Model-Driven Comparison of State-Machine-based and Deferred-Update Replication Schemes

Paweł T. Wojciechowski, Tadeusz Kobus, and Maciej Kokociński
Poznań University of Technology, Poland
{Pawel.T.Wojciechowski,Tadeusz.Kobus,Maciej.Kokocinski}@cs.put.edu.pl

Technical Report RA-01/12
15 April 2012

Abstract

In this paper, we analyze and experimentally compare state-machine-based and deferred-update (or transactional) replication, both relying on atomic broadcast. We define a model that describes the upper and lower bounds on the execution of concurrent requests by a service replicated using either scheme. The model is parametrized by the degree of parallelism in either scheme, the number of processor cores, and the type of requests. We use our model to make a comparison with a non-replicated service, considering separately abcast- and request-execution-dominant workloads. To evaluate transactional replication experimentally, we developed Paxos STM—a novel fault-tolerant distributed software transactional memory with programming constructs for transaction creation, abort, and retry. We used JPaxos for state-machine-based replication. Both systems share the same implementation of atomic broadcast built on the Paxos algorithm. We present the results of performance evaluation of both replication schemes, and a non-replicated (thus prone to failures) service, considering various workloads. The key result of our theoretical and experimental work is that neither system is superior in all cases. We discuss these results in the paper.

1 Introduction

Replication is an important method to increase service reliability and accessibility. It means deployment of a service on several server machines, each of which may fail independently, and coordination of service replicas so that they maintain a consistent state. Each replica maintains service state in memory (and optionally in its local nonvolatile store).

We analytically and experimentally analyze two approaches to fault-tolerant replication of services (or objects). The first approach is a *replicated state machine (SM)* [1] in which a client request is executed on every server. Services must be *deterministic*: any service replica being in the same state always produces the same effect upon the same request. SM coordinates all servers, so that all requests are delivered and processed by every replica in the same order. Thus, concurrent object accesses are consistent. The second approach is *transactional replication (TR)* based on *deferred update* [2] (also known as multi-primary passive replication). Programmers use atomic transactions to access critical objects; such objects are replicated on every server. The transaction's atomicity and serializability guarantee that the concurrent modifications of object replicas are propagated consistently to every server. In the paper, we consider deferred update relying on *atomic broadcast (abcast)* [3, 4]. This technique prevents deadlocks and allows better scalability than using two-phase commitment since transactions are never blocked [5, 6].

The SM scheme is more general than TR since it can also be used for replication of services that require linearizability [7] while transactions in TR normally only guarantee serializability. Thus, it is important to remember that some services can be replicated using SM but not TR. Hence, our comparison is valid only for replicated services, in which concurrent requests satisfy one-copy serializability [8]. We also assume the *crash-recovery* model of failure, in which $\lceil N/2 \rceil - 1$ server crashes can be tolerated, where N is the number of servers. A server can rejoin the system any time after a crash.

The contributions of the paper are twofold. We define a model of SM and TR, and use it to analyze both replication schemes. Our model describes the upper and lower bounds on the execution of concurrent requests parametrized by the degree of parallelism, the number of processor cores, and the type of requests. Next, we present and discuss the performance evaluation results for SM and TR, obtained using two popular microbenchmarks. To our best knowledge, this is the first experimental comparison of the two replication schemes and a non-replicated service in a uniform environment (same abcast), under varying contention loads and requests sizes.

To facilitate experimentation, we developed *Paxos STM—distributed software transactional memory (DSTM)* for TR replication. Paxos STM has programming constructs for transaction creation, abort, and retry. DSTM systems are a follow-up of the work on *software transactional memory (STM)*—a concurrent programming mechanism intended to replace locks (see [9, 10] among others). In STM, an atomic transaction is any piece of code containing memory reads and writes that must be executed atomically. To avoid blocking, STM systems use *optimistic concurrency control*—a transaction that conflicts with another concurrent transaction is rolled back and retried. DSTM is essentially like STM but the transactional memory is replicated on many network nodes. Contrary to the majority of existing DSTM systems, Paxos STM implements the deferred-update replication scheme using a transaction certification protocol which is based on atomic broadcast.

Below is an example code, which has been taken verbatim from the executable source. The program creates a transaction that accesses two objects `accA` and `accB` atomically:

```
@TransactionObject class Account { ... }
Account accA, accB;

new Transaction() {
```

```
    public void atomic() {
      float amount = 100;
      if (accA.balance() >= amount)
        accA.withdraw(amount);
      else
        retry();
      accB.deposit(amount);
    }
  };
```

For the objects to be accessed atomically, their class `Account` has been annotated as transactional. Paxos STM replicates such objects on a group of servers, coordinates the execution of concurrent transactions preserving isolation (one-copy serializability), and maintains a consistent view of object replicas on every server despite server failures. If `retry` is executed, a transaction is rolled back and reexecuted.

To experimentally evaluate SM replication, we used JPaxos [11]—a state-machine-based replication library, implementing Paxos [12] for replica (server) coordination. Paxos STM reuses this protocol code for agreement coordination. Thus, we are able to fairly compare the results of benchmarks that we implemented using both tools. The analytical model helped us to understand and interpret our experimental results precisely.

1.1 Motivations and results

The motivations to do this research were twofold. Firstly, to our best knowledge there was no prior work on rigorous evaluation and comparison of the SM and TR replication schemes, including estimation of the upper and lower time bounds. Secondly, reasoning about advantages, limitations and possible optimization paths of both replication schemes is difficult without a performance model that abstracts away from any uninteresting details. Although the *modus operandi* of SM and TR may appear simple, concurrency, transaction conflicts, and dealing separately with read-only and read-write (update) requests make the model quite subtle.

The main contributions of our work are the following:

- We designed and implemented Paxos STM, a programming tool for TR replication of services;
- We defined a model of SM and TR that describes the upper and lower bounds on processing a set of concurrent requests (assuming no delays); the lower bound is given for unoptimized and optimized schemes, where the latter recognizes read-only requests and treats them differently;
- Our model shows precisely the potential benefits of various means to increase parallelism in SM and TR and so also to increase throughput, such as optimized abcast, dealing with read-only requests differently, detecting conflicts earlier and fully using multi-core CPUs;
- We have shown when concurrent processing of requests by SM and TR can be faster than their sequential execution, considering (in TR) upper bounds on the number of conflicts that cause transactions to be reexecuted;
- We examined throughput and scalability using two microbenchmarks (Hashmap and Bank) and compared SM and TR; the comparison is fair since JPaxos and Paxos STM share the same implementation of abcast.

Both JPaxos and Paxos STM support the crash-recovery model of failure, which means that a server replica can recover after a crash and catch up on the current state automatically (from a local disk and/or other replicas). However, we do not show evaluation results under faulty behavior

scenarios in this paper since our focus is on modeling and comparing normal behaviour of both systems when no failures occur. We leave examining the faulty behaviour for future work.

1.2 Paper structure

The paper has the following structure. First, we define the analytical model of state machine and deferred update replication in §2. Next, we show the results of our evaluation experiments in §3, comparing performance and scalability of the two replication schemes. Then, we discuss related work in §4. Finally, we conclude in §5.

2 Analytical Model

In both SM and TR replication protocols, we can identify some of the following 5 phases [2], each of which takes the amount of time given in brackets:

1. Client request (q)
2. Server coordination (t_{sc})
3. Local execution (e)
4. Agreement coordination (t_{ac})
5. Client response or answer (a)

In our model, each client request consists of a sequence of operations to be executed atomically, that can read or modify the database (i.e., objects). We call a client request a "transaction" both in the TR and SM scheme. A replicated service (or database) satisfies *one-copy serializability (1SR)* [8]: transactions performed on the replicas have an ordering which is equivalent to an ordering obtained when the transactions are performed sequentially in a single centralized database.

We identify two types of requests: *read-only* and *update*, denoted respectively, r and rw. The former describes requests that contain only read operations to be executed by a replicated service, while the latter contains at least one write. Subscripts r and rw denote the corresponding request types. For the model to be tractable, we consider a replicated system in which the time e of processing a request locally by either TR or SM is the same for all requests; the same holds for q and a. Thus, the total time T of processing $n = n_r + n_{rw}$ requests by a service replicated on N machines (servers) using SM or TR is $\mathcal{M}(q, t_{sc}, e, t_{ac}, a)$, where function $\mathcal{M}$ depends on the semantics of SM and TR and the parallelism enabled by the underlying system.

For each replication scheme, we estimate T_{upper} and T_{lower}, which are the upper and lower bounds on time T. These bounds correspond respectively, to the worst and the best cases when computing $n = n_r + n_{rw}$ concurrent requests *without any delay*. We use a superscript $1{\rightarrow}n$ or $1|..|n$ to declare either sequential (non-replicated) or parallel processing. In the estimation of the latter, we consider the number of available processors (or CPU cores) per server, denoted c, and abstract away other hardware restrictions.

