OPPONENT ASSESSMENT OF DOCTORAL THESIS

The Thesis Title:	Fault Tolerance for Big Data Scientific Workflows in Cloud Computing Environments
Workplace:	Department of Informatics and Artificial Intelligence, FAI UTB
Degree Programme:	**P3903** Engineering Informatics
Degree Course:	**3902V037E** Engineering Informatics

Author:	Ammar Nassan Alhaj Ali
Supervisor:	prof. Said Krayem
Reviewer:	prof. Ing. Petr Dostál, CSc.

1. **Topicality of the topic of the doctoral thesis:**

The presented thesis is devoted to an important area of cloud computing. Thesis primarily addresses to optimize the reliability and execution cost of Big Data scientific workflows on cloud computing environments by a model with fault tolerance is offered. The topic of the work is current to solve it.

2. **Fulfilment of the aims set in the doctoral thesis:**

The thesis goals are set and elaborated in great details in the chapter 3 and the student achievement is following:

- He offered a comprehensive survey of the state-of-the-art algorithms of fault tolerance for Big Data scientific workflows in cloud computing environments.
- He offered a critical overview and evaluate former research by comparison in the experiments.
- He identified the important objectives for scheduling scientific workflows on cloud computing according to the latest research.
- He created a model with two fault-tolerant approaches for scheduling scientific workflows on the cloud, the first approach uses a genetic algorithm and the second one uses the greedy algorithm.
- He evaluated the model on different sizes and types of scientific workflows to validate the effectiveness of the proposed methods.
- He performed a deep analysis of the results, summarizing the results, benefits, and drawbacks of the proposed approaches, formalizing the recommendations for future development in the related research.

3. **Approach to the issue dealt with and to the results of the doctoral thesis including a detailed description of the doctoral student's own contribution:**

In this thesis, the student proposes a fault tolerance model with two approaches for optimizing the reliability and execution cost of scientific workflows on heterogeneous virtual machines in clouds. The first approach is Reliability Driven Workflow scheduling with minimum cost using Genetic Algorithm (RDWGA), in this approach, he uses GA to optimize the schedule of workflows and achieve a trade-off between the reliability and the cost.

And the second approach is Dynamic Fault Tolerance with minimum cost using Greedy Algorithm (DFTGA), in this approach, he moves the reliability requirement of the workflow to the sub-reliability requirement of each task and finding replicas that satisfy sub-reliability with minimum execution cost.

4. **Importance for practical implementation or for the development of the relevant scientific discipline:**

In these days the world's movement to Big Data and cloud computing, scientific workflows are increasingly used for Big Data analysis, processing, and management. Ensuring a level of reliability for a scientific workflow execution is a critical task that will tend to increase the cost. The cost of reliability improvements is paid by a reduction in failure, this issue is not quite so simple, where there is never an endless budget for improving the reliability.

5. **Layout and language level of the doctoral thesis:**

The language level of the work is at a very good level. The work is clearly written and understandable. Formally and stylistically, it meets the requirements for a doctoral thesis. The thesis contains the

following chapters: 1. introduction, 2. state of the art, 3. thesis goal, 4. cloud computing environments, 5. scientific workflows, 6. scientific workflows scheduling, 7. fault tolerance for scientific workflows, 8. modelling and simulation of cloud, 9. genetic algorithm, 10. fault tolerance model, 11. reliability driven workflow scheduling using genetic algorithm, 12. dynamic fault tolerance using greedy algorithm, 13. thesis outcomes, 14. conclusions, 15. references.

6. Publication or artistic activities of the doctoral student:

The publishing activity of the doctoral student proves his expertise and is sufficient.

1- Ali, A. A., Jasek, R., Krayem, S., & Zacek, P. (2017, April). Proving The Effectiveness Of Negotiation Protocols KQML In Multi-Agent Systems Using Event-B. *In Computer Science On-line Conference* (pp. 397-406). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-319-57264-2_40

2- Ali, A. A., Jasek, R., Krayem, S., Chramcov, B., & Zacek, P. (2018, April). Improved Adaptive Fault Tolerance Model For Increasing Reliability In Cloud Computing Using Event-B. In *Computer Science On-line Conference* (pp. 246-258). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-319-91192-2_25.

3- Ali, A. A., Vařacha, P., Krayem, S., Žáček, P., & Urbanek, A. (2018). Distributed Data Mining Systems: Techniques, Approaches And Algorithms. In *MATEC Web of Conferences*(Vol. 210, p. 04038). EDP Sciences. https://doi.org/10.1051/matecconf/201821004038.

