A Multi Agent based Architecture for Data Provenance in Semantic Grid

Syed Imran Jami, *Member, IAENG, WSEAS, ACM* and Zubair Ahmed Shaikh Dr.

***Abstract*—This work deals with the recording of provenance in semantic grid. Provenance is the process of recording and tracking the usage of data in a system. It is considered as one of the requirements of semantic grid. We use multi-agent based workflow management systems in grid environment as data architecture. Most of the provenance models are designed for specialized domains which make them vulnerable for semantic grids which propose to use the services for all in seamless manner. The proposed model can work on generic data sets and the use of mobile agents makes it robust in simultaneous recording of provenance by different nodes. We have proposed to use tree based structure for storing the provenance. The model uses Java based Aglets for mobile agents, dotNet based Alchemi for desktop grid and XML for common data format.**

***Index Terms*—Data Provenance, Multi Agent Systems, Semantic Grid.**

I. Introduction

Grids are used for online collaboration and access to resources. Many scientific and business communities [5], [6], [8], [10], [12] and [13] use it to perform processing and experiments and generate huge data that is required to be tracked for transformations in many applications.

Data Provenance is a technique for recording the log of transformation process of data to its derived form [7]. It answers the queries about originator of Data, transformation and its path to the derived form [9]. These queries are necessary to be answered to get the trust on data for its use in simulation and experiments.

Provenance in all open distributed systems is necessary to track since data proliferation is too frequent. All network based data driven applications require provenance of data. This may include Workflows, scientific data analysis and business intelligence.

Semantic Grid is considered as an interaction among semantic web, software agents and grids. It wants to achieve easy-to-use and seamless automation of processes to facilitate flexible collaborations and computations on a global scale grid [17].

We are working towards solving this problem by developing software agents based provenance model with mobility because it provides seamless automation and autonomy to the activity. Researchers in [2] reported that mobile agents reduce the network load and latency and enables execution autonomously in heterogeneous environment by encapsulating protocols.

Our provenance model uses an intelligent and distributed workflow system [11] to record and track the data. In this model apart from other agent, a provenance agent is proposed that record the usage activity log of data in XML files. In case of simultaneous usage by different nodes, the provenance agent of different nodes can record it at the same time. They are later send to the central archive by the use of mobility feature of agents.

It is proposed to store provenance in tree based structure for efficient tracking. On demanding provenance of data, the tree is traversed from leaf to root node which will determine the original source of data. The traversal from root to leaf will determine the usage of data that will show the path of data to its derived form.

This prototype model is tested on sample documents of our system to determine the usage of data. It provides the basis for our ongoing work in evaluating the trust on nodes in Semantic Grids.

The next section provides the related work of other authors in this area while section 3 provides brief overview of our earlier work [11] and its enhancement that will show the data processing architecture on which we implement this provenance model. Section 4 of this paper describes the model that we proposed with its implementation details while the last section concludes with the limitations of proposed model and brief overview of our ongoing work.

II. Related Work

In the recent years, many labs and researchers are working towards the recording and tracking of provenance. Most of them however worked for specialized domain and based on relational data model.

Lanter's work in [12] is one of the earliest works in this area. A lineage system was designed for spatial databases. It was built specifically for their GIS system. Some researchers in [7] question its scalability issue that makes it vulnerable for open distributed environment.

PASOA [5] and myGrid [10] are two well known projects

Manuscript received December 30, 2007. This work was supported in part by the Higher Education Commission (HEC), Pakistan through its indigenous PhD program and Center for Research in Ubiquitous Computing (CRUC), National University, Karachi Pakistan.

S. I. Jami is with the Center for Research in Ubiquitous Computing, National University of Computer & Emerging Sciences, Karachi, Pakistan (92-333-3400656; fax: 92-21-4100541; e-mail: imran.jami@nu.edu.pk).

Z. A. Shaikh, Dr is with the Computer Science Department and patron of CRUC, National University of Computer & Emerging Sciences, Karachi, Pakistan, (e-mail: Zubair.shaikh@nu.edu.pk).

on provenance that worked for e-Science program which is funded by UK worth $500M and comprising 100 projects related to distributed data, computation and collaboration. The program characterizes larger collaboration, geographically distributed and involves wider range of disciplines that leads to the requirement of Semantic Grid [17]. The team of PASOA developed provenance recording protocol (PReP) that defines set of provenance actions on the invocation of services. They worked for e-science communities that will help them in storing and sharing provenance information. The model is developed for service oriented architectures (SOA) based systems. It uses message passing mechanisms to transfer information between actors of the system. It maintains the provenance in the form of assertions made by the actors involved in the processes. myGrid [10] project also provide provenance mechanism for their Grid environment in the specialized domain of biological sciences.

