Future Generation Computer Systems 82 (2018) 89–103

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Semi-online task assignment policies for workload consolidation in cloud computing systems

Vincent Armant *, Milan De Cauwer *, Kenneth N. Brown, Barry O'Sullivan

Insight Centre for Data Analytics, Department of Computer Science, University College Cork, Ireland

HIGHLIGHTS

- We tackle the challenge of semi-online workload consolidation.
- We propose a formal framework capturing the semi-online consolidation problem.
- We propose a dynamic allocation algorithm based on incremental merging of bins.
- We adapt bin packing heuristics enhanced by local search to the semi-online context.
- Our approach "First Merged Fit" saves up to 40% more resources than other policies.

ARTICLE INFO

Article history:
Received 1 September 2017
Received in revised form 11 December 2017
Accepted 22 December 2017
Available online 3 January 2018

Keywords:
Cloud computing
Workload consolidation
Semi-online policies
Stochastic task duration

ABSTRACT

Satisfying on-demand access to cloud computing infrastructures under quality-of-service constraints while minimising the wastage of resources is an important challenge in data centre resource management. In this paper we tackle this challenge in a semi-online workload management system allocating tasks with uncertain duration to physical servers. Our semi-online framework, based on a bin packing approach, allows us to gather information on incoming tasks during a short time window before deciding on their assignments. Our contributions are as follows: (i) we propose a formal framework capturing the semi-online consolidation problem; (ii) we propose a new dynamic and real-time allocation algorithm based on the incremental merging of bins; and (iii) an adaptation of standard bin packing heuristics with a local search algorithm for the semi-online context considered here. We provide a systematic study of the impact of varying time-period size and varying the degrees of uncertainty on the duration of incoming tasks. The policies are compared in terms of solution quality and solving time on a data-set extracted from a real-world cluster trace.

Our results show that, around periods of high demand, our best policy saves up to 40% of the resources compared to the other polices, and is robust to uncertainty in the task durations. Finally, we show that small increases in the allowable time window allows a significant improvement, but that larger time windows do not necessarily improve resource usage for real world datasets.

1. Introduction

As the demand for IT services continues to increase, worldwide deployment of large data centres is continuing to grow. Those data centres consume enormous amounts of electricity, estimated at 70 terawatt hours for the USA alone in 2014, at a cost of 7 billion dollars [1]. It has been estimated that only 6 to 12% of electricity used by data centres can be attributed to productive computation [2]. Opportunities for reducing the energy consumption of data centres include more efficient cooling, enhanced power management for idle running, and aggressive resource sharing through virtualization. The latter strategy is the one explored in this paper. The aim of resource consolidation through virtualization is to increase the utilisation of a subset of servers. Consolidation is usually achieved by allocating multiple tasks on the same physical machine. In turn, workload consolidation allows data centre operators to spread workload over a smaller set of machines so that those remaining unused can be powered down or put into a standby mode. Data centres are usually over-provisioned so they can cope with high fluctuations in demand from clients. Having a much larger pool of servers than needed also allows the design of fault-resilient systems [2]. As a consequence of over-provisioning, the workload can be dispatched without delays due to resource

* Corresponding authors.
E-mail addresses: vincent.armant@insight-centre.org (V. Armant), milan.decauwer@insight-centre.org (M. De Cauwer), ken.brown@insight-centre.org (K.N. Brown), barry.osullivan@insight-centre.org (B. O'Sullivan).

https://doi.org/10.1016/j.future.2017.12.035
0167-739X/

scarcity. In such a context one aims to dispatch tasks on machines so that resource wastage is minimised.

We leverage semi-online optimisation techniques in which workload allocations must be made without full knowledge of future demands. A semi-online formulation of the workload consolidation problem gathers information on incoming tasks for a short period of time. It may allow an operator to take more informed decisions than the fully online formulation while keeping control of delays in task deployments. In a cloud computing production environment, it is often the case that the duration for which a task will lock resources is either approximated or not known at all. This fits the new challenge of on-demand allocations in which demands are guaranteed to be satisfied in real-time. We therefore formalise the workload consolidation problem as a semi-online bin-packing problem whereby each bin maps to a machine and each item maps to a task.

The remainder of this paper is organised as follows. Section 2 discusses relevant work from both the cloud computing and optimisation communities. In Section 3 we provide a mathematical formulation of the on-demand bin packing (ODBP) problem. Section 4 introduces a novel methodology based on bin merging to solve the ODBP. Section 5 shows the performance of our approach compare to adapted heuristics of the related work in terms of solving time and solution quality. We demonstrate that our bin merging policy can achieve reductions in energy use of up to 40% over the comparator approaches. We show that the policy is relatively robust to increased errors in the predicted duration of the tasks. Finally, we show that moving from the pure online problem to the semi-online problem, with relatively small decision time windows, has a significant impact on the solution quality, but that all policies quickly stabilise and do not benefit further from longer time windows.

