A Survey of Resource Allocation Optimization in Cloud Computing

Wentao Kuang
School of Engineering and Computer Science
Victoria University of Wellington
Email: wt.kuang@icloud.com

Hui Ma
Supervisor, Senior Lecturer
School of Engineering and Computer Science
Victoria University of Wellington

ABSTRACT

Cloud computing provides available, scalable processing and storage resources by sharing commodity computer resources which significantly restructuring the IT industry. Small companies can benefit from using state of the art services and infrastructures with a reasonable price, while large organizations are interested in the high processing speed and ultimate storages. However, optimal allocation of resources in cloud environment is a hard problem due to many factors, such as the heterogeneity of resource types, the scale of data centers, and the diversity of objectives from different actors in the cloud. In this paper, we survey many recent literatures, and identify five challenges for the future scope, includes auxiliary objectives, coherent cloud optimization simulator or framework, hybridization, problem scalability, and cross-layer optimization.

Keywords: *CLoud Computing, Resource allocation, Survey.*

1 Introduction

Cloud Computing (CC) delivers hosts of services over the internet and enables companies to consume cloud resources, such as virtual machines (VMs), network,storage, and power. Due to the increasing demands of the cloud resources, it becomes difficult to optimally allocate resources based on the users demands in order to satisfy the their requirement under the constraint of Service Level Agreement(SLA) [1].

CC resource allocation is a novel, interesting and important area to explore. CC resource allocation often refers to the efficient allocation between the resource requests and heterogeneous cloud resources by the different actors in cloud environments [2]. The resource requests can be varied from different actors, for example, cloud clients may request for VMs to deploy web services or applications, and cloud providers may request to allocate their VMs into Physical Machines (PMs) in the data centers. Cloud resources can be classified into different types of resources, such as compute resource, network resource and storage resource and power resource. The allocation objectives also different concerned by different cloud actors, cloud clients often concern their deployment cost and Quality of Service (QoS), and cLoud provider may interested in minimization of the power consumption and resource utilization [3]. Apparently, the cloud resource allocation problems present a variety of mixture due to the complex architecture, and many actors involving. In our survey, our first contribution is to classify these resource allocation problems based on the cloud requests, resource types and cloud actors.

These resource allocation problems are large-scale optimization problems with dynamics an uncertainty due to the multiple actors and users with variety of requirements and objectives. So most of CC resource allocation problem are NP-hard problem which means there no exact and optimal algorithms for optimization, in order to solve these NP-hard problems, Heuristic algorithms are common solutions for CC resource allocation optimization. Computational Intelligence (CI) is also known as soft computing which provides near optimal solutions to NP-hard problems. CI technique deals with many CC resource allocation factors resulting from the multi-tier structure and diverse requirements, such as dynamics, uncertainty and approximation. CI community using techniques such as Evolutionary Computing (EC) and Fuzzy System to solve CC

resource allocation problems [4]. Another contribution of our survey is to identify the gap and challenges of CI techniques in the cloud resource allocation scope.

After reviewing the literatures under the scope of cloud resource allocation optimization with CI techniques, we classified the allocation problems into four main categories: *Cloud Brokering, Service Allocation, VMs Allocation* and *VMs Migration*, which will discuss in the section 2 and section 3. We also identify five challenges in section 4, which include: *auxiliary objectives, coherent cloud optimization simulator or framework, Hybrid algorithms, problem scalability,* and *cross-layer optimization*. Finally, we give our conclusion.

2 Background

CC Architecture consists of different layers and construct to distinct business model which is managed by different cloud actors. Cloud resource allocation challenges are distributed between these layers, and resource allocation objectives are concerned by different actors. These objectives many vary due to different concerns, such as financial, environment, performance, geo-location and so on [5]. This section will introduce the background knowledge of the CC resources allocation includes CC architecture and cloud actors and their objectives.

2.1 Cloud Computing Architecture

The commonly architecture of CC can be separated into four layers: *Application layer, Platform layer, Infrastructure layer* and *Hardware layer* [6].

The hardware layer is also known as server layer, it is the bottom layer of the cloud stack, which represents all the cloud physical resources, such as physical machines, power and cooling systems. This layer is implemented in data centers and consists of thousands of servers. The main resource allocation challenge in hardware level is efficient utilization of the hardware resources, so the power resource usage is minimized.

The infrastructure layer provides virtual resources on top of physical resources by using Virtualization techniques, such as KVM, Xen and VMware [7], so it is also called Virtualization layer. The virtual resources consist of computing and storage resources by partitioning the physical machines. The main challenge here is the efficiently allocate of virtual resource to physical machine to minimize the usage of physical resources.

The platform layer rests on the infrastructure layer, if provides operating systems and application frameworks for customers to develop their application, so the customers do not need to manage their applications directly in Virtual Machine (VM). The platform layer is responsible to manage the platform resources, such as framework and VM. The main challenge in platform layer is optimizing the performance of platform by a given available resources, i.e. Efficiently allocate VMs between data centers in order to maximize the network performance.

