IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

An Analysis of Virtual Machine Scheduling Algorithms in Cloud Computing

Wakekar Anil Laxman[1] and Dr. Arvind Kumar Bhardwaj[2]
Research Scholar, Department of Computer Science and Engineering[1]
Professor, Department of Computer Science and Engineering[2]
Sunrise University, Alwar, Rajasthan, India

Abstract: *Cloud computing is a paradigm that facilitates instantaneous, convenient, and ubiquitous network connectivity to a repository of configurable computing resources, including but not limited to servers, storage, applications, and services. As evidenced by the fact that these resources can be provisioned and released promptly with minimal management effort or service provider interaction, the use of cloud computing is currently expanding at an accelerated rate. Thus, achieving a balance between the cloud and its resources in order to provide improved performance and services to the cloud's end users while simultaneously ensuring that the majority of users are served by application deployments in the cloud provider's environment has become an essential concern. Load balancing in cloud computing entails distributing the workload across three critical phases of request processing. This consists of task scheduling and virtual machine selection at the designated data center. This paper is predominately concerned with the algorithms and techniques that are accessible for the administration of virtual machines. In addition, it provides a transparent view of their attributes for addressing accumulated burden in an effective virtual machine management system*

Keywords: Virtual Machine Scheduling, Cloud Computing

I. INTRODUCTION

Cloud computing is a service that operates on demand, providing clients with shared resources, software, hardware, and other devices at the precise moment they require them. It can be conceptualized as a solution in which data storage and processing occur in a manner that prevents the user from identifying the particular computer that is transporting them. Cloud computing guarantees access to virtualized IT resources that are hosted and shared outside of a data center. Cloud computing is frequently classified into three distinct categories:

Infrastructure as a service (IaaS)

It offers adaptable methods for generating, utilizing, and overseeing virtual devices. As services, IaaS models provision computing resources including storage, networks, and computation resources. The capability to deploy and execute arbitrary software, including operating systems and applications, is granted to consumers. The responsibility of managing and controlling the virtual infrastructure, which is typically constructed using virtual machines hosted by the IaaS vendor, rests with the consumer. While this thesis primarily centers on this particular model, its findings may also be extrapolated and applied to other models.

Platform as a service (PaaS)

Platform as a Service (PaaS) offers comprehensive infrastructure for developing applications and services, eliminating the need for software installation or download. It prioritizes the provision of advanced functionalities that extend beyond mere virtual machines, which are essential for supporting applications. Cloud providers deliver a computing platform and/or solution stacks under the PaaS model [5]. These stacks typically consist of a database, programming language execution environment, operating system, and web server. Application developers have the ability to create and execute their software on a cloud platform without the need to oversee or regulate the foundational hardware and

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

software layers—such as the network, servers, operating systems, or storage. However, they retain authority over the deployed applications and potentially configure the application-hosting environment.

Software as a service (SaaS)

In SaaS, the user utilizes various software applications hosted on separate servers via the Internet. The application is responsible for delivering commercial value to users. Software applications are provided as services that operate on infrastructure that is under the management of the SaaS provider. The service provider maintains the infrastructure online while performing all necessary patches and enhancements. Services are typically made available to consumers via a variety of clients, including programming interfaces and web browsers, for which they are charged on a subscription basis [6]. Consumers have no visibility into the implementation or the underlying cloud infrastructure on which it is hosted. Web content management, customer resource management (CRM), video conferencing, IT service management, and accounting are a few examples.

Deployment Models in Cloud Computing:

Optional components of the cloud computing deployment paradigm that specify where software is executed are as follows: Two primary stakeholders can be discerned in a cloud provisioning scenario according to the categorization of cloud services as SaaS, PaaS, and IaaS: the Infrastructure Provider (IP), which furnishes infrastructure resources including Virtual Machines, networks, storage, and more; these resources are utilised by Service Providers (SPs) to deliver end-user services, including SaaS, to their customers; and the potential development of these services utilising PaaS tools. The following are the four primary categories of cloud scenarios, as delineated in reference [7].

- **Private cloud:** A private cloud is a collection of standardized computational resources that are reserved for a specific organization and are typically located within the data centers of that organization. Private cloud-based services (PaaS) enable organizations to manage data and processes internally, unencumbered by legal obligations, network bandwidth limitations, or security vulnerabilities. It operates in conjunction with the existing capital investment and empowers the new function as a service.
- **Cloud Bursting:** Private clouds may transfer capacity to alternative IP addresses during periods of heavy workload or for other purposes, such as scheduled maintenance of internal servers.
- **Federated Cloud:** Federated clouds are collaborative clouds that share burden, allowing them to transfer capacity to one another in a manner analogous to the way electricity providers swap capacity.
- **Multiple clouds:** The SP assumes the additional responsibility of managing the intricacy associated with coordinating the service across numerous external IP addresses in multi-cloud scenarios. This includes tasks such as service planning, initiation, and monitoring.