4- Ali, A. A., Vařacha, P., Krayem, S., Jašek, R., Žáček, P., & Chramcov, B. (2018). Modeling Of Distributed File System In Big Data Storage By Event-B. In *MATEC Web of Conferences* (Vol. 210, p. 04042). EDP Sciences. https://doi.org/10.1051/matecconf/201821004042.

5- Capek, P., Jasek, R., Kral, E., Ali, A. A., & Senkerik, R. (2018, December). Cross Platform Configurable ERP Framework. (2018) *International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 1456-1457). IEEE.

6- Ali, A. A. , Krayem, S., Lazar,I , Kady ,M, Awwama, E. (2018). Solving NP-complete problem using formal method event-B. *9th Comparative European Research,CER 2018* (issue 1.). http://www.sciemcee.org/library/proceedings/cer/cer2018_proceedings01.pdf

7- Ali, A. A., Krayem, S., Chramcov, B., & Kadi, M. F. (2018). Self-Stabilizing Fault Tolerance Distributed Cyber Physical Systems. *Annals of DAAAM & Proceedings*, 29. https://www.daaam.info/Downloads/Pdfs/proceedings/proceedings_2018/169.pdf.

8- Ali, A. A., Chramcov, B., Jasek, R., Katta, R., & Krayem, S. (2021, April). Fault Tolerant Sensor Network Using Formal Method Event-B. In Computer Science On-line Conference (pp. 317-330). Springer, Cham.

9- Katta, R., Ali, A. A., Chramcov, B., Krayem, S., & Jasek, R. (2021, April). Formal Development of Fault Tolerance by Replication of Distributed Database Systems. In Computer Science On-line Conference (pp. 293-306). Springer, Cham.

10- Awwama, E., Ali, A. A., Jasek, R., Chramcov, B., Krayem, S., & Katta, R. (2021, April). Fault Detection Model for Multi Robotic System Using Formal Method Event-B. In Computer Science On-line Conference (pp. 307-316). Springer, Cham.

11- Ali, A. A., Chramcov, B., Jasek, R., Katta, R., & Krayem, S. (2021, April). Classification of Plant Diseases Using Convolutional Neural Networks. In Computer Science On-line Conference (pp. 268-275). Springer, Cham.

Concluding remarks

I recommend the doctoral thesis for defence and after its successful defence to award Ammar Nassan Alhaj Ali a Ph.D. in the relevant field.

In Brno: 20.08.2021

prof. Ing. Petr Dostál, CSc.
reviewer

VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Review of the thesis submitted by Ammar Nassan Alhaj Ali

Title: Fault Tolerance for Big Data Scientific Workflows in Cloud Computing Environments

The thesis is composed of fifteen chapters. The thesis structure is well defined after the introduction, where the primary goal of the thesis is summarized, and the motivation is explained. The later chapters contain theoretical parts of the thesis. The main results are depicted in chapters 10, 11, and 12.

The thesis topic seems very actual because it tries to improve the ability to parallelize scientific computations that are very expensive in computation power and memory. The goals of the thesis are summarized in chapter 3. The goals are satisfied based on the results presented in chapter 11.

Most of the thesis contains a definition of the problem and all the features relevant to it. All the aspects are discussed in considerable detail with crucial additional information. The main work of the student, summarized in chapters 10 to 12, shows that the algorithm based on the NSGA-II can optimize the reliability and greedy algorithm and fault tolerance of the computation run. The main drawback of the thesis is that the algorithm based on the NSGA-II is not compared with any other algorithms. The NSGA-II algorithm follows the Pareto optimal solution based on the weights set on the cost and reliability. The greedy algorithm is then compared. Both algorithms can achieve good results.

The designed algorithms look very efficient, even though they have computational costs during the optimization phase. Their evaluations on more real-time problems and comparison of the algorithms with state-of-the-art algorithms need to prove their more general ability to improve the reliability of the computation chain.

The formal form of the thesis is on a reasonable level. The language is on the lower level – the text contains many typos and grammatical mistakes, incomplete sentences, and typos. Complete proofreading is needed to improve the quality of the text.

The publication activity of the student consists of seven conference papers where only three are related to the thesis topic. The publications are on the lower level of the acceptable publication records for the Ph.D. candidate.

