Study of Virtual Machine Placement, its Parameters, Challenges, and State of the Art in Cloud Computing

Kumaraswamy S

Associate Professor, Dept. of CSE
Global Academy of Technology, Bengaluru

Abstract

Cloud computing data centers host hundreds of thousands of virtual machines (VMs) in real world scenarios. With the emergence of cloud computing, computing resources are provisioned as metered on-demand services over networks, and can be rapidly allocated and released with minimal management effort. In the cloud computing paradigm, the virtual machine is one of the most commonly used resource carriers in which business services are encapsulated. In this context, Virtual Machine Placement (VMP) is one of the most challenging problems in cloud infrastructure management, considering also the large number of possible optimization criteria and different formulations that could be studied. The primary contribution of this article is the development of our combinatorial optimization approaches to virtual machine placement parameters, challenges, literature and state of the art in cloud environments. This work presents a comprehensive up to date analysis of the nearly relevant VMP published writings in order to identify research opportunities.

I. INTRODUCTION

Cloud end-users (e.g., service consumers and developers of cloud services) can access various services from cloud providers such as Amazon, Google and SalesForce. They are relieved from the burden of IT maintenance and administration and it is expected that their total IT costs will decrease. From a cloud provider’s or an agent’s perspective, however, due to the scale of resources to manage, and the dynamic nature of service behaviors (with rapid demands for capacity variations and resource mobility), as well as the heterogeneity of cloud systems, resource allocation and scheduling are becoming challenging issues, e.g., to find optimal placement schemes.
for resources, and resource reconfigurations in response to the changes of the environment [1].

The process of selecting which virtual machines (VMs) should be located (i.e. executed) at each physical machine (PM) of a datacenter is known as Virtual Machine Placement (VMP). The VMP problem has been extensively studied in cloud computing literature and several surveys have already been presented. Existing surveys focus on specific issues such as: (1) energy-efficient techniques applied to the problem [2][3] (2) particular architectures where the VMP problem is applied, specifically federated clouds [4] and (3) methods for comparing performance of placement algorithms in large on demand clouds [5].

There is a multitude of parameters and considerations (e.g., performance, cost, locality, reliability and availability, etc.) involved in the decision of where and when to place and reallocate data objects and computation resources in cloud environments. Some of the considerations are consistent with one another while others may be contradicting. At the same time, we are witnessing an increasing trend towards hosting soft real-time applications, such as airline reservation systems, virtual reality applications, Netflix video streaming and Coursera online digital learning, on the cloud. These applications demand more stringent performance requirements, e.g., being sensitive to latency and response times. The resource overbooking used by cloud providers may incur negative impact on their performance because multiple collocated VMs caused by resource overbooking can trigger significant performance interference [6][7][8][9] for applications hosted on their respective VMs.

Although there exists prior work on performance isolation [9] among VMs collocated on an overbooked host machine, it is still a challenging task to shield the VMs from its neighbors due to the nature of resource sharing, resource overbooking practices employed, and the fluctuating workload characteristics in the cloud. Therefore, an application running on one VM might impact the performance of another application running on a separate VM on the same host machine. Specifically, network-intensive and compute intensive applications might be affected considerably.

Since performance interference is caused because of how one VM interacts with another collocated VM, addressing the performance interference challenges that stem from resource overbooking and satisfying the response time requirements of soft real-time applications will require effective placement of VMs on host machines by carefully considering the actual workload characteristics of the VMs. Due to the changing dynamics of the workloads on the VMs and also because VMs often tend to migrate from one physical machine to another for a variety of reasons, traditional and offline heuristics such as bin packing will not be applicable for interference-aware VM placement in cloud computing.

II. CLOUD COMPUTING

Cloud Computing provides a paradigm shift following the shift the way in which current enterprises IT infrastructure is constituted and it is a new paradigm in which
computing is delivered as a service rather than a product, whereby shared resources, software, and information are provided to consumers as a utility over networks.