Parameters of interest for cloud services Provider

As with any other performance monitoring metric, monitoring the utilization of physical server infrastructure is critical for the cloud, given that these services comprise the cloud.

Infrastructure response time (IRT): IRT gives the clear picture of the overall performance of the cloud as it checks the time taken for each transaction to complete.

Virtualization metrics: Analogous to the physical machine, it is necessary to gather data on the utilization of resources by the virtual machines. This feature offers insight into the extent to which the virtual machine is being utilized, which facilitates the allocation of resources based on application demands and scale requirements.

Transaction matrices: It is possible to classify it as a derivative of IRT. By utilizing metrics such as success percentage and transaction counts, one can obtain a comprehensive understanding of an application's efficacy in a specific cloud moment.

Cloud computing benefits from numerous advantageous features offered by virtualization technology, including support for suspend/resume operations, consolidation, isolation, and migration. A virtual machine (VM) is a software implementation of a computing environment that permits the installation and execution of an operating system (OS) or program. Critical variables associated with virtual devices include The quantity of virtual machines that applications utilize, Duration required to create a new VM, time required to allocate additional resources to a virtual machine and

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

time required to migrate an application from one VM to another. Virtualization is defined as "something that is not real," yet it provides every advantage of the tangible. The process of virtualization involves the development of an indistinguishable implementation of an object, be it a storage device, operating system, or network resources.

The allocation of fundamental processing units within a computing environment has historically constituted a critical concern [1]. Virtual machines (VMs) must be scheduled in the cloud, similar to any other processing unit, to optimize utilization. Perform the task more quickly, use less energy, and reserve resources (allocation) easily. Elasticity, as it pertains to virtual machines (VMs) in cloud computing, is the capacity of a system to dynamically adapt to changes by automatically provisioning and deprovisioning resources to ensure that the current demand is closely matched with the available resources at all times.

The number of cloud users and, it would appear, the scheduling of virtual machines have increased exponentially. Analysis of the cloud becomes a crucial concern. A user in cloud computing might necessitate a collection of virtual machines that collaborate in order to complete a specific mission. Historically, task interrelationships have not been taken into account. The allocation of system resources to virtual machine flows is accomplished through the scheduling process.

Vm Scheduling Algorithm

In distributed systems, the objective of scheduling algorithms is to distribute processor burden, maximize utilization, and minimize total task execution time. Job scheduling, a renowned optimization problem, is of utmost importance in enhancing the flexibility and dependability of systems. The primary objective is to allocate tasks to flexible resources in accordance with flexible time, which necessitates determining an appropriate job execution sequence that satisfies transaction logic constraints [2].

The two primary classifications of scheduling algorithms exist. The static scheduling algorithm comes first, followed by the dynamic scheduling algorithm. Every individual possesses a unique set of strengths and weaknesses. While dynamic scheduling algorithms exhibit superior efficacy in comparison to static algorithms, they come with a substantial overhead. Load balancing decisions are predicated on the present condition of the systems. It is more effective than a static approach. Static algorithms are predominantly effective in stable and homogeneous environments, where they are capable of generating excellent results. It is not contingent upon the present condition of the system. Prior understanding of the system is required. Nevertheless, they are typically rigid and unable to accommodate dynamic attribute changes that occur during execution. Dynamic algorithms exhibit greater flexibility by accounting for various categories of system attributes both before and during execution. These algorithms are capable of adjusting to dynamic and heterogeneous environments and producing superior results. Nonetheless, in terms of distribution attributes, complexity and dynamism increase. Consequently, certain algorithms may exhibit inefficiency and generate unnecessary overhead, thereby leading to a comprehensive deterioration in the performance of the services.

Gang scheduling Algorithm

C. Reddy [7] describes the application of the gang scheduling algorithm in cloud computing, which is tasked with determining the most appropriate resources for task execution by considering static and dynamic parameters as well as constraints of the virtual machine. Gang scheduling is a parallel system scheduling algorithm that arranges for related virtual machines (VMs) to operate concurrently on distinct machines. Gang Scheduling is a time-sharing algorithm that has been implemented in parallel and distributed systems due to its efficiency. Gang scheduling can be implemented efficiently and cost-effectively in a Cloud Computing environment. Gang scheduling, which permits the scheduling of such virtual machines, is a subset of task scheduling. Gang scheduling is a specific instance of parallel job scheduling in which duties of occupations require frequent communication. Gang scheduling entails a substantial amount of overhead due to the necessity of saving and restoring network status when transitioning between duties.

