International Journal of Enhanced Research in Management & Computer Applications,
ISSN: 2319-7471, Volume 15 Issue 5, May 2026

Virtual Machine Placement on Cloud Infrastructure

Dr. Ravinder[1], Ashok Kumar Adarsha[2]

[1]Assistant Professor, Department of Computer Science Engineering, Rattan Institute of Technology and Management, Haryana, India
[2]Research Scholar, Department of Computer Science Engineering, Rattan Institute of Technology and Management, Haryana, India

ABSTRACT

Cloud computing has transformed the IT landscape by providing scalable, on-demand infrastructure through virtualization. At the core of this infrastructure is Virtual Machine Placement (VMP), the process of mapping virtual instances onto optimal physical machines (PMs) within massive data centers. Efficient VMP is a critical determinant of operational efficiency, directly impacting resource utilization, hardware costs, and energy consumption. However, balancing multi-dimensional resource constraints (CPU, memory, storage, and network bandwidth) while preventing Service Level Agreement (SLA) violations remains an NP-hard optimization challenge. This paper presents a comprehensive analysis of contemporary VMP strategies, classifying them into heuristic, meta-heuristic, and learning-based approaches. We examine the trade-offs between conflicting objectives—such as minimizing energy footprint via server consolidation versus maximizing performance via load balancing. Finally, we outline open research challenges, including traffic-aware placement and real-time dynamic migration under fluctuating workloads, providing a roadmap for future cloud infrastructure management.

INTRODUCTION

General

Cloud computing is the utility computing that has unlimited virtualized resource to create a custom-built infrastructure or platform to run applications or full part services as pay-as-you-use basis. Cloud computing has modified the paradigm of system deployment [2]. The advanced system implementation are abstracted from the end user by the help of the virtualization techniques, resources are virtualized that offers an illusion of infinitely scalable and universally available system [1] [2]. With the assistance of pay-as-you-use model, infinitely scalable and universally available systems, cloud computing makes the long command dream of utility computing possible. Developers having initiative concepts for providing new Internet services do not need massive expenditure to setup the hardware and software necessities. Cloud computing refers to each application that are provided as a service over the Internet and therefore the underlying hardware infrastructure and the platform over that those applications are developed. The underlying hardware infrastructure is technically called data-center, having massive vary of physical devices starting from personal computers to cluster computers to high end server machines. [3].

A major a part of cloud computing's price proposition and its charm is its ability to convert capital expenses to operational expenses through a usage rating theme that's elastic and might be right sized. The conversion of real assets to virtual ones provides a live of protection against an excessive amount of order client infrastructure. Basically, moving expenses onto the operational expenses facet of budget permits a company to transfer risk to cloud computing supplier.

Resource allocation is a topic that has been addressed in several computing areas, like operating systems, grid computing, and datacenter management. A Resource Allocation System in Cloud Computing are often seen as any mechanism that aims to ensure that the application's needs are attended properly by the provider's infrastructure. In conjunction with this guarantee to the developer, resource allocation mechanisms ought to additionally take into account the present status of every resource within the cloud environment, so as to use algorithms to better allocate physical and/or virtual resources to developers' applications, therefore minimizing the operational cost of the cloud environment.

The rest of the thesis is organized as follows. Chapter 2 gives an overview of cloud computing, which includes the terminologies and technologies used in cloud computing. Chapter 3 describes the works that have been done addressing the issue of virtual machine placement along with the strategies followed by open-source technologies. In Chapter 4 the proposed framework for virtual machine placement as well as the Virtual Machine Scheduler algorithm is described. Chapter

5 presents the simulation and results of the proposed framework. Chapter 6 concludes the discussion and gives a direction to future research in this issue we have discussed.

