VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wānanga o te Ūpoko o te Ika a Māui

School of Engineering and Computer Science
Te Kura Mātai Pūkaha, Pūrorohiko

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341
Fax: +64 4 463 5045
Internet: office@ecs.vuw.ac.nz

Energy-efficient Server Consolidation in Container-based Clouds with EC approaches

Boxiong Tan

Supervisors: Dr.Hui Ma, Dr.Yi Mei, and Prof. Mengjie Zhang

Submitted in partial fulfilment of the requirements for PhD.

Abstract

Container-based Cloud is a new trend in Cloud computing that introduces finer granularity management of applications and reduce overhead of virtual machine (VM). Compared with VM-based Cloud, container-based Cloud can further improve the energy efficiency with a finer granularity of server consolidation in data centers. Current VM-based single level of server consolidation cannot be used in container-based cloud because the container-based cloud has two levels of placement: container to VM and VM to PM. Existing research lacks energy model and optimization algorithms that consider the joint allocation of container and VM. This work aims to improve energy efficiency in container-based cloud by proposing a bilevel energy model and three Evolutionary Computation (EC)-based optimization algorithms for three placement decision scenarios: initial placement of application, periodic placement of application, and dynamic placement of application. The novel bilevel energy model and three EC algorithms contribute to a better manage of resources for energy efficiency in container-based cloud.

Contents

Figures

Chapter 1

Introduction

This chapter introduces this research proposal. It starts with the problem statement, then outlines the motivations, research goals and the organisation of this proposal.

1.1 Problem statement

Cloud computing has made a huge impact on modern software industry by offering on-demand computing capacity (e.g storage and computing) [19]. Compare with traditional software industry that applications run on individual hardware, on-demand cloud computing provides convinent architectures for software industry. For example, web service providers such as Google and Neflix deploy their applications on cloud. These web service providers do not need to purchase and maintain hardware resources. In addition, these web service providers do not need to worry about the scalability issue (e.g. dynamically increases the capacity of application) and availability of applications (e.g services are accessible 99.99% of the time) when their applications gradually increase [2]. Moreover, application users can enjoy applications without experiencing breakdown and access the applications from anywhere in the world.

A major issue in cloud computing is the huge energy consumption generated by data centers– a typical data center consumes as much energy as 25,000 households [26]. Huge energy consumption has become the major expense of cloud providers. The reduction of energy bill will be further beneficial to cloud providers and web service providers. Furthermore, people will pay less to applications on the cloud.

Generally, reducing the energy consumption in cloud depends on reducing the number of live physical machines (PMs) (e.g. servers). Studies shows [7, 84], PMs account for the majority – more than 40% – of energy consumption in cloud among other components such as cooling systems, PMs, and network devices. Moreover, these PMs has been not used effectively. Some studies analyzed that the proportion of PMs' average utilization is quite low - from 10% to 50% [42] Therefore, it is needed to reduce energy by improving the utilization of PMs and reducing the usage of PMs in cloud.

The common way to improve the utilization of PMs in cloud is through resource management of PMs [64] (see Figure 1.1). A centralized resource management system in cloud has two main functionalities. First, the management system allocates resources such as CPUs and memories of PMs for cloud users to run applications. Second, the management system handles the workload fluctuations to reduce the potential migration. These two main functionalities deploy and maintain applications in cloud.

The two main functionalities of resource management in cloud involve four steps. First, the management system collects the utilization information of PMs and then analyze the

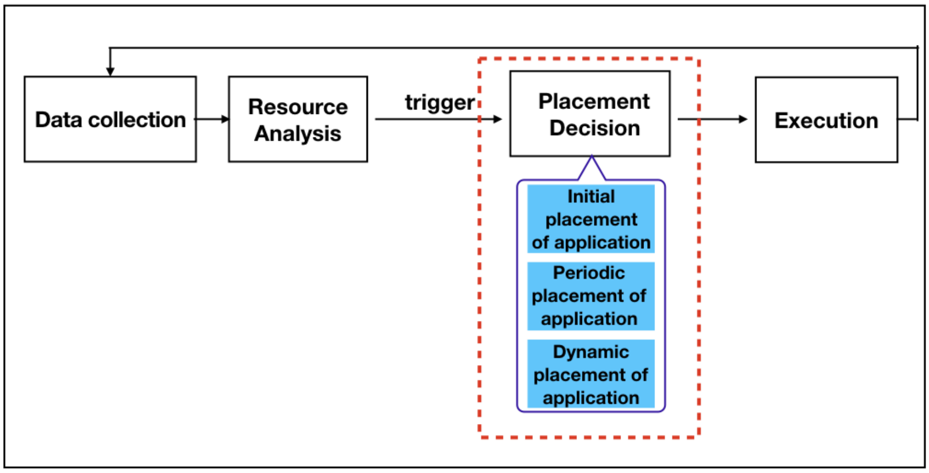


Figure 1.1: A workflow of resource management [68]

usage of PMs and the needed resource of applications. Next, triggered by resource analysis, the management system determines the placement of applications. The placement of applications includes three scenarios: initial placement of application for new applications; periodic placement of application for adjusting the placement periodically; and dynamic placement of application for adjusting applications in a fast manner when abnormal events such as overloading happened. Finally, the management system executes the placement decision of applications to PMs. Hence, better management of resources in cloud contributes towards a better utilization of PMs and thus a reduction of energy consumption.

Server consolidation [99] is a popular strategy in resource management in cloud. Two different types of server consolidation are used in cloud: static and dynamic. Static server consolidation manages resources in an off-line fashion, which is mainly used in the initial placement of applications and periodic placement of applications. Dynamic server consolidation manages resources in an on-line fashion, which is used in the dynamic placement of applications. Both static and dynamic server consolidation aim to place applications in fewer PMs that leads to higher utilization of PMs and lower energy consumption.

Currently, server consolidation in cloud is based on *virtualization* technology [97]. Such virtualization separates the resources (e.g. CPUs and RAMs) of a PM into several parts called *virtual machines (VMs)*. Each VM runs an isolated operating system. This VM-based technology is much different from a traditional cloud that places each application to a single PM and leads to the low reserved utilization of PMs. Compared with traditional clouds, current VM-based clouds significantly improve the utilization of PMs and reduces the energy consumption.

However, in recent years, resource management with VMs cannot catch up with a new trend in software industry – Service Oriented Architecture (SOA) [90]. This SOA has widely used in modern software industry because of its agility and re-usability [90]. SOA separates a centralized application into multiple distributed components called web services. As web services only require a small amount of resources (e.g. 15% of CPU), using a VM for a web service causes resource wastage inside a VM. Consequently, the low utilization of PMs decreases the energy efficiency.

To support the architecture of SOA and further reduce energy consumption, a new virtualization technology: containers [36,88] has been proposed. Containers run on top of VMs called an operating system (OS) level of virtualization [88]. Similar to VM, a container provides performance and resource isolation for a single application. Different to VMs, multiple containers can run in the same VM without interfering each other. In addition, containers naturally support vertical scaling (change its size during runtime) [98]. The vertical scaling provides resilient resources to fluctuate workloads. The container technology provides a new architecture for allocating applications and a finer granularity of resource management. Hence, Containers has the potential to further improve the energy consumption.

Although the efficient use of containers can improve the utilization of VMs, containers bring new challenges and difficulties to server consolidation [1]. We cannot directly apply current VM-based server consolidation strategies on container-based cloud because of the different placement structures of application. Moreover, to support the increasing size of containers, the vertical scaling in cloud requires the VMs to reserve sufficient resources. The interaction between VMs and containers changes the server consolidation into a bilevel problem: VMs-to-PMs, and VMs-to-containers. Bilevel problems are NP-hard [86]. NP-hard means the problem is difficult to find a polynomial algorithm to find the global optimum solution. Therefore, we need to develop better algorithms for the bilevel problem.

This research aims at improving the energy efficiency in container-based cloud by proposing new bilevel energy models and server consolidation algorithms for three placement decision scenarios: initial placement of application, periodic placement of application, and dynamic placement of application.

1.2 Motivation

Container-based Cloud is a promising technology to support SOA, the new trend in software industry. Container-based cloud provides a finer granularity management of applications compared with VM-based Cloud [1]. Such finer granularity management improves the utilization of resources and minimizes energy consumption of data centers. Therefore, container-based cloud not only increases the profit for cloud providers but reduces the cost of cloud users as well.

Even though container-based cloud has many advantages to achieve energy-efficiency, it is more challenging in optimizing the placement of application than in VM-based cloud. **First**, we cannot use existing VM-based energy models for the container-based cloud. Current VM-based energy models represent a single level of placement: VM-PM while the container-based cloud needs a two-level energy model representing the placement between container-VM and the placement between VM-PM. However, the two-level energy model is more complicated than the single-level energy model. **Second**, current VM-based optimization approaches [9, 61] formulate the placement of application as a bin packing problem. Most VM-based approaches use bin-packing heuristics such as First Fit [30] that are designed for the problem of single-level placement. Since the container-based cloud has two-level placement, bin-packing heuristics can rarely achieve the global optimal solution in container-based cloud. **Third**, in recent years, Evolutionary Computation (EC)-based approaches have been applied to bilevel problems in other areas such as logistics distribution centers and have shown promising results [3, 28, 87]. However, EC-based approaches have not been used to solve the placement of application in container-based cloud. Therefore, we need to study how to design specific genetic operators and representations for adapting EC algorithms to container-based cloud. The above three issues exist in all three main scenarios of placement in container-based cloud.

Besides the above issues, other specific issues exist in three main scenarios of placement of application. **Initial placement of application** deploys applications when data centers receive a number of requests. We need to investigate a scalable EC-based optimization algorithm to further minimize the energy consumption when placing over one thousand of

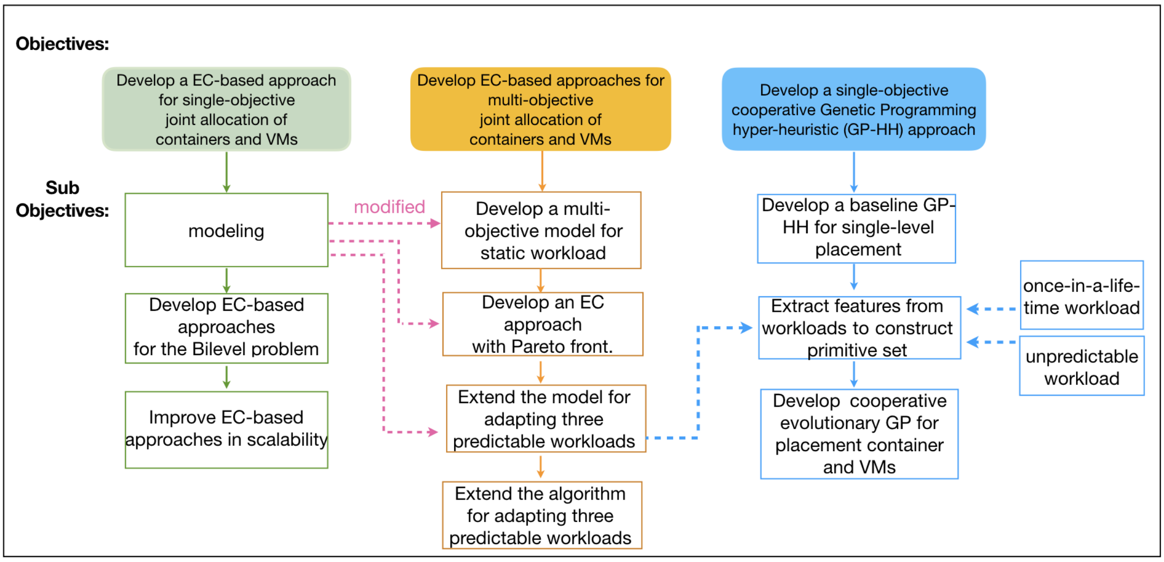


Figure 1.2: Relationship between objectives

applications in PMs. **Periodic placement of application** migrates applications to fewer PMs to optimize energy consumption and migration cost with consideration of workload. We need to take workload into account in a new model for multi-objective (energy and migration cost) optimization. Meanwhile, we will investigate how to design an EC-based approach to solve the multi-objective bilevel placement. **Dynamic placement of application** is capable of dynamically allocating applications when abnormal events such as overloading and under-loading [10] happen in data centers. Unlike the previous two scenarios, dynamic placement requires fast allocation for optimizing energy consumption. A genetic programming [4] hyper-heuristic (GP-HH) is a promising technology that has been used in many dynamic problems such as dynamic job shop scheduling [70]. GP-HH is based on learning patterns from previous good solutions and can automatically generate heuristics. Hence, GP-HH can allocate applications quickly and optimally in terms of energy consumption. However, the complex learning patterns from previous good solutions are challenging in dynamic placement because different types of workload exist and some of them are unpredictable. In summary, all these issues are critical and need solving in order to achieve energy efficiency in container-based cloud. Therefore, we need to design bilevel energy models and applying EC-based optimization algorithms for the placement of application in container-based cloud.

1.3 Research Goals

The overall goal of this research is to optimize energy consumption of a container-based Cloud data center using EC-based approaches for three placement decision scenarios: initial placement of application, periodic placement of application, and dynamic placement of application. The specific research objectives of this work can be itemized as follows.

1.3.1 Objective One: Develop EC-based approaches for the single objective joint placement of containers and VMs for initial placement of application

The goal of objective one is to minimize energy consumption in initial placement of application at container-based cloud. We set three sub objectives to achieve this goal. The first sub objective is to propose a new bilevel energy model for the joint placement of containers and VMs in PMs. The second sub objective is to develop an EC-based optimization algorithm to solve the joint placement of containers and VMs. The third sub objective improves scalability of the proposed EC-based optimization algorithm. Note that, we will use "the bilevel model/problem" to replace "the joint placement of containers and VMs model/problem" in the following content.

1. Develop a new bilevel energy model to represent the relationship between five factors and energy consumption. The five factors involve locations of container, types of VM, locations of VM, overheads of VM, and the balance between memory and CPU. We need to consider the interaction between these factors.

   The major challenge of this sub-objective is that the bilevel energy model is more complicated than a single-level energy model. **First**, in a VM-based model, the only factor is the locations of VMs in PMs. In contrast, in a container-based model, the energy model involves four more variables. Specifically, locations of container decide the utilization of VMs. Types of VM constrain the placement of container. Overheads of VM bring extra workload to the PMs. The balance of CPU and memory may indirectly impact on the energy consumption. Mishra [92] finds better balance leads to a higher probability of allocating more applications. The above mentioned five factors are very likely to affect the energy consumption. **Second**, two pairs of interaction have not been considered. The first pair is the interaction between containers and types of VM. The second pair is the interaction between locations of container and locations of VM. Therefore, the bilevel energy model is very complicated.

   In order to establish a bilevel energy model, we will follow three steps and gradually add factors to the energy model. **First**, we will formulate the relationship between overheads, types of VM, and numbers of VM. **Second**, we will formulate the balance between CPU and memory in both VM and PM levels. **Third**, we will formulate the energy consumption with container, VM, and PM [37,41,113].

   The sub objective is expected to formulate the bilevel energy model to represent the relationship between five factors.

2. Propose a new EC based bilevel optimization approach to solve the initial placement of application to achieve better energy efficiency than VM-based approaches.

   We need to solve two challenges. **The first challenge** is to design a representation of multiple variables for EC-based optimization approach. Current VM-based representation only contains one variable while container-based representation contains three variables: locations of container, locations of VM and types of VM. In addition, it is difficult to design a representation that narrows down the search space of solutions. **The second challenge** is to design genetic operators and search mechanisms for an EC-based optimization approach. Since bilevel optimization problem is known as a NP-hard problem [65], the landscape of search space is ragged because of non-linearity, non-differentiability, non-convexity etc. Therefore, we need to further explore an effective search mechanism that can quickly locate feasible solutions.

   In order to solve these challenges, we can explore some existing approaches. For the representation, direct binary representation [113] and indirect continuous probability representation [112] are often used in VM-based approaches. Based on different forms of representation, we can investigate a few EC-based bilevel algorithms such as Nested Particle Swarm Optimization [59], Differential evolution (DE) based approach [3,116] and Co-evolutionary approach [57].

We will evaluate our approaches in two ways. **First**, in order to find the most suitable representation and EC-based algorithm, we will conduct experiments on different approaches and compare their performance in terms of energy efficiency. **Second**, we will compare our algorithm with existing VM-based approaches on the same benchmark dataset.

The sub objective is expected to propose an EC-based approach for the initial placement of application. We expect the proposed EC-based approach can achieve better energy efficiency than existing VM-based approaches [106, 113].

3. Improve the scalability of the proposed EC-based approach up to one thousand applications.

   The major challenge of this task is the complexity of bilevel problems and the long computation time of EC-based approaches [87].

