International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 5 ǁ May 2023 ǁ PP. 211-217

Virtual Machine Consolidation for Stochastic Load Balancing In Cloud Data Center Management

DEEKSHITHA DAKSHINAMURTHY, SRIMATHI DHARMARAJ, SEETHA KUMAR

Computer Science and Engineering, EGS Pillay Engineering College, 611108, Nagapattinam
Computer Science and Engineering, EGS Pillay Engineering College, 611108, Nagapattinam
Computer Science and Engineering, EGS Pillay Engineering College, 611108, Nagapattinam
This work was supported in part by the Department of Computer Science and Engineering (CSE)

ABSTRACT*- Cloud computing can manage a big volume of expanding work in a planned manner for the advantage of corporate clients. Virtualization is a critical enabling technology for cloud computing since it generalizes infrastructure and makes it easier to maintain. In this proposed work, the virtualization method is used to assign cloud assets based on their needs and to facilitate the computing notion. The "skewness" approach is used here to mix workloads and maximize server utilization by minimizing the equal. Managing consumer requests for resource allocation produces difficult on-call resource allocation situations. It has been engaged for resource provision in order to enforce the Virtual Machine (VM) allocation technique. It is envisaged that the adoption of virtualized environments will minimize typical task response time while also executing tasks in accordance with cloud resource availability. As a result, consumers are given VMs based on task characteristics. The VM and PM (Physical Machine) mapping is achievable while the execution is running, owing to the VM migration technology. The effective and dynamic utilization of cloud sources can aid in weight stabilization and the avoidance of issues such as slow structure run. In order to offer virtual space when multiple requests are made at the same time, this strategy can use a local negotiation-based VM consolidation method to anticipate each task request and reduce overloads. The proposed system uses a co-location method to combine underutilized small rooms to produce more virtual space, hence enhancing server performance. Implement a self-destruction technique based on the time-to-live characteristic to delete incorrect data. The proposed system operates in real time and efficiently allocates resources. In this approach, first build a prediction model that can estimate the partition sizes of reduce commitments at runtime. Furthermore, after dynamically recognizing information skewness, giving higher belongings to reducers with big partitions to aid in their completion.*

Index Terms—*VM resource allocation, VM monitoring, VM migration, Self destruction approach.*

Date of Submission: 26-04-2023 Date of acceptance: 06-05-2023

I. INTRODUCTION

Standard cloud computing techniques such as virtualization, parallel processing, and distributed database and storage have experienced substantial progress and are now extensively employed in a variety of situations due to the rapid rise of cloud computing on a global scale. As one of the main components of cloud computing architecture, virtualization plays a critical role in providing dependable cloud computing services. With the help of a few simulated virtual machines (VMs) grown on a cluster of high-performance network servers, virtualization is a fundamental technique that can be used to recognize the quick deployment, dynamic allocation, and move area management of IT resources. Through these virtual machines, customers can access on-demand services.[5] The variety and workloads of virtual machines change frequently due to the fact that customer preferences are continuously changing, which, incidentally, results in the creation of a new project for source scheduling and migrations of digital machines. According to a unique component of the digital tool migration strategy, choosing the deliver host device and destination host is the most important step in virtual device migrations.

A few researchers propose the so-called "live migration approach for digital machines" to assist in choosing the deliver and destination host computers as well as to prevent needless migrations of digital systems caused by transiently high workload values. There are currently two different digital device migration methodologies that can be found in the literature: one uses the better and lower thresholds of the host device to manage resource usage, and the other uses the workload threshold of the host tool to anticipate the pattern of its upcoming workloads. The problem of aggregation conflict that arises in conventional workload balancing techniques cannot be solved by the prior solution, even though it can handle the issue of aid waste brought on by

the static workload balancing method's resource. The latter technique, on the other hand, can resolve the issue of "false alarm" virtual device migrations caused by some brief height workload values, but it ignores the uncertainty and stochastic character of workload values, as well as the aggregate of each, on host systems.

