Literature Review

Leo Unbekandt
leo@unbekandt.eu

Cranfield University
School of engineering
Msc Computational and Software Techniques in Engineering

May 9, 2014

Contents

1 Background **2**

2 Motivation **2**

3 Algorithms **3**
3.1 Linear Programming 3
3.2 Bin packing 5
3.2.1 Different variants 5
3.2.2 Their application in resource allocation 6
3.3 Others 8
3.3.1 Ant colony algorithms 8
3.3.2 Genetic algorithms 9
3.3.3 Network flows 10

4 Real data analysis **11**

Introduction

These 10 last years, the Internet has grown at an exponential rate. More and more people are connected, and using always more numerous kinds of device. As the demand is raising strongly, service providers have to be able to support an increasingly amount of customers on their infrastructure. In 2006, Amazon launched Amazon Web Services (AWS) is born, popularising the concept of cloud computing. This platform allows developers and companies to allocate resources on demand, and these are provisioned instantly.

The market of cloud computing has grown really quickly and forecasts predict that this expansion won't stop in the near future. The following study gathers information from different surveys[1]. Currently, the public cloud providers represent a market of 47.4 billion of dollars and is expecting to grow up to 107.2 billion of dollars on only 4 years.

It is no coincidence if the industry and the academic environments are really interested by this domain. On the one hand, it is a field where money is present and on the second hand, we can consider that we are only at the beginning of an era: a lot of things have to be developed, to be optimized and to be found.

1 Background

An interesting definition of the Cloud Computing[2] has been written by the National Institute of Standards and Technology:

> "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."

Different kind of clouds are specified, if Amazon is a "public" cloud, this is not the only way to use a cloud infrastructure: private cloud or hybrid clouds mixing private and public cloud infrastructures are being developed more and more. Thanks to open-source projects like Openstack[3], cloud environments can be installed on private infrastructures. This is sometimes necessary or requested for security, performance or data control purposes.

The evolution of the paradigm of cloud computing has been made possible thanks to different technologies. The virtualisation, as explained by Paul Barham et al.[4] allows servers to be splitted in different sub-components, isolated from each other, sharing the resources of the physical machine.

Technologies have been developed to give people much more flexibility in the way to manage their virtual machines, also called instances. Actually, the concept of live migration, which is detailed in the work of Christopher Clark et al.[5], has been built to move instances from one physical host to another without interrupting the activity of anything running in the virtual machine. The memory is kept intact of course, but also the running connections. The instance may seem frozen for a few second when the migration is finalized, but nothing is disrupted.

2 Motivation

The legitimate question is "Why do people migrate their infrastructure to a cloud infrastructure?". The answers are multiple, Valentina Salapura explains how a virtualized environment improves the resiliency of an infrastructure [6]. More precisely, when a service requires to be scalable, highly available and fault tolerant, using cloud technologies is essential. In the case of disaster recovery scenarios, they are highly simplified and cheaper thanks to those environments.

As a result the infrastructures are composed of a certain amount of physical machines (PMs) which could be dispatched among different data centers, and each of these PMs, containers a variable number of virtual machines (VMs). The problematic which is now interesting concerns the assignment of these VMs, what is the optimal distributions of the instances among the different servers? It depends of what characteristic has to be optimized.

Thomas Setzer and Alexander Stage base their study on the statement that energy represents up to 50% of operating costs of an infrastructure[7]. That's why there is a need to optimize it. Using the virtual machine reassignment through live migrations, they are looking at consolidating the VMs on the physical servers. Consolidating an infrastructure consists in reducing the number of PMs which are hosting instances without disturbing the performance of these. After this operation, useless PMs can be suspended and electricity is saved, then when more computational power is required they are resumed dynamically.

In the publication *An adaptive Resource Provisioning for the Cloud Using Online Bin Packing*[8], the authors also introduce their subject by explaining that it has been estimated that Amazon manages more than half a million of physical servers around the world and that it must be a priority for them to reduce their expenses by consolidating their infrastructure.

3 Algorithms

We have seen that cloud computing is a hot topic in the Internet industry which results in a lot of new problematics in computer science. The resource allocation problem is one of them. All over the world, universities have started studying different approaches of allocation optimisation. The different algorithms listed in this document gather publications around the virtual machine assignment and reassignment on a set of physical machines.