A. **Hardware Virtualization**

Virtualization is a technology that combines or divides computing resources to present one or many operating environments using methodologies like hardware and software partitioning, partial or complete machine simulation, time-sharing, and others. Virtualization is a technology that separates computing functions and implementations from physical hardware. It is the foundation of cloud computing, since it enables isolations between hardware and software, between users, and between process and resources. Virtualization technologies find important applications over a wide range of areas such as server consolidation, secure computing platforms, supporting multiple operating systems, kernel debugging and development, system migration, etc., resulting in widespread usage. Most of them present similar operating environments to the end user; however, they tend to vary widely in their levels of abstraction they operate at and the underlying architecture. Hardware virtualization approaches include (1) Full Virtualization (2) Partial virtualization (3) Para Virtualization.

Cloud systems deployable services can be encapsulated in virtual appliances (VAs) [10], and deployed by instantiating virtual machines with their virtual appliances [11]. We have identified the following abstraction levels: instruction set level, hardware abstraction layer (HAL) level, operating system level, library level and application level virtual machines. By decoupling the hardware and operating system infrastructure provider from the application stack provider, virtual appliances allow economies of scale on the one side to be leveraged by the economy of simplicity on the other.

B. **XaaS Service Models**

Commonly associated with cloud computing are the following service models:

1) **Software as a Service (SaaS)**

In the SaaS model, software applications are delivered as services that execute on infrastructure managed by the SaaS vendor. Consumers are enabled to access services over various clients such as web browsers and programming interfaces, and are typically charged on a subscription basis. It is based on the concept of renting an application from a service provider rather than buying, installing and running software yourself.

2) **Platform as a Service (PaaS)**

In the PaaS model, cloud providers deliver a computing platform and/or solution stack typically including operating system, programming language execution environment, database, and web server [12]. Application developers can develop and run their software on a cloud platform without having to manage or control the underlying hardware and software layers, including network, servers, operating systems, or storage, but maintains the control over the deployed applications and possibly
configuration settings for the application-hosting environment [13]. Examples include Force.com, Microsoft Azure, and Google App Engine.

3) **Infrastructure as a Service (IaaS)**
In IaaS model, computing resources such as storage, network, and computation resources are provisioned as services. Consumers are able to deploy and run arbitrary software, which can include operating systems and applications. Consumers do not manage or control the underlying cloud infrastructure but have to control its own virtual infrastructure typically constructed by virtual machines hosted by the IaaS vendor. Examples include Amazon EC2 and S3, Rack space, AT&T, and Verizon.

C. **Cloud computing scenarios**
Based on the cloud services offered, two main stakeholders in a cloud provisioning scenario can be identified: (1) Infrastructure Provider (IP) (2) Service Provider (SP). IP who offers infrastructure resources such as virtual machines, networks, storage etc., which is used by SP to deliver to end-user services such as SaaS to their customers, these services are being developed using PaaS tools. As mentioned in [14] four main types of cloud scenarios are identified:

1) **Private cloud**
And organization provisions services using internal infrastructure and thus runs the roles of both SP and IP. Private clouds can avoid many of the security and privacy concerns related to hosted sensitive information in public clouds, the latter is where the SP leases IaaS resources publicly available IPs. Private cloud as in Fig.1 offers stronger guarantees on control, monitor and performance as the whole infrastructure can be administered within the same domain.

![Figure 1: Private cloud scenario](image)

2) **Cloud Bursting**
Private clouds may offload capacity to other IPs under periods of high workload, or for other reasons, e.g., planned maintenance of the internal servers. Here, the providers form hybrid architecture commonly referred to as cloud bursting as in Fig. 2. Basically, less sensitive tasks are executed in the public cloud instead while tasks that requiring higher levels of security are provisioned the private infrastructure.
3) **Federated Cloud**
Federated clouds are IPs collaborating on a basis of joint load-sharing agreements enabling them to offload capacity to each others [15] in a manner similar to how electricity providers exchange capacity. The federation takes place at the IP level in a transparent manner. In other words, an SP that deploys services to one of the IPs in a federation is not notified if its service is off-loaded to another IP within the federation. However, the SP is able to steer in which IPs the service may be provisioned, e.g., by specifying location constraints in the service manifest, Fig. 3 illustrates a federation between three IPs.

