

Distributed Software Transactional Memory

Abstract

The present paper describes the details of a distributed software transactional memory implementation. Solutions for 2 scenarios are presented:

- a) perfect links and processes*
- b) fault-tolerant environment with replication*

1. Introduction

Due to the cloud computing revolution in computing, today's world has increasingly moved to cloud-based services rather than maintaining their own infrastructure. The cloud service providers are well equipped with the hardware and software infrastructure to provide better services by applying concepts of *Distributed Computing*.

In this model of computing, one of the main problems that providers face is concurrent transactions; how the system behaves when multiple users access/manipulates an object at the same time; how to maintain the proper order of commit and how to resolve conflicting operations. This paper addresses such issues in detail and suggests possible solutions for implementation.

In the following sections the implementation of a distributed software transactional memory library is presented for two scenarios:

- a) perfect Master, Servers and links
- b) perfect Master, links and recoverable, sequentially

Section 2 introduces an overview of each scenario. Section 3 discusses the algorithms used for concurrency control, transaction management and Client-Server communication

2. Architecture

The assumption made for all the sections below is that the Master is a perfect process. The names Object Servers and Servers are used interchangeably and refer to any non-Master and non-Client process that participates in a transaction. The main differences/similarities between the two

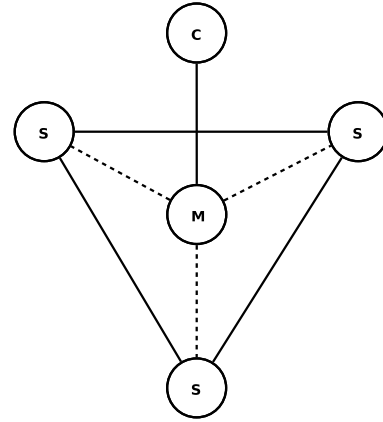


Figure 1: Master, Server and Client entities in a perfect setup

architectures, perfect and fault-tolerant, are the following:

- for perfect, the Coordinator of a transaction can be any Server, not necessarily the Master
- in the fault-tolerant scenario, replication of tentative objects and transactions is assured
- the Master also acts as a failure detector in both scenarios
- the Master and the Coordinator are the same process in the fault-tolerant scenario

Subsection 2.1 details the perfect architecture, while Subsection 2.2 lists the details of the fault-tolerant architecture. Further details on the responsibilities of the Coordinator/Master are in Subsection 2.3.

2.1. Perfect links and processes

As illustrated in Figure 1, the Master server (*M*) is used by the Object Servers (*S*) when bootstrapping. Information such as IP address and port number of the Server is stored by the Master which, then, issues unique identifiers for each Server. Once the Object Server (*S*) has been issued an identifier and notified the Master of joining the group of available servers, it can be connected by Clients (*C*) for handling transactions. Transaction identifiers (TID) should be unique to the whole system therefore the Master server (*M*) will keep track of the global sequence of TIDs.

Note: the number of Object Servers can vary.

Clients (C) will query the Master (M) if their cached list of servers is outdated or attempting to reach an Object Server (S) times out. The Master (M) provides the Client an updated list of the available Object Servers (S) to be used for establishing/handling transactions.

Objects that are going to be stored in Object Servers (S) have unique identifiers (integer) so that each object can be uniquely identified throughout the system. Therefore in this solution we are using the modulo function against the number of Object Servers (S) available in order to find the actual location of the object to be stored or referred.

The clients (C) can connect to multiple Object Servers inside a single transaction, hence it is required to have a Coordinator that manages the transaction operations such as commit and abort. The Coordinator can be any Server from the list provided by the Master in the current, perfect, scenario. On the other hand, in the fault-tolerant scenario, we consider the Master (M) as the coordinator. When a Client (C) needs to manipulate an object it connects to the relevant Object Server, which will contact the Coordinator (M) and join the transaction. Then the manipulation operations (Read/Write) are handled only by the Client and the Object Servers (S).

2.2. Fault Tolerant system

For recoverability in case of failure, an extension to the above system is necessary. Having a replication of the transaction's objects in a particular Object Server (S) across other Servers (S) is necessary to assure recoverability. The Master (M) handles the replica location of each Server. Our proposal for the replica locations is a conceptual ring, where S1's replica is stored in S2, and that of S2 in S3, etc. The Master notifies each Server of its replica location. The transaction model is similar to the phase one, the only extension required is to ask the replica server as well to join the coordinator for the transaction.

The Master server (M) also assumes the role of a *failure detector*, by using the heartbeat messages regularly exchanged between the Master and the Server. In case of an Object Server(S) failure, the Master server(M) will identify it through heartbeats and issue a notification to all the other Object Servers (S) about the failure. No further TIDs are issued for the clients until the failure is recovered. Once the current transactions complete, the Master will stabilize the system with the correct replica locations.

