Replication on the AMGA Metadata Catalogue

Enabling Grids for E-sciencE

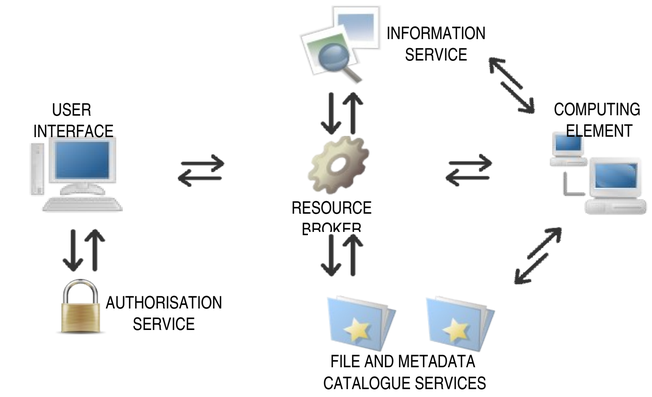
Nuno Santos, Birger Koblitz

Background

When the Large Hadron Collider (LHC) begins operation at CERN in 2007 it will produce data in volumes never before seen. Physicists spread across over two hundred sites around the world will manage, distribute and analyse Petabytes of data using the middleware provided by the LHC Computing Grid.

File and Metadata Catalogues Services are essential for discovering and locating the data required by users and jobs. Therefore, they must be Grid-wide and be prepared to scale to service thousands of users and jobs spread over more than two hundred grid sites. The scalability, dependability and fault-tolerance required to operate successfully on such a challenging environment can only be provided by replication and distribution mechanisms.

ARDA is studying and implementing distribution mechanisms on the AMGA Metadata Catalogue, with focus on the requirements of EGEE applications.



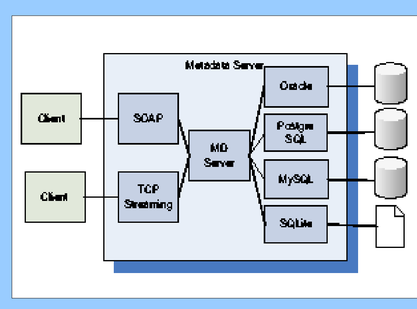
Overview of a typical Grid. File and Metadata Catalogues are used by users and jobs for discovering and locating data among the hundreds of Grid sites.

The AMGA Metadata Catalogue

AMGA is the **gLite Metadata Catalogue** and is currently being used by several groups on different user communities, including High Energy Physics, Biomed and UNOSAT. Main features:

- **Modular back-end** - supports Oracle, PostgreSQL, MySQL and SQLite
- **Modular front-end** – high performance TCP Streaming interface and standard Web-Services frontends.
- **Hiearchical organisation** – metadata organised in a tree-like structure.
- **Dynamic Schemas** – Schemas can be created, modified and delete by clients at run-time.

AMGA Architecture



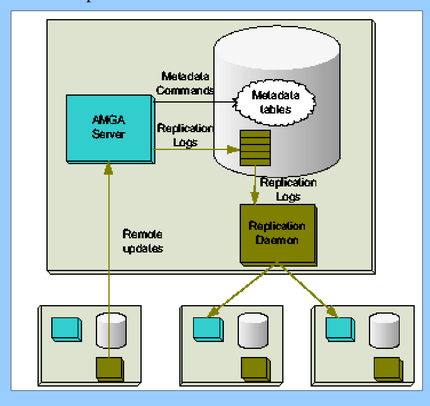
Architecture

Replication in AMGA was designed from the ground-up targeting the requirements of Grid Metadata Catalogues and the challenges of a Grid environment. Its main architectural principles are:

- **Asynchronous replication** – Scale to large number of nodes spread across WANs.
- **Master/Slave** – Writes allowed only at a single node, and replicated to all slaves.
- **Catalogue-level replication –** Implemented by replicating metadata commands. Provides **database independence** and works with any back-end supported by AMGA (Oracle, PostgreSQL, MySQL, SQLite).
- **Partial Replication** – A slave can subscribe only a part of the metadata tree.
- **Federation** – A node can subscribe different parts of the metadata tree from different masters.