Moschakis et al. [8] present an enhanced iteration of gang scheduling along with an assessment of its efficacy and cost. The primary objective of the scheduling methods incorporated in the scheduler is to reduce superfluous delays in order to achieve improved response times and diminished slowdowns. Additionally, the scheduler must consider the lease time costs of virtual machines in an effort to achieve a higher cost-to-performance ratio [8]. Performance and expense are considered in the study's integration of mechanisms for job migration and job depletion management. The quantity

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

of Virtual Machines (VMs) that are accessible at any given time fluctuates and adjusts in response to the requirements of the tasks being executed. The findings demonstrate that this scheduling strategy can be implemented efficiently in the cloud. The number of virtual machines (VMs) that are accessible at any given time is scaled dynamically in response to the workload being executed. To accomplish this, they utilized the "Shortest Queue First" (SQF) algorithm, which distributes assignments to virtual machines (VMs) that have the shortest queue.

The investigation into gang scheduling has demonstrated the capacity of time sharing to increase output [17]. In addition to being generally costly, migrating virtual machines to a new data center does not eradicate malnutrition. Gang scheduling can be implemented efficiently and cost-effectively in a cloud computing environment.

The Response Time Rj denotes the duration in seconds between the job j's arrival and departure. The definition of its mean is [8]:

$$R_j = \sum_{J=1}^{n} (Rj / n)$$

Response Time

Where n is the total number of jobs.

Round Robin algorithm

Virtual Machines are distributed to physical hardware in round robin scheduling in a first-in, first-out (FIFO) fashion, with a time-slice or quantum of CPU time [6]. When a process fails to reach its time quantum, the execution of the corresponding virtual machine is preempted and transferred to the subsequent virtual machine in a queue. Following this, the preempted process is relegated to the end of the ready list. In general, a time quantum ranges between 100 and 1000 milliseconds. Consequently, the RR algorithm permits the initial virtual machine (VM) in the queue to operate until its time quantum terminates, at which point it will transition to the subsequent VM in the queue for the same time quantum. By nature, the RR algorithm is preemptive. RR is among the most effective scheduling algorithms that numerous researchers have developed.

The RR algorithm, also known as maximum throughput scheduling, is a proportionally fair algorithm. One of the primary strengths of this algorithm is its balanced utilization of all available resources. The scheduler advances from one node to the subsequent node once a virtual machine has been allocated to that node. This process is iterated until at least one virtual machine has been assigned to each node, at which point the scheduler resumes at the initial node. Therefore, the scheduler proceeds to the next node without waiting for the resources of the current node to be depleted [6]; this is considered fault tolerance.

Genetic algorithm

Virtual Machines are distributed to physical hardware in round robin scheduling in a first-in, first-out (FIFO) fashion, with a time-slice or quantum of CPU time [6]. When a process fails to reach its time quantum, the execution of the corresponding virtual machine is preempted and transferred to the subsequent virtual machine in a queue. Following this, the preempted process is relegated to the end of the ready list. In general, a time quantum ranges between 100 and 1000 milliseconds. Consequently, the RR algorithm permits the initial virtual machine (VM) in the queue to operate until its time quantum terminates, at which point it will transition to the subsequent VM in the queue for the same time quantum. By nature, the RR algorithm is preemptive. RR is among the most effective scheduling algorithms that numerous researchers have developed.

The RR algorithm, also known as maximum throughput scheduling, is a proportionally fair algorithm. One of the primary strengths of this algorithm is its balanced utilization of all available resources. The scheduler advances from one node to the subsequent node once a virtual machine has been allocated to that node. This process is iterated until at least one virtual machine has been assigned to each node, at which point the scheduler resumes at the initial node. Therefore, the scheduler proceeds to the next node without waiting for the resources of the current node to be depleted [6]; this is considered fault tolerance.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

Content-Based Virtual Machine Scheduling Algorithm

The objective of the content-based VM scheduling algorithms was to reduce the volume of data transmitted between racks in the data center during the copying of disk images for virtual machines to the host node [5]. The algorithm returns the node and virtual machine (VM) that has the most similar content to the one that was selected. We search for potential hosts that have VMs with comparable content to the one being scheduled prior to deploying the virtual machine. Subsequently, the host hosting the virtual machine (VM) with the greatest number of disk blocks that are identical to those in the VM under consideration is chosen.

