Towards Robust Optimistic Approaches

R. Jiménez-Peris, M. Patiño-Martínez

Technical University of Madrid (UPM), Madrid, Spain
{rjimenez, mpatino}@fi.upm.es

1 Introduction

Optimism is a well-known technique to enhance the performance of distributed protocols. Optimistic approaches exploit properties exhibited by the system with certain likelihood, (i.e., that certain kinds of scenarios will prevail over others) to outperform the corresponding conservative protocol. These properties are usually referred as optimistic assumptions (e.g., an optimistic assumption is that reliably multicast messages in a LAN are spontaneously totally ordered). When the optimistic assumption holds, the optimistic approach is more efficient than the conservative one. However, this gain usually implies a tradeoff. That is, if the optimistic assumption does not hold, the optimistic approach is less efficient than the conservative one. This is due to the need to undo or repair the incorrect actions and the dismissal of work already done. This is precisely the Achilles' heel of traditional optimistic approaches. Therefore, what is crucial for an optimistic approach to be successful is that the resulting gains of optimism outweigh the penalties imposed by optimism failures.

Researchers have long recognized the potential benefits of using optimism and have proposed optimistic versions of conventional distributed protocols [BHG87,Ped01]. However, despite the many optimistic approaches suggested in distributed computing, they are not that common in industrial applications. The main reason for this reluctance is that whenever the optimistic assumption fails, the protocol behaves worse that the conventional protocol. This behavior might imply more messages, or undoing part of the work. It is our opinion that to increase the applicability of optimistic protocols they need to be enriched with safeguards that limit the consequences of those scenarios where the optimistic assumptions do not hold. These safeguards make optimistic approaches more robust and efficient, and therefore, more applicable. As a consequence, in those periods during which the optimistic assumptions do not hold frequently enough, the system will not degrade to unacceptable levels.

In this paper, we try to point out the possible causes of the lack of success of some optimistic protocols, and show which directions can be taken to overcome these shortcomings in order to diminish the existing reluctance in industry for this kind of protocols. We think that optimistic protocols will play a crucial role in the upcoming wide-area distributed systems. Despite that bandwidth will grow more and more, latency will always be a problem in WANs due to physical limitations.

2 Traditional Optimistic Approaches

Replication. Computer clusters are the hardware platform of choice for many types of distributed information systems, and more concretely for replicated databases. These systems usually have strong availability and consistency requirements. Reliable multicast [Bir96] has become one of the main abstractions for building fault-tolerant distributed systems. More particularly, it has become a key building block for modern information systems. Unfortunately, and in spite of the intensive research carried out in the area, there is still a lot of reluctance in the information systems community to use such protocols on account of their performance (mainly the high latency). This reluctance has prevented its widespread use in some contexts such as replicated databases. In this paper, we will focus on optimistic protocols for distributed databases [BHG87,LKS91], [GHR97,BKS97], with special emphasis on those based on reliable multicast [KPAS99], [PGS,AT02,KA00a,JPAA01].

Typical measures of efficiency in an information system are throughput and response time. The latency introduced by reliable multicast, especially when total order and/or uniformity are provided, can have a severe effect on these efficiency measurements. Thus, in practice, designers opt for protocols that provide weaker guarantees but have a lower latency, especially in WANs.

The ideal would be a protocol exhibiting the strong guarantees provided by reliable multicast (i.e. providing total order and/or uniformity) but without the latency penalty typically associated with the implementation of these guarantees. One way to achieve this goal is by using an optimistic approach. One of such optimistic approaches is taken in [KPAS99], where the spontaneous total order exhibited by IP multicast in a LAN is exploited to reduce the latency of transaction processing in a replicated database. In [KPAS99] transactions are multicast to all sites. Multicast messages are totally ordered to ensure one copy serializability, the correctness criteria for replicated databases [BHG87]. In this approach, multicast messages are delivered optimistically as soon as they are received [VR02,SPMO02]. Thus, the database can start their processing in an optimistic fashion. When the total order of the multicast message is established, the message is definitively delivered to the database system. The time elapsed between the optimistic delivery and the definitive delivery is used to process (at least, partially) the message (transaction).

