RAW 2023
2nd Workshop on Resource Awareness of Systems and Society

co-located with Euro-par 2023
Limassol, Cyprus, August 28th, 2023

BOOK OF ABSTRACTS

Editors: Maja H. Kirkeby and Gordana Rakić

Editors

Maja H. Kirkeby
Roskilde University, Computer Science
Universitetsvej 1, 4000 Roskilde, Denmark
E-mail: kirkebym@acm.org

Gordana Rakić
University of Novi Sad
Faculty of Sciences, Department of Mathematics and Informatics
SQLab: Software Quality Laboratory
Trg Dositeja Obradovića 3, 21000 Novi Sad, Serbia
E-mail: gordana.rakic@dmi.uns.ac.rs

Publisher

University of Novi Sad
Faculty of Sciences, Department of mathematics and informatics
SQLab: Software Quality Laboratory
Trg Dositeja Obradovića 3, 21000 Novi Sad, Serbia
www.sqlab.org

Copyright

Preface

The 2nd workshop on Resource AWareness of Systems and Society (RAW 2023) is co-located with Euro-par 2023 and aims at bringing together educators, researchers, and engineers from academia and industry to discuss solutions and open problems in the area of resource-aware computing. To this end, RAW 2023 aims at forming a community around resource-aware computing, both for classical resource trade-offs, such as energy vs. performance, but also incorporating novel concepts, such as development time and effort, resilience, or sustainability over product or process life cycles.

The RAW 2023 workshop aims to establish a discussion forum, grounded in the formal submission, review, and publication of workshop papers, discussing early scientific findings and their application in the area of resource-aware computing. We are looking for papers and presentations about resource measurement, monitoring, controlling, and the multi-objective trade-offs that are applicable in the different domains and sectors. We look for experience reports, industrial case studies as well as early and preliminary research results that may later appear as full papers at focused or multidisciplinary conferences or in journals.

We aim at establishing and improving interaction and cooperation between scientists, researchers, educators, innovators, and practitioners from all related sectors and domains that are concerned about the need for more aware resource utilisation in systems and society in general.

The RAW Workshop consists of four technical sessions including the panel discussion at the end. Technical sessions welcome 8 regular presentations and 2 invited talks on state-of-the-art, early ideas, work in progress, preliminary results, case studies, industrial cases, open problems and mature results in the form of research results, reports or demonstrations. Presentation topics are related to different kind of resources (time, energy, space, data, effort, etc.) and their usage in the development and usage of systems in our environment. Presentations are referred to techniques involved in resource measurement, monitoring, controlling, and trade-offs, including the full scope of techniques from background theory to application of innovative solutions in different domains. Finally, presentations tackle different aspect of the mentioned techniques such as education, training, research and practice.

Authors of presented work submitted papers on the presented contributions. Submissions were reviewed by an international program committee, where all submitted papers were reviewed by three reviewers in the first cycle, and by the workshop chairs in the second round checking the completeness of changes and improvements according to the reviewers comments. Finally, 7 full papers are accepted for publishing in LNCS Workshop Proceedings of the Euro-par 2023.

We are grateful to all PC members and sub-reviewers for providing careful and timely opinions on the papers. We are grateful also to CERCIRAS CA19135 and COST Association, as well as to the local organizing team at Euro-par 2023 for technical, logistic and organizational support, including preparation of this proceedings. Finally. we are thankful to all contributors and authors for sharing their ideas and findings with the workshop participants, contributing directly to the success of the first edition of the RAW Workshop.

