Intellectual Property Rights (IPRs) notice

Part of this thesis may be protected by one or more of Indian Copyright Registrations (Pending) and/or Indian Patent/Design (Pending) by Dean, Industrial Research & Development (IRD) unit, Indian Institute of Technology Delhi (IITD) New Delhi-110016, India. IITD restricts the use, in any form, of the information, in part or full, contained in this thesis ONLY on written permission of the competent Authority: Dean, IRD, IIT Delhi OR MD, FITT, IIT Delhi. ©Indian Institute Of Technology Delhi (IITD), New Delhi, 2014.

AFFINITY DRIVEN DISTRIBUTED SCHEDULING FRAMEWORK FOR PARALLEL COMPUTATIONS

by

ANKUR NARANG

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Submitted

in fulfillment of the requirements of the degree of

Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY, DELHI

INDIA

April 2014

To my family,

inspiring teachers in my life and Goddess Gayatri

Certificate

This is to certify that the thesis titled **"AFFINITY DRIVEN DISTRIBUTED SCHEDULING FRAMEWORK FOR PARALLEL COMPUTATIONS"** being submitted by Ankur Narang in the Department of Computer Science and Engineering, Indian Institute of Technology, Delhi, for the award of the degree of Doctor of Philosophy, is a record of bona-fide research work carried out by him.

The results contained in this thesis have not been submitted to any other university or institute for the award of any degree or diploma.

Dr. Sandeep Sen (Caretaker Supervisor)
Professor
Department of
Computer Science & Engineering
Indian Institute of Technology
New Delhi - 110 016

Acknowledgments

I am truly indebted to Dr. Sandeep Sen for the great amount of encouragement, technical guidance and support I have received from him. It is my great pleasure to whole-heartedly thank Prof. R.K. Shyamasundar for his invaluable support and technical guidance all throughout my thesis. Lot of work in this thesis has been done under his direct guidance, support and evaluation. I thank him for introducing some of the key problems in this thesis and providing interesting ideas even for the most difficult problems. Also, would like to express my gratitude to Prof. Naveen Garg and Prof. Amit Kumar who provided me great inspiration to look at the most challenging problems and to learn from each failure while dealing with these challenges. A humble salutation to Prof. Maheshwari for being a strong pillar of support, encouragement and guidance.

I am highly obliged to IBM Research Labs, Delhi for the strong support for my research work during the entire course of PhD and providing a motivating research environment and strong research exposure. I would like to especially thank Prof. Vijay Garg for guidance in research values and research writing skills, Dr. Manish Gupta for showing the light of leadership in research and being a strong inspiration to achieve high quality results and Dr. Ravi Kothari and Dr. Raghuram Krishnapuram for constantly guiding me. I would like to especially thank Abhinav Srivastava for fruitful technical discussions and collaboration.

I feel very grateful to my institute, I.I.T. Delhi for giving me a wonderful learning experience for nearly seven years. I express my gratitude to all the faculty members of the computer science department for their great teachings and helping nature. I feel obliged to Prof. Vivek Borkar and Dr. Onkar Dabeer at T.I.F.R. Mumbai for fruitful and inspiring discussions.

I am grateful to my parents for giving me their immense love, support, the great gift of a life and the encouragement to excel. I express heartiest gratitude to my wife, Shilpi, for precious inspiration and unconditional love that I have received from her. I thank my children , Siddhansh and Anish, who kept me motivated to perform and deliver my best even under strenuous moments. Many thanks to both of my brothers, Girish and Ulind, for being a constant source of encouragement, motivation and guidance and being always there for me.

Above all, I would like to thank God who always helps me achieve difficult targets in life. I hope using the learnings in this effort and all other efforts in life, I can do true service to humanity in terms of enlightment with rich knowledge as well as contributing to health and happiness for all of mankind and nature at large.

New Delhi | Ankur Narang

Abstract

Massively parallel computing is fast becoming a mainstream research area. In order to realize performance on large scale parallel architectures and massive (multiple petabytes/exabytes) amounts of data, it is necessary to have efficient online load balancing and scheduling of large parallel computations with scalable performance on huge number (millions) of cores/processors. The scheduler needs to execute in a pure distributed and online fashion, should maximally follow affinity inherent in the computation and must have low time, message and space complexity. Further, these massive scale parallel computations need to be scheduled under constraints of bounded resources and energy to be economically viable in practical deployments. Bounded resources could lead to physical deadlocks (arising from cyclic dependency of resources) which necessitates the scheduler to avoid deadlocks in a distributed fashion. Simultaneous consideration of these objectives including affinity, time, message and space complexity and constraints such as deadlock-free execution makes affinity driven multi-objective distributed scheduling a very challenging problem.

In this thesis, we present a generic framework for affinity driven multi-objective distributed scheduling. As part of this framework, we consider multiple problems including affinity driven scheduling in unconstrained and constrained space as well as load-balance vs affinity trade-offs in distributed scheduling. We present novel distributed data structures and algorithms (*ADS-US* algorithm for unconstrained space and *ADS-CS* algorithm for constrained space) that address the challenges in multi-objective distributed scheduling. Experimental results using real benchmarks on Intel NUMA architecture, multi-core cluster architecture (Blue Gene/P) and many-core cluster (Intel cluster) demonstrate the efficacy of our approach. Distributed deadlock freedom proof along with space and time complexity analysis have also been provided.

Further, large scale data dependent irregular computations such as sparse matrix benchmarks (e.g. *Conjugate Gradient*) and state-space search benchmarks such as *UTS* (Unbalanced Tree Search) require optimum trade-offs between load balance, scheduling overheads and/or affinity to maximize performance. We present a novel algorithm (referred to as *LDS*) that uses combination of randomized topology-aware inter-place *work pushing* along with inter-place *work-stealing* to achieve such an optimized trade-off. Detailed experimental analysis on benchmarks such as *UTS*, *Conjugate Gradient* and *Heat* demonstrate the superior scalability and performance achieved using our approach as compared to state-of-the-art approaches including:(a) upto $2\times$ improvement over *Cilk-NOW* [BL97], (b) upto 40% improvement over *Charm++* [SZJK11], (c) parallel efficiency and performance improvement over [VDP$^+$11], (d) largest geometric tree with 1 trillion nodes (for UTS benchmark) ever considered by any prior work [SJ08] [VDP$^+$11] [JDP$^+$09]. Further, as compared to custom (hand-written) parallel code we deliver comparable results for these benchmarks demonstrating the low time and message complexity.

Contents