We use the following symbols: N is the number of servers (replicas), n is the number of requests (equal to the number of transactions), t_{abc} is the execution time of a single atomic broadcast (abcast) with a superscript r (requests) for SM and o (object read-sets and write-sets) for TM (t^r_{abc} can differ from t^o_{abc} considerably, depending on the size of requests vs. the size of read/write-sets), β is the degree of the atomic broadcast optimization ($\beta \geq 1$), where the greater the β the higher the throughput of the atomic broadcast; $\beta = 1$ means no optimization. Thus, broadcasting

a continuous stream of n requests by the optimized abcast protocol takes $\frac{n}{\beta} t_{abc}$, where β is assumed to be a small number (say < 5).

Optimizations of abcast are request batching and pipelining [13]. *Batching* means processing of several requests by a single round of abcast. The best results are when there is a continuous stream of requests, so that the protocol does not need to wait for the batch to be filled in. *Pipelining* of instances [12] is an extension of the basic Paxos protocol, allowing the Paxos leader to initiate new instances of the ordering protocol before the previous ones have completed.

2.1 State machine replication (SM)

We first model a non-replicated state machine, and extend it to a replicated SM executing on N servers with c cores each.

Single processing ($N = n = 1$):

$$t_{sc} = t^r_{abc} \tag{1}$$

$$t_{ac} = 0 \tag{2}$$

$$T^1_{SM} = q + t_{sc} + e + t_{ac} + a = q + t^r_{abc} + e + a \tag{3}$$

Sequential processing ($N = 1$, $n > 1$):

$$T^{1 \to n}_{SM} = n T^1_{SM} \tag{4}$$

Parallel processing ($N > 1$, $n > 1$): In a system replicated on N servers using SM:

- all requests must be executed sequentially and in the same order by all servers,
- server coordination (atomic broadcast) and the execution of requests can occur in parallel,
- a client can submit request to any non-faulty server, and the current Paxos leader will broadcast the message.

The above assumptions hold in JPaxos [11]—our reference implementation of SM. An optimized (rare in practice) variant of SM could recognize requests that are read-only and process them differently, replacing the first assumption by:

- update requests must be executed sequentially and in the same order by all servers,
- read-only requests can be executed concurrently by individual servers (i.e., $t^r_{abc} = 0$).

Consider a sequence of three client requests $r(o)$, $w(o)$, and $r(o)$, each containing only a single operation either reading (r) or updating (w) a replicated object o. The unoptimized SM guarantees that the last read will always see object o modified by $w(o)$. In the optimized SM, read-only requests are not ordered, so the last read may not see the update of object o. To solve this problem additional machinery is required (see e.g., [12]), which we neglect in our model since without this the 1SR property is still guaranteed: the effect is equivalent to a sequential execution $r(o)$, $r(o)$, and $w(o)$.

Below we compute the upper and lower bounds on the total time of processing n requests by SM. The lower bound is computed both for the unoptimized and optimized SM.

2.1.1 Upper bound

In the worst case there is no concurrency, which means sequential execution. Thus, we have:

$$T^{1|..|n}_{SM_{upper}} = nT^1_{SM} \tag{5}$$

2.1.2 Lower bound (unopt. SM)

In the best case there is as much concurrency as possible, i.e. server coordination (abcast) and the execution of requests occur in parallel. Concurrently, the client requests and responses are communicated but these times are relatively short. Thus, the total time of processing n requests by an unoptimized SM has only two outcomes, depending on which of the parallel parts will take more time. This is expressed using a function max, where $max(a, b) = a$ if $|a| > |b|$ and if $a = b$ then max returns either a or b:

$$T^{1|..|n}_{SM_{lower}} = T^1_{SM} + max(\frac{n}{\beta}t^r_{abc}, ne) - \delta_{SM} \tag{6}$$

where

$$\delta_{SM} = \begin{cases} t^r_{abc} & \text{if } max(\frac{n}{\beta}t^r_{abc}, ne) = \frac{n}{\beta}t^r_{abc} \\ e & \text{if } max(\frac{n}{\beta}t^r_{abc}, ne) = ne \end{cases}$$

In the former case ($\delta_{SM} = t^r_{abc}$), we say that *abcast time is dominant* (i.e., requests are short). In the latter case ($\delta_{SM} = e$), *request processing time is dominant* (i.e., requests are long).

2.1.3 Lower bound (opt. SM)

In the optimized SM, read-only requests can be processed by any non-faulty replica in parallel with any other requests. However, parallelism is limited by the fact that each of the N servers (replicas) has only c processors (or processor cores), where $c \geq 1$.

Thus, the best case for single-core servers ($c = 1$) is:

$$\begin{aligned} T^{1|..|n}_{SM^{opt}_{lower}} &= max\left(\frac{n_{rw}}{\beta}t^r_{abc}, (n_{rw} + \left\lceil\frac{n_r}{N}\right\rceil)e - \pi\right) \\ &+ T^1_{SM} - \delta_{SM}, \text{ where } \pi \in \{0, t^r_{abc}\} \end{aligned} \tag{7}$$

where π models parallelism between read-only requests and the first abcast (if $n_r = 0$ then $\pi = 0$ else $\pi = t^r_{abc}$). If request processing is dominant, then either requests are long or they are short but many read-only requests are among them.

If servers have many processors (or processor cores) then:

$$\begin{aligned} T^{1|..|n}_{SM^{opt}_{lower}} &\approx max\left(\frac{n_{rw}}{\beta}t^r_{abc}, n_{rw}e, \left\lceil\frac{n_r}{N(c-\theta)}\right\rceil e - \pi\right) \\ &+ T^1_{SM} - \delta_{SM}, \text{ where } c > 1 \text{ and } \theta \in \{0, 1\} \end{aligned} \tag{8}$$

where n_{rw} and n_r denote correspondingly, the number of update and read-only requests, and δ_{SM} equals respectively, t^r_{abc} if max returns the first argument, and e otherwise. When no parallel update requests are present, read-only requests can be executed on c cores instead of $c - 1$. We use $\theta \in \{0, 1\}$ to model this choice. More precisely, the third argument of max should be $\lceil\frac{n'_r}{Nc} + \frac{n''_r}{N(c-1)}\rceil e - \pi$, where $n'_r + n''_r = n_r$. However, we prefer the less precise approximation to avoid cluttering the model with additional parameters.

Below we give the main results for the optimized SM. The proofs of lemmas are available in the Appendix.

Lemma 1. *The speedup of processing requests by the optimized SM when compared to the un-optimized SM is proportional to the number of read-only requests, and—for abcast dominant processing—inversely proportional to β.*

Lemma 2. *In the best case, a service replicated on N single-core processor servers ($N > 1, c = 1$) using the optimized SM scheme is not slower than the non-replicated service if at least one request is read-only and*

$$n_{rw} + \left\lceil \frac{n_r}{N} \right\rceil - \frac{\pi}{e} \leq \frac{n_{rw} t^r_{abc}}{\beta e} \leq n - 1 \tag{9}$$

when abcast is dominant, and

$$\left(\frac{n_{rw}}{\beta} + 1 \right) t^r_{abc} \leq \left\lceil \frac{n_r}{N} \right\rceil e - \pi \leq ne \tag{10}$$

when request execution is dominant.

2.2 Transactional replication (TR)

In TR, each request is processed as a single *atomic transaction* that can read and write a set of objects atomically. All transaction objects are replicated on every server. TR maintains one-copy serializability of distributed object accesses. Concurrent transactions are executed optimistically (objects are not locked) and may conflict. An update transaction (or request) x *conflicts* with some concurrent transaction y that is about to commit, resulting in x being rolled back and reexecuted, if x reads any object modified by y. We call the former transaction *conflicting* and the latter one *committing*.

We denote K to be the number of conflicts while processing n requests by TR ($K \geq 0$). Note that K is also the number of transaction (or request) reexecutions caused by conflicts. K depends on the type of requests and the intersection of objects modified by transactions (the more shared objects, the higher the probability of a conflict). By a conflict definition, write-only requests cannot conflict[1]. Since read-only requests are not causally ordered, they also do not cause conflicts, unless strict one-copy serializability is required. K does not depend on the number of servers (replicas) N. K cannot be statically predicted since whether (or not) a conflict occurs depends on transaction interleaving at runtime. However, the upper bound can be estimated—if n update requests have been submitted by clients concurrently, the number of conflicts cannot be greater than $(n-1) + \ldots + 1 = \frac{(n-1)n}{2}$.

Instead of server coordination, TR requires agreement coordination, which is responsible for *transaction certification*: when a transaction has completed, the effects of its execution are sent to all servers (replicas) using atomic broadcast; if no conflicts with other concurrent transactions are detected locally, the effects are made permanent on every server and the transaction *commits*. Otherwise, the transaction is rolled back and reexecuted.

