Semantic Grid Resource Discovery using DHTs in Atlas*

Manolis Koubarakis, Zoi Kaoudi, Iris Miliaraki, Matoula Magiridou and Antonios Papadakis-Pesaresi

Dept. of Informatic and Telecommunications
National and Kapodistrian University of Athens
Panepistimioupolis, Ilissia
Athens 15784 Greece
{manolis,zoi,iris,matoula,apapadak}@di.uoa.gr

1 Introduction

For the Semantic Grid vision [13] to become a reality, *high quality of service* must be offered to users and applications at all levels of the Grid fabric. In this position paper, we concentrate on high quality of service in the provision of *resource discovery* services in Semantic Grids. Resource discovery is an important problem in Grids in general, and Semantic Grids in particular. We discuss how to achieve *high-performance, scalability, resilience to failures, robustness* and *adaptivity* in the provision of resource discovery services in Semantic Grids, and especially in OntoKit, the Semantic Grid toolkit currently under development in project OntoGrid [22].

OntoGrid (`http://www.ontogrid.net`) is a Semantic Grid project funded by the Grid Technologies unit of the European Commission under the strategic objective "Grid-based systems for Complex Problem Solving" of the Information Society Technologies programme of FP6.

Our basic assumption in this paper is that Semantic Grid resources (e.g., machines, services or ontologies) will be annotated by RDF(S) metadata. Metadata pervades the Semantic Grid and is used to describe Grid resources, the environment, provenance and trust information etc. [13]. The Resource Description Framework (RDF) and RDF Schema (RDFS) are frameworks for representing information about Web resources. RDF(S) consists of W3C recommendations that enable the encoding, exchange and reuse of structured metadata, providing the means for publishing both human-readable and machine-processable information and vocabularies for semantically describing things on the Web. Although RDF(S) was originally proposed in the context of the Semantic Web, it is also a very natural framework for representing information about Grid resources. As a result, it is used heavily in various Semantic Grid projects e.g., ^{my}Grid (`http://www.mygrid.org.uk`) or OntoGrid.

* This work is partially funded by FP6/IST project OntoGrid. The work was performed while the authors were with the Dept. of Electronic and Computer Engineering, Technical University of Crete.

We propose to view resource discovery in Semantic Grids as *distributed RDF query answering* on top of a P2P network of Grid resource *providers* and *requesters*.

Our proposal goes beyond well-known Grid information services such as MDS4 of GT4 in two significant ways:

- We offer service providers and service requesters expressive semantics-based data models and query languages (i.e., RDF(S) and RQL instead of XML and XPath).
- We implement resource discovery using techniques from P2P systems and achieve full distribution, high-performance, scalability, resilience to failures, robustness and adaptivity. Notice that MDS4 implementations are centralized or hierarchical and will never achieve the performance and scalability typically associated with P2P networks.

In the context of OntoGrid, our proposal is realized with the implementation of *Atlas*, a P2P system for the distributed storage and querying of RDF(S) metadata describing Semantic Grid resources e.g., ontologies or services.

The rest of the paper is organized as follows. Section 2 briefly discusses related work at the crossroads of Grid and P2P computing research. Section 3 gives a short description of Atlas. Section 4 shows how to use Atlas for service discovery in OntoKit. Finally, Section 5 concludes the paper.

2 Related Work

Our research can be understood to lie at the intersection of P2P and Grid computing. Although these computing paradigms have different origins and have been developed largely independently, there has been a lot of interesting work lately at the crossroads of these paradigms [12, 28, 10].

Previous papers that explore connections among Grids and P2P networks can be distinguished in the following categories:

1. General papers that discuss the similarities and differences of P2P and Grid systems pointing out important areas where more work is needed [12, 28, 10].
2. Papers where ideas from P2P computing are used in Grid systems. Here, we can further differentiate as follows:
   (a) Works where Grid computing problems are given as a primary motivation, but the contributions are essentially in the P2P domain and can also be applied elsewhere. For example, [3, 21, 6] consider attribute-value data models that can be used to describe Grid resources (e.g., by specifying the CPU power, disk space capacity, operating system and location of a computer) and show how to evaluate queries in these models on top of DHTs (e.g., I am looking for an idle PC that runs Linux and has CPU $\geq$ 3GHz).
   (b) Works where P2P techniques are used to improve functionality in existing Grid systems e.g., resource discovery [17, 18, 16] and replica location management in Globus [7] or flocking in Condor [5].

