Lecture Notes in Computer Science 12985

More information about this subseries at http://www.springer.com/series/7407

Dalibor Klusáček · Walfredo Cirne ·
Gonzalo P. Rodrigo (Eds.)

Job Scheduling Strategies for Parallel Processing

24th International Workshop, JSSPP 2021
Virtual Event, May 21, 2021
Revised Selected Papers

Springer

Editors
Dalibor Klusáček
CESNET
Prague, Czech Republic

Walfredo Cirne
Google
Mountain View, CA, USA

Gonzalo P. Rodrigo
Apple
Cupertino, CA, USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-030-88223-5 ISBN 978-3-030-88224-2 (eBook)
https://doi.org/10.1007/978-3-030-88224-2

LNCS Sublibrary: SL1 – Theoretical Computer Science and General Issues

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume contains the papers presented at the 24th Workshop on Job Scheduling Strategies for Parallel Processing (JSSPP 2021) that was held on May 21, 2021, in conjunction with the 35th IEEE International Parallel and Distributed Processing Symposium (IPDPS 2021). The proceedings of previous workshops are also available from Springer as LNCS volumes 949, 1162, 1291, 1459, 1659, 1911, 2221, 2537, 2862, 3277, 3834, 4376, 4942, 5798, 6253, 7698, 8429, 8828, 10353, 10773, 11332, and 12326.

This year 17 papers were submitted to the workshop, of which we accepted 10. All submitted papers went through a complete review process, with the full version being read and evaluated by an average of 3.4 reviewers. Additionally, one invited keynote paper was included in the workshop. We would like to especially thank our Program Committee members and additional reviewers for their willingness to participate in this effort and their excellent, detailed, and thoughtful reviews.

For the second time in its history the JSSPP workshop was held fully online due to the worldwide COVID-19 pandemic. Despite the obvious logistic problems, all talks were presented live, allowing the participants to interact with the authors of the papers. We are very thankful to the presenters of accepted papers for their participation in the live workshop session. Recordings from all talks at the 2021 edition can be found at the JSSPP's YouTube channel: https://bit.ly/3mXyT8F.

This year, the workshop was organized into three major parts: a keynote, a session containing two papers discussing open scheduling problems and proposals, and a session containing eight technical papers.

The keynote was delivered by Dror Feitelson from the Hebrew University, Israel. In his keynote, Feitelson presented resampling with feedback, a performance evaluation method for job scheduling. The method builds upon previous methods in the field. First works on evaluation used accounting logs as workload data for simulations. These logs were precise, but would only provide data on specific situations and would not allow simulating scenarios different to the original logs. These challenges were solved by workloads models, but models are usually limited to workload insights that researchers know in advance. Resampling combines characteristics of both, partitioning real workloads in different components (job streams from different users) and generating new workload by sampling from the pool of basic components. These workloads keep most of the original structure while adjusting them to simulate desired scenarios. However, they lack realism as patterns in the workloads do not change depending on the behavior of the scheduler. This is solved in resampling with feedback, a model where users are modeled and their behavior adapts to the resulting scheduling decisions, e.g., jobs must start in a particular order, or jobs are not submitted till others complete or start. Resampling with feedback provides the realism of logs while eliminating many of their drawbacks and enables evaluations of throughput effects that are impossible to observe with static workloads.

Papers accepted for this year's JSSPP cover several interesting problems within the resource management and scheduling domains and include two open scheduling problems (OSP). This year's OSPs focus on the artifacts and data formats needed to perform scheduling research. Soysal et al. highlight the lack of availability of source code from past work on runtime prediction. As a consequence, evaluating new methods includes re-implementing past methods and frameworks numerous times. The authors present a framework to describe, evaluate, and store runtime prediction methods within an openly available online collection that will help future research.

In the second OSP, Corbalan et al. discuss the challenge of simulating systems using standard formats that cannot capture characteristics of current systems and workloads. The paper proposes to extend the Standard Workload Format (SWF) with the Modular Workload Format (MWF). MWF allows new semantics to be defined as modules in the header and referred to as units of workload or part of jobs.

The first full technical paper was presented by Minami et al., who proposed over-committing scheduling systems to enable interactive workloads in HPC systems. The paper analyzes the impact of resource and proposes performance prediction when over commitment is not present. These methods reliably predict the performance degradation of collocated applications, becoming a valid source for future collocation of HPC schedulers.

Rao et al. propose a placement scheme to map containers of a micro service within a node to maximize performance by taking into account the architecture of the hardware. Their mechanism reduces the latency and increases throughput of hosted services. At the same time, it coalesces services to increase performance even further.

The third paper presents a learning-based approach to estimate job wait times in high-throughput computing systems. Such systems are usually governed by fair-share schedulers that do not provide estimations on the expected wait times. To correct this, Gombert et al. analyzed the correlation between job characteristics and wait time in real workloads. Based on this study, they evaluated machine learning algorithms to train on the past workloads and produce wait time prediction models based on the more promising job characteristics.

