A Recovery Protocol for Middleware Replicated Databases Providing GSI

J. E. Armendáriz-Iñigo, F. D. Muñoz-Escoí
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Ciudad Politécnica de la Innovación
Edificio 8G - Escalera 3, Planta 2ª
Camino de Vera s/n
46022 Valencia, Spain
{armendariz, fmunyoz}@iti.upv.es

J. R. Juárez-Rodríguez, J. R. González de Mendívil
Dpto. de Matemática e Informática
Universidad Pública de Navarra
Campus de Arrosadía s/n
31006 Pamplona, Spain
{jr.juarez, mendivil}@unavarra.es

Technical Report ITI-ITE-06/10

A Recovery Protocol for Middleware Replicated Databases Providing GSI

J.E. Armendáriz-Iñigo[1], F.D. Muñoz-Escoí[1],
J.R. Juárez-Rodríguez[2], J.R. González de Mendívil[2]
Technical Report ITI-ITE-06/10

[1] Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera s/n
46022 Valencia, Spain
Ph./Fax: (+34) 96 387 72 37 / 72 39
{armendariz, fmunyoz}@iti.upv.es

[2] Dpto. Matemática e Informática
Universidad Pública de Navarra
Campus de Arrosadía s/n
31006 Pamplona, Spain
Ph./Fax: (+34) 948 16 90 93 / 95 21
{jr.juarez, mendivil}@unavarra.es

November 14, 2006

Abstract

Middleware database replication is a way to increase availability and afford site failures for dynamic content websites. There are several replication protocols that ensure data consistency for these systems. The most attractive ones are those providing Generalized Snapshot Isolation (GSI), as read-operation never blocks. These replication protocols are based on the certification process, however, up to our knowledge, they do not cope with the recovery of a replica. In this paper we propose a recovery protocol that ensures GSI (we provide an outline of its correctness) that does not interfere with user transactions and permits the execution of transactions in the recovering node, even though the recovery process has not finished.

1 Introduction

Web servers usually provides dynamic content that is persistently stored in a DBMS. It is well-known that the replication of the database in different replicas geographically distributed improves data availability and scaling performance. On the other hand, data consistency is sacrificed. Due to this, several data consistency criteria has been introduced: One-Copy Serializable (1CS) [5], and Generalized Snapshot Isolation (GSI) [10], among others [9, 18, 21]. 1CS presents some drawbacks for dynamic content web page generation such as read operations may become blocked, which are most of the operations for web commerce application such as TPC-W standard states [22]. As a consequence of this and because of most of commercial and open-source databases provide Snapshot Isolation (SI) [4], the GSI correctness criterion is used.

GSI states that a transaction does not necessarily need to observe the "latest" snapshot, as opposite to SI in the centralized environment. It can observe an older snapshot, and many properties as those in (centralized) snapshot isolation continue to hold. In other words, transaction will see a consistent a snapshot and they will not become blocked, but at the price of increasing the abort rate as updates may collide with other more "recent" transactions. This is not a major drawback since in TPC-W most if operations are read-only (at most 50% in the *Ordering mix* with a negligible conflict rate) [22]. This is of high importance from the point of view of replica failure and its recovery (or even for joining new replicas).

The main concern of recovery techniques is to penalize as little as possible ongoing transactions (i.e. aborting transactions during the recovery process) and transferring the data in a manner that the new joining replica may become available as soon as possible (i.e. issuing new incoming user transactions on it too). This last point is where GSI comes in handy when dealing with the recovery process. The rejoining replica will have a (probably older) consistent snapshot and may accept user transactions as soon as the replica is available.

Replication protocols developed for this GSI consistency level [2, 8, 10, 18, 19, 24], to the best of our knowledge, lack of recovery solutions in case of a rejoining node. Some hints are outlined in [10] but there is not a formal presentation of the solution. In this paper we take the recovery ideas from [10] and the recovery protocol presented in [15] both taking advantage of the facilities provided by a Group Communication System (GCS) [6], mainly strong virtual synchrony [11]). Most of these replication protocols already proposed are to be used in middleware architectures, due to its portability among several DBMSs and the standard interface they offer to user applications (e.g. JDBC) so they do not have to modify their application from a replicated setting. In [15] a recovery protocol for a 1CS replication protocol [20] is introduced. This replication protocol presents the inconvenience that the application programmer is forced to pre-declare the structure of each transaction or to send transactions as full blocks [20]. This will not be the case for our solution, we only take the approach they follow in the recovery data transfer: how it is started, how it goes and how it finishes. In this paper, we present a GSI recovery protocol that works with any of the GSI replication protocols already proposed [2, 18, 10, 19]. Thus, we may obtain that ongoing transactions on previously alive replicas do not need to be aborted, they can continue working as normal and the transaction load can be balanced to the new replica as soon the GCS states it is reachable from the communications point of view. Furthermore, for the rejoining node the recovery data transfer is nothing more than a "delayed" propagation of writesets.

The rest of this paper is organized as follows. Section 2 is devoted to explain the main characteristics of replication protocols providing GSI that could use our recovery proposal. Section 3 introduces the middleware architecture used in this work. The recovery protocol is described, along an outline of its correctness, in Section 4. Some optimizations are shown in Section 5, it includes a variation of the recovery protocol for transferring the whole database to a new replica. Some discussion with previous related works are outlined in Section 6. Finally, conclusions end the paper.