Thus, the server and agreement coordination times are:

$$t_{sc} = 0 \tag{11}$$

$$t_{ac} = t_{cer} + t^o_{abc} \tag{12}$$

where t_{cer} and t^o_{abc} are correspondingly, the local transaction certification time and the time of atomic broadcast, where the latter is executed on commit only. The broadcast data include object

[1]However, in our object-oriented Paxos STM they are treated as regular rw requests, since a transaction normally modifies only a subset of object fields; the other object fields are then "read".

read-sets, write-sets, and changes made to objects. For simplicity, we assume that all these messages have the same size for all update requests (transactions).

The agreement coordination phase also includes any other operations of the transaction processing protocol, such as creation of object shadow copies accessed by transactions. Some of these operations are executed in parallel with abcast, while the rest is assumed to be included in t_{cer}. Since in typical applications the network communication will be the bottleneck, we assume that $t_{cer} << t^o_{abc}$.

Single processing ($N = n = 1$):

$$T^1_{TR} = q + 0 + e + t^o_{abc} + t_{cer} + a = q + t^o_{abc} + e + a + t_{cer} \ . \tag{13}$$

Sequential processing ($N = 1, n > 1$):

$$T^{1 \to n}_{TR} = nT^1_{TR} \ . \tag{14}$$

Note that if $t^o_{abc} = t^r_{abc}$ then we obtain $T^1_{TR} = T^1_{SM} + t_{cer}$ and $T^{1 \to n}_{TR} = nT^1_{SM} + nt_{cer}$.

In single and sequential processing, there are obviously no concurrent transactions, so conflicts cannot occur.

Parallel processing ($N > 1, n > 1$): In a system replicated on N servers using TR:

- a client can submit request to any non-faulty server,
- each request (transaction) is executed by one server only and any object modifications are consistently applied to object replicas on all servers,
- requests (transactions) can be executed in parallel,
- each server is multi-threaded and can execute its requests (transactions) concurrently under an optimistic concurrency control scheme (no blocking),
- conflicts are detected as soon as possible, so a conflicting transaction can be aborted before completion, giving the execution time less than e and $t^o_{abc} = 0$ (but for simplicity, we use e to describe such cases),
- a conflict can also be detected after a transaction completes but before its effects are broadcast, causing an abort of the committing transaction; then t^o_{abc} is also 0.

Paxos STM that we developed for experimental validation allows an optimized variant of TR (which is common for TR):

- read-only requests do not need agreement coordination.

We say that a conflict is detected *early* to describe one of the two cases above. Conflicts can be detected early no matter if the conflicting and committing transactions are executed on the same server or not.

As before, we consider 1SR as the criterion of correctness. If consistency guarantees should reflect any causal relations between requests issued by various clients, they have to be ensured by the replicated service itself.

Below we compute the upper and lower bounds on the total time of processing n requests by TR.

2.2.1 Upper bound

We consider almost sequential execution but allow a bit of concurrency, so before a transaction commits with its effects stored, a new transaction can commence that may conflict with the committing transaction. Then the total time of processing n requests is

$$\begin{aligned} T^{1|..|n}_{TR_{upper}} &\approx n(q + e + t^o_{abc} + t_{cer} + a) + K(e + t_{cer} + t^o_{abc}) \\ &= T^{1 \to n}_{TR} + K(e + t_{cer} + t^o_{abc}) \quad \text{where } 0 \leq K < n \end{aligned} \tag{15}$$

We approximated the almost sequential execution to sequential but admitted K conflicts ($0 \leq K < n$). We assumed the worst case: a conflicting transaction x has completed and broadcast its effects to all servers as part of transaction certification. Not till then did the servers detect that there is a conflict with some other transaction that completed soon after x had commenced but not committed yet. The conflicting transaction x must be rolled back and reexecuted, hence its execution time is $2(e + t_{cer} + t^o_{abc})$.

2.2.2 Lower bound (unoptimized TR)

In the best case, all transactions are executed by TR in parallel but in case of any conflicts, as before, the conflicting transactions must be rolled back and reexecuted. The first naîve version of the lower bound is

$$\begin{aligned} T^{1|..|n}_{TR_{lower}} \approx T^1_{TR} + max\Big(& f(n')(e + t_{cer}) + \frac{K'}{\beta} t^o_{abc}, \\ & \frac{n + K'}{\beta} t^o_{abc} \Big) - \delta_{TR} \end{aligned} \tag{16}$$

where $f(n')$ is a *conflict function* which for a given number of conflicting transactions n' executed in parallel ($0 \leq n' < n$) returns a factor that when multiplied by a time of processing one transaction gives the time of processing all n' transactions. Note that some of the conflicting transactions can be executed several times until they commit, so the number of conflicts K can be larger than n'. Once a conflict is detected the conflicting transaction is aborted but for simplicity we still use the complete times e and t_{cer} to describe its execution. We also use K' which is the number of transaction conflicts that are *not* detected early, hence the transaction effects are abcasted to all servers.

We use function *max* defined in §2.1.2 to get the total time of the parallel execution of local processing and abcast. In order to reflect the order which is imposed by TR phases, the first argument of *max* describes a *sequence* of the 'execute', 'certify' (locally), and 'abcast' operations of a subset of transactions that run into conflicts (not detected early), while the second argument of *max* describes the total of 'abcast' operations of all transactions. Note that abcasts of non-conflicting transactions may occur in parallel with the local execution of retried transactions. We reduce (16) to

$$T^{1|..|n}_{TR_{lower}} \approx T^1_{TR} + max\big(f(n')e', \frac{n}{\beta} t^o_{abc}\big) + \frac{K'}{\beta} t^o_{abc} - \delta_{TR} \tag{17}$$

where $e' = e + t_{cer}$

$$\delta_{TR} = \begin{cases} 0 & \text{if } max(f(n')e', \frac{n}{\beta} t^o_{abc}) = f(n')e' \\ t^o_{abc} & \text{if } max(f(n')e', \frac{n}{\beta} t^o_{abc}) = \frac{n}{\beta} t^o_{abc} \end{cases}$$

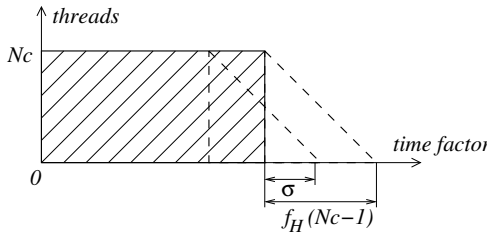

Figure 1: Processing $n + K$ requests on Nc cores, where $K \geq Nc - 1$

The interpretation of the above conditional is as follows. In the former case ($\delta_{TR} = 0$), *request execution is dominant*, i.e. either transactions are relatively long or many conflicts occur that are detected early (which means no abcast is required). In the latter case ($\delta_{TR} = t^o_{abc}$), *atomic broadcast is dominant*, i.e. transactions are relatively short and there are few conflicts.

The more concurrent accesses of shared objects occur and the more parallelism is allowed, the more probable the conflicts are. The actual number of conflicts K and a value $f(n')$ depend on the intersection of objects shared by transactions and their runtime interleaving. In the Appendix, we defined a conflict function $f()$ for a few special cases, and computed the upper/lower bounds for the worst case ($f(n') = n'$).

2.2.3 Lower bound (optimized TR)

In our estimation of TR's lower bound, we neglected hardware restrictions and assumed an unlimited number of processors. Below we approximate the lower time bound of processing n concurrent requests by a service replicated in a system of N servers with c processor cores each, and consider the optimized TR in which read-only requests do not require the agreement coordination phase:

$$\begin{aligned} T^{1|..|n}_{TR^{opt}_{lower}} &\approx T^1_{TR} + max\Big(\Big\lceil \frac{n+K}{Nc} \Big\rceil e' + \Sigma, \frac{n_{rw}+K'}{\beta} t^o_{abc}\Big) + \\ &- \delta_{TR} \text{ where } \Sigma = \sigma e' + \Big\lfloor \frac{\sigma}{x} \Big\rfloor t^o_{abc} \text{ and } \sigma \in \langle 0, f_H(Nc-1)\rangle \end{aligned} \tag{18}$$

where δ_{TR} is equal to e' if the first argument of max is greater, or t^o_{abc} otherwise. The conflict function $f_H()$ is defined as before but the function domain is $\langle 0, f_H(Nc)\rangle$, which means that at the same time there cannot be more than Nc concurrent conflicting transactions. $\Sigma = \sigma e' + \lfloor \frac{\sigma}{x} \rfloor t^o_{abc}$ describes the offset caused by the conflicting transactions that have not managed to be executed in parallel with non-conflicting transactions (see Fig. 1), where the upper bound on σ is equal $f_H(Nc-1)$ and x ($0 < x \leq \sigma + 1$) reflects the fact that some of the conflicts can be detected early and no abcast is used.