Maja H. Kirkeby and Gordana Rakić, editors

RAW Workshop Organization

Organized by *CA19135 CERCIRAS: Connecting Education and Research Communities for an Innovative Resource Aware Society*
Funded by *COST Association* and European Union under the Horizon Europe program

Program Chair

Maja H. Kirkeby, Roskilde University, Computer Science, Denmark
Gordana Rakić, University of Novi Sad, Novi Sad, Serbia

Program Committee

Yasmina Abdeddaïm, Université Gustave Eiffel , LIGM, CNRS, France
Zoran Budimac, University of Novi Sad, Serbia
Marcus Denker, INRIA Lille, France
Martin Frieb, Augsburg University, Germany
Padma Iyenghar, University of Osnabrueck, Germany
Elinda Kajo Mece, Faculty of Information Technology, Albania
Ana Oprescu, University of Amsterdam, Nederlands
Zoran Pandilov, Ss. Cyril and Methodius University in Skopje, North Macedonia
Radu Prodan, University of Klagenfurt, Austria
Christine Rochange, IRIT – Université de Toulouse, France
Oleksandr Zaitsev, CIRAD, UMR SENS, MUSE, Université de Montpellier, France

Sub-reviewers

Kurt Horvath, University of Klagenfurt, Austria
Christian Piatka, Augsburg University, Germany
Alexander Stegmeier, Augsburg University, Germany

Unveiling the Potential of the Edge-Cloud Continuum: A Deep Dive into Video Streaming and Graph Processing Use Cases

INVITED TALK

Reza Farahanicenter

reza.farahani@@aau.at

Abstract

Over the past few years, our society has witnessed a growing digitalization trend, with its procedures undergoing a progressive shift into digital manifestations. As an essential technology for digital society, graphs provide a universal abstraction to represent concepts and objects and the relationships between them. On the other hand, the advancement of networking paradigms and steadily increasing numbers of users who prefer to watch video content over the Internet rather than using classical TV have made video the predominant traffic on the Internet. Fog and edge computing expand the capabilities of high-performance cloud data centers by incorporating nodes in close proximity to the data sources located at the network's edge. The edge-cloud continuum enables the creation of a new type of services spanning across distributed infrastructures. However, processing massive data in a representation of Graph or video sequences on the edge-cloud continuum raises numerous challenges concerning resource management, Quality of Experience (QoE), Quality of Service (QoS), performance, and sustainability metrics. This talk first presents the key concepts of two widely recognized applications, namely, video streaming and graph processing. It also outlines multiple critical challenges for processing these applications within the edge-cloud continuum. It finally introduces two innovative frameworks applied within each use case to address these concerns.

Reza Farahani is a Postdoctoral researcher in the Horizon Europe GraphMassivizer project at the Institute of Information Technology (ITEC), University of Klagenfurt. He received his Ph.D. and M.Sc. degrees in information technology and computer engineering in 2023 and 2019, respectively, from the University of Klagenfurt, Klagenfurt, Austria, and the University of Tehran, Tehran, IRAN. Between October 2019 and February 2023, he was involved in the ATHENA project, funded by the Christian Doppler Forschungsgesellschaft and the industry partner Bitmovin GmbH. From November 2022 to January 2023, he was a visiting scholar at the 5G6G Innovation Centre (5GIC6GIC), Institute for Communication Systems (ICS), at the University of Surrey, UK. He has co-authored more than 20 publications in international conferences and journals. Additionally, he has held the role of technical program chair for several international conferences. He also has been working in the computer networks field in different roles, e.g., network engineer, developer, protocol designer, and instructor (RS, SP) for over six years. His research interests are Network and Service Management, Parallel and Distributed Systems, Multimedia Communication, Edge-Cloud Continuum, Serverless Computing, Network Softwarization/Virtualization, Mathematics Optimization, and Distributed Learning approaches. Further information at https://www.rezafarahani.me/

Meta-heuristic search for the optimization of complex problems and its applications in the context of resource-aware development of cyber-physical systems

INVITED TALK

Aitor Arrieta Marcoscenter

aarrieta@mondragon.edu

Abstract

Meta-heuristic search algorithms, which are often inspired on biological evolution, can be used to optimize extremely complex processes. For instance, in the context of resource-aware systems, it is possible to significantly reduce the amount of resources required to accomplish a goal while keeping an adequate quality of service level. In this talk, I will first introduce what meta-heuristic search algorithms are and how complex problems can be recasted to that of an optimization search problem. Afterwards, based on lessons learned gained during the last few years, I will show a set of examples of the applications of these algorithms to the context of resource-aware development and test of cyber-physical systems. These will include examples and experiences of the applications of these techniques to industrial case studies. Lastly, the talk will discuss future work and challenges ahead in the application of these algorithms in the context of resource-aware systems.

