Development and evaluation of database replication in ESCADA*

A. Sousa L. Soares A. Correia Jr. F. Moura R. Oliveira
Universidade do Minho

Abstract

Software based replication is a highly competitive technique to improve the dependability of database systems. However, the unavoidable trade-off between consistency and performance causes some discredit among database designers with respect to synchronous, strong consistent, replication protocols. This is usually due to performance and scalability problems as classic distributed locking based protocols lead to high resource contention, high transaction latency and high deadlock rates [4]. As a result, commercial database products often privilege asynchronous (or lazy) replication protocols in order to boost performance at the expense of data consistency.

Asynchronous replication is not transparent for the user and therefore cannot be generically applied. Moreover, while strong consistency criteria such as 1-copy-serializability [2] is rigorously defined, relaxed criteria are often ambiguous, hard to formalize, and based on the belief of eventual replica convergence.

To overcome the above problems, a suite of group based communication protocols has emerged and has been the focus of a considerable body of research [1, 8, 10, 5, 7, 3, 6, 9]. Basically, the main and common characteristics of these protocols are the optimistic transaction execution based on deferred updates [2] and the use of total ordered broadcast primitives to enforce a unique sequence of committed transactions.

In some sense, these protocols avoid the efficiency issues of classic distributed locking based protocols by not coordinating the execution of (remote) concurrent transactions and disallow replica divergence of asynchronous replication protocols by aborting transactions that would otherwise violate serializability.

This paper reports our experience on the development and evaluation of group communication based database replication protocols in the ESCADA project.

References

[1] Y. Amir, D. Dolev, P. Melliar-Smith, and L. Moser. Robust and efficient replication using group communication. Technical Report CS94-20, The Hebrew University of Jerusalem, November 1994.

[2] P. Bernstein, V. Hadzilacos, and N. Goodman. *Concurrency Control and Recovery in Database Systems*. Addison-Wesley, 1987.

[3] U. Fritzke and P. Ingels. Systéme transactionnel pour donnés partiellement dupliqués, fondé sur la communication de groupes. Technical Report 1322, INRISA, Rennes, France, April 2000.

[4] J. Gray, P. Helland, P. O'Neil, and D. Shasha. The dangers of replication and a solution. pages 173–182, 1996.

[5] B. Kemme and G. Alonso. A suite of database replication protocols based on communication primitives. In *Proceedings of the 18th International Conference on Distributed Computing Systems*, Amsterdam, The Netherlands, May 1998.

[6] B. Kemme and G. Alonso. Don't be lazy, be consistent: Postgres-R, a new way to implement database replication. In *Proceedings of 26th International Conference on Very Large Data Bases (VLDB 2000)*, pages 134–143. Morgan Kaufmann, 2000.

[7] F. Pedone, R. Guerraoui, and A. Schiper. The database state machine approach. Technical Report SSC/1999/008, École Polytechnique Fédérale de Lausanne, Switzerland, March 1999.

[8] A. Schiper and M. Raynal. From group communication to transactions in distributed systems. 39:84–87, April 1996.

[9] A. Sousa, A. C. Jr, F. Moura, J. Pereira, and R. Oliveira. Evaluating certification protocols in the partial database state machine. Technical report, Univ. do Minho, 2003.

[10] I. Stanoi, A. Agrawal, and E. Abbadi. Using broadcast primitives in replicated databases (abstract). In *Proceeding of the Sixteen Annual ACM Symposium on Principles of Distributed Computing*, page 283, Santa Barbara, USA, August 1997.

* Research funded by FCT, STRONGREP project (POSI / 41285 / CHS / 2001).