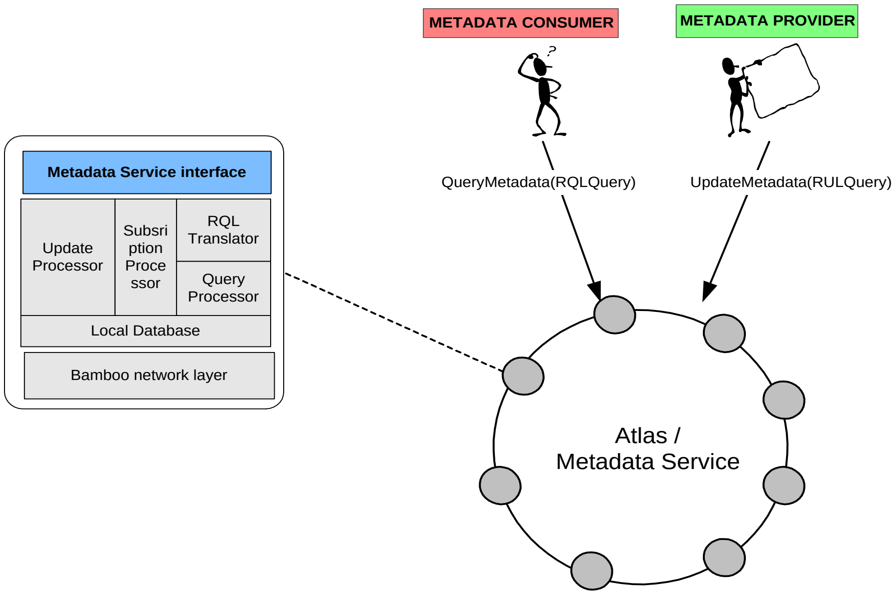

Fig. 1. Atlas and the metadata service

   (c) Service-oriented application development frameworks that enhance existing frameworks for Web or Grid service computing [1, 14] with P2P protocols.

3. Papers where ideas from Grid computing are used in P2P systems. For example, [9] shows how to implement a P2P data integration framework using OGSA-DAI [2].

Our work should be classified in categories 2(b) and 2(c) above. Work with goals similar to ours that uses description logics instead of RDF(S) is reported in [15].

3 A Short Description of Atlas

In Atlas, we use state of the art *distributed hash table (DHT)* technology [4] to implement a distributed system that will be able to scale to hundreds of thousands of nodes and to large amounts of RDF(S) data and queries.

Nodes in an Atlas network are organized under the Bamboo DHT protocol [26]. Nodes can enter RDF(S) data into the network and pose RQL queries. Two kinds of querying functionality are supported by Atlas: *one-time* querying and *publish/subscribe*. Each time a node poses a one-time query, the network nodes cooperate to find RDF(S) data that form the answer to the query. In the publish/subscribe scenario, a node can *subscribe* with a *continuous* query. A continuous query is indexed somewhere in the network and each time matching RDF(S) data is published, nodes cooperate to *notify* the subscriber.

The current implementation of Atlas supports a subset of the query language RQL [20]. The query processing algorithm we use for one-time queries is based on the algorithms proposed in [8] where they were originally used for a smaller class of queries based on triple patterns [8]. Publish/subscribe scenarios in Atlas are handled using the algorithms in [23]. Atlas also supports the recently proposed RDF update language RUL for inserting, deleting and updating RDF metadata [25].

Atlas is used in OntoKit for realizing a fully distributed *metadata service*. A high level view of Atlas and the metadata service of OntoKit is shown in Figure 1.

The implementation of Atlas was started at the Technical University of Crete and is currently continued at the National and Kapodistrian University of Athens. More information on the current version of Atlas is available in [19]. We expect to be able to analyse the performance of Atlas soon on real-world wide area networks using the PlanetLab infrastructure. Currently, we only have very promising simulation results [23] and more experimentation is underway.

4 Atlas in Operation: Service Discovery in OntoKit

In this section, we show how Atlas can be used in OntoKit during service annotation and discovery [22]. The whole scenario is depicted in Figure 2.

OntoGrid is developing annotation technology for Grid services [27]; this technology is deployed as the *annotation service* of OntoKit. For the purposes of this section, it is also important to mention another service of OntoKit, the *ontology service* [11]. The current version of the ontology service provides a Grid interface to an RDFS store where RDFS ontologies are stored (e.g., *service ontologies* or *domain ontologies* etc.).

An ontology for services and various domain ontologies are needed in order to create a service annotation. Let us suppose that the annotation service chooses to search for an ontology about cars in order to annotate a car-repair service (the example comes from a car insurance use case studied in OntoGrid). The annotation service can pose an RQL query to the metadata service and get information about such ontologies e.g., the location and description of a particular ontology – let us call it `car-repair-ontology`. After discovering information about `car-repair-ontology`, the annotation service can retrieve it from the ontology service.

If the annotation service does not know the ontology for annotating services, it has to search for such an ontology as well. An example ontology describing services that could be found in this case is the ^{my}Grid service ontology [24]. We should mention here that this step may be unnecessary if a specific service ontology has been selected for annotating services in OntoKit.