Aitor Arrieta is an Associate Professor at Mondragon University. He got his PhD in software engineering at Mondragon University in 2017. His research interests lie in the field of software testing (e.g., search-based software testing, test oracle problem, regression testing, debugging and repair), especially on testing untestable systems, such as Cyber-Physical Systems, highly configurable systems and Machine-Learning based systems. He has been the Principal Investigator of regional, national and international projects, among which the Adeptness (H2020) stands out.

Towards a Simulation as a Service Platform for the Cloud-to-Things Continuum

Wilson Valdez[0000-0002-2851-7576], Hamza Baniata[0000-0003-1978-3175], Andras Markus[0000-0003-3829-414X] and Attila Kertesz[0000-0002-9457-2928]

University of Szeged, Dugonics ter 13, 6720 Szeged, Hungary
{wilson,baniatah,markusa,keratt}@inf.u-szeged.hu

Abstract

In the past years, we have seen an unprecedented pace of technological development in smart applications. Smart Systems incorporate securely connected sensors, actuators, and data processing resources to provide digital services. They provide a wide range of smart applications using emerging technologies that address governmental or industrial processes or citizen life in smart cities, and many of them have been affected by the COVID-19 pandemic which involved a general lack of trust. Integrating Blockchain-based data management into smart systems can enhance the performance, trust, and privacy of their applications, which are getting more and more crucial. In this paper, we propose a vision for a unified Simulation as a Service platform, which will be able to model and investigate Blockchain-based smart systems exploiting IoT, Fog, and Cloud Computing infrastructures.

Keywords: Simulation, Blockchain, Internet of Things, Fog Computing, Cloud Computing.

Cormas: The Software for Participatory Modelling and its Application for Managing Natural Resources in Senegal

Oleksandr Zaitsev[1], François Vendel[1,2] and Etienne Delay[1,3]

[1]CIRAD, UMR SENS, MUSE, Université de Montpellier, France
[2]Institut Sénégalais de Recherches Agricoles (ISRA), CNRF, Dakar, Senegal
[3]Ecole Supérieur Polytechnique (ESP), UMMISCO, Dakar, Senegal

Abstract

Cormas is an agent-based simulation platform developed in the late 90s by the Green research at CIRAD unit to support the management of natural resources and understand the interactions between natural and social dynamics. This platform is well-suited for a participatory simulation approach that empowers local stakeholders by including them in all modelling and knowledge-sharing steps. In this short paper, we present the Cormas platform and discuss its unique features and their importance for the participatory simulation approach. We then present the early results of our ongoing study on managing pastoral resources in the Sahel region, identify the problems faced by local stakeholders, and discuss the potential use of Cormas at the next stage of our study to collectively model and understand the effective ways of managing the shared agro-sylvo-pastoral resources.

Keywords: Resource management, Agent-based modelling, Participatory simulation, Software

GPPRMon: GPU Runtime Memory Performance and Power Monitoring Tool

Burak Topçu[0000-0002-2462-0509] and Işıl Öz[0000-0002-8310-1143]