   We can use three methods to improve the scalability. **First**, we can reduce the number of variables by combining small containers into groups. We can use clustering approaches such as K-means [111] and decision tree to group containers. **Second**, we can separate a large question into several small questions so that they can be executed in parallel. The divide-and-conquer approach is a possible way but we need to investigate how to split the question. **Third**, we can design a hybrid EC-based algorithm mixed with heuristics. The hybrid algorithm can quickly locate the feasible solutions. For example, we can embed the representation with a First Fit heuristic. We will investigate how to design a heuristic and how the heuristics cooperate with the EC-based approach.

   We will evaluate our approach by comparing energy efficiency and execution time with the previously proposed EC-based approach.

   The sub objective is expected to improve the scalability of the previously proposed EC-based approach up to one thousand applications without decreasing the energy efficiency.

1.3.2 Objective Two: Develop an EC-based approach for the multi-objective periodic placement of application

The goal of this objective is to develop a multi-objective EC-based approach for periodic placement of application at container-based cloud with consideration of various types of workload. Two objectives are minimizing the energy consumption and minimizing the migration cost. We set four sub objectives to achieve this goal. The first sub objective proposes a multi-objective bilevel model for periodic placement of application taking into consideration with static workload. The second sub objective proposes an EC-based multi-objective algorithm with Pareto front approach. The third sub objective extends the previous multi-objective model with consideration of three types of workload: static, continuously increasing, and periodic. The fourth sub objective extends the previous EC-based multi-objective algorithm to adapt to various workloads.

1. Propose a multi-objective bilevel model for periodic placement with consideration of static workload. Two objectives are minimizing the energy consumption and minimizing the migration cost.

   The main challenge is that the additional migration model in container-based cloud is much complicated than VM-based cloud. **First**, in VM-based cloud, many research only considers the number of migrations [69]. In container-based cloud, because the sizes of container vary in a wide range (e.g from MBs to GBs), it is unfair to simplify the migration cost as the number of migrations. Therefore, we must consider more factors such as network bandwidth in modeling the migration cost which makes the model difficult. **Second**, container-based migration model not only considers migration of VM, but also migration of container. Therefore, adding the migration model into our previous bilevel model is difficult.

In order to solve these challenges, **first**, we will explore some VM-based migration models and investigate the relationship between network bandwidth and utilization of resources in VM. **Second**, we will formulate the relationship between migration cost and the utilization of resources in two levels: VM and PM.

The sub objective is expected to formulate the bilevel model to represent both energy consumption and migration cost in container-based cloud.

2. Propose a multi-objective bilevel EC-based algorithm for periodic placement of application with Pareto front approach to achieve better performance than VM-based approach in both energy efficiency and migration cost.

   We need to solve three challenges. **First**, the relationship between the two optimization objectives is very complicated. Since both migration cost and energy consumption can be non-convex, we cannot easily imagine the trade-off between two objectives. **Second**, we must design genetic operators to adapt bilevel multi-objective optimization. Specifically, genetic operators in both levels collaborate with each other. Hence, it is more difficult to design the collaboration of operators between two levels. **Third**, after obtaining a set of non-dominated solutions, we need to further design a fitness function assessment. The assessment selects one solution from a large number of non-dominated solutions based on some predefined preferences from Cloud providers. The selection of solution can be a difficult task for the large size of non-dominated set [118].

   We will follow three steps to solve the above challenges. **First**, to study the relationship between two objectives, we may visualize the relationship by using a technique called Trade-off region map [75]. The trade-off region map can identify local and complex relationships between objectives, gaps in the fitness landscape, and regions of interest. **Second**, for designing genetic operators, based on previously designed representation, we will add trade-off control mechanisms such as non-dominated sorting into the genetic operators. **Third**, in order to design a fitness function assessment, we will review some priori and posteriori decision making support methods such as Guided Multi-objective Genetic Algorithm [46] and a clustering approach [117].

   We will evaluate our algorithm by comparing with existing VM-based approaches on the same benchmark dataset.

   The sub objective is expected to propose a multi-objective bilevel EC-based approach that can achieve better energy efficiency and migration cost than VM-based approaches.

3. Extend the bilevel multi-objective model for adapting three types of predictable workload [34]: static, linear continuously changing, and periodic.

   The major challenge for this objective is the complexity of modeling three types of workload. Even though we can roughly categorize workloads into three types, practically, workloads are still full of uncertainty and noise.

We may try several ways to represent workloads uniformly. **First**, we may consider modeling workload as a function of time. **Second**, we may apply time-series analysis techniques on workloads so that three types of workload can be represented as their features.

The sub objective is expected to extend previous bilevel multi-objective model to adapt to three types of workload.

4. Extend the previous EC-based multi-objective algorithm to adapt to various of workload.

   The major challenge for this objective is to design representations and genetic operators to adapt to three types of workload. Different representations may be suitable for different types of workload. Therefore, it is possible to design three distinct representations for three workloads. In addition, we must design search mechanisms to collaborate with the representations.

   In order to find the most suitable representation and EC-based algorithm, we will conduct experiments on different approaches and compare their performance in terms of energy efficiency.

   The sub objective is expected to extend a multi-objective bilevel EC-based approach to three types of workload.

1.3.3 Objective Three: Develop a hyper-heuristic single-objective Cooperative Genetic Programming Hyper-heuristic (GP-HH) approach for automatically generating dispatching rules for dynamic placement of application.

The goal of this objective is to develop a cooperative GP-based hyper-heuristic algorithm so that the generated dispatching rules can achieve both fast placement and global optimization with five types of workloads. We set three sub objectives to achieve this goal step by step. The first sub objective proposes a GP-based hyper-heuristic (GP-HH) algorithm for automatically generating dispatching rules for the single-level placement: container to VM with static workload. The second sub objective extracts features on the predictable workloads and two additional unpredictable workloads to construct a new primitive set. The third sub objective develops a cooperative GP-HH approach for automatically generating dispatching rules for placing both containers and VMs.

1. Develop a GP-based hyper-heuristic (GP-HH) algorithm for automatically generating dispatching rules for the single-level placement: container to VM with static workload. In order to develop a cooperative GP-based hyper-heuristic to the bilevel problem, it is necessary to develop a GP-HH for the single-level placement first. Then, two GP-HH can collaborate with each other.

   Two major challenges, **first**, it is difficult to design a hyper-heuristics to generate dispatching rules. We must explore meaningful features of static workload, VM, and PM to construct a primitive set. It is uncertain which feature can contribute to the dispatching rule. **Second**, it is even difficult for the generated dispatching rules to reach a global optimal solution because of the trade-off between training accuracy and overfitting.

   To gradually solve above challenges, **first**, we may start from considering the features such as the status of VMs (e.g resource utilization), features of workloads (e.g resource requirement) that will affect the placement decision. The above features are represented as resource utilization. We will construct a primitive set with the selected features. **Second**, we will use the default function set and the basic tree representation for constructing the dispatching rules. **Third**, to train the GP-HH, we will use the solutions from the first objective as the learning sets. We may use ten folds cross-validation to prevent from over-fitting.

In order to evaluate the automatically generated heuristics, we will use a widely used simulator called CloudSim [20]. Since our proposed algorithm is equivalent to the VM-based placement algorithm because they both focus on single-level of placement. We will compare our heuristic to a highly cited work [8] from Beloglazov who proposes a Best Fit Decreasing heuristic for the energy consumption problem.

The sub objective is expected to propose a GP-HH algorithm for automatically generating dispatching rules for the single-level placement with static workload. The generated dispatching rules are expected to achieve equal or better performance than manually designed heuristics in terms of energy efficiency.

2. Conduct feature extraction on the three predictable workloads and two unpredictable workloads to construct a new primitive set.

   The major challenge is to discover useful features that can represent various workloads. The first challenge is to identify features from high dimensionality of datasets. The raw workload dataset may contain a large number of data points. It is difficult to extract meaningful features from the data points. The second challenge is to select useful features from a large number of raw features. It is very time consuming to manually examine raw features. The correlations between these raw features bring additional complexity.

   To solve these challenges, **first**, we will review a number of dimensionality reduction techniques. Some possible techniques are sampling, extrema extraction. **Second**, we may employ the principle component analysis (PCA) technique on the features set to reduce the number of features.

   Classification of various workloads is one of the way to test the extracted features. If we can differentiate workloads according to extracted features with a high accuracy (e.g. 95%), the features can be used in generating new dispatching rules.

   The sub objective is expected to extract a number of features to represent different types of workload. We can apply these features to construct a new primitive set so that the dispatching rules can deal with all kinds of workload.

3. Develop a cooperative GP-HH approach for automatically generating dispatching rules for placing both containers and VMs.
   The major challenge is to develop a cooperating procedure between two GP-HHs on two levels. As the first step, we will consider using two sub-populations to represent two-levels of placement. Each sub-population would be trained to minimize their objectives. The second step,the subpopulation in the container-VM level will be used in training the VM-PM level of placement.

   This sub objective is expected to propose a cooperative GP-HH approach for automatically generating dispatching rules for dynamic placement in container-based cloud.

1.4 Published Papers

During the initial stage of this research, some investigation was carried out on the model of container-based server consolidation [95].

1. Tan, B., Ma, H., Mei, Y. and Zhang, M., "A NSGA-II-based Approach for Web Service Resource Allocation On Cloud". *Proceedings of 2017 IEEE Congress on Evolutioanry Computation (CEC2017).* Donostia, Spain. 5-8 June, 2017.pp.2574-2581

1.5 Organisation of Proposal

The remainder of the proposal is organised as follows: Chapter 2 provides a fundamental background of the resource management in cloud data centers and its the energy consumption problem. It also conducts a literature review covering a range of works in this field; Chapter 3 discusses the preliminary work carried out to explore the techniques and EC-based techniques for the joint allocation of container and VMs; Chapter 4 presents a plan detailing this projects intended contributions, a project timeline, and a thesis outline.

Chapter 2

Literature Review

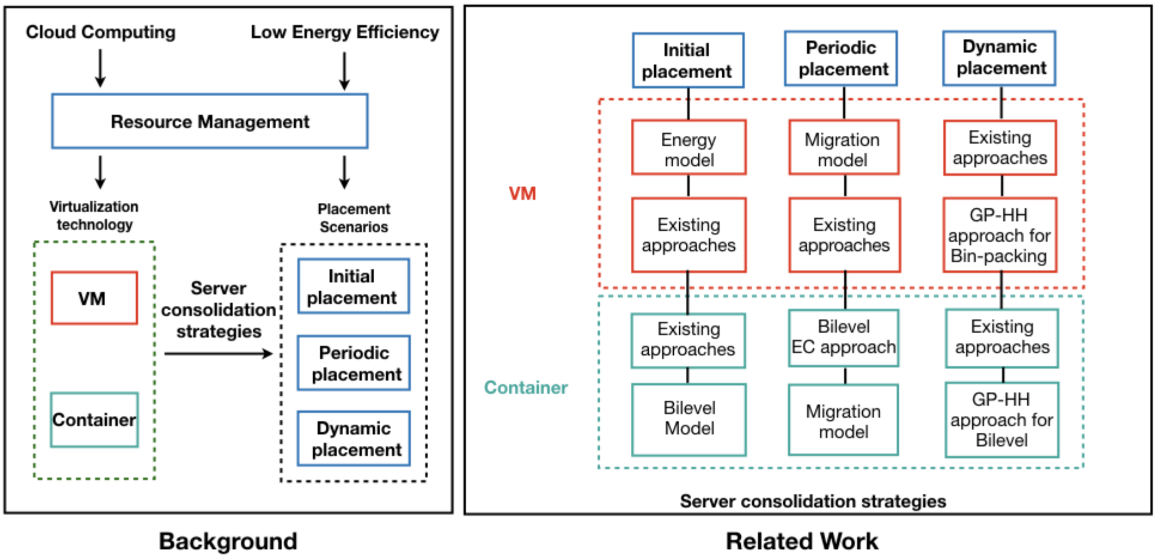


Figure 2.1: Scope of the literature review

The chapter introduces the background of resource management in cloud computing and the related work of server consolidation strategies related to three placement decision scenarios (see Figure 2.1). **Background** introduces Cloud computing, energy efficiency problem and resource management in cloud. We also explain virtualization technologies with three common placement scenarios and corresponding server consolidation strategies. **Related Work** discusses server consolidation strategies in terms of three placement scenarios with two virtualization technologies, VMs and containers.

2.1 Cloud Computing and Energy Efficiency Problem

Cloud computing is a computing model that offers a network of servers to their clients in a on-demand fashion. From NIST's definition [66], *"cloud computing is a model for enabling ubiquitous, 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 effort or service provider interaction."*

To give an example of how Cloud computing works (see Figure 2.2), consider the case: a *Cloud provider* builds a data center contained thousands of servers. These servers connect

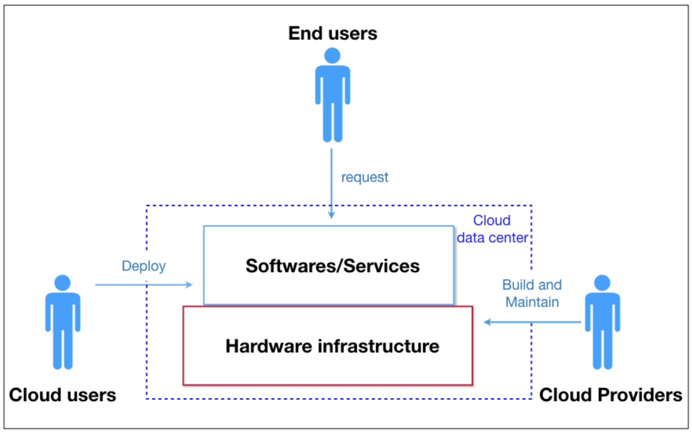


Figure 2.2: Stakeholders of Cloud computing [47]

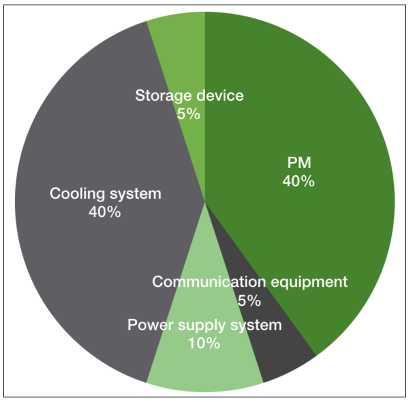


Figure 2.3: Energy consumption distribution of data centers [82]

with network. To use these remote servers, a web-based application provider is also a *Cloud user*, can deploy and access their applications (e.g Endnote, Google Drive and etc.) in these servers from anywhere in the world. Once the applications start serving, *End users* can use them without installing on their local computers. Cloud providers charge fees from Cloud users for infrastructure. Cloud users charge fees from End users for using applications. Therefore, from cloud providers' perspective, both accommodating more applications and reducing energy consumption lead to the increase of profit.

The major expense of a cloud provider is energy consumption [49] and physical machines (PMs) (e.g servers) contribute a majority of the energy. As shown in Figure 2.3, both cooling system and physical machines (PMs) account for 40% of energy. However, PMs' energy efficiency are low on average [42]. The reason for low energy efficiency is the disproportionate between the utilization of PMs and the energy consumption of PMs. For example, when CPU utilization of a PM is 15%, the energy consumption of the PM is 70% of its peak time. Therefore, cloud providers can reduce the energy consumption by improving the utilization of PMs.

In order to solve the low energy efficiency caused by low utilization of PMs, cloud providers use a resource management system to manage the applications.

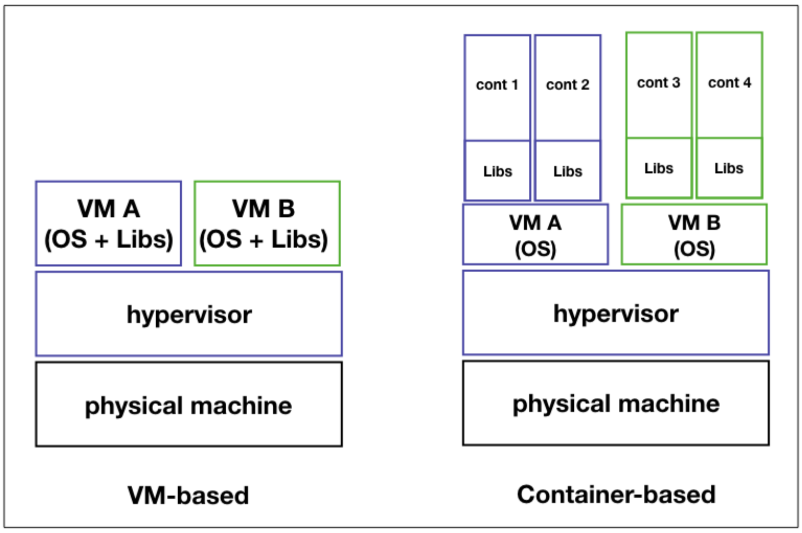


Figure 2.4: A comparison between VM-based and Container-based virtualization [76]

2.2 Cloud Resource Management

Cloud resource management is a process of allocating computation resources (e.g CPU and memory) to Cloud users to run their applications. Meanwhile, resource management aims to achieve low energy consumption in cloud [47]. Resource management applies two types of virtualization: virtual machine (VM) and container on four management steps (see Section 1.1): collecting PMs' utilization data, analyzing available PMs, deciding the placement of applications, and executing the placement. Furthermore, in the third step of placement decision, resource management has three common scenarios: initial placement, periodic placement, and dynamic placement. Resource management uses distinct server consolidation strategies on these placement scenarios based on different virtualiazation technologies to achieve energy efficiency.