2 GSI and the Certification Process

The replication protocols supporting GSI use a certification process [10] for committing a transaction in the system, exchanging only one message per transaction; i.e. they are of constant interaction according to [23]. Nevertheless, read-only transactions will directly commit. They follow the Read One Write All Available (ROWAA) approach.

Each database replica persistently stores the current snapshot version they hold. A transaction T is firstly executed at its master replica (it obtains the current local snapshot version, $T.start$)), the rest of replicas enter in the context of the transaction when an update transaction requests to commit. The protocol collects the updates performed by T in the local database (the writeset of a transaction, $T.WS$) and is sent to the rest of replicas using the total order multicast facility [6]. Upon delivery of this message at each available replica the protocol locally performs a test to decide if T can commit or must abort. It checks whether $T.WS$ intersects with the writeset of any update transaction that committed after $T.start$. If all intersections are empty, T will commit; otherwise, it will aborts. It is important to note that all sites reach the same decision, since all writeset are delivered in the same order to all replicas. Hence, from this previous explanation each database replica k must persistently store, apart from its version number ($Version_k$) , a Log_k that is a set of $\langle \text{snapshot version}, WS \rangle$ tuples.

3 System Model

In this work we took the advantage from our previous works [14] and other middleware architectures providing database replication [18, 20]. On the top of Figure 1 relies the user application that uses a JDBC driver interface to issue transactions in the system. This driver is able to connect to a given replica and in case of failure to redirect the transaction to another available replica. We use a full replicated approach so that each underlying DBMS has a copy of the replicated database and provides SI. The middleware contains a replication (and recovery) protocol module that is in charge of maintaining consistency and communicate with the rest of replicas using a GCS.

While developing replication and recovery protocols, it is a must not to re-implement features provided by the underlying DBMS. The DBMS performs these tasks much more efficiently. Therefore, the database replication middleware architecture (see Figure 1) must follow the "*gray box approach*". The writeset must be efficiently picked up from the DBMS. The writeset contains the updated/inserted/removed tuples identified through the primary key. Furthermore, applying certified remote writesets may become aborted due to deadlocks with local transactions. This can be circumvented by way of reattempting the application of writesets proposed in [18], or it can be used a conflict detection mechanism [19]. This last technique uses the concurrency control support of the underlying DBMS. Thereby, the middleware is enabled to provide a row-level control (as opposed to the usual coarse-grained table control), while all transactions (even those associated to remote write sets) are

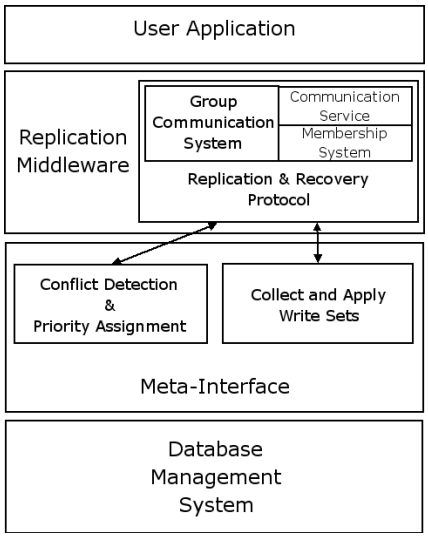

Figure 1: An example of a replica inside a database middleware architecture

subject to the underlying concurrency control support. It periodically looks up blocked transactions in the DBMS metadata (e.g., in the pg_locks view of the PostgreSQL system catalog). It returns a set of pairs consisting of the identifiers of the blocked and blocking transactions.

3.1 About the Group Communication System

A GCS provides a communication and a membership service (see Figure 1), supporting virtual synchrony [6]. It is assumed a partially synchronous system and a *partial amnesia crash* [7] failure model. We consider this kind of failures as we want to deal with node recovery after its failure. The communication service features a total order multicast for message exchange among nodes through reliable channels. Membership services provide the notion of view (current connected and active nodes with a unique view identifier, $\mathcal{V} = \langle id, nodes \rangle$). Changes in the composition of a view (addition or deletion) are delivered to the recovery protocol. We assume a primary component membership [6]. In a primary component membership, views installed by all nodes are totally ordered (there are no concurrent views), and for every pair of consecutive views there is at least one process that remains operational in both views. The GCS groups messages delivered in views [6]. We assume a *strong* virtual synchrony [11], where the view change occurs at two stages, first it stops sending multicast messages and keeps processing previous multicast messages; and once the block process is done, it will deliver the view change event. The uniform reliable multicast facility [12] ensures that if a multicast message is delivered by a node (faulty or not) then it will be delivered to all available node in that view. All these characteristics permit us to know which writesets have been applied in the context of an installed view.

3.2 Assigning Priorities to Transactions

The conflict detection scheme is combined with a transaction priority scheme in the replication protocol [19]. For instance, we might define two priority classes, with values 0 (assigned to local transactions that have not started their commit phase) and 1 (for those local transactions that have started their commit phase and also for those transactions associated to delivered write sets that have to be locally applied (either remote or recovery ones). By default, it aborts the transaction with the smallest priority but takes no action if both transactions have the same priority level. The replication (respective recovery) protocol will take the appropriate action, e.g. aborting the local transaction against a remote (recovery) transaction since the total order of the messages ensures that this transaction is going to be finally aborted.