3.1 Linear Programming

Also known as Linear optimization. It is specialisation of mathematical programming, which is focused on linear functions. The main goal of linear programming is to find a maximum or a minimum to a linear function given a set of constraints, in other words: maximizing profits while minimizing costs. In scope of resource allocation, it is required to define the different variables, the function we want to optimize and the constraints linked to the variables.

In their work, Ruben Van den Bossche, Kurt Vanmechelen and Jan Broeckhov[9] are working with linear programming. The aim of their study is to define a way to optimize the number of allocated virtual machines splitted in different cloud infrastructures. Different constraints are defined to setup the scope of the function to minimize.

Equation 1 Example of linear optimization problem

$$Minimize \sum_{k=1}^{A} \sum_{l=1}^{T_k} \sum_{i=1}^{I} \sum_{j=1}^{C} (y_{klij} \cdot (ni_{kl} \cdot pi_j + no_{kl} \cdot po_j) + \sum_{s=1}^{S} (p_{ij} \cdot x_{klijs}))$$

Equation 1 is the problem they want to solve, in this case a cost minimization problem. How can we minimize for each task t of each application k in each virtual machine i of each cloud infrastructure j the price of the input and output bandwidth ($ni \cdot pi_j$ and $no_{kl} \cdot po_j$) and the price the requested virtual machines ($x_{klijs} \cdot p_{ij}$) at each unit of time (S)

Equation 2 Example of constraints in a linear program

$$\forall j \in [1, C], s \in [1, S] : \sum_{k=1}^{A} \sum_{l=1}^{T_k} \sum_{i=1}^{I} cpu_i \cdot x_{klijs} \leq maxcpu_j$$

The *Equation 2* defines a constraint from the linear problem, which explains that in each cloud, at each unit of time, the sum of all the tasks run on all the virtual machines instantiated should be less than the number of CPUs available. (There is note that in the case of public clouds, the amount of CPU is considered unlimited so this constraint becomes void).

The work of Mark Stillwell, Frédéric Vivien and Henri Casanova, which focuses virtual machine resources allocation in heterogeneous environment [10] also start by defining a formal model based on linear programming. However, as explained in this publication, resolving such a problem requires an exponential time, linked to the amount of allocations to achieve. As a result using directly this solution on an important workload is not feasible.

The work of Neal E. Young about linear optimization relaxation[11] has been used to simplify the original problem and transform it from an exponential complexity to a polynomial complexity. The "random rounding" is a probabilistic approach which modifies some of the constraints by a weaker one.

Equation 3 Application of random rounding

$$\text{constraint before: } 0 \leq x \leq 1$$

$$\text{constraint after: } x_r \in 0, 1$$

$$x_r = 1 \text{ with a probability of } x \text{, otherwise: } 0$$

However, the RRND approaches is quickly discarded as the results are not good enough in the case of resource allocations in heterogeneous environment.

3.2 Bin packing

Bin packing is one of the most common approach to resource allocation or re-allocation in a cloud environment. It consists in representing "bins" associated to a storage capacity and "items" which have to be packed into those bins.

3.2.1 Different variants

Two main types of bin packing algorithms exist. On the one hand, those considered as "offline". They consider that we have access to all the items to find the optimal packing on the different bins. This problem is a NP-hard problem, there is no, to this day, a polynomial way to solve this problem. That is why to answer this problem in a reasonable duration, different heuristics have to be defined. The most common have been studied by David Stiefler Johnson in 1967[12]:

Algorithm Name	Description
First Fit (FF)	Pack the item in the first bin with a large enough capacity
Best Fit	Pack the item in the bin which will have the less capacity after packing
Worst Fit	Opposite of Best Fit: Pack the item in the bin with the biggest capacity
Next Fit	Same as FF except that instead of reconsidering the first bin after packing, the current one then the next one is considered
*-Fit Decreasing	First, sort the items in a decreasing order, then apply any of the *-Fit algorithm

Those different algorithms reduce the complexity of the packing operation to $O(n \log n)$. But as [12]'s title explains: they are "Near-Optimal". The issue is finally to find the best ratio optimality/complexity.