This section will first introduce and compare two types of virtualization: container-based and VM-based and then illustrate three placement decision scenarios.

2.2.1 Virtualization Technologies

Resource management uses virtualization technologies [97] to achieve a finer granularity management than a traditional way. In comparison with traditional management – allocating an application to a PM – virtualized management partitions PM's resources (e.g. CPU, memory and disk) into several independent units and allocates applications into these units. The most common units are virtual machines (VMs) and containers.

Virtualization technology rooted back in the 1960s' was originally invented to enable isolated software testing. Because VMs can provide good isolation for applications running without interfering each other [89]. Soon, people realized that virtualization can improve the utilization of hardware resources: With each application deployed in a VM, a PM can run multiple applications.

Next two sections illustrate two classes of virtualization (see Figure 2.4): VM-based and container-based virtualization and then we give a comparison between two virtualizations.

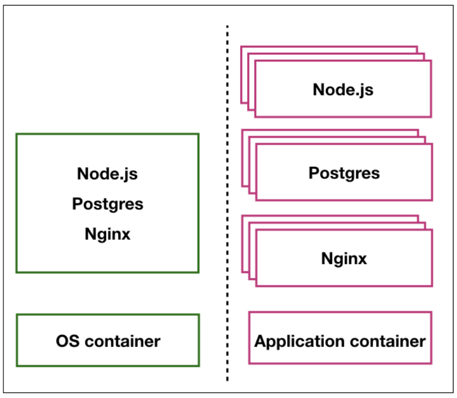


Figure 2.5: A comparison between OS container and Application container [?]

VM-based Virtualization

A VM-based virtualization has a three-layers of structure: PM-Hypervisor-VM (see Figure 2.4 left-hand side). The hypervisor or the virtual machine monitor (VMM) is a software layer on top of PM. A hypervisor arbitrates accesses to the PM's resources so that VMs can share resources on the PM. Some implementations of VM-based hypervisors such as Xen [6], KVM [52], and VMware ESX [102] dominate this field in recent years. On of top of hypervisor, **VMs** are the fundamental resource management units. A VM allows independent Operating system (OS) running on it.

In addition, hypervisors support dynamic migration techniques (e.g pre-copy [22] and post-copy [44]) that can move VMs from one PM to another. Therefore, resource management can improve the utilization of a PM by migrate VMs to that PM.

Container-based Virtualization

A Container-based virtualization has a four-layers of structure: PM-Hypervisor-VM-Container (see Figure 2.4 right-hand side). Container-based virtualization is also addressed as operating-system-level virtualization because containers run on top of VMs. Specifically, container-based virtuliazation includes two classes: OS container and application container [?].

OS containers (Figure 2.5 left-hand side) have a one-on-one relationship with VM. Multiple applications running inside an OS container. Three implementations of OS-level of containers: OpenVZ, Google's control groups, and namespace [83] are widely used in Google and Facebook.

Application containers (Figure 2.5 right-hand side) have a many-to-one relationship with VM. A single application runs on an application container. Major implementations such as Docker, Rocket and Kubernetes [11] are very popular in the software industry.

In comparison with OS container, application container is much more flexible because each container to have its separated environment (e.g. libraries) for applications. Furthermore, application containers provide a finer granularity of resource management by enabling an application level of operations including deployment, scaling, and migration.

Notice that, in the following content, we use "container" to represent "application container". We do not discuss Os container because it is very similar to a VM.

Comparison between Container-based and VM-based virtualization

This section summarizes three disadvantages of VM-based virtualization in terms of energy efficiency and shows container-based can overcome these disadvantages based several research [31,36,108].

Some main disadvantages in VM-based virtualization are listed as follows:

- Resource over-provisioning
  Cloud users tend to reserve over more resources for ensuring the Quality of Service at peak hours [21] which leads to low resource utilization. Cloud users do not completely rely on auto-scaling because auto-scaling is more expensive than reservation. However, the peak hours only account for a short period, therefore, in most of time, resources are wasted.

- Unbalanced usage of resources
  Specific applications consume unbalanced resources which leads to vast amount of resource wastage [96]. For example, computation intensive tasks consume much more CPU than RAM; a fixed type of VM provides much more RAM than it needs. Because the tasks use too much CPU, they prevent other tasks from co-allocating. This also causes wastage.

- Heavy overhead of VM hypervisors and redundant operating systems (OSs)
  Heavy overhead of hypervisors and redundant OSs running in the PM causes huge resource wastage. Traditional VM provides a complete independent environment for deploying software which includes its own OS and libraries. However, as most applications only require a general OS such as Windows or Linux, multiple duplicate OSs running in the system is a waste of resource.

Some key characteristics of containers can help overcome the above disadvantages of VMs.

- Container-based virtualization has a lightweight management which generates much less overhead than a VM hypervisor.

- Container-based virtualization shares OSs which reduces the overhead of multiple OSs while VMs have to run separated OSs.

  The above two characteristics can overcome the heavy overhead of VM hypervisors and reduce the redundant OSs.

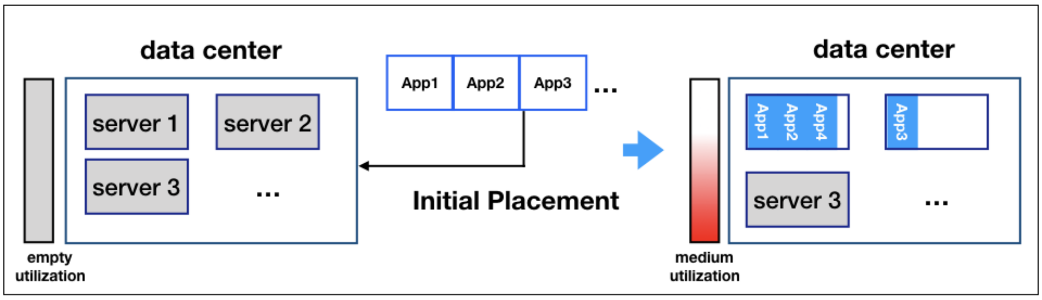
- Container-based virtualization naturally supports vertical scaling while VM-based virtualization does not. Vertical scaling means a container can dynamically adjust its resources under the host's resource constraint.This feature offers a fine granularity management of resources.

  Vertical scaling can overcome the resource over-provisioning by dynamically adjusting the size of containers. Furthermore, the size of container reflects the requirement of the application. We can achieve a balanced usage of resources by using appropriate placement algorithm.

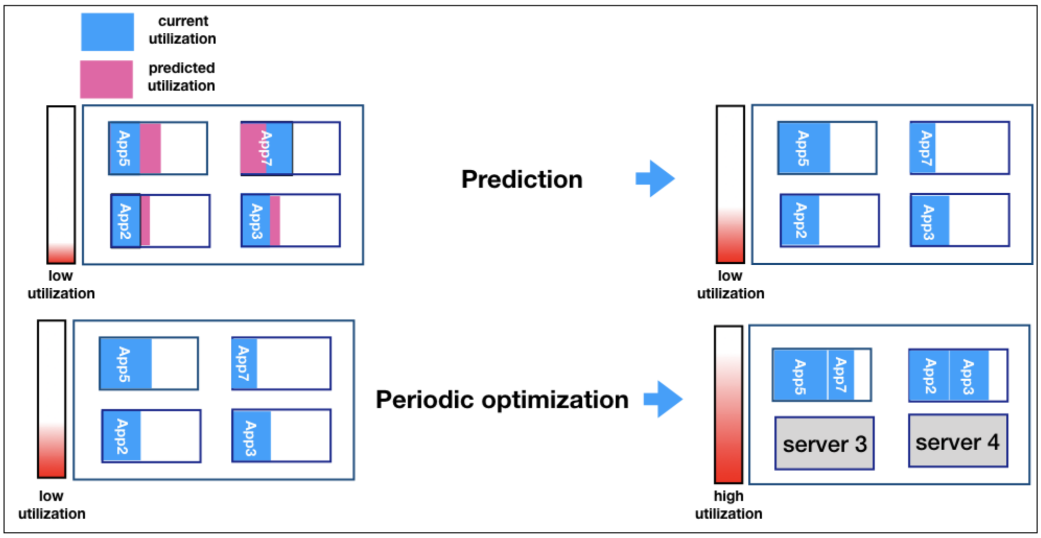
In summary, container-based virtualization has the potential to improve further improve the energy efficiency than VM-based virtualization. No matter cloud uses which virtualization technology, cloud often deals with three placement decision scenarios.

2.2.2 Placement Decision Scenarios

Three placement decision scenarios [68,94]: initial placement of application, periodic placement of application, and dynamic placement of application (see Figure 2.6).



(a) Initial placement of application



(b) Periodic placement of application

(c) Dynamic placement of application

Figure 2.6: Three scenarios of placement decision

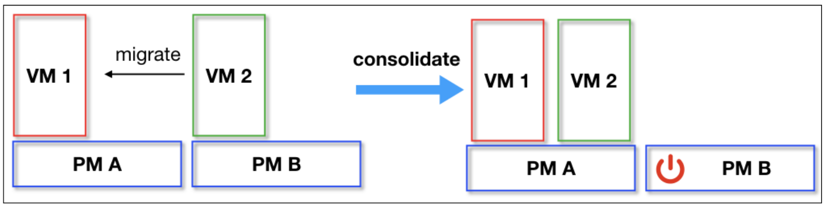


Figure 2.7: A Server Consolidation example: Initially, each PM runs an application wrapped with a VM in a low resource utilization state. After the consolidation, both VMs are running on PM A, so that PM B is turned off to save energy [7].

Initial placement of application is applied when new applications arrived. The task is to place applications into a set of PMs [68] so that PMs satisfy all applications' resource demands and minimize energy consumption.

Periodic placement of application is applied periodically to adjust the current placement of applications. The task is to re-place the current applications so that resource management minimizes the energy consumption and minimizes the cost of migration.

Dynamic placement of application is applied in two scenarios [68]: Overloading and underloading. Overloading is a scenario where the total demands of applications in a PM are higher than the PM's resources. Therefore, the PM causes a degradation in one or more applications' performance. Underloading is a scenario where the PM is running in low utilization. In both scenarios, resource management moves the applications from one PM to another PMs in an on-line fashion [16].

Next section will discuss server consolidation strategies in three scenarios based on the two types of virtualization: VM and container.

2.3 Related Work

Related work discusses the server consolidation strategy and the studies of consolidation strategies on three placement decision scenarios: initial placement, periodic placement, and dynamic placement.

2.3.1 Server consolidation strategy

Server consolidation is a resource management strategy that aims to improve the utilization of PMs and reduce the energy consumption. We use VM-based virtualization as an example: a general step of server consolidation is shown in Figure 2.7, a number of VMs is migrated to fewer number of PMs. Resource management applies Server consolidation to solve the low utilization of PMs called PM sprawl [50].

Types of Server consolidation

Server consolidation can be done in two ways: Static and dynamic [100,110] based on different scenarios. Initial placement and periodic placement involve large number of variables, therefore, they are very time-consuming job and often conducted in an off-line fashion. Dynamic placement requires a fast decision-making to place one application to PMs. Thus, resource management uses a dynamic server consolidation strategy to handle the scenario.

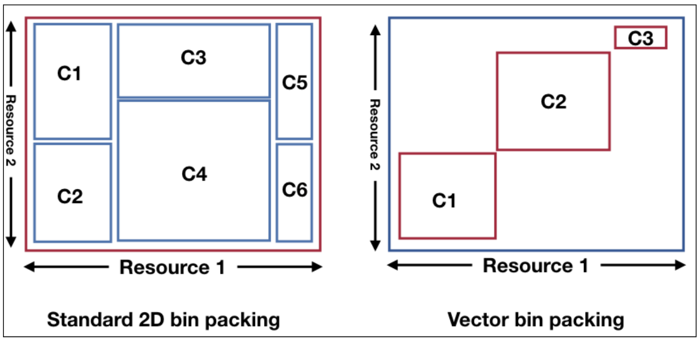


Figure 2.8: A comparison between standard bin packing and vector bin packing

2.3.2 Initial placement of application

This section first discusses server consolidation strategies on container-based virtualization and VM-based virtualization.

Container-based initial placement of application

This section will discuss existing approaches for Container-based initial placement. Then, we will describe a potentially way of modeling energy consumption in container-based virtualization: a bilevel model. Last, we will illustrate a number of algorithms to solve bilevel optimization problems.

Existing approaches Only few container-based research in the literature which mainly follow the VM-based modeling. Piraghaj et al [76] and Mann [63] both consider a linear energy model which is widely adopted in the literature [109]. They consider two resources: CPU and memory. In addition, the constraints are allocated resources cannot exceed the PMs' capacity. In terms of wastage model, so far, they did not use any measure on the residual resources. In contrast, in most VM-based approaches, wastage has been widely considered. It will be discussed in the next section.

The design of three-tier structure: container-VM-PM raises a placement concern, optimize each tier separately and collaboratively. Piraghaj et al [76] consider it separately. Their placement strategy replies on fast and simple heuristic: in each tier of placement, they apply First Fit algorithm. While their main contribution is not the placement strategy but an architecture for container-based resource management. In this architecture, it allows the cloud providers to customize the size or type of VM when allocating applications instead of traditional fixed size VM. They determine the size of VM by two steps, in a pre-execution phase, they perform clustering technique on historical workload data from Google Cluster Data. In this way, they assert that the applications with similar workload pattern can be categorized into the same group. In the execution phase, when a number of deployment requests of container arrives, if there are enough available resources in the existing VMs, they perform First Fit on containers. If new VMs need to be initiated, they first determines the size of VMs by designed policies. Then, the containers are allocated to certain size of VMs.

Mann's research [63] is the earliest study which realizes this two level of placement are interact with each other, therefore, they must be considered collaboratively. Different from Piraghaj's assumption, they consider the container is based on an existing IaaS where fixed types of VM are provided. Therefore, the problem becomes three folds:

1. VM size selection for containers,

2. Place containers,

3. VM placement,

They should be considered together. However, another concern is "why not placement as many containers as possible in a single VM which minimizes the overhead of hypervisor". The paper gives an answer of "Too big VMs limit the consolidation possibilities". In order to prove their interaction, they apply a fixed VM placement algorithm and considering a series of VM selection algorithms such as simple selection [40], Multiple selection, Maxsize, Consolidation-friendly. They discover that the final energy consumption varies with the selection algorithms. They claim that the performance is better when VM selection has more knowledge of the PMs' capacity. However, their study only focuses on the partial placement with fixed VM placement algorithm. The answer of "How these two level of placement interact ?" is still undiscovered.

Three reasons for us to propose a distinct approach from Piraghaj [76] to solve the container-based placement problem. First, their proposed architecture assumes arbitrary size of VM can be created when requests arrive. While our assumption is that the container-based architecture is based on traditional IaaS, where fix-size VMs provide the fundamental resources. Second, from the perspective of energy efficiency, the allocation of container and VM interact with each other. That is, the minimum number of VMs does not necessary lead to the minimum number of PMs, because the type of VMs also affect the results. Therefore, their approach cannot guarantee the energy consumption is near optimal. This inspires us to simultaneously allocate containers and VMs. Third, both approaches did not consider the OS requirement of containers or applications. We argue that this is another critical reason for deploying containers into different VMs besides the "limit the consolidation possibilities" argument states by Mann [63].

In order to solve the problem, we believe one of the promising way is to model the problem as a bilevel optimization [23] as described in detailed in Section 2.3.2. Bilevel optimization naturally models the interaction so that this problem can be tackled by an optimization algorithm. The main challenge is that Bilevel optimization is a strongly NP-hard problem [86], therefore, traditional approaches such as Branch-and-bound [5] can only be applied in very small problems, or convert the problem into a single-level optimization problem. Evolutionary algorithms, on the other hand, becoming popular approaches in this field [86,104].

