European Journal of Operational Research 227 (2013) 62–75

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

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

Discrete Optimization

Using matrix approximation for high-dimensional discrete optimization problems: Server consolidation based on cyclic time-series data

Thomas Setzer[a,*], Martin Bichler[b]

[a] Karlsruhe Institute of Technology, Englerstraße 14, 76131 Karlsruhe, Germany
[b] Technische Universität München, Boltzmannstraße 3, 85748 Garching, Germany

ARTICLE INFO

Article history:
Received 20 December 2011
Accepted 6 December 2012
Available online 29 December 2012

Keywords:
Matrix approximation
Multi-dimensional packing
Server consolidation
Dimensionality reduction

ABSTRACT

We consider the assignment of enterprise applications in virtual machines to physical servers, also known as server consolidation problem. Data center operators try to minimize the number of servers, but at the same time provide sufficient computing resources at each point in time. While historical workload data would allow for accurate workload forecasting and optimal allocation of enterprise applications to servers, the volume of data and the large number of resulting capacity constraints in a mathematical problem formulation renders this task impossible for any but small instances. We use singular value decomposition (SVD) to extract significant features from a large constraint matrix and provide a new geometric interpretation of these features, which allows for allocating large sets of applications efficiently to physical servers with this new formulation. While SVD is typically applied for purposes such as time series decomposition, noise filtering, or clustering, in this paper features are used to transform the original allocation problem into a low-dimensional integer program with only the extracted features in a much smaller constraint matrix. We evaluate the approach using workload data from a large data center and show that it leads to high solution quality, but at the same time allows for solving considerably larger problem instances than what would be possible without data reduction and model transform. The overall approach could also be applied to similar packing problems in service operations management.

1. Introduction

Today's larger data centers are typically hosting thousands of enterprise applications such as Enterprise Resource Planning (ERP) modules or database applications with different workload characteristics from many customers. The underlying IT infrastructures may contain hundreds of physical servers and various other hardware components. There is even a trend to increase the size of data centers to leverage economies of scale in their operation (Digital Realty Trust, 2010).

To reduce the number of required servers, server virtualization has been increasingly adopted during the past few years (Hill et al., 2009). Server virtualization refers to the abstraction of hardware resources and allows the hosting of multiple virtual machines (VM) including business applications plus underlying operating system on a single physical server. The capacity of a physical server is then shared among the VMs, leading to increased utilization of physical servers. While traditional capacity planning in Operations Research relies on queuing theory or simulation to predict the performance of a single dedicated system given available resources and resource demands (Hühn et al., 2010), new challenges arise in the capacity planning for virtualized data centers.

We consider the problem of assigning enterprise applications with time-varying demands for multiple hardware resources such as CPU, memory, or I/O bandwidth efficiently to servers with finite capacities. In this domain, economic efficiency means the use of computing resources so as to minimize the number of servers required to provide a given set of VMs. Applications installed in VMs must get enough resources in order to conform to pre-defined service level agreements. This new capacity planning problem has been described as the *server consolidation problem* (Speitkamp and Bichler, 2010).

1.1. Server consolidation

Data centers usually measure resource utilization for each VM and for each physical server. For example CPU utilization and allocated memory of a server are typically logged in 5-minute intervals. Such log data can then be used for characterizing and predicting future resource demands and determining efficient VM allocations. Industry practice usually follows simple approaches to consolidate servers based on peak workloads observed in the past. For example, administrators collect maximum VM demands for various resources over several weeks and derive a

[*] Corresponding author. Tel.: +49 721 9654 866; fax: +49 721 9654 867.
E-mail addresses: setzer@fzi.de (T. Setzer), bichler@in.tum.de (M. Bichler).

0377-2217/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2012.12.005

respective VM demand vector. VMs are then assigned to a physical server at a time until the demand exceeds the capacity.

Many enterprise applications exhibit predictable workloads with daily or weekly seasonalities (Parent, 2005; Rolia et al., 2005). Enterprise applications with predictable workloads are actually the prime candidates for server consolidation projects as these allow for reliable resource planning and high potential server savings. For example, email servers typically face high loads in the morning and after the lunch break when most employees download their emails, payroll accounting is often performed at the end of the week, while workload of a data warehouse server has a daily peak very early in the morning when managers typically access their reports.

The problem of determining efficient VM allocations can be modeled using a multi-dimensional bin-packing (MBP) problem formulation. We will briefly introduce the formulation as a succinct description of the server consolidation problem, as it will provide a basis for the contribution in this paper. Suppose that we are given J VMs $j = 1, \ldots, J$ to be hosted by I servers $i = 1, \ldots, I$ or fewer. Different types of resources $k = 1, \ldots, K$ such as CPU, I/O, or main memory may be considered and each server has a certain capacity s_{ik} of each resource k. y_i is a binary decision variable indicating if server i is used, c_i describes the cost of a server (for example, energy costs per hour), and the binary decision variable x_{ij} indicates which VM j is allocated to which server i. The planning period is divided into time intervals indexed by $t = 1, \ldots, T$. These intervals might be minutes or hours per day if we have daily periodicity. Let r_{jkt} be the capacity that VM j requires of resource k in time period t. The server consolidation problem can now be formulated as the MBP in (1).