The caveat of such an optimistic approach is that it can result in a high number of transaction aborts (rollbacks) when the optimistic assumption does not hold. More concretely, when the load is high and messages are retransmitted due to buffer overruns, the order in which the messages arrive at different sites differs. Therefore, conflicting transactions can be executed in different orders at different sites. The corrective action needed when the optimism fails consists in aborting those transactions that have been executed optimistically following an order that do not comply with the total order and reexecute them.

Atomic Commitment. Distributed information systems use atomic commit protocols to ensure the atomicity of their operations. Well-known examples of atomic commit protocols are the two phase commit protocol (2PC) [Gra78] or three phase commit protocols (3PC) [Ske82,KD95]. One of the intrinsic limitations of commit protocols is the incurred latency, involving several rounds of messages and forced disk writes. Some optimistic approaches have been proposed to overcome these limitations, such as the optimistic two phase commit protocol [LKS91] or the OPT two-phase commit protocol [GHR97].

The optimistic two-phase commit protocol [LKS91] takes as optimistic assumption that the most likely outcome of the commit protocol is to commit the transaction. The protocol exploits this optimistic assumption by releasing the locks of the committing transaction at each participant (site) when the participant votes "yes" (that is, once the participant is prepared). In this way, conflicting transactions are able to obtain their locks without waiting for the commit to complete. The tradeoff of the protocol is that when the transaction outcome is abort, the transactions that obtained the locks optimistically should be aborted in order to guarantee full atomicity. This is a serious drawback since it could lead to cascading aborts. For that reason, [LKS91] resorts to the application semantics and uses compensating transactions to undo semantically the effects of the transaction that optimistically released the locks. The transactions that have acquired the locks do not abort. In this way, if the optimistic assumption does not hold, there are no cascading aborts. However, unlike the approaches that will be described in the next section this safeguard is introduced by relaxing the problem to be solved. More concretely, this atomic commitment protocol only provides semantic atomicity, instead of guaranteeing full atomicity, since it sacrifices the isolation property of transactions.

Concurrency Control. Database replication protocols are classified as eager or lazy depending on whether they propagate the updates of transactions as part of the original transaction or in a different one [GHOS96]. The advantage of eager protocols is that consistency is kept among replicas. However, transaction latency increases and scalability diminishes as more replicas are added to the system [GHOS96]. Traditional eager replication approaches synchronize the acquisition of each lock at all sites, that is, a transaction gets a lock on a data item when the lock is granted at all sites, what undermines scalability. This locking protocol is used to ensure one-copy serializability. Recent advances in this area have shown that is possible to scale up by using an optimistic concurrency control method that serializes transactions according to the total delivery order of multicast [KA00a,PGS,KA00b]. These approaches extend former traditional optimistic concurrency control protocols [BHG87] with the use of total ordered multicast. In these new optimistic protocols a transaction is executed optimistically at a single site. The optimistic assumption is that transactions executed optimistically at different sites are unlikely to conflict. As part of the protocol, after executing the transaction optimistically, a site propagates the transaction updates to the rest of the sites by means of totally ordered multicast. This information exchange is used to verify whether that transaction conflicts with the transactions executed concurrently at the rest of the sites, and the total order is used as the serializing order. If there is a conflict, the optimistic assumption does not hold, and conflicting transactions must be aborted.

3 Future Direction: Robust Optimistic Approaches

In the previous section we have seen that optimism has been introduced in different distributed protocols in order to enhance the performance of the corresponding conservative protocol. In this section, we propose a new direction to overcome the problems of traditional optimistic approaches. We describe some early steps taken in this direction related to the successful application of robust optimistic distributed protocols. These protocols are enriched with safeguards that prevent a trashing behavior when optimistic assumptions do not hold frequently enough. In the following, a partially synchronous system with crash failures is assumed. More details about the underlying model are found in the given references.