Energy model in Container-based cloud: Bilevel model The joint allocation of container and VM can be modeled as a bilevel optimization. A bilevel optimization [23] is a kind of optimization where one problem is embedded within another. The general formulation of a bilevel optimization problem can be defined as:

$$min_{x \in X, y} \quad F(x,y) \tag{2.1a}$$
$$s.t \quad G(x,y) \leq 0, \tag{2.1b}$$
$$min_{y} \quad f(x,y) \tag{2.1c}$$
$$s.t \quad g(x,y) \tag{2.1d}$$

The lower-level problem is the function $f(x,y)$, where the decision variable is $y \in \mathbb{R}^{n_2}$. The upper-level problem is the function Fx,y where the decision variable is $x \in \mathbb{R}^{n_1}$. The function $F : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \rightarrow \mathbb{R}$ and $f : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \rightarrow \mathbb{R}$ are the *upper-level* and *lower-level objective functions* respectively. The function $G : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \rightarrow \mathbb{R}^{m_1}$ and $g : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \rightarrow \mathbb{R}^{m_2}$ are called the *upper-level* and *lower-level constraints* respectively.

Bilevel optimization problem has a hierarchical structure which may introduce difficulties such as non-convexity and disconnectedness even for simple cases such as bilevel linear programming problems is strongly NP-hard [86].

In practice, there are a number of problems that are bilevel in nature. For example, transportation related: work design, optimal pricing [17, 24], management: network facility location [93], and engineering related: optimal design [51].

Existing approaches for Bilevel optimization A number of studies have been conducted on bilevel optimization [23, 29]. Approximation algorithms such as Karush-kuhn-Tucker approach [12, 43], branch-and-bound [5] are often applied to solve the problem. Most of these approaches are not applicable when the problem size increases.

Evolutionary methods have been applied to bilevel optimization problem since 90s. Mathieu et al [65] proposed an genetic algorithm (GA) based approach. It uses a nested strategy - the lower level is optimized with a linear programming method and the upper level apply a GA.

Oduguwa and Roy [71] proposed a co-evolutionary approach for bilevel problems. Two population are co-operated to find the optimal solution, where each population handles a sub-problem.

Wang et al [104] proposed an evolutionary algorithm based approach with a constraint handling technique. Their approach is able to handle non-differentiability at the upper level objective function, but not in constraints and lower level objective function. Later on, Wang proposed an improved version [105] that shows better performance than the previous version.

Particle Swarm Optimization [59] was also used in solving bilevel problems. A recent work is from Sinha et al [86], they propose a bilevel evolutionary algorithm (BLEAQ) works by approximating the optimal solution mapping between the lower level optimal solutions and the upper level variables. BLEAQ was tested on two sets of test problems and the results were compared with WJL [104] and WLD [105]. The results show BLEAQ is much faster than previous approaches. One major drawback of evolutionary algorithms is its high computation cost which limits the problem size from growing bigger.

In conclusion, as the complexity of the problem, practical problems with bilevel nature are often simplified into a single-level optimization problem which can achieve a satisfactory level instead of optimal. Classic algorithms often fail because of the nature of bilevel problem such as non-linearity, discreteness, no-differentiability, non-convexity etc. EC algorithms have been successfully applied on bilevel problems.

VM-based initial placement of application

This section first describes a commonly used energy model for VM-based virtualization. Then, we review a number of traditional approaches for VM-based initial placement . We mainly study the following five aspects: resources, power model, wastage model, objective, and algorithm.

Energy model in VM-based cloud: Vector Bin Packing Model Server consolidation is typically modeled as a Vector bin packing problem which is a variant of standard bin packing problem (see Figure 2.8). Vector bin packing is also referred as multi-capacity [58] or multi-dimensional bin packing problem [112]. Vector bin packing is particularly suitable for modeling resource allocation problems where there is a set of bins with known capacities and a set of items with known demands [72]. The optimization objective is to minimize the number of bins.

Table 2.1: A Comparison of different models and approaches

Research	Resources	Algorithm	Power model	Wastage model	Objective
Xu et al [113]	CPU and RAM	GGA and Fuzzy multi-objective	Linear	balance resources	three
Gao et al [41]	CPU and RAM	Ant Colony Optimization	Linear	balance resources	Two
Ferdaus et al [37]	CPU, RAM, and IO	Ant Colony Optimization	Linear	Sum of resources	Single
Wang and Xia [103]	CPU and RAM	MIP	Cubical	No	Single
Wilcox et al [106]	CPU and RAM	GGA	Linear	Sum of resources	Single
Xiong and Xu [112]	CPU,RAM,Bandwidth,Disk	PSO	Non-linear	Sum of resources	Single

A d-dimensional Vector Bin Packing Problem (VBP_d), give a set of items $I^1, I^2, \ldots, I^n$ where each item has d dimension of resources represented in real or discrete number $I^i \in R^d$. A valid solution is packing I into bins $B^1, B^2, \ldots, B^k$. For each bin j and each dimension i, the sum of resources can not exceed the capacity of bin. The goal of Vector Bin Packing problem is to find a valid solution with minimum number of bins. Notice that, the items assigned to bins do not consider the positions in the bins, that is, there is no geometric interpretation of the items or bins [48]. Vector bin packing reduces to the classic *bin-packing* problem when $d = 1$. Vector bin packing is an NP-hard problem in strong sense, as it is a generalized bin packing problem.

Existing Approaches Most of the works model VM placement problem as variants of bin packing problem and propose extensions of greedy-based heuristics such as First Fit Decreasing (FFD) [107], Best Fit, Best Fit Decreasing [9] etc. However, as VM placement is an NP-hard problem, greedy-based approaches can not guaranteed to generate near optimal solutions. Mishra and Sahoo's paper [92] further analyzes and discusses the drawbacks of these approaches. They found that, instead of standard bin packing, only vector bin packing is suitable for modeling resource allocation (see Section 2.3.2). Another drawback of traditional bin packing heuristic is that they do not consider the balance among resources which is a critical issue for vector bin packing problem. Their main contribution is that they list five principles for a good design of objective function, specially, the core idea is to capture the balance among resources.

Based on this insight, Gao et al [41] and Ferdaus et al [37] both propose an Ant Colony Optimization based metaheuristic using a vector algebra complementary resource utilization model proposed by Mishra [92]. They considered three resources CPU, memory, and network I/O with two objectives: minimizing power consumption and resource wastage. They apply the *Resource Imbalance Vector* to capture the imbalance among three resources. Meanwhile, they use a linear energy consumption function to capture the relationship between CPU utilization and energy [33]. Their solution was compared with four algorithms: Max-Min Ant System, a greedy-based approach, and two First Fit Decreasing-based methods. The results show that their proposed algorithm has much less wastage than other algorithms.

Xu and Fortes [113] propose a multi-objective VM placement approach with three objectives: minimizing total resource wastage, power consumption and thermal dissipation costs. They applied an improved grouping genetic algorithm (GGA) with fuzzy multi-objective evaluation. Their wastage by calculating as differences between the smallest normalized residual resource and the others. They also applied a linear power model to estimate the power consumption [60].They conduct experiments on synthetic data and compare with six traditional approaches including First Fit Decreasing (FFD), Best Fit Decreasing (BFD) and single-objective grouping GA. The results showed the superior performance than other approaches.

Wilcox et al [106] also propose a reordering GGA approach because GGA can effectively avoid redundancy [32]. They use a indirect representation [80] which represents the packing as a sequence. In order to transform the sequence into a packing, they applied an ordering operator which, in essence, is a first fit algorithm. This design naturally avoids infeasible solution, therefore, there is no need for constraint handling.

Wang and Xia [103] develop a MIP algorithm for solving large-scale VM placement problem under a *non-linear* power consumption model. Instead of considering the power consumption as a linear model like most researchers, they consider the CPU frequency can be adjust by dynamic voltage and frequency scaling (DVFS), therefore, the power consumption is a cubical power function of frequency. In order to solve the non-linear problem, they first use a linear function to approximate the cubical function. Then, they first use the Gruobi MIP solver to solve the relaxed linearized problem. Then, they apply an iterative rounding algorithm to obtain the near optimal solution.

$$\delta = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{d} (u_j^i - ubest_i)^2} \tag{2.2}$$

Xiong and Xu [112] propose a PSO based approach to solve the problem. Their major contribution is using a total Euclidean distance δ to represent the distance between current resource utilization and the optimal resource utilization (see equation 2.2) where d is the dimension of resources, u_j^i is the current resource utilization of j in a PM i, $ubest_i$ is the predefined optimal resource utilization (e.g 70% CPU utilization). Another contribution is their representation used in PSO. They represent the allocation of each VM to a PM as a probability and let particles search through the indirect solution space.

In summary, most of VM-based placement approaches consider two or three resources (I/O has not been considered in many approaches because they assume that network attached storage (NAS) is used as a main storage along the cluster [69]). After Mishra unreal the principles of vector bin packing, most research apply a balance-measure among resources as their objectives. EC approaches are widely used because they are better performed than traditional heuristics and faster than ILP methods.

2.3.3 Periodic placement

Periodic placement (see Section 2.2.2) is an process that optimizes the current allocation of resources in a periodic fashion [68]. This is because the cloud data center is a dynamic environment with continuous deployment and releases that causes degradation of the resource utilization, thus, the allocation needs to be adjusted when the performance degrades to a certain level. In comparison with initial placement of application (see Section 2.3.2), the similarity is that they are both static approaches which consider a batch of applications and PMs. The difference is that periodic placement needs to take the cost of application migration into account, therefore, it is often considered as a multi-objective optimization problem.

This section first discusses VM-based approaches including migration model and existing approaches. So far, no research has focused on container-based approaches, hence, we are motivated to fill the research gap.

VM-based periodic placement

Therefore, this section will first discuss VM-based models, especially the migration models. Secondly, we will discuss the approaches in periodic placement in the VM-context, specifically, in terms of the prediction of workload, these gaps existed in both VM and container context.

Migration Model Murtazaev and Oh [69], Beloglazov et al [8] and Ferreto et al [38] realize that the migration process generates a large overhead so that it should be used as few as possible. In their migration model, they use the number of migration as the optimization objective. Murtazaev's approach minimize this number by developing an algorithm which always chooses a VM in the least loaded PM and attempts to allocate them on the most loaded PMs.

Existing approaches Based on this idea, Murtazaev develops a heuristic based on First and Best Fit. They select a candidate VM based on a surrogate weight of the resources it used. Beloglazov, on the other hand, considers different criteria for selecting candidate VMs. They not only considers the utilization of VMs but also the utilization of the original PM and target PMs. They also propose a simple heuristic: a modified Best Fit Decreasing to solve the problem. However, these two approaches develop their selection criteria in a greedy fashion which may lead to a local optimal. Ferreto proposes a preprocessing step before the placement algorithm. It first orders the VMs according to their workload variation. Then, it only performs placement on those VMs with the highest variability. These three papers provide some insight that a good placement algorithm should consider more than the utilization of host and target PMs, but also the variation of workload. Most previous consolidation approaches [35,101] only consider static workload. That is, they use a peak or average workload as a represented value as the consolidation input. In most of cases, this will lead to either low utilization: peak time only account for small proportion of the total time, or more migrations: extra migration are performed on workload changes. Therefore, the consolidation is more than aggressively concentrate workload on as few PM as possible, but also considers the robustness. The robustness is referred to the capability of enduring the variation of workload without make too many changes.

In order to achieve robustness, the workload variation must be taken into account. Bobroff [13] analyzed a large number of traces from real world data center. They categorize workloads into three main groups:

- Weak variability.
- Strong variability with weak periodic behavior.
- Strong variability with strong periodic behavior.

Workload with weak variability can be directly packed. The only problem is that their long-term workload can also be changed. For the second type of workload, it is hard or even impossible to predict its behavior. The third type of workload can be predicted. However, it is hard to find the applications with compensated workload patterns.

Meng et al [67] proposed a standard time series technique to extract the deterministic patterns (e.g trends, cycles and seasonality) and irregular fluctuating patterns from workloads' CPU utilization; they assume the periodic behavior of workload will preserve in the future and predict the irregular parts with two approaches: with and without explicit forecast error models. Then, applications are paired according to their negative correlation. They evaluate the workload prediction and application selection with a server consolidation task. They use First Fit to allocate paired applications. During the consolidation, The consolidation results show that they use 45% less PMs for hosting the same number of VMs. Furthermore, their approach is more robust since the variation of workload is considered. However, they only consider two complementary applications at a time.

2.3.4 Dynamic placement

This section first discusses server consolidation strategies on VM-based virtualization and container-based virtualization.

VM-based dynamic placement

Forsman et al [39] propose two distributed migration strategies to balance the load in a system. *The push* strategy is applied on overloaded PM; it attempts to migrate *One* VM at a time to less loaded PMs. *The pull* strategy is applied on underutilized PMs request workload from heavier loaded PMs. Each of the strategy is executed on each PM as an intelligent agent. They share their status with each other through a communication protocol. There are several interesting features of their approach. First, they apply an adaptive high-load threshold (e.g 0.7 of overall CPU utilization) so that it considers the environment changes. Second, they use an EWMA algorithm to reduce the unnecessary migration because EWMA [45] is useful in smoothing out variations in the average load. Third, they applied an entropy to model the load distribution which is also applied in some previous approaches [54, 79]. Their system is agent-based which means large amount of communication may occur between nodes, this would certainly cost extra network resources which are not discussed. Therefore, we expect to design a centralized system, where all nodes are controlled by a controller.

Xiao et al [110] make two contributions, first, they build a quadratic energy model for the energy consumption of PM and a linear model for the energy consumption of migration [61]. Second, they propose an algorithm based on Multiplayer random evolutionary game theory to solve the problem. In their approach, VMs are mapped into players that take part in the evolutionary game. In each iteration, all the players choose their best feasible action, i.e, Migrate to a PM which can minimize the energy consumption. Some players will randomly choose PM to avoid being stuck at a local optimal. Their approach is compared with First Fit, Best Fit Increasing, Best Fit Decreasing, Greedy and Load Balance rule. The solutions show their approach can improve energy consumption greatly, especially in the scenario that the distributions of VMs are very centralized.

Genetic Programming-based Hyper-heuristic (GP-HH) for Bin packing Genetic programming [53] is an evolutionary computation technique, inspired by biological evolution, to automatically find computer programs for solving a specific task. In a GP population, each individual represents a computer program. In each generation, these programs are evaluated by a predefined fitness function, which accesses the performance of each program. Then, individuals will go through several genetic operators such as selection, crossover, and mutation. A number of top individuals will survive to the next generation while others will be discarded. The major difference between GA and GP is that, each GP individual is represented as a tree with variant depth instead of a string. This representation is particular suitable for a program. For example, a GP individual is showed in Figure 2.9 which is a program x + max(y $\times$ 2, -2). The variables {x, y} and constraint {-2, 2} are called terminal of the program. The arithmetic operations {+, $\times$, max } are called functions in GP. A GP individual is a specific combination of elements in terminal set and functional set. In order to observe the relationship between a function and its subtrees, the GP programs are usually presented to human users by using the $prefix$ notation similar to a Lisp expression, for example, x + max(y $\times$ 2, -2) can be expressed as (+ (x (min ($\times$ y 2) -2))).

GP-based hyper-heuristics (GP-HH) has been applied in many applications such as Job shop scheduling to evolve dispatching rules [70]. The term hyper-heuristics [25] means

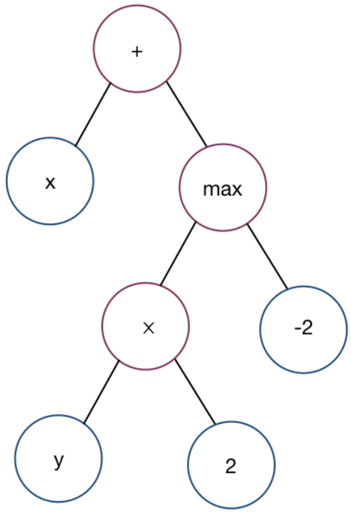


Figure 2.9: GP program that represents x + max(y × 2, -2)

"heuristics to choose heuristics". *Dispatching rule* is essentially a heuristics [73] used in a scheduling context. Resource allocation problems are also in the scheduling category and they are often modeled as bin packing problems. GP-HH has been applied in generating heuristics for bin-packing problems [18,78,85]. These research have shown that GP-HH can generate excellent heuristics which have equal or better performance than human designed heuristics.

In the Cloud computing context, Cloud resource allocation usually has extra constraints such as multi-dimensional resources, migration costs, heterogeneous PMs etc. These constraints make the Cloud resource allocation problem much harder than original bin packing [62]. Therefore, traditional bin packing approaches such as First Fit Decreasing, Best Fit etc cannot perform well in this context. GP-HH, therefore, is a promising technique can be used to automatically generate heuristics under multiple constraints.