The Application layer is on the top of the Cloud Computing architecture hierarchy, it can be web applications and services that serve end user directly. The application layer manages all the software resources and encompasses the largest accessible layer of CC. Their main concern is the Quality of Service (QoS), which is constrained by the service provisioning problem.

2.2 Cloud Computing Business Model

Cloud Providers employ a service driven business model [8], they provide different resources as services. In practice, CC business models are normally described as three main levels of service, that is, *Infrastructure as a Service (IaaS), Platform as a Service (PaaS)* and *Software as a Service (SaaS)* [9]. For CC architecture discussed in the previous section, every layer can be an implementation of service business model or a customer on the top of another layer [10].

Infrastructure as a Service provides computer infrastructure as a service such as computing capacity and storage, customers can enjoy the state of the art technology and pay as growth infrastructure resources from public cloud. Infrastructure providers manage thousands of physical resources and provide virtual resources by sharing and virtualization. In order to optimize the infrastructure resources, VM allocation, load balancing and Capacity planning are some common problems for infrastructure providers.

Platform as a Service is an additional abstract implementation based on IaaS. Platform providers offer OS, framework, storage for customers to develop and host web applications. In practice, PaaS and IaaS are commonly owned by the same organization, such as Google, Amazon and Microsoft. Hence, PaaS and IaaS providers often faced same resource management problems. As a result, Most of studies in the research community are only tackling the resource allocation problem on IaaS and SaaS level.

Software as a Service serves a wide variety of customers with web applications in the cloud. Web Service providers host their services on geographically distributed servers to provide large scaled online services for customers. Service providers always consider the QoS and profit optimization objectives since they only need to consider the service provisioning problem such as cloud brokering and service allocation.

3 Resource Allocation Literature Review

As previously discussed, CC offers services as IaaS, PaaS and SaaS models. In terms of resource allocation, based on the usage and providing of these business models cloud actors can be described as follows: *Cloud Provider, Cloud User, End User* and *Cloud Broker*. Each actor concerns different objectives for the resource allocation optimization.

3.1 Cloud Provider based Resource Allocation

Cloud Provider owns a set of data centers and provides both hardware resources and software resources. A cloud provider can offers many business models depends on the deployment model. In the context of public cloud, cloud provider normally provides IaaS and PaaS to the cloud user, the goal of cloud provider is to maximum their income by minimizing cost under the constraint of SLA. The cost are usually related to the power usage, resource utilization. In terms of private cloud, cloud provider manages all the layers from hardware layer to the software layer, and provides SaaS to the End User, in addition to the cost related objectives, cloud provider also need to consider the QoS objectives such as throughput and response time. The most basic concerns of cloud providers are how to efficiently allocate VMs to the PMs, since they always manage the hardware layer and infrastructure layer. Some static problems like VM allocation and dynamic problem like VM migration are exploited by many researchers.

3.1.1 VM Placement and Allocation

The fundamental technology powers CC is virtualization, which create a pool of virtualized resources on top of physical resources. It enables the infrastructure provider to reduce IT costs while increasing utilization. VM allocation as a concomitant bin packing problem of virtualization, is the process to optimally map virtual machine to appropriate infrastructure and Physical Machine (PM) in a data center or private cloud. Common objectives concerned by cloud provider in VM placement are energy, resource utilization, performance, cost. The static VM allocation differs from VM migration and consolidation, which only consider the initial deployment of the data center. Table 1 is the summary of VM allocation problems.

An single objective energy aware VM allocation problem is solved by Gao et al. [11] with considering both PM constraints and network constraints. They propose an Ant Colony Optimization (ACO) to minimize a single objective energy cost. The PM constraints include CPU utilization, memory capacity and idle PMs, the network constraints are flow conservation, flow routing and bandwidth. They use a 2 dimensional matrix to present the ACO solution of allocating VMs to PMs, that each VM row can only select one element to ensure a VM is hosted on only one PM. They compare ACO with a greedy algorithm, which is First Fit Decreasing Algorithm (FFD). The test results show that ACO always has better performance than FFD for different number of VMs. They also illustrate that the network constraint will limit the optimization result. Finally, they prove that the ACO has much better performance than FFD, since the FDD greedy strategy can easily fall into local optimal for large problem.

