GGF10 OREP Working Group Minutes
March 12, 2004

Ann Chervenak, one of the OREP chairs, led the meeting.

Agenda:

- Brief review of OREP RLS design and specification
- WS-RF Issues
- Prototype implementation of RLS specification, including a discussion of policy issues and some initial performance numbers

RLS Grid Service design

Based on Data Services:
A data service represents and encapsulates a data virtualization, which is an abstract view of some data

- Service data elements (SDEs) describe key parameters of the data virtualization
- OGSI Grid Service Handles globally and uniquely identify data services

Also based on OGSI Service Groups

- ServiceGroups are Grid services that maintain information about a group of other Grid services
- A ServiceGroup contains entries for member services
- Entries are represented as Service Data Elements (SDEs) of the ServiceGroup
- We extend the ServiceGroupRegistration port type's add and remove methods for RLS

Summary of Replica Location Grid Service Design

- Data items are Grid services
- Replicated data items are grouped together in a set
- Want to expose these replica sets as services
- Thus define a ReplicaSet Grid service as a virtualization of a set of replicas
- This ReplicaSet is globally and uniquely identified by a Grid Service Handle
- Effectively, a ReplicaSet provides a mapping from the locator (handle) of the ReplicaSet to one or more locators for member data services
- Represent information about member data services as service data elements (SDEs) of the ReplicaSet service
- ReplicaSet service data may include policy information for ReplicaSet
- E.g., replicas are byte-to-byte copies, have matching checksum, consistency is or is not maintained, etc.
- A client may use standard methods to obtain information about replicaSet members and policies (Inspection, Subscription/notification)

Policies that might be associated with RLS Grid service

- Authorization Policies
    - Who is allowed to create a new ReplicaSet
    - Who is allowed to add/remove entries from ReplicaSet
- Semantic definition of replication
    - Byte-for-byte copies (e.g., same checksum)
    - Versions

    - Copies synchronized within some time frame
    - Same content in different formats
- How and when replica semantics are enforced
    - No enforcement (assume high level of trust)
    - Continuous enforcement (maintain consistency)
    - Enforce at time new member is added to ReplicaSet

Changes to OREP Spec due to WS-RF

- Web Service Resource Framework (WS-RF): An emerging set of standards for specifying stateful resources in Web service environments
- WS-RF standards are intended eventually to supercede the OGSI standards upon which the design of our Replica Location Service is based
- In particular, we need to adapt to changes in:
    - Data Services: being adapted through DAIS working group
    - ServiceGroups: new WS-Service Group specification being developed

Discussion of an RLS Grid Service Implementation that is based on the OREP Specification

- Prototype implementation in Globus Toolkit Version 3 (GT3) environment
- Note that the prototype is a file-based implementation
- The specification is NOT file-based
- Intention is to generalize the implementation as Data Services become available

Implementation Features

1) Identifying replicated data items

- OGSA Data Services have not yet been implemented in the GT3 environment
- Our implementation uses file URLs as locators for data objects
- Thus, our implementation is file-based, although the specification is general and not limited to files
- When data services are eventually implemented, file URLs will be replaced by the Grid Service Handles or WS-Addresses of data services

2) Replica Semantic Policies supported by RLS implementation

- Policy 1
    - GSI Authentication, Gridmap authorization at ReplicaSetFactory and ReplicaSet
    - Require replicas to have same checksum as first member added to a ReplicaSet
    - Enforce at time of addition
- Policy 2
    - GSI Authentication, Gridmap authorization at ReplicaSetFactory and ReplicaSet
    - No other validation of members added to ReplicaSet
    - Assumes a trusted environment
    - Policy 1 requires verifying the checksum for the data file before adding it to a ReplicaSet.
    - Overhead is proportional to the size of the file and can be substantial for large files, as shown in our performance results
    - Comment from WG: Should be able to use more efficient checksum algorithms to reduce the overhead of this calculation. The MD5 checksum is very costly to compute.
    - Alternative policy enforcement implementation would allow a client to assert a checksum value for the file
    - This is based on trust; the ReplicaSet only needs to verify that the asserted checksum matches the checksum of master copy

- Our implementation only verifies checksum when a new member is added to a ReplicaSet; thereafter does not enforce consistency
- Alternate policies could support stronger consistency
- More enforcement incurs additional overheads

3) ReplicaSet Authorization Policies
- Each ReplicaSet determines whether a client is allowed to add or remove ReplicaSet entries
- Restrictions based on a Globus grid-mapfile, which specifies the users allowed to add or remove ReplicaSet members

4) ReplicaSetFactory
- To create a new instance of a ReplicaSet service, a client makes a request to a ReplicaSetFactory
- ReplicaSetFactory extends OGSI Factory service
- ReplicaSet factory determines whether a client is allowed to create a new ReplicaSet instance
- Uses standard Globus grid-mapfile

5) Port types
- Extend ServiceGroupRegistration port type to modify add, remove methods
- Add function (Policy 1) requires:
  - Copying file from remote location to local storage
  - Performing MD5 checksum
  - Verifying match with ReplicaSet checksum
  - Adding member if matches master copy checksum

6) File transfer method
- GridFTP data transport
- Reliable File Transfer Service

7) Service data and associated methods
- numOfReplicas and GetNumOfReplicas method
- Checksum and GetChecksum method
- GetListOfReplicas: returns a list of file URLs for members of the ReplicaSet

8) Faults
- ReplicaAlreadyExists
- CopyException : due to file permission problems, failure of RFT, etc.
- FileTooLarge: ReplicaSet can configure a size limit

Performance Results: See slides
- Transfer and checksum calculation proportional to file size; other overheads fairly constant
- Without checksum verification, adding a member takes about 5 seconds
- Overhead is GT3 including GSI authentication

Working Group comments on implementation
- Concern over the scalability of having every replicated data item be a ReplicaSet service
- Concern over the overhead of creating Grid services in GT3: the memory requirements of the factory during service creation and the memory consumption of a service in a container
- Concern over need to resolve Grid service handles for all the ReplicaSets in the system
- Question: does this scale or can it scale to millions of replica sets?

- Another concern is persistency of the ReplicaSet service; current implementation does not include persistency (saving service state and recovering from service failures); this is clearly needed in the ReplicaSet implementation to avoid losing replica information when services crash

Future Work Planned for RLS Grid Service specification

- More extensive performance measurement of this implementation
  - Remove operations
  - Queries
  - ReplicaSet creation operations
  - Wide area performance
- Will implement higher-level index described in OREP Specification to provide better availability and performance
- Incorporate WS-Agreement in policy specifications
- OREP Specification and our implementation will evolve to accommodate WS-Resource Framework

Summary

- Minimal changes in specifications for current meeting (GGF10)
- Major work required between now and GGF11 to incorporate WS-RF changes
- Will schedule at least one intermediate face-to-face meeting in April or May
  - Watch OREP mailing list for scheduling
- Will continue with implementation

Discussion on Future of OREP Group

- OREP charter envisions a family of specifications
  - RLS is low-level specification that groups together replicas
  - Higher-level specifications would provide replica management, consistency, subscription, etc.
- Inviting additional participants
- Alternative is to allow OREP group to finish after RLS specification is complete