Container-based Cloud

Existing approaches Piraghaj et al [77] propose a framework for container-based resource management including three steps, analyzing resources to trigger migration, deciding which containers to migrate, and placing the container to a VM. In the third step, Piraghaj applies three heuristics: First-Fit, Random, and Least Full. However, this work only reports that their approach can reduce the number of VMs but does not mention how to reduce the number of PMs by migrating VMs. Therefore, this work does not consider the interactions between two levels: VM and PM.

GP-HH approach for Bilevel optimization problem So far, no study has focus on using GP-HH for bilevel optimization problem.

2.4 Summary

This chapter reviewed the main concepts of cloud resource management and server consolidation. The challenges of server consolidation in container-based cloud data center were discussed. This chapter also discussed the limitations of existing work on three placement decision scenarios in both container-based and VM-based data center.

- Current research lacks appropriate model to capture the relationship between containers, VMs and PMs. Hence, most research on container-based initial placement of application conducts the placement in two independent steps: container-VM and VM-PM. These approaches neglected the interaction between two levels of placement, hence, they cannot reach a near optimal energy consumption. A bilevel model for the joint placement of containers and VMs need to be proposed. Related sub models such as energy model, workload model, variables and constraints need to be further investigated.

- Periodic placement of application has not been studied in the container context. A bilevel multi-objective energy model needs to be proposed which considers minimizing migration cost as well as minimizing energy consumption.

- Traditional periodic placement of application mostly consider static workload. Thus, it is very likely lead to large number of adjustment of applications' placement in the future because the fluctuation of workloads. These adjustment will increase the cost for Cloud providers. In order to provide a robust placement of application, various predictable workload patterns such as linear continuously changing can be considered. It needs more investigation on how to represent various workloads and how to combine them in a compact structure.

- Current dynamic placement of application approaches are based on simple bin-packing algorithms and manually designed heuristics. These heuristics are either perform poorly or cannot be applied with specific constraints. A hyper-heuristic approach can learn from previous good placement patterns and automatically generate heuristics. In order to design a hyper-heuristic, features of various workload need to be investigated.

This research aims to address the above-mentioned issues. The next chapter will focus on the initial work conducted in investigating NSGA-II for bilevel initial placement of application.

Chapter 3

Preliminary Work

This chapter presents the initial work conducted in investigating NSGA-II for the joint placement of container and VM. In this work, we consider the web services are deployed in containers, therefore, "web service" is used in the content instead of container. This work investigates the bilevel model including sub models (e.g power model), variables, and constraints. Two optimization objectives were initially considered: minimizing the energy consumption and the total price for the used VMs. A NSGA-II approach is applied for optimizing the problem. The result covers the evaluation of the proposed algorithm along with analysis, and concluding remarks and discussion of future work is given in conclusion (Section 3.5).

3.1 Related models

3.1.1 Workload model

Xavier el al [114] develops a *Resource-Allocation-Throughput (RAT)* model for web service allocation. The *RAT model* mainly defines several important variables for an atomic service which represents a software component. Based on this model, firstly, an atomic service's throughput equals its coming rate if the resources of the allocated VM are not exhausted. Secondly, increasing the coming rate will also increase an atomic service's throughput until the allocated resource is exhausted. Thirdly, when the resource is exhausted, the throughput will not increase as request increasing. At this time, the virtual machine reaches its capacity.

3.1.2 Power Model

Shekhar's research [91] is one of the earliest in energy aware consolidation for cloud computing. They conduct experiments of independent applications running in physical machines. They explain that CPU utilization and disk utilization are the key factors affecting the energy consumption. They also find that only consolidating services into the minimum number of physical machines does not necessarily achieve energy saving, because the service performance degradation leads to a longer execution time, which increases the energy consumption.

Bohra [14] develops an energy model to profile the power of a VM. They monitor the sub-components of a VM which includes: CPU, cache, disk, and DRAM and propose a linear model (Eq 3.1). Total power consumption is a linear combination of the power consumption of CPU, cache, DRAM and disk. The parameters α and β are determined based on the observations of machine running CPU and IO intensive jobs.

$$P_{(total)} = \alpha P_{\{CPU,cache\}} + \beta P_{\{DRAM,disk\}} \quad (3.1)$$

Although this model can achieve an average of 93% of accuracy, it is hard to be employed in solving SRAC problem, for the lack of data.

Beloglazov et al. [?] propose a comprehensive energy model for energy-aware resource allocation problem (Eq 3.2). P_{max} is the maximum power consumption when a virtual machine is fully utilized; k is the fraction of power consumed by the idle server (i.e. 70%); and u is the CPU utilization. This linear relationship between power consumption and CPU utilization is also observed by [55,81].

$$P(u) = k \cdot P_{max} + (1-k) \cdot P_{max} \cdot u \quad (3.2)$$

3.2 Problem Description

We consider the problem as a multi-objective problem with two potentially conflicting objectives, minimizing the overall cost of web services and minimizing the overall energy consumption of the used physical machines.

To solve the SRAC problem, we model an atomic service as its request and requests' coming rate, also known as frequency.

The request of an atomic service is modeled as two critical resources: CPU time $A = \{A_1, A_i, \ldots, A_t\}$ and memory consumption $M = \{M_1, M_i, \ldots, M_t\}$, for each request consumes a A_i amount of CPU time and M_i amount of memory. The coming rate is denoted as $R = \{R_1, R_i, \ldots, R_t\}$. In real world scenario, the size and the number of a request are both variant which are unpredictable, therefore, this is one of the major challenges in Cloud resource allocation. In this paper, we use fixed coming rate extracted from a real world dataset to represent real world service requests.

The cloud data center has a number of available physical machines which are modeled as CPU time $PA = \{PA_1, PA_j, \ldots, PA_p\}$ and memory $PM = \{PM_1, PM_j, \ldots, PM_p\}$. PA_j denotes the CPU capacity of a physical machine and PM_j denotes the size of memory. A physical machine can be partitioned or virtualized into a set of virtual machines; each virtual machine has its CPU time $VA = \{VA_1, VA_n, \ldots, VA_v\}$ and memory $VM = \{VM_1, VM_n, \ldots, VM_v\}$.

The decision variable of service allocation is defined as X_n^i. X_n^i is a binary value (e.g. 0 and 1) denoting whether a service i is allocated on a virtual machine n. The decision variable of virtual machine allocation is defined as Y_j^n. Y_j^n is also binary denoting whether a VM n is allocated on a physical machine j.

In this work, we consider homogeneous physical machine which means physical machines have the same size of CPU time and memory. The utilization of a CPU of a virtual machine is denoted as $U = \{U_1, U_n, \ldots, U_v\}$. The utilization can be calculated by Eq.3.3.

$$U_n = \begin{cases} \frac{\sum_{i=1}^{t} R_i \cdot A_i \cdot X_n^i}{VA_n}, \text{If } \frac{\sum_{i=1}^{t} R_i \cdot A_i \cdot X_n^i}{VA_n} < 1 \\ 1 \qquad\qquad\qquad\quad, \text{otherwise} \end{cases} \quad (3.3)$$

The cost of a type of virtual machine is denoted as $C = \{C_1, C_n \ldots, C_v\}$.

In order to satisfy the performance requirement, Service providers often define Service Level Agreements (SLAs) to ensure the service quality. In this work, we define throughput as a SLA measurement [74]. Throughput denotes the number of requests that a service could successfully process in a period of time. According to *RAT* model, the throughput is equal to the number of requests when the allocated resource is sufficient. Therefore, if a VM reaches its utilization limitation, it means that the services have been allocated exceedingly. Therefore, all services in that VM suffer from performance degradation.

Then we define two objective functions as the total energy consumption and the total cost of virtual machines:

$$\text{minimize}$$
$$Energy = \sum_{j=1}^{p}(k \cdot V_{max} + (1-k) \cdot V_{max} \cdot \sum_{n=1}^{v} U_n \cdot Y_j^n) \tag{3.4}$$

$$Cost = \sum_{j=1}^{p}\sum_{n=1}^{v} C_n \cdot Y_j^n \tag{3.5}$$

Hard constraint

A virtual machine can be allocated on a physical machine if and only if the physical machine has enough available capacity on every resource.

$$\begin{aligned} \sum_{n=1}^{v} VM_n \cdot Y_j^n \leq PM_j \\ \sum_{n=1}^{v} VA_n \cdot Y_j^n \leq PA_j \end{aligned} \tag{3.6}$$

Soft constraint

A service can be allocated on a virtual machine even if the virtual machine does not have enough available capacity on every resource, but the allocated services will suffer from a quality degradation.

$$\sum_{i=1}^{t} M_i \cdot R_i \cdot X_i^n \leq VM_n \tag{3.7}$$

3.3 Methods

As we have discussed, Multi-objective Evolutionary Algorithms are good at solving multi-objective problems and NSGA-II [27] has shown his effective and efficiency. NSGA-II is a well-known MOEA that has been widely used in many real-world optimization problems. In this paper we also adopt NSGA-II to solve the SRAC problem. We first propose a representation and then present a NSGA-II based algorithm with novel genetic operators.

3.3.1 Chromosome Representation

SRAC is a two-level bin-packing problem, in the first level, bins represent physical machines and items represent virtual machines. Whereas, in the second level, a virtual machine acts like a bin and web services are items. Therefore, we design the representation in two hierarchies, virtual machine level and physical machine level.

Figure 3.1 shows an example individual which contains seven service allocations. Each allocation of a service is represented as a pair where the index of each pair represents the number of web service. The first number indicates the type of virtual machine that the service is allocated in. The second number denotes the number of virtual machine. For example, in Figure 3.1, service #1 and service #2 are both allocated in the virtual machine

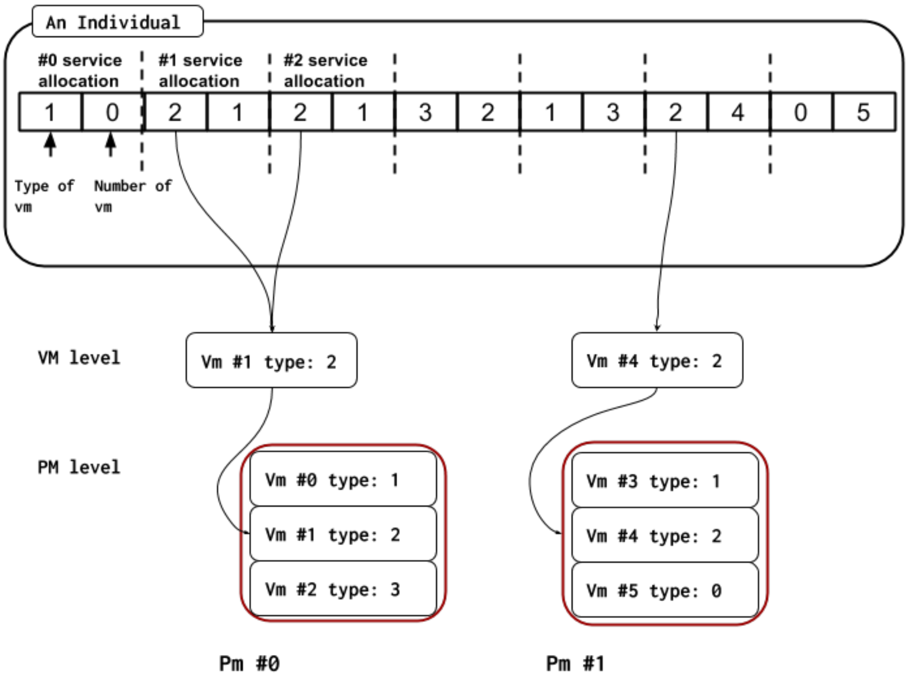


Figure 3.1: An example chromosome representation

#1 while service #1 and service #5 are allocated to different virtual machines sharing the same type. The first hierarchy shows the virtual machine in which a service is allocated by defining VM type and number. Note that, the VM type and number are correlated once they are initialized. With this feature, the search procedure is narrowed down in the range of existing VMs which largely shrinks the search space. The second hierarchy shows the relationship between a physical machine and its virtual machines, which are implicit. The physical machine is dynamically determined according to the virtual machines allocated on it. For example, in Figure 3.1, the virtual machines are sequentially packed into physical machines. The boundaries of PMs are calculated by adding up the resources of VMs until one of the resources researches the capacity of a PM. At the moment, no more VMs can be packed into the PM, then the boundary is determined. The reason we designed this heuristic is because a physical machine is always fully used before launching another. Therefore, VM consolidation is inherently achieved.

Clearly, specifically designed operators are needed to manipulate chromosomes. Therefore, based on this representation, we further developed initialization, mutation, constraint handling and selection method.

3.3.2 Initialization

The initialization (see Alg 1) is designed to generate a diverse population. In the first step, for each service, it is able to find the most suitable VM type which is just capable of running the service based on its resource requirements. In the second step, based on the suitable VM type, a stronger type is randomly generated. If there exists a VM with that type, the service is either deployed in the existing VM or launch a new VM. We design a consolidation factor c which is a real number manually selected from 0 to 1 to control this selection. If a random number u is smaller than c, the service is consolidated in an existing VM.

This design could adjust the consolidation, therefore, controls the utilization of VM.

Algorithm 1 Initialization

Inputs:
VM CPU Time VA and memory VM,
Service CPU Time A and memory M
consolidation factor c
Outputs: A population of allocation of services

```
for Each service t do
    Find its most suitable VM Type
    Randomly generate a VM type vmType which is equal or better than its most suitable type
    if There are existing VMs with vmType then
        randomly generate a number u
        if u < consolidation factor then
            randomly choose one existing VM with vmType to allocate
        else
            launch a new VM with vmType
        end if
    else
        Create a new VM with its most suitable VM type
    end if
end for
```

3.3.3 Mutation

The design principle for mutation operator is to enable individuals exploring the entire feasible search space. Therefore, a good mutation operator has two significant features, the exploration ability and the its ability to keep an individual within the feasible regions. In order to achieve these two goals, firstly, we generate a random virtual machine type which has a greater capacity than the service needs. It ensures the feasible of solutions as well as exploration capability. Then, we consider whether a service is consolidated with the consolidation factor c.

The consolidation is conducted with a roulette wheel method which assigns fitness value to each VM according to the reciprocal of its current utilization. The higher the utilization, the lower the fitness value it is assigned. Therefore, a lower utilization VM has a greater probability to be chosen. At last, if a new VM is launched, it will not be placed at the end of VM lists. Instead, it will be placed at a random position among the VMs. The reason is illustrated in Figure 3.2. In the example, VM #2 is mutated into a new type and be placed at the end of the VM list. However, because of the size of VM #3 is too large for PM #0, the hollow in PM #0 will never be filled. This problem can be solved with the random insertion method.

3.3.4 Violation control method

A modified violation ranking is proposed to deal with the soft constraint, for the hard constraint is automatically eliminated by the chromosome representation. We define a violation number as the number of services which are allocated in the degraded VMs. That is, if there are excessive services allocated in a VM, then all the services are suffered from a degraded in performance. The violation number is used in the selection procedure, where the individuals with less violations are always preferred.

3.3.5 Selection

Our design uses the binary tournament selection with a constrained-domination principle. A constrained-domination principle is defined as following. A solution I is considered constraint-dominate a solution J, if any of the following condition is true:

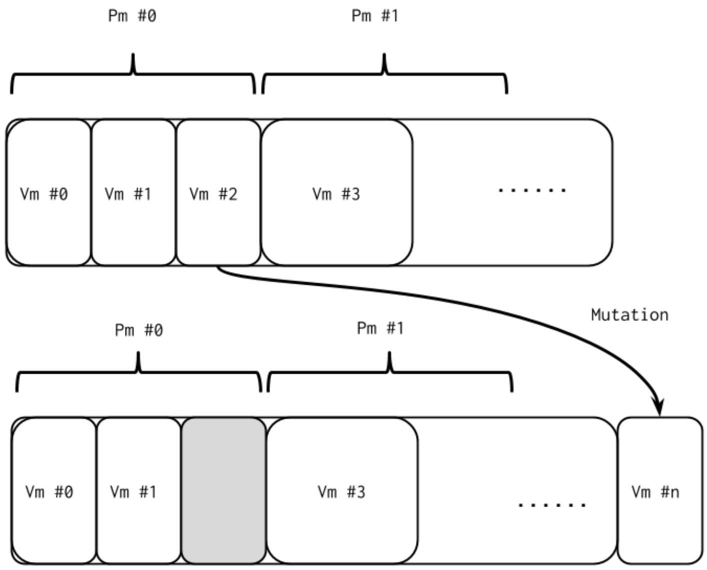


Figure 3.2: An example mutation without insertion that causes a lower resource utilization

---

Algorithm 2 Mutation

Inputs:
An individual VM CPU Time *VA* and memory VM,
Service CPU Time A and memory M
consolidation factor c
Outputs: A mutated individual

for Each service **do**
  Randomly generate a number u
  if $u <$ mutation rate **then**
    find the most suitable VM Type for this service
    Randomly generate a number k
    if $k <$ consolidation factor **then**
      calculate the utilization of used VMs
      assign each VM with a fitness value of 1 / utilization and generate a roulette wheel according to their fitness values
      Randomly generate a number p, select the VM according to p
      Allocate the service
    else
      launch a new VM with the most suitable VM Type
      insert the new VM in a randomly choose position
    end if
  end if
end for

---

1. Solution I is feasible, solution is not,

2. Both solutions are infeasible, I has smaller overall violations,

3. Both solutions are feasible, solution I dominates solution J.

An individual with no or less violation is always selected. This method has been proved effective in the original NSGA-II paper [27].