On the other hand, the "online" algorithms, which, on the contrary, are packing items at the time they are arriving. In this case bins are already partially filled with other items, and it is not always possible to move those. Thus, the main goal is to find the best assignment for the newly coming item. Previous *-Fit could be directly used. However, it is really limited to pack one item in a set of bin, this is why different algorithms have been developed

To answer more precisely to the cloud resource allocation problem, some people have defined some variants of those two main categories of bin packing algorithms. G. Gambosi and A. Postiglione and M. Talamo have developed a "relaxed online bin packing" algorithm[13]. It may be represent as a mix between online and offline bin packing. When a new item has to be packed, it allows an additional limited number of moves among the currently packed items.

Another interesting variant is the dynamic online bin packing defined by Joseph Wun-Tan. It differentiates itself from standard online bin packing by allowing items to be removed from bins. Static online bin packing does not allow these items changes, once an item has been placed it does not move anymore.

3.2.2 Their application in resource allocation

In the scope of virtual machines assignment on physical hosts, the bins are the physical servers and the items are the virtual machines we want to assign, and consequently, the services or tasks we want to host. Some additional aspects have to be considered: virtual machines need different resources like memory, CPU, persistent storage (disk), network input/output. So often, the items we want to pack are multidimensional items, and we speak of multidimensional vector bin packing. Another interesting point is that moving a virtual machine from one host to another has a cost which may be important, and which prevent doing numerous migrations in the meantime.

In the work about online bin packing for virtual machines allocation of Weijia Song et al.[14], they consider first, that a virtual machine only has one dimension, its CPU consumption. From that point they study which algorithm may fit this particular problem. They reject "strict" online bin packing, because in realistic situations it is uncommon to know exactly the future consumption of a virtual machine, so it is necessary to move it afterward, when we can measure it. Moreover, as VMs can be migrated easily, there is no reason not considering it if the resulting performance is better. "Relaxed online bin packing" allows movements when a new item is packed, but an item cannot be resized. "Dynamic online bin packing" is thought inadequate in this context too, but often, when an virtual machine has to move the best solution is not always to remove it then repack it, but to move others instances which are easier to move.

This is why in[14], they decided to build an online bin packing algorithm which suits the virtualisation environment: “Variable Item Size Bin Packing”, its characteristics are the following.

- As relaxed online bin packing, it allows movements when a new item is packed
- Stronger limit of movements, to avoid executing too many migrations
- A `change` operation is defined to modify the size of an item in a bin

They extend their algorithm to multidimensional vectors by considering the biggest value among the different dimensions of a vector, so the problem returns to one-dimension. Using this way to simplify the problem is working in some cases. Commonly when a resource consumption increases the others are following. For example an application having a high network bandwidth requirement, would also have a high CPU consumption. Finally, they admit that this solution would work quite poorly in the case of instances with non-proportional requirements.

In [10], we have seen that the first approach of the author was around linear programming, but the main part of their work is defining a way to apply multidimensional vector bin packing to heterogeneous environments. On a first side, they deal with the multidimensional aspect of this problem. It is necessary to specify how to sort the items because there is not natural way to sort these vector.

- Value of the maximal dimension
- Sum of all dimensions
- Ratio of the max/min
- Difference max-min
- Lexicographic order
- None

Most of the previous algorithms are not considering the way the bins are used. In this publication, as it is targeting heterogeneous infrastructure, the order the bins are sorted when executing any algorithm matters. All the previous way to sort the items can be applied to the set of bins.

All these previous possibilities of ordering among the virtual machines and physical hosts are combined and result in a “meta” algorithm (METAHVP) which takes the best result out of the different combinations of one items ordering and one bins ordering. After individual analysis, some sort types are removed from the meta algorithm to improve its runtime. (METAHVP-LIGHT)

The simulation achieved to test these heuristics are comparing the results to those which have been found using the linear programming method and those obtained using greedy algorithms (*-Fit). The conclusion is that METAHVP has the best results over all the other, and METAHVPLIGHT achieves this result in on tenth of METAHVP's runtime.

Finally, according to what we want to study there are several possible solutions using bin packing. Semantically, it is really comfortable to compare bins with physical servers and items with virtual machines, it allows a very natural vision of this problem.

3.3 Others

To deal with mathematical optimization and approximative solution of NP-complete problems, Ants colony algorithms, genetic algorithms and some other famous methods, based on statistical analysis.

