deliver this vision, a number of computing paradigms have been proposed, of which the latest one is known as Cloud Computing. Cloud is nothing but large pool of easily accessible and usable virtual resources.

Dr. Rajkumar Buyya says "A Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resource(s) based on service-level agreements established through negotiation between the service provider and consumers." [2] Cloud computing is a computing model that maintains statistics and applications, using internet and central secluded servers. This methodology permits end users and businesses to use applications without putting in and entrée their private records at any computer with internet entrée. Cloud computing permits for much more proficient computing by centralizing storage, reminiscence, dispensation and bandwidth. Some examples of cloud computing are Yahoo email, Google, Gmail, or Hotmail etc. The server and email administration software is all on the cloud and is completely

managed by the cloud service supplier. The end user gets to use the software unaccompanied and get pleasure from the benefits. Cloud computing acts as a service moderately than a merchandise, whereby mutual resources, software, and information are provided to computers and other strategies. Cloud computing can be categorized into three services:

- SaaS (software-as-a-service)
- PaaS (platform-as-a-service)
- IaaS (infrastructure-as-service) respectively [3].

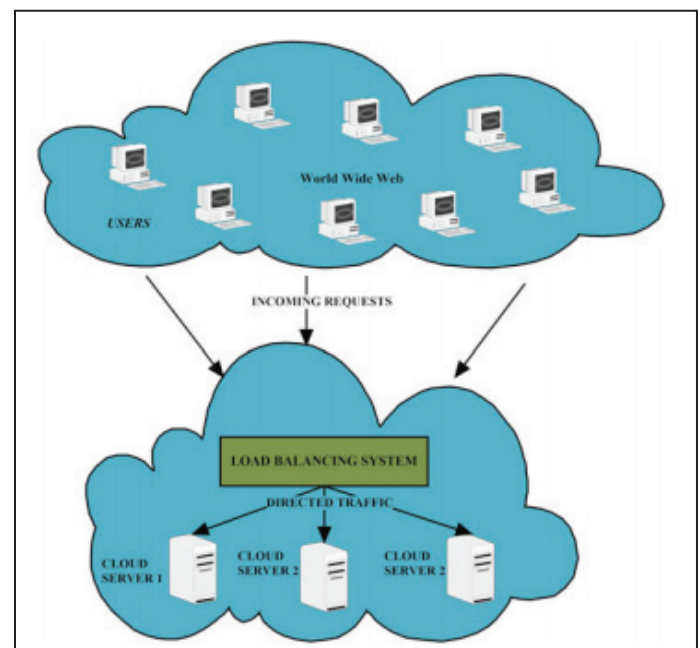


Fig. 1: Load Balancing in Cloud Computing Environment

## A. Cloud Software as a Service (SaaS)

In this service model, instead of using locally run applications the cloud consumer uses the cloud provider's software services running on a cloud infrastructure. It is the job of cloud provider to maintain and manage the software services that are used by the cloud consumer. The cloud provider may charge according to quantity of software and using time. SaaS is the best way to use advanced technology. Salesforce.com and Customer Relationship Management (CRM) are the examples of such service model.

## B. Cloud Platform as a Service (PaaS)

In this service model, the cloud platform offers an environment on which developers create and deploy applications. It provides platform where applications and services can run. The consumers do not need to take care of underlying cloud infrastructure including network, servers, operating system or storage but has a control over deployed application. Google Application Engine, Microsoft Azure and RightScale are the example of such model [4].

## C. Cloud Infrastructure as a Service (IaaS)

In this service model, cloud providers manage large set of

computing resources such as storing and processing capability. Cloud consumer can control operating system; storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls). Sometimes it is also called as a Hardware as a Service (HaaS). The cost of the Hardware can be greatly reduced here. Amazon Web Services, Open Stack, Eucalyptus, GoGrid and Flexiscale offers IaaS.