Replication. [PJKA00] presents an eager data replication protocol based on total order multicast enriched with safeguards to improve its robustness. This protocol uses the same optimistic approach for multicast messages as the one described in [KPAS99]. That is, the calculation of the total order is overlapped with the optimistic execution of transactions, taking advantage of the spontaneous total order exhibited by LANs. The novelty in this approach lies in its robustness. The protocol is enhanced with a reordering algorithm. The reordering acts as a safeguard that prevents the abortion of transactions executed optimistically when the optimistic assumptions do not hold (i.e., when the spontaneous order does not match the total order). If the spontaneous and total orders differ at the site where the transaction is executed a reordering technique is used to prevent the transaction abortion. That site informs the rest of the sites about the new order when it propagates the transaction updates (without any extra messages). The reordering can be done as long as the updates of the transaction optimistically executed are a subset of the updates of the preceding transactions in the total order. A transaction aborts only if both, this property and the optimistic assumption about spontaneous total order do not hold. This reordering has the additional advantage that it does not incur in any significant overhead. The protocol has been used successfully in a replication middleware [JPAK02]. Experimental results showed that in most cases there were no aborts and that under worst case scenarios, aborts never exceeded 0.2%, therefore limiting successfully the cost of scenarios where the optimistic assumption did not hold.

Atomic Commitment. Another recent advance in this direction consists in an optimistic non-blocking atomic commitment protocol [JPAA01]. Non-blocking atomic commitment is known to have an inherent cost of two rounds of messages in a synchronous system [CL02] and, although not proven, it seems that there is a lower bound of three rounds in a partially synchronous system (at least, only protocols with three rounds have been proposed). The extra round that seems to be needed in a partially synchronous system with respect to a synchronous system is due to the possibility of false suspicions. As suggested by [KR01], it is possible to circumvent some lower bounds of agreement problems, such as non-blocking atomic commitment, by using an optimistic approach. One way to model the non-blocking atomic commitment problem is by using uniform multicast [CKV01]. The protocol proposed in [JPAA01] hides the latency introduced by the uniformity by overlapping the stabilization of uniform multicast messages with an optimistic execution of the atomic commitment protocol. Once all the votes have been received optimistically and are yes, the transaction locks are released and they can be granted optimistically to other transactions. In this way, conflicting transactions do not pay for the full cost of non-blocking atomic commitment, three rounds. Instead they are allowed to progress after the second round of messages. In this protocol the safeguard consists in limiting the optimistic execution of conflicting transactions to one level, therefore preventing cascading aborts in the unlikely case the optimism fails (this safeguard was first proposed in the OPT 2PC protocol [GHR97]). The optimistic assumption fails when the message (uniform multicast) with the last vote arrives at a single receiver optimistically and then the sender and the receiver fail before the message reaches any other receiver, and additionally, the receiver committed optimistically the transaction before failing. It should be noted that garbage collection in this protocol can be performed in the same way is done in traditional commit protocols, without requiring to keep a trace of the whole history.

Some recent work is focused on extending optimistic delivery multicast protocols to WANs. [VR02] proposes a uniform total ordered multicast for WANS in which optimistic delivery takes place after the total order is established and before the message is stable. A different approach has been taken in [SPMO02] a spontaneous total order in WANs is tentatively attempted by enforcing a homogeneous delivery time at all sites.

We think that the robust optimistic approaches followed in some of the presented protocols might be extended to other contexts and might help to advance the state of the art in this area and at the same time foster their industrial use. It is our opinion that optimistic protocols will become essential for developing the upcoming wide-area distributed systems to deal with the inherent latency of wide-area networks.

References

[AT02] Y. Amir and C. Tutu. From Total Order to Database Replication. In *ICDCS*, 2002.

[BHG87] P. A. Bernstein, V. Hadzilacos, and N. Goodman. *Concurrency Control and Recovery in Database Systems*. Addison Wesley, Reading, MA, 1987.