Below we give the main results for the optimized TR:

Lemma 3. *A service replicated on N servers, each having c-processor cores, using the optimized TR runs $n_r \frac{t^o_{abc}}{\beta}$ faster than when using the unoptimized TR.*

Lemma 4. *In the best case, a service replicated on N single-core processor servers ($N > 1$, $c = 1$) using the optimized TR scheme can be faster than the non-replicated service if*

$$\Big\lceil \frac{n+K}{N} \Big\rceil e' + \Sigma \leq \frac{n_{rw}+K'}{\beta} t^o_{abc} \leq ne - e' \tag{19}$$

when abcast is dominant, and

$$\frac{n_{rw}+K'}{\beta} t^o_{abc} \leq \Big\lceil \frac{n+K}{N} \Big\rceil e' + \Sigma \leq ne - t^o_{abc} \tag{20}$$

when request execution is dominant.

The proofs of lemmas are available in the Appendix. Note that in the equation (19) n_{rw} (and so t^o_{abc}) cannot equal 0 since we assumed abcast dominance.

If equation (19) or (20) holds, then the TR-replicated service can be faster than a non-replicated service. In particular, if $n_{rw} = 0$, there are no conflicts ($K = K' = \Sigma = 0$) and no abcast ($t^o_{abc} = 0$). So we can reduce (20) to

$$0 \leq \left\lceil \frac{n_r}{N} \right\rceil (e + t_{cer}) \leq n_r e \tag{21}$$

Since normally $t_{cer} << e$, the above equation is mostly true, which agrees with the intuition that a TR-replicated service can be faster than a non-replicated service if there are many read-only requests that are processed in parallel. Note that if $n_r = 1$ or $N = 1$ then the above equation is false. If $n_r \leq N$ then the equation is true only if $n_r \geq \frac{e+t_{cer}}{e}$.

3 Experimental Evaluation

In this section, we empirically evaluate and compare performance and scalability of SM- and TR-based replicated and a non-replicated service, modelled by two benchmarks.

3.1 Programming tools

In order to evaluate SM-based replication, we used JPaxos [11]—an efficient implementation of the Paxos [12] algorithm, with the support of the *crash-recovery* model of failure. To boost performance, JPaxos supports concurrent rounds of consensus and request batching. Request types are not recognized so it implements the unoptimized SM model (see §2.1).

To evaluate TR-based replication, we designed and implemented *Paxos STM*—an object-oriented fault-tolerant DSTM system, which replicates every shared object on each node for increased availability and minimal access latency. Paxos STM supports multi-primary passive replication (similar to multi-master replication in databases) and relies on the optimistic concurrency control scheme. Paxos STM's certification protocol is built on top of JPaxos, with each replica able to propose new updates to the distributed state by abcasting them. Paxos STM supports both the crash-stop and the crash-recovery failure models. Our system also takes advantage of modern multicore hardware by allowing multithreaded processing of transactions. It distinguishes between read-only and updating transactions, thus supporting the optimized version of TR replication (see 2.2). By the use of the object multiversioning optimization read-only transactions are guaranteed to commit successfully.

3.2 Benchmarks

To evaluate SM and TR replication schemes under different workloads and compare with a non-replicated run, we implemented two popular microbenchmarks: Hashtable and Bank.

3.2.1 Hashtable microbenchmark

The hashtable of size n stores key-value integer elements and manages them through *get/put/remove* operations. It is prepopulated with $n/2$ random elements from a defined range, thus giving

Operation (Transaction)	Default	Prolonged	High-Contention
`get` (RO)	100	100	100
`get` (RW)	8	8	40
`put/remove` (RW)	2	2	10
"active wait" 1ms (RO+RW)	0	1	0

Figure 2: The number of transactional operations (Hashtable benchmark)

the saturation of 50%. A single run consists of a series of requests issued to the hashtable. There are two types of requests (or transactions): *read-only (RO)* and *read-write (RW)*, which correspond to request types r and rw in §2. The RO request atomically performs a given number of *get* operations with a randomly chosen set of keys. The RW request executes a defined number of *get* operations followed by updating operations (either *put* or *remove*). To keep the hashtable 50% saturated, the decision whether to insert a new object to the hashtable or remove an existing one depends on the previous *get* operations.

The benchmark parameters in Fig. 2 reflect different kinds of workload: Default, Prolonged, and High-Contention. The RO transactions scan through a vast amount of data using many *get* operations. In contrary, RW transactions involve much fewer operations (two to ten times less, depending on the test), 25% of which are modifying ones. Different levels of contention can be generated by manipulating the size (or the number of involved operations) of the RW transactions and the relative number of RO and RW transactions. In the evaluation we use the same sizes of RO transactions for all tests and two different sizes of RW transactions. For each test, we examine three scenarios consisting of a different mix of RW and RO transactions: 10/90, 50/50, 90/10, denoted respectively: 10%, 50%, and 90% of RWs. In the Prolonged workload each request additionally performs the "active wait" for a given amount of time (1 ms) to simulate computation-heavy workloads.

3.2.2 Bank benchmark

Operations are performed on an array of accounts shared between nodes. We have two types of transactions: An RW transaction performs *transfer* of funds from one account to another, thus executing two read and two write operations on two distinct accounts in total. An RO transaction computes *balance*, which requires reading all of the accounts and summing up the funds. In our tests, we evaluate three scenarios with different percentage of RW transactions, namely 10%, 50%, and 90%. In each scenario, the number of accounts is 10000.

3.3 Evaluation Environment

We used a cluster of eight nodes, each equipped with a Xeon Quad-core X3230 2.66GHz, L2 cache 2x4MB CPU, 4GB RAM ECC DDR2, 800MHz, running OpenSUSE 10.3 (kernel 2.6.22.19) with Sun JRE 1.6.0. The nodes are connected via a private 1Gb Ethernet network.

JPaxos was configured to have at most two concurrent instances of consensus, the maximum batch size 64KB, and no batching delay. All available CPU cores were utilized, so the number of physical on-board cores c (see §2) is four. We experimentally established an optimal number of threads in Paxos STM to be 20 for the Hashtable benchmark and 80 for the Bank benchmark (these values were used in all of our tests). Such a high number of threads (far exceeding the number of physical cores) is necessary to fully exercise the hardware potential due to threads blocking on network I/O operations.

Benchmark service	10% RW	50% RW	90% RW
a) Default Hashtable	110398	119135	161415
b) Prolonged Hashtable	666	667	667
c) High-Contention Hashtable	103666	106608	120830
d) Bank	123183	155615	206042

Figure 3: The results of the non-replicated benchmark execution (req/s)

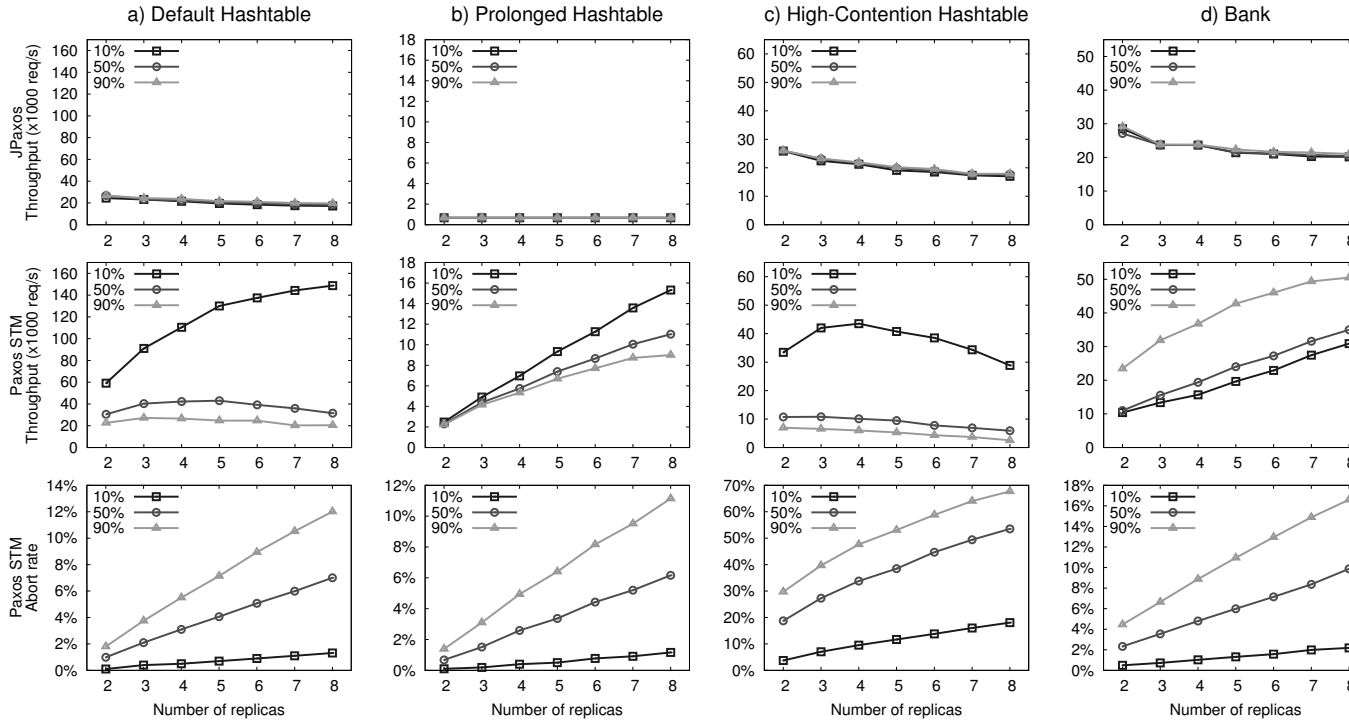

Figure 4: Benchmarks, where 10%, 50%, and 90% denote the percentage of read-write (RW) requests (or transactions).