#### D. Cloud Resource Allocation

The overlapping concept between resource provisioning, resource allocation, and resource scheduling is briefly defined by as: resource provisioning is the allocation of a service provider's resources to a customer, while resource allocation is the process of distributing resources economically between competing groups of programs or users, and resource scheduling is a timetable of allocation of resources where resources are shared and available at certain times, and computational events are planned during these times. In other words, it is the process of defining about when a computational activity should start or end, contingent on its (1) predecessor activities, (2) predecessor relationships, (3) resources allocated, and (4) duration. Further, cloud resource allocation is the process of resource discovery, selection, provisioning, application scheduling, and management of resources.

In addition, cloud resource allocation involves decision making with respect to howmuch, what, when, and where to allocate the available resources to the user (represented as a block diagram in fig. 1). Generally, users determine the amount and type of the resources for the request, and in response, the service providers allocate the requested resource containers in their data centers. For the efficient execution of applications, the type and the numbers of resource containers should be sufficient to meet the constraints (e.g., job completion time deadline) and should match the workload characteristics. The elasticity in a cloud computing environment enables the users to request or return resources dynamically; here it is also worth mentioning to consider how to realize such adjustments.

Therefore, one must take the characteristics and behavior of actors in a cloud computing environment into account to provide efficient cloud services and cloud-based applications. By "efficient," we mean that suitable resources are allocated to an appropriate application at an appropriate time, such that applications can utilize the resources effectively. In other words, efficient resource allocation maximizes throughput (or minimizes job completion time) of an application or minimizes the amount of resources for an application to maintain an acceptable level of service quality.

#### E. Significance of Resource Allocation

In cloud computing, Resource Allocation (RA) [5] is the process of assigning available resources to the needed cloud applications over the internet. Resource allocation starves services if the allocation is not managed precisely. Resource provisioning solves that problem by allowing the service providers to manage the resources for each individual module. Resource Allocation Strategy (RAS) is all about integrating cloud provider activities for utilizing and allocating scarce resources within the limit of cloud environment so as to meet the needs of the cloud application. It requires the type and amount of resources needed by each application in order to complete a user job. The order and time of allocation of resources are also an input for an optimal RAS. An optimal RAS should avoid the following criteria as follows:

- Resource contention situation arises when two applications try to access the same resource at the same time.

- Scarcity of resources arises when there are limited resources.
- Resource fragmentation situation arises when the resources are isolated. [6]
- Over-provisioning of resources arises when the application gets surplus resources than the demanded one.
- Under-provisioning of resources occurs when the application is assigned with fewer numbers of resources than the demand.

Resource users' (cloud users) estimates of resource demands to complete a job before the estimated time may lead to an over-provisioning of resources. Resource providers' allocation of resources may lead to an under-provisioning of resources. To overcome the above mentioned discrepancies, inputs needed from both cloud providers and users for a RAS as shown in table I. From the cloud user's angle, the application requirement and Service Level Agreement (SLA) are major inputs to RAS. The offerings, resource status and available resources are the inputs required from the other side to manage and allocate resources to host applications by RAS. The outcome of any optimal RAS must satisfy the parameters such as throughput, latency and response time. Even though cloud provides reliable resources, it also poses a crucial problem in allocating and managing resources dynamically across the applications.

From the perspective of a cloud provider, predicting the dynamic nature of users, user demands, and application demands are impractical. For the cloud users, the job should be completed on time with minimal cost. Hence due to limited resources, resource heterogeneity, locality restrictions, environmental necessities and dynamic nature of resource demand, we need an efficient resource allocation system that suits cloud environments.

Cloud resources consist of physical and virtual resources. The physical resources are shared across multiple compute requests through virtualization and provisioning. The request for virtualized resources is described through a set of parameters detailing the processing, memory and disk needs which is depicted in Fig.1. Provisioning satisfies the request by mapping virtualized resources to physical ones. The hardware and software resources are allocated to the cloud applications on-demand basis. For scalable computing, Virtual Machines are rented.

The complexity of finding an optimum resource allocation is exponential in huge systems like big clusters, data centers or Grids. Since resource demand and supply can be dynamic and uncertain, various strategies for resource allocation are proposed. This paper puts forth various resource allocation strategies deployed in cloud environments.

