THE UNIVERSITY of ADELAIDE

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

LIST OF FIGURES

LIST OF TABLES

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.