3.4 Evaluation results and analysis

Below we discuss the results of benchmark tests. In Fig. 4, we present *throughput* obtained using JPaxos and Paxos STM, i.e. the number of transactions committed per second. We also present the transaction *abort rate* (in Paxos STM), i.e. the percentage of transactions aborted due to conflicts and reexecuted (equal $\frac{K}{n+K}100\%$ in our model in §2); the abort rate gives useful insight into the level of contention.

3.4.1 Default Hashtable

Hashtable with default configuration executed under JPaxos on two nodes, touches the score of 25000 requests per second (req/s); see Fig. 4-a. With the number of nodes increasing the performance gradually decreases stabilizing at the level of around 17000 req/sec. This drop results from higher coordination costs of maintaining a higher number of replicas that the Paxos leader replica must handle. The differences among scenarios including a various mix of request types are minimal. However, slightly (7-11%) better results are obtained in scenarios including more read-write (RW) requests. The larger size of read-only (RO) requests adds to the execution time,

as well to the abcast time since more data needs to be exchanged between nodes.

Results of Paxos STM evaluation show more differences between various scenarios. Conversely to JPaxos, better results are obtained with a higher percentage of read-only requests, which do not require the server agreement phase and therefore the costly abcast operation. The results in scenarios including 90% and 50% of read-write requests resemble the ones obtained with JPaxos, but scaled up by a certain factor. Paxos STM 50% scenario performance is roughly twice the JPaxos'. One can observe that the best performance is obtained with a small number of nodes and it falls with the increasing number of nodes. The 10% scenario exhibits different characteristics. It scales up with the number of nodes. It is the result of Paxos STM's ability to process read-only requests (which are almost an order of magnitude more frequent than in other scenarios) in a fully parallel manner with no communication overhead. The higher the number of nodes, the more requests can be processed. The top performance obtained for a maximum number of nodes reaches almost 150000 requests per second. In the other scenarios the predominant cost of abcast does not allow to fully exercise the potential level of parallelism. The abort rate in all scenarios is moderate and ranges from 0 to 12%, depending on the scenario.

The results obtained using a non-replicated Default Hash-table greatly surpass the results of both JPaxos and Paxos STM (see Fig. 3). This immense throughput of the latter is, however, achieved at the cost of absolutely no fault tolerance. Note that in the 10% RW scenario, Default Hash-table replicated using Paxos STM outperforms its non-replicated, non-fault-tolerant variant if there are more than 4 replicas.

3.4.2 Prolonged Hashtable

Contrary to the first test, where the cost of abcast in SM as well as TR was predominant, the second test aims at mimicking a computation-heavy workload (which corresponds to the request processing time dominance in §2). The parameters of this test differ only in one aspect compared to the Default Hashtable configuration—the execution of each request is prolonged by 1 ms.

JPaxos's evaluation (see Fig. 4-b) show stunningly uniform performance of 675 requests per second regardless of the number of nodes involved in computation. It indicates that a high execution time of requests entirely covers up any cost of replica coordination. The system throughput is directly limited by the time needed by replicas to actually execute the requests.

On the other hand, Paxos STM exhibits excellent scaling capabilities. Performance increases with the number of nodes. For the 10% RW scenario, it does so almost linearly. In other scenarios the agreement coordination phase required by (more frequent in these cases) RW requests introduces a slight overhead. The fall of performance in case of 50% RW and 90% RW scenarios is small up to 3-4 replicas and slightly raises for a higher number of nodes. Performance achieved for the minimal number of nodes is almost four times higher than in case of JPaxos since Paxos STM takes advantage of the multicore hardware architecture. The abort rate is nearly identical to the one from the previous test.

The throughput of a non-replicated Prolonged Hashtable is similar to JPaxos and limited by the request execution time. Now, the performance of the non-replicated service cannot be matched to Paxos STM which is superior this time (see Fig. 3). This result is justified by Lemma 4-(20), showing precisely when TR can be faster for the execution-dominant workload.

3.4.3 High-Contention Hashtable

This benchmark test aims at examining both replication approaches under high contention. For this, the number of read and write operations in RW requests grew 5 times compared to the Default

Hashtable benchmark configuration. The performance of JPaxos (see Fig. 4-c) is very similar to the one obtained from the first test. It is due to the requests being executed by JPaxos sequentially, thus not conflicting with each other. The main difference is the lower performance of scenarios involving more RW requests. Their performance is now closer to the performance of the read-only dominated scenario, because the lengths of RO and RW requests are now comparable.

The change of contention level has a much more visible impact on Paxos STM. The abort rate now reaches up to 20% for 10% RW scenario (20 times more), 50% for 50% RW scenario and, 70% for 90% RW scenario. Thus, in the case of the highest contention level almost every RW transaction is aborted at least once. The impact of such a high abort rate can be easily observed in the throughput diagram. The 50% RW and 90% RW scenarios, which are the most affected by the contention increase, demonstrate the max. four times throughput downfall. The throughput is diminished even more for a higher number of nodes, where this decrease is even larger—up to eight times. In the 10% RW case, the throughput is roughly halved for a small number of nodes, and then falls drastically when the number of nodes increases. Note that only RW transactions may be aborted since RO transactions are guaranteed to commit successfully. If 90% of transactions are guaranteed to succeed then, in case of the 20% abort rate, the rest of transactions (all RW transactions) are aborted twice on average. The higher is the number of RW transactions performed (including the aborted re-runs), the more this scenario resembles scenarios with a higher base percentage of RW transactions, so its performance is decreased.

Higher contention also has an impact on the execution of a non-replicated High-Contention Hashtable service (see Fig. 3). The lower performance can be attributed to higher packet processing overhead and the execution time of RW requests.

3.4.4 Bank Benchmark

In this test (see Fig. 4-d), the JPaxos throughput ranges from nearly 30000 to slightly above 20000, similarly as in the case of Hashtable. There is no observable difference between various scenarios. Even though RO requests consist of a huge number of operations the increased execution time has no significant impact on the overall throughput. The factor that has the biggest impact on the throughput is the abcast cost, which is similar for both types of requests. Note that although the RO request requires a large number of operations to be performed (all the accounts are scanned), the amount of data being sent is limited to a single word which depicts the type of the request.

In Paxos STM, the observed tendencies are different than before. Contrary to the previous tests, where the scenarios including more RO requests performed better, now the opposite is true. The best results were obtained for 90% RW scenario. This result can be explained by a higher execution time of RO requests that is even more important in the TR approach due to transactional processing overhead. Although RO requests have higher execution times, they scale well with the number of nodes, allowing Paxos STM to improve its performance from 10000 requests per second for two nodes up to 30000 for eight nodes in case of 10% RW scenario, and up to 35000 in case of 50% RW scenario. While the performance of Paxos STM increases with the number of nodes, the performance of JPaxos lowers slightly. In the 10% RW and 50% RW scenarios, both tools obtain roughly equal throughput on 5 nodes. With the lower number of nodes, JPaxos exhibits better performance, while with 5 and more nodes Paxos STM is superior.

In case of 90% RW scenario one can notice exceptionally good behaviour of Paxos STM. Usually in an abcast dominated workload the speed of Paxos STM is limited by the speed of abcast. Clearly, in this case, it is the opposite. It is the result of a number of effects occurring simultaneously: very small requests (each 76 bytes), extremely short transaction execution times, and a high number of updating transactions performed concurrently. In such circumstances replicas might have a lot of

transactions ready to commit at the same moment. In this case, the abcast protocol may broadcast them all at once using a single message. Thus, in practice, by optimizing abcast, we will be able to considerably improve performance of TR.

The non-replicated Bank service again tops all the approaches with the best throughput, ranging from 120000 to over 200000 (see Fig. 3). Yet again, this performance gain is at the expense of no support for fault-tolerance.