Some multi-objectives VM allocation problems are also tackled by recent researchers, Portaluri and Giordano [12] propose a VM allocator for CC that allocate a set of VMs on servers in the data center. They solve this problem with a Multi-Objective GA (MOGA) approach by considering server power, switch power consumption and resource utilization include CPU, RAM, disk and bandwidth. They compared their allocator with a mixed integer linear problem which solve by CPLEX [13]. They set up experiments with different parameters and allocated up to 2000 VMs. The results indicate that, Although MOGA and CPLEX both can find similar optimal solutions, MOGA is more powerful for larger number of VMs than CPLEX since the execution time increased significantly for more than 200 VMs. Pires and Fabio [14] devote to a Multi Objective Memetic Algorithm (MOMA) for optimally selecting PMs as VMs placement by minimizing energy consumption and network traffic and maximizing economical revenue. The experiment is performed in several scenarios with different number of instances. They also implement an Exhaustive Search Algorithm (ESA) for comparison and evaluation their proposed algorithm. Three experimental tests are conducted for different purposes. For the first test, they prove the effectiveness of MOMA. Compared to ESA, MOMA find 100% of the optimal Pareto front. The second test verifies the scalability of MOMA while considering lager solutions search space. The last test shows that by increasing the percentage of VMs with critical Service Level Agreement (SLA) also decrease the execution time, and the MOMA can find at least 95% of feasible solutions.

One emerging objective of the VM allocation problem is environmental impact. Ahvar et al. [15] present a study on the VM allocation problem which aims at minimizing the Carbon emission and cost. They propose a combination of prediction-based A* Algorithm and Fuzzy Sets optimization method called CACEV. Their scenario is focused on allocating the new VM requests to the available PMs from geographically distributed data centers, since the energy prices and carbon emission rates vary by location. As a consequence, their algorithm consists two steps, the first step selects data centers and the second step choose PM to allocate from the selected data center. They construct three different versions of CACEV for comparison, which include CACEV-Cost(only consider cost objective), CACEV-Carbon (only consider carbon emission) and CACEV (consider both objectives). After simulation, they demonstrate that CACEV improves 40-155% in carbon emissions on CACEV-Cost which increases 20-60% on cost. In general, CACEV can improve total cost by 40-60% and carbon emission by 10-30%.

Paper	CI Technique	Request	Resource	Objectives
Gao et al. [11]	ACO	VM	PM	Energy Cost
Portaluri and Giordano [12]	GA	VM	PM	Power Consumption, Resource Utilization
Pires and Fabio [14]	MA	VM	PM	Energy, Network, Revenue
Ahvar et al. [15]	A* + Fuzzy	VM	PM	Cost,Carbon Emission, Energy

Table 1. VM Placement and Allocation Optimizations

3.1.2 VM Migration

VM migration is an online server load balancing problem that scheduling the virtual resources to physical resources with balanced load. Compare to the static VM allocation problem, VM migration is a dynamic problem since the cloud providers often encounter a load imbalance problem when existing hosts are stoped and new hosts are requested dynamically. Technique likes live migration enable the VM migration with minimal disturbance to services that running inside the migration. The aims of VM migration are to shut down idle PMs and increase the utilization of functional PMs, in order to minimize energy cost and resource utilization. Reviewed papers are listed in table 2.

Some single objective VM migration problems have been proposed. Sharma and Guddeti [16] solve an on-demand VM migration problem to optimize energy consumption by a hybrid of Cat Swarm Optimization (CSO) and GA called HGACSO. The design of algorithm is embedded CSO in the middle of GA, The solutions are initialized in GA, then apply CSO algorithm for the selecting 50% fittest solutions, finally, apply GA crossover and mutation on the best cats from the CSO. The experiments consider 1500 of 3 types PMs in the data center and periodical VM request from 50 users. They compare four algorithms in the experiments include First Fit Decreasing (FFD), GA, CSO and HGACSO. Results show that HGACSO has the highest resource utilization and the lowest energy consumption. Sundararajan et al. [17] propose a constrained GA approach for rebalancing of services in cloud data centers with considering affinity and anti-affinity constraints, and the objective is minimizing the number of service migrations. Affinity constrains some VMs should place on the same host for reducing the co-hosting data transfers. Anti-affinity constrains some VMs are needed to allocate on different hosts for backup replica. They conduct simulation based experiments on real world PlanetLab data [18]. The first experiment shows the impact of the number of hosts will decrease the performance for larger number compared to demanded number. The second experiment illustrates that the increasing of services will increase the fitness for larger number of hosts. The last experiment concludes that the number of affinity and anti-affinity constraint services set is no major impact on the fitness.

A traditional multi objectives VM migrating problem is tackled by Zhao et al. [19] by NSGA-II, they consider three resource utilization include CPU, Memory and bandwidth. A set of experiments is constructed to compare NSGA-II with Random Algorithm, static Algorithm and Rank Algorithm mentioned in study [20]. They provide four different strategies since the attributes are non-dominated in NSGA-II, which are CPU dominated, memory dominated, bandwidth dominated and balance strategy. For the comparison of Algorithms, NSGA-II is obviously better than other three algorithms, and the Rank Algorithm is the worst. Ramezani et al. [21] improve a PSO by using Fuzzy Logic System (FLS) to solve VM migration problem with the multiple objectives of minimizing the power consumption and maximizing the resource utilization. In PSO, the particles situation changes dynamically with the environment, hence the choice of the inertia weight can provide a balance between local and global optimal. The author uses a fuzzy rule to select the inertia weight for achieving optimum results. The experiment results indicate that their proposed algorithm achieves better resource utilization and also decreases the bandwidth traffic.