Buneman *et al* in [9] described a framework for understanding the concept of provenance of data in the context of SPJU (select, project, join, union) queries and views. Based on their studies they coined two terms for provenance named, Why and Where provenance. The former one answers the query of the form, "Why is a piece of data in the output?" while the later one answers the query, "Where did a piece of data come from?" They provided a recording framework based on queries for RDBMS.

Chimera [8] is another implementation for provenance model that develops the derivation of data objects. It provides the generic solution for collaboratory environments which makes it applicable to work in our scenario. It can record provenance on user constructed workflow in a high level language. They however lack infrastructure support which some other researchers in [5] and [14] also commented.

The Karma Provenance service [6] proposed a generic framework for recording provenance in a distributed heterogeneous environment. The model is developed for workflow engines by recording input and output of workflow processes.

Our analysis shows that most of the provenance systems work only for specialized domain. This argument is supported by [7] who provides a detailed study on related techniques of provenance. Some of the works propose to use annotations to determine provenance. Most of the approaches lack the use of software agents to implement the model. Researchers in [13] however use agents to record provenance in distributed environment. They extended the health care system provenance aware, by introducing agents in it. They use PASOA based architecture for recording provenance in health system. This system works only for the domain of health care system to develop the existing system provenance aware. The Karma framework [6] is modeled for generic environment but it lacks the use of agents which make it vulnerable in evaluating sub processes.

We, in this work, are proposing to record provenance in a grid environment using mobile agents. [2] reported seven different reasons which is also endorsed by other researchers in [1] for using mobile agents. They are related to reduction in network load and latency by encapsulating protocols. It provides asynchronous and autonomous execution and can work dynamically even in heterogeneous environment. Foster *et al* in their famous work in [19] cited several benefits in integrating the two approaches that will lead to the development of semantic grids.

III. Data Processing System

In our earlier work [11] we implemented an academic management system using multi agent based workflow system. The work is extended to develop distributed workflow for Grid environment. The data generated by workflow is tracked using the proposed provenance mechanism.

The system in [11] uses four agents in all, out of which, three are stationary and one is mobile. Although this system is applicable for a limited domain (academics), but its extension for grid provides us good environment to implement mobile agent based grid computing. It however, only considers the local environment without considering global scale access.

This environment is implemented using MASIF complaint Aglets [2] for agent based processing and communication and XML for data representation [4]. Java based rule engine (JSR 094) is used to represent rule engine.

Mobile agents using Aglets is useful for any open distributed applications since processing is migrated toward resources [2]. In [11] each mobile aglet carries XML data (state) and code for processing. The mobile aglets visit several nodes and perform necessary communication with respective aglets at each node for computation. Each node runs its own aglet server and connected with centralized database. The nodes in the system generate and manipulate XML data that is logged and tracked in our proposed model for its usage. Details of this system can be found in [11].

This agent based environment is extended for resource sharing purpose by developing each node as desktop grid node using Alchemi which is a .NET based Grid platform for creating resource sharing environment [18]. It mainly consists of two services. Alchemi Manager Service aggregates the Windows Desktop machines for power and submits the threads on Alchemi Executor service. Alchemi provides a programming environment (API) for constructing Desktop Grid applications. It works on the paradigm of master-slave model. The manager is responsible for coordinating the execution of tasks sent to its executors.

Desktop Grid provides architecture for distributed systems on which agents are running. The master node sends the job as Grid Thread to executor nodes using Alchemi API. Each executor node executes thread independently from each other. After integration of Alchemi with system in [11] each executor node also contains Aglet based agents that can interact with centralized provenance archive as shown in fig 1.

IV. Provenance Architecture

In Grid systems, the participating nodes of a sub-system are autonomous and possibly heterogeneous. It is also possible that such sub-systems can be under the authority of

different administrators. Each node can independently take decisions and generate data without the need of others which leads to the requirement of provenance system that should be decentralized in nature. In our architecture, agents are used for coordination and cooperation among nodes.

The processes at Grid based workflow model generate data whose usage is recorded in our approach. Figure 1 shows our extended model that implement provenance in agent based workflow in Grid environment.

