Abstract

Nowadays, information technology and computing systems have a great relevance on our lives. Among current computer systems, distributed systems are one of the most important because of their scalability, fault tolerance, performance improvements and high availability.

Replicated systems are a specific case of distributed system. This Ph.D. thesis is centered in the replicated database field due to their extended usage, requiring among other properties: low response times, high throughput, load balancing among replicas, data consistency, data integrity and fault tolerance.

In this scope, the development of applications that use replicated databases raises some problems that can be reduced using other fault-tolerant building blocks, as group communication and membership services. Thus, the usage of the services provided by *group communication systems* (GCS) hides several communication details, simplifying the design of replication and recovery protocols.

This Ph.D. thesis surveys the alternatives and strategies being used in the replication and recovery protocols for database replication systems. It also summarizes different concepts about group communication systems and virtual synchrony. As a result, the thesis provides a classification of database replication protocols according to their support to (and interaction with) recovery protocols, always assuming that both kinds of protocol rely on a GCS.

Since current commercial DBMSs allow that programmers and database administrators sacrifice consistency with the aim of improving performance, it is important to select the appropriate level of consistency. Regarding (replicated) databases, consistency is strongly related to the isolation levels being assigned to transactions.

One of the main proposals of this thesis is a recovery protocol for a replication protocol based on certification. Certification-based database replication protocols provide a good basis for the development of their recovery strategies when a snapshot isolation level is assumed. In that level readsets are not needed in the validation step. As a result, they do not need to be transmitted to other replicas. Additionally, these protocols hold a writeset list that is used in the certification/validation step. That list maintains the set of writesets needed by the recovery protocol. This thesis evaluates the performance of a recovery protocol based on the writeset list tranfer (basic protocol) and of an optimized version that compacts the information to be transferred.

The second proposal applies the compaction principle to a recovery protocol designed for weak-voting replication protocols. Its aim is to minimize the time needed for transferring and applying the writesets lost by the recovering replica, obtaining in this way an efficient recovery. The performance of this recovery algorithm has been checked implementing a simulator. To this end, the Omnet++ simulating framework has been used. The simulation results confirm that this recovery protocol provides good results in multiple scenarios.

Finally, the correction of both recovery protocols is also justified and presented in Chapter 5.