



An Adaptive Provenance Collection Architecture in Scientific Workflow Systems

Thesis submitted in accordance with the requirements of
the University of Adelaide for the degree of Doctor in Philosophy
by

Mehdi Sarikhani

Supervisors:

Dr. Andrew L. Wendelborn

Dr. Bradley Alexander

Faculty of Engineering, Computer and Mathematical Sciences (ECMS)

School of Computer Science

The university of Adelaide

April 2015

ABSTRACT

This thesis investigates adaptive provenance collection in the context of scientific workflow systems. In particular, we show how to design and implement an adaptive provenance system that operates at multiple levels of granularity.

Scientists in different disciplines use scientific workflows as management and representational frameworks for distributed scientific computations. Scientific workflow systems need a scientific workflow management system (SWfMS) to manage the flow of work among (both local and distributed) participants and resources; and to coordinate user and system participants. Scientific workflow systems are run over heterogeneous environments, which see changes over time in resources, requirement and policies (e.g. the cost of resources, or the policy of provenance collection in). Such changes may influence the way in which workflow mechanisms can best operate within the environments, and motivate our consideration of adaptive mechanisms to deal with such changes.

SWfMSs run a scientist's experiments. They manage sequences of complex transformational processes; in particular, they collect provenance information at various levels of abstraction (or granularity). Provenance in SWfMS is important because it enables scientists to have a clear understanding of results, especially to reproduce and verify them.

Provenance information can be collected at different levels of detail, typically coarse, medium and fine grained, using specific provenance collection mechanisms. We define a Model of Provenance (MoP) for each level to make it explicit what is determined as provenance information in each level, and in addition how it is represented.

We explore and survey provenance collection mechanisms and MoP, in order to provide sufficient understanding of the design and development of suitable provenance mechanisms for workflow systems. We emphasize adaptability and interoperability as important and desirable properties of a provenance system, especially those running over distributed environments.

We propose a novel provenance architecture in scientific workflow architectures, which benefit from the notion of *separation of concerns*, which is an important principle in middleware architecture. The design and development of our adaptive provenance architecture untangles the adaptive-granularity and provenance-collection concerns, so that we can more easily offer adaptive provenance collection mechanisms.

We use reflection (MetaObject Protocol (MOP)) and Aspect-Oriented Programming (AOP) as two ways of realizing the separation of concerns in our adaptive provenance collection mechanisms. Both the MOP and AOP oriented adaptive provenance collection mechanisms are explored in our scientific workflow case study, and implemented on a process network based workflow model. The case study demonstrates adaptive collection and representation of multiple levels of provenance granularity, according to our model of provenance (MoP). This MoP represents various levels of provenance granularity in a format compatible with a generic Open Provenance Model, enabling interoperability.

TABLE OF CONTENTS

ABSTRACT	I
LIST OF FIGURES	VII
LIST OF TABLES	X
LIST OF ABBREVIATIONS	XI
DECLARATION	XII
ACKNOWLEDGEMENTS	XIII
1 INTRODUCTION	1
1.1 SCIENTIFIC WORKFLOW MANAGEMENT SYSTEM.....	3
1.2 CASE STUDY OF PROVENANCE IN A WORKFLOW SYSTEM.....	5
1.3 PROVENANCE COLLECTION MECHANISMS IN WORKFLOW SYSTEMS	8
1.3.1 <i>Adaptive provenance collection in workflow system</i>	9
1.4 THESIS CONTRIBUTION	12
1.5 THESIS OUTLINE	12
1.5.1 <i>An “Aspect-oriented” Thesis Outline - the Concerns of the Thesis</i>	14
2 LITERATURE REVIEW	17
2.1 PROVENANCE	18
2.1.1 <i>Provenance Concepts</i>	18
2.1.2 <i>Provenance Systems in eScience</i>	22
2.1.3 <i>Architectural layers of provenance systems</i>	26
2.2 SCIENTIFIC WORKFLOW MANAGEMENT SYSTEM.....	27
2.2.1 <i>Reference Model for SWfMS</i>	28
2.3 PRINCIPLES OF WORKFLOW: EXECUTION AND CONTROL.....	30
2.3.1 <i>Workflow scheduling</i>	33
2.3.2 <i>Workflow Engine</i>	36
2.3.3 <i>Workflow Controller</i>	36
2.4 TOWARDS ADAPTIVE PROVENANCE IN SCIENTIFIC WORKFLOW.....	38
2.4.1 <i>Reflective Architecture</i>	39
2.4.2 <i>Aspect-Oriented Programming Architecture</i>	56
2.5 SUMMARY.....	60