The proposed work suggests a new virtual device migration strategy that is based on time series prediction in cloud precept in order to move closer to incorporating the uncertainty and randomness of workload values into the decision-making process of digital system migrations, and thereby resulting in a better migration method. The basic operation of this method is as follows: it first establishes upper and lower workload thresholds for host machines, uses the cloud concept to forecast the host machine's future workload trend, and then specifies a migration desire criterion, which is then used to choose the supply host, destination host device, and digital machine to carry out the desired Migration [12]. The uncertainty, fuzzyness, and randomness of workload values are addressed by this suggested migration method. It also converts qualitative to quantitative concepts and vice versa, removes the aggravation battle headache brought on by digital machine migrations because of a few transient and short-term top workload values, and helps to achieve dynamic source balancing. A website hosting provider often owns and controls the physical environment for cloud storage, which is defined as information storage in which data is saved in logical pools, physical storage spans a few servers (and frequently locales), and physical storage is spread across a few servers [3,9]. The supplier vendors are in responsibility of maintaining the physical environment's security as well as the user's access to the preserved data. To safeguard consumer, corporate, or software programmed data, people and organizations rent or purchase garage space from suppliers.

A co-located cloud system, an internet provider application programming interface (API), or API-enabled software, such as a cloud storage server, a garage gateway, or Web-based content material cloth control structures, can access cloud storage services. Cloud storage is built on virtualized infrastructure and is comparable to general cloud computing in terms of accessible interfaces, nearly instant elasticity and scalability, multi-tenancy, and metered assets. Cloud storage can be installed locally or accessed from an external source (such as Amazon S3).

Although the phrase "cloud storage" is most generally associated with a hosted object storage provider, it has evolved to encompass other types of data storage that are now available as a service, such as block storage. Garages that can be hosted and deployed using cloud storage characteristics include Amazon S3 and Microsoft Azure Storage, object storage software such as Openstack Swift, object storage systems such as EMC Atmos, EMC ECS, and Hitachi Content Platform, and distributed storage research initiatives such as OceanStore and VISION Cloud[12].

A list of cloud storage services is shown below:

• A federated or cooperative garage cloud structure is composed of several assigned assets that serve as a single entity.
• Due to redundancy and statistical dispersion, it is very fault resistant.
• Extremely long-lasting because versioned copies are used
• When it comes to data replication, they are usually consistent sooner or later.

II. RELATED WORK

Torre,, et.al,…[1] The effective tradeoff between resource wastage and overcommitment is a challenging task in virtualized Clouds and depends on the allocation of virtual machines (VMs) to physical resources. We propose in this paper a multi-objective method for dynamic VM placement, which exploits live migration mechanisms to simultaneously optimize the resource wastage, overcommitment ratio and migration energy. Our optimization algorithm uses a novel evolutionary meta-heuristic based on an island population model to approximate the Pareto optimal set of VM placements with good accuracy and diversity. a multi-objective method and 106 algorithm for dynamic placement of VMs in response to their fluctuating 107 resource demands. Our goal is to minimise the energy consumption in data 108 centres faced with dynamic workloads by dynamically allocating VMs to 109 the minimum number of PMs using a three-fold strategy is reduce the 110 number of PMs by increasing the overcommitment, analyse the effects 111 of the overcommitment and overly reduced number of PMs on the QoS, analyse the effects on live migration, ignored in related work

Alharbi et.Al,…[2] In this paper, we formulate placement of VMs to PMs in a data center as a constrained combinatorial optimization problem and make use of the information from PM and VM profiles to minimize the total energy consumption of all active PMs. An Ant Colony System (ACS) embedded with new heuristics is presented for an energy-efficient solution to the optimization problem. To demonstrate the effectiveness of the ACS, simulation experiments are conducted on small-, medium- and large-scale data centers. The results from our ACS are compared with two existing ACS methods as well as the widely used

First-Fit-Decreasing (FFD) algorithm. To solve the formulated problem, it was 104 proposed to integrate the minimum cut into the Best Fit algorithm to achieve 105 a reduced number of PMs and network elements.