Souza et al. describe a co-scheduler for HPC systems that relies on reinforcement learning to determine the best collocation patterns to increase utilization in some long running HPC workloads. This scheduler applies decision trees to collocate jobs and learns from low and high quality past decisions to improve its collocation logic. This scheduling system increases utilization and reduces wait time together with overall makespan.

In the fifth paper, Jaros et al. propose a set of methods to determine the right resource request for moldable jobs. The methods rely on genetic algorithms that evolve on historical data while aiming to reduce makespan, computation cost, and idling resources. The methods were tested with a set of established workflows, improving their makespan.

The last section of the workshop started with a presentation on methods to to optimize task placement for streaming workloads on many-core CPUs with DVFS. Kessler et al. argue that performance and energy usage can be optimized by taking into account the thermal impact of task placing decisions, the physical structure of a CPU, and its heat propagation patterns. In particular, they show that alternating task executions between disjoint "buddy" cores avoids long term overheating of cores, and thus allows for higher throughput.

In the seventh paper, Zhang et al. show that future clusters will suffer large variations in their available resources due to power constraints or non-predictable supply from cloud providers. The authors modeled this variability and its impact on cluster performance governed by current scheduling systems. They conclude with some ideas on scheduling techniques to reduce the impact of capacity variability.

Last but not least, Hataishi et al. present GLUME, a system that reduces workflow execution times. This system divides the workflow subsections aiming for the shortest combination of runtime and estimated inter wait times, thus providing the shortest makespan. GLUME also re-evaluates its plan when the completion of each job is near. As each job completes, the remaining workflow is shorter and estimations are more precise, reducing the makespan even further.

We hope you can join us at the next JSSPP workshop, this time in Lyon, France, on June 3, 2022. Enjoy your reading!

August 2021

Dalibor Klusáček
Gonzalo P. Rodrigo
Walfredo Cirne

Organization

Program Chairs

Dalibor Klusáček	CESNET, Czech Republic
Gonzalo P. Rodrigo	Apple, USA
Walfredo Cirne	Google, USA

Program Committee

Amaya Booker	Facebook, USA
Stratos Dimopoulos	Apple, USA
Hyeonsang Eom	Seoul National University, South Korea
Dror Feitelson	Hebrew University, Israel
Jiří Filipovič	Masaryk University, Czech Republic
Liana Fong	IBM T. J. Watson Research Center, USA
Bogdan Ghit	Databricks, The Netherlands
Eitan Frachtenberg	Facebook, USA
Alfredo Goldman	University of Sao Paulo, Brazil
Cristian Klein	Umeå Univeristy/Elastisys, Sweden
Bill Nitzberg	Altair, USA
Christine Morin	Inria, France
P-O Östberg	Umeå University, Sweden
Larry Rudolph	Two Sigma, USA
Lavanya Ramakrishnan	Lawrence Berkeley National Lab, USA
Uwe Schwiegelshohn	TU Dortmund, Germany
Leonel Sousa	Universidade de Lisboa, Portugal
Ramin Yahyapour	University of Göttingen, Germany

Additional Reviewers

Ganesh Kamath Nileshwar	TU Dortmund, Germany
Joao C. Martins	Polytechnic Institute of Beja, Portugal
Ricardo Nobre	INESC-ID, Portugal
Abel Souza	Umeå University, Sweden
Diogo Marques	Tecnico Lisboa, Portugal

Contents

Keynote

Resampling with Feedback: A New Paradigm of Using Workload Data for Performance Evaluation: (Extended Version) 3
Dror G. Feitelson

Open Scheduling Problems and Proposals

Collection of Job Scheduling Prediction Methods 35
Mehmet Soysal and Achim Streit

Modular Workload Format: Extending SWF for Modular Systems 43
Julita Corbalan and Marco D'Amico

Technical Papers

Measurement and Modeling of Performance of HPC Applications Towards Overcommitting Scheduling Systems 59
Shohei Minami, Toshio Endo, and Akihiro Nomura

Scheduling Microservice Containers on Large Core Machines Through Placement and Coalescing 80
Vishal Rao, Vishnu Singh, K. S. Goutham, Bharani Ujjaini Kempaiah, Ruben John Mampilli, Subramaniam Kalambur, and Dinkar Sitaram

Learning-Based Approaches to Estimate Job Wait Time in HTC Datacenters 101
Luc Gombert and Frédéric Suter

A HPC Co-scheduler with Reinforcement Learning 126
Abel Souza, Kristiaan Pelckmans, and Johan Tordsson

Performance-Cost Optimization of Moldable Scientific Workflows 149
Marta Jaros and Jiri Jaros

Temperature-Aware Energy-Optimal Scheduling of Moldable Streaming Tasks onto 2D-Mesh-Based Many-Core CPUs with DVFS 168
Christoph Kessler, Jörg Keller, and Sebastian Litzinger

Scheduling Challenges for Variable Capacity Resources 190
Chaojie Zhang and Andrew A. Chien

GLUME: A Strategy for Reducing Workflow Execution Times
on Batch-Scheduled Platforms 210
*Evan Hataishi, Pierre-François Dutot, Rafael Ferreira da Silva,
and Henri Casanova*

Author Index 231