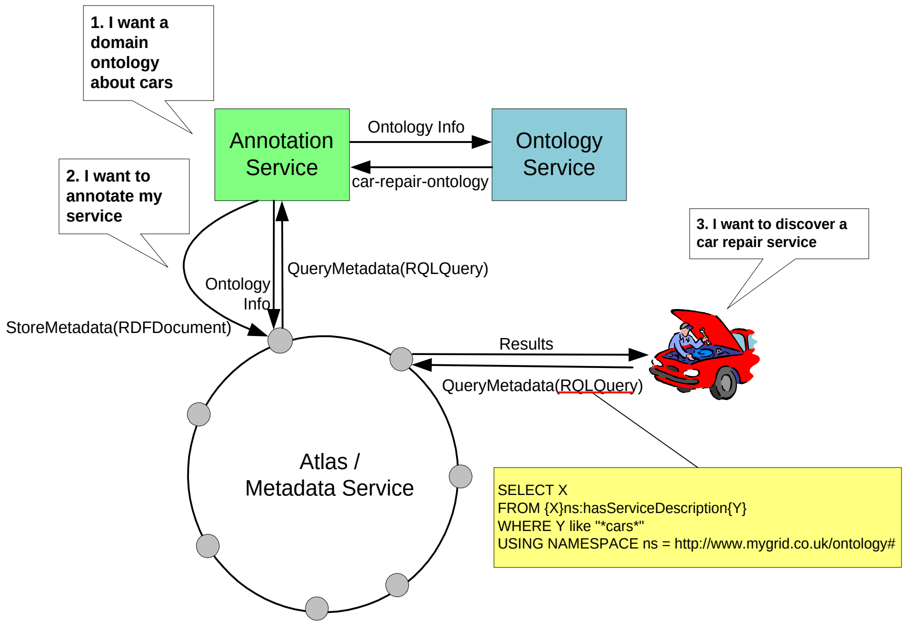

Fig. 2. Using Atlas for Service Annotation and Discovery

Using these ontologies, the annotation service can complete the service annotation process. The result of the annotation process will be stored in Atlas by calling the `UpdateMetadata` operation (see Figure 1). The ontology used for describing the service should have been stored previously in Atlas by calling the `StoreOntology` operation.

Let us suppose now that an OntoKit user wants to discover a service for repairing cars. This is accomplished by submitting RQL queries using appropriate service and domain ontologies (see Figure 2).

Finally, notice that after an annotation is stored, it might be necessary to be able to update it. An appropriate update operation can be expressed in RUL and executed in Atlas.

5 Conclusions

We have argued that resource discovery services for Semantic Grids can be made scalable, fault-tolerant, robust and adaptive, by exploiting distributed RDF query processing algorithms implemented on top of DHTs. We discussed the implementation of our ideas in the system Atlas and its role in the Semantic Grid toolkit OntoKit.

References

1. jxta. http://www.sun.com/software/jxta/.
2. Open Grid Services Architecture Data Integration (OGSA-DAI). http://www.ogsadai.org.uk/.
3. Artur Andrzejak and Zhichen Xu. Scalable, Efficient Range Queries for Grid Information Services. In *the second IEEE International Conference on Peer-to-Peer Computing (P2P2002)*, Linkoping, Sweden, 5-7 September 2002.
4. Hari Balakrishnan, M. Frans Kaashoek, David R. Karger, Robert Morris, and Ion Stoica. Looking up data in P2P systems. *Communications of the ACM*, 46(2):43–48, 2003.
5. Ali Raza Butt, Rongmei Zhang, and Y. Charlie Hu. A Self-Organizing Flock of Condors. In *Proceedings of Supercomputing Conference (SC)*, Phoenix, Arizona, November 2003.
6. M. Cai, M. Frank, and P. Szekely. MAAN: A Multi-Attribute Addressable Network for Grid Information Services. In *Proceedings of the 4th International Workshop on Grid Computing (Grid2003)*, 2003.
7. Min Cai, Ann Chervenak, and Martin Frank. A Peer-to-Peer Replica Location Service Based on A Distributed Hash Table. In *the 2004 ACM/IEEE Conference on Supercomputing (SC2004)*, Pittsburgh, November 2004.
8. Min Cai, Martin R. Frank, Baoshi Yan, and Robert M. MacGregor. A Subscribable Peer-to-Peer RDF Repository for Distributed Metadata Management. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 2(2):109–130, December 2004.
9. Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Riccardo Rosati, and Guido Vetere. Hyper: A framework for peer-to-peer data integration on grids. In *Proceedings of the International Conference on Semantics of a Networked World: Semantics for Grid Databases (ICSNW 2004)*, pages 144–157, 2004.
10. Jon Crowcroft, Tim Moreton, Ian Pratt, and Andrew Twigg. *The GRID2: Blueprint for a New Computing Infrastructure*, chapter Peer-to-Peer Technologies. 2004.
11. M. Esteban Gutirez (ed), S. Bechhofer, O. Corcho, M. Ferndez-Lez, A. Gez-Perez, Z. Kaoudi, I. Kotsiopoulos, M. Koubarakis, M C. Suez-Figueroa, and V. Tamma. Specification and design of ontology grid compliant and grid aware services. Deliverable 3.1 OntoGrid project.
12. Ian Foster and Adriana Iamnitchi. On Death, Taxes, and the Convergence of Peer-to-Peer and Grid Computing. In *2nd International Workshop on Peer-to-Peer Systems (IPTPS'03)*, Berkeley, CA, February 2003.
13. Carole A. Goble and David De Roure. The Semantic Grid: Myth Busting and Bridge Building. In *Proceedings of ECAI*, pages 1129–1135, 2004.
14. A. Harrison and I. Taylor. Dynamic Web Service Deployment Using WSPeer. In *Proceedings of 13th Annual Mardi Gras Conference - Frontiers of Grid Applications and Technologies*, pages 11–16. Louisiana State University, February 2005.
15. Felix Heine, Matthias Hovestadt, and Odej Kao. Towards Ontology-Driven P2P Grid Resource Discovery. In *5th International Workshop on Grid Computing (GRID 2004)*, pages 76–83, Pittsburgh, PA, USA, November 2004.
16. A. Iamnitchi, I. Foster, and D.C. Nurmi. A peer-to-peer approach to resource discovery in grid environments. Technical Report TR-2002-06, University of Chicago, 2002.