Paper	CI Technique	Request	Resource	Objectives
Sharma and Guddeti [16]	CSO + GA	VM	PM	Energy Consumption
Sundararajan et al. [17]	GA	VM	PM	Migration Number
Zhao et al. [19]	GA	VM	PM	CPU, Memory, Bandwidth
Ramezani et al. [21]	PSO + Fuzzy	VM	PM	Power Consumption, Makespan

Table 2. VM Migration Optimization

3.2 Cloud User based Resource Allocation

Cloud User deploys application on the public cloud witch provided by cloud provider, and offers SaaS to the end users. Cloud users are interested in obtaining the best QoS performance for the lowest price while meeting the SLA with the end users. The SaaS market presents incredibly competitive, so most cloud users must consider multiple objectives include QoS factors and cost in order to survive. The cloud provider also scales and redeploy their application based on the demand of end users. The most common resource allocation problem for cloud users is service allocation, which is allocating web services to virtual resources or servers in order to provide a better SaaS with a lower cost.

3.2.1 Service Allocation

Service allocation and composition problems are often concerned by service providers who are trying to optimally allocate web service onto a set of virtualized infrastructure resources in order to optimize the cost and QoS. The QoS is often presented by latency and throughput, since network latency and throughput became increasingly important for QoS while the Internet bandwidth and computing power are guaranteed [22]. The network latency can vary between different geographical infrastructure locations where the service hosted. In order to provide a better service performance, service providers attempt to minimize the network latency by allocation servers in geographically distributed data centers with minimal cost. Table 3 summarizes the reviewed literatures about service allocation and composition problems.

Bao et al. [23] proposed an Orthogonal Genetic Algorithm for QoS objective service composition problem, they add up six QoS objectives into one same weighted fitness function, which includes execution price, execution duration, availability, throughput, successful execution rate and reliability. They also introduce an Orthogonal array design into the solution presentation in order to simplify the complexity. They set up a set of experiments based on the QWS dataset [24, 25] and compare to 20 different advanced optimization algorithms which include nine PSOs, five GA and six Differential Evolution (DE). The results show that OGA has a better solution quality than other algorithms, but often give smaller standard deviation of function values. In terms of execution speed, OGA converge faster than other PSO and GA algorithms.

Tan et al. [26–28] perform a series of studies on the service location allocation problem. They consider the latency as the most critical effect on the QoS, so their objectives are minimizing the total cost and latency add them up into one tradeoff weighted model. They employ the dataset collected by WS-DREAM [29, 30] which describes real world evaluation results between end user and web services such as location and latency. In their first research [26], they propose a NSGA-II approach, and conduct full experiment on the comparison between NSGA-II and GA. The results show that NSGA-II perform effectively to produce near optimal solutions, NSGA-II is also more efficient for problem with big numbers of data, and GA only can tackle small problems. For their following research [27], they devote to the PSO approach with a set of larger experiment datasets. They initial propose a Weighted Sum PSO (WSPSO) since the fitness function is weighted sum up of all the objectives, then another approach is dominance-based fitness assignment schema PSO called Non-dominated PSO (NSPSO). The results show that both WSPSO and NSPSO perform more effective than NSGA-II, and NSPSO can perform better than WSPSO especially on large problems. Their next research [28] devote on a further research on a new PSO approach called Modified Binary PSO (MBPSO) which applied dynamic inertia weight. It returned results show that MBPSO has the best performance on larger problems than their previous approaches.

Paper	CI Technique	Request	Resource	Objectives
Bao et al. [23]	GA	Service	VM	QoS factors
Tan et al. [26–28]	GA and PSO	Service	Server	Cost, Latency

Table 3. Service Allocation Optimization

3.3 Cloud Broker based Resource Allocation

Cloud Broker is an intermediary between cloud users and cloud providers in CC that manages the workload allocation. Sometimes, clients can submit their requests to the cloud broker for multiple reasons. On one hand, cloud broker matches the requests with the best offers of cloud provider in order to maximize profit. On the other hand, cloud customers are satisfied with the better offerings, professional solutions and higher performance. In practice, the cloud brokers are responsible for efficient allocation of the customer requests into pre-booked resources in order to optimize the profit, often coupled with some other QoS requirements, such as response time, user satisfaction.

3.3.1 Cloud Brokering

The cloud community enthusiastically welcomed the concept of cloud brokering, since cloud brokering contributes many resource allocation problems and challenges to investigate and solve. Cloud brokering focuses on the development of brokering and multi-cloud platforms, and provider optimization solutions for the cloud users. From the perspective of resource allocation, a cloud broker normally act as a intermediary in the process of workload submission, that is, the process of allocating the requests from multiple cloud users to the offers from multiple cloud providers. The requests can be different based on the usage of cloud business model, i.e. application requests for SaaS, or VM requests for IaaS. The reviewed literatures about cloud brokering problem are concluded in Table 4.

