A Methodological Construction of an Efficient Sequentially Consistent Distributed Shared Memory

Vicent CHOLVI° Antonio FERNÁNDEZ† Ernesto JIMÉNEZ‡ Pilar MANZANO‡
Michel RAYNAL⋆
° Universitat Jaume I, Castellón, Spain
† LADyR, GSyC, Universidad Rey Juan Carlos, 28933 Móstoles, Spain
‡ EUI, Universidad Politécnica de Madrid, 28031 Madrid, Spain
⋆ IRISA, Université de Rennes, Campus de Beaulieu, 35 042 Rennes, France
vcholvi@lsi.uji.es antonio.fernandez@urjc.es
ernes@eui.upm.es pmanzano@eui.upm.es raynal@irisa.fr

Abstract

A concurrent object is an object that can be concurrently accessed by several processes. Sequential consistency is a consistency criterion for such objects. Informally, it states that a multiprocess program executes correctly if its results could have been produced by executing that program on a single processor system. (Sequential consistency is weaker than atomic consistency -the usual consistency criterion- as it does not refer to real-time.) The paper proposes a simple protocol that ensures sequential consistency when the shared memory abstraction is supported by the local memories of nodes that can communicate only by exchanging messages through reliable channels. Differently from other sequential consistency protocols, the proposed protocol does not rely on a strong synchronization mechanism such as an atomic broadcast primitive or a central node managing a copy of every shared object. From a methodological point of view, the protocol is built incrementally starting from the very definition of sequential consistency. It has the noteworthy property of providing fast write operations (i.e., a process has never to wait when it writes a new value in a shared object). According to the current local state, most read operations are also fast.

Keywords: Distributed Systems, Distributed Shared Memory, Sequential Consistency.

I. INTRODUCTION

The definition of a consistency criterion is crucial for the correctness of a multiprocess program [9]. Basically, a consistency criterion defines which value has to be returned when a read operation on a shared object is invoked by a process. The strongest (i.e., most constraining) consistency criterion is *atomic consistency* [17] (also called *linearizability* [11]). It states that a read returns the value written by the latest preceding write, "latest" referring to real-time occurrence order (concurrent writes being ordered). *Causal consistency* [3], [5] is a weaker criterion stating that a read does not get an overwritten value. Causal consistency allows concurrent writes; consequently, it is possible that concurrent read operations on the same object get different values (this occurs when those values have been produced by concurrent writes). Other consistency criteria (weaker than causal consistency) have also been proposed [1], [23].

This paper focuses on *sequential consistency* [14]. This criterion lies between atomic consistency and causal consistency. Informally, it states that a multiprocess program executes correctly if its results could have been produced by executing that program on a single processor system. This means that an execution is correct if we can totally order its operations in such a way that

1) The order of operations in each process is preserved, and
2) Each read operation obtains the latest previously written value, "latest" referring here to the total order.

This work has been partially supported by the Spanish MEC under grants TIN2005-09198-C02-01, TIN2004-07474-C02-01, TIN2004-07474-C02-02, and TSI2004-02940. A preliminary version of this paper was published in [12]

The difference between atomic consistency and sequential consistency lies in the meaning of the word "latest". This word refers to real-time when we consider atomic consistency, while it refers to a logical time notion when we consider sequential consistency (namely the logical time defined by the total order). The main difference between sequential consistency and causal consistency lies in the fact that (like atomic consistency) sequential consistency orders all write operations, while causal consistency does not require to order concurrent writes.

Atomic consistency is relatively easy to implement in a distributed message-passing system. Each process p_i maintains in a local cache the current value v of each shared variable x, and such a cached value v is systematically invalidated (or updated) each time a process p_j writes x. The conflicts due to multiple accesses to a shared variable x are usually handled by associating a manager M_x with every shared variable x. One of the most known atomic consistency protocols is the invalidation-based protocol due to Li and Hudak [15] that has been designed to provide a distributed shared memory on top of a local area network. An update-based atomic consistency protocol is described in [8].

Due to its very definition, atomic consistency requires that the value of a variable x cached at p_i be invalidated (or updated) each time a process p_j issues a write on x. In that sense, the atomic consistency criterion (that is an abstract property of a computation) is intimately related to an *eager* invalidation (or update) mechanism (that concerns the operational side). Said in another way, atomic consistency is a consistency criterion that can be too *conservative* for some applications.

Differently, sequential consistency can be seen as a form of *lazy* atomic consistency [21]. A cached value has not to be systematically invalidated each time the corresponding shared variable is updated. Old and new values of a shared variable can coexist at different processes as long as the resulting execution could have been produced by running the multiprocess program on a single multiprogrammed processor system. Of course, a protocol implementing sequential consistency can be more involved than a protocol implementing atomic consistency, as it has to keep track of global information allowing it to know, for each process p_i, which old values currently used by p_i have to be invalidated (or updated) and which ones do not have to. This global information tracking, which is at the core of sequential consistency protocols, is the additional price that has to be paid to replace eager invalidation by lazy invalidation, thereby providing the possibility for more efficient runs of multiprocess programs.

Related work: Sequential consistency protocols

Several protocols providing a sequentially consistent shared memory abstraction on top of an asynchronous message passing distributed system have been proposed. The protocol described in [2] implements a sequentially consistent shared memory abstraction on top of a physically shared memory and local caches. It uses an atomic n-queue update primitive. Attiya and Welch [7] present two sequential consistency protocols. Both protocols assume that each local memory contains a copy of the whole shared memory abstraction. They order the write operations using an atomic broadcast facility: all the writes are sent to all processes and are delivered in the same order by each process. Read operations issued by a process are appropriately scheduled to ensure their correctness.