3.3.6 Fitness Function

The cost fitness (Eq.3.5) is determined by the type of VMs at which web service are allocated. The energy fitness is shown in Eq.3.4, the utilizations (Eq.3.3) of VM are firstly converted into the utilizations of PM according to the proportion of VMs and PMs CPU capacity.

3.3.7 Algorithm

The main difference between our approach and the original NSGA-II is that our approach has no crossover operator.

That is, a random switch of chromosome would completely destroy the order of VMs, hence, no useful information will be preserved. Therefore, we only apply mutation as the exploration method. Then, the algorithm becomes a parallel optimization without much interaction between its offspring, which is often addressed as Evolutionary Strategy [56].

Algorithm 3 NSGA-II for SRAC

Inputs:
VM CPU Time VA and memory VM,
PM CPU Time PA and memory PM,
Service CPU Time A and memory M
consolidation factor c
Outputs: A Non-dominated Set of solutions

```
Initialize a population P
while Termination Condition is not meet do
    for Each individual do
        Evaluate the fitness values
        Calculate the violation
    end for
    non-Dominated Sorting of P
    calculate crowding distance
    while child number is less than population size do
        Selection
        Mutation
        add the child in a new population U
    end while
    Combine P and U { for elitism}
    Evaluate the combined P and U
    Non-dominated sorting and crowding distance for combined population
    Include the top popSize ranking individuals to the next generation
end while
```

3.4 Experiment

3.4.1 Dataset and Problem Design

This project is based on both real-world datasets *WS-Dream* [115] and simulated datasets [15]. The *WS-Dream* contains web service related datasets including network latency and service frequency (request coming rate). In this project, we mainly use the service frequency matrix. For the cost model, we only consider the rental of virtual machines with fixed fees (monthly rent). The configurations of VMs are shown in Table 3.2, the CPU time and memory were selected manually and cost were selected proportional to their CPU capacity. The maximum PM's CPU and memory are set to 3000 and 8000 respectively. The energy consumption is set to 220W according to [15].

Table 3.1: Problem Settings

Problem	1	2	3	4	5	6
Number of services	20	40	60	80	100	200

Table 3.2: VM configurations

VM Type	CPU Time	Memory	Cost
1	250	500	25
2	500	1000	50
3	1500	2500	150
4	3000	4000	300

We designed six problems shown in Table 3.1, listed with increasing size and difficulty, which are used as representative samples of SRAC problem.

Selection Method with violation Control vs. without violation control

We conducted two comparison experiments. For the first experiment, we make a comparison between NSGA-II with violation control and NSGA-II without violation control. In second experiment, two mutation operators are compared. The first is the roulette wheel mutation, the second is the mutation with greedy algorithm. The mutation with greedy algorithm is a variant of roulette wheel mutation. The only difference is that instead of selecting a VM to consolidate with fitness values, it always selects the VM with the lowest CPU utilization. Therefore, it is a greedy method embedded in the mutation.

The experiments were conducted on a personal laptop with 2.3GHz CPU and 8.0 GB RAM. For each approach, 30 independent runs are performed for each problem with constant population size 100. The maximum number of iteration is 200. k equals 0.7. We set mutation rate and consolidation factor to 0.9 and 0.01.

3.4.2 Results

Table 3.3: Comparison between two Mutation methods

Problem	roulette wheel mutation		Greedy mutation	
	cost fitness	energy fitness	cost fitness	energy fitness
1	2664.6 ± 66.4	1652.42 ± 18.2	2661.7 ± 56.9	1653.2 ± 18.2
2	6501.1 ± 130.2	4614.0 ± 110.7	6495.37 ± 110.7	4132.5 ± 80.4
3	8939.2 ± 118.5	6140.7 ± 204.0	9020.5 ± 204.0	5739.6 ± 148.6
4	11633.7 ± 301.1	9301.9 ± 254.0	12900.6 ± 243.0	9376.3 ± 120.9
5	14102.0 ± 231.7	10164.8 ± 238.9	14789.2 ± 238.8	9876.3 ± 120.9
6	27194.3 ± 243.0	19914.4 ± 307.5	27654.2 ± 307.5	19187.1 ± 176.6

As we conducted the experiment for 30 runs, we first obtain an average non-dominated set over 30 runs by collecting the results from a specific generation from all 30 runs, and then apply a non-dominated sorting over them.

Firstly, we show the non-dominated solutions evolve along with the evolution process in Figure 3.3. These results come from selection method without violation control. As it illustrated, different colors represent different generations from 0th to 200th. For problem 1, because the problem size is small, the algorithm converged before 100 generations. Therefore,

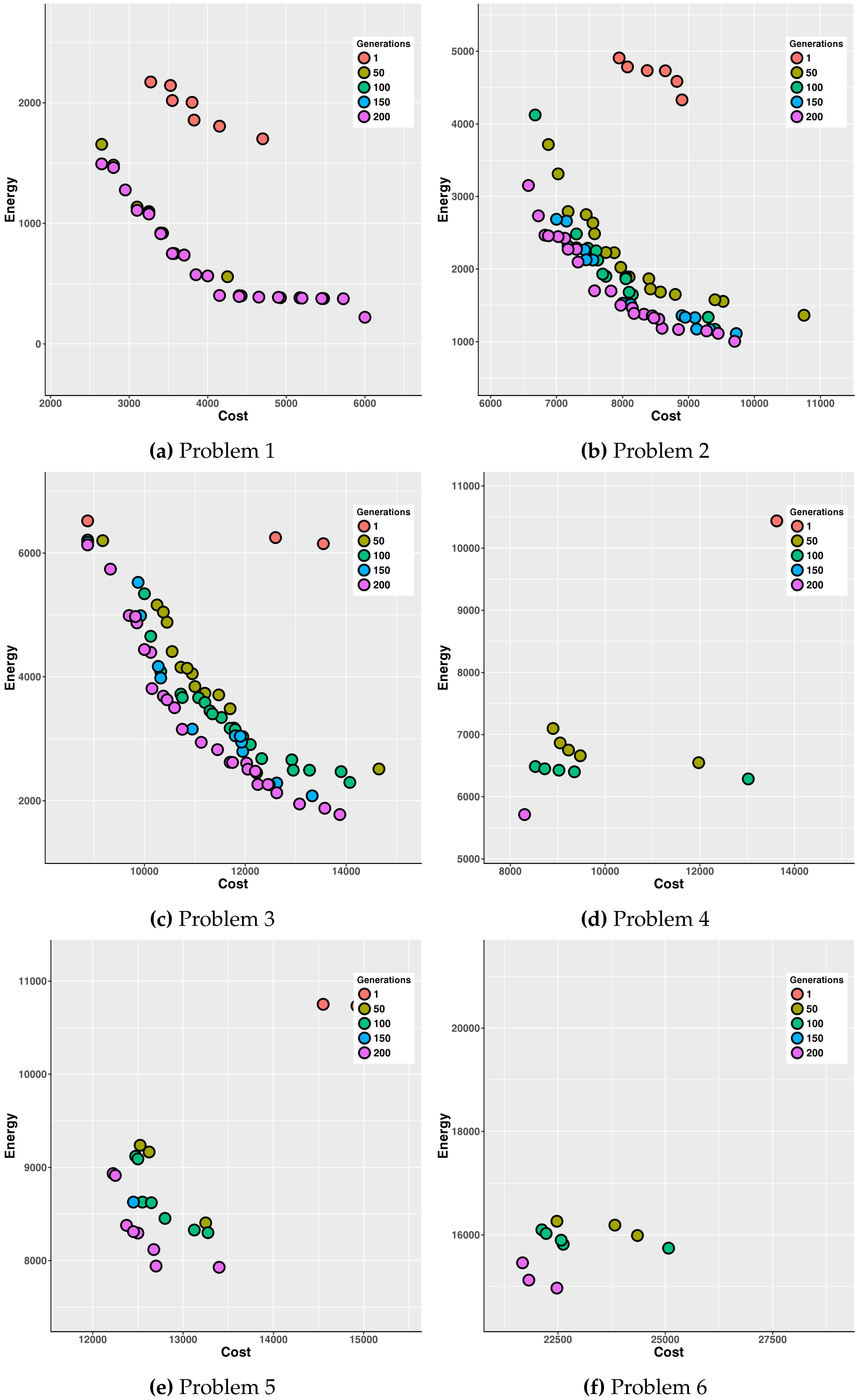
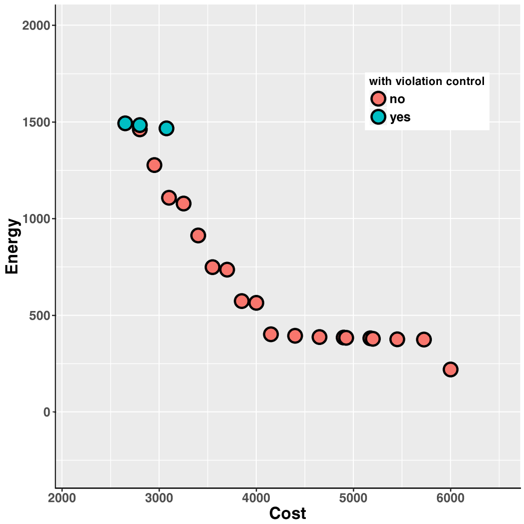
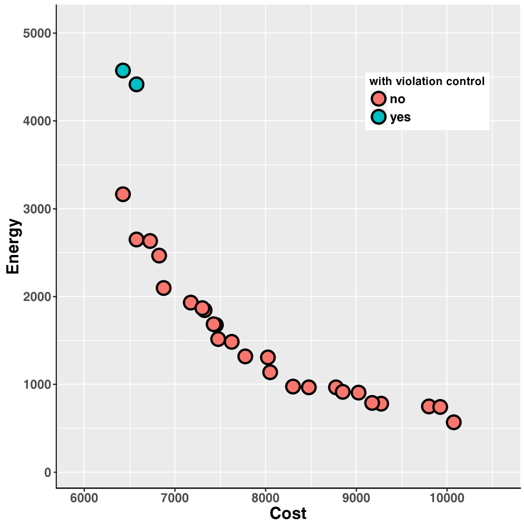


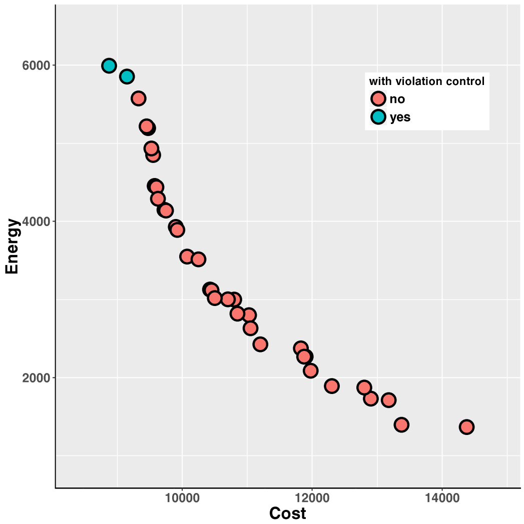
Figure 3.3: Non-dominated solutions evolve along with the generation



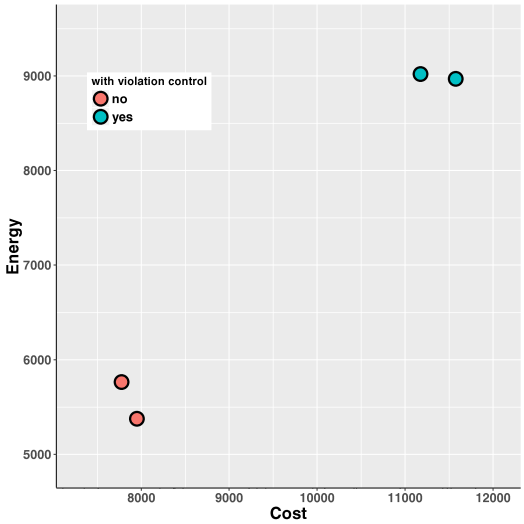
(a) Problem 1



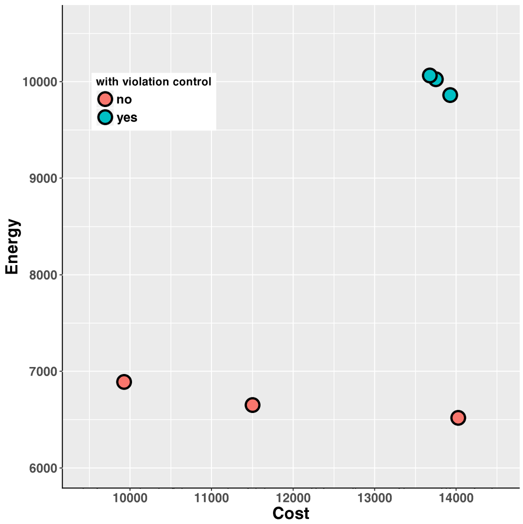
(b) Problem 2



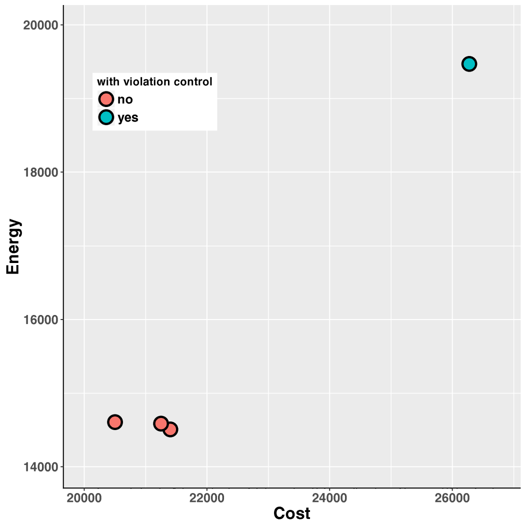
(c) Problem 3



(d) Problem 4



(e) Problem 5



(f) Problem 6

Figure 3.4: non-dominated solutions comparison between selection with violation control and without violation control

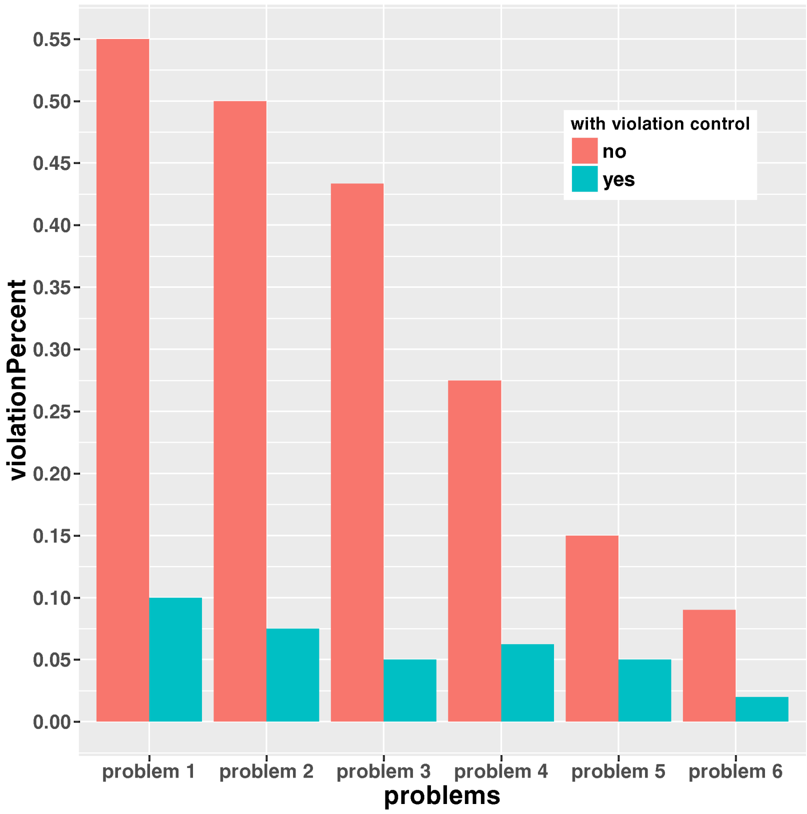


Figure 3.5: Violation Percentage comparison between selection with violation control and without violation control

the non-dominated set from the 100th and 150th generations are overlapping with results from the 200th generation. For problem 2 and problem 3, it clearly shows the improvement of fitness values. For problem 4 onwards, the algorithm can only obtain a few solutions as the problem size is large, it is difficult to find solutions.