3 MODELS OF COMPUTATION IN SCIENTIFIC WORKFLOW SYSTEMS	63
3.1 SCIENTIFIC WORKFLOW SYSTEMS WITH UNDERLYING MODEL OF COMPUTATION	64
3.1.1 <i>Case Study: A Simple Dataflow experiment.....</i>	65
3.2 INFLUENCES OF DATAFLOW MODEL ON IMPORTANT WORKFLOW SYSTEMS	68
3.2.1 <i>The MoC in PtolemyII and Kepler.....</i>	68
3.2.2 <i>Provenance in the Kepler workflow system.....</i>	71
3.3 THE PROCESS NETWORK MODEL AS A FOUNDATION OF WORKFLOW	74
3.4 PROCESS NETWORK APPLICATION.....	77
3.4.1 <i>A simple Producer and Consumer process network in PNA.....</i>	83
3.4.2 <i>Implementing a Process Network case study.....</i>	84
3.5 SUMMARY	85
4 MODEL OF PROVENANCE	87
4.1 INTRODUCTION.....	87
4.2 A GENERALIZED NOTION OF MODEL OF PROVENANCE.....	89
4.2.1 <i>A Simple MoP</i>	90
4.3 REVIEWING MODEL OF PROVENANCE	95
4.3.1 <i>Anand's MoP</i>	96
4.3.2 <i>COMAD MoP.....</i>	98
4.3.3 <i>Muniswamy-Reddy's MoP.....</i>	98
4.4 OPEN PROVENANCE MODEL MOP.....	101
4.4.1 <i>The Open Provenance Model for Workflows MoP.....</i>	105
4.4.2 <i>Interoperability of OPM in workflow systems</i>	106
4.5 A MECHANISM FOR MULTIPLE-GRANULARITY PROVENANCE	109
4.5.1 <i>A design for Multiple-granularity MoP.....</i>	111
4.5.2 <i>Discussion of multiple-granularity provenance</i>	126
4.6 SUMMARY	131
5 MECHANISMS FOR PROVENANCE COLLECTION	134
5.1 INTRODUCTION.....	134
5.1.1 <i>Provenance collection phases in Workflow lifecycle.....</i>	136
5.1.2 <i>Workflow Orientation.....</i>	136
5.1.3 <i>Level of abstraction</i>	137
5.1.4 <i>Prospective and Retrospective provenance information</i>	138
5.1.5 <i>Granularity of Provenance Information.....</i>	138
5.1.6 <i>Accessibility of detailed information</i>	139

