



LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR INFORMATIK
LEHR- UND FORSCHUNGSEINHEIT FÜR KOMMUNIKATIONS-
SYSTEME UND SYSTEMPROGRAMMIERUNG



Prof. Dr. Dieter Kranzlmüller · Institut für Informatik
Ludwig-Maximilians-Universität München · Oettingenstr. 67 · 80538 München

Prof. Dr. Dieter Kranzlmüller
Oettingenstr. 67, 80538 München
Tel: 0 89 / 21 80-9146
Fax: 0 89 / 21 80-9147

kranzlmue@ifi.lmu.de
www.nm.ifi.lmu.de

AGH University of Science and Technology
Faculty of Computer Science, Electronics and Telecommunications
Prof. Dr. Krzysztof Wincza, Prodziekan
al. A. Mickiewicza 30
30-059 Krakow
POLAND

Munich, 20 May 2019

Review of PhD thesis of Mr. Kamil Figiela on „Optimization of scientific workflow execution in the cloud”

Dear Prof. Dr. Wincza,

I am honored and delighted to provide this review for the PhD thesis of Mr. Kamil Figiela with the title „Optimization of scientific workflow execution in the cloud”, submitted to the Faculty Council of Computer Science, Electronics and Telecommunications of the AGH University of Science and Technology, Krakow, Poland.

The thesis addresses the problem of optimal execution of scientific applications in cloud environments. The optimization criteria are resource provisioning and task scheduling with the goal of improved execution times and thereby minimal infrastructure costs. The question at hand is crucial for many IT infrastructures utilized by scientists around the world and in particular for clouds, which are prominent not only in science but as a general service today. The thesis is therefore highly relevant and any form of optimization should be able to reduce costs for its users and cloud providers. However, at the same time, these infrastructures are very volatile and optimization is a challenging task. The candidate has therefore chosen a topic, which is both, relevant and challenging for a broad field of applications.

Furthermore, the solution provided by the candidate is interesting in many technical aspects, starting from its proposal of using a Mixed Integer Linear Programming (MILP) model for the scheduling problem. The optimization itself is then performed for a series of deadlines, which allows to investigate different approaches concerning their costs depending on the chosen deadline. This increases the understanding of both, provider and user, such that a choice for an actual solution is improved. The thesis also demonstrates its usefulness for High Performance Computing (HPC) applications, which are usually not executed with the same efficiency on clouds compared to dedicated HPC resources. The concluding observations indicate that clouds are a viable solution for some applications in this domain.

The thesis is structured in 3 parts, 8 chapters and 5 appendices. The introduction in Chapter 1 provides a motivation for the work, namely the usage of scientific workflows in the growing field of computational science, which is relevant for almost all domains of science and research. The basic assumption is that simulation can perform many tasks which are too expensive or even impossible otherwise. The workflows themselves are series of interconnected tasks, which together deliver the answers to scientific questions. As such, workflows can formally be described as directed acyclic graphs, and the execution of a workflow follows the given graph structure. The other basic ingredient is the prominent usage of virtualization through cloud infrastructures, which represent very powerful IT

resources today. Within this field, the candidate formulates the goal of the thesis as providing and executing efficient plans for arbitrary workflows. In addition, three objectives underlying the thesis are identified, before providing an overview of the structure of the thesis.

Chapter 2 provides the basis for the work by explaining the characteristics of cloud infrastructures. I do appreciate the truthfulness of the author by not providing another definition of cloud computing, but rather a list of pointers to existing definitions. The selected definition by NIST serves its purpose for the goals of the thesis and is probably the most widely used description of clouds. The candidate also explains the different service and deployment models and the cloud building blocks, before introducing the business model for the most representative cloud today, the Amazon Web Services (AWS).

In Chapter 3, the candidate introduces the concept of workflows and bags of tasks as two prominent classes of applications and explains the directed acyclic graph as a representative thereof. Next, the necessity to estimate the resource demand for a particular workflow is explained, which requires substantial knowledge about the underlying infrastructure and its utilization. While graphs are used to describe the structure of processing, a workflow environment is needed to execute it on actual resources. Chapter 3 describes the structure and components of such a workflow environment.