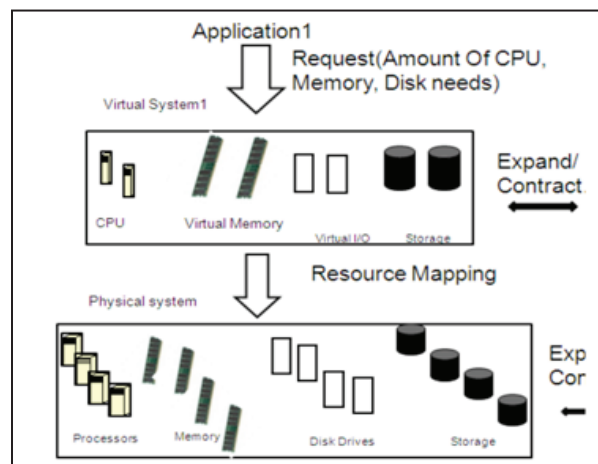


Fig. 1: Mapping of Virtual to Physical Resources

## II. Applications

Applications hooked up on vehicular touch scope from clean transactions of vehicle rank statistics to pretty convoluted huge-scale site visitor's affiliation encompassing foundation integration. As a onset to studying requests, this serving offers an overview of envisioned request corporations for vehicular networks. Even though particular system functions aren't but uniform for maximum requests, and in spite the reality that this type of collection cannot ever be absolutely finished, the evaluation brings frank mechanisms, parts, and constraints encompassed in the arrangement This gives an early expertise of the residences of VANET touch and ends in a extra methodical scrutiny of net traits inside the subsequent phase.

### A. Safety Applications

Safety requests can reduce drastically the variety of road accidents. According to a little studies 60 percentages of accidents is probably evaded if a driving force were endowed along a notice half a next beforehand the moment of collision. There are 3 most important eventualities in that protection requests is probably extremely beneficial.

#### 1. Accidents

Vehicles tour at a expanded pace on principal roads. This offers motive force's extraordinarily mild length to react to the vehicle in front of them. If an mishap occurs, the imminent cars frequently crash beforehand they could come to a forestall. Protection requests is probably utilized to alert cars of an mishap that transpired greater alongside the road, therefore preventing a pile-up from happening. A protection request moreover might be applied to supply drivers along primary warnings and stop an mishap from transpiring inside the early location.

#### 2. Intersections

Steering adjacent and throughout intersections is one of the maximum convoluted trials that drivers face because or extra traffic flows intersect, and the ability of come across is high. According to the U.S. Department of Transportation (DoT), intersection crashes accounted for additional than 45 percentage of all defined crashes and 21 percent of fatalities, this is, 9213 fatalities transpired at intersections inside the United States. The number of injuries needs to reduce if a safety request counseled the driver of an approaching collision. The motive force next might seize deed to prevent it.

#### 3. Road Congestion

Protection requests moreover are probably applied to provide drivers alongside the exceptional paths to their destinations. This should reduce congestion on the street and uphold a flat go with the flow of traffic, therefore rising the capacity of the roads and stopping site visitors jams. It additionally would possibly have the oblique result of slicing traffic injuries due to the fact drivers need to be much less angry and additional inclined to pursue visitor's guidelines.

#### 4. User Applications

User requests can provide avenue customers alongside information, advertisements, and amusement across their adventure. Two frank consumer-associated requests are delineated underneath.

#### 5. Internet Connectivity

Regular Internet admission has turn out to be a day by day necessity

for limitless folks and due to the fact endless person requests additionally needs Internet connectivity, bestowing this capacity to car occupants and supplementary VANET requests is important. Moreover, this way that the usual company framework could be present seamlessly in automobiles, missing a need for specific redevelopment.

### 6. Peer-to-Peer Applications

To alleviate boredom, peer-to-peer requests moreover are a thrilling believed for VANETs. Travelers in the cars might allocate music, films, and so forth and chat along every single supplementary and frolic video games. They additionally would possibly flow track or movies from wonderful servers throughout long trips.