5.1.7	<i>Model of Provenance (MoP)</i>	140
5.1.8	<i>Architectural layers of provenance systems</i>	140
5.1.9	<i>Coupling strategy</i>	143
5.1.10	<i>Storing, accessing and querying provenance infrastructure</i>	144
5.1.11	<i>Provenance representation techniques</i>	144
5.1.12	<i>Types of instrumentation</i>	146
5.2	REVIEW OF PROVENANCE COLLECTION WORKS IN WORKFLOW SYSTEMS.....	148
5.2.1	<i>Kepler</i>	150
5.2.2	<i>Matrioshka</i>	151
5.2.3	<i>Provenance-Aware Storage System</i>	153
5.2.4	<i>SPADE</i>	154
5.2.5	<i>Karma</i>	155
5.2.6	<i>Pegasus</i>	158
5.2.7	<i>VIEW</i>	160
5.2.8	<i>Trident</i>	161
5.3	COMPARISON AND DISCUSSION	163
5.4	SUMMARY.....	168
6	AN ADAPTIVE PROVENANCE ARCHITECTURE IN SCIENTIFIC WORKFLOW	169
6.1	ADAPTIVE PROVENANCE IN SCIENTIFIC WORKFLOW SYSTEMS	170
6.1.1	<i>Adaptive Provenance Collection Mechanisms</i>	175
6.1.2	<i>Desirable design dimensions for adaptive provenance collection mechanisms</i>	176
6.2	PRINCIPLES OF ADAPTIVE WORKFLOW ARCHITECTURES	181
6.2.1	<i>Provenance component in adaptive workflow architecture</i>	181
6.2.2	<i>Workflow Architecture in terms of provenance</i>	183
6.3	A METAOBJECT PROTOCOL DESIGN FOR ADAPTIVE PROVENANCE IN SCIENTIFIC WORKFLOW	185
6.3.1	<i>A MOP for Provenance-collection meta-behaviour</i>	186
6.3.2	<i>A MOP for Distribution meta-behaviour</i>	188
6.3.3	<i>A MOP for Adaptive-granularity meta-behaviour</i>	193
6.4	AOP DESIGN FOR ADAPTIVE PROVENANCE ARCHITECTURE IN SWF	194
6.4.1	<i>Provenance-collection Aspects</i>	195
6.4.2	<i>Adaptive-granularity Aspect</i>	197
6.5	SUMMARY.....	198
7	CASE-STUDY: ADAPTIVE PROVENANCE COLLECTION IN A WORKFLOW SYSTEM	199
7.1	EXPERIMENTAL CONFIGURATION.....	199

7.2	A MOP ORIENTED ADAPTIVE PROVENANCE COLLECTION MECHANISM	202
7.2.1	<i>Enigma MOP</i>	203
7.2.2	<i>Meta-behaviour implementation in Enigma MOP</i>	208
7.3	AOP ORIENTED ADAPTIVE PROVENANCE COLLECTION MECHANISM	216
7.3.1	<i>Implementation of the Provenance Aspect</i>	217
7.3.2	<i>Implementation of Adaptive Aspect</i>	221
7.4	EVALUATION AND COMPARISON	222
7.4.1	<i>Comments on the implementation of fine-grained provenance</i>	225
8	SUMMARY AND CONCLUSIONS AND FUTURE WORK	228
8.1	SUMMARY	228
8.2	CONTRIBUTIONS	229
8.3	FUTURE WORK.....	231
8.4	CLOSING NOTE	232
	APPENDIX A: FINE-GRAINED PROVENANCE	234
	APPENDIX B: MEDIUM-GRAINED PROVENANCE	237
	APPENDIX C: COARSE-GRAINED PROVENANCE (SHORT VERSION)	240
	APPENDIX D: COARSE-GRAINED PROVENANCE	242
	APPENDIX E: MULTIPLE-GRANULARITY PROVENANCE	244
	APPENDIX F: KEPLER PROVENANCE RECORDER CONFIGURATION.....	249
	APPENDIX G: A SIMPLE CASE STUDY USING ENIGMA.....	251
	REFERENCES.....	259