Afterwards, Chapter 4 provides an overview of related work and the state-of-the-art in the respective problem domain. Firstly, the candidate introduces a taxonomy to classify different algorithms according to their specific resource requirements, and explains each of the individual characteristics with an overview of existing solutions for the respective aspects. This chapter clearly demonstrates the knowledge of the candidate about the target science domain and related work in the area, and with the given taxonomy serves as a good basis for the remainder of the work.

In Chapter 5, the first objective of the thesis, the MILP-based planning and scheduling approach is explained. For this, the candidate provides an overview of his respective publications and their relation to the thesis, which emphasizes his abilities as a scientist and underlines the value and importance of his work, as well as his connective to other well-respected scientists who serve as co-authors on his publications. The author explains the characteristics of the MILP approach and its utilization for the general optimization problem itself, followed by an introduction of particular cloud instances for actual estimates. By combining both, the candidate is able to formally specify the optimization problem, which is then used to perform a number of experiments to demonstrate the viability and efficiency of the approach with two MILP-based models. This chapter can be considered the core part of the thesis and is certainly interesting in a number of aspects. In particular, the formal description and the real-world experiments are very useful and impressive.

An example case study and respective experiments are shown in Chapter 6, thereby addressing the second objective of the thesis. The hypothesis of the candidate assumes that certain HPC codes can be represented as workflows, and thus mapped to loosely coupled cloud architectures. The critical issue for this is the minimization of the communication to computation ration, which is described with the PCAM (partitioning, communication, agglomeration and mapping) terminology. The candidate assumes that the flexibility offered by elastic resource provisioning can outperform the batch processing nature of HPC systems through a reduction of waiting times. An additional argument by the candidate is the price-to-performance ratio of cloud systems, which stems from the fact that clouds can be built with off-the-shelf components. While this does not hold for the general case, it is true that certain classes of applications can benefit from these characteristics. While I am critical about the number of applications in this class, I can follow the basic assumption and see potential benefits as explained by the candidate. More details are provided that underline this approach, before the actual experiments are presented. Again, the experimental evaluation is of high quality with a number of interesting results, which are extensively analyzed and critically discussed.

The third objective of the thesis is addressed in Chapter 7 with the investigation of cloud functions as pioneered by AWS Lambda, an increasingly popular approach to distributed computing. The chapter itself focuses on the performance evaluation and how heterogeneity aspects are addressed. I found the cost comparison between Infrastructure-as-a-Service (IaaS) and serverless platforms particularly interesting, especially for the experiments carried out in this work.

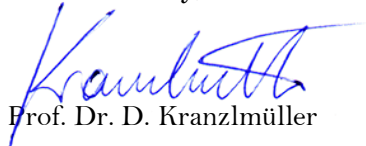
Finally, in Chapter 8 the work is summarized and the contributions are discussed, before a short outlook on future work is given. Here, the auxiliary contributions as provided by the candidate shows his value for the scientific community beyond the scope of the thesis, while the lessons learned are useful indications for applying the new method. In the Future Work section, the candidate identifies a number of novel questions derived from the work and its solution as a follow up to this thesis. The list of publications by the author is quite impressive, and underlines his capabilities as a scientist, which is also shown with the appendices containing relevant papers by the candidate.

All in all, the candidate presents with his thesis an interesting scientific case of high relevance, and demonstrates his abilities in terms of scientific methodology. The actual approach offers a useful solution for a workflow optimization on clouds and particular classes of HPC applications, which may be utilized by other researchers around the world. The thesis itself is well-written and provides a good overview of the candidate's own contribution to the field as well as other related work in the domain.

In conclusion, I am giving the thesis a positive evaluation and support its acceptance as PhD thesis as it provides a substantial contribution to science and research.

If you have any questions, please do not hesitate to contact me. Thank you once again for allowing me to contribute to this important task as a reviewer of the PhD thesis.

Yours sincerely,



Prof. Dr. D. Kranzlmüller