3.5 Evaluation Summary

In most cases, a non-replicated service outperforms its the SM- and TR-based replicated variant. It is however done at the expense of providing absolutely no fault tolerance. In some scenarios, however, the TR-based replicated service is the clear winner. This is justified by our theory: Lemmas 2 and 4 show precisely when SM- and TR-based replicated services can outperform its non-replicated variant, considering abcast or request dominant workloads. As expected (see §2.1.2), JPaxos does not scale at all and is insensitive to high-contention workloads. It performs poorly when the workload is execution dominated since all requests have to be processed sequentially. It is not the case with Paxos STM which takes the advantage of multicore hardware and allows for concurrent distributed processing on several nodes (see 2.2.3). This is especially visible in the case of read-only transactions. However, TR performance suffers under high contention. Abcast overhead should be reduced as much as possible since otherwise it may overshadow gains of parallelism and reduce scalability (e.g., 10% RW scenario in Fig. 4-c). On the other hand, SM outperforms TR in High-Contention Hashtable for 50% and 90% RW. Therefore one can see that no single solution would fit all purposes. However, TR-based replication holds a lot of promise.

4 Related Work

A lot of work was done on replication in distributed systems in the past years (see [2] for a survey), and different models and replication techniques have emerged. Unfortunately, various authors often use different terms to name similar abstractions. Some models (such as state-machine replication, originally proposed in [1]) evolved considerably since their initial formulation. Below we briefly describe some of the work most closely related to ours.

Our SM model describes the performance of the state-machine [1] and also the quorum-based [14] approaches to replication, each relying on a distributed agreement protocol. The key idea of the former approach is processing all requests in the same order by all replicas. On the other hand, the fundamental idea of the quorum-based replication is that a transaction is executed if the majority of sites vote to execute it. Our TR model describes a variant of the primary-copy replication [15] that allows many concurrent master replicas. It is called multi-primary passive replication [2] or, in the database community, deferred update or multi-master replication; in the classification of [6], it is an eager, update everywhere approach. As in the primary-backup replication, update transactions can only be processed on a master replica, with the updates propagated eagerly or lazily to slave replicas [6], but many concurrent master replicas are allowed.

The deferred update systems often employ pessimistic concurrency control based on strict two-phase locking (S2PL) [16]. Contrary, our TR model describes deferred update based on atomic broadcast, which allows transactions to be executed without blocking. Several authors demonstrated advantages of using this technique to replicate databases and make them tolerant to machine crashes (see e.g., [3, 4, 17, 18]). Various optimizations of the basic scheme are possible, e.g. readsets of update transactions do not need to be broadcast if an additional communication phase

is introduced to broadcast the decision regarding committing or restarting a transaction [19]. More recently, deferred update protocols tolerating Byzantine faults are also investigated (see e.g., [20]).

There exists work on analytical performance evaluation of transactional and replicated systems. But there is relatively little work on formalization of replication schemes similar to ours. Yu [21] defines an analytical model of various concurrency control schemes used in transactional processing. Ciciani *et al.* [22] describe an analytical model designed to study the tradeoff between replicating data in database systems using various pessimistic, optimistic, and semi-optimistic concurrency control schemes. However, this study does not include approaches based on group communication. Nicola and Jarke [23] propose a 2D analytical queueing model of replication for performance evaluation of distributed and replicated database systems. Jiménez-Peris *et al.* [24], analytically and experimentally compare various quorum-based data replication schemes. The authors conclude that in most cases the read-only-write-all-available approach outperforms quorum replication.

Paxos STM that we developed is a Distributed Software Transactional Memory (DSTM) system. Most of these systems extend the implementations of some non-distributed STMs with replication protocols, which are often designed *ad-hoc*, providing no fault-tolerance and depending on a central coordinator. In contrast to such systems, we designed Paxos STM from the ground up as a fault-tolerant distributed STM.

DiSTM [25] is an object-level DSTM implementing several coherence protocols. Serialization of concurrent transactions is ensured either by a distributed mutual exclusion algorithm, or by a lease mechanism. Leases are managed by a designated machine, which can be a bottleneck under high load. Anaconda [26] alleviates some of the DiSTM shortcomings by extending it with distributed object replication, caching mechanisms, and a new three-phase pessimistic concurrency control protocol. However, neither DiSTM nor Anaconda provide fault tolerance. The closest system to Paxos STM is D2STM [27], which also implements an optimistic transaction certification based on atomic broadcast and multiversioning. All objects are replicated on each node, thus eliminating the problem of fetching objects from remote locations. However, D2STM is built as a local STM, extended to support replication. More recently, D2STM has been equipped with the lease-based mechanism to limit abort rate under high contention [28].

5 Conclusions and Future Work

We analyzed and experimentally compared two approaches for replication of services (or databases), both based on atomic broadcast: replicated state machine (SM) and transactional (deferred update) replication (TR). The key corollary one can draw from our analytical model is that neither solution is superior in all cases. This is due to the differences between the two approaches in sensitivity to various workloads. Execution dominated workloads are handled much better when using TR since this approach can (inherently) execute multiple requests concurrently, contrary to classical SM. In particular, TR allows higher throughput than SM for read-write requests with a majority of read operations that do not cause conflicts (which is a typical workload of web services). However, performance gains from parallel request execution may be overshadowed by high costs of atomic broadcast, which is especially visible in the abcast-dominated workloads. The predictions given by our model are supported by the results of evaluation. For our experimental evaluation, we have used JPaxos (SM) and developed Paxos STM (TR). The tools are based on the same implementation of the MultiPaxos algorithm, thus ensuring fairness of the comparison. Since only TR exercises the ability to scale, one would expect it to perform better than SM. However, the results show that sometimes the overhead of the transactional machinery makes SM a better choice. One can also observe the high footprint of using either replication scheme compared to the performance of a non-replicated (thus prone to failures) variant. However, the fault-tolerance is worth the

price. Moreover, the costs of expensive inter-node communication can be partially compensated by parallel request execution in TR. In workloads that exhibit high request execution times this may even result in much higher performance of TR compared to a non-replicated service. To conclude, when considering replication as a mean of providing fault tolerance one should carefully choose one or the other solution based on the expected workload. In the future, we would like to compare TR and SM under faulty scenarios, using different protocols for recovery after crashes which are already supported by JPaxos and Paxos STM. Our comparison of SM and TR schemes is valid for services that require 1SR only. It may also be interesting to design TR with support of linearizability, and repeat the comparison.

Acknowledgments The authors would like to thank Jan Kończak, Nuno Santos, Tomasz Żurkowski, and André Schiper for their work on the implementation of JPaxos.

References

[1] F. B. Schneider, *Replication management using the state-machine approach.* ACM Press/Addison-Wesley Publishing Co., 1993, pp. 169–197.

[2] B. Charron-Bost, F. Pedone, and A. Schiper, Eds., *Replication: Theory and Practice*, ser. LNCS. Springer, 2010, vol. 5959.

[3] D. Agrawal, G. Alonso, A. E. Abbadi, and I. Stanoi, "Exploiting atomic broadcast in replicated databases (extended abstract)," in *Proc. of EuroPar '97*, Aug. 1997.

[4] B. Kemme, F. Pedone, G. Alonso, and A. Schiper, "Processing transactions over optimistic atomic broadcast protocols," in *Proc. of ICDCS '99*, Jun. 1999.

[5] A. Schiper and M. Raynal, "From group communication to transactions in distributed systems," *Communications of the ACM*, vol. 39, no. 4, pp. 84–87, Apr. 1996.

[6] J. Gray, P. Helland, P. O'Neil, and D. Shasha, "The dangers of replication and a solution," in *Proc. of SIGMOD '96*, 1996.

[7] M. P. Herlihy and J. M. Wing, "Linearizability: A correctness condition for concurrent objects," *ACM Transactions on Programming Languages and Systems (TOPLAS)*, vol. 12, no. 3, pp. 463–492, 1990.

[8] P. A. Bernstein and N. Goodman, "Serializability theory for replicated databases," *J. Comput. Syst. Sci.*, vol. 31, no. 3, pp. 355–374, Dec. 1985.

[9] N. Shavit and D. Touitou, "Software transactional memory," in *Proc. of PODCS '95*, Aug. 1995.

[10] T. Harris and K. Fraser, "Language support for lightweight transactions," in *Proc. of OOPSLA '03*, Oct. 2003.

[11] J. Kończak, N. Santos, T. Żurkowski, P. T. Wojciechowski, and A. Schiper, "JPaxos: State machine replication based on the Paxos protocol," Faculté Informatique et Communications, EPFL, Tech. Rep. 167765, Jul. 2011.

[12] L. Lamport, "The part-time parliament," *ACM Transactions on Computer Systems (TOCS)*, vol. 16, no. 2, pp. 133–169, May 1998.