3.3.1 Ant colony algorithms

In [15], [16] and [17], the ant colony algorithms are studied. As we can see in the following graph:

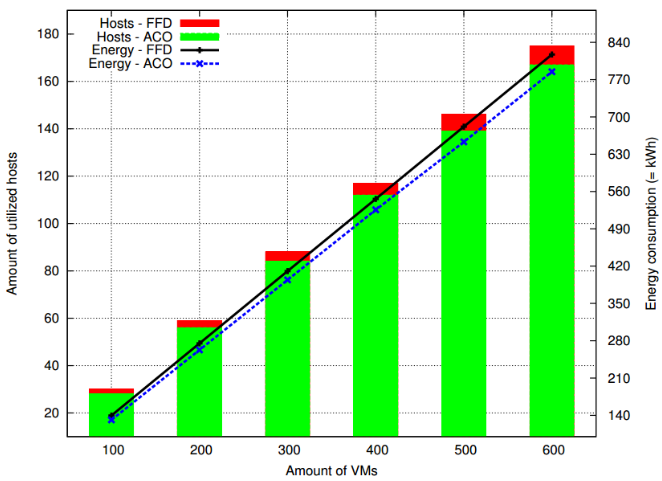

Figure 1: Comparison between First Fit Decreased and Ant Colony algorithms in [16]

The simulation shows that the ant colony gets better performance than a simple greedy First Fit Decreasing, however this improvement is not free:

VMs	Policy	Hosts	Execution time	Energy (= kWh)	Energy gain (= %)
100	FFD	30	0.39 sec	139.62	
	ACO	28	37.46 sec	131.41	5.88
200	FFD	59	0.58 sec	275.13	
	ACO	56	4.51 min	262.83	4.47
300	FFD	88	0.77 sec	410.65	
	ACO	84	15.04 min	394.28	3.98
400	FFD	117	1.03 sec	546.16	
	ACO	112	34.23 min	525.75	3.73
500	FFD	146	1.39 sec	681.67	
	ACO	139	1.17 h	653.17	4.18
600	FFD	175	1.75 sec	817.19	
	ACO	167	2.01 h	784.75	3.96

Figure 2: Runtime of First Fit Decreased and Ant Colony algorithms in [16]

When the number of nodes becomes bigger, the time spent to find the optimal allocation grows hugely, it is thousands times longer than a simple First Fit Decreasing for 3 to 5 percents of improvement. For analysis purpose it is something interesting to get better results, but in a realistic point of view, this operation can not take several hours as it should be repeated often.

3.3.2 Genetic algorithms

Genetic algorithms (GA) are heuristics based on natural selection. Generations of solutions are mutating, inheriting with and from each other to result in close to optimal results. [18] and [19] focused on them to solve the virtual machines assignment problem. In the work of David Wilcox et al.[18], simulations are comparing GA with *-Fit algorithms.

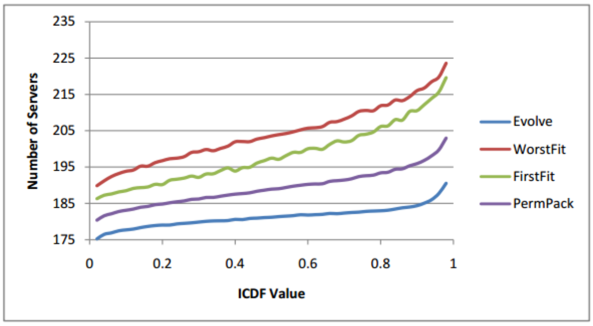

Fig. 7. A comparison of the the number of servers found.

Figure 4: Results of simulations using a genetic algorithm[18]

On the following graphs, ICDF stands for "inverse cumulative distribution function" also known as "quantile function", the authors use it to represent the load: "Using the icdf, we can specify a percentile value and obtain a corresponding load which can be passed to the assignment algorithm".

The conclusion which is that GA tends to consume less physical hosts, at any load, the number of PMs is largely under the amount of servers used by the other bin packing algorithms. As a direct consequence, the PMs which are over-capacitated (where the amount of VMs exceed the resource capacity of the physical sever), is much more high. For this reason, this approach can hardly be used in environment where a SLA (Service Level Agreement) has to be respected, because if there are overloaded servers, some applications or tasks running of them will be slowed by this situation.