In this model, apart from other agents, we introduce a provenance agent in each grid node. The provenance agent has the following roles to play:

1. Monitor executor service for log of data usage.
2. Send the local provenance information to provenance archive by using its mobility feature

In grid, same data item can be accessed by different nodes or same node can process different data. We propose to develop Alchemi API based interface that is responsible for recording the log of data usage in XML file. On completion of job by executor, it tags the file for its end. Provenance Agent (PA) compiles this XML file for necessary conversions at each node. Each data item in a Grid has been assigned a unique ID. On manipulating this data, the nodes log the usage in XML file. After detecting the end tag PA sends local provenance information to central archive. This will ensure seamless automation of the provenance process in a semantic grid.

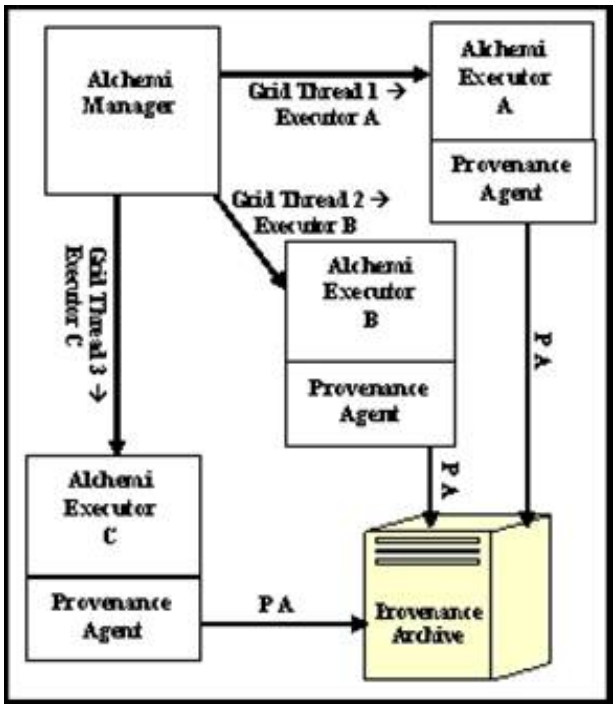

Fig 1: Mobile agent based provenance system

The Semantic Grid proposes to create a universal medium for the exchange of data which leads to the standardization of XML as common data format [4]. XML helps our system to interchange provenance information between dotNet based Alchemi and Java based Aglets. The central archive contains set of XML files received from different nodes. The recording algorithm in centralized system generates derivation tree of data usage by merging local provenance information received from different nodes of same data ID. Each node of a tree contains (source, process) ID that use data. The derivation tree determines parent/child relationship on the basis of node access and time stamp noted in XML file. Root node is the generator of data while the leaf node is the last user of that data. The use of XML and autonomous agents with Alchemi will ensure data interchange in case of heterogeneous environment.

This system is tested on desktop Grid by developing docAPI on C#.Net. This API compares the target document with other documents on executors. The document usage log is maintained in XML file by this API, which is then accessed by Aglets.

The use of mobile agents makes this system autonomous that leads to seamless recording of metadata in semantic grid. The software agent uses message passing paradigm to communicate with other node using agent transfer protocol (ATP). The seamless recording ensures the fulfillment of one of the requirements of Semantic Grid [17].

A. Implementation

As in our data processing architecture, the provenance model is implemented on Alchemi.Net for Grid architecture and MASIF based Aglets [2] for agent based communication while XML for local and global provenance representation.

Alchemi is based on dotNet platform [18] whereas Aglet is Java based platform for agents. Currently, robust dotNet based agent platform is not available that can interact with Alchemi, however different labs are working on it but their results are in initial stage. With this hurdle, instead of intertwining the two technologies we followed the layered approach by placing Aglet layer over Alchemi layer, with XML as an binding platform as shown in fig 2.

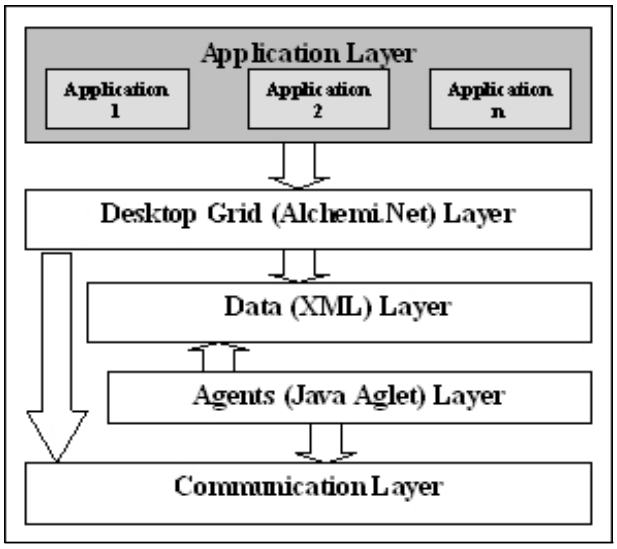

Fig 2: Implementation model of Provenance Architecture in Grid

On manipulation of data by executors, the developed provenance API on Alchemi maintains the log of data usage