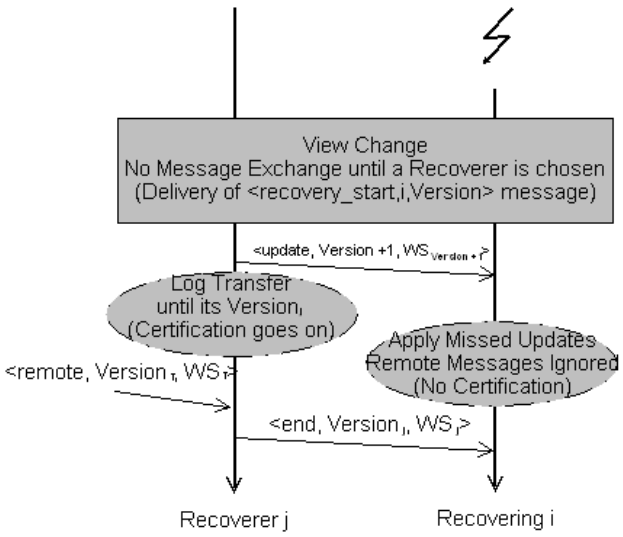

Figure 2: Example of the data transfer between the recoverer replica j and the recovering replica i

4 Recovery Protocol

Protocol Description. As a general overview of the main goal of our recovery protocol, let us say that one node ($recoverer$) will transfer the missed writesets to the recovering node arranged by their respective versions. This means that user application transactions executed on the recovering node will run under GSI in a "slower" replica. As it may be seen there no restrictions to execute user transactions in the replica and transactions executing at other replicas will behave as it nothing happens in the system. To achieve this we take the ideas outlined in [15, 10].

A recovering replica i joins the group (see Figure 2), triggering a view change. As part of this procedure, the recovering protocol instance running in i multicasts a message indicating the $\mathsf{Version}_i$ of the last applied writeset, no message activity is done until this message is delivered. This means that all messages uniformly delivered in the previous view are delivered to all nodes that install the next view. Furthermore, we make no assumption about if these pending messages have been certified at these nodes nor applied. As far as the recovery protocol know, all messages delivered in the old view have been delivered to all available nodes transiting to the new one.

In parallel to this, a procedure takes place to choose a recoverer replica. It is also assumed there is at least one replica with the latest snapshot version in the system. This does not imply that the chosen recoverer has to be this node. This is to emphasize the behavior of the recovery process, we allow a replica to act as the recoverer even though it has not processed all the updates it has pending to apply. Several optimizations can be included during the recovery process (e.g. the selection of multiple recoverer replicas); but in order to keep the protocol description simpler we have included them in Section 5. The recoverer replica (j) starts a recovery thread that sends point-to-point messages starting from $\mathsf{Version}_i + 1$.

Meanwhile, the recovering replica ignores all messages delivered from the total order multicast in that view. In other words, no certification process is performed in the recovering node but those writesets coming from the recoverer replica that are directly applied in the database. This is a better approach than storing in a separate queue total order delivered messages, since they must pass a certification process (and some of the messages will make no sense since their associated writesets have to be finally aborted) which has already been done by the recoverer. Thus, we have to be able to define the point of stopping the recovery process at the recoverer replica.

At some point, assuming that the recovery data transfer is faster than applying writesets, the recoverer reaches its current snapshot version for transferring data. The recoverer replica sends its $\mathsf{Version}_j$ with a flag (see Figures 2 and 3, the $\langle \mathsf{end}, \mathsf{Version}_j, \mathsf{WS}_j \rangle$).When the message is processed at the recovering replica it will multicast (using the total order primitive) a $\langle \mathsf{start\_listening}, i \rangle$ message. The recoverer will store all certified transactions between the $\langle \mathsf{end}, \mathsf{Version}_j, \mathsf{WS}_j$ message has been sent and the delivery of the $\langle \mathsf{start\_listening}, i \rangle$ message. The recovering replica will listen to total order messages, however it will discard every message, until its own $\langle \mathsf{start\_listening}, i \rangle$ message is delivered; after that point, it will enqueue every remote message delivered ($\langle \mathsf{remote}, \mathsf{Version}_\mathsf{T}, \mathsf{WS}_\mathsf{T} \rangle$ in Figure 2 and 3). It will wait for the remaining Log data transfer.