LITERATURE REVIEW

The literature separates VMP strategies based on their driving objectives. While older studies primarily utilized mono-objective optimization, modern research intensely shifts toward multi-objective frameworks. Power & Energy Efficiency (Green Computing): Server consolidation aims to pack VMs onto the fewest possible active physical hosts. Hypervisors shift underutilized or idle PMs into power-saving low states, drastically reducing data center electricity and cooling costs. Quality of Service (QoS) & SLA Adherence: Packing too many VMs onto a single host causes resource contention (CPU/Memory throttling). The literature measures the success of an algorithm by its ability to eliminate Service Level Agreement (SLA) violations while operating at optimal capacity. Network-Aware Traffic Optimization: Modern datacenters suffer from network congestion. Network-aware VMP algorithms analyze communication dependencies between VMs, placing high-bandwidth interacting VMs physically close to each other (same rack or switch) to slash intra-datacenter latency.

Load Balancing & Resource Completeness: Ensuring multi-dimensional balance (e.g., balancing CPU usage alongside RAM usage) protects physical hosts from running out of one specific resource type while others sit wasted.

An Overview of Cloud Computing

There is plenty of dialogue of what cloud computing is. The US National Institute of Standards and Technologies (NIST) have place a shot in process cloud computing. According to NIST [5] Cloud computing is a model for convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management e ort or service provider interaction.

The above definitions often explained brie y as, network access in on-demand basic and in a very convenient manner along with less e ort from management and less service provider's interaction explains quick and straight forward access for potential resources. With resources in a shared pool, illustrates the supply of computing resources from a cloud service provider are combined in a one massive assortment, for serving all users. The frequent provisioning of resources is employed for quickly matching the active resources, once a necessity comes for those resources. This frequent and quick provisioning prevents a scarcity of computing power once the requirement will increase.

Private Cloud

Private cloud (also referred to as internal cloud or company cloud) is a term for a proprietary computing design that gives hosted services to a restricted variety of individuals behind a firewall. The cloud could be managed by that organization or a third party. Private cloud could also be either on or off premises. This offers organization the price advantages of virtualization.

Para Virtualization:

Para virtualization is a virtualization technique that presents a software system interface to the virtual machines that's almost like but not same as that of the underlying hardware. The intent of this changed interface is to cut back the portion of the guest operational system's execution time that's spent performing operations that are substantially tougher to run in exceedingly virtual surroundings compared to non-virtualized environment surroundings. There are specially defined "hooks" that enable the guest and host to request and acknowledge these tough tasks that will rather be executed within the virtual domain, where execution performance is slower.

Virtual Machine Placement Strategies in Cloud computing

Virtual machine placement is the method of mapping virtual machines to physical machines. In different words, virtual machine placement is the method of choosing the foremost appropriate host for the virtual machine. The method involves categorizing the virtual machines hardware and resources necessities and the anticipated usage of resources and therefore the placement goal. The primary goal may either be increasing the usage of accessible resources or it will be saving of power by having the ability to stop working some servers. The involuntary virtual machine placement algorithms are designed keeping in mind the on top of goals.

The chapter is organized as follows. In section 3.2 categories of Virtual Machine placement algorithm is discussed. In section 3.3 Virtual Machine placement based on constraint programming is discussed. In section 3.4 Virtual Machine placement based on stochastic integer programming is discussed. In section 3.5 Virtual Machine placement based on bin packing is discussed. In section 3.6 Virtual Machine placement based on genetic algorithm is discussed. In section 3.7

strategies followed by open-source solutions are discussed.

Virtual Machine Placement Based on Bin packing Approach
The bin packing approach can be used to realize the particular mapping of virtual machines to available physical machines. It's potential to attenuate the value of running the data-center by tightly packing the Virtual Machines needed to be running at a time onto the smallest number of physical machines possible.

Proposed Scheme
To resolve this problem virtual machine manager sends the virtual machine specifications to the virtual machine scheduler (Virtual Machine Scheduler). But in this model instead of sending that request directly, a binary search tree of the requested virtual machines is created and that is sent to the Virtual Machine Scheduler. This is done to improve the resource utilization, which is explained in the next sub section in the algorithm. Now Virtual Machine scheduler will take the maximum requirement node from that virtual machine tree and will search for a host that will best t the requirement. For implementing the best- t strategy all the physical resources are also logically organized in a binary search tree, since searching a host using this data structure will take in average O(log N) time where N is the number of physical hosts.

