Preface

Cloud Computing concerns large-scale interconnected systems and it has the main purpose to aggregate and to effcient exploit the power of widely distributed resources. Resource Management and Task Scheduling play an essential role, in cases where one is concerned with optimized use of resources. The ubiquitous networks are highly dynamic distributed systems so the changes in overlay are frequent. On the other hand, the Cloud systems are highly dynamic in its structure because the user requests must be respected as an agreement rule. When ubiquitous networks become clients for Cloud systems new algorithm for events and tasks scheduling and new methods for resource management should be designed in order to increase the performance of such systems. The adaptive methods used in context are oriented on: self-stabilizing, self-organizing, and autonomic systems; dynamic, adaptive, and machine learning-based distributed algorithms; fault tolerance, reliability, availability of distributed systems.

This volume contains the papers presented at ARMS-CC-2014: Workshop on Adaptive Resource Management and Scheduling for Cloud Computing held in conjunction with PODC 2014 (ACM Symposium on Principles of Distributed Computing) in Paris, France, on July 15, 2014. The papers of this volume have identified several important aspects of the problem addressed by ARMS-CC: foundational models for resource management in Cloud, scheduling algorithms, and services and applications. We strongly believe that the papers included in this volume will serve as reference for the researchers and scientists in the field of Cloud Computing. The selected papers for this volume comprise a variety of successful approaches including: Distributed Scheduling Algorithms; Load-Balancing and Co-Allocation; Dynamic, Adaptive, and Machine Learning-based Distributed Algorithm; Many-Task Computing in the Cloud; Self-* and Autonomic Cloud Systems; Cloud Resource Virtualization and Composition; Fault Tolerance, Reliability, Availability of Cloud Systems; Cloud Workload Profiling and Deployment Control; Cloud Quality Management and Service Level Agreement (SLA); High-Performance Cloud Computing, Mobile Cloud Computing; and Green Cloud Computing.

There were 29 initial submissions. Each submission was peer-reviewed by Program Committee members or invited external reviewers. Each submission was reviewed by three Program Committee members. Finally, 14 high-quality papers were selected (about 48 % acceptance ratio) for publishing in the LNCS Post-Proceedings and presented during the workshop. This volume consists of 15 papers (14 papers from ARMS-CC and 1 short invited paper) and two invited talks, which are organized as follows.

The two invited talks were given by Thilo Kielmann (VU Amsterdam, Netherlands) and Marc Shapiro (INRIA & LIP6, Université Pierre et Marie Curie, Paris, France).

The invited paper, “In-Memory Runtime File Systems for Many-Task Computing” by Alexandru Uta, Andreea Sandu, Ion Morozan, and Thilo Kielmann, presents a distributed, in-memory runtime file system called MemFS that replaces data locality by

uniformly spreading file stripes across all storage nodes. Due to its striping mechanism, MemFS leverages full network bisection bandwidth, maximizing I/O performance while avoiding storage imbalance problems.

The 14 papers presented in ARMS-CC workshop are organized as follows.

In the first paper, titled "A Multi-Capacity Queuing Mechanism in Multi-Dimensional Resource Scheduling," Mehdi Sheikhalishahi et al. present a queuing mechanism based on a multi-resource scheduling technique by modeling multi-resource scheduling as a multi-capacity bin-packing scheduling algorithm at the queue level to reorder the queue in order to improve the packing and as a result improve scheduling metrics. The proposed solution demonstrates performance improvements in terms of wait-time and slowdown metrics.

The second paper, "A Green Scheduling Policy for Cloud Computing," presented by Jordi Vilaplana et al., introduced a power-aware scheduling policy algorithm called Green Preserving SLA (GPSLA) for Cloud Computing systems with high workload variability. GPSLA aims to guarantee the SLA (Service-Level Agreement) by minimizing the system response time and, at the same time, tries to reduce the energy consumption. The authors present a formal solution, based on linear programming, to assign the system load to the most powerful Virtual Machines, while respecting the SLA and lowering the power consumption as far as possible.

Ansuman Banerjee et al. describes in the third paper, "A Framework for Speculative Scheduling and Device Selection for Task Execution on a Mobile Cloud," the problem of opportunistic task scheduling and workload management in a mobile cloud setting considering computation power variation. The authors gathered mobile usage data for a number of persons and applied supervised clustering to show that a pattern of usage exists and that follows a state-based model. The proposed solution is used as a strategy to choose and offoad work on a mobile device.

The fourth paper, named "An Interaction Balance Based Approach for Autonomic Performance Management in a Cloud Computing Environment," was presented by Rajat Mehrotra et al. In this paper, a performance management approach is introduced that provides dynamic resource allocation for deploying a general class of services over a federated Cloud Computing infrastructure. This performance management approach is based on distributed control, and is developed by using an interaction balance methodology, which has previously been successfully used in developing management solutions for traditional large-scale industrial systems.

Jordi Arjona Aroca et al. present the problem Virtual Machine Assignment (VMA) in the fifth paper, "Power-Efficient Assignment of Virtual Machines to Physical Machines." The optimization criterion is to minimize the power consumed by all the physical machines. The authors present in this paper four VMA problems depending on whether the capacity or the number of physical machines is bounded or not.