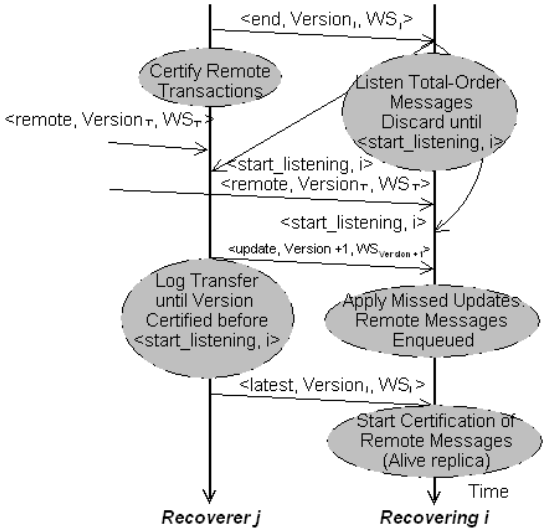

Figure 3: Finalization of the recovery process at recovering replica i

Hence, the recovering replica is alive and is able to apply the enqueued update message. The protocol is described in Figure 4 for the rejoining and leaving process and its state variables are shown in Table 1 respectively. They are described in terms of the procedures executed with every message and event of interest. Since they use shared variables, we assume each one of these procedures is executed in one atomic step. It is important to note that we present the algorithm for a single replica recovery just to simplify its presentation. Its extension to multiple replica recovery is straightforward, the recoverer will have as many recovery threads as nodes being recovered and must individually monitor the delivery of the $\langle \mathsf{start\_listening}, k \rangle$ messages delivery, with k the replica identifier of an element in the set of recovering replicas.

$status$	It contains the state of the replica in the system: crashed, alive, recovering and recoverer.
$curr\_\mathcal{V}$	The current installed view, a $\langle$id, replicas$\rangle$ tuple.
$pre\_\mathcal{V}$	The previous installed view, a $\langle$id, replicas$\rangle$ tuple.
Recoverer	The recoverer replica identifier.
Log	a set of $\langle$snapshot version, WS$\rangle$ tuples.
Version	The current snapshot version.
CurrentVersion	The present snapshot being transferred to the recovering replica.
LatestVersion	The last snapshot version to be transferred by the recoverer, initially 0.

Table 1: State variables kept by each replica i and their description

Replica Failure. Whenever a site failure occurs, a view change event is fired, all GCS activity is stopped. Replicas that are about to instal the new view process those writesets pending to apply (due to the uniform delivery of messages) and will not install this view until the precious process start. Hence, all available replicas have a consistent state to continue processing new incoming transactions, i.e. they can use the GCS to multicast and deliver messages. It is important to note that in case of a failure of a recoverer node, the recovering node has to "restart" the recovering process by sending the latest version it has applied (recall we assume that there is at least one node in the new partition with all the versions installed). Whereas a recovering replica failure only implies the interruption of the recovering thread at the recoverer node.

Replica Recovery. At the time a replica k rejoins the group of replicas, a view change event is fired. As in the previous case of a replica failure, all message exchange is suspended until all writesets have been applied. Along with the new installed view a message containing the $\mathsf{Version}_k$ is sent. One function is used to univocally determine among all previous available nodes one site to act as the recoverer replica. This function could implement many different load balancing and recovery policies thereby making it arbitrarily complicated. For simplicity, we use the simplest possible version of this function: a

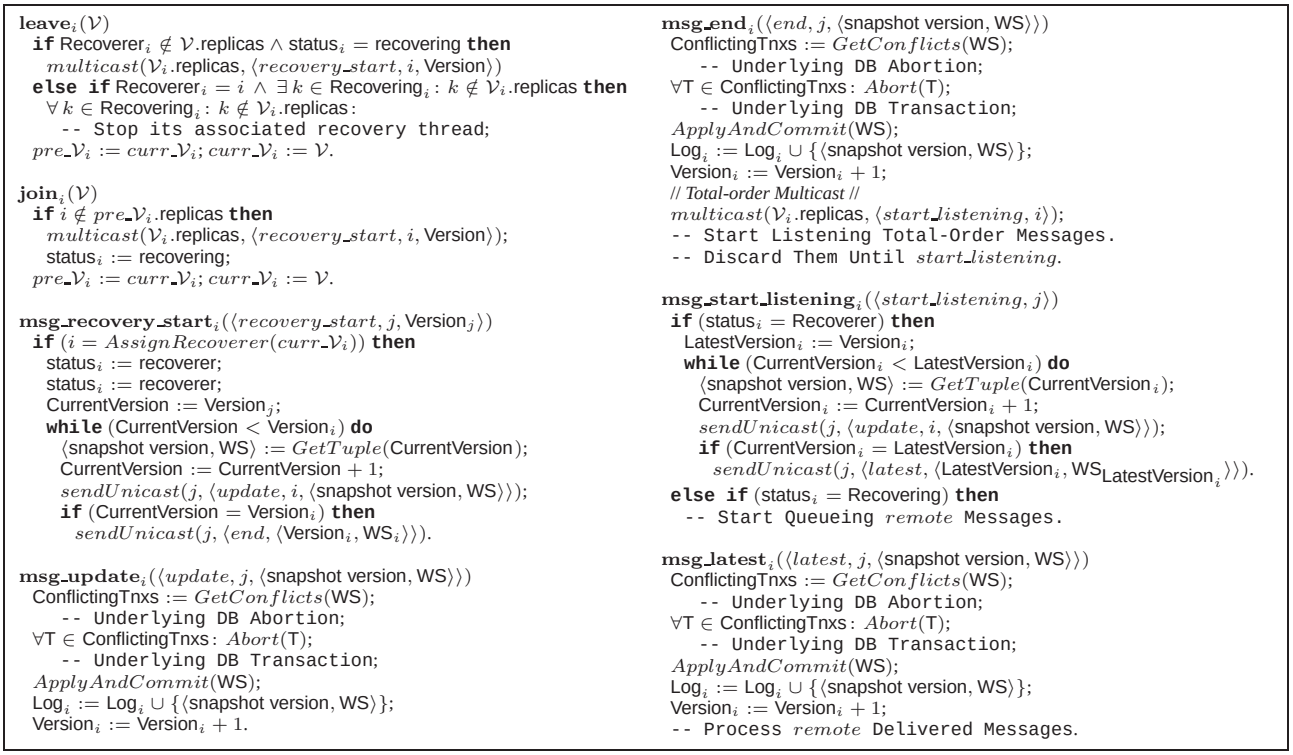

Figure 4: Specific recovery protocols action and message events executed at replica i

single recoverer that recovers data partitions sequentially, and ignores any load balancing issues.