LIST OF FIGURES

FIGURE 1.1. SCIENCE PARADIGMS [1].	2
FIGURE 1.2. (A) UV-CDAT FRAMEWORK; (B) GUI FOR THE UV-CDAT BASED ON VISTRAILS VISUALIZATION NOTATION [39].	6
FIGURE 1.3. VISTRAILS WORKFLOW AND (RIGHT) PROVENANCE BROWSER [43].	7
FIGURE 1.4. A WORKFLOW SYSTEM RUNNING OVER DISTRIBUTED ENVIRONMENTS.	10
FIGURE 2.1. WORKFLOW SYSTEM ARCHITECTURE.	32
FIGURE 2.2. THE RELATIONSHIP BETWEEN REFLECTION AND REFLECTION.	44
FIGURE 2.3. ONE META-OBJECT FOR EACH BASE-LEVEL OBJECT.	47
FIGURE 2.4. META-OBJECTS FOR A CLASS OF BASE-LEVEL OBJECT THAT REIFIES.	47
FIGURE 2.5. SOFTWARE ARCHITECTURE FOR COMPONENT-BASED MIDDLEWARE PLATFORM.	51
FIGURE 2.6. SOFTWARE ARCHITECTURE FOR REFLECTIVE COMPONENT-BASED MIDDLEWARE PLATFORM.	52
FIGURE 3.1. (A) A SIMPLE DATAFLOW; (B) FIREABLE QUEUE IN INTERPRETER.	67
FIGURE 3.2. A WORKFLOW EXAMPLE IN THE KEPLER WORKFLOW SYSTEM EXECUTED UNDER SUPERVISION OF PN DIRECTOR AND FACILITATED BY KEPLER'S PROVENANCE RECORDER.	72
FIGURE 3.3. KEPLER PROVENANCE SCHEMATIC VIEW.	72
FIGURE 3.4. A SIMPLE PN DATAFLOW.	76
FIGURE 3.5. THE HIERARCHY OF PTOLEMY DATAFLOW MODELS [148].	77
FIGURE 3.6. PNA PRODUCER AND CONSUMER PROCESS NETWORK.	80
FIGURE 3.7. HALF-CHANNEL DESIGN.	81
FIGURE 3.8. THE RUN METHOD OF "PROCESS THREAD".	83
FIGURE 3.9. PNA MAIN CLASS FOR PRODUCER AND CONSUMER PROCESS NETWORK.	84
FIGURE 3.10. PNA MAIN CLASS FOR PROCESS NETWORK CASE STUDY.	85
FIGURE 4.1. A DATAFLOW GRAPH	91
FIGURE 4.2. DATA-PROCESS DEPENDENCY.	93
FIGURE 4.3. DATA DEPENDENCY.	93
FIGURE 4.4. PROCESS DEPENDENCY GRAPH.	94
FIGURE 4.5. OPM DEPENDENCIES [49].	102
FIGURE 4.6. OPM DEPENDENCY GRAPH OF FIGURE 4.1.	102
FIGURE 4.7. A BLACK BOX VIEW ON PROCESS NETWORK GRAPH.	113
FIGURE 4.8. OPM STRUCTURE FOR COARSE-GRAINED MOP.	114
FIGURE 4.9. WASGENERATEDBY DEPENDENCY IN COARSE-GRAINED PROVENANCE.	114
FIGURE 4.10. OPM STRUCTURE FOR MEDIUM-GRAINED MOP.	116
FIGURE 4.11. A WHITE BOX VIEW ON PROCESS NETWORK GRAPH WITH INTRA-PROCESS ACTIVITIES.	118
FIGURE 4.12. OPM STRUCTURE FOR FINE-GRAINED MOP.	122
FIGURE 4.13. ONE STEP INFERENCE IN THE PROVENANCE MODEL [93, 175].	127
FIGURE 4.14. OPM STRUCTURE FOR COARSE-GRAINED MOP.	128
FIGURE 4.15. OPM STRUCTURE FOR DATA-ORIENTED COARSE-GRAINED MOP.	129