4) **Multi-Cloud**
In multi-cloud scenarios as in Fig. 4, the SP is responsible for handling the additional complexity of coordinating the service across multiple external IPs, i.e., planning, initiating and monitoring the execution of services.

IV. VIRTUAL MACHINE PLACEMENT

Given a set of admitted services and the availability of local and possibly remote resources, there are a number of placement problems to be solved to determine where to save data and where to execute VMs. The following sections describe the challenges and state of the art of VM placement and scheduling in cloud environments.

A. Parameters and Considerations

There are a multitude of parameters and considerations involved in the decision of where and when to reallocate data objects and computations in cloud environments. An automated placement and scheduling mechanism should take into account the considerations and tradeoffs, and allocate resources in a manner that benefits the stakeholder for which it operates (SP or IP). For both of these, this often leads to the problem of optimizing price or performance given a set of constraints, often including the one of price and performance that is subject to optimization. Among the main considerations are:

1) Performance:

In order to improve the utilization of physical resources, data centers are increasingly employing virtualization and consolidation as a means to support a large number of disparate applications running simultaneously on server platforms. With different placement schemes of virtual machines, the performance achieved may differ a lot [7].

2) Cost:

The price model was dominated by fixed prices in the early phase of cloud adoption. However, cloud market trend shows that dynamic pricing schemes utilization is being increased [16]. Investment decreases by dynamically placing services among clouds or by dynamically reconfiguring services (e.g., resizing VM sizes without harming service performance) become possible. In addition, internal cost for VM placement, e.g., interference and overhead that one VM causes on other concurrently running VMs on the same physical host, should also be taken in to account.

3) Locality:

In general, for considerations of usability and accessibility, VMs should be located close to users (which could be other services/VMs). However, due to e.g., legal issues and security reasons, locality may become constraints for optimal placement.

4) Reliability and continuous availability:

Part of the central goals for VM placement is service reliability and availability. To achieve this, VMs may be placed/replicated/migrated across multiple (at least two)
geographical zones. During this procedure, factors such as the importance of the data/service encapsulated in VMs, its expected usage frequency, and the reliability of the different data centers, must be taken into account.

B. Challenges
Given the variety of deployment scenarios, the range of relevant parameters, and the set of constraints and objective functions of potential interest, there are a number of challenges to the development of broadly applicable placement methods, some of which are presented below.

Firstly, there exists no generic model to represent various scenarios of resource scheduling, especially when users’ requirements are vague and hard to encode through modeling languages.

Secondly, model parameterization, i.e., finding suitable values for parameters in a proposed model is a tedious task when the problem size is large. For example, in for a multi-cloud scenario that includes n cloud providers and m VMs, (m*n) assignments are needed to express the VM migration overheads ignoring possible changes of VM sizes. Therefore, mechanisms that can help to automatically capture those values are required.

Thirdly, the VM placement problem is typically formulated as a variant of the class constrained multiple-knapsack problem that is known to be NP hard [17]. Thus, tradeoffs between quality of solution and execution time must be taken into account. This is a very important issue given the size of real life data centers, e.g., Amazon EC2 [18], the leading cloud provider, has approximately 40,000 servers and schedules 80,000 VMs every day [19].

C. State of the Art
Virtual machine placement in distributed environments has been extensively studied in the context of cloud computing. Such approaches address distinct problems, such as initial placement, consolidation, or tradeoffs between honoring service level agreements and constraining provider operating costs, etc. [5]. Studied scenarios are usually encoded in mathematical models and are finally solved either by algorithms such as approximation, greedy packing and heuristic method, or by existing programming solvers such as Gurobi [20], CPLEX [21] and GLPK [22]. Those related work can be separated into two sets: (1) VM placement in single-cloud environments and (2) VM placement in multi-cloud environments.

