HisT/PLIER: A two-fold Provenance Approach for Grid-enabled Scientific Workflows using WS-VLAM

Michael Gerhards, Sascha Skorupa, Volker Sander
FH Aachen, University of Applied Sciences
Jülich, Germany
{M.Gerhards|Skorupa|V.Sander}@fh-aachen.de

Adam Belloum, Dmitry Vasunin,
Ammar. Benabdelkader
University of Amsterdam, the Netherlands
A.S.Z.Belloum@uva.nl,dvasunin@gmail.com,
ammar@science.uva.nl

Abstract— **Large scale scientific applications are frequently modeled as a workflow that is executed under the control of a workflow management system. One crucial requirement is the validation of the generated results, e.g. the traceability of the experiment execution path. The automated tracking and storage of provenance information during workflow execution could satisfy this requirement.. To collect provenance data using the grid-enabled scientific workflow management system WS-VLAM, experimentations were made with two different implementations of the provenance concepts. The first one, adopts the Open Provenance Model (OPM) using the Provenance Layer Infrastructure for e-Science Resources (PLIER). The second one is the history-tracing XML (HisT). This paper describes how these two provenance models are integrated into WS-VLAM.**

Keywords- Grid computing, data provenance-oriented, process provenance, web service, workflow, workflow management system

I. Introduction

Complex processes are often modeled as workflows, using tools which are based on specific workflow languages. Once a user has modeled a particular workflow he submits it to a workflow management system (WfMS) for execution. The WfMS takes care of the dependencies and the progress of the individual tasks in the workflow. Since workflows are used to automate the processing of complex problems, the actual execution path of a particular workflow instance is typically not known in advance. From the user perspective it is therefore necessary to validate the execution path of each workflow instance. In the domain of WfMS this demand is reflected by the term provenance [1].

A particular challenge arises when workflows are mapped to resources at different organizations, each providing a heterogeneous system with non-uniform interfaces to access these resources. The contribution of different organizations in the ‘partial’ execution of the tasks within a scientific workflow experiment should be determined in a liable way.

The Virtual Laboratory Abstract Machine (WS-VLAM) [2] is a Grid enabled scientific workflow management system developed at the University of Amsterdam (UvA). WS-VLAM supports and coordinates the execution of scientific workflows, which are modeled as Grid-enabled software components. To execute workflows using resources across multiple organizations, WS-VLAM deploys the Globus middleware [3], to establish the connection to the underlying Grid. Two approaches for workflow provenance are implemented within WS-VLAM; both assure the automatic capture of data provenance at run-time.

II. Provenance

The importance of validating and reproducing the outcome of computational processes is fundamental to many application domains. Therefore, the particular demand is to keep track of the execution path of a workflow and to record its provenance. In the scope of the Provenance Aware Service-oriented Architecture (PASOA) project [4], several requirements for a provenance system were identified. Listed requirements are for example the verifiability of actors involved in a process, the reproducibility of the process, the accountability and preservation of provenance over time. It is also a frequent practice to distinguish between data and process provenance. Data provenance is defined to be information that helps determine the derivation history of a data product, starting from its original sources. In contrast, process-oriented provenance collects provenance information in the form of a workflow trace. To cope with these different types of provenance information, we implemented and deployed two complementary provenance approaches within WS-VLAM, namely PLIER and HisT.

A. *Provenance Layer Infrastructure for e-Science Resources (PLIER)*

The open provenance model (OPM) [5] provides a comprehensive set of concepts to capture how things came out to be in a given state. OPM defines three types of nodes (artifacts, processes, and agents), which can be represented in a directed graph with causal dependencies. It is designed to achieve inter-operability between various provenance systems.

The Provenance Layer Infrastructure for e-Science Resources PLIER [6], developed by the information management group within Big Grid [7], provides an implementation of the OPM 1.1 specifications [5]. The PLIER API provides a set of functions to build, store, and share workflow experiments as graphs. It also implements an optimal relational database as back-end storage that captures the concepts of the OPM model, using most recent standards and mechanisms, namely the Java Persistence API (JPA 2.0) and Hibernate [8]. The PLIER API provides specific interfaces,

This work was carried out in the context of HiX4AGWS and BiG Grid projects. HiX4AGWS is supported in part of the Federal Ministry of Education and Research in Germany. Grant No.: 17N3409. Big Grid is financially supported by the Netherlands Organization for Scientific Research (NWO).

using JDO 3.1 [9], to transform, or serialize, the provenance data into specific formats (e.g. RDF, XML, and DOT). In addition, provenance query interfaces are available, which allow the end-user to access the information about the performed experiments, interpret the results, and trace the sources of failure. These interfaces combine the power of the database querying functionalities and the rich graphical representation of experiments generated by the PLIER API.

B. History-Tracing XML (HisT)

The HisT provenance structure was developed within the HiX4AGWS project [10] and provides data/process provenance following an approach that directly maps the workflow graph to a layered structure (Figure 1) of an XML document. To pursue this idea, control flow patterns [11] are mapped to generic data patterns within the XML schema to specify the provenance information for any task of a workflow.