After selecting the host node, the discrepancy between the newly created virtual machine (VM) and the existing VMs on the host is computed. Subsequently, solely the discrepancy is transferred to the destination host. Ultimately, the new virtual machine (VM) can be reconstructed at the destination host using the transmitted difference and the contents of local VMs. An algorithm for scheduling content-based virtual machines (VMs) in a cloud data center that can substantially reduce the network traffic required to migrate VMs from storage racks to host racks.

Adaptive Algorithm

K. Kumar [6] proposed an adaptive algorithm that schedules and assigns virtual machines (VMs) in accordance with the dynamic priority of nodes. The arrangement of virtual machines (VMs) on the nodes is dynamically determined by the priority values, which in turn affect the load factor. The determination of a node's priority is contingent upon its load factor and capacity. This algorithm achieves an optimal trade-off between power efficiency and performance by recalculating the priorities of virtual machines as they are allocated to nodes. The concept of dynamic priority allows for more efficient use of available resources. Adaptive algorithms are effective when it comes to determining the anticipated response time of individual virtual machines. It increases the system's throughput, bandwidth utilization, and probability of outages.

Priority scheduling algorithm

Priority is designated to each Virtual Machine, and only those with that priority are permitted to execute. Instances with equal priority are scheduled using the FCFS method. Priorities are determined by attributes of virtual machines (VMs), including burden magnitude, anticipated execution time, and user-assigned priority. Internally defined priorities determine the priority of a virtual machine using measurable quantities or qualities. Using the concept of aging, a previously designated priority for a virtual machine (VM) can be dynamically modified; in this case, the VM's priority increases in proportion to the total time it spends in the available queue awaiting execution. In the event that the priority of one VM surpasses that of another VM executing on physical hardware, the executing VM preempts the VM with the higher priority. A VM is also preempted from physical hardware when it is created, migrated, or transferred to a system with a higher priority than the VM executing on the hardware.

In their paper, Vignesh V et al. [13] suggested an enhanced algorithm for priority scheduling by incorporating the SJF policy. Strict-Job-First (SJF) policies are implemented as an exception to the general priority scheduling algorithm. Simply put, an SJF algorithm is a priority algorithm in which the priority is notated as the inverse of the subsequent CPU surge. Thus, priority decreases with the duration of the CPU surge, and vice versa [13].

It has a minimal turnover time, a high throughput, and high processor utilization. Priority may be modified in accordance with an entity's age or execution record.

Efficient Resource Utilization Algorithm

Utilization of the vast reservoir of resources is described by R. Nivethitha [9] in terms of the pay-as-you-use policy. The resources are delivered on demand by the cloud by utilizing network resources across various conditions. Users will incur charges for the efficient utilization of resources in accordance with their consumption. The individual suggested an algorithm known as the "Effective Resource Utilization Algorithm (ERUA)" that operates on a three-tier cloud architecture (Consumer, Service Provider, and Resource Provider). This architecture provides advantages to both the service provider and the user (QoS and cost, respectively) by reallocating schedules in an efficient manner according to utilization ratios, thereby improving resource utilization.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

An evaluation of the performance of current scheduling methods reveals that the effective resource utilization algorithm generates a schedule that is more optimized and increases the rate of efficiency [9]. In order to serve customers, the service provider dynamically generates Virtual Machine (VM) instances from resources it acquires from the resource provider.

Renewable Energy Source provisioned algorithm

In their article, D. Hatzopoulos et al. [11] investigate the issue of virtual machine (VM) allocation in a network of geographically dispersed cloud server facilities. He examines the issue pertaining to the system's energy-efficient allocation. The primary aim is to decrease the operator's overall power consumption expenses. Every instance of a task requiring execution in the cloud is accompanied by a virtual machine (VM) request that specifies the necessary resources and a deadline for its completion. The cloud provider must generate a virtual machine (VM) that meets the request's resource specifications and run the VM prior to the specified deadline. An online algorithm is suggested, featuring a look-ahead horizon, in which the grid power price and the output power pattern of the RES are predetermined.

The implementation of algorithm applications powered by renewable energy sources substantially reduces the amount of bandwidth resource wasted. This methodology and design the metering system in order to monitor the conditions of the resource. It establishes the utmost throughput for requests.