FIGURE 4.16. OPM DEPENDENCIES FOR DATA-ORIENTED FINE-GRAINED MOP.....	130
FIGURE 5.1. PROVENANCE PYRAMID FROM [177].....	137
FIGURE 5.2. ARCHITECTURAL LAYERS OF PROVENANCE SYSTEMS.....	140
FIGURE 5.3. MATRIOSHKA PROVENANCE DATA SCHEMA, DERIVED FROM [55].	152
FIGURE 5.4. PASSV2 ARCHITECTURE, DERIVED FROM [72]	153
FIGURE 5.5. INFORMATION MODEL COMPOSED OF REGISTRY AND EXECUTION LAYER FROM [186].	156
FIGURE 5.6. KARMA PUBLISH-SUBSCRIBE ARCHITECTURE FROM [194].	157
FIGURE 5.7. (A) ARCHITECTURE OF VIEW (B) VIEW PROVENANCE MANAGER FROM [7, 97].	160
FIGURE 5.8. PROVENANCE DATA MODEL IN TRIDENT FROM [16].....	162
FIGURE 5.9. BLACKBOARD ARCHITECTURE FROM [202].	163
FIGURE 6.1. WORKFLOW SYSTEM.	171
FIGURE 6.2. MOP ARCHITECTURE FOR PROVENANCE COLLECTION MECHANISM.....	184
FIGURE 6.3. AOP ARCHITECTURE FOR PROVENANCE COLLECTION MECHANISM.....	184
FIGURE 6.4. PROVENANCE META-BEHAVIOUR.....	187
FIGURE 6.5. META-BEHAVIOURS.....	188
FIGURE 6.6. A MODEL REIFYING THE THREE PHASES OF METHOD INVOCATION [111].	190
FIGURE 6.7. THE COMPUTATION META-OBJECT REIFIES A COMPUTATION'S COMPONENTS [111].	191
FIGURE 6.8. META-META-LEVEL FOR CUSTOMIZING ALL COMPUTATION COMPONENTS[111].	192
FIGURE 6.9. ASPECT-ORIENTED PROGRAMING CASE STUDY IN A PROCESS NETWORK.	196
FIGURE 7.1. CASE STUDY ARCHITECTURE: MOP VIEWPOINT.....	203
FIGURE 7.2. ENIGMA UML CLASS DIAGRAM.	205
FIGURE 7.3. INSTANTIATION AND REIFICATION OF AN INPUT PORT IN "PNAMOPFACTORY" CLASS.....	206
FIGURE 7.4. CREATION OF OUTPUT PORT AND REIFICATION OF IT IN META-COMPUTATION.	208
FIGURE 7.5. META-BEHAVIOURS ON ENIGMA MESSAGE DECOMPOSITION.	209
FIGURE 7.6. META-BEHAVIOURS ON ENIGMA MESSAGE DECOMPOSITION WITH META-COMPUTATION NOTATION.....	209
FIGURE 7.7. "PROVENANCEHANDLER" CLASS.	211
FIGURE 7.8. "PROVENANNCEHANDLECLASS" CLASS.	211
FIGURE 7.9. "NEWINSTANCE" METHOD OF "PNAFACTORY" CLASS.....	212
FIGURE 7.10. METHODS FOR CONSTRUCTING OPM DEPENDENCIES IN "MOPMULTIPLEGRAINEDPC" CLASS.	213
FIGURE 7.11. THE CREATEWGB METHOD DECIDES ABOUT THE LEVEL OF PROVENANCE GRANULARITY.....	216
FIGURE 7.12. CASE-STUDY ARCHITECTURE: ASPECT VIEWPOINT.	217
FIGURE 7.13. THE AOP CASE STUDY.	220
FIGURE 7.14. POINTCUTS IN ASPECTJ COLLECTING PROVENANCE INFORMATION FOR OPM DEPENDENCIES.....	220
FIGURE F.1. CONFIGURATION GUI OF PROVENANCE RECORDER.	250
FIGURE G.1. NEWMETALEVELFOR METHOD IN DYNAMICMETAFACTORY CLASS.	252
FIGURE G.2. MAIN CLASS OF THE CASE STUDY.	253
FIGURE G.3. TRACE META-BEHAVIOUR.....	253
FIGURE G.4. CREATION OF META-OBJECT AND META-LEVEL.....	254

