Scientific workflows have emerged as a key technology that assists scientists with the design, management, execution, sharing and reuse of in silico experiments. Workflow management systems simplify the management of scientific workflows by providing graphical interfaces for their development, monitoring and analysis. Nowadays, e-Science combines such workflow management systems with large-scale data and computing resources into complex research infrastructures. For instance, e-Science allows the conveyance of best practice research in collaborations by providing workflow repositories, which facilitate the sharing and reuse of scientific workflows. However, scientists are still faced with different limitations while reusing workflows. One of the most common challenges they meet is the need to select appropriate applications and their individual execution parameters. If scientists do not want to rely on default or experience-based parameters, the best-effort option is to test different workflow set-ups using either trial and error approaches or parameter sweeps. Both methods may be inefficient or time consuming respectively, especially when tuning a large number of parameters. Therefore, scientists require an effective and efficient mechanism that automatically tests different workflow set-ups in an intelligent way and will help them to improve their scientific results.

This thesis addresses the limitation described above by defining and implementing an approach for the optimization of scientific workflows. In the course of this work, scientists’ needs are investigated and requirements are formulated resulting in an appropriate optimization concept. This concept is prototypically implemented by extending a workflow management system with an optimization framework. This implementation and therewith the general approach of workflow optimization is experimentally verified by four use cases in the life science domain. Finally, a new collaboration-based approach is introduced that harnesses optimization provenance to make optimization faster and more robust in the future.

This publication was written at the Jülich Supercomputing Centre (JSC) which is an integral part of the Institute for Advanced Simulation (IAS). The IAS combines the Jülich simulation sciences and the supercomputer facility in one organizational unit. It includes those parts of the scientific institutes at Forschungszentrum Jülich which use simulation on supercomputers as their main research methodology.
Automated Optimization Methods for Scientific Workflows in e-Science Infrastructures

Sonja Holl
# Contents

List of Figures xi

List of Tables xiii

List of Abbreviations xv

List of Publications

## 1 Introduction

1.1 Scientific Workflows in e-Science .......................... 1
1.2 Challenges for Scientists Using Life Science Workflows ........ 3
1.3 Goals of the Thesis ........................................ 5

## 2 Concept Development for State-of-the-Art Workflow Optimization

2.1 General Aspects of Scientific Workflows .................. 9
2.1.1 Scientific Workflows .................................... 9
2.1.2 Scientific Workflow Management Systems ............... 12
2.1.3 e-Science Collaborations ................................ 12
2.2 General Aspects of Optimization and Learning .......... 14
2.2.1 Mathematical Background and Notations ............... 14
2.2.2 Different Optimization Algorithms .................... 15
2.2.3 Design Optimization Frameworks ...................... 17
2.3 State-of-the-Art Scientific Workflow Optimization ........ 18
2.3.1 Runtime Performance Optimization .................. 18
2.3.2 Output Performance Optimization ................... 19
2.3.3 Other Concepts of Workflow Modification ............ 20
2.4 A Concept for Scientific Workflow Optimization .......... 21
3 Enabling Parallel Execution in Scientific Workflow Management Systems 27
3.1 Investigation of Scientific Workflow Management Systems in e-Science 28
3.2 Extension of a Workflow Management System 32
3.2.1 The Taverna Workflow Management System 33
3.2.2 UNICORE Middleware 35
3.2.3 Architecture of the Grid Plugin 35
3.2.4 Development of the Grid Plugin 36
3.2.5 Enhanced Parallel Application Execution 38
3.3 Evaluation by Life Science Use Cases 39
3.4 Discussion 43
3.5 Conclusion 44

4 A Framework for Scientific Workflow Optimization 47
4.1 The Approach of Scientific Workflow Optimization 48
4.1.1 A New Optimization Phase in the Scientific Workflow Life Cycle 48
4.1.2 Investigation of Different Optimization Levels 51
4.1.3 Definition of the Optimization Target 54
4.2 The Usability Compliance of Workflow Optimization 55
4.3 The Taverna Optimization Framework 56
4.4 Enabling Optimization on Distributed Computing Infrastructures 60
4.4.1 Three Tier Execution Architecture 61
4.4.2 Implementation of Parallel Workflow Execution 62
4.4.3 Parallel Optimization Use Case 64
4.5 Discussion 64
4.6 Conclusion 65

5 Optimization Techniques for Scientific Workflow Optimization 67
5.1 Optimization Techniques for Scientific Workflow Parameters 67
5.1.1 Genetic Algorithms 70
5.1.2 A Genetic Algorithm for Scientific Workflows 71
5.2 The Parameter Optimization Plugin 72
5.2.1 Development of the Parameter Optimization Plugin 73
5.2.2 Discussion 76
5.3 Evaluation of the Parameter Optimization Plugin 76
5.3.1 Proteomics Workflows 77
Scientific workflows have emerged as a key technology that assists scientists with the design, management, execution, sharing and reuse of *in silico* experiments. Workflow management systems simplify the management of scientific workflows by providing graphical interfaces for their development, monitoring and analysis. Nowadays, e-Science combines such workflow management systems with large-scale data and computing resources into complex research infrastructures. For instance, e-Science allows the conveyance of best practice research in collaborations by providing workflow repositories, which facilitate the sharing and reuse of scientific workflows. However, scientists are still faced with different limitations while reusing workflows. One of the most common challenges they meet is the need to select appropriate applications and their individual execution parameters. If scientists do not want to rely on default or experience-based parameters, the best-effort option is to test different workflow set-ups using either trial and error approaches or parameter sweeps. Both methods may be inefficient or time consuming respectively, especially when tuning a large number of parameters. Therefore, scientists require an effective and efficient mechanism that automatically tests different workflow set-ups in an intelligent way and will help them to improve their scientific results.

This thesis addresses the limitation described above by defining and implementing an approach for the optimization of scientific workflows. In the course of this work, scientists’ needs are investigated and requirements are formulated resulting in an appropriate optimization concept. This concept is prototypically implemented by extending a workflow management system with an optimization framework. This implementation and therewith the general approach of workflow optimization is experimentally verified by four use cases in the life science domain. Finally, a new collaboration-based approach is introduced that harnesses optimization provenance to make optimization faster and more robust in the future.

This publication was written at the Jülich Supercomputing Centre (JSC) which is an integral part of the Institute for Advanced Simulation (IAS). The IAS combines the Jülich simulation sciences and the supercomputer facility in one organizational unit. It includes those parts of the scientific institutes at Forschungszentrum Jülich which use simulation on supercomputers as their main research methodology.