Zhang, et.al,…[3] This paper proposes a novel and effective evolutionary approach for VM allocation that can maximize the energy efficiency of a cloud data center while incorporating more reserved VMs. Aiming at accurate energy consumption estimation, our approach needs to simulate all the VM allocation updates, which is time-consuming using traditional cloud simulators. To overcome this, we have designed a simplified simulation engine for CloudSim that can accelerate the process of our evolutionary approach. Comprehensive experimental results obtained from both simulation on CloudSim and real cloud environments show that our approach not only can quickly achieve an optimized allocation solution for a batch of reserved VMs, but also can consolidate more VMs with fewer physical machines to achieve better energy efficiency than existing methods

Moges,, et.Al,…[4] Several researches have been conducted on Virtual Machine(VM) consolidation to optimize energy consumption. Among the proposed VM consolidations, OpenStack Neat is notable for its practicality. OpenStack Neat is an open-source consolidation framework that can seamlessly integrate to OpenStack, one of the most common and widely used opensource cloud management tool. The framework has components for deciding when to migrate VMs and for selecting suitable hosts for the VMs (VM placement). The VM placement algorithm of OpenStack Neat is called Modified Best-Fit Decreasing (MBFD). MBFD is based on a heuristic that handles only minimizing the number of servers. The heuristic is not only less energy efficient but also increases Service Level Agreement (SLA) violation and consequently cause more VM migrations. To improve the energy efficiency, we propose VM placement algorithms based on both bin-packing heuristics and servers' power efficiency. In addition, we introduce a new bin-packing heuristic called a Medium-Fit (MF) to reduce SLA violation

Zhou,, et.Al,…[5] The paper's main goal is to decrease energy consumption and ensure the high QoS within cloud data centers (CDCs). To accomplish this, we propose an energy-efficient VM allocation and deployment algorithm based on an adaptive energy-aware framework. Unlike other energy-aware algorithms that only consider energy consumption due to VM deployment's, the proposed algorithm considers VM provision's energy efficiency during VM allocation and deployment. The proposed algorithm can effectively deal with variable load and maintain low energy consumption and SLA violation. All in all, our main contributions can be summarised as: Proposal of the adaptive four thresholds energy-aware framework that can effectively address the variable load named AFED

III. EXISTING SYSTEM

VM MIGRATION TECHNIQUES:

Genetic algorithm based resource allocation:

Unbalance in venture problems can be resolved by solving the scheduling problem in resource allocation; in this situation, a parallel genetic algorithm is used, which is significantly faster than a conventional genetic algorithm. GA is a great option for resolving scheduling challenges because of its advanced asset utilization guidance when VMs are assigned. Genetic algorithms (GA) are effective strategies that can be applied to solve challenging issues in a range of industries. In order to boost performance and scalability, parallel genetic algorithms (PGAs) are genetic algorithms that are implemented in parallel. PGAs are straightforward to construct on parallel mainframes or networks of heterogeneous computer architectures. [13]. Due to the rate at which the enormous allocation series has been acquired with rapid convergence, PGA is intending for them to have cloud sources in a more successful manner. In this instance, every cloud node has a scheduler. Three sports are primarily finished according to the schedule. The machine initially keeps track of idle assets, and the availability of virtual machines is probably updated on a regular basis when new VM requests come in, virtual machines enter shutdown mode, or real assets undergo any changes. Select the allocation collecting sample with the biggest sample size using the PGA. Later, the requested VMs will receive the necessary real machines. When resolving unbalanced project challenges with our PGA, a number of crucial criteria are taken into account. The precise genetic algorithm parameters, such as chromosome example, optimal fitness characteristic design, and the use of the right migration technique, must be kept in mind while allocating the requested sources.

Auction based Models:

Grid computing resource allocation has been studied using the Continuous Double Auction (CDA). One of the most well-known strategies is CDA, and it has long been employed in the virtual stock market. Throughout the public sale, any number of bids from different parties may be made [19]. They contend that the

market-based strategy outperforms non-marketplace alternatives in terms of project utilization and effective resource allocation. The fundamental issue with this strategy is that it only takes into account one workable resource allocation mechanism and is unaware of resource allocation for some sources. In a local grid environment, this approach was used to allocate CPU time. For marketplace-based environments in grid, combinatorial double auction has been proposed as a different resource allocation strategy that added income maximization and economy overall performance. This suggested approach had the benefit of being entirely monetary-based and having lots of flexibility.