Ongoing transactions in previously available nodes behave as normal. Furthermore, new user transactions are allowed to execute in the recovering replica. This last assertion is of key importance, since read-only transactions will be committed as in any other replica and update transactions have to pass through the certification process and will remain blocked. It is important to note that during the recovery process, only the writesets coming from the recoverer are processed. This implies that user transactions trying to update items belonging to a given writeset will be rolled back. Moreover, at the end of the recovery process there will not be any update user transaction blocked. Those that pass the certification process in the remainder replica will come in a recovery data transfer and those that failed will be rolled back by applying a missed writeset.

Recovery Thread. A recovery thread looks for the last snapshot version of the recovering replica. We can assume that if there are several nodes recovering, the recoverer will look for the lowest common last snapshot version for all recovering replicas and starts multicasting updates (using uniform reliable service) from the Log_k from that version on. The recovering nodes discard those recovery messages coming from snapshots they already got. We assume that read and write operations performed by the recovery and replication protocols in the persistent storage are realized in mutual exclusion.

Transferring Missing Updates. The data transfer to the recovering node flows as depicted in the outline of the recovery protocol. The recoverer sends several $msg_update(\langle$snapshot version, WS$\rangle)$ tuples are applied one after the other according to its snapshot version. Each tuple passes the certification phase and the respective $\mathsf{Version}_k$ is increased. Of course, those changes contained in the delivered writeset have to be applied in the underlying database. Before the writeset is applied by way of a recovery transaction (by the $ApplyAndCommit$(WS) function), the recovery protocol must abort those conflicting transactions being executed at the recovering replica, the mechanism used here is the same as the one used in [19] and highlighted in Section 3. Recall that in GSI only write operations do conflict, these local transactions are going to be finally aborted since there are more recent transactions, from the snapshot version point of view, than they are. The process of recovery is continued until the recoverer send the $\langle end, \mathsf{Version}_j, \mathsf{WS}_j\rangle)$ to recovering replica. This version number is stored in $\mathsf{CurrentVersion}_j$ for determining the remainder Log_j transfer.

Finishing the Missed Data Transfer Process. When the previous message is received at the recovering replica, it will multicast (total order primitive) the $start\_listening, k$. This message will be silently discarded at all replicas but the recoverer one. The delivery of this message marks the end of its Log_j data transfer, as the recovery protocol reads the current

$\mathsf{Version}_j$ and assigns it to the variable $\mathsf{LatestVersion}_j$. This second data transfer will comprise from version $\mathsf{CurrentVersion}_j$ to $\mathsf{LatestVersion}_j$ of the Log_j. The data transfer is identically to the first one. However, the last message is flagged as $latest$ just to emphasize, that there is no more pending updates to be transferred. Thus, the recovering replica is alive.

Queueing of Remote Messages. In parallel to the previous paragraph, the recovering replica will start listening from incoming messages of the total order service, although all of them will be discarded till the $\langle start\_listening, k\rangle$ message is delivered. Once it is delivered, it will queue $\langle \mathsf{remote}, \mathsf{Version}_\mathsf{T}, \mathsf{WS}_\mathsf{T}\rangle$ messages (recall from Figures 2 and 3)coming from the GCS. Once all missed updates of the second phase are applied, it will start processing these messages.

4.1 Outline of its Correctness Proof

We make a basic assumption about the system's behavior: there is always a primary partition and at least one replica with all the versions installed transits from one view to the next one.

Lemma 1 (Absence of Lost Updates in Executions without View Changes)**.** *If no failures and view changes occur during the recovery procedure, and the recovery procedure is executed to completion, a node $j \in N$ can resume transaction processing in that partition without missing any update.*

Proof (Outline). Let us denote $\{r_{\nu+1}, r_{\nu_+2}, \ldots, r_{V_j}, \ldots, r_{LV}\}$ as the set of recovery transactions exclusively executed at the recovering replica associated to each set of tuples that must be applied. Thus, the set of tuples to be delivered are: $\{\langle \nu + 1, \mathsf{WS}_{\nu+1}\rangle, \langle \nu + 2, \mathsf{WS}_{\nu+2}\rangle, \ldots, \langle V_j - 1, \mathsf{WS}_{V_j-1}\rangle, \langle V_j, \mathsf{WS}_{V_j}\rangle, \ldots, \langle LV - 1, \mathsf{WS}_{LV-1}\rangle, \langle LV, \mathsf{WS}_{LV}\rangle\}$. In the same way, let us denote $\{t_b, \ldots, t_f\}$ as the set of concurrent committed transactions during the recovery process, assume they are ordered by the way they are inserted in the Log. The values t_b and t_f stand for the begin and the end of the recovery process.