[13] N. Santos and A. Schiper, "Tuning Paxos for high-throughput with batching and pipelining," in *Proc. of ICDCN '12*, 2012.

[14] D. K. Gifford, "Weighted voting for replicated data," in *Proc. of SOSP '79*, Dec. 1979, pp. 150–162.

[15] M. Stonebraker, "Concurrency control and consistency of multiple copies of data in distributed ingres," *IEEE Transactions on Software Engineering (TSE)*, vol. 5, no. 3, pp. 188–194, May 1979.

[16] P. A., Bernstein, V. Hadzilacos, and N. Goodman, *Concurrency control and recovery in database systems.* Addison-Wesley Longman Publishing Co., Inc., 1987.

[17] E. Cecchet, J. Marguerite, and W. Zwaenepoel, "C-JDBC: flexible database clustering middleware," in *Proc. of USENIX ATEC '04*, Jun. 2004, pp. 26–26.

[18] F. Pedone, R. Guerraoui, and A.Schiper, "The database state machine approach," *Distributed and Parallel Databases*, vol. 14, no. 1, pp. 71–98, Jul. 2003.

[19] B. Kemme and G. Alonso, "Don't be lazy, be consistent: Postgres-R, a new way to implement database replication," in *Proceedings of VLDB '00: the 26th International Conference on Very Large Data Bases*, 2000.

[20] F. Pedone, N. Schiper, and J. E. Armendáriz-Iñigo, "Byzantine fault-tolerant deferred update replication," in *Proc. of LADC '11*, Dec. 2011.

[21] P. S. Yu, "Modeling and analysis of transaction processing systems," in *Performance Evaluation of Computer and Communication Systems*, ser. LNCS, vol. 729. Springer-Verlag, 1993.

[22] B. Ciciani, D. M. Dias, and P. S. Yu, "Analysis of replication in distributed database systems," *IEEE Trans. on Knowl. and Data Eng.*, vol. 2, no. 2, pp. 247–261, Jun. 1990.

[23] M. Nicola and M. Jarke, "Increasing the expressiveness of analytical performance models for replicated databases," in *ICDT '99*, Jan. 1999.

[24] R. Jiménez-Peris, M. Patiño-Martínez, G. Alonso, and B. Kemme, "Are quorums an alternative for data replication?" *ACM Trans. Database Syst.*, vol. 28, no. 3, pp. 257–294, Sep. 2003.

[25] C. Kotselidis, M. Ansari, K. Jarvis, M. Luján, C. C. Kirkham, and I. Watson, "DiSTM: A software transactional memory framework for clusters," in *Proc. of ICPP '08*, Sep. 2008.

[26] C. Kotselidis, M. Lujan, M. Ansari, K. Malakasis, B. Kahn, C. Kirkham, and I. Watson, "Clustering JVMs with software transactional memory support," in *Proc. IPDPS '10*, 2010.

[27] M. Couceiro, P. Romano, N. Carvalho, and L. Rodrigues, "D2STM: Dependable Distributed Software Transactional Memory," in *Proc. of PRDC '09*, Nov. 2009.

[28] N. Carvalho, P. Romano, and L. Rodrigues, "Asynchronous lease-based replication of software transactional memory," in *Proc. of Middleware '10*, ser. LNCS, vol. 6452, 2010.

.1 State Machine Replication

Lemma 1. *The speedup of processing requests by the optimized SM when compared to the un-optimized SM is proportional to the number of read-only requests, and—for abcast dominant processing—inversely proportional to β.*

Proof. Below we estimate the speedup $speedup^n_{SM^{opt}}$ of processing n_{rw} update and n_r read-only concurrent requests by the optimized SM to be:

$$speedup^n_{SM^{opt}} = T^{1|..|n}_{SM_{lower}} - T^{1|..|n}_{SM^{opt}_{lower}} \approx max(\frac{n}{\beta}t^r_{abc}, ne) - \Delta \tag{1}$$

where $n = n_{rw} + r_r$ and

$$\Delta = \begin{cases} max\left(\frac{n_{rw}}{\beta}t^r_{abc}, (n_{rw} + \left\lceil\frac{n_r}{Nc}\right\rceil)e\right) & \text{if } c = 1 \\ max\left(\frac{n_{rw}}{\beta}t^r_{abc}, n_{rw}e, \left\lceil\frac{n_r}{N(c-\theta)}\right\rceil e\right) & \text{if } c > 1 \end{cases}$$

Then we obtain for abcast and request dominant processing respectively:

$$speedup^n_{SM^{opt}_{abc}} \approx \frac{n}{\beta}t^r_{abc} - \frac{n_{rw}}{\beta}t^r_{abc} = (n - n_{rw})\frac{t^r_{abc}}{\beta} = n_r\frac{t^r_{abc}}{\beta} \tag{2}$$

$$speedup^n_{SM^{opt}_e} \approx \begin{cases} ne - (n_{rw} + \lceil\frac{n_r}{Nc}\rceil)e \\ \quad = (n - n_{rw})e - \lceil\frac{n_r}{N}\rceil e \\ \quad = (n_r - \lceil\frac{n_r}{N}\rceil)e & \text{if } c = 1 \\ \left(n - max(n_{rw}, \lceil\frac{n_r}{N(c-1)}\rceil)\right)e & \text{if } c > 1\ . \end{cases}$$

Since $n_r \geq \lceil\frac{n_r}{N}\rceil$, the speedups are not negative, proportional to n_r ($n - n_{rw} = n_r$), and inversely proportional to β for abcast dominant processing, which ends the proof. □

Below we give the main result for the optimized SM:

Lemma 2. *In the best case, a service replicated on N single-core processor servers ($N > 1$) using the optimized SM scheme is not slower than the non-replicated service if at least one request is read-only and*

$$n_{rw} + \left\lceil\frac{n_r}{N}\right\rceil - \frac{\pi}{e} \leq \frac{n_{rw}t^r_{abc}}{\beta e} \leq n - 1 \tag{3}$$

when abcast is dominant, and

$$\left(\frac{n_{rw}}{\beta} + 1\right)t^r_{abc} \leq \left\lceil\frac{n_r}{N}\right\rceil e - \pi \leq ne \tag{4}$$

when request execution is dominant.

Proof. We compute below the overhead imposed by SM in the best case (the lower time bound) and for $c = 1$, and test the hypothesis that the overhead is not positive.

$$\begin{aligned} overhead^n_{SM^{opt}} &= T^{1|..|n}_{SM^{opt}_{lower}} - (q + \left\lceil\frac{n}{c}\right\rceil e + a) \\ &= max\left(\frac{n_{rw}}{\beta}t^r_{abc}, (n_{rw} + \left\lceil\frac{n_r}{N}\right\rceil)e - \pi\right) + \\ &\quad + e + t^r_{abc} - \delta_{SM} - ne\ . \end{aligned} \tag{5}$$

We first put forward the hypothesis considering abcast dominance:

$$overhead^{n}_{SM^{opt}_{abc}} = \frac{n_{rw}}{\beta} t^{r}_{abc} + e - ne \leq 0 \tag{6}$$

which implies that

$$\frac{n_{rw}}{\beta} t^{r}_{abc} \leq (n-1)e \quad \text{iff} \quad \frac{n_{rw}}{\beta} t^{r}_{abc} \geq (n_{rw} + \left\lceil \frac{n_r}{N} \right\rceil)e - \pi \ . \tag{7}$$

From the above we obtain

$$n_{rw} + \left\lceil \frac{n_r}{N} \right\rceil - \frac{\pi}{e} \leq \frac{n_{rw} t^{r}_{abc}}{\beta e} \leq n - 1 \ . \tag{8}$$

Note that if $N = 1$ or $n_r = 0$, then the above equation is false, which is intuitively valid since the overhead of replication must slow down the replicated service if no parallelism is possible.