The protocol described in [19] considers a server site that has a copy of the whole shared memory abstraction. The local memory of each process contains a copy of a shared memory abstraction, but the state of some of its objects can be invalid. When a process wants to read an object, it reads its local copy if it is valid. When a process wants to read an object whose state is invalid, or wants to write an object, it sends a request to the server. In that way the server orders all write operations. An invalidation mechanism ensures that the reading by p_i of an object that is locally valid is correct. A variant of this protocol is described in [4]. The protocol described in [20] uses a token that orders all write operations and piggybacks updated values. This protocol, like one of the protocols described in [7], provides fast (i.e., purely local) read operations [10][1].

[1]As shown in [7] atomic consistency does not allow protocols in which all read operations (or all write operations) are fast [11], [18]. Differently, causal consistency allows protocols where all operations are fast [3], [5], [22].

Most of the previous protocols rely on a strong synchronization mechanism that has a scope spanning the whole system (atomic broadcast facility, navigating token, or central manager[2]). Differently, the protocol described in [21] is fully distributed in the sense that it does not rely on an underlying global mechanism: each object x is managed by its own object manager M_x and there is no synchronization primitive whose scope is the entire system.

Content of the paper

This paper presents a methodological construction of a sequential consistency protocol. A variant of this protocol has first been presented in [13] as a dynamically adaptive and parameterized algorithm that implements sequential consistency, cache consistency or causal consistency, according to the setting of some parameter. This parameterized algorithm is presented "from scratch", without exhibiting or relying on basic underlying principles. Here, it is shown that a variant of its sequential consistency instantiation can be obtained from a simple derivation starting from the very definition of sequential consistency.

The algorithm obtained here from this derivation not only is surprisingly simple, but -as it is based on the very essence of sequential consistency- it reveals to be particularly efficient for some classes of applications. The protocol has the nice property to allow the write operations to be fast, i.e., a write operation is always executed locally without involving global synchronization. Differently, some read operations can be fast, while others cannot. The fact that a read operation is fast or not depends on the variable that is read and the set of variables that have been previously written by the process issuing the read operation, so it is context-dependent.

The derived algorithm has been implemented and used to run typical parallel programming applications, namely Finite Differences, Matrix Multiplication, and Fast Fourier transform, in a cluster of workstations. The performance of this implementation of the algorithm in this context has been compared with implementations of the sequential consistency algorithms proposed by Attiya and Welch [7]. The results of this comparison show that the former implementation runs faster and requires smaller number of messages (by two orders of magnitude) than both the latter. Furthermore, unlike with the algorithms from [7], in most cases a large majority of the messages carry information about written values.

The paper is made up of five sections. Section II presents the computation model, and defines sequential consistency. Then, Section III derives the protocol from the sequential consistency definition. Finally, Section IV concludes the paper.

II. The Sequentially Consistent Shared Memory Abstraction

A parallel program defines a set of processes interacting through a set of concurrent objects. This set of shared objects defines a *shared memory abstraction*. Each object is defined by a sequential specification and provides processes with operations to manipulate it. When it is running, the parallel program produces a concurrent system [11]. As in such a system an object can be accessed concurrently by several processes, it is necessary to define consistency criteria for concurrent objects.

A. Shared Memory Abstraction

A shared memory system is composed of a finite set of sequential processes $p_1, \ldots, p_n$ that interact via a finite set X of shared objects. Each object $x \in X$ can be accessed by read and write operations. A write into an object defines a new value for the object; a read allows to obtain a value of the object. A write of value v into object x by process p_i is denoted $w_i(x)v$; similarly a read of x by process p_j is denoted $r_j(x)v$ where v is the value returned by the read operation; op will denote either r (read) or w (write). To simplify the analyses, as in [3], [17], [22], we assume that all values written into an object x are distinct[3]. Moreover, the parameters of an operation are omitted when they are not important. Each

[2]E.g., an atomic broadcast facility allows ordering all the write operations, whatever the processes that issue them.

[3]Intuitively, this hypothesis can be seen as an implicit tagging of each value by a pair composed of the identity of the process that issued the write plus a sequence number. Such a tagging is only conceptual and not required for the correctness of the algorithm.

object has an initial value (it is assumed that this value has been assigned by an initial fictitious write operation).

B. Programs, Histories and Legality

A *program* is a set of read and write operations to be issued by the processes that form the program. The *local program* of process p_i is the set of operations to be issued by p_i. If $op1$ and $op2$ are going to be issued by p_i and $op1$ is going to be issued first, then we say that "$op1$ precedes $op2$ in p_i's process-order", which is denoted $op1 \rightarrow_i op2$. Note that nothing has been said about the read or written values, nor about the order between operations from different processes.

In order to model concrete executions of programs, we introduce the concept of history. The *local history* (or local computation) $\widehat{h}_i$ of p_i is the sequence of operations issued by p_i in process order such that each operation has an associated (read or written) value. If h_i denotes the set of operations executed by p_i then $\widehat{h}_i$ is the total order $(h_i, \rightarrow_i)$.

Definition 1: An *execution history* (or simply history, or computation) $\widehat{H}$ of a shared memory system is a partial order $\widehat{H} = (H, \rightarrow_H)$ such that:

- $H = \bigcup_i h_i$
- $op1 \rightarrow_H op2$ if:

i) $\exists\ p_i\ : op1 \rightarrow_i op2$ (in that case, $\rightarrow_H$ is called *process-order* relation),
or *ii)* $op1 = w_i(x)v$ and $op2 = r_j(x)v$ (in that case $\rightarrow_H$ is called *read-from* relation),
or *iii)* $\exists op3\ : op1 \rightarrow_H op3$ and $op3 \rightarrow_H op2$.

Two operations $op1$ and $op2$ are *concurrent* in $\widehat{H}$ if we have neither $op1 \rightarrow_H op2$ nor $op2 \rightarrow_H op1$.

The legality concept is the key notion on which the definitions of shared memory consistency criteria are based [3], [5], [18], [23]. From an operational point of view, it states that, in a legal history, no read operation can get an overwritten value.

