Brief announcement: Optimistic Algorithms for Partial Database Replication[*]

Nicolas Schiper[1], Rodrigo Schmidt[2], and Fernando Pedone[1]

[1] University of Lugano, Switzerland
[2] EPFL, Switzerland

1   Introduction

Database replication protocols based on group communication have recently received a lot of attention. The main reason for this stems from the fact that group communication primitives offer adequate properties, namely agreement on the messages delivered and on their order, to implement synchronous database replication. Most of the complexity involved in synchronizing database replicas is handled by the group communication layer. Previous work on group-communication-based database replication has focused mainly on full replication. However, full replication might not always be adequate. First, sites might not have enough disk or memory resources to fully replicate the database. Second, when access locality is observed, full replication is pointless. Third, full replication provides limited scalability since every update transaction should be executed by each replica.

2   Partially-replicated Database State Machine

In this paper, we discuss an extension of the Database State Machine (DBSM) [3], a group-communication-based database replication technique, to partial replication. The DBSM is based on the deferred update replication model [1]. Transactions execute locally on one database site and their execution does not cause any interaction with other sites. Read-only transactions commit locally only; update transactions are atomically broadcast at commit time for certification. The certification test ensures *one-copy serializability* [1] and requires every database site to keep the *writesets* of committed transactions. The certification of a transaction T

consists in checking that T's *readset* does not contain any outdated value, i.e., no committed transaction T' wrote a data item x after T read x.

A straightforward way of extending the DBSM to partial replication consists in executing the same certification test as before but having database sites only process update operations for data items they replicate. But as the certification test requires storing the writesets of all committed transactions, this strategy defeats the whole purpose of partial replication since replicas may store information related to data items they do not replicate. Ideally, sites would only store transaction information related to the data items they replicate. However, we do not want to rule out solutions that rely on building blocks (e.g., consensus) that may be oblivious to the data items replicated by the sites. In such cases, sites may momentarily store transactions unrelated to the data items they replicate. Moreover, we want to make sure each transaction is handled by a site at most once. If sites are allowed to completely forget past transactions, this constraint cannot obviously be satisfied. We capture these constraints with the following property:

- *Quasi-Genuine Partial Replication:* For every submitted transaction T, correct database sites that do not replicate data items read or written by T permanently store not more than the identifier of T.[3]

Consider now the following modification to the DBSM, allowing it to ensure Quasi-Genuine Partial Replication. Besides atomically broadcasting transactions for certification, database sites periodically broadcast "garbage collection" messages. When a garbage collection message is delivered, a site deletes all the writesets of previously committed transactions. When a transaction is delivered for certification, if the site does not contain the writesets needed for its certification, the transaction is conservatively aborted. Since all sites deliver both transactions and garbage collection messages in the same order, they will all reach the same outcome after executing the certification test. This mechanism, however, may abort transactions that would be committed in the original DBSM. In order to rule out such solutions, we introduce the following property:

- *Non-Trivial Certification:* If there is a time after which no two submitted transactions conflict, then eventually transactions are not aborted by certification.

---

[3] Notice that even though transaction identifiers could theoretically be arbitrarily large, in practice, 4-byte identifiers are enough to uniquely represent 2^{32} transactions.

In [4], we present two algorithms for partial database replication for clusters of servers. Our algorithms satisfy both Quasi-Genuine Partial Replication and Non-Trivial Certification and are optimistic, i.e., we assume *spontaneous total order*: with high probability messages sent to all servers in the cluster reach all destinations in the same order.

To the best of our knowledge, [2] and [5] are the only papers addressing partial database replication using group communication primitives. In [2], every operation of a transaction on data item x is multicast to its replicas and a final atomic commit protocol ensures transaction atomicity. In [5], the authors extend the DBSM for partial replication by adding an extra atomic commit protocol. Both of our algorithms compare favorably to [2, 5]: they either have a lower latency or make weaker assumptions about the underlying model, i.e., they do not require perfect failure detection.

3   Final Remarks

This short paper defines two properties, Quasi-Genuine Partial Replication and Non-trivial Certification. These properties characterize our view of partial replication in the DBSM. The first property forbids replicas to permanently store information about data items they do not replicate; the second property prevents trivial solutions that would abort unnecessary transactions in an attempt to satisfy the first property. Two algorithms for partial replication in the DBSM that ensure these two properties are presented in [4]. In the future, we intend to better characterize partial replication and devise efficient algorithms that satisfy a stronger property than Quasi-Genuine Partial Replication. Intuitively, Genuine Partial Replication should be defined such that only database sites that replicate data items touched by a transaction T are involved in its certification.

References

1. Philip A. Bernstein, Vassos Hadzilacos, and Nathan Goodman. *Concurrency Control and Recovery in Database Systems*. Addison-Wesley, 1987.
2. Udo Fritzke Jr. and Philippe Ingels. Transactions on partially replicated data based on reliable and atomic multicasts. In *Proceedings of the 21st International Conference on Distributed Computing Systems (ICDCS)*, pages 284–291, 2001.
3. F. Pedone, R. Guerraoui, and A. Schiper. The database state machine approach. *Journal of Distributed and Parallel Databases and Technology*, 14(1):71–98, 2003.
4. N. Schiper, R. Schmidt, and F. Pedone. Optimistic algorithms for partial database replication. Technical Report 2006, University of Lugano, 2006.
5. A. Sousa, F. Pedone, R. Oliveira, and F. Moura. Partial replication in the database state machine. In *Proceedings of the 1st International Symposium on Network Computing and Applications (NCA)*, October 2001.