[Bir96] K. Birman. *Building Secure and Reliable Network Applications*. Prentice Hall, 1996.

[BKS97] Y. Breitbart, H. F. Korth, and A. Silberschatz. Optimistic protocols for replicated databases. Technical Report BL0112370-970227-07, Bell Labs, 1997.

[CKV01] G. V. Chockler, I. Keidar, and R. Vitenberg. Group Communication Specifications: A Comprehensive Study. *ACM Computer Surveys*, 33(4):427–469, December 2001.

[CL02] B. Charron-Bost and F. LeFessant. Validity Conditions in Agreement Problems and Time Complexity. Technical Report 4526, INRIA, 2002.

[GHOS96] J. Gray, P. Helland, P. O'Neil, and D. Shasha. The Dangers of Replication and a Solution. In *Proc. of the SIGMOD*, pages 173–182, Montreal, 1996.

[GHR97] R. Gupta, J. Haritsa, and K. Ramamritham. Revisiting Commit Processing in Distributed Database Systems. In *Proc. of the ACM SIGMOD*, 1997.

[Gra78] J. Gray. *Notes on Database Operating Systems*. Springer, 1978.

[JPAA01] R. Jiménez-Peris, M. Patiño-Martínez, G. Alonso, and S. Arevalo. A Low-Latency Non-Blocking Atomic Commitment. In *Proc. of DISC. LNCS-2180*. Springer, 2001.

[JPAK02] R. Jiménez-Peris, M. Patiño-Martínez, G. Alonso, and B. Kemme. Scalable Database Replication Middleware. In *Proc. of 22nd IEEE ICDCS*, Vienna, Austria, July 2002.

[KA00a] B. Kemme and G. Alonso. Don't be lazy, be consistent: Postgres-R, A new way to implement Database Replication. In *Proc. of VLDB*, 2000.

[KA00b] B. Kemme and G. Alonso. A new approach to developing and implementing eager database replication protocols. *ACM TODS*, 25(3):333–379, September 2000.

[KD95] I. Keidar and D. Dolev. Increasing the Resilience of Atomic Commit at No Additional Cost. In *Proc. of ACM Principles of Database Systems*, 1995.

[KPAS99] B. Kemme, F. Pedone, G. Alonso, and A. Schiper. Processing Transactions over Optimistic Atomic Broadcast Protocols. In *Proc. of ICDCS*, 1999.

[KR01] I. Keidar and S. Rajsbaum. On the Cost of Fault-Tolerant Consensus When There Are No Faults - A Tutorial. Technical Report MIT-LCS-TR-821, MIT CS Lab, 2001.

[LKS91] E. Levy, H. F. Korth, and A. Silberschatz. An optimistic commit protocol for distributed transaction management. In *ACM SIGMOD Conf.*, pages 88–97, 1991.

[Ped01] F. Pedone. Boosting System Performance with Optimistic Distributed Protocols. *IEEE Computer*, pages 80–86, 2001.

[PGS] F. Pedone, R. Guerraoui, and A. Schiper. The Database State Machine Approach. *Journal of Distributed and Parallel Databases and Technology*. to appear.

[PJKA00] M. Patiño-Martínez, R. Jiménez-Peris, B. Kemme, and G. Alonso. Scalable Replication in Database Clusters. In *Proc. of DISC. LNCS-1914*. Springer, 2000.

[Ske82] D. Skeen. A Quorum-Based Commit Protocol. In *Proc. of the Works. on Distributed Data Management and Computer Networks*, pages 69–80, 1982.

[SPMO02] A. Sousa, J. Pereira, F. Moura, and R. Oliveira. Optimistic Total Order in Wide Area Networks. In *Proc. of SRDS*, 2002.

[VR02] P. Vicente and L. Rodrigues. An Indulgent Uniform Total Order Algorithm with Optimistic Delivery. In *Proc. of SRDS*, 2002.