Recovery transactions are sequentially executed at the recovering replica i by what it is stored at the Log_j of the recoverer replica j. Concurrent executing transactions are certified by the replication protocol at the rest of replicas. Hence, they are appended into the Log_j that must be transferred, more precisely the interval of $[t_b, t_{V_j}]$, where $b \leq V_j$, can be already transferred to the recovering node, i.e. may be the recovery process is so fast that all concurrent transactions have been applied or so slow that no transaction has been certified at the recoverer node. In any case, the insertion and application of these transactions is determined by the way they are inserted in the Log_j. The rest of the concurrent transactions ($[V_j + 1, t_f]$) will be applied in the phase just after the total order $\langle start\_listening, i\rangle$ message exchange to determine the finalization of the recovery process. Concurrent transactions from $[t_f + 1, \ldots)$ will be certified by the replication in the recovering node as this has finally achieved the recovery of its missed data. □

Lemma 2 (Absence of Lost Updates after a View Change)**.** *A recovering replica i in view $\mathcal{V}_i$ that transits to view $\mathcal{V}_i + 1$ resumes the recovery process without missing any update.*

Proof (Outline). We have to consider the cases when the recoverer node fails. Otherwise, the recovery process remains unaffected. If the recoverer fails, it will force a "rejoin" to tell its latest recovered $\mathsf{Version}_i$ so that a recovery process will take place for it. Hence, no updates will be missed since we will be under the circumstances of Lemma 1. Assuming that the time length of installed views is stable enough to eventually achieve the recovery of a replica without continuously failing recoverers. □

Theorem 1 (GSI Recovery)**.** *Upon successful completion of the recovery procedure, a recovering replica reflects a state compatible with the GSI execution that took place.*

Proof (Outline). According to [9] it is sufficient to show that if a given replication (or recovery) protocol using SI replicas provides global atomicity and commits update transactions in the same order at all replicas it provides GSI.

To prove that this implementation is deterministic and obeys GSI rules, we need to show two properties. The first property is that at the certification of t, all replicas have the same Log and Version. From the replication protocol point of view, as we are assuming a replication protocol that ensures GSI [2, 10, 18, 19], the concurrent committed transactions during the recovery process are applied in the same order at all alive replicas. On the other hand, by Lemmas 1 and 2, we have shown that the recovery does not produce lost updates and missing updates in the recovering replica (nor at any other available node, since the certification process remains the same at the rest of replicas) are applied in the same order they are committed. □

5 Optimizations

Several hints and optimizations are included in [10, 15]. The first one is the *garbage collection* of the Log. As it can be inferred the Log may become incredible large and, therefore, difficult to manage. There can be a thread at each replica k that periodically multicasts its current version $\mathsf{Version}_k$ and listen for incoming versions. Their respective Log_k can be trimmed to the minimum version collected from all (available or not) replicas. Of course, there exists a trade-off between addition of new replicas (e.g. due to a high workload) and the garbage collection since it must be omitted or, otherwise, some replica musts be stopped to transfer in the background, its current state to the new replica; afterwards, both issue the beginning of a recovery process for both replicas.

Another improvement will be to consider the distribution of the recovery process among several replicas as there is a penalization in the performance of the recovering replica. This can be done by appointing different recoverers for different partitions and/or by using several recovery threads at the recoverer. However, as they may differ in their Version, the function that determines the recoverers must know the latest version in the system. This can be included in the initial message of the recovery process in conjunction with the latest committed version in the recovering replica. The recovery process may be uniformly distributed, with the proper range of version intervals. However, special attention has to be paid during failures in the data recovery process (specially to recoverer replicas) and possible out of order delivery of $msg\_update(\langle \mathsf{snapshot\ version}, \mathsf{WS} \rangle)$ messages. On the other hand, there may exist a grouping process in the recovery thread so that several missed tuples of the Log can be sent in a single message during the missed data transfer.

The last kind of optimization consists in marking transactions (as read-only or update ones). This is specially useful in replicas that belong to a minority partition and have been forced to shutdown due to a network partition such as WAN environments. As we are ensuring GSI we may allow the execution of read-only transactions to applications even though they are "offline". This can be an interesting approach since most of the operations in TPC-W are read statements.

5.1 Adding New Replicas

Up to know we have coped with what can be considered as "short length" failures (due to its duration in time or the number of updated items). If we want to add a new replica with this recovery protocol, no garbage collection can be done. We have to store all changes done from the database initial version. Let us see a rough outline of how we can re-arrange this issue. When a new replica (or an old one that has been crashed for a long period time) joins the system, the recovery process starts in the same way as the previous replication protocol. Instead of transferring the Log_j of the recoverer j we start a read-only transaction (non-blocking) with the current snapshot version $\mathsf{Version}_j$. All the tables are transferred to the recovering (or joining) replica in a similar manner as the missed data transfer of the recovering protocol. Once all tables are transferred it continue with the Log_j transfer of the missed updates as with our original proposal. This can be best seen in Figure 5.