Definition 2: A read operation $r(x)v$ of a history $\widehat{H}$ is *legal* if:
i) $\exists\ w(x)v\ : w(x)v \rightarrow_H r(x)v$
and *ii)* $\nexists\ op(x)u\ : (u \neq v) \wedge (w(x)v \rightarrow_H op(x)u \rightarrow_H r(x)v)$.

A history $\widehat{H}$ is *legal* if all its read operations are legal.

C. Sequential Consistency

Sequential consistency was proposed by Lamport in 1979 to define a correctness criterion for multiprocessor shared memory systems [14]. A system is sequentially consistent with respect to a multiprocess program if *"the result of any execution is the same as if (1) the operations of all the processors were executed in some sequential order, and (2) the operations of each individual processor appear in this sequence in the order specified by its program."*

This informal definition states that the execution of a program is sequentially consistent if it could have been produced by executing this program on a single processor system[4]. More formally, we define sequential consistency in the following way. Let us first recall the definition of *linear extension* of a partial order. A linear extension $\widehat{S} = (S, \rightarrow_S)$ of a partial order $\widehat{H} = (H, \rightarrow_H)$ is a topological sort of this partial order. This means we have the following:
i) $S = H$,

[4]In his definition, Lamport assumes that the *process-order* relations defined by the program (point 2 of the definition) is maintained in the equivalent sequential execution, but not necessarily in the execution itself. As we do not consider programs but only executions, we implicitly assume that the *process-order* relations displayed by the execution histories are the ones specified by the programs which gave rise to these execution histories.

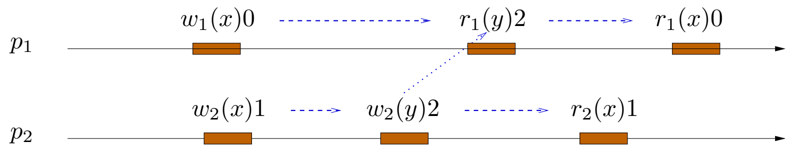


Fig. 1. A sequentially consistent execution $\widehat{H}$. Transitivity edges come from *process–order* relations (represented by dashed arrows) and *read–from* (represented by dotted arrows) relations. Only the edges that are not due to transitivity are shown.

ii) $op_1 \rightarrow_H op_2 \Rightarrow op_1 \rightarrow_S op_2$ ($\widehat{S}$ maintains the order of all ordered pairs of $\widehat{H}$) and
iii) $\rightarrow_S$ defines a total order.

Definition 3: A history $\widehat{H}$ is *sequentially consistent* if it has a legal linear extension $\widehat{S}$. We also say that $\widehat{S}$ is a *base legal sequentially consistent history* of $\widehat{H}$.

As an example let us consider the history $\widehat{H}$ depicted in Figure 1. Each process has issued three operations on the shared objects x and y. The write operations $w_1(x)0$ and $w_2(x)1$ are concurrent. It is easy to see that $\widehat{H}$ is sequentially consistent by building a legal linear extension $\widehat{S}$ including first the operations issued by p_2 and then the ones issued by p_1.

III. The Methodological Construction

The aim of this work is to implement a sequentially consistent shared memory abstraction on top of an underlying message-passing distributed system. Such a system is a distributed system made up of n reliable sites, one per process (hence, without ambiguity, p_i denotes both a process and the associated site). Each p_i has a local memory. The processes communicate through reliable channels by sending and receiving messages. There are no assumptions neither on process speed, nor on message transfer delay. Hence, the underlying distributed system is reliable but asynchronous.

A. The Methodology

The usual approach to design sequential consistency protocols consists in first defining a protocol and then proving it is correct. The approach adopted here is different, in the sense that we start from the very definition of sequential consistency, and *derive* from it a sequential consistency protocol.

More precisely, to ensure that a distributed execution has a base legal sequentially consistent history, we:

1) First define a base legal sequentially consistent history $\widehat{S}$, and
2) Then, design a protocol that controls the execution of the multiprocess program in order to produce an actual distributed execution $\widehat{H}$ that has $\widehat{S}$ as a base legal sequentially consistent history.

The first subsection that follows derives a trivial sequential consistency protocol that works for a very particular type of multiprocess programs; these particular multiprocess programs have the nice property that all operations can be executed locally. Then, by observing that the history of each sequentially consistent process can be decomposed in segments such as those considered in the previous type of multiprocess programs, a new sequential consistency protocol is derived that works for the general case. Finally, the last subsection shows how to enhance such a general protocol in order to achieve higher efficiency.

B. Step 1 of the Construction: The Trivial Case

Let us start with a multiprocess program where the local program of each process p_i has the following very particular structure. Namely, it is formed by a (possibly empty) sequence containing only read operations (denoted SR_i), followed by a (possibly empty) sequence of write and read operations (denoted SWR_i) such that the read operations are issued only on variables that have been previously written by p_i.

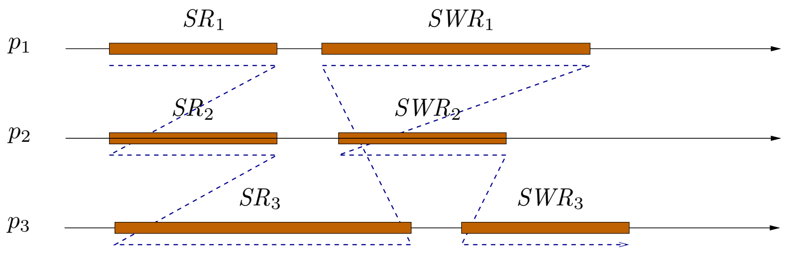


Fig. 2. Example of a program execution $\widehat{H}$ that is "trivially" sequentially consistent. Reads are assumed to return the closest previously written value, by the same process, in the corresponding variable (or the initial value, if it has not been written yet). The ordering of the base legal sequentially consistent history $\widehat{S}$ is indicated with the dashed arrow.

```
**init:**
    **for each** $x \in X$ **do**
        $C_i[x] \leftarrow$ initial value of $x$; % $C_i[x]$ denotes a local cache associated with each shared variable $x$ %
    **end do**

**operation** $w_i(x)v$:
    $C_i[x] \leftarrow v$;
    **return**()

**operation** $r_i(x)$:
    **return**$(C_i[x])$
```

Fig. 3. Trivial Case Protocol for process p_i.