Hence for each virtual machine step4, step 5, step 11 and step 12 will run for $O(\log 2n)$ in average case and O(n) in worst case. step 3 and step 6 will run for O(1) times for each virtual machine in all cases. As we proved Allocate procedure will run for $O(\log 2n)$ in average case and O(n) in worst case for each virtual machine. Hence for each virtual machine placement the Virtual Machine Scheduler algorithm will run for $O(1) + O(\log 2n) + O(\log 2n) + O(1) + O(\log 2n) + O(\log 2n) + O(\log 2n)$ times i.e., $O(\log 2n)$ in average case and $O(1) + O(n) + O(n) + O(1) + O(n) + O(n) + O(n)$ times i.e., O(n) in worst case. Hence for m virtual machines it will run for $O(m \log 2n)$ times in average case and O(m n) times in worst case.

Simulation and Results
The proposed Virtual Machine Scheduler algorithm is simulated along with other three algorithms (Ranking algorithm, Greedy rest t algorithm and Round-Robin algorithm) using Cloud Sims [30].

Cloud Sims is an open-source toolkit developed in Java, for simulation of cloud computing application and environment. Cloud Sims can be used to simulate data-center, Host machines, Virtual Machines, Cloudlet (a dummy application that will run on a virtual machine) etc. by using predefined classes and functions.

Using this simulator we have created some host machines in a data-center and some virtual machines to run on those host machines. We have implemented four Virtual Machine provisioning models (Virtual Machine Scheduler, Greedy First Fit, Ranking and Round-Robin) to allocate the created virtual machines to the host machines depending on the type of algorithm used in each.

CONCLUSION

Virtual machine placement is an important issue in cloud computing; it is because all the requests that arrive for any infrastructure have to be served by creating virtual machines of the requested specification on the underlying physical machines. In case of On-Demand access the virtual machine requests have to served quickly for a small interval of time. In this paradigm to serve more requests at a particular time-frame, the physical machines should be used effectively i.e., the virtual machine placement policy should be good enough to minimize the number of physical machines used, considering the cost and SLA. In this thesis we discussed some virtual machine placement policies adopted by various open-source cloud computing solutions. We discussed our proposed framework for efficiently solve this problem, in which we described our proposed policy named Virtual Machine Scheduler for virtual machine placement. From the results obtained it is clear that the proposed Virtual Machine Scheduler is performing much better than other discussed placement policies in terms of minimizing cost, minimizing allocation time and minimizing SLA violations.

REFERENCES

[1] Shuai Zhang, Shufen Zhang, Xuebin Chen, and Xiuzhen Huo. Cloud computing research and development trend. In Future Networks, 2010. ICFN'10. Second International Conference on, pages 93{97. IEEE, 2010.
[2] Barrie Sosinsky. Cloud computing bible, volume 762. Wiley, 2012.
[3] Armando Fox, Rean Gri th, A Joseph, R Katz, A Konwinski, G Lee, D Patterson, A Rabkin, and I Stoica. Above the clouds: A berkeley view of cloud computing. Dept. Electrical Eng. and Computer. Sciences, University of California, Berkeley, Rep. UCB/EECS, 28, 2009.

[4] Piyush Patel and Arun Kr Singh. A survey on resource allocation algorithms in cloud computing environment. In Golden Research Thoughts Volume 2, Issue. 4. Ashok Yakkaldevi, 2012.
[5] Peter Mell and Timothy Grance. The nist definition of cloud computing (draft). NIST special publication, 800:145, 2011.
[6] Bernard Golden. Cloud computing: Two kinds of agility. www.cloudtweaks.com/2010/07/ cloud-computing-two-kinds-of-agility/, 2010.
[7] Multitenancy. en.wikipedia.org/wiki/Multitenancy, 2013.
[8] Cloud computing demystifying saas, paas and iaas. www.cloudtweaks.com/2010/05/cloud-computing-demystifying-saas-paas-and-iaas/, 2010.
[9] Ryan Sobel. Virtualization vs. cloud computing: There is a di erence. www.hightech-highway.com/virtualize/virtualization-vs-cloud-computing-there-is-a-difference/, 2012.
[10] John Considine. Do Virtual Machines still matter in the cloud? www.cloudswitch.com/page/do-virtual-machines-still-matter-in-the-cloud, 2010.