In single-cloud environments, given a set of physical machines and a set of services (encapsulated within VMs) with dynamically changing demands, on-line placement controllers that decide how many instances to run for each service and where to put and execute them, while observing resource constraints, are NP hard problems. Tradeoff between quality of solution and computation cost is a challenge. To address this issue, various approximation approaches are applied, e.g., by Tang et al. [17] propose an algorithm that can produce within 30 seconds high-quality solutions for hard placement problems with thousands of machines and thousands of VMs. This approximation algorithm strives to maximize the total satisfied application demand, to
minimize the number of application starts and stops, and to balance the load across machines. Hermenier et al. [23] present the Entropy resource manager for homogeneous clusters, which performs dynamic consolidation based on constraint programming and takes migration overhead into account. Entropy chooses migrations that can be implemented efficiently, incurring a low performance overhead. The CHOCO constraint programming solver [24], with optimizations e.g., identifying lower and upper bounds that are close to the optimal value, is employed to solve the problem. To reduce electricity cost in high performance computing clouds that operate multiple geographically distributed data centers, Le et al. [25] study the impact of VM placement policies on cooling and maximum data center temperatures, develop a model of data center cooling for a realistic data center and cooling system, and design VM distribution policies that intelligently place and migrate VMs across the data centers to take advantage of time-based differences in electricity prices and temperatures.

For VM placement across multiple cloud providers, information about the number of physical machines, the load of these physical machines, and the state of resource distribution inside the IP side are normally hidden from SP, and hence not parameters that can be used for placement decisions. Only provision related information such as types of VM instance, price schemes, are exposed to SP. Hence, most works on VM placement across multi-cloud environments are focusing on cost aspects. Chaisiri et al. [26] propose an stochastic integer programming (SIP) based algorithm that can minimize the cost spending in each placement plan for hosting virtual machines in a multiple cloud provider environment under future demand and price uncertainty. Borsches et al. [27] examine the workload outsourcing problem in a multi-cloud setting with deadline constrained, and present cost-optimal optimization to maximize the utilization of the internal data center and to minimize the cost of running the outsourced tasks in the cloud, while fulfilling the applications quality of service constraints. Tordsson et al. [28], propose a cloud brokering mechanisms for optimized placement of VMs to obtain optimal cost-performance tradeoffs across multiple cloud providers. Similarly, Vozmediano et al. [29] [30] explore the multi-cloud scenario to deploy a computing cluster on top of a multi-cloud infrastructure, for solving loosely-coupled Many-Task Computing (MTC) applications. In this way, the cluster nodes can be provisioned with resources from different clouds to improve the cost-effectiveness of the deployment, or to implement high-availability strategies. Fei et al [35] considers convex optimization theory for optimization of virtual machine placement which minimize the network traffic minimization and resource utilization maximization. Wang et al [36] presents an improved virtual machine (VM) placement mechanism, called Energy efficiency and Quality of Service (QoS) aware VM Placement (EQVMP) to overcome the problem of unbalanced traffic load in switching on and off VMs for the purpose of energy saving, three-tier algorithm to take both energy efficiency and QoS into consideration.
al[33] proposed a hybrid algorithm ACO-PSO to optimize algorithm for virtual machine placement in cloud computing. On similar lines, V A K Sarma et al [34] proposed an hybrid algorithm that combines both multi-objective genetic algorithm and ant colony optimization for the virtual machine placement problem

V. FUTURE WORK
Future directions for this work include modeling the interconnection requirements that can precisely express the relationships between VMs to be deployed. In addition, researchers are working on a specific scenario where cloud users can specify hard constraints and soft constraints when demanding resource provisions. A hard constraint is a condition that has to be satisfied when deploying services, i.e., it is mandatory. In contrast, a soft constraint is optional. An optimal placement solution with soft constraints satisfied is preferable over other solutions. The hard and soft constraints can be used to specify collocation or avoidance of co-location of certain VMs. Readers or researchers can also investigate on how to apply multi-objective optimization techniques to this scenario. Another area of future work is approximation algorithms based on problem relaxations and heuristic approaches such as greedy formulation for considerations of tradeoff between quality of solution and execution time.

In conclusion this detailed study has paved way for more research opportunities in the field of cloud computing for VM placement and associated challenges.

REFERENCE


Kumaraswamy S


Study of Virtual Machine Placement, its Parameters, Challenges