1.1 Advantages of VM Migration

Load balancing:

The variation in resource consumption levels between each PM in the cluster is reduced as a result. In the presence of other machines that are only lightly filled and have enough spare capacity, this prevents some machines from getting overloaded. The device can be kept stable by using live migration. The common machine load can be balanced by moving VMs from PMs that are overcrowded to PMs that are under loaded.

Server Consolidation:

In order to limit server sprawl, server consolidation strategies are required at Carrier Company. These techniques are VM packing techniques that aim to cram as many VMs onto a PM as feasible in order to maximize resource utilization and turn off unnecessary machines. Less energy will be consumed as a result of consolidation, which will cut current operating expenses for data centre administrators.

Hotspot & Coldspot Migration:

Hotspot and cold spot detection is constantly dependent on thresholds that can be set using a service provider's resource or based on the SLAs offered by cloud customers. Due to the top threshold, a higher useful resource utilization rate at the maximum is typically prepared, whereas the drop threshold typically prepares a fully low useful resource usage rate. PMs with aid utilization levels above the upper criterion are known as hotspots, whereas those with usage levels below the lower criterion are known as cold spots [15]. Both are true regardless of resource length. The former suggests excessive use, whereas the latter denotes insufficient use.

- **Cold Migration**

Transfer a previously turned off or suspended digital machine to today's host. Consumers have the option of moving configuration and disc documentation for powered off or suspended digital machines to new garage locations at their leisure. Bloodless migration can also be utilized to transfer digital equipment from one digital transfer to another, as well as from one record to another. The user has the option of performing bloodless migration manually or arranging a challenge.

- **Hot Migration**

Move an active virtual records garage to a new server. Customers have the option of flowing into the digital machine discs or folders to a superb statistics shop. "Hot migration" relates to the application of another phrase, such as "live VM migration" or "vMotion." People can use vMotion to transfer their digital devices without harming their availability.

IV. PROPOSED FRAMEWORK USING VM MIGRATION WITH SELF-MONITORING APPROACH

In recent years, the exponential expansion of data in numerous application industries such as e-commerce, social networking, and scientific computing has produced a massive demand for large-scale data processing. As a parallel computing architecture, the VM consolidation method has recently gained a lot of traction in this context. Each request runs a user-specified function on a block of input data to produce an interim scheduling phase. Then, using a user-specified reduce function, each job compiles intermediate requests and generates the final result [12]. A skewed distribution of task decrease can have negative consequences. In this proposed system, a novel resource allocation technique was developed to effectively reduce system overload while minimizing the number of servers used. In order to describe how a server is used unevenly, introduce the concept of "skewness." Data centre may, however, employ these capabilities to merely accept more virtual machines (VMs) than they have physical resources for. This circumstance is frequently referred to as "resource overbooking" or "resource over commitment." The total requested capacity is smaller than the whole available capacity in the resource management process [16]. This is a tried-and-true technique for managing precious and rare resources that has been applied for a very long period in many different fields. In cloud environments, the best method for increasing machine usage is to overload cloud resources. The

application may identify a rising trend in resource usage habits and support location negotiation in order to significantly lower placement churn. The main goal of the initiative is to:

Overload Avoidance: The capacity of a Physical Machine should be adequate to meet the resource requirements of all VMs running on that Physical Machine. In other words, the Physical Machine is overloaded, which affects its own virtual machines' performance.

Green computing: is defined as optimizing the use of physical machines while yet ensuring that they can support all virtual machines. In order to preserve energy, physical machines that are not in use are routinely shut off.

The proposed work investigates stochastic load balancing via VM migration to address the issue of unpredictable demand and changing workloads. The stochastic load balancing methodology, in contrast to existing approaches, probabilistically characterizes VM resource demand and work load states of Physical Machines with the aim of guaranteeing that aggregate cloud resource consumption on each PM does not, with a high likelihood, exceed its capacity. Each PM is made aware of the risk of SLA violations and the probability is estimated in the SLA agreement. The unpredictability and dynamic fluctuations in resource consumption can be handled by stochastic load balancing. By choosing load balancing, you can ensure that the performance of the resulting application is more resistant to highly dynamic workloads and that resources are efficiently multiplexed statistically [18] with a probabilistic assurance for handling overloads. The estimation of stochastic resource demand, the detection of hotspots, and the performance of VM migrations while capturing multidimensional stochastic resource requirements are just a few of the new challenges that stochastic load balancing brings about.

