Approximation Algorithms for Resource Allocation Optimization

by

Kewen Liao

A thesis submitted for the degree of
Doctor of Philosophy

School of Computer Science,
The University of Adelaide.

March, 2014

To my parents and grandparents,
for their endless support.

Contents

List of Figures

List of Tables

List of Acronyms

Acronym	*Meaning*
UFL	Uncapacitated Facility Location (problem)
KM	K-median (problem)
ILP	Integer linear program
$FTFL$	Fault-Tolerant Facility Location (problem)
$SCFL$	Soft Capacitated Facility Location (problem)
k-UFL	K-Uncapacitated Facility Location (problem)
$FTRA_\infty$	Unconstrained Fault-Tolerant Resource Allocation (problem)
$CFTRA_\infty$	Capacitated Unconstrained Fault-Tolerant Resource Allocation (problem)
RRA	Reliable Resource Allocation (problem)
$FTRA$	Constrained Fault-Tolerant Resource Allocation (problem)
k-$FTRA$	K-Constrained Fault-Tolerant Resource Allocation (problem)
QoS	Quality of service
FPT	Fixed parameter tractable
CDN	Content distribution/delivery network
PM	Physical machine
VM	Virtual machine
OR	Operations research
TCS	Theoretical computer science
CC	Computational complexity (theory)
VCO	Vertex cover optimization (problem)
VCD	Vertex cover decision (problem)
TM	Turing Machine
FSC	Finite state control
RAM	Random Access Machine
CU	Control unit
PC	Program counter
SAT	Boolean satisfiability (problem)

Acronym	*Meaning*
IS	Independent set (problem)
UVC	Unweighted vertex cover (problem)
LP	Linear program
CSC	Complementary slackness condition
$\mathcal{APX}$	Approximable
$\mathcal{PTAS}$	Polynomial time approximation scheme
$\mathcal{FPTAS}$	Full polynomial time approximation scheme
TSP	Traveling salesman problem
AP-reduction	Approximation preserving reduction
PCP	Probabilistically checkable proof (theorem)
VC	Weighted vertex cover (problem)
SC	Set cover (problem)
JV	An approximation algorithm by Jain and Vazirani [75] for UFL
MP	An approximation algorithm by Mettu and Plaxton [113] for UFL
MMSV	An approximation algorithm by Mahdian *et al.* [104] for UFL
JMS	An approximation algorithm by Jain *et al.* [74] for UFL
CRR	Clustered randomized rounding (algorithm) in [41]
CSGA	Cost scaling and greedy augmentation (procedures)
MYZ	An approximation algorithm by Mahdian *et al.* [106] for UFL
FLO	Facility Location with Outliers (problems)
$MLFL$	Multi-level Facility Location (problems)
OFL	Online Facility Location (problems)
DR	Dependent rounding (technique)
LC	Laminar clustering (technique)
CFL	Capacitated Facility Location (problem)
$HCFL$	Hard Capacitated Facility Location (problem)
IG	Integrity gap (of an integer linear program)
$CFLS$	Capacitated Facility Location (problem) with splittable demands
$CFLU$	Capacitated Facility Location (problem) with unsplittable demands
$UniFL$	Universal Facility Location (problem)

Acronym	*Meaning*
$HCFLU$	Hard Capacitated Facility Location (problem) with unsplittable demands
GAP	Generalized assignment problem
LR	Lagrangian relaxation (technique)
LMP	Lagrangian multiplier preserving (property)
$FTFA$	Fault-Tolerant Facility Allocation (problem)
$FTFP$	Fault-Tolerant Facility Placement (problem)
SG-1	A star-greedy algorithm for $FTRA_\infty$
PD-1	A primal-dual algorithm for $FTRA_\infty$
SG-2	An improved star-greedy algorithm for $FTRA_\infty$
PD-2	An improved primal-dual algorithm for $FTRA_\infty$
MRR	Minimum reliability requirement
VLSI	Very-large-scale integration
PD-3	A primal-dual algorithm for RRA
APD-3	An accelerated primal-dual algorithm for RRA
IDF	Inverse dual fitting (technique)
ULPR	A unified LP-rounding algorithm for $FTRA$
PD-4	A primal-dual algorithm for $FTRA$
APD-4	An accelerated primal-dual algorithm for $FTRA$
SOC	Sum of contributions
AGA	Acceleration of greedy augmentation (procedure)
PK	Procedures for solving k-$FTRA$
BS	Binary search (procedure)
GP	Greedy pairing (procedure)
RR	Randomized rounding (procedure)
e.g.	For example
i.e.	That is
etc.	And so on
w.r.t.	With respect to
w.l.o.g.	Without loss of generality
s.t.	Such that

Abstract

Nowadays, data storage, server replicas/mirrors, virtual machines, and various kinds of services can all be regarded as different types of resources. These resources play an important role in today's computer world because of the continuing advances in information technology. It is usual that similar resources are grouped together at the same site, and can then be allocated to geographically distributed clients. This is the resource allocation paradigm considered in this thesis. Optimizing solutions to a variety of problems arising from this paradigm remains a key challenge, since these problems are $\mathcal{NP}$-hard.

