Dieter Gawlick, Oracle Corporation
Vitthal Gogate, IBM Almaden Research Center
Cecile Madsen, IBM Silicon Valley Laboratory
Shailendra Mishra, Oracle Corporation
Inderpal Narang, IBM Almaden Research Center
Mahadevan Subramanian, IBM Almaden Research Center

Category: Informational
OGSA-DAI
9/19/2003

Data Distribution in the Grid Environment

Status of This Memo

This memo provides information to the Grid Data Access and the Grid Data Replication communities. It does not define any standards or technical recommendations. Distribution is unlimited.

Copyright Notice

Abstract

Data Grids [1] are distributed environments where applications access, distribute and manage data at a very large scale across organizational boundaries, be it by the size of the data or by geographical distance. Data Grids have requirements of their own not only to protect the critical data they manage (security, authorization, auditing, etc.) but also to guarantee transparent and efficient access and distribution of data with quality of service.

These requirements are answered by the "Grid Data Distribution" here after known as (GDD) model. The GDD model supports dynamic and efficient data distribution of customized information based on consumers interest. The GDD model is based on publish/subscribe paradigm that scales to a large number of consumers.

The outline of the paper is as follows. First we give definition of the Grid Data Distribution components. Second, we position Grid Data Distribution functionality with respect to Grid Data Access [2] and Notification [3]. Then we describe in detail the operations that are introduced to support the new functionality, which are defined as extensions of the existing Data Access and Notification specifications. We describe the use case scenarios of GDD for Data Replication and 3rd party Data delivery. Finally, it outlines the remaining work and open issues in this area.

Contents

1. Introduction

The Grid Data Distribution model (**GDD**) allows one to share data and events between publishers and subscribers. Publishers are sources of data and events. Subscribers are consumers of data and events. Data and events can be pushed or pulled from the publisher to the subscriber. The publisher can notify subscribers of the existence of data or events. Using GDD, one can control how information is published, shared and consumed. The model supports efficient asynchronous distribution of data, so publishers don't necessarily need to know of the target recipients and vice-versa.

In essence, the Data Distribution model provides a functionality extensions to Access and Notification by introducing new interfaces and operations.

1.1 Grid Data Distribution and the OGSA Data Service

Data Distribution portTypes are defined for Data Services as extensions of the Data Management port type to support:

- Operations for both administrative tasks (define rules for Data Publication and Data Subscription) and operational tasks (publish/subscribe/deliver) should implement these portTypes.

We propose the following scoping of Data Distribution with respect to Data Access functionality:

- Define both the synchronous and asynchronous data access *"operations by reference"* as mentioned in [], to be derived as a special case of Data Distribution. For example, in DAIS specifications, the "sqlQueryByRefASync" and "sqlQueryByRefSync are now considered (asynchronous) data distribution operations.

1.2 Grid Data Distribution and OGSI Notification

The Data Distribution model is a subscription-based model that extends the existing Notification specification to support:

- Event-based, scheduled-based, continuous Data Distribution.
- Efficient scaling to a very large number of subscribers (batch/grouping/efficient Replication models, etc.).
- Consistency requirements (transactional, etc.).
- Efficient proprietary data distribution mechanisms, as provided by existing vendors (this will typically imply dedicated data channels between data source and target).
- Existing distribution topologies (pub/sub, brokering, etc.).
- Controlled security, authorization, auditing and tracking of the distributed data.

2. Definition and Requirements

2.1 Data Publication

Data Publication occurs at the Data Service which acts as the logical source of data. The Data publication rules may include what data is to be published, and may further include to whom published data is to be delivered and how.

2.2 Data Subscription

Data Subscription occurs at the Data Service which acts as the logical data source. The Data subscription specifies:

- A rule or a set of rules seeking matching entries.
- Data seeking matching rules.
- Mix of the above.

Data subscription specifications includes QOS (best effort, only once, etc.), auditing, tracking, non-repudiation requirements (how do I track my data), what do I do with the data once it reaches the target site… perform format/type conversions, apply the data etc.

2.3 Rules

Rules are XML elements that are used in the publish and subscribe operations to determine selectivity on data. Thus rules fall into two categories namely:

1. Publication rules:

   The publication rules are invoked before publication on data to determine whether data is eligible for publication or not.

2. Subscription rules – this includes subscription rules involving direct peer-to-peer data movement. The subscription rules may be invoked on publication to compute the subscription list for this data.

2.4 Event Alerts:

Alerts allows clients to subscribe to data from GDD instance. It is a two step process:

- Subscribes to Data service using the data subscription interface. The client is allowed to specify rules as well, as part of this interaction.
- Registers its location with the Data service.

On an event, the Data service invokes the rules applicable to event data and computes the subscription list. The Data service then delivers the alert to the available subscribers.
During the registration step a client may specify some of the following:

- protocol, quality of service etc. regarding how it wants to be notified.
- whether it wants to receive data or notification only.

3. Interfaces and Operations – Work in Progress

4. Scenarios – Work in progress

5. Open Issues – Work in progress

6. Security Considerations – Work in progress

This is a REQUIRED section.

Author Information

Dieter Gawlick
Oracle Corporation
500 Oracle Parkway
Redwood Shores
CA 94065
(650) 506 8706
dieter.gawlick@oracle.com

Vitthal Gogate
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099
(408) 927 1799
gogate@almaden.ibm.com

Cecile Madsen
IBM Silicon Valley Laboratory
555 Bailey Avenue
San Jose, CA 95141
(408) 463 2578
madsen@us.ibm.com

Shailendra Mishra
Oracle Corporation
500 Oracle Parkway
Redwood Shores
CA 94065
(650) 506 9123

Inderpal Narang
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099
(408) 927 1743
narang@almaden.ibm.com

Mahadevan Subramanian
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099
(408) 927 1777
maha@almaden.ibm.com

Glossary

Recommended by not required.

Intellectual Property Statement

The GGF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the GGF Secretariat.

The GGF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this recommendation. Please address the information to the GGF Executive Director.

Full Copyright Notice

This document and the information contained herein is provided on an "AS IS" basis and THE GLOBAL GRID FORUM DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE."

References

1. A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, S. Tuecke, **The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets**, *Journal of Network and Computer Applications*, 23:187-200, 2001 (based on conference publication from Proceedings of NetStore Conference 1999).
2. N.W. Paton, M.P. Atkinson, V. Dialani, D. Pearson, T. Storey and P. Watson, **Database Access and Integration services on the Grid**, Technical Report UKeS-2002-3, National e-Science Centre, 2002.
3. **Open Grid Services Infrastructure (OGSI)** Version 1.0, Global Grid Forum, June 27, 2003.