PhD Activity Report

A. Correia*

February 9, 2006

PhD Student - Alfrânio Tavares Correia Júnior

Thesis - Open Database Replication Based on Group Communication

Adviser - Rui Carlos Mendes Oliveira

Start - 01/01/2005

End - 01/01/2009

*Contact Author. email: alfranio@lsd.di.uminho.pt

1 Introduction

Synchronous database replication based on group communication appears as a solution to circumvent the scalability and performance issues related to traditional replication protocols based on distributed locking.

To fully exploit group communication based replication, the underlying database engine is expected to cooperate with a set of functionalities that can be directly provided in core or, otherwise recreated or indirectly obtained through middleware components. The protocols should also adapt themselves to a variety of workloads in such a manner that a protocol could fall back on a more restrictive solution in terms of performance to avoid conflicts and thus aborts.

In such context, this paper reports on the research, which constitutes the PhD project of the author, on an architecture for generic database replication encompassing a set of components acting at the database or at the middleware level that can efficient and gracefully accommodate a broad set of group communication based replication protocols.

This research is funded by the FCT, reference SFRH / BD / 18852 / 2004 and is inserted in the Gorda Project, reference FP6-IST2-004758.

2 Related Work

In contrast with replication based on distributed locking and atomic commit protocols, group communication based protocols minimize interaction between replicas and the resulting synchronization overhead by relying on total order multicast to ensure consistency. Generically, the approach builds on the classical replicated state machine [Mul89]: The exact same sequence of update operations is applied to the same initial state, thus producing a consistent replicated output and final state. The problem is then to ensure deterministic processing without overly restricting concurrent execution, which would dramatically reduce throughput, and avoid re-execution in all replicas.

These concerns have been addressed by several proposals based on group communication [KA00, PMKA02, Ped99, KA98]. Although all rely on a totally ordered multicast for consistency, they differ mainly in whether transactions are executed conservatively [KA00, PMKA02] or optimistically [Ped99, KA98]. In the former, by a priori coordination among the replicas, it is assured that when a transaction executes there is no concurrent conflicting transaction being executed remotely and therefore its success depends entirely on the local database engine. In the latter ones, execution is optimistic, each replica independently executes its locally submitted transactions and only then, just before committing, sites coordinate and check for conflicts between concurrent transactions.

This difference results in multiple and often subtle performance and resiliency trade-offs. Namely, how does each protocol cope with a large share of update transactions, conflicting updates, high latency in wide area networks, and symmetric load to multiple replicas. Unfortunately, each protocol is presented using a different load scenario and evaluation method which makes it very hard to clearly highlight the main consequences of the approach. In [JSS+05] we address this problem and present comparisons among a set of protocols based on group communication by using a common framework.

3 Developed Activities (First Year)

3.1 Activity 01

Proposal of an architecture for generic database replication and development of a preliminary set of interfaces that are now being evaluated by the Gorda's partners. For further details, see [Cor05].

3.2 Activity 02

Development of a prototype using the PostgreSQL as a the target database. Specifically, we defined and built light-weight triggers that allow us to easily implement different replication protocols based on group communication. Our current implementation is available for download in [gor].

3.3 Activity 03

Study and definition of a minimum set of tuples read during the transaction's execution, such that no proper subset exists that could be used to re-evaluate the transaction's statements, producing the same results, that is, the same resulting database and response. This set, usually called read set, is a core component for the database replication protocols based on group communication and to the best of our knowledge nobody has ever given an intuition on how to obtain it. For further details, see [CO05].

3.4 Activity 04

We pointed out a problem in the current database replication protocols based on group communication. Basically, in order to improve performance, the read-only transactions are not handled by these protocols thus avoiding any interactions among the replicas. Although it seems to be quite intuitive optimization, in a multi-tier environment clients should access stale information. For instance, in a cluster a client could experience to read different values from its previous writes on the same set of values even if no other concurrent transaction changes these values. For further details, see [OPJAar].

3.5 Activity 05

Development and evaluation of a set of database replication protocols in order to figure out their behavior when submitted to OLTP workloads. For further details, see [JSS+05, SPS+05b, SPS+05a].

From this experiment, we devised the ideas to build an hybrid and adaptive protocol to handle different workloads. This protocol was built and provides better results in terms of performance and scalability when compared to others protocols based on group communication.

4 Future Activities (Next Year)

4.1 Activity 01

We intent to prove correctness of the protocol devised in Sect. 3.5 by using TLA+ [Lam03].

We also expect to develop different consistency criteria [JSS+04] in order to improve performance.

4.2 Activity 02

We are validating our architecture and we will implement it inside Derby database [der] and PostgreSQL and most likely in a middleware solution such as the C-JDBC.

References

[CO05] A. Correia and R. Oliveira. Thrifty read sets for resilient database replication. Technical report, Departamento de Informática, Universidade do Minho, 2005.

[Cor05] Alfranio Correia. Towards an Architecture for Generic Database Replication. In *WTD - Third Workshop on Theses and Dissertations in Dependable Computing, collocated at 2nd Latin-American Symposium on Dependable Computing*, 2005.

[der] http://db.apache.org/derby/.

[gor] Gorda Project. http://gorda.di.uminho.pt/community.

[JSS+04] A. Correia Jr., A. Sousa, L. Soares, F. Moura, and R. Oliveira. Revisiting Epsilon Serializability to improve the Database State Machine (Extended Abstract). In *Workshop on Dependable Distributed Data Management, IEEE SRDS*, 2004.

[JSS+05] A. Correia Jr., A. Sousa, L. Soares, J. Pereira, F. Moura, and R. Oliveira. Group-based Replication of On-line Transaction Processing Servers. In *2nd Latin-American Symposium on Dependable Computing*, 2005.

[KA98] Bettina Kemme and Gustavo Alonso. A Suite of Database Replication Protocols Based on Group Communication Primitives. In *IEEE ICDCS*, 1998.

[KA00] B. Kemme and G. Alonso. Don't Be Lazy, Be Consistent: Postgres-R, A New Way to Implement Database Replication. In *VLDB Conference*, 2000.

[Lam03] Leslie Lamport. *Specifying Systems, The TLA+ Laguange and Tools for Hardware and Software Engineers.* Addison-Wesley, 2003.

[Mul89] Sape Mullender. Distributed Systems. In *Distributed Systems.* ACM Press, 1989.

[OPJAar] Rui Oliveira, José Pereira, Alfranio Correia Jr., and Edward Archibald. Revisiting 1-copy equivalence in clustered databases. In *ACM SAC*, 2006 (to appear).

[Ped99] F. Pedone. *The Database State Machine and Group Communication Issues.* PhD thesis, Département d'Informatique, l'École Polytechnique Fédérale de Lausanne, 1999.

[PMKA02] Ricardo Jiménez Peris, M. Patiño Martínez, Bettina Kemme, and Gustavo Alonso. Improving the Scalability of Fault-Tolerant Database Clusters. *IEEE ICDCS*, 2002.

[SPS+05a] A. Sousa, J. Pereira, L. Soares, A. Correia Jr., L. Rocha, R. Oliveira, and F. Moura. Experimental Evaluation of Database Replication Protocols. In *Work-in-Progress Session, collocated at Performance Evaluation*, 2005.

[SPS+05b] A. Sousa, J. Pereira, L. Soares, A. Correia Jr., L. Rocha, R. Oliveira, and F. Moura. Testing the Dependability and Performance of GCS-Based Database Replication Protocols. In *IEEE DSN*, 2005.