FIGURE G.5. MESSAGE DECOMPOSITION IN ENIGMA.	255
FIGURE G.6. TRACE OF METHOD INVOCATION THROUGH ENIGMA MESSAGE DECOMPOSITION.	258

LIST OF TABLES

TABLE 4.1. OPM DEPENDENCIES FOR COARSE-GRAINED MOP.....	114
TABLE 4.2. OPM DEPENDENCIES FOR MEDIUM-GRAINED MOP.	116
TABLE 4.3. OPM DEPENDENCIES FOR FINE-GRAINED MOP.	124
TABLE 4.4. OPM-PROFILE-MAPPING OPM ENTITIES DURING WORKFLOW (PNA).D ATTACHED	125
TABLE 4.5. OPM DEPENDENCIES FOR COARSE-GRAINED MOP.....	128
TABLE 4.6. OPM DEPENDENCIES FOR DATA-ORIENTED COARSE-GRAINED MOP.....	129
TABLE 4.7. OPM DEPENDENCIES FOR DATA-ORIENTED FINE-GRAINED MOP.	131
TABLE 5.1. SUMMARY OF DESIGN DIMENSIONS OF SURVEYED PROVENANCE COLLECTION MECHANISMS.....	164
TABLE 6.1. DESIRABLE DEIGN DIMENSIONS.....	176
TABLE 7.1. DESIGN DIMENSIONS OF OUR MOP AND AOP ORIENTED ADAPTIVE PROVENANCE COLLECTION.....	222

LIST OF ABBREVIATIONS

Advanced Message Queuing Protocol	AMQP	Simple Storage Service	S3
Aspect-Oriented Programming	AOP	Synchronous Dataflow	SDF
Aspect-Oriented Software Development	AOSD	Support for Provenance Auditing in Distributed Environments	SPADE
Collection-oriented modelling and design	COMAD	Simple Queuing Service	SQS
directed acyclic graph	DAG	SQL Server Data Services	SSDS
Dynamic Dataflow	DDF	Scientific Workflow Management System	SWfMS
Description-Driven System	DDS	Scientific Workflow Provenance Data Model	SWPDM
Earth System Science Server	ES3	Transparent Result Caching	TREC
Earth System Science Workbench	ESSW	Ultrascale Visualization Climate Data Analysis Tools	UV-CDAT
Geographic information System	GIS	Visual sciEntific Workflow management system	VIEW
Intra-process Used	I-Used	WasControlledBy	WCB
Intra-process WasDerivedFrom	I-WDF	Windows Communication Foundation	WCF
Intra-process WasGeneratedBy	I-WGB	WasDerivedFrom	WDF
Intra-process WasIndirectlyInvokedBy	I-WIIB	Workflow Management Coalition	WfMC
Lineage File System	LinFS	workflow management system	WfMS
Language Independent Query	LINQ	WasGeneratedBy	WGB
Model of Computation	MoC	WasTriggeredBy	WTB
Modelling Markup Language in XML	MoML	Markup language	XML
Model of Provenance	MoP		
MetaObject Protocol	MOP		
neuGRID for You	N4U		
object-oriented programming	OOP		
Open Provenance Model	OPM		
Open Provenance Model for Workflows	OPMW		
Provenance-aware Storage System	PASS		
Process Networks	PN		
Process Network Application	PNA		
Provenance Interchange Language	PROV		
Quality of Services	QoS		

DECLARATION

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Mehdi Sarikhani

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisors, Andrew and Brad, for their invaluable guidance and assistance with my work. Both have been excellent supervisors, and provided a great deal of insightful advice and feedback that has guided my project.

Many thanks go to my parents, who have done so much over the years to support and encourage me in my studies, and teach me the value of working hard to pursue my goals. To fully describe the ways in which they have helped me get to where I am today would take another three hundred pages. It is to them that I dedicate this thesis. Finally, to those always support me in every second of this journey in various ways -Maman joon, Azar, Mohammad, Ali, and Avin - thanks for making my life enjoyable.