Every workflow task is represented by an XML element, the so-called layer element (lines 1, 3, 8). The transitions between tasks are represented by interleaving these layer elements. The layer element of the successor task (line 1) always includes the layer element of the predecessor tasks (lines 3, 8), which in turn, encapsulates the previous layer stack (line 5). That means that the layer stack represents the logical execution order of dependent tasks. The representation of one task as one XML element is intuitive for the user and can be easily processed by programs and transformed to the original workflow graph.

```
   ...
 1 <successor>
 2   <events> ... DATA ... </events>
 3   <predecessor1>
 4     <events> ... DATA ... </events>
 5     <pre-predecessor1> ... </pre-predecessor1>
 6     <sign-predecessor1> ... </sign-predecessor1>
 7   </predecessor1>
 8   <predecessor2> ... </predecessor2>
 9   <sign-successor> ... </sign-successor>
10 </successor>
   ...
```

Figure 1. Interleaving structure

A unique feature of the interleaving structure is the possibility to interleave embedded XML signatures of layer elements without the usage of complex XPath expressions. In provenance trace in Figure 1, the successor task bases on the contribution of its two predecessor tasks. The embedded signature (line 9) references the whole *successor* element (lines 1-10). That means that both predecessor layer elements (lines 3-7, 8) with their own signatures (line 6) are part of the signature check sum. The interleaving of signatures allows liable countersigning which is interesting for application domains like eGovernment and medical applications where human actors execute tasks where the contribution is based on the contribution of a previous human actor [12]. No actor can dispute his contribution and all contributions can be traced back to the responsible actor. This allows the creation of a liable basis by using the XML-DSig [13] and XAdES [14] standards that follows the European Union Directive 1999/93/EC [15].

III. Conclusion and Future Work

The case study described in this paper using the two-fold provenance approach, demonstrated that both HisT and PLIER provide great common functionalities with regards to the capture and storage of provenance data, during workflow execution. In addition, each solution (HisT or PLIER) provide an additional set of distinctive features, which could be suited to better serve the specific need of certain applications. Furthermore, these features could be combined in a complementary manner, in order to gain the full advantages of both approaches. To combine the advantages of HisT and PLIER, mappings will be defined. This is possible because OPM does not restrict to a file structure.

References

[1] Y. L. Simmhan, B. Plale, and D. Gannon, “A Survey of Data Provenance in e-Science”, *SIGMOD RECORD*, vol. 34, pp. 31-36, 2005.

[2] V. Korkhov, D. Vasyunin, A. Wibisono V. Guevara-Masis, A. Belloum “WS-VLAM: Towards a Scalable Workflow System on the Grid” Workshop on workflows in Support of Large-Scale Science (WORKS 07); In conjunction with HPDC 2007; Monterey Bay, June 2007.

[3] I. Foster. “Globus Toolkit Version 4: Software for Service-Oriented Systems.” IFIP International Conference on Network and Parallel Computing, Springer-Verlag LNCS 3779, pp 2-13, 2006.

[4] L. Moreau, P. Groth, S. Miles, J. Vazquez, J. Ibbotson, S. Jiang, S. Munroe, O. Rana, A. Schreiber, V. Tan, and L. Varga, “The Provenance of Electronic Data”, Communications of the ACM, vol. 51(4), pp. 52-58, April 2008.

[5] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan and J. Van den Bussche, “The Open Provenance Model Core Specification (v1.1*),”* Future Generation Computer Systems, vol.27(6) pp.743-756, June 2011.

[6] Provenance Layer Infrastructure for e-Science Resources - PLIER: http://twiki.ipaw.info/bin/view/OPM/Plier .

[7] The Dutch Grid for e-science: www.biggrid.nl .

[8] G. King, C. Bauer, “Java Persistence with Hibernate (Second ed.)“ Manning Publications, pp. 880, ISBN 1932394885, November 2006.

[9] D. Jordan, C. Russell, “Java Data Objects (1st ed.)” O'Reilly Media. pp. 384. ISBN 0596002769, 2003.

[10] M. Gerhards, A. Belloum, F. Berretz, V. Sander, and S. Skorupa “A History-tracing XML-based Provenance Framework for Workflows”, 5th Workshop on Workflows in Support of Large-Scale Science (WORKS), November 2010.

[11] W. Van Der Aalst, A. Ter Hofstede, B. Kiepuszewiski, and A. Barros, “Workflow Patterns“, Distributed Parallel Databases, Vol. 14, No. 1, pp. 5-51, 2003.

[12] S. Skorupa, F.Berretz, A. Belloum, V. Sander, “Towards an Actor-Driven Workfow Management System for Grids”, *CTS 2010*, pp. 611-617, May 2010.

[13] M. Bartel, J. Boyer, B. Fox, B. LaMacchia, and E. Simon, "XML Signature Syntax and Processing", http://www.w3.org/TR/xmldsig-core/, June 2008.

[14] J. C. Cruellas, G. Karlinger, D. Pinkas, and J. Ross, “XML Advanced Electronic Signatures (XAdES)”, World Wide Web Consortium, Note NOTE-XAdES-20030220, February 2003.

[15] DIRECTIVE 1999/93/EC of the European Parliament and of the council of 13 December 1999 on a community framework for electronic signatures, Official Journal of the European Union, vL 013. 0012-0020.