Tel-Aviv University
Raymond and Beverly Sackler
Faculty of Exact Sciences
School of Computer Science

Scheduling and Load Balancing

This thesis was submitted in partial fulfillment of the requirements
for the doctor of philosophy degree (Ph.D.) at
Tel-Aviv University

Submitted by
Oded Regev

This work was supervised by Prof. Yossi Azar.

January 2001

Acknowledgments

I am indebted to Yossi Azar, my thesis advisor, without whom this thesis would not exist. His great insights guided me through the nights I spent solving the problems presented in this thesis. Returning home late, I always found my parents, Gadi and Zehavit, and my brother, Amit. I thank them and all my family for supporting me and helping me through the times when things seemed a little too difficult.

Credits

The results in this thesis were obtained jointly with my advisor, Yossi Azar. I am grateful to my co-authors for allowing me to include in this thesis results obtained in joint work with them. These include the results in Chapter 2 which are a joint work with with Jiří Sgall and Gerhard J. Woeginger and appear in [16]. The results in Chapter 3 were obtained jointly with Leah Epstein and appear as part of [4]. The results in Chapter 4 appear in [7] and were obtained jointly with Baruch Awerbuch and Stefano Leonardi. The results in Chapter 5 are joint work with Baruch Awerbuch and appear in [9]. Finally, the results in Chapter 6 can be found in [15].

Abstract

In this thesis, we consider scheduling and load balancing problems. In scheduling problems we are given a set of jobs to be assigned to free time slots on one of several processors. In these problems the time axis is the only axis that exists. In load balancing problems on the other hand, each job arrives with its active time period and we have to choose a machine to which we assign it. So in these problems, in addition to the time axis, there exists a load axis which we wish to balance. The goal in both problems can be either to minimize a cost function or to maximize a benefit function. Providing optimal solutions for these problems is usually intractable, hence we are considering approximated solutions for both off-line and on-line aspects of the problems.

A short summary of problems and results in this thesis follows. The first two problems are load balancing problems, the next two are scheduling problems and the last one is a routing problem.

Temporary Task Assignment on Identical Machines

We consider the off-line temporary task assignment problem. Jobs, in addition to their weight, have an arrival and a departure time. The goal is to assign the jobs such that the maximum load over both machines and time is minimized. We show that no polynomial time algorithm can achieve an approximation ratio below 1.5 for this problem. However, for the case where the number of machines is fixed, we present a $PTAS$.

Online Load Balancing with Unrelated Machines

Here we consider on-line load balancing of temporary tasks on unrelated machines. We prove an inapproximability result for the problem and show that a trivial algorithm almost achieves the best competitive ratio possible. In the special case of the related-restricted machines model we show tight results on the competitive ratio for a whole range of speeds. Our results apply to randomized algorithms as well.

Multiprocessor Scheduling without Migration

For the off-line problem of scheduling jobs in a multiprocessor setting in order to minimize the flow time, the $SRPT$ algorithm is known to perform within a logarithmic factor of the optimal schedule. This algorithm both preempts jobs and migrates jobs between machines. Unlike preemption, migration is not known to be necessary for achieving these

low approximation ratios. We show how one can achieve the same approximation ratio without migrating jobs. This result also applies to the on-line setting where the algorithm achieves the best competitive ratio possible.

Benefit Maximization for Online Scheduling

In this on-line scheduling problem jobs arrive over time. Our goal is to maximize the total benefit gained from the scheduling of the jobs. A common model is to give each job its own deadline and to take into account only jobs completed by their deadline. The relatively high competitive ratios encountered in this model motivate the search for more reasonable measures of benefit. We consider a model where the benefit gained from a job is a function of its processing time: the longer a job is delayed the lower the benefit gained. A constant competitive algorithm is shown for this model.

The Unsplittable Flow Problem

The Unsplittable Flow Problem (UFP) is a routing problem where we are given a capacitated graph and a set of connection requests with individual demands and profits. The objective is to route a subset of the requests in order to maximize the total profit of the routed requests. The routing must obey edge capacities and use single flow paths. We present algorithms for several variants of the problem. We identify the three main cases of the problem and either improve or match the previously known approximation ratios for all three. However, unlike previous algorithms, all of our algorithms are both strongly polynomial and combinatorial. While the results above apply to the off-line setting, we also present several results for the on-line setting.

Contents