in XML format. This XML file is continuously monitored by provenance agent (PA) for change. PA is run by aglet server that creates the clone. The completion of grid thread by respective executors leads to sending the XML file created by our Alchemi API using PA to main central archive.

The provenance information maintained in XML file contains

{data ID, process ID, current node, timestamp}

The example XML file generated by local node is as follows

```
<provenance information>
   <data_ID>111</data_ID>
   <process>2</process>
   <current>B</current>
   <time>1011070830</time>
</provenance information>
```

In this example 111 is the document id generated by process id 2 at time 08:30 am dated 10th Nov, 07. B is the id of one of nodes who is the user of this document.

The algorithm working in archive generates tree by merging these XML files on the basis of data access and timestamp of same dataID. The provenance can be tracked by using simple tree traversal algorithm from leaf to root.

The algorithm is run on around 200 documents in our system distributed across 4 nodes.

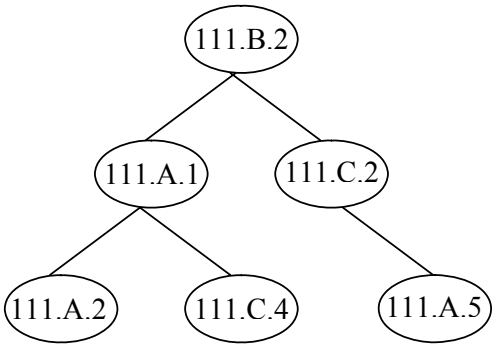

Fig 3: Implementation model of Provenance Architecture in Grid

Fig 3 shows a tree generated by central archive based on the XML file received from different nodes for document id 111 similar to the file shown above. It shows that process ids 2, 4 and 5 of node A, C and A respectively are the last users of document 111. Process 1 and 2 of node A and C are the earlier users while process 2 of node B is the first user of this document. The tree generator is a prototype model that shows its output in a text file and which will later be extended for above graphical display.

V. Conclusion & ongoing work

Semantic Grid is considered as an extension of current grid in which information and services are given well defined meaning. This integration of Grid with semantic web [19] provides simple and seamless automation which enables flexible collaboration and computations on a global scale among projects of distinct domains. David Roure et. al. in [17] state ten different requirements for semantic grid to work. Some of them are:

- Grid Workflow
- Autonomous behavior
- Security and trust
- Provenance and metadata
- Synchronous information streams and fusion
- Communities
- Smart environments

The use of agents in Grid environment provides us autonomous provenance system by introducing provenance agent (PA). This system can be extended for maintaining trust based mechanisms. Multi Agent systems are considered as robust platform that provides mechanisms for maintaining trust and forming dynamic communities [19]. The integration of Grids with Agents help the provenance model to be adapted by semantic grids. It is to be mentioned that the non-availability of dotNet agent framework forced us to use layered approach to integrate the two platforms which makes the processing of XML files slow. A robust dotNet based Agent platform will help us to integrate the two technologies tightly coupled to improve the processing.

The prototype model is tested on a limited number of nodes in our existing small scale grid but it can be scalable to many nodes. More tests are needed to test it on large number of nodes on global scale. Our initial analysis shows that with the use of software agents, the model can accommodate large number of nodes for simultaneous recording of local provenance. The only performance bottleneck is the centralized archive which can be solved by distributing it to among other nodes. [3] provides some algorithms in this regard.

Minor issues related to synchronization area of distributed systems [3] are required to solve, such that the model is dependent on unique ID of data and processes in the whole system with correct timestamp. There are some cases of recording similar timestamps by two nodes in our system. Our model resolves this issue by using the inequality of $\leq$ to determine parent child relationship.

Abstraction is required in dataset which is our next target to address. This system assigns the testing data sets an ID which is applicable under current situation since we are considering small and limited number of nodes. Semantic Grids can accommodate large number of nodes under any domain which will make our model vulnerable. It is therefore required that data items are abstracted at certain level to reduce complexities.