2. Related work

Workload consolidation in data centres (DC) is a key challenge in the operations of cloud computing systems through which operators efficiently dispatch a workload on a pool of servers over which operations are virtualised [3]. Workload consolidation aims at maximising the usage of servers by grouping tasks to run concurrently on fewer machines. This technique is used to maintain control over the potentially high economic and environmental cost [4]. Due to the large spectrum of technologies that implement cloud computing systems, there is a vast body of literature reviewing efforts to optimise task placement (see surveys [5–7]). Depending on the technologies at play, this can be achieved either dynamically or statically. In the dynamic version, the DC management system is allowed to migrate tasks across hosts [8,9]. On the other hand, static workload consolidation does not allow migrations and focuses on consolidating the initial task placement. The operational setting in which our work stands is *on-demand* static placement of tasks. Moreover an efficient placement of tasks may reduce the cost of migrating tasks from a physical host to an other.

Policies for static consolidation have been studied in [10]. The authors leverage machine learning methods to predict the often unknown size of the virtual machines (VMs). They show that different prediction methods lead to assignments with significant differences in terms of resource usage. Our study takes place after this prediction step. We suppose that the size of each task can be predicted by the consolidation system upon its arrival. In the context of cloud gaming, the authors of [11] have considered the play-request dispatching problem where the aim is to minimise the total rental cost of a cloud platform. Although this problem has strong connections to ours, it differs in that it assumes that each task has the same size as well as perfect information concerning the their duration.

We tackle the semi-online formulation of the on-demand workload consolidation where tasks have to be allocated to servers in real-time. While the vast majority of the work carried on workload consolidation considers either the offline [12,13] or online [14–16] setting, we cast the problem in a semi-online framework by considering a short period of a few seconds within which tasks are grouped before being allocated to hosts. More precisely, we build upon the offline workload consolidation problem discussed in [17].

In classical bin packing (BP) (see survey [18]) we are given a set of items along with their sizes and a set of bins having equal capacities. The objective is to find an assignment from items to bins such that the total number of bins is minimised. Many variants of BP have been studied extensively in both offline and online settings. Van Hentenryck et al. [19] considered a number of different approaches to online stochastic optimisation under time constraints, including various packing models. The authors exploited distributions over future task arrivals. Delaying or rejecting tasks is allowed but neither of which applies to the problem we address. The semi-online bin packing problem, also known as batch bin packing, has been described in [20] in which the authors derive lower bounds on the minimum number of bins to be used to accommodate tasks over time. That work does not model the relationship between batches induced by item durations.

Several variants of BP are concerned with items that have a 'lifespan' within the bins. In offline dynamic BP (DBP), we are given items defined by their size and lifespan. The objective is to minimise the maximum number of bins over a given horizon. Here, we study the problem from a different viewpoint where we aim to minimise the cumulative cost. As an extension, fully DBP [21,22] considers rearranging the items across the bins to retain a minimal number of used bins. In our study, instead of rearranging tasks to optimise the consolidation, we explore the possibility of gathering the tasks during a short time window to make more informed decisions for their placement. In [11], the authors revisited the online version of the dynamic BP problem in which items are characterised by size, arrival time, and an arbitrary departure time with the objective being to minimise the maximum number of bins ever used over time.

Finally, the family of online scheduling problems are to some extent related to our work. In [23], the authors propose and evaluate online scheduling policies and define new distance metrics used in the best-Fit family of heuristics. The objective function under consideration is to minimise the queuing delay under constrained overall computational capacity. In our study, we propose a new approach for the semi online context that enforces a fixed delay satisfying the on demand placement. We also compare the efficiency of this heuristic against adapted Bin Packing heuristics.

3. Minimising resource wastage

On-line policies decide the placement of incoming tasks as soon as they arrive in the system. Such a framework must fully satisfy on-demand Quality-of-Service (QoS) requirements, guaranteeing the real-time placements of tasks. However, due to the lack knowledge of which tasks might come next, on-line placement strategies may provide poor consolidation solutions and waste more resources than required.

On the other hand, for efficient resource utilisation, off-line approaches consider the task placement as a batch optimisation problem for which the incoming tasks are known in advance. The knowledge of forthcoming tasks allows a better consolidation but involves more sophisticated techniques that may required an expensive solving time. In the context of on-demand placement, neither the existence of incoming tasks nor their duration can be known in advance.