As a very last factor, requests industrialized for VANETs ought to protect those setbacks inherent in VANETs are invisible to the customers. In the pursuing servings, we gaze at a touch resolution that had been cautioned.

### III. Related Work

**García, A.G., Blanquer, 2015** [7] This paper propose a generic methodology for the representation of Cloud services. This methodology uses the WS-Agreement specification for capturing and manipulation arbitrary services using SLA fragments. SLA fragments are composed on the fly in response to user request. A SLA composition algorithm enables a prototype implementation of the methodology in a SLA-aware Cloud platform. This methodology provides the genericity, extensibility and flexibility to unify the modeling of Cloud services. Finally a use case provides a quantitative measure of the utility provided by the methodology from a Cloud user and Cloud provider point of view.

**Pascual, J.A., Lorido-Bostrán, T., Miguel-Alonso 2015** [8] We have tested three multi-objective optimization algorithms with problem-specific crossover and mutation operators. Simulation-based experiments demonstrate how, in comparison with classic placement techniques, a low-cost optimization results in improved assignments of resources, making applications run faster and reducing the energy consumed by the data center. This is beneficial for both cloud clients and cloud providers.

**Singh, S., Chana 2015** [9] This research depicts a broad methodical literature analysis of autonomic resource management in the area of the cloud in general and QoS (Quality of Service)-aware autonomic resource management specifically. The current status of autonomic resource management in cloud computing is distributed into various categories. Methodical analysis of autonomic resource management in cloud computing and its techniques are described as developed by various industry and academic groups. Further, taxonomy of autonomic resource management in the cloud has been presented. This research work will help researchers find the important characteristics of autonomic resource management and will also help to select the most suitable technique for autonomic resource management in a specific application along with significant future research directions.

**Singh, S., Chana, I., Buyya 2016** [10] Cloud computing has transpired as a new model for managing and delivering applications as services efficiently. Convergence of cloud computing with technologies such as wireless sensor networking, Internet of Things (IoT) and Big Data analytics offers new applications' of cloud services. This paper proposes a cloud-based autonomic information system for delivering Agriculture-as-a-Service (AaaS) through the use of cloud and big data technologies. The proposed system gathers information from various users through preconfigured devices and IoT sensors and processes it in cloud

using big data analytics and provides the required information to users automatically. The performance of the proposed system has been evaluated in Cloud environment and experimental results show that the proposed system offers better service and the Quality of Service (QoS) is also better in terms of QoS parameters.

**Nguyen, Nguyen Cong 2017** [11] This paper reviews applications of the economic and pricing models to develop adaptive algorithms and protocols for resource management in cloud networking. Besides, we survey a variety of incentive mechanisms using the pricing strategies in sharing resources in edge computing. In addition, we consider using pricing models in cloud-based software defined wireless networking. Finally, we highlight important challenges, open issues and future research directions of applying economic and pricing models to cloud networking.

**Yousafzai, A., Gani, A., Noor 2017** [12] In this paper, current state-of-the-art cloud resource allocation schemes are extensively reviewed to highlight their strengths and weaknesses. Moreover, a thematic taxonomy is presented based on resource allocation optimization objectives to classify the existing literature. The cloud resource allocation schemes are analyzed based on the thematic taxonomy to highlight the commonalities and deviations among them. Finally, several opportunities are suggested for the design of optimal resource allocation schemes.

**Weerasiri, Denis, et al 2017** [13] This framework is essential to empower effective research, comprehension, comparison, and selection of cloud resource orchestration models, languages, platforms, and tools. This article provides such a comprehensive framework while analyzing the relevant state of the art in cloud resource orchestration from a novel and holistic viewpoint.