The VM consolidation technique, which serves as the basis for dynamic resource allocation, is implemented in the proposed work. The VM consolidation technique has gained popularity as a large-scale data processing idea in recent years. Due to current scheduling methods, resource mapping activities' output is dispersed unevenly among several systems. This project introduces the VM consolidation technique, a framework that enables run-time partitioning skew reduction. We address the problem of partitioning skew in the VM consolidation method by altering task run-time resource allocation, in contrast to previous approaches that attempted to balance the workload of reducers by repartitioning the intermediate data allocated to each reduce job. We demonstrate how our method can lower the overhead associated with data repartitioning for VM consolidation strategies. There are two major problems with the VM consolidation method that need to be resolved [18]. This proposed study detects partition skew to build a run-time prediction algorithm that predicts the partition size of each reducer. In order to determine the ideal container size for each decreasing job, a task performance model that links task running duration with resource allocation must be created. The partitioning plans upon which the repartitioning techniques are based need the execution of a progressive report each time work is initiated. The partition size prediction can be performed totally online; there is no need to alter how partitioning is implemented in the proposed approach. As a result, we found that our current prediction method is simple but successful in generating outcomes of high caliber. As soon as the necessary resources are available, we can set up a co-located virtual machine. By combining small quantities of idle virtual machine space, we can give users fresh virtual machine space. We can also leverage the time to live function to flush data from the cloud provider using a self-destruction technique. We save personal information in Cloud Storage, which contains details that take up more space when their validity expires. This information is copied and cached by cloud service providers [22]. A self-destruction system's basic purpose is to erase the user's vital data based on the time live property. All data, as well as copies of data, have been destroyed. In this proposed study, present a system that satisfies the integration of active storage techniques. The basic configuration of the VM migration is depicted in Figure 2.

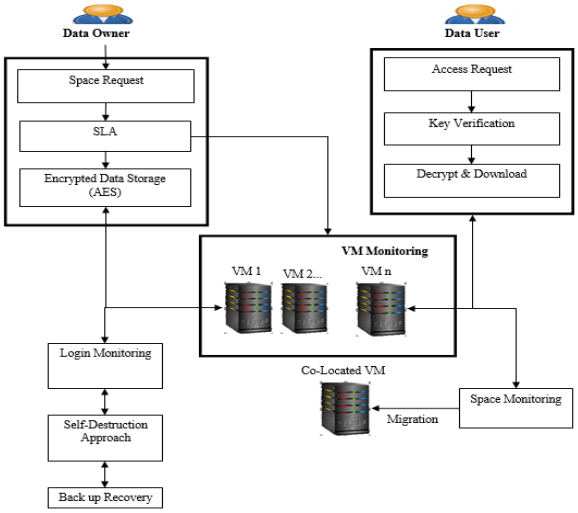

Fig 2: Proposed framework

ALGORITHM IMPLEMENTATION

Instead of total time, the proposed technique is based on projected execution time. As a result, utilizing this strategy to schedule jobs in a cloud environment can result in a shorter make span than using the original set of heuristic criteria. The proposed algorithm is described as follows:

for all tasks Ti in Meta venture Mv
    for all assets Rj
    Cj= Ej+rj
do till all duties in Mv are mapped
    if the number of sources is even then
for each responsibilities in Mv find the earliest
    complete time and the assets that incorporates it
    find the TaksTk to the resources Rk with maximum earliest completion time
    assign assignment Tk to the assets Rk that offers the earliest finishing touch time
delete task Tk from Mv
    updateRk
    updateCij for all I
else
for each tasks in Mv discover the earliest
    complete time and the sources that incorporates it
    locate the TaksTk to the assets Rk with maximum earliest completion time
    assign project Tk to the sources Rk that offers the earliest final touch time
delete assignment Tk from Mv
    updateRk
    updateCij for all I
quit if
cease do