The internal architecture was inspired by Oracle Streams [3] and Slony-I [4] (PostgreSQL). The master generates replication logs and saves them a local table. An external process, the Replication Daemon, monitors this table and ships the logs to all interested slaves, where they are replayed locally.

Replication Architecture in AMGA



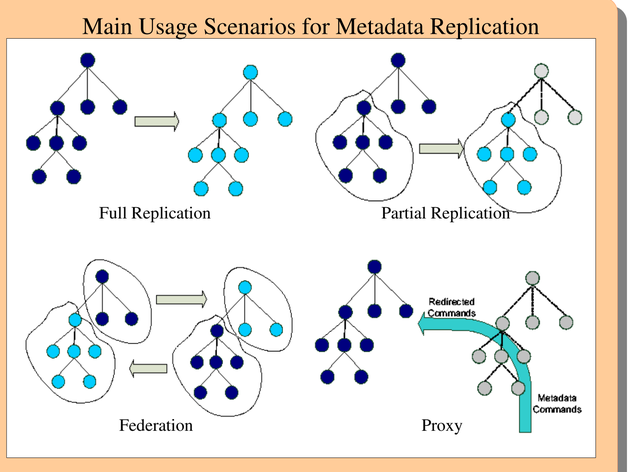
Replication Use Cases

Replication on AMGA is designed to cover a broad range of usage scenarios that are typical of the main user communities of EGEE.

- **High Energy Physics (HEP)** – Large amounts of read-only metadata, produced on a single location and accessed by thousands of physicists spread across hundreds of remote sites. Full and partial replication are required for providing the required scalability and performance.

- **Biomed** – Highly sensitive metadata produced by many sites. Security is the main concern, and therefore replication is not appropriate as it would increase the exposure of the metadata. Federation of individual catalogues into a single virtual distributed catalogue allows data to remain secure on its origin site, while providing transparent access to authorized users regardless of their location.

Main Usage Scenarios for Metadata Replication



Current Status

We have recently finished a prototype supporting all the functionalities mentioned above. It is fully integrated on AMGA and does not require any external middleware.

The prototype is currently undergoing internal testing and soon we will start working with the interested communities, with the goal of better evaluating our ideas and of obtaining user feedback to guide us through further development of the replication mechanisms.

Future Plans

The distribution mechanisms described above are the basis for studying and developing more advanced scalability and dependability techniques suited for Distributed Metadata Catalogues on Data Grids. We are considering the following research topics:

- Distribution of replication logs using **group communication**, to provide the scalability and fault-tolerance required for supporting hundreds of replicas.
- **Decentralized discovery and location** mechanisms to locate a particular metadata collection among a set of distributed catalogues.
- **Dynamic replication** of metadata to cope with changing network and system conditions.
- Extension of the master/slave model to efficiently support writes from any location on the Grid

References

1. N. Santos, and B. Koblitz, "Metadata services on the grid". In Proc. of Advanced Computing and Analysis Techniques (ACAT'05), Zeuthen, Berlin, May 2005
2. AMGA Web Page, http://project-arda-dev.web.cern.ch/project-arda-dev/metadata/
3. Oracle Streams, http://www.oracle.com/technology/products/dataint/htdocs/streams_fo.html
4. Slony-I for PostgreSQL, http://slony.info/

Acknowledgements

The authors wish to thank the ARDA team for all of their help and support. Nuno Santos would also like to thank the Portuguese Foundation for Science and Technology (FCT) for his funding.

Conclusion

The scalability and dependability required from Grid Catalogue Services can only be achieved using replication and distribution mechanisms. We are designing such mechanisms into the AMGA Metadata Catalogue, targeting the requirements of the EGEE user communities. We have completed an initial prototype using the AMGA metadata catalogue and we will begin tests with the interested communities soon. We are also starting to work on mechanisms to further improve the fault-tolerance and scalability of the system.