At the fault recovery, the Master server (M) will issue a notification to the Object Servers (S) asking to rearrange, so that the Object Servers (S) go through the list of objects which it is responsible and apply the modulo function again according to the new Object server count. All the Object Servers (S) process this action simultaneously and the

object movement among servers will stabilize the system. Once the system is stable, the Master (M) will issue another notification to the Object Servers to invalidate the current replica object set and replicate the new object set. After this replication step the system becomes stable and the Master Server (M) starts to issue TIDs again. Furthermore, if a new server is added to the system, exactly the same steps mentioned above will be followed by the system.

2.3. Coordinator/Master responsibilities

The Coordinator's responsibilities and interface are different for the two scenarios (perfect and fault-tolerant) detailed in Section 2. For instance, assuming perfect links and processes, the Coordinator can be any of the Servers, thus reducing the Master's load. The list of servers available to the Client is supplied by the Master and it's updated with every server group membership change.

The Coordinator's responsibilities below are shared in both scenarios:

- assign a transaction identifier to the Client
- handle abort and commit commands from the Client or Object Servers
- interface with the participant servers in the transaction during the voting and commit phases

The Master also assumes the role of a failure detector through the use of heartbeat messages. In the fault-tolerant environment, the Master also assumes the Coordinator role.

3. Algorithms

3.1. Communication

The following communication occurs **before** the Client attempts connection to the Master node - between the Master and the Servers.

1. Master boots, followed by the Object Servers bootstrapping
2. During Object Servers bootstrap:
 - Master identifies the Object Servers trying to connect
 - Master assigns a unique name to the Object Servers to later be used as a reference
3. Object Servers's heartbeat messages periodically being sent to Master. Heartbeats are also used for replica management in the fault-tolerant scenario. Should one of the replicas fail, the objects are replicated in another server in order to maintain a certain level of availability in the system.

Upon *attempting a transaction* the following communication exists between the Client, Master and Object Servers in a **fault-tolerant** environment:

1. Client connects to the Master and requests a transaction identifier
2. Master maintains a global transaction sequence and returns a unique transaction identifier (TID) to the Client. Furthermore, a list of healthy Object Servers is also returned to the Client.
3. Client generates a unique ID (integer) when generating the object and also be used in order to select an available server from the list (used in a modulo function)
4. Client uses the modulo result to connect to the Object Server and store the object representing the tentative versions of the transaction.

The main difference between the above steps and the communication in a **perfect** environment is that the TID is not necessarily assigned by the Master, but any of the Servers.

When the Client attempts to **manipulate an existing object** in the *perfect* environment, the following communication takes place:

1. Client requests and retrieves TID from Master
2. Client retrieves object's previously generated unique ID
3. Client applies modulo on the retrieved ID to locate the relevant Object Server to request for reference
4. With the retrieved object reference, client can manipulate the object value

3.2. Transaction Management and Concurrency Control

The *Flat Transaction* model allows for a client to manipulate objects on multiple servers in a single transaction. With regards to concurrency control, *Timestamp Ordering* will be used for this project. Figure 2 illustrates a flat transaction model setup with a transaction being executed from the client on 3 servers.

The advantages of choosing *Timestamp Ordering* over *Two-Phase Locking* or *Optimistic Concurrency* options are the following:

- Deadlock prevention - common with the use of locks
- Better performance for transactions with predominantly *read* operations [1]
- Faster conflict resolution when compared to locking - transactions are aborted immediately.

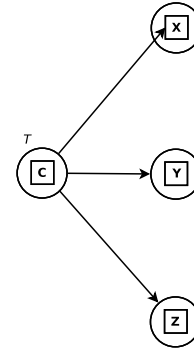


Figure 2: Flat Transaction model between a Client and 3 Servers

Figure 3 illustrates the steps between the Client, Coordinator, Master and Object Servers during a transaction. The steps below offer more details on the communication that occurs in a perfect setup:

- **Steps 3-4:** Client acquires transaction ID from the Coordinator
- **Step 5:** Client passes TID to the Object Server
- **Step 6:** Object Server calls the Coordinator's *interface*
- **Steps 7-8:** Client manipulates objects directly on the Object Server
- **Step 9:** Upon transaction end, Client asks the Coordinator to either *Abort* or *Commit*
- **Step 10:** Coordinator will request each participant server in the transaction to indicate whether it can commit a transaction or not. (*Voting Phase*)
- **Step 12:** If all participants answer/vote *positively*, the Coordinator issues *doCommit(TID)* for each participant to commit its part of the transaction
- **Steps 13-14:** Once completed, all servers acknowledge the commit and Coordinator notifies the Client that it's successful
- Should any of the participant servers be unable or disagree to commit and aborts, the Coordinator will request all the remaining participants to abort. Client will then be notified.