Consider a concrete execution $\widehat{H}$ of such a program, produced by executing sequentially the SR_i sequences in any order, and then the SWR_i sequences also in any order, and make each read operation return the closest previously written value, by the same process, in the corresponding variable (or the initial value, if it has not written any value). Since the SR_i sequences contain only read operations that obtain the initial values of the shared variables and the read operations in the SWR_i sequences read only variables previously written by process p_i, from the very definition of sequential consistency, it is immediate to find a base legal sequentially consistent history $\widehat{S}$ of $\widehat{H}$. Namely,

$$\widehat{S} = SR_1 \ \ldots \ SR_n \ SWR_1 \ \ldots \ SWR_n.$$

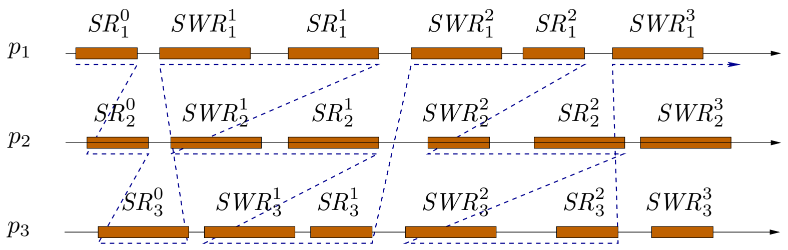
Figure 2 shows an example of a parallel execution of a program made of $n = 3$ processes, as described in the above paragraph. Our goal is now to design a protocol that ensures that any history of the multiprocess programs considered, is of the previously defined form (i.e., has $\widehat{S}$ as a base legal sequentially consistent history).

Implementation of the Trivial Case Protocol: From the above presented reasoning, it follows that an implementation would simply consist on providing each process p_i with a local cache containing all the shared variables and performing both reads and writes locally. Clearly, all the resulting histories will have the above presented $\widehat{S}$ as a base legal sequentially consistent history. Consequently, no additional protocol would be necessary. Figure 3 shows an implementation of the described protocol.

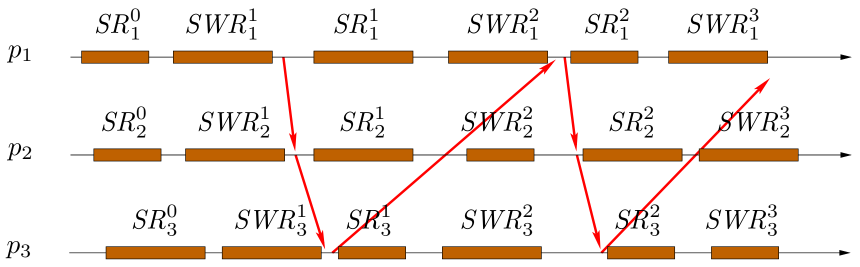
C. Step 2 of the Construction: The General Case (Looking for Correctness)

Let us first observe that, in the general case, the local program of p_i (for each process) can always be expressed as

$$SR_i^0 \ SWR_i^1 \ SR_i^1 \ SWR_i^2 \ SR_i^2 \ \ldots \ SWR_i^k \ SR_i^k \ \ldots$$


SR_1^0 SWR_1^1 SR_1^1 SWR_1^2 SR_1^2 SWR_1^3
p_1
SR_2^0 SWR_2^1 SR_2^1 SWR_2^2 SR_2^2 SWR_2^3
p_2
SR_3^0 SWR_3^1 SR_3^1 SWR_3^2 SR_3^2 SWR_3^3
p_3

(a) Example of a program's execution $\widehat{H}$ that is sequentially consistent. The ordering of the base legal sequentially consistent history $\widehat{S}$ is indicated with the dashed arrow. Reads are assumed to return the closest previously written value (according to $\rightarrow_S$), by any process, in its corresponding variable (or the initial value, if it has not been written yet).


SR_1^0 SWR_1^1 SR_1^1 SWR_1^2 SR_1^2 SWR_1^3
p_1
SR_2^0 SWR_2^1 SR_2^1 SWR_2^2 SR_2^2 SWR_2^3
p_2
SR_3^0 SWR_3^1 SR_3^1 SWR_3^2 SR_3^2 SWR_3^3
p_3

(b) Travel of the token in the implemented protocol. Observe that for the distributed execution $\widehat{H}$ to have $\widehat{S}$ as a base legal sequential history, the values carried by the token when it arrives at a process, say p_2, for the first time, have to be considered only if they have not been overwritten by SWR_2^1.

Fig. 4. Example of a general case program's execution that is sequentially consistent.

where SR_i^k is a (possibly empty) sequence of only read operations and SWR_i^k is a (possibly empty) sequence of write and read operations such that read operations are performed only on variables that have been previously written in SWR_i^k. The superscript k is used to associate a SR_i sequence with its immediately preceding SWR_i sequence.

The decomposition of each process history into sequences and the particular case of a single sequence examined in the previous step of the construction, provides us with some hint on how to proceed. Indeed, let us define a history $\widehat{S}$ formed first by the sequences $SR_1^0, SR_2^0, ..., SR_n^0$, in this order, and then by the sequences $\boxed{SWR_1^1\ SR_1^1}, \boxed{SWR_2^1\ SR_2^1}, ..., \boxed{SWR_n^1\ SR_n^1}$, in this order. Additional subsequent phases, similar to the second one, will take place until completing the execution. Also, make read operations to return the closest previously written value, by any process, in the corresponding variable (or the initial value, if it has not been written yet).