Gaetano et al. [31] propose a GA approach for cloud brokering to find IaaS resources for optimizing the QoS satisfaction include cost, ram, storage, and location, the hypothesis or solution is described by a vector based representation called Chromosome which the index represents a request and the corresponding value represents the assigned VM. A set of experiments is performed for evaluating the validation, tuning and scalability of GA approach. However, the result shows that only cost objective is optimized while the ram and storage satisfaction are not optimized. But it still can be noticed that the cost-effectiveness for multiple providers.

Yacine et al. [32] solve cloud brokering problem by using Multi-Objectives Particle Swarm Optimization (MOPSO). They model the allocation problem by setting clients request and VM as the particle position, then produce a set of Pareto solutions. After comparison with NSGA-II and random search algorithm, the result shows that MOPSO achieve the best Pareto solution.

Some hybrid Algorithms are proposed to improve the effectiveness with a better global optimal result, such as, Max-Min Ant (MMAS) [33] + GA [34] and Simulated Annealing (SA) + GA [35]. For the MMAS + GA approach, the mapping between requests and VMs are modeled as a graph, where the set of nodes represents the requests and the set of edges describes the mapping, then applying the GA on the solutions generated by MMAS. SA + GA uses GA to model the mapping solution and parallelly employed SA as an operator to exploit the search space regions. So GA manages the evolution of a population, and a SA operator called migration is responsible for managing a set of different populations and exchanging individuals among the populations. Lotfi and Salah [34] devote a MMAS+GA for allocation of VM resources to satisfy the users request which is composed of a set of tasks to be executed, objectives of resource allocation are minimizing the execution time and cost. They set up the experiment based on the a benchmark dataset from research [36] which is a matrix mapping the execution time between different task types and VMs. They demonstrate the effectiveness and efficiency among GA, MMAS and MMAS+GA, the result shows that the hybrid approach has the best solution, the GA perform slightly inferior, and the MMAS doesnt dominate any solution. However, the MMAS+GA has the highest execution time, and the MMAS computes fastest. Santiago et al. [35] proposed another hybrid approach that combines the SA and GA in parallel to solve the cloud brokering problem by optimizing the profit and makespan. After a set of experiment with a fixed 90s execution time. Then they compare to a greedy approach on the performance, the profit result improved significantly with an up to 138% and average 17.8% to 43.3% improvement. In terms of makespan result, the improvement is between 4.5% and 17.7%.

Paper	CI Technique	Request	Resource	Objectives
Gaetano et al. [31]	GA	Application	VM	QoS satisfaction
Yacine et al. [32]	MOPSO	Internet of Things	VM	Response time, Energy
Lotfi and Salah [34]	MMAS + GA	Service	VM	Execution time, Cost
Santiago et al. [35]	SA + GA	VM Request	VM	Profit, Makespan

Table 4. Cloud Brokering Optimization

3.4 Summary

3.4.1 CI Approaches

Different CI Algorithms are proposed to solve the CC resource allocation problems with the expectation for efficiency and effectiveness, such as GA, PSO, SA and hybrid.

GA is a well known heuristic algorithm that can stochastic search near-optimal solutions in large search spaces iteratively. GA manages a set of solutions called chromosomes, then evolves the chromosomes during the iterations until reach the expected solution. For most resource allocation problems are mapping the requests to the resources, GA chromosomes can present the mapping clearly. GA also can solve multi objective problems by search a set of named non-dominated solutions [19, 26, 31]. However, one major limitation of GA is the time consuming since the a number of fitness calculation is

performed for each iteration.

PSO is a biologically inspired powerful optimization algorithm that mimics the social behaviors of bird flocking to search the solution [37]. PSO also uses vector-based solutions called particles, so the same mapping problem can be easily modeled by particle. PSO can solve multi objective problems by finding a set of Pareto optimal set [38]. Compared to GA, PSO is easier to setup with fewer parameters, while the configuration of GA parameter may take much more effort, hence PSO also has a faster optimization speed [39].

Many Hybrid Algorithms are constructed by adopting both advantages from different algorithms to improve the performance. A MMAS and a GA are combined by applying the MMAS solutions to GA [34], which gives better results with more execution time. A parallel hybrid EA + SA is proposed by using SA as an operator for exploiting promising search space regions [35], which produce better results in an affordable amount of time. Another hybrid algorithm is applying the CSO on the GA selected solutions than perform the GA evolution, the result shows that the hybrid performs better than GA and CSO for an online problem [16]. The authors [21] use a fuzzy rule to select the PSO inertia weights, which achieves a better global optimal.

3.4.2 Objectives

In summary, resource allocation objectives can be classified into three groups: *Performance, Financial* and *environmental* groups. The summary of objectives and corresponding relation of cloud actors are in the table as follows (See Table 5).