3.3. Timestamp Ordering and Deadlock Detection

Due to the use of timestamp ordering for concurrency control, there is limited probability of deadlock occurrence. Each transaction is assigned a unique timestamp on start. Each operation in a transaction is validated when it is carried out. Should a transaction fail the validation it is immediately aborted, but can be restarted by the Client. The Client is issued with a globally unique transaction timestamp by the Coordinator. The servers are jointly responsible for ensuring serial equivalence. For instance, server S1 access an object before S2, thus server S1 appears before S2 for all objects. The Coordinators must agree on times-

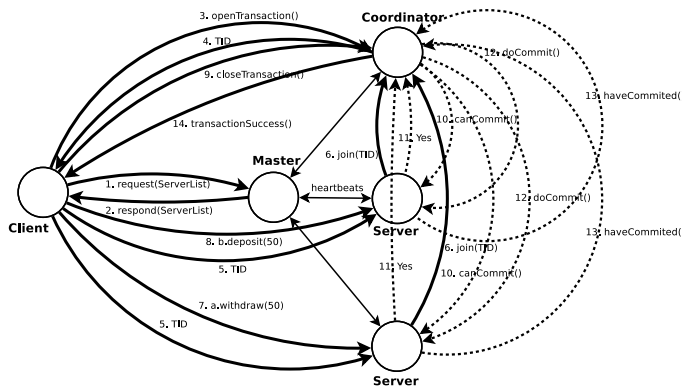


Figure 3: Transaction steps between the Client, Master, Coordinator and Servers. The dotted lines represent the 2-phase distributed commit

tamp ordering to maintain all servers synchronised with the same ordering. The timestamp consists of a (*local times-tamp*, *server-id*) pair and is kept synchronized by the use of local physical clocks coordinated by the Master.

Conflict resolution is performed at each operation. The basic timestamp ordering rule is based on operation conflicts and is very simple: *A transaction's request to write an object is valid only if that object was last read and written by earlier transactions. A transaction's request to read an object is valid only if that object was last written by an earlier transaction.*[2]. Each transaction has its own tentative version of each object it accesses, such that multiple concurrent transactions can access the same object. The tentative versions of each object are committed in the order determined by the timestamps of their transactions by transactions waiting, if necessary, for earlier transactions to complete their writes. Since transactions only wait for earlier ones (thus, no cycle occurs in the wait-for graph), no deadlocks occur.

Conflict resolution is performed at each operation. If the resolution is to abort the transaction the Coordinator is informed and it will abort the transaction for all the participants.

3.4. Replication

The fault tolerant design supports failure of a single Object Server. It extends the perfect architecture by assuring the replication of objects in the adjacent server. For instance, the object stored in Object Server S1 keeps a copy in S2 as a replica. The Master server is responsible for keeping track of the replica servers.

When an Object Server **fails**:

1. Master detects the failure from the absence of heart-beat messages from the server

2. Master server issues a “*failure warning notification*” to the other servers, so that the current transactions related to the absent server be aborted. Others will continue but new transactions will not be accepted. (Master no longer issuing TIDs)
3. Master coordinates the object replication so that the currently available servers share the objects held by the absent server.
4. Once the Master notifies all the servers about the new, stable status, transactions can be take place again.

When the failed server **recovers**:

1. Master detects the arrival of a new server
2. Master server issues a “*New server notification*” for the servers. Once the current pending transactions are completed, servers are stabilized
3. Master server coordinates object replication such that a portion of objects in the current system will be transferred to the new server
4. Once the Master notifies all servers of the stabilized status, transactions can be resumed

4. Conclusion

Across the sections we discussed the solutions proposed in order to implement Distributed Software Transactional Memory System. Our solution addresses two scenarios: perfect links and processes and a fault tolerant system. A very important goal has been to provide a simple solution which would eliminate conflicts in transactions without encountering the inefficiency of a distributed environment. The sections and subsections address each of the two architectures individually in order to provide a clearer distinction of responsibilities and complexity of the system. Furthermore, *Timestamp Ordering* based concurrency control aids in eliminating local and global deadlocks which represents one of the major issues that need to be addressed in DSTMs.

Finally, we believe that the concepts we came across in the discussion of this paper builds a firm foundation for implementing this solution in an efficient manner.

References

- [1] P. A. Bernstein, V. Hadzilacos, and N. Goodman. *Concurrency control and recovery in database systems*, volume 370. Addison-wesley New York, 1987.
- [2] G. F. Coulouris, J. Dollimore, and T. Kindberg. *Distributed systems: concepts and design*. pearson education, 2005.