Clearly, for the definition of sequential consistency, $\widehat{S}$ will be a base legal sequentially consistent history of "some" of the histories of the general program. Figure 4(a) shows an example, in the case where there are $n = 3$ processes, of the parallel execution of a program $\widehat{H}$ that has the above defined history $\widehat{S}$ as a base legal sequential history.

Now, the goal is to design a sequential consistency protocol that ensures that "any" possible program execution has $\widehat{S}$ as a base legal sequentially consistent history.

Implementation of the General Case Protocol: For the design of the protocol, we observe that, in $\widehat{S}$, when p_{i+1} executes SR_{i+1}^1, it can read the value of a variable x that has been written by p_i when it executed SWR_i^1. Hence, p_{i+1} must be informed of these writes before it executes SR_{i+1}^1. A simple way to attain this goal consists of using a token travelling along a logical ring so that no process misses updates (e.g., p_1, $p_2, \ldots, p_n$, p_1) and carrying the latest known value of each shared variable. Therefore, we have to manage the token exactly as if it was received by p_{i+1} just after p_{i+1} executed SR_{i+1}^0 and was sent by p_{i+1} to p_{i+2} just after p_{i+1} terminated SR^1_{i+1}. Logically, the token follows the arrow in Figure 4(a), so that $\widehat{H}$ will have $\widehat{S}$ as a base legal sequentially consistent history. Then, in the algorithm to carry the new values written in SWR^1_i, the token has to be sent after SWR^1_i finishes. Moreover, as SR^1_i modifies no shared variables, the token can be sent by p_i before SR^1_i. So, when a process p_i receives the token, it ends a segment SWR^k_i, sends the token and starts a segment SR^k_i.

The resulting protocol is described in Figure 5. As already indicated, X denotes the set of shared variables, and $C_i[x]$ is p_i's local cache containing the value of the shared variable x. Each process p_i maintains a boolean array $updated_i$ such that $updated_i[x]$ is true iff p_i has updated x since the last visit of the token. The boolean no_change_i is a synonym for $\wedge_{x\in X}(\neg updated_i[x])$ (no_change_i is true iff no shared variable has been updated since the last visit of the token at p_i). The write operation and the statements associated with the token reception are executed atomically. (Let us observe that the arrival of the token at a process always corresponds to the beginning of a new segment SR^k_i for that process.)[5] Figure 4(b) shows the actual travel of the token with this algorithm in the example used in this Step. Observe that, in this protocol the token could be replaced by a list containing only the modifications. This "improvement", together with one dealing with the dissemination of updates, is incorporated in the algorithm presented in the next Step.

D. Step 3 of the Construction: The General Case (Looking for Efficiency)

When we look at the form of the sequences SR^j_i, as defined in the Step 2, we also observe that they can always be decomposed as follows:

$$SR^j_i = \begin{cases} SR^0_{i,1} \;\ldots\; SR^0_{i,i} & \text{when } j = 0 \\ SR^j_{i,i(mod\, n)+1} \;\ldots\; SR^j_{i,(i+n-1)(mod\, n)+1} & \text{when } j > 0 \end{cases}$$

Note that SR^0_i is decomposed in i sequences, whereas $SR^{j>0}_i$ is always decomposed in n sequences.

The rationale behind the form we have decomposed SR^j_i into n subsequences (except for the start–up phase, where it is split into i subsequences) can be explained as follows. By using such a decomposition, the goal is to allow a process p_i, during its sequence of reads in SR^j_i, to obtain the updated values as quickly as possible. Namely, those updates will take place at the beginning of each one of the $SR^j_{i,k}$ subsequences. With this in mind, in the new base legal sequentially consistent history $\widehat{S}$, the updated values within SWR^j_i will have to be "disseminated" to all processes at the same time (contrary to what is done at Step 2, where the updated values were disseminated sequentially). Clearly, this type of "eager" dissemination allows processes to be informed of new values earlier. Furthermore, this also allows processes to disseminate only their own updates (in the protocol in Step 2, the token accumulates all the updates), thus reducing the transfer of data between processes.

Therefore, by substituting the new decomposed sequences into the local history of p_i, we obtain the following,

$$\begin{array}{l}
SR^0_{i,1} \;\ldots\; SR^0_{i,i} \\
SWR^1_i \; SR^1_{i,i(mod\, n)+1} \;\ldots\; SR^1_{i,(i+n-1)(mod\, n)+1} \\
SWR^2_i \; SR^2_{i,i(mod\, n)+1} \;\ldots\; SR^2_{i,(i+n-1)(mod\, n)+1} \\
\ldots \\
SWR^j_i \; SR^j_{i,i(mod\, n)+1} \;\ldots\; SR^j_{i,(i+n-1)(mod\, n)+1} \\
\ldots
\end{array}$$