Izmir Institute of Technology, Turkey
{buraktopcu,isiloz}@iyte.edu.tr

Abstract

Graphics Processing Units (GPUs) serve highly-efficient parallel execution for general-purpose computations at high-performance computing, machine learning, and embedded systems. While performance concerns drive the main optimization efforts, power issues become important for energy-efficient GPU executions. While performance profilers and architectural simulators offer statistics about the target execution, they either present only performance metrics in a coarse kernel function level or lack visualization support that enables performance bottleneck analysis or performance-power consumption comparison. Evaluating both performance and power consumption dynamically at runtime and across GPU memory components enables a comprehensive tradeoff analysis for GPU architects and software developers. This paper presents a novel memory performance and power monitoring tool for GPU programs, GPPRMon, which performs a systematic metric collection and offers useful visualization views to guide power and performance optimizations. Our simulation-based framework dynamically collects memory-centric microarchitectural metrics by monitoring individual instructions and reports achieved performance and power consumption information at runtime. Our visualization interface presents spatial and temporal views of the execution. While the first demonstrates the performance and power metrics across GPU memory components, including NoCs, global memory, and cache, the latter shows the corresponding information at the instruction granularity in a timeline. Our case study reveals the potential usages of our tool in bottleneck identification and power consumption for a memory-intensive graph workload.

Keywords: GPU Computing, Performance monitoring, Power consumption

subMFL: Compatible subModel Generation for Federated Learning in Device Heterogeneous Environment

Zeyneddin Oz[1,[0000-0002-4216-9854]], Ceylan Soygul Oz[3,[0000-0001-6361-821X]], Abdollah Malekjafarian[2,[0000-0003-1358-1943]], Nima Afraz[1,[0000-0002-8422-9878]] and Fatemeh Golpayegani[1,[0000-0002-3712-6550]]

[1]University College Dublin, School of Computer Science, Belfield, Dublin, Ireland
[2]University College Dublin, School of Civil Engineering, Belfield, Dublin, Ireland
[3]National College of Ireland, School of Computing, Dublin, Ireland

Abstract

Federated Learning (FL) is commonly used in systems with distributed and heterogeneous devices with access to varying amounts of data and diverse computing and storage capacities. FL training process enables such devices to update the weights of a shared model locally using

their local data and then a trusted central server combines all of those models to generate a global model. In this way, a global model is generated while the data remains local to devices to preserve privacy. However, training large models such as Deep Neural Networks (DNNs) on resource-constrained devices can take a prohibitively long time and consume a large amount of energy. In the current process, the low-capacity devices are excluded from the training process, although they might have access to unseen data. To overcome this challenge, we propose a model compression approach that enables heterogeneous devices with varying computing capacities to participate in the FL process. In our approach, the server shares a dense model with all devices to train it: Afterwards, the trained model is gradually compressed to obtain submodels with varying levels of sparsity to be used as suitable initial global models for resource-constrained devices that were not capable to train the first dense model. This results in an increased participation rate of resource-constrained devices while the transferred weights from the previous round of training are preserved. Our validation experiments show that despite reaching about 50 percent global sparsity, generated submodels maintain their accuracy while can be shared to increase participation by around 50 percent.

Keywords: Resource-constrained heterogeneous edge devices, Federated learning, Model pruning, Mobile edge devices.

Analysis of approaches to the hardware implementation of a convolutional neural network

Mikhail Tatur and Dmitry Pertsau

Belarusian State University of Informatics and Radioelectronics,
Platonova str., 39, Minsk, 220034, Belarus

Abstract

The problem of hardware support for computing of convolutional neural networks on comprehensive approach is studied in this work. This methodology considers the total number of processing elements, complexity of a processing element, circuit deep for data processing and others as main features for created technical solutions. The LeNet-5 architecture was chosen as an object for research, as one of the most frequently used architectures in comparative studies. For this model, several typical methods of parallelization are considered. The analytical express assessments of some key technical characteristics for each system solution are made.