3.3.3 Network flows

Network flows are basically directed graphs where each edge has a capacity and a flow. The main property is that each node of this graph must have an equal sum of flows from the edges directed to it and leaving from it, except for two particular type of nodes: "the source node" and "the sink node".

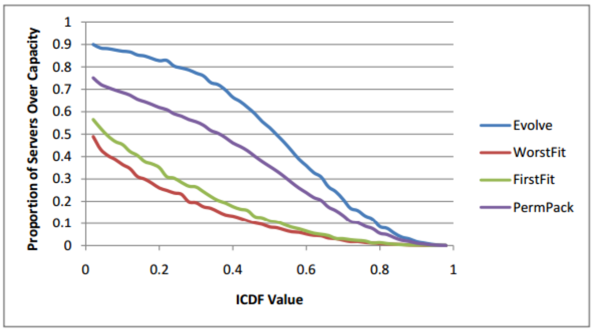

Fig. 6. A comparison of the proportion of servers over capacity.

Figure 3: Results of simulations using a genetic algorithm[18]

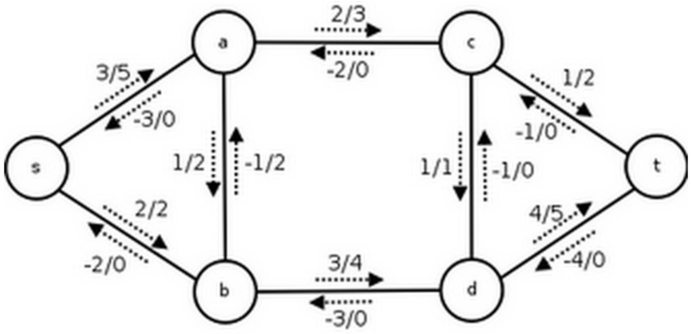

Figure 5: Example of network flow directed graph

Some people have used this concept to build a model to solve the resource allocation problem, to fine a close to optimal solution. Kimish Patel, Murali Annavaram and Massoud Pedram worked on resource assignment in datacenter[20], considering an heterogeneous environment as in [10]. Each set of similar servers, considered as a pool of servers is represented by a node, with a capacity different from each other according to the differences between two pools of servers.

Unfortunately, this technique does not seem to be used for virtual machines allocation, and the link between this method and the problem we are dealing with is not obvious at all.

4 Real data analysis

Most of the cited works in the literature review are basing their work on simulations. In the experiments, simulation tools like SimGrid[21] or CloudSim[22] are used to simulate the behavior of one or multiple cloud infrastructures.

The data may be generated randomly or following some statistical rules, but often, workloads are based of extract of real workload. Typically, Google is releasing workloads of its own production infrastructure.

In 2012, Google sponsored the ROADEF contest (Operational research and decision support French society)[27]. The contest was focusing the machine reassignment problem based on Google workload. Each attendee had to find the best solution find solution. Some of them resulted in an official publication like “Heuristics and matheuristics for a real-life machine reassignment problem” from Ramon Lopes, Vinicius W.C. Morais, Thiago F. Noronha and Vitor A.A. Souza[23]. They based their work on linear programming. However in [24] and [25], the authors have used around the bin packing algorithms. Unfortunately, the work of the winner has not been published so we are not able to see which algorithm has been used to achieve the best reassignment.

Cooonclusion

To conclude, it is obvious that a lot of different approaches to the resource allocation problem in a cloud computing environment have been studied. However, those may be really difficult to compare with each other, as various mathematical tools are used and because each algorithm focuses on something else. But overall we have seen that the two biggest approaches are using linear programming and bin packing. For each of these methods there are a lot of variants which are really different from each other.

Another aspect is that most of the experiments are simulations. When they are not, the experiment on real cloud infrastructure are mostly done on a really limited number of nodes. Building a model is interesting, but how close to the reality are them? How close would be the results from simulated experiences from real life experiences?

References

[1] Louis Colombus. Roundup of cloud computing forecasts update, 2013. http://www.forbes.com/sites/louiscolumbus/2013/11/16/roundup-of-cloud-computing-forecasts-update-2013/, November 2013.

[2] Peter Mell and Timothy Grance. The NIST definition of cloud computing. *Recommandation of National Institute of Standards and Technology*, Spetember 2011.

[3] Openstack Community. Official website for the openstack project. http://www.openstack.org/.