Alexandru-Florian Antonescu and Torsten Braun in the sixth paper, named "Simulation of Multi-Tenant Scalable Cloud-Distributed Enterprise Information Systems," present a simulation approach for validating and comparing SLA-aware scaling policies using the CloudSim simulator, using data from an actual Distributed Enterprise Information System (dEIS). This work extends CloudSim with concurrent and multi-tenant task simulation capabilities.

Vlad Serbanescu et al. present in the seventh paper, “Towards Type-Based Optimizations in Distributed Applications using ABS and JAVA 8,” an API to support modeling applications with Actors based on the paradigm of the Abstract Behavioral Specification (ABS) language. The authors validate this solution through a case study where we obtain significant performance improvements as well as illustrating the ease with which simple high- and low-level optimizations can be obtained by examining topologies and communication within an application.

The eighth paper, “A Parallel Genetic Algorithm Framework for Cloud Computing Applications,” presented by Elena Apostol et al. describes the use of subpopulations for the GA MapReduce implementations. Second, the paper proposes new models for two well-known genetic algorithm implementations, namely island and neighborhood model.

Raphael Gomes et al. discuss in the ninth paper the scalability strategies to enact service choreographies using cloud resources. The authors present efforts at the state-of-the-art technology and an analysis of the outcomes in adopting different strategies of resource scaling. The paper is titled “Analysing Scalability Strategies for Service Choreographies on Cloud Environments.”

Shadi Ibrahim et al., on behalf ok KerData team from Inria Rennes, presents in the 10th paper, “Towards Efficient Power Management in MapReduce: Investigation of CPU-Frequencies Scaling on Power Effciency in Hadoop,” the impact of dynamically scaling the frequency of compute nodes on the performance and energy consumption of a Hadoop cluster. Taking into account the nature of a MapReduce application (CPU-intensive, I/O-intensive, or both) and the fact that its subtasks execute different workloads (disk read, computation, network access), there is significant potential for reducing power consumption by scaling down the CPU frequency when peak CPU performance is not needed. To this end, a series of experiments are conducted to explore the implications of Dynamic Voltage Frequency scaling (DVFS) settings on power consumption in Hadoop-clusters: benefiting from the current maturity in DVFS research and the introduction of governors (e.g., performance, power-save, on-demand, conservative, and user-space).

The 11th paper, “Self-management of Live Streaming Application in Distributed Cloud Infrastructure,” presented by Patricia Endo et al., describes an autonomic strategy that manages the dynamic creation of reectors for reducing redundant traffic in live streaming applications. Under this strategy, nodes continually assess the utilization level by live streaming ows. When necessary, the network nodes communicate and self-appoint a new reector node, which switches to multicasting video ows hence alleviating network links.

Cristina Marinescu et al., in the 12th paper, “Towards the Impact of Design Flows on the Resources Used by an Application,” make the assumption that the presence of design aws in the implementation of a software system may lead to a suboptimal resource usage. The investigations on the impact of several design aws on the amount of resources used by an application indicate that the presence of design aws has an inuence on memory consumption and CPU time, and that proper refactoring can have a beneficial inuence on resource usage.

“Policy-Based Cloud Management Through Resource Usage Prediction” is the 13th paper, presented by Catalin Leordeanu et al. The paper proposes a novel solution,

which offers an efficient resource management mechanism for Clouds. The solution is based on monitoring hosts belonging to the Cloud in order to obtain load data. A policy-based system uses the monitoring information to make decisions about deployment of new virtual machines and migration of already running machines from overloaded hosts.

In the last paper, "An Inter-Cloud Architecture for OpenStack Infrastructures," Stelios Sotiriadis et al. explore an inter-cloud model by creating a new cloud platform service to act as a mediator among OpenStack, FI-WARE datacenter resource management, and Amazon Web Service cloud architectures, therefore to orchestrate communication of various cloud environments.

Florin Pop and Maria Potop-Butucaru acknowledge support by PHC Bilateral Research Project: SideSTEP – Scheduling Methods for Dynamic Distributed Systems: a self-* approach, ID: PN-II-CT-RO-FR-2012-1-0084.

We also express our gratitude and thank to all of the members of the Program Committee, to all of the reviewers, for their hard work in finalizing the reviews on time, as well as the authors for submitting their papers to ARMS-CC-2014. We address our personal warm regards to PODC-2014 organizers, especially to Workshop chairs, Sebastien Tixeuil and Dariusz Kowalski for their support and advices offered during the workshop organization. The editors would like to thank Alfred Hofmann, Peter Steasser, and Anna Kramer for the editorial assistance and excellent cooperative collaboration to produce this valuable scientific work. We appreciate the support offered by EasyChair system team to handle the paper submission, review process, and communications with authors and reviewers. We thank them for this important support.

July 2014

Florin Pop
Maria Potop-Butucaru

http://www.springer.com/978-3-319-13463-5

Adaptive Resource Management and Scheduling for Cloud Computing
First International Workshop, ARMS-CC 2014, held in Conjunction with ACM Symposium on Principles of Distributed Computing, PODC 2014, Paris, France, July 15, 2014, Revised Selected Papers
Pop, F.; Potop-Butucaru, M. (Eds.)
2014, XII, 217 p. 68 illus., Softcover
ISBN: 978-3-319-13463-5