[5]The reader familiar with token-based termination detection protocols [16] can see that the protocol described in Figure 5 and these termination detection protocols share the same underlying mechanism combining token and flags (here, the flags no_change_i). The corresponding flags in a termination detection protocol are usually called $cont\_passive_i$, and are used to know if a process p_i stayed continuously passive between two consecutive visits of the token. This flag is set to *false* when p_i receives a message. It is reset to *true* when p_i owns the token, becomes passive and sends the token to its successor.

```
init:
    for each $x \in X$ do
        $C_i[x] \leftarrow$ initial value of $x$;
        $updated_i[x] \leftarrow false$;
    end do;
    $no\_change_i \leftarrow true$;
    The token (with initial values) is initially at $p_1$ that simulates its arrival at the end of $SWR_1^1$

operation $w_i(x)v$: % $w_i(x)v$ always belongs to some segment $SWR_i^z$ %
    $C_i[x] \leftarrow v$;
    $updated_i[x] \leftarrow true$;
    $no\_change_i \leftarrow false$;
    return()

operation $r_i(x)$:
    wait until $(no\_change_i \vee updated_i[x])$;
    % $no\_change_i \Rightarrow r_i(x) \in SR_i^z \wedge updated_i[x] \Rightarrow r_i(x) \in SWR_i^z$ %
    return $(C_i[x])$

Task $T_i$: (activated upon reception of $token[X]$)
    for each $x \in X$ such that $\neg updated_i[x]$ do
        $C_i[x] \leftarrow token[x]$;
    end do;
    for each $x \in X$ such that $updated_i[x]$ do
        $token[x] \leftarrow C_i[x]$;
        $updated_i[x] \leftarrow false$;
    end do;
    send $token[X]$ to the next process on the logical ring;
    $no\_change_i \leftarrow true$;
    % we have here: $\forall x \in X : \ updated_i[x] = false$ %
```

Fig. 5. General Case Protocol for process p_i.

Now, we define the form of the base legal sequentially consistent history $\widehat{S}$. To do that, first we order the different sequences of the local programs as follows:

$$
\begin{array}{l}
SR^0_{1,1}\ SR^0_{2,1}\ \ldots\ SR^0_{n,1} \\
SWR^1_1\ SR^1_{1,2}\ SR^0_{2,2}\ \ldots\ SR^0_{n,2} \\
SWR^1_2\ SR^1_{1,3}\ SR^1_{2,3}\ SR^0_{3,3}\ \ldots\ SR^0_{n,3} \\
\ldots \\
SWR^1_i\ SR^1_{1,i(mod\,n)+1}\ \ldots\ SR^1_{i,i(mod\,n)+1}\ SR^0_{i+1,i(mod\,n)+1}\ \ldots\ SR^0_{n,i(mod\,n)+1} \\
\ldots \\
SWR^1_n\ SR^1_{1,n(mod\,n)+1}\ SR^1_{2,n(mod\,n)+1}\ \ldots\ SR^1_{n,n(mod\,n)+1} \\
SWR^2_1\ SR^2_{1,1(mod\,n)+1}\ SR^1_{2,1(mod\,n)+1}\ \ldots\ SR^1_{n,1(mod\,n)+1} \\
SWR^2_2\ SR^2_{1,2(mod\,n)+1}\ SR^2_{2,2(mod\,n)+1}\ SR^1_{3,2(mod\,n)+1}\ \ldots\ SR^1_{n,2(mod\,n)+1} \\
\ldots \\
SWR^2_i\ SR^2_{1,i(mod\,n)+1}\ \ldots\ SR^2_{i,i(mod\,n)+1}\ SR^1_{i+1,i(mod\,n)+1}\ \ldots\ SR^1_{n,i(mod\,n)+1} \\
\ldots \\
SWR^2_n\ SR^2_{1,n(mod\,n)+1}\ SR^2_{2,n(mod\,n)+1}\ \ldots\ SR^2_{n,n(mod\,n)+1} \\
\ldots
\end{array}
$$

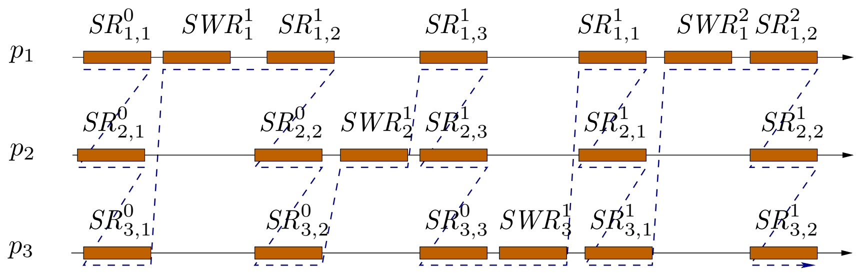
Then, we make read operations to return the closest previously written value, by any process, in the corresponding variable (or the initial value, if it has not been written yet). Clearly, for the definition of sequential consistency, $\widehat{S}$ will be a base legal sequentially consistent history of "some" of the histories of the general program. Figure 6(a) shows an example, in the case where there are $n = 3$ processes, of parallel execution of a program $\widehat{H}$ that has the above defined history $\widehat{S}$ as a base legal sequential history. Figure 6(b) also illustrates how the dissemination of writes is performed. Now, as in the previous cases, the goal is to design a sequential consistency protocol that ensures that "any" possible program execution has $\widehat{S}$ as a base legal sequentially consistent history.

Implementation of the Efficient General Case Protocol: The design of the protocol is based on the protocol in Step 2. However, in order to dissociate the two different roles of the token (namely, dissemination and gathering of updates), the token itself is replaced by the local variables $token_i$. $token_i = j$ means that, from p_i's point of view, p_j is the process that is currently allowed to disseminate updates. So, circulating the token around the logical ring $p_1, p_2, \ldots, p_n, p_1, \ldots$, is realized by having each $token_i$ variable taking successively the values $1, 2, \ldots, n, 1, \ldots$ Note that $token_i = i$ means that p_i (knows that it) has the token and is consequently allowed to disseminate updates.