Assume that m resources Rj(j = 1,...,m) are needed to fulfil n obligations Ti(i = 1,..., n). An agenda for each project is the assignment of one or more time periods to one or more assets. The estimated time of execution When Rj has no load when Ti is assigned, Eij of mission Ti on aid Rj is defined as the length of time it takes Rj to complete Ti. The expected completion time Cij of assignment Ti on assistance Rj is defined as the wall-clock time when Rj completes Ti (after having finished any formerly assigned duties).Let bi signify the start of assignment Ti's execution. Cij=bi+Eij, according to the definitions above. Let Ci be the mission's culminating glory time, and Cij be the time when aid Rj is assigned to carry out endeavour Ti. We can compare the device's performance in terms of reaction time, which is calculated as the time between the end of an inquiry or request on a computer device and the commencement of a reaction.

V. CONCLUSION

In the context of cloud computing, a resource allocation system (RAS) is any technique intended to guarantee that the infrastructure of the issuer is used properly to satisfy the programmers' requirements. In addition to providing the developer with this assurance, practical resource allocation mechanisms should also consider the current reputation of each resource in the cloud environment. In doing so, algorithms can be used to more effectively allocate physical and/or virtual resources to developers' applications, reducing the operational costs of the cloud environment. Based only on the necessity for conversion, our technology adaptively multiplexes digital to physical resources. In order to properly utilize server resources, the proposed artworks mix VMs with various support qualities using the Migration technique. The suggested method achieves overload reduction and environmentally friendly computing for systems with several resource constraints.

References

[1]. Zheng, Jing, Qi Li, GuofeiGu, Jiahao Cao, David KY Yau, and Jianping Wu. "RealtimeDDoSdefense using COTS SDN switches via adaptive correlation analysis." IEEE Transactions on Information Forensics and Security 13, no. 7 (2018): 1838-1853.

[2]. Pham, Thi Ngoc Diep, Chai Kiat Yeo, Naoto Yanai, and Toru Fujiwara."Detecting flooding attack and accommodating burst traffic in delay-tolerant networks." IEEE Transactions on Vehicular Technology 67, no. 1 (2017): 795-808.

[3]. Kohnhäuser, Florian, NiklasBüscher, and Stefan Katzenbeisser."A practical attestation protocol for autonomous embedded systems."In 2019 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 263-278.IEEE, 2019.

[4]. Seth, Bijeta, SurjeetDalal, VivekJaglan, Dac-Nhuong Le, Senthilkumar Mohan, and GautamSrivastava."Integrating encryption techniques for secure data storage in the cloud." Transactions on Emerging Telecommunications Technologies 33, no. 4 (2022): e4108.

[5]. Chen, Min, Wei Li, Giancarlo Fortino, YixueHao, Long Hu, and IztokHumar."A dynamic service migration mechanism in edge cognitive computing." ACM Transactions on Internet Technology (TOIT) 19, no. 2 (2019): 1-15.

[6]. Lv, Liang, Yuchao Zhang, Yusen Li, KeXu, Dan Wang, Wendong Wang, Minghui Li, Xuan Cao, and Qingqing Liang. "Communication-aware container placement and reassignment in large-scale internet data centers." IEEE Journal on Selected Areas in Communications 37, no. 3 (2019): 540-555.

[7]. Tao, Ye, PengXu, and Hai Jin. "Secure data sharing and search for cloud-edge-collaborative storage." IEEE Access 8 (2019): 15963-15972.

[8]. Misra, Sudip, and NiloySaha. "Detour: Dynamic task offloading in software-defined fog for IoT applications." IEEE Journal on Selected Areas in Communications 37, no. 5 (2019): 1159-1166.

[9]. Kim, Namkyu, Yunseong Lee, Chunghyun Lee, The Vi Nguyen, and Sungrae Cho. "GPU-specific task offloading in the mobile edge computing network."In 2020 International Conference on Information and Communication Technology Convergence (ICTC), pp. 1874-1876.IEEE, 2020.

[10]. Chen, Jie, L. Ramanathan, and MamounAlazab. "Holistic big data integrated artificial intelligent modeling to improve privacy and security in data management of smart cities." Microprocessors and Microsystems 81 (2021): 103722.