Keywords: CNN, LeNet-5, hardware implementation

Towards Resource-Efficient DNN Deployment for Traffic Object Recognition: From Edge to Fog

Dragan Stojanovic[0000-0003-4893-3814], Stefan Sentic and
Natalija Stojanovic[0000-0002-3594-1826]

University of Nis, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, Nis, Serbia
{firstname.lastname}@elfak.ni.ac.rs

Abstract

The paper focuses on the challenges associated with deploying deep neural networks (DNNs) for the recognition of traffic objects using the camera of Android smartphones. The main objective of this research is to achieve resource-awareness, enabling efficient utilization of computational resources while maintaining high recognition accuracy. To achieve this, a methodology is proposed that leverages the Edge-to-Fog paradigm to distribute the inference workload across multiple tiers of the distributed system architecture. The evaluation was conducted using a dataset comprising real-world traffic scenarios and diverse traffic objects. The main findings of this research highlight the feasibility of deploying DNNs for traffic object recognition on resource-constrained Android smartphones. The proposed Edge-to-Fog methodology demonstrated improvements in terms of both recognition accuracy and resource utilization, and viability of both edge-only and edge-fog based approaches. Moreover, the experimental results showcased the adaptability of the system to dynamic traffic scenarios, thus ensuring real-time recognition performance even in challenging environments.

Keywords: Traffic object recognition, Edge, Fog, DNN, TensorFlow.

Performance and energy aware training of a deep neural network in a multi-GPU environment with power capping

Grzegorz Koszczał, Jan Dobrosolski, Mariusz Matuszek[0000-0001-7551-256X]
and Paweł Czarnul[0000-0002-4918-9196]

Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology, Narutowicza 11/12, 80-233 Poland
pczarnul@eti.pg.edu.pl

Abstract

In this paper we demonstrate that it is possible to obtain considerable improvement of performance and energy aware metrics for training of deep neural networks using a modern parallel multi-GPU system, by enforcing selected, non-default power caps on the GPUs. We measure the power and energy consumption of the whole node using a professional, certified hardware power meter. For a high performance workstation with 8 GPUs, we were able to find non-default GPU power cap settings within the range of 160–200W to improve the difference between percentage energy gain and performance loss by over 15.0%, EDP1 by over 17.3%, EDS with k=1.5 by over 2.2%, EDS with k=2.0 by over 7.5% and pure energy by over 25%, compared to the default power cap setting of 260W per GPU. These findings demonstrate the potential of today's CPU+GPU systems for configuration improvement in the context of performance-energy consumption metrics.

Keywords: Deep neural network training, power capping, multi GPU, performance-energy optimization.

The Implementation of Battery Charging Strategy for IoT Nodes

Petar Rajković[1,[0000-0003-4998-2036]], Dejan Aleksić[2,[0000-0002-5833-6194]] and Dragan Janković[1,[0000-0003-1198-0174]]

[1]University of Niš, Faculty of Electronic Engineering; Aleksandra Medvedeva 14, Niš, Serbia
petar.rajkovic@elfak.ni.ac.rs, dragan.jankovic@elfak.ni.ac.rs
[2]University of Niš, Faculty of Sciences and Mathematics, Department of Physics
alexa@pmf.ni.ac.rs

Abstract

The Internet of Things (IoT) nodes dedicated to off-grid usage must fully rely on their battery power for continuous operation. In that sense battery charging process design is one of the focal points for the complete system design. Nowadays, battery charging, for such devices, usually relies on solar power which is not, unfortunately, the source of constant energy. Both environmental and constructive elements could easily make a negative impact on the charging process and reduce the amount of collected energy. Furthermore, if the IoT nodes are in hazardous areas, they are less accessible, and the value of effective battery management is even higher. The requirements for the battery charging process implementation are considered as opposite – on one hand, the requirement is to run charging with the lowest possible frequency and not up to 100other hand, the battery should always have enough energy to maintain the regular operation. In this research, we present the structure of the custom-developed IoT node based on the ECS32 system-on-a-chip dedicated to operating in remote industrial areas, and with an accent of its battery charging routine. The current routine is based on standard thresholds approach and improved by including consumption estimates for the predefined periods. This paper presents the first results and should pave the ground for further upgrades. In addition, the comparison with state-of-the-art charging approaches is presented, as the guidelines for future work.

Keywords: Internet of things, resource awareness, industry 4.0, hardware-software codesign.