$$\begin{aligned} \min \quad & \sum_i c_i \cdot y_i \\ \text{s.t.} \quad & \sum_{i \leqslant I} x_{ij} = 1 \qquad \forall j \leqslant J \\ & \sum_{j \leqslant J} r_{jkt} \cdot x_{ij} \leqslant s_{ik} \cdot y_i \qquad \forall i \leqslant I, \quad \forall k \leqslant K, \quad \forall t \leqslant T \\ & y_i, x_{ij} \in \{0,1\} \qquad \forall i \leqslant I, \quad \forall j \leqslant J \end{aligned} \tag{1}$$

The objective function minimizes total server costs. The first set of constraints ensures that each VM is allocated to one of the servers, and the second set of constraints ensures that the aggregated resource demand of multiple VMs does not exceed a server's capacity per host server, time interval, and resource type. As we will show in Section 4, r_{jkt} can typically be estimated with sufficient quality from workload data.

Speitkamp and Bichler (2010) have shown that the consideration of daily workload cycles using an MBP formulation can lead to 31% savings in physical servers. These results are only based on smaller instances of up to 50 Enterprise Resource Planning (ERP) applications or around 200 Web applications, but considering only CPU and aggregating the workload data to 15-minute time intervals. Although the consideration of memory and I/O is essential in most projects, the resulting MIPs would not be tractable except for small problem sizes with 10 or 20 VMs.

When daily workload cycles are assumed and resource requirements for all 5-minute time intervals per day are considered, resource demand behavior is described by 288 distinct time intervals ($T = 288$). So the number of resource constraints in (1) grows with $O(TKI)$ and there are $O(IJ)$ binary variables.

Linear programming-based heuristics, which allow for the consideration of important technical allocation constraints have also been evaluated, but did not lead to a significant increase in scaleability. Although server consolidation based on mathematical programming has shown to yield considerable savings in server cost, this *curse of dimensionality* is a barrier not only to this but any practical application in data centers, where often hundreds of VMs need to be considered. The problem is important for IT (information technology) service managers, because data centers regularly need to assign or reassign a set of VMs to a new hardware infrastructure.

Nowadays, VM managers such as VMware or Xen allow for *VM live migration* – the dynamic reallocation of VMs during runtime without a significant interruption of service (Clark et al., 2005). There has been work on estimating the hypothetical additional energy savings by means of VM live migration. The intuition here is to power off whole servers temporarily after vacating them, i.e., after migrating all hosted VMs to other physical servers. Based on simulations or numerical experiments, Beloglazov and Buyya (2010), Setzer and Wolke (2012), and Ardagna et al. (2012), among others, calculated potential additional energy-savings of up to 30% when using a central controller or scheduler for VM live migration. To mitigate limitations of centralized approaches regarding scalability and reliability (for example crashs of controller nodes), Quesnel (2011), among others, introduced a distributed approach where VMs exchange states with their neighbors in order to determine efficient VM migration schedules.

Interestingly, so far there is little understanding of the reliability of dynamic workload management and benefits achievable in practice through automated VM reallocation. Such reallocations require a significant amount of CPU and memory resources on the physical servers as well as network bandwidth. Repeated daily demand peaks might lead to many reallocations for short time periods when dynamic reallocations are used. Even if VM migration scheduling can theoretically lead to significant energy savings, this might well be compensated by these overheads, if analyzed in a realistic data center environment.

Therefore, stable allocations of virtual machines to servers, which do not require reassignments during a defined period of time, are a preferred means of capacity management by most IT (information technology) service managers. According to our industry partner and the outcomes of recent surveys, e.g., a study by North Bridge Venture Partner (2011), today most enterprises are awaiting market maturity of dynamic VM management approaches before adopting a dynamic provisioning strategy for their business-critical systems.

1.2. Solving high-dimensional resource allocation problems

The sheer volume of workload data of hundreds of VMs is a challenge when applying mathematical optimization to solve the resource allocation problem. There is a significant literature on solving large-scale integer programs with many variables which consist of sparse, specially structured constraint matrices using techniques such as column generation. An overview of such techniques can be found in Barnhart et al. (1996). In the server consolidation problem, however, the constraint matrix has many constraints rather than many variables. Furthermore, the constraint matrix is not sparse and it does not exhibit a special block diagonal structure as is required for Danzig–Wolfe composition algorithms and branch-and-price approaches.

A literature review of Wascher et al. (2007) reveals that research on cutting and packing problems still is rather traditionally oriented, stressing areas which include clearly-defined standard problems with one, two, or (sometimes) three dimensions. Heuristics and approximations with good computational results have been developed for specific packing problems with one up to three dimensions (Lodi et al., 1999, Faroe et al., 2003, Grainic et al., 2009). Polynomial time approximation schemes (PTAS) have also been developed for MBP, but the gaps between upper and lower bounds on the solution quality increase with the number of dimensions, which limits their applicability to problems with a low