The standard objectives of resource allocation is performance related, cloud user often tried to optimize the response time for end user and the throughput for their services in order to provide a better QoS in this competitive market. These objectives can be further defined as makespan witch denotes to the maximum completion time of a batch of tasks. Another group of objectives are cost-related, on the cloud user perspective, cloud users are interested in providing a better QoS with a lower deployment cost, on the cloud provider hand, they aim to maximize their profit by minimizing the operation cost such as power and energy cost. Recently, some emerging objectives can be grouped into environmental objectives, witch driven by government policies. The main contribution is to minimize the environmental impact by minimizing the power usage and CO_2 emission. In addition, some other objectives are concerned by cloud providers when maintaining the data center, i.e. resource utilization and VM migration number are directly related to the maintenance cost.

Objective Groups	Objectives	Cloud Actors
Performance	Response Time, Throughput, Makespan	Cloud User, Cloud Broker
Financial	Cost, Profit, Price	Cloud User, Cloud Broker and Cloud Provider
Environmental	Power, CO_2 Emissions, Energy	Cloud Provider
Others	Resource Utilization, VM Migration Number	Cloud Provider

Table 5. Objective Classification

4 Challenges for Future Scope

CC is still not a mature field, and presents many opportunities and challenges for the research community. After reviewing recent researches, we identified some research challenges and gaps for the future scopes. These unexplored territories include: *auxiliary objectives, coherent cloud optimization simulator or framework, Hybrid algorithms, problem scalability*, and *cross-layer optimization*

4.1 Auxiliary Objectives

CC resource allocation optimization still presents many unexplored areas. Another interesting domain is auxiliary objectives for the optimization. The existing objectives can be divided into three groups, that is, performance, finance and environment. There is still a gap on auxiliary objectives resulting from shared and highly-distributed cloud environments, for example, reliability for allocating on heterogeneous infrastructural resources, availability for distributed resources, security for multi-tenancy distribution, and data storage for legal compliance.

4.2 Coherent Simulator and Framework

Despite lots of cross-layer CC resource allocation optimization problem has been studied, there is still no coherent simulator or framework to leverage the existing approaches. Some potential candidates can be CloudSim [40], CloudAn-

alyst [41], GreenCloud [42] and iCanCloud [43]. CloudSim enable seamless simulating of CC resource allocation from service application layer to infrastructure layer, researcher can model their problem without considering about low level infrastructure details. CloudAnalyst is a UI based simulator that enables researchers to evaluate the experiments of geographical distribution workload-server allocation problems. GreenCloud provides researchers a simulation environment for energy-aware CC data centers with a focus on communications. iCanCloud allows researchers to predict trade-off between cost and performance by executing applications in specific hardware.

4.3 Hybridization

Hybrid algorithms are very common in real-world optimization problems, there are still many more hybrid algorithms can be constructed in order to improve the performance. The combination can be embedded an algorithm in the middle of another algorithm [16], or just running after another algorithm [34]. Another hybrid approach is to adopt some specialized operators for another algorithm [35]. Most of hybrid approach can be easier when they share very similar solution encoding. However, this may introduce more challenges on the algorithm execution time due to the fact of complexity of hybrid algorithm.

4.4 Scalability

Another challenge for CI community is very large scaled optimization problem with more than 10,000 users. Even though these large scaled problems are only concerned by large cloud providers, such as Google and Amazon, some the real world problems with several thousands of users still need to be solved efficiently and quickly. Additionally, online problems normally constrain the solution execution time, so future algorithms should perform more efficiently while the accuracy is guaranteed.

4.5 Cross-layer optimization

There still some cross-layer CC problems can be considered in the future, that is, consider the problems cross multiple cloud layers. For example, some cloud providers may manages hardware, infrastructure, platform and software layer together and provide SaaS to the end users, they many encounter more challenges since they manages multiple layers and concerns more objectives. Research [35] solved a cloud brokering problem that optimally allocates VM requests to VM, and studies [26–28] also consider a service allocation problem between service and geographically distributed servers. A future study may propose a mechanism that combines these two problems into two steps, the first step optimizes the the allocation between VM and services, and the second step allocates the VM to geographically distributed server in order to minimize the latency.

5 Conclusion

This survey reviews a cross layer range of CC resource allocation optimization using CI techniques. Apparently, CI techniques are applicable to both single objective and multi objectives resource allocation problem that exists at all layers of CC. However, there still lots of gaps and challenges for the research community, include algorithm performance, hybridization, exploration. Firstly, the performance of CI approaches for heavy load resource allocation problems is still not good enough in terms of executing efficiently and adapting dynamically, and it is a big challenge to find a good solutions for resource allocation in CC within a time frame. Secondly, there are still lots of opportunities for approach hybridization. CI tools can be hybridized with machine learning or mathematical programming for different purpose, such as, improving the accuracy and increasing the adaptability. Finally, CC is still a developing area, the evolving may impact existing problems and introduce many new problems. For example, cross-layer resource allocation optimization and auxiliary objectives. In conclusion, considering these gaps in CC resource allocation optimization, there are still many opportunities for CI communities to explore.