[4] Barham Paul, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield. Xen and the art of virtualisation. In *19th ACM Symposium on Operating Systems Principles*, 2003.

[5] Clark Christopher, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield. Live migration of virtual machines. In *2nd Symposium on Networked Systems Design and Implementation (NSDI'05)*, 2005.

[6] Valentina Salapura. Cloud computing: Virtualization and resiliency for data center computing. In *30th International Conference of Computer Design*, 2012.

[7] Thomas Setzer and Alexander Stage. Decision support for virtual machine reassignments in enterprise data centers. In *Network Operations and Management Symposium*, 2010.

[8] Weijia Song, Zhen Xiao, Qi Chen, and Haipeng Luo. An adaptive resource provisioning for the cloud using online bin packing. *Transactions on Computers*, 2013.

[9] Kurt Vanmechelen Ruben Van den Bossche and Jan Broeckhov. Cost-optimal scheduling in hybrid iaas clouds for deadline constrained workloads. In *3rd International Conference on Cloud Computing*, 2010.

[10] Mark Stillwell, Frédéric Vivien, and Henri Casanova. Virtual machine resource allocation for service hosting on heterogeneous distributed platforms. In *Parallel & Distributed Processing Symposium*, 2012.

[11] Neal E. Young. Randomized rounding without solving the linear program. In *6th ACM-SIAM Symposium on Discrete Algorithms*, 1995.

[12] David Stifler Johnson. *Near-Optimal Bin Packing Algorithms.* PhD thesis, Massachusetts Institute of Technology, 1967.

[13] G. Gambosi, A. Postiglione, and M. Talamo. Algorithms for the relaxed online bin-packing model. *SIAM Journal on Computing*, 30(5), 2000.

[14] Weijia Song, Zhen Xiao, Qi Chen, and Haipeng Luo. Adaptive resource provisioning for the cloud using online bin packing. In *Transactions on Computers*, 2013.

[15] Lskrao Chimakurthi and Madhu Kumar SD. Power efficient resource allocation for clouds using ant colony framework. *CoRR*, 2011.

[16] Eugen Feller, Louis Rilling, and Christine Morin. Energy-aware ant colony based workload placement in clouds. In *Proceedings of the 2011 IEEE/ACM 12th International Conference on Grid Computing*, 2011.

[17] Linan Zhu, Qingshui Li, and Lingna He. Study on cloud computing resource scheduling strategy based on the ant colony optimization algorithm. *IJCSI International Journal of Computer Science*, 9(5), 2012.

[18] David Wilcox, Andrew McNabb, Kevin Seppi, and Kelly Flanagan. Probabilistic virtual machine assignment. In *The First International Conference on Cloud Computing, GRIDs, and Virtualization*, 2010.

[19] David Wilcox, Andrew McNabb, and Kevin Seppi. Solving virtual machine packing with a reordering grouping genetic algorithm. In *IEEE Congress on Evolutionary Computation (CEC)*, 2010.

[20] Kimish Patel, Murali Annavaram, and Massoud Pedram. Nfra: Generalized network flow-based resource allocation for hosting centers. *IEEE Transactions on Computers*, 62(9), 2013.

[21] INRIA. Simgrid official website. http://simgrid.gforge.inria.fr/.

[22] Melbourne University. Project page of cloudsim. https://code.google.com/p/cloudsim/.

[23] ROADEF members. Roadef contest homepage. http://challenge.roadef.org/.

[24] Ramon Lopes, Vinicius W.C. Morais, Thiago F. Noronha, and Vitor A.A. Souza. Heuristics and matheuristics for a real-life machine reassignment problem. *International Transactions in Operational Research*.

[25] Renaud Masson, Thibaut Vidal, Julien Michallet, Puca Huachi Vaz Penna andVinicius Petrucci, Anand Subramanian, and Hugues Dubedout. An iterated local search heuristic for multi-capacity bin packing andmachine reassignment problem. *Expert Syst. Appl.*, 40(30), 2013.

[26] Brian Kell and Willem-Jan van Hoeve. An mdd approach to multidimensional bin packing. *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems Lecture Notes in Computer Science*, 7874, 2013.

[27] Joseph Wun-Tat, Chan Tak-Wah Lam, and Prudence W.H. Wong. Dynamic bin packing of unit fractions items. *Theoretical Computer Science*, 409(8), 2008.