Literature study
Increasing resource utilization of cloud

You Hu
adolphus.hu@student.vu.nl
VU Amsterdam

1 INTRODUCTION

The cloud computing aims to provide reliable, customized and QoS guaranteed dynamic computing environment for end-users[4]. The key advantage of cloud computing is the elasticity of resource provisioned on demand. It allows developers to deploy their innovative products with much less capital for hardware. Besides, for companies requiring to process batch tasks, cloud computing provides close to unlimited resources enabling the tasks to get accelerated as long as the program is well-scalable.

From the point of financial concern, it is vital for cloud providers to maximize the resource utilization for their assets. The first paradigm of cloud computing is based on virtual machine technologies. After years of development, containerization as a lighter-weight virtualization technology is used to manage applications widely. In this report, we will firstly discuss what is concerned in terms of resource utilization. Then, we will list researches which aims to maximize resource utilization in virtual machine and container environment respectively.

2 WHAT ARE CONCERNED IN CLOUD AREA REGARDING RESOURCE UTILIZATION

Definition of resources

In the cloud environment, there are many kinds of resources and a set of aspects around the cloud economy. Jennings et al.[2] categorize the resources into compute, networking, storage, and power. Manvi et al.[3] summarize that there are physical resources(CPU, memory, storage, network elements and sensors) and logical resources(OS, energy, Network throughput/bandwidth, Load balancing mechanisms and so on). There are overlap between two of them especially in the physical part, while Manvi adds API, OS, load balancing to logical resource concept. In this report, we mainly focus on the resource optimization on physical resources like CPU, memory, storage; and consider the trade-off between utilization rate and QoS.

Quality of Service metrics

QoS (Quality of Service) metrics are import to both cloud provider and consumer. Bardsiri and Hashemi listed detailed metrics from four kinds of features:performance, economic, security and general. [1] Their coverage is comprehensive. There are plenty of features and corresponding metrics that cloud users would concern.

There are few metrics we consider important in terms of resource utilization. For performance features, the CPU load rate and packet loss frequency are what users may concern while the cloud provides needs to make a compromise for utilization of resources. In the economic aspect, the price per resource unit is the key point that cloud providers and users wrestle on. However, from the technical view, the time for VM booting/deleting/suspending/provision attract more attention. Besides, the availability and reliability are very important as well. Cloud providers need to pay effort on fault tolerance to make sure the safety of the cloud. Usually Recovery Point Objective (RPO) and Recovery Time Objective (RTO) are the key parameters for disaster recovery concerned in researches.

In the following sections, we will explore how cloud providers face resource management issues and the metrics shown above play important roles in those researches.

3 VIRTUAL MACHINE FOUNDED CLOUD ERA

After the first introduction by Eric Schmidt in 2006, cloud computing met its sharp increase. The development of cloud computing is based on the maturation of virtualization technologies. The fine-granularity dividable resource enables users to obtain the 'pay-as-you-go' resources, and data centers can profit from selling every portion of the resource.

More detailed, the cloud users and cloud providers make a formal agreement known as SLA(Service level agreement). Both cloud providers and cloud consumers need to formulate their management functions regarding the SLAs as for cloud consumers; they also need to meet the SLA requirement to their end-users. Besides the SLAs, the data center infrastructure-related objectives include load-balancing, fault tolerance, and energy. While for cloud users, they need to make a trade-off: conservative over-provisioning but less profit or aggressive minimizing cost but a higher risk of violating SLAs. Considering the resources, computation, networking, storage, power, and etc., academia and industry developed mechanisms or systems to meet SLAs' requirements and achieve better resource management, thus more financially efficient.

Global scheduling

The most straightforward approach is to optimize the global scheduling of virtualized resources. Mills et al. compared 18 (heuristics) VM placement algorithms by experiments and parameter grid search[**?**] in 2011. The 18 algorithms are made of a combination of 3 criteria for choosing a cluster, times six heuristics for choosing nodes under the two-level taxonomy. The results reflect the no-free-lunch theorem in optimization: the percent-allocated (PAL) cluster-choice criterion leads to higher average loads and utilization, but this benefit of a cloud provider is based on the negative effect on users, for instance, the waiting time; Least-Full first(LF) and Tag&Pack(TP) lead to lower cloud-wide virtual core utilization as these heuristics more often choose empty nodes on which to place VMs. However, LF tends to squeeze out some larger VM types, which leads to yield lower user success rate and higher give-up rate. This comparison gives a benchmark for cloud providers to determine which algorithm to employ. On the other hand, it indicates that the significant outperformance usually exists in the domain-specific scenario. Recent researches support this trending. The [**?**] focuses on energy aware VM placement; the traffic linked placement solution is purposed in [**?**]; and [**?**] considers the VM placement in heterogeneous clusters.

VM placement

Following the VM placement, dynamic placement comes to the stage naturally, which benefits from the VM rescaling and live migration. Sharma et al. [**?**][**?**] formulated the problems of VM rescaling, replication and lived migration under the cost-efficient scenario as Integer Linear Programming problems , and proposed Kingfisher, a set of techniques based on greedy heuristic solutions. This work shows a cost-aware rescaling and lives migration algorithm can reduce the cost of VM transition(paused, serialized, and transferred to a different PM) compared to the cost-oblivious algorithms. In [**?**], Wuhib et al. propose a decentralized solution, utilizing a gossip protocol. It is shown to scale to problem sizes above 150,000 PMs and 350,000 VMs and reduces the number of migrations.