6 Related Works

There are several works in the literature that propose several alternatives to accomplish database recovery, mainly due to the use of GCS for an efficient way to provide database replication (i.e. total order of message deliveries, virtual synchrony, etc.). Most of the replication protocols that have been proposed are best suited for 1CS replication protocols [3, 13, 15, 17] while only hints (up to our knowledge) about recovery procedures for GSI replication protocols [10]. Thus, our work presents as it is presents a novelty approach because it pretends to work using GSI as its default consistency level.

In [17] several solutions to online reconfiguration in replicated databases are proposed making use of view synchrony and enriched view synchrony properties; they do not pretend to present one as the best among the different solutions. In [13], three recovery protocols have been also proposed using the virtual synchrony properties for replication protocols relying on total order message delivery for guaranteeing 1CS [1, 16]. These recovery techniques are based in the use of a *Log* that varies from blocking the system for transferring the missed data from the *Log* to relaxing the blocking need of the system and monitor the state of ongoing transactions during a view change that may include resend the writeset of these transactions during the recovery process in their commit message. In our work we take the advantage of the *Log* used for the certification process in order to determine the data to be transferred, however we do not need to perform additional tasks on ongoing transactions nor blocking the system. Furthermore, we permit the execution of transactions in the recovering node.

We have based our recovery proposal in the ideas proposed in [15]. Thus, its recovery ideas are pretty similar to the ones presented in this paper. However, we want to emphasize the differences. The most important is that the database is split into partitions, such as stored procedures of a web page. Ongoing transactions are never blocked by the execution of the

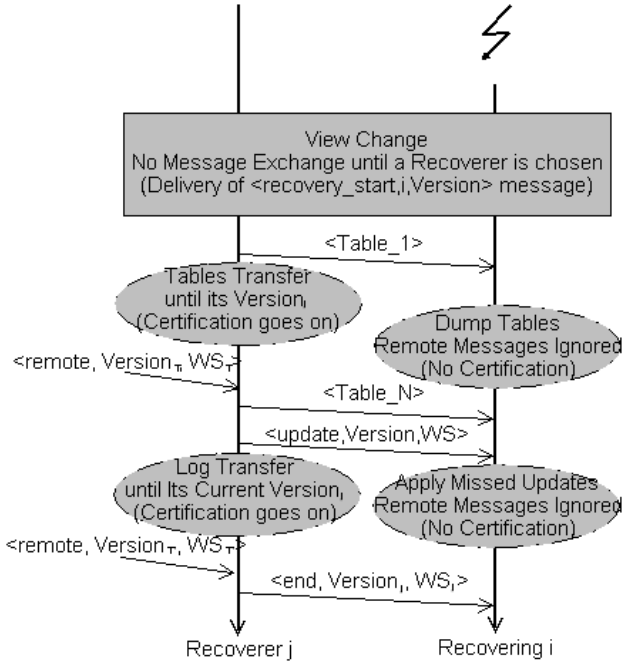

Figure 5: Finalization of the recovery process at recovering replica i

recovery process. This protocol ensures the maintenance of the 1CS due to the construction of database partitions that orders the execution of conflicting transactions due to the total order message delivery. Our solution is not restricted to certain patterns of transactions, the application is free to issue any kind of SQL statements. Moreover, transactions may be issued at any replica, including the recovering one. This is obtained thanks to the GSI consistency level we are offering.

Finally, we want to compare our solution to our previous work [3] that ensures the recovery for 1CS using the properties of uniform delivery [12] of messages combined with the virtual synchrony. These facilities provide an easy way to feature node recovery as it is possible to group the updates missed by a faulty node by the installed view where they happened. This information is stored as recovery metadata in the database. Once a node recovers from a failure, it is established a set of partitions at all available nodes (as many as views missed by the recovering node). Thus, those available nodes that are neither recoverer nor recovering nodes will access these partitions as usual. However they will get blocked when they propagate their updates and they conflict with a recovery partition at the recovering or the recoverer nodes. The recovering node will not be able to access these objects. The recovering node may start accepting user transactions as soon as the partitions are set up on it, even though it is not up to date. We overcome the limitations of blocking update transactions and read-only transactions in the recovering node (we proposed to run these transactions in SI). Moreover, we permit that transactions to be immediately scheduled in the recovering node.

7 Conclusions

In this paper we have presented a middleware database recovery protocol whose main novelty is its non-blocking property for ongoing transactions even at the recovering node (except those updates transactions but only at commit time. This protocol provides GSI [10] in comparison with previous solutions that were suited for 1CS [3, 13, 15, 17]. Furthermore, we have outlined its correctness for providing GSI. Another feature of this protocol is that it proposes two alternatives for achieving data state transfer, either the whole database or only the missed updates. This can be managed by the system administrator, depending on the kind of requirements of the application used.

This recovery protocol does not need any additional metadata from the ones already needed for the certified replication protocols providing GSI [2, 10, 18, 19]. The rejoining replica will send its current snapshot version and one node will be elected as the recoverer replica that will start a recovering thread that transfers data from the Log in the background to

the recovering replica. This permits that the replication protocol to remain unaffected. Finally, we have proposed several optimizations somehow outlined in [10, 15] and some other novelty optimizations.

Acknowledgments

This work has been supported by the Spanish Government under research grant TIN2006-14738-C02. The authors would also like to thank Prof. B. Kemme for her suggestions in the design of the protocol.