References

[1] Dillon, T., Wu, C., and Chang, E., 2010. "Cloud computing: issues and challenges". In Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on, Ieee, pp. 27–33.

[2] Endo, P. T., de Almeida Palhares, A. V., Pereira, N. N., Goncalves, G. E., Sadok, D., Kelner, J., Melander, B., and Mangs, J.-E., 2011. "Resource allocation for distributed cloud: concepts and research challenges". *IEEE network,* **25**(4).

[3] Jennings, B., and Stadler, R., 2015. "Resource management in clouds: Survey and research challenges". *Journal of Network and Systems Management,* **23**(3), p. 567.

[4] Guzek, M., Bouvry, P., and Talbi, E.-G., 2015. "A survey of evolutionary computation for resource management of processing in cloud computing". *IEEE Computational Intelligence Magazine,* **10**(2), pp. 53–67.

[5] Zhang, Q., Cheng, L., and Boutaba, R., 2010. "Cloud computing: state-of-the-art and research challenges". *Journal of internet services and applications,* **1**(1), pp. 7–18.
[6] Jadeja, Y., and Modi, K., 2012. "Cloud computing-concepts, architecture and challenges". In Computing, Electronics and Electrical Technologies (ICCEET), 2012 International Conference on, IEEE, pp. 877–880.
[7] Rodríguez-Haro, F., Freitag, F., Navarro, L., Hernánchez-sánchez, E., Farías-Mendoza, N., Guerrero-Ibáñez, J. A., and González-Potes, A., 2012. "A summary of virtualization techniques". *Procedia Technology,* **3**, pp. 267–272.
[8] Foster, I., Zhao, Y., Raicu, I., and Lu, S., 2008. "Cloud computing and grid computing 360-degree compared". In Grid Computing Environments Workshop, 2008. GCE'08, Ieee, pp. 1–10.
[9] Parikh, S. M., 2013. "A survey on cloud computing resource allocation techniques". In Engineering (NUiCONE), 2013 Nirma University International Conference on, IEEE, pp. 1–5.
[10] Vaquero, L. M., Rodero-Merino, L., Caceres, J., and Lindner, M., 2008. "A break in the clouds: towards a cloud definition". *ACM SIGCOMM Computer Communication Review,* **39**(1), pp. 50–55.
[11] Gao, C., Wang, H., Zhai, L., Gao, Y., and Yi, S., 2016. "An energy-aware ant colony algorithm for network-aware virtual machine placement in cloud computing". In Parallel and Distributed Systems (ICPADS), 2016 IEEE 22nd International Conference on, IEEE, pp. 669–676.
[12] Portaluri, G., and Giordano, S., 2016. "Multi objective virtual machine allocation in cloud data centers". In Cloud Networking (Cloudnet), 2016 5th IEEE International Conference on, IEEE, pp. 107–112.
[13] , 2010. "Efficient modeling in ilog opl cplex development system". In *White Paper*.
[14] Pires, F. L., and Barán, B., 2013. "Multi-objective virtual machine placement with service level agreement: A memetic algorithm approach". In Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing, IEEE Computer Society, pp. 203–210.
[15] Ahvar, E., Ahvar, S., Mann, Z. A., Crespi, N., Garcia-Alfaro, J., and Glitho, R., 2016. "Cacev: a cost and carbon emission-efficient virtual machine placement method for green distributed clouds". In Services Computing (SCC), 2016 IEEE International Conference on, IEEE, pp. 275–282.
[16] Sharma, N. K., and Guddeti, R. M. R., 2016. "On demand virtual machine allocation and migration at cloud data center using hybrid of cat swarm optimization and genetic algorithm". In Eco-friendly Computing and Communication Systems (ICECCS), 2016 Fifth International Conference on, IEEE, pp. 27–32.
[17] Sundararajan, P. K., Feller, E., Forgeat, J., and Mengshoel, O. J., 2015. "A constrained genetic algorithm for rebalancing of services in cloud data centers". In Cloud Computing (CLOUD), 2015 IEEE 8th International Conference on, IEEE, pp. 653–660.
[18] Park, K., and Pai, V. S., 2006. "Comon: a mostly-scalable monitoring system for planetlab". *ACM SIGOPS Operating Systems Review,* **40**(1), pp. 65–74.
[19] Zhao, J., Zeng, W., Liu, M., and Li, G., 2011. "Multi-objective optimization model of virtual resources scheduling under cloud computing and it's solution". In Cloud and Service Computing (CSC), 2011 International Conference on, IEEE, pp. 185–190.
[20] Sotomayor, B., Montero, R. S., Llorente, I. M., and Foster, I., 2009. "Virtual infrastructure management in private and hybrid clouds". *IEEE Internet computing,* **13**(5).
[21] Ramezani, F., Naderpour, M., and Lu, J., 2016. "A multi-objective optimization model for virtual machine mapping in cloud data centres". In Fuzzy Systems (FUZZ-IEEE), 2016 IEEE International Conference on, IEEE, pp. 1259–1265.
[22] Aboolian, R., Sun, Y., and Koehler, G. J., 2009. "A location–allocation problem for a web services provider in a competitive market". *European Journal of Operational Research,* **194**(1), pp. 64–77.
[23] Bao, L., Zhao, F., Shen, M., Qi, Y., and Chen, P., 2016. "An orthogonal genetic algorithm for qos-aware service composition". *The Computer Journal,* **59**(12), pp. 1857–1871.
[24] Al-Masri, E., and Mahmoud, Q. H., 2007. "Discovering the best web service". In Proceedings of the 16th international conference on World Wide Web, ACM, pp. 1257–1258.
[25] Al-Masri, E., and Mahmoud, Q. H., 2008. "Investigating web services on the world wide web". In Proceedings of the 17th international conference on World Wide Web, ACM, pp. 795–804.
[26] Tan, B., Ma, H., and Zhang, M., 2016. "Optimization of location allocation of web services using a modified non-dominated sorting genetic algorithm". In Australasian Conference on Artificial Life and Computational Intelligence, Springer, pp. 246–257.
[27] Tan, B., Mei, Y., Ma, H., and Zhang, M., 2016. "Particle swarm optimization for multi-objective web service location allocation". In European Conference on Evolutionary Computation in Combinatorial Optimization, Springer, pp. 219–234.
[28] Tan, B., Huang, H., Ma, H., and Zhang, M., 2017. "Binary pso for web service location-allocation". In Australasian Conference on Artificial Life and Computational Intelligence, Springer, pp. 366–377.
[29] Zheng, Z., Zhang, Y., and Lyu, M. R., 2014. "Investigating qos of real-world web services". *IEEE Transactions on Services Computing,* **7**(1), pp. 32–39.
[30] Zheng, Z., Zhang, Y., and Lyu, M. R., 2010. "Distributed qos evaluation for real-world web services". In Web Services