**Kumar, C. Ashok, et.al 2017** [14] Cloud computing is a developing technology that enables on-demand network access to the users through a shared pool of cluster computing resources. However, maintaining the stability of processing several tasks in the cloud environment is a complex issue. Hence, it requires a load balancing technique that allocates the task to the Virtual Machines (VMs) without affecting the performance of the system. This paper presents a technique for load balancing, called fractional dragonfly based load balancing algorithm (FDLA), by proposing two selection probabilities and fractional dragonfly algorithm. This paper presented a load balancing technique, named FDLA, in the cloud computing system using a novel load balancing algorithm, named fractional dragonfly, along with the development of two selection probabilities, known as TSP and VSP. The proposed load balancing model is based on the evaluation of parameters, such as capacity, and loads of machines. The proposed FDLA technique in the cloud environment is evaluated based on load and number of tasks reallocated. The performance of FDLA is compared with three existing techniques, such as PSO, HBB-LB, and DA, where the proposed FDLA could attain a minimum load of 0.2133 with the number of tasks reallocated as 14, showing that the proposed technique is effective with significant performance.

**Basu, Sayantani, et. al 2019** [15] Resource utilization and energy need to be carefully handled for achieving virtualization in the cloud environment. An important aspect to be considered is that of load balancing, where the workload is distributed so that a particular node does not become overburdened with tasks. Improper load balancing will lead to losses in terms of both memories as well as energy consumption. The resources should be scheduled in a cloud in such a way that users obtain access at any time and with possibly less energy wastage. Each virtual machine is allocated to a node. The virtual machines on every node correspond to the genes of a chromosome. Crossover and mutation operations have been

performed after which optimization techniques have been used to obtain the resulting allocation of tasks. The proposed approach has proved to be competent with respect to earlier approaches in terms of load balancing and resource utilization. An improved genetic algorithm with local search (GA-LS) has been proposed that handles load balancing and manages resource utilization. In this case, the objective function aimed at reducing the energy consumption and memory usage of virtual machines allocated to nodes in a cloud environment.

**Bhandari, Anmol et.al 2019** [17] Load balancing has been considered as one of the most important aspect of cloud computing in recent times. An increase in the number of users around the world has resulted in a large number of requests at a rapid rate. Researchers around the world have designed many algorithms to carry out the client's request at distributed cloud servers. Based on this, the cloud computing paradigm will automate configuration of servers in order to achieve efficient load balancing. Henceforth, selection of virtual machines has to be scheduled efficiently based on the load balancing algorithm. In this paper, a load balancing algorithm is proposed based on the availability of the VM. Specifically, the Availability Index (AI) value is evaluated for every VM over a given period of time, and therefore a task is assigned to that machine based on the AI value. In order to validate the proposed model, it is compared with three famous load balancing algorithms are compared, namely Round Robin, Throttled and Active Monitoring. The performance of each algorithm was evaluated using Cloud Analyst. Simulation results show that the proposed algorithm is more efficient in load balancing over virtual machines as compared to other algorithms. Load balancing algorithms take into account the principle that workload can be assigned in any situation, whether during compile time or at runtime. Moreover, since the increase in the user load has put a load on the computational capability of the VM, it becomes important to distribute load to appropriate VM. Henceforth, in this paper, a modified throttled balancing technique is proposed and is implemented on Cloud Analyst tool of CloudSim. More, it is validated by comparing it with other load balancing techniques. The above comparison shows that the proposed load balancing algorithm is more effective and efficient as compared to other conventional load balancing algorithms. However, evaluation of appropriate load in real time still remains an open challenge for the future perspective.

#### IV. Conclusion and Future Scope

Diverse techniques for ensuring optimized resource allocation in cloud computing environments have been surveyed and investigated both at the advanced level as well as the stumpy levels. The prose shows the counter actions which have been proposed to conquer the hurdles in mounting the speed and competence of the resource allocation. Though some tangible results have been obtained in ensuring the performance enhancement in dynamic resource allocation, there is scope for further enhancement. However, many issues remain unsolved. In the last two decades, the uninterrupted increase of computational power has produced an irresistible flow of data. The result of this is the manifestation of a clear opening between the quantity of data that is being produced and the capability of customary systems to accumulate, analyze and make the best use of this data. In topical years, cloud computing has gained much thrust due to its monetary advantages. In scrupulous, cloud computing has promised various advantages for its hosting to the exploitations of data-demanding applications. Modern cloud platforms increased the techniques to allocate resources in

a more efficient way. However, several scheduling strategies have been developed for dynamic and optimized resource allocation. Indeed, to appropriately assure applications with QoS demands resource accessibility and handling which directly bang on energy utilizations has to be tracked. Moreover the need for efficient allocation makes the administration of resources and energy saving a challenging design goal.