The proposed model can work in heterogeneous environment as well. Our future plans include to run this model on a high power heterogeneous grid that can analyze thousands of terabytes of data related to chemical structure of molecules.

The tracking algorithm in this model uses simple tree traversal algorithm since we limit the scope of this paper to effective recording only. The future work includes designing an efficient algorithm for tree traversal. One of the suggested approaches on which our lab is working is to use genetics

algorithm [15] for optimized traversal that will work on bit patterns. This will also help us to represent trees in a trie [16] data structure that will improve the efficiency of tree traversal algorithm.

The model presented here is the basis for our ongoing research in the area of data provenance for Semantic Grid to maintain trust among nodes. It will also enable systems to determine the resource usage by tracking the provenance of data.

REFERENCES

[1] Gerhard Weiss, "Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence", The MIT Press, 1999
[2] Danny Lange and Mitsuru Oshima, "Programming and Deploying Java Mobile Agents with Aglets, Addison Wesley - 1998.
[3] Andrew S. Tanenbaum and Maarten van Steen, "Distributed Systems: Principles and Paradigms", 2nd Edition, Prentice Hall, 2007
[4] Norman Walsh "A Technical Introduction to XML", http://www.xml.com
[5] P. Groth, S. Jiang, S. Miles, S. Munroe, V. Tan, S. Tsasakou, L. Moreau, "An Architecture for Provenance Systems" Technical report D3.1.1, Ver 0.6 November 29, 2006, www.pasoa.org
[6] Y. L. Simmhan, B. Plale, D. Gannon, "A Fremwork for Collecting Provenance in Data-Centric Scientific Workflows", IEEE International Conference on Web Service (ICWS '06), Chicago, USA, September 18-22, 2006
[7] Y. L. Simmhan, B. Plale, D. Gannon, "A Survey of Data Provenance Techniques", Computer Science Department, Indiana University, Bloomington IN 47405 Technical Report IUB-CS-TR618
[8] I. Foster, J Vöckler, M. Wilde, and Y. Zhao, "Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation," in proceedings of 14th International Conference on Scientific and Statistical Database Management (SSDBM '02), 2002
[9] P. Buneman, S. Khanna, and W. C. Tan, "Why and Where: A Characterization of Data Provenance," in proceedings of 8th International Conference on Database Theory (ICDT '01), London, U.K., 4-6 January 2001
[10] R. D. Stevens, A. J. Robinson, and C. A. Goble, "myGrid: Personalized Bio-Informatics on the Information Grid," Bioinformatics Journal, vol. 19, 2003, pp. 302 -304.
[11] S. I. Jami and Z. A. Shaikh, "A Workflow based Academic Management System using Multi Agent Approach", WSEAS International Conference on Computers, Crete Island, Greece, July 26-28, 2007
[12] D. P. Lanter, "Design Of A Lineage-Based Meta-Data Base For GIS," Cartography and Geographic Information Systems, vol. 18, 1991, pp. 255-261.
[13] T. Kifor, L. Z. Varga, J. Salceda, S. Alvarez, S. Willmott, S. Miles and L. Moreau, "Provenance in Agent-Mediated Healthcare Systems" IEEE Intelligent Systems, Volume 21 , Issue 6, November 2006, pp. 38-46
[14] M. Szomszor and L. Moreau. "Recording and reasoning over data provenance in web and grid services", Lecture Notes in Computer Science (LNCS), vol 2888, Catania, Sicily, Italy, November 2003, pp 603-620
[15] Michael D. Gordon, "A learning algorithm applied to document redescription" Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval, Montreal, Quebec, Canada, 1985, pp 179 - 186
[16] Dan E. Willard, "New trie data structures which support very fast search operations", Journal of Computer and System Sciences, Vol 28, Issue 3, 1984, pp 379 - 394
[17] David De Roure, Nicholas R. Jennings and Nigel.R. Shadbolt, "The Semantic Grid: Past, Present, and Future", Proceedings of the IEEE, Volume 93, Issue 3, March 2005.
[18] A. Luther, R. Buyya, R. Ranjan, and S. Venugopal. "Peer-to-Peer Grid Computing and a .NET-based Alchemi Framework", High Performance Computing: Paradigm and Infrastructure, Laurence Yang and Minyi Guo (editors), Wiley Press, New Jersey, USA. 2004.
[19] Ian Foster, Nicholas R. Jennings and Carl Kesselman, "Brain Meets Brawn: Why Grid and Agents Need Each Other", in Proceedings of 3rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2004), New York, USA, 2004.