For all the resource allocation problems studied in this thesis, we are given a set of sites containing facilities as resources, a set of clients to access these facilities, an opening cost for each facility, and a connection cost for each allocation of a facility to a client. The general goal is to decide the number of facilities to open at each site and allocate the open facilities to clients so that the total cost incurred is minimized. This class of the problems extends the classical $\mathcal{NP}$-hard facility location problems with additional abilities to capture various practical resource allocation scenarios.

To cope with the $\mathcal{NP}$-hardness of the resource allocation problems, the thesis focuses on the design and analysis of approximation algorithms. The main techniques we adopt are linear programming based, such as primal-dual schema, linear program rounding, and reductions via linear programs. Our developed solutions have great potential for optimizing the performances of many contemporary distributed systems such as cloud computing, content delivery networks, Web caching, and Web services provisioning.

Declaration

I, Kewen Liao, certify that this work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis (as listed on Page x) resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signature

Date

Preface

During my PhD study at the University of Adelaide from 2009 to 2013, I have produced four conference papers and two journal articles related to this thesis (see my homepage at `http://cs.adelaide.edu.au/~kewen` for my bio with a complete list of publications). The thesis topic is theoretical in nature and based on the content presented in the following papers.

Conference Publications:

- Kewen Liao and Hong Shen. Unconstrained and constrained fault-tolerant resource allocation. In *Proceedings of the 17th annual international conference on computing and combinatorics (COCOON)*, pages 555–566, Dallas, Texas, USA, 14-16 August 2011. Springer-Verlag, Berlin

- Kewen Liao and Hong Shen. Fast fault-tolerant resource allocation. In *Proceedings of 12th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*, pages 231–236, Gwangju, Korea, October 20-22 2011. IEEE

- Kewen Liao and Hong Shen. Approximating the reliable resource allocation problem using inverse dual fitting. In *Proceedings of the Eighteenth Computing: The Australasian Theory Symposium (CATS)*, page Vol. 128, Melbourne, Australia, January-February 2012. ACS, Sydney

- Kewen Liao, Hong Shen, and Longkun Guo. Improved approximation algorithms for constrained fault-tolerant resource allocation. In *Fundamentals of Computation Theory (FCT) - 19th International Symposium*, pages 236–247, Liverpool, UK, August 19-21 2013. Springer-Verlag, Berlin (extended version invited to the special issue of Theoretical Computer Science)

Journal Publications:

- Kewen Liao and Hong Shen. Lp-based approximation algorithms for reliable resource allocation. *The Computer Journal*, 57(1):154–164, 2014

- Kewen Liao, Hong Shen, and Longkun Guo. Constrained fault-tolerant resource allocation. *Theoretical Computer Science*, submitted in December 2013, at `http://arxiv.org/abs/1208.3835`, currently under review

Acknowledgments

This PhD study was a real challenge for me, and overcoming this challenge would not have been possible without the support of many people.

First and foremost, I would like to thank my principal supervisor Prof. Hong Shen, for his great supervision over the past few years. I am especially grateful for his trust in me working on tough theoretical problems. Without his constant advice and guidance, this thesis would not be even completed.

I am truly indebted and thankful to my supervisor A/Prof. Michael Sheng. He was also the supervisor of my honors thesis. He is the one who introduced me to research, inspired and encouraged me to pursue a PhD during my honors year. Without him, this thesis might not even exist.

I am sincerely grateful to my supervisor Emeritus Prof. Zbigniew Michalewicz, for his kind help, support, and encouragement in every aspect of my research.

I owe a huge debt of gratitude to my parents and grandparents, for their unconditional love, support, and encouragement. My most heartfelt thanks go to my parents for their spiritual and financial support during my eight years of study in Australia.

I would like to thank some of my colleagues and friends: Changjian, Donglai, Jeff, Lei, Li, Lina, Longkun, Mike, Scott, Shihong, Sim, Xiang, Xiaoqiang, Yidong, Yihong, Yong, Yongrui, Denny, Ke, Leo, Nick, Su, White, Xiaoming, Yibing, Yuguo, and Zilang, for their accompany, inspiration, and influence during my PhD journey.

Finally, I thank the University of Adelaide for providing me scholarship and the School of Computer Science for financially supporting my conference travels. I would also like to express my gratitude to the anonymous examiners of this thesis.