The overall evaluation of the thesis is not easy. Still, the suggested algorithms look very promising. The modification indicated by the student seems very interesting and may be very promising in the later research and evaluation on broader groups of problems. **Therefore, I consider the thesis acceptable for the doctoral degree.**

17. listopadu 2172/15
708 00 Ostrava-Poruba
Czech Republic

attendant: +420 597 321 111
ID data mailbox: d3kj88v

IČO: 61989100
VATIN: CZ61989100

email: university@vsb.cz
www.vsb.cz

Questions:

1. How many iterations your algorithm needs for computing?
2. What is the computation cost of the algorithms?

Ostrava, July 26, 2021

prof. Ing. Jan Platoš, Ph.D.
Department of Computer Science, FEECS
VSB-Technical University of Ostrava

Review of the Ph.D. Thesis

Fault Tolerance for Big Data Scientific Workflows in Cloud Computing Environments

By: Ammar Nassan Alhaj Ali

Overview

The Author deals within this Ph.D. thesis with the process of development, optimization, and performance analysis of the fault tolerant models for big data scientific workflows in HPC/Cloud computing. Currently, many researchers across different science disciplines require the processing of large amounts of data with high reliability for reproducibility of experiments and validity of claims. Therefore, it is necessary to focus on the development of novel models, optimization and algorithms for cost effective scheduling of scientific workflows. Thus, the topic of this thesis can be considered as of high importance and relevant.

The thesis itself is divided into the several essential blocks and one general conclusion section. An introduction into the main ideas, thesis goals and motivation for the research, is followed by the theoretical block comprising chapters 4 – 7. These chapters represent the state-of-the-art review aiming at all aspects related to the cloud computing, scientific workflows, and their scheduling together with fault tolerance issue. Chapters 8 - 10 presents the transition between theory and practical part, as it is describing tools for modelling and simulations of cloud computing, metaheuristic optimizer and fault tolerance model. Overall, many references within these chapters lend weight to the evidence of Author's extensive reading, research and high knowledge about related research tasks, optimization engines, and practical problems.
The next "practical" block of chapters 11 and 12 represents for each case study the brief proof/concept approach, problem design, description of the model architecture and validation/analyses of the results with partial conclusions. The last part concludes all obtained results and knowledge briefly and indicates possible directions for future research and the continuation of the work.
Overall, all parts are logically structured and represent the comprehensive description of the process of development and improvement of optimal models for different studied cases, which are reliability driven workflow scheduling and dynamic fault tolerance.
Regarding the formal remarks, the thesis itself is written in proper English, misprints or wrong cross-refs are not frequent. The organization and quality of typography and graphics is acceptable; however, it is a pity that the theses were not written in LaTex, which would guarantee the uniformity of text blocks, formulas, and graphics. Unfortunately, the author used a misleading name for the dynamic greedy algorithm "DFTGA ", where the last two symbols "GA" may give the impression that this is a reimplementation of the genetic algorithms from the previous chapter. Overall, the thesis can be considered as complete.

Research contribution of the work (Author's own contribution)

The main contribution of this thesis and Author's research contribution to the field of scheduling of scientific (big data) workflows can be summarized as:

- Development of fault tolerance model for big data scientific workflows.
- Customizable and extendable model open for future research (adding more criterions, constraints).
- Robust tests with genetic algorithms (GA) and own greedy algorithm (DFTGA).
- Research on encoding of individuals in GA suitable for given task.
- Single as well as multi criteria optimization experiments.
- Results analysis and comparison with other algorithms have also been carried out.

The developed and described models and algorithms represents the ready to use solution given by presented results, comparisons, and general recommendations. Based on the presented facts, it can be stated that the Author's research contribution is adequate for a Ph.D. candidate.

Questions

- Is it possible to merge and compare the results from chapters 11 (GA) and 12 (DFTGA)? In both chapters, another way of presenting the results is used. (Does not apply to multicriteria optimization, where a different nature of the results - sets of compromise solutions - is expected).
- What is the scalability of the model when using DFTGA algorithm? Are there any upper limits for instance size for a greedy algorithm (i.e. deterministic approach)?
- On page 104 it is stated "We proposed WorkflowSim; it is an extension of CloudSim as a tool for modeling and simulation of cloud computing infrastructures and services." Have any major adjustments been made to the simulation tool? Or it's a printing error proposed -> used.

Final conclusion

In my opinion, the Ph.D. Candidate Ammar Nassan Alhaj Ali has presented the capability of solving the current research tasks and projects from the computer science/computational intelligence research field. The Candidate has performed recognized research work under the guidance of the supervisor Prof. Said Krayem. The research is relevant and proved that the proposed optimization model is efficient. Publications of the Candidate are at a average level, as proved by totally 7 papers in the international conferences. However, there is no publication in the impacted journal. Despite this minor criticism and remarks,

I recommend,

Ammar Nassan Alhaj Ali's Ph.D. thesis to be accepted and defended for the award of the Ph.D degree.

In Zlin: 22.8.2021

Assoc. prof. Roman Senkerik
Tomas Bata University in Zlin
Faculty of Applied Informatics
Department of Informatics and AI
Zlin, Czech Republic