17. Adriana Iamnitchi and Ian Foster. On fully decentralized resource discovery in grid environments. In *International Workshop on Grid Computing*, Denver, Colorado, 2001. IEEE.
18. Adriana Iamnitchi, Ian Foster, and Daniel C. Nurmi. A Peer-to-Peer Approach to Resource Location in Grid Environments. In *Proceedings of the 11th Symposium on High Performance Distributed Computing*, Edinburgh, UK, August 2002.
19. Zoi Kaoudi, Iris Miliaraki, Spiros Skiadopoulos, Matoula Magiridou, Erietta Liarou, Stratos Idreos, and Manolis Koubarakis. Specification and Design of Ontology Services and Semantic Grid Services on top of Self-organized P2P Networks. Deliverable D4.1, Ontogrid project, September 2005.
20. Greg Karvounarakis, Sofia Alexaki, Vassilis Christophides, Dimitris Plexousakis, and Michel Scholl. RQL: A Declarative Query Language for RDF. In *Proceedings of the 11th International World Wide Web Conference*, May 2002.
21. A. Kothari, D. Agrawal, A. Gupta, and S. Suri. Range Addressable Network: A P2P Cache Architecture for Data Ranges. In *Third International Conference on Peer-to-Peer Computing (P2P'03)*, Linkoping, Sweden, 2003.
22. Ioannis Kotsiopoulos, Sean Bechhofer, Pinar Alper, Paolo Missier, Oscar Corcho, Dean Kuo, and Carole Goble. Specification of a Semantic Grid Architecture. Deliverable 1.2, OntoGrid project.
23. Erietta Liarou, Stratos Idreos, and Manolis Koubarakis. Publish-Subscribe with RDF Data over Large Structured Overlay Networks. In *Proceedings of the 3rd International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2005)*, Trondheim, Norway, 28-29 August.
24. Phillip Lord, Pinar Alper, Chris Wroe, and Carole Goble. Feta: A light-weight architecture for user oriented semantic service discovery. In *Proceedings of The Semantic Web: Research and Applications: Second European Semantic Web Conference (ESWC 2005)*, Heraklion, Crete.
25. Matoula Magiridou, Stavros Sahtouris, Vassilis Christophides, and Manolis Koubarakis. RUL: A Declarative Update Language for RDF. In *Proceedings of the Fourth International Semantic Web Conferece (ISWC2005)*, 2005.
26. Sean Rhea, Dennis Geels, Timothy Roscoe, and John Kubiatowicz. Handling Churn in a DHT. In *USENIX Annual Technical Conference*, 2004.
27. José Oliver Segura, Richard Benjamins, José Manuel Gómez Pérez, Jesús Contreras, Robert Salla, Óscar Corcho, Rafael González, Guadalupe Aguado de Cea, Inmaculada Álvarez de Mon y Rego, Antonio Pareja Lora, and Rosario Plaza Arteche. Specification and Design of Annotation Services. Deliverable D5.1, Ontogrid project, March 2005.
28. Domenico Talia and Paolo Trunfio. Toward a Synergy Between P2P and Grids. *IEEE Internet Computing*, 7(4):94–96, 2003.