References

[1] D. Agrawal, G. Alonso, A. El Abbadi, and I. Stanoi. Exploiting atomic broadcast in replicated databases. *LNCS*, 1300:496–503, 1997.

[2] José Enrique Armendáriz-Íñigo, José Ramón Juárez-Rodríguez, José Ramón González de Mendívil, and Francesc D. Muñoz-Escoí. *k*-Bound GSI: A flexible database replication protocol. In *SAC*. ACM Press, 2007. *Accepted for publication*.

[3] José Enrique Armendáriz-Íñigo, Francesc D. Muñoz-Escoí, Hendrik Decker, José Ramón Juárez-Rodríguez, and José Ramón González de Mendívil. A protocol for reconciling recovery and high-availability in replicated databases. In *21st International Symposium on Computer and Information Sciences (ISCIS'06)*, volume 4263 of *LNCS*, pages 634–644. Springer, 2006.

[4] Hal Berenson, Philip A. Bernstein, Jim Gray, Jim Melton, Elizabeth J. O'Neil, and Patrick E. O'Neil. A critique of ANSI SQL isolation levels. In *SIGMOD Conference*, pages 1–10. ACM Press, 1995.

[5] Philip A. Bernstein, Vassos Hadzilacos, and Nathan Goodman. *Concurrency Control and Recovery in Database Systems*. Addison Wesley, 1987.

[6] Gregory Chockler, Idit Keidar, and Roman Vitenberg. Group communication specifications: a comprehensive study. *ACM Comput. Surv.*, 33(4):427–469, 2001.

[7] Flaviu Cristian. Understanding fault-tolerant distributed systems. *Commun. ACM*, 34(2):56–78, 1991.

[8] Khuzaima Daudjee and Kenneth Salem. Lazy database replication with snapshot isolation. In *VLDB*, pages 715–726. ACM, 2006.

[9] J. R. González de Mendívil, J. E. Armendáriz, J. R. Garitagoitia, L. Irún, and F. D. Muñoz. Non-blocking ROWA Protocols Implement GSI Using SI Replicas. Technical Report ITI-ITE-06/04, ITI, 2006.

[10] Sameh Elnikety, Fernando Pedone, and Willy Zwaenopoel. Database replication using generalized snapshot isolation. In *SRDS*. IEEE-CS, 2005.

[11] Roy Friedman and Robbert van Renesse. Strong and weak virtual synchrony in horus. In *SRDS*, pages 140–149, 1996.

[12] Vassos Hadzilacos and Sam Toueg. A modular approach to fault-tolerant broadcasts and related problems. Technical Report TR94-1425, Dep. of Computer Science, Cornell University, Ithaca, New York (USA), 1994.

[13] JoAnne Holliday. Replicated database recovery using multicast communication. In *NCA*, pages 104–107. IEEE-CS, 2001.

[14] Luis Irún, Hendrik Decker, Rubén de Juan, Francisco Castro, Jose E. Armendáriz, and Francesc D. Muñoz. MADIS: A slim middleware for database replication. In *Euro-Par*, volume 3648 of *LNCS*, pages 349–359. Springer, 2005.

[15] Ricardo Jiménez-Peris, Marta Patiño-Martínez, and Gustavo Alonso. Non-intrusive, parallel recovery of replicated data. In *SRDS*, pages 150–159. IEEE-CS, 2002.

[16] Bettina Kemme and Gustavo Alonso. Don't be lazy, be consistent: Postgres-R, a new way to implement database replication. In *VLDB*, pages 134–143, 2000.

[17] Bettina Kemme, Alberto Bartoli, and Özalp Babaoglu. Online reconfiguration in replicated databases based on group communication. In *DSN*, pages 117–130. IEEE-CS, 2001.

[18] Yi Lin, Bettina Kemme, Marta Patiño-Martínez, and Ricardo Jiménez-Peris. Middleware based data replication providing snapshot isolation. In *SIGMOD Conference*, 2005.

[19] Francesc D. Muñoz, Jerónimo Pla, María Idoia Ruiz, Luis Irún, Hendrik Decker, José Enrique Armendáriz, and José Ramón González de Mendívil. Managing transaction conflicts in middleware-based database replication architectures. In *SRDS*, pages 401–410. IEEE-CS, 2006.

[20] Marta Patiño-Martínez, Ricardo Jiménez-Peris, Bettina Kemme, and Gustavo Alonso. MIDDLE-R: Consistent database replication at the middleware level. *ACM Trans. Comput. Syst.*, 23(4):375–423, 2005.

[21] Christian Plattner, Gustavo Alonso, and Michael Tamer-Özsu. Indexing multidimensional time-series. *VLDB J.*, 2006. *Accepted for publication*.

[22] TPC-W. Transaction processing performance council. Accessible in URL: `http://www.tpc.org`, 2005.

[23] M. Wiesmann, A. Schiper, F. Pedone, B. Kemme, and G. Alonso. Database replication techniques: A three parameter classification. In *SRDS*, pages 206–217, 2000.

[24] Shuqing Wu and Bettina Kemme. Postgres-R(SI): Combining replica control with concurrency control based on snapshot isolation. In *ICDE*, pages 422–433. IEEE-CS, 2005.