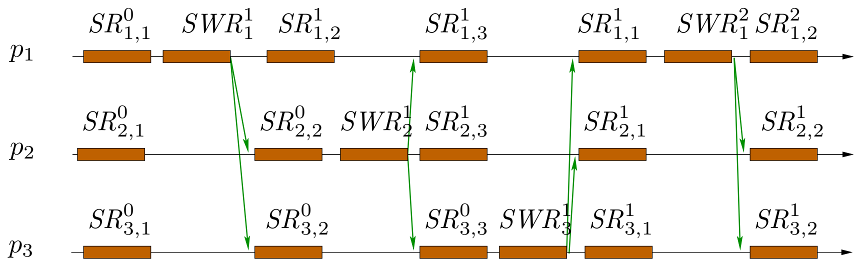
Additionally, the task associated with the management of the token in Figure 5 has changed significantly (see Figure 7). This task defines two distinct behaviors for a process p_i according to the token position. More precisely, when p_i has the token (case $token_i = i$), it is allowed to send to the rest of processes information about all the write operations (*updates*) it has executed since the previous visit of the token (lines 3-4). This set of updates upd is carried in the message UPDATES(upd). After broadcasting its updates, p_i resets its local control variables (lines 5-6).

When p_i does not have the token (case $token_i \neq i$), it waits for an UPDATES() message from the next process allowed to broadcast its updates (p_{token_i}). When it receives that message (line 8), p_i updates accordingly its local cache (as in the previous protocol, lines 9-10). This constitutes an early refreshing of its local cache with the new values provided by p_{token_i}.

Note that, for a process p_i, the token moves from p_j to p_{j+1} when, $token_i$ being equal to j, p_i executes $token_i \leftarrow (token_i \mod n) + 1$ (line 13). All the processes have the same view of the order in which the token visits the processes. Consequently, after it has received and processed an UPDATES() message from p_j, the process p_{j+1} knows that it has the token: no explicit message is necessary to represent the token.



(a) Example of a program's execution $\widehat{H}$ that is sequentially consistent. The ordering of the base legal sequentially consistent history $\widehat{S}$ is indicated with the dashed arrow. Reads are assumed to return the closest previously written value (according to $\rightarrow_S$), by any process, in the corresponding variable (or the initial value, if it has not been written yet).



(b) Eager dissemination of the updates.

Fig. 6. Example of an efficient general case program execution that is sequentially consistent.

It is important to notice that all the processes update their local caches (with the new values coming from the other processes) in the same order. This is an immediate consequence of the fact that each process p_i delivers the UPDATES() messages in the order defined by the successive values of $token_i$. As in the base token-based protocol, p_i's own updates are done at the time p_i issues the corresponding write operations and tracked with the boolean array $updated_i$. These boolean flags are used to maintain the consistency of p_i's local cache each time it receives and processes an UPDATES() message.

Evaluation of the Efficient General Case Protocol: In this section we show that in our efficient general case protocol most of the memory operations are performed in a fast manner. That is, they are mostly executed locally without involving global synchronization. This is clearly a nice property since they can be served almost immediately. An exact analytic evaluation of how many operations the protocol allows to be fast is not feasible, since it depends on the read/write application pattern. Hence, we have used real benchmark implementations to estimate the number of fast operations.

In order to evaluate the performance of our protocol with real applications, we have implemented it in a real scenario[6]. We have chosen as final applications three typical parallel processing applications: Finite Differences (FD) [24], Matrix Multiplication (MM) [24], and Fast Fourier Transform (FFT) [6]. The experiments have used the following data sizes:

1) Finite Differences (FD) with $16,384 \times 1,024$ elements.
2) Matrix Multiplication (MM) with square matrices of $1,600 \times 1,600$.
3) Fast Fourier Transform (FFT) with $262,144$ coefficients.

The executions have been conducted in an experimental environment formed by a cluster of 2, 4 and 8 computers connected with a switched full-duplex 1 Gbps Ethernet network. Each computer is a PC running Linux Red-Hat with a 1.5 GHz AMD CPU, and 512 Mbytes of RAM memory. We have mapped

[6]The source code can be found at *http://luna.dat.escet.urjc.es/˜ernes*.

```
init:
    for each $x \in X$ do
        $C_i[x] \leftarrow$ initial value of $x$;
        $updated_i[x] \leftarrow false$;
    end do;
    $no\_change_i \leftarrow true$;
    $token_i \leftarrow 1$;

operation $w_i(x)v$: % $w_i(x)v$ always belongs to some segment $SWR_i^z$ %
    $C_i[x] \leftarrow v$;
    $updated_i[x] \leftarrow true$;
    $no\_change_i \leftarrow false$;
    return()

operation $r_i(x)$:
    wait until $(no\_change_i \vee updated_i[x])$;
    % $no\_change_i \Rightarrow r_i(x) \in SR_i^z \wedge updated_i[x] \Rightarrow r_i(x) \in SWR_i^z$ %
    return $(C_i[x])$

Task $T_i$:
 (1)  loop
 (2)      case $(token_i = i)$ then
 (3)              $upd = \{(x, C_i[x]) \mid updated_i[x]\}$;
 (4)              for each $j \neq i$ do send UPDATES($upd$) to $p_j$ end do;
 (5)              for each $(x, v_x) \in upd$ do $updated_i[x] \leftarrow false$ end do;
 (6)              $no\_change_i \leftarrow true$;
 (7)           $(token_i \neq i)$ then
 (8)              wait (UPDATES($upd$) from $token_i$);
 (9)              for each $(x, v_x) \in upd$ do
 (10)                 if $(\neg updated_i[x])$ then $C_i[x] \leftarrow v_x$ end if
 (11)             end do;
 (12)     end case;
 (13)     $token_i \leftarrow (token_i \mod n) + 1$;
 (14) end loop
```

Fig. 7. Efficient General Case Protocol for process p_i.

one process to each computer and have restricted our implementation to a maximum of 100 memory operations carried in one single message.

In Table I the percentages of observed fast read and fast write operations per process are shown. As it can be readily seen, all write operations are fast, while in all cases, almost 100% of the read operations are fast. This makes evident that the main goal of our protocol (i.e., to be fast) is certainly achieved.