Then, the non-dominated sets of the last generation from two selection methods are compared in Figure 3.4. There are much fewer results are obtained from the violation control method throughout all cases. For the first three problems, the non-dominated set from the violation control method has similar quality as the no violation control method. From problem 4 onwards, the results from selection with violation control are much worse in terms of fitness values. However, most of the results from non-violation control selection have a high violation rate. That is, the method without violation control is stuck in the infeasible regions and provide high-violation rate solutions.

From figure 3.5, we can observe the violation rate between two methods. It proves violation control has a great ability to prevent the individual from searching the infeasible region. On the other hand, without violation control, although, the algorithm can provide more solutions with better fitness values, most of them have a high violation rate over 10% which are not very useful in reality.

As we mentioned in previous section, the mutation rate and consolidation factor are set differently for the two methods. For the method with violation control, the mutation rate is set to 0.9 and the consolidation factor c is set to 0.01, this is because the feasible region is narrow and scattered. In order to avoid stucking in the local optima, a large mutation rate can help escape local optima. For the factor c, a larger percentage would easily lead the algorithm to infeasible regions. Therefore, it is set to a small number.

Mutation with roulette wheel vs. Mutation with greedy algorithm

Table 3.3 shows the fitness value comparison between mutation methods. According to statistics significant test, there is little difference between methods. The possible reason is the consolidation factor is set to 0.01. In each mutation iteration, there is only 1% probability that a service will be consolidated in an existed VM, therefore, the influence between different consolidation strategies is trivial.

3.5 Findings and Future work

This work investigates the bilevel energy model for the joint placement of container and VM. Several sub models such as workload model, power model were considered. A multi-objective formulation of the bilevel problem were established. Two objectives, minimizing the cost of used VMs and minimizing the energy consumption are achieved. In order to optimize the problem, we propose a NSGA-II based algorithm with specific designed representation which reduces the bilevel placement into one single-level. Genetic operators such as population initialization, mutation, selection were also designed for generating valid solutions and handling the constraints. The results are compared with different variances of the algorithm. The results show our approach can solve the very complicate optimization problem. Current work does not consider the balance between CPU and memory and the overhead of VM. Therefore, in the next step, we will investigate these two factors and add it into the bilevel energy model.

Chapter 4

Proposed Contributions and Project Plan

This thesis will contribute to the field of Cloud Computing by proposing novel solutions for bilevel optimization of the joint allocation of container and VM, and to the field of Evolutionary Computation by proposing new representations and genetic operators in evolutionary algorithms. The proposed contributions of this project are listed below:

1. Two bilevel models for two placement problems: **first**, a new bilevel energy model for initial placement of application; **second**, a new bilevel energy and migration model for periodic placement of application with the consideration of three types of workload. The above two models will address the relationship between five factors and energy consumption. The five factors involve locations of container, types of VM, locations of VM, overheads of VM, and the balance between memory and CPU. In addition, the energy and migration model addresses two more factors, migrations of VM and container, three types of workload. These two bilevel models can be used in optimizing energy consumption in initial placement and periodic placement problems.

2. An EC-based bilevel single-objective optimization algorithm for the initial placement of application based on previously proposed bilevel energy model. This algorithm combines clustering technique and heuristics to achieve the scalability of handling one thousand applications. This algorithm is expected to achieve a better energy efficiency than existing VM-based approaches.

3. An EC-based bilevel multi-objective algorithm with Pareto front approach for the periodic placement of application with consideration of three types of workload. This work proposes specific representations and genetic operators for three types of workload. The algorithm is expected to achieve better energy efficiency and migration cost than VM-based approaches with the consideration of three types of workload.

4. A new genetic programming hyper-heuristic (GP-HH) approach for single-objective dynamic placement of application with various types of workload in **VM-based cloud**. The proposed GP-HH solves the single-level of placement: VM to PM, therefore, the algorithm can be used in VM-based cloud. This work will also extract features from various workloads to construct a new primitive set for the GP-HH approach. The proposed GP-HH is expected to learn from good placement solutions and automatically generate dispatching rules for dynamic placement of VMs. These dispatching rules are expected to fast allocate VMs to PMs and achieve a near-optimal solution in energy consumption.

Table 4.1: Phases of project plan

Phase	Task	Duration (Months)
1	Reviewing literature, overall design, selection of datasets and writing the proposal	12 (Complete)
2	Develop a single-objective EC-based approach for the joint allocation of containers and VMs	7
3	Develop multi-objective EC-based approaches for container-based cloud in periodic placement of application with considering various types of workload	7
4	Develop a cooperative Genetic programming based hyper-heuristic approach for dynamic placement.	7
5	Writing the thesis	6

5. A new cooperative GP-HH approach for single-objective dynamic placement of application with various types of workload in **container-based cloud**. This work is based on the previously proposed GP-HH approach. Two GP-HH approaches cooperate to generate dispatching rules for dynamic placement problem in container-based cloud. The cooperative GP-HH can learn good placement patterns from good solutions and output dispatching rules. These dispatching rules can achieve fast allocation of containers and VMs as well as near optimal solutions in terms of energy consumption.

4.1 Overview of Project Plan

Six overall phases have been defined in the initial research plan for this PhD project, as shown in Table 4.1. The first phase, which comprises reviewing the relevant literature, investigating both VM-based and container-based server consolidation algorithms, and producing the proposal, has been completed. The second phase, which corresponds to the first objective of the thesis, is currently in progress and is expected to be finished on time, thus allowing the remaining phases to also be carried out as planned.

4.2 Project Timeline

The phases included in the plan above are estimated to be completed following the timeline shown in Table 4.2, which will serve as a guide throughout this project. Note that part of the first phase has already been done, therefore the timeline only shows the estimated remaining time for its full completion.

4.3 Thesis Outline

The completed thesis will be organized into the following chapters:

- *Chapter 1: Introduction*
  This chapter will introduce the thesis, providing a problem statement and motivations, defining research goals and objectives, and outlining the structure of the final thesis.

- *Chapter 2: Literature Review*
  The literature review will illustrate the fundamental background of Cloud computing, resource management, and server consolidation. It will examine the existing work on VM-based and container-based server consolidation, discuss concepts in this field in order to provide readers with the necessary background. Multiple sections will consider the issues such as initial placement of application, periodic placement, and

Table 4.2: Time Line

Task	Months											
	2	4	6	8	10	12	13	16	18	20	22	24
Literature Review and Updating	x	x	x	x	x	x	x	x	x	x	x	x
Develop a new bilevel energy model	x											
Propose a new EC based bilevel optimization approach to solve the initial placement of application	x	x										
Improve the scalability of the proposed EC-based approach up to one thousand applications		x	x									
Propose a multi-objective bilevel model for periodic placement with consideration of static workload.				x								
Propose a multi-objective bilevel EC-based algorithm for periodic placement of application with Pareto front approach.					x	x						
Extend the bilevel multi-objective model for adapting three types of predictable workload.						x						
Extend the previous EC-based multi-objective algorithm to adapt to three types of workload.							x					
Develop a GP-based hyper-heuristic (GP-HH) algorithm for automatically generating dispatching rules for the single-level placement.								x	x			
Conduct feature extraction on the three predictable workloads and two unpredictable workloads to construct a new primitive set.									x			
Develop a cooperative GP-HH approach for automatically generating dispatching rules for placing both containers and VMs.									x	x		
Writing the first draft of the thesis										x	x	
Editing the final draft										x	x	x

dynamic placement of application. The focus of this review is on investigating server consolidation techniques.

- *Chapter 3: Develop EC-based approaches for the single objective joint placement of containers and VMs for initial placement of application.*
  This chapter will establish a new bilevel energy model for the joint placement of container and VM problem. Based on this model, this chapter will introduce a new EC-based bilevel algorithm combined with clustering technique and heuristics to solve the initial placement of application problem.

- *Chapter 4: Develop EC-based approaches for the multi-objective joint allocation problem for periodic placement of application*
  This chapter proposes a new bilevel energy and migration model based previously proposed energy model to adapt to multi-objective problem with three types of workload. This chapter will also propose new EC-based approaches for the bilevel multi-objective periodic placement with considering three types of workload. It is then followed by algorithm performance evaluation that contains an experiment design, setting, results and analysis.

- *Chpater 5: Develop a single-objective cooperative Genetic Programming hyper-heuristic (GP-HH) approach for automatically generating dispatching rules for dynamic placement of application*
  This chapter focuses on providing a Genetic Programming-based hybrid heuristic approach to automatic generate dispatching rules to dynamic consolidation problem. This chapter will propose two algorithms. A GP-HH for single-level of placement: VM-PM and A cooperative GP-HH for bilevel placement: container-VM and VM-PM.

- *Chapter 7: Conclusions and Future Work* In this chapter, conclusions will be drawn from the analysis and experiments conducted in the different phases of this research, and the main findings for each one of them will be summarized. Additionally, future research directions will be discussed.

4.4 Resources Required

4.4.1 Computing Resources

An experimental approach will be adopted in this research, entailing the execution of experiments that are likely to be computationally expensive. The ECS Grid computing facilities can be used to complete these experiments within reasonable time frames, thus meeting this requirement.

4.4.2 Library Resources

The majority of the material relevant to this research can be found on-line, using the university electronic resources. Other works may either be acquired at the university library, or by soliciting assistance from the Subject Librarian for the fields of engineering and computer science.

4.4.3 Conference Travel Grants

Publications to relevant venues in this field are expected throughout this project, therefore travel grants from the university are required for key conferences.

Bibliography

[1] 0002, Y. S., White, J., Li, B., Walker, M., and Turner, H. A. Automated QoS-oriented cloud resource optimization using containers. *Autom. Softw. Eng. 24*, 1 (2017), 101–137.

[2] Adhikari, V. K., Guo, Y., Hao, F., Varvello, M., Hilt, V., Steiner, M., and Zhang, Z.-L. Unreeling netflix - Understanding and improving multi-CDN movie delivery. *INFOCOM* (2012).

[3] Angelo, J. S., Krempser, E., and Barbosa, H. J. C. Differential evolution for bilevel programming. In *2013 IEEE Congress on Evolutionary Computation (CEC)* (2013), IEEE, pp. 470–477.

[4] Banzhaf, W., Nordin, P., Keller, R. E., and Francone, F. D. Genetic programming: an introduction, 1998.

[5] Bard, J. F., and Falk, J. E. An explicit solution to the multi-level programming problem. *Computers & OR 9*, 1 (1982), 77–100.

[6] Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T. L., Ho, A., Neugebauer, R., Pratt, I., and Warfield, A. Xen and the art of virtualization. *SOSP* (2003), 164.

[7] Barroso, L. A., and Hölzle, U. The Case for Energy-Proportional Computing. *IEEE Computer 40*, 12 (2007), 33–37.

[8] Beloglazov, A., Abawajy, J. H., and Buyya, R. Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing. *Future Generation Comp. Syst. 28*, 5 (2012), 755–768.

[9] Beloglazov, A., and Buyya, R. Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers. *MGC@Middleware* (2010), 1–6.

[10] Beloglazov, A., and Buyya, R. Managing Overloaded Hosts for Dynamic Consolidation of Virtual Machines in Cloud Data Centers under Quality of Service Constraints. *IEEE Transactions on Parallel and Distributed Systems 24*, 7 (2013), 1366–1379.

[11] Bernstein, D. Containers and Cloud - From LXC to Docker to Kubernetes. *IEEE Cloud Computing* (2014).

[12] Bianco, L., Caramia, M., and Giordani, S. A bilevel flow model for hazmat transportation network design. *Transportation Research Part C: ... 17*, 2 (Apr. 2009), 175–196.

[13] BOBROFF, N., KOCHUT, A., AND BEATY, K. A. Dynamic Placement of Virtual Machines for Managing SLA Violations. *Integrated Network Management* (2007), 119–128.

[14] BOHRA, A. E. H., AND CHAUDHARY, V. VMeter: Power modelling for virtualized clouds. In *Parallel {& Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on* (2010), Ieee, pp. 1–8.

[15] BORGETTO, D., CASANOVA, H., DA COSTA, G., AND PIERSON, J.-M. Energy-aware service allocation. *Future Generation Computer Systems 28*, 5 (2012), 769–779.

[16] BORODIN, A., AND EL, R. Yaniv (1998): Online Computation and Competitive Analysis.

[17] BROTCORNE, L., LABBÉ, M., MARCOTTE, P., AND SAVARD, G. A Bilevel Model for Toll Optimization on a Multicommodity Transportation Network. *Transportation Science 35*, 4 (Nov. 2001), 345–358.

[18] BURKE, E. K., HYDE, M. R., KENDALL, G., AND WOODWARD, J. Automating the Packing Heuristic Design Process with Genetic Programming. *Evolutionary Computation 20*, 1 (Mar. 2012), 63–89.

[19] BUYYA, R., BELOGLAZOV, A., AND ABAWAJY, J. H. Energy-Efficient Management of Data Center Resources for Cloud Computing - A Vision, Architectural Elements, and Open Challenges. *PDPTA cs.DC* (2010), arXiv:1006.0308.

[20] BUYYA, R., RANJAN, R., AND CALHEIROS, R. N. Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit - Challenges and Opportunities. *PDPTA cs.DC*, Chapter 4 (2009), 24.

[21] CHAISIRI, S., LEE, B.-S., AND NIYATO, D. Optimization of Resource Provisioning Cost in Cloud Computing. *IEEE Trans. Services Computing 5*, 2 (2012), 164–177.

[22] CLARK, C., FRASER, K., HAND, S., HANSEN, J. G., JUL, E., LIMPACH, C., PRATT, I., AND WARFIELD, A. Live migration of virtual machines. 273–286.

[23] COLSON, B., MARCOTTE, P., AND SAVARD, G. An overview of bilevel optimization. *Annals OR 153*, 1 (2007), 235–256.

[24] CONSTANTIN, I. Optimizing frequencies in a transit network: a nonlinear bi-level programming approach. *International Transactions in Operational Research 2*, 2 (Apr. 1995), 149–164.

[25] COWLING, P., KENDALL, G., AND SOUBEIGA, E. A Hyperheuristic Approach to Scheduling a Sales Summit. In *Practice and Theory of Automated Timetabling III*. Springer, Berlin, Heidelberg, Berlin, Heidelberg, Aug. 2000, pp. 176–190.

[26] DAYARATHNA, M., WEN, Y., AND FAN, R. Data Center Energy Consumption Modeling - A Survey. *... Surveys & Tutorials* (2016).

[27] DEB, K., PRATAP, A., AGARWAL, S., AND MEYARIVAN, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation 6*, 2 (Apr. 2002), 182–197.

[28] DEB, K., AND SINHA, A. An Efficient and Accurate Solution Methodology for Bilevel Multi-Objective Programming Problems Using a Hybrid Evolutionary-Local-Search Algorithm. *dx.doi.org 18*, 3 (Aug. 2010), 403–449.

[29] DEMPE, S., DUTTA, J., AND LOHSE, S. Optimality conditions for bilevel programming problems. ... *with Multivalued Mappings 55*, 5-6 (Oct. 2006), 505–524.

[30] DÓSA, G., AND SGALL, J. First Fit bin packing - A tight analysis. *STACS* (2013).

[31] DUA, R., RAJA, A. R., AND KAKADIA, D. Virtualization vs Containerization to Support PaaS. In *2014 IEEE International Conference on Cloud Engineering (IC2E)* (2014), IEEE, pp. 610–614.

[32] FALKENAUER, E. A hybrid grouping genetic algorithm for bin packing. *Journal of heuristics 2*, 1 (1996), 5–30.

[33] FAN, X., WEBER, W.-D., AND BARROSO, L. A. Power provisioning for a warehouse-sized computer. *ISCA* (2007), 13.

[34] FEHLING, C., LEYMANN, F., RETTER, R., SCHUPECK, W., AND ARBITTER, P. Cloud Computing Patterns - Fundamentals to Design, Build, and Manage Cloud Applications.

[35] FELLER, E., RILLING, L., AND MORIN, C. Energy-Aware Ant Colony Based Workload Placement in Clouds. *GRID* (2011).

[36] FELTER, W., FERREIRA, A., RAJAMONY, R., AND RUBIO, J. An updated performance comparison of virtual machines and Linux containers. *ISPASS* (2015), 171–172.

[37] FERDAUS, M. H., MURSHED, M. M., CALHEIROS, R. N., AND BUYYA, R. Virtual Machine Consolidation in Cloud Data Centers Using ACO Metaheuristic. *Euro-Par 8632*, Chapter 26 (2014), 306–317.

[38] FERRETO, T. C., NETTO, M. A. S., CALHEIROS, R. N., AND DE ROSE, C. A. F. Server consolidation with migration control for virtualized data centers. *Future Generation Comp. Syst. 27*, 8 (2011), 1027–1034.