II. CONCLUSION AND FUTURE WORK

We have analyzed and surveyed numerous algorithms for effective load balancing and virtual machine management in this paper. While each algorithm possesses its own advantages and disadvantages, it also provides distinct scenarios in which they are most applicable. Additionally, it provides an understanding of the efficacy and diverse performance attributes of these algorithms. These algorithms provide guidance on the parameters that ought to be considered when choosing a virtual machine (VM). As a future endeavor, we are attempting to expand the functionality of VM to include multimedia applications, an emergent technology that intrigues a great number of users. By implementing an effective VMM for the cloud computing environment, multimedia application services can be efficiently extended to a large number of users and managed.

REFERENCES

[1]. Hadi Salimi , “Advantages, Challenges and Optimizations of Virtual Machine Scheduling in Cloud Computing Environments” in International Journal of Computer Theory and Engineering Vol. 4, No. 2, April 2012.

[2]. Pinal Salot , “A Survey Of Various Scheduling Algorithm In Cloud Computing Environment” in M.E, Computer Engineering, Alpha College of Engineering, Gujarat, India , Volume: 2 Issue: 2.

[3]. MR.NISHANT, “Pre-Emptable Shortest Job Next Scheduling In Private Cloud Computing” in journal of information, knowledge and research computer engineering, NOV 12 TO OCT 13 | VOLUME – 02, ISSUE – 02.

[4]. TARUN GOYAL,“Host Scheduling Algorithm Using Genetic Algorithm In Cloud Computing Environment”,

[5]. International Journal of Research in Engineering & Technology (IJRET) Vol. 1, Issue 1, June 2013.

[6]. Sobir Bazarbayev,” Content-Based Scheduling of Virtual Machines (VMs) in the Cloud” in University of

[7]. Illinois at Urbana-Champaign, AT&T Labs Research.

[8]. Kiran Kumar et. al., “An Adaptive Algorithm For Dynamic Priority Based Virtual Machine Scheduling In Cloud” in IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012.

[9]. Dr. Chenna Reddy , “An Efficient Profit-based Job Scheduling Strategy for Service Providers in Cloud Computing Systems” in International Journal of Application or Innovation in Engineering & Management (IJAIEM) , Volume 2, Issue 1, January 2013.

[10]. Ramkumar N, Nivethitha , “Efficient Resource Utilization Algorithm (ERUA) for Service Request Scheduling in Cloud” in International Journal of Engineering and Technology (IJET) , Vol 5 No 2 Apr-May 2013.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, September 2024

[11]. Dimitris Hatzopoulos, "Dynamic Virtual Machine Allocation in Cloud Server Facility Systems with Renewable Energy Sources" at IEEE International Conference on Communications (ICC) 2013, Budapest, Hungary.

[12]. Vignesh V, Sendhil Kumar KS, Jaisankar N , " Resource management and scheduling in cloud environment"

[13]. in International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013.

[14]. Jeongseob Ahn, Changdae Kim, Jaeung Han," Dynamic Virtual Machine Scheduling in Clouds for

[15]. Architectural Shared Resources".

[16]. Tommaso Cucinotta," Providing Performance Guarantees to Virtual Machines using Real-Time Scheduling" in Tommaso Cucinotta, Dhaval Giani, Dario Faggioli, and Fabio Checconi Scuola Superiore Sant'Anna, Pisa, Italy.

[17]. Junliang Chen,Bing Bing Zhou, "Throughput Enhancement through Selective Time Sharing and Dynamic Grouping" in 2013 IEEE 27th International Symposium on Parallel and Distributed Processing.

[18]. Manoranjan Dash , " Cost Effective Selection of Data Center in Cloud Environment" in ISSN, Volume-2, Issue-1, 20131.

[19]. Abirami S.P., Shalini Ramanathan (2012), "Linear Scheduling Strategy for Resource allocation in Cloud Environment", International Journal on Cloud Computing and Architecture, vol.2, No.1, February.

[20]. Nilabja Roy, Abhishek Dubey and Aniruddha Gokhale, "Efficient Autoscaling in the Cloud using Predictive Models for Workload Forecasting".

[21]. Soramichi Akiyama, Takahiro Hirofuchi, Ryousei Takano, Shinichi Honiden (2012), "MiyakoDori: A Memory Reusing Mechanism for Dynamic VM Consolidation", Fifth International Conference on Cloud Computing, IEEE 2012.

[22]. V. Venkatesa Kumar et. al , "Job Scheduling Using Fuzzy Neural Network Algorithm in Cloud Environment", International Journal of Man Machine Interface, Vol. 2, No. 1, March 2012.