## References

- [1] Chana, I., Singh, S., "Quality of service and service level agreements for cloud environments: issues and challenges", In: *Cloud Computing-Challenges, Limitations and R&D Solutions*, pp. 51–72. Springer International Publishing, 2014.
- [2] Buyya, Rajkumar, et al., "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation computer systems* 25.6, pp. 599-616, 2009.
- [3] Weerasiri, Denis, et al., "A Taxonomy and Survey of Cloud Resource Orchestration Techniques." *ACM Computing Surveys (CSUR)* 50.2 (2017): 26.
- [4] García, A.G., Espert, I.B., García, V.H., "SLA-driven dynamic cloud resource management", *Futur. Gener. Comput. Syst.* 31, pp. 1–11, 2014.
- [5] Petcu, D., "Consuming resources and services from multiple clouds", *J. Grid Comput.* 12(2), pp. 321–345, 2014.
- [6] Singh, S., Chana, I., "Formal Specification Language Based IaaS Cloud Workload Regression Analysis", arXiv preprint arXiv:1402.3034. [Online] Available: <http://arxiv.org/ftp/arxiv/papers/1402/1402.3034.pdf> (2014)
- [7] Szabo, C., Sheng, Q.Z., Kroeger, T., Zhang, Y., Jian, Y.: *Science in the cloud: Allocation and execution of data-intensive scientific workflows*. *J. Grid Comput.* 12(2), pp. 245–264, 2014.
- [8] García, A.G., Blanquer, I., "Cloud services representation using SLA composition. *J. Grid Comput.* 13(1), pp. 35–51, 2015.
- [9] Pascual, J.A., Llorido-Bostrán, T., Miguel-Alonso, J., Lozano, J.A., "Towards a greener cloud infrastructure management using optimized placement policies", *J. Grid Comput.* 13(3), pp. 375–389, 2015.
- [10] Singh, S., Chana, I., "QoS-aware autonomic resource management in cloud computing: A systematic review", *ACM Comput. Surv.* 48(3), 39, 2015.
- [11] Singh, S., Chana, I., Buyya, R., "Building and Offering Aneka-based Agriculture as a Cloud and Big Data Service", *Big Data: Principles and Paradigms*, pp. 1–25. Elsevier 2016.
- [12] Nguyen, Nguyen Cong, et al., "Resource management in cloud networking using economic analysis and pricing models: A survey." *IEEE Communications Surveys & Tutorials* (2017).
- [13] Yousafzai, A., Gani, A., Noor, R. M., Sookhak, M., Talebian, H., Shiraz, M., Khan, M. K., "Cloud resource allocation schemes: Review, taxonomy, and opportunities", *Knowledge and Information Systems*, 50(2), pp. 347-381, 2017.
- [14] Kumar, C. Ashok, R. Vimala, KR Aravind Britto, S. Sathya Devi, "FDLA: Fractional dragonfly based load balancing algorithm in cluster cloud model." *Cluster Computing* 22, No. 1, pp. 1401-1414, 2019.
- [15] Basu, Sayantani, G. Kannayaram, Somula Ramasubbareddy, C. Venkatasubbaiah, "Improved Genetic Algorithm for Monitoring of Virtual Machines in Cloud Environment." In

- Smart Intelligent Computing and Applications, pp. 319-326. Springer, Singapore, 2019.
- [16] Bhandari, Anmol, Kiranbir Kaur, "An Enhanced Post-migration Algorithm for Dynamic Load Balancing in Cloud Computing Environment." In *Proceedings of International Ethical Hacking Conference 2018*, pp. 59-73. Springer, Singapore, 2019.