[39] FORSMAN, M., GLAD, A., LUNDBERG, L., AND ILIE, D. Algorithms for automated live migration of virtual machines. *Journal of Systems and Software 101* (2015), 110–126.

[40] GANESAN, R., SARKAR, S., AND NARAYAN, A. Analysis of SaaS Business Platform Workloads for Sizing and Collocation. *IEEE CLOUD* (2012), 868–875.

[41] GAO, Y., GUAN, H., QI, Z., HOU, Y., AND LIU, L. A multi-objective ant colony system algorithm for virtual machine placement in cloud computing. *J. Comput. Syst. Sci. 79*, 8 (2013), 1230–1242.

[42] HAMEED, A., KHOSHKBARFOROUSHHA, A., RANJAN, R., JAYARAMAN, P. P., KOLODZIEJ, J., BALAJI, P., ZEADALLY, S., MALLUHI, Q. M., TZIRITAS, N., VISHNU, A., KHAN, S. U., AND ZOMAYA, A. Y. A survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems. *Computing 98*, 7 (2016), 751–774.

[43] HERSKOVITS, J., LEONTIEV, A., DIAS, G., AND SANTOS, G. Contact shape optimization: a bilevel programming approach. *Structural and Multidisciplinary Optimization 20*, 3 (2000), 214–221.

[44] HINES, M. R., DESHPANDE, U., AND GOPALAN, K. Post-copy live migration of virtual machines. *SIGOPS Oper. Syst. Rev. 43*, 3 (July 2009), 14–26.

[45] HOLT, C. C. Author's retrospective on 'Forecasting seasonals and trends by exponentially weighted moving averages'. *International journal of forecasting 20*, 1 (Jan. 2004), 11–13.

[46] JAT, S. N., AND YANG, S. A Guided Search Non-dominated Sorting Genetic Algorithm for the Multi-Objective University Course Timetabling Problem. *EvoCOP 6622*, 9 (2011), 1–13.

[47] JENNINGS, B., AND STADLER, R. Resource Management in Clouds: Survey and Research Challenges. *Journal of Network and Systems Management 23*, 3 (2015), 567–619.

[48] JOHNSON, D. S. Vector Bin Packing.

[49] KAPLAN, J. M., FORREST, W., AND KINDLER, N. Revolutionizing data center energy efficiency, 2008.

[50] KHANNA, G., BEATY, K. A., KAR, G., AND KOCHUT, A. Application Performance Management in Virtualized Server Environments. *NOMS* (2006).

[51] KIRJNER-NETO, C., POLAK, E., AND DER KIUREGHIAN, A. An Outer Approximations Approach to Reliability-Based Optimal Design of Structures. *Journal of Optimization Theory and Applications 98*, 1 (1998), 1–16.

[52] KIVITY, A., KAMAY, Y., LAOR, D., AND LUBLIN, U. kvm: the Linux virtual machine monitor. *... of the Linux ...* (2007).

[53] KOZA, J. R. Genetic programming - on the programming of computers by means of natural selection. *Complex adaptive systems* (1993).

[54] KUNKLE, D., AND SCHINDLER, J. A Load Balancing Framework for Clustered Storage Systems. *HiPC 5374*, 1 (2008), 57–72.

[55] KUSIC, D., KEPHART, J. O., HANSON, J. E., KANDASAMY, N., AND JIANG, G. Power and performance management of virtualized computing environments via lookahead control. *Cluster computing 12*, 1 (2009), 1–15.

[56] LEE, K. Y., AND YANG, F. F. Optimal reactive power planning using evolutionary algorithms: A comparative study for evolutionary programming, evolutionary strategy, genetic algorithm, and linear programming. *IEEE Transactions on power systems 13*, 1 (1998), 101–108.

[57] LEGILLON, F., LIEFOOGHE, A., AND TALBI, E.-G. CoBRA - A cooperative coevolutionary algorithm for bi-level optimization. *IEEE Congress on Evolutionary Computation* (2012), 1–8.

[58] LEINBERGER, W., KARYPIS, G., AND KUMAR, V. Multi-Capacity Bin Packing Algorithms with Applications to Job Scheduling under Multiple Constraints. *ICPP* (1999), 404–412.

[59] LI, X., TIAN, P., AND MIN, X. A Hierarchical Particle Swarm Optimization for Solving Bilevel Programming Problems. In *Service-Oriented and Cloud Computing*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 1169–1178.

[60] LIEN, C.-H., BAI, Y.-W., AND LIN, M.-B. Estimation by Software for the Power Consumption of Streaming-Media Servers. *IEEE Trans. Instrumentation and Measurement 56*, 5 (2007), 1859–1870.

[61] LIU, H., JIN, H., XU, C.-Z., AND LIAO, X. Performance and energy modeling for live migration of virtual machines. *Cluster computing 16*, 2 (2013), 249–264.

[62] MANN, Z. Á. Approximability of virtual machine allocation: much harder than bin packing.

[63] MANN, Z. Á. Interplay of Virtual Machine Selection and Virtual Machine Placement. In *Service-Oriented and Cloud Computing*. Springer International Publishing, Cham, Aug. 2016, pp. 137–151.

[64] MANVI, S. S., AND SHYAM, G. K. Resource management for Infrastructure as a Service (IaaS) in cloud computing - A survey. *J. Network and Computer Applications 41* (2014), 424–440.

[65] MATHIEU, R., PITTARD, L., AND ANANDALINGAM, G. Genetic algorithm based approach to bi-level linear programming. *RAIRO-Operations Research 28*, 1 (Mar. 2011), 1–21.

[66] MELL, P. M., AND GRANCE, T. The NIST definition of cloud computing. Tech. rep., National Institute of Standards and Technology, Gaithersburg, MD, Gaithersburg, MD, 2011.

[67] MENG, X., ISCI, C., KEPHART, J. O., 0002, L. Z., BOUILLET, E., AND PENDARAKIS, D. E. Efficient resource provisioning in compute clouds via VM multiplexing. *ICAC* (2010), 11.

[68] MISHRA, M., DAS, A., KULKARNI, P., AND SAHOO, A. Dynamic resource management using virtual machine migrations. *IEEE Communications ... 50*, 9 (2012), 34–40.

[69] MURTAZAEV, A., AND OH, S. Sercon: Server Consolidation Algorithm using Live Migration of Virtual Machines for Green Computing. *IETE Technical Review 28*, 3 (Sept. 2014), 212.

[70] NGUYEN, S., ZHANG, M., JOHNSTON, M., AND TAN, K. C. Automatic Design of Scheduling Policies for Dynamic Multi-objective Job Shop Scheduling via Cooperative Coevolution Genetic Programming. *IEEE Transactions on Evolutionary Computation 18*, 2 (2014), 193–208.

[71] ODUGUWA, V., AND ROY, R. Bi-level optimisation using genetic algorithm. In *2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS 2002)* (2002), IEEE Comput. Soc, pp. 322–327.

[72] PANIGRAHY, R., TALWAR, K., UYEDA, L., AND WIEDER, U. Heuristics for vector bin packing. *research microsoft com* (2011).

[73] PANWALKAR, S. S., AND ISKANDER, W. A Survey of Scheduling Rules. *Operations Research 25*, 1 (1977), 45–61.

[74] PASCHKE, A., AND SCHNAPPINGER-GERULL, E. A Categorization Scheme for SLA Metrics. *Service Oriented Electronic Commerce 80*, 25-40 (2006), 14.

[75] PINHEIRO, R. L., SILVA, D. L., AND ATKIN, J. Analysis of Objectives Relationships in Multiobjective Problems Using Trade-Off Region Maps. *GECCO* (2015), 735–742.

[76] PIRAGHAJ, S. F., CALHEIROS, R. N., CHAN, J., DASTJERDI, A. V., AND BUYYA, R. Virtual Machine Customization and Task Mapping Architecture for Efficient Allocation of Cloud Data Center Resources. *Comput. J. 59*, 2 (2016), 208–224.

[77] PIRAGHAJ, S. F., DASTJERDI, A. V., CALHEIROS, R. N., AND BUYYA, R. A Framework and Algorithm for Energy Efficient Container Consolidation in Cloud Data Centers. *DSDIS* (2015), 368–375.

[78] POLI, R., WOODWARD, J., AND BURKE, E. K. A histogram-matching approach to the evolution of bin-packing strategies. In *2007 IEEE Congress on Evolutionary Computation* (2007), IEEE, pp. 3500–3507.

[79] QIN, X., ZHANG, W., 0049, W. W., WEI, J., ZHAO, X., AND HUANG, T. Towards a Cost-Aware Data Migration Approach for Key-Value Stores. *CLUSTER* (2012).

[80] RADCLIFFE, N. J. Forma Analysis and Random Respectful Recombination. *ICGA* (1991).

[81] RAGHAVENDRA, R., RANGANATHAN, P., TALWAR, V., WANG, Z., AND ZHU, X. No power struggles: Coordinated multi-level power management for the data center. In *ACM SIGARCH Computer Architecture News* (2008), ACM, pp. 48–59.

[82] RONG, H., ZHANG, H., XIAO, S., LI, C., AND HU, C. Optimizing energy consumption for data centers. *Renewable and Sustainable Energy Reviews 58* (May 2016), 674–691.

[83] ROSEN, R. Resource management: Linux kernel namespaces and cgroups. *Haifux* (2013).

[84] SHEN, S., VAN BEEK, V., AND IOSUP, A. Statistical Characterization of Business-Critical Workloads Hosted in Cloud Datacenters. *CCGRID* (2015), 465–474.

[85] SIM, K., AND HART, E. Generating single and multiple cooperative heuristics for the one dimensional bin packing problem using a single node genetic programming island model. *GECCO* (2013), 1549.

[86] SINHA, A., MALO, P., AND DEB, K. Efficient Evolutionary Algorithm for Single-Objective Bilevel Optimization.

[87] SINHA, A., MALO, P., AND DEB, K. A Review on Bilevel Optimization: From Classical to Evolutionary Approaches and Applications. *IEEE Transactions on Evolutionary Computation* (2017), 1–1.

[88] SOLTESZ, S., PÖTZL, H., FIUCZYNSKI, M. E., BAVIER, A. C., AND PETERSON, L. L. Container-based operating system virtualization - a scalable, high-performance alternative to hypervisors. *EuroSys 41*, 3 (2007), 275–287.

[89] SOMANI, G., AND CHAUDHARY, S. Application Performance Isolation in Virtualization. *IEEE CLOUD* (2009), 41–48.

[90] SPROTT, D., AND WILKES, L. Understanding service-oriented architecture. *The Architecture Journal* (2004).

[91] SRIKANTAIAH, S., KANSAL, A., AND ZHAO, F. Energy aware consolidation for cloud computing. In *Proceedings of the 2008 conference on Power aware computing and systems* (2008), San Diego, California, pp. 1–5.

[92] SUDEVALAYAM, S., AND KULKARNI, P. On Theory of VM Placement - Anomalies in Existing Methodologies and Their Mitigation Using a Novel Vector Based Approach. *IEEE CLOUD* (2011), 275–282.

[93] SUN, H., GAO, Z., AND WU, J. A bi-level programming model and solution algorithm for the location of logistics distribution centers. *Applied mathematical modelling 32*, 4 (Apr. 2008), 610–616.

[94] SVARD, P., LI, W., WADBRO, E., TORDSSON, J., AND ELMROTH, E. Continuous Datacenter Consolidation. In *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)* (2015), IEEE, pp. 387–396.

[95] TAN, B., MA, H., AND MEI, Y. A NSGA-II-based approach for service resource allocation in Cloud. *CEC* (2017), 2574–2581.

[96] TOMÁS, L., AND TORDSSON, J. Improving cloud infrastructure utilization through overbooking. *CAC* (2013), 1.

[97] UHLIG, R., NEIGER, G., RODGERS, D., SANTONI, A. L., MARTINS, F. C. M., ANDERSON, A. V., BENNETT, S. M., KÄGI, A., LEUNG, F. H., AND SMITH, L. Intel Virtualization Technology. *IEEE Computer 38*, 5 (2005), 48–56.

[98] VAQUERO, L. M., RODERO-MERINO, L., AND BUYYA, R. Dynamically scaling applications in the cloud. *Computer Communication Review 41*, 1 (2011), 45.

[99] VARASTEH, A., AND GOUDARZI, M. Server Consolidation Techniques in Virtualized Data Centers: A Survey. *IEEE Systems Journal* (2015), 1–12.

[100] VERMA, A., AND DASGUPTA, G. Server Workload Analysis for Power Minimization using Consolidation. *USENIX Annual Technical Conference* (2009), 28–28.

[101] VISWANATHAN, B., VERMA, A., AND DUTTA, S. CloudMap - Workload-aware placement in private heterogeneous clouds. *NOMS* (2012), 9–16.

[102] WALDSPURGER, C. A. Memory resource management in VMware ESX server. *ACM SIGOPS Operating Systems Review 36*, SI (Dec. 2002), 181–194.

[103] WANG, Y., AND 0002, Y. X. Energy Optimal VM Placement in the Cloud. *CLOUD* (2016), 84–91.

[104] WANG, Y., JIAO, Y.-C., AND LI, H. An evolutionary algorithm for solving nonlinear bilevel programming based on a new constraint-handling scheme. *IEEE Trans. Systems, Man, and Cybernetics, Part C 35*, 2 (2005), 221–232.

[105] WANG, Y., LI, H., AND DANG, C. A New Evolutionary Algorithm for a Class of Nonlinear Bilevel Programming Problems and Its Global Convergence. *INFORMS Journal on Computing 23*, 4 (2011), 618–629.

[106] WILCOX, D., MCNABB, A. W., AND SEPPI, K. D. Solving virtual machine packing with a Reordering Grouping Genetic Algorithm. *IEEE Congress on Evolutionary Computation* (2011), 362–369.

[107] WOOD, T., SHENOY, P. J., VENKATARAMANI, A., AND YOUSIF, M. S. Sandpiper - Black-box and gray-box resource management for virtual machines. *Computer Networks 53*, 17 (2009), 2923–2938.

[108] XAVIER, M. G., NEVES, M. V., ROSSI, F. D., FERRETO, T. C., LANGE, T., AND DE ROSE, C. A. F. Performance Evaluation of Container-Based Virtualization for High Performance Computing Environments. In *2013 21st Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP 2013)* (2013), IEEE, pp. 233–240.

[109] XAVIER, M. G., ROSSI, F. D., DE ROSE, C. A. F., CALHEIROS, R. N., AND GOMES, D. G. Modeling and simulation of global and sleep states in ACPI-compliant energy-efficient cloud environments. *Concurrency and Computation - Practice and Experience 29*, 4 (Feb. 2017), e3839.

[110] XIAO, Z., JIANG, J., ZHU, Y., MING, Z., ZHONG, S.-H., AND CAI, S.-B. A solution of dynamic VMs placement problem for energy consumption optimization based on evolutionary game theory. *Journal of Systems and Software 101* (2015), 260–272.

[111] XIE, J., JIANG, S., XIE, W., AND GAO, X. An Efficient Global K-means Clustering Algorithm. *JCP 6*, 2 (2011).

[112] XIONG, A.-P., AND XU, C.-X. Energy Efficient Multiresource Allocation of Virtual Machine Based on PSO in Cloud Data Center. *Mathematical Problems in Engineering 2014*, 6 (2014), 1–8.

[113] XU, J., AND FORTES, J. A. B. Multi-Objective Virtual Machine Placement in Virtualized Data Center Environments. *GreenCom/CPSCom* (2010).

[114] YAU, S. S., AND AN, H. G. Adaptive resource allocation for service-based systems. In *Proceedings of the First Asia-Pacific Symposium on Internetware* (2009), ACM, p. 3.

[115] ZHANG, Y., ZHENG, Z., AND LYU, M. R. Exploring Latent Features for Memory-Based QoS Prediction in Cloud Computing. In *Reliable Distributed Systems (SRDS), 2011 30th IEEE Symposium on* (2011), pp. 1–10.

[116] ZHU, X., YU, Q., AND WANG, X. A Hybrid Differential Evolution Algorithm for Solving Nonlinear Bilevel Programming with Linear Constraints. In *2006 5th IEEE International Conference on Cognitive Informatics* (2006), IEEE, pp. 126–131.

[117] ZIO, E., AND BAZZO, R. A clustering procedure for reducing the number of representative solutions in the Pareto Front of multiobjective optimization problems. *European Journal of Operational Research 210*, 3 (2011), 624–634.

[118] ZIO, E., AND BAZZO, R. A Comparison of Methods for Selecting Preferred Solutions in Multiobjective Decision Making. In *Computational Intelligence Systems in Industrial Engineering*. Atlantis Press, Paris, Nov. 2012, pp. 23–43.