(ICWS), 2010 IEEE International Conference on, IEEE, pp. 83–90.

[31] Anastasi, G. F., Carlini, E., Coppola, M., and Dazzi, P., 2014. "Qbrokage: A genetic approach for qos cloud brokering". In Cloud Computing (CLOUD), 2014 IEEE 7th International Conference on, IEEE, pp. 304–311.

[32] Kessaci, Y., Melab, N., and Talbi, E.-G., 2013. "A pareto-based genetic algorithm for optimized assignment of vm requests on a cloud brokering environment". In Evolutionary Computation (CEC), 2013 IEEE Congress on, IEEE, pp. 2496–2503.

[33] Stützle, T., and Hoos, H. H., 2000. "Max–min ant system". *Future generation computer systems,* **16**(8), pp. 889–914.

[34] Hajjem, L., and Benabdallah, S., 2016. "An mmas-ga for resource allocation in multi-cloud systems". In Internet Technology and Secured Transactions (ICITST), 2016 11th International Conference for, IEEE, pp. 421–426.

[35] Iturriaga, S., Nesmachnow, S., Dorronsoro, B., Talbi, E.-G., and Bouvry, P., 2013. "A parallel hybrid evolutionary algorithm for the optimization of broker virtual machines subletting in cloud systems". In P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2013 Eighth International Conference on, IEEE, pp. 594–599.

[36] HCSP, 2000. "Heterogeneous computing scheduling problem". In *https://www.fing.edu.uy/inco/grupos/cecal/hpc/HCSP/index.htm*.

[37] Kennedy, J., 2011. "Particle swarm optimization". In *Encyclopedia of machine learning*. Springer, pp. 760–766.

[38] Reyes-Sierra, M., and Coello, C. C., 2006. "Multi-objective particle swarm optimizers: A survey of the state-of-the-art". *International journal of computational intelligence research,* **2**(3), pp. 287–308.

[39] Bai, Q., 2010. "Analysis of particle swarm optimization algorithm". *Computer and information science,* **3**(1), p. 180.

[40] Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A., and Buyya, R., 2011. "Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms". *Software: Practice and experience,* **41**(1), pp. 23–50.

[41] Wickremasinghe, B., Calheiros, R. N., and Buyya, R., 2010. "Cloudanalyst: A cloudsim-based visual modeller for analysing cloud computing environments and applications". In Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on, IEEE, pp. 446–452.

[42] Kliazovich, D., Bouvry, P., Audzevich, Y., and Khan, S. U., 2010. "Greencloud: a packet-level simulator of energy-aware cloud computing data centers". In Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE, IEEE, pp. 1–5.

[43] Núñez, A., Vázquez-Poletti, J. L., Caminero, A. C., Castañé, G. G., Carretero, J., and Llorente, I. M., 2012. "icancloud: A flexible and scalable cloud infrastructure simulator". *Journal of Grid Computing,* **10**(1), pp. 185–209.