IV. Conclusion

This paper has presented a new sequential consistency protocol. Differently from the previous protocols we are aware of, this one has been derived from the very definition of the sequential consistency criterion. Due to its design principles, the protocol we have obtained is particularly simple. It provides fast write operations: these operations are always executed "locally" (i.e., without requiring any form of global synchronization). Read operations can also be fast when they are on a variable that has just been previously

	2 nodes			4 nodes			8 nodes		
Operations	MM	FD	FFT	MM	FD	FFT	MM	FD	FFT
Reads	99.21%	99.57%	99.46%	99.99%	99.82%	99.95%	99.99%	99.87%	99.98%
Writes	100%	100%	100%	100%	100%	100%	100%	100%	100%

TABLE I

PERCENTAGE OF FAST READ AND WRITE OPERATIONS PER PROCESS

updated by the same process. The proposed protocol is very efficient in terms of achieving a high rate of fast memory operations. Finally, we note that it is possible, from an engineering point of view, to adapt the globally efficient protocol to particular environments. A simple adaptation would consist in allowing some processes p_i to keep the token for some time when they have it. The benefit of such a possibility depends on the read/write access pattern of the upper layer application program.

REFERENCES

[1] Adve S.V. and Garachorloo K., Shared Memory Models: a Tutorial. *IEEE Computer*, 29(12):66-77, 1997.

[2] Afek Y., Brown G. and Merritt M., Lazy Caching. *ACM Transactions on Programming Languages and Systems*, 15(1):182-205, 1993.

[3] Ahamad M., Hutto P.W., Neiger G., Burns J.E. and Kohli P., Causal memory: Definitions, Implementations and Programming. *Distributed Computing*, 9:37-49, 1995.

[4] Ahamad M. and Kordale R., Scalable Consistency Protocols for Distributed Services. *IEEE Transactions on Parallel and Distributed Systems*, 10(9):888-903, 1999.

[5] Ahamad M., Raynal M. and Thia-Kime G., An Adaptive Protocol for Implementing Causally Consistent Distributed Services. *Proc. 18th IEEE Int. Conf. on Distributed Computing Systems*, IEEE Computer Society Press, pp. 86-93, Amsterdam (Netherland), 1998.

[6] Akl S.G., The design and analysis of parallel algorithms. *Prentice-Hall*, 1989.

[7] Attiya H. and Welch J.L., Sequential Consistency versus Linearizability. *ACM Transactions on Computer Systems*, 12(2):91-122, 1994.

[8] Bal H. and Tanenbaum A.S., ORCA: a Language for Parallel Programming of Distributed Systems. *IEEE Transactions on Software Engineering*, 18(3):190-205, 1992.

[9] Cholvi V. and Bernabéu J., Relationships between Memory Models. *Information Processing Letters*, 90(2):53–58, 2004.

[10] Herlihy M.P., Wait-Free Synchronization. *ACM Transactions on Programming Languages and Systems*, 13(1):124-149, 1991.

[11] Herlihy M.P. and Wing J.L., Linearizability: a Correctness Condition for Concurrent Objects. *ACM Transactions on Programming Languages and Systems*, 12(3):463-492, 1990.

[12] Cholvi V., Fernandez A., Jimenez E. and Raynal M., A Methodological Construction of an Efficient Sequential Consistency Protocol. *IEEE International Symposium on Network Computing and Applications*, August 2004.

[13] Jimenez E., Fernandez A. and Cholvi V., A parameterized Algorithm that Implements Sequential, Causal and Cache Consistency. *Proc. 10th EUROMICRO Workshop on Parallel, Distributed and Network-Based Processing (PDP'02)*, Islas Canarias (Spain), 2002.

[14] Lamport L., How to Make a Multiprocessor Computer that Correctly Executes Multiprocess Programs. *IEEE Transactions on Computers*, C28(9):690-691, 1979.

[15] Li K. and Hudak P., Memory Coherence in Shared Virtual Memory Systems. *ACM Transactions on Computer Systems*, 7(4):321-359, 1989.

[16] Mattern F., Algorithms for Distributed Termination Detection. *Distributed Computing*, 2:161-175, 1987.

[17] Misra J., Axioms for Memory Access in Asynchronous Hardware Systems. *ACM Transactions on Programming Languages and Systems*, 8(1):142-153, 1986.

[18] Mizuno M., Nielsen M.L. and Raynal M., An Optimistic Protocol for a Linearizable Distributed Shared Memory System. *Parallel Processing Letters*, 6(2):265-278, 1996.

[19] Mizuno M., Raynal M. and Zhou J.Z., Sequential Consistency in Distributed Systems. *Proc. Int. Workshop on Theory and Practice of Distributed Systems*, Springer Verlag LNCS #938, pp. 224-241, Dagsthul Castle (Germany), 1994.

[20] Raynal M., Token-Based Sequential Consistency. *Int. Journal of Computer Systems Science and Engineering*, 17(6):359-366, 2002.

[21] Raynal M., Sequential Consistency as Lazy Linearizability. *Proc. 14th ACM Symposium on Parallel Algorithms and Architectures (SPAA'02)*, pp. 151-152, Winnipeg, 2002.

[22] Raynal M. and Schiper A., From Causal Consistency to Sequential Consistency in Shared Memory Systems. *Proc. 15th Int. Conf. on Foundations of Software Technology and Theoretical Computer Science (FST&TCS'95)*, Springer-Verlag LNCS #1026, pp. 180-194, Bangalore (India), 1995.

[23] Raynal M. and Schiper A., A Suite of Formal Definitions for Consistency Criteria in Distributed Shared Memories. *Proc. 9th Int. IEEE Conference on Parallel and Distributed Computing Systems (PDCS'96)*, IEEE Computer Society Press, pp. 125-131, Dijon (France), 1996.

[24] Wilkinson B. and Allen M., Parallel Programming: Techniques and Applications using Networked Workstations and Parallel Computers. *Prentice-Hall*, 1999.