Furthermore, the global scheduling for applications and VMs requires resource demand profiling, resource utilization estimation, and resource pricing & profit maximization as a compliment. For demand profiling, a typical model-based solution is [**?**], which employs Fast Fourier Transform(FFT) for resource usage detecting and a discrete-time Markov chain for demand predictions. Furthermore, it was extended by online adaptive padding and reactive error correction to mitigate under-estimation[**?**]. Resource utilization estimation is not popular in research as the metrics for utilization are very clear. The considerable research in this topic may relate to the special metric. For instance, in [**?**], a particular CPU and memory-related utilization metrics used by VMware's Distributed Power Management (DPM) is discussed as it has effects on VMware's power management mechanism. In [**?**], Zhao et al. developed an online algorithm to maximize the profit of the cloud provider over the long run via scheduling job over data centers cross geolocation.

Local scheduling

Following the global scheduling, the local scheduling of VM is vital as well. The explicit solutions may formulate the problem into the allocation of physical machine resources to VMs. Urgaonkar et al. proposed an approach aiming to dynamic resource allocation based on queuing information. Therefore the online control is achieve[**?**]. Each physical machine is equipped with a resource controller and a buffer(queue) containing applications in this approach. Thus, as it is shown in the work [**?**], the sequence of the execution of VMs on PM has a significant impact on response time. The authors propose a local scheduler combining both compute resource allocation and control of VM execution sequencing.

The researches mentioned above are collected due to their representative. Most of them are either published in the early 2010s or very recent. Overall, these researches reflect a trend that after years of development in the cloud area, resource management research at the VM level has been to an in-depth and much more scenario dependent. One remarkable exception is that with the rising of deep learning, the state-of-art on old and general topics has a big step forward. The historical statistic based deep(reinforcement) learning models show substantial advantages on traditional rule-based, fuzzy theory, and heuristic solutions.

Provisioning

Workload management

4 CONTAINER AND ORCHESTRA

Architecture

Scheduling

Placement

VMs enable clouds to achieve elasticity of large-scale shared resources, while it is still a heavyweight solution for dynamic provisioning. Containers as a lightweight technology to virtualize applications have recently been successful and widely used for especially web applications. Docker is the most popular container solution in the industry at this mount. Therefore, efficient management of the container layer, inserted between VMs and applications, is the must for a container-based cloud. The management can be implemented directly or on top of container orchestras like Kubernetes or docker

swarm. Researches may either focus on direct management or the management with the help of orchestra frameworks.

The problems can be considered the same as resource management on VM; global scheduling is the most straightforward approach. Due to the lightweight nature of containerization, the placement of containers can be more flexible. In [**?**], the authors proposed architecture for docker container placement. It especially takes into consideration the collaboration of two kinds(container and VM) of placement. Normally, The placements of the VM to PM and container to VM are based on the best-fit algorithm. This work extends the best-fit algorithm, and the placement of containers should consider the resource utilization of PM. However, this work considers the problem of container placement on VM as a bin pack problem. This assumption makes local scheduling of operating systems useless, and the elasticity of container has been sacrificed. Another work[**?**] is proposed for container placement as well, which is called Resource Stable Placement(RSP). The difference in architecture is that there is no VM placement problem, only the placement of containers on heterogeneous PMs. The resources are modeled via a vector where each element represents the amount of one kind of resource. Therefore, more than CPU and memory can be considered for resource placement. This scheduling optimization shows better performance in response time and utilization of resources when the workload is heavy.

Besides the direct management, using the orchestra tool, especially like Kubernetes, are the most common approach. In [**?**], the authors developed a Reference net-based model that employs real data from Kubernetes. It characterizes the performance of Kubernetes and the lifecycle of containers and Pod. Kubernetes provides a diverse interface for task management and resource management. However, as it is mentioned in [**?**], the scheduling only considers the current optimal node, regardless of the use of resource costs. It provides a better allocation of Pod scheduling, which is based on the Ant Colony Algorithm and Particle Swarm Algorithm. The result of experiments on CloudSIM shows resource utilization, and load balancing can be improved significantly.

REFERENCES

[1] Amid Khatibi Bardsiri and Seyyed Mohsen Hashemi. 2014. QoS Metrics for Cloud Computing Services Evaluation. *International Journal of Intelligent Systems and Applications* 6, 12 (2014), 27–33. https://doi.org/10.5815/ijisa.2014.12.04

[2] Brendan Jennings and Rolf Stadler. 2015. Resource Management in Clouds: Survey and Research Challenges. *Journal of Network and Systems Management* 23, 3 (2015), 567–619. https://doi.org/10.1007/s10922-014-9307-7

[3] Sunilkumar S. Manvi and Gopal Krishna Shyam. 2014. Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey. *Journal of Network and Computer Applications* 41, 1 (2014), 424–440. https://doi.org/10.1016/j.jnca.2013.10.004

[4] Lizhe Wang, Gregor Von Laszewski, Andrew Younge, Xi He, Marcel Kunze, Jie Tao, and Cheng Fu. 2010. Cloud computing: a perspective study. *New generation computing* 28, 2 (2010), 137–146.