Now let us put forward the hypothesis considering request execution dominance, i.e. when $(n_{rw} + \lceil \frac{n_r}{N} \rceil)e - \pi \geq \frac{n_{rw}}{\beta} t^{r}_{abc}$:

$$\begin{aligned} overhead^{n}_{SM^{opt}_{e}} &= (n_{rw} + \left\lceil \frac{n_r}{N} \right\rceil)e - \pi + t^{r}_{abc} - ne \\ &= \left\lceil \frac{n_r}{N} \right\rceil e - \pi + t^{r}_{abc} - n_r e \leq 0 \end{aligned} \tag{9}$$

which implies that

$$\left\lceil \frac{n_r}{N} \right\rceil e \leq n_r e - t^{r}_{abc} + \pi \quad \text{iff} \quad \left\lceil \frac{n_r}{N} \right\rceil e \geq \frac{n_{rw}}{\beta} t^{r}_{abc} - n_{rw} e + \pi \ . \tag{10}$$

From the above we obtain

$$\begin{aligned} \frac{n_{rw}}{\beta} t^{r}_{abc} - n_{rw} e + \pi \leq \left\lceil \frac{n_r}{N} \right\rceil e \leq n_r e - t^{r}_{abc} + \pi \\ \left(\frac{n_{rw}}{\beta} + 1 \right) t^{r}_{abc} \leq \left\lceil \frac{n_r}{N} \right\rceil e - \pi \leq ne \ . \end{aligned} \tag{11}$$

Note that if $n_r = 0$, then this equation is false ($t^{r}_{abc} > 0$). □

.2 Transactional Replication

.2.1 Proofs

Lemma 3. *A service replicated on N servers, each having c-processor cores, using the optimized TR runs $n_r \frac{t^{o}_{abc}}{\beta}$ faster than when using the unoptimized TR.*

Proof. The proof is straightforward since the only difference between the optimized and unoptimized TR with hardware restricted parallelism is that the former does not broadcast read-only

requests. We define speedup as:

$$
\begin{aligned}
&speedup^{n}_{SM^{opt}} = T^{1|..|n}_{TR_{lower}} - T^{1|..|n}_{TR^{opt}_{lower}} \\
&= T^{1}_{TR} + max\left(\left\lceil\frac{n+K}{Nc}\right\rceil e' + \Sigma, \frac{n+K'}{\beta}t^{o}_{abc}\right) - \delta_{TR} \\
&- T^{1}_{TR} + max\left(\left\lceil\frac{n+K}{Nc}\right\rceil e' + \Sigma, \frac{n_{rw}+K'}{\beta}t^{o}_{abc}\right) - \delta_{TR} \\
&= \frac{n+K'}{\beta}t^{o}_{abc} - \frac{n_{rw}+K'}{\beta}t^{o}_{abc} = (n-n_{rw})\frac{t^{o}_{abc}}{\beta} = n_r\frac{t^{o}_{abc}}{\beta}
\end{aligned}
\tag{12}
$$

where $e' = e + t_{cer}$. □

Below we give the main result for the optimized TR:

Lemma 4. *In the best case, a service replicated on N single-core processor servers ($N > 1$) using the TR scheme can be faster than the non-replicated service if*

$$
\left\lceil\frac{n+K}{N}\right\rceil e' + \Sigma \leq \frac{n_{rw}+K'}{\beta}t^{o}_{abc} \leq ne - e' \tag{13}
$$

when abcast is dominant, and

$$
\frac{n_{rw}+K'}{\beta}t^{o}_{abc} \leq \left\lceil\frac{n+K}{N}\right\rceil e' + \Sigma \leq ne - t^{o}_{abc} \tag{14}
$$

when request execution is dominant.

Proof. We compute below the overhead imposed by TR in the best case (the lower time bound) and for $c = 1$, and test the hypothesis that the overhead is not positive.

$$
\begin{aligned}
&overhead^{n}_{TR^{opt}} = T^{1|..|n}_{TR^{opt}_{lower}} - (q + \left\lceil\frac{n}{c}\right\rceil e + a) \\
&= max\left(\left\lceil\frac{n+K}{N}\right\rceil e' + \Sigma, \frac{n_{rw}+K'}{\beta}t^{o}_{abc}\right) + \\
&+ t^{o}_{abc} + e' - \delta_{TR} - ne\ .
\end{aligned}
\tag{15}
$$

We first put forward the hypothesis considering abcast dominance:

$$
overhead^{n}_{TR^{opt}_{abc}} = \frac{n_{rw}+K'}{\beta}t^{o}_{abc} + e' - ne \leq 0 \tag{16}
$$

which implies that

$$
\begin{aligned}
&\frac{n_{rw}+K'}{\beta}t^{o}_{abc} \leq ne - e' \\
\text{iff}\quad &\frac{n_{rw}+K'}{\beta}t^{o}_{abc} \geq \left\lceil\frac{n+K}{N}\right\rceil e' + \Sigma\ .
\end{aligned}
\tag{17}
$$

From the above we obtain

$$
\left\lceil\frac{n+K}{N}\right\rceil e' + \Sigma \leq \frac{n_{rw}+K'}{\beta}t^{o}_{abc} \leq ne - e'\ . \tag{18}
$$

Note that in the above equation n_{rw} (and so t^{o}_{abc}) cannot be equal to 0 since we assumed abcast dominance.

Now let us put forward the hypothesis considering request execution dominance, i.e. when $\lceil \frac{n+K}{N} \rceil e' + \Sigma \geq \frac{n_{rw}+K'}{\beta} t^o_{abc}$:

$$\begin{aligned} overhead^n_{TR^{opt}_e} &= \left\lceil \frac{n+K}{N} \right\rceil e' + \Sigma + t^o_{abc} + e' - e' - ne \\ &= \left\lceil \frac{n+K}{N} \right\rceil e + \Sigma + t^o_{abc} - ne \leq 0 \end{aligned} \tag{19}$$

which implies that

$$\begin{aligned} &\left\lceil \frac{n+K}{N} \right\rceil e + \Sigma \leq ne - t^o_{abc} \\ &\quad \text{iff} \quad \left\lceil \frac{n+K}{N} \right\rceil e' + \Sigma \geq \frac{n_{rw}+K'}{\beta} t^o_{abc} \ . \end{aligned} \tag{20}$$

From the above we obtain

$$\frac{n_{rw}+K'}{\beta} t^o_{abc} \leq \left\lceil \frac{n+K}{N} \right\rceil e' + \Sigma \leq ne - t^o_{abc} \tag{21}$$

□

If equation (13) or (14) holds, then the TR-replicated service can be faster than a non-replicated service. In particular, if $n_{rw} = 0$, there are no conflicts ($K = K' = \Sigma = 0$) and no abcast ($t^o_{abc} = 0$). So we can reduce (14) to

$$0 \leq \left\lceil \frac{n_r}{N} \right\rceil (e + t_{cer}) \leq n_r e \tag{22}$$

Since normally $t_{cer} << e$, the above equation is mostly true, which agrees with the intuition that a TR-replicated service can be faster than a non-replicated service if there are many read-only requests that are processed in parallel. Note that if $n_r = 1$ or $N = 1$ then the above equation is false. If $n_r \leq N$ then the equation is true only if $n_r \geq \frac{e'+t_{cer}}{e}$.

.2.2 Transaction conflicts

Below we define a conflict function $f()$ for a few special cases.

Consider TR that tries to commit n transactions in parallel. Suppose that all transactions can be mutually conflicting (the worst case possible). In this case only the first committing transaction commits and the rest is rolled back and reexecuted. If the aborted transactions will be reexecuted sequentially (no overlap in time), then there will be no more conflicts. So $K = 1$ and $f(n-1) = n-1$. If the aborted transactions are reexecuted in pairs in parallel, then in each pair exactly one transaction commits and another is rolled back and reexecuted, so $K = n-1$ and $f(n-1) = n-1$. If the aborted transactions are reexecuted in parallel, then each time one commits and all but one are rolled back and reexecuted in parallel, so $K = \frac{(n-1)n}{2}$ and again $f(n-1) = n-1$. This is also an upper bound on the number of conflicts when processing n requests ($0 < K \leq \frac{(n-1)n}{2}$).

If the aborted transactions are executed in parallel but they are conflicting pairwise, then each time only half of transactions will be aborted and the rest can commit. So, we have $K = n-1$ and $f(\frac{n}{2}) = log_2(n) = log_2(K+1)$. In general, if transactions are conflicting k-wise, then at first only $\frac{n}{k}$ will commit and $n - \frac{n}{k}$ will be conflicting. Then we have $K = n(k-1)$ and $f(n - \frac{n}{k}) = log_k(n)$. Note that $log_k(n) < n-1$, so committing the conflicting transactions will take less time than before.

Consider the worst case possible, i.e. the conflict function $f(n') = n'$ where n' is the number of

transactions conflicting for the first time ($0 \leq n' < n$) and *none* of the conflicts are detected early. We obtain the bounds by substituting $f(n')$ by n' in the equations inferred in Section 2.2.2:

$$T^{1|..|n}_{TR_{upper}} \approx T^{1 \rightarrow n}_{TR} + n'(e' + t^o_{abc})$$
$$T^{1|..|n}_{TR_{lower}} \approx T^1_{TR} + max(n'e', \frac{n}{\beta} t^o_{abc}) + \frac{n'}{\beta} t^o_{abc} - \delta_{TR}$$

where

$$\delta_{TR} = \begin{cases} 0 & \text{if } max(n'e', \frac{n}{\beta} t^o_{abc}) = n'e' \\ t^o_{abc} & \text{if } max(n'e', \frac{n}{\beta} t^o_{abc}) = \frac{n}{\beta} t^o_{abc} \ . \end{cases}$$